DATA 542 – FINAL REPORT

Name: Ritayu Nagpal

1. **Summary of Dataset:**

Newest Reviews Files: These files had the actual user reviews on the app and data was collected weekly and we had 8 data files for 8 different categories i.e. 64 files in total.

All Details Files: These files had all the other information related to app like its content rating, number of downloads, ads supported, genre etc. There were 8 weekly files each for 8 categories i.e. 64 files in total.

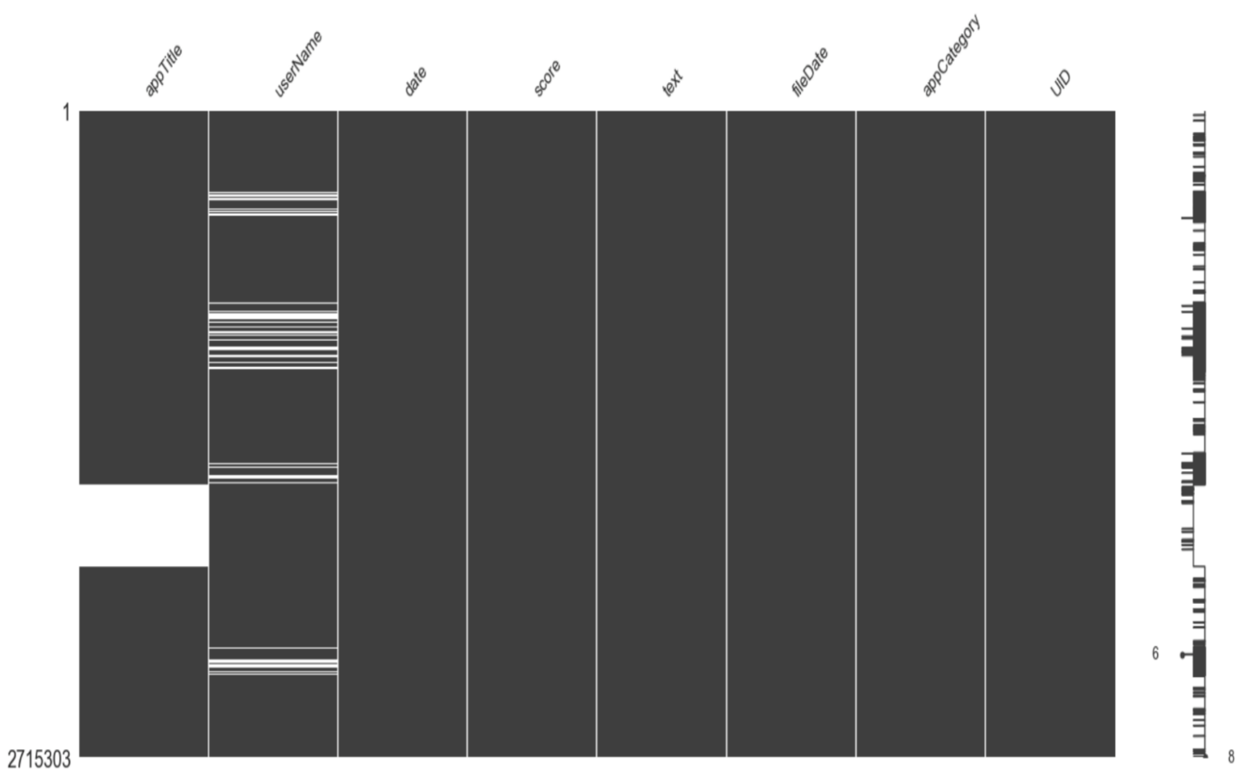
File Consolidation:

File Name	Number of Rows
Newest Reviews	**2,715,303**
All Details	**637**

Files did not include the app category or file date, so these 2 columns were added to the file – **'appCategory' and 'fileDate'.**

<u>DATA SPARSITY MATRIX:</u>

Understanding non-null values using the Missingno library:



NaN numbers from data:

	appTitle	userName	date	score	text
NaN Count	342,211	224,007	0	0	182
% Missing Data	12.6%	8.2%	0%	0%	0.001%

The above table gives us an insight that some people rate the app but not leave a comment or text review for the app. NaN values in appTitle is a cause of concern and we should further investigate the reasons behind this.

