INTRODUCTION:

The data available for analysis comprises of reviews that is extracted from the Appstore. The reviews are collected for the apps each week and comprises of 2 files. The review file holds one line for each review for every app, the date review was posted, the score and comments/reviews. The detailed data file contains the aggregated information of the app and includes additional information like the genre of the app, the content rating group, number of installs until that date etc.

The title of the review file contained the date those reviews were collected and the category the apps belonged to. Hence, string operation was implemented on the title to extract and input the category and date as a column to the data. Due to many files, the package "glob" was used to read the files in the folder and concatenate them into a single data frame for processing in Python.

ANALYSIS & FINDINGS:

As a first step, the total number of records that were available were obtained, which was 2715303. There were around 51013 duplicate rows in the data, which were all dropped, and 2664290 unique reviews were retained. The total number of apps that are available in the data was 87 with 8 different categories. The number of apps under each category are as follows:

Education	10
Entertainment	14
Family	10
Finance	10
Game Action	11
Health and Fitness	10
Lifestyle	12
Music and Audio	11

Entertainment tops the entire list might be since people seek mobiles predominantly for entertainment.

The sum of the number of apps listed in each category is 88 did not reconcile with the total unique apps available in this data, 87 due to the fact, the app “Duolingo: Learn Language Free” was listed in two of the categories viz., Family and Education. During analysis, the number of reviews that were present in each of the app category was found, out of which, app category Health and Fitness has a greater number of reviews with 355183 reviews, followed by the category Finance with 352979. It is good to note that, though, Entertainment category had the greatest number of apps they seem to be the third in the number of reviews with 341127. The overall content rating group are classified as Everyone, Everyone 10+, Mature 17+, Teen. It is interesting to note that, apps in the Education, Finance and Health and Fitness category are accessible by everyone whereas there are certain apps in the Game Action category that can only be accessed by Mature 17+ or Teen content rating groups. Under the Entertainment app category which accounted for the most number of apps, showed that most of these apps are for “Teen” content rating group followed by the entertainment. The Game Action category almost has similar number of apps on offer for Everyone and Teen rating groups. Most of the apps in the Music and Audio category are for the Teen content rating group. This shows that most of the apps are accessible by Everyone after which most of the apps are developed for the Teens. To perform proper contentRating analysis, some of the columns from the detailed data file was imported to the review data file in order to analyze the number of reviews that were available for each of the content rating group in each app category. In order to collect this data, all the records in the details data file were concatenated and defined in a dataframe with the app title and the corresponding content rating group. Further analysis revealed that the app Unicorn Slime Maker and Simulator was listed in two of the content rating group Everyone and Teen. Once the data frame was ready, it was observed that the app title imported from the detailed data file had special characters in them, which had to be handled, to avoid erroneous analysis. Then the data was merged with the data collected from the reviews file, it provided us a

total of 2677056 records as the app Unicorn Slime Maker and Simulator was listed in two content categories. The analysis on the number of reviews for each app category revealed that there were also null values in the app title. Analyzing the Entertainment category which comprised of the maximum number of apps showed that “Teen” content rating group were providing a higher amount of review as they had a greater number of apps in their offering. This proved to be the same with the Game and Action category with “Everyone” content rating group providing more reviews as they had a greater number of apps to offer. An interesting fact to note from the above analysis, app categories with only “Everyone” content rating group tend to receive a greater number of reviews even if the other app categories offer many apps.

TEXT ANALYSIS:

Textual analysis is a method used to describe and interpret the characteristics of a recorded or visual message. The purpose of textual analysis is to describe the content, structure, and functions of the messages contained in texts. The important considerations in textual analysis include selecting the types of texts to be studied, acquiring appropriate texts, and determining which particular approach to employ in analyzing them to provide constructive feedback for any development.

As there are a lot of non-English characters in the text/review, it needs to be removed before performing meaningful text analysis of the data provided. To perform this, `nltk` library with the library `words corpus` was used. It helps in parsing through the text and removes any non-English characters and replaces them with a space. In the similar way, removal of non-ASCII characters, punctuation and repeated characters from the review is necessary as they do not help with a meaningful analysis.

