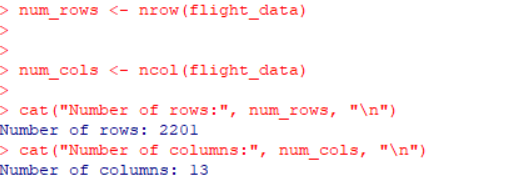
# Flight Delays Analysis

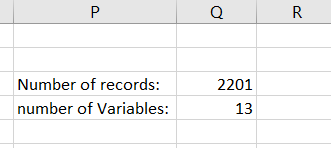
## Preliminary Analysis:

1. Check the **records and variables are the same** in the imported and original flightdelays.xlsx Excel file:

**R Dimension:**

****

**Excel Dimensions:**

****

1. Install or Load necessary **library packages**:

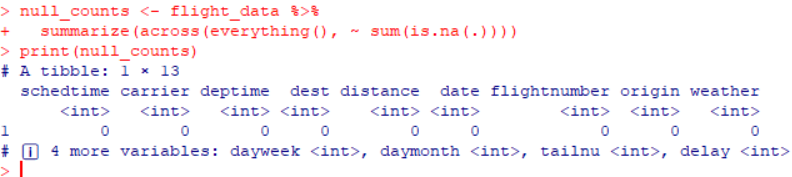
Library **readxl:** import data from Excel spreadsheets into R.

Library **dplyr**: for data manipulation and transformation.

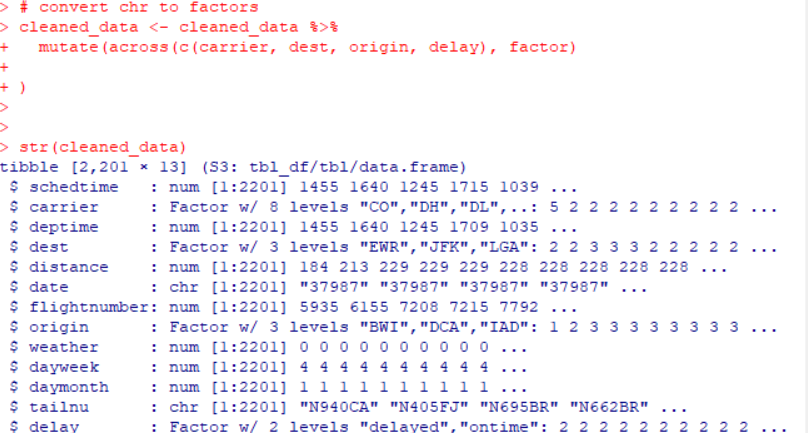
Library **lubridate:** to convert date strings in the "day/month/year"

Library **ggplot2**: for data visualization.

1. Find out the **null values:**

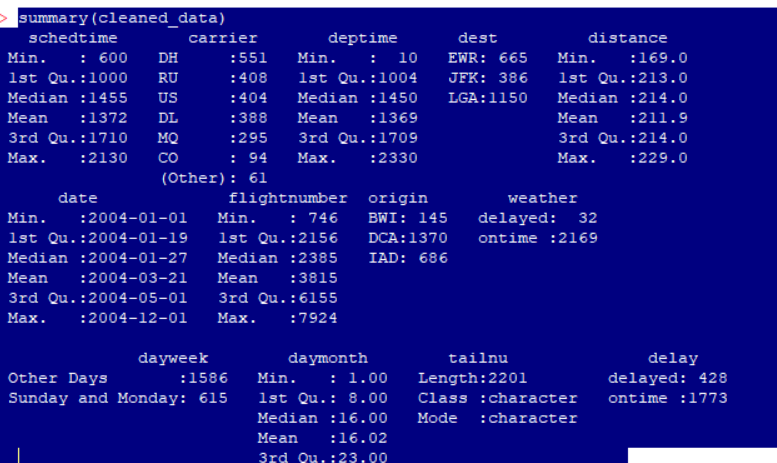


No null values detected

1. **Convert** Categorical variables to Factors enhancing statistical analysis: 
2. **Transforming** columns:

|  |
| --- |
| * **weather:** 0 – on time 1- delayed      * **dayweek**: 1 - Sunday and Monday 0 - for all other days |
|  |  |
|  |  |
|  |  |

1. Understand the **summary** of descriptive statistics:



**schedtime**: Flight times range from **600** (6:00 AM) to **2130** (9:30 PM). The mean is **1372** (1:12 PM).

 **carrier**: The top airline, DH, has **551** flights (25%), while the least popular has only a few.

 **deptime**: Most flights depart between **10:00 AM** and **11:00 PM**.

 **dest**: EWR (665 flights, **30%**), JFK (386 flights, **18%**), and LGA (1150 flights, **52%**) are the main destinations.

 **distance**: Distances range from **169** to **229** miles, with a mean of **211.9** miles.

 **date**: Data covers flights from **January 1, 2004** to **December 1, 2004**.

 **flightnumber**: There are **2201** total flights, indicating a busy schedule.

 **origin**: BWI has **145** flights (7%), DCA has **1370** flights (62%), and IAD has **686** flights (31%).

 **weather**: Out of **2201** flights, **1773** (80%) are on time, and **428** (20%) are delayed.

 **dayweek**: **1586** flights (72%) are on other days, while **615** flights (28%) are on Sunday and Monday.

 **daymonth**: Days range from **1** to **31** in each month, with a median of **16**.

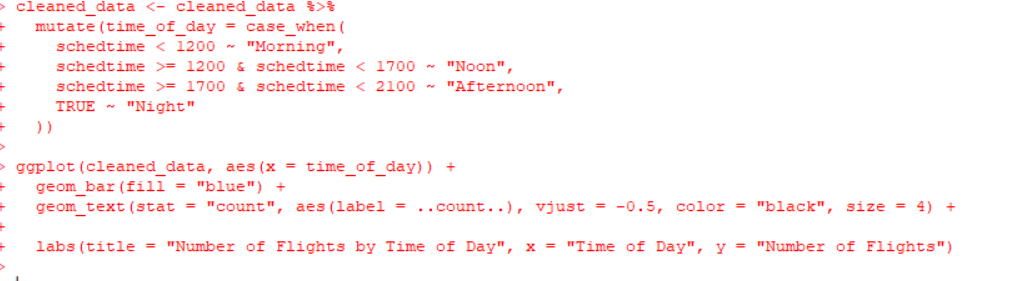
 **tailnu**: The tail numbers are unique identifiers for planes, which can help track performance.

