Analytics in Practice - Assignment 2: Data Visualization

Srushti Padade

April 10, 2020

Problem Statement

An Airport manager is receiving complaints from customers regarding the delays in departure and arrival of flights. He thinks that there are hardly any delays and even if there are any delays, it is mostly due to weather. But he is unable to convince his supervisor. He has now hired you, a Data Scientist, a graduate from Kent State University. Your job is to explain to the Airport manager about the overall delays at the Airport. You have been provided with complete air traffic data to analyze. Your task is to analyze the data and provide insights about the delays. Following are the set of questions, he is looking for an answer from you.

- To Solve the problem faced by the Airport manager we are using Data Visualization techniques to make it easier for the management to have a clear idea of the situation.
- The libraries used would be as below for the data wrangling as well as the data visualization.

```
library(ggplot2)
library(dplyr)
```

```
##
## 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(tidyr)

options(warn = -1)
```

- Here we are loading the data set with the Flight details of the California State in United States.
- The data contained in the compressed file has been extracted from the Marketing Carrier On-Time Performance (Beginning January 2018) data table of the "On-Time" database from the TranStats data library.

```
FlightData <- read.csv("Sample_CA_airtraffic_delays.csv")
dim(FlightData)
```

```
## [1] 6934  123
```

1. What is the pattern of arrival traffic and departure traffic delays with respect to days and weeks?

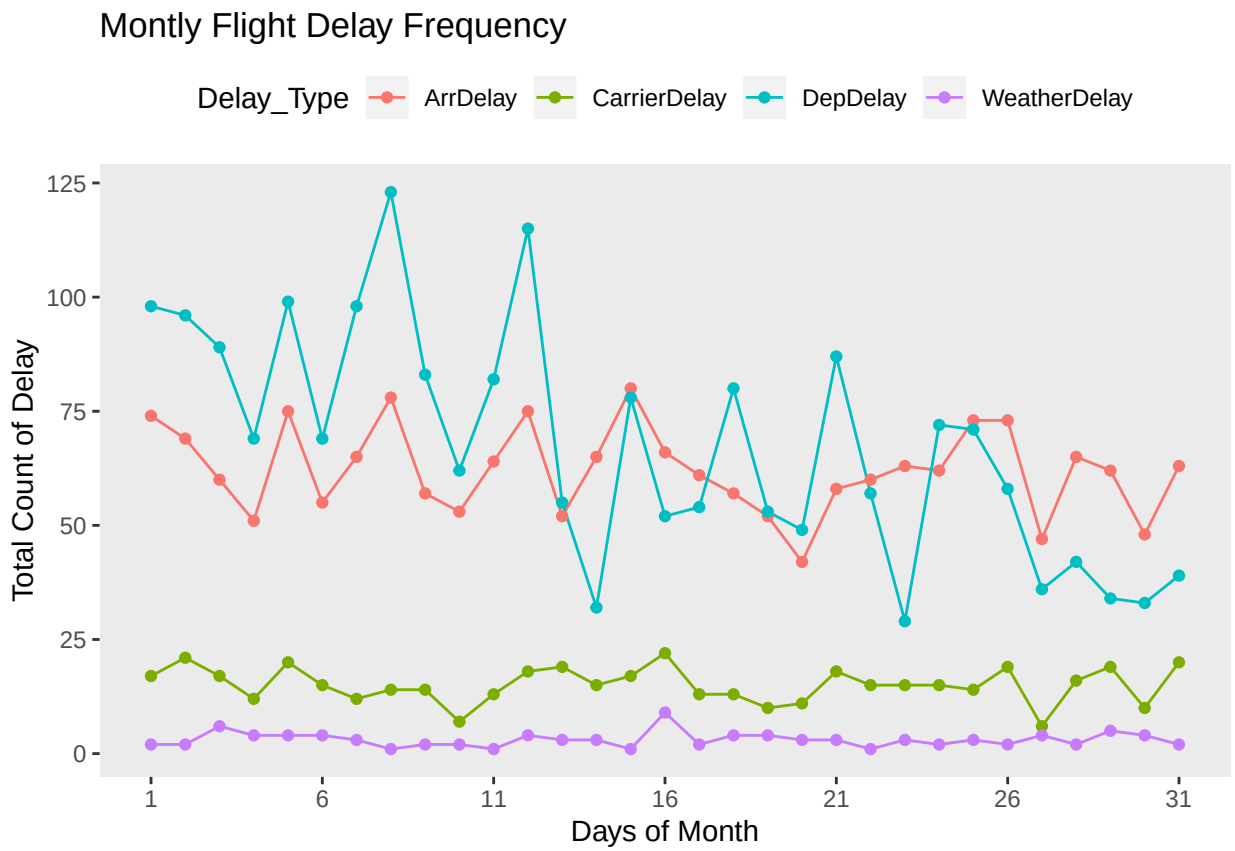
```
DelayOnMonths <- FlightData %>% select("DayofMonth","DepDelay","ArrDelay", "CarrierDelay","WeatherDelay
  filter(Delay_Time > 0) %>% group_by(DayofMonth,Delay_Type) %>% summarise(Frequency = n())

head(DelayOnMonths)
```

```
## # A tibble: 6 x 3
## # Groups:   DayofMonth [2]
##   DayofMonth Delay_Type   Frequency
##        <int> <chr>            <int>
## 1          1 ArrDelay            74
## 2          1 CarrierDelay        17
## 3          1 DepDelay            98
## 4          1 WeatherDelay         2
## 5          2 ArrDelay            69
## 6          2 CarrierDelay        21
```

Answer:

- The above table is a glimpes of data that is wrangled to extract the frequency of arrival as well as departure traffic delay by filtering it from the early arrived or departed flights on DAYS of MONTHs bases.
- Also I have included Carrier and weather type delay where PURPLE represents Weather Delay which is very rare and GREEN represents Carrier Delay which also is minimal.

```
ggplot(DelayOnMonths, aes(DayofMonth, Frequency)) +
  geom_line(aes(color = Delay_Type)) +
  labs(title = "Montly Flight Delay Frequency",  x = "Days of Month",  y ="Total Count of Delay") +
  geom_point(aes(color = Delay_Type)) +
  scale_x_continuous(breaks = seq(1, 31, by = 5)) +
  theme(legend.position="top", panel.grid.major = element_blank(), panel.grid.minor = element_blank())
```



- The graph displayed above represents the pattern of how the Arrival & Departure Traffic Delay is featured based on month.
- As we can see the plot there are 2 lines, RED representing Arrival Traffic Delay pattern and BLUE line representing Departure Traffic Delay pattern.
- Here we observe that the Departure traffic delay frequency varies a lot than that of Arrival Traffic. Also the delay is likely to be more at the begining of the month and then we see delay frequency going down by the end of the month.
- Below is the data that tells us the story about the Flight Delay based on DAYS OF WEEK.

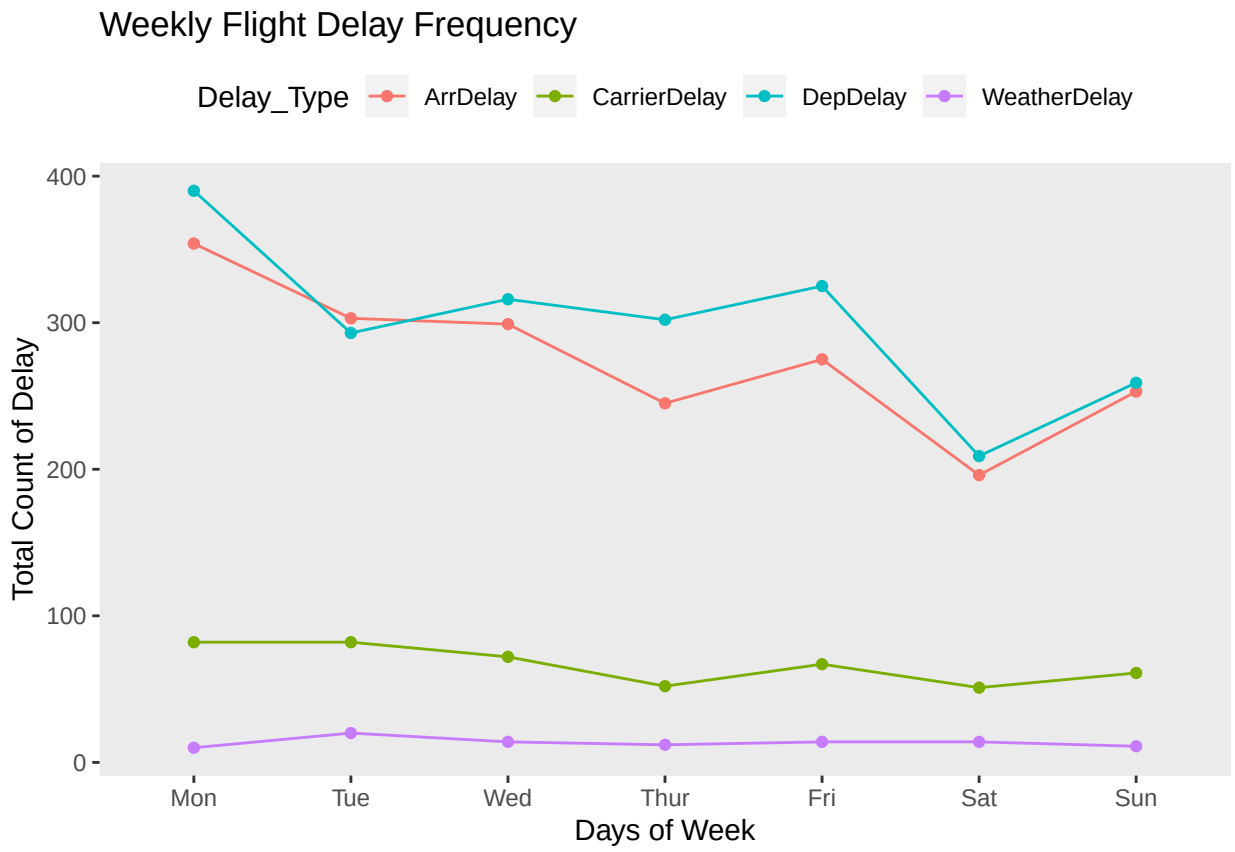
```
DelayOnWeeks <- FlightData %>% select("DayOfWeek","DepDelay","ArrDelay", "CarrierDelay","WeatherDelay")
  gather("Delay_Type","Delay_Time",-1) %>%  filter(Delay_Time > 0) %>%
  group_by(DayOfWeek,Delay_Type) %>% summarise(Frequency = n())

head(DelayOnWeeks)
```

```
## # A tibble: 6 x 3
## # Groups:   DayOfWeek [2]
##   DayOfWeek Delay_Type   Frequency
##       <int> <chr>            <int>
## 1         1 ArrDelay           354
## 2         1 CarrierDelay        82
## 3         1 DepDelay           390
## 4         1 WeatherDelay        10
```

```
## 5          2 ArrDelay           303
## 6          2 CarrierDelay        82
```

