NYC Flight Homework

by Robby Jeffries

Description

This page documents the first homework assignment for the data visualization course in University of Arkansas' M.S. in Economic Analytics. The full R script can be viewed by clicking here.

Load Necessary Libraries

Let's load some necessary libraries for this problem.

```
library(tidyverse)
library(nycflights13)
library(ggthemes)
```

Step 2. Filter the data

We wish to show a df of flights on March 10th, specifically Delta flights that originated at EWR.

```
flights %>%
  filter(month==3,
         day==10,
         carrier=="DL",
         origin=="EWR") %>%
  count()
```

Answer: There are **10** Delta flights that originated at EWR on March 10th, 2013.

Step 3. Generate a histogram

We want a histogram of the departure delays for each NYC airport.

```
ggplot(flights) +
  geom_histogram(aes(x = dep_delay),
                 color="red",
                 fill="red3",
                 alpha = 0.3,
                 bins = 15) +
  ggtitle("Plot of All Departure Delays by Airport") +
  xlab("Departure Delay (min)") +
  ylab("Number of Flights") +
  facet_wrap(~ origin) +
  theme_stata()
```
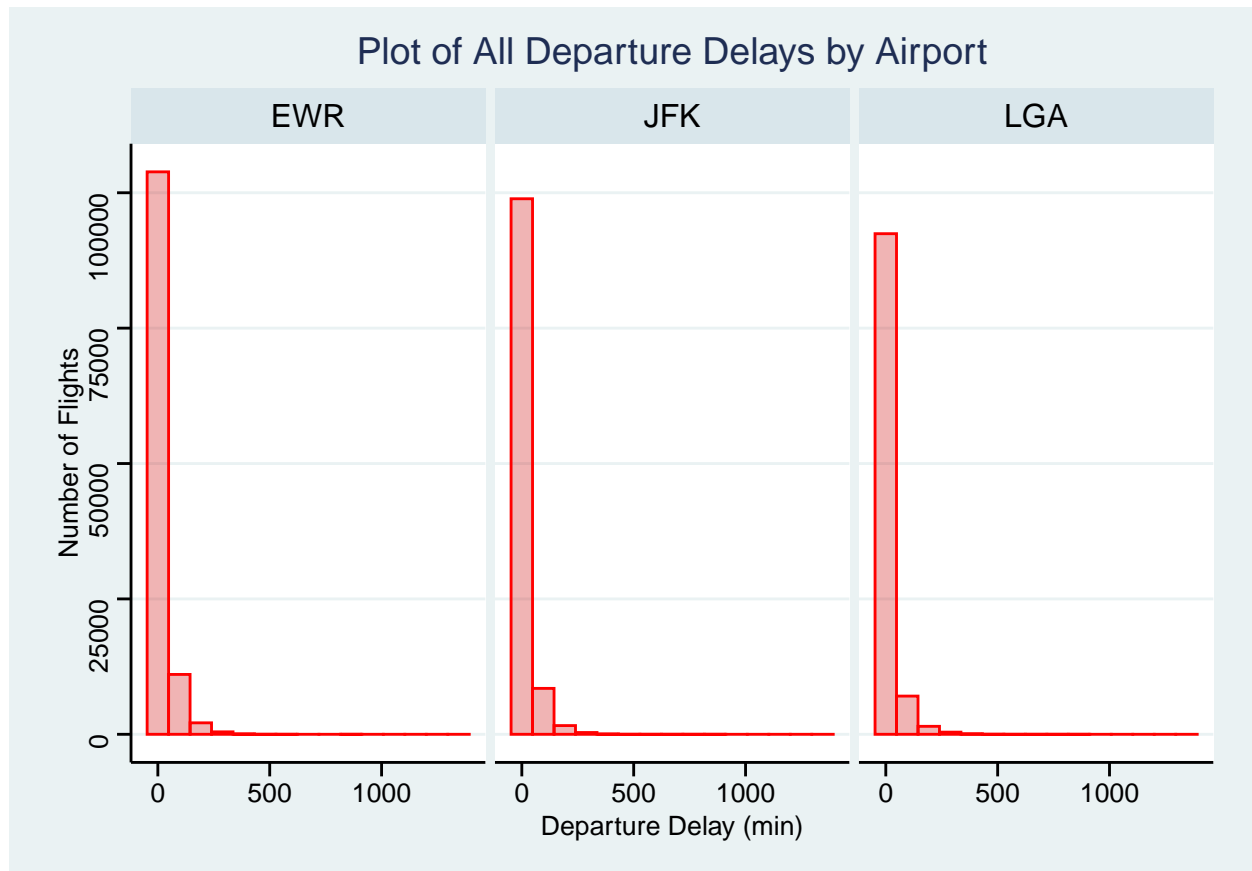
Because there are only 1,311 flights with departure delays above 250 minutes, let's filter out delays greater than 250 minutes to improve the graph's readability. It's important to note that this graph will not be representing all of the data.

```
ggplot(flights) +
  geom_histogram(aes(x = dep_delay),
                 color="red",
                 fill="red3",
                 alpha = 0.3,
                 bins = 15) +
  xlim(c(-50, 251)) +
  ggtitle("Plot of Departure Delays <= 250 min. by Airport") +
  xlab("Departure Delay (min)") +
  ylab("Number of Flights") +
  facet_wrap(~ origin) +
  theme_stata()
```

