Airline Delay Analysis Using R

1. Abstract:

Delays in Airline Networks has become a common problem for passengers while traveling around the world[1]. Airline delay leads to Airlines reputation affecting passenger's money, time, and patience[2]. Factors like weather conditions, flight booking issues, air trafficking, or any unforeseen events contribute to airline delays[2][1][3]. In this project, we mainly aim at doing Statistical Analysis, and Data Visualization on various factors that are responsible for flight delays in the United States of America between the years 2009 and 2019 to analyze flight performance[4].

2. Introduction:

Studies have estimated the cost of delays to the U.S. economy in 2007 ranging from $32.9 billion to $41 billion[2]. Flight delays have been one of the important problems in airport management and flight scheduling, blurring the efficiency of air system operations and the choice of passengers[1]. Although some airports and airlines have put effort into airline management to reduce the possible delays, flight delays become unavoidable in some airports[1].

We look at various reasons for flight delays in the project like carrier delay, weather delay, security delay, NAS delay, and late aircraft delay with a variety of assumptions which will be concluded from the plots during statistical analysis and data visualization[1][2][3]. These plots could be used to minimize delays[5].

3. Data Availability and Description:

The Air delay analysis data set is publicly available at https://www.kaggle.com/sherrytp/airline-delay-analysis The dataset consists of 29 features as X1, Fl_Date, Op_Carrier, Op_Carrier_Fl_Num, Origin, Dest, Crs_Dep_Time, Dep_Time, Dep_Delay, Taxi_Out, Wheels_Off, Wheels_On, Taxi_In, Crs_Arr_Time, Arr_Time, Arr_Delay, Cancelled, Cancellation_Code, Diverted, Crs_Elapsed_Time, Actual_Elapsed_Time, Air_Time, Distance, Carrier_Delay, Weather_Delay, Nas_Delay, Security_Delay, and Late_Aircraft_Delay.

4. Basic Visualizations:

Histogram, Bar Chart, Line Chart, Box plot, Scatter plot, Grid Plots, Maps, Pie Charts.

5. Packages used in Data Visualization:

tidyverse, dplyr, ggplot2, lubridate, corrplot, usmap, ggpubr and, plotrix.

6. Results:

6.1 Objective: To find out what are the factors that cause the flight delay and to visualize delays per carrier based on the delay reason[1][6][7].

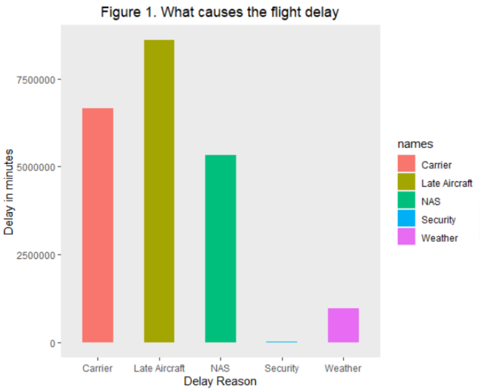


Figure 1. What causes the flight delay

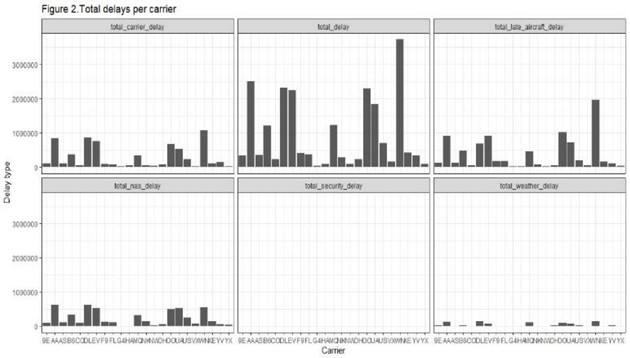


Figure 2.Total delays per carrier

Findings: Flight delays happen due to various reasons such as bad weather, security delays, late aircraft delays, carrier delays, etc. It can be seen from Figure 1, we notice that late aircraft delay contributes the most, and security delay conditions contribute the least to the flight delays. From Figure 2, we can notice that most of the carriers have the late aircraft delay as the most followed by the NAS delay followed by the weather delay followed by the security delay.

6.2 Objective: To find out which are the best and worst days to travel in a month, to find out which month in a year contributes the most /least to the flight delays, which year contributed the most /least to the flight delays and, to analyse which delays are influencing the most per each carrier between the year 2009 and 2018[1][6][7].

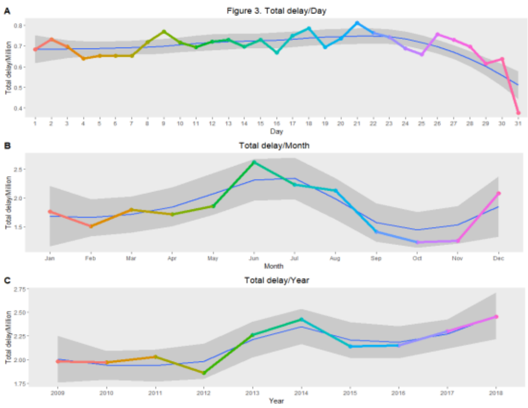


Figure 3. Total delay/Day

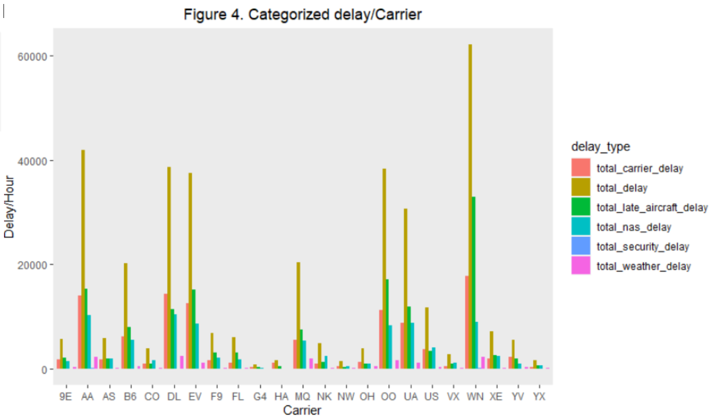
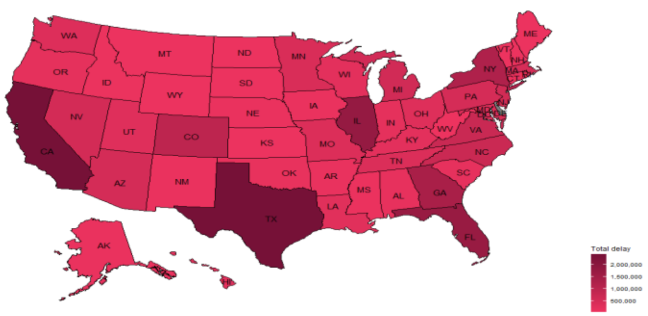


