CSE – 3020

Data Visualization

Lab DA – 1

Name : Anish Desai

Reg. No. : 20BCE0461

Slot : L39 + L40

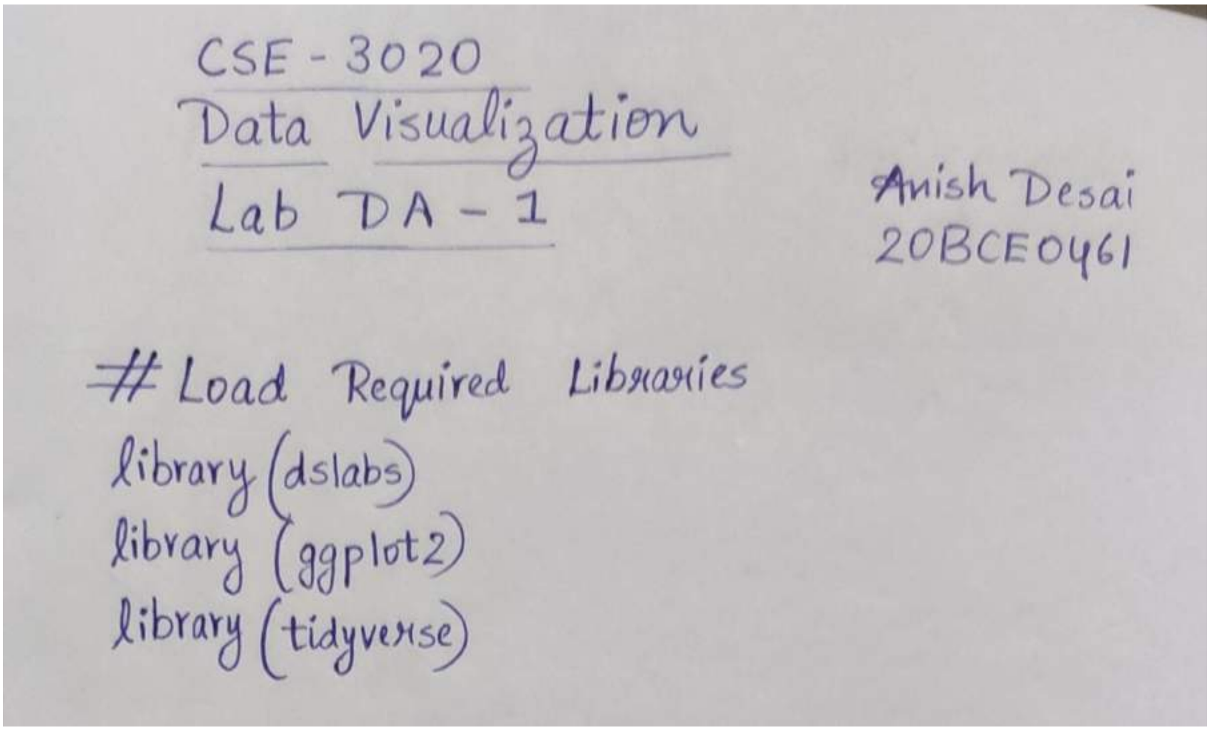
Guided by : Prof. Jyotismita Chaki

VIT®

Vellore Institute of Technology

(Deemed to be University under section 3 of UGC Act, 1956)

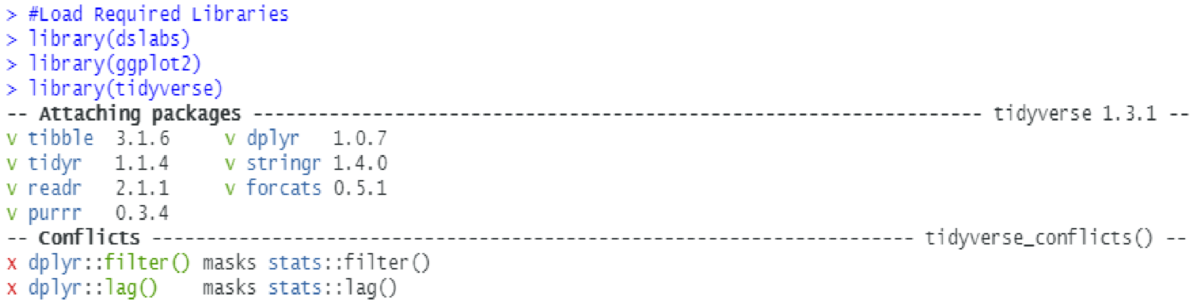
Loading Required Libraries :



CSE - 3020
Data Visualization
Lab DA - 1

Anish Desai
20BCE0461

```
# Load Required Libraries
library(dslabs)
library (ggplot2)
library(tidyverse)
```

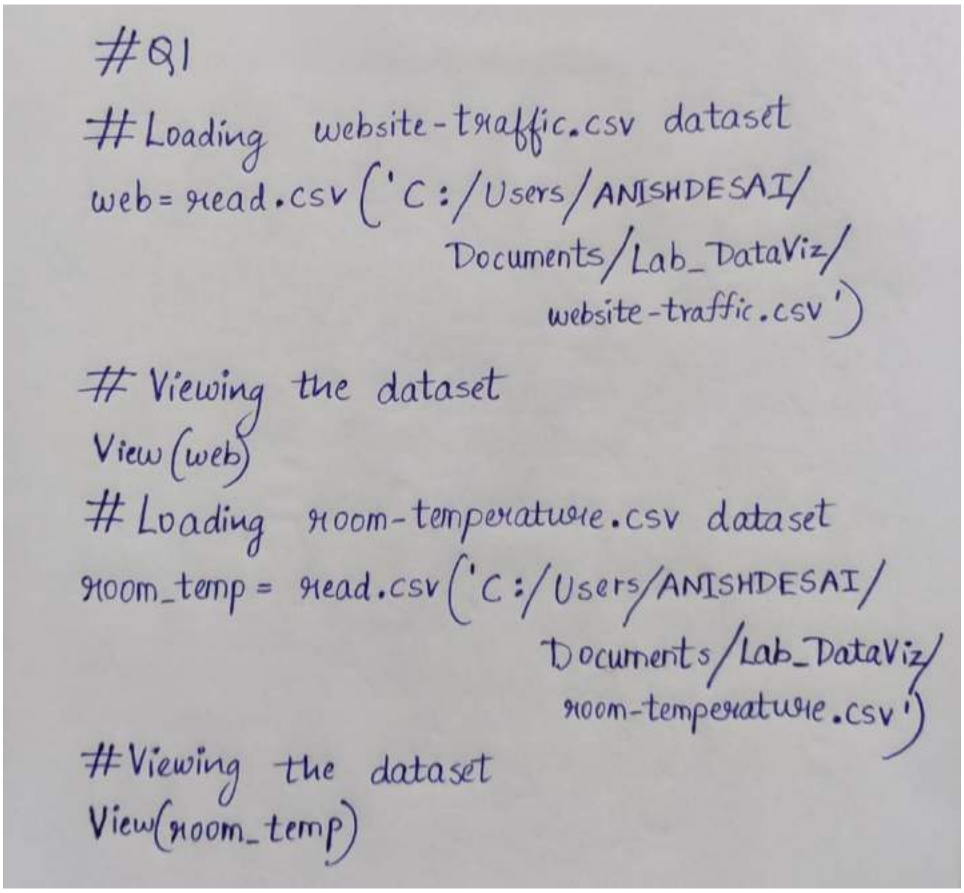


```
> #Load Required Libraries
> library(dslabs)
> library(ggplot2)
> library(tidyverse)
-- Attaching packages --------------------------------------------------------------------- tidyverse 1.3.1 --
v tibble  3.1.6     v dplyr   1.0.7
v tidyr   1.1.4     v stringr 1.4.0
v readr   2.1.1     v forcats 0.5.1
v purrr   0.3.4
-- Conflicts ------------------------------------------------------------------------ tidyverse_conflicts() --
x dplyr::filter() masks stats::filter()
x dplyr::lag()    masks stats::lag()
```

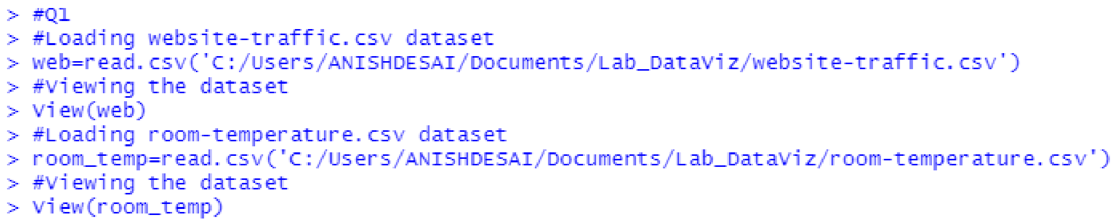
Question 1 :

Load “website-traffic.csv” and “room-temperature.csv” datasets in R (or in similar visualization tool) and view the data.

