Northumbria University
NEWCASTLE

Department of Computer & Information Sciences

STATISTICAL PROGRAMMING QUESTIONS

MODULE TITLE: STATISTICAL PROGRAMMING

MODULE CODE: KL7012

MODULE TUTOR: Dr. Naveed Anwar

NAME: Luke Chugh

COURSE: MSc in Data Science

STUDENT ID: w21054410

Answer 1)

Exercise Classes verses Gym-only Workouts

Introduction:

It is very crucial for any business owner or organization to implement and understand, which kind of practices or approaches can lead it towards maximizing its profit using suitable statistical techniques. The purpose of the following report is to perform descriptive analysis by summarizing the interpretations and findings from the given statistics of the data which was collected over the period of six months, to conclude, which type of exercise, namely “gym only workouts” or “exercise classes” was the most effective in helping individuals to lose weight.

Findings and discussion:

Having a first look at the metrics, “Exercise classes” group had 44 participants with mean weight loss of 1.8 Kgs over the period of six months, mode weight loss of 1.5 Kgs over the period of 6 months and a standard deviation of 1.03. Whereas, “Gym-only Workouts” group had 60 participants with mean weight loss of 2.5 Kgs over the period of six months, mode weight loss of 1.7 Kgs over the period of six months and standard deviation of 1.33. At a first glance, it seems that “ Gym-Only workouts” gave better results than “Exercise Classes” as the mean weight loss is greater which is 2.5 Kilograms. Since the mean is greater than mode for both the categories we can say for sure that the distribution is positively skewed and therefore we can use the relation 3 Median – 2 Mean = Mode to find the Median which comes out to be 2.234 Kg and 1.7 Kg for Gym-Only workouts and Exercise classes respectively. Since the distribution is skewed, median is more preferred measure of central tendency than the mean. The higher median for Gym-Only workouts again indicates that it is better. But if we look at the most repeated amounts of weight loss i.e. modes of these two categories (1.5 Kgs vs 1.7 Kgs) there is not much of a difference which might indicate that the higher mean in the “Gym-Only workouts” category might have been affected by outliers. Standard deviation of 1.03 in “ Exercise – Class “ category indicates that there is less variation in the weight loss of the people included in this group. On the other hand the Standard Deviation of 1.33 for the “ Gym-Only “ category indicates much more variation in the weight loss of the people in this group which means that some people would have lost much more weight and some lost much less.

The fact that was ignored all this while was that there are 60 people in “ Gym-Only “ workouts group, whereas there were only 44 people participating in the “ Exercise – Class “ group which brings us back to elementary statistics that the larger is the sample size, lesser should be the standard deviation of the means (since standard deviation is inversely proportional to the square root of sample size) which surprisingly is the complete opposite of what was observed in the observations because the “ Gym-Only workouts“ category had more data points so it should have had less standard deviation but on the contrary it had more standard deviation. This means that the data points were much far from each other in the “Gym-Only workouts” group as compared to the spread of data points in the “ Exercise – Class “ group; which means that if a sample is picked at random from the confidence interval of both these categories then one can’t be certain about which sample would have better results.

Conclusion:

In conclusion, the given tabular records do not suffice to conclude which exercise is most effective for losing weight; it seems that one has to look at the sample to see if that group which the sample belongs to, has people which had higher weight loss goals which lead to better weight loss results for people in that particular group.

Appendix for answer 1:

	Exercise class	Gym-only workouts
Participants	44	60
Mean weight loss over 6 months	1.8 kgs	2.5 kgs
Mode weight loss over 6 months	1.5 kgs	1.7 kgs
Standard deviation	1.03	1.33

Answer 2)

One of the ways of dealing with missing values while processing the data is to impute the missing values with mean or the median of the values of that particular column. In general, when data is normally distributed we should impute the values with the mean, whereas for skewed distributions missing values should be imputed with the median.

Positives:

- Imputing the mean ensures that the mean of the variable is unchanged. When the average of data in each variable is the only metric we are concerned with, mean substitution might turn out be a good approach
- Imputing with mean does not require us to delete the corresponding rows or columns which means that it does not reduces the sample size as the size of our dataset is same as the original.

Negatives:

- Works only with numerical continuous variables and not for categorical data
- Can cause data leakage which refers to accidentally sharing information between test and training datasets.
- Since we are imputing all the missing values with the mean for a particular variable, we would be very certain about the correctness of our mean which means that the standard error of the mean and the variance of the imputed variables are now biased which will result in narrowed confidence interval around the point of estimation of our mean
- Since most research studies revolve around finding the relationship between the variables, mean imputation is not a good approach, as mean imputation leads to bias in correlation and regression coefficients.

Answer 3) Part a)

Code:

```
data = data.frame(read.table(file = "Data for Question 3.txt", skipNul = T, header = TRUE, sep = " "))

attach(data)
```

# By attaching it to the data frame we can access the variables present in the data framework without calling the data frame

Answer 3) Part b)

```
summary <- as.data.frame(apply(data,2,summary))

View(summary)
```

```
install.packages('writexl')
library('writexl')
write_xlsx(summary,"summary.xlsx")
```

Output:

	age	sex	height	weight	bmp	fev1	rv	frc	tlc	pemax
Min	7	0	109	12.9	64	18	158	104	81	65
1st Qu.	11	0	139	25.1	68	26	188	127	101	85
Median	14	0	156	37.2	71	33	225	139	113	95
Mean	14.48	0.44	152.8	38.404	78.28	34.72	255.2	155.4	114	109.12
3rd Qu.	17	1	174	51.1	90	44	305	183	128	130
Max.	23	1	180	73.8	97	57	449	268	147	195

Code:

```
male <- length(which(sex == 0))
message('No. of males = ', male)
female <- length(which(sex != 0))
message('No. of females = ', female)
```

Output:

```
No. of males = 14
No. of females = 11
```

Code: (For Males)

```
male_data <- subset.data.frame(data, sex == '0')
male_data_summary <- as.data.frame(apply(male_data,2,summary))
View(male_data_summary)
write_xlsx(male_data_summary,"male_data_summary.xlsx")
```

Output:

	age	sex	height	weight	bmp	fev1	rv	frc	tlc	pemax
Min.	7	0	109	13.1	64	22	171	104	95	70
1st Qu.	9.75	0	134	22.4	68.25	33.25	184.75	127.75	101.5	95
Median	15.5	0	165.5	43.65	78.5	38.5	215	135	106	100
Mean	15.21429	0	155.9286	41.36429	79.71429	39.85714	234.9286	148.4286	113.6429	117.5
3rd Qu.	19.75	0	174.75	53.725	92	48	249.75	153.25	125.5	152.5
Max.	23	0	180	73.8	97	57	441	268	147	195

Code: (For Females)

```
female_data <- subset.data.frame(data, sex == '1')
female_data_summary <- as.data.frame(apply(female_data,2,summary))
View(female_data_summary)
write_xlsx(female_data_summary,"female_data_summary.xlsx")
```

Output:

	age	sex	height	weight	bmp	fev1	rv	frc	tlc	pemax
Min.	7	1	112	12.9	65	18	158	118	81	65
1st Qu.	11.5	1	144.5	29.55	67.5	22	210	126.5	102	85
Median	14	1	153	34.8	70	28	253	146	120	90
Mean	13.54545	1	148.8182	34.63636	76.45455	28.18182	281	164.2727	114.4545	98.45455
3rd Qu.	16.5	1	157	39.65	89.5	30	337	194.5	127	115
Max.	19	1	176	60.1	93	45	449	245	136	134

Comments:

The recorded data includes information about **11 females** and **14 males** suffering from Cystic fibrosis.

MALES:

Youngest male is **7 years** old, approximately 25% of the males lie below the age of **10**, 50% of the males lie below the age of **15**, average age of males suffering from Cystic fibrosis is approximately **15**, 75% of the males lie below the age of **20** and the oldest male is of **23** years old. Range of age lies between **7 to 23 years**.

Shortest male is **109 cm** tall, approximately 25% of the males lie below the height of **134 cm**, 50% of the males lie below the height of **165.5 cm**, average height of males suffering from Cystic fibrosis is approx. **156 cm**, 75% of the males lie below the height of **174.75 cm** and the tallest male is of **180 cm** years old. Range of height lies between **109 to 180 cm**.

Minimum weight amongst all males is **13.1 Kg**, approximately 25% of the males lie below the weight of **22.4 Kg**, 50% of the males lie below the weight of **43.65 Kg**, average weight of males suffering from Cystic fibrosis is **41.36 Kg**, 75% of the males lie below the weight of **53.72 Kg** and the heaviest male is of **73.8 Kg**. Range of weight lies between **13.1 to 73.8 Kg**.

Minimum body mass amongst all males is **64 %**, approximately 25% of the males lie below the body mass of **68 %**, 50% of the males lie below the body mass of **71 %**, average body mass of males suffering from Cystic fibrosis is **78.28 %**, 75% of the males lie below the body mass of **90 %** and the highest body mass amongst all males is **97 %**. Range of body mass lies between **64 to 97 %**.

FEV stands for forced expiratory volume. Minimum FEV amongst all males is **22**, approximately 25% of the males lie below the FEV of **33.25**, 50% of the males lie below the FEV of **38.5**, average FEV of males suffering from Cystic fibrosis is **39.85**, 75% of the males lie below the FEV of **48** and the highest FEV amongst all males is **57**. Range of FEV lies between **22 to 57**.

RV stands for residual volume. Minimum RV amongst all males is **171**, approximately 25% of the males lie below the RV of **184.75**, 50% of the males lie below the RV of **215**, average RV of males suffering from Cystic fibrosis is **234.92**, 75% of the males lie below the RV of **249.75** and the highest RV amongst all males is **441**. Range of RV lies between **171 to 441**.

FRC stands for functional residual capacity. Minimum FRC amongst all males is **104**, approximately 25% of the males lie below the FRC of **127.75**, 50% of the males lie below the FRC of **135**, average FRC of males suffering from Cystic fibrosis is **148.42**, 75% of the males lie below the FRC of **153.25** and the highest FRC amongst all males is **268**. Range of FRC lies between **104 to 268**.

TLC stands for total lung capacity. Minimum TLC amongst all males is **95**, approximately 25% of the males lie below the TLC of **101.5**, 50% of the males lie below the TLC of **106**, average TLC of males suffering from Cystic fibrosis is **113.64**, 75% of the males lie below the TLC of **125.5** and the highest TLC amongst all males is **147**. Range of TLC lies between **95 to 147**.

PEMAX stands for maximum expiratory pressure. Minimum PEMAX amongst all males is **70**, approximately 25% of the males lie below the PEMAX of **95**, 50% of the males lie below the PEMAX of **100**, average PEMAX of males suffering from Cystic fibrosis is **117.5**, 75% of the males lie below the PEMAX of **152.5** and the highest PEMAX amongst all males is **195**. Range of PEMAX lies between **70 to 195**.

FEMALES:

Youngest female is **7 years** old, approximately 25% of the females lie below the age of **11**, 50% of the females lie below the age of **14**, average age of females suffering from Cystic fibrosis is approx. **14**, 75% of the females lie below the age of **16** and the oldest female is of **19** years old. Range of age lies between **7 to 19 years**.

Shortest female is **112 cm** tall, approximately 25% of the females lie below the height of **144.5 cm**, 50% of the females lie below the height of **153 cm**, average height of females suffering from Cystic fibrosis is **148.81 cm**, 75% of the females lie below the height of **157 cm** and the tallest female is of **176 cm** years old. Range of height lies between **112 to 176 cm**.

Minimum weight amongst all females is **12.9 Kg**, approximately 25% of the females lie below the weight of **29.55 Kg**, 50% of the females lie below the weight of **34.8 Kg**, average weight of females suffering from Cystic fibrosis is **34.63636 Kg**, 75% of the females lie below the weight of **39.65 Kg** and the heaviest female is of **60.1 Kg**. Range of weight lies between **12.9 to 60.1 Kg**.

Minimum body mass amongst all females is **65 %**, approximately 25% of the females lie below the body mass of **67.5 %**, 50% of the females lie below the body mass of **70 %**, average body mass of females suffering from Cystic fibrosis is **76.45%**, 75% of the females lie below the body mass of **89.5 %** and the highest body mass amongst all females is **93 %**. Range of body mass lies between **65 to 93 %**.

FEV stands for forced expiratory volume. Minimum FEV amongst all females is **18**, approximately 25% of the females lie below the FEV of **22**, 50% of the females lie below the FEV of **28**, average FEV of females suffering from Cystic fibrosis is **28.18**, 75% of the females lie below the FEV of **30** and the highest FEV amongst all females is **45**. Range of FEV lies between **18 to 45**.

