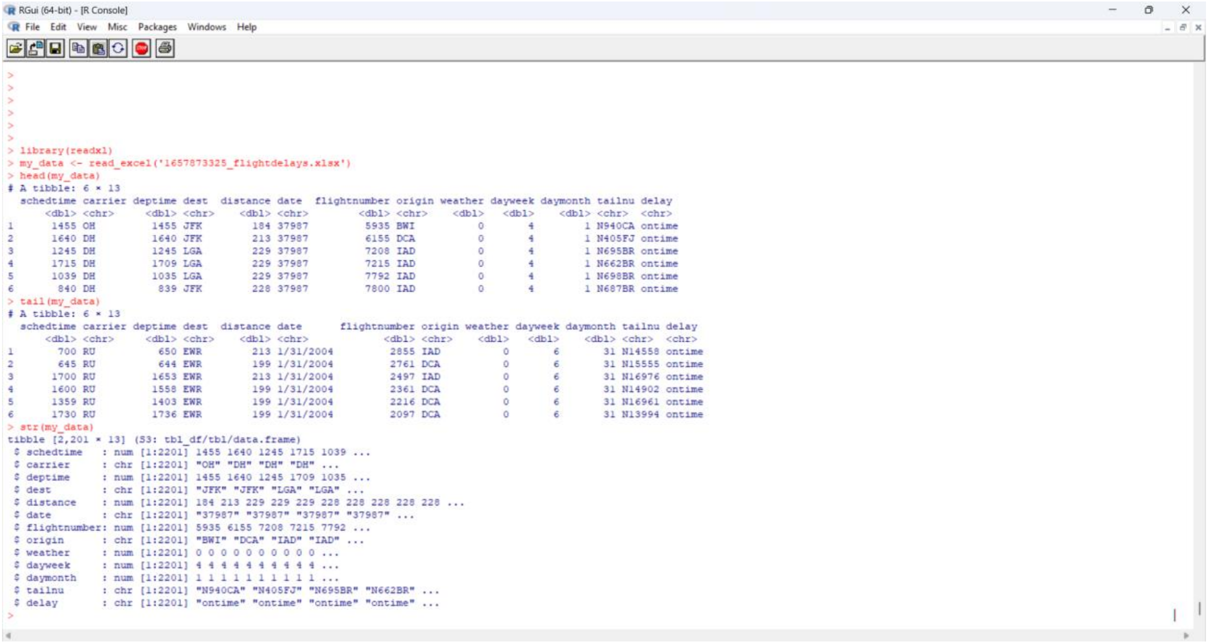
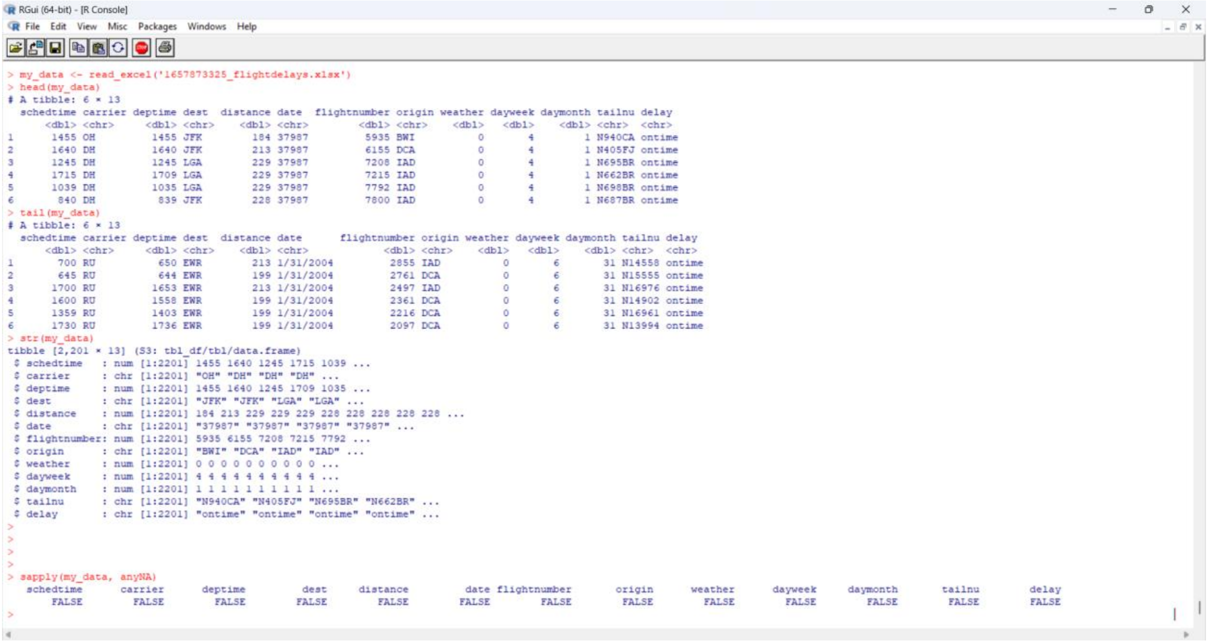
- Read the dataset
- Read the dataset description
- Understand the data



RGui (64-bit) - [R Console]
File Edit View Misc Packages Windows Help

```
>
>
>
>
>
>
> library(readxl)
> my_data <- read_excel('1657873325_flightdelays.xlsx')
> head(my_data)
# A tibble: 6 × 13
  schedtime carrier deptime dest  distance date  flightnumber origin weather dayweek daymonth tailnu delay
      <dbl> <chr>     <dbl> <chr>    <dbl> <chr>        <dbl> <chr>    <dbl>   <dbl>    <dbl> <chr>  <chr>
1      1455 OH         1455 JFK        184 37987         5935 BWI          0       4        1 N940CA ontime
2      1640 DH         1640 JFK        213 37987         6155 DCA          0       4        1 N405FJ ontime
3      1245 DH         1245 LGA        229 37987         7208 IAD          0       4        1 N695BR ontime
4      1715 DH         1709 LGA        229 37987         7215 IAD          0       4        1 N662BR ontime
5      1039 DH         1035 LGA        229 37987         7792 IAD          0       4        1 N698BR ontime
6       840 DH          839 JFK        228 37987         7800 IAD          0       4        1 N687BR ontime
> tail(my_data)
# A tibble: 6 × 13
  schedtime carrier deptime dest  distance date      flightnumber origin weather dayweek daymonth tailnu delay
      <dbl> <chr>     <dbl> <chr>    <dbl> <chr>            <dbl> <chr>    <dbl>   <dbl>    <dbl> <chr>  <chr>
1       700 RU          650 EWR        213 1/31/2004         2855 IAD          0       6       31 N14558 ontime
2       645 RU          644 EWR        199 1/31/2004         2761 DCA          0       6       31 N15555 ontime
3      1700 RU         1653 EWR        213 1/31/2004         2497 IAD          0       6       31 N16976 ontime
4      1600 RU         1558 EWR        199 1/31/2004         2361 DCA          0       6       31 N14902 ontime
5      1359 RU         1403 EWR        199 1/31/2004         2216 DCA          0       6       31 N16961 ontime
6      1730 RU         1736 EWR        199 1/31/2004         2097 DCA          0       6       31 N13994 ontime
> str(my_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" ...
>
```

