5-2024

Predictive Analytics: Assessing Air Pollution’s Influence on Real Estate Prices in Dubai

Hani Khalaf
hk9090@rit.edu

Predictive Analytics: Assessing Air Pollution's Influence on Real Estate Prices in Dubai

by

Hani Khalaf

A Thesis Submitted in Partial Fulfilment of the Requirements for the Degree of Master of Science in Professional Studies: Data Analytics

Department of Graduate Programs & Research

Rochester Institute of Technology

May 2024

RIT

Master of Science in Professional Studies: Data Analytics

Graduate Thesis Approval

Student Name**: Hani Khalaf**

Thesis Title**: Predictive Analytics: Assessing Air Pollution's Influence on Real Estate Prices in Dubai**

Graduate Committee:

Name: **Dr. Sanjay Modak** **Date:**

Chair of committee

Name: **Dr Ioannis Karamitsos** **Date:**

Member of committee

Acknowledgments

I would like to thank all the people who contributed to the work described in this thesis. First and foremost, I thank my academic advisor, Dr. Ioannis Karamitsos, for accepting to supervise my work on this topic, engaging me in new ideas, and guiding me throughout the process.

I would like to thank my family for their support and patience throughout the journey and their continuous encouragement.

Finally, I would like to thank my fellow students in cohort 9, the A-Team, who made this journey fun and enjoyable.

Abstract

This study examines the link between air pollution and property prices in Dubai. It aims to find the relationship between environmental and economic factors in Dubai's rapidly expanding real estate market and concerns about city livability.

The research is centered on investigating how differences in air quality between various neighborhoods in Dubai affect property pricing. To achieve this goal, the thesis adopted the Hedonic Pricing Method (HPM) typical for this type of analysis. The Hedonic Pricing Method (HPM) provides insights on the price of a product from its components. The study utilized records of housing transactions and air quality measurements from 2021 to 2023 obtained from open data portals provided by the government of Dubai.

The methodology adopted in the study is based on linking the house sales transactions with the environmental data from the nearest air quality station. The Air Quality Index (AQI) was adopted as the measure of air quality. The findings of this research reveal that there is no quantifiable relationship between air quality and property values in Dubai.

Several conclusions were drawn from the study. These include the need for more comprehensive datasets, both on the house attributes and the air quality data, to enhance the accuracy of the models.

As future research, the thesis recommends including macroeconomic and local investment factors as data sources affecting the price of property in Dubai.

Keywords: Dubai, hedonic pricing method, air quality, urban planning

Table of Contents

List of Figures

List of Tables

Chapter 1. Introduction

Background

"A city's Environmental, Social and Governance (ESG) profile is fundamental in building a sustainable, inclusive, and resilient future" – International Organization for Standardization.

Most cities world-wide suffer from serious air-quality problems, which have received increasing attention in the past decade (Mayer, 1999). Many cities expanded in unplanned and chaotic ways, where industrial and residential areas were mixed without considering the long-term effects on the livability of the city. As urbanization continued and people had many choices on where to settle, cities realized that they needed to attract business and people to continue to thrive. Competition between cities around the world started to emerge, and one of the major factors that new comers started looking at is the city livability index, a measure of how comfortable and pleasant a city is for its residents to live in, based on broad categories of stability, healthcare, culture and environment, education and infrastructure.

Affordability and availability of housing as well as health and wellbeing are key factors affecting the city livability index. Air quality and pollution have a direct effect on the citizens' health and their well-being. Citizens naturally have a desire to live in areas which are perceived to have less pollution. Consequently, they might be willing to pay higher prices for such properties. Analyzing the relation between air pollution and the price of property in the neighborhood will inform us on two sides: whether there is a relation between property price and air quality in the neighborhood, and second how much extra are people willing to pay for a property in a neighborhood perceived to have less pollution.

In Dubai, the property market has flourished in the past two years with individuals and families taking up residency in the city (Waters, 2023). Many of them look to buy property instead of renting as they are settling in their new home. The Government of Dubai Media office states that "*Dubai's annual real estate transactions have crossed the milestone of* *half a trillion dirhams for the first time in 2022. Maintaining its exponential growth trajectory, the sector witnessed transactions worth a record AED528 billion in 2022, a 76.5% increase from 2021.*" (Dubai Media Office, 2023).

Many studies have been done on the relationship between outdoor air quality and the price of real estate in cities around the world. The first studies on this topic was carried out as early as 1967 (Nourse H. O., 1967). These studies found a direct correlation between air quality and real estate prices: as the area becomes more polluted i.e. air quality decreases, the demand on property decreases and hence price of property decreases.

Cities and municipalities use Internet of Things (IoT) systems to collect information about air quality in neighborhoods. These air quality stations are equipped with sensors that can detect certain pollutants in the air. These pollutants include: Carbon monoxide, ozone, nitrogen dioxide, sulfur dioxide, the concentration of particulates (PM10) and the concentration of particulates (PM2.5). The first four parameters are measured in P.P.M. (parts per million) and the latter two are measured in $\mu g/m^3$ of air. Air Quality Index (AQI), developed by the United States Environmental Protection Agency (USEPA), is a standard metric used to inform the public on the status of contamination of the atmosphere, based on values of the pollutants in the air (Horn & Dasgupta, 2023). The air quality data from Dubai Municipality serves as the first source of data for the thesis.

A smart city uses open data as a catalyst for innovation (Walravens, Breuer, & Ballon, 2014). Open data refers to the publishing of governmental data on public portals so that it's available to interested parties to analyze. As a smart city, Dubai publishes real estate data for all transactions happening in the city. This data serves as a second data source for our thesis.

The results of this thesis could serve as an important input for the city's policy makers in planning their activities to improve air quality in the different areas of Dubai, such as increasing green spaces to improve air quality, relocating existing industries to other parts of the city, or selecting the location of new industrial facilities, which in turn will improve the city livability index.

Problem Statement

With the unprecedented urban growth, the impact of air pollution is becoming one of the most important topics affecting the quality of life in cities. The impact of air pollution on housing prices has emerged as a significant area of study. Dubai has been a hub for real estate investment for more than two decades and the housing demand surged after the 2020 pandemic, where the city has seen a big influx of people wanting to settle in the city. Despite the recognition of the harmful effects of air pollution on wellbeing and living standards, there remains a conspicuous gap in research examining the direct correlation between air pollution and real estate prices in Dubai.

Property price can vary significantly in a city. The price of the property could be affected by many parameters including those related to the property itself like the square foot area of the property, number of bedrooms, number of floors, number of bathrooms, usable space, age of the property, and other external factors related to mortgage rates, neighborhood location, closeness to the transportation network among others. Air quality in the neighborhood has been identified by many research studies as a potential factor affecting property price. Research around the world has been done to study this relationship.

This research aims to study the relation of air quality in Dubai with the house prices and create a prediction model for estimation of house prices. The results of the research could provide insights to several stakeholders, from homeowners and new investors to policy makers, on the economic and environmental impact of air pollution on the housing market. This in turn will help Dubai progress in the direction of a data-driven city.

Research Aim and Objectives

The research aims to:

- Develop two regression models to predict the price of the property as a function of the property parameters, while studying the effect of each property on the price and evaluate the performance and accuracy of the prediction models.

The objective of this study is to investigate the relationship between air quality in Dubai and real estate prices and to create a predictive model for estimation real estate prices. The research findings could provide insights to several stakeholders, from homeowners and new investors to policy makers, on the economic and environmental impact of air pollution on the housing market. This, in turn, will help Dubai move towards becoming a data-driven city.

Research Questions

This research attempts to answer the following questions:

Primary research question:

- What is the relationship between air quality in a neighborhood in Dubai and the property price in that neighborhood?

Secondary research questions:

- Which parameters of a property have a significant effect on its price?
- What is the willingness to pay extra for a property that falls in a non-polluted neighborhood?
- What challenges and limitations are present in the Hedonic Price Method and how can future research be enhanced?

Limitations of the study

This study encountered several limitations which may have impacted the findings. These can be summarized as follows:

- The intention was to do the study on data that spans 5 years or more, as this is the norm from the study of previous literature. However, because of the 2020 pandemic, and the high possibility that the data of the year 2020, whether house sale transactions or air quality data, will not be representative of "normal conditions" data, the decision was made to exclude the year 2020 and start the study from the year 2021 onwards. The data available included the year 2021, 2022 and until April 2023.

- The quarterly air quality data from Dubai Municipality is available publicly for the year 2021 on a yearly basis. To maintain consistency for the air quality data, the air quality index (AQI) reading for each air quality station was considered on a yearly basis. If more granular air quality data is available, this could possibly increase the accuracy of the results.
- Dubai Municipality provides air quality data from 14 weather stations in Dubai. These stations are strategically located in the city but do not cover every neighborhood in the city. The housing transaction on neighborhood data is very granular, hence an association was made between the transaction and the "nearest" air quality station. The assumption was that the air quality data within a circle of 7km radius from the station covers all neighborhoods within the circle.
- Dubai Municipality does not have data on the Concentration of particulates (PM10) and the Concentration of particulates (PM2.5) pollutants for the year 2021, hence only the following pollutants available for the duration of the study were considered: Carbon Monoxide (CO), Ozone (O3), Nitrogen Dioxide (NO2) and Sulphur Dioxide (SO2).

Structure of the Thesis

The following provides guidance on the structure of the thesis document.

Chapter 1 provides a background on the topic of the thesis. It includes the problem that the thesis attempted to address, the research objectives and the limitations of the study.

Chapter 2 is a literature review of thirty journal articles, reports and publications related to the topic. It covers similar research done in other cities of the world and study of the methods used by different researchers. A summary of the takeaways from the literature review is provided at the end of the chapter.

Chapter 3 discusses the research methodology. It details the steps taken to acquire the data, clean the data, link the two data sets (housing transactions and air quality) and the data analysis steps.

Chapter 4 contains details on the clean dataset, the data analysis detailed steps including the types of hedonic regression used. It also lists some of the problems encountered during the analysis.

Chapter 5 is an evaluation of the findings of the study, relating them to the questions and the objective of the research. It also includes relating the findings to the literature review.

Chapter 6 lists the conclusions of the research, summarizing the learning from the research, recommendations on how to build on it and suggestions for future research in the domain.

Chapter 2. Literature Review

Many studies have examined the relationship between property price and air quality. Most studies used the Hedonic Price Method (HPM) also known as hedonic regression, which is the most suitable to examine the price of an item based on the contributions of its attributes (Herath & Maier, 2010). The air quality at the location of the property is considered as an attribute of the property. Some studies have also used the Marginal Willingness-to-Pay (MWTP), which measures the willingness to pay for a specific increase/decrease in an attribute value.

(Wang & Lee, 2022) found in their study in China that the Air Quality Index (AQI) negatively and significantly affects both housing sale prices and rental prices. They used a two-way fixed-effects panel model, fixing both province effects and time effects, to estimate the effects of air quality. They concluded that both home buyers and renters value air quality homogenously, even in the sub-market analysis. They also found that air quality impacts are predominantly documented in large and medium cities, whereas air quality has negligible impacts in small cities. The research provides input to policy makers to consider regional-targeted actions to control air pollution in the cities.

(G, Macpherson, & Zietz, 2005) examined numerous studies (125 in total) in which hedonic models were used to find a relationship between the attributes of a house and its price. The objective was to determine the attributes that are consistently significant to explain the house pricing, compare attributes' coefficients in several geographical locations and the relation of house price to the time-on-market. This research is crucial because all research on to air quality relation with housing sales or rent pricing is based on hedonic pricing models. Therefore, understanding how these models work is key to conducting new research in this area. They concluded that hedonic pricing models are location-specific and cannot be generalized for different locations. For this reason, a study using data from Dubai will provide valuable insights. They also found that square footage and lot size effects did not have a great variation on house price in different regions, and that there is no clear relationship between house price and time-on-market.

(Berezansky, Portnov, & Barzilai, 2010) started from the assumption that property pricing is not affected by “objectively measured” property attributes, but by factors which the sellers and buyers *perceive* as factual. Their research had two sides: objective (based on data) and subjective (based on surveys). In their study attempted to find out if the subjective measures performed better at predicting the property prices and if the variation in subjective air pollution in the population provided a better explanation in the variation of property prices compared to air quality assessments. They used a multi-variate regression analysis to predict property prices. The conclusion was that subjective evaluation of air pollution tended to explain the variation in unit pricing significantly better than the objective measurements. They also found out that an increase in perceived air pollution level from min to max affected the unit price by as much as 30% decrease in value. Hence it’s valuable to include subjective input when doing similar studies.

(Nourse, 1967) study was done in 1967 and is considered one of the first to study the effect of air pollution on housing prices in a scientific way by applying statistical models. He assumed that the people living in a neighborhood were homogeneous, i.e. in terms of income level, education, family stage and thus it’s enough to study the average property value of the neighborhood. Household income was a factor considered in the study, whereas family size, house area size and household occupation where excluded. A regression analysis with the property value as the dependent variable and “quality” and “income” as the independent values was done by the researcher. His research concluded that the actual impact of air pollution will depend on the total available houses in the city as well as the income distribution of the households.

(Chasco & Gallo, 2013) research was different because it combined the study of housing prices for downtown Madrid in relation to both air quality (5 primary pollutants and one secondary) and urban noise together. They used objective and subjective measures, similar to the research done by Berezansky, B., Portnov, B., & Barzilai, B. (2010). Since the researchers had access to hierarchical data (houses, census tracts and neighbourhoods), they opted for “multi-level models” (also known as hierarchical linear models), which a generalization of linear models. They used a variety of visualizations that provided useful insights into the distribution of the data, particularly between subjective and objective data. The researchers concluded that subjective measures have a greater influence on the house prices. They also uncovered hidden determinants (place, desirability) of house prices. They also concluded that noise and air pollution are “place-based” perception variables. Wealthy neighbourhood’s perception of air pollution is skewed because their perception is that other neighbourhoods are more polluted. They found that subjective measures are more accurate in predicting property prices than objective measures.

(Zou, 2019) studied the effect of air pollution on housing prices in China and found that this is becoming a global problem that requires the attention of governments. His research touched on two points that have not studied before: Air pollution may cause housing quality to deteriorate AND Air pollution may cause value depreciation to accelerate. The researcher used a geographically weighted regression (GWR) model to examine the variation effect of air pollution on housing pricing. GWR is a spatial analysis technique that takes non-stationary variables into consideration (e.g., climate; demographic factors; physical environment characteristics) and models the local relationships between these predictors and an outcome of interest. He also used ordinary least squares (OLS) regression in his research. He concluded that air quality is a factor that residents might have willingness to pay for, and confirmed there is a variation in the relation between air pollution and housing prices across the different cities, which requires place-based policies for each city.

(Amini, Nafari, & Singh, 2022) investigated the impact of air pollution on housing prices and rent across 1823 neighborhoods in Iran. The aim of the research was to study the willingness to avoid pollution. In 2010, the US imposed sanctions on Iran, preventing oil exports to Iran. The Iranian government took action to convert many petrochemical plants to produce low-quality gasoline. This resulted in a substantial increase in air pollution. The researchers wanted to study the effect of this pollution on housing pricing and rent using data from 2009-2014 by studying 1 million housing transactions. They used a pollution index for each neighborhood calculated as a distance-weighted average from the closest 3 pollution monitors., where the pollutant being monitored is Nitrogen dioxide.

The research found a relationship between air quality and housing pricing, where a 10 % increase in air pollution reduces house prices by 0.6%–0.8 %.

(Carriazo & Gomez-Mahecha, 2018) conducted their study in Bogota, Colombia, a city with many industrial areas. The pollutant studied was PM10. The researchers applied a *second stage* (SS) hedonic pricing model that allowed them to determine an inverse demand function for PM10 reductions. Their goal was to study whether a hedonic model was suitable to identify a demand function for air quality. They state that most studies on air quality based on the hedonic model, used first stage (FS) estimations. FS estimations have found a negative correlation between air quality and housing prices. They argue that FS studies are not suitable to determine welfare effects from non-marginal changes in air quality and attempted to find a demand function (a.k.a. willingness-to-pay function) that allows to value non-marginal changes in air quality monetarily. The research results confirmed the “demand law” through the SS hedonic model and the hypothesis that air quality is a normal good. They also concluded that the air quality demand function is a very suitable and flexible tool for decision makers to determine the non-marginal benefits of air quality improvement, thus it can guide their policies in this regard.

(Genanew, 2017) examined house prices drivers in Dubai by applying several linear and non-linear regression models. He aimed to explain the nonlinearity and heterogeneity in the house prices in relation to the house attributes. Air quality in the neighborhood was not a factor he studied. His research builds on previous research – which used the Box-Cox test - to address the nonlinearity and heterogeneity by using semi-log, log-log and quantile regression, taking into consideration the time trend and the fact that Dubai experiences house price bubbles. The research demonstrates that using a combination of linear and nonlinear / quantile regression specifications can address the concerns of nonlinearity and heterogeneity. House type was found to be an important factor in house price determination, and that apartments are associated with a higher premium compared to villas. His research informs housing developers that buyers are concerned with house location and willing to pay more if the house is near water, in the city center or in newly developed areas.