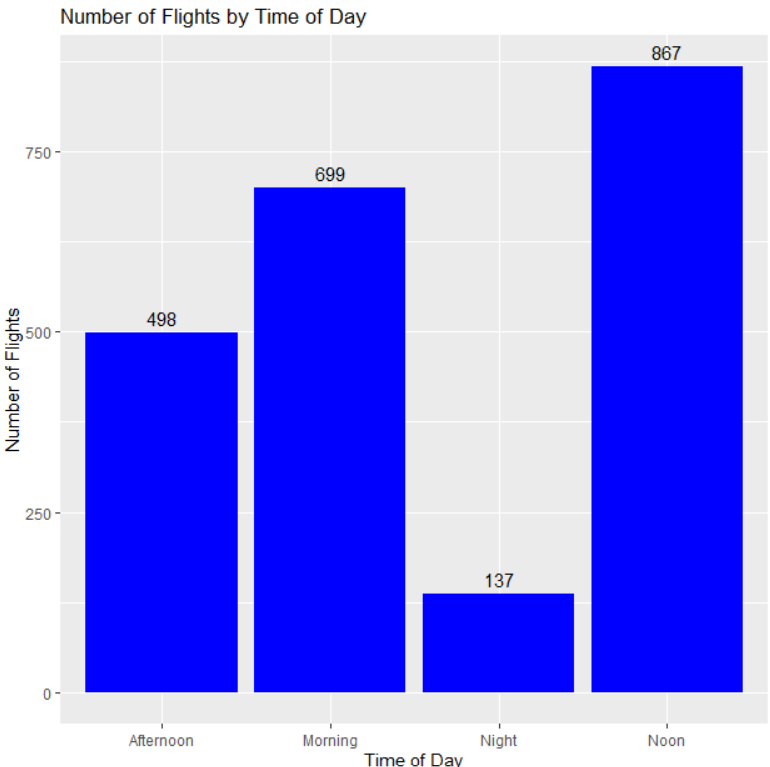
 **delay**: Delayed flights total **428** (20%), while on-time flights total **1773** (80%).

## EDA Analysis

1. Plot the **histograms** to understand the relationships **between scheduled time, carrier, destination, origin, weather, and day of the week**

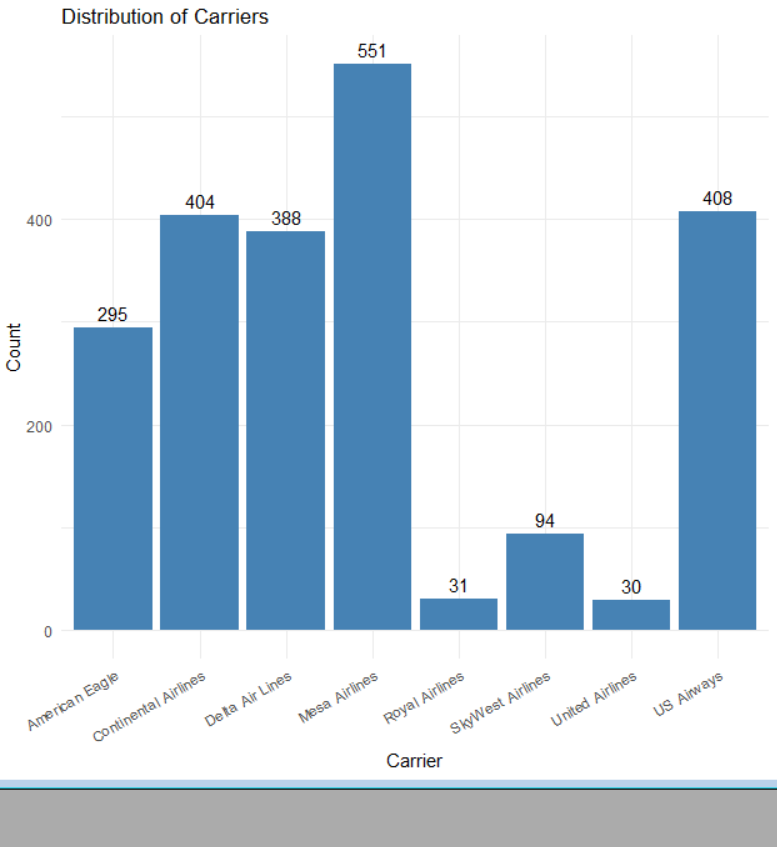
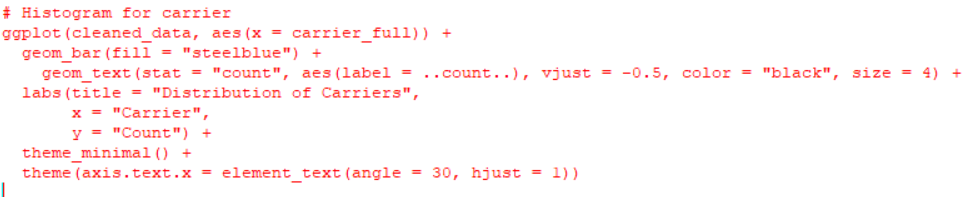
* **Flight per day**

