Price Prediction and Valuation Using Data Mining in Dubai Real Estate Market

Abdulla AlHathboor
ama1463@rit.edu
PRICE PREDICTION AND VALUATION USING DATA MINING IN DUBAI REAL ESTATE MARKET

by

Abdulla AlHathboor

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

DEC 2020
Master of Science in Professional Studies:

Data Analytics

Graduate Capstone Approval

Student Name: Abdulla AlHathboor - 355009810

Graduate Capstone Title: Price Prediction and Valuation Using Data Mining in Dubai Real Estate Market

Graduate Capstone Committee:

Name: Dr. Sanjay Modak

Chair of committee

Name: Dr. Ehsan Warriach

Member of committee
ACKNOWLEDGMENTS

I would like to thank Dr. Ehsan Warriach, my mentor, who has supported me through the capstone project by extending his unlimited support in helping me realize the final outcome of the project. I would also like to thank Dr. Sanjay Modak for his support and valuable feedback prior the start of the project.
ABSTRACT

The purpose of this study is to find out the impact of data mining in predicting prices and values of real estate units in the Dubai real estate market. This market has always been one of the biggest markets in the economy of any nation worldwide and has always been considered one of the biggest indicators on the health of any economy. After the devastating crash of the world economy in 2008, many real estate projects were halted and economies are still recovering from that incident. Real estate brokers and agents found it difficult to sell any property during that period, and they are still following the same valuation procedure that they have been using since. They mainly value the unit based on the cost of the unit, return on investment, price comparison to similar units and are focused on the current market condition. The problem with this valuation procedure is that it fails to take into account historical data, state of the unit, price fluctuation in the location and other valuable pieces of information. With that in mind, data mining could prove to be extremely helpful in determining proper prices and values for real estate units that would allow buyers, sellers, and real estate agents to make more informed decisions.

In this project we will be analyzing the Real Estate Market in the Emirate of Dubai to produce a model that can be predict the prices and valuations of real estate units in the emirate. We will collect demographic and property data, which will then be cleaned and combined for further analysis. The sources for the demographic dataset is Dubai Statistics Center and Google, while the source for the property dataset is PropertyFinder.ae. In the analysis stage useful plots will be produced to help understand the data and the relationship between the attributes. After which, a gradient boosting regression model will be used, and the data will be split into 80% training and 20% testing. The model managed to achieve an accuracy of 90.6%, and in the end our findings were compiled and example use cases were presented.

Keywords: Real Estate, Dubai, UAE, Artificial Intelligence, Machine Learning
# TABLE OF CONTENTS

**ACKNOWLEDGMENTS** ........................................................................................................... 3  
**ABSTRACT** ............................................................................................................................. 4  
**LIST OF FIGURES** .................................................................................................................. 6  
**LIST OF TABLES** ................................................................................................................... 6  
**CHAPTER 1** ............................................................................................................................ 7  

1.1 **BACKGROUND** .................................................................................................................. 7  

1.2 **PROBLEM STATEMENT** ...................................................................................................... 8  

1.3 **METHODOLOGY** ............................................................................................................... 8  

1.4 **LIMITATIONS OF THE STUDY** ....................................................................................... 10  

**CHAPTER 2 – LITERATURE REVIEW** .................................................................................... 11  

**CHAPTER 3 – PROJECT DESCRIPTION** ................................................................................ 13  

**CHAPTER 4 – PROJECT ANALYSIS** ..................................................................................... 14  

4.1 **DATA COLLECTION** .......................................................................................................... 14  

4.1.1 **DEMOGRAPHIC DATASET** ............................................................................................. 14  

4.1.2 **PROPERTY DATASET** .................................................................................................... 19  

4.2 **DATA CLEANING** ............................................................................................................. 19  

4.3 **DATA PREPROCESSING** ..................................................................................................... 28  

4.4 **DATA JOINING** ................................................................................................................... 28  

4.5 **DATA ANALYSIS** .............................................................................................................. 29  

4.6 **MODEL** ................................................................................................................................ 38  

4.7 **FINDINGS** .......................................................................................................................... 40  

**CHAPTER 5 – CONCLUSION** ............................................................................................... 44  

5.1 **CONCLUSION** .................................................................................................................... 44  

5.2 **RECOMMENDATIONS** ...................................................................................................... 45  

5.3 **FUTURE WORK** .................................................................................................................. 45  

**BIBLIOGRAPHY** ..................................................................................................................... 46
LIST OF FIGURES

Figure 1 Villas/Townhouses from Propertymonitor.ae .........................................................14
Figure 2 Apartments from Propertymonitor.ae ......................................................................15
Figure 3 Villas/Townhouses from Bayut.com ..........................................................................15
Figure 4 Apartments from Bayut.com ....................................................................................16
Figure 5 Histogram before removing outliers .......................................................................21
Figure 6 Price vs Type before removing outliers ....................................................................22
Figure 7 Bed vs Type before removing outliers ......................................................................22
Figure 8 Bath vs Type before removing outliers ......................................................................23
Figure 9 Area vs Type before removing outliers ......................................................................23
Figure 10 Histogram after removing outliers ..........................................................................24
Figure 11 Price vs Type after removing outliers .....................................................................25
Figure 12 Bed vs Type after removing outliers ......................................................................25
Figure 13 Bath vs Type after removing outliers .....................................................................26
Figure 14 Area vs Type after removing outliers ......................................................................26
Figure 15 Joined Dataset .........................................................................................................29
Figure 16 No. of Listings vs Population ...................................................................................30
Figure 17 Population growth in each Neighborhood ...............................................................31
Figure 18 Population in each Neighborhood ..........................................................................31
Figure 19 Avg. area for each unit type per location .................................................................31
Figure 20 Avg. area per No. of bedrooms for each type ............................................................32
Figure 21 Common unit type in each Neighborhood ...............................................................33
Figure 22 Avg. price vs Unit type in each Neighborhood ..........................................................34
Figure 23 Avg. price per sqft ....................................................................................................34
Figure 24 Price vs Location ......................................................................................................35
Figure 25 Price vs Unit Type ....................................................................................................35
Figure 26 Correlation Plot .......................................................................................................38
Figure 27 Model Accuracy ......................................................................................................39
Figure 28 Test vs Prediction plot ............................................................................................39
Figure 29 Website .....................................................................................................................40

LIST OF TABLES

Table 1 Demographic dataset data dictionary .......................................................................18
Table 2 Properties dataset data dictionary .............................................................................19
Table 3 Summary statistics before removing outliers ...............................................................27
Table 4 Summary statistics after removing outliers .................................................................27
Table 5 Changing Location from strings to integers ...............................................................36
Table 6 Changing Type from strings to integers .....................................................................36
CHAPTER 1

1.1 BACKGROUND

The real estate market has been in existence for a very long period of time and has always been considered a safe haven for many investors worldwide. In the 48 years since the establishment of the UAE, the real estate market in Dubai has drastically changed, from a couple of towers on what is now known as Sheikh Zayed Road, to a string of sky high towers that line the horizon of this Emirate. From the Dubai World Trade Center, to the tallest skyscraper in the world, Burj Khalifa, and the largest shopping mall in the world, The Dubai Mall.

