Data Analysis (1)

Presentation of Data Analysis Project

Outline

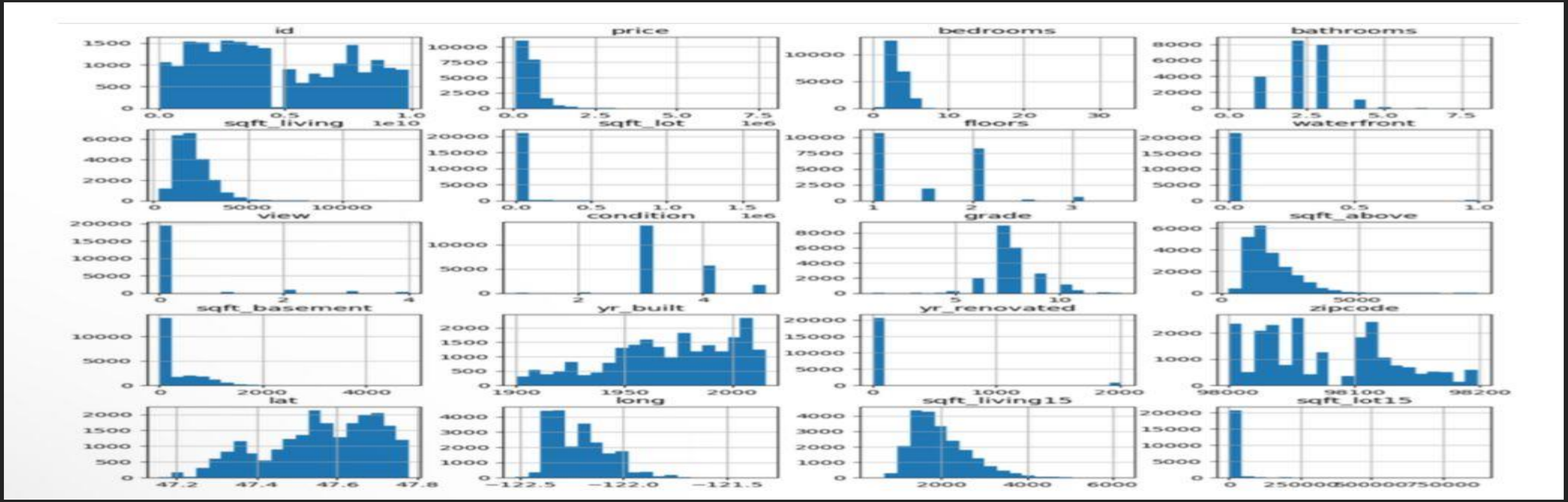
- *Introduction and Exploratory about Dataset*
- *Linear Regression model*
- *Classification by (Decision tree)*
- *Clustering*
- *Anomaly Detection*
- *PCA*

Introduction and Exploratory about Dataset

- First, the data set was chosen, and the choice fell on a data set related to home selling prices in King County. The data covers a period from May 2014 to May 2015. After the selection, the data was formatted, missing values were checked, and some methods were applied that make the data ready to build the model.

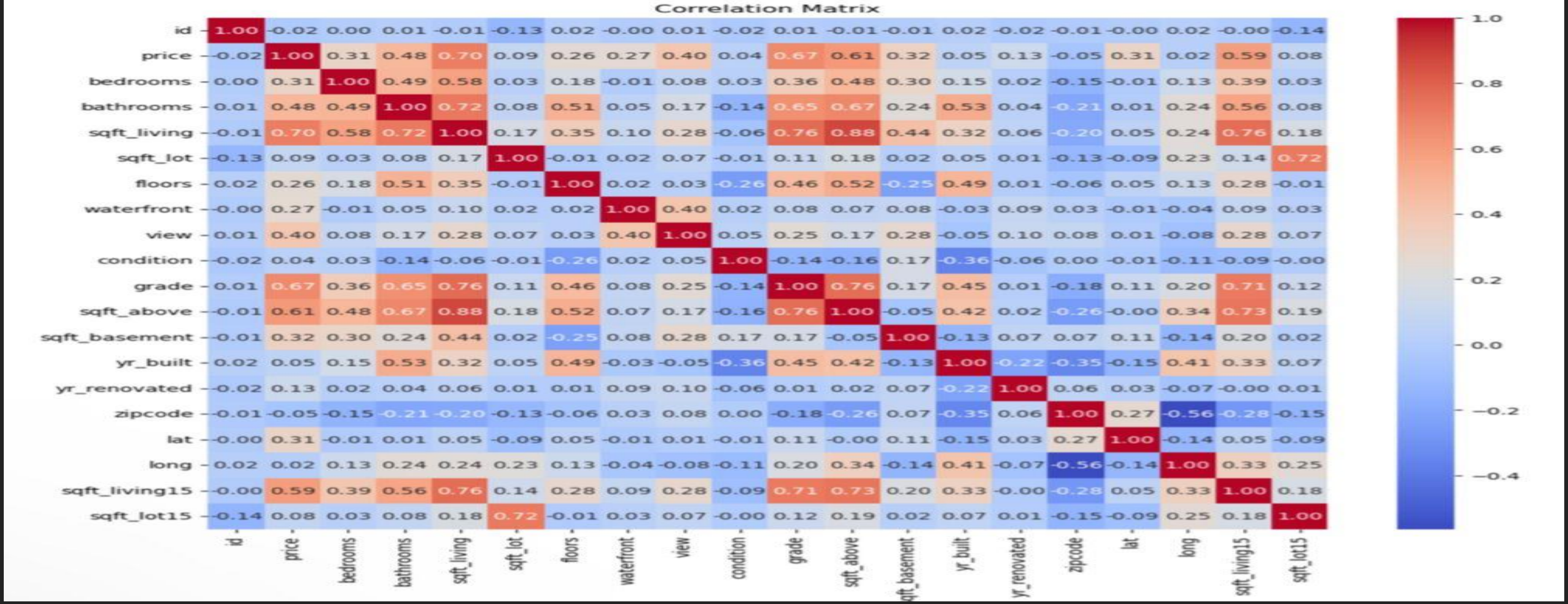
- there are some important features: Price, bedrooms, bathrooms, sqft_living, floors, etc…

