STAT0006: ICA 1

Student number: 21026576

Deadline for submission is 12 noon on Monday 14th November 2022

Submit a (knitted) pdf document only

Question 1

This is a dataset which contains collective data about a random subset of individuals' commutes over the last five years. This explanatory analysis is conducted for comprehension of the effects of various factors on the individual's length of journey, and further estimates the total time taken for an individual to get to/from work. 8 factors that are believed to have varying impacts on commute time are as follows; the direction of travel either to or from work; the number of cups of coffee they drank if they had stopped at the cafe; types of shoes worn; if they had stopped by their kid's school on the way; the amount of rainfall in mm; the average temperature in degrees Celsius; and the traffic index during their journey.

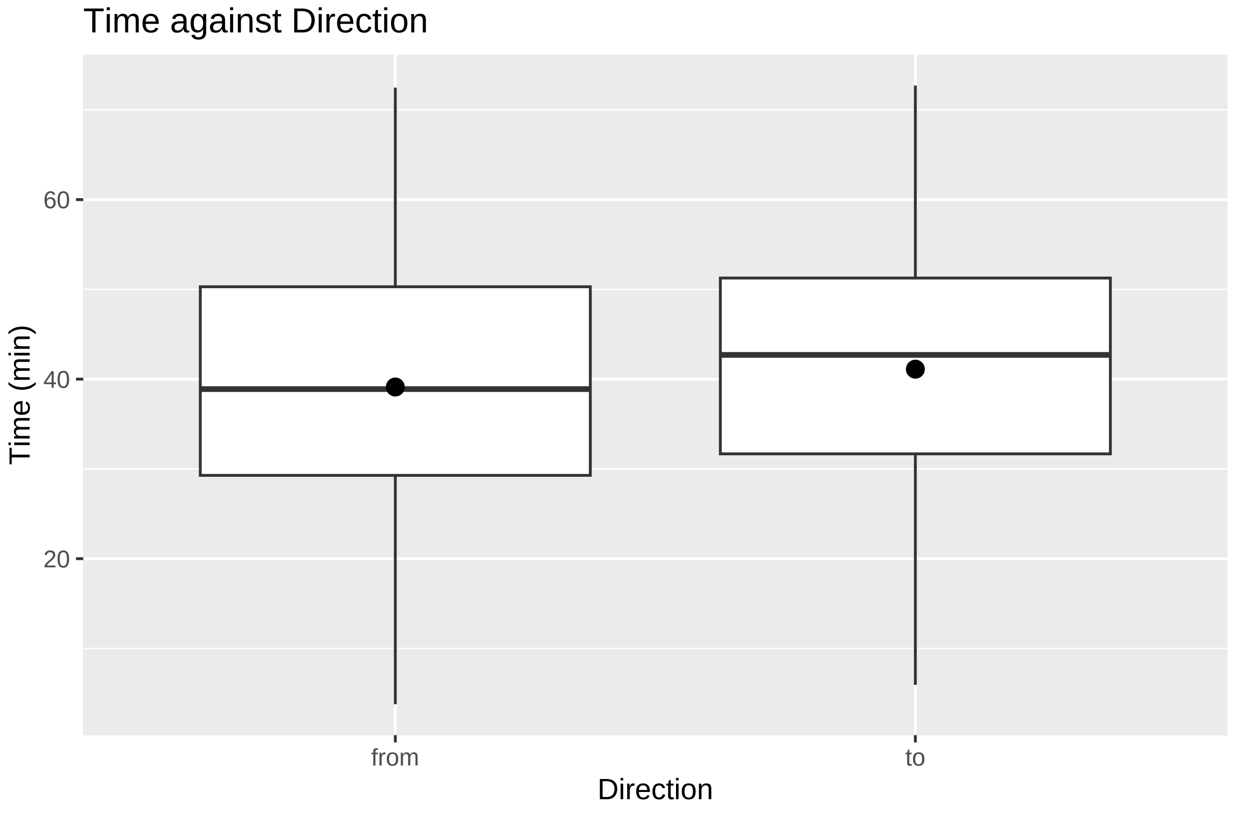


Fig 1

Fig1: A boxplot comparing the commute time from and to work. This indicates that the average journey time from work (38.8minutes) is faster than the average journey time to work (41.0minutes). The mean is represented by black dots in all of the plots.

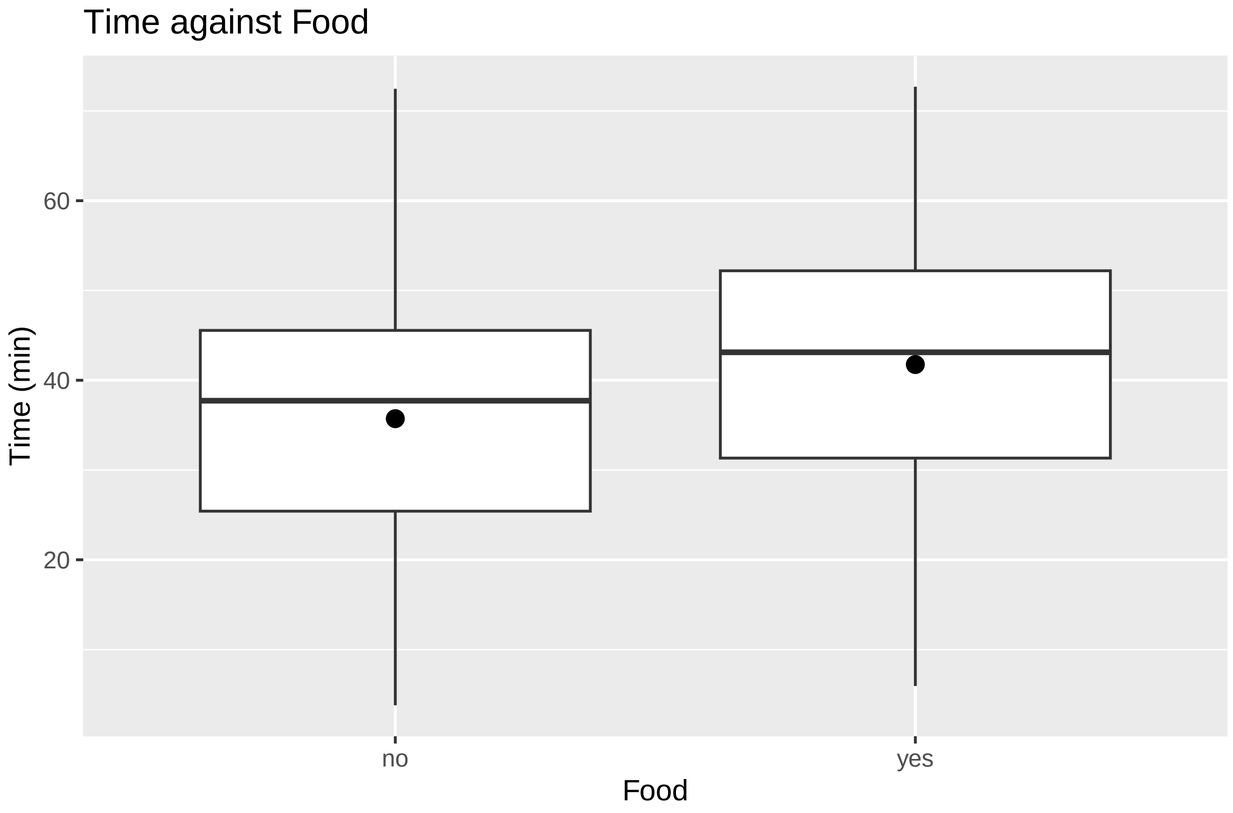


Fig 2

Fig2: A boxplot between the commuting time and whether they stopped for food or not. It demonstrates a longer average commuting time when they had stopped for food by 6minutes compared to when they had not stopped for food.

