**Group 7 DSF-PT05**

**Members**

1. Linet Wangui-Lead
2. Brian Kanyotu
3. Esther Gakio
4. Beryl Wafula
5. Christine Gitau
6. Allan Omollo

### **PROJECT OVERVIEW:**

**What problem is the project aiming at solving?**

* The project purposes to answer the question: "What housing attributes impact the value of houses in King County?"
* In analysis, the prices of houses in the county are dynamic and continuously shifts (according to statistics-over the years). This possess a challenge to buyers and sellers in determining house prices. Both are curious to know the predictors of house values in reference to price.
* Every Door Real Estate Agency assumes the challenges faced by its clients (Home buyers, sellers, renters, investors, and businesses) would wish to provide insightful information into how much pre-approval they need for their ideal property.
* To help the Agency realize its set objectives, we will conduct a linear regression analysis.
* We intend to use the analysis to provide recommendations and conclusion about the business problem using evidences from the analysis.

**Who are the beneficials?**  
Home buyers, sellers, renters, investors, and businesses

**What are the objectives of the analysis?**  
a. Determining house attributes impacting prices of houses  
b. Providing recommendations to guide the process of buying and selling houses  
c. Create awareness of the role of regression analysis in enhancing evidence-based decision making processes

## Dataset

**Briefly provide a description and understanding of the data used in your analysis?**

* **Source:** <https://github.com/learn-co-curriculum/dsc-phase-2-project-v2-3/tree/main/data>
* **Name of data:** kc\_house\_data.csv
* **Nature of data:** House properties sold from 2014 to 2016
* **Data Structure:** 21597 observations and 21 columns describing the property attributes
* **Significance:** appropriate and useful for regression analysis since the data suggests correlations between prices and house attributes

**What do the different columns describe for King County Data Set?**

1. **Id** - Unique identifier for a house
2. **Date** - Date house was sold
3. **Price** - Sale price (prediction target)
4. **Bedrooms** - Number of bedrooms
5. **Bathrooms** - Number of bathrooms
6. **Sqft\_living** - Square footage of living space in the home
7. **Sqft\_lot** - Square footage of the lot
8. **Floors** - Number of floors (levels) in house
9. **Waterfront** - Whether the house is on a waterfront  
   Includes Duwamish, Elliott Bay, Puget Sound, Lake Union, Ship Canal, Lake Washington, Lake Sammamish, other lake, and river/slough waterfronts
10. **View** - Quality of view from house  
    Includes views of Mt. Rainier, Olympics, Cascades, Territorial, Seattle Skyline, Puget Sound, Lake Washington, Lake Sammamish, small lake / river / creek, and other
11. **Condition** - How good the overall condition of the house is. Related to maintenance of house.  
    See the King County Assessor Website for further explanation of each condition code
12. **Grade** - Overall grade of the house. Related to the construction and design of the house.  
    See the King County Assessor Website for further explanation of each building grade code
13. **Sqft\_above** - Square footage of house apart from basement
14. **Sqft\_basement** - Square footage of the basement
15. **Yr\_built** - Year when house was built
16. **Yr\_renovated** - Year when house was renovated
17. **Zipcode** - ZIP Code used by the United States Postal Service
18. **Lat** - Latitude coordinate
19. **Long** - Longitude coordinate
20. **Sqft\_living15** - The square footage of interior housing living space for the nearest 15 neighbors
21. **Sqft\_lot15** - The square footage of the land lots of the nearest 15 neighbors

**What regression models were used in the project?**  
a. simple linear regression to form the baseline  
b. Multiple linear regressions (various models were fit and the best model selected)

