Rochester Institute of Technology

RIT Digital Institutional Repository

Theses

12-2022

House Price Classification using Clustering Algorithms

Hamad Ahli
hsa4486@rit.edu

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

Recommended Citation
Ahli, Hamad, "House Price Classification using Clustering Algorithms" (2022). 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.

House Price Classification using Clustering Algorithms

by

Hamad Ahli

A Capstone Submitted in Partial Fulfillment 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
2022 - December

RIT

Master of Science in Professional Studies:

Data Analytics

Graduate Capstone Approval

Student Name**: Hamad Ahli**

Graduate Capstone Title**:** House Price Classification using Machine Learning Algorithms

Graduate Capstone Committee:

Name: Dr. Sanjay Modak Date:

Chair of committee

Name: Dr.Ioannis Karamitsos Date:

Member of committee

Acknowledgments

First and foremost, I'd like to thank God almighty for blessing me throughout this journey.

I would also like to thank my mentor Dr. Ioannis Karamitsos for his constructive criticism, support, and guidance. His deep knowledge and professionalism have helped me push myself to achieve things I never thought I would be able to and I will always be thankful for that.

Furthermore, I would like to thank Dr. Sanjay Modak, Chair of Graduate Programs and Research for his helpful care and beneficial feedback during all the stages of my capstone.

Finally, I would like to thank my parents and siblings for their support and love during the most stressful times. I am very lucky to have them and will forever be grateful to them.

Thank you to all the teaching staff at RIT Dubai that have helped me by providing me the knowledge necessary to complete this capstone.

Abstract

The housing segment is one of the most lucrative industries in almost all parts of the world, and with an emerging place like Dubai with global attraction the real-estate market is set to expand more. With this, there is an importance to be able to understand such markets with in-depth expertise which not only helps to be a subject matter expert, but also provide recommendations and insights to customers and stakeholders. According to Asteco, UAE would witness an addition of 38,500 apartments and 3,800 villas and Dubai is estimated to account the most with 30,000 flats and 3,500 villas in 2022. Abu Dhabi is expected to see around 2,000 residential units to be given to Reem Island, 2,000 each in Al Raha Beach and Yas Island and 1,200 in Saadiyat Island. In January 2022, more than 53% transactions were for ready properties and 47% for off-plan properties.(Frank, 2022)

There are different methods to segment properties, which is mainly dependent on collecting information about the apartment done through real estate agents or construction groups, who provide the information publicly or on request. With the evolution of the housing market, evidently due to the expansion of population and other development scope in different regions, the need to develop effective marketing strategies with high quality content as well as relevance has become key to sustenance. It becomes a starting point to building relationships with consumers, and this can be done by marketing the appropriate property to the right customer groups through online email marketing, brochures as well as pamphlets.

In this report, we plan to use commonly available datasets, which are basically Dubai property records collected. We plan to implement an unsupervised clustering technique on the data to segment apartments/properties based on different traits. Even before that, we would be going through the details of the dataset through some exploratory data analyses, and then cleaning the data for inconsistency and then finally clustering the apartment ids based on different traits.

Keywords: Price Clustering, Dubai Properties, K-means, Classification, Machine Learning, Random Forest.

Table of Contents

List of Figures

Chapter 1 - Introduction

1.1 Problem Statement

The modern era has seen an increase in customer demands and buying patterns, they have become smarter than ever when it comes to choosing a service or product for purchase. This is one of the key problem areas wherein businesses are trying to identify patterns as well as similarities among users to be able to promote their product and services better. A typical example would be the user base who visit websites like Noon or Amazon on a daily basis. These companies would need to understand the needs of their users to be able to roll out the appropriate offers and products in the long run. Knowing their users is one of the keys to unleashing their business metrics. Without knowing their customers, it would lead to promoting the wrong content and marketing strategies to the incorrect cohort of users. This has a long term business impact and would result in customers churning from the website to look for other alternatives.

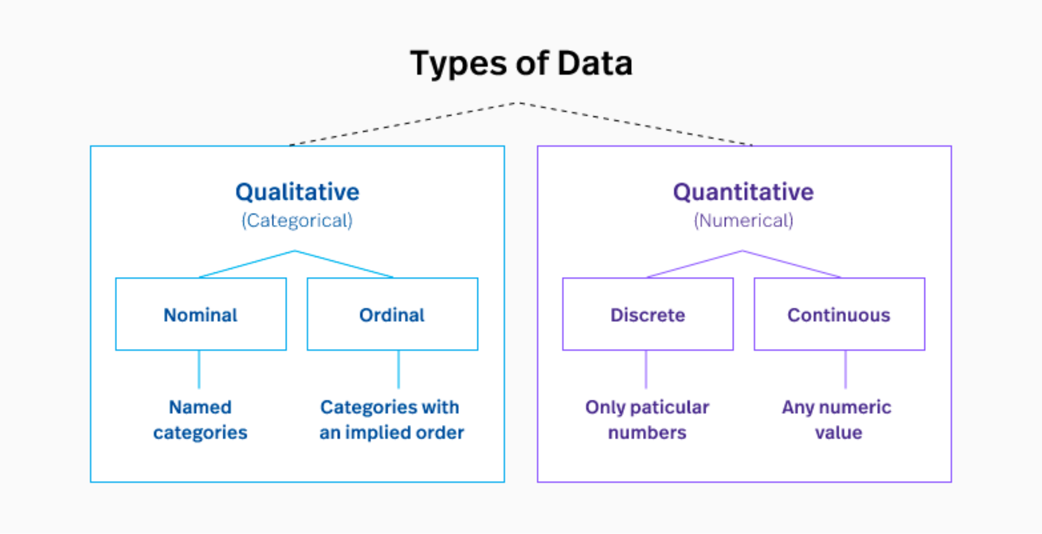
Hence, it is very important that organizations and businesses know their consumers with respect to their interests, purchasing patterns, demographics etc. This is where customer segmentation or customer profiling comes into play, wherein data analysis techniques are implemented to bucket users based on numerous traits like purchasing history, viewing interests, and many other factors. (Morgan, 2021) This enables the identification of users with similar interests, so as to roll out the appropriate program/promotion to them. Sending the right content to the right cohort of users enables better conversion rates and long term retention, which in turn boosts the revenue and other performance metrics for the company. For instance, if we know that users from cohort A love to buy fashion products, we would be better off targeting fashion products to users of cohort A, and not send them promotions related to toys or other goods. The key lies in the process of identifying cohort A, which is very important in the modern industry, be it in airlines, e-commerce, hotel industry, fashion or many other domains. Irrespective of the industry or line of work, it is always important that proper customer segmentation or customer profiling is done to have better context about its users.

1.2 Introduction

There has been a great disruption in the housing market in recent times due to COVID-19, and this has resulted in tremendous business changes and revenue stream impacts. Not only that, the housing purchase trends have seen a drastic change as well. This has necessitated the importance to understand these markets in terms of data, to be able to prepare for further global impacts in the future. When global problems arise, people tend to hold on to their liquidity in terms of financials and tend to invest less on anything including properties. This is when the prices for properties go down with the decreasing demand, and data analyses around such factors help us understand the world better in terms of businesses.

In the considered dataset, we have multiple attributes defining a given property ID which would help us understand the factors that affect the prices of a property at any given time.

The governments have tried to change how the housing industry functions with the citizens as well as the expats in the UAE, with the rise in global tech cultures in places like Abu Dhabi and Dubai. Governments have started to hand out visas for expats that grant ten years citizenships for them to make it convenient to purchase properties, and settle. These have been done to cater to the increasing demand of the property share in the country as well as make it easier for investors and settlers to accommodate in the UAE.



The above figure is indicative of different types of fields that consist of a dataset, in terms of qualitative and quantitative. Different models behave differently to different types of target variables, which can be nominal, ordinal, discrete or continuous. Now, the columns can further be categorized in the following manner -

Id - unique identifier of the property
Neighborhood - the neighborhood where the property belong to
Latitude - the geographical coordinates of the property

Longitude - the geographical coordinates of the property

Price - the price of the property

Size_in_sqft - the property size in square feet

Price_per_sqft - the price per square feet of the property

No_of_bedrooms - the total number of bedrooms in the property

No_of_bathrooms - the number of bathrooms in the property

Quality - the quality of the property scored

Maid_room - the maid rooms present in the property

Unfurnished - indicates if the property is furnished or unfurnished

Balcony - indicates if the property has a balcony or not

Barbecue_area - indicates if the property has a barbeque area

Built_in_wardrobes - indicates if wardrobes are built in the property

Central_ac - indicates if the property has central air conditioning

Childrens_play_area - indicates if the property has children's play area

Childrens_pool - indicates if there is a children’s pool present

Concierge - indicates if concierge service is provided in the apartment or not

Covered_parking - indicates if there is covered parking available in the apartment

Kitchen_appliances - indicates if kitchen appliances are available

Lobby_in_building - indicates if there is a lobby present in the building

Maid_service - indicates if the apartment provides maid service

Pets_allowed - this shows if there are pets allowed into the property

Private_garden - indicates if there are private gardens available

Private_gym - indicates if there are private gyms provided in the property

Clustering dataset involves multiple steps, before even forming the clusters. As shown in the figure given below, the first step is the data acquisition phase wherein we would need to obtain the appropriate dataset for our entire exercise. The next step involves the pre-processing step wherein the dataset is cleaned and is made ready for the unsupervised machine learning algorithm to be able to identify the features appropriately. Messy data is always bad for machine learning algorithms because it can lead to errors and biases, and to avoid such problems the dataset needs to be cleaned and normalized.

