Capturing the Android Market:

Insights from Google Play Store Data

Business Intelligence and Data Mining with SAS Suite (94-832)

Team # 6

Zhe Li

Prasun Shrestha

Wendy Velaquez Ebanks

Alyssa Wroblewski

Eric Yuan

Carnegie Mellon University

December 9, 2019

**Table of Contents**

[**Executive Summary**](#_1g2bnus629q7) **2**

[**Overview**](#_kic3emyel194) **3**

[**Our Business Question**](#_e0xfgwkejii0) **3**

[**The Data**](#_tfx5ciem9c82) **3**

[Basic Overview](#_7nlq4wjaag2w) 3

[Dependent Variable: Ratings and Number of Installs](#_ao7qf4b4da7u) 4

[Independent Variables: Genre and Price](#_8zr2prjdxqoc) 4

[Extra](#_i1hbspsx9s3g) 4

[**Data Cleaning**](#_jtkvto4p10ep) **4**

[Combine Data Sets](#_3cp3a3kursu) 4

[Size](#_7mcjmi3erqpw) 5

[Number of Installations](#_15pc9q36pkoj) 5

[Latest Android Version Needed](#_dd5e9n87wv1u) 5

[App Name Length](#_l9wo5wnp15rd) 5

[Install Class](#_hmf3khtanliw) 6

[Updated Time](#_3isb6nh8jpam) 6

[**Running File in SAS**](#_hrvgec1tikar) **6**

[**Supervised Models**](#_rn3frpd346bv) **7**

[**Unsupervised Models**](#_ki4r1u2lgwj) **14**

[**Conclusions**](#_m58sz8ykh6m4) **16**

[**Bibliography**](#_jus4roqske67) **18**

# **Executive Summary**

Google Play is Google's official store and portal for Android apps, games and other content to Android-powered phones, tablets or [Android TV](https://www.androidcentral.com/android-tv) devices. It presents people with personalized collections of apps and games, based on criteria such as the user’s past activity, actions they’re trying to complete, location, and major events. Google play store captures early consumer interest and accelerate launch performance.

According to a Statista research study, 2.8 million applications from Google Play are available for download this year. However, not all the applications released and published are well received by the consumer for many reasons such as the type of application, the features it contains, who is the target audience, etc. Our study, focus on providing insights to Android developers as to what constitutes a good and marketable application for Google play store. We provide insights in the following aspects: the number of reviews, ratings, and category.

This research is to construct helpful suggestions for app developers to increase the installation or rating of their apps. The dataset we use is retrieved from Kaggle. The influential factors we chose to include app category, review quantity, size of the app, type of app (free or paid), updated time (number of months since last update), and Android version needed. We use both supervised learning (regression and decision tree) and unsupervised learning (association rules) in this research.

Through these models, we find that Gaming, Family, Communication and Productivity apps have more market demand and are more likely to be downloaded. Another two important factors influencing user’s installation are the price of the app and the Android version needed for the app. Additionally, Gaming and Personalization apps are more likely to get high ratings, and updating frequency also impacts app’s rating. We provide these helpful suggestions and insights to Android app developers to increase the installation or rating of the apps developed.

# **Overview**

For Android developers, Google Play Store is the leading platform to publish an application. The amount of applications launched is over 2 million just in the last three years. In addition, Google Play store reaches billions of Android devices in 190 countries and territories around the world. Consequently, for developers who want their app to be successful, will certainly reward any insight that can help their apps to increase the reach. Users may interact with each application and platform differently, and an app that is wildly successful on one platform may not be as successful in another. Therefore, developers have to consider these differences if they want the launch of their app to be successful.

# **Our Business Question**

If you are an Android app developer, then to consider your app successful, you want your app to get the highest review and as many downloads as possible. But what factors constitute the reviews and downloads in Google’s Play Store market? What should app-developers consider if they were to capture the Android market and drive your app-making business to success?

# **The Data**

## Basic Overview

The Google Play Store Apps Downloads data is the main dataset we are analyzing. This data was scraped on September 18, 2018. Each observation is a different app. Factors measured include the average rating, the number of installations, the number of reviews, and the price of the app, to name a few. We chose this dataset to represent our business question because of the amount of information that can be derived from it.

The Google Play Store App Reviews data is our secondary dataset that we are analyzing. This dataset includes written reviews from users, sentiment polarity, sentiment subjectivity, and the positive, neutral, and negative sentiment levels of each app.

We attempted to combine both datasets by using Python and SQL before importing into SAS. At first, this was very tricky because the sentiment values were creating duplicates of the data. If we attempted to remove the duplicates, then data was lost. Our solution was to merge these duplicates in SQL so that each observation had a rounded sentiment ratio.

## Dependent Variable: Ratings and Number of Installs

Within our review, the ratings are the first main variable we will use to determine an app’s success. Our other factor we will use to determine any app’s success include the number of times an app was installed since the data was measured. Because the data labels these numbers as benchmarks rather than each app’s true number, the installs will be categorized as nominal variables. With these two, we will use these as our base criteria to show marketers example of successful apps.

## Independent Variables: Genre and Price

We will not observe the number of installs with the ratings against the titled Genre and Price columns. The Genre column is extremely similar to the Category column already, so including the genre column will not be necessary. This is especially the case when observing that some apps were given two different genres within the same variable. Additionally, in our analysis, we will also observe Type rather than Price. This is because only 801 apps out of the 10,841 apps in the Google Play Data are paid apps, and Type observes them on a binary level of “free” or “paid.” As we’ll observe later on, any apps labeled as “free” made a great impact within our conclusions.

## Extra

Within our analysis, we did notice something interesting when it comes to any apps that are above the price of $200. Any apps between the price of $200 and $400 are “I Am Rich” apps. These apps have no function other than potentially providing an “I am rich” mantra, but their sole purpose is to prove that the purchaser has a lot of money. 16 apps are included in the observation. One final tidbit is that the most expensive app in the data of $400, with a difference of one cent from the next most expensive app, is the “I Am Rich – Trump Edition.” Because of the wide price range, this is additional reasoning for to observe the data in the Type column rather than the Price column.

# **Data Cleaning**

## Combine Data Sets

The Google Play Store Apps Downloads data and the Google Play Store App Reviews data were combined. This is so that we can also make conclusions as well an app did based off of the content of the reviews. We used Python and SQL to combine the datasets before placing them into SAS, plus Excel for additional adjustments.

## Size

The size column indicates an app’s download size. Figures like “M”, “k’, “varies with device”, and “+” were removed to help the process of sorting the data. An M indicates that the download size is in megabytes, a k indicated kilobytes,and a + symbol indicates that the download size is larger than the benchmark amount of 1000 megabytes.

We ran into one discrepancy when removing the M and k figures in the variables: 300 MB is not the same as 300 KB, but removing the letters caused the numbers to be measured the same way. Since 316 out of 10841 variables are measured in KB (about 3% of data), we decided to delete these from our final data for easier cleaning and because each app measured in KB was not a well-known, popular app.

## Number of Installations

The installs column indicates the number of times an app has been downloaded. This data collection is interesting because it is measuring from different benchmark levels rather than the actual number. For example, if an app was downloaded 12,305 times, the data would label the app as 10,000+ downloads.

