# Problem Statement

**Problem Statement:**

Airport arrival performance and other events are affected by weather, which may result in delays or capacity constraints. Almost half of all airport traffic delays are caused by adverse weather conditions. These uncertainties during airport operations can result in significant delays and inconvenience to passengers. Therefore, the airport authority wants to analyze the flights that are delayed and the effect of weather on the delays.

**Objective:** To visualize the data with the help of histograms, scatter plots, box plots, and pie charts and understand the effect of weather conditions and other factors on flight delays

**Data Set:** flightdelays.xlsx

**Data Description**

|  |  |
| --- | --- |
| **Variable** | **Description** |
| schedtime | Scheduled time |
| Carrier | Airline codes |
| deptime | Time of departure |
| dest | Destination of flight |
| distance | Travelling distance |
| date | Date of travel |
| flightnum | Flight number |
| origin | Airport codes |
| weather | Coded as:  0 – ontime  1 - delayed |
| dayweek | Coded as:  1 - Sunday and Monday 0 - for all other days |
| daymonth | Number of days in month |
| tailnu | Tail number of flight |
| delay | Delay status |

**Steps to Perform:**

1. Read the dataset
2. Read the dataset description
3. Understand the data
4. Find out the null values
5. Install the required packages
6. Understand the summary of descriptive statistics
7. Plot the histograms to understand the relationships between scheduled time, carrier, destination, origin, weather, and day of the week
8. Plot the scatter plot for flights on time and delayed
9. Plot the box plot to understand how many days in a month flights are delayed by what time
10. Define the hours of departure
11. Create a categorical representation of data using a table
12. Redefine the delay variables
13. Understand the summary of major variables
14. Plot histograms of major variables
15. Plot a pie chart to see how many flights were delayed

# Task

# Step: Install and load the `readxl` package

# install.packages("readxl")

library(readxl)

# Step 1: Load the dataset

flights <- read\_excel("flightdelays.xlsx")

# Step 2: Attach descriptions using attributes

attributes(flights)$variable.description <- list(

schedtime = "Scheduled time",

Carrier = "Airline codes",

deptime = "Time of departure",

dest = "Destination of flight",

distance = "Traveling distance",

date = "Date of travel",

flightnum = "Flight number",

origin = "Airport codes",

weather = "0 – on time, 1 - delayed",

dayweek = "1 - Sunday and Monday, 0 - for all other days",

daymonth = "Number of days in month",

tailnu = "Tail number of flight",

delay = "Delay status"

)

# Function to print the variable descriptions

print\_variable\_descriptions <- function(df) {

cat("Variable Descriptions:\n")

for (var in names(attributes(df)$variable.description)) {

cat(var, ":", attributes(df)$variable.description[[var]], "\n")

}

}

# Print the variable descriptions

print\_variable\_descriptions(flights)

A screen shot of a computer code

Description automatically generated

# Step 3. Understand the data

str(flights)

A screenshot of a computer

Description automatically generated

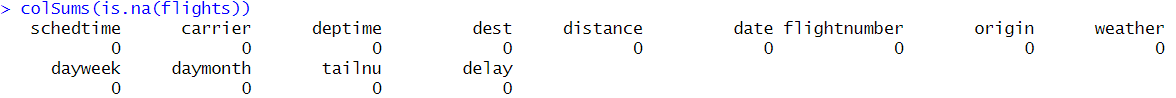
summary(flights)

A screenshot of a computer program

Description automatically generated

# Step 4. Find out the null values

colSums(is.na(flights))



# Step 5. Install the required packages

install.packages(c("ggplot2", "dplyr"))

library(ggplot2)

library(dplyr)

# Step 6. Understand the summary of descriptive statistics

summary(flights)

