HW02

February 2, 2020

0.1 Author: Skylar Trendley

0.2 Assignment: HW02

0.3 Class: Data Mining

0.3.1 Concept Description:

In this exercise, existing data was imported and used for attribute analysis. Using the scales of measurement, four attribute types–Nominal, Ordinal, Interval and Ratio–were used to qualify or quantify data variables presented in a csv.

0.3.2 Data Collection:

N/A

0.3.3 Example Description:

1. audience_freshness: Interval – value ranked on a 0-100 scale. Values have a meaningful order and measureable difference.
2. poster_url: Nominal – each value is unique and has no scale.
3. rt_audience_score: Interval – value is ranked on a 1-5 scale. Values have meaningful order and measureable difference (i.e. a 1 score is considerably worse than a 2 score)
4. rt_freshness: Ratio – value is ranked on a 0-100 scale. Values have meaningful order and measureable difference (i.e. a 50 score is a grade below a 60 score). The significance of a 0 score has meaning, as it is used to depict a movie that the movie was rated unanimously poorly by its audience.
5. rt_score: Ratio – value is ranked on a 0-10 scale. Values have meaningful order and measureable difference (i.e. a movie with a 10 score is considerably better than a movie with a 9 score, as it is the difference between a perfect movie and a movie with very few flaws). The significance of a 0 score has meaning, as it is used t depict a movie that rotten tomatoes deems unwatchable.
6. 2015_inflation: Interval – value is a percentage, has a measureable difference, mathematical operations can be conducted on it. It cannot be ratio as there is not a significant zero value, the zero is not the true starting point as the inflation rate can be negative. The order is not a known scale, but predictively (-1000% - 1000%)
7. adjusted: Interval – value is monetary, has a measurable difference, mathematical operations can be conducted on it. It cannot be ratio as the adjusted amount spent on the movie cannot be zero. The order is a known scale only in terms of the value of money (one dollar versus a billion dollars)

8. genres: Nominal – values are strings and have no scale.
9. genre_1: Nominal – values are strings and have no scale.
10. genre_2: Nominal – values are strings and have no scale.
11. genre_3: Nominal – values are strings and have no scale.
12. imdb_rating: Interval – value is ranked on a 1.0-10.0 scale. Values have meaningful order and measurable difference (ie a 1 score is considerably worse than a 3 score).
13. length: Interval – value is time in chronological order. Values have meaning order and measurable difference. (i.e. a movie that is 120 minutes long feels shorter than a movie that is 160 minutes long)
14. rank_in_year: Interval – value is ranked in a 1-10 scale. Values have meaningful order and measurable difference (i.e. the movie ranked at 3 is considered to be better than the movie ranked at 2).
15. rating: Nominal – values are strings and have no scale.
16. release_date: Nominal – values are strings and have no scale.
17. studio: Nominal – values are strings and have no scale.
18. title: Nominal – values are strings and have no scale.
19. worldwide_gross: Interval – value is monetary, has a measurable difference, mathematical operations can be conducted on it. It cannot be ratio as the adjusted amount spent on the movie cannot be zero. The order is a known scale only in terms of the value of money (one dollar versus a billion dollars)
20. year: Interval - value is a year, has a measureable difference, mathematical operations can be conducted on it. The scale is year (i.e. 1975-2020) and has measurable difference (a movie released in an earlier year would have less time to make profit than a previous year.

0.3.4 Data Import and Wrangling:

Two python3 library imports were used in the analysis of this dataset. The first, being pandas, allowed for reading the data in as a .csv for data processing. The second, matplotlib, allowed for displaying the quantifiable data in histograms.

0.3.5 Exploratory Data Analysis:

In order to look into the data, some excel functions were used to view the minimum and maximum values of some attributes to see if they fell into the significant value criteria. This allowed for scales and ranges to organize the data into certain scales of measurement.

0.3.6 Mining or Analytics:

Visualizations that were used were histograms. These histograms allowed for a visual representation of the order and measureable differences in the data. Each attribute that has significant, scalable data is displayed below.

0.3.7 Evaluation:

By using min and max excel functions to organize the data and find significant zero values, other significant features of the data were found. This included missing data values for attributes such as 2015_inflation, adjusted, worldwide_gross, and year.