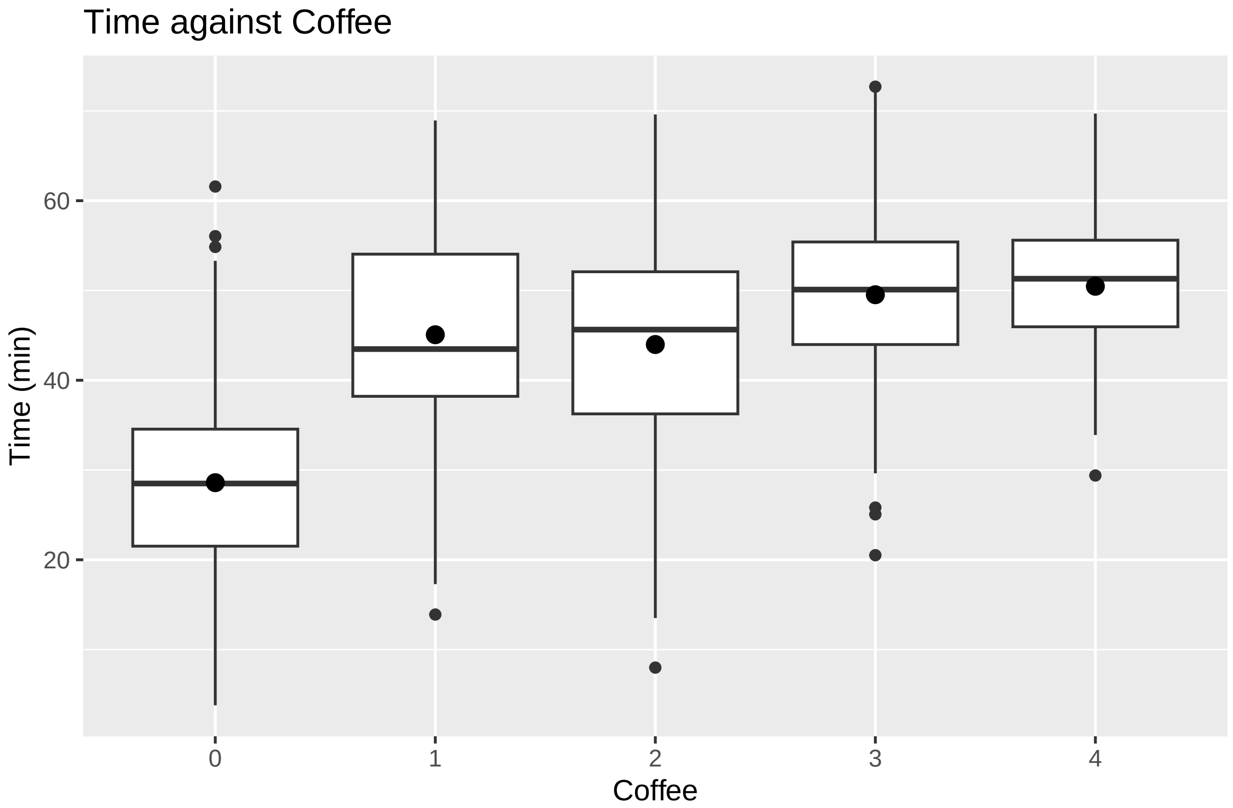


Fig 3

Fig3: A boxplot showing commute time for each number of cups of coffee. The average commute time if they had not stopped for a coffee is around 28minutes while the marginal time taken to get just one cup of coffee increased the commute time by approximately 17minutes. There is an increasing trend of time taken against the number of coffee with an average of 50minutes for four cups of coffee.

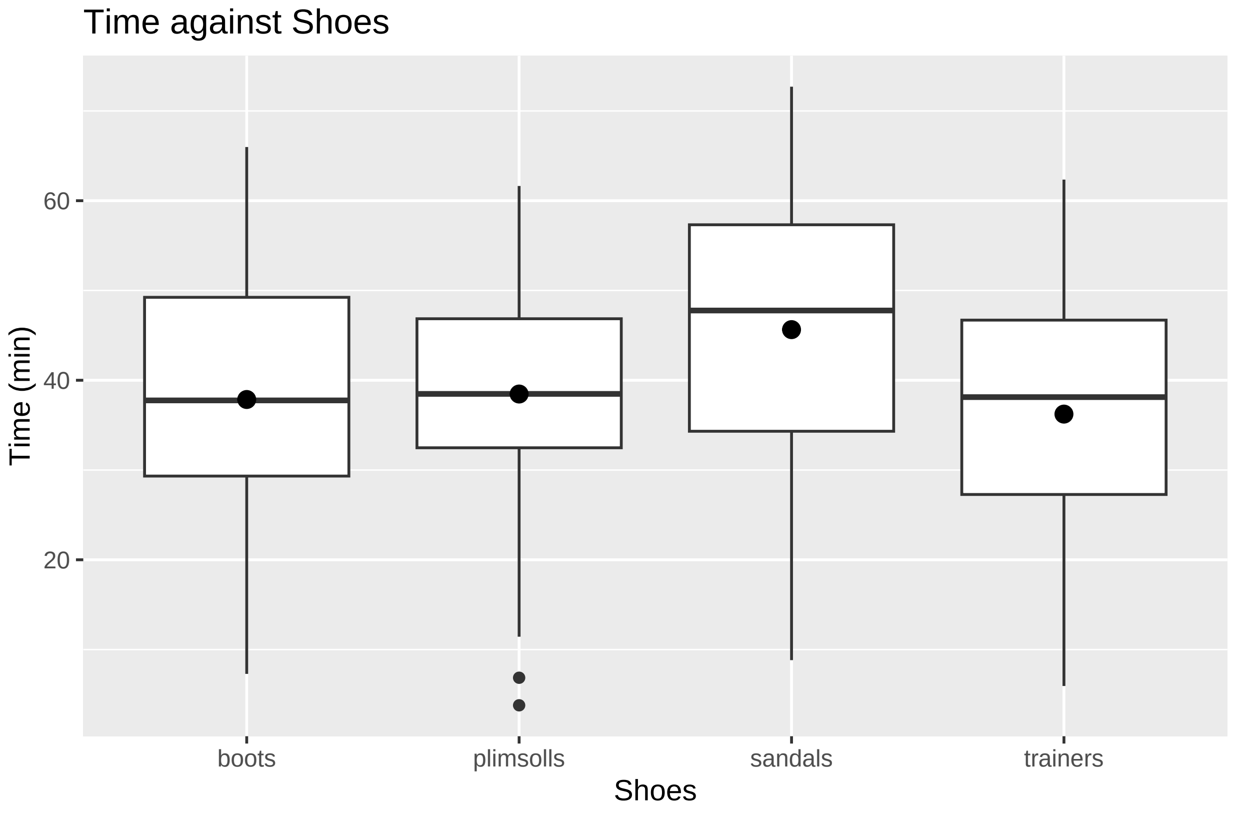


Fig 4

Fig4: This contains boxplots for the time against each type of shoe worn by the individual. From this plot, wearing sandals led to the highest average commuting time of 45.6minutes compared to other footwear. Trainers can be seen with the lowest average commuting time of 35.6minutes.

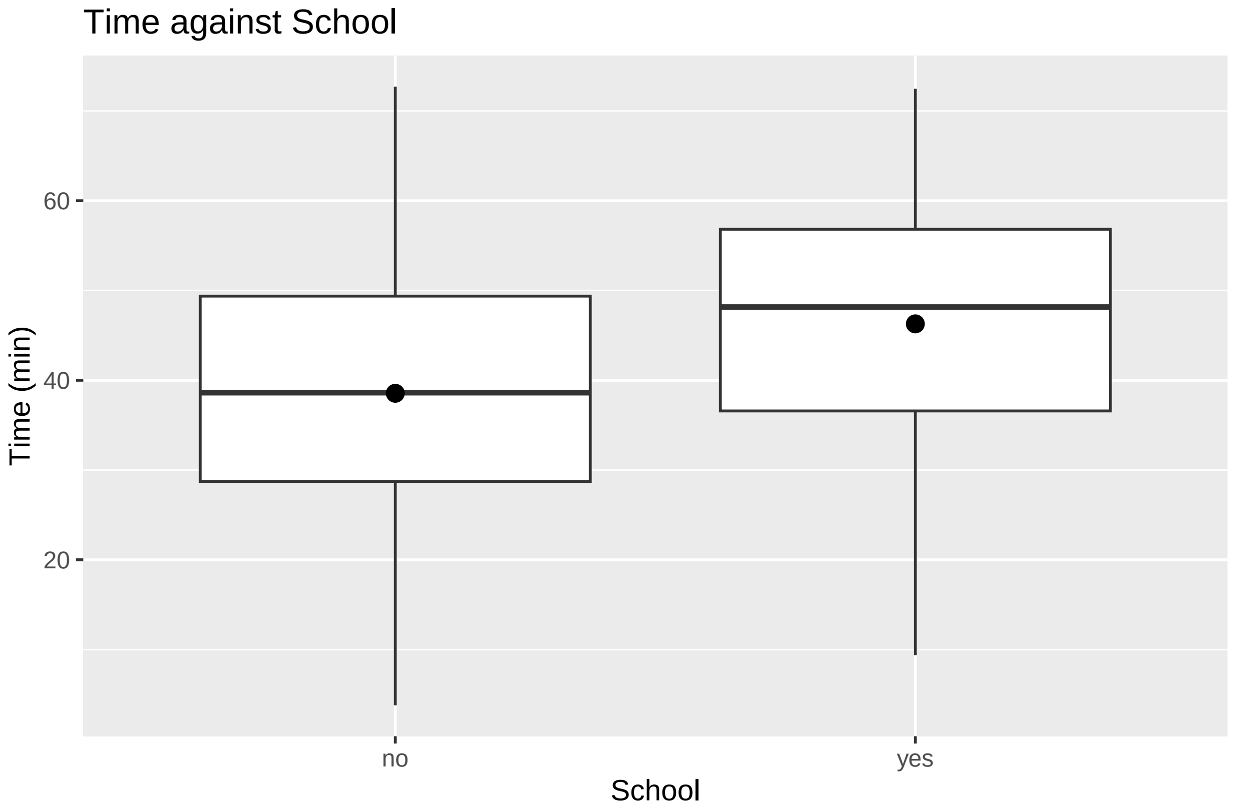


Fig 5

Fig5: A boxplot comparing the difference in time taken whether they had stopped by their children's school on their journey. This shows that visiting the school would increase their average commuting time from around 38minutes to 46minutes.

