PortfolioProject

December 6, 2023

1 Lab 11 - Visualization

1.1 The Question: What are the qualities of highly rated Netflix films?

1.1.1 This question could be very important to a movie producer trying to make it big in the movie industry. By comparing things like genre (fig 1), certificate (fig 2), and duration (fig 3) compared to how people liked it, you can figure out exactly what kind of movie would be most likely to become popular and make you more money.

1.1.2 This data analysis will focus specifically on prevoius highly rated movies. We will look at ratings compared to other traits the film holds, and try to find a connection.

[2]:

```python
from datascience import *
import numpy as np
import matplotlib.pyplot as plt
%matplotlib inline
plt.style.use("ggplot")
```

[3]:

```python
data = Table.read_table("n_moviesALLRAT.csv")
```

This dataset includes information about netflix movies from netflix themselves. It includes the title, year made, certificate, duration, rating, and a short description. This data could be compared to a list of most watched shows to find coorelations between them. For example, we may find in this set that a certain genre is rated higher, or maybe a certain duration period is rated higher than others. This dataset eliminates all rows where rating was null.

[4]:

```python
d = data.sort('rating', descending = True)
d.show(10)
```

```
<IPython.core.display.HTML object>
```

[5]:

```python
avg_com = np.average(d.where('genre', 'Comedy').column('rating'))
```

[6]:

```python
avg_ani = np.average(d.where('genre', 'Animation').column('rating'))
```

[7]:

```python
avg_Dra = np.average(d.where('genre', 'Drama').column('rating'))
```

```
[8]: avg_cri = np.average(d.where('genre', 'Crime').column('rating'))
```

```
[9]: avg_act = np.average(d.where('genre', 'Action').column('rating'))
```

```
[10]: avg_doc = np.average(d.where('genre', 'Documentary').column('rating'))
```

```
[11]: avg_mus = np.average(d.where('genre', 'Musical').column('rating'))
```

```
[12]: avg_hor = np.average(d.where('genre', 'Horror').column('rating'))
```

```
[13]: avg_thr = np.average(d.where('genre', 'Thriller').column('rating'))
```

```
[14]: avg_sci = np.average(d.where('genre', 'Sci-Fi').column('rating'))
```

```
[15]: avg_rea = np.average(d.where('genre', 'Reality-TV').column('rating'))
```

```
[16]: avg_bio = np.average(d.where('genre', 'Biography').column('rating'))
```

The following cell creates an array of all the average ratings for each genre and an array for the genre names

