Predicting Real-Estate Price in Riyadh

Prepared by

Amal Alshahrani
Shahad Aati

Introduction

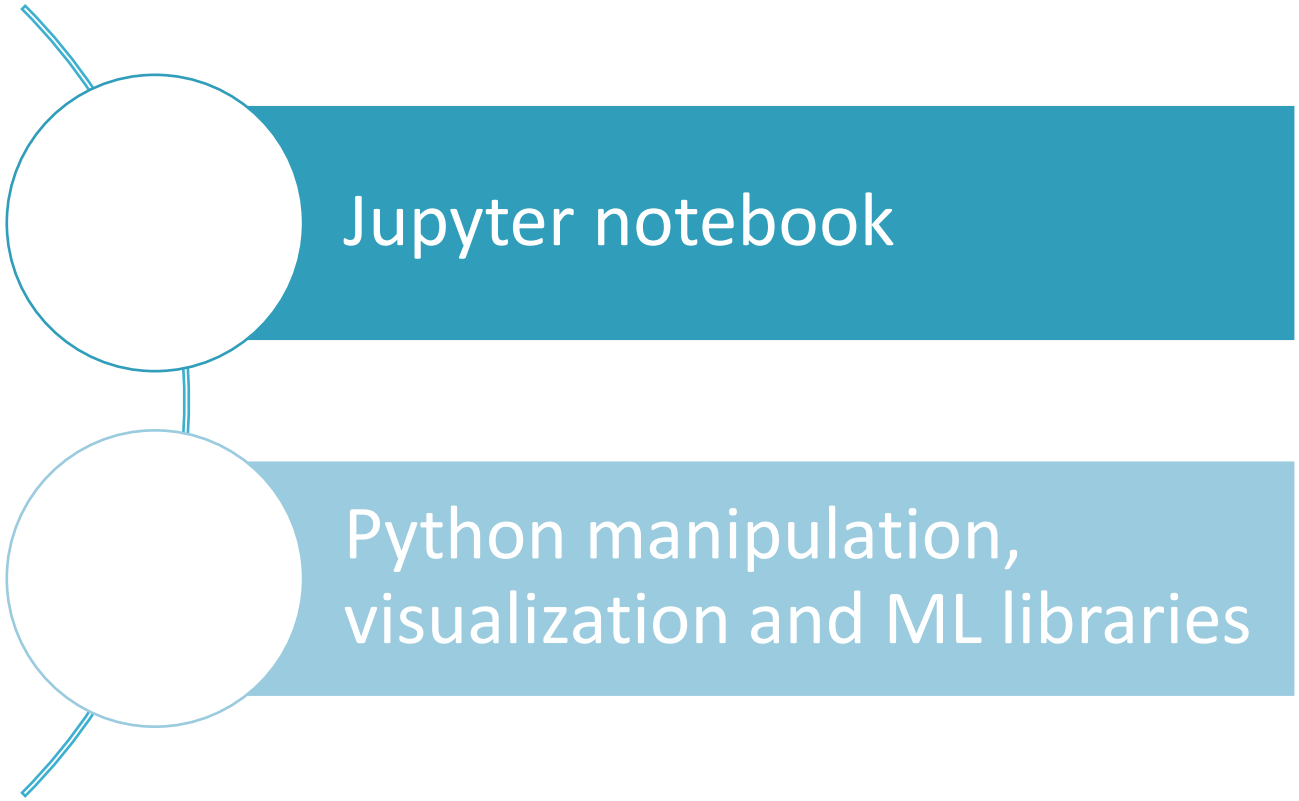
- Saudi Arabia’s residential real estate market, in the future , we expected to there will be a rising demand for housing units.
- It is a non-stable market that attempts to grow or fall periodically.
- Non-stable market and prices are very dynamic