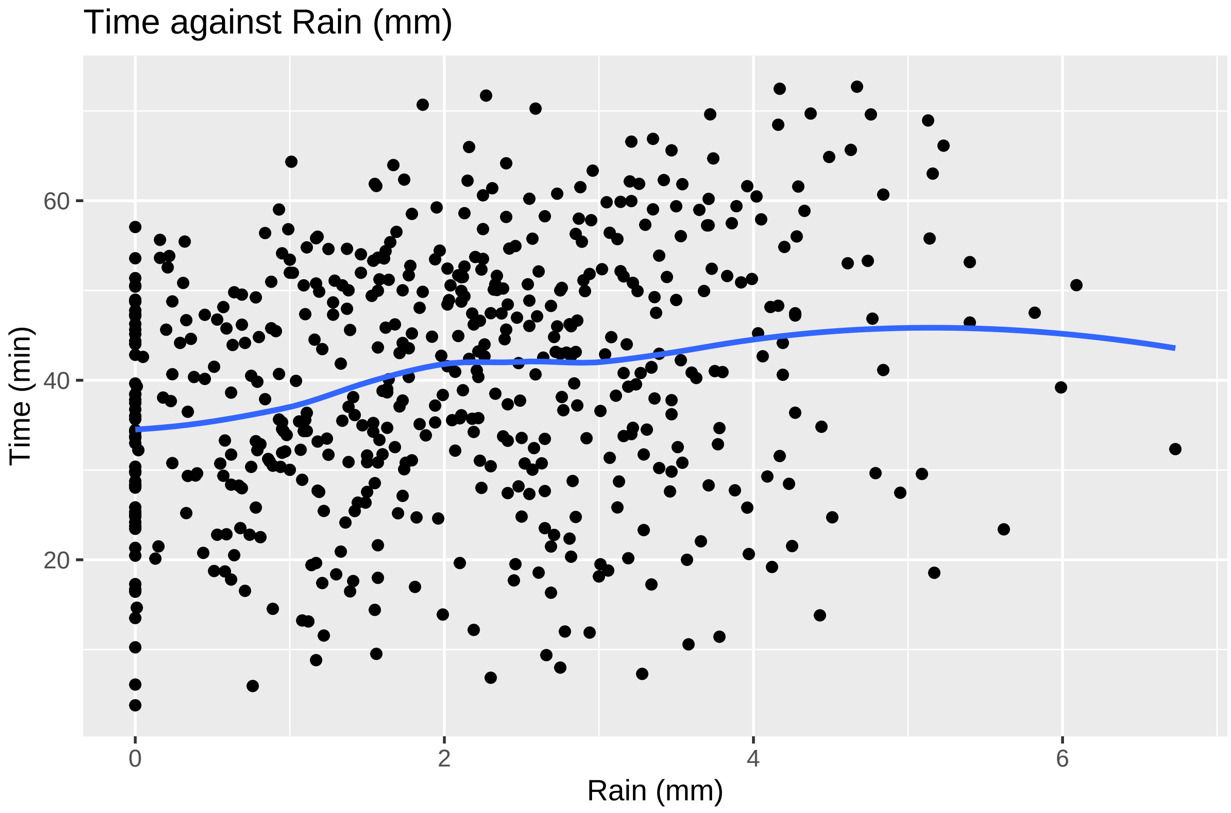


Fig 6

Fig6: A scatter plot demonstrating a weak positive relationship of time against the amount of rainfall in mm using the blue trend line.

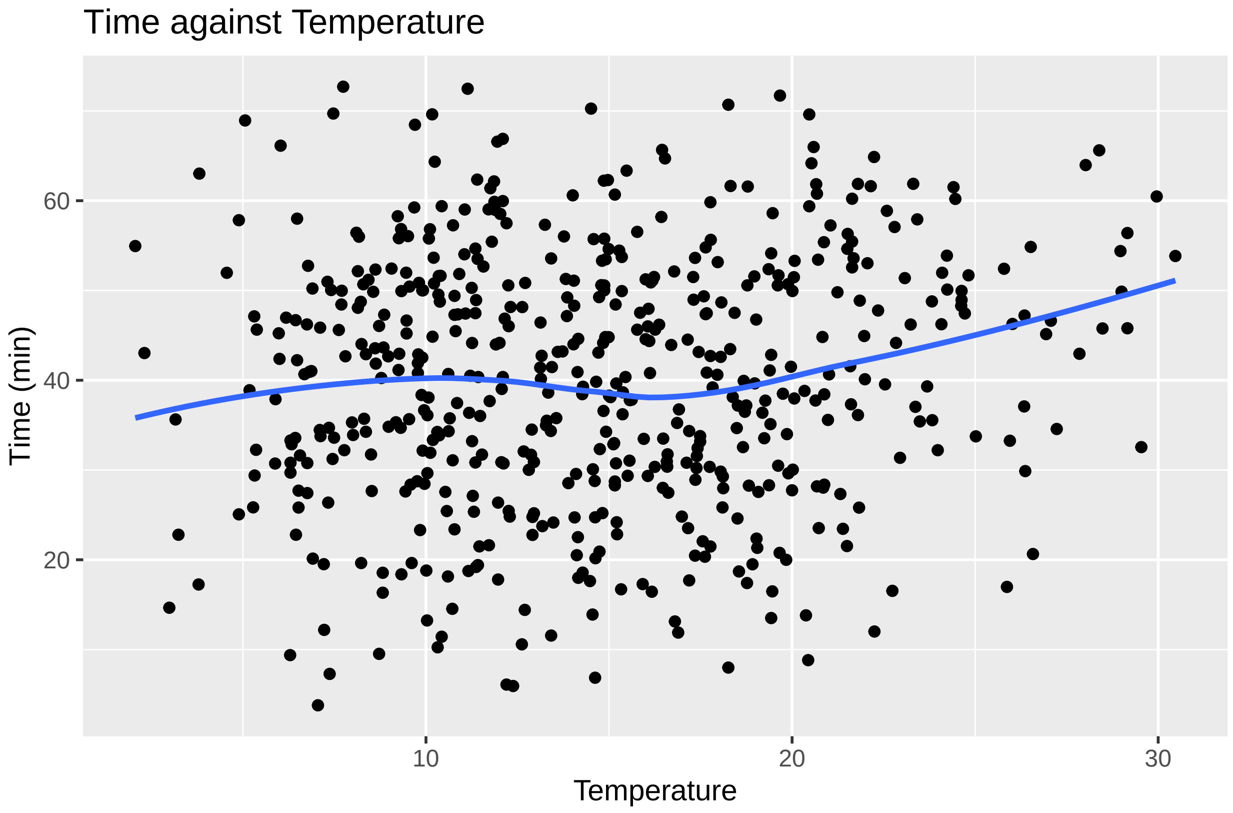


Fig 7

Fig7: This figure has a weak positive relationship between temperature and time which is shown by the blue trend line.

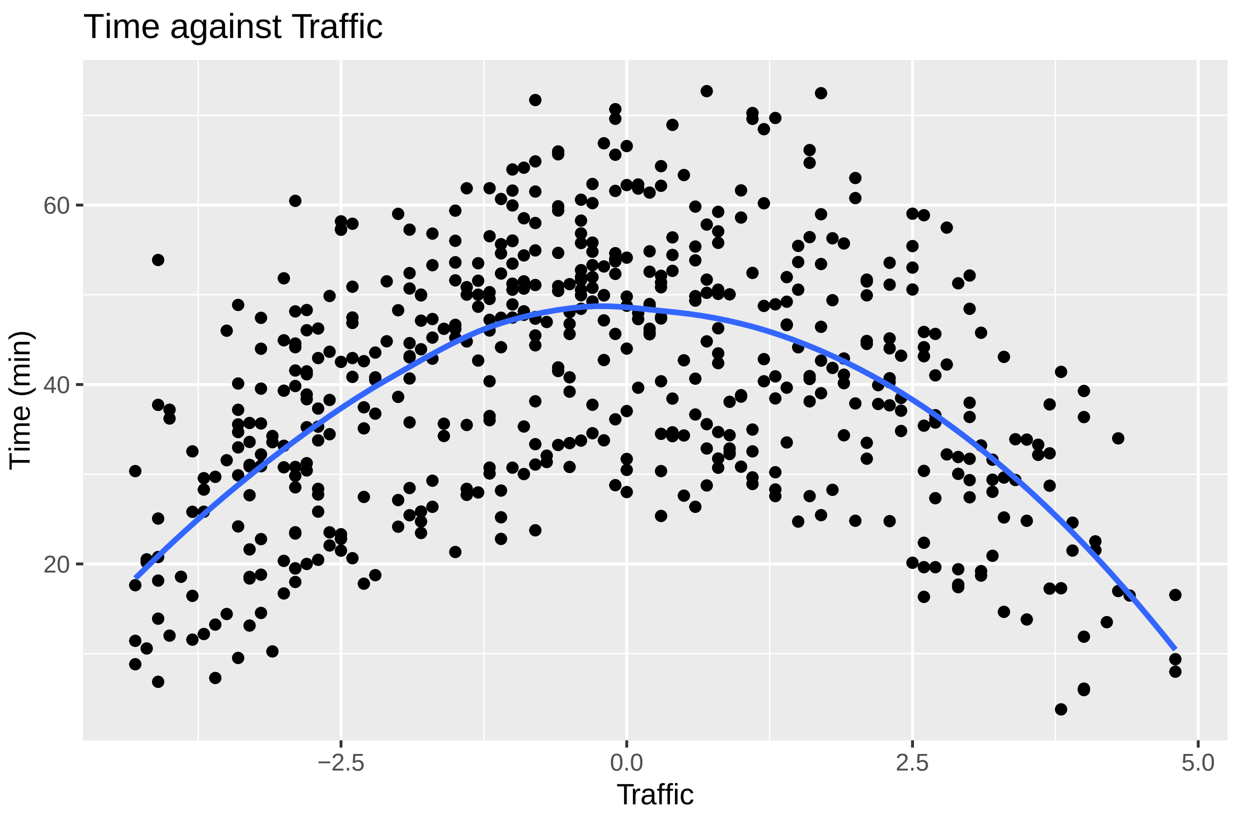


Fig 8

Fig8: A scatter plot demonstrating the effect of traffic in the numeric index on time. There is evidence of shorter commute times when there is either little or a lot of traffic, whereas a longer commute time is when there is moderate traffic.

To conclude, all of the variables establish some pattern of a relationship with respect to the journey time and should be further examined. There are also some entries stating a negative commute time which is impractical in the real world, therefore, we would be omitting these values for a more reliable outcome.