Figure 4. Categorized delay/Carrier

Findings: It can be seen that from the Figure 3. line graph “A” the 18th and 21st day of all the months in all the years between 2009 to 2018 has the highest amount of delay which proves that they are the worst days to travel whereas the 4th and 29th days are the best days to travel as on those days flight delays are experienced less. The day 31st has the least amount of delay (needs to make a note that not all months have 31 days in a year). It can be seen from Figure 3. Line Graph “B” that October has the least delays followed by November whereas June has the highest delays followed by July in all years together. It can be seen from the Figure 3.line graph “C” that the 2012 year had the least delay(less than 1 million minutes) whereas the year 2018 had the highest flight delays (approx. 2.5 million minutes), the year 2014 had the second-highest flight delay(approx. 2.4 million minutes). From Figure 4, we can notice that most of the carriers facing the late aircraft delay more often followed by carrier delay followed by weather delay followed by the NAS delay followed by security delay.

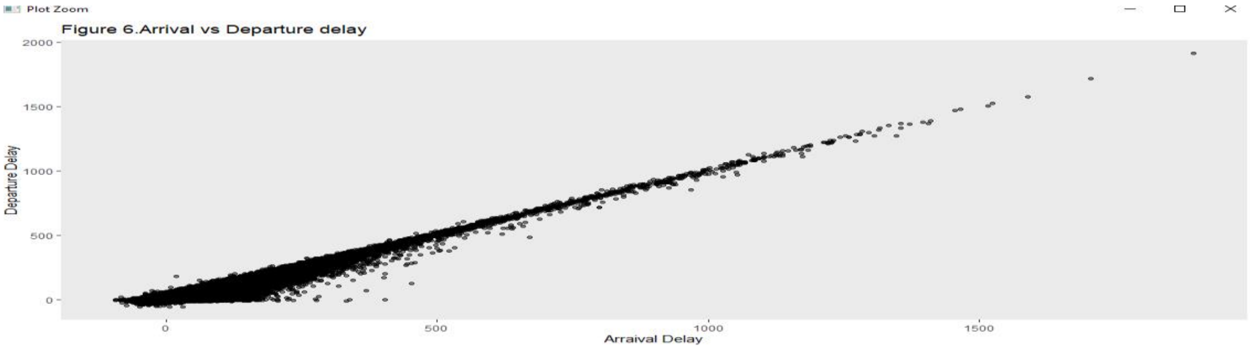
6.3 Objective: Which state in the USA has the most flight delays on a map[8][9][7].

Figure 5. Total delay in each State



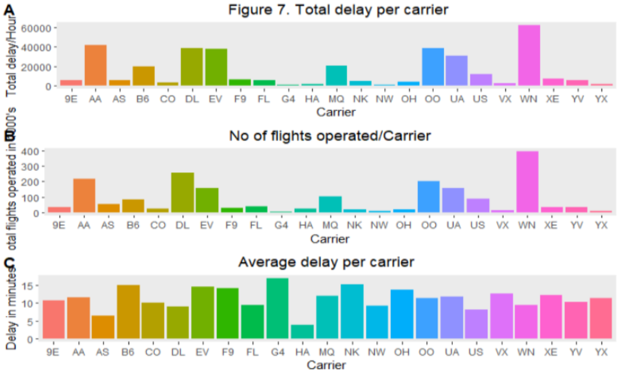
Findings: It can be seen from Figure 5 US Map that Texas (TX) and California State (CA) has the most flight delays followed by Illinois State (IL). The darker the red colour the greater the flight delay. The darker the blue colour the lesser the flight delay.

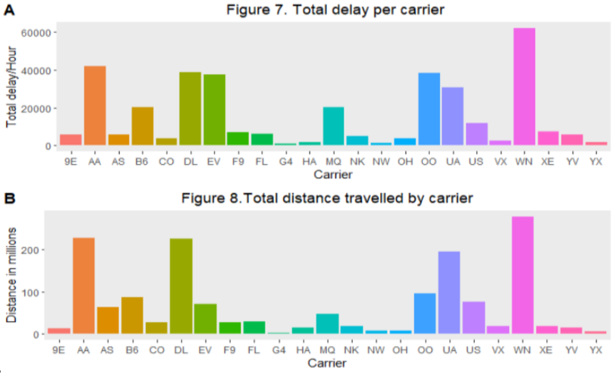
6.4 Objective: A plot between the Arrival Delay and Departure Delay[8][6][10].



Findings: From Figure 6, Scatter plot it can be seen that there exists a Straight linear line between Arrival and Departure delays which means if the flight arrives early at the Airport then it will take off early whereas if an airline arrives late to the airport then it will take off late. (i.e. departure delay).

6.5 Objective: A plot to analyse to get the insights from total delay, total flights operated, and the average flight delay per carrier. Also to visualize does distance influence the total delay per carrier[2][6][10]?





Findings: From Figure 7 and Figure 8., we can notice that the carrier WN has the highest flight delay, but it has also operated the greatest number of flights close to 400,000 between the years 2009 and 2018. Hence, the average flight delay for the carrier WN is not as much as it would have thought to be based on the total delay. With the above comparison, we can conclude that the frequency of flight delay for carrier WN and DL is quite less. It is the other way around for the carriers 9E, AS, B6, CP. G4, F9, etc. which have operated the least but have the more average delays. We can also conclude that the flights that have operated the most have the highest amount of delays.

