INFX 573: Problem Set 1 - Exploring Data

Pierre Augustamar

Due: Monday, October 11, 2016

Collaborators:

Instructions:

Before beginning this assignment, please ensure you have access to R and RStudio.

1. Download the `problemset1.Rmd` file from Canvas. Open `problemset1.Rmd` in RStudio and supply your solutions to the assignment by editing `problemset1.Rmd`.
2. Replace the "Insert Your Name Here" text in the `author:` field with your own full name. Any collaborators must be listed on the top of your assignment.
3. Be sure to include well-documented (e.g. commented) code chucks, figures and clearly written text chunk explanations as necessary. Any figures should be clearly labeled and appropriately referenced within the text.
4. Collaboration on problem sets is acceptable, and even encouraged, but each student must turn in an individual write-up in his or her own words and his or her own work. The names of all collaborators must be listed on each assignment. Do not copy-and-paste from other students' responses or code.
5. When you have completed the assignment and have **checked** that your code both runs in the Console and knits correctly when you click `Knit PDF`, rename the R Markdown file to `YourLastName_YourFirstName_ps1.Rmd`, knit a PDF and submit both the PDF file on Canvas.

Setup:

In this problem set you will need, at minimum, the following R packages.

```
# Load standard libraries
library(tidyverse)
library(nycflights13)
```

Problem 1: Exploring the NYC Flights Data

In this problem set we will use the data on all flights that departed NYC (i.e. JFK, LGA or EWR) in 2013. You can find this data in the `nycflights13` R package.

(a) Importing and Inspecting Data:

Load the data and describe in a short paragraph how the data was collected and what each variable represents. Perform a basic inspection of the data and discuss what you find.

Response Data collection process:

Since the nycflights13 is part of the packages made available by cran-r-project.org, I used the information provided by the reference manualwhich is located at https://cran.r-project.org/web/packages/nycflights13/nycflights13.pdf.

Per this document, this package is a collection of airline data for all flights departing NYC in 2013. It also includes data on airlines, airports, weather, and planes. It provides an explanation of each metada about the main components. I used this information to help direct my thought process on how I will be exploring the data and formulating questions for analysis.

Response basic data inspection:

For an initial inspection of the data, I use the summary function which provides a descriptive statistics of a dataset. Also, I use the str function which provides information about the structure of the objects of a data set. Furthermore, I use the head and tail function to have a sense of the data by retrieving the first and last values of the datasets. Finally, I used a combination of sapply, sum and is.na functions to filter out components that have a large set of missing data. Note that even though I have run script to inspect each of the components, my primary area of interests are mainly with the flights data. Thus, my response will solely related to flights specific data.

```
# Summary results

summary(flights)
```

```
##       year          month             day           dep_time
##  Min.   :2013   Min.   : 1.000   Min.   : 1.00   Min.   :   1
##  1st Qu.:2013   1st Qu.: 4.000   1st Qu.: 8.00   1st Qu.: 907
##  Median :2013   Median : 7.000   Median :16.00   Median :1401
##  Mean   :2013   Mean   : 6.549   Mean   :15.71   Mean   :1349
##  3rd Qu.:2013   3rd Qu.:10.000   3rd Qu.:23.00   3rd Qu.:1744
##  Max.   :2013   Max.   :12.000   Max.   :31.00   Max.   :2400
##                                                  NA's   :8255
##  sched_dep_time   dep_delay          arr_time    sched_arr_time
##  Min.   : 106   Min.   : -43.00   Min.   :   1   Min.   :   1
##  1st Qu.: 906   1st Qu.:  -5.00   1st Qu.:1104   1st Qu.:1124
##  Median :1359   Median :  -2.00   Median :1535   Median :1556
##  Mean   :1344   Mean   :  12.64   Mean   :1502   Mean   :1536
##  3rd Qu.:1729   3rd Qu.:  11.00   3rd Qu.:1940   3rd Qu.:1945
##  Max.   :2359   Max.   :1301.00   Max.   :2400   Max.   :2359
##                 NA's   :8255      NA's   :8713
##    arr_delay          carrier              flight       tailnum
##  Min.   : -86.000   Length:336776      Min.   :   1   Length:336776
##  1st Qu.: -17.000   Class :character   1st Qu.: 553   Class :character
##  Median :  -5.000   Mode  :character   Median :1496   Mode  :character
##  Mean   :   6.895                      Mean   :1972
##  3rd Qu.:  14.000                      3rd Qu.:3465
##  Max.   :1272.000                      Max.   :8500
##  NA's   :9430
##     origin              dest              air_time        distance
##  Length:336776      Length:336776      Min.   : 20.0   Min.   :  17
##  Class :character   Class :character   1st Qu.: 82.0   1st Qu.: 502
##  Mode  :character   Mode  :character   Median :129.0   Median : 872
##                                        Mean   :150.7   Mean   :1040
```

```
##                                      3rd Qu.:192.0   3rd Qu.:1389
##                                      Max.   :695.0   Max.   :4983
##                                      NA's   :9430
##       hour           minute        time_hour
##  Min.   : 1.00   Min.   : 0.00   Min.   :2013-01-01 05:00:00
##  1st Qu.: 9.00   1st Qu.: 8.00   1st Qu.:2013-04-04 13:00:00
##  Median :13.00   Median :29.00   Median :2013-07-03 10:00:00
##  Mean   :13.18   Mean   :26.23   Mean   :2013-07-03 05:02:36
##  3rd Qu.:17.00   3rd Qu.:44.00   3rd Qu.:2013-10-01 07:00:00
##  Max.   :23.00   Max.   :59.00   Max.   :2013-12-31 23:00:00
##
```

```
summary(airports)
```

```
##      faa                name               lat             lon
##  Length:1396        Length:1396        Min.   :19.72   Min.   :-176.65
##  Class :character   Class :character   1st Qu.:34.27   1st Qu.:-119.34
##  Mode  :character   Mode  :character   Median :40.15   Median : -94.92
##                                        Mean   :41.76   Mean   :-103.71
##                                        3rd Qu.:45.26   3rd Qu.: -82.55
##                                        Max.   :72.27   Max.   : 174.11
##       alt               tz              dst
##  Min.   : -54.0   Min.   :-11.000   Length:1396
##  1st Qu.:  70.0   1st Qu.: -8.000   Class :character
##  Median : 481.5   Median : -6.000   Mode  :character
##  Mean   :1006.3   Mean   : -6.422
##  3rd Qu.:1076.2   3rd Qu.: -5.000
##  Max.   :9078.0   Max.   :  8.000
```

```
summary(airlines)
```

```
##    carrier             name
##  Length:16          Length:16
##  Class :character   Class :character
##  Mode  :character   Mode  :character
```

```
summary(planes)
```

```
##    tailnum               year          type           manufacturer
##  Length:3322        Min.   :1956   Length:3322        Length:3322
##  Class :character   1st Qu.:1997   Class :character   Class :character
##  Mode  :character   Median :2001   Mode  :character   Mode  :character
##                     Mean   :2000
##                     3rd Qu.:2005
##                     Max.   :2013
##                     NA's   :70
##     model              engines          seats           speed
##  Length:3322        Min.   :1.000   Min.   :  2.0   Min.   : 90.0
##  Class :character   1st Qu.:2.000   1st Qu.:140.0   1st Qu.:107.5
##  Mode  :character   Median :2.000   Median :149.0   Median :162.0
##                     Mean   :1.995   Mean   :154.3   Mean   :236.8
##                     3rd Qu.:2.000   3rd Qu.:182.0   3rd Qu.:432.0
```

```
##                     Max.   :4.000   Max.   :450.0   Max.   :432.0
##                                                    NA's   :3299
##     engine
##  Length:3322
##  Class :character
##  Mode  :character
##
##
##
##
```

```
summary(weather)
```

```
##     origin               year          month              day
##  Length:26130       Min.   :2013   Min.   : 1.000   Min.   : 1.00
##  Class :character   1st Qu.:2013   1st Qu.: 4.000   1st Qu.: 8.00
##  Mode  :character   Median :2013   Median : 7.000   Median :16.00
##                     Mean   :2013   Mean   : 6.506   Mean   :15.68
##                     3rd Qu.:2013   3rd Qu.: 9.000   3rd Qu.:23.00
##                     Max.   :2013   Max.   :12.000   Max.   :31.00
##
##       hour            temp            dewp           humid
##  Min.   : 0.00   Min.   : 10.94   Min.   :-9.94   Min.   : 12.74
##  1st Qu.: 6.00   1st Qu.: 39.92   1st Qu.:26.06   1st Qu.: 46.99
##  Median :12.00   Median : 55.04   Median :42.08   Median : 61.66
##  Mean   :11.52   Mean   : 55.20   Mean   :41.39   Mean   : 62.35
##  3rd Qu.:18.00   3rd Qu.: 69.98   3rd Qu.:57.92   3rd Qu.: 78.62
##  Max.   :23.00   Max.   :100.04   Max.   :78.08   Max.   :100.00
##                  NA's   :1        NA's   :1       NA's   :1
##     wind_dir       wind_speed       wind_gust          precip
##  Min.   :  0.0   Min.   :   0.000   Min.   :   0.000   Min.   :0.000000
##  1st Qu.:120.0   1st Qu.:   6.905   1st Qu.:   7.946   1st Qu.:0.000000
##  Median :220.0   Median :   9.206   Median :  10.594   Median :0.000000
##  Mean   :198.1   Mean   :  10.396   Mean   :  11.963   Mean   :0.002726
##  3rd Qu.:290.0   3rd Qu.:  13.809   3rd Qu.:  15.892   3rd Qu.:0.000000
##  Max.   :360.0   Max.   :1048.361   Max.   :1206.432   Max.   :1.180000
##  NA's   :418     NA's   :3          NA's   :3
##     pressure          visib          time_hour
##  Min.   : 983.8   Min.   : 0.000   Min.   :2012-12-31 16:00:00
##  1st Qu.:1012.9   1st Qu.:10.000   1st Qu.:2013-04-01 14:00:00
##  Median :1017.6   Median :10.000   Median :2013-07-01 07:30:00
##  Mean   :1017.9   Mean   : 9.205   Mean   :2013-07-01 12:07:20
##  3rd Qu.:1023.0   3rd Qu.:10.000   3rd Qu.:2013-09-30 07:45:00
##  Max.   :1042.1   Max.   :10.000   Max.   :2013-12-30 15:00:00
##  NA's   :2730
```

