NYC Flight Data Analysis

Monica Buczynski

08/06/2021

Note: The purpose of this document is to showcase a sample of skills covered in *NYC Flight Data Analysis* by Soutik Chakraborty. All scripts were taken from http://soutik.github.io/NYC-Flight-Analysis/. The code for each exercise was studied carefully for understanding and then was retyped manually into R to maximize the learning experience; however, many of the original scripts were altered or I added my own original scripts for further experimentation and presentation aesthetics.

1) What was the worst day to fly out of NYC in 2013 if you dislike delayed flights?

The worst day to fly in 2013 (across all 3 airports) would've been the 12th September, 2013 where the average delay was 228 mins i.e approx 3 hrs 50 mins.

```
flights %>%
  select(month, day, arr_delay, dep_delay)  %>%
filter(arr_delay  >= 0, dep_delay >= 0) %>%
  group_by(month, day) %>%
  summarise(avg_delay = mean(arr_delay, na.rm = TRUE) +
              mean(dep_delay, na.rm = TRUE)) %>%
  ungroup() %>%
  arrange(-avg_delay)%>%
  head(1)
```

```
## # A tibble: 1 x 3
##   month   day avg_delay
##   <int> <int>     <dbl>
## 1     9    12      229.
```

2) Is there some particular airport with the highest delay in the 365 operation that needs to be avaoided?

We should have avoided LGA on 2nd September, 2013 and on 12th September, 2013 along with EWR on 12th September, 2013 if we want to avoid the highest delays in flight arrival and departures.

```
data1 <- flights %>%
  select(origin, month, day, arr_delay, dep_delay) %>%
  filter(arr_delay >= 0, dep_delay >= 0) %>%
  group_by(origin, month, day) %>%
  summarise(avg_delay = mean(arr_delay, na.rm = TRUE) +
              mean(dep_delay, na.rm = TRUE)) %>%
  ungroup() %>%
  arrange(-avg_delay)

head(data1,3)
```

```
## # A tibble: 3 x 4
##   origin month   day avg_delay
##   <chr>  <int> <int>     <dbl>
## 1 LGA        9     2      305.
## 2 LGA        9    12      252.
## 3 EWR        9    12      244.
```

3) Which airport has the greatest total average delay?

LGA has the greatest average total delay with 99.3 minutes followed closely by EWR and JFK with 96.9 and 95.5 minutes, respectively.

```
data2 <- flights %>%
  select(origin, arr_delay, dep_delay) %>%
  filter(arr_delay >= 0, dep_delay >= 0) %>%
  group_by(origin) %>%
  summarise(avg_delay = mean(arr_delay, na.rm = TRUE) +
              mean(dep_delay, na.rm = TRUE)) %>%
  ungroup() %>%
  arrange(-avg_delay)
data2
```

```
## # A tibble: 3 x 2
##   origin avg_delay
##   <chr>      <dbl>
## 1 LGA         99.3
## 2 EWR         96.9
## 3 JFK         95.5
```

4) Write a query to find the average arrival delay and the average departure delay at LGA, EWR and JFK.

