PREDICTING HOUSE PRICES WITH LINEAR REGRESSION

by Ruba Alnashwan

METHODOLOGY

- Proplem understanding
- Data collection
- Data cleaning
- Exploratory data analysis (EDA)
- Feature Engineering and selection
- Data modeling

Proplem understanding

Overview:
In this project, we will use data for villas offered for sale in the city of Riyadh from the (Aqar) website, which provides villas for sale and their prices.

Problem statement:
The objective of this project is to predict the price of the villa by providing some features for each vila by using supervised linear regression model

Scope:
The scope of this project was the villas for sale in the areas of Riyadh, whether old or new villas. The dataset contains 2099 rows × 10 columns.

Data cleaning

- check nulls

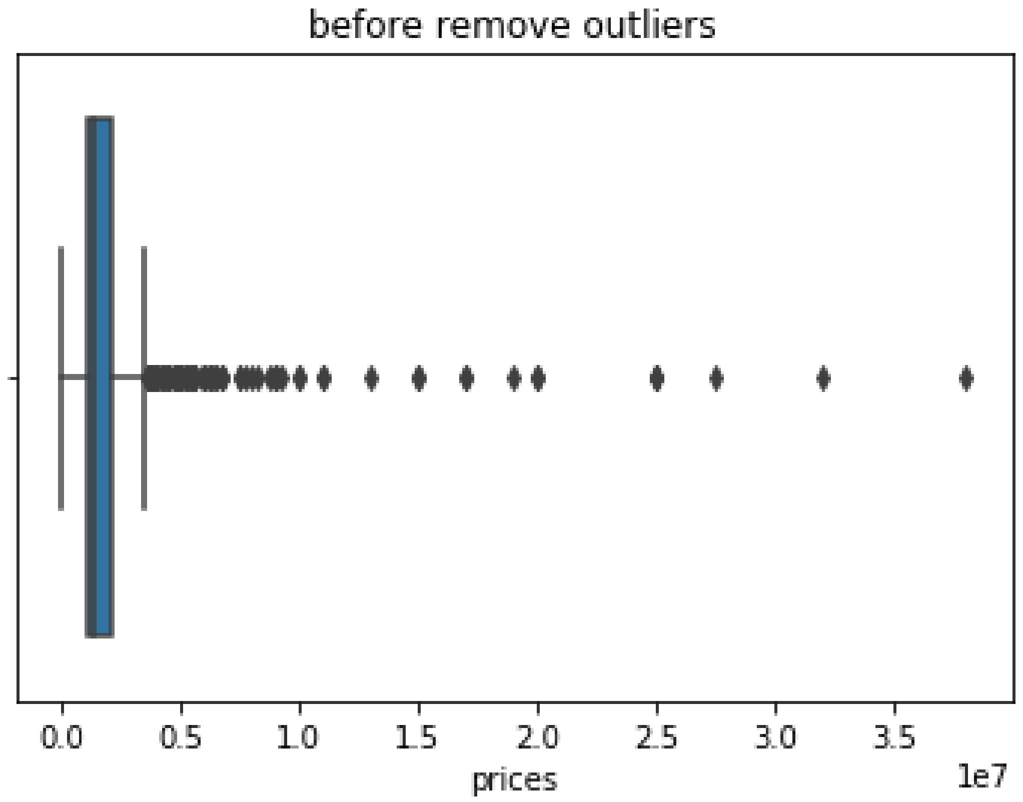
  I filled in the null values in the oldness feature with zero. My notes. If there is no oldness, the villa will be new and filled the street width, number living rooms with mode and median.

