

Rochester Institute of Technology

## RIT Digital Institutional Repository

---

Theses

---

2-2023

### House Price Prediction Using Machine Learning Model

Abdulla Alfalasi  
ara4293@rit.edu

Follow this and additional works at: <https://repository.rit.edu/theses>

---

#### Recommended Citation

Alfalasi, Abdulla, "House Price Prediction Using Machine Learning Model" (2023). Thesis. Rochester Institute of Technology. Accessed from

This Master's Project is brought to you for free and open access by the RIT Libraries. For more information, please contact [repository@rit.edu](mailto:repository@rit.edu).

# **House Price Prediction Using Machine Learning Model**

By:

**Abdulla Alfalasi**

A Capstone 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**

**RIT Dubai**

**2023 - Feb**

# RIT

## Master of Science in Professional Studies: Data Analytics

Student Name: **Abdulla Alfalasi**

Graduate Capstone Title: **House price prediction**

### Graduate Capstone Committee:

Name: **Dr. Sanjay Modak**  
**Chair of committee**

---

Date:

Name: **Dr. Hammou Messatfa**  
**Member of committee**

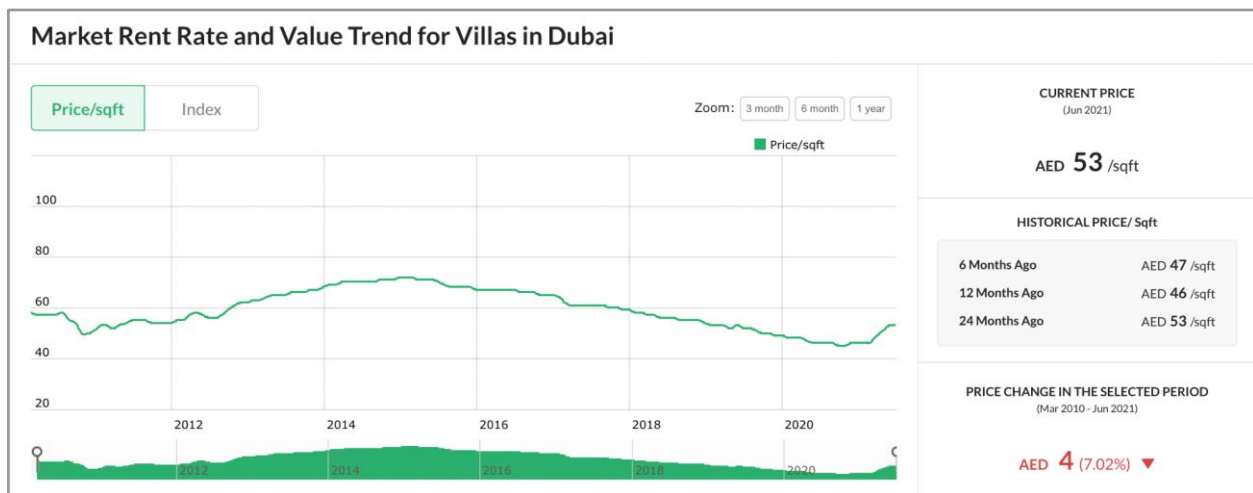
---

Date:

## Abstract

Data Analytics and Machine Learning play an important role in extracting insights and patterns from datasets. The primary objective of using the above techniques for real world problems is to understand the intricacies of the problem, which is impossible with the help of manual human effort.

The below chart shows the price trend (per sq. ft) over the past 10 years in Dubai, with a steady rise during the years from 2014. The 2020 pandemic plummeted the global prices due to the lack of demand and travel restrictions, which is clearly depicted in the chart. But at the very end towards 2022, the price is observed to be increasing steadily which shows the rise in demand due to the opening of travel and ease of restrictions across the globe.(UAE Residential Study, 2022)



In this project, we leverage Data Analytics and Machine Learning to uncover some of the patterns and details about the house price trends in Dubai, and implement a prediction model using different machine learning techniques like Generalized Linear Model, SVM, Neural Networks etc.(Zhang, 2022). to predict the future prices of the properties in Dubai. Different models along with feature combinations would be tested to derive the optimal scores and model results along with different parameters to gauge model performance and results. It has been studied that feature combinations also impact model scores along with the variation of modeling techniques. (Dash, 1997)

**Keywords:** Real estate market, Dubai housing, machine learning, price predictions.

# Table of Contents

<b>Acknowledgments</b>	<b>2</b>
<b>Abstract</b>	<b>3</b>
<b>Chapter 1 - Introduction</b>	<b>5</b>
1.1 Introduction	5
1.2 Project goals	7
1.3 Aims and Objectives	8
1.4 Research Questions	8
<b>Chapter 2 - Literature Review</b>	<b>9</b>
2.1 Literature Review	9
2.2 Takeaways from Literature Review	12
<b>Chapter 3 - Research Methodology</b>	<b>14</b>
3.1 Methodology	14
<b>Chapter 4 - Data Analysis</b>	<b>16</b>
4.1 Data description	16
<b>4.2.1 Data Source</b>	<b>16</b>
<b>4.2.2 Data Audit</b>	<b>16</b>
4.2 Exploratory Data Analysis	18
<b>4.2.1 Data preparation</b>	<b>23</b>
4.3 Machine Learning	27
<b>4.3.1 Model Description</b>	<b>29</b>
<b>4.4 Results and Observations</b>	<b>33</b>
<b>Chapter 5 - Conclusion</b>	<b>36</b>
5.1 Conclusion	36
5.2 Recommendations	36
5.3 Future Work	37
<b>Bibliography</b>	<b>43</b>

## Table of Figures

Fig.1. Comparative analysis YoY residential price per sq ft Dubai

Fig.2. CRISP-DM workflow depiction

Fig.3. Dataset view of top n rows

Fig.4. Descriptive statistics of our dataset view

Fig.5. Histogram of Price variable inclusive of outliers

Fig.6. Summary statistics of Price variable

Fig.7. Price Distribution after removing outlier prices

Fig.8. Relationship of Price and Size of Property

Fig.9. Box plot for price with quality of property

Fig.10. Box plot for price with Neighborhood

Fig.11. Distribution of properties across all neighborhoods

Fig.12. SPSS Mahalanobis Distance Calculation

Fig.13. Table after calculating the Mahalanobis and p-value for outlier detection

Fig.14. Table for outlier and other distribution stats

Fig.15. Mahalanobis scatterplot

Fig.16. Table for model summary

Fig.17. Baseline Modelling workflow using SPSS Modeler

Fig.18. Factor significance from modeling

Fig.19. Scenario 1 model interpretation and workflow

Fig.20. Factor significance using Random Forest

Fig.21. Scenario 2 model interpretation and workflow

Fig.22. Scenario 3 model interpretation and workflow

Fig.23. Scenario 4 model interpretation and workflow

Fig.24. Comparative model study based on parameters for Case 1

Fig.25. Comparative model study based on parameters for Case 2



# Chapter 1 - Introduction

## 1.1 Introduction

The global outbreak of COVID-19 (Deloitte, 2022) has impacted the entire world and UAE's real estate sector happens to be one of those to be affected. In UAE, the economy has experienced a lot of uncertainties and the real estate sector has been the most affected. Dubai's expo for the year 2020 was shifted to 2021 to boost the economic condition of the country and support the growth of the real estate sector. Many real estate groups had introduced relief packages such as rent reliefs, waiving administrative charges, deferred payment plans etc. Both public and private sector developers are trying to execute the projects at a faster pace now and aim to deliver the best quality projects like before. Some of the indicators showcasing the slight boom of the housing sector include new visa initiatives, nationality for selected expatriates, virtual real estate events, vaccine roll out, changes in existing company laws, attractive offerings by lenders and many other initiatives. The government has rolled out many visa initiatives to create a positive impact on the housing scenario in UAE such as ten year residency for expatriates, five year renewable retirement visas to retired residents and new self-sponsored remote work visa allowing employees from all over the world to live and work remotely from the UAE. Recently the UAE government has also amended its citizenship policy allowing investors, professionals and their families to acquire Emirati nationality under certain conditions. This has been done to further expand and boost the demand for housing. Cityscape Intelligence, another initiative that allows investors and buyers to experience the housing investments in the form of physical events, live and in-person. Not only has these, there are other initiatives of which the companies based out of UAE who had to have more than one Emirati shareholder with minimum of 51% share in the company, the requirement mandate has been removed. This policy aims to boost the economy and growth by attracting foreign investors. Apart from these, there have been initiatives around the increase of loan to value ratio, mortgage gap as well as renter conversion to homeowners, along with the interest rates being very attractive for people. These rules promote sustainable development and innovation in building design. It also aims to decrease construction costs by restructuring building rules and requirements. Reducing fuel surcharge for electricity and water is another such initiative. Also, a new law governing the rental agreements in Dubai is likely to be issued in order to fix rents. This will help the landlords to achieve stable tenants and avoid paying an annual increase in the rental amount.

The Dubai residential real estate sector was very competitive, until COVID-19 struck across the entire globe. With the ease of travel restrictions and rise in travel activities, the demands for these properties are set to increase again in the coming months. According to a report by Asteco, UAE is set to see an increase in 38,500 apartments along with the supply of 3,800 villas and Dubai is estimated to account for a large chunk of it with 30,000 flats and 3,500 villas in 2022. The month of January 2022, saw over 53% transactions which were for complete properties along with 47% for off-plan properties. The off-plan market accounted for 2,706 properties while the ready market accounted for 3,091 transactions. A survey forecasted the house prices in Dubai to rise 3.0% in 2021 and 2.5% in 2022 compared with 1.1% and 2.8% compared to three months ago. In August 2021, a sum of AED 4.95 billion was logged by off-plan category of properties, which recorded the highest in the sales for the particular category in December 2013. There were a sum of 187 units that were sold in Arabian Ranches 3 and 157 in Villanova. Of all the locations, there were some more locations including cities like Tilal al Ghaf having above 79 units, Dubai South having 58 units and Mohammed bin Rashid City with more than 16 units. Dubai Marina, Downtown Dubai, Palm Jumeirah, Business Bay and Jumeirah Village Circle are some of the locations with the highest demand for off-plan property types. The recorded average in the same category of properties witnessed a spike of 53% from a sum of AED 1.2 million in August 2020 to a sum total of AED 1.9 million in August 2021.

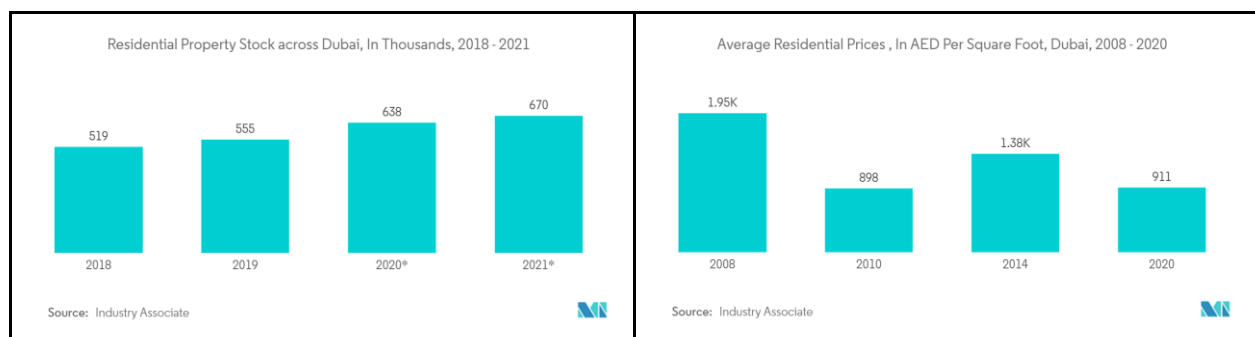


Fig.1. Comparative analysis YoY residential price per sq ft Dubai

With the advent of machine learning and advanced data analytics techniques, the implementation of these tools allow larger scope for experimentation to reduce the gap between real estate and technical know-how. We would be able to use advanced data analysis techniques along with Machine Learning to be able to comprehend price trends and various other details in order to provide a better scope of recommendation to our viewers and customers in the housing market.

## 1.2 Project goals

With the increasing demand of houses in the Dubai market, it is essential that investors understand the current trends and requirements through data. This project will help investors (both long term as well as short term) to understand price trends in different neighborhoods and regions, along with different factors affecting housing prices in the city.

Now, the scope of our project is to understand the different trends of the housing market in UAE through data analysis and be able to define price trends for properties with different attributes like number of rooms, size of the apartment, number of bathrooms etc. This helps us understand the real-estate market with respect to apartment pricing in depth so as to be able to recommend better options to prospects looking for investment into the apartment space. Once we have the insights generated from data analysis techniques, we then clean the dataset and perform the necessary data manipulation techniques to have clean data. Clean data is always important before we plan to train any machine learning model. This ensures that the models return accurate results with respect to predictions or significant factor determination. In our project, we plan to use Linear Regression to predict the house prices based on different traits. So, in future if someone enters the details like number of rooms, property size, number of bathrooms etc. to our model we would be able to predict the price of the apartment for them. This would have a lot of implementation prospects in the future because of the application perspectives that it provides to the users.

