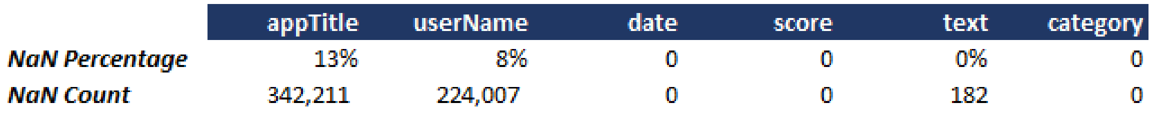
DATA 542 – FINAL REPORT

NAME : NGHIA TRONG PHAN (VINCENT)

A. INITAL STATISTICS

a. Newest Review:

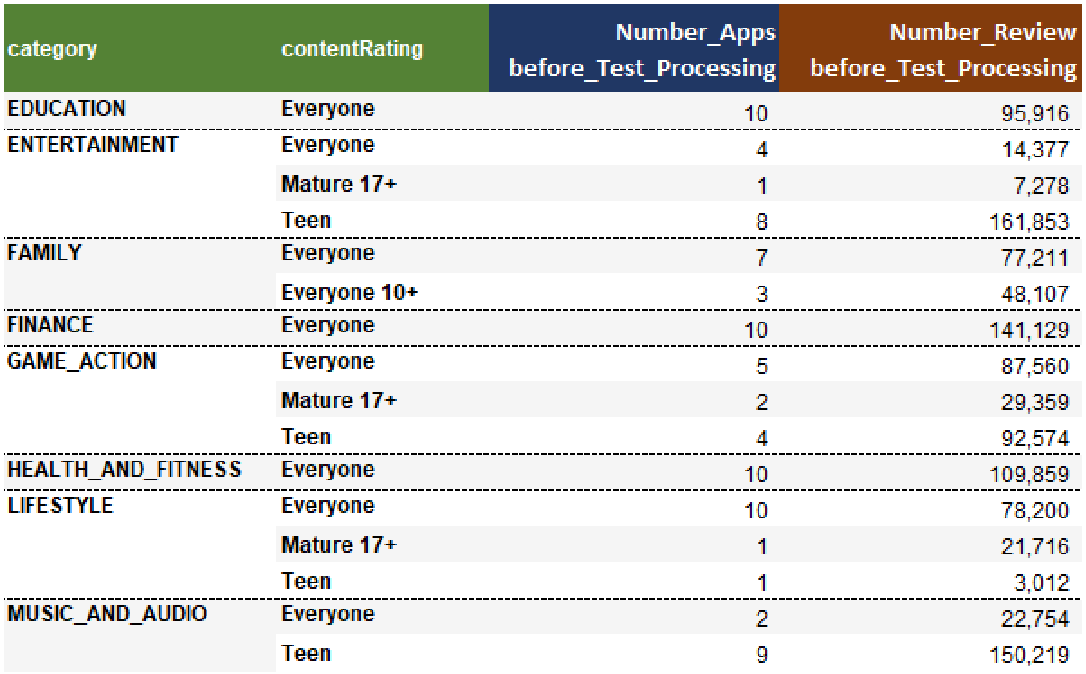
- The dataset including all the Newest review files contains **2.715.303** observations and 5 attributes ***appTitle, username, date, score, text***. The 6th attribute ***category*** is added in according to the folder name where the data is stored. The 7th attribute ***appTitle_lower*** is derived from the ***appTitle*** after applying lower case to the content of ***app_Title***.
- The number of duplicated observations is **1.236.365** about 45% of the data. Those duplicated rows are removed from the database.
- The statistics of `Nan` values is included as below:



	appTitle	userName	date	score	text	category
NaN Percentage	13%	8%	0	0	0%	0
NaN Count	342,211	224,007	0	0	182	0

We find that **appTitle** column has more than 22% NaN values and userName has about 10% NaN value. As the information of ***userName*** is not important, the observations with NaN username value are kept in the dataset while those with Nan appTitle are removed.

