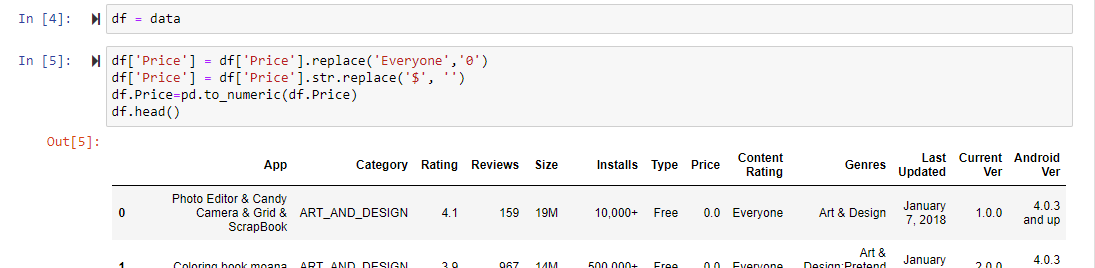
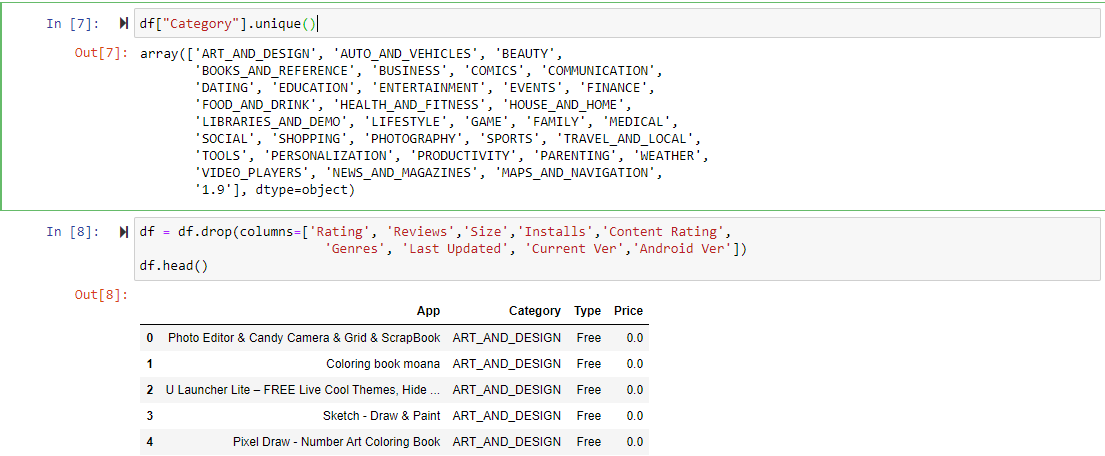
During my live project I worked with the Google Play Store data and performed my analysis with Python utilizing Jupyter Notebooks. I completed four user stories during my ten-day live project course. I was able to answer questions such as, ‘How many apps are within each of the categories, Entertainment, Social and Productivity?” I was tasked to further brake these down by paid vs. free apps.

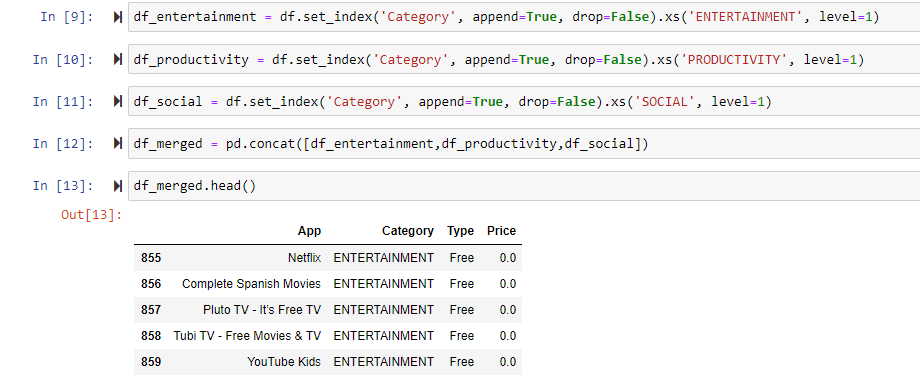
I knew that I would need to convert the ‘Price’ column to numeric in order to perform any analysis with sorting I stripped off the ‘$’ sign and converted the cells to numeric.



