Department of Statistics, SPPU

ST-O13 Statistical Learning and Data Mining

Internal-2 Assignment 2023-24

Name: 1) Devendra Sanjay Patil - 2234

2) Prajakta Pandurang Madane - 2231

Title: "Unsupervised Learning Analysis of PMPML Depot Performance"

1. Introduction

- **Background:**

Public transportation is the backbone of any urban city, and its efficiency directly impacts the lives of its residents. In Pune, the Pune Mahanagar Parivahan Mahamandal Limited (PMPML) provides public transport buses for the Pune and Pimpri Chinchwad region. With numerous depots spread across the city, managing and optimizing the performance of these depots is a challenging task.

This project aims to leverage unsupervised learning algorithms to analyze the monthly statistics of PMPML depots for the year 2021. The objective is to identify patterns, similarities, and differences among the depots based on various parameters and across different quarters.

The project will also focus on identifying outliers with respect to any of the variables or performance measures in any of the quarters. This could provide valuable insights into any anomalies or exceptional cases in the data.

The analysis will be carried out in a systematic manner, starting with data cleaning and preprocessing, followed by the application of suitable unsupervised learning algorithms. The findings from this project could potentially help in improving the efficiency and performance of PMPML depots.

- **Objectives**

The primary objective of this project is to apply unsupervised learning algorithms to analyze the monthly statistics of PMPML depots for the year 2021. The specific objectives are as follows:

1. **Data Cleaning and Preprocessing**: To prepare the data for analysis by cleaning and preprocessing the depot-related monthly statistics.
2. **Depot Similarity Analysis**: To identify which depots are similar to each other based on various parameters and whether they remain similar across all the quarters.
3. **Performance Evaluation**: To evaluate the performance of the depots based on various parameters in different quarters and identify which depots are performing better or worse.
4. **Outlier Detection**: To detect any outliers with respect to any of the variables or performance measures in any of the quarters.

2. Data Cleaning and Preprocessing

- **Data Cleaning:**

The dataset provided for this project comprises monthly statistics for the year 2021, distributed across 12 distinct Excel files. Upon initial examination, it is observed that the dataset contains numerous variables, some of which may not be relevant to our analysis. Additionally, there are variables with a significant number of missing values, which could potentially impact the accuracy of our results.

To ensure the integrity and reliability of our analysis, a crucial step in our data preprocessing will involve a thorough review and cleaning of these variables. This will include the removal of unnecessary variables and those with a high proportion of missing values. We will focus on retaining variables that are appropriate for our analysis and have fewer missing values. This rigorous data cleaning process will help in enhancing the quality of our dataset, thereby leading to more accurate and meaningful insights from our unsupervised learning algorithms.

- **Data Preprocessing:**

Upon completion of the data cleaning process, the next step involves preparing the data for analysis. This includes addressing the issue of missing values in the dataset. For any month that has missing values, an imputation method will be employed where these missing values will be replaced with the average values of the remaining months. This approach helps in maintaining the overall distribution and integrity of the data.

Following the imputation process, the data will be restructured to facilitate our analysis. Two distinct datasets will be created - one that provides a yearly overview and another that offers a quarterly perspective.

The yearly dataset will be a consolidated view of all the months, with each variable represented by its average value across the year. This dataset will provide a broad overview of the depot performance over the entire year.

The quarterly dataset, on the other hand, will be segmented into four parts, each representing a quarter. Each variable in this dataset will be the average value for that quarter. This dataset will allow us to observe trends and patterns on a quarterly basis, providing more granular insights into the depot performance.

These steps ensure that our data is not only clean but also structured in a way that is most suitable for our analysis. This meticulous data preparation process sets the stage for the effective application of unsupervised learning algorithms in the subsequent stages of our project.

In our data there are 13 depots are "Swargate" "N T Wadi" "Kothrud" "Katraj" "Hadapsar" "Marketyard" "P Station" "Nigadi" "Bhosari" "Pimpri" “Bhekarainagar” "Shewalwadi" "Balewadi", but “Bhekarainagar” has too much missing values for every month so removed that depot from the data to make analysis unbiased.

3. Temporal Analysis of Depot Similarities

Our analysis will begin with the examination of the yearly data. We will assess the similarities among the depots based on the selected variables. To achieve this, we will employ clustering methods, specifically Hierarchical Clustering and K-Means Clustering.

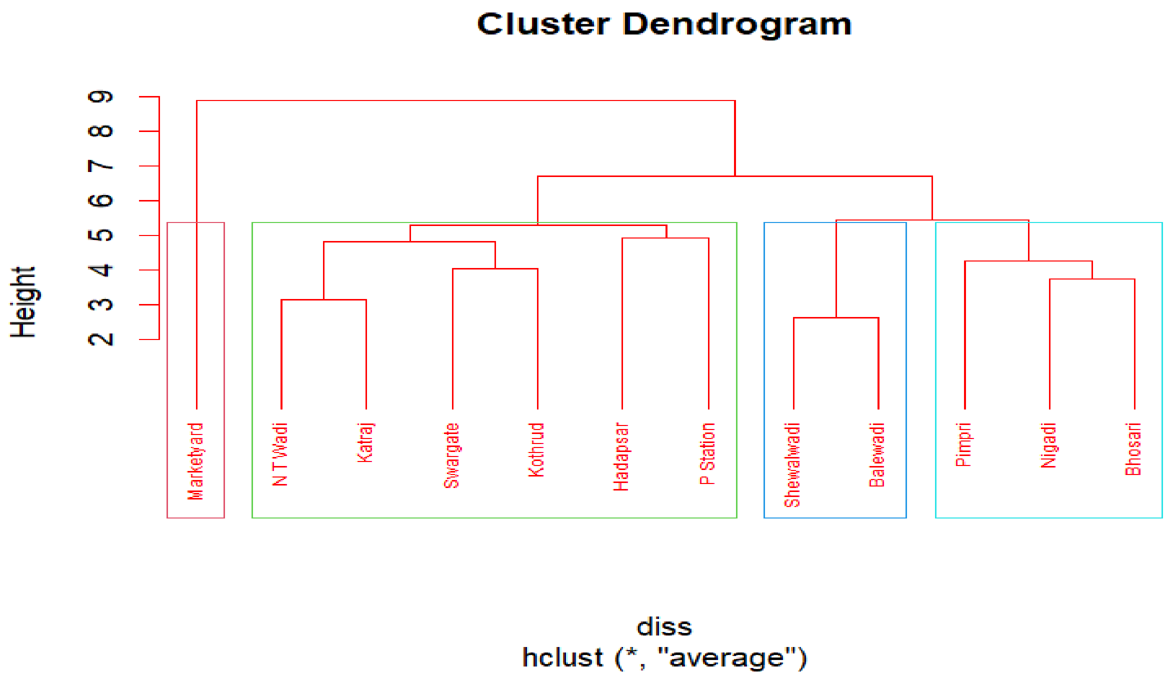
Hierarchical Clustering is an algorithm that groups similar objects into clusters. In this project, we will specifically use the Complete and Average linkage methods in Hierarchical Clustering. The Complete or Maximum linkage method uses the maximum distances between all pairs of observations between two sets. The Average linkage method uses the average of the distances of each observation of the two sets.

