HOUSE SALES IN KING COUNTY REGRESSION

DATA-DRIVEN INSIGHTS AND PREDICTIVE MODELING

CONTENT

ABOUT THE PROJECT

This project consists of prices for house sales which occurred between May 2014 and 2015. The first step was to clean this data and ensure fit for purpose. We then looked at which features would provide interesting insights and answered questions which would help generating insights. Finally we looked into building a model capable of predicting house prices and found out the most suitable model.

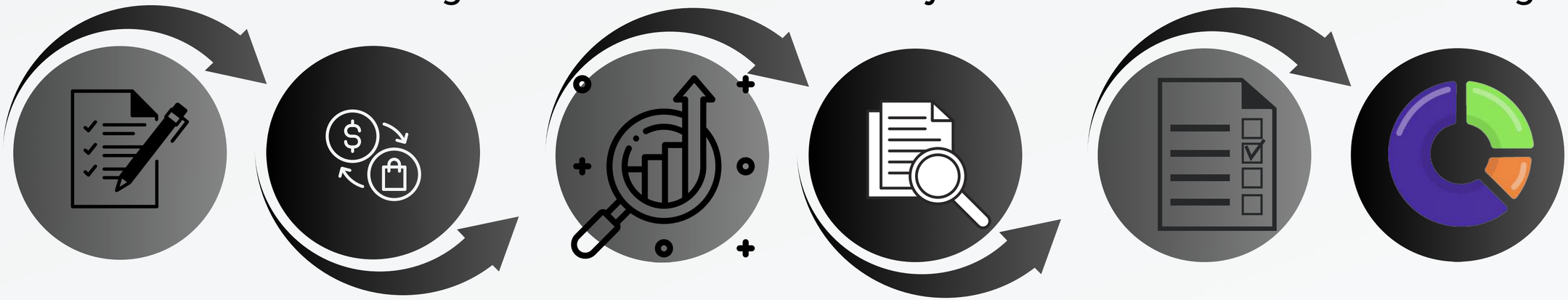
PROJECT TIMELINE

Data Extraction
and
manipulation

Data analytics
and
Insight generation

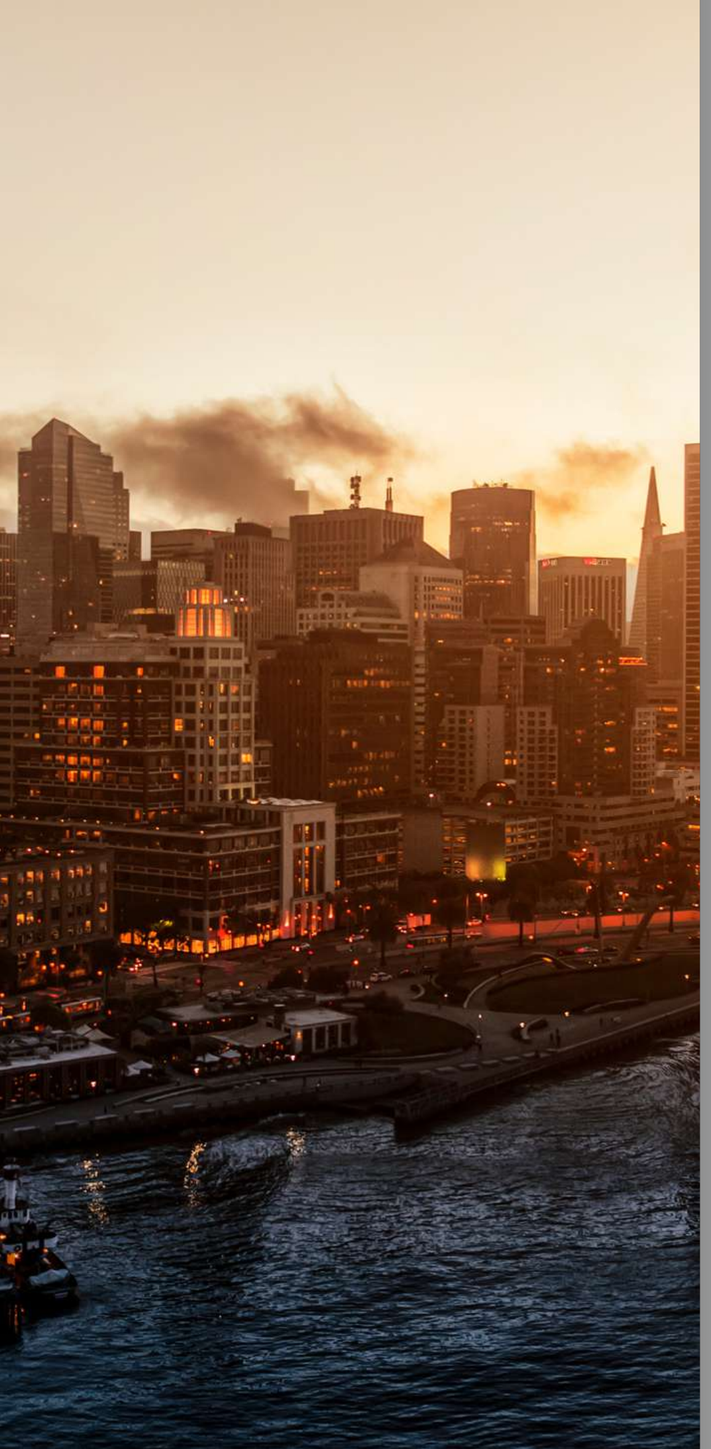
Model Building
and Presentation
preparation