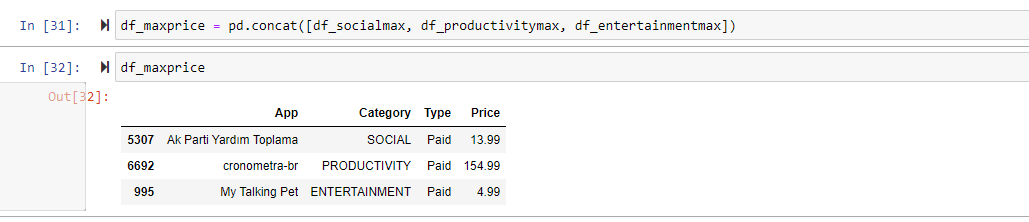
Next, I dropped all of the columns that were irrelevant to my user story. I used unique() to identify all of the columns and then dropped all but App, Category, Type and Price.

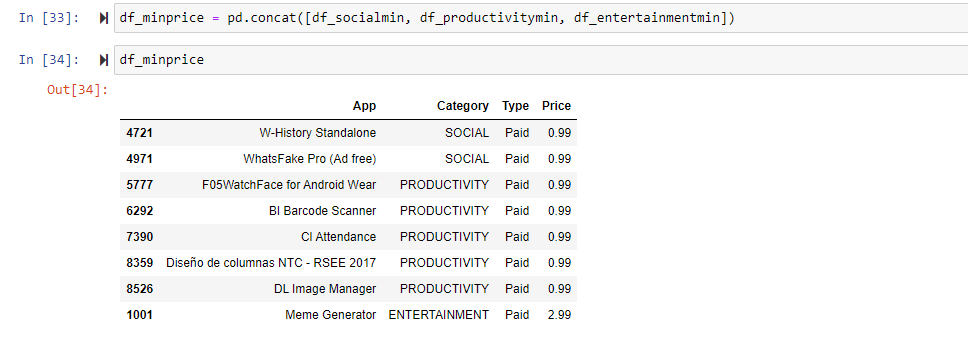


Next, I set my indices on my three categories that the user story focuses on, ‘ENTERTAINMENT’, ‘PRODUCTIVITY’ and ‘SOCIAL’. This was to isolate only these three categories going forward since the user story only wants information within the three. I then used the concat() method to get all three categories and their data into one single dataframe. From there I knew I could answer all of my user story questions.

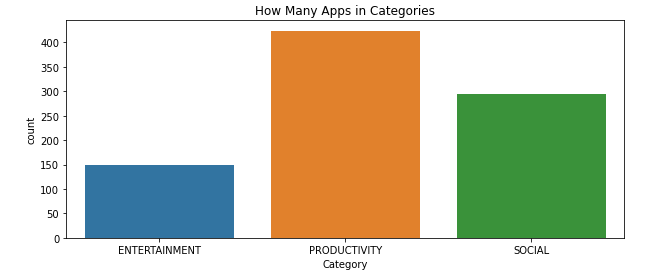


In order to create a visual for the data that encapsulated all three categories I knew I needed to create a couple of dataframes to isolate the necessary data to perform analysis. I decided to create dataframes for each category to show the minimum and maximum priced apps within each. In order to do this, I had to isolate the types of apps by paid or not paid. I used a filter on the dataframe column Price to only show apps that were over $0. Once I had these dataframes setup for each category, I used concat() to string them together and I was able to perform further analysis on the dataset.

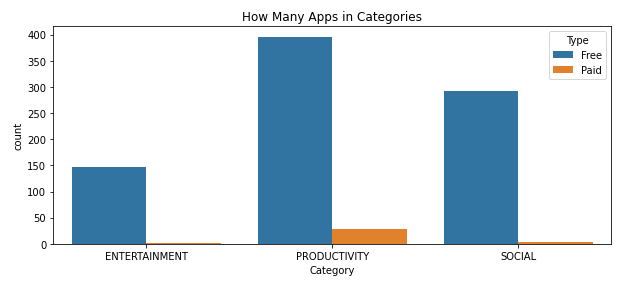




Using a simple bar chart, I was able to that there are 400 apps in the productivity category while there are about 300 in the social category and 150 in the entertainment category.

