Flight Delays Analysis using R

Manish Patwardhan

27-01-2026

Step 1: Install Required Packages

```
install.packages("readxl")
library(readxl)

install.packages("ggplot2")
library(ggplot2)

install.packages("dplyr")
library(dplyr)
```

Explanation: In this step, we install and load the necessary R libraries such as **readxl**, **ggplot2**, and **dplyr** to perform data analysis and visualization.

Step 2: Read the Dataset + Understand Data

```
# Load dataset
flight_data <- read_excel("flightdelays.xlsx")

# Viwe first rows
head(flight_data)
## # A tibble: 6 × 13
##   schedtime carrier deptime dest  distance date  flightnumber origin weath
er
##       <dbl> <chr>     <dbl> <chr>    <dbl> <chr>        <dbl> <chr>    <db
l>
## 1      1455 OH         1455 JFK        184 37987         5935 BWI
0
## 2      1640 DH         1640 JFK        213 37987         6155 DCA
0
## 3      1245 DH         1245 LGA        229 37987         7208 IAD
0
## 4      1715 DH         1709 LGA        229 37987         7215 IAD
0
## 5      1039 DH         1035 LGA        229 37987         7792 IAD
0
## 6       840 DH          839 JFK        228 37987         7800 IAD
0
## # ℹ 4 more variables: dayweek <dbl>, daymonth <dbl>, tailnu <chr>, delay <
chr>
```

```r
# Structure of dataset
str(flight_data)
## tibble [2,201 × 13] (S3: tbl_df/tbl/data.frame)
##  $ schedtime   : num [1:2201] 1455 1640 1245 1715 1039 ...
##  $ carrier     : chr [1:2201] "OH" "DH" "DH" "DH" ...
##  $ deptime     : num [1:2201] 1455 1640 1245 1709 1035 ...
##  $ dest        : chr [1:2201] "JFK" "JFK" "LGA" "LGA" ...
##  $ distance    : num [1:2201] 184 213 229 229 229 228 228 228 228 228 ...
##  $ date        : chr [1:2201] "37987" "37987" "37987" "37987" ...
##  $ flightnumber: num [1:2201] 5935 6155 7208 7215 7792 ...
##  $ origin      : chr [1:2201] "BWI" "DCA" "IAD" "IAD" ...
##  $ weather     : num [1:2201] 0 0 0 0 0 0 0 0 0 0 ...
##  $ dayweek     : num [1:2201] 4 4 4 4 4 4 4 4 4 4 ...
##  $ daymonth    : num [1:2201] 1 1 1 1 1 1 1 1 1 1 ...
##  $ tailnu      : chr [1:2201] "N940CA" "N405FJ" "N695BR" "N662BR" ...
##  $ delay       : chr [1:2201] "ontime" "ontime" "ontime" "ontime" ...

# Dimensions (rows, columns)
dim(flight_data)
## [1] 2201   13

# Check missing values
colSums(is.na(flight_data))
##    schedtime      carrier      deptime         dest     distance         d
ate
##            0            0            0            0            0
0
## flightnumber       origin      weather      dayweek     daymonth       tai
lnu
##            0            0            0            0            0
0
##        delay
##            0

# Dimensions (rows, columns)
dim(flight_data)
## [1] 2201   13
```

