Google Play Store Report and Analysis

Group similar apps based on Rating and 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 task I am asked to perform here is: “Group the apps based on Rating between 3.8 and 4.8 and their Category.” In this report I will be discussing the data, methods used, analysis performed, and the overall results.

# 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.

The data has 34 unique Category values: 'ART\_AND\_DESIGN', ‘AUTO\_AND\_VEHICLES', 'BEAUTY', 'BOOKS\_AND\_REFERENCE', 'BUSINESS', 'COMICS', 'COMMUNICATION', 'DATING', 'EDUCATION', 'ENTERTAINMENT', 'EVENTS', 'FINANCE', 'FOOD\_AND\_DRINK', 'HEALTH\_AND\_FITNESS', 'HOUSE\_AND\_HOME', 'LIBRARIES\_AND\_DEMO', 'LIFESTYLE', 'GAME', 'FAMILY', 'MEDICAL', 'SOCIAL', 'SHOPPING', 'PHOTOGRAPHY', 'SPORTS', 'TRAVEL\_AND\_LOCAL', 'TOOLS', 'PERSONALIZATION', 'PRODUCTIVITY', 'PARENTING', 'WEATHER', 'VIDEO\_PLAYERS', 'NEWS\_AND\_MAGAZINES', 'MAPS\_AND\_NAVIGATION', '1.9'.

Closer inspection showed that the record with 1.9 as Category is corrupt. Since no Category is given, I removed this record. This leaves us with 10,840 apps and 33 unique Categories.

There are no rows without value in the Category column. There are 1,474 rows with no value in the Rating column. I decided to exclude these from my report.

The Rating column has values ranging from 1.0 to 5.0. The user wants a report for the apps with Ratings from 3.8 to 4.8. I interpreted this inclusive. I filtered for these rows and that left me with 7,696 apps.

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

# Method

Because I have a categorical feature and a numerical feature, many clustering methods do not work, because they are designed for numerical data. Converting categorical data into numerical data, using the pd.get\_dummies() method was considered, but I decided to use K-prototypes Clustering for Mixed Categorical and Numerical data, since this is specially designed to work with mixed data.

# Analysis & Results

Running K-prototypes Clustering method results in 8 clusters with the distribution graph shown in figure 1. The graph shows that although some clusters are considerably larger than others, all clusters have a non-trivial number of apps.

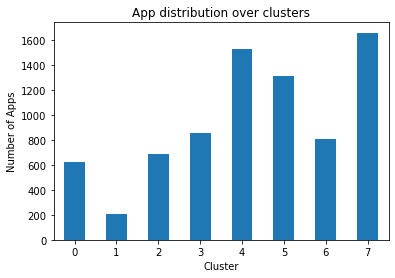
Looking closer at the clusters in figure 2 we see that all clusters except for 2, have apps from almost all 33 Categories. Looking at the distribution over the Categories within each cluster in figure 3, we see aside from the two clusters with only 1 Category in them, no Category or group of Categories clearly dominates a cluster. I ran the K-prototypes Clustering method several times and each time there are one or two clusters with only one Category in them and each time the respective Categories are different. In this case it is HEALTH\_AND\_FITNESS and FAMILY.

Looking closer at the individual Categories in figure 4 we see that most Categories end up in 6 of the 8 clusters. The other Categories end up in 3 or 4 of the clusters. Looking at the distribution of the clusters within each Category in figure 5, we see that only the before mentioned Categories, HEALTH\_AND\_FITNESS and FAMILY, as well as the Categories BOOKS\_AND\_REFERENCE, GAME, PHOTOGRAPHY, and TOOLS show a tendency to be grouped largely in one cluster, but not exclusively in one cluster. Again, which Categories these are, is different for each different run of the K-prototypes Clustering method.

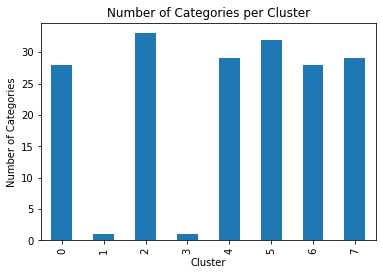
In figure 6 we see an overview of the Ratings and Categories per cluster. It shows that two clusters show only a single Category with a relatively wide range of Ratings. The other clusters are mostly comprised of apps with one, or two adjacent Ratings plus a apps with a wider range of Ratings from one single Category. This is the pattern seen through several different runs of the K-prototypes Clustering method. However, the Categories in the single Category cluster or the Categories grouped with a specific Rating are different each time.

