**Eva Nguyen**
**Student ID: 80476922**

## Question 1

**Drop duplicates and record how many unique reviews are collected?**
There are 1,478,938 unique reviews collected.

## Question 2

**How many unique apps are in the dataset? How many apps in each of the 8 specified app categories?**
There are 86 unique apps. The count of apps in each of the 8 specified app categories is below.

```
category
EDUCATION              10
ENTERTAINMENT          13
FAMILY                 10
FINANCE                10
GAME_ACTION            11
HEALTH_AND_FITNESS     10
LIFESTYLE              12
MUSIC_AND_AUDIO        11
```

## Question 3

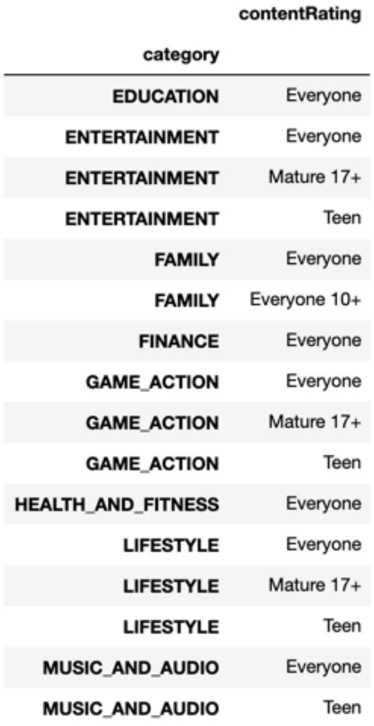
**How many reviews exist in each of the eight app-categories?**
The count of reviews in each of the eight app-categories is below.

```
category
EDUCATION              137227
ENTERTAINMENT          226723
FAMILY                 167172
FINANCE                185512
GAME_ACTION            252986
HEALTH_AND_FITNESS     154330
LIFESTYLE              137751
MUSIC_AND_AUDIO        217237
```

## Question 4

**Based on the contentRating column in the details files for each app, what are the different contentRating groups in each app-category?**

The different contentRating groups in each app-category is to the right.



|  | contentRating |
|---|---|
| **category** |  |
| **EDUCATION** | Everyone |
| **ENTERTAINMENT** | Everyone |
| **ENTERTAINMENT** | Mature 17+ |
| **ENTERTAINMENT** | Teen |
| **FAMILY** | Everyone |
| **FAMILY** | Everyone 10+ |
| **FINANCE** | Everyone |
| **GAME_ACTION** | Everyone |
| **GAME_ACTION** | Mature 17+ |
| **GAME_ACTION** | Teen |
| **HEALTH_AND_FITNESS** | Everyone |
| **LIFESTYLE** | Everyone |
| **LIFESTYLE** | Mature 17+ |
| **LIFESTYLE** | Teen |
| **MUSIC_AND_AUDIO** | Everyone |
| **MUSIC_AND_AUDIO** | Teen |

## Question 5¶

**How many apps exist in each of these contentRating-sub-groups in each app-category?**
The count of apps in each of the contentRating-sub-groups in each app-category is below.

```
category             contentRating
EDUCATION            Everyone         10
ENTERTAINMENT        Everyone          4
                     Mature 17+        1
                     Teen              9
FAMILY               Everyone          7
                     Everyone 10+      3
FINANCE              Everyone         10
GAME_ACTION          Everyone          5
                     Mature 17+        2
                     Teen              4
HEALTH_AND_FITNESS   Everyone         10
LIFESTYLE            Everyone         11
                     Mature 17+        1
                     Teen              1
MUSIC_AND_AUDIO      Everyone          2
                     Teen              9
```

## Question 6

**How many reviews in each contentRating-sub-groups in each app-category?**
The number of reviews in each contentRating-sub-groups in each app-category is below.

```
category             contentRating
EDUCATION            Everyone            95916
ENTERTAINMENT        Everyone            14377
                     Mature 17+           7278
                     Teen               161863
FAMILY               Everyone            77211
                     Everyone 10+        48107
FINANCE              Everyone           141158
GAME_ACTION          Everyone            87560
                     Mature 17+          29360
                     Teen                92577
HEALTH_AND_FITNESS   Everyone           109865
LIFESTYLE            Everyone            81224
                     Mature 17+          21721
MUSIC_AND_AUDIO      Everyone            22754
                     Teen               150227
```