It has been and still is considered one of the biggest markets in any economy in any nation worldwide and is always an indicator on the progress of the economy. Usually, whenever the real estate market is slowing down, we can immediately see the economy of the country also declining, and whenever the market is active and booming, the economy follows as well. The reason for this strong connection between the economy and this market, is the significance of the market and the stakeholders involved in it. People looking for houses, supermarkets, restaurants, and other services, businesses looking for offices, and investors pouring their money to provide them, these are the stakeholders involved in this market. This market creates opportunities to provide the basic necessities that all humans require: food, water and shelter.

This market can be broadly categorized into 2 main groups:

- Residential
- Commercial

Residential real estate includes houses, apartments and any other residential units that serve a purpose of being a house for any person.

Commercial real estate includes offices, storage units, shops and other retail units that are used to provide services and products to people.

The introduction of data analytics in this market will greatly help in sustaining stable growth levels to avoid huge economy crashes such as the one the world suffered in 2008. The process of determining price of retail estate through the use of data mining models will
greatly help in making the market more adaptive and fluid to changes in economy and will also aid in predicting future status of the market.

I am studying the prediction of real estate market prices to find out the outcomes of a data-driven price settings in a market that greatly affects the overall economy of any nation worldwide. It is unclear what the effects of setting price points based on data will be on the economy, and it is even more ambiguous what benefits will be produced. So the question that should be asked, “Will data mining provide a better price prediction and valuation of real estate units than real estate brokers?”

1.2 PROBLEM STATEMENT

The current method that real estate brokers and online platforms are using to price and value real estate units is based upon several factors such as the cost of the unit, unit size, price comparison to similar units, the current market conditions and the possible return on investment. The issue with this method is that it fails to take into account several other important pieces of information such as historical data, state of the unit, price fluctuations, proximity of points of interests and other valuable pieces of information. Buyers and sellers need to make informed decisions, while real estate agents need have a better insight on the real estate market. Using data mining algorithms and models, we may be able to satisfy the interests of all parties involved.

1.3 METHODOLGY

First of all, data for the Dubai residential real estate market must be found. This data must be relevant and should be clean. If the data is not clean, then cleaning must be done to ensure that the data mining process is effective, accurate and will answer the question stated earlier.

Other preprocessing measures must be taken such as dimensionality reduction, aggregation and discretization.

After the data is obtained, a model will be built to predict prices and values of real estate property by using a training and testing set. This way the accuracy of the model can be examined and errors can be reduced. This model might be a regression model or artificial neural network, depending upon the accuracy achieved, and will be built on either R or Python languages as both provide a good platform for the development of the model.
There are many different attributes that will appear in the different data sets to be obtained, however, not all will be relevant in the study. To select the relevant attributes, we must return to the literature review and observe the different studies performed in this topic and find commonalties with the data set we obtained. Indeed, not all the attributes will be similar, since some studies might include distance from railway or subway stations, and this will not be found in the data sets provided for the Emirate of Dubai, since there are no railroad or subway stations.

Preliminarily, the attributes that we are looking for in this study might include the following:

- Unit area
- Price of the unit
- Price per sq/ft
- Neighborhood
- Nearest landmarks
- Distance to metro
- Distance to bus stations
- Age of the unit
- Facilities in the area
- Historical prices

For the study to be more effective, apartments must be segregated from houses or townhouses, since other factors need to be considered such as the floor level of the apartment, the parking spots allocated for each apartment, and other factors as well. After taking into consideration all the points mentioned above, and performing the tasks initially stated, the model will then be tested until an acceptable rate is achieved. Thereafter, the model will then be tested in the real world scenario and a number of units will be sampled and the price and valuation of these units will be predicted by the model. These results will then be compared to the current listing prices and valuation of the units estimated by the realtor agents and brokers. If minute differences are present in the results, more samples will be taken. If there are major differences produced in the results, then we first have to determine the existence of such differences and these results will be presented to experts in the field to determine which of the results is more realistic and better.
1.4 LIMITATIONS OF THE STUDY

The main limitation of this study is the lack of available data on real estate in the Emirate of Dubai. Data on real estate is not publicly available, which will affect the analysis and the model outcome. Moreover, information regarding factors that may affect price (i.e. nearby establishments, services in the neighborhood, unit historical information, and other factors) are not readily available, which means this information must be extracted and might not be very accurate and up to date.
CHAPTER 2 – LITERATURE REVIEW

Dubai has grown to be one of the main hubs, in the region and the world, for investors and tourists alike. The freehold ownership policies have attracted many investors into investing in properties in the emirate, and this has led the real estate market to contribute around 15% to the total GDP of the emirate in 2008 (Hepşen & Vatansever, 2011). However, after the financial crisis of 2008 that hit the world and devasted economies, Dubai was not immune to it and has suffered greatly. Many investors left the country and many projects were halted. Workers also left the country leaving the Emirate in a terrible state and a huge debt to pay (Molotch & Ponzini, 2019).

After the crisis that hit the world and the Emirate as well, the authorities have been more careful despite the substantial growth in the real estate sector the economy has seen. In 2015, the year-on-year house price rate was about 27%, the fastest rate in the world. As mentioned earlier, the authorities have become more cautious and aware of what a large growth rate might bring along, another bubble, they have placed limits on mortgage lending and have imposed high transaction fees to brake to slow down the high price growth (Molotch & Ponzini, 2019).

The major factors that affect the demand of real estate are income, population and finance availability. Where as the factors that affect the supply of real estate are cost of financing and production, technological knowledge and future expectations (Abdelgalil & Bakheet, 2007).

