Phase 2 Project

1. Anthony Thuita
2. Vivian Mosomi
3. Stephen Ndaro
4. Andrew Mutuku
5. Sammy Toroitich

HOUSE SALES ANALYSIS

Business Problem

- Providing advice to homeowners about how home renovations might increase the estimated value of their homes and by what amount.

Business Understanding

- By analysing house sales, we aim to provide factors that influence housing prices. This knowledge can help buyers make better decisions and also assist sellers in setting competitive prices for the real estate properties.

- Analysis of this housing data set can help in predicting future price changes and identifying areas with appreciating property values

Steps we followed

- Data Analysis
- Data Cleaning
- Modelling
- Getting Regression Results.
- Validation of the Model.

The Data Used:

Id - Unique Identifier for a house

Price - Sale Price(Prediction Target)

Bedrooms /Bathrooms - Number of Bedrooms and bathrooms

Sqft_living / Sqft_lot - Square footage of living space and lot in the home

Floors - Number of floors(levels) in house

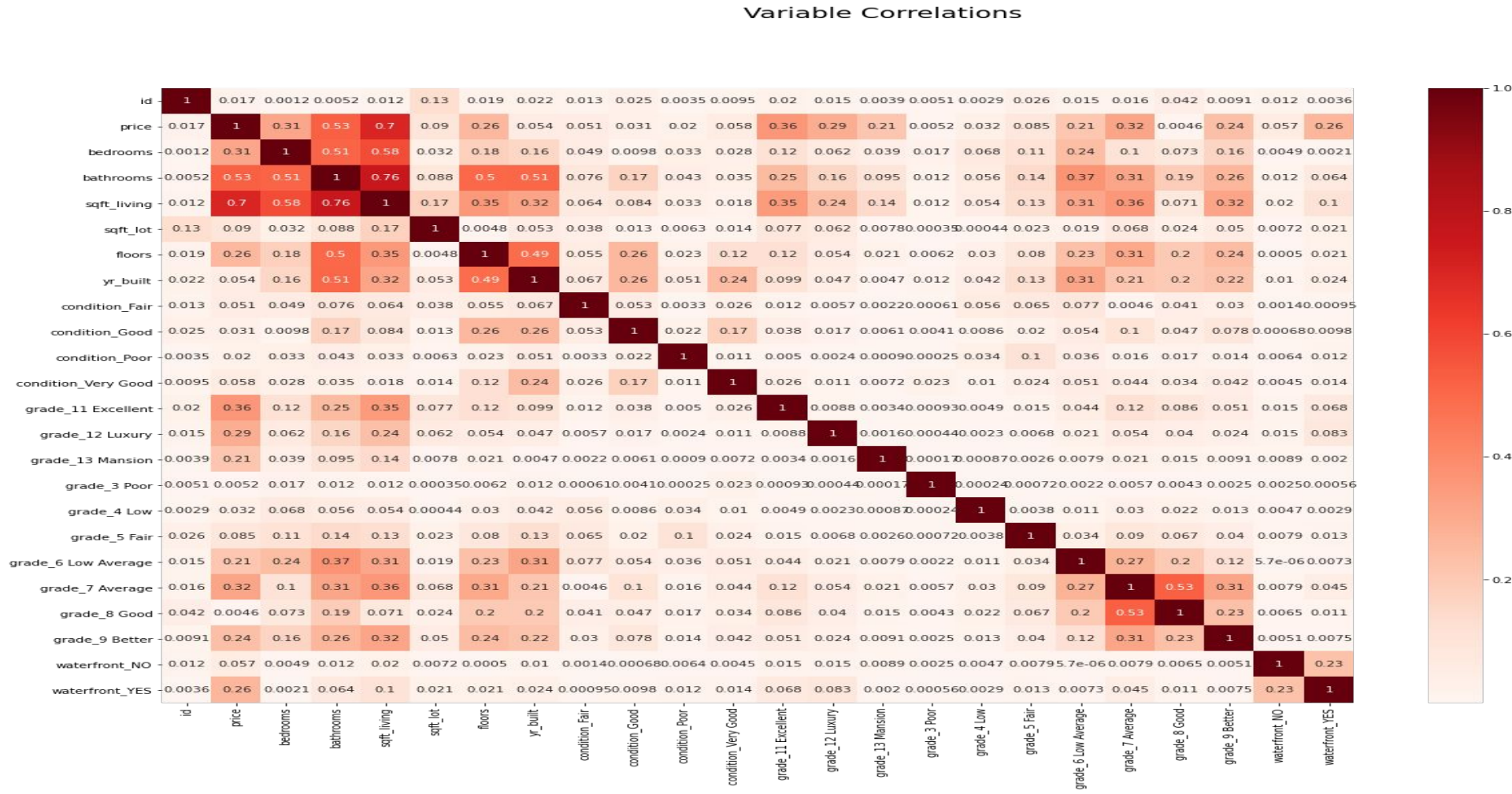
Waterfront - Whether the house is on a waterfront

