General Sir John Kotelawala Defence University

Applied Data Science & Communication Intake-41

Assignment -1

Application of Data Mining in Public Sector Clustering

(Team - Knowledge Excavators)

Authors:

D/ADC/24/0021 - D.P.C.Sadunika

D/ADC/24/0024 - M.M.C.C.Marasingha

D/ADC/24/0033 - E.S.R.Ruparathna

D/ADC/24/0034 - W.D.S.N.Kulasooriya

AeroAi Cluster

A Deep Dive into Airline Delays

Content

01. Introduction
02. Dataset
03. Explanation and preparation of dataset
04. Data Visualizations
05. Data Mining Techniques used
06. Implementation in R
07. Result analysis and Discussion
08. Impact
09. Conclusion
10. References

01.Introduction

Air transport unites people and businesses over vast distances and hence is a fundamental part of overseas transportation. However, air passengers, airlines, and the economy as a whole can all be severely impacted by flight delays. Enhancing productivity and reducing disruption in air transport entail understanding causes and trends of flight delays.

Using actual public-sector information, this study examines flight delays across U.S. planes in December of 2019 and December of 2020. The dataset was provided by the Bureau of Transportation Statistics, where it has a variety of variables that affect aircraft delays including weather, air carrier, security, and national aviation system constraints.

This study uses data mining techniques, such as predictive modeling and clustering, to establish noteworthy trends of airline delays, recognize noteworthy causal factors, and produce valuable information that policymakers, airlines, and airport authorities can benefit from.

The Research Question:
"What were the principal factors that contributed to airline delays during December 2019 and 2020?”

02.Dataset

The Bureau of Transportation Statistics provided the airline_delay.csv dataset, which was used in this study. The dataset includes comprehensive airline delay data for December 2019 and 2020. The dataset includes informative data on flight delays in various American cities, enabling a comprehensive analysis of variables affecting airline delays.

2.1 Dataset Description

The dataset consists of **3,351 rows and 21 variables**, covering information about flights, delays, and cancellations per airline per airport. The key variables include:

2.2 Independent Variables:

- **year** – Year of data collection (2019 or 2020).
- **month** – Numeric representation of the month (December).
- **carrier** – Airline carrier code.
- **carrier_name** – Name of the airline carrier.
- **airport** – Airport code.
- **airport_name** – Name of the airport.
- **arr_flights** – Number of flights arriving at the airport.

2.3 Dependent Variables (Delay-related Features):

- **arr_del15** – Number of flights delayed by more than 15 minutes.
- **carrier_ct** – Number of flights delayed due to air carrier issues (e.g., lack of crew).
- **weather_ct** – Number of flights delayed due to weather conditions.
- **nas_ct** – Number of flights delayed due to National Aviation System (e.g., heavy air traffic).
- **security_ct** – Number of flights canceled due to security breaches.
- **late_aircraft_ct** – Number of flights delayed because of a previous flight on the same aircraft being late.
- **arr_cancelled** – Number of canceled flights.
- **arr_diverted** – Number of diverted flights.
- **arr_delay** – Total delay time in minutes.
- **carrier_delay** – Total delay time due to air carrier issues.
- **weather_delay** – Total delay time due to weather conditions.

- **nas_delay** – Total delay time due to the National Aviation System.
- **security_delay** – Total delay time due to security issues.
- **late_aircraft_delay** – Total delay time caused by previous delayed flights on the same aircraft.

2.4 Data Source & Purpose

This data set was chosen because it is an accurate reflection of flight delays during a period of peak travel demand. We can determine the primary causes of delays for different airports and airlines by thoroughly examining multivariable relationships facilitated by the data.

2.5 Potential Uses of Data Mining

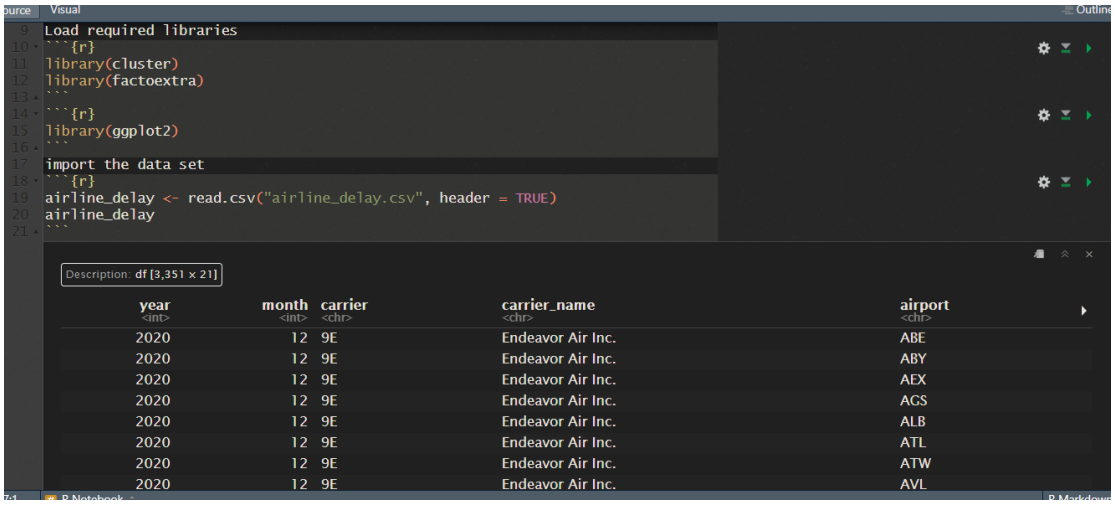
By using predictive modeling and clustering, the data can be utilized to:

- Recognize trends in flight delays for various airlines and airports.
- Identify the primary causes of airline operational delay.
- Create forecasting models to determine the likelihood of future delays.

03.Explanation and preparation Dataset

1. Data Import and Inspection

The dataset **airline_delay.csv** was imported into R and inspected to understand its structure and content.



Load required libraries

```r
library(cluster)
library(factoextra)
```

```r
library(ggplot2)
```

import the data set

```r
airline_delay <- read.csv("airline_delay.csv", header = TRUE)
airline_delay
```

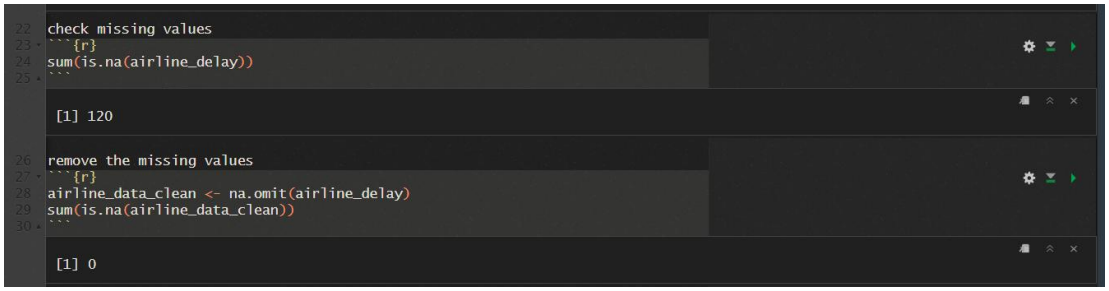
Description: df [3,351 × 21]

year <int>	month <int>	carrier <chr>	carrier_name <chr>	airport <chr>
2020	12	9E	Endeavor Air Inc.	ABE
2020	12	9E	Endeavor Air Inc.	ABY
2020	12	9E	Endeavor Air Inc.	AEX
2020	12	9E	Endeavor Air Inc.	AGS
2020	12	9E	Endeavor Air Inc.	ALB
2020	12	9E	Endeavor Air Inc.	ATL
2020	12	9E	Endeavor Air Inc.	ATW
2020	12	9E	Endeavor Air Inc.	AVL

- The dataset contains **3,351 rows and 21 columns** related to airline delays.
- Variables include **numerical** (e.g., arr_delay, carrier_delay) and **categorical** (carrier, airport).

2. Handling Missing Values

Missing values can impact data quality. The dataset was checked for missing values and cleaned using na.omit().



check missing values

```r
sum(is.na(airline_delay))
```

```
[1] 120
```

remove the missing values

```r
airline_data_clean <- na.omit(airline_delay)
sum(is.na(airline_data_clean))
```

```
[1] 0
```

- Rows with missing values were removed, ensuring a **complete dataset** for analysis.