## Question 7

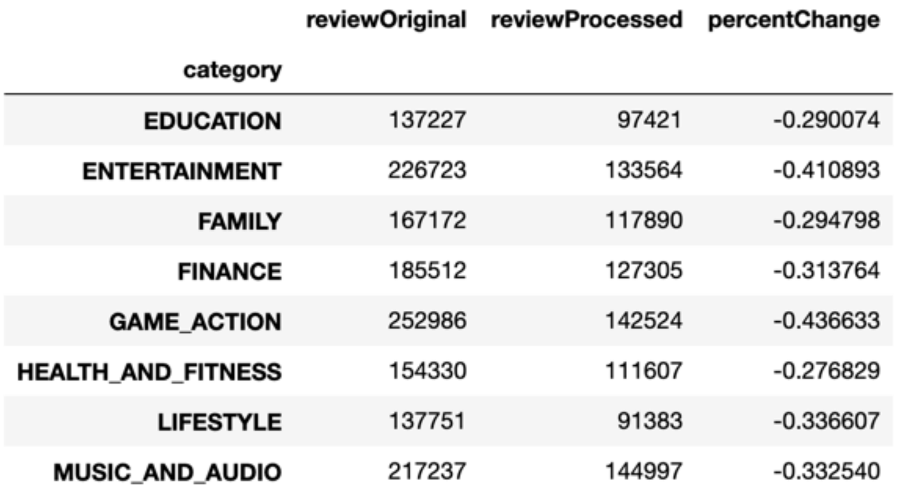
**Should we remove the reviews that contain two or less number of words?**
**a. Justify your answers.**
**b. Can we remove the reviews with two or less words for some score-sub-groups and keep them in the other ones? E.g. if the review has score 1, we should remove such reviews, but we should keep them if the rating is 5. Justify your answers.**

**a.** I do not believe we should remove reviews that contain two or less number of words. The reason is I used to be a data analyst for the customer service department, and we realized people are more likely to submit a review if it is easy and requires little time to do so. For that reason, people are more prone to complete a survey that requires selecting values rather than typing out commentary. Score ratings from reviews with two or less number of words likely contain honest score ratings so they are important to keep in the full data set.

**b.** No, I do not think we should remove reviews with two or less words for some score-sub-groups and keep them in other ones. The reason is similar to the one in part a. Customers are most likely providing honest score ratings but provided two or less words commentary as an obligation to the review requirements. For that reason, even with two or less words in the comments section, the score ratings are important to keep in the full data set.

## Question 8

**How many reviews exist in each of the eight app-categories? Compare with question 3.**
The count of reviews in each of the eight app-categories is below. The percenChange indicates the change when comparing to question 3. Interestingly, the Game Action app category is showing a very large decrease in number of reviews from question 3.

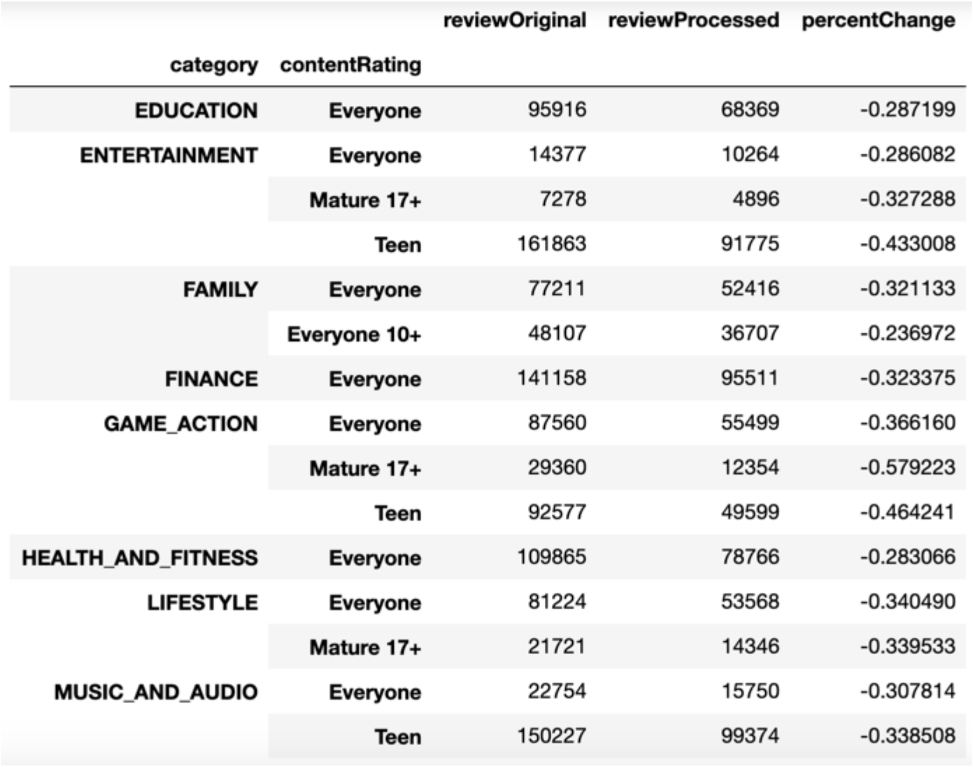


| | reviewOriginal | reviewProcessed | percentChange |
|---|---|---|---|
| **category** | | | |
| **EDUCATION** | 137227 | 97421 | -0.290074 |
| **ENTERTAINMENT** | 226723 | 133564 | -0.410893 |
| **FAMILY** | 167172 | 117890 | -0.294798 |
| **FINANCE** | 185512 | 127305 | -0.313764 |
| **GAME_ACTION** | 252986 | 142524 | -0.436633 |
| **HEALTH_AND_FITNESS** | 154330 | 111607 | -0.276829 |
| **LIFESTYLE** | 137751 | 91383 | -0.336607 |
| **MUSIC_AND_AUDIO** | 217237 | 144997 | -0.332540 |