0.3.8 Results:

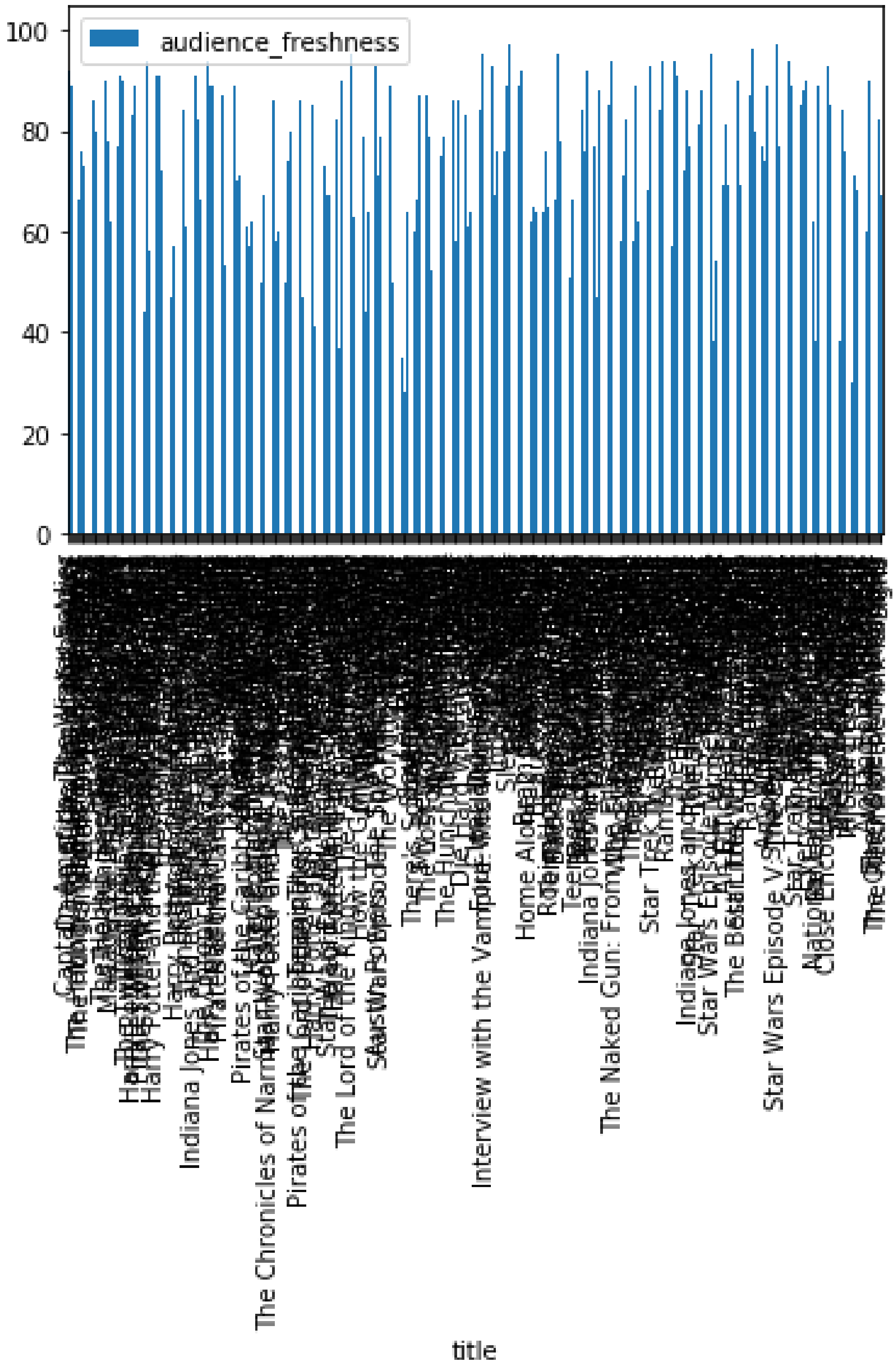
By ordering the ordinal or interval data into histograms, the significance of the measurable differences in each data attribute is visible.

```
In [28]: import pandas as pd
         import matplotlib.pyplot as plt

         df = pd.read_csv("blockbuster-top_ten_movies_per_year_DFE.csv")

         df.set_index('title')[['audience_freshness']].plot.bar()

         #plt.figure(figsize=(10,10))
         plt.show()
```