Each app ranged from 0+ to 1,000,000,000+ downloads. We removed the + signs on all numbers for labeling purposes. In addition, we classified and categorized the data to signify what was a low number of downloads and what was a high number.

## Latest Android Version Needed

This column indicates the version of Android needed to be installed on the phone in order to download the app. Before, results were listed as a certain decimal value with “and up” attached to the end of each variable. An example is “4.0.3. and up,” which means that users needed to have Android version 4.0.3 or higher in order to install the app. We adjusted these numbers for easier processing within SAS. For example, we changed “4.0.3. and up” to “403.”

## App Name Length

This is the number of characters that appear in an app’s name. While this is not an original column within the dataset, we decided to measure this out of curiosity. This is to see if the length of the app’s name had an impact in downloading. To create this column, we simply made a new row and put in a formula to count the number of characters in each app name variable.

## Install Class

In addition to cleaning the number of installations, we also decided to place in a new column called “Install Class.” This is to measure whether an app had been among a low number of installs, a medium amount, or a high amount.

We chose to do this because the benchmarks assigned to the apps are not measured in equal amounts. For example, some of the benchmarks are between 50+ and 100+, while others are between 1,000,000+ and 5,000,000+ with no additional benchmarks in between both observations. Therefore, rather than attempting to measure the benchmark numbers as is, we created Install Class to help with the process.

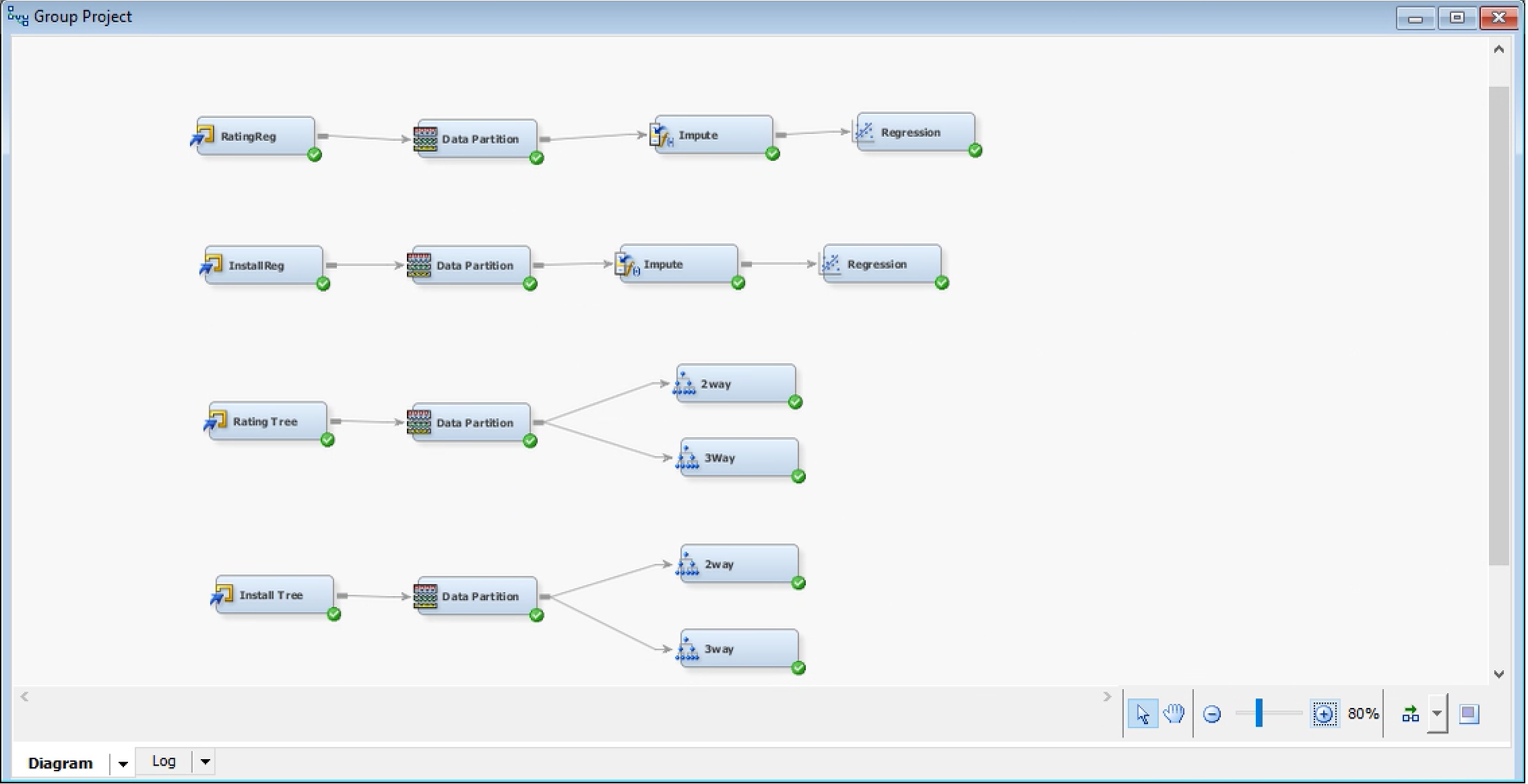
## Updated Time

Most apps have been updated before. Updating an app mainly because the developer wants to fix some problems on the app, so we believe that updating frequency could be one of the factors influencing user’s decision on downloading it or rating it. We calculate the time period after the app was updated last time. We measure the time period in months. That said, a value ‘10’ means that this app was updated 10 months ago.

# **Running File in SAS**

Due to an unknown error, the full dataset file had to be converted from a CSV file to an Excel file. The ID number column was also named “AppID” before the file was converted and imported.

Some variables were missing values within columns such as the ratings. An impute node was added to place in synthetic values. A data partition node was also added to train the data. Our final chart is represented in Figure 1.



