final_project_draft

Omer Shasman

5/3/2022

Introduction

On Time Performance analysis of an airline network - This is an important metric for the airline that is calculated as the percentage of flights which are delayed by more than 14 minutes while the aircraft arrives at the gate. There are multiple reasons which contribute to the variation in OTP. An analysis of the OTP metric breaking it down into its individual components namely different delay and historical delays can provide insights into how the OTP for an airline can be managed by operational/process changes. The Department of Transport releases the flight level, On Time Performance data. This dataset also has various other factors which affect the Arrival Delay of a flight. An exploratory analysis of this data with the Arrival Delay as the response variable analyzed against different dimensions provided in the dataset can reveal several insights to improve the OTP of an Airline.

What

As part of my Final Project, I am planning to use a subset of OTP data to perform analysis of delays on actual file arrivals focusing on one particular Station and Airline. Since airline operations are very complex, the arrival delays itself can be due to varying factors, like weather delay, carrier delays, security delays, Late aircraft delay. . . etc or any combinations of any of these in general. My focus is only on 2 types of delays so that I can minimize the complexities in data structures and limit any repeating processes or steps, and rather focus on how to manipulate and do analysis/inference with few variables. Hence I will be considering only 5 years data ranging from year 2014 till 2019 two types of delays "Weather Delays" and "Carrier Delays"

Why

I thought airline is an interesting business with lot of complex operation/data and business itself is most of us are familiar with. Also, with the time constraint we have, there were few sites like Kaggle and DOT On-Time_performance

This data is presented as yearly file in csv format, I have to merge different years data using dply bind command to append rows at the end and build one file.

How

Use RMarkdown and explore Rfunctions that can integrate some of the topics we learned in the class for flight arrival delay analysis. The following are steps which will be followed as part of the project

- Load data into R Markdown using R chunks
- Merge Data using ddplyr

- Filter/melt/massage data using Tidy Data approach
- Use sampling strategy for identifying sample observations.
- Use Stats function to determine mean, median, IQR,...
- Regression - Find any co-relation between total flights arriving at a particualr airport and delays to identify if it's the airport operational/capacity issue or not.
- Could hosted data and R Markdown interface. Pull data from AWS S3 buckets than loading from local machine.

Body

This project perform analysis of flight arrival delays focusing on one particular Station and Airline. Since airline delays are unavoidable there is always a chance that a particualr flight will be delayed. I think this analysis can be used to further study on why a particular delay happens and if the process/schedule/operations can be enhanced or refined to minimize the delay risk in future flights.

Packages Required

```
library(knitr) ##for printing tables in R Markdown
library(dplyr) ##for data munging
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(ggplot2) ## for charts
library(infer) ## for rep_sample_n used for clustered sampling
```

```
library(readr)
flight.data.y2014 <- read_csv("Data/2014.csv")
```

```
## Rows: 5819811 Columns: 28
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr   (4): OP_CARRIER, ORIGIN, DEST, CANCELLATION_CODE
## dbl  (22): OP_CARRIER_FL_NUM, CRS_DEP_TIME, DEP_TIME, DEP_DELAY, TAXI_OUT, W...
## lgl   (1): Unnamed: 27
## date  (1): FL_DATE
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```r
# head(flight.data.y2014)

flight.data.y2015 <- read_csv("Data/2015.csv")
```

```
## Rows: 5819079 Columns: 28
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr   (4): OP_CARRIER, ORIGIN, DEST, CANCELLATION_CODE
## dbl  (22): OP_CARRIER_FL_NUM, CRS_DEP_TIME, DEP_TIME, DEP_DELAY, TAXI_OUT, W...
## lgl   (1): Unnamed: 27
## date  (1): FL_DATE
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```r
# head(flight.data.y2015)

flight.data.y2016 <- read_csv("Data/2016.csv")
```

```
## Rows: 5617658 Columns: 28
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr   (4): OP_CARRIER, ORIGIN, DEST, CANCELLATION_CODE
## dbl  (22): OP_CARRIER_FL_NUM, CRS_DEP_TIME, DEP_TIME, DEP_DELAY, TAXI_OUT, W...
## lgl   (1): Unnamed: 27
## date  (1): FL_DATE
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```r
# head(flight.data.y2016)

flight.data.y2017 <- read_csv("Data/2017.csv")
```

```
## Rows: 5674621 Columns: 28
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr   (4): OP_CARRIER, ORIGIN, DEST, CANCELLATION_CODE
## dbl  (22): OP_CARRIER_FL_NUM, CRS_DEP_TIME, DEP_TIME, DEP_DELAY, TAXI_OUT, W...
## lgl   (1): Unnamed: 27
## date  (1): FL_DATE
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

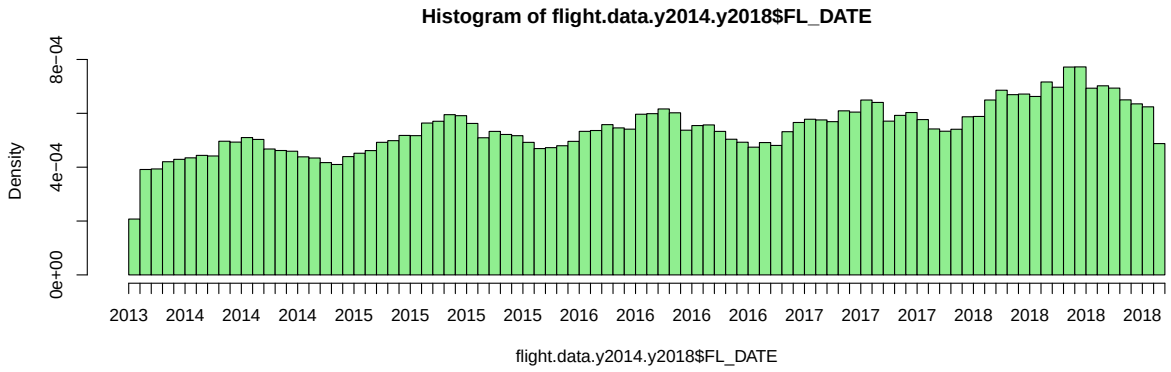
```r
# head(flight.data.y2017)

flight.data.y2018 <- read_csv("Data/2018.csv")
```

```
## Rows: 7213446 Columns: 28
## -- Column specification --------------------------------------------------------
```

