جامعة أم القرى
UMM AL-QURA UNIVERSITY

Data Analysis 1

COURSE INSTRUCTOR
(DR. Idrees Alsolbi)

SUBMITTED BY:
Mohammed Nedal Alshareef
Raed Falah Alhelali
Ziyad omar altalhi
Eyad mohammed alharthi
Aseel Abdulfattah Almaghrabi

STUDENT ID:
444004824
444006561
444000039
444000005
444004296

DEPARTMENT OF (Data Science)
COLLEGE OF COMPUTING
UMM AL-QURA UNIVERSITY

TABLE OF CONTENT

Table of Contents

TASKS DESCRIPTIONS

TASK 1: Data Acquisition and Preparation (Week 3):

- **URL:** https://www.kaggle.com/datasets/harlfoxem/housesalesprediction

- **About Dataset:**

The dataset provides a comprehensive overview of house sale prices within King County, encompassing the vibrant city of Seattle and its surrounding areas. Spanning transactions occurring from May 2014 to May 2015, it serves as a valuable resource for analysing real estate trends and market dynamics during this period. With a diverse range of property types and neighbourhoods represented, from bustling urban districts to serene suburban enclaves, the dataset offers insights into the multifaceted nature of the local housing market. Researchers, analysts, and industry professionals can leverage this rich source of information to gain a deeper understanding of factors influencing home prices, such as location, property size, condition, and economic indicators. Moreover, policymakers and stakeholders may utilize the dataset to inform strategic decisions aimed at promoting sustainable growth and equitable access to housing opportunities across King County.

Important features

Key features included in the dataset provide valuable information about various aspects of the properties:

1) ***price:*** *The sale price of the property, serving as the primary target variable for predictive modeling and analysis.*
2) ***bedrooms:*** *The number of bedrooms in the property, a crucial factor influencing its market value and suitability for different buyers' needs.*
3) ***bathrooms:*** *The number of bathrooms in the property, another essential amenity affecting its desirability and pricing.*
4) ***sqft_living:*** *The total square footage of the living space within the property, indicating its size and functional living area.*
5) ***sqft_lot:*** *The total land area (in square feet) of the property, providing insight into the size of the plot and potential for outdoor amenities or development.*
6) ***floors:*** *The number of floors in the property, influencing its architectural style, layout, and overall appeal.*
7) ***waterfront:*** *A binary indicator (0 or 1) denoting whether the property has a waterfront view, a premium feature often associated with higher property values.*
8) ***view:*** *An index representing the quality of the view from the property, which can significantly impact its market value.*
9) ***condition:*** *An index representing the overall condition of the property, reflecting its maintenance level and structural integrity.*
10) ***grade:*** *An index representing the overall grade assigned to the property, based on factors such as construction quality, design, and finishing materials.*
11) ***yr_built:*** *The year the property was built, offering historical context and insight into its age, construction methods, and architectural style.*

Questions to be answered:

1. What will the average prices be in the coming years?
2. What is the average land area sold in the upcoming years?
3. What factors influence the price of a house in King County, Seattle, and to what extent?
4. Are there distinct groups or clusters of properties with similar characteristics within King County, Seattle?

TASK 2: Exploratory Data Analysis (EDA) (Week 4):

What we use to Exploratory and cleaning the dataset.

We have Exploratory our dataset by using these functions:

.**shape** : we use this function To display the number of rows and columns

describe(): we use this function To display some basic statistics (Mean, STD, MAX ,etc)

Head() and Tail(): we use this two functions To display the some of the first and last rows

.info: we use this function to display summary of the dataset's structure, including the data types of each column, non-null counts, and memory usage

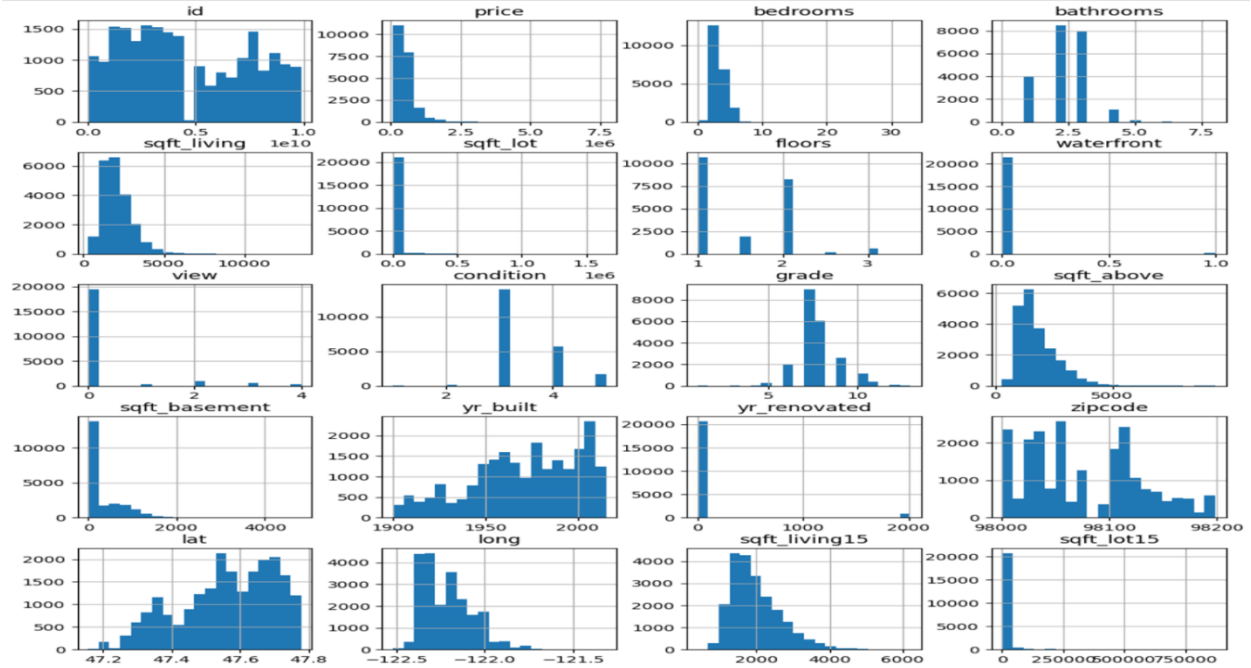
.Columns: we use this function to display list of column names present in the dataset

