Egyptian Movies

Introduction

This is a report for the analysis of a dataset about media available on the website "elcinema" in Egypt. The main purpose of the project is to analyze the dataset in order to identify all possible trends. This was done via visualizing the data to derive conclusions about the different patterns.

Note:

This report describes the identified trends in the dataset and explains the graphs found in the appendix. For the exact numbers of any of the trends, kindly refer to the Jupyter Notebook "Egyptian Movies" which also contains the python codes used throughout the entire project.

Acknowledgments

Authors: Nour Abouseif, Katia Gabriel, Masa Tantawy, Bassel Samy, Mina Kallini.

This project was done by a group of 5 undergraduate students at The American University in Cairo for the course Fundamentals of Data Science II. The authors would like to thank the course professor, Dr Dia Taha, for his enormous efforts and support.

Table of contents

Variables

The dataset originally consisted of 4133 observations and 5 variables. The five variables are:

1. **ID:** a unique numeric identifier for each observation. IDs cannot be repeated. [Numerical Variable]
2. **Year:** the year of publication for the observation. [Numerical Variable]
3. **Type:** the category of the observation, such as movie or series. [Qualitative]
4. **Genre:** the genre of the observation such as comedy or drama. Each observation can have multiple genres. [Multi-valued categorical variable]
5. **Title:** the designated name of the observation. [Qualitative]

Data Wrangling

The first process was data wrangling. There was a typing error in the variable type, which was "*series*" instead of *series*. Observations with this error were adjusted to have the correct type. In addition, the dataset contained repeated observations which were identified through their IDs. Observations with duplicate ID were eliminated from the dataframe, ensuring that each observation is present only once. Lastly, NaNs in genre were changed to missing to allow categorisation of observations with no specific genre.

Trends A: Media Type

1. **How many media types are represented in the dataset?**

The dataset contained 14 types, varying in frequency. In decreasing order of frequency, the 14 types are: *series, movie, program, play, short, radio, sitcom, cartoon, tv, documentary, mini, riddle, seven, and opera*.

