Flights analysis.USA2013

Najimi Sanae

2025-01-16

```
#Analysis objective :

#The primary objective of this analysis is to uncover patterns and trends in U.S. flight delays
#from 2013.By examining factors such as seasonal variations, time of day, and airport performance,
#the analysis aims to identify the key drivers of delays. The findings highlight significant patterns,
#including peak delays during winter and summer,correlations between departure and arrival delays,
#and airport-specific challenges.These insights can inform strategies to optimize scheduling
#and improve overall efficiency in air travel.
```

```
# Data Loading
flights <- nycflights13::flights
airports <- nycflights13::airports

# Data Exploration
head(flights)
```

```
## # A tibble: 6 x 19
##    year month   day dep_time sched_dep_time dep_delay arr_time sched_arr_time
##   <int> <int> <int>    <int>          <int>     <dbl>    <int>          <int>
## 1  2013     1     1      517            515         2      830            819
## 2  2013     1     1      533            529         4      850            830
## 3  2013     1     1      542            540         2      923            850
## 4  2013     1     1      544            545        -1     1004           1022
## 5  2013     1     1      554            600        -6      812            837
## 6  2013     1     1      554            558        -4      740            728
## # i 11 more variables: arr_delay <dbl>, carrier <chr>, flight <int>,
## #   tailnum <chr>, origin <chr>, dest <chr>, air_time <dbl>, distance <dbl>,
## #   hour <dbl>, minute <dbl>, time_hour <dttm>
```

```
ncol(flights)  # Number of columns
```

```
## [1] 19
```

```
nrow(flights)  # Number of rows
```

```
## [1] 336776
```

```r
colnames(flights)  # Column names
```

```
##  [1] "year"           "month"          "day"            "dep_time"
##  [5] "sched_dep_time" "dep_delay"      "arr_time"       "sched_arr_time"
##  [9] "arr_delay"      "carrier"        "flight"         "tailnum"
## [13] "origin"         "dest"           "air_time"       "distance"
## [17] "hour"           "minute"         "time_hour"
```