****



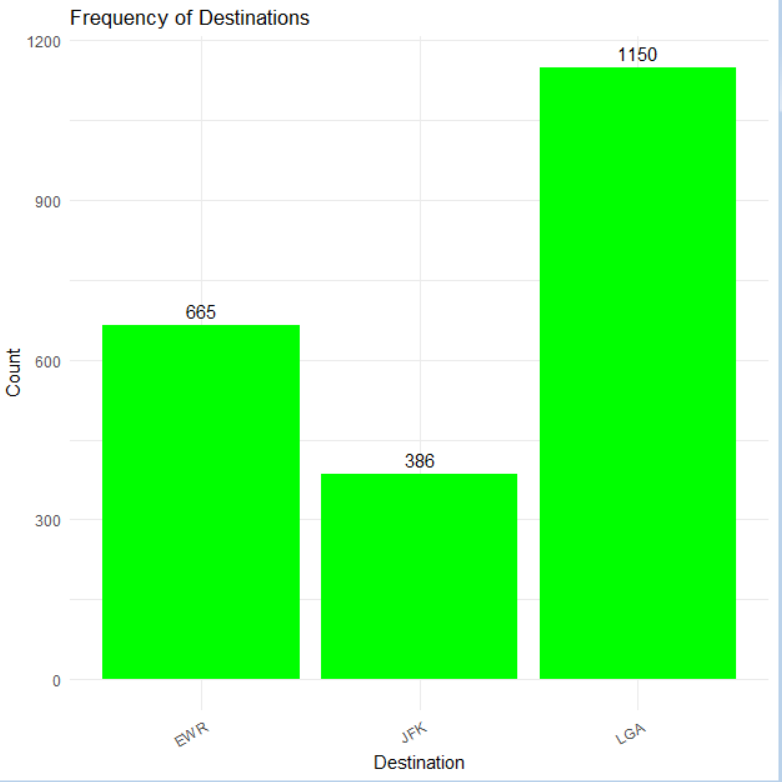
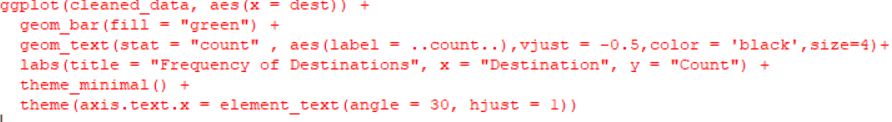
The highest number of flights are during the morning and noon, while the night period had the least flights.

* Carrier



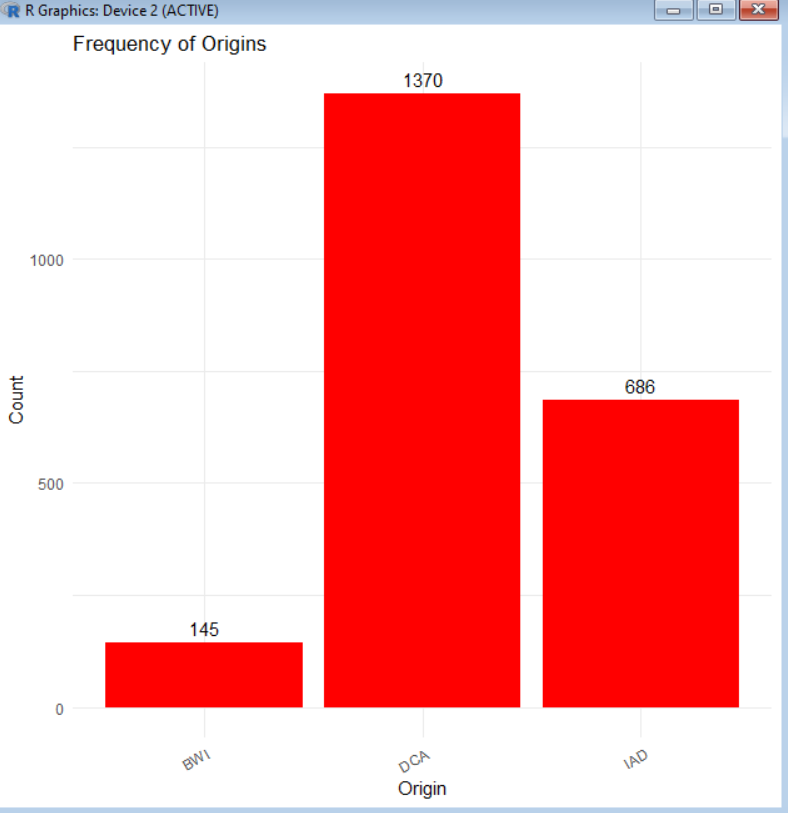
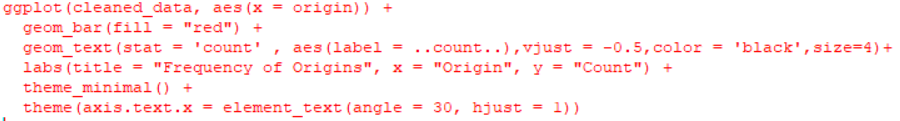
Mesa Airlines had the highest count, followed by Delta Air Lines and Continental Airlines. SkyWest and United Airlines had the lowest counts.