Once we have the model deployed in our code base, there is a future scope of predicting house prices with even more accuracy if we have better data points. Not only this, there is also scope of predicting ahead of the time, i.e, the look ahead duration can be increased to several months if the historical dataset is increased and the model is able to train better on a larger dataset.

## 1.3 Aims and Objectives

The objective of this experiment and study is to understand the house price patterns in Dubai, and also interpret the housing market overall in Dubai. This will help us understand the price trends as well as the factors depending on the price of properties. Throughout the course of the study, we will use one of the popular frameworks called CRISP-DM to solve the problem statement and then identify the key drivers of price points for different properties in different neighborhoods of Dubai. Also, we will study different predictive models and their capabilities with the dataset and pick the best fit model based on its prediction capability. We will use Machine Learning to be able to determine the future prices of the properties in Dubai. This will help real

estate companies determine price trends, both historical and future and be able to make necessary investments for the same.

#### 1.4 Research Questions

During the course of our study, we will be leveraging data to understand the different factors influencing the prices of the properties. Not only this, we would be able to answer some of the questions within this problem statement like -

- Which neighborhoods have the most expensive apartments and which have the cheapest properties?
- What factors do these property prices depend on; size of apartment, number of bathrooms or bedrooms?

Some of these questions are the most frequently asked questions and helping our readers answer the same would help us understand these intricacies in this problem statement.

# Chapter 2 - Literature Review

## 2.1 Literature Review

As per the study conducted by Deloitte (2021), the average sales prices for housing property in Dubai deteriorated by 7% between Q3 of 2019 and Q3 of 2020. There was also a decline in the average amount of rents by 10% in the same period. While the average price per sq ft for apartments reduced from AED 1,090 in 2019 to AED 1,011 as of September 2020. A lot of tenants decided to travel to bigger units with superior amenities that are now more affordable. But even after the decline in prices, project handovers in Dubai continue in Expo 2020. Based on discussions with investors, it is estimated that a total of 24,000 to 25,000 residential units have been handed over during the first nine months of 2020. The amount of transaction volume reduced by 16% YTD September 2020 as compared to the same period in 2019. The demand for secondary market properties outperformed transaction volumes for off-plan units. Developers now are offering discounts, fee waivers and many other incentives to entice the buyers.

As per a survey conducted by Abdulla (2020), the demand was higher for apartments and then followed by villas and number of bedrooms and bathroom requirements was 2 followed by 3. The size of the units lie between 1000-2000 square feet and also, the prices of these housings stands below 2,500,000 AED. This information can be used by investors and also constructors to determine the purchasing trend as well as price points at different times for the building. The minimum price of a unit listed was determined to be 225,000 AED and the average price of the property was evaluated approximately 2,670,000 AED. There is a difference between the price of a particular unit in different locations. The price of penthouses in Dubai Sports City and a townhouse in Palm Jumeirah were found to be almost the same, while the sizes of these two property types differed in both these neighborhoods. The villas in Jumeirah Village Circle turn out to be cheaper than villas in Arabian Ranches and Damac Hills, while they are larger in size. The units that are most expensive are penthouses in Downtown Dubai, whereas, the cheapest are apartments in Dubai Silicon Oasis and Dubai Sports City. With the kind of access to these insights in the housing price trends with relation to the property attributes, Abdulla was able to achieve the required recommendations and research the price trends in Dubai.

As per research conducted by Knight Frank (2021), there exists an excess level of supply and weaker level of demand when it comes to the housing sector in the UAE. Abu Dhabi and Dubai's

housing market are expected to record historic levels in 2021 of new supply of circa 14,000 and 83,000 units respectively. In Dubai, the expectation is that the market will perform across prime and non-prime neighborhoods, wherein the prime neighborhoods segment is expected to outperform the latter. The existing vacancy in Dubai has increased by 1.8% points over the course of 2020 to 18.3%. While in Abu Dhabi, the challenging economic conditions are probable to continue with the current rate of decline throughout the year of 2021. Keeping a long term view on the housing market, there is an expectation of new supply levels to begin from 2022 in Abu Dhabi and from the late 2023 in Dubai. The number of housing launches in Abu Dhabi are expected to rise over the coming years while in Dubai, it is expected to remain below the average as seen over recent years. Assuming that the trend continues, mortgage rates are expected to remain at or around historic lows and loan-to-value ratios at current levels, which are likely to bottom out the prices during 2022.

According to a study conducted by Farhad and group (2020) wherein they tried to establish a relationship between variables such as house price index with exchange rate, inflation rate and supply. The study revealed that in the case of consumer price index and house price index, there exists a direct relationship which means that if consumer price index increases, house price index will rise too. There exists a negative relationship between the exchange rate and housing price index of Dubai. It means that if the exchange rate is high, the housing price index will be lower and vice versa. Money supply is directly related to house price index in Dubai and consumer pricing index will lead to a fall in the housing sector prices. The real estate market analysis reveals that there are several indexes that help measure the efficiency. Through the research done by Farhar, the recommendations came as strong for the following points. Policymakers must increase the number of units and money supply in order to raise the investments by decreasing the interest rate. There is a major portion of investment in UAE's housing sector and this can further be developed by boosting the tourism in UAE to attract more investors. Dubai has developed its housing industry which consists of indexes such as house price index, that helps to measure effectiveness and efficiency of the industry.

In a paper by Quang T. (2020), we understand that one of the most commonly used measures to calculate or estimate the changing housing prices is the House Price Index or HPI. This measure is also used with other extrinsic factors to predict house prices as the HPI is highly correlated with other factors like location of the property, area of the house, population in the location etc. Since a lot of the research works ignore traditional machine learning models during their prediction cycles, Quang dedicated this paper to using both the advanced and traditional Machine Learning

models because according to him the traditional models are popular yet complex but are equally contending modeling techniques. The paper implements various techniques for modeling with regression and determining the pragmatic prices for houses. Some of the different models used in this paper are Random Forest, XGBoost, LightGBM and two techniques in machine learning like Hybrid Regression and Stacked Generalisation Regression for the prediction purposes and a comparative study of all the models. It was observed that Hybrid Regression and Stacked Generalisation Regression provide compelling results, but the time complexity has to be taken into consideration since both the models involve the use of a very complicated modeling technique called Random Forest which is computationally expensive. Hence, the conclusion was that XGBoost and LightGBM provide satisfactory results and their time complexity is not very high, which makes them satisfactory models for performance evaluation as their accuracies were determined to be good enough. It can be rightly said that not only the accuracy is important for a model, but different aspects come into picture like speed of model fitting, extrinsic factor consideration, coupling effect of different models based on regression techniques etc.

The prediction approaches for housing prices have been mostly carried out on big city datasets, while using the traditional approach with linear or spatial linear models. To try a different approach in their paper, Wang L. (2020) uses a dataset based on the prices and factors picked from middle to small tier cities. The prediction approaches have been carried out with a flexible spatiotemporal model (FSTM), which implemented both the spatial and temporal components of the housing price components in the research carried out. It was observed that external factors like government policies had a huge impact on the prices of the properties, and this led to the characteristics of the small-medium cities being very different from the big cities. Some of the other factors include road density, bank density, supermarket density and many others, which are different from the factors that are usually considered in the traditional house price prediction processes. This study has also demonstrated the potential for FSTM in spatiotemporal prediction approaches for residential house prices which are located in the middle-small cities, contrary to the big city predictions usually carried out.

There are numerous papers which implement various techniques to predict the prices of properties by using model optimization techniques or using feature selection processes. Manjula R. (2017) demonstrated there are different sets of features which are helpful in predicting the house prices with accuracy. Her team determined the usage of many techniques based on regression to optimize on the error score (RSS). Feature selection methods are very important in improving the scores of Regression models. A better model fit is obtained using various research

techniques wherein sets of different features or polynomial regression techniques are implemented to obtain the same as end result. But this results in sometimes the model results being over-fitted as a result of which it is better to use ridge regression techniques.

In Turkey, the country market for houses was picked to predict and research on the price trends. Many businesses and countries determine Turkey as a profitable market due to which they tend to invest in the same. Hacievliyagil N. (2021) performed work using a dynamic model averaging (DMA) which was able to predict the monthly house price growth. Of all the features, the residential property price index or the RPPI was determined to be the most significant of all the features in the modeling process. There were several other features along with this like unemployment, exchange rate as well as mortgages. It has been recommended that house prices should be tracked very closely because it contributes as a macroeconomic factor for a country ultimately.

One other core aspect to predicting with accuracy is the amount of historical dataset available to us during the research purpose. In the paper by Chen (2017), they consider the price data from January 2004 to October 2016 which is then used to predict the prices between November till December for the year 2016 for some of the districts in China like Beijing, Shanghai, Guangzhou and Shenzhen. During the experiment they used an autoregressive integrated moving average model as a baseline and then the LSTM network was implemented to build the model to predict the values. The correction measures used are the MSE or Mean Squared Error value, and concluded that the LSTM has very good predictive capabilities to predict the time series.

## 2.2 Takeaways from Literature Review

The idea behind a literature survey is to have an in-depth understanding of the subject matter or problem space. It also helps us understand the trends that have been ongoing in the housing market research space which would allow us to get a better grasp of moving ahead with our project.

- A global disruption like the COVID-19 brought about a lot of trend changes in the housing market from the time it started till date. This unforeseen circumstances has necessitated the use of advanced techniques and tools to better prepare for further trends in the future
- The housing market in Dubai is set to increase steadily as there has been a steady supply of properties, with the demand being nominal. The supply to demand ratio is high which



is leading to higher supply than demand. This needs to be fulfilled at a larger scale and advanced analytics and insights can help identify the gaps

- A lot of property attributes contribute to the prices of the property being high or low, and this is explained in some researches well about the correlation between different factors which might affect the house prices accordingly
- Even though the modern world is inclined towards using complex Machine Learning techniques, the traditional models are equally compelling for such tasks. It was observed that the criteria for computation complexity and times is also very important when considering Machine Learning models, instead of just relying on accuracy because it can cost expensive at a later stage
- House prices are impacted by many factors like exchange rate, unemployment, inflation and many others (Wang, 2019). And the same applies the other way round, because the housing industry also plays a role in the macro-economic factors of the country thereby impacting the ecosystem due to unforeseen circumstances

With the above summarization of our research on different papers and resources, it is understood that house price prediction plays an important function in the modern world. Through this project, we plan to implement some of the most complex techniques to understand the housing market better and be able to predict accurately using different techniques.

# Chapter 3 - Research Methodology

## 3.1 Methodology

In the provided Dubai house price dataset from Kaggle (Data Regress, 2020), we plan to perform the following steps to be able to obtain a comprehensive understanding of the data before moving ahead with any machine learning model. A data analytics/science project comprises the CRISP-DM (Data Solut, 2022) life cycle of working ahead with the project. The steps followed are shown below -

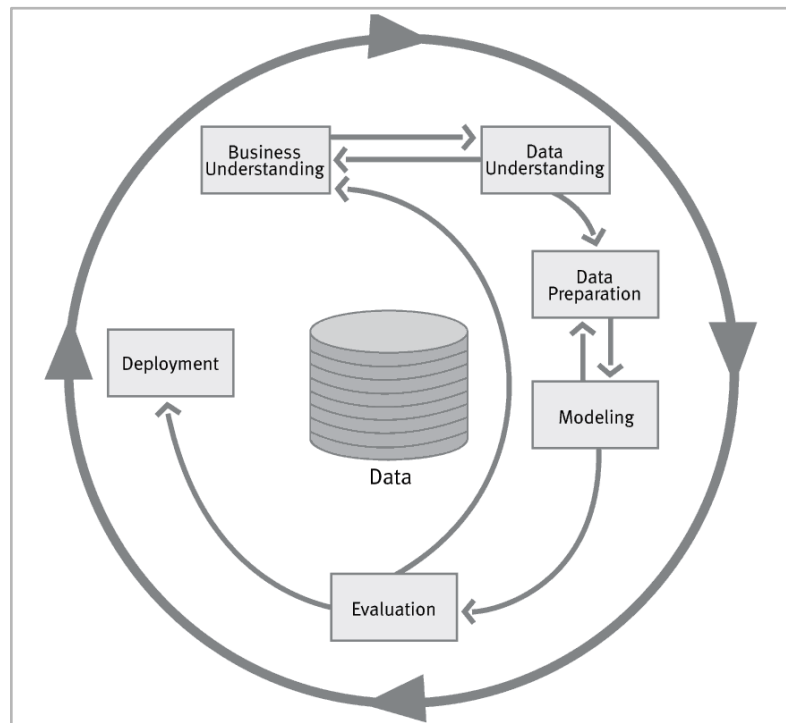


Fig.2. CRISP-DM workflow depiction

### **Business Understanding**

We have to start by understanding our industry and reading about use cases in the housing industry, so that we get a good knowledge about the industry and the housing market in Dubai

### **Data Understanding**

In this step, the dataset is obtained and used to perform exploratory data analysis and different statistical models to understand in-depth about the data. We would start with exploring or getting

an overview of the dataset like the mean prices, standard deviation of the predictor and other continuous features

### **Data Preparation**

In this step of the process, the dataset has to be cleaned of inconsistencies and missing values after an overview exploration for missing values and data types for all features. This helps us perform the analyses better, for example the price column might have some outlier values due to typing errors and this might skew the dataset. Hence we need to clean the data off these problems

### **Data Modelling**

Data Modelling is the process of using the clean and prepared dataset to make a machine learning model, train the data and be able to predict the outcome values in the final step. Once we understand the features in our dataset, we can move ahead to prepare our train and test set splitting (mostly 80-20 for train to test set). Having that split allows us to train the dataset as well as test the dataset for prediction values and understand the accuracies

### **Evaluation**

Now, this is the final step where we would need to validate the constructed model and check if the accuracy and the predicted values are good from the model or not. If there are deviations in the predicted values, what is the deviation? These checks enable us to understand the performance of our model and have a better view of the final results before we convey it to our stakeholders

Following the above mentioned steps from the CRISP-DM lifecycle enables a structured flow to solving the entire problem and have accurate results for our stakeholders as well as recommendations. Hence, using such a process ensures a good project process and we plan to use the same to achieve our results and final output.