Features/Variables of Importance Dataset:

<u>Unique Categories</u>**:** Education, Entertainment, Family, Finance, Game Action, Health & Fitness, Lifestyle, Music & Audio

<u>Unique Content Ratings</u>**:** Everyone**,** Teen, Mature 17+, Everyone 10+

<u>Score</u>**:** Scale for rating app from 1 to 5

App Titles: 86 unique apps in the dataset

Number of Reviews available by category:

Category	Count
EDUCATION	331308
ENTERTAINMENT	347914
FAMILY	322352
FINANCE	357719
GAME ACTION	354058
HEALTH AND FITNESS	358237
LIFESTYLE	285482
MUSIC AND AUDIO	358233

Health and Fitness has the maximum number of reviews and Lifestyle has the least number of reviews. However, on further analyzing data we can observe that we have almost uniform distribution of reviews across all the 8 categories.

2. **Cleaning of Dataset**

The following are the key steps involved in text-processing process:

- In initial stage of the process, we removed non English words from the dataset followed by punctuation and non-ASCII characters as well as characters with more than 2 repetitions
- Removed reviews having 2 or less words in the text column since these words will make no sense in most cases and will be act as noise in data while doing textual analysis

After the text processing, the remaining observations in the database is 855,193.

Changes in dataset after cleaning and preprocessing data:

Category	Initial Number	Processed Number	Ratio
EDUCATION	128984	83650	0.648530
ENTERTAINMENT	226420	119657	0.528474
FAMILY	143707	91325	0.635494
FINANCE	185512	111647	0.601832
GAME ACTION	252872	127026	0.502333
HEALTH AND FITNESS	154168	103444	0.670982
LIFESTYLE	137350	85773	0.624485
MUSIC AND AUDIO	216963	132671	0.611491

appCategory	contentRating	Initial Number of Reviews	Processed Number of Reviews	Ratio
EDUCATION	Everyone	87861	56981	0.648536
ENTERTAINMENT	Everyone	14377	9371	0.651805
ENTERTAINMENT	Mature 17+	7278	4468	0.613905
ENTERTAINMENT	Teen	161863	81751	0.505063
FAMILY	Everyone	58645	34492	0.588149
FAMILY	Everyone 10+	48107	33988	0.706508
FINANCE	Everyone	141158	83311	0.590197
GAME ACTION	Everyone	87560	50096	0.572133
GAME ACTION	Mature 17+	29360	10822	0.368597
GAME ACTION	Teen	92577	43631	0.471294
HEALTH AND FITNESS	Everyone	109865	72887	0.663423
LIFESTYLE	Everyone	81224	48777	0.600524
LIFESTYLE	Mature 17+	21721	13493	0.621196
LIFESTYLE	Teen	3012	1925	0.639110
MUSIC AND AUDIO	Everyone	22754	14538	0.638921
MUSIC AND AUDIO	Teen	150227	90651	0.603427

Game Action category is most significantly impacted by cleaning as it has smallest ratio indicating that maximum percentage of rows deleted by category is highest for this category.

3. Category-wise Analysis

Category-wise analysis is carried out with 2 important parameters: score and review word length.

