Mon-13Mar2023

Report #1

Data Descriptionhttps://www.kaggle.com/datasets/carolzhangdc/imdb-5000-movie-dataset

This report provides a technical overview of the dataset obtained for the 820-machine learning project. The contents of this report provide the source of the data, the features and variables of the raw data, as well as any limitations of the data. The report aims at describing the data in its original form prior to any analytical or feature manipulation techniques.

The data was obtained from Kaggle.com at the following url:

https://www.kaggle.com/datasets/carolzhangdc/imdb-5000-movie-dataset

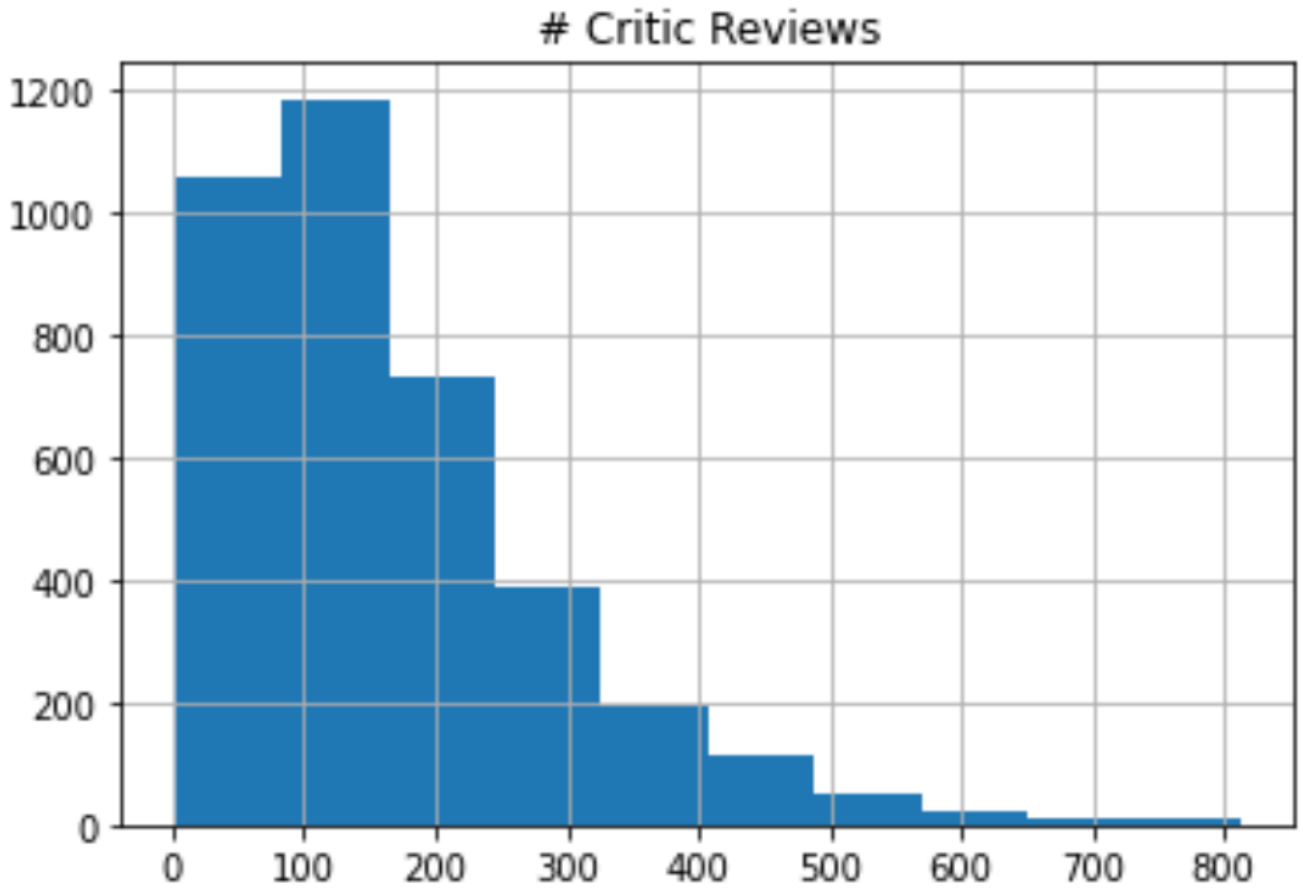
The original dimensions of the data are 5043L x 28W, and contains movies from 1916 to 2016, with the vast majority of movies being released after 1980.
In the table below, the entire dataset is described after removing movie records with missing values. The table also shows the quantity and datatype of each variable.

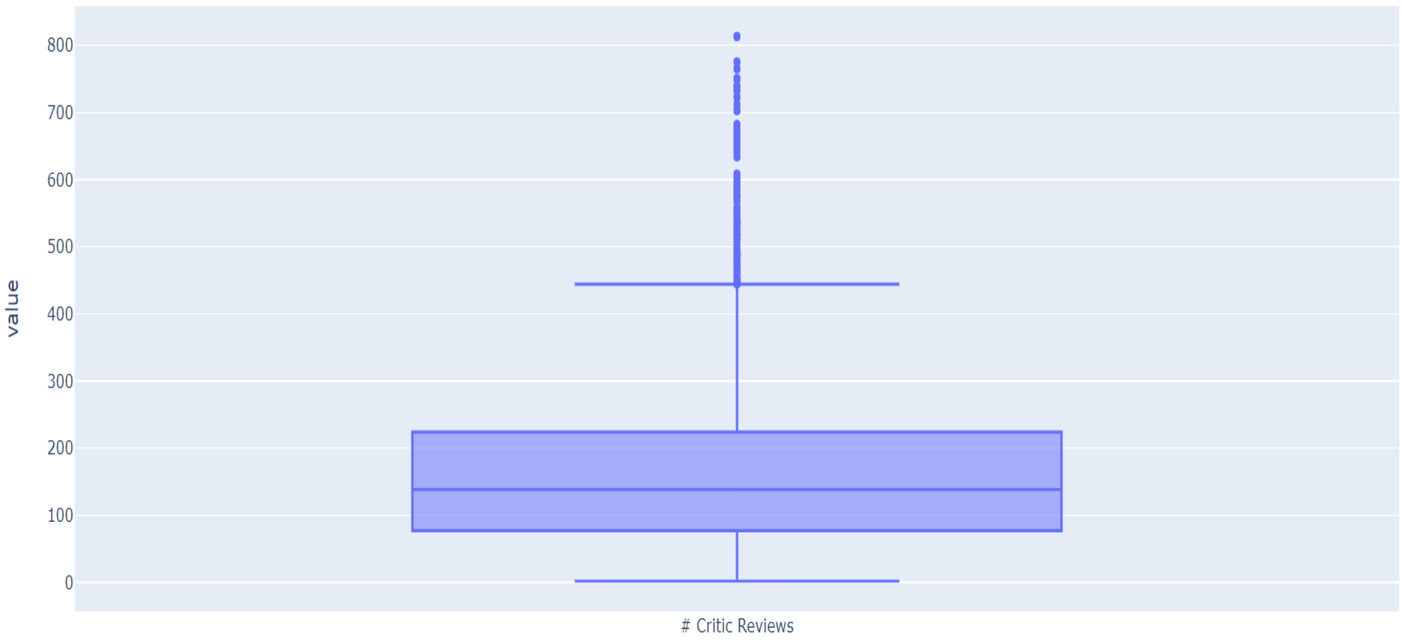
```
Int64Index: 3756 entries, 0 to 5042
Data columns (total 28 columns):
 #   Column               Non-Null Count  Dtype
---  ------               --------------  -----
 0   Color                3756 non-null   object
 1   Director Name        3756 non-null   object
 2   # Critic Reviews     3756 non-null   float64
 3   Duration             3756 non-null   float64
 4   # Director Likes     3756 non-null   float64
 5   # Actor 1 Likes      3756 non-null   float64
 6   Actor 2 Name         3756 non-null   object
 7   # Actor 1 Likes      3756 non-null   float64
 8   Gross                3756 non-null   float64
 9   Genres               3756 non-null   object
 10  Actor 1 Name         3756 non-null   object
 11  Movie Title          3756 non-null   object
 12  # Users Voted        3756 non-null   int64
 13  # Cast Likes         3756 non-null   int64
 14  Actor 3 Name         3756 non-null   object
 15  # FB Poster          3756 non-null   float64
 16  Plot Keywords        3756 non-null   object
 17  Movie Link           3756 non-null   object
 18  # Users for Reviews  3756 non-null   float64
 19  Langauge             3756 non-null   object
 20  Country              3756 non-null   object
 21  Content Rating       3756 non-null   object
 22  Budget               3756 non-null   float64
 23  Title Year           3756 non-null   float64
 24  # Actor 2 Likes      3756 non-null   float64
 25  IMDB Score           3756 non-null   float64
 26  Aspect Ratio         3756 non-null   float64
 27  # Movie Likes        3756 non-null   int64
dtypes: float64(13), int64(3), object(12)
memory usage: 851.0+ KB
```

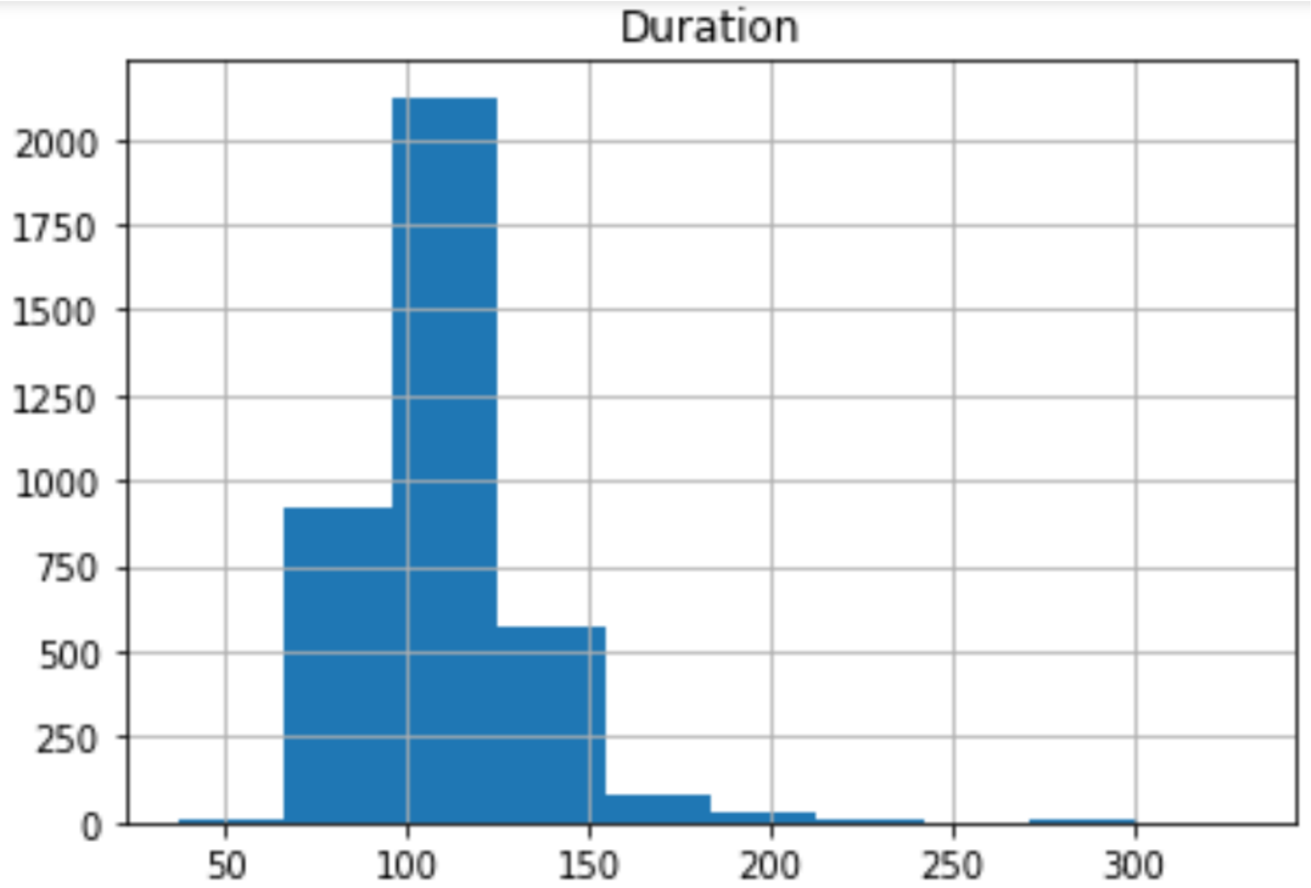
To gain better insight into the data, the below table shows a description of the original dataset’s numerical features.