3. Data Structure and Summary Statistics

The cleaned dataset was analyzed using summary statistics to understand distributions.

```r
Inspect the dataset in R
```{r}
names(airline_data_clean)
head(airline_data_clean)
str(airline_data_clean)
```
```

R Console

data.frame
6 x 21

Description: df [6 × 21]
```

| | year <int> | month <int> | carrier <chr> | carrier_name <chr> | airport <chr> |
|---|---|---|---|---|---|
| 1 | 2020 | 12 | 9E | Endeavor Air Inc. | ABE |
| 2 | 2020 | 12 | 9E | Endeavor Air Inc. | ABY |
| 3 | 2020 | 12 | 9E | Endeavor Air Inc. | AEX |
| 4 | 2020 | 12 | 9E | Endeavor Air Inc. | AGS |
| 5 | 2020 | 12 | 9E | Endeavor Air Inc. | ALB |
| 6 | 2020 | 12 | 9E | Endeavor Air Inc. | ATL |

6 rows | 1-6 of 21 columns

- Variables such as arr_delay and late_aircraft_ct showed **wide variations**, highlighting the need for normalization.
- Airlines had varying delay patterns, which justified clustering analysis.

## 4. Data Dimensionality Check

Before proceeding with normalization, the dataset's **dimensions (rows and columns)** were verified.

```r
Check the dimension and number of points
```{r}
nrow(airline_data_clean)
ncol(airline_data_clean)
dim(airline_data_clean)
```
```

```
[1] 3343
[1] 21
[1] 3343   21
```

- The dataset maintained **a balanced structure** after removing missing values.

5. Data Visualization and Initial Analysis

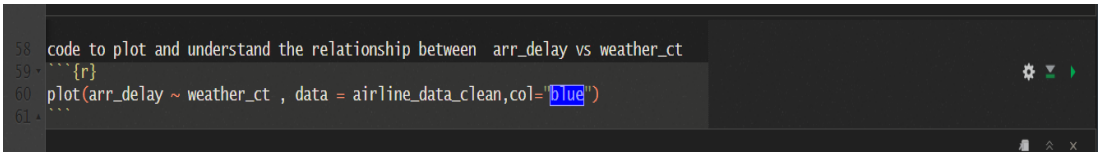
To understand relationships between features, scatter plots were created.

Scatterplot Matrix for Key Variables

```r
Create scatterplot matrix
```{r}
pairs(airline_data_clean[, c("arr_flights", "arr_del15", "carrier_ct", "weather_ct", "nas_ct", "security_ct",
"late_aircraft_ct")])
```
```

- This **identified potential correlations** between delay causes.
```

Scatterplots for Delay Analysis

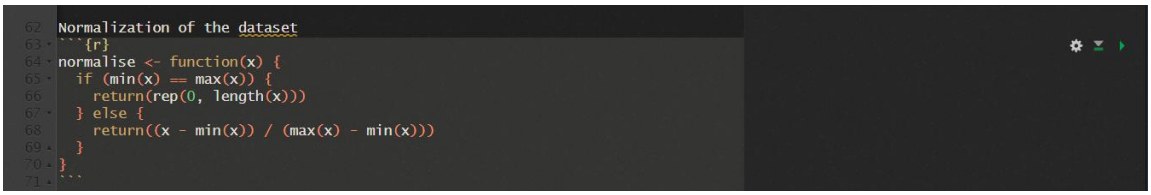


- These plots **visualized trends** in flight delays.

6. Data Normalization

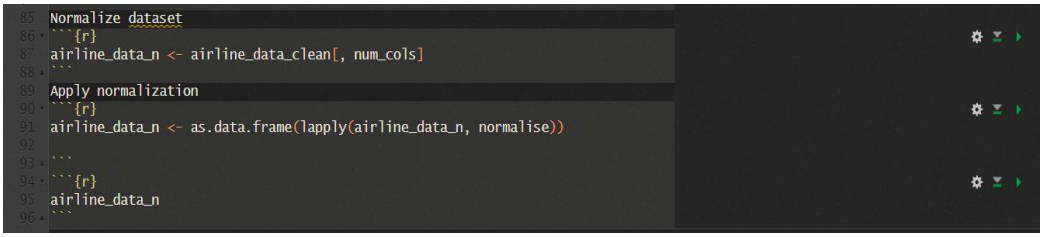
Since delay-related variables had different scales, **min-max normalization** was applied.

Code for Normalization Function



Selecting Numeric Variables for Normalization

Applying Normalization



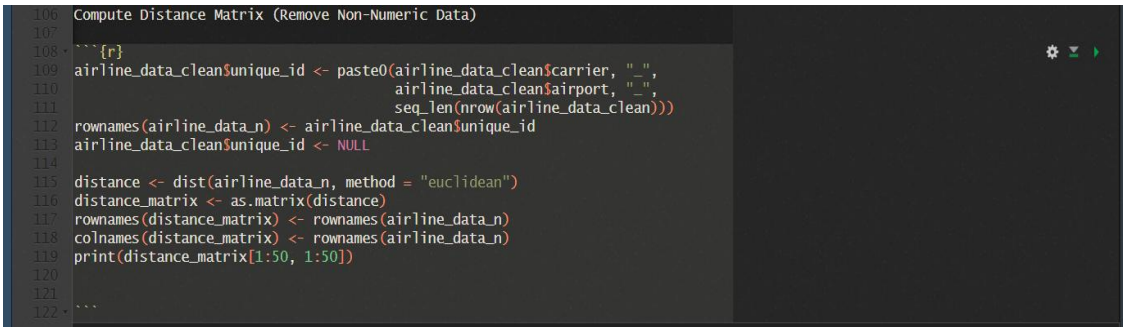
Effect of Normalization:

- Transformed values to a **0–1 range**, making them comparable.
- Prevented variables with larger numerical ranges from **dominating clustering algorithms**.

7. Computing Distance Matrix

To prepare for clustering, a **Euclidean distance matrix** was computed.

Code for Creating Unique Identifiers and Computing Distance



````
Compute Distance Matrix (Remove Non-Numeric Data)

```{r}
airline_data_clean$unique_id <- paste0(airline_data_clean$carrier, "_",
                                       airline_data_clean$airport, "_",
                                       seq_len(nrow(airline_data_clean)))
rownames(airline_data_n) <- airline_data_clean$unique_id
airline_data_clean$unique_id <- NULL

distance <- dist(airline_data_n, method = "euclidean")
distance_matrix <- as.matrix(distance)
rownames(distance_matrix) <- rownames(airline_data_n)
colnames(distance_matrix) <- rownames(airline_data_n)
print(distance_matrix[1:50, 1:50])


```
````

Effect of Distance Computation:

- Converted data into a **format suitable for clustering**.

8. Exporting Distance Matrix

The computed distance matrix was saved as a CSV file.

````
distance as csv file
```{r}
write.csv(distance_matrix, "airline_distance_matrix.csv")
```
````

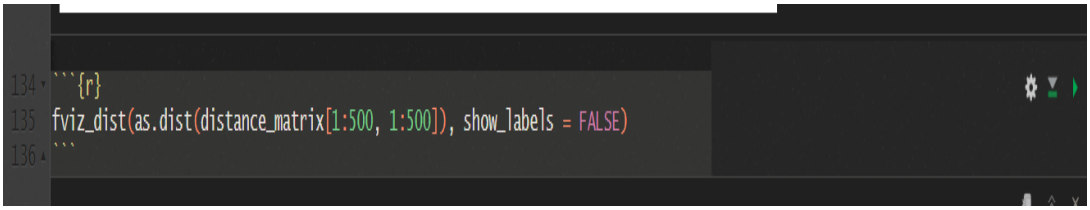
- This allowed for **further analysis and sharing of results**.

9. Data Visualization

The **distance matrix** was visualized using pheatmaps and cluster plots.

Pheatmap of Airline Distance Matrix

Cluster Visualization



- Helped identify **patterns in airline delays**.

