Air Traffic Delays Analysis

Sumanth

4/13/2020

The data 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. (Ref: https://www.transtats.bts.gov/)

I have performed analysis on California State

Importing libraries

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

Load the data :

```
California_delays <- read_csv("~/CLASSROOM/Analytics inn practice/ASS 2/Sample_CA_airtraffic_delays.csv
```

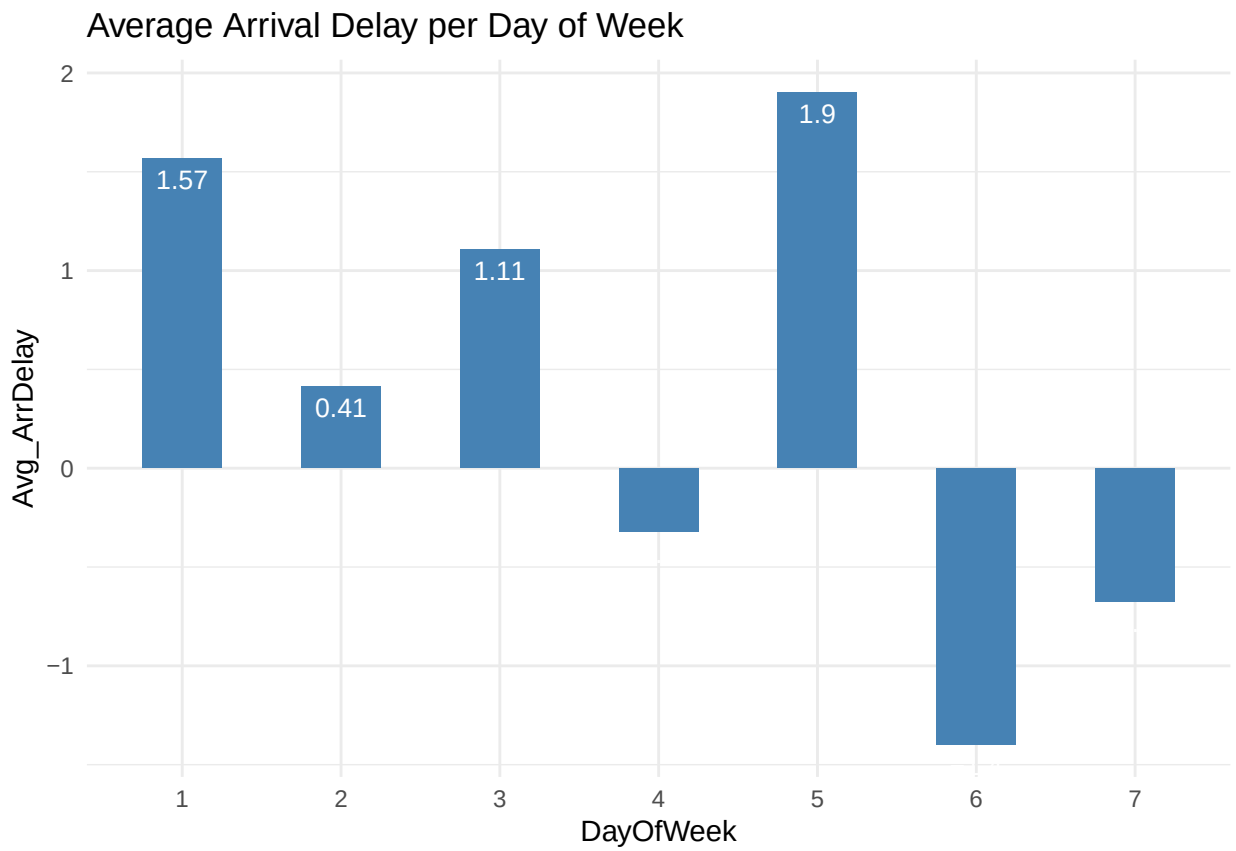
Exploratory Data Analysis

```
ArrDep_delays <- California_delays %>% select("DayofMonth","DayOfWeek","Origin","DepDelay","DepDelayMinu
ArrDep_delays$DayofMonth<-as.Date(ArrDep_delays$DayofMonth, origin = "2018-01-01")
ArrDep_delays$DayOfWeek<-as.factor(ArrDep_delays$DayOfWeek)
ArrDep_delays$DepDel15<-as.factor(ArrDep_delays$DepDel15)
ArrDep_delays$ArrDel15<-as.factor(ArrDep_delays$ArrDel15)
```

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

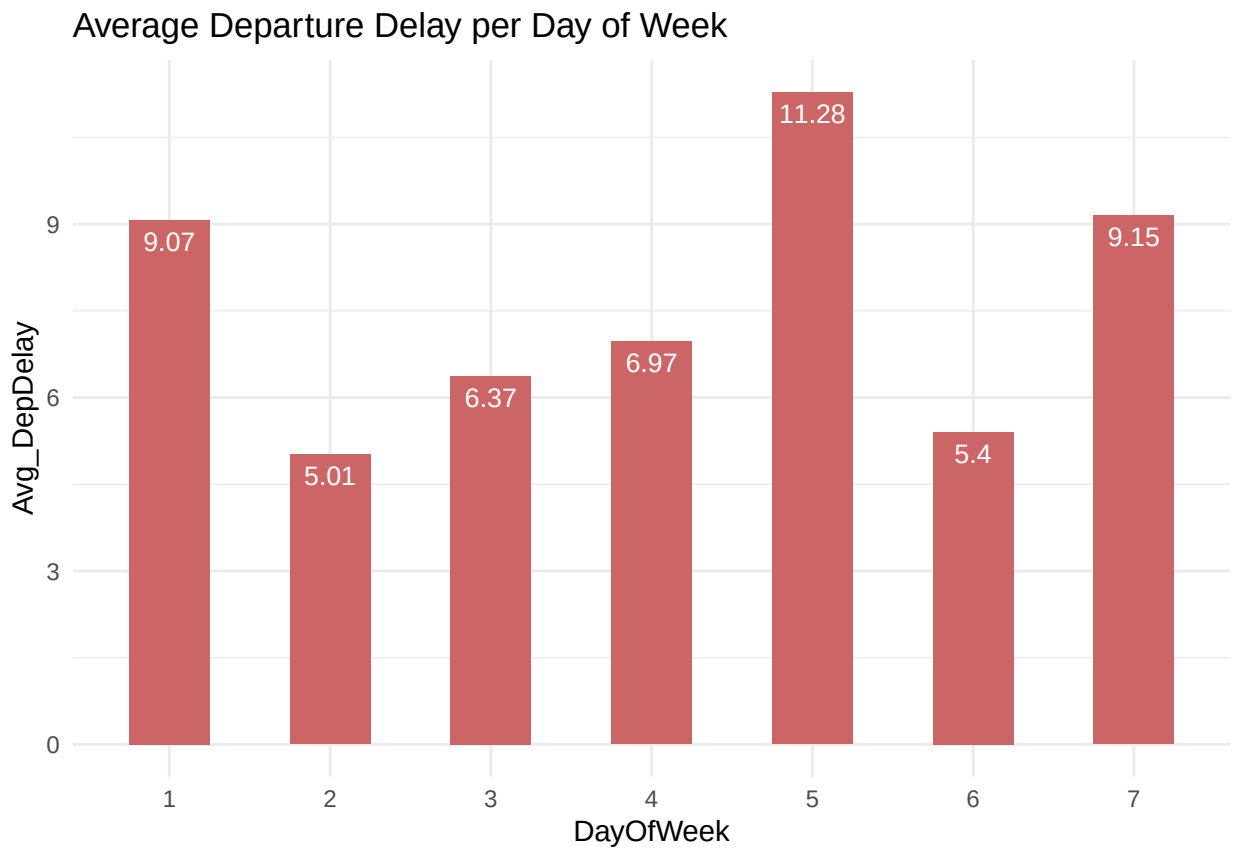
```
ArrDep_delays[is.na(ArrDep_delays)]<-0
Week <- ArrDep_delays %>%
        group_by( DayOfWeek ) %>%
        summarise( Count = n(), Avg_ArrDelay=mean(ArrDelay), Avg_DepDelay=mean(DepDelay))
Month <-ArrDep_delays %>%
        group_by( DayofMonth ) %>%
        summarise( Count = n(), Avg_ArrDelay=mean(ArrDelay), Avg_DepDelay=mean(DepDelay))
```

```
 ggplot(Week, aes(x=DayOfWeek, y=Avg_ArrDelay)) + geom_bar(stat="identity", fill="steelblue",width = 0.5
  geom_text(aes(label=round(Avg_ArrDelay, 2)), vjust=1.6, color="white",
            position = position_dodge(0.9), size=3.5)+
  scale_fill_brewer(palette="Paired")+
 theme_minimal()+labs(title="Average Arrival Delay per Day of Week")
```



