Group Work

```
Rows: 460
Columns: 24
$ FeatureCode    <chr> "S12000005", "S12000006", "S12000008", "S12000010", "S1~
$ DateCode       <dbl> 2005, 2005, 2005, 2005, 2005, 2005, 2005, 2005, 2005, 2~
$ Expenditure    <dbl> 3541, 20466, 12145, 6818, 11185, 7428, 12049, 29490, 81~
$ Killed_Injured <dbl> 25, 144, 53, 50, 17, 20, 85, 197, 38, 62, 39, 82, 8, 14~
$ Cards          <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA,~
$ Congestion     <dbl> 13.8, 5.1, 12.9, 14.9, 15.5, 1.1, 12.7, 11.1, 9.5, 14.5~
$ Repair         <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA,~
$ Mileage        <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA,~
$ Work_Bus       <dbl> 6.2, 5.8, 10.4, 8.2, 7.8, 18.1, 10.0, 7.3, 9.4, 20.0, 8~
$ Business       <dbl> 6, 6, 4, 5, 3, 7, 4, 8, 2, 8, 6, 4, 14, 4, 8, 4, 6, 5, ~
$ School         <dbl> 53.3, 46.4, 41.9, 60.5, 44.2, 20.0, 46.8, 48.9, 38.3, 5~
$ Commuting      <dbl> 23, 27, 27, 29, 19, 28, 24, 29, 21, 28, 33, 23, 25, 31,~
$ Work_Cycling   <dbl> 2.7, 0.9, 0.0, 2.7, 2.4, 0.0, 1.2, 2.0, 0.0, 0.0, 5.4, ~
$ Education      <dbl> 2, 2, 2, 2, 4, 3, 3, 2, 3, 6, 2, 1, 1, 3, 1, 2, 2, 5, 3~
$ Health         <dbl> 2, 3, 3, 1, 2, 3, 3, 2, 4, 1, 2, 2, 3, 3, 2, 2, 2, 2, 3~
$ Shopping       <dbl> 20, 24, 17, 23, 12, 21, 26, 19, 27, 18, 18, 20, 21, 18,~
$ Work_Train     <dbl> 3.1, 0.0, 3.2, 6.0, 8.4, 0.0, 6.5, 1.3, 9.5, 0.0, 0.5, ~
$ Work_Walking   <dbl> 11.4, 20.9, 11.6, 12.6, 3.8, 10.7, 9.8, 16.3, 11.6, 7.9~
$ Train_Stations <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA,~
$ Satisfaction   <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA,~
$ One_Car        <dbl> 45.9, 47.6, 44.8, 43.4, 39.5, 48.9, 43.1, 49.8, 39.2, 4~
$ More_Car       <dbl> 28.8, 28.6, 25.9, 27.3, 41.7, 24.7, 28.4, 28.9, 21.9, 3~
$ Without_Car    <dbl> 25.4, 23.8, 29.3, 29.3, 18.8, 26.4, 28.5, 21.3, 38.9, 2~
$ Petrol_Diesel  <dbl> 19, 165, 79, 61, 60, 13, 104, 165, 38, 43, 46, 53, 8, 7~
```