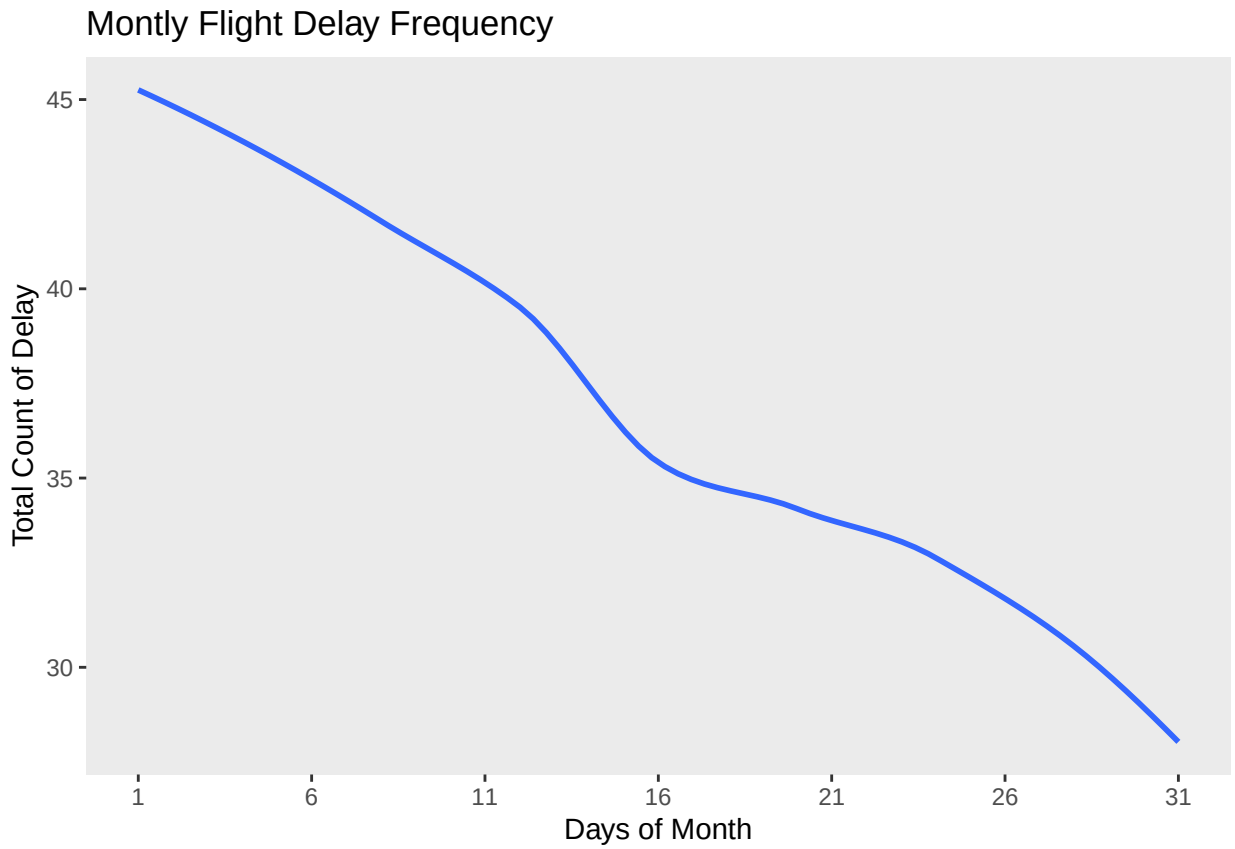
- The plot shows the frequency count of Arrival Traffic delay in RED and Departure Traffic Delay in BLUE.
- Based on the plot we can conclude that the delay rate is much higher at the start of the week and slowly goes down as the week comes to an end. Also the Departure Delay is much higher than Arrival Delay.
- Also i have included Carrier and weather type delay where PURPLE represents Weather Delay which is very rare and GREEN represents Carrier Delay which also is minimal.

```
ggplot(DelayOnWeeks, aes(DayOfWeek, Frequency)) +
  geom_point(aes(color = Delay_Type)) + geom_line(aes(color = Delay_Type)) +
  labs(title = "Weekly Flight Delay Frequency",  x = "Days of Week",  y ="Total Count of Delay") +
  scale_x_discrete(breaks=c("1","2","3", "4", "5", "6","7"),
  labels=c("Mon", "Tue", "Wed", "Thur", "Fri", "Sat", "Sun"), limits = c(1,2,3,4,5,6,7)) +
  theme(legend.position="top", panel.grid.major = element_blank(), panel.grid.minor = element_blank())
```

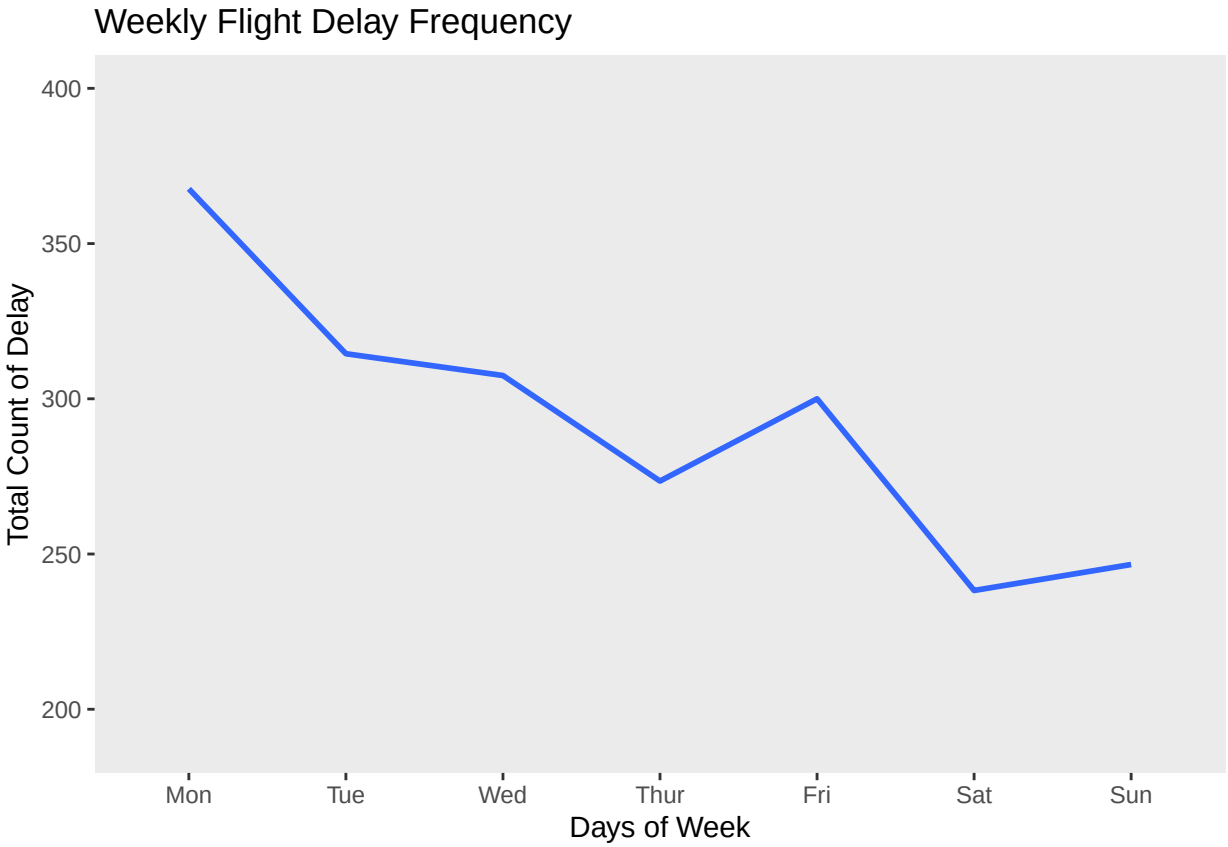


2. Can you interpret the traffic delays?

```
# Montly Delay
ggplot(DelayOnMonths, aes(DayofMonth, Frequency)) +
  labs(title = "Montly Flight Delay Frequency",  x = "Days of Month",  y ="Total Count of Delay") +
  geom_smooth(method = "loess", se = FALSE) +
  scale_x_continuous(breaks = seq(1, 31, by = 5)) +
  theme(legend.position="top", panel.grid.major = element_blank(), panel.grid.minor = element_blank())
```




```
# Weekly Delay
ggplot(DelayOnWeeks, aes(DayOfWeek, Frequency)) +
  geom_smooth(method = "loess", se = FALSE, size = 1) +
  labs(title = "Weekly Flight Delay Frequency",  x = "Days of Week",  y ="Total Count of Delay") +
  scale_x_discrete(breaks=c("1","2","3", "4", "5", "6","7"),
  labels=c("Mon", "Tue", "Wed", "Thur", "Fri", "Sat", "Sun"), limits = c(1,2,3,4,5,6,7)) +
  ylim(190, 400) +
  theme(legend.position="top", panel.grid.major = element_blank(), panel.grid.minor = element_blank())
```