<u>Score vs Number of Reviews:</u>

score	Number of Reviews
1	167017
2	40272
3	52779
4	92302
5	502823

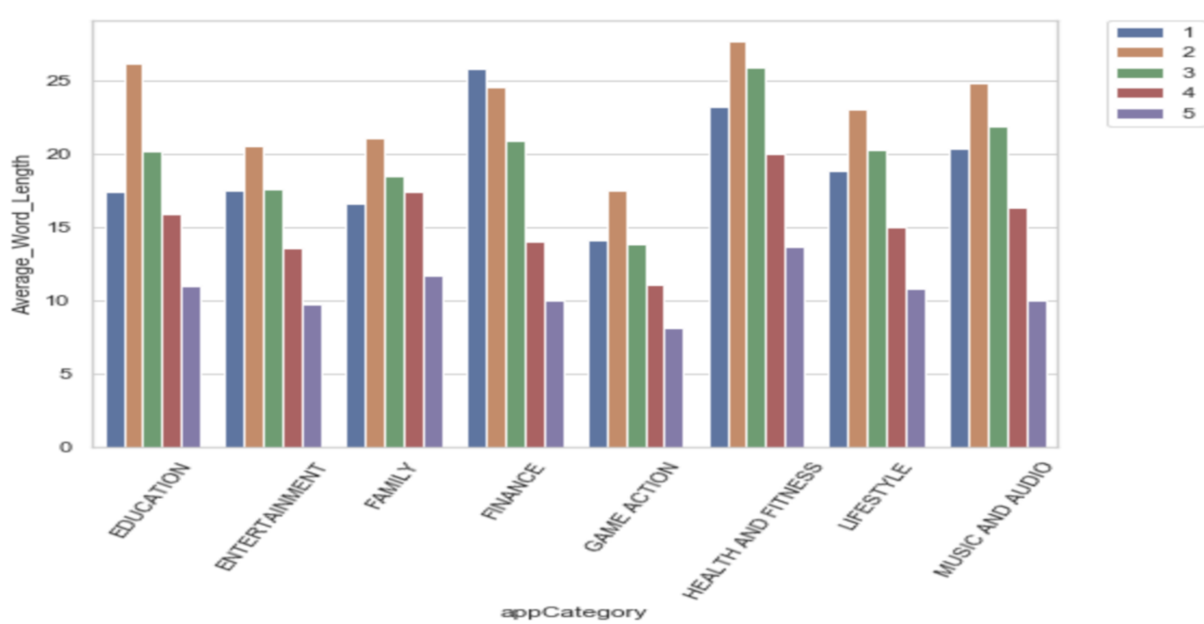
Maximum number of english reviews available are for score/rating of 5. This means most people who give a rating of 5 also write and explain why they like a particular app.

<u>Score vs Average Word Length:</u>

score	Average Review Length
1	19.834197
2	23.500546
3	20.067849
4	15.375626
5	10.470020

Average word length per review is higher for low score reviews (2 and 3). We believe that people tend to be more specific and detailed when they do not like the app or have complains about the app to try. The same can be observed from the above numbers as well.

<u>Average Word Length vs Category by Score:</u>



4. Correlation and Time-Based Analysis:

The numbers calculated on last page point us to investigate if there is any correlation between score and word length i.e is it a pattern that users rating an app with high score write shorter review and for low score with longer reviews.

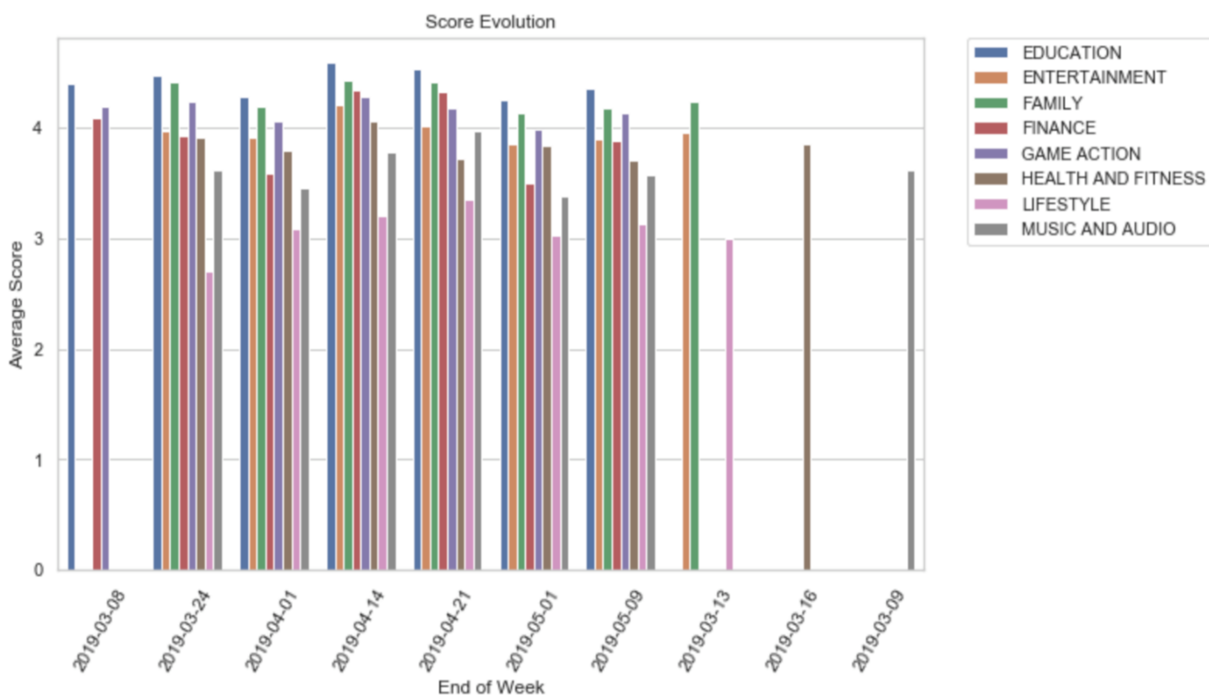
Here are the results of correlation between score and word length by each category:

```
appCategory
EDUCATION              -0.207587
ENTERTAINMENT          -0.253039
FAMILY                 -0.149600
FINANCE                -0.403002
GAME ACTION            -0.210923
HEALTH AND FITNESS     -0.248917
LIFESTYLE              -0.232061
MUSIC AND AUDIO        -0.308771
Name: score, dtype: float64
```