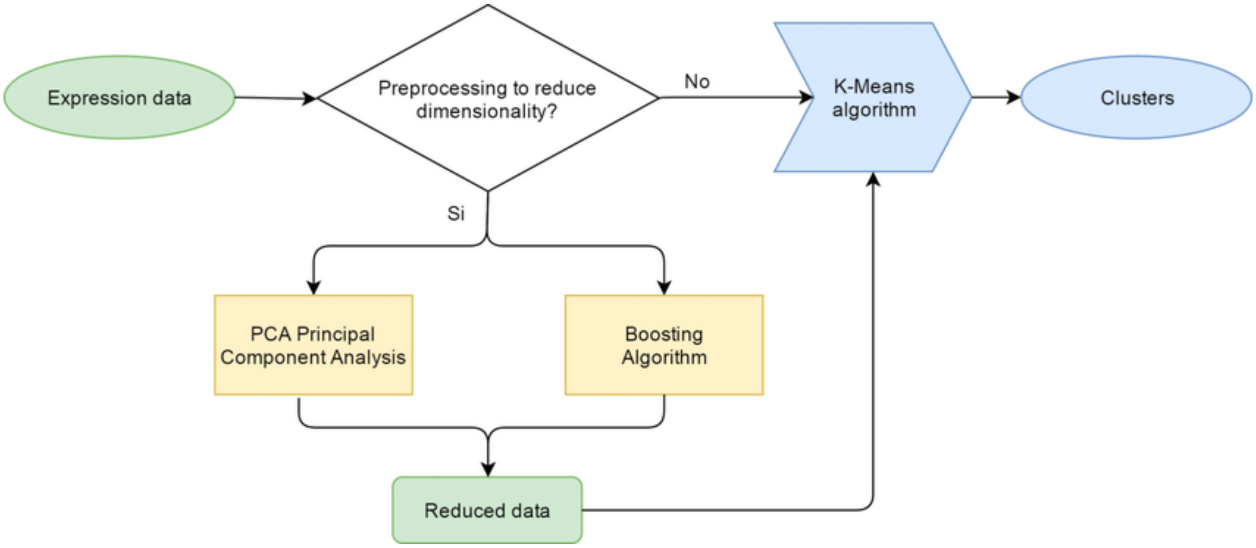


Fig.1. Dataset clustering pipeline with K-Means Algorithm

Once the process of cleaning and normalization is in place, we can move ahead to creating the cluster using different methods like k-means, hierarchical clustering etc. Once the clusters are formed, we move ahead to identify the different matrices of the different clusters. In a typical clustering technique, the method to identify the appropriate number of clusters is done by different techniques like Within Sum of Squares (WSS) or Silhouettes coefficient technique. This ensures that we consider an optimal value for the number of clusters even before moving to creating the model.

Once we have the clustered dataset, we would be able to understand the behavior and traits of different clusters with respect to their buying patterns, behavior as well as different other traits.

1.3 Project goals

In the project, we plan to use one of the popular housing dataset regarding Dubai, wherein we use the different property attribute information to bucket them into different clusters. This clustering will help us understand which are High, Medium and Low cost apartments. Our end result should be that we have clusters of different houses based on their attributes. For example, a house costing 200,000 AED and having two rooms and one bathroom compared to another one costing 100,000 AED with one room and one bathroom are tagged as High cost and Medium cost apartments based on different other factors in the dataset. Clustering techniques are widely used in the industry to obtain different groups/segmentations for any problem statement so that the relevant marketing campaigns or information can be rolled out. In this highly competitive market, it is very important that the right people get the right information for better conversions and attention from these customers.

The scope of our project involves collecting the appropriate dataset related to house attributes as well as different traits that explain the property characteristics. In our case, we will use a Dubai property dataset available on a public repository online and then identify different clusters as well as performing data analysis and different visualizations on these clusters. This would help us in understanding the properties of each cluster even after performing the clustering technique, and post-hoc analysis is very important in any clustering technique.

Having different clusters of the properties (as shown in Fig.2.) would enable us to perform data analyses on these clusters to understand the mean prices, number of rooms and other attributes of each of the clusters. Once we understand the trends of each of these clusters, we are at an advantage to promote the properties to the right cohort of users.

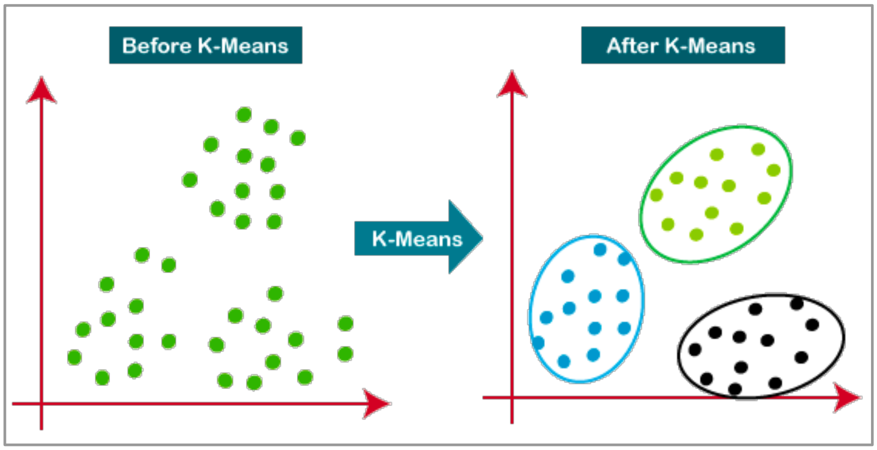


Fig.2. K-means clustering results before-after

The idea behind any clustering algorithm is to group data points based on similarities, as shown above, clusters are obtained on the right which are based on similarities in different attributes from the dataset. This is the main goal that we try to achieve from the dataset shown above as well as our explanation on the subject.

1.4 Research Questions

During the phase of the coming section, we will perform some univariate and bivariate analysis to understand more about the data points and their distribution. We will understand some key factors like -

- What are the different trends of prices of properties in Dubai?

- Which property specifications are more expensive than others and what factors determine the property prices?
- Which neighborhoods are costlier than one another and the price distribution in each neighborhood?

The above hypothesis, when answered with the help of data, can help us understand the relationship of the features that we have in the dataset which we can then use for the modeling purpose.

Chapter 2 - Literature Review

2.1 Literature Review

The changes in house prices are commonly estimated using the metric called House Price Index (HPI) and the housing prices are highly correlated with factors like location, area, population and other attributes of the property, and this necessitates the dependency on other metrics apart from the HPI to be able to predict and cluster individual housing prices. In his paper, (Truong, 2020) has research involving the application of both traditional and advanced machine learning approaches to deep-dive into differences among the many advanced models. The paper also investigates the comparison for each of these models to perform a comparative analyses amongst all of those. According to their paper (Varma, 2018), the housing market is very obscure when it comes to the underlying functionalities involved in the prices and many other factors. The housing prices are very dynamic in nature and tend to be hyped too much rather than the valuation itself. Using the real factors as data points is the main crux of their research and they aim to make evaluations based on parameters which are very basic to determine the prices of the properties at the end. They consider the weighted mean of various models to conclude on the pricing, rather than depending on a single model to have a comprehensive outlook of the predictions. This resulted in minimum error and maximum accuracy than the model results when they are used individually. Wang in 2019 (Wang, 2019) defined the nonlinear relationship that lies in the influential factors as well as the house prices and inadequate number of sample size, which might result in the poor performance of the traditional models. The daily data of the housing market is always on the rapid rise and the traditional approaches for predicting house prices lack the capacity for immense data analysis which results in low utilization of the data. In order to be able to address these problems, a deep learning approach is proposed in their paper using the TensorFlow framework. The Adam optimizer involved in Deep Learning is used in the model training phase wherein Relu function is implemented for the activation function. Once this is done, the ARIMA model is used to forecast the house price trends. This experimental result shows the individual house prices predicted by the above mentioned approach is better than the other SVR method.

Multi-level models and Artificial Neural Networks which are two of the most advanced modeling approaches have been implemented in their research approach to predicting house prices (Feng, 2015). A comparative analysis is done between the standard Hedonic Price Model as well as the above mentioned approaches to compare in terms of predictive accuracy, also the capability to capture the location information inherently contained in the data as well as their explanatory power. The models have been applied on the house prices of 2001-2013 for the Greater Bristol Area using secondary data from Land Registry, the Population Census as well as the Neighbourhood Statistics so that a national approach is involved in the model application. The results are indicative of the fact that MLM is a good model in terms of predictive accuracy providing high explanatory power, especially when the neighborhood effects are explored at multiple spatial scales.

A lot of the papers available around house price prediction are either for model optimisation or factor consideration for the modeling purpose. In another paper by Manjula R. (2017), it has been demonstrated about the various features that can be considered necessary for house price prediction accuracy. The team used various regression techniques using differing sets of features to land at lower Residual Sum of Squares error (RSS error). It is often noticed that for prediction using Regression models, the best features are required for the accuracy to be good. In different research approaches, a lot of the set of features or polynomial regression techniques are used which can make the model fit better. But a downside is that these models are highly susceptible to over-fitting and hence an approach with ridge regression is implemented to reduce the same. Their research work directs the audience towards the best techniques in the market to optimize on the accuracy results using different Machine Learning models.

This type of research on house prices was examined in Turkey using the housing market data in the Turkey economy. Being a highly profitable market in the housing industry, Turkey is sought as an investment option by many for research purposes as well as investments. In a research work by Hacievliyagil N. (2021), dynamic model averaging (DMA) was implemented to predict the monthly house price growth. The Residential Property Price Index (RPPI) was considered as the target variable while twelve other independent variables were considered for the modeling purpose. Some of the factors that impact the house price included the level of mortgages, unemployment, exchange rates and google trend index being some of the few. The researchers want to recommend that many stakeholders like researchers, consumers, and policymakers should monitor the housing market through the use of the house price index as the prices are a contributing factor to the other macroeconomic factors of the country that the housing industry resides in.

In a paper by Chen (2017), the team used the data from house prices through January 2004 and October 2016 which was then implemented on an Autoregressive Integrated Moving Average model to predict the price for the houses from November - December 2016 in various districts throughout China. The baseline was created using a popular model called LSTM to build the prediction model. The model evaluation was performed on the basis of Mean Squared Error which showed that LSTM had the best results in terms of predicting time series. This indicated that research around house prices was carried out in different corners of the world, which helped us understand the pricing trends and properties across different parts of the geography.

In our report, we plan to use one of the most popular clustering algorithms called K-Means clustering. To understand the model better we read in one of the papers (Likasa, 2003), wherein the K-Means algorithm was presented based on an incremental approach. This approach added a cluster center dynamically from the appropriate initial position of the cluster center. The team also worked on a proposal to modify this complex method of dynamic cluster generation to reduce the computational complexity which inherently affects the system load, and this would be done without impacting the quality of the model approach. Related studies around the modeling approach solidifies our approach in the modeling section and it is imperative to study some related work in the same field.

During a study conducted (Morgan, 2021), it was observed that the average sale price of properties in Dubai dropped by 7% during the time 2019 Q3 and 2020 Q3. The average rents also dropped by 10% during that time period. Since the price depreciated between the period 2019 and 2020, during that time a lot of tenants decided to move to bigger units with bigger amenities which had become affordable due to the backdrop. During the Expo 2020, an estimated total of 24,000-25,000 residential units had been handed over in the first couple of months of 2020. The transaction volume also reduced by 16% YTD September 2020 compared to the same period in 2019. To work on a problem related to an industry, it is important the we identify the trends and key imperatives within the same domain so as to have a better understanding of the yearly, monthly or weekly trends overall.

