Project Report

Mobile network preference of Langara students of spring 2019

Abhishek Singh, Amit Verma, Ravinder Singh, Sparsh Sharma

|April 03, 2019|

Contents

Introduction

This project is basically based on analysis of data which has been collected in Langara College through online questionnaire from students enrolled in spring semester 2019. The dataset which is analyzed contains data related to mobile network used by the students. It includes variables that gives the type of student, network name, cost, data plan, preferred time of usage, preferred purpose of usage and customer satisfaction rating.

This dataset is great for evaluating simple regression model, as this method and variables described in the dataset can be used practically by mobile network operators to decide what features they need to add to their network plan depending upon the different variables. For instance, there can be a separate plan for students which allows use of data for social networking sites only targeting those whose preferred purpose is social networking.

Variables were studied to find which network is most preferred by students of Langara studying in spring semester 2019, what is their most preferred data plan along with the most preferred time and purpose of use. It was also studied if there exist a linear correlation between cost per month and the data plan through correlation analysis and further this correlation was quantified through linear regression analysis. Chi square analysis was done to check if there is any difference between international students and domestic students of Langara College studying of 2019 spring semester in terms of preferred time when they use data the most and their preferred network.

Dataset Details

The data used for analysis consist of 100 rows and 8 Columns

Variable Explanation

Sr. No.	Variable Name	Description	Values	Type	Scale
1	Id	Unique id for each student who participated in survey			
2	Mobile Network	Name of mobile network used by student	*Fido, Telus, Rogers, Bell, Freedom, Chatr*	Categorial	Nominal
3	Data Plan	Number of data bytes of pack being used by the student (in GB per month)	NA	Quantitative	Ratio
4	Cost	Cost of the data plan being used by student (in CAD per month)	NA	Quantitative	Ratio
5	Purpose	Preferred purpose of usage	*Email, Social Network, News/Banking/Music/Email, Gaming, Downloading, Telephone/ Video Chat*	Categorial	Nominal
6	Student Type	Type of student	Domestic, International	Categorial	Nominal
7	Preferred Time	Preferred time of day when the student prefers to use data the most	*Morning, Afternoon, Evening, Night*	Categorial	Ordinal
8	Customer_Satisfaction	Satisfaction level of students with their mobile network	*Very satisfied, Satisfied, Neutral, Dissatisfied, Very Dissatisfied*	Categorial	Ordinal

Software & Tools

Python, RStudio, Excel, MS WORD

Sampling Design

Convenience and voluntary response sampling were used in which students studying in Langara College (spring semester 2019) were surveyed through an online questionnaire. As it was not possible to reach every student of each class the questionnaire link was posted on several whats app groups of different classes of 2019 spring semester of Langara College.

Sample size n = 100.

Convenience sampling technique was used as it was easy to approach the students of certain classes due to time and cost limitations.

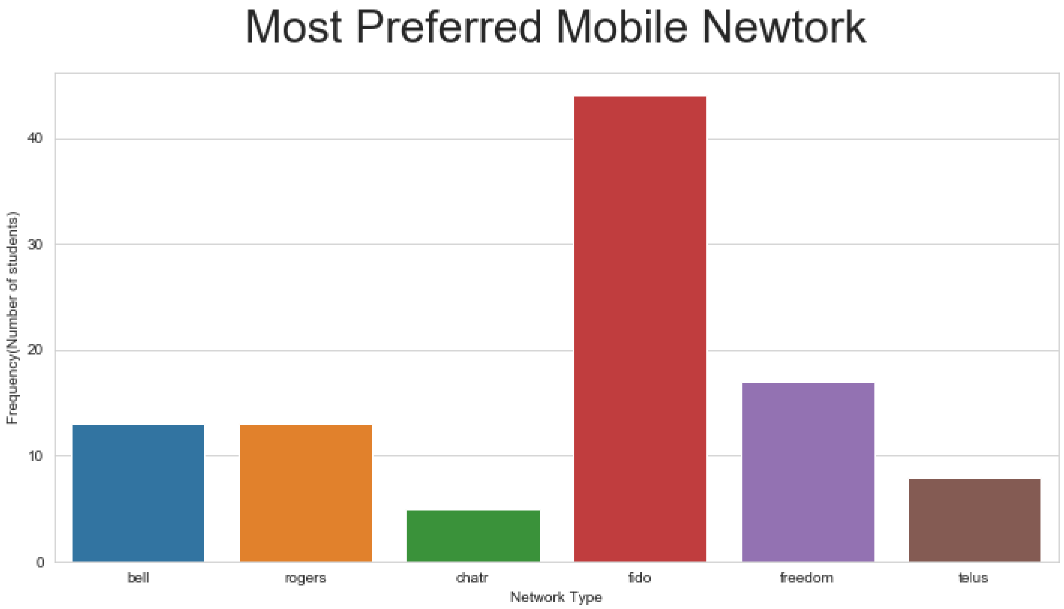
Bias

1. Since the sample is subset of entire Langara students of 2019 spring semester, naturally there will be sampling error
2. Only those classes were approached which were easy to reach so there are high chances that the results produced may favor certain outcomes
3. The individuals who took trouble to respond to our survey may not represent a clearly defined population
4. Since, online questionnaire was used as medium of survey, it may result in under coverage of certain section of population and non- response bias (some students may not respond to our survey at all)

Analysis & Discussion

Univariate Analysis

1. Mobile Network: The most preferred mobile network by the Langara students enrolled in spring semester 2019 is Fido. More than 40% of the sample population use Fido as their preferred network. Around 18% students use Freedom mobile network while Chatr is the least preferred mobile network for the students comprising of just 5% of the sample population.