From Figure 8, we can notice that WN travelled close to 300 million km which is the most followed by AA close to 228 million km followed by the carrier DL close to 226 million km, etc. The carriers that travelled the most have greater delays but carriers EV, MQ, and OO have travelled less but have a greater delay. The carrier US is the only carrier that has travelled more but has less delay.

6.6 Objective: To visualize the cancellation reason in a 3D Pie Chart and to find out which airport in the USA has the highest/lowest cancellations with a cancellation reason. visualize flight cancellations per each month in the year 2017. To find out which airport in the USA has the highest/lowest cancellations with a cancellation reason[5][6][10][3].

Figure 9.Reason of cancelled flights

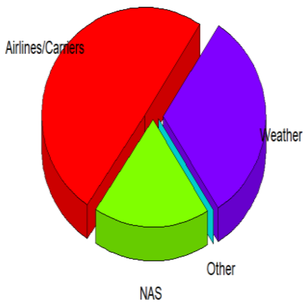


Figure 10. No of Cancelled flights by month in 2017

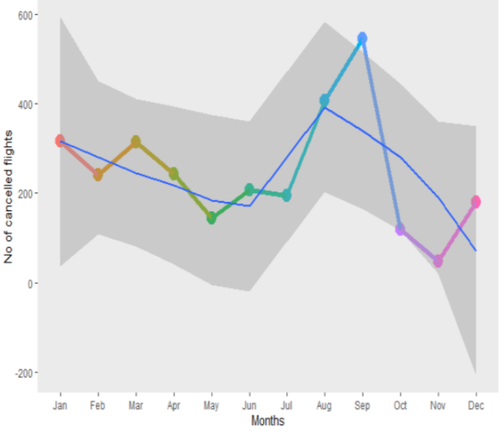
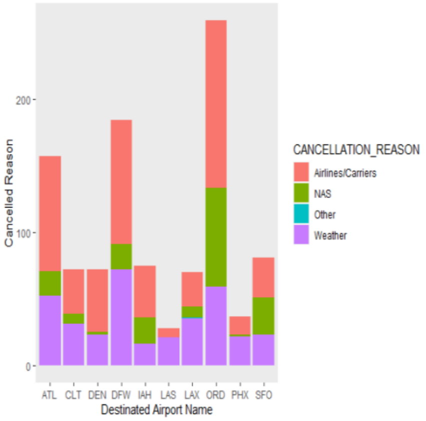


Figure 11. Condition led to cancellations in airport



Findings: We can note from Figure 9. 3D Pie chart that approximately 50% of the flights were cancelled with a reason Airlines/Carriers, followed by Weather, followed by NAS, followed by the cancellation reason other which contributes to the least amount of flight cancellations. We can note in Figure 10 that November experienced fewer flight cancellations followed by Oct followed by May etc in the year 2017. We can note that September experienced higher flight cancellations followed by July followed by January etc in the same year.

It can be seen from Figure 11. that Chicago O' Hare (Airport code -ORD) had the highest number of cancellations among all the Airports in the USA, whereas McCarran International Airport (Airport Code –LAS) had the least Flight Cancellations. Furthermore, the four different cancellation reasons are:1) Airlines/Carriers cancellations, 2) NAS cancellations and 3) Other Cancellations 4) Weather Cancellations, Respectively.

6.7 Objective: To visualize which are the best and worst days to travel based on the data plotted from the years 2009 and 2018 and to visualize which are the best and worst days to travel based on the data plotted from the years 2009 and 2018 for the state of Texas[2][6][10][7].

Figure 12.Categorized delay/Day

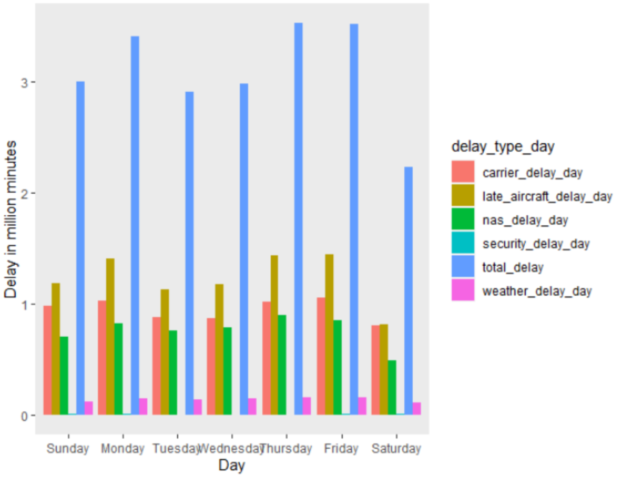
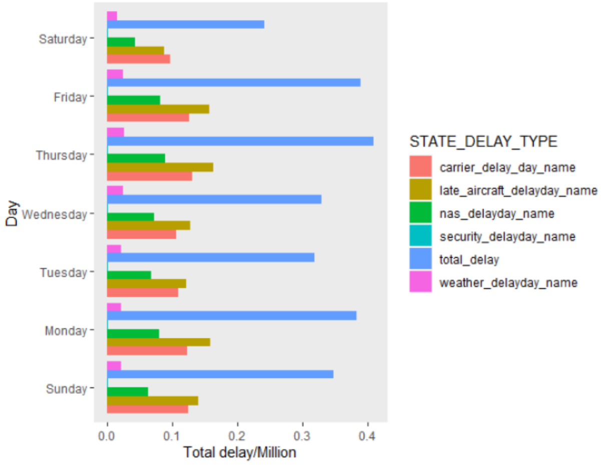


Figure 13. Day of the week delay for Texas



Findings: From Figure 12, we can notice that Saturday is the best day to travel as the total delay for Saturday is close to 2.5 million minutes while Thursday and Friday are the worst days to travel as the total delay for these days is close to 4 million minutes. From Figure 13, we can notice still Saturday is the best day to travel in Texas which has a delay close to 0.22 million minutes while Thursday and Friday are the worst days to travel as the total delay for these days is close to 0.4 million minutes.

