REAL ESTATE INDUSTRY

Kings County House Presentation

Mary Murugami

Executive Summary

The Kings County House data will be used to analyze house sales in a Northwestern county..

The dataset contains information on features of the houses such as the number of bedrooms, bathrooms, and the square footage of the houses, among others which will be used to advise homeowners on how to increase the estimated value of their homes

The data will be analyzed using a regression model.

Business Problem

The stakeholder needs to advice home owners on how the variables below can increase the estimated values of their homes

01 **Zip Code**

Does the zip code influence price
Do more affluent houses with better
zip codes sell faster?

02 **Year Renovated**

Do renovations influence
house market rate

03 **Square Footage**

How does the square footage
Influence the price, zip code and
no of bedrooms in th
estate market for kin ty
Houses?

Research Questions

1. How do various factors affect the price of houses in Kings County?
2. Does the location of a house in Kings County have an impact on its price?
3. Is there a significant linear relationship between the square footage of a house and its price in Kings County?

Hypothesis

Alternate Hypothesis

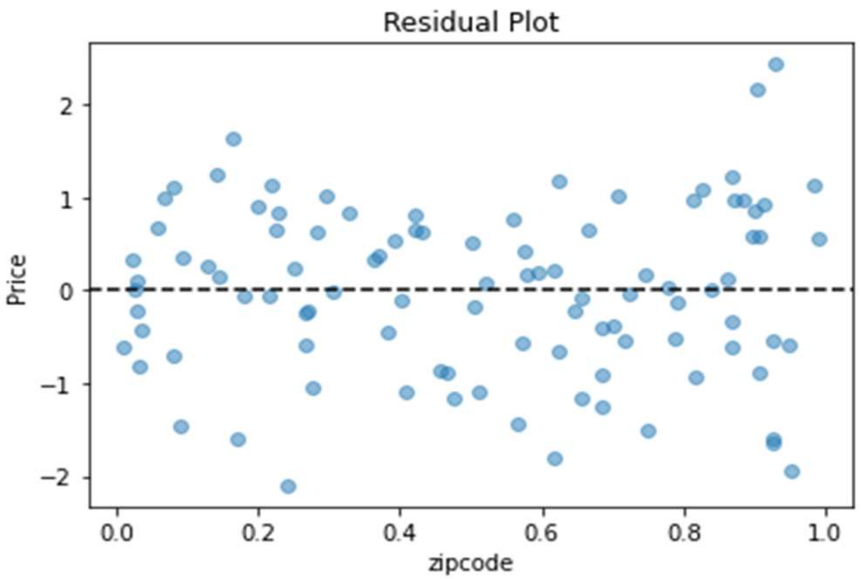
- **There is a positive linear relationship between the square footage of a house and its price in Kings County, such that as the square footage increases, the price of the house also increases. This relationship is statistically significant at the 5% level of significance.**

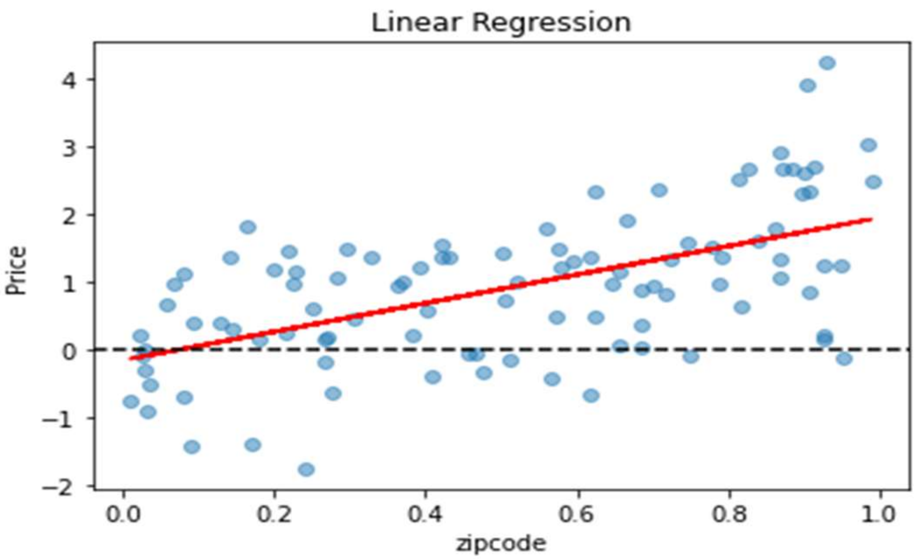
Null Hypothesis

- **There is no linear relationship between sqft_living and price in the Kings County housing data.**

A t-statistic of 16.12 with a p-value of 4.04e-58 indicates strong evidence against the null hypothesis that there is no linear relationship between sqft_living and price in the Kings County housing data.
Therefore, we can reject the null hypothesis and conclude that there is a significant linear relationship between these two variables.