According to Itedal (2013), the use of neural network to predict real estate market prices will greatly help in filling information gaps in the market and will also improve the efficiency of the market. In his controlled study of the use of neural network, the data mining model was able to produce an accuracy rate exceeding 95% in predicting prices of real estate units. This shows a positive outcome and provides hope for implementation of the network in real estate markets.

Sawant, Jangid, Tiwari, Jain, & Gupta (2018) highlight that a price prediction model will greatly help in making real estate agents, buyers and sellers more informed when making decisions and will also aid in avoiding overestimation or underestimation of the property prices, especially places that are growing and attracting more population to the area.
The customary way of valuing a real estate unit is through cost of the unit, return on investment of the unit and the price comparison between similar units and the issue with this customary way is the possibility of manipulation by the parties involved with a real estate transaction. Moreover, the standard way is focused on the current market condition of the market and the existing data, failing to take into account historical data and other valuable information such as the state of the unit, the price fluctuation in the neighborhoods, and other pieces of information (Hromada, 2016).

According to the study conducted by Fan, Ong, & Koh (2006), there are differences between the determinants home buyers look for when looking at either apartments or luxury units. When looking to buy an apartment, the home buyers look for the basic characteristics of the unit such as the area, age and the model of the unit. This is true for 2,3 or 4 room apartments, however, when it comes to 5 room apartments, they look for the floor level of the unit as well. The luxury apartment owners, on the other hand, pay more attention to the amenities, facilities and the quality of the overall apartment, and pay less attention to the basic conditions of the unit.

Another study conducted by Gan, Agarwal and Kim (2015) found that the real estate market provides good indication on the health of economy. Further, in their study they also implemented a decision tree model and a neural network model to forecast real estate prices, and they discovered that, overall, the neural network model was more accurate than the decision tree model.

In another study by Cacho (2010), it was found that most online real estate platforms utilize the Naïve Neighborhood algorithm when it comes to determining the prices of real estate units. This algorithm, despite its simplicity, has shown good performances when compared to other algorithms in this study.
CHAPTER 3 – PROJECT DESCRIPTION

This project is aimed at finding a model that helps users (i.e. Authorities and Individuals) find useful price related insights on the Real Estate market in Dubai. The current issue is that pricing and valuation is done based on the market demands and with a surplus of supply compared to demand, it becomes even more difficult and competitive to price these properties. Some areas, that are highly attractive, have higher prices compared to other properties that might be even better in terms of age, transportation, and services available around it. The model that is proposed, helps the users to make informed decisions when it comes to purchasing or regulating the prices so that prices are not inflated, and constructions is distributed more evenly across the Emirates. This might provide the Emirate with a more stable real estate market and the residents with better information at hand. This might also aid investors in selecting valuable properties to invest in and locations for construction. It can also help property managers in properly pricing their units based on their location and give them a competitive advantage. Moreover, new properties can use the tool to determine the prices they should place on their properties based on several factors such as location, age, unit type, floor level and other factors as well. Other benefits from the model could be extracted to provide other valuable pieces of information to the stakeholders.
CHAPTER 4 – PROJECT ANALYSIS

The project has been done through the use of Python programming language because of its flexibility, ease of use and the various machine learning algorithm modules it has.

4.1 DATA COLLECTION

The data for this project compromises of demographic data and property data.

4.1.1 DEMOGRAPHIC DATASET

The demographic data was collected from Dubai Statistics Center (DSC), and it included data on the population of each neighborhood in the emirate of Dubai, and from Google to incorporate the services/establishments within a 5km radius of each neighborhood.

The population dataset of DSC contained the population of 226 neighborhoods in the emirate of Dubai as of 2019. Not all these neighborhoods were selected, only the top 10 neighborhoods were selected for analysis. The selection process was based on the information obtained from online real estate platforms that show the popularity of the neighborhoods in Dubai. Popular neighborhoods for apartments and villas/townhouses were selected to make the analysis more comprehensive and to ensure the model works for either of the aforementioned residential property units. The following figures show the popular neighborhoods from 2 of the most popular online real estate platforms.

![Figure 1 Villas/Townhouses from Propertymonitor.ae](image-url)
Figure 2 Apartments from Propertymonitor.ae

Figure 3 Villas/Townhouses from Bayut.com
The neighborhoods that were selected are:

**Villas/Townhouses**
1. Arabian Ranches
2. Damac Hills (Akoya by Damac)
3. Jumeirah Park
4. Jumeirah Village Circle
5. Palm Jumeirah

**Apartments/Penthouses**
1. Dubai Marina
2. Downtown Dubai
3. Palm Jumeirah
4. Dubai Sports City
5. Dubai Silicon Oasis

The reason for selecting Palm Jumeirah in both unit types is because of the popularity of the area and its appeal for both apartments and villas/townhouses residents.
The process of obtaining the population data was not straightforward, this is because the neighborhood names in the dataset from DSC were different than the conventional naming of the neighborhoods. For example, Arabian Ranches community is the conventional name of the neighborhood, however in DSC dataset this area is called Wadi Al Safa 6. In order to obtain the population of the selected neighborhoods, Makani website, similar to Google Maps but belonging to Dubai Municipality, was used to obtain the official names of the neighborhoods. After that, these names were searched for in the DSC population data set, and the populations of the selected neighborhoods were obtained.

Apart from collecting demographic information, data on facilities and services around each neighborhood was collected. This data included the below facilities and services that are within a radius of 5Km from each neighborhood:

1. Schools
2. Universities
3. Hospitals & Clinics
4. Shopping Malls
5. Supermarkets
6. Cafes
7. Restaurants

This data was collected by using Google Places API to scrape data from Google Maps database. It was noticed that Google only allows a maximum of 60 establishments to be scraped for each category using their API. Hence, the number of establishments can be more than what was obtained. The reason for not going for a manual method of obtaining this data (i.e. manually counting each of the aforementioned categories by looking at Google Maps) is because when trying to search for these establishments on Google Maps the accuracy of the radius is not ensured, and when zooming in to count the establishments in the area not all of them appear and when refreshing the search other establishments appear and some of the older ones disappear, so the process was not reliable.

The distance of each neighborhood to the nearest metro station was also obtained by using the measure tool on Google Maps.

