Rentomojo
Assignment

Submitted by:-

Rishabh kapoor
rishabhkapoor101@gmail.com

Acknowledgement

First and foremost I would like to thank “RentMojo” for entrusting me with a project that gave me a chance to prove my value that I can provide to the company if I am chosen for the position. Also I would like to thank all my professors that taught me the tools necessary for such work. And last but not the least I would like to thank my family that has been pushing me forward since day one.

Without any further ado let's have a gander on the project.

(Also for a comfortable read please download full in pdf format from my github :- https://github.com/rishabhkapoor101/MoviesAnalysis)

Dataset

So the data set was a csv formatted file that contained the data about the movies.
It had 24 columns namely:-

1. index
2. budget
3. genres
4. homepage
5. id
6. keywords
7. original_language
8. original_title
9. overview
10. popularity
11. production_companies
12. production_countries
13. release_date
14. revenue
15. runtime
16. spoken_languages
17. status
18. tagline
19. title
20. vote_average
21. vote_count
22. cast
23. crew
24. director

With a row count of 4803.

The data set with a skim, seemed a piece of cake to work with, but after just digging a little, I realised that it was not the case.

I had to clean the data as there were “nan” values that essentially means “not a number” even when a number was expected.

This was not everything that was a challenge about the dataset, in fact we will see much more as we will progress through the dataset.

Toolset

The tools I used for this assignment are as follows:-

- Python3.8
- Pandas
- Jupyter notebook
- Microsoft Excel
- Tableau
- Pdf viewer
- github