In the course of the reporting and literature review process, we have gone through numerous research work performed in different cities for the house price trends to understand the domain better. In a research conducted by (Lim, 2016), the property prices were forecasted for Singapore to help buyers make informed decisions on the properties around the city. The paper aligns with the use of two algorithms which were used to predict the Singapore housing market, and ultimately compare the performances of both these models Artificial Neural Network (ANN) and Autoregressive Integrated Moving Average (ARIMA). Out of both these models, the better one was used to also predict the House Price Index or the HPI (which has been termed as CPI in the paper and essentially means the same thing). The evaluation was done on the basis of mean square error (MSE) for the ANN model which showed superiority compared to the other models in terms of performance. On a similar note, a study conducted by Ahmed (Ahmed, 2020) indicated the relationship between selected macroeconomic indicators that impacted the real estate industry of Dubai. The team used a time series regression and also monthly level dataset from 2008 to 2017 to study the price trends and impact factors for the same. It was determined that there was a negative impact of the exchange rate and oil prices on the house price index along with a direct proportional impact on the inflation and money supply in the market.

Now, coming back to the UAE property pricing and the real estate industry, we study about a paper which aimed to explore the house prices in Dubai and the various factors that determine the prices of these properties, thereby, addressing the intrinsic and extrinsic factors which affect the prices of the properties in the long run. It was identified in Zaabi's work (Zaabi, 2019) that there are several factors which impacted the house prices like the house size, availability of bathrooms, bedrooms, waterfronts, and pools. During the course of the research it was determined that several extrinsic factors like policy making processes, and many others impacted the house prices during the time frame and the data for the entire study has been used based on information provided by the Dubai land Department.

An interesting study carried out by Mbazia (Mbazia, 2017) tried to determine the relationship between housing and money demand empirically for countries from the Middle East and North Africa. The plan was to use quarterly data from 2007 Q3 and ending year 2014 Q4 including the house prices as a feature to determine the house price development over the years. A technique which is termed as Pool Mean Group Estimation was used during the course which estimated that there is a dynamic relationship with money demand both in the short term and long term. It was estimated that the higher the prices of the houses, there is an increase in the money demand for both long-term and short-term requirements. This very well indicated the impact that the housing market had on the monetary policies within the Middle Eastern and North African countries. It was also identified that the structural features of the properties had a role to play in the entire ecosystem of money demand and house prices overall. This is a key finding to be able to understand the external effects that the real estate industry can have and it is quite important to understand the various factors or impact measures in-depth.

2.2 Takeaways from Literature Review

Through the above exploration of different research approaches, we learn a lot about different technological approaches taken to solve some of the problems around house price prediction and machine learning. Understanding different approaches in depth prepares one for their own research or problem statement to enhance the subject matter knowledge. Some of the key takeaways from the above literature review are as follows -

1. The housing market is very dynamic, with a lot of external factors contributing to the pricing. Hence, robust research and modeling techniques need to be considered while exploring the real estate data and market so that the results and findings are accurate and to the point (Coskun, 2020)
2. With the increase in the demand supply curve for real estate in UAE, the need to understand this subject also increases. Having gone through numerous prediction capabilities and approaches in our literature survey, we understand various implementation approaches to be able to solve such problems in the future (Abbas, 2022)
3. Prices of property are dependent on various factors, and the above research materials have not failed to consider these factors and explain about them in-depth. And reading the research materials have successfully explained to us the need of such robust research background when conducting such projects

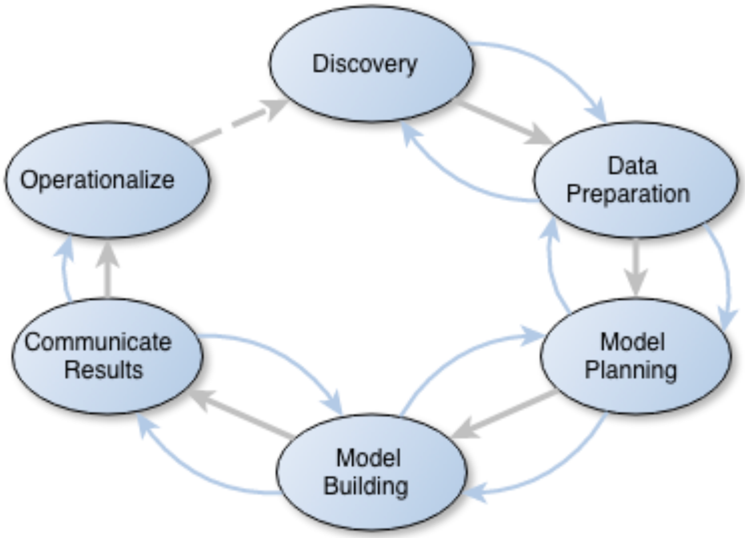
The above summarisation has been done on the basis of our literature review and background information of the subject in hand. Through some of these understandings, we plan to use and implement the same for our research which has prepared us better to take the project further.

Chapter 3 - Research Methodology

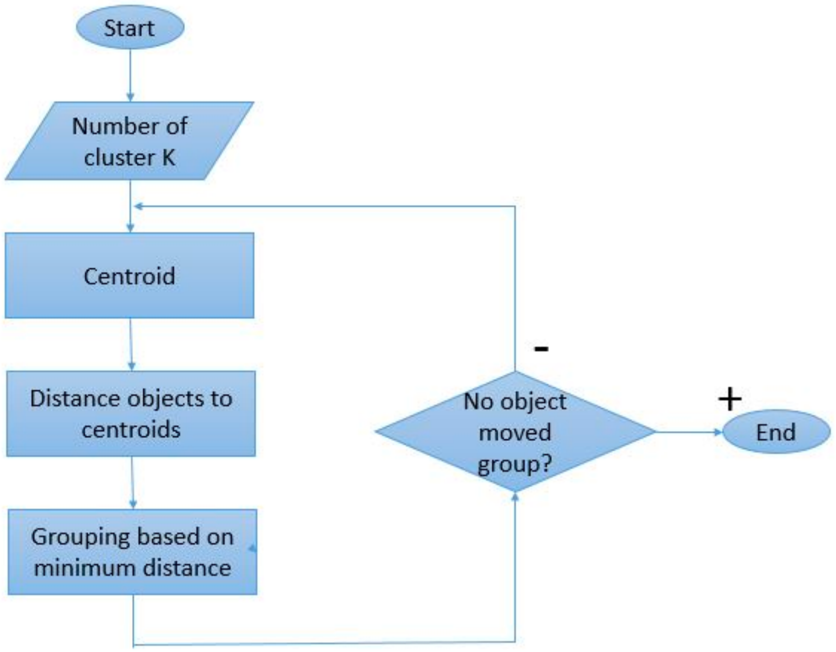
3.1 Methodology

Here in the project, we start with obtaining the appropriate dataset of apartment descriptions and attributes. There are different steps to an analytics project, and we can discuss about it as below -

- The first step to any data analytics project is preparing the dataset or collecting the data for further analyses.
- Now, when we have the dataset ready the next step is to prepare the data, and this can be through cleaning, manipulation or even aggregations as per our need. There might be a lot of inconsistencies in the dataset like missing values, outliers, data type errors etc. which may deter us from creating the models appropriately. When we treat the dataset, it is a clean dataset which can further be used for aggregations and manipulations.
- Once we have the dataset ready, we can use it to build our model or in the case of the project, we can perform the clustering technique using either k-means or hierarchical clustering. Before building any clustering model, it is very crucial that the optimal number of clusters are determined before doing it.



Once we have the prepared dataset for clustering, we can move ahead to the next steps. The below figure gives a gist about the steps involved in a K-Means clustering algorithm. The first step is to identify the optimal number of clusters for the model, and once that is done we move ahead to identify the centroid for each cluster. The centroid formation then allows the data points to be grouped for each centroid and distance from the same. This process is repeated till all the data points are appropriately grouped to each of these clusters.


Start
Number of cluster K
Centroid
Distance objects to centroids
Grouping based on minimum distance
No object moved group?
-
+
End

There are different KPIs involved to understand the results of our clustering method and this can be achieved through data analysis steps. We finally move ahead to operationalise the model and present our findings and results to the relevant stakeholders.

This is the summary of the project implementation lifecycle which we plan to use in this project, and after having the appropriate cohort of houses we can identify the right marketing strategies for the customers which would help us have better conversion rates as end results. Better conversion rates mean better profits for the company.

3.2 Deliverables

The project deliverables would contain different aspects of our analyses. In this section, we discuss the different deliverables involved with the project submission.

1. Report containing the entire analysis along with findings and insights to understand the data and problem statement better
2. Presentation containing a gist and summarization of the entire process along with the relevant results to walk the audience through our project in brief
3. Clustering results based on the unsupervised machine learning technique that we use

The above deliverables are the primary materials that would be submitted for the project along with production ready code base, which can be implemented on similar types of datasets.

3.3 Tools and Data Used

In the due course of our report, we will be using one of the popular datasets from Kaggle which contains the prices of properties in Dubai, and these properties belong to different neighborhoods. In this section, we would like to depict the dataset to the readers along with the description of the tools that we plan to use within the process. The dataset is constructed of 1000 rows of property information with 38 columns explaining each property. Below we have the data dictionary describing each of the features that we have in the dataset. Each row of the data defines a particular property which we will try to visualize and model using different tools and techniques to be able to cluster the properties based on similarity.

Column Name	Description
id	Unique identifier of the property
neighborhood	The neighborhood where the property is located
latitude	The latitude of the property location
longitude	The longitude of the property location
price	The price of the property in AED
size in sqft	The size of the property
price per sqft	The price per square feet area of the property
no of bedrooms	The number of bedrooms available
no of bathrooms	The number of bathrooms available
maid rooms	Flag for maid room availability
unfurnished	Flag for furnish status of the property
balcony	Flag for balcony availability
barbeque area	Flag for barbeque area availability
built in wardrobes	Flag for built in wardrobe availability
central ac	Flag for central ac availability

childrens play area	Flag for children’s play area availability
children's pool	Flag for children’s pool availability
concierge	Flag for concierge availability
covered parking	Flag for covered parking availability
kitchen appliances	Flag for kitchen appliances availability
lobby in building	Flag for building lobby availability
maid service	Flag for maid service availability
networked	Flag for network connection availability
pets allowed	Flag if pets are allowed into the property
private garden	Flag if a private garden is available
private gym	Flag if a private gym is available
private jacuzzi	Flag if a private jacuzzi is available
private pool	Flag if a private pool is available
security	Flag if security is available
shared gym	Flag for shared gym availability
shared pool	Flag for shared pool availability
shared spa	Flag for shared spa availability
study	Flag for the availability of a study
vastu compliant	Flag if the property is vastu compliant
view of landmark	Flag if a view of landmark is available
view of water	Flag if a view of water body is available
walk in closet	Flag for walk in closet availability
quality	Categorical feature describing the quality of the property as ultra high, high, medium, low

Table.1. Data dictionary of the dataset

The above table helps readers understand a bit more about the dataset before we delve into the EDA and Modeling aspect of things. In this section we would also like to explain about the tools that we plan to use within the scope of our project.