The dataset consists of 5 records and 17 attributes and the data dictionary of the dataset is presented in the table below:
<table>
<thead>
<tr>
<th>Si</th>
<th>Variable</th>
<th>Variable Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Si</td>
<td>Si of the records</td>
</tr>
<tr>
<td>2</td>
<td>Community Code</td>
<td>Community code of the neighborhood from DSC demographic dataset</td>
</tr>
<tr>
<td>3</td>
<td>Neighborhood</td>
<td>Name of the neighborhood</td>
</tr>
<tr>
<td>4</td>
<td>Pop2019</td>
<td>Population of the neighborhood in 2019</td>
</tr>
<tr>
<td>5</td>
<td>Pop2018</td>
<td>Population of the neighborhood in 2018</td>
</tr>
<tr>
<td>6</td>
<td>Pop2017</td>
<td>Population of the neighborhood in 2017</td>
</tr>
<tr>
<td>7</td>
<td>PopGrowth2018-19</td>
<td>Population growth in the neighborhood from 2018 to 2019</td>
</tr>
<tr>
<td>8</td>
<td>PopGrowth2017-18</td>
<td>Population growth in the neighborhood from 2017 to 2018</td>
</tr>
<tr>
<td>9</td>
<td>Nearest Metro Station</td>
<td>Name of the nearest metro station</td>
</tr>
<tr>
<td>10</td>
<td>Dist. To Nearest Metro Station (Km)</td>
<td>Distance to the nearest metro station</td>
</tr>
<tr>
<td>11</td>
<td>Schools</td>
<td>Number of schools within 5km radius of the neighborhood</td>
</tr>
<tr>
<td>12</td>
<td>Universities</td>
<td>Number of universities within 5km radius of the neighborhood</td>
</tr>
<tr>
<td>13</td>
<td>Hospitals &amp; Clinics</td>
<td>Number of hospitals &amp; clinics within 5km radius of the neighborhood</td>
</tr>
<tr>
<td>14</td>
<td>Cafes</td>
<td>Number of cafes within 5km radius of the neighborhood</td>
</tr>
<tr>
<td>15</td>
<td>Restaurants</td>
<td>Number of restaurants within 5km radius of the neighborhood</td>
</tr>
<tr>
<td>16</td>
<td>Shopping Malls</td>
<td>Number of shopping malls within 5km radius of the neighborhood</td>
</tr>
<tr>
<td>17</td>
<td>Supermarkets</td>
<td>Number of supermarkets within 5km radius of the neighborhood</td>
</tr>
</tbody>
</table>

Table 1 Demographic dataset data dictionary
4.1.2 PROPERTY DATASET

The property data was collected by scraping propertyfinder.ae for the selected neighborhoods. The units that were scraped are the ones listed as for sale. Penthouses have been included in the apartments search as they are considered as apartments in nature. Duplex units were not collected due to some listings showing them as apartments while others show them as villas/townhouses. To avoid confusion during the analysis process, they were eliminated. Moreover, the number of duplexes listed is very small in comparison to the number of villas/townhouses and apartments.

The dataset consists of 9171 records and 7 attributes and the data dictionary of the dataset is presented in the table below:

<table>
<thead>
<tr>
<th>Si</th>
<th>Variable</th>
<th>Variable Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Property</td>
<td>Name of the property listed</td>
</tr>
<tr>
<td>2</td>
<td>Location</td>
<td>Name of the neighborhood</td>
</tr>
<tr>
<td>3</td>
<td>Type</td>
<td>Type of the listed unit</td>
</tr>
<tr>
<td>4</td>
<td>Bed</td>
<td>Number of beds in the listed unit</td>
</tr>
<tr>
<td>5</td>
<td>Bath</td>
<td>Number of baths in the listed unit</td>
</tr>
<tr>
<td>6</td>
<td>Area</td>
<td>Area of the unit listed</td>
</tr>
<tr>
<td>7</td>
<td>Price</td>
<td>Price of the listed unit</td>
</tr>
</tbody>
</table>

Table 2 Properties dataset data dictionary

4.2 DATA CLEANING

In the demographic dataset, a few attributes were dropped due to their irrelevance in the analysis. These attributes were the Si, Community Code and the Nearest Metro Station. No further cleaning was required.

In the property dataset, cleaning was required. This included cleaning the dataset by replacing certain missing values, removing irrelevant attributes, deleting other NA values, and removing outliers.

When scraping the data, some listings did not show the number of bathrooms, hence I had to manually input this number based on either similar properties listed by different agents or similar properties in the building with approximately similar size. If no similar properties
are available, the number of bathrooms was selected to be 1 more than the number of bedrooms. This is because, many listings have the bathrooms to be one more than the bedroom, while other listings have equal number of bedrooms and bathrooms. The listings with equal number of beds and baths do not include bathrooms without showers (i.e. guest toilets). I have noticed that the number of listings without the bathrooms listed is 1% of the entire listings within the area (e.x in Dubai Silicon Oasis there were only 2 properties, out of 213, that did not have the number of bathrooms listed. Similar percentage was seen in all other locations.) The only exception to this case were studios, where the number of bathrooms will always be 1.

The property name attribute includes names of the properties and hence was removed from the dataset because it provides no value in the analysis.

Some of the listings did not mention the price, however, the statement ‘Ask for Price’ was placed in the price field. Hence, these properties were removed from the dataset. The total number of properties that failed to mention the price were 138 and they were all removed.

Outliers were present in the property data set and hence they were removed to enhance the accuracy of the model to be built.

In the figures below we can see a histogram of the attributes and the outliers present in the dataset.
Figure 5 Histogram before removing outliers
Figure 6 Price vs Type before removing outliers

Figure 7 Bed vs Type before removing outliers
Figure 8 Bath vs Type before removing outliers

Figure 9 Area vs Type before removing outliers
After removing the outliers from the dataset we produce similar plots to visualize the change. The figures below show the histogram and attributes after removing the outliers.

*Figure 10 Histogram after removing outliers*
Figure 11 Price vs Type after removing outliers

Figure 12 Bed vs Type after removing outliers
Figure 13 Bath vs Type after removing outliers

Figure 14 Area vs Type after removing outliers
As we can see from the plots, we have managed to remove many of the outliers present in the dataset to ensure that we have a higher accuracy model. Before removing the outliers we had a total of 9033 records, and after removing the outliers we ended up with 7555 records. The tables below show a summary statistics of the dataset before and after removing the outliers.

