

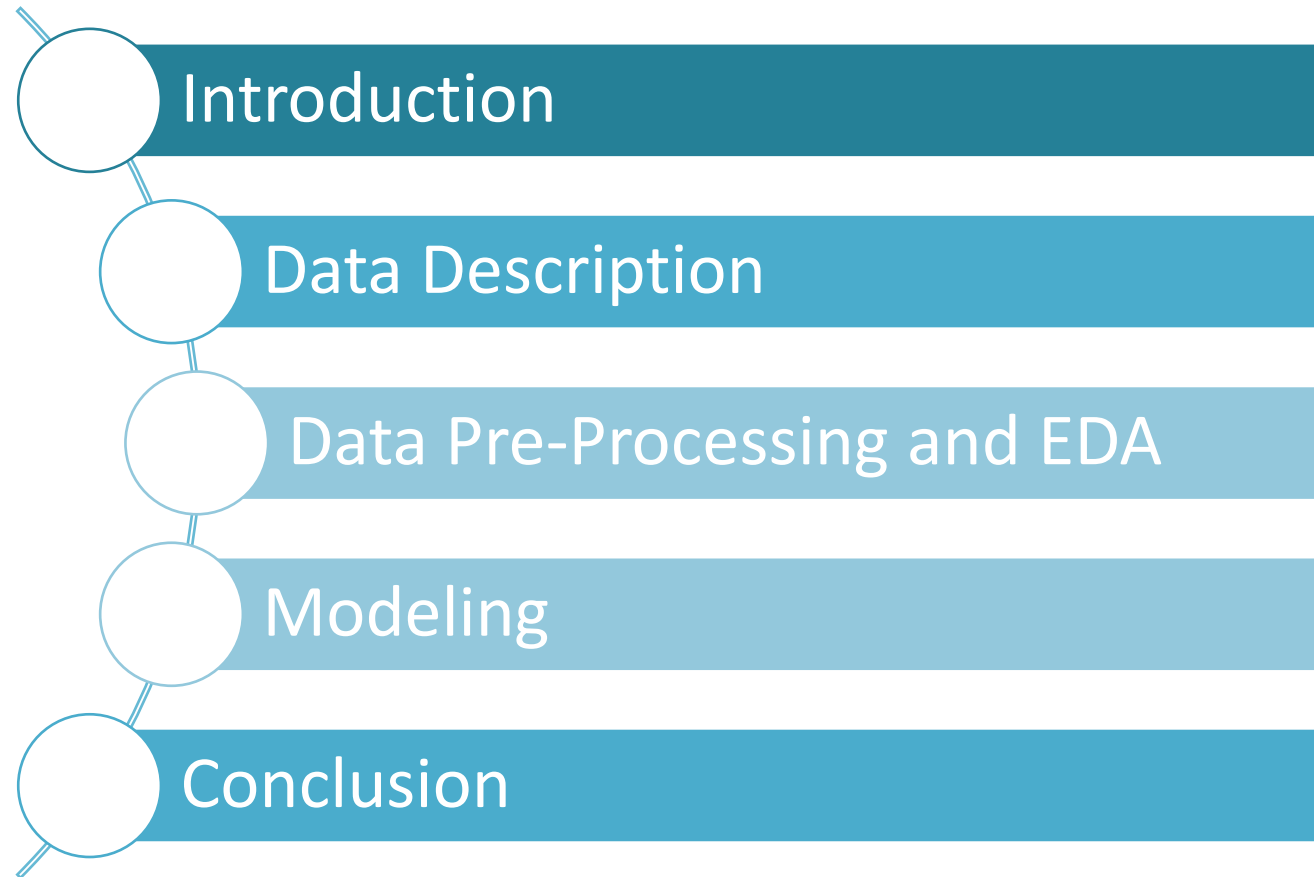
Predicting Real-Estate Price in Riyadh

Prepared by

Hayat Aldhahri
Muneera Alshunaifi



Content




Introduction

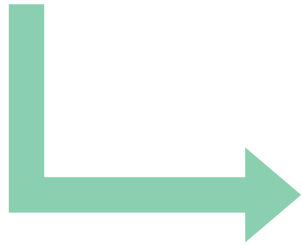
- Saudi Arabia's residential real estate market, is expected to grow further in the coming years on the back of rising demand for housing units.
- It is a non-stable market that attempts to grow or fall periodically.
- Real-estate prices change dynamically in which it is hard for real-estate owners to measure the pricing criteria well.