## Question 9

**How many reviews in each contentRating-sub-groups in each app-category? Compare with question 6.**

The number of reviews in each contentRating-sub-groups in each app-category is below. The percenChange indicates the change when comparing to question 6. The contentRating groups assist with providing insight to the interesting decrease in Question 8 for the Game Action app category. The large decrease can be due to Mature 17+ and Teen contentRating groups because they both have large percent changes.



| | | reviewOriginal | reviewProcessed | percentChange |
|---|---|---|---|---|
| **category** | **contentRating** | | | |
| **EDUCATION** | **Everyone** | 95916 | 68369 | -0.287199 |
| **ENTERTAINMENT** | **Everyone** | 14377 | 10264 | -0.286082 |
| | **Mature 17+** | 7278 | 4896 | -0.327288 |
| | **Teen** | 161863 | 91775 | -0.433008 |
| **FAMILY** | **Everyone** | 77211 | 52416 | -0.321133 |
| | **Everyone 10+** | 48107 | 36707 | -0.236972 |
| **FINANCE** | **Everyone** | 141158 | 95511 | -0.323375 |
| **GAME_ACTION** | **Everyone** | 87560 | 55499 | -0.366160 |
| | **Mature 17+** | 29360 | 12354 | -0.579223 |
| | **Teen** | 92577 | 49599 | -0.464241 |
| **HEALTH_AND_FITNESS** | **Everyone** | 109865 | 78766 | -0.283066 |
| **LIFESTYLE** | **Everyone** | 81224 | 53568 | -0.340490 |
| | **Mature 17+** | 21721 | 14346 | -0.339533 |
| **MUSIC_AND_AUDIO** | **Everyone** | 22754 | 15750 | -0.307814 |
| | **Teen** | 150227 | 99374 | -0.338508 |

## Question 10

**What is the number of reviews for each score (score column)? For example, 35000 reviews have a score of 1, etc.**
The number of reviews for each score is below.

```
score
1    173270
2     41358
3     55655
4    102145
5    594263
```

