Data 220 Lab3 Report

Zhaohua Huang
Teng Gao
09152019

Dataset Overview:

In United States, nearly 81% of adults own a smartphone [(1)]. They can download their Apps either on iTunes or Google Play. For Android planform users, what types of App they will install on their phones? For Android developers, which Android version will be the most cost effective? For this lab assignment, we will analyze the App installation popularities and how Android version support will affect that.

Based on these points, the dataset we are using is **Google Play Store** in 2018, which is chosen from Kaggle (https://www.kaggle.com/lava18/google-play-store-apps). It contains 10842 rows and 13 columns.

The columns of this dataset are:
(1) App-Application name
(2) Category-Category the app belongs to
(3) Rating-Overall user rating of the app (as when scraped)
(4) Reviews-Number of user reviews for the app (as when scraped)
(5) Size-Size of the app (as when scraped)
(6) Installs-Number of user downloads/installs for the app (as when scraped)
(7) Type-Paid or Free
(8) Price-Price of the app (as when scraped)
(9) Content Rating-Age group the app is targeted at - Children / Mature 21+ / Adult
(10) Genres-An app can belong to multiple genres (apart from its main category). For example, a musical family game will belong to Music, Game, Family genres.
(11) Last Updated-Date when the app was last updated on Play Store (as when scraped)
(12) Current Ver-Current version of the app available on Play Store (as when scraped)
(13) Android Ver-Min required Android version (as when scraped)

For our analysis, we chose the columns of 'Category', 'Installs' and 'Android Ver' as this assignment dataset.

Data Cleaning:

We need to clean the data first before doing analysis, making them more portable to be used.
There are two mean problems we need tackle:

1. Clean nan values and wrong formatted data
2. Convert string data to numeric

For 'Installs', we stripped the '+' at the end of the string and eliminated the ',' inside the numbers.

```
googleplay.iloc[:,5]=googleplay['Installs'].map(lambda x: x.rstrip('+')).values
googleplay.iloc[:,5]=googleplay['Installs'].map(lambda x: ''.join(x.split(',')))
```

We also deleted the missing/wrong values by using df. drop ():

```
googleplay=googleplay.drop(googleplay.index[[10471,10472]])
#remove row 10471 due to missing value of Category
#remove row 10472 due to "Free" value in Installs
```

Convert 'Installs' values from strings to editable numeric numbers

```
googleplay['Installs'] = pd.to_numeric(googleplay['Installs']) #change install type to numeric
```

Convert Android Version to numeric

```
googleplay['AndroidV']=googleplay['Android Ver'].str[0] #only look a the big version for this time
googleplay['AndroidV']=pd.to_numeric(googleplay['AndroidV'],errors='coerce') #change the Reviews column to float an
```

Hope to Discover:

By analyzing the data for Google Play Store, we are hoping to get some insights and possible solutions from two sides:

(1) Google Play App installation popularity in different categories

(2) Google Play App installation under different Android version supports

We speculate that Android version support may affect the App installation. Android developer may need to consider this factor in order to achieve a balance between App popularity and cost effectiveness.

Approach of Analysis:

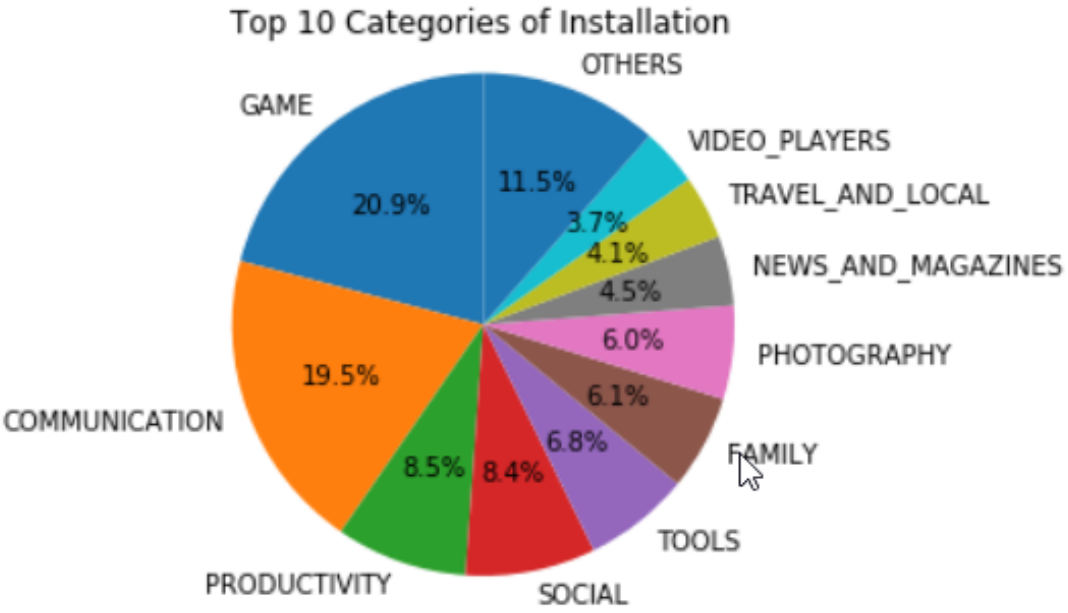
'Category' vs 'Installs':

