



# 

# Machine Learning and Data Mining

# 5DATA001C.2

# Course work Report

# 

**Module Leader’s Name- Mr. Nipuna Senanayaka**

**Name**: K.K. Isuru Kuruppu

**UOW ID**: w187045

**IIT ID**: 20210172

# Partitioning Clustering Part

## Objectives/Deliverables (partitioning clustering)

1st Subtask Objectives:

(a)

Scaling: - Scaling is the process of converting a dataset's numerical features to a standard scale. Due of the sensitivity of many machine learning algorithms to the magnitude of the input features, this is done. Results may be skewed if the features are not weighted equally; when this happens, some features will predominate over others. This bias can be avoided and the machine learning algorithm's performance can be enhanced by scaling the features. Scaling can also hasten the optimization process during training, hastening convergence and improving outcomes.

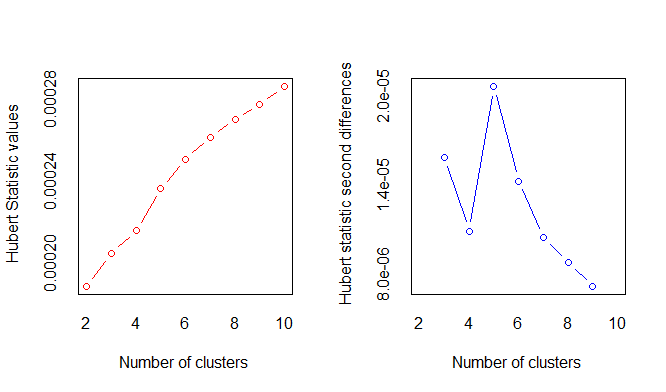
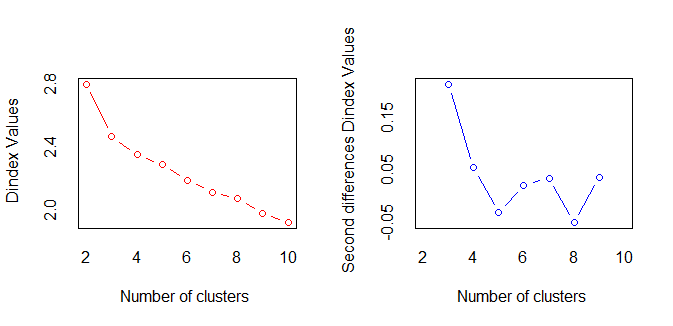
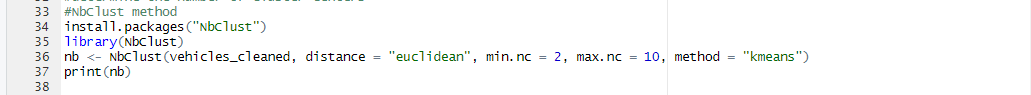
Outliers Detection: - Data points known as outliers differ significantly from other data points in the dataset. They may appear for a number of reasons, including measurement errors, data corruption, or just the basic nature of the data itself. Outliers can significantly affect how well machine learning algorithm’s function since they can distort the findings and produce false models. In order to improve data collecting and processing procedures, outliers can be used to discover data quality problems and abnormalities.

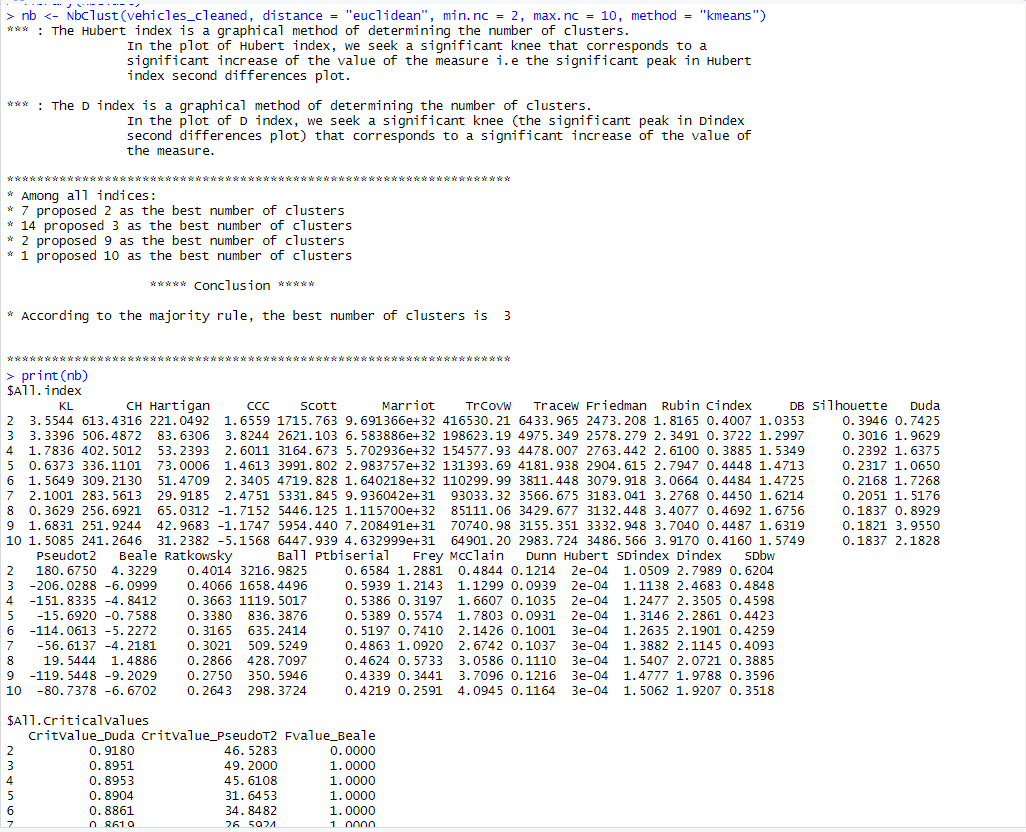
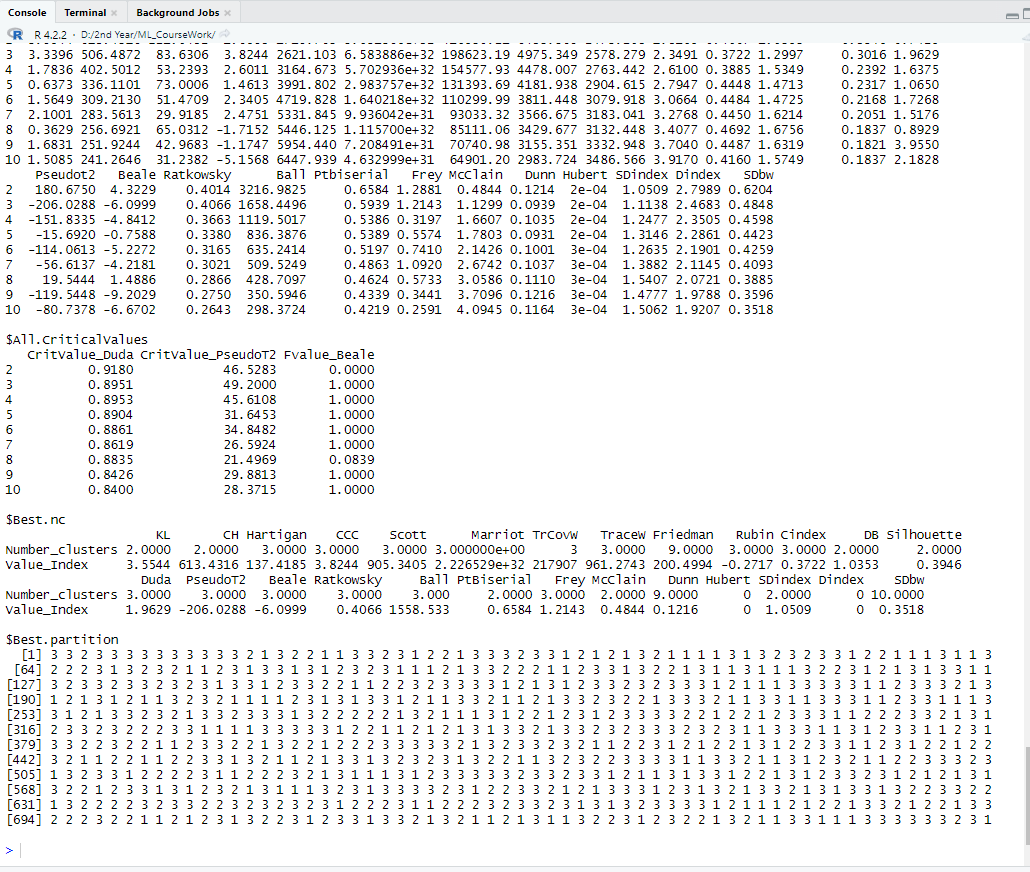
Removal: - To increase the model's accuracy and resilience, outliers can be eliminated. However, it's crucial to remember that eliminating too many outliers can also result in the loss of important data, so a careful balance must be struck. Removing outliers can aid in discovering and fixing problems with data collection and processing, which can assist to improve data quality and reliability. Additionally, eliminating outliers can aid in enhancing the model's interpretability by assisting in the discovery of data patterns and relationships that may have been hidden by the presence of outliers.

