CLUSTERING BASED REGRESSION

GROUP-6

|  |  |
| --- | --- |
| TEAM MEMBERS | ROLL NO: |
| BODDU MADAN GOPAL | AM.EN.U4AIE22112 |
| MONINGI AYUSH | AM.EN.U4AIE22134 |
| POOJA S PILLAI | AM.EN.U4AIE22140 |
| PARVATHY SURESH KUMAR | AM.EN.U4AIE22141 |

Abstract

Traditional linear regression uses simple linear relationship across the entire dataset. When it comes to real world problems, the data is not captured well by a single linear model. Using linear regression model also is not good at capturing the complexities in the real world data. To overcome this problem, a clustering based linear regression is done. This project is based on the application of cluster based linear regression which helps to improve prediction accuracy. By using clustering techniques, we divide the data into meaningful groups and fit separate linear models for each cluster. We will apply different linear models for each clustering method. The cluster-based linear regression method is a technique that gives us the opportunity to deal with differences in a much more accurate way. This project will involve the K-means, Hierarchical, and DBSCAN methods of creating clusters by using linear regression. We seek to demonstrate that the application of cluster-based approach can yield better and more meaningful models as opposed to the traditional linear regression method by using clusters.  
  
  
Table Of contents

* Introduction
* Literature Review
* Implementation
* Data Analysis
* Results
* Conclusion

Introduction

Linear regression is the simplest technique applied in describing the relationship that exists between a target variable and one or more features, or a dependent variable against one or more independent variables. Traditional linear regression used only a single linear relationship across the total dataset and often failed at capturing complex relationships prevailing in most real-world data. To combat this weakness, we use clustering-based linear regression, which combines techniques of clustering with linear regression to result in an accurate prediction. In the clustering-based regression, rather than one regression model for the whole data set, the data is first clustered into a similar group, and then the training of separate regression models for each cluster. One such assumption is that data observations belonging to the same clusters would more likely reflect similar relationships between their independent and dependent variables. In this way, each regression model can model more localized patterns and relationships more effectively. The major processes involved are data preprocessing, application of clustering, training individual regression models for every cluster, and testing. Generally, clustering-based regression increases the prediction accuracy compared to simple linear regression. Furthermore, model interpretability within each of the clusters may give more insight into what drives the outcomes of each group's factors, therefore resulting in more targeted essential interventions for business strategies.  
  
Other areas, ranging from econometrics through wireless sensor networks to energy use forecasting and finally healthcare, have also combined clustering techniques with linear regression to great success. This approach is referred to as clustering-based linear regression, where the strengths of two methodologies are leveraged to improve predictive accuracy and the robustness of a model.  
  
In many fields of studies recently, clustering-based linear regression has been revealed to find its application. For instance, MacKinnon et al. (2023) constructed tests for the cluster structure of the error variance matrix in linear regression models. This provides a robust framework for validating clustered robust variance estimators [1]. Zhang et al. (2020) developed a fuzzy time series forecasting model using multiple linear regression and time series clustering to improve the prediction accuracy by clustering data based on their linear relationship[2]. Similarly, threshold-based adaptive iterative linear regression-based clustering algorithm was proposed by Hemavathi and Sudha in 2016 for Wireless Sensor Networks to counteract the complexity of hierarchical clustering[3].  
  
In the domain of energy consumption forecasting itself, Chen et al. established a density-based clustering multiple linear regression model to predict the energy consumption by electric vehicles, and such a model performed better than traditional methods. Besides, Goia et al. adopted functional clustering with linear regression to predict peak energy demand by performing functional linear regression on daily data clustering data to enhance its predictive accuracy.   
  
It has also benefited the healthcare application. Nagwani and Deo proposed a cluster regression-based technique for estimating concrete compressive strength, where they illustrated this using hard and fuzzy clustering methods in 2014, showing it to have handled outliers and noise [6]. More recently, Huang et al., in 2020, utilized the K-means-based LWLR model in predicting morbidity in chronic obstructive pulmonary disease, indicating that it had predictive power and reliability over classical regression methods [7].  
  
That is to say, the combination of clustering techniques and the traditional linear model in a regression setting makes it versatile and powerful enough to work with different kinds of datasets in various problem domains. Continued research and improvement in this approach promise huge strides in predictive modeling and data analysis.

Literature Review  
  
Implementation

* Data Preprocessing

The process of transforming raw data into clean and usable format which can be used for analysis and modeling. It enhances the quality of data, which helps to find more accurate and reliable models, which also helps in reducing the computational complexity and improves the overall efficiency of the whole data analysis process. The process includes the data clearing process, such as removing duplicates, standardizing error normalization, handling empty records, and the integration of data collected from various sources. This includes the transformation of the dataset (normalization, standardization, categorical variable encoding/discretization), feature engineering, establishing the dimensionality reduction and feature scaling.

* Clustering

This step involves applying clustering algorithms to partition the dataset into meaningful groups.

1. **K –means clustering**

This is an unsupervised learning algorithm which deals with the petitioning method that divides the dataset into k non overlapping clusters. The main aim of this algorithm is to minimize the variance within each cluster. The process starts with the initialization steps where each data point is assigned to nearest center based on a distance metric forming K clusters. Then the new centroids are calculated by taking the mean of all data points in each cluster. This process is followed until the centroids do not change or a maximum number of iterations are reached. To determine the optimal number of clusters, we have used the Elbow method. This method involves plotting the within- cluster sum of squares against the number of clusters and then identifies the elbow point where the rate of decrease in WCSS slows significantly. This point suggests the optimal number of clusters. Along with this the silhouette score can be calculated to assess the cluster quality, providing a measure of how similar each data point is to its own cluster. K- means is simple and fast and it works well with spherical, equally sized clusters.

1. **Hierarchical Clustering**

Hierarchical clustering is a method that builds a hierarchy of clusters organized as a hierarchical tree which can be visualized as a dendrogram, where each data point starts as its own clusters. It is the method in which there is no need of predefining the number of clusters. The algorithm iteratively merges the closest clusters based on linkage distances(minimum distance, maximum distance, mean distance etc). This process continues until all data points belong to a single cluster or a stopping criterion is met. By checking the increase of the linkage distances we can determine the optimal number of clusters from the dendrogram which indicates distinct clusters. The cut of point in the dendrogram where this increase occurs defines the optimal number of clusters.

1. **DBSCAN(Density Based Spatial Clustering of Applications with Noise)**

Density based Clustering algorithm which identifies clusters based on the density of data points. It categorizes each data point as a core point, border point or noise based within a specific radius(epsilon) and a minimum number of neighboring points. DBSCAN identifies clusters based on density reachability which becomes difficult when it comes to varying densities or non – spherical shapes. The DBSCAN algorithm clusters the training data. For each cluster a separate linear regression model is trained. For new data the algorithm clusters the data again. If a data point is considered noise an average prediction from all model is used. This approach, by capturing the local structures in data and separate regression models to different clusters, improves the accuracy of prediction.

* Regression with Cluster

Here clustering methods are combined with regression models to improve predictive accuracy.