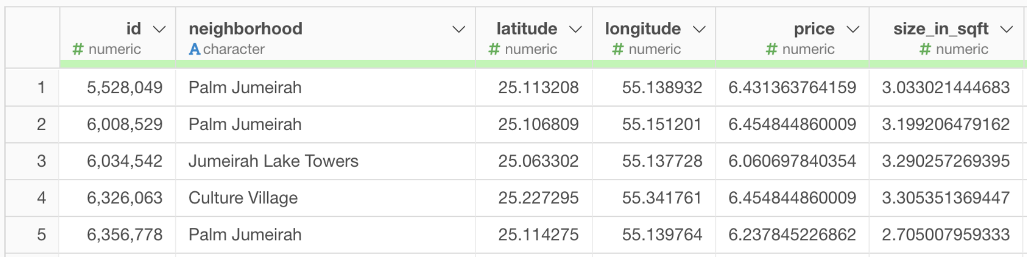
Jupyter Notebook - It is an open-source web application which allows the users to create and share documents that contain codes, markdown as well as visualizations and other machine learning implementations. It can be used for data cleaning, data manipulation and visualization using different coding languages like Python, R etc.

Exploratory - This is a proprietary tool based on R language and this can be used to view, manipulate, visualize and model with datasets. In this report, we use this tool to perform quick data manipulations and analysis along with clustering to have a better holistic understanding of the data.

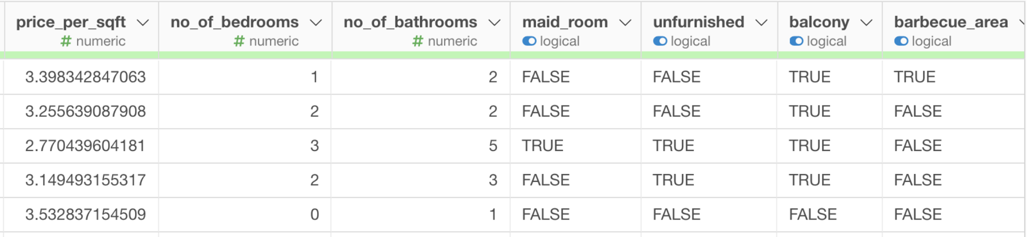
Chapter 4 - Analysis and Clustering

4.1 Data Preparation

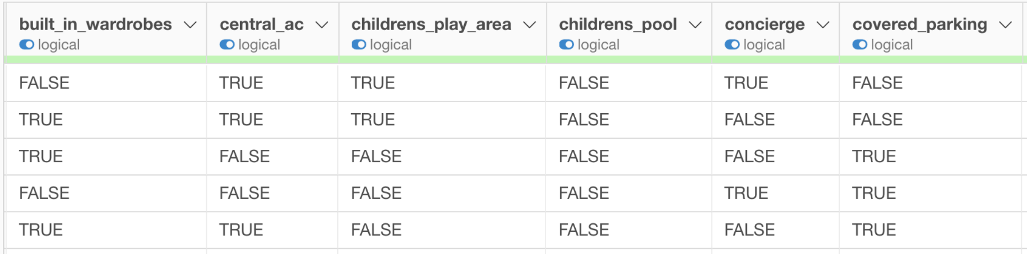
We will use this section of the report to explain some of the data preparation and cleaning processes that we followed to have standardized data. We start off with showing the top few rows of the data for understanding.



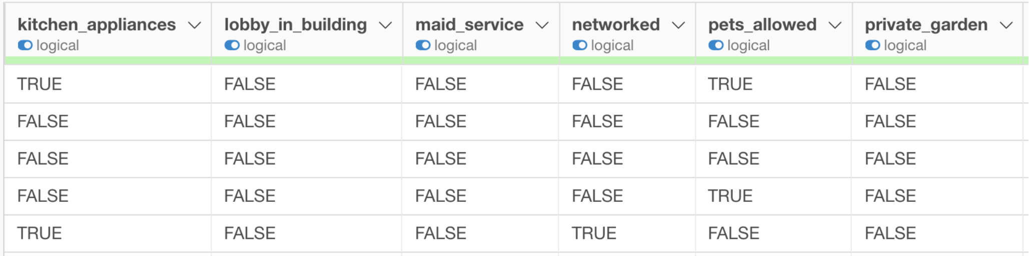
	id	neighborhood	latitude	longitude	price	size_in_sqft
	# numeric	A character	# numeric	# numeric	# numeric	# numeric
1	5,528,049	Palm Jumeirah	25.113208	55.138932	6.431363764159	3.033021444683
2	6,008,529	Palm Jumeirah	25.106809	55.151201	6.454844860009	3.199206479162
3	6,034,542	Jumeirah Lake Towers	25.063302	55.137728	6.060697840354	3.290257269395
4	6,326,063	Culture Village	25.227295	55.341761	6.454844860009	3.305351369447
5	6,356,778	Palm Jumeirah	25.114275	55.139764	6.237845226862	2.705007959333



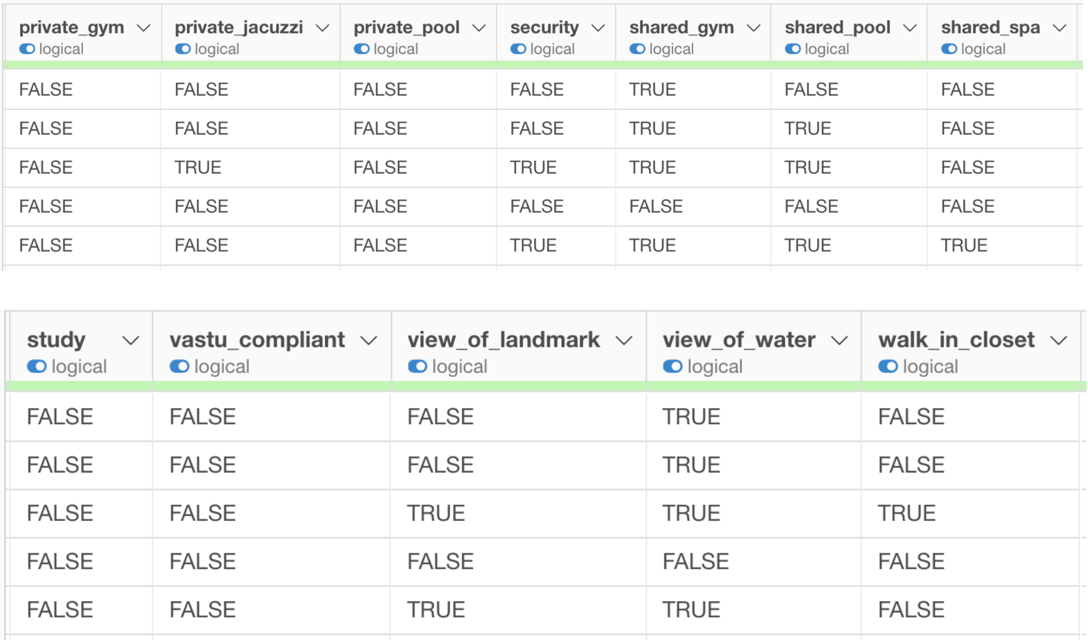
price_per_sqft	no_of_bedrooms	no_of_bathrooms	maid_room	unfurnished	balcony	barbecue_area
# numeric	# numeric	# numeric	logical	logical	logical	logical
3.398342847063	1	2	FALSE	FALSE	TRUE	TRUE
3.255639087908	2	2	FALSE	FALSE	TRUE	FALSE
2.770439604181	3	5	TRUE	TRUE	TRUE	FALSE
3.149493155317	2	3	FALSE	TRUE	TRUE	FALSE
3.532837154509	0	1	FALSE	FALSE	FALSE	FALSE



built_in_wardrobes	central_ac	childrens_play_area	childrens_pool	concierge	covered_parking
logical	logical	logical	logical	logical	logical
FALSE	TRUE	TRUE	FALSE	TRUE	FALSE
TRUE	TRUE	TRUE	FALSE	FALSE	FALSE
TRUE	FALSE	FALSE	FALSE	FALSE	TRUE
FALSE	FALSE	FALSE	FALSE	TRUE	TRUE
TRUE	TRUE	FALSE	FALSE	FALSE	TRUE



kitchen_appliances	lobby_in_building	maid_service	networked	pets_allowed	private_garden
logical	logical	logical	logical	logical	logical
TRUE	FALSE	FALSE	FALSE	TRUE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	TRUE	FALSE
TRUE	FALSE	FALSE	TRUE	FALSE	FALSE