(b)

1. NbClust methods.

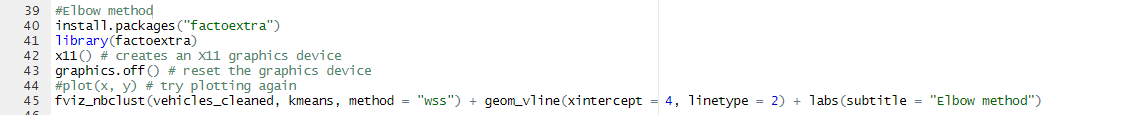
The R package NbClust is used to determine the ideal number of clusters to include in a dataset. It offers a number of indices and techniques, including the Silhouette coefficient, Calinski-Harabasz index, Dunn index, and Gap statistic, to assess clustering solutions. In order to facilitate the selecting process, it also offers graphical representations of the indexes.

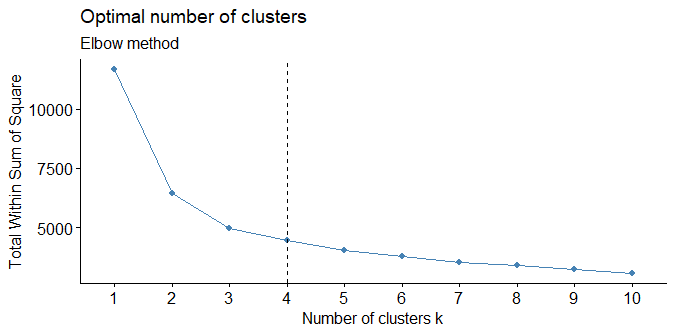




2.Elbow methods.

Based on the idea that as the number of clusters increases, the within-cluster sum of squares (WCSS) drops and the between-cluster sum of squares (BCSS) increases, the Elbow approach is a heuristic methodology used to determine the ideal number of clusters in a clustering algorithm. Plotting the WCSS versus the number of clusters and choosing the number of clusters where the rate of WCSS decline begins to level off are both required. The Elbow method can be combined with other techniques to conduct a more thorough study.

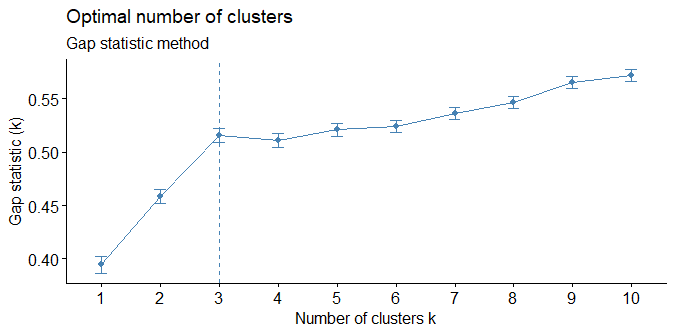




3. Gap statistics methods.

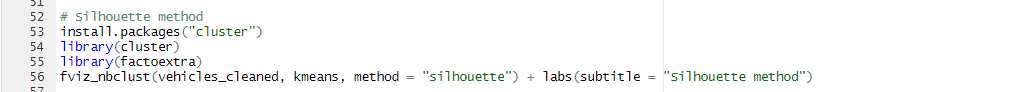
The ideal number of clusters in a dataset can be found using the statistical method known as gap statistics. It compares the overall within-cluster variation with what would be predicted under a null reference distribution for various values of k (number of clusters). The amount of k that maximizes the gap statistic, or the amount of clustering that is most noticeably improved above the null reference distribution, is the optimal number of clusters.

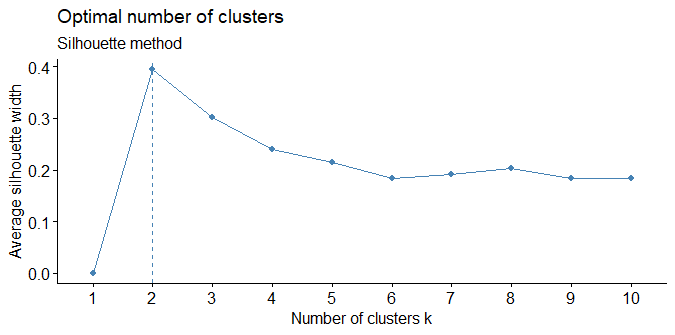




4.silhouette methods.

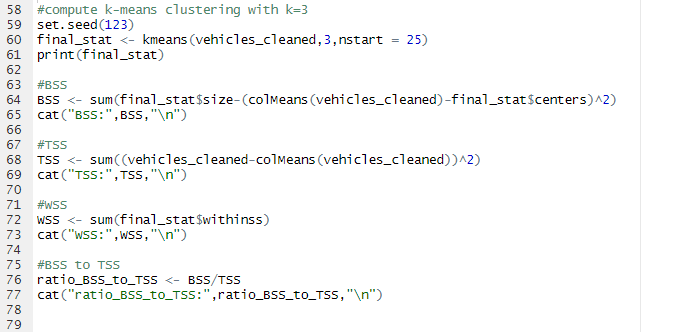
When comparing an observation to other clusters, the silhouette method, a clustering evaluation technique, determines how similar the observation is to its own cluster. For each observation, it generates a silhouette coefficient that spans from -1 to 1. The overall average is generated for the entire dataset and includes the average silhouette coefficient for each observation in each cluster.

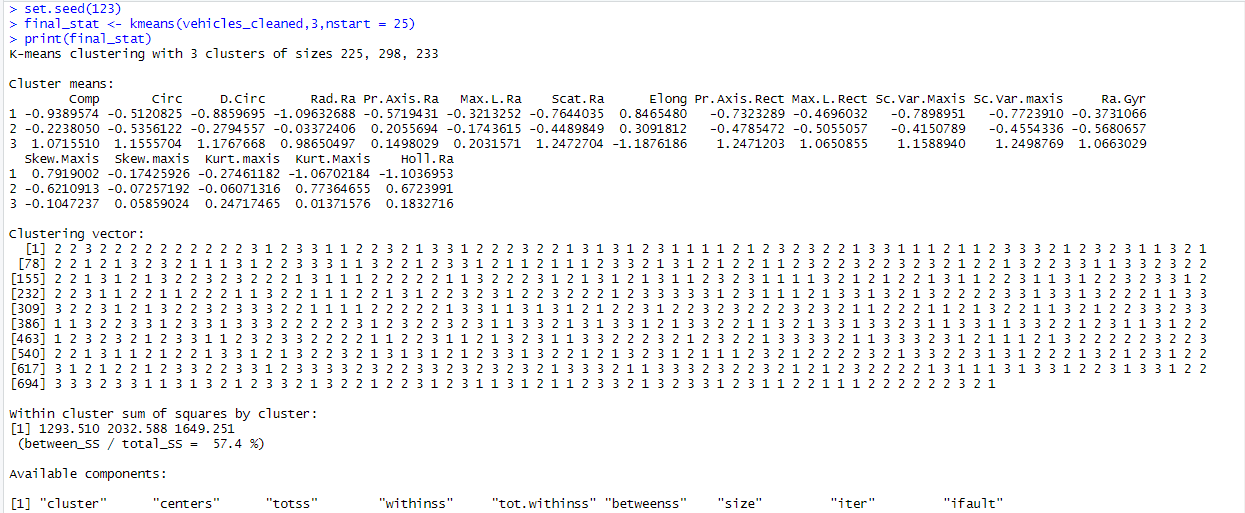


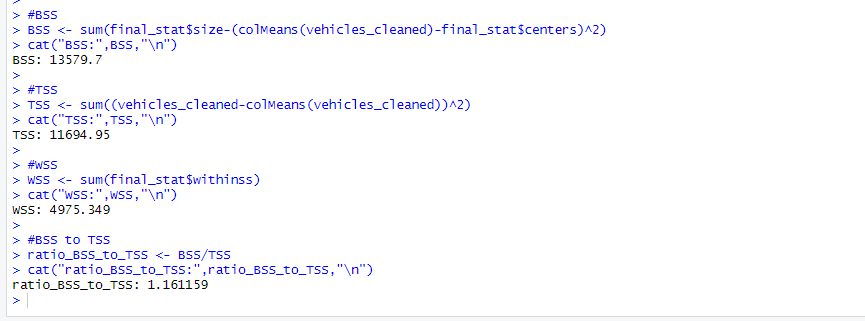


(c)

Elbow methods are useful for identifying the optimal number of clusters but have limitations, silhouette methods are useful for determining the quality of the clusters, NbClust methods are useful for determining the optimal number of clusters and the best clustering method, and gap statistics methods are useful for identifying the optimal number of clusters but can be computationally intensive. The particular requirements of the investigation and the features of the dataset will determine which clustering validation technique is used.