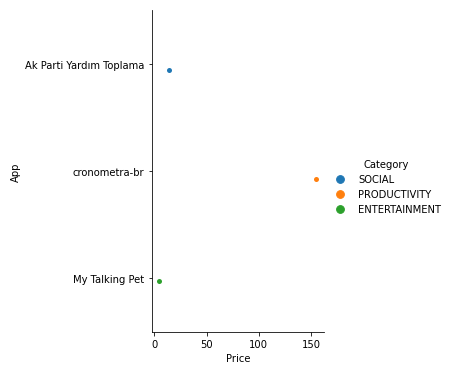


Using a simple bar chart, I was able to answer this question with the below visual. We can see the number of apps that are free far outweigh the number of apps that are paid in every category. While productivity appears to have the highest number of paid apps, it also has the highest number of overall apps.

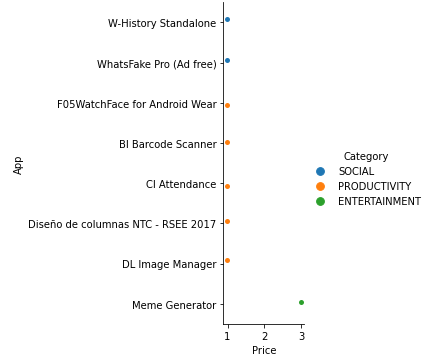


Below we can see the highest priced app in each category and the price. The app in productivity, ‘cronometra-br’ is $154.99, far higher than any other app in the dataset.

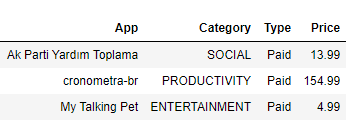
## 



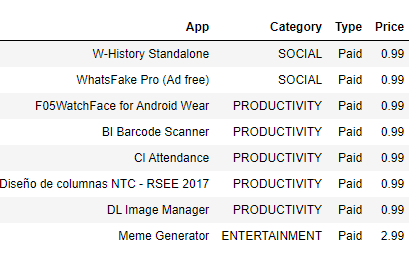
## Below we can see the lowest priced apps within each category. There are 5 apps in the productivity category and two in the social category that are $0.99.



Below we can see the highest priced apps in the categories Social, Productivity and Entertainment listed out along with their respective prices.



Below we can see the lowest priced apps in the categories Social, Productivity and Entertainment listed out along with their respective prices.



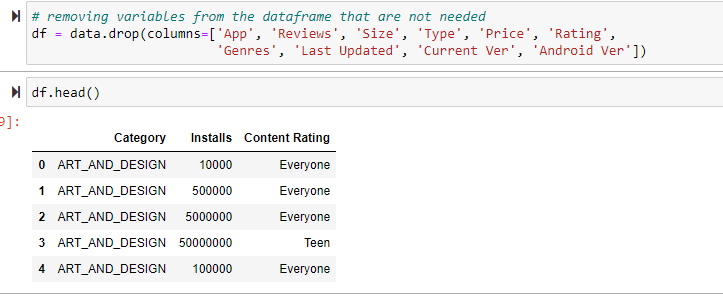
Next I was tasked with determining which content rating has the highest number of installs and how many categories it belongs to

I needed to convert the ‘Installs’ column to numeric in order to perform any analysis with sorting, I stripped off the ‘+’ signs and commas and converted the cells to numeric.

In order to perform this anlysis I needed to clean my data. I needed to do a number of things, first was to convert my Insalls column to numeric.

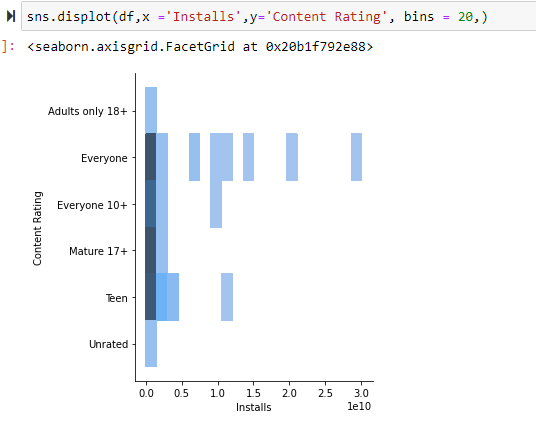
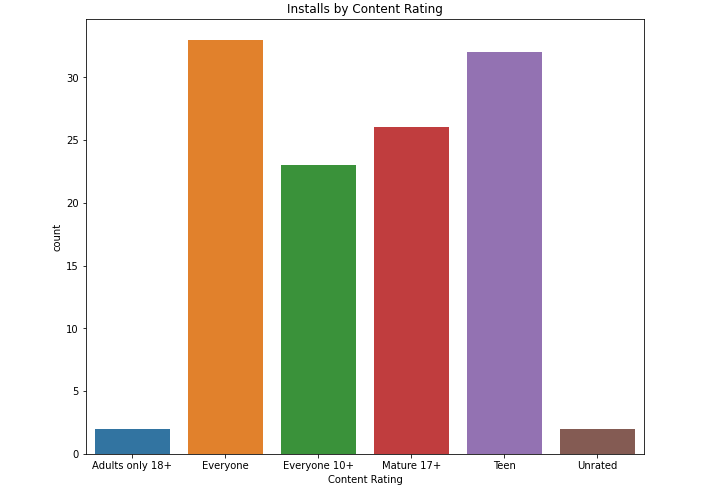