RV stands for residual volume. Minimum RV amongst all females is **158**, approximately 25% of the females lie below the RV of **210**, 50% of the females lie below the RV of **253**, average RV of females suffering from Cystic fibrosis is **281**, 75% of the females lie below the RV of **337** and the highest RV amongst all females is **449**. Range of RV lies between **158 to 449**.

FRC stands for functional residual capacity. Minimum FRC amongst all females is **118**, approximately 25% of the females lie below the FRC of **126.5**, 50% of the females lie below the FRC of **146**, average FRC of females suffering from Cystic fibrosis is **164.27**, 75% of the females lie below the FRC of **194.5** and the highest FRC amongst all females is **245**. Range of FRC lies between **118 to 245**.

TLC stands for total lung capacity. Minimum TLC amongst all females is **81**, approximately 25% of the females lie below the TLC of **102**, 50% of the females lie below the TLC of **120**, average TLC of females suffering from Cystic fibrosis is **114.45**, 75% of the females lie below the TLC of **127** and the highest TLC amongst all females is **136**. Range of TLC lies between **81 to 136**.

PEMAX stands for maximum expiratory pressure. Minimum PEMAX amongst all females is **65**, approximately 25% of the females lie below the PEMAX of **85**, 50% of the females lie below the PEMAX of **90**, average PEMAX of females suffering from Cystic fibrosis is **98.45**, 75% of the females lie below the PEMAX of **115** and the highest PEMAX amongst all females is **134**. Range of PEMAX lies between **65 to 134**.

Answer 4) Part a)

Code (for correlation matrix to find strong relationships between variables):

```
correlation <- cor(data)

correlation <- as.data.frame(round(correlation, 2))

View(correlation)

write_xlsx(correlation,"correlation.xlsx")
```

Output:

(**one can just look at the upper triangle of the correlation matrix. For 30 unique correlations.**)

	age	sex	height	weight	bmp	fev1	rv	frc	tlc	pemax
age	1	-0.16712	0.926052	0.905867	0.377764	0.294488	-0.55194	-0.63936	-0.46937	0.613474
sex	-0.16712	1	-0.16755	-0.19044	-0.13756	-0.52826	0.271352	0.183605	0.024235	-0.28857
height	0.926052	-0.16755	1	0.920695	0.440762	0.316664	-0.56952	-0.62428	-0.45708	0.59922
weight	0.905867	-0.19044	0.920695	1	0.672546	0.448839	-0.62151	-0.61726	-0.41847	0.635222
bmp	0.377764	-0.13756	0.440762	0.672546	1	0.54552	-0.58237	-0.43439	-0.3649	0.229515
fev1	0.294488	-0.52826	0.316664	0.448839	0.54552	1	-0.66586	-0.66511	-0.44299	0.453376
rv	-0.55194	0.271352	-0.56952	-0.62151	-0.58237	-0.66586	1	0.910603	0.589139	-0.31555
frc	-0.63936	0.183605	-0.62428	-0.61726	-0.43439	-0.66511	0.910603	1	0.7044	-0.41721
tlc	-0.46937	0.024235	-0.45708	-0.41847	-0.3649	-0.44299	0.589139	0.7044	1	-0.18162
pemax	0.613474	-0.28857	0.59922	0.635222	0.229515	0.453376	-0.31555	-0.41721	-0.18162	1

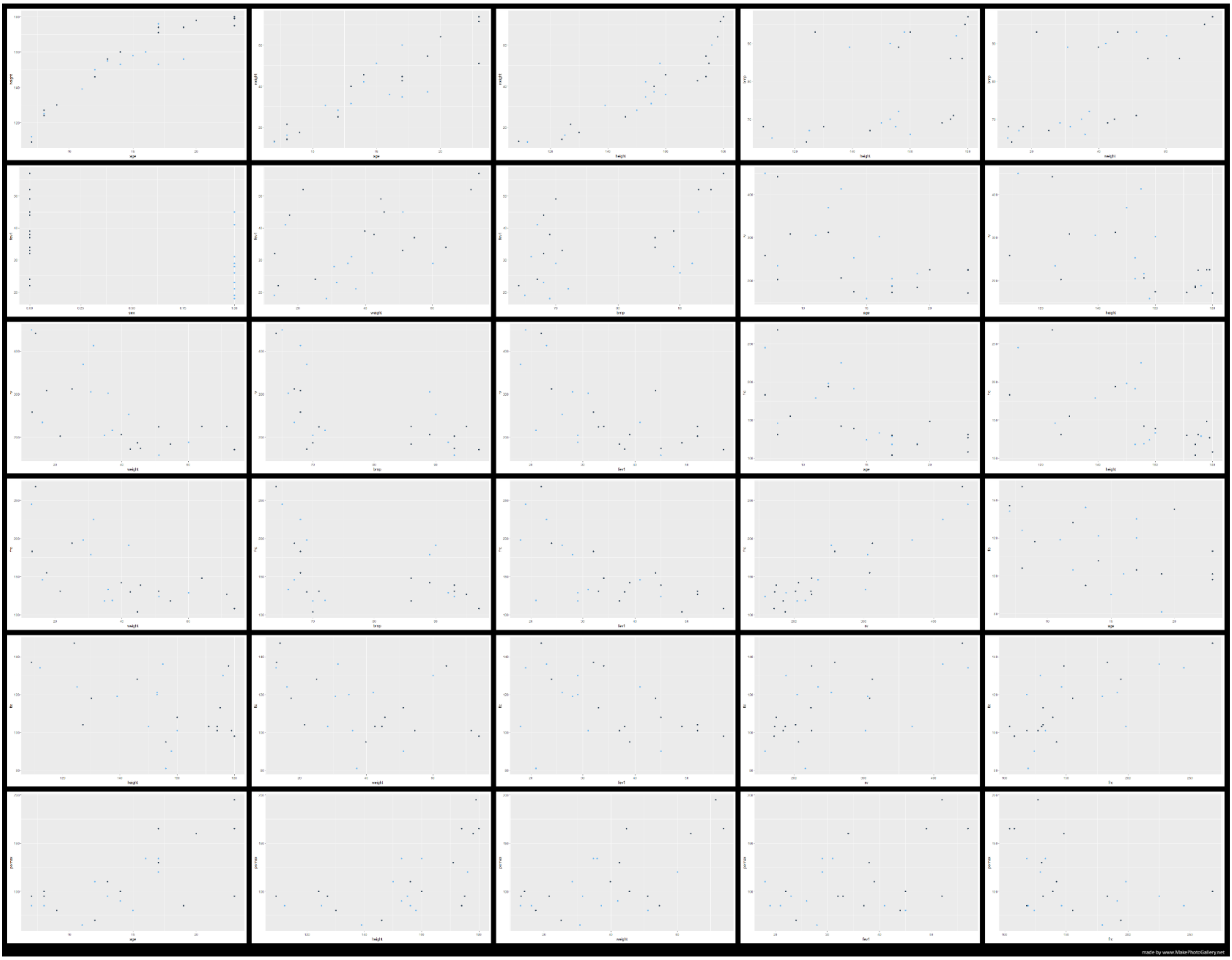
The 🟨 ones represent **strong correlation** and the 🟦 ones represent **moderate correlation**.

I can't comment on all of the 30 unique correlations but I will be discussing relationships between variables which have correlation above 0.55 and below -0.55. Filtering them out we get: (Filtered Table)

Feature 1	**Feature 2**	**Correlation**
age	height	0.926052
age	weight	0.905867
age	pemax	0.613474
age	frc	-0.63936
age	rv	-0.55194
height	weight	0.920695
height	pemax	0.59922
height	frc	-0.62428
height	rv	-0.56952
weight	bmp	0.672546
weight	pemax	0.635222
weight	rv	-0.62151
weight	frc	-0.61726
rv	frc	0.910603
rv	tlc	0.589139
rv	bmp	-0.58237
rv	fev1	-0.66586
frc	tlc	0.7044
frc	fev1	-0.66511

Filtered Table of Correlations

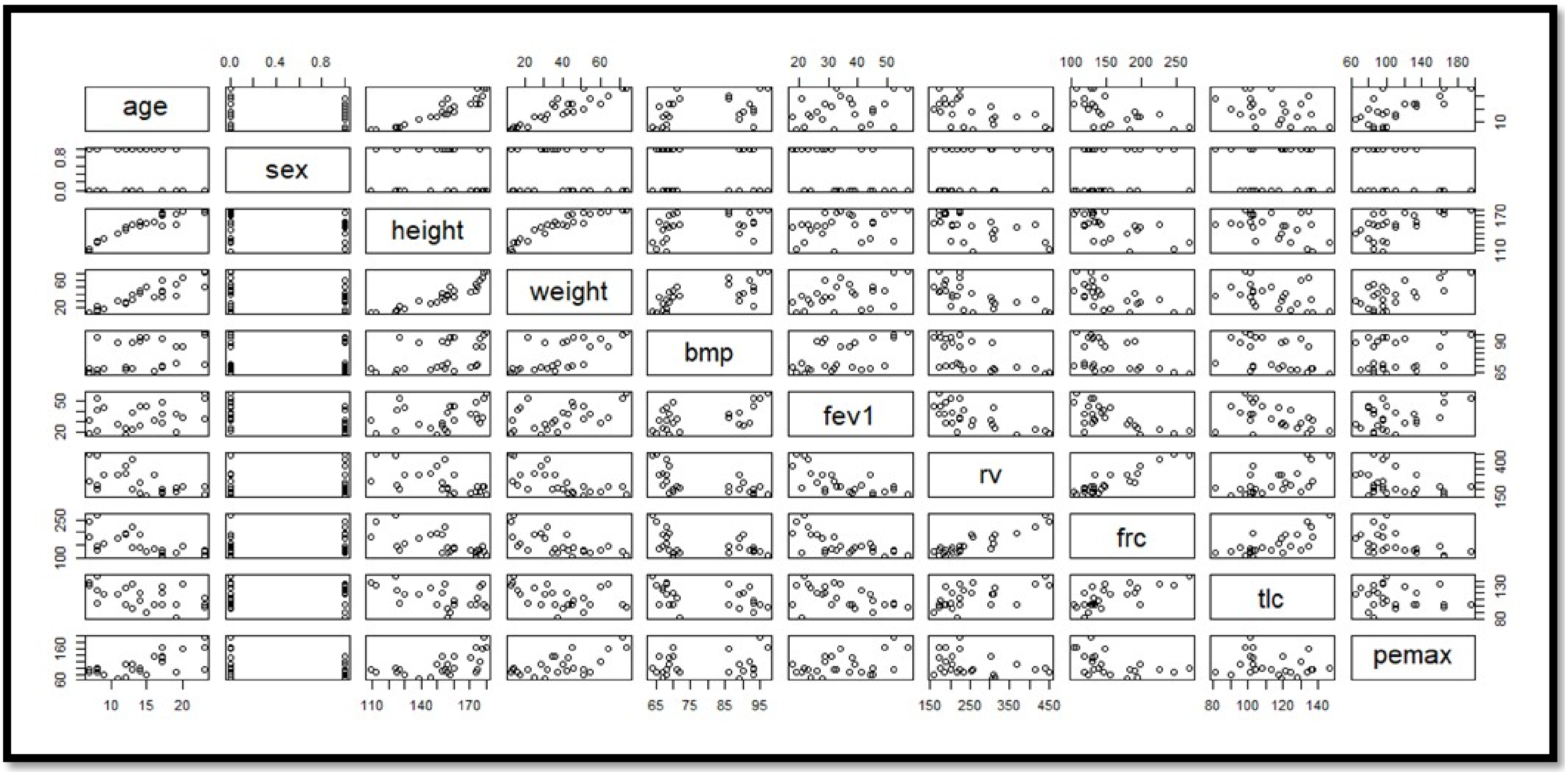
FOR GRAPHS:



In the above combined plot, **BLACK** points represent **MALES** and **BLUE** points represent **FEMALES**.

