****

**DATA WAREHOUSING AND DATA MINING**

**REAL ESTATE PRICE PREDICTION**

**REPORT – 3**

By

Arooba Siddiqi               321618

Emaan Bashir          296190

Rimsha Mirza 285718

Faraz Shah 303087

**CONTENTS**

[**INTRODUCTION:** 3](https://d.docs.live.net/fa426205b6217983/Desktop/Data%20Warehousing%20and%20Data%20Mining/Project/Report%202.docx#_Toc123332075)

[**PROJECT SOLUTION:** 4](https://d.docs.live.net/fa426205b6217983/Desktop/Data%20Warehousing%20and%20Data%20Mining/Project/Report%202.docx#_Toc123332075)

[**IMPLEMENTATION:** 4](https://d.docs.live.net/fa426205b6217983/Desktop/Data%20Warehousing%20and%20Data%20Mining/Project/Report%202.docx#_Toc123332076)

[**DATA COLLECTION:** 4](https://d.docs.live.net/fa426205b6217983/Desktop/Data%20Warehousing%20and%20Data%20Mining/Project/Report%202.docx#_Toc123332077)

[**DATA PRE-PROCESSING:** 4](https://d.docs.live.net/fa426205b6217983/Desktop/Data%20Warehousing%20and%20Data%20Mining/Project/Report%202.docx#_Toc123332078)

[**DATA VISUALIZATION:** 5](https://d.docs.live.net/fa426205b6217983/Desktop/Data%20Warehousing%20and%20Data%20Mining/Project/Report%202.docx#_Toc123332078)

[**PRICE PREDICTION:** 9](https://d.docs.live.net/fa426205b6217983/Desktop/Data%20Warehousing%20and%20Data%20Mining/Project/Report%202.docx#_Toc123332078)

**INTRODUCTION**

Predicting real estate prices is a crucial task that impacts both buyers and sellers in the real estate industry. For buyers, accurate price predictions can assist with determining whether or not to purchase a property and at what price point. For sellers, price predictions can aid in setting a suitable price for their property.

In Pakistan, the real estate market has seen significant growth in recent years, but there are also many challenges and uncertainties that can affect the prices of properties. These challenges include : economic factors such as inflation, unemployment, and changes in interest rates, as well as local factors such as the availability of land, the quality of infrastructure, and the overall demand for real estate.

Our project aims to provide a data-driven approach for real estate price prediction in Pakistan. By collecting and analyzing relevant data, we can build a model that accurately predicts the prices of properties in different locations and under different market conditions.

To conduct our analysis, we will collect data on the characteristics of specific properties from ‘Zamee.com’. We will then apply appropriate data mining techniques to identify patterns and relationships in the data, and use these insights to develop a predictive model, and some visualizations.

Our project has the potential to provide valuable insights and assistance to both buyers and sellers in the real estate market. For buyers, our model can help them make informed decisions about whether to purchase a property and at what price, potentially saving them money and avoiding overpaying for a property. For sellers, our model can help them set appropriate prices for their properties, potentially increasing their chances of a successful sale and maximizing their profits.

Overall, our data warehousing and data mining project aims to address a real-world problem that has significant practical applications in the Pakistani real estate market. By providing accurate price predictions and helping buyers and sellers make informed decisions, we hope to contribute to a more efficient and effective real estate market in Pakistan.

**IMPLEMENTATION**

**DATA COLLECTION**

The first step in the project implementation was data collection. This was done using web scraping techniques to extract relevant information from the zameen.com website.

**THE DATASET**

The dataset consisted of information about 280,046 properties from 189 different cities. The following information about each property was extracted from the website.

