**Google Play Store: Does correlation content rating affects how many installs there are per category**

**Report and Analysis**

**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 questions we are tasked with are: The Google Play Store user wants to know if there is a correlation, if content rating affects how many installs there are per category. Also, to look over each content rating and number of installs. And then determine which content rating has the highest number of installs and how many categories it belongs to.

**Body:**

**Data:**

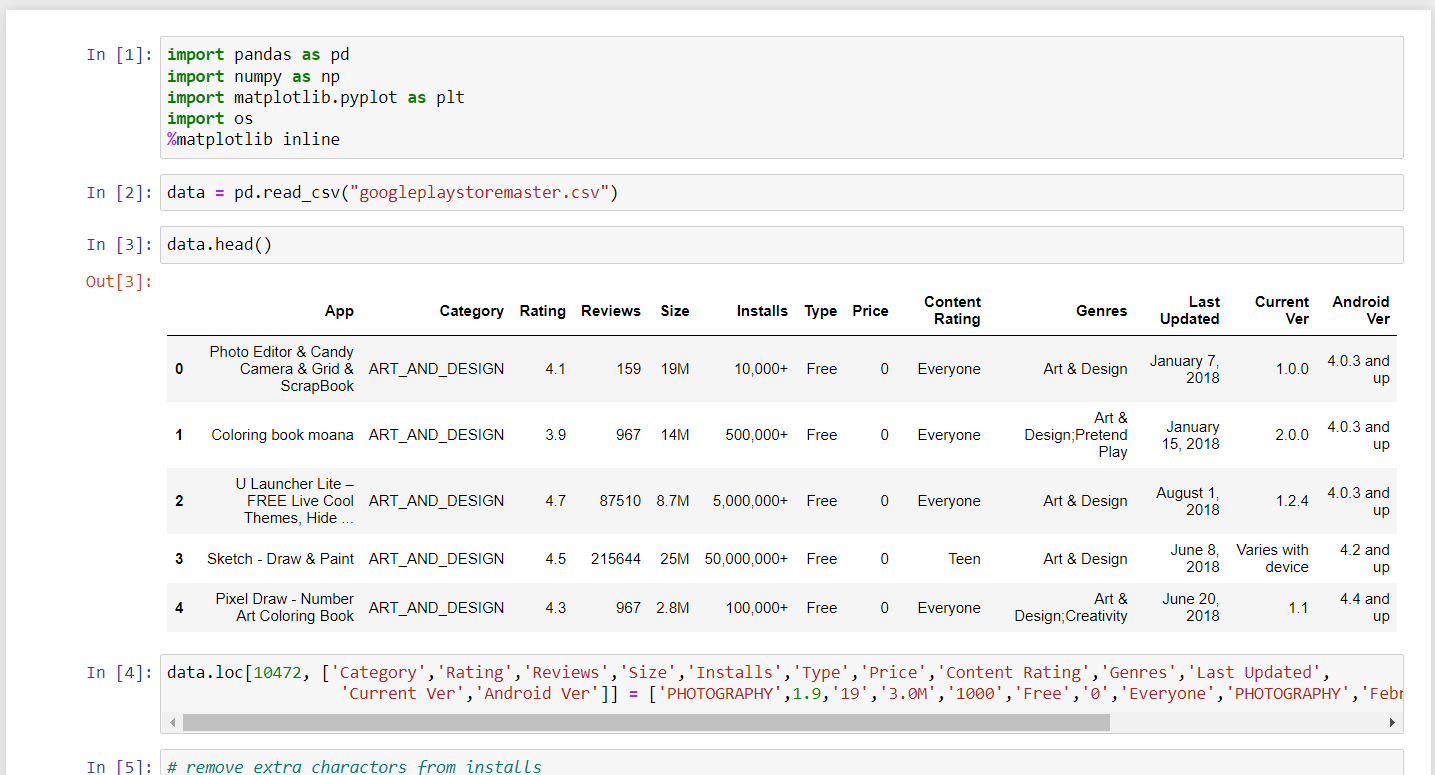
To start off this analysis I initially gathered all the data into a Data Frame for inspection of rows which will be relevant to the user story which will be tested. The next step taken was to start cleaning the data and preparing it for later analysis. One of the first things I noticed was an issue in the Installs column. This is a numerical column and would be needed for mathematical operations, however, the data was categorical and had symbols (i.e., 1,000,000+). After this column was changed to an integer data type as well as removing symbols, it was now available to be used for mathematical operations later.

