# 1014SCG – Milestone 2

**Himath Ratnayake**

## A Study of The Interaction Between Mode of Transport, School of Study, And Weekly Transport Cost

**Signature:** 

s5209861

himath.ratnayake@griffithuni.edu.au

## Part 1

### Explanation and Justification of Variables

For this study, the factors (independent/explanatory variables) whose interactions were explored were the mode of transport used by students and the school of study that they were a part of. Each independent variable was qualitative (categorical) and possessed 3 nominal levels, meaning the levels were in no particular order (Table 1).

The response variable for this study was the numerical continuous variable cost; the amount students paid for transport each week in dollars, giving the unit $/week. This variable was continuous as cost can take any value in a given range, including decimals.

**Table 1:** Summary of all variables

| Table 1 | | | |
|---|---|---|---|
| **Variable Name** | **Variable Type** | **Notes** | **Used in Study?** |
| **Mode of Transport** | Qualitative Nominal | 3 Levels: Public, Private, Other | Yes – independent variable 1 |
| **School** | Qualitative Nominal | 3 Levels: Env. Science, Business, Other | Yes – independent variable 2 |
| **Cost** | Quantitative Continuous | Units = $/week | Yes – dependent variable |
| **Frequency** | Quantitative Discrete | Units = # of times certain transport is used per week | No |
| **Age** | Qualitative Ordinal | 4 Levels: <20, 20-24, 25-29, 30+ (Units = Years) | No |

In this study, a factorial ANOVA (2-way analysis of variance) test will be employed to determine if group means are different between the independent variables selected, with regards to average weekly expenditure on transport. This is because accommodating for student's travel commitments is an important aspect of achieving a sustainable and organised campus. By analysing the interactions between modes of transport, school, and cost, Universities will be able to better understand the importance of facilitating for particular transport modes around specific campus faculties to improve convenience for students.

**Exploratory Data Analysis:**

Using the "Describe By" function in the "psych" R library, Table 2 was plotted with key information pertaining to this study.

**Table 2:** Summary statistic showing information for each treatment (See Appendix 1.2 for R Code)

| Treatm. | Mode | Sch. | N | Mean | SD | Median | Min | Max | Range | Skew | Kurtosis | SE |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | Public | ENS | 24 | 61.77 | 10.43 | 59.34 | 46.07 | 78.09 | 32.02 | 0.28 | -1.39 | 2.13 |
| 2 | Private | ENS | 24 | 54.24 | 9.8 | 53.74 | 37.05 | 74.49 | 37.44 | 0.38 | -0.94 | 2 |
| 3 | Other | ENS | 24 | 55.67 | 7.26 | 55.53 | 43.94 | 70.55 | 26.61 | 0.43 | -0.6 | 1.48 |
| 4 | Public | BUS | 24 | 60.95 | 11.34 | 58.19 | 44.82 | 87.44 | 42.62 | 0.48 | -0.73 | 2.31 |
| 5 | Private | BUS | 24 | 55.31 | 10.04 | 50.88 | 42.25 | 76.19 | 33.94 | 0.72 | -0.77 | 2.05 |
| 6 | Other | BUS | 24 | 55.24 | 6.68 | 55.78 | 42.9 | 69.5 | 26.6 | 0.37 | -0.38 | 1.36 |
| 7 | Public | OTH | 24 | 61.93 | 10.29 | 59.8 | 47.26 | 82.62 | 35.36 | 0.43 | -1.05 | 2.1 |
| 8 | Private | OTH | 24 | 52.1 | 9.82 | 51.92 | 35.85 | 73.03 | 37.18 | 0.48 | -0.4 | 2.01 |
| 9 | Other | OTH | 24 | 56.85 | 6.98 | 56.27 | 46.24 | 69.22 | 22.98 | 0.15 | -1.22 | 1.42 |