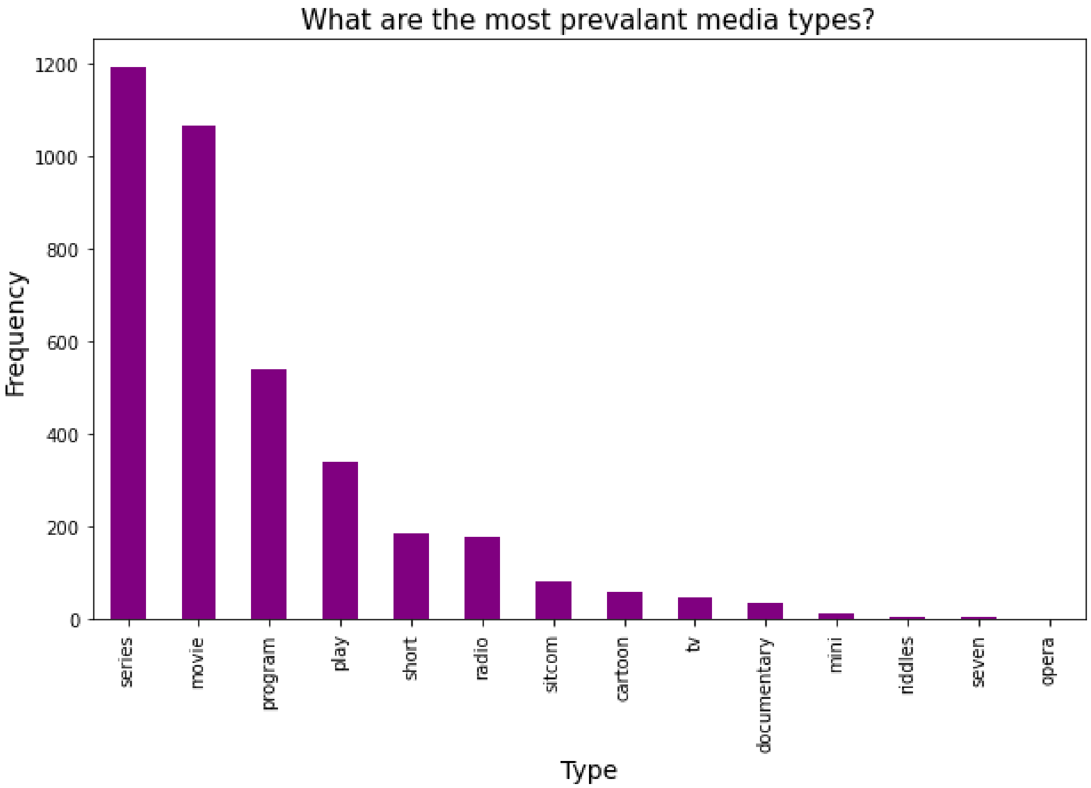

Figure 1

Figure 1 illustrates the number of occurrences of each type. The most common type is *series* followed by *movies*. On the other hand, the least frequent type among the observations is *Operas* as there were only 2 observations with this type. Both operas were arabic and were relatively old, for they were not produced in the last decade.

They are:

- أوبريت القدس ح ترجع لنا produced in year 2000
- ملكة القطن و الشمس produced in year 2010

Three types were unclear, so there was a need for clarification. *Short* is a type that refers to short movies. *Mini* is a type that refers to short sitcoms, and lastly, *seven* is a type that refers to a series or programme or any media consisting of exactly 7 episodes or parts.

2. **How have media types changed over the years?**

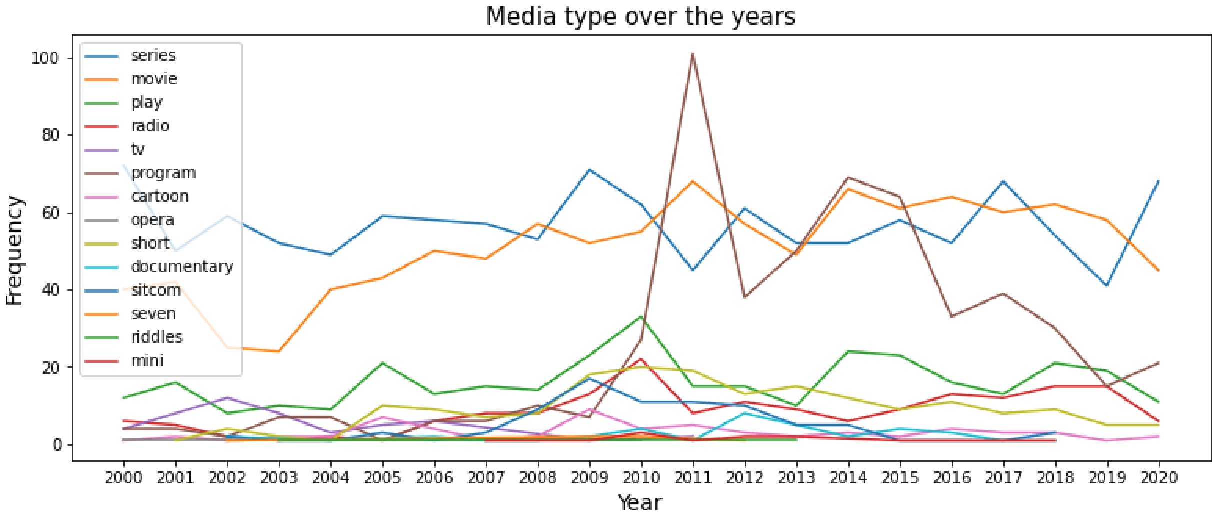

Figure 2

Figure 2 shows how the frequency of different types changed from the year 2000 to 2020. The graph depicts that *series* and *movies* have the highest numbers produced over all the years.

One notable trend is that starting 2019, the number of *series* inclined, whereas *movies* experienced a slight decline in number. This change in production can be explained by the current COVID-19 pandemic, which led to the closing of cinemas due to the mandated lockdown and health safety measures. As a result, movies are less suitable while series became more appealing as they can be watched at home.

