The NYC Flights Dataset

Akshay Prakash

13 February 2015

Contents

Introduction

In this report, the `flights` dataset is described and further analyzed using techniques in **R**. The `flights` dataset is a part of the `nycflights13` data package. The `nycflights13` package contains the data regarding flights that departed New York City (hereafter referred to as NYC) all through the year 2013. The source of the data is the U.S. Department of Transportation's (DOT) Bureau of Transportation Statistics (BTS). Detailed documentation can be found at CRAN:nycflights13. The package includes five data frame objects, the name and function of each is described below.

Data Frame Objects

- **`flights`**

  This dataset provides data related to the on-time performance of all domestic flights departing NYC in 2013. Some of the columns in the data frame include the arrival & departure times and delays, airine carrier abbreviation, flight number, tail number, origin and destination airport code, etc.

- **`airlines`**

  Metadata that acts as a lookup for all the domestic carrier names based on the two letter carrier code from the `flights` dataset.

- **`airports`**

  Metadata about all the domestic airports based on the airport code from the `flights` dataset. The columns include the airport code, airport name, the geographical coordinates, time zone, etc.

- **`planes`**

  Plane metadata for the plane tail numbers from `flights` dataset found in the Federal Aviation Administration (FAA) registry. This excludes American Airways (AA) and Envoy Air (MQ). The metadata includes the manufacture year, number and type of engines, seating capacity, average cruise speed, etc.

- **`weather`**

  Hourly meteorological data at the three airports in NYC with airport codes **JFK**, **LGA** and **EWR**. The data includes the timestamp, temperature, pressure and visibility amongst other meteorological observations.

The names of the airports corresponding to the three airport codes are obtained using the following code chunk.

```
# Load the dplyr and nycflights13 package
library(dplyr)
library(nycflights13)
# Filter rows from airports dataset for the three airport codes
filter.nyc = filter(airports, faa == 'JFK' | faa == 'LGA' | faa == 'EWR')
```

```
# Print the names of the airports in NYC with the FAA codes
select(filter.nyc, faa, name)
```

```
## Source: local data frame [3 x 2]
##
##     faa                name
##   (chr)               (chr)
## 1   EWR Newark Liberty Intl
## 2   JFK John F Kennedy Intl
## 3   LGA          La Guardia
```

In the subsequent sections, the `flights` dataset is described and analyzed in detail.

Dataset description

In this section, the `flights` dataset is described in detail, begining with descriptions of the variables (columns of the data frame). Following this, what each row of the data frame represents is described. The structure of the `flights` data frame is obtained using the `str` function, as shown below.

```
str(flights)
```

```
## Classes 'tbl_df', 'tbl' and 'data.frame':    336776 obs. of  16 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 ...
##  $ 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 ...
##  $ arr_delay: num  11 20 33 -18 -25 12 19 -14 -8 8 ...
##  $ carrier  : chr  "UA" "UA" "AA" "B6" ...
##  $ tailnum  : chr  "N14228" "N24211" "N619AA" "N804JB" ...
##  $ flight   : int  1545 1714 1141 725 461 1696 507 5708 79 301 ...
##  $ 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 5 5 5 5 5 5 ...
##  $ minute   : num  17 33 42 44 54 54 55 57 57 58 ...
```

From the first line of the output, it can be seen that there are **336,776 rows**, with each row having information regarding **16 variables** (or columns). Each line of the output, begining from the second line, provides the name, variable type and the first few values of a variable (column) in the dataset. The 16 variables are described, as below.

Year

1. **`year`** (datatype: *integer*)

   The column indicates the year of departure, which is 2013 for the entire dataset.

Month

2. **`month`** (datatype: *integer*)

   The column records the month of departure, with integer values between 1 and 12.

Day

3. **`day`** (datatype: *integer*)

   This variable records the day of the month of departure.

Dep_Time

4. **`dep_time`** (datatype: *numeric*)

   This variable records the time of departure, in the local time zone of the airport. The last two digits should represent the *minutes*, as a number between 0 and 59, while the first two (or one) should represent the *hour* as a number between 0 and 23.

Dep_Delay

5. **`dep_delay`** (datatype: *numeric*)

   This variable records the departure delays, in *minutes*. Negative times represent early departures.

Arr_Time

6. **`arr_time`** (datatype: *numeric*)

   This variable records the time of arrival, in the local time zone of the destination airport. The last two digits should represent the *minutes*, as a number between 0 and 59, while the first two (or one) should represent the *hour* as a number between 0 and 24.

Arr_Delay

7. **`arr_delay`** (datatype: *numeric*)

   This variable records the arrival delays, in *minutes*. Negative times represent early arrivals.

Carrier

8. **`carrier`** (datatype: *character*)

   Represents the domestic carrier (airline) as a two-letter abbreviation (code). The `airlines` dataset can be looked up to get the complete name of the carrier for the two-letter code.

TailNum

9. **`tailnum`** (datatype: *character*)

   A six-letter code associated with the tail number of the aircraft. Further details of the plane can be looked up from the `planes` dataset, based on the tail number.

Flight

10. **`flight`** (datatype: *integer*)

    The flight number, an integer. A departing flight can be uniquely identified by a combination of the carrier code followed by the flight number.

Origin

11. **`origin`** (datatype: *character*)

    The three-letter FAA airport code of the airport of departure. The airport name and other information can be looked up from the `airports` dataset.

Dest

12. **`dest`** (datatype: *character*)

    The three-letter FAA airport code of the destination airport. The airport name and other information can be looked up from the `airports` dataset.

Airtime

13. **`air_time`** (datatype: *numeric*)

    Represents the amount of time spent in the air, in *minutes*.

Distance

14. **`distance`** (datatype: *numeric*)

    The distance flown (may not be the distance between airports), in *miles*.

hour

15. **hour** (datatype: *numeric*)

    The hour of departure, number betweeen 0 and 24 (instead of between 0 and 23).

minute

16. **minute** (datatype: *numeric*)

    The minute of departure, a number between 0 and 59.

The first three rows of the 'flights' dataset are displayed using the code below.

```
# To display all the columns without truncation
options(dplyr.width = Inf)
# To display the first 3 rows
head(flights, 3)
```

```
## Source: local data frame [3 x 16]
##
##    year month   day dep_time dep_delay arr_time arr_delay carrier tailnum
##   (int) (int) (int)    (int)     (dbl)    (int)     (dbl)   (chr)   (chr)
## 1  2013     1     1      517         2      830        11      UA  N14228
## 2  2013     1     1      533         4      850        20      UA  N24211
## 3  2013     1     1      542         2      923        33      AA  N619AA
##   flight origin  dest air_time distance  hour minute
##    (int)  (chr) (chr)    (dbl)    (dbl) (dbl)  (dbl)
## 1   1545    EWR   IAH      227     1400     5     17
## 2   1714    LGA   IAH      227     1416     5     33
## 3   1141    JFK   MIA      160     1089     5     42
```

Each row represents the information regarding the travel of a particular scheduled flight, on a specific date and time. For example, row 2 shows that the flight number 1714, operated by the airlines UA, departed from LGA airport at 05:33 hrs (EST)with a delay of 4 minutes, covered a distance of 1416 miles in 227 minutes, arriving at IAH (Houston) airport at 8:50 hrs (CST), 20 minutes later than scheduled arrival time.

Dataset preparation

Based on understanding of the variables of the `flights` dataset described in the previous section, it makes sense to make the following variables as a `factor` variable type.

- **month**
- **day**

- **hour**
- **carrier**
- **origin**
- **dest**

This is done using the `factor` command, as shown in the code block below. From now on, the modified version of `flights` dataset is stored as a new dataframe `flights.dataset`.

```
# New data frame by the name flights.dataset is created from flights
flights.dataset = flights
# Change to factor variable type
flights.dataset$month = factor(flights.dataset$month)
flights.dataset$hour = factor(flights.dataset$hour)
flights.dataset$day = factor(flights.dataset$day)
flights.dataset$carrier = factor(flights.dataset$carrier)
flights.dataset$origin = factor(flights.dataset$origin)
flights.dataset$dest = factor(flights.dataset$dest)

# Print the levels of the factor variable month
levels(flights.dataset$month)     # 12 levels expected
```

```
##  [1] "1"  "2"  "3"  "4"  "5"  "6"  "7"  "8"  "9"  "10" "11" "12"
```

There are 12 levels for the *month* variable, as expected.

```
# Print the levels of the factor variable hour
levels(flights.dataset$hour)      # 24 levels expected
```

```
##  [1] "0"  "1"  "2"  "3"  "4"  "5"  "6"  "7"  "8"  "9"  "10" "11" "12" "13"
## [15] "14" "15" "16" "17" "18" "19" "20" "21" "22" "23" "24"
```

It is interesting to observe that there are 25 levels for the `hour` variable, from 0 to 24. Further analysis (refer code block below) shows that hour is 24 only when the `dep_time` is exactly **2400** (looking at max and min).

```
# Summary of dep_time where hour = 24 in flights.dataset
summary(select(filter(flights, dep_time >= 2400),dep_time))
```

```
##     dep_time
##  Min.   :2400
##  1st Qu.:2400
##  Median :2400
##  Mean   :2400
##  3rd Qu.:2400
##  Max.   :2400
```