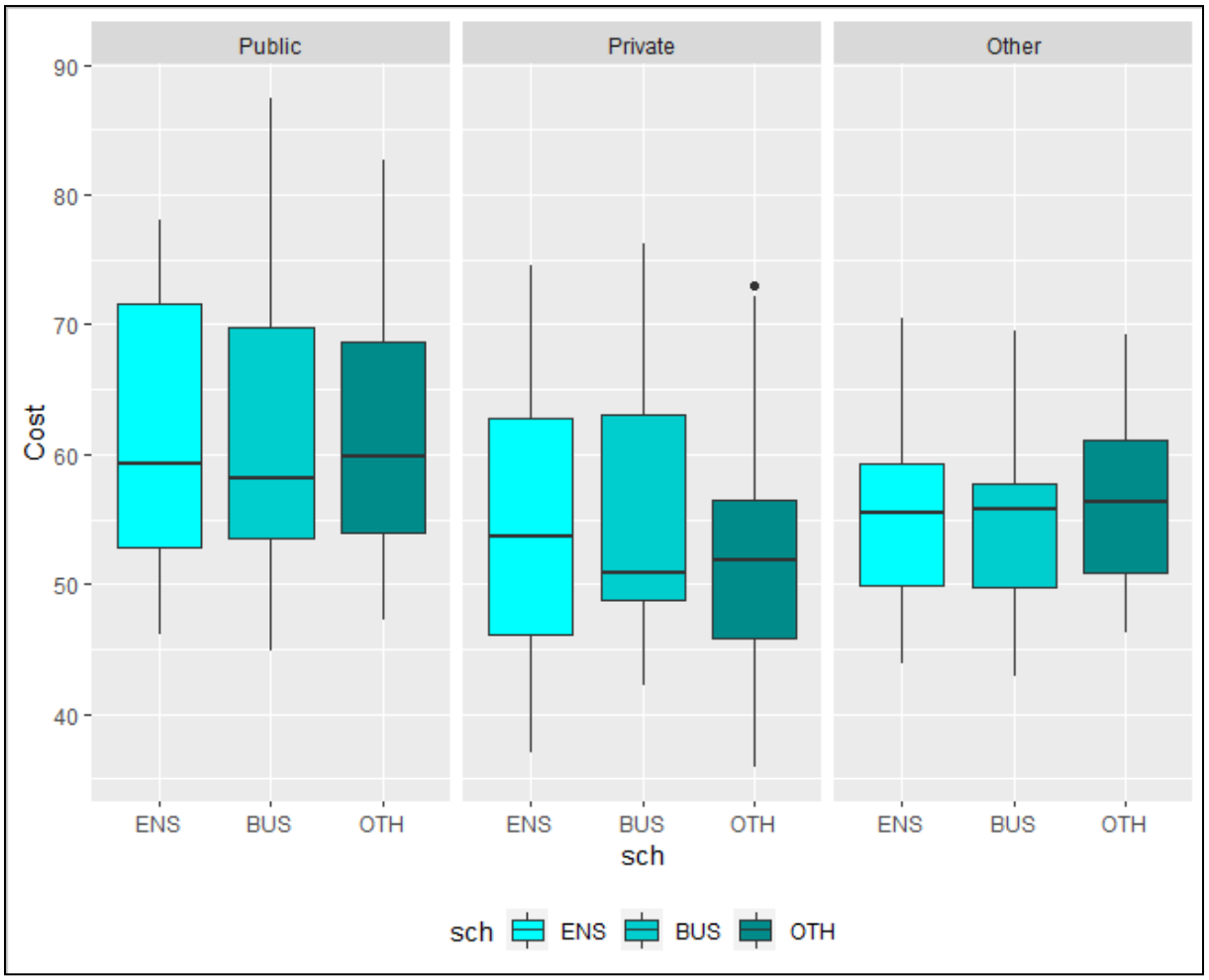


**Figure 1:** Diagram depicting boxplots for each treatment in the data (See Appendix 2.1 for R Code)

A boxplot (Figure 1) was plotted for each treatment with the purpose of visually depicting the interquartile range of each treatment, as well as aiding in easily identifying any outliers. There was one found within this data set for treatment 6 (Other, BUS), which can affect the mean and range within the treatment; in this case it could increase the treatment's mean and range as it is above the 4th quartile.

Treatment 5 (Private, BUS) had a mean of 55.31. This mean was significantly greater than the median of 50.88, which in turn also meant this treatment was moderately skewed at 0.72 and not as symmetrical (See Figure 2). The median for this data set was also the lowest of all treatments – a reason for this low median could be because of a small Quartile 2 in the treatment's interquartile range, as it was only between 48.72 and 50.88 (a 2nd quartile range of only $2.16). There was central tendency and high density around this area as seen in figure 2's graph shape for private, BUS. In contrast, Quartile 3 was significantly more spread out, ranging from 50.88 to 63.05 (a range of $12.17 – approximately 5.6 times greater than the range of the 2nd quartile). Therefore, these larger, more dispersed higher values increase the mean greatly, yet the median remains lower in comparison due to the high density occurring around the median and 2nd quartile.

Treatment 7 (Public, OTH) possesses the highest mean at 61.93 and highest median at 59.8 of all the data.
As seen in Figure 2 a gap was present between $70-75 as no student paid between these values for weekly transport. This contributes to a large 4th quartile in figure 1. Overall, students in this category tend to have the most expensive average weekly transport expenditure.

