Data Preprocessing and Cleaning

In [1]:
```
'''
To better prepare the anime searching questions and their sequence, I have created the
    1. What is the anime genre distribution like from the csv file?
    2. What is the anime score distribution like from the csv file?
    3. What is the anime rating level distribution like from the csv file?
    4. What year or time period does the most anime come out?
'''
```

Out[1]:
```
'\nGuiding Questions: \n1. What kind of Genres would mostly likely to score high(above 8)?\n2. What are the top 10-rated TV anime after 2010?\n3. What are the top 10-rated anime movie from 2000 to 2010?\n'
```

In [2]:
```
#Load the anime data
import pandas as pd
df=pd.read_csv("anime.csv")
df
```

Out[2]:

	MAL_ID	Name	Score	Genres	English name	Japanese name	Type	Episodes	Aire
0	1	Cowboy Bebop	8.78	Action, Adventure, Comedy, Drama, Sci-Fi, Space	Cowboy Bebop	カウボーイビバップ	TV	26	Apr 1998 t Apr 2 199
1	5	Cowboy Bebop: Tengoku no Tobira	8.39	Action, Drama, Mystery, Sci-Fi, Space	Cowboy Bebop:The Movie	カウボーイビバップ 天国の扉	Movie	1	1-Sep-0
2	6	Trigun	8.24	Action, Sci-Fi, Adventure, Comedy, Drama, Shounen	Trigun	トライガン	TV	26	Apr 1998 t Sep 3 199
3	7	Witch Hunter Robin	7.27	Action, Mystery, Police, Supernatural, Drama, ...	Witch Hunter Robin	Witch Hunter ROBIN (ウイッチハンターロビン)	TV	26	Jul 2002 t Dec 2 200
4	8	Bouken Ou Beet	6.98	Adventure, Fantasy, Shounen, Supernatural	Beet the Vandel Buster	冒険王ビィト	TV	52	Sep 3 2004 t Sep 2 200
...	...	...	...	...	...	...	...	...	...

	MAL_ID	Name	Score	Genres	English name	Japanese name	Type	Episodes	Aire
17557	48481	Daomu Biji Zhi Qinling Shen Shu	Unknown	Adventure, Mystery, Supernatural	Unknown	盗墓笔记之秦岭神树	ONA	Unknown	Apr 2021 to
17558	48483	Mieruko-chan	Unknown	Comedy, Horror, Supernatural	Unknown	見える子ちゃん	TV	Unknown	2021 to
17559	48488	Higurashi no Naku Koro ni Sotsu	Unknown	Mystery, Dementia, Horror, Psychological, Supe...	Higurashi:When They Cry – SOTSU	ひぐらしのなく頃に卒	TV	Unknown	Jul, 202 to
17560	48491	Yama no Susume: Next Summit	Unknown	Adventure, Slice of Life, Comedy	Unknown	ヤマノススメ Next Summit	TV	Unknown	Unknow
17561	48492	Scarlet Nexus	Unknown	Action, Fantasy	Unknown	SCARLET NEXUS	TV	Unknown	Jul, 202 to

17562 rows × 35 columns

In [14]:
```
#Data cleaning
#Drop animes that have unknown scores,episodes, premiered year, source
df.loc[:, ~(df == 'Unknown').any()]
df.drop(df.loc[df['Genres'] == 'Unknown'].index, inplace=True)
df.drop(df.loc[df['Score'] == 'Unknown'].index, inplace=True)
df.drop(df.loc[df['Premiered'] == 'Unknown'].index, inplace=True)
df.drop(df.loc[df['Source'] == 'Unknown'].index, inplace=True)
```

In [15]:
```
#separate the production year from the Premiered column
df["Year"] = df["Premiered"].str.extract('(\d+)').astype(int)
df["Year"].head()
#create a column in the dataframe that has pre
df.to_csv("anime_updated.csv")
df
```

Out[15]:

	MAL_ID	Name	Score	Genres	English name	Japanese name	Type	Episodes	Aired	Premiered
0	1	Cowboy Bebop	8.78	Action, Adventure, Comedy, Drama, Sci-Fi, Space	Cowboy Bebop	カウボーイビバップ	TV	26	Apr 3, 1998 to Apr 24, 1999	Spring 1998

