ST511 FINAL PROJECT (SUMMER-2018)

Author: Hanif Khan

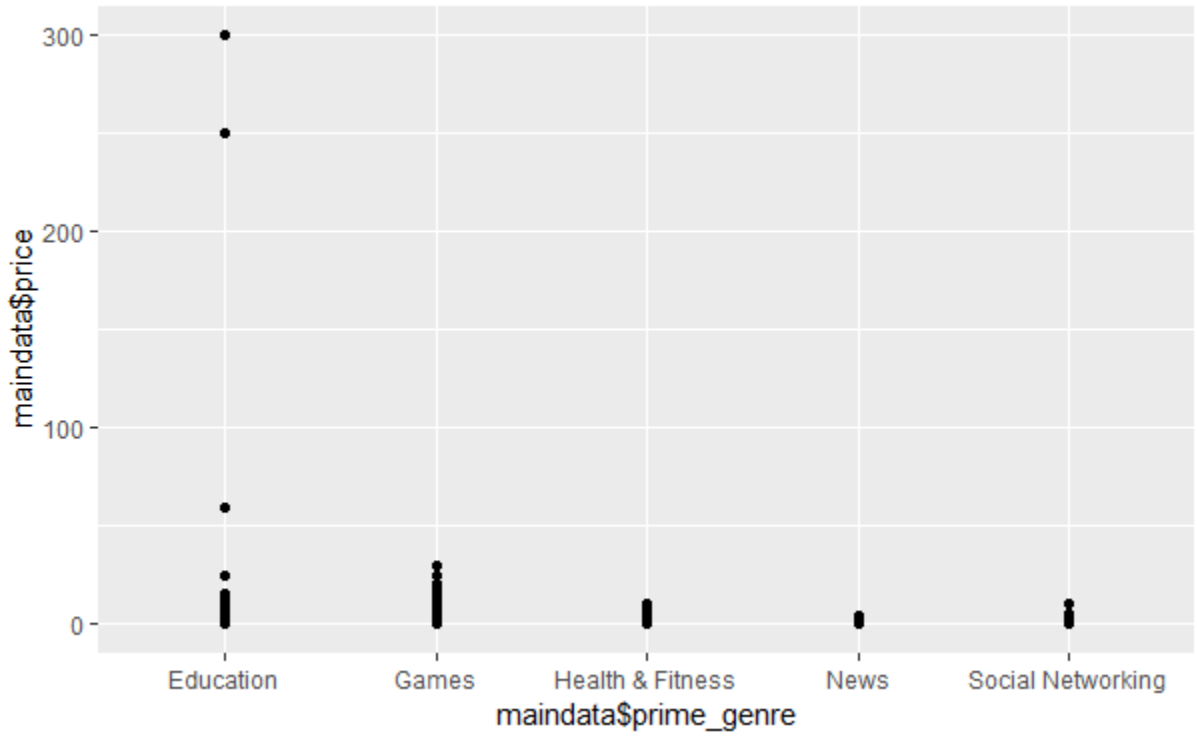
Abstract

Performed an Analysis of variance test on the mobile applications data to test whether the amount spend on the all the type of application is same or not.

Hanif Khan
[Email address]

Introduction:

I have performed the ANOVA test on the Apple Store data (source Kaggle). This data had name of the application, application version, ratings, amount, genre, content ratings etc. I choose this data because I would like to know the people's interest on the different kind of application and my focus was Games, Education, Social-Networking, and News. For me the interesting column was price to know the people's interest. Please check the following graph representing the price for each group of application.



From the graph above we can see that the amount spends for the education is more than the other four types of application whereas the amount spend on the news application is lowest in all the types.

Method:

I am using ANOVA to compare the above group and find the answer for the question whether the mean amount spend on the Games, Educations, News, Health & Fitness, and Social Networking applications are same or not? Since I need to compare more than one group ,hence I believe it will be the good option to go for the ANOVA test.

Results:

Null hypothesis for the ANOVA test is $\mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu_5$ where μ_1 is mean amount spend for the games, μ_2 mean spend of the education and so on. Whereas the

alternative hypothesis is opposite of it as $\mu_1 \neq \mu_2 \neq \mu_3 \neq \mu_4 \neq \mu_5$. From calculation shown in the R-code using function as follows.

```r
anova(lm(maindata$price ~ maindata$prime_genre, data=maindata))
summary(aov(maindata$price ~ maindata$prime_genre, data=maindata))
kruskal.test(maindata$price ~ maindata$prime_genre, data=maindata)
```

Result of the above code as follows.

```
Analysis of Variance Table

Response: maindata$price
                     Df Sum Sq Mean Sq F value    Pr(>F)
maindata$prime_genre    4   3140  784.92  20.248 < 2.2e-16 ***
Residuals            4732 183437   38.77
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
                     Df Sum Sq Mean Sq F value Pr(>F)
maindata$prime_genre    4   3140   784.9   20.25 <2e-16 ***
Residuals            4732 183437    38.8
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

        Kruskal-Wallis rank sum test

data:  maindata$price by maindata$prime_genre
Kruskal-Wallis chi-squared = 257.96, df = 4, p-value < 2.2e-16
```

Please follow the ANOVA test result only. Since above figure represents the results for the ANOVA and Kruskal test.

Conclusion:

From the result table we can see that the p value is very small and hence we will reject the null hypothesis and say that the mean amount spend on one of all the type of the application is at-least different than amount spend on other type of application.