(Monson, 2009) studied the hedonic pricing method on the price of buildings, as an alternative to discounted cash flow models. He argued that this method is a good alternative in the absence of a market, when no similar property is available to compare with and for non-income generating buildings. He lists in detail all the property characteristics which should be considered when building a hedonic pricing model. He also lists three cases on which he applied the hedonic pricing model and discusses the results. He concludes that the method is a valuable tool in the real estate industry to understand the correlation between property characteristics and its transaction price, as well as predict the future price of the property.

(Saphores & Wei, 2012) presented a study on around 20,660 single family homes for the years 2003 and 2004 in the city of Los Angeles where they analyzed the effect of urban trees, irrigated grass, and non-irrigated grass areas on the price of the property in the neighborhood. They followed the standard hedonic framework which states that the property price is a function of structural, neighborhood and environmental factors. Moreover, they used two models in their study: the geographically weighted regression (GWR) and the Cliff–Ord model with fixed effects. While their analysis is an advanced version of the hedonic pricing model analysis, their methodology is valuable overall to guide the research in the field. Their conclusion was that the addition of irrigated grass in the property or at the neighborhood level would benefit most properties. Their study is valuable to the urban planners in their green planning exercises.

(Bazyl, 2009) applied the hedonic pricing method on the Warsaw property market based on 2006 data, including air pollution parameters such has NO_2 and SO_2. The researcher used two models: the *spatial autoregression model* and *spatial error model*, where she found that both models confirmed there is a significant special autocorrelation in the basic hedonic model. The research assumed that the price of the property is only dependent on its characteristics and not correlated with prices of nearby properties.

She concluded that the price of the property has a positive correlation with the presence of a metro station within 1km of the property and with the presence of green areas next to the property. Moreover, she concluded that the presence of industrial area decreases prices of flats.

(Graves, Murdoch, Thayer, & Waldman, 1988) studied the robustness of the hedonic-based methods by addressing the issues of variable selection, the measurement error, error distribution and functional form. This was applied to property prices in California in relation to urban air quality, for the purpose on guiding the public policy. The researchers used an iso-pleth curve to assign air quality data to each property, which is a method I plan to use in my research. They concluded that in order to properly estimate the effect of air quality on the house value, using a hedonic pricing method, would require four items, namely, a complete set of independent variables, with accurate measures of the variables, selecting the appropriate relation between price and the variables and finally, having the right stochastic assumptions.

(Brécard, Le Boennec, & Salladarré, 2019) studied the effects of special and environmental variables on the prices of property in Nantes, France. They argued that air quality and closeness to the mobility network had no effect on the prices of the property but that closeness to the city center affects positively the property price. They also found results that are consistent with previous research done in the French market, where the property surface area plays a major role in the pricing. The result of this research provides guidance to city planners on the sustainable urban mobility plans. Their methodology is based on using a hedonic price model that takes into account spatial autocorrelation and spatial heterogeneity. They also state that the hedonic method of analysis is almost used unanimously by researchers in the housing pricing field.

(Neill, Hassenzahl, & Assane, 2007) studied the effects of air quality on the property pricing and compared *spatial* versus *traditional* hedonic models. They compared the spatial MLE method with the traditional ordinary least squares (OLS) method. They addressed the limitation of the maximum likelihood estimation (MLE) method in hedonic housing pricing to small data sets, by coupling MLE with a technique called block bootstrapping. Their findings showed that spatial MLE is far more superior than the traditional OLS in performance and that air quality matters regardless of the method used in analysis. Moreover, they found out that carbon monoxide and particular matter (PM10) are the most significant variables from an air quality perspective on the price of the property.

(Cebula, 2009) study in a direct application of the hedonic pricing model on house prices in the city of Savannah, using 24 potential variables that could affect the price. Although air quality is not among these parameters, the methodology of the researcher and the steps shown in the study provide excellent guidance on how to conduct similar research, especially that it took into consideration interior and exterior features of the property. The research used seasonal control variables to study the effect of the "time of the sale" on the price, which provides guidance to potentially consider that in my research. The researcher built three models and did a comparison between the three, where each model has a different set of variables. His 1st model included all parameters, then he further reduced the number of variables based on the significance of the input parameters. He concluded that there is a positive correlation between price and the following variables: number of bathrooms, existence of fireplaces, bedrooms, garage spaces, number of stories, and the number of square feet of livable space in the property.

(Zhang, Mao, & Wang, 2021) adopted a different approach to studying the effect of air quality on house prices, where they studied the willingness of people to pay for clean air following the establishment of the "Smog Free Tower (SFT)" project in Xi'an, China, and the release of an assessment report about it. The project included measures to purify polluted air and the researchers' goal was to find out if property prices would be affected by cleaner air, as measured by the closeness of the property to the SFT project. Their research was based on the hedonic model, as this is the universal model used to capture the buyers' willingness to pay for various housing features. The researchers drew a circle of 5Km around the SFT project and studied properties within this circle. My approach will be similar where I will associate the properties around an air quality station by drawing a circle around each station.

They concluded that, following the release of the SFT assessment report, the relationship between housing prices and distance to SFT changed, where the distance to SFT was negatively related to the housing price. This indicates that people are willing to pay for clean air. They also found out that access to transportation had more significance on the price than clear air, which means people put more emphasis on transportation accessibility than on air quality when buying a house.

(Ilvessalo, 1995) reported on a new method to calculate the air quality index, based on values from different pollutants, which he states as a simplified way to express air quality. His driver for the new method is the fact that it's getting more and more difficult to make use of the different concentration values of the pollutants, and argues that the index presentation of air quality data is more easily understandable. This air quality index is calculated using the concentrations of all the contaminants measured, which is an exercise that would largely simplify my analysis if the air quality data is represented by a single value, rather than a set of contaminant values.

(Rusmawati, Maharani, & Surahman, 2020) used regression analysis to explore the factors that influence house prices in the cities of Surabaya and Gresik in Indonesia. They studied attributes such as land area, building area, number of bedrooms, number of bathrooms, electrical features, and others to determine their impact on house prices. They identified key attributes that influence house prices in each city. Their findings were different for the two cities. In Surabaya, land area, building area, number of bedrooms, and number of bathrooms were found to be key factors affecting the house prices, whereas in Gresik, electrical features, land area, building area, material, and carport were found to be the effective attributes. Their research provided insights into understanding the dynamics of house price determination in different urban areas, showing that regional factors were driving house prices. Their research is important in showing the importance of considering local factors in real estate analysis and decision-making and their findings would help policymakers and investors in making informed decisions about the property market.

(Wang, Lee, & Shirowzahn, 2021) studied how air quality affects property values in China through a Meta-Regression analysis. They joined findings from 117 observations in different studies to understand the relationship between air quality and housing prices.

Their study confirmed that air quality is a significant factor that impacts housing prices, and that various factors such as types of air quality parameters, the data sources, control variables, and estimation methods greatly affect these estimates. Their analysis shed some light on the complexity of the housing market in relation to environmental factors. In terms of contribution, their research helps policymakers by highlighting the economic significance of air quality on housing prices, assisting policymakers and homeowners in making informed decisions. Their research also is useful to guiding scholars on the choice of control variables and the estimation approaches used in their analyses.

(Chiarazzo, Coppola, Dell'Olio, Ibeas, & Ottomanelli, 2014) investigated the relationship between environmental conditions in a given area and the residential location choices, emphasizing the effects of environmental quality and landscaping on house values. They used hedonic Multiple Linear Regression (MLR) models to estimate housing prices in a metropolitan area as a function of real estate, environmental, and accessibility attributes. The results of their study indicated the importance of including environmental variables in the MLR model specification. The models showed a complex relation between accessibility, location choices, and real estate values, providing valuable insights for urban planners, transport policy makers and investors.

The researchers used both quantitative and qualitative research methods. They identified significant attributes such as the number of bedrooms, number of bathrooms, presence of a parking or garden, and transport accessibility indicators, as influencers of housing values. They also revealed the trade-offs between air quality, environmental conditions, and accessibility to workplaces.

(Chay & Greenstone, 2005) studied the relationship between air quality regulations and property values, through the analysis of county-level data from 1970 to 1990. They first discussed the regulatory framework established by the 1970 and 1977 Clean Air Act Amendments, which categorized counties as "non-attaining" or "attaining" based on air quality standards. The study utilized this attainment status as a variable to determine the impact of pollution regulations on housing prices. They showed that there was a significant decline in pollution and a corresponding rise in housing values in regulated counties during the 1970s and 1980s. In addition, the researchers highlighted the complex relationship between air quality and housing prices, presenting evidence of the economic gains for homeowners by pollution regulations.

(Murdoch & Thayer, 1988) did an interesting study which is based on the fact that environmental quality, unlike static housing attributes, fluctuates over time. They examined the validity of using mean levels of environmental quality in the hedonic model. They showed that a “probability model” based on the distribution of environmental quality values outperforms the mean model in estimating the housing prices. They also showed that the benefit estimates derived from the traditional mean model are likely to be biased, emphasizing the need for more complete measures of environmental quality to improve the accuracy of hedonic methods. They recommended further research in understanding the variable nature of environmental quality before utilizing hedonic price method for environmental policy decisions. Their research was questioning the traditional approach of using mean environmental quality in the hedonic method.

(Bayer, Keohane, & Timmins, 2009) focused on the use of hedonic valuation methods to estimate the willingness to pay for air quality. They used a discrete-choice model to deduce the utility associated with living in different areas and then analyzed the relationship between the utilities and air pollution concentrations. In their hedonic analysis, they addressed the endogeneity problem by using an innovative instrumental variables approach, highlighting the importance of taking into consideration endogeneity (where the predictor is correlated with the error term in the regression) and mobility costs. Their study showed that estimates of willingness-to-pay for air quality may be biased downward if not taking into consideration the migration costs. They also discussed the importance of particulate matter (PM) as the standard measure of air pollution and its relation to health issues, highlighting the importance of considering far emissions as a natural instrument for local air pollution.

(Yang, Zhou, & Ding, 2018) used an approach to predicting air quality in urban residential using machine learning classification algorithms. They used a hedonic approach in their study – thus it was worth looking at their research. Traditional air quality prediction typically relies on sensors and complex algorithms, which can be costly and time-consuming. The researchers proposed a new approach that uses housing prices and characteristics of urban residences as indicative variables for air quality prediction. Their study used the Support Vector Machine (SVM), Naive Bayes, and K-Nearest Neighbor (KNN) algorithms to do the mapping between feature variables and air quality, which enabled them to predict air quality with high accuracy. SVM was the most accurate in predicting air quality (88% accuracy). Their research is valuable because it could guide future research in the domain, where air quality is predicted based on house attributes, and not vice-versa as typically done by researchers of the topic.

(Saptutyningsih & Ma'ruf, 2015) studied the economic impact of air pollution on housing prices in Yogyakarta City. Their study focused on ozone (O3) levels and examined how air quality influenced property values and the willingness of consumers to pay for air quality improvements. Their study was based on a hedonic price model to establish the relationship between air quality and property prices. They found that finding that a 1% increase in O3 levels resulted in a 0.063% increase in property prices, which would be against expectation as more ozone means more pollution and hence would result in a decrease in property price. Their study also utilized a health production function to assess the impact of air pollution on health-related workday losses and medical expenses. They concluded that individual medical histories significantly affect the number of workdays lost due to air pollution, and that higher Ozone levels lead to increased medical expenses.

(Nguyen, 2020) applied the hedonic pricing model to estimate house prices in the housing market of Vietnam. The goal of the study was to provide insights into the factors that affect house prices in a country with a developing housing market. The approach involved collecting data through surveys of housing projects in Ho Chi Minh and Ha Noi city. The researcher applied the "Ordinary Least Squares" (OLS) regression and robustness statistics to ensure reliable estimation results. The results showed that the hedonic pricing model can be effectively utilized to estimate house prices in Vietnam, with factors such as house area, number of bedrooms, amenities, and house structure significantly influence the prices. Interestingly, the study shows a negative correlation between the proximity to the city center and house prices, which is in contrast to the trend found in many markets. The results of the study are not only applicable for Vietnam but also for other countries with similar housing market characteristics. The study confirms the advantages of the hedonic pricing model in estimating house prices, stressing out the need for selecting the right attributes that influence the house price.

(Zietz, Zietz, & Sirmans, 2008) used quantile regression to explore the pricing of residential real estate, as a different approach from the traditional OLS regression method. The researchers studied the impact of housing characteristics on the selling price. They emphasized the importance of quantile regression in understanding how housing attributes are valued differently across the distribution of house prices. For example, they found that buyers of higher-priced homes value certain housing characteristics, such as house area and the number of bathrooms, differently from buyers of lower-priced homes. The study also covered the implications of spatial autocorrelation and its effects on the coefficients of various attributes. This study was challenging the Ordinary Least Squares (OLS) regression methodology and argued that quantile regression provides a better understanding of the house price at different points of the house prices.

(Fernandez, 2019) did an analysis of how hedonic pricing models are applied to the New Zealand housing market. His goal was to highlight the use of these models to evaluate the impact of various environmental and urban features on house prices, and he showed how different factors like environmental features, urban features, and policy changes influence the house prices. His report highlighted the importance of hedonic models in guiding urban planners and policymakers on “the value individuals place on features” and the impacts this has on housing prices.

His report also covered the methodological aspects of hedonic models, including the use of quantile regressions to explore heterogeneous responses across price distributions and the value of non-market amenities, similar to the research by (Zietz, Zietz, & Sirmans). Fernandez also stated that hedonic models have evolved as essential tools for decision-making, cost-benefit analysis, and policy formulation in cities.

(Hill, 2011) did a detailed study on the hedonic pricing method and explored the application of hedonic pricing models in the domain of housing markets. His focus was on quality-adjusting house prices by considering the unique characteristics of each house, such as size, location, and amenities. His study explored the importance of hedonic methods in developing accurate house price indexes, which are important for real-estate practitioners, investors as well as policy makers. An evolution of hedonic models was provided as well as their methodologies, and their implications for understanding housing market dynamics. The researcher also listed the weaknesses of the different hedonic approaches and compares it to alternative approaches like repeat-analysis, showing that the hedonic method is superior in capturing the value of the house attributes. He also studied the role of hedonic pricing in addressing substitution bias and the challenges associated with missing observations in attributes (an issue faced in the data set for this research), providing examples, and discussing the implications of unexpected coefficient signs in the linear regression.

(Zeng, Fahad, Wang, Nassani, & Binsaeed, 2023) did an interesting study to explore the causal impact of house prices on air quality in Chinese cities from 2009 to 2018. This is the exact opposite study of this research which studies the impact of air quality on house prices. Their study used instrumental variable methods and a two-stage least squares regression analysis to examine the impact “mechanism” of housing prices on air quality. Their findings show a negative impact of housing prices on air quality, where a 1% increase in house prices lead to a 0.1485% increase in air pollution, specifically PM2.5 concentration.

Additionally, the study investigated the collection of administrative levels and identified that housing prices in general administrative level cities significantly inhibit air quality, while the impact in high administrative level cities is not as significant. The study also covered the mechanisms through which housing prices affect air quality, highlighting the promotion of real estate investment and the inhibitory effect on urban innovation and development.

Fenwick (2013) discussed in a dedicated chapter the hedonic price method. He discussed the application of hedonic regression methods, how to incorporate a time dummy variable and imputation approaches for estimating price indices. His handbook highlights the theoretical foundations, model specifications, and practical considerations for using hedonic regression to estimate the marginal contributions of property characteristics and construct quality-adjusted price indices. He emphasizes the use of least squares regression (OLS) to estimate the hedonic models, demonstrating its relevance in academic studies and its potential applicability in statistical agencies. He also lists the log-linear regression as a good alternative to the OLS regression.

2.1 Takeaways from the Literature Review

As a summary of the literature review, the following points should be considered in the research thesis:

- The hedonic pricing model is the method used by most researchers when conducting similar research on the relationship between the property value and its characteristics, internal and external. Air quality is considered as an external characteristic of the property.
- To simplify the study, an Air Quality Index (AQI) value can be preferred to the values of the individual air contaminants in the calculations. The AQI is an indicative value that takes into consideration the individual contaminants and can be easily calculated from the values of individual contaminants.
- The researchers used data spanning at least 5 years for their study.
- It is important to include particulate matter (PM) as a key factor in hedonic studies of air pollution.
- At least two models need to be created and compared, as is done by almost all researchers in this field. Most researchers used the Ordinary Least Squares (OLS) method in their hedonic studies of the housing market, which has proven to be reliable in terms of results. Other researchers used quantile regression models as an alternative way.