Answer:

- Considering the same data wrangled todetermine traffic delay for days of week and days of month above graphs are interpreted.
- As we look at the graph we can see the clear picture of how the Arrival and departure traffic delay have occured over the Days of the Month and Days of week.
- We can say that the delay in approximately higher at the start of the Month i.e. dated as 1st of the month or it is Monday of the week.
- As the days pass by the delay frequency is less by the end of the Month 30th or 31th of Month or Saturday/Sunday of the Weeks.
- Thus we can interpret the Highest delay frequency of more than 350 times on Mondays and as high as 80 times on 1st of the Months in 2018 in California.

3. Which Airport ('Origin Airport') has highest departure delay?

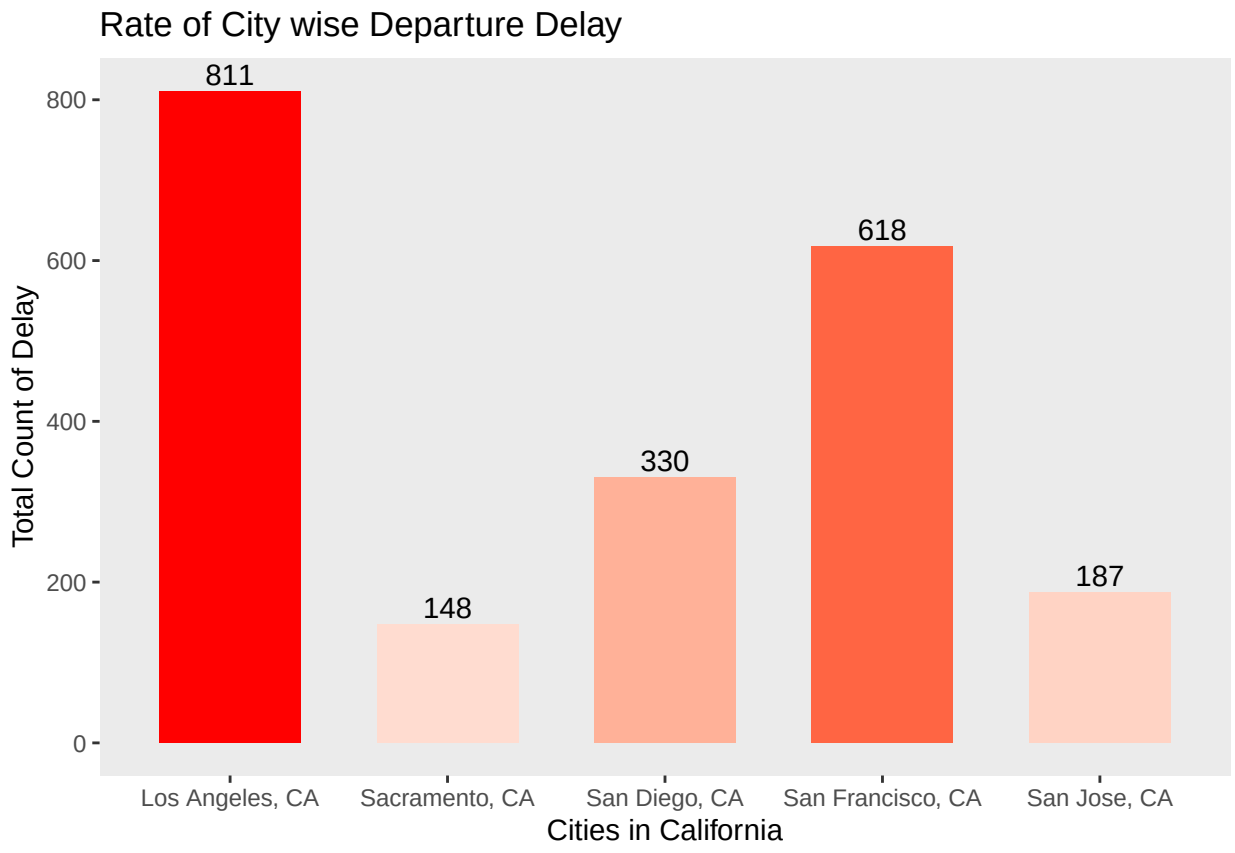
```
Departure_Delay <- FlightData  %>% filter(DepDelay > 0)%>% select("Origin","OriginCityName") %>%
  group_by(OriginCityName) %>% summarise(Count=n())

Departure_Delay
```

```
## # A tibble: 5 x 2
##   OriginCityName    Count
##   <fct>             <int>
## 1 Los Angeles, CA     811
## 2 Sacramento, CA      148
## 3 San Diego, CA       330
## 4 San Francisco, CA   618
## 5 San Jose, CA        187
```

Answer:

- By some data extraction we have found the total count of Flight departure delay among all the California City Airports.
- Below graph represents the scale at which the Departure Delay frequency is based on the Cities. Here, Los Angles have highest departure delay and Lowest in Sacramento.

```
ggplot(Departure_Delay, aes(OriginCityName, Count)) +
  geom_col(aes(fill = Count), width = 0.65) +
  scale_fill_gradient2(low="white", high="red") +
  geom_text(aes(OriginCityName, Count, label = Count), vjust = -0.3) +
  theme(legend.position = "none", panel.grid.major = element_blank(), panel.grid.minor = element_blank()
  labs(title = "Rate of City wise Departure Delay",  x = "Cities in California",  y ="Total Count of Del
```



4. Which Airport has highest Arrival delay?

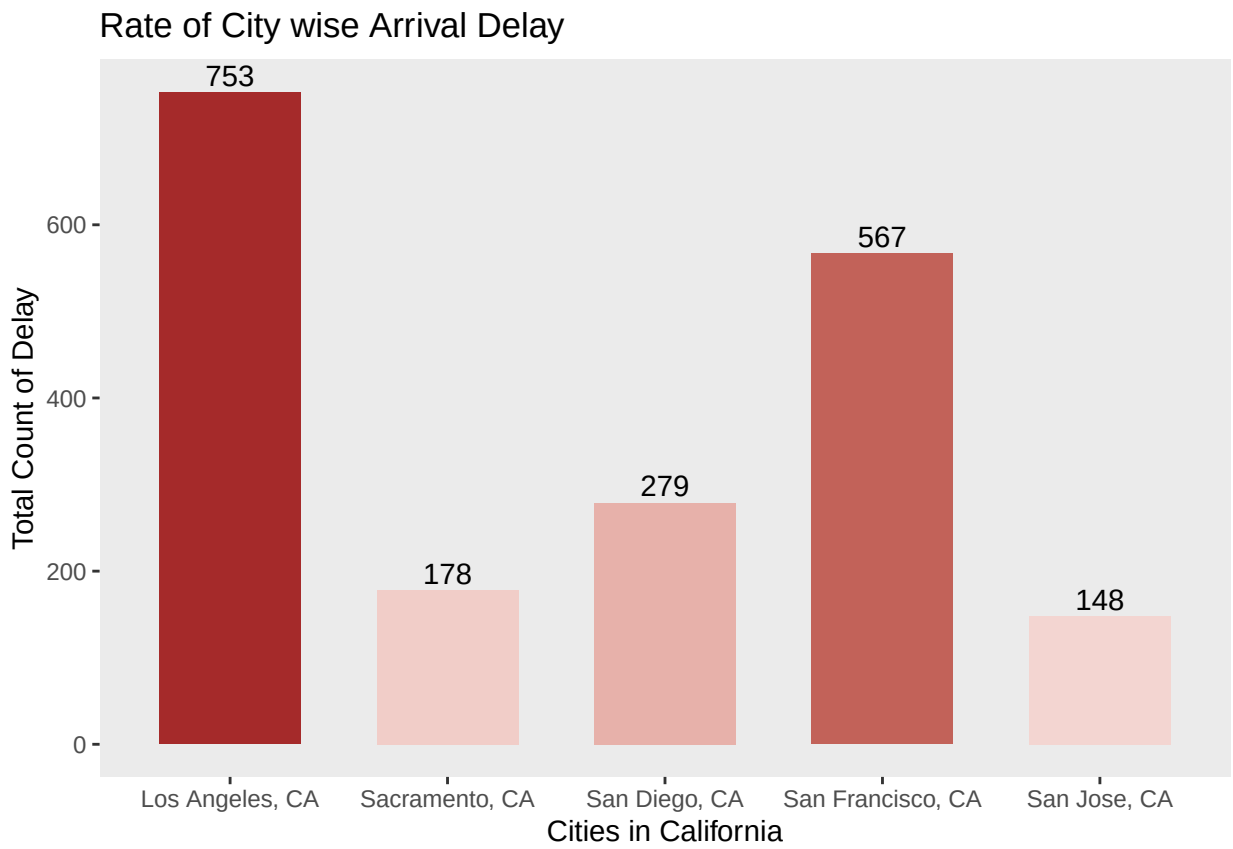
```
Arrival_Delay <- FlightData  %>% filter(ArrDelay > 0)%>% select("Origin","OriginCityName") %>%
  group_by(OriginCityName) %>% summarise(Count=n())

Arrival_Delay
```

```
## # A tibble: 5 x 2
##   OriginCityName    Count
##   <fct>             <int>
## 1 Los Angeles, CA     753
## 2 Sacramento, CA      178
## 3 San Diego, CA       279
## 4 San Francisco, CA   567
## 5 San Jose, CA        148
```