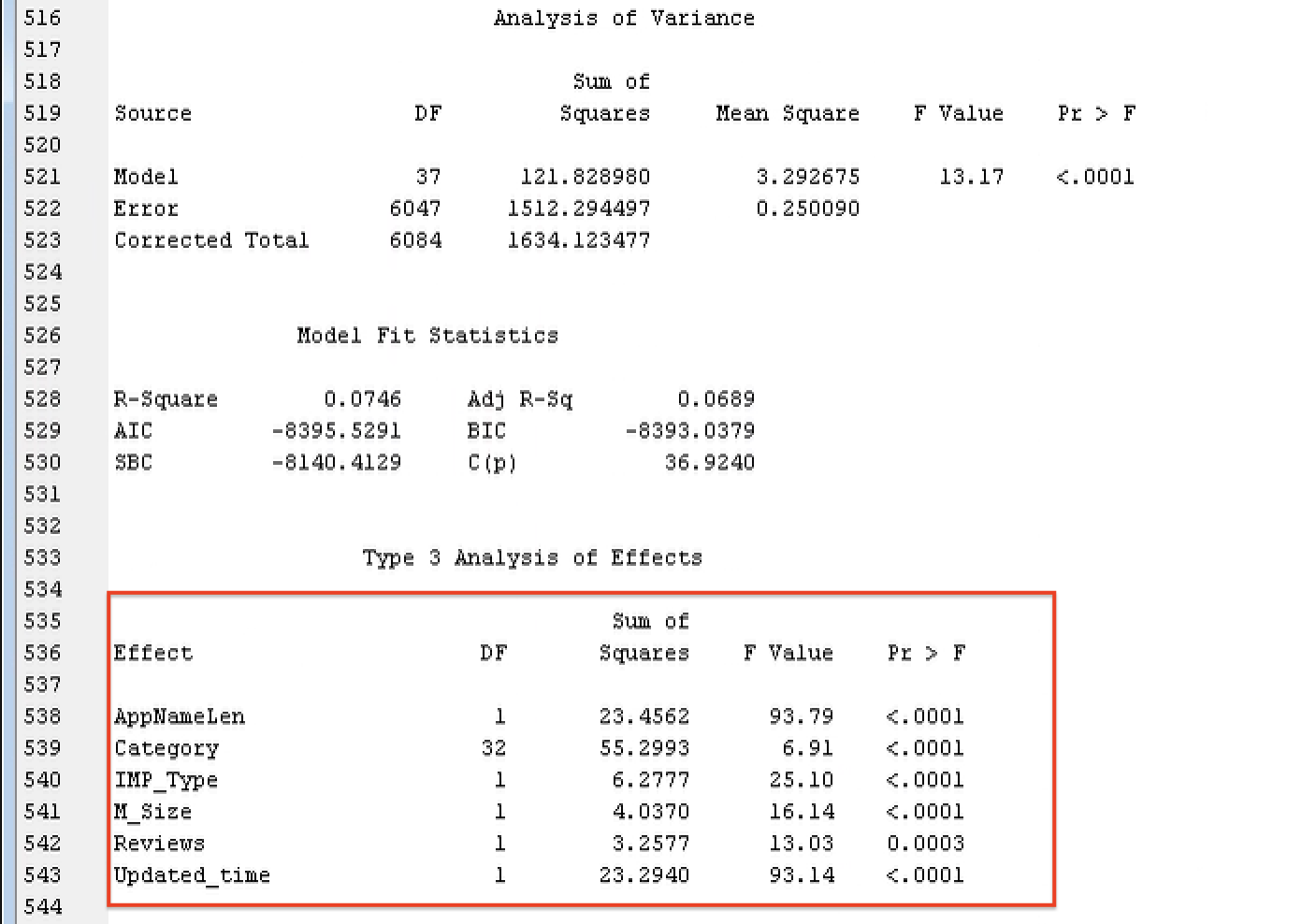


# **Supervised Models**

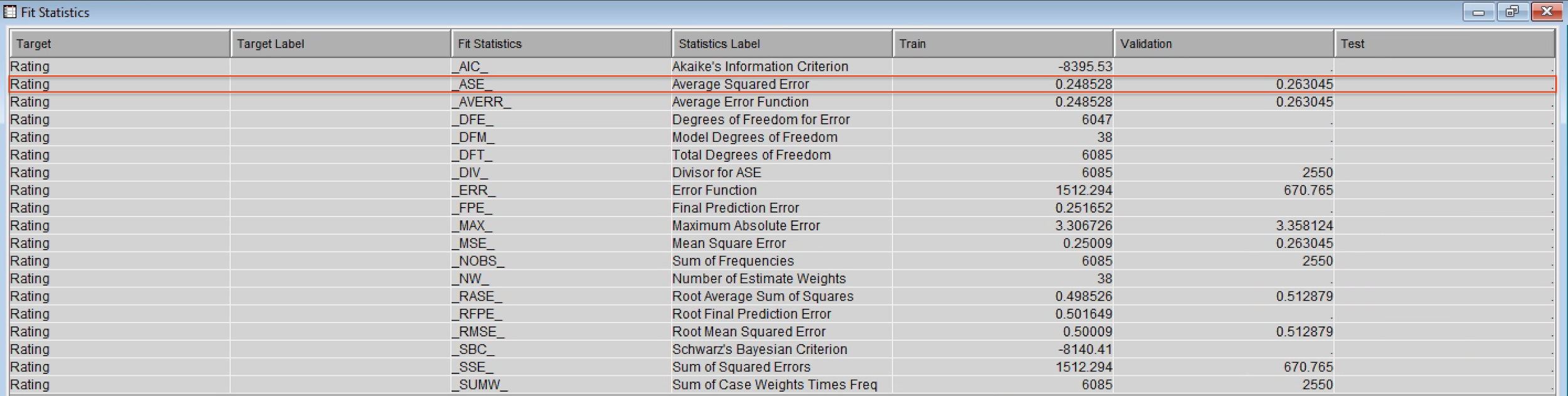
Two tests were run on both the complete and reduced version of the dataset. The complete dataset is the combination of both the Downloads and Reviews dataset, while the reduced version has all duplicates removed. Two-way decision trees were run for both the ratings and number of installs. In addition, linear regression models were run for both ratings and number of installs.

Algorithm 1: Regression

Our first linear regression was run in consideration of app ratings. The length of the app’s name, the categories, the type (free or paid), the size, the number of reviews, and the time were all statistically significant.



