Google Play Store Report and Analysis

Correlation between Content Rating and Installs per Category

February 2021 - Catharina van Veen

# Introduction

DigitTech Advertising has acquired a new mobile gaming company client. This client wants to broaden its targeted market in the mobile gaming app industry. But before they can settle on the types of games, they need to know what kind of games are being played and by whom.

The question we are tasked with here is: “Is there a correlation between content rating and the number of installs per category?” In this report I will be discussing the data, methods used, analysis performed, and the overall results. The null hypotheses I will be testing in this report will be: 1) All Content Ratings have equal Number of Installs, 2) All Categories have equal Number of Installs, and 3) There is no interaction between Content Ratings and Categories.

# Data

For this report I used the data and information about Google Play Store contained in the provided file, googleplaystoremaster.csv. The data consists of 10,841 rows and 13 columns all related to Google Play Store statistics and metrics. There are 10,841 apps with the following information per app contained within the data: Name, Category, Rating, Number of Reviews, Size, Number of Installs, Type, Price, Content Rating, Genres, Last Updated, Current Version, Android Version.

For the purpose of this report, I needed only the columns: Category, Content Rating, and Installs.

Quick inspection of the data showed that there was only one row with a null value, namely in the Content Rating. Closer look at this row makes clear that the row is corrupted, the values in the other columns are incorrect too. I decided to drop this row.

The data has 33 unique Category values and 6 unique Content Rating values.

The data has 21 unique Installs values: 0 /0+, 1+, 5+, 10+, 50+, ……., 1,000,000,000 stored as text values. I decided to convert these to numeric values.

# Method

First, I compared the number of Installs, grouped by Category and again, grouped by Content Rating, using box plots.

Then I did a regression analysis on to see if there is a linear relationship between Category and number of Installs and between Content Rating and the number of Installs.

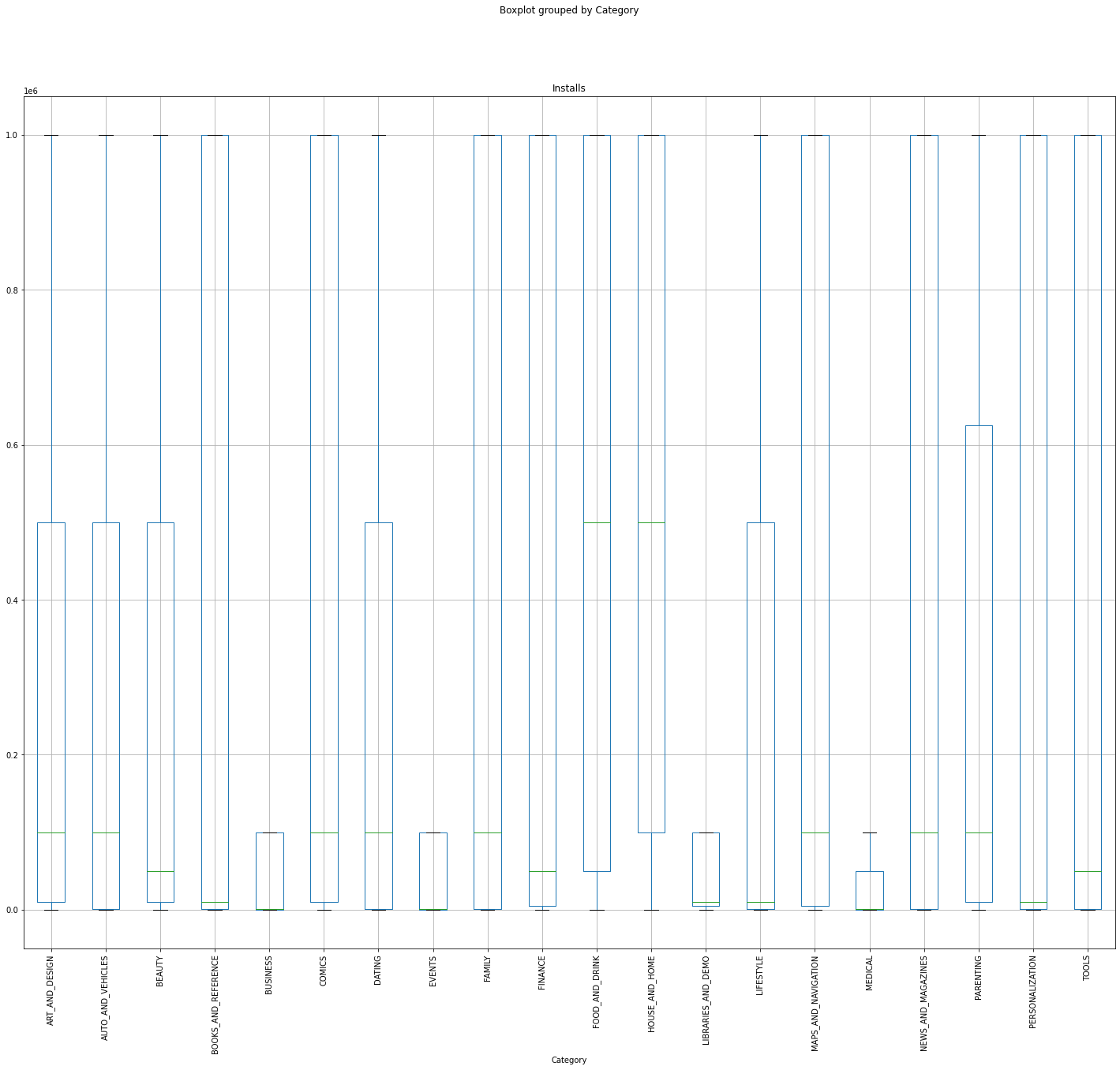
Since I had two categorical independent variables, Content rating and Category, I used the Two Way ANOVA test to analyze the data and see if there is a correlation between the independent variables and the number of Installs.

