



Introduction

- We **aim** to find factors related to passengers' satisfaction with public transport to help operator to improve its service.
- 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.

Methods

- Summarize table and density plots are illustrated to detect data patterns. The scatter and correlation plots are proposed to explore the relationship among variables. Potential factors are identified through **EDA**.
- The *Satisfaction* is the response variable, the *DateCode* is the control variable and others are independent variables. A **linear regression** model is applied as [Eq.\(1\)](#):

$$y = \beta_0 + \beta_1 x_1 + \dots + \beta_p x_p + \varepsilon \quad (1)$$

- Model diagnosis** are carried out to check model assumptions. **Stepwise regression** is applied to select variables with **AIC** as the criterion. Compare the selected model with the full model in $adj R^2$, AIC and BIC.
- The uncertainty of the parameters is determined via **bootstrap** method. The significant variables are verified and their 95% CI are estimated.

Results

Model diagnosis

[Fig.1](#) shows the regression diagnosis results of model1 with all variables.

- Residuals vs Fitted plot** (left top) shows that there is no systematic correlation between the residual value and the fitting value.
- Normal Q-Q plot** (right top) shows the points on the graph fall on a straight line with an angle of 45 degrees, which indicates the assumption of normality is not violated.
- Scale-location plot** (left bottom) displays the points around the horizontal line are randomly distributed. The invariant variance assumption is satisfied.
- Residuals vs leverage** (right bottom) figures out special observations.

- According to the correlation plot and VIF, there is obvious **multicollinearity** among variables.