Condition - How good the overall condition of the house is

Grade - The overall grade of the house

Yr_built - Year when the house was built

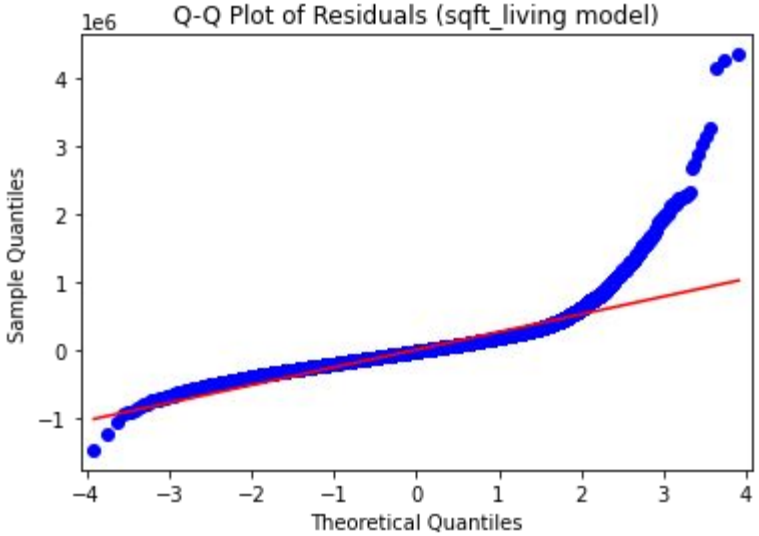
Correlation of different columns using a heat map



The following is after data preparation(data cleaning and one hot encoding)

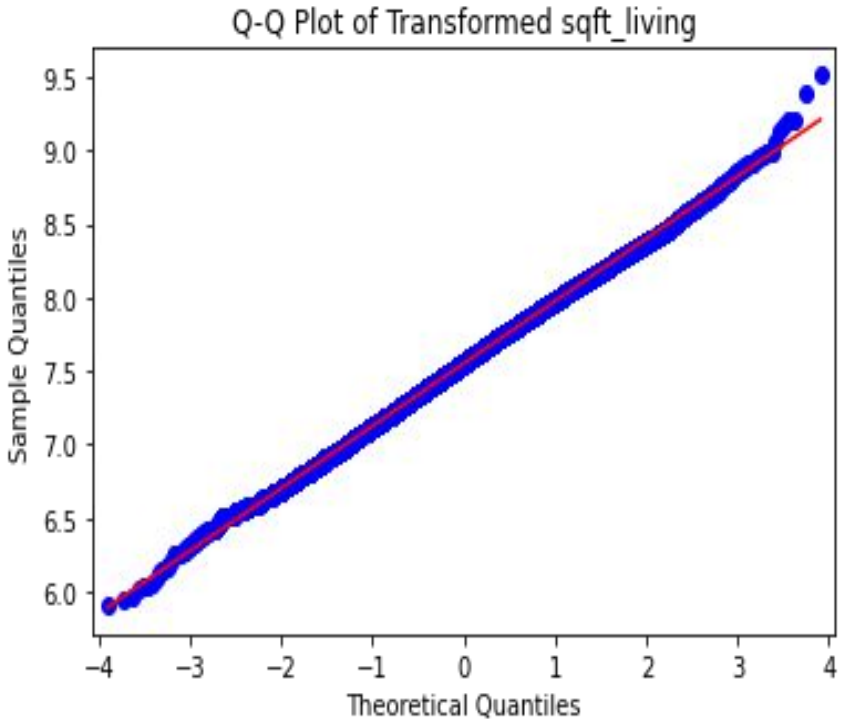
Simple Linear Regression

i) Price vs sqft_living(Most Correlated)



