nyc flight data exploration ,time series and regression model

Aritra Banerjee

2024-03-30

`` installing important libraries

```
library(reshape2)
library(tidyverse)
```

```
## ── Attaching core tidyverse packages ──────────────────────── tidyverse
2.0.0 ──
## ✓ dplyr     1.1.4     ✓ readr     2.1.5
## ✓ forcats   1.0.0     ✓ stringr   1.5.1
## ✓ ggplot2   3.5.0     ✓ tibble    3.2.1
## ✓ lubridate 1.9.3     ✓ tidyr     1.3.1
## ✓ purrr     1.0.2
## ── Conflicts ──────────────────────────────────────────
tidyverse_conflicts() ──
## ✕ dplyr::filter() masks stats::filter()
## ✕ dplyr::lag()    masks stats::lag()
## ℹ Use the conflicted package (<http://conflicted.r-lib.org/>) to force all
conflicts to become errors
```

```
library(ggplot2)
library(GGally)
```

```
## Registered S3 method overwritten by 'GGally':
##   method from
##   +.gg   ggplot2
```

```
library(caret)
```

```
## Loading required package: lattice
##
## Attaching package: 'caret'
##
## The following object is masked from 'package:purrr':
##
##     lift
```

```
library(dplyr)
library(forecast)
```

```
## Registered S3 method overwritten by 'quantmod':
##   method            from
##   as.zoo.data.frame zoo
```

DATA LOADING AND EXPLORATION

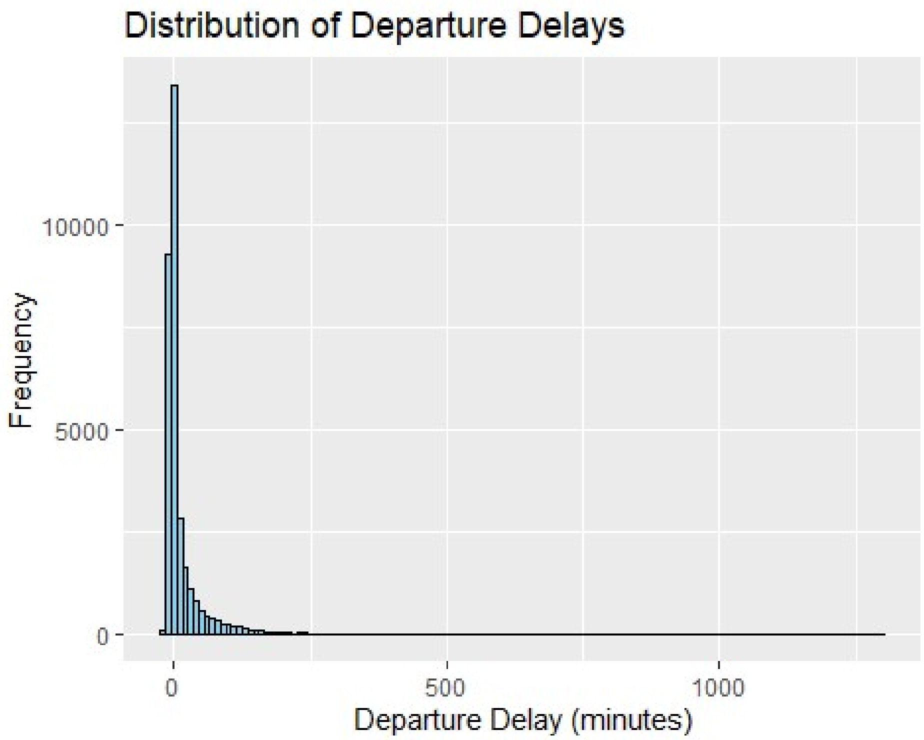
```
flights<-read.csv("nycflights.csv")
#exploring the dataset
head(flights)
summary(flights)
names(flights)
glimpse(flights)
```

DATA CLEANING

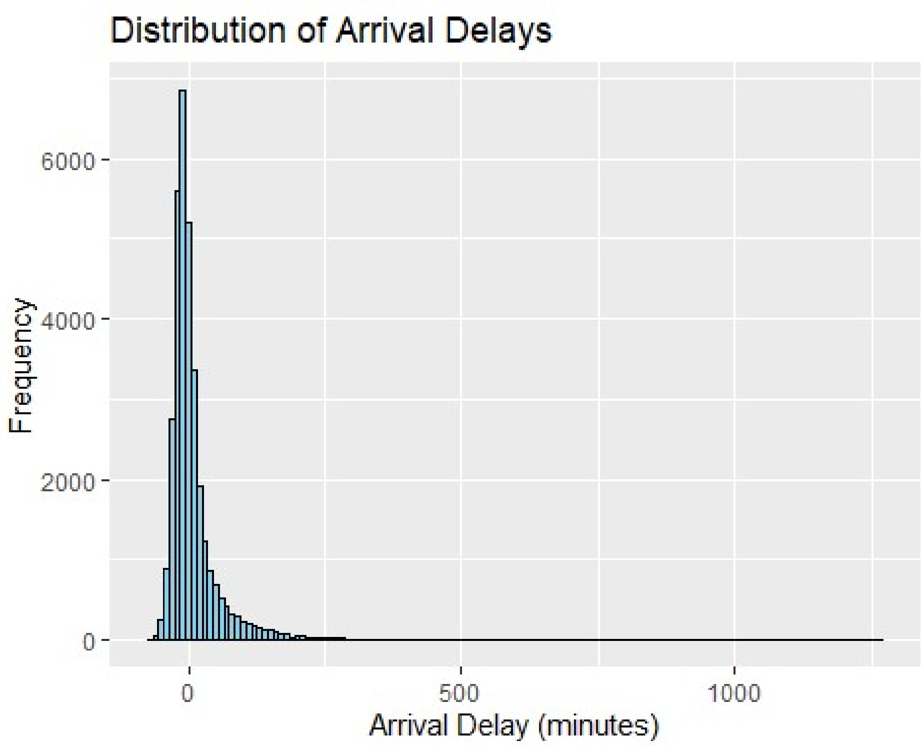
```
#Checking for na values
sum(is.na(flights))

## [1] 0

#removing na values
flights<-na.omit(flights)
```

