Maddie Kelsch and Ben Spilo (Group 1) SI 206 Final Project

IMDB: https://github.com/Benspilo/Final-Project-Movie-Analysis.git

Original vs. achieved goals

General: Budget and Box office adjusted for inflation. Correlations between various factors and box office.

Graphs: The budget per minute effect on the rating at the box office; how the release date affects the rating at the box office; how different budget tiers affect the discrepancy between user and critic ratings

General: Didn't account for inflation. We got info such as: box office $, budget $, rank, release month, rating, and compared them.

Graphs: Top 250 Movies By Month of Release, Rank of Films By Budget Per Minute, Top 250 Movies Rank Compared to Budget, Budget Per Minute of Films By Box Office, Box Office By Individual Rating

Problems we ran into

- Overused daily API key amount :(→ used different email to sign up
- **Originally wanted to use wikipedia but the tags weren't consist and we couldn't scrape the data → changed to the Box Office Mojo site**
- **Missing data from the Box Office Mojo Space → used try and except to put 0's into missing data points**
  - Had to do this individually for each type of datapoint
- **Data didn't conform to our expectations (i.e. 45 mins or 2 hrs) → changed our regex expression**
- Brief confusion on how to limit to 25 items → used SQL MAX() function

Calculation Files:

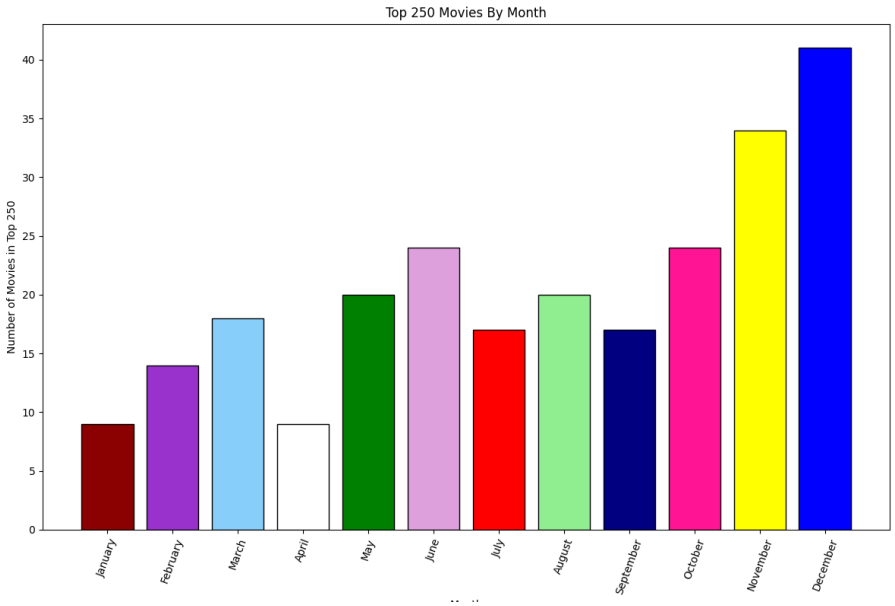
1.

```
Film Count by Month
"('January', 9)"
"('February', 14)"
"('March', 18)"
"('April', 9)"
"('May', 20)"
"('June', 24)"
"('July', 17)"
"('August', 20)"
"('September', 17)"
"('October', 24)"
"('November', 34)"
"('December', 41)"
```