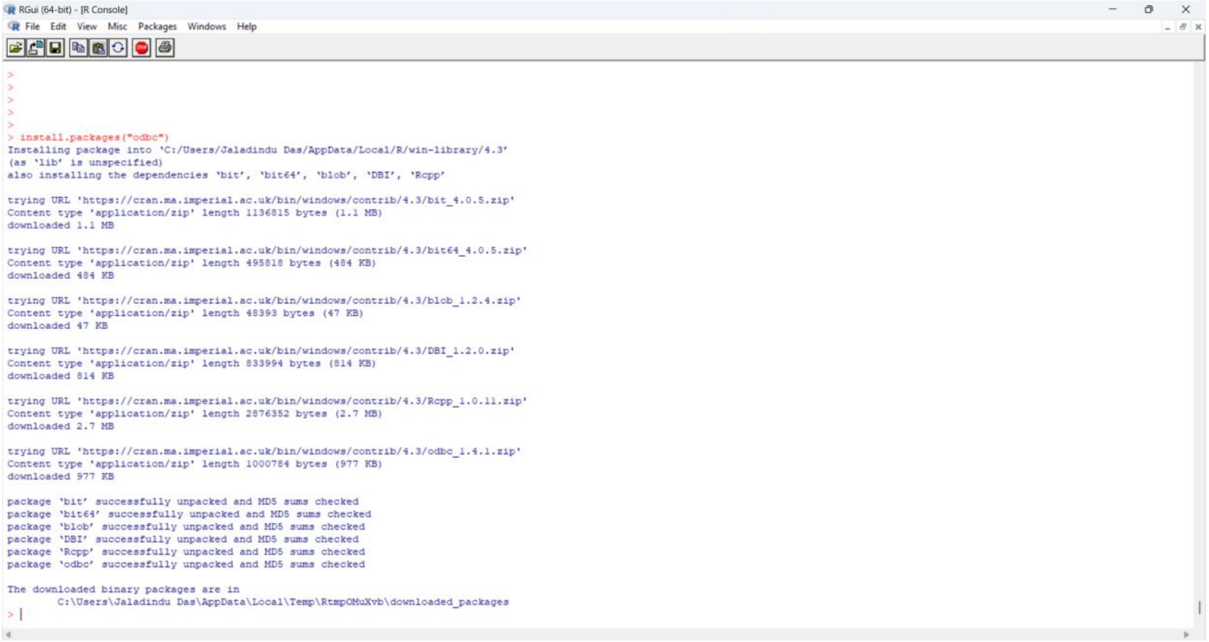
- Find out the null values



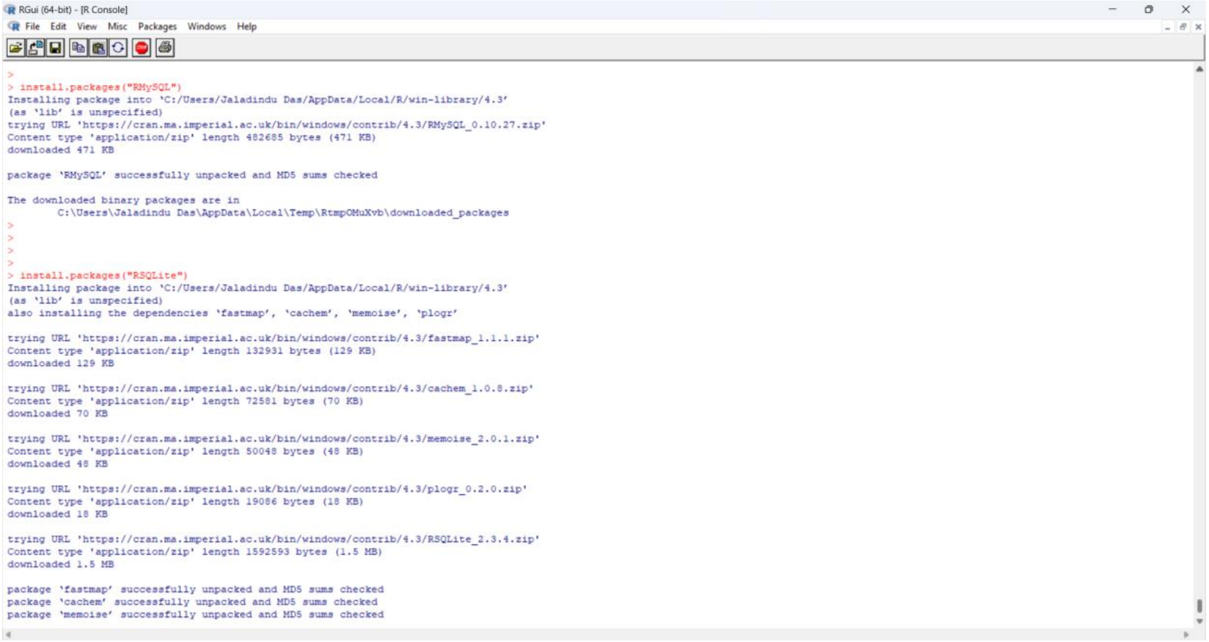
RGui (64-bit) - [R Console]
File Edit View Misc Packages Windows Help

```
> my_data <- read_excel('1657873325_flightdelays.xlsx')
> head(my_data)
# A tibble: 6 × 13
  schedtime carrier deptime dest  distance date  flightnumber origin weather dayweek daymonth tailnu delay
      <dbl> <chr>     <dbl> <chr>    <dbl> <chr>        <dbl> <chr>    <dbl>   <dbl>    <dbl> <chr>  <chr>
1      1455 OH         1455 JFK        184 37987         5935 BWI          0       4        1 N940CA ontime
2      1640 DH         1640 JFK        213 37987         6155 DCA          0       4        1 N405FJ ontime
3      1245 DH         1245 LGA        229 37987         7208 IAD          0       4        1 N695BR ontime
4      1715 DH         1709 LGA        229 37987         7215 IAD          0       4        1 N662BR ontime
5      1039 DH         1035 LGA        229 37987         7792 IAD          0       4        1 N698BR ontime
6       840 DH          839 JFK        228 37987         7800 IAD          0       4        1 N687BR ontime
> tail(my_data)
# A tibble: 6 × 13
  schedtime carrier deptime dest  distance date      flightnumber origin weather dayweek daymonth tailnu delay
      <dbl> <chr>     <dbl> <chr>    <dbl> <chr>            <dbl> <chr>    <dbl>   <dbl>    <dbl> <chr>  <chr>
1       700 RU          650 EWR        213 1/31/2004         2855 IAD          0       6       31 N14558 ontime
2       645 RU          644 EWR        199 1/31/2004         2761 DCA          0       6       31 N15555 ontime
3      1700 RU         1653 EWR        213 1/31/2004         2497 IAD          0       6       31 N16976 ontime
4      1600 RU         1558 EWR        199 1/31/2004         2361 DCA          0       6       31 N14902 ontime
5      1359 RU         1403 EWR        199 1/31/2004         2216 DCA          0       6       31 N16961 ontime
6      1730 RU         1736 EWR        199 1/31/2004         2097 DCA          0       6       31 N13994 ontime
> str(my_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" ...
>
>
>
>
> sapply(my_data, anyNA)
   schedtime      carrier      deptime         dest     distance         date flightnumber       origin      weather      dayweek     daymonth       tailnu        delay
       FALSE        FALSE        FALSE        FALSE        FALSE        FALSE        FALSE        FALSE        FALSE        FALSE        FALSE        FALSE        FALSE
>
```

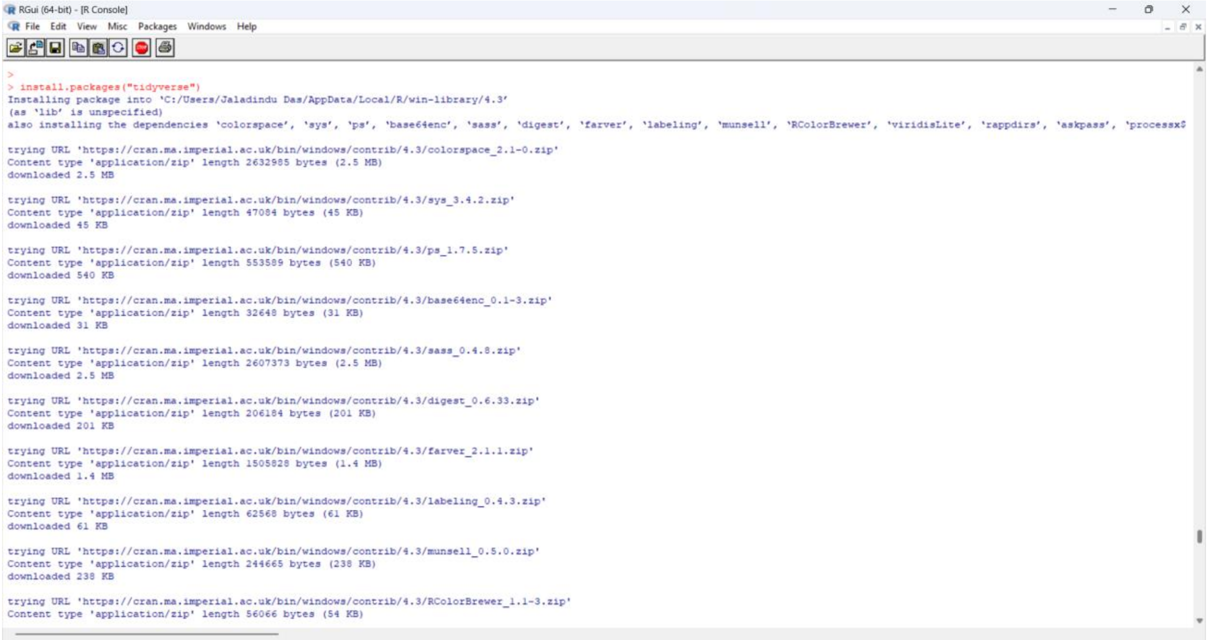
- Install the required packages


RGui (64-bit) - [R Console]
File Edit View Misc Packages Windows Help

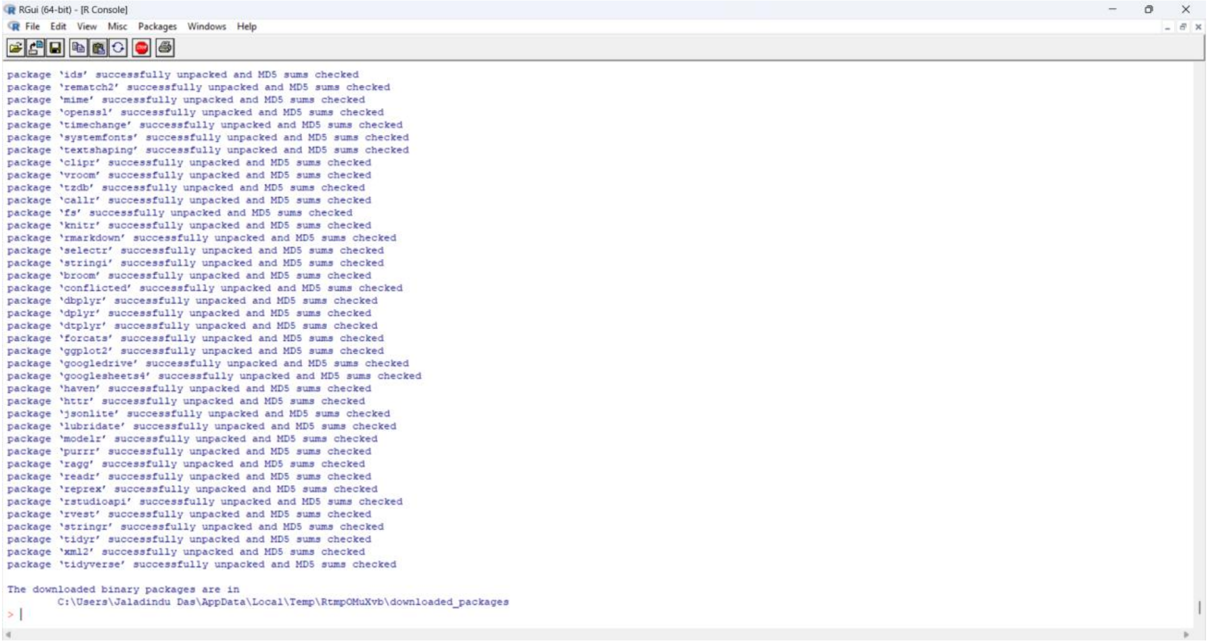
```
>
>
>
>
>
> install.packages("odbc")
Installing package into ‘C:/Users/Jaladindu Das/AppData/Local/R/win-library/4.3’
(as ‘lib’ is unspecified)
also installing the dependencies ‘bit’, ‘bit64’, ‘blob’, ‘DBI’, ‘Rcpp’

trying URL 'https://cran.ma.imperial.ac.uk/bin/windows/contrib/4.3/bit_4.0.5.zip'
Content type 'application/zip' length 1136815 bytes (1.1 MB)
downloaded 1.1 MB

trying URL 'https://cran.ma.imperial.ac.uk/bin/windows/contrib/4.3/bit64_4.0.5.zip'
Content type 'application/zip' length 495818 bytes (484 KB)
downloaded 484 KB

trying URL 'https://cran.ma.imperial.ac.uk/bin/windows/contrib/4.3/blob_1.2.4.zip'
Content type 'application/zip' length 48393 bytes (47 KB)
downloaded 47 KB

trying URL 'https://cran.ma.imperial.ac.uk/bin/windows/contrib/4.3/DBI_1.2.0.zip'
Content type 'application/zip' length 833994 bytes (814 KB)
downloaded 814 KB

trying URL 'https://cran.ma.imperial.ac.uk/bin/windows/contrib/4.3/Rcpp_1.0.11.zip'
Content type 'application/zip' length 2876352 bytes (2.7 MB)
downloaded 2.7 MB

trying URL 'https://cran.ma.imperial.ac.uk/bin/windows/contrib/4.3/odbc_1.4.1.zip'
Content type 'application/zip' length 1000784 bytes (977 KB)
downloaded 977 KB

package ‘bit’ successfully unpacked and MD5 sums checked
package ‘bit64’ successfully unpacked and MD5 sums checked
package ‘blob’ successfully unpacked and MD5 sums checked
package ‘DBI’ successfully unpacked and MD5 sums checked
package ‘Rcpp’ successfully unpacked and MD5 sums checked
package ‘odbc’ successfully unpacked and MD5 sums checked

The downloaded binary packages are in
        C:\Users\Jaladindu Das\AppData\Local\Temp\RtmpOMuXvb\downloaded_packages
> |
```