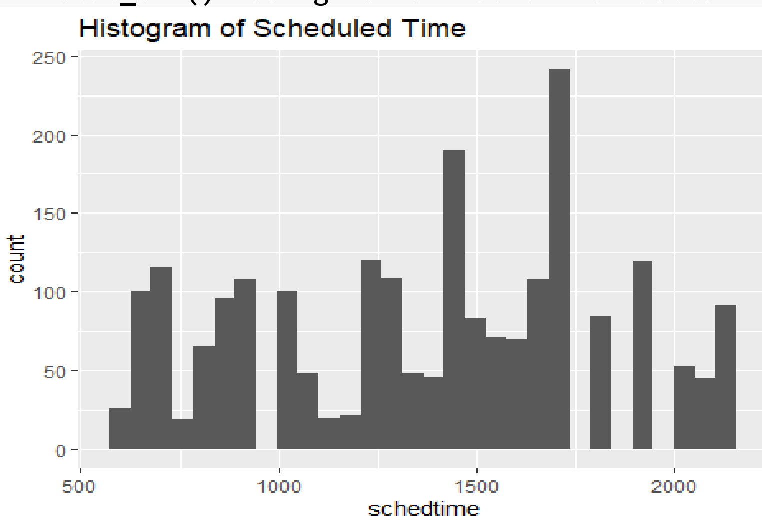
Explanation: In this step, we import the flight delays dataset into R and examine its structure, column names, and sample rows to understand the data.

Step 3: Summary of Descriptive Statistics

```r
# Summary of Descriptive Statistics
summary(flight_data)
##    schedtime      carrier              deptime         dest
##  Min.   : 600   Length:2201        Min.   :  10   Length:2201
##  1st Qu.:1000   Class :character   1st Qu.:1004   Class :character
```

```
##  Median :1455   Mode  :character   Median :1450   Mode  :character
##  Mean   :1372                      Mean   :1369
##  3rd Qu.:1710                      3rd Qu.:1709
##  Max.   :2130                      Max.   :2330
##     distance          date            flightnumber     origin
##  Min.   :169.0   Length:2201        Min.   : 746   Length:2201
##  1st Qu.:213.0   Class :character   1st Qu.:2156   Class :character
##  Median :214.0   Mode  :character   Median :2385   Mode  :character
##  Mean   :211.9                      Mean   :3815
##  3rd Qu.:214.0                      3rd Qu.:6155
##  Max.   :229.0                      Max.   :7924
##     weather            dayweek         daymonth        tailnu
##  Min.   :0.00000   Min.   :1.000   Min.   : 1.00   Length:2201
##  1st Qu.:0.00000   1st Qu.:2.000   1st Qu.: 8.00   Class :character
##  Median :0.00000   Median :4.000   Median :16.00   Mode  :character
##  Mean   :0.01454   Mean   :3.905   Mean   :16.02
##  3rd Qu.:0.00000   3rd Qu.:5.000   3rd Qu.:23.00
##  Max.   :1.00000   Max.   :7.000   Max.   :31.00
##     delay
##  Length:2201
##  Class :character
##  Mode  :character
##
##
##
```

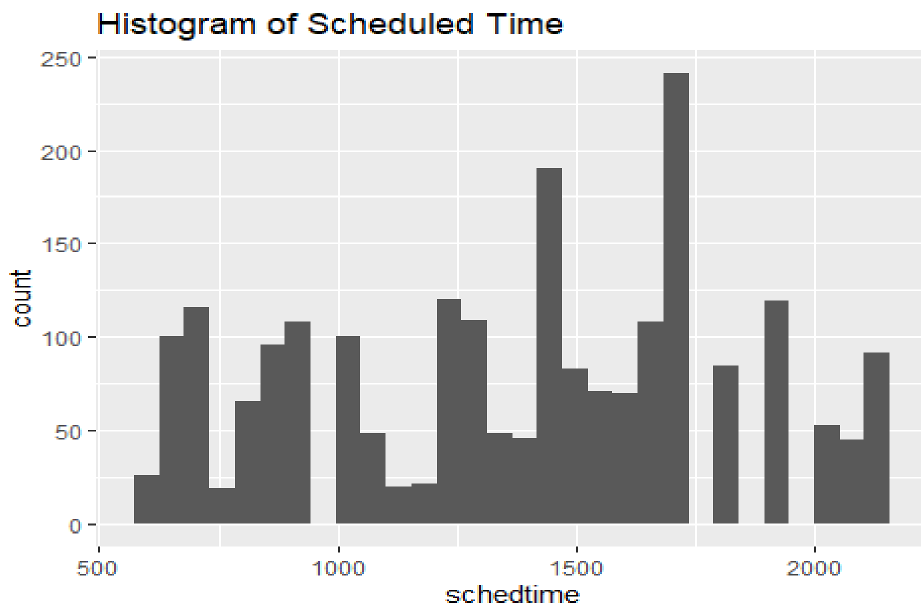
```r
# Histograms (Relationships)
ggplot(flight_data, aes(x = schedtime)) +
  geom_histogram() +
  ggtitle("Histogram of Scheduled Time")
## `stat_bin()` using `bins = 30`. Pick better value `binwidth`.
```