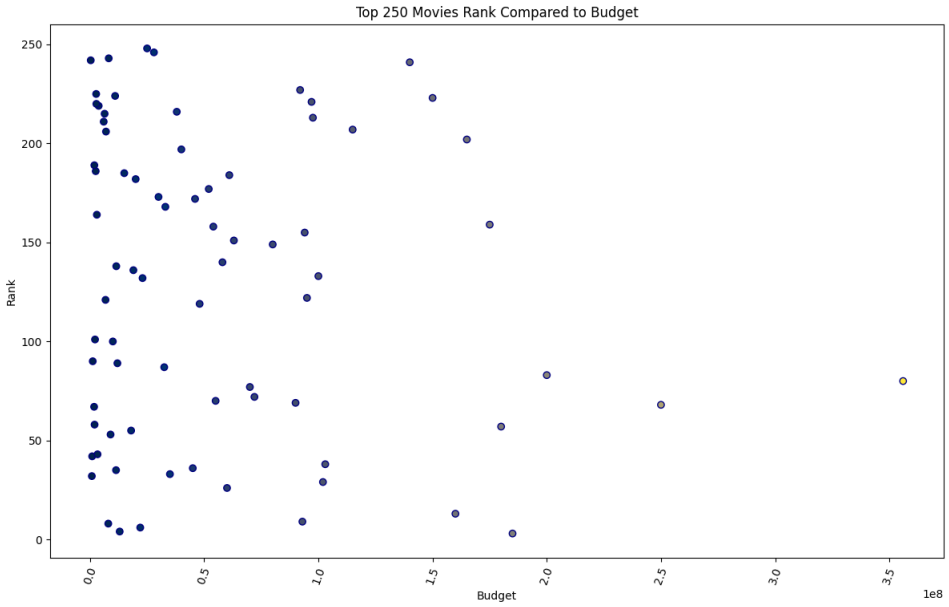
2.

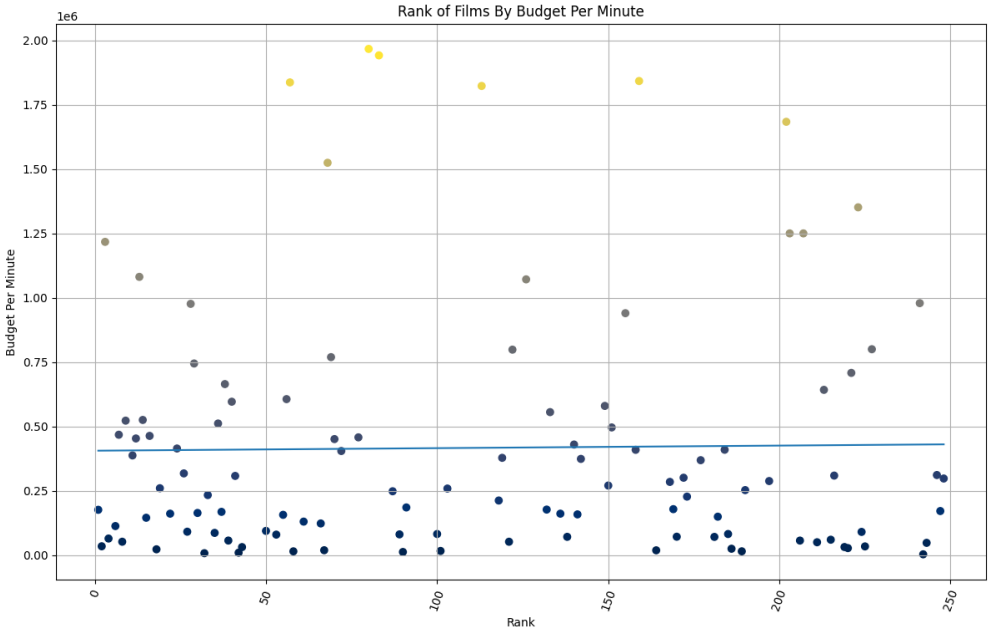
```
Rank of Films By Budget Per Minute
"(1, 176056)"
"(2, 34286)"
"(3, 1217105)"
"(4, 64356)"
"(6, 112821)"
"(7, 467662)"
"(8, 51948)"
"(9, 522472)"
"(11, 387324)"
"(12, 453237)"
"(13, 1081081)"
"(14, 525140)"
"(15, 145161)"
"(16, 463235)"
"(18, 22556)"
"(19, 259843)"
"(22, 161017)"
"(24, 414201)"
"(26, 317460)"
"(27, 90909)"
"(28, 976331)"
"(29, 744526)"
"(30, 163793)"
"(32, 7403)"
"(33, 233333)"
"(35, 86364)"
"(36, 511364)"
"(37, 168067)"
"(38, 664516)"
"(39, 56604)"
"(40, 596026)"
"(41, 307692)"
"(42, 9314)"
"(43, 31132)"
"(50, 94017)"
"(53, 79646)"
"(55, 156522)"
"(56, 606061)"
"(57, 1836735)"
"(58, 14599)"
"(61, 130137)"
"(66, 122951)"
"(67, 18947)"
"(68, 1524390)"
"(69, 769231)"
"(70, 450820)"
"(72, 404494)"
"(77, 457516)"
"(80, 1966851)"
"(83, 1941748)"
"(87, 248092)"
"(89, 80537)"
"(90, 12121)"
"(91, 185185)"
"(100, 81967)"
"(101, 16176)"
"(103, 258621)"
"(113, 1822917)"
"(118, 212121)"
"(119, 377953)"
"(121, 51908)"
"(122, 798319)"
"(126, 1071429)"
"(132, 176923)"
"(133, 555556)"
"(136, 161017)"
"(138, 70988)"
"(140, 429630)"
"(141, 158228)"
"(142, 373832)"
"(149, 579710)"
"(150, 270270)"
"(151, 496063)"
"(155, 940000)"
"(158, 409091)"
"(159, 1842105)"
"(164, 18634)"
"(168, 284483)"
"(169, 178571)"
"(170, 71429)"
"(172, 300654)"
"(173, 227273)"
"(177, 368794)"
"(181, 70755)"
"(182, 149254)"
"(184, 409396)"
"(185, 81967)"
"(186, 24752)"
"(189, 14729)"
"(190, 252525)"
"(197, 287770)"
"(202, 1683673)"
"(203, 1250000)"
"(206, 56452)"
"(207, 1250000)"
"(211, 50000)"
"(213, 642105)"
"(215, 59813)"
"(216, 308943)"
"(219, 31405)"
"(220, 27225)"
```