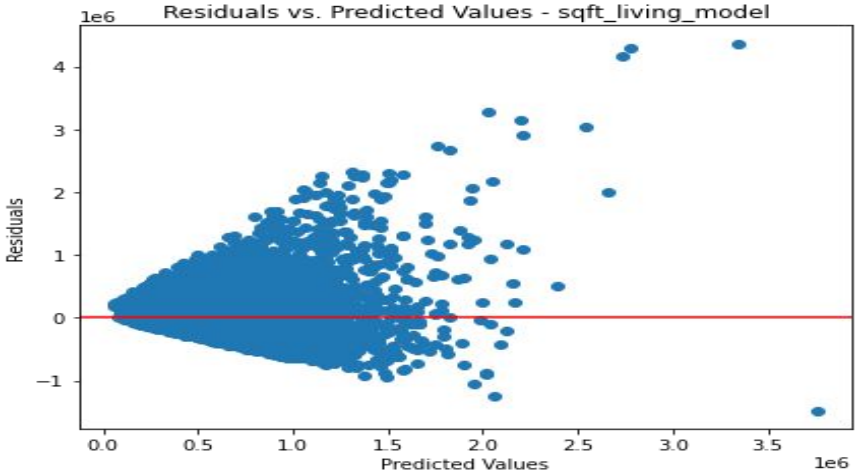
- Our model for this explained 49.3% of the variance in prices
- But the residuals displayed homoscedasticity