** EXPLORATORY DATA ANALYSIS**

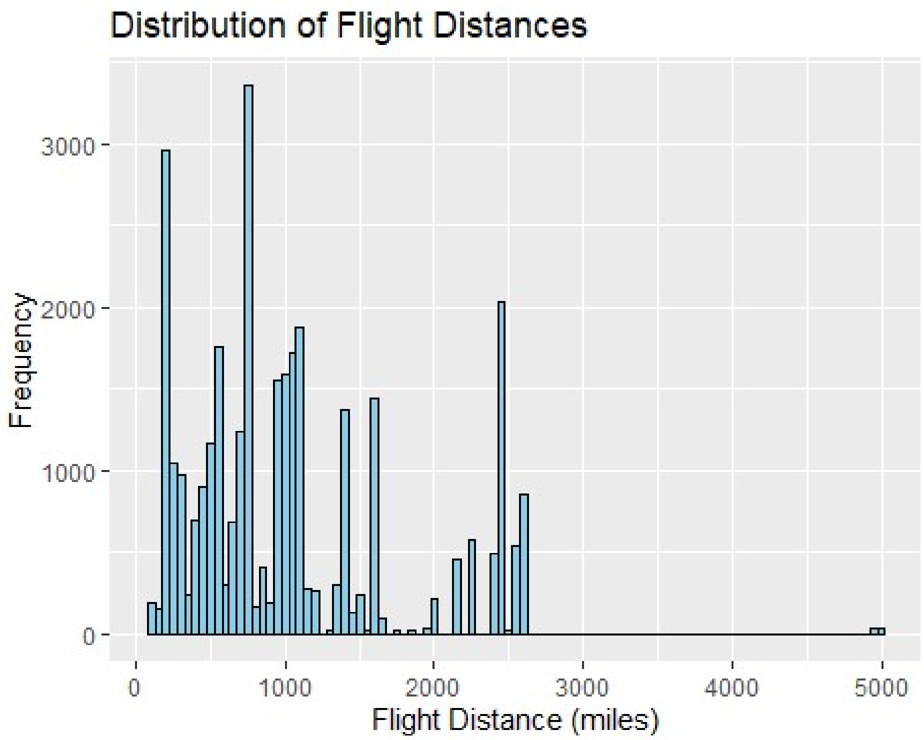
```
#creating data frame
data <- data.frame(
  DepartureDelay = flights$dep_delay,
  ArrivalDelay = flights$arr_delay,
  FlightDistance = flights$distance
)


#HISTOGRAM PLOT
ggplot(data, aes(x = DepartureDelay)) +
  geom_histogram(binwidth = 10, fill = "skyblue", color = "black") +
  labs(title = "Distribution of Departure Delays", x = "Departure Delay
(minutes)", y = "Frequency")
```



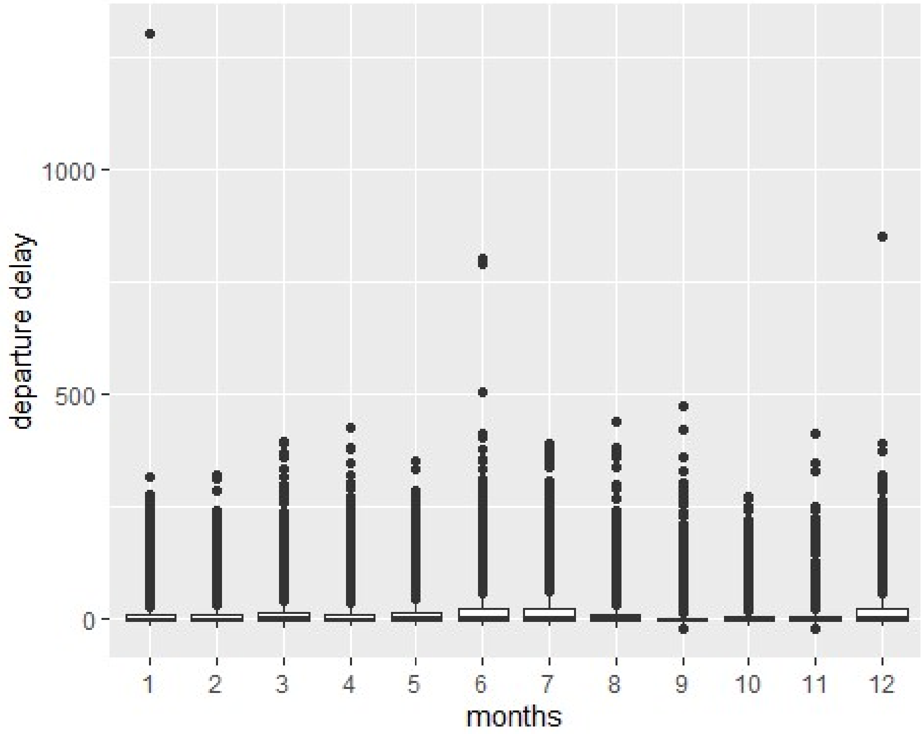
```
ggplot(data, aes(x = ArrivalDelay)) +
  geom_histogram(binwidth = 10, fill = "skyblue", color = "black") +
  labs(title = "Distribution of Arrival Delays", x = "Arrival Delay
(minutes)", y = "Frequency")
```



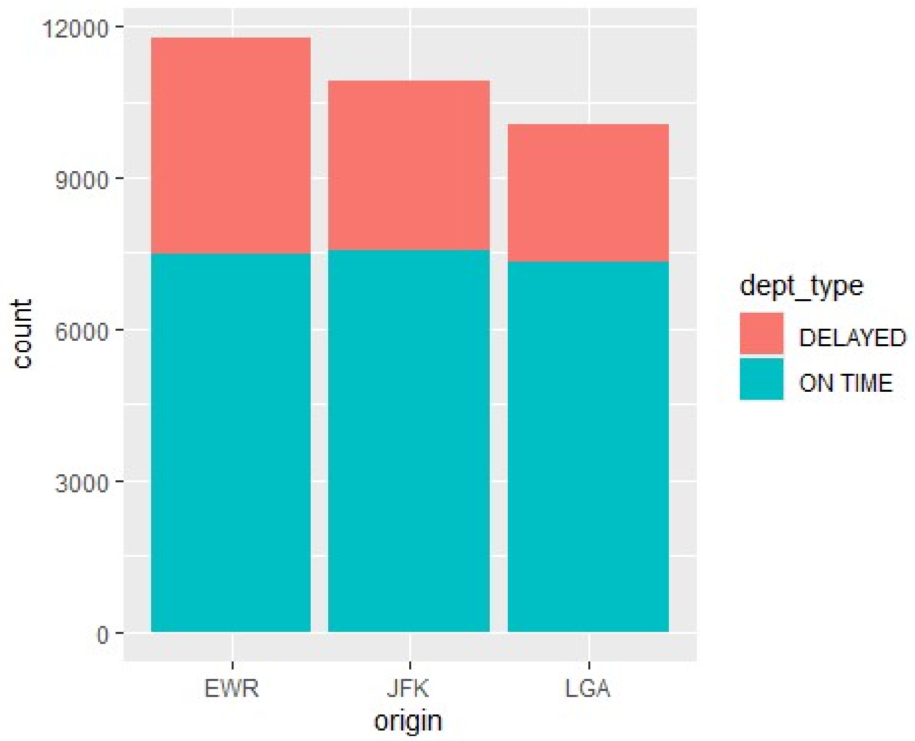
```r
ggplot(data, aes(x = FlightDistance)) +
  geom_histogram(binwidth = 50, fill = "skyblue", color = "black") +
  labs(title = "Distribution of Flight Distances", x = "Flight Distance
(miles)", y = "Frequency")
```



