Yonghao Xu Project 1

Based on the data displayed by HappyDB, I want to study the relationship between the entertainment part and this data.

Next, I will introduce the files used in this project and what information they contain;

1"demographic.csc" This csv contains the personal information of all people whose information is collected in HappyDB, such as gender, nationality, age, etc.

```
demo = pd.read_csv("demographic.csv")# A brief view of demographic which include the person's information
demo
```

	wid	age	country	gender	marital	parenthood
0	1	37.0	USA	m	married	y
1	2	29.0	IND	m	married	y
2	3	25	IND	m	single	n
3	4	32	USA	m	married	y
4	5	29	USA	m	married	y
...	...	...	...	...	...	...
10839	13835	25.0	USA	m	single	n
10840	13836	31	USA	m	single	y
10841	13837	22.0	USA	f	single	n
10842	13838	38	USA	f	married	y
10843	13839	24	USA	f	single	y

2"entertainment-dict.csv" This csv file contains keywords related to entertainment.

```
ent = pd.read_csv("entertainment-dict.csv")
ent
```

0	tv
1	film
2	television
3	show
4	book
...	...
108	game of thrones
109	rick and morty
110	comedy
111	nintendo
112	cinema

113 rows × 1 columns

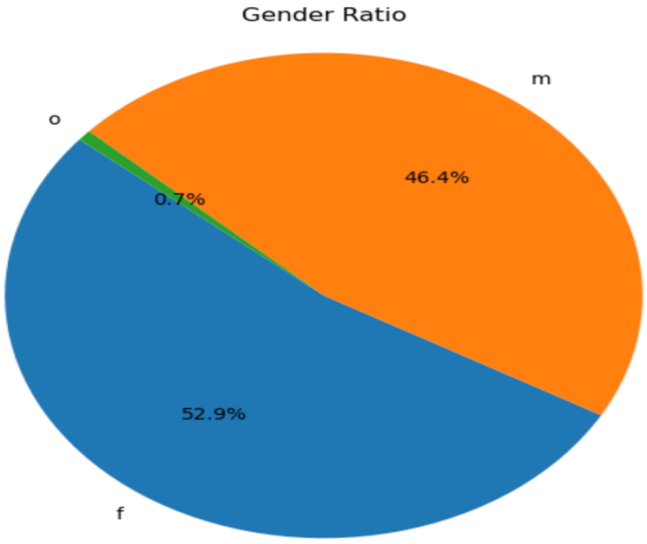
3 “filtered_dataset.csv” I generated a csv file based on all data containing only the entertainment keyword through methods such as merge and streamlining.

```
new_filtered_dataset_df = pd.read_csv("filtered_dataset.csv")# A brief view of demographic which include the person'
new_filtered_dataset_df
```

	hmid	wid	reflection_period	original_hm	cleaned_hm	modified	num_sentence	ground_truth_category	predicted_category	id	text
0	27679	195	24h	I made a new recipe for peasant bread, and it ...	I made a new recipe for peasant bread, and it ...	True	1	NaN	achievement	7	bread peasant recipe spectacular
1	27683	7334	24h	I came in 3rd place in my Call of Duty video g...	I came in 3rd place in my Call of Duty video g...	True	1	NaN	leisure	11	called duty game video
2	27687	586	24h	Hearing Songs It can be nearly impossible to g...	Hearing Songs It can be nearly impossible to g...	False	2	NaN	enjoy_the_moment	15	angry angry direction direction ease feel happ...
3	27696	737	24h	There are two types of people in the world: th...	There are two types of people in the world: th...	False	2	NaN	enjoy_the_moment	24	belief choose choose contrary doesnt fame fort...
4	27699	156	24h	I napped with my husband on the bed this after...	I napped with my husband on the bed this after...	True	1	NaN	affection	27	afternoon bed close cuddled husband nap sweet

