Assignment 6 (Mid) - Part ii

DATAWAREHOUSING BY DR. IMRAN KHAN

Part 1: Data Scrapping

I scrapped the data with help of python library BeautifulSoap, I used Jupyter Lab to use IMDB URLs to scrape data one by one with different years. I scrapped data 5 times each of different year, getting the data of 50 movies at one time. When I got 5 .csv files, I manually combined the data of all 5 of them through excel.

URL for 2015: https://www.imdb.com/list/ls071776440/

URL for 2016: https://www.imdb.com/list/ls031269618/

URL for 2017: https://www.imdb.com/list/ls062905646/

URL for 2018: https://www.imdb.com/list/ls021105452/

URL for 2019: https://www.imdb.com/list/ls041125816/

Some of the difficulties I faced while scrapping were:

1: Directors, and Movie stars were written as part of one String, therefore I had to use the Cast function to separate them. Same was the case with movie runtime and movie gross.

2: Some of the data were missing like not every movie had gross amount present, and there were many stars, so it was not practically possible to take every one of them, therefore I only took the 2 most important ones.

Part 2: Data Cleaning

I did all the cleanings manually through excel. Initially I used a list to store genres as there were more than one for each movie, same was the case with stars. Therefore, I had to use LEFT, RIGHT, ROW to COLUMN functions to divide them as different values, furthermore I only took the 2 most important stars and movie genres.

The .csv file was not properly first loading on Tableau this was the reason I emailed you regarding using Power BI, but Hamza(our TA) helped me out and it turned out that the problem was solved by using the excel file instead of csv, therefore I saved it as the excel file before going towards Data visualization.

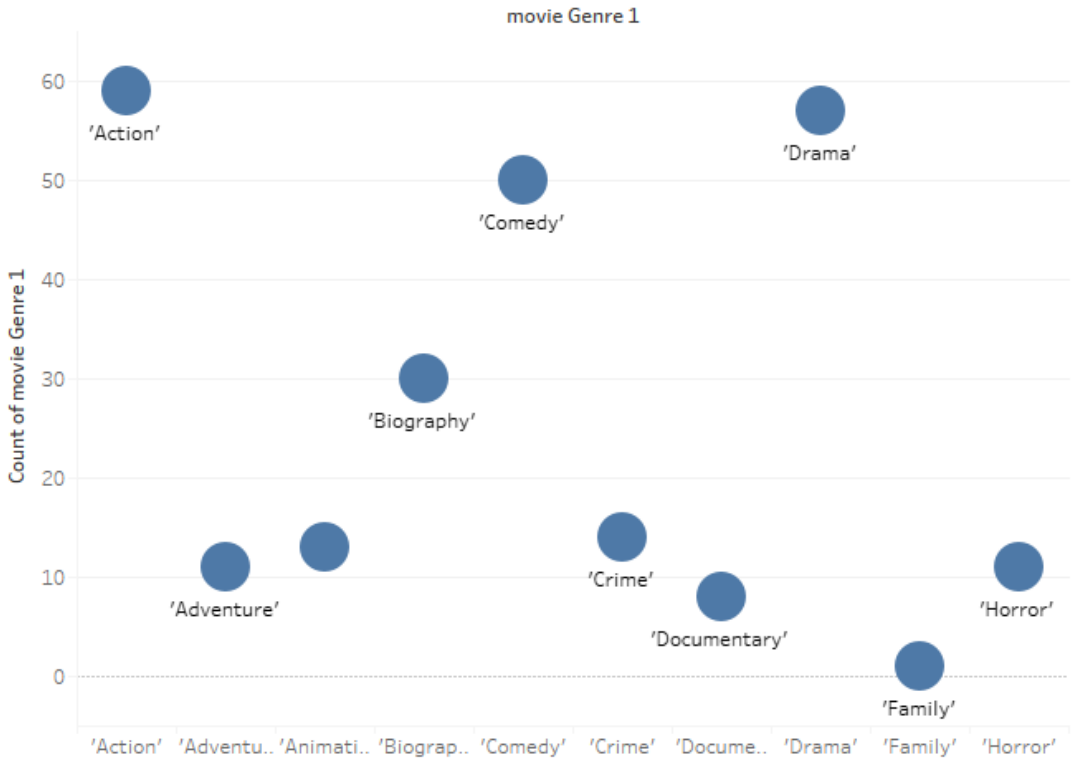
Following is the code snippet of one of the .ipynb file, I will attach all 5 of them along with this report.

```
import pandas as pd
import numpy as np
import re
import lxml

from bs4 import BeautifulSoup
from requests import get
%matplotlib inline

url1="https://www.imdb.com/list/ls071776440/"

page = get(url1)
soup = BeautifulSoup(page.content, 'lxml')
content = soup.find(id="main")

articleTitle = soup.find("h1", class_="header").text.replace("\n","")
print(articleTitle)

movieTitle= []
movieDate= []
movieRunTime= []
movieGenre= []
movieRating= []
movieDesc= []
movieVotes= []
movieGross= []
movieStars= []
movieDirector= []
movieFrame = content.find_all("div", class_="lister-item mode-detail")
w=0
while w < 51:
    movieFirstLine = movieFrame[w].find("h3", class_="lister-item-header")
    movieTitle.append(movieFirstLine.find("a").text)
    movieDate.append( re.sub(r"[()]","", movieFirstLine.find_all("span")[-1].text.strip()))
    movieRunTime.append(movieFrame[w].find("span", class_="runtime").text[:-4])
    movieGenre.append(movieFrame[w].find("span", class_="genre").text.rstrip().replace("\n","").split(","))
    movieRating.append(movieFrame[w].find("span", class_="ipl-rating-star__rating").text)
    movieDesc.append(movieFrame[w].find_all("p",class_="")[0].text)
    movieNumbers=movieFrame[w].find_all("span", attrs={"name": "nv"})
    if len(movieNumbers) == 2:
        movieVotes.append(movieNumbers[0].text)
        movieGross.append(movieNumbers[1].text)
    else:
        movieVotes.append(movieNumbers[0].text)
        movieGross.append(np.nan)
    movieCast = movieFrame[w].find_all("p", class_="text-muted")[1]
        #movieDirector = movieCast.find("a").text
        #print(movieDirector)
    try:
        cast = movieCast.text.replace("\n","").split('|')
        cast = [x.strip() for x in cast]
        cast = [cast[i].replace(j, "") for i,j in enumerate(["Director:", "Stars:"])]
        movieDirector.append(cast[0])
        movieStars.append([x.strip() for x in cast[1].split(",")])
    except:
        casts = movieCast.text.replace("\n","").strip()
        movieDirector.append(np.nan)
        movieStars.append(([x.strip() for x in casts.split(",")]))
    print(w)
    w+=1

Movies check = { "movie_title" : movieTitle, "movie_date" : movieDate,"movie_Runtime" : movieRunTime, "movie_Genre" : movieGenre, "movie_Ratings" : movieRating,
"movie_Storyline" : movieDesc, "movie_Votes" : movieVotes, "movie_Stars" : movieStars, "movie_Director" : movieDirector, "movie_Gross" : movieGross}

pd.DataFrame(Movies check).to_csv('E:\IBA resoures\Semester 8\DW\imdb\Movies check.csv')
```

