STAT0006: ICA 1

Student number: 21000790

11th November 2022

Question 1

Import data set and check for missing values

The data set in csv format could be imported as df, displayed as below:

```
##    time direction food coffee   shoes school rain temperature traffic
## 1 22.52      from  yes      2 sandals     no 0.81       14.15     4.1
## 2 53.65        to  yes      1   boots    yes 1.57       10.21     1.5
## 3 53.31      from  yes      2 sandals     no 1.54       14.81    -0.3
## 4 61.87      from  yes      4   boots    yes 1.55       21.80    -1.4
## 5 34.32      from  yes      0 sandals     no 0.96       10.62     0.5
## 6 46.00      from  yes      3 sandals     no 2.73       16.06    -3.5
```

There are 9 predictor variables and 547 observations. Missing values could be checked as below:

```
sum(is.na(df))
```

```
## [1] 0
```

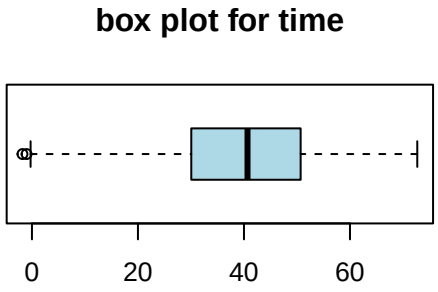
There is no missing data found in the data set.

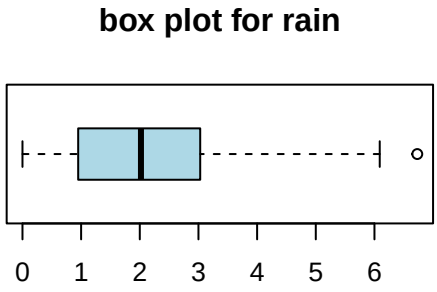
It is important to gain insights into the distribution and summary statistics of each variable, so that potential issues such as outlier can be removed. The numeric and categorical variables can be considered separately.

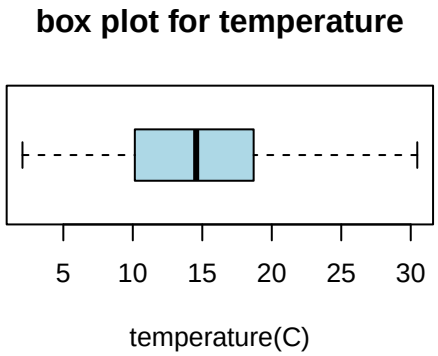
Numeric Variable

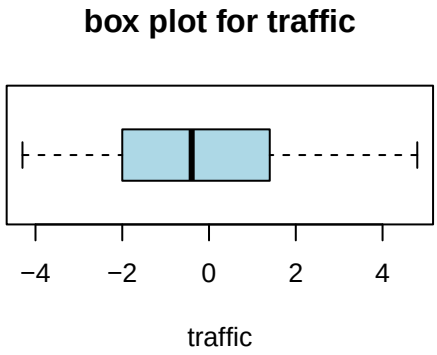
```
##       time             rain        temperature       traffic       
##  Min.   :-1.79   Min.   :0.000   Min.   : 2.06   Min.   :-4.3000  
##  1st Qu.:30.06   1st Qu.:0.950   1st Qu.:10.14   1st Qu.:-2.0000  
##  Median :40.67   Median :2.020   Median :14.56   Median :-0.4000  
##  Mean   :39.93   Mean   :2.052   Mean   :14.64   Mean   :-0.2558  
##  3rd Qu.:50.70   3rd Qu.:3.030   3rd Qu.:18.70   3rd Qu.: 1.4000  
##  Max.   :72.70   Max.   :6.730   Max.   :30.47   Max.   : 4.8000
```

We can see that the minimum value of "time" variable is negative, which is not valid. We could filter out all observations where time is negative, yet it was mentioned in the ICA Forum that this was not necessary. We can also observe the distributions of the numeric variables through a series of box plots:





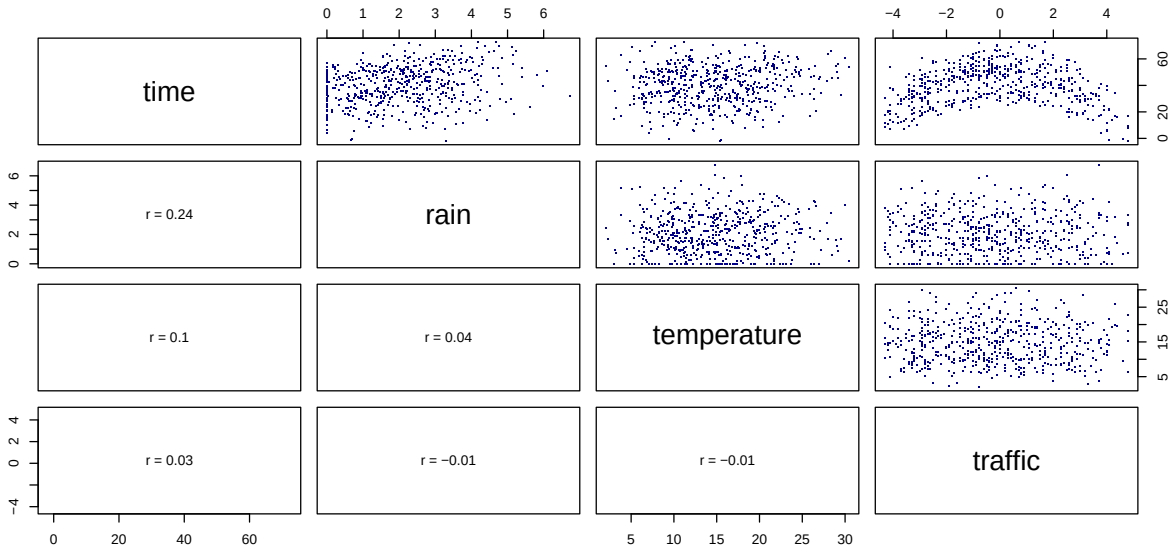




From the box plots of the numeric variables, there is little skewness in the distribution of each variable. Of all the variables, outliers can be observed in "rain" and "time".

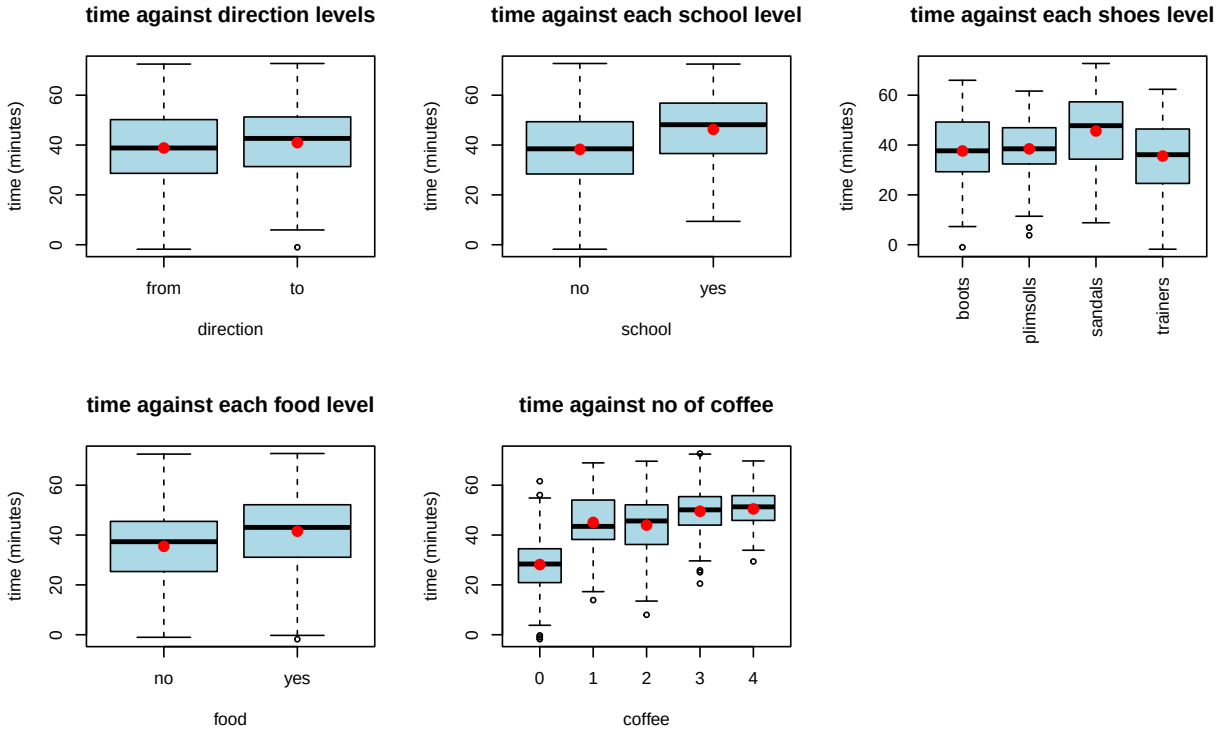
We can also observe if there is any relationship between each pair of variable from "time","rain","coffee","temperature" and "traffic". This will support our selection of predictor variables when it comes to building linear models later on. The scatter plot of the variables and their Pearson Correlation Coefficients can be seen from the matrix below.