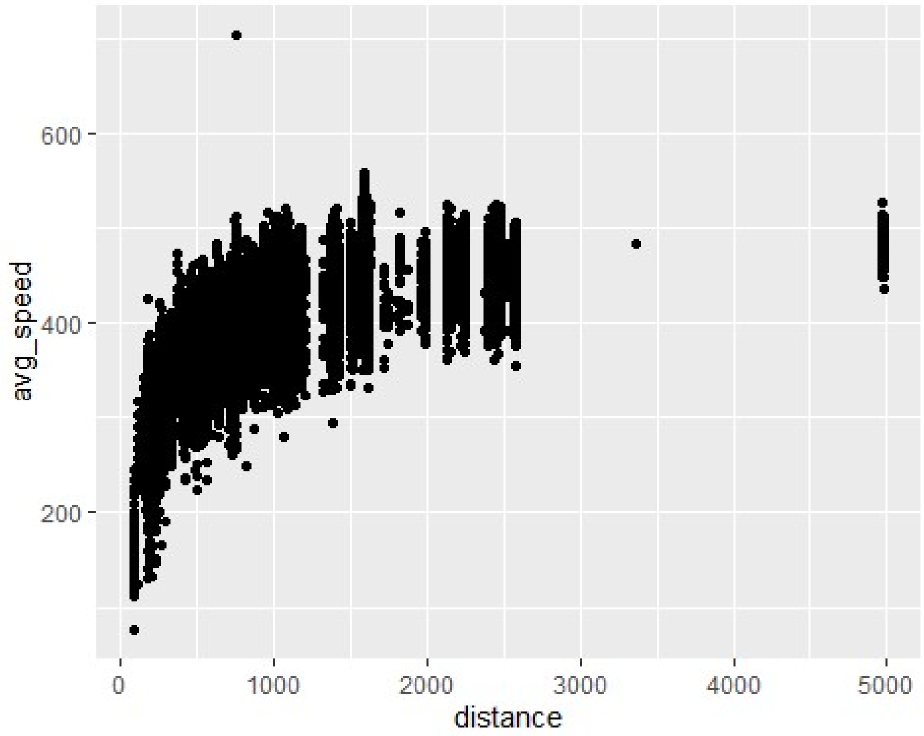
```
#UNDERSTANDING DATA USING BOXPLOT
ggplot(flights,aes(x=factor(month),
y=dep_delay))+geom_boxplot()+labs(x="months",y="departure delay")
```



```
#SUMMARIZING MORE FACTORS

flights %>%
  select(dep_delay) %>%
  summary()

##    dep_delay
##  Min.   : -21.00
##  1st Qu.:  -5.00
##  Median :  -2.00
##  Mean   :  12.71
##  3rd Qu.:  11.00
##  Max.   :1301.00

flights %>%
  group_by(month) %>%
  summarise(mean_dd=mean(dep_delay)) %>%
  arrange(desc(mean_dd))

## # A tibble: 12 × 2
##    month mean_dd
##    <int>   <dbl>
##  1     7   20.8
##  2     6   20.4
##  3    12   17.4
##  4     4   14.6
##  5     3   13.5
##  6     5   13.3
##  7     8   12.6
##  8     2   10.7
##  9     1   10.2
## 10     9    6.87
## 11    11    6.10
## 12    10    5.88

flights %>%
  group_by(month) %>%
  summarise(median_dd=median(dep_delay)) %>%
  arrange(desc(median_dd))

## # A tibble: 12 × 2
##    month median_dd
##    <int>     <dbl>
##  1    12         1
##  2     6         0
##  3     7         0
##  4     3        -1
##  5     5        -1
##  6     8        -1
##  7     1        -2
##  8     2        -2
##  9     4         -2
## 10    11         -2
## 11     9         -3
## 12    10         -3

#flight delay rate
flights<-flights %>%
  mutate(arrival_type=ifelse(arr_delay<=0,"ON TIME","DELAYED"))
flights<-flights %>%
  mutate(departure_type=ifelse(dep_delay<=0,"ON TIME","DELAYED"))

flights %>% select(departure_type,arrival_type) %>% table

##                arrival_type
## departure_type DELAYED ON TIME
##        DELAYED    9291    3508
##        ON TIME    4171   15765

#proportion of flight on time arrival
(3508/(9291+3508))*100

## [1] 27.40839

#flight delay visualization
flights<-flights %>%
  mutate(dept_type=ifelse(dep_delay<5,"ON TIME","DELAYED"))

flights %>%
  group_by(origin) %>%
  summarise(on_time_dept_rate=sum(dept_type=="ON TIME")/n()) %>%
  arrange(desc(on_time_dept_rate))

## # A tibble: 3 × 2
##   origin on_time_dept_rate
##   <chr>              <dbl>
## 1 LGA                0.728
## 2 JFK                0.694
## 3 EWR                0.637

ggplot(data=flights,aes(x=origin,fill=dept_type))+geom_bar()
```



```
#SPEED ANALYSIS
flights<-flights %>%
  mutate(avg_speed=distance/(air_time/60))

flights %>% select(avg_speed,tailnum) %>% arrange(desc(avg_speed)) %>%
filter(avg_speed==max(flights$avg_speed))

##   avg_speed tailnum
## 1  703.3846  N666DN

ggplot(flights,aes(x=distance,y=avg_speed))+geom_point()
```



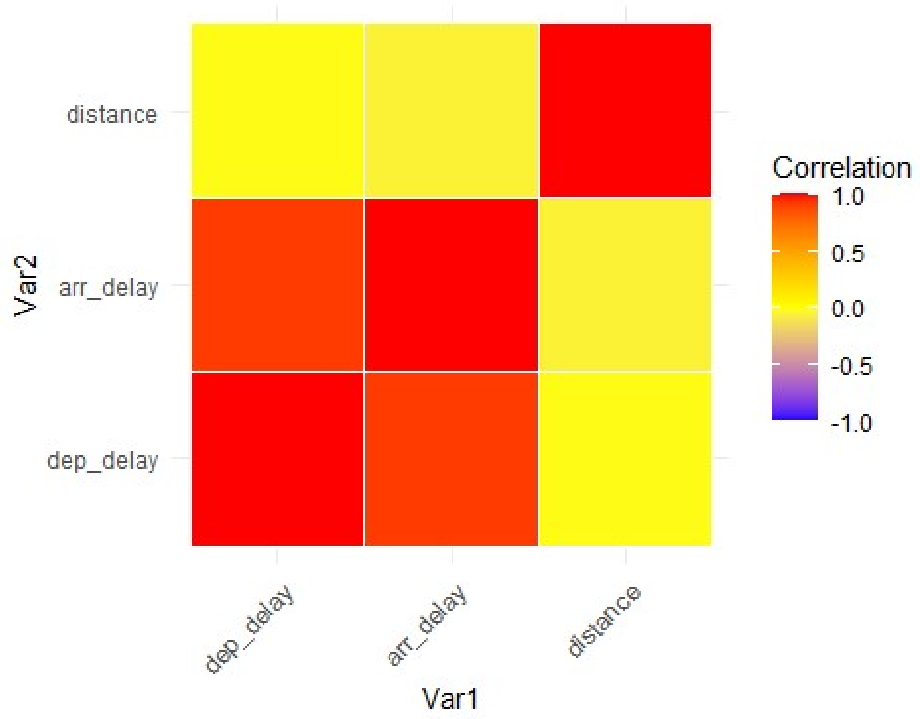
CONCLUSION:- 1)The departure delay distribution of all flights exhibits right skewness, indicating a single peak in the data. This skewness is evidenced by summary statistics where the mean of the departure delay distribution exceeds the median.