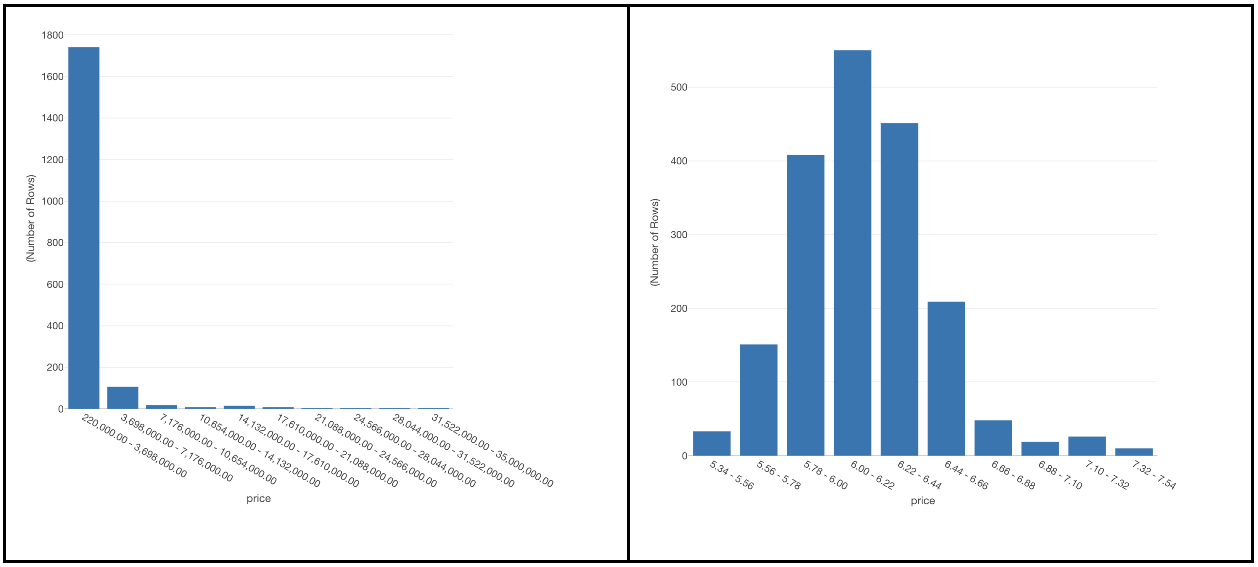
private_gym logical	private_jacuzzi logical	private_pool logical	security logical	shared_gym logical	shared_pool logical	shared_spa logical
FALSE	FALSE	FALSE	FALSE	TRUE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	TRUE	TRUE	FALSE
FALSE	TRUE	FALSE	TRUE	TRUE	TRUE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	TRUE	TRUE	TRUE	TRUE

study logical	vastu_compliant logical	view_of_landmark logical	view_of_water logical	walk_in_closet logical
FALSE	FALSE	FALSE	TRUE	FALSE
FALSE	FALSE	FALSE	TRUE	FALSE
FALSE	FALSE	TRUE	TRUE	TRUE
FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	TRUE	TRUE	FALSE

Table.2. Data view of the top n rows

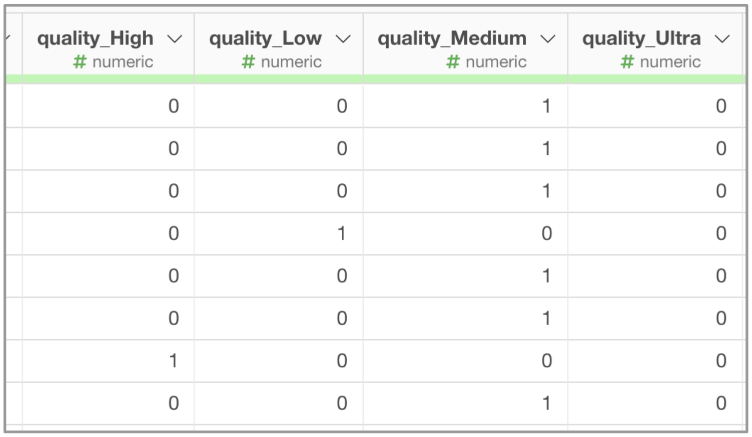
Log Transformation

For some of the features with extreme values (especially the property price and size of the property), we used log10 transformation to normalize the data and have it in a standardized manner. Below we show the distribution of the price feature, and on the left the price variable is right skewed. After doing the log10 transformation we have a normal distribution of the data.



One-hot encoding

This is required for the dimension reduction process, and has been carried out for the quality feature in the data. The below table shows some of the top rows of the result of one hot encoding of the quality feature which contains levels like Low, Medium, High and Ultra.



quality_High # numeric	quality_Low # numeric	quality_Medium # numeric	quality_Ultra # numeric
0	0	1	0
0	0	1	0
0	0	1	0
0	1	0	0
0	0	1	0
0	0	1	0
1	0	0	0
0	0	1	0

Table.3. One-hot encoding of quality feature

4.2 Exploratory Data Analysis

Exploratory Data Analysis is the process of leveraging data to draw insights and findings through visualizations, they can be univariate analyses or bivariate analyses. In this section, we plan to use the same technique to plot different charts and visualize the data to understand the factors and their relationships with one another. For example, if the price of the house is high what is the reason behind it? We want to be able to compare the prices of properties in different neighborhoods in Dubai to compare the cheap properties with the expensive ones. This will also help us understand if the similarity between features are very high and avoid high correlation problems for any form of modeling techniques. (Brownlee, 2020) Different types of distance based correlation metrics and visuals will help us understand the features in depth. There is also the method of univariate and multivariate analyses that would help us understand feature relations with one another for the Dubai property dataset.

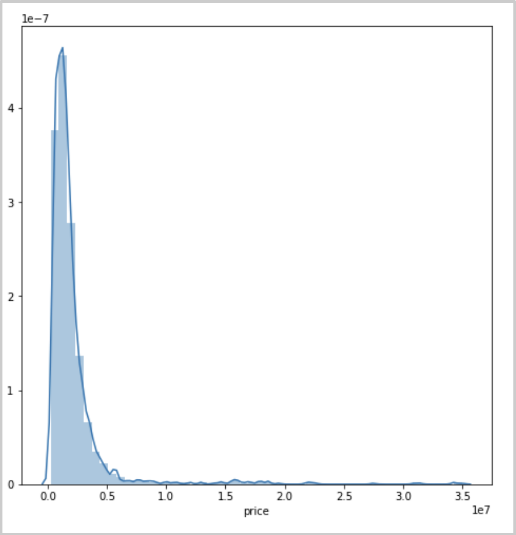


Fig.3. Histogram of price feature (skewed)

First we plot the first variable, i.e., the house prices using python and see that the variable suffers from Skewness and Kurtosis (which would impact our analysis and models in the later stage). The Skewness value is determined to be 6.14 and the Kurtosis is 48.8 and we can observe that the data suffers from high skewness and shows a long-tail positive skewness. This uneven distribution can harm the analysis as well as clustering model in the later stage of our process. We will now clean the outliers and then plot the variable again to have a better representation of the price variable.

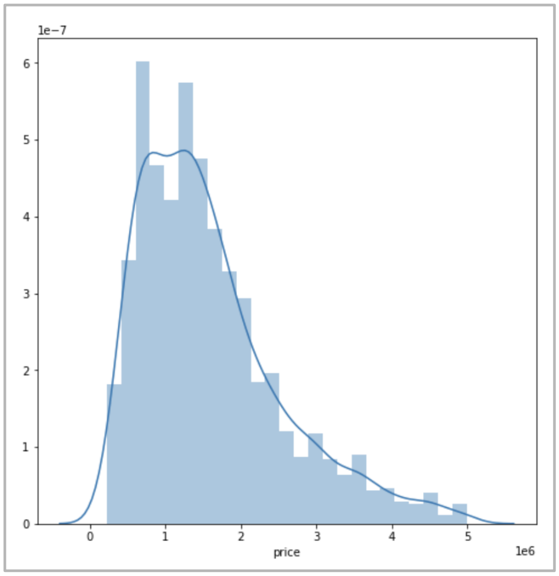


Fig.4. Histogram of price feature (normalized)

Based on the above histogram representation, after removing the outlier values, the skewness and kurtosis values are both 1.1 approximately. Removing outliers and skewness from the dataset helps in a normalized representation of the features and also prevents any form of biases during the modeling phase. Let's say we have a property with a price at 250,000 AED while another property is at 100,000,000 AED. If both of these property prices are input into a learning model, the model would interpret these extreme values to a globalized mean and provide clusters accordingly with extreme data point understanding. Hence, we follow the process of normalizing and standardizing some of the key features within the dataset.

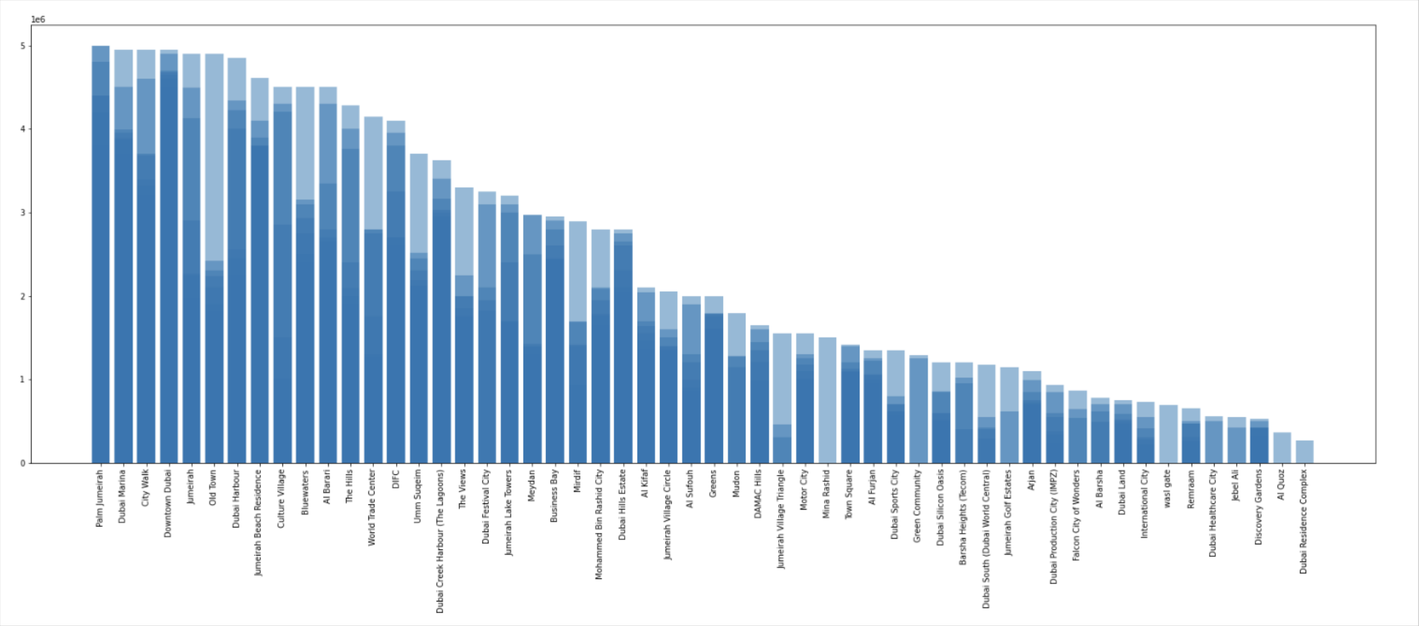


Fig.5. Plot of the property neighborhoods and their price distribution

In the above visualization of the property prices spread across different neighborhoods, we see that Palm Jumeirah, Dubai Marina and CityWalk are some of the most expensive areas/neighborhoods in terms of property prices. While Discovery Gardens, Al Quoz and Dubai Residential Complex are the cheapest based on the data point distribution above. This helps us understand the neighborhoods that have high prices of properties and help us distinguish between the expensive vs the cheap neighborhoods.(Alrawi, 2019)