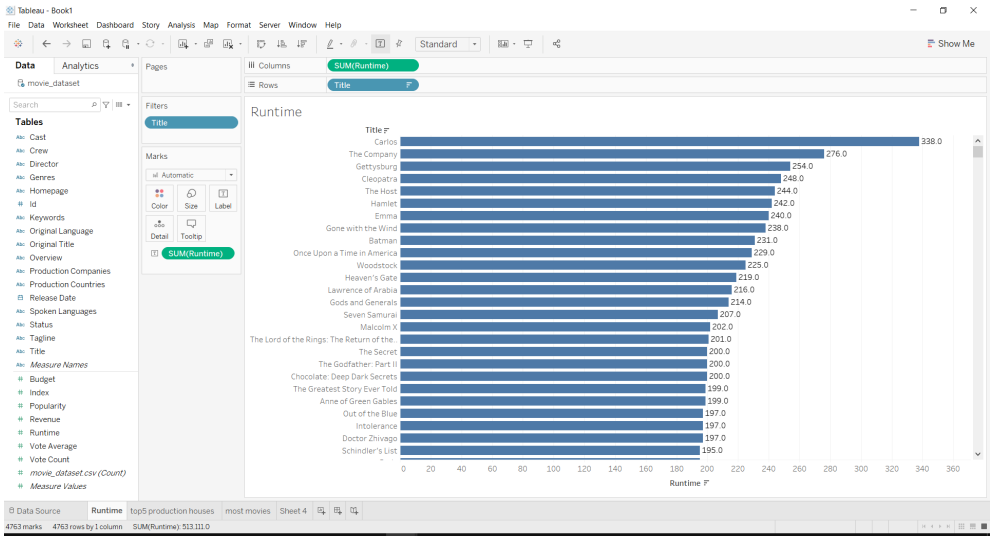
Task 1

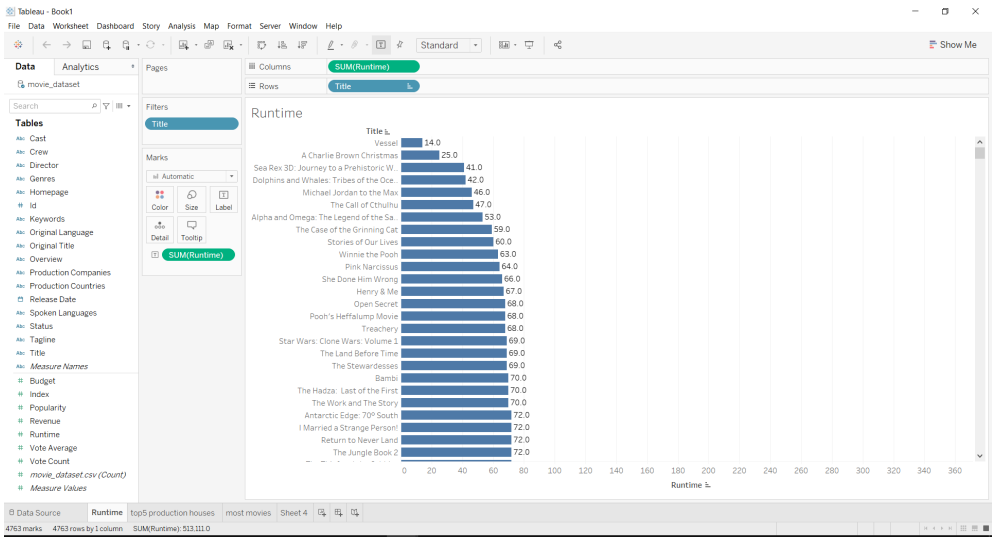
For task 1, I had to find the movies at contrasting ends of the runtime spectrum, i.e. I had to find the movies which had the highest and least runtime.

I did this without any difficulty in Tableau.

<u>Result</u> :- Movie with least runtime was **Vessel** with a runtime of **14 min** and
Movie with highest runtime was **Carlos** with a runtime of **338 min**

I have also attached the screenshot of my findings as follows...





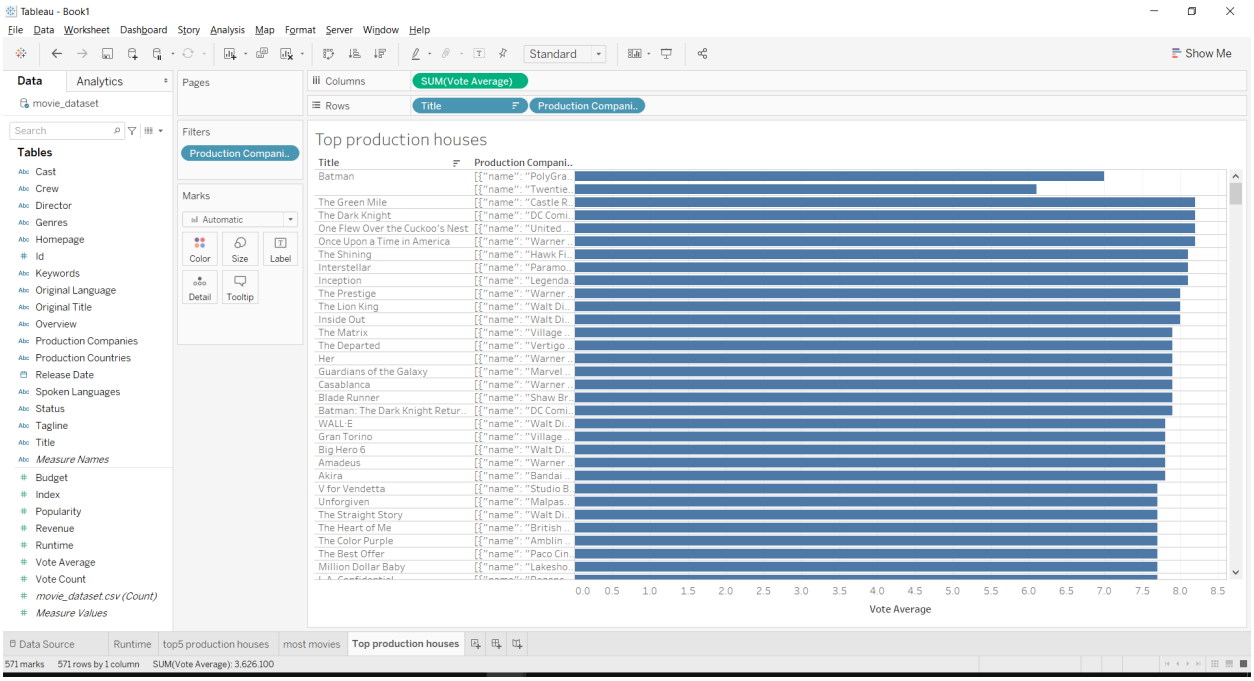
Task 2

For the second task I had to find the top five production houses with highest vote average, revenue and then find their 5 movies each.