- In term of information about the number of App and Reviews, the below table shows that even though the numbers of apps in categories are quite but the number of reviews in ***Entertainment, Game_Action*** and ***Music_and_Audio*** are significant higher than the other groups. It means the numbers of users in these category are high; moreover, from ***Content Rating*** index, we find that the major users in for these three categories are **Teen**. It is worth to mention that Finance apps also have significant number of users regardless the fact is that target customers of these apps are not **Teen**.



category	contentRating	Number_Apps before_Test_Processing	Number_Review before_Test_Processing
EDUCATION	Everyone	10	95,916
ENTERTAINMENT	Everyone	4	14,377
	Mature 17+	1	7,278
	Teen	8	161,853
FAMILY	Everyone	7	77,211
	Everyone 10+	3	48,107
FINANCE	Everyone	10	141,129
GAME_ACTION	Everyone	5	87,560
	Mature 17+	2	29,359
	Teen	4	92,574
HEALTH_AND_FITNESS	Everyone	10	109,859
LIFESTYLE	Everyone	10	78,200
	Mature 17+	1	21,716
	Teen	1	3,012
MUSIC_AND_AUDIO	Everyone	2	22,754
	Teen	9	150,219

B. TEXT PROCESSING

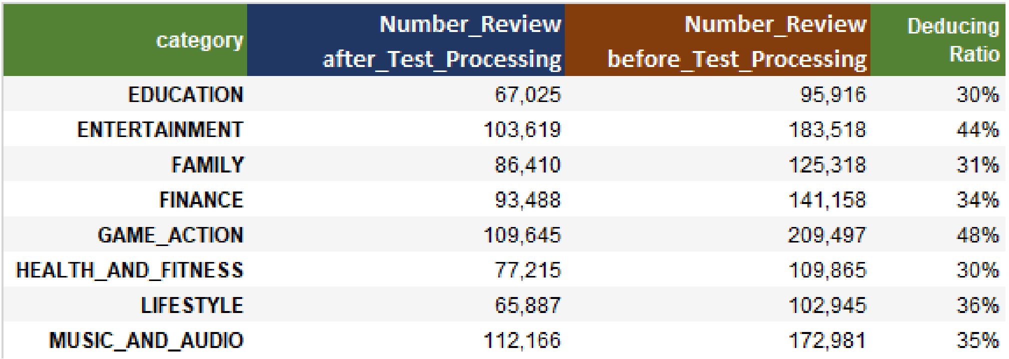
These are the steps in text-processing process:

- In initial stage of the process, we remove punctuation and non-ASCII characters as well as multiple characters in the reviews of `Newest Reviews` data.
- In the next steps, we drop observations with short reviews (less than 3 words). Due to the fact there are many scam reviews which can be made by the owers to promote the product or the competitors to degrade it. The significant evidence for this is the review is often very short. Moreover, if someone give a genuine review about a product with 5 stars or 1 stars, they usually have a long comment to express good/bad feeling about the product.
- If a customer has no impression with the product, their review is utterly short. Therefore, I suggest that we should remove less-than-two-word comments for 1-score and 5-score groups
- After that, we remove non-English reviews in the current database and the number of observations.

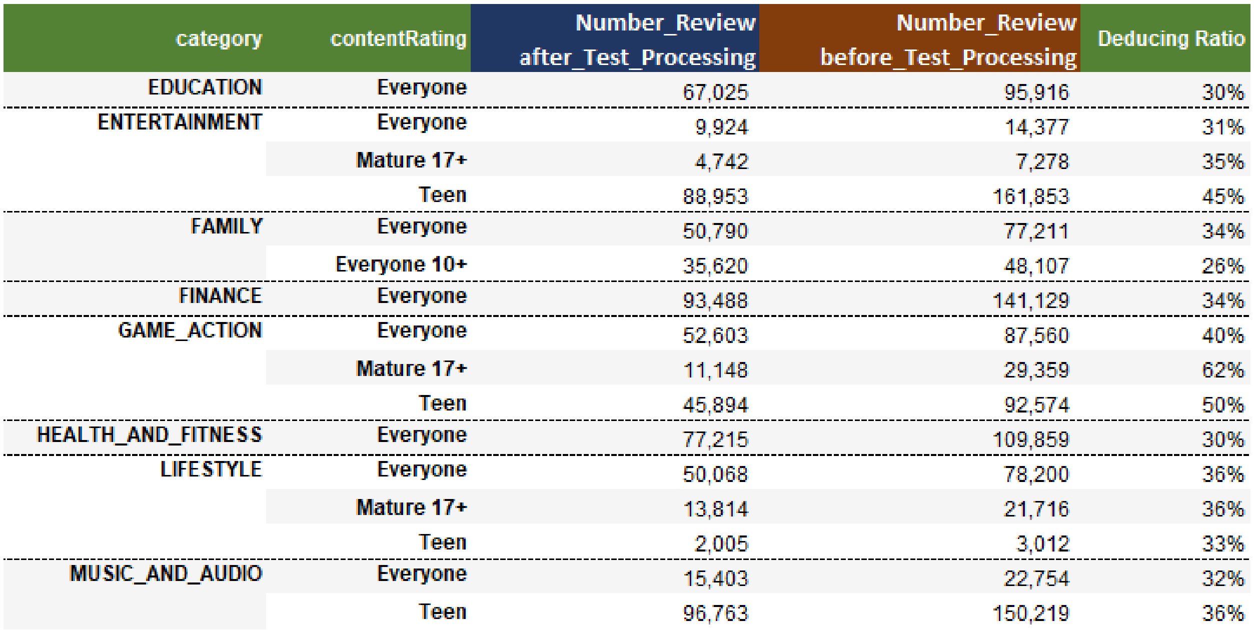
After the text processing, the remaining observations in the database is 715.455.

