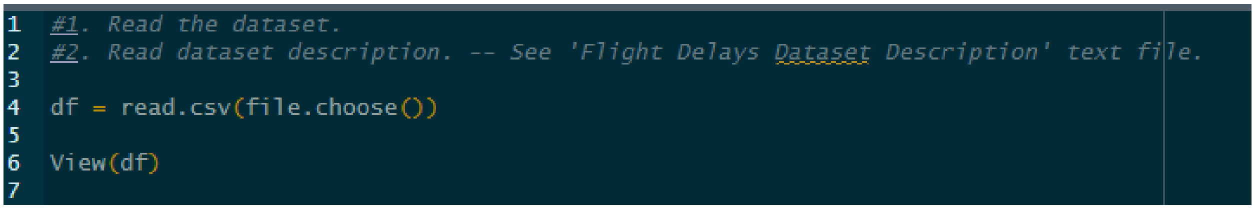
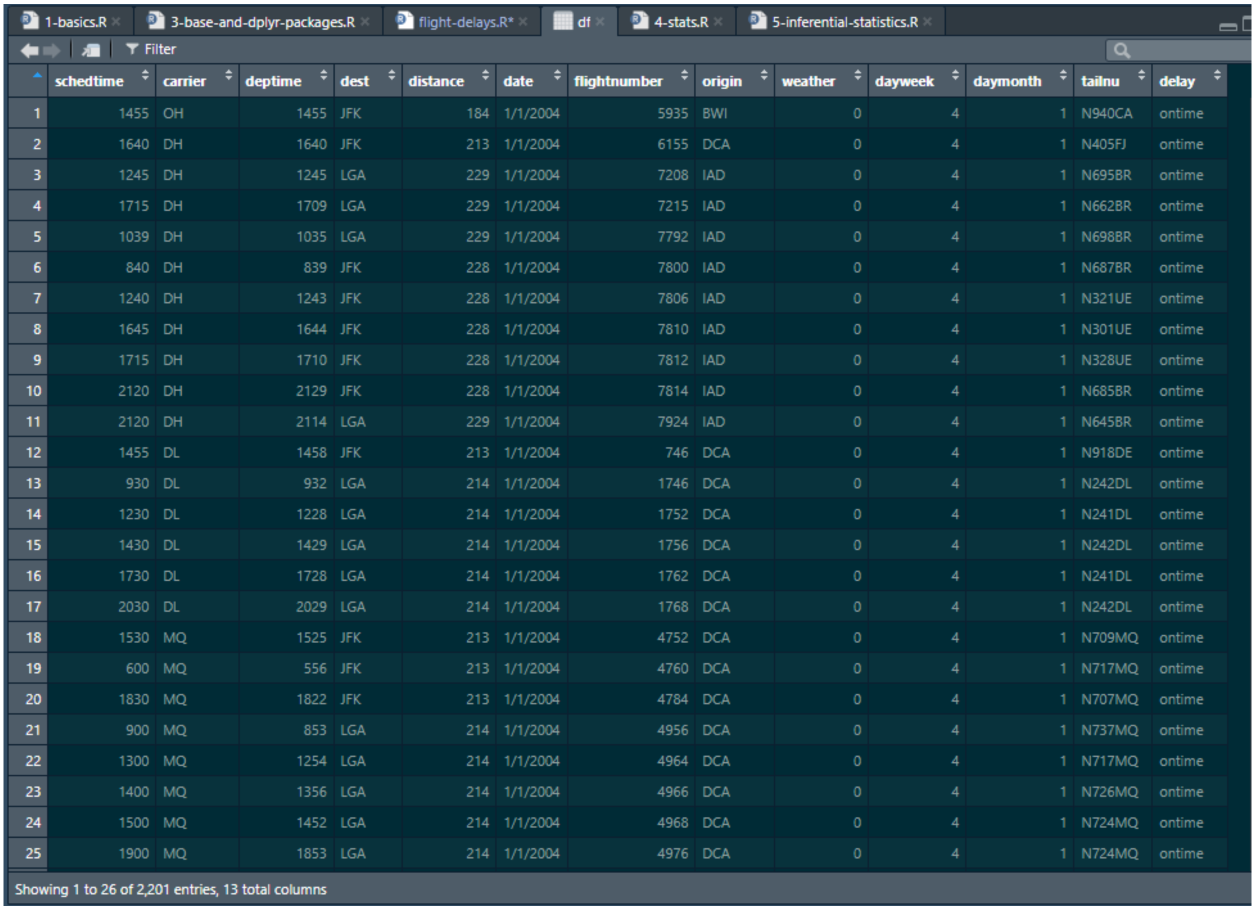
Flight Delays Analysis

