Junfan Chen
IST652
MiniProject1

Analysis Based on TMDB 5000 Movie Dataset

I. Introduction

In this project, I will explore a dataset from Kaggle (https://www.kaggle.com/tmdb/tmdb-movie-metadata). And dealing with the following questions with python:

1. Compute the total number of shots in each movie category, plot the number of the five types of movies with the most shots over time, and calculate the total average score of these five genres.

2. Calculate the total output of films produced by Universal Pictures, Columbia Pictures, Warner Bros., and Paramount Pictures, and compare the profits of their based on novel films and those are not based on.

II. Method

i. Preparing Data

There are two csv files in this Kaggle project, so the first thing I did was to join them together, then I got a dataframe with 4803 rows and 23 columns.

```
Data columns (total 23 columns):
 #   Column                Non-Null Count  Dtype
---  ------                --------------  -----
 0   id                    4803 non-null   int64
 1   title_x               4803 non-null   object
 2   cast                  4803 non-null   object
 3   crew                  4803 non-null   object
 4   budget                4803 non-null   int64
 5   genres                4803 non-null   object
 6   homepage              1712 non-null   object
 7   keywords              4803 non-null   object
 8   original_language     4803 non-null   object
 9   original_title        4803 non-null   object
 10  overview              4800 non-null   object
 11  popularity            4803 non-null   float64
 12  production_companies  4803 non-null   object
 13  production_countries  4803 non-null   object
 14  release_date          4802 non-null   object
 15  revenue               4803 non-null   int64
 16  runtime               4801 non-null   float64
 17  spoken_languages      4803 non-null   object
 18  status                4803 non-null   object
 19  tagline               3959 non-null   object
 20  title_y               4803 non-null   object
 21  vote_average          4803 non-null   float64
 22  vote_count            4803 non-null   int64
dtypes: float64(3), int64(4), object(16)
```

Figure 1. overview of the origin dataframe

The next step was data preprocessing, and I decided to begin with removing the duplicate columns and the columns which were not related to the questions I need to handle. With observation, I dropped the columns 'original_title', 'title_x', 'id', 'cast', 'crew', 'homepage', 'original_language', 'spoken_languages','overview', and 'tagline', and renamed the column 'title_y' as 'title'.

```
Data columns (total 12 columns):
 #   Column                Non-Null Count  Dtype
---  ------                --------------  -----
 0   budget                4803 non-null   int64
 1   genres                4803 non-null   object
 2   keywords              4803 non-null   object
 3   popularity            4803 non-null   float64
 4   production_companies  4803 non-null   object
 5   production_countries  4803 non-null   object
 6   release_date          4802 non-null   object
 7   revenue               4803 non-null   int64
 8   status                4803 non-null   object
 9   title                 4803 non-null   object
 10  vote_average          4803 non-null   float64
 11  vote_count            4803 non-null   int64
dtypes: float64(2), int64(3), object(7)
```

Figure 2. overview of the dataframe in this step

And the following step is dealing with null values in column 'release_date', and my solution was dropping the line with null value under 'release_date'. When further observing the data, I found the data type of some columns need to be changed: data in columns 'genres', 'keywords', 'production_companies', and 'production_countries' are in json format, and data under 'release_date' need to be changed into datetime.

```
Data columns (total 13 columns):
 #   Column                Non-Null Count  Dtype
---  ------                --------------  -----
 0   budget                4802 non-null   int64
 1   genres                4802 non-null   object
 2   keywords              4802 non-null   object
 3   popularity            4802 non-null   float64
 4   production_companies  4802 non-null   object
 5   production_countries  4802 non-null   object
 6   release_date          4802 non-null   datetime64[ns]
 7   revenue               4802 non-null   int64
 8   status                4802 non-null   object
 9   title                 4802 non-null   object
 10  vote_average          4802 non-null   float64
 11  vote_count            4802 non-null   int64
 12  year                  4802 non-null   int64
dtypes: datetime64[ns](1), float64(2), int64(4), object(6)
```

Figure 3. overview of the final version dataframe

ii. Modeling

Question 1:

To answer this question, I only need to focus on columns 'year', 'genres', and 'vote_average', so I build a new dataframe only contains these three columns.