- **Number of Review**

Now, we review the change of number review after text processing:



category	Number_Review after_Test_Processing	Number_Review before_Test_Processing	Deducing Ratio
EDUCATION	67,025	95,916	30%
ENTERTAINMENT	103,619	183,518	44%
FAMILY	86,410	125,318	31%
FINANCE	93,488	141,158	34%
GAME_ACTION	109,645	209,497	48%
HEALTH_AND_FITNESS	77,215	109,865	30%
LIFESTYLE	65,887	102,945	36%
MUSIC_AND_AUDIO	112,166	172,981	35%



category	contentRating	Number_Review after_Test_Processing	Number_Review before_Test_Processing	Deducing Ratio
EDUCATION	Everyone	67,025	95,916	30%
ENTERTAINMENT	Everyone	9,924	14,377	31%
	Mature 17+	4,742	7,278	35%
	Teen	88,953	161,853	45%
FAMILY	Everyone	50,790	77,211	34%
	Everyone 10+	35,620	48,107	26%
FINANCE	Everyone	93,488	141,129	34%
GAME_ACTION	Everyone	52,603	87,560	40%
	Mature 17+	11,148	29,359	62%
	Teen	45,894	92,574	50%
HEALTH_AND_FITNESS	Everyone	77,215	109,859	30%
LIFESTYLE	Everyone	50,068	78,200	36%
	Mature 17+	13,814	21,716	36%
	Teen	2,005	3,012	33%
MUSIC_AND_AUDIO	Everyone	15,403	22,754	32%
	Teen	96,763	150,219	36%

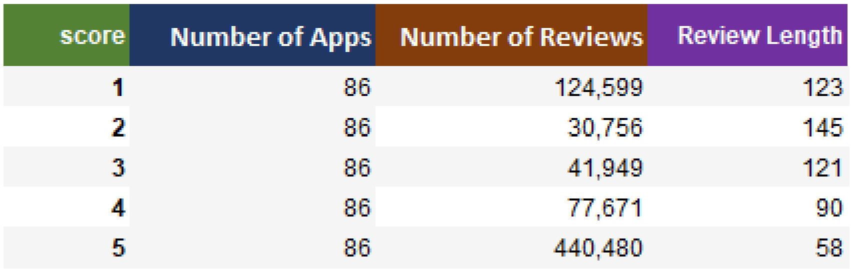
- In term of category, we find GAME_ACTION has highest drop with 48%, next is ENTERTAINMENT with 44%, the other categories drop about 30%

- In term of content rating, we can see that the most significant drop is belong to group ***Mature 17+*** of ***Game_Action*** with 62% the lowest drop of 26% is belong to ***Everyone 10+*** in ***Family*** category, the others are from 32% to 35%.