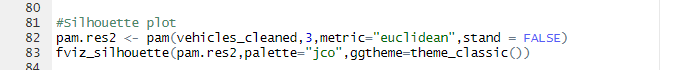
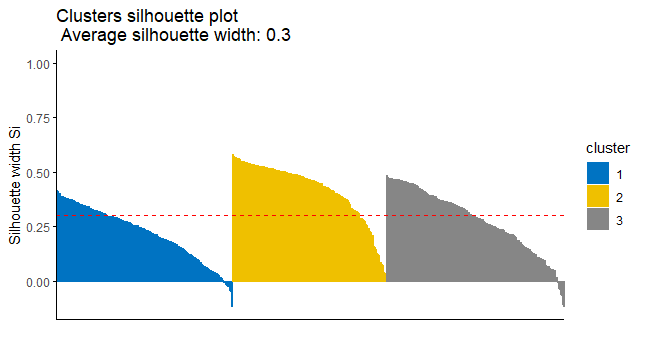


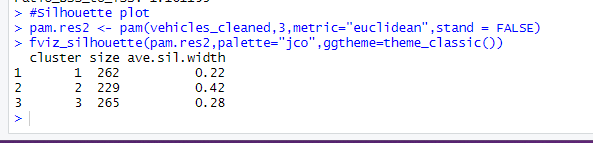


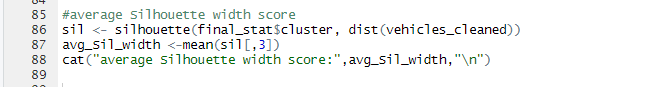


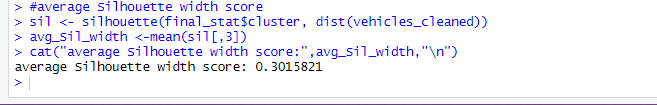
(d)

ytruyruru





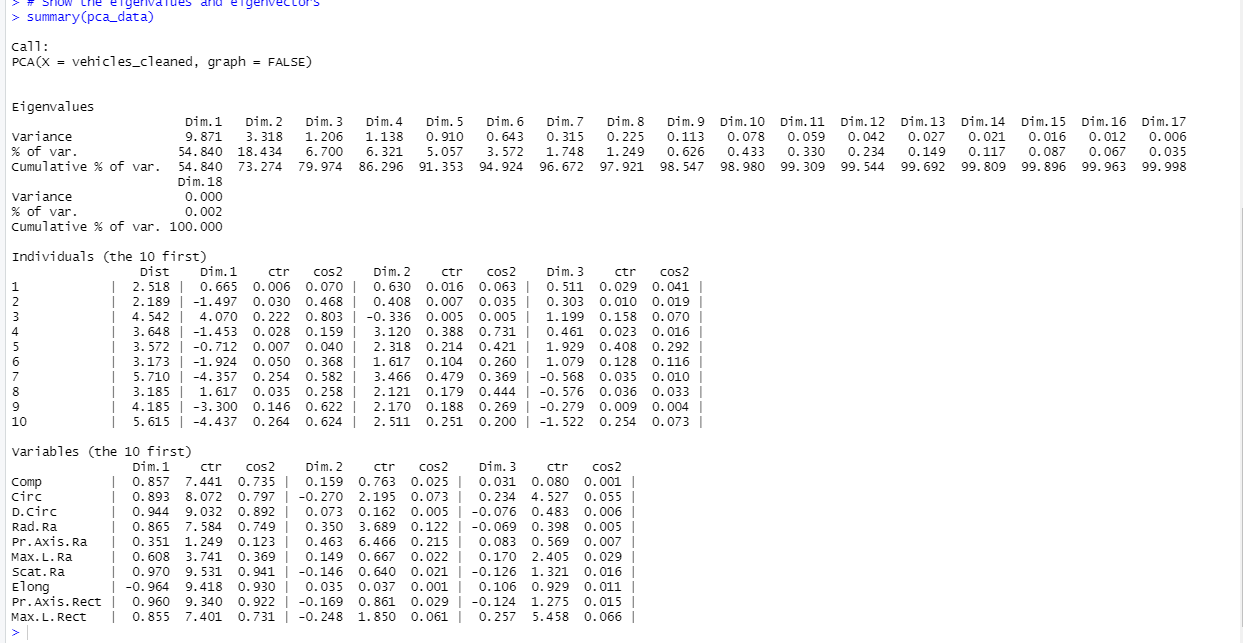




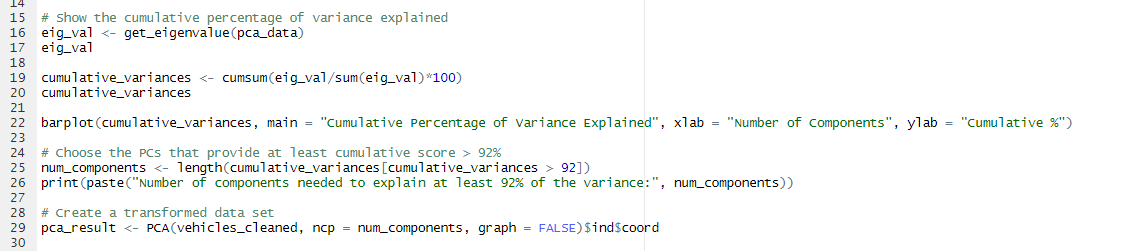
2nd Subtask Objectives:

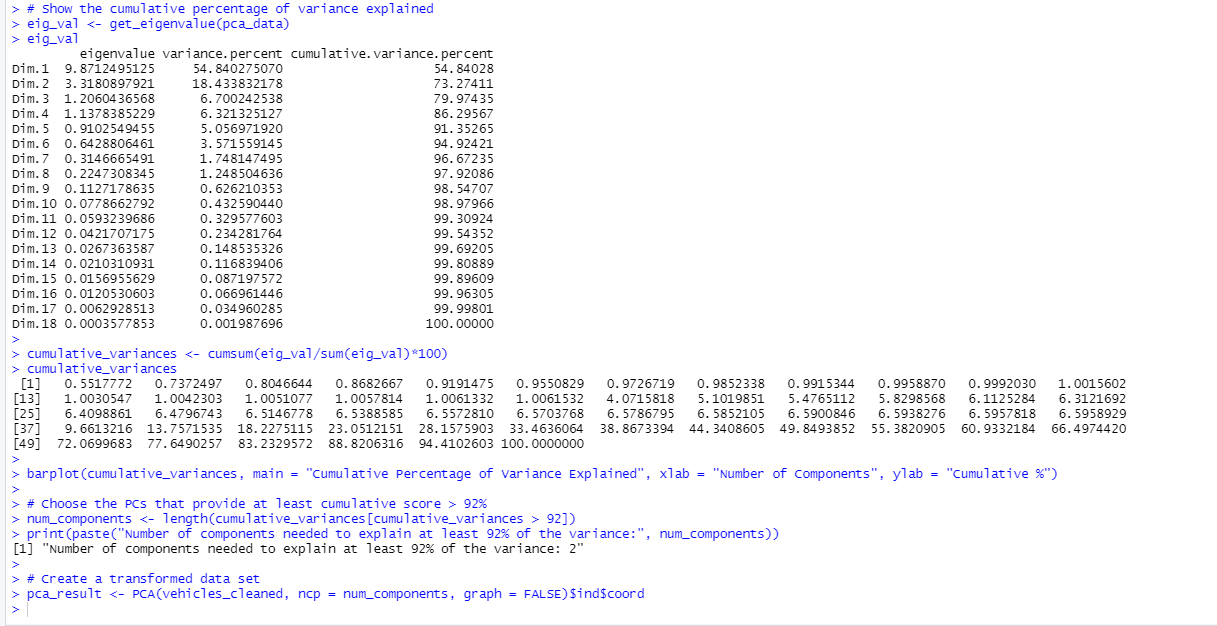
(e) The amount of variance explained by the main components and the complexity of the model must be balanced when deciding how many principal components to keep. In this instance, we decided to keep the eight major components out of the original 18 features that provided a total score of at least 92%. This decision strikes a compromise between the complexity of the modified dataset and the amount of variance that the model can explain.