7. R Code:

Report_1_Code.pdf

8. Conclusions and discussion:

Airline Delay is an unpleasant condition that has caused the huge cost to the Airline Networks, Air Passengers in many Ways[1]. Airline Delays is a crucial subject because of its economic impact[2]. Airline Delays might cause a hike in Flight Ticket charges. Furthermore, also causes an increase in the Operational /Maintenance costs for the Airline Companies. To date, a huge effort has been made to minimize the Airline Delays in Airports[1]. In our project, some Data Analysis and Data Visualization techniques have been used to find Various factors responsible for Airline Delay Analysis at Airports[5][2][1][8]. Also in which years the Flight delay was maximum and minimum was found out[5]. Furthermore, data visualization in Which Airports Airline Cancellations have been maximum, minimum was found out. Taking all these factors into considerations some new solutions must be found to minimize the Airline delays in the Future. Minimizing the Airline delays is not only going the reduce Air Ticket costs and other costs to the Air Passengers but also it improves the economy of the Country[1]. Moreover, minimizing the Airline Delays brings huge Reputation and profits to the Airline Network Companies.

9.References

[1] W. Wu, C. L. Wu, T. Feng, H. Zhang, and S. Qiu, ‘Comparative Analysis on Propagation Effects of Flight Delays: A Case Study of China Airlines’, *J. Adv. Transp.*, vol. 2018, 2018, doi: 10.1155/2018/5236798.

[2] J. Ferguson, A. Q. Kara, K. Hoffman, and L. Sherry, ‘Estimating domestic US airline cost of delay based on European model’, *Transp. Res. Part C Emerg. Technol.*, vol. 33, pp. 311–323, 2013, doi: 10.1016/j.trc.2011.10.003.

[3] L. Ionescu, C. Gwiggner, and N. Kliewer, ‘Data Analysis of Delays in Airline Networks’, *Bus. Inf. Syst. Eng.*, vol. 58, no. 2, pp. 119–133, 2016, doi: 10.1007/s12599-015-0391-3.

[4] L. Sherry, G. Calderon-Meza, and A. Samant, ‘Trends in airline passenger trip delays (2007-2009)’, *2010 Integr. Commun. Navig. Surveill. Conf. Proceedings, ICNS 2010*, no. February, 2010, doi: 10.1109/ICNSURV.2010.5503234.

[5] E. R. Mueller and G. B. Chatterji, ‘Analysis of aircraft arrival and departure delay characteristics’, *AIAA’s Aircr. Technol. Integr. Oper. 2002 Tech. Forum*, no. October, pp. 1–14, 2002, doi: 10.2514/6.2002-5866.

[6] T. Create, E. Data, V. Using, and G. Description, ‘Package “ggplot2”’, 2020.

[7] T. Package, ‘Package “ dplyr ”’, 2020.

[8] C. Y. Hsiao and M. Hansen, ‘Econometric analysis of U.S. airline flight delays with time-of-day effects’, *Transp. Res. Rec.*, no. 1951, pp. 104–112, 2006, doi: 10.3141/1951-13.

[9] A. Paolo, D. Lorenzo, M. Paolo, and D. Lorenzo, ‘Package “ usmap ”’, 2020.