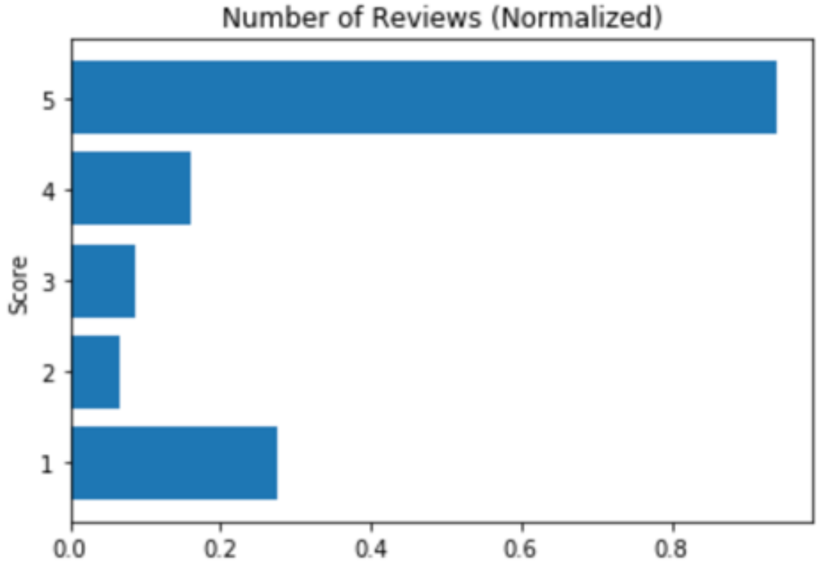
## Question 11¶

**How many apps exist in each score-sub-group?**
As shown below, all apps exist in each score-sub-group.

```
score  category
1      EDUCATION              10
       ENTERTAINMENT          13
       FAMILY                 10
       FINANCE                10
       GAME_ACTION            11
       HEALTH_AND_FITNESS     10
       LIFESTYLE              12
       MUSIC_AND_AUDIO        11
2      EDUCATION              10
       ENTERTAINMENT          13
       FAMILY                 10
       FINANCE                10
       GAME_ACTION            11
       HEALTH_AND_FITNESS     10
       LIFESTYLE              12
       MUSIC_AND_AUDIO        11
3      EDUCATION              10
       ENTERTAINMENT          13
       FAMILY                 10
       FINANCE                10
       GAME_ACTION            11
       HEALTH_AND_FITNESS     10
       LIFESTYLE              12
       MUSIC_AND_AUDIO        11
4      EDUCATION              10
       ENTERTAINMENT          13
       FAMILY                 10
       FINANCE                10
       GAME_ACTION            11
       HEALTH_AND_FITNESS     10
       LIFESTYLE              12
       MUSIC_AND_AUDIO        11
5      EDUCATION              10
       ENTERTAINMENT          13
       FAMILY                 10
       FINANCE                10
       GAME_ACTION            11
       HEALTH_AND_FITNESS     10
       LIFESTYLE              12
       MUSIC_AND_AUDIO        11
```

## Question 12

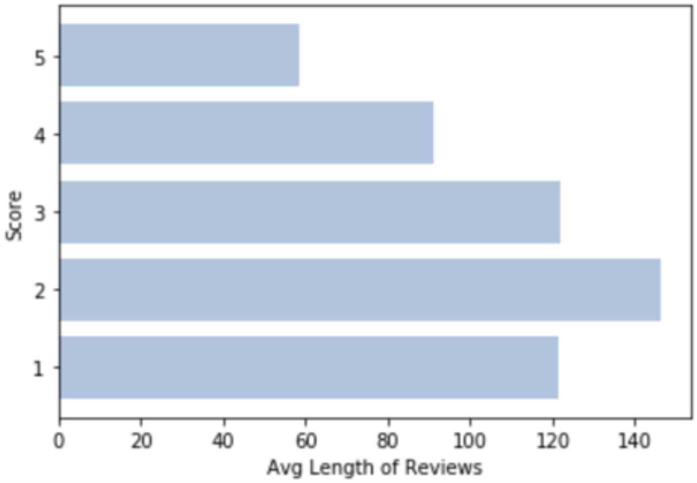
**Compare the number of reviews for each score in a plot (Remember to normalize the numbers when you are comparing them).**
After normalizing the numbers, there is a great proportion of reviews with score 5 and score 1. This is a pattern with many customer reviews surveys, because people tend to complete reviews if they absolutely love or hate a product.



## Question 13

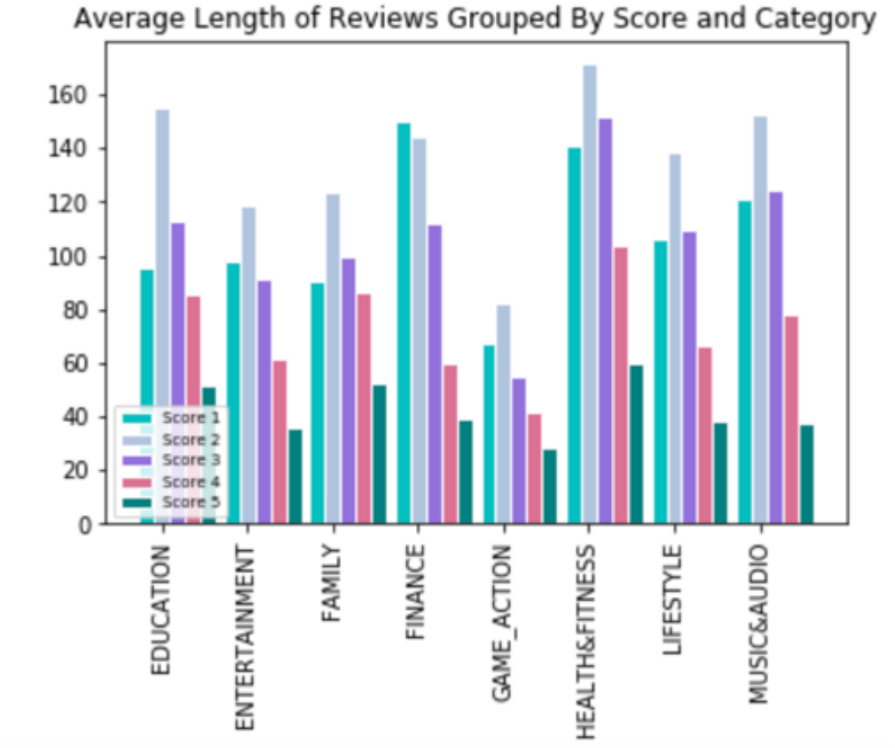
**What is the average length of the reviews in each score-sub-group?**
The average length of the reviews in each score-sub-group is below. The three lowest scores have the highest average length of reviews. This makes sense because people that dislike a product may be more vocal in explaining their distaste.

```
   score  mean_len_text
0      1     121.446748
1      2     146.420515
2      3     121.957724
3      4      90.944379
4      5      58.398489
```