04.Data Visualization

Visualizing data is a crucial step in understanding patterns, relationships, and trends. Various plots and graphs were generated using **ggplot2, pheatmap, and clustering visualization tools** to explore the dataset effectively.

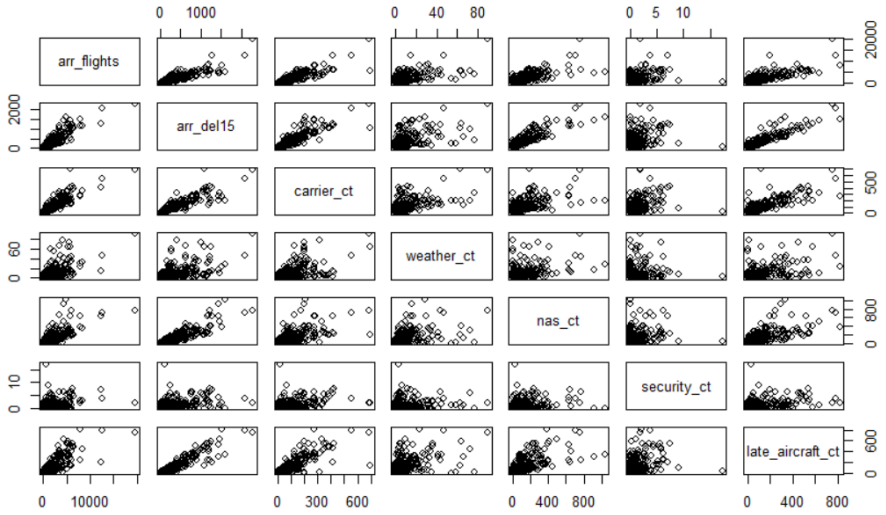
1. Scatterplot Matrix for Key Features

A **scatterplot matrix** was created to analyze relationships between key flight delay factors such as **arr_flights (total arrivals), arr_del15 (flights delayed by more than 15 minutes), carrier delays, weather-related delays, and NAS delays**.

Create scatterplot matrix

Hide

```
pairs(airline_data_clean[, c("arr_flights", "arr_del15", "carrier_ct", "weather_ct", "nas_ct", "security_ct", "late_aircraft_ct")])
```



Findings:

- This visualization helped identify **correlations between different delay causes**.
- **Late aircraft delays** appeared to be a major contributor to total delays.

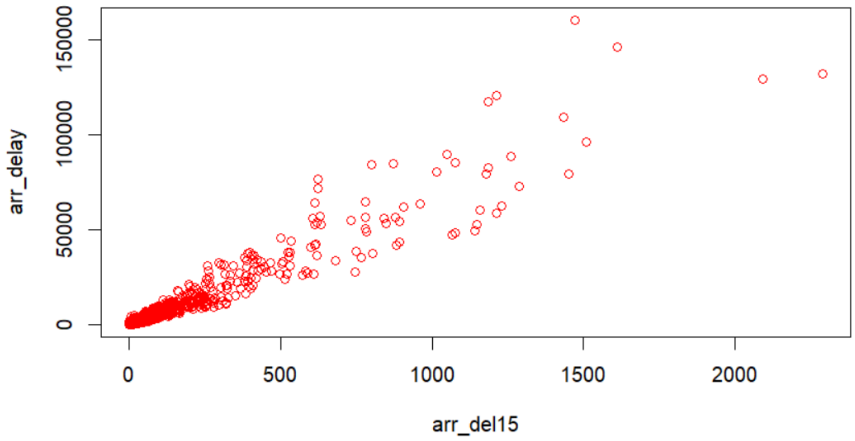
2. Relationship between Arrival Delays and Delay Causes

To understand how different delay factors contribute to total arrival delays, scatter plots were generated.

code to plot and understand the relationship between arr_delay vs arr_del15

Hide

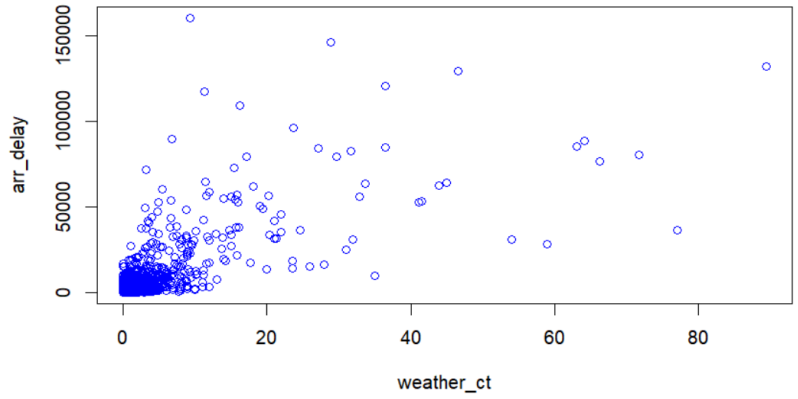
```
plot(arr_delay ~ arr_del15, data = airline_data_clean,col="red")
```



code to plot and understand the relationship between arr_delay vs weather_ct

Hide

```
plot(arr_delay ~ weather_ct , data = airline_data_clean,col="blue")
```



Findings:

- **A positive correlation was observed** between total delay (arr_delay) and the number of delayed flights (arr_del15).
- **Weather-related delays had a moderate impact** on total delays, but other factors such as **carrier delays** and **NAS-related delays** also played significant roles.

3. Heatmap of Distance Matrix

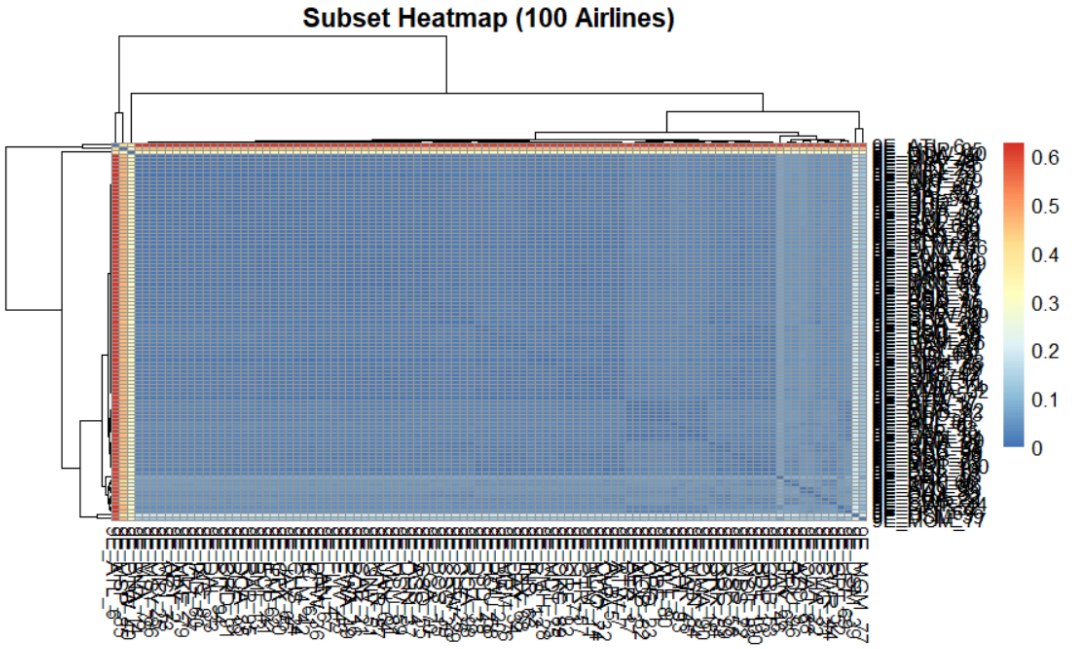
A **Heatmap** was created to visualize airline delay patterns based on **computed Euclidean distances** between different airline delay profiles.

visualization

Hide

```
library(pheatmap)

pheatmap(distance_matrix[1:100, 1:100], main = "Subset Heatmap (100 Airlines)")
```



Findings:

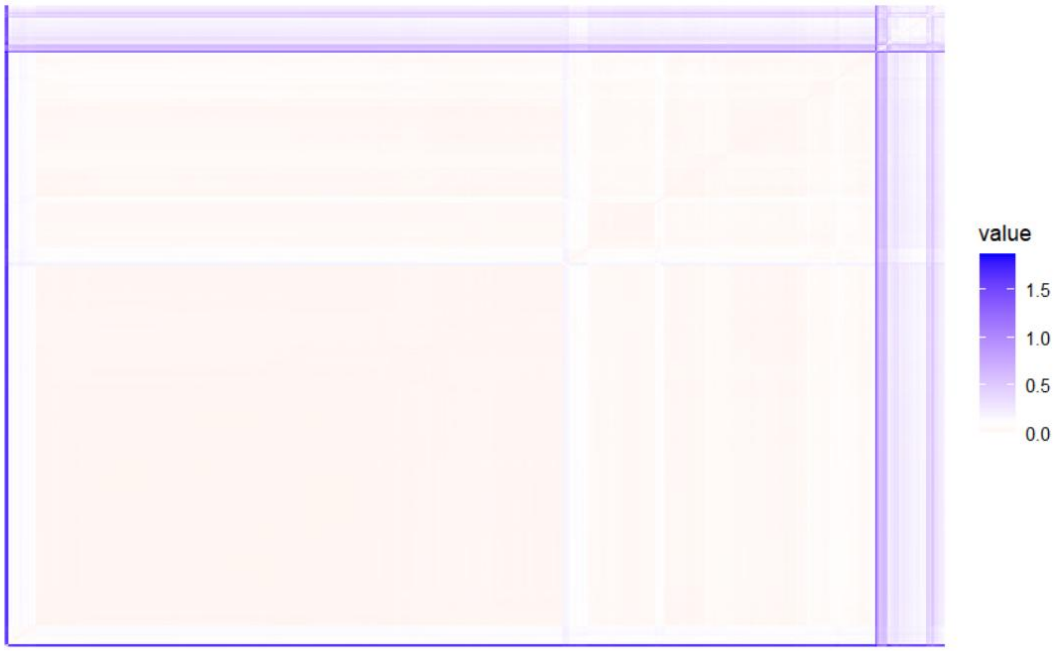
- Airlines with similar **delay characteristics were clustered together**.
- Some airlines consistently exhibited **higher delay times**, forming distinct clusters.

4. Distance Matrix Visualization

A **distance matrix visualization** was created using **factoextra** to better understand airline delay groupings.

Hide

```
fviz_dist(as.dist(distance_matrix[1:500, 1:500]), show_labels = FALSE)
```



Findings:

- Clear **groupings of airlines based on delay profiles** were observed.
- Some clusters had **consistently higher delay times**, making them potential targets for improvement.

5. K-Means Clustering Visualization

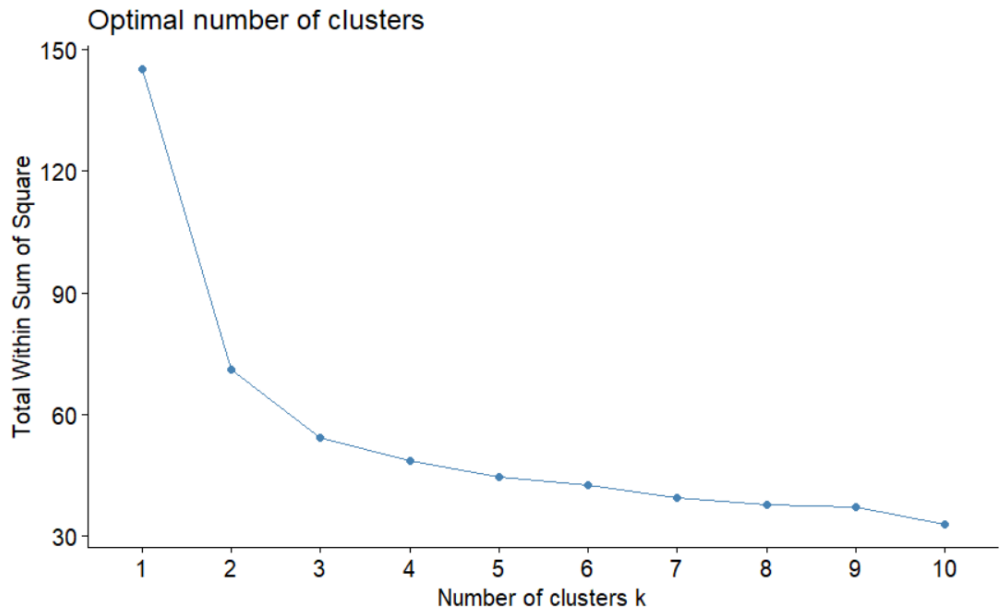
To identify delay patterns, **K-Means clustering** was applied to the dataset, grouping airlines based on delay characteristics.

Determining Optimal Clusters (Elbow Method)

K-Means clustering

Hide

```
fviz_nbclust(airline_data_n, kmeans, method = "wss")
```



Hide

Applying K-Means Clustering

Hide

```
set.seed(123)
k <- 3
kmeans_result <- kmeans(airline_data_n, centers = k, nstart = 30)
```

Hide

Hide

```
set.seed(123)
kc <- kmeans(airline_data_n, centers = 3, nstart = 30)
print(kc)
```

```
K-means clustering with 3 clusters of sizes 36, 3174, 133

Cluster means:
  arr_flights  arr_del15 carrier_ct  weather_ct      nas_ct security_ct late_aircraft_ct arr_cancelled arr_diverte
d
1 0.305583739 0.49455124 0.42013510 0.324498621 0.372346316  0.16543103      0.512705810    0.24565972  0.22949735
4
2 0.007688087 0.01080536 0.01266715 0.008514051 0.006949456  0.00353373      0.009586601    0.00637293  0.00821406
1
3 0.112870829 0.16824499 0.16333308 0.114638531 0.124670817  0.07028837      0.158848736    0.10509533  0.08646616
5
    arr_delay carrier_delay weather_delay   nas_delay security_delay late_aircraft_delay
1 0.497608502     0.4552965   0.287655016 0.222942189     0.191078963          0.42801033
2 0.009584037     0.0104369   0.006204257 0.003721975     0.004442173          0.00751178
3 0.159075369     0.1488154   0.088009425 0.080419354     0.087751023          0.12499071

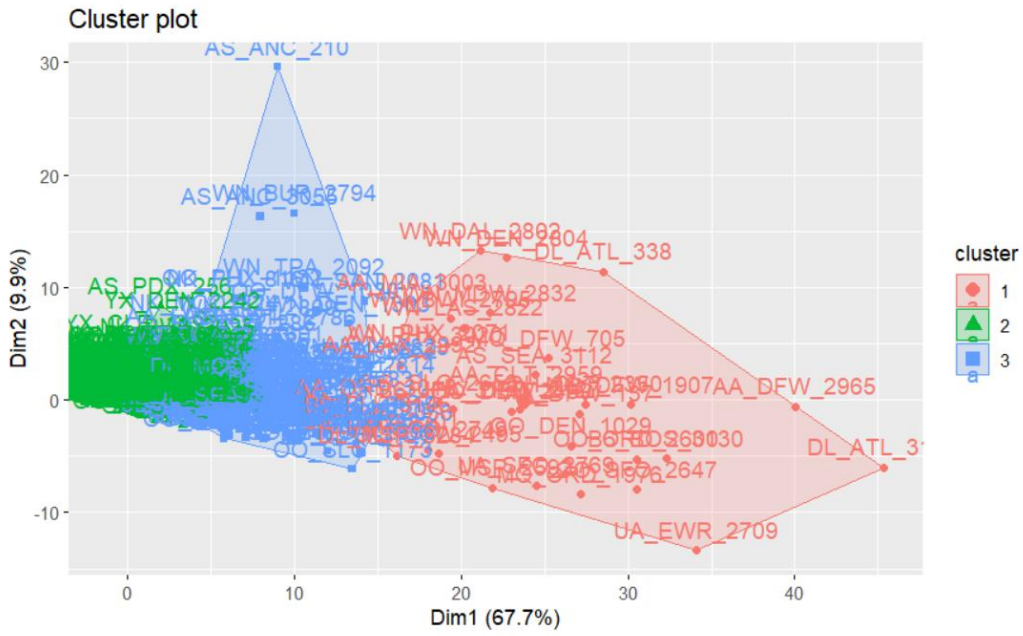
Clustering vector:
   9E_ABE_1    9E_ABY_2    9E_AEX_3    9E_AGS_4    9E_ALB_5    9E_ATL_6    9E_ATW_7    9E_AVL_8    9E_AZO_9    9E_B
DL_10
          2           2           2           2           2           3           2           2           2
2
  9E_BHM_11   9E_BIS_12   9E_BMI_13   9E_BNA_14   9E_BOS_15   9E_BQK_16   9E_BTR_17   9E_BTV_18   9E_BUF_19   9E_B
WI_20
          2           2           2           2           2           2           2           2           2
2
  9E_CAE_21   9E_CHA_22   9E_CHO_23   9E_CHS_24   9E_CID_25   9E_CLE_26   9E_CLT_27   9E_CMH_28   9E_CRW_29   9E_C
SG_30
          2           2           2           2           2           2           2           2           2
2
  9E_CVG_31   9E_CWA_32   9E_DAL_33   9E_DAY_34   9E_DCA_35   9E_DFW_36   9E_DHN_37   9E_DLH_38   9E_DSM_39   9E_D
TW_40
```