Code :

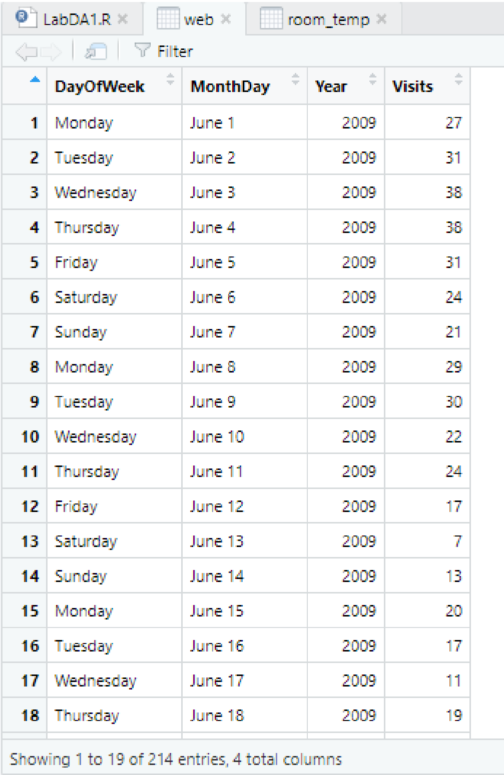


```
#Q1
#Loading website-traffic.csv dataset
web= read.csv('C:/Users/ANISHDESAI/
                       Documents/Lab_DataViz/
                             website-traffic.csv')

# Viewing the dataset
View(web)
#Loading room-temperature.csv dataset
room_temp = read.csv('C:/Users/ANISHDESAI/
                           Documents/Lab_DataViz/
                            room-temperature.csv')

#Viewing the dataset
View(room_temp)
```



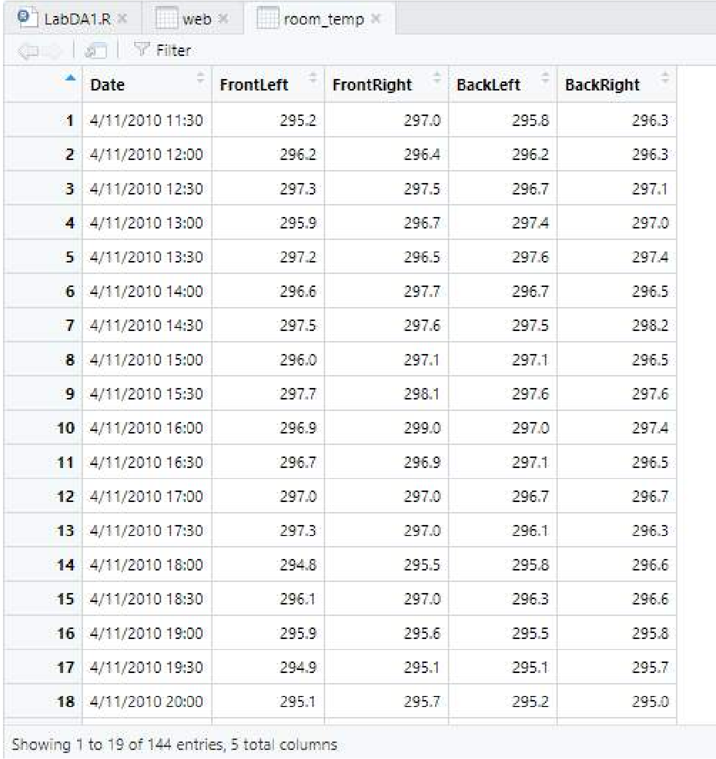
```
> #Q1
> #Loading website-traffic.csv dataset
> web=read.csv('C:/Users/ANISHDESAI/Documents/Lab_DataViz/website-traffic.csv')
> #Viewing the dataset
> View(web)
> #Loading room-temperature.csv dataset
> room_temp=read.csv('C:/Users/ANISHDESAI/Documents/Lab_DataViz/room-temperature.csv')
> #Viewing the dataset
> View(room_temp)
```

Output :



	DayOfWeek	MonthDay	Year	Visits
1	Monday	June 1	2009	27
2	Tuesday	June 2	2009	31
3	Wednesday	June 3	2009	38
4	Thursday	June 4	2009	38
5	Friday	June 5	2009	31
6	Saturday	June 6	2009	24
7	Sunday	June 7	2009	21
8	Monday	June 8	2009	29
9	Tuesday	June 9	2009	30
10	Wednesday	June 10	2009	22
11	Thursday	June 11	2009	24
12	Friday	June 12	2009	17
13	Saturday	June 13	2009	7
14	Sunday	June 14	2009	13
15	Monday	June 15	2009	20
16	Tuesday	June 16	2009	17
17	Wednesday	June 17	2009	11
18	Thursday	June 18	2009	19

Showing 1 to 19 of 214 entries, 4 total columns



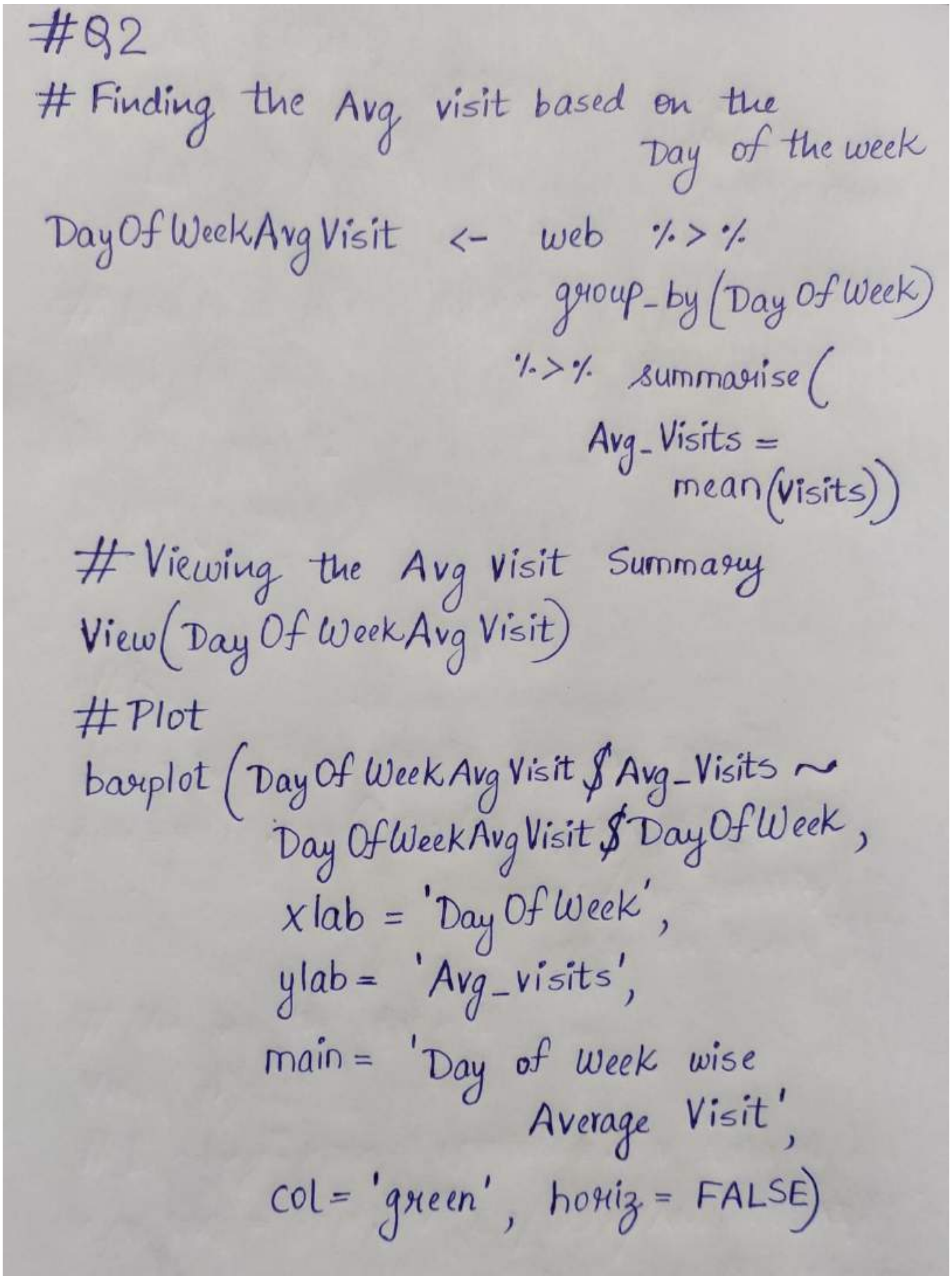
	Date	FrontLeft	FrontRight	BackLeft	BackRight
1	4/11/2010 11:30	295.2	297.0	295.8	296.3
2	4/11/2010 12:00	296.2	296.4	296.2	296.3
3	4/11/2010 12:30	297.3	297.5	296.7	297.1
4	4/11/2010 13:00	295.9	296.7	297.4	297.0
5	4/11/2010 13:30	297.2	296.5	297.6	297.4
6	4/11/2010 14:00	296.6	297.7	296.7	296.5
7	4/11/2010 14:30	297.5	297.6	297.5	298.2
8	4/11/2010 15:00	296.0	297.1	297.1	296.5
9	4/11/2010 15:30	297.7	298.1	297.6	297.6
10	4/11/2010 16:00	296.9	299.0	297.0	297.4
11	4/11/2010 16:30	296.7	296.9	297.1	296.5
12	4/11/2010 17:00	297.0	297.0	296.7	296.7
13	4/11/2010 17:30	297.3	297.0	296.1	296.3
14	4/11/2010 18:00	294.8	295.5	295.8	296.6
15	4/11/2010 18:30	296.1	297.0	296.3	296.6
16	4/11/2010 19:00	295.9	295.6	295.5	295.8
17	4/11/2010 19:30	294.9	295.1	295.1	295.7
18	4/11/2010 20:00	295.1	295.7	295.2	295.0

Showing 1 to 19 of 144 entries, 5 total columns

Question 2 :

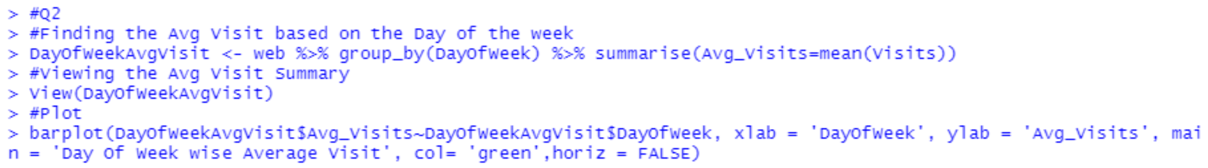
Plot day of week wise average visit. Include proper title, x and y labels. Write the interpretation as well.