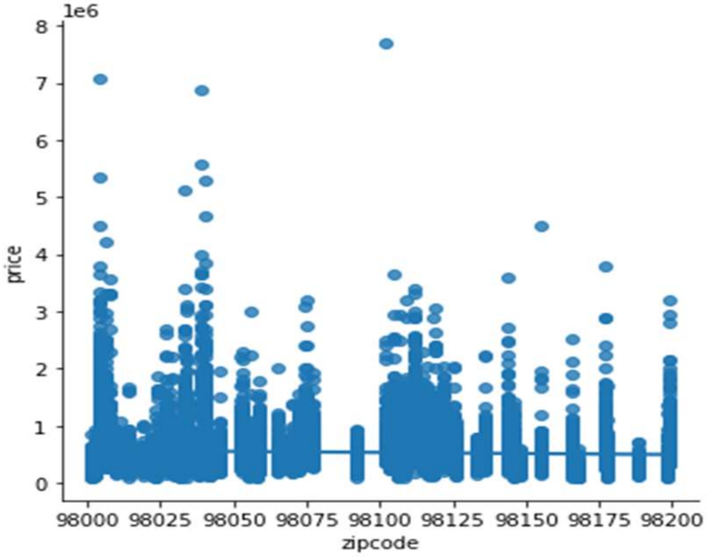
Linear Regression

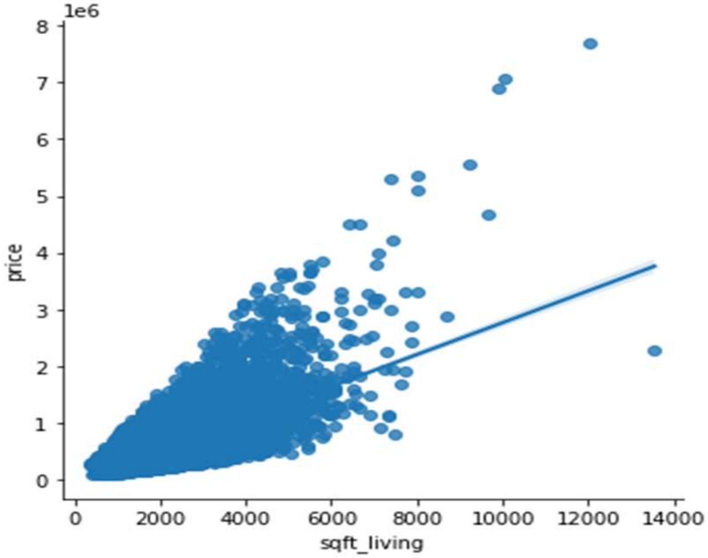




- The relationship between Zip code (Predictor) and Price (Response) variables is represented by a regression line which showcases a linear relationship between the two.
- The slope of our line is indicative of strength and direction of the relationship, the price is influenced by the zip code (the better the zip code, the higher the price)
- The data points are close to the regression line, making the model a better fit.
- There is a random pattern around the zero- line, variance is constant throughout hence it meets the assumptions i.e. the linearity, residual normality and homoscedasticity

