Group 3
presentation

FINAL PROJECT

Group members

- Calvine Dasilver **01**
- Jack Otieno **02**
- Salahudin Salat **03**
- Hellen Samuel **04**
- Sandra Kiptum **05**

06

DEMYSTIFYING HOUSE SALES ANALYSIS WITH REGRESSION MODELING IN A NORTHWESTERN COUNTY

PROJECT OVERVIEW

Business Understanding

The real estate market is a vital component of regional economic health and stability. This project delves into the dynamics of house sales in a specific northwestern county in the United States, aiming to unravel the key factors influencing property valuation in this area.

Problem Statements

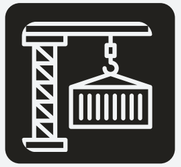
01 What are the most significant factors influencing house prices in this northwestern county?

02 How can we quantify the relationship between these factors and property value?

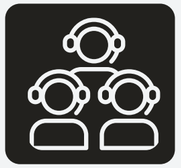
03 Can we develop a reliable model to predict house prices based on relevant characteristics?

Challenges

Real estate data complexity, encompassing diverse property features and local market trends.



Accurately identifying and quantifying the impact of each factor on house prices.



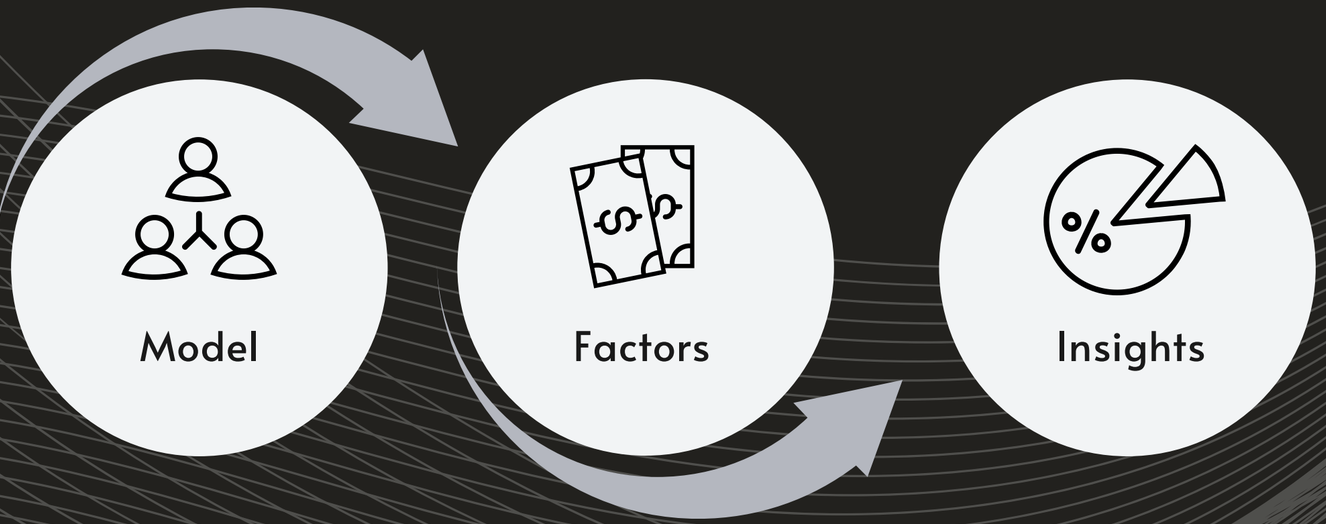
Consideration of external factors like economic conditions and interest rates.

Proposed Solutions

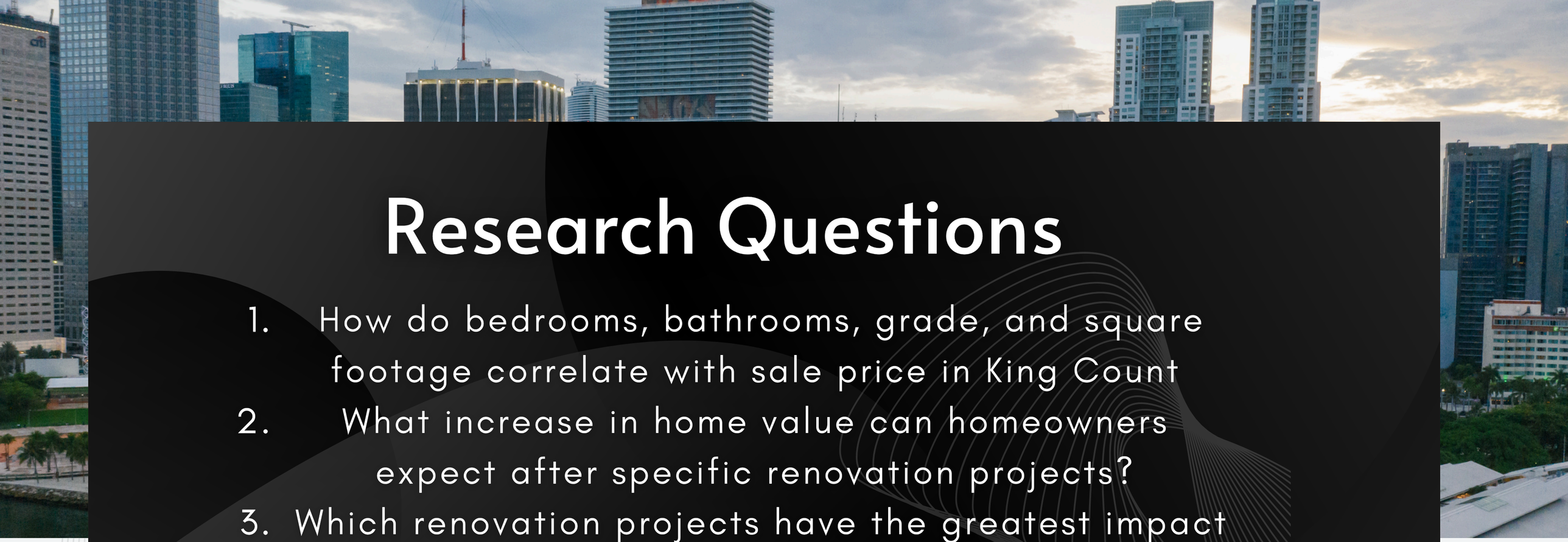
Utilizing multiple linear regression, a powerful machine learning technique, to analyze a large dataset of house sales and identify statistical relationships between property features and sale prices.