.dtypes: we use this function to display the data types of each column in the dataset.

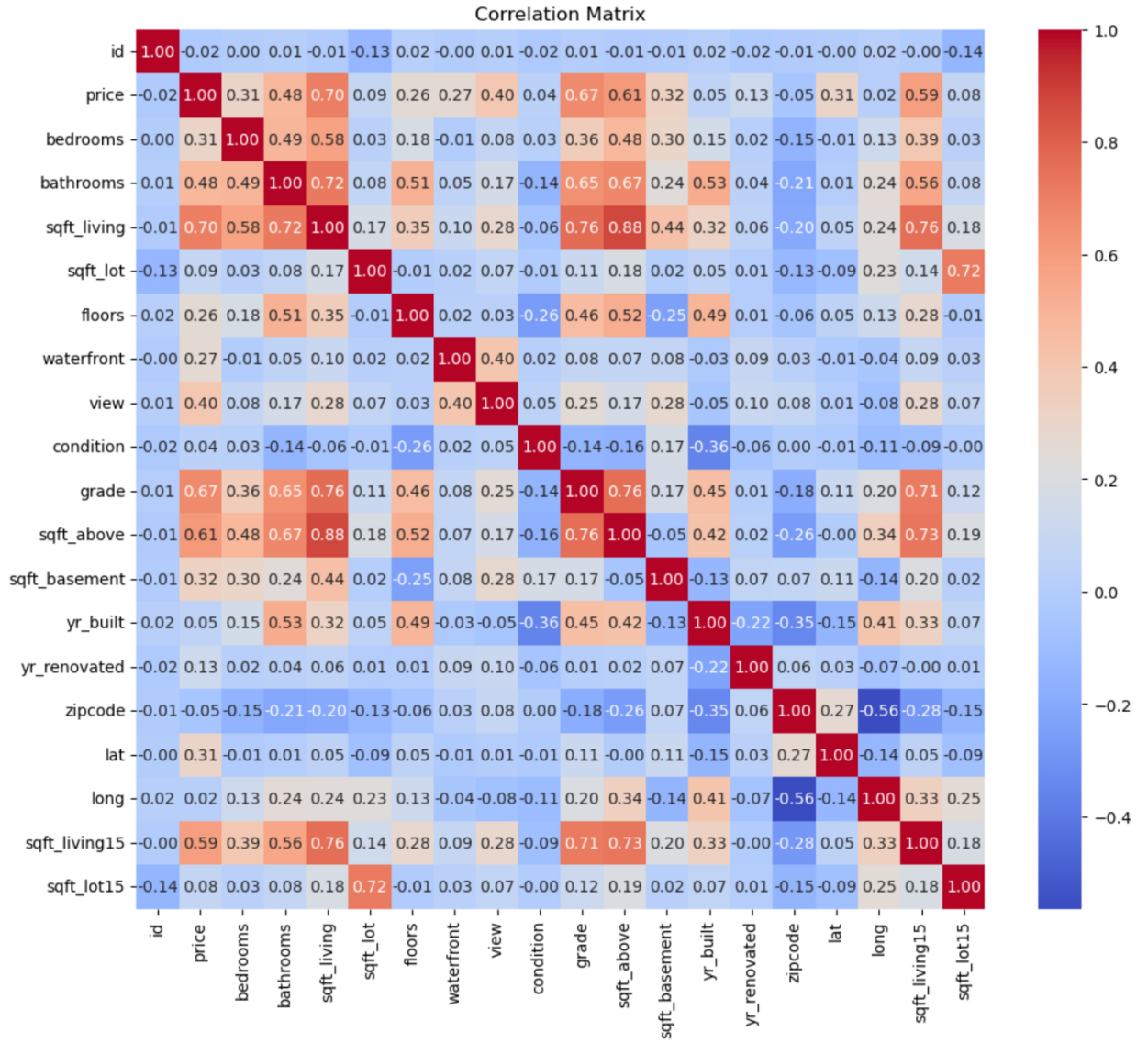
We have cleaned our dataset by using these functions:

Isnull().sum(): we used this function to calculate how many missing values in each features.