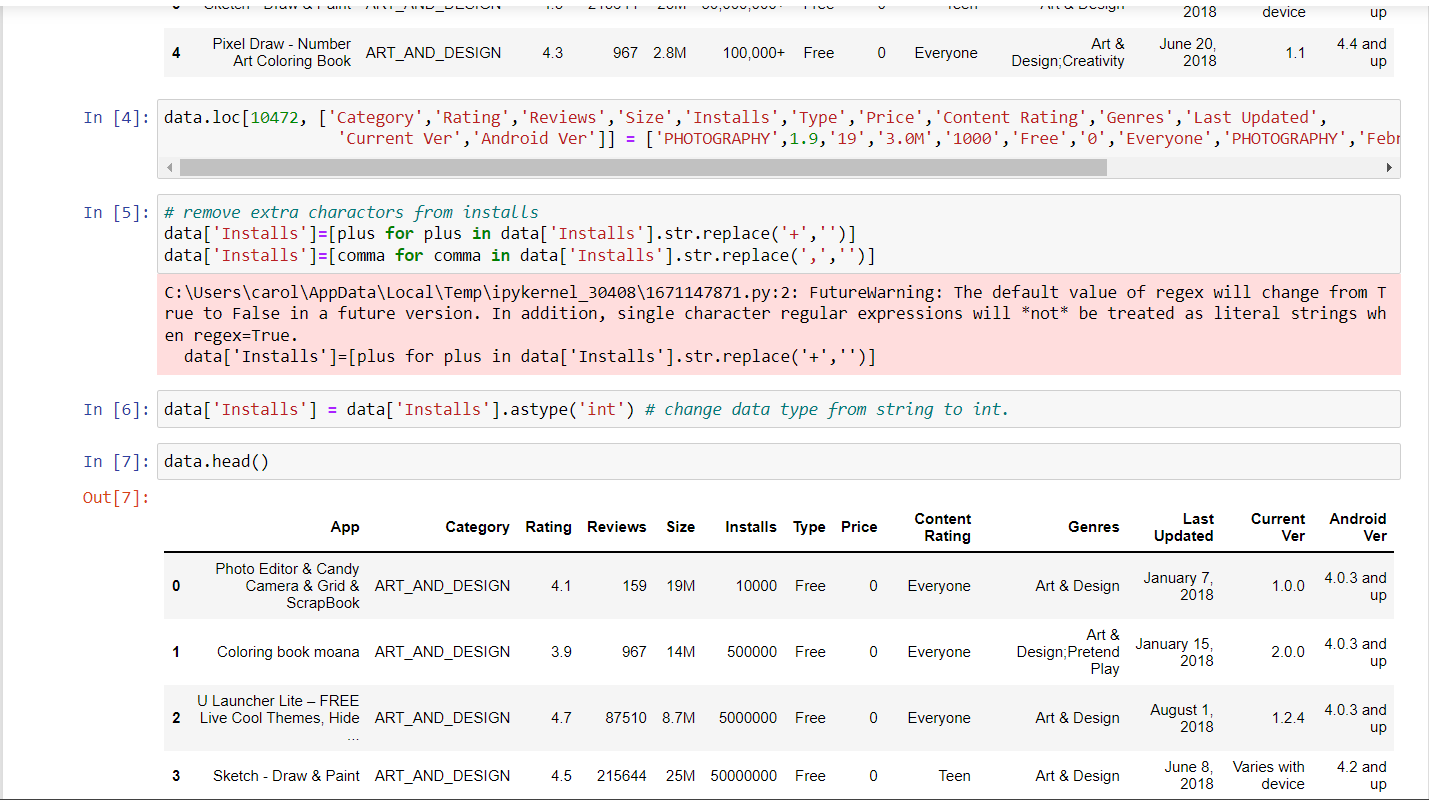
**Method:**

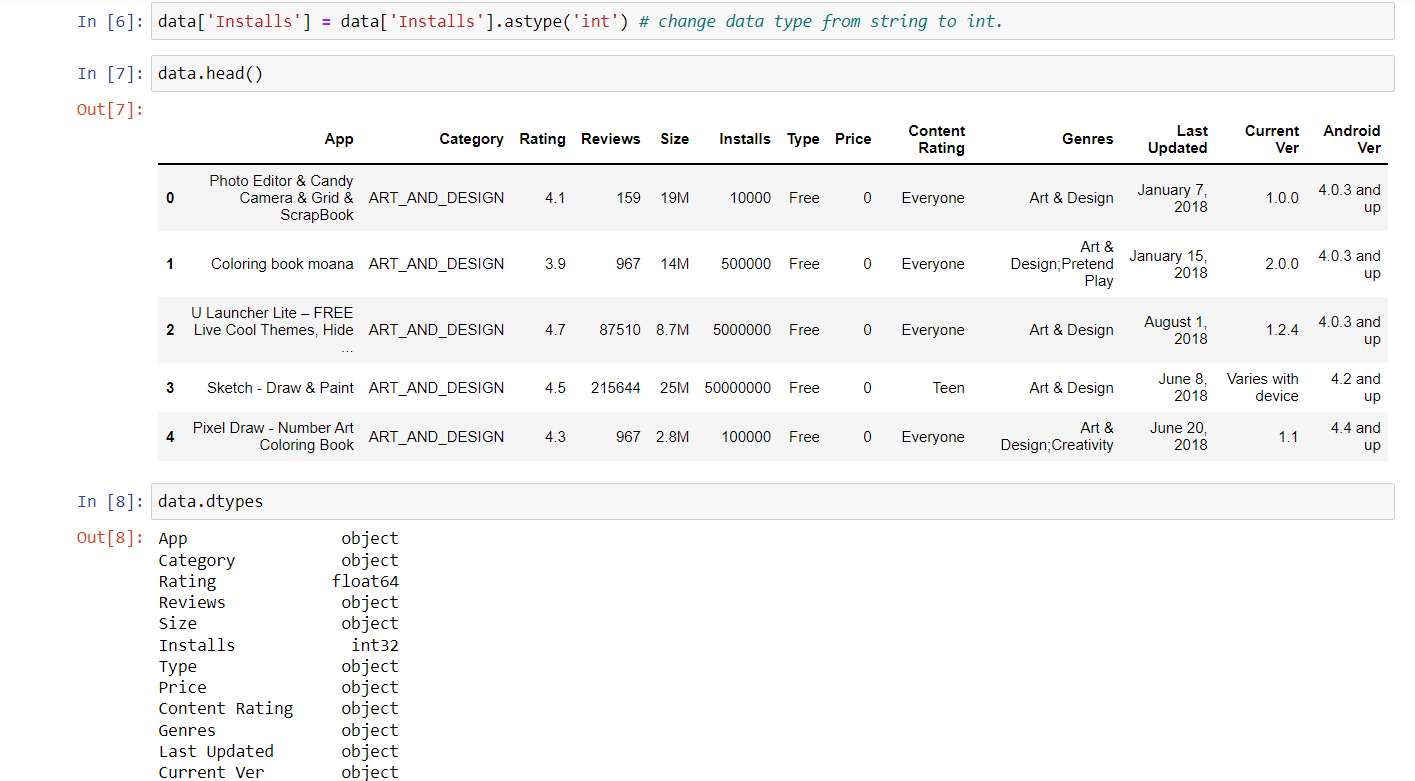
To organize the data a little more I used scale map and assigned a scale number to every category of the content rating, because the more explicit the content is, the higher the number drops unrated.