RGui (64-bit) - [R Console]
File Edit View Misc Packages Windows Help

```
>
> install.packages("RMySQL")
Installing package into ‘C:/Users/Jaladindu Das/AppData/Local/R/win-library/4.3’
(as ‘lib’ is unspecified)
trying URL 'https://cran.ma.imperial.ac.uk/bin/windows/contrib/4.3/RMySQL_0.10.27.zip'
Content type 'application/zip' length 482685 bytes (471 KB)
downloaded 471 KB

package ‘RMySQL’ successfully unpacked and MD5 sums checked

The downloaded binary packages are in
        C:\Users\Jaladindu Das\AppData\Local\Temp\RtmpOMuXvb\downloaded_packages
>
>
>
>
> install.packages("RSQLite")
Installing package into ‘C:/Users/Jaladindu Das/AppData/Local/R/win-library/4.3’
(as ‘lib’ is unspecified)
also installing the dependencies ‘fastmap’, ‘cachem’, ‘memoise’, ‘plogr’

trying URL 'https://cran.ma.imperial.ac.uk/bin/windows/contrib/4.3/fastmap_1.1.1.zip'
Content type 'application/zip' length 132931 bytes (129 KB)
downloaded 129 KB

trying URL 'https://cran.ma.imperial.ac.uk/bin/windows/contrib/4.3/cachem_1.0.8.zip'
Content type 'application/zip' length 72581 bytes (70 KB)
downloaded 70 KB

trying URL 'https://cran.ma.imperial.ac.uk/bin/windows/contrib/4.3/memoise_2.0.1.zip'
Content type 'application/zip' length 50048 bytes (48 KB)
downloaded 48 KB

trying URL 'https://cran.ma.imperial.ac.uk/bin/windows/contrib/4.3/plogr_0.2.0.zip'
Content type 'application/zip' length 19086 bytes (18 KB)
downloaded 18 KB

trying URL 'https://cran.ma.imperial.ac.uk/bin/windows/contrib/4.3/RSQLite_2.3.4.zip'
Content type 'application/zip' length 1592593 bytes (1.5 MB)
downloaded 1.5 MB

package ‘fastmap’ successfully unpacked and MD5 sums checked
package ‘cachem’ successfully unpacked and MD5 sums checked
package ‘memoise’ successfully unpacked and MD5 sums checked
```

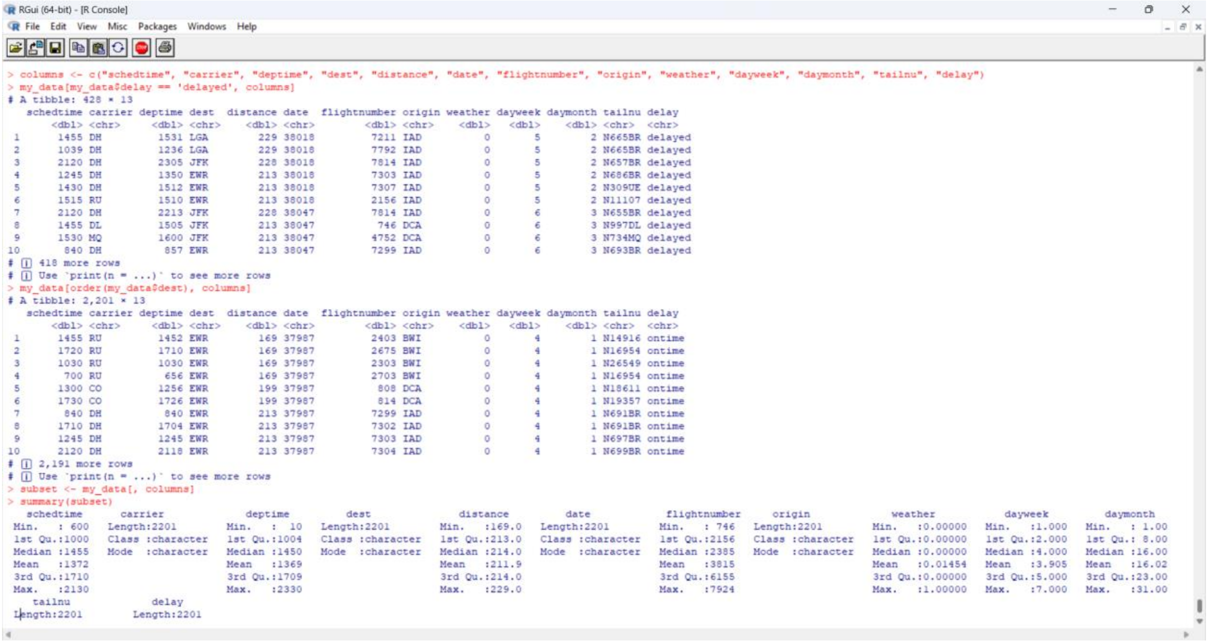

RGui (64-bit) - [R Console]
File Edit View Misc Packages Windows Help

```
>
> install.packages("tidyverse")
Installing package into 'C:/Users/Jaladindu Das/AppData/Local/R/win-library/4.3'
(as 'lib' is unspecified)
also installing the dependencies 'colorspace', 'sys', 'ps', 'base64enc', 'sass', 'digest', 'farver', 'labeling', 'munsell', 'RColorBrewer', 'viridisLite', 'rappdirs', 'askpass', 'processx$

trying URL 'https://cran.ma.imperial.ac.uk/bin/windows/contrib/4.3/colorspace_2.1-0.zip'
Content type 'application/zip' length 2632985 bytes (2.5 MB)
downloaded 2.5 MB

trying URL 'https://cran.ma.imperial.ac.uk/bin/windows/contrib/4.3/sys_3.4.2.zip'
Content type 'application/zip' length 47084 bytes (45 KB)
downloaded 45 KB

trying URL 'https://cran.ma.imperial.ac.uk/bin/windows/contrib/4.3/ps_1.7.5.zip'
Content type 'application/zip' length 553589 bytes (540 KB)
downloaded 540 KB

trying URL 'https://cran.ma.imperial.ac.uk/bin/windows/contrib/4.3/base64enc_0.1-3.zip'
Content type 'application/zip' length 32648 bytes (31 KB)
downloaded 31 KB

trying URL 'https://cran.ma.imperial.ac.uk/bin/windows/contrib/4.3/sass_0.4.8.zip'
Content type 'application/zip' length 2607373 bytes (2.5 MB)
downloaded 2.5 MB

trying URL 'https://cran.ma.imperial.ac.uk/bin/windows/contrib/4.3/digest_0.6.33.zip'
Content type 'application/zip' length 206184 bytes (201 KB)
downloaded 201 KB

trying URL 'https://cran.ma.imperial.ac.uk/bin/windows/contrib/4.3/farver_2.1.1.zip'
Content type 'application/zip' length 1505828 bytes (1.4 MB)
downloaded 1.4 MB

trying URL 'https://cran.ma.imperial.ac.uk/bin/windows/contrib/4.3/labeling_0.4.3.zip'
Content type 'application/zip' length 62568 bytes (61 KB)
downloaded 61 KB

trying URL 'https://cran.ma.imperial.ac.uk/bin/windows/contrib/4.3/munsell_0.5.0.zip'
Content type 'application/zip' length 244665 bytes (238 KB)
downloaded 238 KB

trying URL 'https://cran.ma.imperial.ac.uk/bin/windows/contrib/4.3/RColorBrewer_1.1-3.zip'
Content type 'application/zip' length 56066 bytes (54 KB)
```