**PROJECT METHODOLOGY**

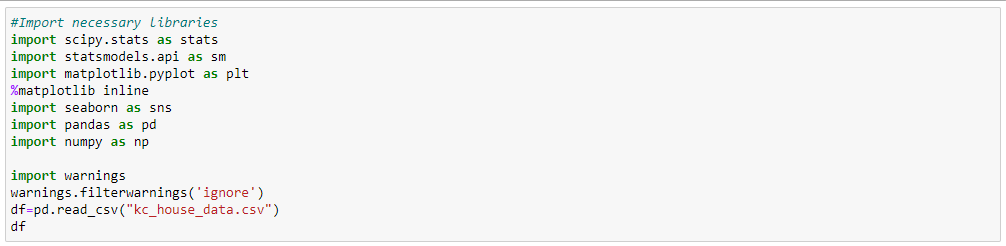
**What procedures, tools, and methods did you use?**  
The project methodology is divident into a pre-preliminary, preliminary, and post-preliminary procedures.  
**Pre-preliminary procedures:** In this phase, all the necessary libraries were imported and the data loaded using pandas library. The data was then cleaned and wrangled to make it more significant. Additional cleaning including replacing missing values, dropping insignificant data, filling null values, and dropping duplicates using relevant functions and methods was also done to further make the data more valuable.  
**Preliminary procedures:** Data correlation was checked. A baseline constituting simple linear regression model was fit. The model helped develop a multiple linear regression. These models helped predicted the value of houses. Graphical representation of the analysis further enhanced to establish correlation towards determining how house attributes impacted price.  
**Post-preliminary procedures:** In reference to best model, recommendations, explanations, and conclusions were generated in this phase.  
The main tools used in the project includes libaries like matplotlib, numpy, pandas, seaborn, scipy, and others.

**RESULTS**

The variables 'sqft\_living', 'bathrooms', 'sqft\_above', 'sqft\_living15’, and 'view are the best predictors of house prices in the county. These variables were used to modeling the final multiple regression. The model satisfied all multiple regression assumptions, and p-values for each predictor variable were below .05. The model contained an R-squared of 50.6% and an (MAE) Mean Absolute Error of 0.2994. This tells the model is off by 0.299 thousand dollars in a given prediction

The project followed the following steps to achieve set goals and objectives

**Pre-preliminary procedures**



**Provide a detailed description of the data preparation process?**

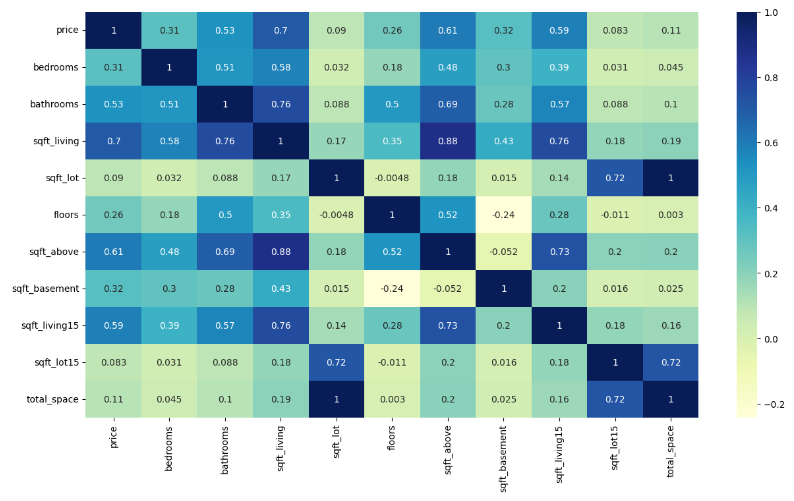
The process of data cleaning and wrangling entailed formatting the data to remove duplicates, filled missing values, deleting undesirable symbols, and selecting required property variables for further modelling. From the observations, the data did not have any duplicates.  
However, the variables:  
a. Waterfront, yr\_renovated, and view contained missing values which were filled with appropriate data. The missing values in waterfront were replaced with “NO” to assume the households lacked waterfronts. The missing values in view were replaced with “NONE” to assumed a lack of view. The missing values in yr\_renovated were replaced with 0 to mean the houses were never renovated.  
b. The values in the variable bathroom were standardized to decimal places because it was observed the values lacked uniformity of decimal places.  
c. The date section was standardized to: day, month, and year format for uniformity.  
d. The sqft\_basement contained undesirable symbols which were removed, replaced with mean, and the data converted to numerical values. The data was also converted to 2 decimal places for uniformity.  
e. The values in the latitude and longitude variables were converted to 2 decimal places for effective analysis and modelling.  
f. The whole data set was confirmed clean and ready for modelling using the info() method.  
g. During wrangling, the columns 'zipcode','id', 'lat', 'long', 'yr\_renovated', 'yr\_built' were observed to be insignificant and consequently were dropped.

