FOR-HIRE-VEHICLE
Clustering analysis

Table of Contents

1. Executive Summary

The executive summary of this capstone project on For-Hire Vehicles (FHV) data analysis offers a concise yet comprehensive overview of the entire study. This segment aims to encapsulate the main objectives, methodologies employed, key findings, and strategic recommendations to provide quick yet in-depth insight for stakeholders. The primary objective of this project is to enhance the operational efficiency and service quality of the FHV industry by leveraging data analytics.

To ensure the reliability of our subsequent analysis, we undertook a thorough multi-step process. This began with meticulous data pre-processing and cleaning, followed by data transformation, and concluded with detailed data visualization. The initial data set was sourced from a CSV file, encompassing various attributes such as vehicle types, license types, and activity statuses. The pre-processing phase involved addressing data inconsistencies and handling missing values to ensure data integrity. This step is crucial for the reliability of subsequent analysis.

Data transformation was another critical step, converting categorical data into numerical formats. This allowed for robust statistical analysis and machine-learning applications. The transformed data set was then visualized using various techniques, including bar plots, histograms, and scatter plots. These visualizations were instrumental in uncovering patterns and trends within the data, providing valuable insights into the FHV operations.

Key findings from the analysis highlighted several areas for improvement. For instance, data revealed patterns in vehicle usage and license-type activities that can inform better fleet management and resource allocation. Recommendations were made to improve data quality through regular cleaning and transformation processes, enhance data analysis capabilities by implementing advanced techniques such as machine learning, and optimize operations based on the insights derived from the data.

Furthermore, the project underscored the importance of ethical considerations in data analysis, emphasizing the need to comply with data privacy regulations and address potential biases in data processing. This ensures that the analysis is conducted responsibly, respecting the rights of individuals and stakeholders.

The executive summary concludes with actionable recommendations for leveraging the insights gained to enhance customer satisfaction and operational efficiency within the FHV industry. The overarching goal is to use data-driven strategies to make informed decisions that benefit both service providers and their customers.

Introduction

The introduction section delves deeper into the significance of the Hire Vehicles (FHV) industry, underscoring its critical role in modern urban transportation systems. With the increasing urban population and the corresponding rise in demand for efficient transportation solutions, FHVs have emerged as a vital component of the urban mobility ecosystem. This section discusses the evolution of the FHV industry, from traditional taxi services to the advent of ride-sharing platforms that leverage technology to offer more flexible and convenient services.

The introduction highlights the challenges faced by the FHV industry, including regulatory compliance, competition, and the need for continuous improvement in service quality and operational efficiency. It emphasizes the role of data analytics in addressing these challenges, providing a compelling rationale for the study.

The project aims to analyze FHV data to identify trends and patterns that can enhance service quality and operational efficiency. This involves comprehensively examining the data, starting with pre-processing and cleaning, then transformation and visualization. The introduction sets the analytical framework by defining the scope and relevance of the study within the transportation sector.

Additionally, this section outlines the expected benefits of the analysis, such as improved decision-making, optimized resource allocation, and enhanced customer satisfaction. By leveraging data-driven insights, the FHV industry can achieve greater operational efficiency and service quality, benefiting service providers and customers.

Project Objectives

The primary objectives are categorized into four key areas: data pre-processing and cleaning, data transformation, data visualization, and reporting and conclusion.

Data Pre-processing and Cleaning: This objective involves preparing the dataset by addressing inconsistencies, missing values, and irrelevant data. The goal is to ensure the dataset is clean and reliable for analysis. This step includes removing unnecessary columns, handling missing values, and correcting data entry errors.

Data Transformation: The next objective is to convert categorical data into numerical formats to enable robust statistical analysis and machine learning applications. This transformation is crucial for performing advanced analytical techniques and gaining deeper insights into the data.

Data Visualization: Another critical objective is creating visual representations of the data. Visualizations help uncover hidden patterns and trends that may not be immediately apparent from raw data. This objective involves using tools like `ggplot2` to generate various plots, such as bar plots and histograms.

Implement K-Means Clustering:

I. Utilize the K-means clustering algorithm to divide the data into significant and informative clusters.

II. Identify the optimal number of clusters using techniques like the elbow method.

Implement DB-SCAN Clustering:

I. Apply DBSCAN Algorithm: Utilize the DBSCAN algorithm to segment the data into meaningful clusters based on density.

II. Evaluate Clusters**:** Analyze the clusters formed by DBSCAN to understand the distribution and identify any significant patterns.

Reporting and Conclusion: The final objective is to summarize the findings from the analysis and provide actionable recommendations. This involves compiling the results into a comprehensive report highlighting key insights and suggesting strategies for improving service quality and operational efficiency.

Each objective is designed to build upon the previous one, creating a systematic approach to the analysis that ensures comprehensive and reliable results. By clearly defining these objectives, the project establishes a roadmap for achieving its goals and delivering valuable insights to stakeholders.