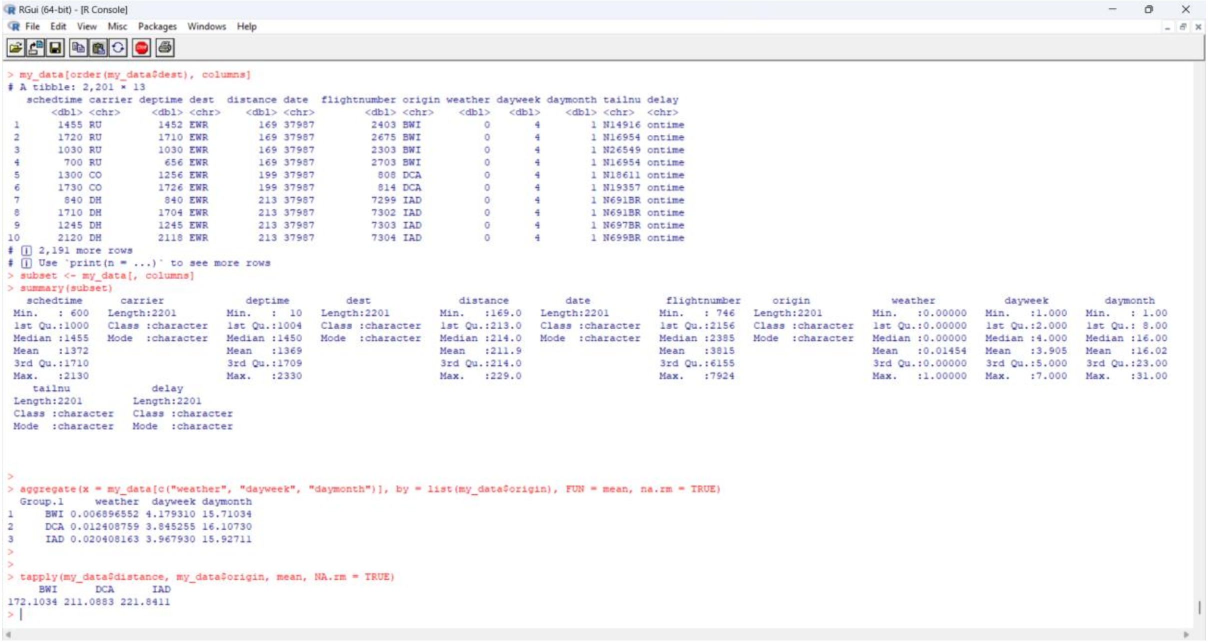

RGui (64-bit) - [R Console]
File Edit View Misc Packages Windows Help

```
package 'ids' successfully unpacked and MD5 sums checked
package 'rematch2' successfully unpacked and MD5 sums checked
package 'mime' successfully unpacked and MD5 sums checked
package 'openssl' successfully unpacked and MD5 sums checked
package 'timechange' successfully unpacked and MD5 sums checked
package 'systemfonts' successfully unpacked and MD5 sums checked
package 'textshaping' successfully unpacked and MD5 sums checked
package 'clipr' successfully unpacked and MD5 sums checked
package 'vroom' successfully unpacked and MD5 sums checked
package 'tzdb' successfully unpacked and MD5 sums checked
package 'callr' successfully unpacked and MD5 sums checked
package 'fs' successfully unpacked and MD5 sums checked
package 'knitr' successfully unpacked and MD5 sums checked
package 'rmarkdown' successfully unpacked and MD5 sums checked
package 'selectr' successfully unpacked and MD5 sums checked
package 'stringi' successfully unpacked and MD5 sums checked
package 'broom' successfully unpacked and MD5 sums checked
package 'conflicted' successfully unpacked and MD5 sums checked
package 'dbplyr' successfully unpacked and MD5 sums checked
package 'dplyr' successfully unpacked and MD5 sums checked
package 'dtplyr' successfully unpacked and MD5 sums checked
package 'forcats' successfully unpacked and MD5 sums checked
package 'ggplot2' successfully unpacked and MD5 sums checked
package 'googledrive' successfully unpacked and MD5 sums checked
package 'googlesheets4' successfully unpacked and MD5 sums checked
package 'haven' successfully unpacked and MD5 sums checked
package 'httr' successfully unpacked and MD5 sums checked
package 'jsonlite' successfully unpacked and MD5 sums checked
package 'lubridate' successfully unpacked and MD5 sums checked
package 'modelr' successfully unpacked and MD5 sums checked
package 'purrr' successfully unpacked and MD5 sums checked
package 'ragg' successfully unpacked and MD5 sums checked
package 'readr' successfully unpacked and MD5 sums checked
package 'reprex' successfully unpacked and MD5 sums checked
package 'rstudioapi' successfully unpacked and MD5 sums checked
package 'rvest' successfully unpacked and MD5 sums checked
package 'stringr' successfully unpacked and MD5 sums checked
package 'tidyr' successfully unpacked and MD5 sums checked
package 'xml2' successfully unpacked and MD5 sums checked
package 'tidyverse' successfully unpacked and MD5 sums checked

The downloaded binary packages are in
        C:\Users\Jaladindu Das\AppData\Local\Temp\RtmpOMuXvb\downloaded_packages
> |
```

- Understand the summary of descriptive statistics

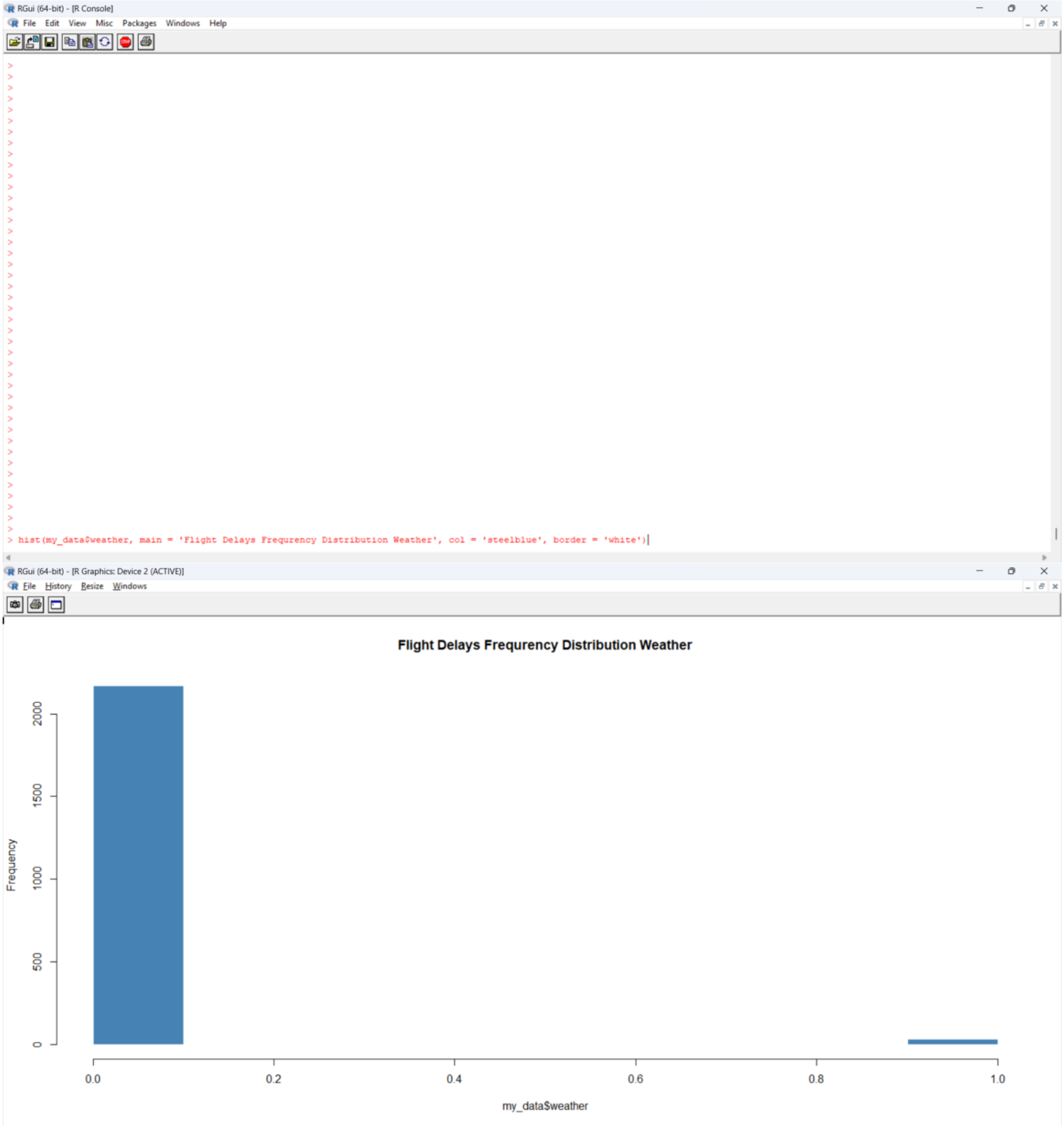

RGui (64-bit) - [R Console]
File Edit View Misc Packages Windows Help

```
> columns <- c("schedtime", "carrier", "deptime", "dest", "distance", "date", "flightnumber", "origin", "weather", "dayweek", "daymonth", "tailnu", "delay")
> my_data[my_data$delay == 'delayed', columns]
# A tibble: 428 × 13
   schedtime carrier deptime dest  distance date  flightnumber origin weather dayweek daymonth tailnu delay
       <dbl> <chr>     <dbl> <chr>    <dbl> <chr>        <dbl> <chr>    <dbl>   <dbl>    <dbl> <chr>  <chr>
 1      1455 DH         1531 LGA        229 38018         7211 IAD          0       5        2 N665BR delayed
 2      1039 DH         1236 LGA        229 38018         7792 IAD          0       5        2 N665BR delayed
 3      2120 DH         2305 JFK        228 38018         7814 IAD          0       5        2 N657BR delayed
 4      1245 DH         1350 EWR        213 38018         7303 IAD          0       5        2 N686BR delayed
 5      1430 DH         1512 EWR        213 38018         7307 IAD          0       5        2 N309UE delayed
 6      1515 RU         1510 EWR        213 38018         2156 IAD          0       5        2 N11107 delayed
 7      2120 DH         2213 JFK        228 38047         7814 IAD          0       6        3 N655BR delayed
 8      1455 DL         1505 JFK        213 38047          746 DCA          0       6        3 N997DL delayed
 9      1530 MQ         1600 JFK        213 38047         4752 DCA          0       6        3 N734MQ delayed
10       840 DH          857 EWR        213 38047         7299 IAD          0       6        3 N693BR delayed
# ℹ 418 more rows
# ℹ Use `print(n = ...)` to see more rows
> my_data[order(my_data$dest), columns]
# A tibble: 2,201 × 13
   schedtime carrier deptime dest  distance date  flightnumber origin weather dayweek daymonth tailnu delay
       <dbl> <chr>     <dbl> <chr>    <dbl> <chr>        <dbl> <chr>    <dbl>   <dbl>    <dbl> <chr>  <chr>
 1      1455 RU         1452 EWR        169 37987         2403 BWI          0       4        1 N14916 ontime
 2      1720 RU         1710 EWR        169 37987         2675 BWI          0       4        1 N16954 ontime
 3      1030 RU         1030 EWR        169 37987         2303 BWI          0       4        1 N26549 ontime
 4       700 RU          656 EWR        169 37987         2703 BWI          0       4        1 N16954 ontime
 5      1300 CO         1256 EWR        199 37987          808 DCA          0       4        1 N18611 ontime
 6      1730 CO         1726 EWR        199 37987          814 DCA          0       4        1 N19357 ontime
 7       840 DH          840 EWR        213 37987         7299 IAD          0       4        1 N691BR ontime
 8      1710 DH         1704 EWR        213 37987         7302 IAD          0       4        1 N691BR ontime
 9      1245 DH         1245 EWR        213 37987         7303 IAD          0       4        1 N697BR ontime
10      2120 DH         2118 EWR        213 37987         7304 IAD          0       4        1 N699BR ontime
# ℹ 2,191 more rows
# ℹ Use `print(n = ...)` to see more rows
> subset <- my_data[, columns]
> summary(subset)
   schedtime      carrier             deptime         dest              distance         date            flightnumber     origin             weather            dayweek         daymonth    
 Min.   : 600   Length:2201        Min.   :  10   Length:2201        Min.   :169.0   Length:2201        Min.   : 746   Length:2201        Min.   :0.00000   Min.   :1.000   Min.   : 1.00  
 1st Qu.:1000   Class :character   1st Qu.:1004   Class :character   1st Qu.:213.0   Class :character   1st Qu.:2156   Class :character   1st Qu.:0.00000   1st Qu.:2.000   1st Qu.: 8.00  
 Median :1455   Mode  :character   Median :1450   Mode  :character   Median :214.0   Mode  :character   Median :2385   Mode  :character   Median :0.00000   Median :4.000   Median :16.00  
 Mean   :1372                      Mean   :1369                      Mean   :211.9                      Mean   :3815                      Mean   :0.01454   Mean   :3.905   Mean   :16.02  
 3rd Qu.:1710                      3rd Qu.:1709                      3rd Qu.:214.0                      3rd Qu.:6155                      3rd Qu.:0.00000   3rd Qu.:5.000   3rd Qu.:23.00  
 Max.   :2130                      Max.   :2330                      Max.   :229.0                      Max.   :7924                      Max.   :1.00000   Max.   :7.000   Max.   :31.00  
    tailnu              delay          
 Length:2201        Length:2201       
```