	MAL_ID	Name	Score	Genres	English name	Japanese name	Type	Episodes	Aired	Premiered
2	6	Trigun	8.24	Action, Sci-Fi, Adventure, Comedy, Drama, Shounen	Trigun	トライガン	TV	26	Apr 1, 1998 to Sep 30, 1998	Spring 1998
3	7	Witch Hunter Robin	7.27	Action, Mystery, Police, Supernatural, Drama, ...	Witch Hunter Robin	Witch Hunter ROBIN (ウイッチハンターロビン)	TV	26	Jul 2, 2002 to Dec 24, 2002	Summer 2002
4	8	Bouken Ou Beet	6.98	Adventure, Fantasy, Shounen, Supernatural	Beet the Vandel Buster	冒険王ビィト	TV	52	Sep 30, 2004 to Sep 29, 2005	Fall 2004
5	15	Eyeshield 21	7.95	Action, Sports, Comedy, Shounen	Unknown	アイシールド21	TV	145	Apr 6, 2005 to Mar 19, 2008	Spring 2005
...	...	...	...	...	...	...	...	...	...	...
17178	42941	Uma Musume: Pretty Derby (TV) Season 2	7.21	Slice of Life, Comedy, Sports	Unknown	ウマ娘プリティーダービー Season 2	TV	13	Jan 5, 2021 to ?	Winter 2021
17224	43299	Wonder Egg Priority	8.32	Psychological, Drama, Fantasy	Unknown	ワンダーエッグ・プライオリティ	TV	12	Jan 13, 2021 to ?	Winter 2021
17229	43350	Gebäude Bäude	6.33	Sci-Fi, Comedy	Unknown	ゲボイデ＝ボイデ	TV	10	Nov 8, 2020 to Dec 10, 2020	Fall 2020

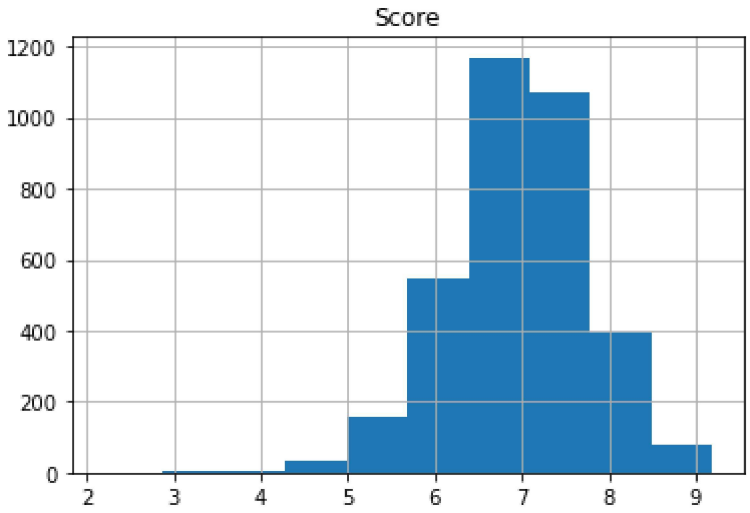
	MAL_ID	Name	Score	Genres	English name	Japanese name	Type	Episodes	Aired	Premiered
17328	44044	Jimihen!!: Jimiko wo Kaechau Jun Isei Kouyuu!!	6.12	Romance, Ecchi	Unknown	じみへんっ!!~地味子を変えちゃう純異性交遊~	TV	8	Jan 4, 2021 to Feb 22, 2021	Winter 2021
17469	46118	Wave!!: Surfing Yappe!! (TV)	6.05	Slice of Life, Sports	WAVE!! - Let's go surfing!!-	WAVE!!~サーフィンやっぺ!!~	TV	12	Jan 12, 2021 to ?	Winter 2021

3456 rows × 36 columns

Data Visualization

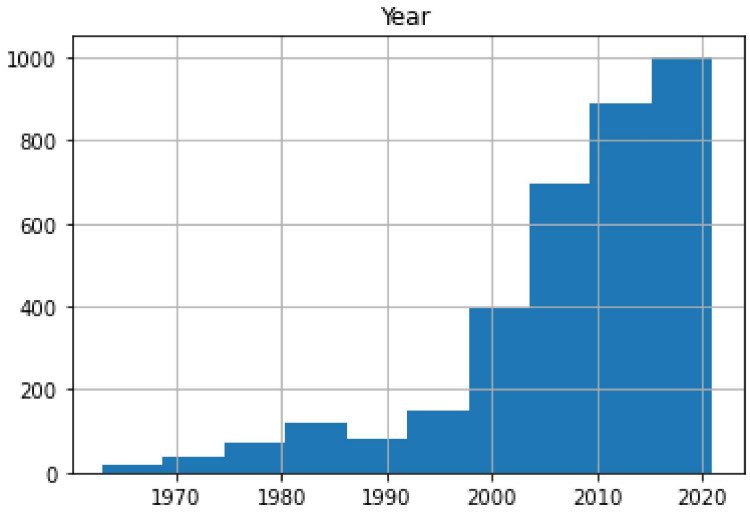
In [8]:
```
#Histogram for Score
df["Score"]=df["Score"].astype(float)
#Histogram
df.hist(column="Score")
```