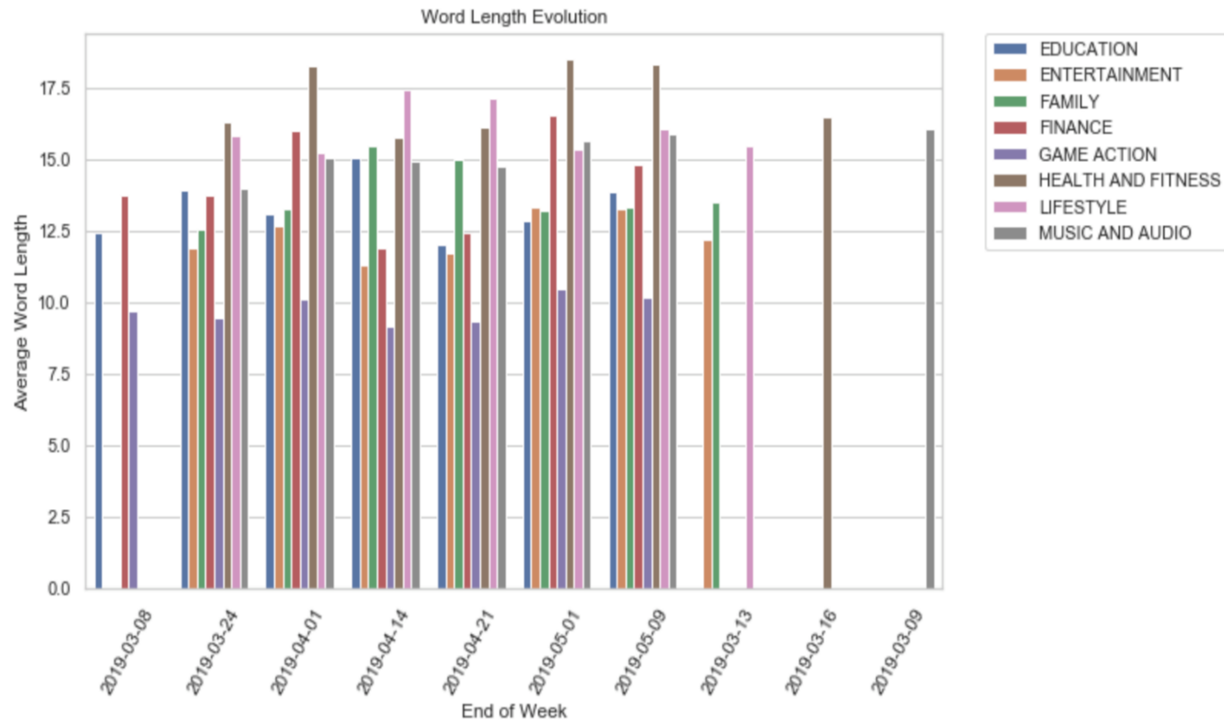
The negative sign is suggesting and supporting our initial assumption but since there is very weak.

Also, the plots for score and average word length by category over time do not show any major trend and we do not have any clear picture around evolution of apps by category.