What I found was that the biggest corporations according to the budget were:

- Walt disney
- Warner bros
- Dc comics
- Marvel
- BBc

And their respective movies are attached as follows:-

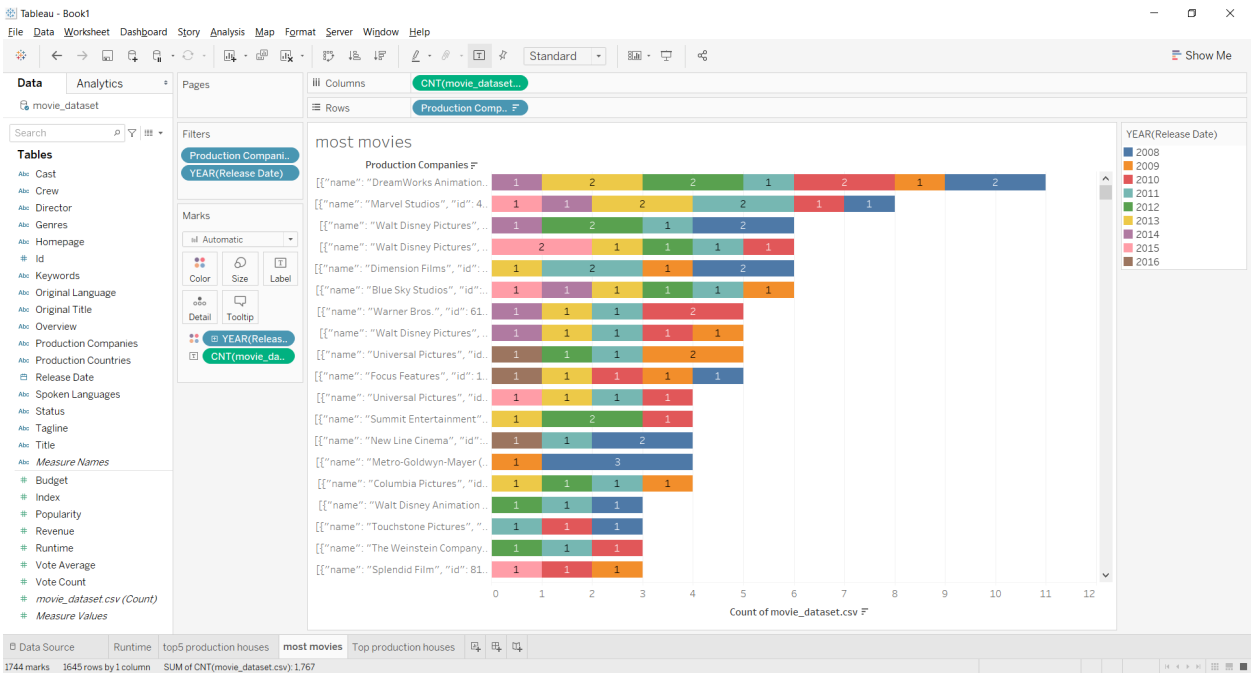


(for a cleaner look, please refer my book1.twb file)

Task 3

For task three, I had to list the production houses that released most movies in the period of 2006-2016

That was fairly easy and the sheet below will be self explanatory.



(I may sound a bit redundant, but for a closer and cleaner look, have a look at my book1.twb)

