Declaration on Plagiarism

Name:	Anshika Sharma
Student Number:	19210993
Programme:	MSc. In Computing, Data Analytics
Module Code:	CA682
Assignment Title:	Data Visualisation
Submission Date:	16 Dec 2019
Module Coordinator:	Dr Suzanne Little

I declare that this material, which I now submit for assessment, is entirely my own work and has not been taken from the work of others, save and to the extent that such work has been cited and acknowledged within the text of my work. I understand that plagiarism, collusion, and copying are grave and serious offences in the university and accept the penalties that would be imposed should I engage in plagiarism, collusion or copying. I have read and understood the Assignment Regulations. I have identified and included the source of all facts, ideas, opinions, and viewpoints of others in the assignment references. Direct quotations from books, journal articles, internet sources, module text, or any other source whatsoever are acknowledged and the source cited are identified in the assignment references. This assignment, or any part of it, has not been previously submitted by me or any other person for assessment on this or any other course of study.

I have read and understood the referencing guidelines found at http://www.dcu.ie/info/regulations/plagiarism.shtml, https://www4.dcu.ie/students/az/plagiarism and/or recommended in the assignment guidelines

Name: Anshika Sharma Date: 13th December 2019

Analysis of Google Play Store applications based on their Category and other characteristics!

As we are moving into the technological era, everything is becoming accessible through the internet via various mobile applications, from paying bills through apps to reminding yourself to drink water. E-commerce and review sites are full of untapped data with prominent potential to transform into meaningful insig hts that can help to make robust decisions. Google play store, is a pre-installed application that enables us to download different applications, books, movies, and the rest, on the mobile device. It's easy to create mobile applications and it is worthwhile. There are more and more applications being created because of these two factors. In this task, by analyzing over ten thousand applications in Google Play across different categories, I conducted a comprehensive analysis of the Android app industry. The purpose of this project is to analyze detailed information on apps in the Google Play Store in order to provide insights on app features. Through these visualizations, you'll be able to comprehend how various categories are performing and contributing in the android market. In the end, you'll be able to clearly understand how each category contrasts from one another based on several factors.

Dataset

The google play store data was retrieved through web scrapping. The dataset is available on Kaggke.com. There were two csv files one with 10842 rows and 13 columns and the other with 37482 rows and 5 columns. Only one of the files were used for the analysis; the first one. Numerous applications are developed on a daily basis, which states that it represents the velocity of Big Data. The fields present in the dataset are:

S. No.	Columns	Datatype	Description
1	App	String	Name of the applications
2	Category	Decimal	Category in which the app belongs to
3	Rating	Integer	Overall user rating of the app (as when scraped)
4	Reviews	Decimal	Number of user reviews for the app (as when scraped)
5	Size	String	Size of the app (as when scraped)
6	Installs	String	Number of user downloads/installs for the app (as when scraped)
7	Type	String	Free/ Paid
8	Price	String	Price of the app (as when scraped)
9	Content Rating	String	Age group the app is targeted at - Children / Mature 21+ / Adult
10	Genres	String	An app can belong to multiple genres (apart from its main category). For eg, a musical family game will belong to Music, Game, Family genres.
11	Last Updated	Date	Date when the app was last updated on Play Store (as when scraped)
12	Current Ver	String	Current version of the app available on Play Store (as when scraped)
13	Android Ver	String	Min required Android version (as when scraped)

2. Data Exploration, Processing, and Cleaning

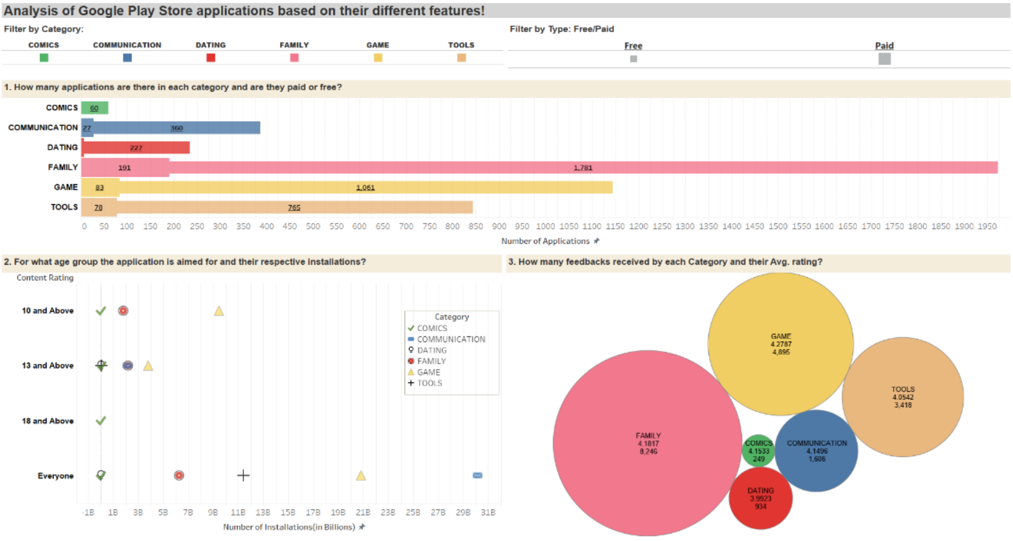
The dataset was a little bit messy, I used a data cleaning tool OpenRefine to clean the data. The column wise changes made in the dataset are:

- **Category**: Remove a numeric category as 1.9.