Introduction



- Saudi Arabia housing market is expected to grow in the upcoming years.
- Non-stable market and prices are very dynamic 



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

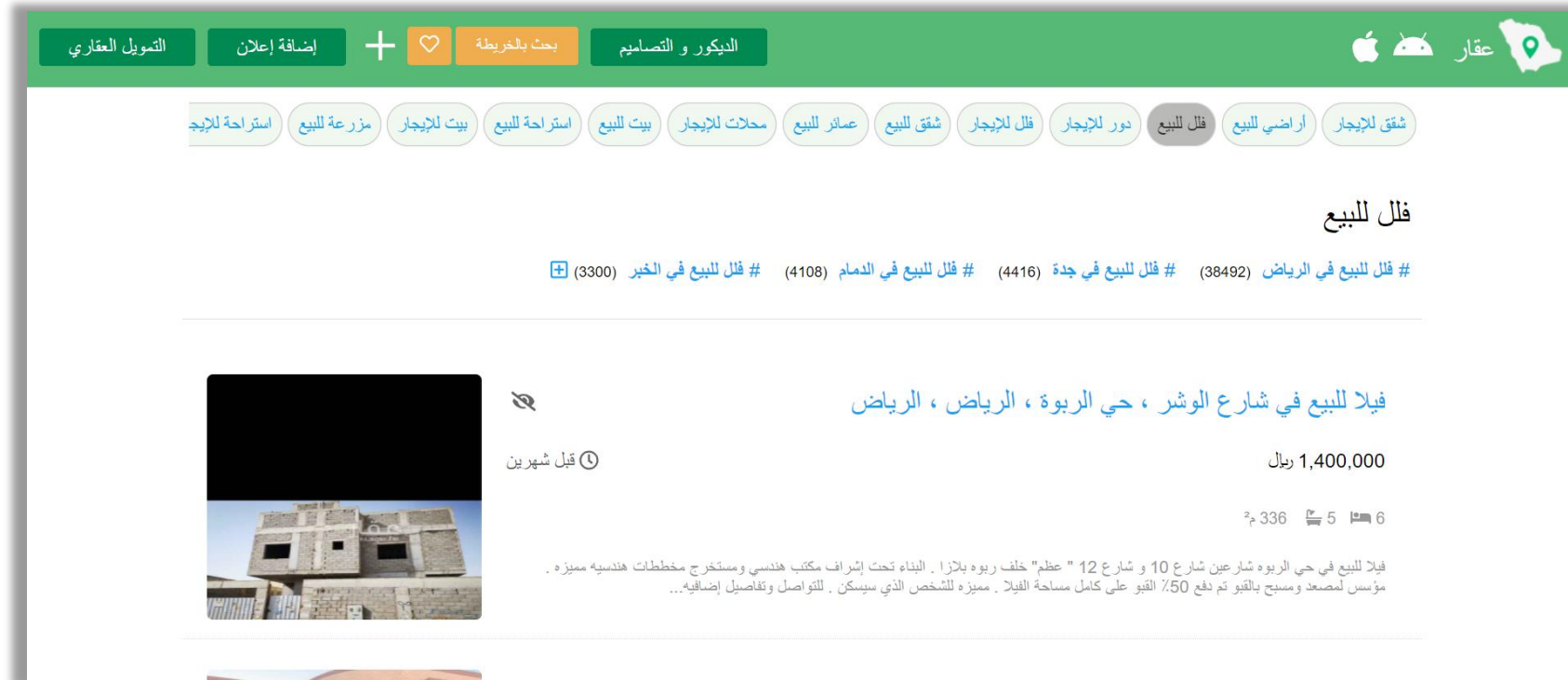
Introduction



Data Description



- **Data source:** www.sa.aqar.fm
- **Real-estate type:** VILLA's at Riyadh
- **Size of Extracted data:** 5504 rows & 6 columns/features



Data Description



Extracted Features

Feature	Description	Data Type
Price	The price of the villa	float
#BEDROOMS	The number of bedrooms for each villa	Integer
#BATHROOMS	The number of bathrooms for each villa	Integer
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.
- Fill nulls “Street Width” with the median.
- Filter data only in Riyadh city
- Remove unnecessary words from the “district” feature such as “حي” & “الرياض”.
- Remove “م²” from size(m²) data list.
- Check nulls on the null and ensured data properly cleaned.

2 Exploratory Data Analysis - EDA

- Estimating number of houses for each districts.
- Exploring the highest districts in terms of average prices.
- Exploring Features correlations

Data Cleaning



Price	#Bedrooms	#Bathrooms	Size(m²)	District
250000.0	4	2	100.0	حي الخالدية - الرياض
720000.0	5	5	420.0	حي بدر - الرياض
290000.0	5	4	88.0	حي جرير - الرياض
240000.0	5	3	77.0	حي الديرة - الرياض
180000.0	1	2	81.0	حي الرفيعة - الرياض
...
970000.0	3	3	360.0	حي الجنادرية, شرق الرياض, الرياض
6600000.0	5	7	600.0	الملقا, شمال الرياض, الرياض
4800000.0	7	7	566.0	حي العقيق, شمال الرياض, الرياض
3500000.0	6	7	1000.0	الملقا, الرياض
1270000.0	4	5	300.0	حي الرمال, شرق الرياض, الرياض