Hierarchical + Linear Regression

Hierarchical clustering is used to groups data points into clusters based on their similarity, without the a predefined number of clusters. After finding out the optimal number of clusters using a dendrogram, each cluster is treated as a separate dataset. For each cluster, a linear regression model is fitted to predict the target variable based on the input features within that cluster. This way it is easy to capture different linear relationships within each cluster, which improves the accuracy.

K- means + Linear regression

In K-means clustering partitions the data into predefined clustered where separate linear regression or Random Forest models are used to capture unique patterns within each cluster. After using the elbow method to determine the optimal number of clusters, each cluster is treated separately. In each cluster a linear regression model is trained to predict the target variable. This algorithm allows to capture unique relationships within each cluster, thus improving the overall predictive performance.

K – means + Random Forest

First the data set is divided into clusters. Here instead of fitting a linear regression model, a random forest is trained for each cluster. Random forest then builds multiple decision trees and aggregate their predictions.

Hierarchical clustering + Random Forest

After determining the optimal number of clusters a random forest model is fitted to each cluster. This combination allows for capturing the complex and non-linear relationships within each cluster.

DBSCAN + Linear Regression

DBSCAN identifies clusters based on the density of data points enabling distinct linear regression models for dense regions and also making the low density areas as noise. After clustering using DBSCAN a linear regression model is fitted to each identified cluster. This method benefits from DBSCANS’s ability to find clusters of arbitrary shape and handle noise. By fitting separate regression models, it captures distinct linear relationships within dense regions of the data, providing accurate predictions of each cluster.

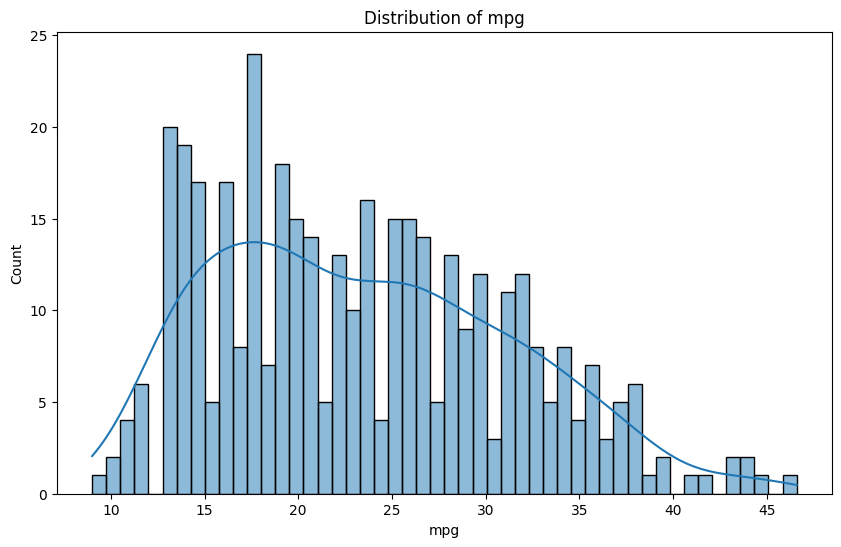
* Prediction and Evaluation

Regression models are applied to each cluster through various clustering techniques to improve prediction accuracy. In hierarchical clustering linear regression models are fitted to each cluster, capturing the unique relationship within them. K- means clustering allows for the application of linear regression or Random Forest models to its predefined clusters, capturing distinct patterns. DBSCAN identifies clusters based on the density of data points enabling distinct linear regression models for dense regions, also handling noise by taking the average of all prediction across all models. The combination of clustering and regression techniques ensure that the models are well-suited to the specific characteristics of each clusters, leading to, more accurate predictions. Also evaluation matrices such as the silhouette score assess the cluster quality while performance is measured through metrices like R- squared and mean squared error ensuring effective predictions.

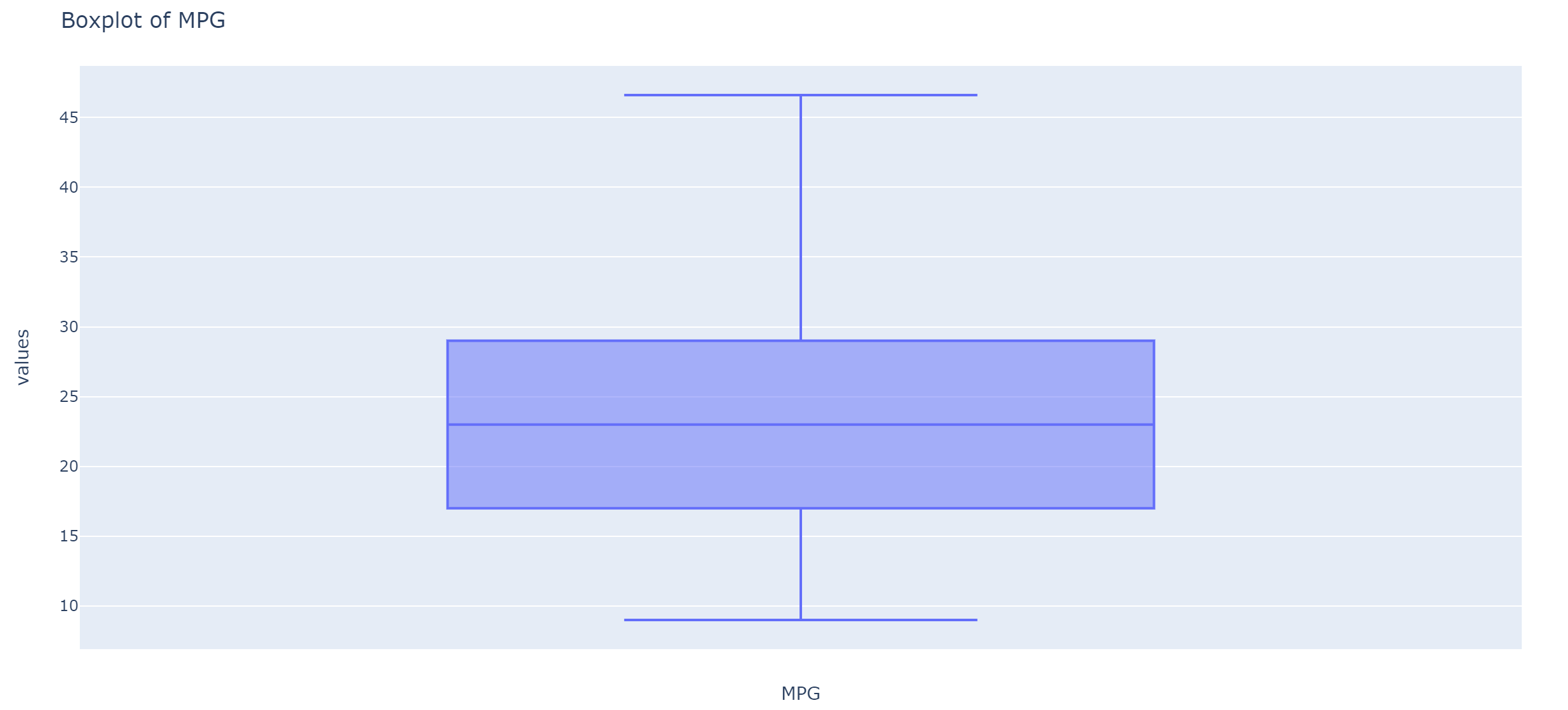
Data Analysis

1. Auto MPG prediction dataset

Histogram is used to analyze the distribution of mpg.