Objectives

1. Develop a robust multiple linear regression model for accurate house price prediction.
2. Identify significant factors influencing property value in the specific market.
3. Provide insights into regional housing market dynamics.



Preservation



Research Questions

1. How do bedrooms, bathrooms, grade, and square footage correlate with sale price in King Count
2. What increase in home value can homeowners expect after specific renovation projects?
3. Which renovation projects have the greatest impact on a home's market value?
4. Are there specific combinations of renovation projects that provide an interdependent effect on home value?

DATA UNDERSTANDING

Dataset Description

The analysis utilizes the King County House Sales dataset, comprising over 21,500 records and 20 distinct features. Spanning house sales from May 2014 to May 2015, the dataset offers a comprehensive snapshot of the housing market.

Key Columns

- id: Unique identifier for a house
- date: Date of house sale
- price: Sale price (prediction target)
- bedrooms, bathrooms, sqft_living, sqft_lot, floors, view, condition, grade, sqft_above, sqft_basement, yr_built, yr_renovated, zipcode, sqft_living15, sqft_lot15, sell_yr

Constraints and Considerations

Data may contain anomalies or inconsistencies necessitating careful examination.

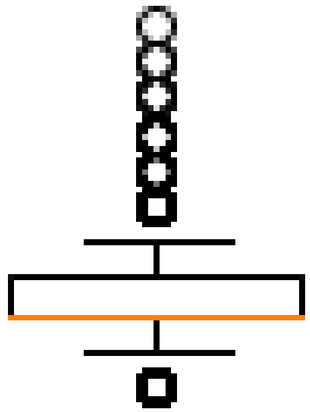
- Time frame (May 2014 - May 2015) may not fully reflect current market dynamics.
- Scope of data may not capture external factors such as interest rates or economic climate influencing property values.

Data preparation

we import the necessary functions and clean the data in the following ways

1. checking the data and null values
2. deleting the columns with null values
3. checking for non-numeric columns
4. checking for duplicates
5. creating the necessary columns
6. checking for outliers using the box plot and deleting the outliers

checking for outliers using the box plot and deleting the outliers



EXPLANATORY DATA ANALYSIS

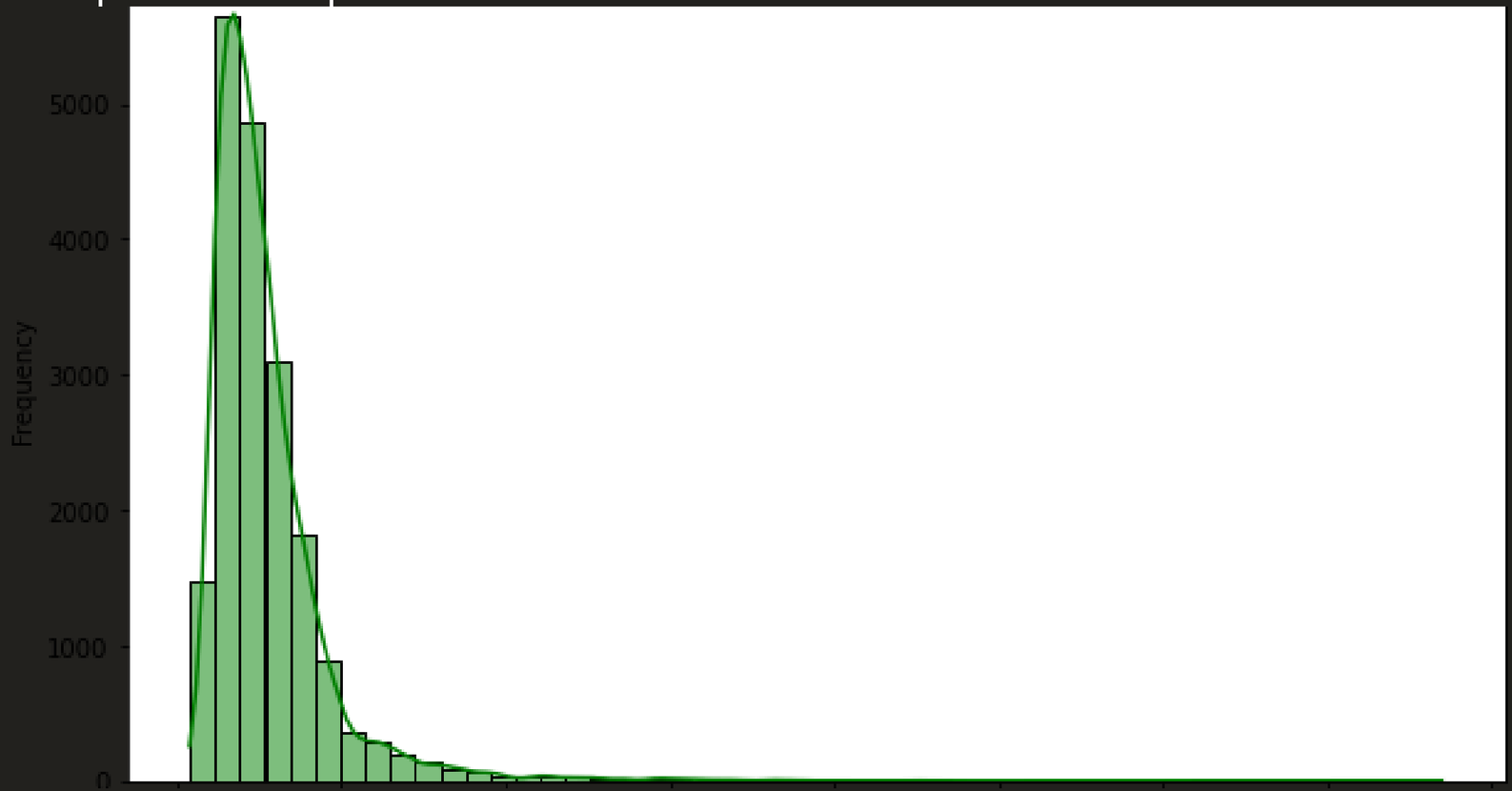
IN THIS SECTION, WE WILL PERFORM EXPLORATORY DATA ANALYSIS (EDA) TO UNDERSTAND THE DATA BETTER AND DISCOVER ANY PATTERNS, TRENDS USING UNIVARIATE,BIVARIATE AND MULTIVARIATE ANALYSIS