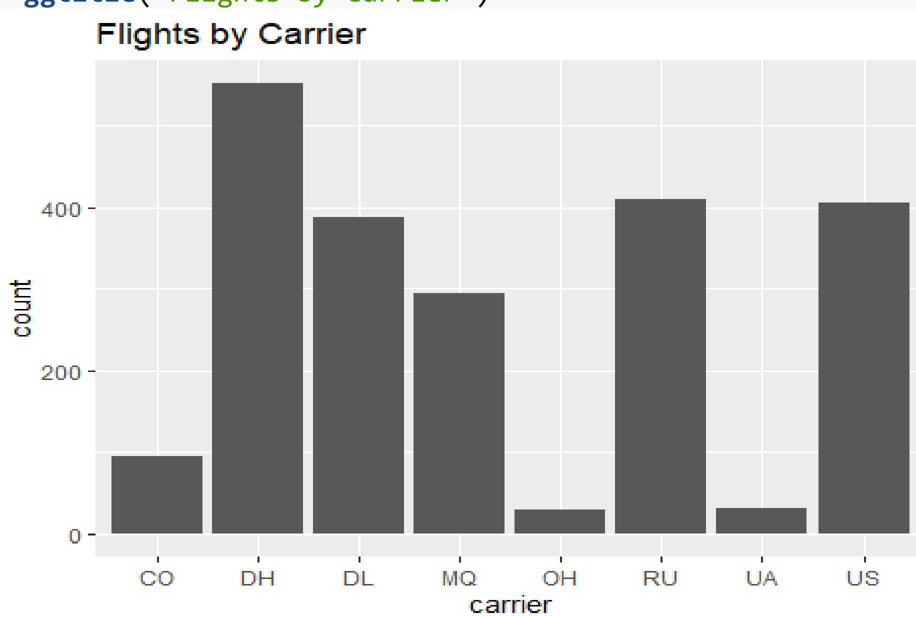
Explanation: In this step, we generate summary statistics such as minimum, maximum, mean, and median values to understand the overall distribution of key variables.

Step 4: Histograms (Relationships)

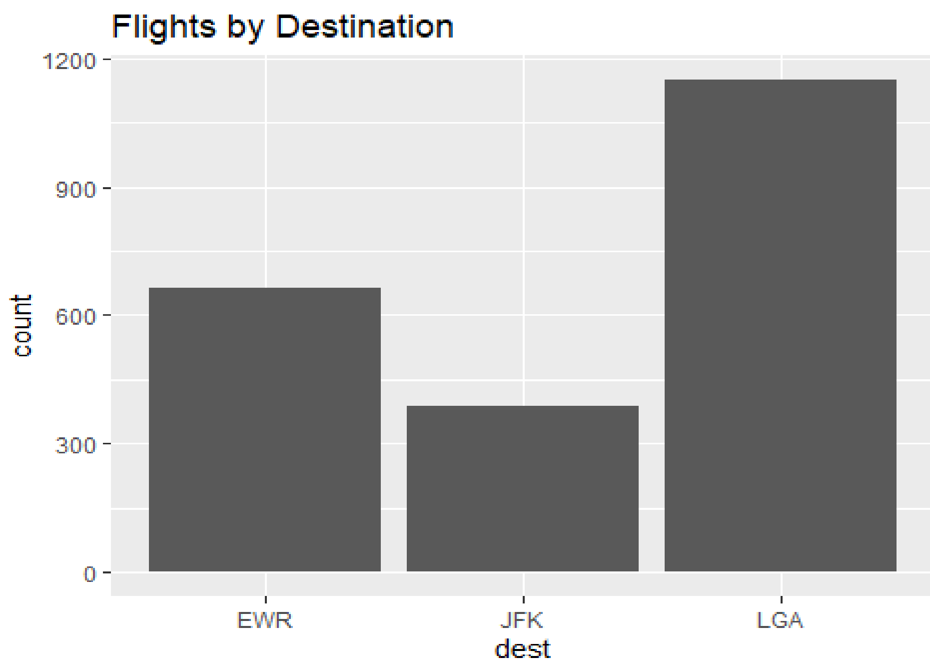
```
# Scheduled Time
ggplot(flight_data, aes(x = schedtime)) +
  geom_histogram() +
  ggtitle("Histogram of Scheduled Time")
## `stat_bin()` using `bins = 30`. Pick better value `binwidth`.
```