Hence, we modify the `flights` dataset to represent hour **24** as hour **00**, using the following code.

```
# Dimensions to look at the number of rows.
dim(filter(flights.dataset, hour == 0))
```

```
## [1] 881  16
```

```
dim(filter(flights.dataset, hour == 24))
```

```
## [1] 29 16
```

```
# Set hour 24 as hour 0 and dep_time 2400 as dep_time 0
flights.dataset$hour[flights.dataset$hour == 24] = 0
flights.dataset$dep_time[flights.dataset$dep_time == 2400] = 0
dim(filter(flights.dataset, hour == 0))
```

```
## [1] 910  16
```

It can be seen that 29 rows have been modified from the dimensions of the data frame before and after modifying the dataset. The levels of the other factor variables, using the following code, are analyzed.

```
# Print the levels of the factor variable carrier, origin and dest
levels(flights.dataset$carrier)   # 16 levels expected
```

```
##  [1] "9E" "AA" "AS" "B6" "DL" "EV" "F9" "FL" "HA" "MQ" "OO" "UA" "US" "VX"
## [15] "WN" "YV"
```

```
levels(flights.dataset$origin)    # 3 levels expected
```

```
## [1] "EWR" "JFK" "LGA"
```

```
levels(flights.dataset$dest)      # 1393 levels or less expeccted
```

```
##   [1] "ABQ" "ACK" "ALB" "ANC" "ATL" "AUS" "AVL" "BDL" "BGR" "BHM" "BNA"
##  [12] "BOS" "BQN" "BTV" "BUF" "BUR" "BWI" "BZN" "CAE" "CAK" "CHO" "CHS"
##  [23] "CLE" "CLT" "CMH" "CRW" "CVG" "DAY" "DCA" "DEN" "DFW" "DSM" "DTW"
##  [34] "EGE" "EYW" "FLL" "GRR" "GSO" "GSP" "HDN" "HNL" "HOU" "IAD" "IAH"
##  [45] "ILM" "IND" "JAC" "JAX" "LAS" "LAX" "LEX" "LGA" "LGB" "MCI" "MCO"
##  [56] "MDW" "MEM" "MHT" "MIA" "MKE" "MSN" "MSP" "MSY" "MTJ" "MVY" "MYR"
##  [67] "OAK" "OKC" "OMA" "ORD" "ORF" "PBI" "PDX" "PHL" "PHX" "PIT" "PSE"
##  [78] "PSP" "PVD" "PWM" "RDU" "RIC" "ROC" "RSW" "SAN" "SAT" "SAV" "SBN"
##  [89] "SDF" "SEA" "SFO" "SJC" "SJU" "SLC" "SMF" "SNA" "SRQ" "STL" "STT"
## [100] "SYR" "TPA" "TUL" "TVC" "TYS" "XNA"
```

It can be seen that, there are 16 different carriers that operated flights out of the three airports in NYC, meaning all the 16 airlines (from the `airlines` dataset) operated flights out of at least one of the airports in NYC. The 3 levels of the origin airport code makes obvious sense. However, it is interesting to note that there are only 105 levels of the factor variable `dest`, pertaining to the destination airports, out of the 1397 airports in the `airports` dataset. This means that **there were flights to a total of 105 unique destinations from NYC**.

An additional column called `madeup_time` is added to the `flights.dataset` dataset to record the time made up or lost during flight based on the departure and arrival delays. The variable is called `madeup_time` given by the following code. The information provided by this variable is presented in the next section when variable summaries are discussed.

```
# Add column using PIPE (%>%) function
flights.dataset <- flights.dataset%>%
  mutate(madeup_time = dep_delay - arr_delay)
```

We also add another variable `date` that stores the date of departure of flight, using the code chunk below.

```
# Load library to use date
library(lubridate)
# Add new variable date for date of departing flight
flights.dataset <- flights.dataset %>%
  mutate(date=ymd(paste(year, month, day, sep="-"))) %>% na.omit()
```

An additional column`dep_delay_type` is created to represent the positive, negative departure delays and on-time departure as *delay*, *early* and *ontime* respectively.

```
flights.dataset <- flights.dataset %>%mutate(dep_delay_type = 'NA')
flights.dataset$dep_delay_type[flights.dataset$dep_delay == 0] = 'ontime'
flights.dataset$dep_delay_type[flights.dataset$dep_delay > 0] = 'delay'
flights.dataset$dep_delay_type[flights.dataset$dep_delay < 0] = 'early'
flights.dataset$dep_delay_type = factor(flights.dataset$dep_delay_type)
```

Variable summaries

There have been a total of **336,776** flights out of NYC in the year 2013. In this section, various summaries of the variables (of the dataset prepared in the previous section) are presented.

1. **`year`**

   The `year` variable has only one value of 2013.

2. **`month`**

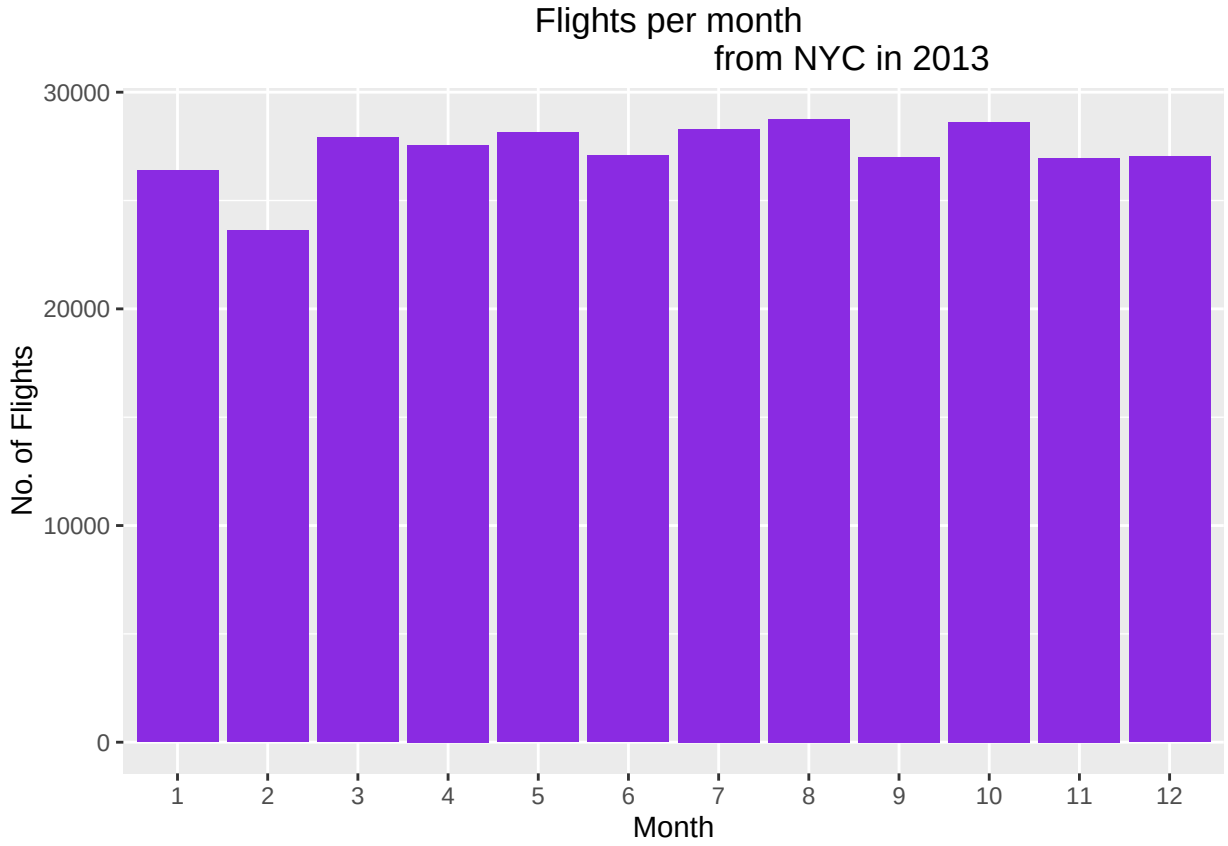
   The analysis of the air traffic of flights flying out of NYC, per month is presented. This is based on `month` variable as a factor variable because a numeric variable in this case will not

be providing meaningful insights using `summary`. The summary of this variable and a bar plot showing the number of flights out of NYC on a monthly basis is presented using the code below.

```
# Load the ggplot2 package
library(ggplot2)
# Variable summary for month (a factor variable)
summary(flights.dataset$month)
```

```
##     1     2     3     4     5     6     7     8     9    10    11    12
## 26398 23611 27902 27564 28128 27075 28293 28756 27010 28618 26971 27020
```

```
# Plot summary as a bar chart
ggplot(data = flights.dataset, mapping = aes(x=month))+
  geom_bar(fill="blueviolet") + labs(title = "Flights per month
                                     from NYC in 2013",
                            x = "Month", y = "No. of Flights")+
  theme(plot.title = element_text(colour = "black"))
```



Based on the bar chart above, it can be inferred that the number of flights that departed NYC is uniformly distributed across the different months of the year 2013.