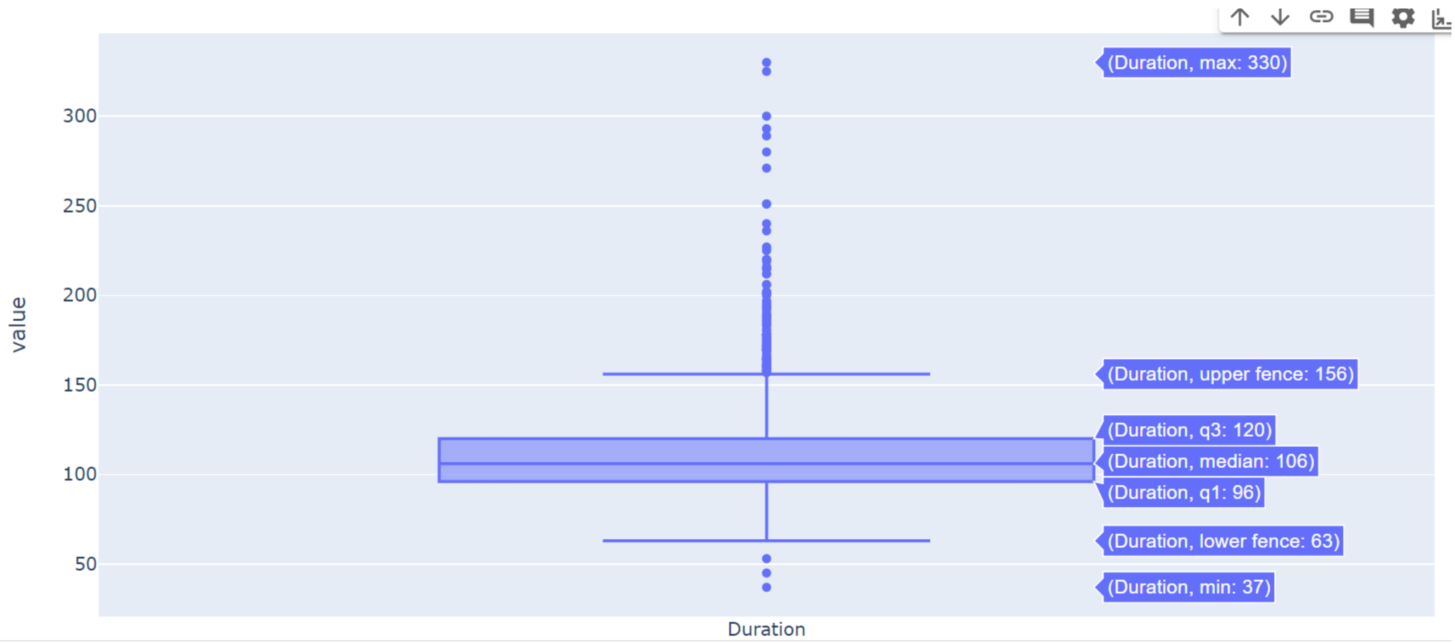
```
    # Critic Reviews  Duration  # Director Likes  # Actor 1 Likes  \
count          3756.00   3756.00           3756.00          3756.00
mean            167.38    110.26            807.34           771.28
std             123.45     22.65           3068.17          1894.25
min               2.00     37.00              0.00             0.00
25%              77.00     96.00             11.00           194.00
50%             138.50    106.00             64.00           436.00
75%             224.00    120.00            235.00           691.00
max             813.00    330.00          23000.00         23000.00

       # Actor 1 Likes       Gross  # Users Voted  # Cast Likes  # FB Poster
count          3756.00      3756.00        3756.00       3756.00      3756.00
mean           7751.34  52612824.24      105826.73      11527.10         1.38
std           15519.34  70317866.91      152035.40      19122.18         2.04
min               0.00       162.00          91.00          0.00         0.00
25%             745.00   8270232.75       19667.00       1919.75         0.00
50%            1000.00  30093107.00       53973.50       4059.50         1.00
75%           13000.00  66881940.75      128602.00      16240.00         2.00
max          640000.00 760505847.00     1689764.00     656730.00        43.00

       # Users for Reviews         Budget  Title Year  # Actor 2 Likes
count              3756.00        3756.00     3756.00          3756.00
mean                336.84    46236849.64     2002.98          2021.78
std                 411.23   226010288.48        9.89          4544.91
min                   4.00         218.00     1927.00             0.00
25%                 110.00    10000000.00     1999.00           384.75
50%                 210.00    25000000.00     2004.00           685.50
75%                 398.25    50000000.00     2010.00           976.00
max                5060.00 12215500000.00     2016.00        137000.00

       IMDB Score  Aspect Ratio  # Movie Likes
count     3756.00       3756.00        3756.00
mean         6.47          2.11        9353.83
std          1.06          0.35       21462.89
min          1.60          1.18           0.00
25%          5.90          1.85           0.00
50%          6.60          2.35         227.00
75%          7.20          2.35       11000.00
max          9.30         16.00      349000.00
```

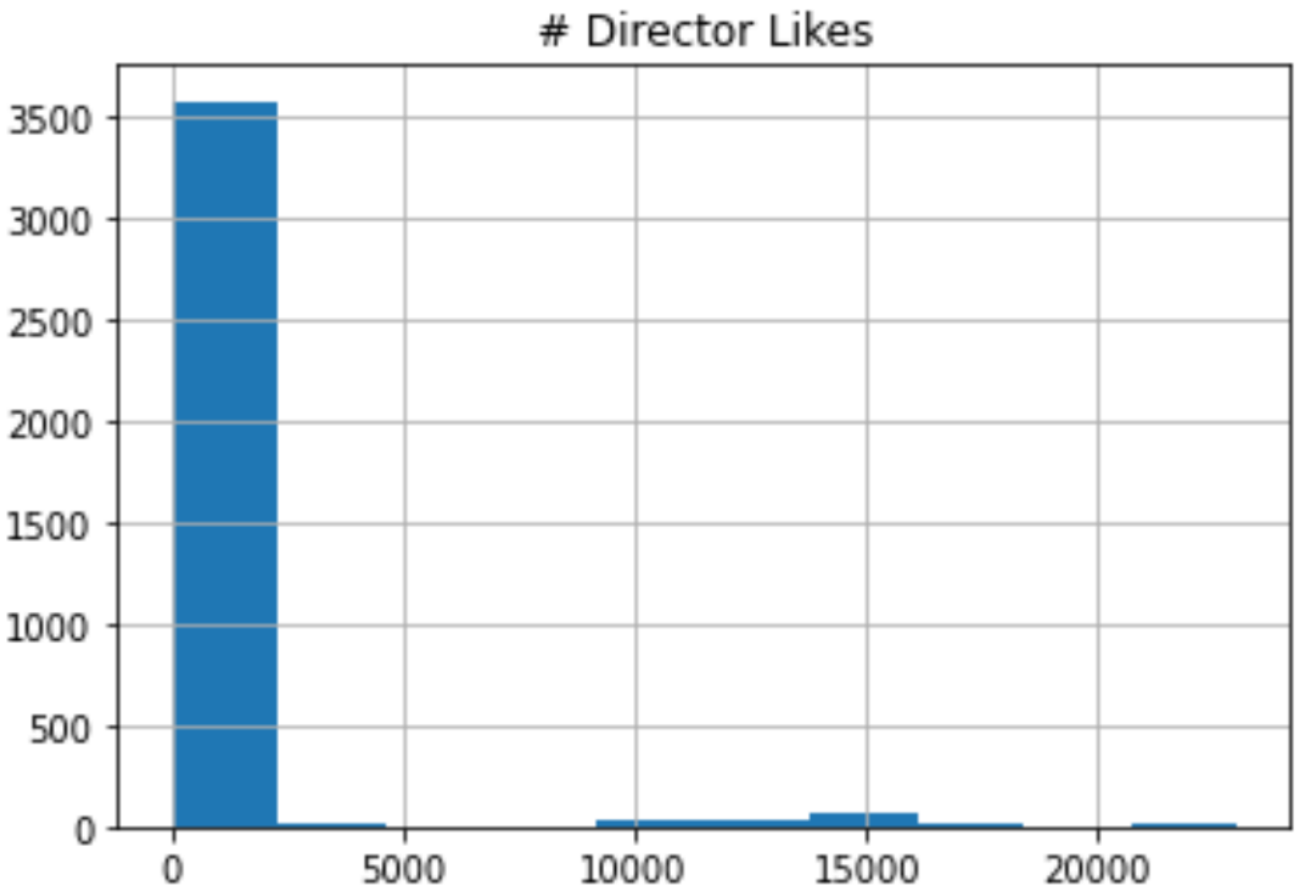
In addition to the above- displayed tables, the histogram- box & whisker combination helps provide a visualization of the numerical features.