- **Ratings**: There were a lot of nan; null values, in the data so they were replaced with the average of rest of the ratings in the dataset; 4.1.
- **Size:** This field had the size of the apps in three dimensions; namely Mega Bytes, Kilo Bytes, and Varies with device All the records with size in kbs were converted to a common dimension mbs. *(The cleaning of this field is demonstrated in the video.)*
- **Installs**: + Sign was removed from all the values in this column.
- **Genres**: " ; " replaced by " : "
- **Type**: Removed Null values.
- **Price**: The " $ " was removed, and the name of the column was changed to *Price in $.*
- **Content** Rating: This column had 6 to 7 categories which were then short listed to 10 and Above, 13 and Above, 18 and Above, Everyone and Unrated.
- 3 columns were removed: Last Updated, Android Ver, and Current Ver.

The columns that I've worked on are: App, Category, Rating, Reviews, Installs, Type, Size and, Content Rating. I chose these fields because each of them contributes in creating great impact to analyse a particular application or category. For example, to decide whether an application is befitting or not, one might look at the configurations like what is it's rating, how many people have rated the application, is it paid or free, how much space is it going to occupy, is it relevant according to one's age or not. On the basis of these figures we can make decisions accordingly

3. **Visualisation**



Analysis of Google Play Store applications based on their different features!

I've created a Tableau dashboard which consists of:

1) Two Filters:
   a) **Category Wise**: Comics, Communication, Dating, Family, Game and Tools: This highlights anf filters each category in all three graphs.
   b) **Type- Paid/Free**: Only highlights the bars of the first graph. The thick part if it's paid a category and the thin part if it's free.

2) **Horizontal Bar Graph with varying size of the bars: How many applications are there in each category and are they paid or free?**

I wanted to show the number of applications in each category along with the count of applications denoting paid/ free. The y axis represents Category whereas x axis gives the count of applications. The varying size of the bars makes it easier to understand visually and differentiate all the categories on the basis of count of free and paid applications in all categories respectively. The best match for this would be a bar graph. The 6 categories are chosen randomly to show the versatile nature of the data.

3) **Bar Graph with shapes representing each category: For what age group the application is aimed for and their respective installations?**

In this graph, the total number of installs in each category are displayed and in what category they fall in. On y axis content rating is displayed and total number of installations in every category on the x axis. This graph helps us understand that which category in spite of high or low number of installations is available for what age group. Shapes added to the graph with the matching colour coding makes it less complicated to decode the information.

4) **Scatter Plot with no axes: How many feedbacks received by each Category and their Avg. rating?**

This graph represents the total number of feedbacks received by the users and how does the ratings vary with the number of reviews received. Even though the average ratings may be similar but the size of the circles makes it visually possible to male judgements on the comparison of the categories.

Use of colour: A white background is used to make the visualizations stand out. The headings are given light background colour to make it subtly stand out. The categories have a specific colour, Dating-red, because it somewhere represents love, Comics-green because they are groovy, Communication-blue because one of the most used application in communication category is blue(Facebook), Family-pink, pink also represents love, Game-yellow as games are usually funky, and Tools- brown as brown is a fitting colour for a toolkit.

Shapes and Marks: In 2nd graph, the visualization depends entirely on the shape and colour assigned to each category. I've tried to maintain the colour coding throughout the dashboard. Each symbol either represent the category or the colour assigned to it. For example, communication's symbol is a blue coloured envelope which matches both the meaning and colour of the category, where as Game is assigned a yellow triangle, the triangle can be referred as a play button but not exactly so in spite of any random triangle a yellow coloured triangle is usd to match the colour of Game.

Layout and Structure: The graphs are organized in a way that all three graphs have equal impact and contribution in the dashboard.

Font and Labels: A common theme has been followed in the dashboard, however, some of the specific figures have comparative big size to make them stand out in the visualization.

Interactivity: All the graphs in the dashboard follow the same pattern. All the analysis of applications is done based of different categories. The interactivity makes it easier to have a closer look at each category individually. Either you can hightlight a particular category or select it to see only that particular category.

Tools: Open refine for cleaning the data and Tableau for making the visualizations are the two tools that were used in this process.

4. Conclusion [½ page]

Through this visualization we can clearly understand that:

- Family has the maximum number of applications as well as reviews.
- Comics has only Free applications.
- Dating had negligible amount of paid applications.

- Tools is an average performing Category.
- All categories have some applications for Everyone.
- All the paid applications have installation below 1 Billion.

Criticism:
Rating of this visualization would be above average, approximately around 4.4.
There were a few of aspects that could have been improved:

- The graph choices could have been better, regarding the bar graph usage redundancy in two graphs.
- In graph 2, symbols representing their respective categories are not that relevant. The values between 0-1 Billion, are all over lapping each other which makes it more difficult to understand the message.
- Because of few paid applications in Dating, the visualization fails to state it clearly
- In graph 3, Comics' circle is so small and is unable to keep the labels inside the circle, which makes the graph look untidy.
- Categories should have been short listed on the basis of some agenda, and not randomly.

The watermark in the video spoils the entire video of visualization.

As mentioned above, even though Tableau is one of the best visualization tools, it has its own limitations. Depending on the fields selected; measures and dimensions, only limited graph choices are available. The symbols used in graph 3, Tableau had very few options available.

References:

https://www.kaggle.com/reginashay/data-pre-processing-and-feature-selection-using-rf
https://www.kaggle.com/namandave/kernel79b479f654
https://public.tableau.com/en-gb/s/resources
https://public.tableau.com/profile/ken.flerlage#!/vizhome/SampleDashboardActions/Filter