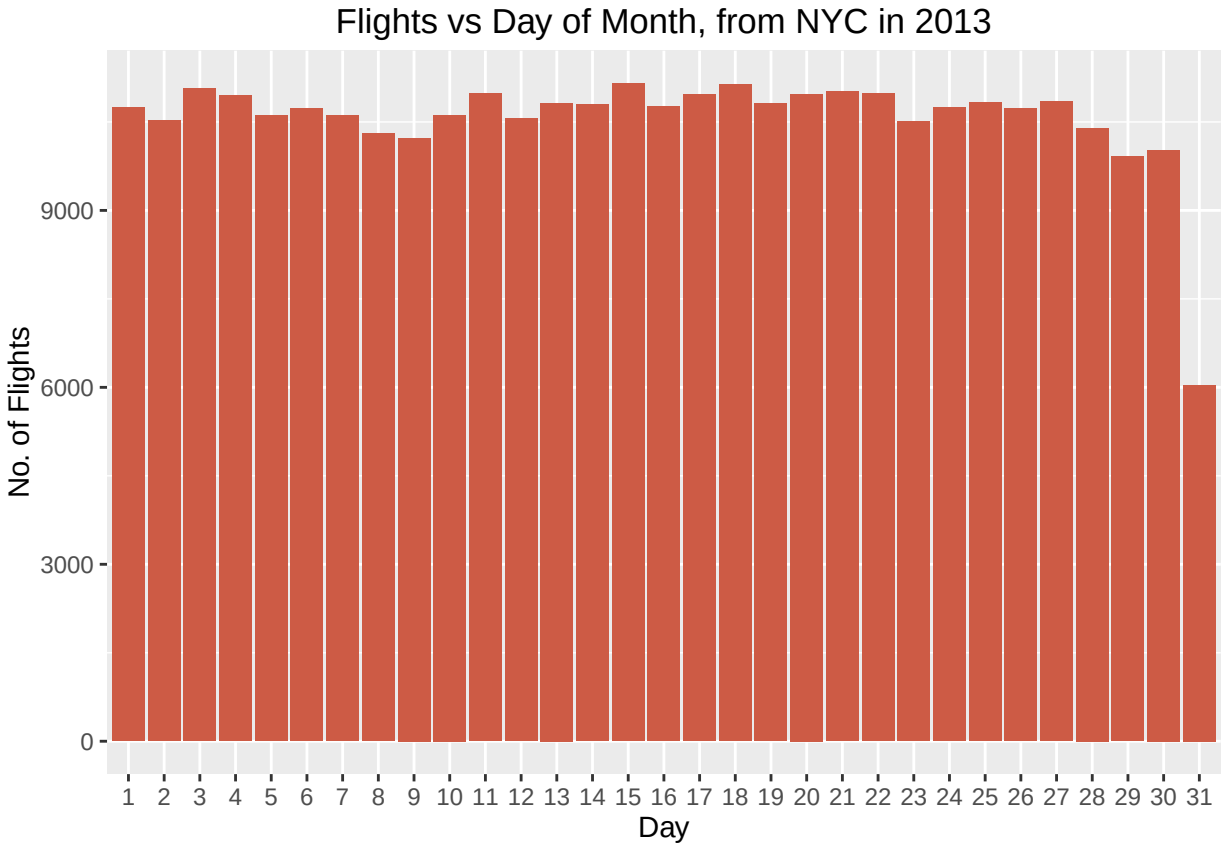
3. **day**

The flights out of NYC as a function of the day of a month is analyzed using the code chunk below wherein the summary and a bar chart are presented below.

```r
# Variable summary for day (as a factor variable)
summary(factor(flights.dataset$day))
```

```
##     1     2     3     4     5     6     7     8     9    10    11    12
## 10748 10524 11070 10949 10609 10725 10611 10308 10231 10614 10983 10556
##    13    14    15    16    17    18    19    20    21    22    23    24
## 10818 10794 11150 10758 10961 11131 10811 10974 11017 10985 10511 10741
##    25    26    27    28    29    30    31
## 10827 10724 10845 10394  9916 10023  6038
```

```r
# Plot summary as a bar chart
ggplot(data = flights.dataset, mapping = aes(x=day))+
  geom_bar(fill="coral3") +
labs(title = "Flights vs Day of Month, from NYC in 2013", x = "Day",
     y = "No. of Flights")+
  theme(plot.title = element_text(colour = "black"))
```



Again, it can be seen that the number of flights that departed NYC is uniformly distributed across the days of every month. The exceptions, in the bar plot, is for the days 29 through 31, which can be attributed to the fact that 2013 was a leap year (February having only 28 days) and only seven months have the 31st day.

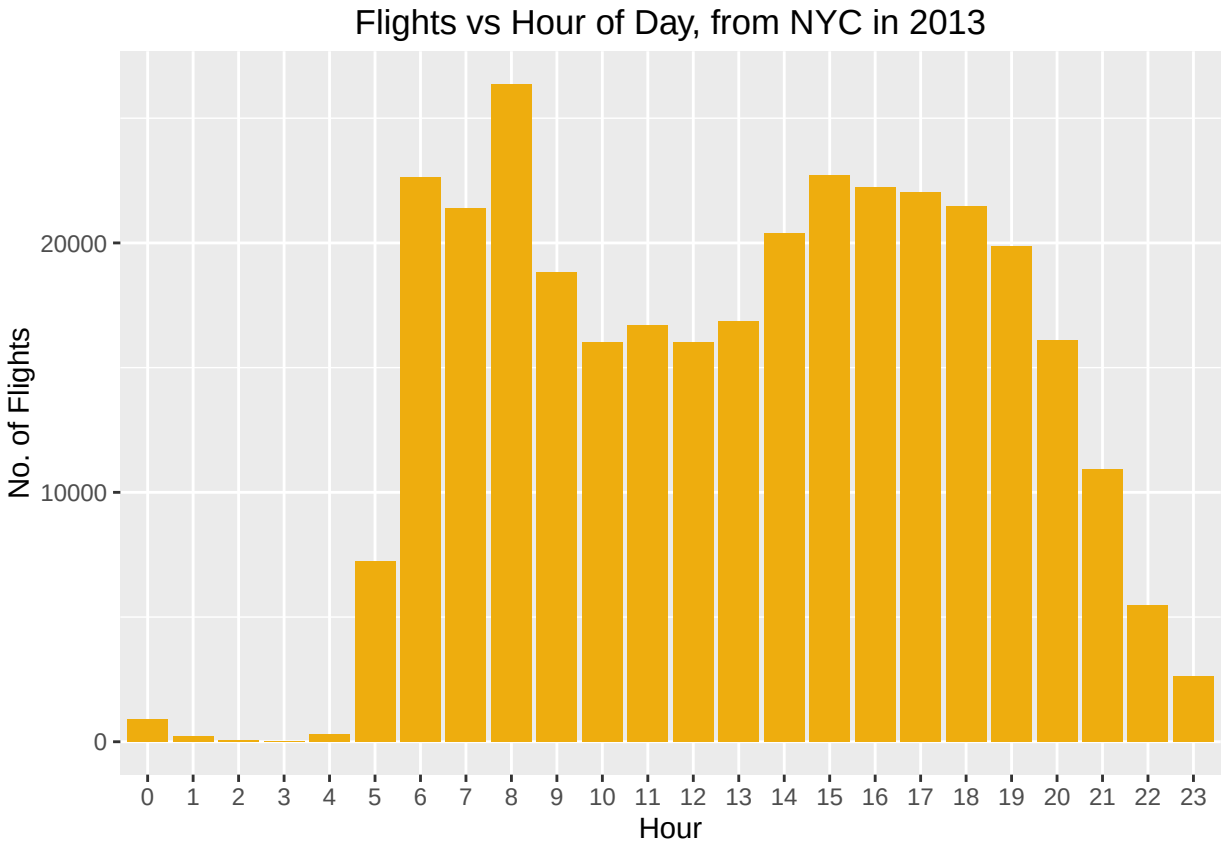
4. **hour**

The code chunk to analyze the `hour` variable is below.

```r
# Variable summary for hour (as a factor variable)
summary(flights.dataset$hour)
```

```
##     0     1     2     3     4     5     6     7     8     9    10    11
##   905   221    63    11   305  7230 22627 21409 26367 18820 16031 16712
##    12    13    14    15    16    17    18    19    20    21    22    23
## 16024 16845 20394 22703 22236 22039 21453 19871 16097 10914  5461  2608
##    24
##     0
```

```r
# Plot summary as a bar chart
ggplot(data = flights.dataset, mapping = aes(x=hour))+
  geom_bar(fill="darkgoldenrod2") +
  labs(title = "Flights vs Hour of Day, from NYC in 2013",
       x = "Hour", y = "No. of Flights") +
  theme(plot.title = element_text(colour = "black"))
```



It can be seen that there are only a few flights that departed during the first four hours of the day. Starting from the fifth hour on, it gradually picks up and had the maximum number of flights during the 8^{th} hour of the day (26,367 flights). Then it drops and remains uniform until after noon. Again, it gradually increases, peaking during the fifteenth and eighteenth hours and dropping thereafter. The observed pattern makes sense, expectedly. Interestingly, it is also to be noted that there are 8,255 flights that has *NA* in the `hour` column.