Chapter 3. Research Methodology

3.1 Introduction

Cross-Industry Standard Process for Data Mining (CRISP-DM) is selected for this thesis. The methodology of CRISP-DM is comprised of six steps as depicted in the following figure:

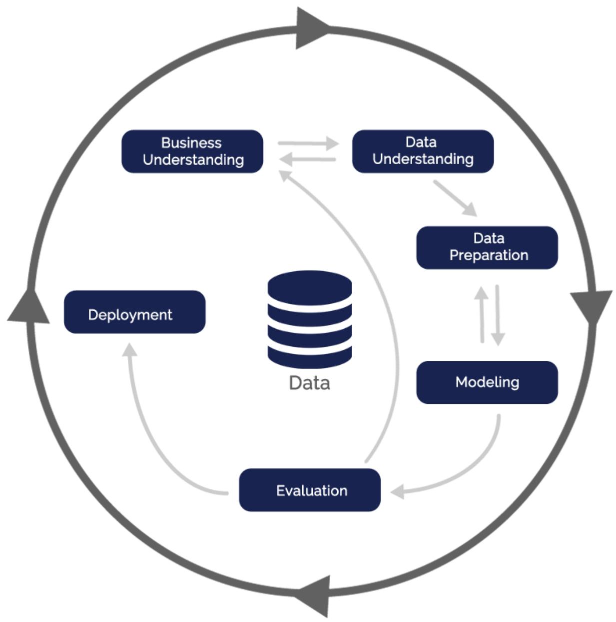


Figure 1 CRISP-DM Methodology

Business Understanding: It is essential to understand the business at the beginning of the project and recognize the limitations and problems associated with specified study.

Data Understanding: This step is an exploration of the data files for data cleaning. Using the pollutants data, the Air Quality Index (AQI) value will be calculated so as to deal with a single value for air quality rather than a set of six values.

Data Preparation: Then using the location of each Dubai Municipality monitoring station, a circle of (5km) radios will be drawn in order to assign all neighborhoods within the circle in that monitoring station data. A join will be done with the data in the property sales file and the air quality data based on the neighborhood and date. The result of this join is a data set which contains the transactions and the air quality data at the time of the transaction.

Modeling: A regression model will be built to assess the relation between input variables of the property, including AQI, and the output variable being sale price. A model will be built to predict the house price based on the input attributes.

Evaluation: The results of the modelling will be used to draw conclusions.

This research has to the objective of finding if there is a relation between the house price in different neighborhoods in Dubai and the air quality in that district. It will attempt to use the hedonic pricing method to build a model that can predict the price of the house from its characteristic including the area of the house, the number of bedrooms, the number of parking spots in addition to the Air Quality Index (AQI) of the area. The hedonic price method is the most common method used in this type of research and this research will attempt to apply it on the Dubai property and air quality data. At least two types of models will be used in the study, with an assessment of their accuracy and power of prediction. This kind of research has not been done in Dubai data in the last 5 years.

Chapter 4. Data Analysis

The research focused on using quantitative methods for the analysis. Dubai is a city that has adopted the “open data” strategy, hence the data for this research is available through government portals, namely Dubai Pulse and Dubai Statistics Centre. The data provided through the portals has the basic attributes of the sales transactions needed for the research for the years 2021-2023. The air quality data – from Dubai Municipality air quality stations – is also available for the same period. The philosophy of the research was based on the fact that the housing sale transaction can be linked with the air quality index for the nearest air quality station for that period. By linking the housing sales transactions and the air quality datasets, a combined dataset which has the housing sales attributes and the Air Quality index (AQI) reading for the transaction was obtained.

4.1 Data Acquisition and Preparation

- Obtaining of the housing transactions and air quality datasets from government portals for the period from Jan 2021 to April 2023. The original data file had 47 attributes and 173,643 records.
- Performing a data cleaning on the housing transactions data sets as follows:
    a. Removal of all data attributes that are not relevant to the study. This included all attributes in Arabic (duplicates of the English attributes), attributes which had “null” for most of the records, and attributes which can be deduced from other attributes (such as sale price per meter).
    b. Removal of attributes with more than 25% of the records of missing values and records with missing values for which data cannot be filled (example: nearest metro station, nearest landmark)
    c. Filtering the data to include the transaction dates from Jan 2021 to April 2023.
    d. Extracting the year of the transaction in a separate attribute.

e. Processing the “number of bedrooms” attribute to remove the text “BR” and converting it to a numerical integer.
f. Processing the “parking data” to convert the parking spot names into a number of parking spots attribute as a numerical integer.

The output of this step is the sales transactions records as follows:

Table 1 Housing Sales Transaction Records after Initial Data Cleaning

transaction_id	trans_date	trans_year	area_name	price	unit_area	numb_rm	Total_park
1-11-2021-7316	04/05/2021	2021	Marsa Dubai	1600000	201.99	3	1
1-11-2021-114	04/01/2021	2021	Marsa Dubai	1600000	190.28	3	1
1-11-2022-26348	26/10/2022	2022	Marsa Dubai	2175000	164.13	3	1
1-11-2021-18989	26/10/2021	2021	Marsa Dubai	2153976	189.53	3	1
1-11-2021-21530	30/11/2021	2021	Marsa Dubai	2050000	144.68	3	1
1-11-2022-25984	21/10/2022	2022	Marsa Dubai	2050000	201.84	3	1
1-11-2021-18385	14/10/2021	2021	Marsa Dubai	1925000	177.3	3	1
1-11-2022-2950	17/02/2022	2022	Marsa Dubai	2150000	174.87	3	1
1-11-2021-7572	09/05/2021	2021	Marsa Dubai	4100000	223.55	3	1
1-11-2021-20114	10/11/2021	2021	Marsa Dubai	2978999	400.04	3	1
1-11-2022-12084	02/06/2022	2022	Marsa Dubai	8000000	180.13	3	1
1-11-2022-26327	25/10/2022	2022	Marsa Dubai	3450000	175.98	3	1
1-11-2022-8392	20/04/2022	2022	Marsa Dubai	2700000	178.9	3	1
1-11-2022-218	06/01/2022	2022	Marsa Dubai	6180000	172.89	3	1
1-11-2021-22842	20/12/2021	2021	Marsa Dubai	23000	1.82	3	1
1-11-2022-26839	31/10/2022	2022	Marsa Dubai	4000000	420.76	3	1
1-11-2022-18476	04/08/2022	2022	Marsa Dubai	7500000	180.48	3	1
1-11-2021-17853	07/10/2021	2021	Marsa Dubai	2900000	321.07	3	1
1-11-2022-19069	11/08/2022	2022	Marsa Dubai	8700000	197.78	3	1

1-11-2022-29587	24/11/2022	2022	Marsa Dubai	2025000	175.04	3	1
1-11-2021-21736	06/12/2021	2021	Marsa Dubai	6240000	173.52	3	1
1-11-2022-16537	14/07/2022	2022	Marsa Dubai	3710000	161.88	3	1
1-11-2021-13567	05/08/2021	2021	Marsa Dubai	6199000	173.52	3	1
1-11-2021-16363	19/09/2021	2021	Marsa Dubai	3300000	224.68	3	1

- Obtaining the air quality data for the 14 air quality stations. The data includes minimum, maximum and average values for the following pollutants: Carbon Monoxide (CO), Ozone (O3), Nitrogen Dioxide (NO2), Sulphur Dioxide (SO2), Concentration of particulates (PM10) and Concentration of particulates (PM2.5). Since only *Carbon Monoxide (CO), Ozone (O3), Nitrogen Dioxide (NO2), Sulphur Dioxide (SO2)* data was available for the study period, only these pollutants were considered in the air quality data. The *average* values for these pollutants were used from the dataset as it is the most representative. Since the average values were almost constant for each station per pollutant, the average values for the station were used to represent the pollutant values for the year.

Table 2 List of Air Quality Stations and Air Pollutants Average Levels per Year

Station	Year	CO	O3	NO2	SO2
Deira	2021	0.37	0.03	0.02	0
	2022	0.36	0.04	0.01	0
	2023	0.34	0.04	0.01	0
Al Karama	2021	0.38	0.03	0.02	0
	2022	0.38	0.04	0.02	0
	2023	0.38	0.04	0.02	0
Zabeel Park	2021	0.39	0.03	0.01	0
	2022	0.4	0.04	0.02	0

	2023	0.38	0.03	0.01	0
DIP	2021	0.35	0.03	0.02	0
	2022	0.328	0.028	0.024	0.003
	2023	0.32	0.04	0.02	0.003
Emirates Hills	2021	0.37	0.03	0.01	0
	2022	0.33	0.03	0.01	0
	2023	0.31	0.04	0.02	0
Dubai Airport	2021	0.39	0.03	0.02	0
	2022	0.37	0.03	0.02	0
	2023	0.39	0.06	0.02	0

- Calculating the Air Quality Index (AQI) as a single value that represents the pollutants values of the air quality station. The AQI combines the data of the pollutants into a single value that can be associated with the housing sale transaction. The U.S. EPA's AQI calculation was used since this is the standard used by Dubai Municipality. The calculation involves converting pollutant concentrations into a sub-index and then taking the maximum of these sub-indices as the AQI. The U.S. EPA provides specific breakpoints for different pollutants. The assumption was to use standard conditions (e.g., 8-hour averages for O3, 1-hour averages for NO2 and SO2, and 8-hour averages for CO). For example, the breakpoints for CO are:
    - 0.0 to 4.4 ppm: AQI 0 to 50 (Good)
    - 4.5 to 9.4 ppm: AQI 51 to 100 (Moderate)
    - 9.5 to 12.4 ppm: AQI 101 to 150 (Unhealthy for Sensitive Groups)
    - 12.5 to 15.4 ppm: AQI 151 to 200 (Unhealthy)
    - 15.5 to 30.4 ppm: AQI 201 to 300 (Very Unhealthy)
    - 30.5 to 40.4 ppm: AQI 301 to 400 (Hazardous)
    - 40.5 to 50.4 ppm: AQI 401 to 500 (Hazardous)

For each pollutant, its concentration was converted to its respective AQI value and then taking the highest of these as the overall AQI for that period. The formula used for the AQI is:

$$AQI = (I_{High} - I_{Low}) \,/\, (C_{High} - C_{Low}))\times(C - C_{Low}) + I_{Low}$$

where C is the pollutant concentration, C_{Low} and C_{High} are the concentration breakpoints that C falls between, and I_{Low} and I_{High} are the AQI breakpoints corresponding to C_{Low} and C_{High}.

The output of this step was a data file containing the air quality stations names, the year and the AQI associated with the station as follows:

Table 3 Air Quality Stations and Air Quality Index per Year

Air_quality_stn	Year	AQI
Deira	2021	27.78
Deira	2022	37.04
Deira	2023	37.04
Al_Karama	2021	27.78
Al_Karama	2022	37.04
Al_Karama	2023	37.04
Zabeel_Park	2021	27.78
Zabeel_Park	2022	37.04
Zabeel_Park	2023	27.78
DIP	2021	27.78
DIP	2022	25.93
DIP	2023	37.04
Emirates_Hills	2021	27.78

Emirates_Hills	2022	27.78
Emirates_Hills	2023	37.04
Dubai_Airport	2021	27.78
Dubai_Airport	2022	27.78
Dubai_Airport	2023	67.33
. . .	. . .	. . .

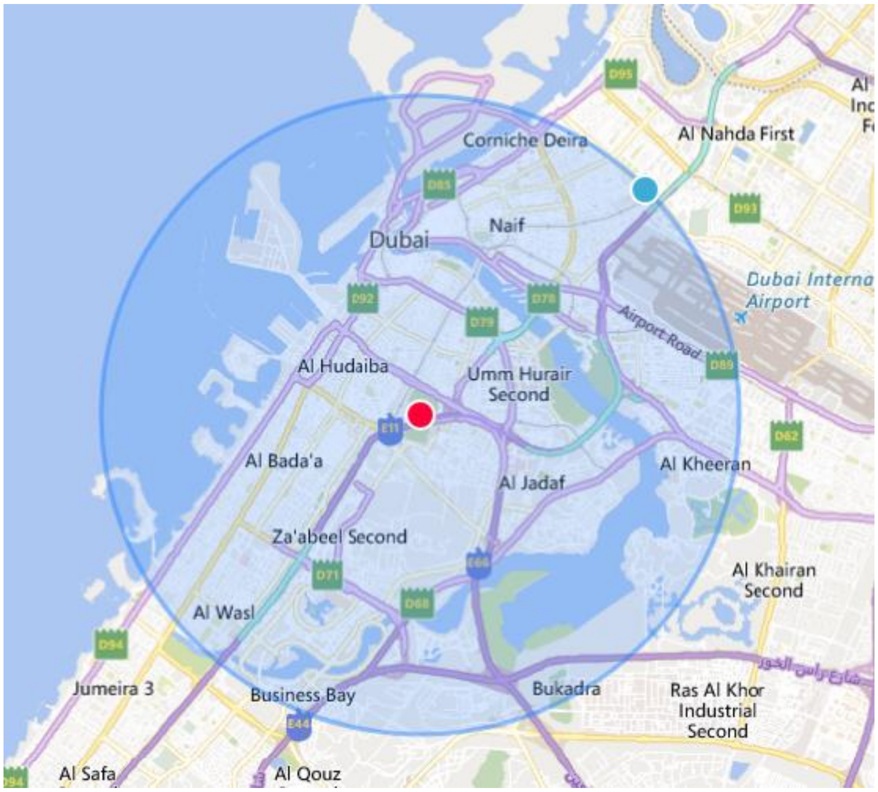
- Determining the area name for the transaction and the corresponding air quality station. For this step, a circle of 7KM radius was drawn around each air quality station based on the following GIS data. The latitude and longitude values were converted to a format acceptable for Tableau:

Table 4 Air Quality Stations with Latitude and Longitude Coordinates

Air quality_stn	Lt	Ln	Lt_dd	Ln_dd
Dubai Airport	25°15'11.675	55°21'49.698	25.25324306	55.363805
Nad_Al-Shiba	25°9'12.741	55°20'23.617	25.15353917	55.33989361
Al_Qusais	25°16'39.501	55°21'59.341	25.27763917	55.36648361
Deira	25°15'49.764	55°18'37.576	25.26382333	55.31043778
DIP	24°59'55.229	55°9'47.876	24.99867472	55.16329889
Emirates_Hill	25°4'16.158	55°9'55.538	25.071155	55.16542722
Jebel_Ali	25°1'23.300	55°6'16.202	25.02313889	55.10450056
Mushrif_Park	25°13'1.623	55°27'14.314	25.2171175	55.45397611
Sh_MBZ	25°3'9.485	55°16'16.996	25.05263472	55.27138778
Sh_Zayed	25°9'24.475	55°13'48.394	25.15679861	55.23010944
Zabeel_Park	25°13'58.534	55°17'55.423	25.23292611	55.29872861
Warsan	25°9'5.542	55°25'31.558	25.15153944	55.42543278
Al_Karama	25°14'46.104	55°18'24.569	25.24614	55.30682472

The following shows an example of the circles drawn and the associated area names with the air quality station:

Figure 2 Mapping the association of Air Quality Station with Area Names



The following step was to associate the area name in the sales transaction with the air quality station. This step was done manually by visually including all area names within a circle with the center point (the air quality station). The output of this step is the sales transactions data file with the air quality stations in it:

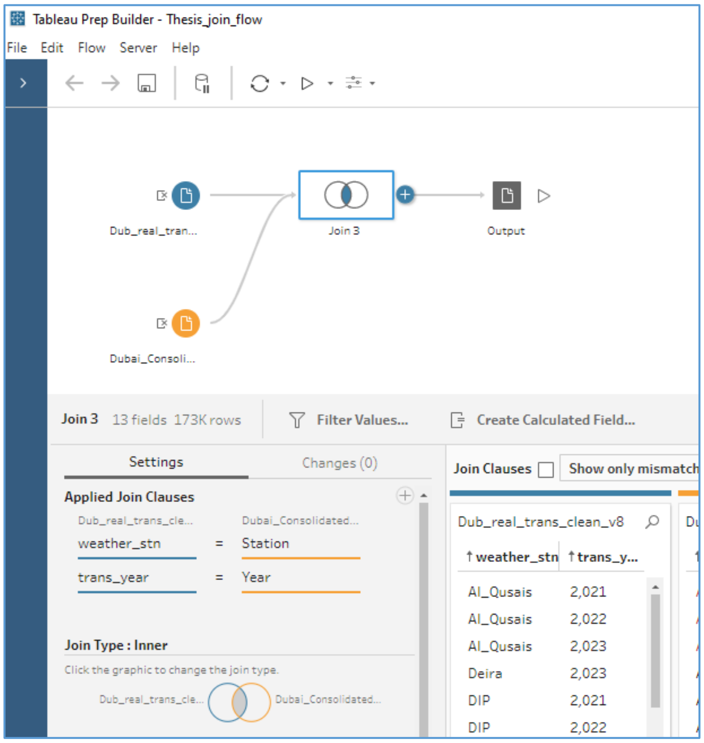
Table 5 Housing Sales Transactions Records with Air Quality Station Data

transaction_id	trans_date	trans_year	area_name	Air quality_stn	price	unit_area	numb_rm	Total_park
1-11-2021-7316	04/05/2021	2021	Marsa Dubai	Emirates_Hills	1600000	201.99	3	1
1-11-2021-114	04/01/2021	2021	Marsa Dubai	Emirates_Hills	1600000	190.28	3	1
1-11-2022-26348	26/10/2022	2022	Marsa Dubai	Emirates_Hills	2175000	164.13	3	1
1-11-2021-18989	26/10/2021	2021	Marsa Dubai	Emirates_Hills	2153976	189.53	3	1
1-11-2021-21530	30/11/2021	2021	Marsa Dubai	Emirates_Hills	2050000	144.68	3	1