WE WILL USE DESCRIPTIVE STATISTICS AND VISUALIZATIONS TO SUMMARIZE THE MAIN CHARACTERISTICS OF THE DATA AND EXAMINE THE RELATIONSHIPS BETWEEN THE FEATURES AND THE TARGET VARIABLE.

WE WILL ALSO CHECK THE DISTRIBUTION AND CORRELATION OF THE VARIABLES AND IDENTIFY ANY POTENTIAL PROBLEMS OR OPPORTUNITIES FOR THE ANALYSIS.

Univariate Analysis

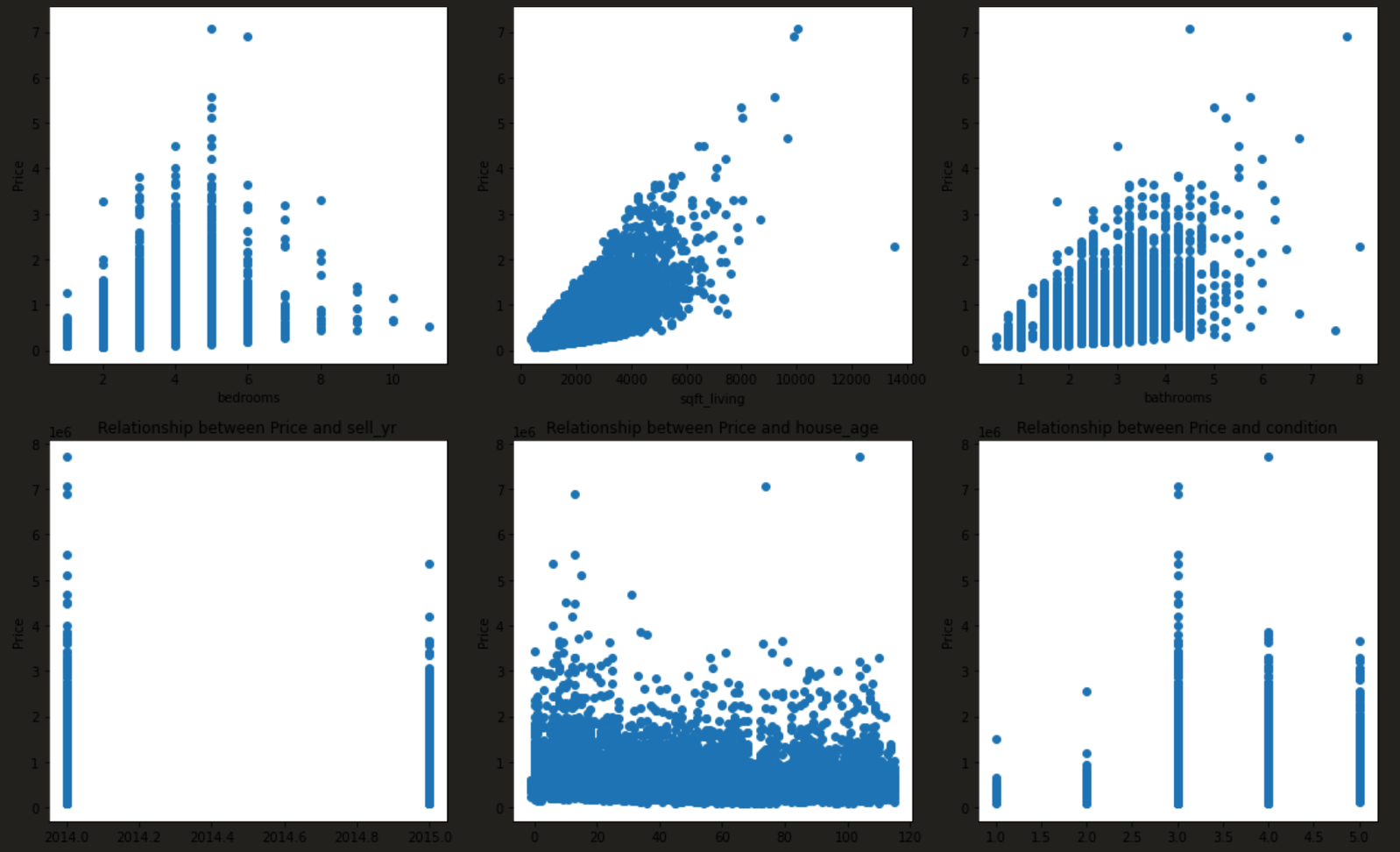
Univariate analysis involves the examination of single variables.We focus in the summary statistics of target variable-price to help us undersatand the distribution and skewness of house prices.



