KING COUNTY HOUSE PRICES

PREDICTION MODEL

OVERVIEW

This project was completed using a dataset acquired from Kaggle. The data for the dataset was provided by King County, Washington. The data includes sold homes between may 2014 and may 2015.

The goal for this project is to come with a suitable model for a real estate company in Washington that will be used to predict house prices which will enable them maximize on the profits.

BUSINESS UNDERSTANDING

- A real estate agent of a company in Seattle wants to know which factors significantly impact the prices of a house in King County. This will aid in strategizing on the best criteria to take in order to maximize on profit.

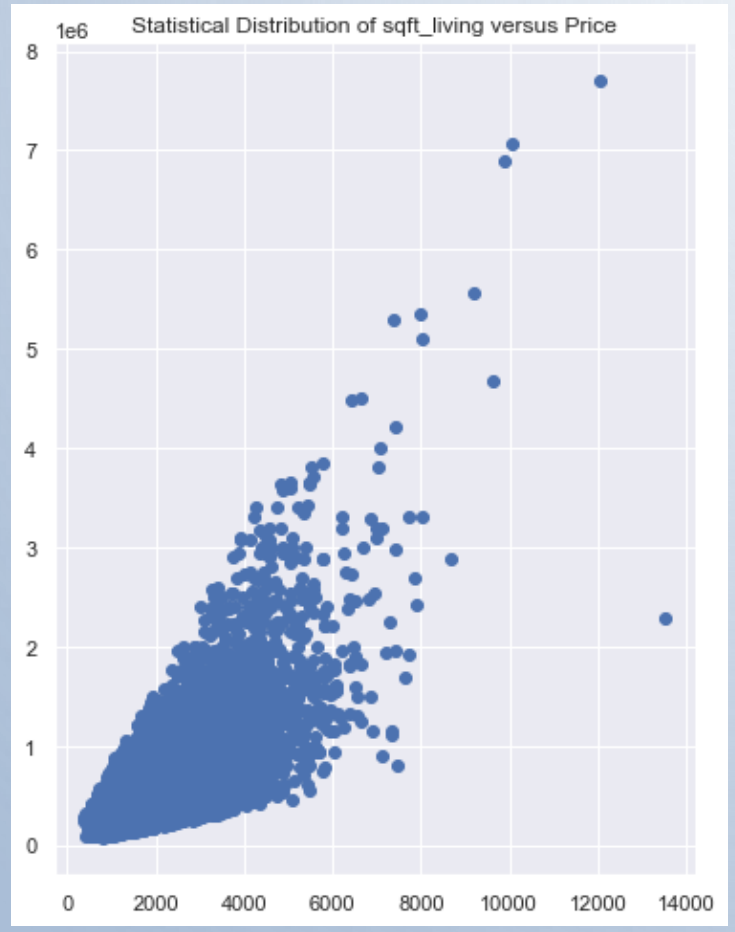
METHODOLOGY AND TECHNIQUE

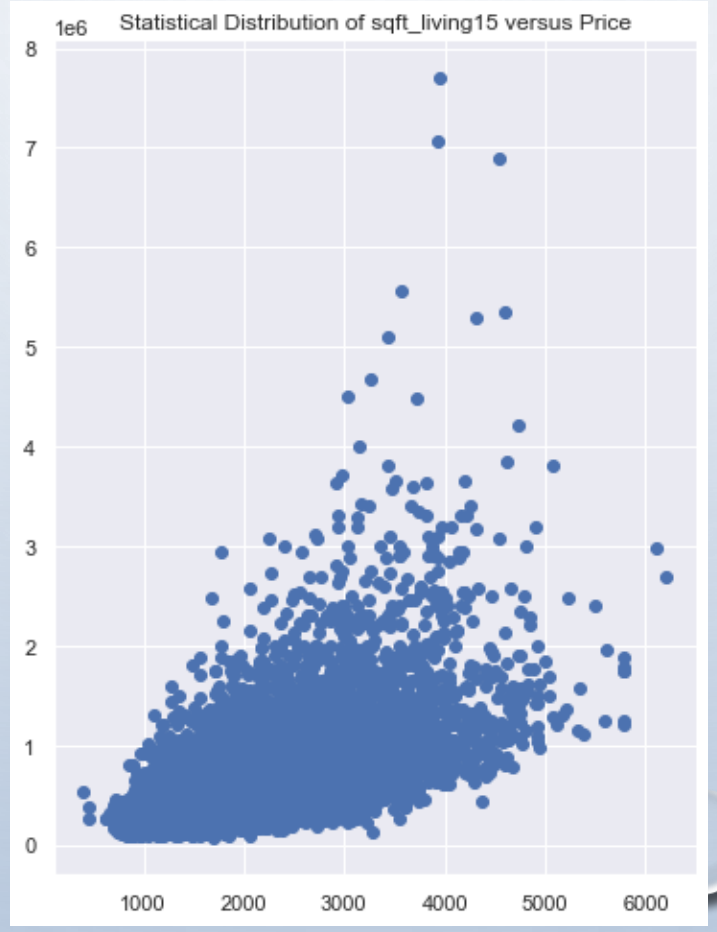
The project involved an iterative approach of visualizing relationships features against price, building a multiple linear regression model with Python and StatsModels to predict sale prices for houses and identifying features that affect prices of houses in King County.