On the other hand, K-Means Clustering is a method that aims to partition n observations into k clusters in which each observation belongs to the cluster with the nearest mean.

These clustering methods will allow us to identify which depots exhibit similar characteristics based on the selected parameters. By doing so, we can gain valuable insights into the performance and operational patterns of these depots.

Analysis on Yearly data

- **Hierarchical Clustering:**



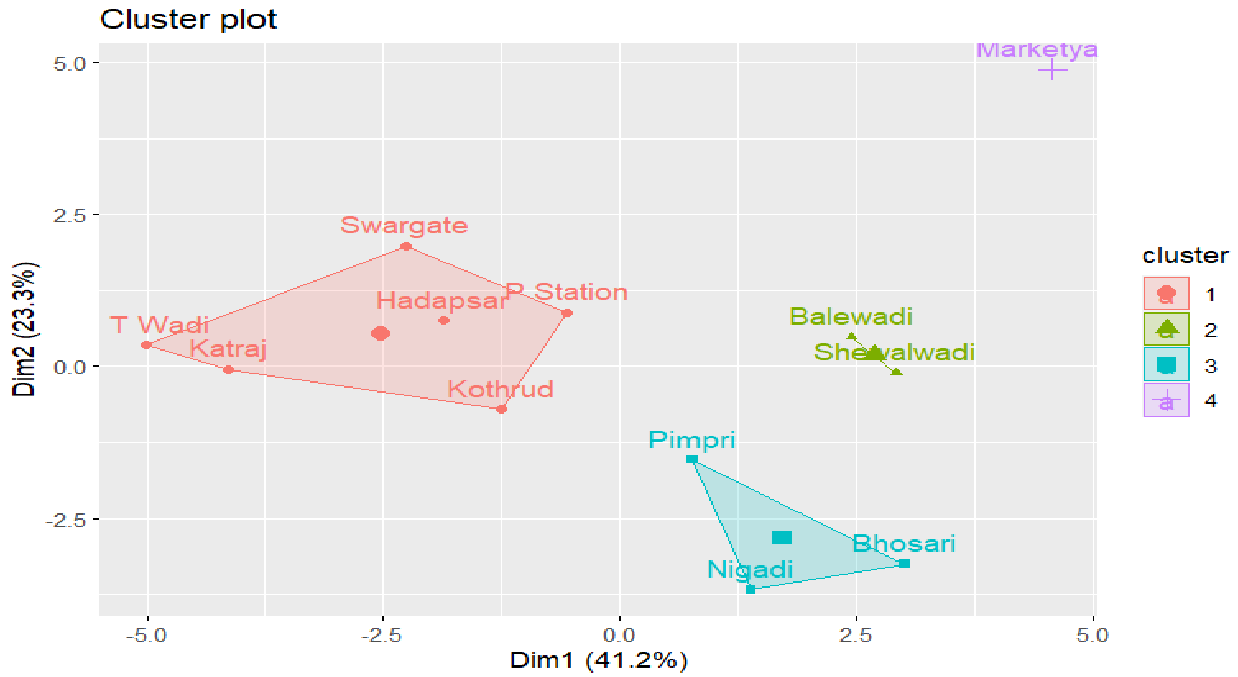
From the Hierarchical clustering analysis, we observe distinct groupings among the depots based on their operational characteristics:

1. **NT Wadi, Katraj, Swargate, Kothrud, Hadapsar, and P Station** form one cluster, indicating a high degree of similarity in their operational parameters.
2. **Shewalwadi and Balewadi** form another cluster, suggesting similar patterns in their operations.
3. **Pimpri, Nigadi, and Bhosari** also show similarity, constituting another cluster.

Interestingly, **Marketyard** does not align closely with any other depot, suggesting unique operational characteristics that set it apart from the rest.

These findings provide valuable insights into the operational dynamics of the PMPML depots. Understanding these similarities can help in benchmarking performance, sharing best practices, and driving operational efficiencies across depots.

- **K-means clustering:**

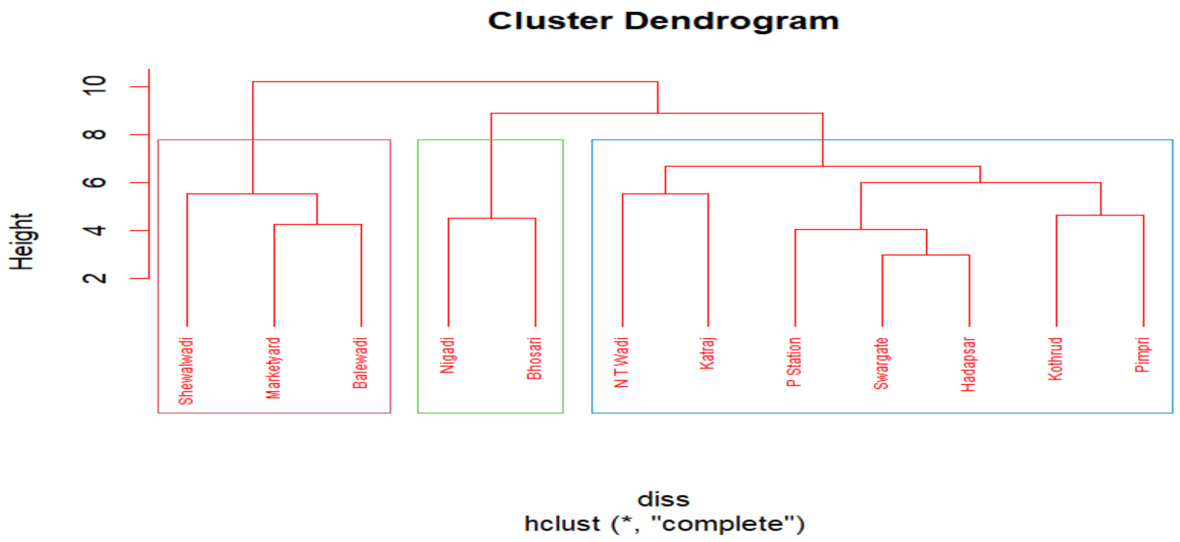


Same clusters we can see by using the k-means clustering method.