Interval estimation of population i.e Langara students of 2019 spring semester for Mobile Network Fido:

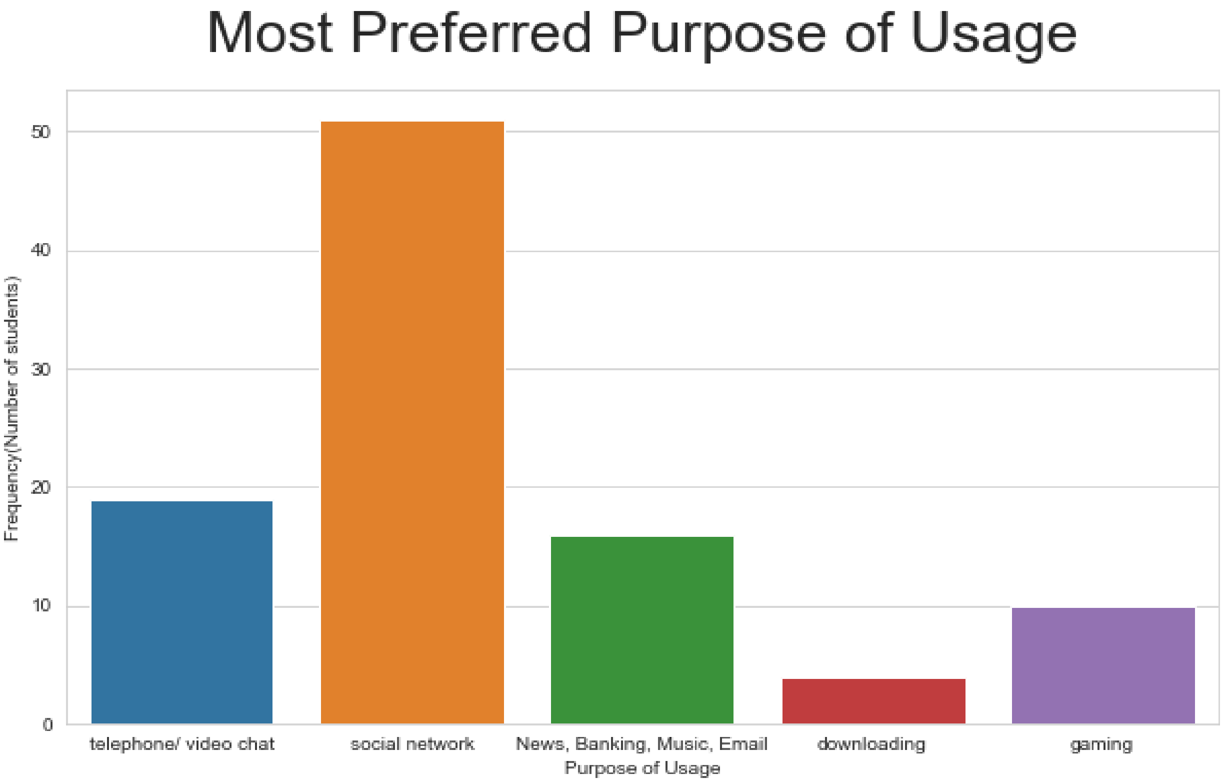
$$\bar{p} \pm z_{\alpha/2}\sqrt{\frac{\bar{p}(1-\bar{p})}{n}}$$

where: 1 – a is the confidence coefficient = 0.95 and $z_{\alpha/2}$ is the z value providing an area of a/2 in the upper tail of the standard normal probability distribution = 1.96 and $\bar{p}$ is the sample proportion = 42/n = 42/100 = 0.42

Interval estimate = 0.42 $\pm$ 0.0967

Hence, we are 95% confident that all the students in population who prefer Fido as their mobile network will lie between 32.33% and 51.67%

2. Purpose of Usage: The most preferred purpose for using internet by the Langara students enrolled in spring semester 2019 is Social Networking. 51% of the students in the sample population use mobile phones for accessing social networking sites. 18% of the students use mobile phones for telephone/video chat while 16 % use for accessing news, music, email and banking services. The least purpose of usage was downloading with only 5% Langara students using mobile phones for downloading purpose.



Interval estimation of population i.e Langara students of 2019 spring semester for purpose of usage 'social networking':

