**Google Play Store Report and Analysis: 104-DB-Story#3**

**Introduction:**

This report will include data and information regarding the Google Play Store and will be used to answer a user story. The data utilized for answering the user story consists of 10,841 rows and 13 columns all related to Google Play Store statistics and metrics. These ranged from the category of the app, the overall rating, installs, current version, price, and other similar factors, meaning there are 10,841 apps and relevant information contained within the data.

The question I was tasked with is, “User would like to know if price affects the amount of downloads for the ENTERTAINMENT category.” A dataframe containing all the cleaned data from the Price and Installs columns in the ENTERTAINMENT category was used to present a statistically valid conclusion. My hypothesis was that Price would affect the number of Installs. For example, I think if the Price is free, people would be more likely to Install the app. I used the Chi-Square test to test my hypothesis, assuming that my variables were categorical. I also used an additional test in the case that my variables could be considered continuous (floats).

**Data:**

The first task in evaluating the data involved taking all rows from the category ENTERTAINMENT and cleaning the data in order to set a dataframe with the columns Price and Installs. In order to clean the Installs column, I first had to convert its numeric string values (i.e, 1,000,000+) from the object data type to the float data type. This required the stripping of the plus sign from the end of each number in the cell and the removal of the commas using the replace() method for both. I then converted the object data type to the float data type using the astype() method. The Installs column was then checked for unique data types using the unique() method. This indicated all unique numbers except for the value Free, so I used the loc[] method to search the dataframe for this value. Since this was only one record consisting of a relatively low install value (1000), I used the drop() method to remove the record from the dataframe. The final task for dataframe construction was to assign a variable to all the records in the category ENTERTAINMENT using the loc[] method. The data for this story was now considered cleaned and ready for statistical analysis.

To perform my statistical analysis, I first developed a null hypothesis. My null hypothesis was that the Price of a download does not affect the number of Installs. The p-value was set at 0.05.

I chose two different statistical tests to test the hypotheses. The Chi-Square (χ²) Test is a Goodness-of-fit test that determines whether or not a data set is dependent upon one or more independent variables. It also gives a value to determine how dependent the data set is upon these variables.

In order to run a Chi-Square test in Python, I wanted to test if there was a statistically significant difference in Installs for apps between Prices of $0.00, $2.99 and $4.99 in my df1 dataframe. I converted the data into a contingency table with frequencies using the crosstab command from pandas:

contingency= pd.crosstab(df1['Price'], df1['Installs'])

After I built the contingency table, I passed it to the chi2\_contingency function from the scipy package:

c, p, dof, expected = chi2\_contingency(contingency)

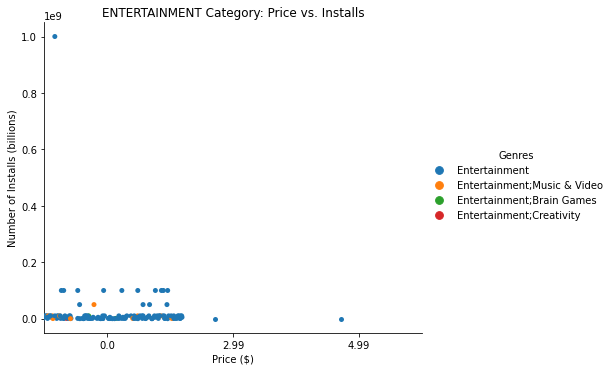
The above function returned the following values:

* **c=chi2:** The Chi-Square test statistic
* **p:** The p-value of the test
* **dof:** Degrees of freedom
* **expected:** The expected frequencies, based on the marginal sums of the table

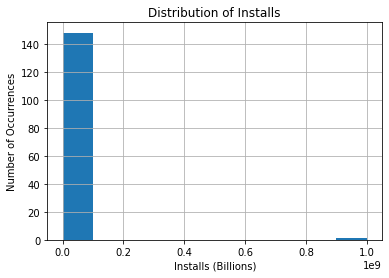
In addition to the Chi-Square test, which is for categorical variables, I also performed research that suggested I may want to check the results using a test intended for continuous variables since my data are floats. For a test of three or more unmatched groups, a normally distributed group could undergo the One-way ANOVA test using the function scipy.stats.f\_oneway(). However, I first checked to see if my data was normally distributed. I then ran statistical tests of normality using both the Jarque-Bara test (jb\_stats = jarque\_bera(df1['Installs'])), and Normality test (norm\_stats = normaltest(df1['Installs'])).

