KINGS COUNTY HOUSING ANALYSIS

BY JAMLECK MATHENGE N.

BUSINESS UNDERSTANDING:

During this project, I assumed the role of a data scientist whereby I used our dataset("kc_house_data.csv") to generate various business problems that will be the founding blocks of our linear regression models.

- **STAKEHOLDER**

I assumed the role of a data scientist at Vision Real Estate Agency which deals with the valuation of houses. I'm tasked at coming up with models that can predict house prices based on certain features in our dataset.

- **Business problems:**

Model 1 :

Provide insight to Vision Real Estate Agency on how the footage of a home and can cause changes in house prices and by how much.

Model 2 :

Provide insight to Vision Real Estate Agency on how the footage of a home and the footage of the nearest 15 neighbors can cause changes in house prices and by how much.

Model 3 :

Provide insight to Vision Real Estate Agency on how the footage of a home, the footage of the nearest 15 neighbors, footage of the lot, footage of the nearest 15 neighbors, number of floors, number of bathrooms, footage of basement and footage of above can cause changes in house prices and by how much.

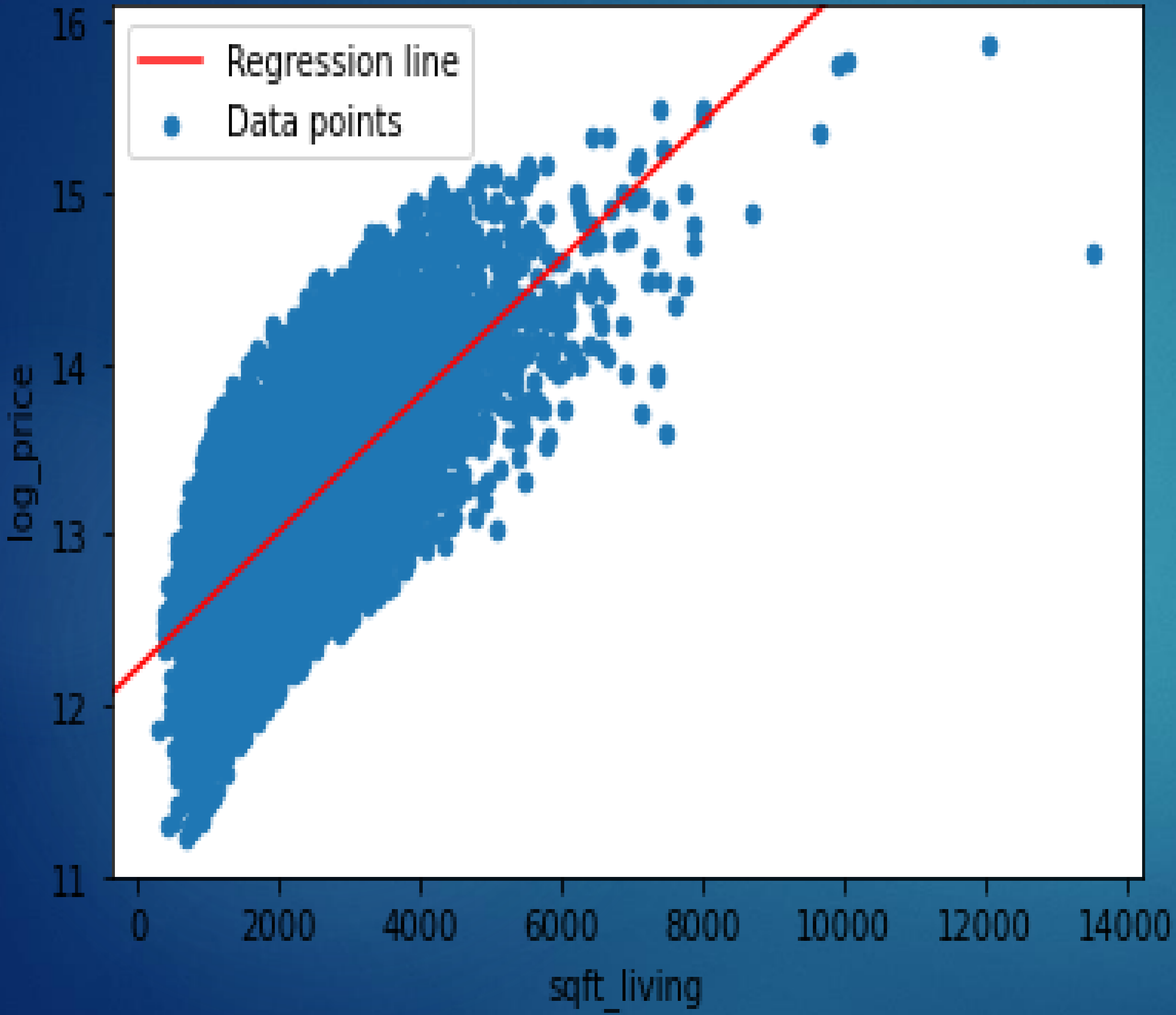
Model 4 :

Provide insight to Vision Real Estate Agency on how the footage of a home and condition of the home can cause changes in house prices and by how much.