Code :

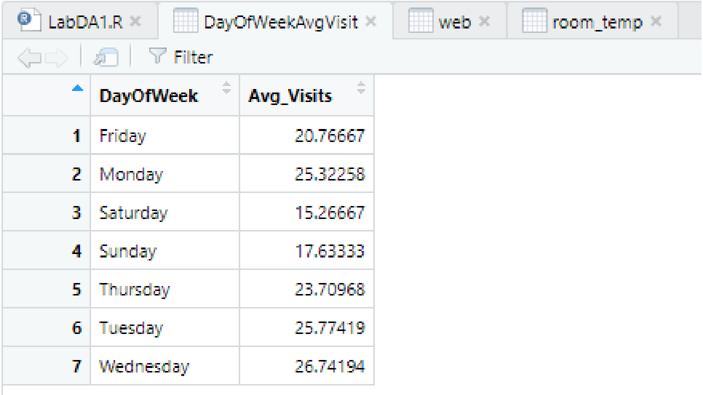


```
#Q2
# Finding the Avg visit based on the
                              Day of the week

DayOfWeekAvgVisit <- web %>%
                         group_by(Day Of Week)
                       %>% summarise(
                            Avg_Visits =
                                mean(visits))
  # Viewing the Avg visit Summary
  View(Day Of WeekAvg Visit)
  #Plot
  barplot(Day Of WeekAvg Visit $ Avg_Visits ~
          DayOfWeekAvgVisit $ DayOfWeek,
          xlab = 'Day Of Week',
          ylab= 'Avg_visits',
          main= 'Day of Week wise
                       Average Visit',
          col= 'green', horiz = FALSE)
```

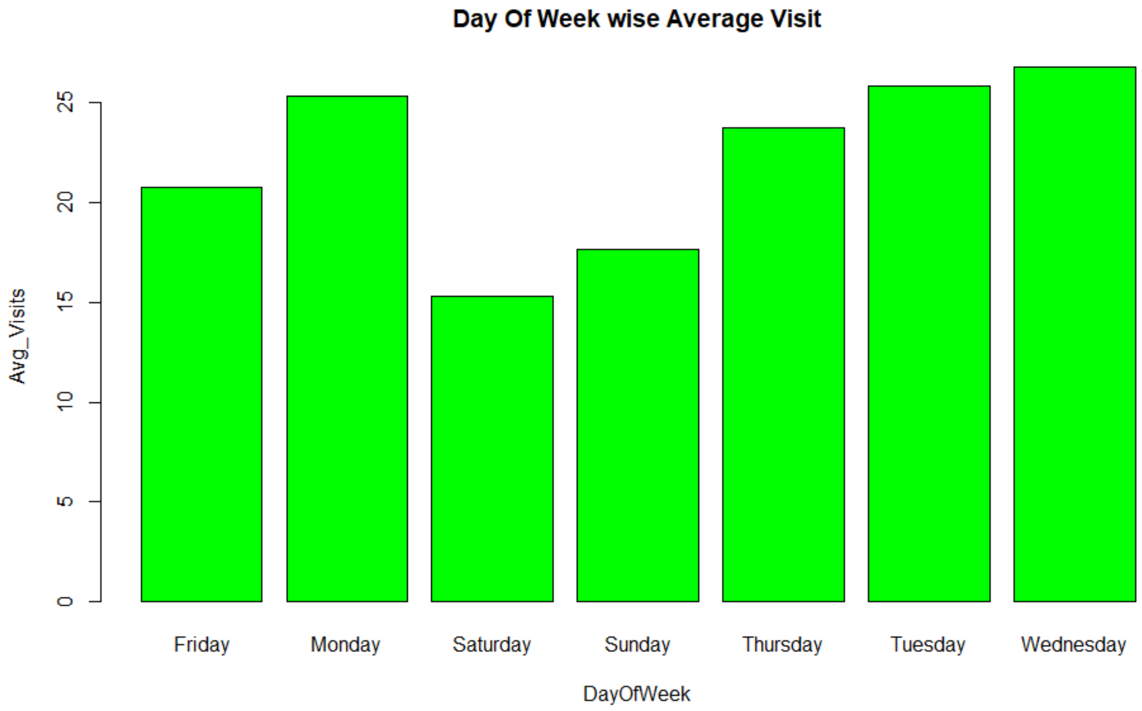
Output :



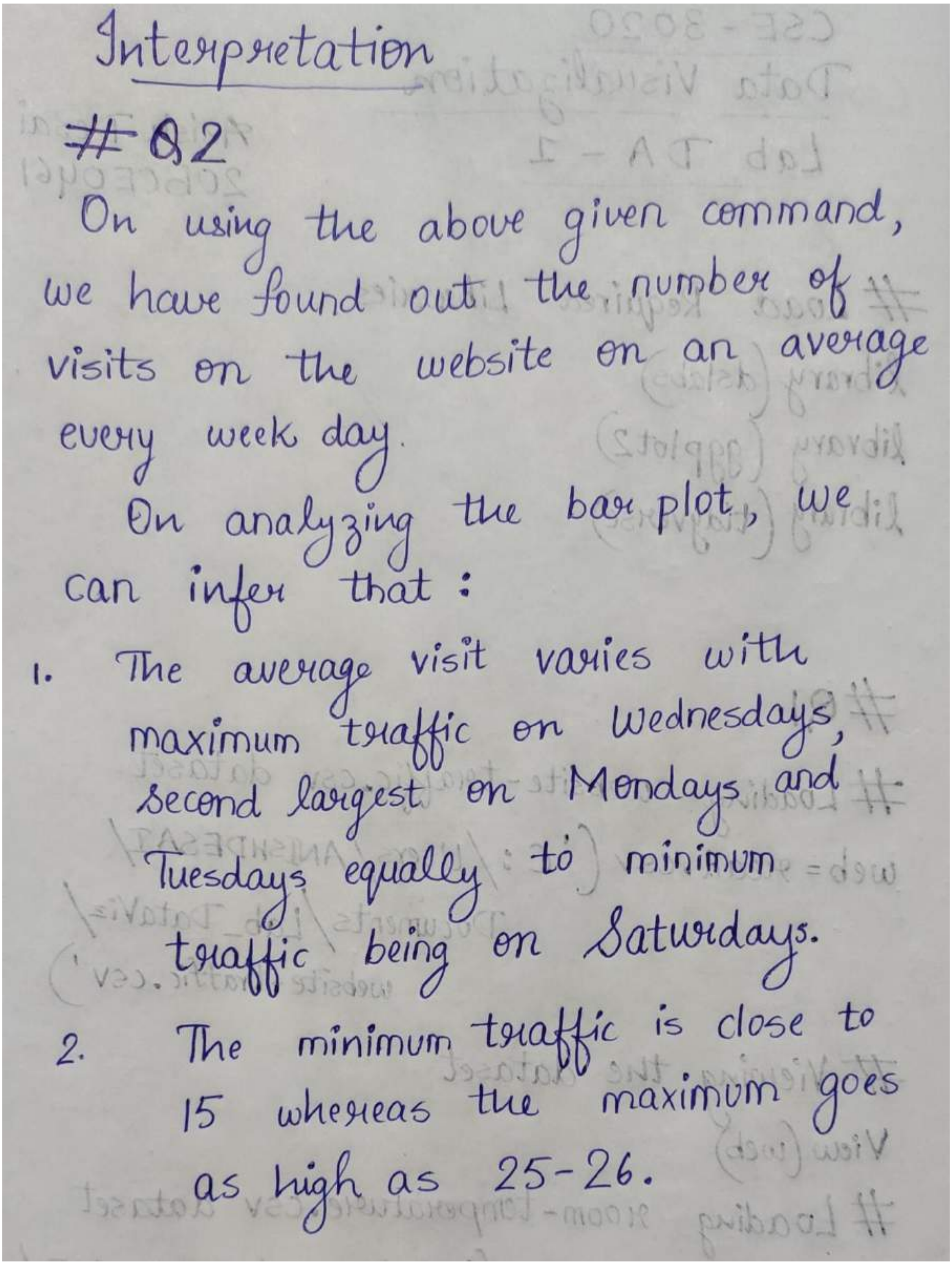
```
> #Q2
> #Finding the Avg Visit based on the Day of the week
> DayOfWeekAvgVisit <- web %>% group_by(DayOfWeek) %>% summarise(Avg_Visits=mean(Visits))
> #Viewing the Avg Visit Summary
> View(DayOfWeekAvgVisit)
> #Plot
> barplot(DayOfWeekAvgVisit$Avg_Visits~DayOfWeekAvgVisit$DayOfWeek, xlab = 'DayOfWeek', ylab = 'Avg_Visits', mai
n = 'Day Of Week wise Average Visit', col= 'green',horiz = FALSE)
```



	DayOfWeek	Avg_Visits
1	Friday	20.76667
2	Monday	25.32258
3	Saturday	15.26667
4	Sunday	17.63333
5	Thursday	23.70968
6	Tuesday	25.77419
7	Wednesday	26.74194



Interpretation :



Interpretation

#Q2

On using the above given command, we have found out the number of visits on the website on an average every week day.