PROJECT CRONOLOGY



VARIABLES

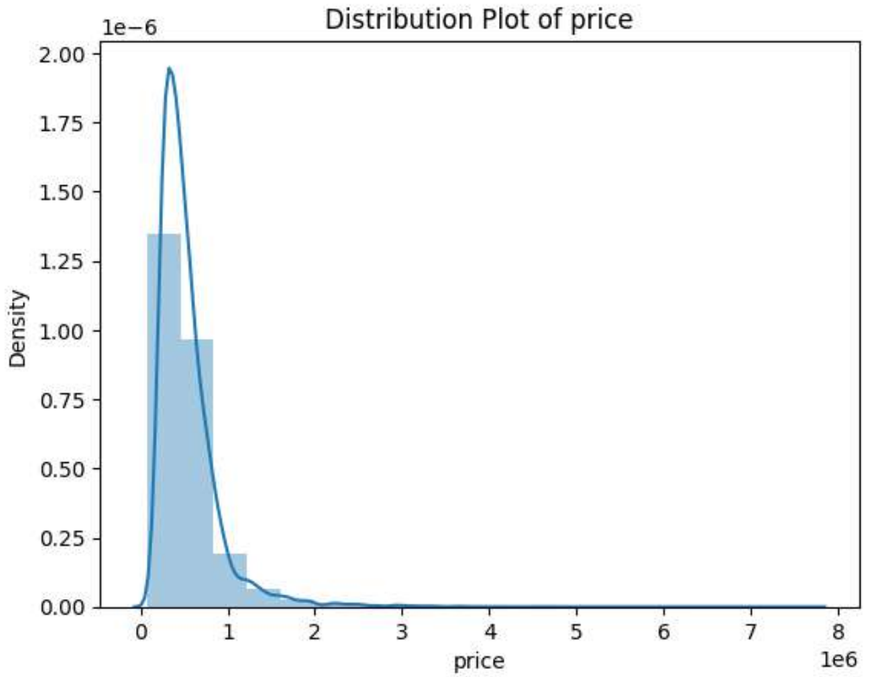
- **Yid** - Unique ID for each home sold
- **Date** - Date of the home sale
- **Price** - Price of each home sold
- **Bedrooms** - Number of bedroom
- **Bathrooms** - Number of bathrooms
- **Sqft_living** - Square footage of the apartments interior living space
- **Sqft_lot** - Square footage of the land space
- **Floors** - Number of floors
- **Waterfront** - A dummy variable for whether the apartment was overlooking the waterfront or not
- **View** - An index from 0 to 4 of how good the view of the property was



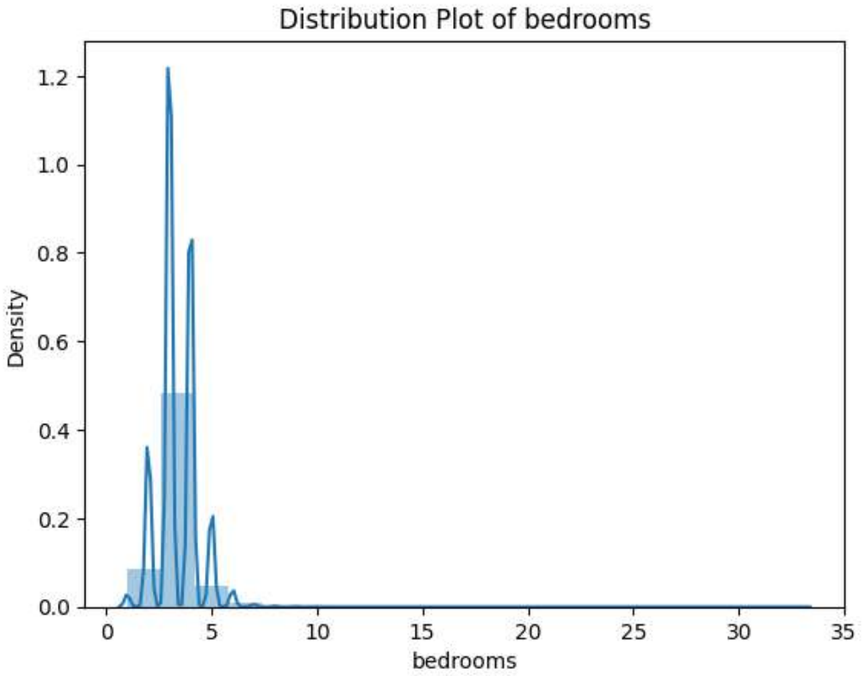
- **Condition** - An index from 1 to 5 on the condition of the apartment,
- **Grade** - An index from 1 to 13, where 1-3 falls short of building construction and design, 7 has an average level of construction and design, and 11-13 have a high quality level of construction and design.
- **Sqft_above** - The square footage of the interior housing space that is above ground level
- **Sqft_basement** - The square footage of the interior housing space that is below ground level
- **Yr_built** - The year the house was initially built
- **Yr_renovated** - The year of the house's last renovation
- **Zipcode** - What zipcode area the house is in
- **Lat** - Lattitude
- **Long** - Longitude
- **Sqft_living15** - The square footage of interior housing living space for the nearest 15 neighbors
- **Sqft_lot15** - The square footage of the land lots of the nearest 15 neighbors