PART 3 DATA VISUALISATION

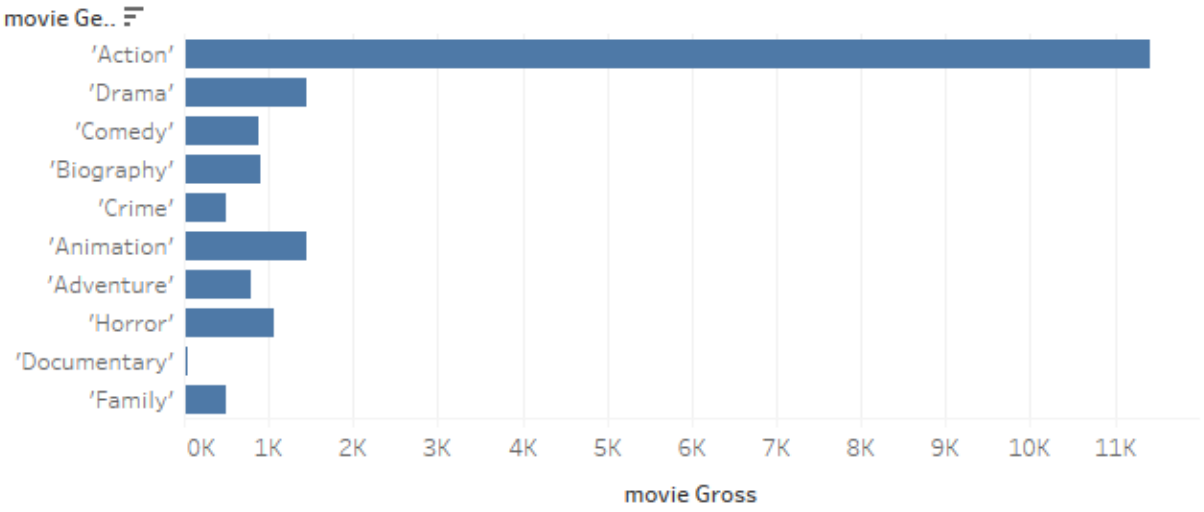
I made some of the insights through the data I collected. Screenshot of Tableau data visualizations are attached to support the visualizations.