<table>
<thead>
<tr>
<th></th>
<th>Bed</th>
<th>Bath</th>
<th>Area</th>
<th>Price</th>
</tr>
</thead>
<tbody>
<tr>
<td>count</td>
<td>9033.000000</td>
<td>9033.000000</td>
<td>9033.000000</td>
<td>9.033000e+03</td>
</tr>
<tr>
<td>mean</td>
<td>2.445588</td>
<td>3.185653</td>
<td>2544.031772</td>
<td>4.319357e+06</td>
</tr>
<tr>
<td>std</td>
<td>1.421223</td>
<td>1.494219</td>
<td>2751.974793</td>
<td>7.826231e+06</td>
</tr>
<tr>
<td>min</td>
<td>0.000000</td>
<td>1.000000</td>
<td>3.000000</td>
<td>2.250000e+05</td>
</tr>
<tr>
<td>25%</td>
<td>1.000000</td>
<td>2.000000</td>
<td>1048.000000</td>
<td>1.327500e+06</td>
</tr>
<tr>
<td>50%</td>
<td>2.000000</td>
<td>3.000000</td>
<td>1667.000000</td>
<td>2.099000e+06</td>
</tr>
<tr>
<td>75%</td>
<td>3.000000</td>
<td>4.000000</td>
<td>3063.000000</td>
<td>3.700000e+06</td>
</tr>
<tr>
<td>max</td>
<td>7.000000</td>
<td>7.000000</td>
<td>88264.000000</td>
<td>1.150000e+08</td>
</tr>
</tbody>
</table>

*Table 3 Summary statistics before removing outliers*

<table>
<thead>
<tr>
<th></th>
<th>Bed</th>
<th>Bath</th>
<th>Area</th>
<th>Price</th>
</tr>
</thead>
<tbody>
<tr>
<td>count</td>
<td>7555.000000</td>
<td>7555.000000</td>
<td>7555.000000</td>
<td>7.555000e+03</td>
</tr>
<tr>
<td>mean</td>
<td>2.176042</td>
<td>2.881668</td>
<td>2015.016942</td>
<td>2.671627e+06</td>
</tr>
<tr>
<td>std</td>
<td>1.304656</td>
<td>1.350443</td>
<td>1556.935519</td>
<td>3.029786e+06</td>
</tr>
<tr>
<td>min</td>
<td>0.000000</td>
<td>1.000000</td>
<td>65.000000</td>
<td>2.250000e+05</td>
</tr>
<tr>
<td>25%</td>
<td>1.000000</td>
<td>2.000000</td>
<td>938.000000</td>
<td>1.220000e+06</td>
</tr>
<tr>
<td>50%</td>
<td>2.000000</td>
<td>3.000000</td>
<td>1450.000000</td>
<td>1.800000e+06</td>
</tr>
<tr>
<td>75%</td>
<td>3.000000</td>
<td>4.000000</td>
<td>2446.000000</td>
<td>2.883888e+06</td>
</tr>
<tr>
<td>max</td>
<td>7.000000</td>
<td>7.000000</td>
<td>9366.000000</td>
<td>2.950000e+07</td>
</tr>
</tbody>
</table>

*Table 4 Summary statistics after removing outliers*
We can see that after removing outliers almost all the statistics have reduced in value, especially the standard deviation in the Price attribute. Also the maximum values in both the Area and Price attributes have reduced greatly.

4.3 DATA PREPROCESSING

In the preprocessing stage we had to change all the string values to integers so we are able to visualize them and create our model. This step was broken down into stages, the first stage was to convert the Bed, Area and Price attributes of the Property dataset into integers because of the string values present in these attributes. The Bed attribute had ‘Studio’ as one of the entries and so we changed it to 0 and not 1 since the 1 was used to describe a 1 bedroom apartment. The Areas and Price attributes had commas in them and ‘sqft’ and ‘AED’ in them, respectively. We had to remove them from the attributes so we can perform the analysis. This stage of preprocessing was done during the cleaning stage, and this is because when we had done this process we had to change the ‘Ask for price’ entries in the Price attribute to NA when converting the attribute to integers. The second stage of preprocessing involved label encoding the Location and Type attributes of the Property dataset from strings to integers so we can prepare the data for our model and create some visualizations. The change of these 2 attributes will be done before plotting the correlation matrix and the creating model so that we are able to produce plots that contain the names of these locations and the unit types as well. Hence this change process will be shown prior to starting the correlation plot in the Data Analysis step.

4.4 DATA JOINING

We joined the 2 datasets together to be able to go ahead and build our model. We want to be able and see what attributes affect property prices, such as the number of schools in the neighborhood or the population growth in the area, and so on.

We joined the dataset on the neighborhood attribute, since they contain the same values, the names of the neighborhood. However, due to the attribute being labeled as Location in the Property data set and Neighborhood in the Demographic dataset, we had to drop one of these 2 attributes.

Hence, we dropped the Neighborhood attribute after merging the dataset.

The table below shows a sample head of the joined dataset.
4.5 DATA ANALYSIS

In this step we produce plots and visualizations that will help us to better understand the dataset we have and to explore relationships. We want to investigate the joined dataset to discover patterns and to obtain important information.

