ASSIGNMENT 7

GSI Intro to Big Data and Data Mining

The University of Texas at Austin

Zhaowen Fan
Rafael Ignacio Gonzalez Chong

Table of Contents

Create at least two of the following Visualization Ideas or create your own visualization idea using ggplot2 package.

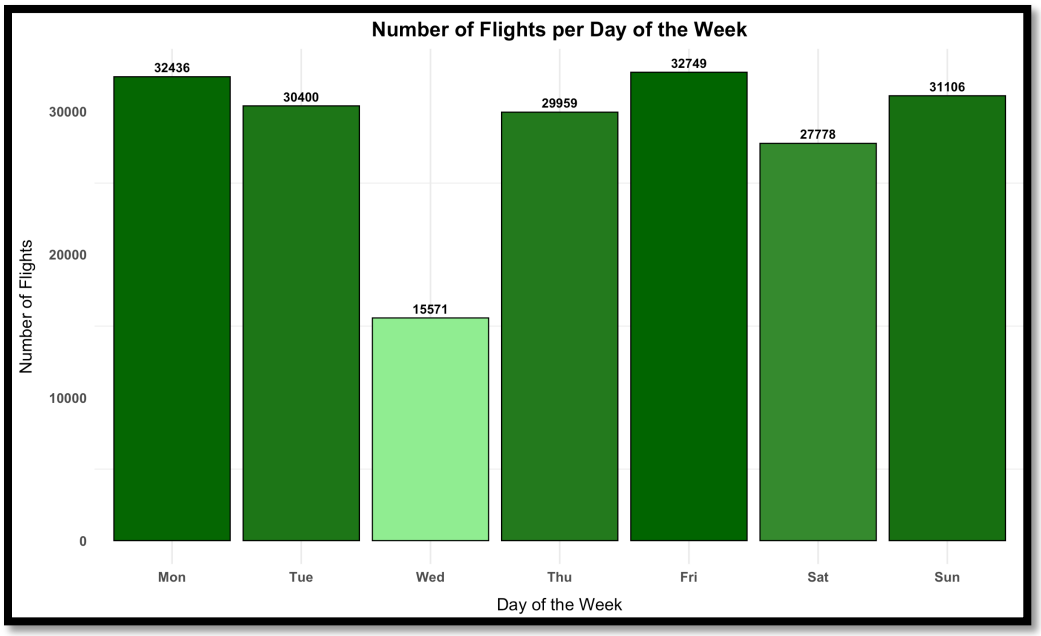

Fig 1. Number of flights per day of the week

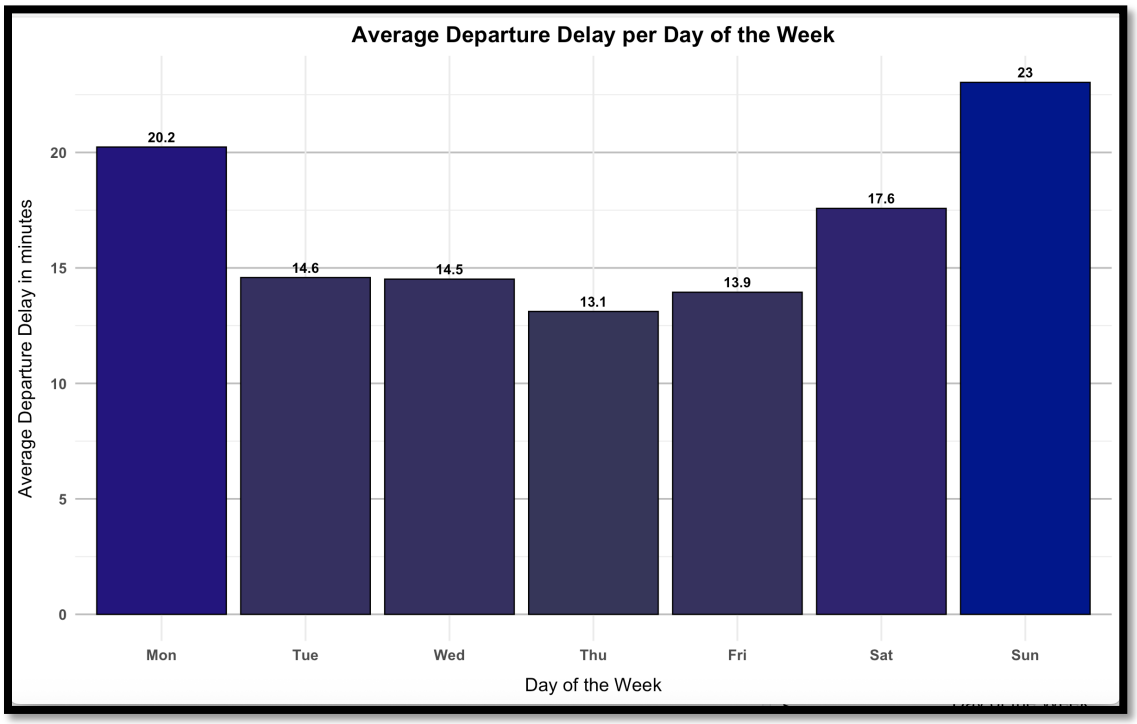

Fig. 2 Average departure delay per day of the week

Appendices (Code)

```
#ASSIGNMENT 7

#GSI Intro to Big Data and Data Mining

#Zhaowen Fan

#Rafael Ignacio Gonzalez Chong


library(dplyr)

library(ggplot2)


flights.file  <- "/Users/rafaelgonzalez/Desktop/assignment7/flights-small.csv"

flights <- read.csv(flights.file, stringsAsFactors = FALSE)


#Create at least two of the following Visualization Ideas,

#or create your own visualization idea using ggplot2 package.


#Number of flights per day of the week

flights %>%

  group_by(DAY_OF_WEEK) %>%

  summarise(num_flights = n()) %>%

  ggplot(aes(

    x = factor(DAY_OF_WEEK, levels = 1:7, labels = c("Mon", "Tue", "Wed", "Thu", "Fri", "Sat",
"Sun")),

    y = num_flights,
```

```
  fill = num_flights
)) +
geom_bar(stat = "identity", color = "black", show.legend = FALSE) +
geom_text(aes(label = num_flights), vjust = -0.5, size = 4, fontface = "bold") +
scale_fill_gradient(low = "lightgreen", high = "darkgreen") +
labs(
  title = "Number of Flights per Day of the Week",
  x = "Day of the Week",
  y = "Number of Flights",
) +
theme_minimal(base_size = 15) +
theme(
  plot.background = element_rect(fill = "white", color = NA),
  panel.grid.major.y = element_blank(),