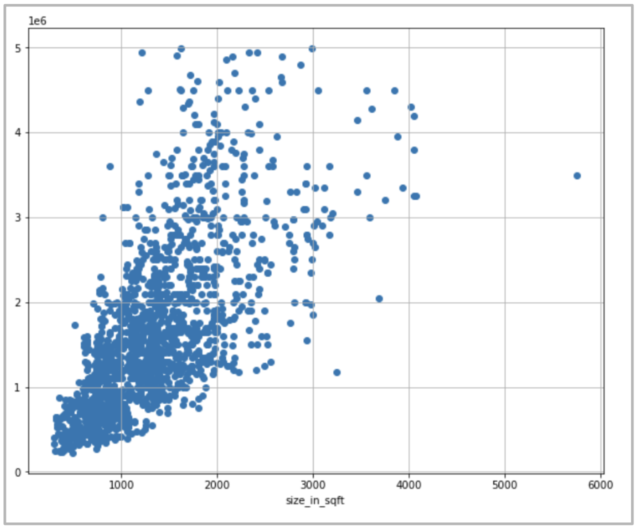


Fig.6. Scatter Plot of the size in sq ft feature

Now, we plot the price of the properties with the area of the property to determine if there is any relationship between both the variables. We see that they have a linear relationship between both. As the area of the property increases, the price of the property also increases obviously. This means that if we plan to buy properties the size of the property is an important factor to be considered. We would also want to understand the other features of the houses to understand the impact or pieces along with other properties of the houses.

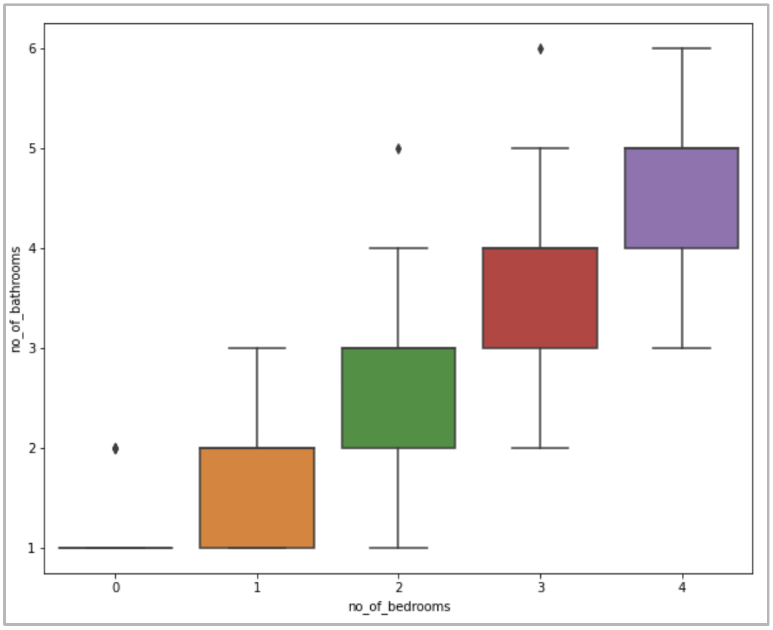


Fig.7. Box plot of number of bedrooms and bathrooms

In the above box plot, we plot the number of bedrooms with the number of bathrooms to understand the property dimensions with respect to the bathrooms and bedrooms numbers. We observe that higher the number of bedrooms in the property, the more bathrooms there are obviously. This also indirectly indicates that if the property sizes increase, so does the number of bedrooms and bathrooms. To summarize the above finding, if a buyer wants to have more bedrooms and bathrooms, they would have to spend more on the property. For a cheap property with more bathrooms and bedrooms, they might have to explore neighborhoods with less property prices in general. Now, we will also try to understand the relationship of all the features in the dataset with one another through a very popular method of plotting called correlation plot which helps us understand the relationship between two features in terms of being directly or indirectly proportional to one another.

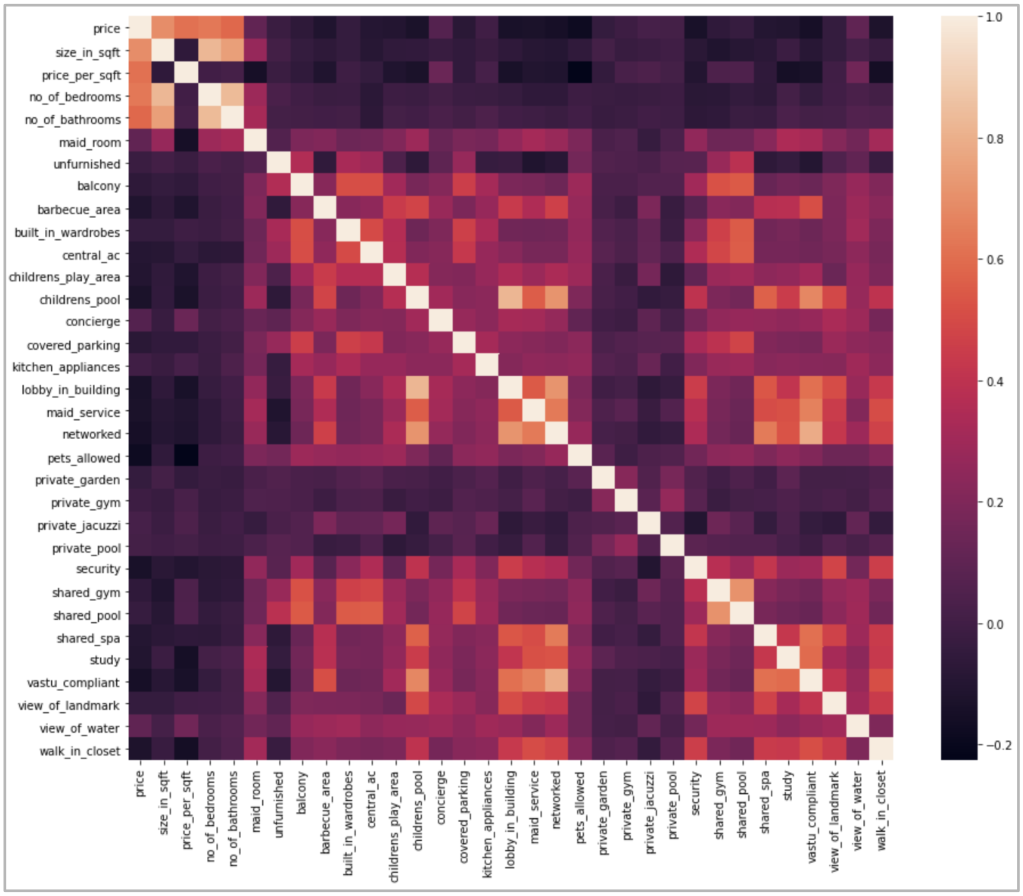


Fig.8. Correlation plot of all the features in the data

In the above correlation plot of all the features, we observe some of the combination of features like price, size in sqft, price per sqft, number of bedrooms and bathrooms, maid room etc. being highly correlated to one another. In the above chart, a value tending towards 1 indicates high positive correlation with one another while a value tending towards -1 indicates a negative correlation. With the help of the above analysis and insights, it is easier for us to pick the required features to use in the clustering algorithm to find the different clusters and their properties.

4.3 Model Description

In the below section, we will describe some of the models that we would implement in the due course of the report. These different modeling techniques help us solve the problem in hand, one of which was used to identify factor significance in our data while the other was used to cluster our dataset to understand the different segments of the properties in Dubai and be able to describe and differentiate each segment of properties (Wang, 2019).

Random Forest

The random forest model on the other hand is an extension of the decision tree, whereby instead of fitting a single decision tree, multiple decision trees are fit on random samples from the training data hence the name Forest. The number of samples (number of trees) can be adjusted although the tricky part is that the range is too wide to come up with a reasonable search grid. Using this model, we identified the factor significance of the different features within the dataset

K-Mean Clustering

This method is a vector quantization method which aims to partition different observations (in our case n) into various clusters k. In these clusters, the observations belong to the cluster with the nearest mean, serving as a view of the overall cluster.

4.4 Modeling

In this section of the project, we use the modeling techniques to determine the key drivers of the property price trends and their characteristics to be able to cluster them into segments based on similar traits. Before doing that, we will also use different models to determine the factor significance of the features in the data. In the below plot, we see that size in square feet is the most important factor determined by Random Forest mode. (Truong, 2020) This means that this factor plays a very key role in determining the property characteristics of Dubai.