Looking at the Pearson Correlation Coefficients, it is fair to claim that there is no strong linear relationship between all pairs of potential predictor variable (r = -0.01, r = 0.04, r = 0.02) and there is a stronger linear relationship between the pairs response-predictor variables (r = 0.24,r = 0.6, r= 0.1). However, from the scatter plot, it can be seen that there is strong non-linear relationship between "time" and "traffic" and a relatively low value of correlation coefficient (r = 0.03) between them.



Categorical Variable

The box plots below show how the distribution of "time" varies among each level of each categorical predictor variable. We can therefore investigate the effect of each predictor variable on commute time by looking at where the range (height of box) and the mean (the red dot) shift.



The following observations can be made:

- direction: The box plot suggests that direction has an effect on commute time. Traveling to work

generally leads to longer commute time.

- school: The box plot suggests that school has an effect on commute time. Having to drop kids off for school generally leads to longer commute time.
- shoes: The box plot suggests that shoes has an effect on commute time, as different types of shoes generally leads to different ranges of values (wider for sandals, narrower for plimsolls) as well as the average commute time (highest for sandals, lowest for trainers).
- food: The box plot suggests that food has an effect on commute time. Having to stop for food generally leads to longer commute time.
- coffee: The box plot suggests that coffee has a stronger effect on commute time. More coffees ordered generally leads to longer commute time.

Word count for Q1: insert word count here.

Question 2

Summary of Model 1a:

```
##
## Call:
## lm(formula = df$time ~ df$coffee, data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -34.881  -7.301   0.531   7.258  32.405
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  30.1988     0.7453   40.52   <2e-16 ***
## df$coffee     6.3360     0.3627   17.47   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 11.58 on 545 degrees of freedom
## Multiple R-squared:  0.359,  Adjusted R-squared:  0.3578
## F-statistic: 305.2 on 1 and 545 DF,  p-value: < 2.2e-16
```

Summary of Model 1b:

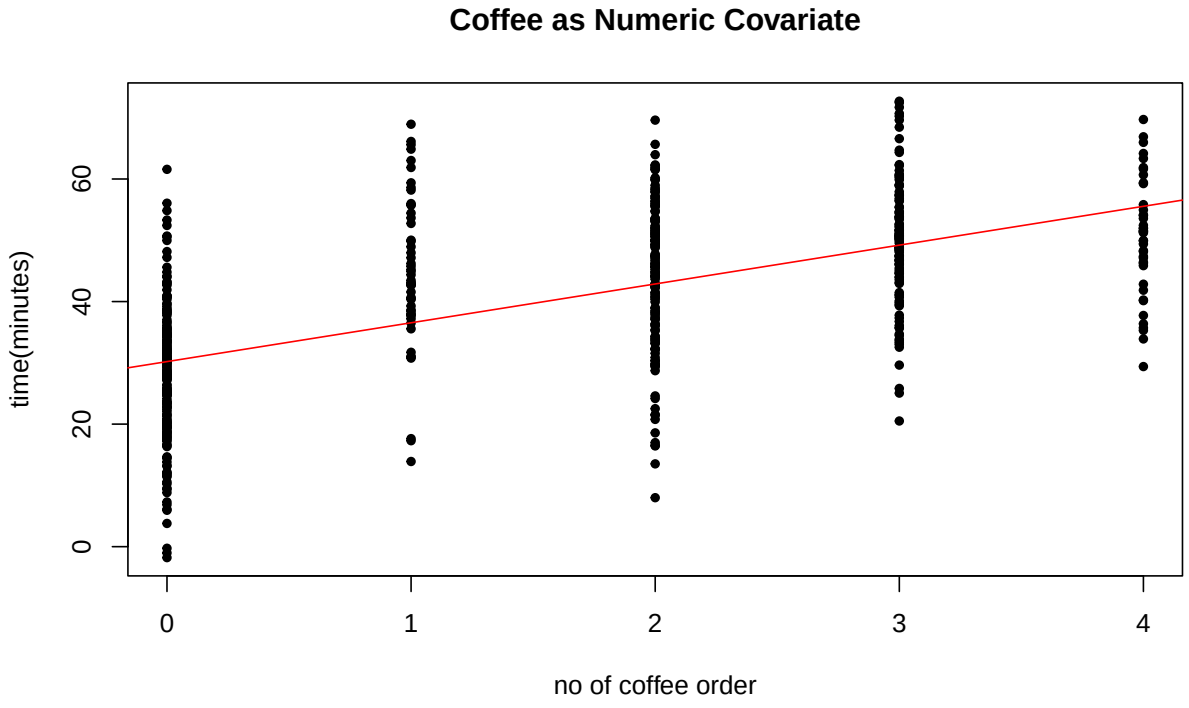
```
##
## Call:
## lm(formula = df$time ~ as.factor(df$coffee), data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -35.991  -6.809   0.572   6.996  33.432
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)             28.148      0.786  35.810   <2e-16 ***
```