The histogram shows that the distribution of house price is positively skewed suggesting that while most houses are concentrated around lower prices, there are some properties with significantly higher prices.

Bivariate Analysis

We perform bivariate analysis to examine the relationship between the target variable - price and the other numeric and continuous features in the data using the scatter plots to show the direction, strength, and shape of the relationship between two numeric variables.

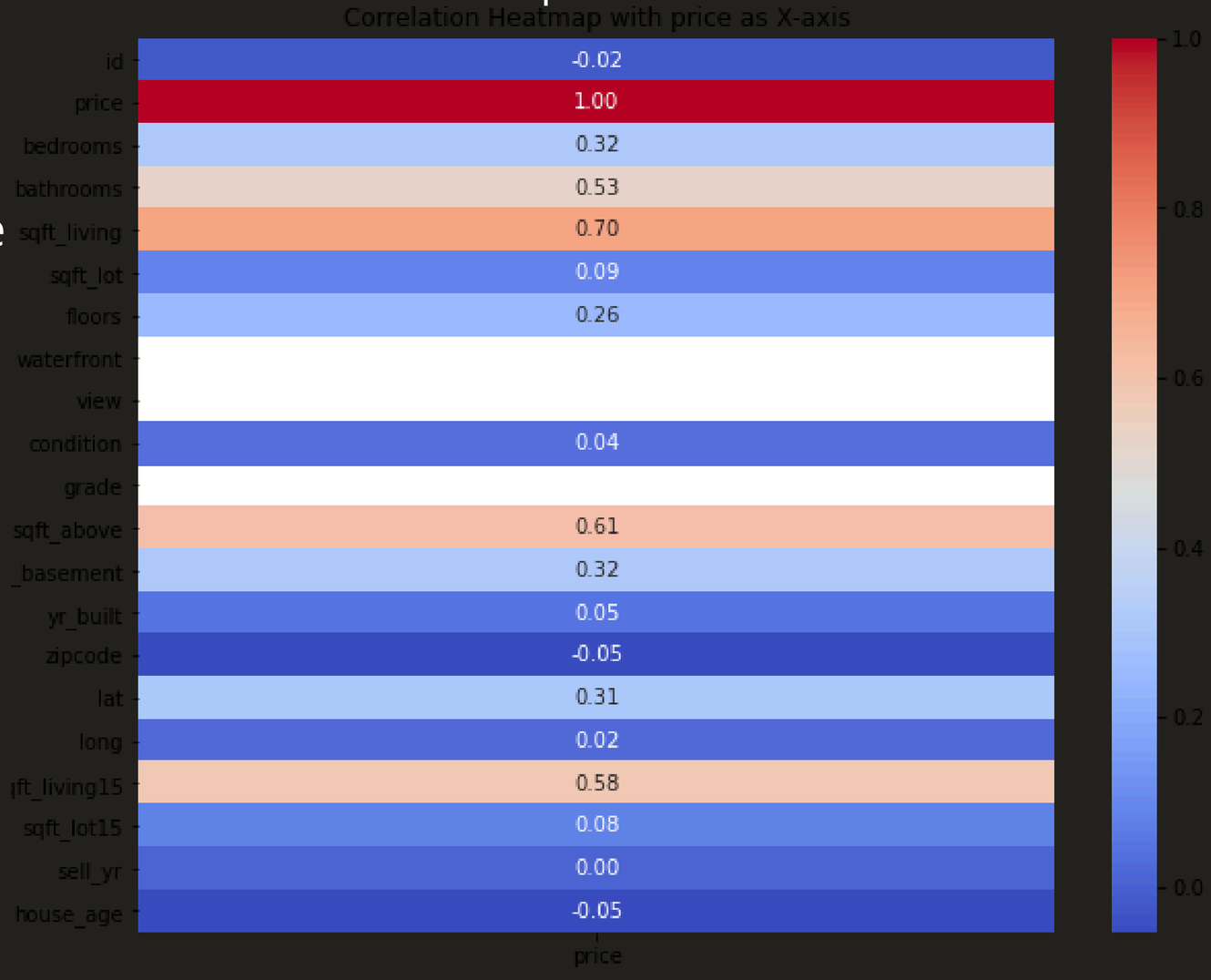


The scatter plots show that there is a positive relationship between most of the independent variables and the price of a house. This means that houses with higher values for these variables tend to be more expensive

Multivariate Analysis

In this section, we will perform multivariate analysis to examine the relationship between the target variable - price and multiple features in the data. We will use heatmap to visualize the correlation matrix of the features and see how they are related to each other and to the price.