+~60%

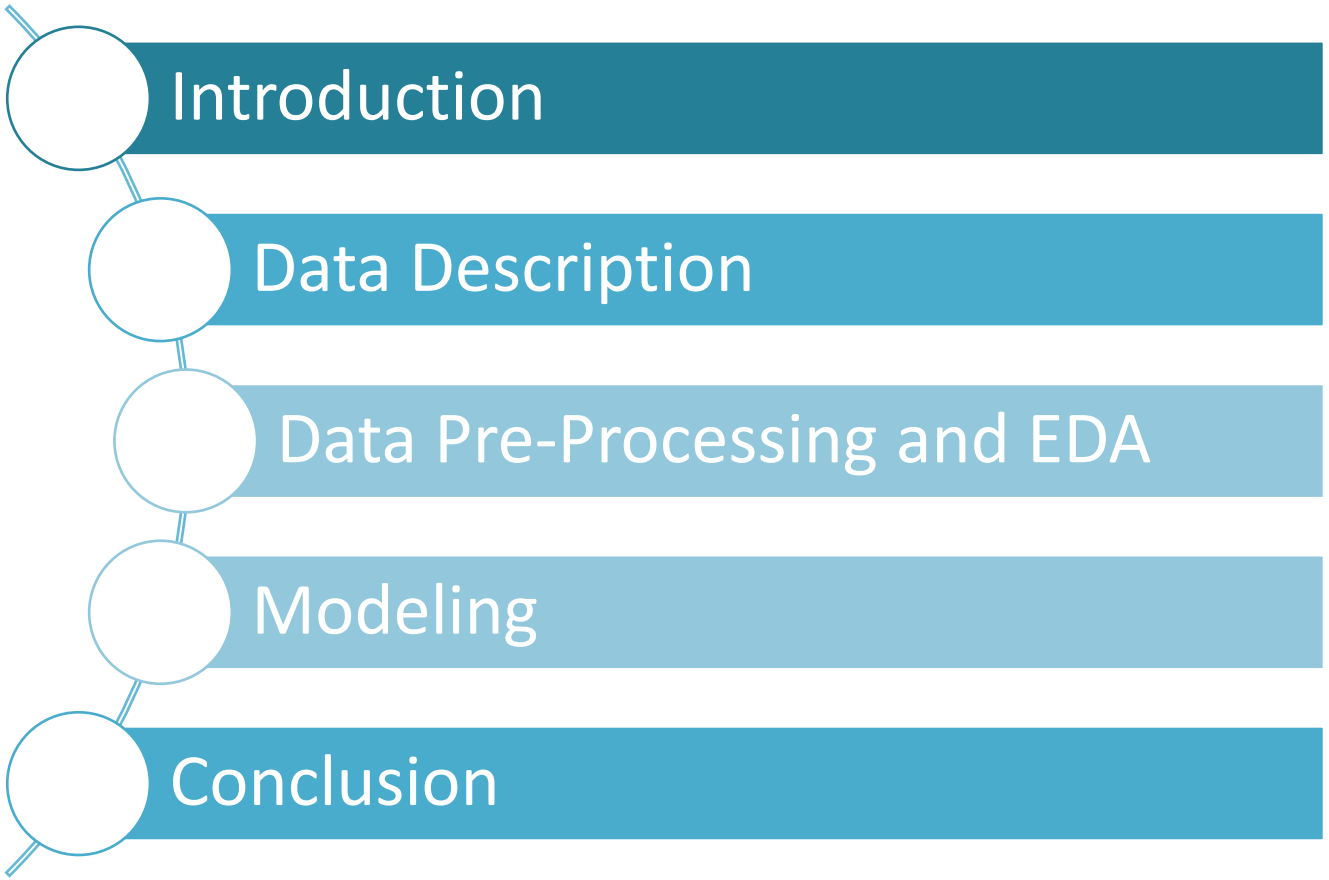
2021  **2030**

Develop a model to predict housing prices in across Riyadh’s District

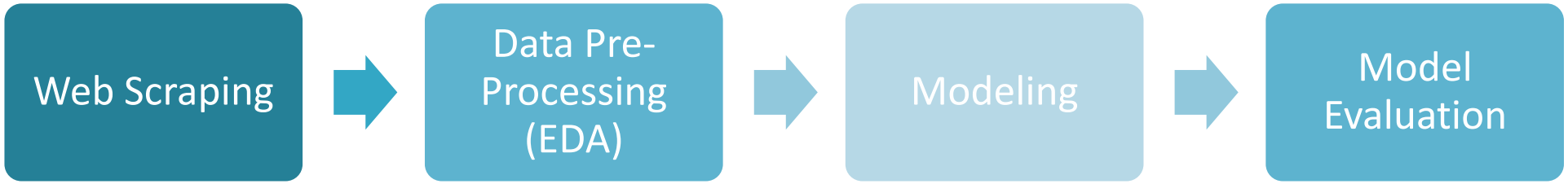
Tools used


Jupyter notebook
Python manipulation, visualization and ML libraries

Content

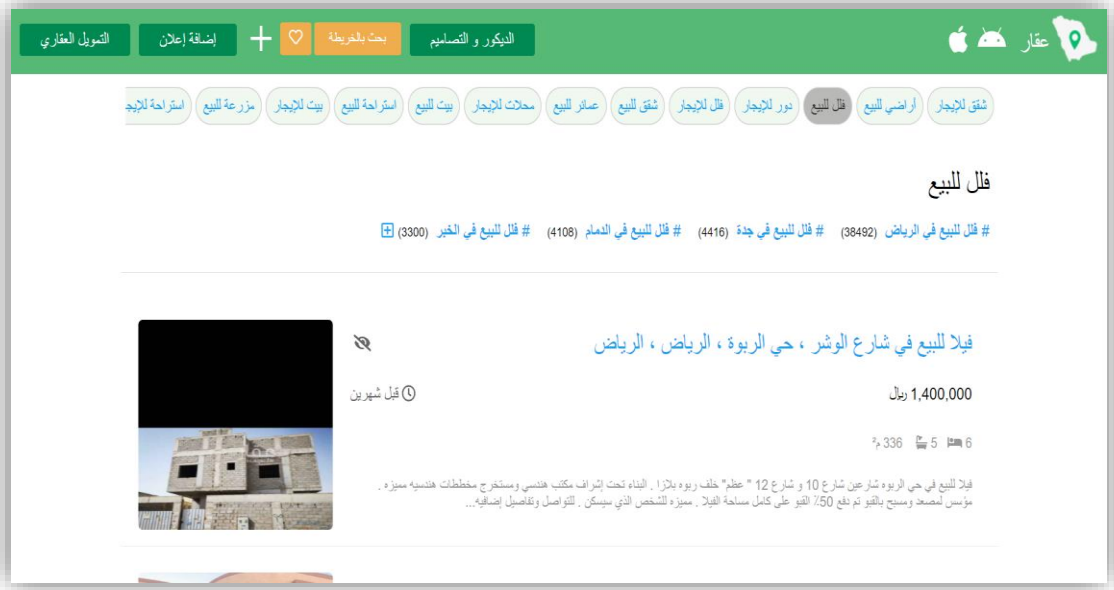

Introduction
Data Description
Data Pre-Processing and EDA
Modeling
Conclusion

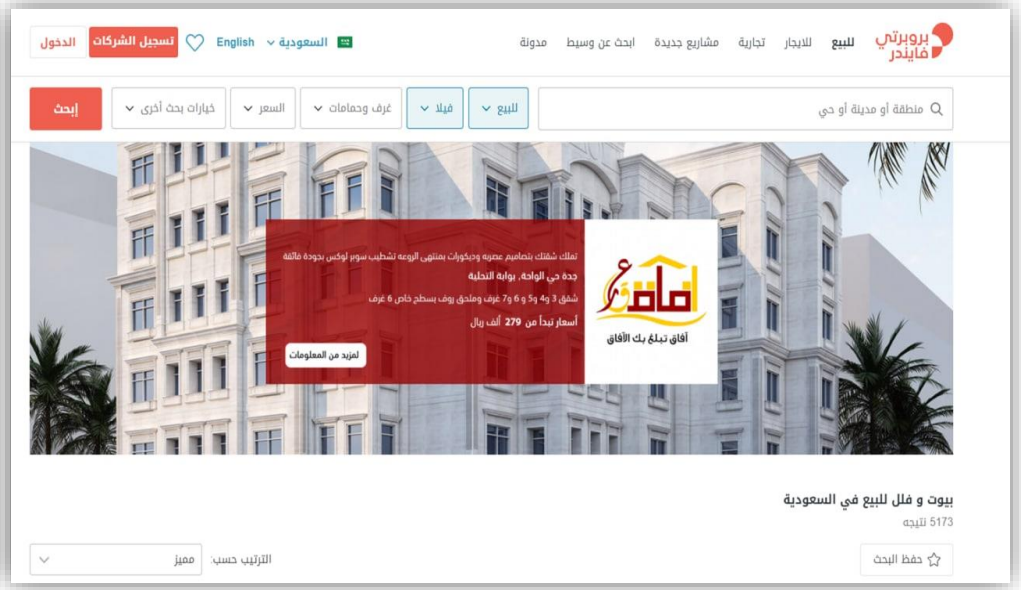
Introduction



Data Description

- **Data source:** www.sa.aqar.fm and https://www.propertyfinder.sa/
- **Real-estate type:** VILLA's at Riyadh
- **Size of Extracted data:** 1981 rows & 6 columns/features





Data Description

Extracted Features

Feature	Description	Data Type
PRICE	The price of the villa	float
NUM_OF_BEDROOMS	The number of bedrooms for each villa	float
NUM_OF_BATHROOMS	The number of bathrooms for each villa	float
SIZE (m^2)	The total building area in squared meter of each villa	float
DISTRICT	The district's name where the villa is located	Object
STREETWIDTH	The street's width where the villa is located in meter	float