Clear relationships between variables:

As we can explicitly observe the above filtered table, with increase in age -> height , weight and pemax increases whereas frc and rv decreases. With increase in height -> weight and pemax increases whereas frc and rv decreases. With increase in weight -> bmp and pemax increases whereas rv and frc decreases. With increase in rv -> frc and tlc increases whereas bmp and fev1 decreases. With increase in frc -> tlc increases whereas fev1 decreases. With increase in fev1 -> tlc decreases however this relationship was not very strong. Since correlation measures the strength of linear relationships between 2 numeric variables, all these relationships were strong linear relationships. But there could have been strong non-linear relationships which correlation couldn’t find. When checked by human supervision, all kinds of strong relationships were already covered.
(These linear relationships can be verified from the upper triangle of the figure below):

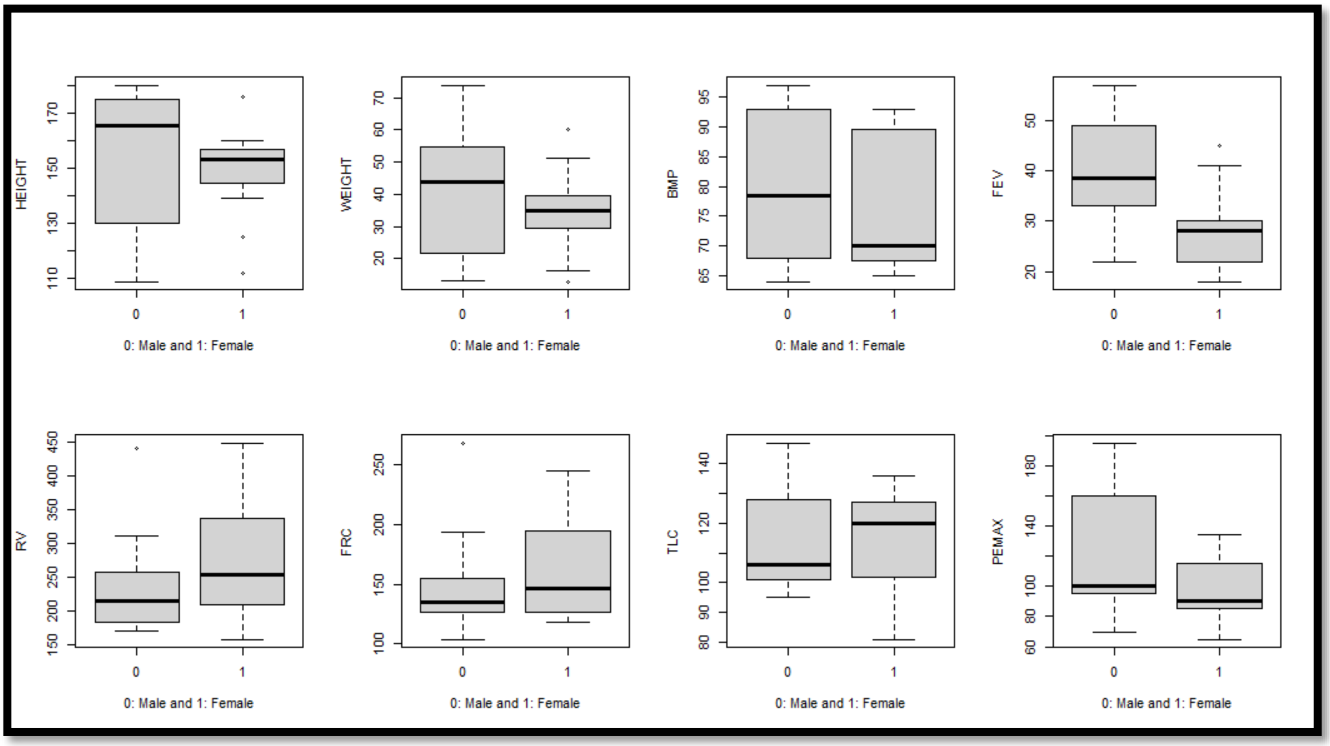


Answer 4) Part b)

Code:

```
par(mfrow = c(2,4))

boxplot(data$height~data$sex, xlab = '0: Male and 1: Female', ylab = 'HEIGHT')

boxplot(data$weight~data$sex, xlab = '0: Male and 1: Female', ylab = 'WEIGHT')

boxplot(data$bmp~data$sex, xlab = '0: Male and 1: Female', ylab = 'BMP')

boxplot(data$fev1~data$sex, xlab = '0: Male and 1: Female', ylab = 'FEV')

boxplot(data$rv~data$sex, xlab = '0: Male and 1: Female', ylab = 'RV')

boxplot(data$frc~data$sex, xlab = '0: Male and 1: Female', ylab = 'FRC')

boxplot(data$tlc~data$sex, xlab = '0: Male and 1: Female', ylab = 'TLC')

boxplot(data$pemax~data$sex, xlab = '0: Male and 1: Female', ylab = 'PEMAX')
```

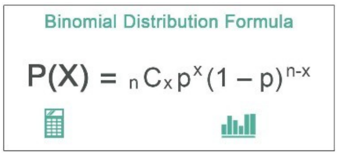
Output:



Outliers are represented by dots in boxplots. Outliers are those values which are above and below 1.5 times Inter Quartile Range. Height has 3 outliers for females. Weight has 2 outliers for female. FEV has 1 outlier for females. RV has 1 outlier for male. FRC has 1 outlier for male. BMP, TLC and PEMAX don't have any outliers

Answer 5)

Since the total number of trials are fixed, each trial has 2 possible outcomes, P(success) is same for each trial, none of the trials have effect on the probability of the next trial and exact probability of an event happening is given and we are asked to calculate the probability of this event happening x times out of n The given problem is of binomial distribution. With parameters x = 5, total trials (n) = 8 and probability of success (p) = 0.88



Binomial Distribution Formula

$$P(X) = {}_{n}C_{x} p^{x} (1-p)^{n-x}$$

Code:

```
dbinom(x = 5, size = 8, prob = 0.88)
```

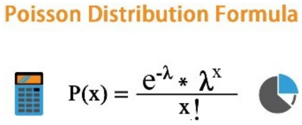
Output:

```
0.05106756
```

Therefore, the probability that exactly 5 of the next 8 patients having this operation survive is 0.05106756

Answer 6)

Since events are independent and the average probability of an event happening per unit time is given which is lambda and we are asked to calculate the probability of *x* events happening in a given time. The given problemis of Poisson's distribution with parameters x = 7/min and lambda = 5/min



Poisson Distribution Formula

$$P(x) = \frac{e^{-\lambda} * \lambda^{x}}{x!}$$

Code:

```
dpois(x = 7, lambda = 5 , log = FALSE)
```

Output:

```
0.1044449
```

Therefore, the probability of receiving 7 emails in a given minute is 0.1044449

Answer 7)

Since the mean and standard deviation is already given and we are required to calculate the area under the curve above a given threshold, The given problem is of Normal Distribution as only the Normal distribution requires mean and standard deviation. The parameters are x >= 10000, Mean = 14,500 and Standard Deviation = 2,500.

Code:

```
pnorm(q = 10000, mean = 14500, sd = 2500, lower.tail = F)
```

Output:

```
0.9640697
```

Therefore, the probability that more than 10000 liters will be sold is 0.9640697

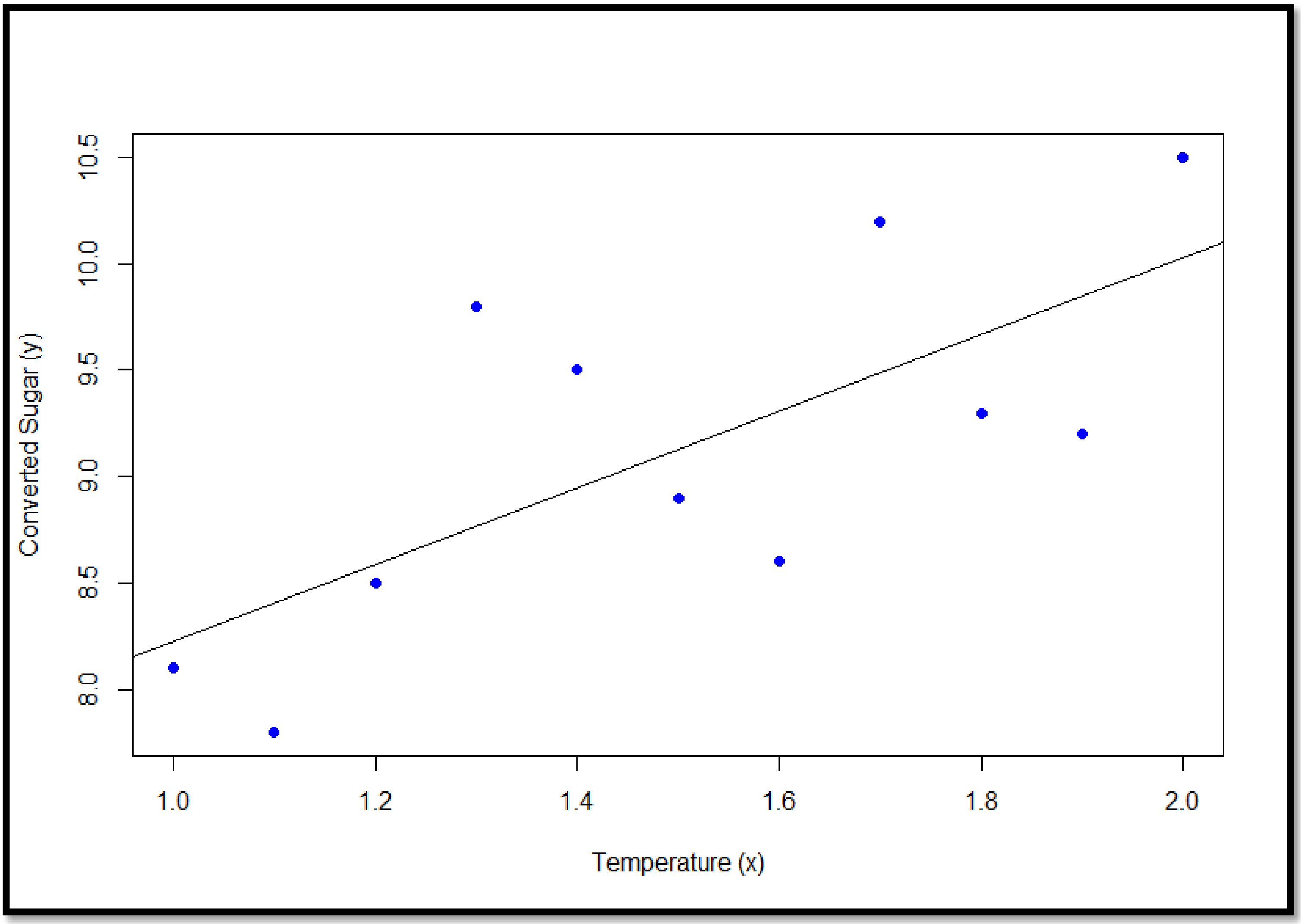
Answer 8) Part a)

Plotting a Linear Regression Line on the data

Code:

```
par(mfrow = c(1,1))

data2 <- as.data.frame(read.csv("Temperature_Sugar_Data.csv"))

lin_reg <- lm(data2$Converted_Sugar..y.~data2$Temperature..x. , data = data2)

plot(data2, pch = 16, col = "blue", xlab = 'Temperature (x)', ylab = 'Converted Sugar (y)')

abline(lin_reg)
```

Output:



To get the summary statistics:

Code:

```
lin_reg_summary <- summary(lin_reg)

View(lin_reg_summary)
```

Output:

```
Call:

lm(formula = data2$Converted_Sugar..y. ~ data2$Temperature..x., data = data2)

Residuals:

   Min       1Q     Median    3Q       Max

-0.7082  -0.4868  -0.1227  0.5109  1.0346

Coefficients:

                  Estimate Std. Error t value Pr(>|t|)

(Intercept)                    6.4136     0.9246   6.936     6.79e-05 ***

  data2$Temperature..x.   1.8091     0.6032  2.999  0.015 *

---

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.6326 on 9 degrees of freedom

Multiple R-squared:  0.4999, Adjusted R-squared:  0.4443

F-statistic: 8.996 on 1 and 9 DF,  p-value: 0.01497
```

To get coefficients and Residuals (separately):

Code:

```
lin_reg_residuals <- as.data.frame(lin_reg_summary[["residuals"]])

write_xlsx(lin_reg_residuals,"lin_reg_residuals.xlsx")

lin_reg_coefficients <- as.data.frame(lin_reg_summary[["coefficients"]])

write_xlsx(lin_reg_coefficients,"lin_reg_coefficients.xlsx")
```

Output:

Linear Regression Coefficients

Estimate	Std. Error	t value	Pr(>|t|)
6.413636	0.924638	6.936375	6.79E-05
1.809091	0.603167	2.999318	0.014973

	lin_reg_summary[["residuals"]]
1	-0.122727273
2	-0.603636364
3	-0.084545455
4	1.034545455
5	0.553636364
6	-0.227272727
7	-0.708181818
8	0.710909091
9	-0.37
10	-0.650909091
11	0.468181818

Answer 8) Part b)

To estimate the mean amount of converted sugar produced when the coded temperature is 1.76:

Code:

```
prediction <- lin_reg$coef[1] + lin_reg$coef[2]*1.76

print(prediction)
```

Output:

```
9.597636
```

Therefore, mean amount of converted sugar produced when the coded temperature is 1.76, is **9.597636**

Answer 9)

Code:

```
Number.of.Advertisements = c(0,6,4,5,2,7,3,10)

Purchases = c(4,8,5,10,1,3,4,12)
print(cor(Number.of.Advertisements, Purchases))
```

Output:

```
0.6790033
```

The correlation coefficient was 0.6790033. This indicates that "Number of Advertisements" has strong positive linear relationship with "Purchases". The number of data points were very less, but the strong positive linear relationship indicates that increasing Number of Advertisements will lead to more Purchases.