* URL
* Title
* Type
* Price
* Location
* Baths
* Beds
* Area
* Purpose
* Date Added
* Description

The different types of properties found to be in the dataset were:

* Residential Plot
* House
* Flat
* Commercial Plot
* Shop
* Plot File
* Office
* Building
* Agricultural Land
* Warehouse
* Industrial Land
* Factory
* Farm House
* Room
* Penthouse

Each property was uploaded for one of the following two purposes:

* For Sale
* For Rent

**WEB SCRAPING**

Web scraping was done using the *BeautifulSoup* library. The Steps followed were as follows:

* Analyze the URLs of the website and come up with a general format for the URLs

**URL for property for Sale:**

[https://www.zameen.com/{type}/{city}-{page}.html](https://www.zameen.com/%7btype%7d/%7bcity%7d-%7bpage%7d.html)

**URL for House for Rent:**

[https://www.zameen.com/Rentals/{city}-{page}.html](https://www.zameen.com/Rentals/%7bcity%7d-%7bpage%7d.html)

**URL for Plot for Rent:**

[https://www.zameen.com/Rentals\_Plots/{city}-{page}.html](https://www.zameen.com/Rentals_Plots/%7bcity%7d-%7bpage%7d.html)

**URL for Commercial Property for Rent:**

[https://www.zameen.com/Rentals\_Commercial/{city}-{page}.html](https://www.zameen.com/Rentals_Commercial/%7bcity%7d-%7bpage%7d.html)

* Make a request to get the HTML content of each page
* Parse the HTML code and find the required data
* Implement threads to extract data from multiple pages at a time
* Store the extracted data in a .csv file

**DATA PRE-PROCESSING**

Once the data was collected, it was cleaned and pre-processed for further analysis. For this purpose, we used *pandas* and *numpy* libraries. The steps followed were as follows:

* Get the numerical value of price from the price strings
* Get the city and province name from the location column
* Parse the strings to get the number of beds and baths from the corresponding columns
* Convert the units of area into square feet
* Get the estimated year, month and date from the strings in the Date Added column (The column consisted of strings such as ‘4 months ago’)
* Handle null and zero values through the methods of dropping, default values etc
* Set data type for each column
* Assign same property type to duplicate property types (e.g. house and home)
* Remove outliers based on the area
  + The outliers were removed separately for each property type
* Remove outliers based on price
  + The outliers were removed separately for each combination of property type, city and purpose (rent or sale)
* Save the pre-processed data in a csv file

**DATA VISUALIZATION**

After the data was cleaned and pre-processed, it was visualized using *Tableau*. This helped to better understand the relationships and patterns in the data, and to identify any trends.

The following 5 types of graphs were used for the visualization:

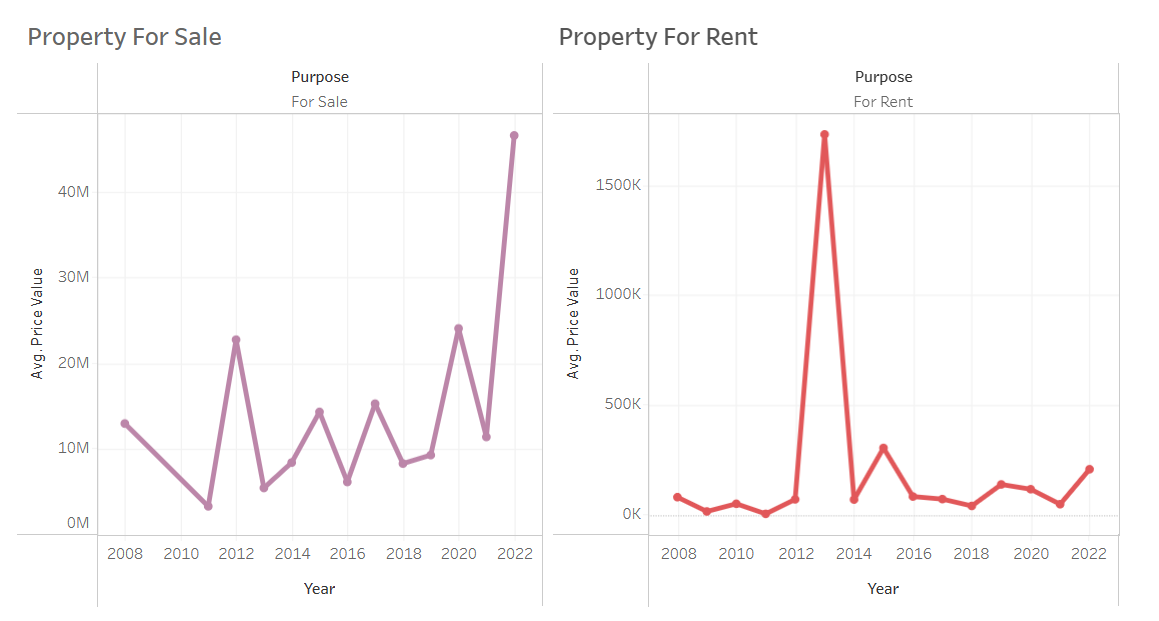
* Histogram
* Line Graph
* Map Chart
* Pie Chart
* Heat Map