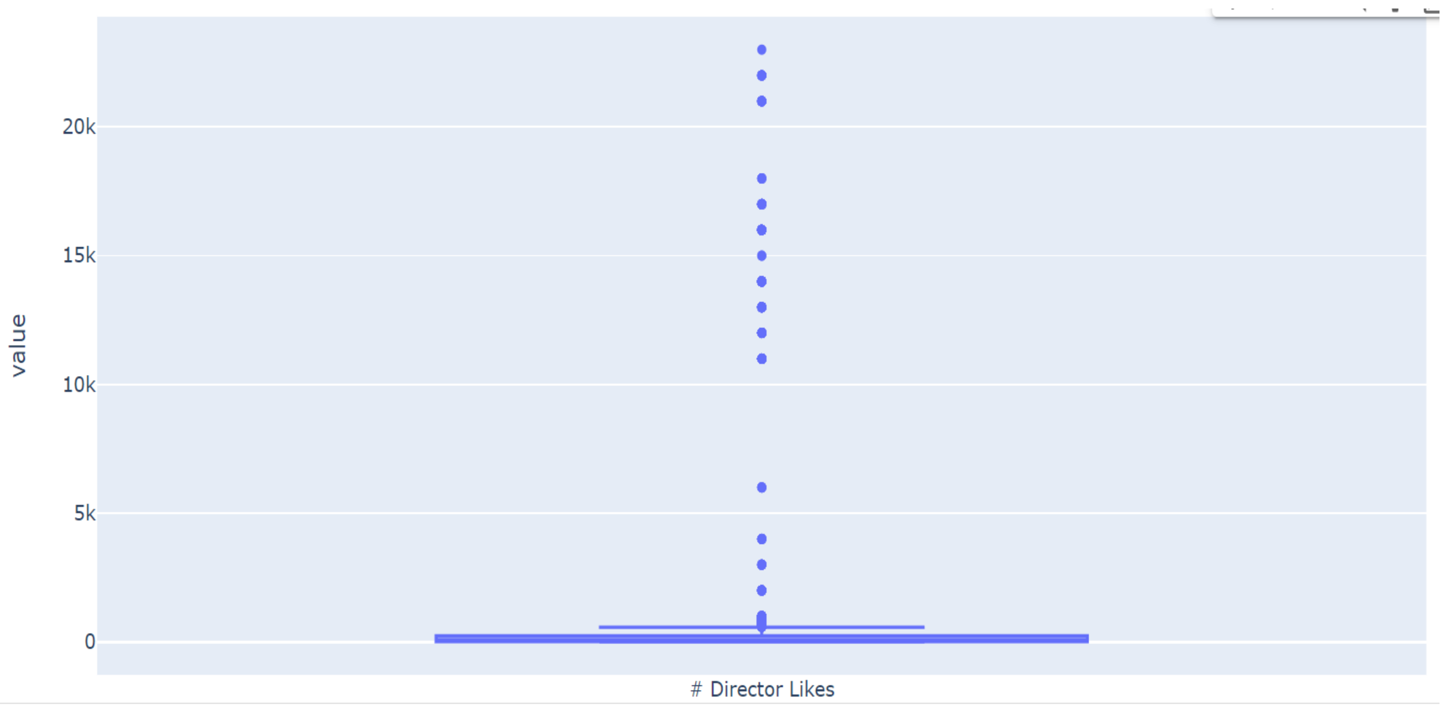

# Critic Reviews
1200
1000
800
600
400
200
0
0
100
200
300
400
500
600
700
800

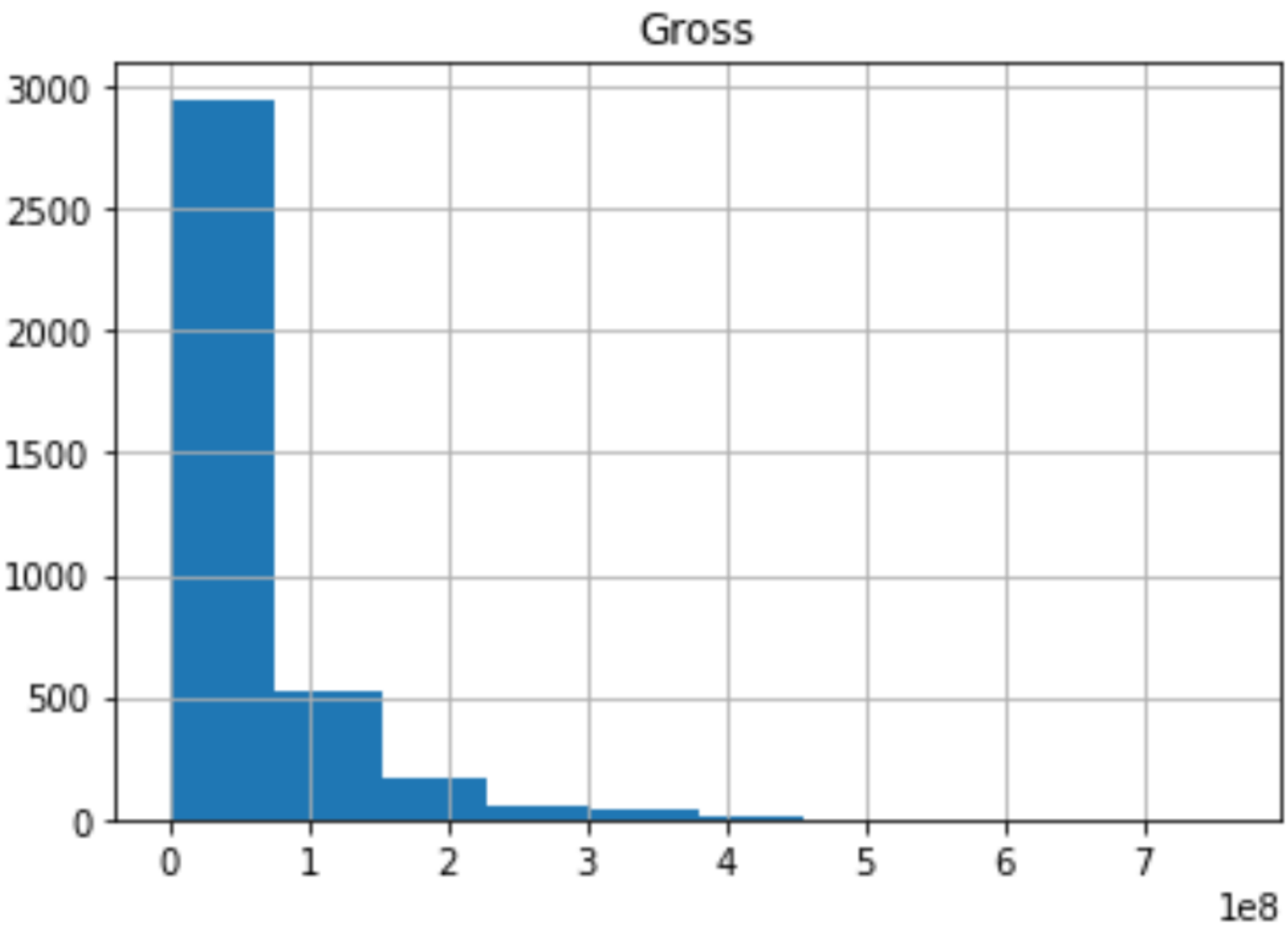

value
800
700
600
500
400
300
200
100
0
# Critic Reviews

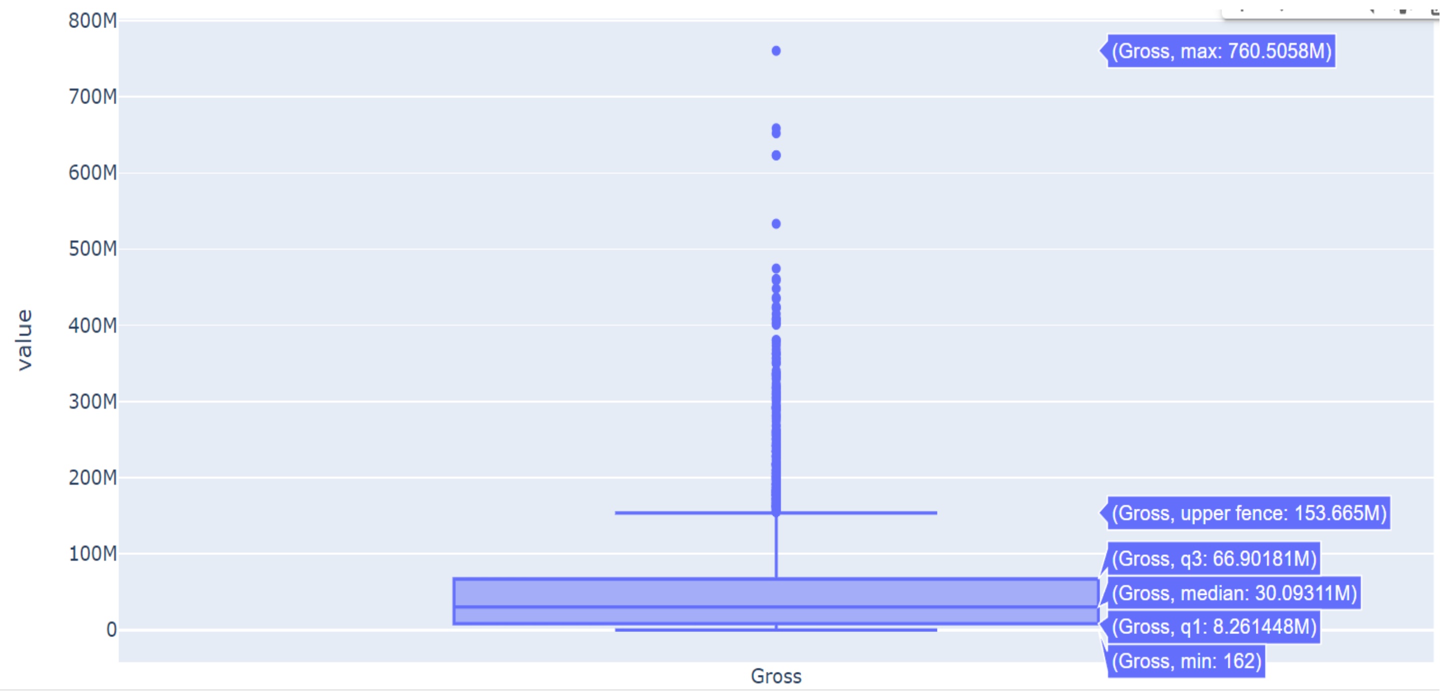

Duration
2000
1750
1500
1250
1000
750
500
250
0
50
100
150
200
250
300

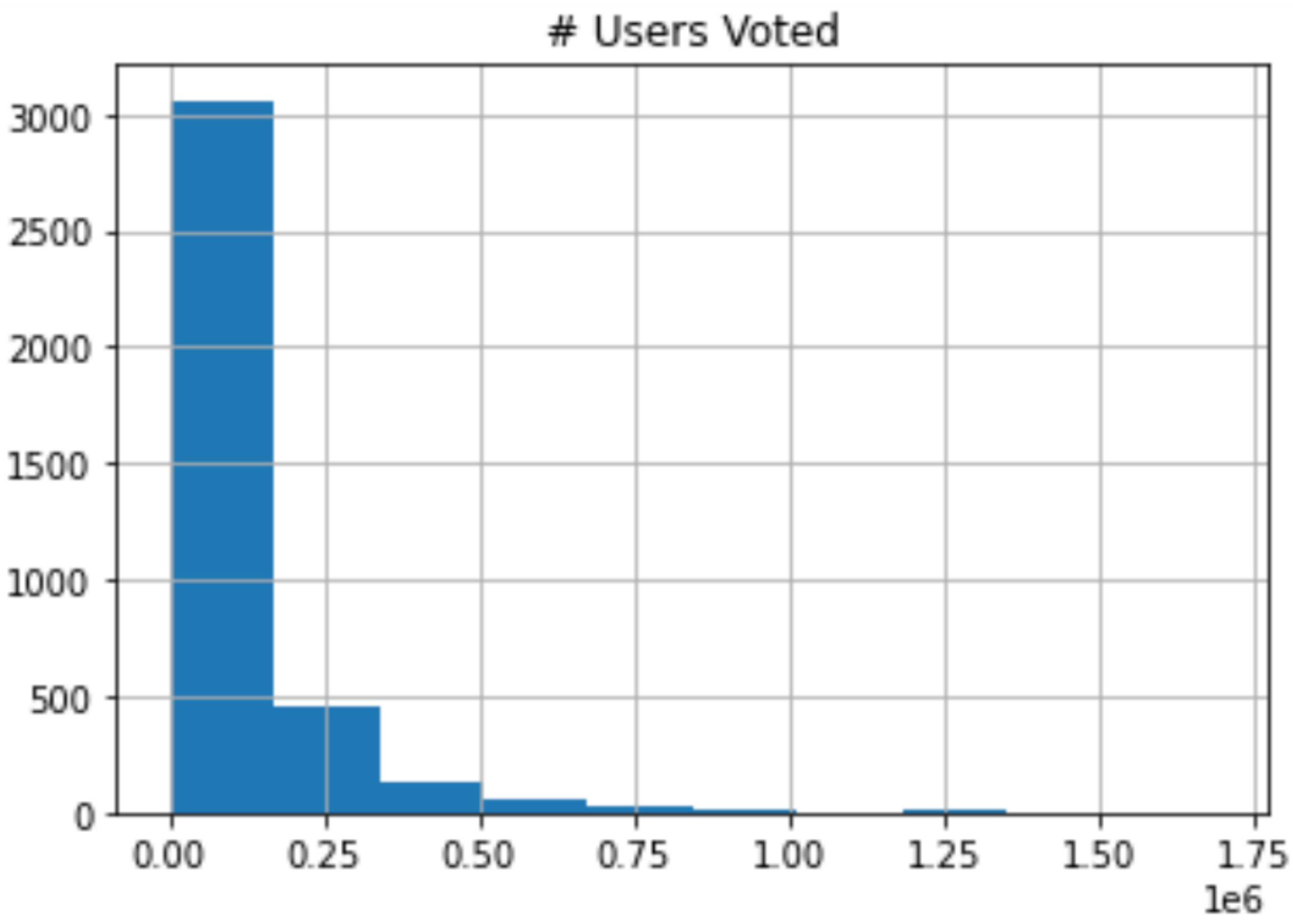

value
300
250
200
150
100
50
Duration
(Duration, max: 330)
(Duration, upper fence: 156)
(Duration, q3: 120)
(Duration, median: 106)
(Duration, q1: 96)
(Duration, lower fence: 63)
(Duration, min: 37)