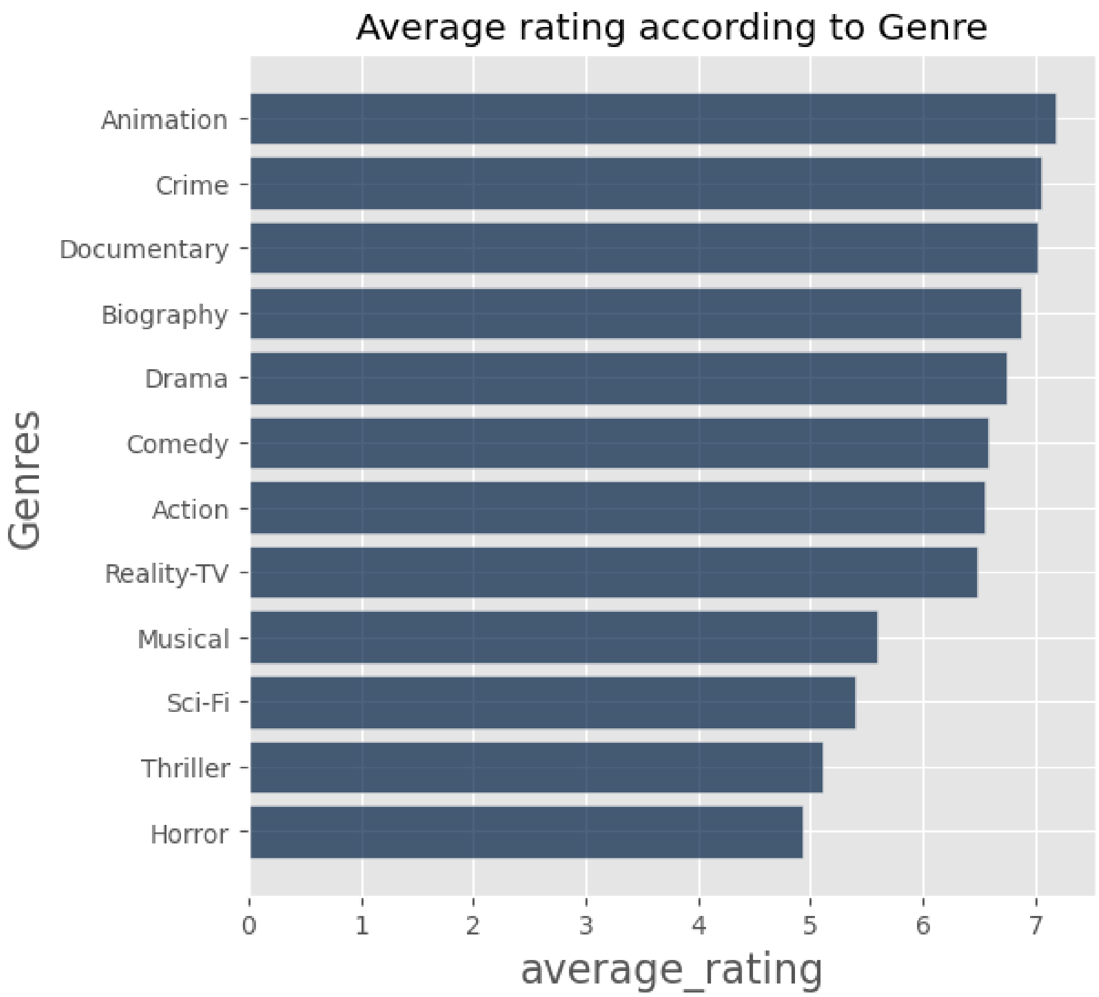
```
[17]: avg_ratings = make_array(avg_com,avg_ani, avg_Dra, avg_cri, avg_act, avg_doc,
      avg_mus, avg_hor, avg_thr, avg_sci, avg_rea, avg_bio)
      genres = make_array('Comedy', 'Animation', 'Drama', 'Crime', 'Action',
      'Documentary', 'Musical', 'Horror', 'Thriller', 'Sci-Fi',
      'Reality-TV','Biography')
```

```
[18]: # This cell created a table with the average rating along with the
      # genre title and sorts it in descending order
      k = Table().with_columns('Genres', genres)
      F = k.with_columns('average_rating', avg_ratings)
      h = F.select('Genres', 'average_rating').sort('average_rating', descending =
      True)
```

1.2 Figure 1

```
[41]: # Creates a bar graph that shows average ratings according to genre
      h.barh('Genres', 'average_rating')
      plt.title("Average rating according to Genre")
```

```
[41]: Text(0.5, 1.0, 'Average rating according to Genre')
```



It appears as if Animated films are the most highly rated. This could mean Netflix is trying to cater to more families, or maybe there is something about animated movies that people find relevant to their own childhood and experiences so they may rate it higher.

The following cells create average ratings based on certificate

```
[20]: TVMA = np.average(d.where('certificate', 'TV-MA').column('rating'))
```

```
[21]: TV14 = np.average(d.where('certificate', 'TV-14').column('rating'))
```

```
[22]: PG = np.average(d.where('certificate', 'PG').column('rating'))
```

```
[23]: PG13 = np.average(d.where('certificate', 'PG-13').column('rating'))
```

```
[24]: R = np.average(d.where('certificate', 'R').column('rating'))
```

```
[25]: G = np.average(d.where('certificate', 'G').column('rating'))
```

[26]:
```
MA17 = np.average(d.where('certificate', 'MA-17').column('rating'))
```

[27]:
```
TVPG = np.average(d.where('certificate', 'TV-PG').column('rating'))
```

[28]:
```
NC17 = np.average(d.where('certificate', 'NC-17').column('rating'))
```

[29]:
```
TVG = np.average(d.where('certificate', 'TV-G').column('rating'))
```

[30]:
```
NR = np.average(d.where('certificate', 'Not Rated').column('rating'))
```

