STAT 184 Final Project: Exploring Flight Status During the Week of Christmas

Felicia Vijayarangam, Varsha Giridharan

2024-12-18

Figure 1: Flight Departure Board

Introduction

The holiday season is one of the busiest times of the year because everyone is traveling to see family, going on vacations, and getting away from the everyday chaos of life. With this, one of the biggest hurdles is transportation. Millions of individuals are either driving or flying during this time, creating an increase in traffic and delays nationwide. In particular, how many of

you have seen headlines on the news stating, "Hundreds of Flights are Delayed Due to Holiday Rush?" Whether it be for Memorial Day weekend, Thanksgiving, Christmas, or any other holiday, various factors play a part in delays such as a flight's departure delay, arrival delay, weather delay, or carrier delay. There could also be overall flight diversions or cancellations that play a part in this.

In our report, we investigated flight delays during the Christmas Season. We have two specific datasets from the United States Department of Transportation compiled by the Bureau of Transformation Statistics. One dataset is from December 2020, and the other is from December 2023.

Both these datasets adhere to the FAIR principles. The datasets are findable and accessible to the general public since it is posted on an open-platform government website: Flight Information. The website provides a way to filter through what months, years, and columns a researcher wants to look into and descriptions for each, making the dataset reusable for various purposes. Then, after selecting the desired attributes, researchers can click the download button, saving to file as T_ONTIME_MARKETING.csv. This CSV format makes the file interoperable.

Both these datasets adhere to the CARE principles. By looking at the dataset, there are collective benefits since there are several research questions such as what flights/airlines have the most delays, etc that could be answered, helping travelers decide what airline to fly. Since this is a government dataset, the airlines have given consent to share flight information with individuals. Our data sets do not have sensitive data, allowing it to adhere to the responsible and ethical criteria.

Within the December 2020 dataset, there are 397,208 cases, each representing individual flights. In comparison, December 2023 had 606,218 cases of flights. Notice how in December 2023 the number of flights almost doubled. This is because, in 2020, there was COVID-19, so the number of individuals traveling greatly decreased. Throughout our analysis, we want to compare these years and look at 34 attributes (later described in detail) such as delay times.

These are our main research questions:

1. What is the average delay time for the week of Christmas 2020 for the top 10 most popular U.S. airlines?
2. What is the average delay time for the week of Christmas 2023 for the top 10 most popular U.S. airlines?
3. How does the average delay time for the week of Christmas for both years compare to each other?

Data Inventory

In order to start answering these research questions, we first need to load the December 2020 and December 2023 data set and the libraries needed for our data visualization. In this case, we will use libraries such dplyr, tidyr, kableExtra, stringr, ggplot2, scales, knitr, readr. Once we have these libraries, we can proceed to the data exploration step.

Data Exploration

Throughout our analysis, we want to compare both these years together and look at various attributes relationships. One of the challenges we faced was that looking at the entire month of dataset was hard to do because of how big the data files are. December 2020 contains 397802 flights and December 2023 contains 606218 flights. Each of these data sets have the following 34 attributes:

```
 [1] "YEAR"                 "DAY_OF_MONTH"         "FL_DATE"
 [4] "OP_UNIQUE_CARRIER"    "OP_CARRIER"           "OP_CARRIER_FL_NUM"
 [7] "ORIGIN_AIRPORT_ID"    "ORIGIN"               "ORIGIN_CITY_NAME"
[10] "DEST_AIRPORT_ID"      "DEST"                 "DEST_CITY_NAME"
[13] "DEP_DELAY_NEW"        "ARR_DELAY_NEW"        "CANCELLED"
[16] "DIVERTED"             "FLIGHTS"              "CARRIER_DELAY"
[19] "WEATHER_DELAY"        "NAS_DELAY"            "SECURITY_DELAY"
[22] "LATE_AIRCRAFT_DELAY" "DIV_AIRPORT_LANDINGS" "DIV_ARR_DELAY"
[25] "DIV1_AIRPORT"         "DIV1_AIRPORT_ID"      "DIV2_AIRPORT"
[28] "DIV2_AIRPORT_ID"      "DIV3_AIRPORT"         "DIV3_AIRPORT_ID"
[31] "DIV4_AIRPORT"         "DIV4_AIRPORT_ID"      "DIV5_AIRPORT"
[34] "DIV5_AIRPORT_ID"
```

Since this governmental dataset looks at all the flights traveling nationwide, we decided it was best to create a new dataset to analyze. Since we are concerned about the holiday rush, we decided to focus on the Christmas season and make a specific subset that looks at the week of Christmas.