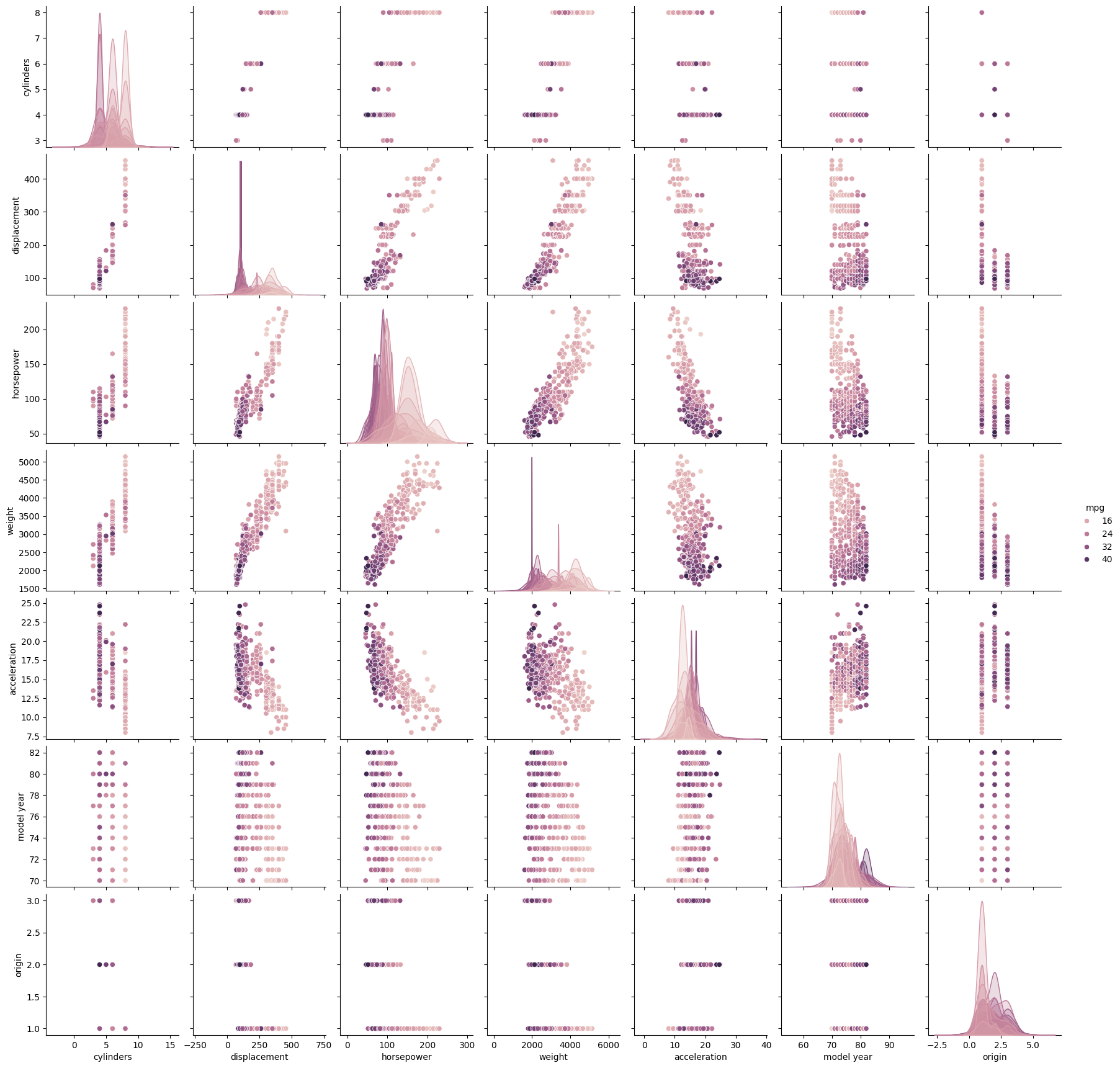


Box plot



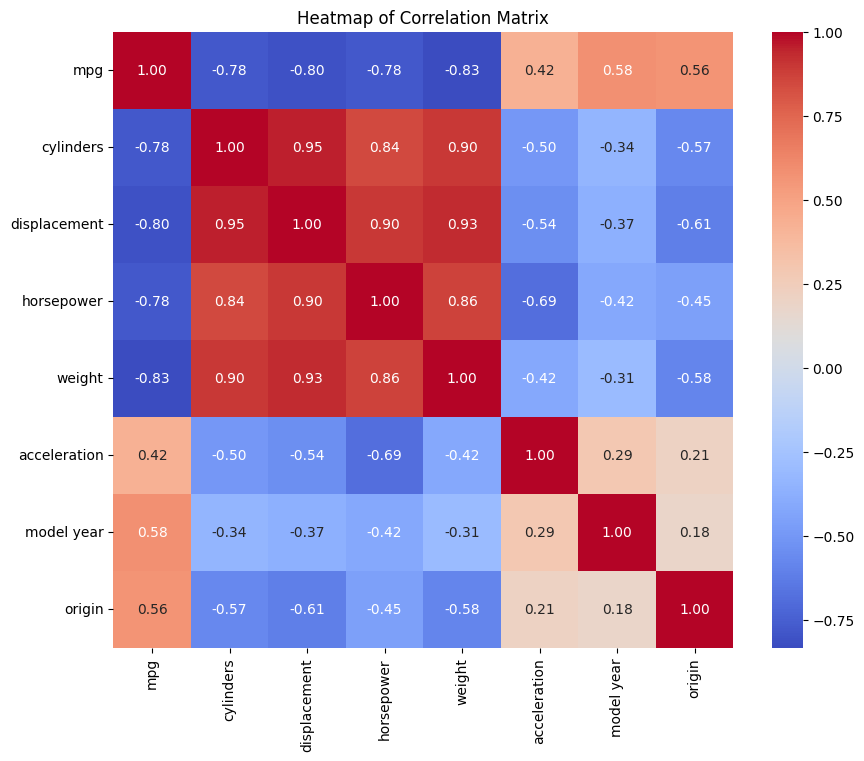
Pair plot analysis

It shows the distribution and correlation of different features. It helps in comparing the distribution of each feature against the others (ie correlation) and selecting the target variable.