UNIVARIATE ANALYSIS

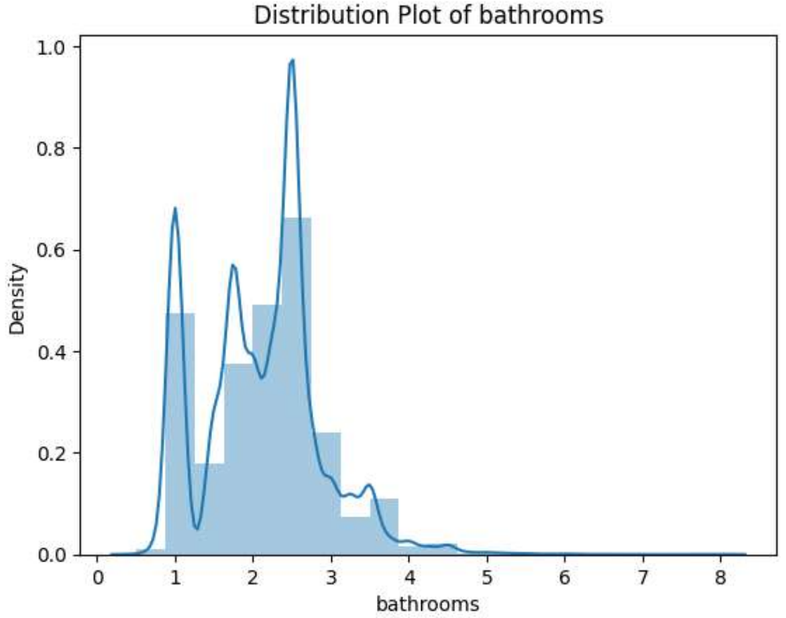
	price	bedrooms	bathrooms	sqft_living	sqft_lot	floors	waterfront	view	condition	grade	sqft_above	sqft_basement	yr_built	yr_renovated	zipcode	lat	long	sqft_living15	sqft_lot15	age_of_property
count	21597	21597	21597	21597	21597	21597	21597	21597	21597	21597	21597	21597	21597	21597	21597	21597	21597	21597	21597	21597
mean	540202.9	3.37	2.12	2080.32	15099.4	1.53	0.01	0.23	3.41	7.66	1788.6	291.73	1971	84.46	98078	47.56	-122.2	1986.62	12758.28	52
min	78000	1	0.5	370	520	1	0	0	1	3	370	0	1900	0	98001	47.16	-122.5	399	651	8
25%	322000	3	1.75	1430	5040	1	0	0	3	7	1190	0	1951	0	98033	47.47	-122.3	1490	5100	26
50%	450000	3	2.25	1910	7618	2	0	0	3	7	1560	0	1975	0	98065	47.57	-122.2	1840	7620	48
75%	645000	4	2.5	2550	10685	2	0	0	4	8	2210	560	1997	0	98118	47.68	-122.1	2360	10083	72
max	7700000	33	8	13540	1651359	4	1	4	5	13	9410	4820	2015	2015	98199	47.78	-121.3	6210	871200	123
std	367133.7	0.93	0.77	918.11	41412.6	0.55	0.09	0.77	0.65	1.17	827.76	442.67	29.38	401.82	53.51	0.14	0.14	685.23	27274.44	29.38



