STOR 455 Group Project (Due 5pm on November 24th)

Marvel Revenue Avengers

Liujie Zheng (730427441), Yuhan Zhou (730621041), Zhetao Zhang (730491003), Eliana Li (730521478)

The Prediction (Required)

Our prediction of the cumulative domestic box office of “The Marvels” by December 8, 2023 is …

Summary of Justification (Required)

Data (Required)

For this prediction task, we collected the **daily box office revenues** and the **Rotten Tomatoes scores** of all 33 Marvel movies. The daily box office revenues are extracted from Box Office Mojo into the **data_raw** folder (scrapper code in **scripts/data_scrapper.py**); the Rotten Tomatoes scores are collected from Wikipedia into the **metadata.csv**.

Metadata.csv

```
metadata = read.csv("metadata.csv")
dim(metadata)
```

```
## [1] 33  4
```

metadata.csv contains 33 rows and 4 columns.

```
head(metadata)
```

```
##                                  Movie Release.Date          Rotten.Tomato
## 1                             Iron Man     2-May-08 94% (281 reviews)[288]
## 2                  The Incredible Hulk    13-Jun-08 67% (239 reviews)[291]
## 3                           Iron Man 2     7-May-10 72% (304 reviews)[294]
## 4                                 Thor     6-May-11 77% (296 reviews)[297]
## 5   Captain America: The First Avenger    22-Jul-11 80% (275 reviews)[300]
## 6                Marvel’s The Avengers     4-May-12 91% (368 reviews)[303]
##                                                                     Url
## 1                     https://www.boxofficemojo.com/release/rl1482327553/
## 2 https://www.boxofficemojo.com/release/rl2791015937/?ref_=bo_tt_gr_1
## 3 https://www.boxofficemojo.com/release/rl1515881985/?ref_=bo_tt_gr_1
## 4                 https://www.boxofficemojo.com/release/rl3094644225/
## 5                 https://www.boxofficemojo.com/release/rl1900578305/
## 6                  https://www.boxofficemojo.com/release/rl709199361/
```

Each row of **metadata.csv** represents a Marvel movie. Explanation of each column:

Movie: Name of the Marvel movie

Release.Date: Release date of the movie in day-month-year

Rotten.Tomato: Rotten Tomatoes score

Url: Link to the domestic daily box office revenue

Raw Data

Each "Movie Name.csv" file in **data_raw** folder gives the **domestic daily box office revenue** of that movie. Take the movie "Ant-Man and the Wasp" as an example:

```
ant_man = read.csv("data_raw/Ant-Man and the Wasp.csv")
dim(ant_man)
```

```
## [1] 119  11
```

data_raw/Ant-Man and the Wasp.csv contains 119 rows and 11 columns

```
head(ant_man)
```

```
##                        Date       DOW Rank       Daily X...YD X...LW Theaters
## 1  Jul 6World Cup (Russia)    Friday    1 $33,725,082      -      -    4,206
## 2  Jul 7World Cup (Russia)  Saturday    1 $23,555,372 -30.2%      -    4,206
## 3  Jul 8World Cup (Russia)    Sunday    1 $18,531,751 -21.3%      -    4,206
## 4  Jul 9World Cup (Russia)    Monday    1  $6,983,824 -62.3%      -    4,206
## 5 Jul 10World Cup (Russia)   Tuesday    1 $10,042,976 +43.8%      -    4,206
## 6 Jul 11World Cup (Russia) Wednesday    1  $5,852,591 -41.7% -82.6%    4,206
##      Avg     To.Date Day Estimated
## 1 $8,018 $33,725,082   1     false
## 2 $5,600 $57,280,454   2     false
## 3 $4,406 $75,812,205   3     false
## 4 $1,660 $82,796,029   4     false
## 5 $2,387 $92,839,005   5     false
## 6 $1,391 $98,691,596   6     false
```

Each row represents a consecutive day. Explanation of each column:

Date: Screening day

DOW: Day of the week

Rank: The movie's rank in terms of box office revenue compared to other movies showing on the same day