Out[8]:
```
array([[<AxesSubplot:title={'center':'Score'}>]], dtype=object)
```



In [5]:
```
#Histogram for Year
df.hist(column="Year")
```

Out[5]:
```
array([[<AxesSubplot:title={'center':'Year'}>]], dtype=object)
```



In [13]:
```
#Histogram for Source
source= df["Source"].explode().value_counts()
source
```

Out[13]:
```
Manga            1335
Original          848
Light novel       340
Game              234
Visual novel      155
4-koma manga      132
Novel             124
Other             100
Web manga          79
Card game          40
Book               32
Music              14
Picture book       13
Digital manga       7
Radio               3
Name: Source, dtype: int64
```

In [10]:
```
#Counts for Rating
rating= df["Rating"].explode().value_counts()
rating
```

Out[10]:
```
PG-13 - Teens 13 or older         2051
R - 17+ (violence & profanity)     457
G - All Ages                       402
R+ - Mild Nudity                   259
PG - Children                      255
Unknown                             32
Name: Rating, dtype: int64
```

In [7]:
```
#Genres
genre= df["Genres"].str.split(", ").explode().value_counts()
genre
```

Out[7]:
```
Comedy           1672
Action           1175
Drama             801
```

```
Adventure         782
Fantasy           778
Romance           755
Shounen           747
Sci-Fi            721
School            681
Slice of Life     635
Supernatural      516
Magic             374
Mecha             315
Seinen            313
Ecchi             301
Mystery           282
Shoujo            273
Historical        246
Sports            240
Harem             203
Super Power       188
Military          158
Music             154
Demons            141
Kids              139
Game              135
Psychological     134
Space             126
Horror            119
Parody            111
Martial Arts       97
Samurai            61
Vampire            57
Police             57
Josei              53
Thriller           47
Shoujo Ai          41
Shounen Ai         29
Cars               22
Dementia           17
Name: Genres, dtype: int64
```

Anime Categorization

In [16]:

```
# get user preference for genre
Action= df[df["Genres"].str.contains("Action")]
Adventure = df[df["Genres"].str.contains("Adventure")]
Comedy = df[df["Genres"].str.contains("Comedy")]
Drama = df[df["Genres"].str.contains("Drama")]
SciFi = df[df["Genres"].str.contains("Sci-Fi")]
Adventure = df[df["Genres"].str.contains("Adventure")]
```

In [17]:

```
#get user preference for anime type
#pick TV as preference
TV= df[df["Type"].str.contains("TV")]
Movie= df[df["Type"].str.contains("Movie")]
TV= df[df["Type"].str.contains("TV")]
Special= df[df["Type"].str.contains("Special")]
```

Find the top 10 genre, type, year of production

In [18]:
```
#find the most popular genre
df_popular_genre=pd.Series(' '.join(df['Genres']).lower().split(",")).value_counts()[:1
df_popular_genre
```

Out[18]:
```
comedy          825
romance         642
drama           568
fantasy         469
adventure       425
sci-fi          424
school          416
magic           306
supernatural    276
ecchi           238
dtype: int64
```

In [19]:
```
#find the most popular year of production
df_popular_year=pd.Series(' '.join(df['Premiered']).lower().split()).value_counts()[:10
df_popular_year
```

Out[19]:
```
spring    1107
fall      1028
winter     704
summer     617
2018       219
2016       217
2017       207
2015       185
2014       180
2019       169
dtype: int64
```

Fun question: What kind of genre would be most likely to receive a high score?

In [20]:
```
#convert data type from object to float
df["Score"]=df["Score"].astype(float)
df["Score"].dtypes
#filter out the anime that receive score above 8

df["high score"]=df["Score"]>= 8.0
df["high score"]
groupdf =  df.groupby("Genres")
groupdf
#df["high_score_df"]=df["Genres"]["high score"].groupby()
#df["high score"]
#df.loc[df["Score"]>8,"Name"]
```

Out[20]:
```
<pandas.core.groupby.generic.DataFrameGroupBy object at 0x0000026317D748E0>
```

In [21]:

```
# filter out the score that is greater then 8
groupdf =  df[df.Score>8]
#group the anime name and anime genres together
groupdf = groupdf[['Name','Genres']]
groupdf
#find the genre that appears the most(top 5)
groupdf['Genres'].value_counts().idxmax()
```

Out[21]: 'Slice of Life, Demons, Supernatural, Drama, Shoujo'