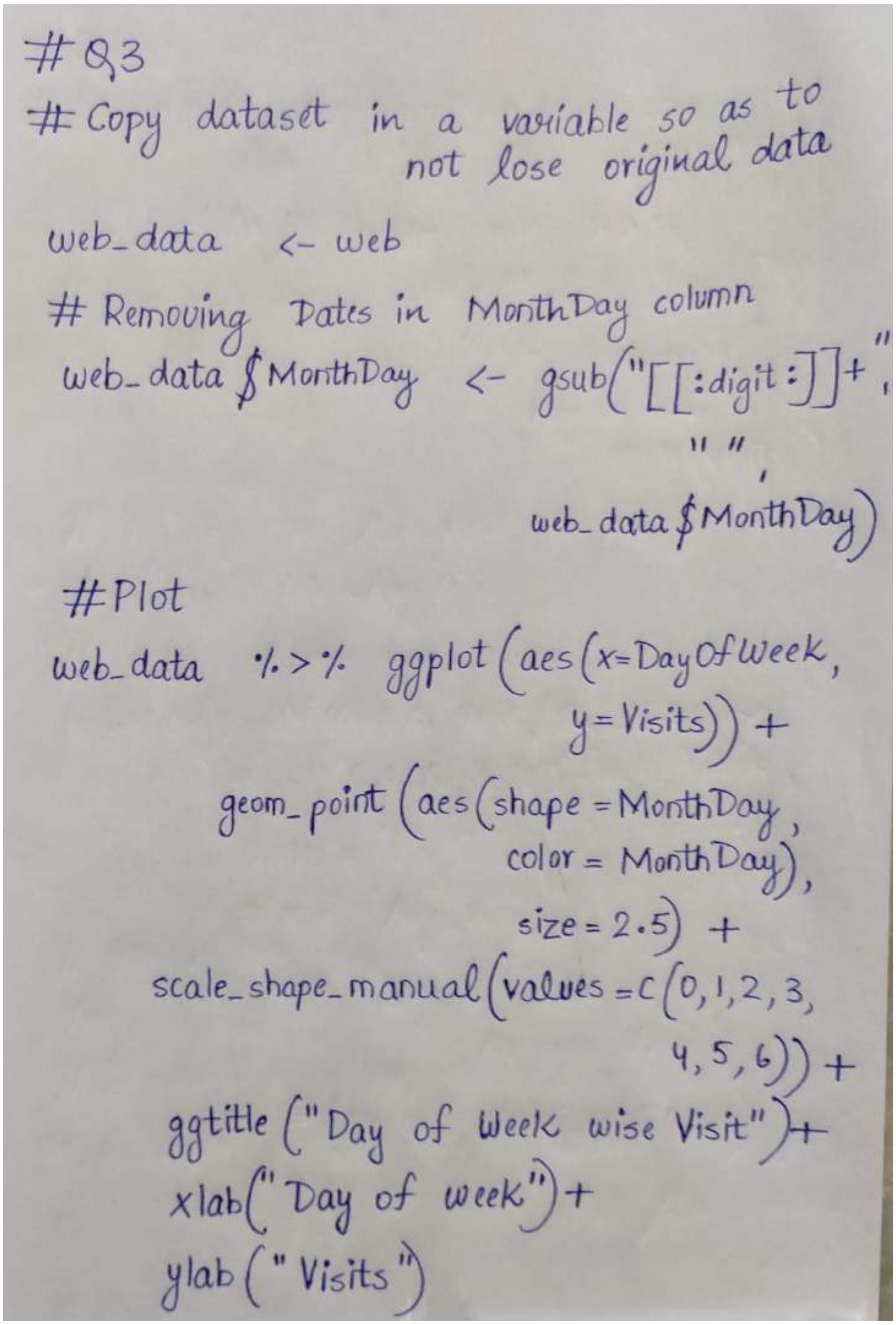
On analyzing the bar plot, we can infer that :

1. The average visit varies with maximum traffic on Wednesdays, second largest on Mondays and Tuesdays equally to minimum traffic being on Saturdays.
2. The minimum traffic is close to 15 whereas the maximum goes as high as 25-26.

Question 3 :

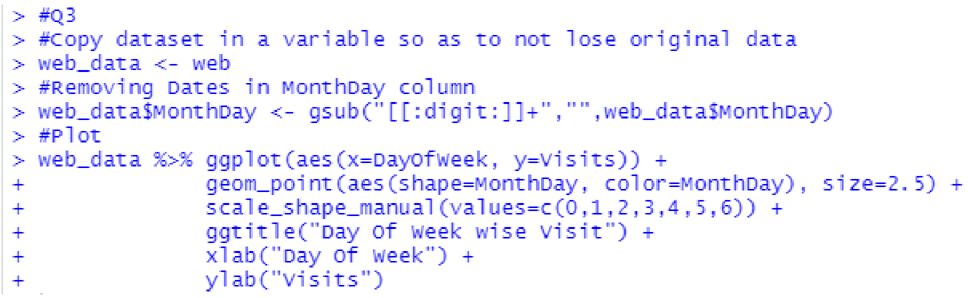
Plot day of week wise visit and include monthday data as color and shape feature. Write the interpretation as well. [hint: use ggplot]

Code :

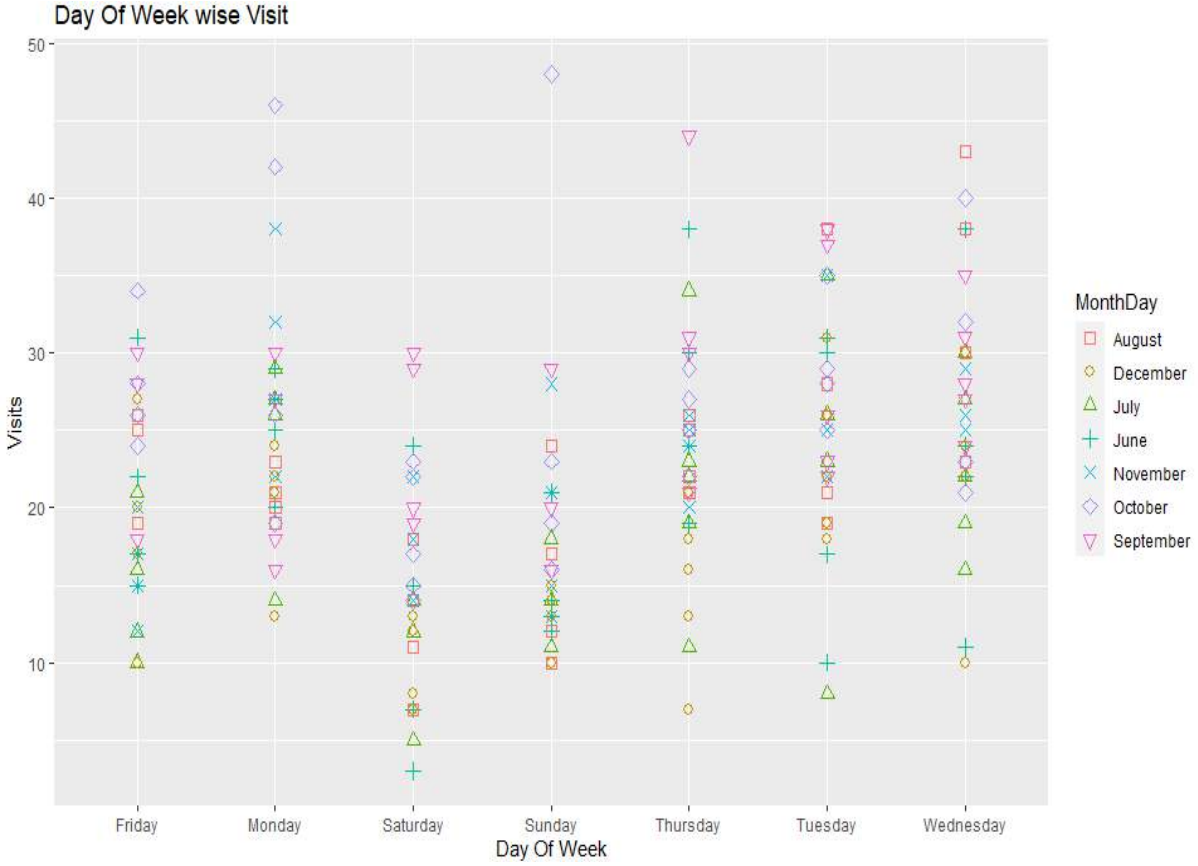


```
#Q3
# Copy dataset in a variable so as to
                    not lose original data

web_data  <- web

# Removing Dates in MonthDay column
web_data$MonthDay  <- gsub("[[:digit:]]+",
                          " ",
                   web_data$MonthDay)

#Plot
web_data  %>%  ggplot(aes(x=DayOfWeek,
                           y=Visits)) +
          geom_point(aes(shape =MonthDay,
                        color = MonthDay),
                        size = 2.5) +
      scale_shape_manual(values =c(0,1,2,3,
                                  4,5,6)) +
      ggtitle("Day of Week wise Visit")+
      xlab("Day of week")+
      ylab(" Visits")
```

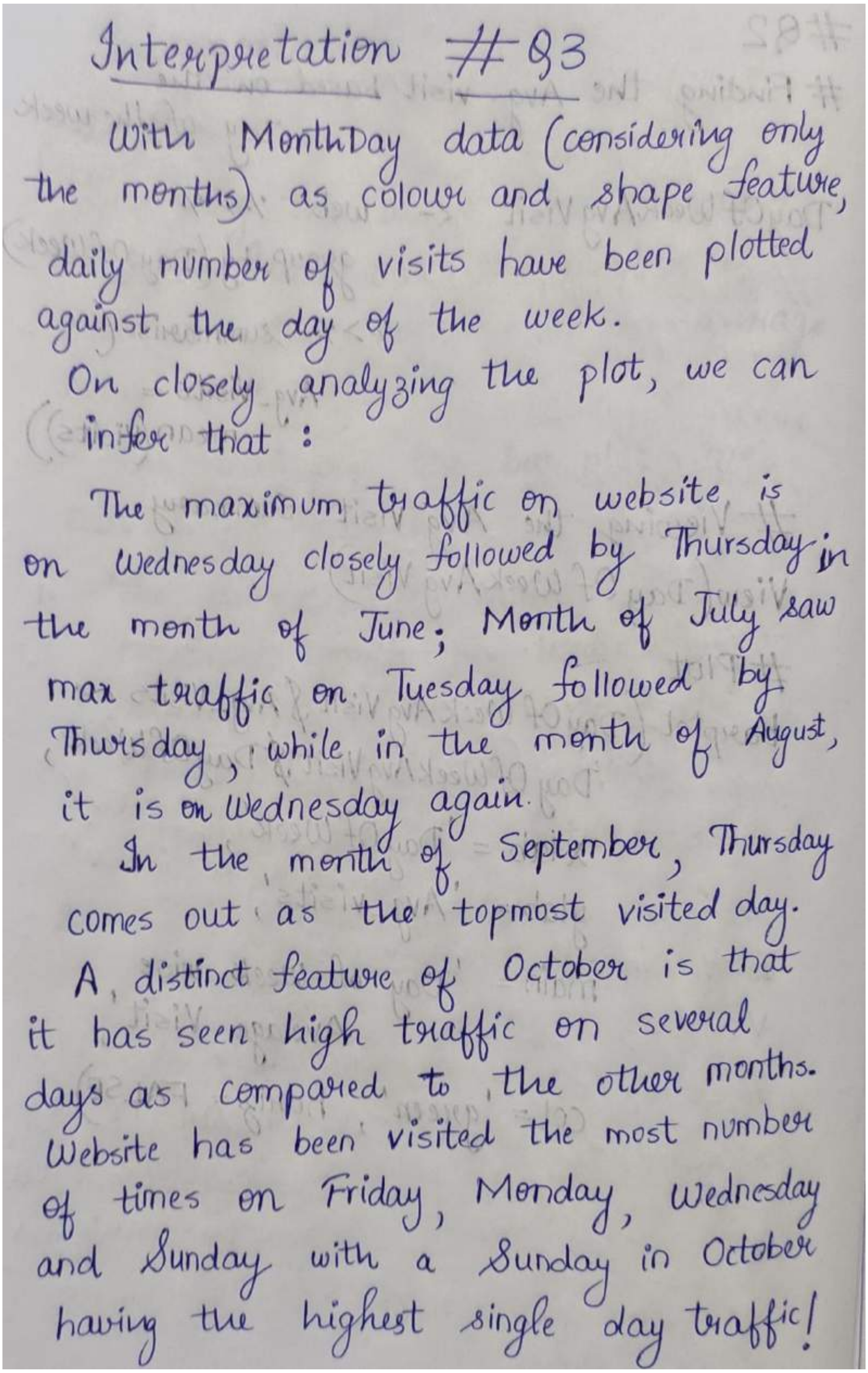
Output :



```
> #Q3
> #Copy dataset in a variable so as to not lose original data
> web_data <- web
> #Removing Dates in MonthDay column
> web_data$MonthDay <- gsub("[[:digit:]]+","",web_data$MonthDay)
> #Plot
> web_data %>% ggplot(aes(x=DayOfWeek, y=Visits)) +
+              geom_point(aes(shape=MonthDay, color=MonthDay), size=2.5) +
+              scale_shape_manual(values=c(0,1,2,3,4,5,6)) +
+              ggtitle("Day Of Week wise Visit") +
+              xlab("Day Of Week") +
+              ylab("Visits")
```