The heatmap shows that Positive correlations are typically represented by shades of red, and negative correlations by shades of blue. We note that bathrooms and sqft_living are highly positively correlated.



REGRESSION MODELLING

Why did we do regression?

- **Regression is done to build models that will help us predict house prices given some house features for a stakeholder.**
- **It will also a renovator know what to improve on to increase the value of their house**

For our regression models we built 3 models

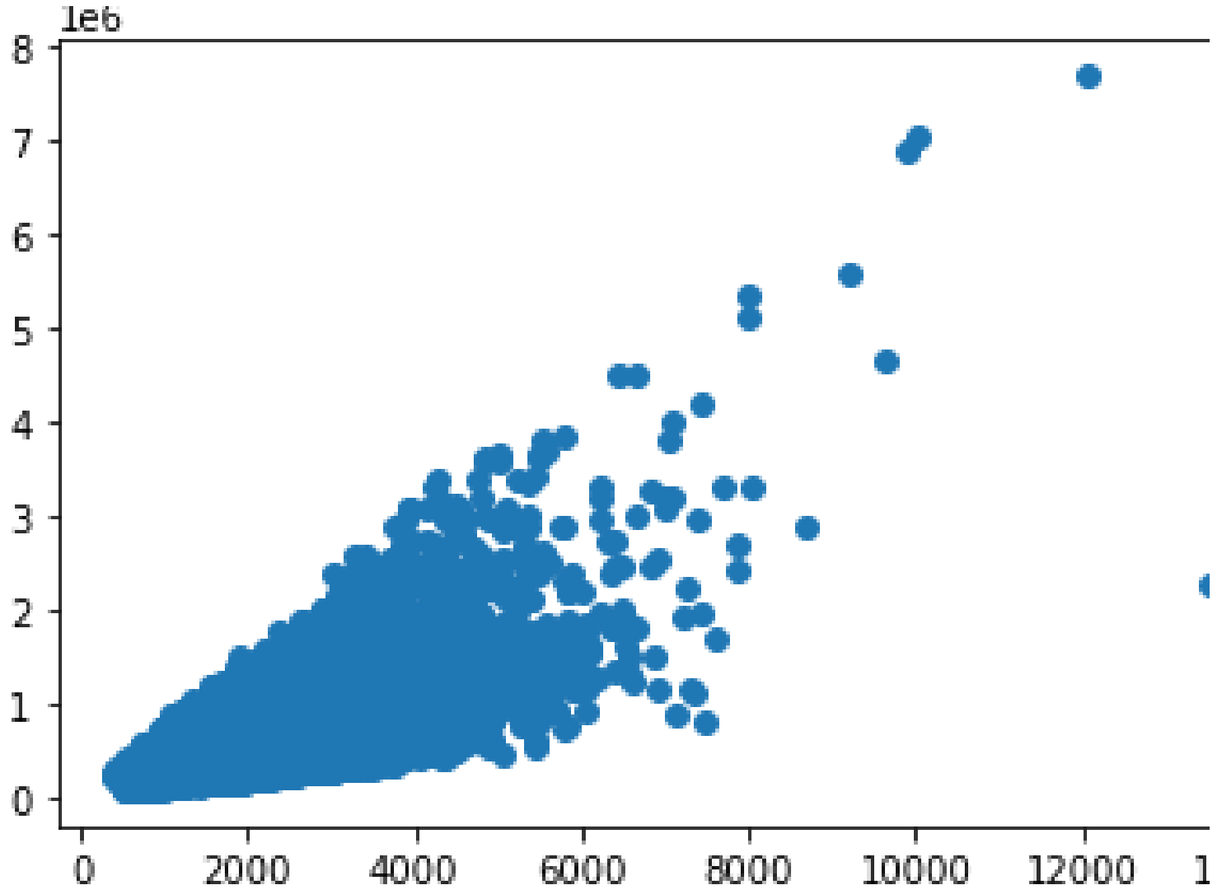
- **1 simple linear**
- **2 multiple linear**

First we will start with our simple linear model

SIMPLE LINEAR REGRESSION

What is simple linearregression?
Simple linear regression is building models from one independent variable or predictor

We built our model based off of the sqft_iving column as it has the highest correlation to price(0.704441)



From this we can see the linear relationship

BUILDING OUR MODEL

We built an OLS model and these were our results

- We can see from ou R-Squared that of model covers 49.6% of the variance in our data Which is not really good thus necessitating use of more predictors
- form the prob(fstatistic) we can tell that our model is statistically significant as the p_value is well below our alpha(0.05)
- our intercept has a p_value less than our alpha(5%) thus it is statistically significant, from this we can say that if the square square footage of the living room in the home is zero the price will be about -44000 USD, which doesn't really make sense but it is useful for our modeL
- our slope coefficient is 280.8688 showing that an increase of 1 square foot living is associated with an increase of about 280.8688 USD in price

We also need to check if our model also meets theLINE specifications;