Linear Regression



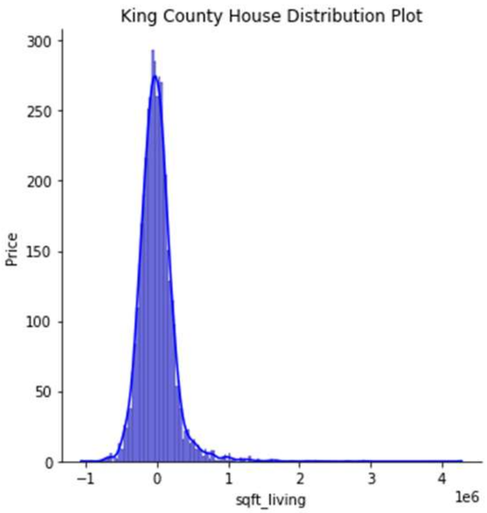


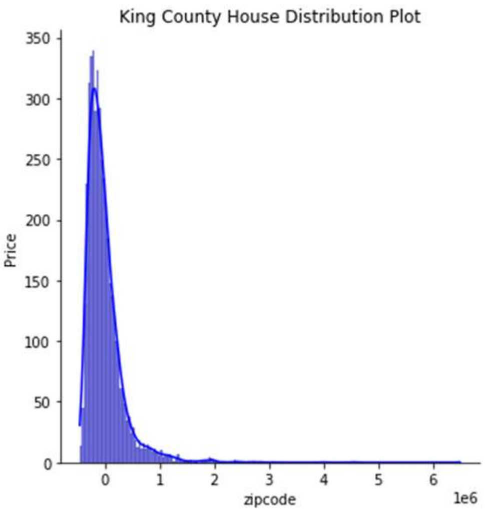
Categorical Variable= zip code and Target/Response Variable = Price

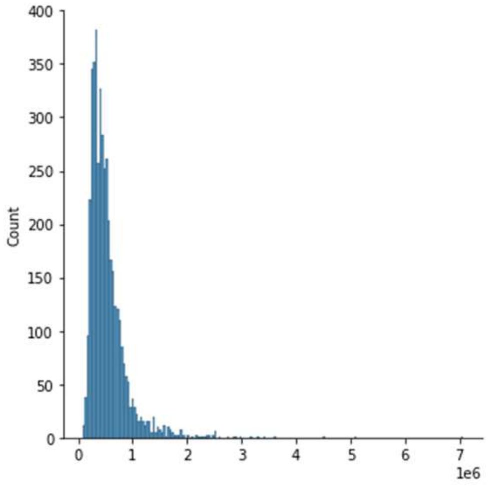
- The linear regression is represented by the blue line which demonstrates the relationship between the zip code and the price
- The shading around the line shows the 95% confidence interval of the regression line meaning the true regression lies around the line
- The shaded area around the line represents the 95% confidence interval of the regression line. This means that there is a 95% chance that the true regression line lies within this shaded area.
- A change in price is shown by an increase in the zip code hence, the positive slope
- The funnel shape indicates that homoscedasticity assumptions have been violated so the model may not be a good fit for the data.

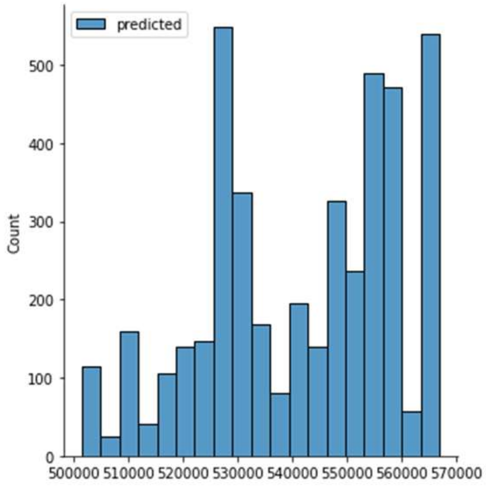
Distribution Plots

- The fitted regression model demonstrates the square foot living versus the price / zip code vs Price
- 80% was training data while 20% was testing
- The model performance is evaluated by comparing predicted versus actual price values in our test set and visually represented by a histogram.
- The square foot living versus price illustrates that as the square footage increases, the price increases
- When we compare the MSE of Kings County House Data model Mean squared error: 129602842224.89789 to that of our baseline model MSE of baseline model: 137510751694.9588
- The MSE of Kings county house data is lower than the MSE of the baseline model
- The data model is performing better than our simple model