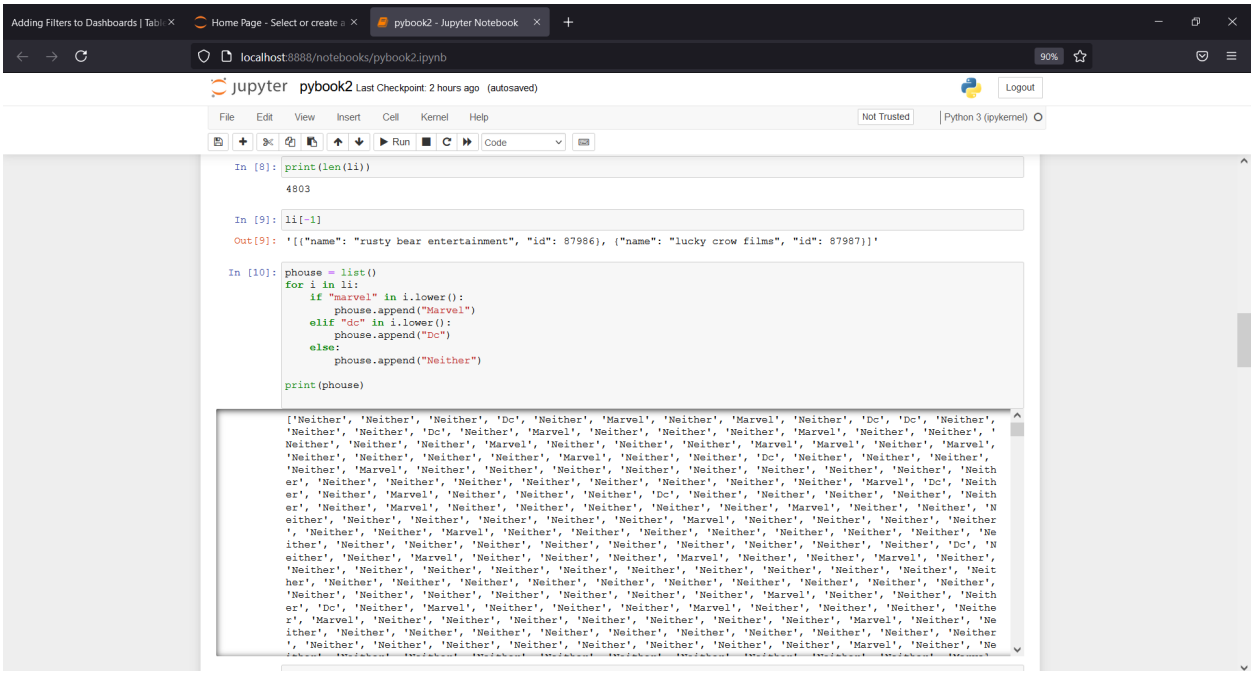
Taks 4

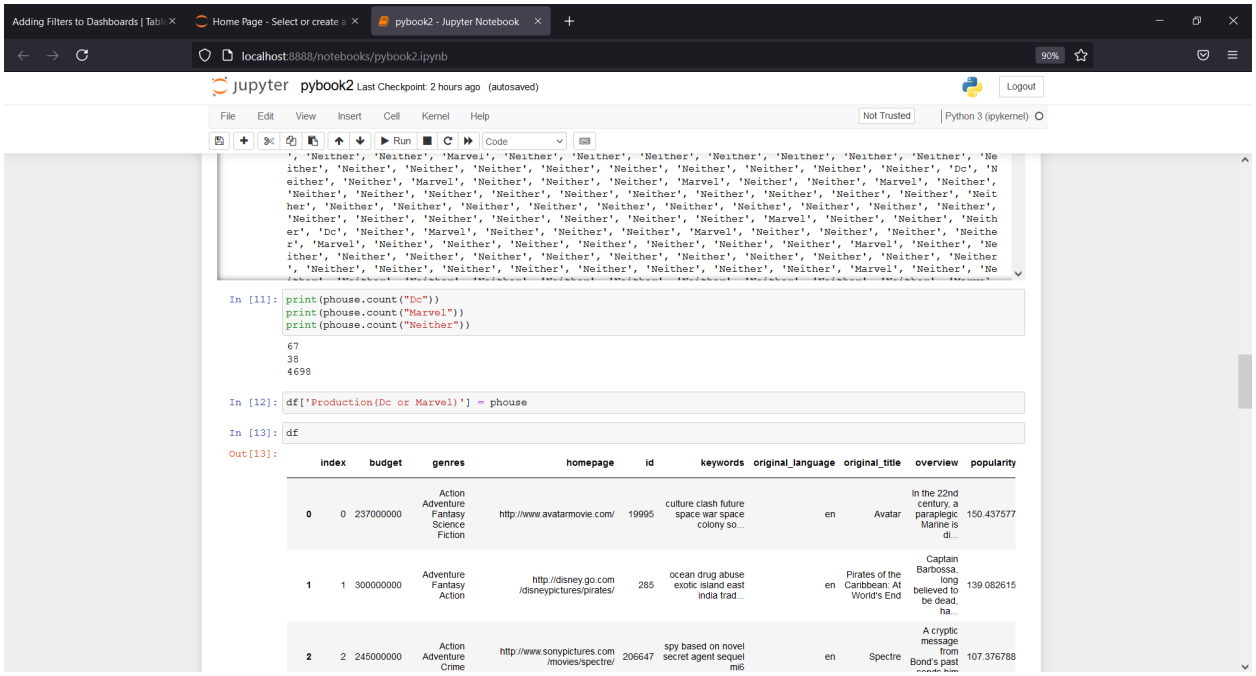
For this task I had to choose a company between DC comics and Marvel studios where I would invest all my money had I been given the option.
This was the question where I felt the most heat.