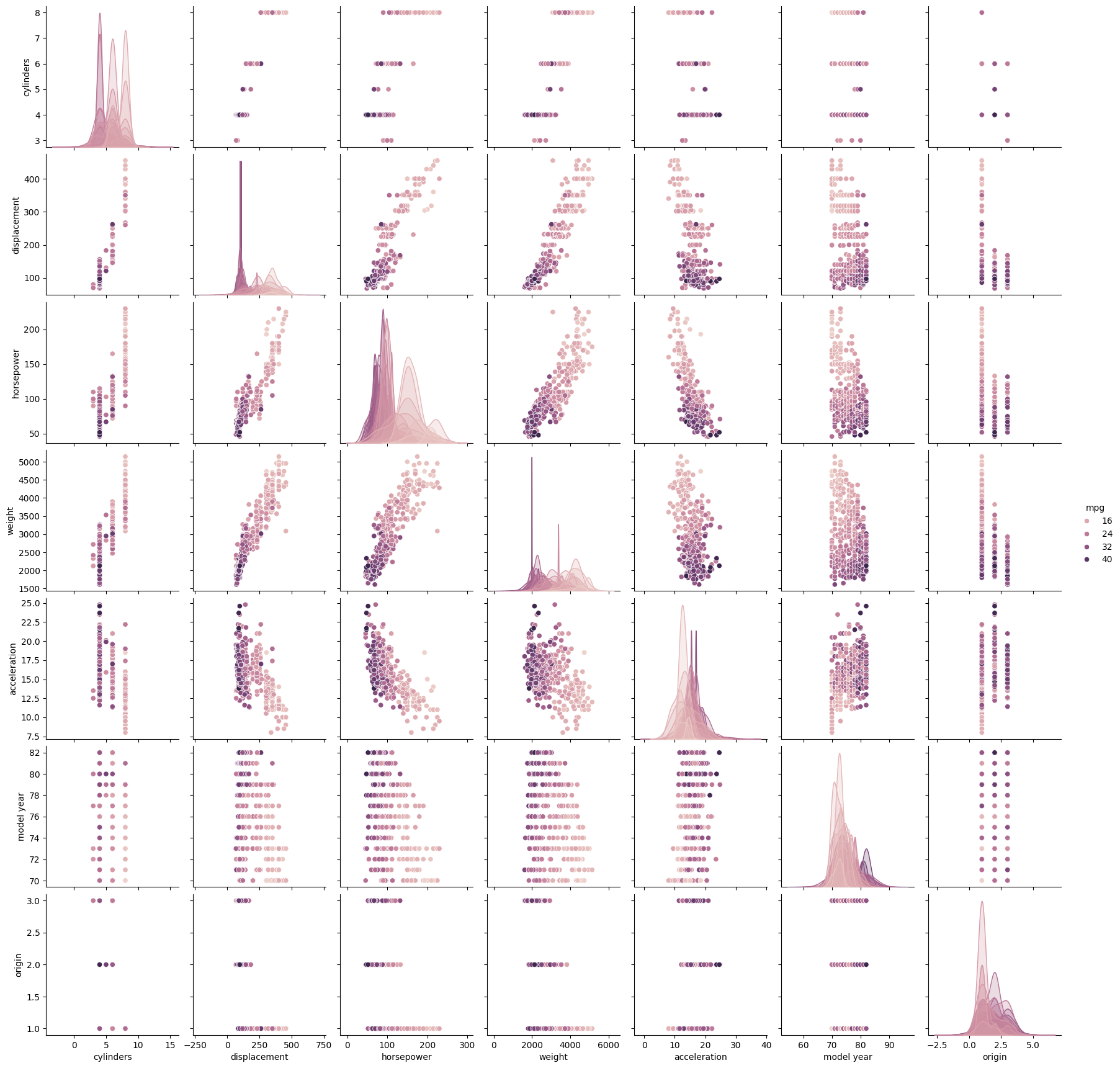


Correlation Matrix of mpg

It represents the correlation coefficients (linear relationship) between every other attribute in the dataset.ie if one is correlation coefficient is positive it means that if one variable increase then other one increases.

Here in this figure, mpg to origin value is positive which indicates that both are in positive linear relationship. Elbow method

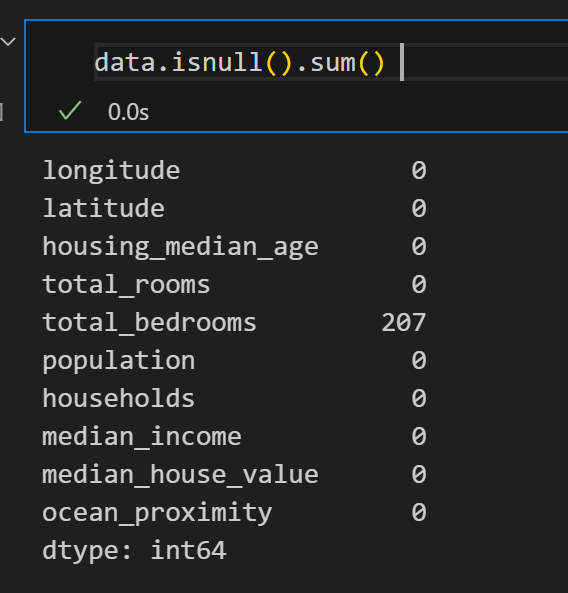
The Elbow method helps to find out the number of clusters. We have used the k-means clustering method to find the inertia which indicates the compactness of the cluster. The bend represents the approximate number of clusters to be formed.



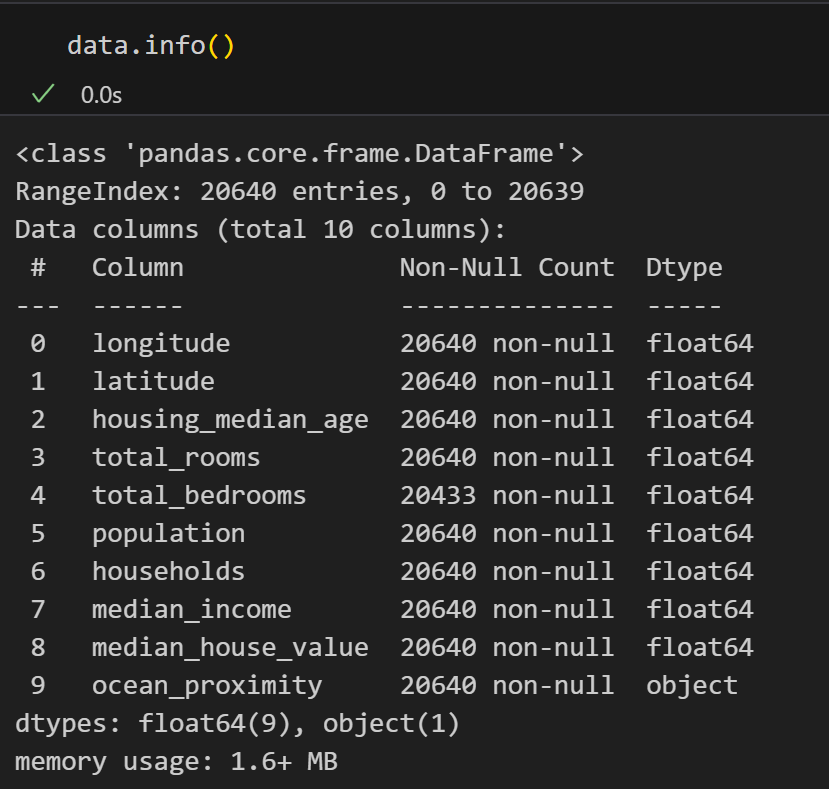
Silhouette score= 0.4090

1. House price prediction dataset

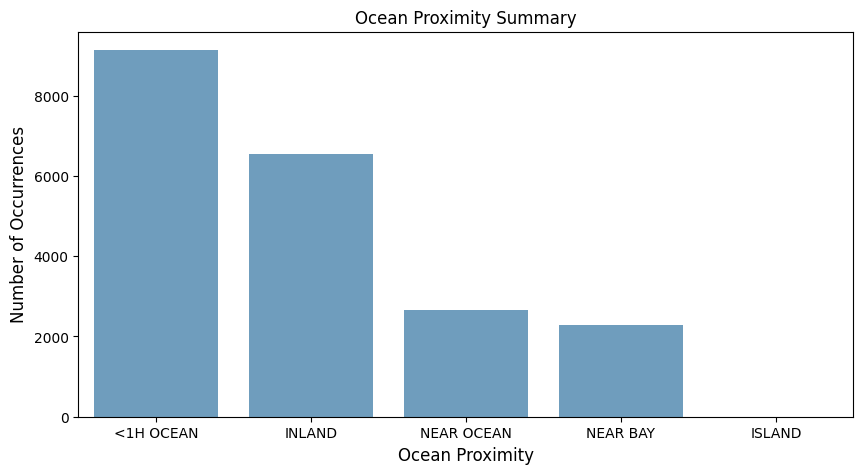
Only one attribute has null values, so we are ignoring and not handling it.



