ASSIGNMENT SUBMISSION COVER SHEET

Programme Title:	**MSc Data Analytics**
Module Code and Title:	**B9DA109 Machine Learning and Pattern Recognition**
Assessment Title:	**Group Report: House Prices Prediction with Linear Regression**

Student Names
1. **Chisimdiri Anyaogu**
2. **Saima Khan**
3. **Michael Anthony Nse**

TABLE OF CONTENTS

TABLE OF FIGURES

1. INTRODUCTION

1.1 PROBLEM STATEMENT

Leading real estate firm in Washington, USA, Chants wants to change the buying and selling of real estate property using complex algorithms for predicting house prices. The firm acknowledges how vital it is to calculate residential property values correctly in order to make wise investment choices, maximize rewards and reduce risks.

However, the current approaches of figuring out property prices sometimes rely on subjective evaluations or sparse data sources, creating uncertainties and substantial monetary losses. By using cutting-edge machine learning techniques and in-depth data analytics strategies to build a reliable predictive model, Chants aims to overcome this difficulty.

The main goal of this project is to create a machine learning algorithm that can predict the value of real estate with accuracy using a variety of important variables, such as number of bedrooms, bathrooms, floors, and property size among other variables. Chants strives to offer its clients accurate and data-driven property valuation estimations by using sophisticated regression models and structured data sources.

If this initiative is carried out successfully, Chants will be able to provide its clients an advantage in the real estate trade. The firm will be able to maximize its investment methods, find undervalued assets, better negotiate contracts, and eventually increase profits.

1.2 DATASETS

To meet the needs of Chants, I decided to use a dataset from the popular data repository and community for data scientists, Kaggle. The dataset contains previous information of houses sold by the firm in Washington, USA(IBM, 2017). For this dataset, here are the variables and their descriptions.

ID	A notation distinct to each house
Date	Date house was sold
Price	Price house was sold for, also the target (dependent) variable for this dataset
Bedrooms	Number of bedrooms
Bathrooms	Number of bathrooms
Sqft_living	Square footage of home
Sqft_lot	Square footage of lot
Floors	Total number of floors in the house
Waterfront	Indication of house having a view to a waterfront
View	Number of how many customers have viewed the property before it was purchased
Condition	Overall condition of the property
Grade	Overall grade given to property, by grading system
Sqft_above	Square footage of the house apart from basement
Sqft_basement	Square footage of the basement
Yr_built	Year property was built
Yr_renovated	Year house was last renovated
Zip code	Zip code
Lat	Latitude coordinates
Long	Longitude coordinates

Sqft_living15	Living area size in 2015 – implies some renovations on the property
Sqft_lot15	Lot size area in 2015 – implies some renovations on the property

To create a complex machine learning algorithm, the dataset must be processed, explored, scaled, and split between dependent and independent variables as is the case when dealing with machine learning algorithms.

1.3 ALGORITHM SELECTION

The choice of machine learning algorithm for the task at hand relies on the data type of the dependent (target) variable. In this case, the dependent (target) variable, price, is a continuous variable and for that reason we are to use a Regression model for machine learning predictions. Regression is a statistical technique that relates a dependent variable to one or more independent variables. A regression model can identify changes in the dependent variable and how they are associated with changes in one or more of the independent variables (Beers et al., 2023). One of the underlying tasks in this project is to build multiple models and highlight which of them would work best in the real world, to achieve this, there would be a Regression model using various optimisation algorithms like Gradient Descent. Gradient Descent is an iterative first-order optimisation algorithm used to find a local minimum/maximum of a given function(Kwiatkowski, 2021). It is commonly used to minimise cost/loss functions.

Another task we added on is to tweak the model with Regularisation methods like L1 Lasso regularisation and L2 Ridge regularisation techniques which will be discussed in later sections of this report.

2. DATA PREPARATION

2.1 DATA PRE-PROCESSING

Data pre-processing mainly deals with data cleaning. Data cleaning can be broken down into four tasks namely.

1. Handling Missing Values: rarely does real-world data appear without missing values, we cannot expect to build accurate machine learning models while having missing values in our data set. These missing values must be addressed depending on the data type and the discretion of the user.
2. Removing Duplicate Rows: rows appearing more than once introduces bias in our data, so it is important duplicate rows be removed from the data set for an accurate model.
3. Dealing with Outliers: extreme isolated incidents provide insights on the data set, but they can also mislead analysis and skew our model if they are not handled correctly. There are different ways to handle these incidents also called outliers, so they don't affect the model's performance.
4. Data Consistency: it is important to check data types/formats and categorical variables in the data set. Data consistency is necessary for building an accurate model.

Next, I describe how each data pre-processing steps made in this project.

I. HANDLING MISSING VALUES

After writing Python code we determined there were no missing values in any of the features and rows of the data set, so no missing value imputation techniques was needed in this data set. Below is the result from the Python code used to check for missing values.

II. REMOVING DUPLICATE ROWS

After writing Python code to check for duplicated rows in the data set, it returned a result of none of the rows possessing any duplicates, subsequently, there was no need to apply any duplicates removal techniques.

III. DEALING WITH OUTLIERS

To detect outliers in this data set we defined a Python function to calculate the percentage of outliers possessed by each feature. The function worked using the Boxplot technique of outlier detection. In statistics, a boxplot is a method for graphically depicting groups of numerical data through their quartiles. Boxplots are separated into 5 points.

- Q1 / 25th percentile: the middle value between the smallest number and the middle of the data set.
- Q2 / 50$^{th:}$ percentile: the middle point of the data set.
- Q3 / 75th percentile: the middle value between the median and highest value of the data set, this is not the maximum.

Before we visit the remaining two points on the boxplot, we must first define Interquartile Range (IQR), which is the difference between Q3 and Q1 of the data set.

- Maximum is simply derived by calculating Q3 + 1.5 * IQR.
- Minimum is derived by calculating Q1 – 1.5 * IQR.

The figure below shows the results from the Python function on the numerical features of the data set.

Figure 1- Outlier Percentages

We determined that the 'price', 'sqft_lot', 'view', 'grade', and 'sqft_lot15' features in the data set presented significant percentages of outliers. The next course of action is to take out these outliers in the respective features using the Z-score method of outlier removal.

IV. DATA CONSISTENCY

The first task to test for data consistency was to check the data types of all the features to make sure they aligned with the data they presented. The data types of all features were consistent with the data presented.

Below is the figure representing the results from that experiment done with Python.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 21613 entries, 0 to 21612
Data columns (total 21 columns):
 #   Column          Non-Null Count  Dtype
---  ------          --------------  -----
 0   id              21613 non-null  int64
 1   date            21613 non-null  object
 2   price           21613 non-null  float64
 3   bedrooms        21613 non-null  int64
 4   bathrooms       21613 non-null  float64
 5   sqft_living     21613 non-null  int64
 6   sqft_lot        21613 non-null  int64
 7   floors          21613 non-null  float64
 8   waterfront      21613 non-null  int64
 9   view            21613 non-null  int64
 10  condition       21613 non-null  int64
 11  grade           21613 non-null  int64
 12  sqft_above      21613 non-null  int64
 13  sqft_basement   21613 non-null  int64
 14  yr_built        21613 non-null  int64
 15  yr_renovated    21613 non-null  int64
 16  zipcode         21613 non-null  int64
 17  lat             21613 non-null  float64
 18  long            21613 non-null  float64
 19  sqft_living15   21613 non-null  int64
 20  sqft_lot15      21613 non-null  int64
dtypes: float64(5), int64(15), object(1)
```

Figure 2 – Dataset Info

Next, we checked for the unique values present in the categorical variables to ensure there were no mistakes in the entries. Below is the figure that shows the results from the Python code to check for unique values.