Before



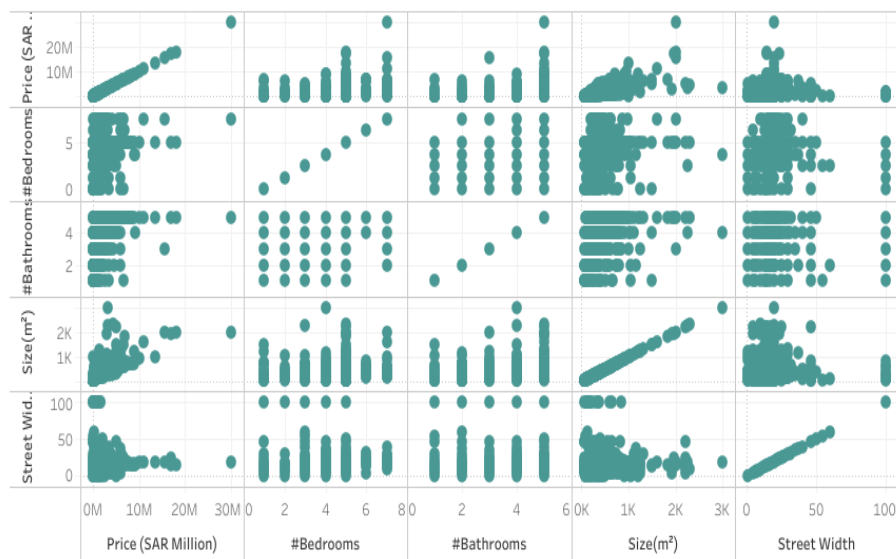
Price	#Bedrooms	#Bathrooms	Size(m²)	District
250000.0	4	2	100.0	الخالدية
720000.0	5	5	420.0	بدر
290000.0	5	4	88.0	جرير
240000.0	5	3	77.0	الديرة
180000.0	1	2	81.0	الرفيعة
...
970000.0	3	3	360.0	الجنادرية
6600000.0	5	7	600.0	الملقا
4800000.0	7	7	566.0	العقيق
13500000.0	6	7	1000.0	الملقا
1270000.0	4	5	300.0	الرمال