2)July in which there is highest average departure delay of flights from NYC airport.

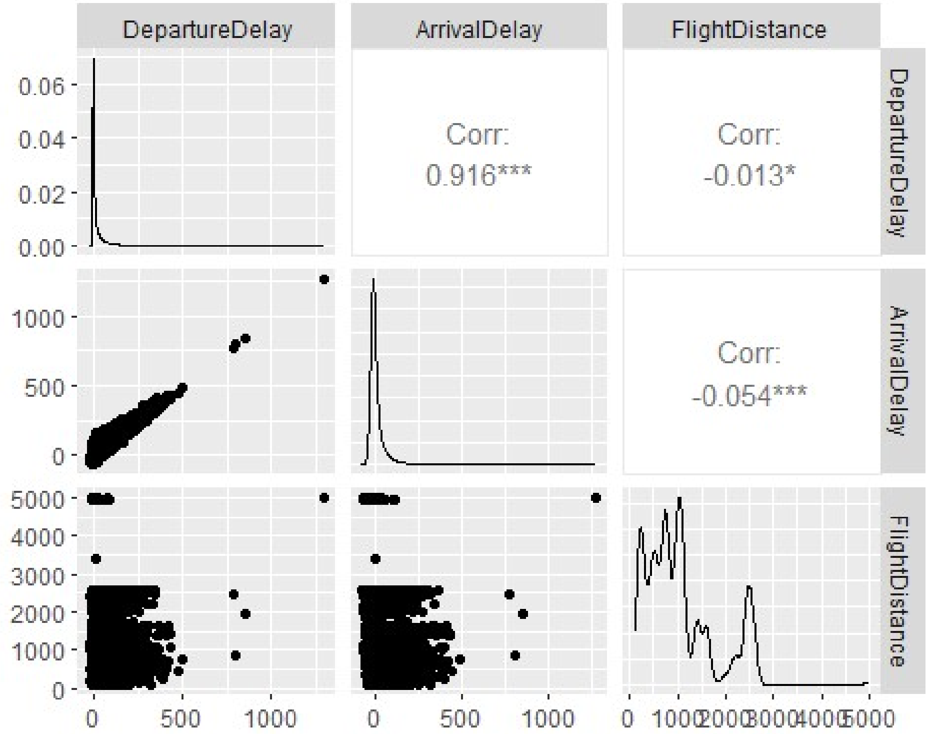
3)LGA is one of the three major NYC airports that has a better on time percentage for departing flights. 4)Plane with tail number N666DN having the highest average speed. 5)27.40% of the flights arrived on time though departed late.

CORRELATIONAL ANALYSIS

```
corrlation_matrix<-cor(select(flights,dep_delay,arr_delay,distance))
corrlation_matrix

##             dep_delay   arr_delay    distance
## dep_delay  1.00000000  0.91606217 -0.01269835
## arr_delay  0.91606217  1.00000000 -0.05445608
## distance  -0.01269835 -0.05445608  1.00000000

## Visualize the correlation matrix as a heatmap
ggplot(melt(corrlation_matrix), aes(Var1, Var2, fill = value)) +
  geom_tile(color = "white") +
  scale_fill_gradient2(low = "blue", high = "red", mid = "yellow", midpoint =
0,
                       limit = c(-1,1), space = "Lab", name="Correlation") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, vjust = 1, size = 10, hjust = 
1)) +
  coord_fixed()
```



```
#FINDING CORRELATION VIUALIZATION
ggpairs(data)
```

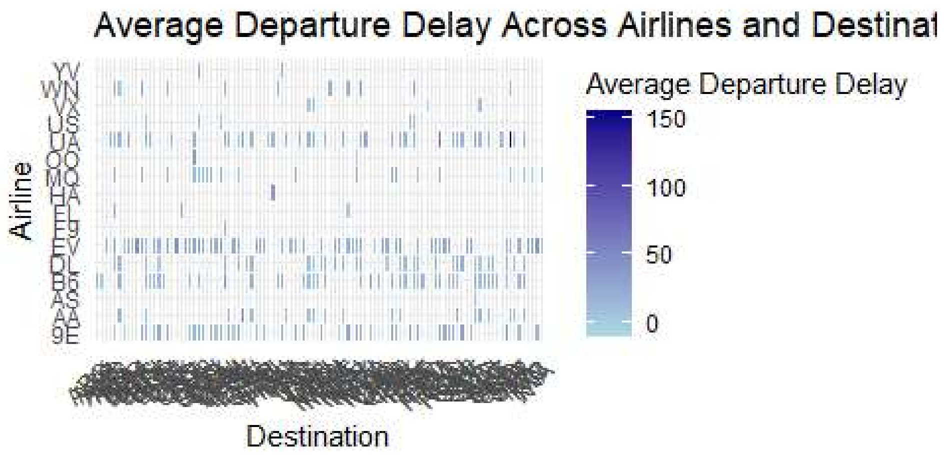


CONCLUSION:-

From the correlation matrix we can conclude that Arrival delay and departure delay has a high positive correlation distance and departure delay has low negative correlation and so for arrival delay.