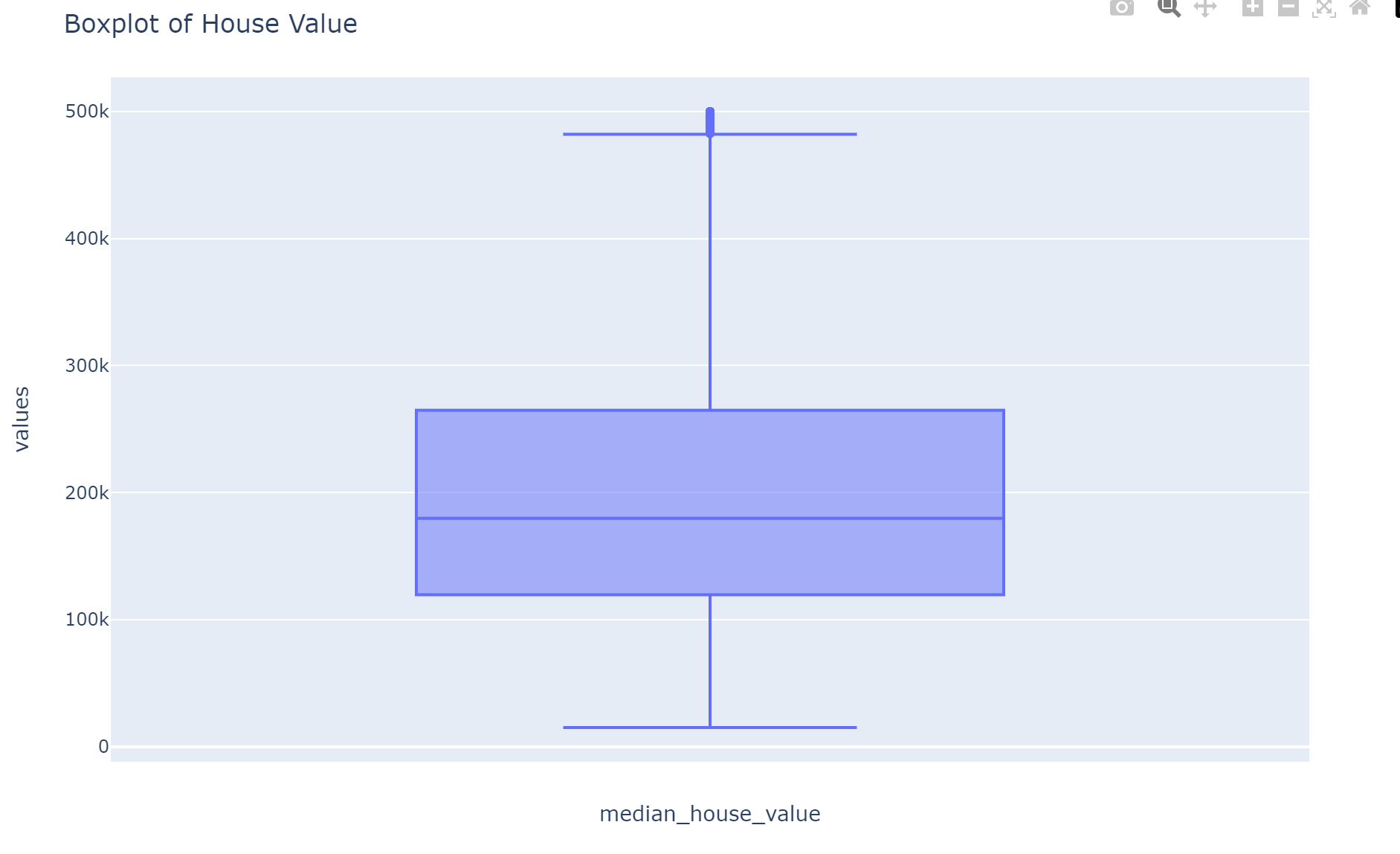
The feature, ‘ocean\_proximity’ data type is categorical value so we must label encode that feature.



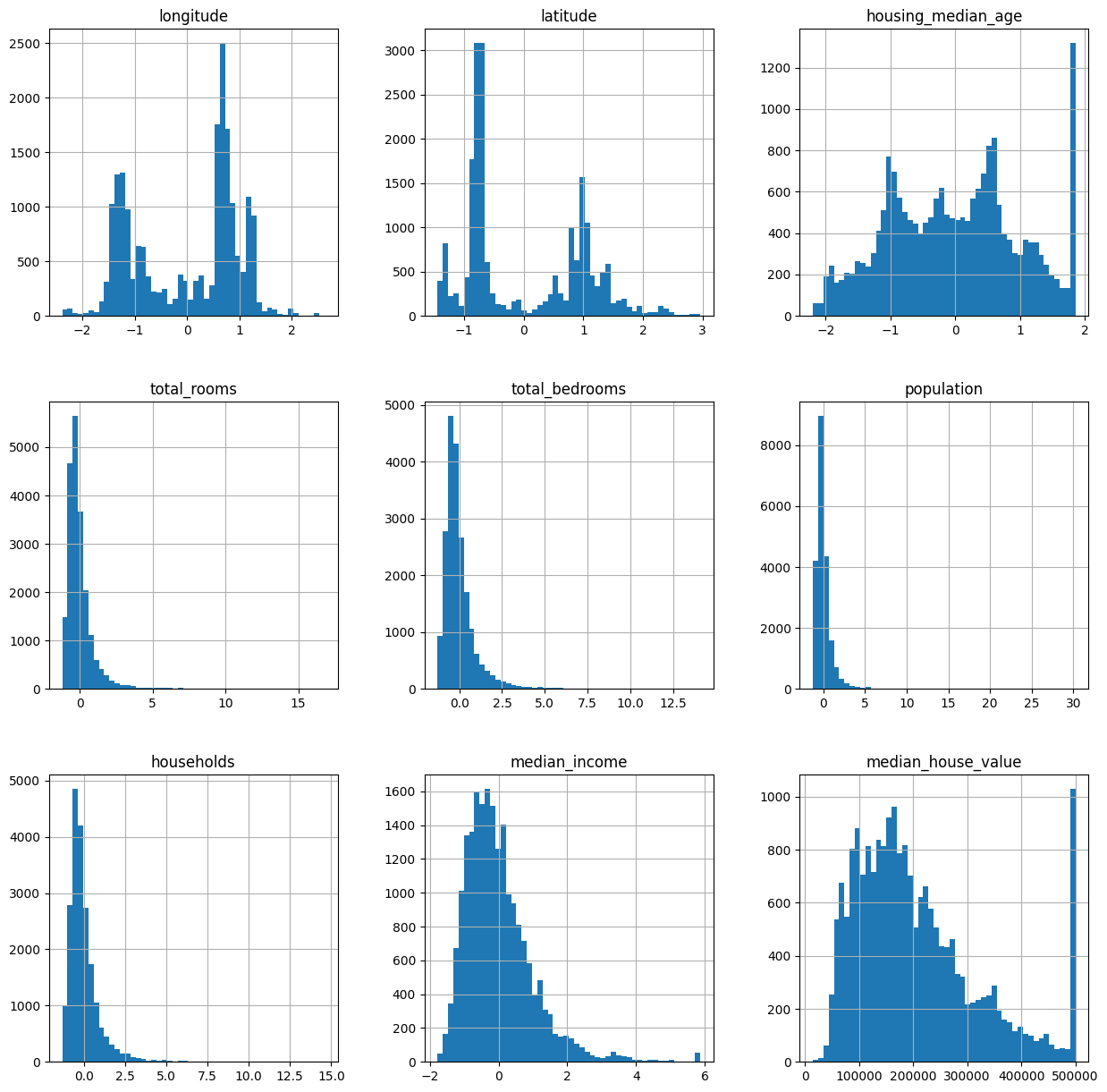
This graph represents the number of occurrences of the different ocean\_proximity values in the dataset.