Exploratory about Dataset



The histogram has been applied to all of our features in the dataset to make it easier to build the model so that we can better understand the distribution of the data and identify any potential issues or outliers that may need to be addressed before training the model.

Heatmap of our Dataset



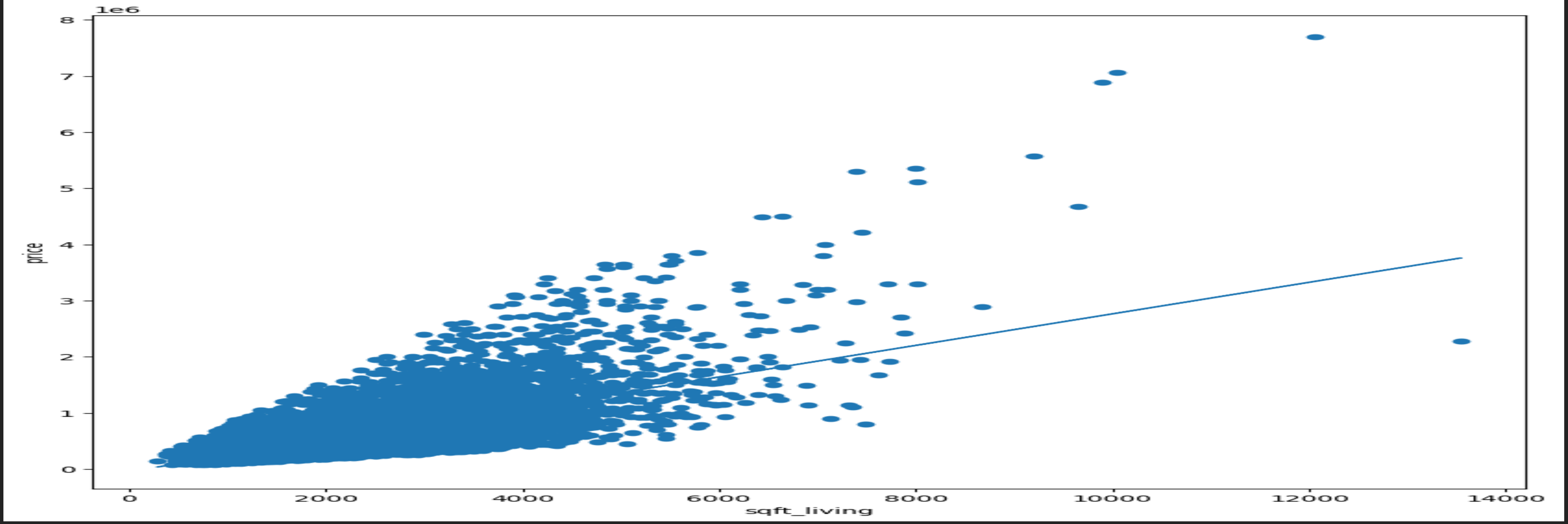
	id	price	bedrooms	bathrooms	sqft_living	sqft_lot	floors	waterfront	view	condition	grade	sqft_above	sqft_basement	yr_built	yr_renovated	zipcode	lat	long	sqft_living15	sqft_lot15
id	1.00	-0.02	0.00	0.01	-0.01	-0.13	0.02	-0.00	0.01	-0.02	0.01	-0.01	-0.01	0.02	-0.02	-0.01	-0.00	0.02	-0.00	-0.14
price	-0.02	1.00	0.31	0.48	0.70	0.09	0.26	0.27	0.40	0.04	0.67	0.61	0.32	0.05	0.13	-0.05	0.31	0.02	0.59	0.08
bedrooms	0.00	0.31	1.00	0.49	0.58	0.03	0.18	-0.01	0.08	0.03	0.36	0.48	0.30	0.15	0.02	-0.15	-0.01	0.13	0.39	0.03
bathrooms	0.01	0.48	0.49	1.00	0.72	0.08	0.51	0.05	0.17	-0.14	0.65	0.67	0.24	0.53	0.04	-0.21	0.01	0.24	0.56	0.08
sqft_living	-0.01	0.70	0.58	0.72	1.00	0.17	0.35	0.10	0.28	-0.06	0.76	0.88	0.44	0.32	0.06	-0.20	0.05	0.24	0.76	0.18
sqft_lot	-0.13	0.09	0.03	0.08	0.17	1.00	-0.01	0.02	0.07	-0.01	0.11	0.18	0.02	0.05	0.01	-0.13	-0.09	0.23	0.14	0.72
floors	0.02	0.26	0.18	0.51	0.35	-0.01	1.00	0.02	0.03	-0.26	0.46	0.52	-0.25	0.49	0.01	-0.06	0.05	0.13	0.28	-0.01
waterfront	-0.00	0.27	-0.01	0.05	0.10	0.02	0.02	1.00	0.40	0.02	0.08	0.07	0.08	-0.03	0.09	0.03	-0.01	-0.04	0.09	0.03