In our subset, we want to use the filter() in order to filter the original datasets with the whole month of December to the specific week of Christmas. Based on the columns, we decided to filter based on the Day of the Month. Now, we will have two new datasets with only the flights for Christmas week.

```r
#Filter December 2020:
christmas_week2020 <- flights2020 %>%
  filter(DAY_OF_MONTH >= 20 & DAY_OF_MONTH <= 27)

#Filter December 2023:
christmas_week2023 <- flights2023 %>%
  filter(DAY_OF_MONTH >= 20 & DAY_OF_MONTH <= 27)
```

We want to create summary() functions to understanding the data types and distributions of each of the attributes represented. In order to have the summary() properly formatted and have all 34 attributes fit into the pdf, we flipped the rows and the column values. With this, now the columns are the summary statistics and the rows are the data sets attributes. Therefore we used the tranpose function t(). To create a summary table, we used the kable(), kable_styling(), column_spec(), and row_spec() for styling. We also needed the package kableExtra().

In the first table, we see the flight summary table for the week of Christmas in 2020.

```
Warning in styling_latex_scale(out, table_info, "down"): Longtable cannot be
resized.
```

Table 1: Flight Summary Table of Christmas Week 2020

YEAR	Min. :2020	1st Qu.:2020	Median :2020	Mean :2020	3rd Qu.:2020	Max. :2020	NA
DAY_OF_MONTH	Min. :20.00	1st Qu.:21.00	Median :23.00	Mean :23.43	3rd Qu.:26.00	Max. :27.00	NA
FL_DATE	Length:11021	Class :character	Mode :character	NA	NA	NA	NA
OP_UNIQUE_CARRIER	Length:110216	Class :character	Mode :character	NA	NA	NA	NA
OP_CARRIER	Length:11021	Class :character	Mode :character	NA	NA	NA	NA
OP_CARRIER_FL_NUM	Min. : 1	1st Qu.:1108	Median :2366	Mean :2787	3rd Qu.:4575	Max. :8802	NA
ORIGIN_AIRPORT_ID	Min. :10135	1st Qu.:11292	Median :12884	Mean :12663	3rd Qu.:14098	Max. :16869	NA
ORIGIN	Length:110216	Class :character	Mode :character	NA	NA	NA	NA
ORIGIN_CITY_NAME	Length:11021	Class :character	Mode :character	NA	NA	NA	NA
DEST_AIRPORT_ID	Min. :10135	1st Qu.:11292	Median :12884	Mean :12662	3rd Qu.:14098	Max. :16869	NA
DEST	Length:11021	Class :character	Mode :character	NA	NA	NA	NA