1: Action, Drama and Comedy are the top 3 genres which came in the most in top 50 list for the last years. While Family coming in the last.

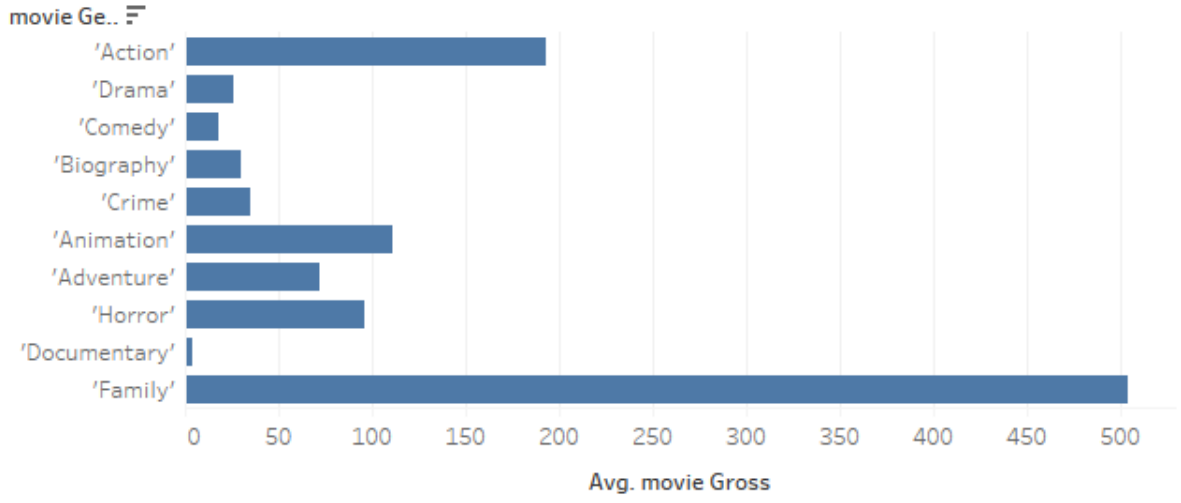


Sheet 1

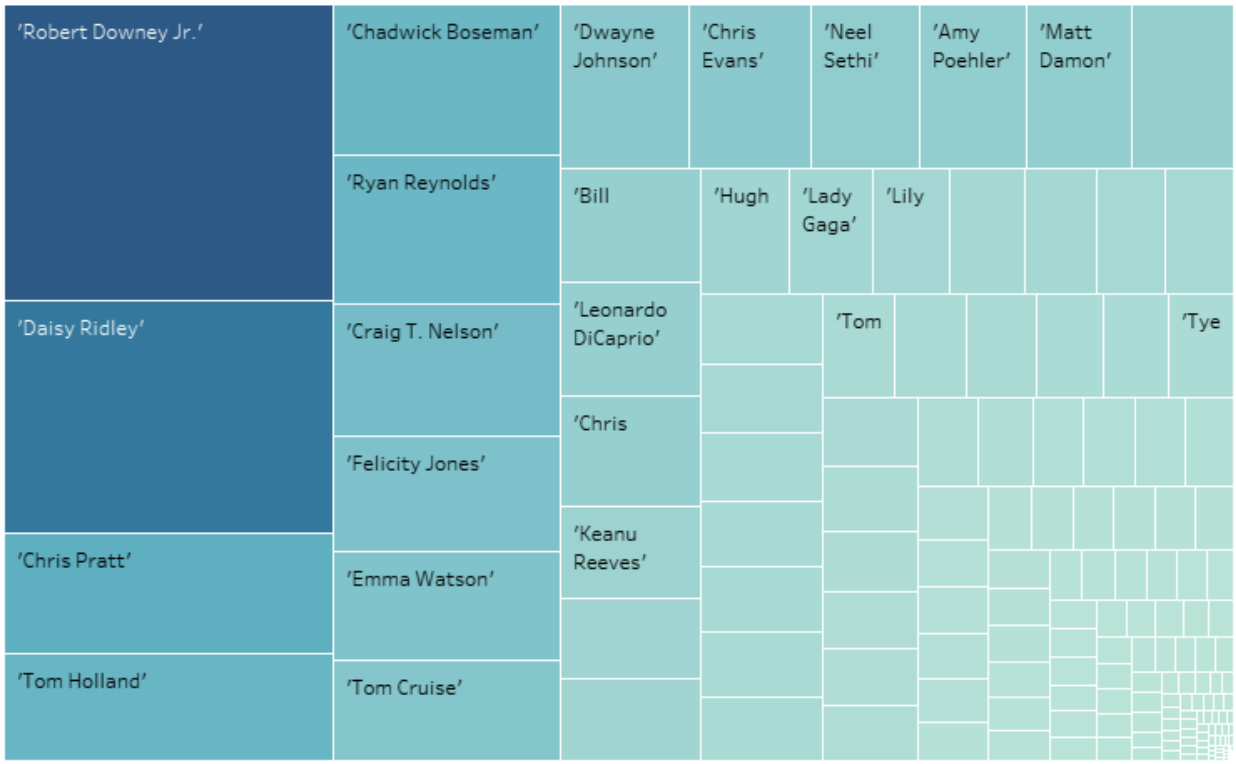
2: There was certainly a reason Action movie were preferred than the other genres and that was because they grossed the most too.