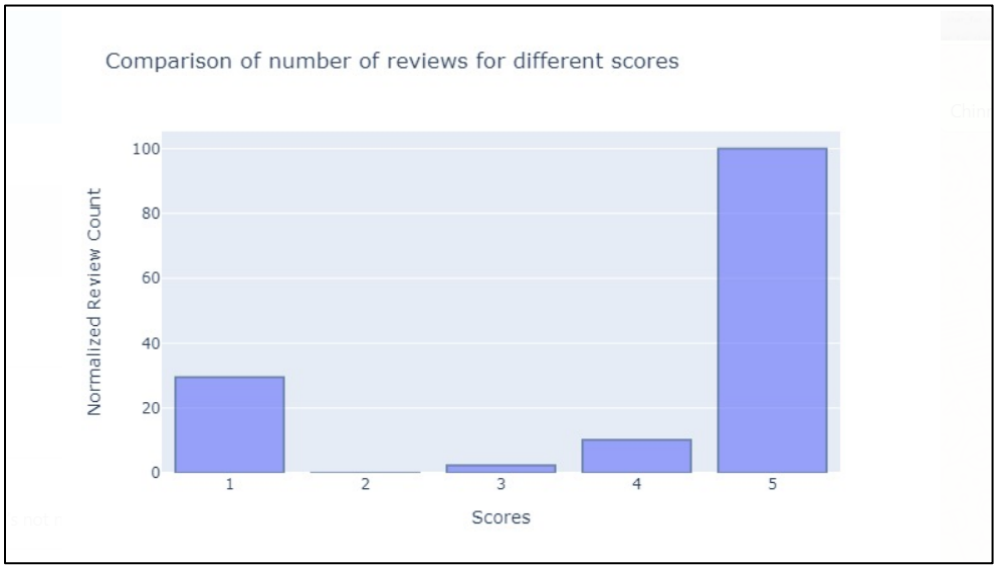
It is a good option to retain most of the data before to perform a complete analysis, however in this situation, we remove the rows that do not contain more than 2 words considering that they fail to deliver a constructive message to the developer for further development of the app. It is recommended that we consider some of

these reviews as they deliver a strong sentiment related to the app. For Eg: two-word reviews like "good, great, bad, cool, very nice, the best, time pass....." strongly help in analyzing the sentiments of the reviews. To perform efficient and faster analysis we add unique key as a column to the data.

Analyzing the number of reviews that is lost due to bad data quality by content rating group, it is noted that 60% of the review by Mature 17+ is lost in the Game and Action category which is followed by “Teen” from the same Category. We can derive that people using the apps in the Game and Action category do not tend to provide a meaningful review most of the situations. The category which has retained most of its reviews are Everyone 10+ in family followed by “Everyone” Health and Fitness which has retained 72% and 71% of the reviews respectively.

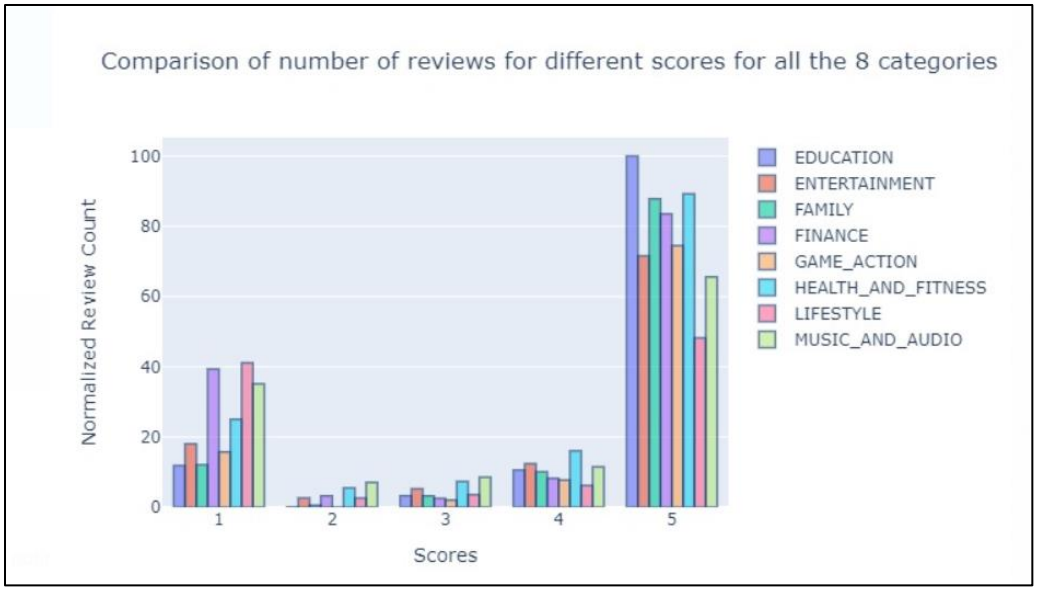
VISUALIZATION:

Plotting the analysis visually provides more clearly understand:

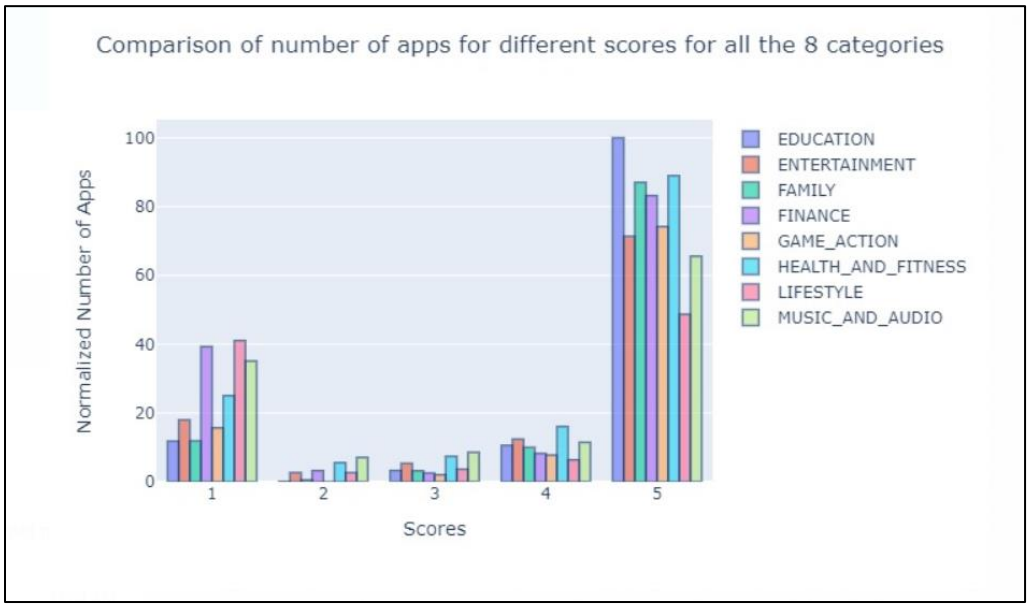


This graph plots the number of reviews plotted against each of the score group. This shows most of the reviews receive a 5-star rating followed by 1 star. This signifies a pattern that people tend to review either when they really like the app or when they totally do not like the app.