1. Read the dataset
2. Read the dataset description

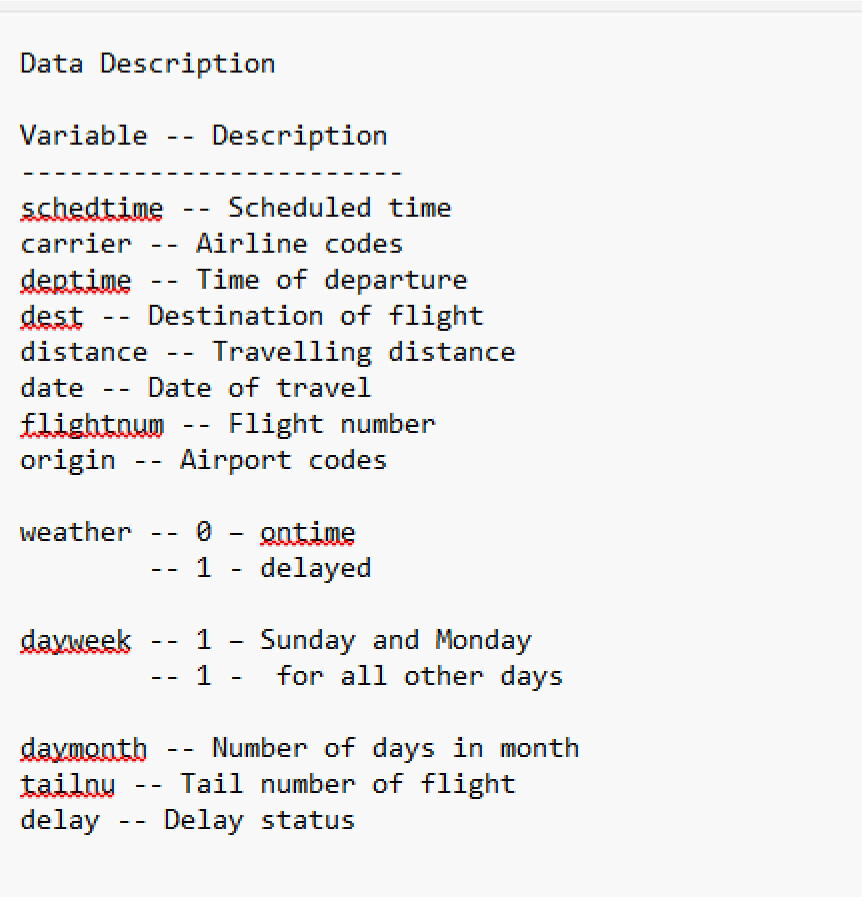


```
#1. Read the dataset.
#2. Read dataset description. -- See 'Flight Delays Dataset Description' text file.

df = read.csv(file.choose())

View(df)

```



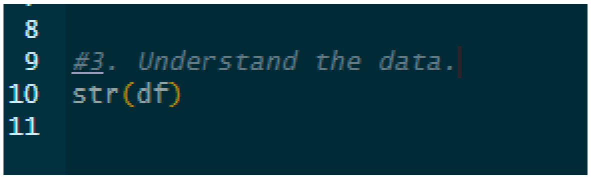
	schedtime	carrier	deptime	dest	distance	date	flightnumber	origin	weather	dayweek	daymonth	tailnu	delay
1	1455	OH	1455	JFK	184	1/1/2004	5935	BWI	0	4	1	N940CA	ontime
2	1640	DH	1640	JFK	213	1/1/2004	6155	DCA	0	4	1	N405FJ	ontime
3	1245	DH	1245	LGA	229	1/1/2004	7208	IAD	0	4	1	N695BR	ontime
4	1715	DH	1709	LGA	229	1/1/2004	7215	IAD	0	4	1	N662BR	ontime
5	1039	DH	1035	LGA	229	1/1/2004	7792	IAD	0	4	1	N698BR	ontime
6	840	DH	839	JFK	228	1/1/2004	7800	IAD	0	4	1	N687BR	ontime
7	1240	DH	1243	JFK	228	1/1/2004	7806	IAD	0	4	1	N321UE	ontime
8	1645	DH	1644	JFK	228	1/1/2004	7810	IAD	0	4	1	N301UE	ontime
9	1715	DH	1710	JFK	228	1/1/2004	7812	IAD	0	4	1	N328UE	ontime
10	2120	DH	2129	JFK	228	1/1/2004	7814	IAD	0	4	1	N685BR	ontime
11	2120	DH	2114	LGA	229	1/1/2004	7924	IAD	0	4	1	N645BR	ontime
12	1455	DL	1458	JFK	213	1/1/2004	746	DCA	0	4	1	N918DE	ontime
13	930	DL	932	LGA	214	1/1/2004	1746	DCA	0	4	1	N242DL	ontime
14	1230	DL	1228	LGA	214	1/1/2004	1752	DCA	0	4	1	N241DL	ontime
15	1430	DL	1429	LGA	214	1/1/2004	1756	DCA	0	4	1	N242DL	ontime
16	1730	DL	1728	LGA	214	1/1/2004	1762	DCA	0	4	1	N241DL	ontime
17	2030	DL	2029	LGA	214	1/1/2004	1768	DCA	0	4	1	N242DL	ontime
18	1530	MQ	1525	JFK	213	1/1/2004	4752	DCA	0	4	1	N709MQ	ontime
19	600	MQ	556	JFK	213	1/1/2004	4760	DCA	0	4	1	N717MQ	ontime
20	1830	MQ	1822	JFK	213	1/1/2004	4784	DCA	0	4	1	N707MQ	ontime
21	900	MQ	853	LGA	214	1/1/2004	4956	DCA	0	4	1	N737MQ	ontime
22	1300	MQ	1254	LGA	214	1/1/2004	4964	DCA	0	4	1	N717MQ	ontime
23	1400	MQ	1356	LGA	214	1/1/2004	4966	DCA	0	4	1	N726MQ	ontime
24	1500	MQ	1452	LGA	214	1/1/2004	4968	DCA	0	4	1	N724MQ	ontime
25	1900	MQ	1853	LGA	214	1/1/2004	4976	DCA	0	4	1	N724MQ	ontime