Another trend visible in the above graph is the rapid increase in the number of *programs* in the years 2011 and 2013. The reason for this might be due to the fact that in these 2 years, Egypt experienced the 2011 Revolution followed by political turbulence until 2013, meaning that there was a high need for programs to cover the ongoing events in the country.

3. Which media types have the most missing data?

The new dataframe (after data wrangling) did not contain missing data. This is because missing data were only in genre after removing repeated observations. Observations with no genre were categorized to have genre.

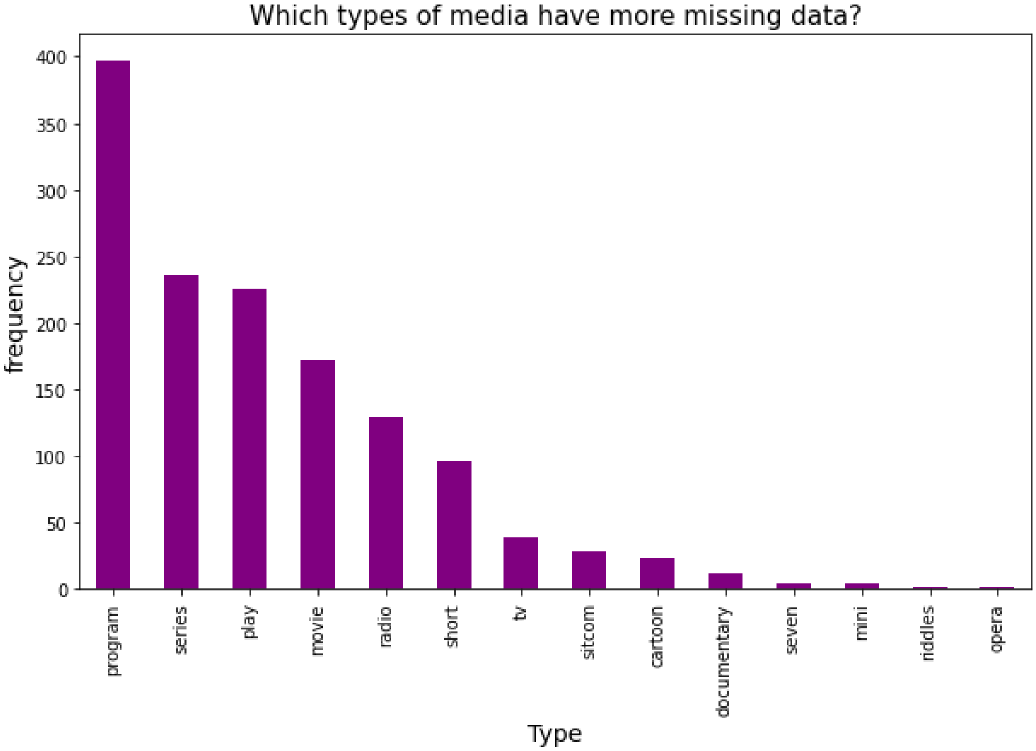

Figure 3

Figure 3 shows the number of observations with missing data among each type. *Programs* have the highest number of missing data, as there were almost 400 observations. *Series* and *plays* followed, for each had over 200 observations, then *movies* and *radios* each with over 100 observations. The media types with the least missing data are operas with 2 observations with missing, followed by sevens and minis with only 4 observations with missing values.