Introduction

Public transport is necessary for most Scotlands. Therefore, its comfort and customer satisfaction are important for operators. To help improve public transport, The purpose is to discover factors having correlation with passengers' satisfaction in this project. The data are from Scotland's official statistics. The theme of Transport contains seven datasets, Road Transport Expenditure, Public Transport, Road Vehicles, Concessionary Travel Cards, Road Network and Traffic, Travel to Work and Other Purposes. There are 460 observations of 24 variables in this research.
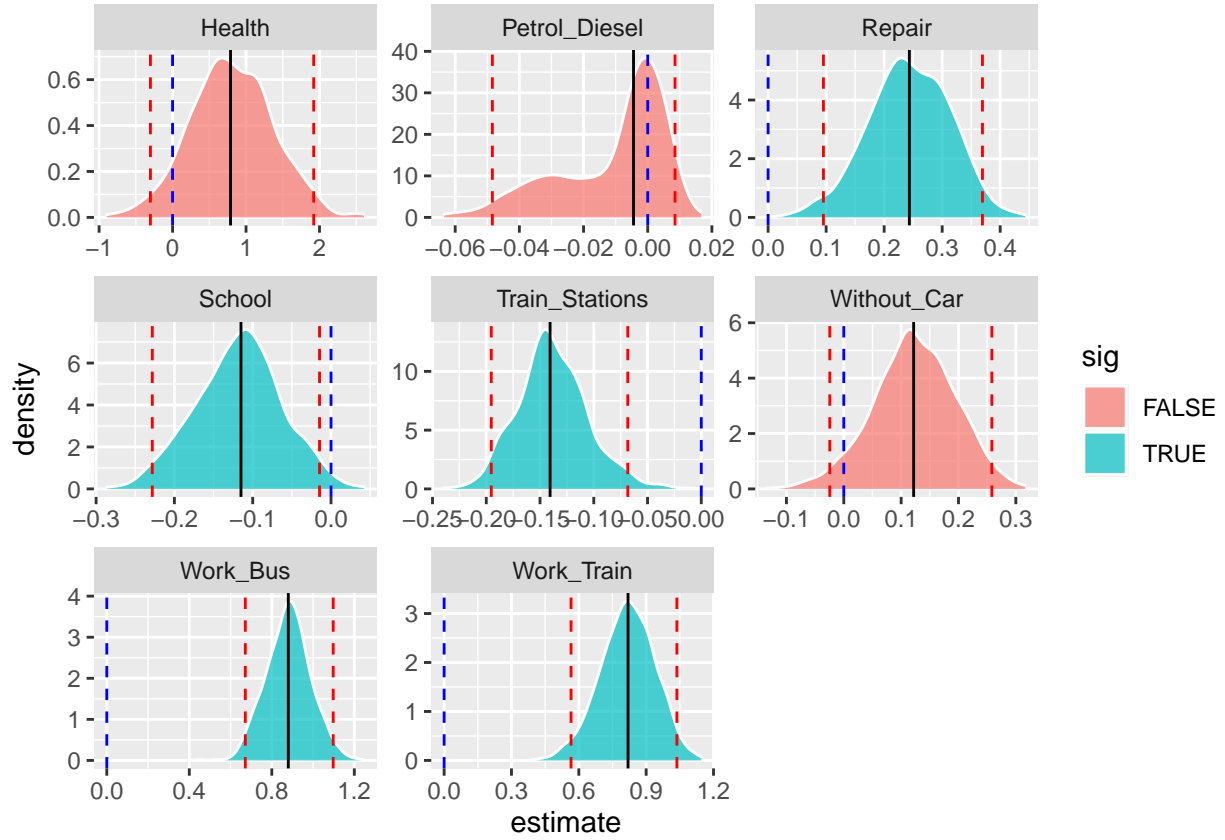
Exploratory Data Analysis

The research is based on a poll by the Scottish Household Survey (SHS). The whole dataset, combined from seven datasets obtained from website of Scottish government statistics, has 24 factors and 460 observations. The explanatory variables are Road Casualties, Road Transport Expenditure, Public Transport, Road Vehicles, Concessionary Travel Cards, Road Network and Traffic and Travel to work and other purposes.

Variable	Mean	SD	min	max
Expenditure	25735.16	69857.32	-10844.0	502630.00
Killed_Injured	124.17	338.85	1.0	2842.00
Cards	73852.76	207392.71	3979.0	1436617.00
Congestion	10.53	4.99	0.0	23.60
Repair	37.10	8.20	19.5	62.00
Mileage	2653.78	7353.99	131.0	48714.21
Work_Bus	8.64	5.64	0.0	28.50
Business	2.40	1.79	0.0	14.00
School	51.99	11.86	15.8	84.00
Commuting	24.25	4.77	14.8	43.60
Work_Cycling	1.88	2.01	0.0	11.80
Education	4.97	2.46	0.6	12.60
Health	2.32	0.94	0.0	5.70
Shopping	22.66	3.70	12.0	35.10
Work_Train	4.01	3.90	0.0	19.00
Work_Walking	12.22	5.54	0.5	29.00
Train_Stations	20.42	57.62	0.0	357.00
Satisfaction	69.56	10.28	35.0	93.00
One_Car	43.96	3.80	33.3	53.50
More_Car	29.79	7.35	10.8	50.30
Without_Car	26.21	8.50	8.7	52.20
Petrol_Diesel	181.39	501.18	8.0	3331.00

Formal Data Analysis

Linear regression model, a model for examining and discovering relations between response variable and explanatory variable(s), is applied in this project. The formula is presented below:

$$y = \beta_0 + \beta_1 x_1 + \dots + \beta_p x_p + \varepsilon \tag{1}\label{eq:(1)}$$

where y is the response variable, x_1 to x_p is the number of columns selected from 1 to p, β_0 is the intercept of data, β_1 to β_p is the coefficients of corresponding columns of X from 1 to p. The *varepsilon* is the error terms of the estimations. A total of four regressions were performed. First, all the independent variables were regressed, and then the highly correlated factors were sequentially removed, and finally four models were obtained. step AIC method is applied as a certeria for model selection. model 2 is the best choice. Using Bootstrap to select significant variables at 95% level and obtain its confidence interval of parameter estimation and their observations.This process should be repeated 1000 times.

Conclusions