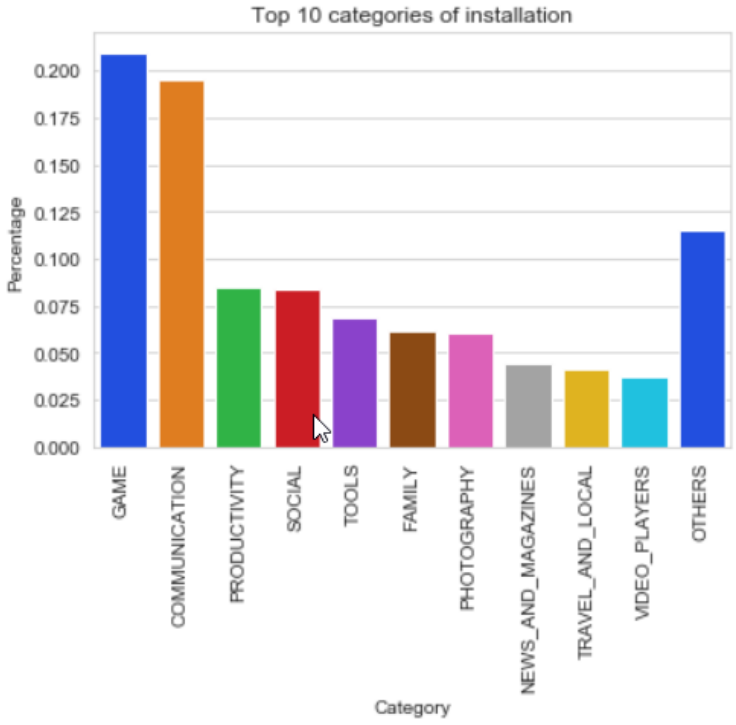
First, we use **zip** function to generate a list of installations in each category, and list of percentage installations

By printing out the results, there are 34 different categories, some of them with very low installation rate. So, we decided to list top 10 installation rate of categories and make rest to be 'OTHERS' under 'Category'.

Then we use **zip** function again, zip up category_name_list, installs_in_category and percentage_in_category. We are using **sorted**, **slicing** methods to generate the target lists.

At last, we use pie chart and bar chart to show our analysis results on Top 10 categories for installs:



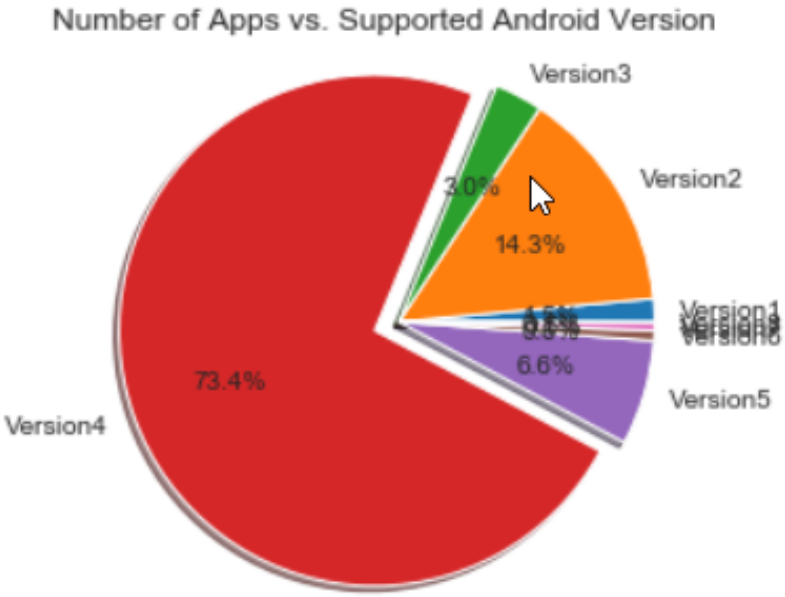


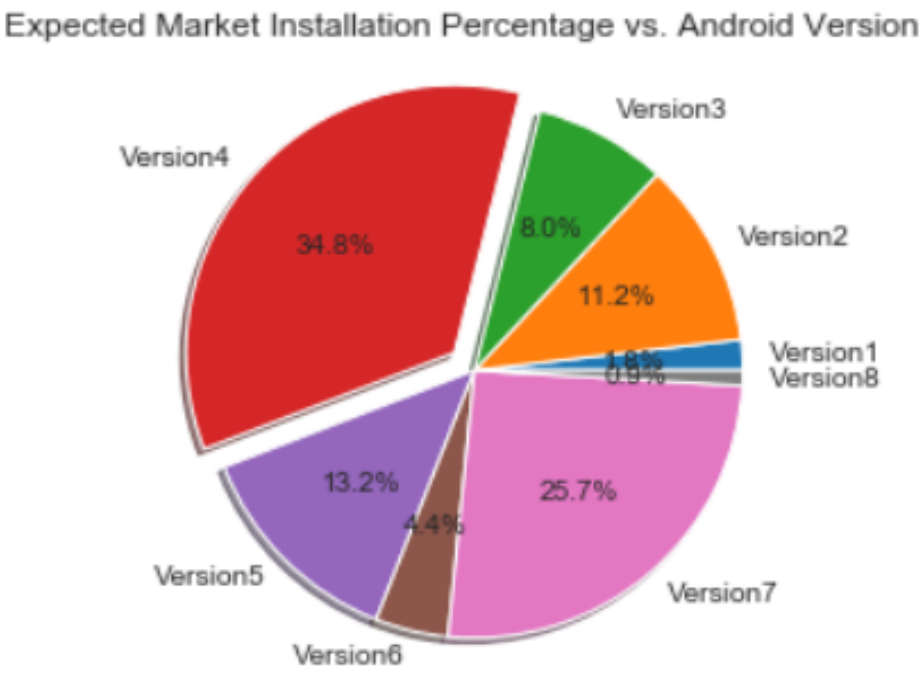
From these two graphs, we can clearly see the difference in installation for different categories.

Android Version:

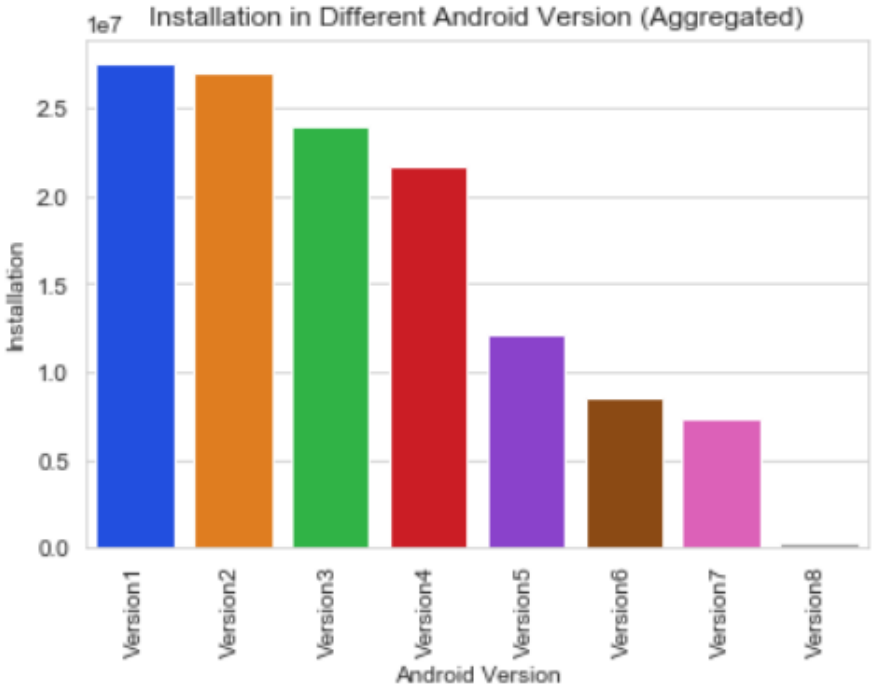
We made two pie charts for Android Version vs. Installation.

One is number of Apps in different Android versions. The other one is the average installations one app in different Android versions. We call that as the expected installation for the app in different Android version because it is defined in statistics.





However, these two pie charts only show the relationship among each Android version. We made a bart charts by using aggregated values, which can give us a better view:



This is the aggregated installations under different Android version. Based on the data we have, the legend “version 1” in graph means version 1 and above, so we also included version 2 and above, version 3 ad above and so on. Therefore, it is normal to see the installation is decreasing from version 1 to 8. However, we must notice the reality that some higher Android version will not support low Android version apps. If we will take this into consideration, the graph could be totally different.

Central Tendency and Variability Test:

At last, we did the distribution and variability test to see how the data for installation looks like. Because category can not be sorted, its standard deviation is large. Installation vs. version seems to act like a normal distribution.

Insights and Conclusion:

Games and Communication are two dominant categories in Google App Store. The following categories are Productivity, Tools, Social, Family, and so on. This distribution of Apps, will help businesses to determine their target market.

For this lab, we want to see how many Apps are installed in different Android version. We want to further explore if Android version is a factor that may affect the installation times of an App.

We find that about 73% of Apps are listed as Android version 4 and above. We also made a chart and a bar chart for average installation in different versions. If we can get the data from the companies to see the developing cost for each version, we may find the most cost-effective way to develop an App (only under the Android version factor).