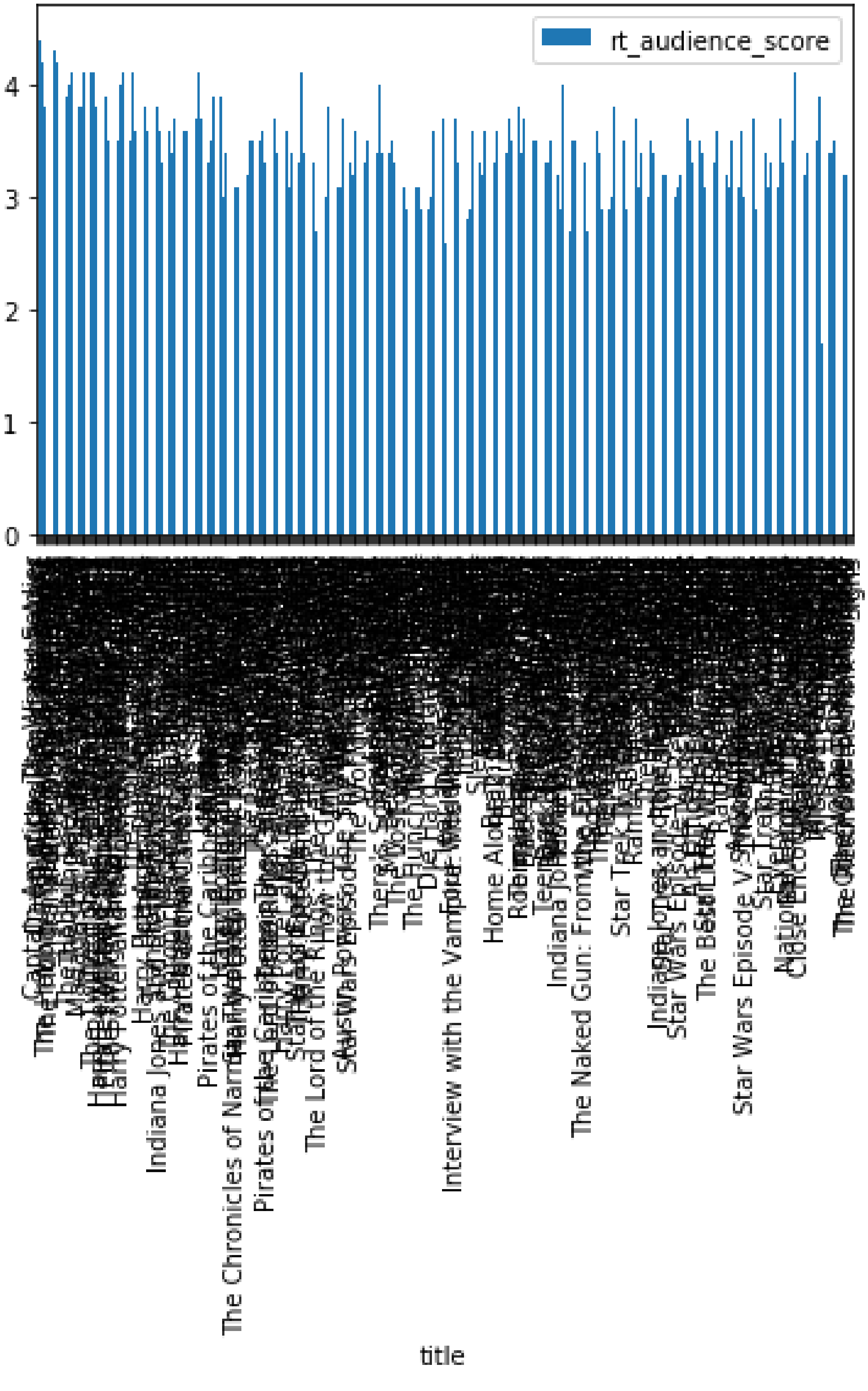


```
In [31]: import pandas as pd
         import matplotlib.pyplot as plt
```

```
df = pd.read_csv("blockbuster-top_ten_movies_per_year_DFE.csv")

df.set_index('title')[['rt_audience_score']].plot.bar()

#plt.figure(figsize=(10,10))
plt.show()
```