From the above graph, It is clearly evident that the Arrival delay for day 1, 5 are high compared to the rest. This analysis suggests there is a greater scope to improve on those days.

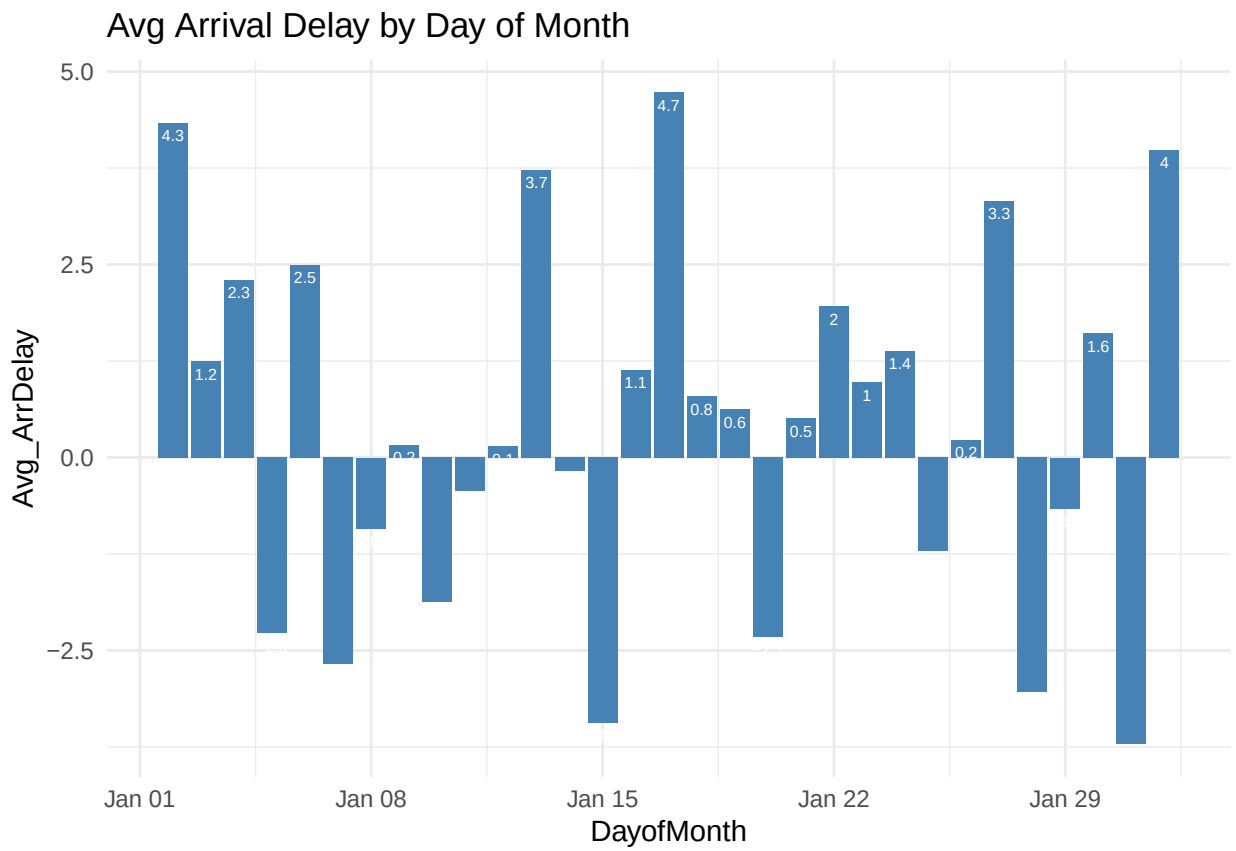
```
ggplot(data=Week, aes(x=DayOfWeek, y=Avg_DepDelay)) +geom_bar(stat="identity", fill="#CC6666", width = 
  geom_text(aes(label=round(Avg_DepDelay, 2)), vjust=1.6, color="white",
            position = position_dodge(0.9), size=3.5)+
  scale_fill_brewer(palette="Paired")+
  theme_minimal()+labs(title="Average Departure Delay per Day of Week")
```



The Departure delays in the California state at all the origins is higher compared to the arrival departure delays.