**PRICE TRENDS OVER THE YEARS**

A line graph was plotted to observe the price trends of properties over the years. The user can apply filters on

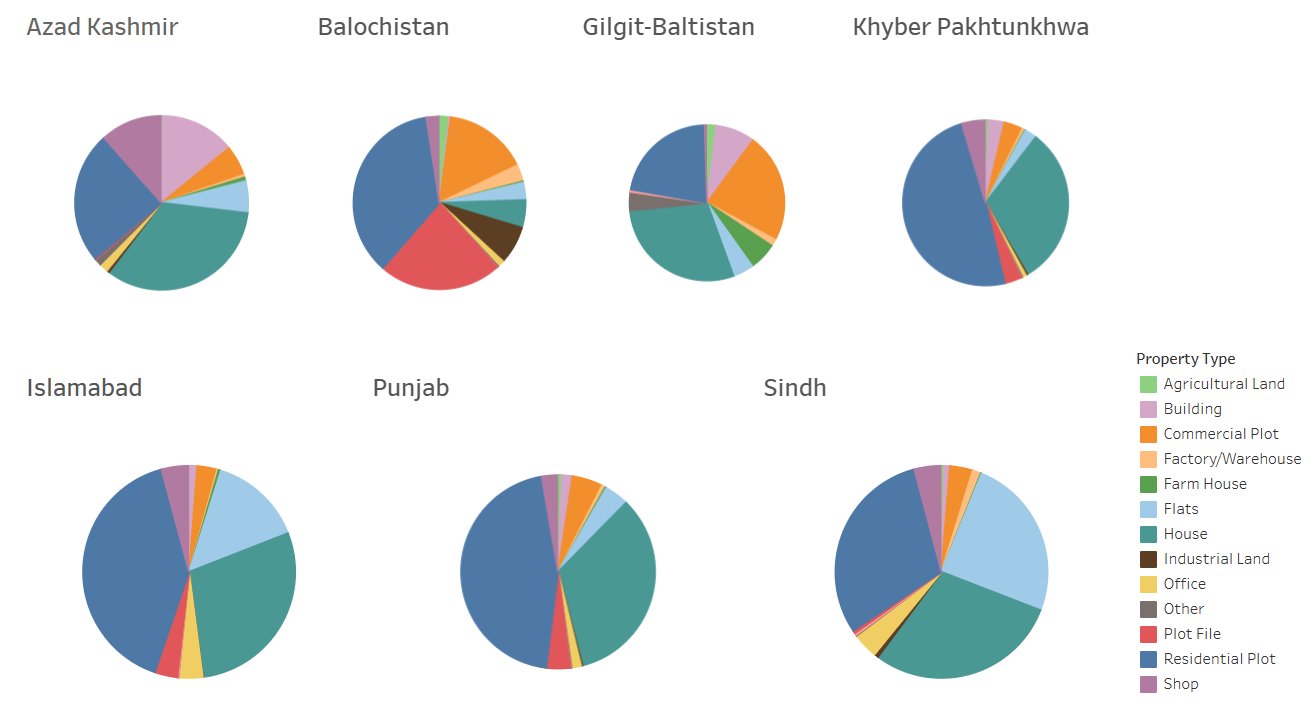
* Province
* City
* Property type
* Area

Using these filters the user can plot graphs showing the price trends of different property types in different locations. A graph showing the average price of a house, considering all the provinces is as follows:



**PROPERTY TYPE DISTRIBUTION BASED ON PROVINCES**

Pie Charts were plotted in order to observe the number of each property type available in each province. The following charts were obtained:

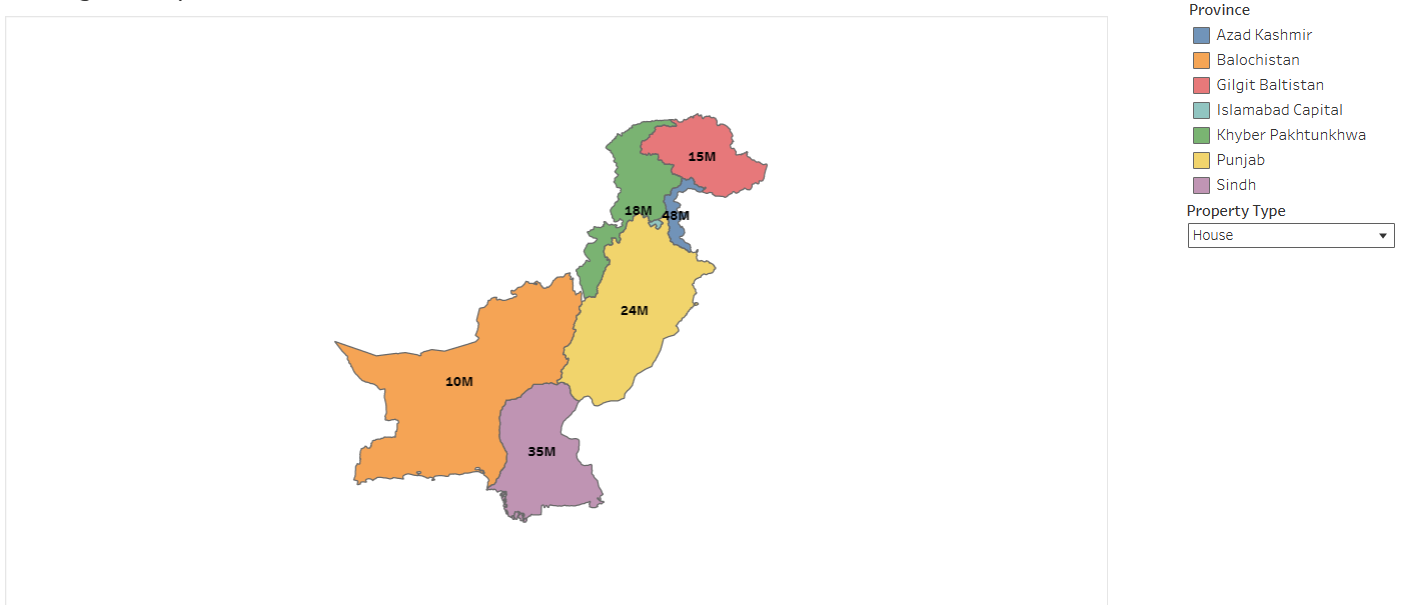


**AVERAGE PROPERTY PRICES BASED ON PROVINCES**

A Map Chart was used to display the average price of each property type in each province. The user can apply filters on

* Property Type

The chart obtained for houses in each province is as follows:

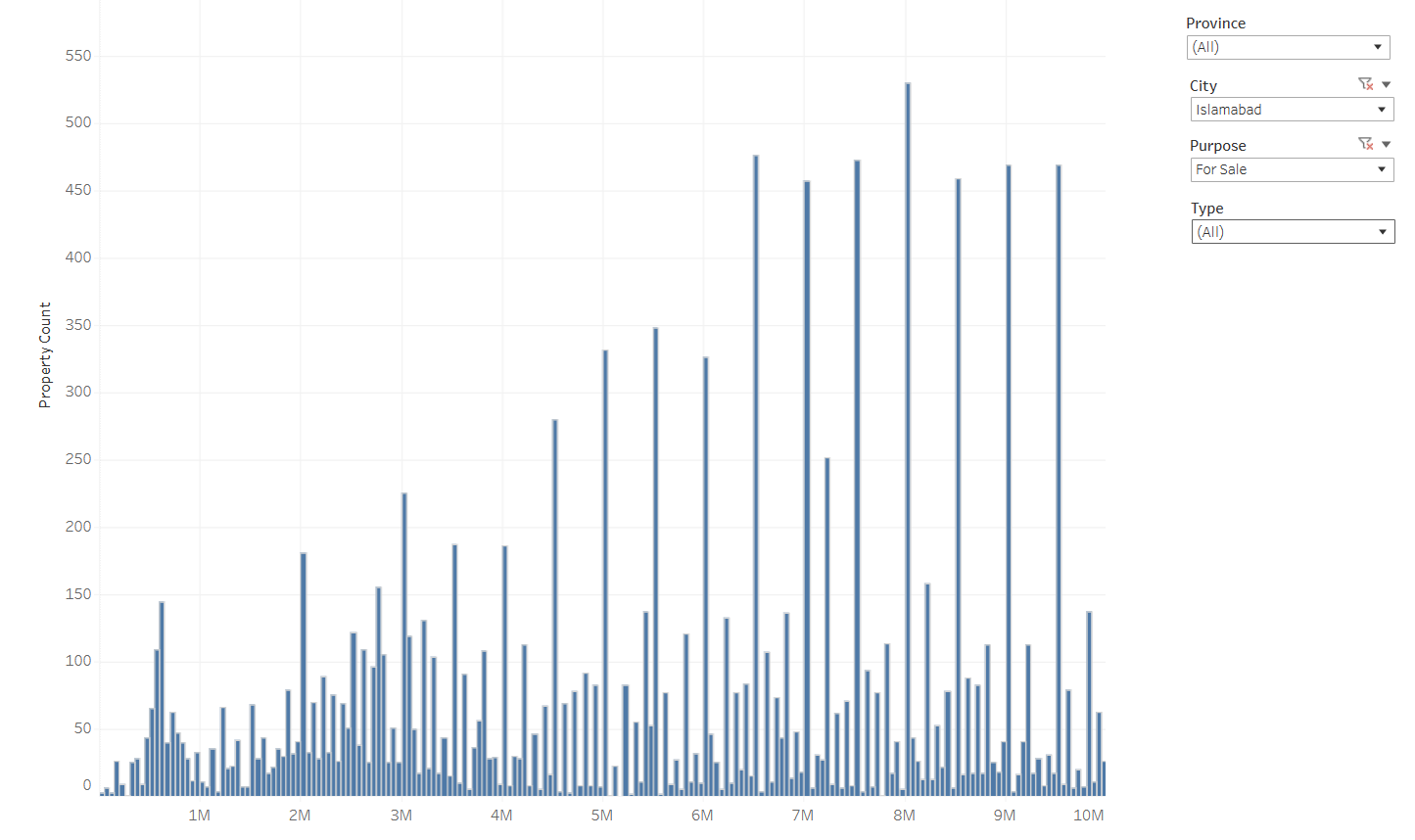


**NUMBER OF PROPERTIES AGAINST EACH PRICE RANGE**

A histogram was used to plot the number of properties against each price range. The user can apply filters on