Showing 1 to 26 of 2,201 entries, 13 total columns

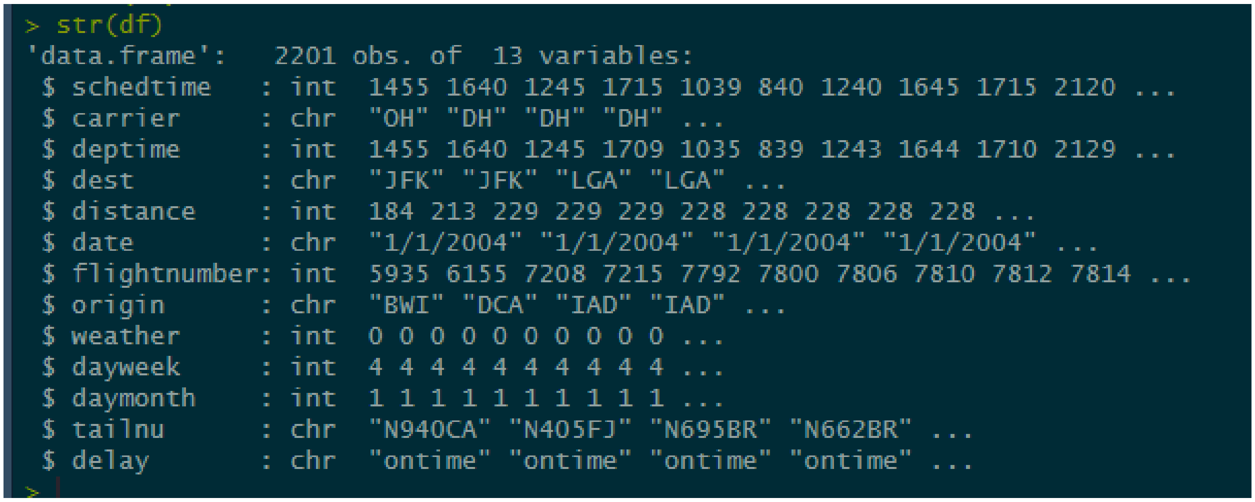


```
Data Description

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 -- 0 – ontime
        -- 1 - delayed

dayweek -- 1 – Sunday and Monday
        -- 1 -  for all other days

daymonth -- Number of days in month
tailnu -- Tail number of flight
delay -- Delay status
```

3. Understand the data

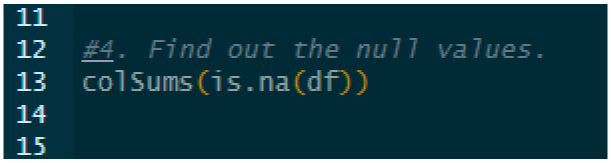


```
#3. Understand the data.
str(df)
```

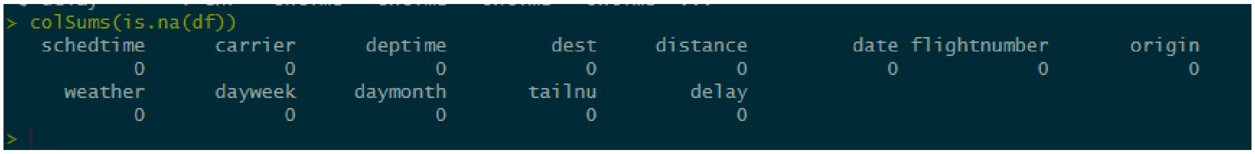


```
> str(df)
'data.frame':	2201 obs. of  13 variables:
 $ schedtime   : int  1455 1640 1245 1715 1039 840 1240 1645 1715 2120 ...
 $ carrier     : chr  "OH" "DH" "DH" "DH" ...
 $ deptime     : int  1455 1640 1245 1709 1035 839 1243 1644 1710 2129 ...
 $ dest        : chr  "JFK" "JFK" "LGA" "LGA" ...
 $ distance    : int  184 213 229 229 229 228 228 228 228 228 ...
 $ date        : chr  "1/1/2004" "1/1/2004" "1/1/2004" "1/1/2004" ...
 $ flightnumber: int  5935 6155 7208 7215 7792 7800 7806 7810 7812 7814 ...
 $ origin      : chr  "BWI" "DCA" "IAD" "IAD" ...
 $ weather     : int  0 0 0 0 0 0 0 0 0 0 ...
 $ dayweek     : int  4 4 4 4 4 4 4 4 4 4 ...
 $ daymonth    : int  1 1 1 1 1 1 1 1 1 1 ...
 $ tailnu      : chr  "N940CA" "N405FJ" "N695BR" "N662BR" ...
 $ delay       : chr  "ontime" "ontime" "ontime" "ontime" ...
>
```

4. Find out the null values
    a. There were no null values found.

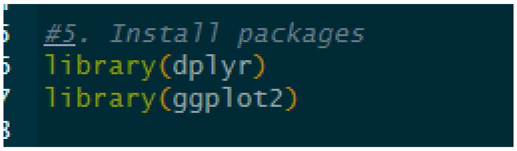


```
#4. Find out the null values.
colSums(is.na(df))
```



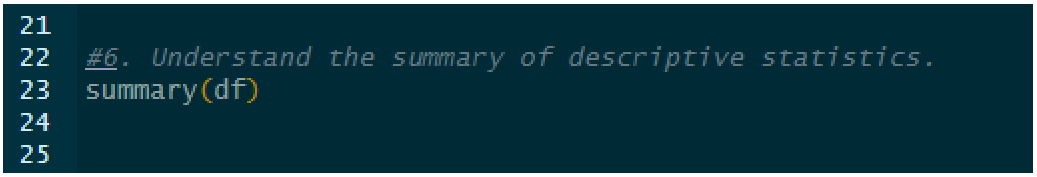
```
> colSums(is.na(df))
   schedtime      carrier      deptime         dest     distance         date flightnumber       origin
           0            0            0            0            0            0            0            0
     weather      dayweek     daymonth       tailnu        delay
           0            0            0            0            0
>
```

