Predict Profitable App Categories from Google Play Store Data

Rayhan Hossain , CS MS Student
Course Instructor: Dr. Michela Taufer

Motivation

For Android App development based startups, it is essential to explore the App market and identify the profitable app categories. To help these startups at the very beginning, we tried to answer the following questions-

- Which App categories have the maximum app downloads?
- Which one might be profitable for business, free apps or paid apps?
- Which age group of users should be the target of the business?
- How many different Android versions should be supported?

Google Play Data Set

- Data is collected from Kaggle.com *https://www.kaggle.com/lava18/google-play-store-apps*
- 10,841 13-dimensional data points
- 5 Numerical and 8 Categorical data columns

DATA COLUMNS
- App
- Category
- Rating
- Reviews
- Size
- Installs
- Type
- Price
- Content Rating
- Genres
- Last Updated
- Current Version
- Android Version

Data Analysis Workflow

Tools and Algorithms

- Python with Jupyter Notebook
- PySpark, a Python API for Spark
- Pandas
- Matplotlib for Plotting
- Seaborn, a Python Data Visualization Library

- Map Reduce
- Counting and Plotting
- K-means Algorithm
- Correlation of data and Heatmap

Raw Data

Raw Data

Preprocessing
- Data are fed into Spark's RDD in order to make our project scalable
- Replaces the missing values with NP Nan
- Filters out the irrelevant columns
- Handles unexpected data format

Preprocessed Data

Clean data with no unexpected or unformatted values in RDD format

Map-Reduce
- Map-Reduced is applied on the preprocessed data
- Mapped data are reduced by different key values for different analysis, like- Category, Type, Content Rating, and Android Version

Data Analysis and Visualization
- Number of total App downloads are counted against several categorical columns, and plotted with bar and pie diagrams
- Correlation among the Numerical columns are calculated
- K-Means algorithm applied to cluster and see the relationships of numerical data columns- Reviews, Rating, Installs, Price, and App size

Result and Findings

Relationships among data Columns

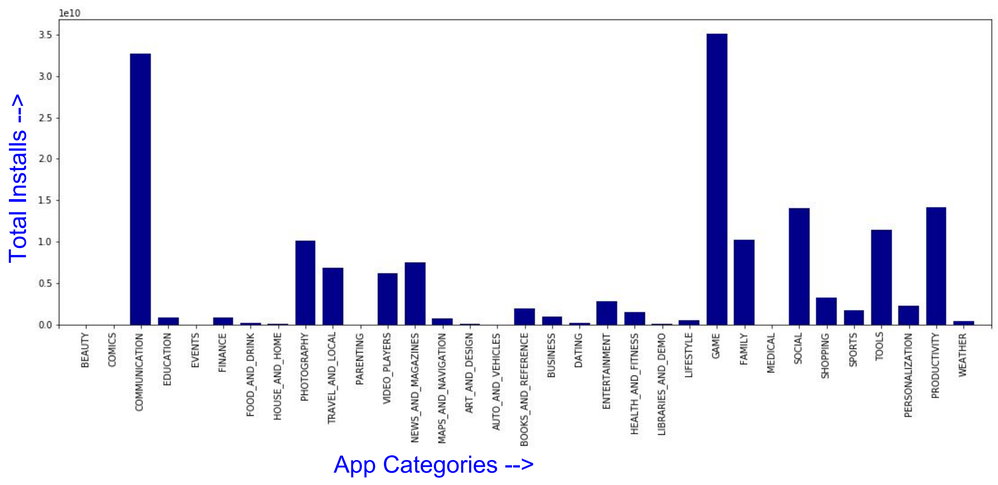

Fig-1:App Category vs Total Installs

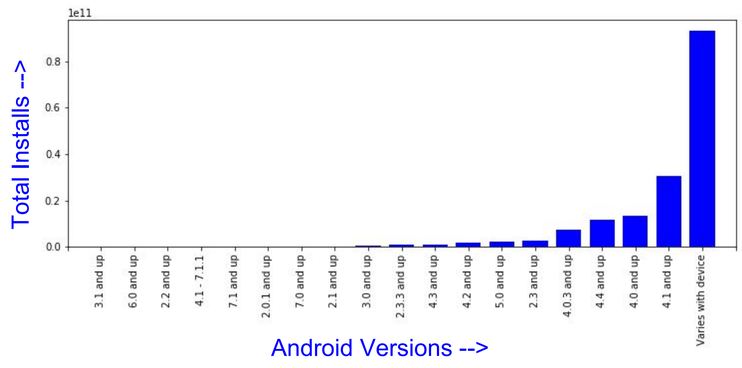

Fig-2: Android Version vs Total Installs

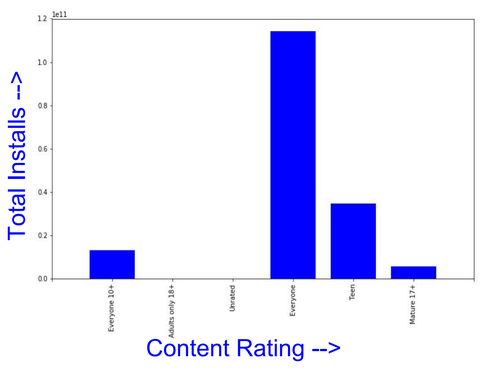

Fig-3: Content Rating vs Total Installs

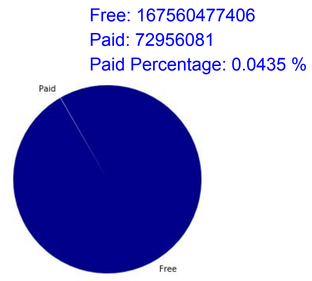

Fig-4: Free vs Paid Installs

Fig-5: Correlation Heatmap

Four numerical columns are considered for clustering- Rating, Reviews, App Size, and Total Installs.

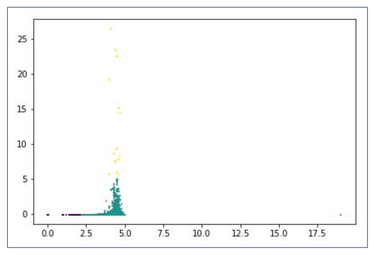

Fig-6: K-Means with K = 3

Result Analysis

Here, we consider the total number of installs as the prime factor to be a profitable application. The number of total installs is strongly related with the categorical data columns rather than numeric columns. After counting and plotting, we can predict the followings.

- From Fig-1, Game, Communication, Social, Photography, and Tools are the most popular App categories
- From Fig-2, Supporting the devices with Android version 4.0 and up might maximize the possibility of downloads
- From Fig-3, Targeting users from all ages or Teenage is the best option
- From Fig-4, Possibility of gaining massive success with paid App is very low
- From Fig-5 and Fig-6, Numerical data columns are not strongly correlated. The K-Means algorithm is applied after doing the data standardization

Future Work

- Analyze the local App Market
- Build a tool to provide localized App market suggestions

THE UNIVERSITY OF TENNESSEE KNOXVILLE