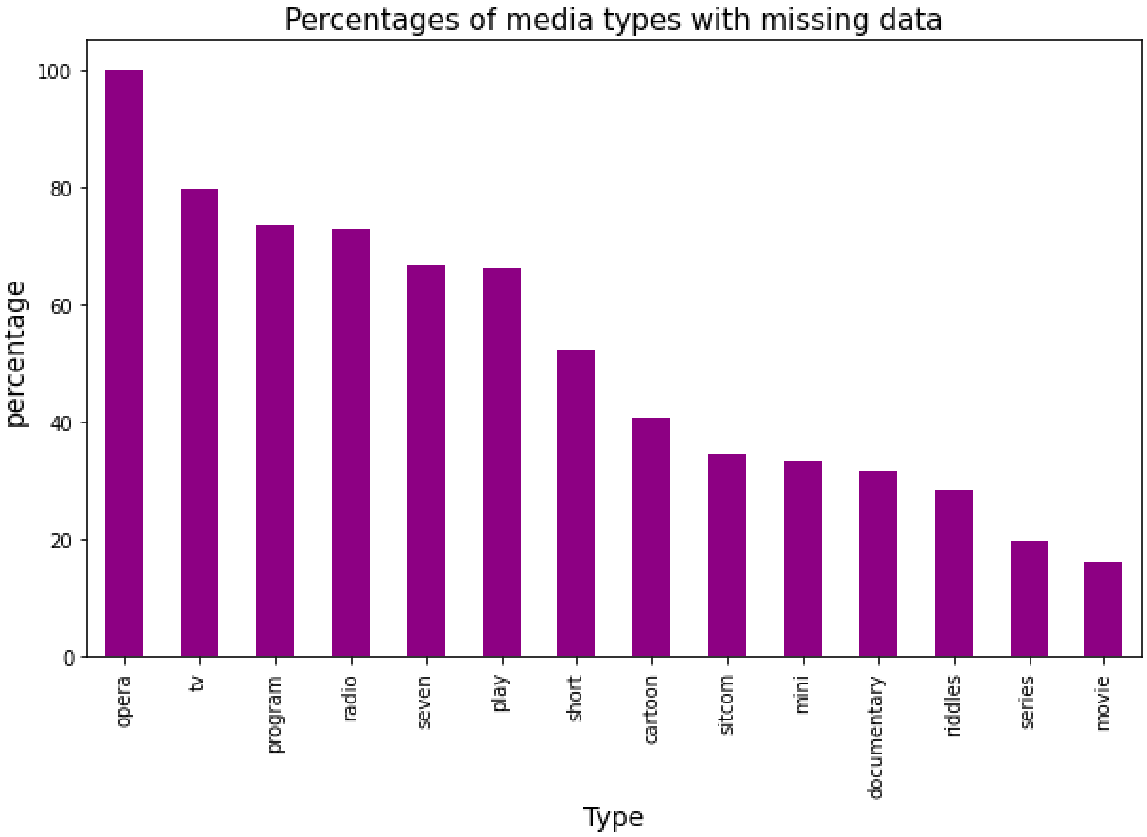

Figure 4

Comparing the numbers relative to the counts of observations for each type, figure 4 depicts the percentages of media types with missing data. The type with the highest percentage of missing data is *operas,* for 100% of the observations have no genre. *TVs*, *programs* and *radios* follow as around 75% of these types contain missing data.

Trends B: Genre

1. **How many genres are presented in the dataset?**

As genre is a multivalued categorical variable, multiple observations have more than one genre. If an observation has no genre, its genre is categorized as missing. In total, there are 21 unique genres. These genres are: *Documentary, Action, Adventure, Animation, Biography, Comedy, Crime, Drama, Family, Fantasy, History, Horror, Musical, Mystery, Religious, Romance, Science Fiction, Short, Sport, Thriller, War*.