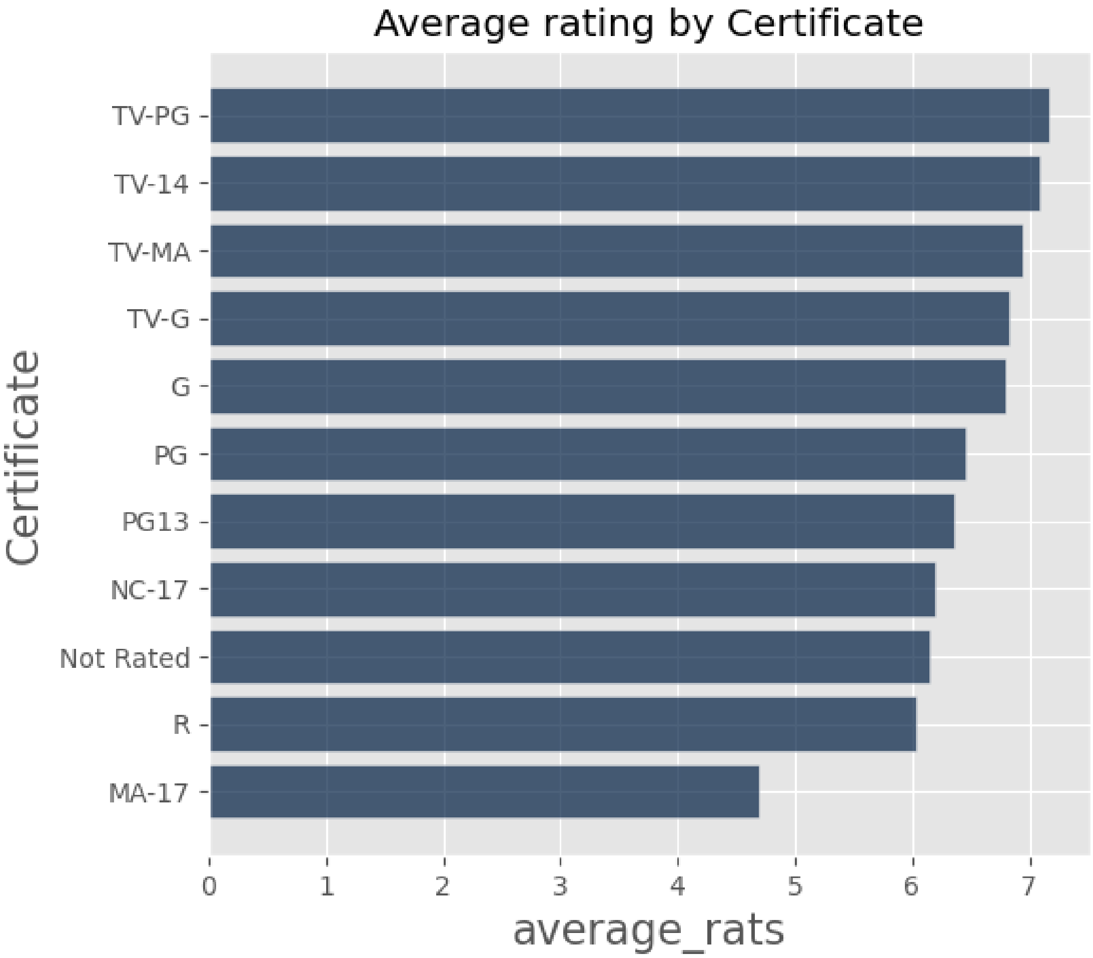
[31]:
```
avg_rats = make_array(TVMA, TV14, PG, PG13, R, G, MA17, TVPG, NC17, TVG, NR)
Certificate = make_array('TV-MA', 'TV-14', 'PG', 'PG13','R','G',
'MA-17','TV-PG','NC-17','TV-G','Not Rated')
```

[32]:
```
#This cell created a table with the average rating along with the genre title
#and sorts it in descending order
b = Table().with_columns('Certificate', Certificate)
c = b.with_columns('average_rats', avg_rats)
m = c.select('Certificate', 'average_rats').sort('average_rats', descending =
True)
```

1.3 Figure 2

[40]:
```
# Creates a bar graph that shows the rating compared to the average rating for
#that certificate
m.barh('Certificate', 'average_rats')
plt.title("Average rating by Certicate")
```

[40]:
```
Text(0.5, 1.0, 'Average rating by Certificate')
```



This graph shows that TV-PG films are typically rated the highest, which cooresponds to a lot of Animated films and also seconds the hypothesis that Netflix targets families for some of their more popular shows.

```
[34]: durdata = Table.read_table("n_moviesDURATION.csv")
```

This dataset is the same as the one before, however this one has been modified to remove all null duration values and also separate the unit min from the numeric value so it could be used