DATA UNDERSTANDING.

- Our dataset("kc_house_data.csv") contains Kings County housing data. The data set consists of 21 columns and 21612 rows.
- The King County Housing Data Set contains information about the size, location, condition, and other features of houses in Kings County
- Our data type is mainly made up of numerical and categorical variables.

MODEL 1



- **Simple Linear Regression Results**

Looking at the summary above, we can see that the regression line we found was

log_price = (12.2185) + 0.0004sqft_living $$

* The model is statistically significant overall, with an F-statistic p-value well below our alpha of 0.05.

* The model explains about 48.5% of the variance in price.

* The model coefficients (`const` and `sqft_living`) are both statistically significant, with p-values well below 0.05 . The y-intercept is at (12.2185).

* If a house added at least 1sqft living room space, we would expect it to have about (0.0004 * 100)% increase increase in price.

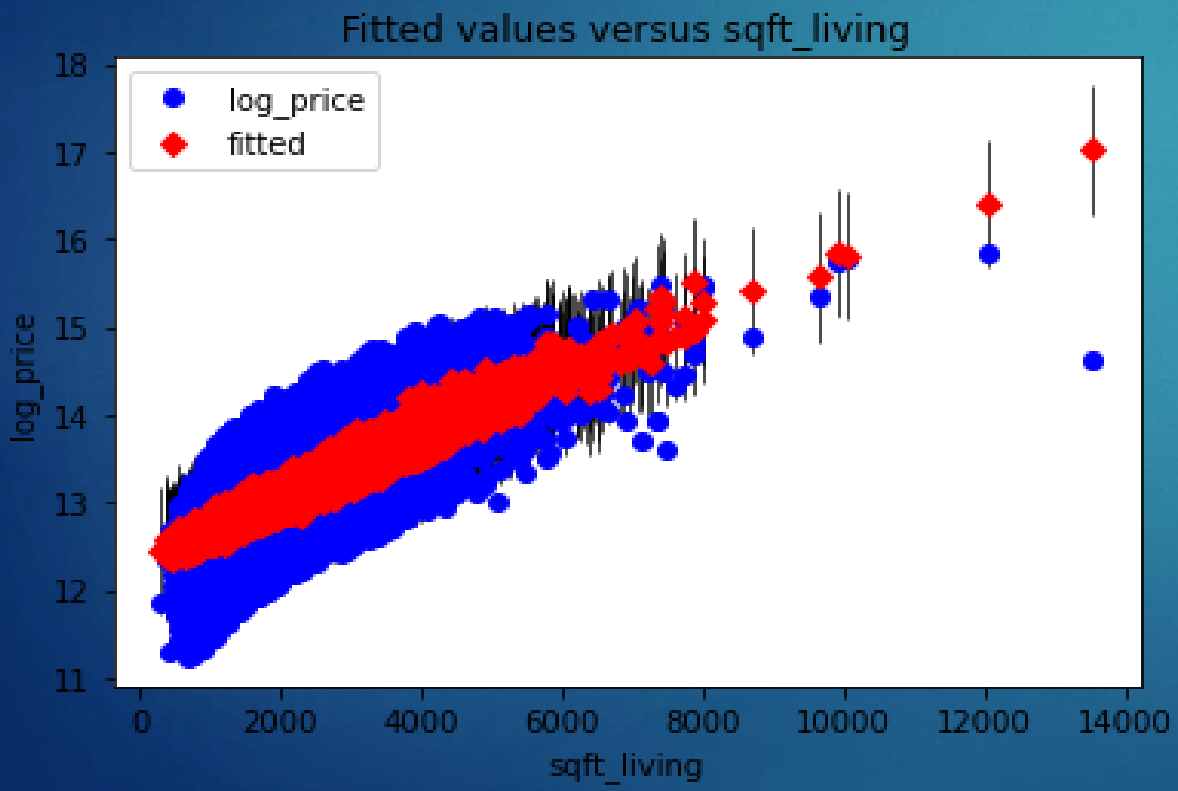
MODEL 2

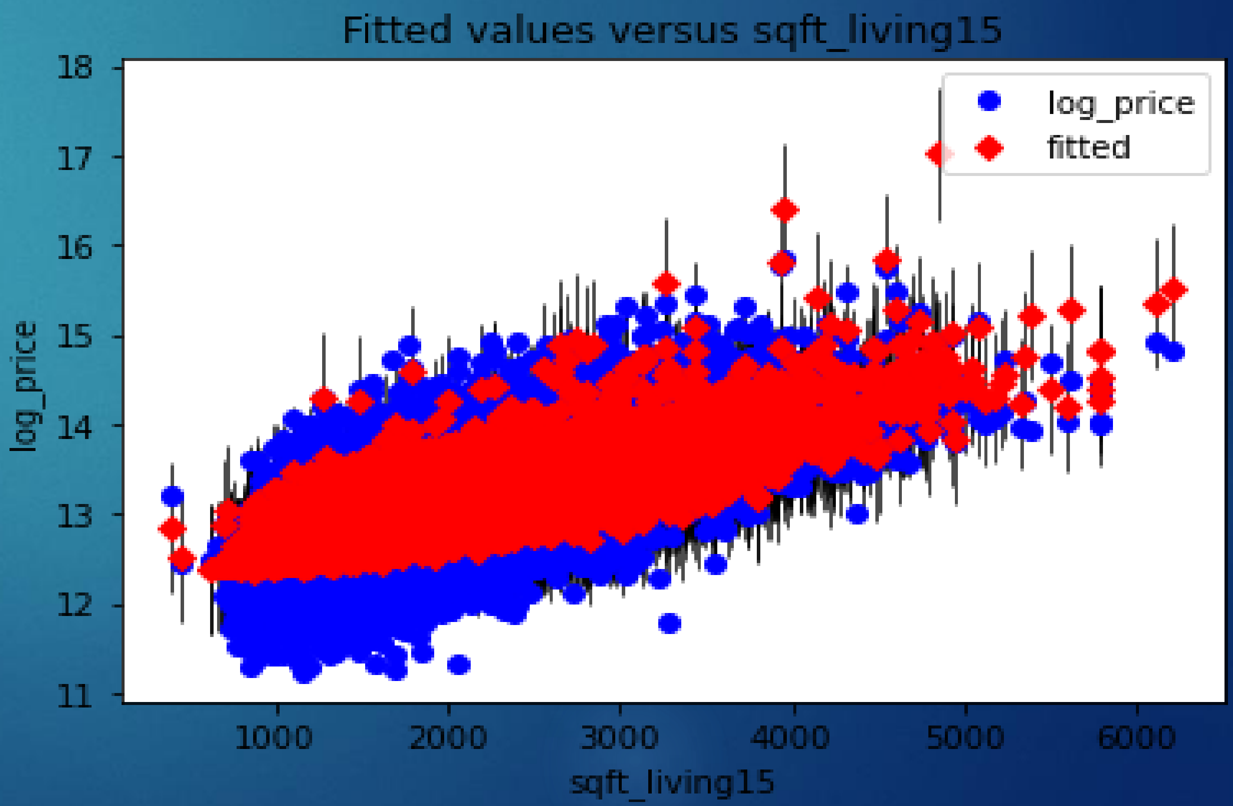
Model with Two Features Results

This time, the model we built was:

log_price = (12.0815) + 0.0002sqft_living15 + 0.0003 sqft_living

* The model is statistically significant overall, with an F-statistic p-value well below 0.05
* The model explains about 50.5% of the variance in price which is an increase from our linear regression.
* The model coefficients(`const`, `sqft_living15`, and `sqft_living`)are all statistically significant, with t-statistic p-values   well below 0.05
* For each increase of 1 in sqft_living15, we see an associated increase in price of about 0.02%
* For each increase of 1 in the sqft_living, we see an associated increase in price of about 0.03%





MODEL 3

- **Model with Multiple Features Results**

This time, the model we built was:

log_price = (12.0553) + (-1.09e-06) sqft_lot15 + 0.0002 sqft_living + (3.129e-07)sqft_lot + 0.1089floors + 0.0002sqft_living15 0.0165bathrooms + (-0.0454)bedrooms + (4.059e-05)sqft_above + 0.0002sqft_basement

* The model is statistically significant overall, with an F-statistic p-value well below 0.05

* The model explains about 51.2% of the variance in price which is an increase from our linear regression.

* The model coefficients(`const`, `sqft_living15`, `sqft_living`,`sqft_lot15`,,`floors`,`bathrooms` ,`sqft_basement`,`above` and `bedrooms`)are all statistically significant, with t-statistic p-values below 0.05

* A change in any of the above features causes a (coeeficient of the feature * 100)% change to price of a house.