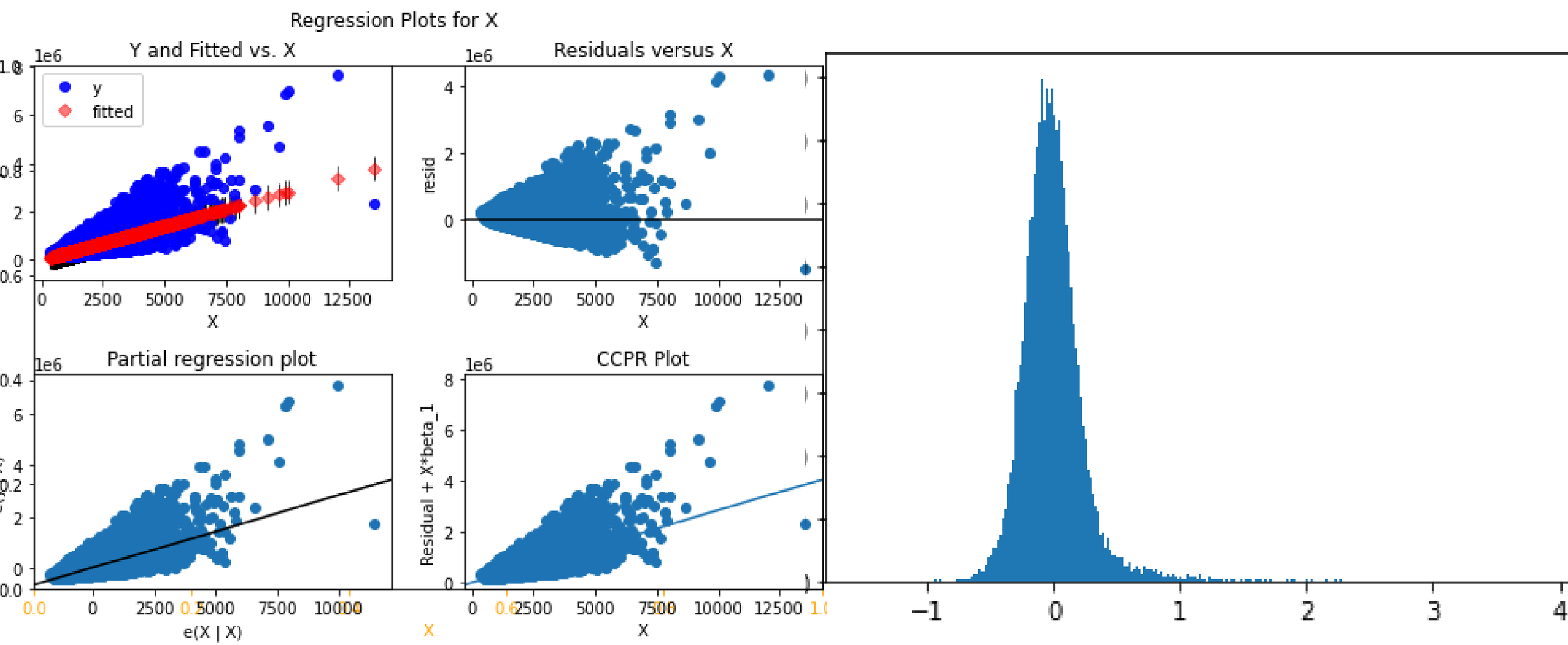
- L - Linearity(Relationship between x and y should be linear)
- I - Independence(the observations and errors are independent of each other)
- N - Normality(the residuals are normarly distribute)

OLS Regression Results

Dep. Variable:	y	R-squared:	0.496
Model:	OLS	Adj. R-squared:	0.496
Method:	Least Squares	F-statistic:	1.887e+04
Date:	Tue, 30 Apr 2024	Prob (F-statistic):	0.00
Time:	23:49:21	Log-Likelihood:	-2.6636e+05
No. Observations:	19163	AIC:	5.327e+05
Df Residuals:	19161	BIC:	5.327e+05
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	-4.863e+04	4697.231	-10.353	0.000	-5.78e+04	-3.94e+04
X	283.4081	2.063	137.385	0.000	279.365	287.452

Omnibus:	13129.896	Durbin-Watson:	1.987
Prob(Omnibus):	0.000	Jarque-Bera (JB):	481917.036
Skew:	2.817	Prob(JB):	0.00
Kurtosis:	26.913	Cond. No.	5.62e+03



from this we can see that
our residuals are not normally distributed.
There is no homoscedasticity