Answer 10)

TRENDS OF SPEED ON THE M1 MOTORWAY

ABSTRACT:

The study of road traffic and speed may involve sampling problems; in this report adaptations of fairly simple methods have been used to obtain representative samples in order to draw meaningful insights about the trends of speed on the M1 Motorway.

INTRODUCTION:

This report describes two sample investigations in which the "population" being sampled was the M1 Motorway way of the Great Britain. Both samples were based upon number of hinderances, type of hinderances and the average of the speeds on each junction. Aim of the first investigation was to find out the best days to travel, the second investigation was done with the aim of finding the best time to travel on these best days. The best days and best time to travel was not only based upon which day or time is the fastest but also on, which day or time is safest and smoothest to travel. Although the methods of sampling were fairly simple ones, their application has presented a number of insights which would be very beneficial for a manufacturing company based in London to carry out its logistic operations through delivery lorries while travelling through the length of the **M1**.

DATA COLLECTION:

Since the peak hours or the time at which data is collected can possibly result in biasing of the trends of traffic and speed of vehicles on each junction, data was collected over period of thirty-five days at the same time (7:30 P.M.) on each day. To investigate the best time to travel on the best days, the data was collected four times a day (8:30 A.M. to 9 A.M. , 5:30 P.M. to 6:00 P.M. , 7:30 P.M. to 8:00 P.M. and 9:30 P.M. to 10 P.M.) for each day over the period of 21 days. This data was collected from http://www.trafficengland.com/traffic-report. The raw data was then transformed into a tabular format which had ten variables namely "day" , "northbound", "junction", "southbound", "avg_speed", "UR", "Incidents", "Congestion", "total_hinderances" and "Time". Every row in the tabular data contained information about the speed, time, number of hinderances and type of hinderances for every two contiguous junctions. Since we calculated average speed by taking the average of speeds at northbound and southbound, we also calculated the average of number of hinderances at northbound and southbound for each type of hinderances and their values were filled in the cells corresponding to the respective contiguous junctions. There were 53 rows for a single day since there are 53 contiguous junctions on the M1 motorway.

INVESTIGATION 1: Best days to travel

Since data on https://www.trafficengland.com/traffic-report is changing with time, the data that was collected over the period of 35 days is in itself a big enough sample of the population dataset. In order to determine the best days, subsets were created for each day out of this sample dataset. Since the distribution of the average speed for each type of day was negatively skewed (Figure 1), median was preferred as a measure of central tendency to compare the average speed for each day.

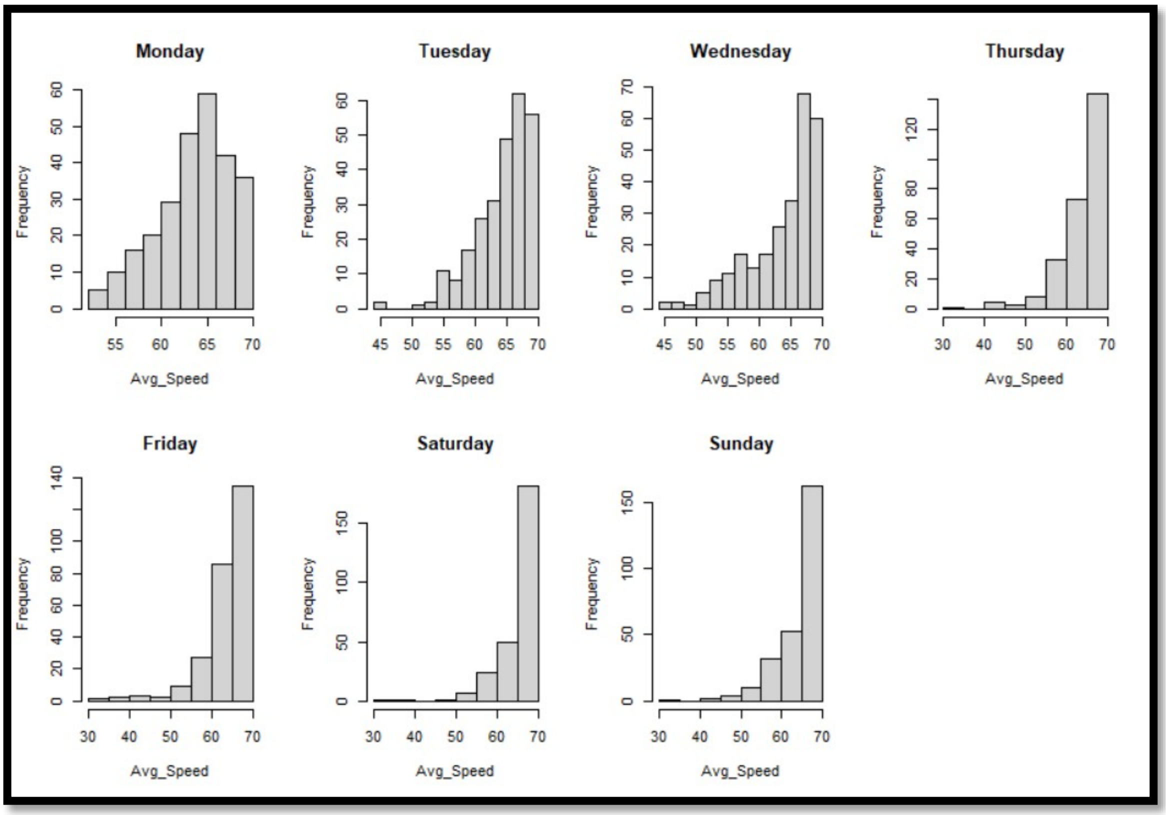


Figure 1: Distribution of Average Speed for each type of day

Summary of Average Speed for each day							
	Monday	Tuesday	Wednesday	Thursday	Friday	Saturday	Sunday
Minimum	53	45	44.5	34	30	30	30
1st Quartile	61	62	61	62.5	62	64.5	63
Median	64.5	65.5	66	65.5	65.5	67	66.5
Mean	63.82	64.58	63.97	64.28	63.81	65.59	64.82
3rd Quartile	67	68	68	68	67.5	69	69
Maximum	70	70	70	70	70	70	70

Table 1: Summary of Average Speed for each day

The maximum average speed for each kind of day was 70 miles per hour.
Ranking the days in descending order on the basis of median of average-speed we get:
List 1: Saturday > Sunday > Wednesday > Tuesday = Thursday = Friday > Monday.

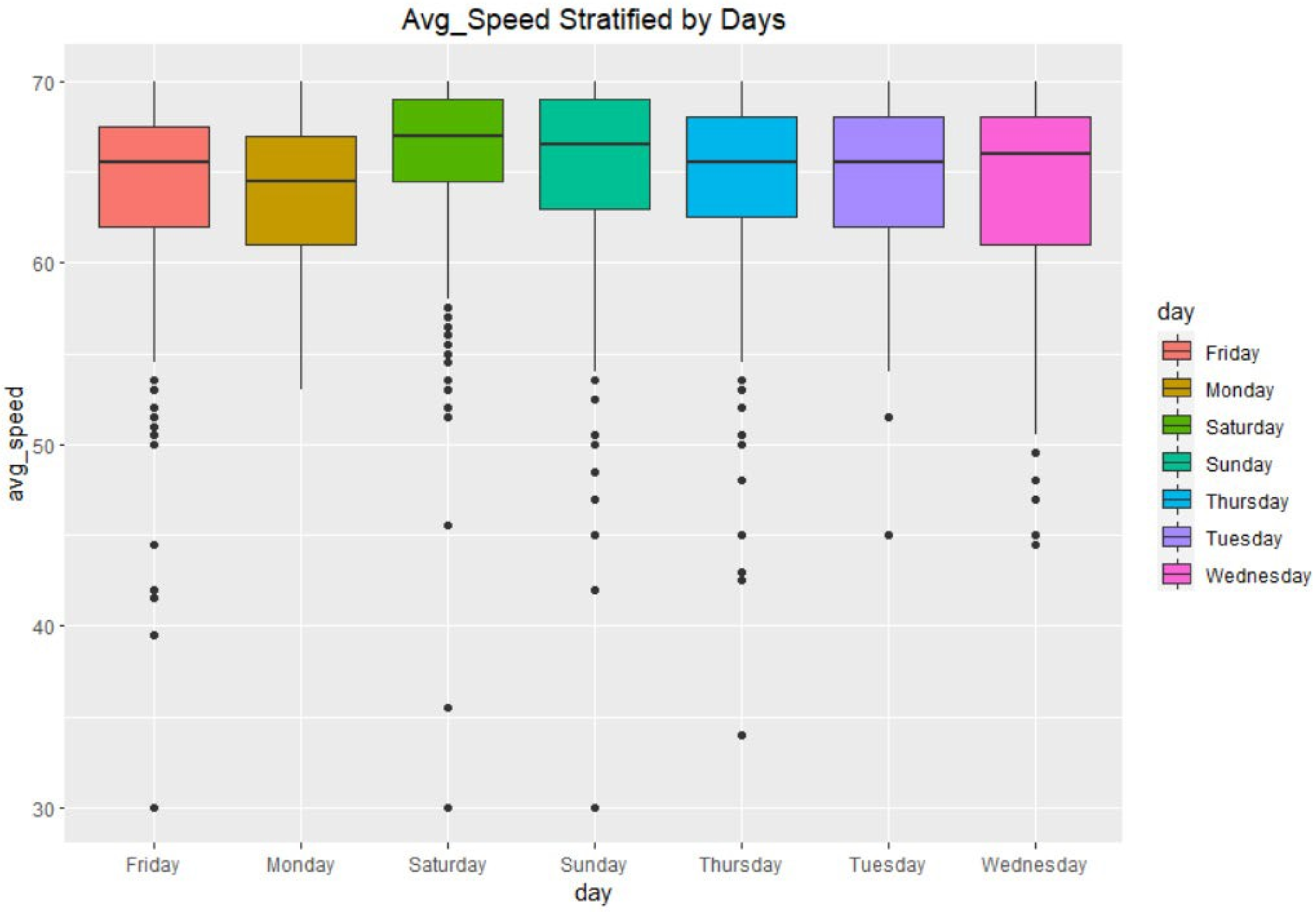


Figure 2: Boxplots for average speech stratified by type of day

As we can explicitly observe in Figure 2, there are a lot of outliers for each kind of day except Monday, since outliers affect the mean value of the data but **have little effect on the median** or mode of a given set of data; determining the best day to travel on the basis of median as a measure of central tendency for average speed for each day would still be the right criteria.

Intuitively, number of hinderances might be the factors responsible for these outliers.