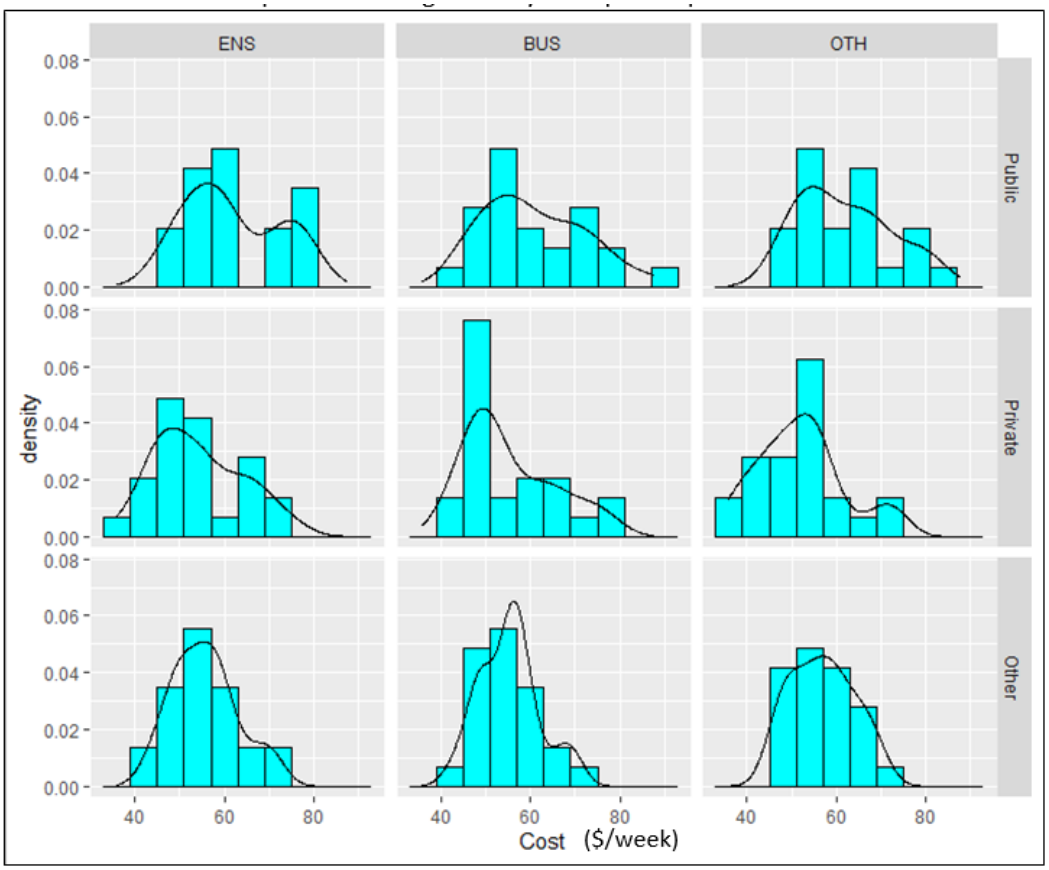


**Figure 2:** Diagram depicting histograms and density plots for each treatment in the data (See Appendix 2.2 for R Code)

Histograms were plotted for each treatment with the purpose of illustrating cost per week (x axis) against the frequency of each cost value and showing the shape of each curve. Density was also plotted above each graph to further illustrate data dispersion.
Within the data, there were several consistencies. In all treatments except treatment 6 (other, business), the mean cost of transport per week was greater than the median; this meant these treatments displayed right (positive) skew, which was further demonstrated by all values in the "skew" column of table 3 being positive numbers.
Ideally, skew values should be between -0.5 and 0.5 to be considered fairly symmetrical in shape. In this data, all treatments were fairly symmetrical except for treatment 5 (private, BUS), which was moderately right skewed with a value of 0.72 (>0.5).

The kurtosis values for all treatments were also all negative. This suggests that the distribution shapes for each graph are flatter than a normal distribution curve that has the same respective mean and standard deviation for the treatment.
Treatment 1 (Public, ENS) possesses the furthest kurtosis value from 0, at -1.39. This means that out of all treatments this treatment least follows the ideal normal distribution graph shape for its mean and standard deviation, which can visually be seen in the graph – as normality was an assumption for an ANOVA, this must also be verified later on in the report.