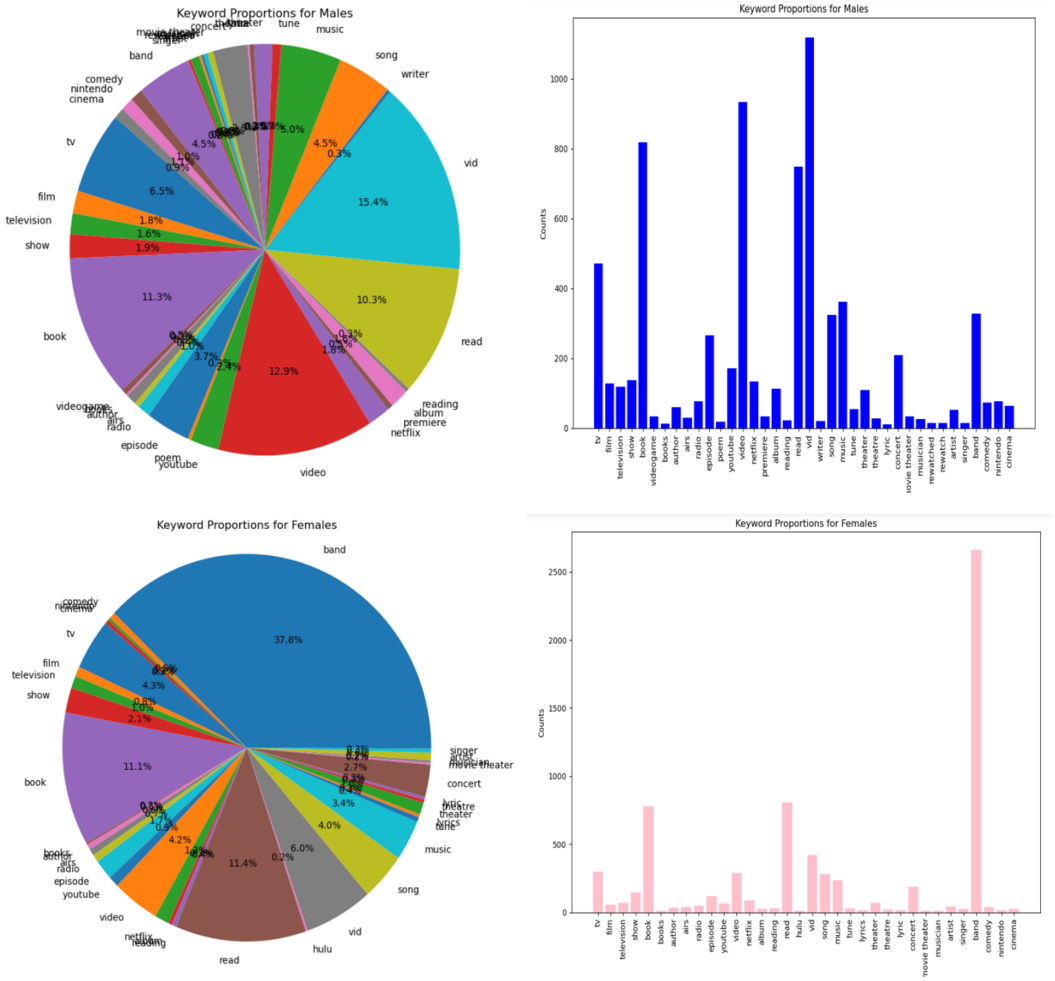
Question1: Is there a difference in the distribution of entertainment keywords between men and women?

First of all we analysis the ratio between the genders:



From the obtained figure, we can easily see that the proportion of men and women is almost the same, so it does not affect our further analysis.

Next, we analyze the Male and Female separately:



Pie Chart Analysis for Males: The first image is a pie chart showing the distribution of various entertainment-related keywords for males. It reveals that 'video' is the most prominent category with 15.4%, followed by 'book' and 'read' with 12.9% and 11.3%, respectively. Smaller categories such as 'cinema' and 'Nintendo' are also present, indicating a diverse range of interests. The multitude of thin slices suggests that males discuss a wide variety of topics within entertainment, though with less concentration in any single category other than 'video.'