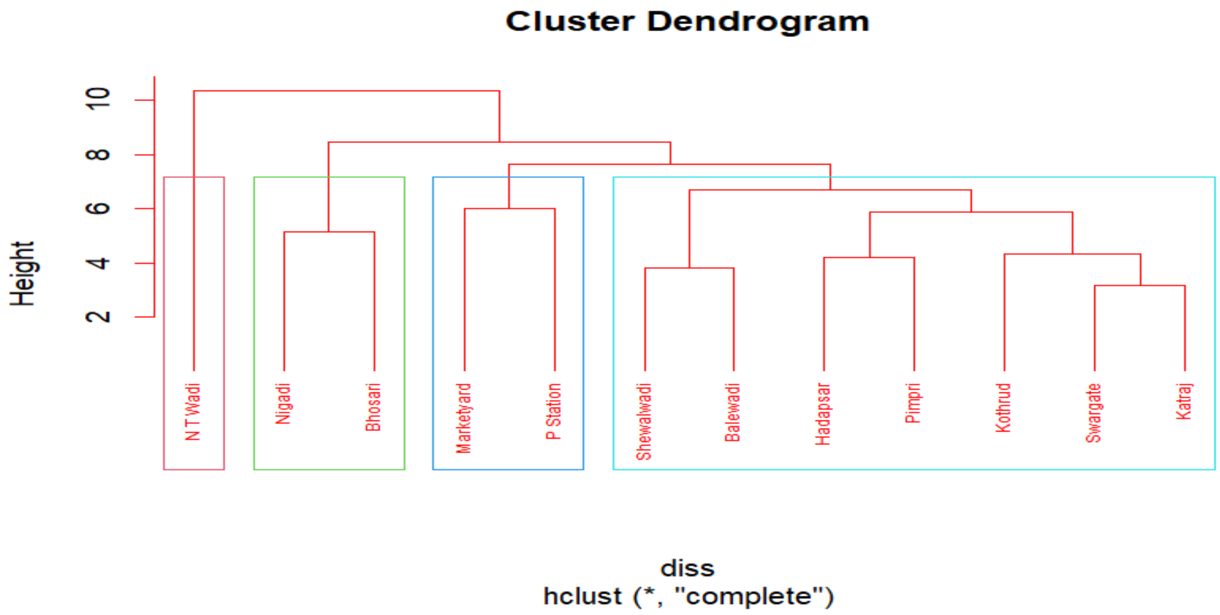
Analysis on Quarterly Data

- **Hierarchical Clustering**

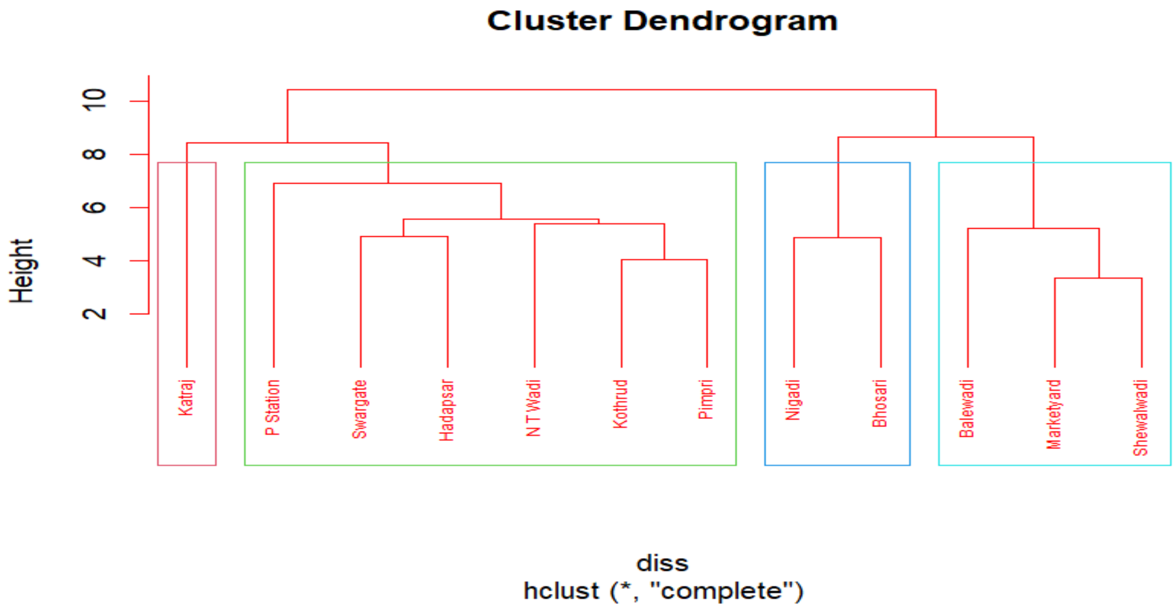
1. **Quarter one:**



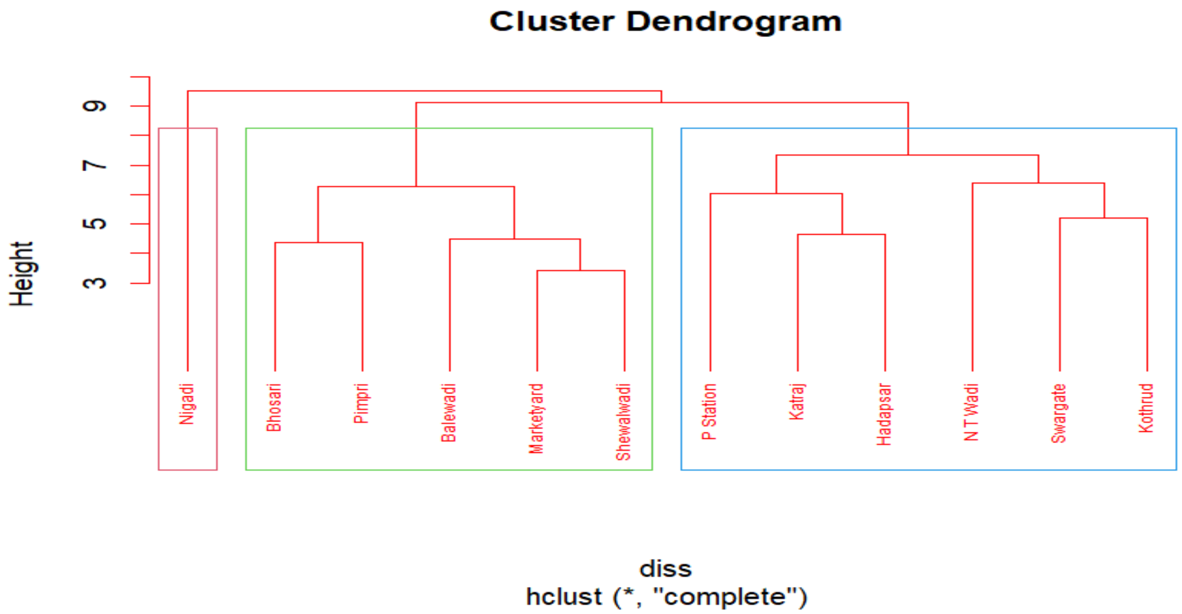
2.Quarter two:



3.Quarter three:



4.Quarter four:



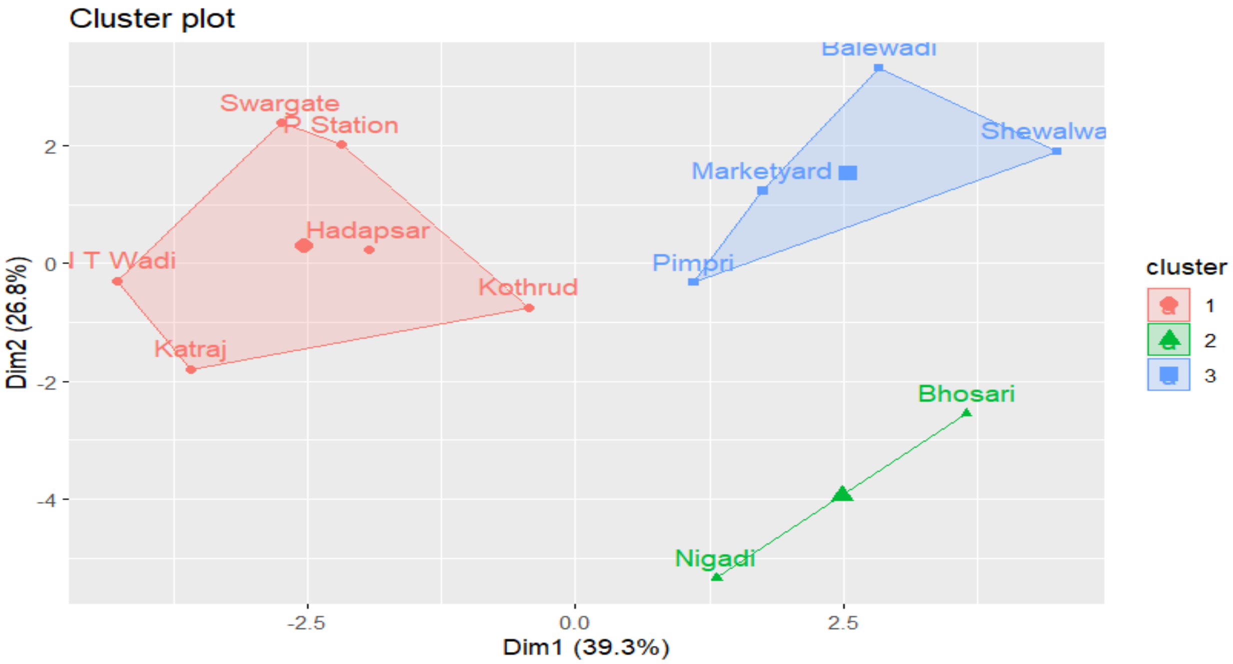
From the dendrogram analysis conducted for all quarters, it is observed that the behavior of the depots, based on the selected variables, remains largely consistent across each quarter. The clusters formed are similar across all quarters, indicating that the operational patterns of certain depots remain stable over time.