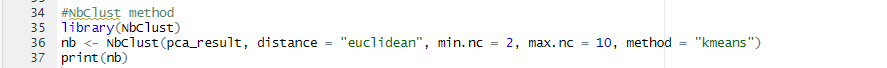
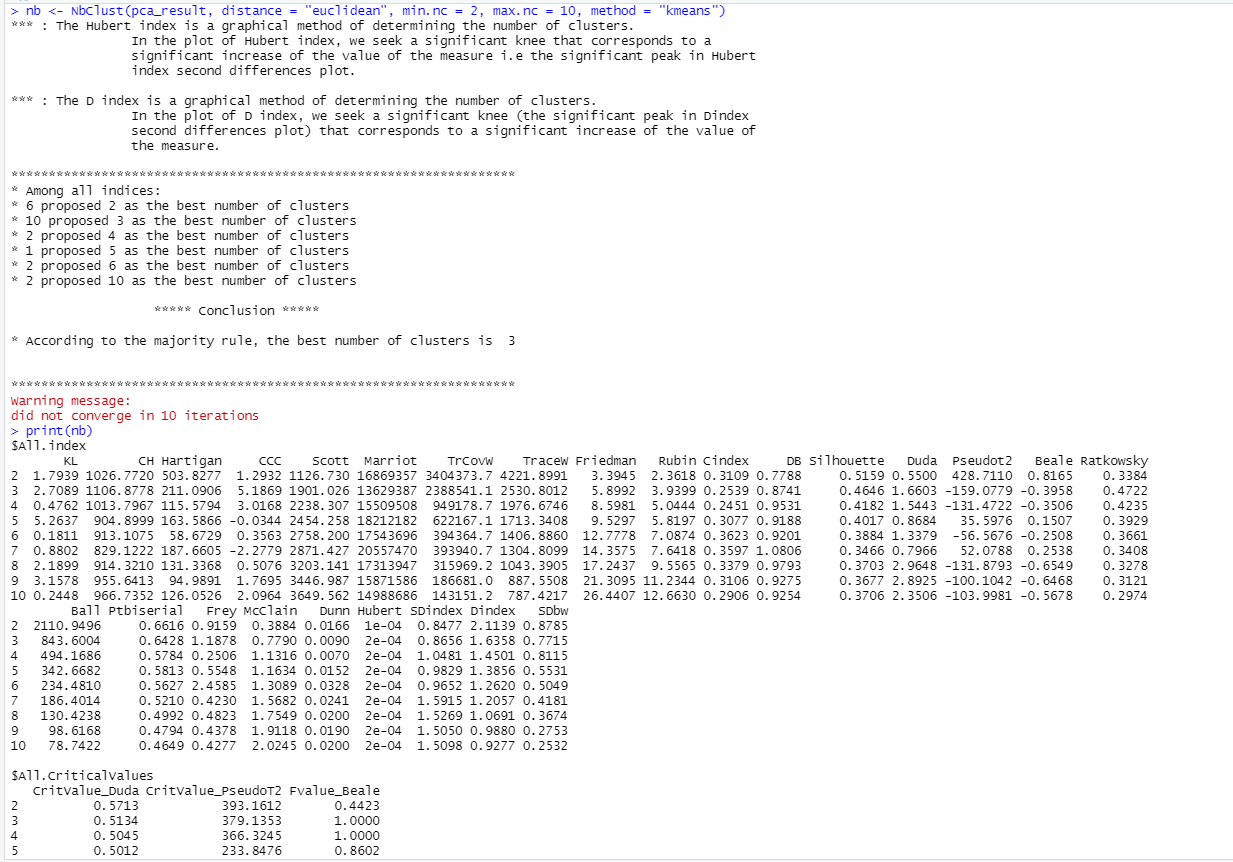
cumulative score per principals

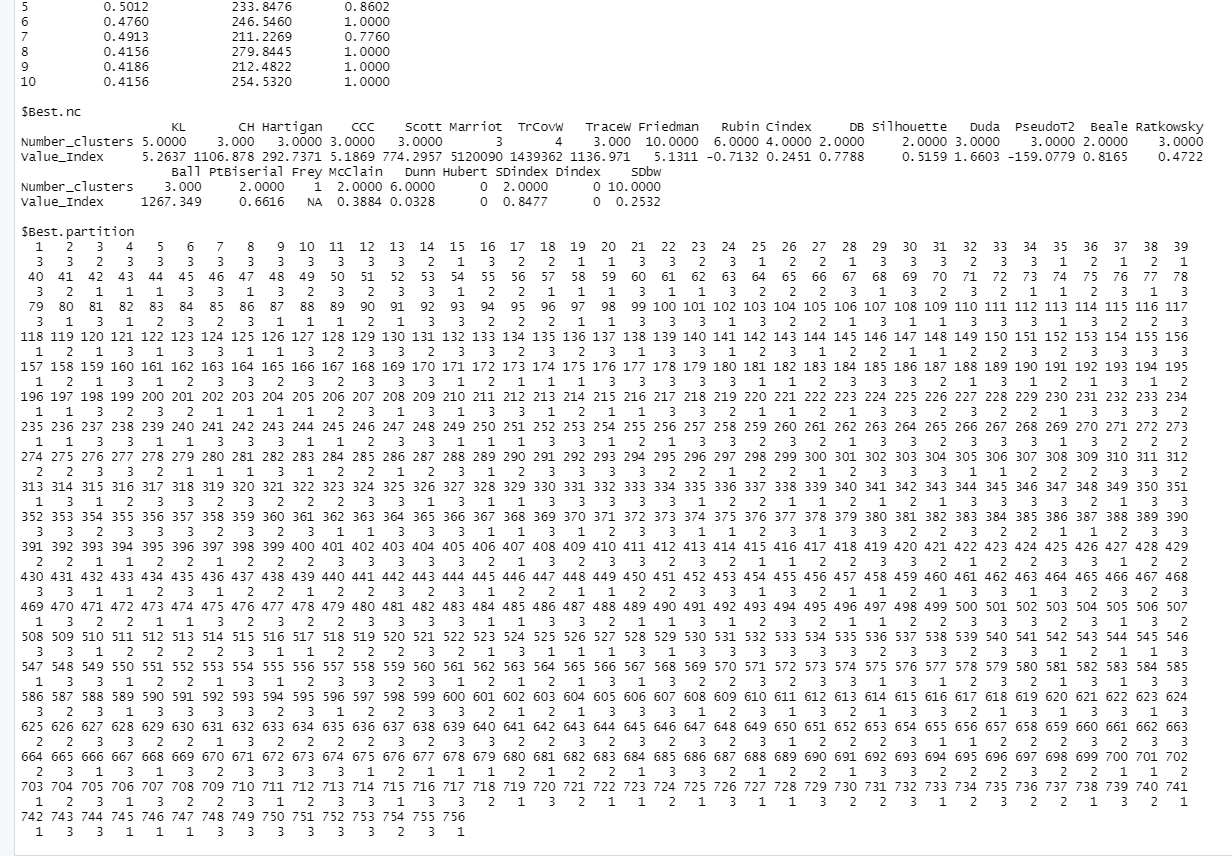
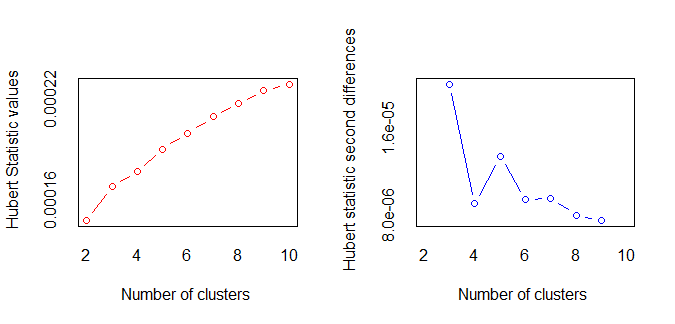


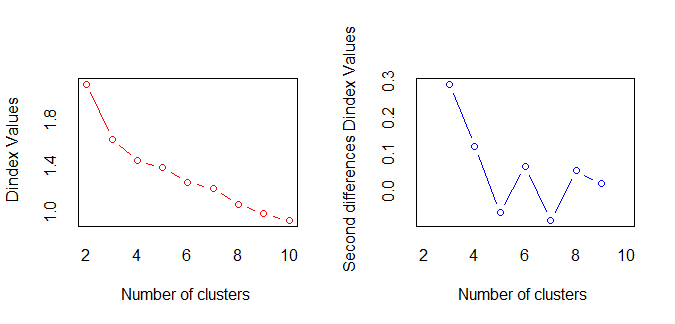


(f)

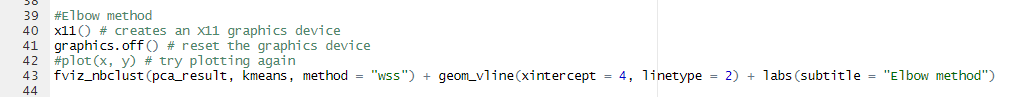
1.NbClust methods.