5. Install the required packages

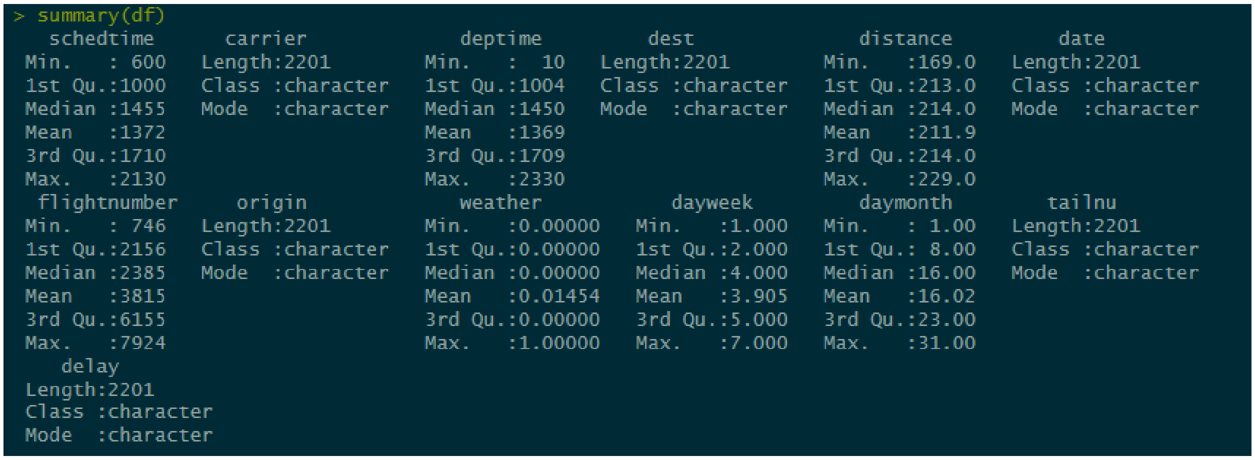


```
#5. Install packages
library(dplyr)
library(ggplot2)
```

6. Understand the summary of descriptive statistics

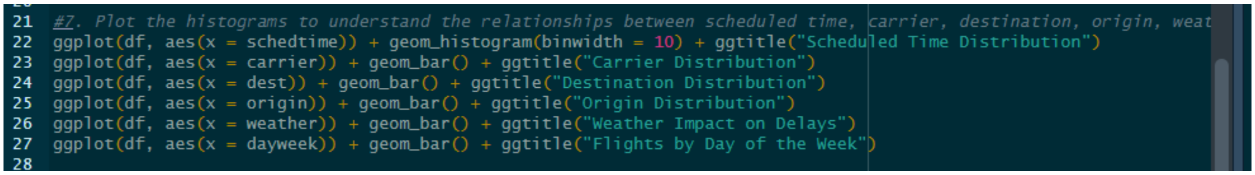


```
#6. Understand the summary of descriptive statistics.
summary(df)
```

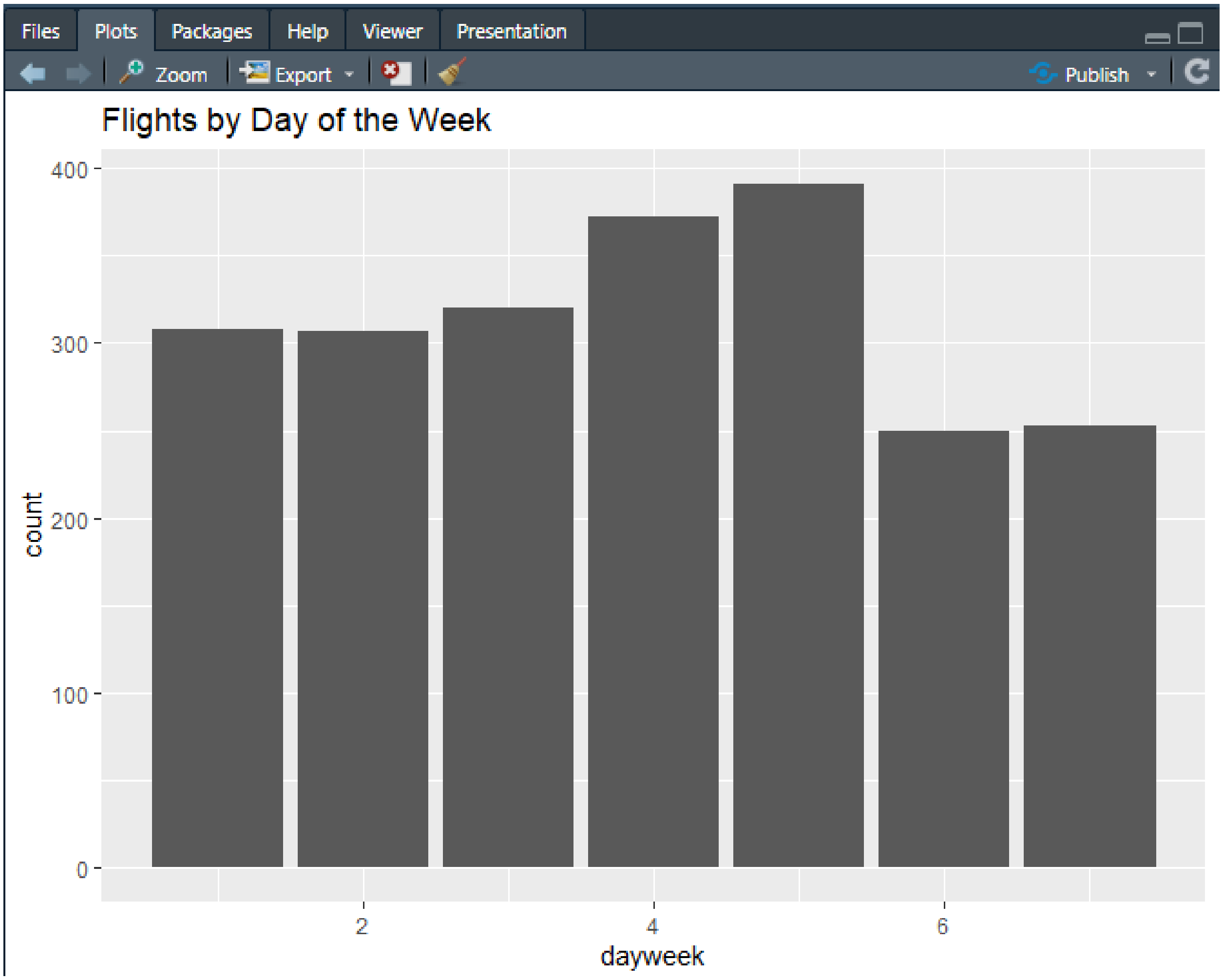


```
> summary(df)
   schedtime      carrier             deptime         dest              distance          date          
 Min.   : 600   Length:2201        Min.   :  10   Length:2201        Min.   :169.0   Length:2201       
 1st Qu.:1000   Class :character   1st Qu.:1004   Class :character   1st Qu.:213.0   Class :character  
 Median :1455   Mode  :character   Median :1450   Mode  :character   Median :214.0   Mode  :character  
 Mean   :1372                      Mean   :1369                      Mean   :211.9                     
 3rd Qu.:1710                      3rd Qu.:1709                      3rd Qu.:214.0                     
 Max.   :2130                      Max.   :2330                      Max.   :229.0                     
  flightnumber     origin             weather           dayweek         daymonth        tailnu          
 Min.   : 746   Length:2201        Min.   :0.00000   Min.   :1.000   Min.   : 1.00   Length:2201       
 1st Qu.:2156   Class :character   1st Qu.:0.00000   1st Qu.:2.000   1st Qu.: 8.00   Class :character  
 Median :2385   Mode  :character   Median :0.00000   Median :4.000   Median :16.00   Mode  :character  
 Mean   :3815                      Mean   :0.01454   Mean   :3.905   Mean   :16.02                     
 3rd Qu.:6155                      3rd Qu.:0.00000   3rd Qu.:5.000   3rd Qu.:23.00                     
 Max.   :7924                      Max.   :1.00000   Max.   :7.000   Max.   :31.00                     
     delay          
 Length:2201       
 Class :character  
 Mode  :character  
```