# Conclusion

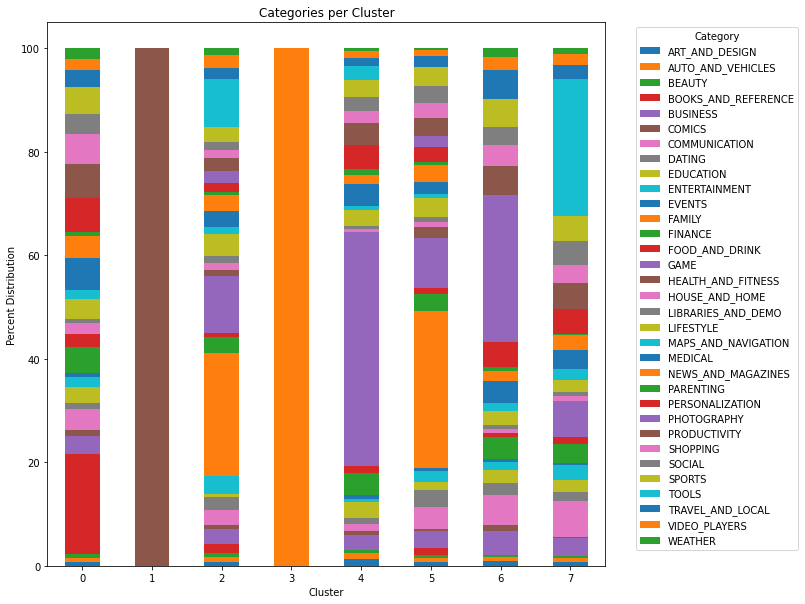
The task I set out to accomplish was to “Group the apps based on Rating between 3.8 and 4.8 and their Category.” Running the K-prototypes Clustering method several times leads to the conclusion that there is no meaningful grouping to be made of the apps with that Rating range, based on their Ratings and their Categories.