However, it’s important to note that this consistency is not universal across all depots. While some depots exhibit a degree of similarity in their operational characteristics across different quarters, others show more variability. These differences suggest that the performance and operational dynamics of certain depots are influenced by a variety of factors that may change significantly from quarter to quarter.

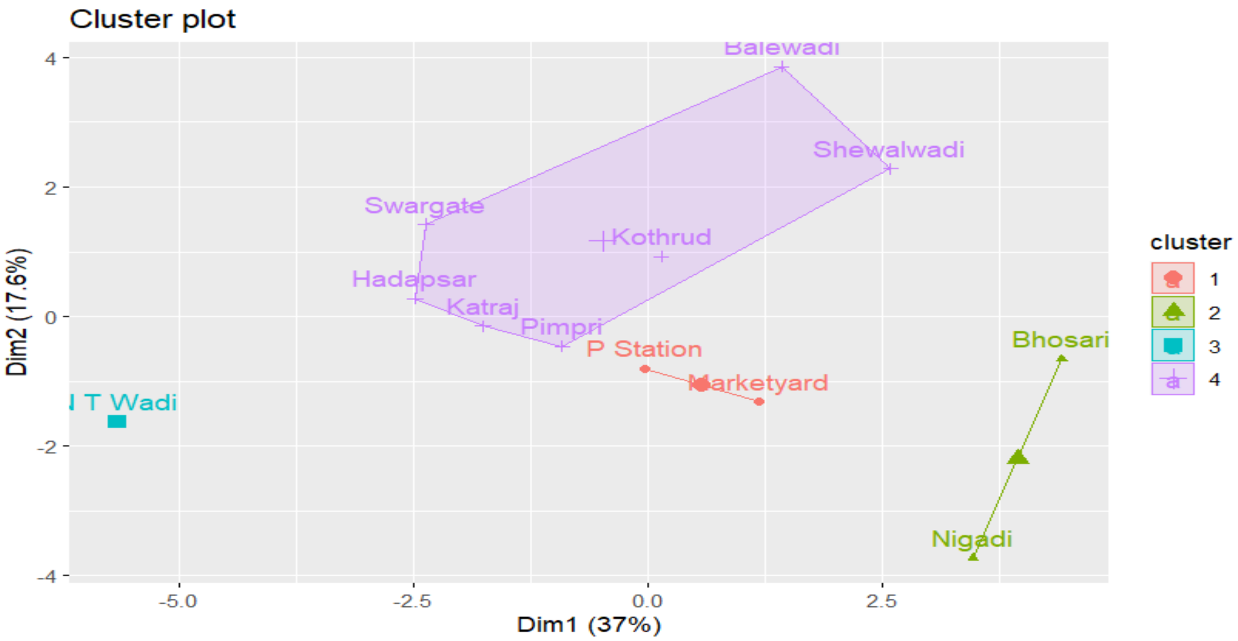
This mix of temporal stability and variability in depot operations underscores the complexity of managing public transportation systems. It highlights the need for both consistent strategies for those depots showing stability, and dynamic, adaptive strategies for those showing variability, in order to optimize depot performance