DEST_CITY_NAME	Length:110216	Class :character	Mode :character	NA	NA	NA	NA
DEP_DELAY_NEW	Min. : 0.000	1st Qu.: 0.000	Median : 0.000	Mean : 9.032	3rd Qu.: 1.000	Max. :1779.000	NA's :1108
ARR_DELAY_NEW	Min. : 0.000	1st Qu.: 0.000	Median : 0.000	Mean : 9.311	3rd Qu.: 3.000	Max. :1776.000	NA's :1335
CANCELLED	Min. :0.00000	1st Qu.:0.00000	Median :0.00000	Mean :0.01047	3rd Qu.:0.00000	Max. :1.00000	NA
DIVERTED	Min. :0.000000	1st Qu.:0.000000	Median :0.000000	Mean :0.001642	3rd Qu.:0.000000	Max. :1.000000	NA
FLIGHTS	Min. :1	1st Qu.:1	Median :1	Mean :1	3rd Qu.:1	Max. :1	NA
CARRIER_DELAY	Min. : 0.00	1st Qu.: 0.00	Median : 4.00	Mean : 27.16	3rd Qu.: 26.00	Max. :1776.00	NA's :95279
WEATHER_DELAY	Min. : 0.00	1st Qu.: 0.00	Median : 0.00	Mean : 5.53	3rd Qu.: 0.00	Max. :1593.00	NA's :95279
NAS_DELAY	Min. : 0.00	1st Qu.: 0.00	Median : 1.00	Mean : 12.47	3rd Qu.: 18.00	Max. :1224.00	NA's :95279
SECURITY_DELAY	Min. : 0.00	1st Qu.: 0.00	Median : 0.00	Mean : 0.14	3rd Qu.: 0.00	Max. :127.00	NA's :95279
LATE_AIRCRAFT_DELAY	Min. : 0.00	1st Qu.: 0.00	Median : 0.00	Mean : 15.31	3rd Qu.: 14.00	Max. :1209.00	NA's :95279
DIV_AIRPORT_LANDIN	Min. :0.000000	1st Qu.:0.000000	Median :0.000000	Mean :0.003684	3rd Qu.:0.000000	Max. :9.000000	NA
DIV_ARR_DELAY	Min. : 44.0	1st Qu.: 125.0	Median : 174.0	Mean : 261.6	3rd Qu.: 253.0	Max. :1192.0	NA's :110063
DIV1_AIRPORT	Length:11021	Class :character	Mode :character	NA	NA	NA	NA
DIV1_AIRPORT_ID	Min. :10135	1st Qu.:11282	Median :12265	Mean :12642	3rd Qu.:14089	Max. :16101	NA's :110010
DIV2_AIRPORT	Length:11021	Class :character	Mode :character	NA	NA	NA	NA
DIV2_AIRPORT_ID	Min. :12889	1st Qu.:12889	Median :12889	Mean :12889	3rd Qu.:12889	Max. :12889	NA's :110215
DIV3_AIRPORT	Mode:logical	NA's:110216	NA	NA	NA	NA	NA
DIV3_AIRPORT_ID	Mode:logical	NA's:110216	NA	NA	NA	NA	NA
DIV4_AIRPORT	Mode:logical	NA's:110216	NA	NA	NA	NA	NA
DIV4_AIRPORT_ID	Mode:logical	NA's:110216	NA	NA	NA	NA	NA
DIV5_AIRPORT	Mode:logical	NA's:110216	NA	NA	NA	NA	NA
DIV5_AIRPORT_ID	Mode:logical	NA's:110216	NA	NA	NA	NA	NA

In the second table, we see the flight summary table for the week of Christmas in 2023.

```
Warning in styling_latex_scale(out, table_info, "down"): Longtable cannot be
resized.
```