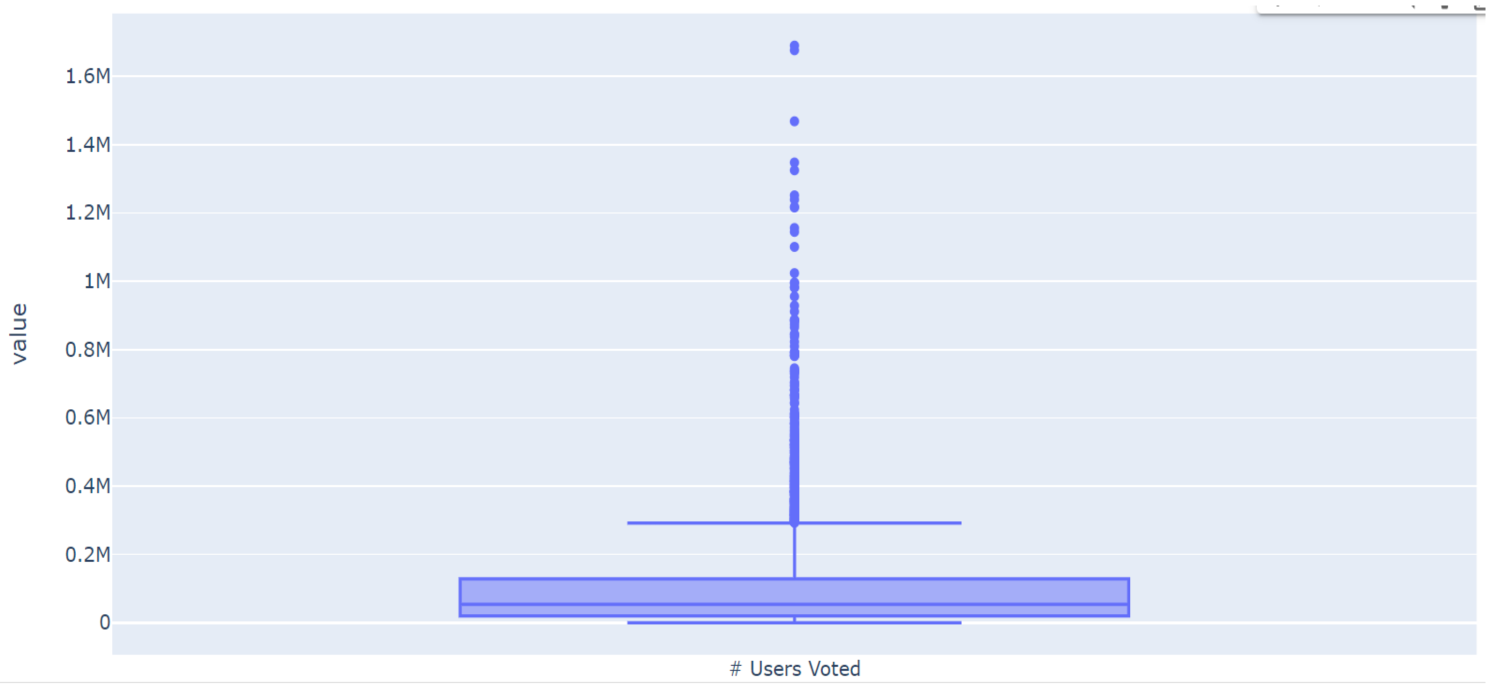

# Director Likes
3500
3000
2500
2000
1500
1000
500
0
0
5000
10000
15000
20000

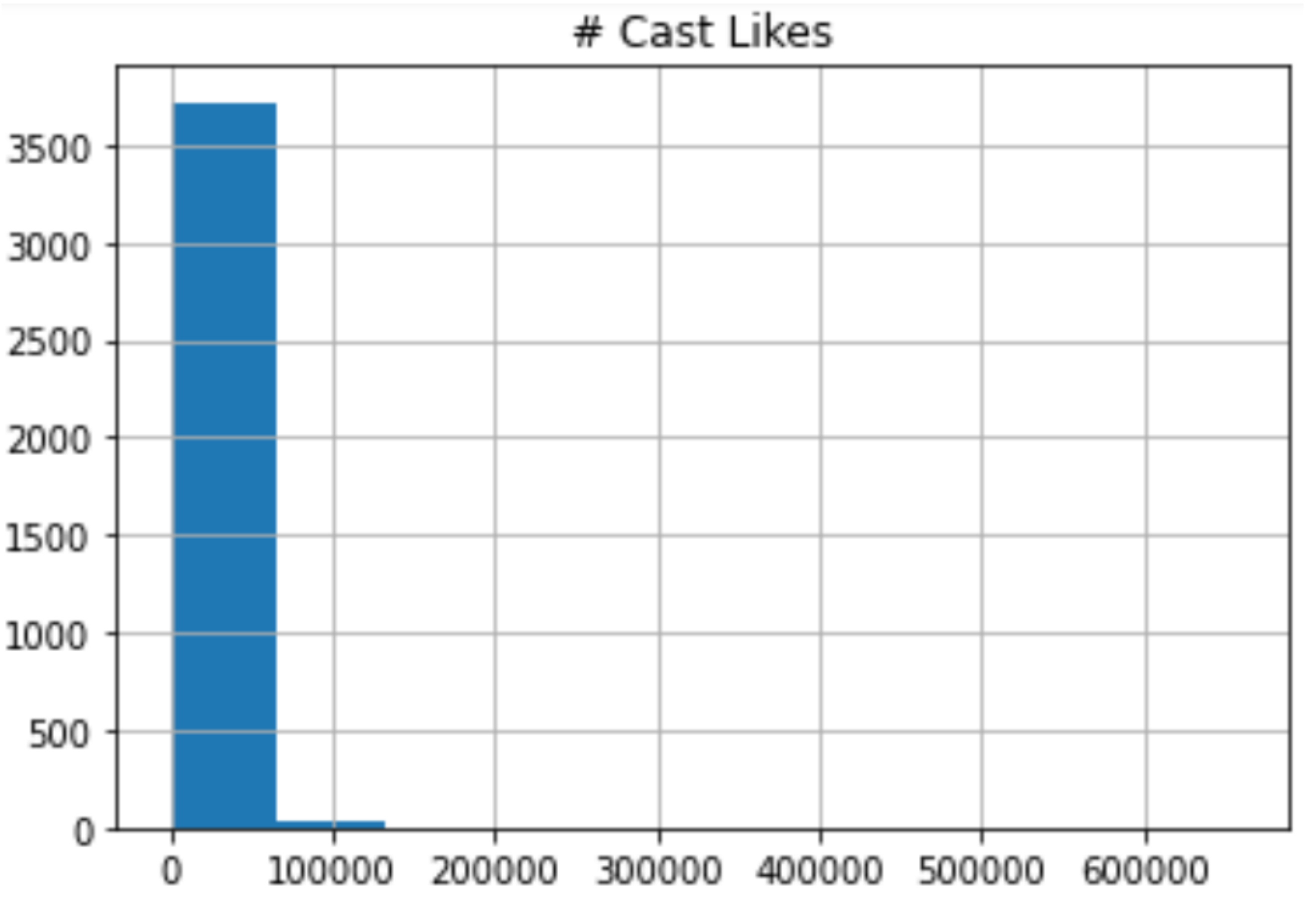

20k
15k
10k
5k
0
value
# Director Likes

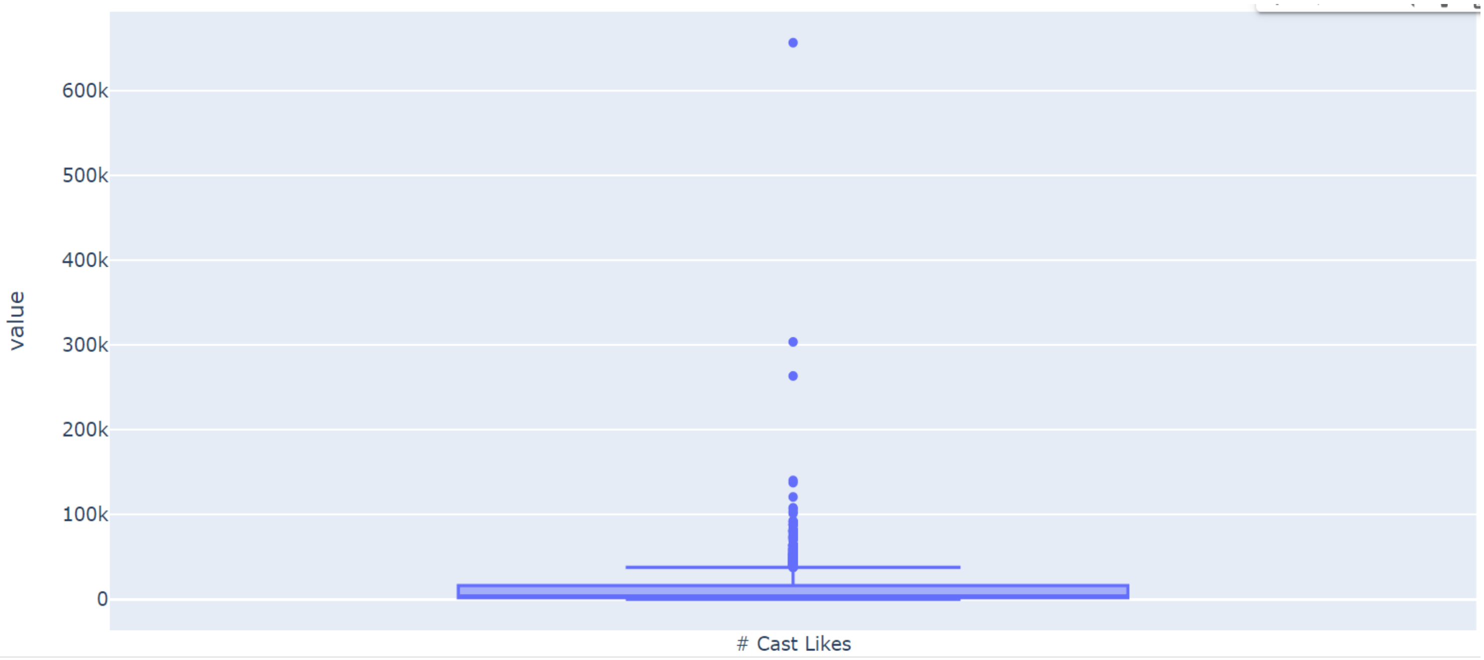

Gross
3000
2500
2000
1500
1000
500
0
0
1
2
3
4
5
6
7
1e8

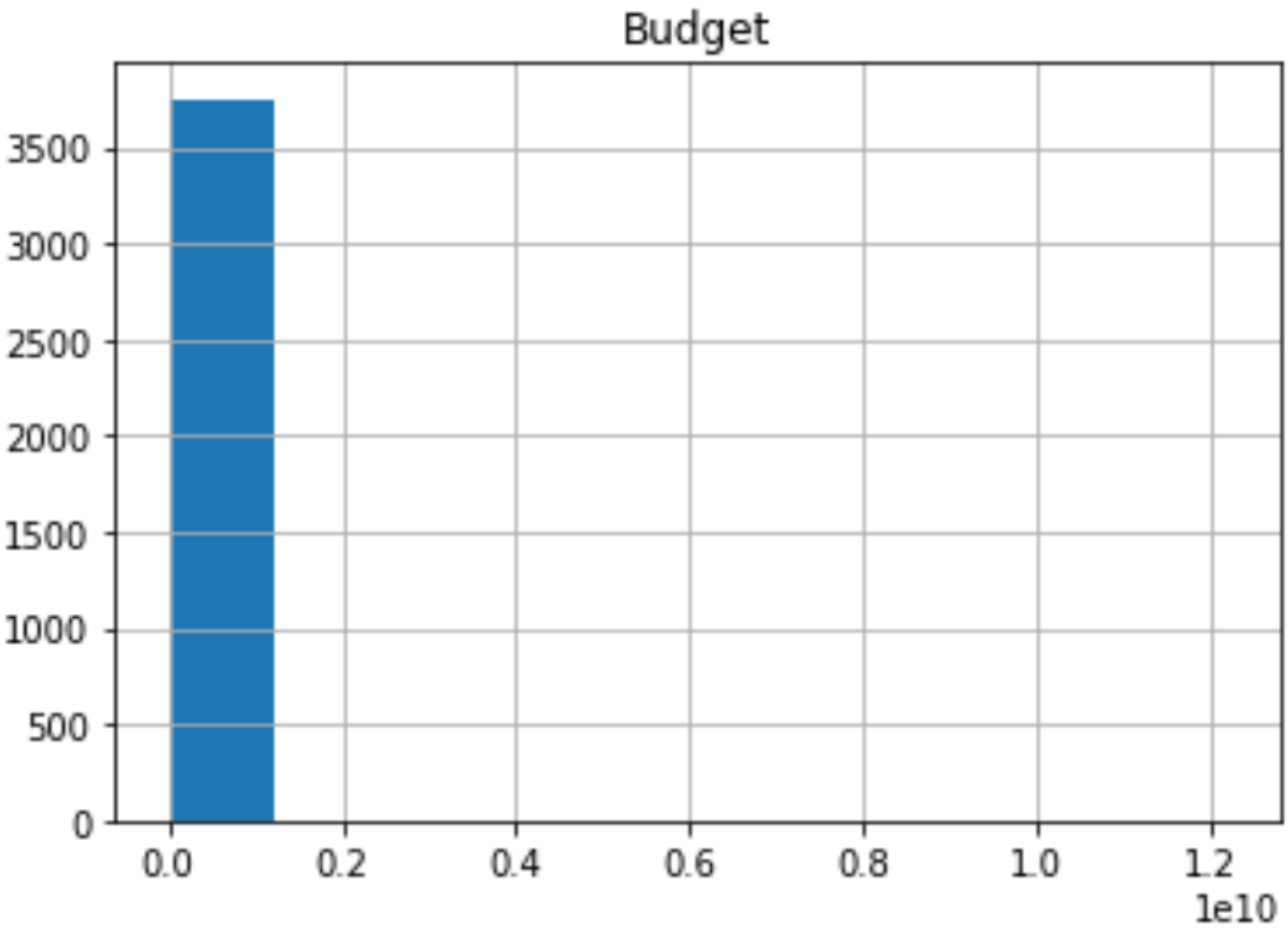

800M
700M
600M
500M
400M
300M
200M
100M
0
value
Gross
(Gross, max: 760.5058M)
(Gross, upper fence: 153.665M)
(Gross, q3: 66.90181M)
(Gross, median: 30.09311M)
(Gross, q1: 8.261448M)
(Gross, min: 162)