The summary reports for fligths show that on average 12.6 flights had a departure delay, 1556 flights arrived at their destinations at the scheduled time, and 6.9 flights had a delayed arrival. It also shows that all the flights that left New York did arrive at their destinations at some point.

```
# str results

str(flights)
```

```
## Classes 'tbl_df', 'tbl' and 'data.frame':    336776 obs. of  19 variables:
##  $ year          : int  2013 2013 2013 2013 2013 2013 2013 2013 2013 2013 ...
##  $ month         : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ day           : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ dep_time      : int  517 533 542 544 554 554 555 557 557 558 ...
##  $ sched_dep_time: int  515 529 540 545 600 558 600 600 600 600 ...
##  $ dep_delay     : num  2 4 2 -1 -6 -4 -5 -3 -3 -2 ...
##  $ arr_time      : int  830 850 923 1004 812 740 913 709 838 753 ...
##  $ sched_arr_time: int  819 830 850 1022 837 728 854 723 846 745 ...
##  $ arr_delay     : num  11 20 33 -18 -25 12 19 -14 -8 8 ...
##  $ carrier       : chr  "UA" "UA" "AA" "B6" ...
##  $ flight        : int  1545 1714 1141 725 461 1696 507 5708 79 301 ...
##  $ tailnum       : chr  "N14228" "N24211" "N619AA" "N804JB" ...
##  $ origin        : chr  "EWR" "LGA" "JFK" "JFK" ...
##  $ dest          : chr  "IAH" "IAH" "MIA" "BQN" ...
##  $ air_time      : num  227 227 160 183 116 150 158 53 140 138 ...
##  $ distance      : num  1400 1416 1089 1576 762 ...
##  $ hour          : num  5 5 5 5 6 5 6 6 6 6 ...
##  $ minute        : num  15 29 40 45 0 58 0 0 0 0 ...
##  $ time_hour     : POSIXct, format: "2013-01-01 05:00:00" "2013-01-01 05:00:00" ...
```

```
str(airports)
```

```
## Classes 'tbl_df', 'tbl' and 'data.frame':    1396 obs. of  7 variables:
##  $ faa : chr  "04G" "06A" "06C" "06N" ...
##  $ name: chr  "Lansdowne Airport" "Moton Field Municipal Airport" "Schaumburg Regional" "Randall Airp
##  $ lat : num  41.1 32.5 42 41.4 31.1 ...
##  $ lon : num  -80.6 -85.7 -88.1 -74.4 -81.4 ...
##  $ alt : int  1044 264 801 523 11 1593 730 492 1000 108 ...
##  $ tz  : num  -5 -5 -6 -5 -4 -4 -5 -5 -5 -8 ...
##  $ dst : chr  "A" "A" "A" "A" ...
```

```
str(airlines)
```

```
## Classes 'tbl_df', 'tbl' and 'data.frame':    16 obs. of  2 variables:
##  $ carrier: chr  "9E" "AA" "AS" "B6" ...
##  $ name   : chr  "Endeavor Air Inc." "American Airlines Inc." "Alaska Airlines Inc." "JetBlue Airways
```

```
str(planes)
```

```
## Classes 'tbl_df', 'tbl' and 'data.frame':    3322 obs. of  9 variables:
##  $ tailnum     : chr  "N10156" "N102UW" "N103US" "N104UW" ...
##  $ year        : int  2004 1998 1999 1999 2002 1999 1999 1999 1999 1999 ...
##  $ type        : chr  "Fixed wing multi engine" "Fixed wing multi engine" "Fixed wing multi engine"
##  $ manufacturer: chr  "EMBRAER" "AIRBUS INDUSTRIE" "AIRBUS INDUSTRIE" "AIRBUS INDUSTRIE" ...
##  $ model       : chr  "EMB-145XR" "A320-214" "A320-214" "A320-214" ...
##  $ engines     : int  2 2 2 2 2 2 2 2 2 2 ...
##  $ seats       : int  55 182 182 182 55 182 182 182 182 182 ...
##  $ speed       : int  NA NA NA NA NA NA NA NA NA NA ...
##  $ engine      : chr  "Turbo-fan" "Turbo-fan" "Turbo-fan" "Turbo-fan" ...
```

```r
str(weather)
```

```
## Classes 'tbl_df', 'tbl' and 'data.frame':    26130 obs. of  15 variables:
##  $ origin    : chr  "EWR" "EWR" "EWR" "EWR" ...
##  $ year      : num  2013 2013 2013 2013 2013 ...
##  $ month     : num  1 1 1 1 1 1 1 1 1 1 ...
##  $ day       : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ hour      : int  0 1 2 3 4 6 7 8 9 10 ...
##  $ temp      : num  37 37 37.9 37.9 37.9 ...
##  $ dewp      : num  21.9 21.9 21.9 23 24.1 ...
##  $ humid     : num  54 54 52.1 54.5 57 ...
##  $ wind_dir  : num  230 230 230 230 240 270 250 240 250 260 ...
##  $ wind_speed: num  10.4 13.8 12.7 13.8 15 ...
##  $ wind_gust : num  11.9 15.9 14.6 15.9 17.2 ...
##  $ precip    : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ pressure  : num  1014 1013 1013 1013 1013 ...
##  $ visib     : num  10 10 10 10 10 10 10 10 10 10 ...
##  $ time_hour : POSIXct, format: "2012-12-31 16:00:00" "2012-12-31 17:00:00" ...
```