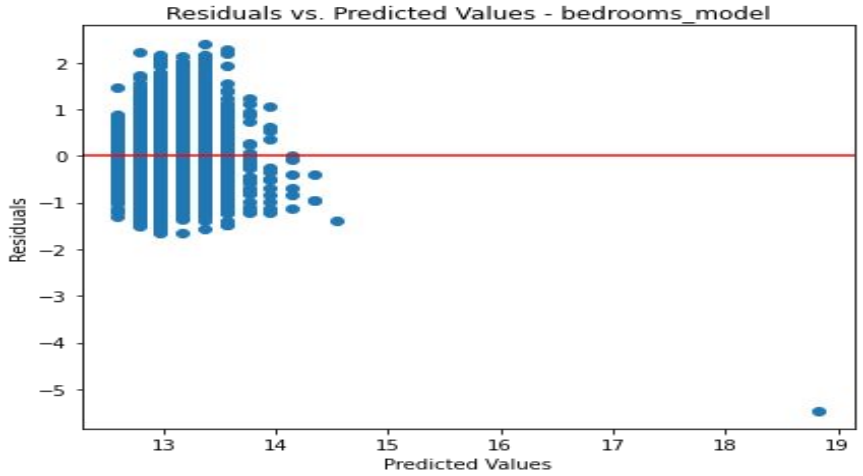
Correcting Homoscedasticity

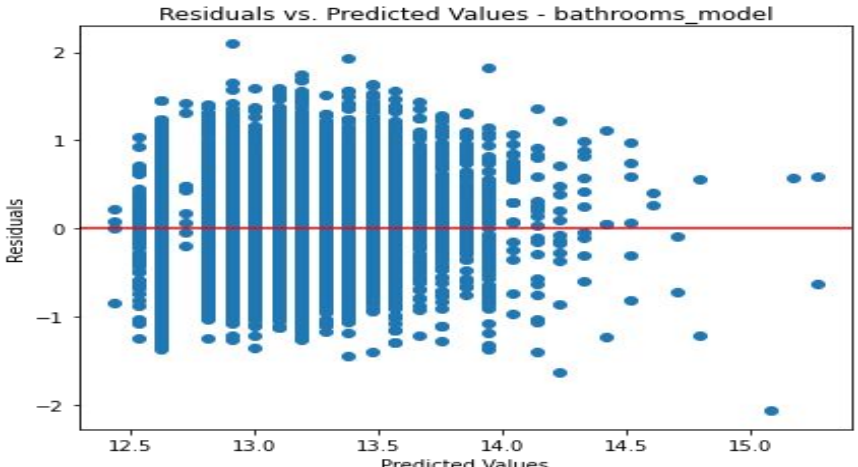


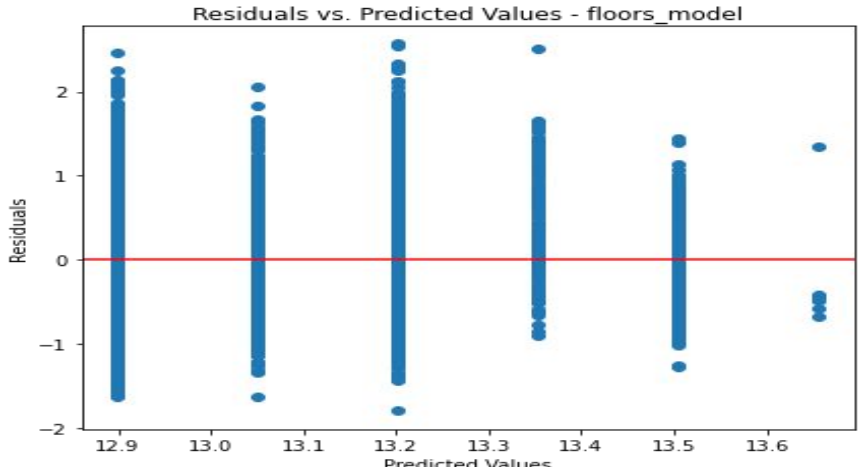
We created other simple linear regression models for bedrooms, bathrooms and floors. These are the residual plots:

Residual plots for 4 most correlated columns with price



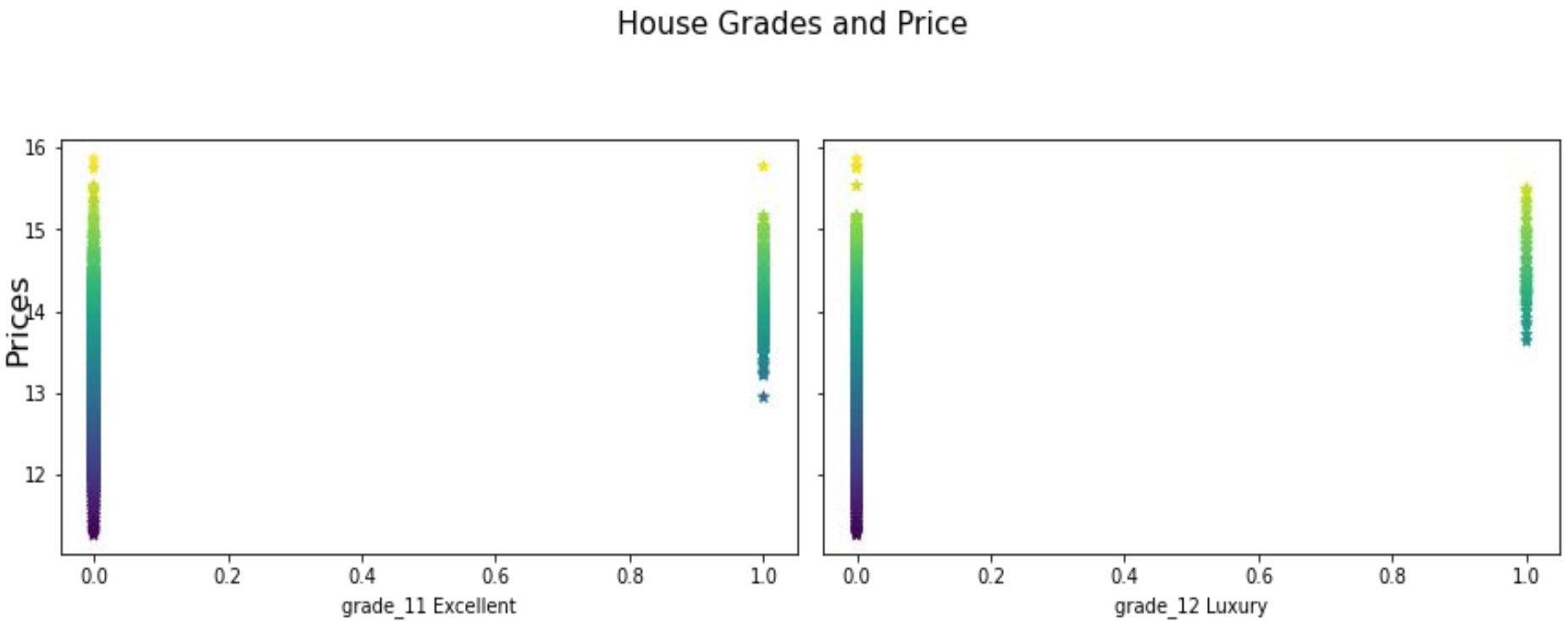




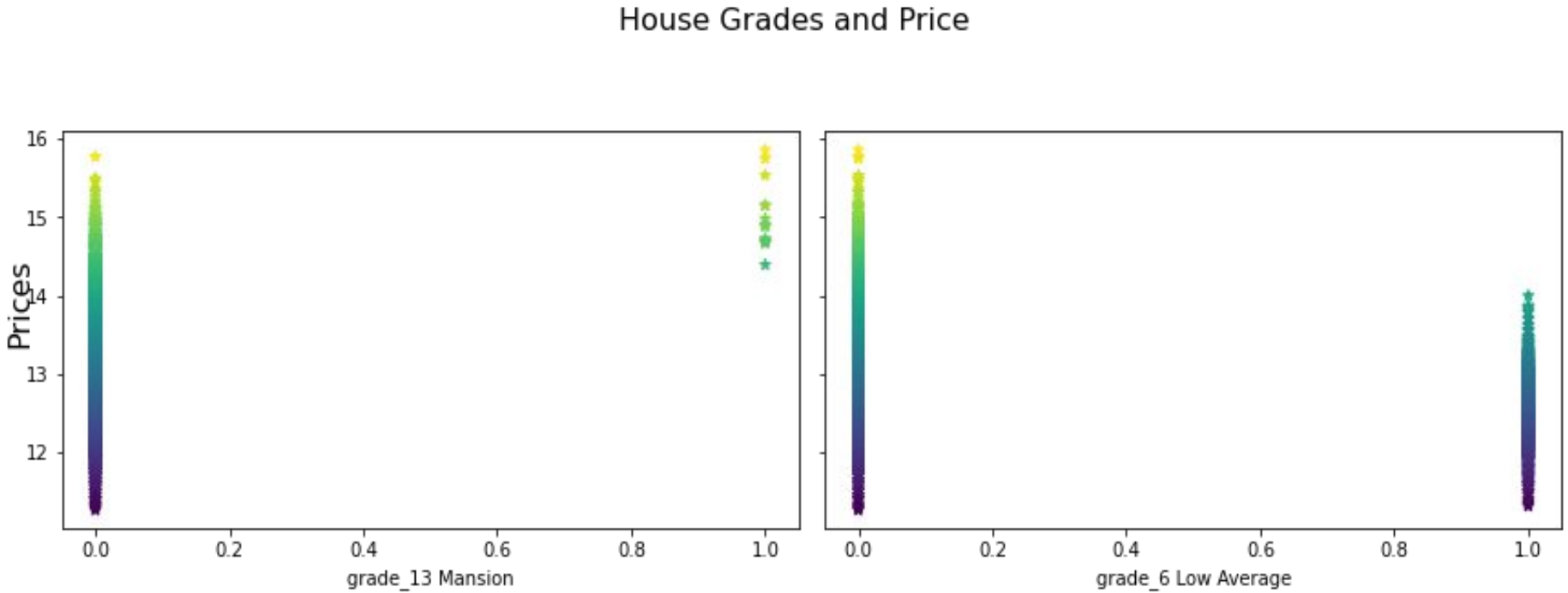


From the plots above, our residuals are normally distributed passing the linear regression assumption of normality of residuals.