[10] T. E. Install, ‘Package “ tidyverse ”’, pp. 1–5, 2019.

```
# Air delay analysis code ( Report_1 Code )

library(tidyverse)
library(lubridate)
library(dplyr)
library(ggplot2)
library(corrplot)
library(usmap)
library(ggpubr)
library(plotrix)
x <-
  c("09", "10", "11", "12", "13", "14", "15", "16", "17", "18", "19")
for (i in 1:length(x)) {
  current_year <- x[i]
  path <-
    paste(
      "C:\\StFX\\Bigdata\\Project1\\airline delay analysis\\20",
      current_year,
      ".csv",
      sep = ""
    )
  df <- read_csv(path)
  set.seed(1)
  sdf = df[sample(nrow(df), 200000),]
  assign(paste("sample", current_year, "data", sep = "") , sdf)
}
air_delay_df <-
  rbind(
    sample09data,
    sample10data,
    sample11data,
    sample12data,
    sample13data,
    sample14data,
    sample15data,
    sample16data,
    sample17data,
    sample18data
  )
write.csv(
  air_delay_df ,
  " C:\\StFX\\Bigdata\\Project1\\airline delay analysis\\final_master_df.csv"
```

```
)
air_delay_df <-
  read_csv(" C:\\StFX\\Bigdata\\Project1\\airline delay analysis\\final_master_df.csv")
airline_codes_df <-
  read_csv(" C:\\StFX\\Bigdata\\Project1\\airline delay analysis\\Airline_Codes_Csv.csv")
air_delay_df <-
  merge(air_delay_df,
        airline_codes_df,
        by.x = "ORIGIN",
        by.y = "CODE")
cancellation_codes_df <-
  data.frame(
    "CANCELLATION_CODE" = c("A", "B", "C", "D", NA),
    "CANCELLATION_REASON" = c("Weather", "Airlines/Carriers", "NAS", "Other", NA)
  )
air_delay_df <-
  merge(air_delay_df,
        cancellation_codes_df,
        by.x = "CANCELLATION_CODE",
        by.y = "CANCELLATION_CODE")
##################### SPLITTING DATE START ############
dates <- air_delay_df$FL_DATE
air_delay_df$FLIGHT_DAY <- day(dates)
air_delay_df$FLIGHT_MONTH <- month(dates)
air_delay_df$FLIGHT_YEAR <- year(dates)
air_delay_df$DAY_NAME <- weekdays(as.Date(air_delay_df$FL_DATE))
air_delay_df$MONTH_NAME <- month.abb[air_delay_df$FLIGHT_MONTH]
##################### SPLITTING DATE END ############
#Start of Figure 1 Causes of flight delay
###################### START OF BAR PLOT REASON FOR FILGHT DELAY
#####################
sum_carrier_delay <- sum(air_delay_df$CARRIER_DELAY, na.rm = TRUE)
sum_weather_delay <- sum(air_delay_df$WEATHER_DELAY, na.rm = TRUE)
sum_nas_delay <- sum(air_delay_df$NAS_DELAY, na.rm = TRUE)
sum_security_delay <- sum(air_delay_df$SECURITY_DELAY, na.rm = TRUE)
sum_late_aircraft_delay <-
  sum(air_delay_df$LATE_AIRCRAFT_DELAY, na.rm = TRUE)
############ STORING CATOGORIZED TOTAL DELAY IN A DATA FRAME TO PLOT
THESE REASONS ################
data <- data.frame(
  names = c("Carrier", "Weather", "NAS", "Security", "Late Aircraft") ,
  values = c(
    sum_carrier_delay,
    sum_weather_delay,
    sum_nas_delay,
    sum_security_delay,
    sum_late_aircraft_delay
  )
)
data %>%
  ggplot(aes(x = names, y = values, fill = names)) +
  geom_bar(stat = "identity",
           position = "dodge",
           width = 0.5) +
  labs(title = "Figure 1. What causes the flight delay", x = "Delay Reason", y = "Delay in
minutes") +
  theme(panel.grid = element_blank(), plot.title = element_text(hjust = 0.5))
#End of Figure 1 Causes of flight delay
#start of Figure 2.Total delays per carrier
delay_summed_df <- air_delay_df %>%
  group_by(OP_CARRIER) %>%
  summarize(
    total_carrier_delay = sum(CARRIER_DELAY, na.rm = TRUE),
    total_weather_delay = sum(WEATHER_DELAY, na.rm = TRUE),
    total_nas_delay = sum(NAS_DELAY, na.rm = TRUE),
    total_security_delay = sum(SECURITY_DELAY, na.rm = TRUE),
    total_late_aircraft_delay = sum(LATE_AIRCRAFT_DELAY, na.rm = TRUE),
    total_delay = sum(
      total_carrier_delay,
      total_weather_delay,
      total_nas_delay,
      total_security_delay,
      total_late_aircraft_delay
    )
  )
delay_type_in_rows_df <-
  delay_summed_df %>% gather(key = delay_type,
                             value = Value,
                             total_carrier_delay:total_delay)
plot_six <- delay_type_in_rows_df %>%
  ggplot(aes(x = OP_CARRIER, y = Value)) +
  geom_bar(stat = "identity", position = "dodge") +
  facet_wrap(~ delay_type) +
  labs(title = "Figure 2.Total delays per carrier",
       x = "Carrier",
       y = "Delay type") +
  theme_bw()
plot_six
#End of Figure 2.Total delays per carrier
#Start of Figure 3 Code for Total Delay
#################### START FOR CREATING DATA FRAMES TO PLOT DELAY BY
DAY, MONTH AND YEAR ############
daily_delay_df <- air_delay_df %>%
  group_by(FLIGHT_DAY) %>%
  summarize(
    daily_delay = sum(
      CARRIER_DELAY,
      WEATHER_DELAY,
      NAS_DELAY,
      SECURITY_DELAY,
      LATE_AIRCRAFT_DELAY,
      na.rm = TRUE
    )
  )
daily_delay_df$FLIGHT_DAY <- factor(
  daily_delay_df$FLIGHT_DAY,
  levels = c( "1", "2",  "3", "4", "5", "6",  "7", "8", "9", "10", "11", "12", "13", "14", "15","16",
"17", "18", "19", "20", "21", "22", "23", "24",  "25", "26", "27", "28",  "29", "30", "31"
  )
)
daily_delay_df <- daily_delay_df[order(daily_delay_df$FLIGHT_DAY), ]
monthly_delay_df <- air_delay_df %>%
  group_by(FLIGHT_MONTH, MONTH_NAME) %>%
  summarize(
    monthly_delay = sum(
      CARRIER_DELAY,
      WEATHER_DELAY,
      NAS_DELAY,
      SECURITY_DELAY,
      LATE_AIRCRAFT_DELAY,
      na.rm = TRUE
    )
  )
yearly_delay_df <- air_delay_df %>%