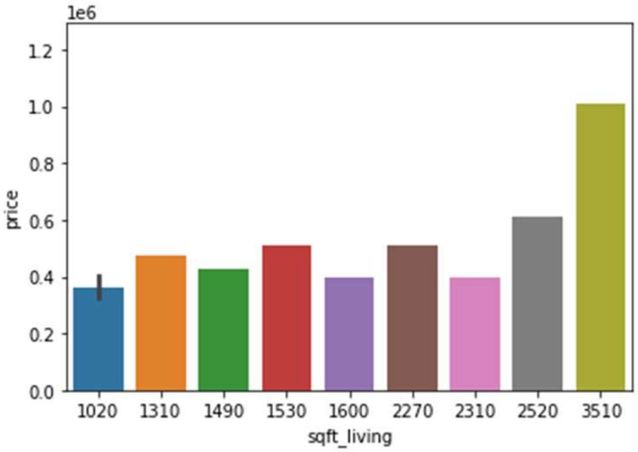


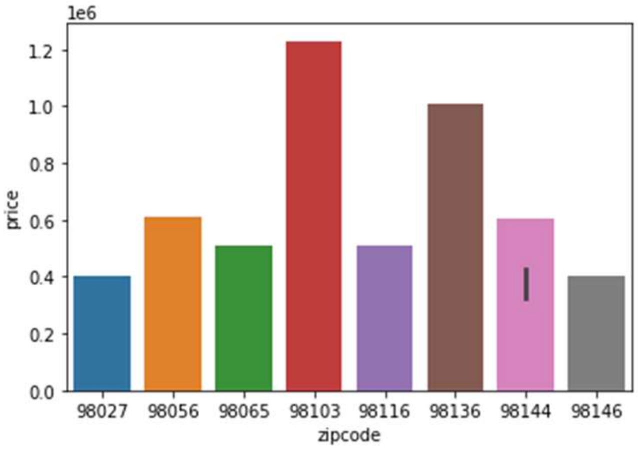


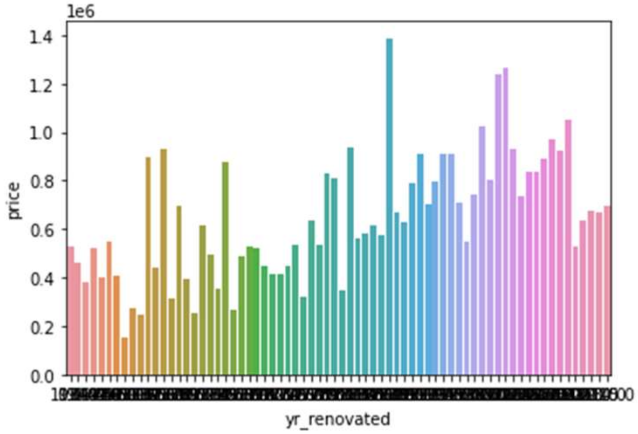




Bar Plots



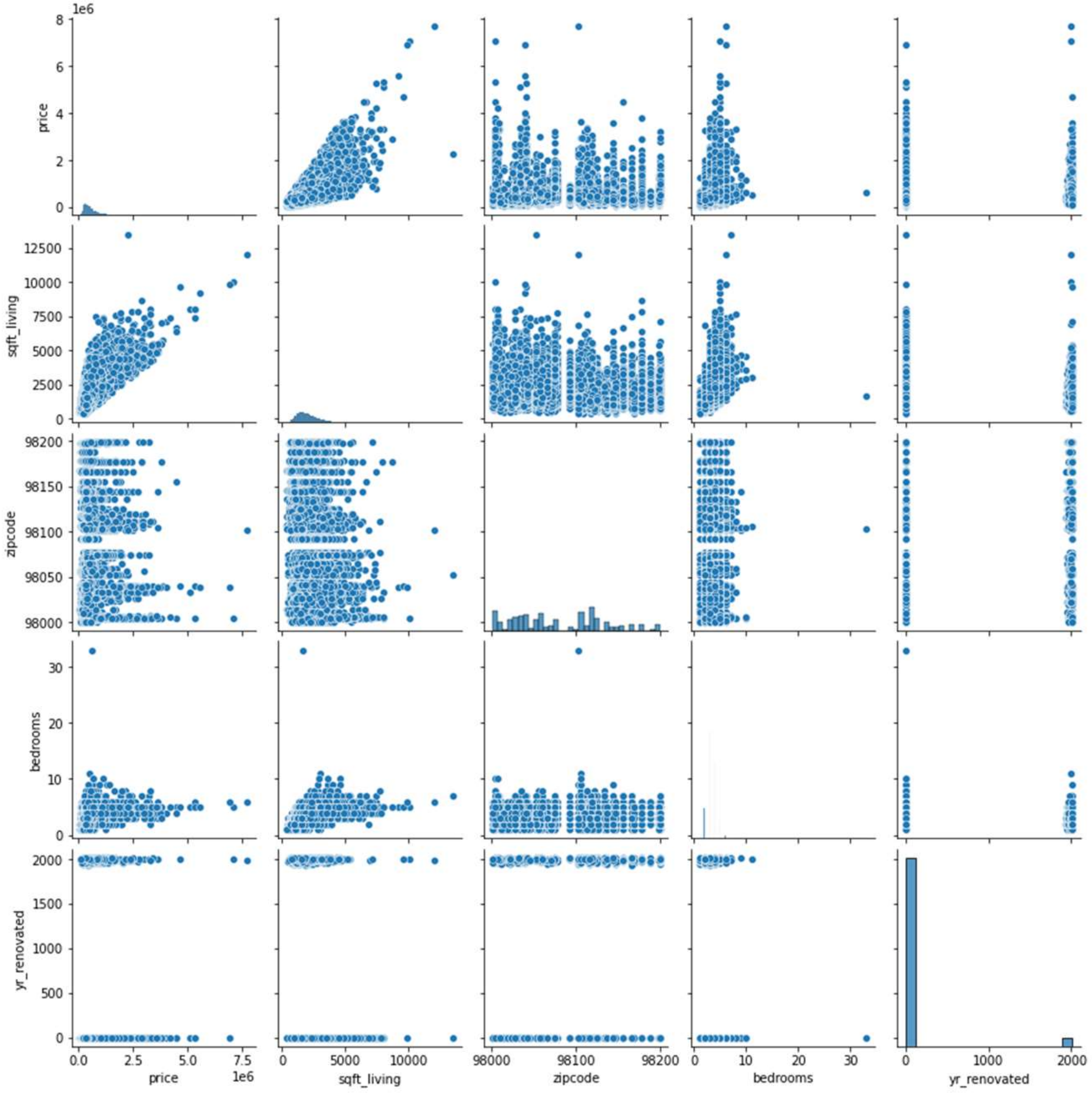




- The square foot living versus price illustrates that as the square footage increases, the price increases
- The Price increases based on year renovated meaning the newer the house, the higher the market value
- The more affluent zip codes demonstrated higher prices
- The bigger the house (square footage, no of bedrooms/bathrooms) and in a affluent zip code the better the price