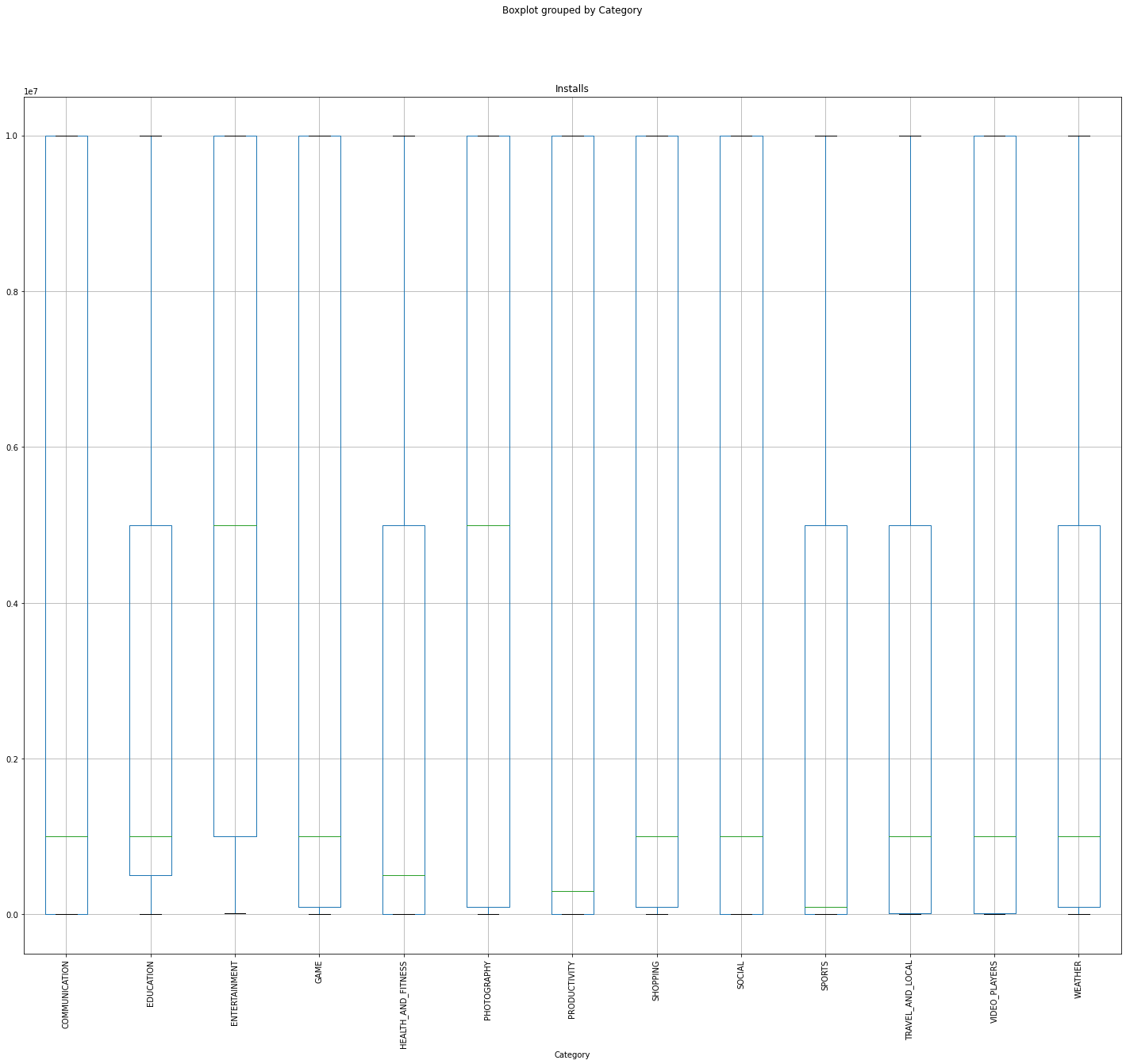
For these I needed the number of Installs to be a continuous variable which is why I converted it into a numeric variable.