RGui (64-bit) - [R Console]
File Edit View Misc Packages Windows Help

```
> my_data[order(my_data$dest), columns]
# A tibble: 2,201 × 13
   schedtime carrier deptime dest  distance date  flightnumber origin weather dayweek daymonth tailnu delay
       <dbl> <chr>     <dbl> <chr>    <dbl> <chr>        <dbl> <chr>    <dbl>   <dbl>    <dbl> <chr>  <chr>
 1      1455 RU         1452 EWR        169 37987         2403 BWI          0       4        1 N14916 ontime
 2      1720 RU         1710 EWR        169 37987         2675 BWI          0       4        1 N16954 ontime
 3      1030 RU         1030 EWR        169 37987         2303 BWI          0       4        1 N26549 ontime
 4       700 RU          656 EWR        169 37987         2703 BWI          0       4        1 N16954 ontime
 5      1300 CO         1256 EWR        199 37987          808 DCA          0       4        1 N18611 ontime
 6      1730 CO         1726 EWR        199 37987          814 DCA          0       4        1 N19357 ontime
 7       840 DH          840 EWR        213 37987         7299 IAD          0       4        1 N691BR ontime
 8      1710 DH         1704 EWR        213 37987         7302 IAD          0       4        1 N691BR ontime
 9      1245 DH         1245 EWR        213 37987         7303 IAD          0       4        1 N697BR ontime
10      2120 DH         2118 EWR        213 37987         7304 IAD          0       4        1 N699BR ontime
# ℹ 2,191 more rows
# ℹ Use `print(n = ...)` to see more rows
> subset <- my_data[, columns]
> summary(subset)
   schedtime      carrier             deptime         dest              distance         date            flightnumber     origin             weather            dayweek         daymonth    
 Min.   : 600   Length:2201        Min.   :  10   Length:2201        Min.   :169.0   Length:2201        Min.   : 746   Length:2201        Min.   :0.00000   Min.   :1.000   Min.   : 1.00  
 1st Qu.:1000   Class :character   1st Qu.:1004   Class :character   1st Qu.:213.0   Class :character   1st Qu.:2156   Class :character   1st Qu.:0.00000   1st Qu.:2.000   1st Qu.: 8.00  
 Median :1455   Mode  :character   Median :1450   Mode  :character   Median :214.0   Mode  :character   Median :2385   Mode  :character   Median :0.00000   Median :4.000   Median :16.00  
 Mean   :1372                      Mean   :1369                      Mean   :211.9                      Mean   :3815                      Mean   :0.01454   Mean   :3.905   Mean   :16.02  
 3rd Qu.:1710                      3rd Qu.:1709                      3rd Qu.:214.0                      3rd Qu.:6155                      3rd Qu.:0.00000   3rd Qu.:5.000   3rd Qu.:23.00  
 Max.   :2130                      Max.   :2330                      Max.   :229.0                      Max.   :7924                      Max.   :1.00000   Max.   :7.000   Max.   :31.00  
    tailnu              delay          
 Length:2201        Length:2201       
 Class :character   Class :character  
 Mode  :character   Mode  :character  



>
> aggregate(x = my_data[c("weather", "dayweek", "daymonth")], by = list(my_data$origin), FUN = mean, na.rm = TRUE)
  Group.1     weather  dayweek daymonth
1     BWI 0.006896552 4.179310 15.71034
2     DCA 0.012408759 3.845255 16.10730
3     IAD 0.020408163 3.967930 15.92711
>
>
> tapply(my_data$distance, my_data$origin, mean, NA.rm = TRUE)
     BWI      DCA      IAD 
172.1034 211.0883 221.8411 
> 
```

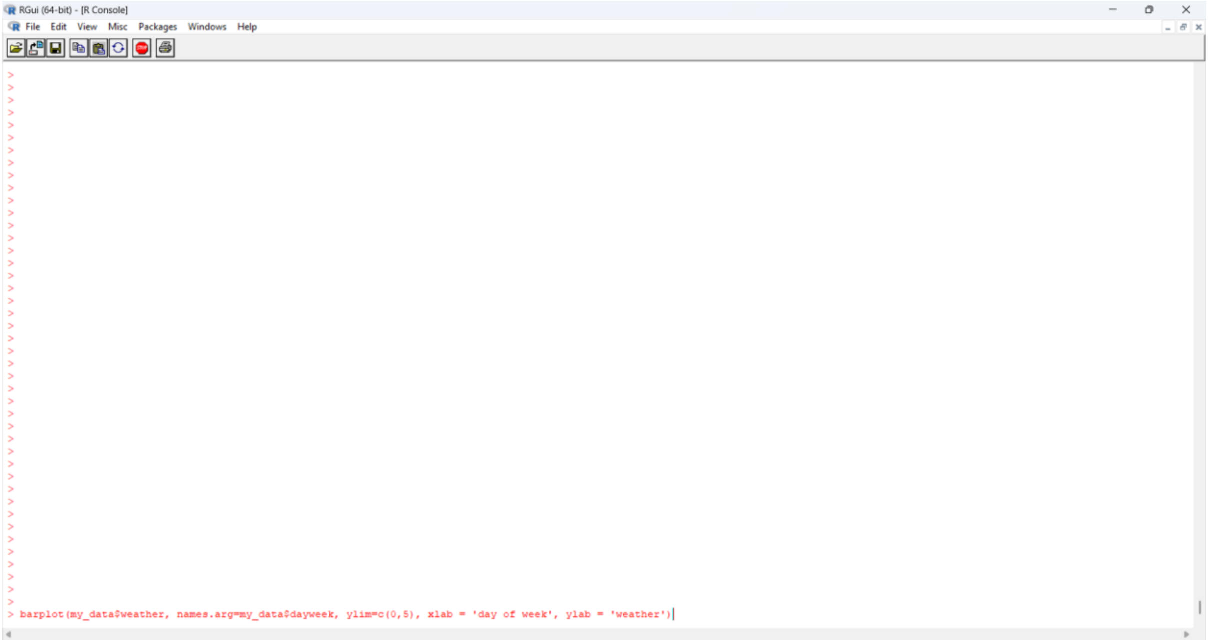
- Plot the histograms to understand the relationships between scheduled time, carrier, destination, origin, weather, and day of the week



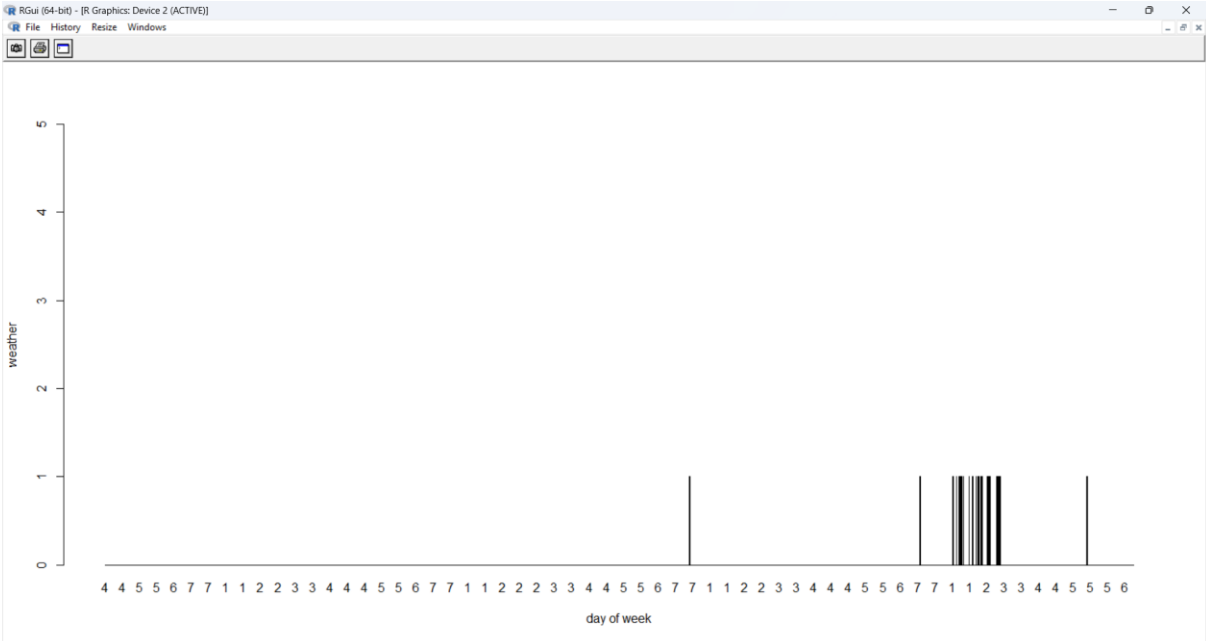
RGui (64-bit) - [R Console]
File Edit View Misc Packages Windows Help

```
>
>
>
>
>
>
>
>
>
>
>
>
>
>
>
>
>
>
>
>
>
>
>
>
>
>
>
>
>
>
>
>
>
>
>
>
>
>
>
>
>
>
>
> hist(my_data$weather, main = 'Flight Delays Frequrency Distribution Weather', col = 'steelblue', border = 'white')
```