- **K-Means Clustering**

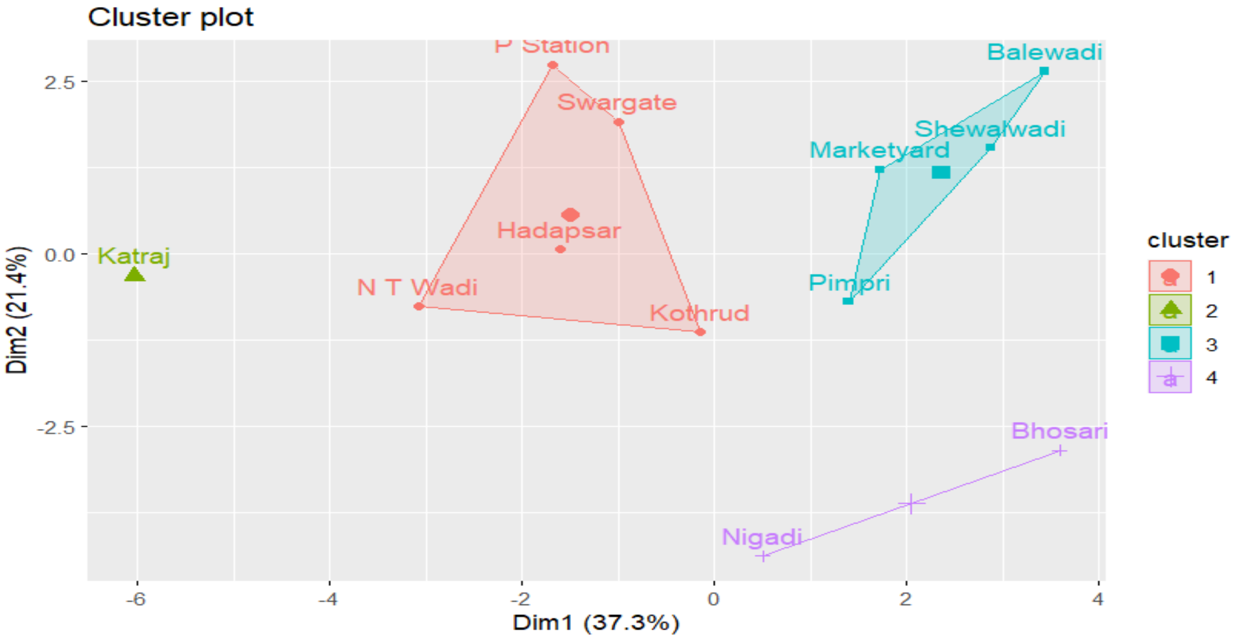
1. **Quarter One**



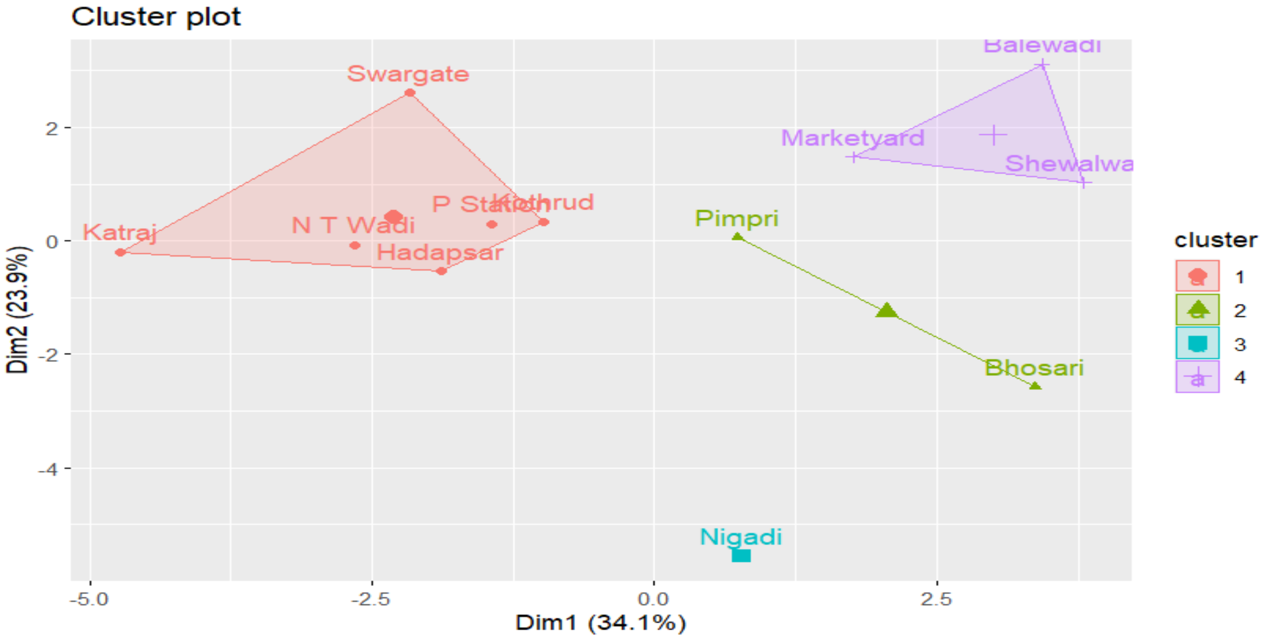
2. **Quarter Two**



3. Quarter Three



4. Quarter Four



From the K-means clustering plots generated for all four quarters, we observe certain consistencies in the behavior of the depots across different quarters. Despite some variations, the clusters formed exhibit a degree of similarity, suggesting that certain depots maintain consistent operational patterns throughout the year.