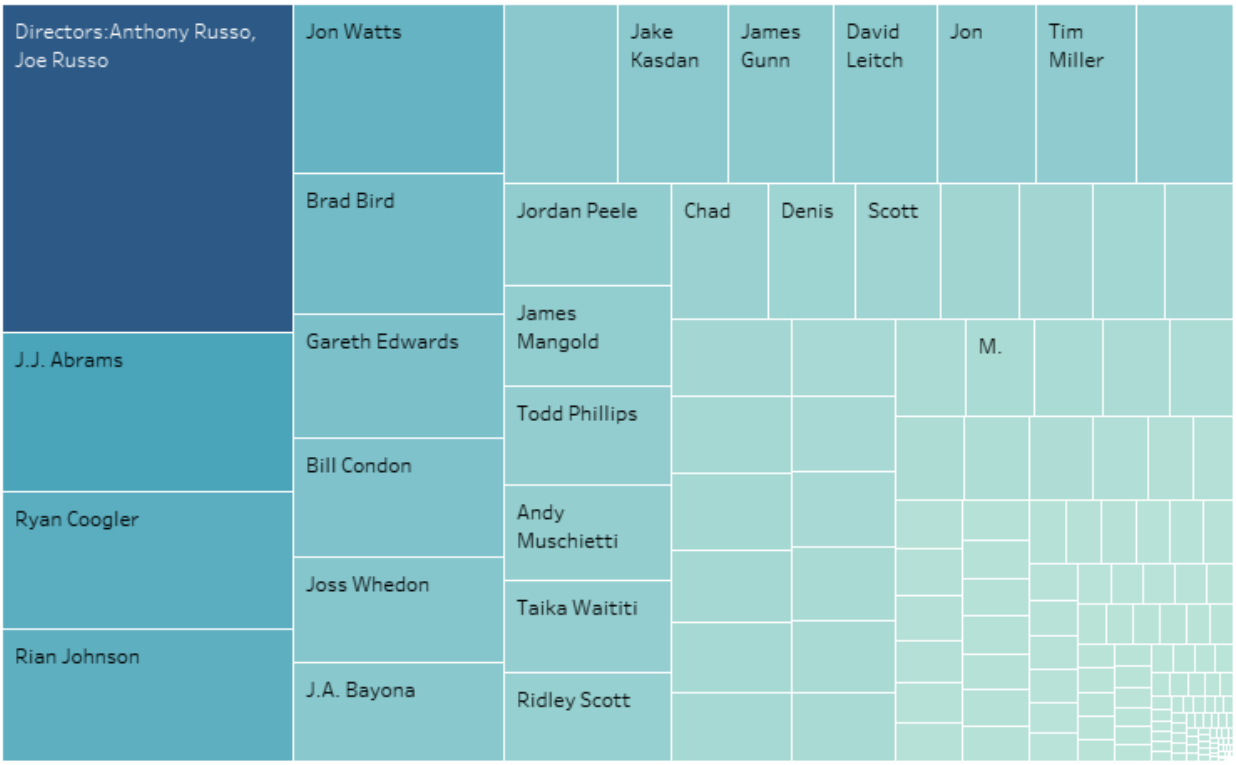


3: But there is something interesting, though action movies grossed the most, and family movies were roughly the 2nd least grossing movies in the last year, they were the safest gross as the avg gross was the highest of family movies.



4: Robert Downey junior was the highest grossing actor, while the Russo brothers were the highest grossing directors.





5: Animated movies got the best avg rating among all the IMDB voters, though the total rating were more with action movies but that was obviously because more number of action movie released in these 5 years.