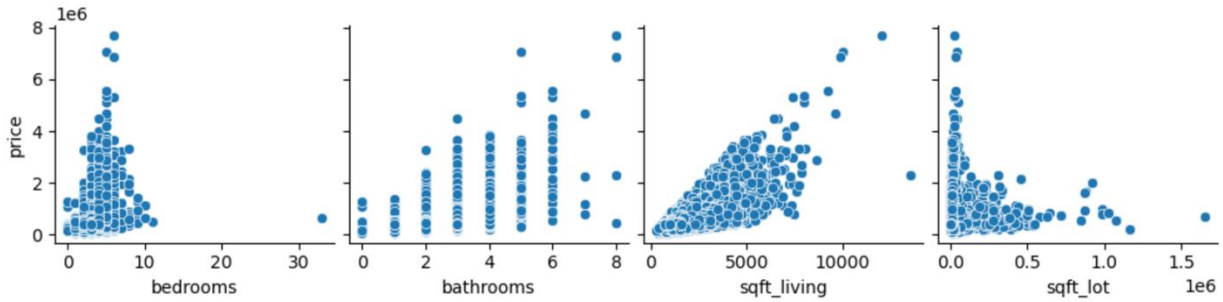
Drop(): we used this function to delete some rows.



The purpose of using histograms to visualize feature distributions is to gain insights into the spread and concentration of data points within each feature. Histograms provide a visual representation of the frequency distribution of values, showing how often certain values occur within the dataset. By examining these distributions, analysts can identify patterns, trends, and potential outliers in the data, which can inform further analysis and modeling decisions.

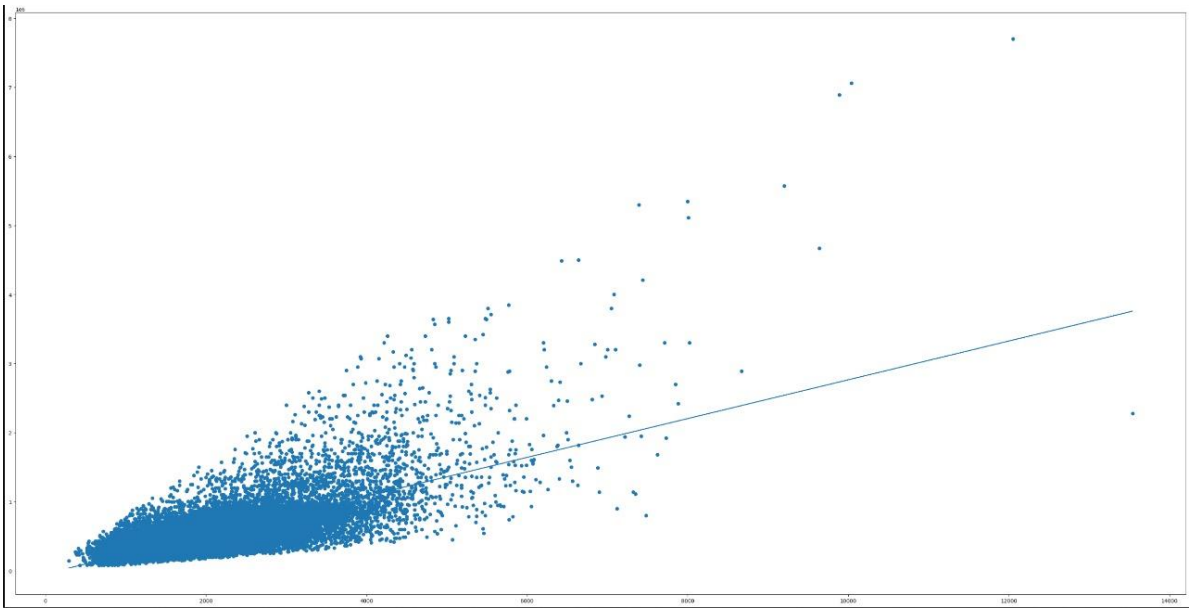


This graph expresses the strength of the relationship between each attribute, and when we approach the number 1 and the colour red, the relationship between the attribute becomes stronger, also This graph helps us in choosing the strongest relationships to use in linear regression.