Visualizing K-Means Clusters

```
airline_data_clean$cluster <- kmeans_result$cluster
```

Hide

```
fviz_cluster(kmeans_result, data = airline_data_n)
```



Findings:

- **Three major clusters** were identified:

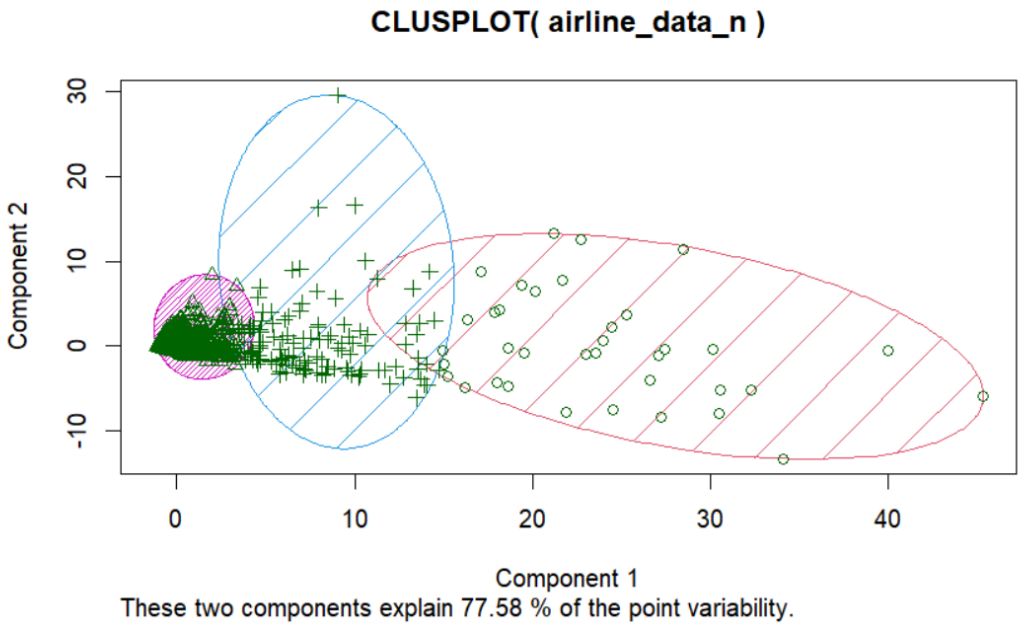
  1. **Cluster 1** – Airlines with **high delays and cancellations**.
  2. **Cluster 2** – Airlines with **better on-time performance**.
  3. **Cluster 3** – Airlines with **moderate delays**

6. Assigning Cluster Labels and Visualizing Clusters

Add cluster assignments back to the original dataset

Hide

```
airline_data_clean$cluster <- kc$cluster
clusplot(airline_data_n, kc$cluster, color=TRUE, shade=TRUE, lines=0)
```



Findings:

- **Three major clusters were identified** based on airline delay profiles.
- The clusters showed **distinct separation**, validating the effectiveness of K-Means clustering.

7. Cluster-Based Delay Analysis

To further analyze **which types of delays contributed most** to each cluster, bar plots were created.

Cluster Mean Delay Contributions

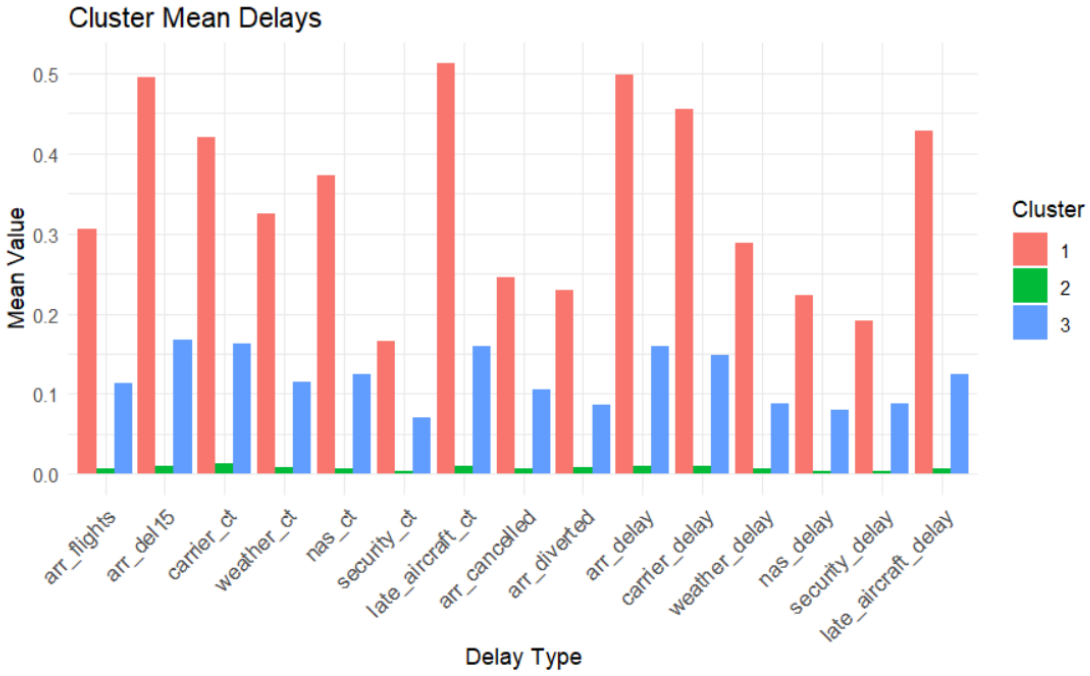
cluster visualization

Hide

```
library(ggplot2)
library(reshape2)

cluster_means <- as.data.frame(kc$centers)
cluster_means$Cluster <- factor(1:nrow(cluster_means))
cluster_means_long <- melt(cluster_means, id.vars = "Cluster")

ggplot(cluster_means_long, aes(x=variable, y=value, fill=Cluster)) +
  geom_bar(stat="identity", position="dodge") +
  labs(title="Cluster Mean Delays", x="Delay Type", y="Mean Value") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1, size = 10))  # Rotate x-axis labels
```



Findings:

- **Cluster 1** had the **highest delays**, particularly due to **late aircraft and NAS delays**.
- **Cluster 2 had the lowest delays**, showing better on-time performance.