Table 2: Flight Summary Table of Christmas Week 2020

YEAR	Min. :2023	1st Qu.:2023	Median :2023	Mean :2023	3rd Qu.:2023	Max. :2023	NA

DAY_OF_MONTH	Min. :20.00	1st Qu.:21.00	Median :23.00	Mean :23.45	3rd Qu.:26.00	Max. :27.00	NA
FL_DATE	Length:15721	Class :character	Mode :character	NA	NA	NA	NA
OP_UNIQUE_CARRIER	Length:157219	Class :character	Mode :character	NA	NA	NA	NA
OP_CARRIER	Length:15721	Class :character	Mode :character	NA	NA	NA	NA
OP_CARRIER_FL_NUM	Min. : 1	1st Qu.:1090	Median :2115	Mean :2423	3rd Qu.:3502	Max. :9658	NA
ORIGIN_AIRPORT_ID	Min. :10135	1st Qu.:11292	Median :12889	Mean :12674	3rd Qu.:14057	Max. :16869	NA
ORIGIN	Length:157219	Class :character	Mode :character	NA	NA	NA	NA
ORIGIN_CITY_NAME	Length:15721	Class :character	Mode :character	NA	NA	NA	NA
DEST_AIRPORT_ID	Min. :10135	1st Qu.:11292	Median :12889	Mean :12674	3rd Qu.:14057	Max. :16869	NA
DEST	Length:15721	Class :character	Mode :character	NA	NA	NA	NA
DEST_CITY_NAME	Length:157219	Class :character	Mode :character	NA	NA	NA	NA
DEP_DELAY_NEW	Min. : 0.00	1st Qu.: 0.00	Median : 0.00	Mean : 15.11	3rd Qu.: 10.00	Max. :3786.00	NA's :1157
ARR_DELAY_NEW	Min. : 0.00	1st Qu.: 0.00	Median : 0.00	Mean : 13.96	3rd Qu.: 8.00	Max. :3795.00	NA's :1660
CANCELLED	Min. :0.000000	1st Qu.:0.000000	Median :0.000000	Mean :0.007684	3rd Qu.:0.000000	Max. :1.000000	NA
DIVERTED	Min. :0.000000	1st Qu.:0.000000	Median :0.000000	Mean :0.002875	3rd Qu.:0.000000	Max. :1.000000	NA
FLIGHTS	Min. :1	1st Qu.:1	Median :1	Mean :1	3rd Qu.:1	Max. :1	NA
CARRIER_DELAY	Min. : 0.0	1st Qu.: 0.0	Median : 4.0	Mean : 22.6	3rd Qu.: 21.0	Max. :3786.0	NA's :127460
WEATHER_DELAY	Min. : 0.00	1st Qu.: 0.00	Median : 0.00	Mean : 5.36	3rd Qu.: 0.00	Max. :1496.00	NA's :127460
NAS_DELAY	Min. : 0.00	1st Qu.: 0.00	Median : 0.00	Mean : 11.06	3rd Qu.: 15.00	Max. :1260.00	NA's :127460
SECURITY_DELAY	Min. : 0.00	1st Qu.: 0.00	Median : 0.00	Mean : 0.19	3rd Qu.: 0.00	Max. :242.00	NA's :127460
LATE_AIRCRAFT_DELAY	Min. : 0.0	1st Qu.: 0.0	Median : 4.0	Mean : 28.4	3rd Qu.: 31.0	Max. :1606.0	NA's :127460
DIV_AIRPORT_LANDIN	Min. :0.000000	1st Qu.:0.000000	Median :0.000000	Mean :0.004929	3rd Qu.:0.000000	Max. :9.000000	NA
DIV_ARR_DELAY	Min. : 41.0	1st Qu.: 154.5	Median : 220.0	Mean : 366.5	3rd Qu.: 452.5	Max. :1388.0	NA's :156872
DIV1_AIRPORT	Length:15721	Class :character	Mode :character	NA	NA	NA	NA
DIV1_AIRPORT_ID	Min. :10140	1st Qu.:11292	Median :12953	Mean :12887	3rd Qu.:14492	Max. :15380	NA's :156732
DIV2_AIRPORT	Length:15721	Class :character	Mode :character	NA	NA	NA	NA
DIV2_AIRPORT_ID	Min. :10466	1st Qu.:11041	Median :11941	Mean :12284	3rd Qu.:13351	Max. :14869	NA's :157203
DIV3_AIRPORT	Mode:logical	NA's:157219	NA	NA	NA	NA	NA
DIV3_AIRPORT_ID	Mode:logical	NA's:157219	NA	NA	NA	NA	NA

DIV4_AIRPORT	Mode:logical	NA's:157219	NA	NA	NA	NA	NA
DIV4_AIRPORT_ID	Mode:logical	NA's:157219	NA	NA	NA	NA	NA
DIV5_AIRPORT	Mode:logical	NA's:157219	NA	NA	NA	NA	NA
DIV5_AIRPORT_ID	Mode:logical	NA's:157219	NA	NA	NA	NA	NA

At a high level, this table helps us pinpoint what attributes are most meaningful for us to analyze and compare performance against. There are various categorical variables such as year, day of the month, OP unique carrier, OP carrier, OP carrier flight number, origin, city name, destination, and such. For numerical values, keep in mind what they represent. For example, ID is a numerical value, but it is categorical in this case. Each airport has a specific ID number, so the mean, median, and such do not mean anything. However, the numerical numbers such as departure delays, arrival delays, flights carrier delays, weather delays, NAS delays, security delays, late aircraft delays, and diverted airport delays do matter. In our research questions, we can use these numerical attributes as checkpoints to see how airlines performance is compared to each other.

With these thoughts in mind, let us begin answering our research questions.

Research Question 1: What is the average departure delay time for the week of Christmas 2020 for the top 10 most popular U.S. airlines?

Before we answer this question. Let us understand what we are being asked to do. We want to find the average departure delay time for each day of Christmas week based on the top ten airlines. In the dataset that we have, the airlines are represented by the OP_UNIQUE_CARRIER column. However, our dataset does not provide a key to what letters and numbers represents a specific airline. Therefore, we found a new data source with two column, one of airline codes and the other with airline names.