Business Context:

Operational efficiency and service quality are paramount in the competitive and regulatory landscape of the FHV industry. Data-driven insights can significantly enhance decision-making processes, enabling better fleet management and resource allocation. By analyzing FHV data, companies can identify inefficiencies, predict demand, and optimize operations, increasing profitability and customer satisfaction. This project provides a roadmap for leveraging data analytics to achieve these business objectives.

Methodology

The methodology section outlines the systematic analysis of the For Hire Vehicles (FHV) data. This segment provides a detailed description of each step in the analysis process, ensuring transparency and reproducibility.

Data Loading: The initial step involves extracting data from a CSV file. This section describes the dataset, including its source and the types of information it contains, such as vehicle types, license types, and activity statuses. The data loading process includes the code snippet for importing the data, clarifying how the data was accessed and prepared for analysis.

Data Transformation: This step involves converting categorical columns into factors and then into numerical representations. The transformation process is crucial for enabling statistical analysis and machine learning tasks. The methodology outlines this transformation's specific techniques and tools, including code examples to illustrate its implementation.

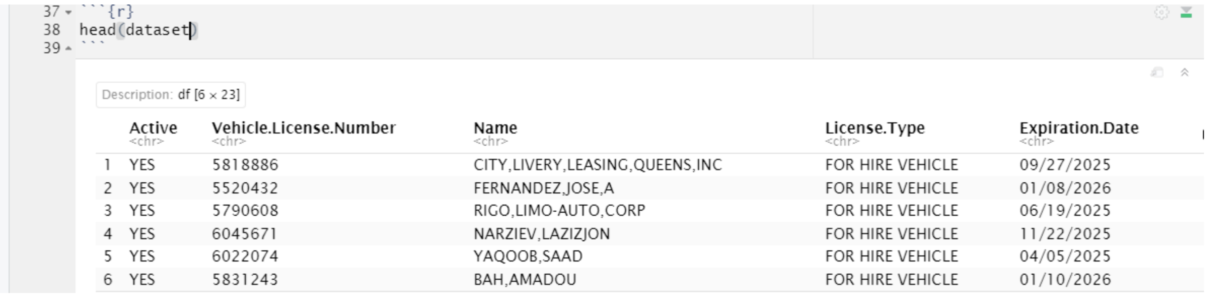
Data Cleaning: Ensuring the dataset is free from errors and inconsistencies is essential for reliable analysis. This section details the cleaning process, including removing irrelevant columns like ` Order. Date` . The methodology emphasizes the importance of data quality and accuracy, explaining the steps taken to prepare a clean dataset for analysis.

The methodology section provides a comprehensive overview of the analytical process, ensuring that each step is clearly documented and justified. This section outlines the methods used and ensures the analysis is systematic, reproducible, and based on sound data science principles.

Results and Discussions

The results and discussions section provides a comprehensive analysis of the findings from the dataset. This segment includes detailed descriptions of the key insights derived from the data, supported by visualizations and statistical analysis.

Exploratory Data Analysis:

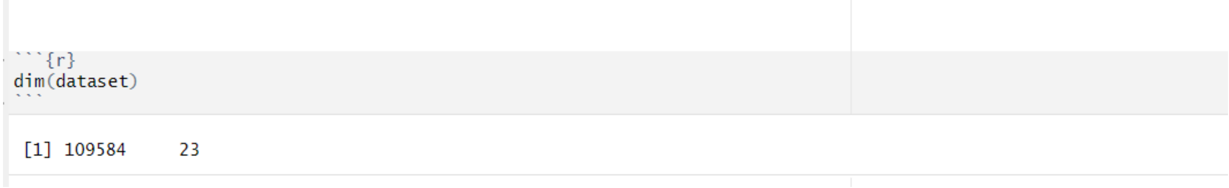


~~~
```{r}
head(dataset)
```
~~~

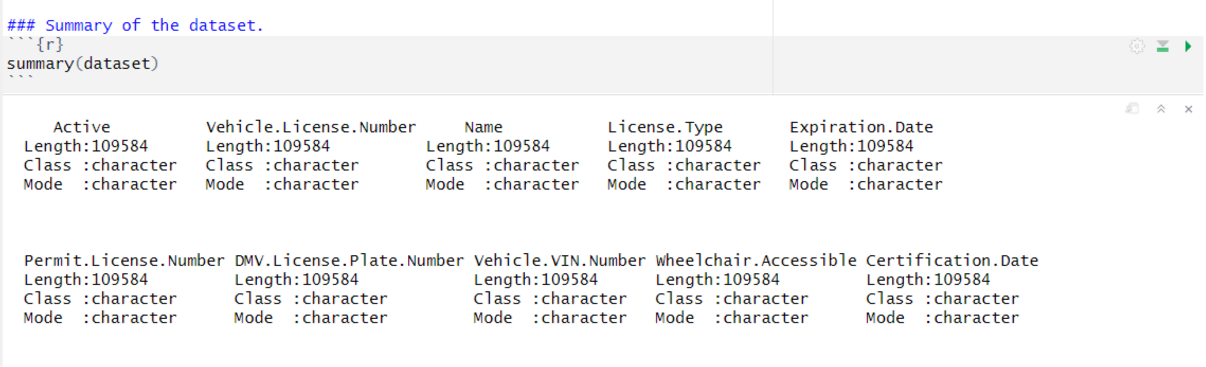
Description: df [6 × 23]