view	0.01	0.40	0.08	0.17	0.28	0.07	0.03	0.40	1.00	0.05	0.25	0.17	0.28	-0.05	0.10	0.08	0.01	-0.08	0.28	0.07
condition	-0.02	0.04	0.03	-0.14	-0.06	-0.01	-0.26	0.02	0.05	1.00	-0.14	-0.16	0.17	-0.36	-0.06	0.00	-0.01	-0.11	-0.09	-0.00
grade	0.01	0.67	0.36	0.65	0.76	0.11	0.46	0.08	0.25	-0.14	1.00	0.76	0.17	0.45	0.01	-0.18	0.11	0.20	0.71	0.12
sqft_above	-0.01	0.61	0.48	0.67	0.88	0.18	0.52	0.07	0.17	-0.16	0.76	1.00	-0.05	0.42	0.02	-0.26	-0.00	0.34	0.73	0.19
sqft_basement	-0.01	0.32	0.30	0.24	0.44	0.02	-0.25	0.08	0.28	0.17	0.17	-0.05	1.00	-0.13	0.07	0.07	0.11	-0.14	0.20	0.02
yr_built	0.02	0.05	0.15	0.53	0.32	0.05	0.49	-0.03	-0.05	-0.36	0.45	0.42	-0.13	1.00	-0.22	-0.35	-0.15	0.41	0.33	0.07
yr_renovated	-0.02	0.13	0.02	0.04	0.06	0.01	0.01	0.09	0.10	-0.06	0.01	0.02	0.07	-0.22	1.00	0.06	0.03	-0.07	-0.00	0.01
zipcode	-0.01	-0.05	-0.15	-0.21	-0.20	-0.13	-0.06	0.03	0.08	0.00	-0.18	-0.26	0.07	-0.35	0.06	1.00	0.27	-0.56	-0.28	-0.15
lat	-0.00	0.31	-0.01	0.01	0.05	-0.09	0.05	-0.01	0.01	-0.01	0.11	-0.00	0.11	-0.15	0.03	0.27	1.00	-0.14	0.05	-0.09
long	0.02	0.02	0.13	0.24	0.24	0.23	0.13	-0.04	-0.08	-0.11	0.20	0.34	-0.14	0.41	-0.07	-0.56	-0.14	1.00	0.33	0.25
sqft_living15	-0.00	0.59	0.39	0.56	0.76	0.14	0.28	0.09	0.28	-0.09	0.71	0.73	0.20	0.33	-0.00	-0.28	0.05	0.33	1.00	0.18
sqft_lot15	-0.14	0.08	0.03	0.08	0.18	0.72	-0.01	0.03	0.07	-0.00	0.12	0.19	0.02	0.07	0.01	-0.15	-0.09	0.25	0.18	1.00