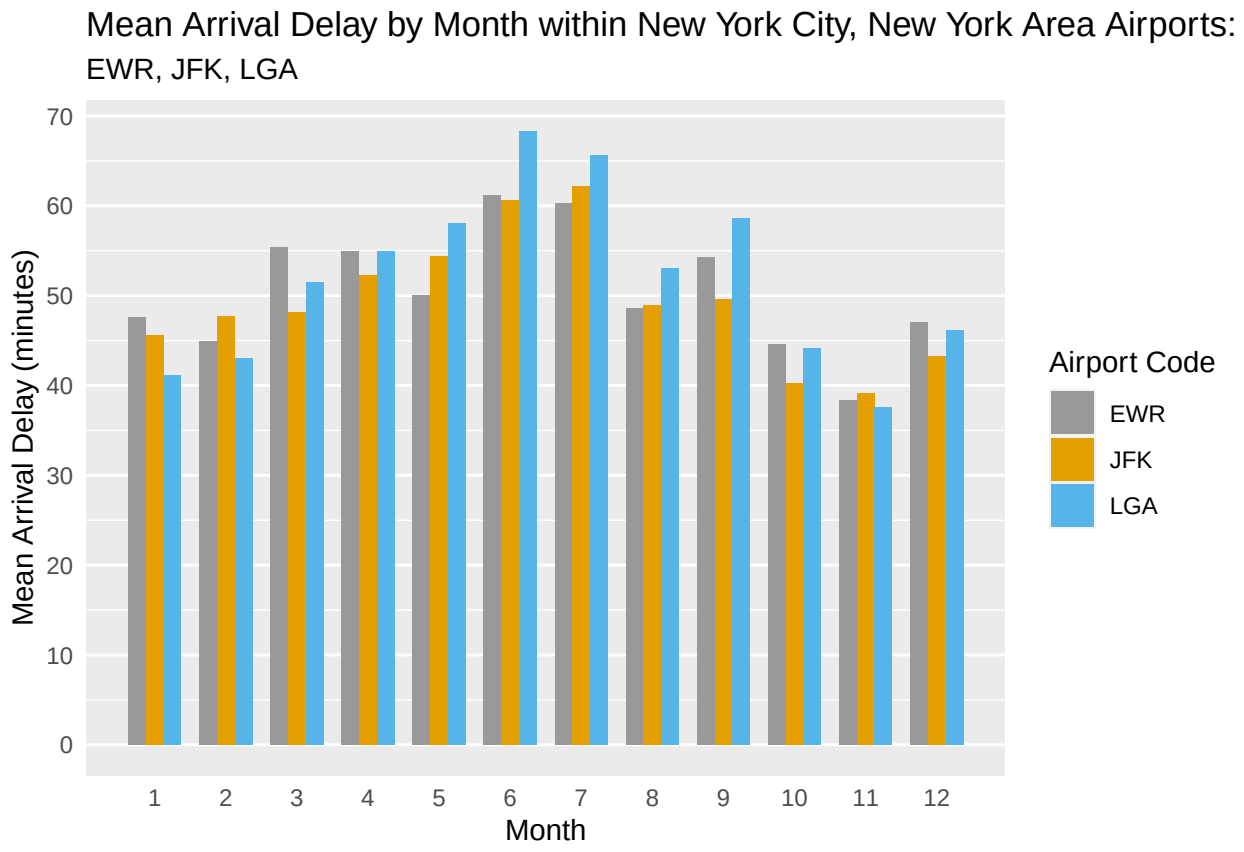
The average departure delay and average arrival delay times are no greater than 4 minutes apart for any flight originating in the NYC airport.

```
data3 <- flights %>%
  select(origin, arr_delay, dep_delay) %>%
  filter(arr_delay >= 0, dep_delay >= 0) %>%
  group_by(origin) %>%
    summarise(avg_arr_delay = mean(arr_delay, na.rm = TRUE),
              avg_dep_delay = mean(dep_delay, na.rm = TRUE))%>%
  ungroup() %>%
  arrange(-avg_arr_delay)
data3
```

```
## # A tibble: 3 x 3
##   origin avg_arr_delay avg_dep_delay
##   <chr>          <dbl>         <dbl>
## 1 LGA             50.6          48.7
## 2 EWR             48.8          48.1
## 3 JFK             48.8          46.8
```

```r
flights %>%
  select(origin, month, day, arr_delay, dep_delay) %>%
  filter(arr_delay > 0, dep_delay > 0) %>%
  group_by(origin, month) %>%
  summarise(mean_dep_delay = mean(dep_delay, na.rm = TRUE),
            mean_arr_delay = mean(dep_delay, na.rm = TRUE)) %>%
  pivot_longer(cols = c(mean_dep_delay,mean_arr_delay), names_to = "delay_type", values_to = "delay_amou
  filter(delay_type == "mean_arr_delay") %>%
ggplot(aes(x=month, y = delay_amount, fill = origin, width = .75)) +
  geom_bar(stat = "identity", position = position_dodge()) +
  scale_x_continuous(breaks = seq(1,12, by = 1)) +
  scale_y_continuous(breaks = seq(0,70, by = 10)) +
  theme(axis.ticks = element_blank(),
        panel.grid.major.x = element_blank(),
        panel.grid.minor.x = element_blank()) +
  labs(x = "Month",
       y = "Mean Arrival Delay (minutes)",
       title = "Mean Arrival Delay by Month within New York City, New York Area Airports:",
       subtitle = "EWR, JFK, LGA") +
  scale_fill_manual(values=c("#999999", "#E69F00", "#56B4E9"),
                    name="Airport Code")
```

```
## `summarise()` has grouped output by 'origin'. You can override using the `.groups` argument.
```



5) Are there any seasonal patterns in departure delays for flights from NYC?

Delays tend to peak in June, July, August and then decrease in winter months - may be due to increase of travelers in a hub city like NYC for domestic and international flights or possible due to repairs after the winter.