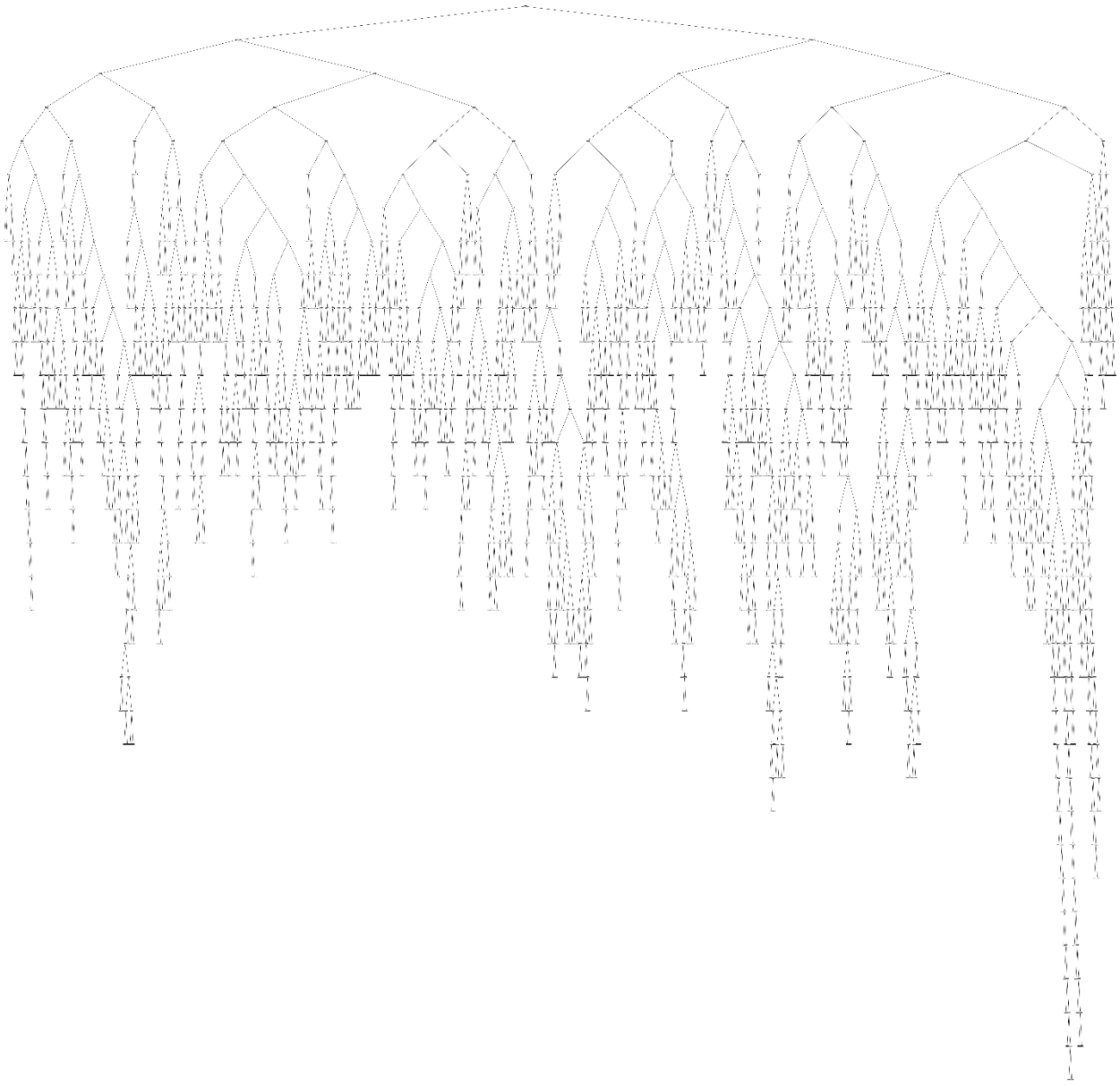
This scatter plots between the target variable (price) and selected predictor variables (bedrooms, bathrooms, sqft_living, sqft_lot) from the dataset. The purpose is to visually explore the relationship between house prices and these specific features. Scatter plots allow us to observe any potential patterns or trends in the data, such as linear or non-linear relationships, correlations, and outliers. This visualization helps in understanding how changes in predictor variables correspond to changes in house prices, aiding in model interpretation and decision-making in predictive modeling tasks.

TASK 3: Linear and Nonlinear Regression (Week 5):