  group_by(FLIGHT_YEAR) %>%
  summarize(
    yearly_delay = sum(
      CARRIER_DELAY,
      WEATHER_DELAY,
      NAS_DELAY,
      SECURITY_DELAY,
      LATE_AIRCRAFT_DELAY,
      na.rm = TRUE
    )
  )
##################### END FOR CREATING DATA FRAMES TO PLOT DELAY BY
DAY, MONTH AND YEAR ############
############ START, SORT BASED ON MONTH NAME AND TO PLOT ALL YEAR
VALUES ################
monthly_delay_df$MONTH_NAME <- factor(
  monthly_delay_df$MONTH_NAME,
  levels = c(
    "Jan","Feb","Mar", "Apr", "May", "Jun", "Jul", "Aug", "Sep", "Oct", "Nov", "Dec"
  )
)
monthly_delay_df <-
  monthly_delay_df[order(monthly_delay_df$MONTH_NAME), ]
yearly_delay_df$FLIGHT_YEAR <- factor(
  yearly_delay_df$FLIGHT_YEAR,
  levels = c(
    "2009",  "2010", "2011", "2012",  "2013", "2014", "2015", "2016", "2017", "2018"
  )
)
yearly_delay_df <-
  yearly_delay_df[order(yearly_delay_df$FLIGHT_YEAR), ]
############ END, SORT BASED ON MONTH NAME AND TO PLOT ALL YEAR
VALUES ################
plot_daily_delay <- daily_delay_df %>%
  ggplot(aes(
    x = FLIGHT_DAY,
    y = daily_delay / 1000000,
    group = 1,
    color = FLIGHT_DAY,
    size = 3
  )) +
  geom_line(size = 1,
            linetype = 1) +
  geom_smooth(size = 1) +
  geom_point(size = 3) +
  labs(title = "Figure 3. Total delay/Day",
       x = "Day",
       y = "Total delay/Million") +
  theme(panel.grid = element_blank(), plot.title = element_text(hjust = 0.5)) + geom_line(size = 2)
plot_daily_delay
plot_monthly_delay <- monthly_delay_df %>%
  ggplot(aes(
    x = MONTH_NAME,
    y = monthly_delay / 1000000,
    group = 1,
    color = MONTH_NAME,
    size = 3
  )) +
  geom_line(size = 1,
            linetype = 1) +
  geom_smooth(size = 1) +
  geom_point(size = 3) +
  labs(title = "Total delay/Month",
       x = "Month",
       y = "Total delay/Million") +
  theme(panel.grid = element_blank(), plot.title = element_text(hjust = 0.5)) + geom_line(size = 2)
plot_monthly_delay
plot_yearly_delay <- yearly_delay_df %>%
  ggplot(aes(
    x = FLIGHT_YEAR,
    y = yearly_delay / 1000000,
    group = 1,
    color = FLIGHT_YEAR,
    size = 3
  )) +
  geom_line(size = 1) +
  geom_smooth(size = 1) +
  geom_point(size = 3) +
  labs(title = "Total delay/Year",
       x = "Year",
       y = "Total delay/Million") + theme(panel.grid = element_blank()) + geom_line(size = 2) +
  theme(panel.grid = element_blank(), plot.title = element_text(hjust = 0.5))
plot_yearly_delay
plot_club_dmy <-
  ggarrange(
    plot_daily_delay,
    plot_monthly_delay,
    plot_yearly_delay,
    labels = c("A", "B", "C"),
    ncol = 1,
    nrow = 3
  ) +
  theme(panel.grid = element_blank())
plot_club_dmy
#End of Figure 3 Code for Total Delay
#Start of Figure 4. Categorized delay/Carrier
plot_categorzied_by_delay <- delay_type_in_rows_df %>%
  ggplot(aes(x = OP_CARRIER, y = Value / 60, fill = delay_type)) +
  geom_bar(stat = "identity", position = "dodge") +
  labs(title = "Figure 4. Categorized delay/Carrier", x = "Carrier", y = "Delay/Hour") +
  theme(panel.grid = element_blank(), plot.title = element_text(hjust = 0.5))
plot_categorzied_by_delay
#End of Figure 4. Categorized delay/Carrier
#Start of Figure 5. Total delay in each State
delay_by_state_df <- air_delay_df %>%
  group_by(STATE) %>%
  summarize(
    total_delay = sum(
      CARRIER_DELAY,
      WEATHER_DELAY,
      NAS_DELAY,
      SECURITY_DELAY,
      LATE_AIRCRAFT_DELAY,
      na.rm = TRUE
    )
  )
new_r <-
  merge(statepop, delay_by_state_df, by.x = "full", by.y = "STATE")
plot_usmap(
  data = new_r,
  values = "total_delay",
  color = "black",
  labels = TRUE
) +
  scale_fill_continuous(
    low = "#ED335F",
    high = "#761137",
    name = "Total delay",
    label = scales::comma
  ) +
  theme(legend.position = "right",
        plot.title = element_text(hjust = 0.5)) +
  labs(title = "Figure 5. Total delay in each State")
#End of Figure 5. Total delay in each State
#Start of Figure 6.Arrival vs Departure delay TAKES TIME TO RUN
air_delay_df %>%
  ggplot(aes(x = ARR_DELAY, y = DEP_DELAY)) +
  labs(title = "Figure 6.Arrival vs Departure delay",
       x = "Arraival Delay",
       y = "Departure Delay") + theme(panel.grid = element_blank()) +
  geom_point(alpha = 0.5)
#End of Figure 6.Arrival vs Departure delay
#Start of Figure 7. Total delay per carrier
##################### START OF TOTAL DELAY PER CARRIER VS NO OF FLIGHTS
OPERATED VS AVERAGE DELAY PER CARRIER #####################
flights_operated_df <- air_delay_df %>%
  count(OP_CARRIER)
plot_overall_delay_carrier <- delay_summed_df %>%
  ggplot(aes(x = OP_CARRIER, y = total_delay / 60, fill = OP_CARRIER)) +
  geom_bar(stat = "identity", position = "dodge") +
  labs(title = " Figure 7. Total delay per carrier", x = "Carrier", y = "Total delay/Hour") +
  theme(panel.grid = element_blank(), plot.title = element_text(hjust = 0.5))
plot_overall_delay_carrier
plot_flights_operated <- flights_operated_df %>%
  ggplot(aes(x = OP_CARRIER, y = n / 1000, fill = OP_CARRIER)) +
  geom_bar(stat = "identity", position = "dodge") +
  labs(title = "No of flights operated/Carrier", x = "Carrier", y = "Total flights operated in
1000's") +
  theme(panel.grid = element_blank(), plot.title = element_text(hjust = 0.5))
plot_flights_operated
plot_avg_delay <- delay_summed_df %>%
  ggplot(aes(
    x = OP_CARRIER,