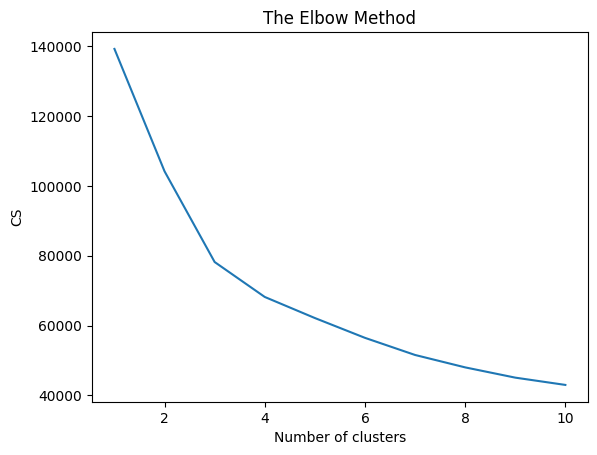
Box plot



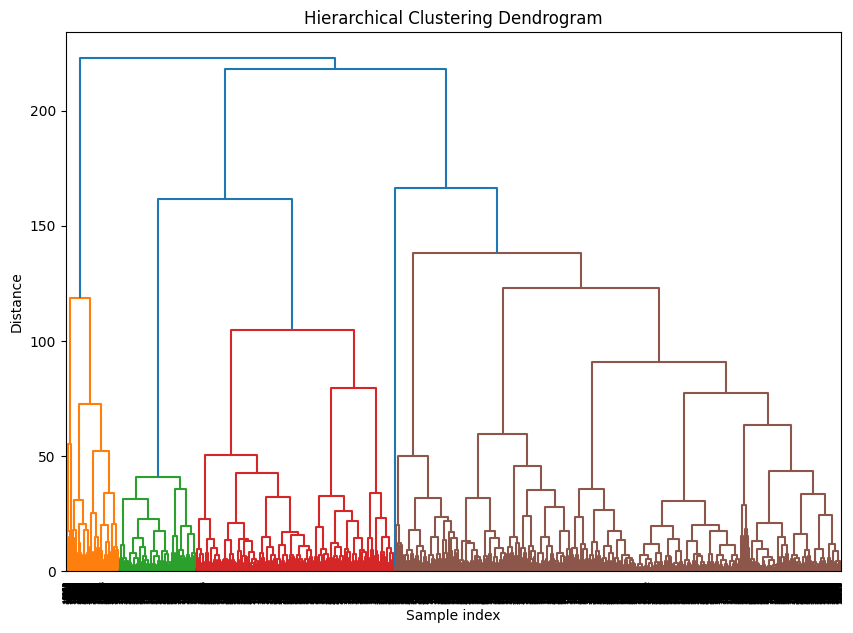
Distribution of the data (values) in the dataset in the features



Number of clusters calculated using the elbow method:

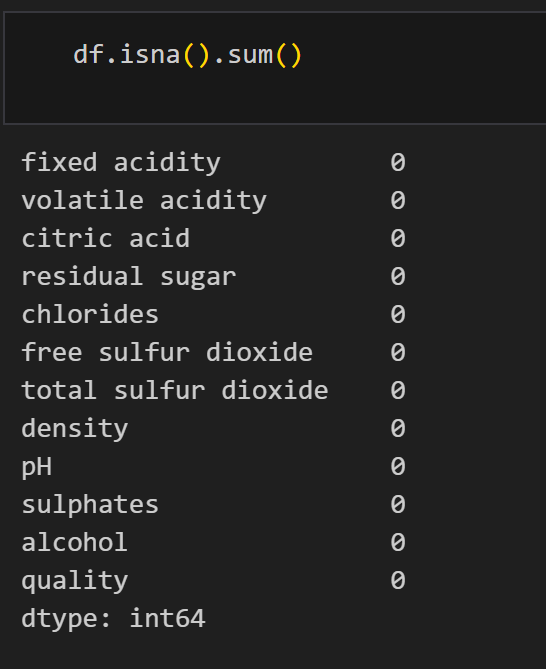


Hierarchical Clustering

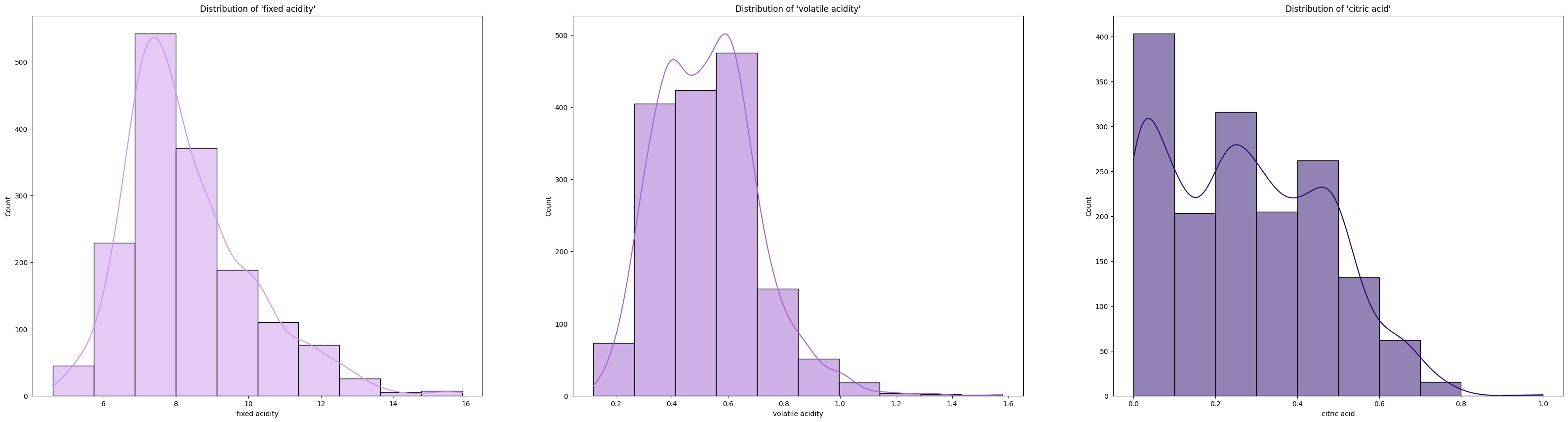


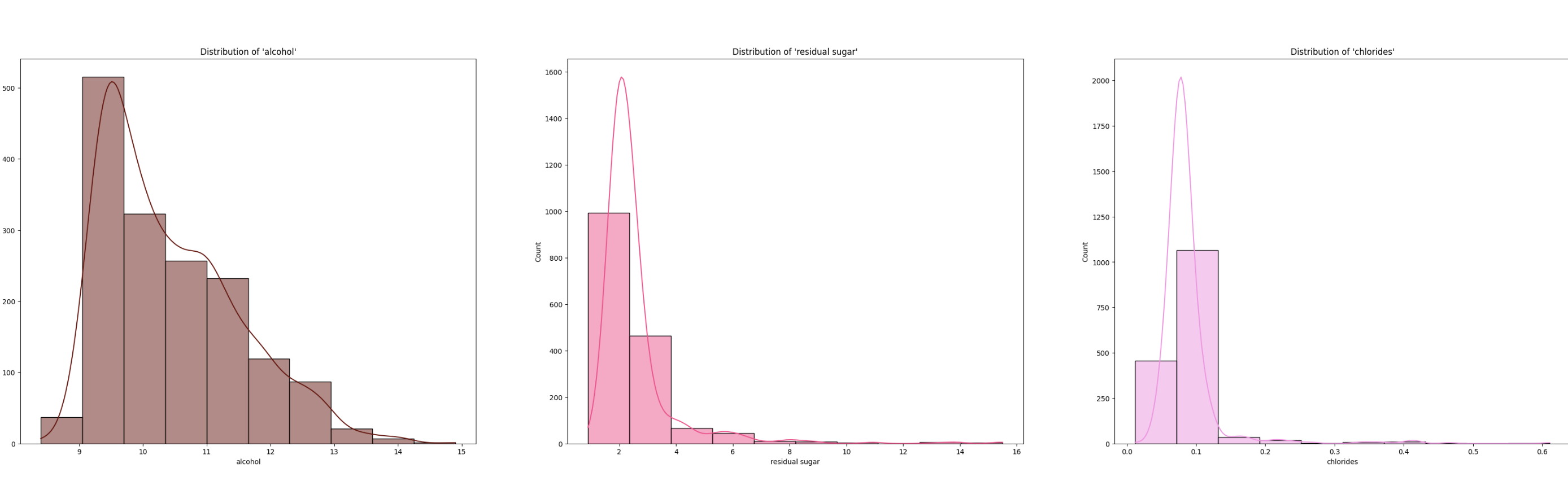
1. Wine Quality prediction dataset

Here since there are no null values, we don’t have to handle any null values.



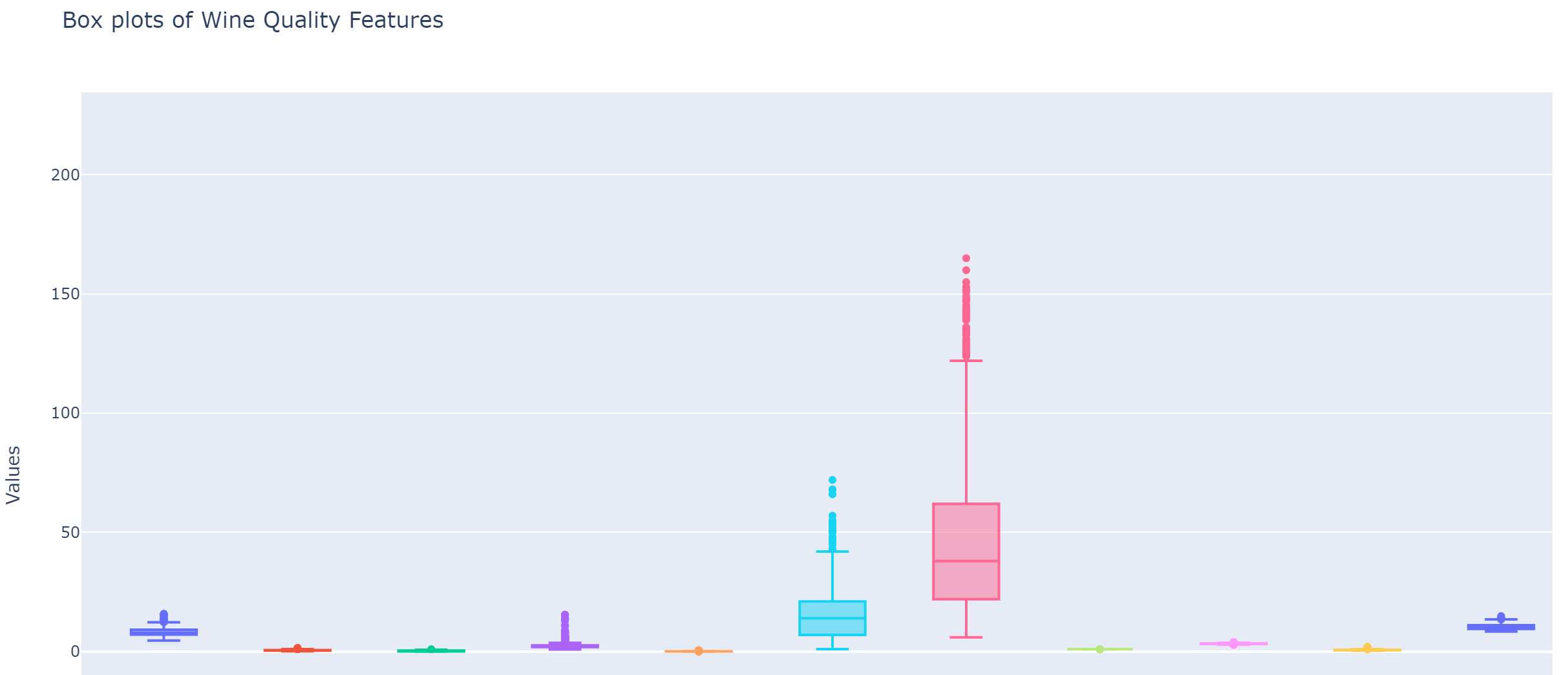
Distribution of data (fixed acidity, volatile acidity, citric acid, alcohol, residual sugar, chlorides):