Having summarized `hour` variable, summarizing the `dep_time` variable does not necessarily provide additional information.

5. **dep_delay**

   The code chunk to analyze the `dep_delay` variable is below.

```
summary(flights.dataset$dep_delay)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  -43.00   -5.00   -2.00   12.56   11.00 1301.00
```

The summary shows that on an average flights have been delayed by 13 minutes. There have been flights departing as early as 43 minutes to as late as 22 hours. The discrepancy between the mean and the median indicates the effect of outliers. So, it makes sense to analyze the positive and negative delays separately. For this, we provide a summary of the `dep_delay` grouped by the `dep_delay_type`.

```
summarize(group_by(flights.dataset,dep_delay_type),
          Mean = mean(dep_delay, na.rm = TRUE),
          Median = median(dep_delay, na.rm = TRUE),
          Max = max(dep_delay, na.rm = TRUE),
          Min = min(dep_delay, na.rm = TRUE),
          Total_Flights = n(), SD = sd(dep_delay, na.rm = TRUE))
```

```
## Source: local data frame [3 x 7]
##
##   dep_delay_type      Mean Median   Max   Min Total_Flights        SD
##           (fctr)     (dbl)  (dbl) (dbl) (dbl)         (int)     (dbl)
## 1          delay 39.237473     19  1301     1        127745 54.143382
## 2          early -4.928119     -5    -1   -43        183135  2.833452
## 3         ontime  0.000000      0     0     0         16466  0.000000
```

Based on the first row of the results above, flights have been delayed as high as 22 hours to as low as a 1 minute and the average delay is about 40 minutes. However, a 22 hour delay is an outlier (due to extreme circumstances). Hence, the median of 19 minutes may be considered a better statistic to understand the delay. In this case, it shows that about half of 128,432 flights had a delay of less than 19 minutes.

The second row shows that flights have departed early by an average of 5 minutes, with the earliest being 43 minutes. Also, the median indicates that about half of 183,575 flights departed earlier than 5 minutes.

The third row shows that 16,514 flights departed on-time. However, it is to be noted that there were 8,255 flights that do not have the dep_delay information. This number of 8,255, observed even in the last variable summary of `hour`, could have been flights that were cancelled. However, to be certain about it further analysis need to be done later (not in this report).

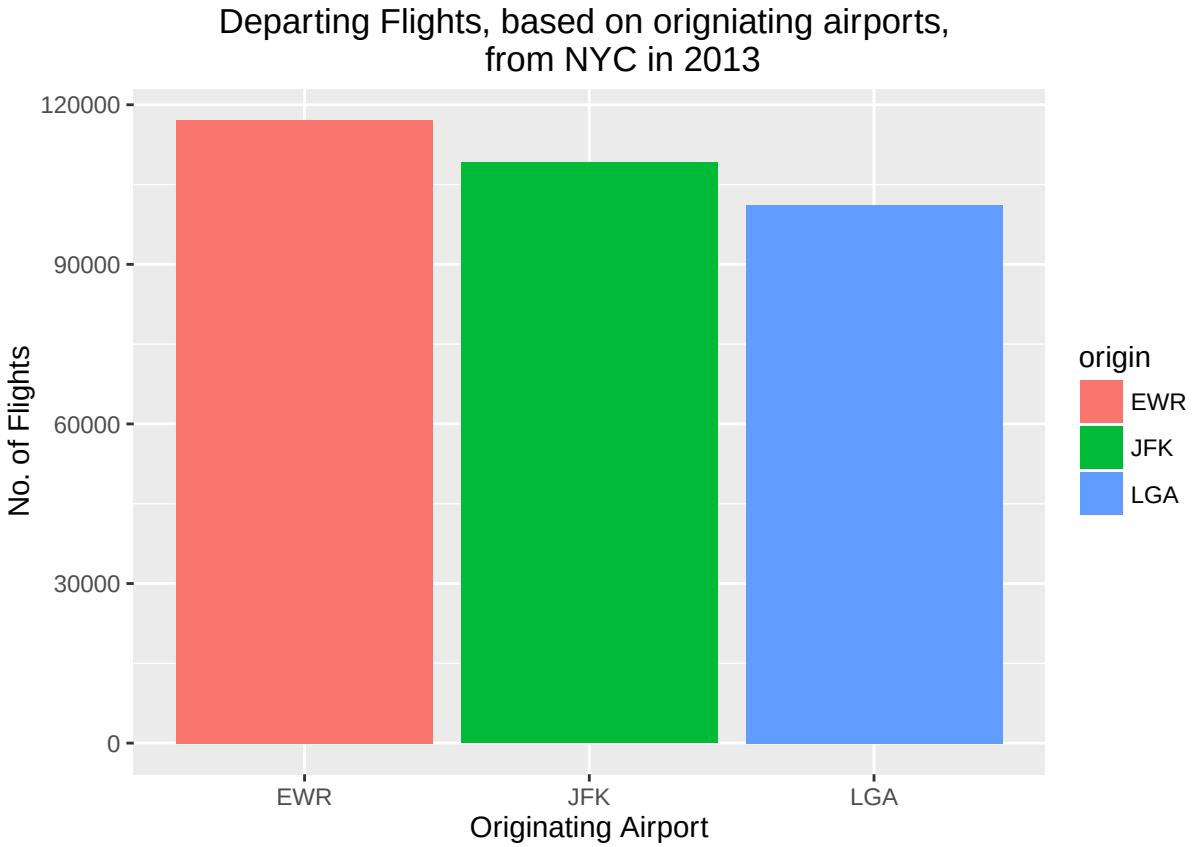
6. **origin**

For the three airports in NYC, the summary of the number of flights departed are presented along with a bar plot showing the number of flights that departed based on the originating airport.

```
summary(flights.dataset$origin)
```

```
##    EWR    JFK    LGA
## 117127 109079 101140
```

```
# Plot summary as a bar chart
ggplot(data = flights.dataset, mapping = aes(x=origin))+
  geom_bar(aes(fill=origin)) +
  labs(title = "Departing Flights, based on origniating airports,
       from NYC in 2013",
       x = "Originating Airport", y = "No. of Flights") +
  theme(plot.title = element_text(colour = "black"))
```



It is clear that Newark airport had the maximum number of flights from the NYC region compared to JFK and La Guardia, although the numbers are comparable within a few thousand flights over the year.

7. **carrier**

For the `carrier` variable, we already know that there are 16 levels (airlines) that operate out of NYC. The names of the 16 airlines operating out of NYC region are listed below along with the summary sorted in ascending order of the number of flights that each carrier operated in 2013 out of NYC. A bar plot the number of flights operated by each airline is presented as well.

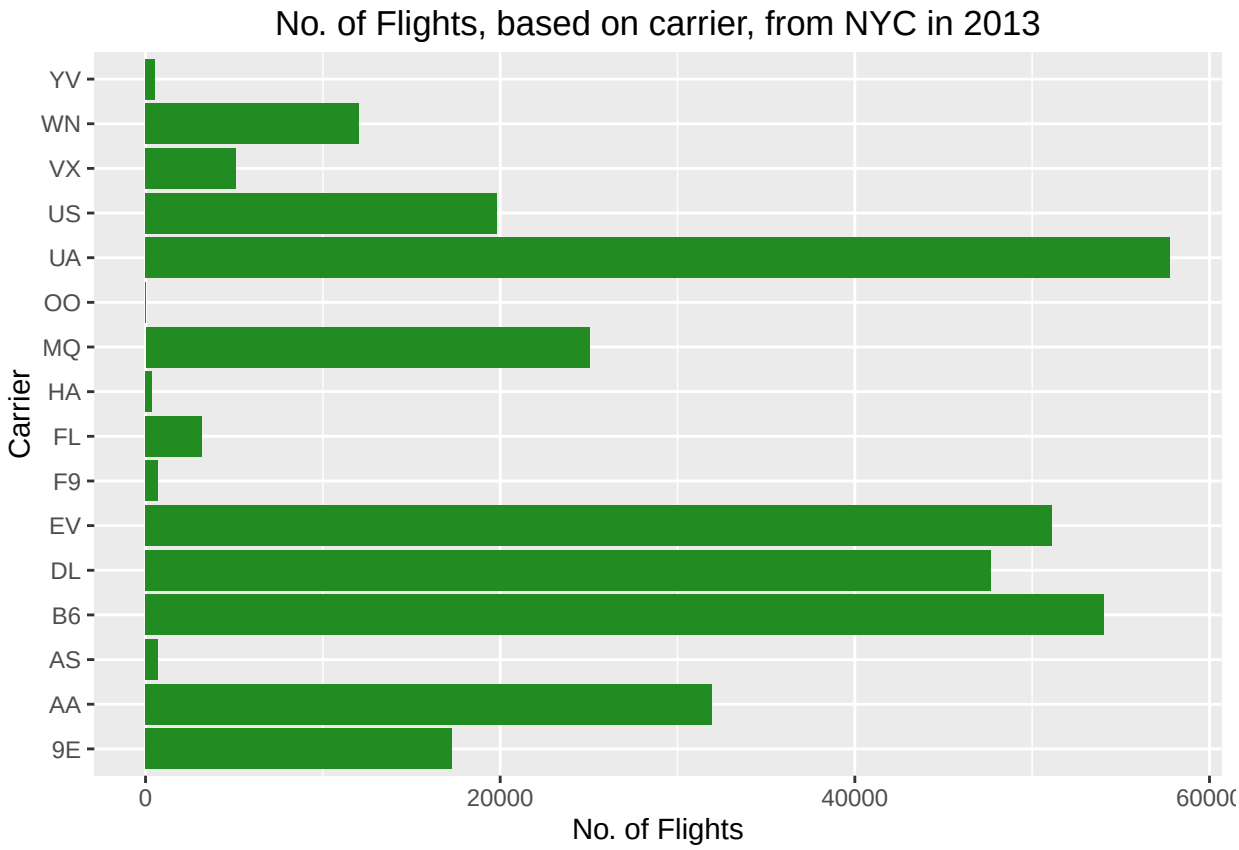
```
airlines
```

```
## Source: local data frame [16 x 2]
##
##    carrier                       name
##     (fctr)                     (fctr)
## 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.
```

```
sort(summary(flights.dataset$carrier))
```

```
##    OO    HA    YV    F9    AS    FL    VX    WN    9E    US    MQ    AA
##    29   342   544   681   709  3175  5116 12044 17294 19831 25037 31947
##    DL    EV    B6    UA
## 47658 51108 54049 57782
```

```
# Plot summary as a bar chart
ggplot(data = flights.dataset, mapping = aes(x=carrier))+
  geom_bar(fill="forestgreen") +
  labs(title = "No. of Flights, based on carrier, from NYC in 2013",
       x = "Carrier", y = "No. of Flights") +
  coord_flip()+
  theme(plot.title = element_text(colour = "black"))
```