A screenshot of a computer screen

Description automatically generated

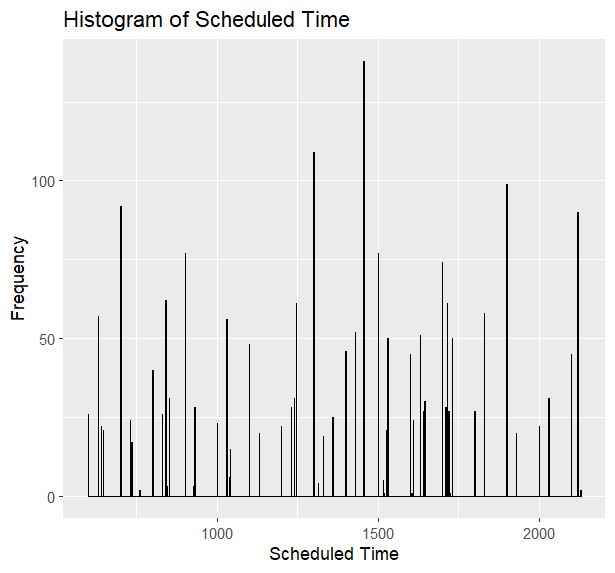
# Step 7. Plot the histograms

# Histogram for scheduled time

ggplot(flights, aes(x = schedtime)) +

geom\_histogram(binwidth = 1, fill = "blue", color = "black") +

labs(title = "Histogram of Scheduled Time", x = "Scheduled Time", y = "Frequency")

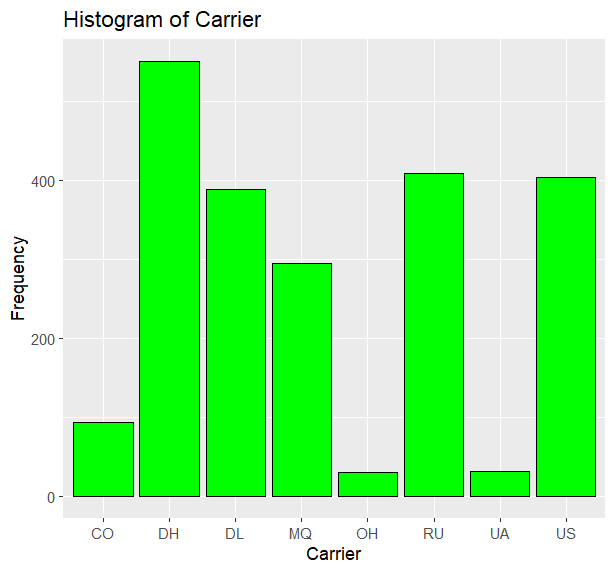


# Histogram for carrier

ggplot(flights, aes(x = carrier)) +

geom\_bar(fill = "green", color = "black") +

labs(title = "Histogram of Carrier", x = "Carrier", y = "Frequency")

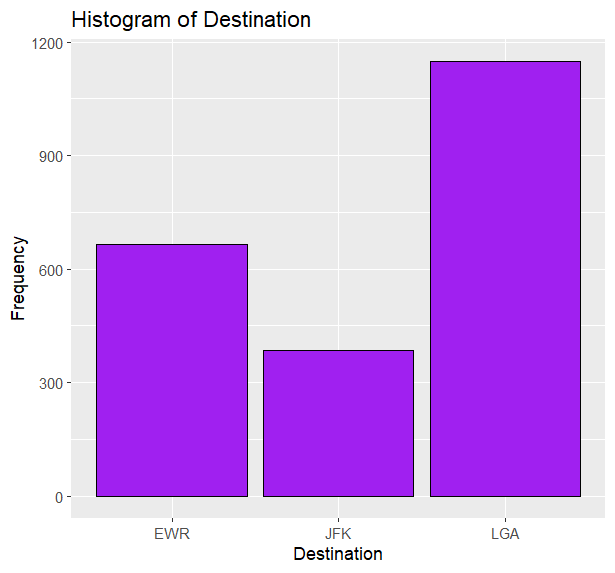


# Histogram for destination

ggplot(flights, aes(x = dest)) +

geom\_bar(fill = "purple", color = "black") +

labs(title = "Histogram of Destination", x = "Destination", y = "Frequency")