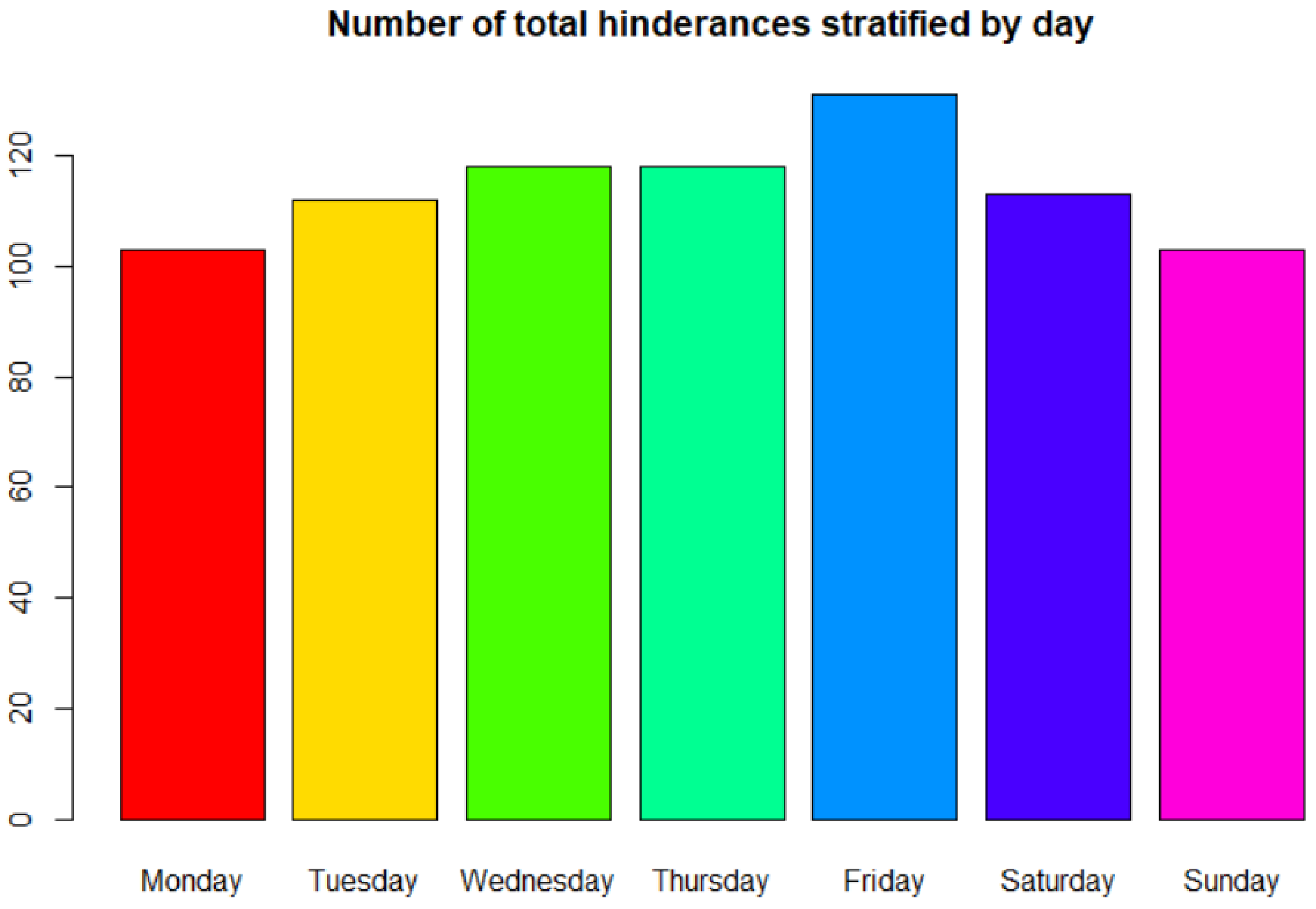


Figure 3: Bar-Plot of Total hinderances stratified by the type of days

Arranging number of total hinderances for each day in descending order: (Figure 3)

List 2: Friday > Thursday = Wednesday > Saturday > Tuesday > Monday = Sunday

Surprisingly since Monday did not have any outliers and number of total hinderances on Monday is same as on Sunday, this forces us re-consider our assumption. It seems that not only the number of hinderances but the type of hinderances also needed be taken into consideration for our analysis.

To identify which kind of hinderances (features) have a strong relationship with the average speed, feature selection was performed using Boruta library [1]. The way it works is that, first it will create duplicates of all independent variables. It will then shuffle the values of added duplicate copies in order to remove their correlation with the target variable which in our case is the average speed.

These duplicate copies are known as shadow variables. Then it runs a random forest classifier [2] on the combined dataset and performs a variable importance measure. After this step Z-Score is computed amongst shadow attributes to calculate MZSA that is Maximum Z Score Amongst Shadow Attributes. After arranging these shadow attributes in descending order based upon their MZSA values, shadow attributes which have a score below a certain threshold are declared as un-important.

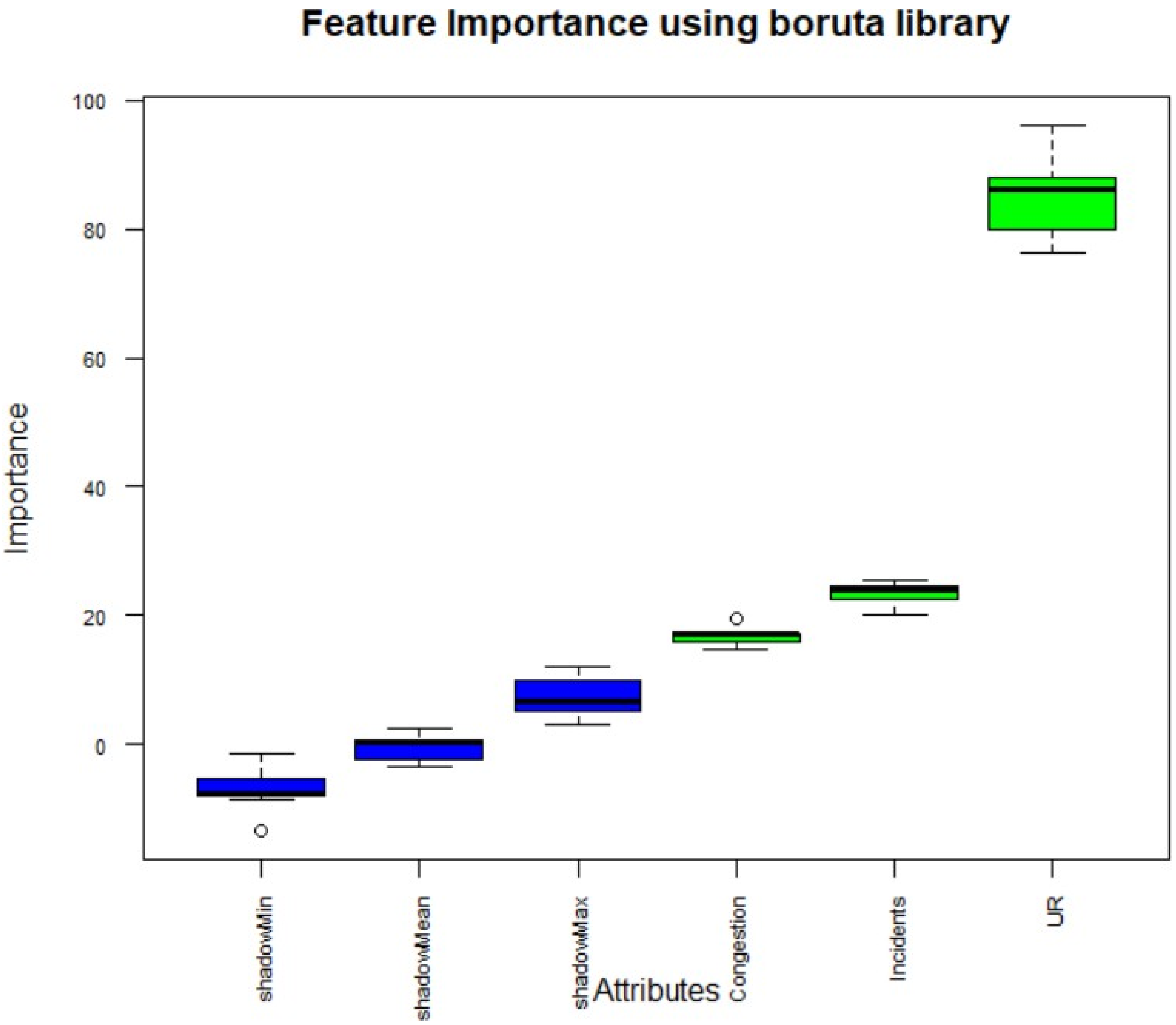


Figure 4: Hinderances that affect average speed the most.

“Unconfirmed Roadworks” and “Incidents” are the most important features that affects average speed the most.

	Monday	Tuesday	Wednessday	Thursday	Friday	Saturday	Sunday
UR	103	108	116	113	125	106	99
Congestion	0	1	0	2	2	4	2
Incidents	0	3	2	3	4	3	2

Table 2: Number of hinderances for each type of hinderance for each kind of day

It is very unsafe to travel on a day which has a lot of incidents, arranging days in ascending order with respect to total number of “Incidents”:

List 3: Monday < Sunday = Wednesday < Tuesday < Thursday < Saturday < Friday

Arranging the days in ascending order with respect to total number of “ Unconfirmed Roadworks “.

List 4: Sunday < Monday < Saturday < Tuesday < Thursday < Wednesday < Friday.

Arranging the days in ascending order with respect to total number of “ Congestions “.

List 5: Monday = Wednesday < Tuesday < Thursday = Friday = Sunday < Saturday

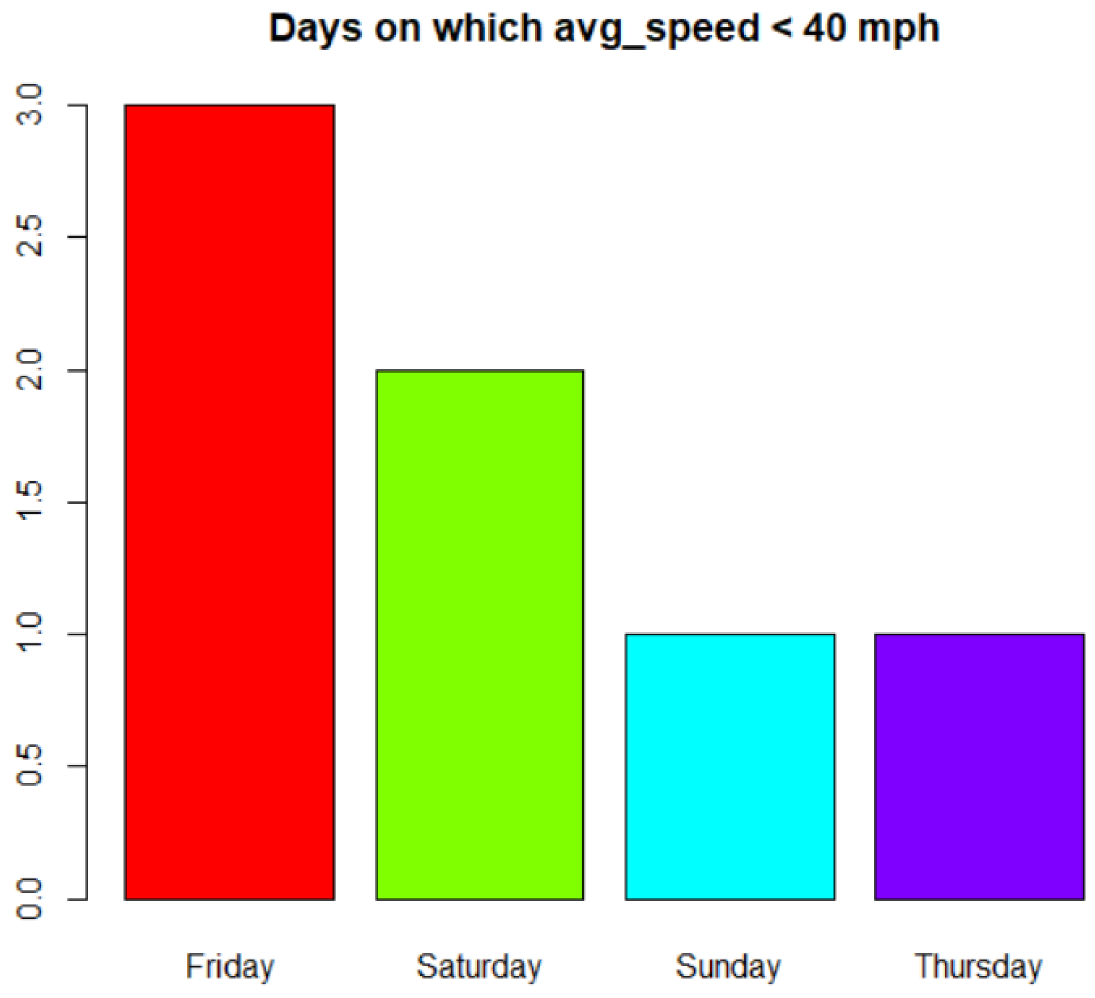


Figure 5: Bar-Plot for days which had instances when average speed was < 40 mph.

From the data illustrated in Figure 5, arranging the days in ascending order based upon the instances when the average speed went less than 40 miles per hour, we get:

List 6: Monday = Tuesday = Wednesday = Sunday < Thursday = Sunday < Saturday < Friday

Note that the maximum speed for each day was 70 miles per hour and difference between the median of average speeds for each day was not significant so other parameters were taken into consideration such as hinderances to make report the final list of rankings. These rankings not only include the information about which day is the fastest but also about which day is the safest to travel.

Now, in order to conclude all these ascending and descending order lists, if we score each day according to their respective ranks in these ascending and descending order lists and then calculate the final composite score for each day and arranging them descending order we get our final rankings:

Final List: Sunday > Monday > Wednesday > Saturday > Tuesday > Thursday > Friday.

The final list might be somewhat surprising because median speed for Monday was the lowest but since Monday had the least number of total hinderances, zero congestion and zero incidents. Monday proved out to be the safest and smoothest day to travel because of which it got second rank in the final rankings of days. Friday is the worst day to travel because of both very, low median of average speed and maximum number of incidents and congestions. All in all, Sunday came out to be the fastest and the best day to travel amongst all other days.