```
## Delimiter: ","
## chr   (4): OP_CARRIER, ORIGIN, DEST, CANCELLATION_CODE
## dbl  (22): OP_CARRIER_FL_NUM, CRS_DEP_TIME, DEP_TIME, DEP_DELAY, TAXI_OUT, W...
## lgl   (1): Unnamed: 27
## date  (1): FL_DATE
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
# head(flight.data.y2018)

# Since this is a large dataset, sampling/manipualting on all the observations is throwing memory error
flight.data.y2014 <- flight.data.y2014[flight.data.y2014$DEST %in% 'MSP', ]
flight.data.y2015 <- flight.data.y2015[flight.data.y2015$DEST %in% 'MSP', ]
flight.data.y2016 <- flight.data.y2016[flight.data.y2016$DEST %in% 'MSP', ]
flight.data.y2017 <- flight.data.y2017[flight.data.y2017$DEST %in% 'MSP', ]
flight.data.y2018 <- flight.data.y2018[flight.data.y2018$DEST %in% 'MSP', ]
```

```
# Combine the 5 vectors to a single vector
flight.data.y2014.y2018 <- dplyr::bind_rows(flight.data.y2014,flight.data.y2015,flight.data.y2016,flight

# replace all na in the fields we interested in to 0
flight.data.y2014.y2018$ARR_DELAY <- flight.data.y2014.y2018$ARR_DELAY %>% replace(is.na(.), 0)

flight.data.y2014.y2018$LATE_AIRCRAFT_DELAY <- flight.data.y2014.y2018$LATE_AIRCRAFT_DELAY %>% replace(
flight.data.y2014.y2018$SECURITY_DELAY <- flight.data.y2014.y2018$SECURITY_DELAY %>% replace(is.na(.), 0
flight.data.y2014.y2018$WEATHER_DELAY <- flight.data.y2014.y2018$WEATHER_DELAY %>% replace(is.na(.), 0)
flight.data.y2014.y2018$CARRIER_DELAY <- flight.data.y2014.y2018$CARRIER_DELAY %>% replace(is.na(.), 0)
hist(flight.data.y2014.y2018$FL_DATE, flight.data.y2014.y2018$ARR_DELAY, breaks = 100, col = "lightgreen
```




```
summary(flight.data.y2014.y2018)
```

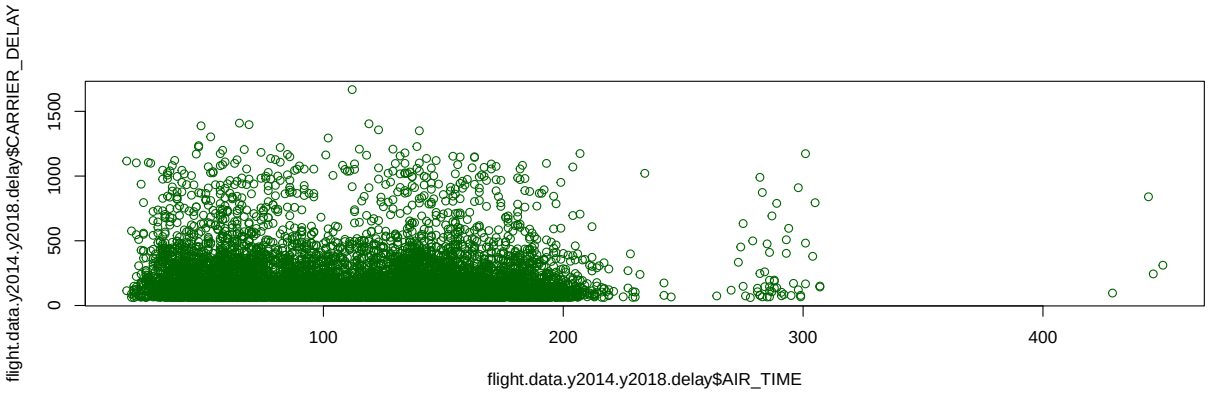
```
##     FL_DATE            OP_CARRIER        OP_CARRIER_FL_NUM    ORIGIN
##  Min.   :2014-01-01   Length:652037      Min.   :   2      Length:652037
##  1st Qu.:2015-06-25   Class :character   1st Qu.:1154      Class :character
##  Median :2016-09-29   Mode  :character   Median :2039      Mode  :character
##  Mean   :2016-09-08                      Mean   :2596
##  3rd Qu.:2017-12-20                      3rd Qu.:4486
```