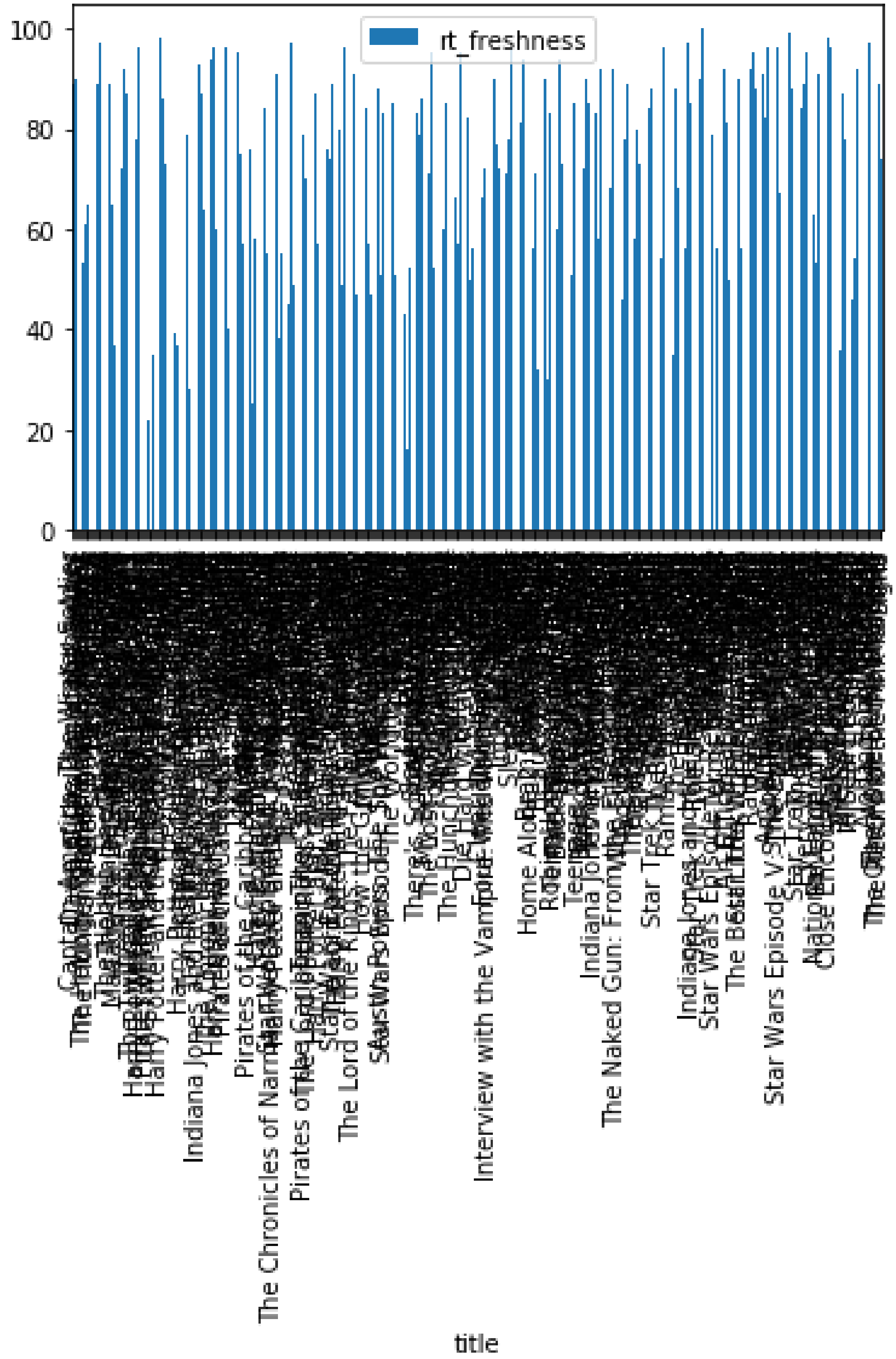


```
In [32]: import pandas as pd
         import matplotlib.pyplot as plt
```

```
df = pd.read_csv("blockbuster-top_ten_movies_per_year_DFE.csv")

df.set_index('title')[['rt_freshness']].plot.bar()

#plt.figure(figsize=(10,10))
plt.show()
```