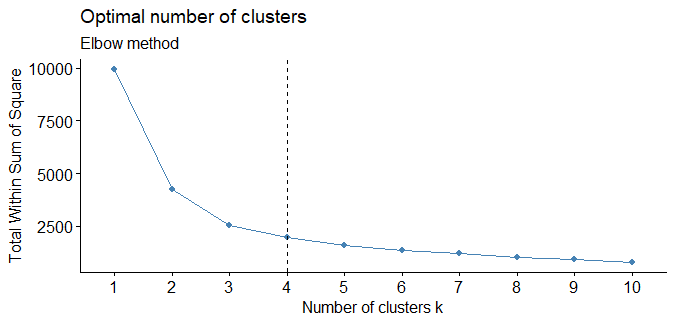






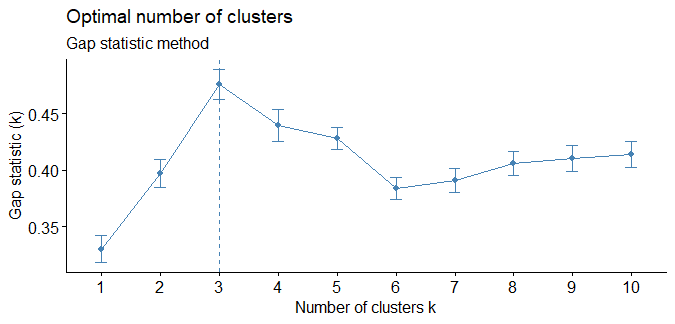
2.Elbow methods.



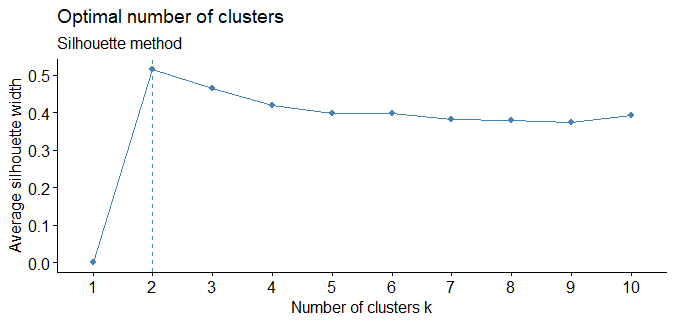


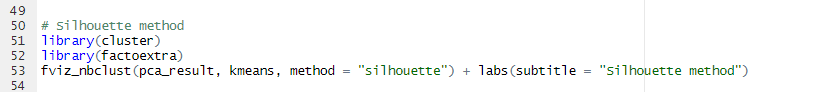
3.Gap statistics methods



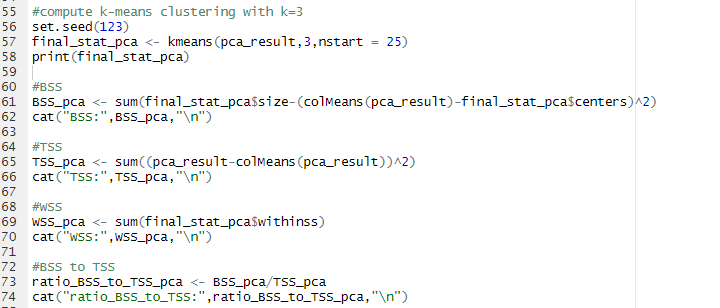


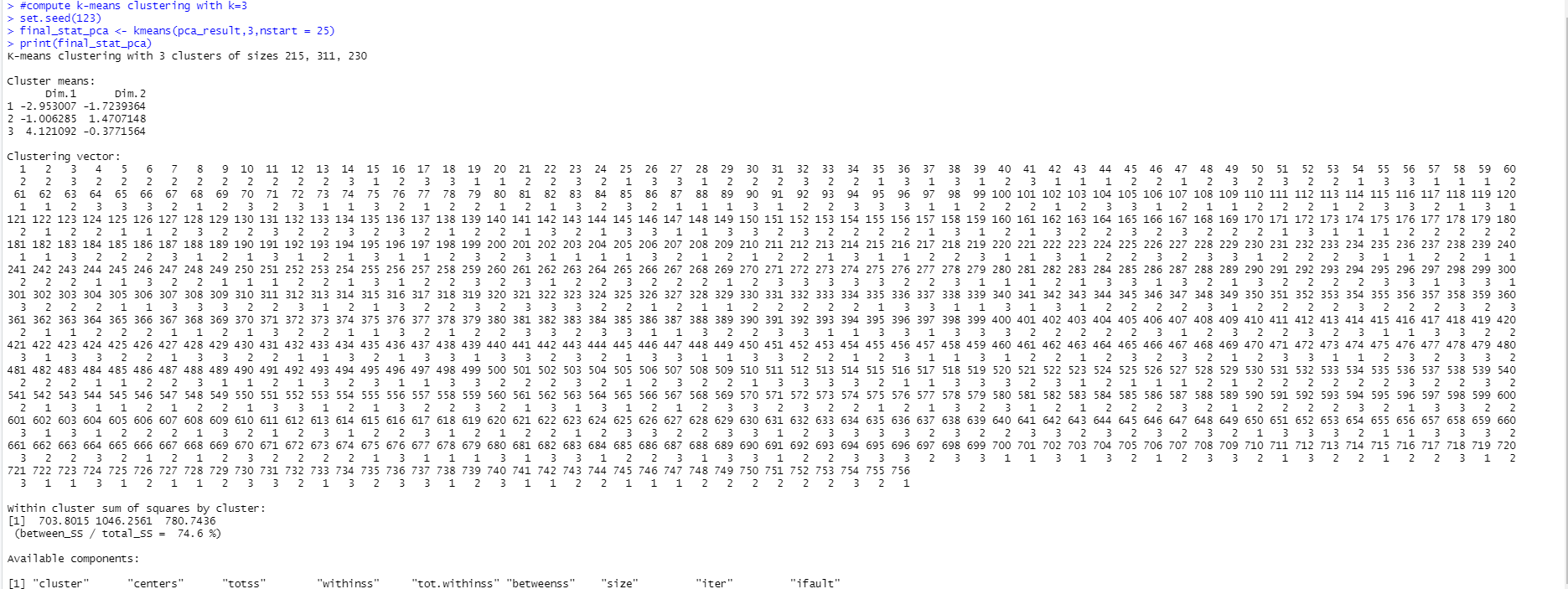
4.silhouette methods.

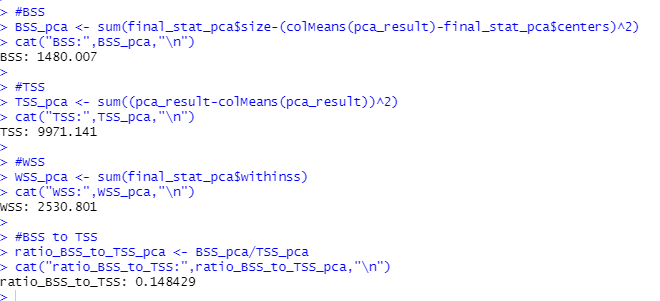


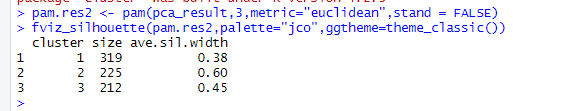


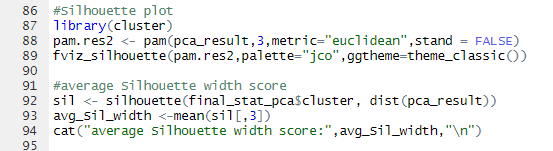
(g)