The internal structure of flights shows that Month and day are collected as integer values. Thus, they will need to be converted to the actual name values for any analysis that has months or days to make any sense. Also, the arr_delay which identifies the arrival delay is stored as a numeric object, which means that the data contains a precision value of the reporting of this variable.

```r
# head and tail results

head(flights, 5)
```

```
## # A tibble: 5 × 19
##    year month   day dep_time sched_dep_time dep_delay arr_time
##   <int> <int> <int>    <int>          <int>     <dbl>    <int>
## 1  2013     1     1      517            515         2      830
## 2  2013     1     1      533            529         4      850
## 3  2013     1     1      542            540         2      923
## 4  2013     1     1      544            545        -1     1004
## 5  2013     1     1      554            600        -6      812
## # ... with 12 more variables: sched_arr_time <int>, 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>
```

```r
tail(flights, 5)
```

```
## # A tibble: 5 × 19
##    year month   day dep_time sched_dep_time dep_delay arr_time
##   <int> <int> <int>    <int>          <int>     <dbl>    <int>
## 1  2013     9    30       NA           1455        NA       NA
## 2  2013     9    30       NA           2200        NA       NA
## 3  2013     9    30       NA           1210        NA       NA
## 4  2013     9    30       NA           1159        NA       NA
## 5  2013     9    30       NA            840        NA       NA
## # ... with 12 more variables: sched_arr_time <int>, 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>
```

```
head(airports, 5)
```

```
## # A tibble: 5 × 7
##     faa                          name      lat       lon   alt    tz   dst
##   <chr>                         <chr>    <dbl>     <dbl> <int> <dbl> <chr>
## 1   04G             Lansdowne Airport 41.13047 -80.61958  1044    -5     A
## 2   06A Moton Field Municipal Airport 32.46057 -85.68003   264    -5     A
## 3   06C           Schaumburg Regional 41.98934 -88.10124   801    -6     A
## 4   06N               Randall Airport 41.43191 -74.39156   523    -5     A
## 5   09J         Jekyll Island Airport 31.07447 -81.42778    11    -4     A
```

```
tail(airports, 5)
```

```
## # A tibble: 5 × 7
##     faa                       name      lat        lon   alt    tz   dst
##   <chr>                      <chr>    <dbl>      <dbl> <int> <dbl> <chr>
## 1   ZUN                Black Rock 35.08323 -108.79178  6454    -7     A
## 2   ZVE    New Haven Rail Station 41.29867  -72.92599     7    -5     A
## 3   ZWI Wilmington Amtrak Station 39.73667  -75.55167     0    -5     A
## 4   ZWU  Washington Union Station 38.89746  -77.00643    76    -5     A
## 5   ZYP              Penn Station 40.75050  -73.99350    35    -5     A
```

```
head(airlines, 5)
```

```
## # A tibble: 5 × 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.
```

```
tail(airlines, 5)
```

```
## # A tibble: 5 × 2
##   carrier                  name
##     <chr>                 <chr>
## 1      UA  United Air Lines Inc.
## 2      US        US Airways Inc.
## 3      VX         Virgin America
## 4      WN Southwest Airlines Co.
## 5      YV     Mesa Airlines Inc.
```

```
head(planes, 5)
```

```
## # A tibble: 5 × 9
##   tailnum  year                    type     manufacturer     model engines
##     <chr> <int>                   <chr>            <chr>     <chr>   <int>
## 1  N10156  2004 Fixed wing multi engine          EMBRAER EMB-145XR       2
## 2  N102UW  1998 Fixed wing multi engine AIRBUS INDUSTRIE  A320-214       2
## 3  N103US  1999 Fixed wing multi engine AIRBUS INDUSTRIE  A320-214       2
## 4  N104UW  1999 Fixed wing multi engine AIRBUS INDUSTRIE  A320-214       2
## 5  N10575  2002 Fixed wing multi engine          EMBRAER EMB-145LR       2
## # ... with 3 more variables: seats <int>, speed <int>, engine <chr>
```

```
tail(planes, 5)
```

```
## # A tibble: 5 × 9
##   tailnum  year                    type                  manufacturer
##     <chr> <int>                   <chr>                         <chr>
## 1  N997AT  2002 Fixed wing multi engine                        BOEING
## 2  N997DL  1992 Fixed wing multi engine MCDONNELL DOUGLAS AIRCRAFT CO
## 3  N998AT  2002 Fixed wing multi engine                        BOEING
## 4  N998DL  1992 Fixed wing multi engine MCDONNELL DOUGLAS CORPORATION
## 5  N999DN  1992 Fixed wing multi engine MCDONNELL DOUGLAS CORPORATION
## # ... with 5 more variables: model <chr>, engines <int>, seats <int>,
## #   speed <int>, engine <chr>
```

```
head(weather, 5)
```

```
## # A tibble: 5 × 15
##   origin  year month   day  hour  temp  dewp humid wind_dir wind_speed
##    <chr> <dbl> <dbl> <int> <int> <dbl> <dbl> <dbl>    <dbl>      <dbl>
## 1    EWR  2013     1     1     0 37.04 21.92 53.97      230   10.35702
## 2    EWR  2013     1     1     1 37.04 21.92 53.97      230   13.80936
## 3    EWR  2013     1     1     2 37.94 21.92 52.09      230   12.65858
## 4    EWR  2013     1     1     3 37.94 23.00 54.51      230   13.80936
## 5    EWR  2013     1     1     4 37.94 24.08 57.04      240   14.96014
## # ... with 5 more variables: wind_gust <dbl>, precip <dbl>,
## #   pressure <dbl>, visib <dbl>, time_hour <dttm>
```

```
tail(weather, 5)
```

```
## # A tibble: 5 × 15
##   origin  year month   day  hour  temp  dewp humid wind_dir wind_speed
##    <chr> <dbl> <dbl> <int> <int> <dbl> <dbl> <dbl>    <dbl>      <dbl>
## 1    LGA  2013    12    30    19 35.96 19.94 51.78      340   13.80936
## 2    LGA  2013    12    30    20 33.98 17.06 49.51      330   17.26170
## 3    LGA  2013    12    30    21 32.00 15.08 49.19      340   14.96014
## 4    LGA  2013    12    30    22 30.92 12.92 46.74      320   17.26170
## 5    LGA  2013    12    30    23 28.94 10.94 46.41      330   18.41248
## # ... with 5 more variables: wind_gust <dbl>, precip <dbl>,
## #   pressure <dbl>, visib <dbl>, time_hour <dttm>
```