```
##  Max.   :2018-12-31                    Max.   :7439
##
##      DEST            CRS_DEP_TIME     DEP_TIME      DEP_DELAY
##  Length:652037      Min.   :   1   Min.   :   1   Min.   :-204.000
##  Class :character   1st Qu.: 810   1st Qu.: 815   1st Qu.:  -6.000
##  Mode  :character   Median :1230   Median :1232   Median :  -3.000
##                     Mean   :1225   Mean   :1231   Mean   :   7.931
##                     3rd Qu.:1610   3rd Qu.:1621   3rd Qu.:   3.000
##                     Max.   :2359   Max.   :2400   Max.   :1676.000
##                                    NA's   :5421   NA's   :5614
##     TAXI_OUT        WHEELS_OFF     WHEELS_ON       TAXI_IN
##  Min.   :  1.00   Min.   :   1   Min.   :   1   Min.   :  1.000
##  1st Qu.: 12.00   1st Qu.: 832   1st Qu.:1015   1st Qu.:  4.000
##  Median : 15.00   Median :1247   Median :1416   Median :  5.000
##  Mean   : 17.32   Mean   :1257   Mean   :1415   Mean   :  5.799
##  3rd Qu.: 20.00   3rd Qu.:1637   3rd Qu.:1824   3rd Qu.:  7.000
##  Max.   :166.00   Max.   :2400   Max.   :2400   Max.   :168.000
##  NA's   :5597     NA's   :5597   NA's   :5725   NA's   :5725
##   CRS_ARR_TIME     ARR_TIME      ARR_DELAY          CANCELLED
##  Min.   :   1   Min.   :   1   Min.   :-119.000   Min.   :0.000000
##  1st Qu.:1030   1st Qu.:1020   1st Qu.: -16.000   1st Qu.:0.000000
##  Median :1428   Median :1420   Median :  -7.000   Median :0.000000
##  Mean   :1433   Mean   :1419   Mean   :   1.781   Mean   :0.008625
##  3rd Qu.:1835   3rd Qu.:1829   3rd Qu.:   4.000   3rd Qu.:0.000000
##  Max.   :2359   Max.   :2400   Max.   :1668.000   Max.   :1.000000
##                 NA's   :5725
##  CANCELLATION_CODE     DIVERTED        CRS_ELAPSED_TIME ACTUAL_ELAPSED_TIME
##  Length:652037      Min.   :0.000000   Min.   : 22.0    Min.   : 28.0
##  Class :character   1st Qu.:0.000000   1st Qu.: 91.0    1st Qu.: 88.0
##  Mode  :character   Median :0.000000   Median :141.0    Median :136.0
##                     Mean   :0.001708   Mean   :140.8    Mean   :134.8
##                     3rd Qu.:0.000000   3rd Qu.:184.0    3rd Qu.:176.0
##                     Max.   :1.000000   Max.   :489.0    Max.   :509.0
##                                                         NA's   :6738
##     AIR_TIME        DISTANCE      CARRIER_DELAY      WEATHER_DELAY
##  Min.   : 12.0   Min.   :  76.0   Min.   :   0.000   Min.   :   0.0000
##  1st Qu.: 64.0   1st Qu.: 386.0   1st Qu.:   0.000   1st Qu.:   0.0000
##  Median :112.0   Median : 852.0   Median :   0.000   Median :   0.0000
##  Mean   :111.7   Mean   : 800.8   Mean   :   3.806   Mean   :   0.6441
##  3rd Qu.:152.0   3rd Qu.:1050.0   3rd Qu.:   0.000   3rd Qu.:   0.0000
##  Max.   :478.0   Max.   :3972.0   Max.   :1668.000   Max.   :1308.0000
##  NA's   :6738
##    NAS_DELAY      SECURITY_DELAY     LATE_AIRCRAFT_DELAY Unnamed: 27
##  Min.   :   0.0   Min.   :  0.0000   Min.   :   0.000    Mode:logical
##  1st Qu.:   0.0   1st Qu.:  0.0000   1st Qu.:   0.000    NA's:652037
##  Median :   2.0   Median :  0.0000   Median :   0.000
##  Mean   :  12.7   Mean   :  0.0134   Mean   :   3.762
##  3rd Qu.:  18.0   3rd Qu.:  0.0000   3rd Qu.:   0.000
##  Max.   :1238.0   Max.   :593.0000   Max.   :1289.000
##  NA's   :553520
```

```
flight.data.y2014.y2018.delay <- flight.data.y2014.y2018[flight.data.y2014.y2018$CARRIER_DELAY > 60, ]
summary(flight.data.y2014.y2018.delay)
```

```
##     FL_DATE            OP_CARRIER        OP_CARRIER_FL_NUM    ORIGIN
##  Min.   :2014-01-01   Length:9608        Min.   :   8      Length:9608
##  1st Qu.:2015-07-02   Class :character   1st Qu.:1280      Class :character
##  Median :2016-09-16   Mode  :character   Median :2172      Mode  :character
##  Mean   :2016-08-29                      Mean   :2784
##  3rd Qu.:2017-12-03                      3rd Qu.:4580
##  Max.   :2018-12-31                      Max.   :7439
##     DEST            CRS_DEP_TIME     DEP_TIME      DEP_DELAY
##  Length:9608        Min.   :   5   Min.   :   1   Min.   :   3.0
##  Class :character   1st Qu.: 714   1st Qu.: 958   1st Qu.:  95.0
##  Mode  :character   Median :1105   Median :1341   Median : 139.0
##                     Mean   :1166   Mean   :1391   Mean   : 207.7
##                     3rd Qu.:1606   3rd Qu.:1843   3rd Qu.: 240.0
##                     Max.   :2359   Max.   :2400   Max.   :1676.0
##     TAXI_OUT        WHEELS_OFF     WHEELS_ON       TAXI_IN        CRS_ARR_TIME
##  Min.   :  2.00   Min.   :   1   Min.   :   1   Min.   : 1.000   Min.   :   1
##  1st Qu.: 12.00   1st Qu.:1012   1st Qu.:1039   1st Qu.: 4.000   1st Qu.: 912
##  Median : 15.00   Median :1350   Median :1429   Median : 5.000   Median :1332
##  Mean   : 18.95   Mean   :1402   Mean   :1402   Mean   : 5.778   Mean   :1360
##  3rd Qu.: 22.00   3rd Qu.:1851   3rd Qu.:1923   3rd Qu.: 7.000   3rd Qu.:1833
##  Max.   :132.00   Max.   :2400   Max.   :2400   Max.   :82.000   Max.   :2359
##     ARR_TIME      ARR_DELAY        CANCELLED CANCELLATION_CODE     DIVERTED
##  Min.   :   1   Min.   :  61.0   Min.   :0   Length:9608        Min.   :0
##  1st Qu.:1041   1st Qu.:  89.0   1st Qu.:0   Class :character   1st Qu.:0
##  Median :1429   Median : 135.0   Median :0   Mode  :character   Median :0
##  Mean   :1399   Mean   : 201.4   Mean   :0                      Mean   :0
##  3rd Qu.:1924   3rd Qu.: 234.0   3rd Qu.:0                      3rd Qu.:0
##  Max.   :2400   Max.   :1668.0   Max.   :0                      Max.   :0
##  CRS_ELAPSED_TIME ACTUAL_ELAPSED_TIME    AIR_TIME        DISTANCE
##  Min.   : 42.0    Min.   : 30.0       Min.   : 18.0   Min.   :  76.0
##  1st Qu.: 94.0    1st Qu.: 90.0       1st Qu.: 65.0   1st Qu.: 386.0
##  Median :143.0    Median :138.0       Median :113.0   Median : 852.0
##  Mean   :142.1    Mean   :135.8       Mean   :111.1   Mean   : 795.5
##  3rd Qu.:183.0    3rd Qu.:174.2       3rd Qu.:149.0   3rd Qu.:1034.0
##  Max.   :473.0    Max.   :494.0       Max.   :450.0   Max.   :3972.0
##  CARRIER_DELAY    WEATHER_DELAY        NAS_DELAY       SECURITY_DELAY
##  Min.   :  61.0   Min.   :  0.0000   Min.   :   0.00   Min.   : 0.00000
##  1st Qu.:  81.0   1st Qu.:  0.0000   1st Qu.:   0.00   1st Qu.: 0.00000
##  Median : 120.0   Median :  0.0000   Median :   0.00   Median : 0.00000
##  Mean   : 189.4   Mean   :  0.5843   Mean   :   4.02   Mean   : 0.00999
##  3rd Qu.: 218.0   3rd Qu.:  0.0000   3rd Qu.:   1.00   3rd Qu.: 0.00000
##  Max.   :1668.0   Max.   :689.0000   Max.   :1023.00   Max.   :79.00000
##  LATE_AIRCRAFT_DELAY Unnamed: 27
##  Min.   :  0.000     Mode:logical
##  1st Qu.:  0.000     NA's:9608
##  Median :  0.000
##  Mean   :  7.418
##  3rd Qu.:  0.000
##  Max.   :928.000
```