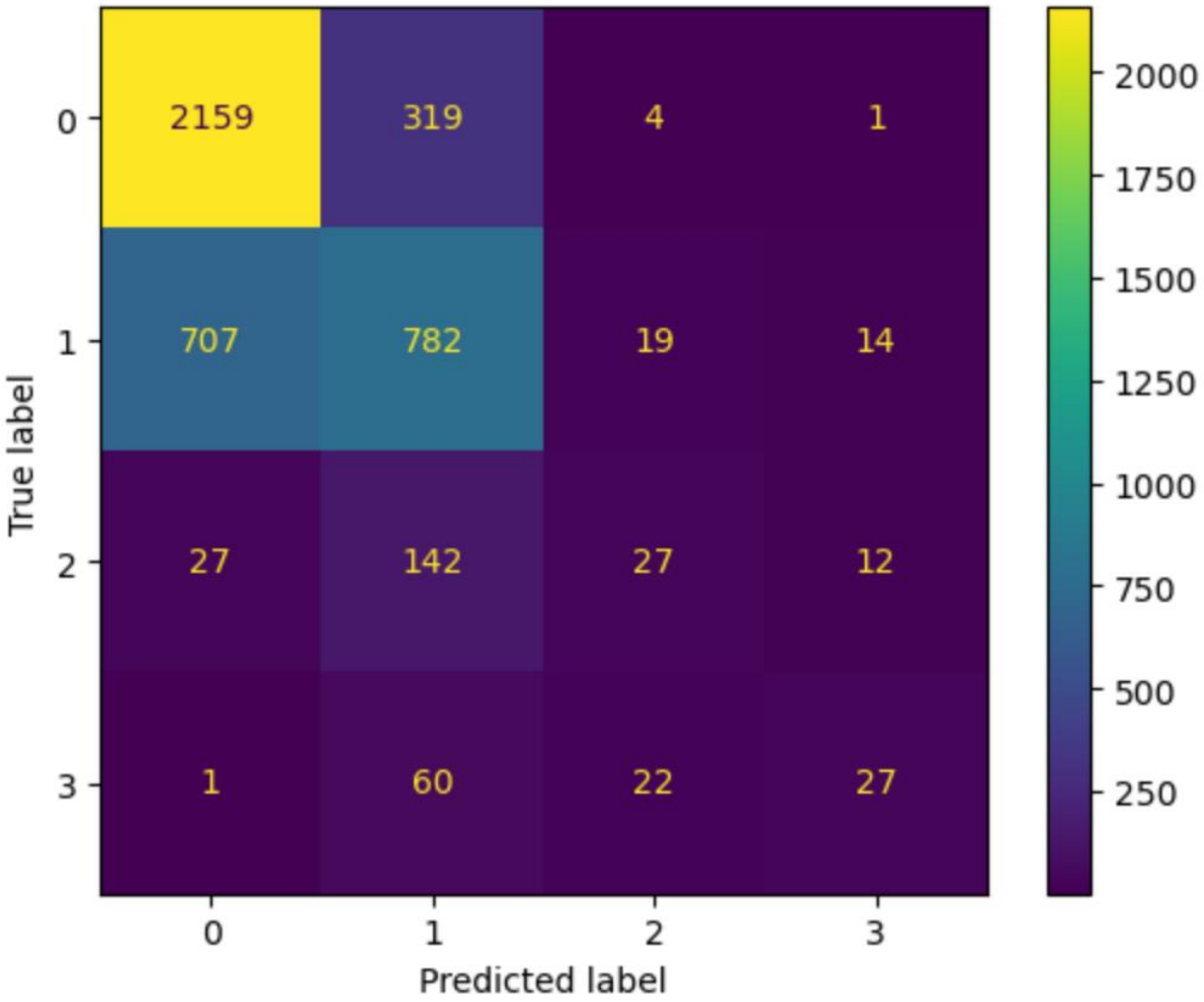


Linear regression is a fundamental statistical method used in data science and machine learning. Its goal is to find the linear relationship between a dependent variable (the variable we want to predict or explain) and one or more independent variables (the variables used to predict the dependent variable).
Based on linear regression analysis, there is a linear relationship between living area (square feet) and property price. As the living area per square foot increases or decreases widely, the price of the property tends to rise or fall as well.

TASK 4: Logistic Regression and Classification (Week 6):