* Destination Histogram



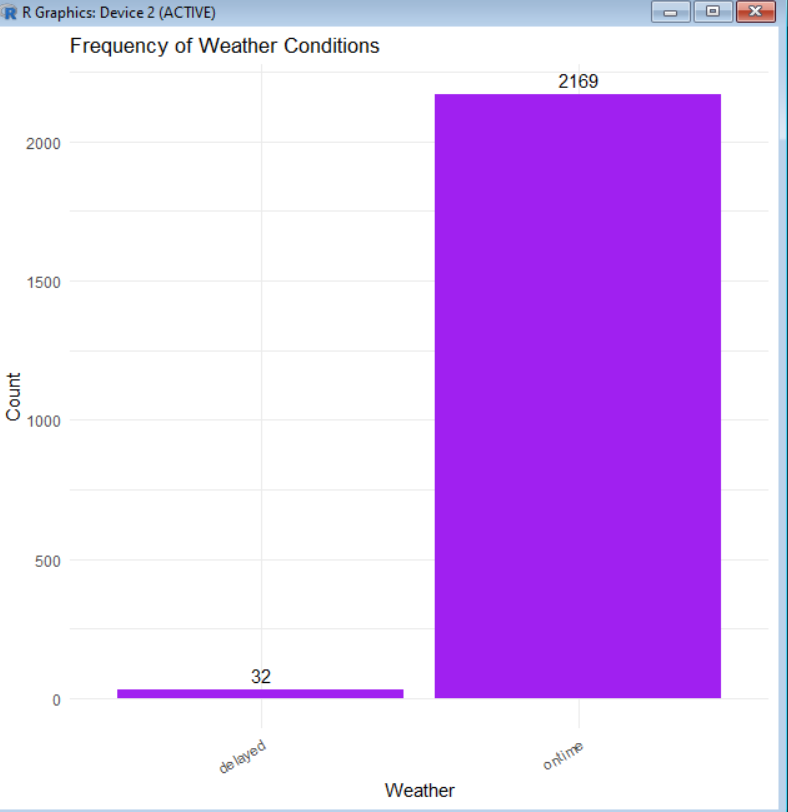
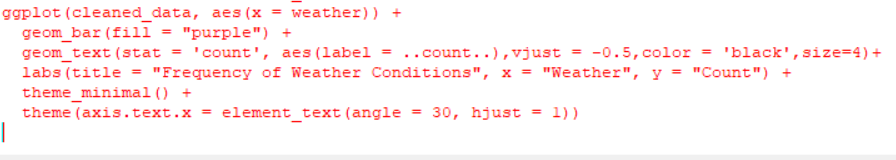
**LGA** is the most designated destination followed by **EWE** and **JFK**.

* Origin Histogram



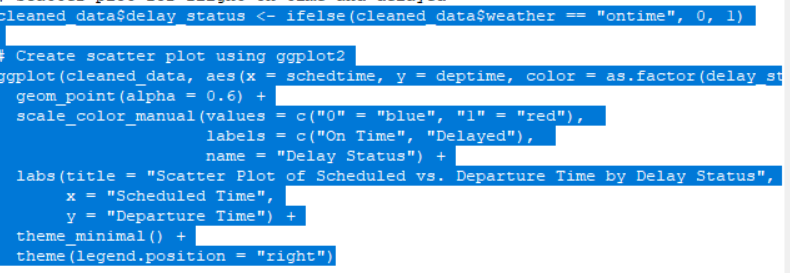
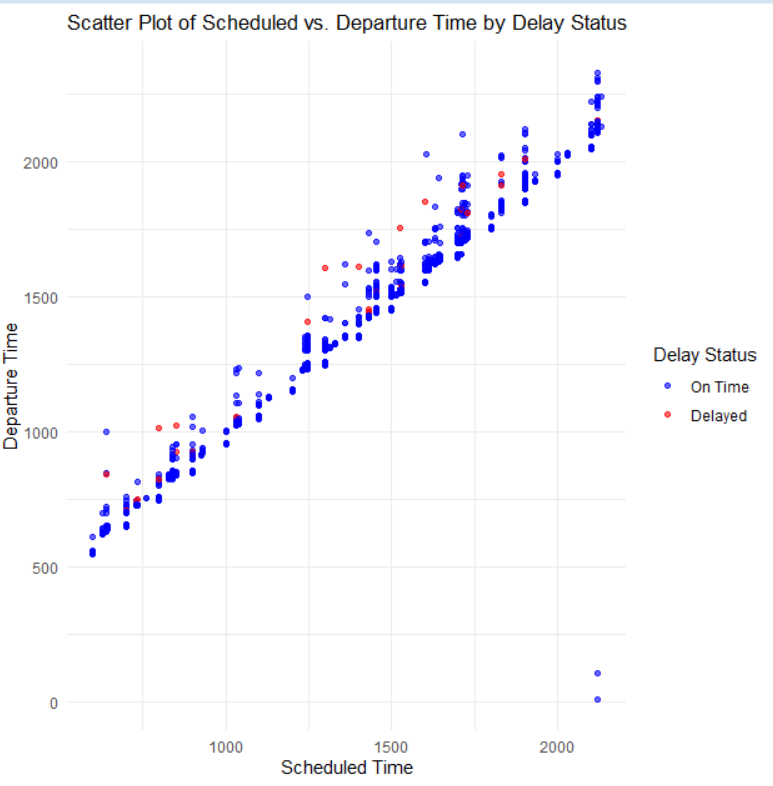
**Most** of the airplanes are originated from **DCA and BWI** has the **least** airplanes going from.

* Weather Histogram

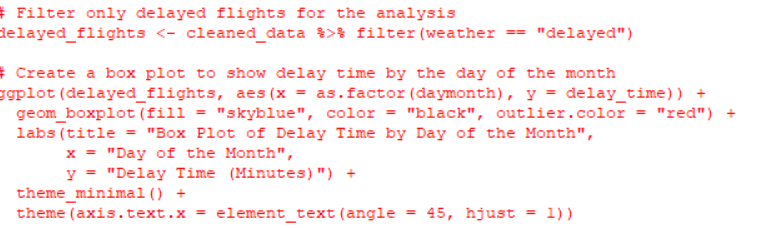
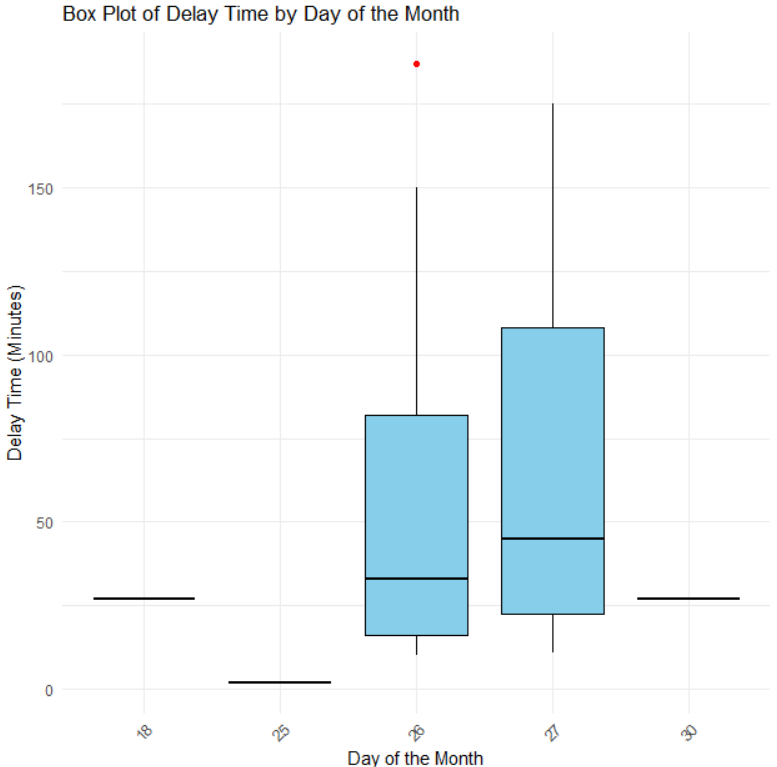