```
plot(flight.data.y2014.y2018.delay$AIR_TIME, flight.data.y2014.y2018.delay$CARRIER_DELAY, col = "darkgr
```



Topics From Class

Topic 1:

R Markdown - I will be presenting the project in R Markdown and knit the file to a pdf document. Will be using R chunks to demonstrate and build the project components.

Topic 2:

GitHub - Will host the project in github repository for others to view my project components.

Topic 3:

Sampling strategies for an Observational study - Will be using sampling strategies - Simple random sampling, Strtified sampling, Cluster sampling and multistage sampling to group the data together by using different variables from the dataset and then use one of the sampling result to build topic#4 and 5.

```
simple.sampling <- dplyr::sample_n(flight.data.y2014.y2018, 1000, replace=FALSE)
# View(simple.sampling)
simple.sampling
```

```
## # A tibble: 1,000 x 28
##    FL_DATE    OP_CARRIER OP_CARRIER_FL_NUM ORIGIN DEST  CRS_DEP_TIME DEP_TIME
##    <date>     <chr>                  <dbl> <chr>  <chr>        <dbl>    <dbl>
##  1 2018-03-18 9E                      3310 DFW    MSP           1200     1211
##  2 2016-01-19 OO                      4572 MKE    MSP           1207     1228
##  3 2016-08-14 DL                      1864 SAN    MSP            620      622
##  4 2015-04-19 DL                      1361 BWI    MSP           1908     1921
##  5 2017-06-17 WN                      3421 PHX    MSP            605      603
##  6 2016-05-17 AA                      2578 ORD    MSP           1210     1234
##  7 2018-01-21 DL                      1438 LAX    MSP           1515     1507
##  8 2015-09-13 DL                      1631 LGA    MSP           1849     1846
##  9 2014-02-03 AA                      1077 ORD    MSP           2025     2101
## 10 2015-11-04 OO                      7371 BJI    MSP           1253     1248
## # ... with 990 more rows, and 21 more variables: DEP_DELAY <dbl>,
## #   TAXI_OUT <dbl>, WHEELS_OFF <dbl>, WHEELS_ON <dbl>, TAXI_IN <dbl>,
```

```
## #   CRS_ARR_TIME <dbl>, ARR_TIME <dbl>, ARR_DELAY <dbl>, CANCELLED <dbl>,
## #   CANCELLATION_CODE <chr>, DIVERTED <dbl>, CRS_ELAPSED_TIME <dbl>,
## #   ACTUAL_ELAPSED_TIME <dbl>, AIR_TIME <dbl>, DISTANCE <dbl>,
## #   CARRIER_DELAY <dbl>, WEATHER_DELAY <dbl>, NAS_DELAY <dbl>,
## #   SECURITY_DELAY <dbl>, LATE_AIRCRAFT_DELAY <dbl>, `Unnamed: 27` <lgl>
```

```
# Here I am making a cluster of where Airline code is the strata
DL <- flight.data.y2014.y2018[flight.data.y2014.y2018$OP_CARRIER %in% 'DL', ]
UA <- flight.data.y2014.y2018[flight.data.y2014.y2018$OP_CARRIER %in% 'UA',]
AA <- flight.data.y2014.y2018[flight.data.y2014.y2018$OP_CARRIER %in% 'AA',]
WN <- flight.data.y2014.y2018[flight.data.y2014.y2018$OP_CARRIER %in% 'WN',]

stratified.sampling <- dplyr::sample_n((UA), 1000, replace=FALSE)
dim(stratified.sampling)
```

```
## [1] 1000   28
```

```
#randomly choose 4 10 groups out of the n
clusters <- sample(unique(flight.data.y2014.y2018$OP_CARRIER), size=10, replace=FALSE)

#define sample as all members who belong to one of the 10 operated carriers
clustered_by_op_carrier <- flight.data.y2014.y2018[flight.data.y2014.y2018$OP_CARRIER %in% clusters, ]

#view how many observations came from each tour
table(clustered_by_op_carrier$OP_CARRIER)
```

```
##
##     AA     AS     B6     DL     EV     FL     OH     WN     YV     YX
##  33675   3583    699 315560  22864   1111     94  41173   1723   5196
```

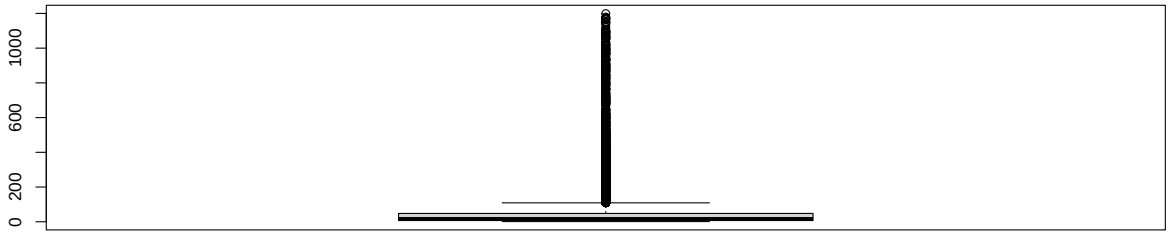
```
clustered_by_op_carrier
```