We start of by finding the number of listings versus the number of population in each area. Places that have a high population and high number of listings indicate that the area has room to accommodate more people and has not yet reached its limit in terms of occupation. On the other hand, locations that have a low population and low number of listings indicate that the area can accommodate more people but it’s reaching its limit. Similarly, locations that have a high population and low number of listings indicate that these areas have almost reached their limits in terms of occupation. In the plot below we can see that Dubai Marina, Palm Jumeirah and Downtown Dubai have a high number of listings, which means that these places can accommodate more people. The other locations all have low number of listings which means that they are close to their occupation limits, especially Jumeirah Park and Dubai Silicon Oasis as these locations have a high population and low number of listings.
We then plot the population and population growth in each area to better understand the influx of people into these locations, as can be seen in the plots below. As we can see, the population from 2017 to 2019 is increasing in all these locations. However, this increase in population is smaller from 2018 to 2019 in comparison to 2017 to 2018. This can be seen by looking at the population growth in these areas. The growth has decreased in all the locations except for 2, Dubai Marina and Dubai Silicon Oasis. We can relate this growth to our previous claim that certain locations have reached their limits in terms of accommodating people, such as in Jumeirah Park, we can see that the population growth in 2018-2019 has decreased in comparison to 2017-2018. Which means that there is not enough space for more people to live in this area. However, in the case of Dubai Silicon Oasis, we can see that the population growth has increased, but this increase is small, and the location might seem to have reached its limit because the number of listings is small. It is expected that the growth from 2019 to 2020 to decrease in this location. In the case of Dubai Marina, the growth is increasing and the number of listings is high which means that there is more space for more people and the population growth to increase from 2019 to 2020. For all other locations the population growth will decrease until there is no more room for extra people.
The next thing we are looking for is the average area size for each unit type in each Neighborhood, which is shown in the plot below. We can see that penthouses are generally higher larger in size in comparison to the other unit types, followed by villas, then townhouses and finally apartments.
Continuing on our investigation of area, we would like to understand what is the most common number of bedrooms and the associated size of these units. Hence, the plot below shows this information. As we can see, the 3 bedroom units seems to be the size that all unit types can come in. An interesting observation that can be seen is that 6 bedroom villas are smaller in size compared to the 5 bedroom villas, and the 2 and 3 bedroom villas are almost similar in size. Another observation that can be seen is that there is a linear increase in size of apartments as the number of bedrooms increase, which is not the case for the other unit types.

![Figure 20 Avg. area per No. of bedrooms for each type](image)

Afterwards, we wanted to investigate what is the most common unit type in each of the locations. The plot below shows that in Arabian Ranches and Damac Hills we can see more villas than townhouses. While in Jumeirah Village Circle we can see more townhouses than villas. In Dubai Silicon Oasis and Jumeirah Park we can see that there is only apartments and villas, respectively. For the other locations, apartments seem to be the most common unit types.
The relationship between price of units in each location was of interest, this is mainly due to the fact that there is a difference between the price of a certain unit in different locations. The plot below shows this difference. As we can see, penthouses have the highest price variations depending upon the location. We can see that we can get a penthouse in Dubai Sports City for almost the same price as a townhouse in Palm Jumeirah, and as we saw in the area plots previously, the size of the penthouse in Dubai Sports City is larger than townhouses in Palm Jumeirah, which may seem as a good value to some people. Another observation we can see is that villas in Jumeriah Village Circle are cheaper than villas in Arabian Ranches and Damac Hills, and they are larger in size as we saw in the area plots earlier. The most expensive units are penthouses in Downtown Dubai, whereas, the cheapest units are apartments in Dubai Silicon Oasis and Dubai Sports City. The plot below it shows the price per sqft for each unit type in all locations. From this plot we can extract more insight on price information, from the previous plot we could only get the average prices of the units. However, the price per sqft plot shows us how much area we are getting for the price we are paying. From the Avg. price vs unit type plot we can see that townhouses in Palm Jumeirah are more expensive than the apartments in the same location. However, if we look at the Avg. price per sqft plot we can see that the apartments are more expensive than townhouses if we consider the same area for both unit types. The
Avg. price per sqft is a better indicator of the value received for the money paid in terms of area of the unit.

**Figure 22** Avg. price vs Unit type in each Neighborhood

**Figure 23** Avg. price per sqft
The price plot above begs a question, what affects the prices and why are there differences between the locations? To answer this question, a correlation plot was created to better understand the relationship between price and the other attributes present in the dataset. To create the correlation plot we had to convert the string attributes to integers, namely the Location and Type attributes. To determine the numbers that should be assigned to each location and unit type, the location/type with the lowest average price will be assigned the number 0, and accordingly, the location/type with the highest average price will be assigned the highest number. The plots below show the average price for each location and type.
These changes to be made are the second stage of the preprocessing step as mentioned earlier, and they are:

<table>
<thead>
<tr>
<th>From</th>
<th>To</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dubai Silicon Oasis</td>
<td>0</td>
</tr>
<tr>
<td>Dubai Sports City</td>
<td>1</td>
</tr>
<tr>
<td>Dubai Marina</td>
<td>2</td>
</tr>
<tr>
<td>Jumeirah Village Circle</td>
<td>3</td>
</tr>
<tr>
<td>Downtown Dubai</td>
<td>4</td>
</tr>
<tr>
<td>Damac Hills</td>
<td>5</td>
</tr>
<tr>
<td>Arabian Ranches</td>
<td>6</td>
</tr>
<tr>
<td>Jumeirah Park</td>
<td>7</td>
</tr>
<tr>
<td>Palm Jumeirah</td>
<td>8</td>
</tr>
</tbody>
</table>

*Table 5 Changing Location from strings to integers*

<table>
<thead>
<tr>
<th>From</th>
<th>To</th>
</tr>
</thead>
<tbody>
<tr>
<td>Apartment</td>
<td>0</td>
</tr>
<tr>
<td>Townhouse</td>
<td>1</td>
</tr>
<tr>
<td>Villa</td>
<td>2</td>
</tr>
<tr>
<td>Penthouse</td>
<td>3</td>
</tr>
</tbody>
</table>

*Table 6 Changing Type from strings to integers*

From the correlation plot we can see that price has a positive correlation with Location (0.39), albeit not a strong correlation compared to Type (0.63), Bed & Bath (0.57), and Area (0.73). So from this correlation we can see that price is strongly related to the size of the unit, the type of the unit and the number of beds and baths. However, Location is not strongly correlated to Price, and therefore we can deduce that Location is the attribute with the least effect on Price and Area is the greatest.

Other interesting correlations that appear in the plot is the positive relation between Location and Restaurants (0.64), and Location and Area (0.44). We can see that Area is
strongly positively correlated to Bed (0.85), Bath (0.82) and Type (0.83), and this comes as no surprise. As the area gets bigger the unit tends to be larger and hence it has more bedrooms, bathrooms and is generally either a villa or penthouse.

Another interesting correlation is between schools and restaurants (-0.74) and schools and shopping malls (0.77). This shows that in school areas there are little restaurants because of the negative correlation, however, what is more surprising is the strong positive relation between the schools and shopping malls. This means that there are more malls nearby school areas.

The correlation between universities and hospitals & clinics is 0.65, universities and cafes is 0.68, and universities and supermarkets is 0.73. These coefficients are understandable since university students and faculty will usually require supermarkets nearby, especially if they are living in dormitories. They need cafes to study in during the day or take a break from the university life, and hospitals & clinics in case of emergencies, especially for students/faculties living on campus.