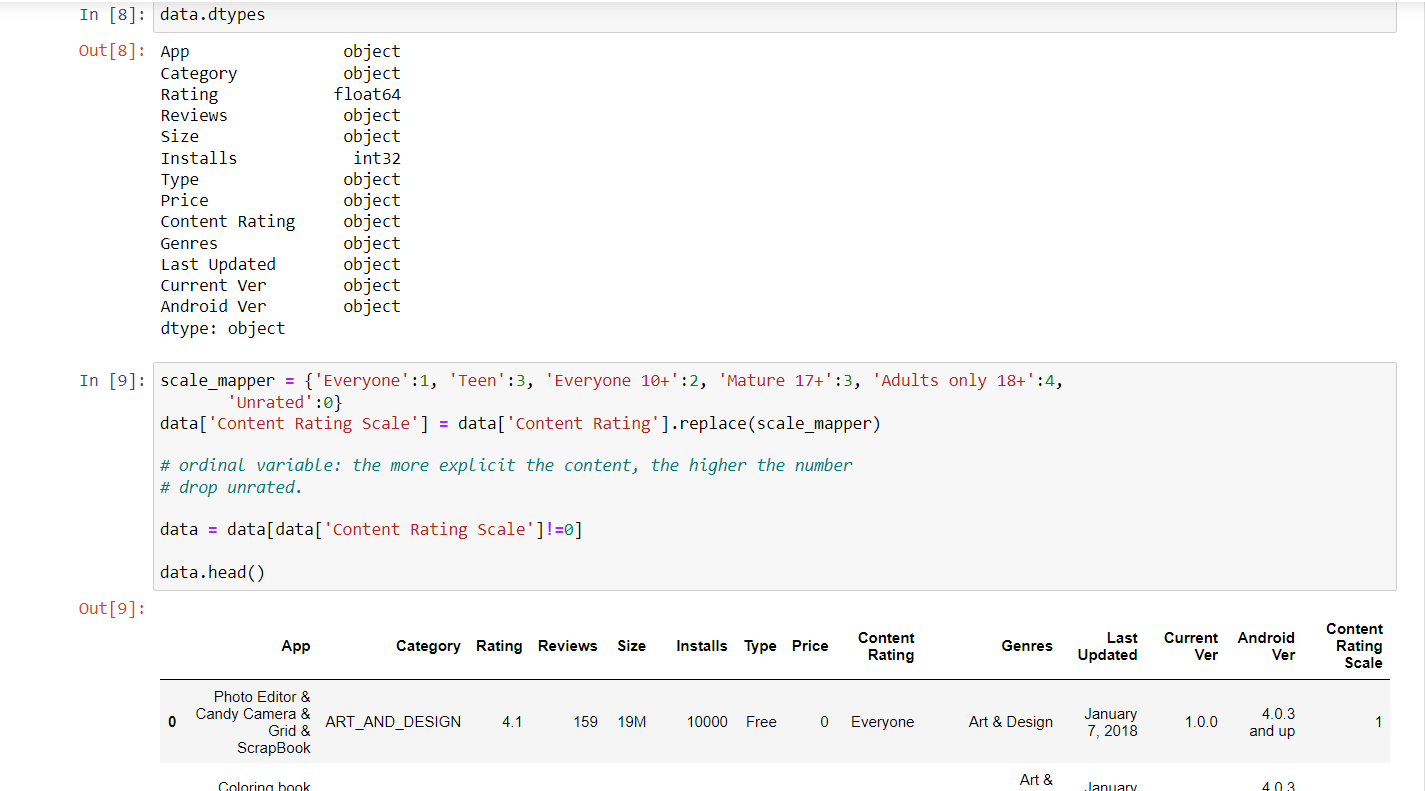
After this I did a histogram plot to look at the distribution of installs and we can see that it is very skewed. Because of this I did a log transform, which can be used to make highly skewed distributions