However, it’s important to note that not all depots follow this trend. Some depots show distinct behaviors in each quarter, indicating a dynamic operational pattern influenced by various factors that may change from quarter to quarter.

This analysis underscores the value of temporal data in understanding the operational dynamics of the depots. It also highlights the potential for using such insights to inform strategies for depot management and performance optimization.

4. Performance Evaluation of Depots Across Quarters

Certainly, here’s the description without hyperlinks:

The Technique for Order of Preference by Similarity to Ideal Solution (TOPSIS) is a multi-criteria decision analysis method. It was originally developed by Ching-Lai Hwang and Yoon in 1981, with further developments by Yoon in 1987, and Hwang, Lai, and Liu in 1993.

TOPSIS is based on the concept that the chosen alternative should have the shortest geometric distance from the positive ideal solution (PIS) and the longest geometric distance from the negative ideal solution (NIS).

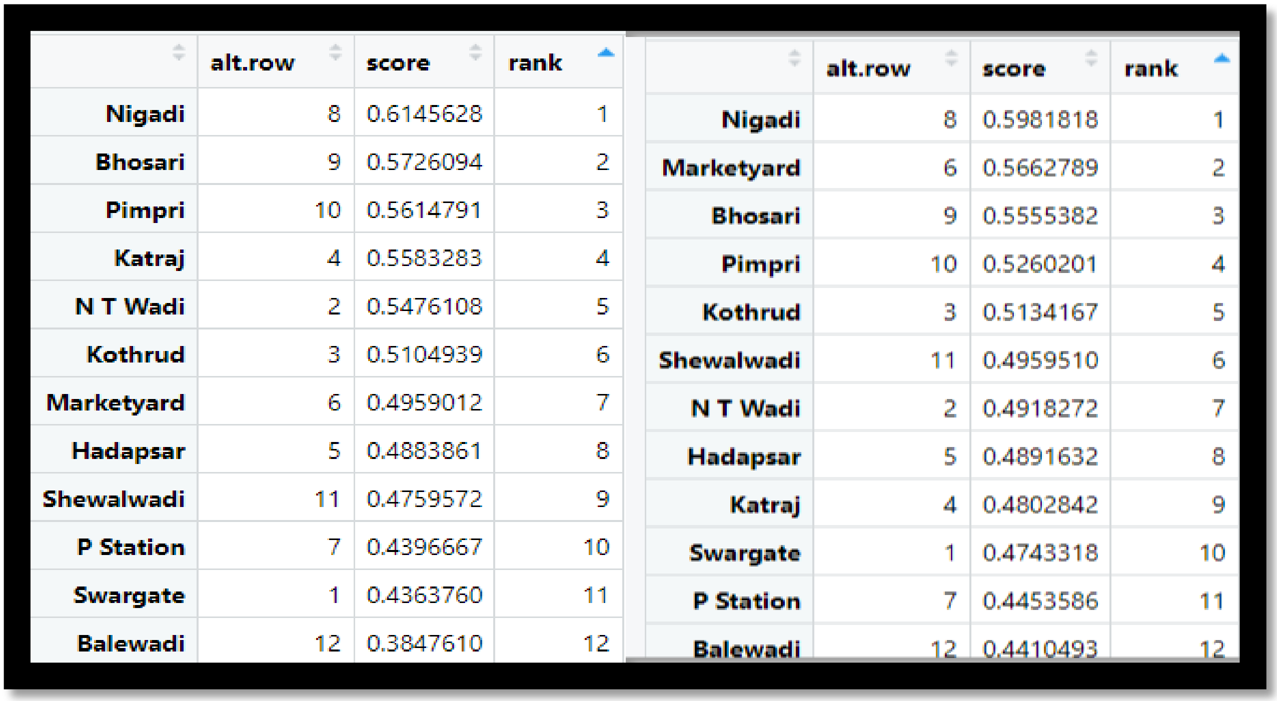
Here’s a brief overview of the TOPSIS process:

1. Create an evaluation matrix: This matrix consists of m alternatives and n criteria.
2. Calculate the normalized decision matrix: Each element is divided by the square root of the sum of the squares of all elements in that column.
3. Calculate the weighted normalized decision matrix: Each element of the normalized decision matrix is multiplied by the corresponding weight of the criterion.
4. Determine the worst alternative (NIS) and the best alternative (PIS).
5. Calculate the distance between the target alternative and the worst condition (NIS) and the distance between the target alternative and the best condition (PIS).
6. Calculate the similarity to the worst condition: The similarity to the worst condition is the distance from the NIS divided by the sum of the distances from the PIS and NIS.
7. Rank the alternatives according to their similarity to the worst condition.

TOPSIS is a compensatory method, which means it allows trade-offs between criteria. A poor result in one criterion can be compensated by a good result in another criterion. This provides a more realistic form of modeling than non-compensatory methods, which include or exclude alternative solutions based on hard cut-offs.