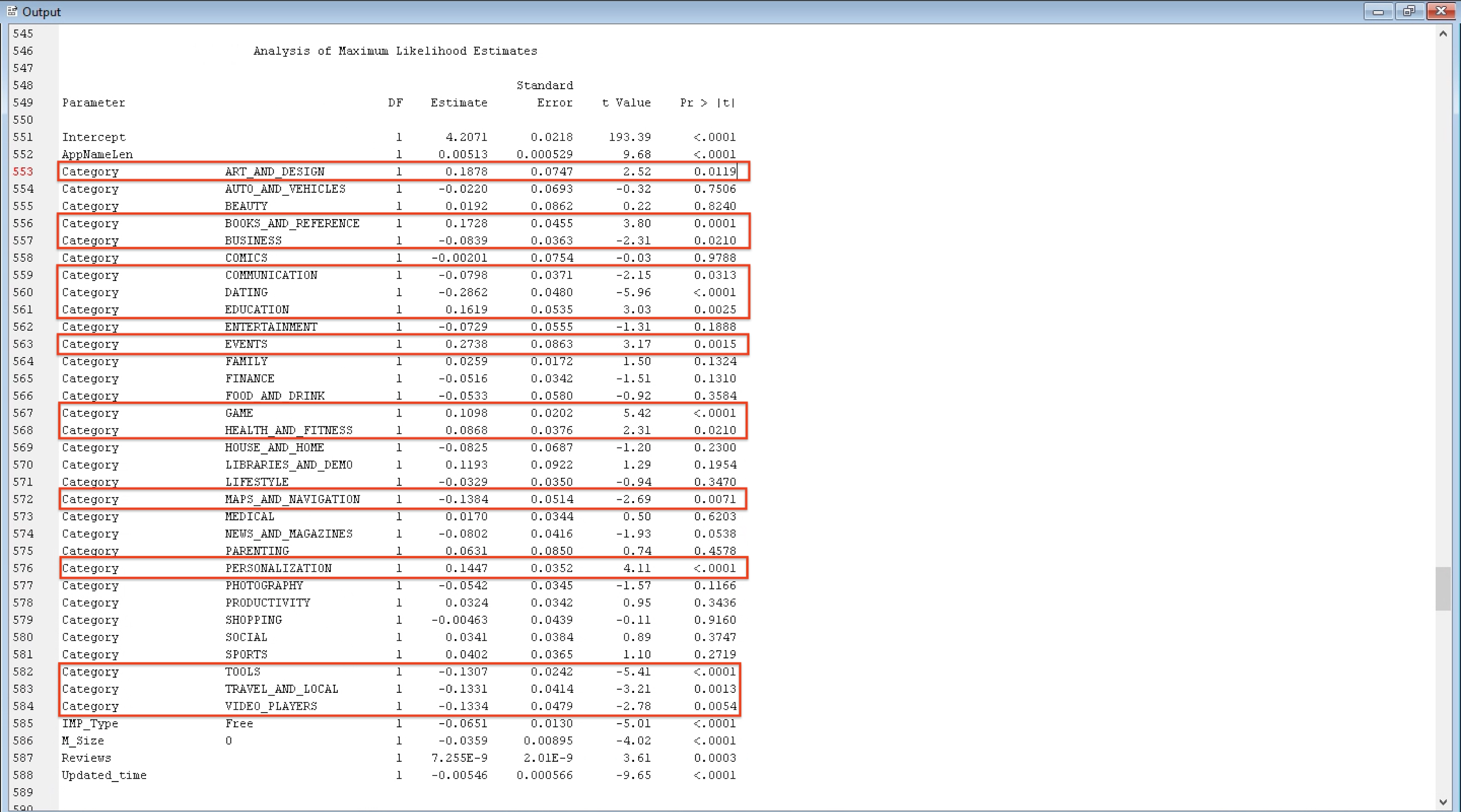






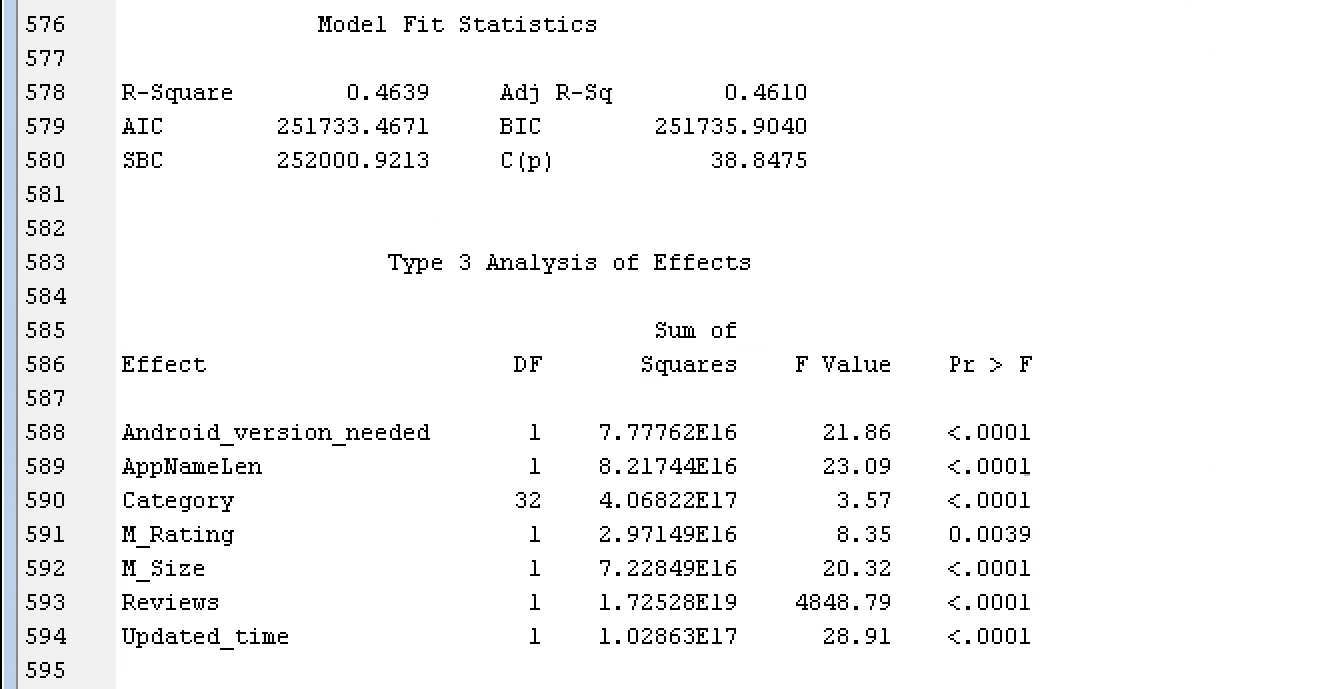
The variables in the regression have different coefficients. AppNameLen and Reviews have positive coefficient, indicating their positive relationship with app rating. Type, Size, and Updated\_time have negative coefficients. Therefore, the more frequent the app is updated, the more ratings the app might get; the smaller size the app has, the more ratings it might get; if the app is free, it might get lower ratings. These findings do not indicate causality but there are indeed some reasonality behind them. For example, free apps get lower ratings possibly because they don’t have enough funding to make it perfect.

In taking a further look at our categories, we found some that had a positive correlation and others that had a negative. Out of all of these categories, Gaming, Dating, Personalization, and Tools were the most statistically significant. The results indicate that developing game apps and personalization apps might have higher rating but developing dating apps and tool apps would suffer the risk of getting lower rating.