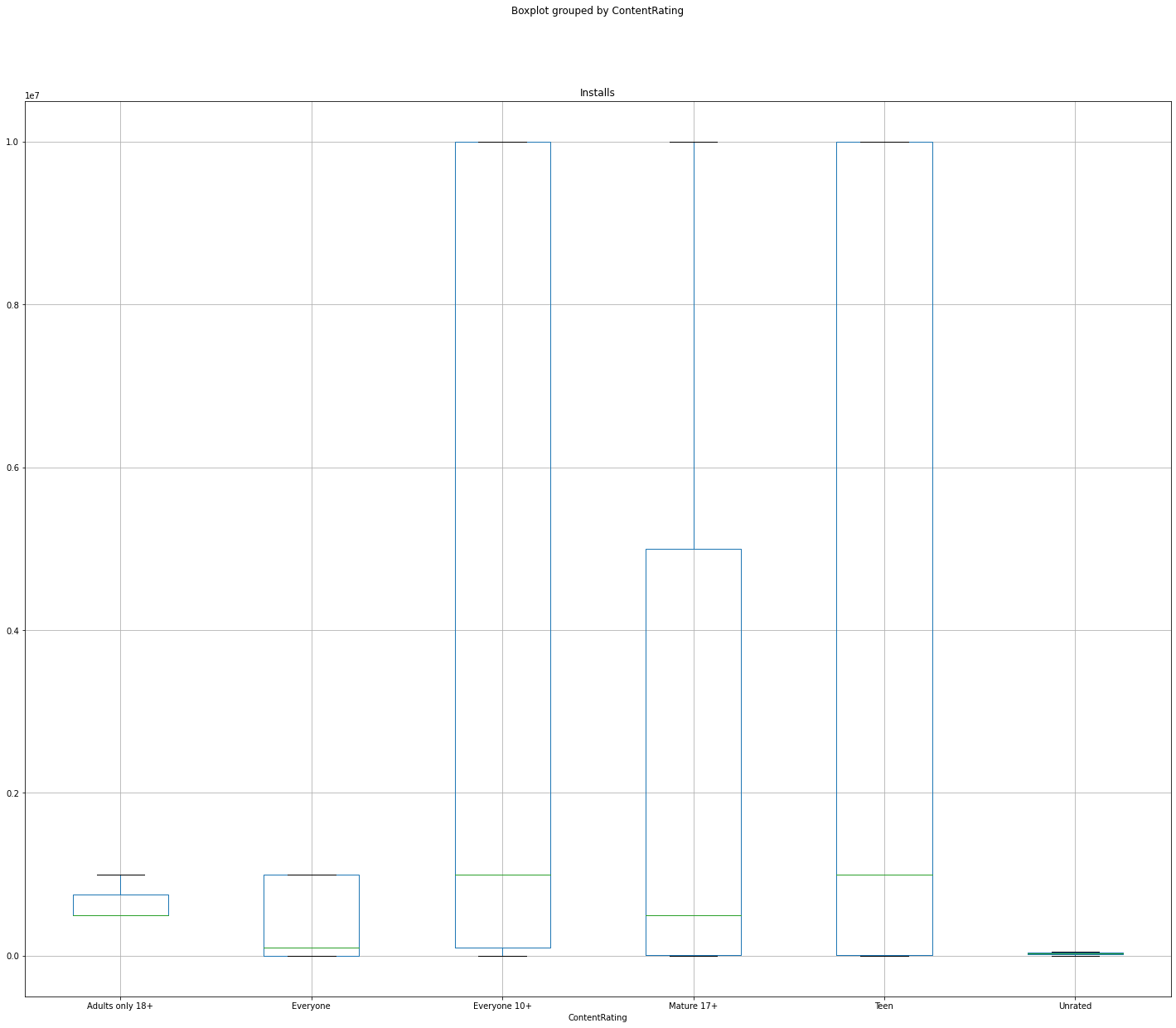
# Analysis

## Box plots

Since there is a huge range in the number of Installs per app, I split the data in two groups for the purpose of showing their box plots here. Group one: Categories with the 3rd quartile for number of Installs at or below 1,000,000. Group 2: Categories with the 3rd quartile above 1,000,000. Below are the box plots for both groups. For readability, the outliers are not shown.



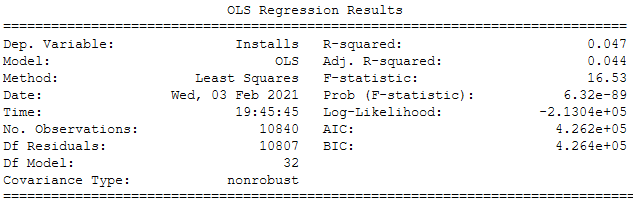




## Regression analysis

### Category

model\_category = ols('Installs ~ C(Category)', data).fit()

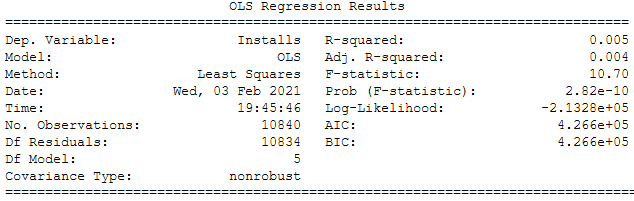


Although the p-value is almost zero, for most of the individual Category values the p-value is high. For the Categories listed below however the p-value was below 5%, showing a statistical significance.

|  | **coef** | **std err** | **t** | **P>|t|** | **[0.025** | **0.975]** |
| --- | --- | --- | --- | --- | --- | --- |
| **C(Category)[T.COMMUNICATION]** | 82450000.0 | 11100000.0 | 7.397 | 0.000 | 60600000.0 | 104000000.0 |