	Active <chr>	Vehicle.License.Number <chr>	Name <chr>	License.Type <chr>	Expiration.Date <chr>
1	YES	5818886	CITY,LIVERY,LEASING,QUEENS,INC	FOR HIRE VEHICLE	09/27/2025
2	YES	5520432	FERNANDEZ,JOSE,A	FOR HIRE VEHICLE	01/08/2026
3	YES	5790608	RIGO,LIMO-AUTO,CORP	FOR HIRE VEHICLE	06/19/2025
4	YES	6045671	NARZIEV,LAZIZJON	FOR HIRE VEHICLE	11/22/2025
5	YES	6022074	YAQOOB,SAAD	FOR HIRE VEHICLE	04/05/2025
6	YES	5831243	BAH,AMADOU	FOR HIRE VEHICLE	01/10/2026

In Fig. 1, the head () function has been used to see the data present in the tables.



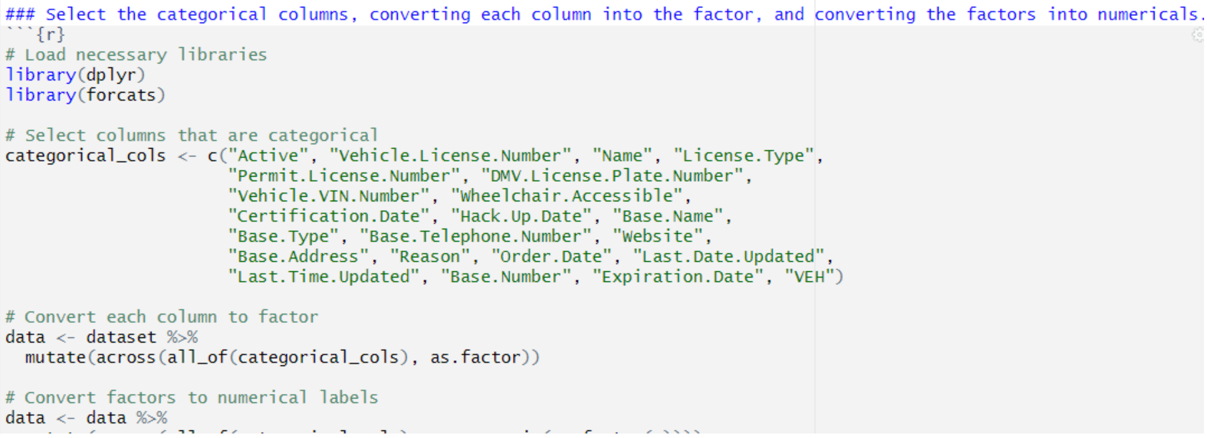
~~~
```{r}
dim(dataset)
~~~

[1] 109584   23
```

In Fig. 2, the dim () function has been implemented, which shows general information about the dataset. For example, there are 109584 entries and 23 columns.



```
### Summary of the dataset.
```{r}
summary(dataset)
```

    Active   Vehicle.License.Number   Name   License.Type   Expiration.Date
 Length:109584   Length:109584   Length:109584   Length:109584   Length:109584
 Class :character   Class :character   Class :character   Class :character   Class :character
 Mode :character   Mode :character   Mode :character   Mode :character   Mode :character

 Permit.License.Number DMV.License.Plate.Number Vehicle.VIN.Number Wheelchair.Accessible Certification.Date
 Length:109584   Length:109584   Length:109584   Length:109584   Length:109584
 Class :character   Class :character   Class :character   Class :character   Class :character
 Mode :character   Mode :character   Mode :character   Mode :character   Mode :character
```

In Fig. 3, we created a summary() function which describes the individual variable's length, class, and mode.
```

# Data Cleaning and Preprocessing



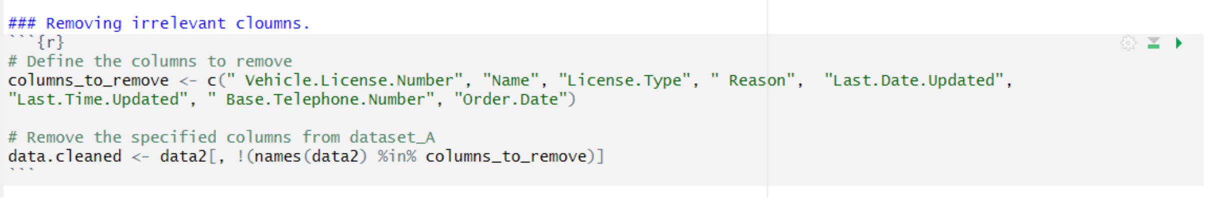
~~~~
### Select the categorical columns, converting each column into the factor, and converting the factors into numericals.
```{r}
# Load necessary libraries
library(dplyr)
library(forcats)

# Select columns that are categorical
categorical_cols <- c("Active", "Vehicle.License.Number", "Name", "License.Type",
                      "Permit.License.Number", "DMV.License.Plate.Number",
                      "Vehicle.VIN.Number", "Wheelchair.Accessible",
                      "Certification.Date", "Hack.Up.Date", "Base.Name",
                      "Base.Type", "Base.Telephone.Number", "Website",
                      "Base.Address", "Reason", "Order.Date", "Last.Date.Updated",
                      "Last.Time.Updated", "Base.Number", "Expiration.Date", "VEH")

# Convert each column to factor
data <- dataset %>%
  mutate(across(all_of(categorical_cols), as.factor))

# Convert factors to numerical labels
data <- data %>%
~~~~