INVESTIGATION 2: Best time to travel on top 4 days

Since we already know the best days to travel, the aim of this investigation is to figure out the best time to travel for the top four days, which are Sunday, Monday, Wednesday and Saturday. Data was collected 4 times for these best four days at 8:30 A.M. to 9:00 A.M. , 5:30 P.M. to 6:00 P.M. , 7:30 P.M. to 8:00 P.M. and 9:30 P.M. to 10:00 P.M. .Since data on https://www.trafficengland.com/traffic-report is changing with time, the data that was collected at these 4 time periods over the period of 21 days is in itself a big enough sample of the population dataset. Since the distribution of speed was negatively skewed, median was preferred as a measure of central tendency to compare the average speed for each of these time periods.

Summary of Average Speed for each time period				
	8:30 A.M. to 9:00 A.M.	5:30 P.M. to 6:00 P.M.	7:30 P.M. to 8:00 P.M.	9:30 P.M. to 10:00 P.M.
Minimum	30	37	30	28
1st Quartile	62.5	58	62.38	61.5
Median	66	64.5	66	65
Mean	64.55	62.11	64.46	64.03
3rd Quartile	68	68	68	68
Maximum	70	70	70	70

Table 3: Summary of Average Speed for each time period

The maximum average speed for each time period was 70 miles per hour.
Ranking the time periods in descending order on the basis of median of average-speed we get:
List 7: 8:30 A.M. to 9:00 A.M. = 7:30 P.M. to 8:00 P.M. > 9:30 P.M. to 10:00 P.M. > 5:30 P.M. to 6:00 P.M.

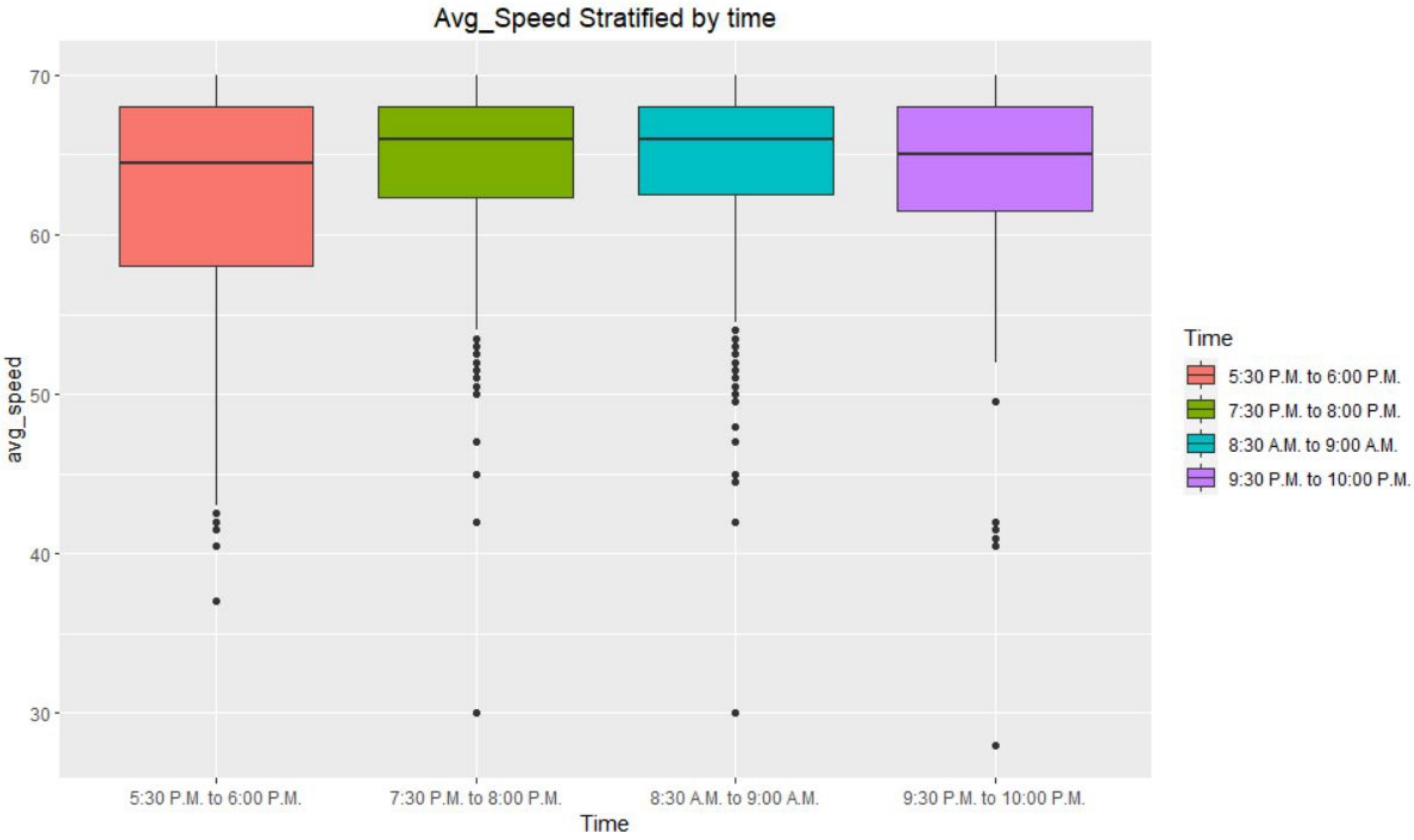


Figure 6: Boxplots of average speed stratified by time periods

Again, outliers were observed for average speeds at each of these time periods, but, since median, unlike the mean is not much influenced by outliers, our comparison of time periods on the basis of median is still quite legitimate.

Next, a comparison was done on the basis of instances when the average speed for each time period went below a certain threshold which was in the following case: 55 miles per hour.

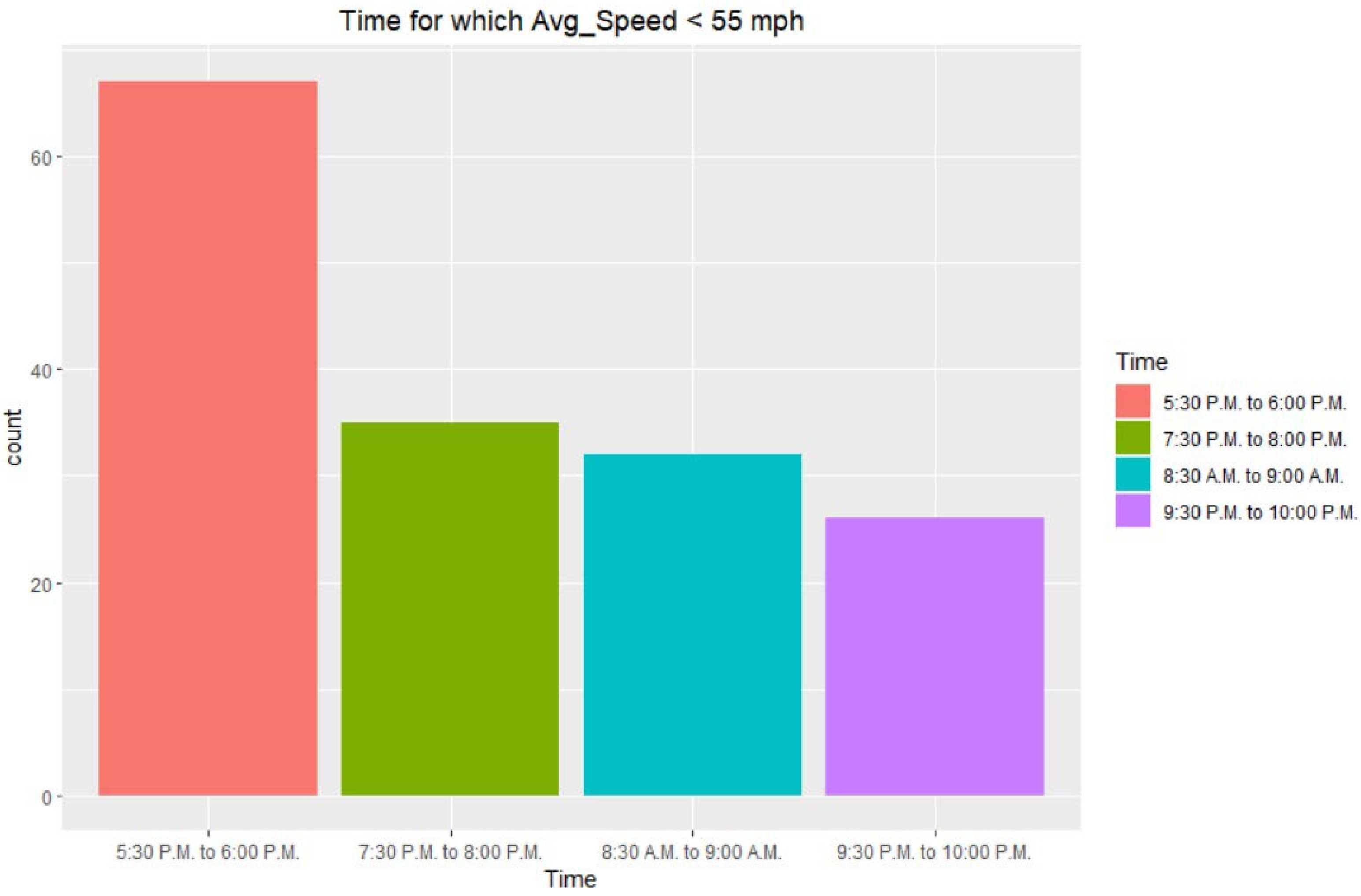


Figure 7: Histogram for number of instances for which average speed went below 55 mph

As per the trends illustrated in Figure 6, arranging time periods on the basis of number of instances where the average speed for each time period went below 55 mph, in ascending order we get:
List 8: 9:30 P.M. to 10:00 P.M. < 8:30 A.M. to 9:00 A.M. < 7:30 P.M. to 8:00 P.M. < 5:30 P.M. to 6:00 P.M.

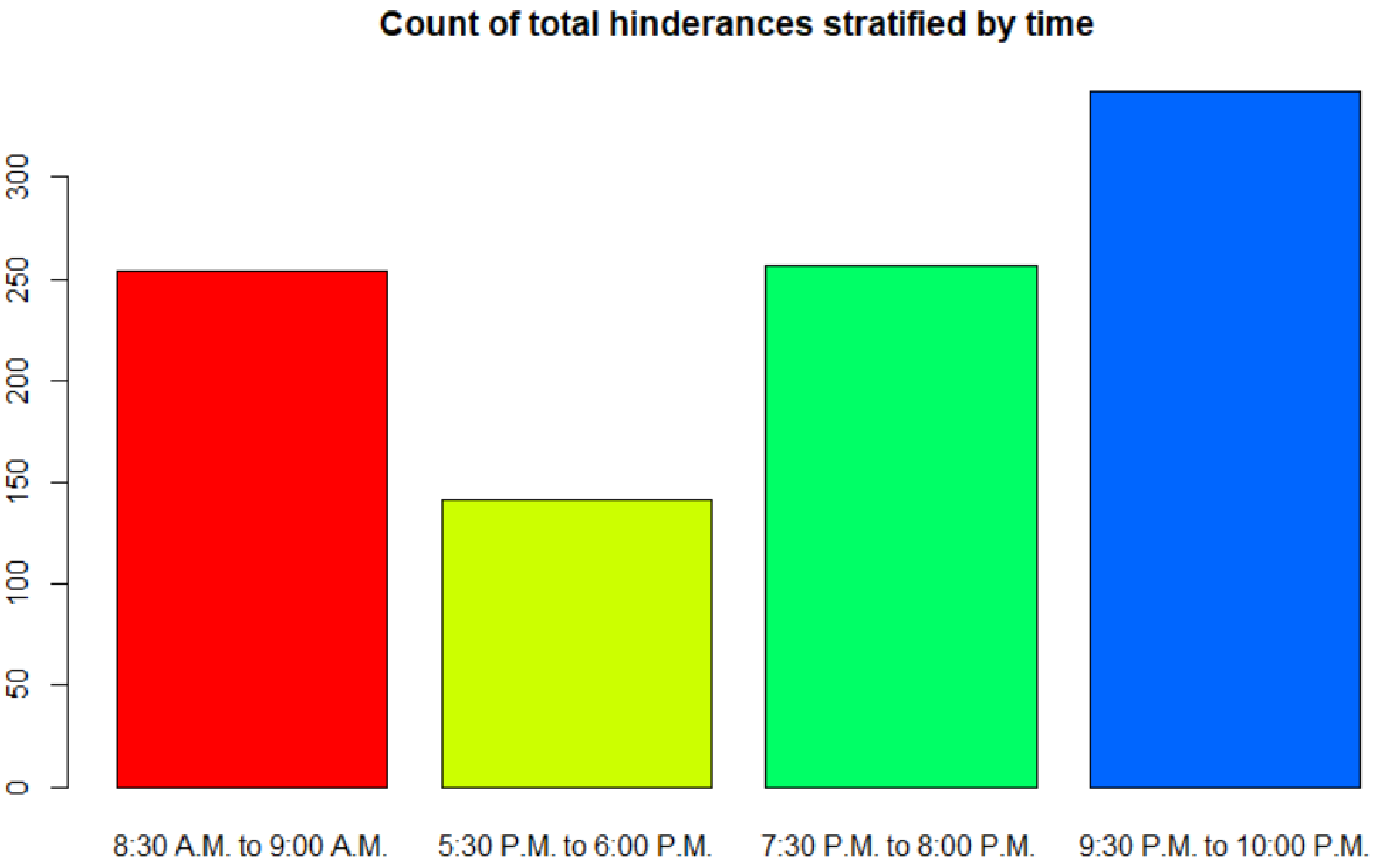


Figure 8: Bar-Plot of count of total number of hinderances stratified by time periods

Arranging the time periods in ascending order on the basis of counts of total number of hinderances for each time period we get:

List 9: 5:30 P.M. to 6:00 P.M. < 8:30 A.M. to 9:00 A.M. < 7:30 P.M. to 8:00 P.M. < 9:30 P.M. to 10:00 P.M.