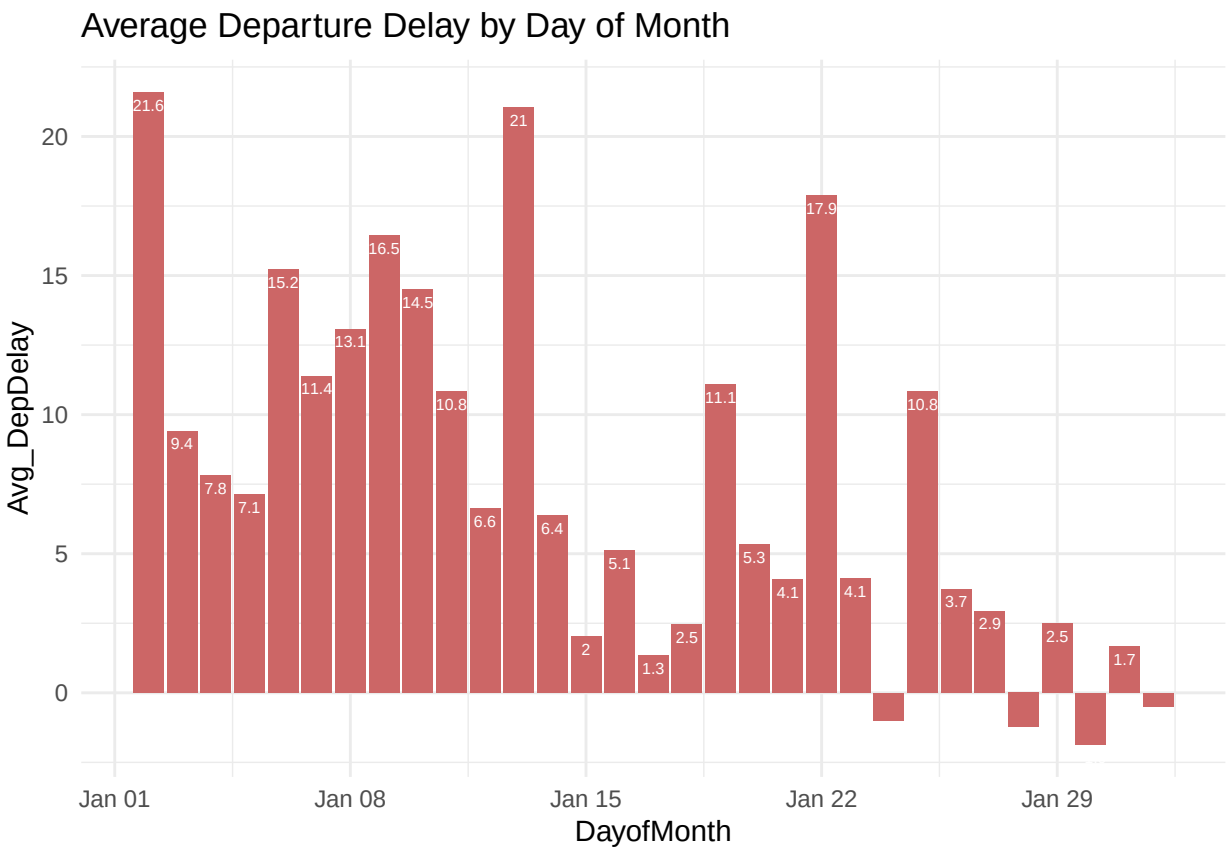
- There are various factors that influence delays. eg: carrier delay, security delay, weather delay etc.,.
- The departure delay is much higher on days 1, 5 and is equally correlated with the arrival delay

```
ggplot(data=Month, aes(x=DayofMonth, y=Avg_ArrDelay)) +geom_bar(stat="identity", fill="steelblue")+
  geom_text(aes(label=round(Avg_ArrDelay, 1)), vjust=1.6, color="white",
            position = position_dodge(0.9), size=2)+
  scale_fill_brewer(palette="Paired")+
  theme_minimal()+labs(title="Avg Arrival Delay by Day of Month")
```



From the above plot it could be inferred that week 1 has the highest the highest number of delays i.e., from jan 01 to jan 08.

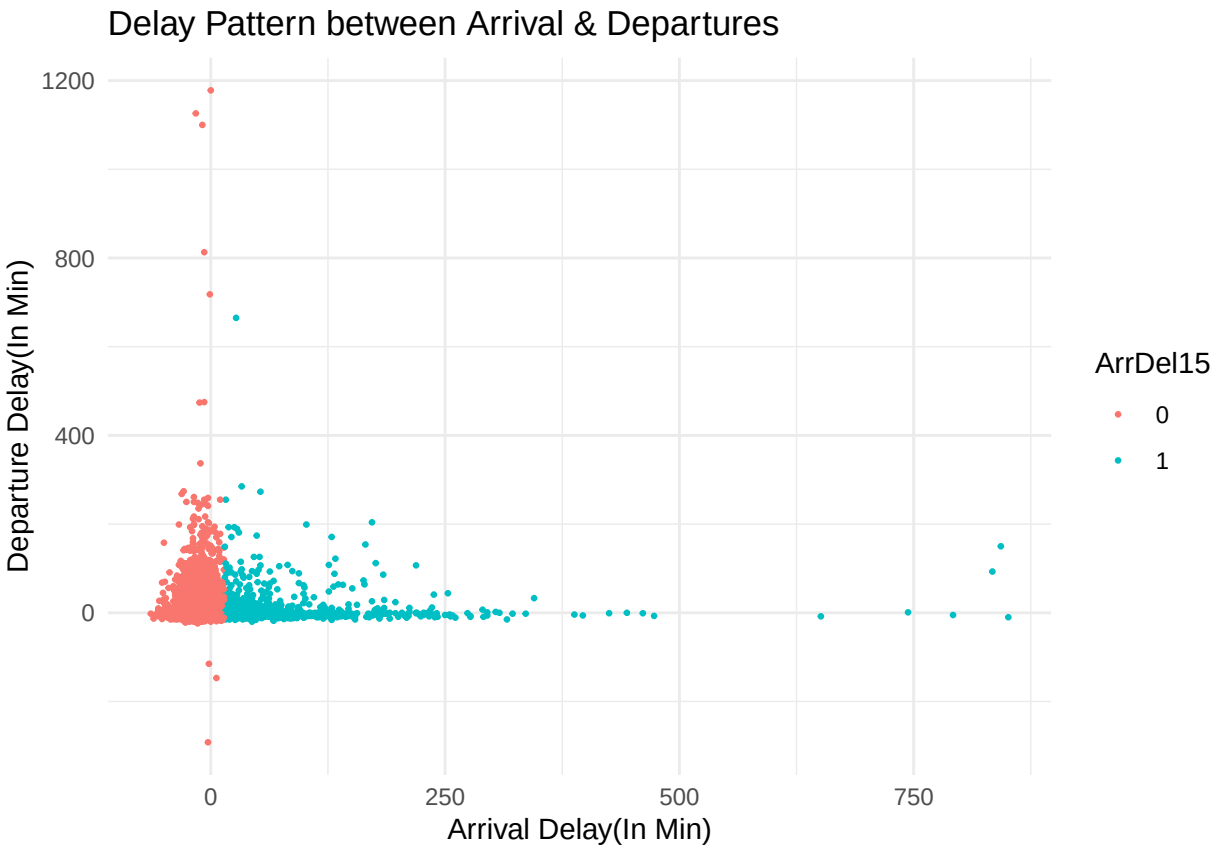
```
ggplot(data=Month, aes(x=DayofMonth, y=Avg_DepDelay)) +geom_bar(stat="identity", fill="#CC6666")+
  geom_text(aes(label=round(Avg_DepDelay, 1)), vjust=1.6, color="white",
            position = position_dodge(0.9), size=2)+
  scale_fill_brewer(palette="Paired")+
  theme_minimal()+labs(title="Average Departure Delay by Day of Month")
```



If we observere the trend, the Departure delays are more frequent and higher in the starting 15 days than the rest.