Retrieving the first and last records of flights data show how the data is reported out in addition to the related data type. Interestingly, the tail shows data for the month of September. I was expecting to see data

all the way to December instead. Also, most of the tail results for departure and arrival time were all NA. This tells me as part of data cleaning and analysis, It may be important to eliminate data for the month of September if the analysis requires departure and arrival times.

```
# missing Data results

sapply(flights, function(x) sum(is.na(x)))
```

```
##           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
```

```
sapply(planes, function(x) sum(is.na(x)))
```

```
##      tailnum         year         type manufacturer        model
##            0           70            0            0            0
##      engines        seats        speed       engine
##            0            0         3299            0
```

```
sapply(airports, function(x) sum(is.na(x)))
```

```
##  faa name  lat  lon  alt   tz  dst
##    0    0    0    0    0    0    0
```

```
sapply(airlines, function(x) sum(is.na(x)))
```

```
## carrier    name
##       0       0
```

```
sapply(weather, function(x) sum(is.na(x)))
```

```
##     origin       year      month        day       hour       temp
##          0          0          0          0          0          1
##       dewp      humid   wind_dir wind_speed  wind_gust     precip
##          1          1        418          3          3          0
##   pressure      visib  time_hour
##       2730          0          0
```

For flights, it shows that 8255 departure flights have an NA value and 9430 arrival delay flights also have an NA value. Considering that there were 336776 flights out of NY during that year, the difference between the missing values is then very negligible. Thus, the missing data will not impact any analysis done on either the departure or arrival delay time.

(b) Formulating Questions:

Consider the NYC flights data. Formulate two motivating questions you want to explore using this data. Describe why these questions are interesting and how you might go about answering them.

Response to Formulating Questions:

Arrival delays of flights can be a nightmare for love ones excited to see their friends or family members arriving safely. And we saw in the past couple years multiple issues with flights delays that cause significant anxiety. To that end, my interest is to analyze the data reported when it comes to arrival delays. My questions are as follow: Question 1: How do flights towards East and West coast differ in arrival delays? Question 2: How does each of the three origin airports in New York fair when it comes to arrival delays?

In question 1, I attempt to understand if there are any significant differences between arrival delays of flights flying to either a city on the West coast or the East coast. To run this analysis, I decided to analyze MIA (Miami) for East Coast data and SEA (Seattle) for West Coast data.

In question 2, I attempt to examine how each of the origin airports compares as far as the arrival delays. To do this, I will generate a density report for total flights out each origin airports and as well a density arrival delays from the same originated airports. I will then use the data to find out which airport or airports have the least amount of arrival delays.

(c) Exploring Data:

For each of the questions you proposed in Problem 1b, perform an exploratory data analysis designed to address the question. At a minimum, you should produce two visualizations related to each question. Be sure to describe what the visuals show and how they speak to your question of interest.

Response to Exploring Data for question 1:

First, I explore the flights arrival delays between SEA and MIA in hours

```
ggplot(subset(flights, dest %in% c("SEA", "MIA")),
       aes(x=hour,
           y=arr_delay,
           color=dest))+
     geom_point() + ggtitle("Arrival delays comparison between SEA and MIA") +
     labs(x="Hour",y="Arrival delays")
```

Second, I explore the flights arrival delays between SEA and MIA on a monthly basis

```
 ggplot(subset(flights, dest %in% c("SEA", "MIA")),
        aes(x=month.abb[month],
            y=arr_delay,
           color=dest))+
     geom_point(size=2, alpha=0.5) +
     ggtitle("Arrival delays comparison between SEA and MIA by month") +
     labs(x="Month",y="Arrival delays")
```

It shows that the East Coast city has more delays and has the highest peak of the arrival of delays as well. When looking at the hourly data, it shows that the flights bound for Miami tend to have a lot more arrival delays than flights that are bound to Seattle. However, when looking at the monthly data, it appears that the arrival delays are very much in sync for both coasts. But in both cases, whether it's for the hourly or monthly data, there were numerous outliers of arrival delays for flights going toward Miami than Seattle. Also, the hourly data shows that between the 10th and 15th hours of the day, the East part of the country is primarily heavy with arrival delays.