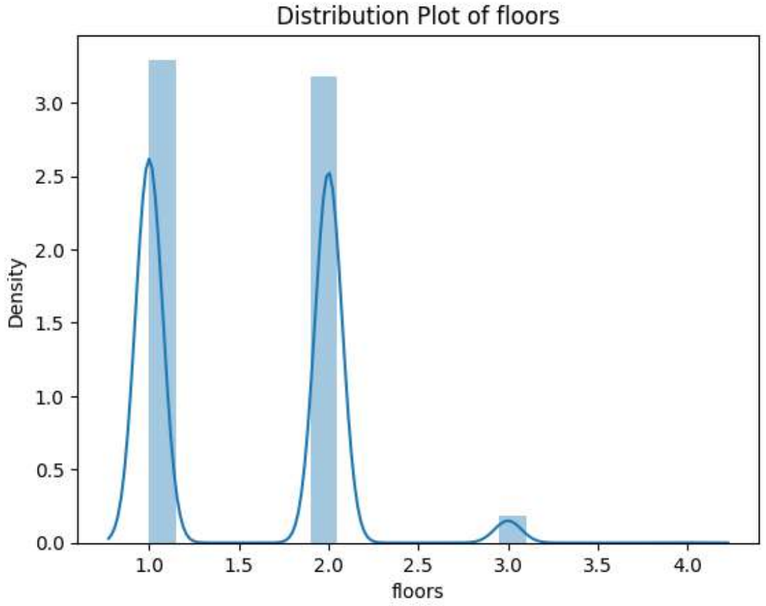
Variation of price



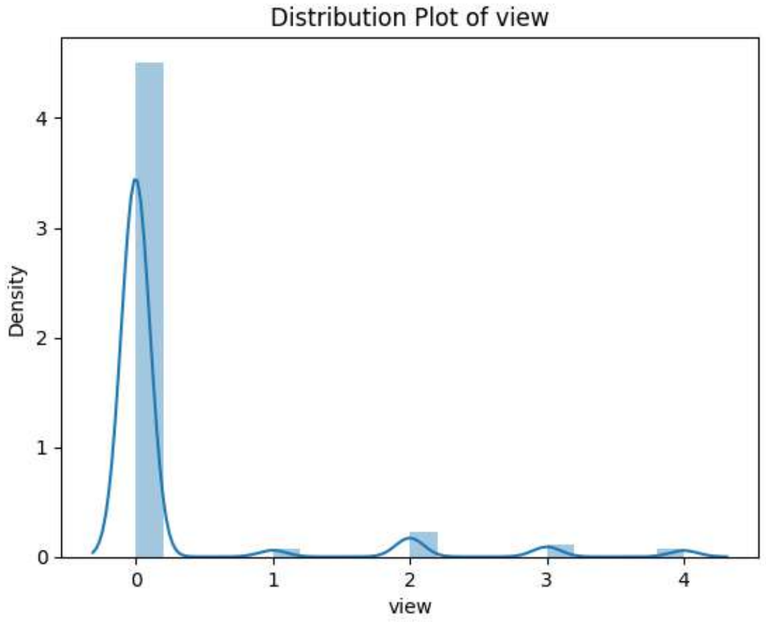
Variation of number of bedrooms



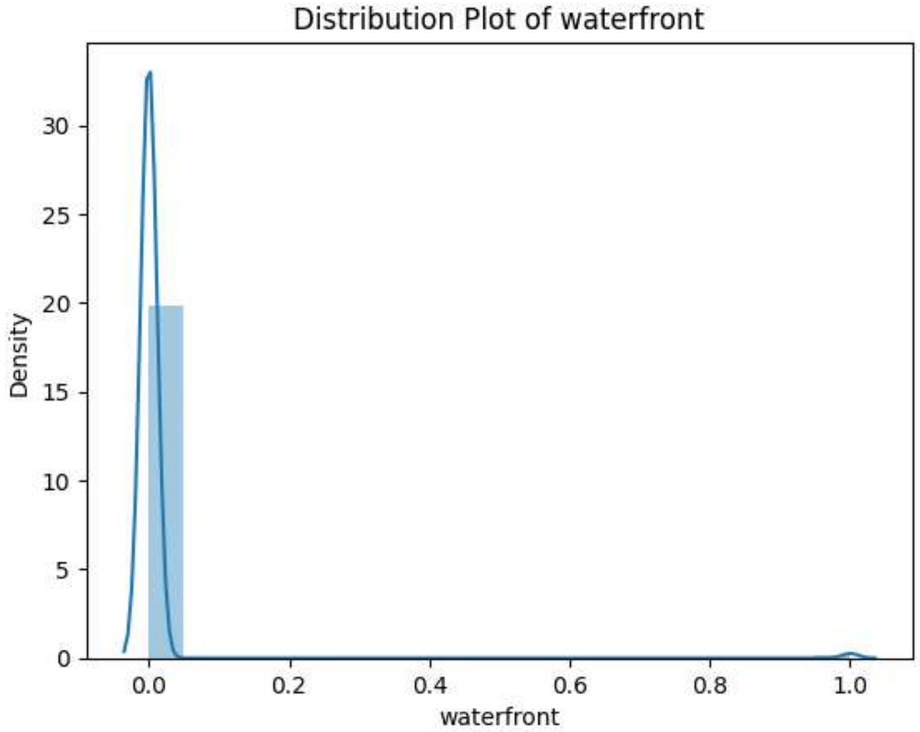
Variation of number of bathrooms