```
unique values for bedrooms : [ 3  2  4  5  1  6  7  0  8  9 11 10 33]
unique values for bathrooms : [1.   2.25 3.   2.   4.5  1.5  2.5  1.75 2.75 3.25 4.   3.5  0.75 4.75
 5.   4.25 3.75 0.   1.25 5.25 6.   0.5  5.5  6.75 5.75 8.   7.5  7.75
 6.25 6.5 ]
unique values for view : [0 3 4 2 1]
unique values for floors : [1.  2.  1.5 3.  2.5 3.5]
unique values for condition : [3 5 4 1 2]
unique values for grade : [ 7  6  8 11  9  5 10 12  4  3 13  1]
```

Figure 3 - Unique Values for Discrete Variables

We found that ‘bathroom’ and ‘floors’ features had decimal figures which should not be the case, to solve this we rounded the values to the nearest whole numbers and changed the data types to int64.

2.2 EXPLORATORY DATA ANALYSIS

In this section of the project, we performed Exploratory Data Analysis, commonly known as EDA. Exploratory Data Analysis refers to the critical process of performing initial investigations on data to discover patterns, to spot anomalies, to test hypothesis and to check assumptions with the help of summary statistics and graphical representations(P. Patil, 2018). We divided EDA into four separate tasks.

- Descriptive Statistics
- Data Visualization
- Feature Engineering

I. DESCRIPTIVE STATISTICS

This is a summary statistics table of the features contents in the data set. It features important statistical values such as count, mean, standard deviation, and the 5 Boxplot points we discussed in earlier sections. Below is a figure of the summary statistics of the data set.

	id	price	bedrooms	bathrooms	sqft_living	sqft_lot	floors	waterfront	view	condition	grade	sqft_above
count	2.161300e+04	2.161300e+04	21613.000000	21613.000000	21613.000000	2.161300e+04	21613.000000	21613.000000	21613.000000	21613.000000	21613.000000	21613.000000
mean	4.580302e+09	5.400881e+05	3.370842	2.058715	2079.899736	1.510697e+04	1.534956	0.007542	0.234303	3.409430	7.656873	1788.390691
std	2.876566e+09	3.671272e+05	0.930062	0.755524	918.440897	4.142051e+04	0.554742	0.086517	0.766318	0.650743	1.175459	828.090978
min	1.000102e+06	7.500000e+04	0.000000	0.000000	290.000000	5.200000e+02	1.000000	0.000000	0.000000	1.000000	1.000000	290.000000
25%	2.123049e+09	3.219500e+05	3.000000	2.000000	1427.000000	5.040000e+03	1.000000	0.000000	0.000000	3.000000	7.000000	1190.000000
50%	3.904930e+09	4.500000e+05	3.000000	2.000000	1910.000000	7.618000e+03	2.000000	0.000000	0.000000	3.000000	7.000000	1560.000000
75%	7.308900e+09	6.450000e+05	4.000000	2.000000	2550.000000	1.068800e+04	2.000000	0.000000	0.000000	4.000000	8.000000	2210.000000
max	9.900000e+09	7.700000e+06	33.000000	8.000000	13540.000000	1.651359e+06	4.000000	1.000000	4.000000	5.000000	13.000000	9410.000000

Figure 4 - Descriptive Statistics Sample

After studying the values in the table, we determined there were no errors in the data.

II. DATA VISUALIZATION

We aimed to use graphs and charts to understand individual elements in the data set. To do this we divided data visualization into three sections of univariate, bivariate and multivariate analyses.

Univariate Analysis: is the simplest form of analysing data. It means your data has only one variable. It doesn't deal with causes or relationships. The major purpose of univariate analysis is to summarize and find patterns of single variables in the data.

Firstly, I looked at the price column which is our target variable to find out the distribution of the prices with a histogram.

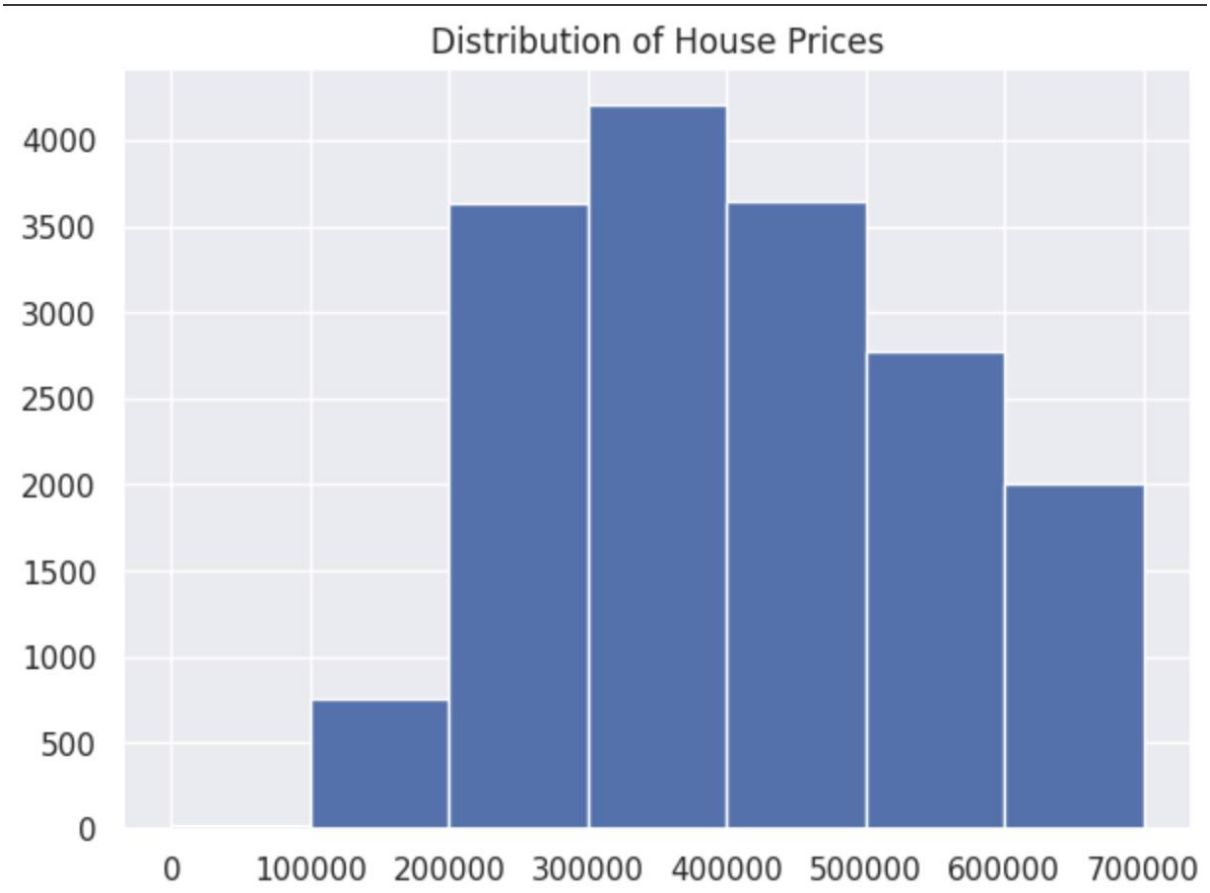

Figure 5 - Prices Histogram

Here are my inferences from that visual,