| **C(Category)[T.GAME]** | 28760000.0 | 10600000.0 | 2.712 | 0.007 | 7980000.0 | 49500000.0 |
| **C(Category)[T.NEWS\_AND\_MAGAZINES]** | 24580000.0 | 11400000.0 | 2.149 | 0.032 | 2160000.0 | 47000000.0 |
| **C(Category)[T.PHOTOGRAPHY]** | 28200000.0 | 11300000.0 | 2.503 | 0.012 | 6110000.0 | 50300000.0 |
| **C(Category)[T.PRODUCTIVITY]** | 31520000.0 | 11100000.0 | 2.846 | 0.004 | 9810000.0 | 53200000.0 |
| **C(Category)[T.SOCIAL]** | 45780000.0 | 11400000.0 | 4.019 | 0.000 | 23500000.0 | 68100000.0 |
| **C(Category)[T.TRAVEL\_AND\_LOCAL]** | 24710000.0 | 11500000.0 | 2.141 | 0.032 | 2090000.0 | 47300000.0 |
| **C(Category)[T.VIDEO\_PLAYERS]** | 33640000.0 | 12100000.0 | 2.786 | 0.005 | 9970000.0 | 57300000.0 |

### Content Rating

model = ols('Installs ~ C(ContentRating)', data).fit()



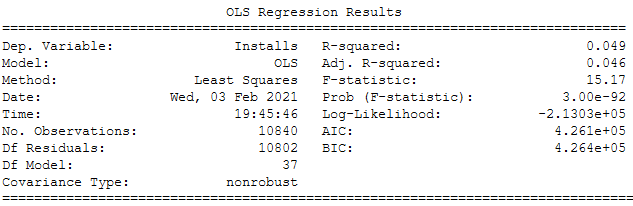
Again, the p-value is almost zero, but for all the individual Content Rating values the p-value is high, showing there is no statistical significance.