This correlation matrix expresses the strength of the relationship between each attribute. When the value approaches 1 and the color is red, the relationship between the attributes becomes stronger. This correlation matrix helps us in choosing the strongest relationships to use in the linear regression model.

Outline

- ***Introduction and Exploratory about Dataset***
- ***Linear Regression model***
- ***Classification by (Decision tree)***
- ***Clustering***
- ***Anomaly Detection***
- ***PCA***

Linear Regression model



Linear regression is a fundamental statistical method used in data science and machine learning. Based on the linear regression analysis, there is a linear relationship between the living area (square feet) and the property price. As the living area per square foot increases or decreases, the price of the property tends to rise or fall accordingly.

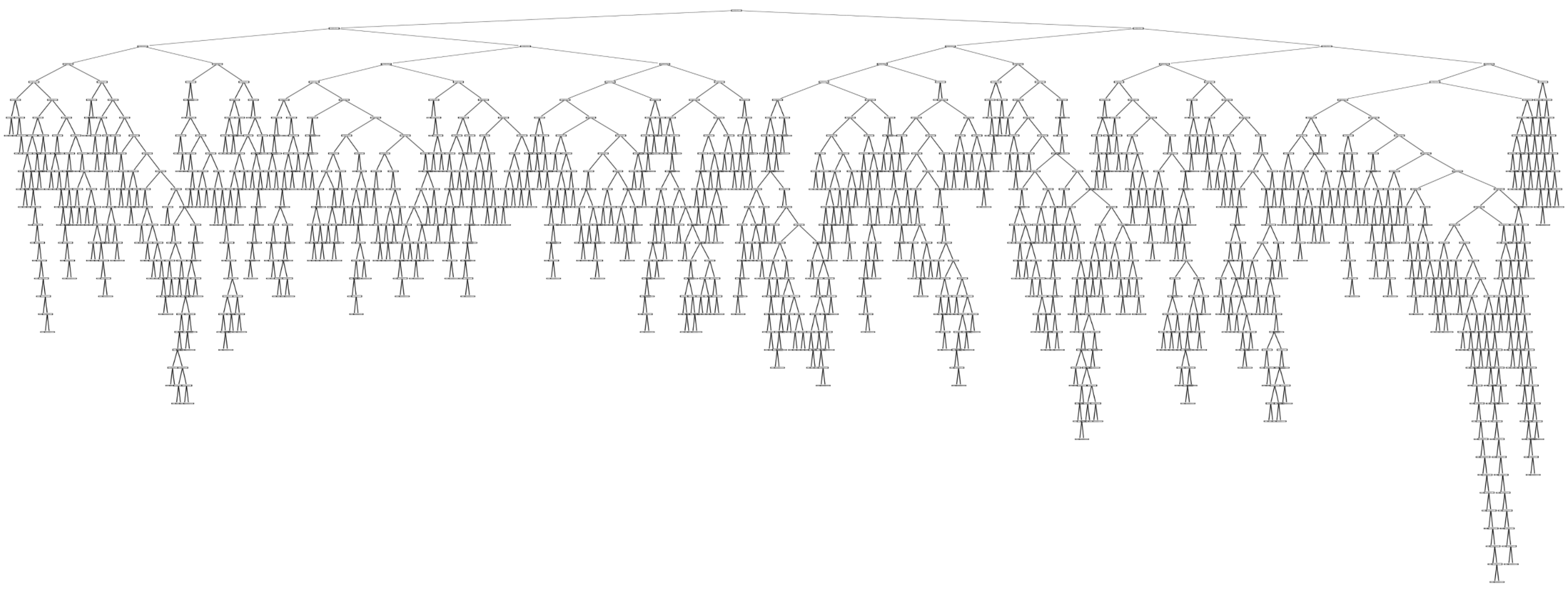
The model was also tested and the results shown are:

RMSE=1.3933333680162217 and R^2=066705049227774556

Outline

- *Introduction and Exploratory about Dataset*
- *Linear Regression model*
- *Classification by (Decision tree)*
- *Clustering*
- *Anomaly Detection*
- *PCA*

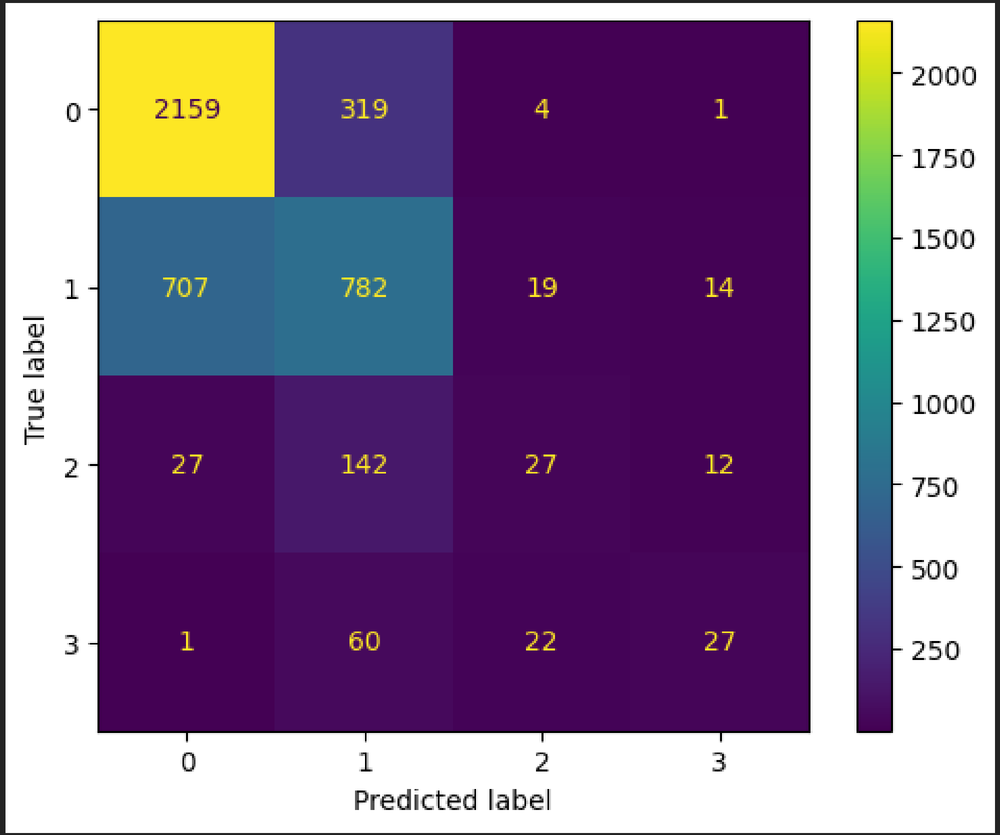
Classification by (Decision tree)



The Decision Tree model has been executed on the variable sqft_living. This allows us to understand how the number of sqft_living in a house impacts the predicted sale price.