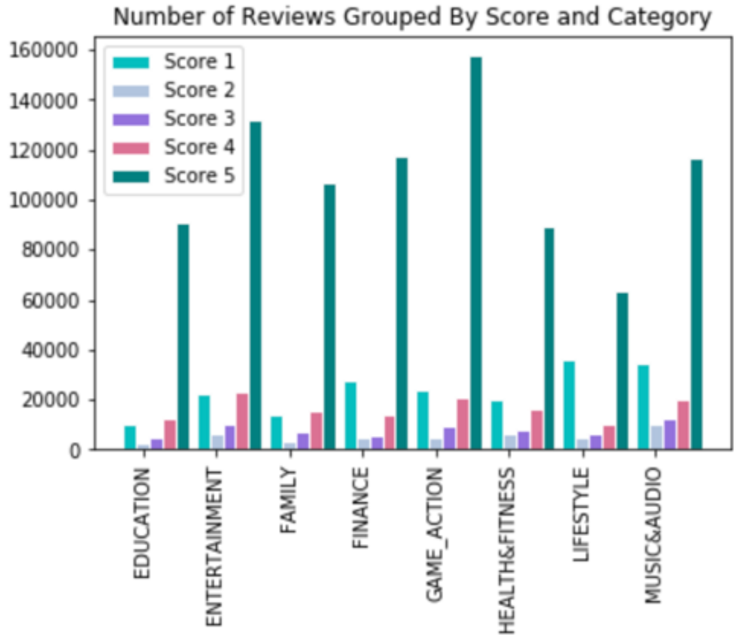
## Question 14¶

**Compare the average length of reviews in each score-sub-group in the 8 app categories (draw a plot).**
The average length of reviews in each score-sub-group in 8 app categories is below. All the app categories have the last three lowest scores with the highest average length of reviews, similar to question 13. Finance is the only app category where Score 2 does not have the highest average length but Score 1 does.



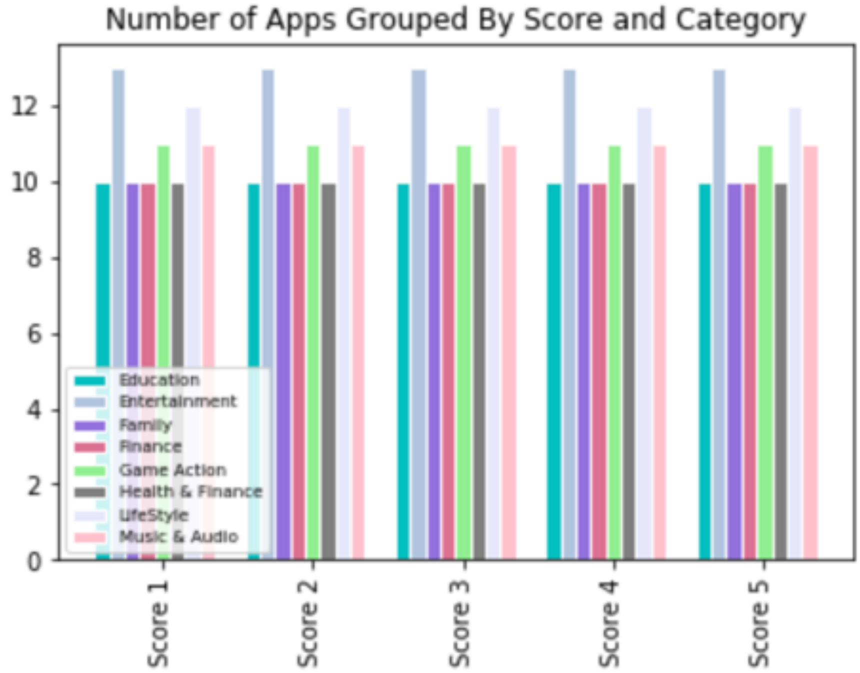
## Question 15

**Compare the number of reviews in each score-sub-group in the 8 app categories (draw a plot).**
Each category has the largest proportion in Score 5 and Score 1. This is a pattern with many customer reviews surveys, because people normally complete reviews they feel strongly about a product. High proportions in Score 1, because they love it, even higher proportions for Score 5 when they're really unhappy with a product.



## Question 16

**Compare the number of apps in each score-sub-group in the 8 app categories (draw a plot).**
All the apps are evenly distributed in each score-sub-group. This is representative in previous questions where we found that all apps exist in each score-sub-group.



## Question 17

**Is there any correlation between the length of the reviews and the score in each app-category?**
Based on the correlation, there is a negative correlation between length of reviews and the score in each app-category. This is in line with our previous analysis. Previously, we noticed the lower scores are higher in average review length for all app categories, which is indictive of a negative correlation.