1-11-2022-25984	21/10/2022	2022	Marsa Dubai	Emirates_Hills	2050000	201.84	3	1
1-11-2021-18385	14/10/2021	2021	Marsa Dubai	Emirates_Hills	1925000	177.3	3	1
1-11-2022-2950	17/02/2022	2022	Marsa Dubai	Emirates_Hills	2150000	174.87	3	1
1-11-2021-7572	09/05/2021	2021	Marsa Dubai	Emirates_Hills	4100000	223.55	3	1
1-11-2021-20114	10/11/2021	2021	Marsa Dubai	Emirates_Hills	2978999	400.04	3	1
1-11-2022-12084	02/06/2022	2022	Marsa Dubai	Emirates_Hills	8000000	180.13	3	1
1-11-2022-26327	25/10/2022	2022	Marsa Dubai	Emirates_Hills	3450000	175.98	3	1
1-11-2022-8392	20/04/2022	2022	Marsa Dubai	Emirates_Hills	2700000	178.9	3	1
1-11-2022-218	06/01/2022	2022	Marsa Dubai	Emirates_Hills	6180000	172.89	3	1
1-11-2021-22842	20/12/2021	2021	Marsa Dubai	Emirates_Hills	23000	1.82	3	1
1-11-2022-26839	31/10/2022	2022	Marsa Dubai	Emirates_Hills	4000000	420.76	3	1
1-11-2022-18476	04/08/2022	2022	Marsa Dubai	Emirates_Hills	7500000	180.48	3	1
1-11-2021-17853	07/10/2021	2021	Marsa Dubai	Emirates_Hills	2900000	321.07	3	1
1-11-2022-19069	11/08/2022	2022	Marsa Dubai	Emirates_Hills	8700000	197.78	3	1
1-11-2022-29587	24/11/2022	2022	Marsa Dubai	Emirates_Hills	2025000	175.04	3	1
1-11-2021-21736	06/12/2021	2021	Marsa Dubai	Emirates_Hills	6240000	173.52	3	1
1-11-2022-16537	14/07/2022	2022	Marsa Dubai	Emirates_Hills	3710000	161.88	3	1

- Associating the air quality index value for the year / station with the sales transaction. For this step, Tableau Prep Builder software was used to perform an inner join on the air quality AQI data with the sales transactions data. The join condition was done on the fields ‘air quality stn’ and the ‘transaction year’ as follows:

Figure 3 Data Joining in Tableau between Housing Sales Transactions and AQI Data



The output of this step is the sales transactions records, with the AQI for each record:

Table 6 Housing Sales Transactions Data Association to Air Quality Index

transaction_id	trans_date	trans_year	area_name	Air_quality_stn	price	unit_area	numb_rm	Total_park	AQI
1-11-2021-13325	02/08/2021	2021	Al Jadaf	Dubai_Airport	1250000	119.74	2	1	27.78
1-102-2021-13291	15/08/2021	2021	Al Jadaf	Dubai_Airport	852600	82.22	2	1	27.78
1-102-2021-13288	15/08/2021	2021	Al Jadaf	Dubai_Airport	850000	122.62	2	1	27.78
1-102-2021-13263	15/08/2021	2021	Al Jadaf	Dubai_Airport	840000	123.37	2	1	27.78
1-102-2021-13260	15/08/2021	2021	Al Jadaf	Dubai_Airport	879103	148.85	2	1	27.78
1-102-2021-13261	15/08/2021	2021	Al Jadaf	Dubai_Airport	954000	148.85	2	1	27.78
1-11-2021-14347	18/08/2021	2021	Al Jadaf	Dubai_Airport	1870000	285.4	2	1	27.78
1-102-2021-14306	24/08/2021	2021	Al Jadaf	Dubai_Airport	820000	82.22	2	1	27.78
2-13-2021-10730	26/08/2021	2021	Al Jadaf	Dubai_Airport	684000	97.12	2	1	27.78
1-11-2021-15233	01/09/2021	2021	Al Jadaf	Dubai_Airport	1400000	130.86	2	1	27.78

2-13-2021-11168	06/09/2021	2021	Al Jadaf	Dubai_Airport	618750	96.04	2	1	27.78
1-11-2021-16625	22/09/2021	2021	Al Jadaf	Dubai_Airport	1400000	100.35	2	1	27.78
1-102-2021-17983	06/10/2021	2021	Al Jadaf	Dubai_Airport	850000	122.62	2	1	27.78
1-102-2021-17984	06/10/2021	2021	Al Jadaf	Dubai_Airport	850000	123.37	2	1	27.78
1-11-2021-17867	07/10/2021	2021	Al Jadaf	Dubai_Airport	800000	118.92	2	1	27.78
1-11-2021-18376	14/10/2021	2021	Al Jadaf	Dubai_Airport	2891662	187.4	2	1	27.78
1-11-2021-18476	17/10/2021	2021	Al Jadaf	Dubai_Airport	2938248	187.75	2	1	27.78

4.2.2 Regression Model: The Hedonic Pricing Method (HPM)

The Hedonic Pricing Method (HPM) is an economic model used to analyze how the characteristics of an item (in our case, a house) affect its price. This approach is excellent for analysing housing markets, where the cost of a house depends on its size, location, number of rooms and other external factors such as the neighbourhood quality and environmental features. In this research, which focuses on the relationship between housing prices and air quality, the Hedonic Pricing Method allows us to identify the specific impact of air pollution on the valuation of housing. In order to provide an in-depth analysis of the impact of various characteristics like air quality on house price, HPM decomposes the total price of a house into attributes contributing to it. It therefore enables us to better understand how much people are willing to pay for cleaner air in their homes by breaking down the price of houses into the key elements including air quality.

The significance of the Hedonic Pricing Method in studying housing prices and air quality results from its ability to quantify the hidden value of environmental attributes (i.e. air quality). Unlike other commodities that have a direct market price, air quality does not have a direct market value. It determines the attractiveness of neighbourhoods and therefore the price of property. The use of HPM enables us to put monetary values on air pollution. This is achieved by comparing houses with similar features but different degrees of air quality. The understanding of how environmental factors impact real estate’s prices will inform policies regarding urban planning, environmental regulation and public health.

The Hedonic model is a regression model in which the dependent variable is the house price, and the independent variables are the various attributes of the house, such as the area of the house, the number of bedrooms, the number of parking spots and others, including the air quality attribute – expressed as the Air Quality Index (AQI). The model can take different forms, such as linear, logarithmic, or polynomial, depending on the nature of the relationship between the house prices and the attributes. For this research, several multiple regression models were created and their accuracy compared. The format of the regression is as follows:

$$\textbf{House Price} = \beta 0 + \beta 1\ \textbf{unit\_area} + \beta 2\ \textbf{numb\_rm} + \beta 3\ \textbf{Total\_park} + \beta 4\ \textbf{AQI} + \varepsilon$$

where:

House Price is the dependent variable.

β_i are the coefficients of the attributes.

ε is the error term

unit_area: is the size of the house in square meter (real number)

numb_rm: is the number of bedrooms in the house (integer)

Total_park: is the number of parking spots for the house (integer)

AQI: is the air quality index value for the transaction (real number)

Once the model forms are specified, we will use the appropriate statistical techniques to assess the robustness of the model. This includes running a regression analysis where the coefficients of the model indicate the “marginal price contribution” of each attribute. In our research, the coefficient of the AQI variable would indicate how changes in air quality are associated with changes in house prices.

In this chapter, the research methodology was discussed along with the steps performed for exploratory data analysis, data clean-up and data preparation, formatting the data into a file that can be used with the data analysis tools. The hedonic price method (HPM) was explained and the research hypothesis defined.

In the next chapter, the data analysis steps will be discussed.

Chapter 5. Data Findings

5.1 Introduction

Starting from the data set in Chapter 3, which contains the quantitative data to be used for the analysis, the following fields were further eliminated as they are not relevant or correlated with other fields: “transaction_id” , “trans_date” and “air_quality_stn”. For “transaction_id”, this attribute is not relevant to the analysis. For “trans_date”, our interest is in the year value which was extracted into a separate attribute and for the “air_quality_stn”, it has a 1-1 correlation with the AQI attribute. Accordingly, the resulting data set sample is shown below.

Table 7 Housing Sales dataset ready for analysis

trans_year	area_name	price	unit_area	numb_rm	Total_park	AQI
2021	Al Jadaf	1250000	119.74	2	1	27.78
2021	Al Jadaf	852600	82.22	2	1	27.78
2021	Al Jadaf	850000	122.62	2	1	27.78
2021	Al Jadaf	840000	123.37	2	1	27.78
2021	Al Jadaf	879103	148.85	2	1	27.78
2021	Al Jadaf	954000	148.85	2	1	27.78
2021	Al Jadaf	1870000	285.4	2	1	27.78
2021	Al Jadaf	820000	82.22	2	1	27.78
2021	Al Jadaf	684000	97.12	2	1	27.78
2021	Al Jadaf	1400000	130.86	2	1	27.78
2021	Al Jadaf	618750	96.04	2	1	27.78
2021	Al Jadaf	1400000	100.35	2	1	27.78
2021	Al Jadaf	850000	122.62	2	1	27.78
2021	Al Jadaf	850000	123.37	2	1	27.78

2021	Al Jadaf	800000	118.92	2	1	27.78
2021	Al Jadaf	2891662	187.4	2	1	27.78
2021	Al Jadaf	2938248	187.75	2	1	27.78
2021	Al Jadaf	850000	164.44	2	1	27.78
2021	Al Jadaf	850000	96.04	2	1	27.78
2021	Al Jadaf	1333760	118.03	2	1	27.78
2021	Al Jadaf	1600000	118.03	2	1	27.78
2021	Al Jadaf	562500	152.92	2	1	27.78
2021	Al Jadaf	562500	152.92	2	1	27.78

Further plots using Tableau were generated to understand the distribution of the transactions on the different areas.