With this information, we could use the left_join() to create a new dataset with a new column of airline names. The logic behind this was that we needed to left_join() by the condition that the new data sources code matched the orginal's **unique carrier code**.

```
# Merge the datasets based on the airline code
airlineInfo2020 <- christmas_week2020 %>%
  left_join(airlineCode, by = c("OP_UNIQUE_CARRIER" = "Code"))
```

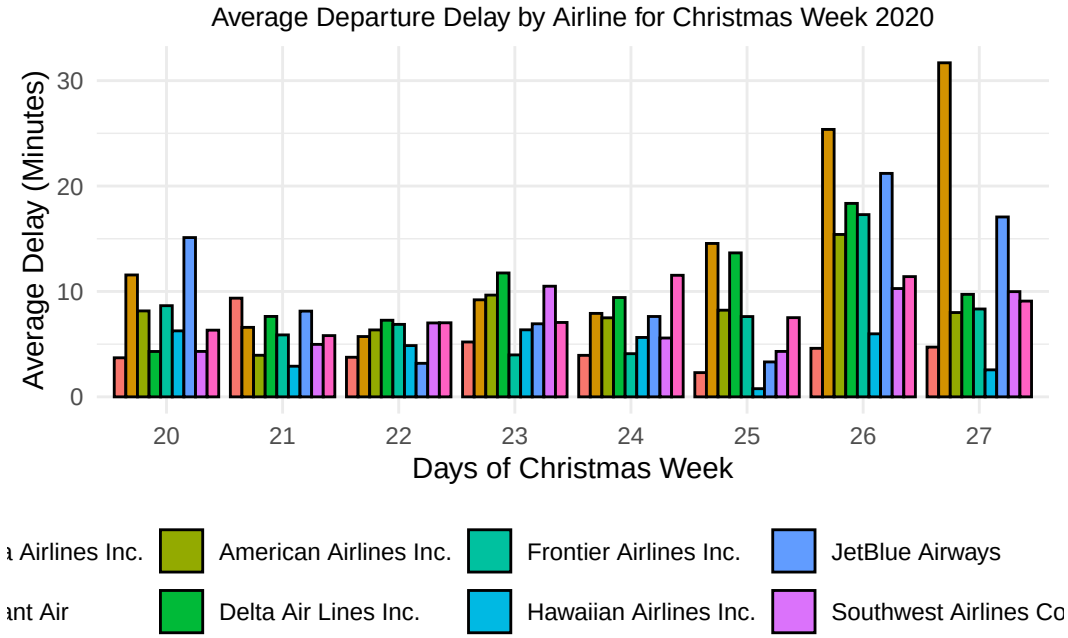
Now that we have the specific airline names, we need to calculate the average airline delay based on the top 10 airlines. We need to group_by() the airline name and days of the week because we want to differentiate what departure delays the various airlines are experiencing the week of Christmas.

The average delay can be found by applying the mean() to departure delay. In order to ensure that our graph is clear to read, we decided to only plot a few airlines. In this case, we choose the top ten airlines according to the based on the https://airadvisor.com/en/top-us-airlines-rating website.

```
# Calculate the average delay time by airline
avgDelay_airlineDay2020 <- airlineInfo2020 %>%
  group_by(Description, DAY_OF_MONTH) %>%
  summarize(AverageDelay = mean(DEP_DELAY_NEW, na.rm = TRUE), .groups = "drop")

specific_airlines <- c("Delta Air Lines Inc.", "Alaska Airlines Inc.", "Hawaiian Airlines In
                       "United Air Lines Inc.", "American Airlines Inc.", "JetBlue Airways",
                       "Southwest Airlines Co.", "Spirit Airlines", "Allegiant Air",
                       "Frontier Airlines Inc.")
```

Then, we created a bar graph using ggplot that filtered the graph based on the airline name and graphed 10 various colored bars for each day represents a different airline.



