**KC DATASET.**

**INTRODUCTION.**

The aim of this project was to put into practice our newly acquired skills in data science to formulate and solve a real business problem. In this project, we were given a raw dataset which is the house sales prices in King County and we were supposed to formulate a real business case study and use the data to solve the business problem and provide a useful business advisory which can be used to formulate real and valuable business decisions.

**PROBLEM STATEMENT.**

The business problem was to provide guidance to Kings Wajenzi Developers, a prospective real estate developer in King County, on the most profitable types of properties to target in the current market. The guidance included recommendations on optimal locations to develop properties, features most desired by customers and had the greatest impact on prices of properties, and season of year when homeowners are most likely to purchase properties. Our objective as Wataalamu Analytics Advisors was to analyze house sales data in King County which were collected between May 2014 and May 2015 and leverage the insights generated to guide the developer in selecting the most profitable properties to develop and maximize their profits. With King County's population on the rise and income stabilizing after the COVID-19 pandemic, demand for new homes in the area is at an all-time high.

**OBJECTIVES.**

* To determine whether the time of year affected the price of a house and identified the most profitable seasons/months.
* To investigate whether location affected house prices in King County and identified the areas that attracted the highest prices.
* To determine which features/attributes had the highest impact on the sales price of houses/properties in King County and identified the features that developers should focus on when developing upcoming projects in the area.

**METHODOLOGY**

***This project followed the OCEMiN Data Science framework for data analysis***

1.**Data Importation**

* We imported data from King County House Data provided as csv, using Pandas library.

2. **Data Cleaning**

* We employed several data cleaning methods to modify and prepare the dataset for analysis.

3.**Data Exploration**

* We analyzed and visualized the cleaned data to gain insights, identify patterns and relationships, and formulate hypotheses.

4. **Data Modelling**

* We developed and evaluated predictive models using the data.

5. **Data Interpretation**

We drew conclusions and made decisions based on the results of the analysis

**BUSINESS UNDERSTANDING.**

The business problem was to provide guidance to Kings Wajenzi Developers, a prospective real estate developer in King County, on the most profitable types of properties to target in the current market. The guidance included recommendations on optimal locations, features, and time of year when home owners are most likely to purchase properties. Our objective as junior data analysts at a Real Estate Agency was to analyze house sales data in King County and leverage the insights generated to guide the developer in selecting the most profitable properties to develop and maximize their profits. With King County's population on the rise and income stabilizing after the COVID-19 pandemic, demand for new homes in the area were at an all-time high.

**DATA UNDERSTANDING**

The whole dataset was stored as pandas data frame. We had three types of data types for columns which were integers, floats and objects. Since we were interested in running a linear regression model, we changed all the columns to be numeric data types. However, some columns like date, waterfront, view, condition, grade, square foot basement were objects and most likely categorical variables and needed to be transformed to dummy variables using one-hot encoding function of python. Intergers were also examined to determine if they were binary variables or numeric data types. It was very important to note that square foot basement needed to be of value type float based on the data preview done before. Therefore, we investigated that column to find out why it appeared as object instead of float. The date column was transformed to month and year columns as we seek to engineer a new feature named seasons to answer our first question. The columns waterfront, view, condition, were investigated and necessary transformations done in order to make them dummy variables for our model.

**DATA CLEANING**

We checked for the missing values in the dataset and from our output, we realized that waterfront, view and year renovated had missing values. We then checked the percentage of missing values in the dataset. We calculated the percentage of missing values to see their magnitude and determine whether to drop the columns or just replace the missing values with appropriate replace values.

Waterfront, view and year renovated columns had 11.0%, 0.29%, and 17.79% respectively of missing data. Waterfront and view were of data type objects and therefore, we replaced the missing values with the mode. The year renovated showed the year in which renovations occurred and therefore, we interpreted the missing values to mean that there was no renovation that occurred for those particular properties. Therefore, we still replaced the missing values with zeros.  
We had two duplicates and upon closer examination, the two rows were not duplicates but errors may have risen as a result of data updates, data merging or entry errors. We then proceeded to investigate the integer and object data types and performed the necessary transformations for our multiple regression model.

Bathrooms appeared as float and we expected them to be integers. Upon researching from common real estate jargon in the United states, bathrooms can be described with floats or fractions to indicate not a full bathroom or with an additional feature such as a sink. We then checked for outliers. After thorough data clean up, we fine-tuned the data for plotting, visualization and subsequent modelling. We dropped view, grade, condition, zip code because they are categorical variables and therefore cannot be reasonably visualized using either box plots or histograms. Floors, age, year Sold, month sold do not appear to have outliers. Therefore, we considered potential outliers to be in bedrooms, bathrooms, square foot living, square lot, floors and squire above features.

The outliers were removed by applying the statistical method that calculates the z-score of the feature columns. The function took in the data frame, calculated the z-score per column then removed any value that fell within z-score >|3|.

**EXPLORATORY DATA ANALYSIS**

We proceeded to investigate price and its relationship with house features, however since price had outliers, for example, houses with very high prices and others with very low prices. We grouped price into three categories; Upper price, Median price and Low price where, upper price had 75% percentile, median price ,50% percentiles and the low price, 25% percentile. We plotted a histogram to analyze these findings.

From the Histogram we observed that the majority of median and high-end properties had a higher density with two bathrooms and an additional bathtub, toilet or sink. From the high end the density was more towards the median showing that there is little difference that distinguished them from the median. For the low price range the density was more on one bedroomed houses indicating a prevalence of studio apartments. The histograms suggest that the number of bathrooms is positively correlated with the price of properties. However, we performed statistical analysis to confirm this relationship and identify the strength of the relationship.