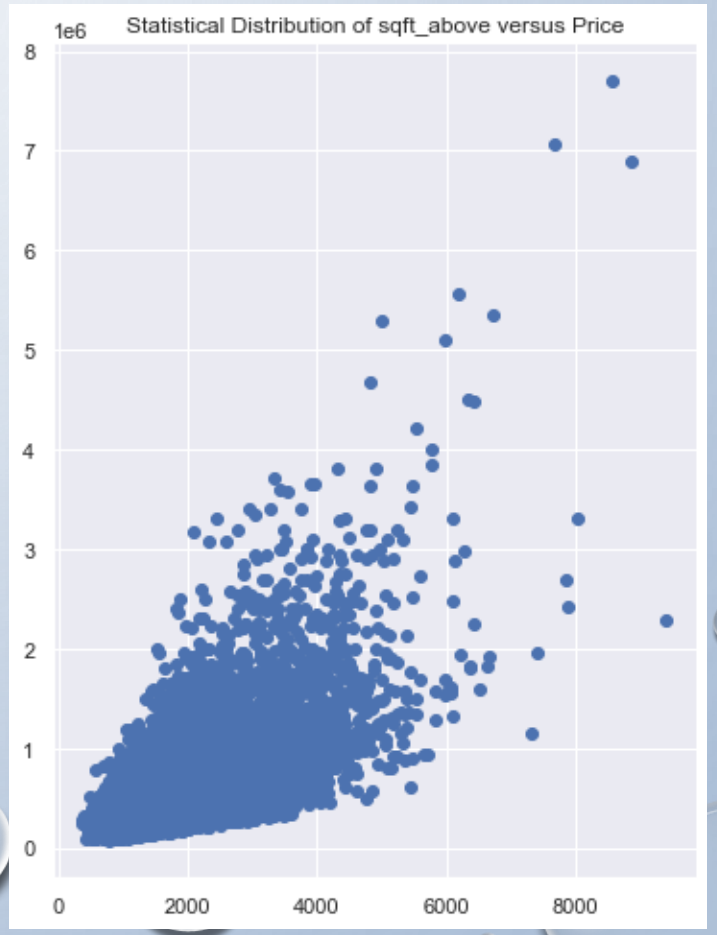
DATA UNDERSTANDING

- King County data contains 21597 records with 20 features.
- The data contains some null values in waterfront, view and year renovated.
- Categorical data include waterfront, condition, grade and view.
- This data set is on house sales prices for king county, covering homes sold between may 2014 to may 2015 most houses have 1 floor
- The average home price is about 540,000 dollars with the highest being sold for nearly 800,000 dollars.
- The data contains houses built between 1900 and 2015.
- The average grade of the houses is 7, meaning most houses sold are above average grade.
- The maximum number of floors of houses you can find in king county is 3.5.