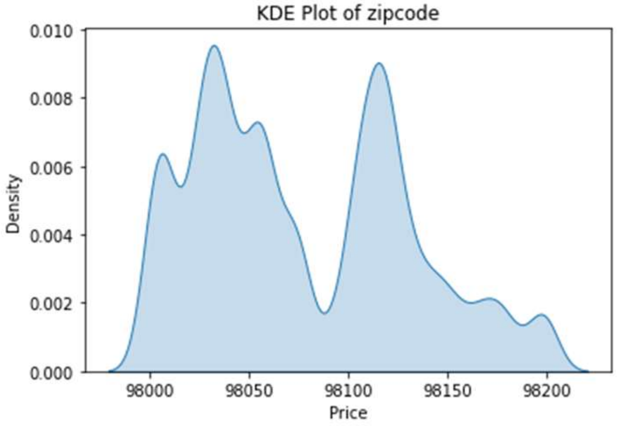
Scatter Plot Matrix

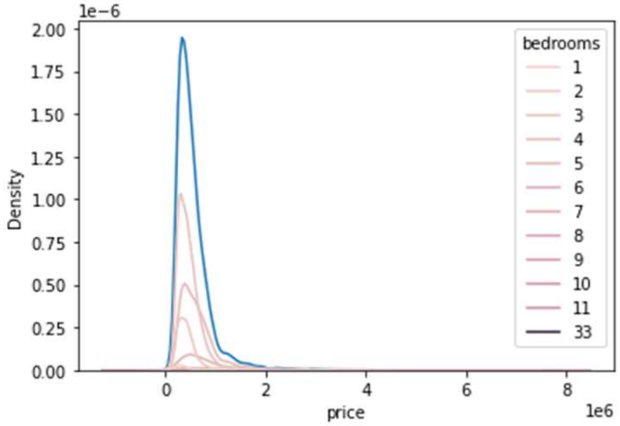
- The scatter plot matrix has been plotted using the Seaborn library.
- It demonstrates pairwise relationships between the variables price, zip code, year renovated, bedrooms and square foot living
- Each variable is plotted against itself and all other variables in the dataset.
- There is a linear relationship between price and sqft_living/zip code
- Outliers and clusters are present in the data showing that the performance of the model may not be the best fit for this data