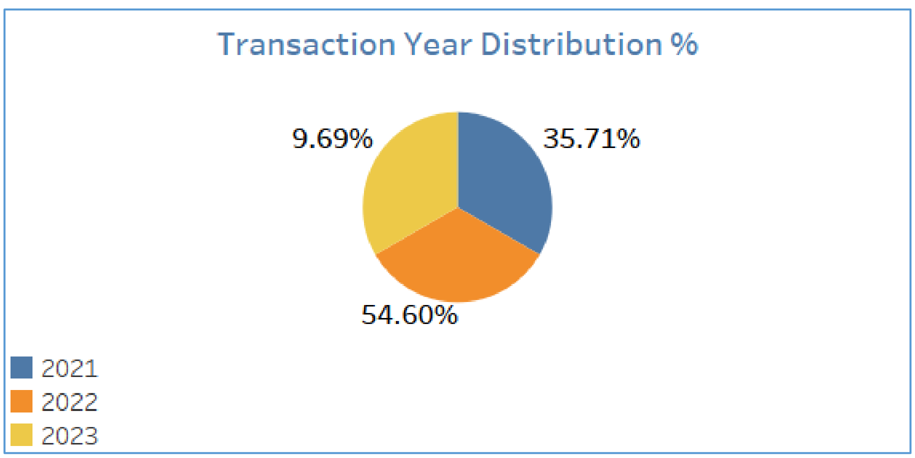


Figure 4 Transactions Distribution per Year

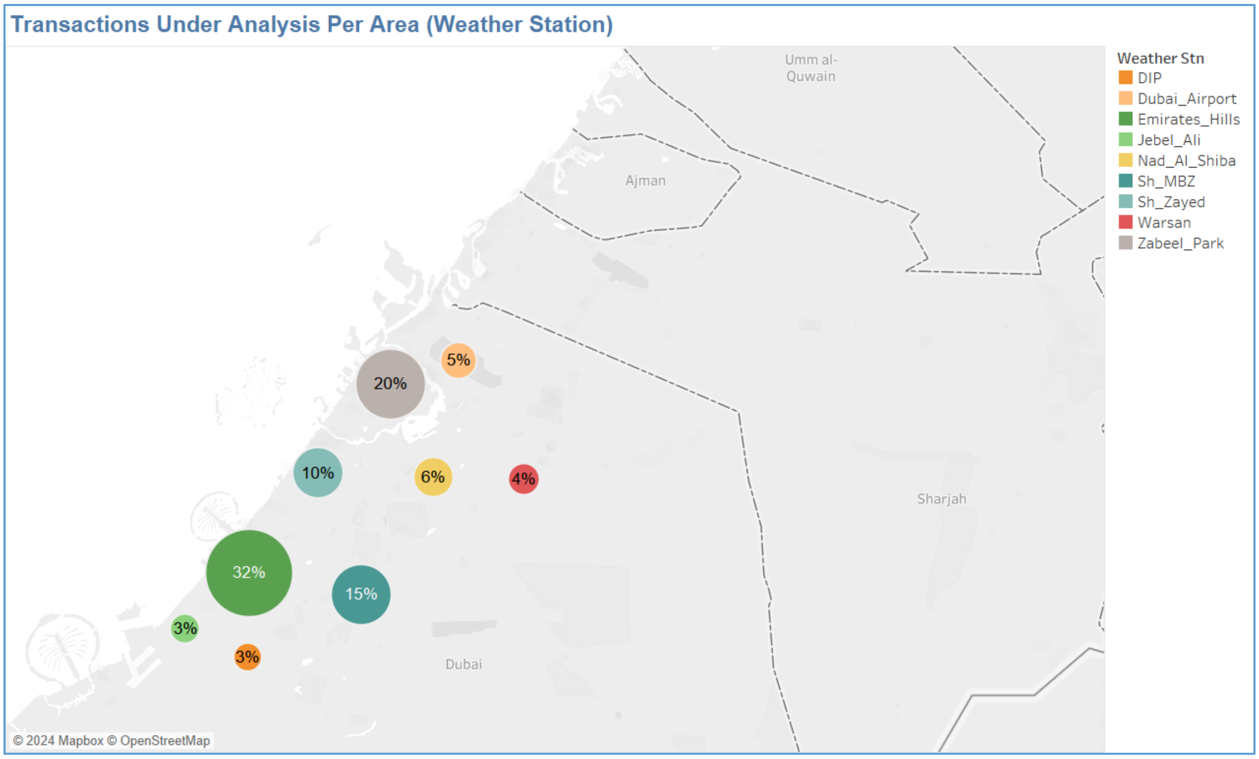


Figure 5 Transactions Under Analysis Per Weather Station

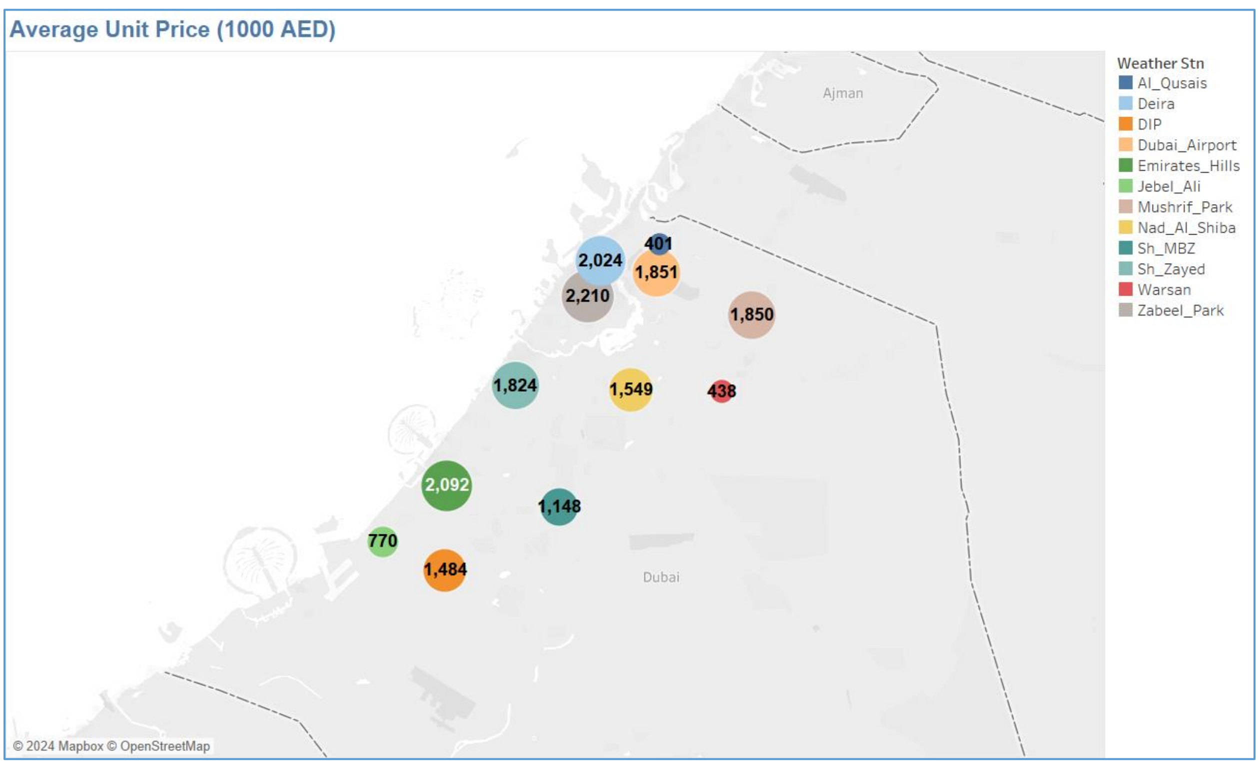
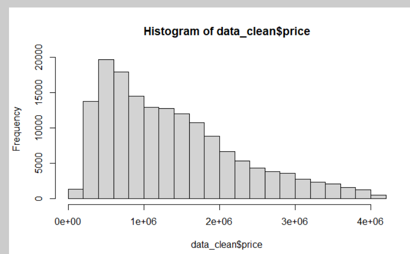
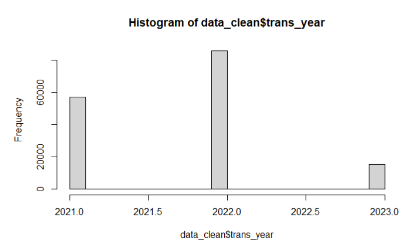
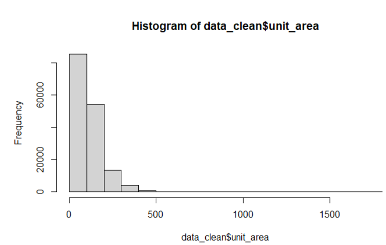


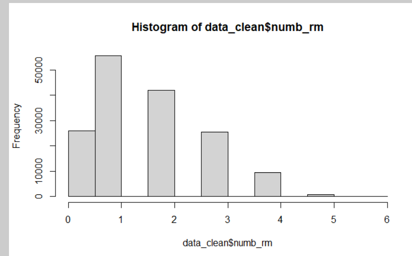
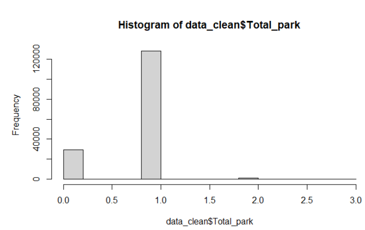
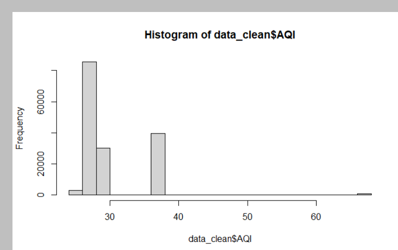
Figure 6 Average Unit Price (1K AED) per Weather Station Area

5.2 Data Attributes Analysis

- Conversion of the categorical variable “area_name” into a factor type.
- Since the goal is to understand how air quality (AQI) affects house prices, “trans_year” will be considered as a control variable in the regression models. This will take into account any general trends in housing prices over time that are not related to air quality. For the purpose of this analysis, “trans_year” will be considered as a continuous variable.
- Plotting a histogram for each attribute. Results as follows showing the data is skewed:

Table 8 Histograms of the Attributes

Attribute	Distribution Type	Histogram
price	Right-skewed distribution	
Trans_year	The highest number of transactions are from the year 2022	
Unit_area	Right-skewed distribution	

Attribute	**Distribution Type**	**Histogram**
Numb_rm	Right-skewed distribution	
Total_park		
AQI	Right-skewed distribution	

5.3 Outlier Analysis

Outlier analysis and removal: the 'ggplot' function in R was used to visualize the outliers. Result of the plot is as follows, showing the presence of outliers in "price".

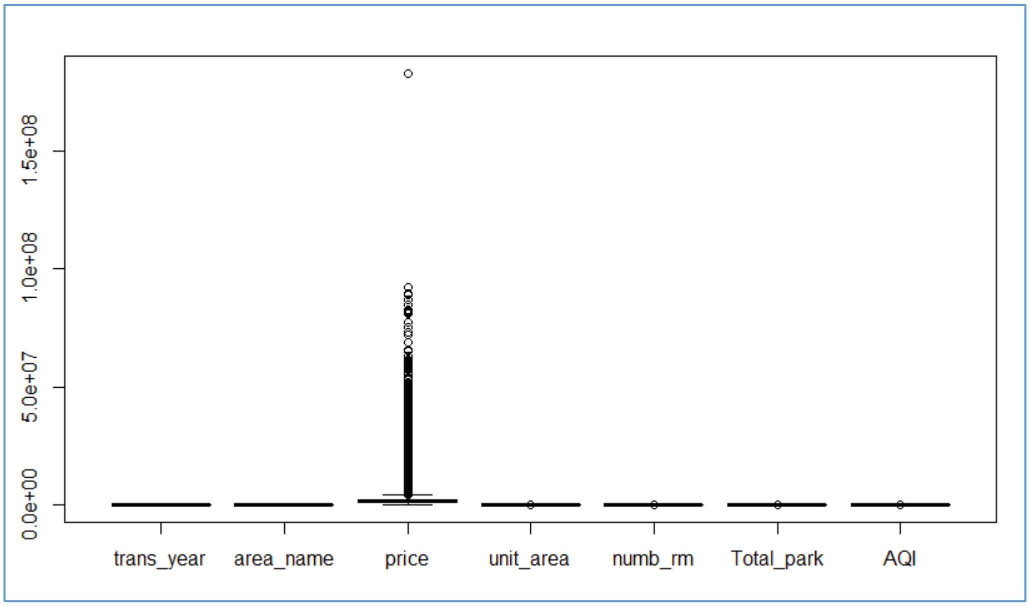


Figure 7 Outlier Identification: Boxplot of Attributes

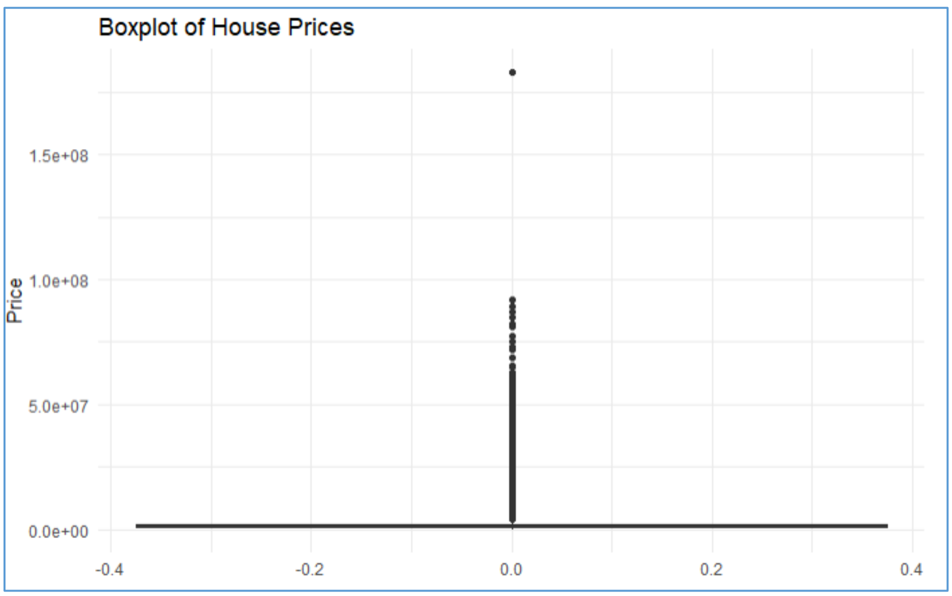


Figure 8 Boxplot for House Prices before Outlier Removal

Since this the data is showing skewed distributions, the Inter-Quartile Range (IQR) method is used to remove the outliers. Below the boxplot of the "price" after outlier removal.

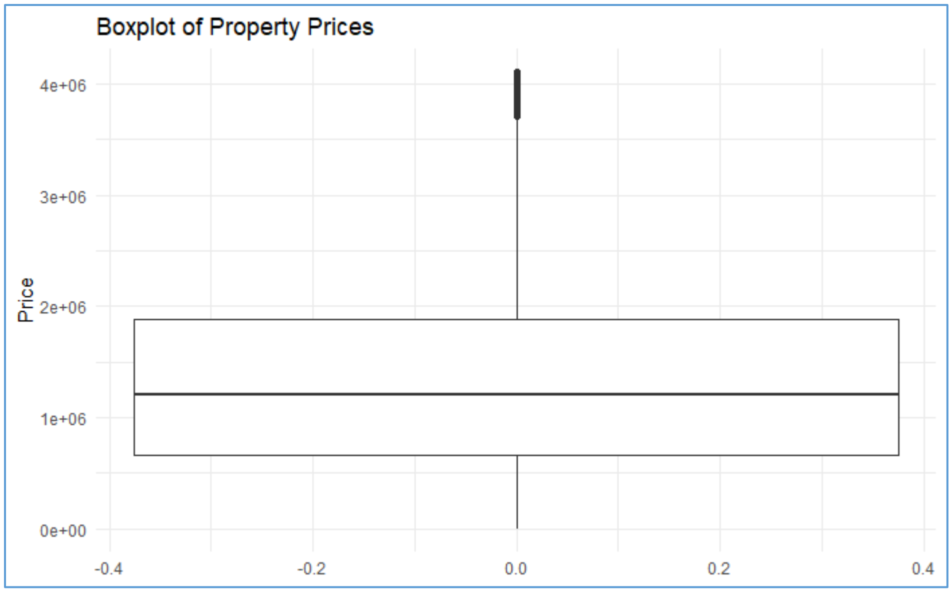


Figure 9 Boxplot for House Prices after Outlier Removal

5.4 Correlation Analysis

Create the correlation matrix between the numerical variables to check whether there is a correlation. Results were as shown in the table below. A strong correlation is observed between the "unit area" and the "number of rooms", which is logical as a larger area means more available rooms.

Table 9 Attributes Correlation Matrix

	trans_year	unit_area	numb_rm	Total_park	AQI
trans_year	1.00000000 -	0.04449603	0.05806134	0.16140812 -	0.43191114 -
unit_area	0.04449603 -	1.00000000	**0.75333566**	0.32334927 -	0.09784799 -
numb_rm	0.05806134	**0.75333566**	1.00000000 -	0.38454327	0.08550906
Total_park	0.1614081	0.3233493	0.3845433 -	1.0000000	0.1946215
AQI	0.43191114	0.09784799	0.08550906	0.19462151	1.00000000

Create a correlation plot for the numerical values to visualize the correlation. The plot shows a strong positive correlation between "unit_area" and the number of bedrooms, and shows a weak positive correlation between the AQI and the price of the unit. The plot also shows a positive correlation between the "trans_year" and the AQI but a weak positive correlation with the price, which is counter intuitive because higher AQI (more pollution) should result in a lower house price.

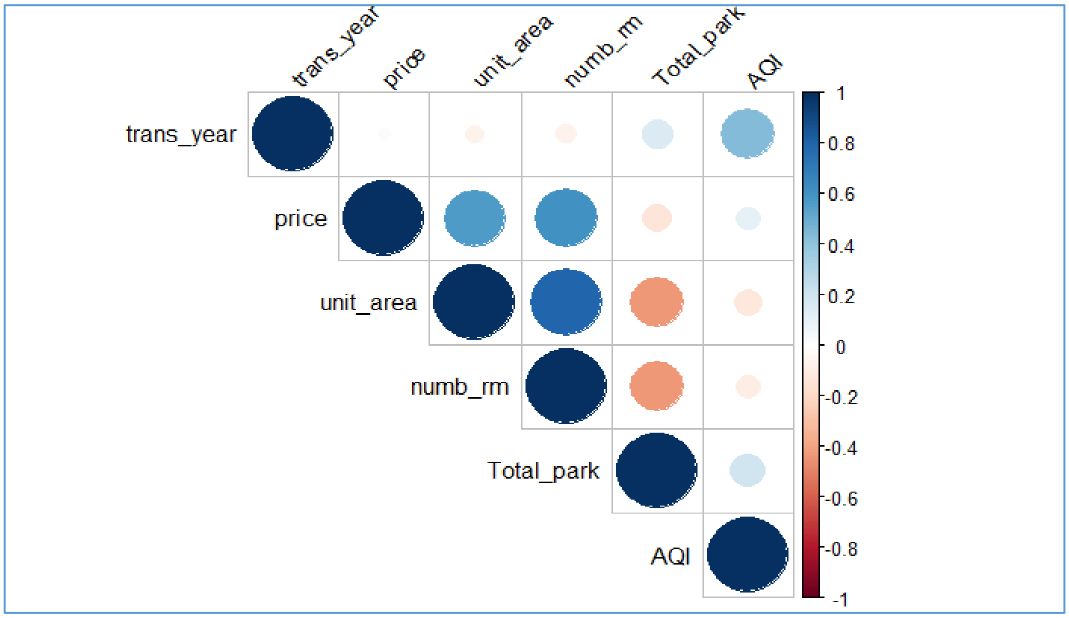


Figure 10 Correlation Plot of Attributes

- Splitting the dataset into a Training dataset (80%) and a Test dataset (20%). This will allow us to test for any overfitting on the training data.

5.5 Building the Hedonic Pricing Models

5.5.1 Linear Regression Model 1 (Ordinary Least Squares)

Using the R-studio software, the following OLS linear regression model was created for the dataset above. As there is a strong correlation between "unit_area" and "numb_rm" and to avoid Multicollinearity, the "unit_area" was dropped from the linear regressions. The following is the output showing the coefficients of the attributes and the OLS model statistics:

```
Residuals:
      Min         1Q     Median         3Q        Max
-14854940    -580532     -68239     394405  173078623
```