1.Quarter One Ranking of Depots

2. Quarter Two Ranking of Depots

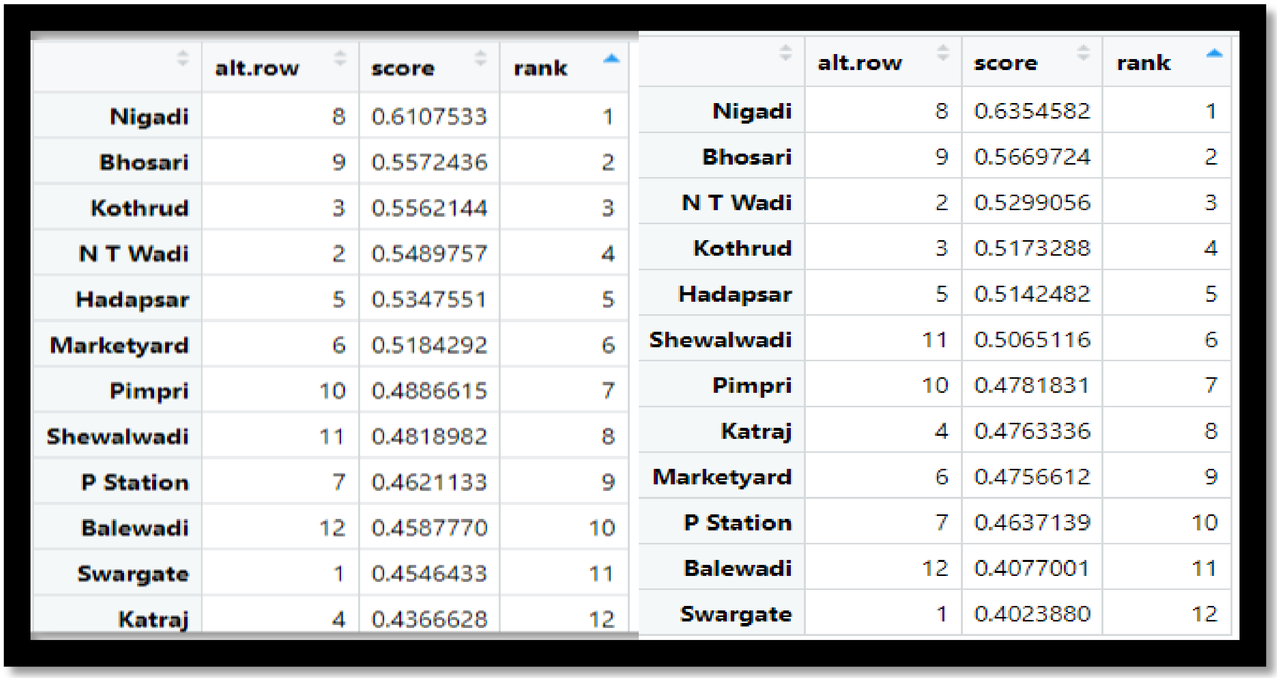


	alt.row	score	rank
Nigadi	8	0.6145628	1
Bhosari	9	0.5726094	2
Pimpri	10	0.5614791	3
Katraj	4	0.5583283	4
N T Wadi	2	0.5476108	5
Kothrud	3	0.5104939	6
Marketyard	6	0.4959012	7
Hadapsar	5	0.4883861	8
Shewalwadi	11	0.4759572	9
P Station	7	0.4396667	10
Swargate	1	0.4363760	11
Balewadi	12	0.3847610	12

	alt.row	score	rank
Nigadi	8	0.5981818	1
Marketyard	6	0.5662789	2
Bhosari	9	0.5555382	3
Pimpri	10	0.5260201	4
Kothrud	3	0.5134167	5
Shewalwadi	11	0.4959510	6
N T Wadi	2	0.4918272	7
Hadapsar	5	0.4891632	8
Katraj	4	0.4802842	9
Swargate	1	0.4743318	10
P Station	7	0.4453586	11
Balewadi	12	0.4410493	12

3.Quarter three Ranking of Depots

4. Quarter four Ranking of Depots



	alt.row	score	rank
Nigadi	8	0.6107533	1
Bhosari	9	0.5572436	2
Kothrud	3	0.5562144	3
N T Wadi	2	0.5489757	4
Hadapsar	5	0.5347551	5
Marketyard	6	0.5184292	6
Pimpri	10	0.4886615	7
Shewalwadi	11	0.4818982	8
P Station	7	0.4621133	9
Balewadi	12	0.4587770	10
Swargate	1	0.4546433	11
Katraj	4	0.4366628	12

	alt.row	score	rank
Nigadi	8	0.6354582	1
Bhosari	9	0.5669724	2
N T Wadi	2	0.5299056	3
Kothrud	3	0.5173288	4
Hadapsar	5	0.5142482	5
Shewalwadi	11	0.5065116	6
Pimpri	10	0.4781831	7
Katraj	4	0.4763336	8
Marketyard	6	0.4756612	9
P Station	7	0.4637139	10
Balewadi	12	0.4077001	11
Swargate	1	0.4023880	12

Based on the TOPSIS method analysis of the provided images, we observe the following performance-based rankings of the depots across each quarter:

1. **Nigadi Depot**: This depot consistently ranks at the top in every quarter, indicating superior performance throughout the year.
2. **Balewadi Depot**: This depot ranks last in the first and second quarters, suggesting areas for improvement in its operations during these periods.
3. **Katraj and Swargate Depots**: These depots rank last in the third and fourth quarters respectively, indicating a need for performance enhancement during these quarters.

These rankings provide valuable insights into the operational efficiency of each depot and can guide targeted strategies for performance improvement. It's important to note that these rankings are dynamic and can change with varying operational parameters and conditions. Therefore, continuous monitoring and analysis are crucial for maintaining and improving depot performance.

5.Identifying Outliers in Variables and Performance Measures Across Quarters

A boxplot is a powerful tool used in statistical analysis for identifying outliers in a dataset. It graphically represents the dataset's minimum, first quartile (Q1), median, third quartile (Q3), and maximum. The Interquartile Range (IQR), calculated as IQR = Q3 - Q1, is used to define the 'whiskers' of the boxplot. Any data point that falls below Q1 - 1.5*IQR or above Q3 + 1.5*IQR is considered an outlier and is often represented as individual points outside the whiskers in the plot. This method provides a quick and efficient way to detect outliers and understand the distribution of your data.

➢ **Quarter One:**

```
Variables                                                          Outliers
[1,]"Hire.Vehicles.Per.Day"                                        "Nigadi"
[2,]"Daily.average.of.Cancelled.km."                               "Katraj"
[3,]"Daily.average.of.Cancelled.km."                               "Nigadi"
[4,]"Avg..kms..per.new.tyres"                                   "Shewalwadi"
[5,]"Avg..kms.per.retreaded.tyres"                              "Shewalwadi"
```

1. **Nigadi Depot**: This depot has been identified as an outlier in terms of 'Hire Vehicles Per Day' and 'Daily Average of Cancelled km.' This could suggest that Nigadi depot has an unusually high number of vehicles hired per day and a high average of cancelled kilometers daily.
2. **Katraj Depot**: Katraj depot shows up as an outlier in the 'Daily Average of Cancelled km.' This might indicate that this depot has a higher than usual average of cancelled kilometers.
3. **Shewalwadi Depot**: Shewalwadi depot is an outlier in both 'Avg. kms. per new tyres' and 'Avg. kms. per retreaded tyres.' This could mean that the average kilometers driven per new and retreaded tyres at this depot are significantly different from the other depots.

