The Agony of Delays: A Look at Airline Clusters

James Down & Julia Hollis
Faculty Advisor: Dr. Xinyan Zhang

KENNESAW STATE UNIVERSITY
COLLEGE OF COMPUTING AND SOFTWARE ENGINEERING
School of Data Science and Analytics

INTRODUCTION

Last year, there were a total of 7 million domestic flights, with over 20% of them being delayed. An additional 3% were outright cancelled. Which airlines are the best at keeping delays at a lower rate? Are there any differences between the airlines in the COVID era? There is anecdotal evidence that the general population likes to rely on but in this study, we attempt to identify clusters of airlines when modeling to delays and cancellations.

METHODS

Flight delay and cancellation data from March 2020-November 2022 were obtained from the Bureau of Transportation Website.

- More recent data were not available at time of collection.

8 Variables were selected to be analyzed in a Clustering Model

- Arrival Delay, Departure Delay, Cancelled Flight, Carrier Delay, Security Delay, NAS Delay, and Late Aircraft Delay
- Airline was selected as the reference factor.

Observations for each variable were summed to create an overall snapshot per Airline.

- Summations were then standardized to the normal distribution.

Hierarchical clustering was chosen over k-means clustering due to the small number of observations.

3 types of clustering techniques were used, Complete, Single and Ward's to determine the best clustering model.

Ward's method was chosen as the best clustering model.

- Complete and Single method clustering models were not as distributed as the Ward's method model.
- Clusters were grouped based on the Ward's method dendrogram.

To determine how close the clusters were in similarity, a linear discrimination analysis was conducted.

- Plotting the first function versus the second function shows the distinctive shapes and relative distance between clusters.

Airline factors were then averaged and standardized by cluster and then graphed to determine the differences between each factor.

James Down
(MSAS)
Graduation: December 2023

Julia Hollis
(MSAS)
Graduation: May 2023

SCAN ME SCAN ME

Figure 1. Cluster Dendrogram of Airlines

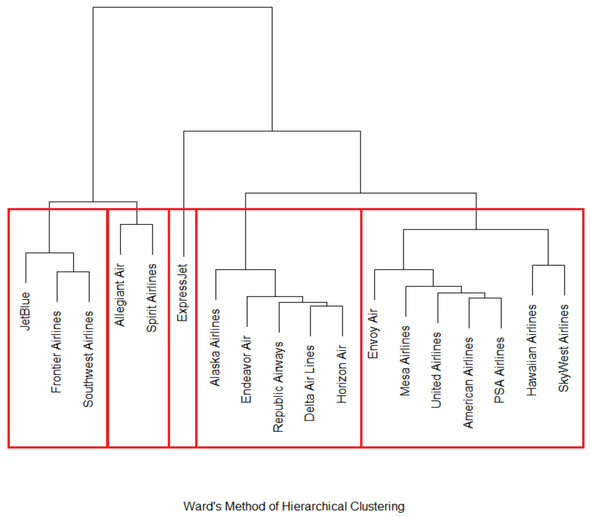


Figure 2: 2D Representation of Airline Clusters

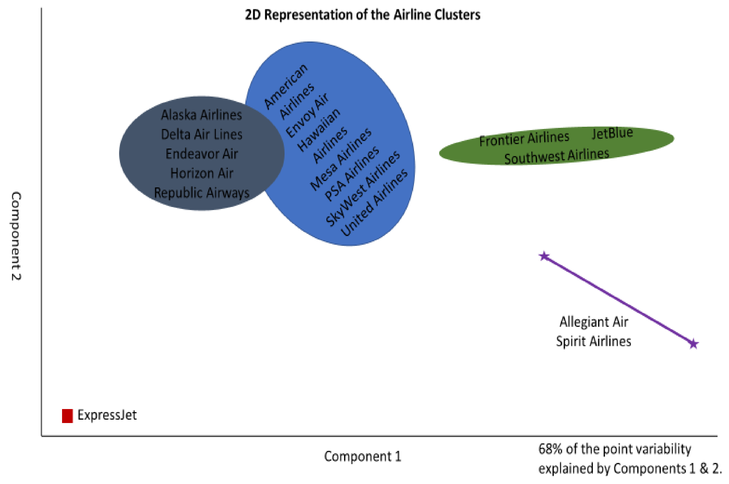
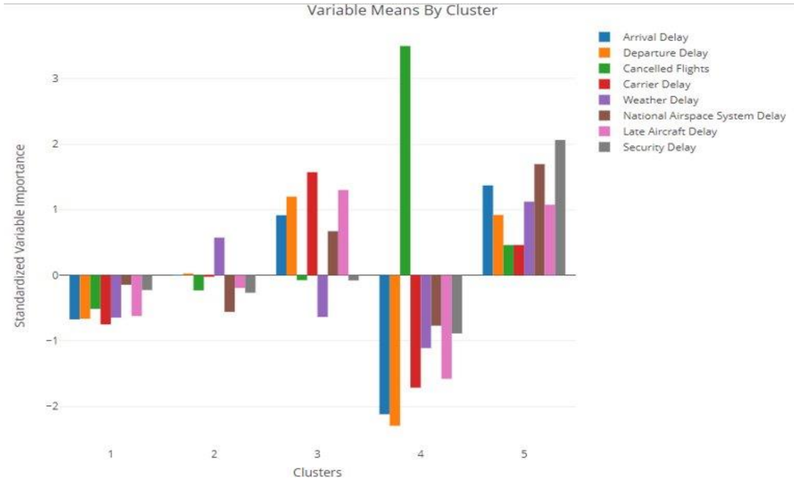