```
## # A tibble: 425,678 x 28
##    FL_DATE    OP_CARRIER OP_CARRIER_FL_NUM ORIGIN DEST  CRS_DEP_TIME DEP_TIME
##    <date>     <chr>                  <dbl> <chr>  <chr>        <dbl>    <dbl>
##  1 2014-01-01 EV                      4214 CLE    MSP           1220       NA
##  2 2014-01-01 EV                      4380 EWR    MSP            828      930
##  3 2014-01-01 EV                      4472 IAH    MSP           1156     1154
##  4 2014-01-01 EV                      4667 EWR    MSP           1400     1400
##  5 2014-01-01 EV                      5003 CLE    MSP           1200     1159
##  6 2014-01-01 EV                      5009 SYR    MSP            825      820
##  7 2014-01-01 EV                      4981 RIC    MSP            720      710
##  8 2014-01-01 EV                      4685 IAH    MSP           1917     1914
##  9 2014-01-01 EV                      5407 OMA    MSP           1712     1837
## 10 2014-01-01 EV                      5353 IND    MSP           1735     1742
## # ... with 425,668 more rows, and 21 more variables: DEP_DELAY <dbl>,
## #   TAXI_OUT <dbl>, WHEELS_OFF <dbl>, WHEELS_ON <dbl>, TAXI_IN <dbl>,
## #   CRS_ARR_TIME <dbl>, ARR_TIME <dbl>, ARR_DELAY <dbl>, CANCELLED <dbl>,
## #   CANCELLATION_CODE <chr>, DIVERTED <dbl>, CRS_ELAPSED_TIME <dbl>,
## #   ACTUAL_ELAPSED_TIME <dbl>, AIR_TIME <dbl>, DISTANCE <dbl>,
## #   CARRIER_DELAY <dbl>, WEATHER_DELAY <dbl>, NAS_DELAY <dbl>,
## #   SECURITY_DELAY <dbl>, LATE_AIRCRAFT_DELAY <dbl>, `Unnamed: 27` <lgl>
```

Topic 4:

Detailing Summary statistics (Min. , 1st Qu., Median, Mean, 3rd Qu., Max.) of a variable and plotting graphs using ggplot2

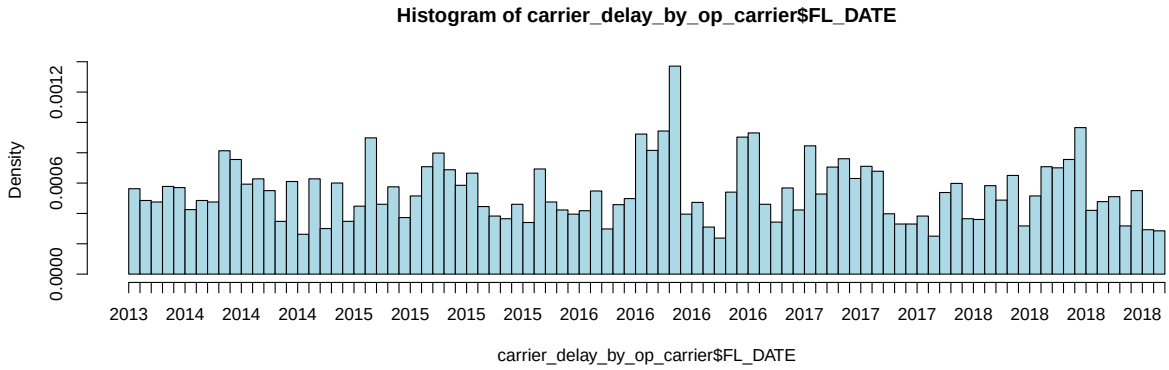
```
# First select the sub-vectors which contains only he coulums we are interested in
carrier_delay_by_op_carrier <- clustered_by_op_carrier %>% select(c(FL_DATE, OP_CARRIER, CARRIER_DELAY))
boxplot(carrier_delay_by_op_carrier$CARRIER_DELAY)
```



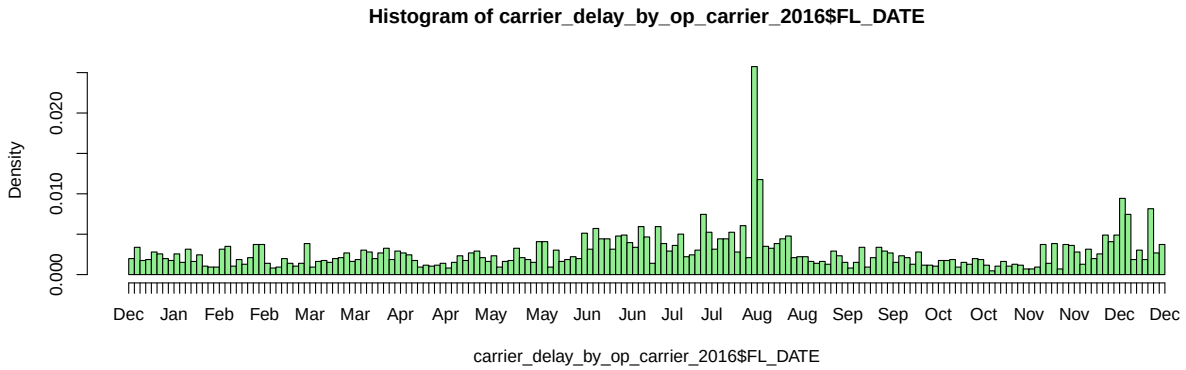

```
IQR(carrier_delay_by_op_carrier$CARRIER_DELAY)
```

```
## [1] 41
```

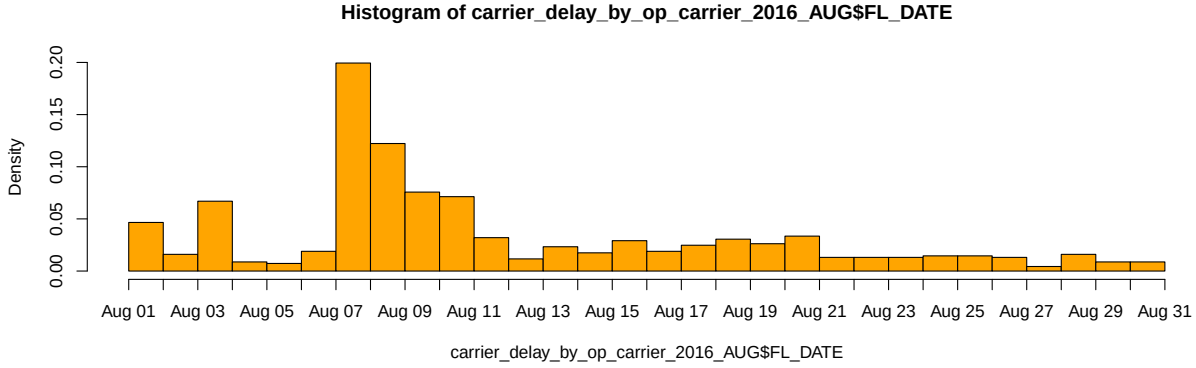
```
hist(carrier_delay_by_op_carrier$FL_DATE, carrier_delay_by_op_carrier$CARRIER_DELAY, breaks = 100, col="
```