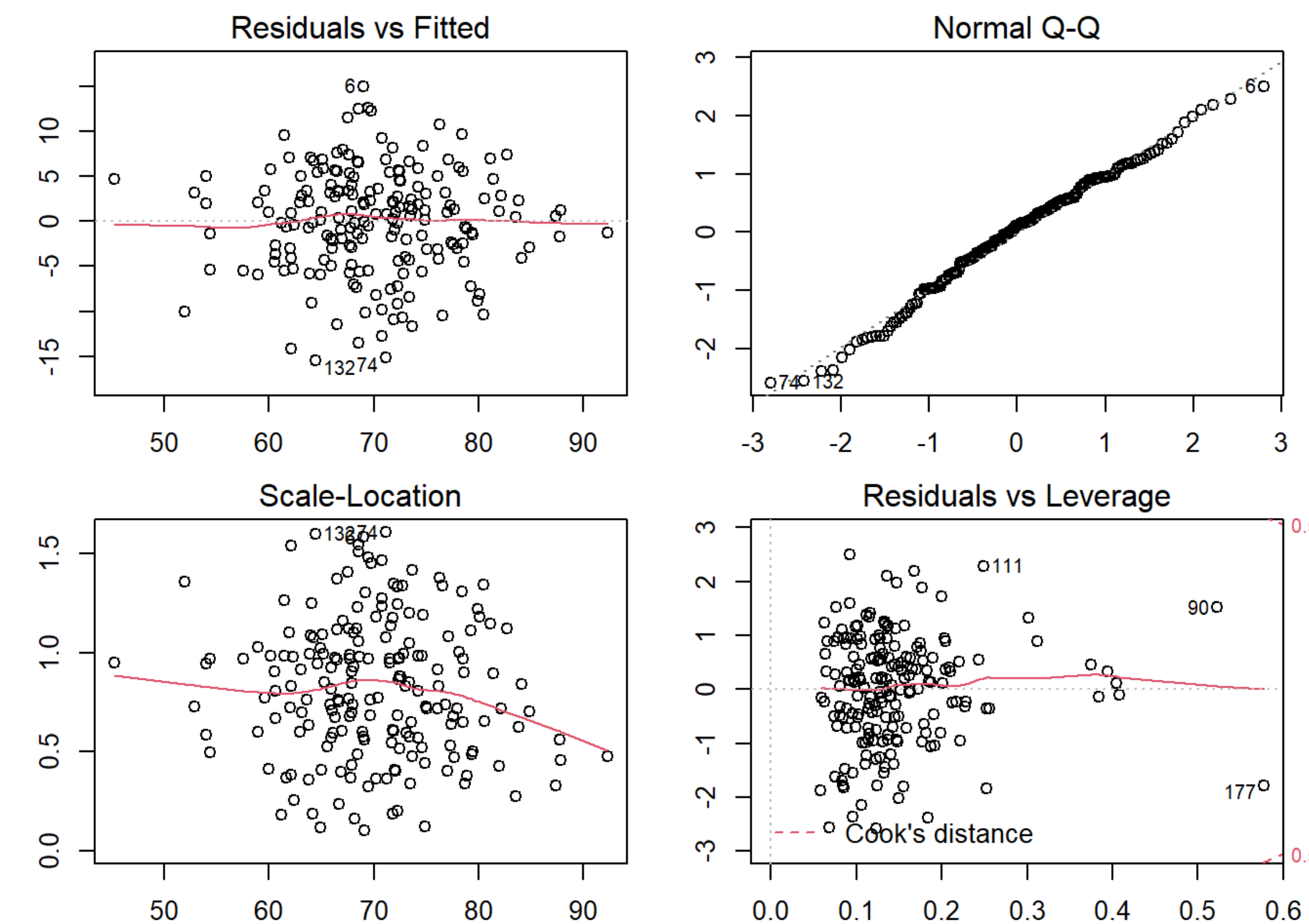


Figure 1: Regression diagnosis results

Stepwise regression

Stepwise regression technique is applied to model selection. The selected model is as [Table. 1](#).

Table 1: Model2, selected by Stepwise regression

term	estimate	std.error	statistic	p.value
(Intercept)	49.1373307	4.0389144	12.165975	0.0000000
Cards	0.0001125	0.0000344	3.269305	0.0012981
Repair	0.2415064	0.0629651	3.835558	0.0001747
Work_Bus	0.8573720	0.1037070	8.267251	0.0000000
School	-0.1186535	0.0456380	-2.599885	0.0101222
Health	0.7652727	0.5091104	1.503157	0.1346011
Work_Train	0.8083991	0.1270588	6.362402	0.0000000
Train_Stations	-0.1437217	0.0439176	-3.272534	0.0012842
Without_Car	0.1023895	0.0713492	1.435047	0.1530586
Petrol_Diesel	-0.0309497	0.0142438	-2.172855	0.0311356

The variables with high correlation are removed by stepwise regression.

Model compare

Table 2: Compare selected model by Stepwise regression with the full model

model	adj.r.squared	AIC	BIC
model 1	0.5478475	1271.288	1365.604
model 2	0.5625876	1254.526	1309.814

Model2 has higher $adj R^2$ and smaller AIC and BIC compared with the model1 as [Table. 2](#). Therefore the **model2** is better.

Bootstrap

To obtain robust results, a bootstrap is developed for the estimation of parameters.