Finally, the correlation coefficient between distance to metro stations and cafes is -0.83, and distance to metro stations and hospitals & clinics is -0.96, and distance to metro stations and supermarkets is -0.78. These strong negative correlation coefficients indicate that as we move further away from metro stations the less of the aforementioned services are present in the neighborhood.

These correlations shows us interesting relationships between different attributes, which might provide value to certain investors, government entities and people looking for a place to settle. For example, neighborhoods that are far away from metro stations may require more hospitals or clinics, and hence the Health Authority in the emirate might provide public clinics in the area, or investors may want to build cafes in these neighborhoods.

People who are looking for a place to settle might want to look into the relationship between distance from metro station and other services, the closer the place is to metro stations the more cafes, supermarkets, universities, hospitals & clinics and shopping malls. Other interesting relationships can be drawn from this correlation and people, investors and government authorities can make well and informed decisions using this plot.
Initially, a decision tree or artificial neural network model was suggested as the machine learning algorithm. However, after trying to use these models with the given datasets, the models did not perform as well as expected. Hence, another model was used for this project, gradient boosting regression model.

We first had to define what our input and output attributes would be. The output attribute was the price and all other attributes were considered as inputs. We then split our dataset into 80% training and 20% testing.

Using the gradient boosting regression we configured the number of estimators (number of boosting stages performed by the model) to be 400 and the max depth of the model to be 3 (it is the max depth of the decision tree estimator in the model). We also used the least squared regression as our loss function and the learning rate to be 0.1 (determines the step size at each iteration). We configured our minimum samples split to be 2 (the minimum number of samples to split and internal node).

**4.6 MODEL**

Finally, a decision tree or artificial neural network model was suggested as the machine learning algorithm. However, after trying to use these models with the given datasets, the models did not perform as well as expected. Hence, another model was used for this project, gradient boosting regression model.

We first had to define what our input and output attributes would be. The output attribute was the price and all other attributes were considered as inputs. We then split our dataset into 80% training and 20% testing.

Using the gradient boosting regression we configured the number of estimators (number of boosting stages performed by the model) to be 400 and the max depth of the model to be 3 (it is the max depth of the decision tree estimator in the model). We also used the least squared regression as our loss function and the learning rate to be 0.1 (determines the step size at each iteration). We configured our minimum samples split to be 2 (the minimum number of samples to split and internal node).
We then used our training dataset and fit it to the model, after which we tested the model using the testing dataset and compared the results to the testing output attribute. We managed to obtain an accuracy of around 90.6%.

```python
# fit and predict
clf_join.fit(x_j_train, y_j_train)
clf_join.score(x_j_test, y_j_test)
```

0.9064787998692496

*Figure 27 Model Accuracy*

The figure below shows the test samples vs prediction samples to visualize the performance and accuracy of our model. We have used the area attribute as the test input variable since it has the highest correlation coefficient with price.

*Figure 28 Test vs Prediction plot*

With the help of Flask, the model was linked to a simple HTML website that produced a simple UI for users to input the details of the unit they want a price estimate of. The website was connected to the model that was saved as pickle file, and using a JS file, the predictions were computed in a simple python script that contained the model and the code to push the results back to the website for users to view. From the image below, we can see that the website allows the users to input the area of the unit, select the number of beds and baths, choose the unit type and in which location using a drop down menu. The estimated price is
then generated once the user clicks on the ‘Estimate Price’ button at the bottom of the page.

Figure 29 Website

4.7 FINDINGS

From the data sets that we obtained and the analysis we have done, we managed to obtain useful insight on the Dubai Real Estate Market. This information may help the public, investors and government authorities make informed decisions when it comes to real estate and the population in the emirate.

From the histogram, after cleaning the property dataset, we found that that most of the unit sizes lie between 1,000 and 2,000 sqft and the most common unit type are apartments, followed by villas. Also, the most common number of bedrooms is 2 followed by 3, and for bathrooms its 2 followed by 3. The difference between common number of bedrooms and bathrooms is because in some listings bathrooms are equal to number of bedrooms, in other listings the bathrooms will be one more than the bedrooms. We also found out that the most listings appear in Dubai Marina and most of the prices are below 2,500,000 AED. This information can help investors when planning to purchase or build residential units by looking at the average area sizes and the most common number of beds and baths while keeping the prices within the acceptable range. For example, an investor might want to
invest in a 1,500 sqft apartment with 2 bedrooms, 2 bathrooms and a price tag less than 2,500,000 AED.

From the boxplots of the property dataset we found out that Palm Jumeirah has the most expensive villas, townhouses and apartments with largest areas for both these unit types, most and least amount of bedrooms for townhouses and the largest number of bedrooms for penthouses. Jumeirah Village Circle had the least expensive villas and townhouses, while Arabian Ranches had the largest number of bedrooms for villas. Downtown Dubai had the most expensive penthouse, however, Dubai Marina had the largest area for penthouses. Dubai Silicon Oasis and Dubai Sports City had the cheapest apartments. People looking to buy units to live in can benefit from this information by taking into account the previous insight that has been provided. For example someone wanting to buy a villa on a budget might want to consider purchasing one in Jumeirah Village Circle as it has the cheapest villas.

The summary statistics provide useful information on the property dataset as a whole. We can see that the minimum price of a unit listed was 225,000 AED, while the mean price of units was around 2,670,000 AED. For area, we can see that mean is 2,015 sqft and the minimum available in the dataset is 65 sqft. For bedrooms and bathrooms the mean is 2 and 3, respectively. This information can also be useful to investors.

In the listings vs population plot we found that Dubai Marina has the highest population and highest number of listings. While the lowest population in a neighborhood can be seen in Damac Hills and the lowest number of listings can be seen Jumeirah Village Circle. In Dubai Silicon Oasis and Jumeirah Park we can see a high population count but a small number of listings, while in Downtown Dubai and Palm Jumeirah we have a medium population and high number of listings. From this information we can determine whether a location has reached its limit in terms of occupancy by looking at the population and listings. In the locations that have a high population and listings, we know that the neighborhood has not reached its limits. Neighborhoods that have a high population and low listings may indicate these locations have reached their limits. In the plot, we can see that Dubai Marina has a high number of listings and a high population, which means that this area is large and can accommodate more people. Similarly, Downtown Dubai and Palm Jumeirah have a medium population count but a high number of listings, which also indicates that they can accommodate more people. Between these 3 locations, Downtown Dubai is further away
from its occupancy limits than the other 2 locations as it has a medium population count but a higher number of listings compared to Palm Jumeirah. This information can be beneficial to both investors and government entities. For investors, locations that have a high number of listings but a low population count could indicate that people are not willing to live in that area. For government entities, the previously mentioned condition could indicate that these areas need to have more services and facilities nearby to attract more people to the areas, such as more electricity and water services.