2. Can you interpret the traffic delays?

```
ArrDep_delays %>%
 filter(!is.na(DepDelay)) %>%
 filter(!is.na(ArrDelay)) %>%
 filter(!is.na(ArrDel15)) %>%
 ggplot() +
 aes(x = ArrDelay, y = DepDelay, colour = ArrDel15) +
 geom_point(size = 0.5) +
 scale_color_hue() +
 labs(x = "Arrival Delay(In Min)", y = "Departure Delay(In Min)", title = "Delay Pattern between Arrival
 theme_minimal()
```

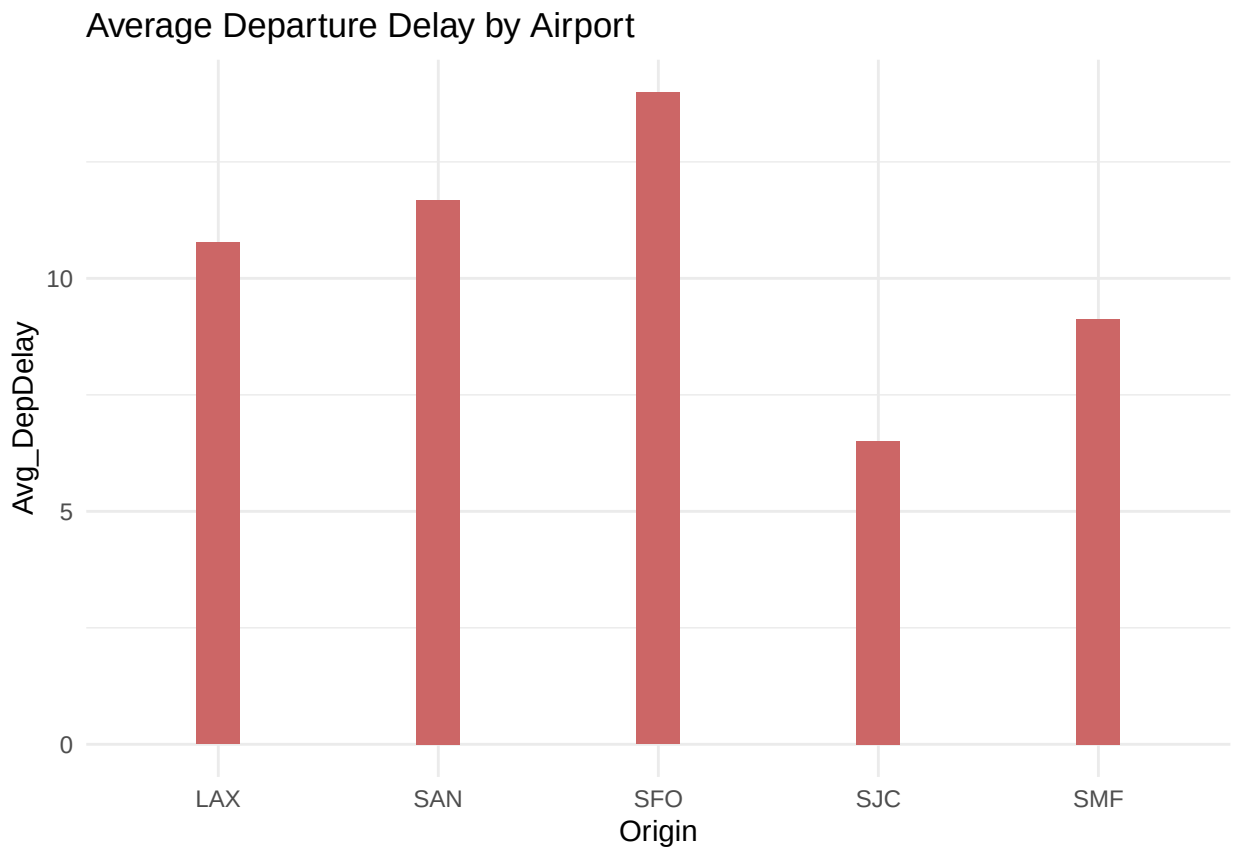


- The pattern of the traffic delays data is normally distributed.
- In specific, The probability of departure delay is higher when there is an early arrival of flights than scheduleed.

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

```
Highest_ArrDep<-ArrDep_delays %>%
        group_by( Origin ) %>%
        summarise( Count = n(), Avg_ArrDelay=mean(ArrDelayMinutes), Avg_DepDelay=mean(DepDelayMinutes))
```

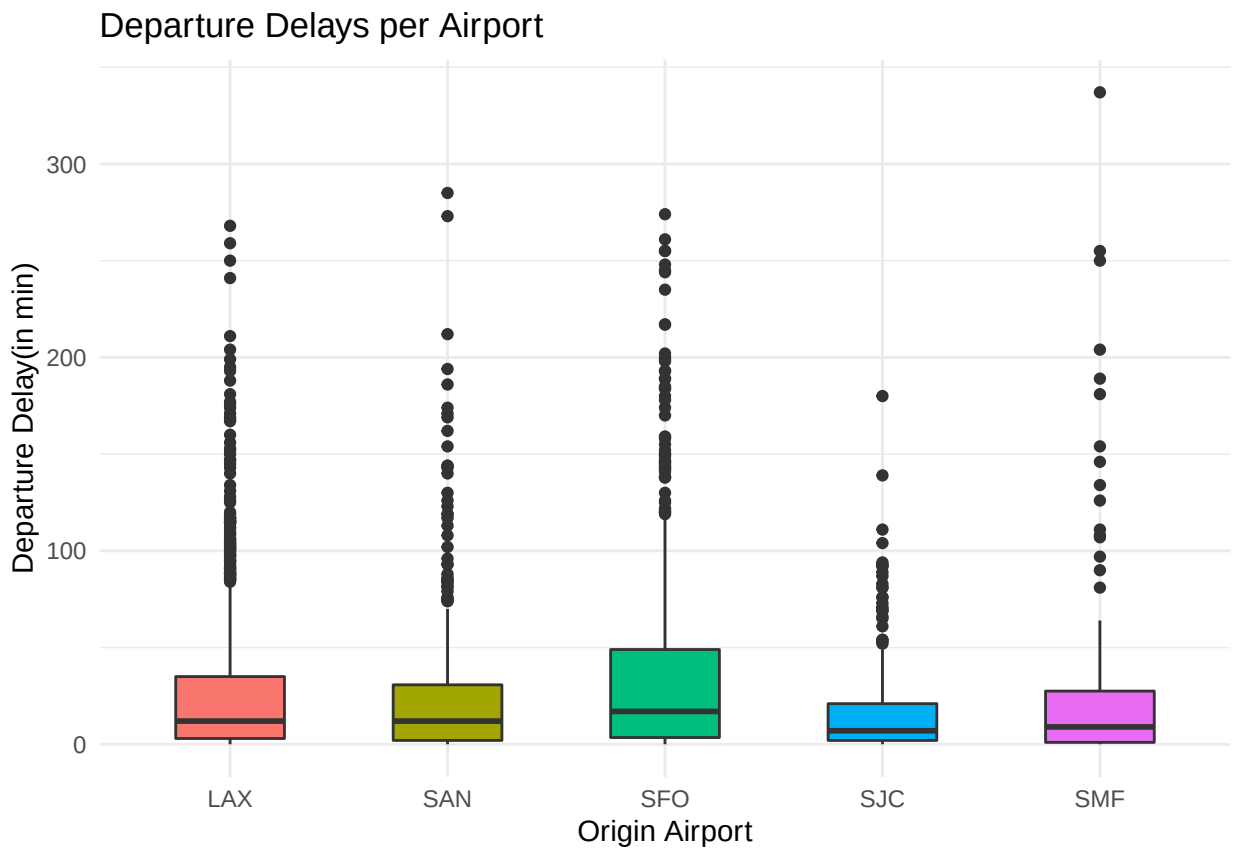
```
ggplot(Highest_ArrDep, aes(x=Origin, y=Avg_DepDelay)) +geom_bar(stat="identity", fill="#CC6666",width =
  theme_minimal()+labs(title="Average Departure Delay by Airport")
```



The SFO (San Francisco International Airport) has highest number of Departure delays among all the airports in the caliifornia State.