- **Score:**

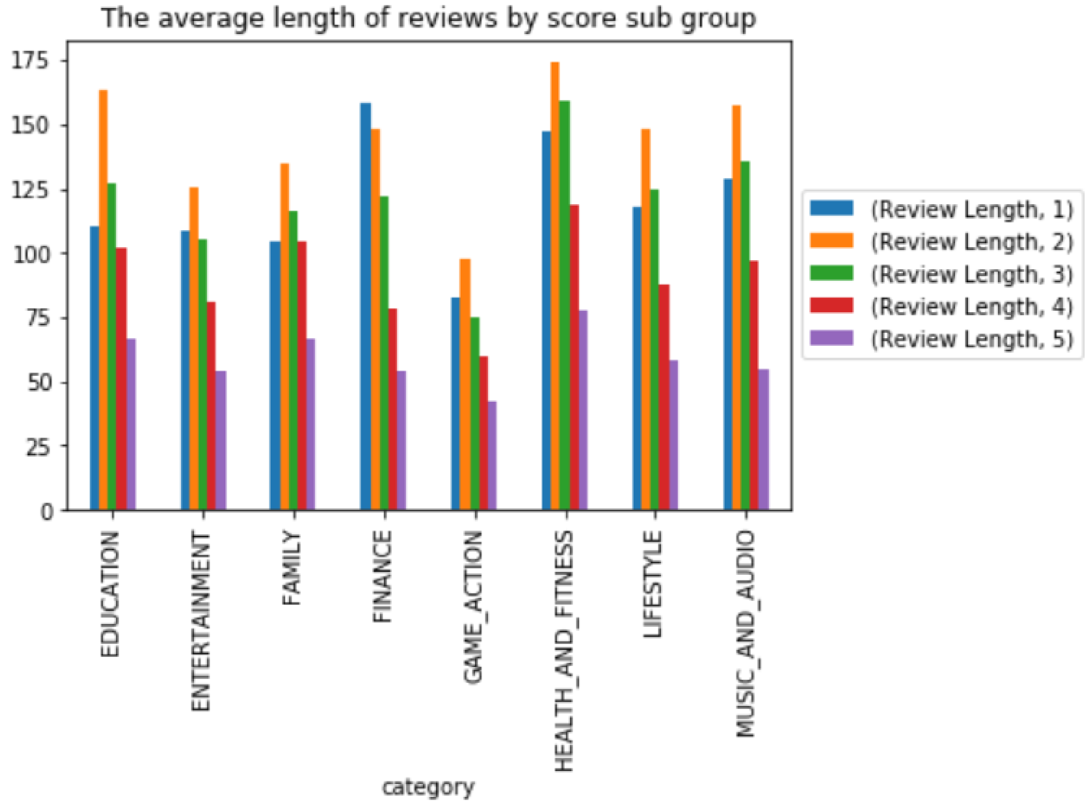
  In the score aspect, all apps are graded from score 1 to 5 showing the diversification of customers' opinion; there is no apps only receiving good score or bad score. More than 2/3 numbers of reviews are score 5 while the second is score 1 with less than 50% the number review of score 5 and the others score have very low reviews. The fact shows the significant bias in the way customers rate an app.

  However, the length of the review noticeable increases when the score reduces (regardless that the score 2 has longer review than score 1). It seems negative experience makes customers write longer to express their disappointment about the app.



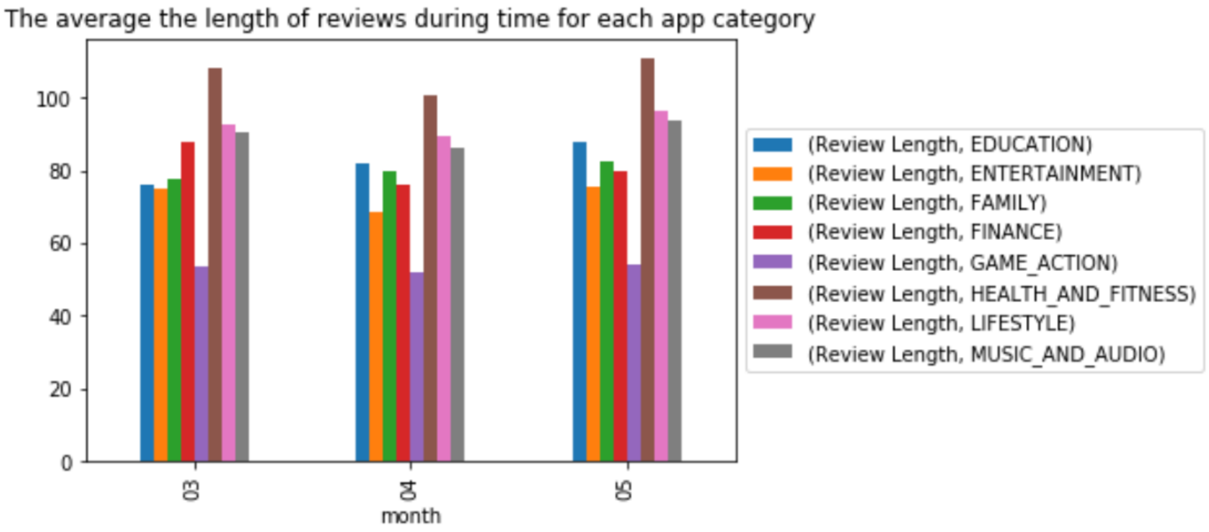
score	Number of Apps	Number of Reviews	Review Length
1	86	124,599	123
2	86	30,756	145
3	86	41,949	121
4	86	77,671	90
5	86	440,480	58