Answer

- Similar to the Departure delay, after data extraction we have found the total count of Flight arrival delay among all the California City Airports.
- Below graph represents the scale at which the Arrival Delay frequency is based on the Cities. Here, Los Angles have highest arrival delay and Lowest in San Jose.

```
ggplot(Arrival_Delay, aes(OriginCityName, Count)) +
  geom_col(aes(fill = Count), width = 0.65) +
  scale_fill_gradient2(low="white", high="brown") +
  geom_text(aes(OriginCityName, Count, label = Count), vjust = -0.3) +
  theme(legend.position = "none", panel.grid.major = element_blank(), panel.grid.minor = element_blank()
  labs(title = "Rate of City wise Arrival Delay",  x = "Cities in California",  y ="Total Count of Dela
```



5. How do you relate the delay pattern to the distance travelled?

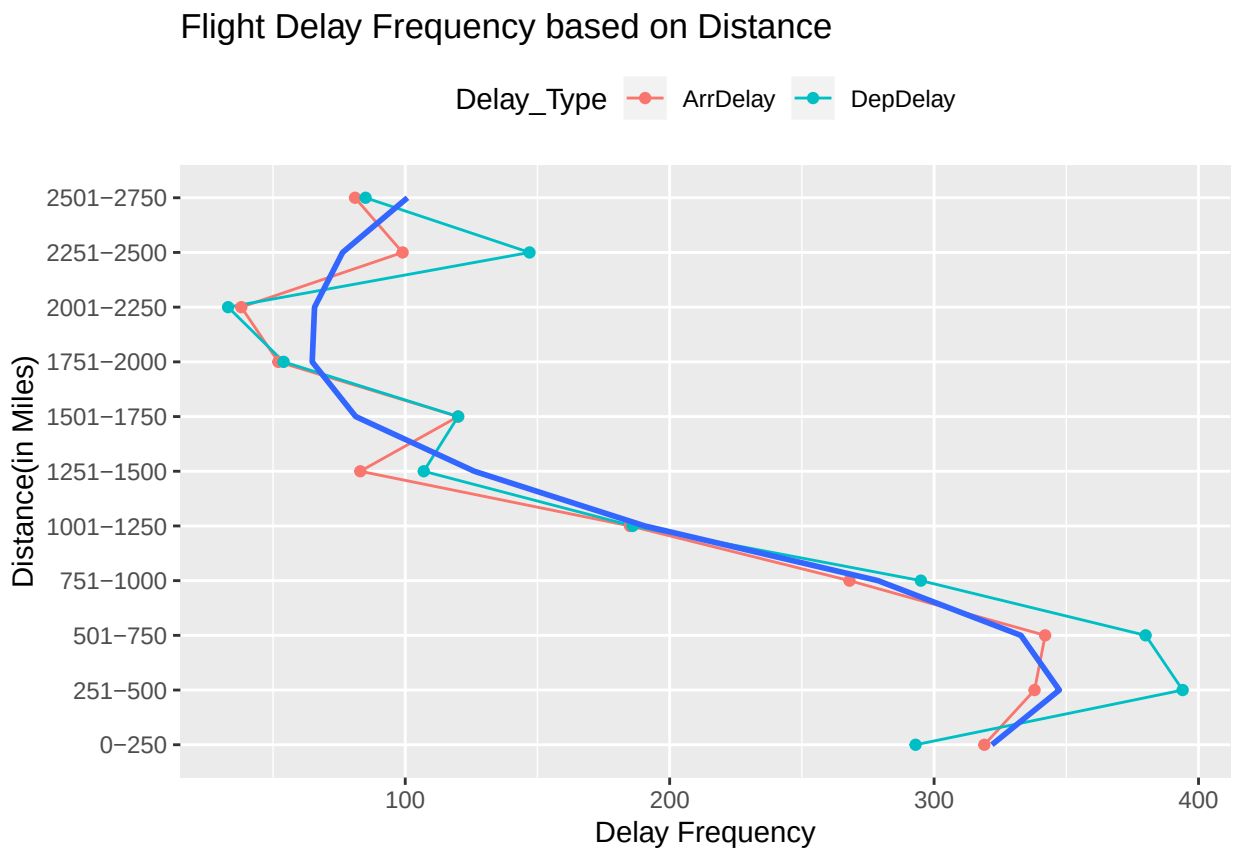
```
DelayByDistance <- FlightData %>% select("DistanceGroup","DepDelay","ArrDelay") %>%
  gather("Delay_Type","Delay_Time", -1)%>% group_by(DistanceGroup,Delay_Type)%>% filter(Delay_Time > 0)
  group_by(DistanceGroup,Delay_Type) %>% summarise(Frequency = n())

head(DelayByDistance)
```

```
## # A tibble: 6 x 3
## # Groups:   DistanceGroup [3]
##   DistanceGroup Delay_Type Frequency
##           <int> <chr>          <int>
## 1             1 ArrDelay         319
## 2             1 DepDelay         293
## 3             2 ArrDelay         338
## 4             2 DepDelay         394
## 5             3 ArrDelay         342
## 6             3 DepDelay         380
```

Answer

- We have gathered the group of distances travelled by the Flights and the filtered them based on the Traffic delay for both Arrival and departure flights to know the frequency of flight delay at certain distance.
- The below graph displays the Distance to type of Flight delay in Miles.
- RED line graph displays the Arrival delay pattern based on distance whereas BLUE plot displays the departure delay.
- The DARK BLUE line shows the pattern of how exactly the flight delay is affected based on the distance travelled.
- Thus looking at the plot we can conclude that the flights with nearby locations are facing highers delays and as the distance increases the delay in flight become lesser and lesser in an approximation.
- The highest delay is over the 250 to 750 miles.

```
ggplot(DelayByDistance, aes(DistanceGroup, Frequency)) +
  geom_line(aes(color = Delay_Type)) + geom_point(aes(color = Delay_Type)) +
  labs(title = "Flight Delay Frequency based on Distance",  x = "Distance(in Miles)",  y ="Delay Freque
  geom_smooth(method = "loess", se = FALSE, size = 1) + theme(legend.position="top") +
  scale_x_discrete(limit=c(1:11), labels=c("0-250", "251-500", "501-750", "751-1000", "1001-1250", "125
  coord_flip()
```



6. Is there any correlation between weather delay and carrier delay?

```
DataWC <- na.omit(cbind.data.frame(Weather = round(FlightData$WeatherDelay/60), Carrier = round(FlightDa
summary(DataWC)
```

```
##     Weather            Carrier
##  Min.   : 0.00000   Min.   : 0.0000
##  1st Qu.: 0.00000   1st Qu.: 0.0000
##  Median : 0.00000   Median : 0.0000
##  Mean   : 0.07285   Mean   : 0.3254
##  3rd Qu.: 0.00000   3rd Qu.: 0.0000
##  Max.   :14.00000   Max.   :13.0000
```

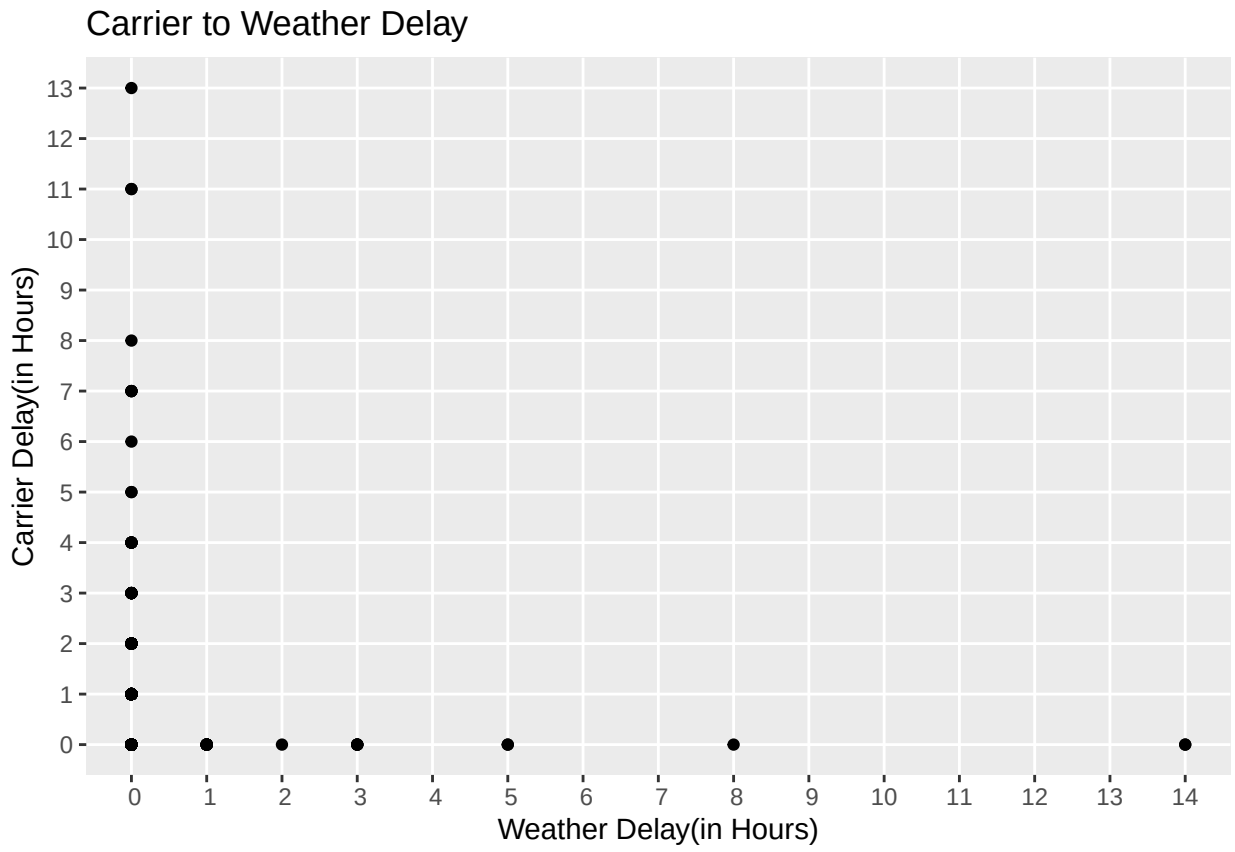
```
Correlation <- cor(DataWC)
Correlation
```

```
##             Weather    Carrier
## Weather  1.0000000 -0.0315184
## Carrier -0.0315184  1.0000000
```