BIVARIATE ANALYSIS(CORRELATION BETWEEN DEPENDENT AND INDEPENDENT VARIABLES)

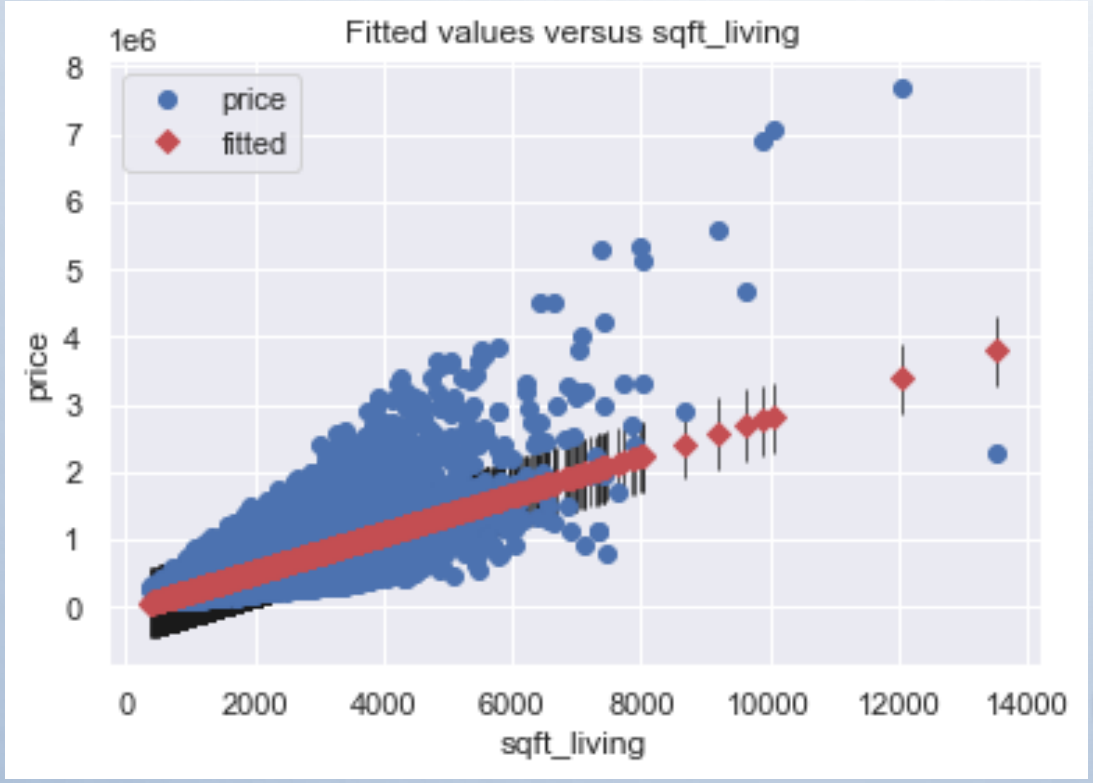