### **Tools and Technologies**

This project would require some of the advanced tools and technologies in terms of Data Analysis and visualizations. Below are some of the details about the tools -

SPSS Statistics - This tool will be used to manipulate and shape the dataset. Not only that, we would also be using the tool to perform different statistical tasks like outlier detection, statistical hypothesis testing, data cleaning etc.

SPSS Modeller - This tool will be used to create modeling pipelines with the cleaned dataset. It can be used to clean, shape and split the dataset for Machine Learning model implementation.

# Chapter 4 - Data Analysis

## 4.1 Data description

### 4.2.1 Data Source

The dataset is available in an open platform called Kaggle, (Kaggle, 2020) which hosts problem statements and datasets around Data Science for different topics and industries. The below snapshots for the dataset have been obtained using Python functions, to depict an example of how the dataset looks like for our use case. There are 20 different columns for each property ID, which depicts the different attributes of the property like neighborhood, latitude, longitude, price, size in square feet, price per square feet etc.

Annotations																			
id	neighborhood	latitude	longitude	price	size_in_sqft	price_per_sqft	no_of_bedrooms	no_of_bathrooms	quality	maid_room	unfurnished	balcony	barbecue_area	built_in_wardrobes	central_ac	childrens_play_area	childrens_pool	concierge	covered_parking
7586332....	Palm Jumeirah	25.104		55.16934314000....	9576.000	3583.330	4.000	5.000 Medi...	False	\$null\$ True	False	True	False	True	False	False	True	True	True
7702221....	Palm Jumeirah	25.104		55.15034340000....	8722.000	3937.170	4.000	6.000 High	True	\$null\$ True	False	True	False	True	False	False	True	True	True
7433552....	Business Bay	25.181		55.26575000000....	7344.000	1021.240	3.000	5.000 Medi...	True	\$null\$ True	False	True	True	False	False	False	False	True	True
7545873....	Business Bay	25.188		55.28930950000....	7922.000	3906.840	4.000	4.000 Medi...	False	\$null\$ False	False	True	True	False	False	False	False	True	True
7636368....	Palm Jumeirah	25.104		55.16935000000....	7346.000	4764.500	4.000	5.000 Low	False	\$null\$ True	False	True	False	False	False	False	True	True	True
7703691....	Palm Jumeirah	25.104		55.15031440000....	6542.000	4805.870	4.000	6.000 High	True	\$null\$ True	False	True	False	True	False	False	True	True	True
7677586....	Jumeirah	25.215		55.23627500000....	6735.000	4083.150	4.000	5.000 Medi...	False	\$null\$ True	False	True	True	False	False	False	False	False	False
7682462....	Palm Jumeirah	25.110		55.11117798550....	6160.000	2889.380	3.000	5.000 Medi...	True	\$null\$ True	True	True	True	True	True	False	False	False	False
7626662....	Business Bay	25.189		55.28916950000....	4907.000	3454.250	3.000	2.000 High	True	\$null\$ True	False	True	True	True	False	False	True	True	True
7557465....	Palm Jumeirah	25.104		55.16921990000....	6252.000	3517.270	4.000	5.000 Medi...	False	\$null\$ True	False	True	True	True	False	False	True	True	True

Annotations																			
	kitchen_appliances	lobby_in_building	maid_service	networked	pets_allowed	private_garden	private_gym	private_jacuzzi	private_pool	security	shared_gym	shared_pool	shared_spa	study	vastu_compliant	view_of_landmark	view_of_water	walk_in_closet	
1	True	False	False	False	False	False	False	False	True	False	False	False	False	False	False	False	False	False	False
2	True	False	True	False	True	True	False	True	True	True	True	True	True	False	False	True	True	True	True
3	True	False	False	False	True	False	False	False	False	True	True	True	True	False	False	True	True	True	True
4	False	False	False	False	False	False	False	False	False	False	True	True	False	False	False	True	False	False	False
5	False	False	False	False	False	False	False	False	True	False	False	False	False	False	False	False	False	False	False
6	True	False	True	False	True	False	False	True	True	True	True	True	True	False	False	True	True	True	True
7	False	False	False	False	False	False	False	False	False	False	True	True	False	False	False	True	True	True	True
8	True	False	False	False	False	False	False	True	True	False	True	False	False	False	False	False	True	False	False
9	True	False	False	False	False	False	False	False	True	True	True	True	True	True	False	False	False	True	True
10	False	False	False	False	False	False	False	False	True	False	False	False	False	False	False	False	False	False	False

Fig.3. Dataset view of top n rows

### 4.2.2 Data Audit

SPSS Statistics would be used to perform exploratory data analysis for univariate and bivariate analysis to interpret the relationship between the variables in the dataset. The SPSS software package was created for complex statistical analyses and research purposes in 1968 by SPSS Inc., which was eventually acquired by IBM. This powerful tool is used by marketing organizations, data miners, government bodies and many more for data processing and analyzing survey data which are created by various platforms.

Types   Format   Annotations					
Read Values   Clear Values   Clear All Values					
Field	Measurement	Values	Missing	Check	Role
id	Continuous	[5528049.0, 77066...		None	Input
neighborhood	Nominal	"Al Barar", "Al Barsh...		None	Input
latitude	Continuous	[24.865992, 25.27...		None	Input
longitude	Continuous	[55.069311, 55.44...		None	Input
price	Continuous	[220000.0, 3.5E7]		None	Target
size_in_sqft	Continuous	[294.0, 9576.0]		None	Input
price_per_sqft	Continuous	[361.87, 4805.87]		None	Input
no_of_bedrooms	Nominal	0.0, 1.0, 2.0, 3.0, 4.0...		None	Input
no_of_bathrooms	Nominal	1.0, 2.0, 3.0, 4.0, 5.0...		None	Input
quality	Nominal	High, Low, Medium, UL...		None	Input
maid_room	Flag	True/False		None	Input
unfurnished	Typeless			None	None
balcony	Flag	True/False		None	Input
barbecue_area	Flag	True/False		None	Input
built_in_wardrobes	Flag	True/False		None	Input
central_ac	Flag	True/False		None	Input
childrens_play_area	Flag	True/False		None	Input
childrens_pool	Flag	True/False		None	Input
concierge	Flag	True/False		None	Input
covered_parking	Flag	True/False		None	Input
kitchen_appliances	Flag	True/False		None	Input
lobby_in_building	Flag	True/False		None	Input
maid_service	Flag	True/False		None	Input
networked	Flag	True/False		None	Input
pets_allowed	Flag	True/False		None	Input
private_garden	Flag	True/False		None	Input
private_gym	Flag	True/False		None	Input
private_jacuzzi	Flag	True/False		None	Input
private_pool	Flag	True/False		None	Input
security	Flag	True/False		None	Input
shared_gym	Flag	True/False		None	Input
shared_pool	Flag	True/False		None	Input
shared_spa	Flag	True/False		None	Input
study	Flag	True/False		None	Input
vastu_compliant	Flag	True/False		None	Input
view_of_landmark	Flag	True/False		None	Input
view_of_water	Flag	True/False		None	Input
walk_in_closet	Flag	True/False		None	Input

Univariate Statistics							
	N	Mean	Std. Deviation	Missing		No. of Extremes <sup>a</sup>	
				Count	Percent	Low	High
id	1905	7573308.19	192525.172	0	.0	173	0
latitude	1905	25.1165381	.062647107	0	.0	1	0
longitude	1905	55.2123383	.068794406	0	.0	0	0
price	1905	2085829.87	2913199.962	0	.0	0	128
size_in_sqft	1905	1417.05	891.488	0	.0	0	88
price_per_sqft	1905	1327.2438	668.47356	0	.0	0	82
no_of_bedrooms	1905	1.79	.949	0	.0	0	75
no_of_bathrooms	1905	2.51	1.063	0	.0	0	88
unfurnished	0	.	.	1905	100.0	0	0
quality	1905			0	.0		
maid_room	1905			0	.0		
balcony	1905			0	.0		
barbecue_area	1905			0	.0		
built_in_wardrobes	1905			0	.0		
central_ac	1905			0	.0		
childrens_play_area	1905			0	.0		
childrens_pool	1905			0	.0		
concierge	1905			0	.0		
covered_parking	1905			0	.0		
kitchen_appliances	1905			0	.0		
lobby_in_building	1905			0	.0		
maid_service	1905			0	.0		
networked	1905			0	.0		
pets_allowed	1905			0	.0		
private_garden	1905			0	.0		
private_gym	1905			0	.0		
private_jacuzzi	1905			0	.0		
private_pool	1905			0	.0		
security	1905			0	.0		
shared_gym	1905			0	.0		
shared_pool	1905			0	.0		
shared_spa	1905			0	.0		
study	1905			0	.0		
vastu_compliant	1905			0	.0		
view_of_landmark	1905			0	.0		
view_of_water	1905			0	.0		
walk_in_closet	1905			0	.0		

a. Number of cases outside the range (Q1 - 1.5\*IQR, Q3 + 1.5\*IQR).

Fig.4. Descriptive statistics of our dataset view

Through the above view of the dataset, we see that there are many properties/factors defining a property in depth like the size of the apartment, the price per sqft, if gyms, private gardens etc. and many other amenities are available in the apartment. Now, having viewed the data we can move ahead with the exploratory data analysis to understand the data better. In the subsequent steps, we will shape and prepare the dataset better for easier interpretability.

In the above shown step, we are mapping the appropriate feature to the target feature in SPSS Modeler, and also type cast some of the features for easier interpretability. For example, we define the ordinal features as either True or False and then define the numerical values accordingly, for the tool to understand the dataset better.

## 4.2 Exploratory Data Analysis

This would help us understand the distribution of the target values along with the other features. We will use some of the most popular forms of data visualizations through pie charts, box plots, histograms and bar charts to interpret the distribution of the dataset.

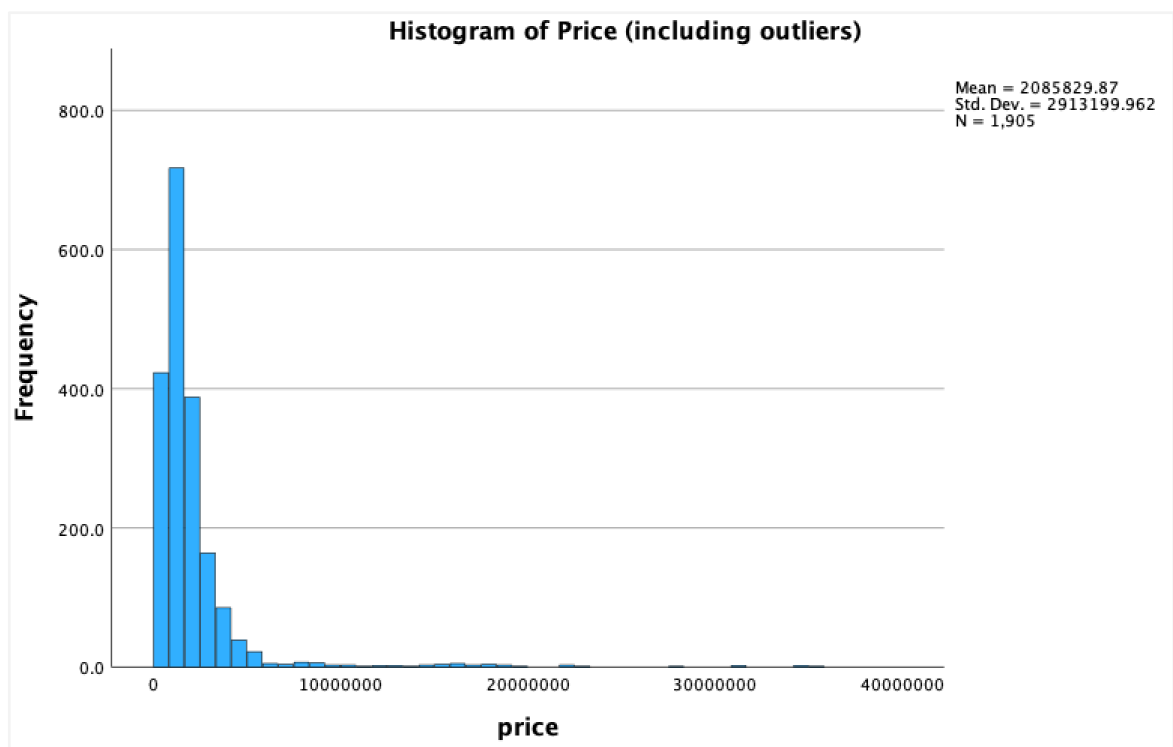


Fig.5. Histogram of Price variable inclusive of outliers

For starters, we plot the histogram of the predictor variable to understand the distribution. It seems like we have a lot of outlier prices in the dataset which are above 10000000 AED, and this skews the entire analysis as there only 5.5% of the dataset which is above 4500000 AED (which is the threshold for the outlier values)

Descriptives			
		Statistic	Std. Error
price	Mean	2085829.87	66745.625
	95% Confidence Interval for Mean	Lower Bound	1954927.64
		Upper Bound	2216732.11
	5% Trimmed Mean	1642895.05	
	Median	1400000.00	
	Variance	8.487E+12	
	Std. Deviation	2913199.962	
	Minimum	220000	
	Maximum	35000000	
	Range	34780000	
	Interquartile Range	1310000	
	Skewness	6.147	.056
	Kurtosis	48.857	.112

Fig.6. Summary statistics of Price variable

The summary table on the top shows the distribution of the dataset. The Kurtosis value should be between -2 and +2, and our variable has 48.8 which is very very high. This means that the data points towards the tails are very heavy. Hence to have a standardized representation of the data, we will remove these 5.5% outlier values and then make the necessary plots.