Answer

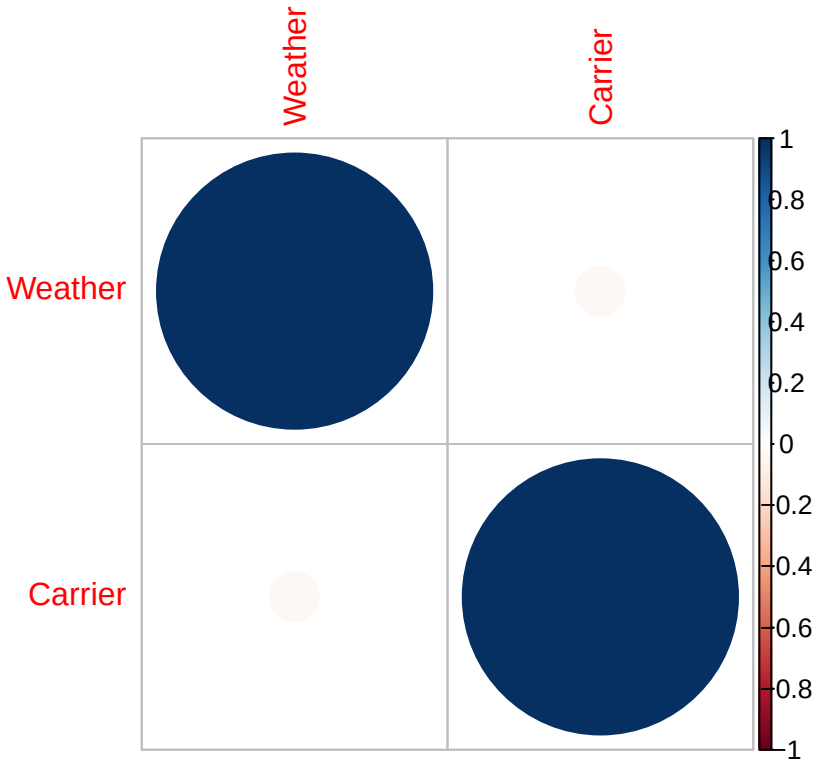
- Looking at the raw data it is very difficult to extract any information. Hence we have altered the data by changing the time frame from minutes to hours.
- Also I have computed the correlation value between the Carrier Delay and Weather Delay, which is -0.0315184 i.e. they are negatively correlated.
- Thus by looking at the graph we could see the the the carrier delay and the weather delay gives a right angle graph and doesn't give much information.

```
ggplot(DataWC, aes(Weather, Carrier)) + geom_point() + scale_x_discrete(limit=c(0:14))+ scale_y_discrete
  labs(title = "Carrier to Weather Delay", x = "Weather Delay(in Hours)",  y ="Carrier Delay(in Hours)"
```



- Also by calculatinng the correlation value of the data, below correlation plot is drawn.

```
corrplot::corrplot(Correlation)
```



7. What is the delay pattern you can find in respective states?

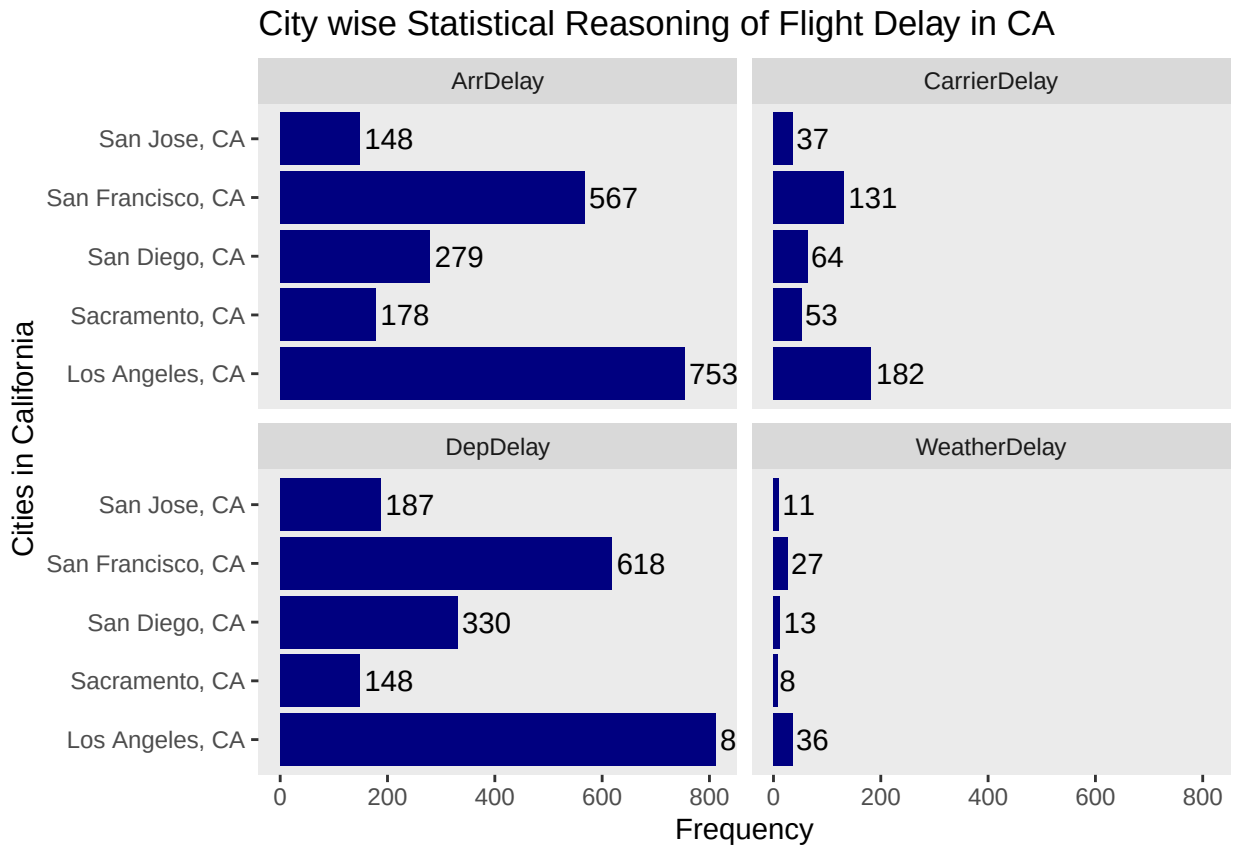
```
DelayPatterns <- FlightData %>% select("OriginCityName","ArrDelay","DepDelay","CarrierDelay","WeatherDel
  gather("Delay_Type","Delay_Time",-1) %>% filter(Delay_Time > 0) %>%
  group_by(OriginCityName,Delay_Type)  %>% summarise(Frequency=n())

head(DelayPatterns)
```

```
## # A tibble: 6 x 3
## # Groups:   OriginCityName [2]
##   OriginCityName  Delay_Type   Frequency
##   <fct>           <chr>            <int>
## 1 Los Angeles, CA ArrDelay           753
## 2 Los Angeles, CA CarrierDelay       182
## 3 Los Angeles, CA DepDelay           811
## 4 Los Angeles, CA WeatherDelay        36
## 5 Sacramento, CA  ArrDelay           178
## 6 Sacramento, CA  CarrierDelay        53
```

Answer

- To extract delay patterns we have computed the delay frequency in each city of California state based on the delay types.
- Above is a glimps of the data that is extracted.
- Thus, looking at the result below graph is computed having a grid of 4 parts each having a delay pattern of eachh city based on the type of delay.
- Hence we see the major delay is occured due to the Departure delay and in the city of Los Angeles.

```
ggplot(DelayPatterns, aes(OriginCityName, Frequency)) +
  geom_bar(stat = "identity", position = position_stack(reverse = TRUE), fill="navy") +
  labs(title = "City wise Statistical Reasoning of Flight Delay in CA",  x = "Cities in California",  y
  geom_text(aes(OriginCityName, Frequency, label = Frequency),  hjust = -0.1) +
  theme(legend.position="top", panel.grid.major = element_blank(), panel.grid.minor = element_blank()) +
  facet_wrap("Delay_Type") + coord_flip()
```