Plot of Departure Delays <= 250 min. by Airport
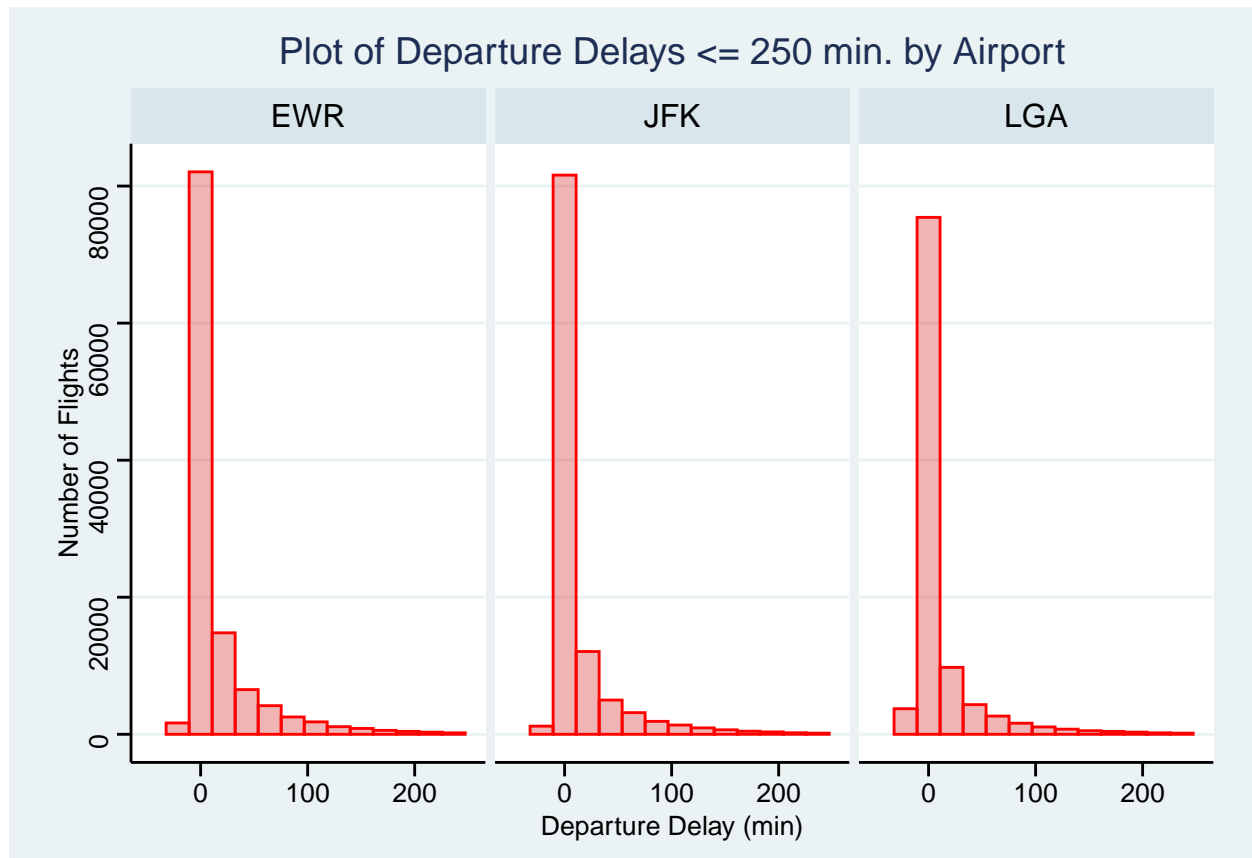
Step 4. Generate a custom tibble