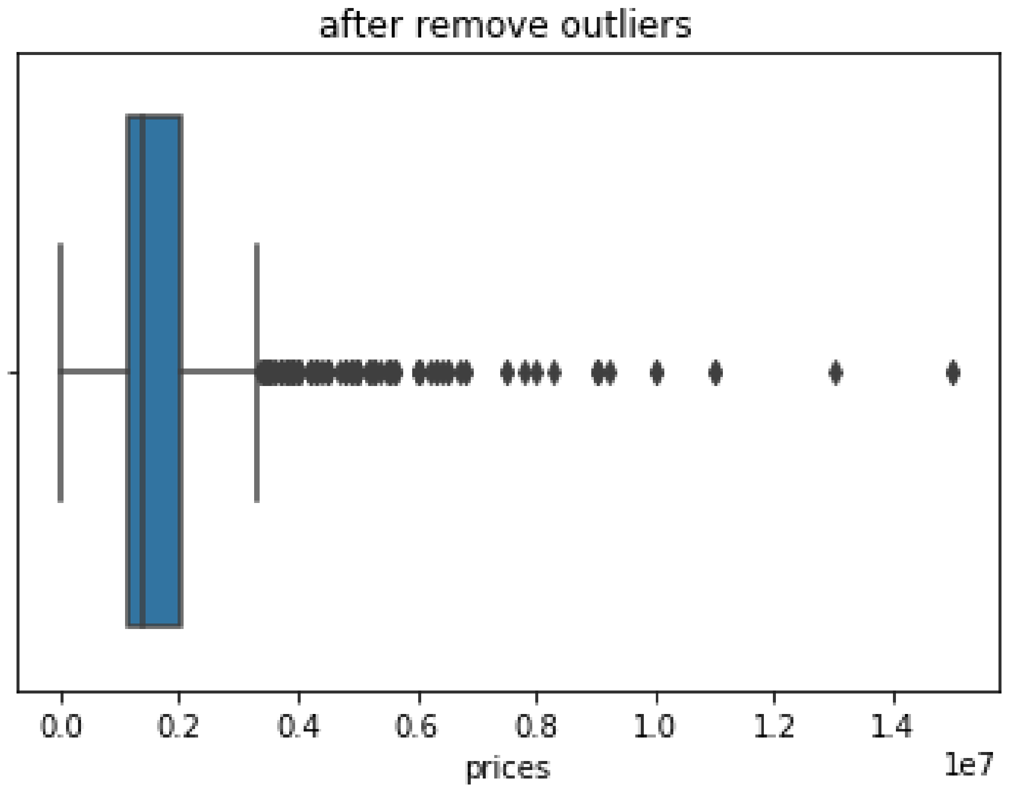
- check duplicates

Data cleaning

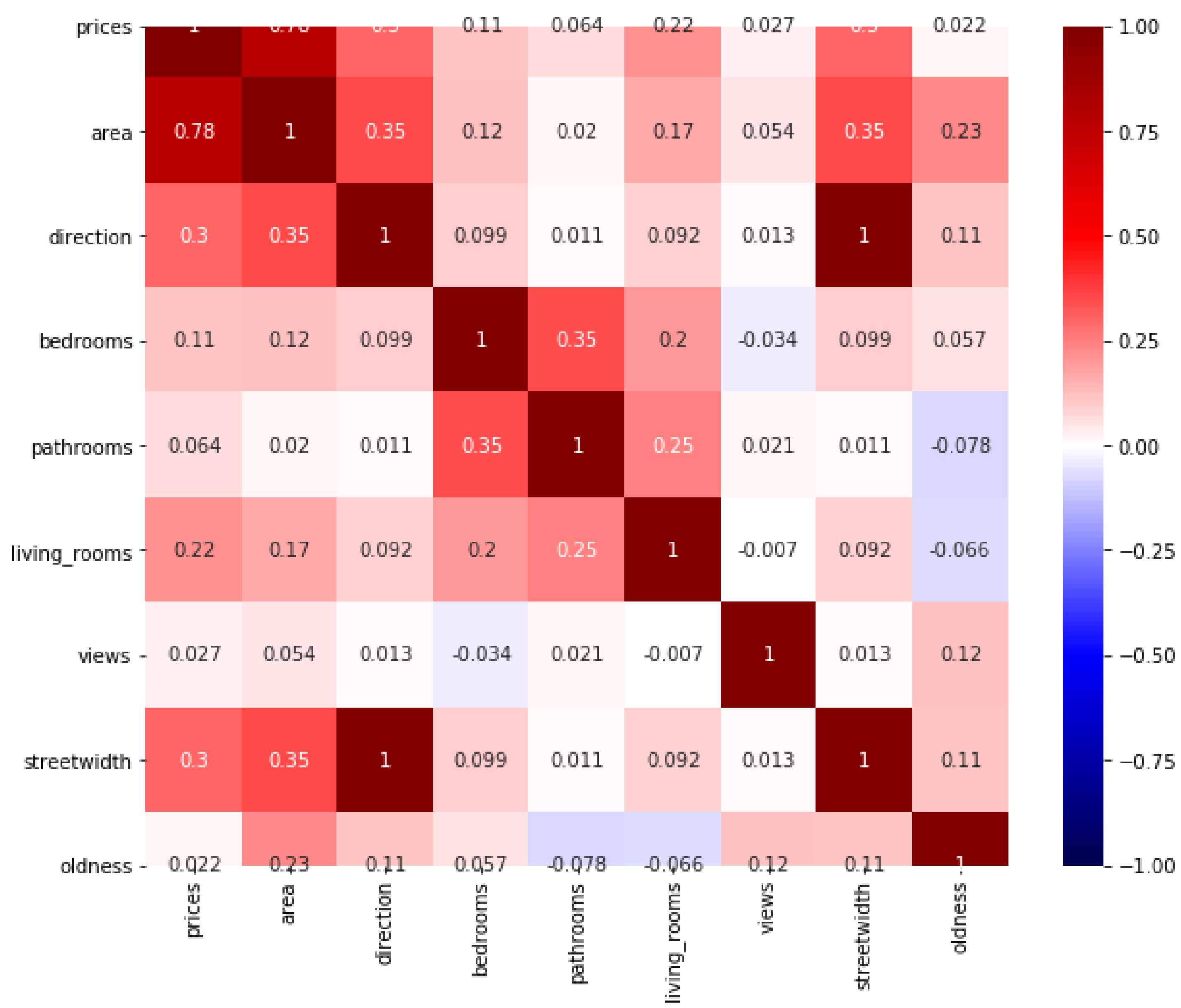
- replace some character and arabic words like (شمال الرياض ، غرب الرياض) and (م٢، سنة)