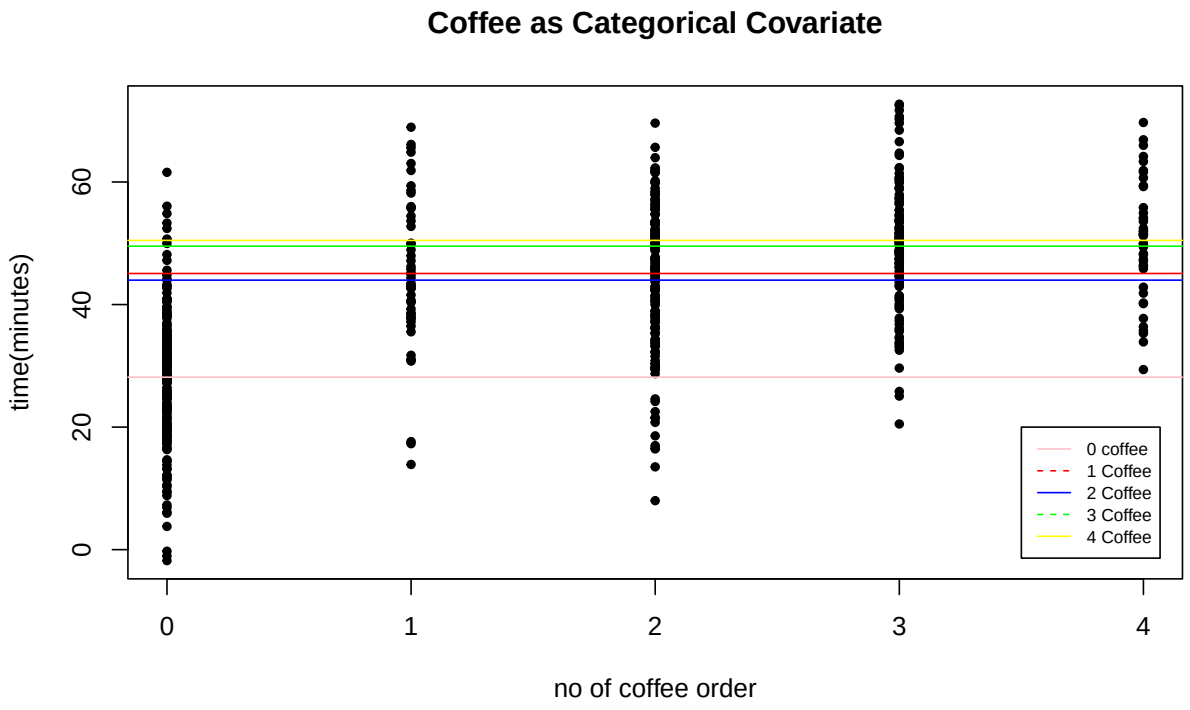
```
## as.factor(df$coffee)1   16.922      1.747   9.685   <2e-16 ***
## as.factor(df$coffee)2   15.833      1.232  12.852   <2e-16 ***
## as.factor(df$coffee)3   21.380      1.303  16.409   <2e-16 ***
## as.factor(df$coffee)4   22.322      1.891  11.806   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 11.14 on 542 degrees of freedom
## Multiple R-squared:   0.41,  Adjusted R-squared:  0.4057
## F-statistic: 94.16 on 4 and 542 DF,  p-value: < 2.2e-16
```

In this context, Model 1a is better because the "coffee" variable makes sense as a numeric variable: 3 coffees is in fact one unit larger than 4 coffees. Looking at the fitted model, it can be seen that if one more coffee is ordered, commute time increases by roughly 6.3 minutes, which is valid to say assuming that each coffee takes roughly the same amount of time to make.

Model 1a is also more effective when more than 4 cups of coffee are ordered, for example when future data collection include values larger than 4, where one extra cup will increase time by 6.3 minutes. Model 1b considers all cases for above 4 cups of coffee to have the same effect on time (the intercept), which is not necessarily true in reality.