Interpretation :



Interpretation #Q3

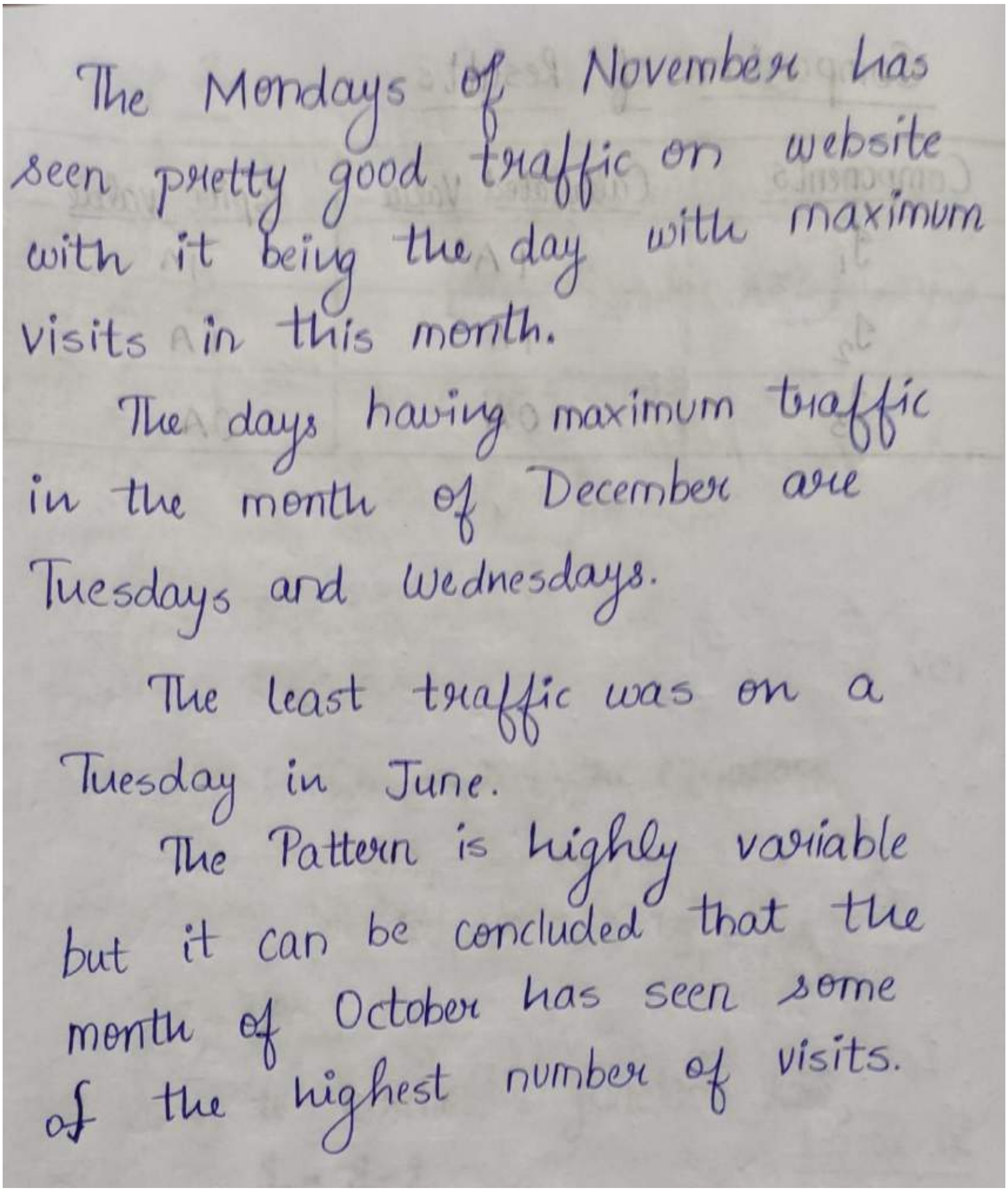
With MonthDay data (considering only the months) as colour and shape feature, daily number of visits have been plotted against the day of the week.

On closely analyzing the plot, we can infer that :

The maximum traffic on website, is on Wednesday closely followed by Thursday in the month of June; Month of July saw max traffic on Tuesday followed by Thursday, while in the month of August, it is on Wednesday again.

In the month of September, Thursday comes out as the topmost visited day.

A distinct feature of October is that it has seen high traffic on several days as compared to the other months. Website has been visited the most number of times on Friday, Monday, Wednesday and Sunday with a Sunday in October having the highest single day traffic!



The Mondays of November has seen pretty good traffic on website with it being the day with maximum visits in this month.

The days having maximum traffic in the month of December are Tuesdays and Wednesdays.

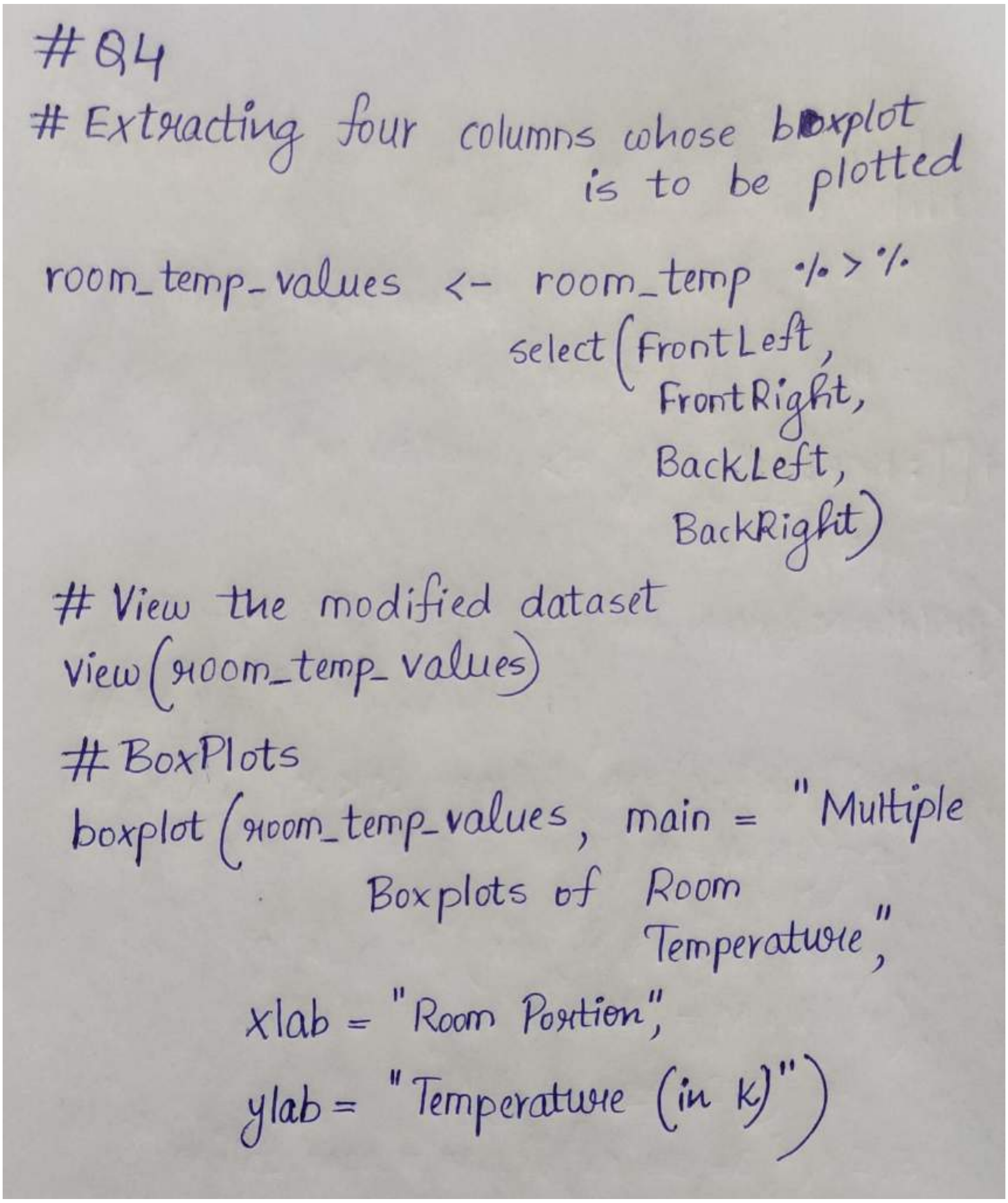
The least traffic was on a Tuesday in June.