For flights from **JFK only**, let's generate a tibble showing the average flight time (time in the air), grouped by the hour of the day it took off. Additionally, we'll show the average number of flights occuring at that hour every day.

```
# Merge the three date columns (year, month, day) into a single date column
flights$date <- as.Date(with(flights, paste(year, month, day, sep="-")), "%Y-%m-%d")
# flights$date

# Find the total number of flights from JFK. Use this to check work.
# flights %>% filter(origin=="JFK") %>% arrange(hour)

# To find the average number of flights occurring at each hour every day,
# create a variable that lists the number of flights per hour for each day.
flights_per_hour <- flights %>%
  filter(origin=="JFK") %>%
  group_by(date, hour) %>%
  count()
# flights_per_hour

# Next, aggregate the 'flights_per_hour' data to show the average number of flights for
# each hour across all days. I'll call this variable, 'test2'.
test2 <- aggregate(flights_per_hour$n~flights_per_hour$hour,
                   data=flights_per_hour,
```

```r
                      FUN=mean)
# test2

# We now have the average number of flights per hour at JFK. I will put them
# into a tibble.
avg_per_hour_every_day <- as_tibble(test2$`flights_per_hour$n`)
# avg_per_hour_every_day

# Find the average flight time
avg_flight_time <- flights %>%
  filter(origin=="JFK") %>%
  group_by(hour) %>%
  summarize(mean(air_time, na.rm=TRUE))
# avg_flight_time

# Place hour, avg_flight_time, and avg_per_hour_every_day into a single tibble
answer <- avg_flight_time %>%
  mutate(avg_per_hour_per_day=avg_per_hour_every_day)
answer
```

```
## # A tibble: 19 x 3
##     hour `mean(air_time, na.rm = TRUE)` avg_per_hour_per_day$value
##    <dbl>                          <dbl>                      <dbl>
## 1      5                           174.                       2.05
## 2      6                           167.                      17.2
## 3      7                           220.                      19.2
## 4      8                           157.                      29.5
## 5      9                           196.                      17.7
## 6     10                           247.                      12.8
## 7     11                           164.                       8.21
## 8     12                           149.                      13.8
## 9     13                           176.                      11.6
## 10    14                           124.                      21.5
## 11    15                           175.                      22.4
## 12    16                           184.                      23.7
## 13    17                           208.                      23.9
## 14    18                           202.                      21.7
## 15    19                           176.                      23.9
## 16    20                           178.                      17.0
## 17    21                           180.                       9.48
## 18    22                            69.0                      6.44
## 19    23                           186.                       2.84
```

Step 5. Generate a column chart

Using the dataframe from Step 4, let's generate a col chart with the hour of the day on the x axis and average flight time on the y axis.

```r
ggplot(answer) + geom_col(aes(x = answer$hour, y = answer$`mean(air_time, na.rm = TRUE)`), fill="dodgerl
  theme_economist() +
  ggtitle("Plot of Average Flight Time by Hour of the Day") +
  xlab("Hour of the Day") +
  ylab("Average Flight Time")
```

```
## Warning: Use of `answer$hour` is discouraged. Use `hour` instead.
```

```
## Warning: Use of `answer$`mean(air_time, na.rm = TRUE)`` is discouraged. Use
## `mean(air_time, na.rm = TRUE)` instead.
```

Plot of Average Flight Time by Hour of the Day