Exploring Young People

Cassie Lo

DATASET/SUMMARY

I'm using the dataset from Kaggle (https://www.kaggle.com/miroslavsabo/young-people-survey#responses.csv), which is a young people survey. It provides 150 questions (preferences, interests, habits, opinions, and fears……) for 1010 Slovakian students, aged between 15-30. Below are the variables I used in this project:

Opera
Don't enjoy at all 1-2-3-4-5 Enjoy very much (integer)

Romantic movies
Don't enjoy at all 1-2-3-4-5 Enjoy very much (integer)

Shopping
Not interested 1-2-3-4-5 Very interested (integer)

Spiders
Not afraid at all 1-2-3-4-5 Very afraid of (integer)

Life struggles
I cry when I feel down or things don't go the right way.: Strongly disagree 1-2-3-4-5 Strongly agree (integer)

Age
(integer)

Gender
Female - Male (categorical)

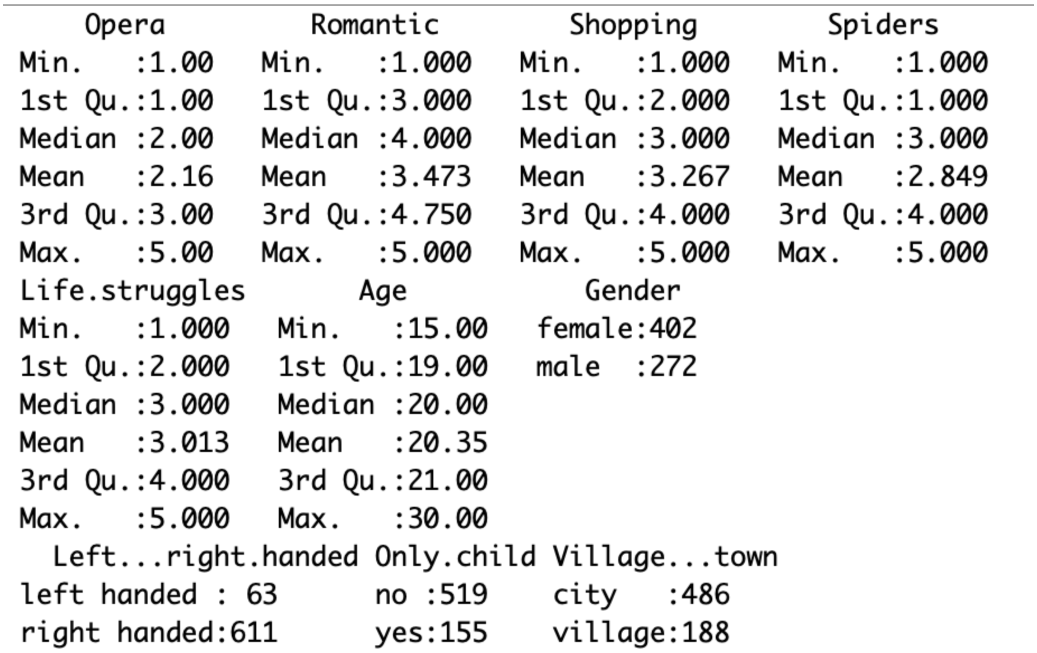
Left - right handed
I am: Left handed - Right handed (categorical)

Only child
I am the only child: No - Yes (categorical)

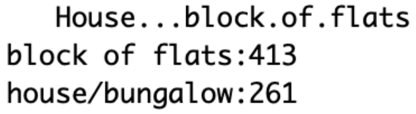
Village - town
I spent most of my childhood in a: City - village (categorical)

House - block of flats
I lived most of my childhood in a: house/bungalow - block of flats (categorical)

Summary statistics for the variables I used can be found below:



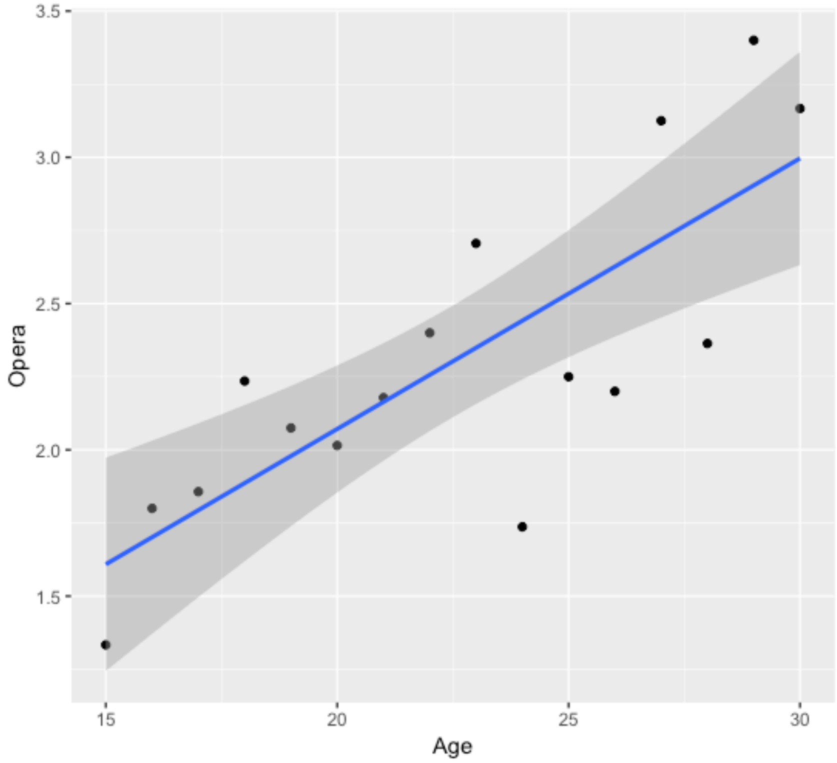
```
     Opera          Romantic        Shopping        Spiders     
 Min.   :1.00   Min.   :1.000   Min.   :1.000   Min.   :1.000  
 1st Qu.:1.00   1st Qu.:3.000   1st Qu.:2.000   1st Qu.:1.000  
 Median :2.00   Median :4.000   Median :3.000   Median :3.000  
 Mean   :2.16   Mean   :3.473   Mean   :3.267   Mean   :2.849  
 3rd Qu.:3.00   3rd Qu.:4.750   3rd Qu.:4.000   3rd Qu.:4.000  
 Max.   :5.00   Max.   :5.000   Max.   :5.000   Max.   :5.000  
 Life.struggles       Age             Gender   
 Min.   :1.000   Min.   :15.00   female:402  
 1st Qu.:2.000   1st Qu.:19.00   male  :272  
 Median :3.000   Median :20.00               
 Mean   :3.013   Mean   :20.35               
 3rd Qu.:4.000   3rd Qu.:21.00               
 Max.   :5.000   Max.   :30.00               
   Left...right.handed Only.child Village...town
 left handed : 63      no :519    city   :486   
 right handed:611      yes:155    village:188   
```

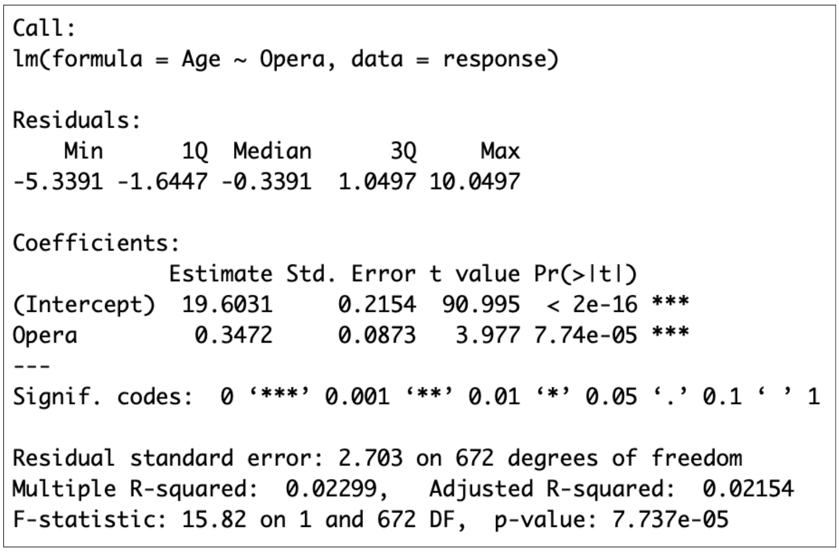


```
    House...block.of.flats
 block of flats:413       
 house/bungalow:261       
```