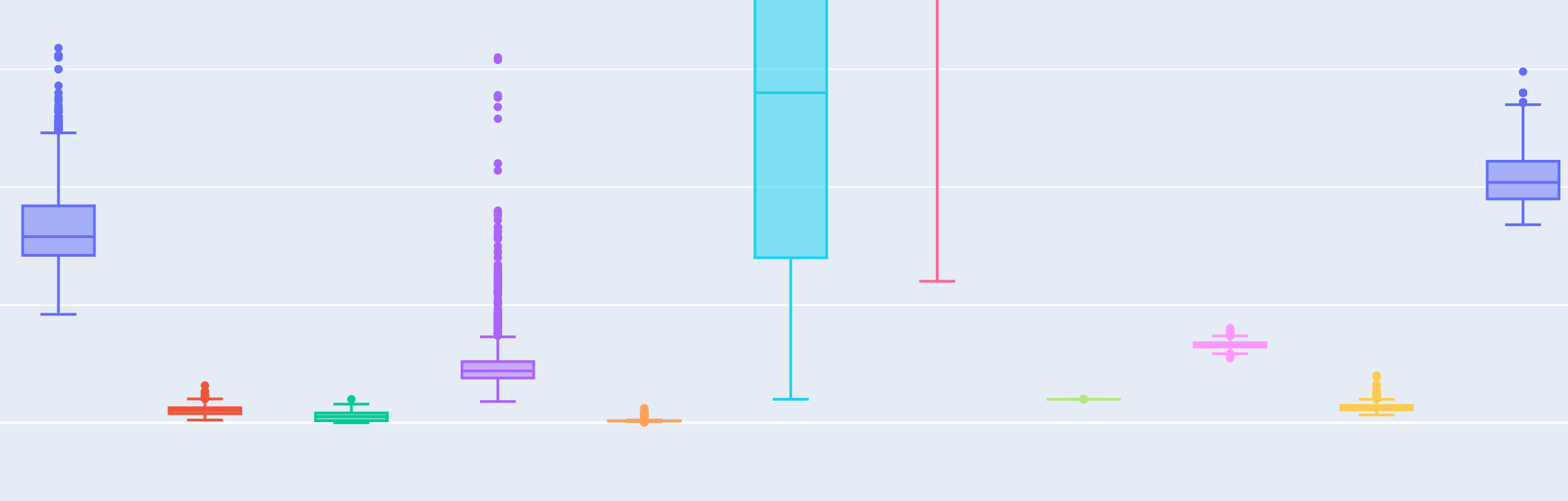


Box plot of the features of the dataset

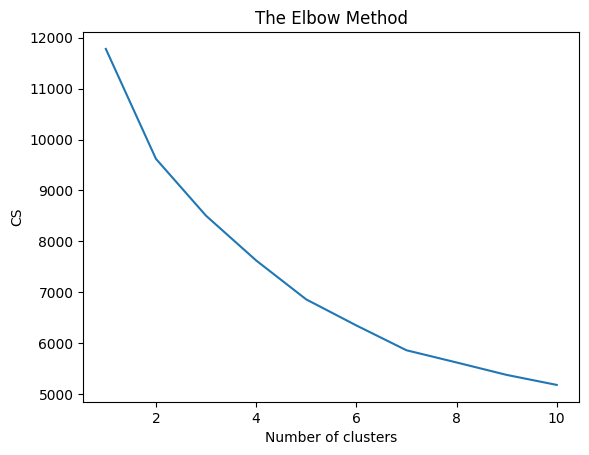
The datapoints out of the inter-quartile region represent the outliers present in the dataset.



Zoom-in



Elbow method



**Results**

We have compared R2 score of:

1. Linear Regression on K-means vs Linear Regression
2. Linear Regression on Hierarchical clustering vs Linear Regression
3. Random Forest on K-means clustering vs Random Forest
4. Random Forest on Hierarchical vs Random Forest.
5. Linear regression on DBSCAN.

**Dataset-1 Prediction of Millions per Gallon in Automobile**

R2 score for Linear regression on K-means is 0.8696 whereas Linear regression has an accuracy of 0.7928

R2 score for Linear regression on Hierarchical clustering is 0.8538 whereas Linear regression has an accuracy 0.7928

R2 score for Random Forest on K-means is 0.9109 whereas Random Forest has an accuracy of 0.9029

R2 score for Random Forest on Hierarchical clustering is 0.8912 whereas Random Forest has an accuracy of 0.9029

R2 score of linear regression on DBSCAN is –13.6118

**Dataset-2 Prediction of House price**

R2 score for Linear regression on K-means is 0.6385 whereas Linear regression has an accuracy of 0.6254

R2 score for Linear regression on Hierarchical clustering is 0.6437 whereas Linear regression has an accuracy 0.6254

R2 score for Random Forest on K-means is 0.8187 whereas Random Forest has an accuracy of 0.8165

R2 score for Random Forest on Hierarchical clustering is 0.8088 whereas Random Forest has an accuracy of 0.8165

R2 score of linear regression on DBSCAN is 0.6316

**Dataset-3 Prediction of Wine Quality**

R2 score for Linear regression on K-means is 0.3788 whereas Linear regression has an accuracy of 0.3594

R2 score for Linear regression on Hierarchical clustering is 0.3528 whereas Linear regression has an accuracy 0.3594

R2 score for Random Forest on K-means is 0.4514 whereas Linear regression has an accuracy of 0.4539

R2 score for Random Forest on Hierarchical clustering is 0.4375 whereas Random Forest has an accuracy of 0.4539

R2 score of linear regression on DBSCAN is –19.7241

**Analysis**

For Linear Regression on K-means and Hierarchical clustering methods showed higher accuracy compared to Linear Regression without clustering.

In prediction of automobile MPG and house prices using cluster-based regression seemed more efficient whereas for prediction of Wine Quality cluster-based regression methods showed poor performance over direct Linear Regression.

**Conclusion**

Our project mainly aims at the effectiveness of cluster-based regression models in improving the accuracy compared to normal linear regression model. By splitting the data into similar clusters using algorithms such as K-means clustering, Hierarchical clustering and DBSCAN we are able to divide the data into meaningful groups and apply separate linear regression models to each cluster.  
From the three data sets used , benefits of using this method are completely visible. For predicting MPG in automobiles, we can see from the above results that clustering – based regression models both hierarchical and K-means have achieved more accuracy than regular linear regression. When it comes to the case of predicting house price the cluster based regression model have shown more accuracy. When it comes to wine quality, looked further into the superiority of clustering-based models, particularly when using Random Forest with K-means and hierarchical.

The K-means clustering combined with linear regression and Random Forest models improved their performance by capturing unique patterns within each cluster. Similarly hierarchical clustering helped to identify the optimal clusters without predefined numbers. DBSCAN’s ability to handle noise and identify clusters of different shapes have also contributed to more accurate models, especially when it comes to datasets with varying densities. The results from our evaluations were measured through metrices like R-squared and the use of silhouette score and other evaluation matrices further validated the quality and effectiveness of the clustering techniques applied. The use of Random Forest models within the clusters also helped us gain additional accuracy, showing the importance of combining clustering with more complex regression techniques.

Overall the project comes to conclusion that clustering-based linear regression is a powerful approach for handling complex and heterogenous datasets. It enhances predictive performance by capturing unique relationships within clusters, making it effective for various real-world problems. Future work could explore the addition of more advances clustering algorithms and machine learning techniques to further enhance the predictive capabilities of this approach.