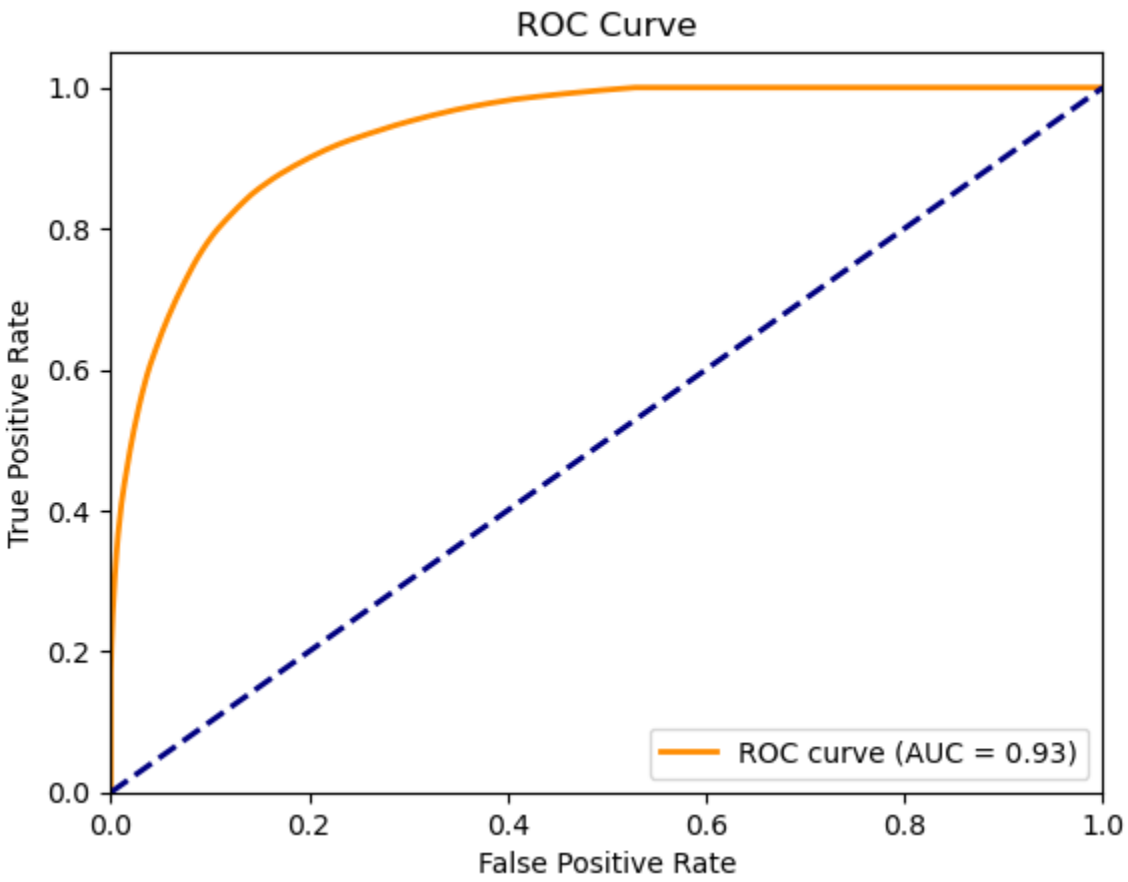


"This decision tree visualizes the relationship between the square footage of living area ('sqft_living') and the price of the properties. Each node in the tree represents a decision based on the value of 'sqft_living', leading to different price outcomes. The size of the tree reflects the complexity of the decision-making process involved in predicting property prices based on living area."



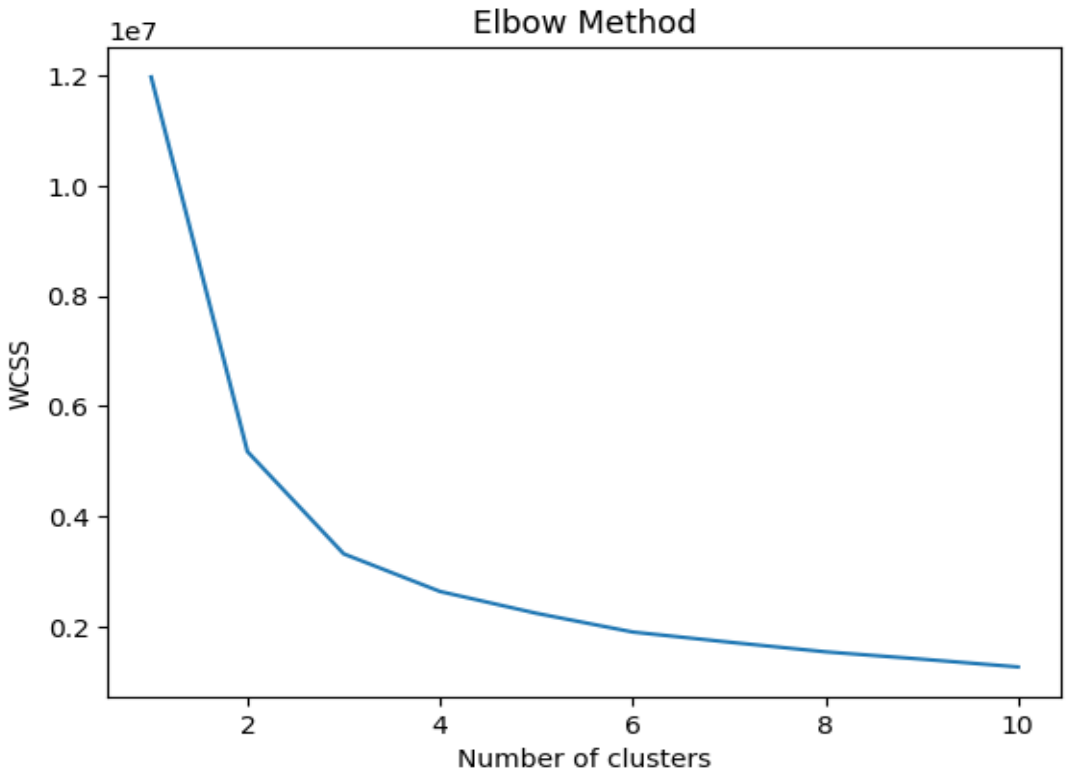
"This confusion matrix illustrates the model's performance in classifying across different price categories. Each row in the matrix represents the true category, while each column represents the predicted category by the model. This plot can be used

to determine whether the model is classifying the data correctly or not across the test set."