```r
sapply(flights, typeof)  # Data types of columns
```

```
##           year          month            day       dep_time sched_dep_time
##      "integer"      "integer"      "integer"      "integer"      "integer"
##      dep_delay       arr_time sched_arr_time      arr_delay        carrier
##       "double"      "integer"      "integer"       "double"    "character"
##         flight        tailnum         origin           dest       air_time
##      "integer"    "character"    "character"    "character"       "double"
##       distance           hour         minute      time_hour
##       "double"       "double"       "double"       "double"
```

```r
colSums(is.na(flights))  # Missing values per column
```

```
##           year          month            day       dep_time sched_dep_time
##              0              0              0           8255              0
##      dep_delay       arr_time sched_arr_time      arr_delay        carrier
##           8255           8713              0           9430              0
##         flight        tailnum         origin           dest       air_time
##              0           2512              0              0           9430
##       distance           hour         minute      time_hour
##              0              0              0              0
```

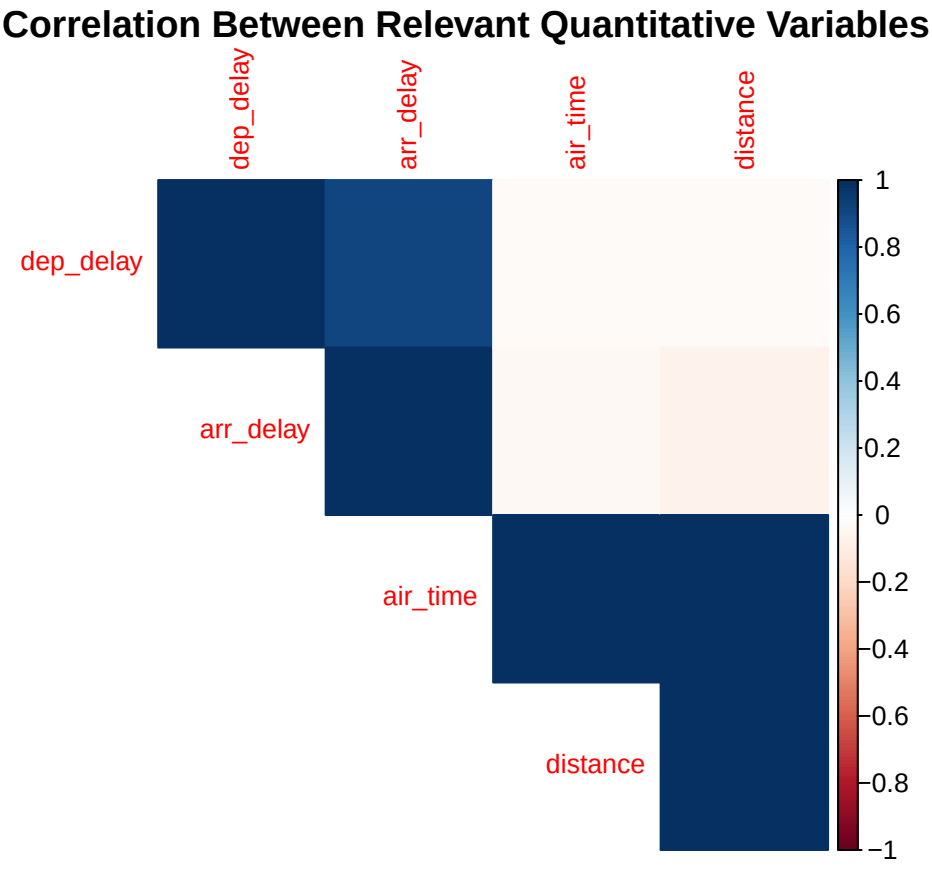
```r
# Data Cleaning
flights_clean <- flights %>%
  filter(!is.na(dep_time), !is.na(arr_time), !is.na(dep_delay),
         !is.na(arr_delay), !is.na(air_time))

# Statistical Summary of Numeric Data
numeric_data <- flights_clean %>% select_if(is.numeric)
summary(numeric_data)
```

```
##       year          month             day           dep_time    sched_dep_time
##  Min.   :2013   Min.   : 1.000   Min.   : 1.00   Min.   :   1   Min.   : 500
##  1st Qu.:2013   1st Qu.: 4.000   1st Qu.: 8.00   1st Qu.: 907   1st Qu.: 905
##  Median :2013   Median : 7.000   Median :16.00   Median :1400   Median :1355
##  Mean   :2013   Mean   : 6.565   Mean   :15.74   Mean   :1349   Mean   :1340
##  3rd Qu.:2013   3rd Qu.:10.000   3rd Qu.:23.00   3rd Qu.:1744   3rd Qu.:1729
##  Max.   :2013   Max.   :12.000   Max.   :31.00   Max.   :2400   Max.   :2359
##    dep_delay          arr_time    sched_arr_time   arr_delay
##  Min.   : -43.00   Min.   :   1   Min.   :   1   Min.   : -86.000
##  1st Qu.:  -5.00   1st Qu.:1104   1st Qu.:1122   1st Qu.: -17.000
##  Median :  -2.00   Median :1535   Median :1554   Median :  -5.000
##  Mean   :  12.56   Mean   :1502   Mean   :1533   Mean   :   6.895
```

```
##  3rd Qu.:  11.00   3rd Qu.:1940   3rd Qu.:1944   3rd Qu.:  14.000
##  Max.   :1301.00   Max.   :2400   Max.   :2359   Max.   :1272.000
##      flight        air_time        distance         hour           minute
##  Min.   :   1   Min.   : 20.0   Min.   :  80   Min.   : 5.00   Min.   : 0.00
##  1st Qu.: 544   1st Qu.: 82.0   1st Qu.: 509   1st Qu.: 9.00   1st Qu.: 8.00
##  Median :1467   Median :129.0   Median : 888   Median :13.00   Median :29.00
##  Mean   :1943   Mean   :150.7   Mean   :1048   Mean   :13.14   Mean   :26.23
##  3rd Qu.:3412   3rd Qu.:192.0   3rd Qu.:1389   3rd Qu.:17.00   3rd Qu.:44.00
##  Max.   :8500   Max.   :695.0   Max.   :4983   Max.   :23.00   Max.   :59.00
```

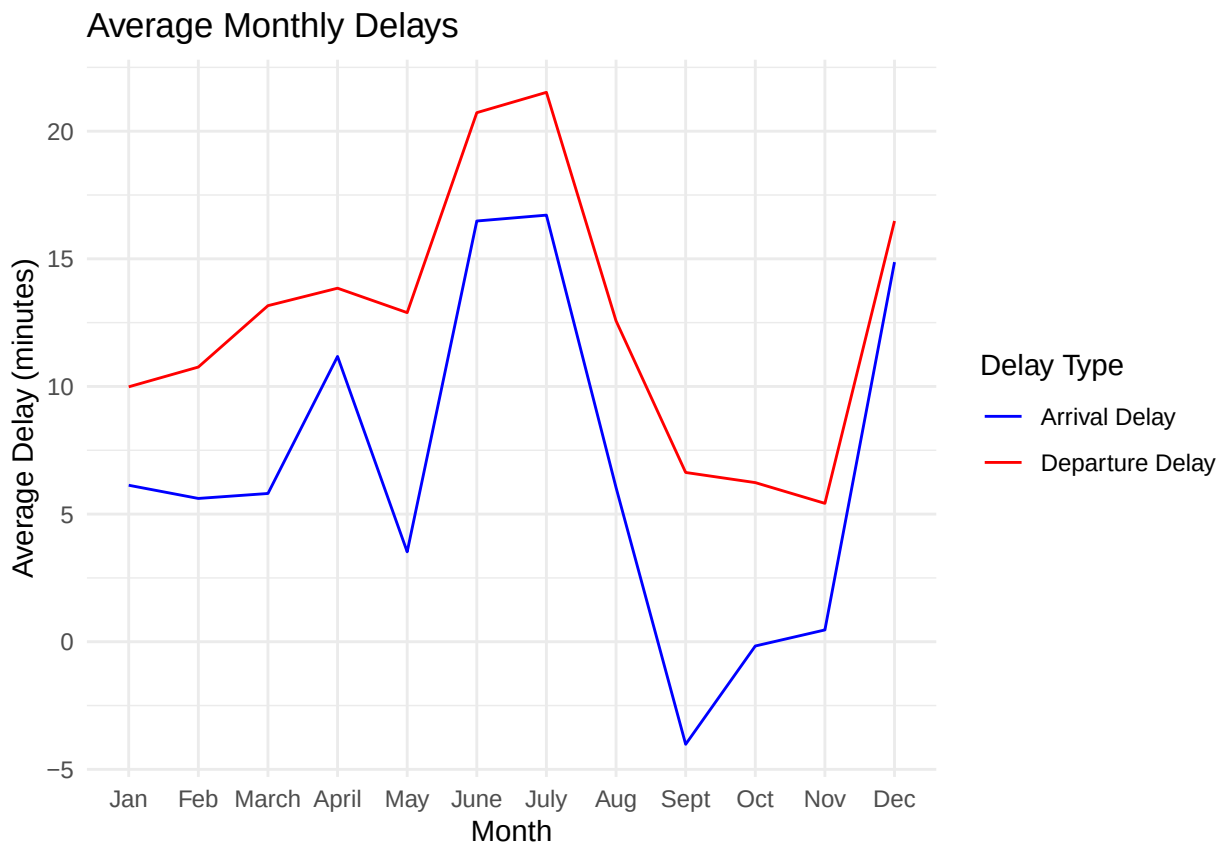
```
#This summary helps identify potential outliers (e.g., extreme delays) and irregular distributions.
```

```
# Correlation Analysis
#Examining relationships between numeric variables, such as departure delays, arrival delays,
#airtime, and distance.
numeric_data_relevant <- flights_clean %>% select(dep_delay, arr_delay, air_time, distance)
cor_matrix <- cor(numeric_data_relevant, use = "complete.obs")

#graphic representation :
corrplot(cor_matrix, method = "color", type = "upper", tl.cex = 0.8,
         title = "Correlation Between Relevant Quantitative Variables",
         mar = c(0, 0, 1, 0))
```



```r
# Strong correlations are expected between `dep_delay` and `arr_delay`, indicating a causal
#relationship where late departures lead to late arrivals.