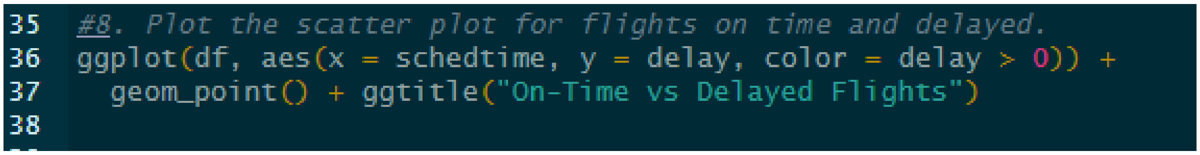
7. Plot the histograms to understand the relationships between scheduled time, carrier, destination, origin, weather, and day of the week



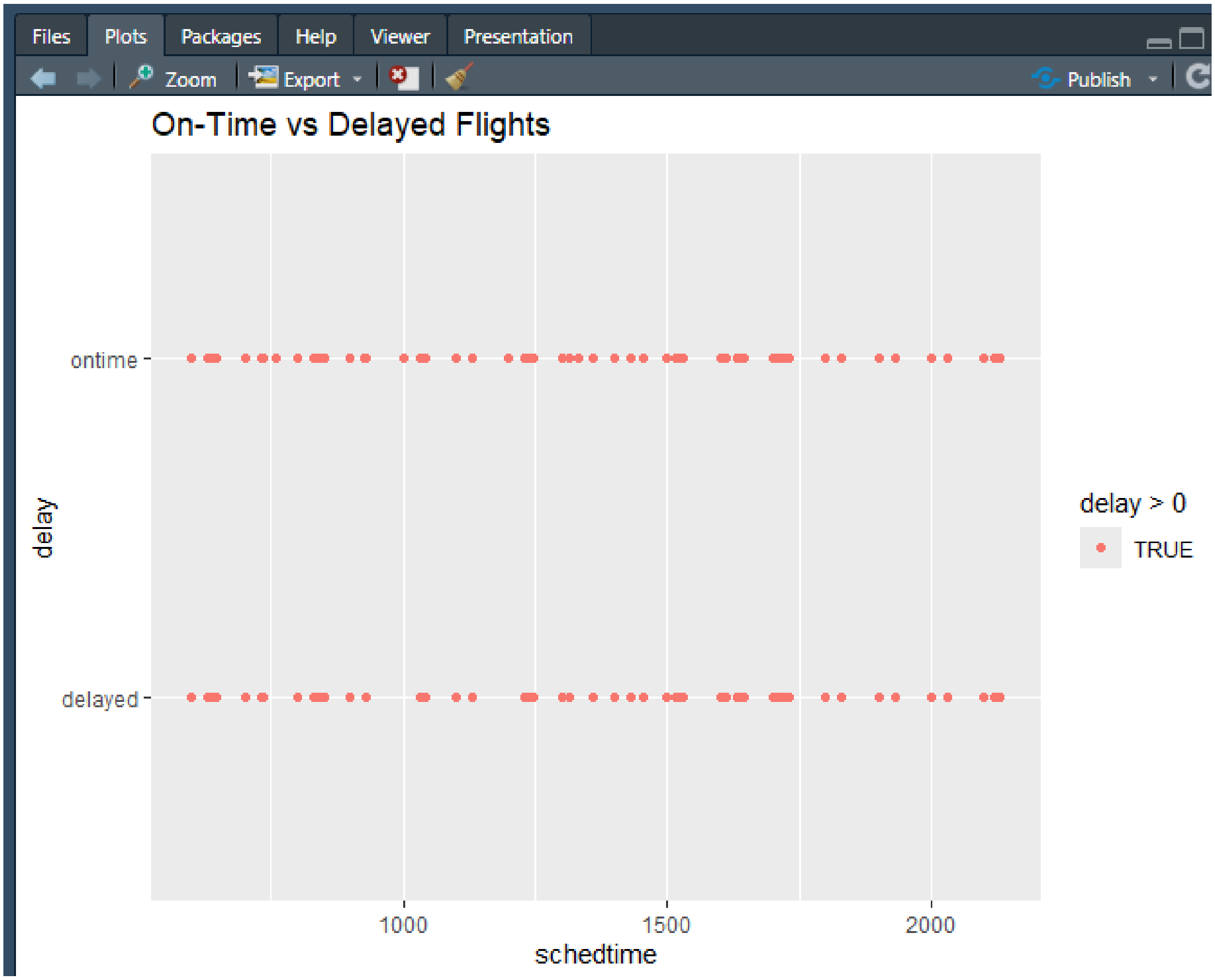
```
#7. Plot the histograms to understand the relationships between scheduled time, carrier, destination, origin, weat
ggplot(df, aes(x = schedtime)) + geom_histogram(binwidth = 10) + ggtitle("Scheduled Time Distribution")
ggplot(df, aes(x = carrier)) + geom_bar() + ggtitle("Carrier Distribution")
ggplot(df, aes(x = dest)) + geom_bar() + ggtitle("Destination Distribution")
ggplot(df, aes(x = origin)) + geom_bar() + ggtitle("Origin Distribution")
ggplot(df, aes(x = weather)) + geom_bar() + ggtitle("Weather Impact on Delays")
ggplot(df, aes(x = dayweek)) + geom_bar() + ggtitle("Flights by Day of the Week")
```