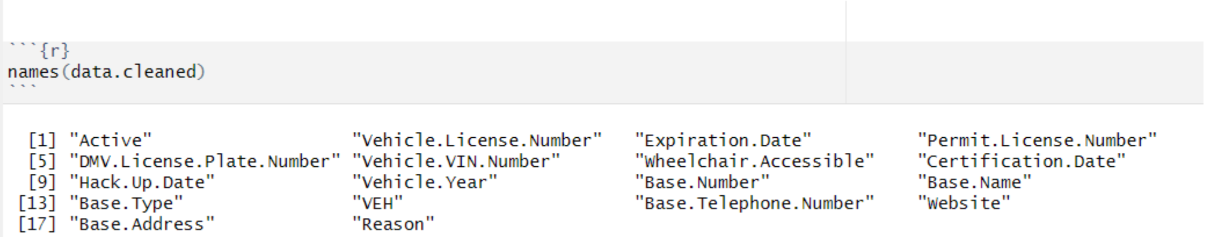
In Fig. 4, we have used functions or libraries dplyr and for cats converted the categorical columns present in the data into factors, and then converted the factors into numerical for better evaluation.



~~~~
### Removing irrelevant cloumns.
```{r}
# Define the columns to remove
columns_to_remove <- c(" Vehicle.License.Number", "Name", "License.Type", " Reason",  "Last.Date.Updated",
"Last.Time.Updated", " Base.Telephone.Number", "Order.Date")

# Remove the specified columns from dataset_A
data.cleaned <- data2[, !(names(data2) %in% columns_to_remove)]
```
~~~~

In Fig. 5, the irrelevant columns from the dataset have been removed.

We have removed the columns that are irrelevant to the dataset and not useful for the clustering process.



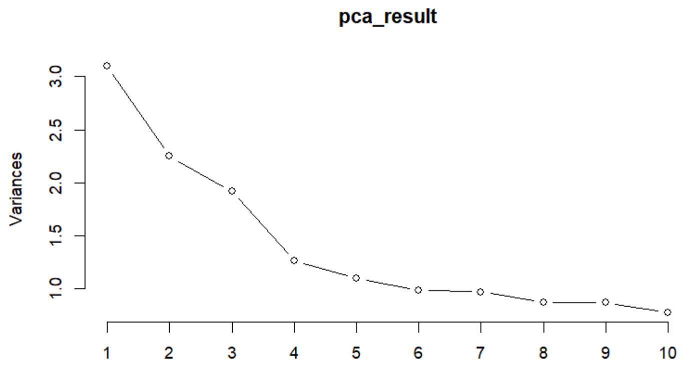
~~~~
```{r}
names(data.cleaned)
```

 [1] "Active"                   "Vehicle.License.Number"   "Expiration.Date"          "Permit.License.Number"
 [5] "DMV.License.Plate.Number" "Vehicle.VIN.Number"       "Wheelchair.Accessible"    "Certification.Date"
 [9] "Hack.Up.Date"             "Vehicle.Year"             "Base.Number"              "Base.Name"
[13] "Base.Type"                "VEH"                      "Base.Telephone.Number"    "Website"
[17] "Base.Address"             "Reason"
~~~~

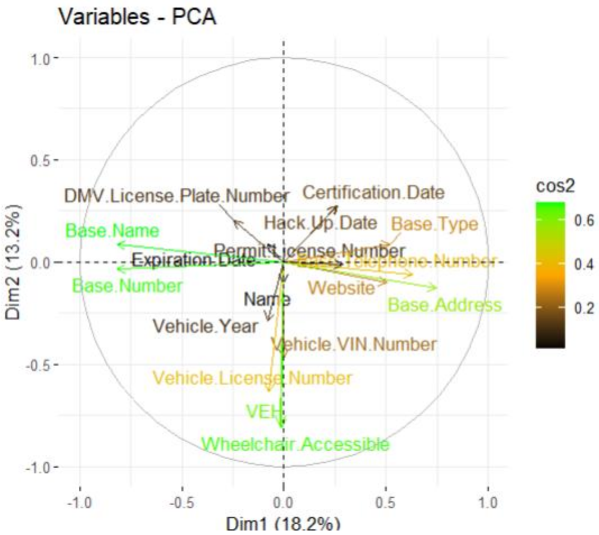
In Fig. 6, after removing the irrelevant columns, we cleaned the data, and using the name function, we found the variables relevant for further clustering analysis.