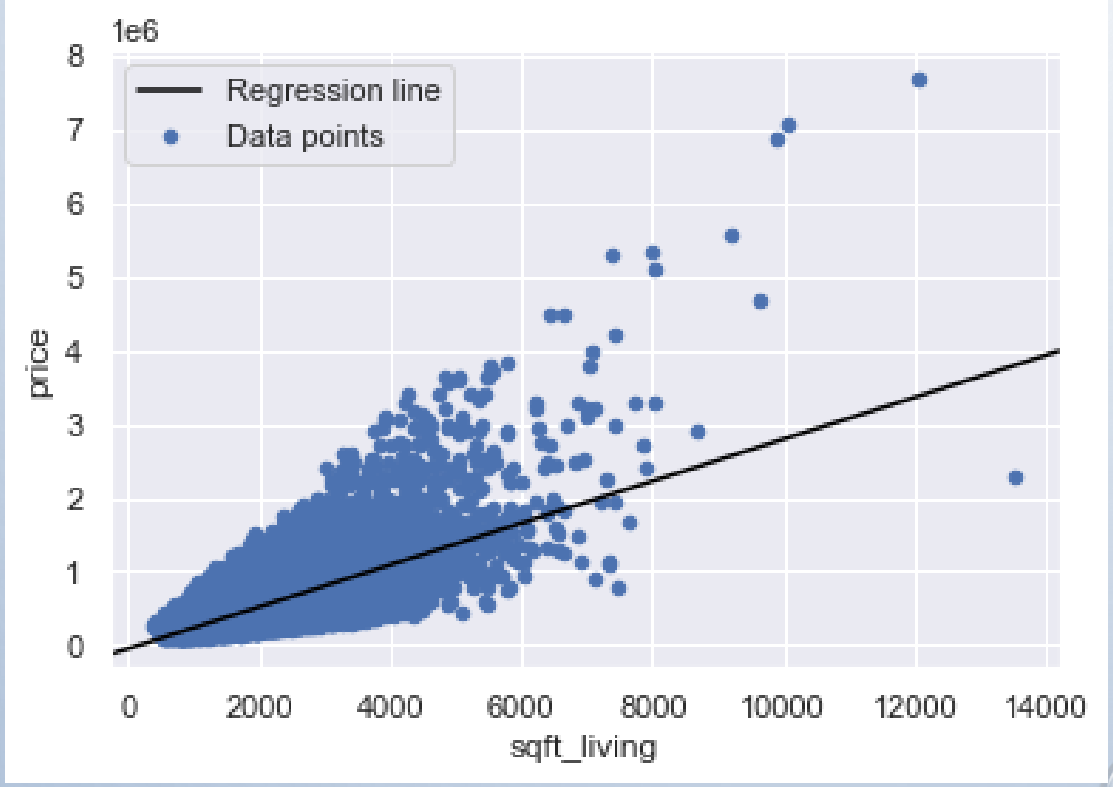


MODELING

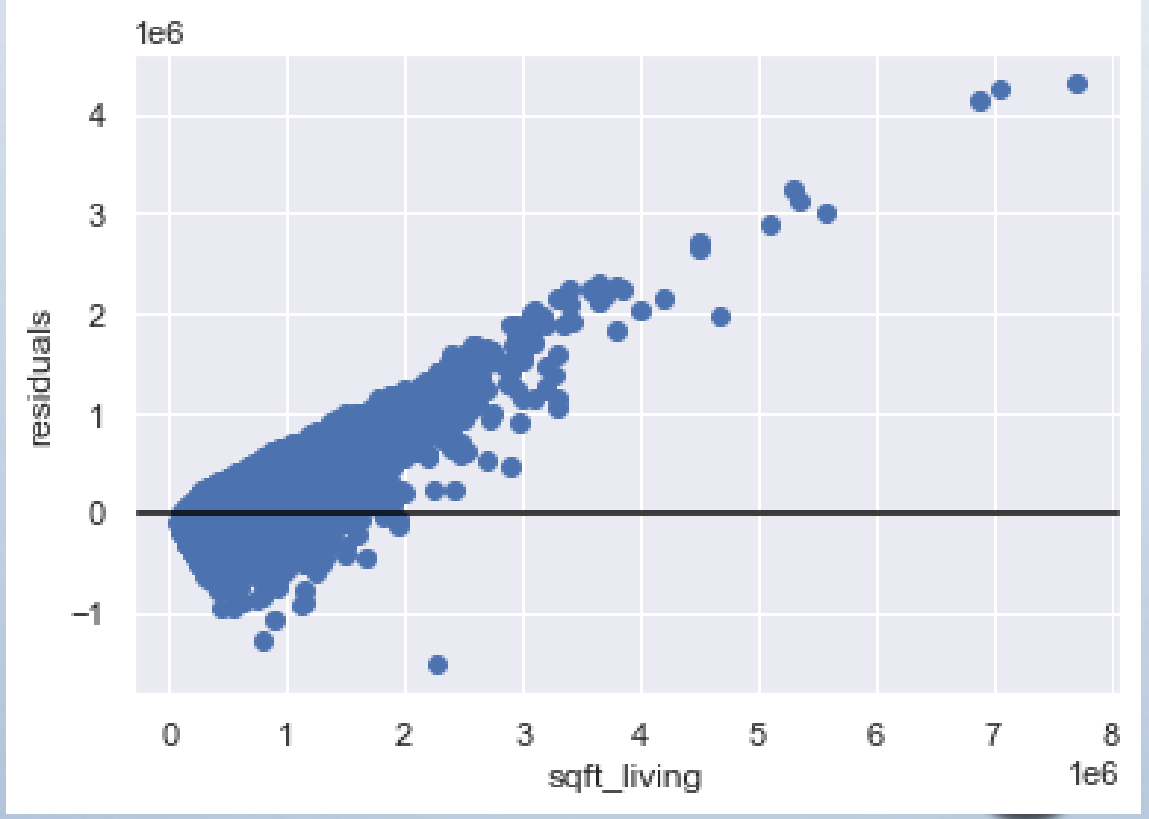
- A) LINEAR REGRESSION
- PRICE = -5.045E+04 + 284.3280SQFT_LIVING