Word count for Q1: 500

Question 2

```
##
## Call:
## lm(formula = time ~ coffee, data = commute)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -34.988  -7.156   0.234   7.261  32.155
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  30.5914     0.7367   41.52   <2e-16 ***
## coffee        6.1933     0.3575   17.32   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 11.38 on 542 degrees of freedom
## Multiple R-squared:  0.3564, Adjusted R-squared:  0.3552
## F-statistic: 300.1 on 1 and 542 DF,  p-value: < 2.2e-16
```

```
##
## Call:
## lm(formula = time ~ as.factor(coffee), data = commute)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -35.991  -6.826   0.441   6.923  32.990
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)         28.5896     0.7787  36.714   <2e-16 ***
## as.factor(coffee)1  16.4802     1.7206   9.578   <2e-16 ***
## as.factor(coffee)2  15.3914     1.2151  12.667   <2e-16 ***
## as.factor(coffee)3  20.9381     1.2847  16.298   <2e-16 ***
## as.factor(coffee)4  21.8799     1.8615  11.754   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 10.96 on 539 degrees of freedom
## Multiple R-squared:  0.4064, Adjusted R-squared:  0.402
## F-statistic: 92.26 on 4 and 539 DF,  p-value: < 2.2e-16
```

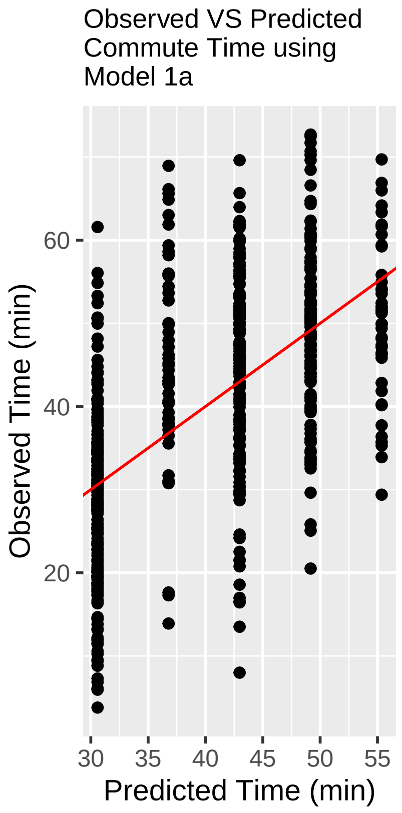
Model 1b is preferable because the predictions it produces are more representative of the distribution of commute time for each amount of coffee.

Figure3 shows the mean commute time for not stopping by the cafe(coffee=0) is much lower than stopping for coffee(coffee=1,2,3,4). Thus, we should not assume a linear relationship between cups of coffee and time.

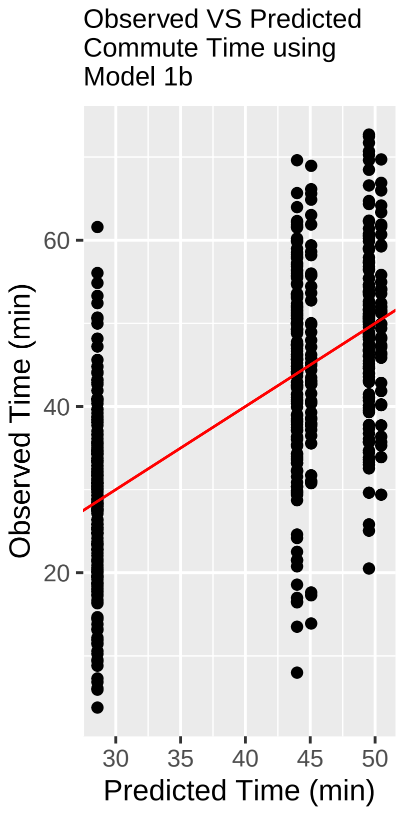
Model 1a, each marginal increase in time from the marginal increase in coffee are equal, shown in the table below. This model assumes a linear relation between coffee and time.

Model 1b, the predicted commute time for coffee=0 is lower at 28minutes while the predicted time for the journey with coffee=1,2,3,4 ranges between 40 to 50minutes as seen in the table. This is similar to the distributions of means observed in Figure3.

Moreover, the calculated root mean squared error further supports the argument that Model 1b is better as the value of residuals is lower.



Plot 1



Plot 2

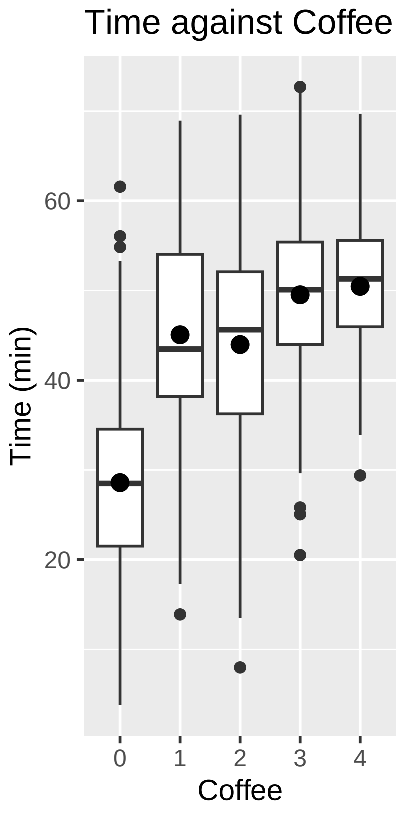


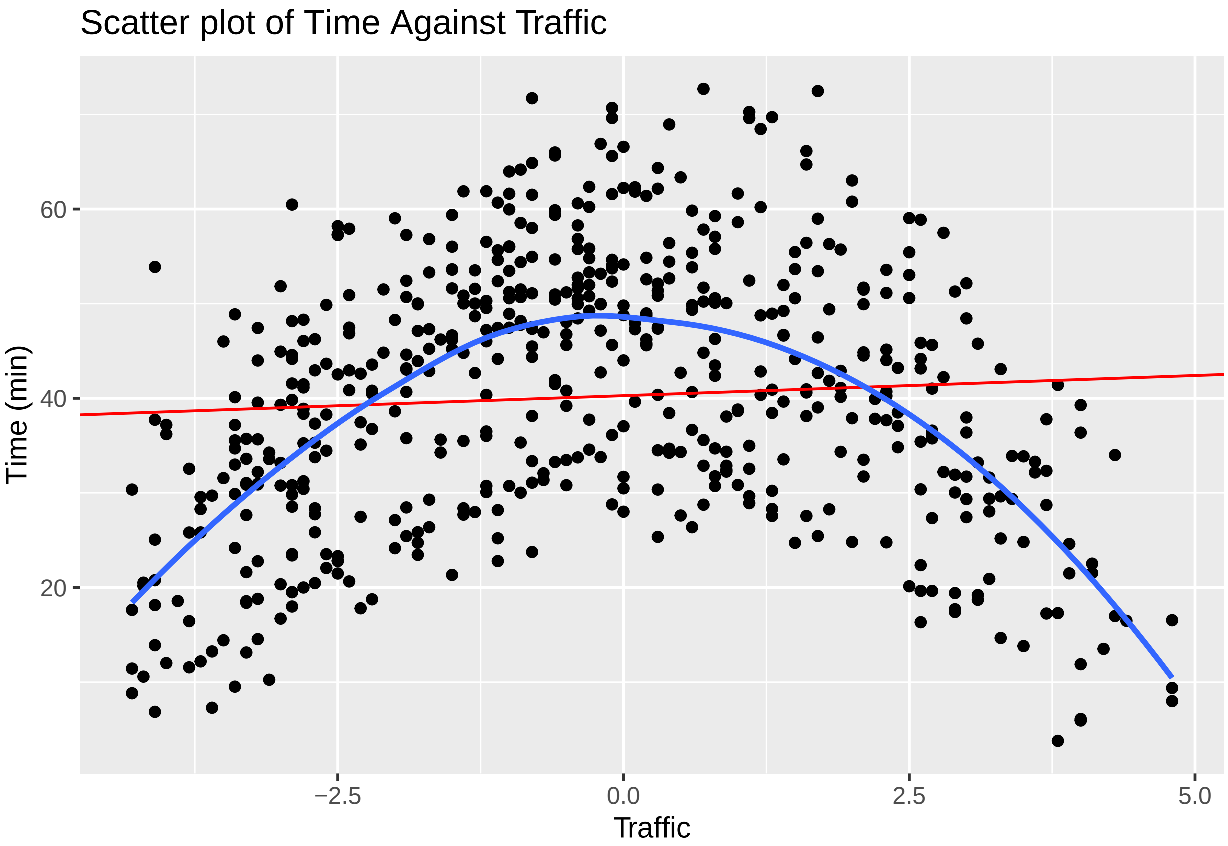
Fig 3

	Coffee	Model_1a_Predicted_Time	Model_1b_Predicted_Time
1	0	30.59	28.59
2	1	36.78	45.07
3	2	42.98	43.98
4	3	49.17	49.53
5	4	55.36	50.47

Word count for Q2: 150

Question 3

```
##
## Call:
## lm(formula = time ~ traffic, data = commute)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -38.105  -9.841   0.613  10.522  32.127
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  40.2748     0.6116  65.847   <2e-16 ***
## traffic       0.4264     0.2744   1.554    0.121
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 14.15 on 542 degrees of freedom
## Multiple R-squared:  0.004434,   Adjusted R-squared:  0.002597
## F-statistic: 2.414 on 1 and 542 DF,  p-value: 0.1209
```



Yes, there are concerns about the linearity of this model. The scatter plot of journey time against traffic shows points scattered along a curve. This is highlighted by the blue trend line.

The red line shows the regression line from Model 2a. In this plot, we can see that the distribution of points does not follow the red line.