MODEL 4

- **Multiple linear Regression Results with categorical variables.**

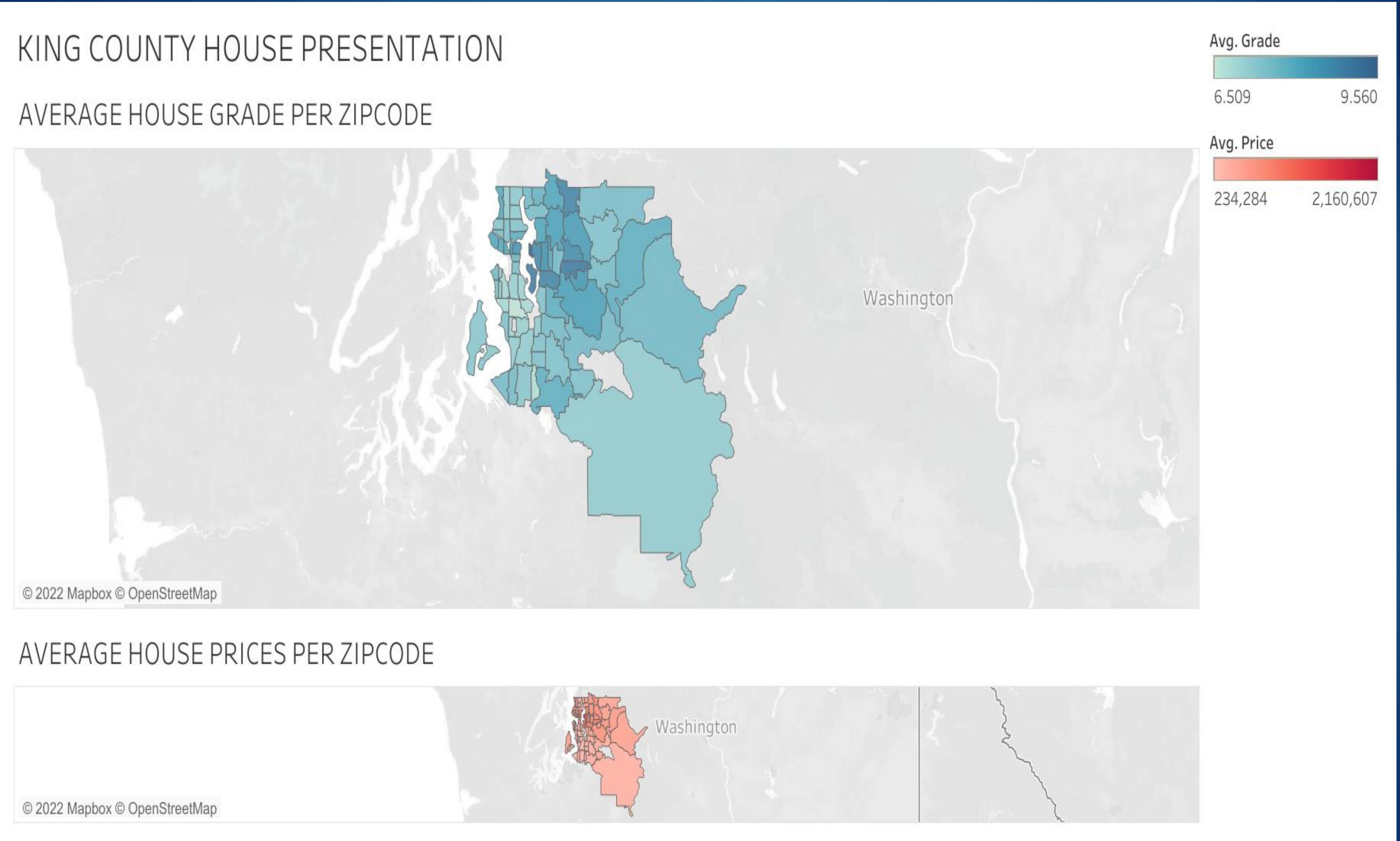
* Most features in the model are statistically significant overall, with an F-statistic p-value well below our alpha of 0.05.

* The model explains about 49.3% of the variance in price.

* The model coefficients (`const`,`sqft_living`,`condition_3`,`condition_4` and `condition_5`) are both statistically significant, with p-values well below 0.05 . The y-intercept is at (11.9950).

* A change in any of the above features causes a (coeeficient of the feature * 100)% change to price of a house.

TABLEAU DASHBOARD



KING COUNTY HOUSE PRESENTATION

AVERAGE HOUSE GRADE PER ZIPCODE

AVERAGE HOUSE PRICES PER ZIPCODE

RECOMMENDATIONS

- From the model results and tableau dashboard visualization we can come up with the following recommendations.

From Model 1.

The square footage of a home has an effect on the price of a house whereby price will increase with increase in footage as per the equation given.

From Model 2.

Areas where houses are similar are likely to have a higher prices for houses. This applies in the real world today where people prefer living in gated controlled development estates.

From Model 3.

The numeric variables of Kings County housing have an effect on the price of a house.

From Model 4.

The condition of a house has an effect on the price of the house. House prices increase as the condition level increases.

From Tableau Dashboard.

The location of a house has an effect on its price. The company should consider a house’s location in calculating its price.

CONCLUSION

- This presentation was made to create recommendations for Vision Real Estate Agency on the valuation of houses in Kings County.
- THANK YOU