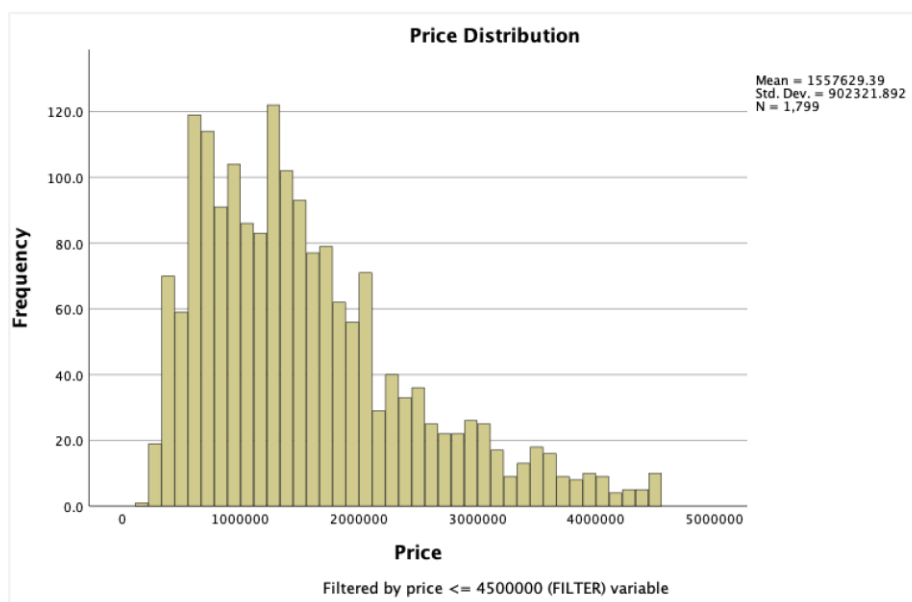




Fig.7. Price Distribution after removing outlier prices

In Fig.3. we can see a clear distribution of the dataset having removed the outlier prices (i.e., anything above 4500000). This also helps us understand that most of the house prices are distributed across 1000000 and 2000000. We see that the average prices are reduced by 33% after removing the outliers.

When we plot the distribution of the prices with that of the size of the properties, as shown in the scatterplot in Fig.4. we see that there is a linear relationship between both the variables. As the size of the properties increase, the price gradually increases for them as well. This helps us understand the features that can be used during the modeling phase, to understand the linear and non-linear dependency of each of them.



Fig.8. Relationship of Price and Size of Property

In the following boxplot, we want to understand the trend of prices with the quality of the properties. It is seen that the Low and Medium quality properties have a higher average price range. While the Ultra quality properties have the lowest average.

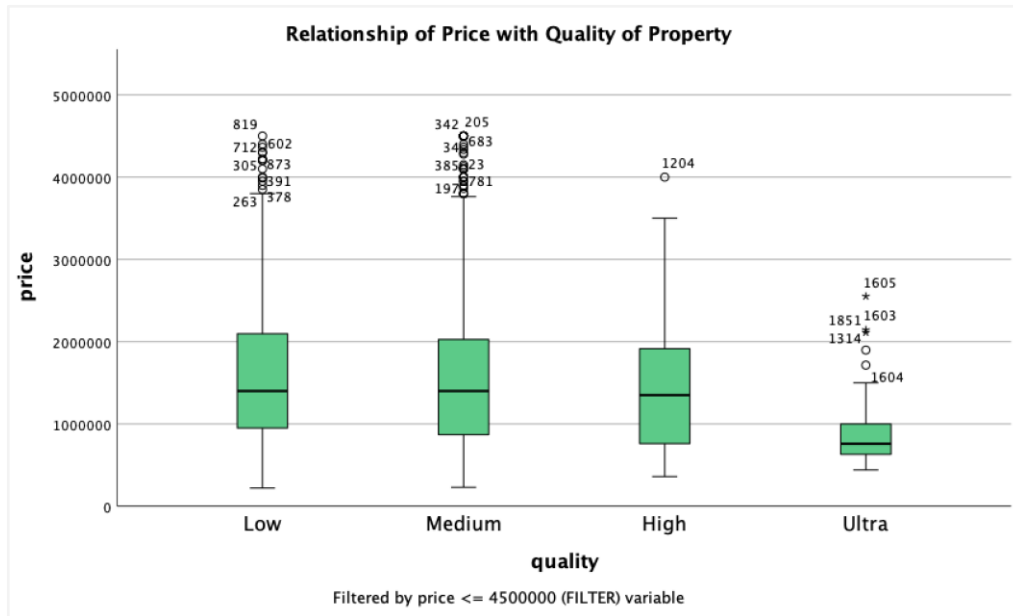


Fig.9. Box plot for price with quality of property

Again, in the chart below we have the price distribution for properties for different neighborhoods. It is seen that 1,557,629 being the average price Bluewaters, Culture Village and The Hills have higher than average prices, highest average prices of all neighborhoods. Hence, this helps us understand and differentiate the cheap from the expensive neighborhoods.

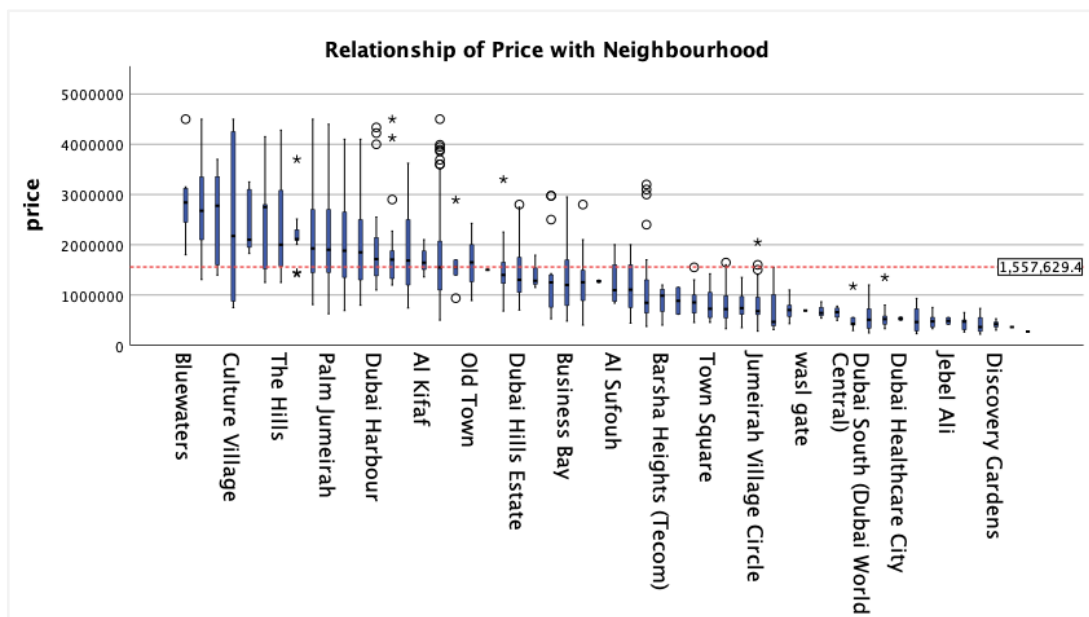


Fig.10. Box plot for price with Neighborhood

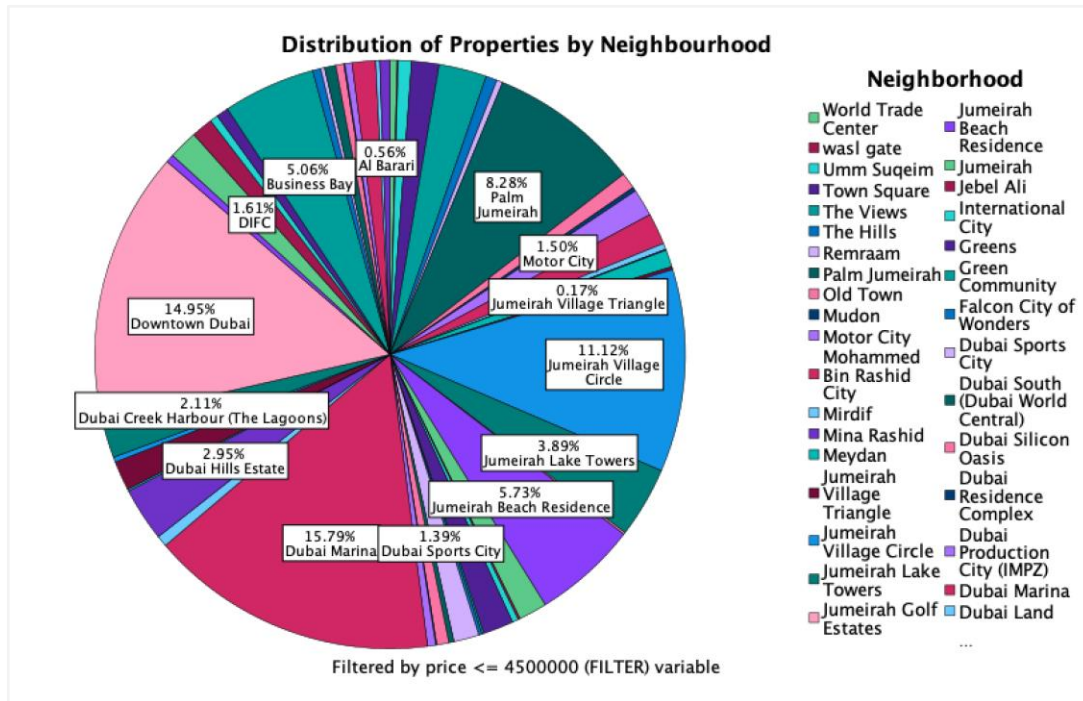


Fig.11. Distribution of properties across all neighborhoods

Most of the properties are available in Dubai Marina and Downtown Dubai, constituting 15.8% and 14.95% respectively of all the properties available in Dubai. This helps us understand where most of the property information is coming from, how they are distributed across the city.

Through the above analysis, we come to understand a lot about some of the factors in the dataset like the average prices, the outlier prices and cleaning them, the expensive vs the cheap neighborhoods for property prices and many others. This helps us prepare for the modeling of the dataset along with building the right hypothesis while doing different types of tests.

#### 4.2.1 Data preparation

Having performed data cleaning and filtering steps (using outlier removal process) along with removing high correlation features like size of the property and price per square feet, we have a dataset that looks like below. It is imperative to remove features that explain the target variable too well, and in this case the size of property and the price per square feet are similar to the target feature. Hence, we remove the same to ultimately derive the below observations.

Through the above processes, we have a standardized dataset which is free from outlier values and any sort of biases. This helps us perform an analysis without extreme results and output and in turn have a normalized understanding of the problem statement.

Mahalanobis Distance (Prabhakaran, 2019)

Mahalanobis distance is the process of determining the distance between two points in a multivariate space. For simple approaches like Euclidean or Manhattan distances, the distance between two points can be marked and the distance can be interpreted accordingly. Now, in the case of the MD, uncorrelated variables and their distances can be computed using the Mahalanobis Distance. This distance metric can be used to measure the distance between two distinct points even if the points are correlated for multi variable problem space.