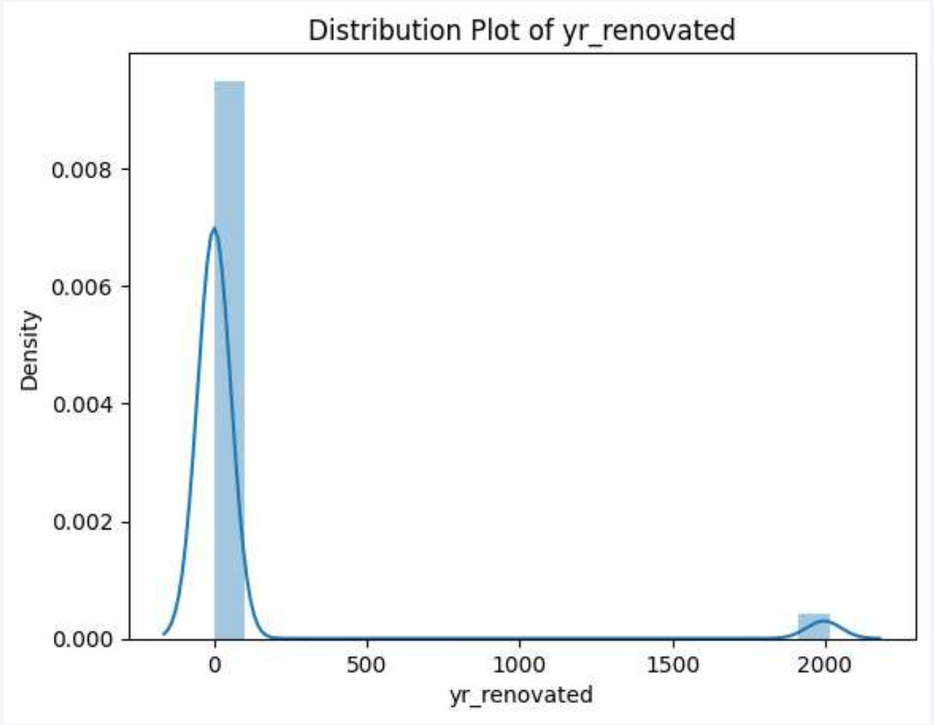
Number of floors



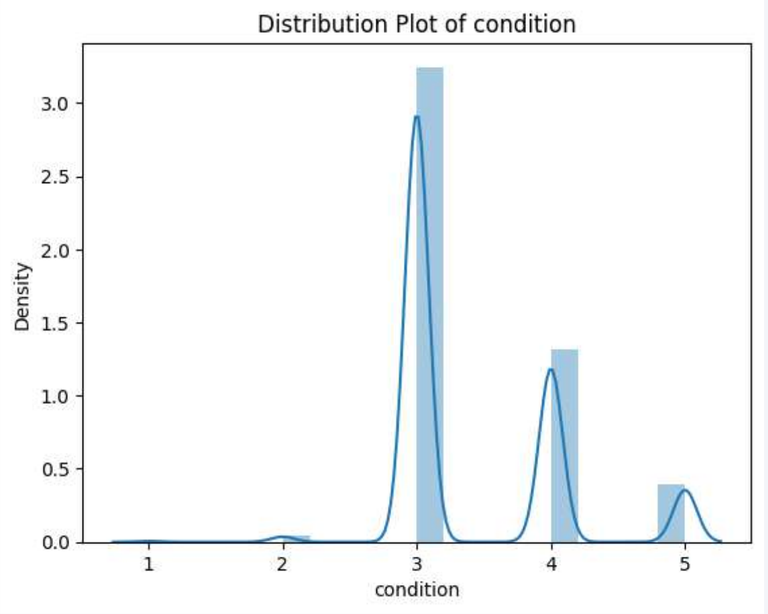
An index from 0 to 4 of how good the view of the property was



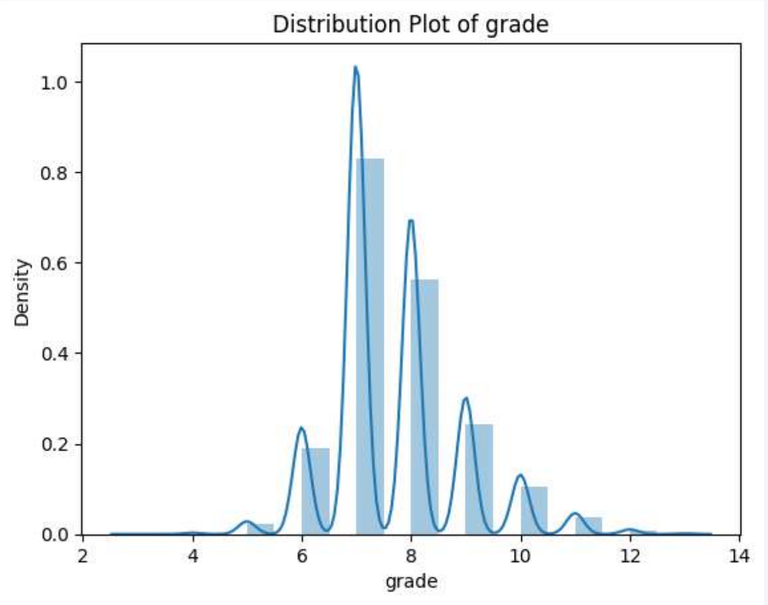
Apartment overlooking the waterfront or not