Model Evaluation by Confusion matrix

As we can see, most of the data is in the true positive position, which means that the model's predictive ability is high

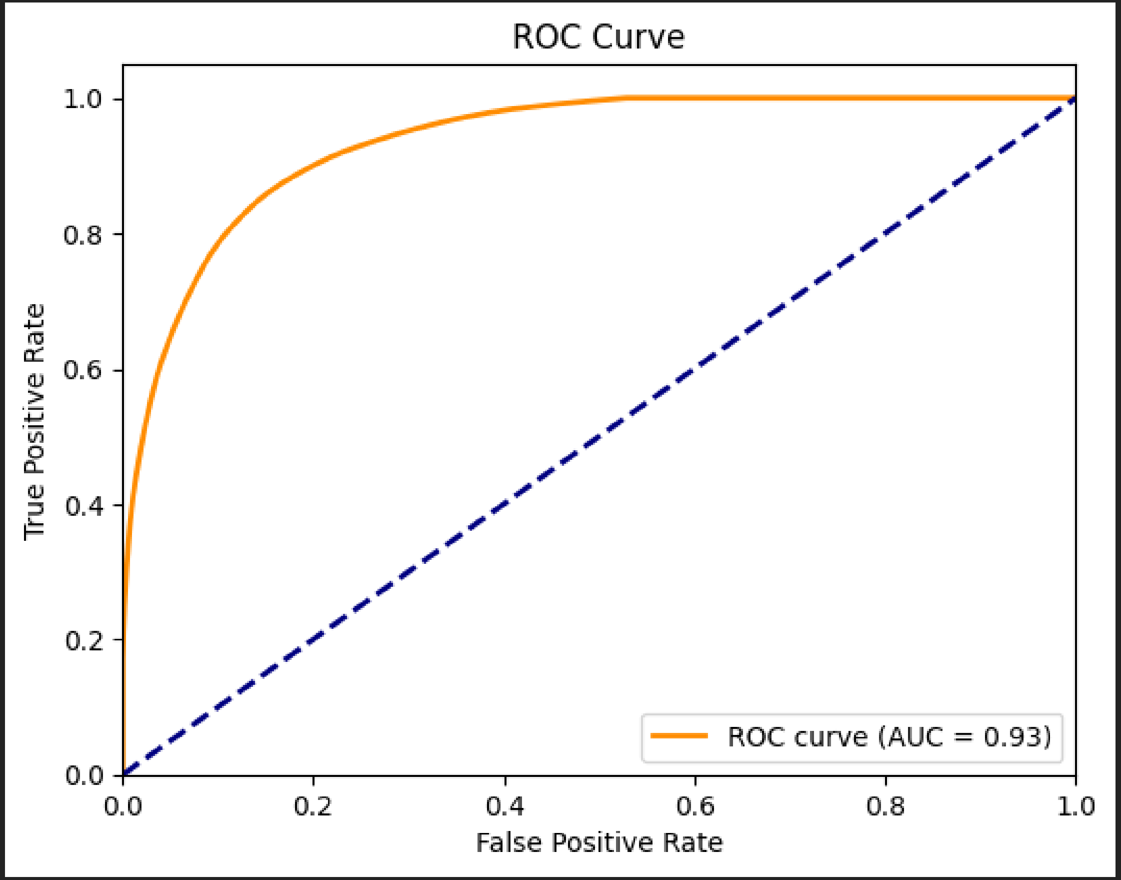


True label \ Predicted label	0	1	2	3
0	2159	319	4	1
1	707	782	19	14
2	27	142	27	12
3	1	60	22	27

Model Evaluation by ROC Method

AUC (Area Under the Curve): 0.93
Indicates excellent performance.

The model performs significantly better than random guessing and is highly effective at classification.