2. **How do genres change over the years?**

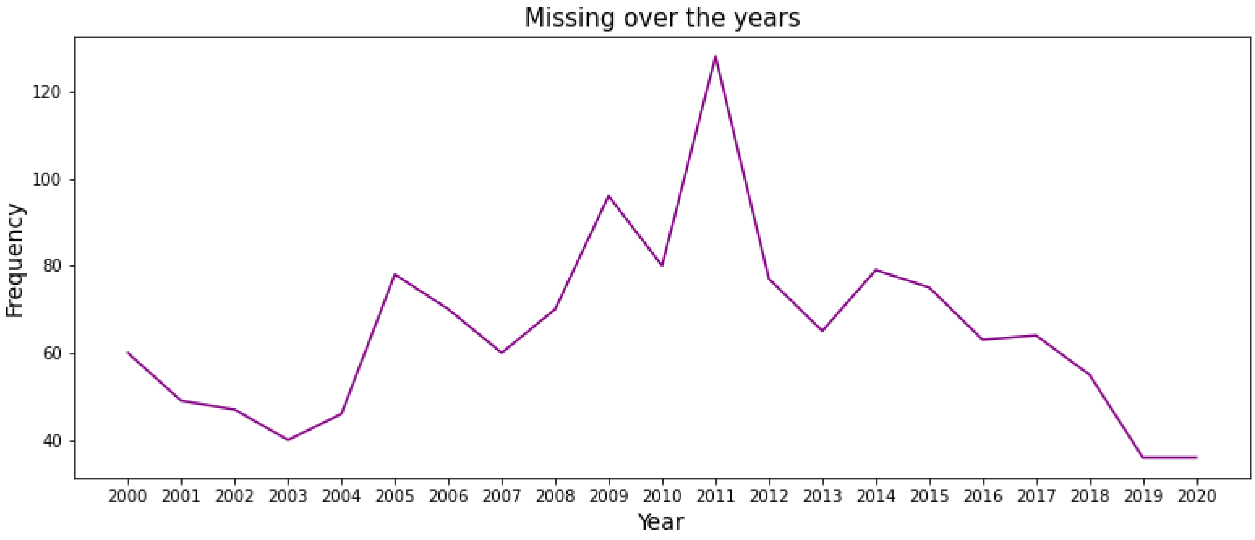

Figure 5

Primarily, observations with the genre *missing* were found to be highest in 2011 as shown in figure 5. However, they experienced a drastic decrease ever since.

Next, the number of occurrences for each genre among all the observations across the years was observed.

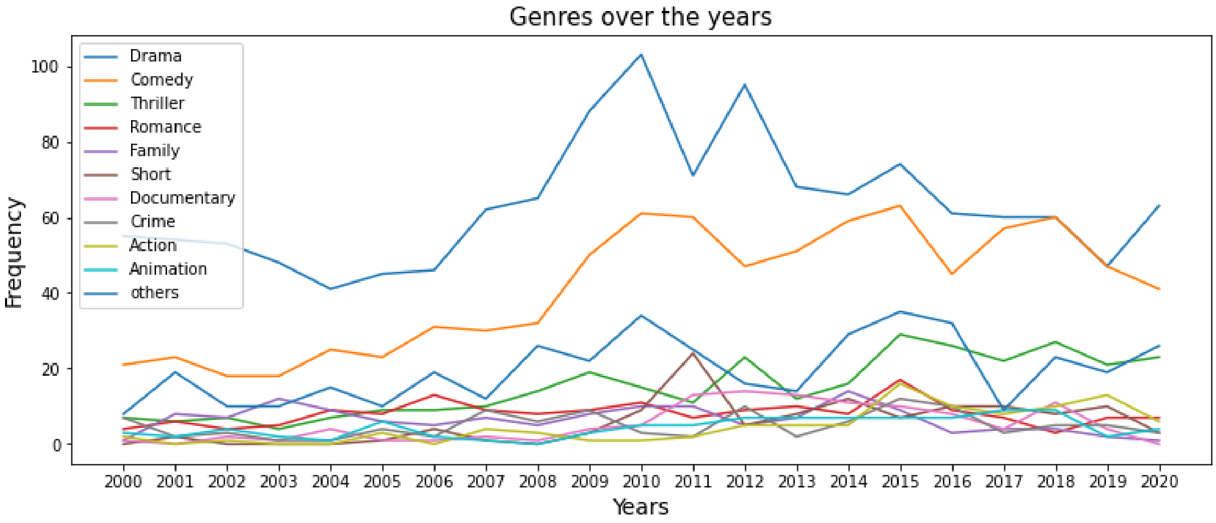

Figure 6

It is conspicuous that *drama* followed by *comedy* are the most popular genres in all years. *Thriller, romance* and *family* are the genres that follow in abundance. Two visible patterns in the line graph are: first, *thriller* started gaining popularity, for the number of observations with this genre began increasing. Second, *short* peaked in 2011, yet it quickly dropped again in the following year.