Count of hinderances for each type of hinderances for each time period				
	8:30 A.M. to 9:00 A.M.	5:30 P.M. to 6:00 P.M.	7:30 P.M. to 8:00 P.M.	9:30 P.M. to 10:00 P.M.
UR	246	120	249	333
Congestion	4	11	4	6
Incidents	4	10	4	3

Table 4: Count of hinderances for each type of hinderances for each time period

In order to ensure a safe and smooth journey while travelling, it is very crucial to identify the time periods which has maximum Incidents and Congestions. Since, from previous investigation using Boruta library, it is already known that Unconfirmed Roadworks are the most important factors which affect average speed the most; a series of comparisons on the following metrics was performed:

Arranging the time periods on the basis of number of unconfirmed roadworks and ranking them in ascending order we get:

List 10: 5:30 P.M. to 6:00 P.M. < 8:30 A.M. to 9:00 A.M. < 7:30 P.M. to 8:00 P.M. < 9:30 P.M. to 10:00 P.M.

Arranging the time periods on the basis of number of Congestions and ranking them in ascending order we get:

List 11: 8:30 A.M. to 9:00 A.M. = 7:30 P.M. to 8:00 P.M. < 9:30 P.M. to 10:00 P.M. < 5:30 P.M. to 6:00 P.M.

Arranging the time periods on the basis of number of Incidents and ranking them in ascending order we get:

List 12: 9:30 P.M. to 10:00 P.M. < 8:30 A.M. to 9:00 A.M. = 7:30 P.M. to 8:00 P.M. < 5:30 P.M. to 6:00 P.M.

Now, in order to conclude List 6 to List 11, we again score each day according to their respective ranks in these ascending and descending ordered lists and then calculate the final composite score for each time period and arranging them descending order we get our final rankings:

Final List: 8:30 A.M. to 9:00 A.M. > 9:30 P.M. to 10 P.M. > 7:30 P.M. to 8:00 P.M. > 5:30 P.M. to 6:00 P.M.

The final list is again very surprising as it was expected that the time period between 8:30 A.M. to 9:00 A.M. would be undesirable as employees travel to reach their offices and children go to school during this time period, this time period still made its place to the first rank. On the contrary, the time period between 5:30 P.M. to 6:00 P.M. matched our expectations, since most employees leave their offices and travel back home during this time period and therefore, had the maximum number of Congestions and Incidents during this time period.

All in all, considering the trend observed in the Final List, it is recommended for the lorries to travel between 9:30 P.M. in the night to 8:30 A.M. in the morning as travelling within this period would ensure not only speed but also safety of travel during their journey and would avoid most of the hinderances such as congestions and especially incidents that might cause loss in the business revenue.

RESULTS AND CONCLUSION:

As discussed in both the investigations, the best days and time periods were determined not only on the basis of the average speeds but also considering the safety and smoothness of journey while travelling on the M1 motorway. The best days to travel were ranked and arranged in descending order as follows:

Sunday > Monday > Wednesday > Saturday > Tuesday > Thursday > Friday.

The best time periods to travel on top 4 best days were ranked and arranged in descending order as follows:

8:30 A.M. to 9:00 A.M. > 9:30 P.M. to 10 P.M. > 7:30 P.M. to 8:00 P.M. > 5:30 P.M. to 6:00 P.M.

The ideal scenario for the lorries to travel would be on Sunday between 9:30 P.M. in the night to 8:30 A.M. in the morning as it would ensure speed, safety, and smoothness of journey. Whereas the worst case would be to travel on Friday between 5:30 P.M. to 8:00 P.M.

REFERENCES:

[1] Miron B. Kursa, Witold R. Rudnicki (2010). Feature Selection with the Boruta Package. Journal of Statistical Software, 36(11), p. 1-13. URL: http://www.jstatsoft.org/v36/i11/

[2] Ho, T.K., 1995. Random decision forests. In *Proceedings of 3rd international conference on* document analysis and recognition. pp. 278–282

APPENDIX:

The table below illustrates data collected for 2 consecutive days at the same time for both the days.

After the table, code for performing all the calculations and making charts is given

day	northbound	junction	southbound	avg_speed	UR	Incidents	Congestion	total_hinderances	Time
Saturday	70	48,47	70	70	0	0	0	0	7:30 P.M. to 8:00 P.M.
Saturday	70	47,46	69	69.5	0	0	0	0	7:30 P.M. to 8:00 P.M.
Saturday	67	46,45	66	66.5	0	0	0	0	7:30 P.M. to 8:00 P.M.
Saturday	68	45,43|44	63	65.5	0	0	0	0	7:30 P.M. to 8:00 P.M.
Saturday	64	43|44,42	68	66	0	0	0	0	7:30 P.M. to 8:00 P.M.
Saturday	66	42,41	68	67	0	0	0	0	7:30 P.M. to 8:00 P.M.
Saturday	68	41,40	70	69	0	0	0	0	7:30 P.M. to 8:00 P.M.
Saturday	67	40,39	70	68.5	0	0	0	0	7:30 P.M. to 8:00 P.M.
Saturday	70	39,38	70	70	0	0	0	0	7:30 P.M. to 8:00 P.M.
Saturday	67	38,37	68	67.5	0	0	0	0	7:30 P.M. to 8:00 P.M.
Saturday	69	37,36	70	69.5	0	0	0	0	7:30 P.M. to 8:00 P.M.
Saturday	70	36,35A	68	69	0	0	0	0	7:30 P.M. to 8:00 P.M.
Saturday	65	35A,35	67	66	0	0	0	0	7:30 P.M. to 8:00 P.M.
Saturday	61	35,34	56	58.5	0	0	0	0	7:30 P.M. to 8:00 P.M.
Saturday	53	34,33	59	56	1	0	0	1	7:30 P.M. to 8:00 P.M.
Saturday	58	33,32	65	61.5	1	0	0	1	7:30 P.M. to 8:00 P.M.
Saturday	66	32,31	67	66.5	1	0	0	1	7:30 P.M. to 8:00 P.M.
Saturday	69	31,30	70	69.5	0	0	0	0	7:30 P.M. to 8:00 P.M.
Saturday	70	30,29A	68	69	0	0	0	0	7:30 P.M. to 8:00 P.M.
Saturday	70	29A,29	70	70	0	0	0	0	7:30 P.M. to 8:00 P.M.
Saturday	66	29,28	70	68	0	0	0	0	7:30 P.M. to 8:00 P.M.
Saturday	68	28,27	68	68	0	0	0	0	7:30 P.M. to 8:00 P.M.
Saturday	69	27,26	68	68.5	0	0	0	0	7:30 P.M. to 8:00 P.M.
Saturday	68	26,25	69	68.5	1	0	0	1	7:30 P.M. to 8:00 P.M.
Saturday	70	25,24A	70	70	1	0	0	1	7:30 P.M. to 8:00 P.M.
Saturday	70	24A,24	70	70	0	0	0	0	7:30 P.M. to 8:00 P.M.
Saturday	70	24,23A	67	68.5	0	0	0	0	7:30 P.M. to 8:00 P.M.
Saturday	70	23A,23	70	70	0	0	0	0	7:30 P.M. to 8:00 P.M.
Saturday	70	23,22	70	70	0	0	0	0	7:30 P.M. to 8:00 P.M.
Saturday	70	22,21A	67	68.5	0	0	0	0	7:30 P.M. to 8:00 P.M.
Saturday	66	21A,21	64	65	0	0	0	0	7:30 P.M. to 8:00 P.M.
Saturday	70	21,20	70	70	1	0	0	1	7:30 P.M. to 8:00 P.M.
Saturday	70	20,19	70	70	1	0	0	1	7:30 P.M. to 8:00 P.M.
Saturday	69	19,18	70	69.5	1	0	0	1	7:30 P.M. to 8:00 P.M.
Saturday	70	18,17	70	70	1	0	0	1	7:30 P.M. to 8:00 P.M.
Saturday	69	17,16	67	68	2	0	0	2	7:30 P.M. to 8:00 P.M.
Saturday	54	16,15A	58	56	1	0	0	1	7:30 P.M. to 8:00 P.M.
Saturday	50	15A,15	57	53.5	1	1	0	2	7:30 P.M. to 8:00 P.M.
Saturday	57	15,14	56	56.5	4	0	0	4	7:30 P.M. to 8:00 P.M.
Saturday	61	14,13	59	60	3	0	0	3	7:30 P.M. to 8:00 P.M.
Saturday	66	13,12	67	66.5	3	0	0	3	7:30 P.M. to 8:00 P.M.
Saturday	70	12,11A	65	67.5	2	0	0	2	7:30 P.M. to 8:00 P.M.
Saturday	32	11A,11	59	45.5	0	0	0	0	7:30 P.M. to 8:00 P.M.
Saturday	12	11,10	59	35.5	0	1	1	2	7:30 P.M. to 8:00 P.M.
Saturday	64	10,9	68	66	0	0	1	1	7:30 P.M. to 8:00 P.M.
Saturday	63	9,8	65	64	0	0	0	0	7:30 P.M. to 8:00 P.M.
Saturday	65	8,7	61	63	0	0	0	0	7:30 P.M. to 8:00 P.M.
Saturday	66	7,6A	66	66	0	0	0	0	7:30 P.M. to 8:00 P.M.
Saturday	64	6A,6	67	65.5	0	0	0	0	7:30 P.M. to 8:00 P.M.
Saturday	65	6,5	67	66	0	0	0	0	7:30 P.M. to 8:00 P.M.
Saturday	66	5,4	66	66	0	0	0	0	7:30 P.M. to 8:00 P.M.
Saturday	70	4,2	67	68.5	1	0	0	1	7:30 P.M. to 8:00 P.M.
Saturday	70	2,1	50	60	0	0	0	0	7:30 P.M. to 8:00 P.M.
Sunday	69	48,47	69	69	0	0	0	0	7:30 P.M. to 8:00 P.M.
Sunday	70	47,46	70	70	0	0	0	0	7:30 P.M. to 8:00 P.M.
Sunday	69	46,45	70	69.5	0	0	0	0	7:30 P.M. to 8:00 P.M.
Sunday	70	45,43|44	68	69	0	0	0	0	7:30 P.M. to 8:00 P.M.
Sunday	70	43|44,42	69	69.5	0	0	0	0	7:30 P.M. to 8:00 P.M.
Sunday	70	42,41	67	68.5	0	0	0	0	7:30 P.M. to 8:00 P.M.
Sunday	70	41,40	69	69.5	0	0	0	0	7:30 P.M. to 8:00 P.M.
Sunday	69	40,39	64	66.5	0	0	0	0	7:30 P.M. to 8:00 P.M.
Sunday	70	39,38	70	70	0	0	0	0	7:30 P.M. to 8:00 P.M.
Sunday	70	38,37	70	70	0	0	0	0	7:30 P.M. to 8:00 P.M.
Sunday	70	37,36	69	69.5	0	0	0	0	7:30 P.M. to 8:00 P.M.
Sunday	70	36,35A	63	66.5	0	0	0	0	7:30 P.M. to 8:00 P.M.
Sunday	67	35A,35	69	68	0	0	0	0	7:30 P.M. to 8:00 P.M.
Sunday	66	35,34	50	58	0	0	0	0	7:30 P.M. to 8:00 P.M.
Sunday	51	34,33	58	54.5	0	0	0	0	7:30 P.M. to 8:00 P.M.
Sunday	64	33,32	69	66.5	0	0	0	0	7:30 P.M. to 8:00 P.M.
Sunday	68	32,31	70	69	1	0	0	1	7:30 P.M. to 8:00 P.M.
Sunday	70	31,30	69	69.5	0	0	0	0	7:30 P.M. to 8:00 P.M.
Sunday	70	30,29A	69	69.5	0	0	0	0	7:30 P.M. to 8:00 P.M.
Sunday	70	29A,29	70	70	0	0	0	0	7:30 P.M. to 8:00 P.M.
Sunday	70	29,28	70	70	0	0	0	0	7:30 P.M. to 8:00 P.M.
Sunday	69	28,27	70	69.5	0	0	0	0	7:30 P.M. to 8:00 P.M.
Sunday	68	27,26	70	69	1	0	0	1	7:30 P.M. to 8:00 P.M.
Sunday	70	26,25	70	70	1	0	0	1	7:30 P.M. to 8:00 P.M.
Sunday	70	25,24A	70	70	0	0	0	0	7:30 P.M. to 8:00 P.M.
Sunday	70	24A,24	65	67.5	0	0	0	0	7:30 P.M. to 8:00 P.M.
Sunday	70	24,23A	60	65	0	0	0	0	7:30 P.M. to 8:00 P.M.
Sunday	70	23A,23	60	65	0	0	0	0	7:30 P.M. to 8:00 P.M.
Sunday	70	23,22	69	69.5	0	0	0	0	7:30 P.M. to 8:00 P.M.
Sunday	70	22,21A	70	70	0	0	0	0	7:30 P.M. to 8:00 P.M.
Sunday	70	21A,21	67	68.5	0	0	0	0	7:30 P.M. to 8:00 P.M.
Sunday	70	21,20	66	68	1	0	0	1	7:30 P.M. to 8:00 P.M.
Sunday	70	20,19	63	66.5	1	0	0	1	7:30 P.M. to 8:00 P.M.
Sunday	70	19,18	65	67.5	1	0	0	1	7:30 P.M. to 8:00 P.M.
Sunday	68	18,17	66	67	0	0	0	0	7:30 P.M. to 8:00 P.M.
Sunday	67	17,16	66	66.5	2	0	0	2	7:30 P.M. to 8:00 P.M.
Sunday	55	16,15A	56	55.5	1	0	0	1	7:30 P.M. to 8:00 P.M.
Sunday	49	15A,15	56	52.5	0	0	0	0	7:30 P.M. to 8:00 P.M.
Sunday	57	15,14	54	55.5	3	0	0	3	7:30 P.M. to 8:00 P.M.
Sunday	59	14,13	61	60	4	0	0	4	7:30 P.M. to 8:00 P.M.
Sunday	68	13,12	67	67.5	2	0	0	2	7:30 P.M. to 8:00 P.M.
Sunday	67	12,11A	69	68	0	0	0	0	7:30 P.M. to 8:00 P.M.
Sunday	67	11A,11	69	68	0	0	0	0	7:30 P.M. to 8:00 P.M.
Sunday	69	11,10	67	68	0	0	0	0	7:30 P.M. to 8:00 P.M.
Sunday	69	10,9	66	67.5	0	0	0	0	7:30 P.M. to 8:00 P.M.
Sunday	67	9,8	63	65	0	0	0	0	7:30 P.M. to 8:00 P.M.
Sunday	65	8,7	64	64.5	0	0	0	0	7:30 P.M. to 8:00 P.M.
Sunday	69	7,6A	68	68.5	0	0	0	0	7:30 P.M. to 8:00 P.M.
Sunday	70	6A,6	70	70	0	0	0	0	7:30 P.M. to 8:00 P.M.
Sunday	70	6,5	63	66.5	0	0	0	0	7:30 P.M. to 8:00 P.M.
Sunday	67	5,4	66	66.5	1	0	0	1	7:30 P.M. to 8:00 P.M.
Sunday	66	4,2	63	64.5	0	0	0	0	7:30 P.M. to 8:00 P.M.
Sunday	70	2,1	58	64	0	0	0	0	7:30 P.M. to 8:00 P.M.