Junfan Chen
IST652
MiniProject1

	year	genres	vote_average
0	2009	Action,Adventure,Fantasy,Science Fiction	7.2
1	2007	Adventure,Fantasy,Action	6.9
2	2015	Action,Adventure,Crime	6.3
3	2012	Action,Crime,Drama,Thriller	7.6
4	2012	Action,Adventure,Science Fiction	6.1

Figure 4. dataframe for this question

In order to do the statistics, I need to extract all the genres from the 'genres' column.

```
Data columns (total 23 columns):
 #   Column           Non-Null Count  Dtype
---  ------           --------------  -----
 0   year             4802 non-null   int64
 1   genres           4802 non-null   object
 2   vote_average     4802 non-null   float64
 3   History          4802 non-null   int64
 4   Western          4802 non-null   int64
 5   TV Movie         4802 non-null   int64
 6   Adventure        4802 non-null   int64
 7   Thriller         4802 non-null   int64
 8   Science Fiction  4802 non-null   int64
 9   Mystery          4802 non-null   int64
 10  Crime            4802 non-null   int64
 11  War              4802 non-null   int64
 12  Animation        4802 non-null   int64
 13  Fantasy          4802 non-null   int64
 14  Horror           4802 non-null   int64
 15  Documentary      4802 non-null   int64
 16  Drama            4802 non-null   int64
 17  Foreign          4802 non-null   int64
 18  Family           4802 non-null   int64
 19  Romance          4802 non-null   int64
 20  Comedy           4802 non-null   int64
 21  Music            4802 non-null   int64
 22  Action           4802 non-null   int64
dtypes: float64(1), int64(21), object(1)
```

Figure 5. overview of the dataframe in this step

After that, I perform a sum() function to compute the total number of each genres, and find 'Drama', 'Comedy', 'Thriller', 'Action', and 'Romance' are the top five. So, I will focus on these five genres to do the further exploration.

```
Drama              2297
Comedy             1722
Thriller           1274
Action             1154
Romance             894
Adventure           790
Crime               696
Science Fiction     535
Horror              519
Family              513
Fantasy             424
Mystery             348
Animation           234
History             197
Music               185
War                 144
Documentary         110
Western              82
Foreign              34
TV Movie              8
```

Figure 6. genres sorted by number of shots

Question 2:

As for this question, I only need to use columns 'production_companies', 'keywords', 'budget', and 'revenue'. Like the previous question, I need to extract the word I need from columns 'production_companies' and 'keywords'. After finishing that, I can do the further mining.

	budget	revenue	based on novel	Universal Pictures	Columbia Pictures	Warner Bros.	Paramount Pictures
0	237000000	2787965087	0	0	0	0	0
1	300000000	961000000	0	0	0	0	0
2	245000000	880674609	1	0	1	0	0
3	250000000	1084939099	0	0	0	1	0
4	260000000	284139100	1	0	0	0	0

Figure 7. dataframe in this step

III. Results & Conclusion

Question 1:

To find the annual trend of 'Drama', 'Comedy', 'Thriller', 'Action', and 'Romance', I firstly called groupby() function to group them by year, then used lineplot() function in seaborn to draw the plot.