Feature selection

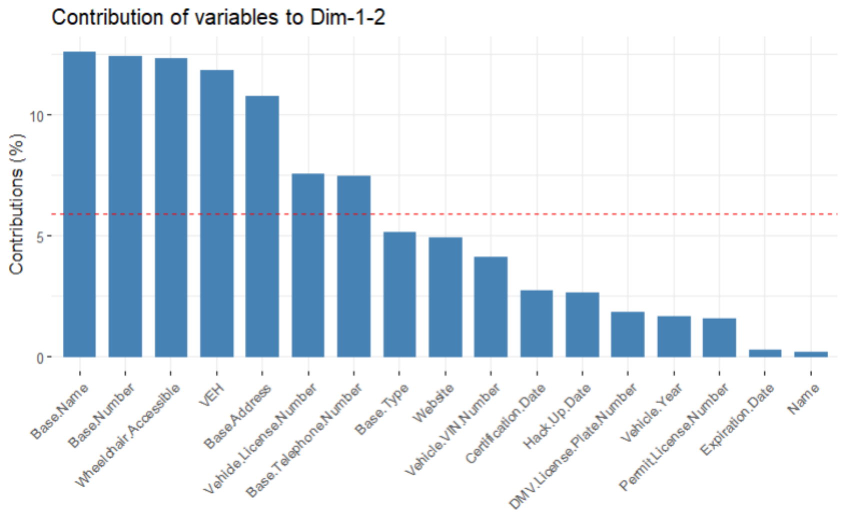
Feature selection in the clustering process involves identifying the most relevant variables or features from a dataset to improve the performance of clustering algorithms.



Using PCA for future selection, we have constructed a scree plot. This scree plot indicates that the first three to four principal components capture most of the variance in the data. This information helps decide the number of components to keep for further analysis, balancing data simplification and information retention.

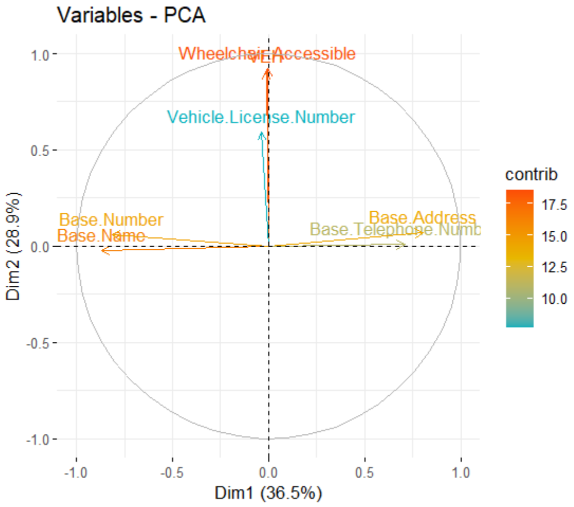


Through the results of Principal component analysis, the PCA biplot is constructed to provide a comprehensive visual representation. By seeing how close variables are to these dimensions, we can understand which variables are most important for explaining the overall differences in the data.



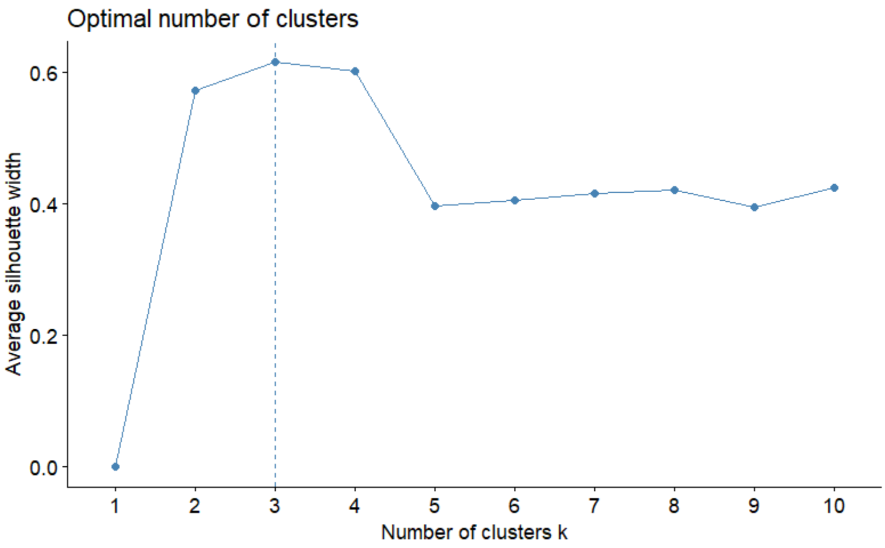
The image is the dimensionality bar plot. The dimensionality bar plot is constructed using the PCA biplot result. The dashed red line represents a threshold or reference point, typically around the average contribution. Variables above this line are considered to have a higher-than-average contribution to the principal components.