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	-525680551	29731172	-17.681	< 2e-16	***
trans_year	260340	14715	17.693	< 2e-16	***
area_nameAl Barsha First	-1430810	470529	-3.041	0.002360	**

area_nameAl Barsha South Fifth	-1165784	390294	-2.987	0.002818	**
area_nameAl Barsha South Fourth	-1181616	382923	-3.086	0.002031	**
area_nameAl Barshaa South Second	-954451	392274	-2.433	0.014971	*
area_nameAl Barshaa South Third	-1023447	383990	-2.665	0.007693	**
area_nameAL FURJAN	-1707973	395030	-4.324	1.54e-05	***
area_nameAl Goze Fourth	-1264552	408879	-3.093	0.001984	**
area_nameAl Hebiah Fifth	-1794238	388846	-4.614	3.95e-06	***
area_nameAl Hebiah First	-1677862	387695	-4.328	1.51e-05	***
area_nameAl Hebiah Fourth	-1319498	384734	-3.430	0.000605	***
area_nameAl Hebiah Second	-978560	399323	-2.451	0.014265	*
area_nameAl Hebiah Sixth	-1652206	392446	-4.210	2.56e-05	***
area_nameAl Hebiah Third	-975792	385943	-2.528	0.011462	*
area_nameAl Jadaf	-1356026	386132	-3.512	0.000445	***
area_nameAL KHAIL HEIGHTS	-2318577	575674	-4.028	5.64e-05	***
area_nameAl Khairan First	-778041	383630	-2.028	0.042552	*
area_nameAl Kheeran	-978673	432853	-2.261	0.023762	*
area_nameAl Kifaf	-767902	386697	-1.986	0.047058	*
area_nameAl Merkadh	-706943	383896	-1.841	0.065551	.
area_nameAl Qusais Industrial Fourth	-1658844	773771	-2.144	0.032047	*
area_nameAl Safouh First	-1594272	449603	-3.546	0.000391	***
area_nameAl Safouh Second	2032837	395496	5.140	2.75e-07	***
area_nameAl Thanayah Fourth	-1164497	387622	-3.004	0.002663	**
area_nameAl Thanyah Fifth	-1266154	383924	-3.298	0.000974	***
area_nameAl Thanyah First	-764380	398309	-1.919	0.054978	.
area_nameAl Thanyah Third	-1057451	385266	-2.745	0.006057	**
area_nameAL WAHA	-1981908	1226290	-1.616	0.106057	
area_nameAl Warsan First	-1174738	385048	-3.051	0.002282	**
area_nameAl Wasl	56110	385481	0.146	0.884271	
area_nameAl Yelayiss 1	-2026237	390321	-5.191	2.09e-07	***
area_nameAl Yelayiss 2	-2032205	385307	-5.274	1.34e-07	***
area_nameAl Yufrah 1	-2187688	389184	-5.621	1.90e-08	***

area_nameARABIAN RANCHES I	-2381944	444251	-5.362	8.26e-08	***
area_nameARABIAN RANCHES II	-2402401	568252	-4.228	2.36e-05	***
area_nameARABIAN RANCHES III	-2510907	439449	-5.714	1.11e-08	***
area_nameARJAN	-1461848	406837	-3.593	0.000327	***
area_nameBARSHA HEIGHTS	-1495735	495553	-3.018	0.002542	**
area_nameBLUEWATERS	4406382	472243	9.331	< 2e-16	***
area_nameBurj Khalifa	361240	383706	0.941	0.346476	
area_nameBURJ KHALIFA	134106	388537	0.345	0.729978	
area_nameBusiness Bay	-414366	383487	-1.081	0.279912	
area_nameBUSINESS BAY	-921897	387231	-2.381	0.017279	*
area_nameCITY OF ARABIA	-1480660	1477457	-1.002	0.316264	
area_nameCITY WALK	-13769	453256	-0.030	0.975766	
area_nameDAMAC HILLS	-1198049	398927	-3.003	0.002672	**
area_nameDISCOVERY GARDENS	-995963	422901	-2.355	0.018520	*
area_nameDOWN TOWN JABAL ALI	-1540168	583382	-2.640	0.008290	**
area_nameDUBAI CREEK HARBOUR	-958044	406907	-2.354	0.018551	*
area_nameDUBAI HARBOUR	725662	454946	1.595	0.110704	
area_nameDUBAI HEALTHCARE CITY - PHASE 1	303903	2054273	0.148	0.882393	
area_nameDUBAI HEALTHCARE CITY - PHASE 2	-1443337	436418	-3.307	0.000942	***
area_nameDUBAI HILLS	-1209901	393924	-3.071	0.002131	**
area_nameDUBAI INDUSTRIAL CITY	-2916158	1226534	-2.378	0.017429	*
area_nameDubai Investment Park First	-1323145	390607	-3.387	0.000706	***
area_nameDUBAI INVESTMENT PARK FIRST	-1371397	444793	-3.083	0.002048	**
area_nameDubai Investment Park Second	-1886940	456840	-4.130	3.62e-05	***
area_nameDUBAI INVESTMENT PARK SECOND	-1906980	583359	-3.269	0.001080	**
area_nameDUBAI LAND RESIDENCE COMPLEX	-1819076	407266	-4.467	7.96e-06	***
area_nameDUBAI MARINA	-1139527	387158	-2.943	0.003248	**
area_nameDUBAI MARITIME CITY	-632306	545044	-1.160	0.246010	
area_nameDUBAI PRODUCTION CITY	-1303465	407514	-3.199	0.001381	**
area_nameDUBAI SCIENCE PARK	-1561016	473352	-3.298	0.000975	***
area_nameDUBAI SOUTH	-1651366	431100	-3.831	0.000128	***

area_nameDUBAI SPORTS CITY	-1594030	398113	-4.004	6.23e-05	***
area_nameDUBAI STUDIO CITY	-1515967	600057	-2.526	0.011526	*
area_nameDUBAI WATER CANAL	49496326	2054399	24.093	< 2e-16	***
area_nameDUBAI WATER FRONT	-1895124	853431	-2.221	0.026380	*
area_nameEMAAR SOUTH	-2637437	420233	-6.276	3.48e-10	***
area_nameEMIRATE LIVING	-1413373	413332	-3.419	0.000628	***
area_nameGRAND VIEWS	-1177627	773646	-1.522	0.127967	
area_nameHadaeq Sheikh Mohammed Bin Rashid	-1099497	384272	-2.861	0.004220	**
area_nameHessyan First	-1649833	576105	-2.864	0.004187	**
area_nameINTERNATIONAL CITY PH 1	-1404791	393722	-3.568	0.000360	***
area_nameINTERNATIONAL CITY PH 2 & 3	-1679103	497579	-3.375	0.000740	***
area_nameIsland 2	11690051	433501	26.967	< 2e-16	***
area_nameJabal Ali First	-1220877	384819	-3.173	0.001511	**
area_nameJabal Ali Industrial Second	-798536	402278	-1.985	0.047142	*
area_nameJADDAF WATERFRONT	-1488873	478198	-3.114	0.001849	**
area_nameJUMEIRA BAY	9651038	853383	11.309	< 2e-16	***
area_nameJUMEIRAH BEACH RESIDENCE	-130391	401293	-0.325	0.745237	
area_nameJumeirah First	587418	388681	1.511	0.130712	
area_nameJUMEIRAH GOLF	-2186624	495506	-4.413	1.02e-05	***
area_nameJUMEIRAH HEIGHTS	-1527169	600463	-2.543	0.010982	*
area_nameJUMEIRAH LAKES TOWERS	-1460972	395452	-3.694	0.000220	***
area_nameJUMEIRAH LIVING	46688	523057	0.089	0.928875	
area_nameJumeirah Second	20870589	479572	43.519	< 2e-16	***
area_nameJUMEIRAH VILLAGE CIRCLE	-1456859	387118	-3.763	0.000168	***
area_nameJUMEIRAH VILLAGE TRIANGLE	-1527209	434372	-3.516	0.000438	***
area_nameLA MER	820294	446325	1.838	0.066082	.
area_nameLIVING LEGENDS	-2378922	853000	-2.789	0.005290	**
area_nameLIWAN	-2237119	468732	-4.773	1.82e-06	***
area_nameMadinat Al Mataar	-2037172	385645	-5.283	1.28e-07	***
area_nameMadinat Dubai Almelaheyah	-701087	401665	-1.745	0.080908	.
area_nameMADINAT HIND 4	-1171087	401007	-2.920	0.003497	**

area_nameMAJAN	-2170846	462134	-4.697	2.64e-06	***
area_nameMarsa Dubai	-104441	382720	-0.273	0.784938	
area_nameMBR DISTRICT 1	-762458	523071	-1.458	0.144938	
area_nameMe'Aisem First	-1211400	385334	-3.144	0.001668	**
area_nameMEYDAN AVENUE	-1418744	438133	-3.238	0.001203	**
area_nameMEYDAN ONE	-986041	461662	-2.136	0.032694	*
area_nameMIRA	-2541382	421685	-6.027	1.68e-09	***
area_nameMirdif	-1218315	391080	-3.115	0.001838	**
area_nameMOTOR CITY	-2126340	417267	-5.096	3.48e-07	***
area_nameMUDON	-1493794	2054274	-0.727	0.467127	
area_nameMuhaisanah First	-1084236	392206	-2.764	0.005703	**
area_nameNad Al Shiba First	-1153494	392136	-2.942	0.003266	**
area_nameNadd Hessa	-1389505	385505	-3.604	0.000313	***
area_namePalm Jumeirah	2340517	383375	6.105	1.03e-09	***
area_namePALM JUMEIRAH	1667612	392183	4.252	2.12e-05	***
area_namePEARL JUMEIRA	3371758	2054243	1.641	0.100725	
area_nameRega Al Buteen	-1821726	545528	-3.339	0.000840	***
area_nameREMRAAM	-2067882	427418	-4.838	1.31e-06	***
area_nameSaih Shuaib 1	-2177596	2054566	-1.060	0.289201	
area_nameSaih Shuaib 2	-2428743	562511	-4.318	1.58e-05	***
area_nameSERENA	-2276616	437006	-5.210	1.90e-07	***
area_nameSILICON OASIS	-1680920	398513	-4.218	2.47e-05	***
area_nameSOBHA HEARTLAND	-1380463	404183	-3.415	0.000637	***
area_nameSUFOUH GARDENS	-1850888	610361	-3.032	0.002426	**
area_nameSUSTAINABLE CITY	-2000795	526527	-3.800	0.000145	***
area_nameTECOM SITE A	229482	2054409	0.112	0.911060	
area_nameTHE GREENS	-1471461	401291	-3.667	0.000246	***
area_nameTHE LAKES	-711500	473809	-1.502	0.133187	
area_nameTILAL AL GHAF	-2777734	809345	-3.432	0.000599	***
area_nameTOWN SQUARE	-2485245	405736	-6.125	9.08e-10	***
area_nameTrade Center Second	-1132244	466436	-2.427	0.015207	*

area_nameUm Hurair Second	-536461	774308	-0.693	0.488420	
area_nameUm Suqaim Third	83660	387624	0.216	0.829123	
area_nameVILLANOVA	-2730398	403190	-6.772	1.28e-11	***
area_nameWadi Al Safa 2	-1617350	392778	-4.118	3.83e-05	***
area_nameWadi Al Safa 3	-809377	387834	-2.087	0.036898	*
area_nameWadi Al Safa 4	-2128107	980201	-2.171	0.029926	*
area_nameWadi Al Safa 5	-1803478	384428	-4.691	2.72e-06	***
area_nameWadi Al Safa 6	-1378699	393618	-3.503	0.000461	***
area_nameWadi Al Safa 7	-1541562	385809	-3.996	6.45e-05	***
area_nameWarsan Fourth	-1178292	398705	-2.955	0.003124	**
area_nameWorld Islands	1085547	434830	2.496	0.012544	*
area_nameZaabeel First	5045738	428485	11.776	< 2e-16	***
area_nameZaabeel Second	-430460	394248	-1.092	0.274901	
numb_rm	1016447	6162	164.968	< 2e-16	***
Total_park	315698	21341	14.793	< 2e-16	***
AQI	-4429	2586	-1.712	0.086809	.

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2018000 on 127016 degrees of freedom

Multiple R-squared: **0.368**,   Adjusted R-squared: **0.3672**

F-statistic: 520.7 on 142 and 127016 DF, p-value: < 2.2e-16

Generating the Q-Q (Quantile-Quantile) plot to assess whether the residuals of the OLS regression model follow a normal distribution, which is an important assumption of the linear regression.

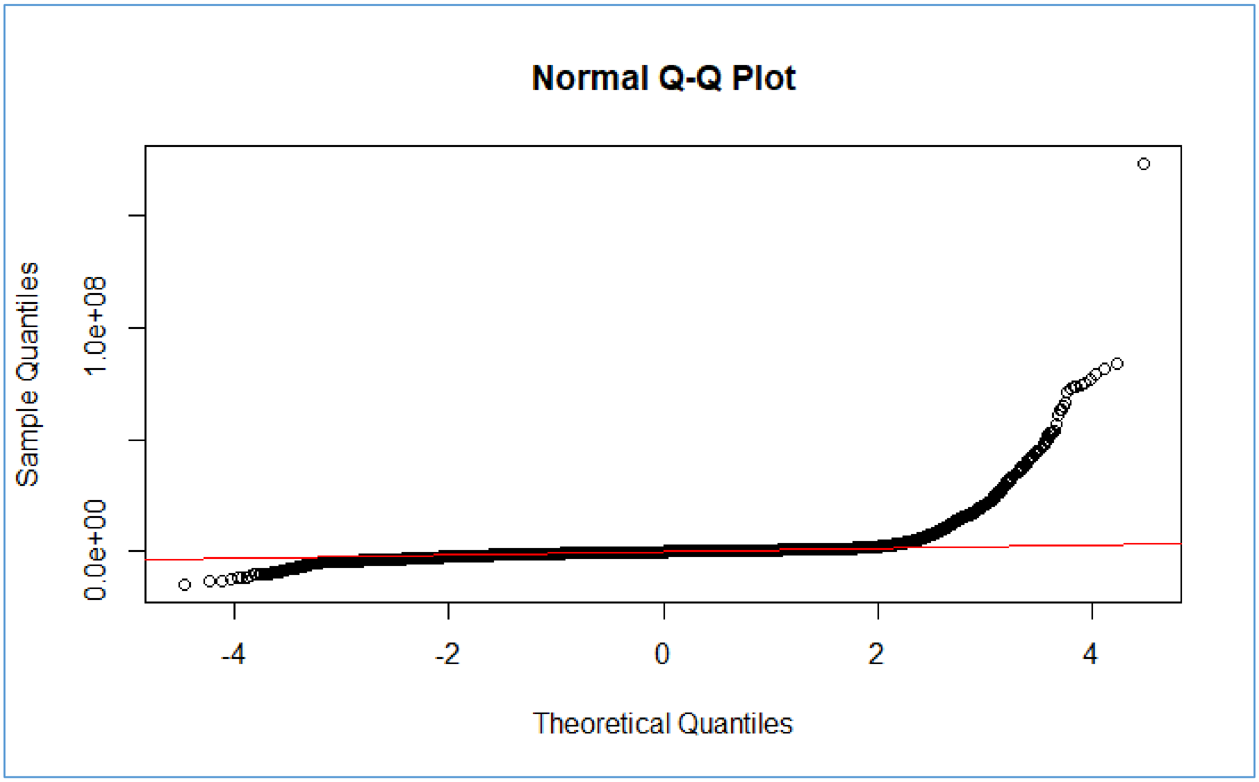


Figure 11 Q-Q Plot of the OLS Regression

5.5.2 Linear Regression Model 2 (Log-Linear regression)

As per Fenwick (2013), when certain variables are lacking in the dataset, the log-linear regression has been found to perform reasonably well. Assuming that the relationship between the independent variables and the log of the dependent variable (price) is linear, a log-Linear regression was created for the dataset. The "unit_area" attribute was also dropped from this regression to avoid multicollinearity. The statistics of the model are as follows:

```
Residuals:
     Min        1Q    Median        3Q       Max
-14.5022   -0.1889   0.0181    0.2283    1.5485
```

	Estimate	Std. Error	t value	Pr(>|t|)
Coefficients (2 not defined because of singularities)				
(Intercept)	-1.418e+02	5.663e+00	-25.049	< 2e-16 ***

trans_year	7.692e-02	2.805e-03	27.421	< 2e-16	***
area_nameAL.BARARI	-6.604e-02	8.354e-02	-0.791	0.429230	
area_nameAl.Barsha.First	-8.433e-01	5.378e-02	-15.680	< 2e-16	***
area_nameAl.Barsha.South.Fifth	-6.904e-01	2.223e-02	-31.050	< 2e-16	***
area_nameAl.Barsha.South.Fourth	-8.361e-01	1.763e-02	-47.426	< 2e-16	***
area_nameAl.Barshaa.South.Second	-6.338e-01	2.418e-02	-26.214	< 2e-16	***
area_nameAl.Barshaa.South.Third	-8.125e-01	1.853e-02	-43.850	< 2e-16	***
area_nameAL.FURJAN	-1.014e+00	2.591e-02	-39.154	< 2e-16	***
area_nameAl.Goze.Fourth	-1.077e+00	3.139e-02	-34.322	< 2e-16	***
area_nameAl.Hebiah.Fifth	-1.271e+00	2.187e-02	-58.124	< 2e-16	***
area_nameAl.Hebiah.First	-7.366e-01	2.074e-02	-35.509	< 2e-16	***
area_nameAl.Hebiah.Fourth	-9.998e-01	1.860e-02	-53.754	< 2e-16	***
area_nameAl.Hebiah.Second	-8.920e-01	2.813e-02	-31.709	< 2e-16	***
area_nameAl.Hebiah.Sixth	-5.694e-01	2.403e-02	-23.698	< 2e-16	***
area_nameAl.Hebiah.Third	-6.257e-01	1.993e-02	-31.396	< 2e-16	***
area_nameAl.Jadaf	-6.993e-01	1.971e-02	-35.473	< 2e-16	***
area_nameAL.KHAIL.HEIGHTS	-1.092e+00	7.870e-02	-13.875	< 2e-16	***
area_nameAl.Khairan.First	-1.407e-01	1.821e-02	-7.723	1.15e-14	***
area_nameAl.Kheeran	-3.195e-01	4.383e-02	-7.290	3.12e-13	***
area_nameAl.Kifaf	-1.286e-01	2.068e-02	-6.217	5.07e-10	***
area_nameAl.Merkadh	-3.504e-01	1.781e-02	-19.677	< 2e-16	***
area_nameAl.Qusais.Industrial.Fourth	-1.420e+00	1.395e-01	-10.184	< 2e-16	***
area_nameAl.Safouh.First	-7.684e-01	4.650e-02	-16.527	< 2e-16	***
area_nameAl.Safouh.Second	4.054e-01	2.794e-02	14.511	< 2e-16	***
area_nameAl.Thanayah.Fourth	-4.910e-01	2.040e-02	-24.066	< 2e-16	***
area_nameAl.Thanyah.Fifth	-6.375e-01	1.811e-02	-35.205	< 2e-16	***
area_nameAl.Thanyah.First	-5.231e-01	2.674e-02	-19.565	< 2e-16	***
area_nameAl.Thanyah.Third	-4.815e-01	1.949e-02	-24.702	< 2e-16	***
area_nameAL.WAHA	-7.912e-01	2.267e-01	-3.490	0.000484	***
area_nameAl.Warsan.First	-1.469e+00	1.896e-02	-77.460	< 2e-16	***
area_nameAl.Wasl	8.272e-02	1.915e-02	4.320	1.56e-05	***