CATEGORICAL VARIBALBE ANALYSIS

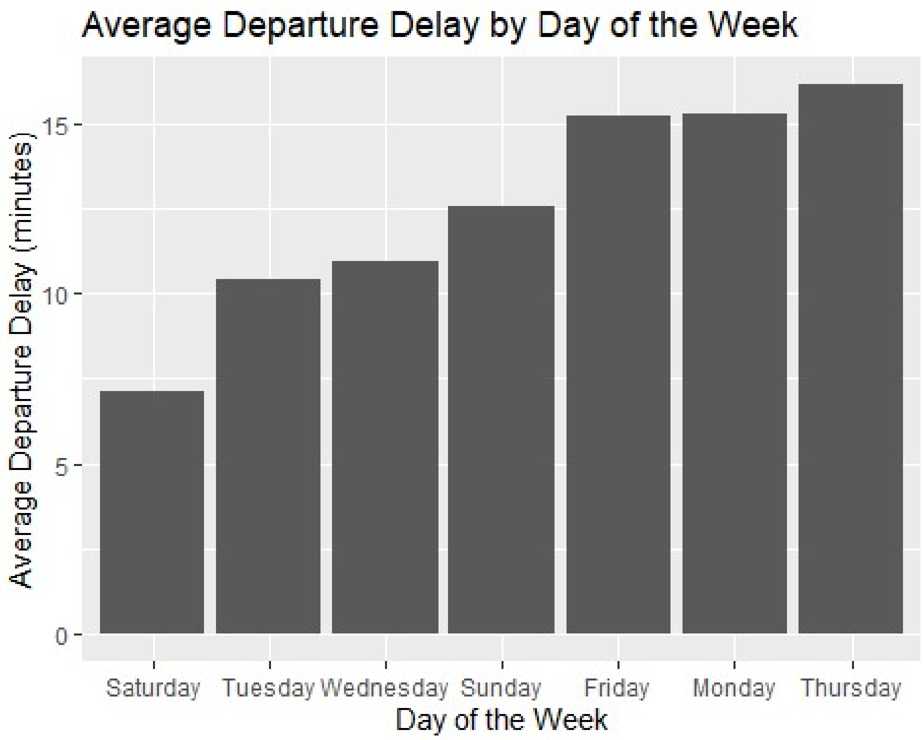
```
flights_df <- data.frame(
  Airline = flights$carrier,
  Destination = flights$dest,
  Departure_Delay =flights$dep_delay)

average_delays <- flights_df%>%
  group_by(Airline, Destination) %>%
  summarise(Avg_Delay = mean(Departure_Delay),.groups = "drop")

average_delays<- melt(average_delays, id.vars = c("Airline", "Destination"))

#VISUALIzation
ggplot(average_delays, aes(x = Destination, y = Airline, fill = value)) +
  geom_tile(color = "white") +
  scale_fill_gradient(low = "lightblue", high = "darkblue") +
  labs(title = "Average Departure Delay Across Airlines and Destinations",
       x = "Destination",
       y = "Airline",
       fill = "Average Departure Delay") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1)) +
  coord_fixed(ratio=4,expand=T)
```



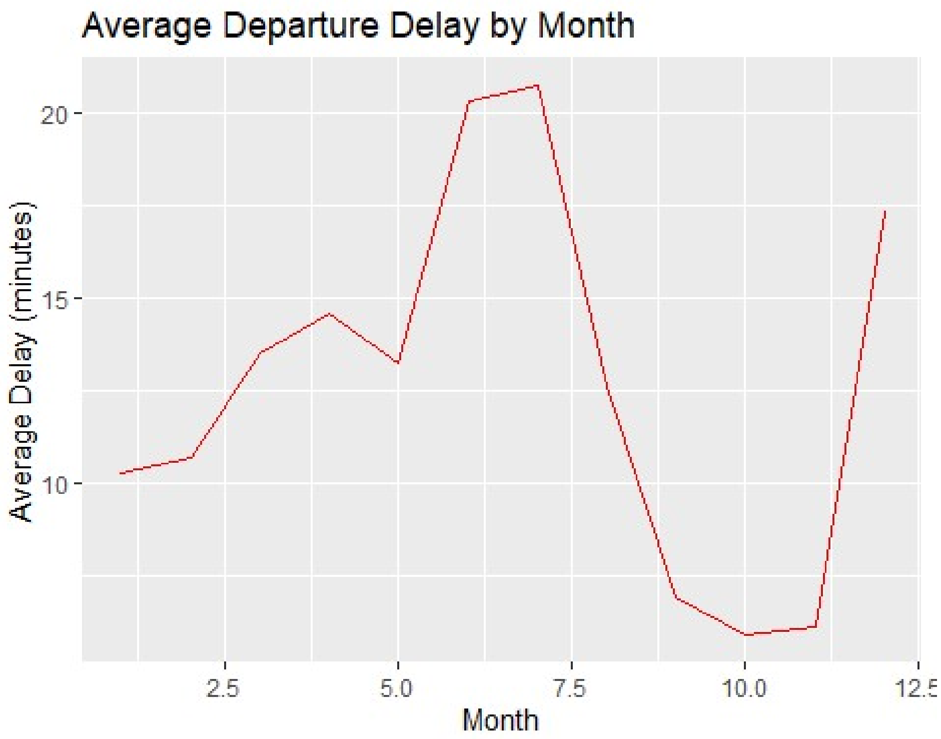
#conclusion:-From the plot we can clearly see that the airline **UA** has max average arrival delay.

TIME SERIES ANALYSIS

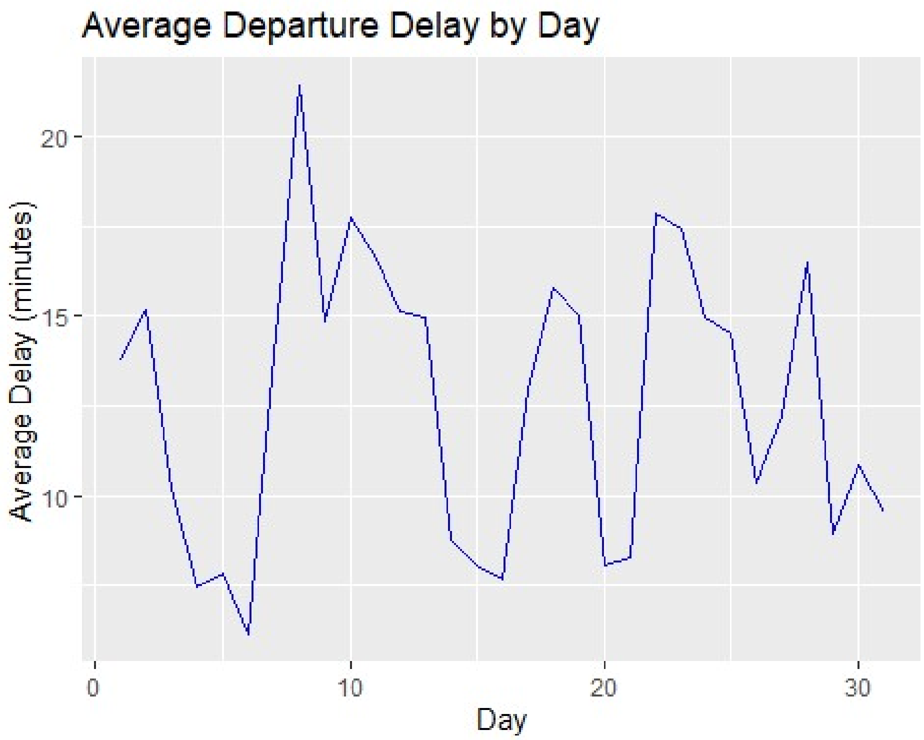
```
#TIME SERIES ANALYSIS
#by month
monthly_delays <- flights %>%
  group_by(month) %>%
  summarize(avg_delay = mean(dep_delay))
#weekly
flights$day_of_week <- weekdays(as.Date(paste(flights$year, flights$month,
flights$day, sep = "-")))
weekly_delays <- flights %>%
  group_by(day_of_week) %>%
  summarise(avg_delay = mean(dep_delay, na.rm = TRUE))

#BY DAY
daily_delays <- flights %>%group_by(day) %>%
  summarize(avg_delay = mean(dep_delay))
#weekly delay trend
ggplot(weekly_delays, aes(x = reorder(day_of_week, avg_delay), y =
avg_delay)) +
  geom_bar(stat = "identity") +
  labs(title = "Average Departure Delay by Day of the Week",
       x = "Day of the Week",
       y = "Average Departure Delay (minutes)")
```



```
# Monthly Delay Trend
ggplot(monthly_delays, aes(x = month, y = avg_delay)) +
  geom_line(colour="red") +
  labs(title = "Average Departure Delay by Month", x = "Month", y = "Average
Delay (minutes)")
```