Model 1b is not valid, as if two coffees are ordered instead of 1, it takes less commute time ($15.833 < 16.922$), which is not true in reality. This can be checked using the second plot from below:





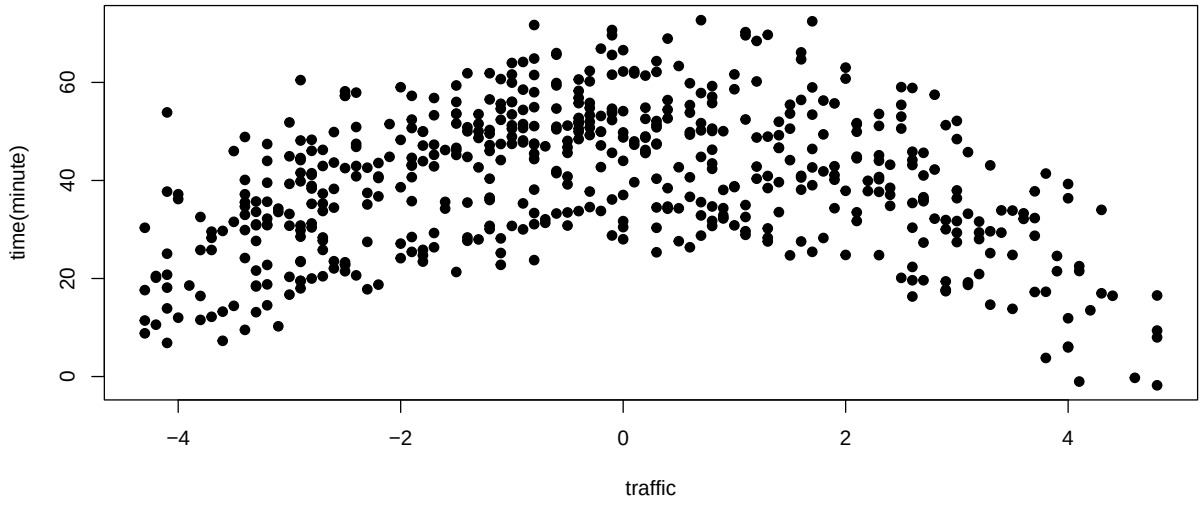
Word count for Q2: insert word count here.

Question 3

Model 2a can be summarised as below. It is concerning that the p-value of traffic is high (0.47>0.05), meaning the relationship between time and traffic is insignificant in the linear model.

```
##
## Call:
## lm(formula = df$time ~ df$traffic, data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -42.731  -9.669   0.640  10.755  32.580
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  39.9799     0.6224  64.240   <2e-16 ***
## df$traffic    0.2002     0.2769   0.723     0.47
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 14.46 on 545 degrees of freedom
## Multiple R-squared:  0.0009582,  Adjusted R-squared:  -0.0008749
## F-statistic: 0.5227 on 1 and 545 DF,  p-value: 0.47
```

To evaluate the linearity assumption, a scatter plot of time against traffic can be used:



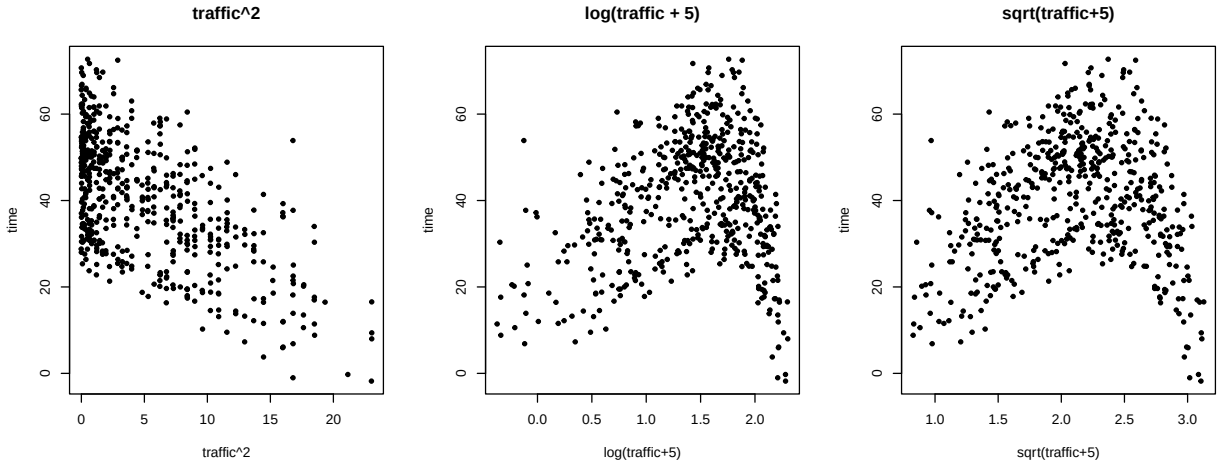
The scatter plot displays a strong non-linear (quadratic) relationship between commute time and traffic. This leads to the fact that assumption of linearity will not be appropriate in this scenario.

Word count for Q3: insert word count here.