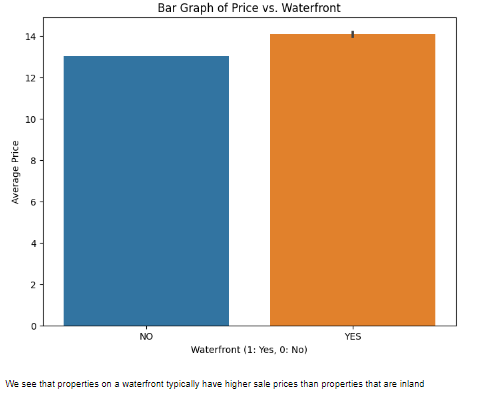
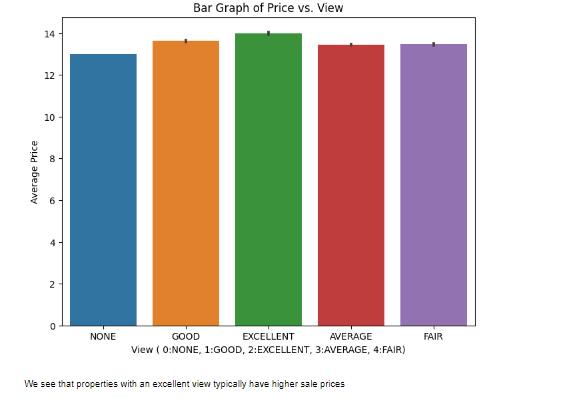
**Were there any data transformed in the pre-preliminary phase?**  
While most of the variables were deemed fit for modeling, some data had to be transformed for healthy consumption. Some data had to be transformed to suitable data types. Additional columns were included forming new features. For instance, a column, total\_space, was created to include sqft\_living and sqft\_lot variables. An additional season column was added to analyse how housing prices across the different seasons. The columns were then retained along with the transformed data.

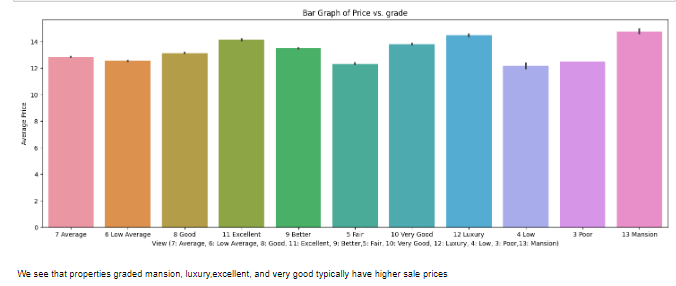
The data was then check for correlation visually and statistically to determined relevant columns for use in modelling. When checking for correlation, variables than were greater than .5 were considered more correlated to price (with a correlation of 1.0). These variables were retained for machine learning and modelling through a multiple regression analysis. Additionally, the selected variables were checked for skewness. All the variables were skewed and the data was standardized to normal distribution. The outliers were checked and removed to make the regression model for effective. Lastly, variables with categorical data were converted to dummy variables and the data added to the dataframe

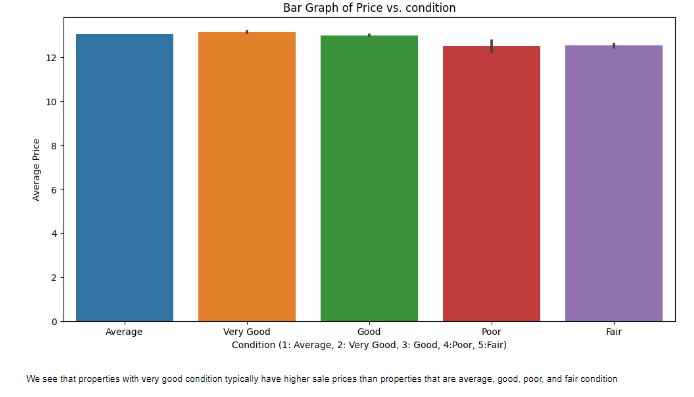
**Preliminary Procedures**

* The summary of the transformed data was checked using describe method
* Correlation was checked to establish relationship between variables
* The aim of the preliminary procedures was to investigate the relationship between property attributes and prices. The relationship between selected variables is summarized in the heat map and graphs seen below.
* Correlation heatmap shows a summary of the correlation between the different variables in the data set in relation to price

