REPORT OF PREDECTING REAL-ESTATES PRICES IN RIYADH

PREPARED BY

Hayat Aldhahri
Muneera Alshunaifi

DATE

14 Oct 2021

Abstract:

Saudi Arabia's residential real estate market, which is majorly concentrated across major cities such as Riyadh, Jeddah, Makkah, and Dammam Metropolitan Area (DMA), is expected to grow further in the coming years on the back of rising demand for housing units. Also, it is a non-stable market that attempts to grow or fall periodically. Therefore, the real-estate prices change dynamically in which it is hard for real-estate owners to measure the pricing criteria well. In order to facilitate the pricing process for the real-estate owners, we decided to develop a solution that aims to predict the real-estate prices based on a machine learning algorithm; specifically, we will be using a linear regression model. To do so, we will use a scraped date from the AQAR website using comprehensive analysis tools. AQAR is a website that targets people willing to sell/rent their property and people searching for properties.

Design:

This project originates from the Data Science Bootcamp (T5) to predict real estates prices in Riyadh from AQAR website: https://sa.aqar.fm/ dataset through Linear Regression. The model developed will assist people willing to sell their real-estate to find the best prices based on their needs.

Data:

The data set for the model will be scraped from the website (aqar.sa), which offers range of real estate for sales. The below table outlines the features that were used for the linear regression. The real estate type will be Villa that are located in Riyadh. The scraped data contains 5504 rows and 6 columns.

Feature	Description	Data Type
Price	The price of the villa	float
#BEDROOMS	The number of bedroom for each villa	Integer
#BATHROOMS	The number of bathrooms for each villa	Integer
SIZE(m^2)	The total building area of each villa	float
DISTRICT	The name of the district	Object
STREETWIDTH	The width of the street on which the villa is located	float

Algorithms:

Data Pre-Processing:

- Data cleaning was performed to replace null values with median for “streetwidths” and districts name were re-formatted.
- Remove duplicated values for the “Price” feature.
- Convert data type for the “District” to dummy variables to be able to feed the model.
- EDA was performed and results are shown in visualization section.

Modeling and Evaluation:

In order to perform modeling to the data, we split the dataset into 70% train and 30% test then we specify the X independent features which are all our dataset excluding the Price feature, and the Y dependent feature was Price which is our target. Moreover, we perform 3 models in order to know which model can perform better and solve our case accurately. The Main metric used to evaluate each model performance was R^2.
We started applying the baseline model which is Linear Regression then we applied Random Forest Regressor, lastly, Ridge regression was applied.
The results below shows the results for each model.

Linear Regression

Training score	Training error	CV score	Testing score	Testing error
0.86	0.14	0.75	0.84	0.1

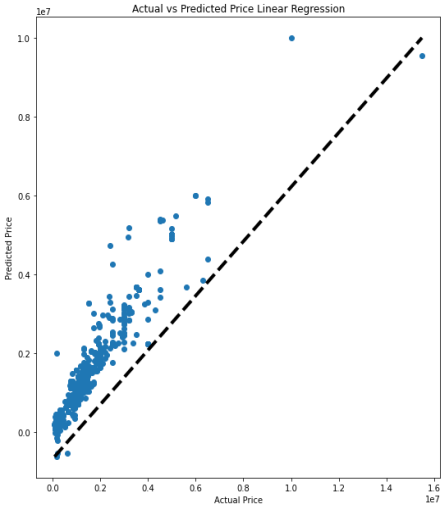
Random Forest Regressor

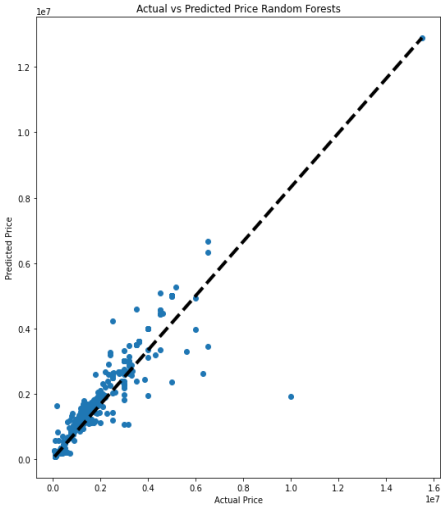
Training score	Training error	CV score	Testing score	Testing error
0.98	0.02	0.78	0.89	0.1