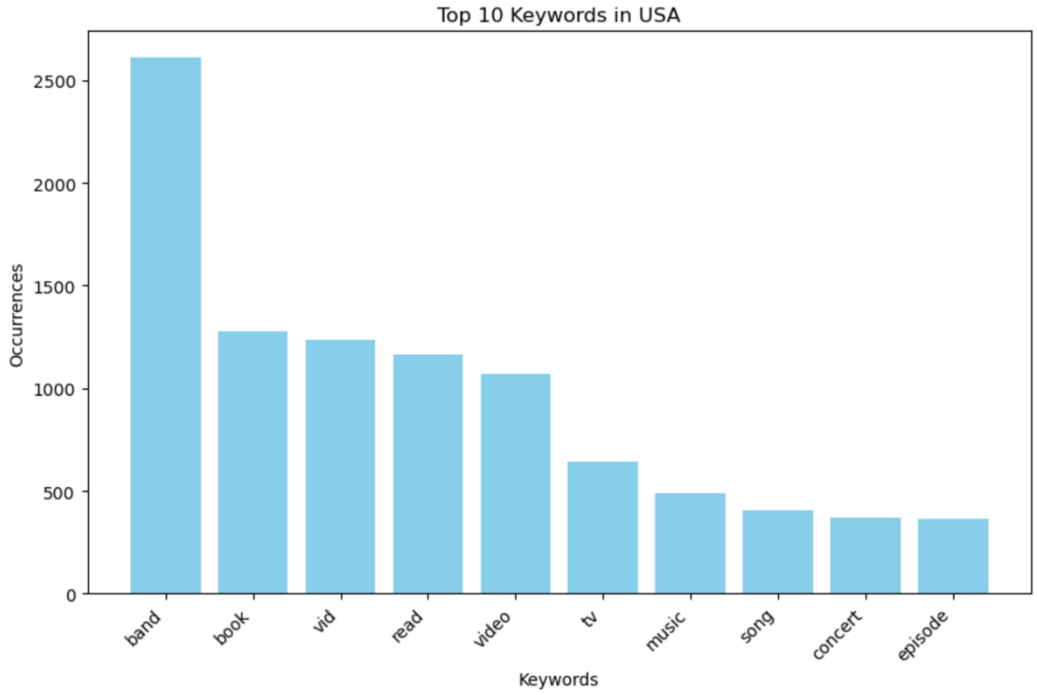
Bar Graph Analysis for Males: The second image is a bar graph that appears to quantify the frequency of keywords mentioned in the pie chart. The 'video' category stands out with the highest count, correlating with the pie chart's indication of its popularity. This is followed by significant counts for 'book' and 'read.' The bars vary widely in height, suggesting significant differences in the frequency of different keywords being mentioned.

Pie Chart Analysis for Females: The third image shows a similar pie chart for females. Here, 'music' is overwhelmingly the largest category with 37.8%, followed by 'video' at 11.4% and 'book' at 11.1%. This indicates a more concentrated interest among females in music-related entertainment, which is a stark contrast to the more evenly distributed interests of males.

Bar Graph Analysis for Females: The final image is a bar graph for females, showing the counts of keywords. The 'music' category dominates, aligning with the pie chart data, indicating that females mention music significantly more than any other entertainment category. The bar for 'music' is substantially taller than the others, which have much lower counts.

From these graphs, we can infer that males have a more varied distribution of entertainment interests with a slight preference for video content, while females show a strong preference for music. Both genders show interest in books and reading, but these interests are more balanced among males compared to the pronounced preference for music among females. The graphical representations suggest that there might be gender-based differences in discussing entertainment choices, with females possibly having more conversations around music, while males have a broader range of topics with less dominance by a single category.

