Riley Breske
CSC-161

Exploratory Data Analysis: Airline Delays

Project Overview:

In this project, I conducted an exploratory data analysis (EDA) to investigate the relationship between various factors, such as flight volume and delay times, in the airline industry. My goal was to determine if the overall flight volume affected the average delay times and to explore potential correlations between these variables.

Key Findings:

1. **Overall Flight Volume vs. Average Delay:**
    - After analyzing the data, it was found that the overall flight volume had no significant effect on average delay times. Contrary to the initial assumption, larger airlines like United Airlines and JetBlue, despite having higher volumes, did not exhibit a correlation between flight volume and delay time.
2. **Impact of Day of the Week on Delays:**
    - The analysis revealed a strong correlation (97%) between the number of flights per day of the week and the average delay time for that day. Specifically, Thursdays and Saturdays, which had the highest number of flights, also experienced the most significant delays.
    - A potential solution to mitigate these delays could involve increasing ticket prices on high-traffic days to discourage travel, thereby reducing the number of flights and associated delays.
3. **Airline Rankings Based on Delay:**
    - United Airlines and JetBlue were found to have the highest average delays among the analyzed airlines. This led to the hypothesis that larger airlines might have higher delays, but the analysis showed no significant correlation between the size of the airline (in terms of flight volume) and delay times.

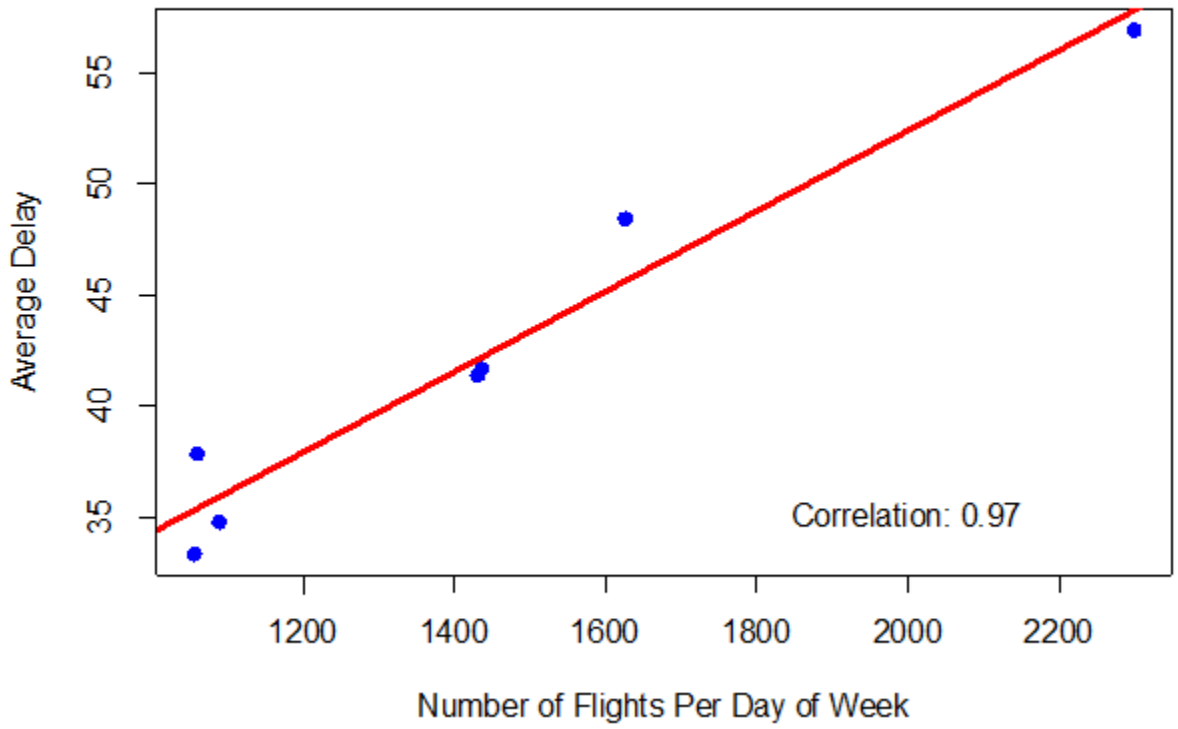
Technical Details:

- **Data Preparation:**
    - The dataset was read in using `read.csv`, and necessary packages like `Hmisc`, `tidyr`, and `ggplot2` were installed and loaded.
    - Data was aggregated to calculate the average delay per airline and per day of the week. Counts for the number of flights per carrier and per day of the week were also computed.
- **Visualization and Analysis:**
    - Various visualizations were created using `ggplot2`, including bar charts to rank airlines by average delay and to illustrate the relationship between flight volume and delay times.

    - Correlations were analyzed using Pearson correlation coefficients, and linear regression models were fitted to determine the significance of these correlations.
- **Key Statistical Insights:**
    - **Correlation Between Day of Month and Departure Delay:** No significant correlation was found.
    - **Correlation Between Flight Volume and Average Delay:** No significant relationship was observed between the number of flights an airline operates and the average delay experienced.
    - **Correlation Between Day of Week and Delay:** A statistically significant positive correlation was observed, with an R^2 value of 0.95, indicating a strong relationship between the number of flights on a given day and the average delay experienced.