- Negative skew: From our observation we determined there is a slight negative skew in the histogram which signifies the distribution of prices lean towards higher values. This also means a larger portion of the prices in the data set are on the expensive end, with fewer properties having lower prices.
- Peak at 300,000 – 400,000: The histogram peaks at the 300,000 – 400,000 range suggesting that most prices fall between this range in the data set. Also, indicates that on average, properties in Washington, USA fall within these prices.
- Absence of significant outliers: Since we already took out the outliers in the pre-processing phase of data preparation, we notice the absence of isolated bars or significant outliers in the histogram. It indicates the remaining data points are within a reasonable range and that influential outliers that can affect model development have been removed from the data set.

Next, I used a bar chart to view the number of houses built by decades.

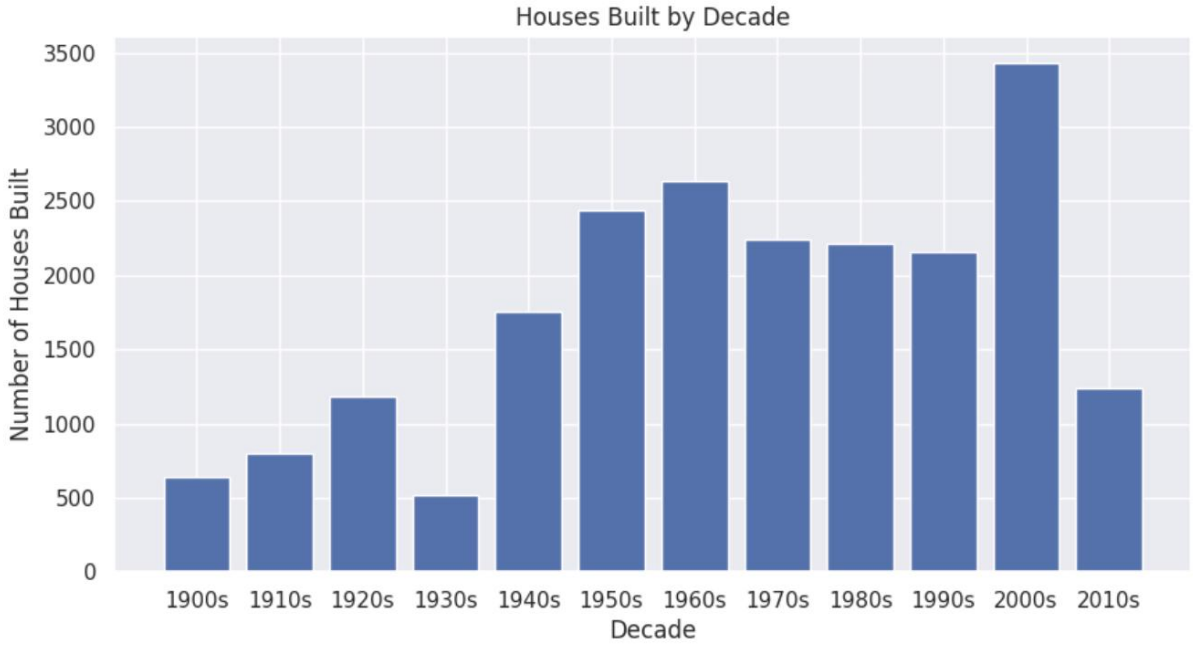

Figure 6 - House Built by Decades

Here are my inferences from that visual,

- Frequency of house construction: First we see a rise in house construction from 1900s till the 1920s and then a steep decline in the 1930s. There was a steady increase between 1940s to the 1960s, later there began a decline in construction of houses between the 1970s and 1990s, till the sudden rise in the 2000s where we saw a moderate number of houses listed in the data set were built in the 2000s and then a huge decline in the 2010s.
- Dominant decades: The construction sector enjoyed its highest increase in house construction in the 2000s and the 1960s.
- Periods of low construction: The 1930s had the lowest number of houses constructed with about 500 houses built in that time.
- We went a bit further in research to find out what exactly happened during periods of low and high amounts of constructions. The increase in house construction between the 1940s and 1960s is mostly attributed to the post-war boom, the Baby Boom era in the United States, government policies that made homeownership easier for people. While the decline between the 1970s and 1990s was due to economic challenges, that included periods of recession and high inflation rates. Another factor was demographic changes, the population growth rate during the 1970s and 1980s because the children from the Baby Boom era was over, and the children from that era had grown into adulthood, other factors like also that also contributed to the decline of house construction were changing preferences, people opted to live in rented condominiums, and townhouses. Also land availability and zoning restrictions

Afterwards, we plot bar charts of the discrete categorical variables, to check the various representations.

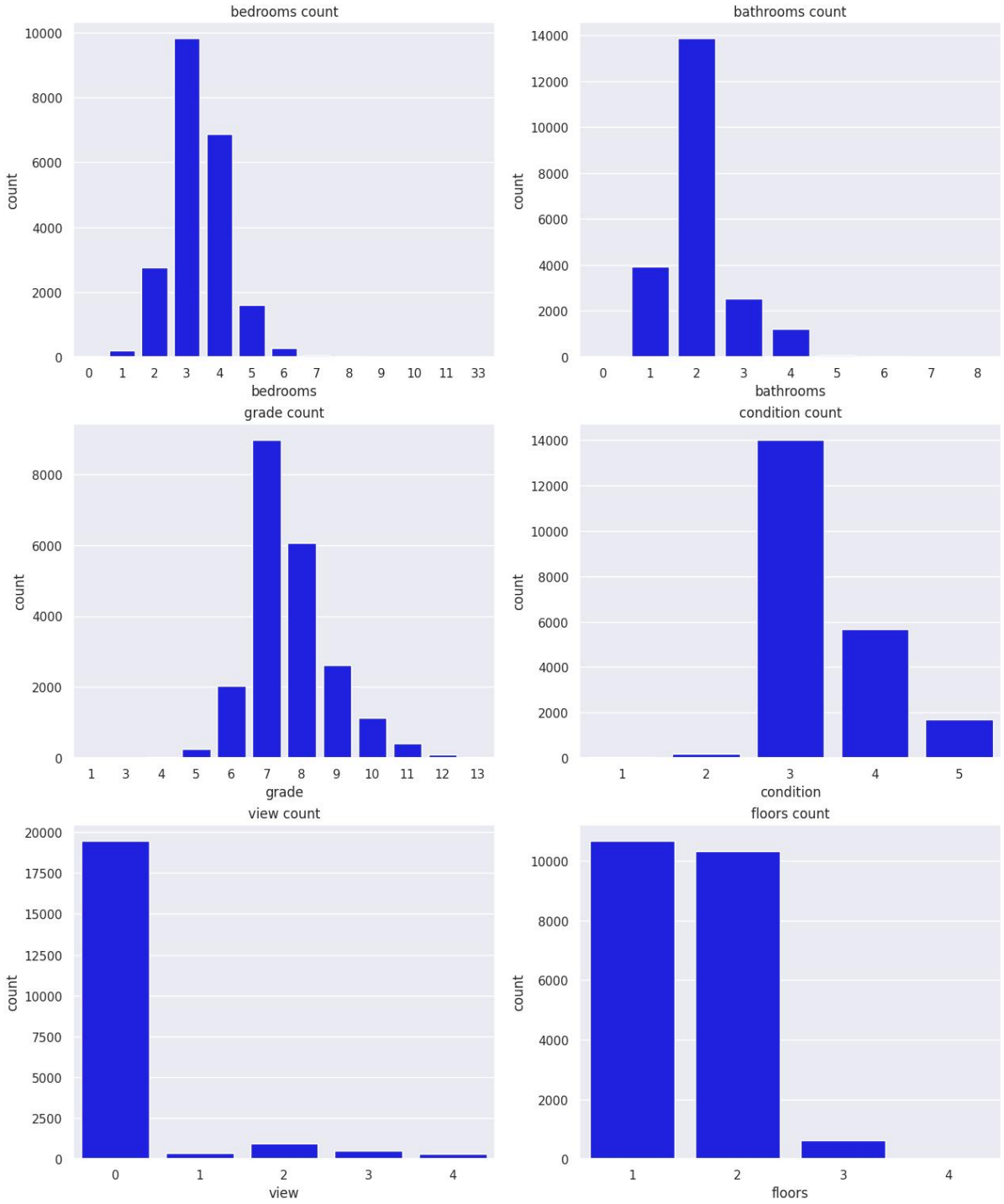

Figure 7 - Bar Charts of Discrete Variables

To do this we defined a Python function that allowed column names and subplot positions as inputs, which then put out bar charts of the mentioned columns as outputs in their respective subplot positions. Here are a few inferences we were able to attain from the plots.

- Bedrooms: The most common bedroom numbers for houses in Washington are three with around 9,500 houses in the data set having this number, the second highest

number is four-bedroom houses with a number around 7,000 houses in the data set. This suggests that most of the houses in the data set are family houses, with family houses typically ranging from three to four rooms. It also tells us about the people in Washington favour family houses.

- Bathrooms: Most houses in the data set possess two-bathrooms with almost 14,000, bringing us to the same conclusion from the bedroom's charts about most houses in the data set are family houses, because two-bathrooms are typical amount for family homes.