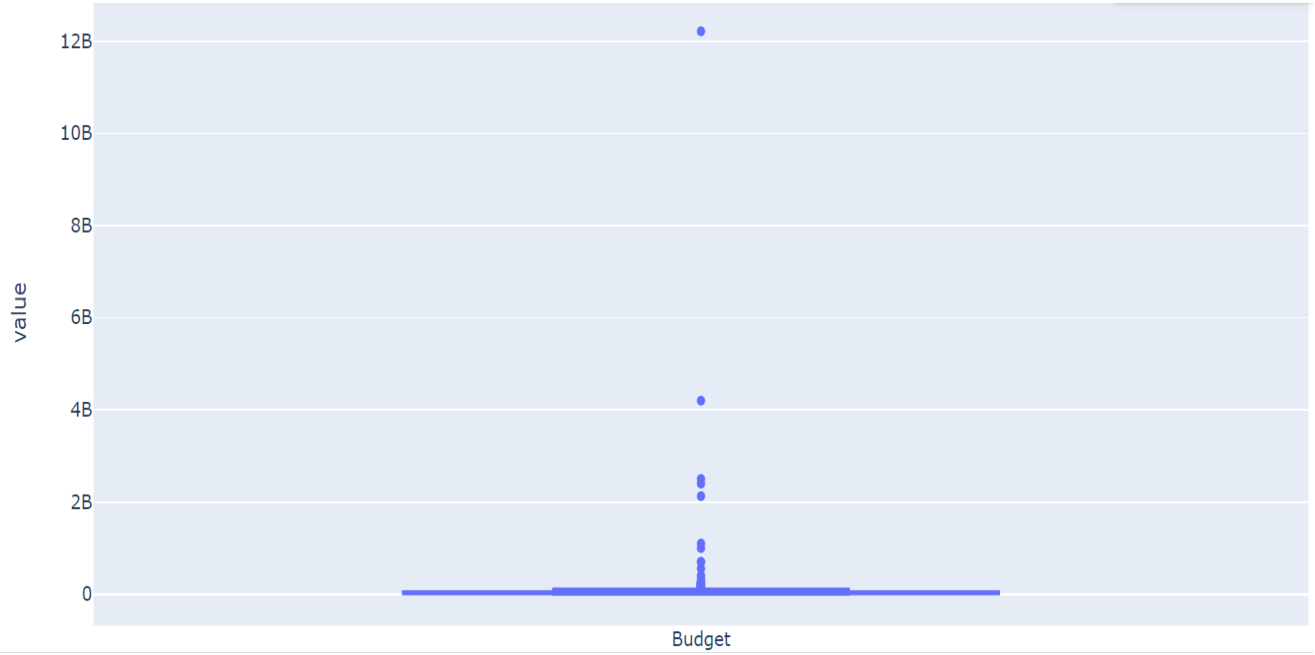

# Users Voted
3000
2500
2000
1500
1000
500
0
0.00
0.25
0.50
0.75
1.00
1.25
1.50
1.75
1e6

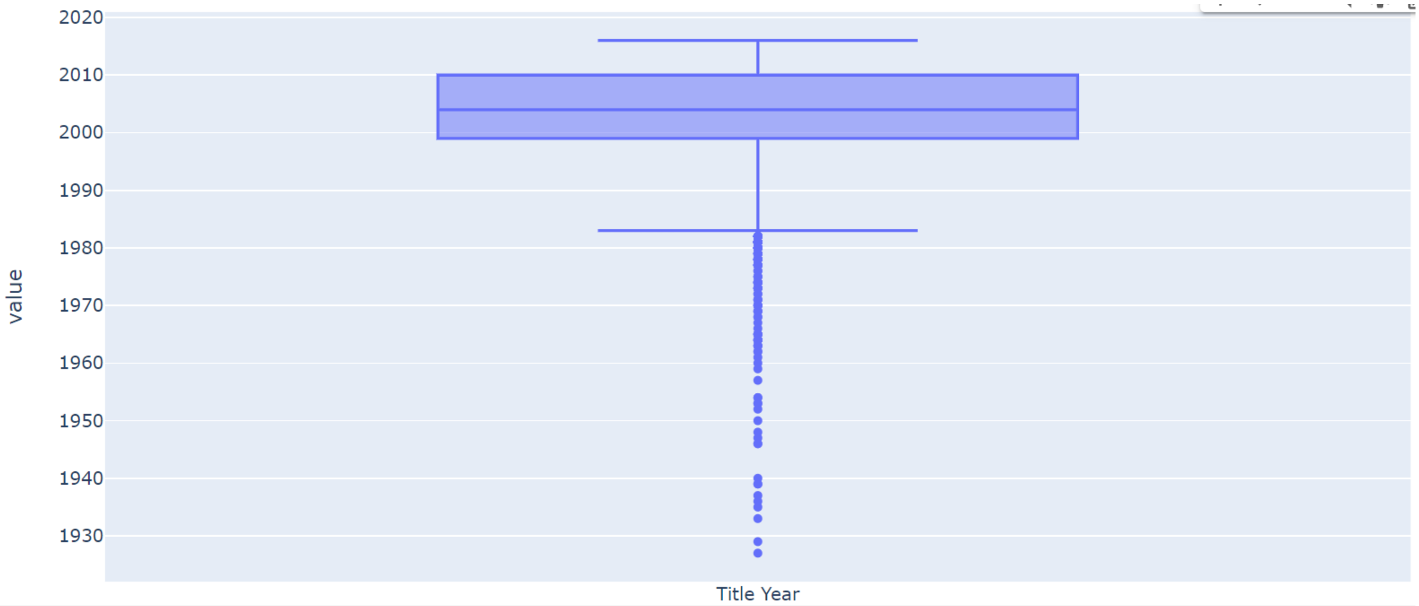

1.6M
1.4M
1.2M
1M
0.8M
0.6M
0.4M
0.2M
0
value
# Users Voted

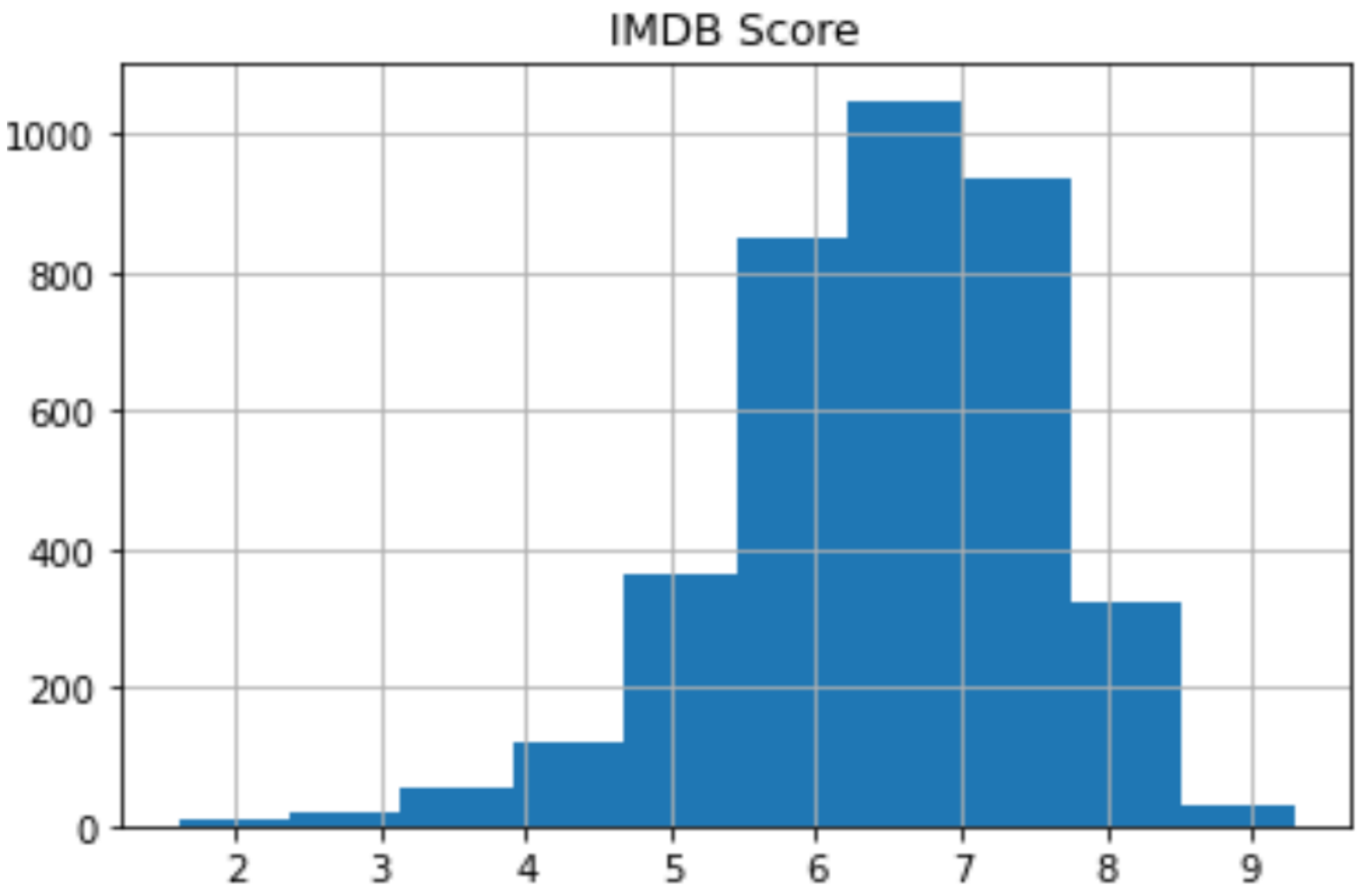

# Cast Likes
0
500
1000
1500
2000
2500
3000
3500
0
100000
200000
300000
400000
500000
600000

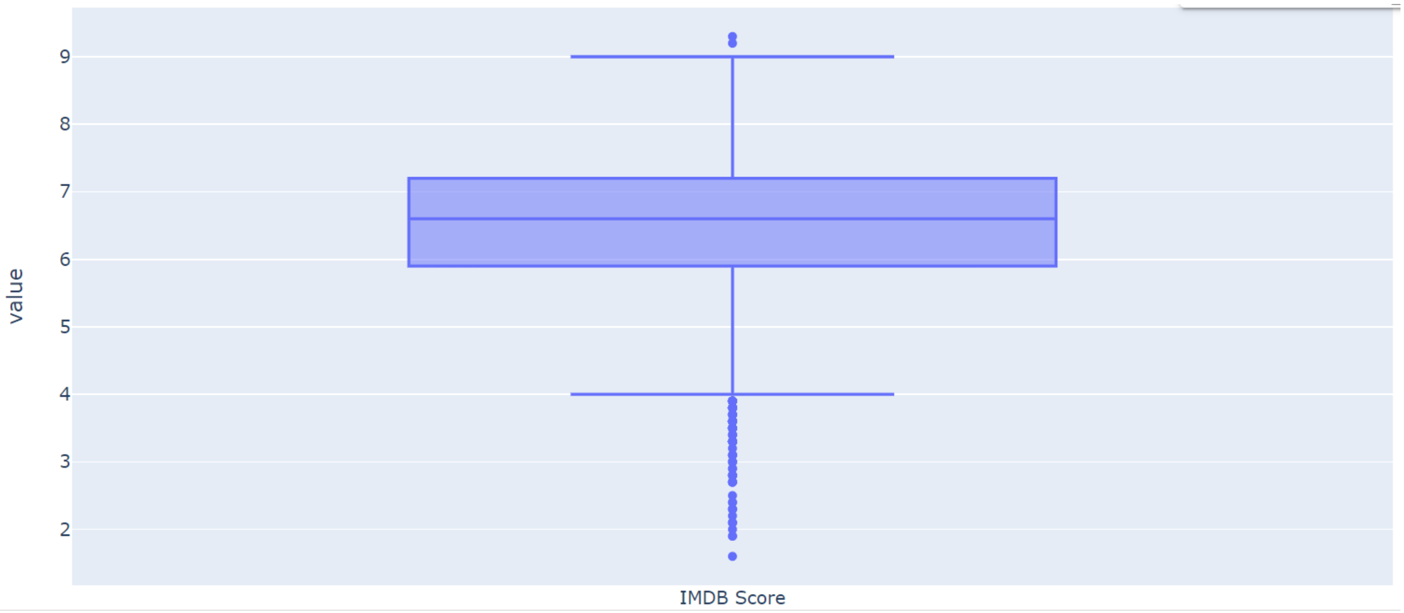

value
0
100k
200k
300k
400k
500k
600k
# Cast Likes

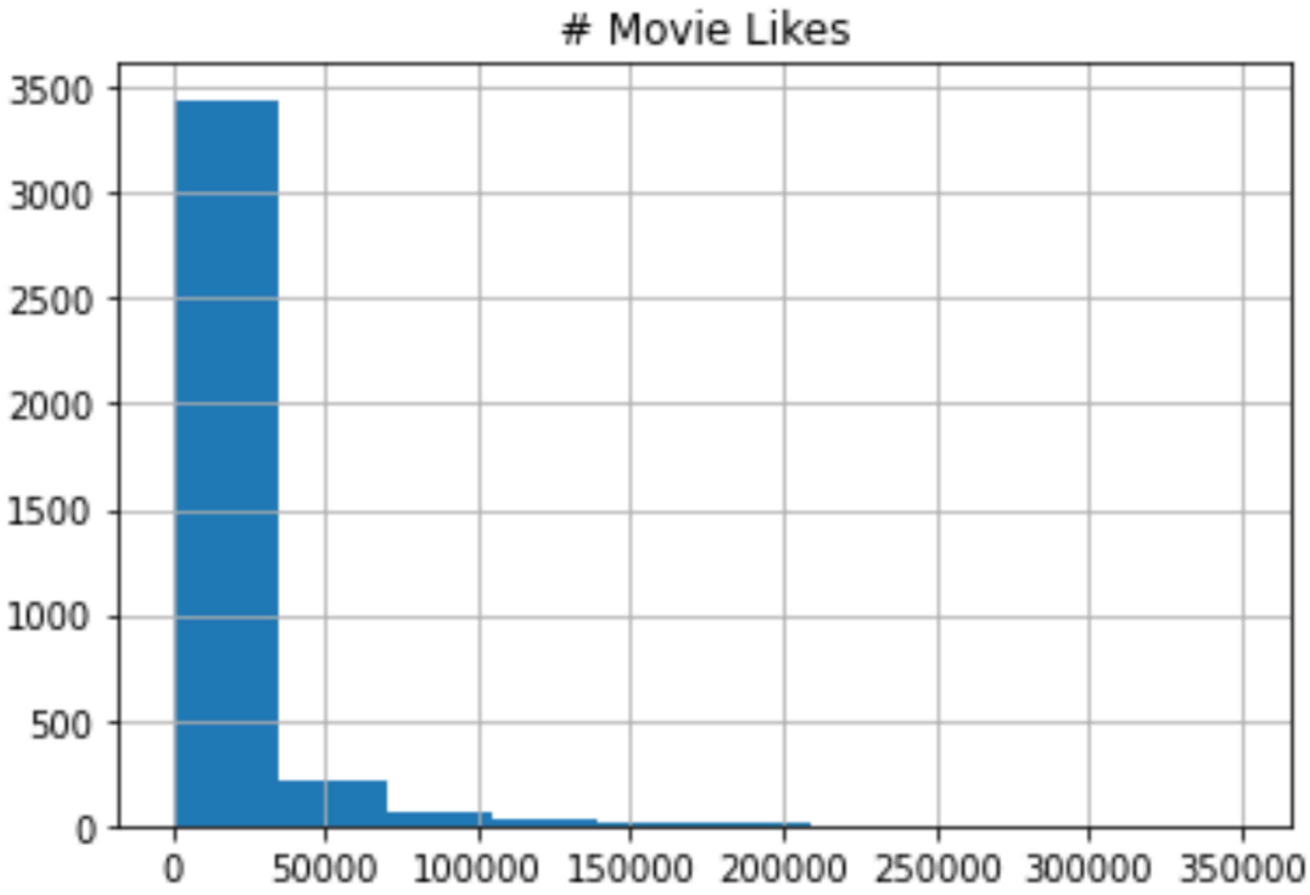

Budget
3500
3000
2500
2000
1500
1000
500
0
0.0
0.2
0.4
0.6
0.8
1.0
1.2
1e10

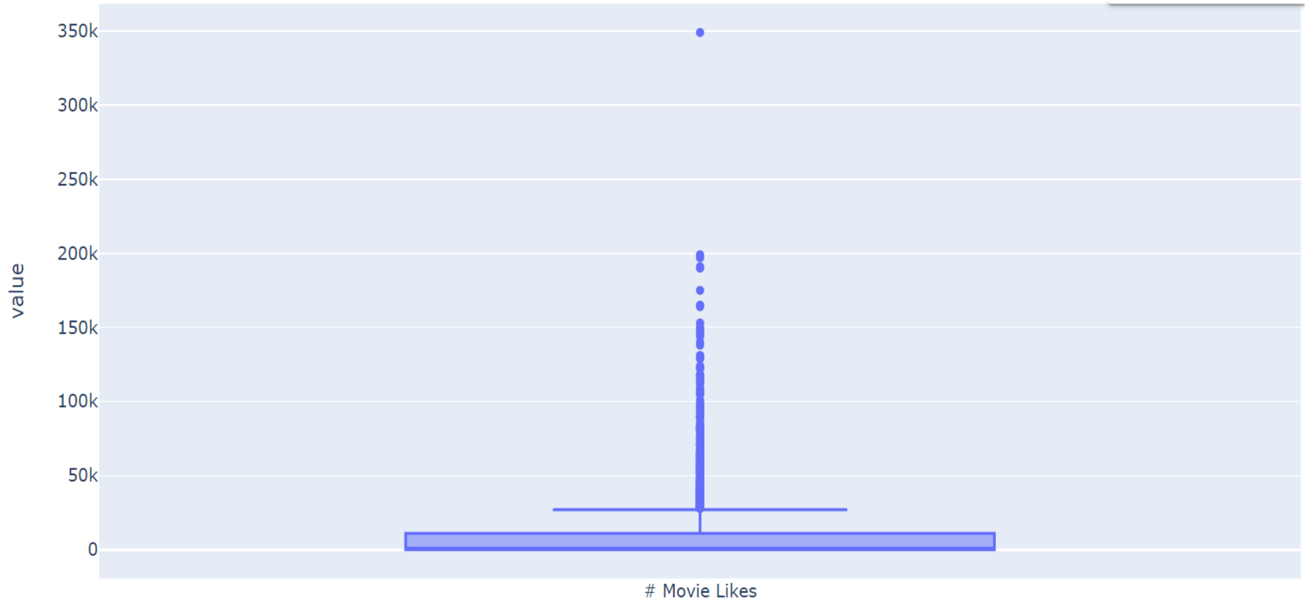

12B
10B
8B