  axis.title.x = element_text(margin = margin(t = 12)),
  axis.title.y = element_text(margin = margin(r = 12)),
  plot.title = element_text(face = "bold", size = 18, hjust = 0.5, color = "black"),
  plot.subtitle = element_text(size = 12, hjust = 0.5),
  axis.text = element_text(face = "bold")
)
```

```
#Average departure delay per day of the week
```

```
flights.delays <- flights %>%
  filter(CANCELLED == 0, !is.na(DEPARTURE_DELAY))


avg.delay <- flights.delays %>%
  group_by(DAY_OF_WEEK) %>%
  summarise(avg.departure.delay = mean(DEPARTURE_DELAY))


ggplot(avg.delay, aes(
  x = factor(DAY_OF_WEEK, levels = 1:7, labels = c("Mon", "Tue", "Wed", "Thu", "Fri", "Sat", "Sun")),
  y = avg.departure.delay, fill = avg.departure.delay
)) +
  geom_bar(stat = "identity", color = "black", show.legend = FALSE) +
  geom_text(aes(label = round(avg.departure.delay, 1)), vjust = -0.5, size = 4, fontface = "bold") +
  scale_fill_gradient2(low = "forestgreen", mid = "darkgreen", high = "darkblue", midpoint = 0) +
  labs(
    title = "Average Departure Delay per Day of the Week",
    x = "Day of the Week",
    y = "Average Departure Delay in minutes",
  ) +
  theme_minimal(base_size = 15) +
  theme(
    plot.background = element_rect(fill = "white", color = NA),
```

```
  panel.grid.major.y = element_line(color = "grey"),
  axis.title.x = element_text(margin = margin(t = 12)),
  axis.title.y = element_text(margin = margin(r = 12)),
  plot.title = element_text(face = "bold", size = 18, hjust = 0.5, color = "black"),
  plot.subtitle = element_text(size = 12, hjust = 0.5),
  axis.text = element_text(face = "bold")
)
```