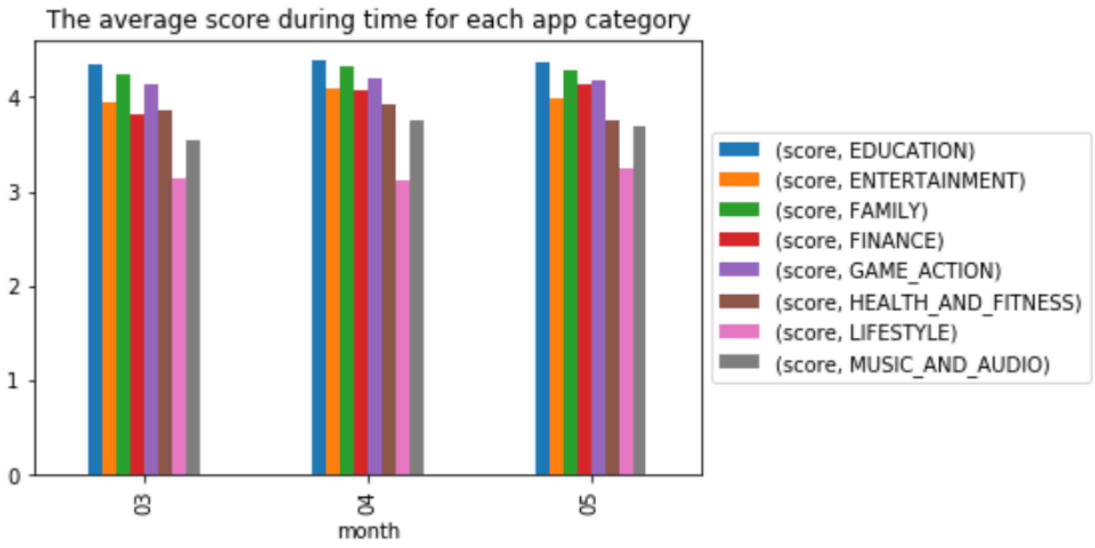
The finding about the negative correlation between score and length review is still true for each app category



The

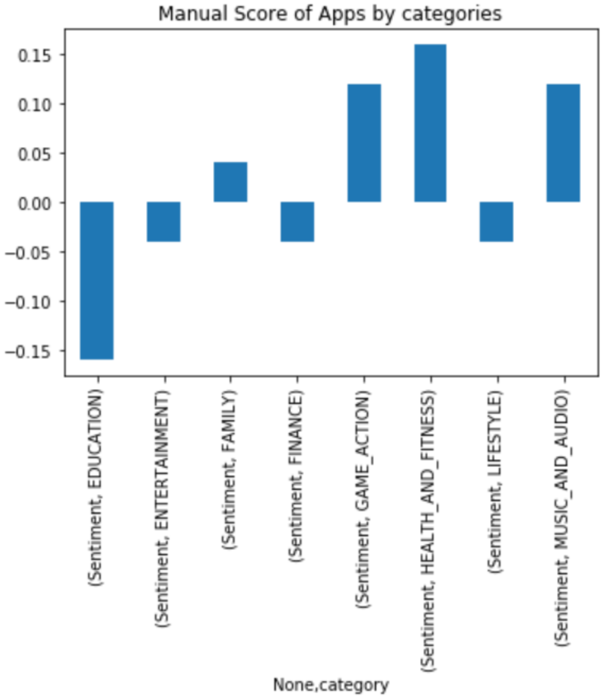
- **Timing:** now we look at the length of reviews and score from March to May. The two graphs below show the stable in structure of review's length and score by categories are quite stable. However, three month period is quite short for us to confirm if this is a long term trend or not.





C. MANUAL VALIDATION

We also pick up randomly 200 reviews and revalidate all the apps on the scale of [-5 , 5] based on the original review of customers. Our valuation shows that in average app categories have close sentiment score via the manual rating and quite neutral with the values are close to 0.



The manual validation show the average grades of App categories are quite evenly and close to 0. There is no app category having outstanding score. This result is similar to the result coming the data found above.

D. CONCLUSION:

The finding from the data as well as the manual validation all confirm that there is no difference in the reviews for 8 app categories and in long term the overall reviews are quite neutral.

During the project, I have change to experience with **langid** library to detect the non-English reviews, even though, the accuracy of the **langid** is not high so if there is more time I would like to experience with other library. Besides, I also have chance to try with Google Colaboratory which can be useful for my study and work in future.