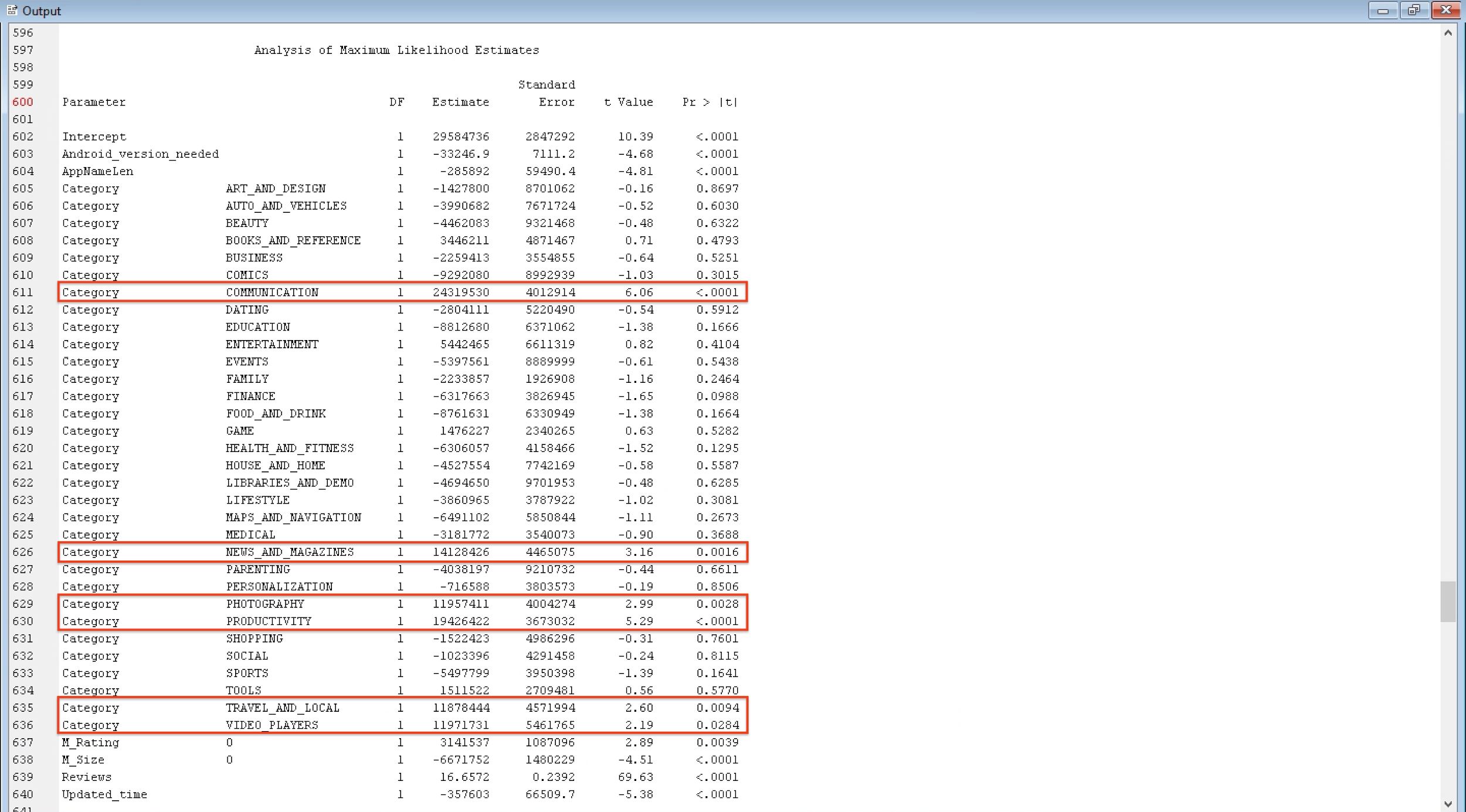


When running another linear regression with the number of installations in consideration, we found that the length of the app’s name, the categories, the size, the number of reviews, and time were all statistically significant again. However, instead of the type (free or paid) being statistically significant, the latest Android version needed was a statistically significant factor when looking at the number of installs. This indicates that the developer needs to consider the universality of their app.





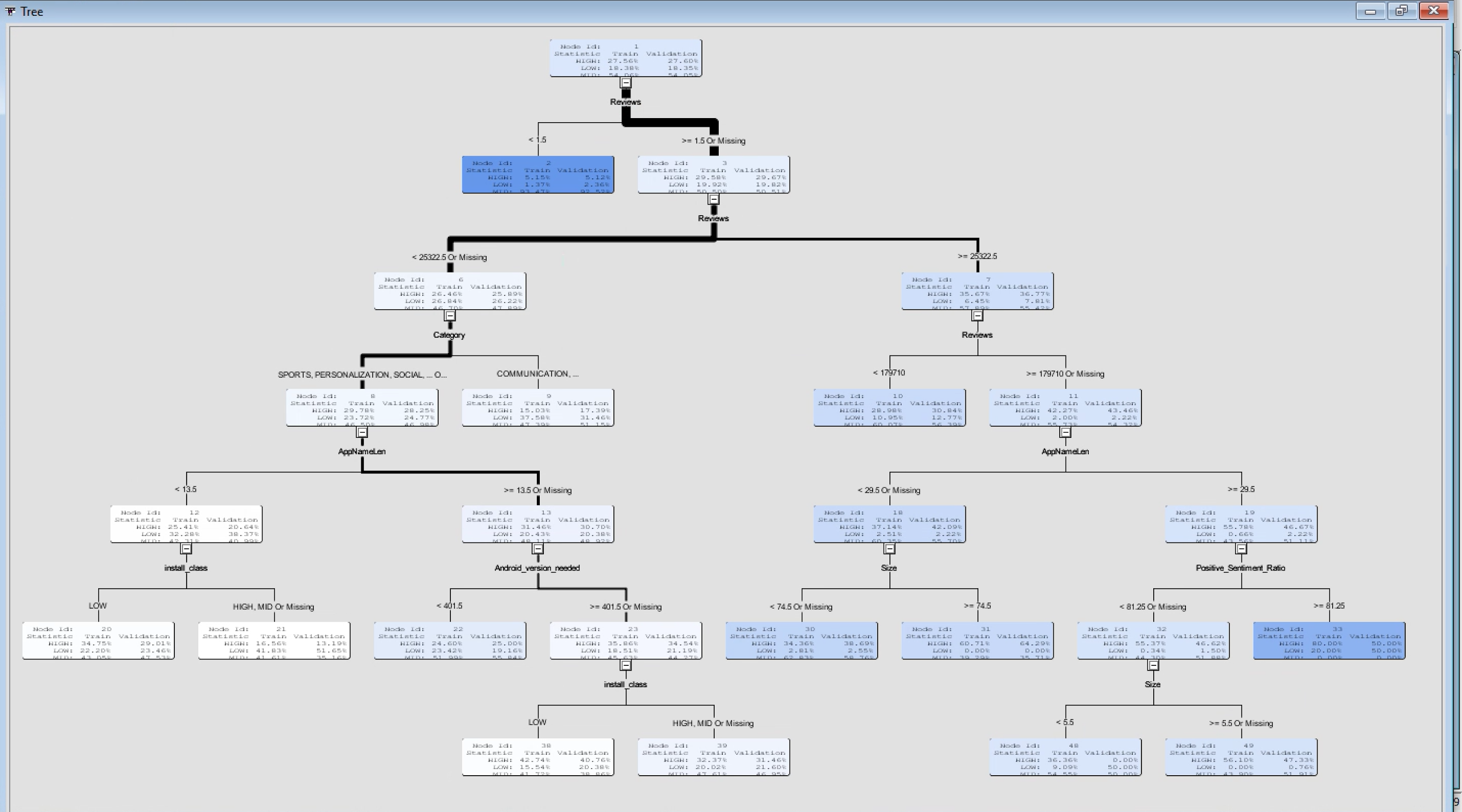
In taking a closer look at the categories for the number of installs, Communication and Productivity were the categories with the most statistical significance and they both have positive coefficient, indicating their positive relationship with app installation. So, these types of apps might have higher market demand.





Algorithm 2: Decision Tree

Additionally, we measured the App Ratings within a two-way decision tree. Apps ranked at a moderate level made up the majority of data at 54.05%, while highly rated app were at 27.6% of the data and poorly rated apps were at 18.35% of the data. In measuring apps rated 1.5 or higher (or apps with missing review variables), review ranked at a moderate level dominated at 50.51%. The tree has a total of 13 leaves with a misclassification rate of 0.4459.