Finding the categorical column with the most linear relationship with price



Finding the categorical column with the most linear relationship with price



Multiple Linear Regression Model

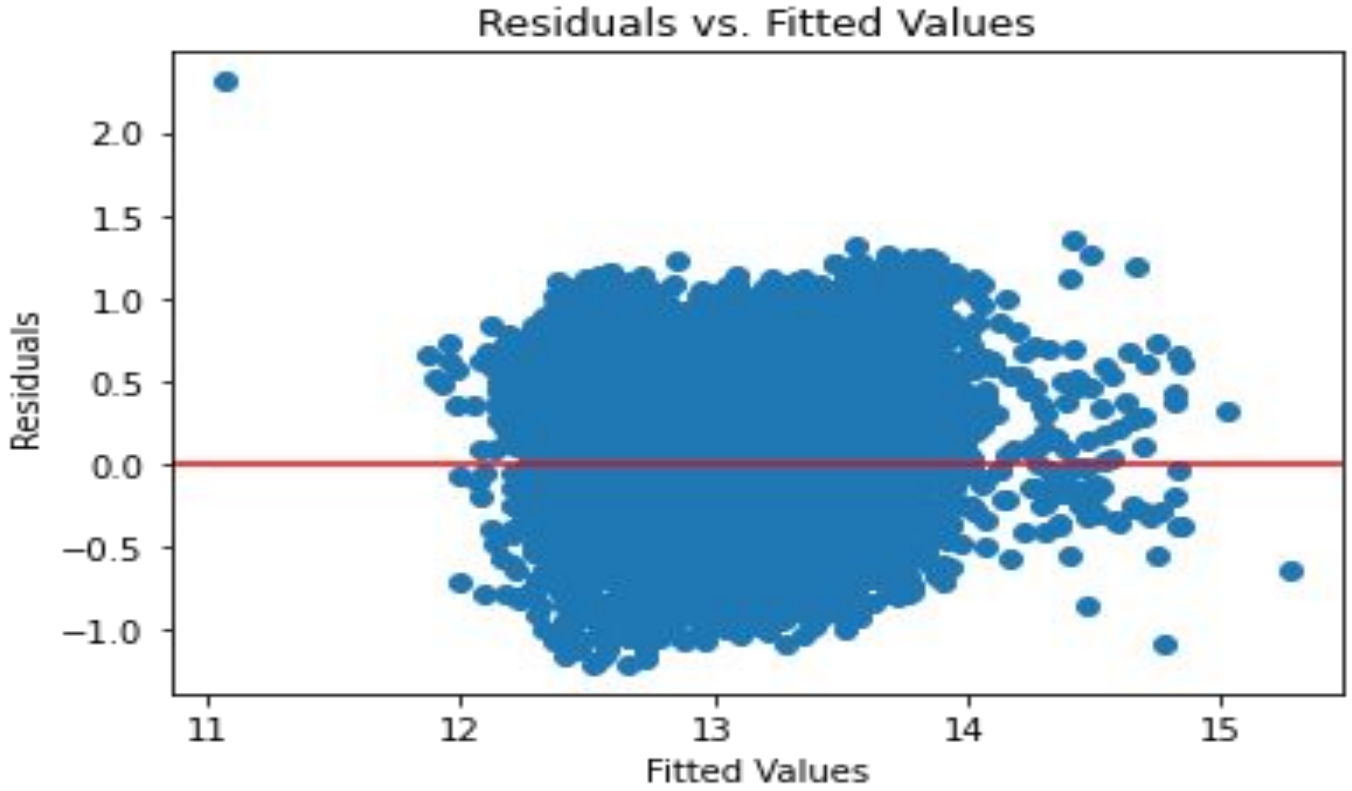
Model 1

- Multiple Linear Regression with **sqft_living, bathrooms, bedrooms and floors** gives and r-squared value of **47.1%**

