hw2-sela-amir

Amir Sela 2024-10-05

Loading the data and tidyverse libraries

```
  #| label: Loading data and libraries
#if library not installed uncomment next line to install the library
#install.packages("nycflights13")
library(nycflights13)
library(tidyverse)
```

Quick view and summary of the data so we know what we are dealing with

There are ***336776*** rows of data, meaning a total of ***336776*** flights

```
   #| label: View the data
summary(flights) # quick info about the flights data
```

```
##       year          month             day           dep_time    sched_dep_time
##  Min.   :2013   Min.   : 1.000   Min.   : 1.00   Min.   :   1   Min.   : 106
##  1st Qu.:2013   1st Qu.: 4.000   1st Qu.: 8.00   1st Qu.: 907   1st Qu.: 906
##  Median :2013   Median : 7.000   Median :16.00   Median :1401   Median :1359
##  Mean   :2013   Mean   : 6.549   Mean   :15.71   Mean   :1349   Mean   :1344
##  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
##                                                  NA's   :8255
##    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.:1124   1st Qu.: -17.000
##  Median :  -2.00   Median :1535   Median :1556   Median :  -5.000
##  Mean   :  12.64   Mean   :1502   Mean   :1536   Mean   :   6.895
##  3rd Qu.:  11.00   3rd Qu.:1940   3rd Qu.:1945   3rd Qu.:  14.000
```

```
##  Max.   :1301.00   Max.   :2400   Max.   :2359   Max.   :1272.000
##  NA's   :8255      NA's   :8713                  NA's   :9430
##    carrier              flight       tailnum             origin
##  Length:336776      Min.   :   1   Length:336776      Length:336776
##  Class :character   1st Qu.: 553   Class :character   Class :character
##  Mode  :character   Median :1496   Mode  :character   Mode  :character
##                     Mean   :1972
##                     3rd Qu.:3465
##                     Max.   :8500
##
##      dest              air_time        distance         hour
##  Length:336776      Min.   : 20.0   Min.   :  17   Min.   : 1.00
##  Class :character   1st Qu.: 82.0   1st Qu.: 502   1st Qu.: 9.00
##  Mode  :character   Median :129.0   Median : 872   Median :13.00
##                     Mean   :150.7   Mean   :1040   Mean   :13.18
##                     3rd Qu.:192.0   3rd Qu.:1389   3rd Qu.:17.00
##                     Max.   :695.0   Max.   :4983   Max.   :23.00
##                     NA's   :9430
##      minute        time_hour
##  Min.   : 0.00   Min.   :2013-01-01 05:00:00
##  1st Qu.: 8.00   1st Qu.:2013-04-04 13:00:00
##  Median :29.00   Median :2013-07-03 10:00:00
##  Mean   :26.23   Mean   :2013-07-03 05:22:54
##  3rd Qu.:44.00   3rd Qu.:2013-10-01 07:00:00
##  Max.   :59.00   Max.   :2013-12-31 23:00:00
##
```

```
airlines # we know with what airlines we are dealing with and their carrier id
```

```
## # A tibble: 16 × 2
##    carrier name
##    <chr>   <chr>
##  1 9E      Endeavor Air Inc.
##  2 AA      American Airlines Inc.
##  3 AS      Alaska Airlines Inc.
##  4 B6      JetBlue Airways
##  5 DL      Delta Air Lines Inc.
##  6 EV      ExpressJet Airlines Inc.
##  7 F9      Frontier Airlines Inc.
##  8 FL      AirTran Airways Corporation
##  9 HA      Hawaiian Airlines Inc.
## 10 MQ      Envoy Air
## 11 OO      SkyWest Airlines Inc.
## 12 UA      United Air Lines Inc.
## 13 US      US Airways Inc.
```

```
## 14 VX      Virgin America
## 15 WN      Southwest Airlines Co.
## 16 YV      Mesa Airlines Inc.
```