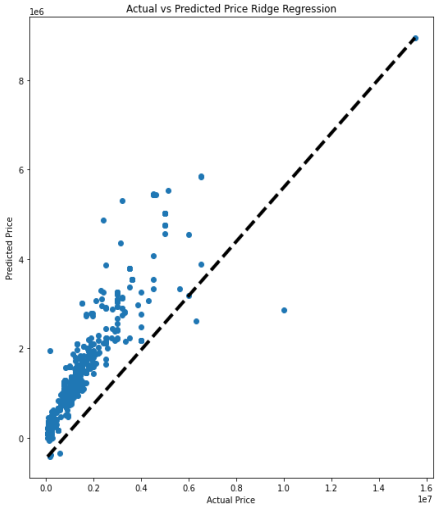
Ridge Regression

Training score	Training error	CV score	Testing score	Testing error
0.85	0.15	0.75	0.85	0.15

Fit line for each model.

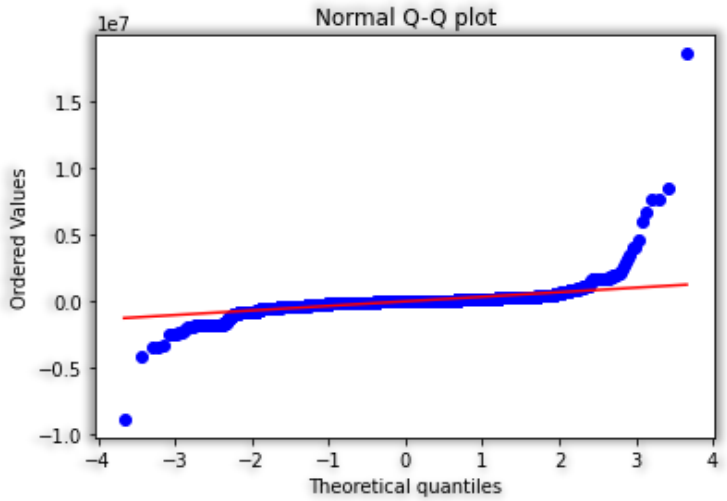




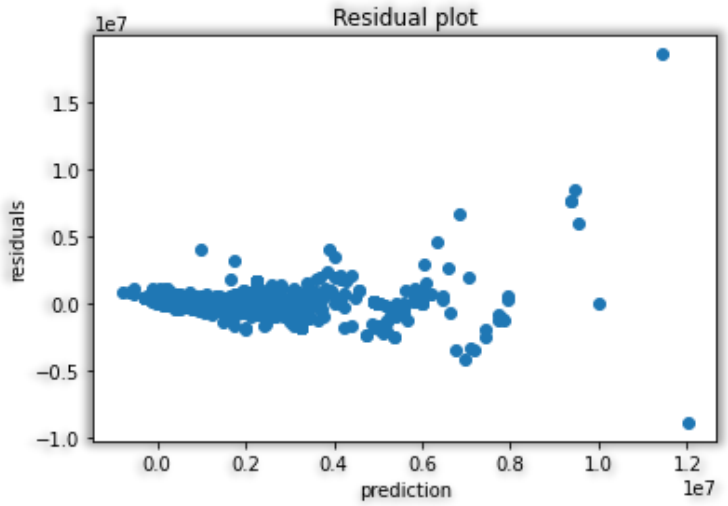


Assumptions:

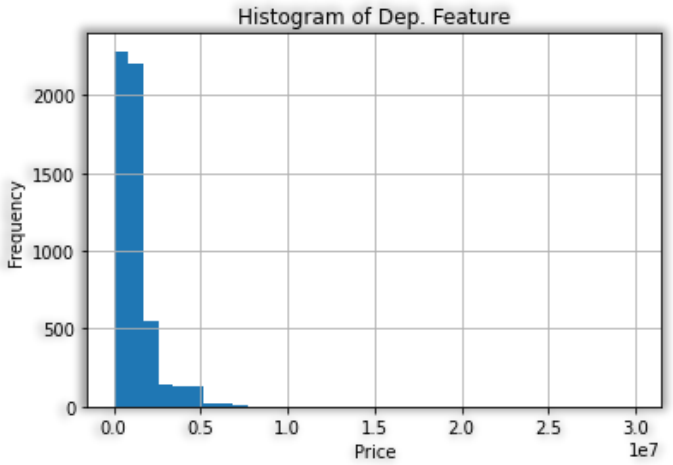
1. Regression is linear in parameters and correctly specified

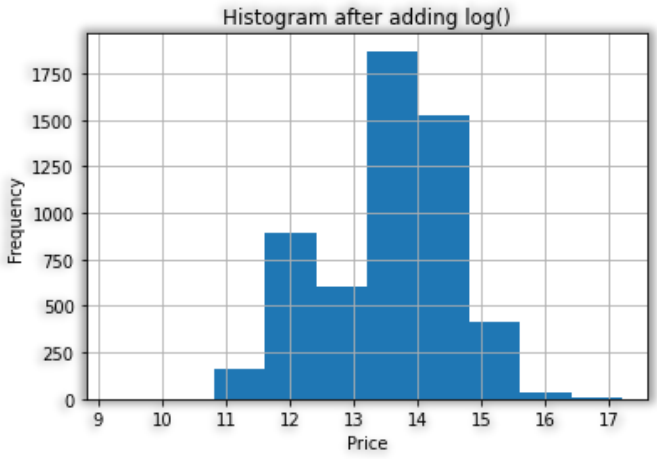


2. Residuals should be normally distributed with 0 mean



3. Error Terms must have constant variance

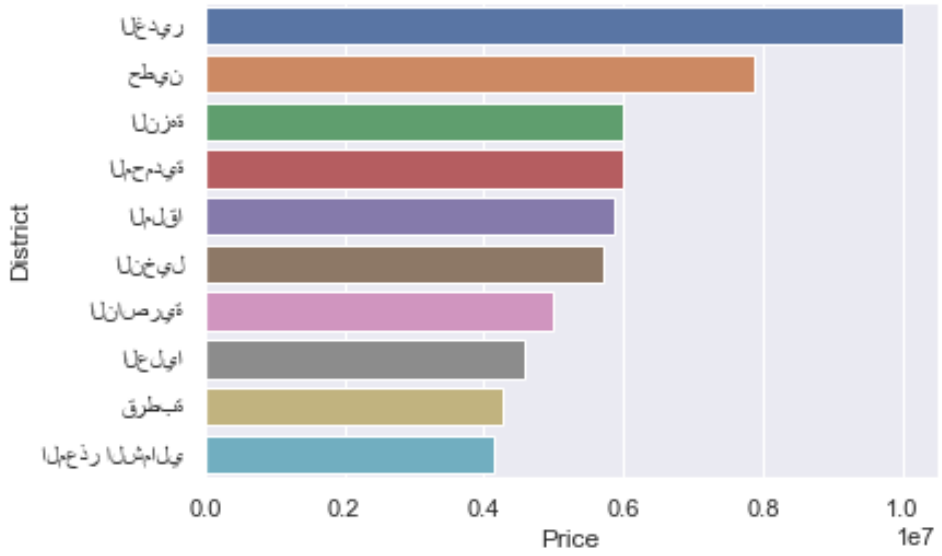




As a result, Price feature is affected the most by all extracted features. However, Random forest wins the best perfomance among all models althouogh the rest of models produce excellent results.

