King County Houses Multiple Regression Analysis

Business Overview:
According to the 2020 census, King County was the most populous county in Washington, and the 13th-most populous in the United States. It is in Seattle, the state's most populous city. Give the statistics, Tella Real Estate Agency, based in King County, has undertaken a research to find out the best performing metrics when it comes to house sale prices determination. As such, we went all out to do multiple regression analysis to gain insights into the home sales market and improve the sellers' chances to make sales. By identifying the key factors that affect the sale prices of houses, the agency can develop more effective marketing strategies, help sellers target the right buyers, and make better investment decisions.

Problem Statement:
Tello Real Estate Agency has a dataset of home features collected from the county houses. They include the number of bathrooms, number of bedrooms, number of floors, square footage of living space amongst others, as seen in the dataset attached. The agency wants to understand the relationship between these features and the sale prices of houses in King County, and particularly, which features most affect the prices.

Objectives:
The following are the questions whose answers we have come up with in this analysis:

- To understand the relationships between the various features and the sale price of a house.
- To determine the peak house sale season.
- To build a multiple linear regression model that identifies the most important factors that influence the sale price of a house.

- To use the model to gain insights into the home sales market and improve decision-making processes.

Data Understanding:

- The data is King County House Sales dataset which contains 21 columns and 21597 rows.
- We loaded the dataset and checked for missing values, outliers, duplicates and inconsistencies in our data.
- We discovered that some of our data was categorical, so we proceeded to perform feature engineering described below.
- We did univariate analysis to understand the distribution of our individual variables, and later compared these to our target variable, price, in bivariate and multivariate analysis.

Feature Engineering

- This involved performing one-hot encoding and label encoding on our categorical columns to transform them to numerical columns.
- We also checked for correlations of several variables with our target variable, price, and the multicollinearity between the independent variables.
- For the non-linear columns, we performed log transformations so that we could fit them into our model.

Modelling:

- We kickstarted our regression analysis using a linear regression model as our baseline model and the indepedent we used, sqft_living, had the highest correlation with ther target variable.
- Using the forward-filling approach, we added more independent variables in our model and compared the RMSE and adjusted R-squared values with those of our baseline model as our choice metrics of success.
-

Model Summary:

- The model is that of bedrooms, sqft_living, floors, grade,sqft_basement,yr_renovated, age, waterfront_YES, view_AVERAGE,view_EXCELLENT, view_FAIR, view_GOOD, condition_Average,condition_Fair, condition_Good, condition_Very Good,log(sqft_lot), log(sqft_lot15) and price.
- The model is statistically significant since the F-statistic p-value is less than 0.05 and it explains 61.7%% of the total variation of price which has improved from the previous models making our model more accurate.
- Most of the predictor variables are statistically significant apart from condition fair and condition average.
- The model is off by $227171 in price which has reduced from the previous models.

- An increase of 1 square foot in the living area leads to an increase of approximately $308.73 in price.
- An increase of 1 square foot in the basement area leads to a decrease of approximately $36.72 in price.
- An increase of 1 bedroom leads to an decrease of approximately $409200 in price.
- A house graded higher by one unit leads to a decrease of approximately $19990 in price.
- An increase of 1 year in the age of the house leads to an increase of approximately $2154.77 in price.
- Renovating a house leads to an increase of your price by $63110 in price.
- A house on a waterfront compared to that not on a waterfront leads to an increase in $502500 in price.
- A house with an average view compared to that with no view leads to an increase of an $90700 in price.
- A house with an excellent view compared to that with no view leads to an increase of an $340600 in price.
- A house with a good view compared to that with no view leads to an increase of an $159900 in price.
- A house with a fair view compared to that with no view leads to an increase of an $140200 in price.
- An increase of one more floor in a house leads to an increase of $31440 in price.
- A house in an average condition compared to that in poor condition leads to an increase of an $80020 in price.
- A house in fair condition compared to that in poor condition leads to an increase of an $41180 in price.
- A house in good condition compared to that in poor condition leads to an increase of an $103100 in price.
- A house in very good condition compared to that in poor condition leads to an increase of an $138300 in price.
- For each increase of 1% in square foot lot there is decrease of $386.8 in price.
- For each increase of 1% in square foot lot15 there is decrease of $135.6 in price.

Challenges

- The model does have some limitations: 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.
- Additionally, given regional differences in housing prices, the model's applicability to data from other counties may be limited.
- Due to the presence of high multicollinearity among several predictor variables, we had to remove some of those columns from our analysis.

Conclusion

- The variables that have a major influence on the price of the house are; square foot living, age of the house,good condition of the house,if the house is on a waterfront and has an excellent view.
- The variables that has the least influence on the price of the house are; grade, number of bedrooms, sqft lot, sqft basement and sqft lot 15.
- For those looking for economical housing options, it might be wise to consider sacrificing spacious living quarters or a scenic waterfront view.

We can also see that:

- The highest number of house sales are made in the second quarter of the year (Q2: April 1 - June 30) which fall in the Spring season
- The lowest number of house sales are made in the first quarter of the year (Q1: January 1 - March 31) which fall mostly in the Winter season.

Recommendations

- Revonate their house since this increases the value of the house
- Ensure that the houses are in good condition before putting it into the market for sale
- Increase square footage of living space
- Put up their houses for sale in peak season-Spring

Future work

- Reducing noise in the data to improve the accuracy of our model.
- Additionally investigate certain features, such as constructional/architectural values of the house, to see what trends we could discern from that.

Column Names and Descriptions for King County Data Set

- id - Unique identifier for a house
- date - Date house was sold
- price - Sale price (prediction target)
- bedrooms - Number of bedrooms
- bathrooms - Number of bathrooms
- sqft_living - Square footage of living space in the home
- sqft_lot - Square footage of the lot
- floors - Number of floors (levels) in house
- 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
- 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
- condition - How good the overall condition of the house is. Related to maintenance of house. See the [King County Assessor Website](https://info.kingcounty.gov/assessor/esales/Glossary.aspx?type=r) for further explanation of each condition code
- grade - Overall grade of the house. Related to the construction and design of the house. See the [King County Assessor Website](https://info.kingcounty.gov/assessor/esales/Glossary.aspx?type=r) for further explanation of each building grade code
- sqft_above - Square footage of house apart from basement
- sqft_basement - Square footage of the basement
- yr_built - Year when house was built
- yr_renovated - Year when house was renovated
- zipcode - ZIP Code used by the United States Postal Service
- lat - Latitude coordinate
- long - Longitude coordinate
- 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