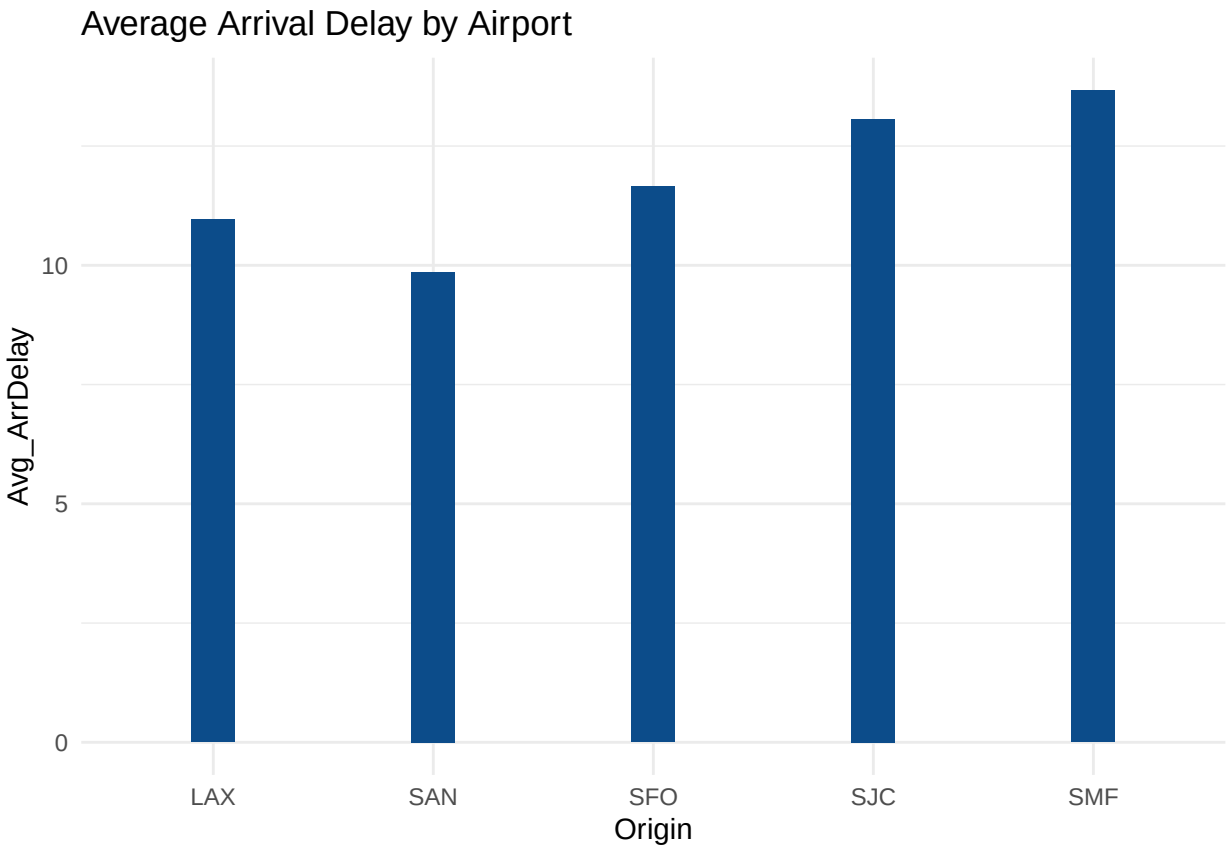
I have also created Boxplots to look at the distribution of the data for a better visual understanding.

```
ArrDep_delays %>%
 filter(DepDelay >= 0 & DepDelay <= 403L & !is.na(DepDelay)) %>%
 ggplot() +
 aes(x = Origin, y = DepDelay, fill = Origin) +
 geom_boxplot(width = 0.5) +
 scale_fill_hue() +
 labs(x = "Origin Airport", y = "Departure Delay(in min)", title = "Departure Delays per Airport") +
 theme_minimal() +
 theme(legend.position = "none")
```



4. Which Airport has highest Arrival delay?

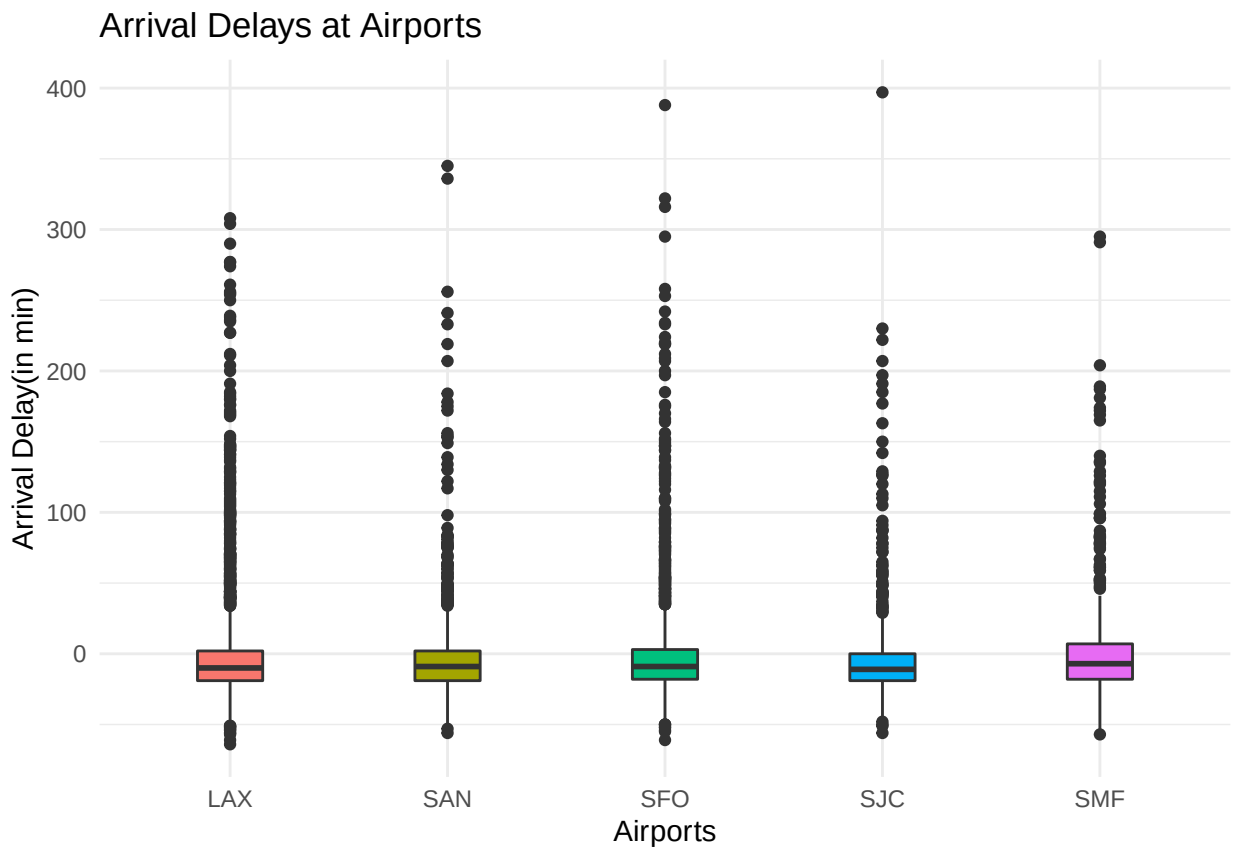
```
ggplot(Highest_ArrDep, aes(x=Origin, y=Avg_ArrDelay)) +geom_bar(stat="identity", fill="#0c4c8a",width =
  scale_fill_brewer(palette="Paired")+
  theme_minimal()+labs(title="Average Arrival Delay by Airport")
```