Trends C: The correlation between type and genre

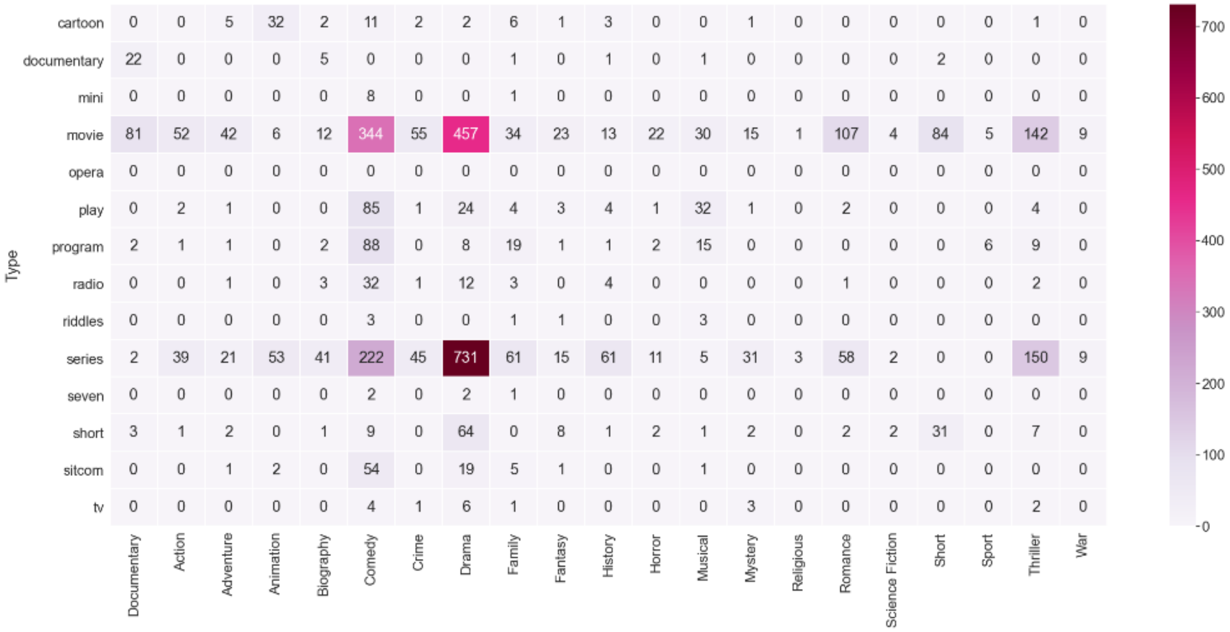

Figure 7

Figure 7 is a heatmap that displays the correlation between all genres and media types, depicting the amounts of genre for each media type. This plot is particularly useful as the upcoming trends focus on the correlation between only 4 types and associated genres.

From the heatmap, it is evident that the most prevalent combination of genres and types is *drama with series*. This could be due to the fact that dramatic media benefits from techniques such as the use of cliffhangers to engage the audience and cliffhangers are better suited to a media type with multiple episodes, to leave the audience waiting for the next episode. The second most common genre was *comedy* which leaned more towards *movies* rather than series.

1. Genres for movies

Among *movies*, the top 4 genres are drama, comedy, thriller, and romance. It can be seen from the pie chart named figure 7a (See Appendix 1 for the graph) that more than a quarter of the movies are categorized to have the genre drama or comedy. This type is one of the highest in adopting the genre *comedy* compared to other types.

2. Genres for series

As evident in figure 7b (See Appendix 2 for the graph), almost half the *series* are *drama*; this makes this genre the highest among this type, leading to *series* being the highest in adopting the genre *drama* compared to other media types . Similarly to *movies*, *comedy* is the genre that follows in abundance in this type. The rest of the top 4 genres in this type are *thriller* and *history*.

3. Genres for plays

For plays, more than half the observations of this type are *comedy*. This is visible in figure 7c (See Appendix 3 for the graph). In decreasing order, *musical, drama* and *history* are the most popular genres for *plays*.

4. Genres for programs

Figure 7d (See Appendix 4 for the graph) depicts the top 4 genres among *programs*, which are *comedy, family, musical*, and *thriller*. More than half of the *programs* are categorized as *comedy*, like plays. The genre *family,* which was not visible among the top genres in the 3 previous types, is one of the most frequently found among this type,

Conclusion

In conclusion, almost all the graphs experienced a drastic shift in 2011. This shows the impact of the Egyptian Revolution/Arab Spring on the Egyptian entertainment industry. The impact of the pandemic has also been visible among a trend. Lastly, the popularity of the genres *comedy* and *drama* have not changed over the years, reflecting the nature of the Egyptian viewers.