The airlines that operated a high frequency (more than 40,000 in 2013) of flights from NYC were UA, EV, DL and B6 while AA, 9E, MQ and US operated a moderate frequency (between 19,000 and 40,000 yearly). The remaining airlines operated less than 10,000 flights out of NYC in 2013.

8. **distance**

The summary of `distance` is presented below.

```
summarize(flights.dataset, Mean = mean(distance,na.rm=TRUE),
          Min = min(distance,na.rm=TRUE),
          Max = max(distance,na.rm=TRUE),
          Median = median(distance,na.rm=TRUE))
```

```
## Source: local data frame [1 x 4]
##
##       Mean   Min   Max Median
##      (dbl) (dbl) (dbl)  (dbl)
## 1 1048.371    80  4983    888
```

Based on the summary, the maximum distance of a flight that departed NYC is about 5000 miles, with the minimum being 17 miles (this flight detail is presented below). The average distance of flights were about 1040 miles. The median indicates that about 50 percent of flights were to destinations less than 872 miles.

```
filter(flights.dataset, distance == 17)
```

```
## Source: local data frame [0 x 19]
##
## Variables not shown: year (int), month (fctr), day (fctr), dep_time (dbl),
##   dep_delay (dbl), arr_time (int), arr_delay (dbl), carrier (fctr),
##   tailnum (chr), flight (int), origin (fctr), dest (fctr), air_time (dbl),
##   distance (dbl), hour (fctr), minute (dbl), madeup_time (dbl), date
##   (time), dep_delay_type (fctr)
```

It is quite possible that this happened to be an erroneous data entry or it could have been a chartered flight.

9. **madeup_time**

   The summary of `madeup_time` is presented below.

```
summary(flights.dataset$madeup_time)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## -196.00   -3.00    7.00    5.66   17.00  109.00
```

Based on the summary median, it is clear that about 50 percent of the flights made up more than 7 minutes during flight.

Relationships between variables

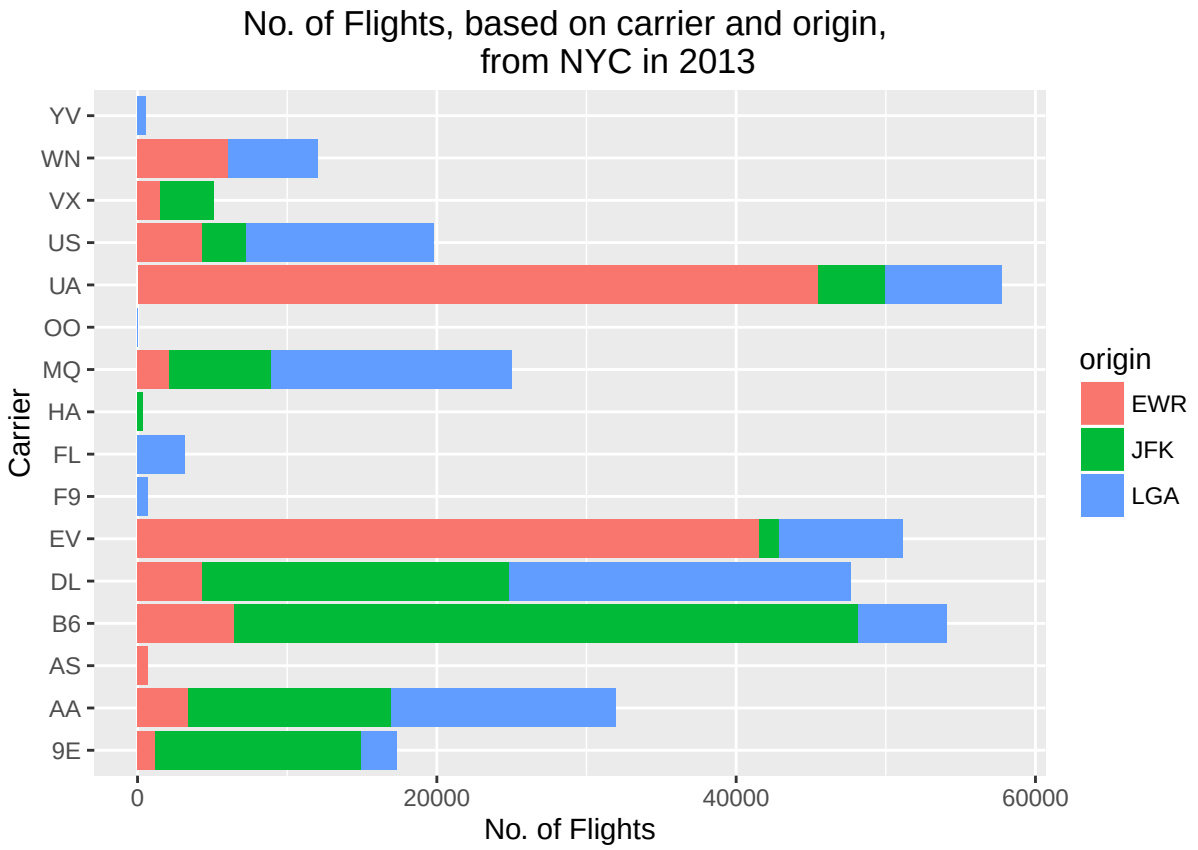
The `carrier` variable is chosen and its relationship with several other variables are analyzed. These analysis are primarily based on visualization based on the set of plots presented in this section.

Carrier VS Originating Airport

- **Carrier VS Originating Airport**

  A bar plot of the number of flights operated by a carrier is presented as a function of the originating airport using the code chunk below.

```
ggplot(data = flights.dataset, mapping = aes(x=carrier))+
  geom_bar(aes(fill = origin)) +
   labs(title = "No. of Flights, based on carrier and origin,
        from NYC in 2013",
        x = "Carrier", y = "No. of Flights") +
  coord_flip()+
   theme(plot.title = element_text(colour = "black"))
```



Out of these, B6 operated the maximum number of flights out of JFK yearly, while UA and EV operated the bulk of flights out of EWR. The airlines DL and MQ operated the maximum number of flights out of LGA. Also, F9, OO, YV and FL flew only from the LGA airport, HA flew only from JFK while AS flew only from EWR.