Figure 3. Variable Means by Cluster



RESULTS

Table 1: Cluster Analysis

Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5
Alaska Airlines	American Airlines	Frontier Airlines	ExpressJet	Allegiant Air
Delta Air Lines	Envoy Air	JetBlue		Spirit Airlines
Endeavor Air	Hawaiian Airlines	Southwest Airlines		
Horizon Air	Mesa Airlines			
Republic Airways	PSA Airlines			
	Skywest Airlines			
	United Airlines			

The clustering analysis results in the identification of five distinctive groups of airlines.

Figure 1: Cluster Dendrogram
The Ward's method dendrogram provides a visual representation of the hierarchical relationships between the different airlines.

Figure 2: 2D Cluster Representation
Using the clusplot function, a two-dimensional graph was derived from Ward's method cluster analysis. This plot easily illustrates which airlines belong to which cluster through the different shapes and colors.

Figure 3: Variable Means
In the 'Variable Means by Cluster' graph, the sign of the standardized values indicates whether a delay variable is above or below the mean for all the clusters combined. For example, a positive value for the 'Departure Delay' variable indicates that flights in cluster 5 tend to have longer departure delays than the mean departure delay time.

Flights from cluster 1 exhibit the most favorable outcomes in terms of departure and arrival delay times. It is recommended to fly with Alaska, Delta, Endeavor, Horizon, and Republic for fewer delays and cancellations.

R CODE

```
library(readr)
library(tidyverse)
library(dplyr)
library(cluster)
library(plotly)

#Standardized Variables in the Dataset
df5 %>%
  mutate(across(where(is.numeric), scale)) -> df5.s
 #Created distances between Arilines.
distance = dist(df5.s)
#Ran a Ward's D hierarchical agglomerative approach to cluster the Airlines
d.hc.ward <- hclust(distance, method = "ward.D")
plot(d.hc.ward, labels=df5.s$Reporting_Airline,
     main = "Airline Cluster Dendrogram",
     sub = "Ward's Method of Hierarchical Clustering",
     axes = "False", ylab = NULL, xlab = " ")

#Determined number of clusters based on dendrogram and displayed the clusters
cutree(d.hc.ward, 5)
rect.hclust(d.hc.ward, k= 5)

#Plotted the clusters in a 2D dimension to show the differences between clusters.
clusplot(df6.merge, df6.merge$cluster,
main='2D Representation of the Airline Clusters',
         color=TRUE, shade=FALSE,
         labels=4, lines=0, col.p=df6.merge$cluster,
        col.clus=df6.merge$cluster)

#Average factors across clusters to determine differences between clusters.
df8 <- df7 %>% group_by(cluster)%>%
 summarise(across(everything(),mean),
           .groups = 'drop') %>%
 as.data.frame()

#Created bar graph showing factor differences between clusters.
fig <- plot_ly(df8, x = ~cluster, y = ~ArrDelay,
type = 'bar', name = 'Arrival Delay')
fig <- fig %>% add_trace(y = ~DepDelay, name
= 'Departure Delay')
fig <- fig %>% add_trace(y = ~Cancelled, name
= 'Cancelled Flights')
fig <- fig %>% add_trace(y = ~CarrierDelay,
name = 'Carrier Delay')
fig <- fig %>% add_trace(y = ~WeatherDelay,
name = 'Weather Delay')
fig <- fig %>% add_trace(y = ~NASDelay, name
= 'National Airspace System Delay')
fig <- fig %>% add_trace(y = ~LateAircraftDelay,
name = 'Late Aircraft Delay')
fig <- fig %>% add_trace(y = ~SecurityDelay,
name = 'Security Delay')
fig <- fig %>% layout(title = "Variable Means
By Cluster",
       yaxis = list(title = 'Standardized
Variable Importance'),xaxis = list(title
= 'Clusters'), barmode = 'group')
```

DISCUSSION

We were surprised by the diverse clustering of airlines, which did not follow our initial hypothesis that big brand names and regional lines would be clustered together. ExpressJet's high rate of cancellations could be explained by their recent bankruptcy, which was discovered through further research. Southwest Airlines also had higher than expected delays, but future analysis with more recent data may show them belonging to a different cluster especially with the higher number of cancellations they experienced in December 2022. To improve predictions of delays, future studies should analyze additional flight factors. Customers should book with Alaska, Delta, Endeavor, Horizon, or Republic, and avoid Allegiant, Frontier, JetBlue, Southwest, and Spirit Airlines if possible.