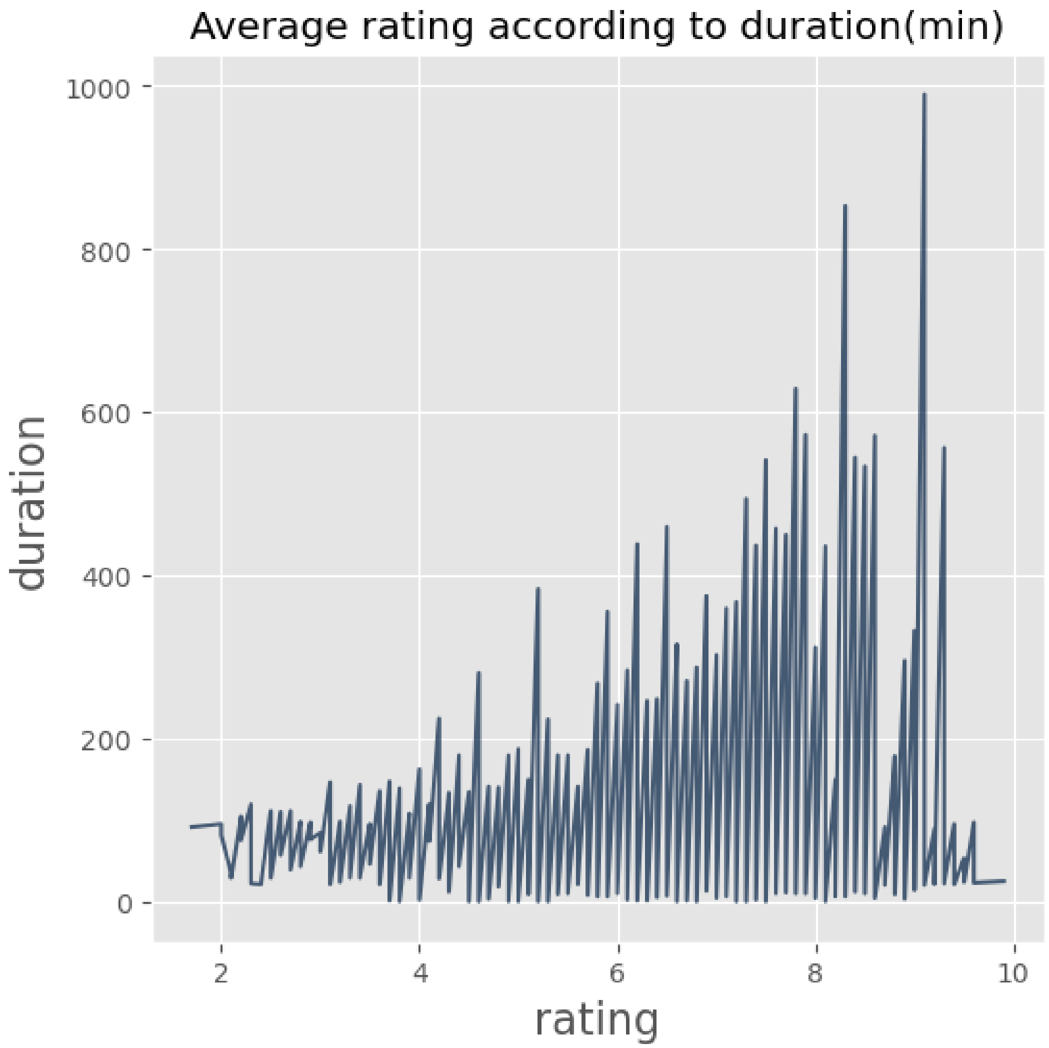
```
[35]: dur = durdata.sort('duration', descending = True)
      dur.show(10)
```

```
<IPython.core.display.HTML object>
```

1.4 Figure 3

```
[39]: dur.plot('rating','duration')
      plt.title("Average rating according to duration(min)")
```

```
[39]: Text(0.5, 1.0, 'Average rating according to duration(min)')
```



This finding is quite interesting…it appears as if it is higher rated, it is more likely to be longer in duration than if it is lower rated. Someone in the industry may use this as evidence to claim a longer movie will be more successful. It is also important to note that it is unclear how they collected the duration, it could potentially be a TV show that was rated high which would mean the more episodes or streaming content in general that you have the more likely you are to have a successful film.

Based on the bar graphs, if a producer wants to have the best success they should focus on creating a long animated film with a TV-PG certificate. Based on this data, this will result in the most highly rated movie/show.