area_nameAl.Yelayiss.1	-8.671e-01	2.167e-02	-40.007	< 2e-16	***
area_nameAl.Yelayiss.2	-9.954e-01	1.890e-02	-52.671	< 2e-16	***
area_nameAl.Yufrah.1	-9.146e-01	2.123e-02	-43.093	< 2e-16	***
area_nameARABIAN.RANCHES.I	-6.576e-01	4.674e-02	-14.070	< 2e-16	***
area_nameARABIAN.RANCHES.II	-6.713e-01	7.740e-02	-8.673	< 2e-16	***
area_nameARABIAN.RANCHES.III	-7.231e-01	4.578e-02	-15.795	< 2e-16	***
area_nameARJAN	-1.057e+00	3.191e-02	-33.135	< 2e-16	***
area_nameBARSHA.HEIGHTS	-6.940e-01	6.354e-02	-10.922	< 2e-16	***
area_nameBLUEWATERS	3.791e-01	1.395e-01	2.718	0.006572	**
area_nameBurj.Khalifa	-1.439e-02	1.775e-02	-0.811	0.417514	
area_nameBURJ.KHALIFA	-2.241e-02	2.366e-02	-0.947	0.343582	
area_nameBusiness.Bay	-2.424e-01	1.746e-02	-13.878	< 2e-16	***
area_nameBUSINESS.BAY	-3.333e-01	2.182e-02	-15.274	< 2e-16	***
area_nameCITY.OF.ARABIA	-8.116e-01	2.267e-01	-3.579	0.000344	***
area_nameCITY.WALK	2.571e-01	5.814e-02	4.423	9.75e-06	***
area_nameDAMAC.HILLS	-6.352e-01	2.857e-02	-22.236	< 2e-16	***
area_nameDISCOVERY.GARDENS	-1.354e+00	4.036e-02	-33.559	< 2e-16	***
area_nameDOWN.TOWN.JABAL.ALI	-1.220e+00	8.717e-02	-13.995	< 2e-16	***
area_nameDUBAI.CREEK.HARBOUR	-4.878e-01	3.034e-02	-16.079	< 2e-16	***
area_nameDUBAI.HARBOUR	1.424e-01	6.215e-02	2.291	0.021986	*
area_nameDUBAI.HEALTHCARE.CITY...PHASE.1	4.641e-01	3.919e-01	1.184	0.236347	
area_nameDUBAI.HEALTHCARE.CITY...PHASE.2	-7.731e-01	4.390e-02	-17.611	< 2e-16	***
area_nameDUBAI.HILLS	-5.276e-01	2.607e-02	-20.237	< 2e-16	***
area_nameDUBAI.INDUSTRIAL.CITY	-1.689e+00	1.965e-01	-8.595	< 2e-16	***
area_nameDubai.Investment.Park.First	-6.688e-01	2.173e-02	-30.777	< 2e-16	***
area_nameDUBAI.INVESTMENT.PARK.FIRST	-5.127e-01	5.426e-02	-9.449	< 2e-16	***
area_nameDubai.Investment.Park.Second	-1.223e+00	4.836e-02	-25.281	< 2e-16	***
area_nameDUBAI.INVESTMENT.PARK.SECOND	-1.434e+00	8.524e-02	-16.817	< 2e-16	***
area_nameDUBAI.LAND.RESIDENCE.COMPLEX	-1.259e+00	3.360e-02	-37.479	< 2e-16	***
area_nameDUBAI.MARINA	-4.929e-01	2.097e-02	-23.502	< 2e-16	***
area_nameDUBAI.MARITIME.CITY	-1.235e-01	7.473e-02	-1.653	0.098288	.

area_nameDUBAI.PRODUCTION.CITY	-1.298e+00	3.210e-02	-40.423	< 2e-16	***
area_nameDUBAI.SCIENCE.PARK	-8.588e-01	5.478e-02	-15.678	< 2e-16	***
area_nameDUBAI.SOUTH	-1.252e+00	4.145e-02	-30.206	< 2e-16	***
area_nameDUBAI.SPORTS.CITY	-1.232e+00	2.829e-02	-43.571	< 2e-16	***
area_nameDUBAI.STUDIO.CITY	-1.003e+00	8.724e-02	-11.496	< 2e-16	***
area_nameDUBAI.WATER.FRONT	-1.576e+00	1.490e-01	-10.579	< 2e-16	***
area_nameEMAAR.SOUTH	-1.008e+00	3.766e-02	-26.768	< 2e-16	***
area_nameEMIRATE.LIVING	-4.149e-01	3.435e-02	-12.076	< 2e-16	***
area_nameGRAND.VIEWS	-4.463e-01	1.760e-01	-2.536	0.011225	*
area_nameHadaeq.Sheikh.Mohammed.Bin.Rashid	-4.987e-01	1.789e-02	-27.878	< 2e-16	***
area_nameHessyan.First	-1.451e+00	8.340e-02	-17.398	< 2e-16	***
area_nameINTERNATIONAL.CITY.PH.1	-1.543e+00	2.568e-02	-60.085	< 2e-16	***
area_nameINTERNATIONAL.CITY.PH.2...3	-1.408e+00	6.099e-02	-23.081	< 2e-16	***
area_nameIsland.2	2.415e-01	2.267e-01	1.065	0.286738	
area_nameJabal.Ali.First	-9.884e-01	1.835e-02	-53.877	< 2e-16	***
area_nameJabal.Ali.Industrial.Second	-1.016e+00	2.925e-02	-34.747	< 2e-16	***
area_nameJADDAF.WATERFRONT	-4.836e-01	5.924e-02	-8.164	3.26e-16	***
area_nameJUMEIRAH.BEACH.RESIDENCE	-2.924e-01	3.164e-02	-9.242	< 2e-16	***
area_nameJumeirah.First	1.507e-01	2.208e-02	6.825	8.82e-12	***
area_nameJUMEIRAH.GOLF	-8.340e-01	6.665e-02	-12.513	< 2e-16	***
area_nameJUMEIRAH.HEIGHTS	-3.442e-01	8.926e-02	-3.856	0.000115	***
area_nameJUMEIRAH.LAKES.TOWERS	-7.526e-01	2.639e-02	-28.518	< 2e-16	***
area_nameJUMEIRAH.LIVING	1.818e-01	7.598e-02	2.393	0.016726	*
area_nameJUMEIRAH.VILLAGE.CIRCLE	-9.626e-01	2.092e-02	-46.012	< 2e-16	***
area_nameJUMEIRAH.VILLAGE.TRIANGLE	-9.686e-01	4.433e-02	-21.851	< 2e-16	***
area_nameLA.MER	2.202e-01	5.567e-02	3.956	7.64e-05	***
area_nameLIVING.LEGENDS	-9.211e-01	1.395e-01	-6.601	4.10e-11	***
area_nameLIWAN	-1.440e+00	5.614e-02	-25.649	< 2e-16	***
area_nameMadinat.Al.Mataar	-1.019e+00	1.912e-02	-53.273	< 2e-16	***
area_nameMadinat.Dubai.Almelaheyah	-1.477e-01	2.848e-02	-5.186	2.15e-07	***
area_nameMADINAT.HIND.4	-1.070e+00	2.939e-02	-36.418	< 2e-16	***

area_nameMAJAN	-1.180e+00	5.133e-02	-22.978	< 2e-16	***
area_nameMarsa.Dubai	-1.512e-01	1.748e-02	-8.651	< 2e-16	***
area_nameMBR.DISTRICT.1	-1.809e-01	7.134e-02	-2.535	0.011240	*
area_nameMe.Aisem.First	-9.032e-01	1.924e-02	-46.953	< 2e-16	***
area_nameMEYDAN.AVENUE	-4.723e-01	4.469e-02	-10.569	< 2e-16	***
area_nameMEYDAN.ONE	-6.723e-01	5.343e-02	-12.581	< 2e-16	***
area_nameMIRA	-7.074e-01	3.851e-02	-18.368	< 2e-16	***
area_nameMirdif	-4.322e-01	2.335e-02	-18.509	< 2e-16	***
area_nameMOTOR.CITY	-8.701e-01	3.719e-02	-23.395	< 2e-16	***
area_nameMUDON	-8.125e-01	3.919e-01	-2.073	0.038152	*
area_nameMuhaisanah.First	-3.047e-01	2.366e-02	-12.880	< 2e-16	***
area_nameNad.Al.Shiba.First	-5.001e-01	2.313e-02	-21.618	< 2e-16	***
area_nameNadd.Hessa	-1.153e+00	1.965e-02	-58.643	< 2e-16	***
area_namePalm.Jumeirah	1.246e-01	1.844e-02	6.759	1.40e-11	***
area_namePALM.JUMEIRAH	4.564e-02	2.594e-02	1.760	0.078494	.
area_nameRega.Al.Buteen	-8.129e-01	8.520e-02	-9.541	< 2e-16	***
area_nameREMRAAM	-1.332e+00	4.069e-02	-32.737	< 2e-16	***
area_nameSaih.Shuaib.1	NA	NA	NA	NA	
area_nameSaih.Shuaib.2	-1.575e+00	8.719e-02	-18.060	< 2e-16	***
area_nameSERENA	-6.566e-01	4.193e-02	-15.662	< 2e-16	***
area_nameSILICON.OASIS	-1.239e+00	2.863e-02	-43.268	< 2e-16	***
area_nameSOBHA.HEARTLAND	-3.930e-01	3.197e-02	-12.293	< 2e-16	***
area_nameSUFOUH.GARDENS	-6.709e-01	8.524e-02	-7.871	3.54e-15	***
area_nameSUSTAINABLE.CITY	-4.802e-01	7.484e-02	-6.416	1.40e-10	***
area_nameTECOM.SITE.A	3.038e-01	3.919e-01	0.775	0.438231	
area_nameTHE.GREENS	-5.961e-01	2.921e-02	-20.406	< 2e-16	***
area_nameTHE.LAKES	-1.775e-01	6.667e-02	-2.663	0.007741	**
area_nameTILAL.AL.GHAF	-1.030e+00	1.608e-01	-6.406	1.50e-10	***
area_nameTOWN.SQUARE	-1.032e+00	3.160e-02	-32.663	< 2e-16	***
area_nameTrade.Center.Second	-2.980e-01	4.979e-02	-5.984	2.18e-09	***
area_nameUm.Hurair.Second	-2.732e-02	1.395e-01	-0.196	0.844707	

area_nameUm.Suqaim.Third	5.220e-02	2.105e-02	2.480	0.013137	*
area_nameVILLANOVA	-8.639e-01	3.125e-02	-27.644	< 2e-16	***
area_nameWadi.Al.Safa.2	-1.262e+00	2.415e-02	-52.258	< 2e-16	***
area_nameWadi.Al.Safa.3	-5.009e-01	2.181e-02	-22.963	< 2e-16	***
area_nameWadi.Al.Safa.4	-4.750e-01	2.267e-01	-2.095	0.036148	*
area_nameWadi.Al.Safa.5	-9.165e-01	1.822e-02	-50.295	< 2e-16	***
area_nameWadi.Al.Safa.6	-6.903e-01	2.342e-02	-29.479	< 2e-16	***
area_nameWadi.Al.Safa.7	-8.682e-01	1.926e-02	-45.066	< 2e-16	***
area_nameWarsan.Fourth	-1.400e+00	2.792e-02	-50.137	< 2e-16	***
area_nameWorld.Islands	2.322e-01	4.456e-02	5.210	1.89e-07	***
area_nameZaabeel.Second	NA	NA	NA	NA	
numb_rm	4.028e-01	1.211e-03	332.721	< 2e-16	***
Total_park	-9.355e-02	4.126e-03	-22.673	< 2e-16	***
AQI	7.053e-03	5.014e-04	14.066	< 2e-16	***

```
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
Residual standard error: 0.3915 on 127022 degrees of freedom
Multiple R-squared:  0.7059,   Adjusted R-squared:  0.7055
F-statistic:  2241 on 136 and 127022 DF,  p-value: < 2.2e-16
[1] "Calculated R-squared for the training set: 0.61110241411037"
[1] "Calculated R-squared for the test set: 0.608754206624008"
```

Generating the Q-Q (Quantile-Quantile) plot to assess whether the residuals of the Log-Linear regression model follow a normal distribution.

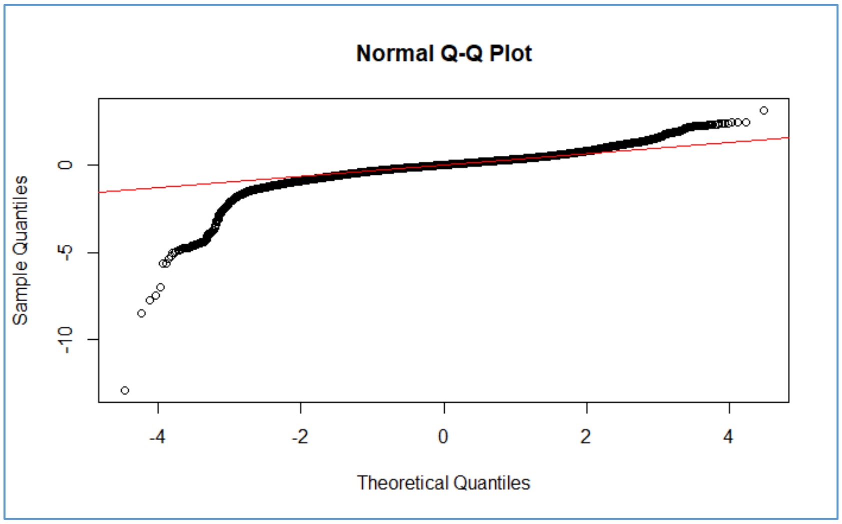


Figure 12 Q-Q Plot of the Log-Linear Regression

Chapter 6 Discussion

Referring to the research questions, this chapter will answer the questions based on the analysis and findings in chapter 4. For the primary research question, "*What is the relationship between air quality in a neighbourhood in Dubai and the property price in that neighbourhood?",* different models showed different answers to the question:

- Based on the output of linear regression 1 (the OLS model), the coefficient of the 'AQI' is negative (-4429), which indicates a negative relation between the house price and the air quality index. However, the p-value for AQI = 0.086809 (larger than 0.05) which indicates that AQI is statistically insignificant to the price.

  The Multiple R-squared for the training data is: 0.368, which is a low value indicating that the OLS model can only explain 36.8% of the variance in the house price. This is a low value for R^2 and means the model does not fit the data very well. Accordingly, the OLS model is not capturing the underlying relationships effectively, which will lead to poor predictions.

  The Test R-squared for the test data is: 0.3791481419829, which is very close to the R^2 of the training data indicating that there is no overfit of the model on the training data.

  The Quantile-Quantile (Q-Q) plot of the model shows that quantiles values (-2 to 2) show points that roughly follow the red reference line, suggesting that the middle portion of the data is approximately normally distributed. However, at the tails there is deviation from the red line, suggestion that the residuals have heavier tails than a normal distribution. This in turn is a violation of the normality assumption for the residuals of the OLS model and can affect the reliability of regression coefficients and standard errors. One potential way to address this situation is to transform the dependent variable using a log transformation to achieve a more normal distribution of residuals.

The next step was to build a model based on the log of the predicted value (price) as the relationship between the independent variable (price) and the dependent variables was non-linear.

- Based on the output of the linear regression 2 (log-linear model), the coefficient of ‘AQI’ is positive but very small (+0.007053), which indicates a very wek relation between the house price and the air quality index. The p-values for “numb_rum”, “total_park” and AQI are all < 0.05 which indicates that all of them are statistically significant to the price.
  The Multiple R-squared for the training data is: 0.7059, which is a relatively high value indicating that the model explains 70.59% of the variance in the house price. The calculated R-square for the training and test dataset is very similar indicating that there is no overfit for the model on the training data and that it can generalize on new data. The Q-Q plot of this model indicates that the residuals do not follow a perfect normal distribution. This could imply that the log-linear model may not be capturing all the nuances of the data, or there could be influential points that are affecting the model fit.