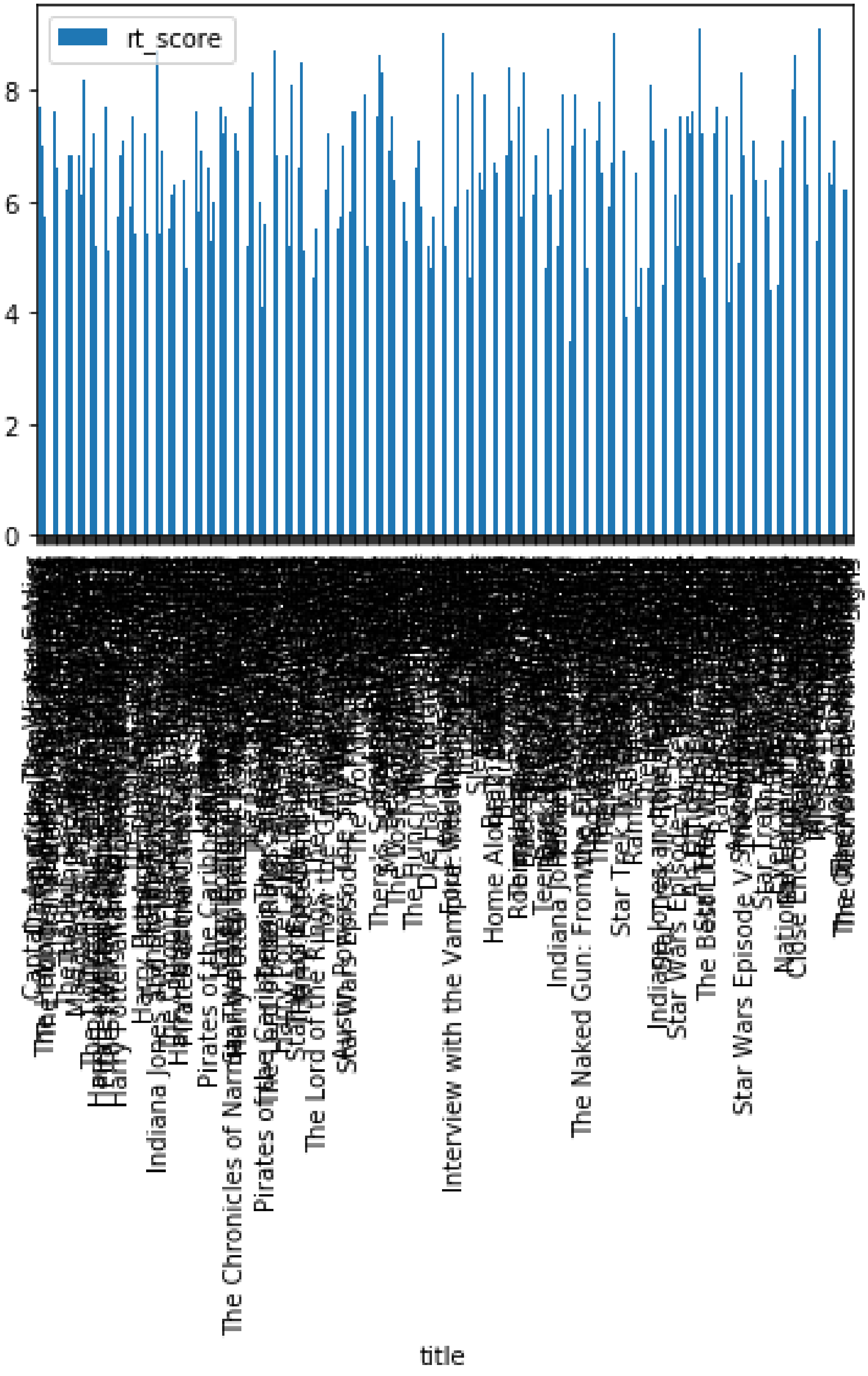


```
In [33]: import pandas as pd
         import matplotlib.pyplot as plt
```

```python
df = pd.read_csv("blockbuster-top_ten_movies_per_year_DFE.csv")

df.set_index('title')[['rt_score']].plot.bar()

#plt.figure(figsize=(10,10))
plt.show()
```

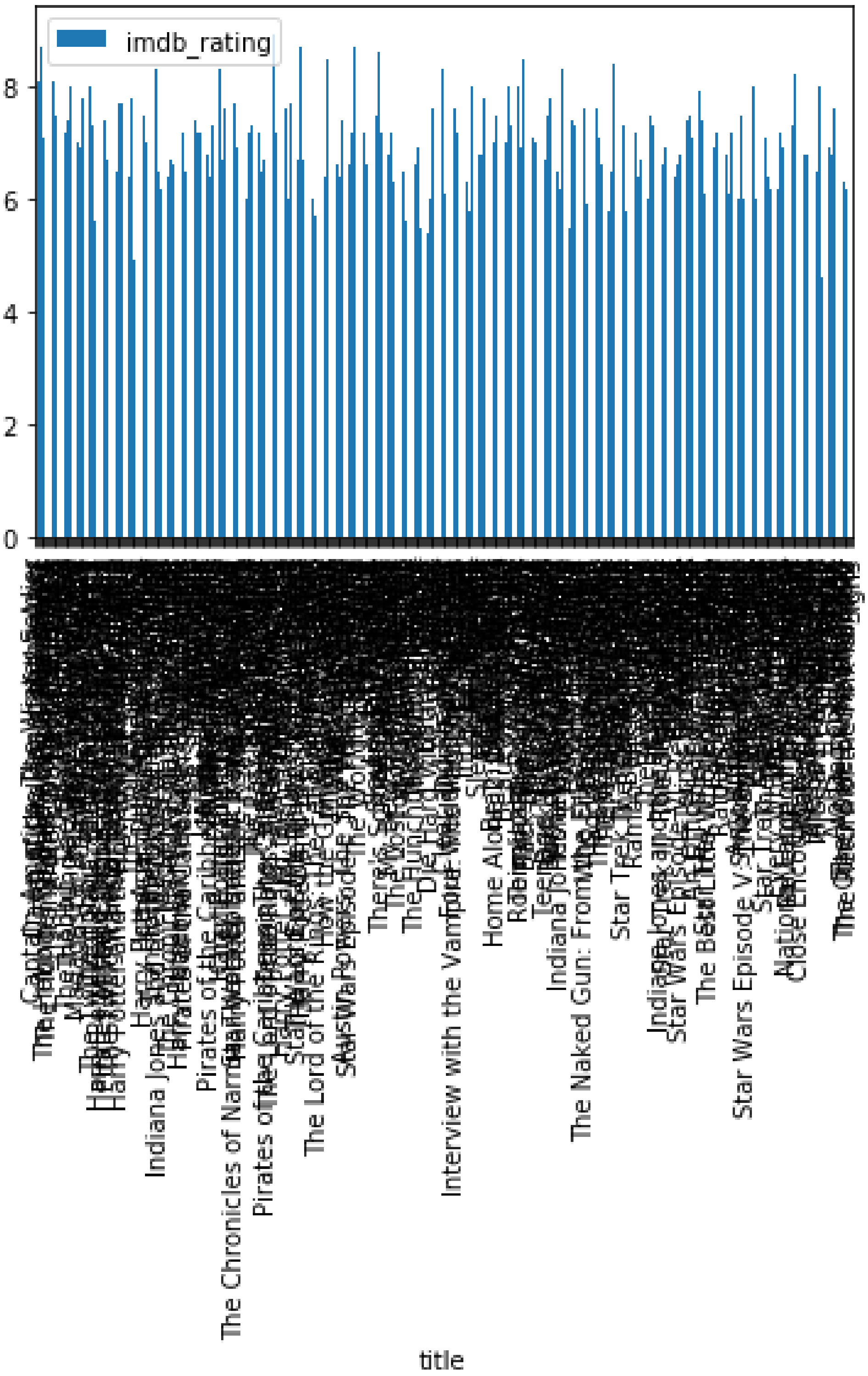


```
In [36]: import pandas as pd
         import matplotlib.pyplot as plt
```

```
df = pd.read_csv("blockbuster-top_ten_movies_per_year_DFE.csv")

df.set_index('title')[['imdb_rating']].plot.bar()

#plt.figure(figsize=(10,10))
plt.show()
```