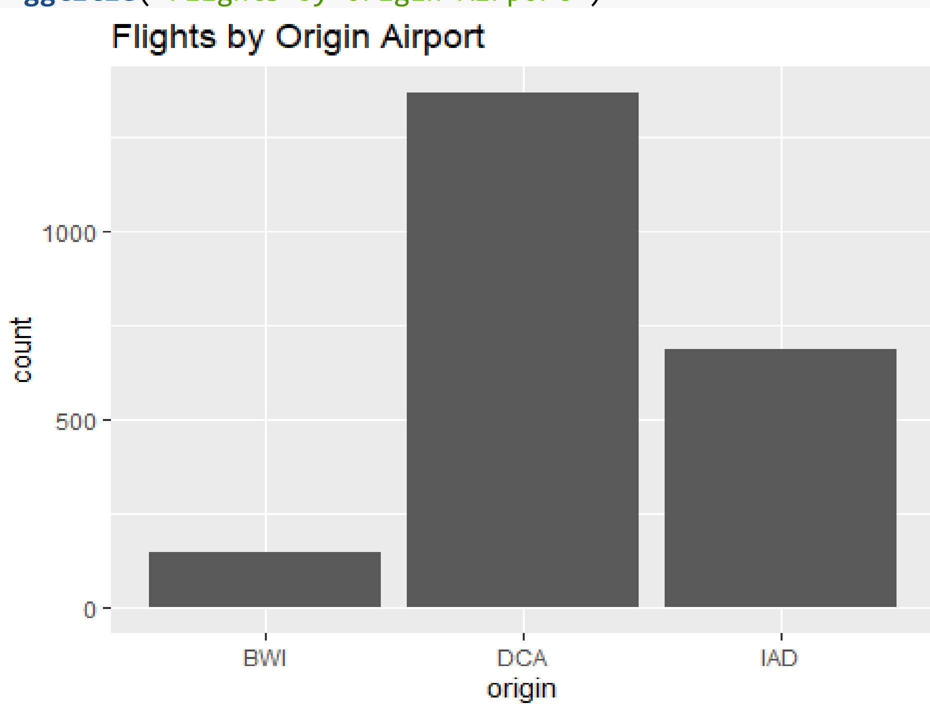

```
# Carrier
ggplot(flight_data, aes(x = carrier)) +
  geom_bar() +
  ggtitle("Flights by Carrier")
```