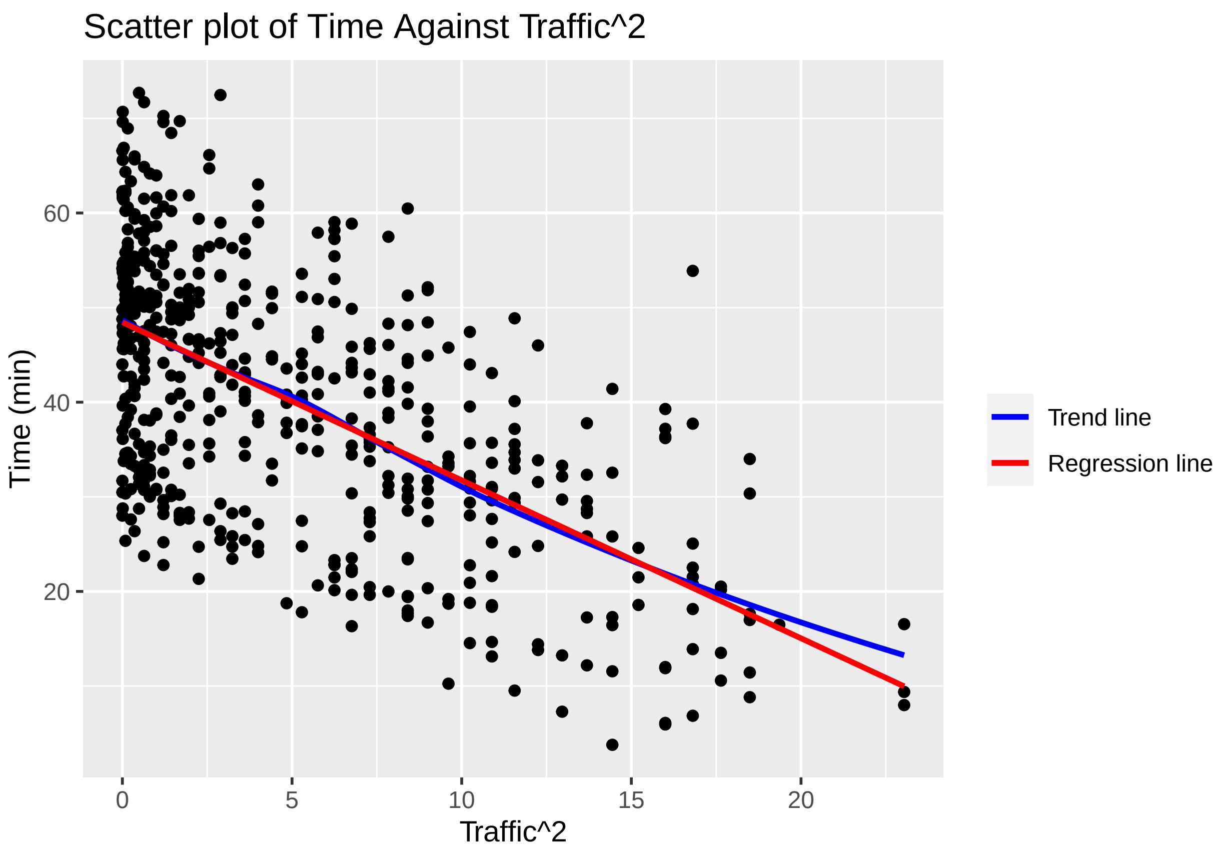
Furthermore, the p-value for the coefficient of traffic is high and does not show a significant linear correlation between traffic and time.

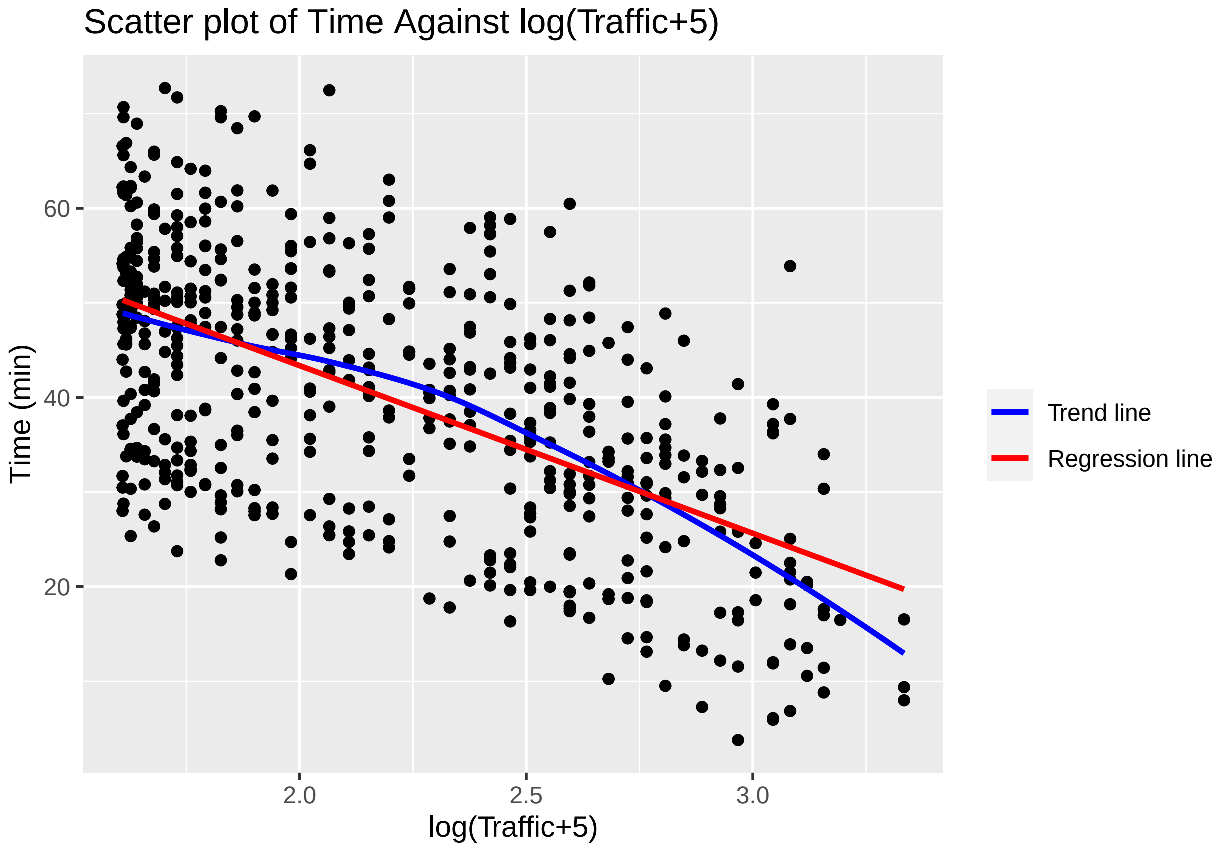
Thus we should not assume linearity for this model.

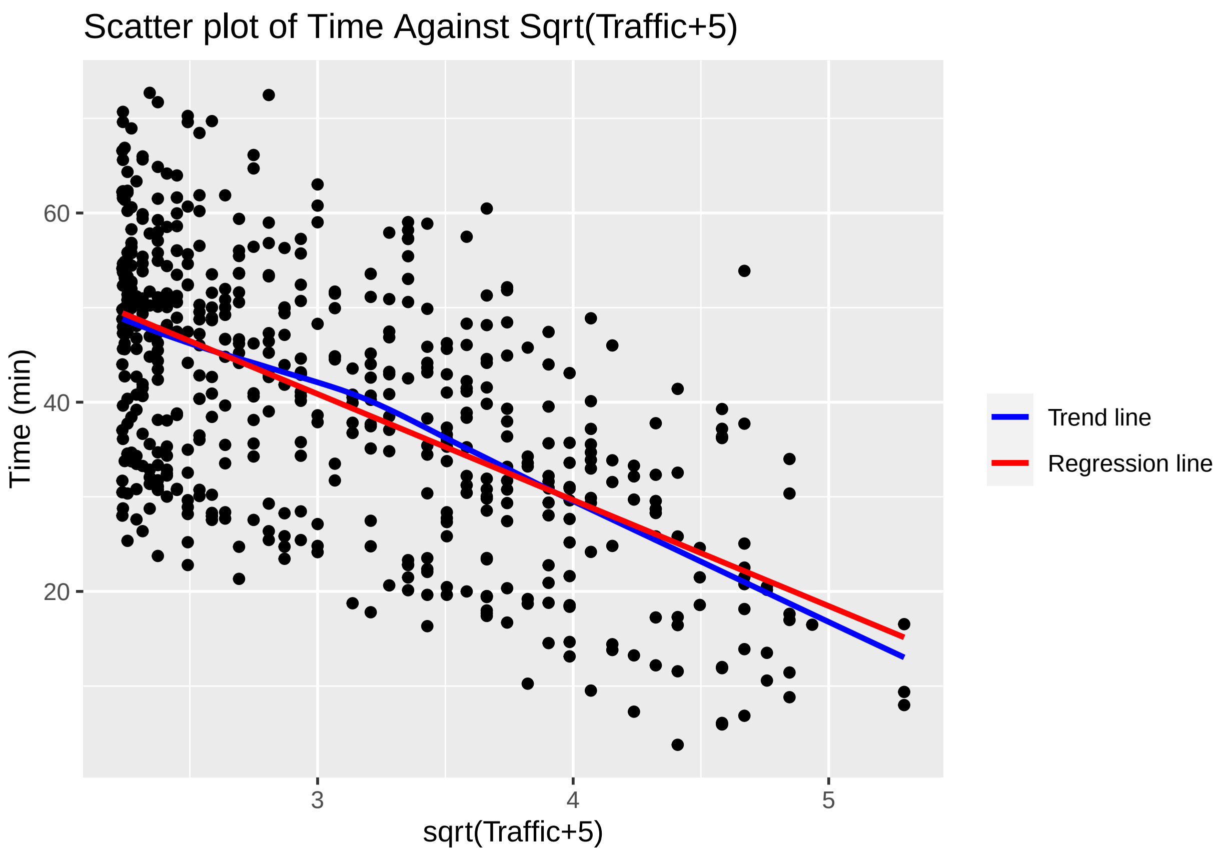
Word count for Q3: 90

Question 4

- Because the range of the numeric index for traffic is between -5 and 5, the values below 0 using the second transformation (*log* function) and third transformation(*square root* function) would result in an imaginary number.