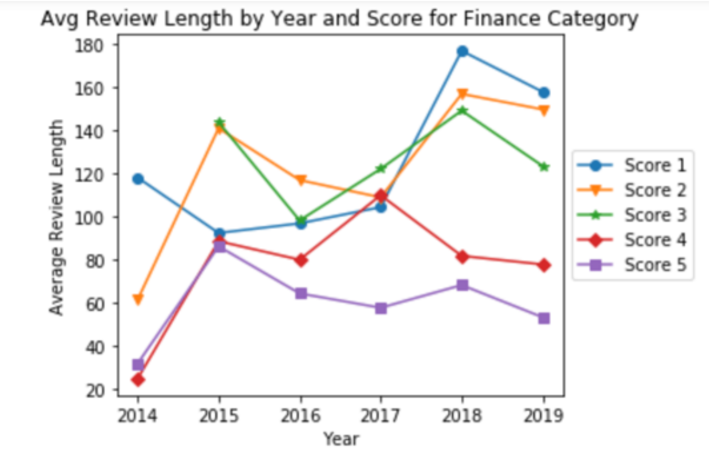
```
category
EDUCATION            -0.632214
ENTERTAINMENT        -0.861607
FAMILY               -0.638768
FINANCE              -0.974997
GAME_ACTION          -0.859411
HEALTH_AND_FITNESS   -0.792682
LIFESTYLE            -0.788759
MUSIC_AND_AUDIO      -0.824784
```

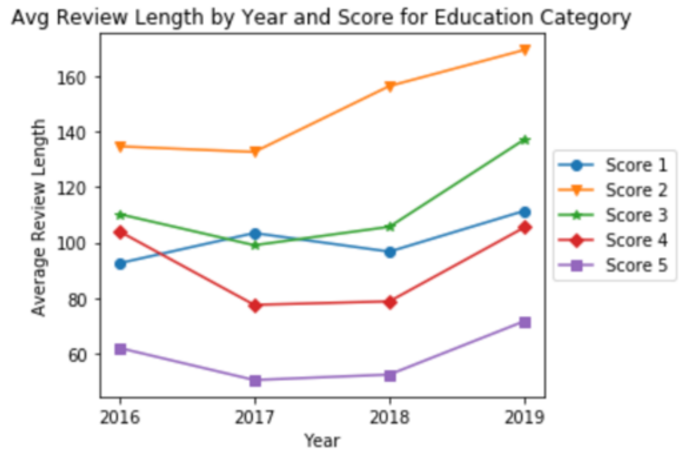
## Question 18

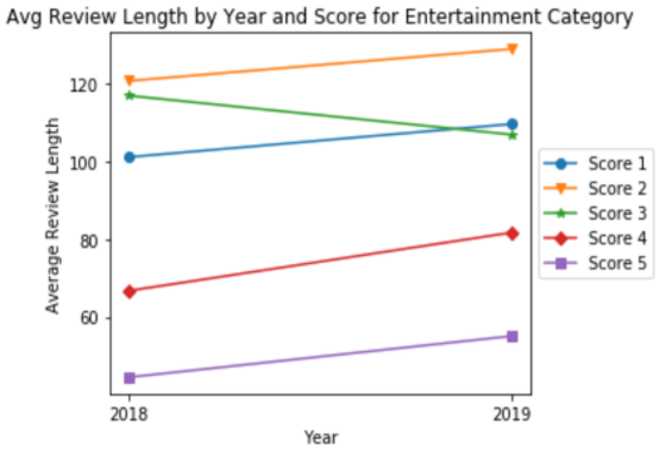
**Find the evolution/changes of the star rating and length of reviews during time for each app category. Draw plots.**
**a. Can we use the date column for this question?**
**b. If not, what is the solution based on the data you have?**

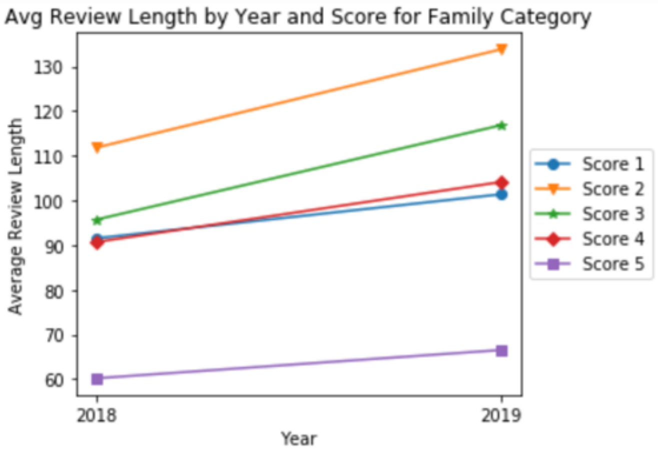
**a.** Yes, I was able to confirm Python properly converted all values in the date columns to dates (even the ones that are represented as a numerical value in Microsoft Excel). The reason is the numerical value in Excel is the number of days since January 1, 1900. In Pandas read_csv, the function is able to decode this numerical value to represent it as a date.
**b.** If I was not able to use all the dates in the date column, I would take the non-dates from the date column and re-assign it to the date from the filename.