- Grade: The grades are given by the Kings County official grading system which ranges between 1 and 13, with 13 being the highest grade. Roughly, 9,000 houses have the average grade of seven (7), followed by 6,000 with a grade of eight (8). It's safe to say that majority of the houses in Washington are of average grade according to the official grading system. The least grade assigned to a house is one (1) with only 1 house in that category, while the highest grade is thirteen (13) with 13 houses in the category.
- Condition: This has to do with the overall condition of the house. It ranges between 1 and 5, with 5 being the best condition. Here, like the grades category, most of the houses are average in condition, with three (3) with 14,000 in this category, followed by four (4) with 6,000. Thirty houses have a condition of one (1).
- View: Refers to how many times the property in question was viewed before purchase. 19,000 of the 21,000 houses listed in the data set were never viewed by anyone ese before the purchase. The second highest is 963 homes viewed two times before purchase. This shows the disparity between the view counts in the data set.

- Floors: Represents the number of floors in each house in the data set. Majority of the houses in the data set have between 1 and 2 houses, with 10,680 houses having one (1) floor and 10,312 houses having two (floors).

Lastly, for the univariate analysis we examined the continuous variables in the data set to check the representations.

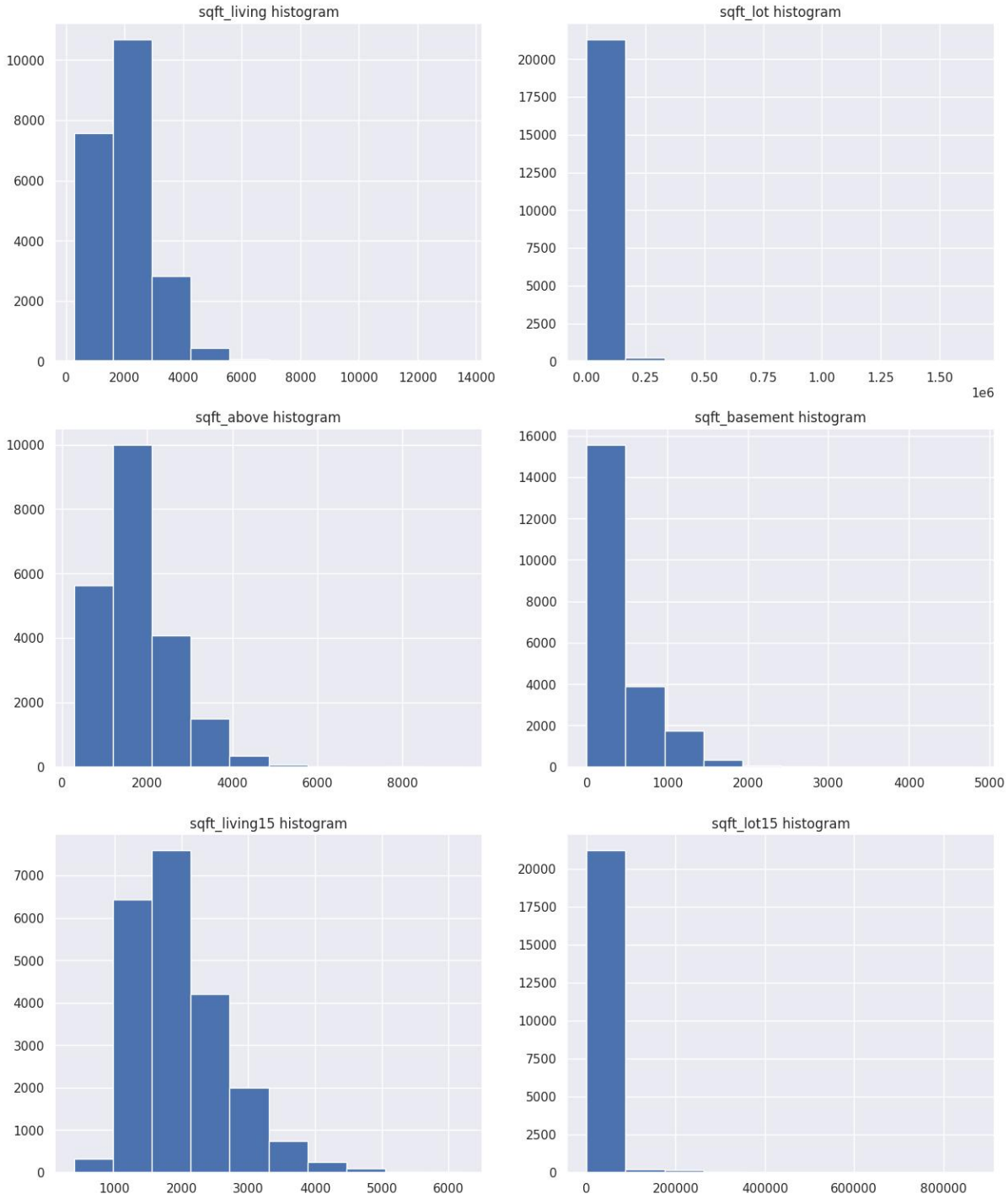

Figure 8 - Histograms of Continuous Variables

- Sqft living: We noticed the histogram was positively skewed (left skewed), with the square feet living area of most houses ranges between 2,000 and 3,000 square feet.
- Sqft lot: The histogram is highly positively skewed, with almost all the houses in the data set falling between 5,000 and 100,000 square feet.
- Sqft above: Also positively skewed, with majority of the houses having a square feet area aside the basement between 1,000 and 2,000 square feet.
- Sqft basement: The histogram is also highly positively skewed, with 15,800 houses have a basement square foot ranging between 0 and 500.
- Sqft living 15: This variable means renovations were done on the property. The histogram is positively skewed with the peak square feet area of living area ranging between 1,500 and 2,000 square feet.
- Sqft lot 15: Like the sqft living 15 variables, this also implies renovations were done on the property before sale. Histogram is positively skewed with 22,000 homes seeing ranges between 600 and 100,00 square feet.

Bivariate Analysis: Bivariate analysis is the statistical process of analysing two variables at once to examine their relationships. To comprehend the type, degree, and direction of their link, it includes evaluating how changes in one variable are related to changes in another variable.

Firstly, we looked at the relationship between the categorical variables and the target variable using box plots. We defined a Python function that looked very much like the ones we used for our univariate analysis.

Below is the figure of the results from the function.

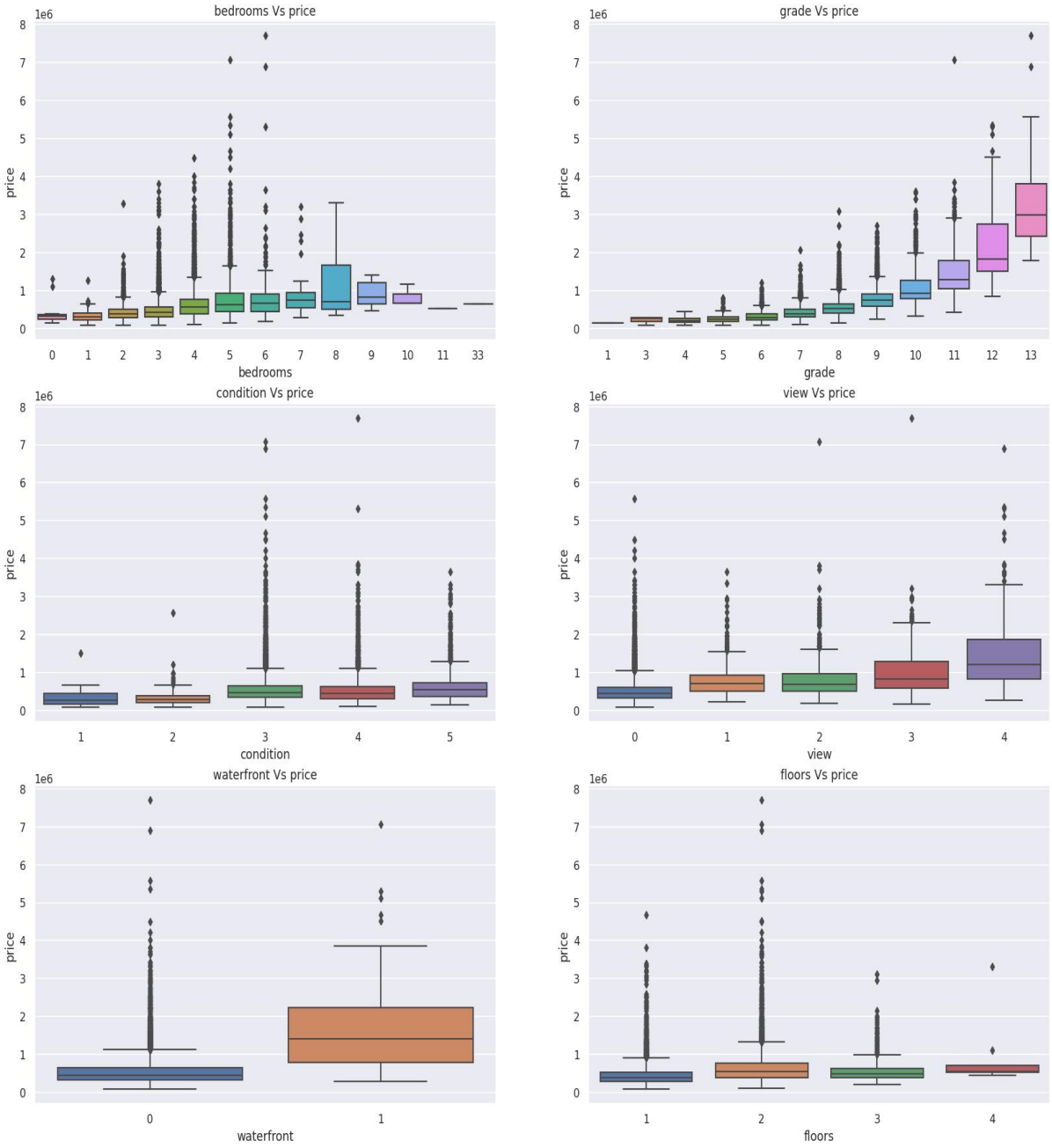

Figure 9 - Boxplots of Discrete Variables by Price

- Bedrooms: The price ranges for various bedroom counts are also shown in the box plot. There may be a link between the number of bedrooms and more expensive properties because homes with eight bedrooms had the broadest range of higher costs. On the other hand, the little representation of these categories in the dataset may be the cause of the narrow price range for homes with 11 and 33 bedrooms. When evaluating these data, it is crucial to take the sample size into account. It is noticeable that there are a substantial number of outliers for homes with two, three, four, five,