less skewed. And then, I got the correlation data of Installs, log installs, and the content rating scale to see if there is any correlation between them.

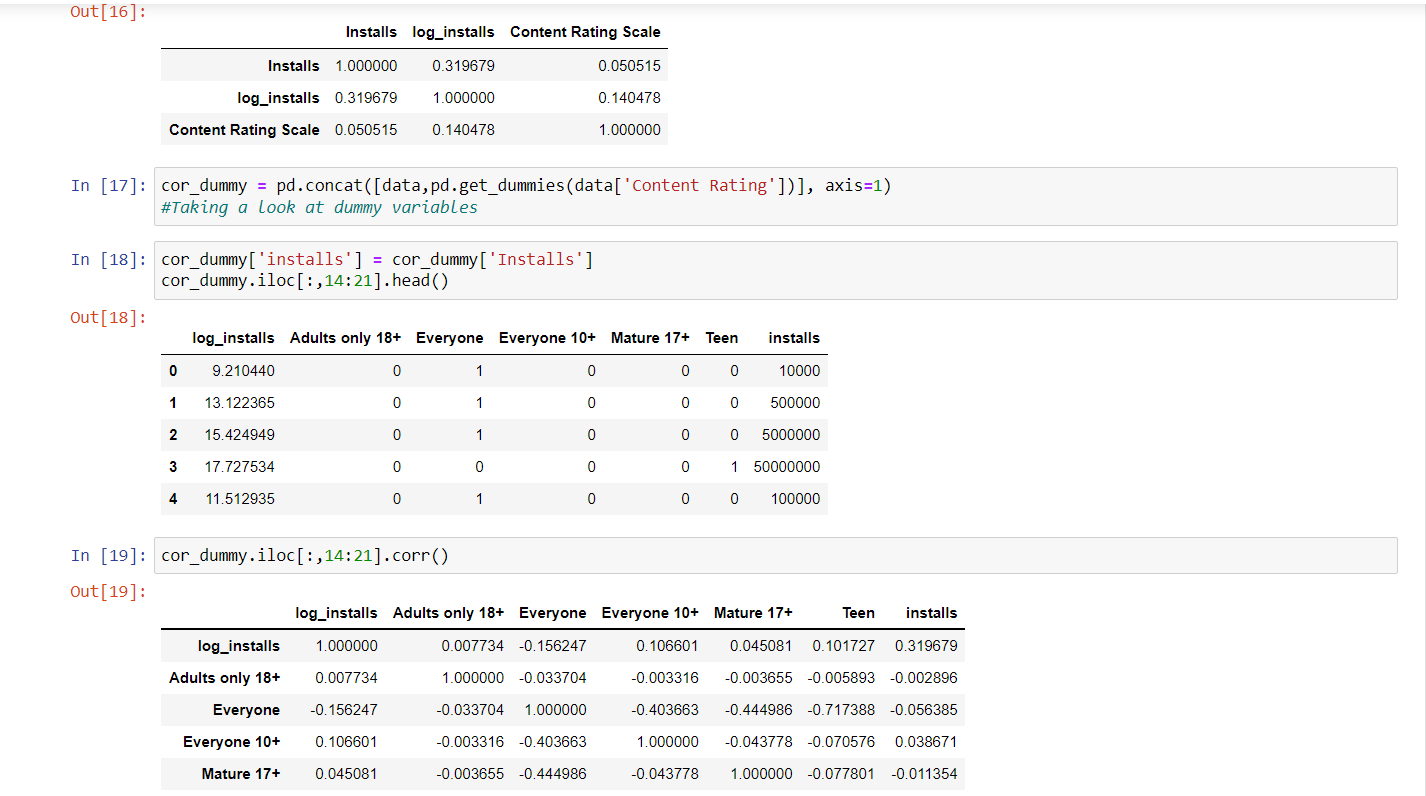
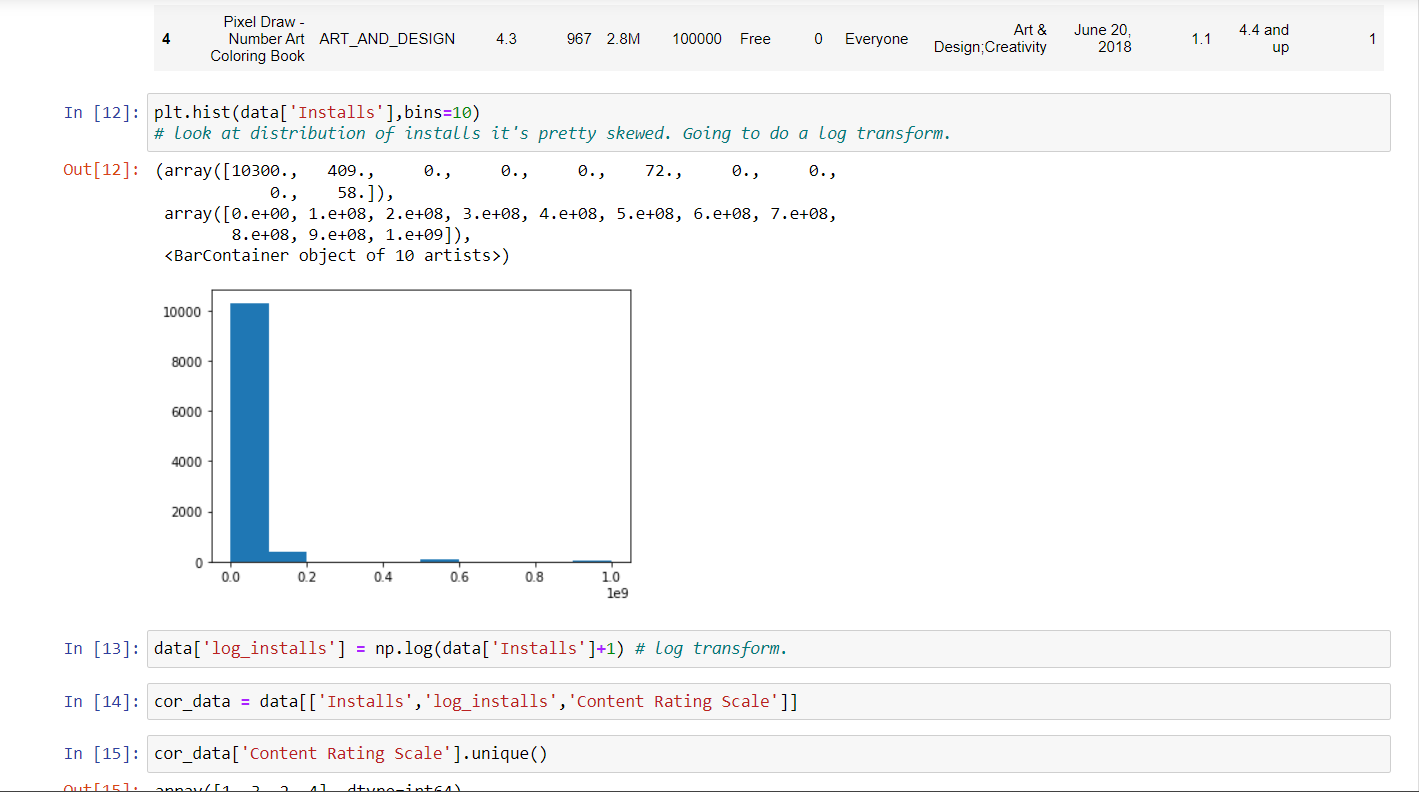