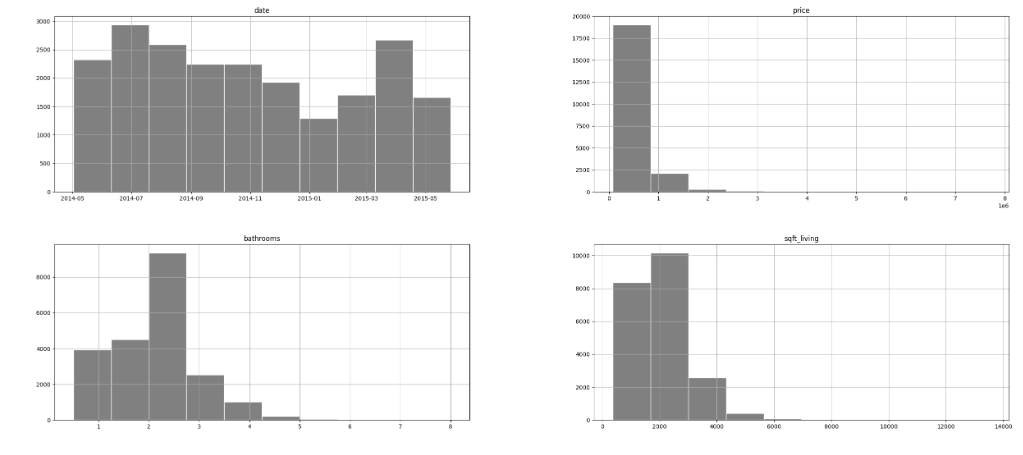


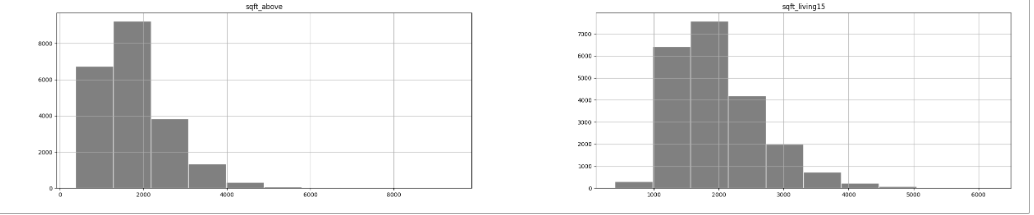
 



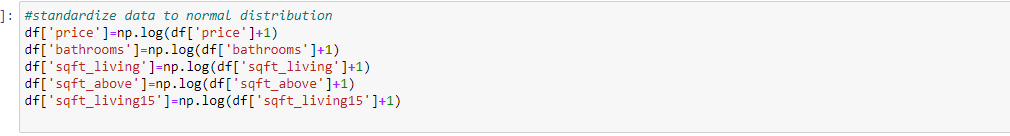


* The least correlated variables were dropped and data checked for normal distribution as seen in the figures below
* Basic correlations in the data

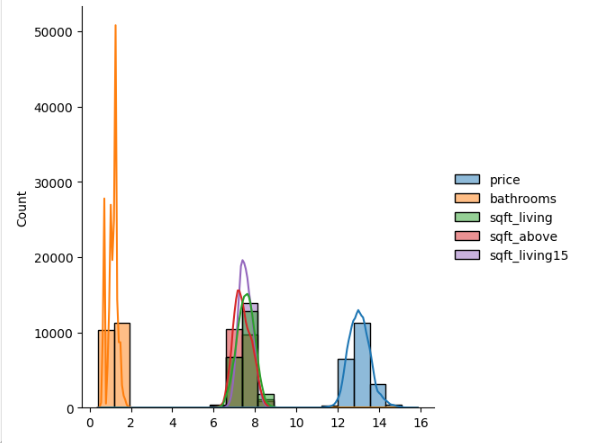




* The data set was then standardized for effective analysis and make the data variables assume normal distribution curves



Distribution Displot

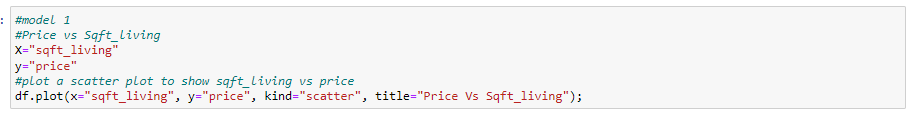


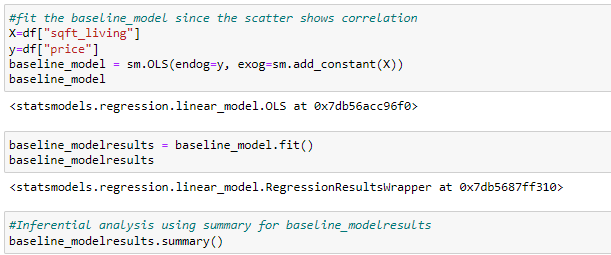
The bathroom variable showed an abnormal