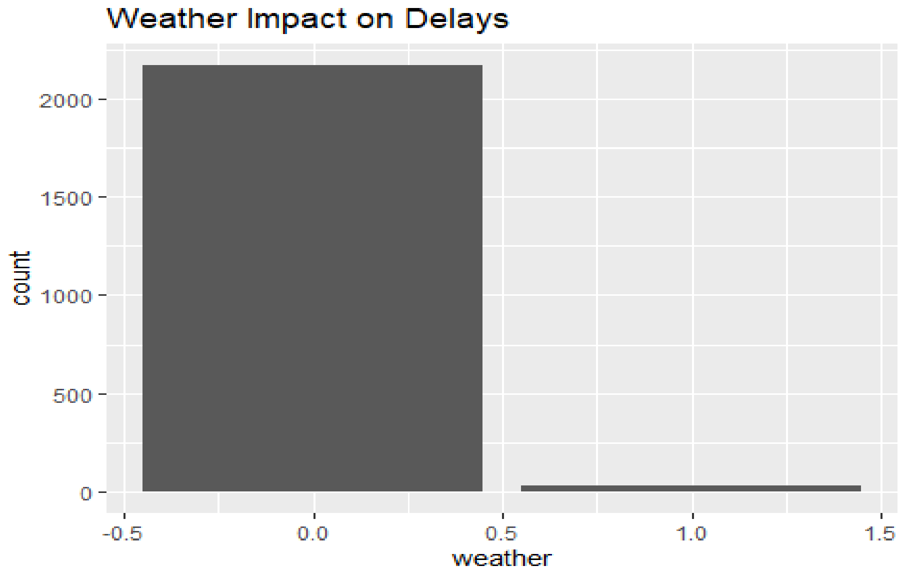
```r
# Destination
ggplot(flight_data, aes(x = dest)) +
  geom_bar() +
  ggtitle("Flights by Destination")
```



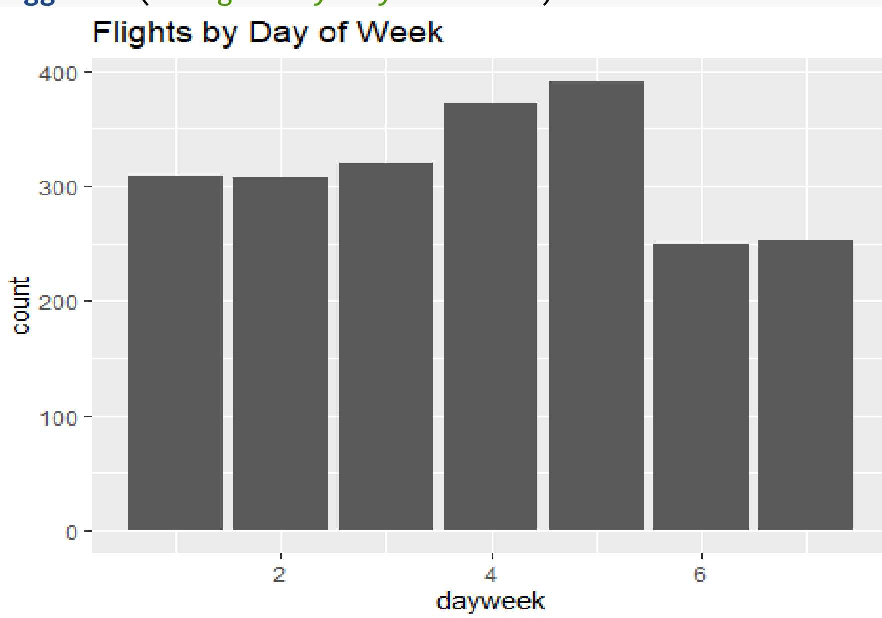

```r
# Origin
ggplot(flight_data, aes(x = origin)) +
  geom_bar() +
  ggtitle("Flights by Origin Airport")
```



```r
# Weather (0=On time, 1=Delayed)
ggplot(flight_data, aes(x = weather)) +
  geom_bar() +
  ggtitle("Weather Impact on Delays")
```