ANALYSES/GRAPHS

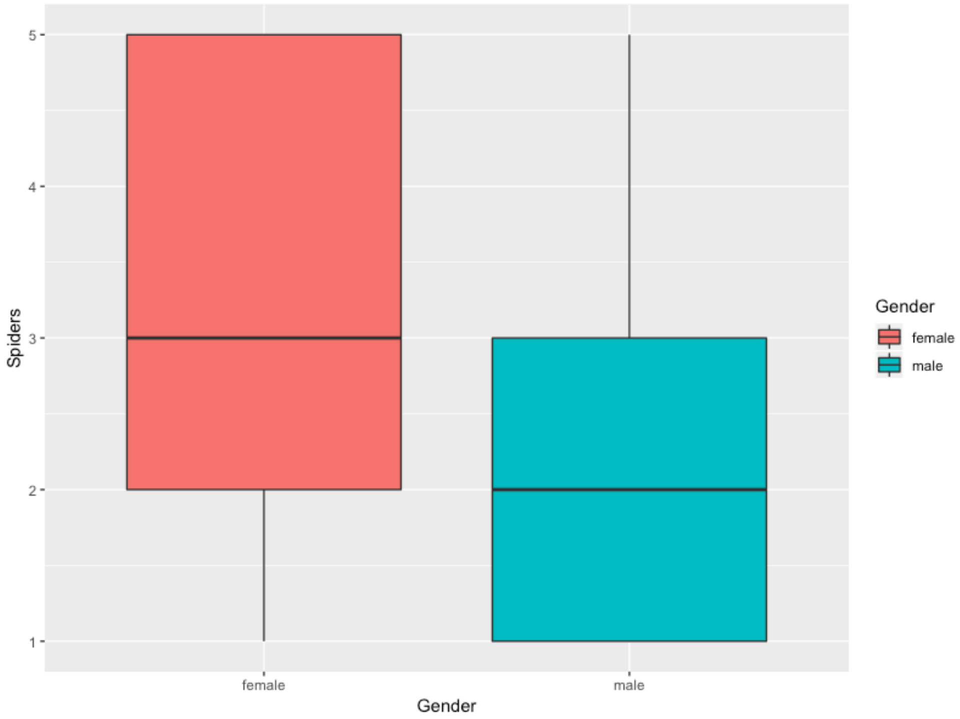
First, I ran a **regression model** in R to explore the relationship between the enjoyment of opera and age.There was a significant (p < 0.05) relationship between age and the enjoyment of opera. The relationship between age and the enjoyment of opera is positive. As age goes on, the level of how they enjoy opera goes up. So older people tend to enjoy opera more than younger people. This makes sense to me cause my friends usually prefer movies compares to opera, but my parents sometimes prefer opera more than movie.

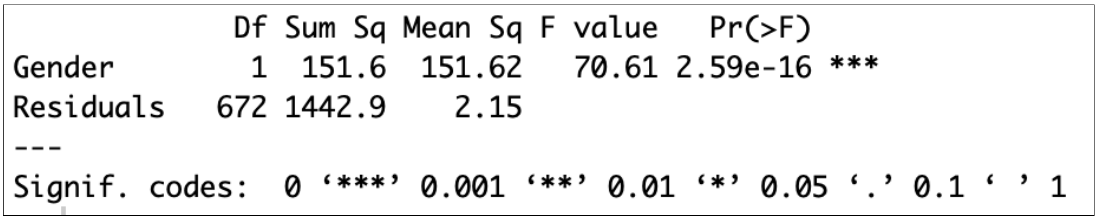




```
Call:
lm(formula = Age ~ Opera, data = response)

Residuals:
    Min      1Q  Median      3Q     Max 
-5.3391 -1.6447 -0.3391  1.0497 10.0497 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  19.6031     0.2154  90.995  < 2e-16 ***
Opera         0.3472     0.0873   3.977 7.74e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.703 on 672 degrees of freedom
Multiple R-squared:  0.02299,	Adjusted R-squared:  0.02154 
F-statistic: 15.82 on 1 and 672 DF,  p-value: 7.737e-05
```