The SMF (Sacramento International Airport) has the highest Departure delays in the state followed by SJC(San Jose International Airport)

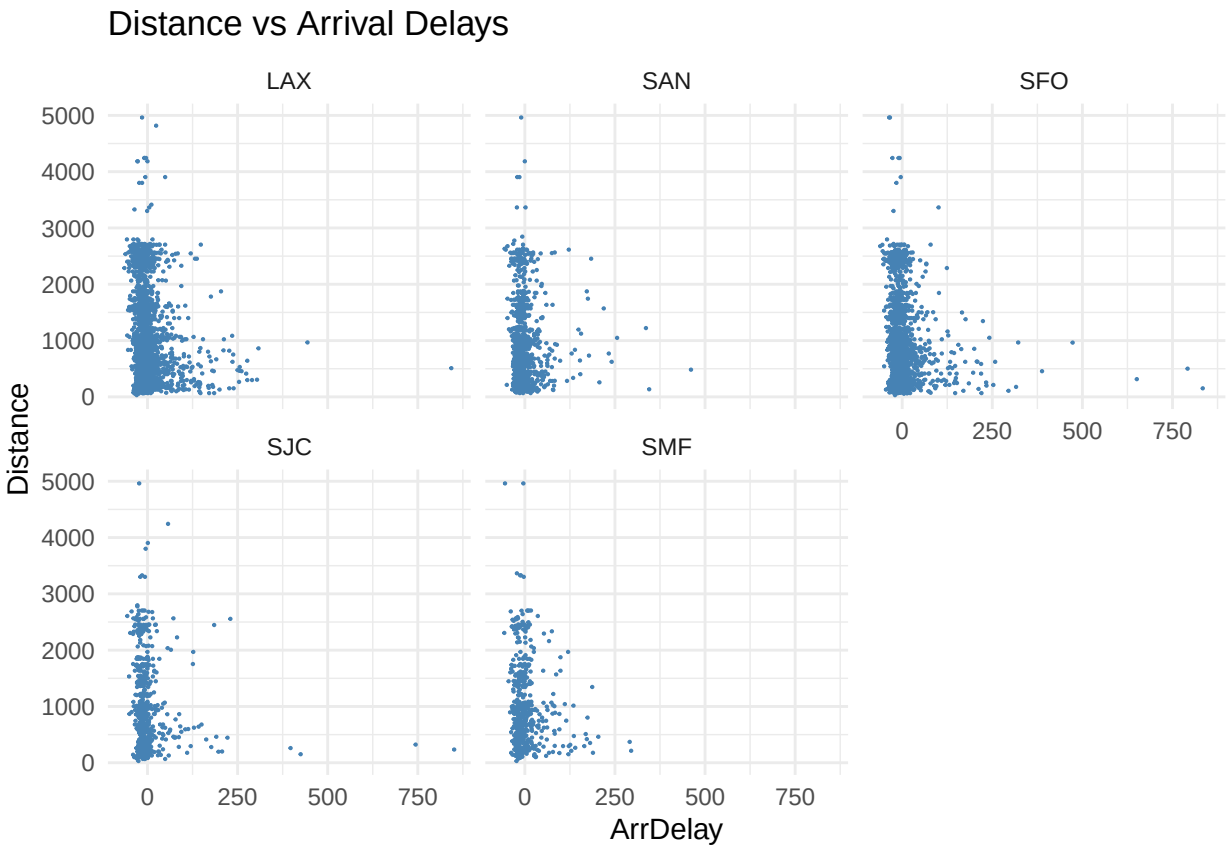
Below, We can find the distribution the data with boxplots.

```
ArrDep_delays %>%
 filter(ArrDelay >= -64L & ArrDelay <= 405L & !is.na(ArrDelay)) %>%
 ggplot() +
 aes(x = Origin, y = ArrDelay, fill = Origin) +
 geom_boxplot(width = 0.3) +
 scale_fill_hue() +
 labs(x = "Airports", y = "Arrival Delay(in min)", title = "Arrival Delays at Airports") +
 theme_minimal() +
 theme(legend.position = "none")
```