8. How many delayed flights were cancelled? (approximation)

```
Cancelled <- FlightData %>% select("Cancelled","ArrDelay","DepDelay","CarrierDelay","WeatherDelay") %>%
  gather("Delay_Type","Delay_Time",-1)%>%  filter(Delay_Time > 0) %>%
```

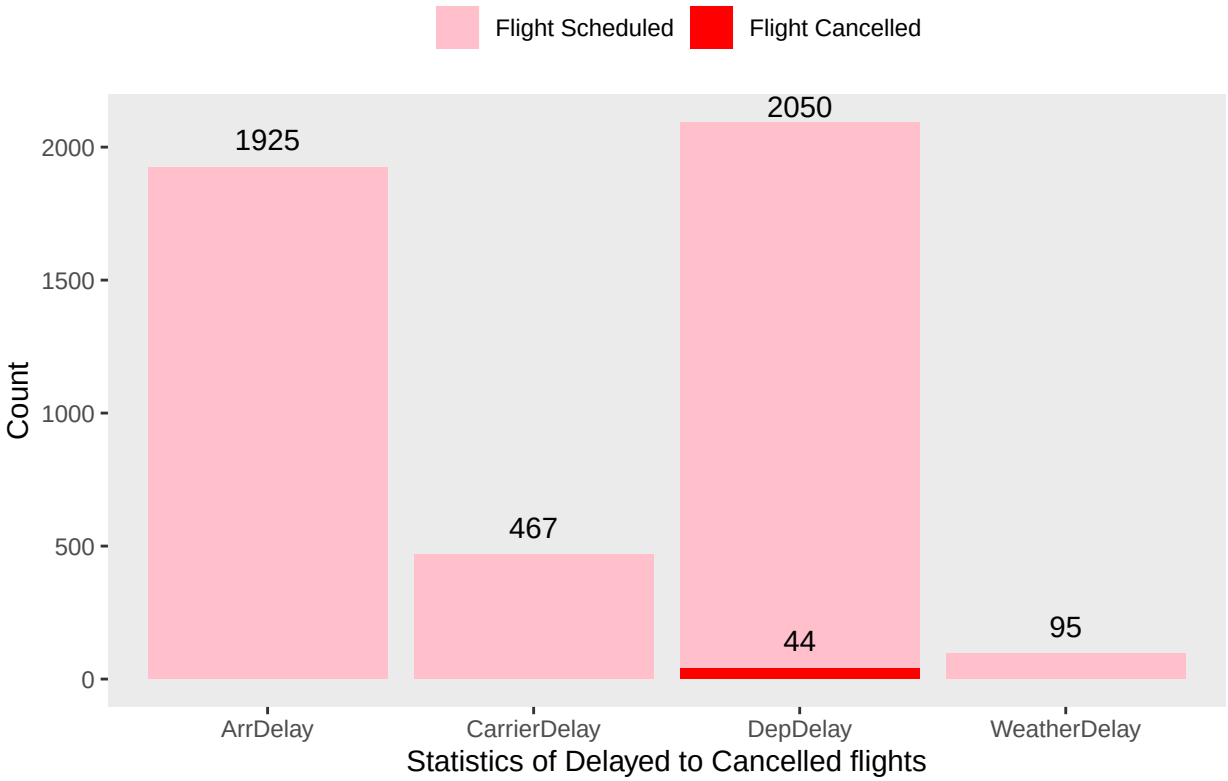
```
  group_by(Cancelled,Delay_Type) %>% summarise(Count = n())

head(Cancelled)
```

```
## # A tibble: 5 x 3
## # Groups:   Cancelled [2]
##   Cancelled Delay_Type   Count
##       <int> <chr>        <int>
## 1         0 ArrDelay      1925
## 2         0 CarrierDelay   467
## 3         0 DepDelay      2050
## 4         0 WeatherDelay    95
## 5         1 DepDelay        44
```

Answer

- We have computed the total count of Delayed flights that were cancelled eventually.
- The below graph displayed all the type of Delays and due to which the actual count of Cancelled flights can be distinguished.
- The RED plot determines the Cancelled flights after the delay. Here total of 44 flights were cancelled which were originally having Departure delay.
- Hence we have focus mainly on the Cancelled flights in the plot.

```
ggplot(Cancelled, aes(Delay_Type, Count)) + geom_col(aes(fill = Cancelled ==1)) +
  geom_text(aes(Delay_Type, Count, label = Count),  vjust = -0.8) +
  theme(legend.position="top", panel.grid.major = element_blank(), panel.grid.minor = element_blank()) +
  labs(title = "",  x = "Statistics of Delayed to Cancelled flights",  y ="Count") +
  scale_fill_manual(name = " ", values = c('pink', 'red'), breaks = c("FALSE","TRUE"), label = c("Fligh
```



9. How many delayed flights were diverted? (approximation)

```
Diverted <- FlightData %>% select("Diverted","ArrDelay","DepDelay","CarrierDelay","WeatherDelay") %>%
  gather("Delay_Type","Delay_Time",-1)%>%  filter(Delay_Time > 0) %>% group_by(Diverted,Delay_Type) %>%

head(Diverted)
```

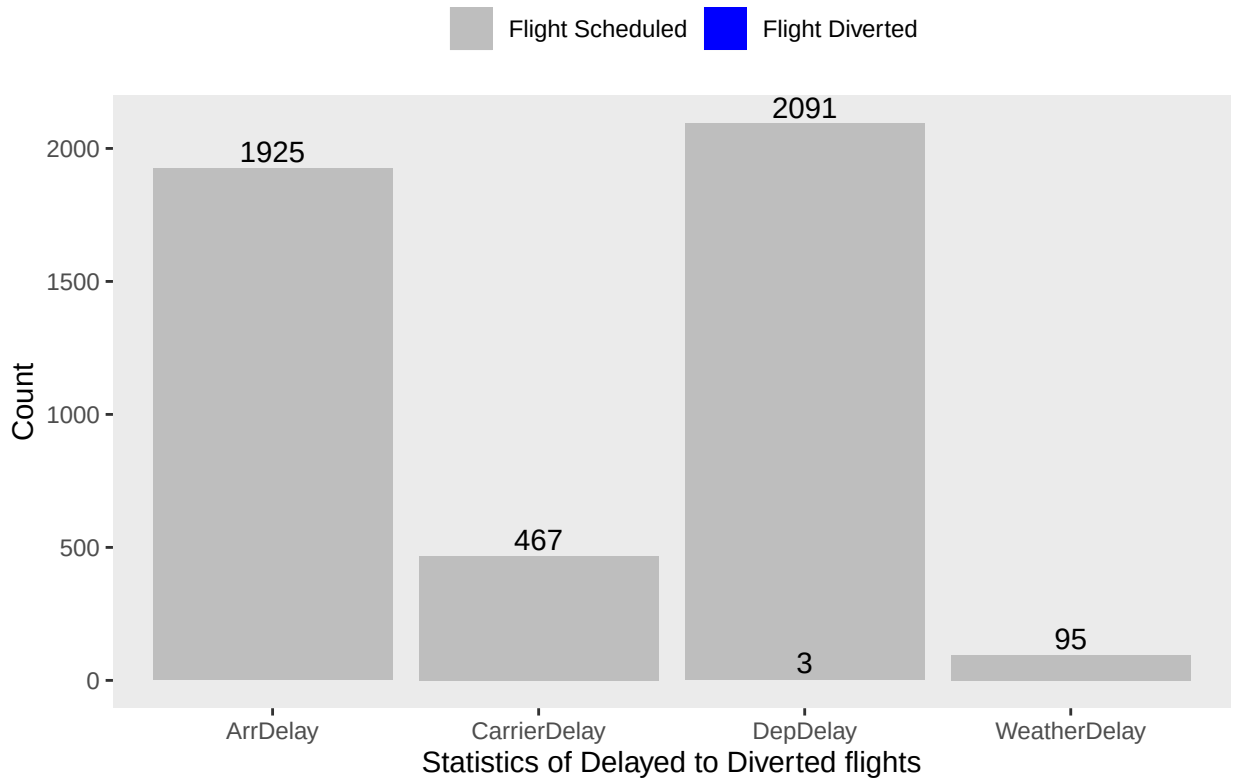
```
## # A tibble: 5 x 3
## # Groups:   Diverted [2]
##   Diverted Delay_Type   Count
##      <int> <chr>        <int>
## 1        0 ArrDelay      1925
## 2        0 CarrierDelay   467
## 3        0 DepDelay      2091
## 4        0 WeatherDelay    95
## 5        1 DepDelay         3
```

Answer

- We have computed the total count of Delayed flights that were diverted eventually.