Question 4

Adding 5 was necessary as traffic takes a range [-5,5], but log() and sqrt() only takes non-negative values. +5 will map traffic onto [0,10], which is compatible with these transformations

Plots of time against traffic^2, log(traffic+5) and sqrt(traffic+5) can be shown below. A strong linear relationship can be seen between time and traffic^2. That is, however, not the case for log(traffic+5) and sqrt(traffic+5), which makes sense as we can observe previously that there was a quadratic relationship between time and traffic in the EDA. traffic^2 will therefore be the best transformation for the linearity assumption of the model.



```
##
## Call:
## lm(formula = df$time ~ traffic1, data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -23.745  -8.847   0.923   7.300  34.120
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 48.59271    0.67225   72.28   <2e-16 ***
## traffic1    -1.71522    0.09234  -18.57   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 11.32 on 545 degrees of freedom
## Multiple R-squared:  0.3876, Adjusted R-squared:  0.3865
## F-statistic:   345 on 1 and 545 DF,  p-value: < 2.2e-16
```

As we recall the two model:

- Model 2a: time = 39.9799 + 0.2002*traffic
- Model 2b: time = 53.2512 - 7.0731*(traffic^2)

In general, the estimated intercept increased from 39.9799 to 53.2512, while the estimated gradient changed in sign and increased from 0.2002 to 7.0731.The p-value of the gradient coefficient also decreased dramatically from 0.47 to <2e-16.

Given the stronger linear relationship after transforming traffic, Model 2b has a significant gradient coefficient after fitting. Compared to Model 2a where p-value is large ($0.47 > 0.05$), we fails to reject the null hypothesis that the gradient coefficient is 0.

Model 2b also displays a better R-squared value (0.3607) compared to Model 2a (0.0009582). This means Model 2b fits the observations better than Model 2a.

Word count for Q4: insert word count here.

Question 5

```
##
## Call:
## lm(formula = df$time ~ df$direction + df$food + df$shoes + df$school +
##     df$coffee + df$rain + df$temperature + df$traffic, data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -31.390  -5.690   1.565   6.575  22.383
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)       13.54313    1.60331   8.447 2.82e-16 ***
## df$directionto     0.26340    0.80028   0.329  0.74218
## df$foodyes         5.61034    0.90620   6.191 1.19e-09 ***
## df$shoesplimsolls  2.28389    1.22686   1.862  0.06321 .
```

```
## df$shoessandals    8.05173    1.00613   8.003 7.58e-15 ***
## df$shoestrainers  -0.94277    1.13831  -0.828  0.40791
## df$schoolyes       9.49593    0.98713   9.620  < 2e-16 ***
## df$coffee          6.49993    0.29284  22.196  < 2e-16 ***
## df$rain            2.30748    0.28878   7.990 8.29e-15 ***
## df$temperature     0.18567    0.06987   2.657  0.00811 **
## df$traffic         0.18793    0.17807   1.055  0.29175
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9.28 on 536 degrees of freedom
## Multiple R-squared:  0.5953, Adjusted R-squared:  0.5878
## F-statistic: 78.86 on 10 and 536 DF,  p-value: < 2.2e-16
```

From the summary, it can be seen that 10 regression coefficients are estimated.

The regression coefficients can be interpreted in context:

- Rain coefficient: time is expected to increase by approximately 2.31 minutes rain increase by 1mm, given all coffee, temperature, traffic as constant and holding shoes, food, school, and direction being constant.
- shoessandals dummy variable: time is expected to increase by approximately 8.05 minutes if Sandals are worn instead of other types of shoes, given all other covariates staying constant
- Intercept: the expected amount of time taken is approximately 13.54 minutes, given the individual travel from work, did not buy food, wore boots, did not drop kids at school, did not stop for coffee, average temperature was zero, and traffic index was 0.

Word count for Q5: insert word count here.