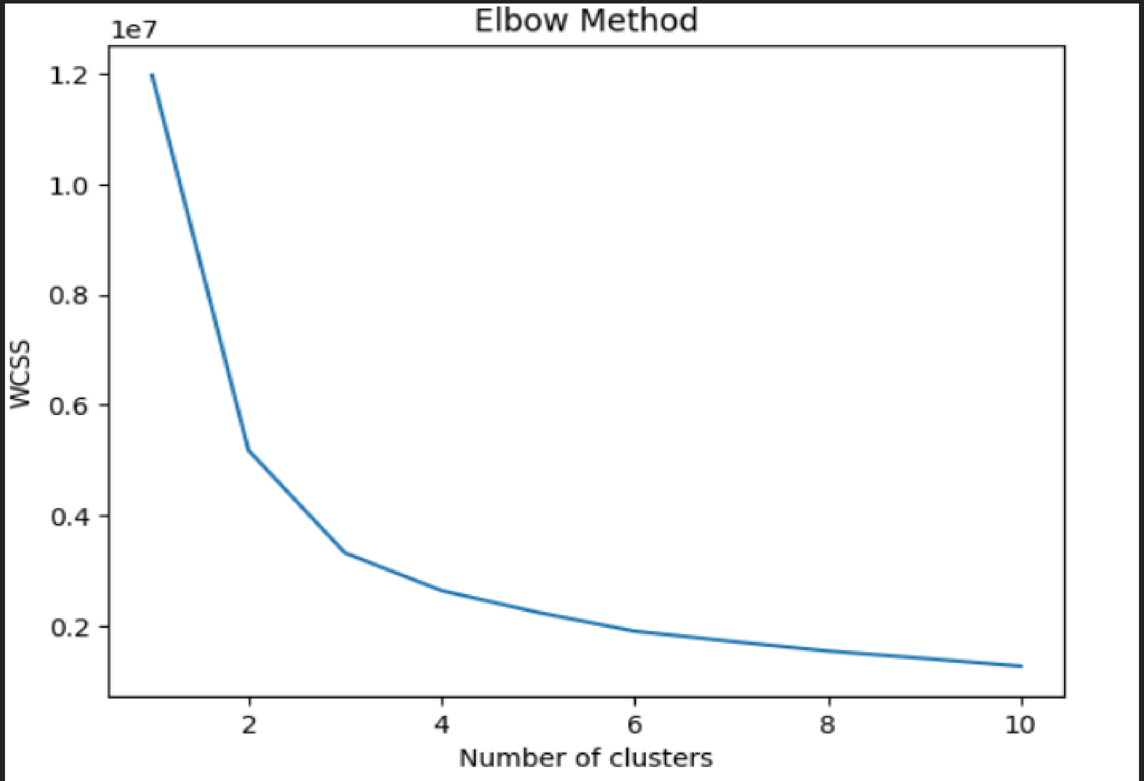


Outline

- ***Introduction and Exploratory about Dataset***
- ***Linear Regression model***
- ***Classification by (Decision tree)***
- ***Clustering***
- ***Anomaly Detection***
- ***PCA***

Clustering

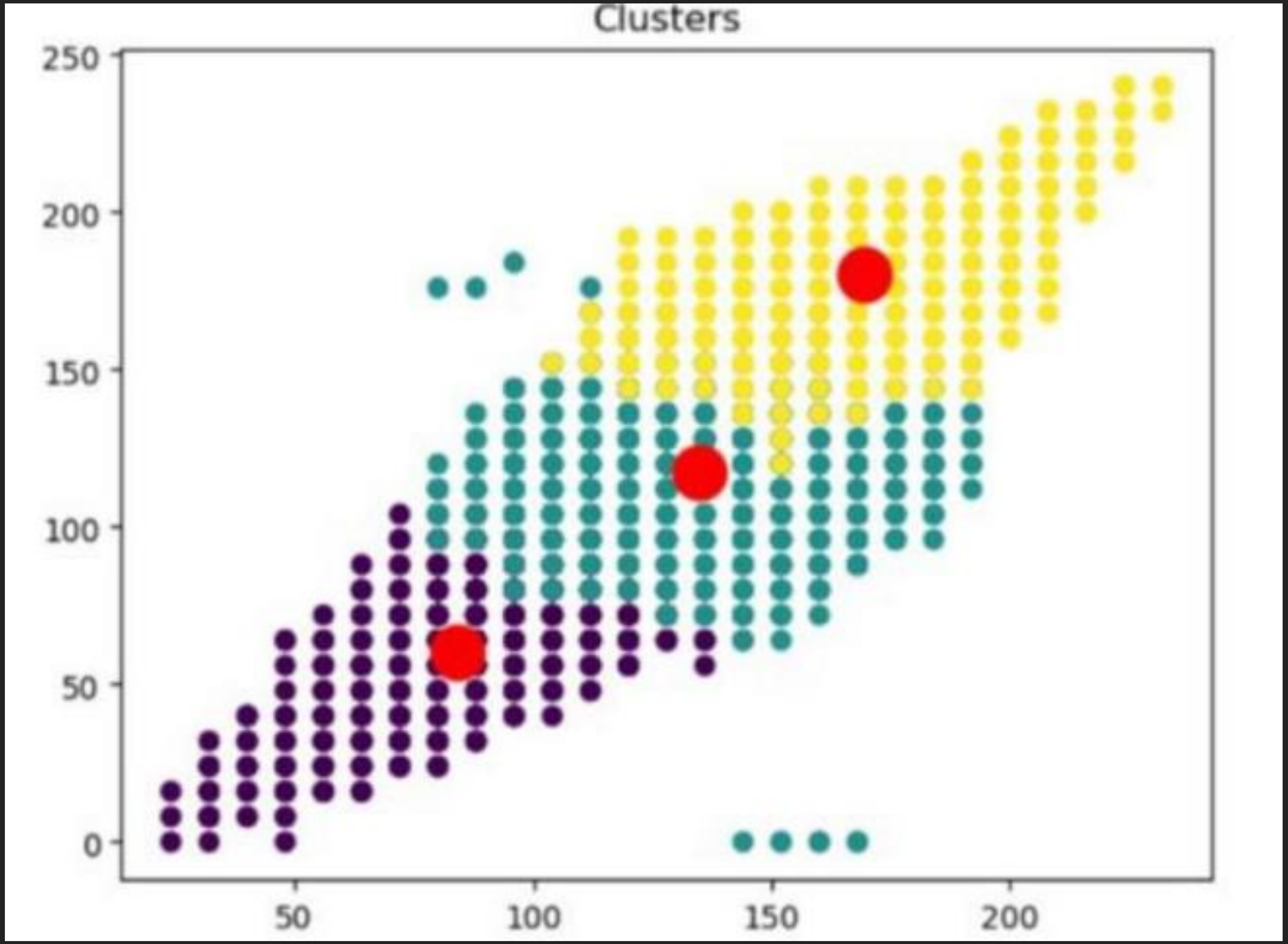
Elbow Method for Optimal Cluster Selection: Plot demonstrating the within-cluster sum of squares (WCSS) as a function of the number of clusters. The elbow point indicates an optimal number of clusters for K-means clustering.