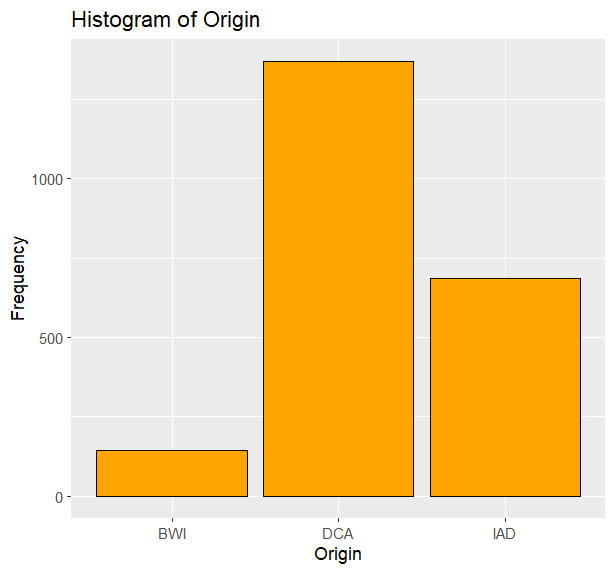


# Histogram for origin

ggplot(flights, aes(x = origin)) +

geom\_bar(fill = "orange", color = "black") +

labs(title = "Histogram of Origin", x = "Origin", y = "Frequency")

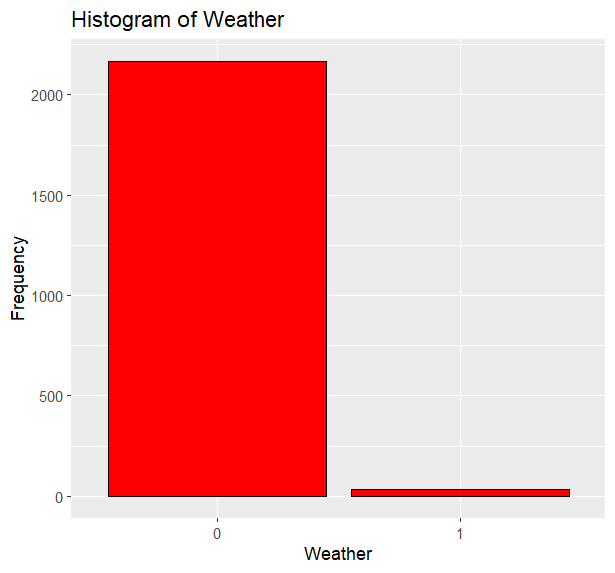


# Histogram for weather

ggplot(flights, aes(x = factor(weather))) +

geom\_bar(fill = "red", color = "black") +

labs(title = "Histogram of Weather", x = "Weather", y = "Frequency")

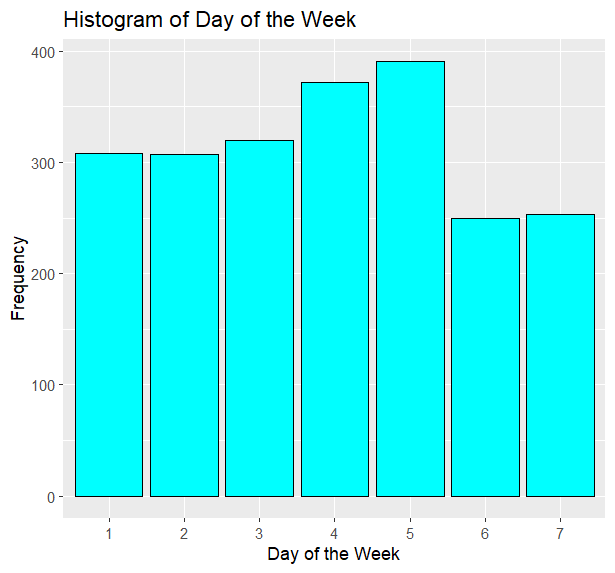


# Histogram for day of the week

ggplot(flights, aes(x = factor(dayweek))) +

geom\_bar(fill = "cyan", color = "black") +

labs(title = "Histogram of Day of the Week", x = "Day of the Week", y = "Frequency")