Scatter plot of Time Against Traffic^2
Time (min)
60
40
20
0
5
10
15
20
Traffic^2
Trend line
Regression line


Scatter plot of Time Against log(Traffic+5)
Time (min)
60
40
20
2.0
2.5
3.0
log(Traffic+5)
Trend line
Regression line



The transformation traffic^2 is the most suitable as the distribution of points best follows a straight line with the trend line and best-fit line coinciding the most.

The most suitable transformation should have points concentrated along a straight line with points evenly distributed above and below it.

This can be more easily illustrated by the trend line and best fit line.

The transformed variable that would be most suitable to assume linearity for should have its best fit line coinciding with the trend line, showing that points are scattered along a linear path.

Thus, from the three plots, the first transformation (traffic^2) is the most suitable transformation.

```
##
## Call:
## lm(formula = time ~ traffic, data = commute)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -38.105  -9.841   0.613  10.522  32.127
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  40.2748     0.6116  65.847   <2e-16 ***
## traffic       0.4264     0.2744   1.554    0.121
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 14.15 on 542 degrees of freedom
## Multiple R-squared:  0.004434,   Adjusted R-squared:  0.002597
## F-statistic: 2.414 on 1 and 542 DF,  p-value: 0.1209
```

```
##
## Call:
## lm(formula = time ~ traffic2, data = commute)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -23.639  -8.669   0.758   7.190  33.504
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 48.44975    0.67404   71.88   <2e-16 ***
## traffic2    -1.67008    0.09441  -17.69   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 11.29 on 542 degrees of freedom
## Multiple R-squared:  0.366,  Adjusted R-squared:  0.3648
## F-statistic: 312.9 on 1 and 542 DF,  p-value: < 2.2e-16
```

The main difference is the p-value. Model 2b has a very small p-value (<2e-16) for the coefficient of the transformed traffic variable while Model 2a's is very large (0.121).

This shows that there is a significant linear correlation between traffic^2 and time (Model 2b) while there is no significant linear correlation between traffic and time (Model 2b).

The estimated values of the coefficient relating to traffic and the intercept are also different. The coefficient of traffic^2 in model 2b (-1.67) shows a negative linear correlation with a 1.67minute decrease in commute time for every unit increase in traffic^2. In Model 2a, there is a non-significant linear correlation.

Word count for Q4: 250

Question 5

```
##
## Call:
## lm(formula = time ~ as.factor(direction) + as.factor(food) +
##     as.factor(shoes) + as.factor(school) + coffee + rain + temperature +
##     traffic, data = commute)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -31.157  -5.612   1.300   6.554  22.233
##
## Coefficients:
##                            Estimate Std. Error t value Pr(>|t|)
## (Intercept)                 14.1133     1.5805   8.930  < 2e-16 ***
## as.factor(direction)to       0.1488     0.7876   0.189  0.85018
## as.factor(food)yes           5.5812     0.8915   6.260 7.90e-10 ***
## as.factor(shoes)plimsolls    2.1505     1.2046   1.785  0.07478 .
## as.factor(shoes)sandals      7.9484     0.9882   8.043 5.69e-15 ***
## as.factor(shoes)trainers   -0.6007     1.1232  -0.535  0.59303
## as.factor(school)yes        9.2949     0.9692   9.590  < 2e-16 ***
## coffee                      6.3838     0.2883  22.142  < 2e-16 ***
## rain                        2.2898     0.2839   8.065 4.88e-15 ***
## temperature                 0.1808     0.0686   2.635  0.00866 **
## traffic                     0.3230     0.1769   1.826  0.06840 .
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 9.103 on 533 degrees of freedom
## Multiple R-squared:  0.5949, Adjusted R-squared:  0.5873
## F-statistic: 78.26 on 10 and 533 DF,  p-value: < 2.2e-16
```

There are 11 coefficients which include the intercept, 6 from categorical covariates and 4 from numeric covariates.

For every categorical covariate, there is one less coefficient than the number of categories, for example, shoes, there are 4 types of shoes but there are only 3 coefficients. One shoe type (i.e. boots) acts as the reference category so there are only 3 dummy variables for the remaining types of shoes.

Numeric covariates each have one coefficient.

- Interpret estimates:
  - rain; There is a positive linear correlation between rain and time such that an increase in rain by 1 mm would lead to an increase of 2.28 minutes in commute time.
  - shoessandals; It would take 8.05 minutes longer in commute time if one wears sandals instead of boots, given that all other variables remain the same.
  - Intercept; Commute time is estimated to be 14.1 minutes given that one is
    - travelling to work
    - does not stop to buy food
    - worn boots on the walk
    - did not have to pick up their children from work
    - faced 0 mm of rain
    - travelling in 0 degrees Celsius temperature
    - and encountered 0 indexes amount of traffic on the roads

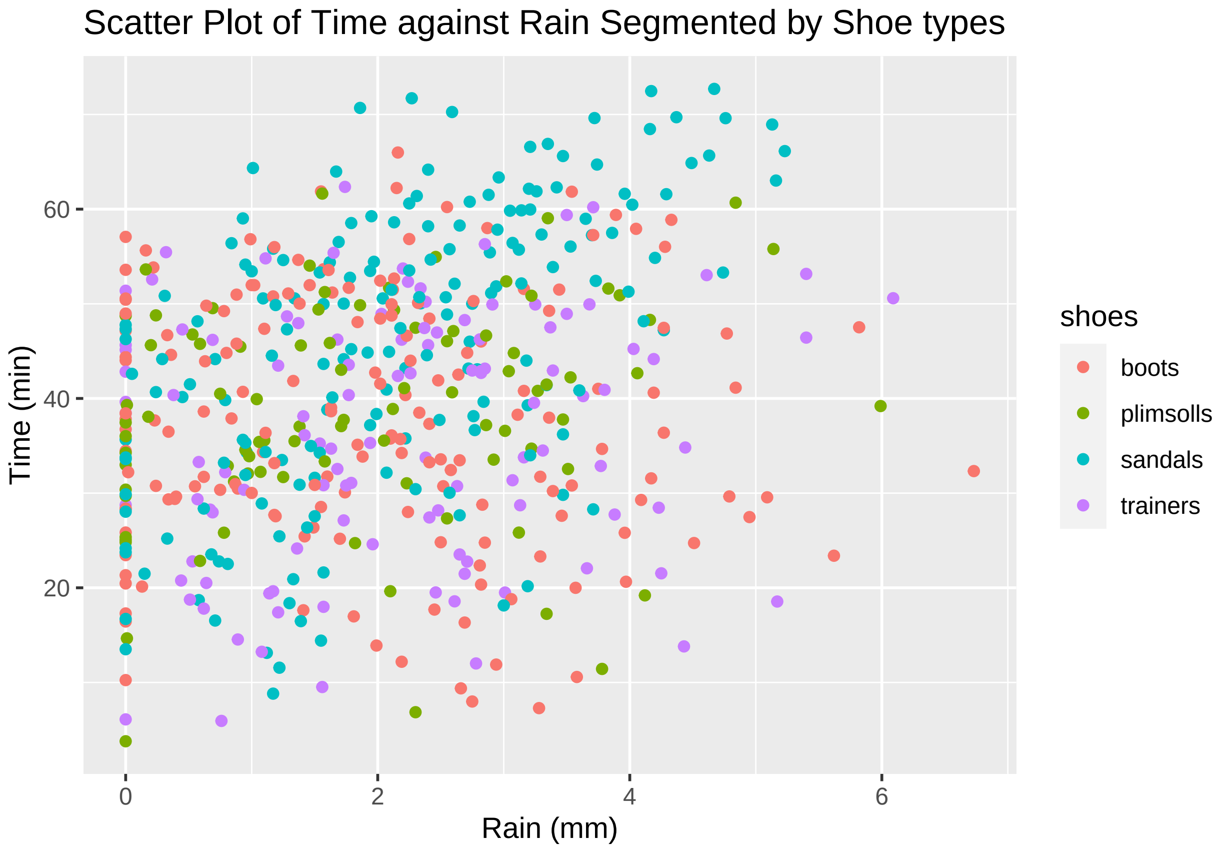
Word count for Q5: 191

Question 6

All of the coefficients would be multiplied by 60. In the initial model, we measure time in minutes. Since all coefficients are related to time, a model with time in seconds would have all coefficients multiplied by 60.

Word count for Q6: 38

Question 7



There is sufficient evidence to support the claim that the effect of rain on time is different depending on shoe type.

From the scatter plot, we can see that points for sandals are clustered along an upward-sloping line. However, for other types of shoes, there seems to be no particular pattern in the scattering of points.