- The below graph displayed all the type of Delays and due to which the actual count of Diverted flights can be distinguished.
- The RED plot determines the Cancelled flights after the delay. Here total of 44 flights were diverted which were originally having Departure delay.
- Hence we have focus mainly on the Diverted flights in the plot.

```
ggplot(Diverted, aes(Delay_Type, Count)) + geom_col(aes(fill = Diverted == 1)) +
  geom_text(aes(Delay_Type, Count, label = Count),  vjust = -0.3) +
  theme(legend.position="top", panel.grid.major = element_blank(), panel.grid.minor = element_blank()) +
  labs(title = "",  x = "Statistics of Delayed to Diverted flights",  y ="Count") +
  scale_fill_manual(name = " ", values = c('grey', 'blue'), breaks = c("FALSE","TRUE"), label = c("Flig
```



10. What time of the day do you find Arrival delays?

```
Arrival_DelayByTime <- FlightData  %>% filter(ArrDelay > 0)%>% select("ArrDelay","ArrTimeBlk") %>%
  group_by(ArrTimeBlk) %>% summarise(Count=n())

head(Arrival_DelayByTime)
```

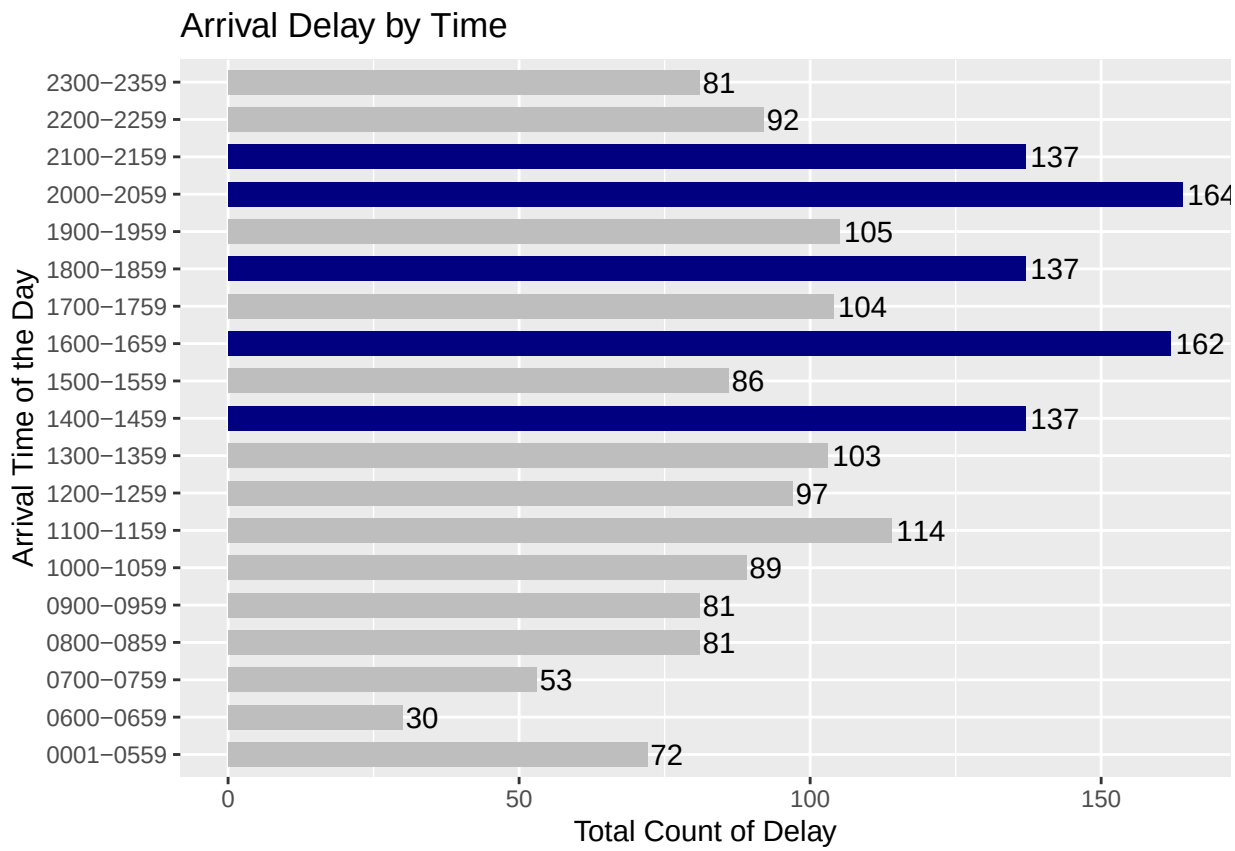
```
## # A tibble: 6 x 2
##   ArrTimeBlk Count
```

```
##   <fct>      <int>
## 1 0001-0559     72
## 2 0600-0659     30
## 3 0700-0759     53
## 4 0800-0859     81
## 5 0900-0959     81
## 6 1000-1059     89
```

Answer

- The data is filtered out in a way where we extract the count of arrival delay at particular time of the day.
- Thus the below graph can help us display a clear picture on what time of the day the arrival delay is prominent.
- To give a focus point for the management to take decisions I have highlighted the top Arrival delay timestamps.
- Thus we see that the maximum delay on arrival traffic is during the 8 to 10 PM whereas it is minimum during the late night and earlt mornings.

```
ggplot(Arrival_DelayByTime, aes(ArrTimeBlk, Count)) +
  geom_col(aes(fill = Count > 125), width = 0.65) +
  scale_fill_manual(name = " ", values = c('grey', 'navy'))+
  geom_text(aes(ArrTimeBlk, Count, label = Count), hjust = -0.1) +
  theme(legend.position = "none") + coord_flip() +
  labs(title = "Arrival Delay by Time",  x = "Arrival Time of the Day",  y ="Total Count of Delay")
```



11. What time of the day do you find Departure delays?

```
Departure_DelayByTime <- FlightData  %>% filter(DepDelay > 0)%>% select("DepDelay","DepTimeBlk") %>%
  group_by(DepTimeBlk) %>% summarise(Count=n())

head(Departure_DelayByTime)
```

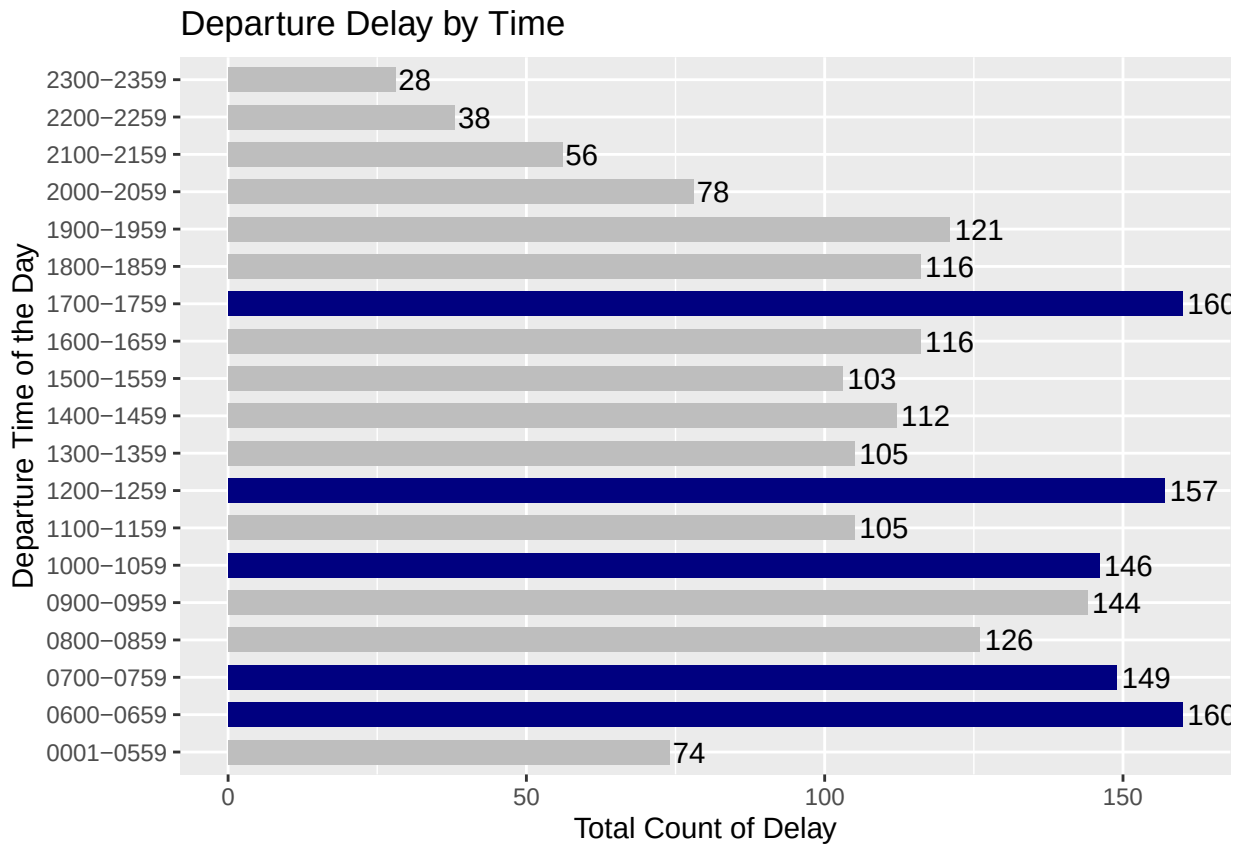
```
## # A tibble: 6 x 2
##   DepTimeBlk Count
##   <fct>      <int>
## 1 0001-0559     74
## 2 0600-0659    160
## 3 0700-0759    149
## 4 0800-0859    126
## 5 0900-0959    144
## 6 1000-1059    146
```

Answer

- The data is filtered out in a way where we extract the count of departure delay at particular time of the day.

- Thus the below graph can help us display a clear picture on what time of the day the arrival delay is prominent.
- To give a focus point for the management to take decisions I have highlighted the top Arrival delay timestamps.
- Thus we see that the departure delay takes place at anytime time of the day, not a particular part of the day but throughout.

```
ggplot(Departure_DelayByTime, aes(DepTimeBlk, Count)) +
  geom_col(aes(fill = Count > 145), width = 0.65) +
  scale_fill_manual(name = " ", values = c('grey', 'navy'))+
  geom_text(aes(DepTimeBlk, Count, label = Count), hjust = -0.1) +
  theme(legend.position = "none") + coord_flip() +
  labs(title = "Departure Delay by Time",  x = "Departure Time of the Day",  y ="Total Count of Delay")
```



To conclude we can say that the arrival traffic and departure traffic takes place at the diffent time of the day.