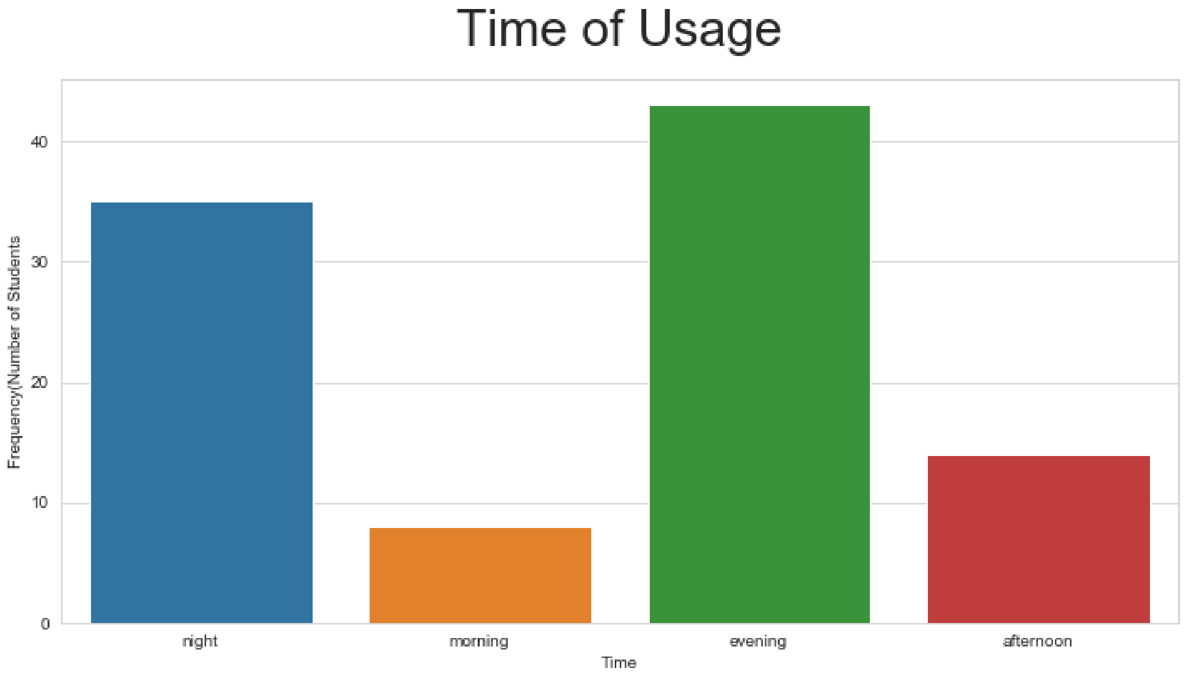
$$\bar{p} \pm z_{\alpha/2}\sqrt{\frac{\bar{p}(1-\bar{p})}{n}}$$

where: 1 – a is the confidence coefficient = 0.95 and $z_{\alpha/2}$ is the z value providing an area of a/2 in the upper tail of the standard normal probability distribution = 1.96 and $\bar{p}$ is the sample proportion = 50/n = 50/100 = 0.5

Interval estimate = 0.50 $\pm$ 0.098

Hence, we are 95% confident that all the students in population whose purpose of data usage is social networking will lie between 40.2 % and 59.8 %

3. Time of Usage: The most preferred time of data usage by the students of Langara enrolled in spring semester 2019 is Evening. 35% of the Langara students use mobile phones at night while 13% use at afternoon. The graph below shows the least preferred time of usage was in the morning.



Interval estimation of population i.e Langara students of 2019 spring semester for preferred time 'evening':

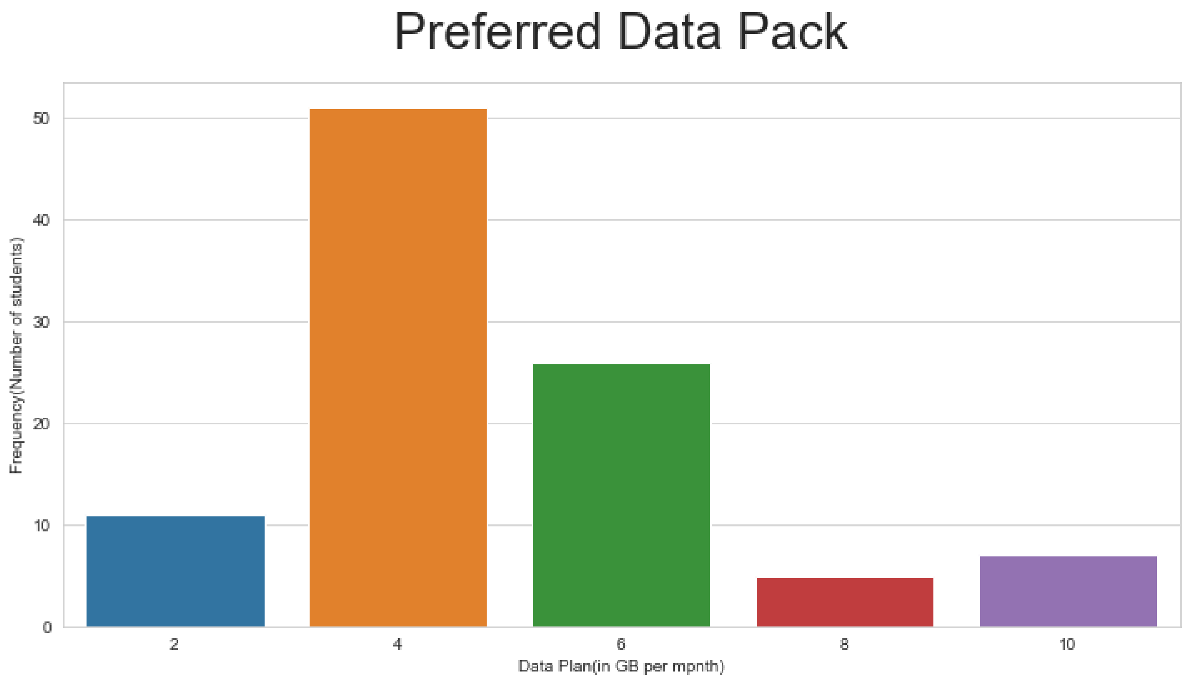
$$\bar{p} \pm z_{\alpha/2}\sqrt{\frac{\bar{p}(1-\bar{p})}{n}}$$

where: 1 – a is the confidence coefficient = 0.95 and $z_{\alpha/2}$ is the z value providing an area of a/2 in the upper tail of the standard normal probability distribution = 1.96 and $\bar{p}$ is the sample proportion = 42/n = 42/100 = 0.42

Interval estimate = 0.42 $\pm$ 0.0967

Hence, we are 95% confident that all the students in population who prefer evening as their preferred time to use data the most will lie between 32.33 % and 51.67 %

4. Preferred Data Pack: The most preferred data pack by the students of Langara enrolled in spring semester 2019 is 4GB per month with more than half of the sample population opting for this data pack. 26% of students opted for 6GB per month data pack while the least preferred data pack as of 8GB per month with only 4% students choosing this data pack.