I then dropped all of the irrelevant columns from my data and grouped the data using groupby() to get the sum of installs for each ‘Content Rating’ and ‘Category’. I also needed to reset my index afterwards in order to keep my columns useable for the analysis.

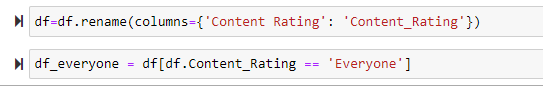




Using sns displot I was able to see that the content rating with the most installs is the Everyone category. I also confirmed my finding with a bar chart.

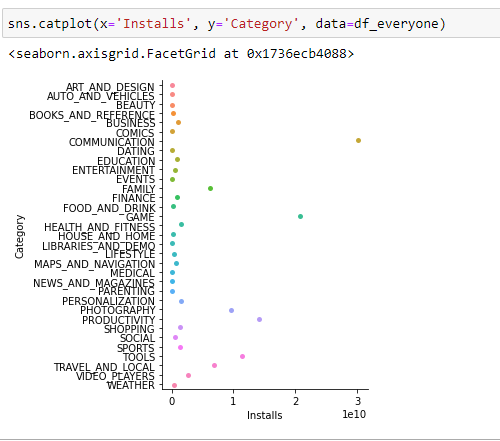
After determining that the ‘Everyone’ content rating is has the highest number of installs I created a dataframe which only contained the ‘Everyone’ content rating and then renamed the ‘Content Rating’ column to Content\_Rating in order to make it usable in analysis.



Using the count() function to see the number of categories that ‘Everyone’ belongs to.

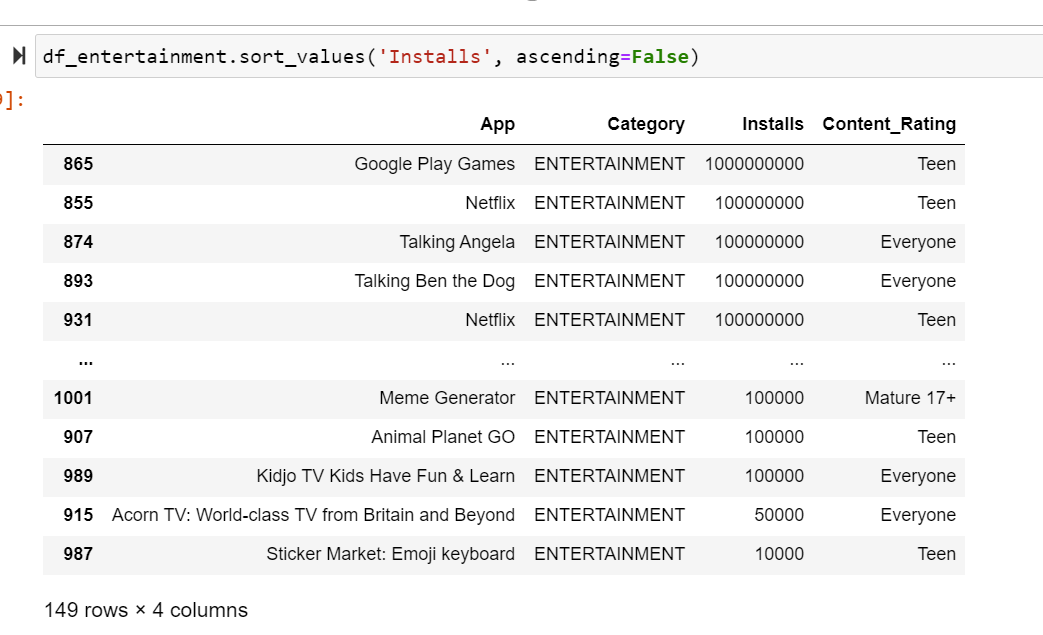


To represent this visually I used sns.catplot to show the categories and corresponding number of installs within the ‘Everyone’ content rating.