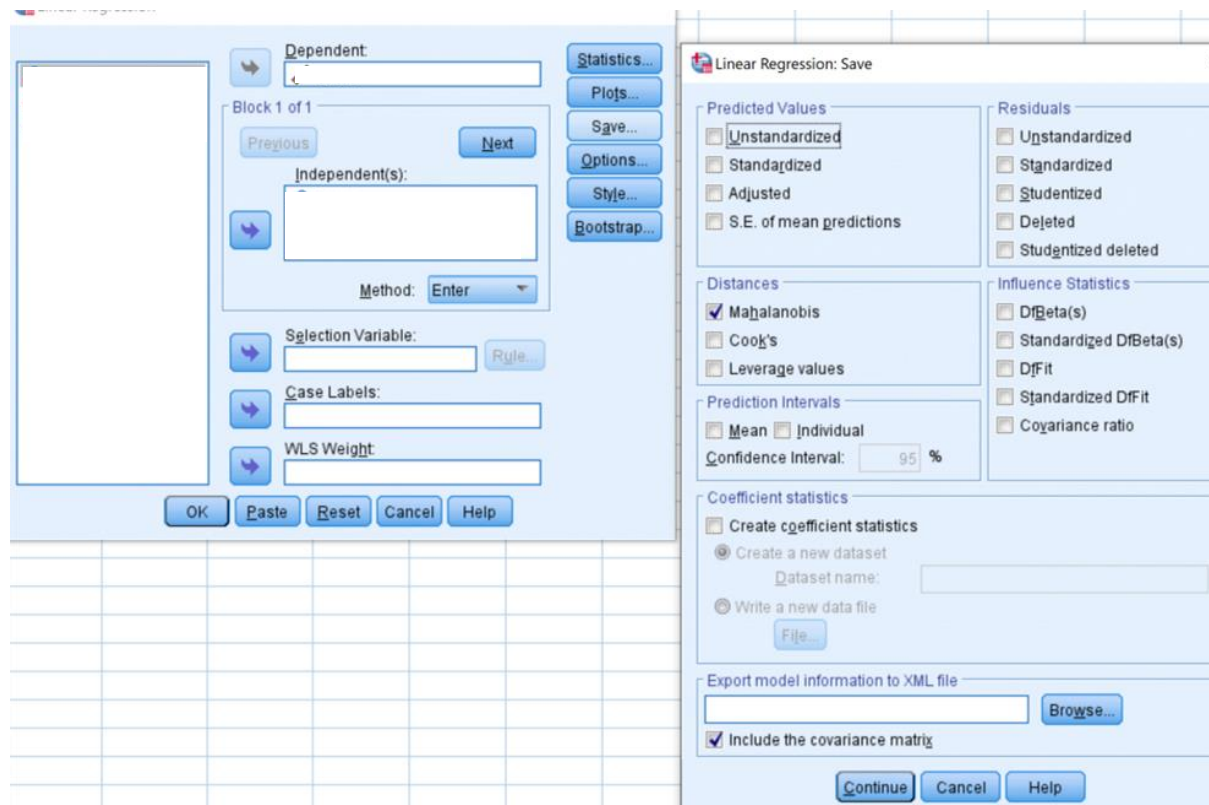


Fig.12. SPSS Mahalanobis Distance Calculation

In the screenshot of SPSS Statistics shown above, the Mahalanobis distance can be derived using the above steps. The dependent and independent variables need to be chosen in order to derive the MAH score (which has to be chosen in the distances tab). After this we derive different MAH scores and we also observe that one score is higher or lower than the other. To determine the

statistical significance of these differences, p value can be computed using the chi square formula (shown below).

MAH_1	PVALUE	OUTLIER
6.92220	.22649	0
2.67962	.74923	0
10.62368	.05937	0
1.61850	.89900	0
29.11152	.00002	1
13.49115	.01919	0
19.08853	.00185	0
13.49115	.01919	0

Fig.13. Table after calculating the Mahalanobis and p-value for outlier detection

From the above table it is interpreted that any p-value that has value less than 0.001, that is considered to be an outlier and can be removed in the next step.

Below is the table that shows the distribution of the outliers.

OUTLIER					
		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	0	1846	96.9	96.9	96.9
	1	59	3.1	3.1	100.0
	Total	1905	100.0	100.0	

Fig.14. Table for outlier and other distribution stats

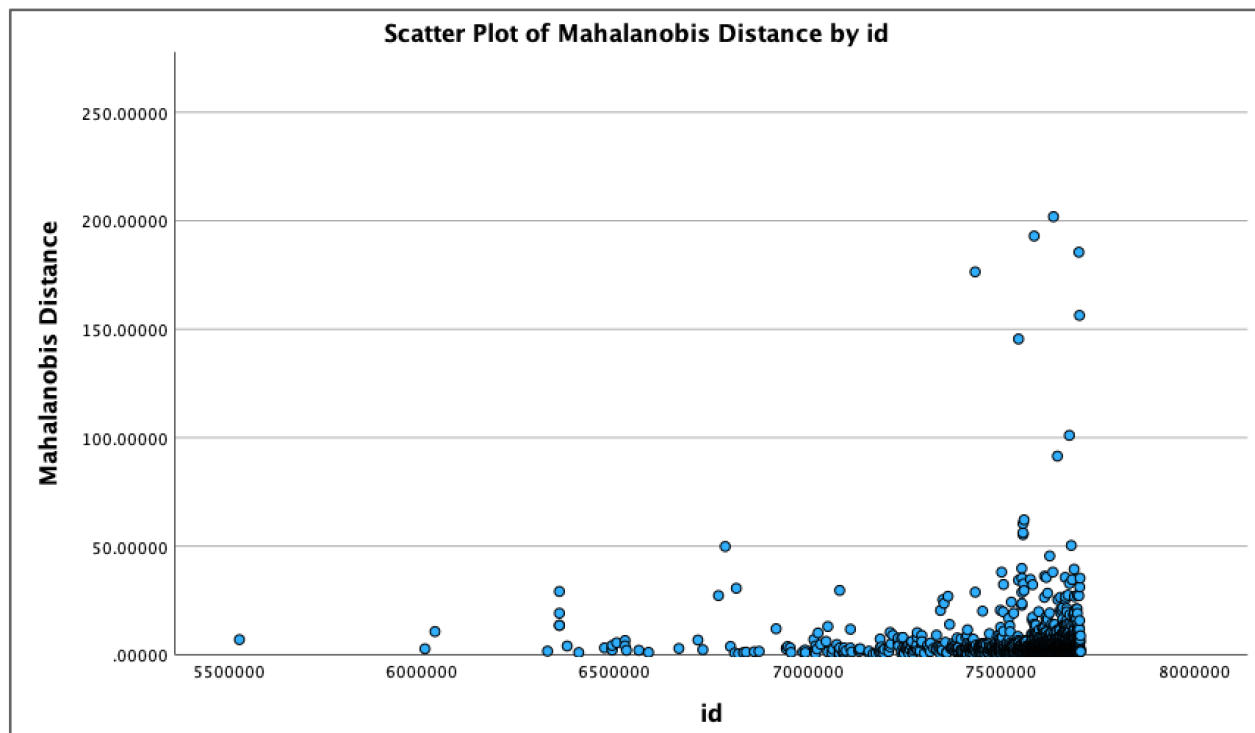


Fig.15. Mahalanobis scatterplot

### Dimension reduction

We will perform factor analysis (Brownlee, 2020) to reduce the number of dimensions, the factor analysis was performed on the following variables:

Maid\_room balcony barbecue\_area built\_in\_wardrobes central\_ac childrens\_play\_area childrens\_pool concierge covered\_parking kitchen\_appliances lobby\_in\_building  
 maid\_service networked pets\_allowed private\_garden private\_gym private\_jacuzzi private\_pool  
 Security shared\_gym shared\_pool shared\_spa study vastu\_compliant  
 View\_of\_landmark view\_of\_water/walk\_in\_closet

Below the program to perform the factor analysis

```

MULTIPLE CORRES VARIABLES=maid_room balcony barbecue_area built_in_wardrobes central_ac
childrens_play_area childrens_pool concierge covered_parking kitchen_appliances lobby_in_building
maid_service networked pets_allowed private_garden private_gym private_jacuzzi private_pool
security shared_gym shared_pool shared_spa study vastu_compliant view_of_landmark view_of_water
walk_in_closet
/ANALYSIS=maid_room(WEIGHT=1) balcony(WEIGHT=1) barbecue_area(WEIGHT=1)
built_in_wardrobes(WEIGHT=1) central_ac(WEIGHT=1) childrens_play_area(WEIGHT=1)
childrens_pool(WEIGHT=1) concierge(WEIGHT=1) covered_parking(WEIGHT=1) kitchen_appliances(WEIGHT=1)
lobby_in_building(WEIGHT=1) maid_service(WEIGHT=1) networked(WEIGHT=1) pets_allowed(WEIGHT=1)
private_garden(WEIGHT=1) private_gym(WEIGHT=1) private_jacuzzi(WEIGHT=1) private_pool(WEIGHT=1)
security(WEIGHT=1) shared_gym(WEIGHT=1) shared_pool(WEIGHT=1) shared_spa(WEIGHT=1) study(WEIGHT=1)
vastu_compliant(WEIGHT=1) view_of_landmark(WEIGHT=1) view_of_water(WEIGHT=1)
walk_in_closet(WEIGHT=1)
/MISSING=maid_room(PASSIVE,MODEIMPU) balcony(PASSIVE,MODEIMPU) barbecue_area(PASSIVE,MODEIMPU)
built_in_wardrobes(PASSIVE,MODEIMPU) central_ac(PASSIVE,MODEIMPU) childrens_play_area(PASSIVE,MODEIMPU)
childrens_pool(PASSIVE,MODEIMPU) concierge(PASSIVE,MODEIMPU) covered_parking(PASSIVE,MODEIMPU)
kitchen_appliances(PASSIVE,MODEIMPU) lobby_in_building(PASSIVE,MODEIMPU) maid_service(PASSIVE,MODEIMPU)
networked(PASSIVE,MODEIMPU) pets_allowed(PASSIVE,MODEIMPU) private_garden(PASSIVE,MODEIMPU)
private_gym(PASSIVE,MODEIMPU) private_jacuzzi(PASSIVE,MODEIMPU) private_pool(PASSIVE,MODEIMPU)
security(PASSIVE,MODEIMPU) shared_gym(PASSIVE,MODEIMPU) shared_pool(PASSIVE,MODEIMPU) shared_spa(PASSIVE,MODEIMPU)
study(PASSIVE,MODEIMPU) vastu_compliant(PASSIVE,MODEIMPU) view_of_landmark(PASSIVE,MODEIMPU)
view_of_water(PASSIVE,MODEIMPU) walk_in_closet(PASSIVE,MODEIMPU)
/DIMENSION=7
/NORMALIZATION=VPRINCIPAL
/MAXITER=100
/CRITERIA=.00001
/PRINT=CORR DISCRIM
/PLOT=OBJECT(20) DISCRIM(20).

```

OBSCO1_1	OBSCO2_1	OBSCO3_1	OBSCO4_1	OBSCO5_1	OBSCO6_1	OBSCO7_1
.10	-.25	-1.12	-1.88	.65	-.07	-.49
-.17	-.76	-.94	-.24	-.36	.33	.70
.38	-.63	1.10	.02	.13	.70	-2.73
-.70	.83	.08	-.32	.35	-.21	-.67
.75	.09	.26	1.43	.28	1.25	.45
.75	.09	.26	1.43	.28	1.25	.45
1.24	.37	1.63	1.39	2.09	-.91	-1.52
.75	.09	.26	1.43	.28	1.25	.45
-1.16	1.54	.04	-.17	.26	-.14	-.06
-.07	-1.03	-.42	.17	-.28	.65	-.36
-.55	.31	-.28	-.29	.30	.25	-.48
-.53	-.12	-.46	.30	-.45	-.04	.81
-.54	.09	-.36	-.05	.10	-.19	.24
.00	-.64	-1.02	-.73	-.16	1.71	.19
.00	-.64	-1.02	-.73	-.16	1.71	.19
.00	-.64	-1.02	-.73	-.16	1.71	.19
.00	-.64	-1.02	-.73	-.16	1.71	.19
-.33	-.34	-.16	-.22	-.12	.36	-.68
-.22	-.50	-.99	-.37	-.75	.15	1.07
.42	-.70	2.24	1.28	-.90	2.36	.09
-.10	-.24	.31	1.30	-.71	-1.58	.12

In the below table, we have the total variance explained using Factor analysis. The first column component is indicative of the number of factors in the dataset that we input. From the table, we see that there are a total of 7 factors. For the next component of the table which is the Eigenvalues indicates the variances among all the factors. The variables have been standardized, indicating that each variable contains a variance of 1 while the total variance is 7 (the count of total factors in the dataset).

Model Summary				
Dimension	Cronbach's Alpha	Variance Accounted For		
		Total (Eigenvalue)	Inertia	% of Variance
1	.903	7.641	.283	28.298
2	.694	3.013	.112	11.158
3	.411	1.655	.061	6.129
4	.352	1.512	.056	5.599
5	.097	1.104	.041	4.087
6	-.047	.957	.035	3.544
7	-.190	.845	.031	3.130
Total		16.726	.619	
Mean	.604 <sup>a</sup>	2.389	.088	8.850

a. Mean Cronbach's Alpha is based on the mean Eigenvalue.