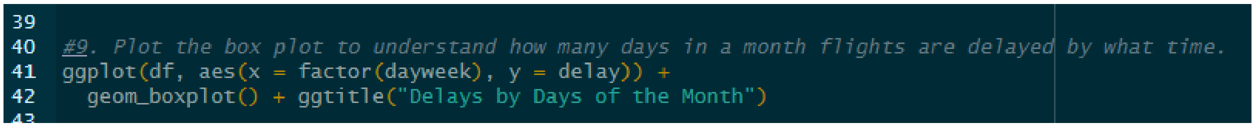
8. Plot the scatter plot for flights on time and delayed



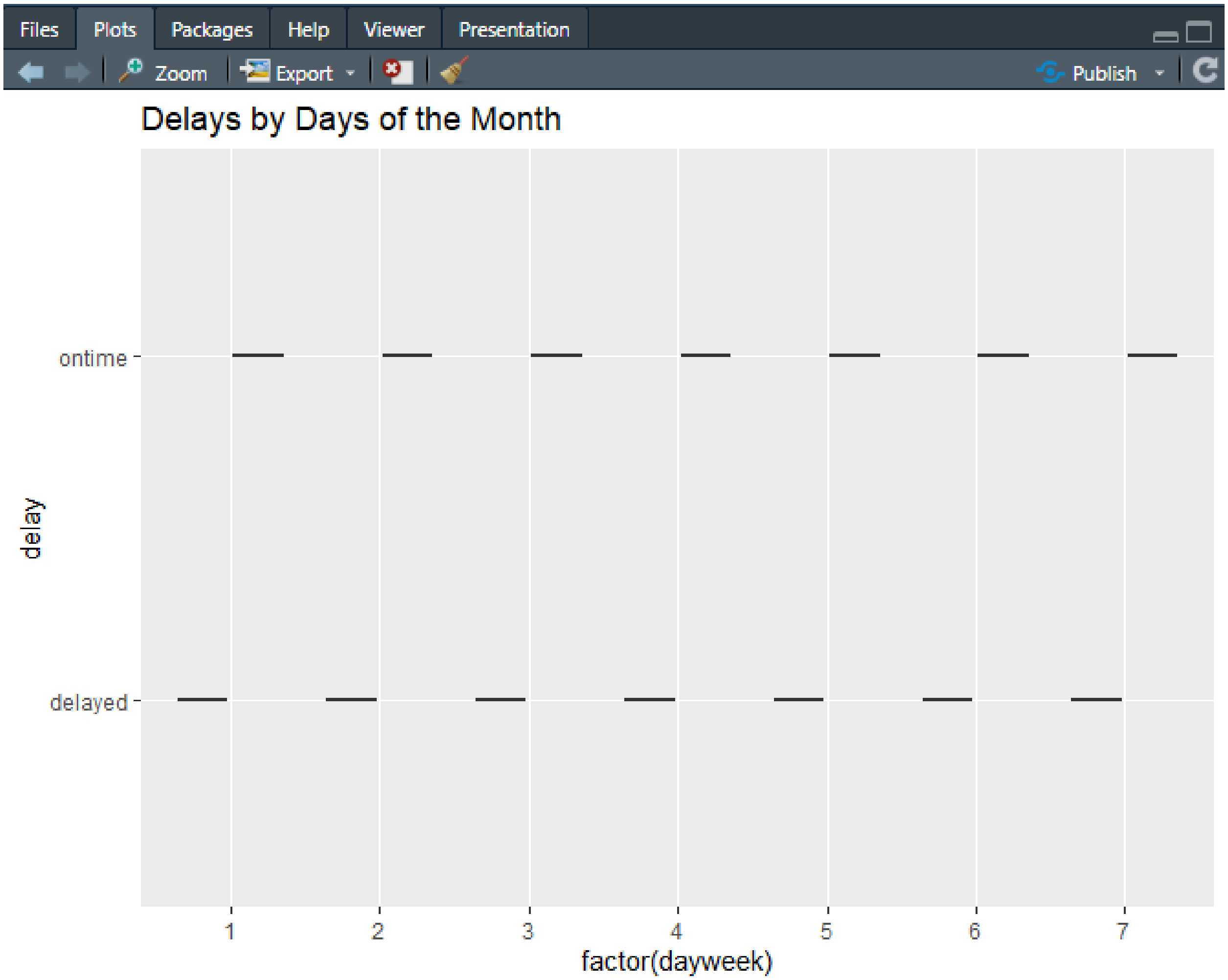
```
#8. Plot the scatter plot for flights on time and delayed.
ggplot(df, aes(x = schedtime, y = delay, color = delay > 0)) +
  geom_point() + ggtitle("On-Time vs Delayed Flights")