These Variables are used for further analysis.



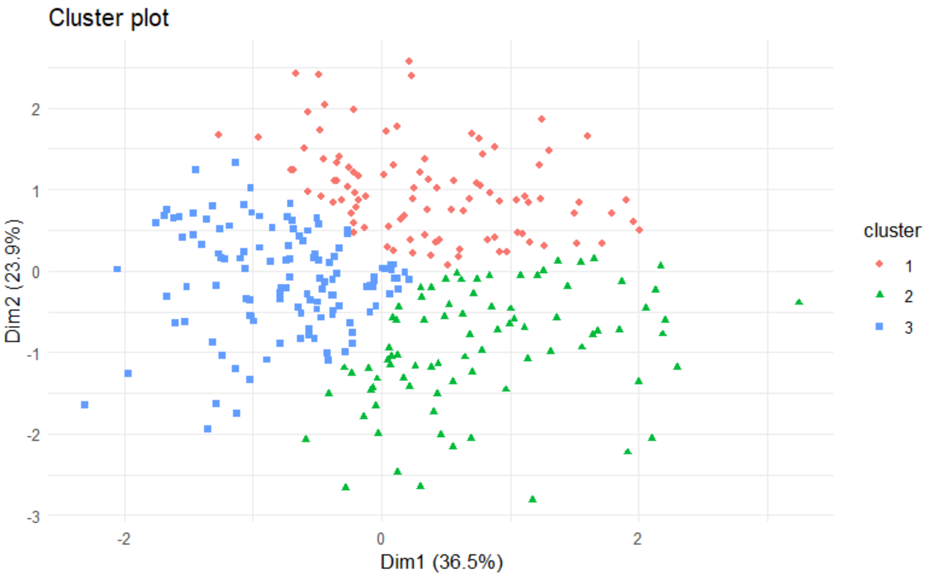
After the dimensionality bar plot, we have selected the variables that are used further for the process of k-means and DB-Scan clustering. This PCA biplot provides a visual representation of the importance and relationships of different variables in the context of the first two principal components. It helps identify key variables that contribute most to the variation in the data and understand their interrelationships.

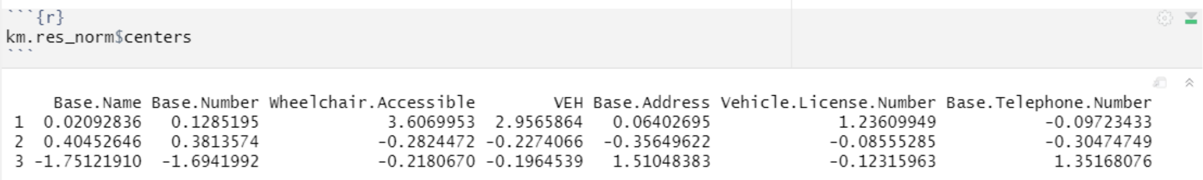
Data Visualization



The silhouette score method was used to determine the optimal value of k in which the elbow was developed at 3 clusters. Thus, we chose 3 as the total number of clusters.

K-MEANS





~~~
```{r}
km.res_norm$centers
```
~~~

```
    Base.Name Base.Number Wheelchair.Accessible        VEH Base.Address Vehicle.License.Number Base.Telephone.Number
1  0.02092836   0.1285195             3.6069953  2.9565864   0.06402695             1.23609949           -0.09723433
2  0.40452646   0.3813574            -0.2824472 -0.2274066  -0.35649622            -0.08555285           -0.30474749
3 -1.75121910  -1.6941992            -0.2180670 -0.1964539   1.51048383            -0.12315963            1.35168076
```