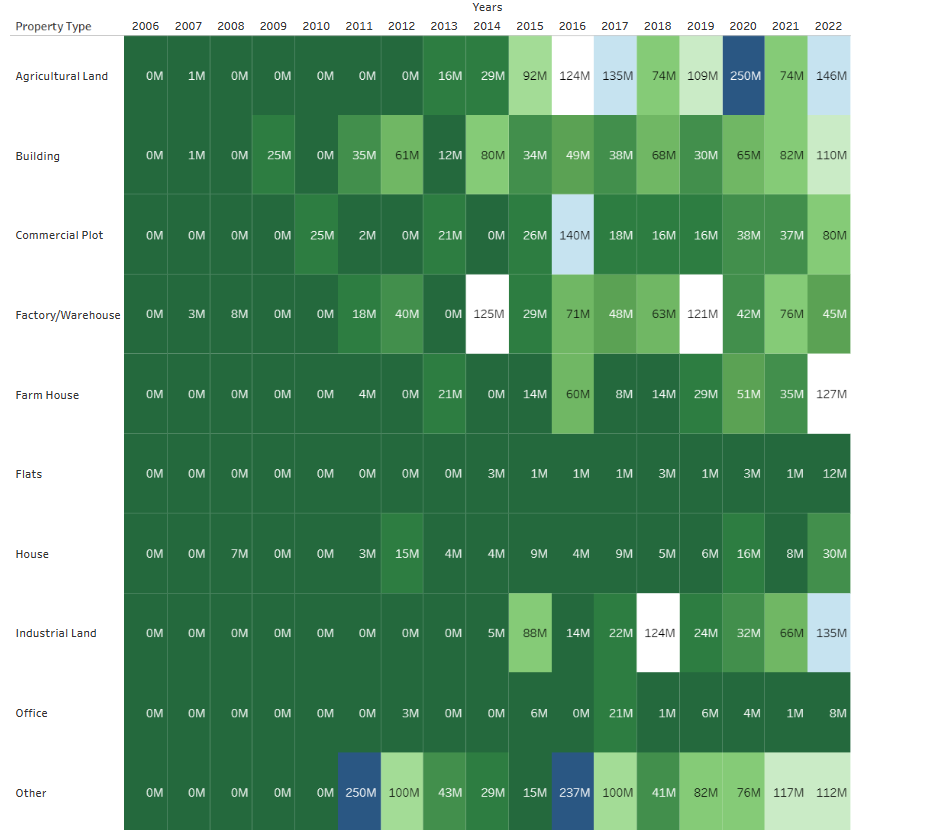
* Province
* City
* Purpose
* Type

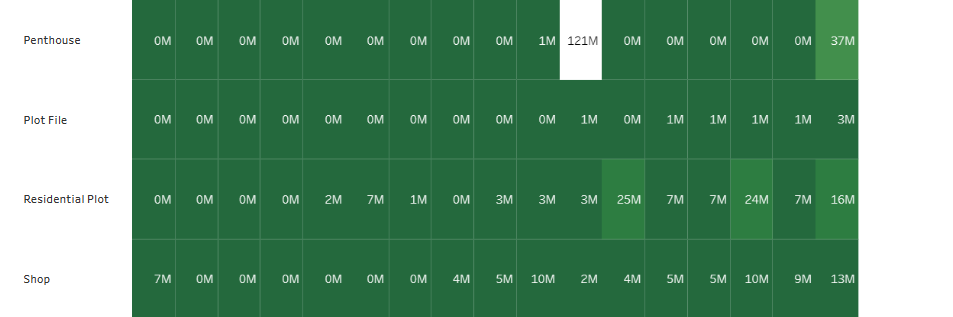
The following histogram was obtained for the total properties for sale in Islamabad



**AVERAGE PRICE FOR VARIOUS PROPERTY TYPES AGAINST YEARS**

A heat map was used to show the average prices of each property type against the years. A part of the heatmap is as follows:





**PRICE PREDICTION**

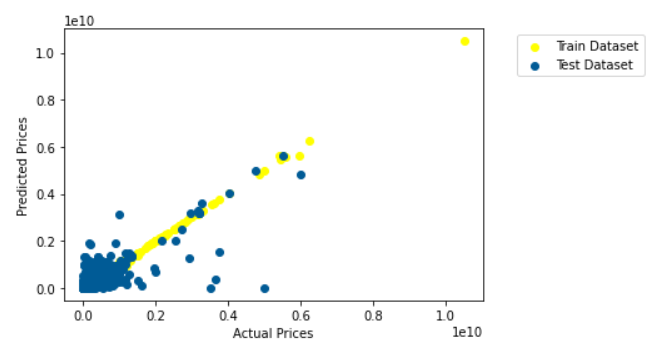
The final step in the project was to develop a model for predicting real estate prices in Pakistan. This was done using the *DecisionTreeRegression* model from the *sklearn* library. The model was trained on the pre-processed data and then used to make predictions on new, unseen data.

The model was trained and tested with two different criteria.

**SQUARED ERROR:**

The squared error criterion is a measure of the difference between the predicted values and the true values in a regression model. It is calculated by taking the difference between the predicted value and the true value for each data point, squaring the differences, and then taking the average of the squared differences.

For this criterion we achieved an accuracy of 72.2% and the following graph between the predicted and actual values was plotted for train and test datasets.



**FRIEDMAN MSE:**

The Friedman MSE (mean squared error) criterion is a variation of the squared error criterion that is used in decision tree regression. It is calculated by taking the squared difference between the predicted value and the true value for each data point, dividing the sum of the squared differences by the number of data points, and then taking the square root of the result.

For this criterion we achieved an accuracy of 70.38% and the following graph between the predicted and actual values was plotted for train and test datasets.