Interval estimation of population i.e Langara students of 2019 spring semester for 4 GB data plan:

$$\bar{p} \pm z_{\alpha/2}\sqrt{\frac{\bar{p}(1-\bar{p})}{n}}$$

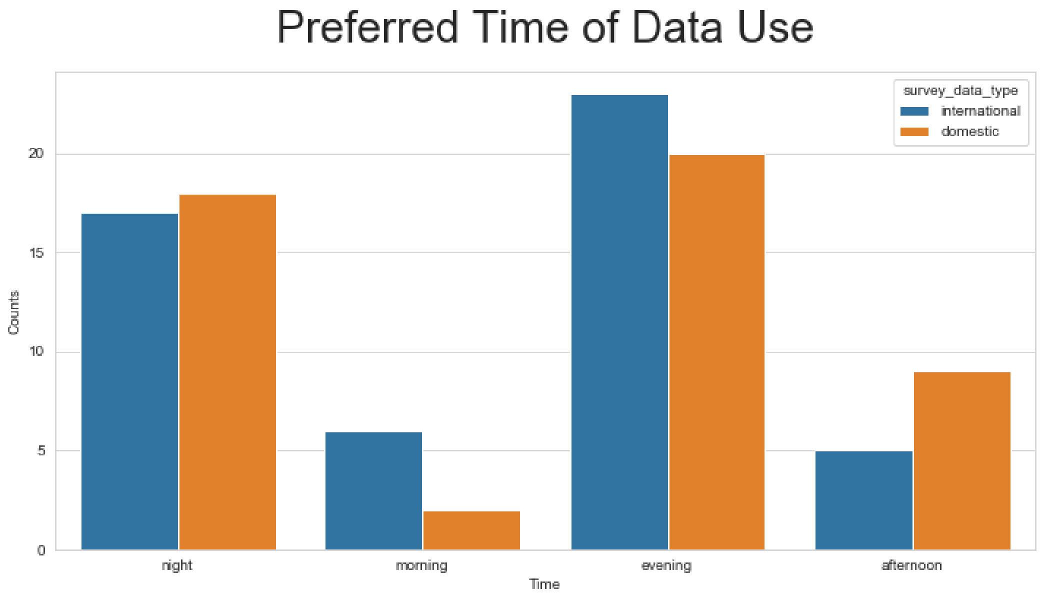
where: 1 – a is the confidence coefficient = 0.95 and $z_{\alpha/2}$ is the z value providing an area of a/2 in the upper tail of the standard normal probability distribution = 1.96 and $\bar{p}$ is the sample proportion = 50/n = 50/100 = 0.5

Interval estimate = 0.50 $\pm$ 0.098

Hence, we are 95% confident that all the students in population who prefer 4GB data per month will lie between 40.2% and 59.8%.

Bivariate Analysis

1. Chi-square test | Student Type and Preferred Time : A chi-square test of independence was performed to examine the relation between type of student and preferred time of use.
   The relation between these variables was significant, X2 (7.21, N = 100) = 3.3421, both Domestic and International students preferred same times for using data which is quite evident from the side by side bar chart given below:



The chart shows that he most preferred time of data use is evening for both international as well as domestic students. 28% of international students and 20% domestic students in Langara college use data in the evening respectively. The second most preferred time of data use for both international and domestic students is at night with 17% and 18% students using data at night respectively. The least preferred time of data use is at morning with only 6% international student and 2% domestic students.

Analysis in details:

Column Labels	Afternoon	evening	morning	night	Grand Total
Row Labels					
Domestic	9	20	2	18	49
International	5	23	6	17	51
Grand Total	**14**	**43**	**8**	**35**	**100**

Column Label	Morning	Afternoon	Evening	Night	Total
Row Label					
Domestic	3.92	6.86	21.07	17.15	**49**
International	4.08	7.14	21.93	17.85	**51**
Total	**8**	**14**	**43**	**35**	**100**

- **Null Hypothesis (H_0):** There is no difference between domestic and international student, in terms of their preferred time of internet usage, in the population
- **Alternative Hypothesis (H_a):** There is a difference between domestic and international students in terms of their preferred time of internet usage, in the population.

Calculation of expected frequency $e = \frac{(\boldsymbol{Row\ Total})(\boldsymbol{Column\ Total})}{(\boldsymbol{Size\ of\ Sample})}$

Calculating the χ^2 statistic:

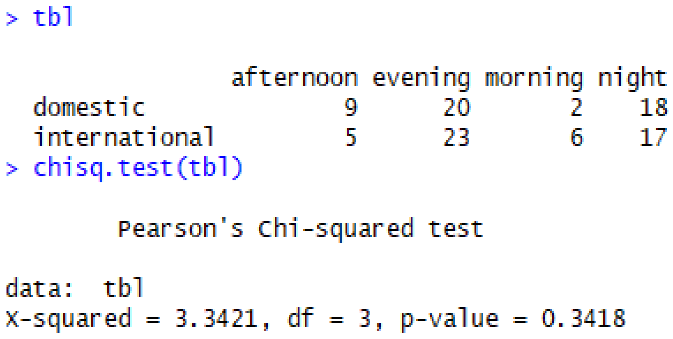
$\chi^2 = \sum \frac{(f-e)^2}{e}$ = 3.34

Degree of freedom, dof = (#rows -1) x (# columns -1) = (2-1) *(4-1) =3

The decision point for Dof= 3 is 7.21

Since, the calculated chi square is less than the DP (χ2 < DP). Therefore, we fail to reject H0 and conclude **H_a** is true. There is a no strong differences between domestic and international student, in terms of their preferred time of internet usage.