Conclusion:

The analysis provided valuable insights into the factors contributing to flight delays, particularly the impact of the day of the week on delay times. While overall flight volume did not significantly affect delays, the concentration of flights on certain days resulted in longer delays. These findings could be used to develop strategies for mitigating delays, such as adjusting ticket prices on high-traffic days.



```r
# Reads in data
MyData = read.csv("Airlines.csv")

# Installs necessary packages
install.packages("Hmisc")
library(Hmisc)
install.packages("tidyr")
library(tidyr)
install.packages("ggplot2")
library(ggplot2)
install.packages("remotes")
remotes::install_github("easystats/report")
library(report)

# Creates a count for flights per carrier
NumFlights = MyData %>% count(UniqueCarrier)

# Creates count for number of flights per day of the week
NumFlightsWk = MyData %>% count(DayOfWeek)

# Creates average flight delay data set, then adds total observations for
number of flights per airline to that.
AvgDelay = aggregate(MyData$DepDelay, list(MyData$UniqueCarrier), mean)
AvgDelay$NumFlights = as.integer(NumFlights$n)

# Creates data set to analyze average flight delay for each day of the
week.
AvgDelayWk = aggregate(MyData$DepDelay, list(MyData$DayOfWeek), mean)
AvgDelayWk$NumFlights = as.integer(NumFlightsWk$n)

# Brief overview of data
summary(MyData)
ls(MyData)
summary(MyData$ArrDelay)

# Correlation between day of month and departure delay
set.seed(1)
x <- MyData$DayofMonth
y <- MyData$DepDelay

# Creating the plot
plot(x, y, xlab="Day of Month", ylab = "Departure Delay", pch = 19, col =
"lightblue")
```

```r
# Regression line
abline(lm(y ~ x), col = "red", lwd = 3)

# Pearson correlation
text(paste("Correlation:", round(cor(x, y), 2)), x = 25, y = 95)
## No correlation

# Ranking of Departure Delay by Carrier (Average not considered)
ggplot(MyData, aes(x=reorder(UniqueCarrier, ArrDelay), y=ArrDelay)) +
geom_bar(stat ='identity', aes(fill=ArrDelay)) +
coord_flip() +
theme_grey() +
scale_fill_gradient(name="Arrival Delay") +
labs(title = 'Ranking of Arrival Delay by Airline', y='Arrival Delay',
x='Airlines') +
geom_hline(yintercept = mean(MyData$ArrDelay), size = 1, color = 'blue')
## Southwest Airlines has the most delays (likely due to pure volume)

# Ranking of AVERAGE Departure Delay by Carrier
ggplot(AvgDelay, aes(x=reorder(Group.1, x), y=x)) +
geom_bar(stat ='identity', aes(fill=x)) +
coord_flip() +
theme_grey() +
scale_fill_gradient(name="Average Arrival Delay") +
labs(title = 'Ranking of Average Arrival Delay by Airline', y='Avg. Arrival
Delay', x='Airlines') +
geom_hline(yintercept = mean(AvgDelay$x), size = 1, color = 'blue')
## United Airlines has the highest average delay

# Correlation between number of flights per airline and average delay
set.seed(1)
x <- AvgDelay$NumFlights
y <- AvgDelay$x

# Creating the plot
plot(x, y, xlab="Number of Flights Per Airline", ylab = "Average Delay",
pch = 19, col = "blue")

# Regression line
abline(lm(y ~ x), col = "red", lwd = 3)

# Pearson correlation
```

```
text(paste("Correlation:", round(cor(x, y), 2)), x = 1230, y = 33)
## No correlation

summary(lm(y ~ x))
report(lm(y ~ x))
## Based on the report, the more flights an airline has does not increase
the average delay they have.

# Ranking of AVERAGE Departure Delay by Day of the Week
ggplot(AvgDelayWk, aes(x=reorder(Group.1, x), y=x)) +
geom_bar(stat ='identity', aes(fill=x)) +
coord_flip() +
theme_grey() +
scale_fill_gradient(name="Average Arrival Delay") +
labs(title = 'Ranking of Average Arrival Delay by Day of Week', y='Avg.
Arrival Delay', x='Day of Week') +
geom_hline(yintercept = mean(AvgDelayWk$x), size = 1, color = 'red')
## Day 7 and 5, respectively, have the average delays that are the highest,
while also being above the average.

# Correlation between number of flights per day and average delay
set.seed(1)
x <- AvgDelayWk$NumFlights
y <- AvgDelayWk$x

# Creating the plot
plot(x, y, xlab="Number of Flights Per Day of Week", ylab = "Average
Delay", pch = 19, col = "blue")

# Regression line
abline(lm(y ~ x), col = "red", lwd = 3)

# Pearson correlation
text(paste("Correlation:", round(cor(x, y), 2)), x = 2000, y = 35)

summary(lm(y ~ x))
report(lm(y ~ x))
```