The Pattern is highly variable but it can be concluded that the month of October has seen some of the highest number of visits.

Question 4 :

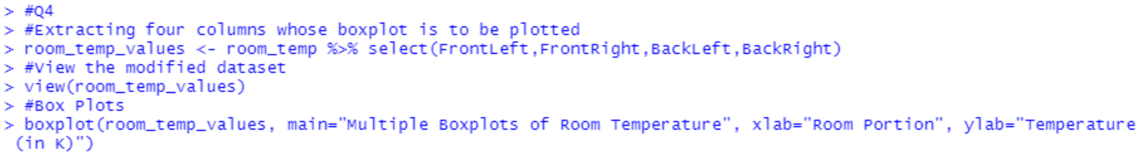
In a single representation, Boxplot the FrontLeft, FrontRight, BackLeft and BackRight data from the room-temperature dataset. Also write the interpretation.

Code :

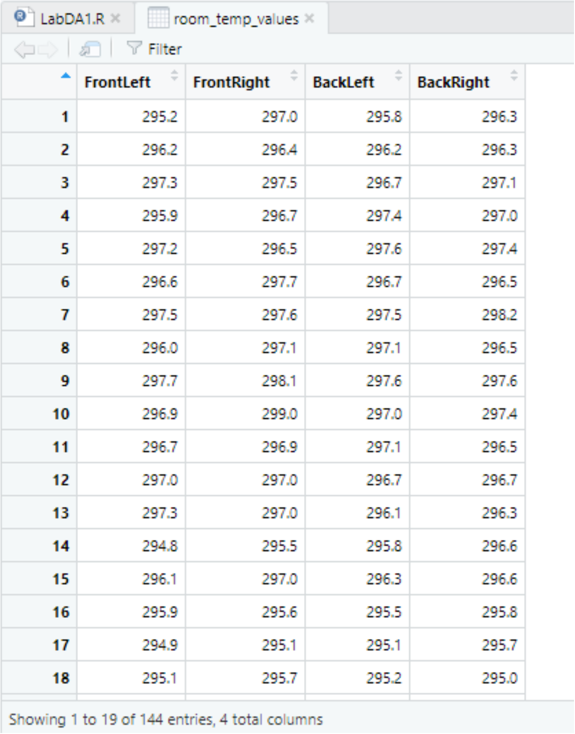


```
#Q4
# Extracting four columns whose boxplot
                              is to be plotted

room_temp_values <- room_temp %>%
                    select(FrontLeft,
                           FrontRight,
                           BackLeft,
                           BackRight)
# View the modified dataset
View(room_temp_values)
# BoxPlots
boxplot(room_temp_values, main = "Multiple
                Boxplots of Room
                            Temperature",
        xlab = "Room Portion",
        ylab = "Temperature (in K)")
```

Output :

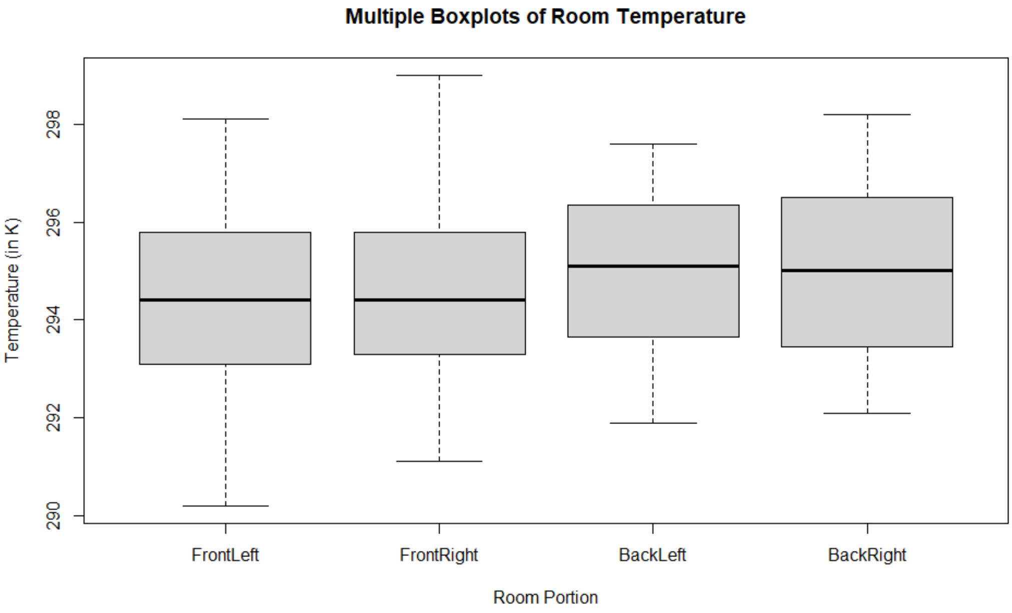


```
> #Q4
> #Extracting four columns whose boxplot is to be plotted
> room_temp_values <- room_temp %>% select(FrontLeft,FrontRight,BackLeft,BackRight)
> #View the modified dataset
> view(room_temp_values)
> #Box Plots
> boxplot(room_temp_values, main="Multiple Boxplots of Room Temperature", xlab="Room Portion", ylab="Temperature
 (in K)")
```

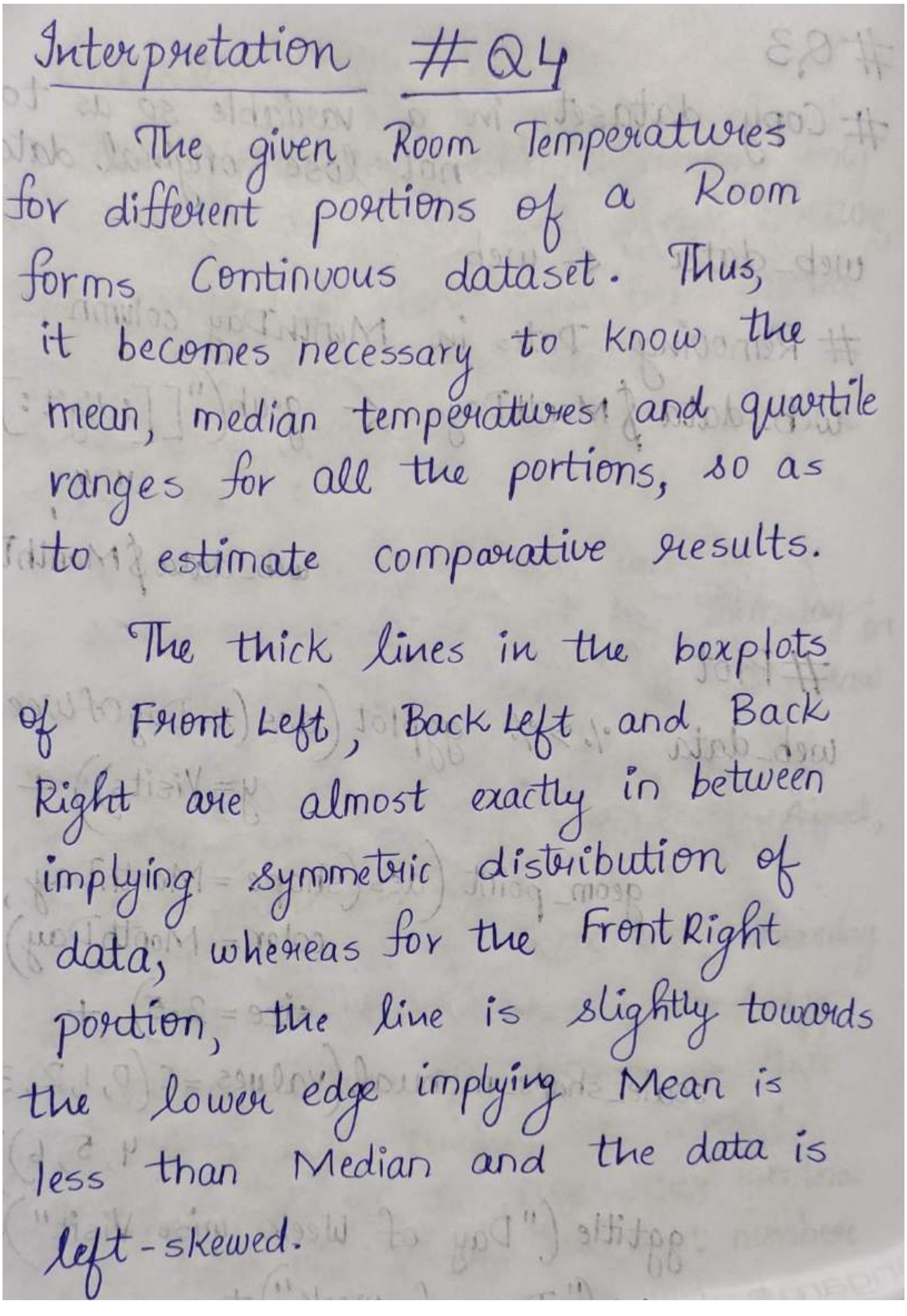


	FrontLeft	FrontRight	BackLeft	BackRight
1	295.2	297.0	295.8	296.3
2	296.2	296.4	296.2	296.3
3	297.3	297.5	296.7	297.1
4	295.9	296.7	297.4	297.0
5	297.2	296.5	297.6	297.4
6	296.6	297.7	296.7	296.5
7	297.5	297.6	297.5	298.2
8	296.0	297.1	297.1	296.5
9	297.7	298.1	297.6	297.6
10	296.9	299.0	297.0	297.4
11	296.7	296.9	297.1	296.5
12	297.0	297.0	296.7	296.7
13	297.3	297.0	296.1	296.3
14	294.8	295.5	295.8	296.6
15	296.1	297.0	296.3	296.6
16	295.9	295.6	295.5	295.8
17	294.9	295.1	295.1	295.7
18	295.1	295.7	295.2	295.0