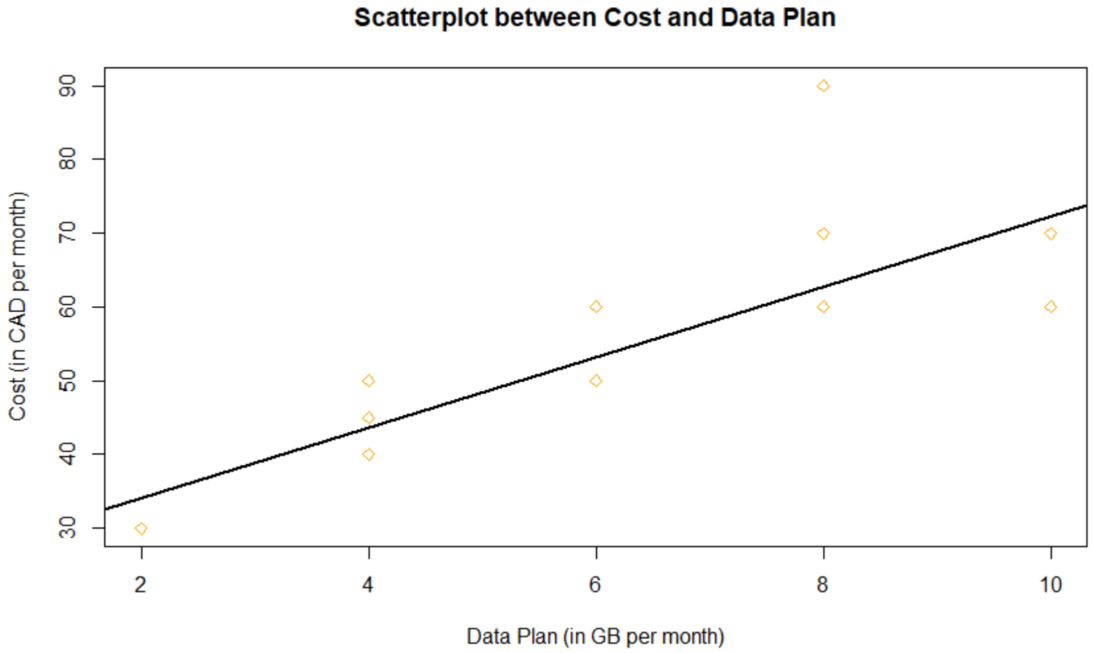
Chi square analysis with R:



```
> tbl

                afternoon evening morning night
  domestic              9      20       2    18
  international         5      23       6    17
> chisq.test(tbl)

        Pearson's Chi-squared test

data:  tbl
X-squared = 3.3421, df = 3, p-value = 0.3418
```

2. Correlation Analysis | Cost and Data Plan:

A correlation analysis was done to check the association between the response variable ‘Cost’ and the explanatory variable ‘Data Plan’.



Insights from scatterplot: The scatterplot between Cost and Data Plan suggest a positive linear relationship with few outliers. Higher GBs per month are associated with higher cost. The relation is not perfect thought, there are few points in the scatter diagram which do no fall on straight line.

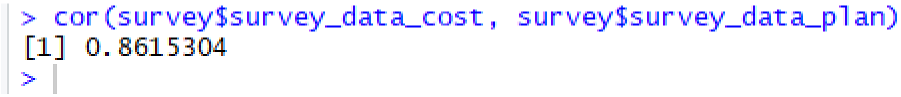
Analysis in detail:

- **Null Hypothesis (H_0):** There is no linear correlation between the cost per month and data plan in the population of all Langara student studying in spring semester 2019.

- **Alternative Hypothesis (H_a):** There is a linear correlation between the cost per month and data plan in the population of all Langara student studying in spring semester 2019.

Decision point for n = 0.197

Correlation coefficient using Rstudio = 0.8615

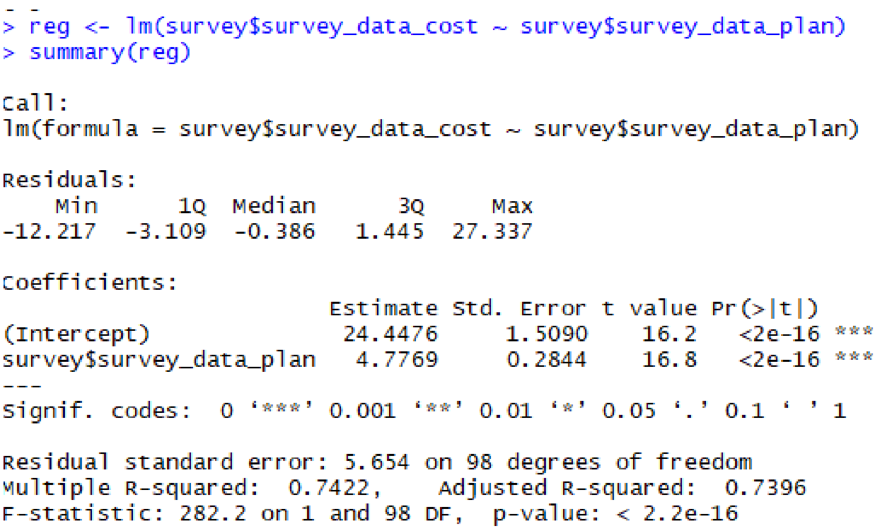


```
> cor(survey$survey_data_cost, survey$survey_data_plan)
[1] 0.8615304
> 
```