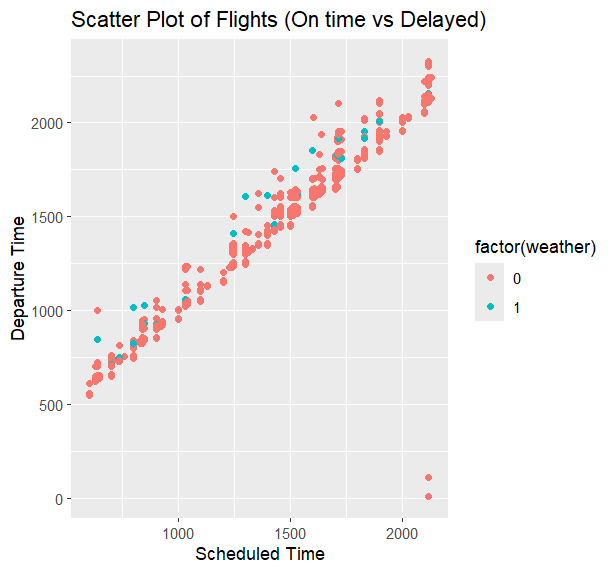


# Step 8. Plot the scatter plot for flights on time and delayed

ggplot(flights, aes(x = schedtime, y = deptime, color = factor(weather))) +

geom\_point() +

labs(title = "Scatter Plot of Flights (On time vs Delayed)", x = "Scheduled Time", y = "Departure Time")

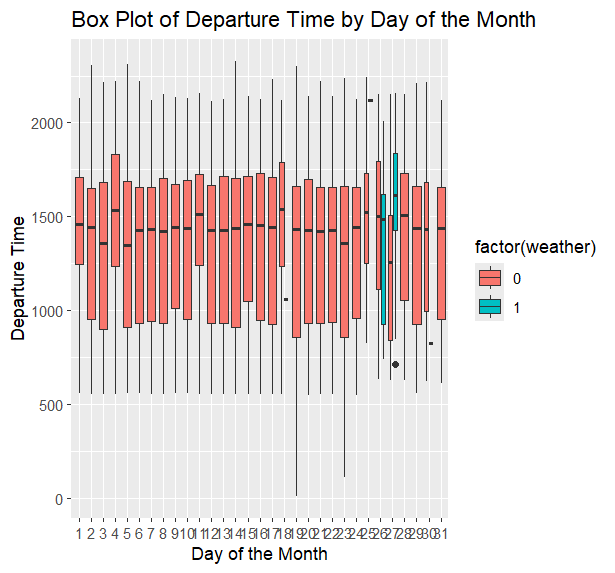


# Step 9. Plot the box plot to understand how many days in a month flights are delayed by what time

ggplot(flights, aes(x = factor(daymonth), y = deptime, fill = factor(weather))) +

geom\_boxplot() +

labs(title = "Box Plot of Departure Time by Day of the Month", x = "Day of the Month", y = "Departure Time")



# Step 10. Define the hours of departure

flights$hour\_of\_departure <- as.integer(flights$deptime / 100)

# Step 11. Create a categorical representation of data using a table

table(flights$weather, flights$dayweek)

A number and text on a white background

Description automatically generated

# Step 12. Redefine the delay variables

flights$delay\_status <- ifelse(flights$delay == 1, "Delayed", "Ontime")

# Step 13. Understand the summary of major variables

summary(flights[c("schedtime", "deptime", "distance", "hour\_of\_departure", "weather")])