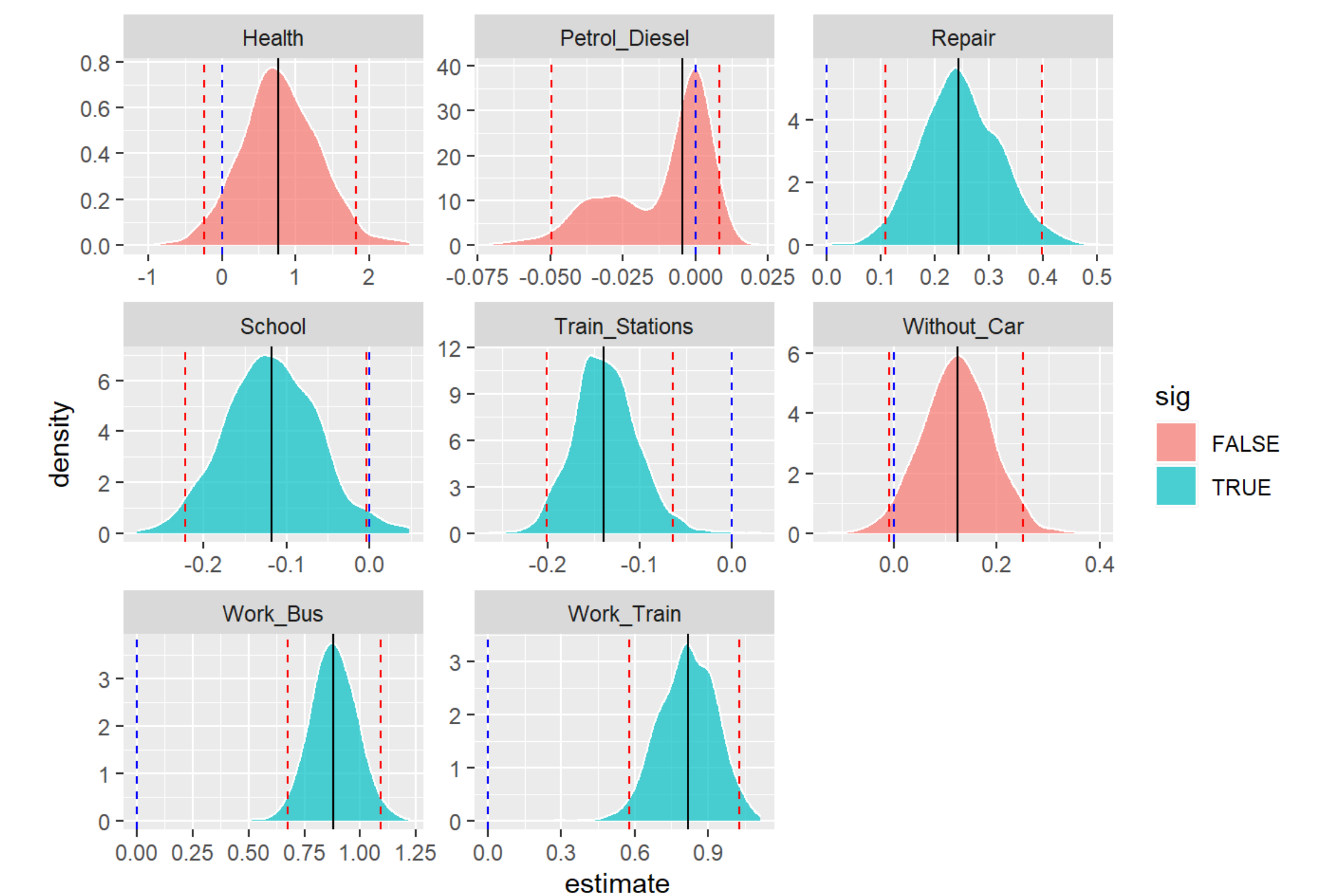


Figure 2: Density plots of parameters via bootstrap

The density plots of parameters are displayed in [Fig.3](#). The variables with orange are not significant while the variables with blue are significant at the $\alpha = 0.05$. The blue dashed lines are zero and the orange dashed lines are 95% CI of parameters.

Conclusion

We can conclude that the **Repair** (*The Percentage Of Roads Needing Repairs*), **School** (*Child Journeys To School By Walking/Cycling*), **Train_Stations** (*Number Of Train Stations*), **Work_Bus** (*Bus Journeys To Work*), **Work_Train** (*Train Journeys To Work*) variables have great influence on satisfaction with public transport.

References

- Jim Hester and Hadley Wickham, (2020). *fs: Cross-Platform File System Operations Based on 'libuv.'* R package version 1.5.0
- Kuhn et al., (2020). Tidymodels: a collection of packages for modeling and machine learning using tidyverse principles.
- Wickham et al., (2019). *Welcome to the tidyverse. Journal of Open Source Software*, e, 4(43), 1686