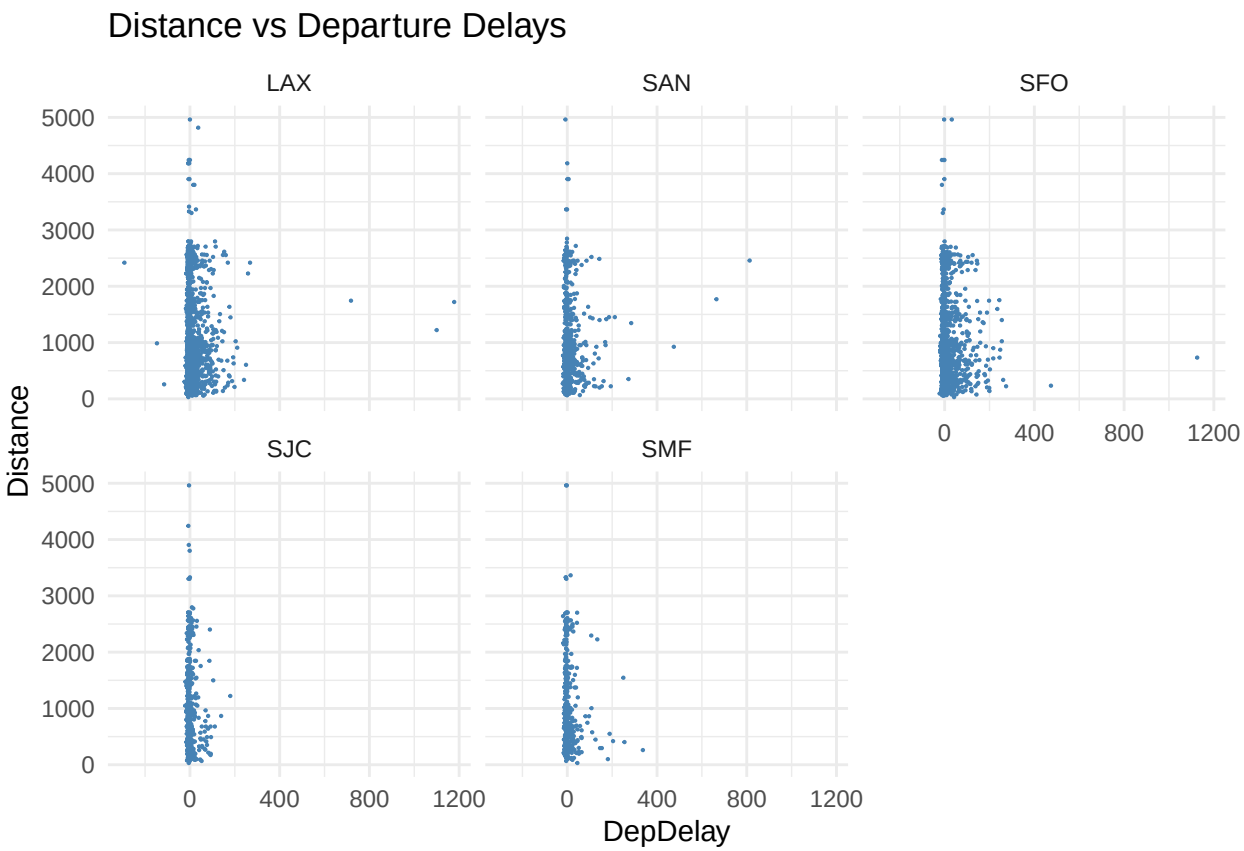


5. How do you relate the delay pattern to the distance travelled?

```
ggplot(ArrDep_delays, aes(ArrDelay, Distance)) +
        geom_point(color='steelblue',size = 0.1)+theme_minimal()+
        facet_wrap(~Origin)+
        labs(title="Distance vs Arrival Delays")
```




```
ggplot(ArrDep_delays, aes(DepDelay, Distance)) +
        geom_point(color='steelblue',size = 0.1)+theme_minimal()+
        facet_wrap(~Origin)+
        labs(title="Distance vs Departure Delays")
```

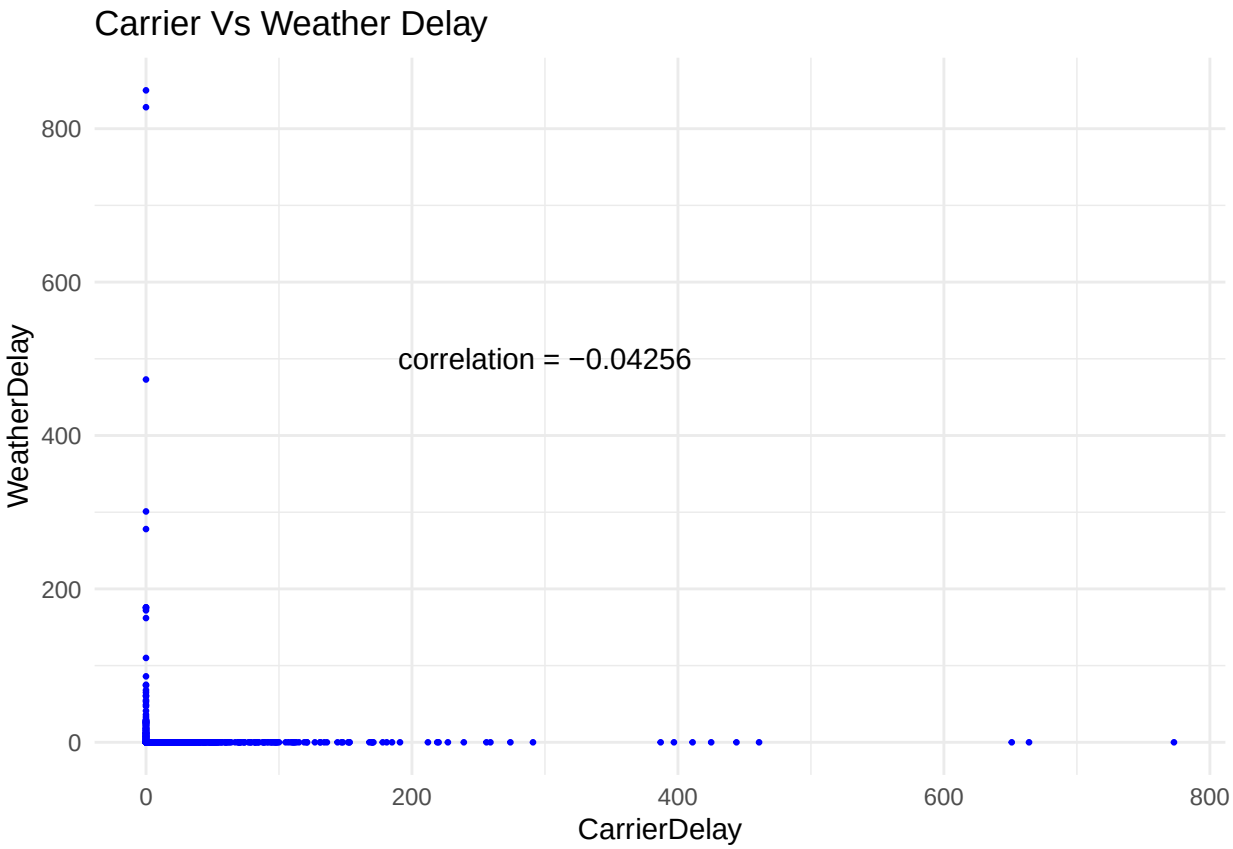


There is no relation between the delays to distance.

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

According to the data, The correlation is almost close to zero by which we can conclude there is no correlation.

```
traffic_delay<-California_delays %>% select('DayofMonth','CarrierDelay','WeatherDelay','NASDelay','Secu

traffic_delay[is.na(traffic_delay)]<-0

 ggplot(traffic_delay) +
 aes(x = CarrierDelay, y = WeatherDelay) +
 geom_point(size = 0.5, colour = "blue") +
 labs(title = "Carrier Vs Weather Delay") +
 theme_minimal() + annotate("text", x = 300, y = 500, label = "correlation = -0.04256")
```



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

As mentioned, Due to to limited capability of my PC, I have done my analysis on California state only. Airports with Highest number of flights count tend to have more number of delays when compared to other airports.

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

```
#filtering the data :
cancel_delays<-California_delays %>% select('DepDelay','ArrDelay','Cancelled','Diverted')
cancel_delays %>% filter(Cancelled==1 & !is.na(Cancelled) & DepDelay > 0 & !is.na(DepDelay)) %>% count(C
```

```
## # A tibble: 1 x 2
##   Cancelled     n
##       <dbl> <int>
## 1         1    44
```

There are a total number of 44 flights cancelled due to delay.