```
# In 2016 we have a increase in delay. Let's find out month by analysis to approximate on which month i
carrier_delay_by_op_carrier_2016 <- carrier_delay_by_op_carrier
carrier_delay_by_op_carrier_2016$YEAR <- format(carrier_delay_by_op_carrier$FL_DATE, "%Y")
carrier_delay_by_op_carrier_2016 <- carrier_delay_by_op_carrier_2016 %>% filter(YEAR == '2016')
carrier_delay_by_op_carrier_2016$MONTH <- format(carrier_delay_by_op_carrier_2016$FL_DATE, "%m")
hist(carrier_delay_by_op_carrier_2016$FL_DATE, carrier_delay_by_op_carrier_2016$CARRIER_DELAY, breaks =
```




```
carrier_delay_by_op_carrier_2016_AUG <- carrier_delay_by_op_carrier_2016
carrier_delay_by_op_carrier_2016_AUG <- carrier_delay_by_op_carrier_2016_AUG %>% filter(MONTH == '08')
carrier_delay_by_op_carrier_2016_AUG$DAY <- format(carrier_delay_by_op_carrier_2016_AUG$FL_DATE, "%d")
hist(carrier_delay_by_op_carrier_2016_AUG$FL_DATE, carrier_delay_by_op_carrier_2016_AUG$CARRIER_DELAY, 
```

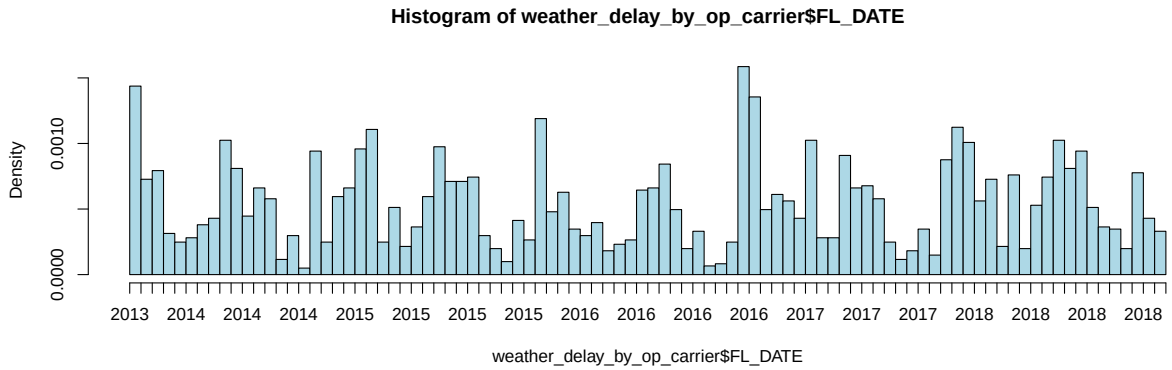



```
summary(carrier_delay_by_op_carrier_2016_AUG$CARRIER_DELAY)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##    1.00    9.00   22.00   70.74   76.00 1167.00
```

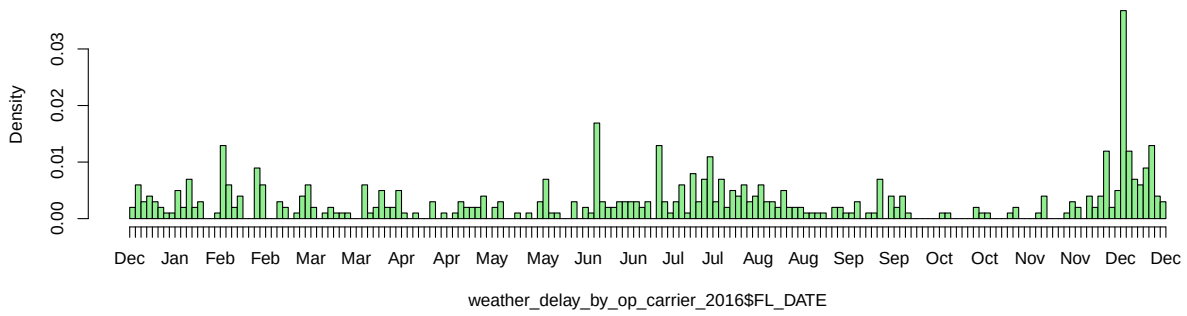
WEATHER DELAY

```
# First select the sub-vectors which contains only he coulums we are interested in
weather_delay_by_op_carrier <- clustered_by_op_carrier %>% select(c(FL_DATE, OP_CARRIER, WEATHER_DELAY)
hist(weather_delay_by_op_carrier$FL_DATE, weather_delay_by_op_carrier$WEATHER_DELAY, breaks = 100, col=
```




```
# In 2016 we have a increase in delay. Let's find out month by analysis to approximate on which month it
weather_delay_by_op_carrier_2016 <- weather_delay_by_op_carrier
weather_delay_by_op_carrier_2016$YEAR <- format(weather_delay_by_op_carrier$FL_DATE, "%Y")
weather_delay_by_op_carrier_2016 <- weather_delay_by_op_carrier_2016 %>% filter(YEAR == '2016')
weather_delay_by_op_carrier_2016$MONTH <- format(weather_delay_by_op_carrier_2016$FL_DATE, "%m")
hist(weather_delay_by_op_carrier_2016$FL_DATE, weather_delay_by_op_carrier_2016$weather_delay, breaks =1
```