The year of the house’s last renovation

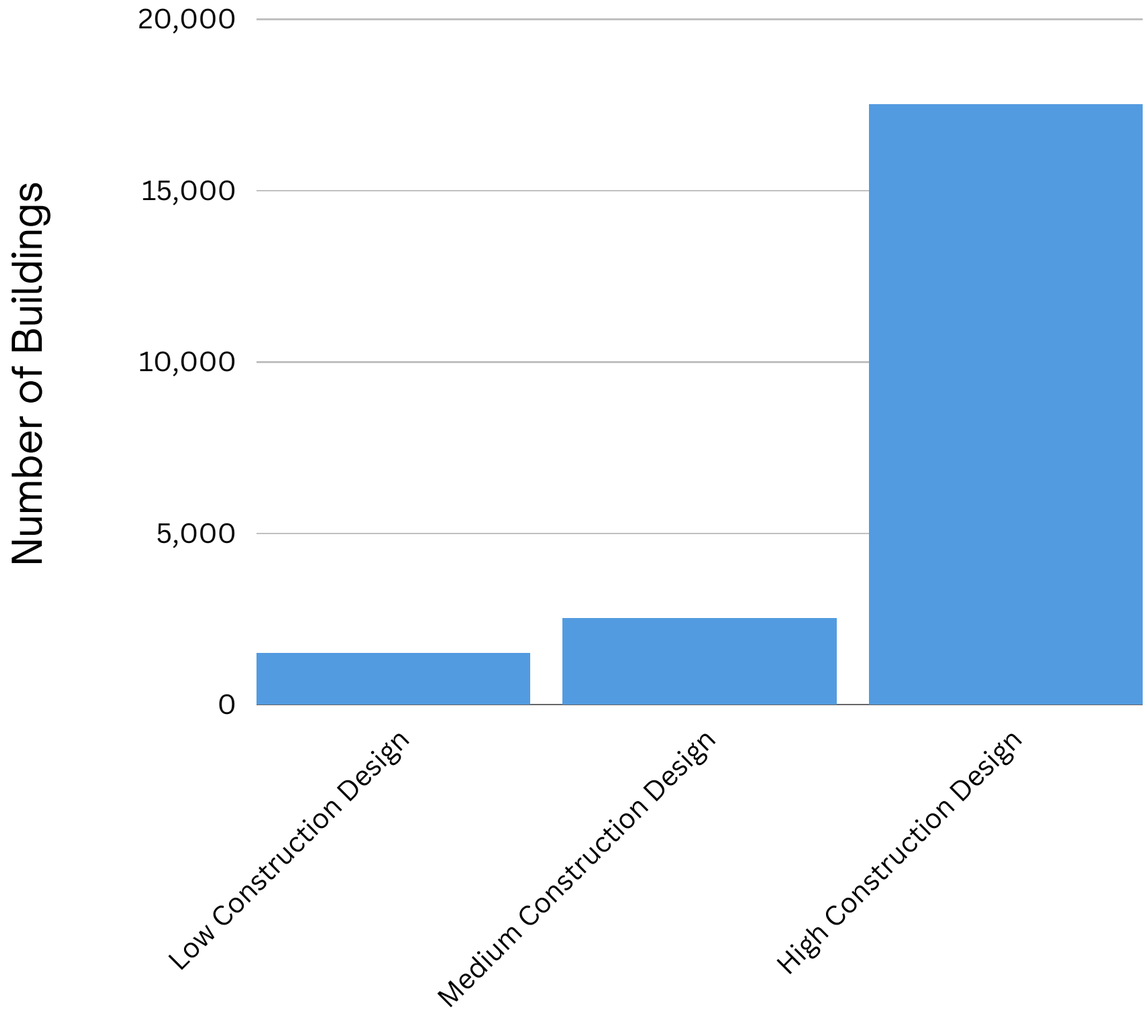


An index from 1 to 5 on the condition of the apartment,

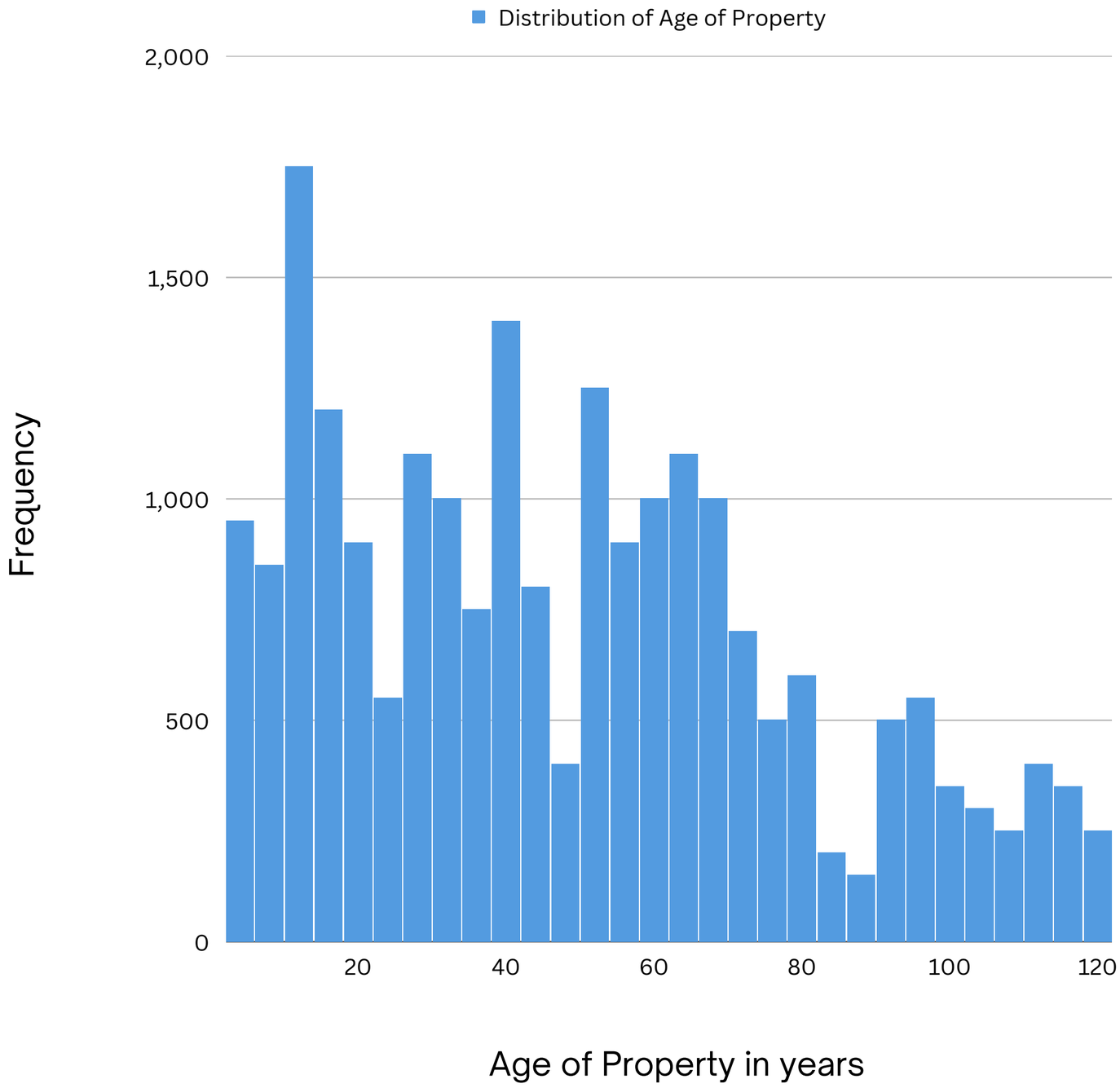


1-3 falls short of building construction and design
7 has an average level of construction and design
11-13 have a high quality level of construction and design.