Fig.16. Table for model summary

### 4.3 Machine Learning

This section would be dedicated to using the prepared dataset to build models for prediction purposes and understanding the comparison of different models. SPSS Modeler allows researchers to build and validate complex models which can be used for predictions. The visual interface allows users to create complex data manipulation and preparation, along with modeling pipelines in a simple drag and drop manner. This can then be leveraged to run the end to end pipeline to draw the necessary results from the modeling phase. In the subsequent steps, we would be using SPSS Modeler to prepare, split and model the data for prediction and then compare the results of different models to pick the best fit.

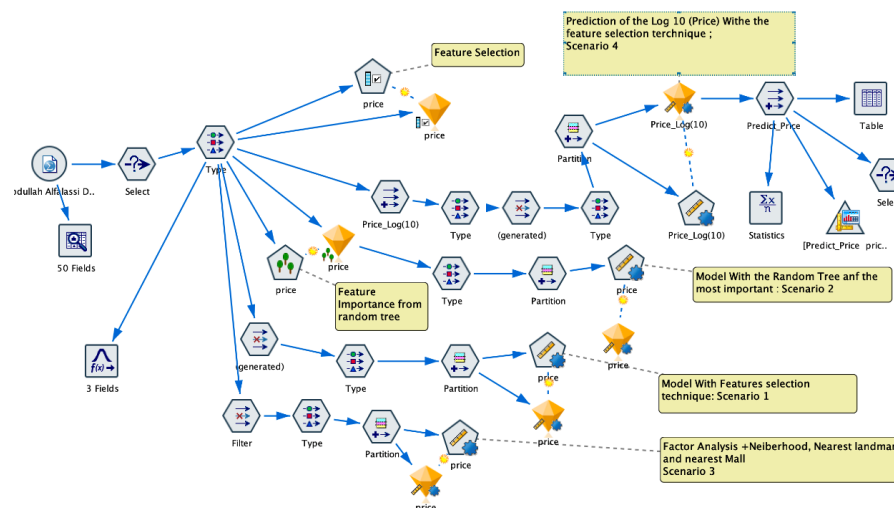




Fig.17. Baseline Modelling workflow using SPSS Modeler

This part of the report is about SPSS Modeler which has been used to prepare the dataset, make selections based on certain criteria as well as some of the modeling aspects to understand data and their different relationships. In the section below, we describe each of the modules that have been used in the modeling process.

- Dataset import: This section of the function has been used to get the raw standard dataset from any data source into the work environment which is a crucial step in the process
- Type: The function here denotes the usage of methods to typecast the variables that are present in the data. We can also use this function to set the data type of the variable like continuous, categorical, nominal or ordinal.
- Select: Similar to the SQL select function, we would be using this to select the appropriate subset of the data. There are numerous conditions that can be defined using this function and the data is subsetted.
- Variable Importance: In this module, we determine the factor importance of the variables in the dataset. Factor importance tells us the features which are important during the model building process for the target feature and is a very important step in the modeling process. Based on that we see the top features below based on their variable importance scores, of them number of bedrooms and bathroom, neighborhood and size in sqft being the top most important features.

#### 4.3.1 Model Description

To understand a few details about the predictive models in our subsequent sections, we would like to describe about the same in an overview.

##### LSVM

LSVM is a machine learning technique and it has a linear nature which can be used for data classification. If the dataset has numerous predictor fields, this modeling technique is better suited in such scenarios. This form of model is like SVM but can handle a larger number of records (Jair, 2020). In a simple example of the LSVM model, if we have a dataset with two levels A and B, the classifier can be used to segregate both these categories as either of the two groups coming from the A and B. (Zhang, 2012) For any 2-d space, the data points can be divided by a straight line into two distinct groups which is the purpose of the SVM technique. There is also a nonlinear version of SVM which is separated by a straight line but for non-linear datasets. (javatpoint, 2022)

## Generalized Linear Regression

Generalized Linear Model (IBM, 2022) is a form of modeling technique and this is kind of a parent technique which covers many sub-models within it. GLM allows a predictor variable  $y$  to obtain an error distribution other than a standard distribution. GLM includes many models under its umbrella like Linear Regression, Logistic Regression and Poisson Regression. For the generalized linear regression model, there are mainly three main components of which the first one is random component  $Y$ , a linear predictor and a final link function. These components invariably define the input and output functions of the model as a prediction property at the final stage of the data input functions. (Zhang, 2022)

## Neural Networks

Neural Networks on the other hand is a collection of algorithms which determine underlying relationships in a dataset in the same way a human brain operates. (Emmert-Streib, 2020) The interesting fact and power of the neural networks is that it can adapt to the varying input layers in order to generate the best possible outputs without any interventions. The roots of Neural Networks are derived from Artificial Intelligence and have been consistently used for advanced predictions and analysis. (Hardesty, 2017) These Neural nets are a connection between the actions and effects of different inputs along with their output layers. The setup of the system is such that it replicates the human brain for interpretations. (Schmidhuber, 2015)

We will perform 4 scenarios:

- Scenario 1: The model we build with features selection that are the most correlated to the price
- Scenario 2: The model is built with features selected using a random tree
- Scenario 3: The model is built with the factor analysis components
- Scenario 4: The price was skewed then we transformed the price to  $\log_{10}(\text{price})$ . We predicted the Log 10 price instead of the price

### Scenario 1

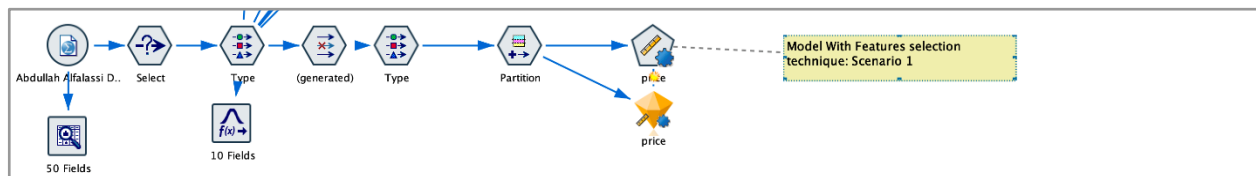
The feature selection process from the modeling helped us determine the best features to be chosen for the modeling purpose. We see that top features recommended by the model are

number of bedrooms and bathrooms, pets allowed, quality, security and lobby in building and use them accordingly for our modeling phase in the first case.

Model Summary Annotations					
<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	Rank		
Rank	Field	Measurement	Importance	Value	
1	no_of_bedrooms	Nominal	Important	1.0	
2	no_of_bathrooms	Nominal	Important	1.0	
4	pets_allowed	Flag	Important	1.0	
5	quality	Nominal	Important	1.0	
6	security	Flag	Important	1.0	
7	lobby_in_building	Flag	Important	1.0	
8	shared_spa	Flag	Important	1.0	
9	barbecue_area	Flag	Important	1.0	
10	walk_in_closet	Flag	Important	1.0	
11	view_of_water	Flag	Important	1.0	
12	childrens_play...	Flag	Important	1.0	
13	central_ac	Flag	Important	1.0	
14	maid_room	Flag	Important	1.0	
15	study	Flag	Important	1.0	
16	covered_parking	Flag	Important	1.0	
17	concierge	Flag	Important	0.999	
18	balcony	Flag	Important	0.996	
19	built_in_wardro...	Flag	Important	0.971	
20	shared_gym	Flag	Important	0.964	

Fig.18. Factor significance from modeling

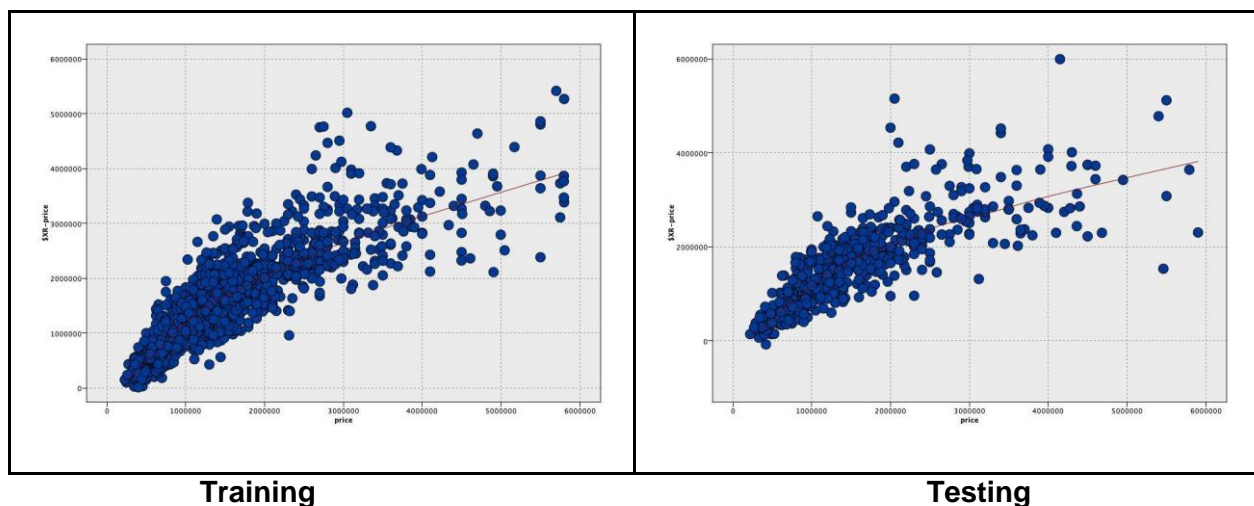
In the above feature selection process done by our modeling technique, based on the importance score, we observe the top 18 features. It is observed that the number of bedrooms, bathrooms, pet restrictions, quality and security are some of the top features chosen by our feature importance process. On the following chart we can also see the feature importance ranking based on their scores. While these features can be used for the modeling process, it is also worthwhile to understand the model performances having chosen all the features (and considering a baseline performance of the model).



<div>  File            Generate            View            Preview            ?            </div>						
<div> <b>Model</b> Graph Summary Settings Annotations </div>						
<div> Sort by: Use Ascending Descending Delete Unused Models View: Testing set </div>						
Use?	Graph	Model	Build Time (mins)	Correlation	No. Fields Used	Relative Error
<input checked="" type="checkbox"/>		Generalized Linear 1	< 1	0.775	21	0.4
<input checked="" type="checkbox"/>		LSVM 1	< 1	0.769	21	0.411
<input checked="" type="checkbox"/>		Linear 1	< 1	0.766	8	0.414
<input checked="" type="checkbox"/>		Neural Net 1	< 1	0.749	21	0.440
<input checked="" type="checkbox"/>		Linear-AS 1	< 1	0.730	21	0.495

Fig.19. Scenario 1 model interpretation and workflow

In this scenario, the correlation is found to be the highest for Generalized Linear 1 model with a score of 77%, while second is LSVM with 76.9% correlation. It might be the case that the Generalized Linear Model with the explanatory features were subject to measurement error and hence we achieved a collectively lower score band for the model evaluation. (Armstrong, 2007) In the subsequent scenarios, we will perform the same steps with different use cases to compare the model scores in each phase.



In the above train test predicted data spread, it is observed that the predicted data points are spread well around the regression line (red line). This indicates that our predicted values have

been correctly determined by the model otherwise the data points would have been spread far away from the regression line.

## Scenario 2

In this scenario, we use a popular modeling called Random Trees to determine the top most important features which we will use for the model prediction step. (Jean-François, 2005) It is observed that the number of bathrooms, nearest mall, number of bedrooms, quality and private jacuzzi are some of the top features which have been deemed best by the model. Hence, we will use some of them to build our next predictive model.

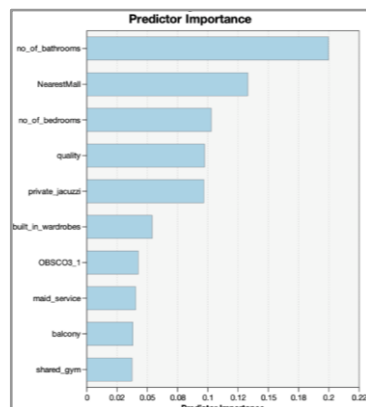
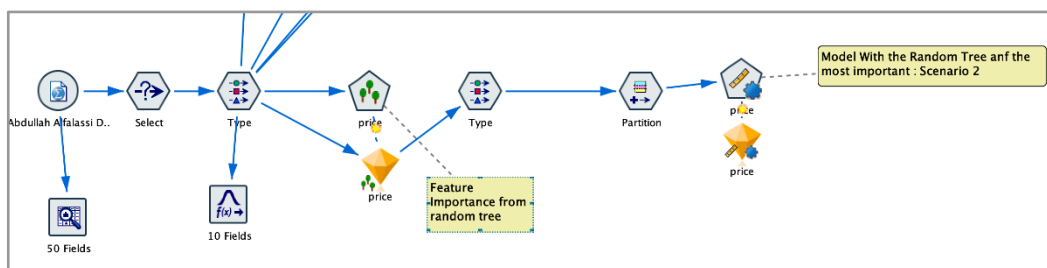


Fig.20. Factor significance using Random Forest

Based on the above selection of features, we create a model using Random Trees to determine the baseline performance of the model as well as some other summary statistics, having chosen some of the top features from the above summary.