CODE FOR QUESTION 10:

```
# Question 10
# Determining the best days
data4 <- read.csv('Days Data for Question 10.csv')

monday <- subset(data4, data4$day == 'Monday')
tuesday <- subset(data4, data4$day == 'Tuesday')
wednesday <- subset(data4, data4$day == 'Wednesday')
thursday <- subset(data4, data4$day == 'Thursday')
friday <- subset(data4, data4$day == 'Friday')
saturday <- subset(data4, data4$day == 'Saturday')
sunday <- subset(data4, data4$day == 'Sunday')

# How is avg_speed distributed ?

#hist(data4$avg_speed)
par(mfrow = c(2,4))
hist(monday$avg_speed, main = 'Monday', xlab = 'Avg_Speed')
hist(tuesday$avg_speed, main = 'Tuesday', xlab = 'Avg_Speed')
hist(wednesday$avg_speed, main = 'Wednesday', xlab = 'Avg_Speed')
hist(thursday$avg_speed, main = 'Thursday', xlab = 'Avg_Speed')
hist(friday$avg_speed, main = 'Friday', xlab = 'Avg_Speed')
hist(saturday$avg_speed, main = 'Saturday', xlab = 'Avg_Speed')
hist(sunday$avg_speed, main = 'Sunday', xlab = 'Avg_Speed')

# Since the distribution of speed in all days is negatively skewed
# We would compare the avg_speeds on the basis of median

par(mfrow = c(1,1))
# Which days have max avg_speed and outliers ?
install.packages('ggplot2')
```

```
library(ggplot2)
ggplot(data4, aes(group = day,x = day, y = avg_speed, fill = day)) + geom_boxplot() +
ggtitle('Avg_Speed Stratified by Days') + theme(plot.title = element_text(hjust = 0.5))

# After looking at the boxplots of avg_speeds stratified by days, we can observe that
# there are a lot of outliers in the dataset so just determining the best days on the
# basis of median speeds for each day won't be the right criteria, we need have a look upon
# which factors led to these outliers, which are infact hindrances

# Which days have maximum total hindrances ?
monday_total <- sum(monday$total_hinderances)
tuesday_total <- sum(tuesday$total_hinderances)
wednesay_total <- sum(wednesday$total_hinderances)
thursday_total <- sum(thursday$total_hinderances)
friday_total <- sum(friday$total_hinderances)
saturday_total <- sum(saturday$total_hinderances)
sunday_total <- sum(sunday$total_hinderances)

vector <- c(monday_total, tuesday_total, wednesay_total, thursday_total, friday_total,
saturday_total, sunday_total)
days <- c("Monday", "Tuesday", "Wednesday", "Thursday", "Friday", "Saturday", "Sunday")
barplot(height=vector, names=days, col = rainbow(7), main = 'Number of total hinderances stratified
by day')
# Monday, Friday and Saturday have the most hindrances whereas, Tuesday, Thursday and
Wednesday
# have the least number of hindrances. Wednesday, Saturday and Sunday are the fastest.
# But Monday doesn't has any outliers which indicates
# that there might be other reasons apart from hindrances which are responsible for these outliers

# Which type of hindrances effect avg_speed the most ?
# There are more hindrances in southbound than in northbound
# Feature Selection using Boruta library:
```

```
install.packages('Boruta')
library(Boruta)
par(mfrow = c(1,1))
data5 <- data4[c('avg_speed', 'UR', 'Congestion', 'Incidents')]
boruta <- Boruta(data5$avg_speed ~. , data = data5, doTrace = 2, maxRuns = 500)
plot(boruta, las = 2, cex.axis = 0.7, main = 'Feature Importance using boruta library')

# Since Unconfirmed Roadworks effect speed the most we should focus more on that

# Which day has maximum congestion ?
monday_congestion <- sum(monday$Congestion) # 0
tuesday_congestion <- sum(tuesday$Congestion) # 1
wednesday_congestion <- sum(wednesday$Congestion) # 0
thursday_congestion <- sum(thursday$Congestion) # 2
friday_congestion <- sum(friday$Congestion) # 2
saturday_congestion <- sum(saturday$Congestion) # 4
sunday_congestion <- sum(sunday$Congestion) # 2

# Which day has maximum incidents ?
monday_incident <- sum(monday$Incidents) # 0
tuesday_incident <- sum(tuesday$Incidents) # 3
wednesday_incident <- sum(wednesday$Incidents) # 2
thursday_incident <- sum(thursday$Incidents) # 3
friday_incident <- sum(friday$Incidents) # 4
saturday_incident <- sum(saturday$Incidents) # 3
sunday_incident <- sum(sunday$Incidents) # 2

# Which day has maximum unconfirmed roadworks ?
monday_ur <- sum(monday$UR) # 103
tuesday_ur <- sum(tuesday$UR) # 108
wednesday_ur <- sum(wednesday$UR) # 116
```

```
thursday_ur <- sum(thursday$UR) # 113
friday_ur <- sum(friday$UR) # 125
saturday_ur <- sum(saturday$UR) # 106
sunday_ur <- sum(sunday$UR) # 99

# Monday is the safest day to travel but the slowest amongst all other days
# Friday is the most dangerous day to travel due to max hindrances and especially Incidents
# Determining the best time for all days
# Time 1: 8:30 A.M. to 9:00 A.M.
# Time 2: 5:30 P.M. to 6:00 P.M.
# Time 3: 7:30 P.M. to 8:00 P.M.
# Time 4: 9:30 P.M. to 10:00 P.M.

data6 <- read.csv('Time Data for Question 10.csv')
install.packages('dplyr')
library(dplyr)
install.packages("ggplot2")
library(ggplot2)
data6 <- data6 %>% filter(day %in% c("Monday", "Wednesday", "Sunday", "Saturday"))

Time_1 <- subset(data6, data6$Time == '8:30 A.M. to 9:00 A.M.')
Time_2 <- subset(data6, data6$Time == '5:30 P.M. to 6:00 P.M.')
Time_3 <- subset(data6, data6$Time == '7:30 P.M. to 8:00 P.M.')
Time_4 <- subset(data6, data6$Time == '9:30 P.M. to 10:00 P.M.')

data7 <- subset(data6, data6$avg_speed<55)
ggplot(data7, aes(x = Time, group = Time)) + geom_histogram(aes(fill = Time), stat="count") + ggtitle('Time for which Avg_Speed < 55 mph') + theme(plot.title = element_text(hjust = 0.5))
```

```
par(mfrow = c(1,1))
ggplot(data6, aes(group = Time,x = Time, y = avg_speed, fill = Time)) + geom_boxplot(outlier.shape = NA) + coord_cartesian(ylim = quantile(data5$avg_speed, c(0.011, 1.0))) + ggtitle('Avg_Speed Stratified by time') + theme(plot.title = element_text(hjust = 0.5))
ggplot(data6, aes(group = Time,x = Time, y = avg_speed, fill = Time)) + geom_boxplot() + ggtitle('Avg_Speed Stratified by time') + theme(plot.title = element_text(hjust = 0.5))


summary(Time_1$avg_speed)
summary(Time_2$avg_speed)
summary(Time_3$avg_speed)
summary(Time_4$avg_speed)
Time1_total <- sum(Time_1$total_hinderances)
Time2_total <- sum(Time_2$total_hinderances)
Time3_total <- sum(Time_3$total_hinderances)
Time4_total <- sum(Time_4$total_hinderances)
vector <- c(Time1_total, Time2_total, Time3_total, Time4_total)
time <- c('8:30 A.M. to 9:00 A.M.', '5:30 P.M. to 6:00 P.M.', '7:30 P.M. to 8:00 P.M.', '9:30 P.M. to 10:00 P.M.')
barplot(height=vector, names=time, col = rainbow(5), main = 'Count of total hinderances stratified by time')
# Which time has maximum Congestion
Time_1_congestion <- sum(Time_1$Congestion) # 4
Time_2_congestion <- sum(Time_2$Congestion) # 11
Time_3_congestion <- sum(Time_3$Congestion) # 4
Time_4_congestion <- sum(Time_4$Congestion) # 6


# Which time has maximum incidents ?
Time_1_incident <- sum(Time_1$Incidents) # 4
Time_2_incident <- sum(Time_2$Incidents) # 10
Time_3_incident <- sum(Time_3$Incidents) # 4
Time_4_incident <- sum(Time_4$Incidents) # 3
```

```
# Which time has maximum unconfirmed roadworks ?

Time_1_ur <- sum(Time_1$UR) # 246

Time_2_ur <- sum(Time_2$UR) # 120

Time_3_ur <- sum(Time_3$UR) # 249

Time_4_ur <- sum(Time_4$UR) # 333
```