Treatment 6 (Other, BUS) had central tendency around the median of 55.78. The mean of this treatment was 55.24; this was the only treatment in the data set where the mean of the treatment was actually less than the median. Usually, this may suggest that the data set was negatively skewed, but due to the similarity in the mean and median values, the skewness was still positive at 0.37. Treatment 6 also had the kurtosis value closest to 0, at -0.38, meaning it most closely follows the expected normal distribution shape for its mean/standard deviation. Additionally, the standard deviation and in turn standard error was the lowest in the data set at 6.68 and 1.36, respectively. Therefore, the sample mean in treatment 6 was more likely to be an accurate representation of the population mean.

In this data, treatment 4 (Public, BUS) has the largest standard deviation at 11.34. This treatment's standard error was also the largest in the data set, which was to be expected since as standard deviation increases, so does standard error, as illustrated by the equation below:

$$SE = \frac{\sigma}{\sqrt{n}}$$

$\sigma$ ← Standard deviation
$n$ ← Number of samples

Standard error is an indicator of sample mean accuracy in comparison to the population mean, so within this data it must be noted that the standard error should ideally be lowered, such as by increasing the sample size, "n" (number of people surveyed for this study).
Therefore, the population mean of this treatment has a higher chance of being inaccurate compared to all other treatments.

For a factorial ANOVA, there are several assumptions that must be met due to it being a parametric test.

## Assumption #1: Residuals follow a normal distribution

The normality of residual distribution can be depicted through a QQPlot, where the standardized residuals should form a roughly straight line that fits closely to the trendline.

However, when the residuals of the original data were plotted, not all values fitted closely to the trendline as can be seen by Figure 3. To properly verify this visual test, a Shapiro-Wilk test of normality can be undertaken in R. (See appendix 2.3 for R code used to generate figure 3/4 graphs)

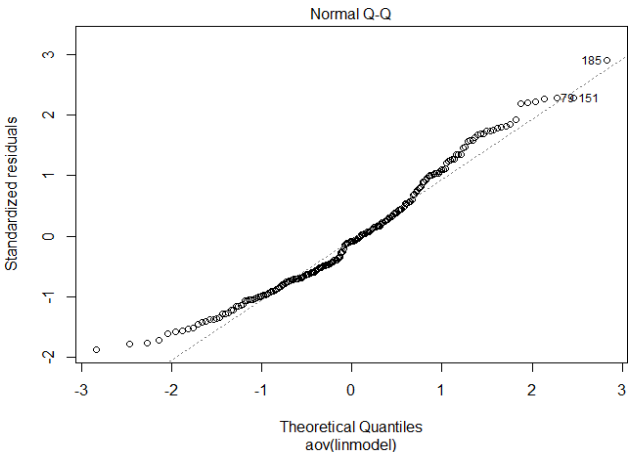


**Figure 3:** A QQPlot demonstrating the distribution of standardized residuals before transformation applied

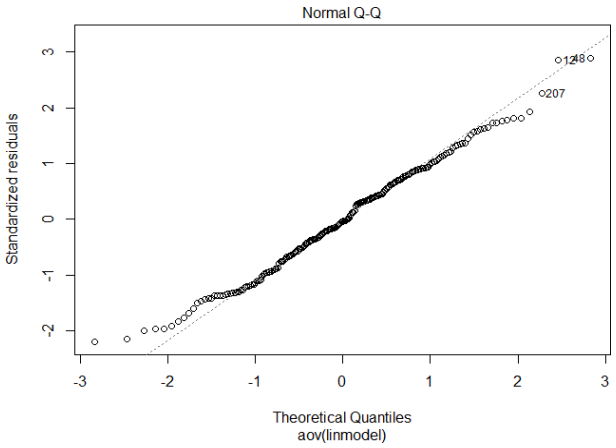


**Figure 4:** QQPlot demonstrating the distribution of standardized residuals after inverse transformation

**Null Hypothesis ($H_0$):** the residuals follow a normal distribution
**Alternate Hypothesis ($H_A$):** the residuals do not follow a normal distribution
**Significance Level, α:** 0.05
**Decision Rule**: Reject $H_0$ if the p value is < α
**Test Statistic:** Shapiro-Wilk Normality Test
**Test Statistic Value:** p value (see appendix 3.1 for R code)

```
data:  res
W = 0.97031, p-value = 0.0001637
```

**Conclusion 1:**

Since the p value = 0.0001637 < α = 0.05, the null hypothesis is rejected – that is to say that this data does not follow a normal distribution. Although the ANOVA test is quite robust to this, as using the normal cost values fails the normality of residuals test, transforming the data or following a non-parametric framework is recommended. In this data, a natural log transformation was not strong enough to make the p value greater than 0.05, so an inverse transformation method was used instead.