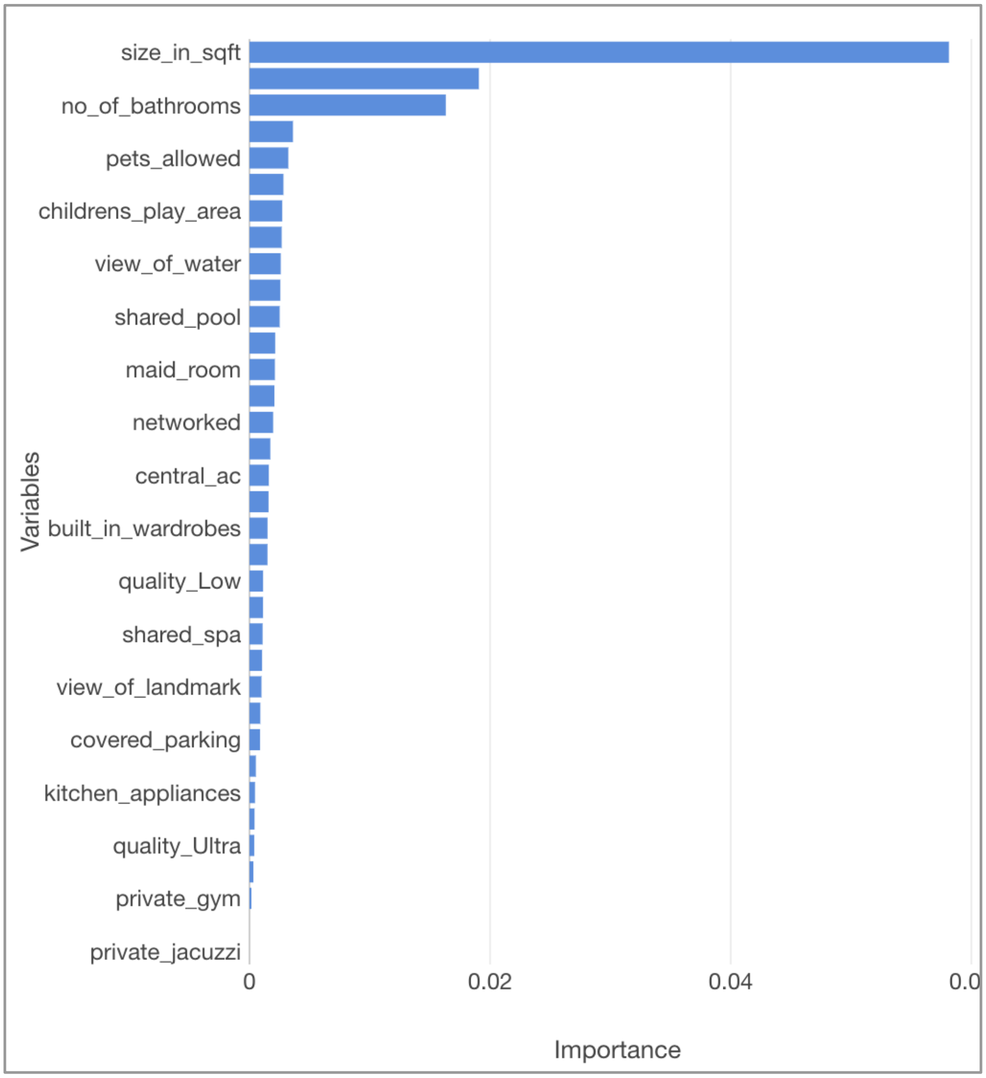


Fig.9. Factor significance plot from random forest

Moreover, we also observe that some other factors are also important along with size in square feet of the property like number of bedrooms and bathrooms, if security is available in the property along with the restrictions on pets. In the next step, we will implement a distance calculation to determine the relative nessness of each feature with one another. In the context of our experiment we will use the Euclidean distance calculation method to determine the same.(Madhuri, 2019)

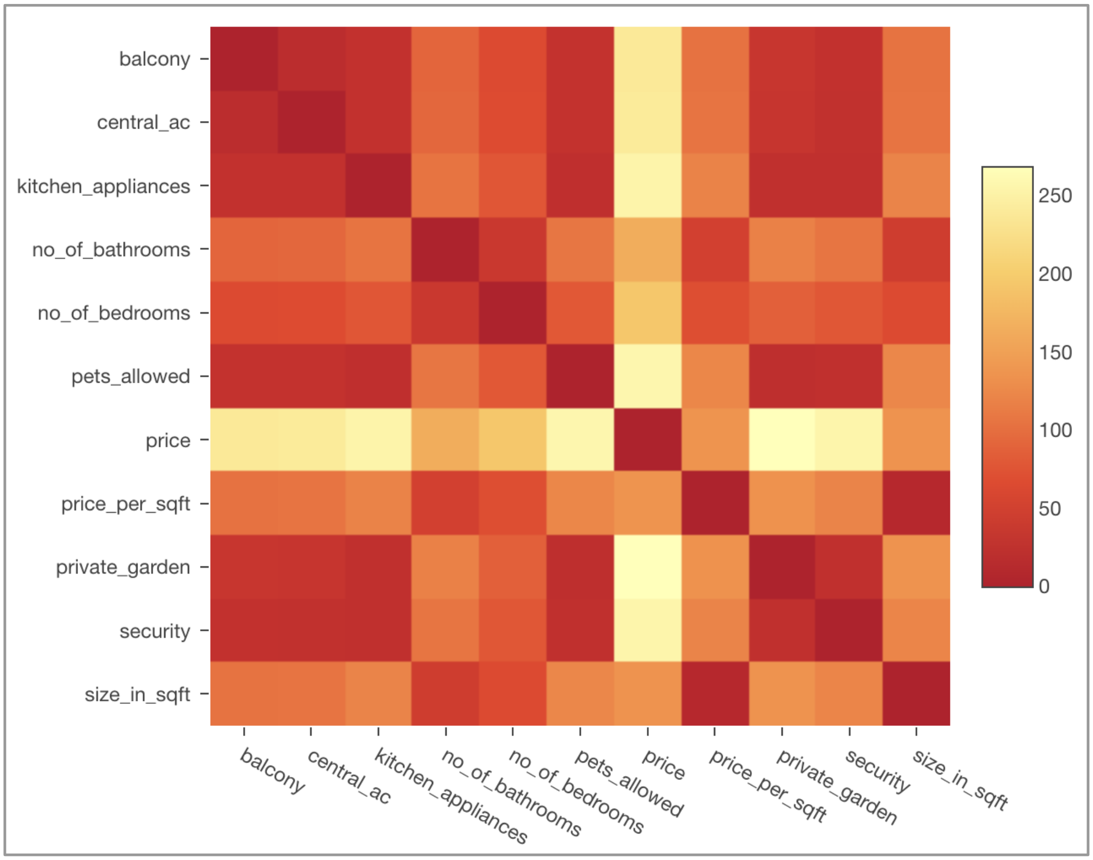
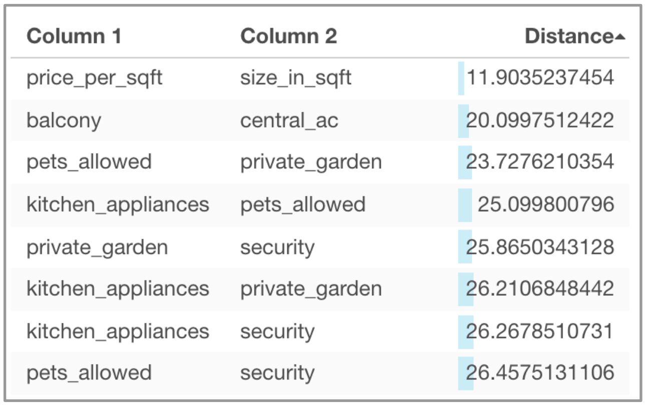


Fig.10. Correlation plot based on Euclidean distance

The above correlation plot essentially indicates that if two features are similar to each other in terms of the Euclidean distance, the value tends towards 0. We see that price per square feet and size in square feet have a very low Euclidean distance with each other.(Jason, 2020)



Column 1	Column 2	Distance▴
price_per_sqft	size_in_sqft	11.9035237454
balcony	central_ac	20.0997512422
pets_allowed	private_garden	23.7276210354
kitchen_appliances	pets_allowed	25.099800796
private_garden	security	25.8650343128
kitchen_appliances	private_garden	26.2106848442
kitchen_appliances	security	26.2678510731
pets_allowed	security	26.4575131106

Table.4. Euclidean distance based feature pairs

The above table shows the distance between some of the features based on low to high. These combinations of features explain each other better than others and hence the Euclidean distance is determined to be very low. Based on their similarity index, we also see the features which are related to one another using the cluster plot below.

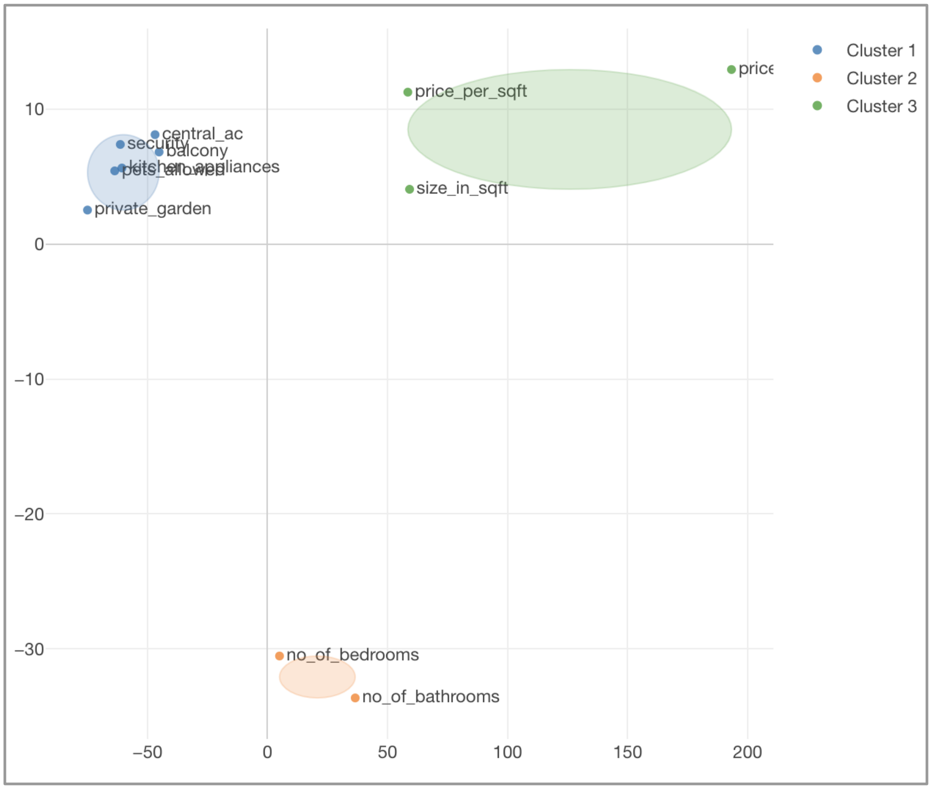


Fig.11. Similarity plot of features based on similarity index

In the subsequent sections, we will implement K-Means clustering with cluster number 5 and then study each of the cluster of properties and their trends based on different features that we have in the database. The below representation is a biplot of PCA components derived from the K-Means clustering algorithm. (Ngo, 2018) PCA captures the level of variation in the dataset and helps explain the similarity in the data points.