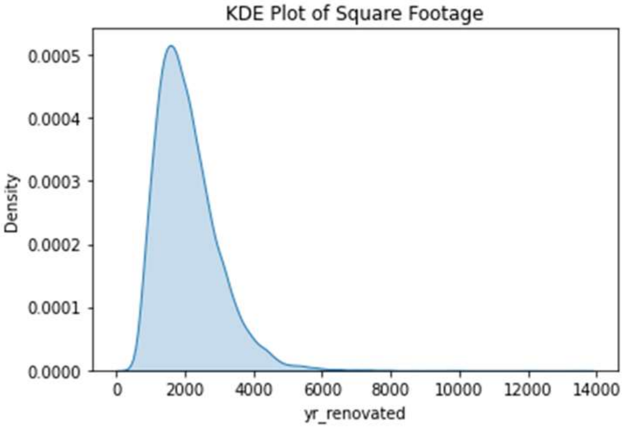


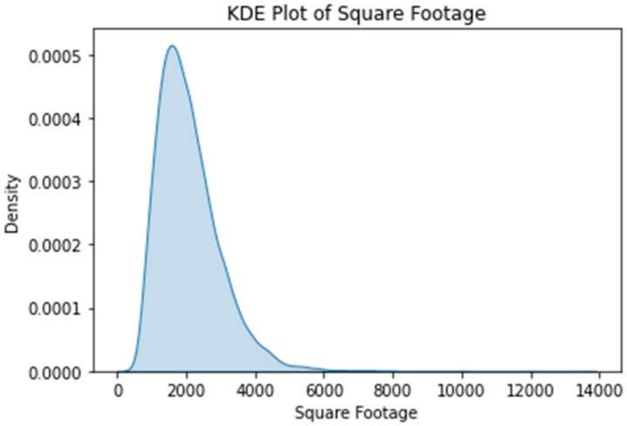
KDE plots

- Year renovated and square footage have a symmetrical bell-shaped curve indicates a normal distribution
- The Price KDE has a skewed curve indicating a non-normal distribution.
- Positive skewness illustrates that the tail of the distribution extends to the right (higher values), while negative skewness means that the tail extends to the left (lower values).
- The height of the curve at a given point represented the estimated probability density of the price variable at that point.
- The higher the curve, the more likely it is that the different variable took on the values indicated