Daily: The total box office revenue generated by the movie on that specific day

X...YD: Percent Change from Yesterday

X...LW: Percent Change from Last Week

Theaters:The number of theaters in which the movie was shown on that day

Avg: Average Revenue per Theater

To.Date: The total cumulative box office revenue for the movie up to and including that day

Day: The number of days since the movie was first released

Estimated: Whether the box office numbers are actual (false) or estimated (true) figures. 'True' suggests that the figures are preliminary and may be adjusted later, while 'False' indicates that the figures are final

For "The Marvels" movie, we have the daily box office revenue of the **first 10 days**.

```
the_marvels = read.csv("data_raw/The Marvels.csv")
dim(the_marvels)
```

```
## [1] 10 11
```

```
head(the_marvels)
```

```
##     Date       DOW Rank       Daily X...YD X...LW Theaters    Avg     To.Date
## 1 Nov 10    Friday    1 $21,603,104      -      -    4,030 $5,360 $21,603,104
## 2 Nov 11  Saturday    1 $15,260,052 -29.4%      -    4,030 $3,786 $36,863,156
## 3 Nov 12    Sunday    1  $9,247,703 -39.4%      -    4,030 $2,294 $46,110,859
## 4 Nov 13    Monday    1  $2,372,375 -74.3%      -    4,030   $588 $48,483,234
## 5 Nov 14   Tuesday    1  $3,300,946 +39.1%      -    4,030   $819 $51,784,180
## 6 Nov 15 Wednesday    1  $1,789,239 -45.8%      -    4,030   $443 $53,573,419
##   Day Estimated
## 1   1     false
## 2   2     false
## 3   3     false
## 4   4     false
## 5   5     false
## 6   6     false
```

Data Cleaning

To get the data ready for analysis, we:

1) separated **Date** and **Event**

2) turned **DOW** into 1 to 7

3) turned `-` mark into `NA`

4) turned columns like **Daily**, **X...YD** and others into numeric

The data cleaning code is in **scripts/data_cleaner.py**. Here is cleaned data of "Ant-Man and the Wasp" :

```
ant_man_cleaned = read.csv("data_cleaned/Ant-Man and the Wasp_cleaned.csv")
head(ant_man_cleaned)
```

```
##       Date              Event DOW Rank    Daily   YD.   LW. Theaters  Avg
## 1  Jul 6 World Cup (Russia)   5    1 33725082    NA    NA     4206 8018
## 2  Jul 7 World Cup (Russia)   6    1 23555372 -30.2    NA     4206 5600
## 3  Jul 8 World Cup (Russia)   7    1 18531751 -21.3    NA     4206 4406
## 4  Jul 9 World Cup (Russia)   1    1  6983824 -62.3    NA     4206 1660
## 5 Jul 10 World Cup (Russia)   2    1 10042976  43.8    NA     4206 2387
## 6 Jul 11 World Cup (Russia)   3    1  5852591 -41.7 -82.6     4206 1391
##    To.Date Day Estimated
## 1 33725082   1     False
## 2 57280454   2     False
## 3 75812205   3     False
## 4 82796029   4     False
## 5 92839005   5     False
## 6 98691596   6     False
```

And the cleaned data of “The Marvels”:

```r
the_marvels_cleaned = read.csv("data_cleaned/The Marvels_cleaned.csv")
head(the_marvels_cleaned)
```

```
##       Date Event DOW Rank    Daily   YD. LW. Theaters  Avg  To.Date Day Estimated
## 1 Nov 10    NA   5    1 21603104    NA  NA     4030 5360 21603104   1     False
## 2 Nov 11    NA   6    1 15260052 -29.4  NA     4030 3786 36863156   2     False
## 3 Nov 12    NA   7    1  9247703 -39.4  NA     4030 2294 46110859   3     False
## 4 Nov 13    NA   1    1  2372375 -74.3  NA     4030  588 48483234   4     False
## 5 Nov 14    NA   2    1  3300946  39.1  NA     4030  819 51784180   5     False
## 6 Nov 15    NA   3    1  1789239 -45.8  NA     4030  443 53573419   6     False
```

Data Visualization

Eliana Li’s part

Analysis (Required)

Simple Linear Regression

Haoyang Li’s part