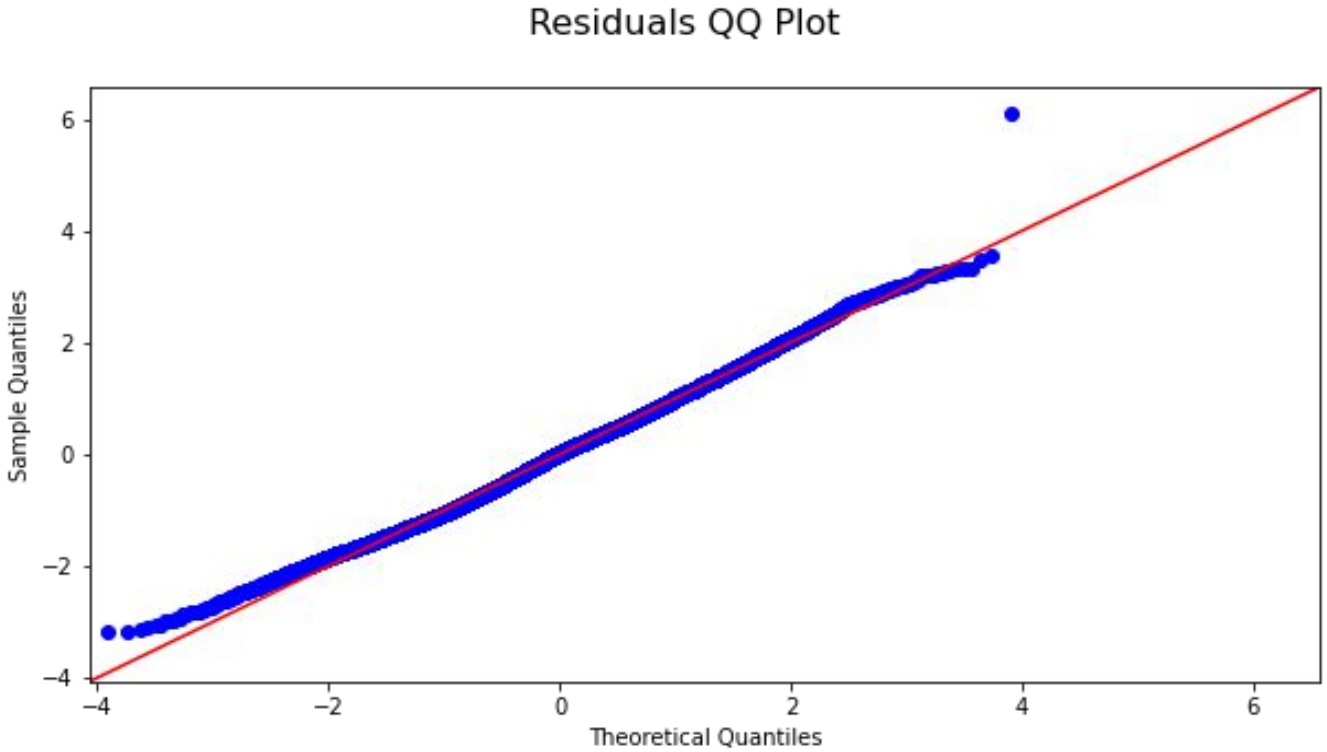
Model 2

- And multiple Regression with s**qft_living, bathrooms, bedrooms , floors, average_7 and Luxury_12 from the grade column** gives an r-squared value of **48.6%**