and six bedrooms, indicating exceptionally high prices for those specific categories. These anomalies could be special or exceptional properties with distinguishing characteristics or prime locations.

- Grade: The box plot indicates a positive relationship between grade score and prices, higher grade scores see higher house prices. Although houses with five, six, seven, eight, nine, ten, and eleven include exceptionally high number of outliers for those categories. These differences could be due to other deciding price factors.
- Condition: The box plot does not show much difference between the price ranges and conditions, with their medians being so closely related, but there are significant outliers in three, four and five that could be due to other factors. Although, five has the widest range of higher prices excluding outliers, while three and four have the same range.
- View: The box plot indicates a positive relationship between view count and prices, higher view counts. Higher view counts have wider ranges of higher prices excluding outliers. We also see significant outliers in houses with zero (0) view counts, which could be due to other deciding price factors.
- Waterfront: This variable indicates the presence of a view to waterfront, with 1 indicating yes and 0 indicating no. the box plot indicates a positive relationship between waterfront presence and higher prices. Houses with waterfront presence has a wider range of higher prices than houses without waterfront. Although, houses without waterfront presence have exceptionally high number of outliers, that could be due to other factors.
- Floors: The box plot does not show much difference between the price ranges and floors count, with their medians being so closely related. Although, houses with one,

two, and three floors showed significant outliers in respect to prices, which could be due to other factors.

Next, we looked at the relationship between the continuous variables and the target variable using scatter plots. We defined a Python function that looked very much like the ones we used for our univariate analysis.

Below is the figure of the results from the function.

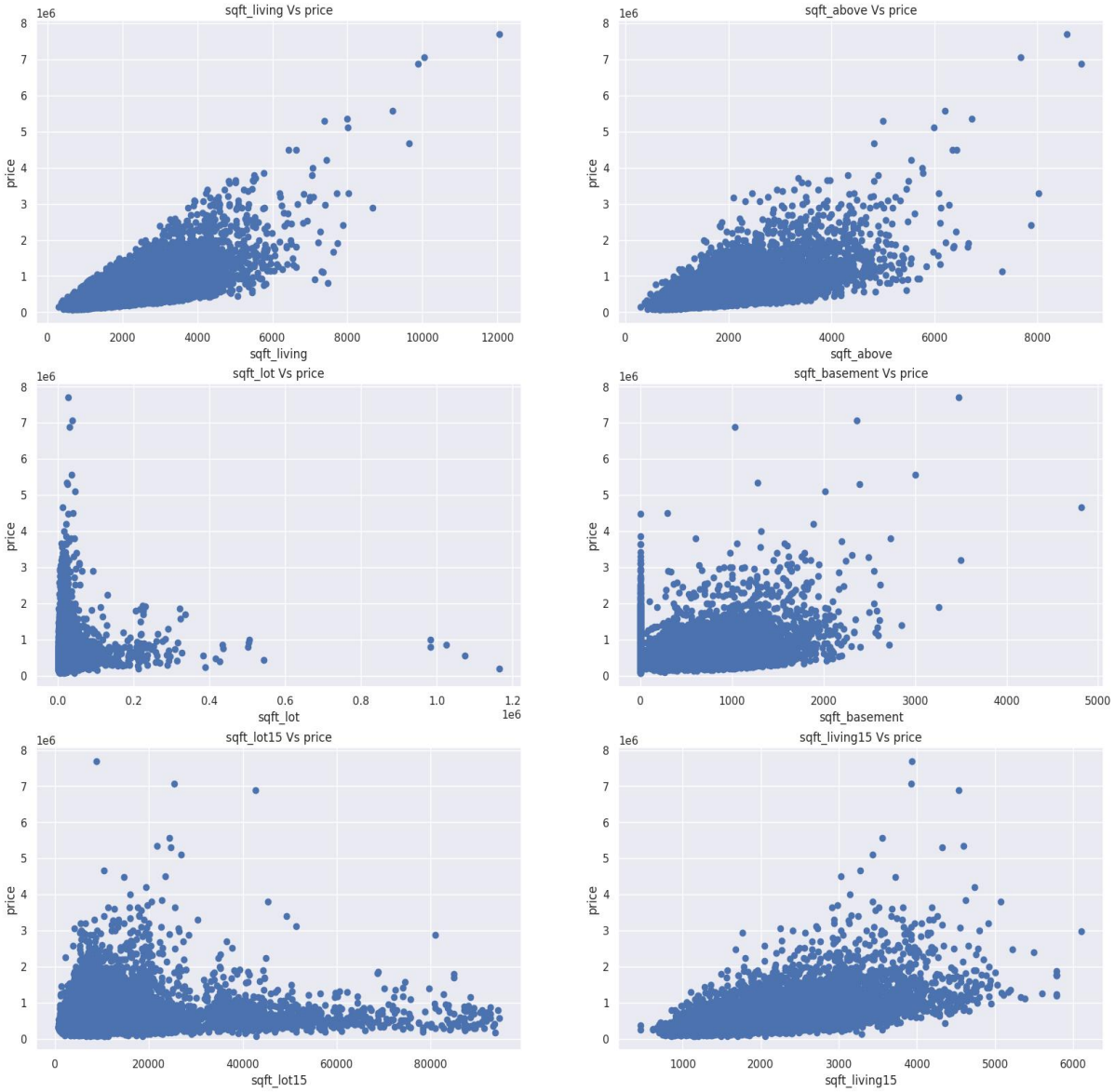

Figure 10 - Scatter Plots of Continuous Variables by Price

- Sqft living: There is a positive relationship between square feet of living area and price of the property. The data points follow a straight line; hence the relationship is linear.
- Sqft above: Same goes for sqft above, there is a positive relationship between square feet of area asides the basement and price of the property. The data points follow a straight line; hence the relationship is linear.
- Sqft lot: There is a negative relationship between area of the lot and price of the property, the relationship appears to be non-linear.
- Sqft basement: There is little to no relationship between the basement area and price of the property.
- Sqft lot 15: There is little to no relationship between the renovated lot area and price of the property.
- Sqft living 15: There is a positive non-linear relationship between the renovated living area and the price of the property.

Multivariate Analysis: This involves analysing the relationships between multiple variables (i.e., multivariate data) and understanding how they influence each other. It is an important tool that helps us better understand complex data sets to make data-driven and informed decisions(Mahmoodinobar, n.d.).