```

```r
# Average Delays by Month :
# Aggregating and visualizing monthly trends in average departure and arrival delays.
monthly_delays <- flights_clean %>%
  group_by(month) %>%
  summarise(
    mean_dep_delay = mean(dep_delay, na.rm = TRUE),
    mean_arr_delay = mean(arr_delay, na.rm = TRUE),
    .groups = 'drop'
  )

#graphic representation :
ggplot(monthly_delays, aes(x = factor(month))) +
  geom_line(aes(y = mean_dep_delay, color = "Departure Delay"), group = 1) +
  geom_line(aes(y = mean_arr_delay, color = "Arrival Delay"), group = 1) +
  scale_color_manual(values = c("Departure Delay" = "red", "Arrival Delay" = "blue")) +
  labs(
    title = "Average Monthly Delays",
    x = "Month",
    y = "Average Delay (minutes)",
    color = "Delay Type"
  ) +
  scale_x_discrete(labels = c("Jan", "Feb", "March", "April", "May", "June", "July", "Aug", "Sept", "Oc
  theme_minimal()
```



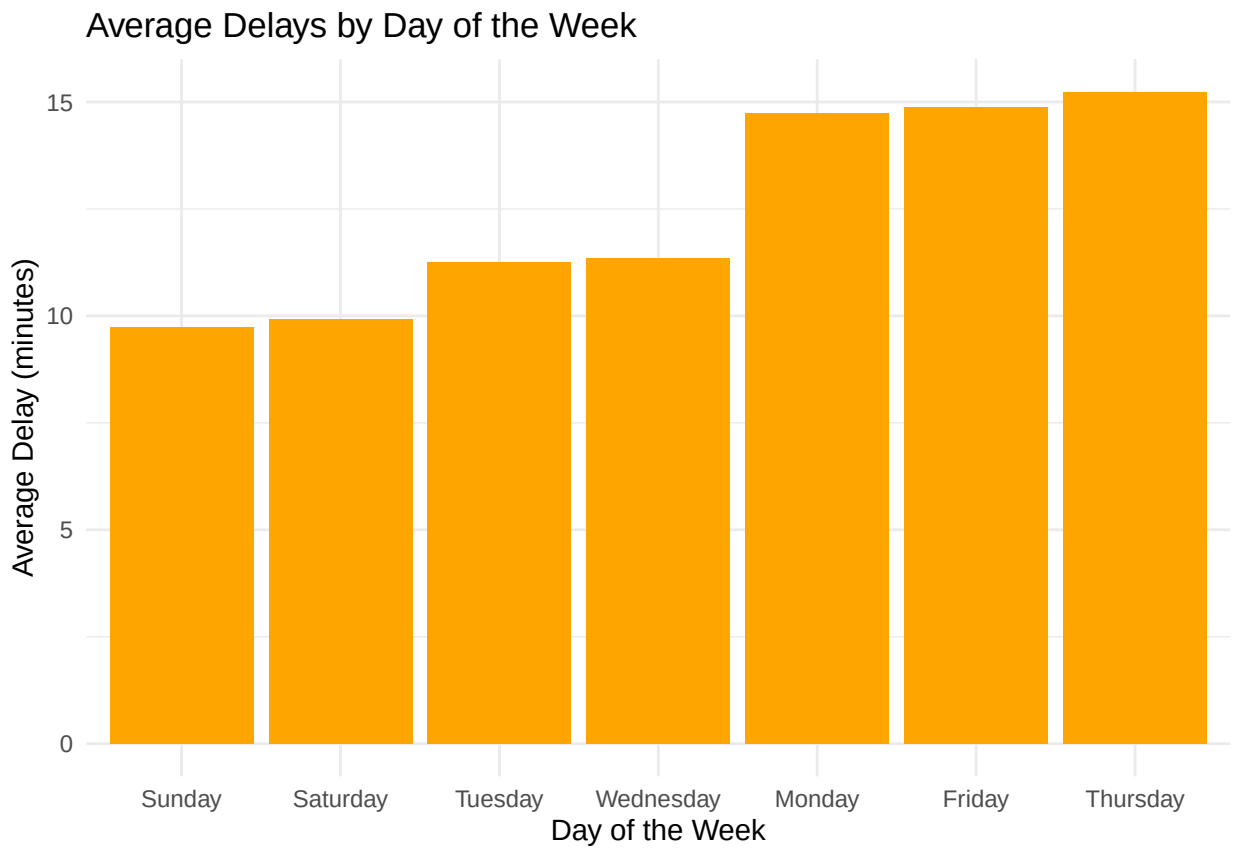

```
# Interpretation:
# Summer months (June and July)  show higher delays due to increased air traffic.
```