A screen shot of a computer

Description automatically generated

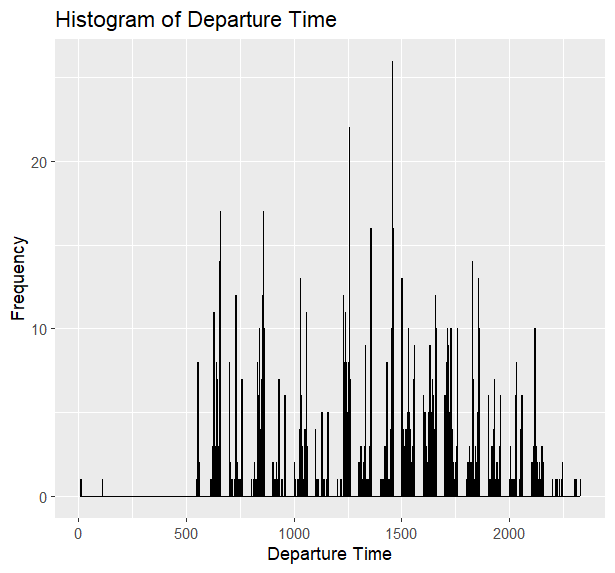
# Step 14. Plot histograms of major variables

# Histogram for departure time

ggplot(flights, aes(x = deptime)) +

geom\_histogram(binwidth = 1, fill = "blue", color = "black") +

labs(title = "Histogram of Departure Time", x = "Departure Time", y = "Frequency")

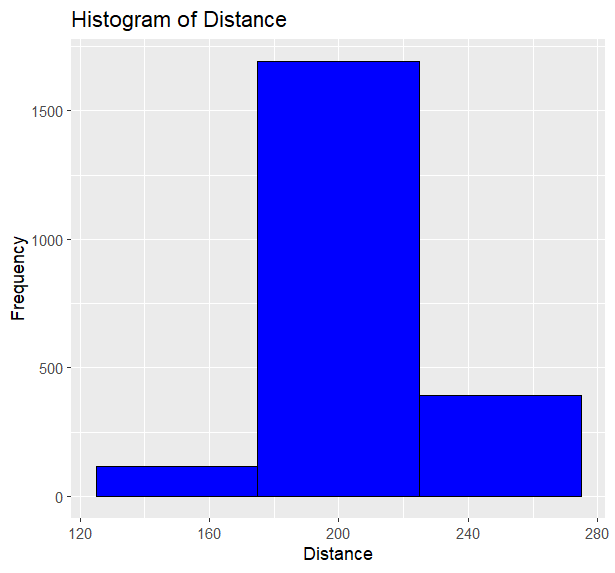


# Histogram for distance

ggplot(flights, aes(x = distance)) +

geom\_histogram(binwidth = 50, fill = "blue", color = "black") +

labs(title = "Histogram of Distance", x = "Distance", y = "Frequency")

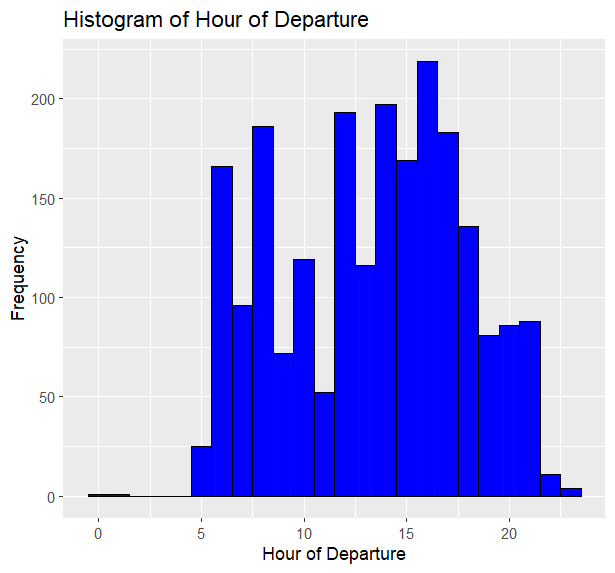


# Histogram for hour of departure

ggplot(flights, aes(x = hour\_of\_departure)) +

geom\_histogram(binwidth = 1, fill = "blue", color = "black") +

labs(title = "Histogram of Hour of Departure", x = "Hour of Departure", y = "Frequency")



# Step 15. Plot a pie chart to see how many flights were delayed

# Count the number of delayed flights

delayed\_flights <- table(flights$delay)

# Pie chart of delayed flights

pie(delayed\_flights,

labels = paste(names(delayed\_flights), "(", delayed\_flights, " flights)", sep = " "),

main = "Pie Chart of Flight Delays")