Since, the correlation coefficient is greater than decision point we can reject **H_0** and conclude there is a linear relationship between cost per month and data plan in the population of Langara students of 2019 spring semester.

3. Linear Regression | Cost and Data Plan:

The correlation analysis does suggest a linear relationship between cost per month and data plan but to see if there is a cause and effect relationship, we performed linear regression analysis with the response variable ‘cost’ (per month) and explanatory variable ‘data plan’. Linear regression quantifies the relationship between a response variable and explanatory variable.



```
> reg <- lm(survey$survey_data_cost ~ survey$survey_data_plan)
> summary(reg)

Call:
lm(formula = survey$survey_data_cost ~ survey$survey_data_plan)

Residuals:
    Min      1Q  Median      3Q     Max 
-12.217  -3.109  -0.386   1.445  27.337 

Coefficients:
                        Estimate Std. Error t value Pr(>|t|)    
(Intercept)              24.4476     1.5090    16.2   <2e-16 ***
survey$survey_data_plan   4.7769     0.2844    16.8   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5.654 on 98 degrees of freedom
Multiple R-squared:  0.7422,	Adjusted R-squared:  0.7396 
F-statistic: 282.2 on 1 and 98 DF,  p-value: < 2.2e-16
```

The estimated linear regression equation is derived as:

$\hat{y} = 24.45 + 4.78x$

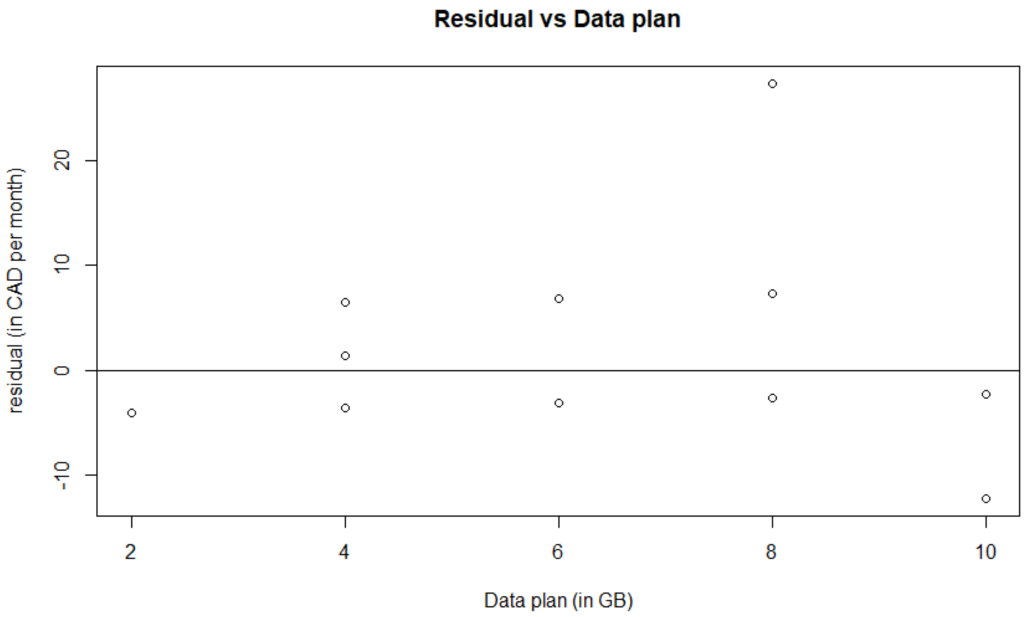
where $\hat{y}$ is predicted value of cost per month for x GB of data per month

Insights:

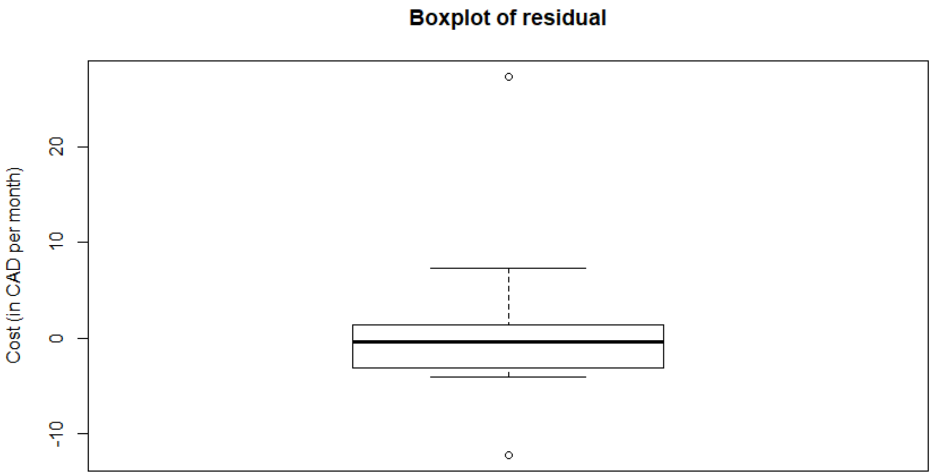
1. For every additional GB of data to plan we can expect an increase of 4.78 CAD in cost per month.
2. When the student does not opt for internet facility (x = 0), the basic cost of network is 24.45 CAD per month.
3. The residual standard error is 5.654 which tells us the typical distance a value is from the regression line or the average prediction error.