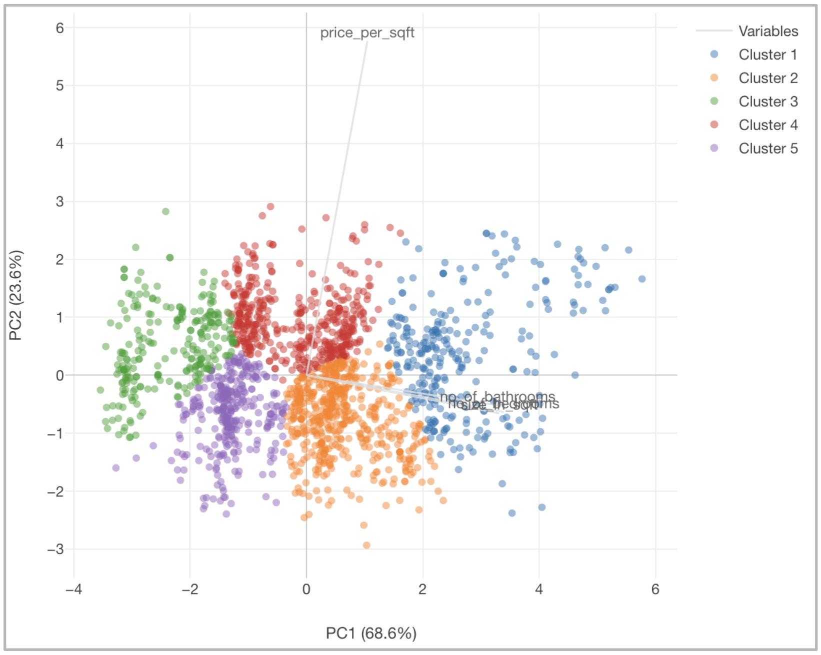
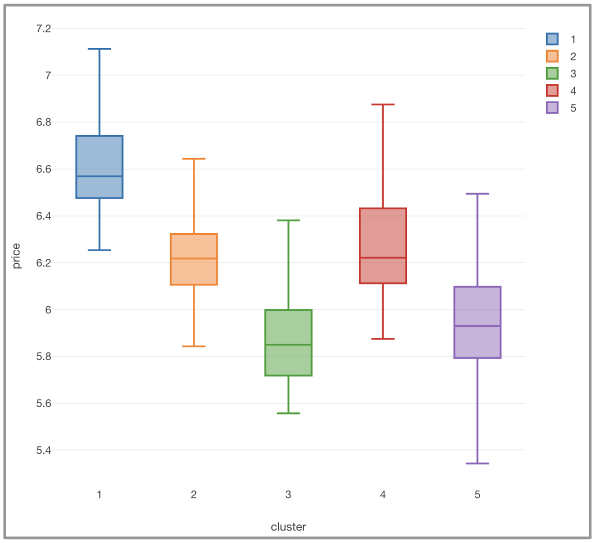


Fig.12.PCA plot from the K-means clustering

We will now analyze the clusters and various properties to understand the derivations of the K-Means clustering technique implemented on the properties of Dubai. For the 5 cluster of properties that we obtained, we see that cluster #1 has the most expensive properties with a higher average for property prices, followed by cluster #4 and cluster #2. It is observed that cluster #3 has the cheapest average of property prices comparatively. In the next plot, we draw the scatterplot of the clusters based on the size of the apartment and the price of the property, and our previous observation has been bolstered that cluster #3 are the cheap and the small sized properties evidently.



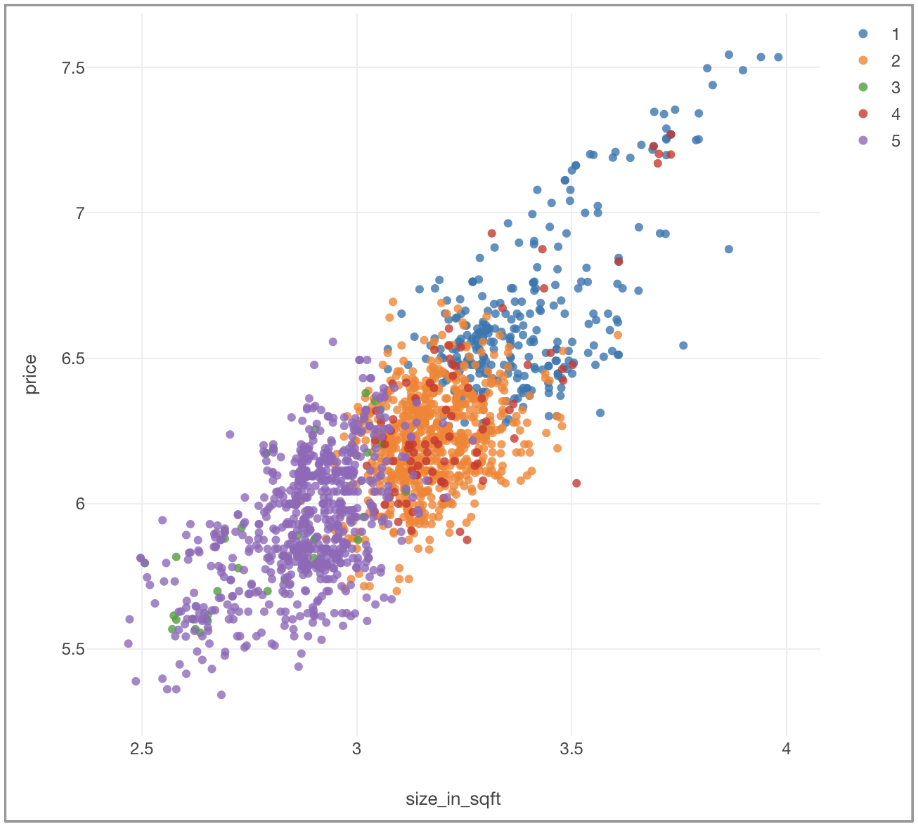


Fig.13. Scatterplot between size and price of apartment with cluster groups

From the above steps, we were able to clean and prepare a dataset to solve a problem statement. We have identified a problem statement that has been trending in the market, i.e., clustering/classification of properties in Dubai based on similarity index and various other factors. Moreover, we were also able to perform many data preparation techniques like one hot encoding, log transformation for normalizing features within the dataset etc.

Chapter 6 - Conclusion

6.1 Conclusion

The learnings of the entire project were manifold starting from defining the problem statement to solving it using data analytics and machine learning models in depth. We used different techniques to understand the business problem at hand and also researched various related work to extend our knowledge in the problem statement. We then used a dataset from Kaggle to answer various questions based on our findings and insights from the dataset. Through the course of the project, we are now better prepared to identify real world problem areas and identify the best dataset to tackle the problem independently. Not only that, we are also able to report on our research for our readers to be able to deliver them new ideas and knowledge about the housing market in Dubai. This will help our readers understand the pain points and opportunity areas of the property market in Dubai which is also one of the most booming industries in the region.

6.2 Future Work

For future work and reference, we have other aspects of the problem area that can be tackled and solved using Data Analytics and Machine Learning. For starters, we can extend our problem to use an even more extensive dataset that has more data points to explain a property characteristic better. Moreover, other data points like buyer characteristics, transaction history and many others would help us approach the problem from more angles and be able to identify and solve more problem areas within the real estate domain.

A second scope of future work is using even more modeling approaches for predictive capabilities. The research can be extended to use even more complex models like Neural Networks and Artificial Intelligence (Hardesty, 2017) which could help us predict the property prices in the future instead of clustering for similarity index of the properties.

The above ways are some of the ideas that can be thought of for future work, but there are plenty of research work done in this problem area and industry which can be studied and solved in different approaches.

Bibliography

[1] Ahmed, Farhan & Maheshwari, Sandhia & Mirani, Sajid. (2020). Dubai House Prices and Macroeconomic Fluctuations: A Time Series Analysis. 13. 61-73.

[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] 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

[4] 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

[5] Varma, A., Sarma, A., Doshi, S., & Nair, R. (2018). House price prediction using Machine Learning and Neural Networks. *2018 Second International Conference on Inventive Communication and Computational Technologies (ICICCT)*. https://doi.org/10.1109/icicct.2018.8473231

[6] Wang, F., Zou, Y., Zhang, H., & Shi, H. (2019). House price prediction approach based on Deep Learning and Arima model. *2019 IEEE 7th International Conference on Computer Science and Network Technology (ICCSNT)*. https://doi.org/10.1109/iccsnt47585.2019.8962443

[7] Feng, Y., & Jones, K. (2015). Comparing multilevel modelling and artificial neural networks in house price prediction. *2015 2nd IEEE International Conference on Spatial Data Mining and Geographical Knowledge Services (ICSDM)*. https://doi.org/10.1109/icsdm.2015.7298035

[8] 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

[9] Lim W. T., Wang L. and Wang Y. 2016 Singapore Housing Price Prediction Using Neural Networks Int. Conf. Nat. Comput. Fuzzy Syst. Knowl. Discov. 12 518-522

[10] Yener Coskun, Unal Seven, H. Murat Ertugrul & Ali Alp (2020) Housing price dynamics and bubble risk: the case of Turkey, Housing Studies, 35:1, 50-86, DOI: 10.1080/02673037.2017.1363378

[11] Hacievliyagil N., Drachal K, Eksi I. (2021). *Economies, MDPI.* https://doi.org/10.3390/economies10030064

[12] Linh Ngo, Principal Component Analysis Explained simply, Bio Turing, 14 June 2018, https://blog.bioturing.com/2018/06/14/principal-component-analysis-explained-simply/

[13] 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

[14] Omar Alrawi, Safak Bayram, Sami G., Al-Ghamdi, Muammer Koc, (2019), 14 October 2019, High-Resolution Household Load Profiling and Evaluation of Rooftop PV Systems in Selected Houses in Qatar, MDPI, https://doi.org/10.3390/en12203876

[15] 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/ICCSNT47585.2019.8962443

[16] Chen X., Wei L., Xu J., (2017) House Price Prediction Using LSTM, Arxiv, Cornell University, https://doi.org/10.48550/arxiv.1709.08432

[17] Aristidis Likasa, Nikos Vlassis, Jakob J. Verbeek, February 2003, The global k-means clustering algorithm, Elsevier, https://doi.org/10.1016/S0031-3203(02)00060-2

[18] Jason Brownlee (2020), Introduction to Dimensionality Reduction for Machine Learning, May 6 2020, https://machinelearningmastery.com/dimensionality-reduction-for-machine-learning/

[19] Larry Hardesty (2017), Explained: Neural networks, MIT News Office, https://news.mit.edu/2017/explained-neural-networks-deep-learning-0414

[20] 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

[21] 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

[22] 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

[23] 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

[24] Al Zaabi, Y. H., & Bekele, G. (2019). The House Price Dynamics and the Macro-Economy: an Empirical Perspective of the UAE. International Conference on Advances in Business and Law (ICABL), 2(1), 28–33. https://doi.org/10.30585/icabml-cp.v2i1.209

[25] Mbazia, N. & Djelassi, M. (2019). Housing Prices and Money Demand: Empirical Evidence in Selected MENA Countries. Review of Middle East Economics and Finance, 15(1), 20170034. https://doi.org/10.1515/rmeef-2017-0034