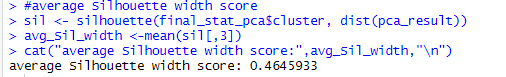


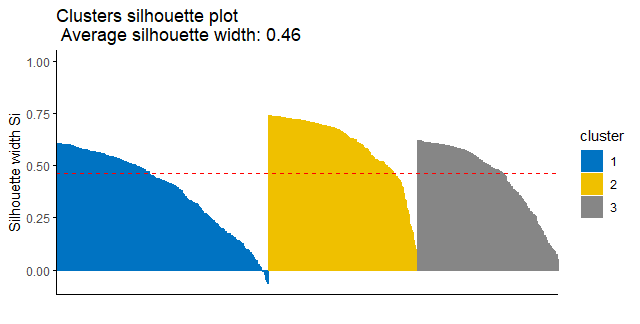


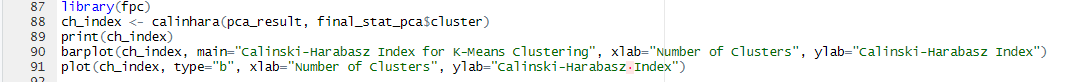


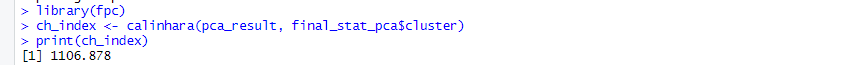
(h)







(i) Calinski-Harabasz Index



# Energy Forecasting Part (part of Work Based Learning Activity)

## Objectives/Deliverables (Multi-layers Neural Network)

1st Subtask Objectives:

(a)

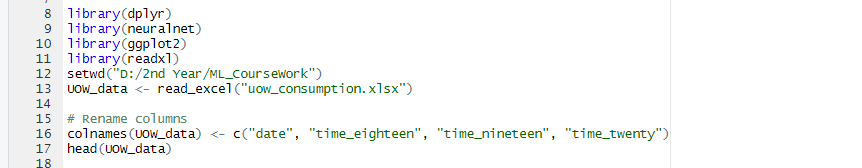
**Time-based Approach: -** Time-based characteristics can be an effective tool for defining the input vector when dealing with predicting issues for electricity load. By incorporating elements like the day of the week, the hour of the day, and the month of the year, the neural network can detect seasonality and temporal trends in the data. If the dataset contains hourly electricity consumption data for several years, the neural network, for example, can use time-based characteristics to distinguish between weekday and weekend patterns, morning and evening peaks, and seasonal swings in demand.

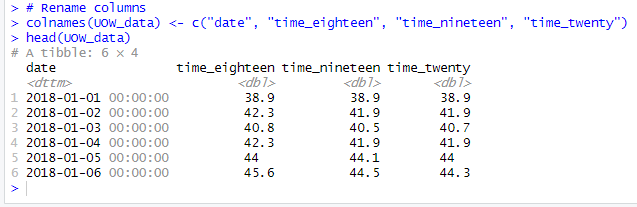
**Calendar-based Approach: -** In this method, the input variables are calendar elements like the day of the week, the month of the year, and the holiday indicators. This strategy acknowledges that the amount of electricity consumed has a seasonal rhythm and is influenced by occasions on the calendar, such as vacations. This strategy has been demonstrated to increase the precision of load forecasting models.

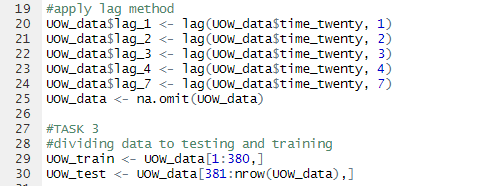
**Weather-based Approach: -** This method makes use of weather input variables like temperature, humidity, and wind speed. This strategy acknowledges how the demand for electricity, particularly for heating and cooling, is influenced by the weather. This strategy has also been demonstrated to increase the precision of load forecasting models.

**Economic-based Approach: -** Economic input factors used in this strategy include GDP, employment, and industrial production. This strategy acknowledges how the demand for electricity is influenced by the economy, particularly in the industrial and commercial sectors. This strategy has also been demonstrated to increase the precision of load forecasting models.

(b)



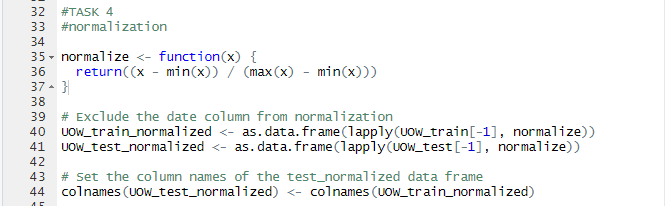


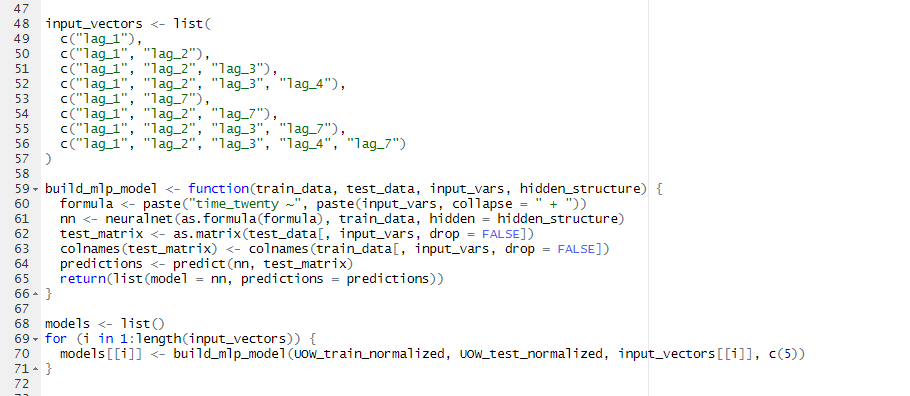


(c)

Normalization

normalizing data before using them in an MLP structure can help improve the convergence, avoid bias, and improve the performance of the MLP. Normalization is a standard pre-processing step in machine learning, and it is important to carefully choose the normalization method that is appropriate for the specific problem and data set.



(d)

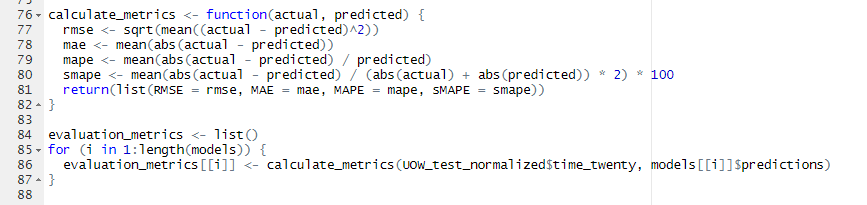
(e)

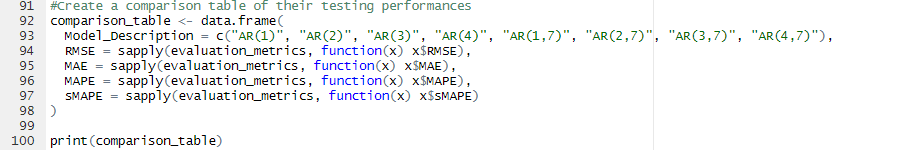
**RMSE: -** The average difference between predicted values and actual values is measured by the term "RMSE," which stands for root mean squared error. The square root of the average of the squared differences between the predicted and actual values is used to calculate RMSE. To assess the precision of predictions, regression analysis frequently uses RMSE.

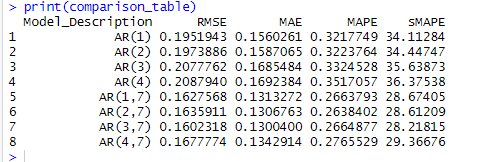
**MAE: -** MAE, or mean absolute error, is another way to quantify the average difference between projected and actual data. The average of the absolute discrepancies between projected and actual values is used to determine MAE. To assess the precision of predictions in regression analysis, MAE is frequently utilized.

**MAPE: -** Mean Absolute Percentage Error, or MAPE, is a measurement of the typical percentage difference between predicted values and actual values. By averaging the absolute percentage deviations between projected and actual values, MAPE is calculated. When the scale of the data varies significantly from prediction to prediction, MAPE is frequently used to assess the accuracy of forecasts.

**sMAPE: -** A modification of MAPE that overcomes some of its drawbacks is called sMAPE, or symmetric MAPE. sMAPE accounts for the magnitude of the projected and actual values when calculating the average percentage difference between predicted and actual values. The problem of MAPE creating infinite or undefined values when the actual values are zero is helped by this.