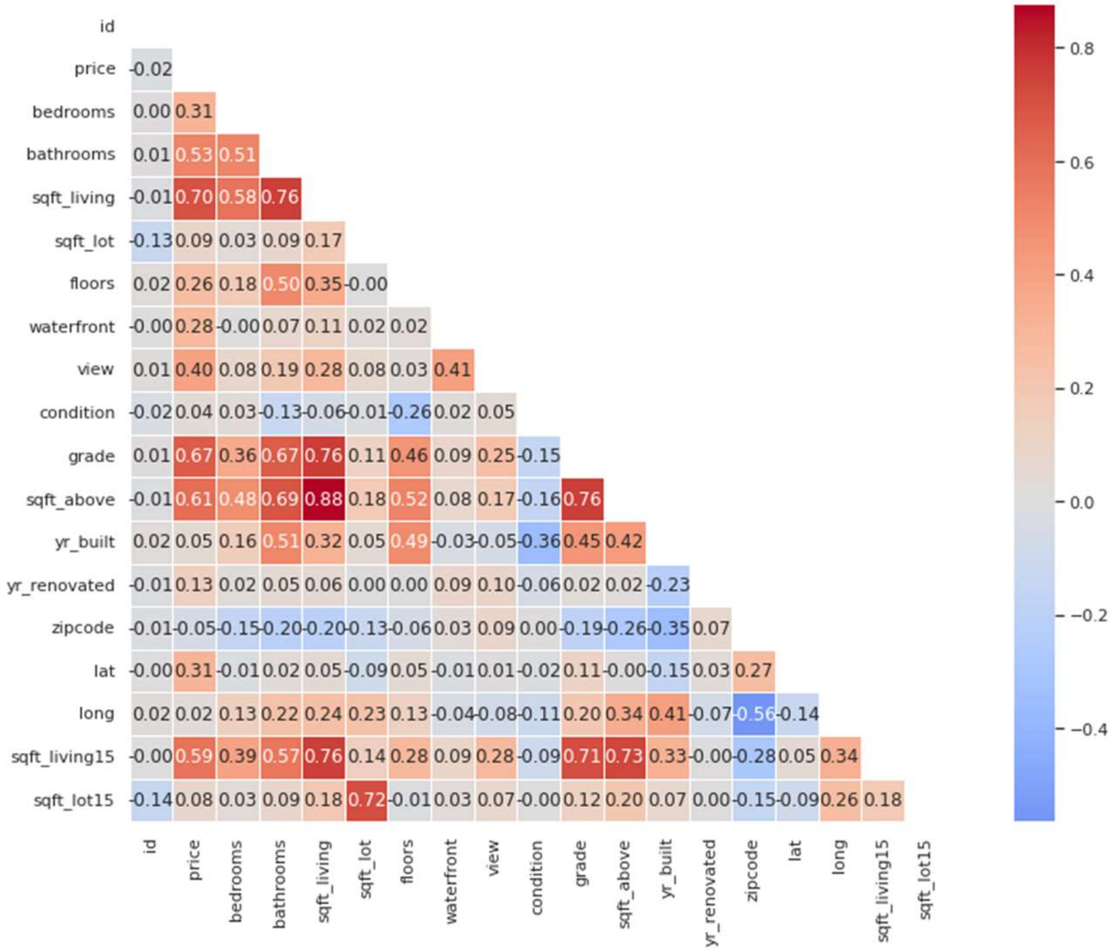






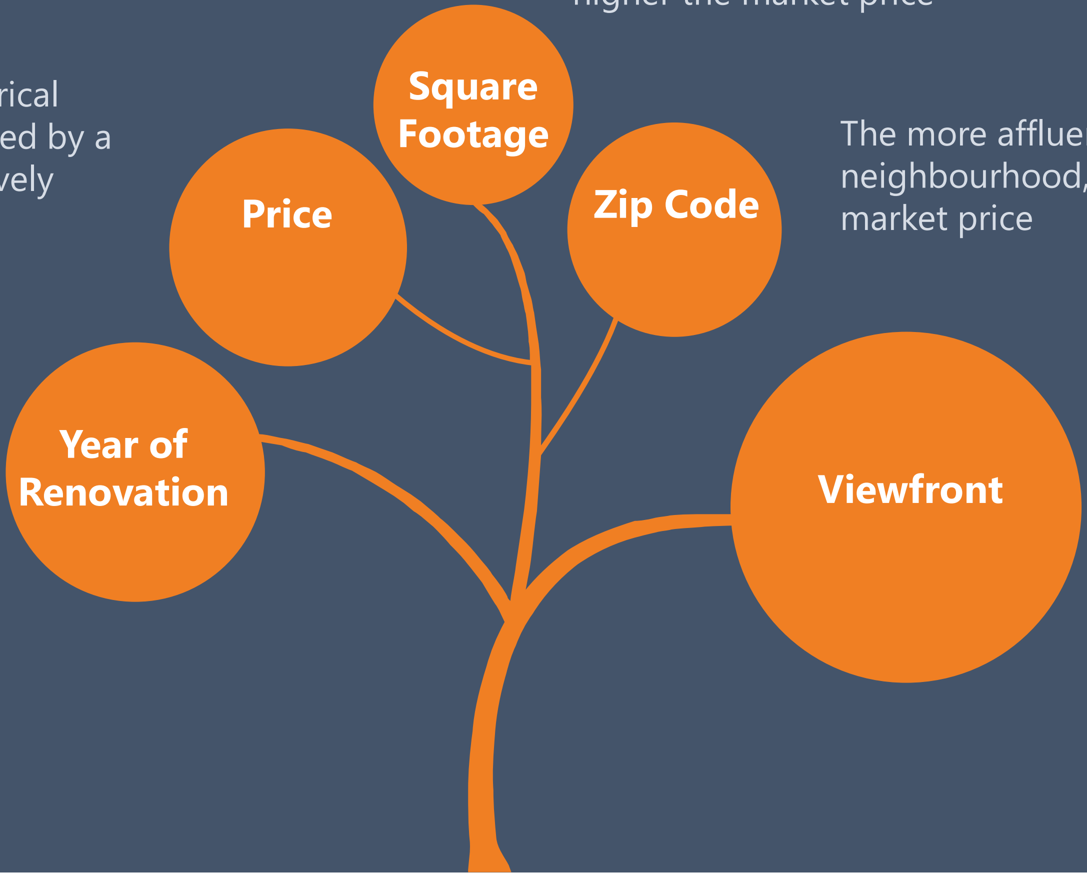
Heatmap

- Here we can visualize the correlation between different variables
- .The heatmap helps to identify the strength and direction of the correlation between the different pairs of variables, as well as identify any potential multicollinearity issues
- A high positive correlation coefficient (close to 1) between the price and square footage this variables indicate that they are positively related, meaning that as the square footage increases, the other price tends to increase as well
- There is a negative correlation between the zip code and the price, the less affluent the zip code, the lower the market rate, the coefficient is close to 0



	id	price	bedrooms	bathrooms	sqft_living	sqft_lot	floors	waterfront	view	condition	grade	sqft_above	yr_built	yr_renovated	zipcode	lat	long	sqft_living15	sqft_lot15
id																			
price	-0.02																		
bedrooms	0.00	0.31																	
bathrooms	0.01	0.53	0.51																
sqft_living	-0.01	0.70	0.58	0.76															
sqft_lot	-0.13	0.09	0.03	0.09	0.17														
floors	0.02	0.26	0.18	0.50	0.35	-0.00													
waterfront	-0.00	0.28	-0.00	0.07	0.11	0.02	0.02												
view	0.01	0.40	0.08	0.19	0.28	0.08	0.03	0.41											
condition	-0.02	0.04	0.03	-0.13	-0.06	-0.01	-0.26	0.02	0.05										
grade	0.01	0.67	0.36	0.67	0.76	0.11	0.46	0.09	0.25	-0.15									
sqft_above	-0.01	0.61	0.48	0.69	0.88	0.18	0.52	0.08	0.17	-0.16	0.76								
yr_built	0.02	0.05	0.16	0.51	0.32	0.05	0.49	-0.03	-0.05	-0.36	0.45	0.42							
yr_renovated	-0.01	0.13	0.02	0.05	0.06	0.00	0.00	0.09	0.10	-0.06	0.02	0.02	-0.23						
zipcode	-0.01	-0.05	-0.15	-0.20	-0.20	-0.13	-0.06	0.03	0.09	0.00	-0.19	-0.26	-0.35	0.07					
lat	-0.00	0.31	-0.01	0.02	0.05	-0.09	0.05	-0.01	0.01	-0.02	0.11	-0.00	-0.15	0.03	0.27				
long	0.02	0.02	0.13	0.22	0.24	0.23	0.13	-0.04	-0.08	-0.11	0.20	0.34	0.41	-0.07	-0.56	-0.14			