- check outliers using EDA and remove





Exploratory data analysis (EDA)

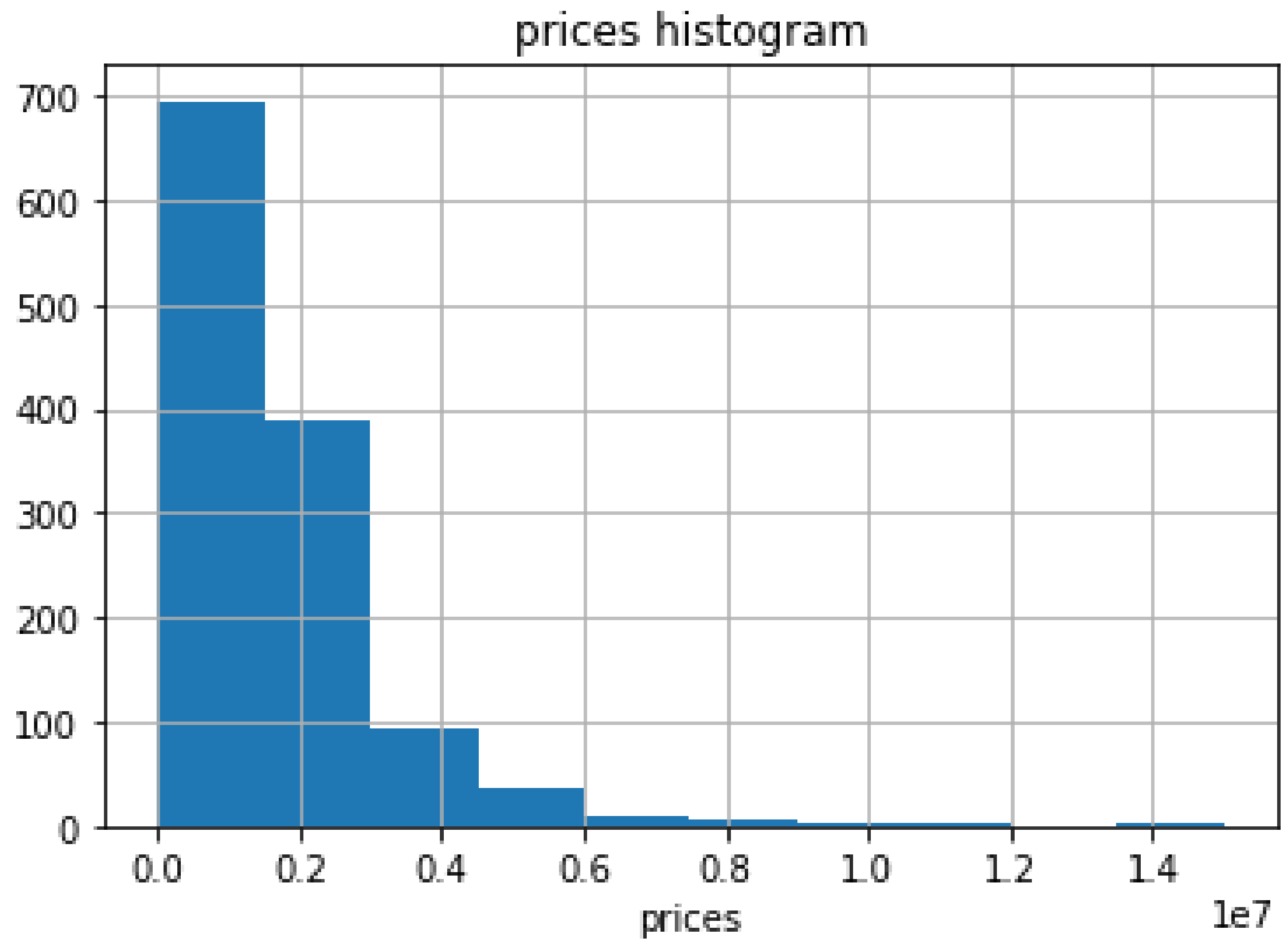


	prices	area	direction	bedrooms	pathrooms	living_rooms	views	streetwidth	oldness
prices	1	0.78	0.3	0.11	0.064	0.22	0.027	0.3	0.022
area	0.78	1	0.35	0.12	0.02	0.17	0.054	0.35	0.23
direction	0.3	0.35	1	0.099	0.011	0.092	0.013	1	0.11
bedrooms	0.11	0.12	0.099	1	0.35	0.2	-0.034	0.099	0.057
pathrooms	0.064	0.02	0.011	0.35	1	0.25	0.021	0.011	-0.078
living_rooms	0.22	0.17	0.092	0.2	0.25	1	-0.007	0.092	-0.066
views	0.027	0.054	0.013	-0.034	0.021	-0.007	1	0.013	0.12
streetwidth	0.3	0.35	1	0.099	0.011	0.092	0.013	1	0.11
oldness	0.022	0.23	0.11	0.057	-0.078	-0.066	0.12	0.11	1

Observations:

This heat map shows that there is no strong relationship between the dependent variable and features except for the area.