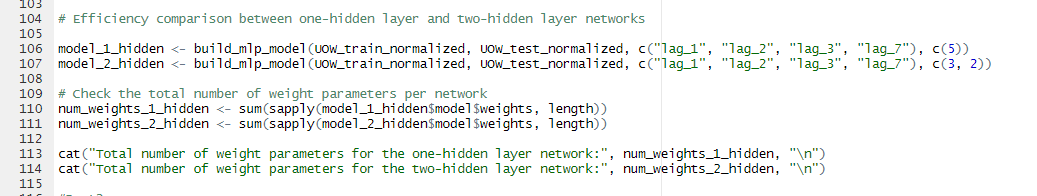
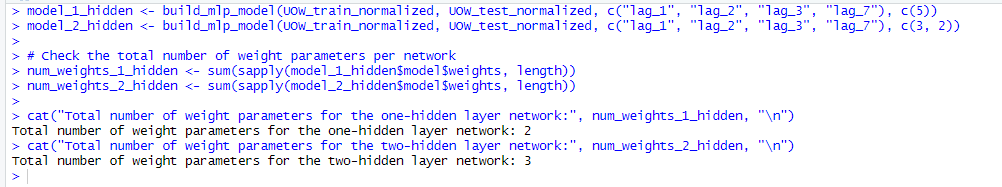
****

(d)



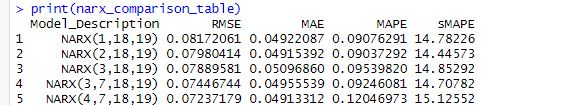
(g)

The one-hidden layer network in this situation has fewer weight parameters (2 vs. 3) than the two-hidden layer network. As a result, the one-hidden layer network uses fewer weight parameters than other networks.

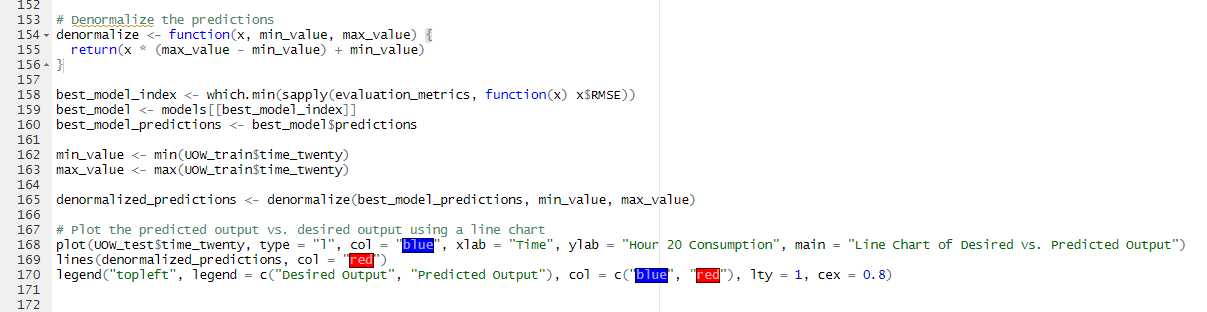


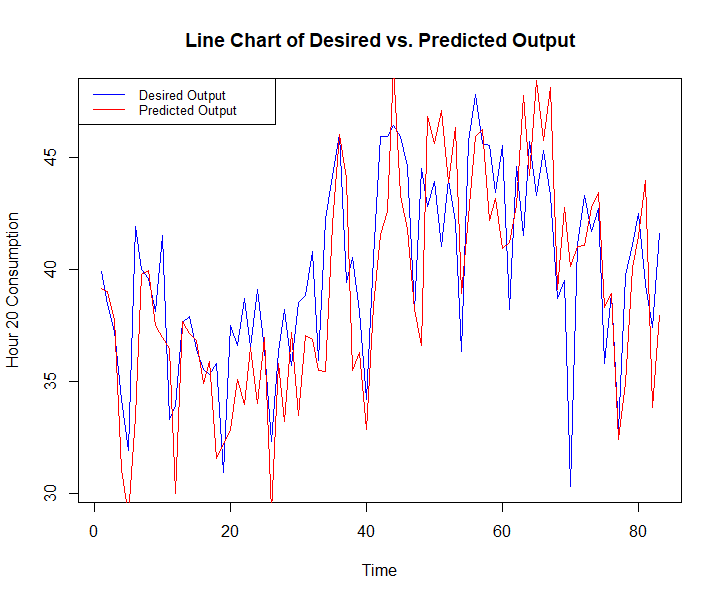
2nd Subtask Objectives:

(h)

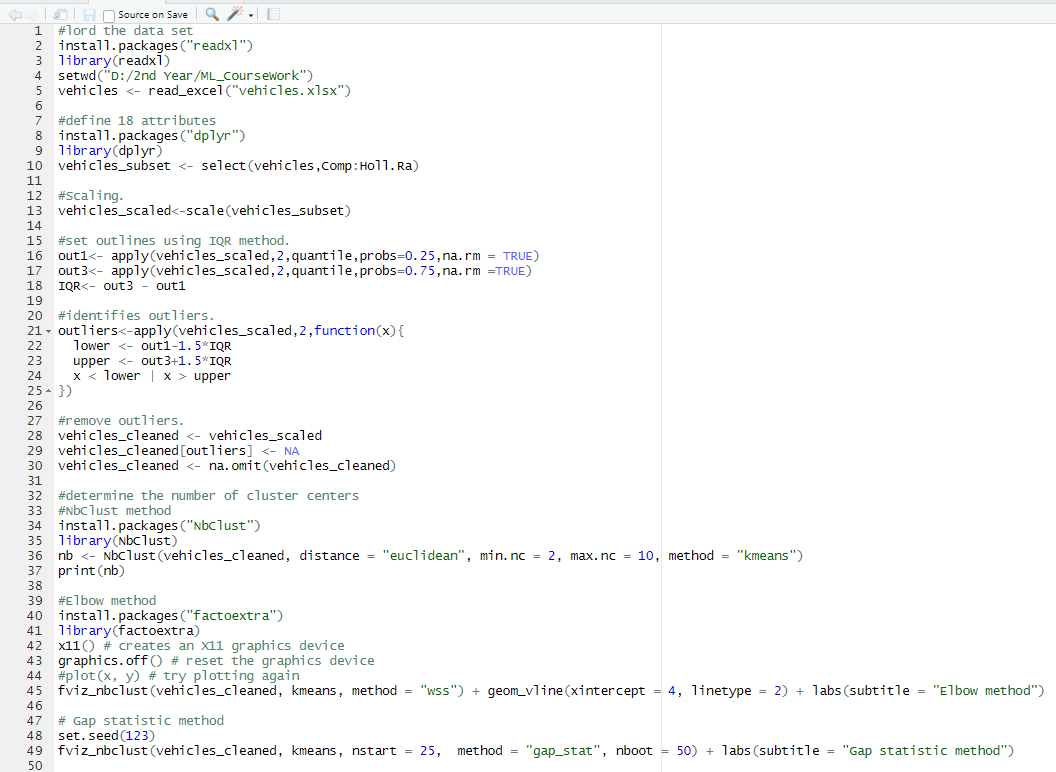


(i)