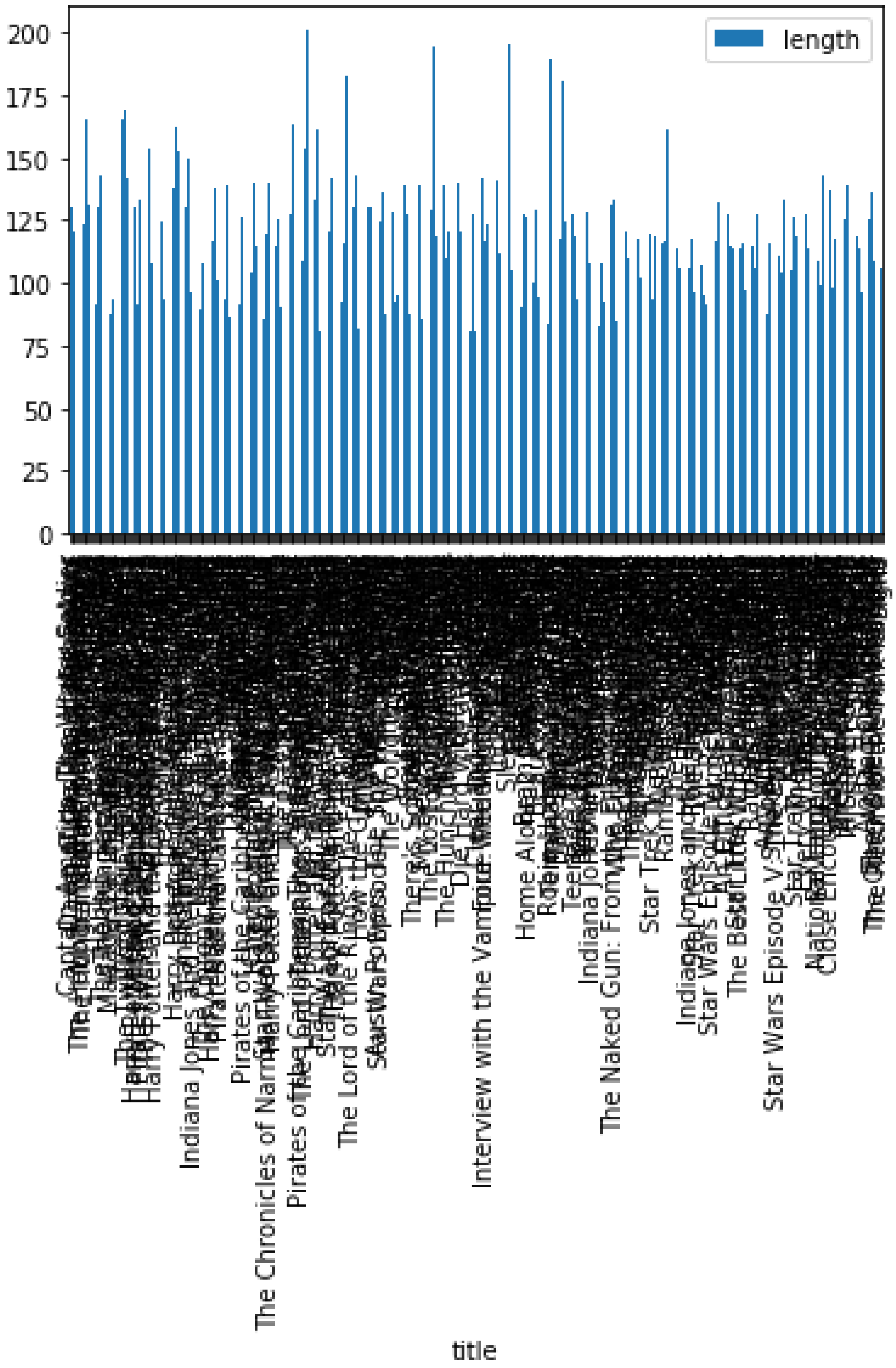


```
In [37]: import pandas as pd
         import matplotlib.pyplot as plt
```

```
df = pd.read_csv("blockbuster-top_ten_movies_per_year_DFE.csv")

df.set_index('title')[['length']].plot.bar()

#plt.figure(figsize=(10,10))
plt.show()
```