Junfan Chen
IST652
MiniProject1

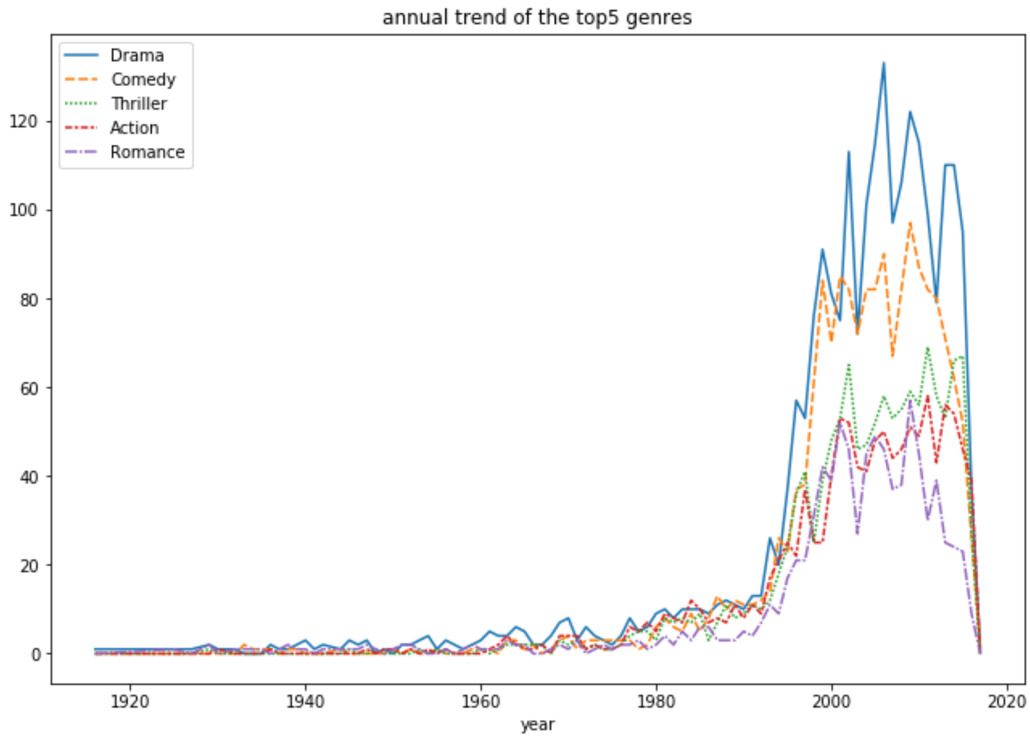

Figure 8. annual trend of the top5 genres

As the plot shown, the number of movies surged after 1980. Drama film reached its peak between 2000 and 2010, while other films reached their peak around 2010. The number of Romance and Comedy generally went down after their peak, and the number of Drama movies went through several big fluctuations, but the number of Thriller and Action seems more stable. It reflects that the market for Thriller movies and Action movies is more stable than its 3 types of movies.
And the next part of this question is to divide total vote by total number to calculate the average vote of these five genres.

```
Drama       6.388594
Comedy      5.945587
Thriller    6.010989
Action      5.989515
Romance     6.207718
```

Figure 9. average vote of top 5 genres

Which shows that people are more likely to give a high score to Drama movies and it also reflects the good quality of such movies at the same time.

Junfan Chen
IST652
MiniProject1

Question 2:

In order to find the total output of films produced by Universal Pictures, Columbia Pictures, Warner Bros., and Paramount Pictures, I firstly call the sum() function, then used barplot() function in seaborn to draw the plot.

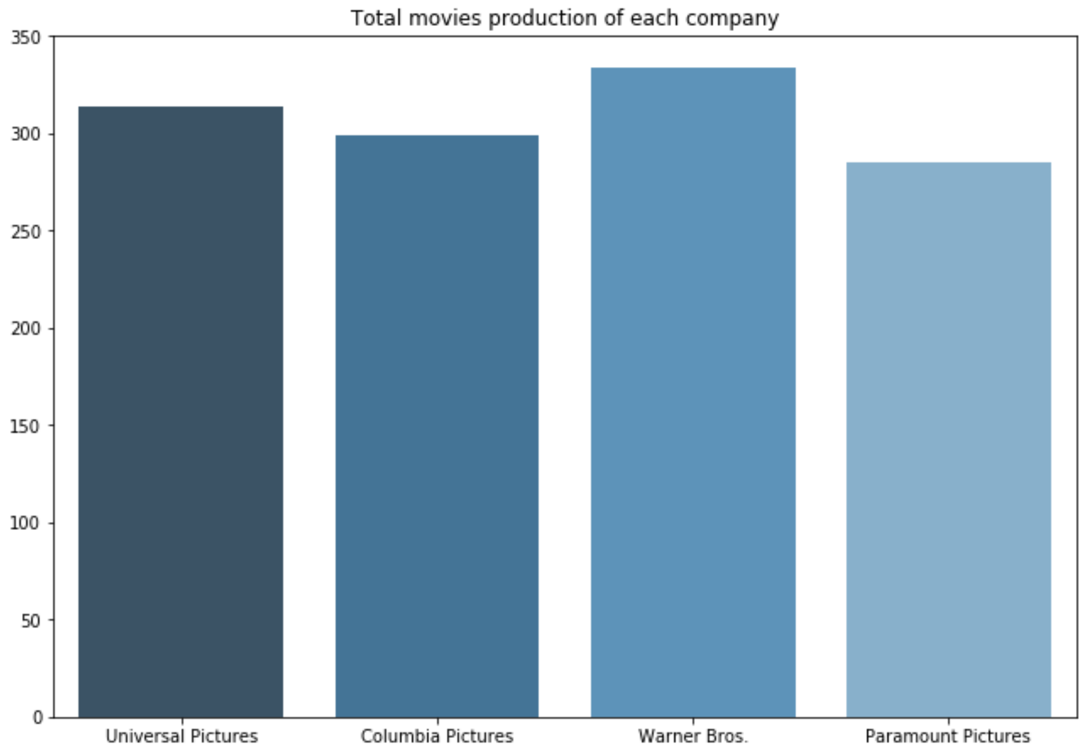

Figure 10.1. bar plot of each company's film production

```
Universal Pictures    314
Columbia Pictures     299
Warner Bros.          334
Paramount Pictures    285
```

Figure 10.2. total movies production of each company

As the plot and the calculate output shown, the total film production of these four companies is not much different, and Warner Bros.has the most movies production among them.

And the next part of this question is to compare the profits (profit = revenue - budget) of their based on novel films and those were not based on.

Table 1. comparison of each company

campany	profit of 'based on novel'	number of 'based on novel'	profit of 'not based on novel'	number of 'not based on novel'
Universal Pictures	920423828	9	28069924768	305
Columbia Pictures	2223450867	15	20559873778	284
Warner Bros.	1110535922	17	32052201795	317
Paramount Pictures	1325626259	13	26243283899	272

Junfan Chen
IST652
MiniProject1

According to the table above, each company produces more ‘not based on novel’ movies than ‘based on novel’ movies, which is why the total profit of based on novel films are less than those are not based on.