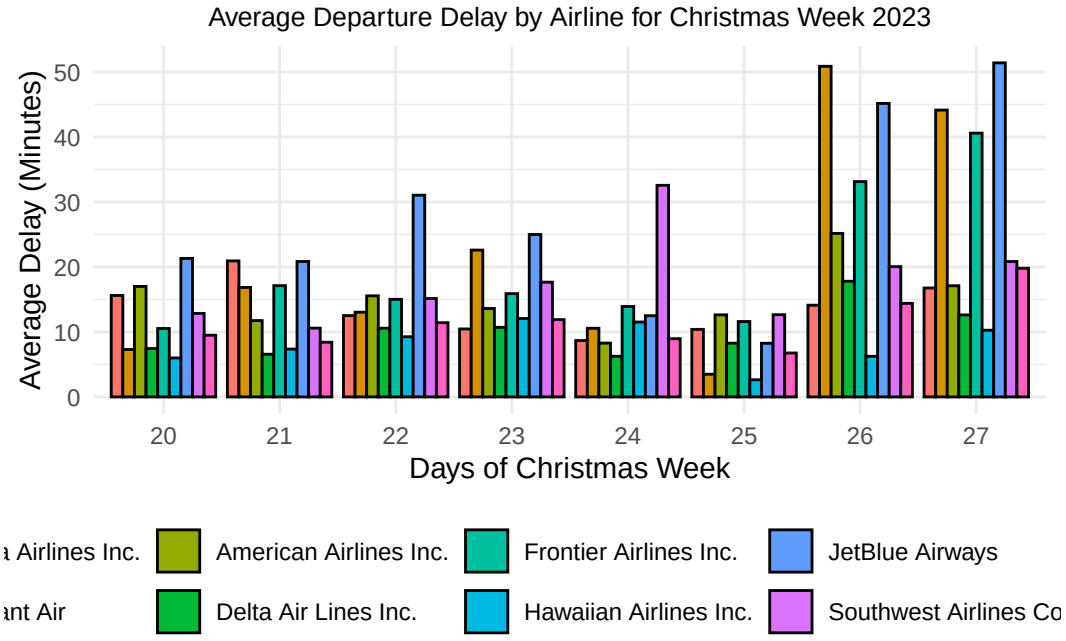
Analysis: Looking at our x and y axis, we can see that Alaska Airlines Inc. had consistently low average departure delays that were less that 5 minutes. the Hawaiian Airlines Inc were also on the low end when it came to average departure delays since they were less than 10 minutes. The the other hand, Allegiant Air had the most delays by far, leading to an average more than 30 minutes on December 27. JetBlue Airways flight delays were also high compared to other flights, especially since they had an average departure delay of more than 20 minutes

on December 27. Another aspect to consider is that the the average delays across airlines were the lowest on the 22. Also, the average delay rates spiked Christmas day and the days after.

Research Question 2: What is the average departure delay time for the week of Christmas 2023 for the top 10 most popular U.S. airlines?

To answer this research question, we follow a similar format as the previous question. First, we need to merge the 2023 Christmas week using the left_join() to get the airline name that corresponded to the existing columns airline code.

The next step is to use the group_by() and summarize() to find the average delay each day based on the airlines in 2023.



Research Question 3: How does the average delay time for the week of Christmas for both years compare to each other?