6B
4B
2B
0
value
Budget

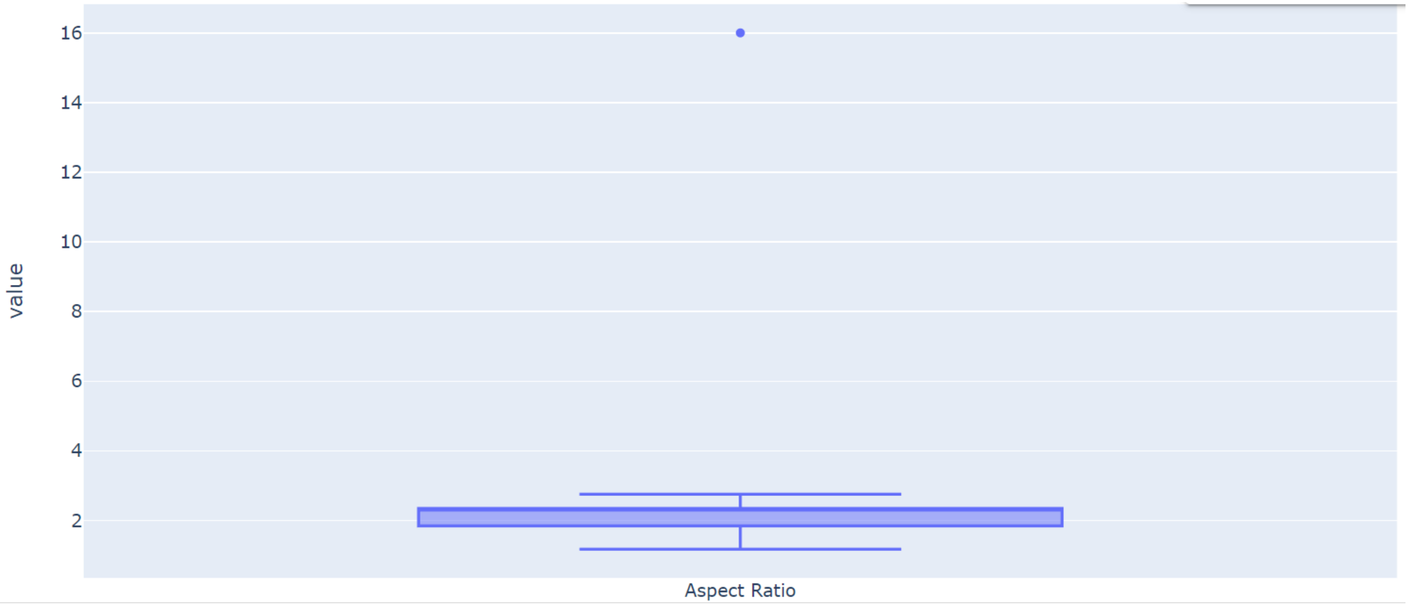

value
2020
2010
2000
1990
1980
1970
1960
1950
1940
1930
Title Year


IMDB Score
1000
800
600
400
200
0
2
3
4
5
6
7
8
9

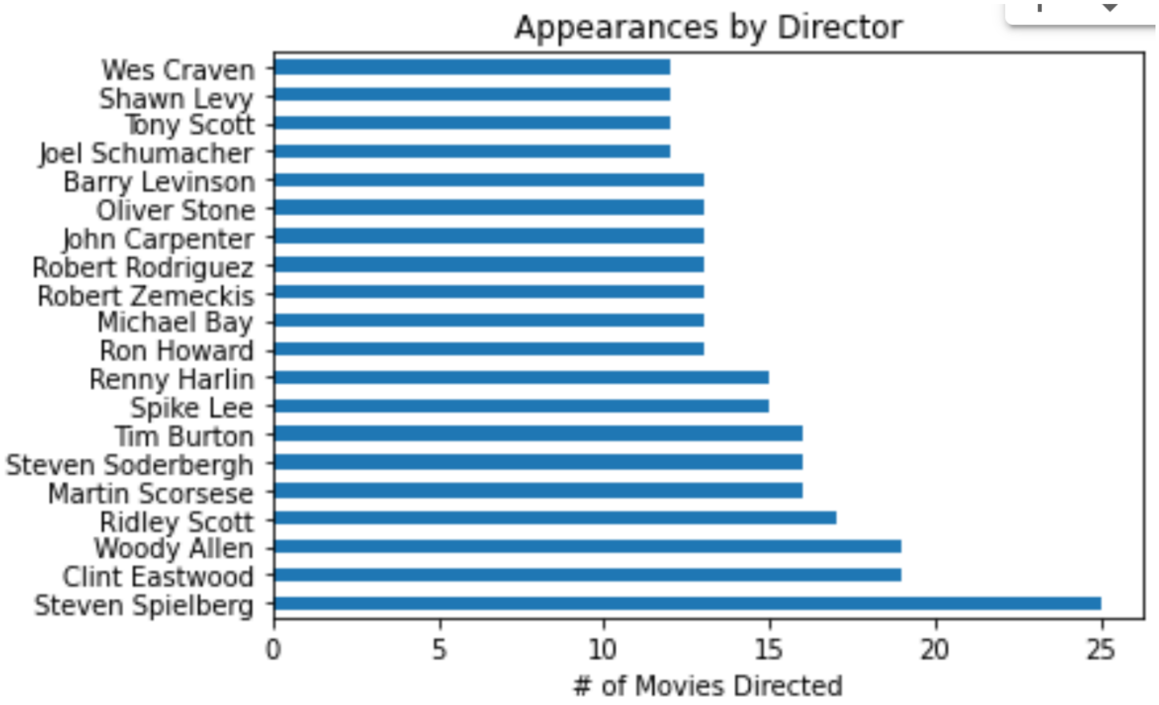

9
8
7
6
5
4
3
2
value
IMDB Score


# Movie Likes
3500
3000
2500
2000
1500
1000
500
0
0
50000
100000
150000
200000
250000
300000
350000

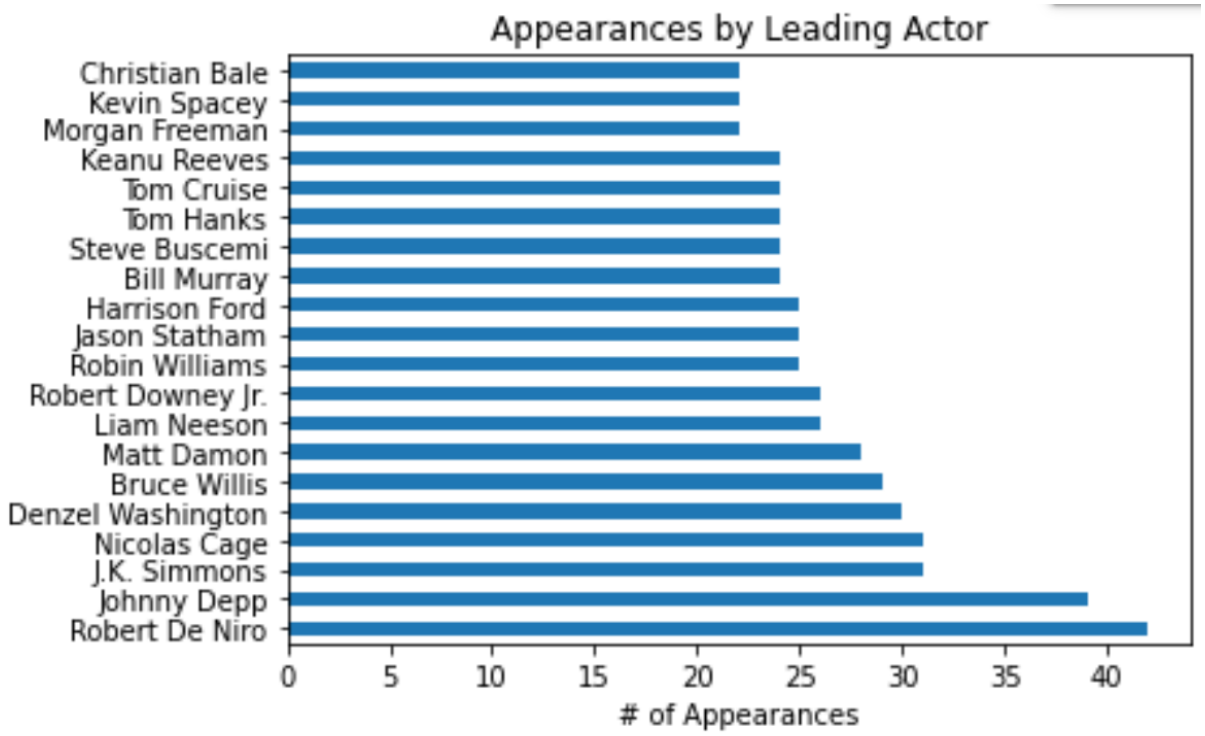

350k
300k
250k
200k
150k
100k
50k
0
value
# Movie Likes

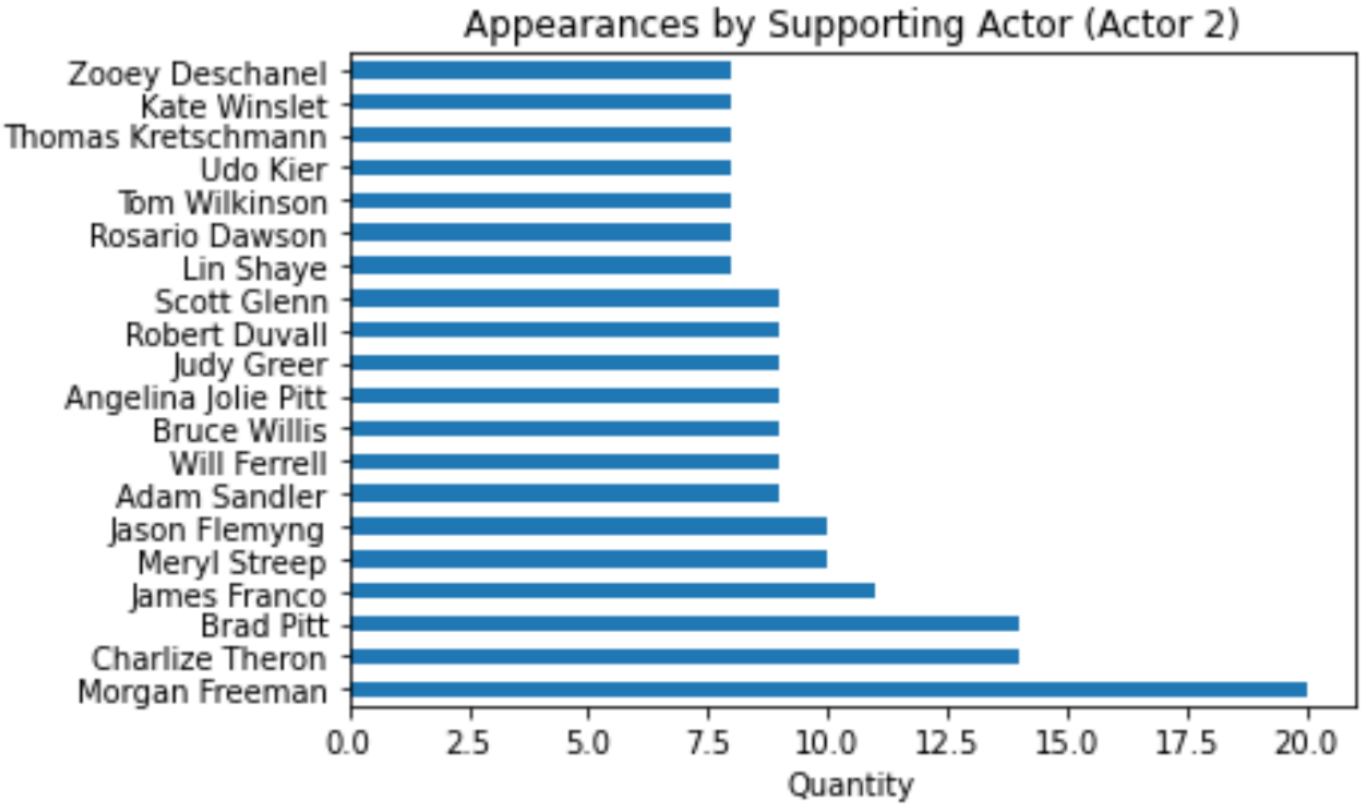

16
14
12
10
8
6
4
2
value
Aspect Ratio

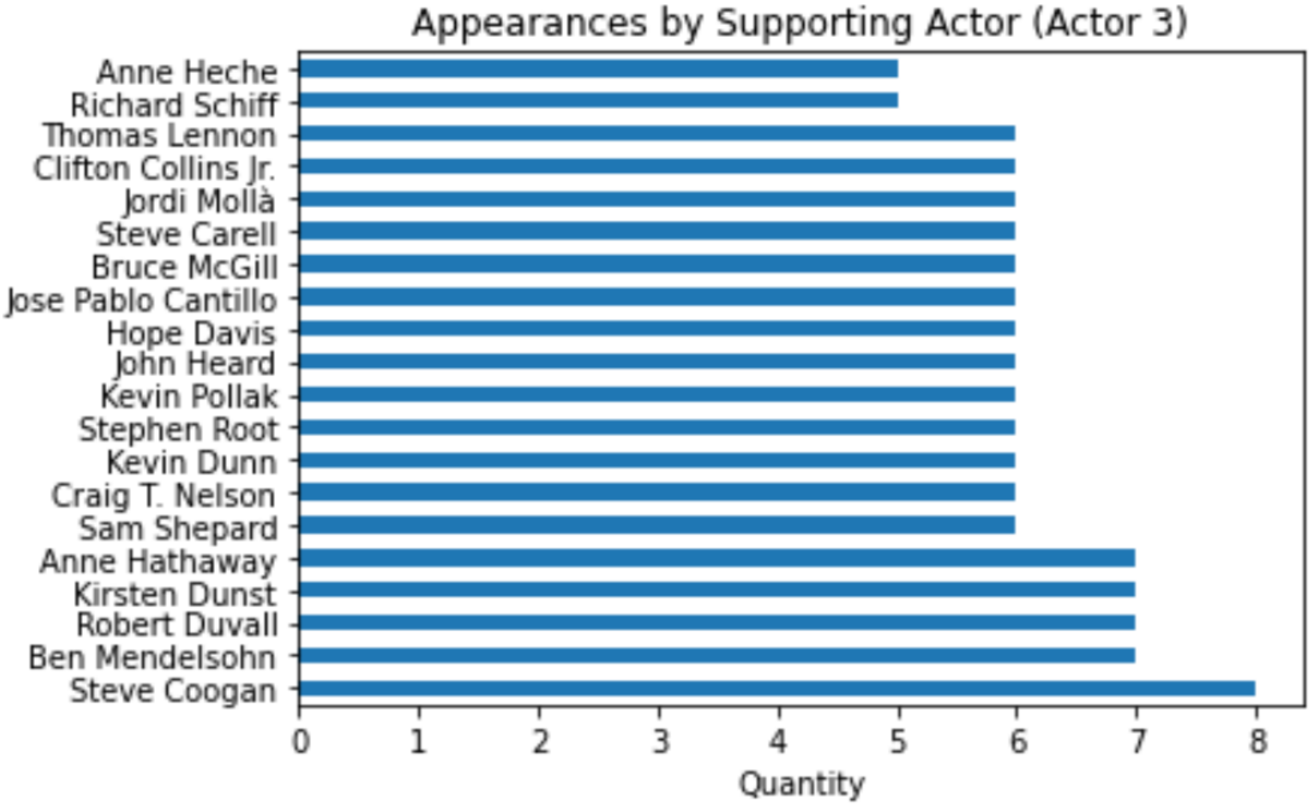
