



# KING COUNTY HOUSE PREDICTION

## Multiple Linear Regression

By: kandie Phelma





# 1. BUSINESS UNDERSTANDING

## 1.1. UNDERSTANDING THE PROBLEM

- Selling and buying a home can be a hectic and tiring process. Homeowners rely on real estate agent's opinions and experience to guide them to the right decisions on how to market and sell their houses.
- What is more troublesome is not knowing what kinds of houses generate a lot of income or what features in a house makes the house valuable.
- There are also issues with poorly designed or condition houses that are in great geographical locations but no one to buy
- My project aims to predict the sale price of houses using the King County housing dataset. The training dataset was provided by Flatiron learning management for our study.

## 1.2. PROBLEM STATEMENT

The King County House Dataset contains a wealth of information about the price, size, location, condition and various other features of houses in Washington's King County. In this article, I'll present how I built a multiple linear regression model in Python to predict house prices.

### 1

#### Objectives

- To analyse the various variables such as living space, bathrooms ,conditions and geographical location and know their relationship with price, and use that to estimate/predict the value of houses
- We are interested in knowing how house renovation would help homeowners to be able to predict the current and future prices of their houses so that they be aware of what best time they can buy or sell the houses.
- Give advice to the stakeholders about the future prices of the houses by building or creating a model that will predict the price of these houses.
- To use different models to predict future house pricing; simple regression,multiple linear regression

A decorative graphic on the left side of the slide consists of a large cyan hexagon in the center. Surrounding it are several smaller hexagons in various shades of blue and cyan. Some of these hexagons contain white icons: a lightbulb, a thumbs-up, a network of nodes, a smartphone, a magnifying glass, a gear, and a speech bubble.

# Technical tools:

Matplotlib to visualize correlations and regression model,  
Heatmap,  
Tableau to visualize relationship between price and  
independent variables,  
Statsmodels and Scikit\_Learn to run multiple linear  
regression model

# DATA UNDERSTANDING






## 2.1. DATA COLLECTION

The dataset was provided by our institution, It contained data from King County Housing in Northwestern County



## DATA CLEANING

- we checked for duplicates, missing data and the data types
  - the sqft\_basement also has values that are 0.0 which is equal to no, null, It also has 454 ? values we are unsure of what it is, so i dropped it.
  - We can see that waterfront, view and the year of renovation has a lot of null values , so i replaced it with the mode
  - Date, sqft basement has misplaced data types: I dropped the variables
  - After that the data was also found to be consistent there being no duplicated data.
- 

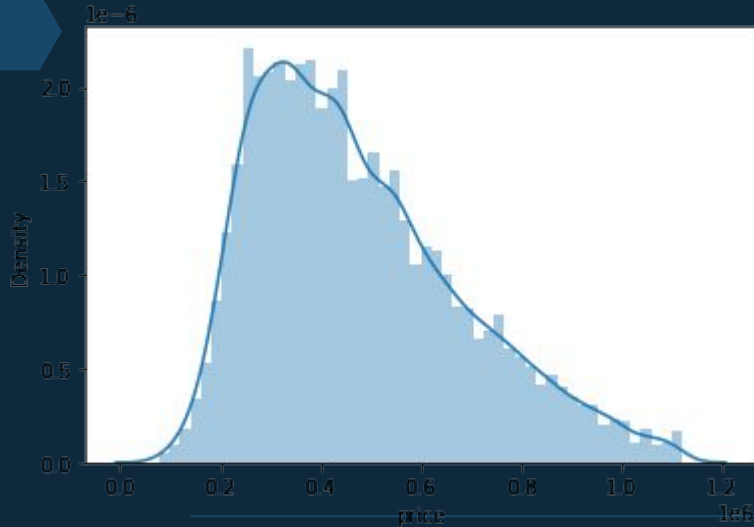
# 4. DATA ANALYSIS

## 4.1. EXPLORATORY DATA ANALYSIS

After discussing the structure of the data and cleaning the data, we have to visualize our data so that it can help us create a good model for our prediction.