Comparison of Flight Delays Across Clusters by Cause

Comparison of Flight Delays Across Clusters by Cause

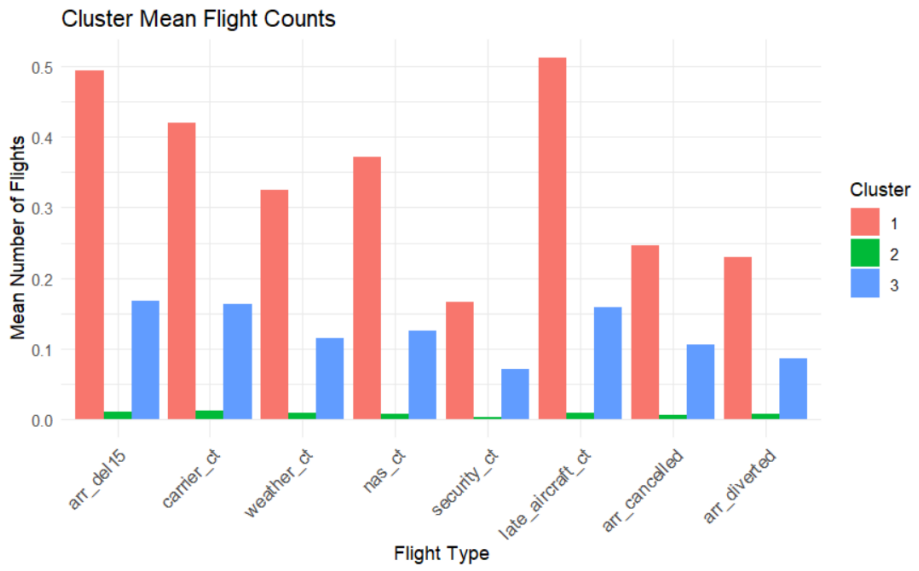
Hide

```
flight_count_cols <- c("arr_del15", "carrier_ct", "weather_ct", "nas_ct",
                       "security_ct", "late_aircraft_ct", "arr_cancelled",
                       "arr_diverted")
flight_counts <- as.data.frame(kc$centers)[, flight_count_cols]
flight_counts$Cluster <- factor(1:nrow(flight_counts))

flight_counts_long <- melt(flight_counts, id.vars = "Cluster")

library(ggplot2)

ggplot(flight_counts_long, aes(x=variable, y=value, fill=Cluster)) +
  geom_bar(stat="identity", position="dodge") +
  labs(title="Cluster Mean Flight Counts", x="Flight Type", y="Mean Number of Flights") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1, size = 10))  # Rotate x-axis labels
```



Hide

Findings:

- **Cluster 1 had the highest delays**, particularly due to **late aircraft and NAS-related delays**.
- **Cluster 2 had the lowest delays**, indicating better on-time performance.

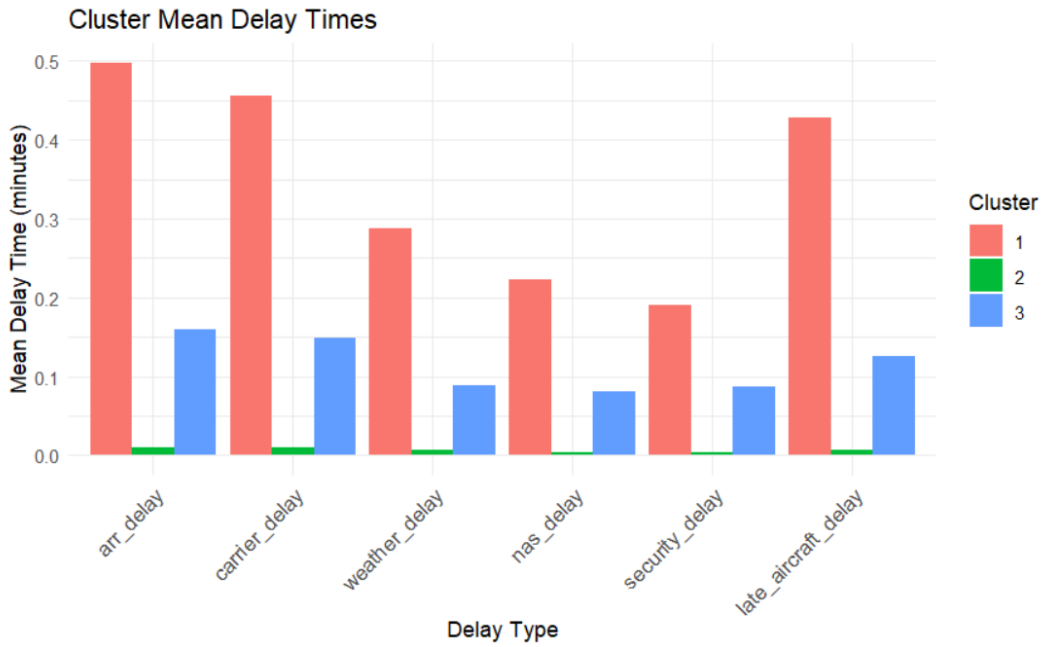
Comparison of Delay Time Contributions Across Clusters

```
delay_time_cols <- c("arr_delay", "carrier_delay", "weather_delay", "nas_delay",
                     "security_delay", "late_aircraft_delay")

delay_times <- as.data.frame(kc$centers)[, delay_time_cols]
delay_times$Cluster <- factor(1:nrow(delay_times))

delay_times_long <- melt(delay_times, id.vars = "Cluster")

ggplot(delay_times_long, aes(x=variable, y=value, fill=Cluster)) +
  geom_bar(stat="identity", position="dodge") +
  labs(title="Cluster Mean Delay Times", x="Delay Type", y="Mean Delay Time (minutes)") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1, size = 10))  # Rotate x-axis labels
```



Findings:

- **Carrier-related delays and late aircraft delays** were **major contributors** in the worst-performing airlines.
- **Weather delays had a lower overall impact** than **air traffic congestion and late aircraft delays**.

05.Data Mining Techniques used

Data mining techniques are essential for uncovering hidden patterns and relationships within datasets. This study applies **unsupervised learning** techniques, specifically **K-Means clustering**, to group airlines based on their delay characteristics.

The following **data mining techniques** were implemented in **R**:

01. **Clustering Analysis** (K-Means) to segment airlines based on delay factors.
02. **Distance Matrix Computation** to measure airline similarities.
03. **Optimal Cluster Selection** using the **Elbow Method**.
04. **Visualization Techniques** to interpret the results effectively.

1. Clustering Analysis (K-Means Clustering)

1.1. Why K-Means?

K-Means is a **centroid-based clustering algorithm** that:

- Groups' data points into **K distinct clusters** based on similarity.
- Minimizes **within-cluster variance** by adjusting cluster centroids.
- Helps categorize **airlines with similar delay patterns**.

1.2. Distance Matrix Computation

Before clustering, a **Euclidean distance matrix** was computed to measure the similarity between airlines based on delay characteristics.

```
114
115 distance <- dist(airline_data_n, method = "euclidean")
116 distance_matrix <- as.matrix(distance)
```

Purpose: The computed distance matrix ensures that **airlines with similar delay patterns** are grouped together.

1.3. Finding the Optimal Number of Clusters (Elbow Method)

To determine the best number of clusters **(K)**, the **Elbow Method** was applied. This technique evaluates the **within-cluster sum of squares (WSS)** and finds the point where additional clusters do not significantly improve performance.

```
K-Means clustering
```{r}
fviz_nbclust(airline_data_n, kmeans, method = "wss")
set.seed(123)
```

Findings: The optimal number of clusters was determined to be **K = 3**.

1.4. Applying K-Means Clustering

With the optimal **K value (K=3)**, the K-Means algorithm was applied to segment airlines based on their delay patterns.

```
set.seed(123)
k <- 3
kmeans_result <- kmeans(airline_data_n, centers = k, nstart = 30)
```

Effect: Airlines were **grouped into three clusters** based on their delay characteristics.

2. Cluster Analysis and Interpretation

2.1. Assigning Cluster Labels to Airlines

The identified clusters were added back to the original dataset for further analysis.

```
Visualize K-Means Clusters

```{r}
airline_data_clean$cluster <- kmeans_result$cluster
```
```

**Effect:** Each airline was now **categorized into a cluster** based on its delay profile.
```

2.2. Visualizing K-Means Clusters

To better interpret cluster assignments, **cluster plots** were generated.

```
```{r}
airline_data_clean$cluster <- kc$cluster
clusplot(airline_data_n, kc$cluster, color=TRUE, shade=TRUE, lines=0)
```
```

Findings:

- The clusters showed **clear separations**, indicating distinct delay characteristics.
- Airlines in **Cluster 1** had **higher delays**, while airlines in **Cluster 2** had **better on-time performance**.