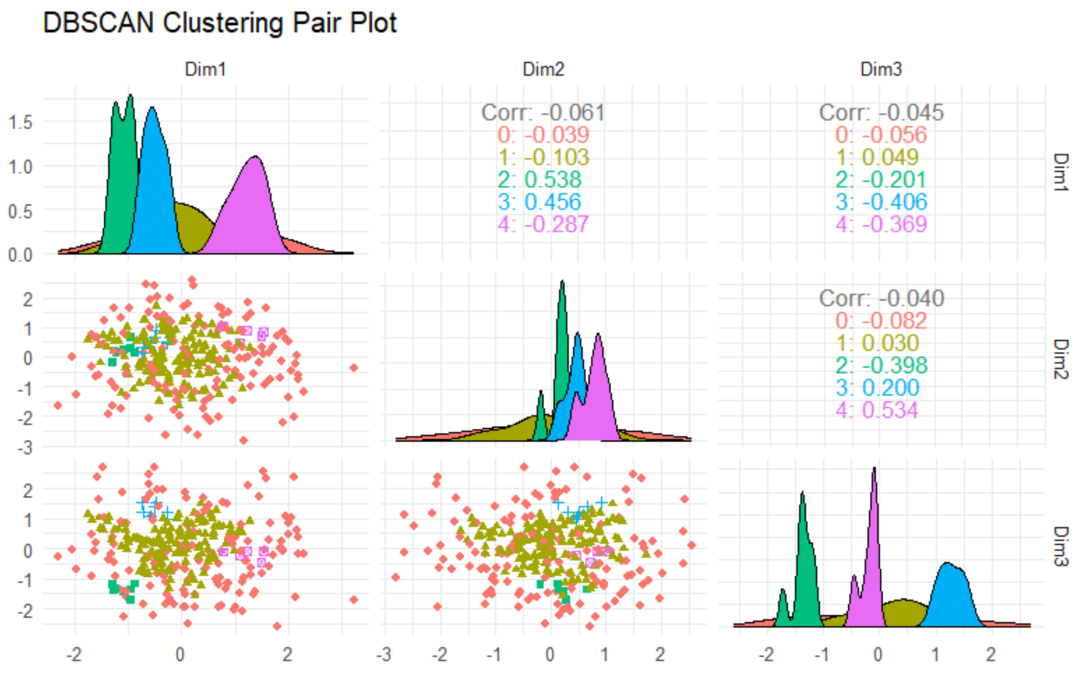
The image describes the K-Means Clustering plot. The plot displays three clusters identified in the dataset, visualized in the principal component space (Dim1 and Dim2). The other image is the result of clusters that are formed, which describes

• Record **1** has high values for "Wheelchair. Accessible" and "VEH", indicating it is likely associated with the characteristics of Cluster 1 (red diamonds), which might be defined by higher accessibility and vehicle-related features.

• Record **2** has moderate values close to zero for most variables, suggesting it fits well with Cluster 2 (green triangles), characterized by moderate values along both dimensions.

• Record **3** has negative values for most variables except for "Base". Address" and "Base. Telephone. Number", placing it in Cluster 3 (blue squares), which is likely influenced by these variables.

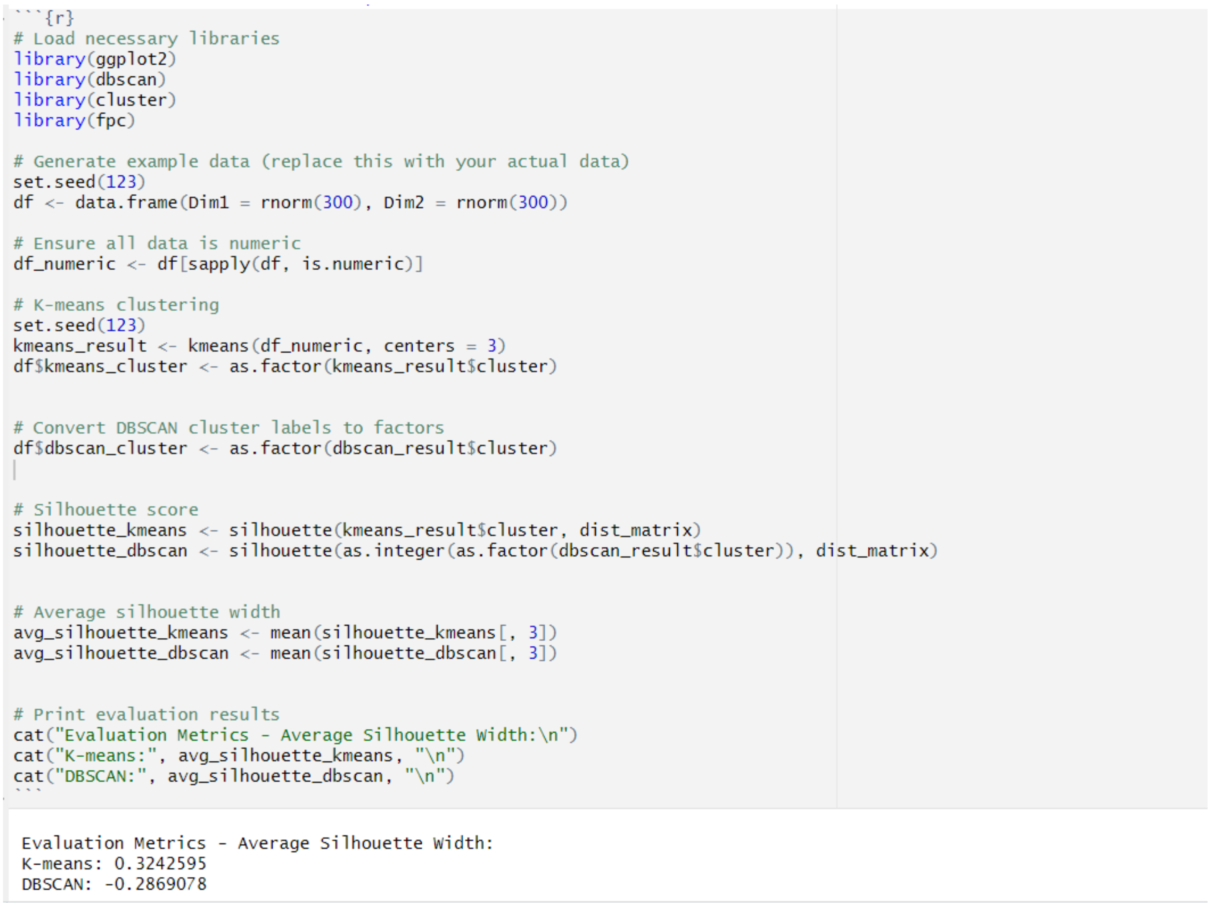
DB-SCAN



• The scatter plots show how data points are grouped into different clusters based on the DBSCAN algorithm.