```



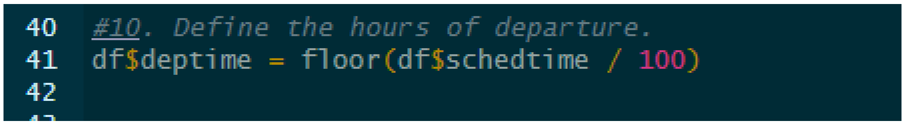
9. Plot the boxplot to understand how many days in a month flights are delayed by what time.



```
#9. Plot the box plot to understand how many days in a month flights are delayed by what time.
ggplot(df, aes(x = factor(dayweek), y = delay)) +
  geom_boxplot() + ggtitle("Delays by Days of the Month")
```

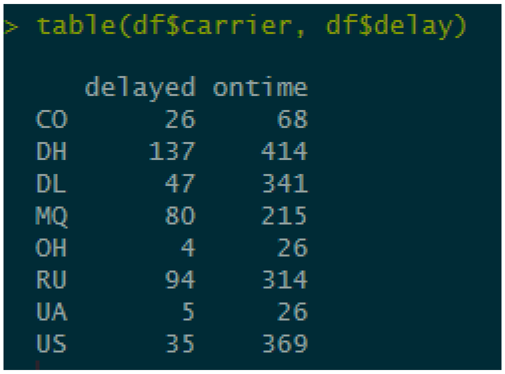


10. Define the hours of departure



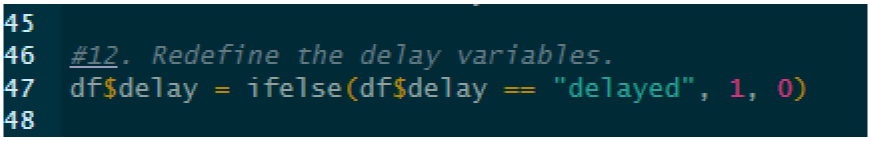
```
#10. Define the hours of departure.
df$deptime = floor(df$schedtime / 100)

```

11. Create a categorical representation of data using a table



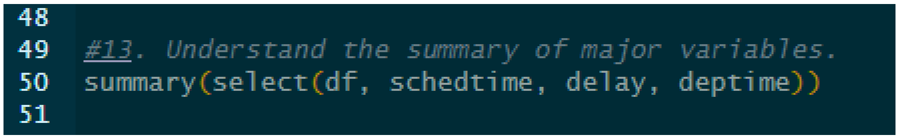
```
> table(df$carrier, df$delay)

     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
```

12. Redefine the delay variables

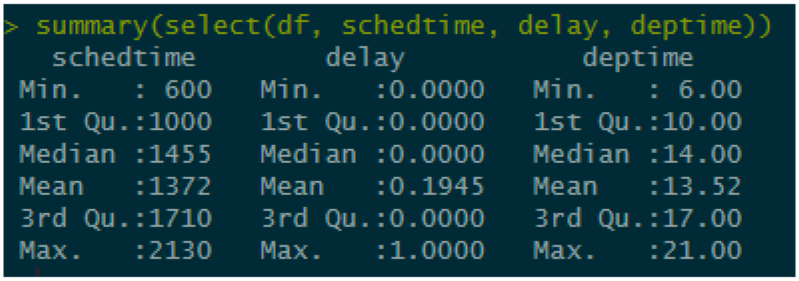


```

#12. Redefine the delay variables.
df$delay = ifelse(df$delay == "delayed", 1, 0)

```

13. Understand the summary of major variables

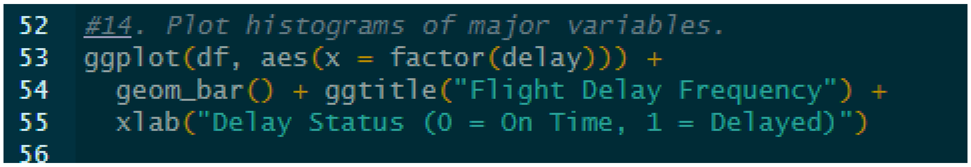


```

#13. Understand the summary of major variables.
summary(select(df, schedtime, delay, deptime))