    y = total_delay / flights_operated_df$n,
    fill = OP_CARRIER
  )) + geom_bar(stat = "identity", position = "dodge") +
  labs(title = "Average delay per carrier", x = "Carrier", y = "Delay in minutes") +
  theme(panel.grid = element_blank(), plot.title = element_text(hjust = 0.5))
plot_avg_delay
plot_club <-
  ggarrange(
    plot_overall_delay_carrier,
    plot_flights_operated,
    plot_avg_delay,
    labels = c("A", "B", "C"),
    ncol = 1,
    nrow = 3
  ) +
  labs(title = "Total delay in hours/No flights operatedr", y = "Total flights operated in 1000's") +
  theme(panel.grid = element_blank())
plot_club
#End of Figure 7. Total delay per carrier
# Start of Figure 8.Total distance travelled by carrier
delay_by_distance_df <- air_delay_df %>%
  group_by(OP_CARRIER) %>%
  summarize(
    total_disance = sum(DISTANCE),
    total_delay = sum(
      CARRIER_DELAY,
      WEATHER_DELAY,
      NAS_DELAY,
      SECURITY_DELAY,
      LATE_AIRCRAFT_DELAY,
      na.rm = TRUE
    )
  )
plot_distance_per_carrier <- delay_by_distance_df %>%
  ggplot(aes(
    x = OP_CARRIER,
    y = total_disance / 1000000,
    fill = OP_CARRIER
  )) + geom_bar(stat = "identity", position = "dodge") +
  labs(title = "Figure 8.Total distance travelled by carrier", x = "Carrier", y =
         "Distance in millions") +
  theme(panel.grid = element_blank(), plot.title = element_text(hjust = 0.5))
plot_distance_per_carrier
plot_club_2 <-
  ggarrange(
    plot_overall_delay_carrier,
    plot_distance_per_carrier,
    labels = c("A", "B"),
    ncol = 1,
    nrow = 2
  ) +
  theme(panel.grid = element_blank())
plot_club_2
#End of Figure 8.Total distance travelled by carrier
#Start of Figure 9.Reason of cancelled flights
```

```
Can_Reason <- air_delay_df %>%
  group_by(CANCELLATION_REASON) %>%
  summarise(COUNT_OF_CAN_REAS = n())
Can_Reason <- Can_Reason[complete.cases(Can_Reason), ]
pie3D(
  Can_Reason$COUNT_OF_CAN_REAS,
  labels = Can_Reason$CANCELLATION_REASON,
  explode = 0.1,
  theta = 1,
  start = 2,
  main = "Figure 9.Reason of cancelled flights"
)
#End of Figure 9.Reason of cancelled flights
#Start of Figure 10. No of Cancelled flights by month in 2017
options(scipen = 999)
cancellations_2017_df <- air_delay_df %>%
  filter(FLIGHT_YEAR == "2017") %>%
  group_by(MONTH_NAME) %>%
  summarize(NO_FLIGHTS_DELAYED = sum(CANCELLED))
cancellations_2017_df$MONTH_NAME <-
  factor(
    cancellations_2017_df$MONTH_NAME,
    levels = c(
      "Jan", "Feb", "Mar",  "Apr", "May", "Jun", "Jul",  "Aug", "Sep", "Oct", "Nov", "Dec"
    )
  )
cancellations_2017_df <-
  cancellations_2017_df[order(cancellations_2017_df$MONTH_NAME), ]
cancellations_2017_df %>%
  ggplot(aes(
    x = MONTH_NAME,
    y = NO_FLIGHTS_DELAYED,
    group = 1 ,
    color = MONTH_NAME,
    size = 2
  )) +
  labs(title = "Figure 10. No of Cancelled flights by month in 2017",
       x = "Months",
       y = "No of cancelled flights") +
  theme(panel.grid = element_blank()) +
  geom_point(alpha = 1) + theme(panel.grid = element_blank(),
                                plot.title = element_text(hjust = 0.5)) + geom_line(size = 2) +
  geom_smooth(size = 1)
```

```
#End of Figure 10. No of Cancelled flights by month in 2017
#Start of Figure 11 Causes of flight delay Condition led to cancellations in airport
data <- data.frame(count(air_delay_df, vars = DEST))
data <- head(data[order(-data$n),], n = 10)
tom_cancelled_dest <- subset(air_delay_df, DEST == data$vars)
tom_cancelled_dest =
tom_cancelled_dest[!is.na(tom_cancelled_dest$CANCELLATION_REASON), ]
tom_cancelled_dest %>%
  ggplot(aes(x = DEST)) +
  labs(title = "Figure 11. Condition led to cancellations in airport",
       x = "Destinated Airport Name",
       y = "Cancelled Reason") + theme(panel.grid = element_blank()) +
  geom_bar(aes(fill = CANCELLATION_REASON)) #, position = "dodge")
#End of Figure 11. Causes of flight delay Condition led to cancellations in airport
#Start of Figure 12.Categorized delay/Day
delay_by_day_name_df <- air_delay_df %>%
  group_by(DAY_NAME) %>%
  summarize(
    carrier_delay_day = sum(CARRIER_DELAY, na.rm = TRUE),
    weather_delay_day = sum(WEATHER_DELAY, na.rm = TRUE),
    nas_delay_day = sum(NAS_DELAY, na.rm = TRUE),
    security_delay_day = sum(SECURITY_DELAY, na.rm = TRUE),
    late_aircraft_delay_day = sum(LATE_AIRCRAFT_DELAY, na.rm = TRUE),
    total_delay = sum(
      CARRIER_DELAY,
      WEATHER_DELAY,
      NAS_DELAY,
      SECURITY_DELAY,
      LATE_AIRCRAFT_DELAY,
      na.rm = TRUE
    )
  )
############ START, MERGING DELAYS IN A SINGLE COLUMN TO PLOT AND
SETTING THE LEVELS TO PLOT IN A DAY SERIES ############
delay_day_in_rows_df <-
  delay_by_day_name_df %>% gather(key = delay_type_day,
                                  value = Value,
                                  carrier_delay_day:total_delay)
delay_day_in_rows_df$DAY_NAME <-
  factor(
    delay_day_in_rows_df$DAY_NAME,
    levels = c(
      "Sunday", "Monday", "Tuesday", "Wednesday", "Thursday",  "Friday", "Saturday"
    )
  )
delay_day_in_rows_df <-
  delay_day_in_rows_df[order(delay_day_in_rows_df$DAY_NAME), ]
plot_categorzied_day_delay <- delay_day_in_rows_df %>%
  ggplot(aes(
    x = DAY_NAME,
    y = Value / 1000000,
    fill = delay_type_day
  )) +
  geom_bar(stat = "identity", position = "dodge") +