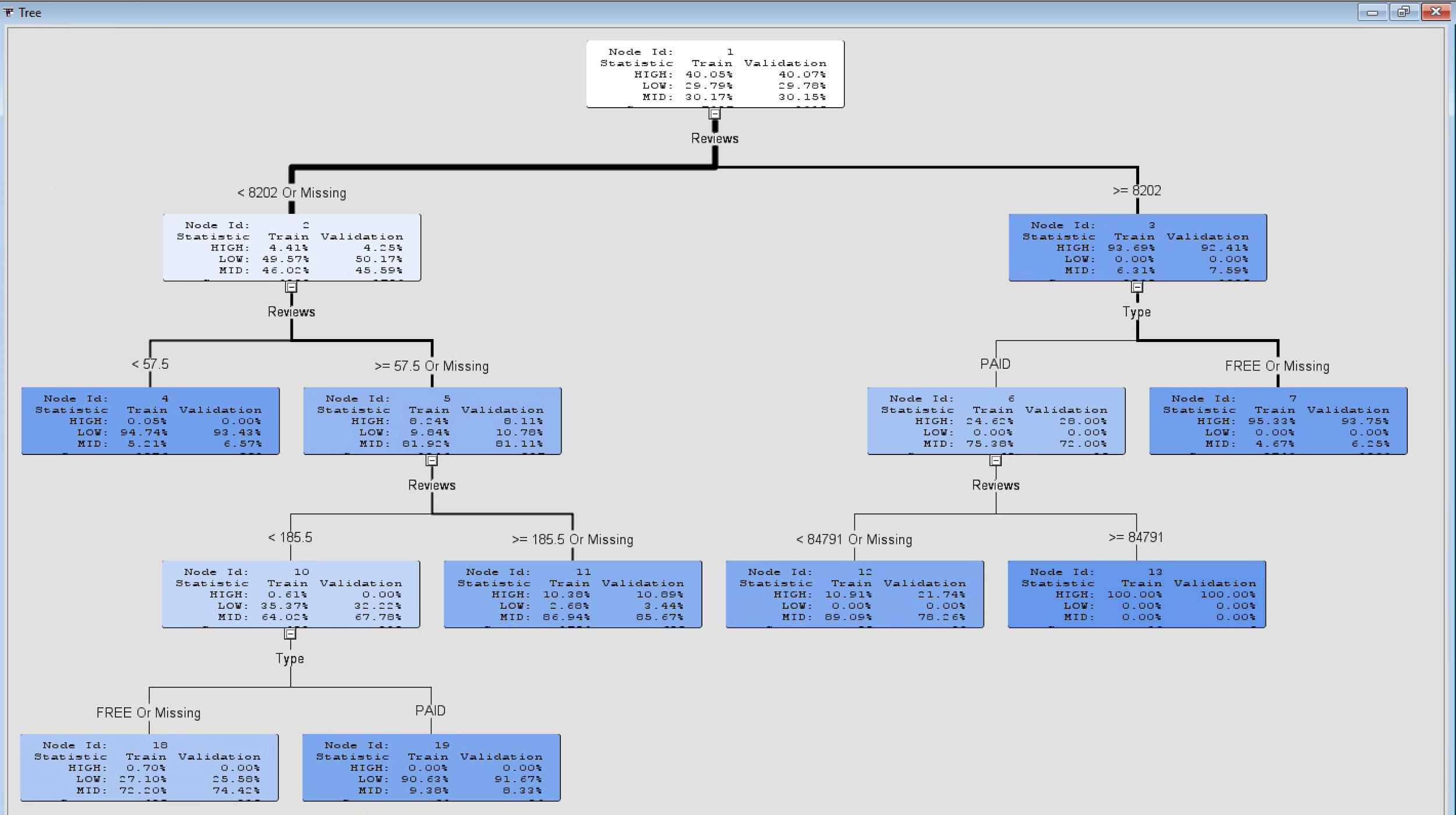




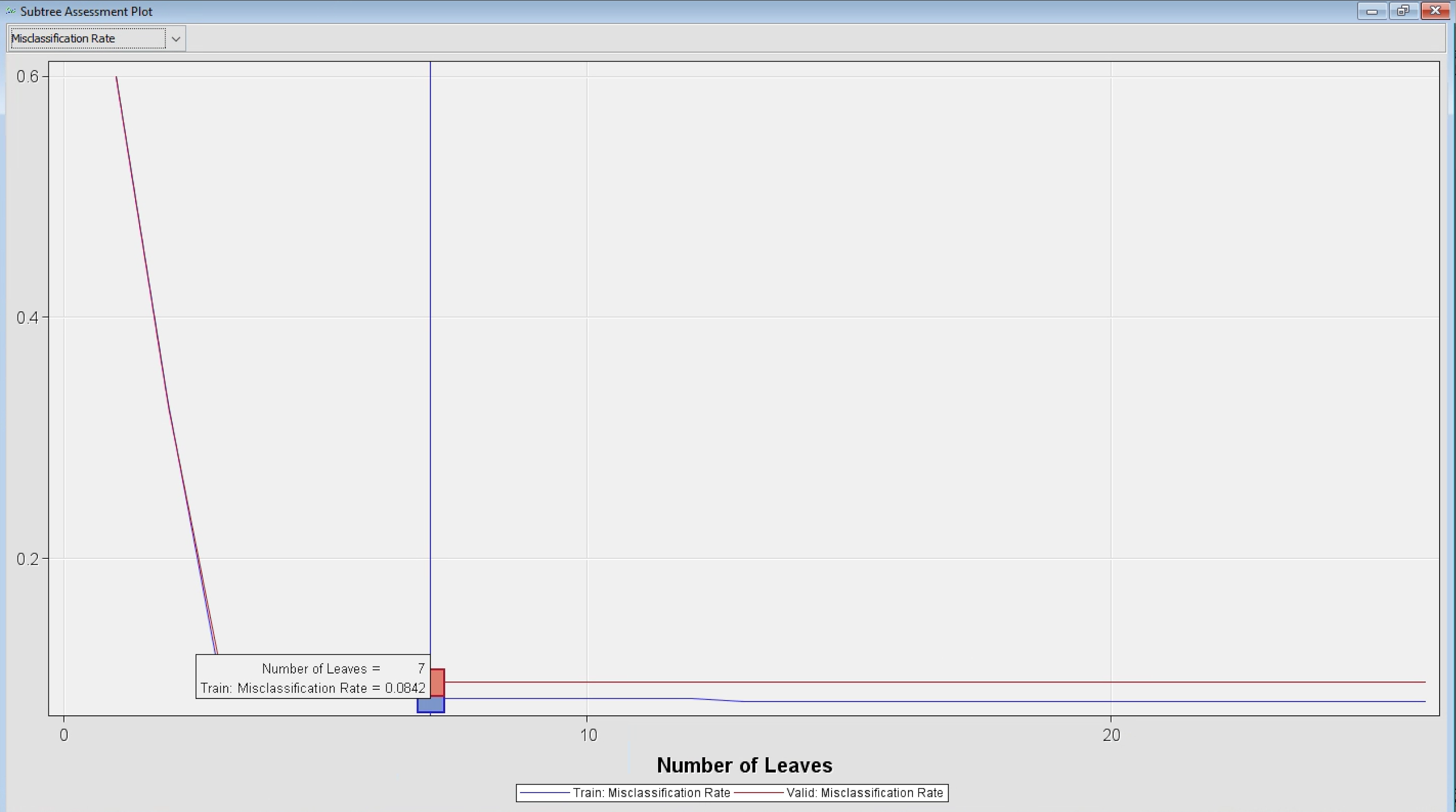




We also measured the number of installations with another 2-way decision tree. Apps with a high number of installations dominated the data at 40.07%, while a moderate number of installations were at 30.15% and a low number of installations were at 29.78%. The tree has a total of 7 leaves with a misclassification rate of 0.0842.