This simple comparison between Miami and Seattle require additional information to understand why there are such a discrepancies in the hourly delays. Is it possible that time zone differential may have caused that flights bound for the east coast are affected by other? Or is this a matter of volume of flights that are

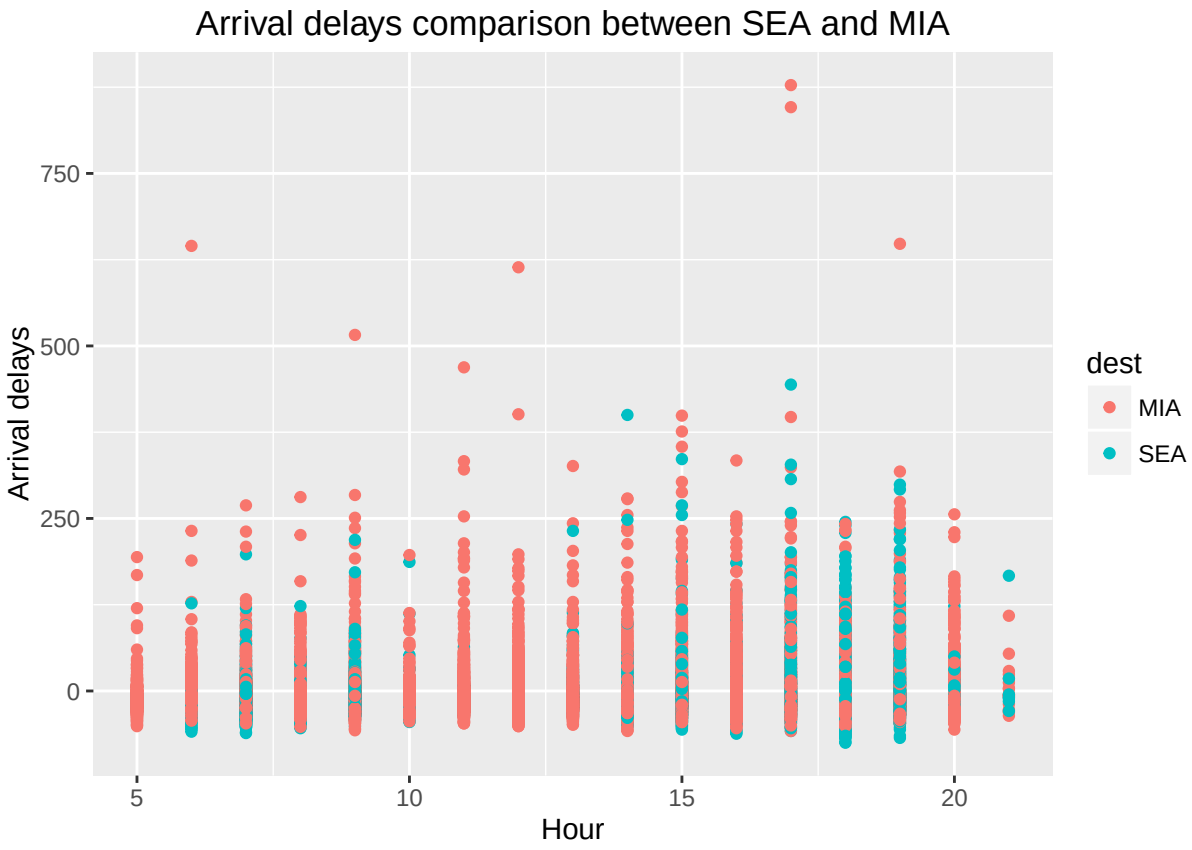

Figure 1: Hours Arrival delays comparison between SEA and MIA

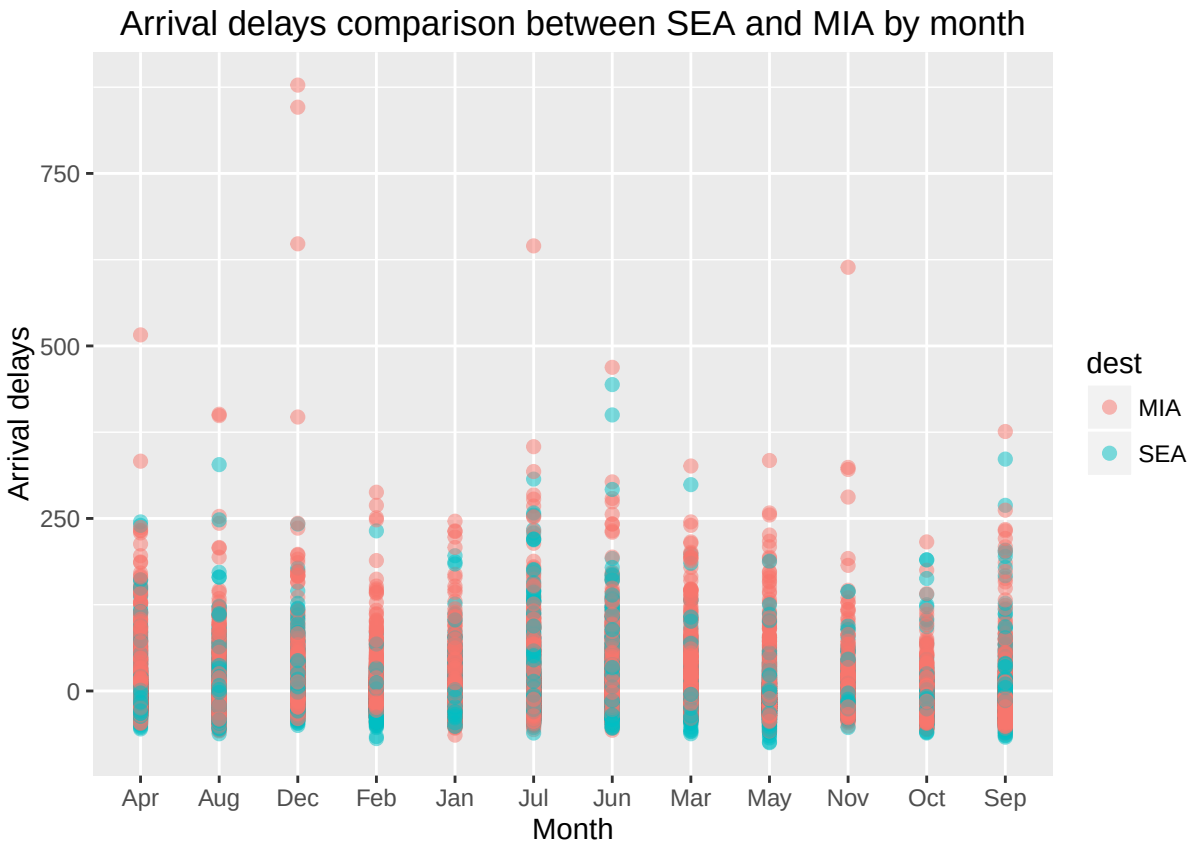

Figure 2: Monthly Arrival delays comparison between SEA and MIA

consistently flying toward the East Coast as opposed to the West Coast. But either the reason for delays, it appears that at the end both coasts are pretty close as far as the number of delays when comparing the data monthly

Response to Exploring Data for question 2:

First, I explore the densify of flights departure from the three originated airports in NY

```
ggplot(flights) +
    geom_bar(aes(origin)) +
    ggtitle("Density of flights originated from NY") +
    labs(x="flight origin",y="total flights")
```

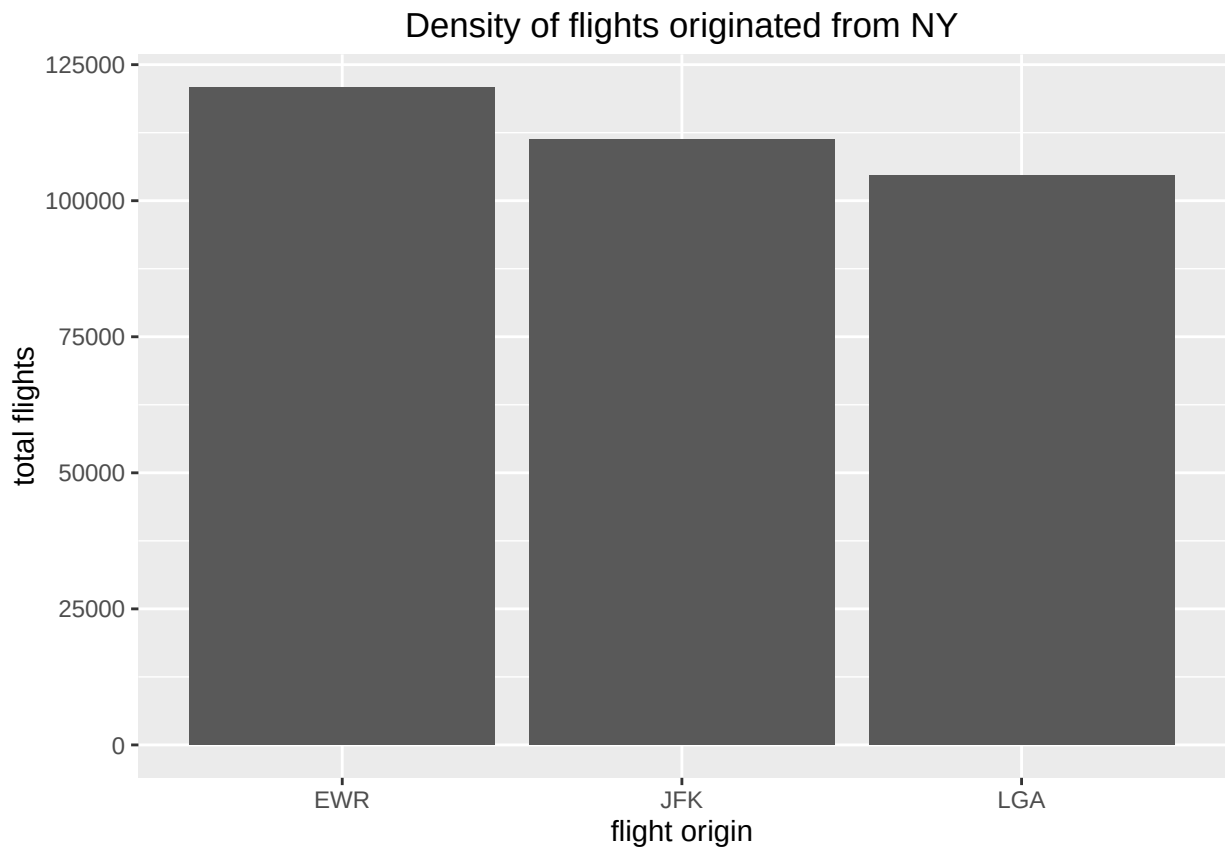

Figure 3: density of flights coming out of NY

Second, I explore the density of arrival delays from the three originated airports in NY

```
ggplot(flights, aes(origin, arr_delay) ) +
    geom_line() +
     ggtitle("Density of delayed flights from NY - line graph") +
     labs(x="Flights Origin",y="Arrival Delays")
```

Figure 3, shows that EWR (Newark) airport has the highest number of flights out of New York followed by JFK and LGA(La Guardia). That said, figure 4 shows that JFK had the largest number of arrival