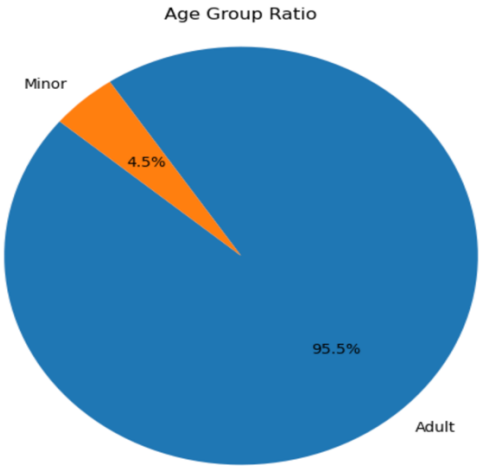
Question 2: what is the top 10 entertainment in USA



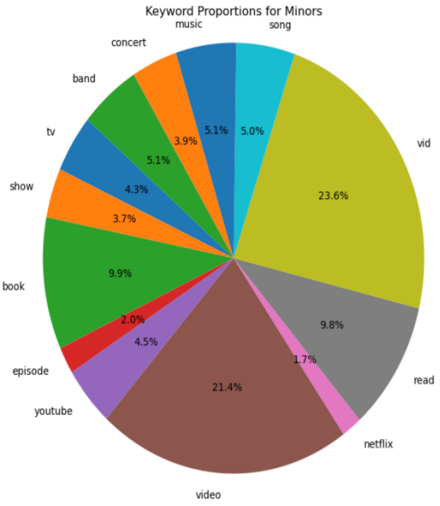
（In order to analyze this problem, we first need to further simplify the analysis text so that it only retains the data showing "USA" in the country column, and then analyze it） The provided bar graph titled "Top 10 Keywords in USA" displays the frequency of occurrence for various entertainment-related keywords. "Band" is the most frequently occurring keyword, with just over 2500 mentions, indicating a high level of interest or discussion around music groups. This is followed by "book," "vid," and "read," each with occurrences between approximately 1500 and 2000, suggesting a strong interest in reading and video content. "Video," "tv," "music," "song," "concert," and "episode" have lower frequencies, ranging from about 500 to 1000 occurrences. The distribution suggests a

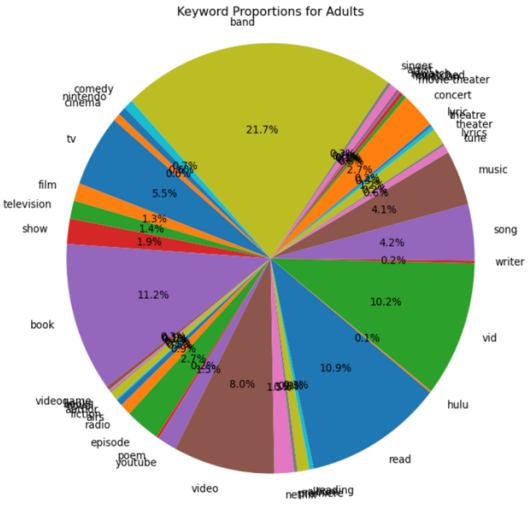
diverse range of entertainment interests in the USA, with a notable emphasis on music bands and literature.

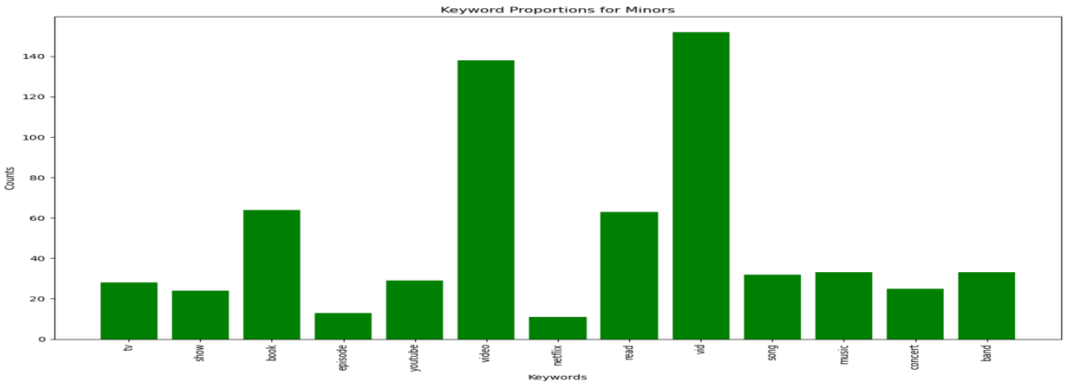
Question 3: What about the entertainment for adults and minors?