sqft_living15	-0.00	0.59	0.39	0.57	0.76	0.14	0.28	0.09	0.28	-0.09	0.71	0.73	0.33	-0.00	-0.28	0.05	0.34		
sqft_lot15	-0.14	0.08	0.03	0.09	0.18	0.72	-0.01	0.03	0.07	-0.00	0.12	0.20	0.07	0.00	-0.15	-0.09	0.26	0.18	

Market Analysis



The larger square footage, the higher the market price

The price was the categorical variable and was influenced by a the other variables positively or negatively

Square Footage

Price

Zip Code

The more affluent the neighbourhood, the higher the market price

The more recent the renovation, the better the market price

Year of Renovation

Viewfront

The better the view,
The faster the house sold but the price influenced the sale positively or negatively

Real Estate Opportunities

Square foot Living

The larger the house, the better the market price for the homeowners and greater ROI for stakeholder

Zip Code

The more affluent the neighbourhood, the higher the price

Renovation

The more recent and updated the house is, the higher the price and faster the sale.

Price

The more affordable the house, the faster it is for home owners to sell

Factors Affecting Real Estate Values

Location

Zip code can negatively influence the price and sale of a house .

Condition

Outdated houses that have no plans of renovations negatively influence the market rate and sale of a house

Market

The market value of a house is negatively influenced by the zip code, square foot living and year of renovation.

Improvements

Home owners who recently renovated their houses stand a better chance of selling their houses faster and at a great market price

Neighbourhood

The more affluent a neighbourhood, the better the condition and price of the house

Seller's Time Frame

The more willing a home owner is to renovate and follow the stakeholder's advice on improving their house market value, the easier and faster it is to sell the house