```
# Average Delays by Day of the Week :Analyzing how delays vary across weekdays.
Sys.setlocale("LC_TIME", "en_US.UTF-8")
```

```
## [1] "en_US.UTF-8"
```

```
#first let's add a column that contains extracted weekdays from time_hour column
flights_clean <- flights_clean %>%
  mutate(
    date = as.Date(time_hour),
    weekday = weekdays(date)
  )

# Summarizing delays by weekday
weekday_delays <- flights_clean %>%
  group_by(weekday) %>%
  summarise(
    mean_dep_delay = mean(dep_delay, na.rm = TRUE),
    mean_arr_delay = mean(arr_delay, na.rm = TRUE),
    .groups = 'drop'
  )

# graphic representation :
```

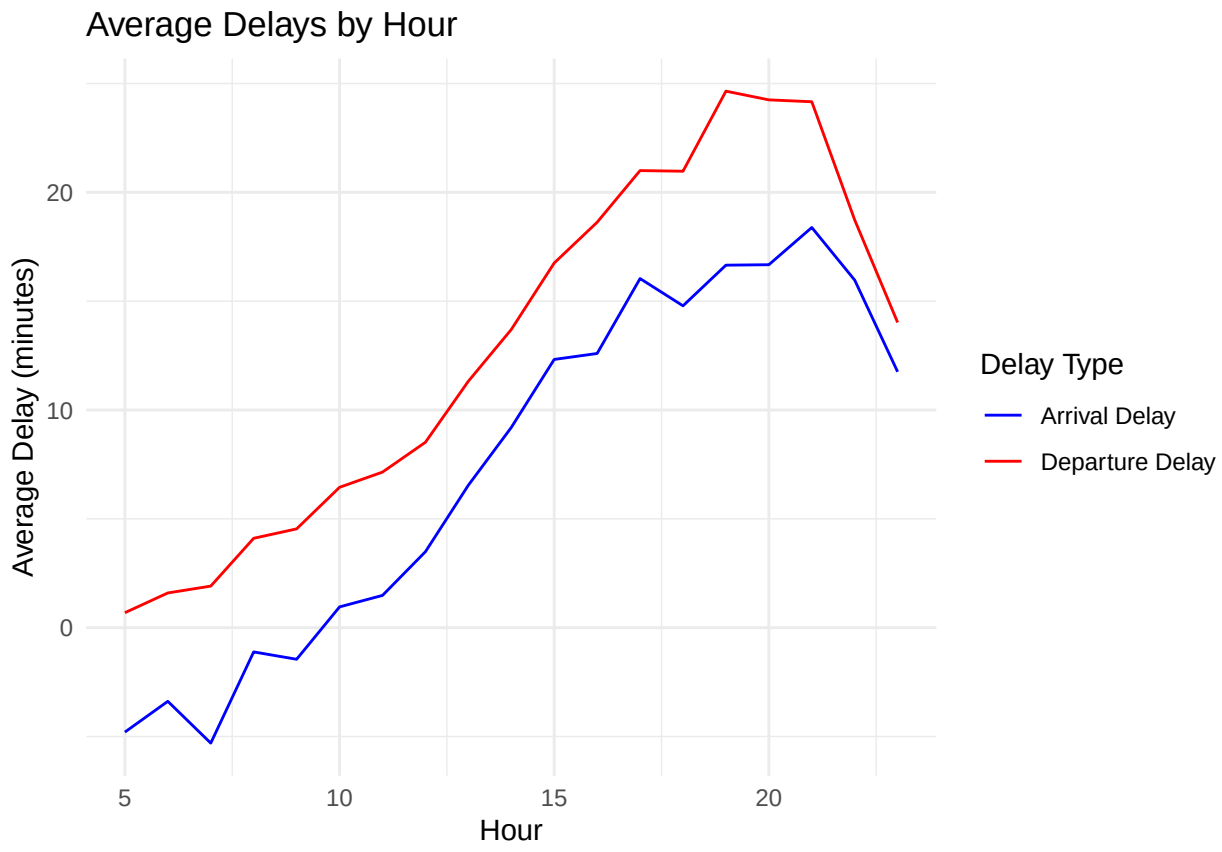
```r
ggplot(weekday_delays, aes(x = reorder(weekday, mean_dep_delay), y = mean_dep_delay)) +
  geom_col(fill = "orange") +
  labs(
    title = "Average Delays by Day of the Week",
    x = "Day of the Week",
    y = "Average Delay (minutes)"
  ) +
  theme_minimal()
```