**Inverse Transformation:**

A new variable "InvCost" was created that contained the values of 1/Cost. Now, when the residuals are plotted, they fit around the QQPlot with less scatter as can be seen by Figure 4 above.

```
        Shapiro-Wilk normality test

data:  res
W = 0.99106, p-value = 0.2059
```

Transformed Data P-Value

## Final Conclusion:

When running the Shapiro test on this new transformed data, it was found that the p value = 0.2059 > α = 0.05. Therefore, at 5% significance, there is now sufficient evidence to conclude that the residuals follow a normal distribution using this transformed continuous data. These findings are further strengthened through visual evidence in figure 4 which shows that through inverse transformation, the normality of the data has increased due to less scatter present within the QQPlot.

### Assumption #2: There is a homogeneity of variance

It is assumed that the variances for each treatment are equal. Visually, this assumption can be explored by plotting the residuals of the linear mode. As can be seen by figure 5 below, the close correlation of the residuals to the fitted line was an indication that there was homogeneity of variance. To confirm this, a Levene's Test was conducted.

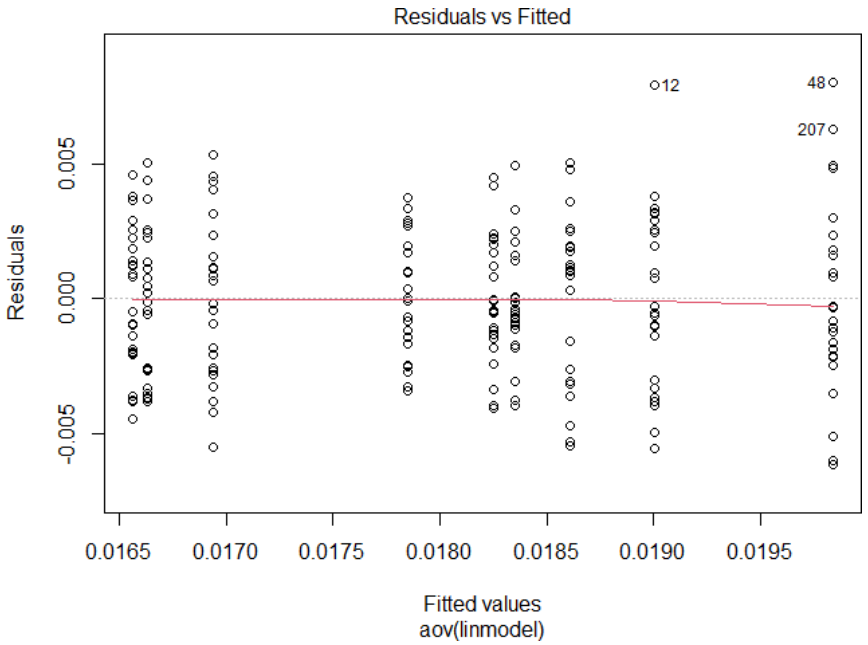


**Figure 5:** A plot of residuals vs the fitted line (See Appendix 2.3 for R Code)

**Null Hypothesis ($H_0$):** there is equal variance
**Alternate Hypothesis ($H_A$):** variance is not equal
**Significance Level, α:** 0.05
**Decision Rule**: Reject $H_0$ if the p value is < α
**Test Statistic:** Levene's Test for Homogeneity of Variance
**Test Statistic Value:** p value (see appendix 3.2 for R code)

```
Levene's Test for Homogeneity of Variance (center = median)
       Df F value Pr(>F)
group   8  1.4886 0.1629
      207
```

**Conclusion:**

As the p value = 0.1629 > α = 0.05, the null hypothesis is not rejected. Therefore, at 5% significance, there is sufficient evidence to conclude that there is indeed homogeneity of variance within this investigation.

### Assumption #3: Random and independent sampling

For this report, it is assumed that data acquired is random, and treatments are sampled independently from each other. It is assumed that this data collected from a seeded script factored in this consideration.

## Part 2

**Research Question:** What is the impact of the mode of transport, school of study or both factors on the average cost of weekly transport for students?