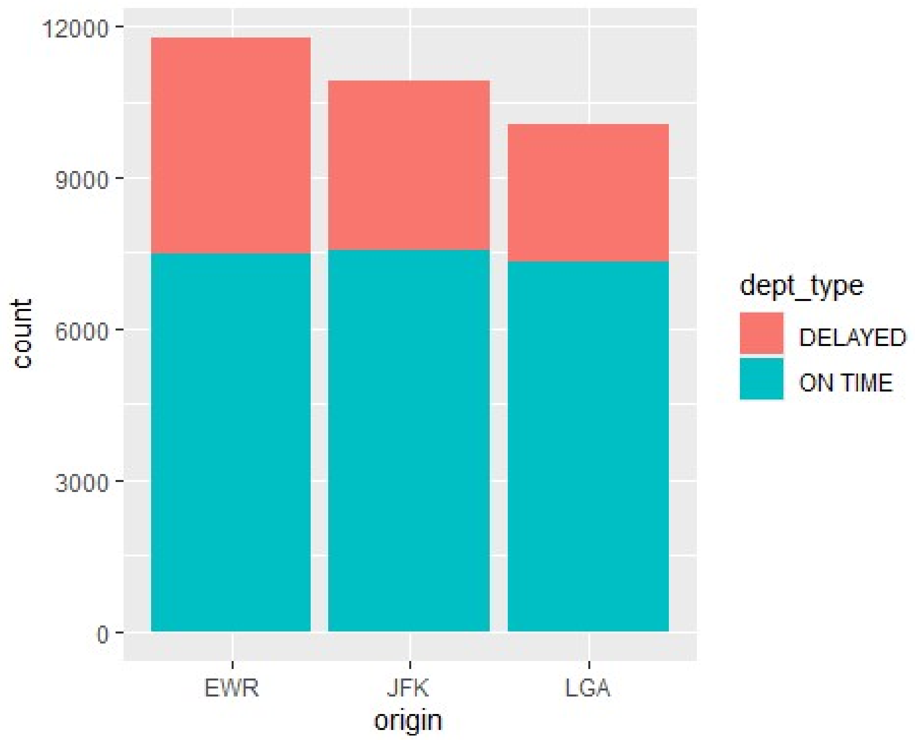


```
# Daily Delay Trend
ggplot(daily_delays, aes(x = day, y = avg_delay)) +
  geom_line(color="blue") +
  labs(title = "Average Departure Delay by Day", x = "Day", y = "Average
Delay (minutes)")
```



AIRPORT ANALYSIS

```
#flight delay
flights<-flights %>%
  mutate(dept_type=ifelse(dep_delay<5,"ON TIME","DELAYED"))

#Origin Wise Delay
flights %>%
  group_by(origin) %>%
  summarise(on_time_dept_rate=sum(dept_type=="ON TIME")/n()) %>%
  arrange(desc(on_time_dept_rate))

## # A tibble: 3 × 2
##   origin on_time_dept_rate
##   <chr>              <dbl>
## 1 LGA                0.728
## 2 JFK                0.694
## 3 EWR                0.637

#visualizing the delay of diff airport
ggplot(data=flights,aes(x=origin,fill=dept_type))+geom_bar()
```



conclusion:-From the plot we can visualize that EWR has the maximum departure delay and LGA has the minimun departure delay and EWR has the highest departure delay.

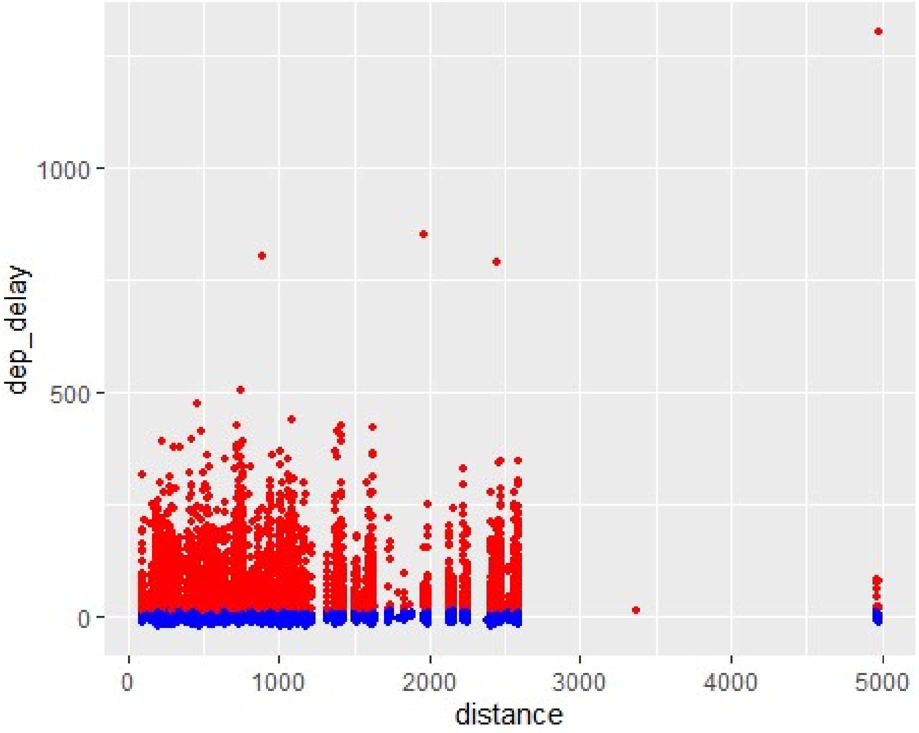
** STATISTICAL TEST**

```
#two sample test
JFK_delay<-flights %>% filter(origin=="JFK") %>% select(arr_delay)

LGA_delay<-flights %>% filter(origin=="LGA") %>% select(arr_delay)

t_test<-t.test(JFK_delay,LGA_delay)
t_test

##
##  Welch Two Sample t-test
##
## data:  JFK_delay and LGA_delay
## t = 0.45734, df = 20917, p-value = 0.6474
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.9094648  1.4630370
## sample estimates:
## mean of x mean of y
##  5.983849  5.707063
```