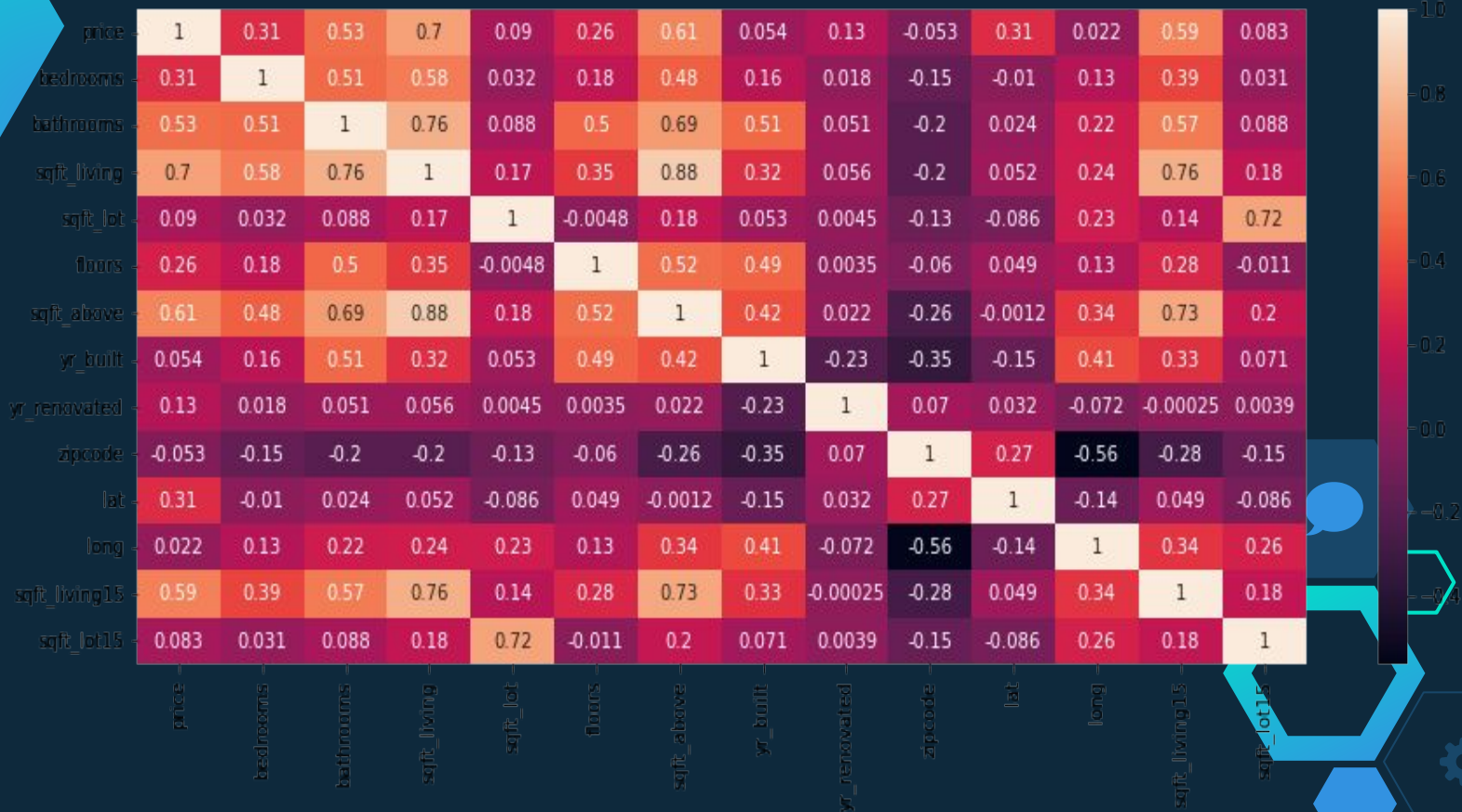
# Univariate Analysis



- I checked the price variable, analysed for outliers and then removed the outliers
- I checked for skewenes before and after removing our outliers
- The histogram is not normal; positively skewed graph