Data Manipulation

The data will be joined so we can view the airline name as well, and the we will manipulate it by using the `dplyr` package of tidyverse, where we will select top 5 airlines by **number of flights**

```
  #| label: Data Manipulation
full_flights_data <- flights %>%
  left_join(airlines, by  = "carrier") %>% # joined data so we can see carrier nam
  relocate(name, .after = carrier) #by default R adds the name column in the end,
totalN_of_flights <- full_flights_data %>%
  count(carrier,name, sort = TRUE) #counts the total times the airline has had a f

top_five_airlines <- totalN_of_flights %>%
  slice_head( n = 5)# view the top 5 rows only
# i created a seperate table for top 5 because i wont want the totalN_of_flights t
top_five_airlines
```

```
## # A tibble: 5 × 3
##   carrier name                        n
##   <chr>   <chr>                   <int>
## 1 UA      United Air Lines Inc.    58665
## 2 B6      JetBlue Airways          54635
## 3 EV      ExpressJet Airlines Inc. 54173
## 4 DL      Delta Air Lines Inc.     48110
## 5 AA      American Airlines Inc.   32729
```

```
#another way we can do this is to first select the top 5 airlines and then join ta
```

Summary statistics

Each airline will have its summary statistics on *avg arrival delay*, *% of flights delayed* and *n of total flights*. This is great to make a comparison among airlines and their **delays**

```
  #| label: Summary Statistics

summaryStats_airlines <- full_flights_data %>%
  group_by(carrier, name) %>% # grouping by both carrier and name so we can see bo
  summarise(
    avg_arrival_delay = mean(arr_delay, na.rm = TRUE),
    percent_flights_delayed = (sum(arr_delay > 0, na.rm = TRUE) / n())* 100,
    total_flights = n()
    )
summaryStats_airlines_arranged <- summaryStats_airlines %>%
  arrange(percent_flights_delayed) # so we can see which airline has lowest percen

summaryStats_airlines_arranged
```