Firstly, we looked at a correlation matrix of the data set with the aim of trying to pinpoint relationships between each variable.

Below is the figure of the correlation matrix.

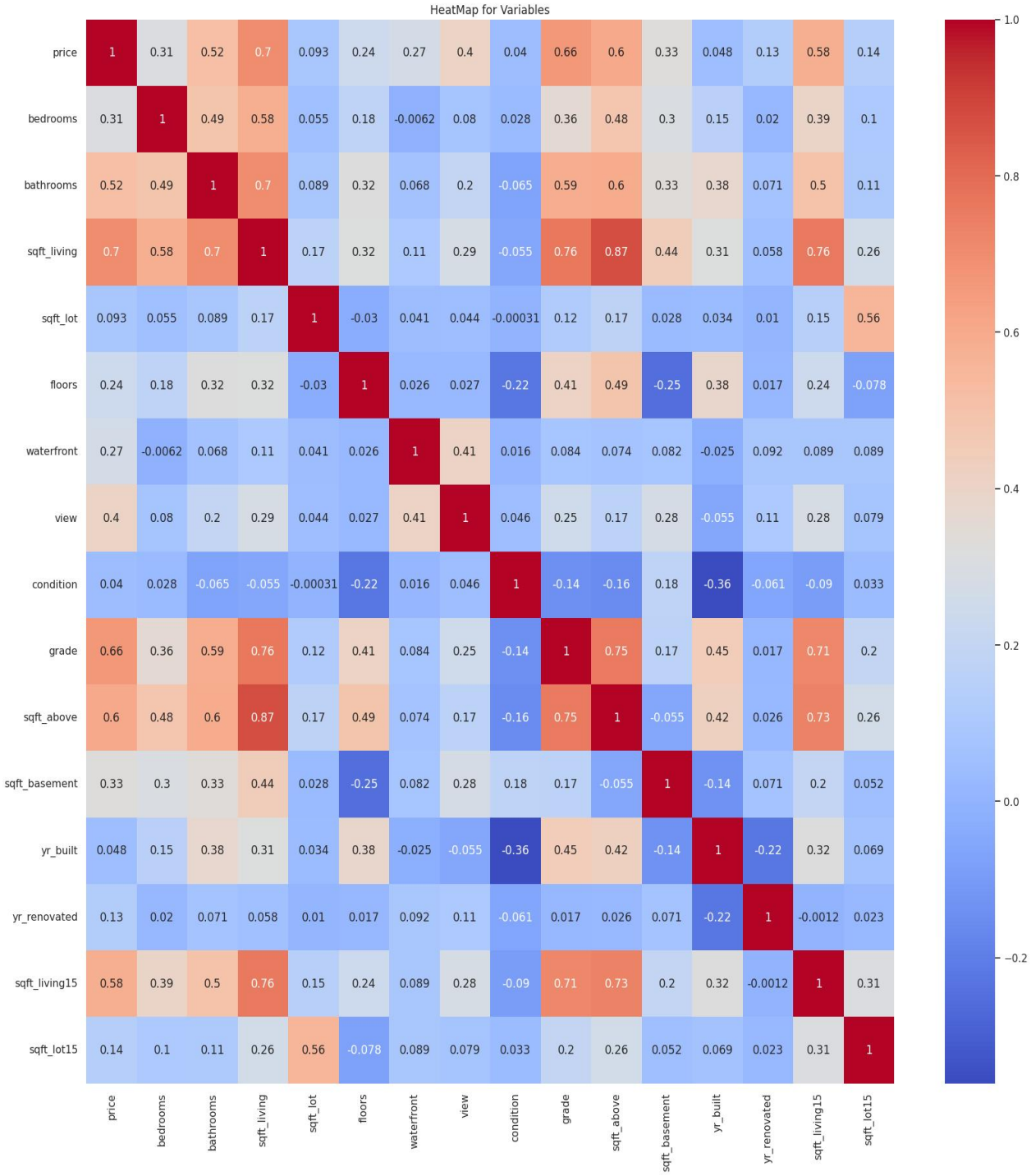

Figure 11 - Confusion Matrix

This plot helps us figure out the relationship between independent variables and the target variable. From this we can determine if the relationships are negative or positive. The closer the value is to 1 (one) the greater the relationship with the matching variable.

We can also use this to check for multicollinearity between independent variables.

For more multivariate analysis, we checked the relationship of the latitude and longitude variables relationship to the price variable using a scatter plot.

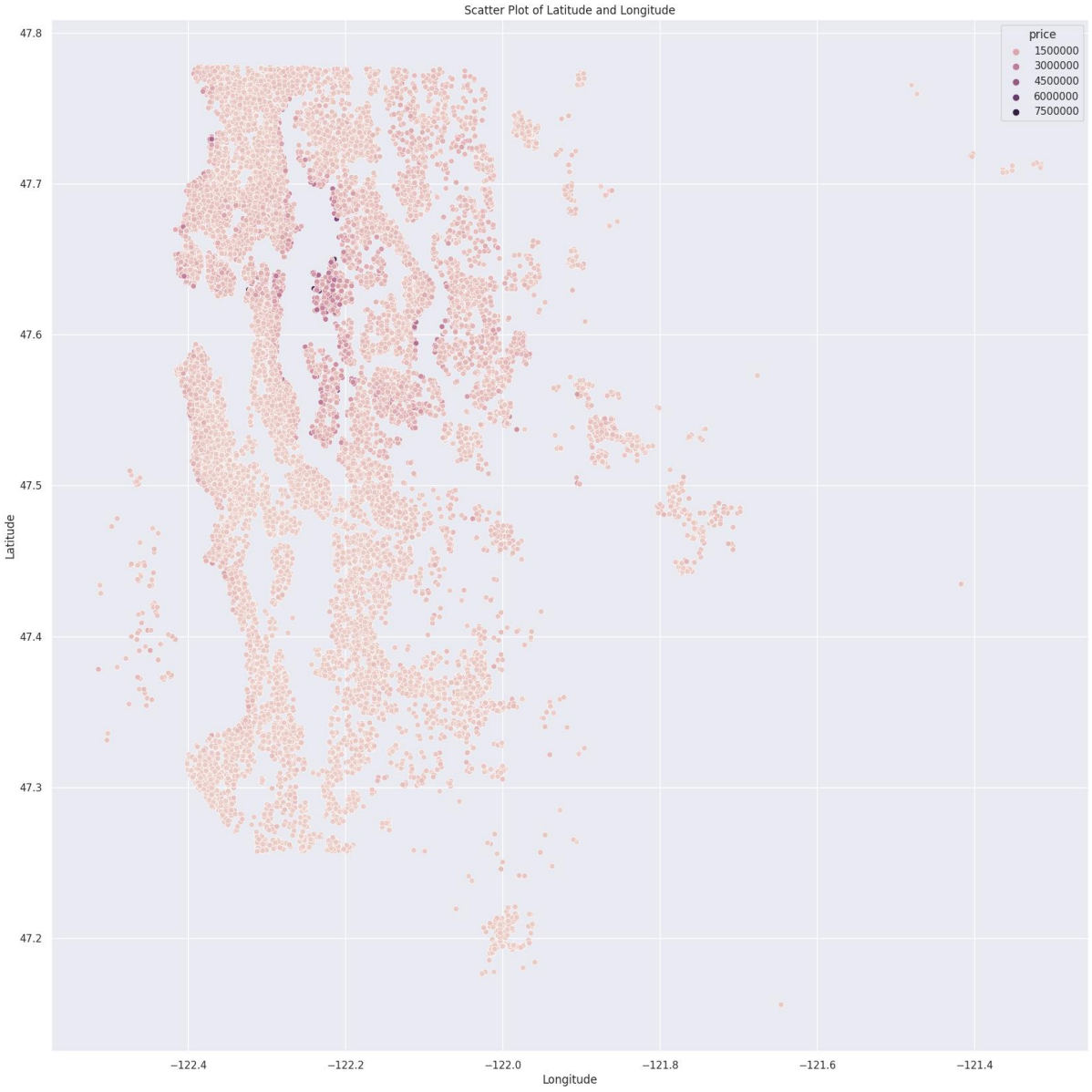

Figure 12 - Longitude by Latitude, with Price as hue

Observations: We see that properties increase in price as they get closer towards the capital of the state, which is Seattle with coordinates 47.6062, -122.3321. So, it is obvious location has a relationship with the price of the property.

III. FEATURE ENGINEERING

Feature engineering is the process of selecting, manipulating, and transforming raw data into features that can be used in supervised learning. In order to make machine learning work well on new tasks, it might be necessary to design and train better features. Feature engineering, in simple terms, is the act of converting raw observations into desired features using statistical or machine learning approaches (H. Patil, 2021).

We were able to create new features from existing ones that could be crucial to the performance of our model.

- Distance: because we found that location influences the price of the property, we created the distance feature by calculating proximity of the property to Seattle, implying the closer the property the higher the price.
- Age: We created a feature that tells us the age of the property, modern houses should cost more than older ones.
- Age renovated: We created a feature that tells us how long ago the property was last renovated.
- Renovated: We created a feature that indicates if a property has been renovated or not. 1 for yes and 0 for no.
- Total square feet: This feature is the addition of living area square feet and the lot area square feet.
- Year sold: This feature derived by taking the first four characters in the data column.

3. FEATURE SELECTION

One of the core concepts in machine learning, Feature Selection has a significant impact on how well your model performs. The performance you can get depends greatly on the data attributes you use to train your machine learning models(Shaikh, 2018).

The model's performance may be

adversely affected by irrelevant or only partially relevant features.

Here are a few benefits of performing Feature Selection on the data set.

- It reduces the chances of overfitting.
- It helps improve the accuracy of the model.
- Lesser training time

For this project we used Univariate Selection for our Feature Selection technique.

Univariate Selection: This technique performs statistical tests and selects features that have strong relationships with the target variable.

Univariate selection chooses features based on their relationships with the dependent (target) variable. It utilizes statistics tests like Chi-Square, correlation coefficients and ANOVA to score and rank features.

Before we performed Feature Selection we divided the data set in two, for target (y) and predictor variables (X).

Below is the code and the results from the Univariate Selection.