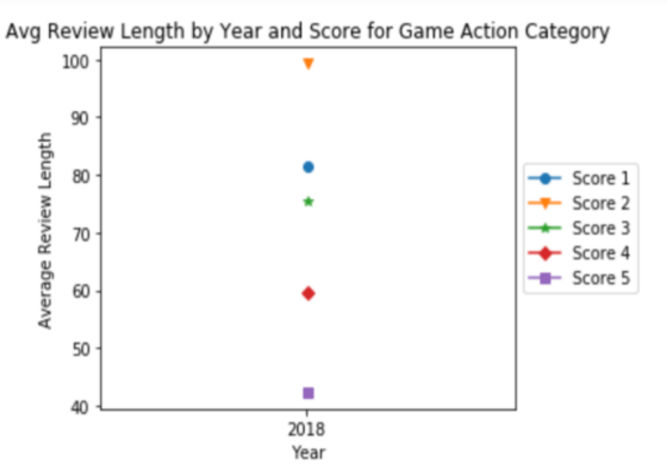
**Analysis**
I would expect all categories to show the same trend for every year as shown in my previous analysis. My previous analysis had been lower scores have larger average review lengths overall. After reviewing all 8-line graphs for each of the app categories, the trend remains the same throughout the years. Score 4 and Score 5 remain the lowest in Average Review Lengths while Score 1, Score 2, and Score 3 remain the largest in Average Review Lengths. Score 2 stays as the top Average Review Length throughout the years for all the application categories, but not for Finance in the year 2014, 2018, and 2019.