REGRESSION LINE



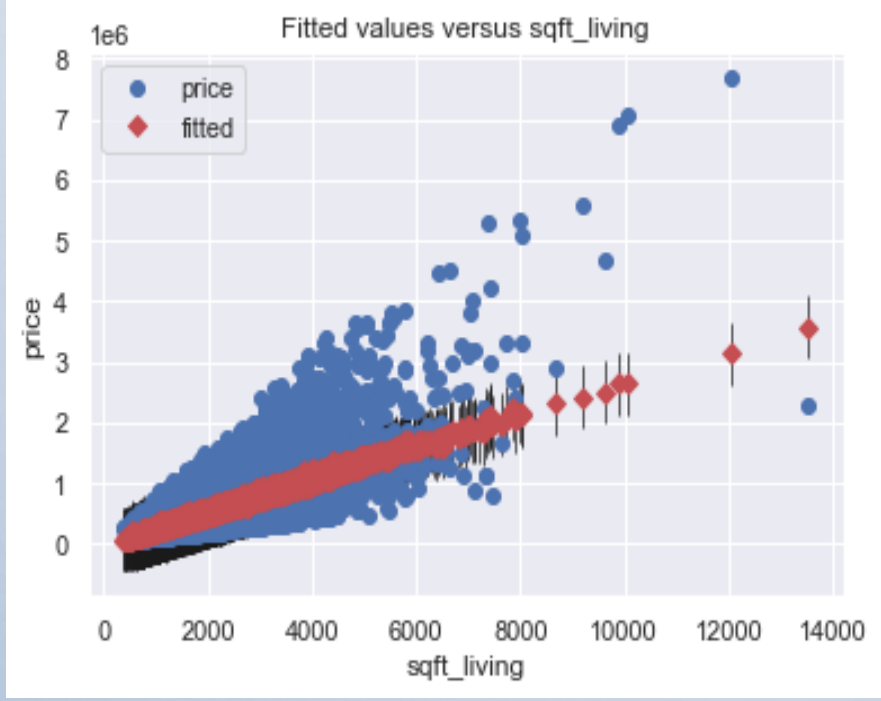
RESIDUAL



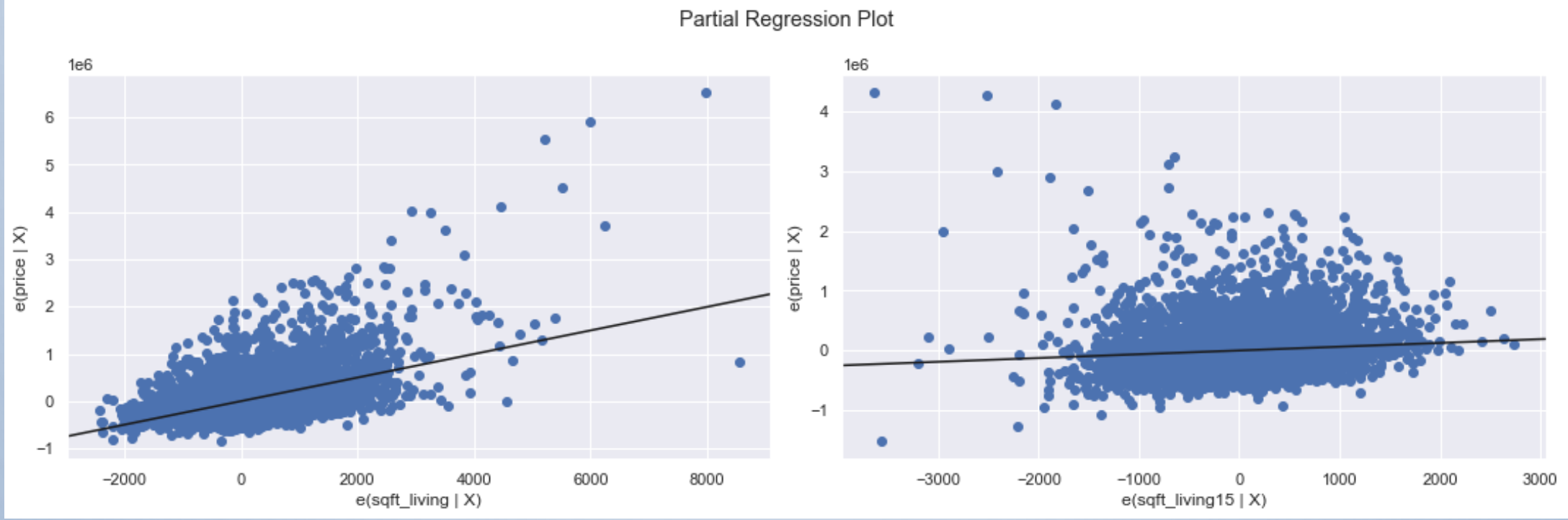
- 49.7% variation in price can be explained by the square footage of living space in the home
- An rsquared of 49.7% is a low variation and indicates the square footage of living space in the home is not explaining much variation of the price.
- We solve this by doing a multiple linear regression, adding more variables to our model to increase variability.