```

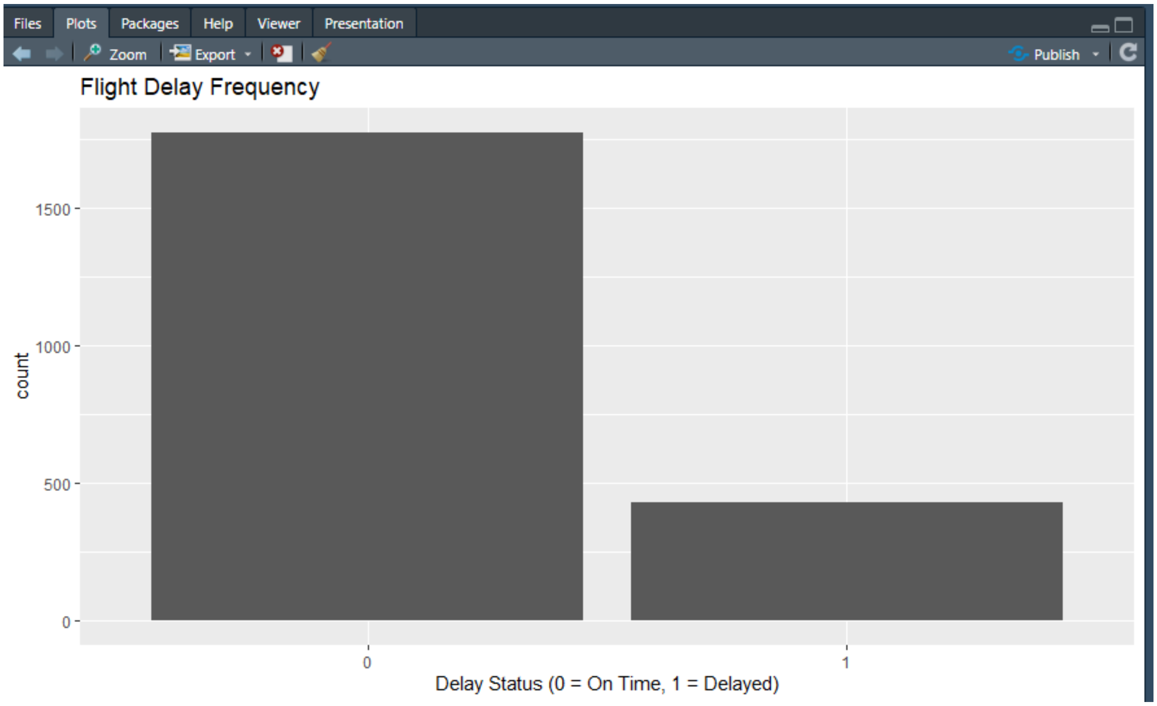


```
> summary(select(df, schedtime, delay, deptime))
   schedtime        delay           deptime     
 Min.   : 600   Min.   :0.0000   Min.   : 6.00  
 1st Qu.:1000   1st Qu.:0.0000   1st Qu.:10.00  
 Median :1455   Median :0.0000   Median :14.00  
 Mean   :1372   Mean   :0.1945   Mean   :13.52  
 3rd Qu.:1710   3rd Qu.:0.0000   3rd Qu.:17.00  
 Max.   :2130   Max.   :1.0000   Max.   :21.00  
```

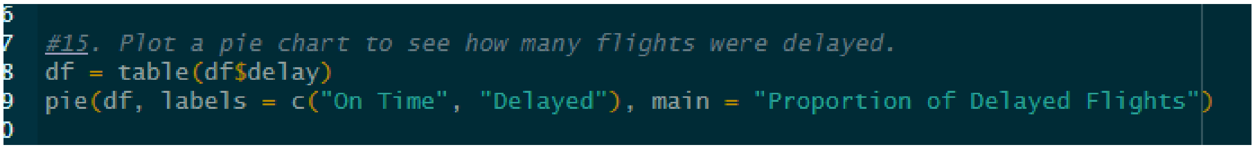
14. Plot histograms of major variables



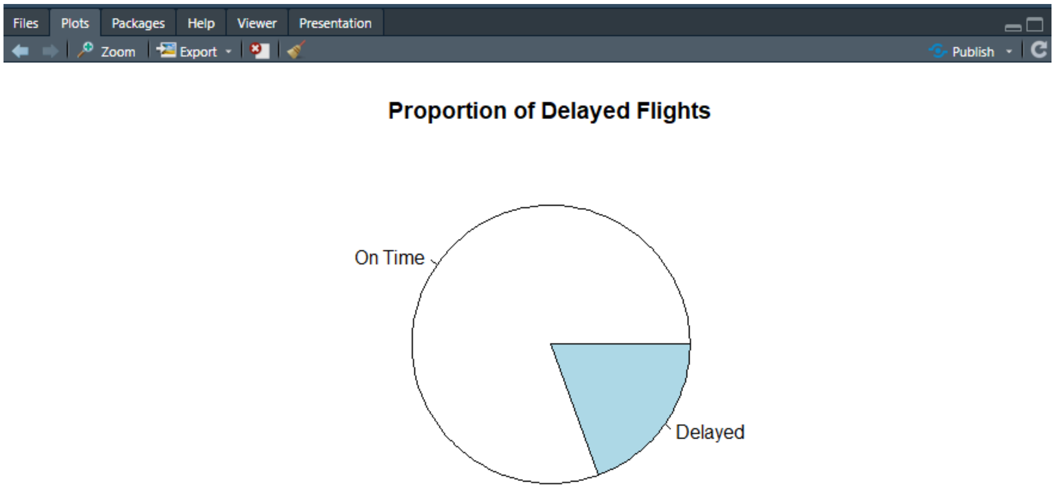
```
#14. Plot histograms of major variables.
ggplot(df, aes(x = factor(delay))) +
  geom_bar() + ggtitle("Flight Delay Frequency") +
  xlab("Delay Status (0 = On Time, 1 = Delayed)")

```



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



```r
#15. Plot a pie chart to see how many flights were delayed.
df = table(df$delay)
pie(df, labels = c("On Time", "Delayed"), main = "Proportion of Delayed Flights")
```


Files
Plots
Packages
Help
Viewer
Presentation
Zoom
Export
Publish
Proportion of Delayed Flights
On Time
Delayed