Clean data without any NA values.

```r
Clean_data <- FlightData[!sapply(FlightData, anyNA)]
dim(Clean_data)
```

```
## [1] 6934   54
```

```r
summary(Clean_data)
```

```
##      Random                Year         Quarter      Month     DayofMonth
##  Min.   :0.0000173   Min.   :2018   Min.   :1   Min.   :1   Min.   : 1.00
##  1st Qu.:0.2564224   1st Qu.:2018   1st Qu.:1   1st Qu.:1   1st Qu.: 8.00
##  Median :0.5111270   Median :2018   Median :1   Median :1   Median :16.00
##  Mean   :0.5051431   Mean   :2018   Mean   :1   Mean   :1   Mean   :15.83
##  3rd Qu.:0.7541251   3rd Qu.:2018   3rd Qu.:1   3rd Qu.:1   3rd Qu.:24.00
##  Max.   :0.9998639   Max.   :2018   Max.   :1   Max.   :1   Max.   :31.00
##
##    DayOfWeek         FlightDate   Marketing_Airline_Network
##  Min.   :1.000   1/15/2018: 271   WN     :1702
##  1st Qu.:2.000   1/25/2018: 261   UA     :1695
##  Median :4.000   1/8/2018 : 253   AA     :1127
##  Mean   :3.742   1/11/2018: 251   DL     : 997
##  3rd Qu.:5.000   1/3/2018 : 247   AS     : 546
##  Max.   :7.000   1/22/2018: 245   VX     : 420
##                  (Other)  :5406   (Other): 447
##  Operated_or_Branded_Code_Share_Partners DOT_ID_Marketing_Airline
##  WN          :1702                       Min.   :19393
##  UA          :1030                       1st Qu.:19690
##  AA          : 742                       Median :19805
##  UA_CODESHARE: 665                       Mean   :19870
##  DL          : 585                       3rd Qu.:19977
##  VX          : 420                       Max.   :21171
##  (Other)     :1790
##  IATA_Code_Marketing_Airline Flight_Number_Marketing_Airline
##  WN     :1702                Min.   :   1
##  UA     :1695                1st Qu.: 779
##  AA     :1127                Median :1763
##  DL     : 997                Mean   :2512
##  AS     : 546                3rd Qu.:4746
##  VX     : 420                Max.   :6937
##  (Other): 447
##  Originally_Scheduled_Code_Share_Airline
##    :6933
##  CP:   1
##
##
##
##
##
##  IATA_Code_Originally_Scheduled_Code_Share_Airline Operating_Airline
##    :6933                                           WN     :1702
##  CP:   1                                           OO     :1093
##                                                    UA     :1030
##                                                    AA     : 742
```

```
##                                                 DL     : 585
##                                                 CP     : 505
##                                                 (Other):1277
##  DOT_ID_Operating_Airline IATA_Code_Operating_Airline  Tail_Number
##  Min.   :19393            WN     :1702                N200NN :  24
##  1st Qu.:19687            OO     :1093                N207AN :  24
##  Median :19930            UA     :1030                N204NN :  21
##  Mean   :20027            AA     : 742                N205NN :  20
##  3rd Qu.:20304            DL     : 585                N216NN :  20
##  Max.   :21171            CP     : 505                N215NN :  17
##                           (Other):1277                (Other):6808
##  Flight_Number_Operating_Airline OriginAirportID OriginAirportSeqID
##  Min.   :   1                    Min.   :12892   Min.   :1289208
##  1st Qu.: 779                    1st Qu.:12892   1st Qu.:1289208
##  Median :1763                    Median :14679   Median :1467903
##  Mean   :2511                    Mean   :14028   Mean   :1402800
##  3rd Qu.:4746                    3rd Qu.:14771   3rd Qu.:1477104
##  Max.   :6937                    Max.   :14893   Max.   :1489302
##
##  OriginCityMarketID Origin              OriginCityName OriginState
##  Min.   :32457      LAX:2747   Los Angeles, CA  :2747   CA:6934
##  1st Qu.:32457      SAN:1032   Sacramento, CA   : 548
##  Median :32575      SFO:1984   San Diego, CA    :1032
##  Mean   :32727      SJC: 623   San Francisco, CA:1984
##  3rd Qu.:32575      SMF: 548   San Jose, CA     : 623
##  Max.   :33570
##
##  OriginStateFips   OriginStateName   OriginWac  DestAirportID
##  Min.   :6       California:6934   Min.   :91   Min.   :10140
##  1st Qu.:6                         1st Qu.:91   1st Qu.:11697
##  Median :6                         Median :91   Median :13232
##  Mean   :6                         Mean   :91   Mean   :13143
##  3rd Qu.:6                         3rd Qu.:91   3rd Qu.:14679
##  Max.   :6                         Max.   :91   Max.   :15919
##
##  DestAirportSeqID  DestCityMarketID      Dest                DestCityName
##  Min.   :1014005   Min.   :30140    LAX    : 432   Los Angeles, CA  : 432
##  1st Qu.:1169706   1st Qu.:30713    SEA    : 380   Seattle, WA      : 380
##  Median :1323202   Median :32211    LAS    : 347   Las Vegas, NV    : 347
##  Mean   :1314338   Mean   :32021    SFO    : 325   San Francisco, CA: 325
##  3rd Qu.:1467903   3rd Qu.:32575    PHX    : 295   Phoenix, AZ      : 295
##  Max.   :1591904   Max.   :35041    DEN    : 274   Denver, CO       : 274
##                                     (Other):4881   (Other)          :4881
##    DestState    DestStateFips      DestStateName     DestWac
##  CA     :2121   Min.   : 2.00   California:2121   Min.   : 1.00
##  TX     : 565   1st Qu.: 6.00   Texas     : 565   1st Qu.:43.00
##  NV     : 414   Median :16.00   Nevada    : 414   Median :85.00
##  WA     : 403   Mean   :22.84   Washington: 403   Mean   :70.28
##  AZ     : 356   3rd Qu.:37.00   Arizona   : 356   3rd Qu.:91.00
##  CO     : 340   Max.   :56.00   Colorado  : 340   Max.   :93.00
##  (Other):2735                   (Other)   :2735
##    CRSDepTime       DepTimeBlk     CRSArrTime       ArrTimeBlk
##  Min.   :   5   1700-1759: 529   Min.   :   1   1600-1659: 530
##  1st Qu.: 907   0600-0659: 520   1st Qu.:1103   2000-2059: 487
```

```
##  Median :1304   0700-0759: 497   Median :1514   1400-1459: 461
##  Mean   :1332   1200-1259: 480   Mean   :1474   1800-1859: 453
##  3rd Qu.:1738   0900-0959: 470   3rd Qu.:1923   1100-1159: 451
##  Max.   :2359   1400-1459: 418   Max.   :2359   2100-2159: 430
##                 (Other)  :4020                  (Other)  :4122
##    Cancelled       CancellationCode    Diverted        CRSElapsedTime
##  Min.   :0.00000    :6785           Min.   :0.000000   Min.   : 24.0
##  1st Qu.:0.00000   A:  20           1st Qu.:0.000000   1st Qu.: 95.0
##  Median :0.00000   B: 123           Median :0.000000   Median :143.0
##  Mean   :0.02149   C:   6           Mean   :0.001586   Mean   :165.1
##  3rd Qu.:0.00000                    3rd Qu.:0.000000   3rd Qu.:213.0
##  Max.   :1.00000                    Max.   :1.000000   Max.   :683.0
##
##     Flights     Distance    DistanceGroup    CarrierDelay_mod
##  Min.   :1   Min.   :  31   Min.   : 1.000   Min.   :  0.000
##  1st Qu.:1   1st Qu.: 401   1st Qu.: 2.000   1st Qu.:  0.000
##  Median :1   Median : 794   Median : 4.000   Median :  0.000
##  Mean   :1   Mean   : 979   Mean   : 4.357   Mean   :  3.268
##  3rd Qu.:1   3rd Qu.:1400   3rd Qu.: 6.000   3rd Qu.:  0.000
##  Max.   :1   Max.   :4962   Max.   :11.000   Max.   :773.000
##
##  WeatherDelayMod    DivAirportLandings  Div1Airport    Div1TailNum   Duplicate
##  Min.   :  0.0000   Min.   :0.000000          :6919          :6927   N:6934
##  1st Qu.:  0.0000   1st Qu.:0.000000   LAX    :   3   N14235 :   1
##  Median :  0.0000   Median :0.000000   SEA    :   3   N16703 :   1
##  Mean   :  0.7877   Mean   :0.006778   OAK    :   2   N37267 :   1
##  3rd Qu.:  0.0000   3rd Qu.:0.000000   PDX    :   2   N622QX :   1
##  Max.   :850.0000   Max.   :9.000000   ABQ    :   1   N625QX :   1
##                                        (Other):   4   (Other):   2
```

References

1. https://www.transtats.bts.gov/
2. Data Visualization: druhrao-ml-data–visualization-08242017_310311.pdf - by Professor Dr. Umesh R Hodeghatta
3. Data Visualization - ggplots.pdf - by Professor Dr. Umesh R Hodeghatta
4. https://www.datacamp.com/courses/data-visualization-with-ggplot2-1
5. https://www.datacamp.com/courses/data-visualization-with-ggplot2-2