```
# calculate pearson coefficients
correlation_scores = X.corrwith(y)
#sort correlation scores
sort_scores = correlation_scores.abs().sort_values()
sort_scores
```

```
<ipython-input-39-b778bdd523ad>:2: FutureWarning: The
  correlation_scores = X.corrwith(y)
long               0.014039
id                 0.015183
condition          0.040022
age                0.048178
yr_built           0.048215
zipcode            0.051826
sqft_lot           0.093255
age_renovated      0.100947
total_sqft         0.116210
renovation_done    0.129202
yr_renovated       0.129547
sqft_lot15         0.141847
floors             0.241433
waterfront         0.269217
distance           0.299065
bedrooms           0.307339
lat                0.311285
sqft_basement      0.327539
view               0.403072
bathrooms          0.515440
sqft_living15      0.583589
sqft_above         0.601878
grade              0.664566
sqft_living        0.701508
```

Figure 13 - Univariate Selection Scores

This method of feature selection uses correlation coefficients of predictor variables with the target variable. From this we were able to select variables that would be helpful in the model building stage.

4. MODEL DEVELOPMENT AND EVALUATION

MODELLING

The modelling phase is where the train split of the dataset is used to train several machine learning models. The result of this phase is the predictions from the models we built.

Here are the features we decided to use for the final model.

- Bathrooms
- Bedrooms
- Sqft living
- Sqft lot
- Waterfront
- View
- Grade
- Latitude
- Longitude
- Sqft living 15
- Distance
- Year sold
- Age
- Age renovated
- Total square feet
- Renovation done

After scaling the features in the data set with min max scaler, we tried a range of different linear models to find which works best at predicting the price of the properties. After tuning hyperparameters, and using different regularisation techniques, none of the models were able

to pass an r square value of 62 percent, so we decided to use Polynomial regression to fit the data, which gave a better r square value of 82 percent.

EVALUATION

Evaluating the model's performance on a test set gives an indication about the model's expected performance on unseen real-world data. Although the model is scored in training, the best indicator of a model's performance is how well it predicts the label in the real-world scenarios. In this phase the trained models will be used to make predictions on the test set which contains 30% of the original dataset.

Evaluation Metrics

- R-square value: This is a statistical measure that indicates how much of the variation of a dependent variable is explained by an independent variable in a regression model(Fernando et al., 2023). An R-squared of 100% means that all movements of a security (or other dependent variable) are completely explained by movements in the index (or whatever independent variable you are interested in).

MODEL	R-SQUARE
Linear Regression	62%
Linear Regression with Stochastic Gradient Descent	62%
Linear Regression with Stochastic Gradient Descent and Lasso Regularisation	62%
Linear Regression with Stochastic Gradient Descent and Ridge Regularisation	61%
Polynomial Regression	80%

- Mean Square Error: The average squared difference between the estimated values and what is estimated is measured by the mean squared error (MSE) of an estimator, which is used in statistics. The randomness or the estimator's failure to take into consideration data that could lead to a more accurate are the reasons why MSE is nearly never precisely positive (Binieli Moshe, 2018).

MODEL	**MSE**
Linear Regression	37402537252.03
Linear Regression with Stochastic Gradient Descent	37378582988.67
Linear Regression with Stochastic Gradient Descent and Lasso Regularisation	37520796328.53
Linear Regression with Stochastic Gradient Descent and Ridge Regularisation	37538466621.44
Polynomial Regression	22584924698.85

Below is the result from the models we built.

We added figures to show how well each model was able to fit the test data.

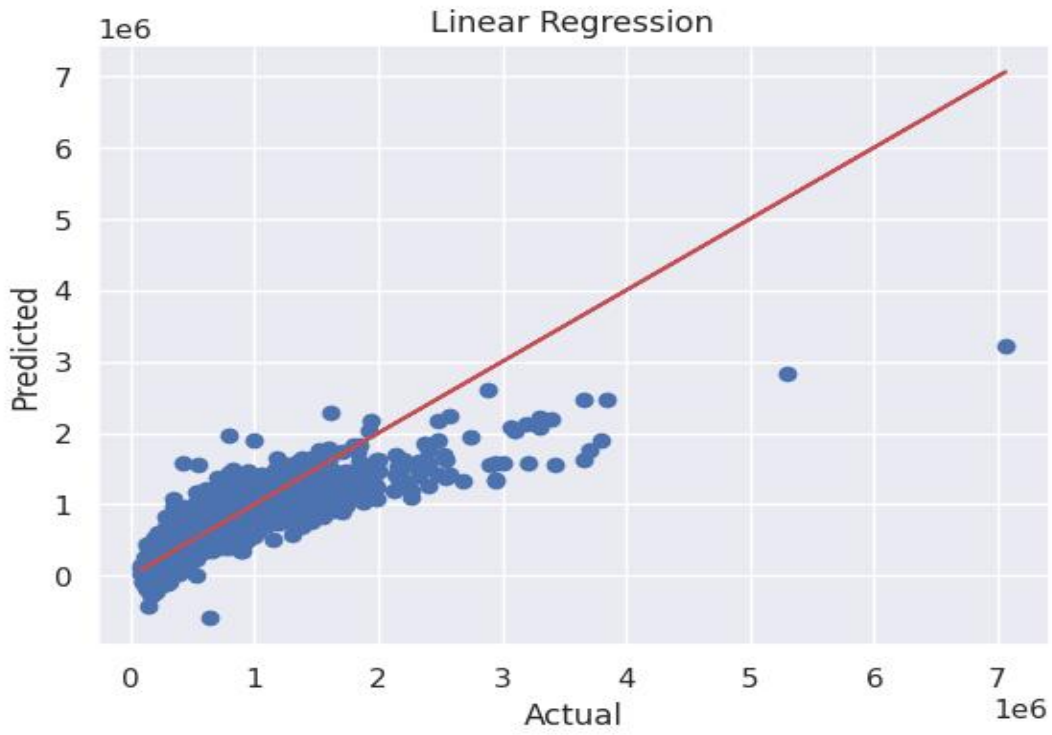

Figure 14 - Linear Regression Model Test Performance

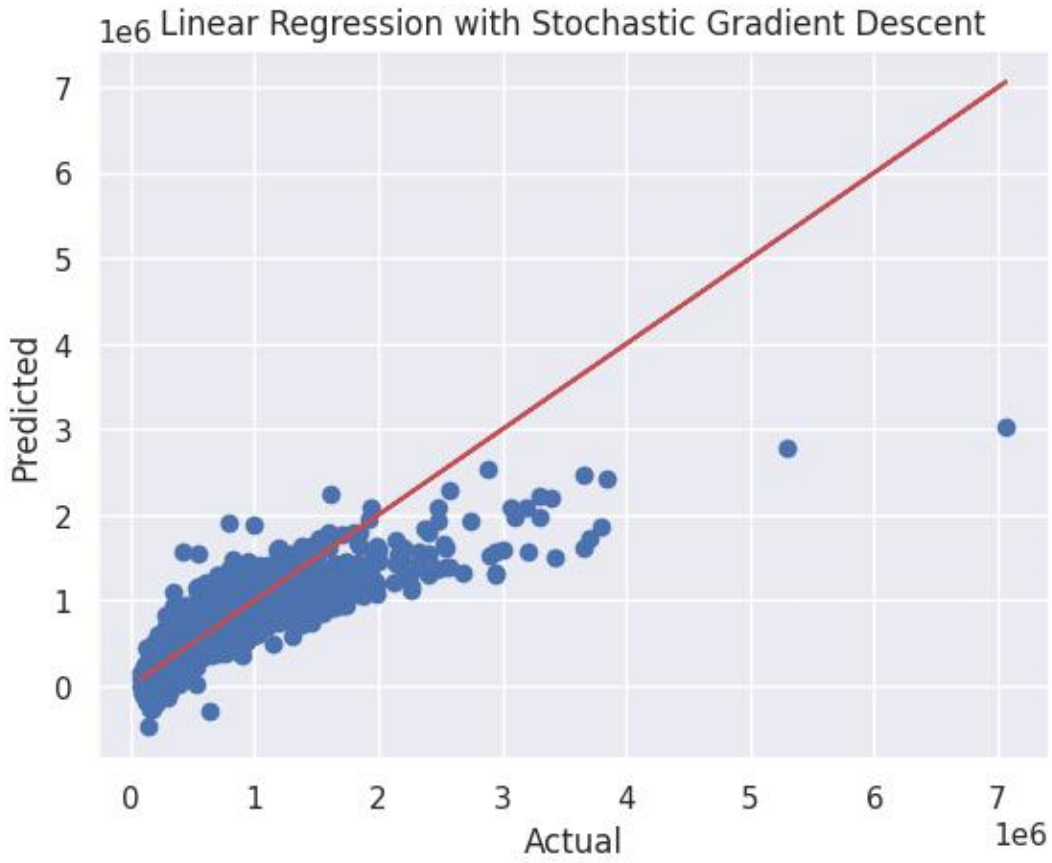

Figure 15 - Linear Regression with Stochastic Gradient Descent Model Test Performance

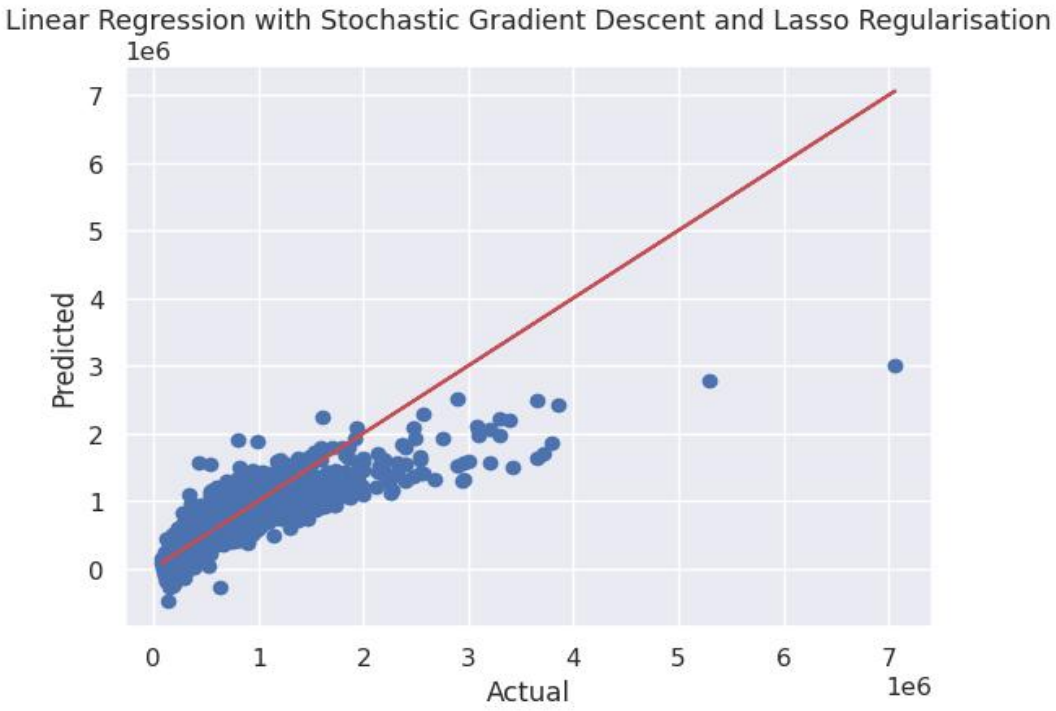

Figure 16 - Linear Regression with Stochastic Gradient Descent and L1 Regularisation

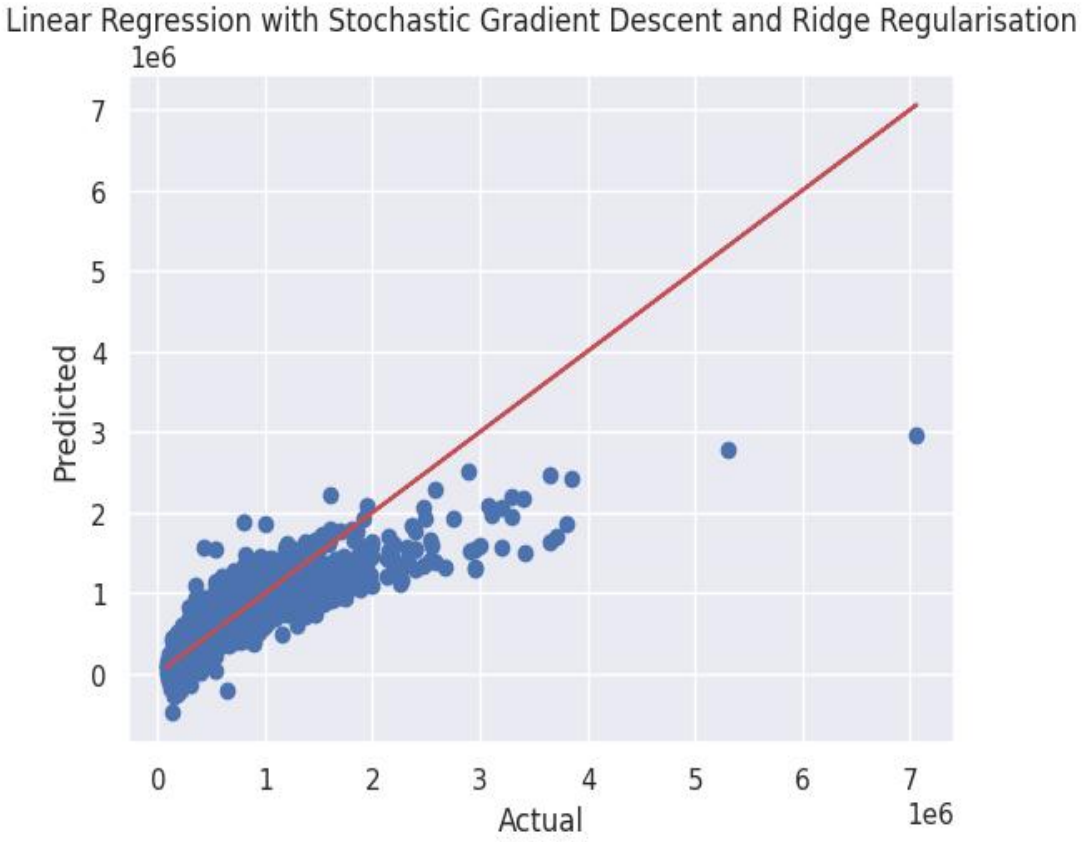

Figure 17 - Linear Regression with Stochastic Gradient Descent and L2 Regularisation

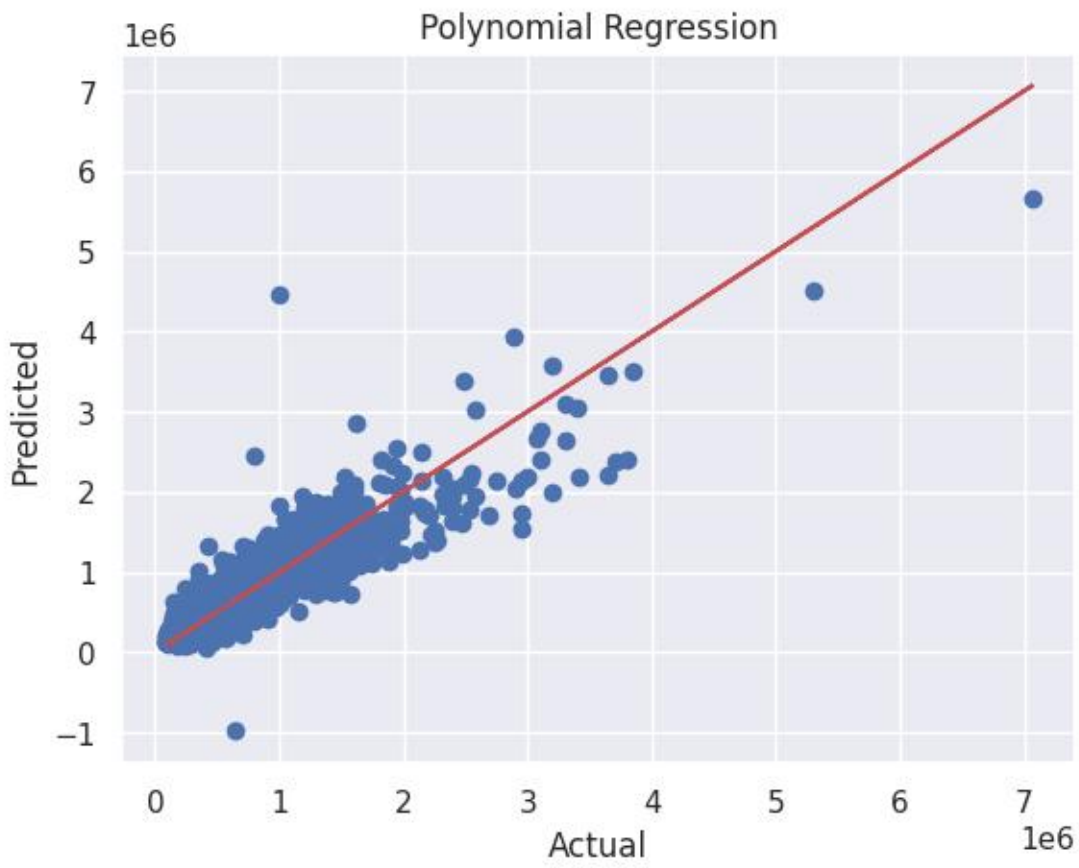

Figure 18 - Polynomial Regression Model Test Performance

The polynomial regression model fit the line more than the other models did, with an R-Square value of 80 percent. This is the model we would advise Chants to use because it meets the objectives better than the other models would.

5. MODEL COMPARISON

In this project, we aimed to develop a machine learning model to accurately predict house prices. We experimented with various regression models and evaluated their performance. The models considered include linear regression, linear regression with gradient descent, linear regression with gradient descent and lasso regularization, linear regression with gradient descent and ridge regularization, and polynomial regression. Among these models, the polynomial regression model demonstrated the best performance.