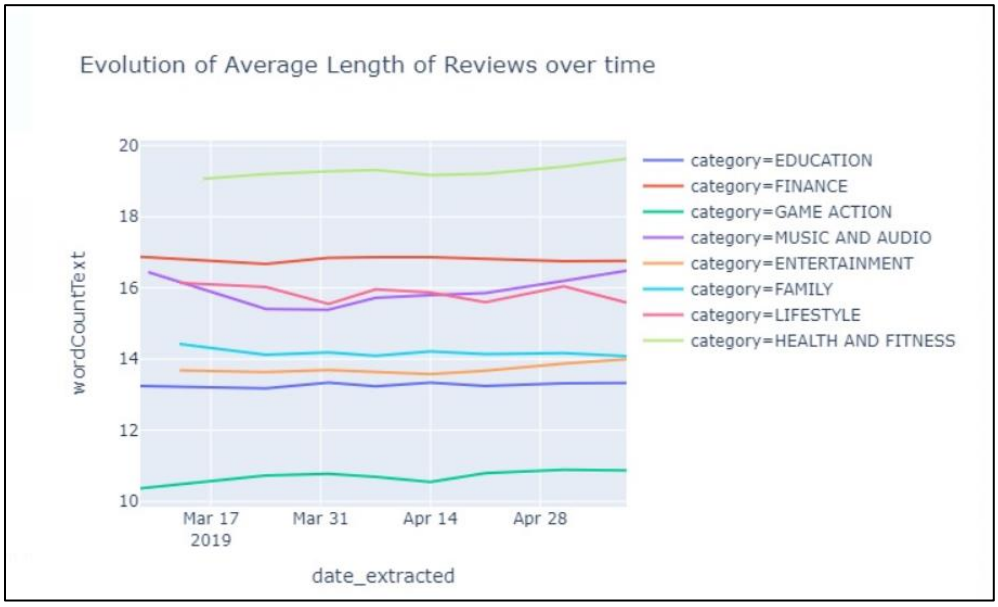
Plotting the length of reviews for each score sub group for each category shows us that Health and Fitness has topped in four score categories. This graph shows us that when people post review with good rating they do not tend to explain but when they do not like the app they post a longer review. We can clearly see a gradual decrease in the length of the review as the score increases.



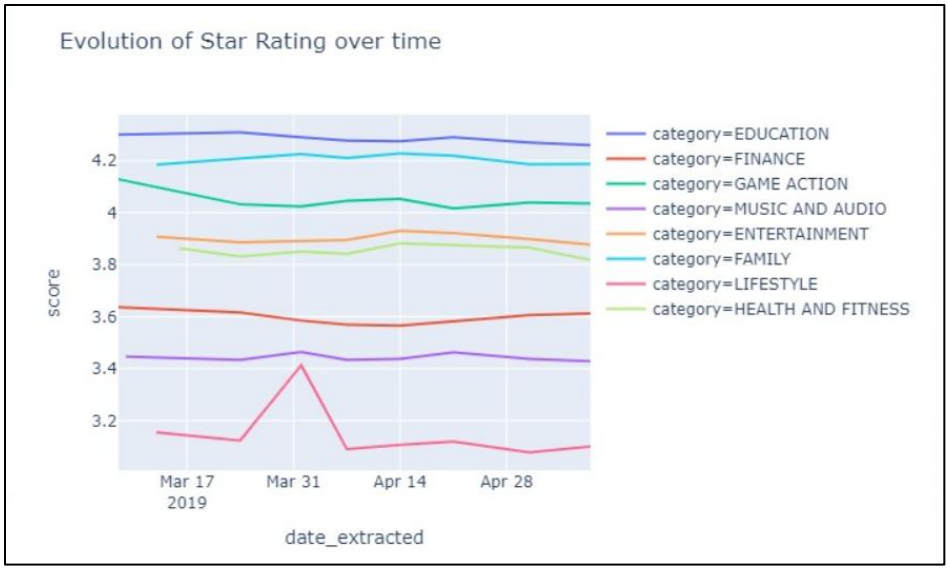
This is a plot portraying the number of reviews for each app category for every sub group. Education app category top the list obtaining the most positive review also it is one amongst the lowest with score rating 1. This shows that the quality of apps in the Education category are good. The Lifestyle app seems to have a lower quality as we see that they top the list for 1 start and have also secured the last spot in the 5-star list.



The count of appTitle and the count of Processed Text have a little difference due to NULL values in the appTitle but as we normalize the data both the graphs look almost similar.



This trend graph plots the evolution of average length of review for each app category. It is noted from this graph that only the category Health and Fitness and Entertainment has seen an increase in the average length of review. The categories Finance, Music and Audio, Education have seen some variations but have remained constant over the time. The Lifestyle category has seen a lot of variations over the period of time and seems to be on the decreasing trend.



This chart shows the evolution of stars over time. Education category has always maintained a higher star rating which relates the number of 5-star review that we have seen in the category. Most of the app seem to have maintained a stable rating over time except for Lifestyle and Game and Action. Lifestyle seems to have had a spike indicating an improvement in the app, but it eventually deteriorates again. The scores for Game and Action are in a decreasing trend.