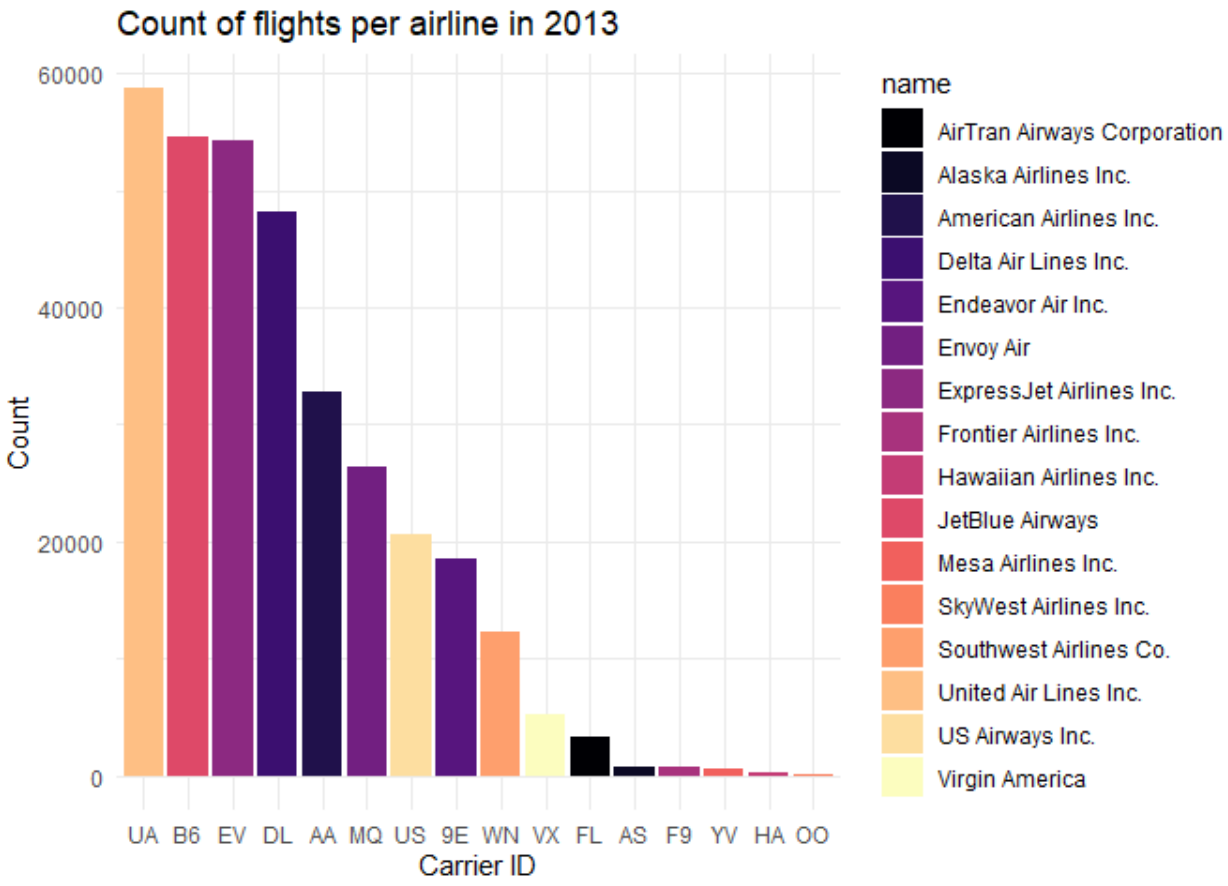
```
## # A tibble: 16 × 5
## # Groups:   carrier [16]
##    carrier name           avg_arrival_delay percent_flights_dela…¹ total_flight
##    <chr>   <chr>                      <dbl>                 <dbl>         <int
##  1 AS      Alaska Airlin…            -9.93                  26.5           71
##  2 HA      Hawaiian Airl…            -6.92                  28.4           34
##  3 OO      SkyWest Airli…            11.9                   31.2            3
##  4 AA      American Airl…             0.364                 32.7         3272
##  5 VX      Virgin America             1.76                  33.8          516
##  6 DL      Delta Air Lin…             1.64                  34.1         4811
##  7 US      US Airways In…             2.13                  35.8         2053
##  8 9E      Endeavor Air …             7.38                  36.0         1846
##  9 UA      United Air Li…             3.56                  37.9         5866
## 10 YV      Mesa Airlines…            15.6                   42.9           60
## 11 WN      Southwest Air…             9.65                  43.2         1227
## 12 B6      JetBlue Airwa…             9.46                  43.2         5463
## 13 MQ      Envoy Air                 10.8                   44.3         2639
## 14 EV      ExpressJet Ai…            15.8                   45.2         5417
## 15 F9      Frontier Airl…            21.9                   57.2           68
## 16 FL      AirTran Airwa…            20.1                   58.1          326
## # ℹ abbreviated name: ¹percent_flights_delayed
```

Plots

We will explore a couple graphs and also explain what can we learn from these plots/ graphs

Bar Plot

```
ggplot(full_flights_data[c("carrier", "name")], aes(x = reorder(carrier, -table(ca
  geom_bar(stat = "count") +
  scale_fill_viridis_d(option = "magma") + # this code gives more distinguishable
  labs(
    x = "Carrier ID",
    y = "Count",
    title = "Count of flights per airline in 2013"
  ) +
  theme_minimal()
```




```
# NOTE : i added the legend on the right, but instead of using carrier ID i used n
```

Now we have the bar plot, we dont really know the excact count per airline, so to make it look nice in a table format, i will use a new library called `knitr`

Table of counts per airline

```
#| label: Loading Table

library(knitr) # loading the library

kable(totalN_of_flights, col.names = c("Carrier","Name", "Count"))
```