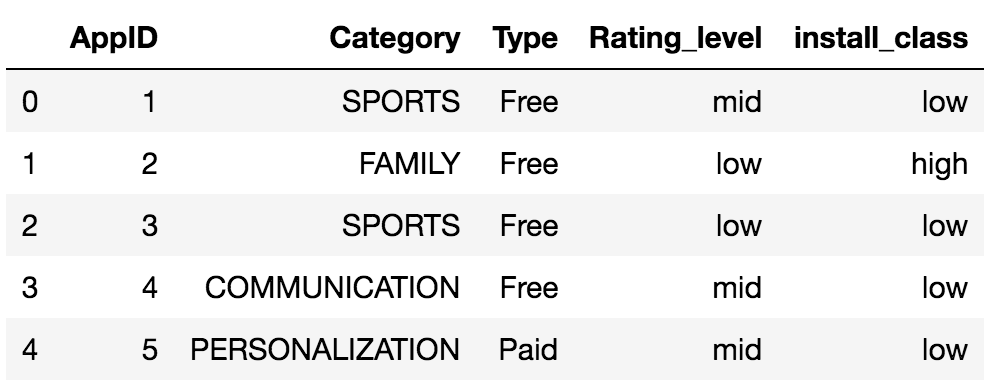




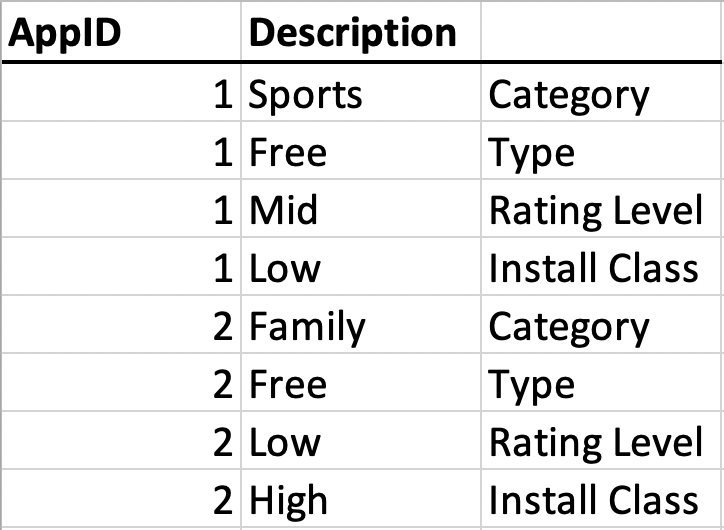


# **Unsupervised Models**

In order to build an unsupervised model, we took the Category, Type, Rating Level, and Install Class and did Data Transposition to fit it into SAS. The first chart shows a sample of the original data, while the second shows the fixed model to run for association.





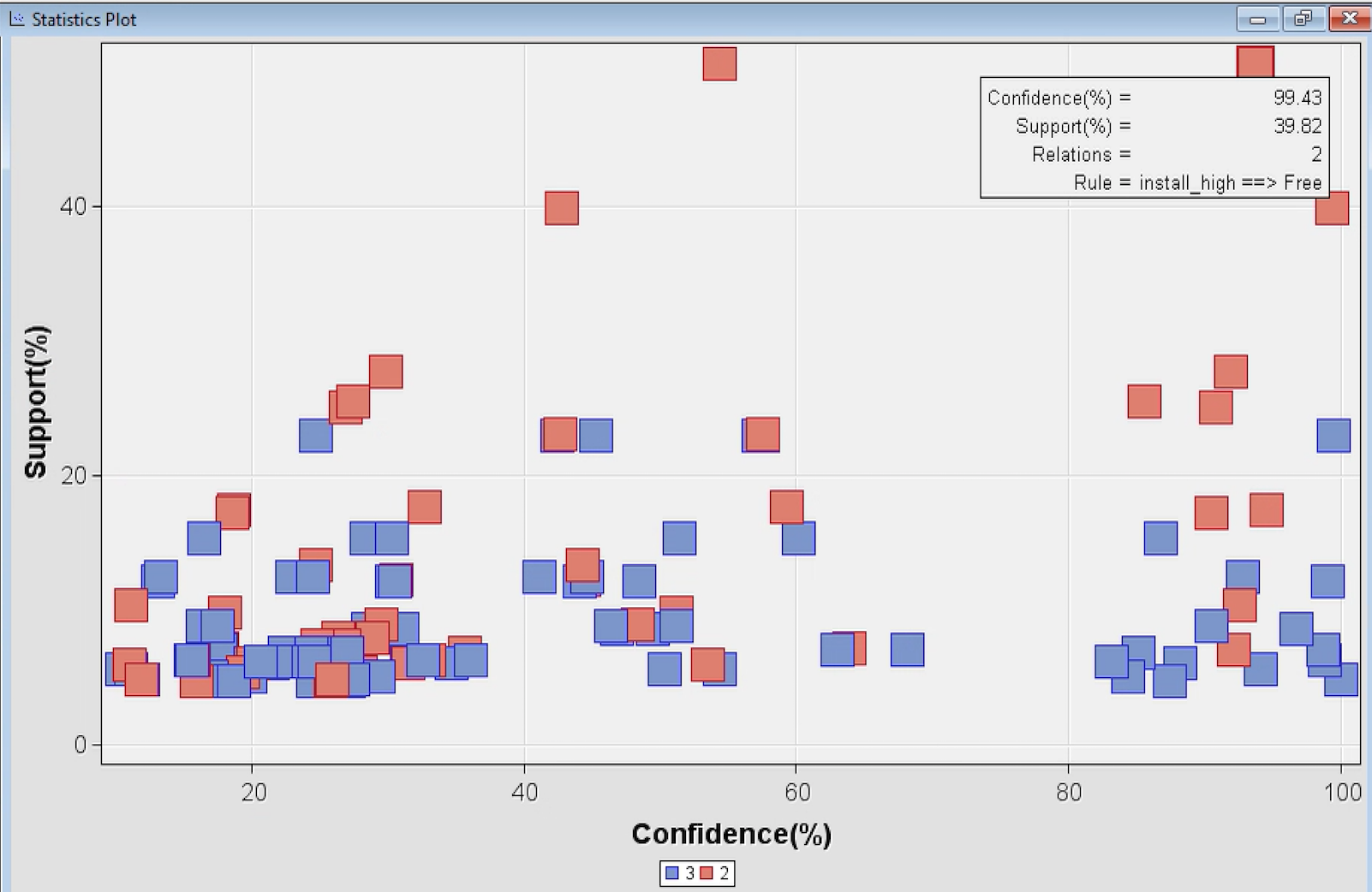




The two charts below show the confidence and support of all possible associations. The red dots towards the top left region of both charts involve the Type column with the “free” variable. In other words, there is a higher lift for some of the app’s factors if they are associated with being free to download.