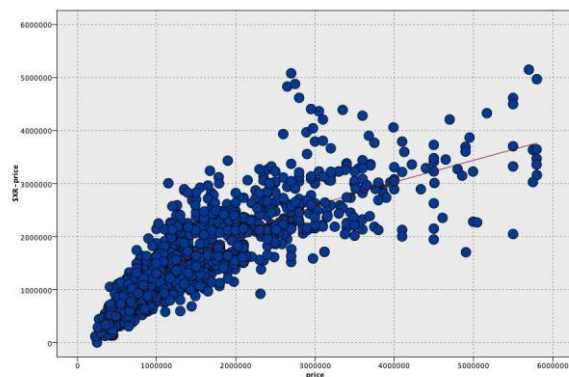
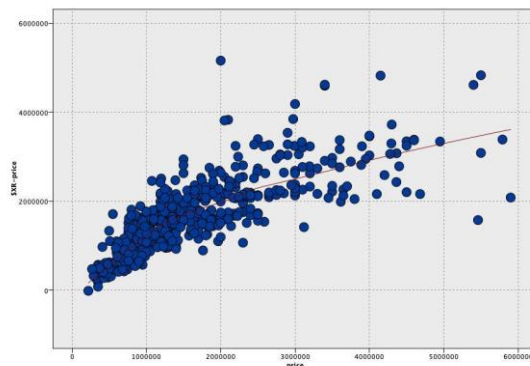
Sort by: <span>Use</span> <span>Ascending</span> <span>Descending</span> <span>Delete Unused Models</span> View: <span>Testing set</span>						
Use?	Graph	Model	Build Time (mins)	Correlation	No. Fields Used	Relative Error
<input checked="" type="checkbox"/>		 Generalized Linear 1	< 1	0.774	11	0.402
<input checked="" type="checkbox"/>		 LSVM 1	< 1	0.767	11	0.414
<input checked="" type="checkbox"/>		 Linear 1	< 1	0.764	6	0.417
<input checked="" type="checkbox"/>		 Neural Net 1	< 1	0.751	11	0.437
<input checked="" type="checkbox"/>		 Regression 1	< 1	0.611	3	0.627

Fig.21. Scenario 2 model interpretation and workflow

In this second scenario, we used a Random Tree model to determine the factor significance of our dataset and use the same features to build our next models. We see that Generalised Linear 1 has again the highest correlation with 77.4% while second is LSVM with 76.7% score. This clearly shows us the model which needs to be considered in this case even with high significance features considered for the step.



Training



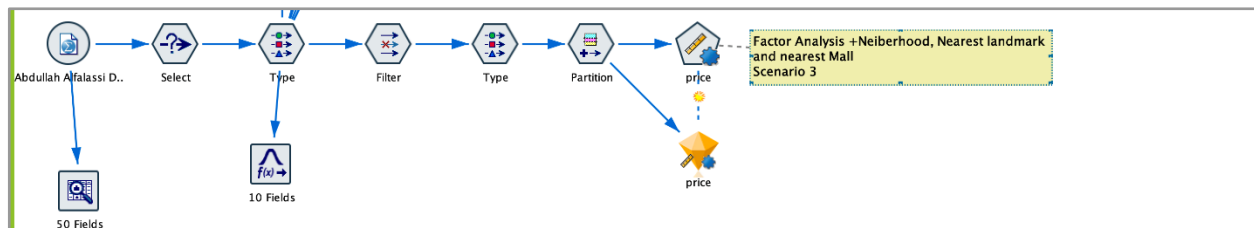
Testing

In this scenario, the predicted values around the regression line are spread evenly, though some of the data points are deviated a little. For the test set, the data points on the far right are deviated which indicates the outlier prices or indicates that the model might not be accurate enough for extreme values when predicting.

### Scenario 3:

The model is built with the factor analysis components performed in the previous step. We perform different scenarios using different combinations to determine the best fit candidate model for the prediction step. In this process we choose the factors that we obtained from factor analysis and then build the workflow on SPSS Modeler.

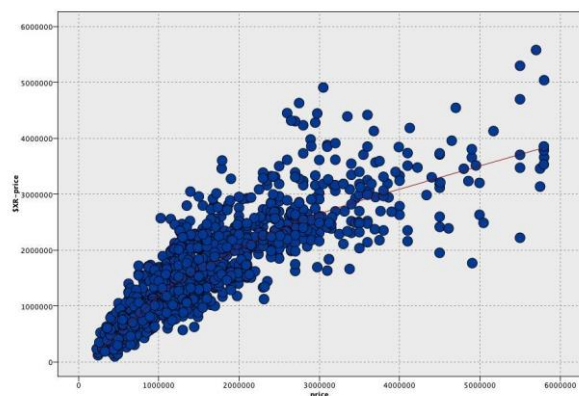
Field	Measurement	Values	Missing	Check	Role
neighborhood	Nominal	"Al Barari", "Al B...		None	Input
price	Continuous	[220000.0,3.5E7]		None	Target
no_of_bedrooms	Nominal	0.0,1.0,2.0,3.0,...		None	Input
no_of_bathrooms	Nominal	1.0,2.0,3.0,4.0,...		None	Input
OBSCO1_1	Continuous	[-1.161626731...		None	Input
OBSCO2_1	Continuous	[-1.504006464...		None	Input
OBSCO3_1	Continuous	[-1.605510004...		None	Input
OBSCO4_1	Continuous	[-6.305381163...		None	Input
OBSCO5_1	Continuous	[-5.032413013...		None	Input
OBSCO6_1	Continuous	[-4.197952016...		None	Input
OBSCO7_1	Continuous	[-3.937655334...		None	Input
NearestLandmark	Nominal	"Al Makhtoum I...		None	Input
NearestMall	Nominal	"City Centre Mir...		None	Input



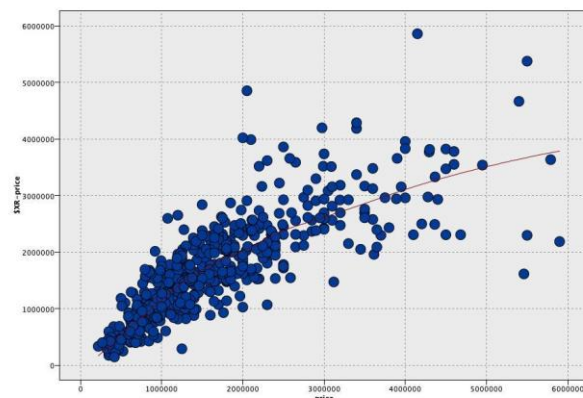
Sort by:	Use	Ascending	Descending	Delete Unused Models	View:	Testing set
Use?	Graph	Model	Build Time (mins)	Correlation	No. Fields Used	Relative Error
✓		Generalized Linear 1	< 1	0.771	12	0.406
✓		Neural Net 1	< 1	0.767	12	0.414
✓		LSVM 1	< 1	0.763	12	0.419
✓		Linear-AS 1	< 1	0.743	12	0.459
✓		CHAID 1	< 1	0.724	8	0.477

Fig.22. Scenario 3 model interpretation and workflow

In the above scenario, we observe again that the Generalized Linear 1 model is the best performing model with 77% correlation while in the second is 76.7%. Hence, we see that the Generalized Linear model has been consistently performing the best in all the above performed scenarios.



Training



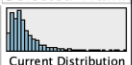

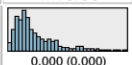
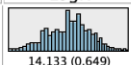
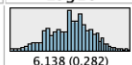
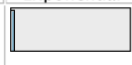
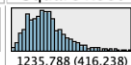


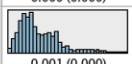
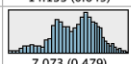
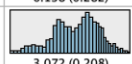

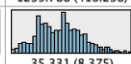
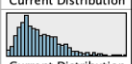

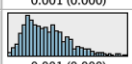
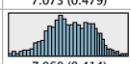
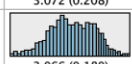
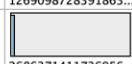
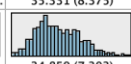
Testing

The same is noticed in the above predicted variable spread around the regression line again, for both the train and test set. For the test set, there are some data points spread a bit farther away from the line at the extreme right which indicate again the problem with extreme values.

#### Scenario 4:

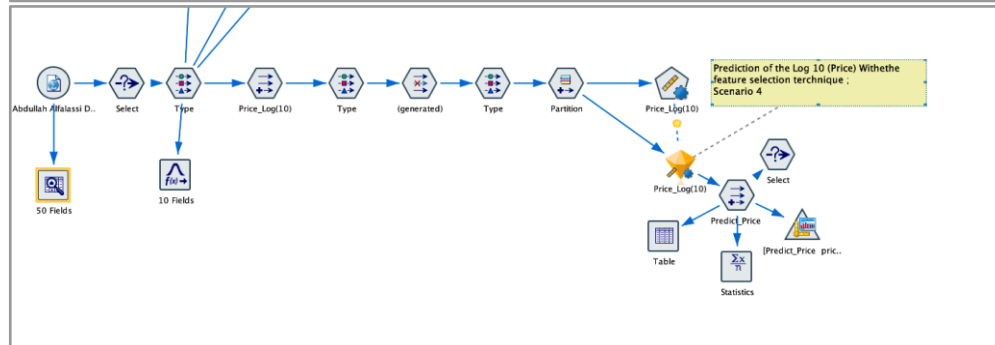
In the 4th scenario, taking the skewness of the target variable into consideration, we use a normalization technique to standardize the feature. The price was skewed which we then transformed using log transformation technique to  $\log_{10}(\text{price})$ . Subsequently, we predicted the Log 10 price instead of the price in this scenario to determine the model performances.

In the table below the distribution of the price is skewed, which is the reason we then chose to predict the  $\log_{10}$  price which has better distribution. It is a rule of thumb to always normalize the dataset using different techniques like square root transformation, log transformation etc. which helps in removing all sorts of biases from the modeling step.

Field	Selected Tran...	Current Distri...	Inverse	LogN	Log10	Exponential	Square Root
price	 Current Distribution	 1700333.243 (1260...)	 0.000 (0.000)	 14.133 (0.649)	 6.138 (0.282)		 1235.788 (416.238)
size_in_sqft	 Current Distribution	 1318.376 (632.823)	 0.001 (0.000)	 7.073 (0.479)	 3.072 (0.208)	 1269098728391863...	 35.331 (8.375)
price_per_sqft	 Current Distribution	 1268.494 (545.847)	 0.001 (0.000)	 7.060 (0.414)	 3.066 (0.180)	 2686371411736956...	 34.859 (7.303)



Field	Measurement	Values	Missing	Check	Role
neighborhood	Nominal	"Al Barari", "Al...		None	Input
price	Continuous	[220000.0, 3...		None	None
no_of_bedrooms	Nominal	0,0,1,0,2,0,3...		None	Input
no_of_bathroo...	Nominal	1,0,2,0,3,0,4...		None	Input
quality	Nominal	High, Low, Me...		None	Input
maid_room	Nominal	False, True		None	Input
barbecue area	Nominal	False, True		None	Input
childrens_play ...	Nominal	False, True		None	Input
concierge	Nominal	False, True		None	Input
lobby in buildi...	Nominal	False, True		None	Input
pets allowed	Nominal	False, True		None	Input
security	Nominal	False, True		None	Input
shared_gym	Nominal	False, True		None	Input
shared_pool	Nominal	False, True		None	Input
view_of_water	Nominal	False, True		None	Input
Price_Log(10)	Continuous	[5.34242268...		None	Target



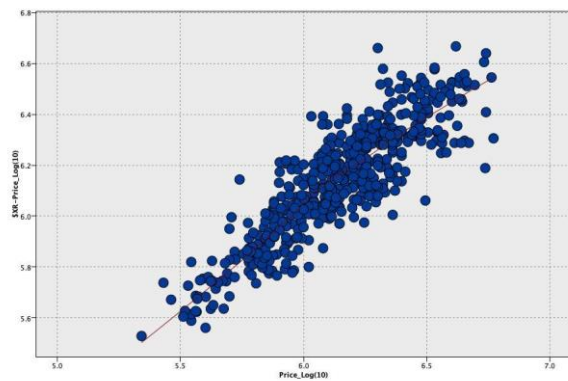
Sort by: <div>Use</div>		<div><input checked="" type="radio"/> Ascending</div> <div><input type="radio"/> Descending</div>	<div></div>	<div>Delete Unused Models</div>	View: <div>Testing set</div>	
Use?	Graph	Model	Build Time (mins)	Correlation	No. Fields Used	Relative Error
<input checked="" type="checkbox"/>		<div> Generalized Linear 1</div>	< 1	0.891	14	0.206
<input checked="" type="checkbox"/>		<div> LSVM 1</div>	< 1	0.881	14	0.224
<input checked="" type="checkbox"/>		<div> Neural Net 1</div>	< 1	0.869	14	0.247
<input checked="" type="checkbox"/>		<div> Regression 1</div>	< 1	0.690	2	0.525