Question 6

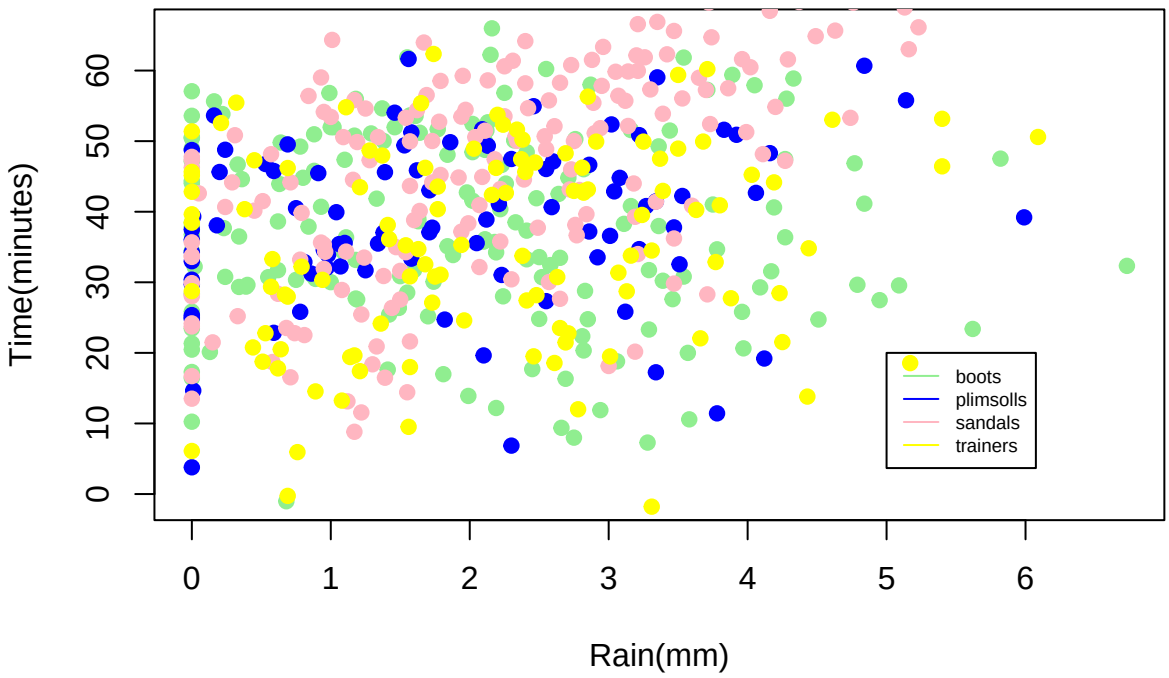
After fitting in Model 3, the estimated coefficients of the predictor covariates means that if the covariate increases by 1, the response variable will increase by an according value of the coefficient.

When time is scaled by 60, to change from minutes to seconds, all the coefficients must also be multiplied with 60 to preserve the linear relationship between the predictor and the response covariates. That relationship does not change.

Word count for Q6: insert word count here.

Question 7

Plot of time against rain, varying with shoes type:



It can be seen that the scatter plots of time against rain vary across different of shoes. For example, that of sandals (light pink) is generally higher than that of boots (light green). It is thus fair to say that shoes has an effect on the time-rain relationship.

An interaction term can be added to the model:

```
##
## Call:
## lm(formula = df$time ~ df$direction + df$food + df$shoes + df$school +
##     df$coffee + df$rain + df$temperature + df$traffic + (df$rain *
##     df$shoes), data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -29.167  -4.912   1.749   6.213  16.197
##
## Coefficients:
##                            Estimate Std. Error t value Pr(>|t|)
## (Intercept)                18.33537    1.63693  11.201  < 2e-16 ***
## df$directionto              0.37584    0.73252   0.513 0.608110
## df$foodyes                  5.39091    0.82991   6.496  1.9e-10 ***
## df$shoesplimsolls          -1.99594    1.87701  -1.063 0.288099
## df$shoessandals            -5.81685    1.66448  -3.495 0.000514 ***
## df$shoestrainers           -2.92547    1.87068  -1.564 0.118445
## df$schoolyes               10.21050    0.90772  11.249  < 2e-16 ***
## df$coffee                   6.41664    0.26809  23.935  < 2e-16 ***
## df$rain                    -0.22135    0.45357  -0.488 0.625740
```

```
## df$temperature             0.20392    0.06398   3.187 0.001521 **
## df$traffic                 0.08461    0.16341   0.518 0.604808
## df$shoesplimsolls:df$rain  2.16244    0.78991   2.738 0.006397 **
## df$shoessandals:df$rain    6.67188    0.67009   9.957  < 2e-16 ***
## df$shoestrainers:df$rain   1.13980    0.73933   1.542 0.123750
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.49 on 533 degrees of freedom
## Multiple R-squared:  0.6633, Adjusted R-squared:  0.655
## F-statistic: 80.75 on 13 and 533 DF,  p-value: < 2.2e-16
```

The new linear model accounts for the effects of shoes on the relationship between rain and time, which supported the suggestion. If we isolate coefficient of rain by partial differentiation, the coefficient would be beta*(shoes type), where beta is the estimated coefficient.This varies with the dummy variable of the shoes type.