We analyzed the relationship between season and prices of properties using scatter plots and the results indicated that based on the count of properties sold in each season, it appeared that spring and summer are the most popular seasons for home buying, while fall and winter have comparatively fewer sales.

The scatter plots of high-end, medium, and low-price properties against the season showed that the prices of properties were generally consistent across all seasons. However, there was a slight increase in the number of high-end properties sold in the spring and a decrease in the number of low-priced properties sold in the winter.

This pattern may be due to various factors, including the perceived desirability of certain seasons for buying and selling homes, the availability of listings during different seasons, or even the preferences of buyers and sellers. For example, the spring season may be associated with better weather conditions and more opportunities for home viewing, which may increase demand for high-end properties. Conversely, the winter season may be associated with lower sales of low-priced properties due to holiday expenses and colder weather. However, we carried out further analysis using bar to determine the exact reasons for these trends.

From the results, it is evident that the spring months of March, April, and May were the best for selling houses with a total of 6518 sales. This was likely due to the fact that Spring is a time of renewal and growth, and people may be more willing to make big changes such as buying a new home during the time. May was the most popular month for selling homes within Spring, with 2414 sales. Summer, which comprises the months of June, July, and August, had a slightly lower total sale of 6328 compared to spring. July had the highest sales with 2211 followed by June with 2178 and August with 1939. The fall season of September, October, and November had a total of 5056 sales, with October having the highest sales of 1876 followed by September with 1771 and November with 1409. Lastly, the winter season of December, January, and February had the lowest total sales of 3695. December had the highest sales with 1470 followed by February with 1247 and January with 978. This may be due to the fact that Winter is a time when people tend to stay indoors and may be less inclined to go through the hassle of buying or selling a home.

We also analyzed how location could possibly affect price of houses by geographical distribution and from the visualizations, the highest house prices are concentrated in the area with latitude around 47.6 and longitude around -122.25. There was a disparity with southern locations achieving lower house prices.

We analyzed the relationship between waterfront features and the prices using box plot and from the plot, most houses lacked this feature, however the ones that had, showed highest price rate.

**DATA MODELING**

Here, we developed a model that accurately predicted sales price of properties in King County, given various exogenous variables. The modeling procedure encompassed the selection of suitable algorithms, feature engineering, fine-tuning of hyper parameters, and evaluation of the model's performance.

**CHALLENGES.**

We faced the following challenges in our data analysis

**1. The problem of missing data in 3 columns, view, year renovated and waterfront:** We resolved this by replacing the missing values in columns in view and waterfront by mode and in the year renovated, we replaced the missing vales with 0.

**2. The problem of place orders in the square foot basement Outliers:** The dataset contained extreme values that could skew the results of the analysis.

**3. Non-linear relationships:** The relationships between the independent variables and the target variable may not be linear, which makes it difficult to fit a linear regression model.

**4. Multicollinearity:** The independent variables may be highly correlated with one another, which can lead to multicollinearity which makes it difficult to interpret the coefficients of the model.

**RECOMMENDATIONS AND CONCLUSIONS**

**1. Season of the year**

Sales of properties is highest in spring and lowest in winter. Property developers such as our client Kings Wajenzi developers should target to develop properties that will be market in the season of spring and Summer.

**2. Location**

Location within King County is also a crucial factor, with significant price disparities among zip codes. The location with zip code 98112 has the most priced properties with 98001 having the least priced properties. We suggest that developers targeting high end clients should develop homes in zip codes 98112, 98109 and 98105. Those targeting low income earners should target to develop properties in zip codes 98001, 98106 and 98108. Those targeting middle income earners should target other zip codes.

**3. House features**

The most important predictor of house price is view. We observed that house prices with waterfront view are the most desirable. Houses that have a waterfront view have median price that is almost twice as those that don’t have waterfront view. We suggest that developers target to develop homes in front of water bodies such as lakes and rivers as these will fetch highest prices.

4. **Bedrooms**

Number of bedrooms is the second most important predictor of house prices. Properties with higher number of bedrooms fetch high prices and hence developers should target high number of bedrooms where they wish to fetch high price for properties. However, incremeament beyond 6 bedrooms does not guarantee highest prices. The optimal number of bedrooms is 5 bedroom houses which fetch highest prices.

**5. Bathroom**

Number of bathrooms is not an important predictor of houses. Both high and medium priced properties have an optimal number of bathrooms which is 2.5. Low priced properties have 1 bathroom as the optimal. Therefore, developers should ensure high and medium priced properties have an optimum of 2.5 bathrooms.

**6**. **Grade**

The median house price increases with grade, suggesting a positive correlation. We believe that grade is a reliable indicator of price, so we recommend focusing on homes with a grade of 10 or higher.

**7. Square footage and condition**

Square footage and the condition of the property are also significant factors that affect the price of a house and should be considered by developers when developing their properties.

**FURTHER ANALYSIS**

Our model had shortcomings initially as it only explained less than 50% but after improvements, it was able to predict almost 100% of variations in the response variable. This is because of proper data understanding, cleanup and fine tuning it. However, to gain new insights, we can consider additional factors such proximity to the capital city of Washington State, the income per zip code, presence of schools, hotels, commuter trains, and entertainment facilities. Longer time span should also be considered as the data used here was for 1 year only.