Based on the above, the OLS regression model determined that there is no relation between AQI and the house price. In the log-linear model, the AQI is statistically significant to the house price however, the relation between them is very weak.

For the secondary research questions:

- Which parameters of a property have a significant effect on its price?
- What challenges and limitations are present in the Hedonic Price Method and how can future research be enhanced?

The outputs of the OLS linear regression shows that most of the neighborhoods names attribute (area_nm) is statistically significant to the house price since their p-values are < 0.05. This is expected as the price of the house will depend on the neighborhood it is in. The transaction year (trans_year), number of rooms (numb_rm) and number of parking spaces (Total_park) are also statistically significant to the house price and all have positive coefficients in this model. This is logical as more rooms and parking spaces will increase the price of the house, and the price of the house might increase over time.

The outputs of the log-linear regression also show that most of the neighborhood's names attribute (area_nm) are statistically significant. The transaction year (trans_year) and number of rooms (numb_rm) have positive coefficients as expected. However, the number of parking spaces (Total_park) has a negative coefficient, which is counter-intuitive as more parking space should result in a higher price for the house.

Although the Hedonic Price Method (HPM) is considered the de-facto method in analyzing housing price against their characteristics based on the research conducted in this paper, this method has challenges and limitations. These can be summarized as follows:

- The HPM assumes that the price of the house is a function of its characteristics, and in our case, air quality was the characteristic of choice. The functional form of the relationship between house price and air quality is not known and could be non-linear. If this is not taken into consideration, the impact of air quality on house price may not be accurate. Therefore, the choice of model type is very important.
- The implementation of HPM requires access to accurate and comprehensive data on house prices, their characteristics, and external data such as air quality. If data points are missing, or the data is not accurate or not granular, this can affect the reliability of the results.
- The characteristics of a house can exert collinearity between attributes, which affects the accuracy of the model. This has been experienced in this research as the air quality data was correlated with the location data (area_name), leading to dropping one of the attributes altogether.
- HPM requires accurate data on house prices which are influenced by many market dynamics and can fluctuate over time, complicating the analysis. It also requires accurate data on air quality, which can vary widely over time and acroos

locations, and may not be available for all locations analyzed. Air quality is a characteristic of an area and the values collected will apply to many houses in the area, which complicates the analysis.

Chapter 7 Conclusions

8.1 Conclusions

In this study, the researcher succeeded in applying the hedonic pricing method model to the data on real estate prices and air quality in Dubai. The results show that there is a very weak correlation between air quality and the price of houses. However, these results should be considered as initial findings but could form the basis for further research in this area. The housing market in Dubai is a complex market that is influenced by many local and international factors. Further research in the domain needs to take these factors into account. The house characteristics used in the research are the basic characteristics of the unit, which is a limitation as the price of the house is influenced by several characteristics. These can be internal factors such as the age of the property, the height of the unit, the availability of services, as well as external factors such as proximity to landmarks, proximity to transportation systems and distance to the city center.

8.2 Recommendations

The availability of "open data" through the official Dubai government portals greatly facilitated this research, although the data attributes shared can be further enhanced for both the house sales and the air quality datasets. Getting more granular monthly data for the air quality stations would enable better association of the house price with the associated environmental data. For the house pricing dataset, having the GIS coordinates of the property would greatly help in associating the property with the nearest air quality station based on GIS proximity.

8.3 Future Work

It is important for future work to take into consideration a wider span of time, with at least 5 years of data. It is also important to consider the macroeconomic conditions, the real estate cycle and the view of investors in Dubai as factors affecting the house prices.

Further research in the field might benefit from qualitative data collection, such as surveys to study people's perception of air quality in their neighbourhoods. As shown in the literature review, this can be a valuable source of information for research in this area.

References

Amini, A., Nafari, K., & Singh, R. (2022). Effect of air pollution on house prices: Evidence from sanctions on Iran. *Regional Science and Urban Economics*, 93(103720), 103720. doi:https://doi.org/10.1016/j.regsciurbeco.2021.103720

Bayer, P., Keohane, N., & Timmins, C. (2009). Migration and Hedonic Valuation: The Case of Air Quality. *Journal of Environmental Economics and Management, 58(1)*, 1-14.

Bazyl, M. (2009). *Hedonic price model for Warsaw housing market.* Warsaw School of Economics, Department of Applied Econometrics. Warsaw: Department of Applied Econometrics.

Berezansky, B., Portnov, B., & Barzilai, B. (2010). PErceived Air Pollution as a Factor of Housing Pricing: A Case Study of the Greater Haifa Metropolitan Area. 18(1), 99–122. doi:https://doi.org/10.1080/10835547.2010.12090266

Brécard, D., Le Boennec, R., & Salladarré, F. (2019). Accessibility, local pollution and housing prices. Evidence from Nantes Métropole, France. *Economics and Statistics*, 97-115.

Carriazo, F., & Gomez-Mahecha, J. (2018). The demand for Air quality: evidence from the housing market in Bogotá, Colombia. *Environment and Development Economics*, 23(2), 121–138. doi:https://doi.org/10.1017/s1355770x18000050

Cebula, R. J. (2009). The Hedonic Pricing Model Applied to the Housing Market of the City of Savannah and Its Savannah Historic Landmark District. *The Review of Regional Studies; New Brunswick*, 9-22.

Chasco, C., & Gallo, J. (2013). The Impact of Objective and Subjective Measures of Air Quality and Noise on House Prices:A Multilevel Approach for Downtown Madrid. *Economic Geography*, 89(2), 127–148. doi:https://doi.org/10.1111/j.1944-8287.2012.01172.x

Chay, K., & Greenstone, M. (2005). Does air quality matter? Evidence from the housing market. *Journal of political Economy, 113(2)*, 376-424.

Chiarazzo, V., Coppola, P., Dell'Olio, L., Ibeas, A., & Ottomanelli, M. (2014). The Effects of Environmental Quality on Residential Choice Location. *Procedia-Social and Behavioral Sciences*(162), 178-187.

Dubai Media Office, O. (2023, January 16). *Dubai Media Office*. Retrieved from Dubai Media Office: https://www.mediaoffice.ae/en/news/2023/January/16-01/Dubais-annual-real

Fenwick, D. (2013). Hedonic Regression Methods. In D. Fenwick, *Real Estate Prices: methodological frameworks–the international handbooks on Residential Property Price Indices and Commercial Property Price Indices* (pp. 49-64).

Fernandez, M. A. (2019). *A Review of Applications of Hedonic Pricing.* Research and Evaluation Unit (RIMU). Auckland Council.

G, S., Macpherson, D., & Zietz, E. (2005). The Composition of Hedonic Pricing Models. *Journal of Real Estate Literature*, 13(1), 3-43. Retrieved from https://ezproxy.rit.edu/login?url=https://www.proquest.com/scholarly-journals/composition-hedonic-pricing-models/docview/200098809/se-2

Genanew, B. W. (2017). House price drivers in Dubai: nonlinearity and heterogeneity. *International Journal of Housing Markets and Analysis*, 10(3), 384-409. doi:https://doi.org/10.1108/IJHMA-06-2016-0048

Graves, P., Murdoch, J., Thayer, M., & Waldman, D. (1988). The Robustness of Hedonic Price Estimation: Urban Air Quality. *Land Economics*, 220-233.

Herath, S., & Maier, G. (2010). The hedonic price method in real estate and housing market research: a review of the literature. *Institute for Regional Development and Environment*, 1-21.

Hill, R. (2011). Hedonic Price Indexes for Housing. *OECD Statistics Working Papers*. doi:https://doi.org/10.1787/5kghzxpt6g6f-en

Horn, S., & Dasgupta, P. K. (2023). The Air Quality Index (AQI) in historical and analytical perspective a tutorial review. *The International Journal of Pure and Applied Analytical Chemistry*.

Ilvessalo. (1995). A new method for calculation of an air quality index. *Air Quality Department, Finnish Meteorological Institute*, (pp. 527-533).

Mayer, H. (1999). Air pollution in cities. *Atmospheric Environment*, 4029-4037.

Monson, M. (2009). Valuation Using Hedonic Pricing Models. *Cornell Real Estate Review, 7*.

Murdoch, J., & Thayer, M. (1988). Hedonic price estimation of variable urban air quality. *Journal of Environmental Economics and Management, 15(2)*, 143-146.

Neill, H., Hassenzahl, D., & Assane, D. (2007). Estimating the Effect of Air Quality: Spatial versus Traditional Hedonic Price Models. *Southern Economic Journal*, 1088-1111. Retrieved from https://www.jstor.org/stable/20111943

Nguyen, M. (2020). The Hedonic Pricing Model Applied to the Housing Market. *International Journal of Economics and Business Administration, 8(3)*, 416-428.

Nourse, H. O. (1967). The Effect of Air Pollution on House Values. *Land Economics*, 43(2), 181. doi:https://doi.org/10.2307/3145241

Rusmawati, Z., Maharani, R., & Surahman, D. (2020). Determinant Factors of House Price Using Regression. *Proceedings of the 2nd International Conference of Business, Accounting and Economics.* Purwokerto: ICBAE.

Saphores, J.-D., & Wei, L. (2012). Estimating the value of urban green areas: A hedonic pricing analysis of the single family housing market in Los Angeles, CA. *Landscape and Urban Planning, 104*, 373-387. doi:https://doi.org/10.1016/j.landurbplan.2011.11.012

Saptutyningsih, E., & Ma'ruf, A. (2015). Measuring the Impact of Urban Air Pollution: Hedonic Price Analysis and Health Production Function. *Jurnal Ekonomi Pembangunan: Kajian Masalah Ekonomi Dan Pembangunan*, 146-157.

Walravens, N., Breuer, J., & Ballon, P. (2014). Open Data as a Catalyst for the Smart City as a Local Innovation Platform. *COMMUNICATIONS & STRATEGIES*, 15.

Wang, J., & Lee, C. (2022). The value of air quality in housing markets: A comparative study of housing sale and rental markets in China. *Energy Policy*, 160(112601), 112601. doi:https://doi.org/10.1016/j.enpol.2021.112601

Wang, J., Lee, C., & Shirowzahn, S. (2021). Macro-impacts of air quality on property values in China—A meta-regression analysis of the literature. *Buildings, 11(2)*(48). doi:https://doi.org/10.3390/buildings11020048

Waters, M. (2023). *The Essential Guide to the Dubai Real Estate Market.* Routledge.

Yang, R., Zhou, H., & Ding, D. (2018). Air Quality Prediction Method in Urban Residential Area. *11th International Symposium on Computational Intelligence and Design (ISCID). 1*, pp. 16-20. Hangzhou, China: IEEE. doi:10.1109/ISCID.2018.00010

Zeng, B., Fahad, S., Wang, G., Nassani, A., & Binsaeed, R. (2023). Unleashing the Casual Impact of House Prices on Air Quality: Evidence from Chinese Cities. *Indoor Air*. doi:https://doi.org/10.1155/2023/1338261

Zhang, H., Mao, S., & Wang, X. (2021). How Much Are People Willing to Pay for Clean Air? Analyzing Housing Prices in Response to the Smog Free Tower in Xi'an. *International Journal of Environmental Research and Public Health*, 10210.

Zietz, J., Zietz, E., & Sirmans, G. (2008). Determinants of House Prices: A Quantile Regression Approach. *The Journal of Real Estate Finance and Economics, 37*, 317-333.

Zou, Y. (2019). Air Pollution and Housing Prices across Chinese Cities. *Journal of Urban Planning and Development*, 145(4), 04019012. doi:https://doi.org/10.1061/(asce)up.1943-5444.0000517

Appendix

Appendix 1: R Code Used for Models Creation

```
---

title: "DA Thesis Analysis"

author: "Hani Khalaf"

output:

  word_document: default

  pdf_document: default

  html_document: default

---

```{r setup, include=FALSE}

knitr::opts_chunk$set(echo = TRUE)

```

```{r eval=TRUE, echo=FALSE, error=FALSE}

##install.packages

library(ggplot2)

library(dplyr)

library(ggpubr)

library(tidyverse)

library(corrplot)

library(car)

library(glmnet)

library(caret)

```

##Read Dataset
```

````
```{r echo=TRUE}

hdata <- read.csv("Dub_real_trans_AirQ_clean_v4.csv")

head(hdata)

summary(hdata)

```

 # Convert categorical variables to factor type

```{r echo=TRUE}

hdata$area_name <- as.factor(hdata$area_name)

```

# Correlation matrix

```{r echo=TRUE}

correlation_matrix <- cor(hdata[c("trans_year", "unit_area", "numb_rm", "Total_park", "AQI")])

print(correlation_matrix)

```

# Outlier analysis and removal in price

```{r echo=TRUE}

# Boxplot to visualize outliers

ggplot(hdata, aes(y = price)) +

  geom_boxplot() +

  theme_minimal() +

  labs(title = "Boxplot of House Prices", y = "Price")

# Calculate IQR and determine bounds for outliers

Q1 <- quantile(data$price, 0.25)

Q3 <- quantile(data$price, 0.75)
````
````

````
IQR <- Q3 - Q1


lower_bound <- Q1 - 1.5 * IQR
upper_bound <- Q3 + 1.5 * IQR
# Remove outliers
data_clean <- hdata %>%
  filter(price >= lower_bound & price <= upper_bound)
# Boxplot after outliers removal
ggplot(data_clean, aes(y = price)) +
  geom_boxplot() +
  theme_minimal() +
  labs(title = "Boxplot of Property Prices", y = "Price")
```
# Plot a histogram of each attribute
```{r echo=TRUE}
hist(data_clean$price)
hist(data_clean$unit_area)
hist(data_clean$trans_year)
hist(data_clean$numb_rm)
hist(data_clean$Total_park)
hist(data_clean$AQI)
```
# Plot the correlation matrix
```{r echo=TRUE}
````

~~~
# Calculate the correlation matrix

cor_matrix <- cor(data_clean[, sapply(data_clean, is.numeric)])


# Generate the correlation plot

corrplot(cor_matrix, method = "circle", type = "upper",

        tl.col = "black", tl.srt = 45,

        )

```

# Splitting the data into training (80%) and test (20%) sets

```{r echo=TRUE}

set.seed(123) # Setting a seed for reproducibility

splitIndex <- createDataPartition(data_clean$price, p = 0.80, list = FALSE)

data_train <- data[splitIndex,]

data_test <- data[-splitIndex,]

```

## Perform the linear regression analysis OLS

```{r echo=TRUE}

# Building the Linear Regression Model

mlr_model <- lm(price ~ . -unit_area, data = data_train)


# Summary of the model

summary(mlr_model)

# Predicting on the test set

predictions <- predict(mlr_model, newdata = data_test)
~~~

```
# Compute the Mean Squared Error (MSE)

mse <- mean((data_test$price - predictions)^2)

print(paste("Mean Squared Error: ", mse))


# To calculate R-squared for the test set

ss_total <- sum((data_test$price - mean(data_test$price))^2)

ss_residual <- sum((data_test$price - predictions)^2)

r_squared_test <- 1 - (ss_residual / ss_total)

print(paste("Test R-squared: ", r_squared_test))

```

```{r echo=FALSE}

# Remove outliers from the 'price' field

Q1 <- quantile(hdata$price, 0.25)

Q3 <- quantile(hdata$price, 0.75)

IQR <- Q3 - Q1

lower_bound <- Q1 - 1.5 * IQR

upper_bound <- Q3 + 1.5 * IQR

data_filtered <- hdata %>% filter(price >= lower_bound & price <= upper_bound)


# Drop the 'unit_area' column

data_filtered <- select(data_filtered, -unit_area)


# Split the data into training and testing sets
```

```
set.seed(123) # for reproducibility

training_rows <- createDataPartition(data_filtered$price, p = 0.8, list = FALSE)

train_data <- data_filtered[training_rows, ]

test_data <- data_filtered[-training_rows, ]


# Create a dummy model using the full dataset to ensure consistency in dummy variables

dummy_model <- dummyVars("~ .", data = data_filtered)

full_data_transformed <- predict(dummy_model, newdata = data_filtered)


# Convert to dataframe

full_data_df <- data.frame(full_data_transformed)

full_data_df$price <- data_filtered$price


# Split the transformed data back into training and testing sets

train_data_df <- full_data_df[training_rows, ]

test_data_df <- full_data_df[-training_rows, ]


# Perform log-linear regression

model <- lm(log(price) ~ ., data = train_data_df)

summary(model)

# Predict and calculate R-squared for training and test sets

train_pred <- predict(model, newdata = train_data_df)

test_pred <- predict(model, newdata = test_data_df)
```

```
r2_train <- cor(train_data_df$price, exp(train_pred)) ^ 2

r2_test <- cor(test_data_df$price, exp(test_pred)) ^ 2

# Output the R-squared values

print(paste("R-squared for the training set:", r2_train))

print(paste("R-squared for the test set:", r2_test))
```