BUSINESS UNDERSTANDING

White Oak Realty is a real estate company based in Vancouver, Washington DC that would like to expand its business operations by venturing into other cities in Washington. As a junior data scientist at the company, I have been tasked with analyzing house sales data in King County and building a model that would predict sale prices.

The product design team at the company is in charge of coming up with different ways to package properties suitable for the various classes of customers. For this expansion project their plan is to start by purchasing existing houses in the area and remodeling them for re-sale as well as developing new houses and properties.

In order to achieve this, they need to know;

a) The classes of properties in King County.
b) The common and the unique house features.
c) What features affect house prices most?
d) Other underlying factors to focus on that might increase sale prices.

The sales & marketing team will use the classes of houses to come up with segments of the potential customers and design marketing plans that best highlight the features of each property and appeal to the right segment which is key in making successful sales.

The aim of this project is to help the firm make data driven decisions on their plans to expand their business operations into other parts of Washington. This project will provide insights to the sales &product design teams to help them better understand the real estate market in King County. This project will also build pricing models that will help to best price houses based on their features which will maximize sale revenues and profits.

DATA PREPARATION

This project will use house data from King County, Washington that contains data about the features of the houses located in various parts of the county. The dataset contains more than 20,000 observations of houses including their features such as the size, number of bedrooms and bathrooms and other features which will be used to draw useful insights.

I performed initial cleaning of the dataset by following the following steps:

1. Investigating the datatypes of each column to ensure all numerical columns were stored as integers or floats and converting date columns to datetime objects.
2. Checking for missing / null values - I found that 3 columns ('waterfront', 'view' and 'yr_renovated' had null values). Using the column description file I found that the waterfront column represented binary variables and filled the null values with 0's appropriately. I also chose to fill the null values in the view column with 0's given that the values ranging from 0 to 5 represented the number of times a house was shown. 0 was the most frequent value and therefore replacing the null values with 0 would not distort the mean. Finally, I chose to drop the yr_renovated column as it would not be essential for my analysis.
3. Checking for outliers - I used scatter plots to look for outliers in the various variables and handled them appropriately.

Checking for multicollinearity

I made use of a heatmap and correlation table to investigate multicollinearity between the different features in the dataset. Both the heatmap and correlation table showed a high collinearity between *sqft_above* and *sqft_living*. To minimize multicollinearity, I dropped *sqft_above*. The column contains information on the living space in the house excluding the basement. This should not negatively affect the model because *'sqft_living'* also contains information about the size of the house excluding the basement and garage areas.