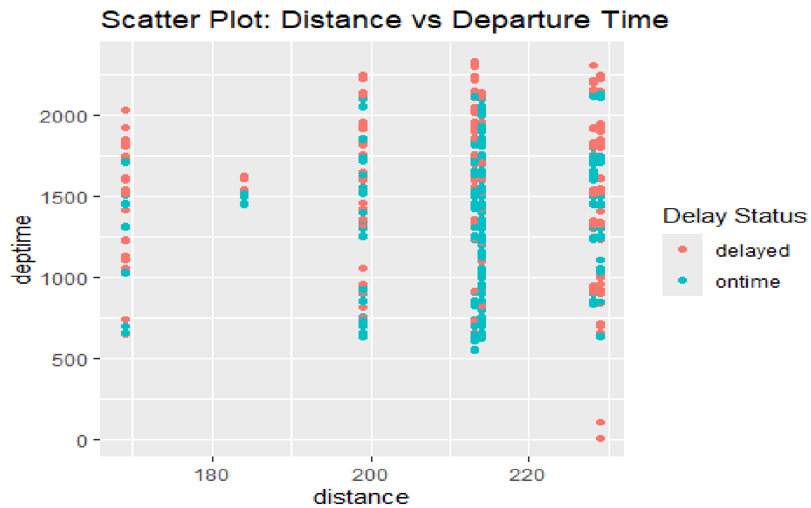

```r
# Day of Week
ggplot(flight_data, aes(x = dayweek)) +
  geom_bar() +
  ggtitle("Flights by Day of Week")
```



Explanation: In this step, we create histograms and bar charts to visualize flight patterns based on scheduled time, carrier, destination, origin, weather conditions, and day of the week.

Step 5: Scatter Plot (On Time vs Delayed Flights)

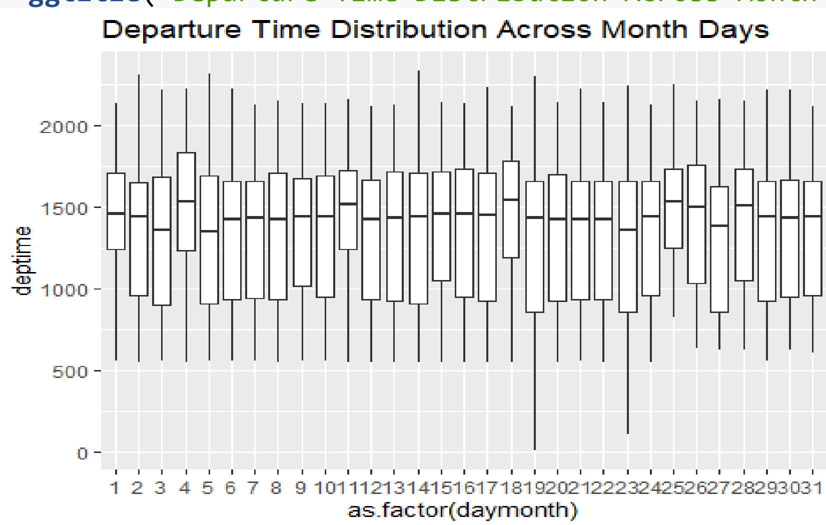
```
ggplot(flight_data, aes(x = distance, y = deptime, color = as.factor(delay)))
+
  geom_point() +
  ggtitle("Scatter Plot: Distance vs Departure Time") +
  labs(color = "Delay Status")
```



Explanation: In this step, we use a scatter plot to compare flights that are delayed versus on-time based on departure time and distance.

Step 6: Box Plot (Delay by Day of Month)

```
ggplot(flight_data, aes(x = as.factor(daymonth), y = deptime)) +
  geom_boxplot() +
  ggtitle("Departure Time Distribution Across Month Days")
```



Explanation: In this step, we create a box plot to analyze how flight delays vary across different days of the month.

Step 7: Define Hours of Departure

```
flight_data$dep_hour <- floor(flight_data$deptime / 100)

head(flight_data$dep_hour)
## [1] 14 16 12 17 10  8
```

Explanation: In this step, we extract the hour from the departure time to categorize flights based on morning, afternoon, or evening departures.