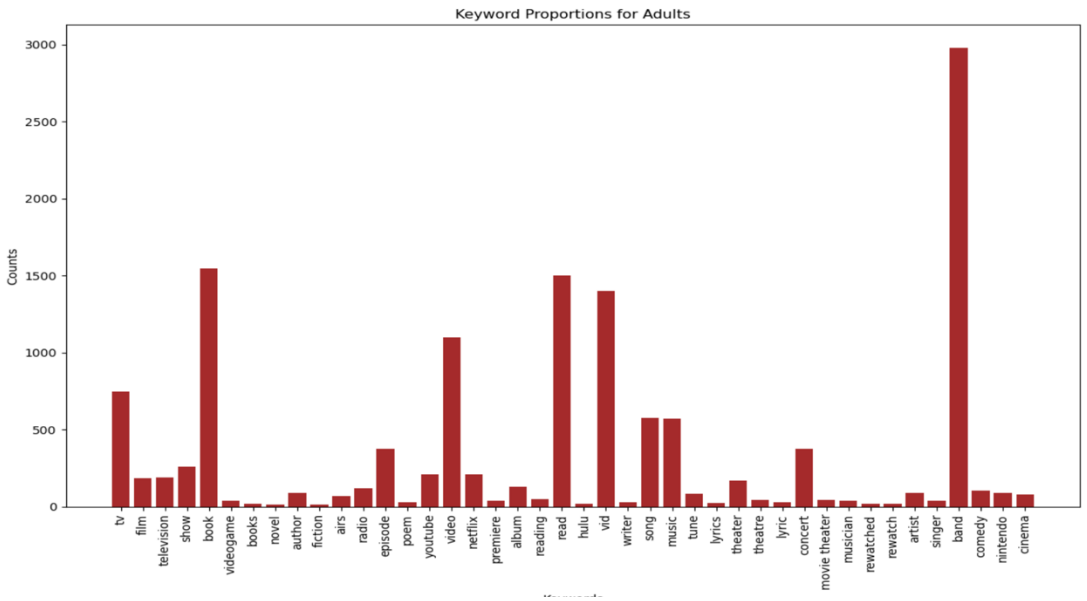


We first analyze the proportion of adults and minors in the sample to get this picture. We know that the sample data of minors is relatively small, and the subsequent statistical results may not be representative, but I will still be concerned about the minors. Analyze entertainment situation









For adults, the pie chart indicates that 'band' is the largest category, accounting for 21.7% of the keywords, which suggests that musical groups are a significant topic of interest. Other notable categories include 'read' and 'book,' which represent 10.9% and 11.2% respectively, indicating a healthy interest in reading. 'Music' and related terms like 'concert' and 'song' also have considerable proportions, reinforcing the strong interest in music-related activities.

The bar chart for adults complements this, with the 'band' category showing the highest count, significantly more than any other category. 'Read' and 'book' also have high counts, supporting the pie chart data. The distribution of counts across different entertainment categories suggests adults have a varied range of interests with a particular emphasis on music and reading.

For minors, the pie chart demonstrates that 'vid' is the dominant category, comprising 23.6% of the keywords, suggesting that videos are a primary area of interest. This is followed by 'video' and 'Netflix,' which have 21.4% and 9.8% respectively, indicating a strong inclination towards streaming and video content. 'Book' accounts for 9.9%, which is substantial but less than the corresponding interest observed in adults.

The bar chart for minors shows a similar trend with 'vid' and 'video' having the highest counts. Unlike adults, minors seem to have a stronger focus on visual and streaming media, with less emphasis on reading and music.

In summary, the statistical analysis reveals that adults in the sample tend to have a broader distribution of interests with a particular focus on music and literature, while minors show a pronounced preference for video and streaming content. This could reflect generational differences in entertainment consumption, with minors perhaps more engaged with digital media and adults maintaining a strong connection with traditional forms such as music bands and books.