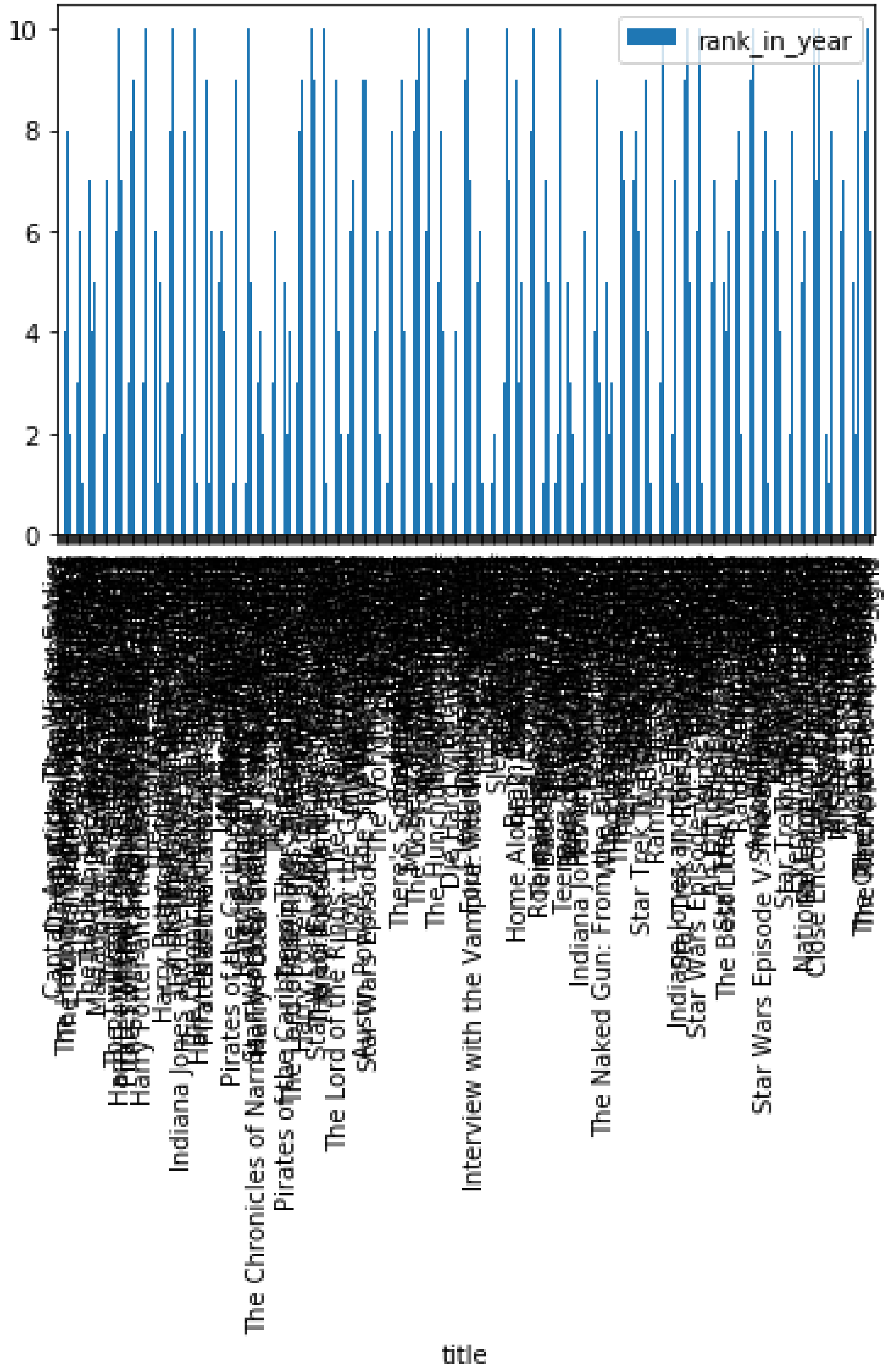


```
In [38]: import pandas as pd
         import matplotlib.pyplot as plt
```

```
df = pd.read_csv("blockbuster-top_ten_movies_per_year_DFE.csv")

df.set_index('title')[['rank_in_year']].plot.bar()

#plt.figure(figsize=(10,10))
plt.show()
```