According to Weather Histogram **weather** has **small effect on airplanes delay**.

1. Plot the **scatter plot** for **flights on time** and **delayed**



1. Plot the **box** **plot** to understand how many days in a month flights are delayed by what time



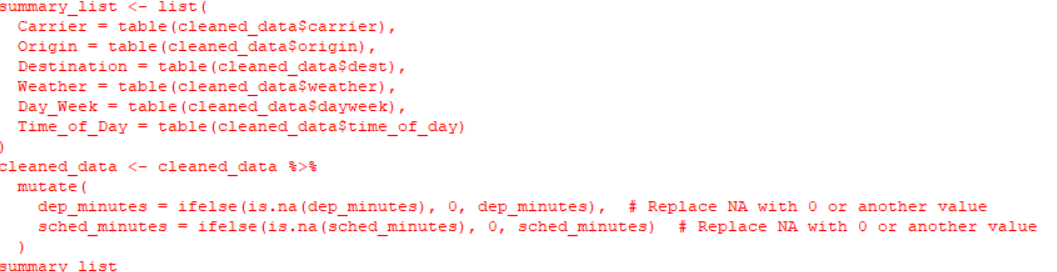
Some days show larger variation in delays **day 19 and 21**, with wider boxes and longer whiskers, while other **days 18 and 25** have small or no delay variation.

1. hours of **departure**

head(cleaned\_data$departure\_hour)

[1] 12 21 14 9 16 12

1. **Categorical** representation of data using a table



$Carrier

CO DH DL MQ OH RU UA US

24 215 139 100 10 110 6 34

$Origin

BWI DCA IAD

55 344 239

$Destination

EWR JFK LGA

190 152 296

$Weather

delayed ontime

32 606

$Day\_Week

Other Days Sunday and Monday

440 198

$Time\_of\_Day

Night Morning Afternoon Evening

1 148 360 129

> summary\_list

$Carrier

CO DH DL MQ OH RU UA US

24 215 139 100 10 110 6 34

$Origin

BWI DCA IAD

55 344 239

$Destination

EWR JFK LGA

190 152 296

$Weather

delayed ontime

32 606

$Day\_Week

Other Days Sunday and Monday

440 198

$Time\_of\_Day

Night Morning Afternoon Evening

1 148 360 129

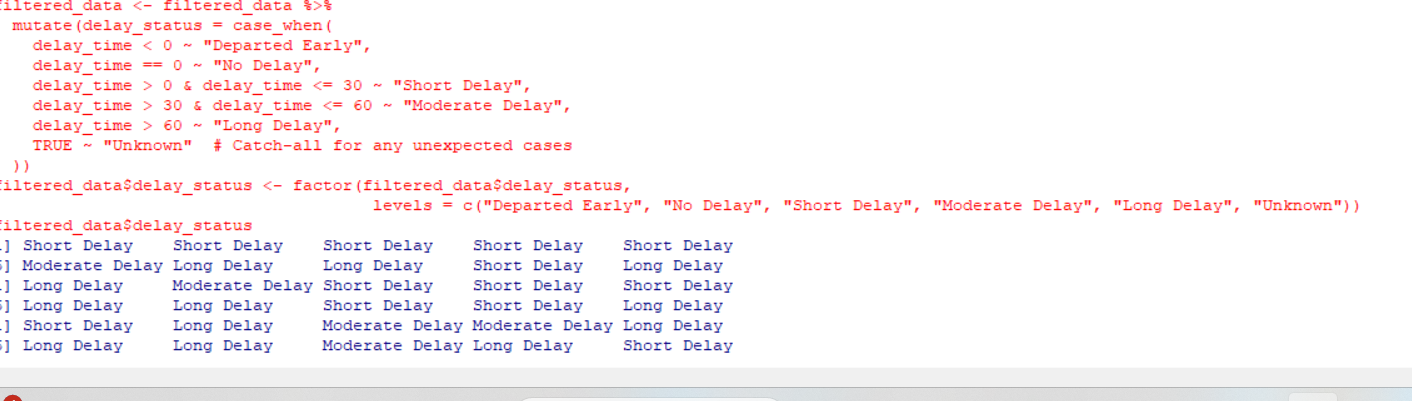
 **Carrier Performance**: Delta Airlines has many flights, so it is important to look at how often their flights are delayed compared to other airlines.

 **Airport Operations**: DCA has a lot of flights, which could cause delays. Checking the delays at this airport might help understand any problems.

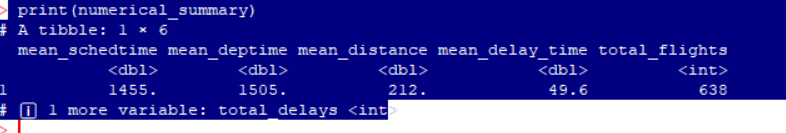
 **Weather Impact**: There are not many delayed flights because of weather, so other reasons might be causing the delays.