Carrier	Name	Count
UA	United Air Lines Inc.	58665
B6	JetBlue Airways	54635
EV	ExpressJet Airlines Inc.	54173
DL	Delta Air Lines Inc.	48110
AA	American Airlines Inc.	32729
MQ	Envoy Air	26397
US	US Airways Inc.	20536
9E	Endeavor Air Inc.	18460
WN	Southwest Airlines Co.	12275
VX	Virgin America	5162
FL	AirTran Airways Corporation	3260
AS	Alaska Airlines Inc.	714
F9	Frontier Airlines Inc.	685
YV	Mesa Airlines Inc.	601
HA	Hawaiian Airlines Inc.	342
OO	SkyWest Airlines Inc.	32

```
# this chunk basically take the data and puts it in a table format
```

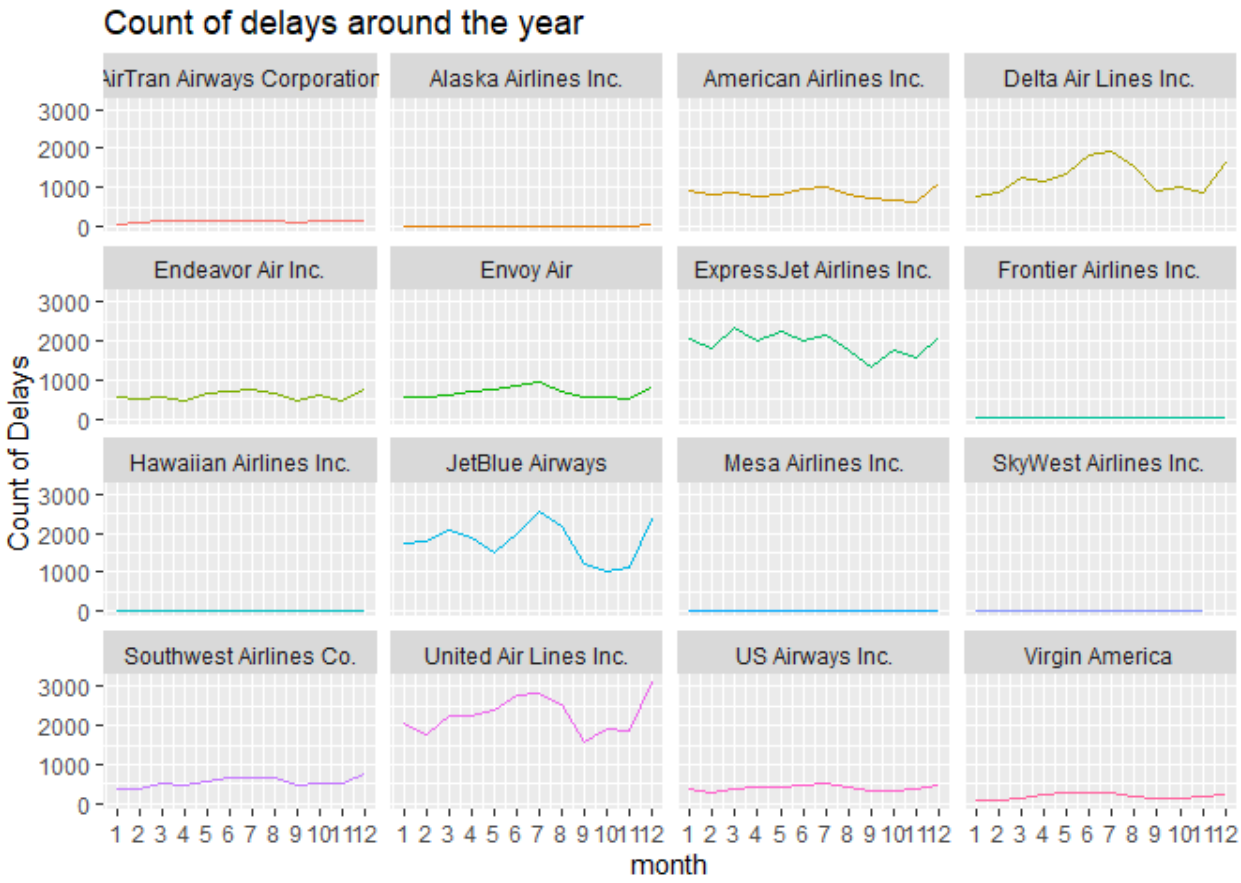
Line graph

We will explore ***trends*** over time (monthly) with a line plot

```
monthly_delays <- full_flights_data %>%
  filter(dep_delay > 0) %>% # we are filtering the df to only include rows where d
  group_by(name, month) %>%
  summarise(count_of_delays = n())
```

```
## `summarise()` has regrouped the output.
## ℹ Summaries were computed grouped by name and month.
## ℹ Output is grouped by name.
## ℹ Use `summarise(.groups = "drop_last")` to silence this message.
## ℹ Use `summarise(.by = c(name, month))` for per-operation grouping
##   (`?dplyr::dplyr_by`) instead.
```

```
ggplot(monthly_delays, aes(x = month, y = count_of_delays, color = name, group = n
  geom_line() +
  scale_x_continuous(breaks = seq(1,12, by = 1)) + # i added this line beacuse wit
  facet_wrap(~name, ncol = 4)+# the reason i faceted is because if we show all the
  labs(
    x = "month",
    y = "Count of Delays",
    title = "Count of delays around the year"
  )+
  theme(legend.position = "none") # we dont need legend, it only takes space and m
```



Hex Plot