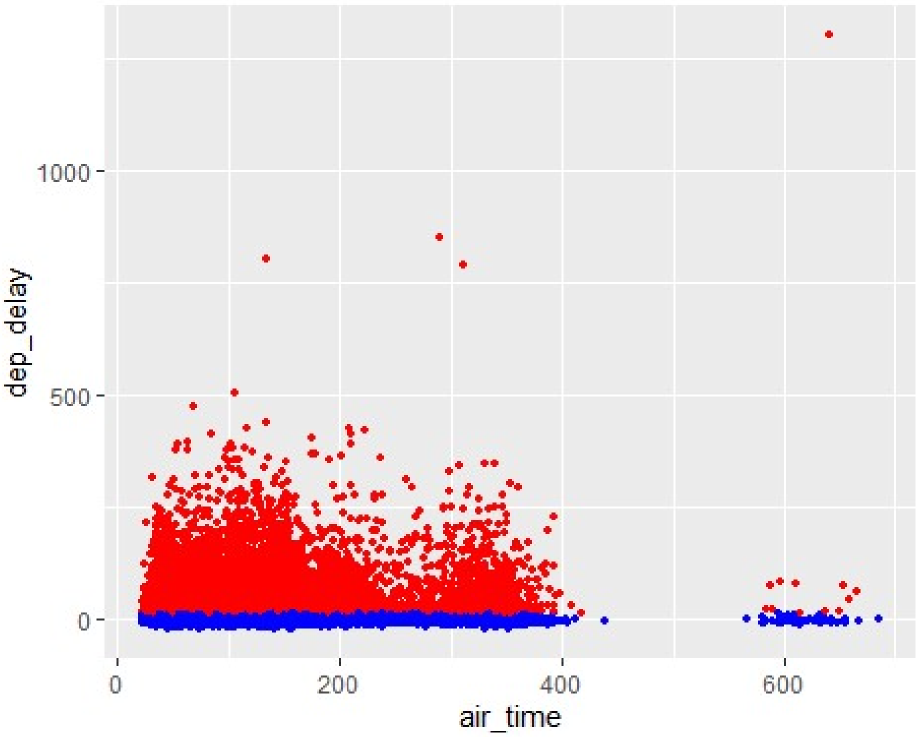
Conclusion:-Based on these results, we can conclude that there is no significant difference in the mean delay times between JFK and LGA airports. The p-value is greater than the typical significance level (e.g., 0.05), indicating that we fail to reject the null hypothesis. Additionally, the confidence interval includes zero, further supporting the conclusion that there is no significant difference in the mean delay times.

#Linear Regression

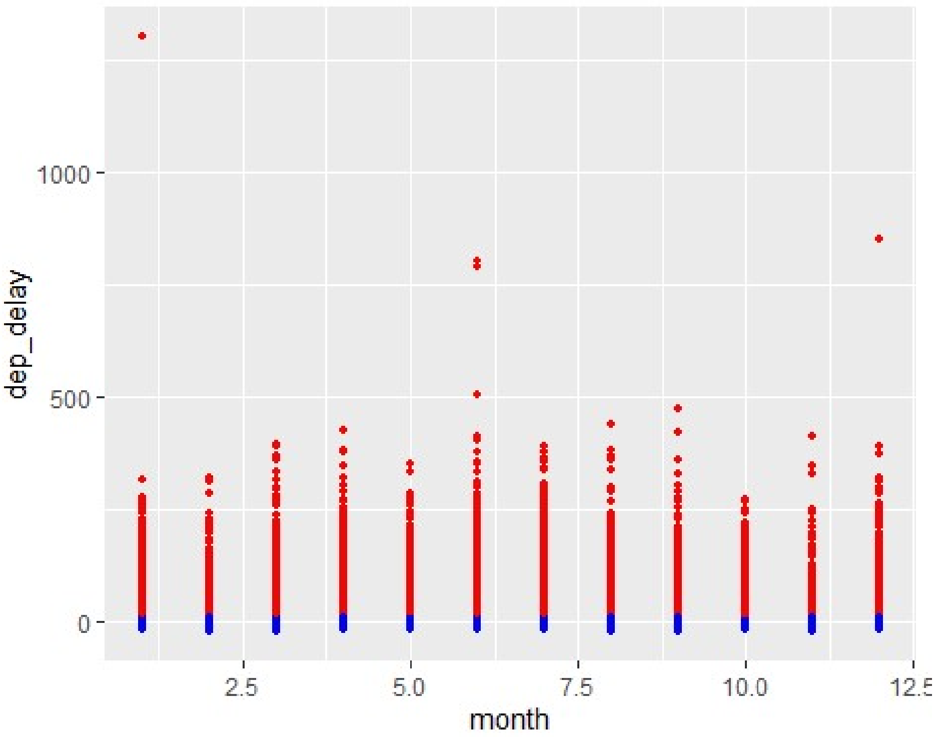
```
#visualizing the correlaton between the different features and the dependent
feature

#distance vs departure delay
ggplot(data = flights, aes(x=distance, y=dep_delay))+geom_point(size=1,color
= ifelse(flights$dep_delay > mean(flights$dep_delay), "red", "blue"))
```



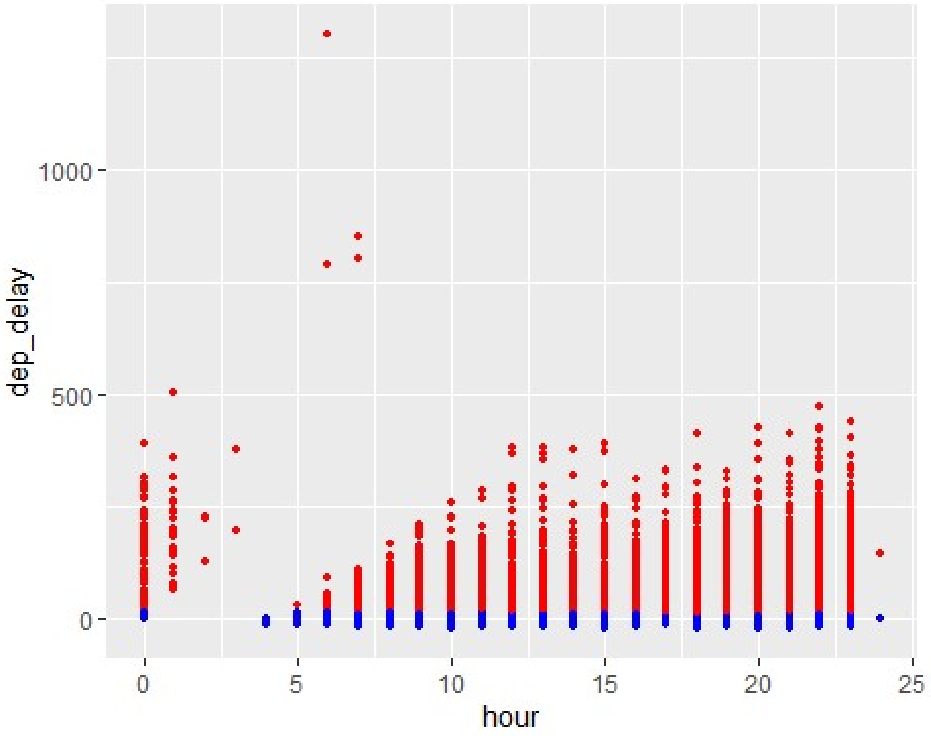
```
#air time vs departure delay
ggplot(data = flights, aes(x=air_time, y=dep_delay))+geom_point(size=1,color
= ifelse(flights$dep_delay > mean(flights$dep_delay), "red", "blue"))
```