  labs(title = "Figure 12.Categorized delay/Day", x = "Day", y = "Delay in million minutes") +
  theme(panel.grid = element_blank(), plot.title = element_text(hjust = 0.5))
plot_categorzied_day_delay
############ END, MERGING DELAYS IN A SINGLE COLUMN TO PLOT AND
SETTING THE LEVELS TO PLOT IN A DAY SERIES ############
#End of Figure 12.Categorized delay/Day
#Start of Figure 13. Day of the week delay for Texas
############ START OF DELAY PER EACH DAY FOR EACH STATE ############
delay_by_state_per_day_name <- air_delay_df %>%
  group_by(STATE, DAY_NAME) %>%
  summarize(
    carrier_delay_day_name = sum(CARRIER_DELAY, na.rm = TRUE),
    weather_delayday_name = sum(WEATHER_DELAY, na.rm = TRUE),
    nas_delayday_name = sum(NAS_DELAY, na.rm = TRUE),
    security_delayday_name = sum(SECURITY_DELAY, na.rm = TRUE),
    late_aircraft_delayday_name = sum(LATE_AIRCRAFT_DELAY, na.rm = TRUE),
    total_delay = sum(
      CARRIER_DELAY,
      WEATHER_DELAY,
      NAS_DELAY,
      SECURITY_DELAY,
      LATE_AIRCRAFT_DELAY,
      na.rm = TRUE
    )
  )
########## MERGING DELAY TYPES IN A SINGLE COLUMN FOR VISUALIZATION
##################
delay_by_state_per_day_name_row <-
  delay_by_state_per_day_name %>% gather(key = STATE_DELAY_TYPE,
                                         value = Value,
                                         carrier_delay_day_name:total_delay)
```

```
############ SORTING DAYS BASED ON THE DAY OF THE WEEK
#####################
delay_by_state_per_day_name_row$DAY_NAME <-
  factor(
    delay_by_state_per_day_name_row$DAY_NAME,
    levels = c(
      "Sunday", "Monday", "Tuesday", "Wednesday", "Thursday",  "Friday", "Saturday"
    )
  )
delay_by_state_per_day_name_row <-
  delay_by_state_per_day_name_row[order(delay_by_state_per_day_name_row$DAY_NAME),
]
delay_by_day_name_texas <-
  delay_by_state_per_day_name_row %>% filter(STATE == "Texas")
############# VISUALIZING WEEKDAY DELAYS FOR EACH STATE
##########################
delay_by_day_name_texas %>%
  ggplot(aes(
    x = DAY_NAME,
    y = Value / 1000000,
    fill = STATE_DELAY_TYPE
  )) +
  geom_bar(stat = "identity", position = "dodge") +
  labs(title = "Figure 13. Day of the week delay for Texas", x = "Day", y = "Total delay/Million")
+
  theme(panel.grid = element_blank(), plot.title = element_text(hjust = 0.5)) +
  coord_flip()
#End of Figure 13. Day of the week delay for Texas
#Start of Figure 14. Flights operated/State
state_flights_count_df <-
  air_delay_df %>% group_by(STATE) %>% count(STATE)
state_flights_count_df %>%
  ggplot(aes(x = STATE, y = n, fill = STATE)) +
  geom_bar(stat = "identity", position = "dodge") +
  labs(title = "Figure 14. Flights operated/State", x = "State", y = "No of flights operated") +
  theme(panel.grid = element_blank(), plot.title = element_text(hjust = 0.5)) +
  coord_flip()
#End of Figure 14. Flights operated/State
#Start of Figure 15. Monthly delay for Texas
############ START OF DELAY PER EACH MONTH FOR EACH STATE ############
delay_by_state_per_month <- air_delay_df %>%
  group_by(STATE, MONTH_NAME) %>%
  summarize(
    carrier_delay_month = sum(CARRIER_DELAY, na.rm = TRUE),
    weather_delay_month = sum(WEATHER_DELAY, na.rm = TRUE),
    nas_delay_month = sum(NAS_DELAY, na.rm = TRUE),
    security_delay_month = sum(SECURITY_DELAY, na.rm = TRUE),
    late_aircraft_delay_month = sum(LATE_AIRCRAFT_DELAY, na.rm = TRUE),
    total_delay = sum(
      CARRIER_DELAY,
      WEATHER_DELAY,
      NAS_DELAY,
      SECURITY_DELAY,
      LATE_AIRCRAFT_DELAY,
      na.rm = TRUE
    )
  )
########## MERGING DELAY TYPES IN A SINGLE COLUMN FOR VISUALIZATION
#################
delay_by_state_per_month_row <-
  delay_by_state_per_month %>% gather(key = delay_type_state_month,
                                      value = Value,
                                      carrier_delay_month:total_delay)
############ SETTING LEVELS TO PLOT MONTHS IN A MONTH SERIES
#####################
delay_by_state_per_month_row$MONTH_NAME <-
  factor(
    delay_by_state_per_month_row$MONTH_NAME,
    levels = c(
      "Jan", "Feb", "Mar",  "Apr", "May", "Jun", "Jul",  "Aug", "Sep", "Oct", "Nov", "Dec"
    )
  )
delay_by_state_per_month_row <-
  delay_by_state_per_month_row[order(delay_by_state_per_month_row$MONTH_NAME), ]
delay_by_month_texas <-
  delay_by_state_per_month_row %>% filter(STATE == "Texas")
############# VISUALIZING MONTHLY DELAYS FOR TEXAS STATE
##########################
delay_by_month_texas %>%
  ggplot(aes(x = MONTH_NAME, y = Value / 100000, fill = delay_type_state_month)) +
  geom_bar(stat = "identity", position = "dodge") +
  labs(title = "Figure 15. Monthly delay for Texas", x = "Month", y = "Total delay/Million") +
  theme(panel.grid = element_blank(), plot.title = element_text(hjust = 0.5)) +
  coord_flip()
############ END OF DELAY PER EACH MONTH FOR EACH STATE ############
#End of Figure 15. Monthly delay for Texas
```

```
d_2018 <- air_delay_df %>%
  filter(CANCELLED == "1")
hist(d_2018$FLIGHT_DAY)
hist(d_2018$FLIGHT_MONTH)
#Start of Figure 16. pie chart percentage of reason of cancelled flights
ggplot(Can_Reason,
       aes(x = "", y = COUNT_OF_CAN_REAS, fill = CANCELLATION_REASON)) +
  geom_bar(width = 1, stat = "identity") +
  coord_polar("y", start = 2)
#End of Figure 16. pie chart percentage of reason of cancelled flights
```