## Two-way ANOVA

### No interaction between the two independent variables

I did a regression analysis considering both the Content Rating and the Category not taking into account the interaction between the two independent variables:

model\_CategoryPlusContent = ols('Installs ~ C(ContentRating) + C(Category)', data).fit()



Although the p-value is almost zero, for all the individual Content Rating values and most of the individual Category values the p-value is high, showing no statistical significance. For the Categories listed below however the p-value was below 5%, showing a statistical significance.

|  | **coef** | **std err** | **t** | **P>|t|** | **[0.025** | **0.975]** |
| --- | --- | --- | --- | --- | --- | --- |
| **C(Category)[T.COMMUNICATION]** | 82450000.0 | 11100000.0 | 7.406 | 0.000 | 60600000.0 | 104000000.0 |
| **C(Category)[T.GAME]** | 24630000.0 | 10600000.0 | 2.317 | 0.021 | 3790000.0 | 45500000.0 |
| **C(Category)[T.PHOTOGRAPHY]** | 28650000.0 | 11300000.0 | 2.546 | 0.011 | 6590000.0 | 50700000.0 |
| **C(Category)[T.PRODUCTIVITY]** | 32030000.0 | 11100000.0 | 2.896 | 0.004 | 10300000.0 | 53700000.0 |
| **C(Category)[T.SOCIAL]** | 42900000.0 | 11500000.0 | 3.731 | 0.000 | 20400000.0 | 65400000.0 |
| **C(Category)[T.TRAVEL\_AND\_LOCAL]** | 25250000.0 | 11500000.0 | 2.191 | 0.028 | 2660000.0 | 47800000.0 |
| **C(Category)[T.VIDEO\_PLAYERS]** | 32830000.0 | 12100000.0 | 2.721 | 0.007 | 9180000.0 | 56500000.0 |

sm.stats.anova\_lm(model\_CategoryPlusContent)

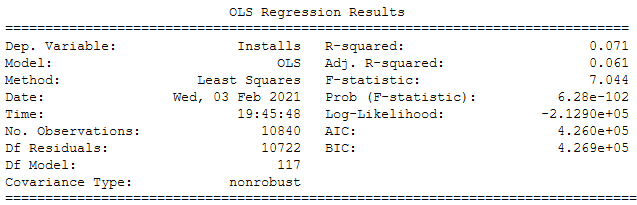
|  | **df** | **sum\_sq** | **mean\_sq** | **F** | **PR(>F)** |
| --- | --- | --- | --- | --- | --- |
| **C(ContentRating)** | 5.0 | 3.850541e+17 | 7.701082e+16 | 11.166922 | 9.375111e-11 |
| **C(Category)** | 32.0 | 3.486638e+18 | 1.089574e+17 | 15.799327 | 2.342911e-84 |
| **Residual** | 10802.0 | 7.449419e+19 | 6.896334e+15 | NaN | NaN |

As the p-values are very close to zero, the ANOVA test confirms that the differences between the means are statistically significant.

### Interaction between the two independent variables

I did a regression analysis considering both the Content Rating and the Category taking into account the interaction between the two independent variables:

model\_CategoryByContent = ols('Installs ~ C(ContentRating) \* C(Category)', data).fit()



Although the p-value is almost zero, for all the individual combinations of Content Rating values and Category values the p-value is high, showing no statistical significance.

# Conclusion

Based on the analysis I:

* Accept the null hypothesis 1) All Content Ratings have equal Number of Installs.
* Reject the null hypothesis 2) All Categories have equal Number of Installs with regards to the categories: COMMUNICATION, GAME, PHOTOGRAPHY, PRODUCTIVITY, SOCIAL, TRAVEL\_AND\_LOCAL, and VIDEO\_PLAYERS.
* Accept the null hypothesis 3) There is no interaction between Content Ratings and Categories.

In answer to the question asked, “Is there a correlation between content rating and the number of installs per category?”, I conclude that there is no relation between the number of Installs and the Content Rating per Category. There is a correlation between number of Installs and Category for some of the categories, but that falls outside the scope of this question.