Carrier VS Departure Delay

- **Carrier VS Departure Delay**

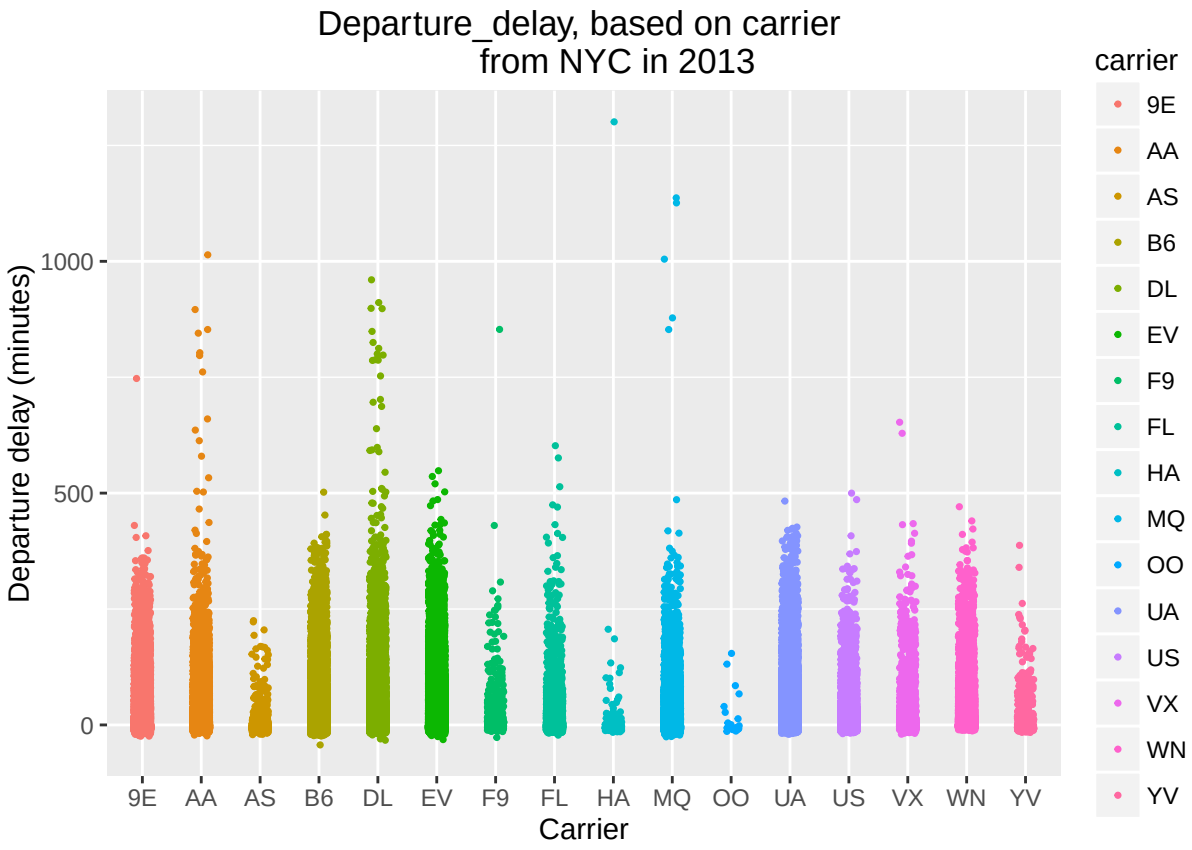
  The dataset is grouped by `carrier` and the mean and median `dep_delay` of these groups are calculated. Also, a scatter plot is presented to visualize.

```
carrierVSdep_delay<-flights.dataset %>%
  group_by(carrier) %>%
  summarize(mean_delay=mean(dep_delay, na.rm = TRUE),
            median_delay =median(dep_delay, na.rm = TRUE))
carrierVSdep_delay
```

```
## Source: local data frame [16 x 3]
##
##    carrier mean_delay median_delay
##     (fctr)      (dbl)        (dbl)
## 1       9E  16.439574           -2
## 2       AA   8.569130           -3
```

```
## 3        AS   5.830748           -3
## 4        B6  12.967548           -1
## 5        DL   9.223950           -2
## 6        EV  19.838929           -1
## 7        F9  20.201175            0
## 8        FL  18.605984            1
## 9        HA   4.900585           -4
## 10       MQ  10.445381           -3
## 11       OO  12.586207           -6
## 12       UA  12.016908            0
## 13       US   3.744693           -4
## 14       VX  12.756646            0
## 15       WN  17.661657            1
## 16       YV  18.898897           -2
```

```
  ggplot(flights.dataset, aes(carrier, dep_delay)) +
geom_point(aes(color=carrier), size = 0.6,
       position = position_jitter(width = 0.35))+
labs(title = "Departure_delay, based on carrier
     from NYC in 2013",
        x = "Carrier", y = "Departure delay (minutes)") +
theme(plot.title = element_text(colour = "black"))
```



The scatter plot, with jitter, shows the departure delay (in minutes) for the various carriers.

Based on the plot, it is quite clear that HA and OO had the lowest frequencies of delays compared to other aircraft carriers.

Carrier VS Unique Destinations

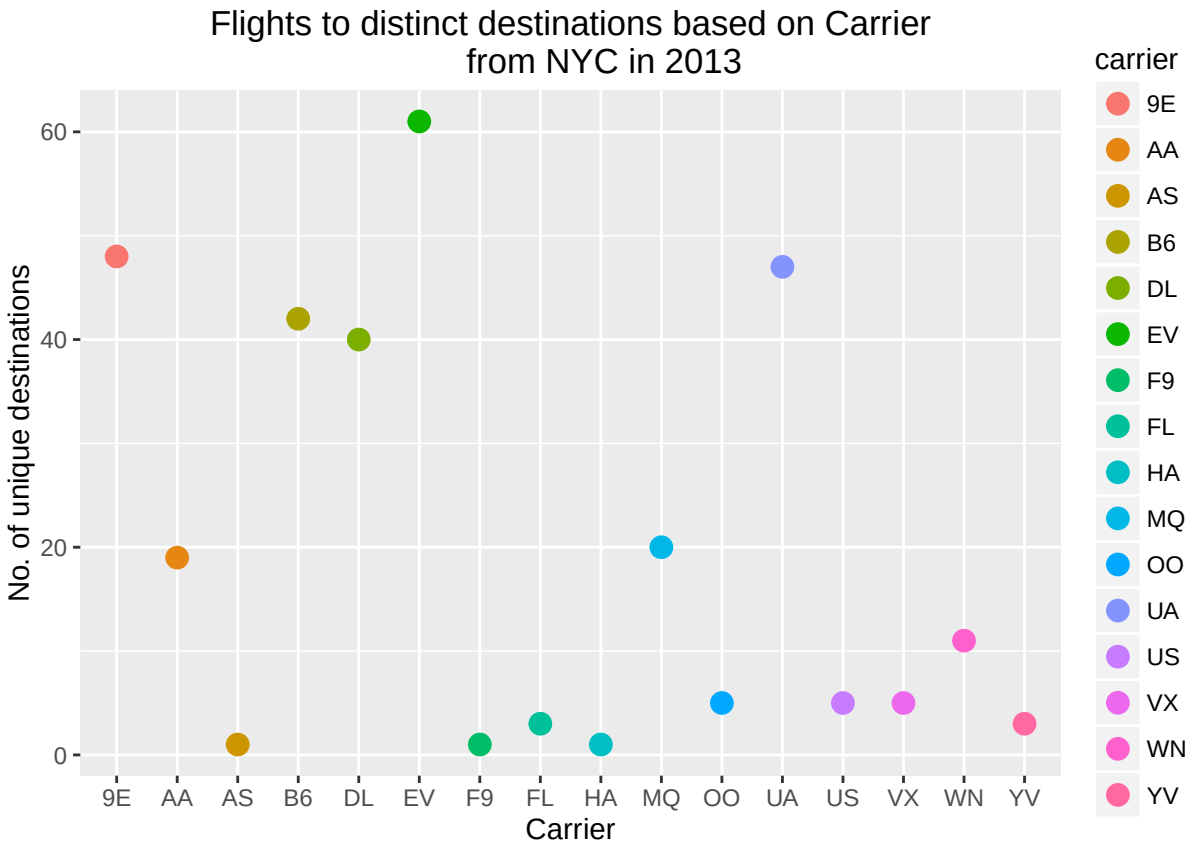
- **Carrier VS Unique Destinations**

  The `flights.dataset` dataset is grouped by the carrier and the number of unique destinations per carrier is counted. This is presented further as a point plot.

```
carrierVSuniquedest<- flights.dataset %>%
  group_by(carrier) %>%
  summarize(unique_dest = n_distinct(dest))
carrierVSuniquedest
```

```
## Source: local data frame [16 x 2]
##
##    carrier unique_dest
##     (fctr)       (int)
## 1       9E          48
## 2       AA          19
## 3       AS           1
## 4       B6          42
## 5       DL          40
## 6       EV          61
## 7       F9           1
## 8       FL           3
## 9       HA           1
## 10      MQ          20
## 11      OO           5
## 12      UA          47
## 13      US           5
## 14      VX           5
## 15      WN          11
## 16      YV           3
```

