Homework 5
CIS 192

Instructor: Jorge Mendez

Due: March 20th 2020

1 Data anaylsis and visualization

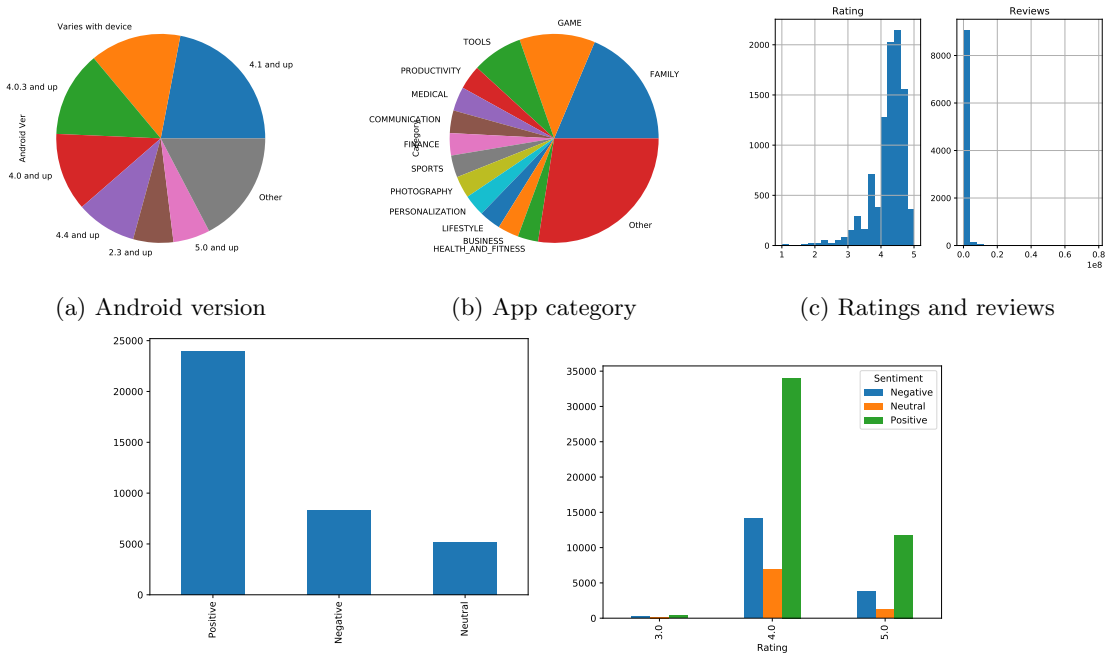
Python with Pandas is a tremendously powerful tool for data analysis and visualization. In this homework, you will use Pandas to analyze the Google Play Store data set, available at `https://www.kaggle.com/lava18/google-play-store-apps/download`. We will ask you to manipulate the data in certain ways in order to show different interesting visualizations. Specifically, you will:

1. Load the `googleplaystore_user_reviews.csv` (reviews) and `googleplaystore.csv` (apps) data sets as pandas DataFrames. You should use the `pd.read_csv()` function for this purpose.
2. Delete any review that does not contain either a `Translated_Review` or a `Sentiment`. The `pd.dropna` function will be helpful for this.
3. Remove any apps whose `Rating` is invalid (> 5).
4. Produce a pie chart with the `Android Ver` requirements for the different apps. Group together all versions that make up less than 5% of the total apps into a single 'Others' category. This should look similar to Figure 1a. You will find the `df.value_counts()` function useful for solving this problem.
5. Create a similar pie chart for app `Category`. In this case, group together categories that make up less than 3% of the apps. The resulting graph should look something like Figure 1b.
6. Show histograms of the `Rating` and `Reviews` across all apps, with 20 bins each. Example histograms can be seen in Figure 1c.
7. Plot a bar chart with the different `Sentiment`. The sentiments chart should look similar to Figure 1d.
8. Combine the two DataFrames into a single one, based on the `App` names. You should make sure that all apps from the apps DataFrame are kept, and no app beyond those is added. The `pd.merge` function will be useful for this.
9. Group the `Sentiment` by rounded `Rating`, and produce a bar chart where you display the different sentiments grouped by rating. You might find the `pd.groupby`, `np.round`, and `df.unstack` functions helpful for this task. The grouped sentiments plot should look like Figure 1e.

Requirements:

- Your code must be properly vectorized. *Hint:* no for loops are necessary to solve the assignment.
- You are free to use any NumPy, Pandas, and Matplotlib functions.
- Efficiency matters! Your code will be evaluated for speed.
- Style also matters. You may leverage a style checker (e.g., `https://pypi.org/project/pep8/`) to ensure your code has proper style.

Additional details on the functions required are provided with the skeleton code, available at `https://seas.upenn.edu/~cis192/jorge/hw5/hw5.py`. Feel free to create additional functions as you see fit to modularize your code.



(a) Android version (b) App category (c) Ratings and reviews

(d) Sentiment (e) Sentiment by rating

Figure 1: Figures created with Pandas for this assignment.

2 Testing your code

The easiest way to test your code is to run it on `eniac`. For this, you will write a `main()` function as explained in the skeleton file, copy your file to `eniac`, and run it on the terminal by executing `python3 hw5.py` from the directory where the file is stored. Note that `eniac` contains version 3.6.5 of Python, and not 3.7. If you wish to make use of any functionality introduced as of version 3.7, feel free to install Python on your own machine. There are multiple resources online describing how to do this.

3 Submission instructions

Submit a single file named `hw5.py` to Canvas. This file must contain implementations for all functions given in the skeleton file provided to you with the same name. Make sure to add your name and PennKey, as well as the number of hours you spent working on this assignment, at the top of the file where indicated.