### Statistical Hypotheses

**Null Hypothesis ($H_0$):**

1. School: $\mu_{ENS} = \mu_{BUS} = \mu_{OTH}$
2. Mode: $\mu_{PUBLIC} = \mu_{PRIVATE} = \mu_{OTHER}$
3. Interaction: There is no interaction effect

**Alternate Hypothesis ($H_A$):**

1. School: $\mu_i \neq \mu_j$ where $i \neq j$ - at least one is different from the others
2. Mode: $\mu_i \neq \mu_j$ where $i \neq j$
3. Interaction: There is an interaction effect

The general linear model for an ANOVA is:

$$\textbf{Observation} = \textbf{Overall Mean} + (\textbf{Factor 1})_i + (\textbf{Factor 2})_j + (\textbf{Factor 1} \times \textbf{Factor 2})_{ij} + \textbf{Error}_{ijk}$$

When applied to this scenario, we can derive the following model:

$$\textbf{Cost}_{ijk} = \mu + \textbf{Mode}_i + \textbf{School}_j + (\textbf{Mode} \times \textbf{School})_{ij} + \xi_{ijk}$$

**Where**

- **i =** 1, 2, 3
- **j =** 1, 2, 3
- **k (replications) =** 1, 2, …, 23, 24

A table was first created in R to showcase the frequency distribution of different treatment combinations (school and mode) – See Appendix 1.1 for this contingency table. The replications between all the treatments were the same at 24, so this data was confirmed as a balanced design However, that table is reminiscent of a Chi-Squared contingency table and was only used to help create the experimental design below (Table 3). As can be seen, this study had a 3x3 factorial design with 9 treatments. For table 3's factorial design, each cell with μ corresponds to the treatment mean of the dependent variable (Cost), as this is an Analysis of Variance test.

**Table 3:** Table showcasing the data's 3x3 Factorial Design (See Appendix 1.1 for preliminary table and R-code)

| **Mode** | **School** | | |
|---|---|---|---|
| | **Level 1: ENS** | **Level 2: BUS** | **Level 3: OTH** |
| **Level 1: Public** | $\mu_{T1}$ | $\mu_{T4}$ | $\mu_{T7}$ |
| **Level 2: Private** | $\mu_{T2}$ | $\mu_{T5}$ | $\mu_{T8}$ |
| **Level 3: Other** | $\mu_{T3}$ | $\mu_{T6}$ | $\mu_{T9}$ |

### Projected ANOVA

**Table 4:** Projected ANOVA table

| **Source of Variation** | **General Rule** | **Working** | **Degrees of Freedom** |
|---|---|---|---|
| **Mode** | (i - 1) | 3 – 1 | **2** |
| **School** | (j - 1) | 3 – 1 | **2** |
| **Mode × School** | (i - 1) × (j - 1) | (3 – 1) × (3 – 1) | **4** |
| **Error (Residual)** | i × j × (reps – 1) | 3 × 3 × (24 - 1) | **207** |
| **Total:** | | 2 + 2 + 4 + 207 | **215** |

## Part 3 - Results

**Statistical Hypotheses**

**Null Hypothesis ($H_0$):**

- School: $\mu_{PUBLIC} = \mu_{PRIVATE} = \mu_{OTHER}$
- Mode: $\mu_{ENS} = \mu_{BUS} = \mu_{OTH}$
- Interaction: There is no interaction

**Alternate Hypothesis ($H_A$):**

- School: $\mu_i \neq \mu_j$ where $i \neq j$
- Mode: $\mu_i \neq \mu_j$ where $i \neq j$
- Interaction: There is an interaction

**Significance Level, α:** 0.05

**Decision Rule**: Reject $H_0$ if the p value is $< \alpha = 0.05$

**Test Statistic:** F-Statistic

**Test Statistic Value:** R Output (See appendix 3.3 for R Code)

```
               Df    Sum Sq   Mean Sq F value   Pr(>F)
sch             2 0.0000007 3.400e-07   0.042    0.959
mode            2 0.0002169 1.084e-04  13.318 3.62e-06 ***
sch:mode        4 0.0000237 5.920e-06   0.727    0.574
Residuals     207 0.0016855 8.140e-06
```

**Conclusion:**

1. **Interaction:**
   As the p value = 0.574 > α = 0.05, the null hypothesis is not rejected. This means that at 5% significance, sufficient evidence is present to conclude that there is no interaction between the school and mode of transport.
   - As no significant interaction has been found, we must now look at the main effects.
2. **Main Effect 1: School**
   As the p value = 0.959 > α = 0.05, the null hypothesis is not rejected. Therefore, at 5% significance, all means within the factor school are not significantly different.
3. **Main Effect 2: Mode**
   As the p value = $3.62*10^{-6}$ < α = 0.05, we reject the null hypothesis in favour of the alternate. That is to say that at 5% significance, there is evidence to conclude that there may be some difference in mean within the levels in mode
   - As the null hypothesis has been rejected, a post-hoc test must be conducted to confirm if differences occurred between groups (Figure 6)

```
LSD t Test for InvCost

Mean Square Error:  8.142443e-06

mode,  means and individual ( 95 %) CI

           InvCost         std  r        LCL        UCL        Min        Max
Other   0.01815021 0.002216410 72 0.01748722 0.01881320 0.01417434 0.02331002
Private 0.01915265 0.003380587 72 0.01848966 0.01981563 0.01312508 0.02789400
Public  0.01671118 0.002782360 72 0.01604820 0.01737417 0.01143641 0.02231147

Alpha: 0.05 ; DF Error: 207
Critical Value of t: 1.97149

least Significant Difference: 0.0009376069

Treatments with the same letter are not significantly different.

           InvCost groups
Private 0.01915265      a
Other   0.01815021      b
Public  0.01671118      c
```