Data Pre-processing and EDA

1 Data cleaning

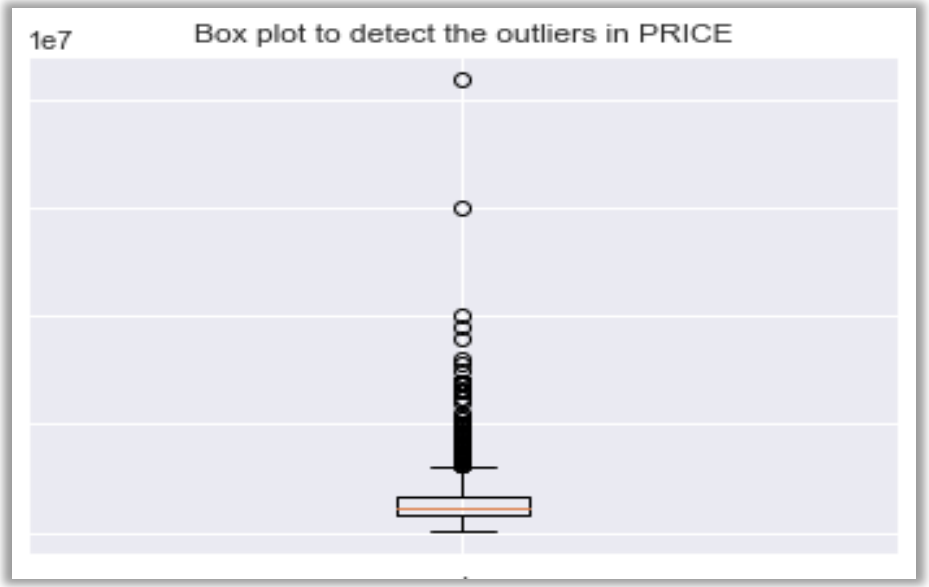
- Ensure data frame are of the same lengths.
- Remove unnecessary words from the “district” feature such as “حي".
- Remove “م^2” from area (size(m^2)) data list.
  Fill nulls “Street Width” with the zero values .
- Filter data only in Riyadh city
- Check nulls to ensure data properly cleaned.
- Converting categorical values to numerical

2 Exploratory Data Analysis - EDA

- Estimating number of houses for each districts.

EDA

- Exploring Features correlations



Insight: Price has Heavy outliers

Treating outliers: Drop only the house prices which are inconsistent ,because removing all outliers will remove data so It will not help us fitting as many values as possible.