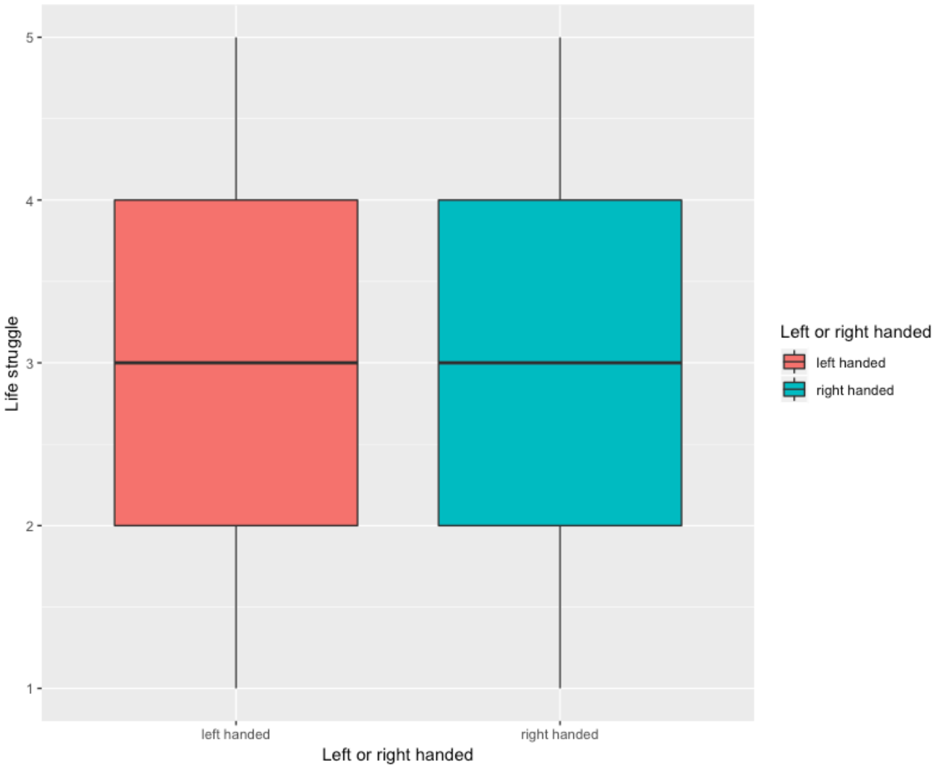
Secondly, I was interested in whether gender was related to the fears of spider. Using an **ANOVA**, I discovered that there was a statistically significant relationship between gender and the fears of spider(p<0.05). Female had a significantly higher fears of spider compared to male, which really surprised me.