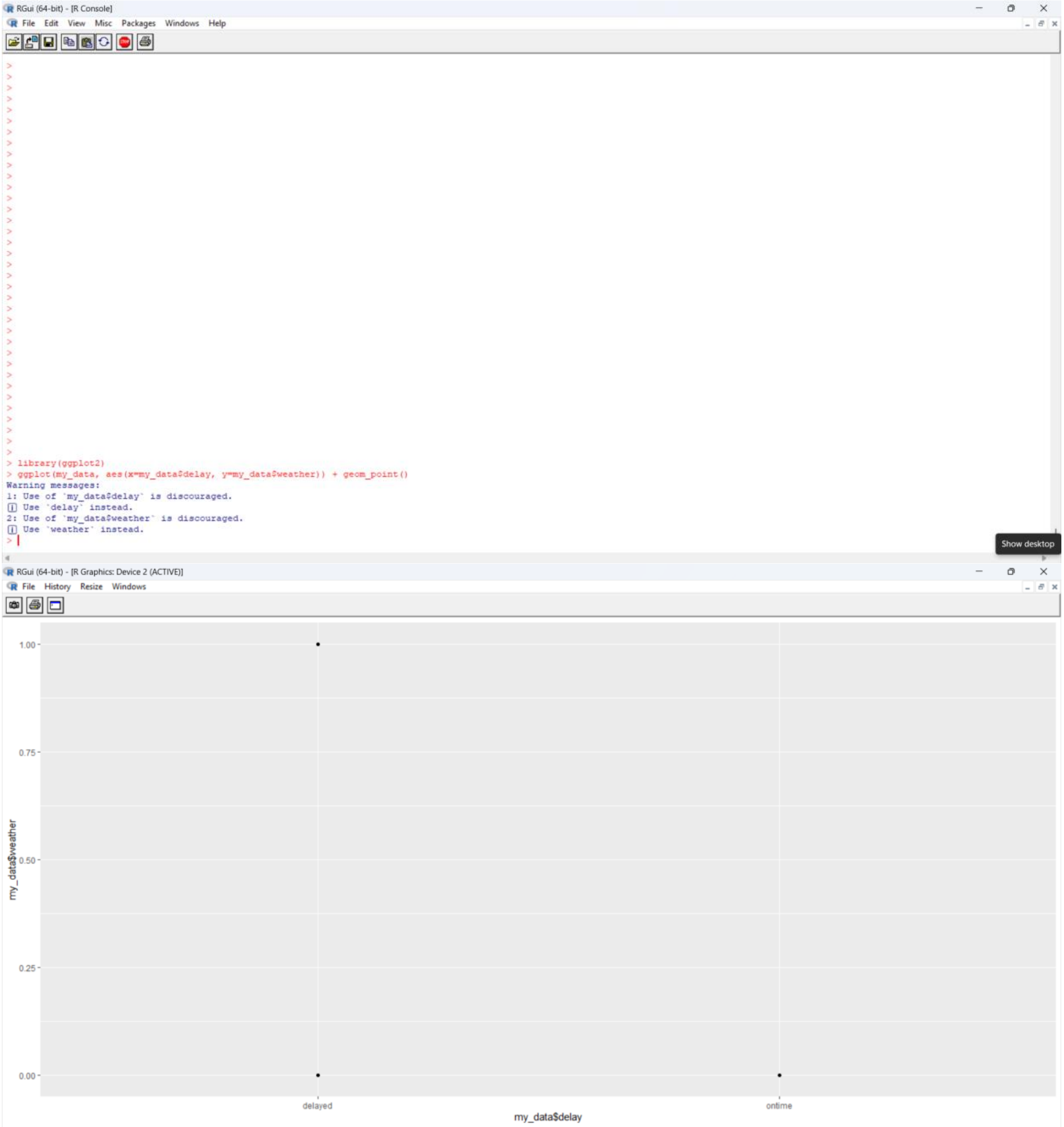
RGui (64-bit) - [R Graphics: Device 2 (ACTIVE)]
File History Resize Windows
Flight Delays Frequrency Distribution Weather
Frequency
0
500
1000
1500
2000
0.0
0.2
0.4
0.6
0.8
1.0
my_data$weather



```
>
>
>
>
>
>
>
>
>
>
>
>
>
>
>
>
>
>
>
>
>
>
>
>
>
>
>
>
>
>
>
>
>
>
>
>
>
>
>
>
>
>
>
> barplot(my_data$weather, names.arg=my_data$dayweek, ylim=c(0,5), xlab = 'day of week', ylab = 'weather')
```

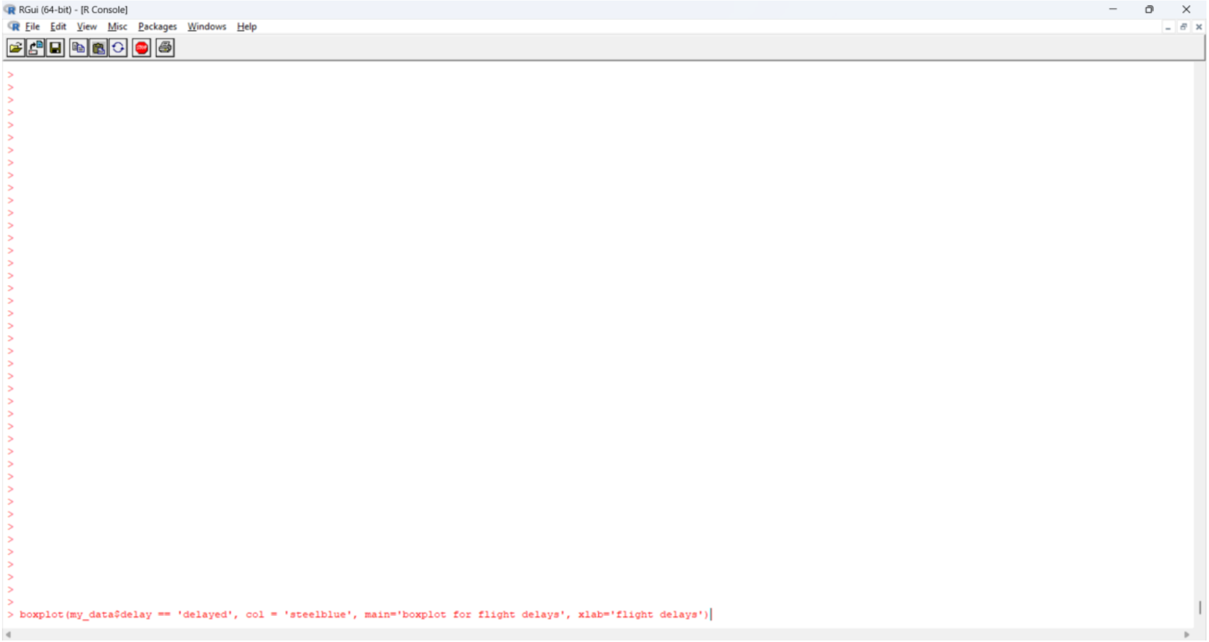


- Plot the scatter plot for flights on time and delayed

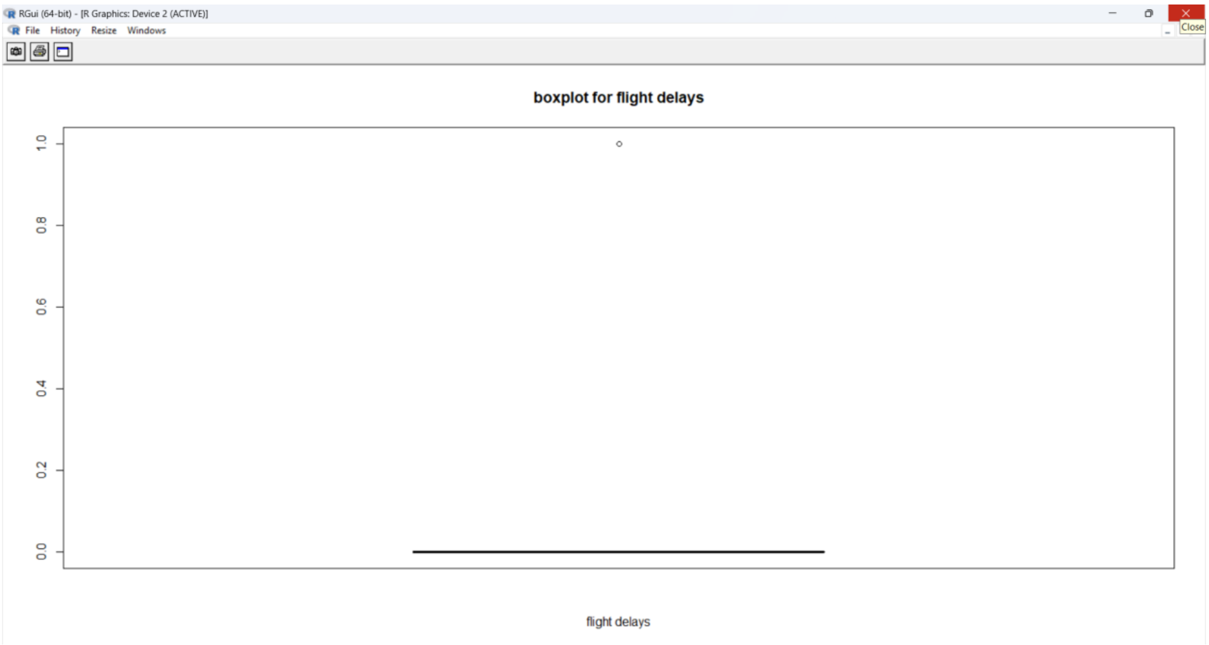

RGui (64-bit) - [R Console]
File Edit View Misc Packages Windows Help
Show desktop
RGui (64-bit) - [R Graphics: Device 2 (ACTIVE)]
File History Resize Windows
1.00
0.75
0.50
0.25
0.00
my_data$weather
delayed
ontime
my_data$delay

```
>
>
>
>
>
>
>
>
>
>
>
>
>
>
>
>
>
>
>
>
>
>
>
>
>
>
>
>
>
>
>
>
>
>
>
>
> library(ggplot2)
> ggplot(my_data, aes(x=my_data$delay, y=my_data$weather)) + geom_point()
Warning messages:
1: Use of `my_data$delay` is discouraged.
ℹ Use `delay` instead.
2: Use of `my_data$weather` is discouraged.
ℹ Use `weather` instead.
>
```