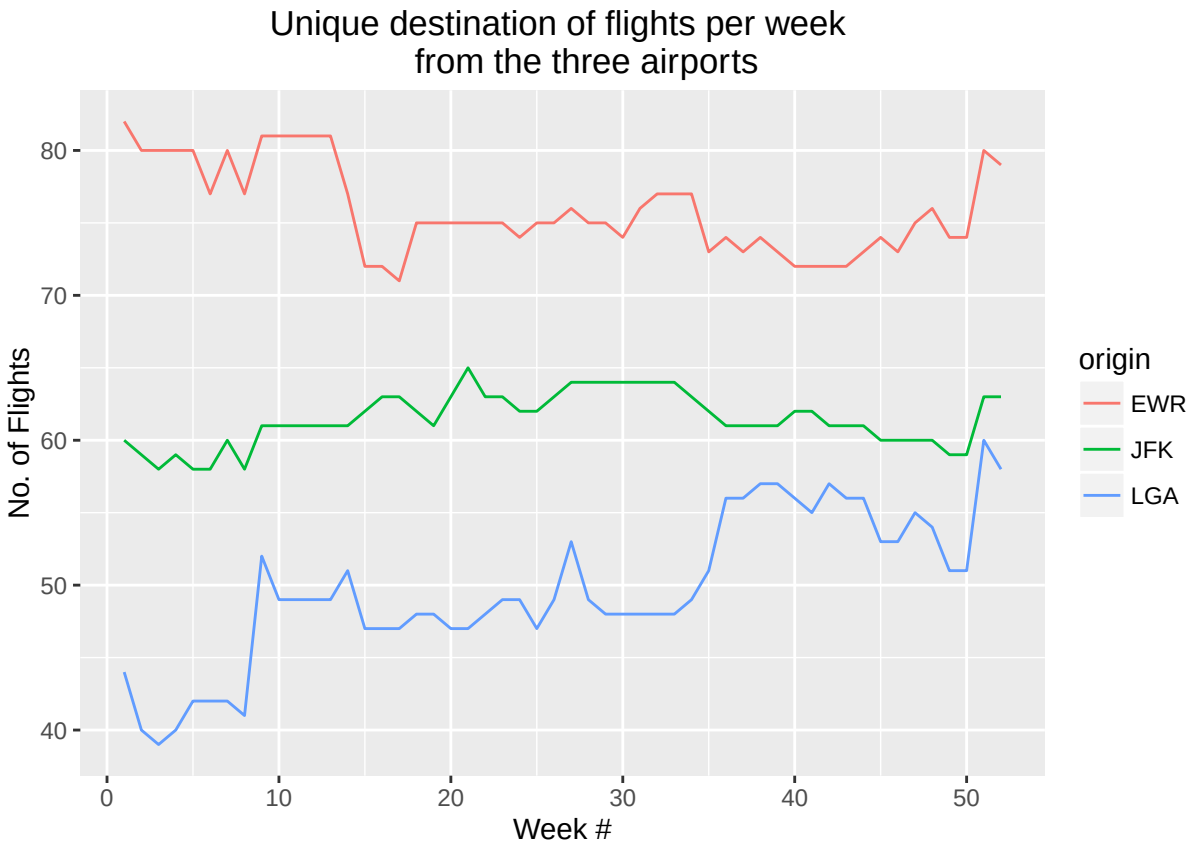
```
ggplot(carrierVSuniquedest, aes(carrier, unique_dest)) +
  geom_point(aes(color=carrier), size = 3.5)+
  labs(title = "Flights to distinct destinations based on Carrier
       from NYC in 2013",
          x = "Carrier", y = "No. of unique destinations") +
  theme(plot.title = element_text(colour = "black"))
```



From the plot, it can be seen that EV operated flights to the maximum number of unique destinations from NYC in 2013, followed by 9E and UA. F9, AS and HA operated flights only to one destination from NYC. To further this analysis, the number of unique destination of flights per week from the three airports are presented, as a time series, using the code chunk below.

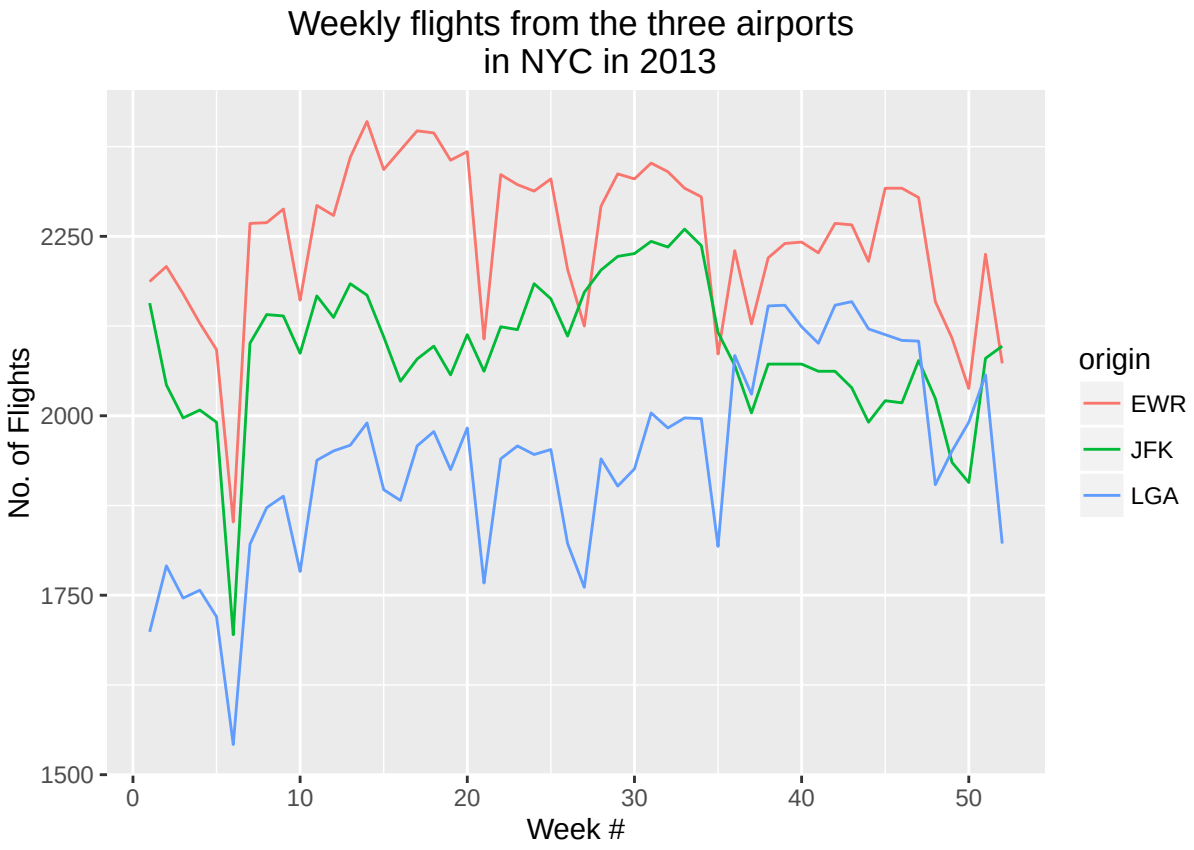
```
flightsByWeek<- flights.dataset %>%
  group_by(week=week(date), origin) %>%
  summarize(no_of_flights=n(), ndests=n_distinct(dest))

#display the graph for unique destinations from origin
flightsByWeek %>% filter(week<=52) %>%
  ggplot(aes(week, ndests, colour=origin)) +
  geom_line() +
labs(title = "Unique destination of flights per week
     from the three airports",
       x = "Week #", y = "No. of Flights") +
  theme(plot.title = element_text(colour = "black"))
```



Consistently, on a weekly basis, LGA operated the least number of flights to unique destinations and EWR operated the highest number of flights to unique destinations. However, there was a significant increase in the number of destinations from LGA between the weeks 35 and 45. To observe the impact of this increase (if any) on the total number of weekly flights out of the three airports, the following code is used to present a time series of the total weekly flights out o these three airports in NYC.

```
#display graph which shows no of flights from origin (week-wise)
#52 Weeks Total in a year
flightsByWeek %>% filter(week<=52) %>%
  ggplot(aes(week, no_of_flights, colour=origin)) +
  geom_line() +
labs(title = "Weekly flights from the three airports
     in NYC in 2013",
       x = "Week #", y = "No. of Flights") +
  theme(plot.title = element_text(colour = "black"))
```



The plot shows that during the weeks between 35 and 45, the total number of flights that departed from LGA were higher than the total number of flights that departed from JFK. This, to a reasonable extent, could be attributed to the increased number of destinations offered from LGA during this time frame.

Carrier VS Distance

- **Carrier VS Distance**

  The dataset is grouped by `carrier` and the mean, maximum, minimum and median `distance` of these groups are calculated. Also, a scatter plot is presented to visualize.