Step 8: Categorical Representation Using Table

```
table(flight_data$carrier, flight_data$delay)
##
##      delayed ontime
##   CO      26     68
##   DH     137    414
##   DL      47    341
##   MQ      80    215
##   OH       4     26
##   RU      94    314
##   UA       5     26
##   US      35    369
```

Explanation: In this step, we create a frequency table to observe how delays are distributed across different airline carriers.

Step 9: Redefine Delay Variable

```
flight_data$delay_status <- ifelse(flight_data$delay == 1, "Delayed", "On Tim
e")

table(flight_data$delay_status)
##
## On Time
##    2201
```

Explanation: In this step, we convert the delay column into a readable categorical format such as **Delayed** and **On Time** for easier analysis.

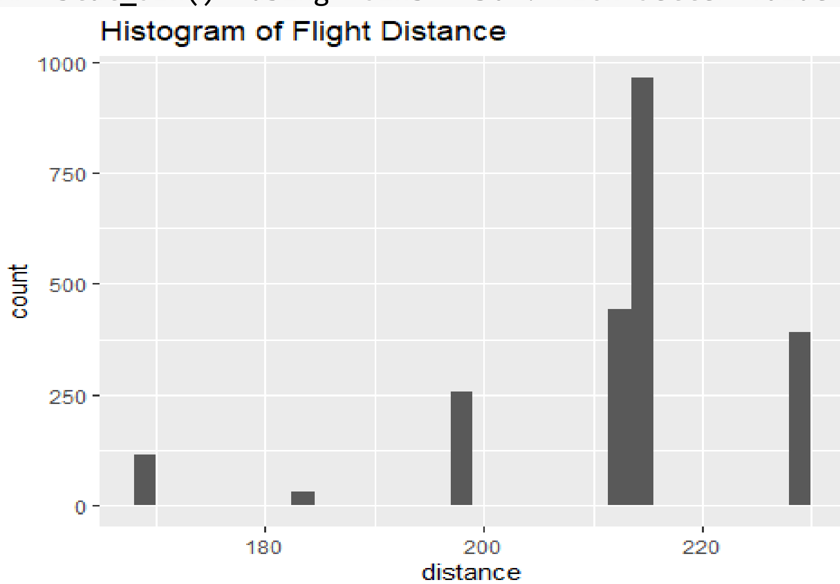
Step 10: Summary of Major Variables

```
summary(flight_data$distance)
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   169.0   213.0   214.0   211.9   214.0   229.0
summary(flight_data$schedtime)
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     600    1000    1455    1372    1710    2130
summary(flight_data$deptime)
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##      10    1004    1450    1369    1709    2330
```

Explanation: In this step, we summarize important numerical variables like distance, scheduled time, and departure time to identify major trends.

Step 11: Histograms of Major Variables

```
ggplot(flight_data, aes(x = distance)) +
  geom_histogram() +
  ggtitle("Histogram of Flight Distance")
## `stat_bin()` using `bins = 30`. Pick better value `binwidth`.
```

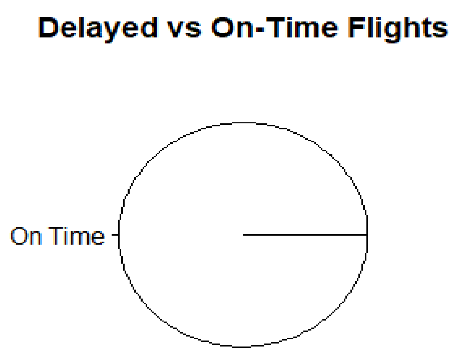


Explanation: In this step, we plot histograms for key numerical variables to better understand their distribution across flights.

Step 12: Pie Chart (Flights Delayed)

```r
delay_count <- table(flight_data$delay_status)

pie(delay_count,
    main = "Delayed vs On-Time Flights")
```



Explanation: In this step, we create a pie chart to visualize the proportion of flights that were delayed compared to flights that were on time.

✅ Final Conclusion (Short)

Overall, this project helps identify patterns in flight delays and highlights the impact of weather, carrier, and scheduling factors on airport performance.