MULTIPLE LINEAR REGRESSION

We built this model to hopefully fill the gaps that the simple model left

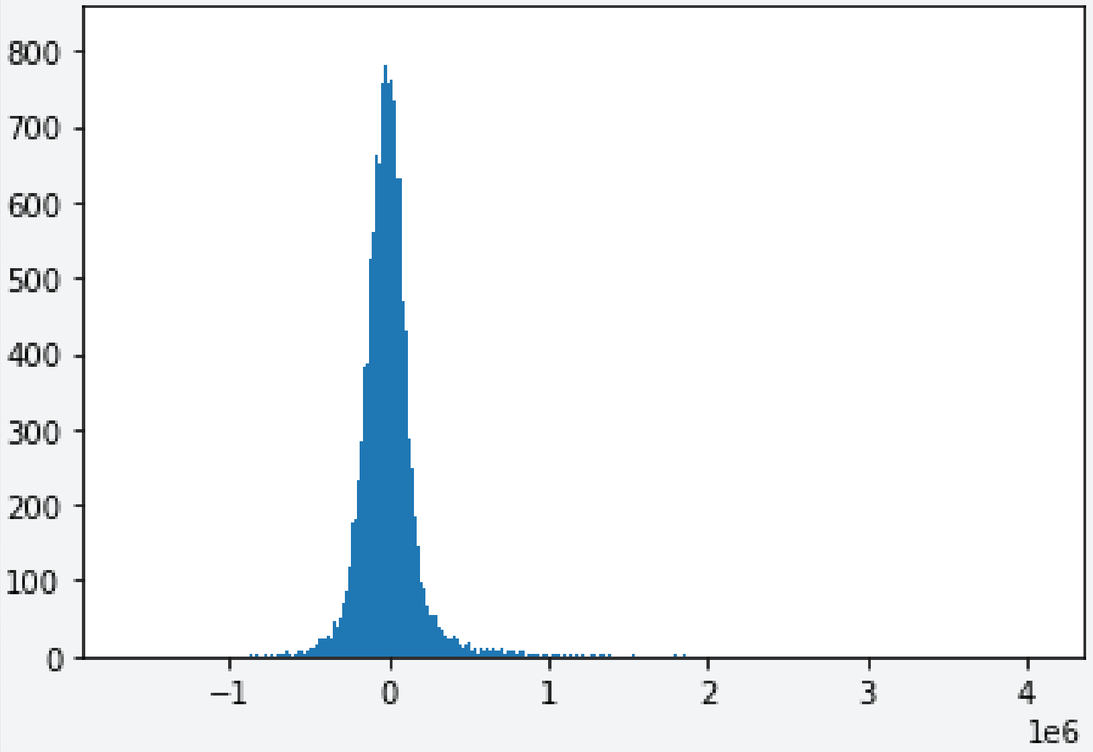
These are the predictors that we used
'bedrooms', 'bathrooms', 'sqft_living', 'sqft_lot', 'floors', 'sqft_above', 'sqft_basement', 'yr_built', 'lat', 'long', 'sqft_living15', 'sqft_lot15', 'sell_yr', 'house_age', 'condition_2', 'condition_3', 'condition_4', 'condition_5', 'view_1', 'view_2', 'view_3', 'view_4', 'grade_4', 'grade_5', 'grade_6', 'grade_7', 'grade_8', 'grade_9', 'grade_10', 'grade_11', 'grade_12', 'grade_13', 'Waterfront_1'

As you can see we had to use some dummies on our categorical data and drop one of the dummy columns to remove perfect multi collinearity

OLS Regression Results
Dep. Variable: price R-squared: 0.728
Model: OLS Adj. R-squared: 0.728
Method: Least Squares F-statistic: 1602.
Date: Tue, 30 Apr 2024 Prob (F-statistic): 0.00
Time: 23:49:35 Log-Likelihood: -2.6045e+05
No. Observations: 19163 AIC: 5.210e+05
Df Residuals: 19130 BIC: 5.212e+05
Df Model: 32
Covariance Type: nonrobust