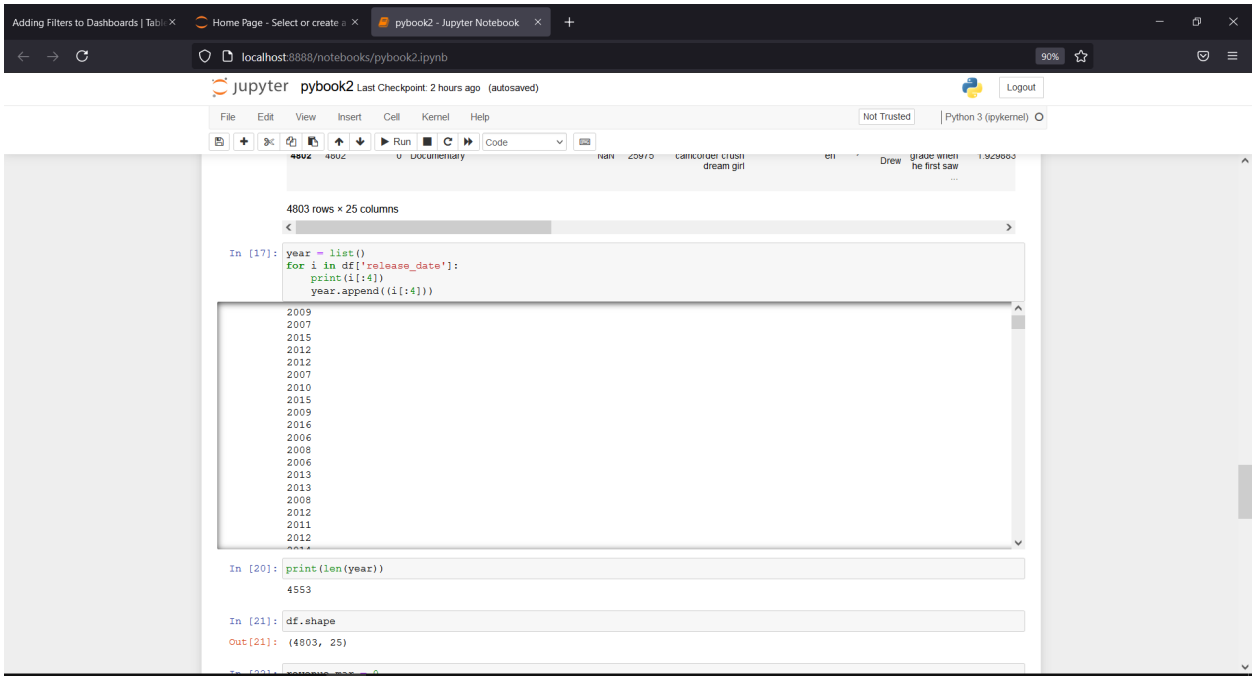
My approach was to compare the growth rate of the two companies rather than the raw revenue and that was a challenge in itself, I had to separate years from full dates, I had to classify them in production_house categories, which essentially were Dc, Marvel or Neither. And then I had to plot new info in the tableau. For which I had to make a new dataframe from modified values for plotting which I did make with the name of **“modified.csv”**

Everything except for the plotting was done in python3.8 and pandas in jupyter notebook.