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

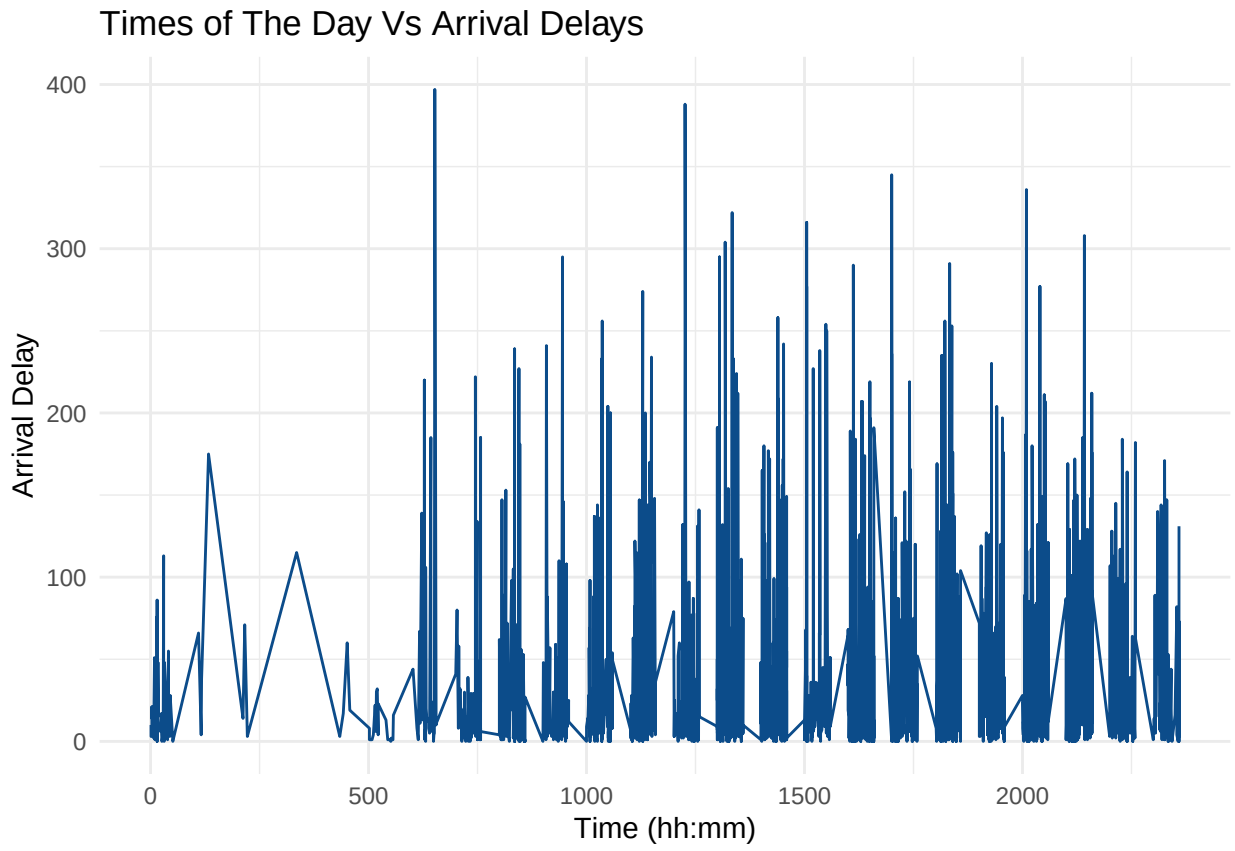
```
cancel_delays %>% filter(Diverted==1 & !is.na(Diverted) & DepDelay > 0 & !is.na(DepDelay)) %>% count(Di
```

```
## # A tibble: 1 x 2
##   Diverted     n
##      <dbl> <int>
## 1        1     3
```

A total number of 3 flights were diverted due to delay.

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

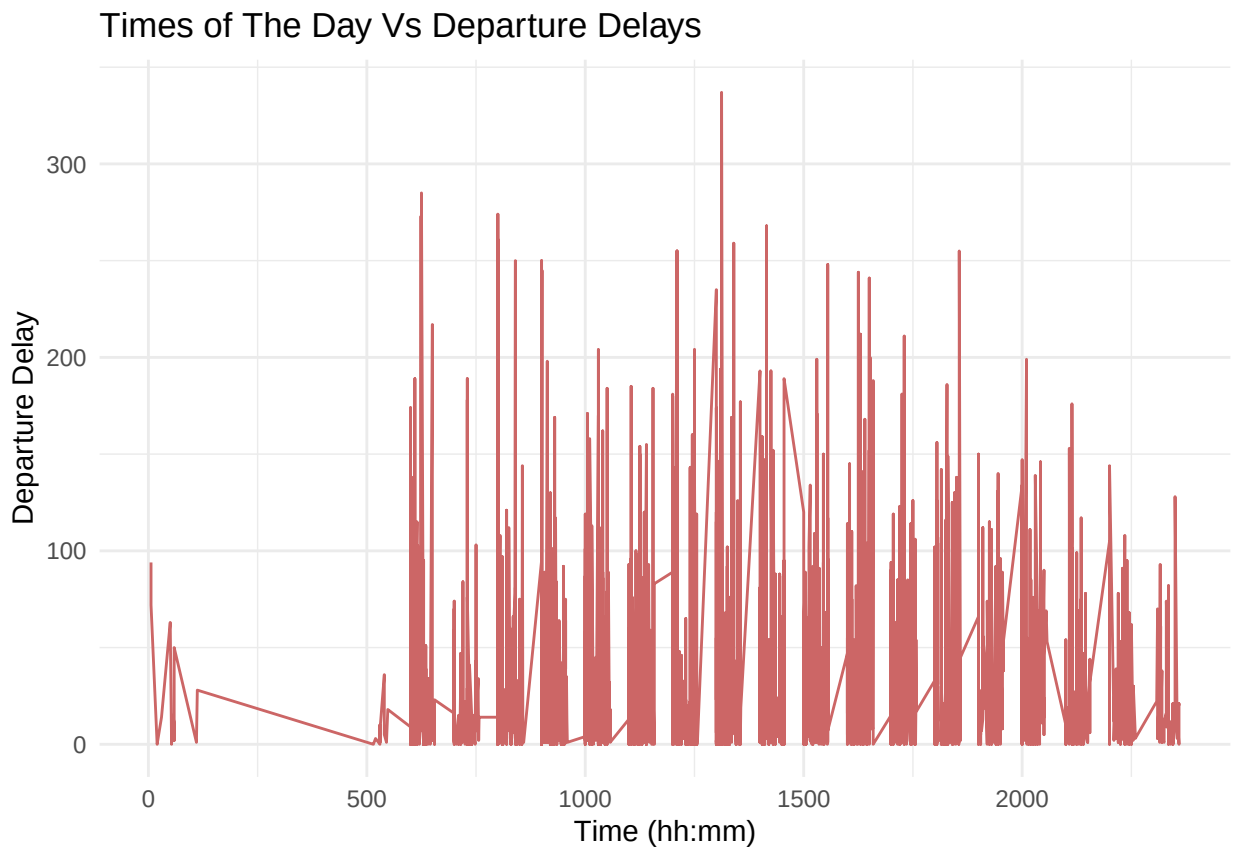
```
time_delays<-California_delays %>% select('CRSDepTime','DepDelay','CRSArrTime','ArrDelay')

time_delays %>%
 filter(ArrDelay >= 0 & ArrDelay <= 410L & !is.na(ArrDelay)) %>%
 ggplot() +
 aes(x = CRSArrTime, y = ArrDelay) +
 geom_line(size = 0.5, colour = "#0c4c8a") +
 labs(x = "Time (hh:mm)", y = "Arrival Delay", title = "Times of The Day Vs Arrival Delays") +
 theme_minimal()
```



The Above plot is the time series visual graph wiwth respect to Arrival delays. As we can infer, The Afternoons of a day is usually having Arrival delays i.e., from 12:00 pm to 18:00 pm

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

```
time_delays %>%
 filter(DepDelay >= 0 & DepDelay <= 417L & !is.na(DepDelay)) %>%
 ggplot() +
 aes(x = CRSDepTime, y = DepDelay) +
 geom_line(size = 0.5, colour = "#CC6666") +
 labs(x = "Time (hh:mm)", y = "Departure Delay", title = "Times of The Day Vs Departure Delays") +
 theme_minimal()
```



The Afternoons of a day is usually having Arrival delays i.e., from 12:00 pm to 18:00 pm