Methodology

Results

Conclusions

References

Code Appendix

```
#Load December Datasets and Additional Packages
library(ggplot2)
library(dplyr)
library(tidyr)
library(knitr)
library(readr)

original2020 <- "C:/Users/felic/Downloads/DL_SelectFields (4)/DecDataset2020.csv"
flights2020 <- read.csv(original2020)

original2023 <- "C:/Users/felic/Downloads/DL_SelectFields (3)/DecDataset2023.csv"
flights2023 <- read.csv(original2023)

colnames(flights2023)
#Filter December 2020:
christmas_week2020 <- flights2020 %>%
  filter(DAY_OF_MONTH >= 20 & DAY_OF_MONTH <= 27)

#Filter December 2023:
christmas_week2023 <- flights2023 %>%
  filter(DAY_OF_MONTH >= 20 & DAY_OF_MONTH <= 27)

# Load required libraries
library(kableExtra)
```

```r
# Flight Summary of Christmas week in 2020
summary2020 <- summary(christmas_week2020)
summary2020_transposed <- t(summary2020)

# PDF output Table
kable(summary2020_transposed, caption = "Flight Summary Table of Christmas Week 2020") %>%
  #Style the format such as table heading font and position
  kable_styling(
    latex_options = c("striped", "scale_down"),
    font_size = 8,
    position = "center"
  ) %>%
  #Looks at the columns and rows and adjust the lengths accordingly
  column_spec(1, bold = TRUE, width = "4cm") %>%
  column_spec(2:ncol(summary2020_transposed), width = "1.5cm") %>%
  row_spec(0, bold = TRUE, background = "#D3D3D3")
# Flight Summary Table for Christmas week in 2023
summary2023 <- summary(christmas_week2023)
summary2023_transposed <- t(summary2023)

# PDF output Table
kable(summary2023_transposed, caption = "Flight Summary Table of Christmas Week 2020") %>%
  kable_styling(
    latex_options = c("striped", "scale_down"),  # Use LaTeX-specific options
    font_size = 8,                               # Adjust font size
    position = "center"                          # Align to center by default
  ) %>%
  column_spec(1, bold = TRUE, width = "4cm") %>%  # Adjust the first column width
  column_spec(2:ncol(summary2023_transposed), width = "1.5cm") %>% # Adjust other columns
  row_spec(0, bold = TRUE, background = "#D3D3D3") # Style header row
# Data Source 2
code <- "C:/Users/felic/Downloads/L_UNIQUE_CARRIERS.csv"
airlineCode <- read.csv(code)
# Merge the datasets based on the airline code
airlineInfo2020 <- christmas_week2020 %>%
  left_join(airlineCode, by = c("OP_UNIQUE_CARRIER" = "Code"))
# Calculate the average delay time by airline
avgDelay_airlineDay2020 <- airlineInfo2020 %>%
  group_by(Description, DAY_OF_MONTH) %>%
  summarize(AverageDelay = mean(DEP_DELAY_NEW, na.rm = TRUE), .groups = "drop")

specific_airlines <- c("Delta Air Lines Inc.", "Alaska Airlines Inc.", "Hawaiian Airlines Inc
```

```
                        "United Air Lines Inc.", "American Airlines Inc.", "JetBlue Airways",
                        "Southwest Airlines Co.", "Spirit Airlines", "Allegiant Air",
                        "Frontier Airlines Inc.")
#Plotting Average
ggplot(avgDelay_airlineDay2020 %>% filter(Description %in% specific_airlines),
       aes(x = factor(DAY_OF_MONTH), y = AverageDelay, fill = Description)) +
  geom_bar(stat = "identity", position = "dodge", color = "black") +
  labs(title = "Average Departure Delay by Airline for Christmas Week 2020",
       x = "Days of Christmas Week",
       y = "Average Delay (Minutes)",
       fill = "Airline") +
  theme_minimal() +
  theme(
     plot.title = element_text(hjust = 0.5, size = 10),
    legend.position = "bottom")
# Merge the datasets based on the airline code
airlineInfo2023 <- christmas_week2023 %>%
  left_join(airlineCode, by = c("OP_UNIQUE_CARRIER" = "Code"))
# Calculate the average delay time by airline
avgDelay_airlineDay2023 <- airlineInfo2023 %>%
  group_by(Description, DAY_OF_MONTH) %>%
  summarize(AverageDelay2023 = mean(DEP_DELAY_NEW, na.rm = TRUE), .groups = "drop")

specific_airlines2023 <- c("Delta Air Lines Inc.", "Alaska Airlines Inc.", "Hawaiian Airlines
                        "United Air Lines Inc.", "American Airlines Inc.", "JetBlue Airways",
                        "Southwest Airlines Co.", "Spirit Airlines", "Allegiant Air",
                        "Frontier Airlines Inc.")
#Plotting Average
ggplot(avgDelay_airlineDay2023 %>% filter(Description %in% specific_airlines2023),
       aes(x = factor(DAY_OF_MONTH), y = AverageDelay2023, fill = Description)) +
  geom_bar(stat = "identity", position = "dodge", color = "black") +
  labs(title = "Average Departure Delay by Airline for Christmas Week 2023",
       x = "Days of Christmas Week",
       y = "Average Delay (Minutes)",
       fill = "Airline") +
  theme_minimal() +
  theme(
     plot.title = element_text(hjust = 0.5, size = 10),
    legend.position = "bottom")
```