# Bivariate Analysis



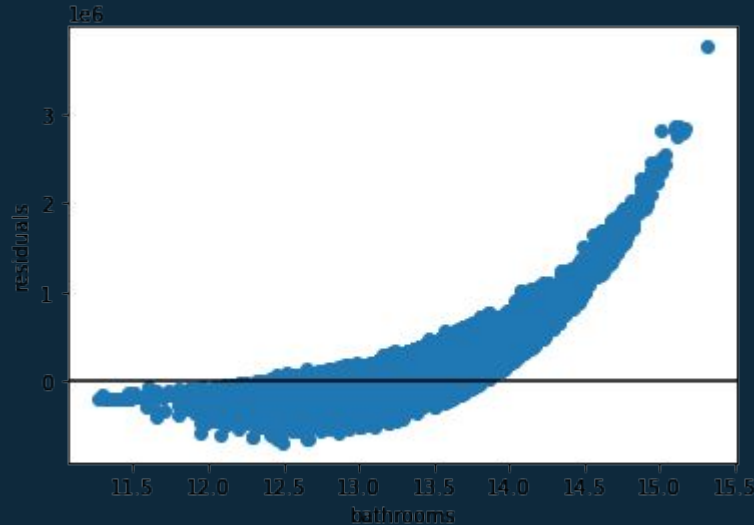


## bivariate summary:

- Sqft\_living has high correlation between the independent variables: Generally, we don't want to include any two x variables whose correlation exceeds .80 in the same model. Therefore will drop it. We do, however, want to include variables that are correlated with the y variable.
- The sqft\_living and sqft\_living15 have a high correlation of 0.76
- sqft\_living has the most correlation with price followed by sqft\_above

# MODELLING:

## Simple Linear Regression



- F\_statistics is less than 0.05, therefore the model is significant.
- The bathroom variables does help predict prices
- Our model has 24% variance in price
- For every increase in 1 bathroom the price of the house increases by 229172.669988 dollars
- Mean Absolute error is 194701 meaning that there is a difference of 194701 between the actual price and our predicted price



# Multiple Linear Regression

## First Model

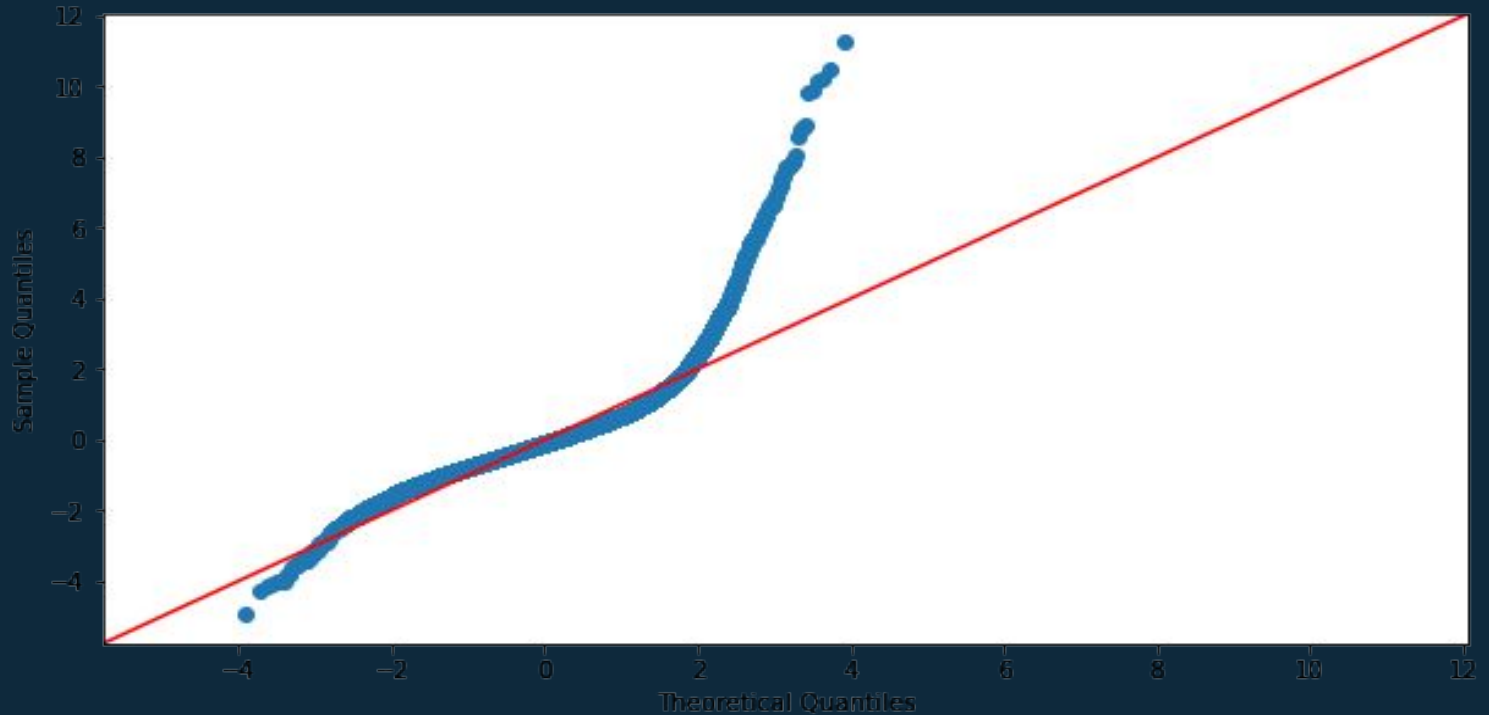
- $F_{\text{statistics}}$  is less than 0.05, therefore the model is significant.
- Our model has 48% variance in price which shows an improvement from our previous model
- For every increase in 1 bedrooms the price of the house increases by 229172.669988 dollars
- For every increase in 1 square foot living in 15 neighbourhoods, the price of the house increases by 99.703009dollars
- For every increase in 1 square foot living in 15 neighbourhoods, the price of the house increases by 99.703009dollars
- Mean Absolute error is 160924 meaning that there is a difference of 160924 between the actual price and our predicted price: Our MAE has reduced showing that the more you add independent variables the more accurate your prediction will be