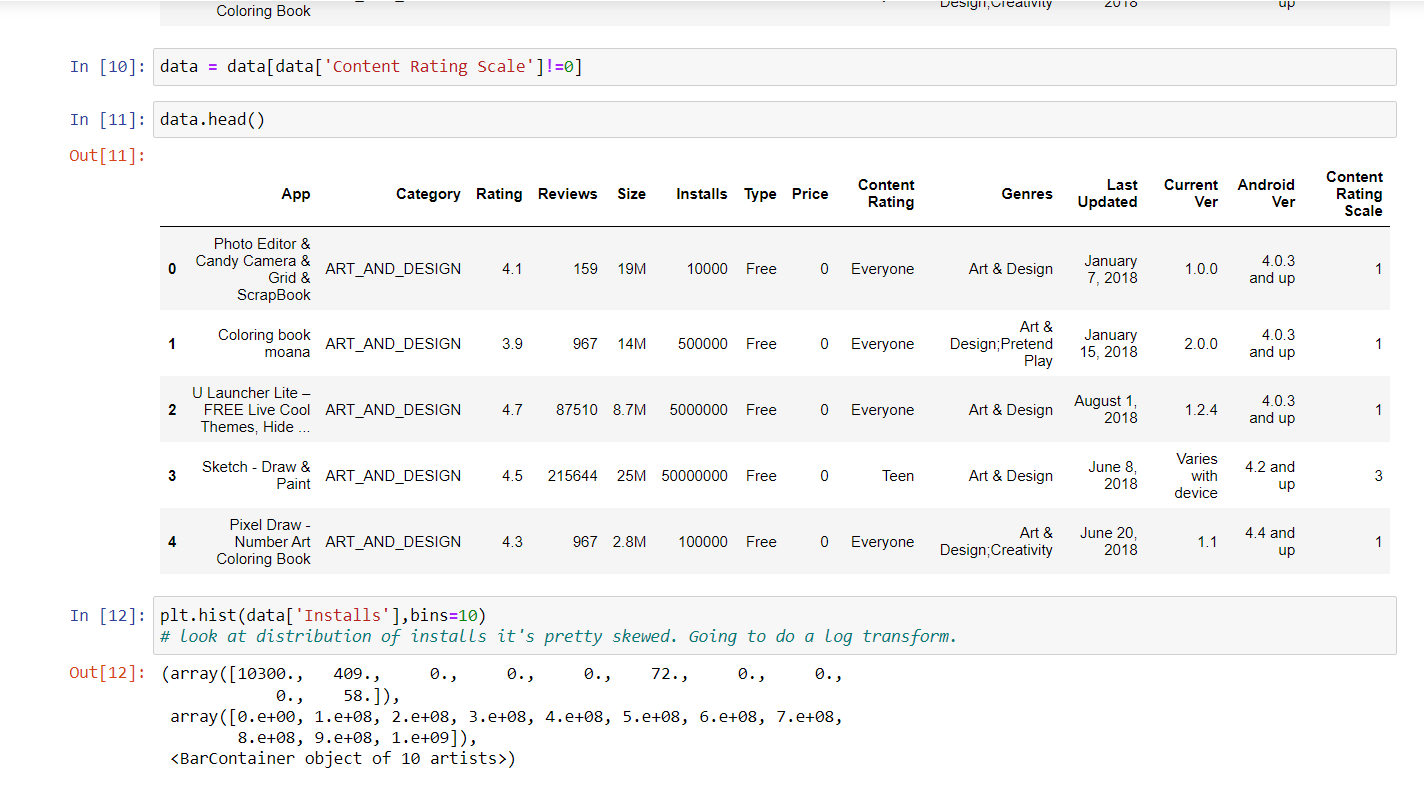
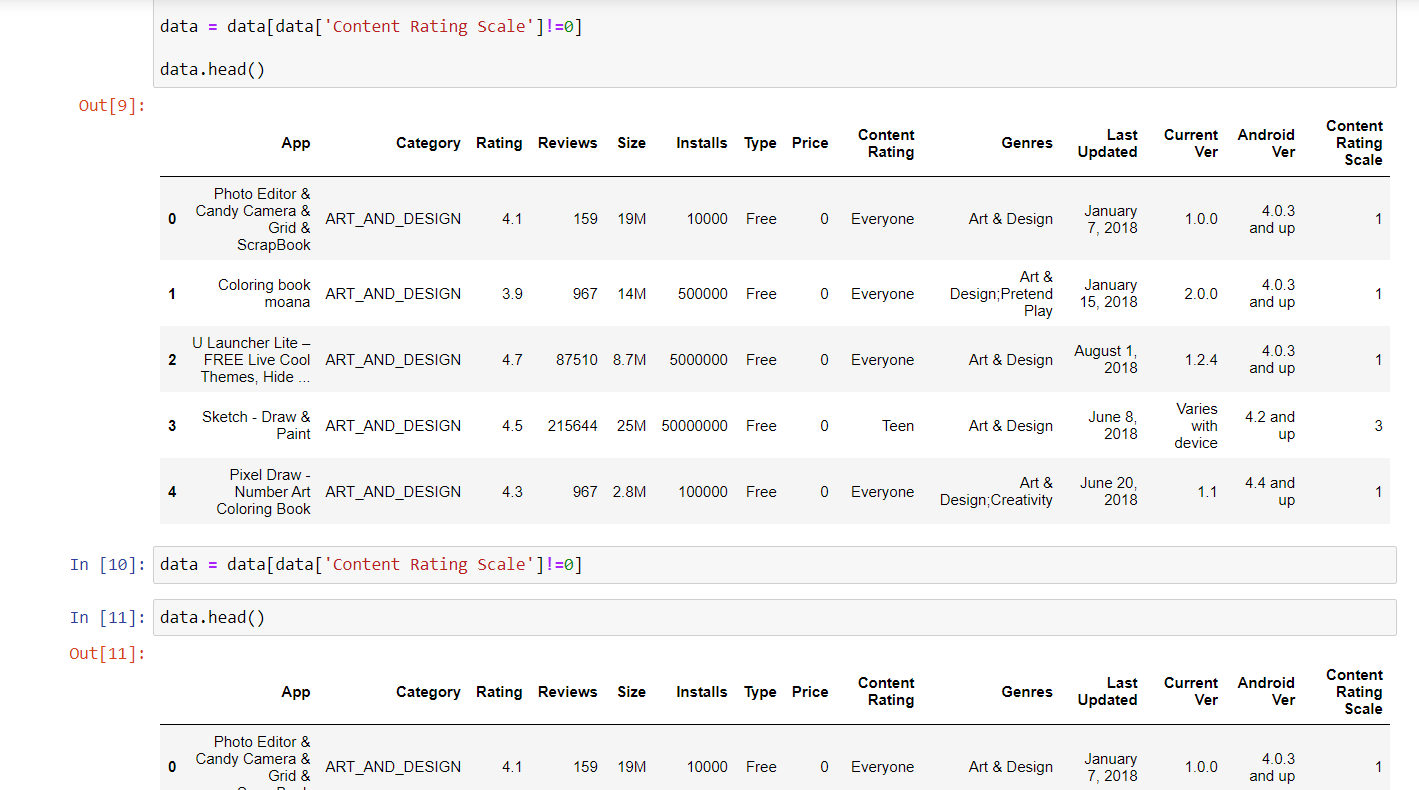
Finally, I looked at the dummy variables. A dataset may contain various types of values, sometimes it consists of categorical values. So, to use those categorical values for programming efficiently we create dummy variables. A dummy variable is a binary variable that indicates whether a separate categorical variable takes on a specific value. And with the dummy variable I checked for correlation.