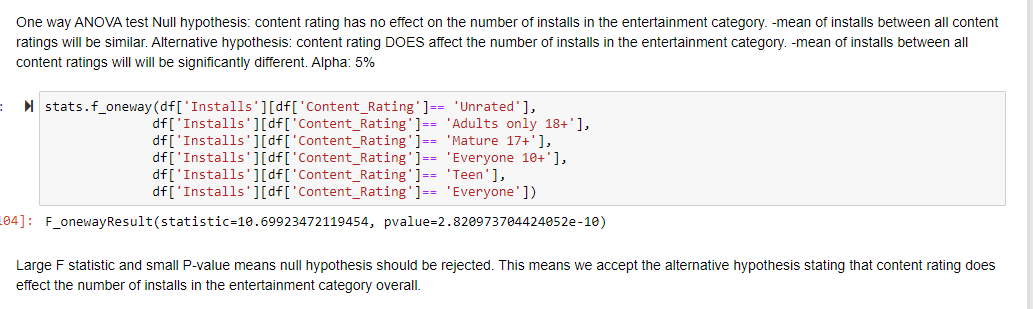
I was then tasked with determining whether the content rating affects the number of installs in the Entertainment category by performing a statistical analysis of my choice. I started working with my data in the same way that I had for my previous tasks, cleaning my data.

I created a new dataframe with only data for the Entertainment category since that is all I need for this user story. I also dropped all of the columns that were irrelevant to my user story. I dropped all columns except App, Category, Install and Content Rating.



The dataframe I created has all of the apps within the Entertainment category and lists their number of installs as well as their content rating.

Based on the data I waw working with, I chose to do a one way ANOVA test using the statsmodel package.



Since my p-value returns as <5% I reject my null hypothesis and determine that in fact, the content rating does have an effect on the number of installs.

## I wanted to a bit further to see if there was an effect on the number of installs between the content ratings. Since I have already determined that overall, the content rating does have an effect on the number of installs, I wanted to see if this this holds true among all content ratings within the entertainment category or does it vary depending on the content rating given to the app.

## To do this, I performed a multicomparison of means using tukeyhsd(). Here are the results:

## 

## We can see here that when we look between content ratings, there are 4 where we should reject the null hypothesis, and 11 comparisons where we should accept the null hypothesis. Visualized another way we can see the same data:

## 

## It is difficult to see because the range of installs is so large, but confidence levels do not overlap for Everyone vs. Everyone 10+, Everyone vs. Teen, Everyone 10+ vs. Mature 17+, and Mature 17+ vs. Teen.

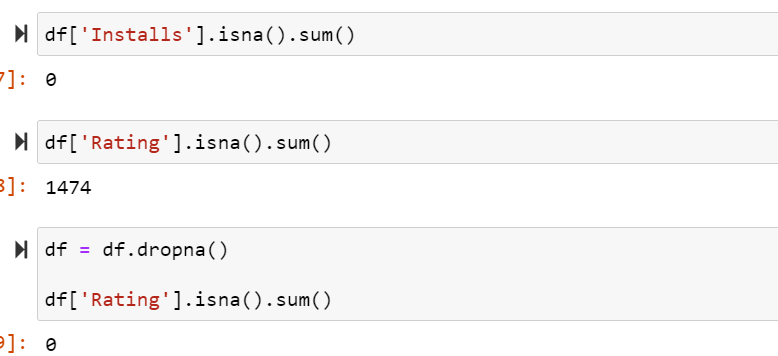
## In conclusion, overall, we reject the null hypothesis that there is no significant correlation between the content rating and the number of installs for apps within the Entertainment category on the Google Play Store. This means there is in fact a correlation between the content rating and the number of installs for apps within the Entertainment category on the Google Play Store.