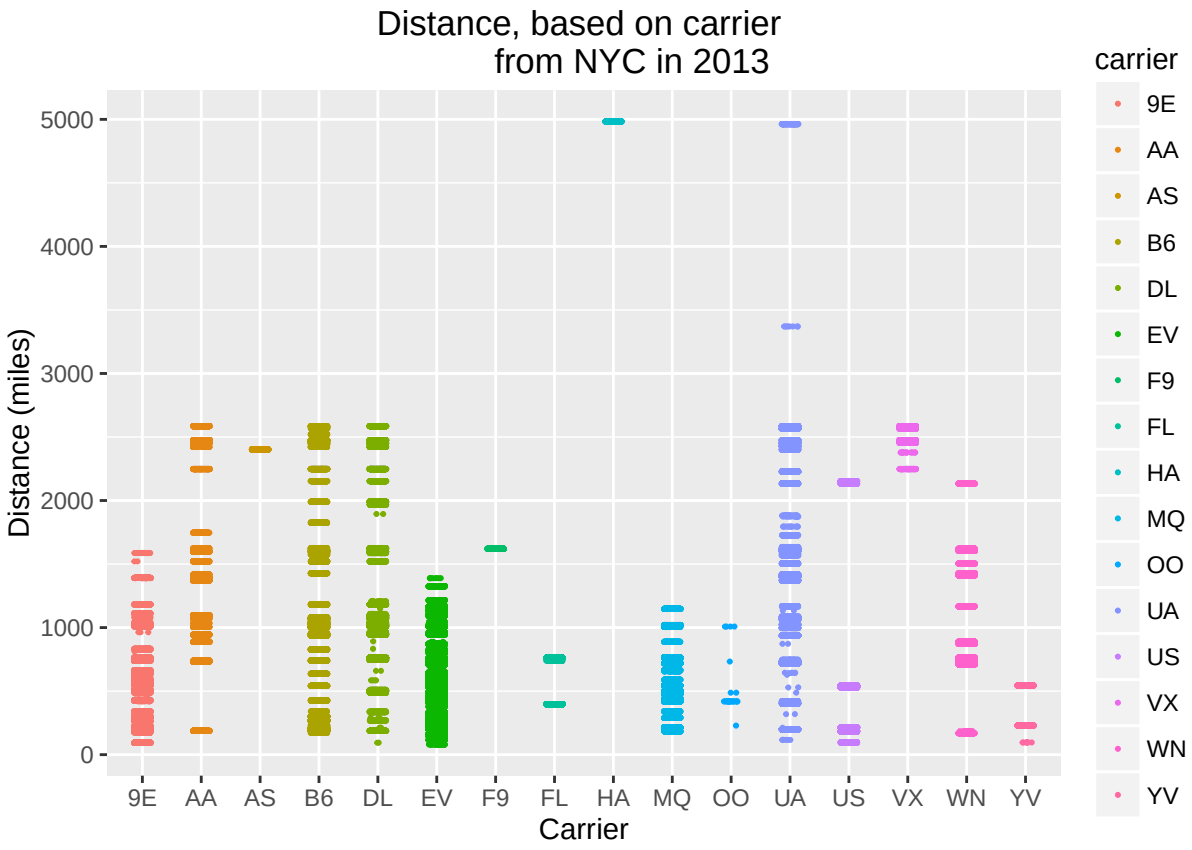
```
carrierVSdistance<-flights.dataset %>%
  group_by(carrier) %>%
  summarize(mean_distance=mean(distance, na.rm = TRUE),
            median_distance =median(distance, na.rm = TRUE),
            max_distance = max(distance, na.rm = TRUE,
            min_distance = min(distance, na.rm = TRUE)))

carrierVSdistance
```

```
## Source: local data frame [16 x 4]
##
##    carrier mean_distance median_distance max_distance
```

```
##     (fctr)         (dbl)           (dbl)        (dbl)
## 1       9E      529.8896             509         1587
## 2       AA     1343.2799            1096         2586
## 3       AS     2402.0000            2402         2402
## 4       B6     1069.6896            1023         2586
## 5       DL     1237.9791            1020         2586
## 6       EV      562.8650             533         1389
## 7       F9     1620.0000            1620         1620
## 8       FL      664.7874             762          762
## 9       HA     4983.0000            4983         4983
## 10      MQ      570.3746             502         1147
## 11      OO      509.2759             419         1008
## 12      UA     1531.3214            1400         4963
## 13      US      560.8259             529         2153
## 14      VX     2499.4326            2475         2586
## 15      WN      996.9714             748         2133
## 16      YV      376.4375             229          544
```

```
  ggplot(flights.dataset, aes(carrier, distance)) +
geom_point(aes(color=carrier), size = 0.4,
       position = position_jitter(width = 0.35))+
labs(title = "Distance, based on carrier
        from NYC in 2013",
        x = "Carrier", y = "Distance (miles)") +
theme(plot.title = element_text(colour = "black"))
```



The scatter plot, with jitter, shows the distance (in miles) of the flights operated by the various carriers. It can be seen that HA (Hawaiian Airlines) operated the longest flight, possibly to Hawaii, which happens to be its only flight. UA also operated a long haul flight travelling close to 5000 miles. If one looks at destinations less than 800 miles from NYC, EV and 9E operated the maximum number of flights for these distances. For any range of distances, the plot shows the carriers that operated a higher number of flights to a reasonably good extent.

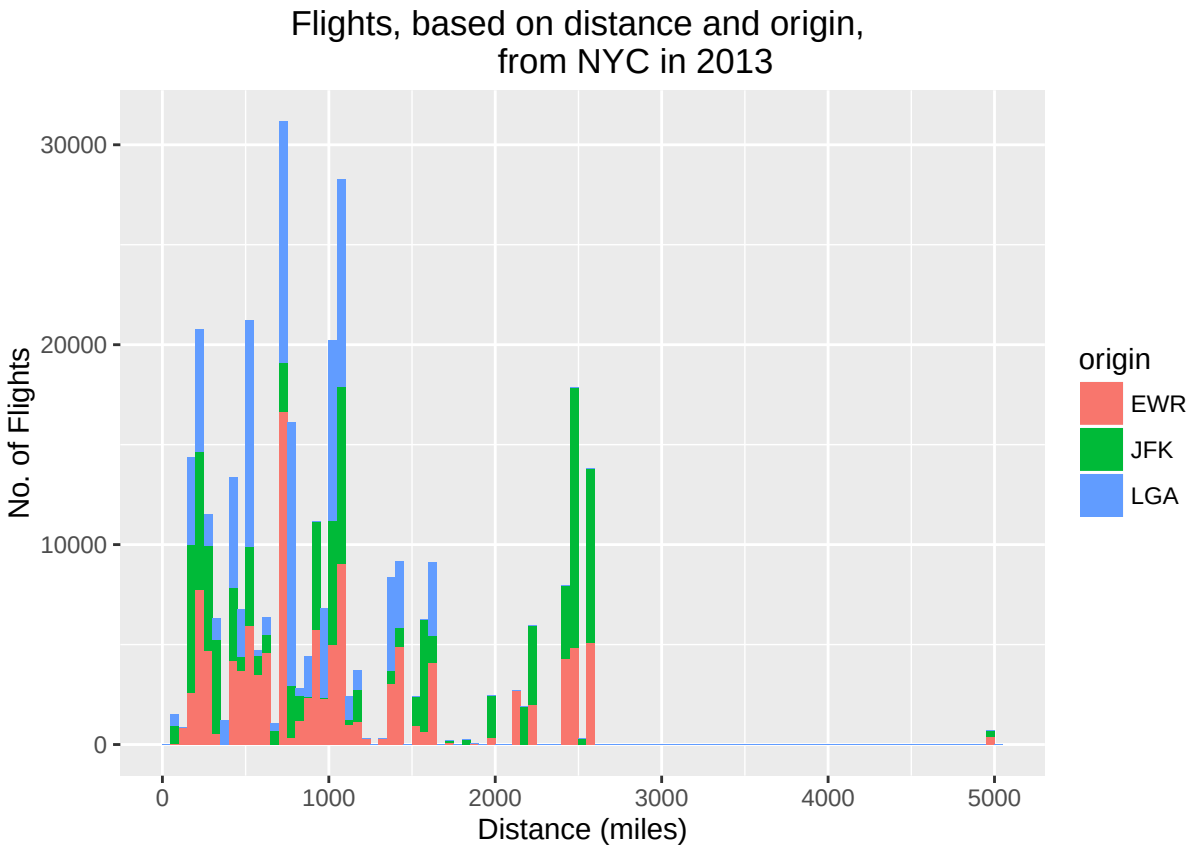
Distance VS Originating Airport

- **Distance VS Originating Airport**

  A bar plot showing the relationship between the distance of the flight and the originating airport is presented using the code below.

```
# Plot summary as a bar chart
    ggplot(data = flights.dataset, mapping = aes(x=distance))+
      geom_bar(aes(fill=origin), binwidth = 50) +
      labs(title = "Flights, based on distance and origin,
           from NYC in 2013",
           x = "Distance (miles)", y = "No. of Flights") +
      theme(plot.title = element_text(colour = "black"))
```

```
## Warning: `geom_bar()` no longer has a `binwidth` parameter. Please use
## `geom_histogram()` instead.
```



So, the bar plot (with a bin size of 50 miles) provides insight as to the choice of airports for long haul flights. It is clear that for distances greater than 1700 miles, either one of JFK or Newark airports were preferred. Between JFK and Newark, JFK was the airport that had more number of long haul flights. LGA, on an average had a higher number of flights that travelled less than 1700 miles.

Conclusion

We have prepared the dataset by performing various functions, analyzed it and finally showing the relationships we've come to a conclusion.

There were several interesting findings in this project they can be summerized as follows:

1. Flights out of NYC are uniformly distributed across the different months of the year, a small reduction is observed in February due to the fact that it had only 28 days in 2013
2. The number of flights tend to increase during the rush-hours of the day i.e 6am to 9am and 3pm to 7pm
3. United Airlines has the maximum number of flights flying out of the NYC area particularly from Newark International Airport followed by JetBlue Airways which operates majorly from JFK airport. Delta airlines (DL) majorly operates from LGA
4. Highest average delay is by Frontier Airlines

5. ExpressJet Airlines(EV) has the largest number of unique destinations on offer to passengers
6. From week number 35 to week 47 number of flights from JFK continued to decline. However number of flights from LGA experienced a sharp rise
7. For long haul flights the preferred airports are Newark and JFK. LGA is preferred for short distance flights