Next we are going to ***analyze*** is there a relationship between `departure time` and `departure delay` per airline. One might think that this is useles, but this **hex plot** will tell us if there are less/more delays depending on what time of the day the airplane is departing

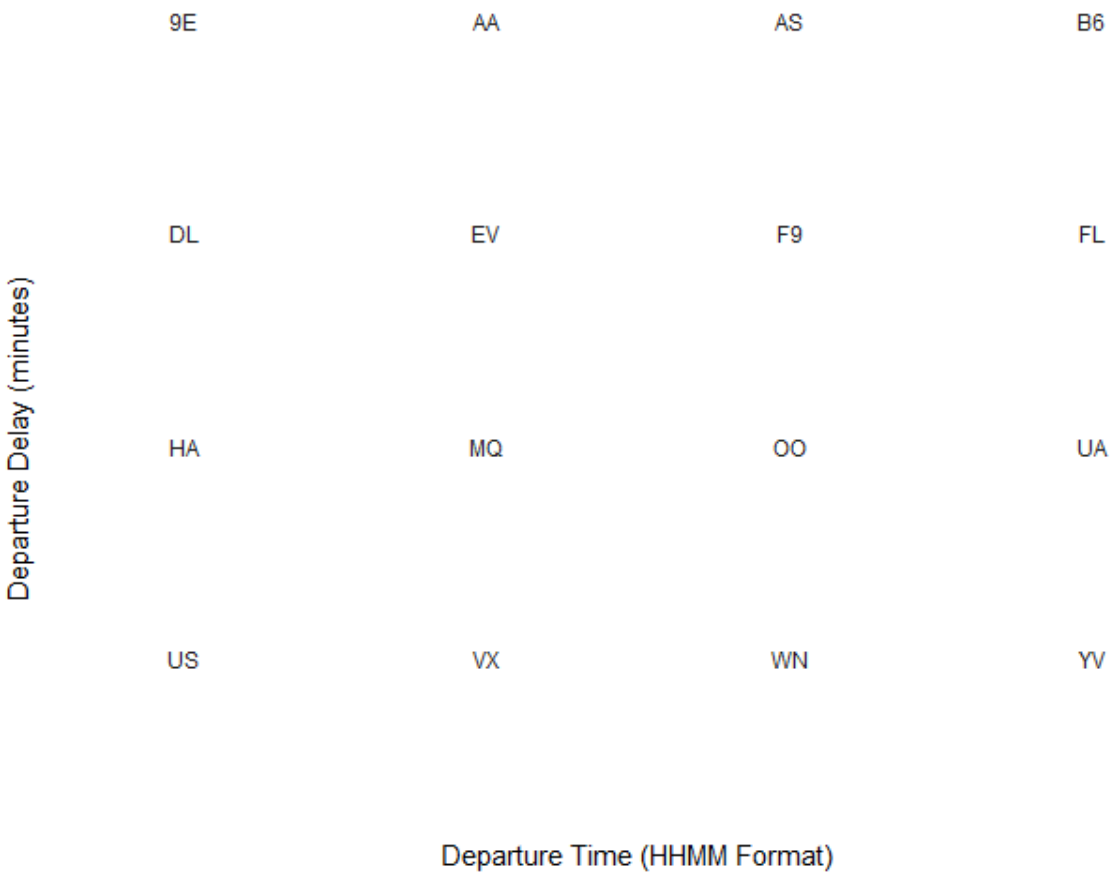
```
#| label: Hex Plot
#first create a df where we only need those three columns - carrier, dep_time and (

depTable <- full_flights_data %>%
  select(carrier,dep_time,dep_delay) %>%
  filter(!is.na(dep_time), !is.na(dep_delay)) # filter out NA
```

```
ggplot(depTable, aes(x = dep_time, y = dep_delay, group = carrier)) +
  geom_hex(bins = 30) +
  facet_wrap(~carrier) +
  scale_fill_gradient(low = "lightpink", high = "darkred") +# Adjusting color scal
```

```
  scale_x_continuous(breaks = seq(0,2400, by = 800))+ #adjusting x scale so it mor
  labs(x = "Departure Time (HHMM Format)", y = "Departure Delay (minutes)") +
  theme_minimal()
```

```
## Warning: Computation failed in `stat_binhex()`.
## Computation failed in `stat_binhex()`.
## Computation failed in `stat_binhex()`.
## Computation failed in `stat_binhex()`.
## Computation failed in `stat_binhex()`.
## Computation failed in `stat_binhex()`.
## Computation failed in `stat_binhex()`.
## Computation failed in `stat_binhex()`.
## Computation failed in `stat_binhex()`.
## Computation failed in `stat_binhex()`.
## Computation failed in `stat_binhex()`.
## Computation failed in `stat_binhex()`.
## Computation failed in `stat_binhex()`.
## Computation failed in `stat_binhex()`.
## Computation failed in `stat_binhex()`.
## Caused by error in `compute_group()`:
## ! The package "hexbin" is required for `stat_bin_hex()`.
```



```
# the reason i did a hex plot over a scatter plot is because graphing the scatter
```