Fig.23. Scenario 4 model interpretation and workflow

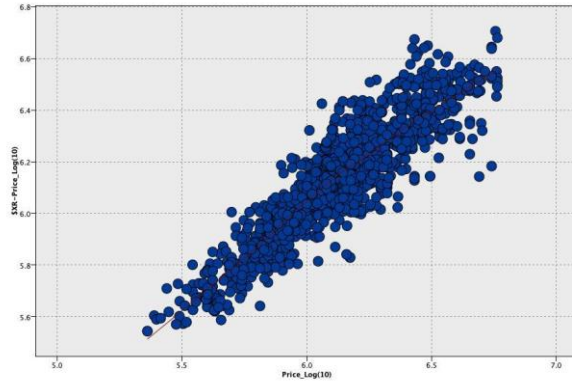
In this modeling process, we observe that the Generalized Linear 1 model has the highest score again with 89% correlation value. This significantly improves the model scores with the Generalized Linear 1 showing consistency again. The formula for performing the transformation on the target feature is as follows -

- Predicted price = exponential (predicted log<sub>10</sub>(price)\*log<sub>10</sub>).
- Correlation (log<sub>10</sub>(price), predicted(log<sub>10</sub>(price)))=0.89
- Correlation (price, predicted (price)) =0.81

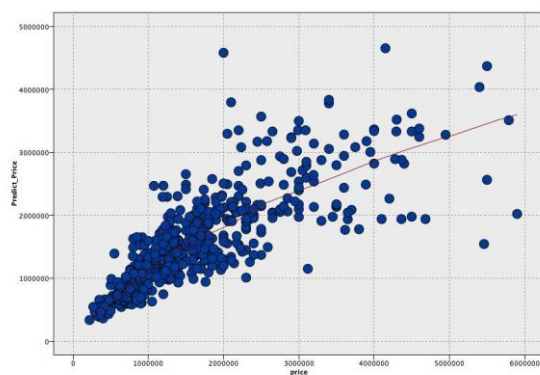
Hence, to clearly summarize, we move ahead with the Generalized Linear 1 model which has shown great performance in all the 4 scenarios while scenario 4 being the best out of all with the target feature normalization technique.



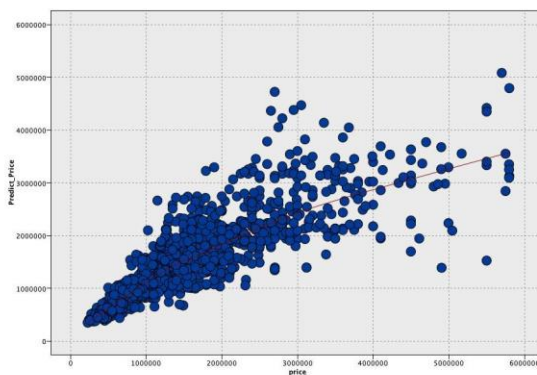
**Training**



**Testing**



**Training**



**Testing**

In the above plot for the predicted value plot around the regression line for both the target variable with and without normalization, it is observed that the data spread is significantly less around the predictor line for the first figure with the log10 transformation. It clearly indicates that normalizing the target feature has improved the predicted values significantly and is not spread farther from the regression line as shown in figure 2 without the normalization

## Chapter 5 - Conclusion

### 5.1 Conclusion

To summarize, we determined the best dataset for our problem statement, i.e., Dubai House Price list for multiple years. We then leveraged SPSS tools like Statistics and Modeller to determine the statistical significance of different values and fields in the dataset. This helped us determine the best fit features and significance factors to build our predictive models. In the later section, we leverage SPSS Modeler to build the baseline pipeline to determine the significant features as well as shape the dataset based on some modeling requirements. Finally, we use SPSS Modeler again to build the final pipeline based on the best fit features for our modeling and then implement an auto numeric model, which uses models like Linear Regression, LSVM, Generalized Linear Model (GLM) and Neural Networks to compare the predictive scores of all the models and provide us the best one at the end. Approaching the modeling phase with different scenarios also helped us understand the better performing scenario in the above cases. In our opinion, the modeling phase should be exhaustive and all cases should be considered while doing the process which we did above.

In this way, we applied the CRISP-DM process of data analytics and machine learning problem solving approach to finally land at the most viable approach for the problem solution.

Scenario considered	Algorithm	Number of features	Correlation
Scenario 1	Generalized Linear 1	21	77.5%
Scenario 2	Generalized Linear 1	11	77.4%
Scenario 3	Generalized Linear 1	12	77.1%
Scenario 4	Generalized Linear 1	14	89.1%

A summary of our modeling phase comparison is shown on the above table with the corresponding correlation scores for the same to conclude that Scenario 4 with target feature normalization along with factor significance consideration yielded the best results among all the processes.

## 5.2 Recommendations

Based on the above experiment and observations, we have used SPSS Statistics and Modeler to prepare our dataset to solve our problem statement. Even within the process, we also performed various statistical analyses to understand the dataset and the insights in-depth. Hence, after we

have the data we were able to implement Machine Learning techniques to provide a comparative analysis of the model performances. After the study, we were able to determine the Generalized Linear 1 model as the best performing model of all and we would like to recommend the same to be used to solve similar problem statements for better predictive abilities. This would help researchers and developers implement the same framework to predict the house prices in other regions of the world, whilst following the above mentioned steps and processes throughout the course of the problem.

### 5.3 Future Work

In the scope of future work, the real estate business has plenty of technology and other progressive implementation scope. Because of the growing nature of the business, organizations and individuals need robust models and analysis to derive insights and understanding of the housing market even better. For future tasks, one can also implement many other models to derive even better predictive power from the same. Not only that, data can be collected in various other dimensions to understand the impact of factors even in depth. This can then be implemented into pipelines and subsequently the entire process can be automated to implement in different places without the need for human interventions.

## Bibliography

- [1] *UAE Residential Real Estate Market: 2022 - 27: Industry share, size, growth - mordor intelligence*. UAE Residential Real Estate Market | 2022 - 27 | Industry Share, Size, Growth - Mordor Intelligence. (n.d.). Retrieved April 9, 2022, from <https://www.mordorintelligence.com/industry-reports/residential-real-estate-market-in-uae#:~:text=A%20survey%20involving%20property%20analysts,a%20modest%20rise%20in%20prices> .
- [2] Abbas, W. (2022, February 17). *UAE property prices are likely to continue rising in 2022 but at a slower pace*. Khaleej Times. Retrieved April 9, 2022, from <https://www.khaleejtimes.com/property/uae-property-prices-like-to-continue-rising-in-2022-but-at-a-slower-pace>
- [3] Morgan, O. (2021, February 20). *Deloitte Real Estate Predictions - Dubai 2021: Deloitte United Arab Emirates: Real estate and construction: Perspectives*. Deloitte. Retrieved April 9, 2022, from <https://www2.deloitte.com/ae/en/pages/real-estate/articles/deloitte-real-estate-predictions-dubai-2021.html>
- [4] *The UAE real estate sector displays signs of recovery*. (n.d.). Retrieved April 8, 2022, from [https://www2.deloitte.com/content/dam/Deloitte/xs/Documents/About-Deloitte/mepovdocuments/mepov35/uae-real-estate\\_mepov35.pdf](https://www2.deloitte.com/content/dam/Deloitte/xs/Documents/About-Deloitte/mepovdocuments/mepov35/uae-real-estate_mepov35.pdf)
- [5] *Middle East Real Estate Predictions: Dubai 2021 - deloitte*. (n.d.). Retrieved April 8, 2022, from [https://www2.deloitte.com/content/dam/Deloitte/xs/Documents/realestate/me\\_deloitte-real-estate-predictions-dubai-2021.pdf](https://www2.deloitte.com/content/dam/Deloitte/xs/Documents/realestate/me_deloitte-real-estate-predictions-dubai-2021.pdf)
- [6] Ahmed, Farhan & Maheshwari, Sandhia & Mirani, Sajid. (2020). Dubai House Prices and Macroeconomic Fluctuations: A Time Series Analysis. 13. 61-73.
- [7] *Price prediction and valuation using ... - rit scholar works*. (n.d.). Retrieved April 8, 2022, from <https://scholarworks.rit.edu/cgi/viewcontent.cgi?article=11834&context=theses>
- [8] Frank, K. (n.d.). UAE Market Review and forecast. Retrieved April 8, 2022, from <https://content.knightfrank.com/research/1064/documents/en/uae-market-review-forecast-2021-7801.pdf>
- [9] Truong, Q., Nguyen, M., Dang, H., & Mei, B. (2020). Housing price prediction via Improved Machine Learning Techniques. *Procedia Computer Science*, 174, 433–442. <https://doi.org/10.1016/j.procs.2020.06.111>

- [10] Madhuri, C. H. R., Anuradha, G., & Pujitha, M. V. (2019). House price prediction using regression techniques: A comparative study. *2019 International Conference on Smart Structures and Systems (ICSSS)*. <https://doi.org/10.1109/icsss.2019.8882834>
- [11] Wang L., Wang G., Yu H., Wang F. (2020). Prediction and analysis of residential house price using a flexible spatiotemporal model. *Journal of Applied Economics, Taylor & Francis Online*. <https://doi.org/10.1989/15140326.2022.2045466>
- [12] Manjula R., Jain S., Srivastava S., Kher P.R. (2017). Real estate value prediction using multivariate regression models. *IOP Conference Series: Materials Science and Engineering*. <https://doi.org/10.1088/1757-899X/263/4/042098>
- [13] Hacievliyagil N., Drachal K, Eksi I. (2021). *Economies, MDPI*. <https://doi.org/10.3390/economies10030064>
- [14] Wang F., Zou Y., Zhang H., (2019) Shi H., House Price Prediction Approach based on Deep Learning and ARIMA Model, *2019 IEEE 7th International Conference on Computer Science and Network (ICCSNT), IEEE*, <https://doi.org/10.1109/ICCSNT47585.2019.8962443>
- [15] Chen X., Wei L., Xu J., (2017) House Price Prediction Using LSTM, *Arxiv, Cornell University*, <https://doi.org/10.48550/arxiv.1709.08432>
- [16] Jair C., Farid Garcia L, Lisbeth R., Asdrubal L., A comprehensive survey on support vector machine classification: Applications, challenges and trends, Elsevier, 30 September 2020, <https://www.sciencedirect.com/science/article/abs/pii/S0925231220307153>
- [17] IBM (2022), IBM Cloud Pak for Data, IBM Documentation for SPSS, <https://www.ibm.com/docs/en/cloud-paks/cp-data/4.6.x>
- [18] Data Solut (2022), CRISP-DM: Basics, goals and the 6 phases of the data mining process, <https://datasolut.com/crisp-dm-standard/>
- [19] Kaggle (2020), Dubai Property Prices, <https://www.kaggle.com/datasets/dataregress/dubai-properties-dataset>
- [20] Selva Prabhakaran (2019), Mahalanobis Distance – Understanding the math with examples (python), <https://www.machinelearningplus.com/statistics/mahalanobis-distance/>
- [21] Jason Brownlee (2020), Introduction to Dimensionality Reduction for Machine Learning, May 6 2020, <https://machinelearningmastery.com/dimensionality-reduction-for-machine-learning/>
- [22] Larry Hardesty (2017), Explained: Neural networks, MIT News Office, <https://news.mit.edu/2017/explained-neural-networks-deep-learning-0414>
- [23] JavaTPoint (2022), Support Vector Machine Algorithm, <https://www.javatpoint.com/machine-learning-support-vector-machine-algorithm>

- [24] Jürgen Schmidhuber (2015), Deep learning in neural networks: An overview, January 15 2015, Science Direct, <https://doi.org/10.1016/j.neunet.2014.09.003>
- [25] Xichu Zhang (2022), An introduction to the generalized linear model (GLM), April 8 2022, Towards Data Science, <https://towardsdatascience.com/an-introduction-to-the-generalized-linear-model-glm-e32602ce6a92>
- [26] M. Dash (1997), Feature Selection for Classification, 21 March 1997, Intelligent Data Analysis , Elsevier, <http://machine-learning.martinsewell.com/feature-selection/DashLiu1997.pdf>
- [27] Jean-François Le Gall (2005), Random trees and applications, 2005, <https://doi.org/10.1214/154957805100000140>
- [28] Ben Armstrong (2007), Measurement error in the generalized linear model, 27 June 2007, Communications in Statistics - Simulation and Computation, Taylor & Francis Online, <https://doi.org/10.1080/03610918508812457>
- [29] Zhang, Y. (2012). Support Vector Machine Classification Algorithm and Its Application. In: Liu, C., Wang, L., Yang, A. (eds) Information Computing and Applications. ICICA 2012. Communications in Computer and Information Science, vol 308. Springer, Berlin, Heidelberg. [https://doi.org/10.1007/978-3-642-34041-3\\_27](https://doi.org/10.1007/978-3-642-34041-3_27)
- [30] Frank Emmert-Streib, Zhen Yang, Han Feng, Shailesh Tripathi, Matthias Dehmer, (2020), An Introductory Review of Deep Learning for Prediction Models With Big Data, 28 February 2020, <https://doi.org/10.3389/frai.2020.00004>