Clustering(.cont)

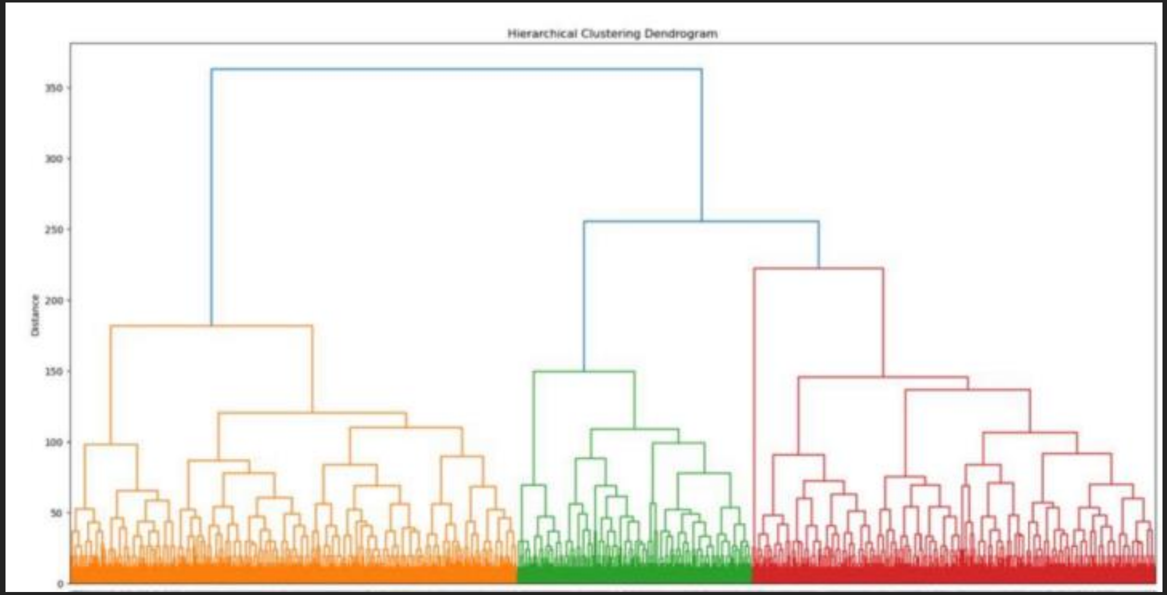
Cluster Visualization: Scatter plot illustrating the clusters generated by K-means clustering algorithm. Each point represents a data instance colored by its assigned cluster, with centroids highlighted in red.

Sum of Squared Error (SSE): 3318367.90223293

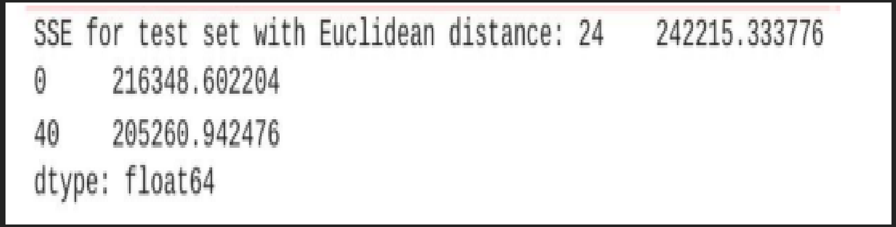


Clustering(.cont)

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.