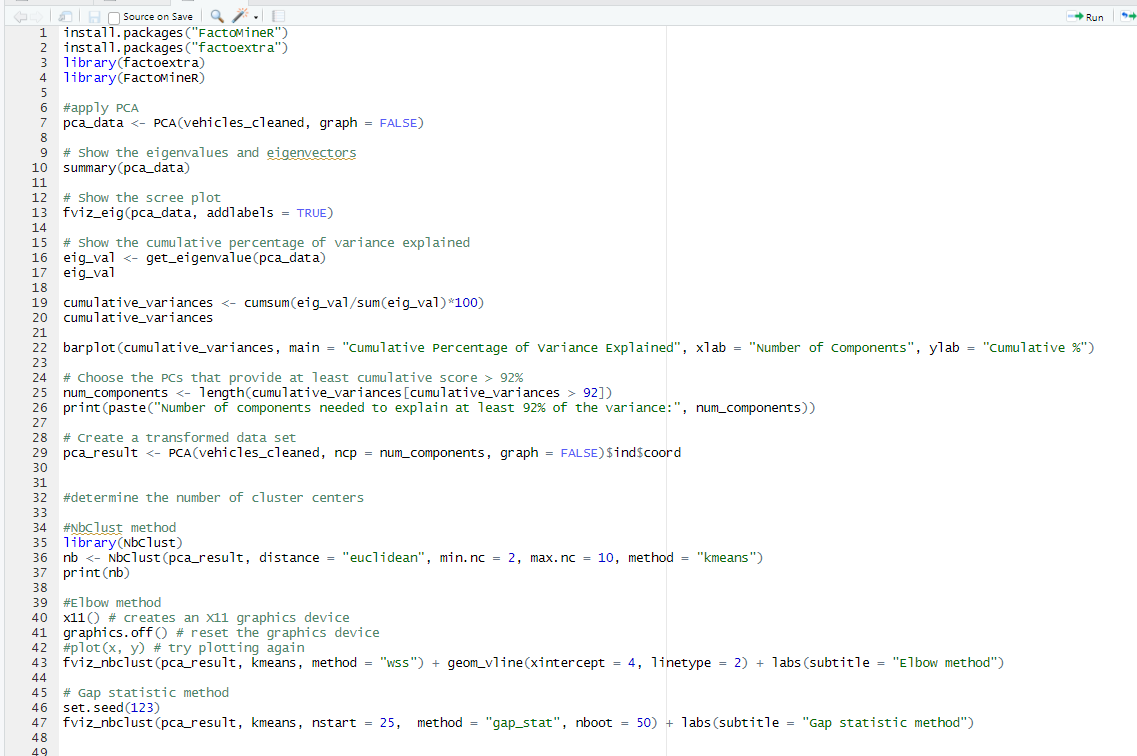


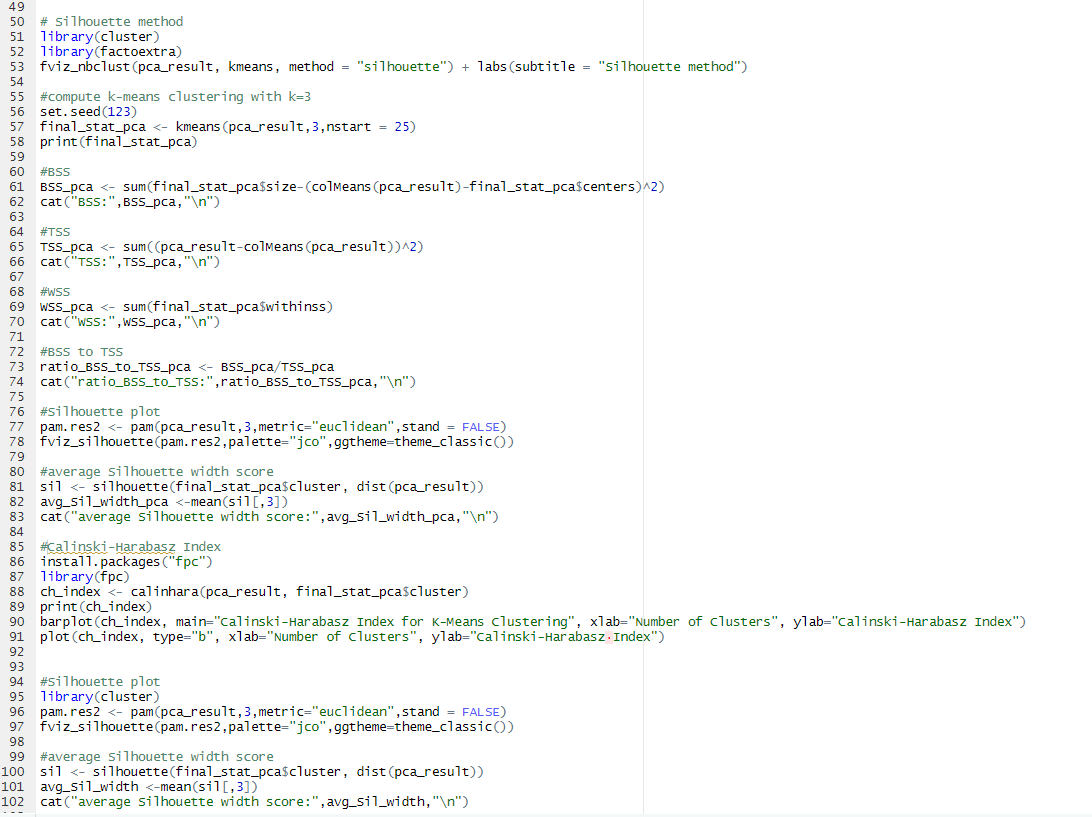


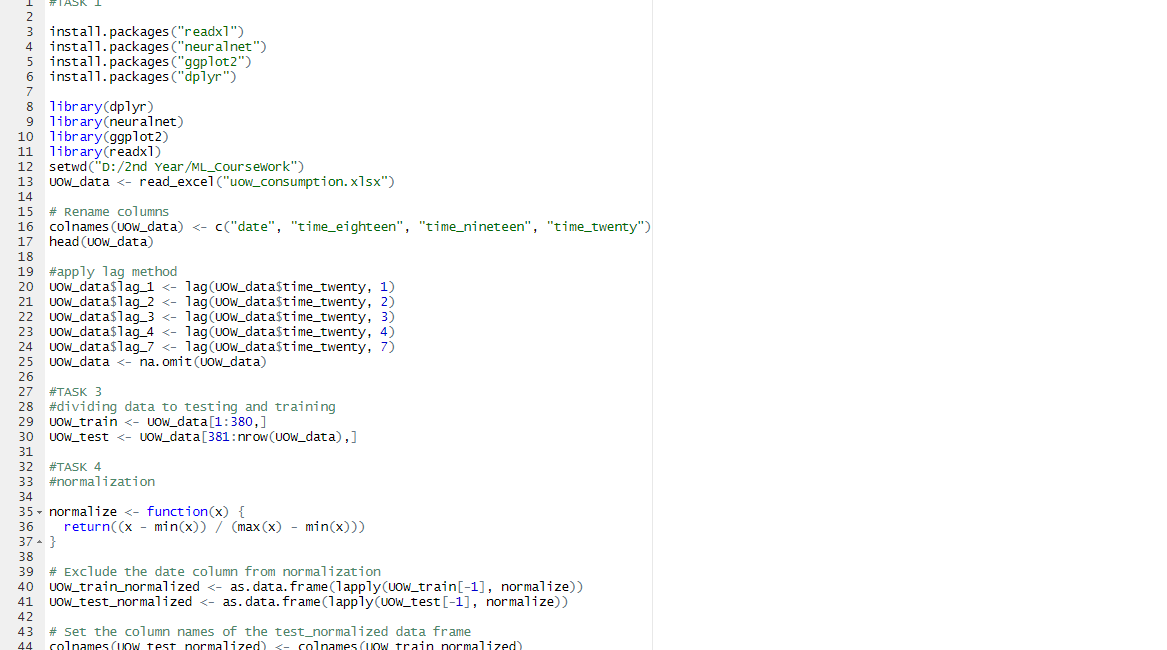
**Appendix**

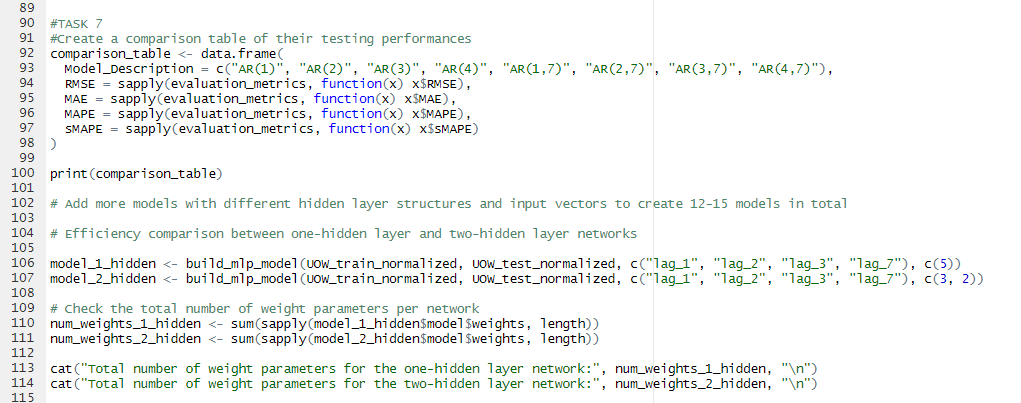
****Part 1 sub task 1,

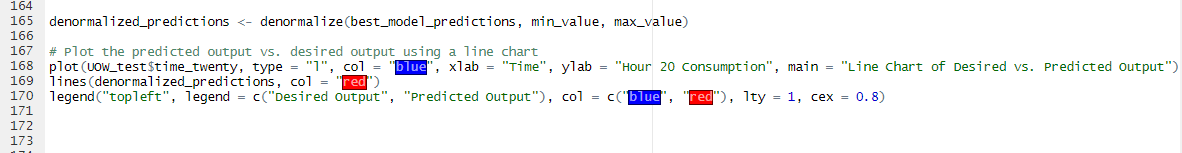
****

Part 1 sub task 2,



Part 2





References

Implementation with R language - ScienceDirect. (no date). Available from https://www.sciencedirect.com/science/article/abs/pii/B9780128131275000072?via%3Dihub [Accessed 4 May 2023].

K-Means Clustering: Concepts and Implementation in R for Data Science | by Maria Gulzar | Towards Data Science. (no date). Available from https://towardsdatascience.com/k-means-clustering-concepts-and-implementation-in-r-for-data-science-32cae6a3ceba [Accessed 4 May 2023].

Performance · Advanced R. (no date). Available from http://adv-r.had.co.nz/Performance.html [Accessed 4 May 2023].