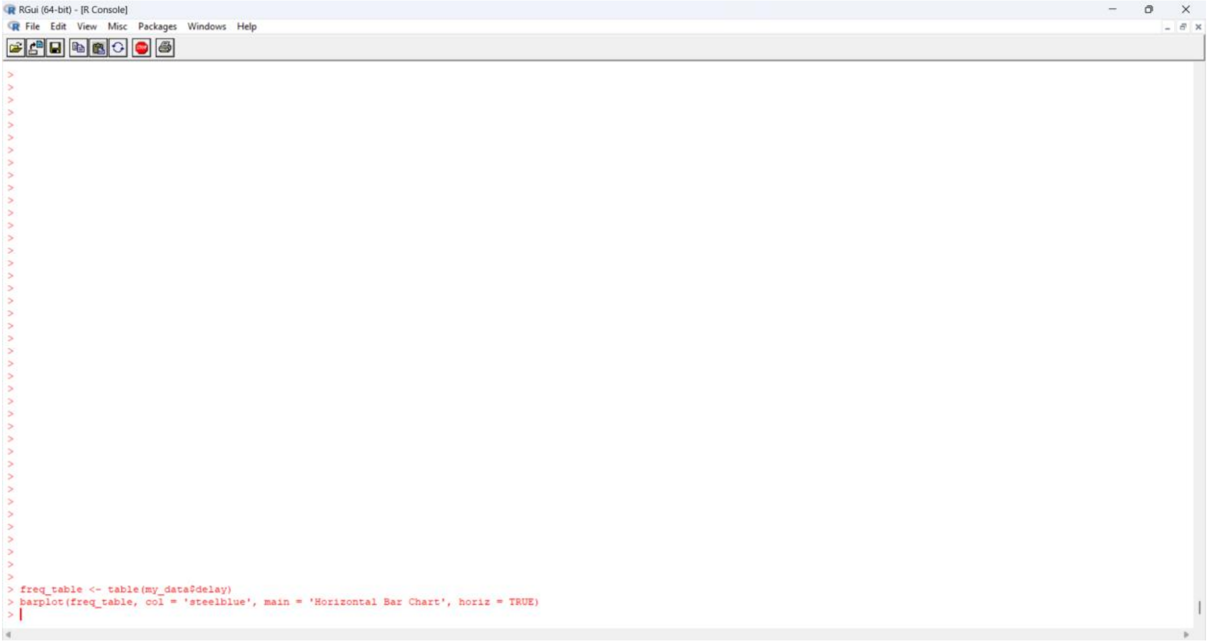
- Plot the box plot to understand how many days in a month flights are delayed by what time



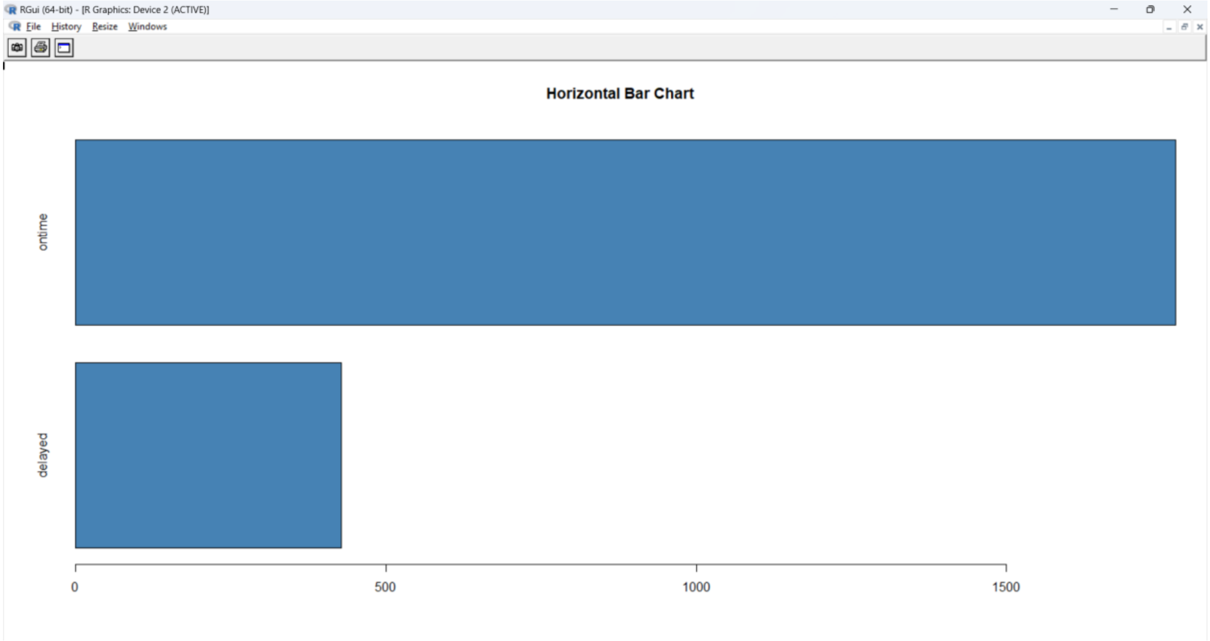
```
> boxplot(my_data$delay == 'delayed', col = 'steelblue', main='boxplot for flight delays', xlab='flight delays')
```

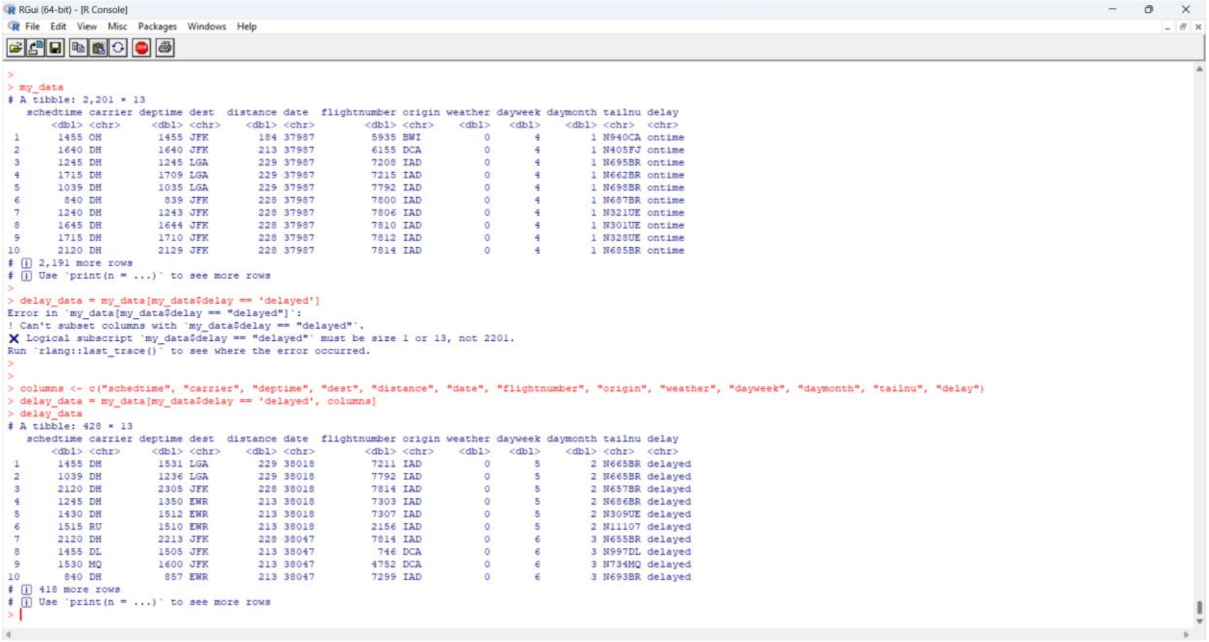

RGui (64-bit) - [R Graphics: Device 2 (ACTIVE)]
File
History
Resize
Windows
Close
boxplot for flight delays
1.0
0.8
0.6
0.4
0.2
0.0
flight delays

- Define the hours of departure
- Create a categorical representation of data using a table
- Redefine the delay variables
- Understand the summary of major variables
- Plot histograms of major variables
- Plot a pie chart to see how many flights were delayed