```
## Warning: Unknown or uninitialised column: `weather_delay`.
```




```
weather_delay_by_op_carrier_2016
```

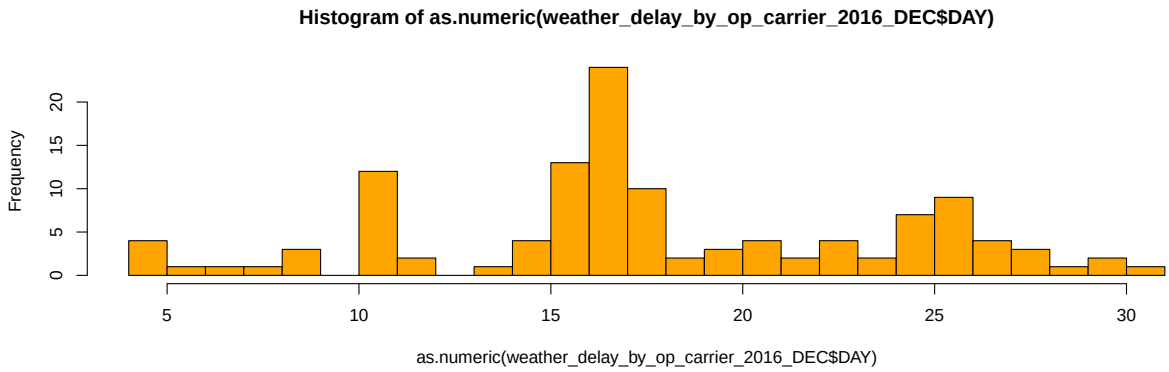
```
## # A tibble: 503 x 5
##    FL_DATE    OP_CARRIER WEATHER_DELAY YEAR  MONTH
##    <date>     <chr>              <dbl> <chr> <chr>
##  1 2016-01-01 DL                     8 2016  01
##  2 2016-01-02 DL                     4 2016  01
##  3 2016-01-03 DL                    20 2016  01
##  4 2016-01-03 DL                    17 2016  01
##  5 2016-01-03 DL                    59 2016  01
##  6 2016-01-03 DL                    12 2016  01
##  7 2016-01-03 DL                   108 2016  01
##  8 2016-01-04 DL                    28 2016  01
##  9 2016-01-06 DL                    33 2016  01
## 10 2016-01-06 DL                    23 2016  01
## # ... with 493 more rows
```

```
weather_delay_by_op_carrier_2016_DEC <- weather_delay_by_op_carrier_2016 %>% filter(MONTH == '12')
weather_delay_by_op_carrier_2016_DEC$DAY <- format(weather_delay_by_op_carrier_2016_DEC$FL_DATE, "%d")
weather_delay_by_op_carrier_2016_DEC
```

```
## # A tibble: 120 x 6
##    FL_DATE    OP_CARRIER WEATHER_DELAY YEAR  MONTH DAY
##    <date>     <chr>              <dbl> <chr> <chr> <chr>
##  1 2016-12-04 DL                    28 2016  12    04
##  2 2016-12-04 DL                     7 2016  12    04
##  3 2016-12-05 DL                     3 2016  12    05
##  4 2016-12-05 DL                   122 2016  12    05
##  5 2016-12-06 DL                   819 2016  12    06
##  6 2016-12-07 DL                     6 2016  12    07
##  7 2016-12-08 DL                    17 2016  12    08
##  8 2016-12-09 DL                   163 2016  12    09
##  9 2016-12-09 DL                    35 2016  12    09
## 10 2016-12-09 DL                    29 2016  12    09
## # ... with 110 more rows
```

```
hist(as.numeric(weather_delay_by_op_carrier_2016_DEC$DAY), weather_delay_by_op_carrier_2016_DEC$weather
```

```
## Warning: Unknown or uninitialised column: 'weather_delay'.
```



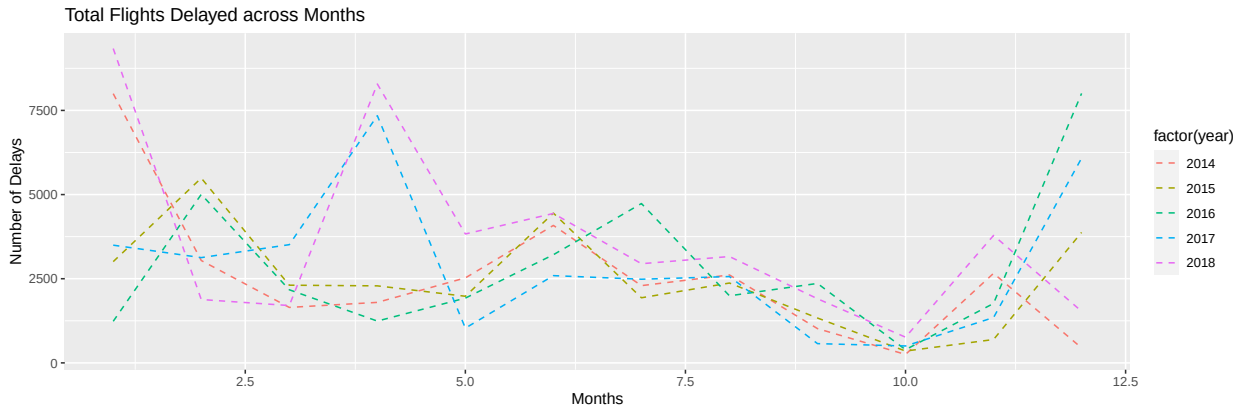

```
summary(weather_delay_by_op_carrier_2016_DEC$weather_delay)
```

```
## Warning: Unknown or uninitialised column: 'weather_delay'.
```

```
## Length  Class   Mode
##      0   NULL   NULL
```

```
weather_delay_by_op_carrier_2014.gorupby <- flight.data.y2014 %>% select(c(FL_DATE, OP_CARRIER, WEATHER

weather_delay_by_op_carrier_2015.gorupby <- flight.data.y2015 %>% select(c(FL_DATE, OP_CARRIER, WEATHER

weather_delay_by_op_carrier_2016.gorupby <- flight.data.y2016 %>% select(c(FL_DATE, OP_CARRIER, WEATHER
```

```
weather_delay_by_op_carrier_2017.gorupby <- flight.data.y2017 %>% select(c(FL_DATE, OP_CARRIER, WEATHER_

weather_delay_by_op_carrier_2018.gorupby <- flight.data.y2018 %>% select(c(FL_DATE, OP_CARRIER, WEATHER_

month_Delay<-rbind(weather_delay_by_op_carrier_2014.gorupby, weather_delay_by_op_carrier_2015.gorupby, w

ggplot(month_Delay, aes(x = `as.numeric(format(FL_DATE, "%m"))`, y = total_delayed, color = factor(year)
geom_line(linetype = 2) +
  labs(title="Total Flights Delayed across Months",y = 'Number of Delays',x = 'Months', fill='YEAR')
```



Topic 5:

Regression (if an increase in number of schedules has any impact/variace on carrier delays).