By isolating the coefficient of interest, each of the additional covariate can be interpreted as follow:

- shoesplimsolls & rain coefficient in model4: when the amount of rainfall increase by 1mm, the expected change in commute time is approximately 2.16 minutes, given they wear plimsolls.
- shoessandals & rain coefficient in model4: when the amount of rainfall increase by 1mm, the expected change in commute time is approximately 6.67 minutes, given they wear sandals.
- shoestrainers & rain coefficient in model4: when the amount of rainfall increase by 1mm, the expected change in commute time is approximately 1.14 minute, given they wear trainers.

Word count for Q7: insert word count here.

Question 8

Following the previous analysis, I would propose the model to be the same as Model 4, apart from two modifications:

- Remove direction as it p-value is 0.608110, which implies that the covariate has an insignificant relationship with the response variable in the model (We fail to reject the null hypothesis that the coefficient is zero).
- Replace traffic with traffic^2 as there is a stronger linear relationship between time and traffic^2.

No further interaction term was included because we are not well-informed of the relationships between the other predictor covariates. Adding more terms could also increase the risk of overfitting.

The inflated p-values of the predictor covariates (rain and dummy variables of shoes) is not a concern because it is the result of the interaction term. It can also be seen that the coefficient of traffic1 (or traffic^2) has p-value of $< 2e^{-16}$, which is significant.

Model 5 can be fitted as below. It can be seen that the Adjusted R-squared did not drop by a large amount compared to the Multiple R-squared value, implying that this model does not overfit.

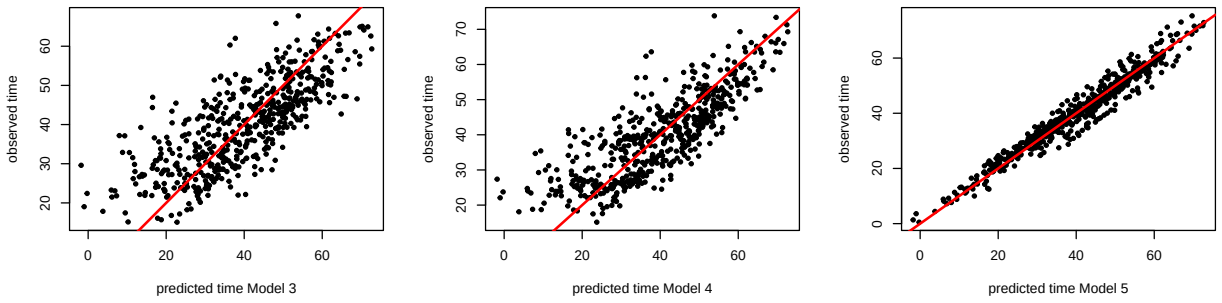
```
##
## Call:
## lm(formula = df$time ~ df$food + df$shoes + df$school + df$coffee +
```

```
##     df$rain + df$temperature + traffic1 + df$rain * df$shoes,
##     data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7.8659 -1.9812 -0.2182  1.5687  9.9222
##
## Coefficients:
##                            Estimate Std. Error t value Pr(>|t|)
## (Intercept)                25.59920    0.61963  41.313  < 2e-16 ***
## df$foodyes                  5.31698    0.31290  16.992  < 2e-16 ***
## df$shoesplimsolls          -0.81606    0.70922  -1.151  0.25039
## df$shoessandals            -2.00416    0.63042  -3.179  0.00156 **
## df$shoestrainers           -1.07519    0.70708  -1.521  0.12895
## df$schoolyes                9.89172    0.34289  28.848  < 2e-16 ***
## df$coffee                   5.91503    0.10143  58.314  < 2e-16 ***
## df$rain                     0.51657    0.17155   3.011  0.00272 **
## df$temperature              0.18351    0.02417   7.594 1.40e-13 ***
## traffic1                   -1.49927    0.02649 -56.588  < 2e-16 ***
## df$shoesplimsolls:df$rain   1.64773    0.29857   5.519 5.33e-08 ***
## df$shoessandals:df$rain     4.94892    0.25445  19.450  < 2e-16 ***
## df$shoestrainers:df$rain    0.46256    0.27914   1.657  0.09809 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.208 on 534 degrees of freedom
## Multiple R-squared:  0.9518, Adjusted R-squared:  0.9507
## F-statistic: 879.2 on 12 and 534 DF,  p-value: < 2.2e-16
```

Word count for Q8: insert word count here.

Question 9

The scatter plots of observed-predicted values of time between three models can be compared as below:



If all the points lie perfectly on the y=x line, it means that all the predicted values are the same as the observed values, and the model has highly accurate predictability. Therefore, the more the points align with the line, the better the model. It can be concluded that:

- The plots of Model 3 and Model 4 have larger spread around y=x compared to Model 5, implying less accurate prediction.
- Model 5 is therefore the best of the three, which agrees with the analysis and selection of its predictor covariates in Question 8.

Word count for Q9: insert word count here.