B) MULTIPLE LINEAR REGRESSION

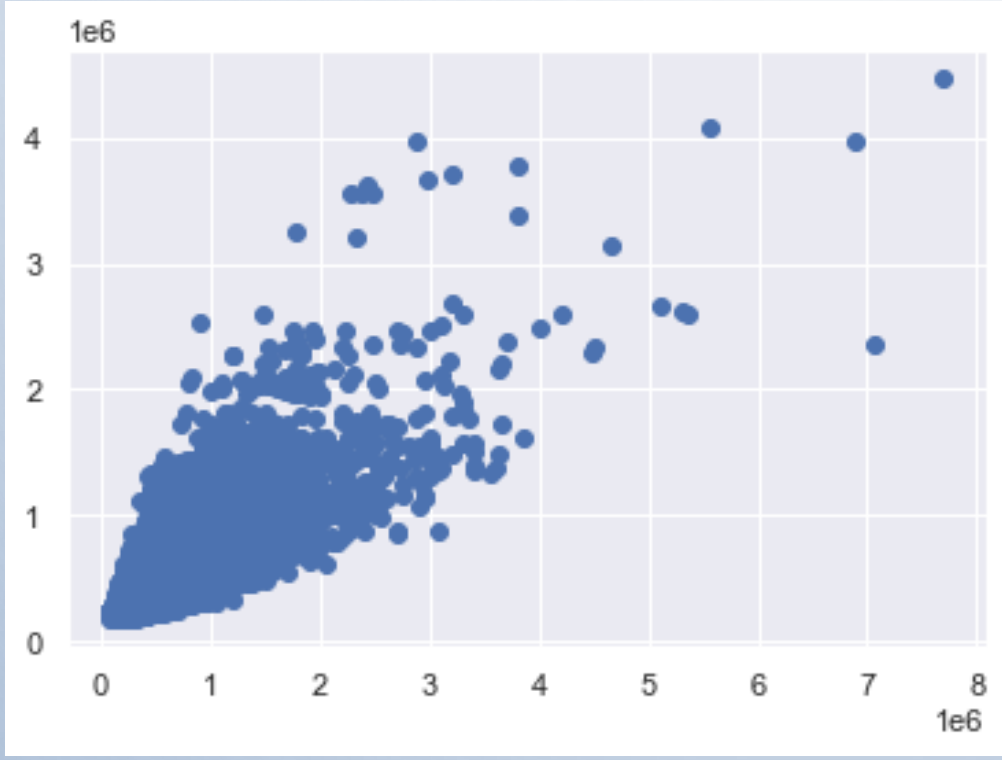
The model built is: price = -1.021e+05 + 249.0055sqft_living + 62.9802sqft_living15



PARTIAL REGRESSION PLOT



ACTUAL PRICE VERSUS PREDICTED PRICE



MODEL RESULTS

- The model explains 50.2 % variation in price.
- The model coefficients (const, sqft_living, and sqft_living15) are all statistically significant, with t-statistic p-values well below 0.05
- For each increase of squarefoot, we see an associated increase in price of about 249.0055.
- This is a little bit smaller of a increase than we saw with the simple model, but not a big change.
- For each increase of 1 in sqft_living15, we see an associated increase in price of about 62.9802.
- All of the p-values round to 0.