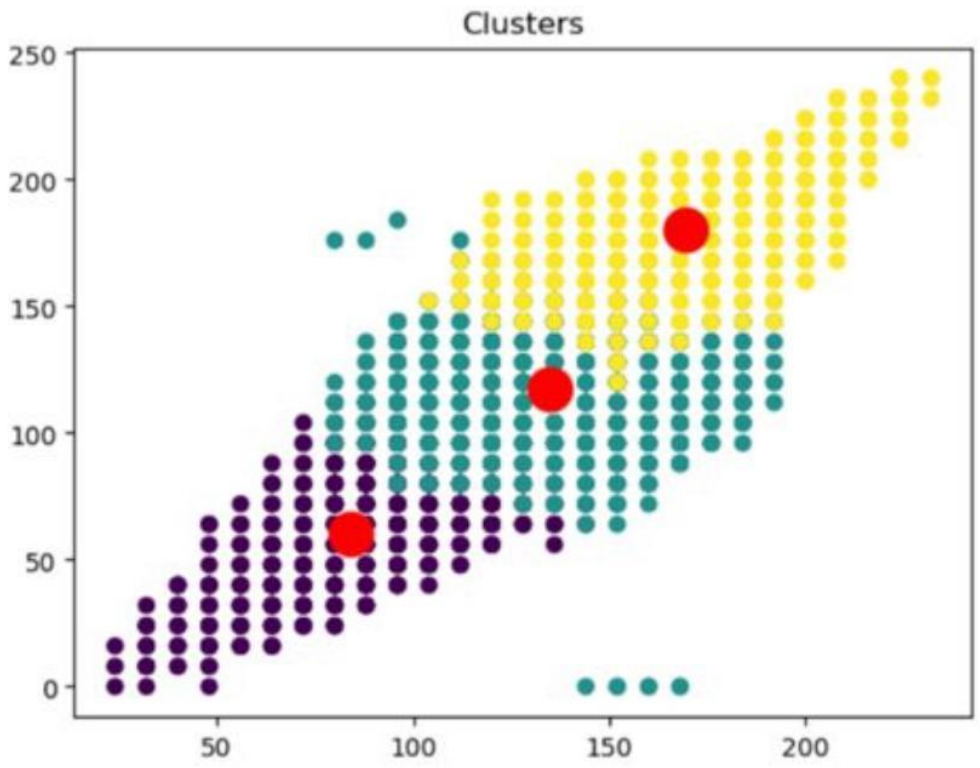


"The Receiver Operating Characteristic (ROC) curve illustrates the model's performance in classifying properties with prices exceeding one million dollars. The AUC value is X, indicating the model's quality in distinguishing between positive and negative"

TASK 5: Clustering (Week 7)

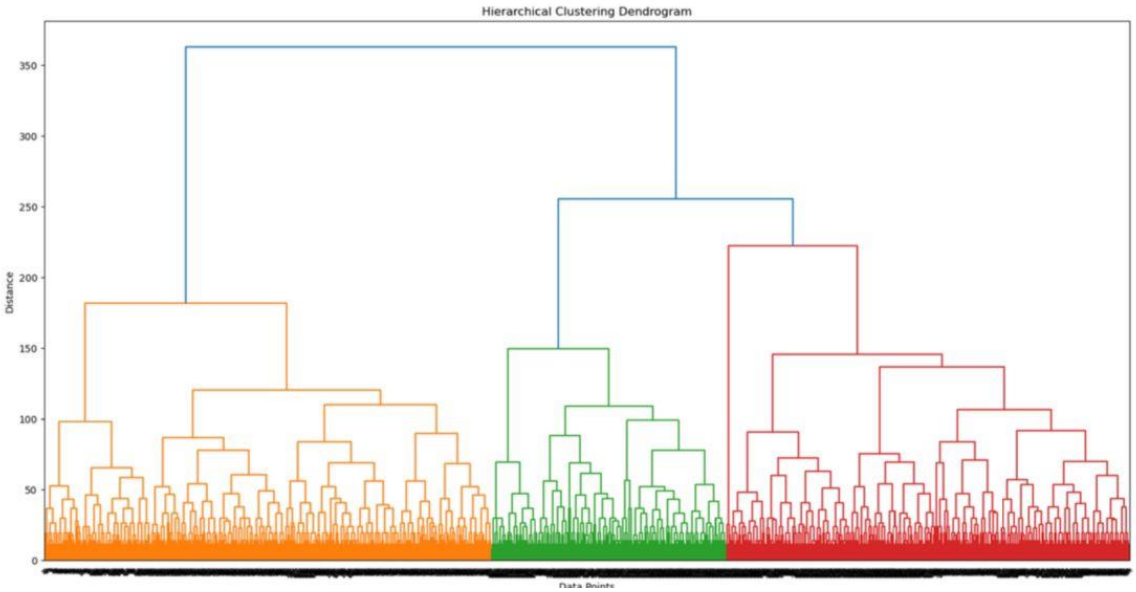


Elbow Method for Optimal Cluster Selection: Plot demonstrating the within-cluster sum of squares (WCSS) as a function of the number of clusters. The elbow point indicates an optimal number of clusters for K-means clustering.



Cluster Visualization: Scatter plot illustrating the clusters generated by K-means clustering algorithm. Each point represents a data instance colored by its assigned cluster, with centroids highlighted in red.