Showing 1 to 19 of 144 entries, 4 total columns



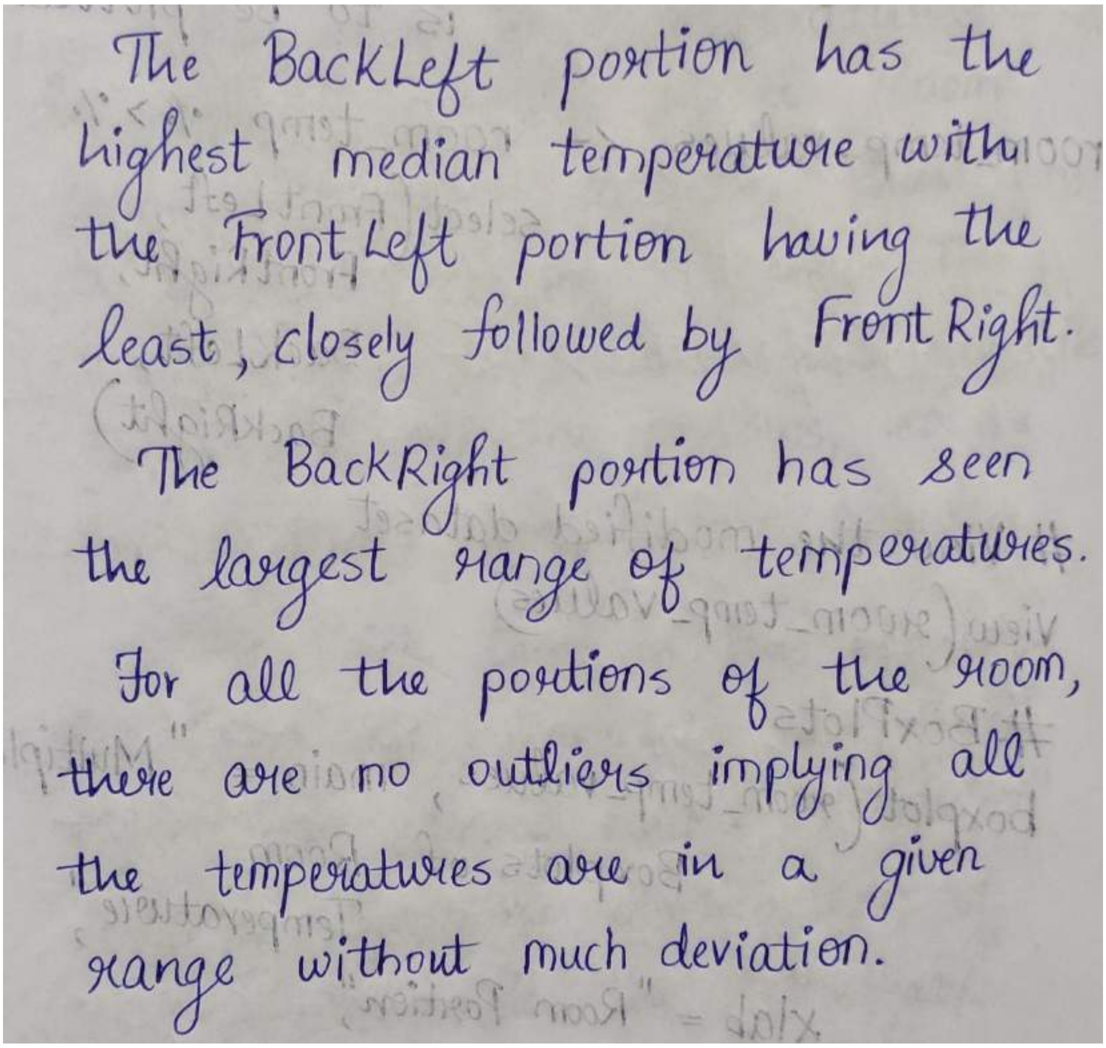
Interpretation :



Interpretation #Q4

The given Room Temperatures for different portions of a Room forms Continuous dataset. Thus, it becomes necessary to know the mean, median temperatures and quartile ranges for all the portions, so as to estimate comparative results.

The thick lines in the boxplots of Front Left, Back Left and Back Right are almost exactly in between implying symmetric distribution of data, whereas for the Front Right portion, the line is slightly towards the lower edge implying Mean is less than Median and the data is left-skewed.



The BackLeft portion has the highest median temperature with the Front Left portion having the least, closely followed by Front Right.

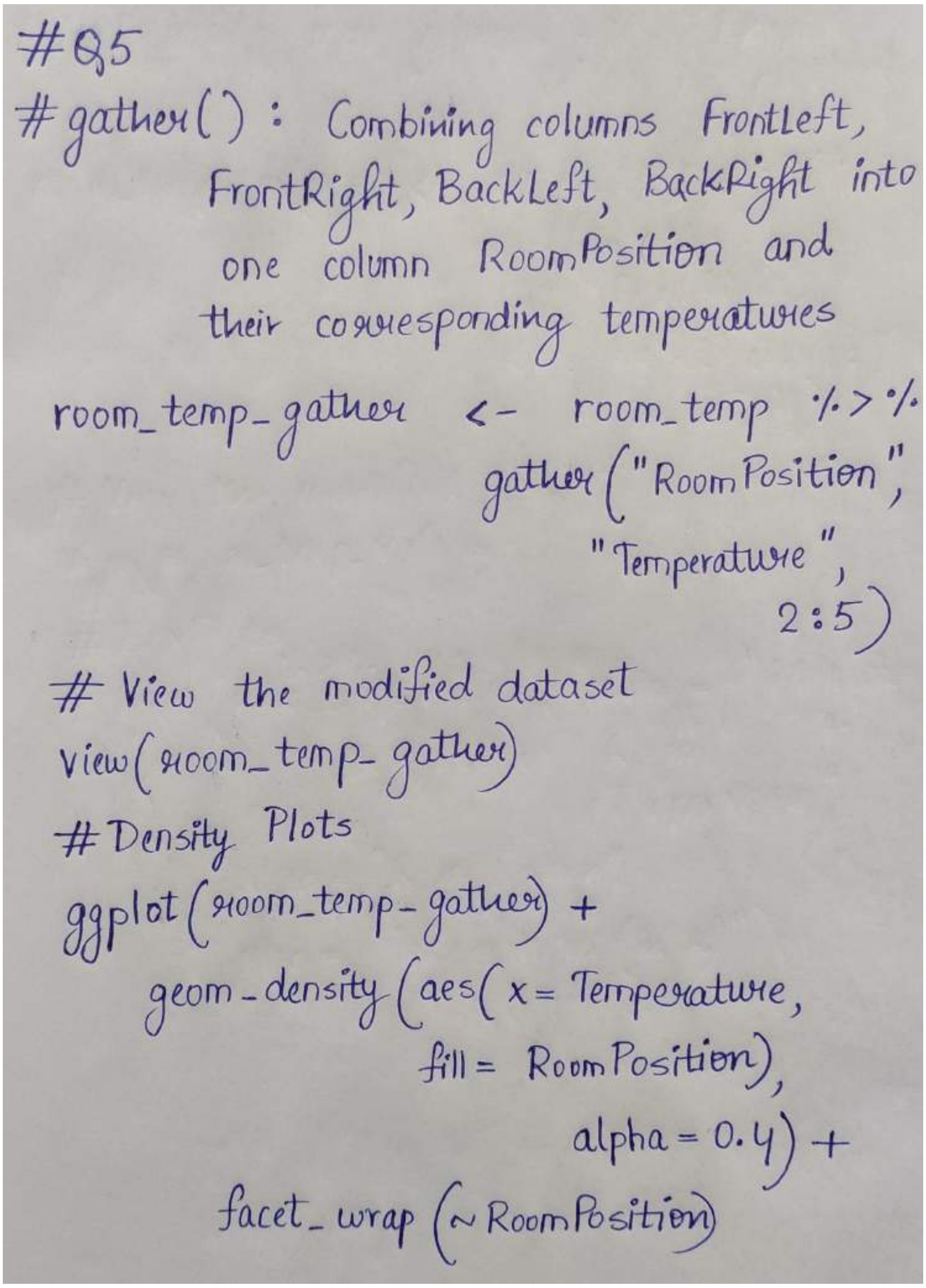
The BackRight portion has seen the largest range of temperatures.

For all the portions of the room, there are no outliers implying all the temperatures are in a given range without much deviation.

Question 5 :

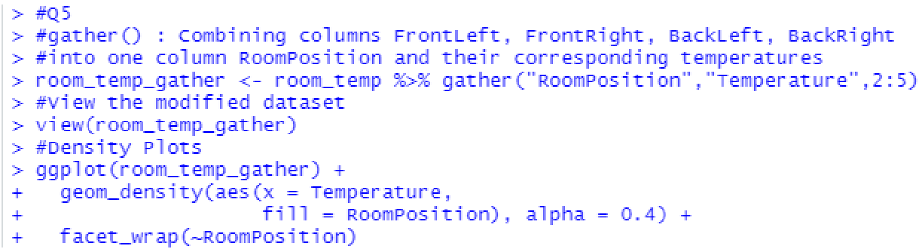
Using density plot, plot the FrontLeft, FrontRight, BackLeft and BackRight data from the room-temperature dataset. In a single representation, four separate plots(density) should be there. Also write the interpretation. [hint: use facet_wrap() function].