 **Traffic Patterns**: The day of the week and time of day may affect delays. Afternoons have more flights, which could lead to more delays.

1. Redefine delay variables



1. **summary** of major variables



 **Mean Scheduled Time**: The average scheduled time for flights is 2:55 PM. Many flights are scheduled during busy afternoon hours, which could lead to delays.

 **Mean Departure Time**: The average departure time is 3:05 PM. Flights are mostly leaving on time, which is a good sign for operations.

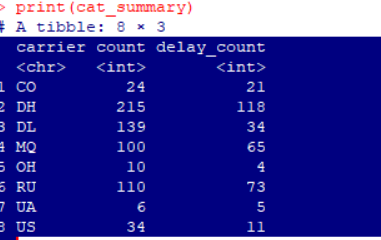
 **Mean Distance**: The average flight distance is 212 miles, indicating that most flights are short, likely regional or domestic. Short flights may have different delay patterns than long flights.

 **Mean Delay Time**: The average delay is almost 50 minutes. This is significant and suggests that while flights are scheduled on time, many experience long delays.

 **Total Flights**: There are 638 flights in the dataset. This is a good sample size for analysis and can lead to reliable results.

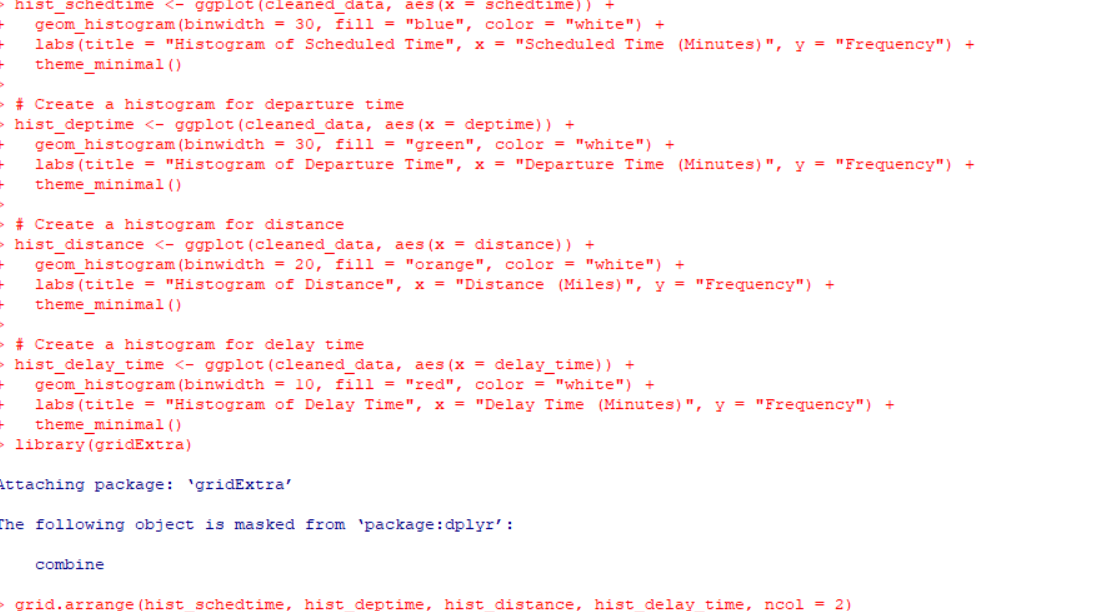
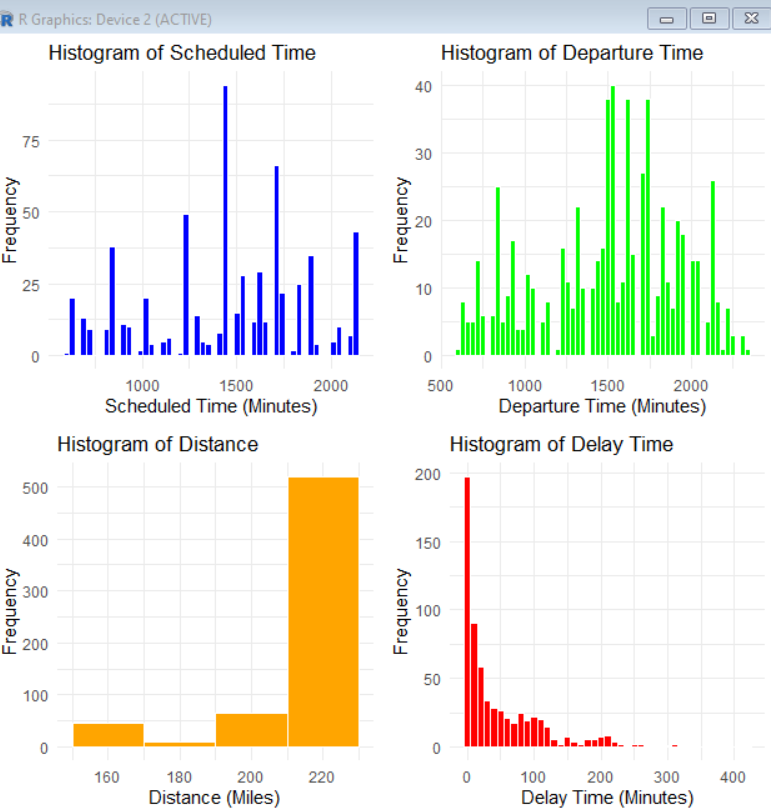
 **Total Delays**: The total number of delayed flights is important to know. With an average delay of nearly 50 minutes, many flights are not on time, which could indicate problems in operations.