Checking the validity of Regression Model:

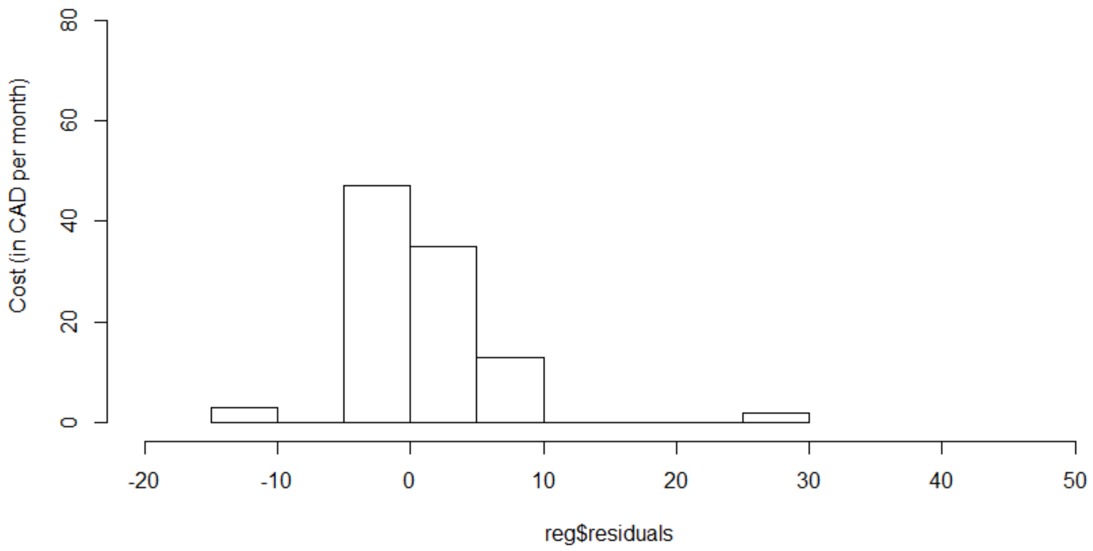
1. Linearity: The scatterplot between residual and Data plan does not show any curve, therefore the graph is linear.



2. Normality: The boxplot and histogram of residuals do not show normal distribution.



Histogram of residuals



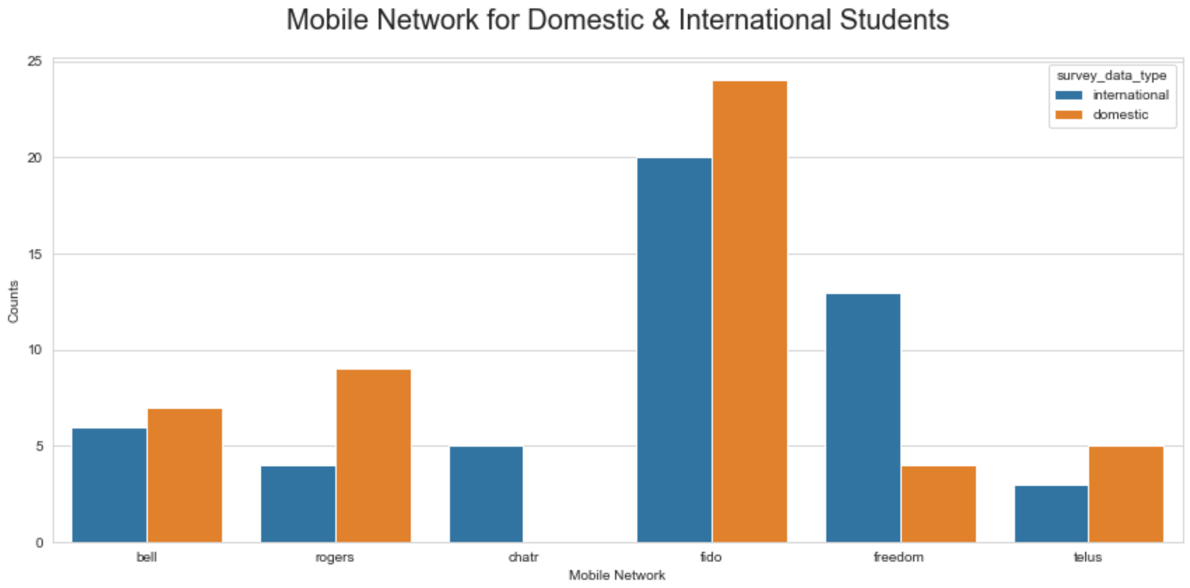
3. Homoscedasticity:

The residual plot does not show any fan shaped patterns; therefore, the graph can be assumed to be homoscedastic.

All three assumptions of linear regression were not satisfied; hence this regression model is not valid.

4. Chi-square Analysis | Student Type and Mobile Network:

   A chi-square test of independence was also performed to examine the relation between type of student and preferred mobile network. The relation between these variables was significant, X2 (11.07, N = 100) = 12.59, therefore, there is a significant difference between Domestic and International students in terms of preferred mobile network which is quite evident from the side by side bar chart given below:



While Fido was most preferred by both international and domestic students but the second preferred network for international student was freedom while domestic students preferred rogers. The least preferred network by international students is Telus and by domestic students is Chatr.