After

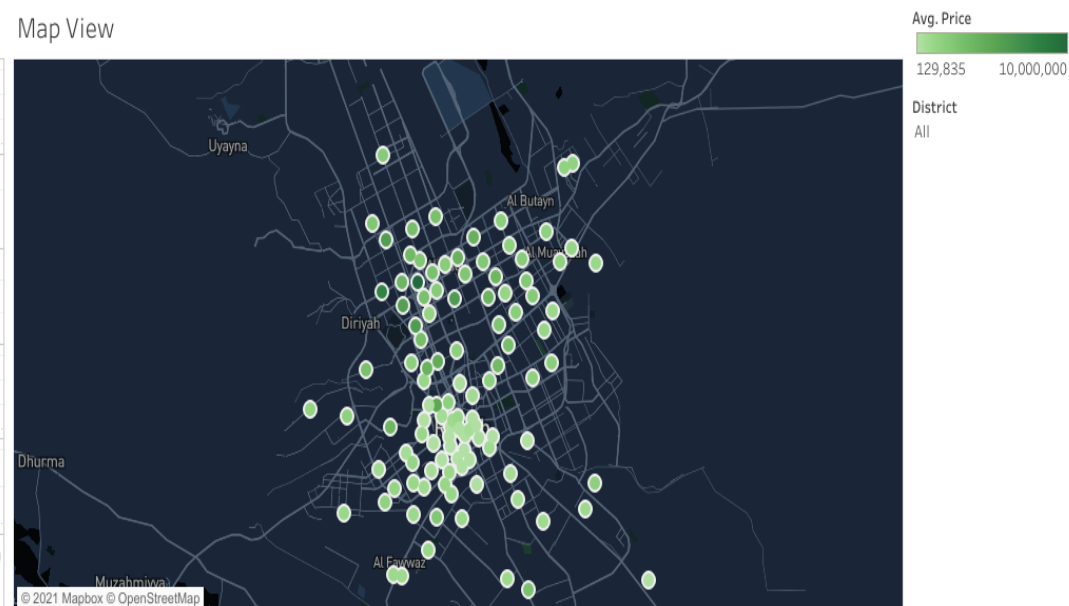
EDA Data presented using Tableau tool



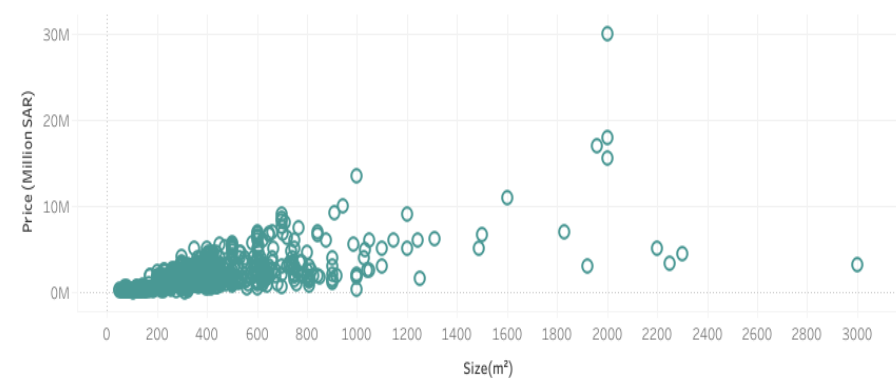
Pairplot



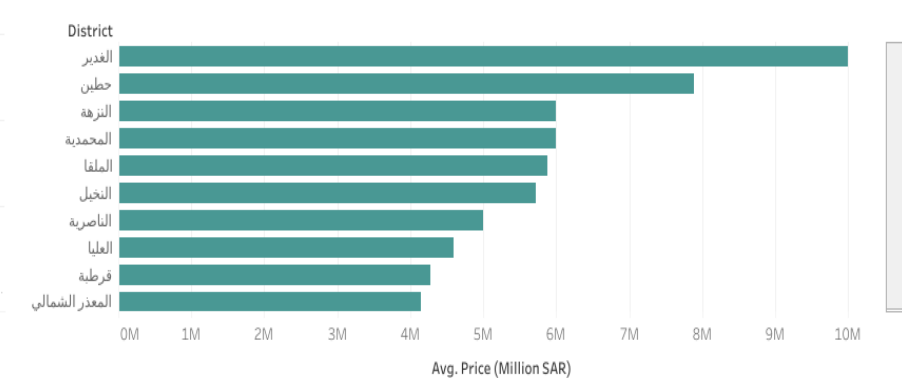
Map View

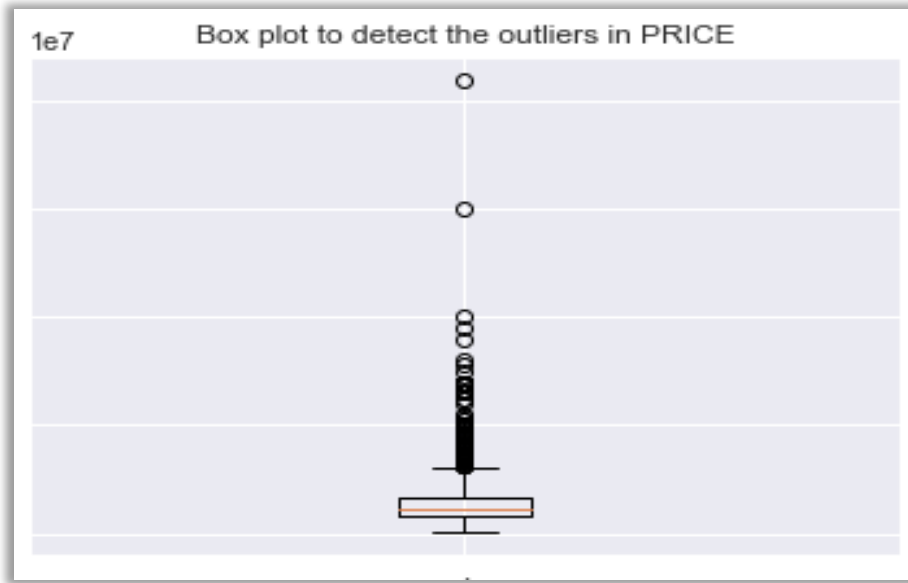


Price vs Size



Highest 10 on average prices





Insight: Price has Heavy outliers

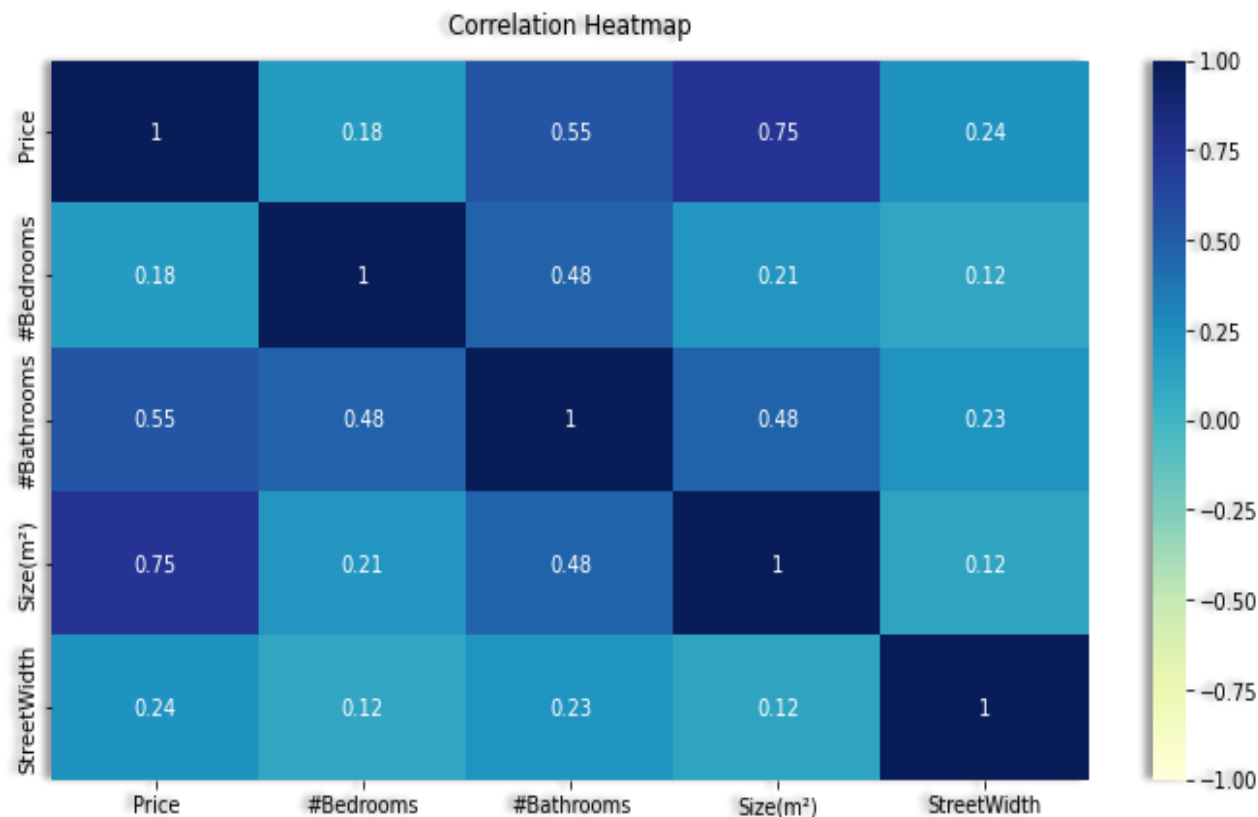
	Price	#Bedrooms	#Bathrooms	Size(m ²)	StreetWidth
count	5.504000e+03	5504.000000	5504.000000	5504.000000	5504.000000
mean	1.193790e+06	4.414608	3.863009	363.307776	15.480923
std	1.276224e+06	1.108494	1.251908	318.800361	9.261991
min	9.000000e+02	1.000000	1.000000	50.000000	1.000000
25%	3.000000e+05	4.000000	3.000000	183.000000	10.000000
50%	1.050000e+06	5.000000	4.000000	307.000000	15.000000
75%	1.400000e+06	5.000000	5.000000	400.000000	20.000000
max	3.000000e+07	7.000000	5.000000	3000.000000	100.000000