```r
# Interpretation:
# High travel days like Mondays and Fridays may show increased delays due to heavy air traffic.
```

```r
# Average Delays by Hour :Analyzing how delays vary through the day
time_delays <- flights_clean %>%
  group_by(hour) %>%
  summarise(
    mean_dep_delay = mean(dep_delay, na.rm = TRUE),
    mean_arr_delay = mean(arr_delay, na.rm = TRUE),
    .groups = 'drop'
  )

#graphic representation :
ggplot(time_delays, aes(x = hour)) +
  geom_line(aes(y = mean_dep_delay, color = "Departure Delay")) +
  geom_line(aes(y = mean_arr_delay, color = "Arrival Delay")) +
```

```
  scale_color_manual(values = c("Departure Delay" = "red", "Arrival Delay" = "blue")) +
  labs(
    title = "Average Delays by Hour",
    x = "Hour",
    y = "Average Delay (minutes)",
    color = "Delay Type"
  ) +
  theme_minimal()
```



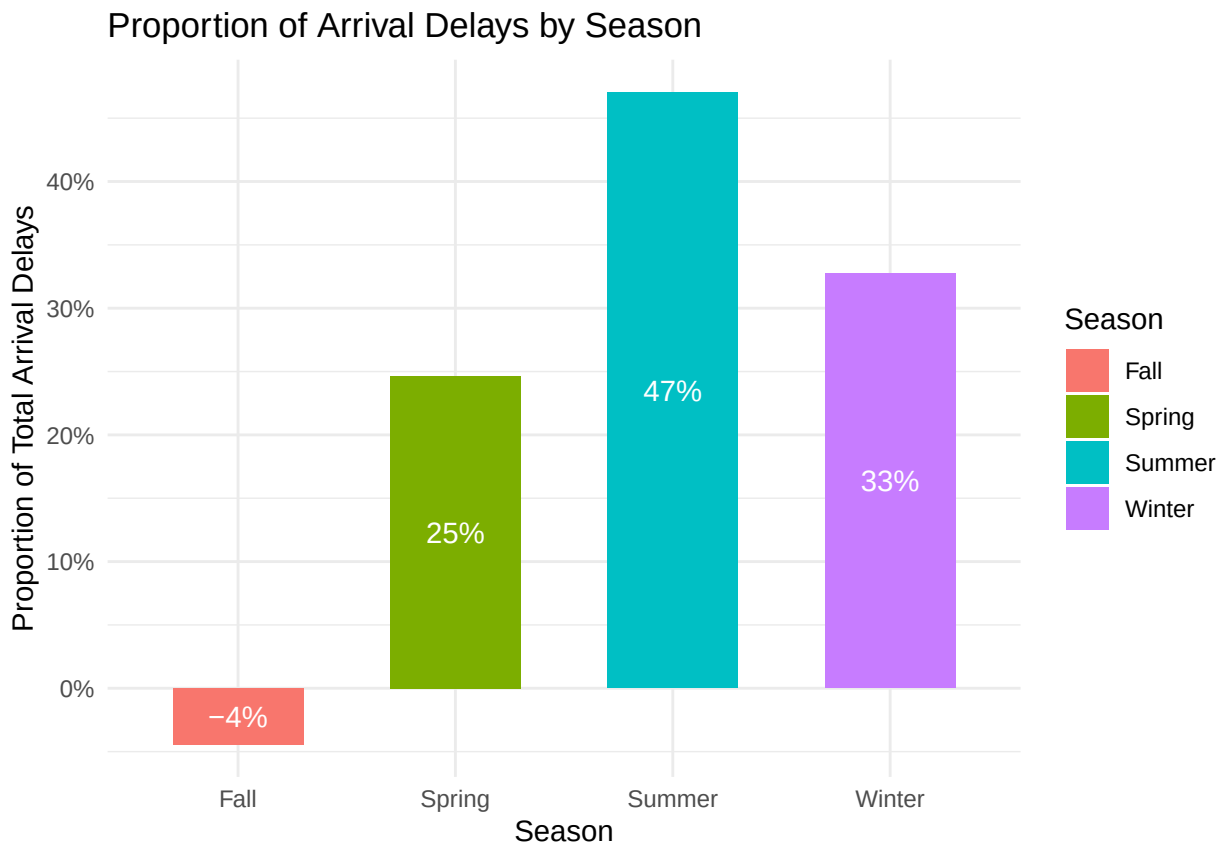

```
# Interpretation:
# Delays are often more significant in the late afternoon and evening (18h -19h) due to the
#accumulation of earlier delays.