Histogram

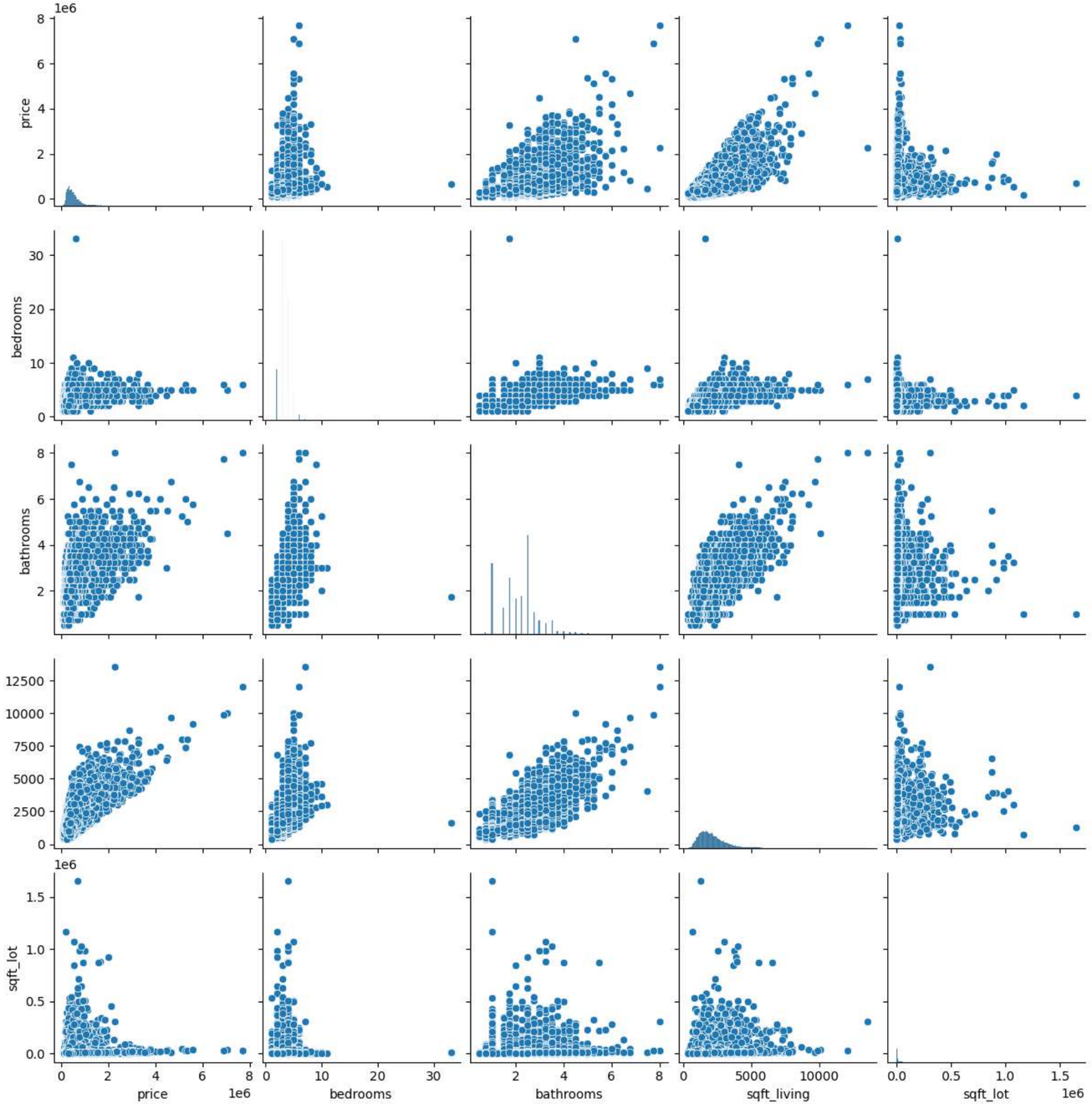


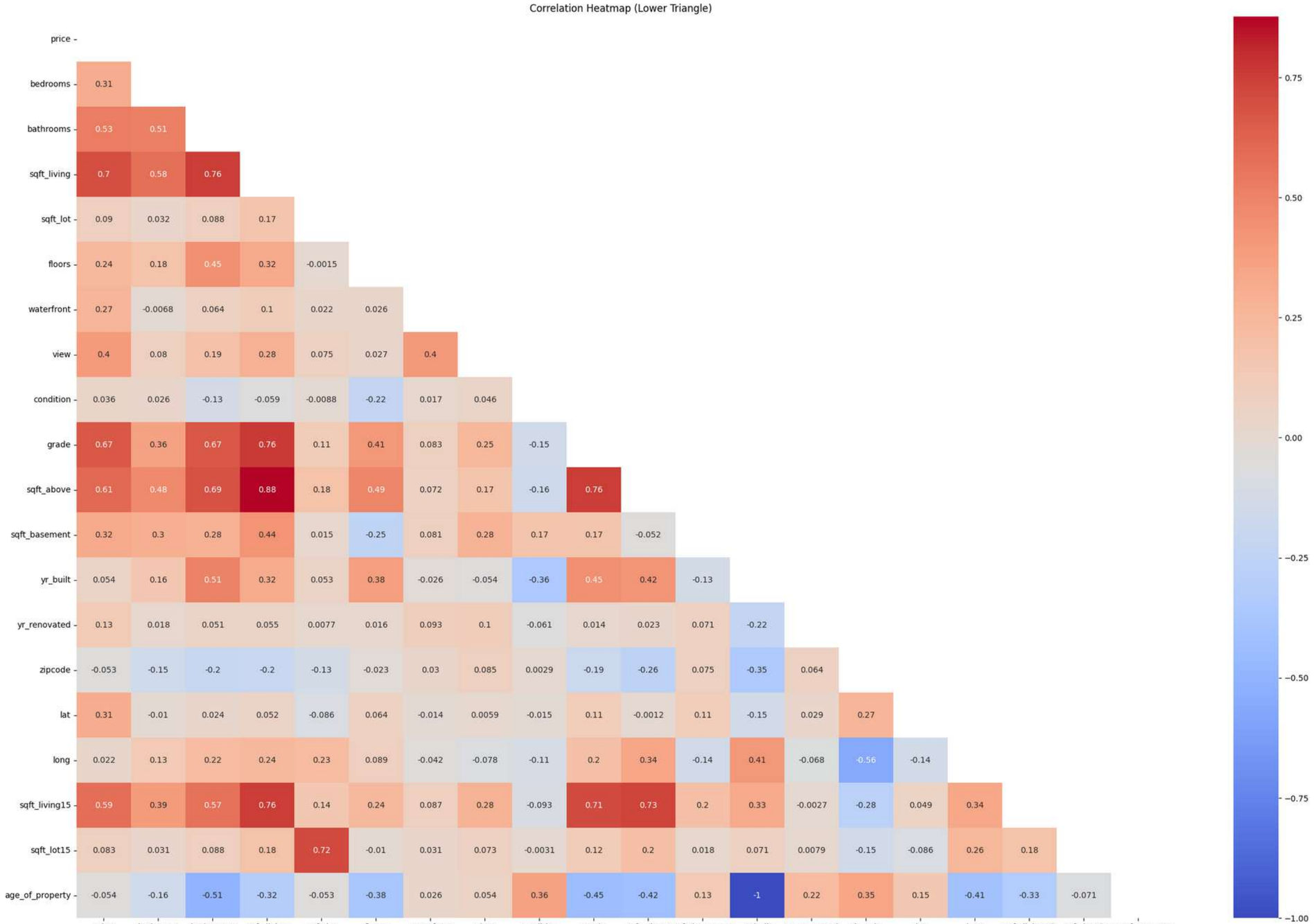
Variation of low, medium and high level of construction and design


Distribution of Age of Property
2,000
1,500
1,000
500
0
Frequency
20
40
60
80
100
120
Age of Property in years

Pair plots

- The scatter plot gives us an idea of correlation between the number of bedrooms and the price of the property.
- This plot provides insights into how the size of the living area impacts the price.
- This plot helps us understand on how the size of the lot has any influence on the price.




Correlation Heatmap (Lower Triangle)
price
bedrooms 0.31
bathrooms 0.53 0.51
sqft_living 0.7 0.58 0.76
sqft_lot 0.09 0.032 0.088 0.17
floors 0.24 0.18 0.45 0.32 -0.0015
waterfront 0.27 -0.0068 0.064 0.1 0.022 0.026
view 0.4 0.08 0.19 0.28 0.075 0.027 0.4
condition 0.036 0.026 -0.13 -0.059 -0.0088 -0.22 0.017 0.046
grade 0.67 0.36 0.67 0.76 0.11 0.41 0.083 0.25 -0.15
sqft_above 0.61 0.48 0.69 0.88 0.18 0.49 0.072 0.17 -0.16 0.76
sqft_basement 0.32 0.3 0.28 0.44 0.015 -0.25 0.081 0.28 0.17 0.17 -0.052
yr_built 0.054 0.16 0.51 0.32 0.053 0.38 -0.026 -0.054 -0.36 0.45 0.42 -0.13
yr_renovated 0.13 0.018 0.051 0.055 0.0077 0.016 0.093 0.1 -0.061 0.014 0.023 0.071 -0.22
zipcode -0.053 -0.15 -0.2 -0.2 -0.13 -0.023 0.03 0.085 0.0029 -0.19 -0.26 0.075 -0.35 0.064
lat 0.31 -0.01 0.024 0.052 -0.086 0.064 -0.014 0.0059 -0.015 0.11 -0.0012 0.11 -0.15 0.029 0.27
long 0.022 0.13 0.22 0.24 0.23 0.089 -0.042 -0.078 -0.11 0.2 0.34 -0.14 0.41 -0.068 -0.56 -0.14
sqft_living15 0.59 0.39 0.57 0.76 0.14 0.24 0.087 0.28 -0.093 0.71 0.73 0.2 0.33 -0.0027 -0.28 0.049 0.34
sqft_lot15 0.083 0.031 0.088 0.18 0.72 -0.01 0.031 0.073 -0.0031 0.12 0.2 0.018 0.071 0.0079 -0.15 -0.086 0.26 0.18
age_of_property -0.054 -0.16 -0.51 -0.32 -0.053 -0.38 0.026 0.054 0.36 -0.45 -0.42 0.13 -1 0.22 0.35 0.15 -0.41 -0.33 -0.071
price bedrooms bathrooms sqft_living sqft_lot floors waterfront view condition grade sqft_above sqft_basement yr_built yr_renovated zipcode lat long sqft_living15 sqft_lot15 age_of_property
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00