The below tables show the categorical features of interest. For personnel- related features, the top 20 highest ranking (measured in number of appearances), are displayed.

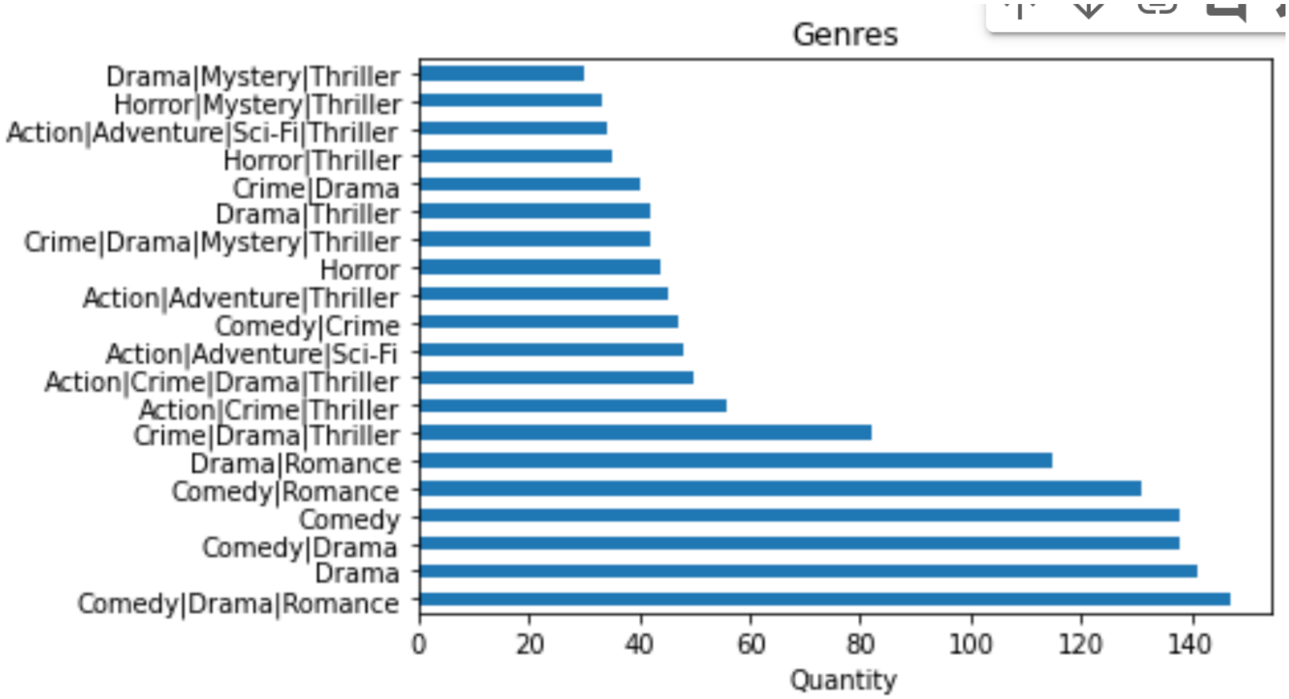


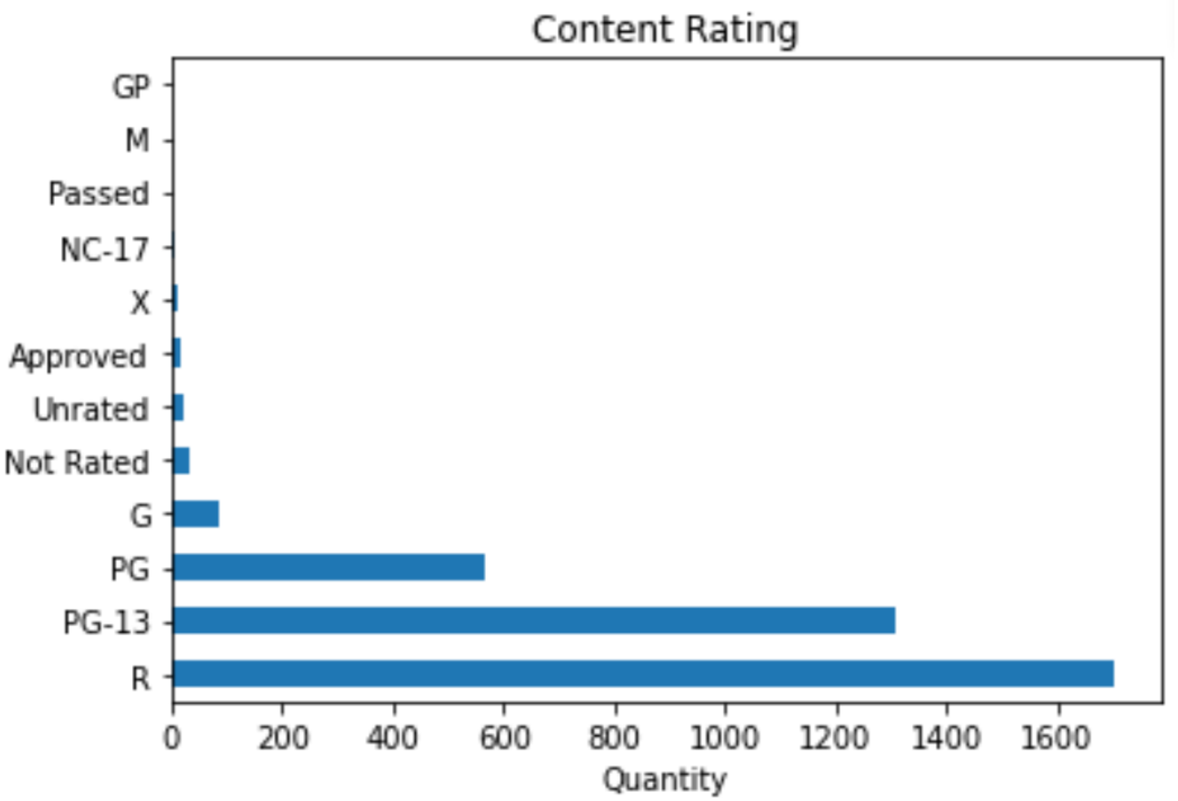
Actor 1:




Appearances by Supporting Actor (Actor 2)
Zooey Deschanel
Kate Winslet
Thomas Kretschmann
Udo Kier
Tom Wilkinson
Rosario Dawson
Lin Shaye
Scott Glenn
Robert Duvall
Judy Greer
Angelina Jolie Pitt
Bruce Willis
Will Ferrell
Adam Sandler
Jason Flemyng
Meryl Streep
James Franco
Brad Pitt
Charlize Theron
Morgan Freeman
0.0
2.5
5.0
7.5
10.0
12.5
15.0
17.5
20.0
Quantity


Appearances by Supporting Actor (Actor 3)
Anne Heche
Richard Schiff
Thomas Lennon
Clifton Collins Jr.
Jordi Mollà
Steve Carell
Bruce McGill
Jose Pablo Cantillo
Hope Davis
John Heard
Kevin Pollak
Stephen Root
Kevin Dunn
Craig T. Nelson
Sam Shepard
Anne Hathaway
Kirsten Dunst
Robert Duvall
Ben Mendelsohn
Steve Coogan
0
1
2
3
4
5
6
7
8
Quantity





This report is designed to provide an overview of the original IMDB dataset sourced from Kaggle.com. The key features of the data were explored within the data. In the next report, “Report #2- Data Preparation”, the data is further analyzed and feature manipulation is explored, in preparation for applying key machine learning algorithms.