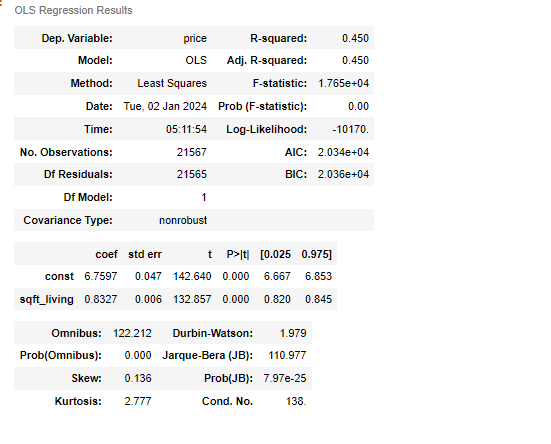
**REGRESSION MODELS**

1. Baseline Model

The model composed of a simple linear regression the based on price vs sqft\_living.  
sqft\_living showed the highest and strongest correlation hence fit to create a baseline model for the multiple linear regression

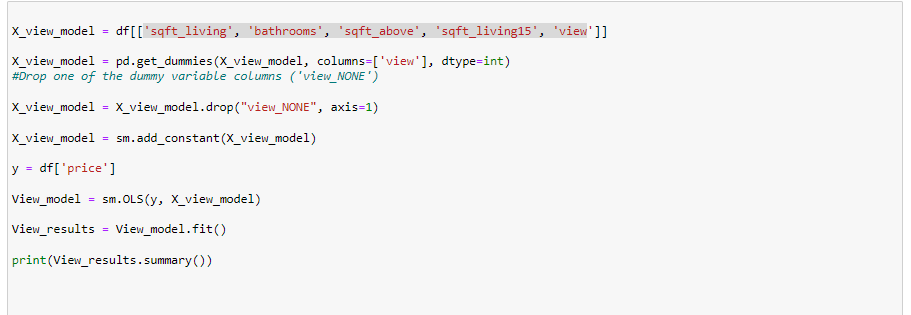


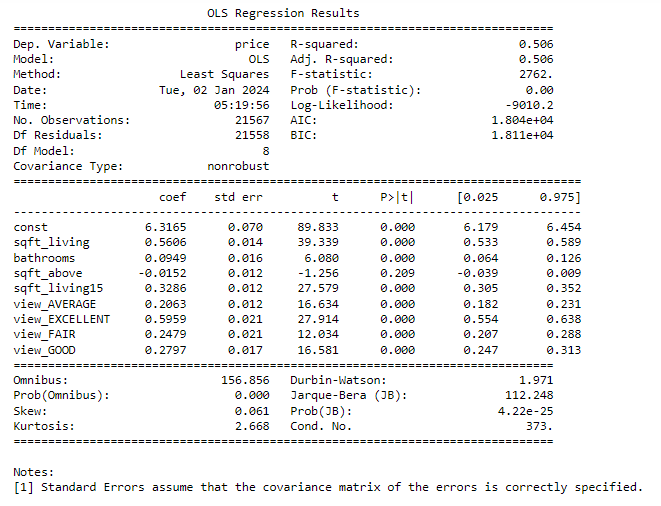




1. Final Model

Multiple linear Regression





The final model accounts to 50.6% of the variance in price and it’s a good model for use. The mean absolute error is 0.29 providing more evidences that the model is best and fit for use since it is a better improvement of the first model. The model strongly supports alternative and null hypothesis. We found that the p<0.05 which insists that we reject the Null Hypothesis that Price of house are independent.

**RECOMMENDATION**

It is highly recommended that the audience focus on the attributes of 'sqft\_living', 'bathrooms', 'sqft\_above', 'sqft\_living15’, and 'view’ when deciding on prices of houses. These variables are the key predictors of house prices in King County. The variables showed strong positive correlations suggesting that an improvement in the variables increases prices.

Model Selection. When choosing a model, it's important to balance complexity and performance. Depending on your objectives, you might opt for a simpler model with fewer features, especially if interpretability is a key priority.

**LIMITATIONS**

The choice of variables included in the models may not be exhaustive. There could be other important factors influencing house prices that are not considered in the current models.

**CONCLUSION**

House properties influence prices of houses. The model suggests 'sqft\_living', 'bathrooms', 'sqft\_above', 'sqft\_living15’, and 'view’ influence prices of houses in the county.