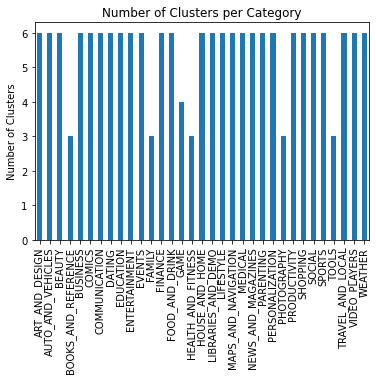
Figure



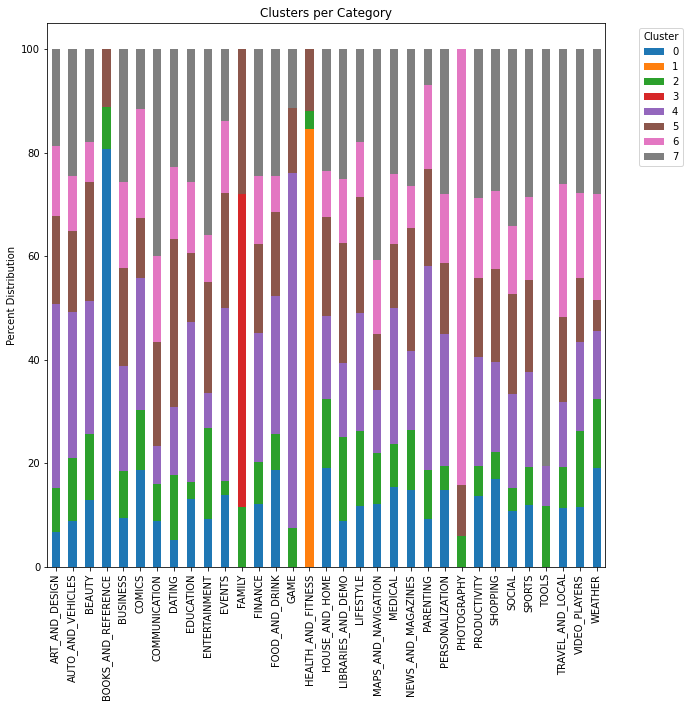
Figure



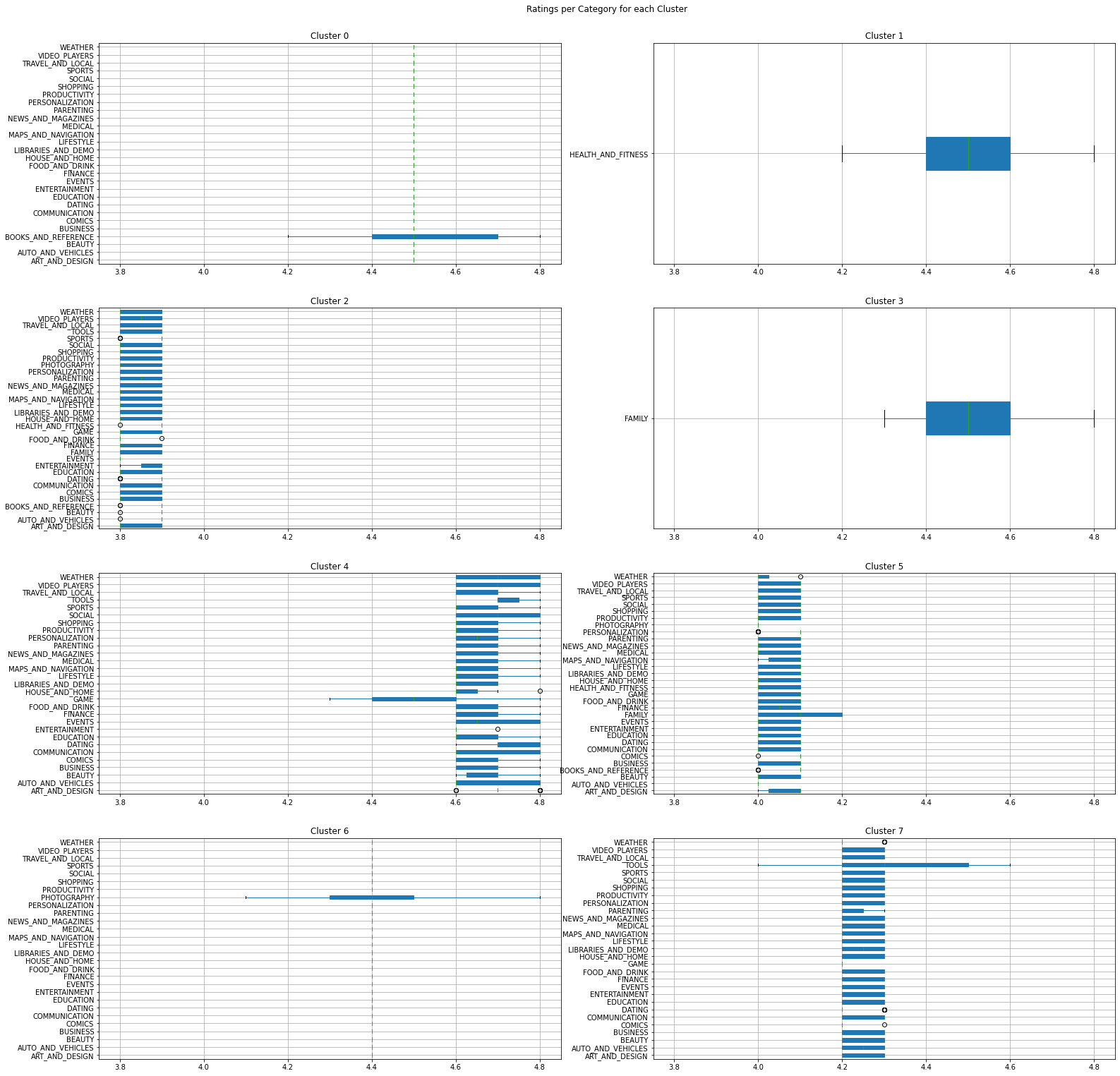
Figure



Figure



Figure



Figure