# Second model

condition	grade	
<b>16458</b>	Average	7 Average
<b>6405</b>	Average	8 Good
<b>9148</b>	Good	8 Good
<b>21561</b>	Average	9 Better
<b>9720</b>	Good	7 Average
<b>12517</b>	Average	6 Low Average
<b>17361</b>	Good	7 Average
<b>4660</b>	Average	7 Average
<b>6430</b>	Average	9 Better
<b>19370</b>	Average	7 Average

# QQ PLOT

Residuals QQ Plot





## 5. REGRESSION RESULTS:

- Nearly all of the model coefficients have p-values less than 0.05 and are thus statistically significant.
- The bathroom variables does help predict prices
- Our model has 58% variance in price which shows an improvement from our previous model
- The more the variables the higher the significance/Accuracy level
- For every increase in 1 square foot living in 15 neighbourhoods, the price of the house increases by 99.703009dollars
- Mean Absolute error is 145001 meaning that there is a difference of 145001 between the actual price and our predicted price: \* Our MAE has reduced showing that the more you add independent variables the more accurate your prediction will be



# Limitations:

- This data is only suited in one geographical region therefore it can be hard to use the same model in different locations
- Messy data can be harder to create a model that has high significance
- given that some of the variables needed to be log-transformed to satisfy regression assumptions, any new data used with the model would have to undergo similar preprocessing.



## 6. RECOMMENDATION

- ◇ It would be great if the homeowners increase the size of their houses: this includes having more bathrooms as this will increase the price they have to earn
- ◇ Homeowners should renovate and make the conditions of their houses to be very good as this attracts more cash flow compared to poor conditioned housing
- ◇ If the homeowners can invest more on excellent grade housing, the higher the grade the more the price of the house
- ◇ There is no need for building square feet above space in the house since it does not really relate to the prices
- ◇ Have small number of bedrooms
- ◇



# Thank you!

## Any questions?

You can find me at:

- ◇ @phelma.kandie@student.moringaschool.com
- ◇ Github: Kandy372





# Adios!