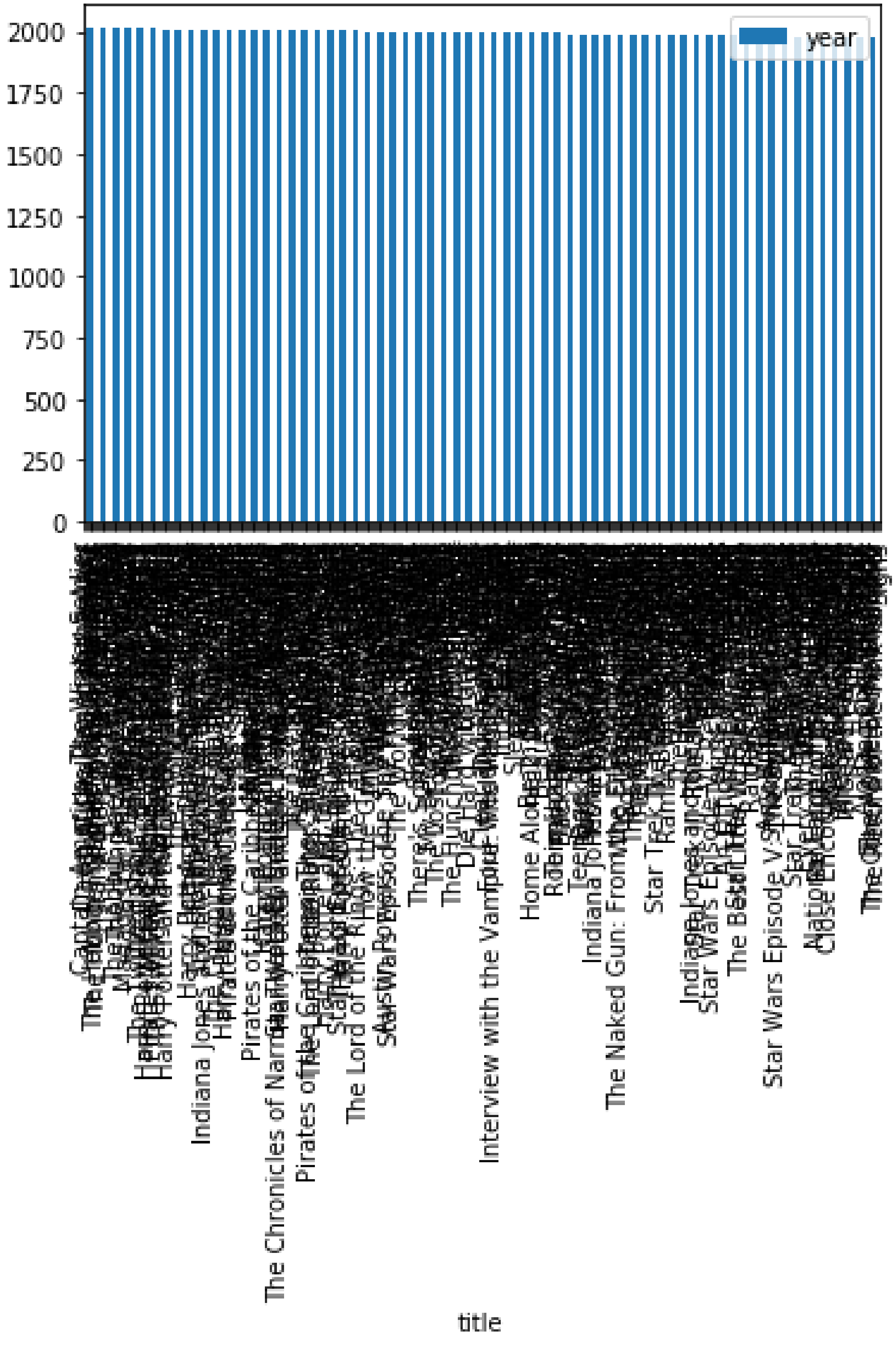


```
In [40]: import pandas as pd
         import matplotlib.pyplot as plt
```

```
df = pd.read_csv("blockbuster-top_ten_movies_per_year_DFE.csv")

df.set_index('title')[['year']].plot.bar()

#plt.figure(figsize=(10,10))
plt.show()
```



```
In [ ]:
```