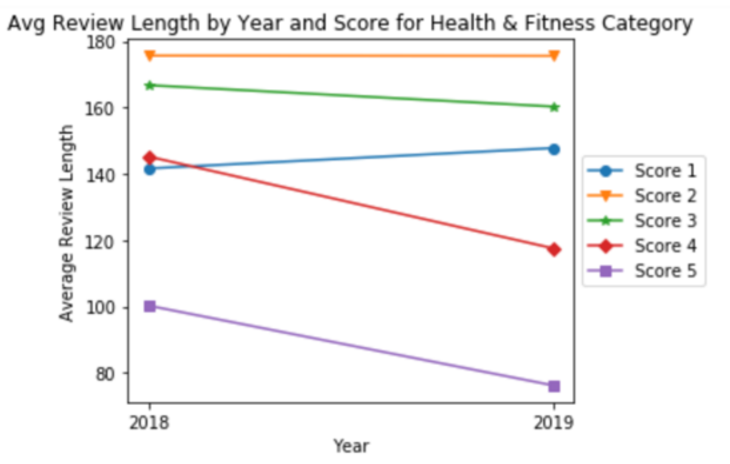


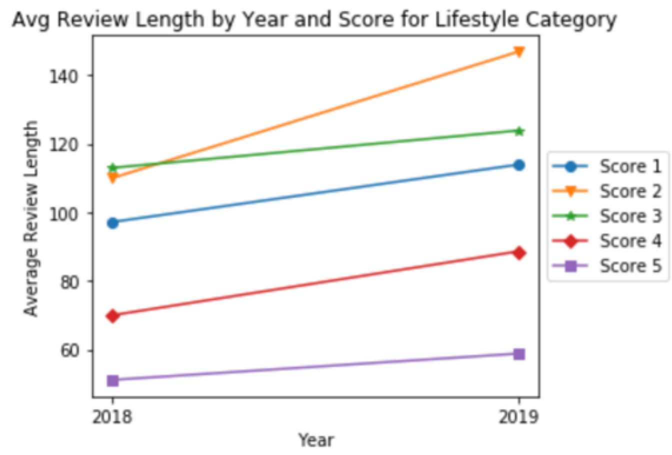


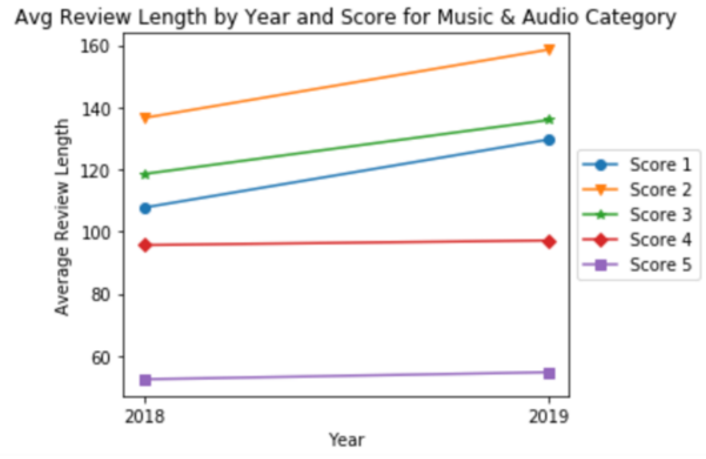












## Question 19

**Compare your scores among the app categories. Do you see a difference between app categories?**

**Sentiment Analysis**

Finance, Lifestyle, Music & Audio, and Entertainment all show Sentiment averages below 1.0. Education, Game Action, Family, and Health & Fitness averages are above 1.0. All categories have averages below 2.0. The reason the 4 category averages were below 1.0 is most likely due to my own subjected ranking system where I gave many 0s (shown in the median dataframe) to those categories. The reason is most users gave opinions where they thought the applications were okay but also thought there could be some improvements in some aspects. However, they did not state if they liked or disliked the application. I can see why Education, Game Action, Family, and Health & Fitness categories have averages above 1.0, because I mainly gave out 3s (shown in the median dataframe). I believe those applications are downloaded more than the other categories; hence more people have sentiments about it.

**Constructive Analysis**

Reading many of the reviews, I found most of them did not have constructive feedback. Most of the reviews either had short comments that they just liked or loved the application, or they would go on a tangent if they really disliked the application and claimed it to be the worst application ever. However, those users were not able to explain the reasons for why they liked or disliked the application that could be useful for developers. I can understand that, because explaining processes of using technology and/or troubleshooting online is extremely difficult. Game Action, Education, and Lifestyle categories had constructive score averages around 2.0. Family and

Entertainment had constructive score averages around 3.0. Music & Audio, Health & Fitness, and Finance had constructive averages around 4.0. I gave a higher proportion of 0s for Game Action and a higher proportion of 8s for Finance. I think the reason Game Action had more 0s is due to the difficulties of explaining app issues when so many glitches occur on games. I believe Finance had more constructive reviews, because there is not a lot you can do with a Finance app, so it is easier to explain and narrow down the issues.

Mean Averages of Sentiment Score

```
Category
FINANCE                 0.00
LIFESTYLE               0.36
MUSIC_AND_AUDIO         0.44
ENTERTAINMENT           0.48
EDUCATION               1.12
GAME_ACTION             1.20
FAMILY                  1.32
HEALTH AND FITNESS      1.76
```

Median of Sentiment Score

```
category
FINANCE               -1
ENTERTAINMENT          0
LIFESTYLE              0
MUSIC_AND_AUDIO        0
EDUCATION              1
FAMILY                 3
GAME_ACTION            3
HEALTH AND FITNESS     3
```

Mean Averages of Constructive Score

```
category
GAME_ACTION             2.08
EDUCATION               2.36
LIFESTYLE               2.92
FAMILY                  3.36
ENTERTAINMENT           3.72
MUSIC_AND_AUDIO         4.12
HEALTH_AND_FITNESS      4.48
FINANCE                 4.96
```

Median of Constructive Score

```
category
GAME_ACTION            0
EDUCATION              1
FAMILY                 2
ENTERTAINMENT          3
LIFESTYLE              3
MUSIC_AND_AUDIO        3
HEALTH_AND_FITNESS     5
FINANCE                8
```

**Is the any difference in the reviews from the 8 app categories that you have investigated?**

Yes, there is a difference in the reviews from the 8 app categories. The way each app is described for its greatness and its weaknesses differ between categories. The reason is the different categories are promoting different features so there would be different opinions about them. However, I did notice a common pattern when it came to reviews with a score of 1 and a score of 5 across the 8 app categories. Many reviews with a score of 5 were very short about how much the users loved the application. Reviews with a score of 1 had a lot more negative language and were on average much greater in length than reviews with score of 5. This corresponds to our previous analyses that lower scores have higher average review lengths while higher scores have lower average review lengths.

**Share your learnings and thoughts in 1 or 2 paragraphs. Any new technique, libraries that you previously did not know, aspects of data analysis that you did not consider before, etc.**

I really enjoyed this project, because I learned how to use Python for text processing that I was not aware was possible. I used the nltk and itertools libraries for the first time. The nltk library helped me distinguish English words from non-English words. I used the itertools library to group letters and count the number of consecutive letters in a word to remove them. I used it in addition to the nltk library to make sure to not remove letters from words like ‘Hello’. During this experience, I had not realized how important optimizing code is when you have large amounts of data. Initially, I used nltk by running multiple and nested for loops to go through each word of each sentence. However, my script would run for a few hours and it would eventually overheat my computer. I changed methods and used the apply function and the script ran within a few minutes. The data analysis project made me realize there is a lot of libraries and functions out there that can be very powerful for any initiative.