Insight: Minimum house price: 900 SAR !

Treating outliers: Drop only the house that costs 900 SAR, because removing all outliers will remove data so It will not help us fitting as many values as possible.



- The "Size" feature has high positive correlation **0.75** with target "Price"
- The "Bathrooms" feature has high positive correlation **0.55** with target "Price"
- The "Bedrooms" feature has low positive correlation **0.18** with target "Price"
- The "StreetWidth" feature has low positive correlation **0.24** with target "Price"



Modeling and Evaluation



Modeling and Evaluation



1 Linear Regression (Baseline model)

- **Preprocessing:** Convert “District” feature to dummy variables
- **Split data:** 30% Test - 70% Train
- **Results:**

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

Modeling and Evaluation



2 Random Forest Regressor

- **Split data:** 30% Test - 70% Train
- **Results:**

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

Modeling and Evaluation



3 Ridge Regression

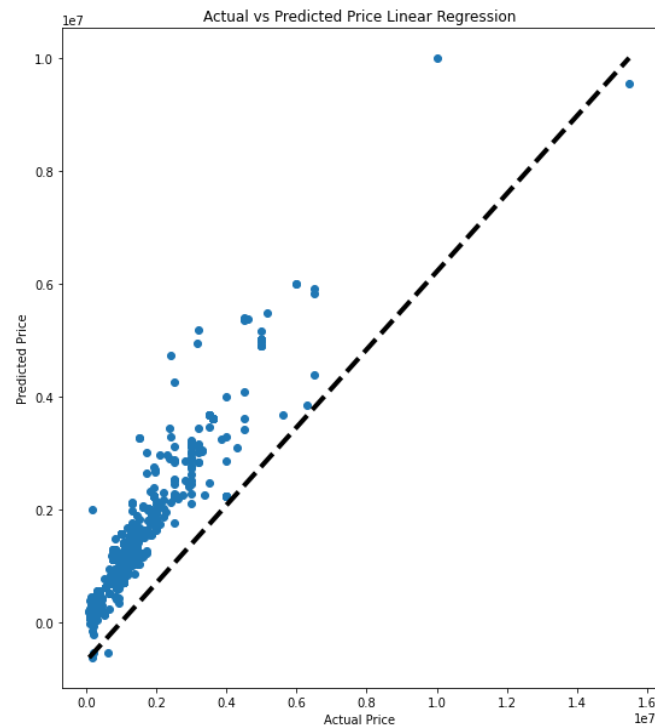
- **Split data:** 30% Test - 70% Train
- **Results:**