<u>Score vs Time:</u>



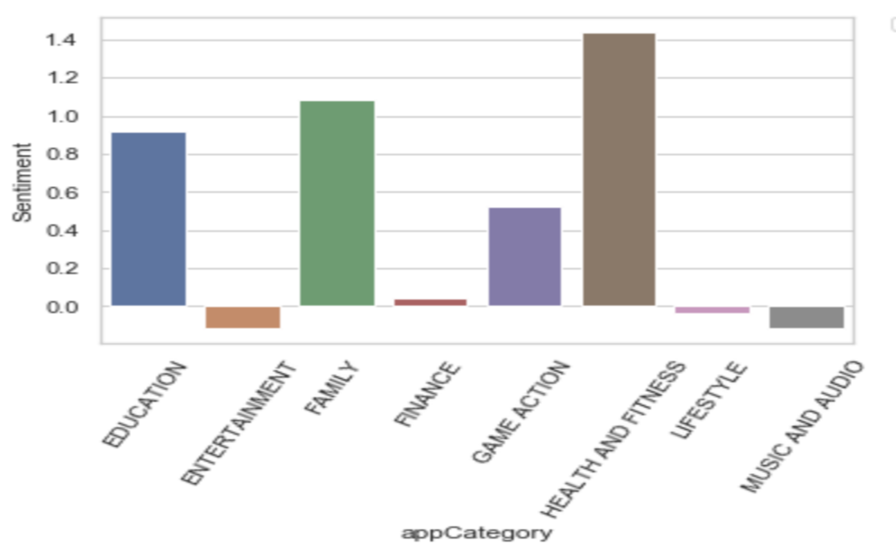
Average Word Length vs Time:



5. Manual Evaluation:

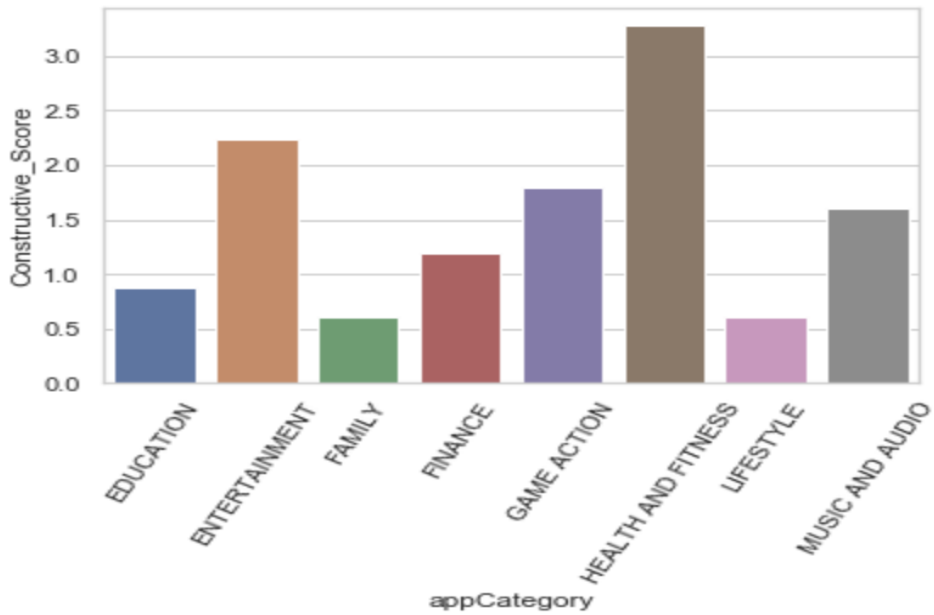
From each category, randomly we selected 25 reviews in a manner that 5 reviews should be picked up from each score group. Therefore, for given 8 categories, we picked up 200 reviews randomly and carried out sentiment analysis for the apps on the scale of [-5 , 5] based on the original review of customers.

Overall, we have observed positive sentiment for most app categories with Health and Fitness having the best sentiment score. Apps under Entertainment and Music and Audio have a slightly negative sentiment.



The overall average of the sentiment analysis was ~0.5 which suggests that overall the sentiment was neutral for the selected data and it should be the case as well since we uniformly sampled data from each of the scores.

Constructive Feedback analysis shows that most of the users actually don't give good constructive feedback to the developers on which they can work and then come up with an updated better version of the app.



6. Conclusion:

After analyzing the data, we do not have any concrete evidence in support of the argument that reviews differ across categories. Therefore, we need to gather more data from more users and we also need to deep dive more into the textual reviews to understand the behavior between different app categories.

So far, the finding from the data as well as from the manual evaluation are inconclusive and suggest that there is no difference in the reviews for 8 app categories.

The user sentiment seems neutral overall with marginally positive sentiment for Health and Fitness related apps and marginally negative for Entertainment and Music and Audio apps.

7. Learnings:

While working on this project, I have become more familiar with the pandas framework. Also, I have worked a bit with the NLTK package for removing Non-English words from the reviews.

Both these packages have vast applications and having worked on these packages will further help me in exploring these packages more in future.