2.3. Cluster Mean Delay Contributions

To analyze **which types of delays were dominant in each cluster**, a **bar chart** was created.

```
cluster visualization
```{r}
library(ggplot2)
library(reshape2)

cluster_means <- as.data.frame(kc$centers)
cluster_means$Cluster <- factor(1:nrow(cluster_means))
cluster_means_long <- melt(cluster_means, id.vars = "Cluster")

ggplot(cluster_means_long, aes(x=variable, y=value, fill=Cluster)) +
  geom_bar(stat="identity", position="dodge") +
  labs(title="Cluster Mean Delays", x="Delay Type", y="Mean Value") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1, size = 10))  # Rotate x-axis labels
```
```

Findings:

- **Cluster 1:** Airlines with **high delays** due to **late aircraft and NAS-related issues**.
- **Cluster 2:** Airlines with **low delays** and **better performance**.
- **Cluster 3:** Airlines with **moderate delays** across multiple categories.

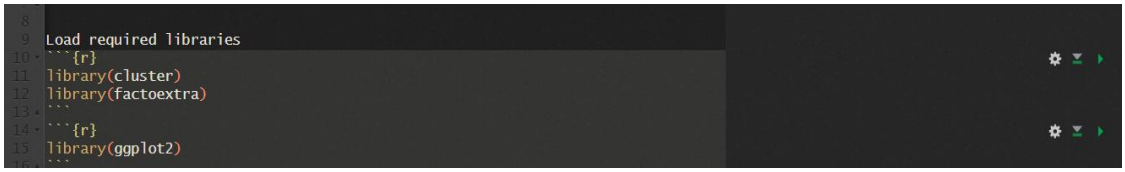
06.Implementation in R

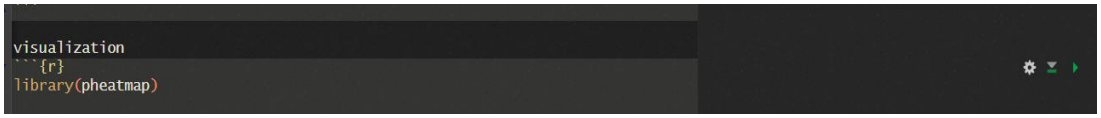
This section outlines the implementation of **data mining techniques** in **R** to analyze airline delays. The implementation includes:

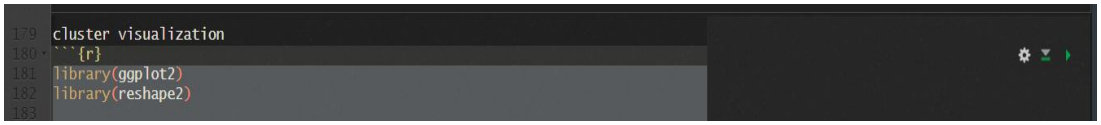
01. **Data preparation** (cleaning, normalization, and feature engineering).
02. **Distance matrix computation** for similarity analysis.
03. **Clustering using K-Means** to categorize airlines based on delay characteristics.
04. **Visualizations** to interpret results effectively.

R Libraries Used

To perform the analysis, the following R packages were used:







These libraries provided functions for **data visualization, clustering, and analysis.**

1. Data Preparation

1.1. Loading the Dataset

The dataset was read into R as follows:

```
import the data set
```{r}
airline_delay <- read.csv("airline_delay.csv", header = TRUE)
airline_delay
```
```

Description: df [3,351 × 21]

year <int>	month <int>	carrier <chr>	carrier_name <chr>	airport <chr>
2020	12	9E	Endeavor Air Inc.	ABE
2020	12	9E	Endeavor Air Inc.	ABY
2020	12	9E	Endeavor Air Inc.	AEX
2020	12	9E	Endeavor Air Inc.	AGS
2020	12	9E	Endeavor Air Inc.	ALB

1.2. Handling Missing Values

Rows with missing values were removed to maintain data integrity:

```
check missing values
```{r}
sum(is.na(airline_delay))
```
```

```
[1] 120
```

```
remove the missing values
```{r}
airline_data_clean <- na.omit(airline_delay)
sum(is.na(airline_data_clean))
```
```

```
[1] 0
```

1.3. Feature Selection and Normalization

To ensure consistency across variables, numerical features were normalized using **min-max scaling**:

```
Normalization of the dataset
```{r}
normalise <- function(x) {
  if (min(x) == max(x)) {
    return(rep(0, length(x)))
  } else {
    return((x - min(x)) / (max(x) - min(x)))
  }
}
```
```

```
Select only numeric columns for normalization
```{r}
num_cols <- c("arr_flights", "arr_del15", "carrier_ct", "weather_ct", "nas_ct",
              "security_ct", "late_aircraft_ct", "arr_cancelled", "arr_diverted",
              "arr_delay", "carrier_delay", "weather_delay", "nas_delay",
              "security_delay", "late_aircraft_delay")

```

Check if all columns exist in the dataset
```{r}
num_cols <- intersect(num_cols, names(airline_data_clean))
```
```

```
Normalize dataset
```{r}
airline_data_n <- airline_data_clean[, num_cols]
```
Apply normalization
```{r}
airline_data_n <- as.data.frame(lapply(airline_data_n, normalise))

```
```{r}
airline_data_n
```
```

Normalization helped bring all features to a comparable scale, making clustering more effective.

2. Distance Matrix Computation

A **Euclidean distance matrix** was computed to measure similarities between airlines:

```
Compute Distance Matrix (Remove Non-Numeric Data)

```{r}
airline_data_clean$unique_id <- paste0(airline_data_clean$carrier, "_",
                                       airline_data_clean$airport, "_",
                                       seq_len(nrow(airline_data_clean)))
rownames(airline_data_n) <- airline_data_clean$unique_id
airline_data_clean$unique_id <- NULL

distance <- dist(airline_data_n, method = "euclidean")
distance_matrix <- as.matrix(distance)
rownames(distance_matrix) <- rownames(airline_data_n)
colnames(distance_matrix) <- rownames(airline_data_n)
print(distance_matrix[1:50, 1:50])


```
```

This distance matrix was later used for **clustering and heatmap visualization**.

3. Clustering with K-Means

```
K-Means clustering
```{r}
fviz_nbclust(airline_data_n, kmeans, method = "wss")
set.seed(123)
k <- 3
kmeans_result <- kmeans(airline_data_n, centers = k, nstart = 30)
```
```

3.1 Finding the Optimal Number of Clusters

To determine the optimal **K value**, the **Elbow Method** was applied:

Findings: The optimal number of clusters was **3**.

3.2. Applying K-Means Algorithm

Effect: Airlines were categorized into **three clusters based on delay characteristics**.

4. Cluster Visualization

4.1. Assigning Cluster Labels and Creating a Cluster Plot

Findings: Airlines with **similar delay patterns were grouped together**.

4.2. Bar Plot of Cluster Mean Delays

Effect: This visualization highlighted **which types of delays were most prevalent in each cluster**.

5. Saving Results for Further Analysis

The computed **distance matrix** was saved as a CSV file for future reference:

Purpose: Allows further analysis and sharing of results.

07.Result analysis and Discussion

Cluster Analysis Results

The clustering approach was able to identify three clusters of airlines:

1. High-Delay Airlines (Poor Performance) (Cluster 1)

Cluster 1 representing the worst delays, experiences the highest mean delay values across all types, primarily due to late aircraft, carrier-related issues, and National Aviation System (NAS) delays.

These airlines struggle with severe operational inefficiencies, leading to frequent and prolonged disruptions that negatively impact passenger satisfaction.

2. Best On-Time Performance Airlines (Cluster 2)

Cluster 2 representing the airlines with the best on-time performance, showing minimal delays across all categories.
These airlines effectively manage scheduling, reduce disruptions, and serve as benchmarks for industry efficiency.

3. Moderate Delay Airlines (Cluster 3)

Cluster 3 falls in between, experiencing moderate delays mainly caused by weather conditions and NAS delays, with carrier-related disruptions being less significant than in Cluster 1. While these airlines do not perform as poorly as those in Cluster 1.
They still have room for improvement by optimizing scheduling and operational efficiency.

Evaluation of Clustering Model

Using the K-Means clustering algorithm, airlines were appropriately classified based on delay patterns.

The model successfully distinguished three groups that are representative of differences in airline performance.