Business Analysis

Based on my findings, i will discuss which airline has the most consistent on-time performance.

Based on the summary statistics table, the airline with the most consistent on-time perfomance is **Alaska Airlines Inc.** with the lowest percentage of flights delayed at **26.4705882**. But this doesnt mean that they still are the best when it comes to on-time perfomance beacuse they only have **714** total of flights. When comparing to **American Airlines Inc.** that has a percentage of delayed flights at **32.7110514** which is slightly higher, but it has a lot more total flight at **32729**. Now beacuse they have a lot more flights, propability is that they have a little more delayed flights.

How this performance could affect profitability (e.g., customer satisfaction, repeat business)?

The perfomance of having a low percentage of flights delayed flights could result in having more repeat business and costumer satisfaction. But this analysis doesnt include of how happy costumers are with the airlines in-flight services. As long as the airline keeps the costumer satisfaction in good shape, when combined with low-delay percentage their profitability will increase in time.

A brief summary of my findings, including recommendations for airlines looking to improve their on-time performance.

From **Bar Plot** - My findings concluded that the airline with the most flights is United Air Lines Inc. at a # of total flights being 58665.

From **Line Graphs** - We found that half of the airlines have a steady continious amount

of delays around the year, but the other half have increasing and decreasing # of delays around the year. We can see that the airlines which have a high number of delayes all year round, they also tend to increase their # of delays when it reaches March.

From **Hex Plot** - This is a very interesting finding. According to the hex plot for almost all airlines after estimated 6 AM the number of delays increases until estimated 3AM. This could give a costumer some very useful information when booking their flight, such as if they should increase their chances of expecting a delay depending by the time of the day.