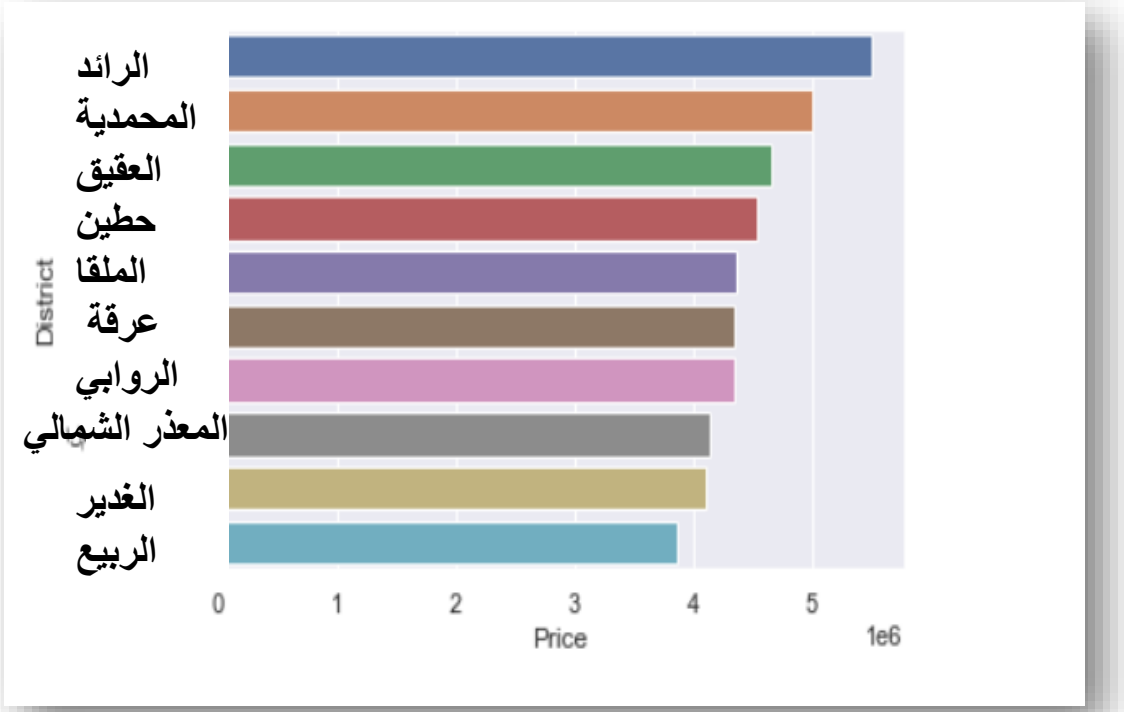
EDA

Heatmap – Features Correlation

- the "Size" feature has high positive correlation 0.67 with our target "Price“.
- - The "Num_Of_Bathrooms" feature has positive correlation 0.33 with our target "Price“.
- - The "Num_Of_Bedrooms" feature has low positive correlation 0.037 with our target "Price"

Data Cleaning

- Exploring the highest 10 districts in terms of average prices:



Modeling and Evaluation

- Linear Regression
- Polynomial Regression
- Ridge Regression
- Lasso Regression

Modeling and Evaluation

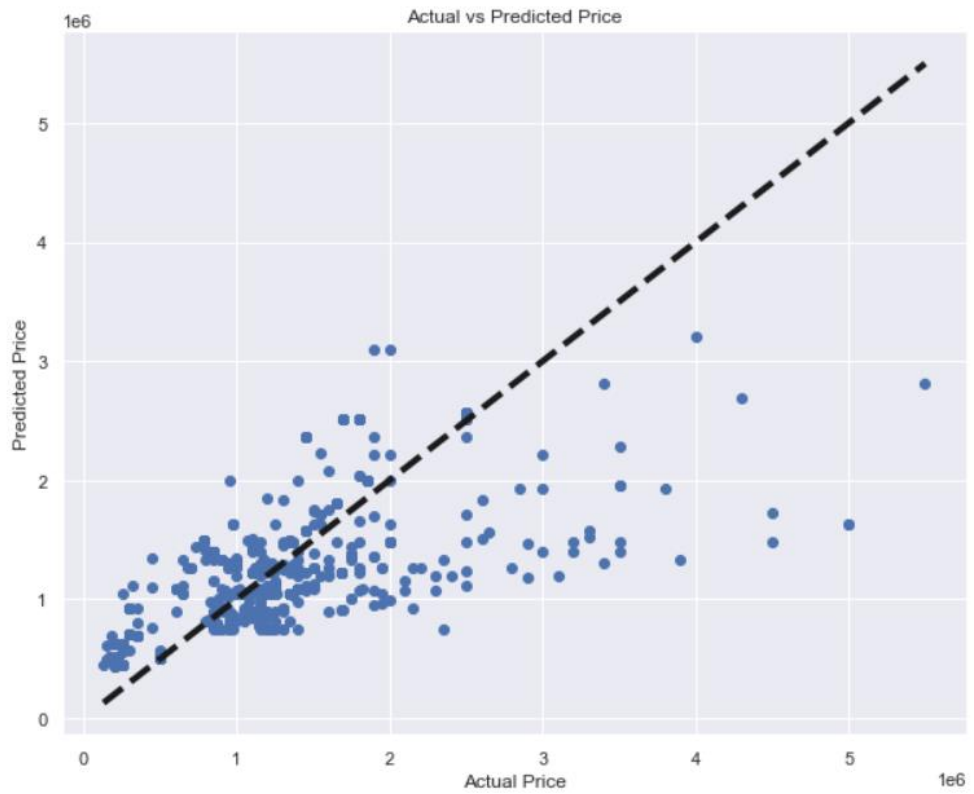
- **Split data:** data split into 3 portions: 60% for training, 20% for validation, 20% for final testing evaluation
- **Preprocessing**: Convert “District” feature to dummy variables
- **Results:**

Model	Training	Validation	R^2
Stats model(OLS)	-	-	0.96
Linear Regression	0.909	0.873	0.786
Ridge Regression	0.909	0.874	-84.541
Lasso Regression	0.909	0.875	-172.772

Polynomial Regression	D2	D3	D4	D5
Training	1.0	0.999	0.992	0.94
Validation	1.0	0.999	0.992	0.94

Modeling and Evaluation

- Fit line for each model



Linear Regression

Conclusion

House price is affected by: # of bedrooms, # of bathrooms, Size, Street width and District

Ridge model wins the best performance model among all models

Future work

- Automate aqar and propertyfinder website.
- Enhance the performance of the estimate.

Thank you!

Any Questions?