**Categorical summary**

****

 **High Delays**: CO and DH have many delays and should improve.

 **Better Performance**: DL has fewer delays and is more reliable.

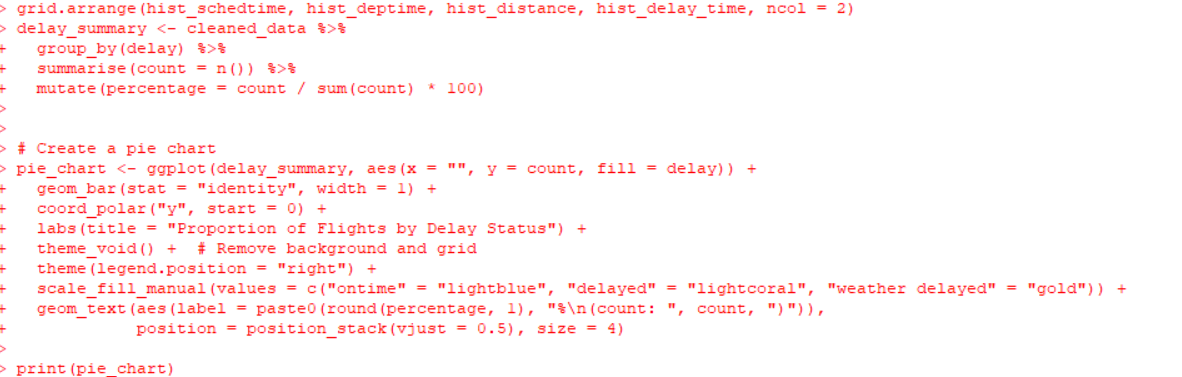
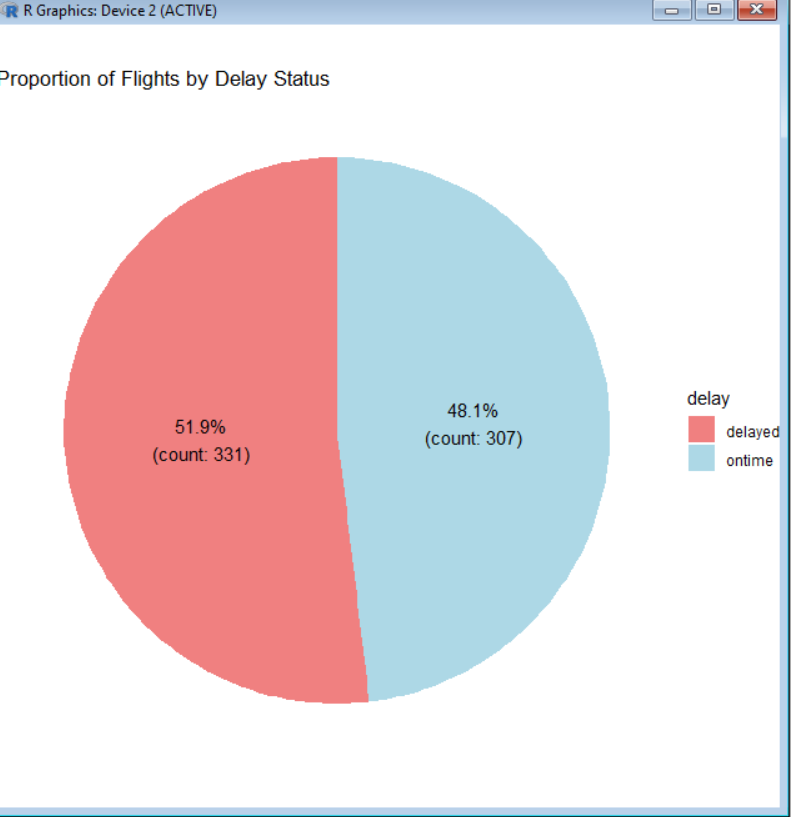
1. **Histograms from Major Variables. **

**Peak Times and Delays**: Flights scheduled between 1000-1500 minutes (morning to early afternoon) are during busy periods. Delays in these times can increase congestion and make operations harder to manage.

**Short-Haul Flights:** Many flights cover short distances (around 220 miles). These shorter flights are more likely to face delays. Improving turnaround times, weather planning, and ground operations can help reduce these delays.

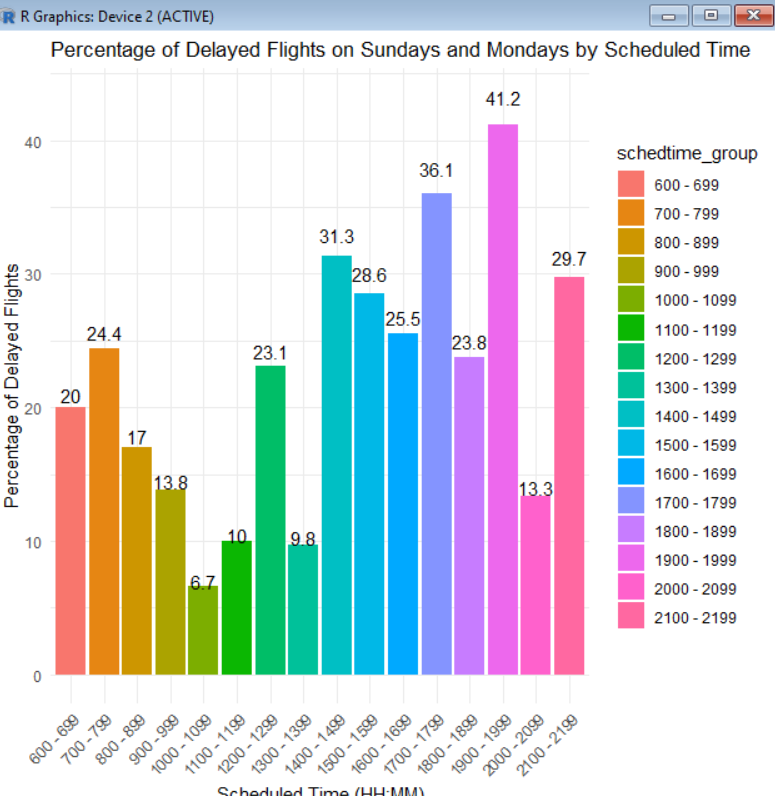
**Outlier Delays:** Most delays are small, but some flights have delays over 100 minutes. These long delays may be due to weather or operational issues, and need focused solutions.