Residual Plots for Model to Check for Homoscedasticity

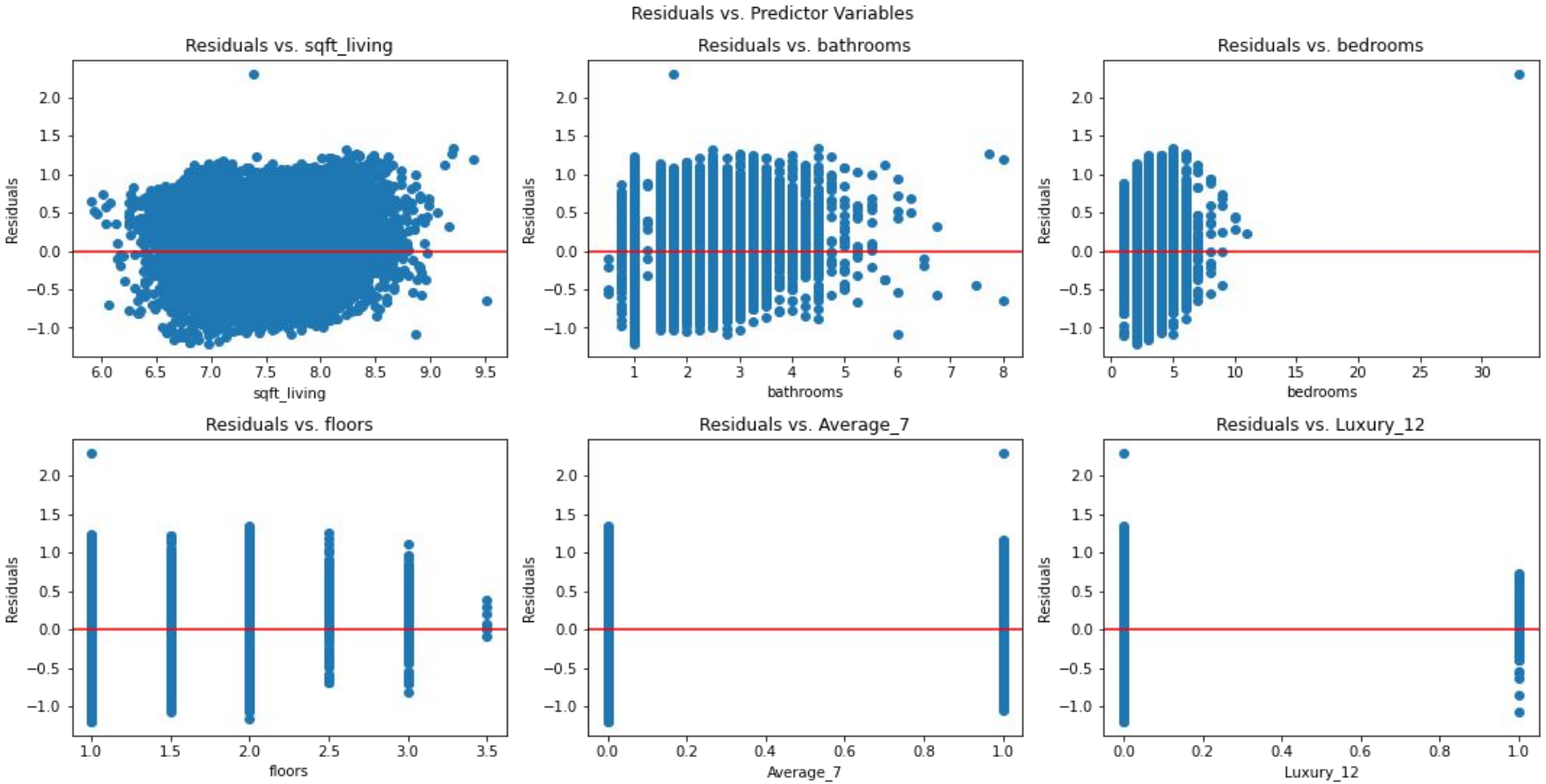


Testing for Normality of Residuals



Our model passess the assumption of normality since our residuals are normal

Residuals vs Predictor Variables



Recommendations.

1. The best predictor for house sales is sqft_living and grade. Hence investors can prioritize those two for increase in sales.
2. The model's applicability may differ due to regional differences(zip codes)
3. The model has its limitations as the need to log-transform one of the variables to satisfy the regression assumptions, any new data input needs to go through the same preprocessing