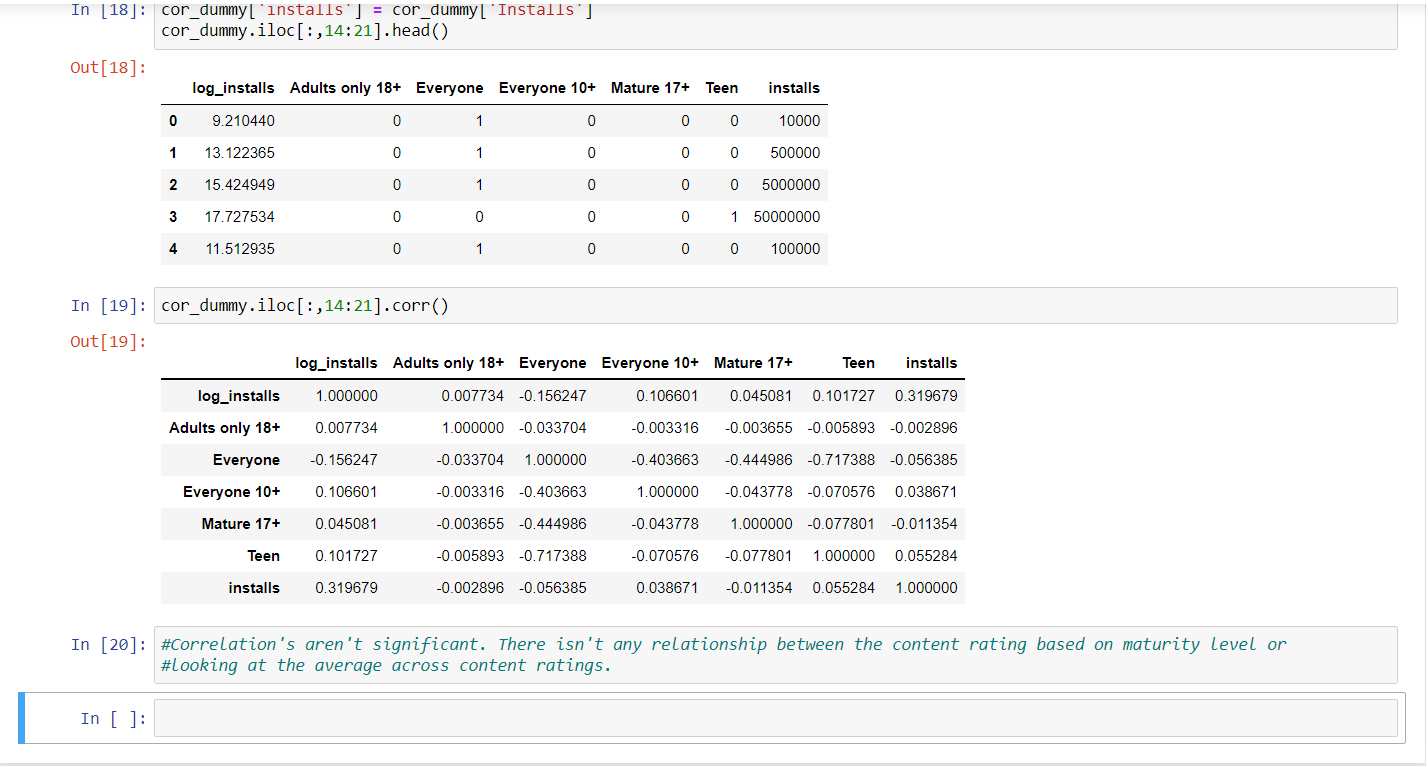












**Result and Conclusion:**

The data revealed that the content rating of Everyone had the highest number of installs. Along with those results, the category with a content rating of Everyone with the highest number of installs was Family. And the correlations aren't significant. There isn't any relationship between the content rating based on maturity level or looking at the average across content ratings.