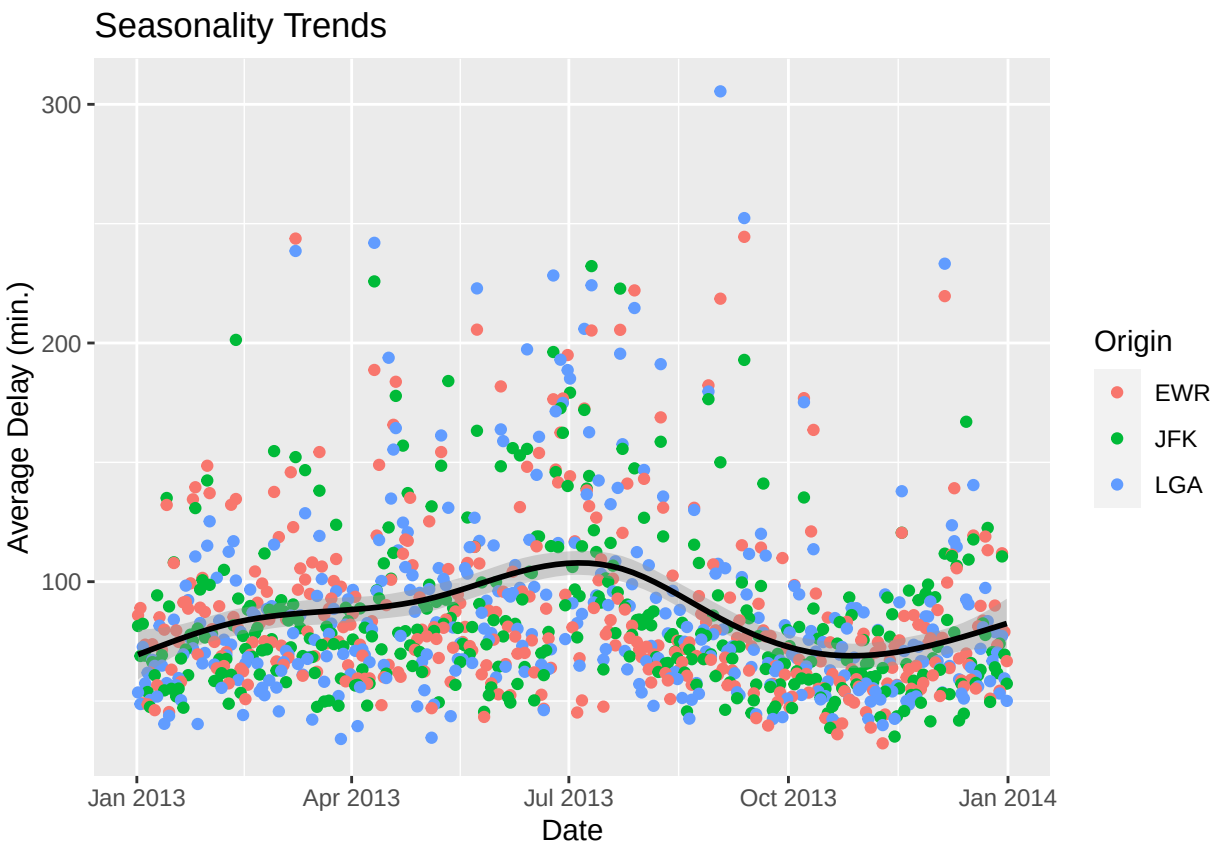
```
data <- flights %>%
  select(origin, month, day ,arr_delay, dep_delay) %>%
  filter(arr_delay >= 0, dep_delay >= 0) %>%
  group_by(origin, month, day) %>%
  summarise(avg_delay =  mean(arr_delay, na.rm = TRUE) +
              mean(dep_delay, na.rm = TRUE)) %>%
  ungroup() %>%
  arrange(-avg_delay)
head(data, 3)
```

```
## # A tibble: 3 x 4
##   origin month   day avg_delay
##   <chr>  <int> <int>     <dbl>
## 1 LGA        9     2      305.
## 2 LGA        9    12      252.
## 3 EWR        9    12      244.
```

```
data$date <- with(data, ISOdate(year = 2013, month, day))
```

```r
#Creating a ggplot function with x-axis = Date and y-axis = Average delay

ggplot(data, aes(x =date, y = avg_delay, color = origin)) +
  geom_point() +
  geom_smooth(color = "Black") +
  labs(x = "Date",
       y = "Average Delay (min.)",
       title = "Seasonality Trends",
       color = "Origin")
```



```
ggplot(data, aes(x=date, y=avg_delay)) +
  geom_smooth(color = "blue") +
  labs (x = "Date",
        y = "Average Delay (min.)",
      title = "Seasonality Trends")
```