```
flight.data.y2018
```

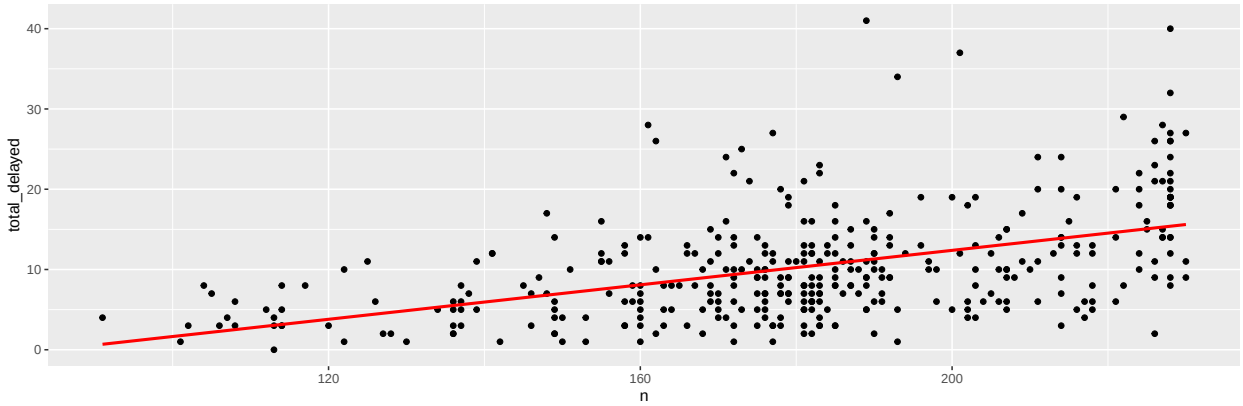
```
## # A tibble: 159,365 x 28
##    FL_DATE    OP_CARRIER OP_CARRIER_FL_NUM ORIGIN DEST  CRS_DEP_TIME DEP_TIME
##    <date>     <chr>                  <dbl> <chr>  <chr>        <dbl>    <dbl>
##  1 2018-01-01 UA                      2118 DEN    MSP           1245     1239
##  2 2018-01-01 UA                      1728 SFO    MSP           2320     2319
##  3 2018-01-01 UA                       878 IAH    MSP           1955     2032
##  4 2018-01-01 UA                       774 ORD    MSP           2245     2244
##  5 2018-01-01 UA                       669 DEN    MSP           2027     2026
##  6 2018-01-01 UA                       573 DEN    MSP            945      944
##  7 2018-01-01 UA                       215 DEN    MSP            756      746
##  8 2018-01-01 AS                        28 SEA    MSP           1750     1748
##  9 2018-01-01 AS                        36 SEA    MSP           1000      951
## 10 2018-01-01 9E                      3615 GFK    MSP           1310     1302
## # ... with 159,355 more rows, and 21 more variables: DEP_DELAY <dbl>,
## #   TAXI_OUT <dbl>, WHEELS_OFF <dbl>, WHEELS_ON <dbl>, TAXI_IN <dbl>,
## #   CRS_ARR_TIME <dbl>, ARR_TIME <dbl>, ARR_DELAY <dbl>, CANCELLED <dbl>,
## #   CANCELLATION_CODE <chr>, DIVERTED <dbl>, CRS_ELAPSED_TIME <dbl>,
## #   ACTUAL_ELAPSED_TIME <dbl>, AIR_TIME <dbl>, DISTANCE <dbl>,
## #   CARRIER_DELAY <dbl>, WEATHER_DELAY <dbl>, NAS_DELAY <dbl>,
## #   SECURITY_DELAY <dbl>, LATE_AIRCRAFT_DELAY <dbl>, 'Unnamed: 27' <lgl>
```

```r
carrier_delay_by_op_carrier_2018.regression <- flight.data.y2018 %>% select(c(FL_DATE, OP_CARRIER, CARR

carrier_delay_by_op_carrier_2018.regression2 <- flight.data.y2018 %>% select(c(FL_DATE, OP_CARRIER, CAR

dataset <- bind_cols(carrier_delay_by_op_carrier_2018.regression, carrier_delay_by_op_carrier_2018.regr
```

```
## New names:
## * `FL_DATE` -> `FL_DATE...1`
## * `year` -> `year...3`
## * `FL_DATE` -> `FL_DATE...4`
## * `year` -> `year...6`
```

```r
ggplot(dataset, aes(x=n, y=total_delayed)) +
  geom_point() +
  geom_smooth(method='lm', se=FALSE, col="red", size=1)
```

```
## `geom_smooth()` using formula 'y ~ x'
```




```r
linear_model <- lm(total_delayed ~ n, data=dataset)
summary(linear_model)
```

```
##
## Call:
## lm(formula = total_delayed ~ n, data = dataset)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -13.1745  -3.8807  -0.8355   3.1588  29.7973
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -9.08524    1.89497  -4.794 2.38e-06 ***
## n            0.10734    0.01031  10.412  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.017 on 363 degrees of freedom
## Multiple R-squared:   0.23,  Adjusted R-squared:  0.2278
## F-statistic: 108.4 on 1 and 363 DF,  p-value: < 2.2e-16
```

Conclusion

I designed this project as a way to review some of the topics we learned in the class/homework/assignments to reinforce some topics learned and also as an opportunity to refer back some of the materials. Hence I thought of picking a variety of topics like sampling strategies, summary statistics, ANOVA and regressions will be the best approach and most I can get from this project. If I have more time, I would have included some more topics (like binom, dbinom, geom...etc distributions) and see if my dataset have variables that can fit these distributions. Given only a academic background in statistics almost almost 20 years ago, I think this subject has given me much learning experience in statistics and I appreciate how these topics are applicable to find solutions in reality.

Access to aws s3 bucket

```
library("aws.s3")
Sys.setenv(
  "AWS_ACCESS_KEY_ID" = "AKIAUTK5NLVJF67UNMH5",
  "AWS_SECRET_ACCESS_KEY" = "",
  "AWS_DEFAULT_REGION" = "us-east-1"
)
```

```
bucketlist()
```

```
## data frame with 0 columns and 0 rows
```