Sum of Squared Error (SSE): 3318367.90223293

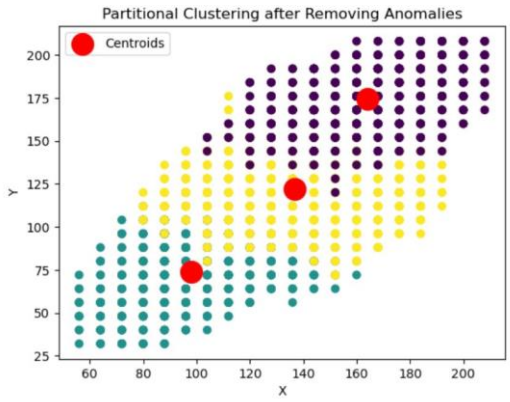


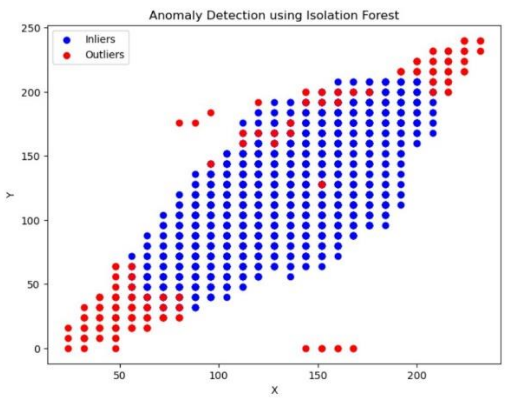
Hierarchical clustering groups similar items together in a tree-like structure based on their similarity. It doesn't require predefining the number of clusters and is used to discover patterns in various fields. Dendrograms visualize the cluster hierarchy, Each tree represents a type of data in a specific order.

Sum of Squared Error (SSE):

```
SSE for test set with Euclidean distance: 24    242215.333776
0     216348.602204
40    205260.942476
dtype: float64
```

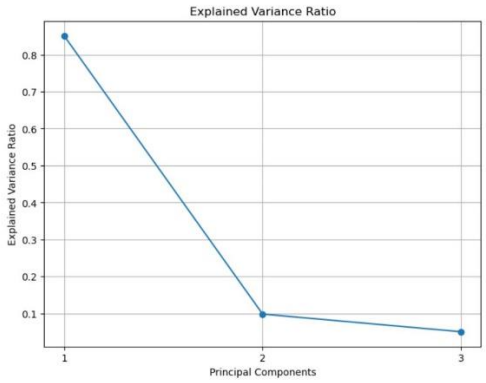
TASK 6: Anomaly Detection (Week 8):

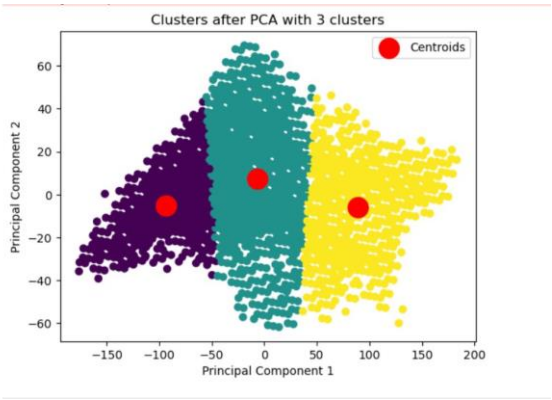




Anomaly detection is finding unusual patterns in data. It helps to identify outliers or distortions that deviate from the norm. In this task, we used it well and were able to detect and recognize outliers, as shown in the graphics

TASK 7: Principal Component Analysis (PCA) (Week 9):





PCA (Principal Component Analysis) is a technique used to reduce the dimensionality of data while preserving important information. It helps in finding the most important patterns in data by extracting key features. In this task, we used PCA to reduce the dimensions available in the data set, and we modified it and reduced the dimensions. This process facilitates the process of understanding the data set and making a specific decision.

TASK 8: Data Summarization and Visualization (Week 10):

- In our project, we embarked on a comprehensive analysis of house sale prices within King County, encapsulating the bustling city of Seattle and its neighboring regions. The dataset, covering transactions from May 2014 to May 2015, emerged as a crucial resource for understanding real estate dynamics during this timeframe. Offering a diverse array of property types and neighborhood landscapes, ranging from urban cores to suburban retreats, the dataset illuminated the intricate tapestry of the local housing market.

- Delving into the labeled data, we uncovered a treasure trove of insights free from missing values. Through extensive exploratory data analysis (EDA), we revealed potent relationships using techniques like heatmaps and regression. Notable discoveries included robust correlations such as price with square footage of living space and price with square footage of living space above ground. Our foray into classification culminated in the creation of a sophisticated decision tree, boasting an impressive area under the ROC curve (AUC) of 0.93.

- Transitioning to unlabeled data, our journey continued with an exploration of clustering methodologies. Utilizing the elbow method, we discerned that partitioning our data into three clusters proved optimal. This insight propelled us into partitional clustering with Euclidean distance and hierarchical clustering, unraveling distinct groupings within our dataset.

- Furthermore, we delved into anomaly detection and principal component analysis (PCA) to uncover outliers and reveal underlying patterns in the data. Notably, in regression analysis, we identified a significant positive correlation between house prices and square footage of living space. By utilizing the regression model with x = data['sqft_living'] and y =

data['price'], we achieved a predictive model with an R-squared value of 0.69. This finding underscores the importance of square footage in determining house prices, offering valuable insights for stakeholders in the real estate market.

- Overall, our analyses have provided invaluable insights for researchers, industry professionals, and policymakers alike. These findings stand poised to inform strategic decisions aimed at fostering sustainable growth and equitable access to housing opportunities across King County

THE EXPERIENCES AND SKILLS ACQUIRED BY THE TEAM MEMBERS

Student Name	Experiences	Skills

CONCLUSION

REFERENCES