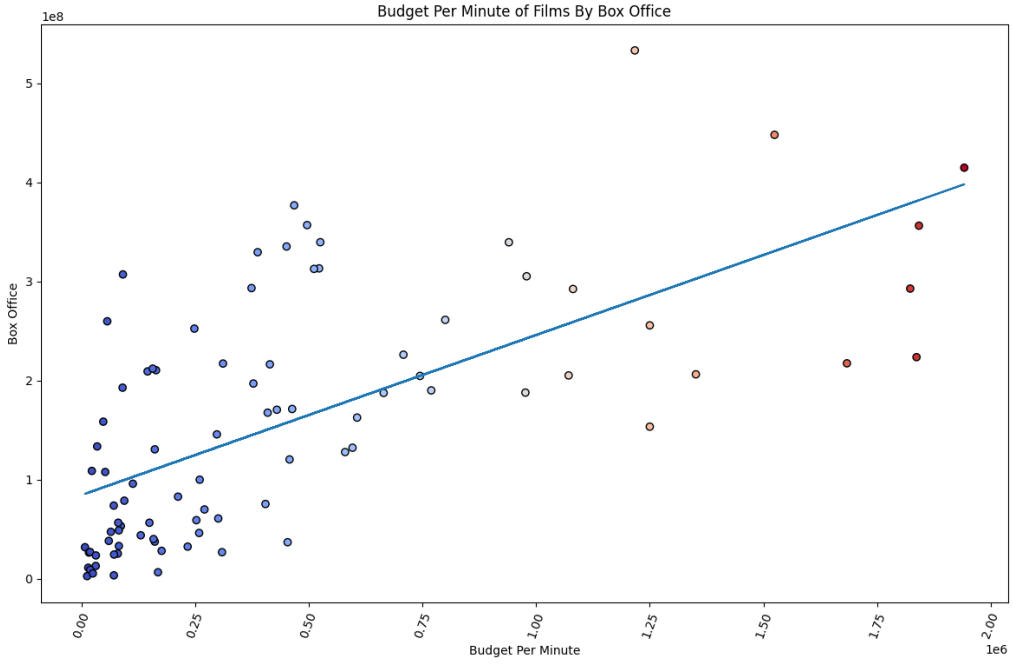
Visualizations:

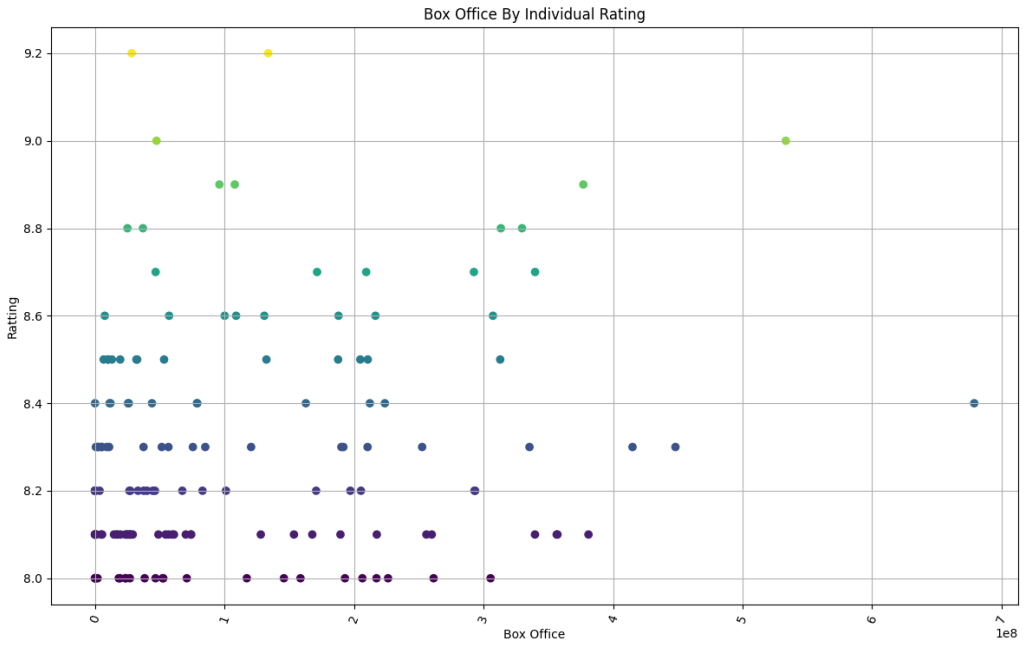


*** month was cut off in the screenshot but is there in the actual graph.




Rank of Films By Budget Per Minute
1e6
2.00
1.75
1.50
1.25
1.00
0.75
0.50
0.25
0.00
Budget Per Minute
0
50
100
150
200
250
Rank


Budget Per Minute of Films By Box Office
1e8
5
4
3
2
1
0
Box Office
0.00
0.25
0.50
0.75
1.00
1.25
1.50
1.75
2.00
1e6
Budget Per Minute



How to run our project

- Our Project has two primary files
  - API-Interaction.py: to build the database
  - Graphs-analysis.py: to create the graphs from the info in the database
- We start by running API-Interaction.py 10x to get all the data in our tables
- Then, simply run Graphs-analysis.py once and you will have all five graphs

Code Documentation inside the two files (API-Interaction.py and Graphs-analysis.py)

Works Cited

https://www.statology.org/line-of-best-fit-python/

https://stackoverflow.com/questions/36596118/how-to-sort-a-python-dictionary-having-month-names-as-keys

https://www.analyticsvidhya.com/blog/2020/02/beginner-guide-matplotlib-data-visualization-exploration-python/

https://www.statology.org/matplotlib-scatterplot-color-by-value/

https://python-graph-gallery.com/3-control-color-of-barplots

https://stackoverflow.com/questions/22023992/beautifulsoup-parent-tag

https://phrase.com/blog/posts/beginners-guide-to-locale-in-python/