## If we want to look at this more granularly, then we can say that for the four different comparisons (Everyone vs. Everyone 10+, Everyone vs. Teen, Everyone 10+ vs. Mature 17+, and Mature 17+ vs. Teen) we reject the null hypothesis as these groups appear to have a correlation between their content rating and their respective number of installs. For the other 11 content rating comparison groups, we would in fact accept the null hypothesis stating that we see no correlation between the content rating given to the app and the number of installs for that app within the Entertainment category.

This report will utilize data from the Google Play Store including app, rating, categories, and number of installs. This report is intended to answer the following user story:

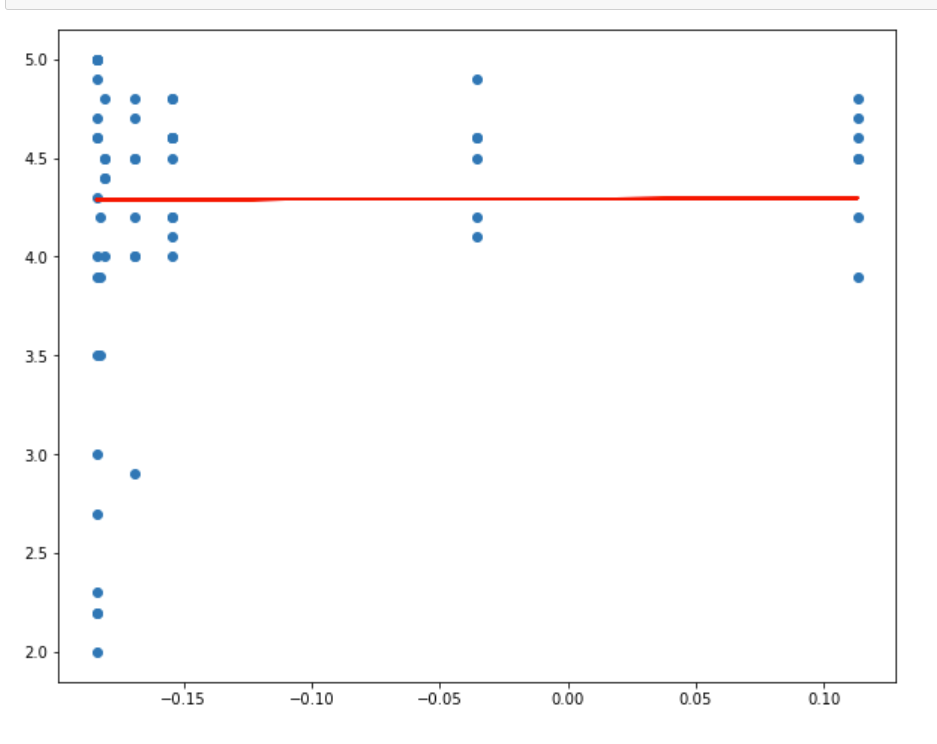
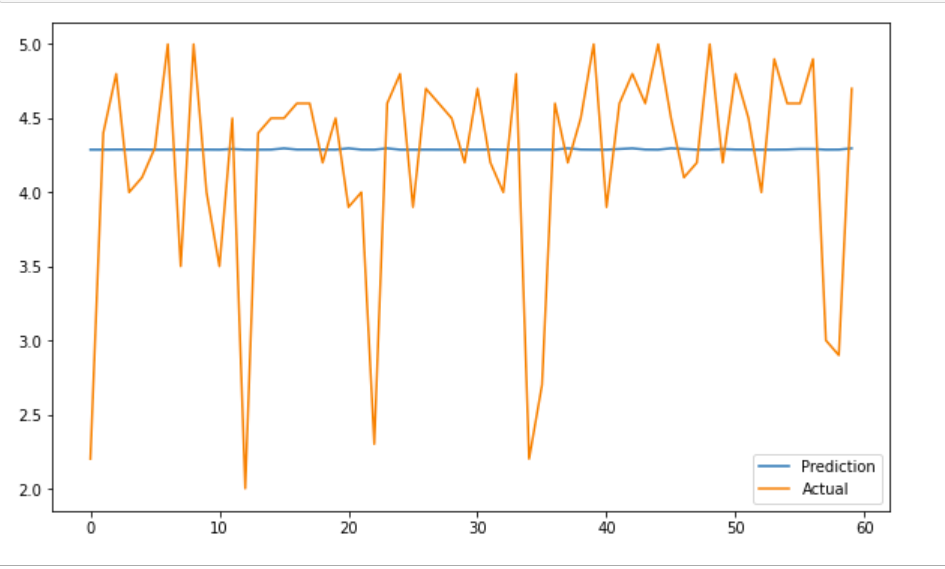
My last task with this dataset was to determine if the number if installs affects the average rating for the HEALTH\_AND\_FITNESS category by performing a statistical test of my choice.