Visualization:

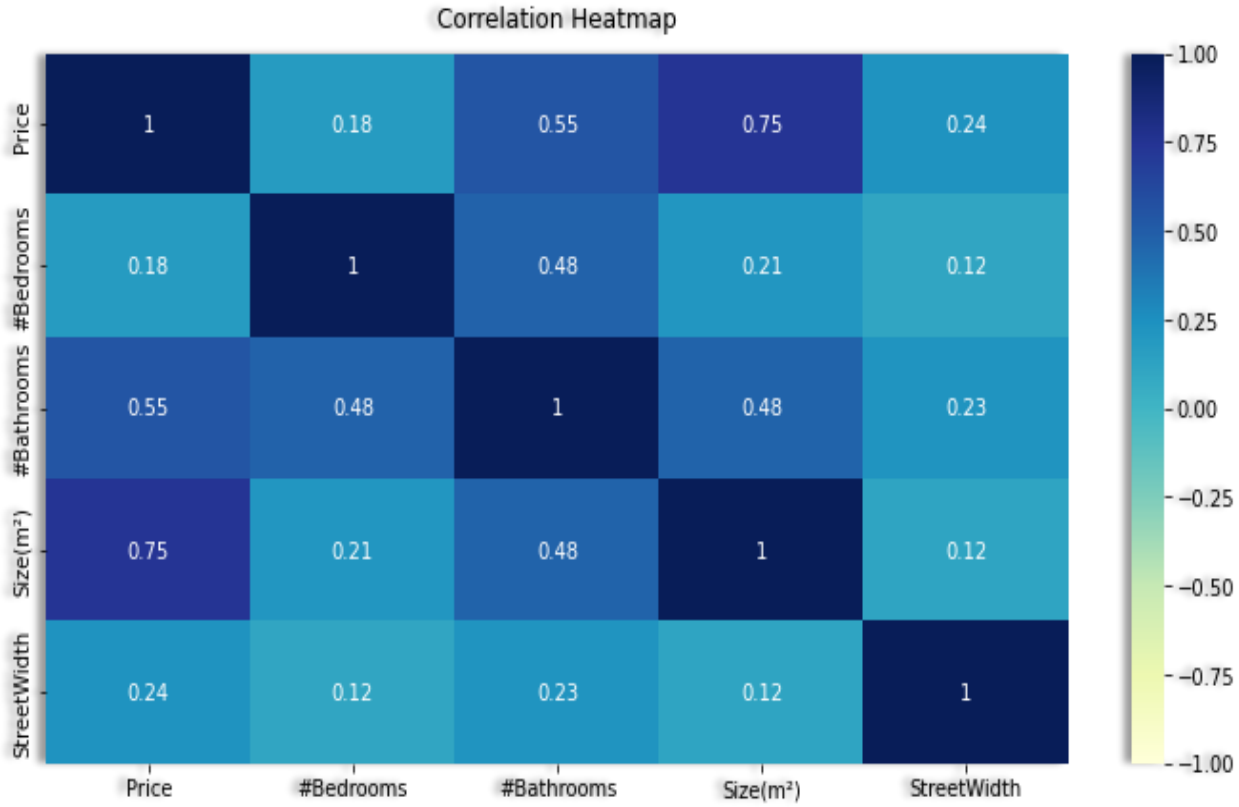
Histogram plot: Highest 10 district based on average price.