```

```
# Seasonal Analysis for delays:
#first step : adding a column that indicates the according season for each month
flights_clean <- flights %>%
  filter(!is.na(dep_delay), !is.na(arr_delay)) %>%
  mutate(
    season = case_when(
      month %in% c(12, 1, 2) ~ "Winter",
      month %in% c(3, 4, 5) ~ "Spring",
      month %in% c(6, 7, 8) ~ "Summer",
      month %in% c(9, 10, 11) ~ "Fall"
    )
  )
```

```r
# Average delays by season
seasonal_delays <- flights_clean %>%
  group_by(season) %>%
  summarise(
    mean_dep_delay = mean(dep_delay, na.rm = TRUE),
    mean_arr_delay = mean(arr_delay, na.rm = TRUE),
    num_flights = n(),
    .groups = 'drop'
  )

print(seasonal_delays)
```

```
## # A tibble: 4 x 4
##   season mean_dep_delay mean_arr_delay num_flights
##   <chr>           <dbl>          <dbl>       <int>
## 1 Fall             6.10          -1.22       82599
## 2 Spring          13.3            6.81       83594
## 3 Summer          18.2           13.0        84124
## 4 Winter          12.5            9.04       77029
```

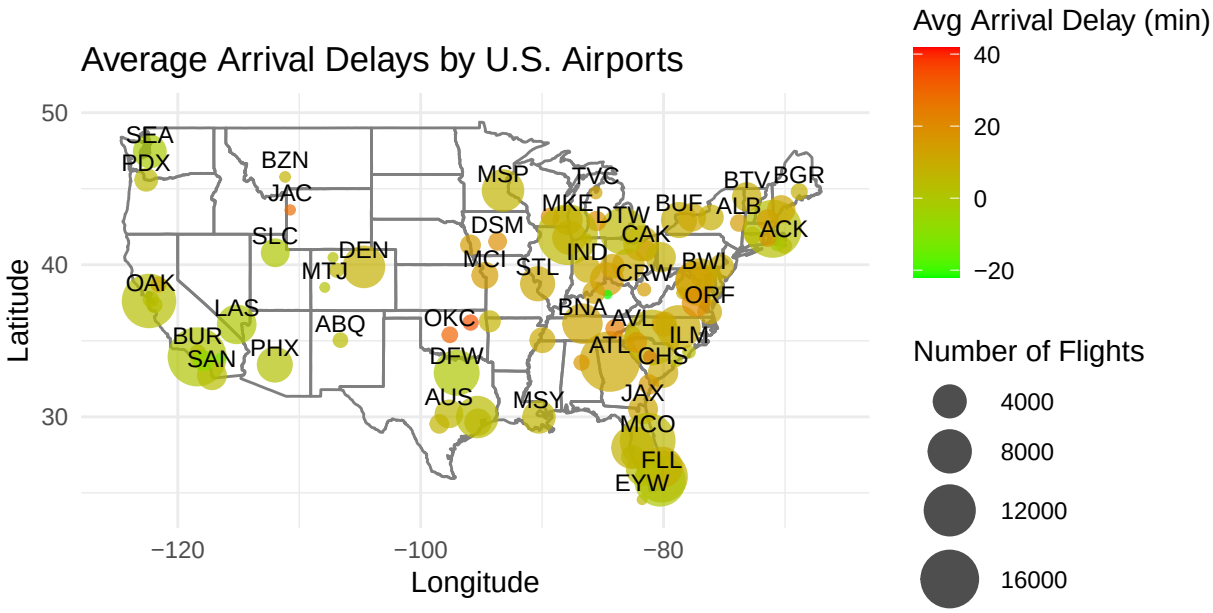
```r
# graphic representation of seasonal trends :
 ggplot(seasonal_delays, aes(x = season, y = mean_arr_delay / sum(mean_arr_delay), fill = season)) +
  geom_bar(stat = "identity", position = "stack", width = 0.6) +
  scale_y_continuous(labels = scales::percent) +
  labs(
    title = "Proportion of Arrival Delays by Season",
    x = "Season",
    y = "Proportion of Total Arrival Delays",
    fill = "Season"
  ) +
  geom_text(
    aes(
      label = scales::percent(mean_arr_delay / sum(mean_arr_delay), accuracy = 1),
      y = (mean_arr_delay / sum(mean_arr_delay)) / 2
    ),
    color = "white",
    size = 4
  ) +
  theme_minimal()
```