**Figure 6:** Post-Hoc Test R Output for Mode (The Relevant Main Effect) – See Appendix 3.4 for R Code

**Interpretation of Results**

Upon conducting the post-hoc test, it can be seen in the R output above that the means in each treatment in mode are all significantly different from each other, as denoted by each treatment having a different letter. These results from the post-hoc were plotted in figure 7. The lines extending from each of the levels within mode correspond to standard error. Thus, it can be seen visually that the means in private, public, and other modes are all significantly different.

*Therefore:* $T_{MODE} \neq T_{PUBLIC} \neq T_{OTHER}$

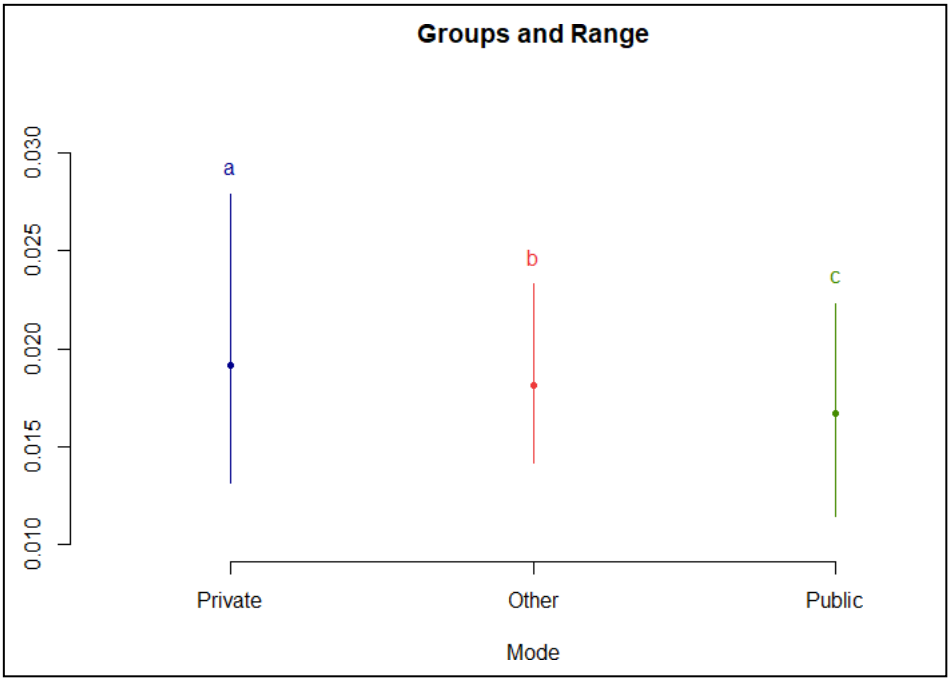


**Figure 7:** A standard error plot of each group in mode and its standard error range (See Appendix 2.4 for R Code)

**Conclusion**

Therefore, to address the research question, this factorial ANOVA study has found that there is no significant interactive effect between the mode of transport and the school of study in regard to the average cost of weekly transport. A post-hoc test was conducted on one of the main effects (mode) which was significant; this post-hoc found that the means of private, public, and other modes of transport were significantly different from each other.

These results are reflected in a 2019 academic study on the transport behaviour of academic communities (Romanowska et al, 2019) , which found that weekly transport cost was subject to a multitude of complex factors beyond the two independent variables chosen for this study. They included car availability, trip distance, accessibility of certain transport modes, and personal values; all of which are factors that differ between individual people. However, environmental science students were found to be more cognizant of the importance of ecologically friendly transport options and were more likely to spend more on public transport rather than on private methods – a relationship that was not clearly exhibited in this particular study. Further research into this area of study may help Universities to better understand the behaviours of University students and cater specific infrastructure around different schools of the campus to sustainably manage demand for different transport modes.