Thus, it seems that rain has an effect on time when wearing sandals but no effect on time when wearing other types of shoes.

```
##
## Call:
## lm(formula = time ~ as.factor(direction) + as.factor(food) +
##     as.factor(shoes) + as.factor(school) + coffee + rain + temperature +
##     traffic + rain * as.factor(shoes), data = commute)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -27.548  -4.922   1.644   6.063  15.728
##
## Coefficients:
##                                Estimate Std. Error t value Pr(>|t|)
## (Intercept)                    19.03746    1.60523  11.860  < 2e-16 ***
## as.factor(direction)to          0.27888    0.71700   0.389 0.697464
## as.factor(food)yes              5.33923    0.81203   6.575 1.17e-10 ***
## as.factor(shoes)plimsolls      -2.28897    1.83432  -1.248 0.212633
## as.factor(shoes)sandals        -6.03975    1.62725  -3.712 0.000228 ***
## as.factor(shoes)trainers       -2.74208    1.83889  -1.491 0.136514    
## as.factor(school)yes           10.00160    0.88641  11.283  < 2e-16 ***
## coffee                          6.29522    0.26253  23.979  < 2e-16 ***
## rain                           -0.28719    0.44346  -0.648 0.517518    
## temperature                     0.19901    0.06247   3.186 0.001529 ** 
## traffic                         0.22206    0.16139   1.376 0.169412    
## as.factor(shoes)plimsolls:rain  2.23517    0.77124   2.898 0.003909 ** 
## as.factor(shoes)sandals:rain    6.72262    0.65439  10.273  < 2e-16 ***
## as.factor(shoes)trainers:rain   1.20458    0.72554   1.660 0.097454 .  
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
## 
## Residual standard error: 8.282 on 530 degrees of freedom
## Multiple R-squared:  0.6665, Adjusted R-squared:  0.6583 
## F-statistic: 81.48 on 13 and 530 DF,  p-value: < 2.2e-16
```

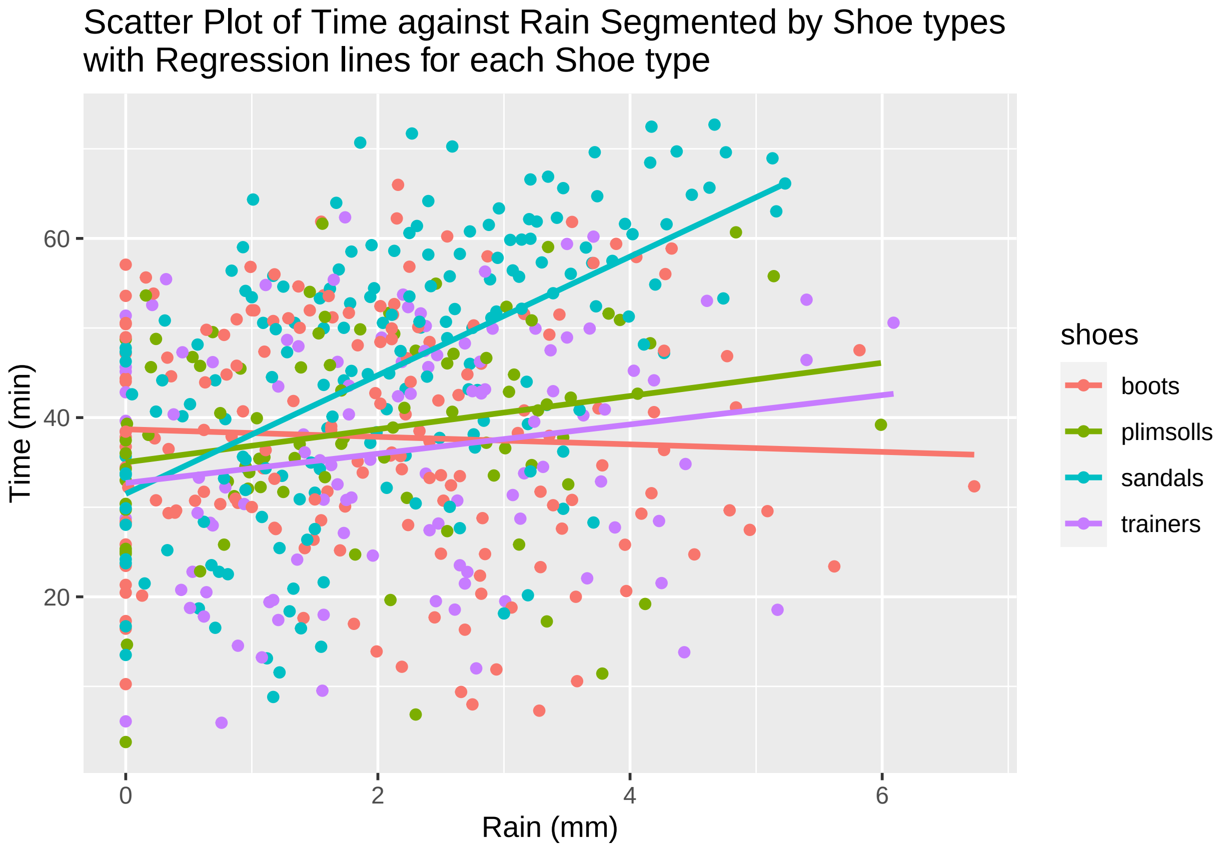
Knowing that the effect of rain on time varies with shoe type, we should not assume the same general model relating rain and time across all shoe types and should instead have different linear regression models for each shoe type. This would illustrate the effect of rain on time for each shoe type.

In this model, we include the variable `rain*as.factor(shoe)` which would relate to rain and shoe type. This produced 4 different sets of intercepts and slopes, one for each shoe type which can be seen in the plot below.

Interpretation: The intercept indicates that with 0 mm of rain, it takes 19.0 minutes to commute with boots. With each additional mm of rain, it would take 0.287 less commute time.

It would take 16.8 minutes, 13 minutes, and 16.3 minutes to commute with plimsolls, sandals and trainers respectively with 0 mm of rain. With each additional mm of rain, it would take 2.2, 6.7 and 1.2 minutes longer for plimsolls, sandals and trainers respectively.

This shows that rain has an effect on time when wearing sandals as suspected earlier.