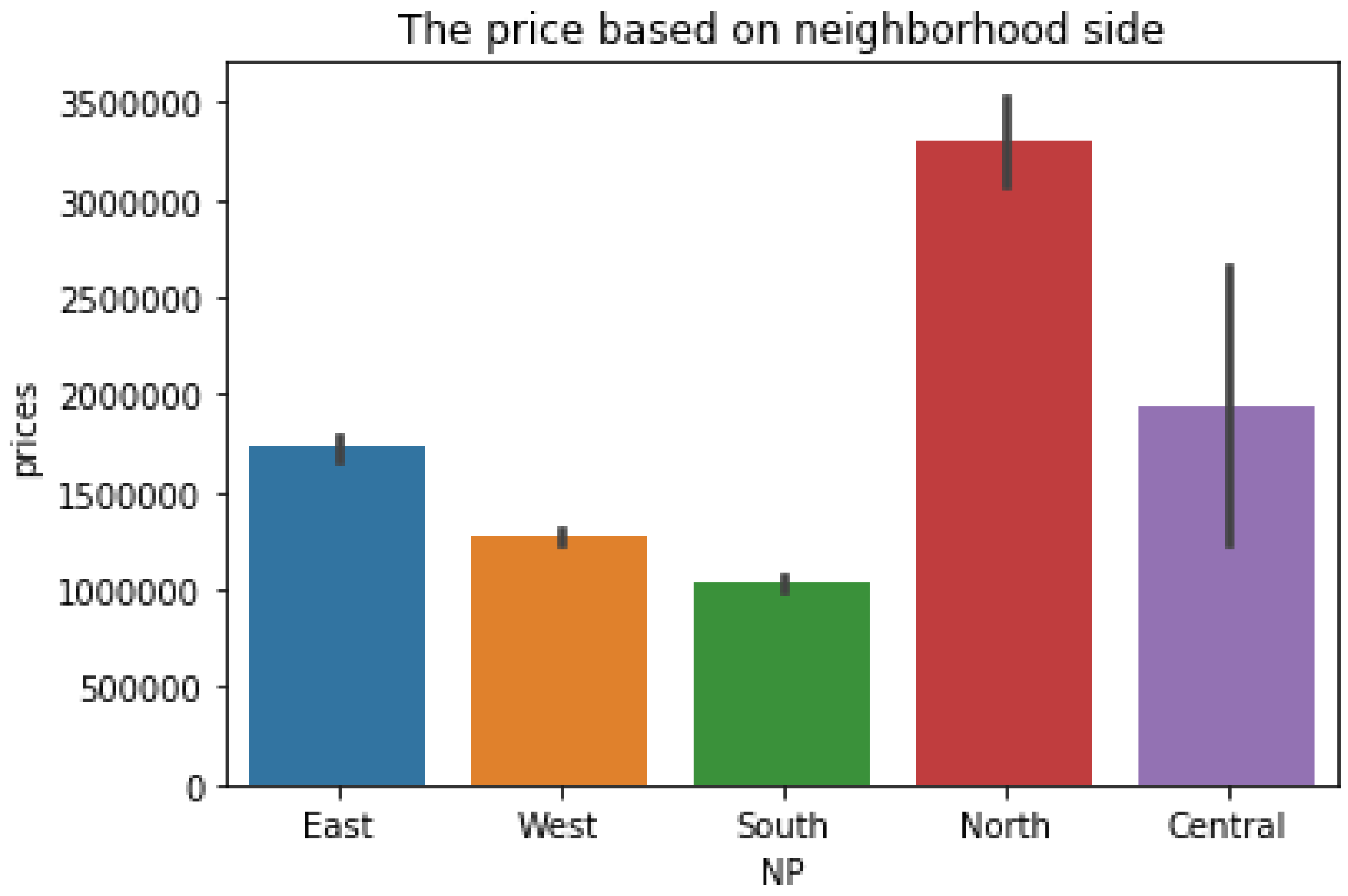
Exploratory data analysis (EDA)



Observations:

In this histogram shows the most of the density lies between 1M and 2M, but there appears to be a lot of outliers on the pricier side.

Exploratory data analysis (EDA)



Observations:

The par plot shows the highest prices are located in the north

Feature Engineering and selection

Principal component analysis (PCA) to reduce the data

Polynomial Features But the results were not good

Dummies for Neighborhood Feature ,the model has been greatly improved

Scaling: MinMaxScaler, StandardScaler

Feature selection: lasso, ridge

Data modeling

MAIN METRIC USED FOR EVALUATION : R2, RMSE

MODEL USED: LINEAR REGRESSION

basline	one feature power of 2	dummies	log target	Polynomial Features	PCA
TRAINING SCORE: 0.37746	TRAINING SCORE: 0.37963	TRAINING SCORE: 0.60967	TRAINING SCORE: 0.60571	TRAINING SCORE: 0.56041	TRAINING SCORE: 0.6097
VALIDATION SCORE: 0.5955	VALIDATION SCORE: 0.59218	VALIDATION SCORE: 0.56537	VALIDATION SCORE: 0.59849	VALIDATION SCORE: 0.63017	VALIDATION SCORE: -0.10154

Data modeling

MAIN METRIC USED FOR EVALUATION : R2, RMSE

MODEL USED: LINEAR REGRESSION

basline

TRAINING SCORE:
0.37746

VALIDATION SCORE:
0.5955

one feature power of 2

TRAINING SCORE:
0.37963

VALIDATION SCORE:
0.59218

dummies

TRAINING SCORE:
0.60967

VALIDATION SCORE:
0.56537

log target

TRAINING SCORE:
0.60571

VALIDATION SCORE:
0.59849

Polynomial Features

TRAINING SCORE:
0.56041

VALIDATION SCORE:
0.63017

PCA

TRAINING SCORE:
0.6097

VALIDATION SCORE:
-0.10154

CONCLUSION

result the best model :

Training score: 0.60555

Validation score: 0.59771

Difficulties:

Real estate is always changing

Prices in each neighborhood are different

Thank you

Any questions?