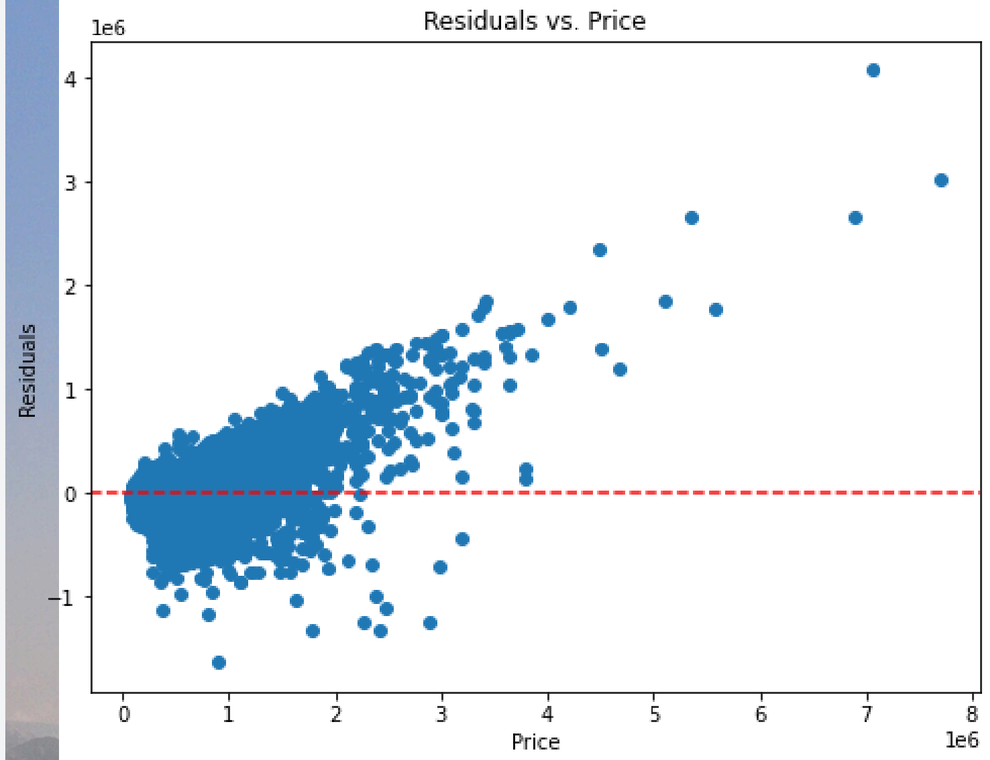
Checking for line features

Graph 1

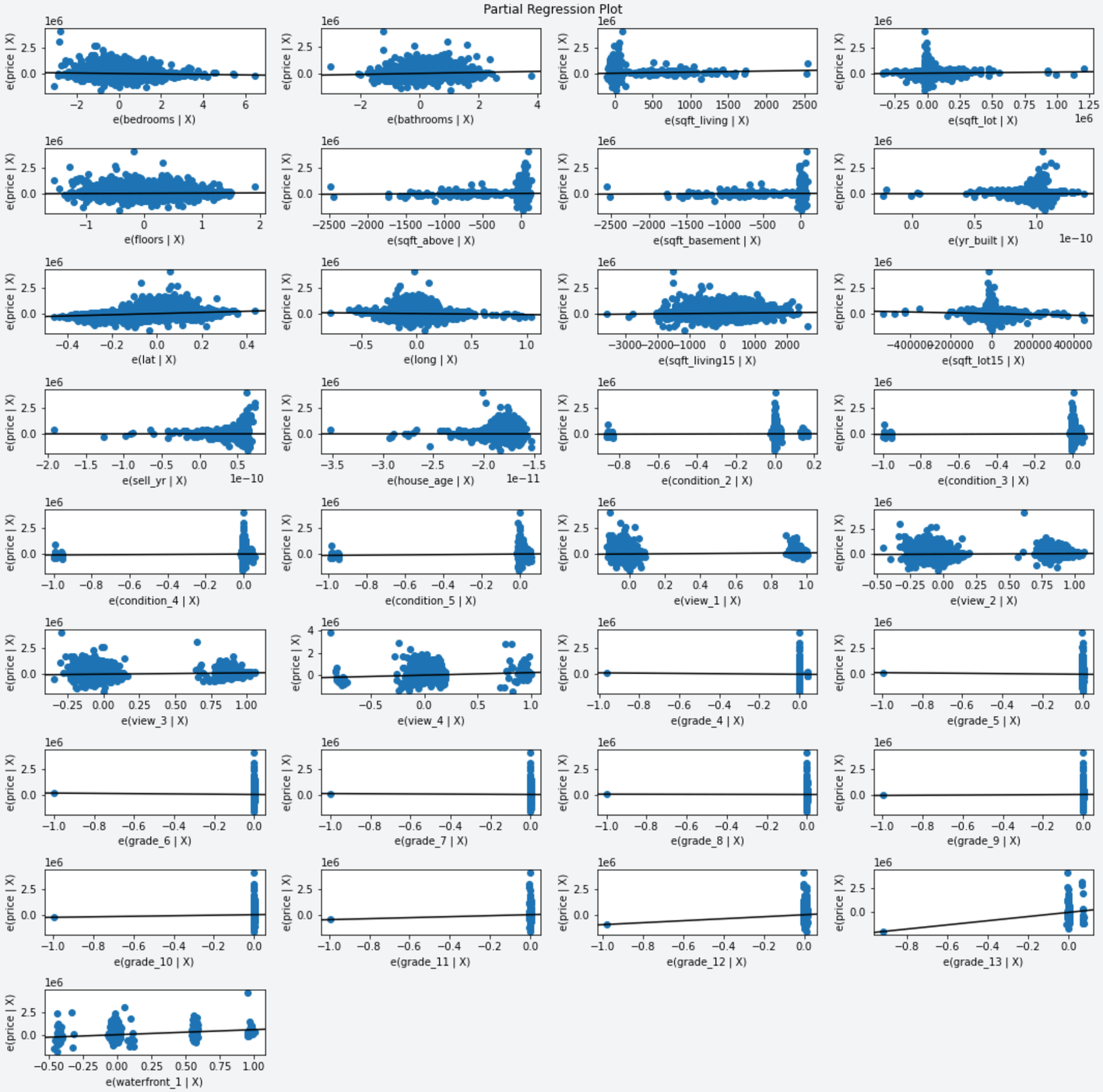


Errors are still not normally distributed from the diagram above

Graph 2



From this(above) there is better distribution of errors but there is still homoskedacity



We have very few linear distributions and a terrible mse of 1.9281434833350915e+17
We will do better with this in the next model

MODEL 3

The predictors we used are bathrooms','sqft_living','grade', 'sqft_above','sqft_living15' as the have better correlation thus a more linear relationship with price

We now have a new mse of 230179.555569266 which is much better

From the calculated MAE and RMSE we can tell that the third model is way accurate compared to the previous 2 models and we recomend using 3rd model in predicting the price of a house.

MODEL SUMMARY

OLS Regression Results
Dep. Variable: price R-squared (uncentered): 0.870
Model: OLS Adj. R-squared (uncentered): 0.870
Method: Least Squares F-statistic: 9176.
Date: Wed, 01 May 2024 Prob (F-statistic): 0.00
Time: 00:04:44 Log-Likelihood: -2.6430e+05
No. Observations: 19163 AIC: 5.286e+05
Df Residuals: 19149 BIC: 5.287e+05
Df Model: 14
Covariance Type: nonrobust

RECCOMENDATIONS

1. FOR STAKE HOLDERS

- **Consider the square footage of the house if you are buying a house for resale**
- **If renovating and buying as well an increase in the number of bathrooms will increase the value off the house**

_Those are the 2 factors that greatly impact the cost of a house

2. FOR MODEL DEVELOPERS

- **Handle all outliers for better models**
- **Use linearly related data for better models**

THANK YOU

group 3 presentation