I have attached some screenshots of the same…

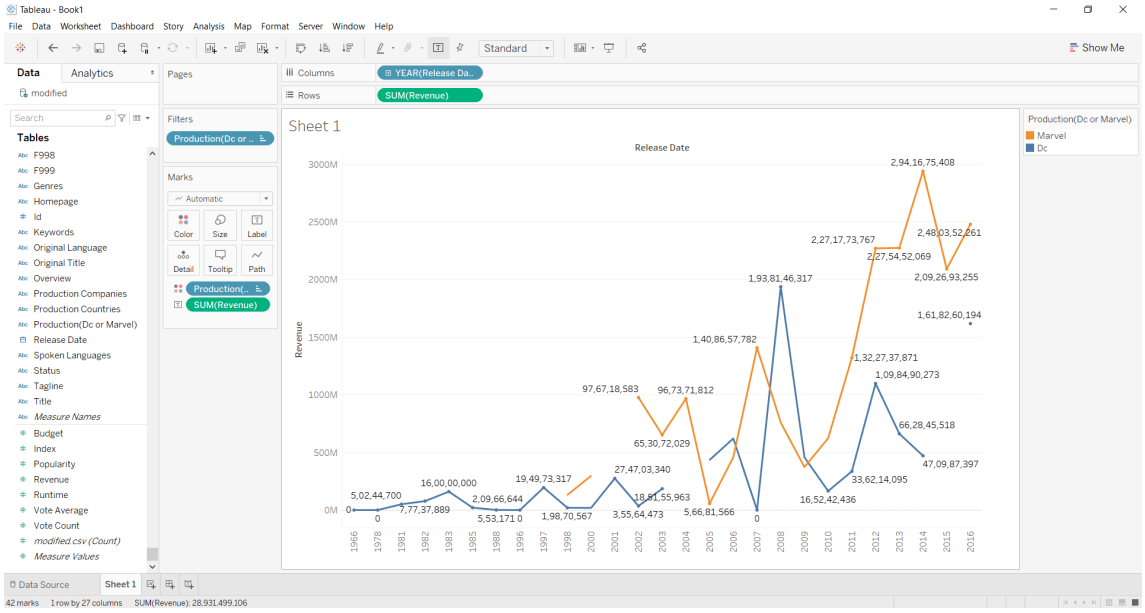


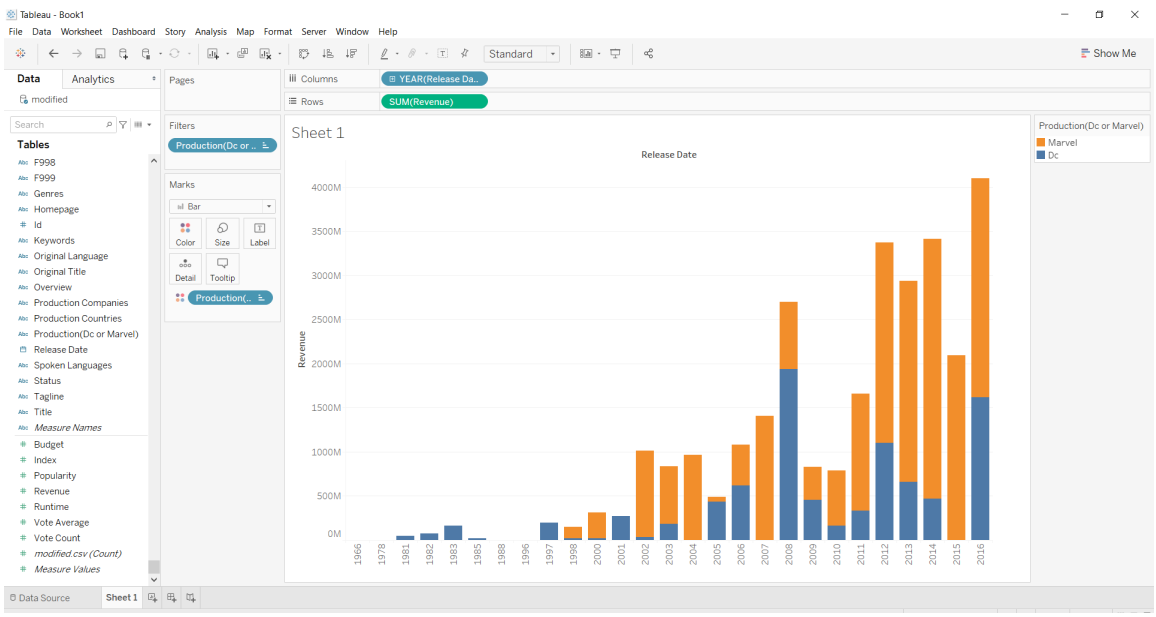

Adding Filters to Dashboards | Tabl
Home Page - Select or create a
pybook2 - Jupyter Notebook
localhost:8888/notebooks/pybook2.ipynb
jupyter pybook2 Last Checkpoint: 2 hours ago (autosaved)
Logout
File Edit View Insert Cell Kernel Help
Not Trusted
Python 3 (ipykernel)
Run
Code
', 'Neither', 'Neither', 'Marvel', 'Neither', 'Neither', 'Neither', 'Neither', 'Neither', 'Neither', 'Neither', 'Ne
ither', 'Neither', 'Neither', 'Neither', 'Neither', 'Neither', 'Neither', 'Neither', 'Neither', 'Neither', 'Dc', 'N
either', 'Neither', 'Marvel', 'Neither', 'Neither', 'Neither', 'Marvel', 'Neither', 'Neither', 'Marvel', 'Neither',
'Neither', 'Neither', 'Neither', 'Neither', 'Neither', 'Neither', 'Neither', 'Neither', 'Neither', 'Neither', 'Neit
her', 'Neither', 'Neither', 'Neither', 'Neither', 'Neither', 'Neither', 'Neither', 'Neither', 'Neither', 'Neither',
'Neither', 'Neither', 'Neither', 'Neither', 'Neither', 'Neither', 'Marvel', 'Neither', 'Neither', 'Neith
er', 'Dc', 'Neither', 'Marvel', 'Neither', 'Neither', 'Neither', 'Marvel', 'Neither', 'Neither', 'Neither', 'Neithe
r', 'Marvel', 'Neither', 'Neither', 'Neither', 'Neither', 'Neither', 'Neither', 'Neither', 'Marvel', 'Neither', 'Ne
ither', 'Neither', 'Neither', 'Neither', 'Neither', 'Neither', 'Neither', 'Neither', 'Neither', 'Neither', 'Neither
', 'Neither', 'Neither', 'Neither', 'Neither', 'Neither', 'Neither', 'Neither', 'Neither', 'Marvel', 'Neither', 'Ne
In [11]:
print(phouse.count("Dc"))
print(phouse.count("Marvel"))
print(phouse.count("Neither"))
67
38
4698
In [12]:
df['Production(Dc or Marvel)'] = phouse
In [13]:
df
Out[13]:
index budget genres homepage id keywords original_language original_title overview popularity
0 0 237000000 Action Adventure Fantasy Science Fiction http://www.avatarmovie.com/ 19995 culture clash future space war space colony so... en Avatar In the 22nd century, a paraplegic Marine is di... 150.437577