I cleaned my data in the same manner that I previously used for my earlier tasks. I checked for any missing values in the Installs and the Rating columns. There were no missing values in the Installs column, but there were 1,474 missing values in the Rating column. I decided it was best to drop the rows that contained no ratings because they would not be useful in this analysis. I decided against inputting the mean as a replacement value because I didn’t want to skew my data.

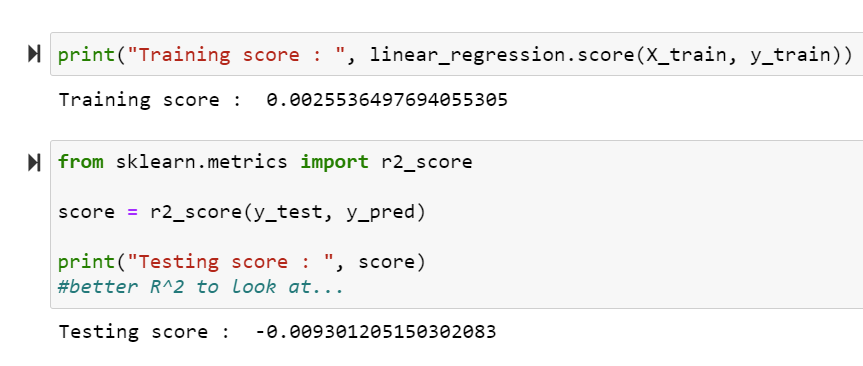


Next, I created a new dataframe that dropped all categories except for HEALTH\_AND\_FITNESS.

I decided to perform a simple linear regression which showed there is clearly no correlation between the number of installs and the rating for the app. I performed the linear regression with machine learning as well as analytically with a single predictor. First, I will show you linear regression with machine learning.

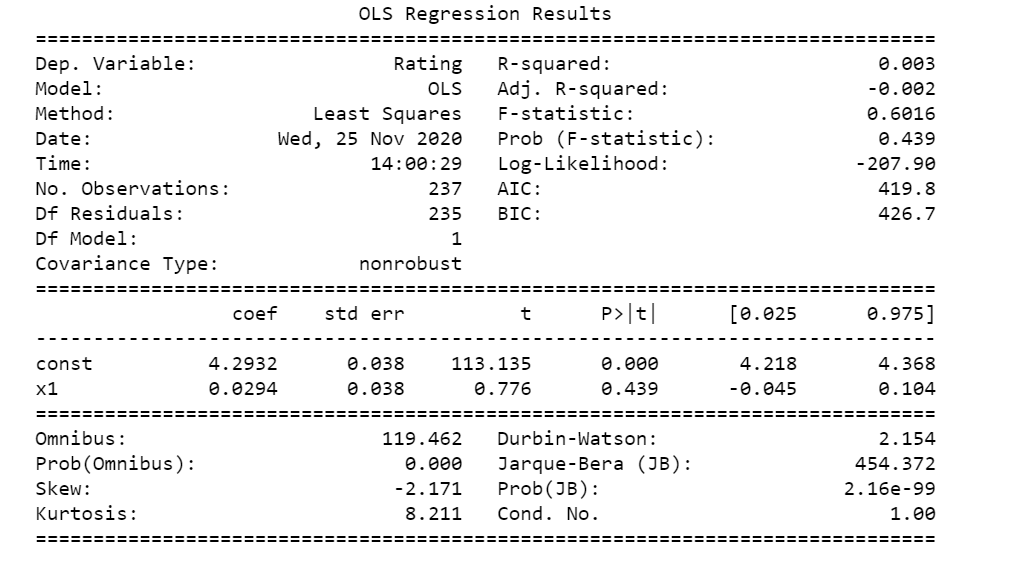
 

We can see in both of these graphical representations that there is no relationship between Installs and Rating. I calculated the R2 for both training and test data; results shown below:



Both scores are very, small so we accept the null hypothesis that Installs has no effect on the Rating of the app.

I then calculated the R2 value using a single predictor to perform linear regression analytically.



Our P-value is 43.9% which affirms our acceptance of the null hypothesis. The number of Installs does not affect the Rating for an app.