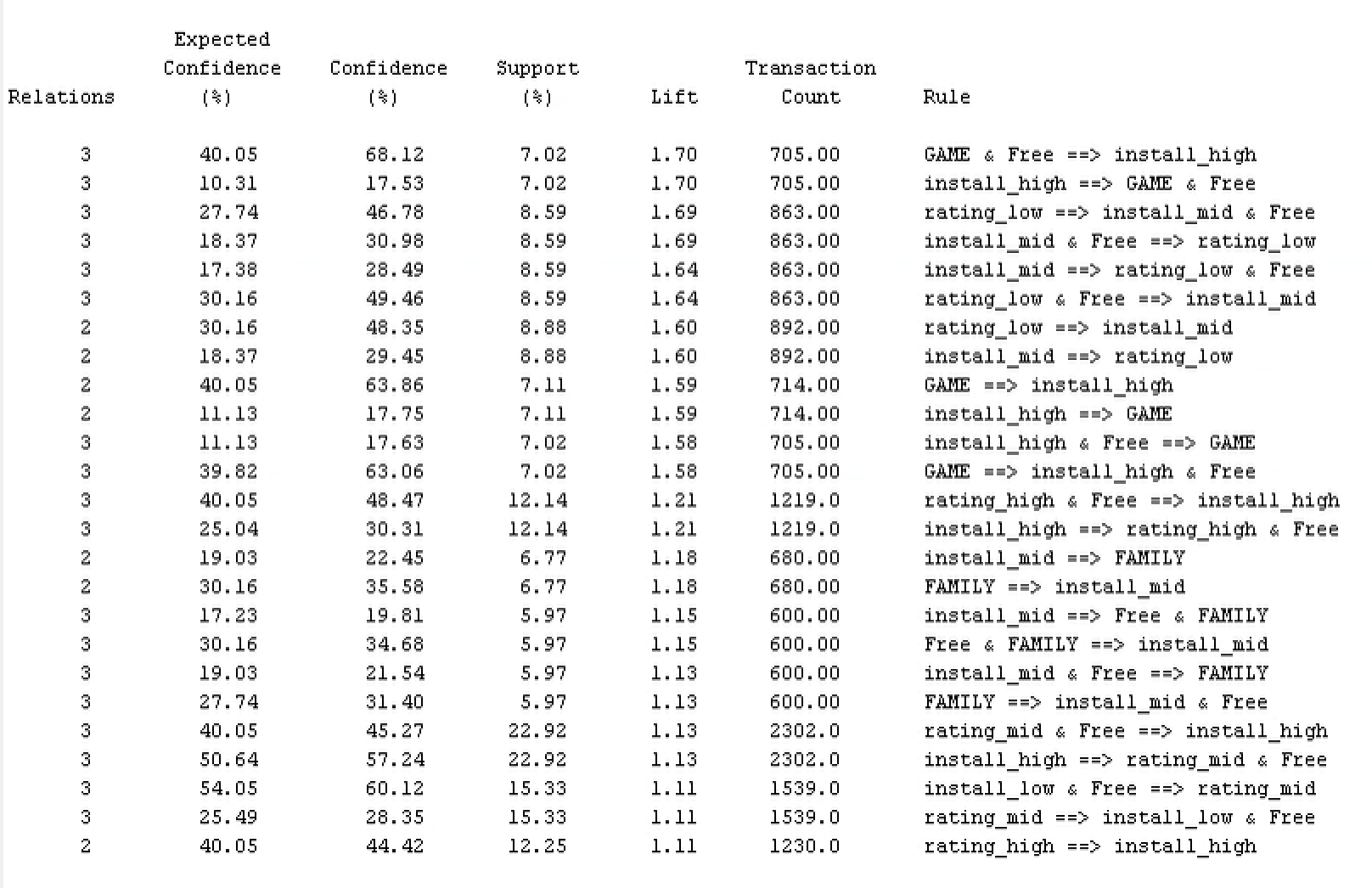








The additional chart below confirms the relations between some of the variables, particularly if the app is free. For example, if a Gaming app is free, it is 1.7 times more likely to have a high number of installations. The reverse of this statement is also true: if an app has a high number of installations, it is 1.7 times more likely to be a free Gaming app. The final example is that if an app is rated poorly, it is 1.69 times more likely to be free app with a moderate number of installations. Among the top rules, we could find that ‘Free’, ‘Game’ and ‘Family’ appear a lot and are associated with more installs. In this way, we could conclude that app users have higher demand on free game apps or free family apps. Although the rating of the app could be influenced by many other factors, developing a free game app or a fee family app has higher possibility of capturing the market.





# **Conclusions**

In our analysis, we found that the number of installations is dependent on ratings. However, the opposite is not true: the ratings of an app is not dependent upon the number of installations. We believe this is because a user are required to download an app onto their phone first before they can rate the app.

The app category makes a difference in whether an app is more likely to be downloaded, rated highly, or not. Gaming and Personalization (for updating your wallpaper, themes, and ringtones) are the most popular categories of highly rated apps. In addition, Communication and Productivity apps are the most popular categories of apps downloaded.

The latest Android version that the app requires for downloading is significant in whether users download an app or not. We believe this is because users are unable to download an updated version of Android on their phone due to owning an older phone or lack of memory space. In addition, some users may not be willing to download an updated version of Android on their phone because of the amount of time it may take. Another reason may be because users are afraid of changes that may occur on their phone from the Android update. Therefore, users may feel it is worth it to forgo downloading a desired app rather than go through the Android software update process.

The more reviews an app has, the more likely it is to be downloaded. This is likely because a user is more likely to use an app that many people have tried. In other words, if an app has only been reviewed by a few people, a user has less information about if an app indeed works. Without a great amount of information, this causes the user to be hesitant in downloading the app. In addition, if there are little reviews, a user may assume that this is because something is wrong with the app. Therefore, they may not download the app at all or choose another app with more reviews.

Regarding the number of reviews an app has, there are a couple of outside statistics that developers should keep in mind. According to an article by Huseyin Y. Yildirimturk on ReputationBuilder.US, customers are about two to three times more likely to write a bad review if they had a negative experience. Nevertheless, according to an article by Caroline Beaton of the New York Times, there are many more positive reviews online than there are negative ones. Therefore, developers should keep their apps at a high quality, satisfactory experience while reviews for the app continue to build.

An app that is free is more likely to be downloaded than an app that costs any amount to use. A significant reason for this is because of the sheer number of free apps available to download on Google Play. Therefore, there may be a high amount of users who may not be willing to pay for an app or put in their credit card information to buy an app when there are other free options. Additionally, most of these free apps have advertisements that pop up during the app’s use. Our conclusion is that many users are more willing to have advertisements present in an app than they are willing to pay to get rid of the advertisements.

# **Bibliography**

Android Developers. “Google Play Store.” *Google Play*, accessed December 8, 2019.<https://developer.android.com/distribute/google-play>.

Beaton, Caroline. “Why You Can’t Really Trust Negative Online Reviews.” *The New York Times*, June 13, 2018.<https://www.nytimes.com/2018/06/13/smarter-living/trust-negative-product-reviews.html>.

Statista. “Google Play Store: Number of Apps 2019.” *Statistica.com*, accessed December 8, 2019.<https://www.statista.com/statistics/266210/number-of-available-applications-in-the-google-play-store/>.

Yildirimturk, Huseyin Y. “Why Are Angry Customers More Likely to Post a Bad Review?.” *Reputation Builder for Small Business*, May 3, 2016.<https://www.reputationbuilder.us/angry-customers-likely-post-bad-review-happy-customers-good-one/>.