1 1 300000000 Adventure Fantasy Action http://disney.go.com/disneypictures/pirates/ 285 ocean drug abuse exotic island east india trad... en Pirates of the Caribbean: At World's End Captain Barbossa, long believed to be dead, ha... 139.082615
2 2 245000000 Action Adventure Crime http://www.sonypictures.com/movies/spectre/ 206647 spy based on novel secret agent sequel mi6 en Spectre A cryptic message from Bond's past sends him 107.376788


Adding Filters to Dashboards | Tabl
Home Page - Select or create a
pybook2 - Jupyter Notebook
localhost:8888/notebooks/pybook2.ipynb
jupyter pybook2 Last Checkpoint: 2 hours ago (autosaved)
Logout
File Edit View Insert Cell Kernel Help
Not Trusted
Python 3 (ipykernel)
Run
Code
4803 rows × 25 columns
In [17]:
year = list()
for i in df['release_date']:
print(i[:4])
year.append((i[:4]))
2009
2007
2015
2012
2012
2007
2010
2015
2009
2016
2006
2008
2006
2013
2013
2008
2012
2011
2012
In [20]:
print(len(year))
4553
In [21]:
df.shape
Out[21]:
(4803, 25)

CHARTS

Below, the charts compare the revenue year wise ie growth and needless to say Marvel is the winner even though Dc started prior to Marve, so if I am given a chance to invest my money, without a shred of doubt it is going to be marvel.





(for a clearer look, refer to book2.twb)

Workflow…

As I already, have mentioned the tools I used, lets see how were they used,

- Python3.8 was very helpful in data cleaning and other manipulations.
  Also it was a massive help in getting to know the data on a very intimate level.

- Pandas was used via python to read and manipulate data frames easily.

- Jupyter notebooks were used for better formation visual representation of dataframes.

- And Github was used to make a public repo for this very project.

Thank you.

I hope you enjoyed reading this report.

A report by:-
Rishabh kapoor.
rishabhkapoor101@gmail.com