These outliers could be due to various reasons such as operational differences, data entry errors, or exceptional events. It’s important to investigate these outliers further to understand their causes and implications for your analysis.

- **Quarter Two:**

```
Variables                                                       Outliers
[1,]  "Avg..vehicles.held...Per.Day.PMPML  in  2021"            "N  T  Wadi"
[2,]  "Total.Daily.Avg..Diesel.CNG."                            "Marketyard"
[3,]  "Diesel.Conusmption.per.Day.in.litres..PMPML"             "P  Station"
[4,]  "Breakdown.rate.per.10000.Kms."                           "P  Station"
[5,]  "Breakdown.rate.per.10000.Kms."                             "Balewadi"
[6,] "Avg..kms..per.new.tyres"                                    "Bhosari"
```

1. **N T Wadi Depot**: This depot has been identified as an outlier in terms of ‘Avg. vehicles held Per Day PMPML in .2021’. This could suggest that N T Wadi depot has an unusually high average number of vehicles held per day.
2. **Marketyard Depot**: Marketyard depot shows up as an outlier in the ‘Total Daily Avg. Diesel CNG.’ This might indicate that this depot has a higher than usual average daily consumption of Diesel and CNG.
3. **P Station Depot**: P Station depot is an outlier in both ‘Diesel Consumption per Day in litres PMPML’ and ‘Breakdown rate per 10000 Kms.’ This could mean that the average daily diesel consumption and the breakdown rate at this depot are significantly different from the other depots.
4. **Balewadi Depot**: Balewadi depot is identified as an outlier in the ‘Breakdown rate per 10000 Kms.’ This could suggest a higher than usual breakdown rate at this depot.
5. **Bhosari Depot**: Bhosari depot is an outlier in ‘Avg. kms. per new tyres.’ This could mean that the average kilometers driven per new tyres at this depot are significantly different from the other depots.

These outliers could be due to various reasons such as operational differences, data entry errors, or exceptional events. It’s important to investigate these outliers further to understand their causes and implications for your analysis.

- **Quarter Three:**

```
 [1,]  "Hire.Vehicles.Per.Day"                                                  "Nigadi"
 [2,]  "Avg.Workshop.Vehicles..Per.Day"                                         "Katraj"
 [3,]        "Vehicle.utilisation.in.kms...Gross...pmpml.only."               "Swargate"
 [4,]  "Breakdown.rate.per.10000.Kms."                                          "Katraj"
 [5,]  "Engine.Oil.Cons.in.litres.perday"                                       "Katraj"
 [6,]  "Avg..kms..per.new.tyres"                                               "Bhosari"
 [7,]  "Avg..kms..per.new.tyres"                                              "Balewadi"
 [8,]  "Avg..kms.per.retreaded.tyres"                                       "P  Station"
 [9,] "Total.no..of.default.cases.reported..DEO"                                "Katraj"
```

1. **Nigadi Depot**: This depot has been identified as an outlier in terms of ‘Hire Vehicles Per Day’. This could suggest that Nigadi depot has an unusually high number of vehicles hired per day.
2. **Katraj Depot**: Katraj depot shows up as an outlier in the ‘Avg. Workshop Vehicles Per Day’, ‘Breakdown rate per 10000 Kms.’, ‘Engine Oil Consumption in litres per day’, and ‘Total number of default cases reported DEO’. This might indicate that this depot has a higher than usual average of workshop vehicles per day, breakdown rate, engine oil consumption, and default cases.
3. **Swargate Depot**: Swargate depot is an outlier in ‘Vehicle utilisation in kms Gross PMPML only.’ This could mean that the vehicle utilisation at this depot is significantly different from the other depots.
4. **Bhosari Depot**: Bhosari depot is an outlier in ‘Avg. kms. per new tyres.’ This could suggest a higher than usual average kilometers driven per new tyres at this depot.
5. **Balewadi Depot**: Balewadi depot is identified as an outlier in the ‘Avg. kms. per new tyres.’ This could mean that the average kilometers driven per new tyres at this depot are significantly different from the other depots.
6. **P Station Depot**: P Station depot is an outlier in ‘Avg. kms. per retreaded tyres.’ This could suggest a higher than usual average kilometers driven per retreaded tyres at this depot.

These outliers could be due to various reasons such as operational differences, data entry errors, or exceptional events. It’s important to investigate these outliers further to understand their causes and implications for your analysis.

- **Quarter Four:**

```
[1,]"Hire.Vehicles.Per.Day"                                    "Nigadi"
[2,]"Avg..Spare.Vehicles.Per.Day"                            "Hadapsar"
[3,]"Avg..Spare.Vehicles.Per.Day"                           "P Station"
[4,]"Avg..kms..per.new.tyres"                                "Swargate"
[5,]"Avg..kms.per.retreaded.tyres"                         "P Station"
```

1. **Nigadi Depot**: This depot has been identified as an outlier in terms of ‘Hire Vehicles Per Day’. This could suggest that Nigadi depot has an unusually high number of vehicles hired per day.
2. **Hadapsar Depot**: Hadapsar depot shows up as an outlier in the ‘Avg. Spare Vehicles Per Day’. This might indicate that this depot has a higher than usual average of spare vehicles per day.
3. **P Station Depot**: P Station depot is an outlier in both ‘Avg. Spare Vehicles Per Day’ and ‘Avg. kms. per retreaded tyres.’ This could mean that the average spare vehicles per day and the average kilometers driven per retreaded tyres at this depot are significantly different from the other depots.
4. **Swargate Depot**: Swargate depot is an outlier in ‘Avg. kms. per new tyres.’ This could suggest a higher than usual average kilometers driven per new tyres at this depot.

These outliers could be due to various reasons such as operational differences, data entry errors, or exceptional events. It’s important to investigate these outliers further to understand their causes and implications for your analysis