In this project, the goal was to develop a machine learning model that can correctly predict house prices. We modelled the data set with various regression models and evaluated the performance of each model. The models we used include, linear regression, linear regression with stochastic gradient descent, linear regression with stochastic gradient descent and lasso regularisation, linear regression with stochastic gradient descent and ridge regularisation, and polynomial regression. Among these models, the polynomial regression model had the best performance.

MODEL DESCRIPTIONS

1. Linear Regression:
    - Linear regression is a basic regression model that creates a linear relationship between the selected features and the target variable.
    - Linear regression does not consider non-linearities.
2. Linear Regression with Stochastic Gradient Descent:
    - Linear regression with stochastic gradient descent is an optimization algorithm that uses a single random sample from the data set to calculate the gradient in each iteration, and updates the parameters based on the gradient of the sample.
    - It is more efficient that batch gradient descent because it only requires a single sample in each iteration rather than the entire data set.

3. Linear Regression with Gradient Descent and Lasso Regularization:
    - Lasso regularization is a technique that adds a penalty term to the linear regression cost function based on the absolute value of the weights.
    - It helps in feature selection by shrinking less important features towards zero.
4. Linear Regression with Gradient Descent and Ridge Regularization:
    - Ridge regularization is similar to lasso regularization but uses a different penalty based on the square of the weights.
    - It can be effective when all features are potentially relevant.
5. Polynomial Regression:
    - Polynomial regression extends linear regression by incorporating polynomial features, allowing for non-linear relationships between the features and the target variable.
    - It captures more complex patterns and can fit curved decision boundaries.

MODEL PERFORMANCE

To evaluate the performance of the models, we used R-squared (R^2) coefficient as our metric. The polynomial regression model outperformed the other models on based on this metrics.

COMPARING RESULTS

The R^2 coefficient of the polynomial regression model was highest, indicating a better fit to the data and explaining more variance in house prices.

CONCLUSION

Based on the evaluation results, the polynomial regression model demonstrated superior performance in predicting house prices compared to the other models considered. Its ability to capture non-linear relationships through the inclusion of polynomial features contributed to

its improved accuracy and predictive power. The findings suggest that incorporating higher-order feature interactions can significantly enhance the predictive capabilities of the model. It is important to note that the selection of the best model may depend on the specific dataset, available features, and problem context. Further experimentation and validation on different datasets are recommended to confirm the generalizability of the findings.

REFERENCES

Beers, B., Potters, C., & Schmitt, K. R. (2023, March 31). *What is Regression? Definition, Calculation, and Example*. Investopedia. https://www.investopedia.com/terms/r/regression.asp

IBM. (2017). *House Sales in King County, USA | Kaggle*. Kaggle. https://www.kaggle.com/datasets/harlfoxem/housesalesprediction?utm_medium=Exinfluencer&utm_source=Exinfluencer&utm_content=000026UJ&utm_term=10006555&utm_id=NA-SkillsNetwork-wwwcourseraorg-SkillsNetworkCoursesIBMDeveloperSkillsNetworkDA0101ENSkillsNetwork20235326-2022-01-01

Kwiatkowski, R. (2021, May 22). *Gradient Descent Algorithm — a deep dive | by Robert Kwiatkowski | Towards Data Science*. Medium. https://towardsdatascience.com/gradient-descent-algorithm-a-deep-dive-cf04e8115f21

Mahmoodinobar, F. (n.d.). *Multivariate Analysis — Going Beyond One Variable At A Time*. Towards Data Science. Retrieved June 2, 2023, from https://towardsdatascience.com/multivariate-analysis-going-beyond-one-variable-at-a-time-5d341bd4daca

Patil, H. (2021). *What is Feature Engineering — Importance, Tools and Techniques for Machine Learning*. Towards Data Science. https://towardsdatascience.com/what-is-feature-engineering-importance-tools-and-techniques-for-machine-learning-2080b0269f10

Patil, P. (2018, March 23). *What is Exploratory Data Analysis? | by Prasad Patil | Towards Data Science*. Towards Data Science. https://towardsdatascience.com/exploratory-data-analysis-8fc1cb20fd15

Shaikh, R. (2018, October 28). *Feature Selection Techniques in Machine Learning with Python*. Towards Data Science. https://towardsdatascience.com/feature-selection-techniques-in-machine-learning-with-python-f24e7da3f36e

Binieli Moshe. (2018, October 16). *Machine learning: an introduction to mean squared error and regression lines*. Free Code Camp. https://medium.com/free-code-camp/machine-learning-mean-squared-error-regression-line-c7dde9a26b93

Fernando, J., Smith, A., & Perez, Y. (2023). *R-Squared: Definition, Calculation Formula, Uses, and Limitations*. Investopedia. https://www.investopedia.com/terms/r/r-squared.asp