1. Pie chart for delayed flights

****

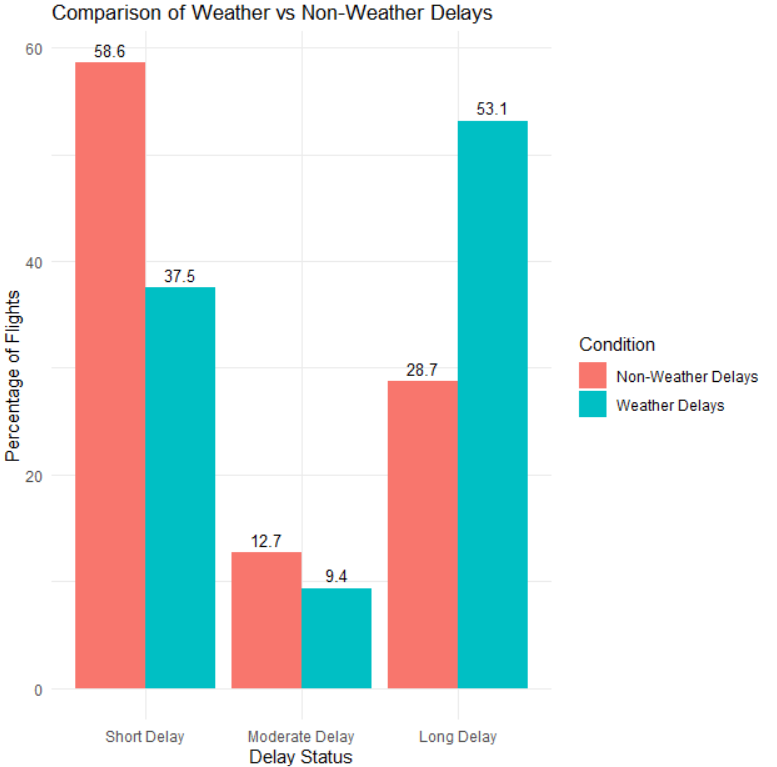
## Further Analysis

1. **Weekends Delays** major cause



Many delays happen in the late afternoon and evening, with over **30%** of flights delayed. This suggests busy times due to lots of passengers and potential operational issues.

1. Comparison of weather vs non weather delays

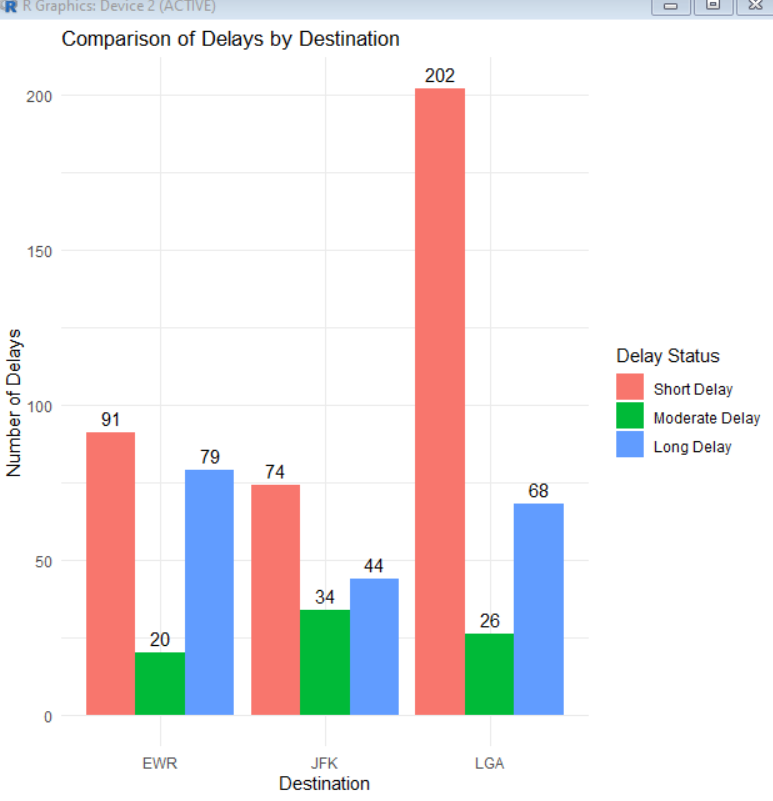


* **Weather-Related Delays:**
  + **Long Delays (more than 1 hour):** 53% of weather delays last over an hour. This shows that when weather causes a delay, it’s often serious.
  + **Short Delays (up to 30 minutes):** Only 37.5% of weather delays are short, meaning most weather delays are longer.
* **Non-Weather-Related Delays:**
  + **Short Delays (up to 30 minutes):** Almost 59% of non-weather delays are short, which means most of these delays aren’t too long.
  + **Long Delays (more than 1 hour):** Around 29% of non-weather delays are long, much lower than for weather-related delays.

### Key Insight:

**Weather tends to cause longer delays**, while **most non-weather delays are short** and under 30 minutes.

1. **Delay** caused by **destinations**



1. **LGA (LaGuardia Airport)**
   * **Total Flights**: 1,150
   * **Total Delays**: 296 (202 Short, 26 Moderate, 68 Long)
   * LGA has the most flights and the highest number of short delays (202).
2. **EWR (Newark Airport)**
   * **Total Delays**: 190 (91 Short, 20 Moderate, 79 Long)
   * EWR has many short delays (91) and a lot of long delays (79).
3. **JFK (Kennedy Airport)**
   * **Total Delays**: 152 (74 Short, 34 Moderate, 44 Long)
   * JFK has some short delays (74) and moderate delays (34).

### Key Points

* **LGA** has the most flights and the most short delays.
* **EWR** and **JFK** have many long delays.

### Conclusion

LGA has a lot of short delays because it has many flights. EWR and JFK have many long delays. Improving these delays can help passengers.