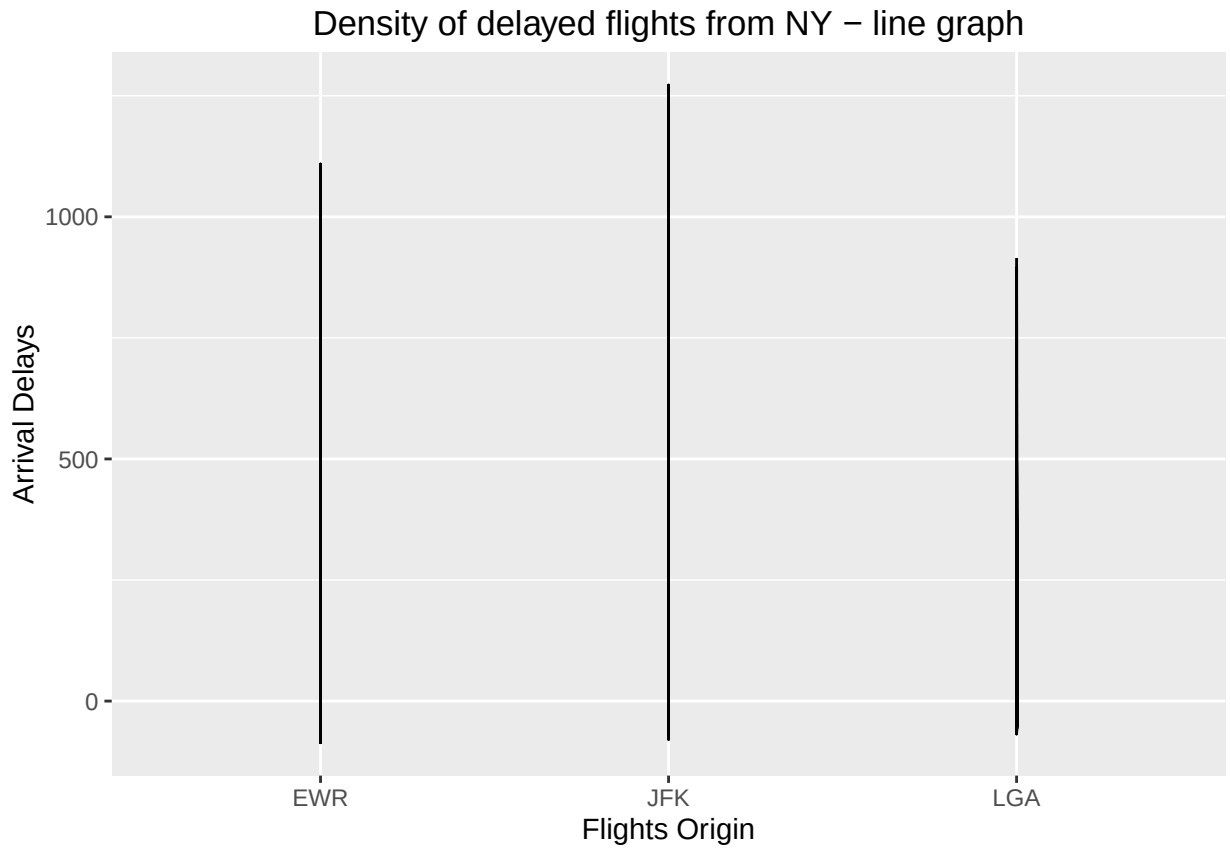

Figure 4: density of delayed flights coming out of NY line chart

delays among all three airports. It's interesting though that even though EWR handled more traffic to other destinations, it came second as far as the arrival delay to a target location. Does EWR have the infrastructure to handle heavy traffic as far as on time departure? To that end, it is important to note that additional logistical data will need to be investigated to understand the site(infrastructure, support workforce, flights' maintenance crew skills, etc) of each airport to formulate a concrete explanation on why some of these airports have more arrival delays than other.

One interesting note that came about from the visualization report is that flying out from La Guardia will be the best option to avoid arrival delay. Since I live on the West Coast, I would need to perform additional analysis on whether this is true for flights which are targeted to land in the West Coast.

(d) Challenge Your Results:

After completing the exploratory analysis from Problem 1c, do you have any concerns about your findings? Comment on any ethical and/or privacy concerns you have with your analysis.

Response to challenge results

Data Concerns

Although La Guardia seems to be best airport to depart from, it will be important to analyze the number of flights that are scheduled to leave each airports during peak seasonal period. For instance, we will need to know the volume of flights that were affected during snow storm. We would also need to analyze data for the same destination for each of the origin airports for the same period of time to have a clear picture on how well each airport performs. Furthermore, it would be important to perform a deeper analysis on the data and correlate whether the flights delay is due to security concerns, technical system, availability of gates to park the planes, average pilot's experience when comparing each airport, number of flights bound for either the East or West Coast from each of these airports.

Privacy and ethical concerns

Origin's airport like La Guardia could use the fact that it has the least number of delays to lure advertisers to post ads at this airport. It can also use this data to discredit the other originated airports and as well as forcing an unfair competition with the other airports. It could potentially lure Commercial airplanes to have their home-based in that airport as opposed the other two.

Airlines companies can even use the arrival delay data as a bargaining chip to grant or deny pilots a promotion, and use it as a leverage to force the pilot union to accept a deal that otherwise they would not have accepted.

Further analysis

We have seen in the last couple years' major flights delay that are caused by technical issues such as computer system being down, flights data are retrieved through ancient technology have limited tech support. Unfortunately, the data did not have any such information. It would have been a good exercise to go beyond the typical weather issue, or flight's maintenance issues to also analyze from an end-to-end where the bottleneck happened that caused arrival delays. I would have liked to get data regarding turn around time to onboard a traveler, to security check delay, flight departure delay, weather situations, gates availability to park the airplane on arrival.