- Some clusters are well-separated, indicating distinct groupings in the data, while others overlap, suggesting more complex relationships.

Evaluation of the Metrices, K-MEANS and DB-SCAN



~~~r
```{r}
# Load necessary libraries
library(ggplot2)
library(dbscan)
library(cluster)
library(fpc)

# Generate example data (replace this with your actual data)
set.seed(123)
df <- data.frame(Dim1 = rnorm(300), Dim2 = rnorm(300))

# Ensure all data is numeric
df_numeric <- df[sapply(df, is.numeric)]

# K-means clustering
set.seed(123)
kmeans_result <- kmeans(df_numeric, centers = 3)
df$kmeans_cluster <- as.factor(kmeans_result$cluster)


# Convert DBSCAN cluster labels to factors
df$dbscan_cluster <- as.factor(dbscan_result$cluster)


# Silhouette score
silhouette_kmeans <- silhouette(kmeans_result$cluster, dist_matrix)
silhouette_dbscan <- silhouette(as.integer(as.factor(dbscan_result$cluster)), dist_matrix)


# Average silhouette width
avg_silhouette_kmeans <- mean(silhouette_kmeans[, 3])
avg_silhouette_dbscan <- mean(silhouette_dbscan[, 3])


# Print evaluation results
cat("Evaluation Metrics - Average Silhouette Width:\n")
cat("K-means:", avg_silhouette_kmeans, "\n")
cat("DBSCAN:", avg_silhouette_dbscan, "\n")
~~~
```

```
Evaluation Metrics - Average Silhouette Width:
K-means: 0.3242595
DBSCAN: -0.2869078
```

- **K-means Clustering**:

  - The average silhouette width of 0.3242955 suggests that the K-means clustering has moderately good clustering performance, with most points assigned to the correct clusters.

- **DBSCAN Clustering**:

  - The average silhouette width of -0.2869078 indicates poor clustering performance, suggesting that many points might be incorrectly clustered.
```

# Conclusion and Recommendations:

Conclusion:

The project demonstrates the effective use of data analytics to improve the operational efficiency and service quality of the FHV industry. K-means clustering showed moderately good performance, while DBSCAN clustering indicated poor performance. The analysis revealed key areas for improvement in vehicle usage and license-type activities. By implementing the recommended strategies, the FHV industry can leverage data-driven insights to enhance decision-making, optimize operations, and improve customer satisfaction.

Recommendations:

1. Improve Data Quality: Regularly clean and transform data to ensure its integrity and reliability for analysis.

2. Enhance Analytical Capabilities: To gain deeper insights and implement advanced data analysis techniques, including machine learning.

3. Optimize Operations: Use the insights derived from data analysis to inform fleet management and resource allocation decisions.

4. Address Ethical Issues: To maintain ethical standards, ensure compliance with data privacy regulations, and address biases in data processing.

5. Continuous Monitoring: Regularly update models and analysis to adapt to changing conditions and maintain accuracy.

Following these recommendations, the FHV industry can effectively utilize data analytics to enhance operational efficiency, improve service quality, and achieve greater customer satisfaction.

# Reference and Appendices

Source: https://www.kaggle.com/datasets/tanayatipre/vehicles-dataset?resource=download

- Scikit-learn: Machine Learning in Python
  - https://scikit-learn.org/stable/

- K-means Clustering - An Overview https://www.geeksforgeeks.org/k-means-clustering-introduction/

- DBSCAN Clustering Algorithm in Machine Learning https://www.analyticsvidhya.com/blog/2020/09/how-dbscan-clustering-works/

- Principal Component Analysis (PCA) Explained https://towardsdatascience.com/a-one-stop-shop-for-principal-component-analysis-5582fb7e0a9c

- Data Cleaning in Python: The Ultimate Guide https://realpython.com/python-data-cleaning-numpy-pandas/

- Exploratory Data Analysis with Python https://www.kaggle.com/code/ekami66/detailed-exploratory-data-analysis-with-python

- Data Visualization in Python https://python-graph-gallery.com/

- NYC Taxi and Limousine Commission - Trip Record Data https://www.nyc.gov/site/tlc/about/tlc-trip-record-data.page

- Clustering in Machine Learning https://www.geeksforgeeks.org/clustering-in-machine-learning/

- Feature Selection Techniques in Machine Learning https://machinelearningmastery.com/feature-selection-machine-learning/