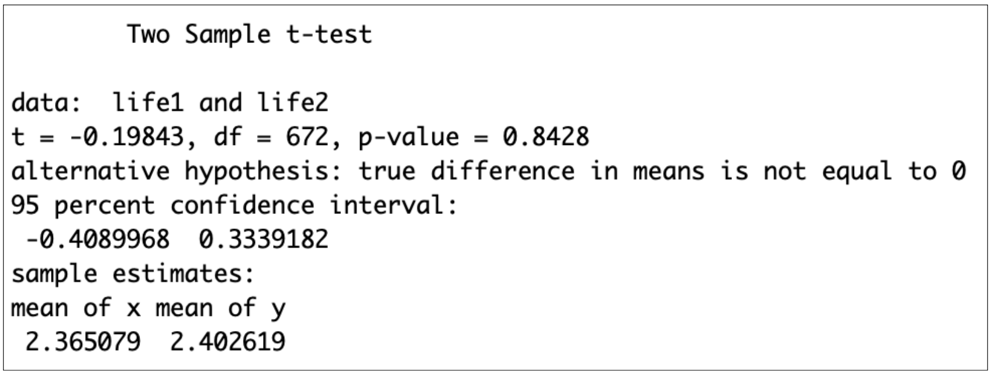




```
             Df Sum Sq Mean Sq F value   Pr(>F)    
Gender        1  151.6  151.62   70.61 2.59e-16 ***
Residuals   672 1442.9    2.15                     
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

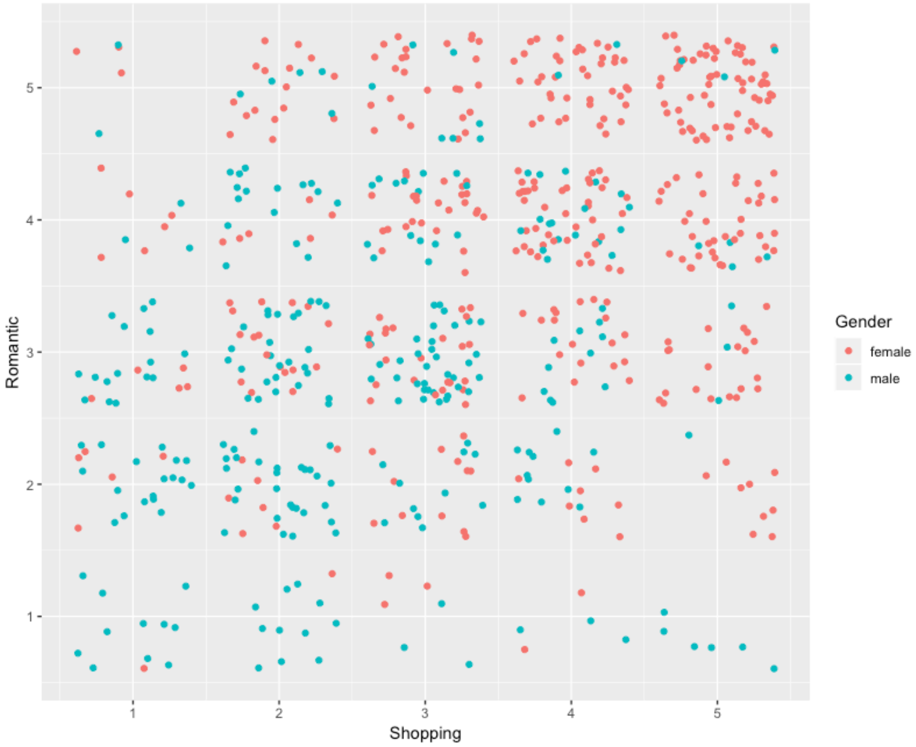
Next, I also wanted to know if there’s any life struggle differences between left-handed people and right-handed people. I used a **t-tset** to determine the life struggle level between left-handed people and right-handed people. The result shows that there was no significant difference of life struggles between two groups of people(p>0.05). In contrast, the distribution of life struggle level was really similar in tow groups.

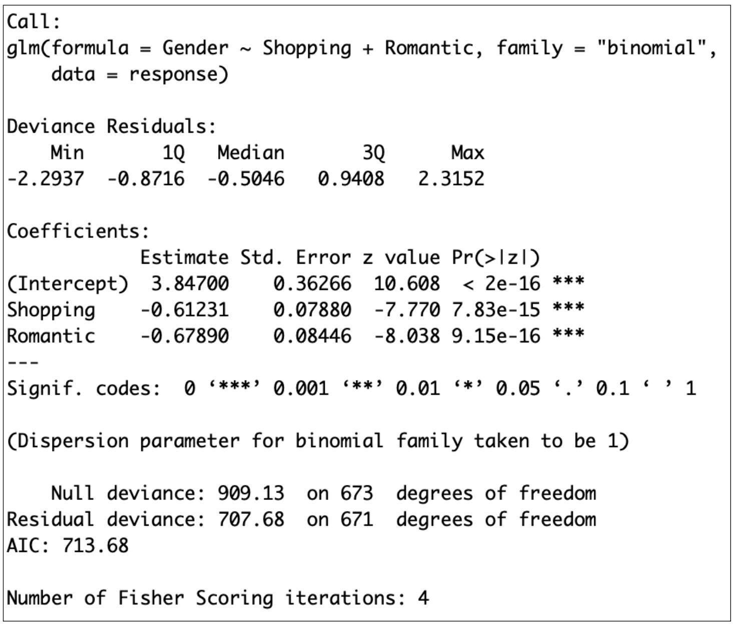




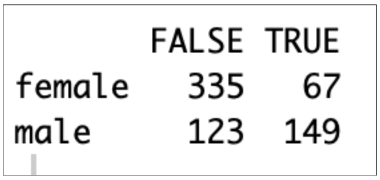
```
        Two Sample t-test