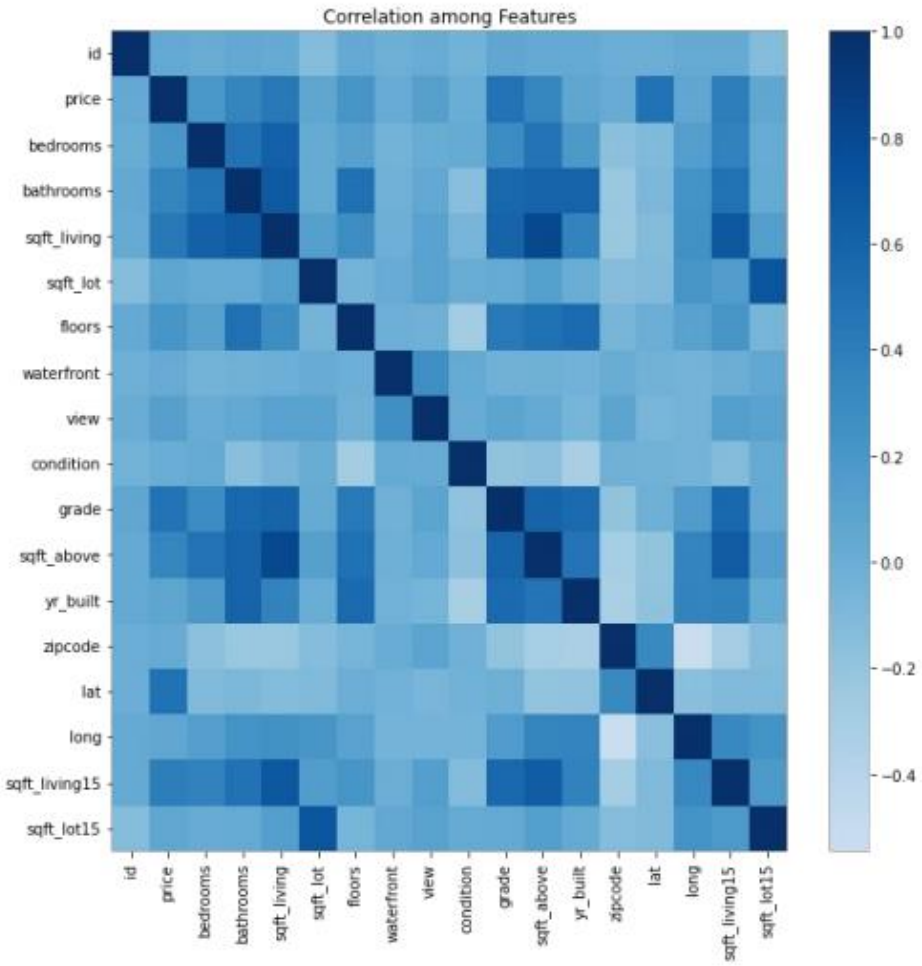

Fig 1: Multicollinearity among the features in the dataset

Modelling

I proceeded to build my first model which would serve as a simple baseline model with little manipulation done to the dataset. I plotted a second heatmap to visualize correlation of each feature against the price and used the features with the highest correlation as predictors for this model. Features used were 'grade', 'Sqft_living', 'sqft_living15' and 'bathrooms'.

For the second model I first performed some modifications to the data to better prepare it before fitting the model. I followed the following steps;

1. Performing log transformations on the numerical variables.

2. Scaling the numerical variables to normalize their distributions.

3. One - hot encoding categorical variables. The second model has performed significantly better than the baseline model.

Evaluation

The baseline model's performance was not very good with an r-squared of 0.27, explaining 27% of the variance. The model also had an RMSE score mean of 108055 and std of 1498. The mean error can be interpreted to mean that the prices predicted using this model with be less or more by about $ 108,055.74 with a standard deviation of 1498.71.
The second model had a r-squared score indicating that the model explains 97.5 % of the variance. Even with a good score, the model used many features and poses a risk of overfitting.
To cater for the risk of overfitting in the second model, this final model will further employ feature selection to reduce the number of predictors used to fit the model and hence build a model that will perform just as accurately on the test set as it did use the training data. I performed feature selection to select the best features for the model by considering the p - value of each feature. I ran a series of multiple models here with each round eliminating the feature with the highest p value that represented the lowest significance to the model and eventually ended up with the best features.

Findings and Recommendations

Based on my findings, I made the following conclusions and subsequent recommendations based on the findings; 1. Bedrooms are bathrooms affect house sale prices - the findings show that houses with at least 4 bedrooms and two bathrooms sell at significantly higher prices than those with less. My recommendation would be to focus on acquiring houses with 4 or more bedrooms and at least 2 bathrooms. 2. The King County house grading system is key to a house's selling price. There is a clear linear relationship between the grade and price and as the grade went up so did the selling price. Although exact determinants of this grade are not clear in the study, it is very significant. My recommendation would be to look into the grading system and to choose houses with a grade of at least 7. 3. Waterfront feature -

Houses without a waterfront sold at lower prices than those with one. My recommendation would be to acquire houses with a close proximity to a waterfront in order to maximize on the demand for this feature. 4. The overall living size area of the house i.e., excluding the basement is very significant to the price. The study shows a linear relationship between the size of the living space and price while the size of the basement played a very insignificant role to the houses' selling price. My recommendation would therefore be to acquire houses with a relatively larger living space as compared to the basement