Heatmap

Model Performances

LINEAR REGRESSION

R2-Score: 0.6807
RMSE: 203796.0159

R2-Score: 0.6767
RMSE: 205070.5268

DECISION TREE

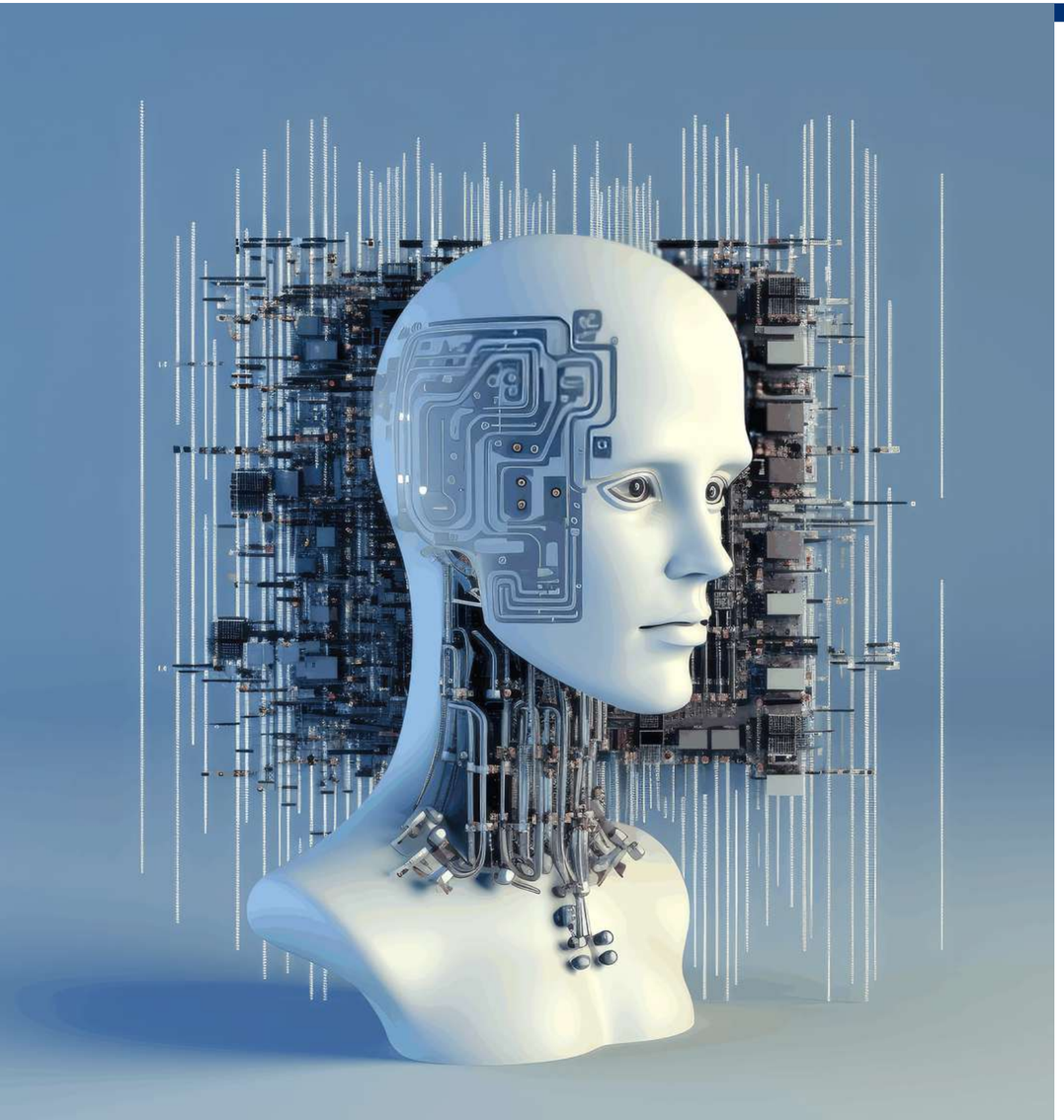
RANDOM FOREST

R2-Score: 0.8727
RMSE: 128690.3932

R2-Score: -0.0534
RMSE: 370186.7632

SVR (SUPPORT VECTOR REGRESSION)

Best Model



The best model is Random Forest

- Random Forest with the accuracy of 87% comes out to be the best model as compared to Linear Regression, Decision tree and SVR models.
- with the accuracy of 68% , Linear Regression performance is not good.

Model performances after removing multi-collinearity

R2-Score: 0.4143
RMSE: 286615.7797

Linear Regression

R2-Score: 0.0820
RMSE: 358847.1861

Decision Tree

R2-Score: 0.4076
RMSE: 288259.5133

Random Forest

SVR

R2-Score: -0.0557
RMSE: 384806.1989

Hyperparameters

Best Hyperparameters:
max_depth: None
min_samples_split: 5
n_estimators: 150

- **Max_depth**: This parameter determines the maximum depth of each decision tree in the random forest ensemble. Setting it to None means there is no maximum depth limit, and the trees can expand until all the leaves are pure or until they contain a minimum number of samples specified by another parameter.
- **Min_samples_split**: It sets the minimum number of samples required to split an internal node during the construction of a tree. In this case, the value is set to 5, meaning a node must have at least 5 samples to be eligible for splitting.
- **n_estimators**: This parameter defines the number of decision trees to be created in the random forest ensemble. Increasing the number of estimators generally improves the model's performance, but it also increases computation time. In this case, 150 trees are used to form the random forest ensemble.

FINAL CONCLUSION

Linear Regression:
R2-Score: 0.6807
RMSE: 203796.0159

Decision Tree:
R2-Score: 0.7043
RMSE: 196142.2947

Random Forest:
R2-Score: 0.8710
RMSE: 129540.4012

Polynomial Regression (Degree 2):
MSE: 29888549258.824467
R2 Score: 0.7869156618096037

Polynomial Regression (Degree 3):
MSE: 8.76672289925631e+26
R2 Score: -6250056939565539.0

Here is the comparision of the all model from this we can conclude that Random Forest regression model giving us a best accurancy


Thank
you!