<u>Sum of Squared Error (SSE):</u>

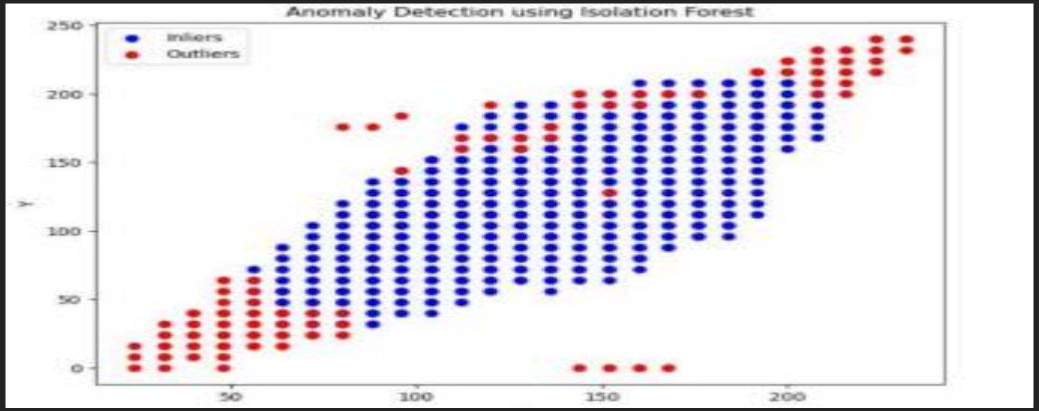


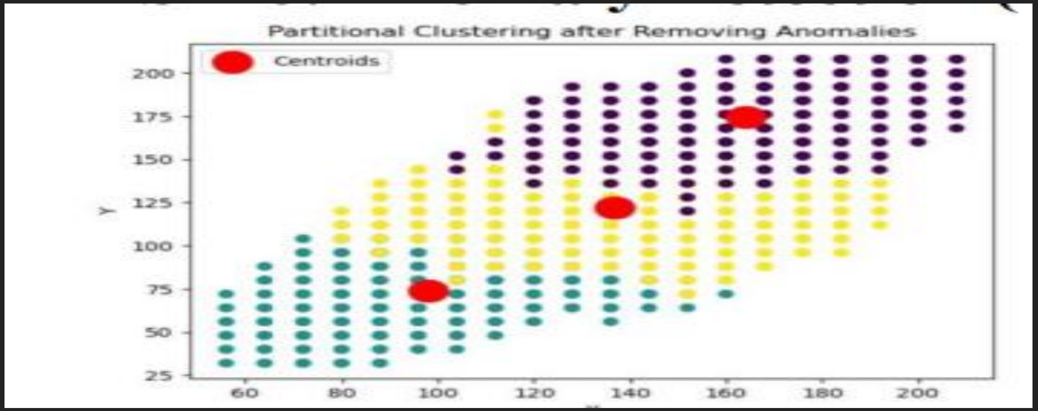
```
SSE for test set with Euclidean distance: 24    242215.333776
0     216348.602204
40    205260.942476
dtype: float64
```

Outline

- ***Introduction and Exploratory about Dataset***
- ***Linear Regression model***
- ***Classification by (Decision tree)***
- ***Clustering***
- ***Anomaly Detection***
- ***PCA***

Anomaly Detection





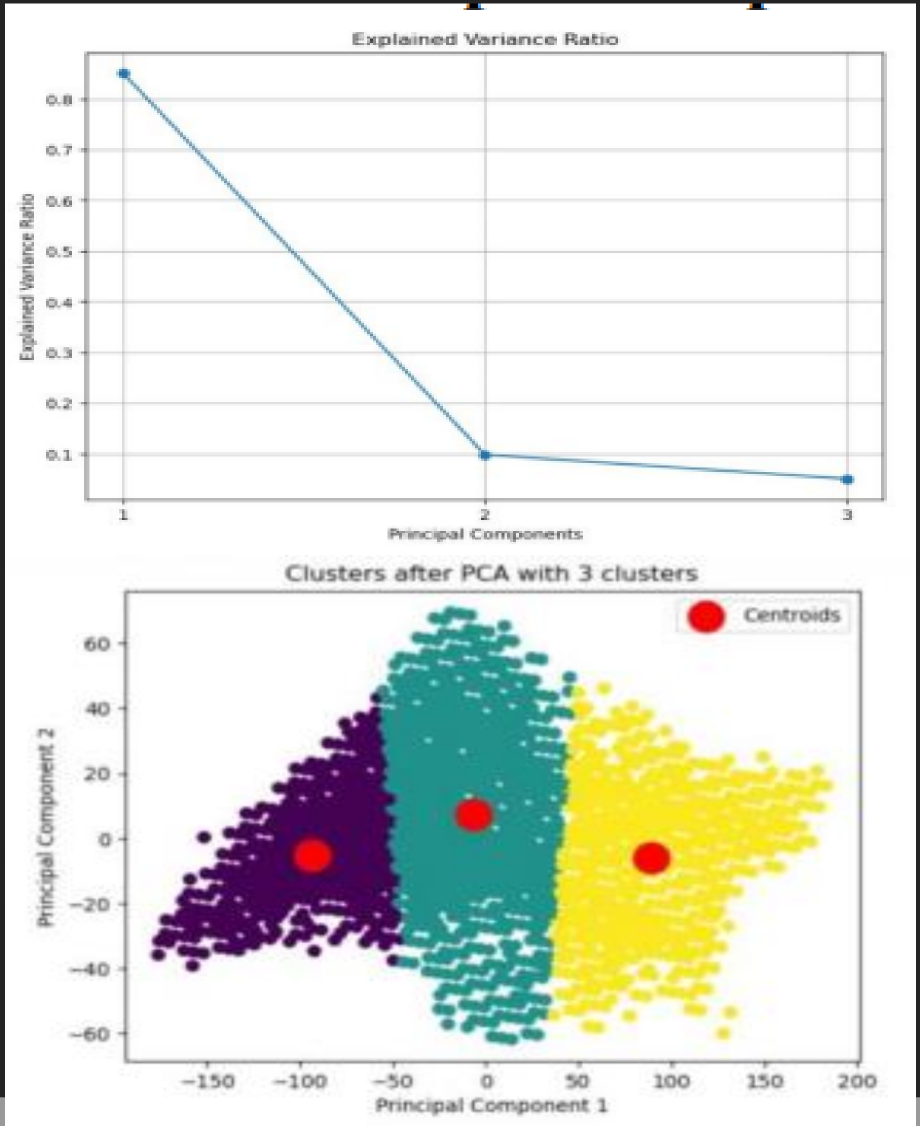
Anomaly detection is finding unusual patterns in data. It helps to identify outliers or distortions that deviate from the norm. In this task, we used it well and were able to detect and recognize outliers, as shown in the graphics

Outline

- ***Introduction and Exploratory about Dataset***
- ***Linear Regression model***
- ***Classification by (Decision tree)***
- ***Clustering***
- ***Anomaly Detection***
- ***PCA***

Principal Component Analysis (PCA)

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



SUBMITTED BY:

Eyad Mohammed Alharthi:444000005

Mohammed Nedal Alshareef:444004824

Ziyad omar altalhi:444000039

Aseel Abdulfattah Almaghrabi:444004296

Raed Falah Alhelali:444006561

Thanks for listening!