Appendix 1

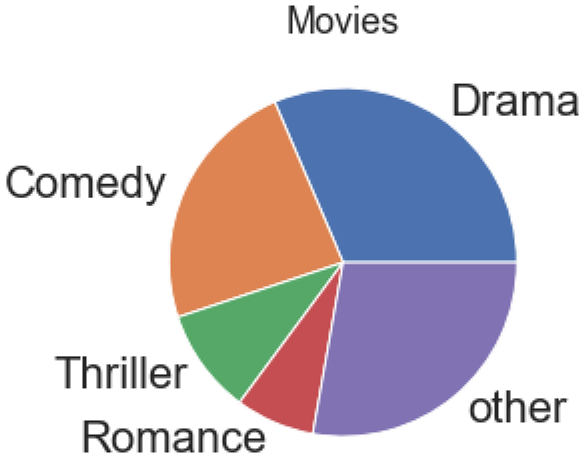

Figure 7a

Appendix 2

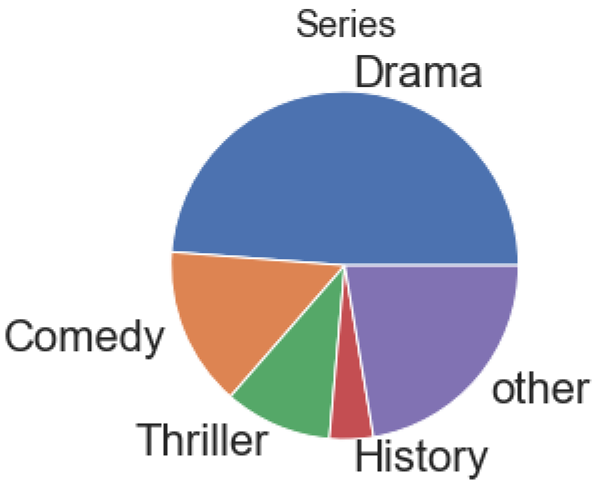

Figure 7b

Appendix 3

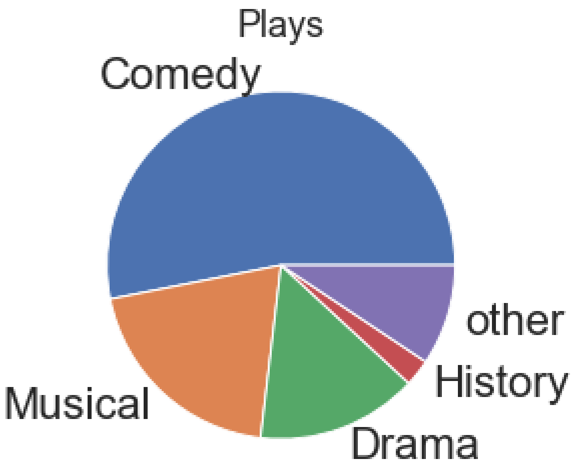

Figure 7c

Appendix 4

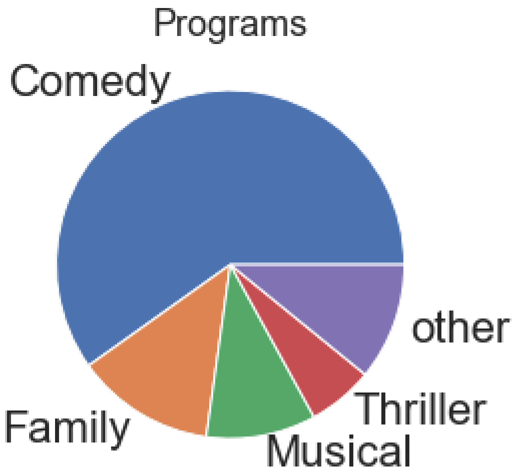

Figure 7d