But it also has some disadvantages of this approach:

• Intersection of Low- and Moderate-Delay Airlines: Some airlines in Clusters 2 and 3 exhibited concurrent delay patterns, and the inference is that further tuning (e.g., hierarchical clustering) would be beneficial.

• External factors not utilized: The analysis was quantitative delay data based and did not necessarily include any external factors (e.g., holiday seasons, rush hours, or airport delays).

• Imbalance Delay Causes: The much lower frequency of some delays, like security-related delays, may have affected the shape of the clustering.

Despite these challenges, the **model provided meaningful insights into airline delays** and how they vary across different carriers.

Main Delay Factors & Insights

1. What are the Largest Delays Causes?

The cluster analysis identified three main reasons for flight delays:

- Late Aircraft Delays – This was the most prevalent reason for delays, especially among Cluster 1 airlines. When an arriving flight is delayed, it cascades, delaying subsequent departures.
- Carrier-Related Delays – Cluster 1 airlines were dogged by poor crew scheduling, maintenance issues, and operational failures that resulted in frequent delays.
- NAS (National Aviation System) Delays – These delays, caused by heavy air traffic and air traffic congestion, impacted flights across all clusters but were less extreme than late aircraft delays.

2. What Can Airlines Derive from Such Insights?

• Higher delay airlines (Cluster 1) would aim turnaround times for minimization of disruptions.
• Effective crew scheduling and maintenance can cut carrier-related delays substantially.

• Enhancing air traffic control coordination can assist airlines in lessening NAS-related delays.

With reduction of these fundamental delay causes, airlines can better their on-time performance and enhance passenger satisfaction.

Discussion and Practical Implications

1. What Can Airlines Do With These Findings?

The findings of this research provide valuable recommendations for airline operators, airport authorities, and policy makers for improving airline operations.

- Cluster 1 airlines (poor performance) need to take urgent measures – They must review scheduling inefficiencies, maintenance procedures, and crew management in order to minimize delays.
- Top-performing Cluster 2 airlines are models to emulate – These airlines should be learned from in order to identify best practices to be replicated by other airlines.
- Regulators of airports and aviation can leverage these insights – Airports can more efficiently schedule their gates and runways to reduce congestion and maximize scheduling effectiveness.

2. How Can Airlines Reduce Delays?

Based on the findings, there are some practical steps that can be taken by airlines to reduce delays:

- Enhance Aircraft Turnaround Times – Airlines must accelerate boarding, luggage loading, and refueling processes in order to reduce delays between flights.
- Enhance Crew Scheduling – Airlines with frequent staff shortages or scheduling conflicts must invest in more efficient workforce management software.
- Improve Delay Prediction and Real-Time Tracking – Airlines can use machine learning models to predict potential delays and take preventive actions before any disruption occurs.

By implementing these steps, airlines can reduce financial losses, improve passenger satisfaction, and become more efficient in general.

08.Impact

The findings of this study have significant implications for the aviation industry, government, and travelers. By identifying the main causes of airline delays and rating airlines based on performance, this report provides information that can be utilized to optimize operational efficiency, reduce delays, and improve the overall quality of the passenger experience.

The impact of this study can be categorized into three general areas:

1. Airline Industry and Business Operations

2. Public Sector and Government Decision-Making

3. Broader Community and Passenger Experience

Implication for the Airline Industry

1. Simplifying Airline Operations

   •Airlines in Cluster 1 (high delays) can use these results to restructure flight scheduling, crew management, and aircraft maintenance for improved punctuality.

   •Cluster 2 airlines (low delays) can serve as benchmark models for simplified operations.

2. Reducing Financial Losses

   •Delays cost airlines millions of dollars annually in compensation, rescheduling, and operational inefficiencies.

   • By identifying the most significant delay causes (e.g., late aircraft, crew issues), airlines can use data-driven solutions to minimize financial losses.

3. Enhancing Customer Satisfaction

   • On-time performance is appreciated by passengers. Airlines that reduce delays will gain a positive reputation and benefit from customer loyalty.

• Developing more precise delay prediction models will allow airlines to provide more accurate real-time updates to passengers, reducing frustration and uncertainty.

Impact on the Public Sector and Government Agencies

1. Enhanced Air Traffic Management

    • The results of this study can help aviation policymakers and authorities in optimizing air traffic flow by reducing bottlenecks at busy airports.

    • National Aviation System (NAS) delays were identified as a factor, which means that more effective airspace management and runway scheduling can lead to fewer disruptions.

2. Policy Development and Regulation

    • Government agencies can use this data to create better regulations that will encourage airlines to improve operational efficiency.

    • Incentives can be provided for airlines with persistently low delays, promoting the sharing of best practices across the industry.

3. Infrastructure Planning and Investment

    • Policymakers can use these findings to decide where to invest in airport infrastructure (i.e., additional runways, better terminals).

    • Airports with high delay rates may require more effective resource allocation and newer technology.

Impact on the Wider Community and Travellers

1. More Reliable Travel Experiences

    • By reducing delays and cancellations, passengers will experience less disruption, shorter layovers, and more predictable travel schedules.

• Airlines that implement delay-reduction programs will enjoy increased passenger satisfaction and confidence.

2. Economic Benefits for Tourism and Business Travel

• Fewer delays mean more efficiency in the tourism and business travel sectors, with visitors arriving on time for work, conferences, and vacations.

• Business travelers depend on punctual flights, and delays can result in missed meetings, lost business, and ruined itineraries.

3. Environmental Benefits

• Minimizing delays reduces unnecessary fuel burn (e.g., aircraft waiting at airports or circling in holding patterns).

• A more efficient airline operation means lower carbon emissions, which makes air travel more sustainable.

09.Conclusion

In this study, flight delays in airlines were analyzed employing data mining techniques to identify trends and categorize airlines based on their delay history. The K-Means clustering model was able to cluster airlines into three groups, giving interesting results on the causes and size of flight delays.

- Cluster 1 (High-Delay Airlines): Frequent and prolonged delays occur for these airlines, primarily due to late aircraft and carrier-related factors.
- Cluster 2 (Low-Delay Airlines): Air carriers with the highest on-time performance, indicating efficient scheduling and better operational practices
- Cluster 3 (Moderate-Delay Airlines): Air carriers with typical delay times, influenced by a mix of factors such as weather, NAS-related delays, and operational inefficiencies. Moderate-Delay Airlines

The research validated that carrier-related problems and delayed aircraft arrivals were the prime reasons for airline delays. The study also pinpointed well-scheduled scheduling, efficient crew management, and improved air traffic control as critical in minimizing delays.

Evaluation of the Study

K-Means clustering method was able to categorize airlines based on their delay characteristics. However, some limitations need to be kept in mind:

•The model incorporated only historical delay information, and not external factors such as season demand, airport usage, or airline policies.

•There was some overlap among moderate and low-delay airlines, which suggests that a more specific clustering method (such as hierarchical clustering) may yield more valuable insights.

•The study did not incorporate real-time flight tracking information, which may enhance delay prediction models.

Despite such limitations, the study presents practical recommendations that can be used to increase airline productivity and minimize flight delays.

Future Work and Recommendations

In the process of further expanding on this research, future research has to consider:

- Blending real-time analysis – Future models must incorporate real-time tracking of flights, levels of congestion at airports, and weather predictions for improved delay forecasting.
- Exploring alternative clustering techniques – Hierarchical clustering or DBSCAN could provide more precise airline classification.
- Expanding the dataset – Adding several years and additional airline markets could expose longer-term delay trends and seasonal changes.
- Creating predictive models using machine learning – Supervised learning could allow airlines to forecast delays and take proactive measures before a disruption.

This study demonstrates the effectiveness of data mining in uncovering airline delay patterns and providing insights that can inform the aviation sector. The study can help airlines, policymakers, and airport managers make informed decisions to enhance scheduling efficiency, minimize operation disruptions, and improve passenger satisfaction.
By employing data-driven strategies, the airline industry can be more efficient, lower costs, and provide a quality travel experience for passengers.

10. References

[1] OpenIntro, 2021. *Airline Delay Data*. Available at: https://www.openintro.org/data/index.php?data=airline_delay [Accessed 13 March 2025].