Code :

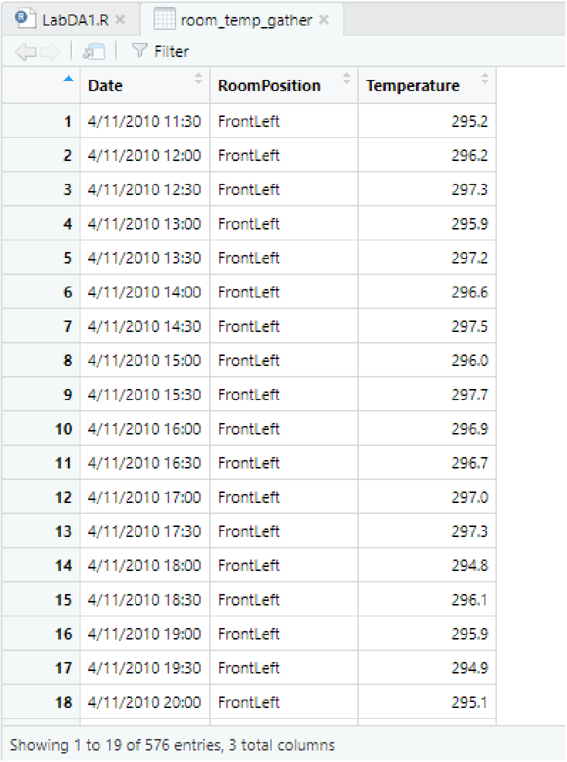


```
#Q5
#gather() : Combining columns FrontLeft,
         FrontRight, BackLeft, BackRight into
         one column RoomPosition and
         their corresponding temperatures

room_temp-gather <- room_temp %>%
                    gather("RoomPosition",
                           "Temperature",
                                  2:5)

# View the modified dataset
view(room_temp-gather)
#Density Plots
ggplot(room_temp-gather) +
   geom-density(aes(x= Temperature,
                   fill= RoomPosition),
                         alpha= 0.4) +
      facet_wrap (~RoomPosition)
```

Output :

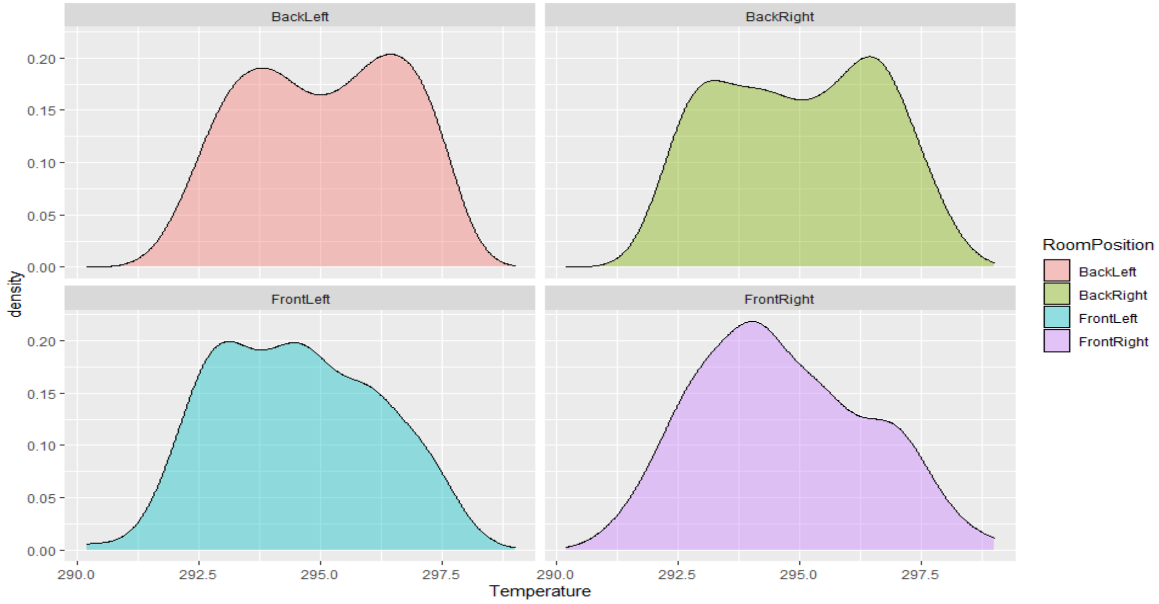


```
> #Q5
> #gather() : Combining columns FrontLeft, FrontRight, BackLeft, BackRight
> #into one column RoomPosition and their corresponding temperatures
> room_temp_gather <- room_temp %>% gather("RoomPosition","Temperature",2:5)
> #View the modified dataset
> view(room_temp_gather)
> #Density Plots
> ggplot(room_temp_gather) +
+   geom_density(aes(x = Temperature,
+                    fill = RoomPosition), alpha = 0.4) +
+   facet_wrap(~RoomPosition)
```

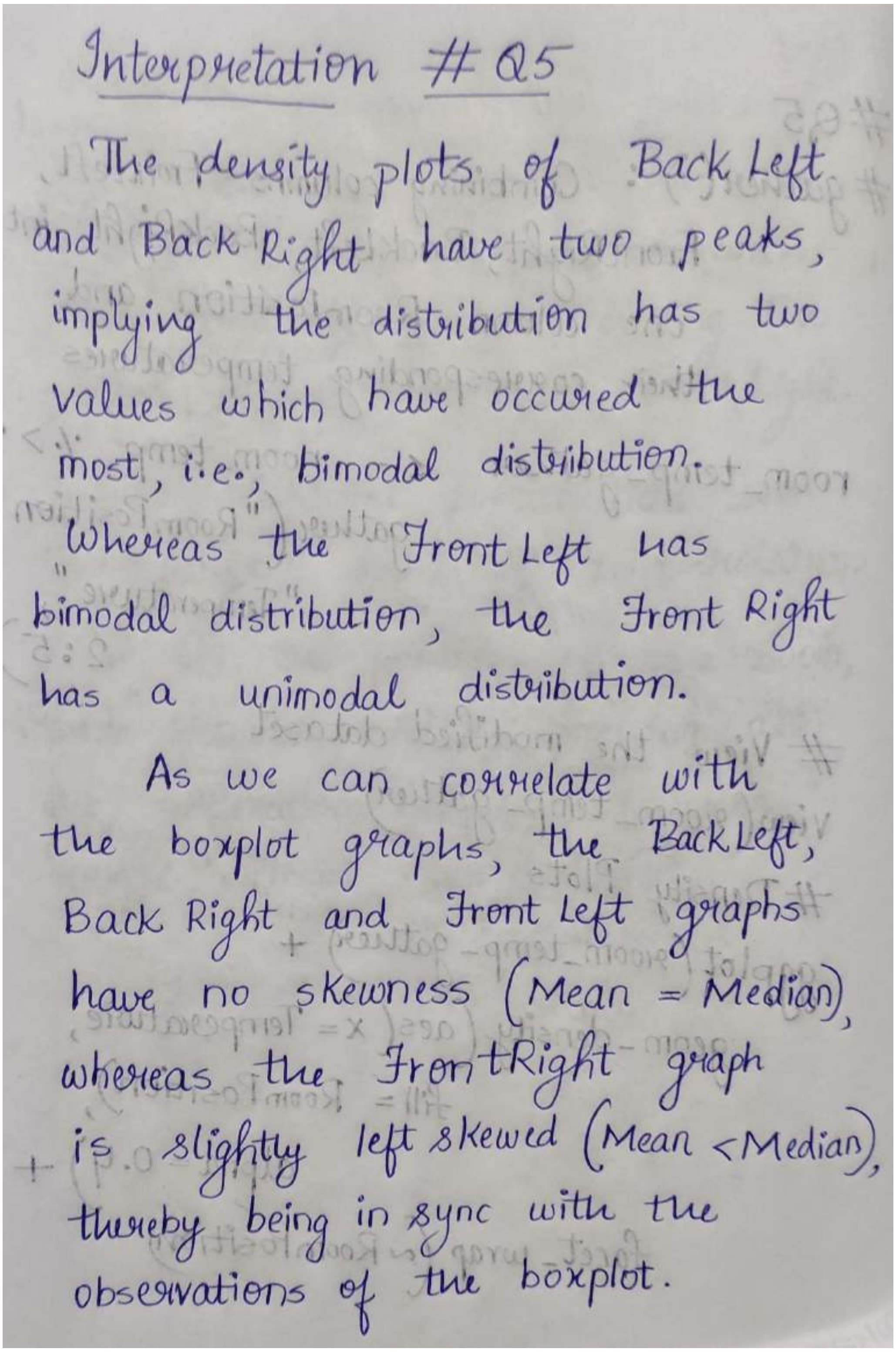


	Date	RoomPosition	Temperature
1	4/11/2010 11:30	FrontLeft	295.2
2	4/11/2010 12:00	FrontLeft	296.2
3	4/11/2010 12:30	FrontLeft	297.3
4	4/11/2010 13:00	FrontLeft	295.9
5	4/11/2010 13:30	FrontLeft	297.2
6	4/11/2010 14:00	FrontLeft	296.6
7	4/11/2010 14:30	FrontLeft	297.5
8	4/11/2010 15:00	FrontLeft	296.0
9	4/11/2010 15:30	FrontLeft	297.7
10	4/11/2010 16:00	FrontLeft	296.9
11	4/11/2010 16:30	FrontLeft	296.7
12	4/11/2010 17:00	FrontLeft	297.0
13	4/11/2010 17:30	FrontLeft	297.3
14	4/11/2010 18:00	FrontLeft	294.8
15	4/11/2010 18:30	FrontLeft	296.1
16	4/11/2010 19:00	FrontLeft	295.9
17	4/11/2010 19:30	FrontLeft	294.9
18	4/11/2010 20:00	FrontLeft	295.1

Showing 1 to 19 of 576 entries, 3 total columns



Interpretation :



Interpretation #Q5

The density plots of Back Left and Back Right have two peaks, implying the distribution has two values which have occured the most, i.e., bimodal distribution.

Whereas the Front Left has bimodal distribution, the Front Right has a unimodal distribution.

As we can correlate with the boxplot graphs, the Back Left, Back Right and Front Left graphs have no skewness (Mean = Median), whereas the FrontRight graph is slightly left skewed (Mean < Median), thereby being in sync with the observations of the boxplot.

-------------------------------------**Thank you**---