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

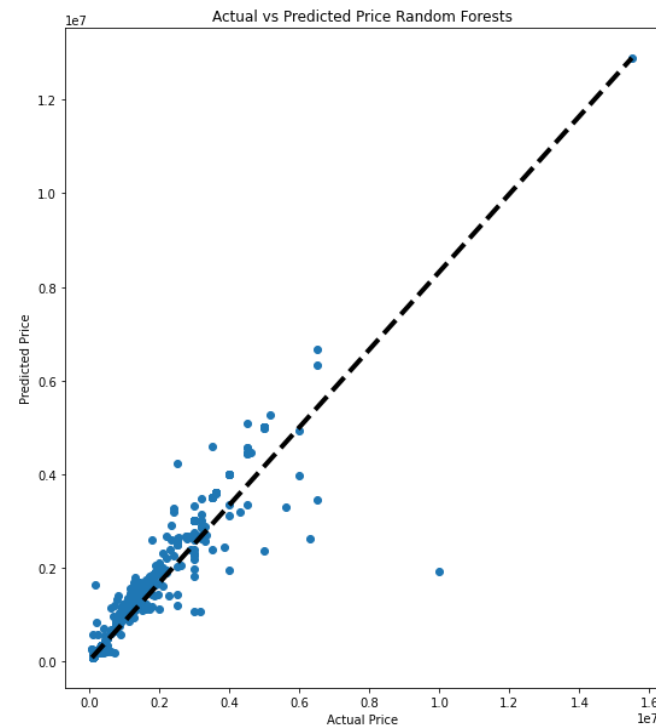
Modeling and Evaluation



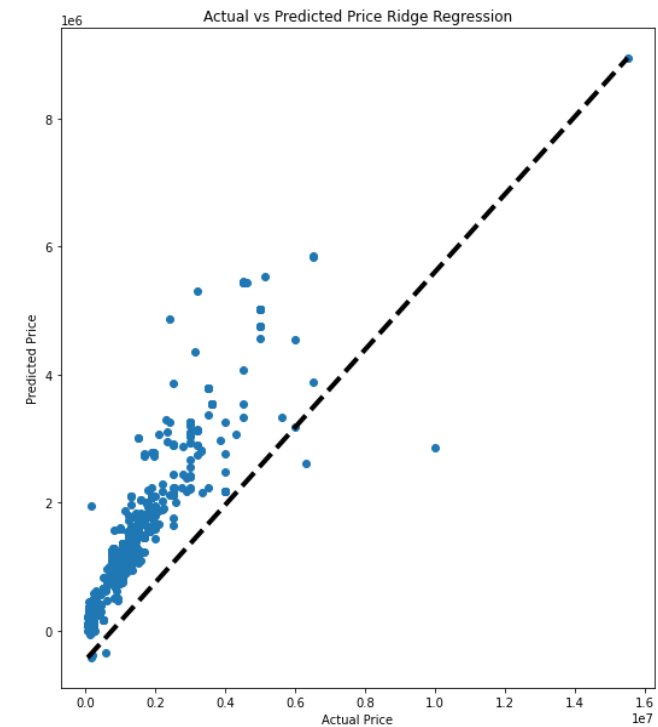
● Fit line for each model



Linear Regression



Random Forest

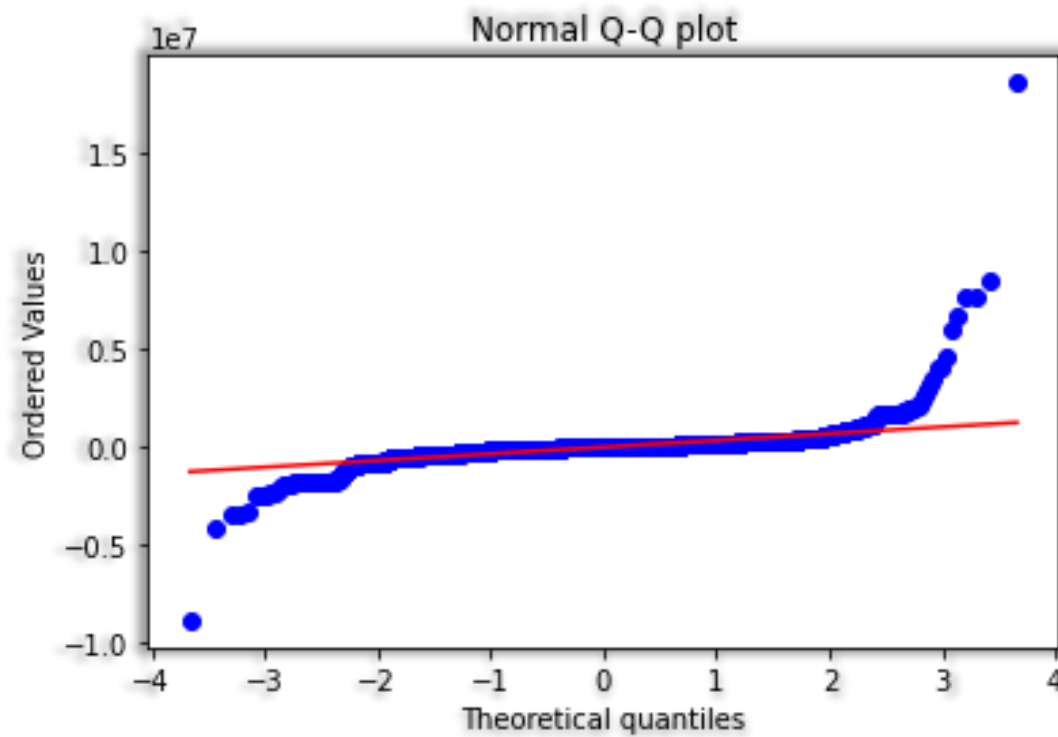


Ridge Regression

Modeling and Evaluation