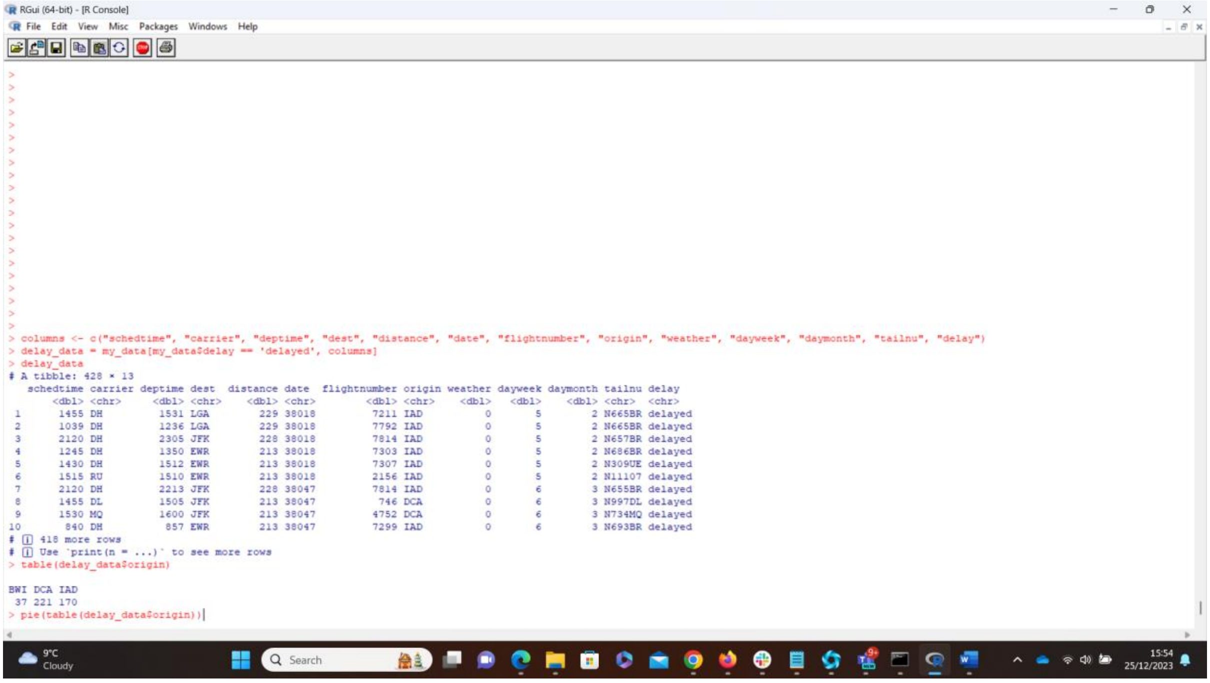


```
>
>
>
>
>
>
>
>
>
>
>
>
>
>
>
>
>
>
>
>
>
>
>
>
>
>
>
>
>
>
>
>
>
>
>
>
>
>
>
>
>
> freq_table <- table(my_data$delay)
> barplot(freq_table, col = 'steelblue', main = 'Horizontal Bar Chart', horiz = TRUE)
> 
```

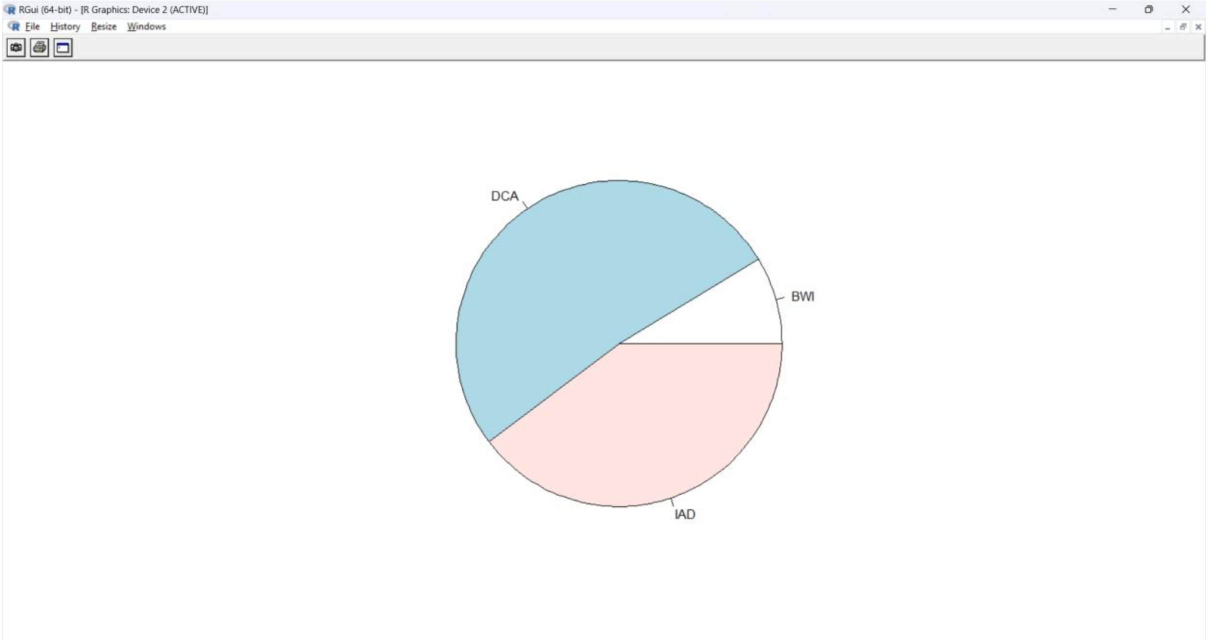



RGui (64-bit) - [R Console]
File Edit View Misc Packages Windows Help

```
>
> my_data
# A tibble: 2,201 × 13
   schedtime carrier deptime dest  distance date  flightnumber origin weather dayweek daymonth tailnu delay
       <dbl> <chr>     <dbl> <chr>    <dbl> <chr>        <dbl> <chr>    <dbl>   <dbl>    <dbl> <chr>  <chr>
 1      1455 OH         1455 JFK        184 37987         5935 BWI          0       4        1 N940CA ontime
 2      1640 DH         1640 JFK        213 37987         6155 DCA          0       4        1 N405FJ ontime
 3      1245 DH         1245 LGA        229 37987         7208 IAD          0       4        1 N695BR ontime
 4      1715 DH         1709 LGA        229 37987         7215 IAD          0       4        1 N662BR ontime
 5      1039 DH         1035 LGA        229 37987         7792 IAD          0       4        1 N698BR ontime
 6       840 DH          839 JFK        228 37987         7800 IAD          0       4        1 N687BR ontime
 7      1240 DH         1243 JFK        228 37987         7806 IAD          0       4        1 N321UE ontime
 8      1645 DH         1644 JFK        228 37987         7810 IAD          0       4        1 N301UE ontime
 9      1715 DH         1710 JFK        228 37987         7812 IAD          0       4        1 N328UE ontime
10      2120 DH         2129 JFK        228 37987         7814 IAD          0       4        1 N685BR ontime
# ℹ 2,191 more rows
# ℹ Use `print(n = ...)` to see more rows
>
> delay_data = my_data[my_data$delay == 'delayed']
Error in `my_data[my_data$delay == "delayed"]`:
! Can't subset columns with `my_data$delay == "delayed"`.
✖ Logical subscript `my_data$delay == "delayed"` must be size 1 or 13, not 2201.
Run `rlang::last_trace()` to see where the error occurred.
>
>
> columns <- c("schedtime", "carrier", "deptime", "dest", "distance", "date", "flightnumber", "origin", "weather", "dayweek", "daymonth", "tailnu", "delay")
> delay_data = my_data[my_data$delay == 'delayed', columns]
> delay_data
# A tibble: 428 × 13
   schedtime carrier deptime dest  distance date  flightnumber origin weather dayweek daymonth tailnu delay
       <dbl> <chr>     <dbl> <chr>    <dbl> <chr>        <dbl> <chr>    <dbl>   <dbl>    <dbl> <chr>  <chr>
 1      1455 DH         1531 LGA        229 38018         7211 IAD          0       5        2 N665BR delayed
 2      1039 DH         1236 LGA        229 38018         7792 IAD          0       5        2 N665BR delayed
 3      2120 DH         2305 JFK        228 38018         7814 IAD          0       5        2 N657BR delayed
 4      1245 DH         1350 EWR        213 38018         7303 IAD          0       5        2 N686BR delayed
 5      1430 DH         1512 EWR        213 38018         7307 IAD          0       5        2 N309UE delayed
 6      1515 RU         1510 EWR        213 38018         2156 IAD          0       5        2 N11107 delayed
 7      2120 DH         2213 JFK        228 38047         7814 IAD          0       6        3 N655BR delayed
 8      1455 DL         1505 JFK        213 38047          746 DCA          0       6        3 N997DL delayed
 9      1530 MQ         1600 JFK        213 38047         4752 DCA          0       6        3 N734MQ delayed
10       840 DH          857 EWR        213 38047         7299 IAD          0       6        3 N693BR delayed
# ℹ 418 more rows
# ℹ Use `print(n = ...)` to see more rows
>
```


RGui (64-bit) - [R Console]
File Edit View Misc Packages Windows Help

```
>
>
>
>
>
>
>
>
>
>
>
>
>
>
>
>
>
>
>
>
>
> columns <- c("schedtime", "carrier", "deptime", "dest", "distance", "date", "flightnumber", "origin", "weather", "dayweek", "daymonth", "tailnu", "delay")
> delay_data = my_data[my_data$delay == 'delayed', columns]
> delay_data
# A tibble: 428 × 13
   schedtime carrier deptime dest  distance date  flightnumber origin weather dayweek daymonth tailnu delay
       <dbl> <chr>     <dbl> <chr>    <dbl> <chr>        <dbl> <chr>    <dbl>   <dbl>    <dbl> <chr>  <chr>
 1      1455 DH         1531 LGA        229 38018         7211 IAD          0       5        2 N665BR delayed
 2      1039 DH         1236 LGA        229 38018         7792 IAD          0       5        2 N665BR delayed
 3      2120 DH         2305 JFK        228 38018         7814 IAD          0       5        2 N657BR delayed
 4      1245 DH         1350 EWR        213 38018         7303 IAD          0       5        2 N686BR delayed
 5      1430 DH         1512 EWR        213 38018         7307 IAD          0       5        2 N309UE delayed
 6      1515 RU         1510 EWR        213 38018         2156 IAD          0       5        2 N11107 delayed
 7      2120 DH         2213 JFK        228 38047         7814 IAD          0       6        3 N655BR delayed
 8      1455 DL         1505 JFK        213 38047          746 DCA          0       6        3 N997DL delayed
 9      1530 MQ         1600 JFK        213 38047         4752 DCA          0       6        3 N734MQ delayed
10       840 DH          857 EWR        213 38047         7299 IAD          0       6        3 N693BR delayed
# ℹ 418 more rows
# ℹ Use `print(n = ...)` to see more rows
> table(delay_data$origin)

BWI DCA IAD 
 37 221 170 
> pie(table(delay_data$origin))
```


RGui (64-bit) - [R Graphics: Device 2 (ACTIVE)]
File
History
Resize
Windows
DCA
BWI
IAD