The population growth plot shows that all locations have a decreased growth rate between 2018 and 2019 compared to 2017 and 2018, except for 2 locations. These two locations are Dubai Marina and Dubai Silicon Oasis. This means that the aforementioned locations seem to be more appealing for people to live in, and this could be quite useful for investors and government entities. Investors might want to buy units there to rent to people, and government entities might want to increase these services in the area and keep them well maintained.

From the average area plot for each unit type, we can see that penthouses have the highest average area, followed by villas, then townhouses and finally apartments. We can see from the area vs bed plot that 5 bedroom villas are larger than 6 bedroom villas and the 2 and 3 bedroom villas are almost similar in area. Another useful piece of information we can observe is that 2 and 3 bedroom townhouses and villas are almost similar in size, and from the price boxplot we can see that townhouses are cheaper than villas. This could be useful for residents and investors. Residents may want to buy 5 bedroom villas, as they are larger in size than 6 bedroom villas, or they may want to buy 2 bedroom villas over a 3 bedroom villas as they might have more space. They might also want to buy townhouses if they are considering getting 2 or 3 bedroom villas, as they are cheaper. Investors can make 3 or 6 bedroom villas that are larger in size than the 2 and 5 bedroom villas, or they may want to invest in 2 and 5 bedroom villas as people are likely to live in them. They might want to invest in townhouse when it comes to a smaller number of bedrooms, as people are likely to select townhouses over villas due to the cheaper price.

From the average area and average price of unit types in each neighborhood plots we can see that penthouses in Dubai Sports City are cheaper and larger in size than townhouses in Palm Jumeirah. Also, villas in Jumeirah Village Circle are cheaper and larger than villas in Damac Hills and Arabian Ranches. People looking for places to live might want to consider
living in a penthouses in Dubai Sports City or villas in Jumeirah Village Circle. From the average price per sqft plot we can have better indication on the actual value of the unit in terms of the price paid for the area. As mentioned earlier, townhouses in Palm Jumeirah are more expensive than apartments in that location, however, if both units have the same area then townhouses become cheaper than apartments.

The correlation plot shows many interesting relationships and provides useful information to people, investors and government entities. We can see that price has strong positive correlation with area, type, bed and bath. While the correlation between price and location is not strong, yet positive. Shopping malls and schools have a strong positive correlation as well, and the distance to metro stations has a strong negative correlation with cafes, hospitals and clinics, and supermarkets. This relationship between schools and shopping malls should be looked by educational institutions if they have a case of many students disappearing from schools in between classes. Or establishments in the shopping malls might want to open up earlier for parents who have dropped their kids at school and are looking for a place to grab breakfast or do some shopping early in the day. The relationship between distance to metro stations and cafes, hospitals and clinics, and supermarkets means that the further away a neighborhood is from metro stations the lower number of establishments these areas have. This might be useful for government entities as they may want to build more hospitals or clinics in these areas, or they may want to create roads that provide quick and direct routes to these establishments. Investors may want to open more supermarkets in these areas, and residents need to consider the fact that buying a car while living in these areas may be required.
5.1 CONCLUSION

In this project, we have managed to collect data on property listings in 9 neighborhoods in Dubai that contain villa, townhouse, apartment and penthouse residential unit types. We selected 5 neighborhoods for villas and townhouses, and 5 neighborhoods for apartments and penthouse. However, 1 neighborhood appeared to be popular for both these categories. The popularity of these neighborhoods were gathered from 2 major online real estate platforms. We also managed to collect demographic data from Dubai Statistics Center to include in the previous dataset. Both datasets were cleaned by replacing certain missing values, removing irrelevant attributes, deleting NA values, changing types of certain attributes and removing outliers. We than began the preprocessing step in 2 stages, the first stage was involved converting certain attributes from strings to integers and was done during the cleaning stage, and the second stage also involved converting a couple of attributes from strings to integers, and was done at later stage during the analysis. The two datasets were then joined and the analysis was performed on the joined dataset. We managed to create various plots to visualize our data to provide valuable insight. A correlation plot was created to determine the factors that affect the price of units. We found out that price is strongly related to the size of the unit, the type of the unit and the number of beds and baths. The price was not correlated with other factors such as schools, supermarkets, cafes, hospitals & clinics, restaurants, universities, shopping malls and distance to metro station. These factors are taken into consideration by home owners when looking for a place to live in, however, they do not affect the valuation of the residential units. This proves our statement in the problem statement section, where we mention that real estate agents and their current valuation method fail to consider other important factors that help stakeholders make informed decisions. We then built our gradient boosting regression model and split our dataset into 80% training and 20% testing. After training the model, we tested its performance and managed to achieve an accuracy of 90.6%. In the final step of the project, we highlighted the findings of the project and showcased some examples where they might be useful to the public, investors or government entities.
5.2 RECOMMENDATIONS

Since there are no historical data for the properties available in Dubai, the data utilized in this model must be updated on a regular basis, quarterly, semi-annually or annually. This is to ensure that the data is accurate and up to date. Moreover, a new attribute will be included in the dataset, namely ‘YEAR’, this to ensure that we are recording the data as we are updating the dataset. By creating and keeping historical information we will be able to train our model better and help increase its accuracy and performance, and provide insight and information on real estate year-wise.

5.3 FUTURE WORK

To take this project to next level and make it a viable tool for everyone to use, all neighborhoods in the Emirate of Dubai must be included in the model. Moreover, the web platform that has been created should be made public and accessible for everyone. After these steps have been taken, real estate information for the other Emirates must be analyzed and incorporated in the model. This way, the prediction tool and website will be available nation-wide. The dataset also would be updated on a regular basis as mentioned in section 5.2 ‘Recommendations’.
BIBLIOGRAPHY