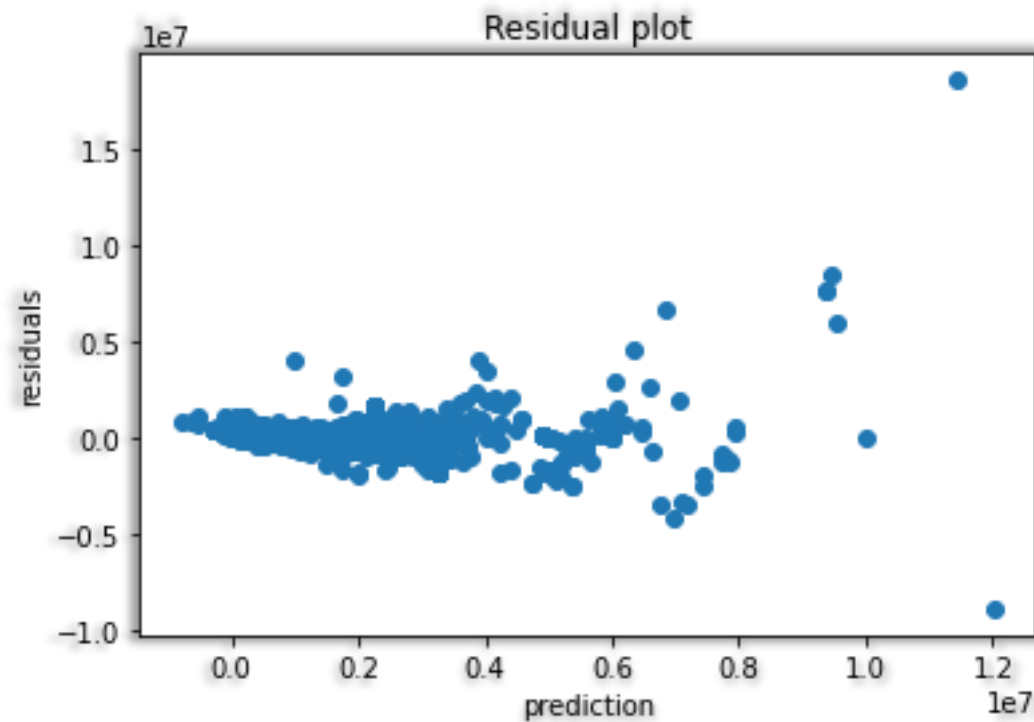
- **Assumption 1:** Regression is linear in parameters and correctly specified



Modeling and Evaluation



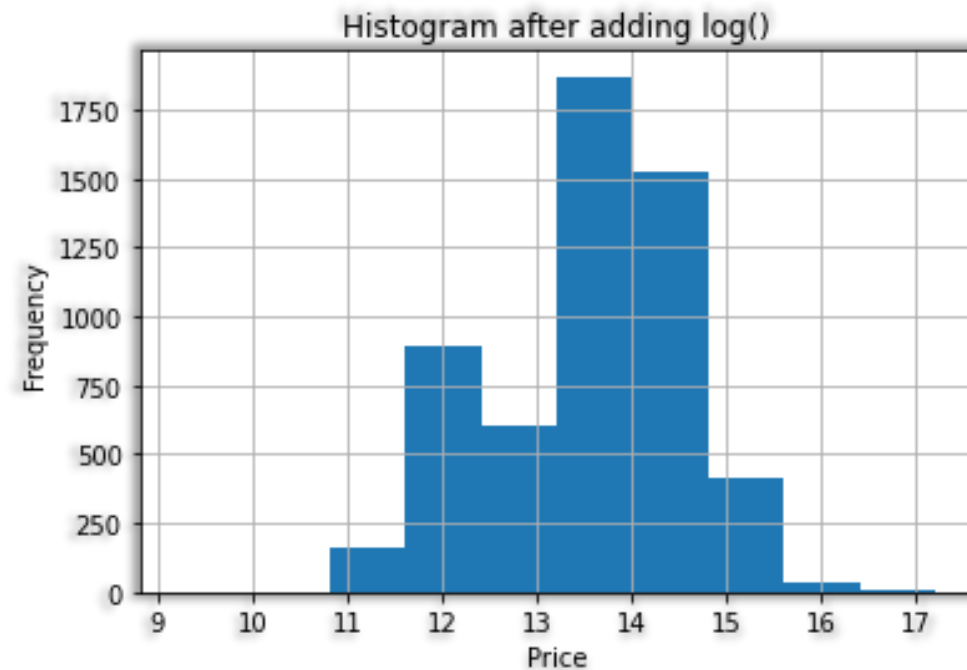
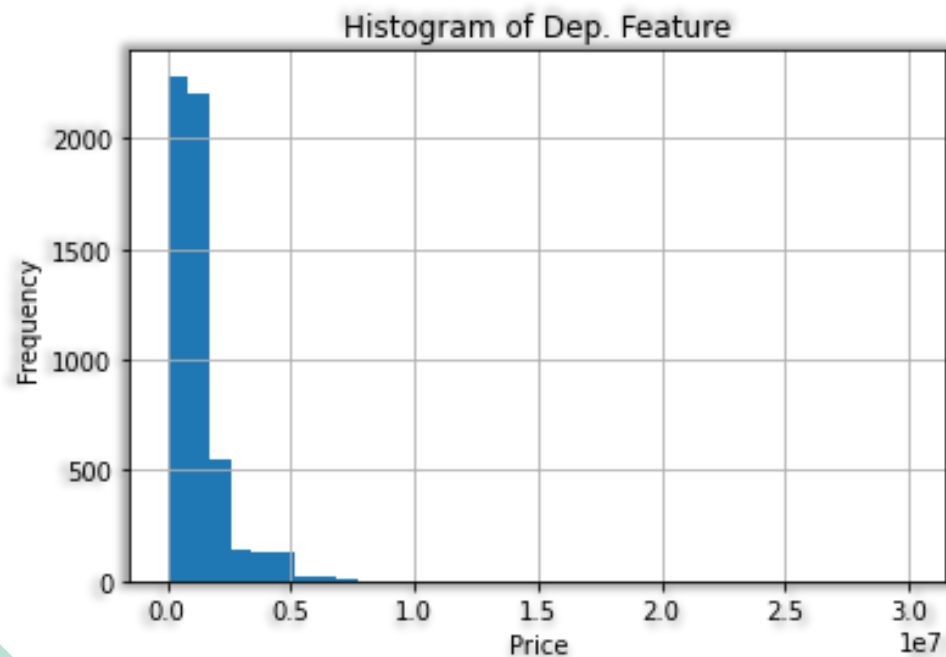
- **Assumption 2:** Residuals should be normally distributed with 0 mean