```
# Interpretation:
# Winter : often experiences higher delays, potentially due to weather conditions like snow and storms.
# Summer: also shows elevated delays due to increased travel demand and thunderstorms.
# Spring and Fall tend to have fewer delays, reflecting calmer weather and moderate travel demand.
```

```
#Delays by Airport : Summarizing average delays for each airport
airport_delays <- flights %>%
  group_by(dest) %>%
  summarise(
    avg_arr_delay = mean(arr_delay, na.rm = TRUE),
    avg_dep_delay = mean(dep_delay, na.rm = TRUE),
    num_flights = n(),
    .groups = 'drop'
  )

# Merging with airport locations
joined_airp_flights <- airport_delays %>% left_join(airports, by = c('dest'='faa'))

# Removing airports with no delay data or missing locations
joined_airp_flights <- joined_airp_flights %>%
  filter(!is.na(lat) & !is.na(lon) & !is.na(alt) & !is.na(avg_arr_delay) & !is.na(avg_dep_delay))

#graphic Visualization: Average Arrival Delays on a U.S. Map

# Visualization: Average Arrival Delays on a U.S. Map
```

```r
ggplot(data = joined_airp_flights) +
  borders("state") +
  geom_point(
    mapping = aes(x = lon, y = lat, colour = avg_arr_delay, size = num_flights),alpha = 0.7) +
  geom_text(
    mapping = aes(x = lon, y = lat, label = dest),
    size = 3, vjust = -0.5, hjust = 0.5, check_overlap = TRUE) +
  scale_colour_gradient(
    low = "green", high = "red", name = "Avg Arrival Delay (min)") +
    scale_size_continuous(
    name = "Number of Flights", range = c(1, 10))+
  labs(
    title = "Average Arrival Delays by U.S. Airports",
    x = "Longitude",
    y = "Latitude"
  ) +
  coord_quickmap(xlim = c(-125, -66), ylim = c(24, 50)) +
  theme_minimal()
```




```r
# Interpretation:
# The map highlights airports with the highest average arrival delays.
#Larger red circles indicate airports with more severe delays.
```

```r
# Airports with the Highest Delays
top_delayed_airports <- airport_delays %>%
  arrange(desc(avg_arr_delay)) %>%
  head(10)

top_delayed_airports
```

```
## # A tibble: 10 x 4
##    dest  avg_arr_delay avg_dep_delay num_flights
##    <chr>         <dbl>         <dbl>       <int>
##  1 CAE            41.8          35.6         116
##  2 TUL            33.7          34.9         315
##  3 OKC            30.6          30.6         346
##  4 JAC            28.1          26.5          25
##  5 TYS            24.1          28.5         631
##  6 MSN            20.2          23.6         572
##  7 RIC            20.1          23.6        2454
##  8 CAK            19.7          20.8         864
##  9 DSM            19.0          26.2         569
## 10 GRR            18.2          19.5         765
```

```r
# Interpretation:
# The table shows the 10 airports with the highest average arrival delays in 2013,
#In first place comes CAE: Columbia Metropolitan Airport, followed by Tul : Tulsa International Airport
#ad OKC: Will Rogers World Airport

#Study of correlation between flights volume and delays :

#The previous analysis shows a potential correlation between arrival delays & number of flights,
#some airports could experience mole delay due to busy flights , in order to study the association
#between the two variables we will conclude this project with a study of correlation betweeen
#flights volume and arrival delays of Usa flights in 2013

airport_delays_total<- flights%>%
  group_by(dest) %>%
  summarise(
    total_arr_delay = sum(arr_delay, na.rm = TRUE),
    total_dep_delay = sum(dep_delay, na.rm = TRUE),
    num_flights = n(),
    .groups = 'drop'
  )
cor(airport_delays_total$total_arr_delay, airport_delays_total$num_flights)
```

```
## [1] 0.7461309
```

```r
#interpretation :
#As the number of flights increases, the total arrival delays tend to increase
#as well (corr = 0.76) , This suggests that airports with higher flight volumes experience more
#delays overall, but it's important to note that this result doesnt indicate causation,
#some other factors like weather could affect too the arrival delays .
```

‘‘‘