data:  life1 and life2
t = -0.19843, df = 672, p-value = 0.8428
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.4089968  0.3339182
sample estimates:
mean of x mean of y
 2.365079  2.402619
```

I also built a **logistic regression** model to predict gender using the interest in shopping and the preference in romantic movies. Overall, this model performed relatively well. Using the interest in shopping and romantic movies, the model was able to predict 484 out of 674 genders in the dataset correctly. The accuracy is about 70%, and it seems to have higher accuracy in female than male. To improve this model, I might want to add some other variables like the fears of spider cause we know there was significant relationship between gender and the fears of spider from the anova test above. Also, we can see from the plot that female tend to have more interests in shopping and romantic movies than male.



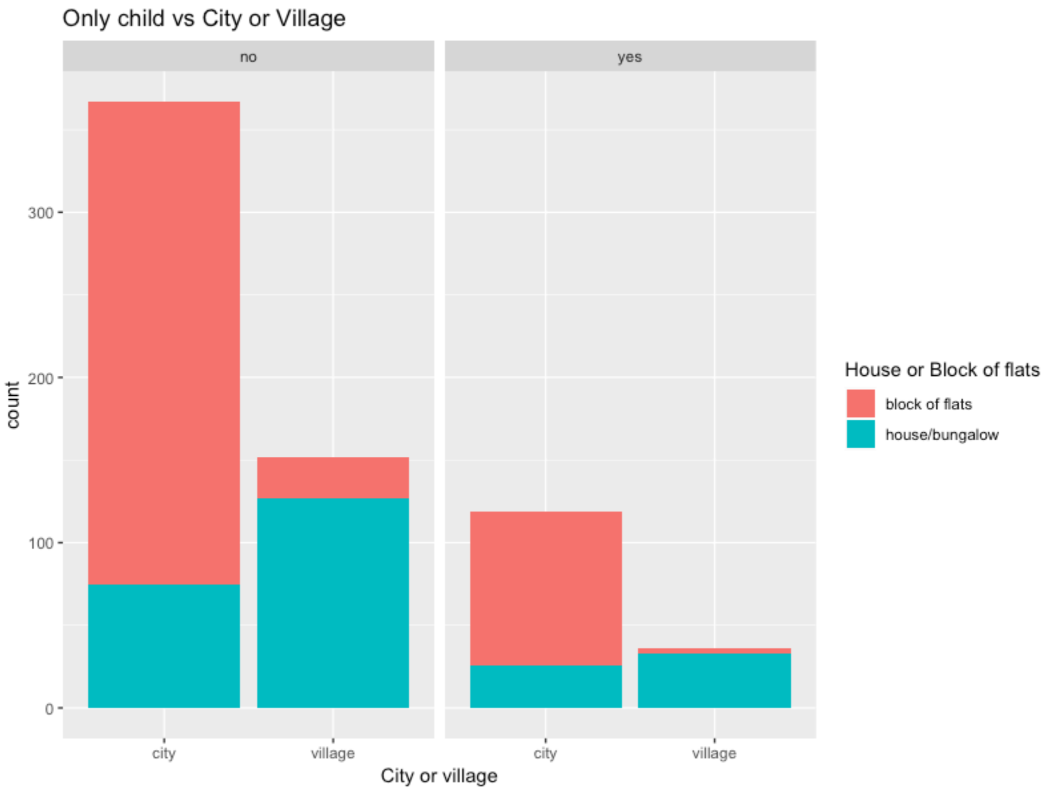


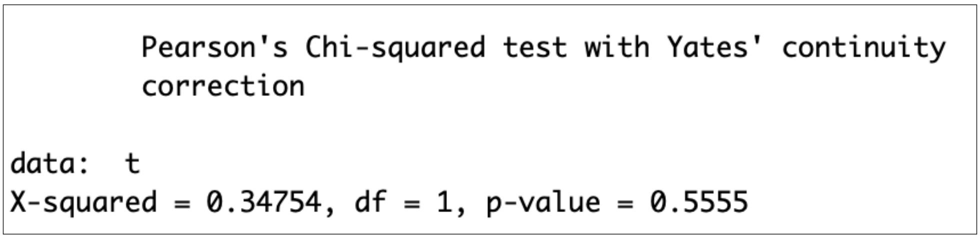
```
Call:
glm(formula = Gender ~ Shopping + Romantic, family = "binomial",
    data = response)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.2937  -0.8716  -0.5046   0.9408   2.3152

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  3.84700    0.36266  10.608  < 2e-16 ***
Shopping    -0.61231    0.07880  -7.770 7.83e-15 ***
Romantic    -0.67890    0.08446  -8.038 9.15e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 909.13  on 673  degrees of freedom
Residual deviance: 707.68  on 671  degrees of freedom
AIC: 713.68

Number of Fisher Scoring iterations: 4
```



```
       FALSE TRUE
female   335   67
male     123  149
```

Lastly, I used a **chi square test** to determine if being an only child was independent of the living place in childhood. The result shows that being an only child is not significantly related to spending most of childhood time in city or village(p>0.05). However, the bar chart tells that people spending most of childhood time in city tend to live most of childhood in a block of flats more than house or bungalow, and people spending most of childhood time in village tend to live most of childhood in house or bungalow more than a block of flats, which makes sense.





```
        Pearson's Chi-squared test with Yates' continuity
        correction

data:  t
X-squared = 0.34754, df = 1, p-value = 0.5555
```