Word count for Q7: 261

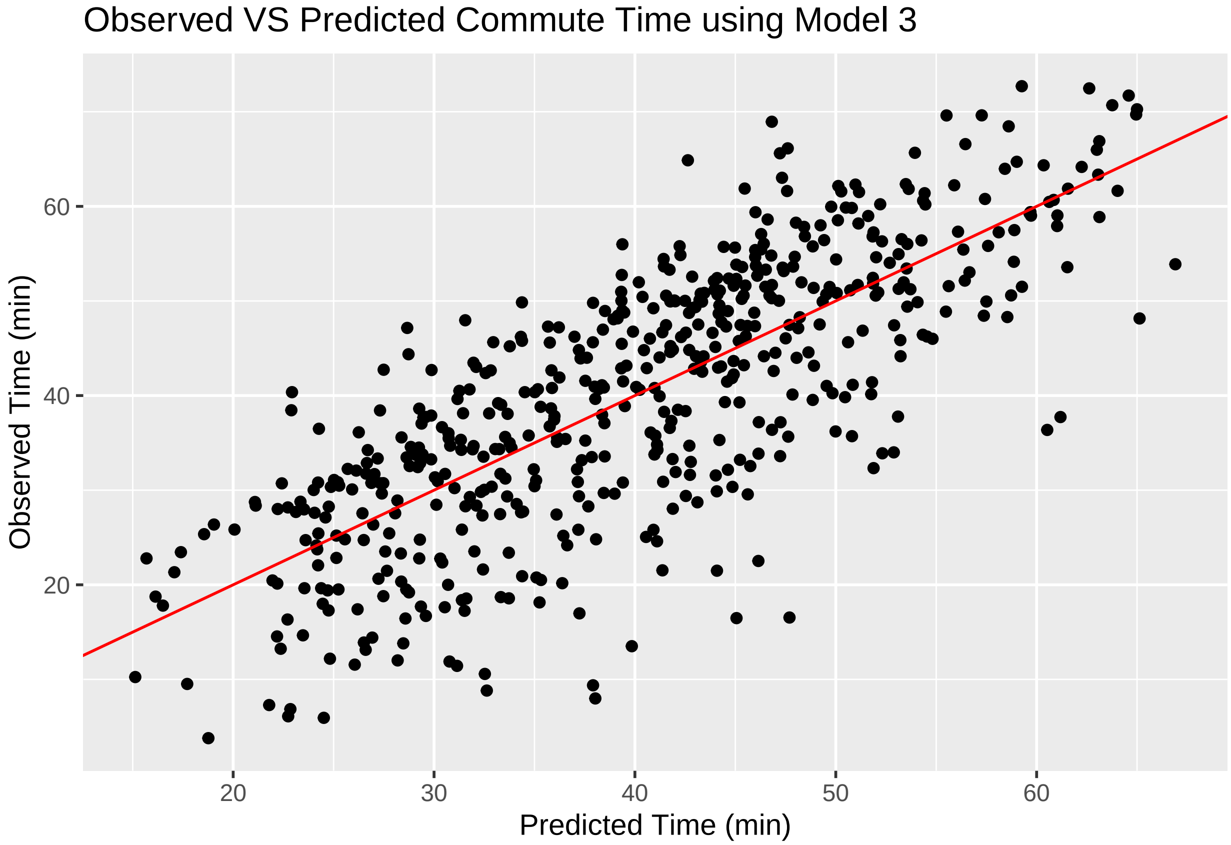
Question 8

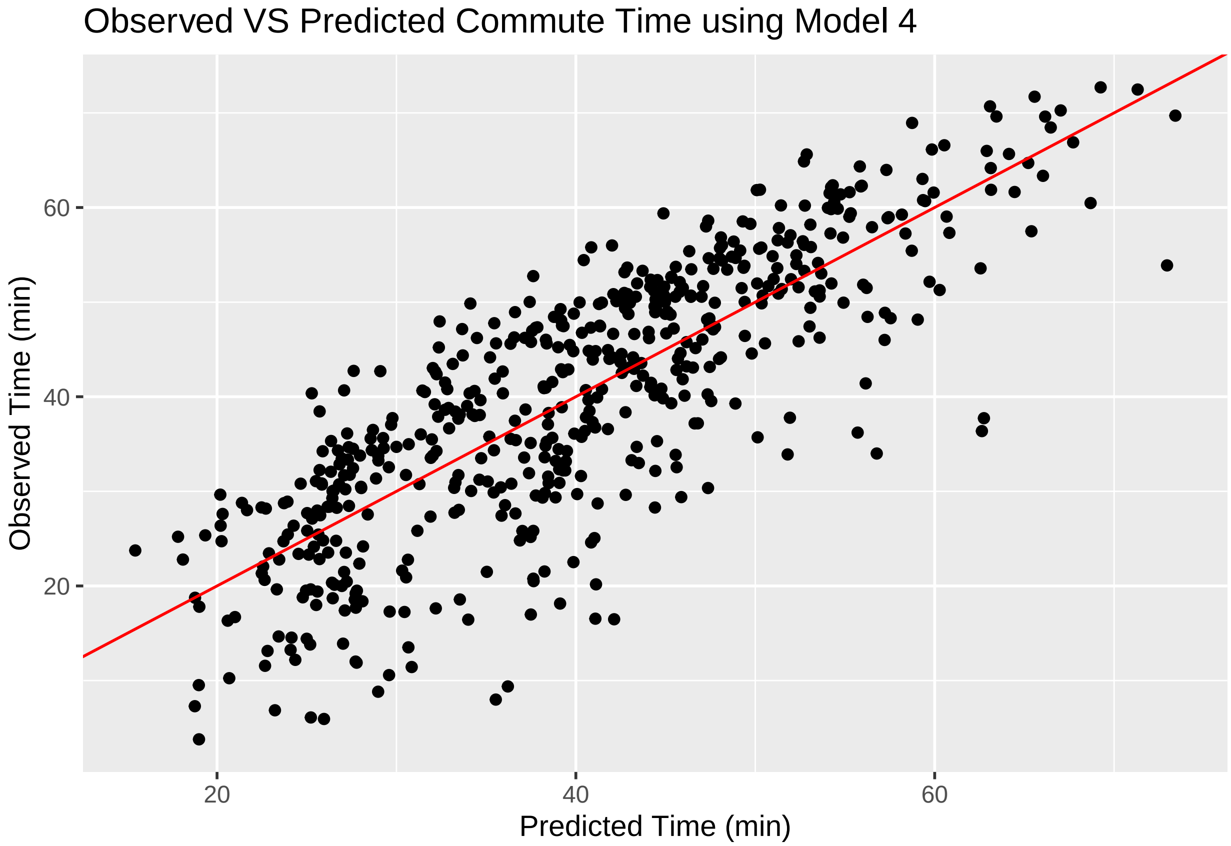
```
##
## Call:
## lm(formula = time ~ as.factor(direction) + as.factor(food) +
##     as.factor(shoes) + as.factor(school) + as.factor(coffee) +
##     rain + temperature + traffic^2 + rain * as.factor(shoes),
##     data = commute)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -28.564  -4.477   2.462   5.976  11.367
##
## Coefficients:
##                              Estimate Std. Error t value Pr(>|t|)
## (Intercept)                  16.99886    1.55052  10.963  < 2e-16 ***
## as.factor(direction)to        0.11131    0.68223   0.163  0.87046
## as.factor(food)yes            5.17455    0.77115   6.710 5.03e-11 ***
## as.factor(shoes)plimsolls    -1.33022    1.74844  -0.761  0.44712
## as.factor(shoes)sandals      -4.49972    1.56574  -2.874  0.00422 **
## as.factor(shoes)trainers     -1.92269    1.75091  -1.098  0.27266
## as.factor(school)yes         10.29028    0.84261  12.212  < 2e-16 ***
## as.factor(coffee)1           14.93529    1.25960  11.857  < 2e-16 ***
## as.factor(coffee)2             14.99644    0.88308  16.982  < 2e-16 ***
## as.factor(coffee)3             21.24253    0.93677  22.676  < 2e-16 ***
## as.factor(coffee)4             22.50178    1.34345  16.749  < 2e-16 ***
## rain                            0.10007    0.42485   0.236  0.81388
## temperature                     0.17169    0.05954   2.884  0.00409 **
## traffic                         0.23928    0.15382   1.556  0.12040
## as.factor(shoes)plimsolls:rain  1.94664    0.73661   2.643  0.00847 **
## as.factor(shoes)sandals:rain    5.95909    0.63149   9.436  < 2e-16 ***
## as.factor(shoes)trainers:rain   0.93063    0.69162   1.346  0.17902
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.859 on 527 degrees of freedom
## Multiple R-squared:  0.7015, Adjusted R-squared:  0.6924
## F-statistic: 77.39 on 16 and 527 DF,  p-value: < 2.2e-16
```

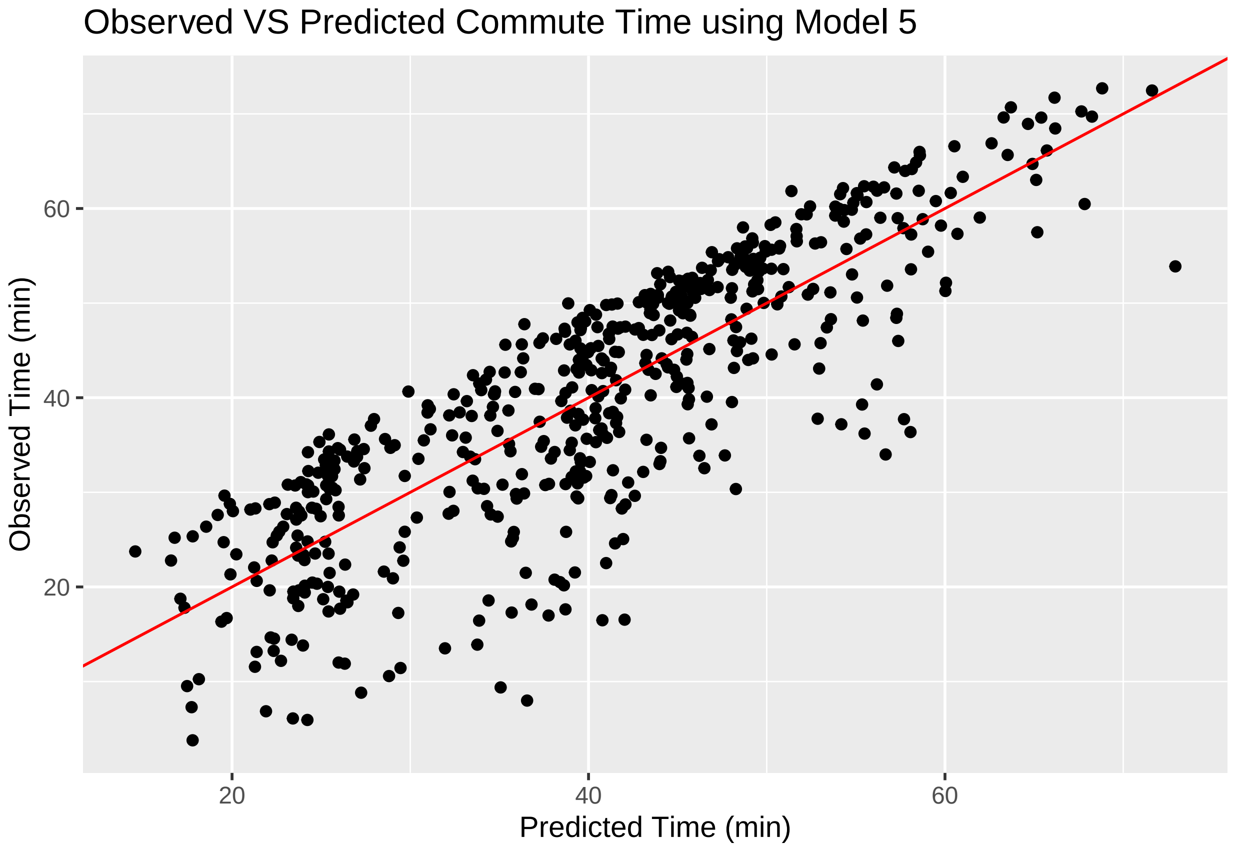
Model 5 would be an adaptation of Model 3 with the following changes: 1. Changing coffee into a categorical covariate; as shown in question 2, the number of cups of coffee taken should not assume linearity. 2. Using transformation traffic^2 instead of traffic; this has been demonstrated in question 4 that the transformation traffic^2 is more suitable for the linear model as time and traffic^2 have a stronger linear correlation. 3. Introducing variable `rain*as.factor(shoes)`; there are associations observed between the amount of rain and shoe types worn where this variable accounts for the different effects on time.

Word count for Q8: 97

Question 9


Observed VS Predicted Commute Time using Model 3
Observed Time (min)
60
40
20
20
30
40
50
60
Predicted Time (min)


Observed VS Predicted Commute Time using Model 4
Observed Time (min)
60
40
20
20
40
60
Predicted Time (min)


Observed VS Predicted Commute Time using Model 5

	Model	RMSE
1	3	9.010527
2	4	8.175104
3	5	7.735060

From Model 3 to 5, value points are increasingly concentrated along the line `y=x` and are closer to the line, meaning that the residuals are the smallest.

There are no issues occurred in these three plots

The red line in each plot indicates the line `y=x`. I prefer Model 5 as the data points on the plot are the most closely distributed along the line, meaning that predicted values are the closest to observed values. This shows that Model 5 is the most accurate.

Furthermore, the smallest root mean squared error in Model 5 reinforces the fact that the residual in Model 5 is the smallest.

Word count for Q9: 106