Modeling and Evaluation



● **Assumption 3:** Error Terms must have constant variance



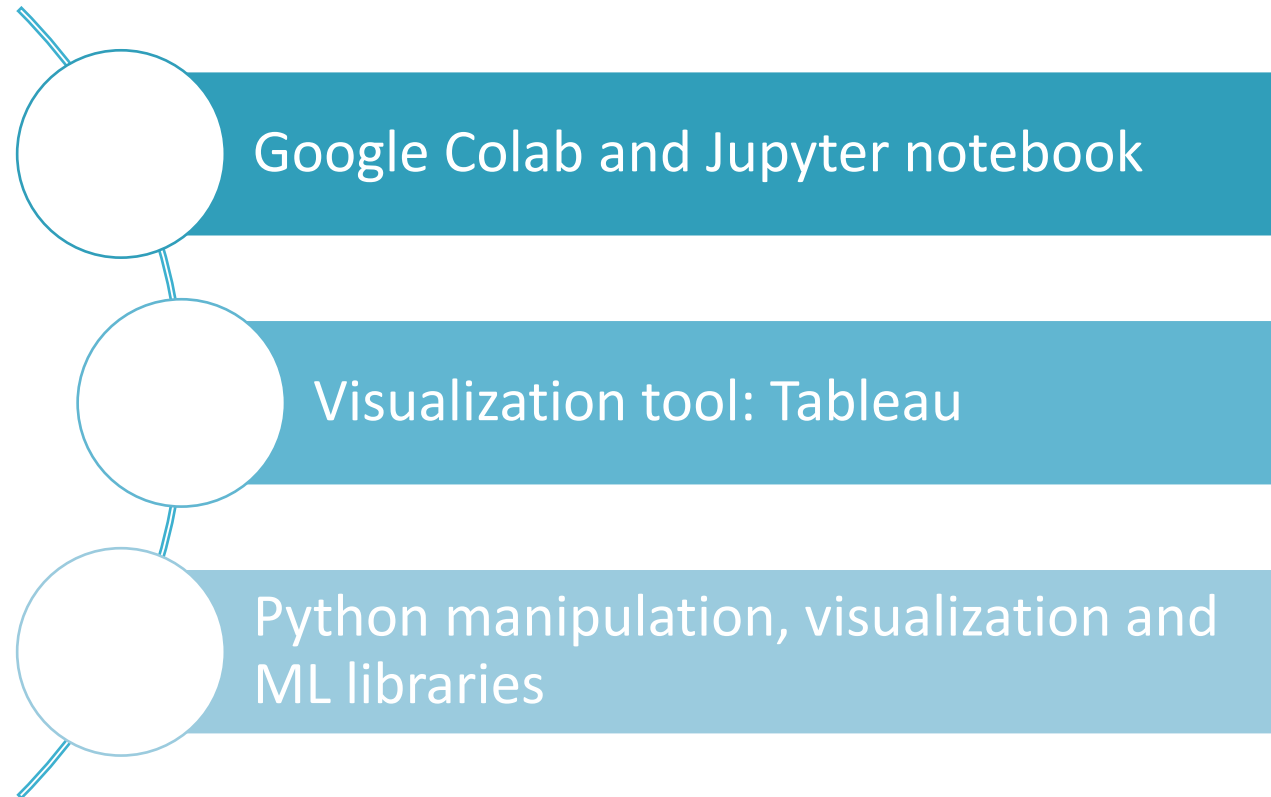
Conclusion

-  House price is affected by: # of bedrooms, # of bathrooms, Size, Street width and District
-  All models produce very good results even without using feature selection methodologies
-  Random forest wins the best performance model among all models



Conclusion

Tools used



Thank you!

Any Questions?