## References


- Revelle, W. (2020) psych: Procedures for Personality and Psychological Research, Northwestern University, Evanston, Illinois, USA, https://CRAN.R-project.org/package=psych Version = 2.0.8,.
- Felipe de Mendiburu (2020). agricolae: Statistical Procedures for Agricultural Research. R packag e version 1.3-3. https://CRAN.R-project.org/package=agricolae
- John Fox and Sanford Weisberg (2019). An {R} Companion to Applied Regression, Third Edition. T housand Oaks CA: Sage.URL:https://socialsciences.mcmaster.ca/jfox/Books/Companion/
- Jeroen Ooms (2020). writexl: Export Data Frames to Excel 'xlsx' Format. R package version 1.3.1. https://CRAN.R-project.org/package=writexl
- Romanowska, A., Okraszewska, R. and Jamroz, K., (2019). A Study of Transport Behaviour of Aca demic Communities. *Sustainability*, 11(13), p.3519.
- Script was provided by James McBroom from Learning at Griffith, Griffith University


## Appendix

### Appendix 1: R code for summary statistic generation

#### 1.1 R Code and table used to create preliminary design, so that factorial design can be made

| Mode | School | | |
|---|---|---|---|
| | ENS | BUS | OTH |
| Public | 24 | 24 | 24 |
| Private | 24 | 24 | 24 |
| Other | 24 | 24 | 24 |

```
table(mode, sch) # Create a Table
```

#### 1.2 Table showing information for each treatment

```
library(psych) # Install Psych Library
Results = describeBy(Cost ~ mode + sch, # Cost with regards to mode and sch
           mat = TRUE, # For better formatting
           data = Transport, digits = 2) # Generate table
```

#### 1.3 Code for exporting data to excel for formatting

```
#Export to Excel
library(writexl) # Install WriteXL library
# include correct path for document to be written to
write_xlsx(Results, "C:\\Users\\himat\\OneDrive\\Desktop\\EDA_petrol.xlsx")
```

## Appendix 2: Graph R Code

### 2.1 Diagram depicting boxplots for each treatment in the data

```
library(ggplot2) # Install GGPlot2 library
ggplot(data = Transport) +
  aes(x=sch, y = Cost, fill = sch) +
  geom_boxplot(varwidth = TRUE) + # specify to create boxplot
  facet_wrap(~mode) + theme(legend.position = "bottom") + # formatting
  scale_fill_manual(values = c("cyan1", "cyan3", "cyan4")) # colours for each variable
```

### 2.2 Diagram depicting histograms for each treatment in the data

```
library(ggplot2) # Install GGPlot2 library
ggplot(data = Transport) + # specify data
  aes(x = Cost, y = ..density..) +
  geom_histogram(binwidth = 6, fill = "lightblue", colour = "black") + # bar colours & outline colors
  geom_density() + facet_grid(mode ~ sch) # include density plot
```

### 2.3 Code for graphs used in assumption testing

```
plot(avmodel, 1) # 1 = Residuals vs Fitted Graph for Variance
plot(avmodel, 2) # 2 = QQPlot for Normality
```

### 2.4 Post Hoc test graph

```
library(agricolae) # Post Hoc Test
posthoc = LSD.test(avmodel, "mode", console = TRUE) # Least Significant Difference Test
# "mode" = main effect test is being done on
plot(posthoc, xlab = "Mode") # Plot PostHoc results
```

## Appendix 3: Hypothesis Testing

### 3.1 Shapiro Test of Normality

```
res=avmodel$residuals # Isolate residuals from model
shapiro.test(res) # Normality Test code
```

### 3.2 Levene's Test of Variance

```
library(car) # Install car library for Levene's test
leveneTest(linmodel) # Variance Test
```

### 3.3 R Code for ANOVA Test

```
#Using the Cost fails normality of residuals
#Run Inverse Transformation on Data
#Run the ANOVA test again with the transformed variable "InvCost)
InvCost = 1/(Cost) #transforming Cost using inverse (1/x)
linmodel = lm(InvCost ~ sch*mode, data = Transport) # Create Linear Model
avmodel = aov(linmodel) # Run ANOVA test on linear model
summary(avmodel) # Summarize model
```

### 3.4 R Code for Post-Hoc Test

```
library(agricolae) # Post Hoc Test
posthoc = LSD.test(avmodel, "mode", console = TRUE) # Least Significant Difference Test
# "mode" = main effect test is being done on
plot(posthoc, xlab = "Mode") # Plot PostHoc results
```

## Appendix 4: R Code to generate citations for relevant libraries used

```
# Generate Citations
citation("psych")
citation("agricolae")
citation("car")
citation("writexl")
```