**Results:**

The catplot method was used for data visualization of all the Installs vs. Price data for the ENTERTAINMENT category. There can possibly be considered one outlier in the 0.0 category. There were many instances of Installs when Price=0, but not for the other Price levels.



I determined that the distribution of my Installs variable was not normal because the visualization of the Distribution of Installs plot below did not appear to be a Gaussian (normal) distribution.



The results of the Jarque-Bara and Normality tests were essentially zero, meaning the variable Installs is not normally distributed.

The contingency= pd.crosstab(df1['Price'], df1['Installs']) command resulted in the following frequency table.

Installs 1.000000e+04 5.000000e+04 1.000000e+05 5.000000e+05 1.000000e+06 5.000000e+06 1.000000e+07 5.000000e+07 1.000000e+08 1.000000e+09

Price

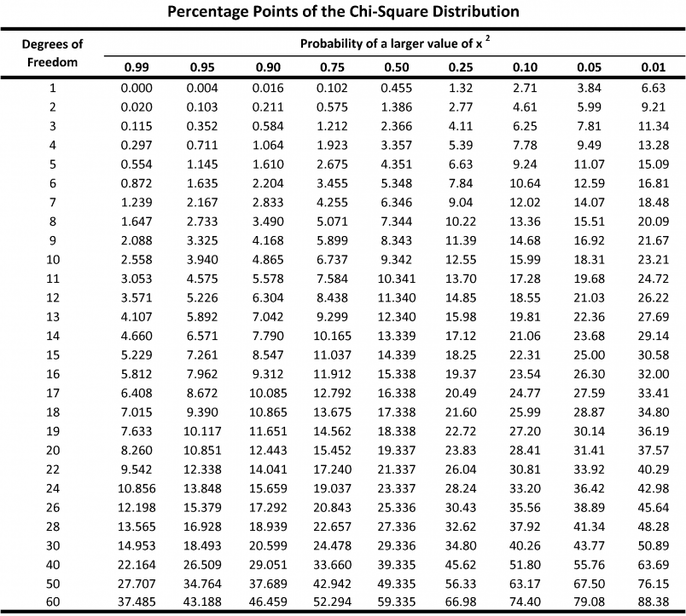
0.00 1 1 9 4 51 17 48 5 10 1

2.99 0 0 1 0 0 0 0 0 0 0

4.99 0 0 1 0 0 0 0 0 0 0

The Chi-Square test returned a p-value of 0.11, which means that I could not reject the null hypothesis at 95% level of confidence.

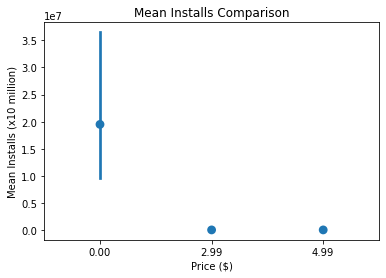
Also, my Chi-Square test value was 25.43. At 18 degrees of freedom, the critical value was 28.87, as presented in the table below. Since my Chi-Square value was less than the critical value, once again, I could not reject the null that there is no difference in Installs when Price was considered.



The results of the Kruskal-Wallace test are shown below. Once again, my p value was greater than 0.05, indicating that I could not reject the null hypothesis. The two statistical tests that I conducted were in agreement.

KruskalResult(statistic=5.286371761708344, pvalue=0.0711342828223119)

To visualize the results, I plotted the mean Installs for each Price level. Here I see there is a wide range of Installs at the 0.00 Price, while the 2.99 and 4.99 Price levels are more closely grouped.



Conclusions:

The test for normality showed that the Installs variable was not normally distributed, so the Kruskal-Wallace test was chosen in addition to the Chi-Square test. I ran both of these tests to compare tests for categorical and continuous data. Although my data were floats, they could be considered discrete categories. Both tests were in agreement, and I concluded that this set of data showed no difference between Installs at three different price levels. A catplot of the frequencies of occurrence provided a clue as to why the Chi-Square test was not ideal in this instance since I learned from my reading that it is recommended that at least five frequencies are required; two of my data points only had one. From a plot of the mean Installs, I could see that the mean Installs had a wide range, which could possibly affect the ultimate significance of my tests.

Further research could include obtaining more data from more years because economic conditions during the years data was collected could have impacted the results. I might also suggest removing a potential outlier in the 0.00 Price category.