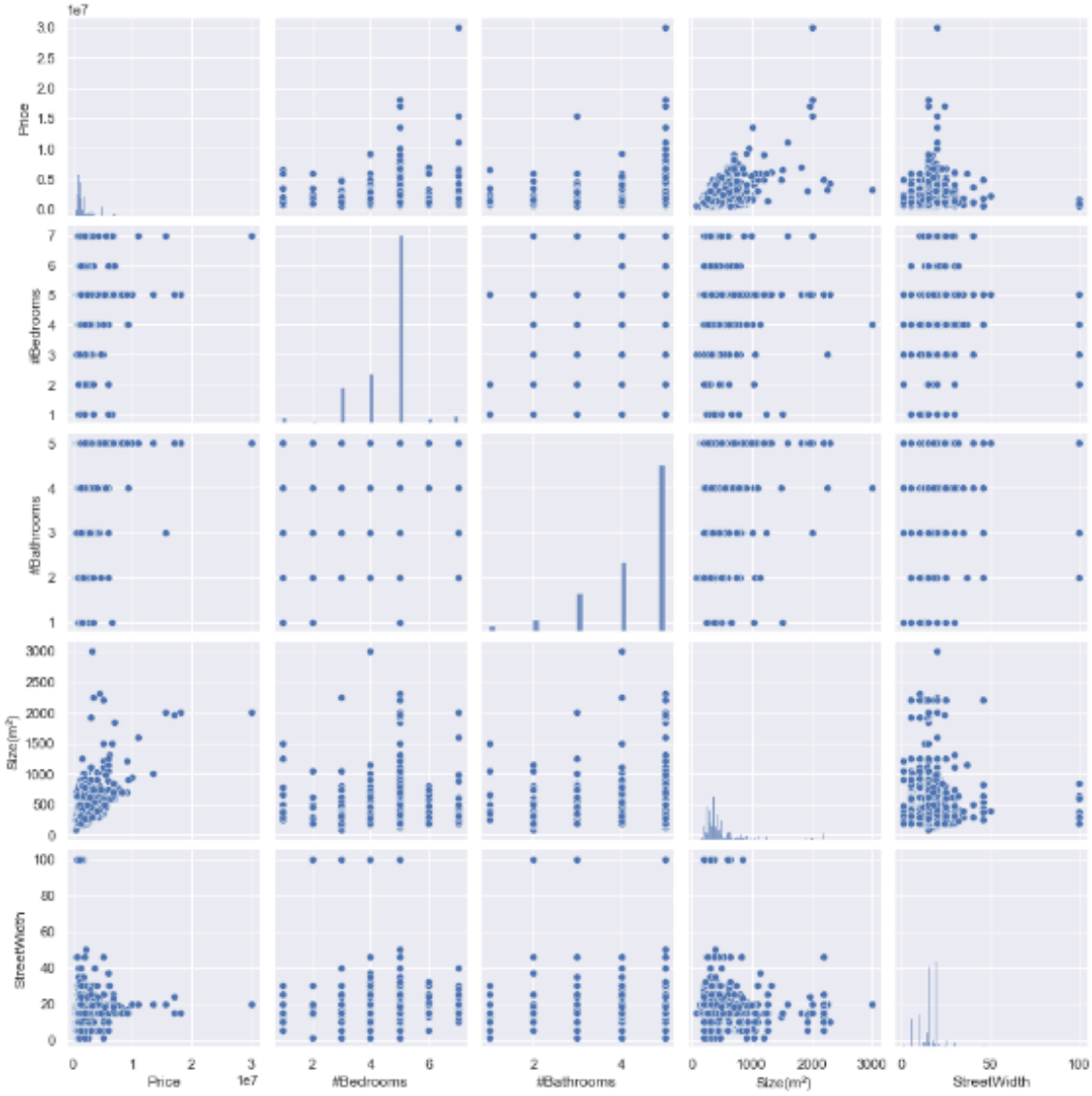
The districts from top to bottom are:

- الغدير
- حطين
- النزهة
- المحمدية
- الملقا
- النخيل
- الناصرية
- العليا
- قرطبة
- المعذر الشمالي

Heatmap: Correlation heatmap for all features. It is clearly seen that “Price” is strongly correlated to “Size (m^2)”. Also with “#Bathrooms”



Pairplot: Representing the distribution for all features.



Useful dashboard of the scraped data

Tools:

- Main language used for web scraping and model development is Python Language.
- For data exploratory and analysis, Jupyter notebook is used for python code execution.

Python libraries, such as:

- Pandas and NumPy packages for data manipulation.
- Matplotlib, seaborn library for data visualization.
- BeautifulSoup library for web scraping.
- LinearRegression from the sklearn.linear_model package for applying the linear regression model.

Visualization tools:

- Tableau

Communication

- Presentation.
- GitHub
- Tableau Dashboard