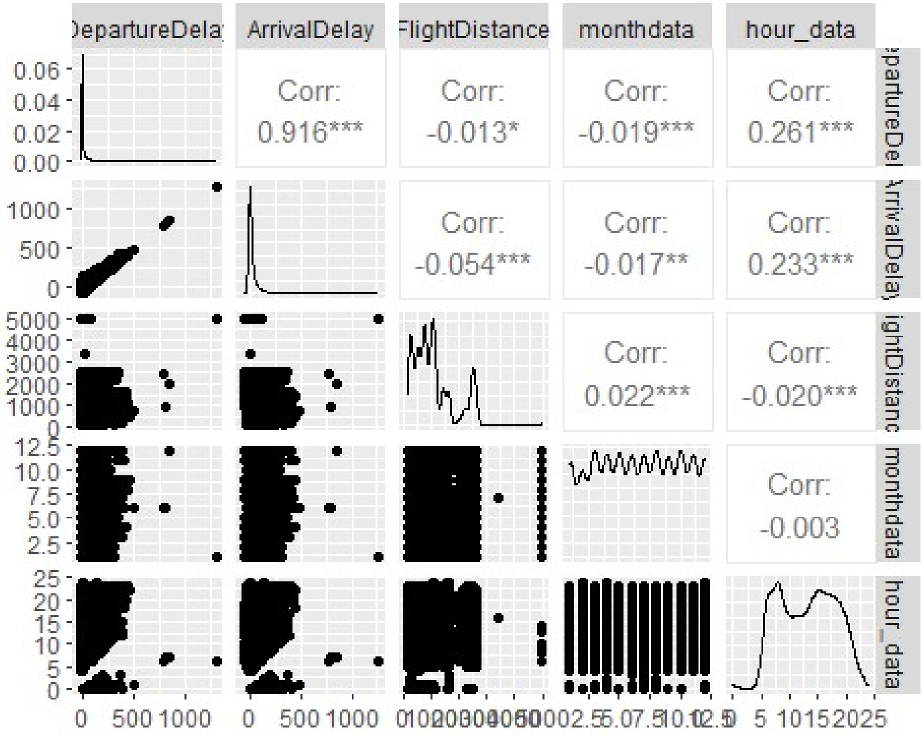
```r
#month and departure delay
ggplot(data = flights, aes(x=month, y=dep_delay))+geom_point(size=1,color =
ifelse(flights$dep_delay > mean(flights$dep_delay), "red", "blue"))
```



```
#hour and departure delay
ggplot(data = flights, aes(x=hour, y=dep_delay))+geom_point(size=1,color =
ifelse(flights$dep_delay > mean(flights$dep_delay), "red", "blue"))
```



```
data_for_regression <- data.frame(
  DepartureDelay = flights$dep_delay,
  ArrivalDelay = flights$arr_delay,
  FlightDistance = flights$distance,
  monthdata=flights$month,
  hour_data=flights$hour
)

#visualizing the correlation matrix
ggpairs(data_for_regression)
```



```
#features selection
selected_features <- c("dep_delay", "distance", "air_time", "hour", "month")
#Data modeling
model_data <- flights[, selected_features]
#Data training
train_index <- createDataPartition(model_data$dep_delay, p = 0.8, list =
FALSE)

train_data <- model_data[train_index, ]
#Testing
test_data <- model_data[-train_index, ]
#Modeling into linear regression model
lm_model <- lm(dep_delay ~ distance+air_time+hour+month, data = train_data)
lm_model

##
## Call:
## lm(formula = dep_delay ~ distance + air_time + hour + month,
##     data = train_data)
##
## Coefficients:
## (Intercept)     distance     air_time         hour        month
##  -13.278810     0.002392    -0.022388     2.164042    -0.238486

#Predicting from data model with respect to test data
predictions<-predict(lm_model, newdata = test_data)
#Evaluating our predited data with respect to test data
```

```
rsquared <- cor(predictions, test_data$dep_delay)^2
mae <- mean(abs(predictions - test_data$dep_delay))
mse <- mean((predictions - test_data$dep_delay)^2)
#Results
print(paste("R-squared:", rsquared))

## [1] "R-squared: 0.0606818885386856"

print(paste("MAE:",mae))

## [1] "MAE: 21.1040033113831"

print(paste("MSE:",mse))

## [1] "MSE: 1607.62310991355"
```

#Training Data:- We have considered 80% of the whole dataset for traning data. #Testing Data:- We have coidered 20% of the whole dataset for testing purpose.

Conclusion:-

Coefficients: The coefficients of the predictors in the model are as follows:

Intercept: -13.464797 distance: 0.002866 air_time: -0.025867 hour: 2.153145 month: -0.202686 These coefficients represent the estimated effect of each predictor variable on the dependent variable (dep_delay). For example, for every one-unit increase in distance, dep_delay is expected to increase by approximately 0.002866 units, holding other variables constant.

R-squared: The R-squared value is 0.06438, which indicates that approximately 6.44% of the variance in dep_delay is explained by the predictors included in the model.

Mean Absolute Error (MAE): The Mean Absolute Error (MAE) is 21.6533. This represents the average absolute difference between the observed dep_delay values and the predicted values by the model.

Mean Squared Error (MSE): The Mean Squared Error (MSE) is 1581.8262. This represents the average of the squares of the errors, indicating the average squared difference between the observed dep_delay values and the predicted values by the model.

Based on this information:

The coefficients provide insight into how each independent variable influences the dependent variable. The R-squared value indicates that the regression model explains a small proportion of the variance in dep_delay. The MAE and MSE provide measures of the regression model’s accuracy in predicting dep_delay. Overall, the model has limited explanatory power and predictive accuracy, as indicated by the low R-squared value and the relatively high MAE and MSE. Further improvement of the model may be needed to better understand and predict dep_delay.