Analysis in details:

Labels	Bell	Chatr	Fido	Freedom	Rogers	Telus	Total
Domestic	7	0	24	4	9	5	49
International	6	5	20	13	4	3	51
Total	**13**	**5**	**44**	**17**	**13**	**8**	**100**

- **Null Hypothesis (H_0):** For all the student types in the population, there is no difference between domestic and international students, in terms of their preferred mobile network.
- **Alternative Hypothesis (H_a):** For all the student types in the population, there is a difference between domestic and international students, in terms of their preferred mobile network.

Calculation of expected frequency

$$\boldsymbol{e} = \frac{(\boldsymbol{Row\ Total})(\boldsymbol{Column\ Total})}{(\boldsymbol{Size\ of\ Sample})}$$

Labels	Bell	Chatr	Fido	Freedom	Rogers	Telus	Total
Domestic	6.37	2.45	21.56	8.33	6.37	3.92	49
International	6.63	2.55	22.44	8.67	6.63	4.08	51
Grand Total	**13**	**5**	**44**	**17**	**13**	**8**	**100**

Calculating the $\boldsymbol{\chi^2}$ statistic

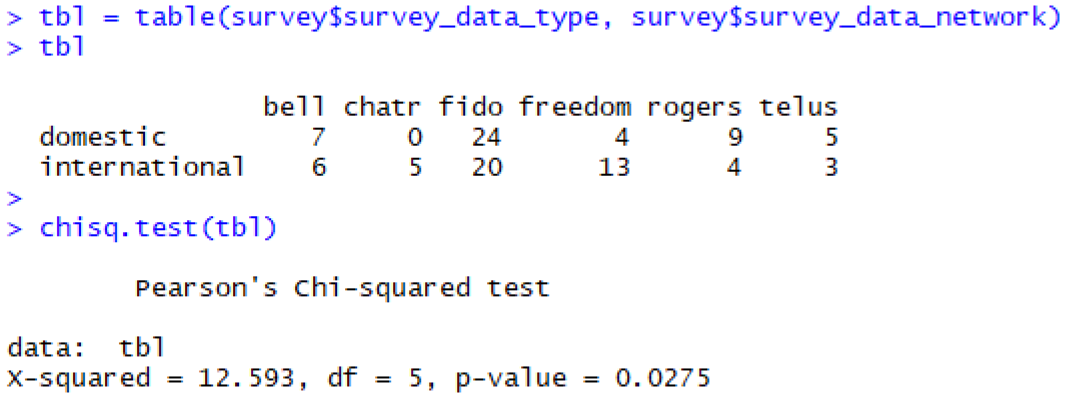
$\boldsymbol{\chi^2} = \sum \frac{(\boldsymbol{f-e})^2}{\boldsymbol{e}}$ =12.59

Degree of freedom, dof = (#rows -1) x (# columns -1) = (2-1) *(6-1) =5

The decision point for Dof= 5 is 11.07

Since, the calculated chi square is greater than the DP (χ2 > DP). Therefore, we reject H0 and conclude **H_a** is true. There are significant differences between domestic and international student, in terms of their preferred network type.

Chi square analysis with R:



```
> tbl = table(survey$survey_data_type, survey$survey_data_network)
> tbl

                bell chatr fido freedom rogers telus
  domestic         7     0   24       4      9     5
  international    6     5   20      13      4     3
>
> chisq.test(tbl)

        Pearson's Chi-squared test

data:  tbl
X-squared = 12.593, df = 5, p-value = 0.0275
```

Conclusion

- Both international and domestic students of Langara 2019 spring semester preferred same
- times to use their mobile internet the most.
- But there were differences observed in the choice of mobile network, while Fido was most preferred by both international and domestic students but the second preferred network for international student was freedom while domestic students preferred rogers. The least preferred network by international students is Telus and by domestic students is Chatr.
- The cost per month has a positive linear association with data plan. The strength of this relation was quite significant.
- The overall most preferred mobile network is Fido among the Langara students of 2019 spring semester.
- Most preferred purpose of usage was Social Networking and most preferred time of usage was evening.

Limitation:

- The correlation analysis does suggest that there is a positive linear correlation between cost per month and data plan, but it does not suggest that changing the data plan will result in proportional change in cost per month. For confirming if there is a cause and effect relationship, we did linear regression analysis.
- All three assumptions of regression analysis have not been met; therefore, the regression model is not valid.
- Since, the sample size n = 100 is very small as compared to the whole population of Langara students of 2019 spring semester, the study consists of sampling error (which can be reduced by increasing the sample size).
- Online questionnaire was used as medium of survey; it may result in under coverage of certain section of population and non- response bias (some students may not have responded to our survey at all).

Appendix

1. Online Survey Link: http://danasurvey.x10host.com/
2. Python Documentation for seaborn: https://seaborn.pydata.org/
3. RStudio: https://www.rstudio.com/

Glossary

Keyword	Description
Population	The set of all elements/subjects of interest in a study
Sample	The subset of population that we study
Variables	Characteristic of a case
Univariate	One variable of observations
Bivariate	Pairs of linked numerical observations
Convenience Sample	A sample selected by taking the members of the population that are easiest to reach
Categorical Data	Labels or names used to identify an attribute of each element
Quantitative Data	Quantitative data indicate how many or how much