**CS 181-01 Final Project Proposal**

*Ichhit Joshi*

**Observing and exploring the trends in box office (movie theater) revenues and attributes of yearly top 100 highest grossing movies from 1995 to 2022 (present)**

**Central Questions:** How have the box office revenues of movies changed over the years from 1995 to 2022? What is the most profitable genre of movie to produce? What are the top 10 most expensive movies ever made from 1995 to 2022? What are the top 10 all-time highest grossing movies and what were their profits? Is there any correlation between production budget and profit? Which theater distributor has made the most profit over the years?

These are the questions I am looking forward to answering through this project.

**Database Sources:**

<https://www.the-numbers.com>

**© 1997-2022 Nash Information Services, LLC.**

**Example Link:**

Getting data of top 100 highest grossing movies of 2022:

https://www.the-numbers.com/movies/report/All/All/All/All/All/All/All/All/All/None/None/2022/2022/None/None/None/None/None/None?show-release-date=On&view-order-by=domestic-box-office&show-release-year=On&view-order-direction=desc&show-production-budget=On&show-domestic-box-office=On&show-inflation-adjusted-domestic-box-office=On&show-international-box-office=On&show-worldwide-box-office=On&show-theatrical-distributor=On&show-genre=On&show-source=On&show-production-method=On&show-creative-type=On

To get data of a specific year, I will use the above link but use string formatting to change the value of year in the URL in the section: “[All/None/None/2022/2022/None/None/](https://www.the-numbers.com/movies/report/All/All/All/All/All/All/All/All/All/None/None/2022/2022/None/None/None/None/None/None?show-release-date=On&view-order-by=domestic-box-office&view-order-direction=desc&show-production-budget=On&show-domestic-box-office=On&show-inflation-adjusted-domestic-box-office=On&show-international-box-office=On&show-worldwide-box-office=On&show-theatrical-distributor=On&show-genre=On&show-source=On&show-production-method=On&show-creative-type=On)” . So, I just have to pass any year starting from 1995 to 2022. For example, I will pass “1995/1995” to get the correct URL for the 1995 data page.

**Database Description:**

The Numbers is a film industry data website that tracks box office revenue in a systematic, algorithmic way. The company also conducts research services and forecasts incomes of film projects. Their website has tabular movie report data for all the top 100 highest grossing movies of every year starting from 1995 to 2022 (present), all free for data analysis use as per their terms of service.

This project is only about the box office performance of movies in theaters and doesn’t include any streaming / home-DVD revenues.

I will be fetching yearly data from the years 1995 to 2022, which will equal **27 tables of 100 rows each (data sets)**, which can be combined into one huge dataset with **2700 rows.**

Every yearly table from the source website will contain the following **columns**:

**Rank:**

Data Type: int

Rank of the movie ranging from 1 to 100 based on highest revenue to lowest revenue

**Release Year:**

Data Type: int

Ranges from 1995 to 2022

It is the year the movie was released in the domestic market.

**Release Date:**

Data Type: string

Ranges from “January 1, 1995” to “December 31, 2022”

It is the date when the movie was released in the domestic market.

**Movie Title:**

Data Type: string

Official name of the movie released in the theaters

**Theatrical Distributor:**

Data Type: string

Company responsible for the marketing of a film. The distribution company may be the same with, or different from, the production company. For example, Sony Pictures is a Theatrical Distributor.

**Genre:**

Data Type: string (categorical)

Stylistic categories where a particular movie can be placed based on the setting, characters, plot, mood, tone, and theme.

*Can be one of the following:*

Adventure

Action

Drama

Comedy

Thriller/Suspense

Horror

Romantic Comedy

Musical

Documentary

Black Comedy

Western

Concert/Performance

Multiple Genres

Reality

Educational

**Source:**

Data Type: string (categorical)

Source from which the plot of the movie is based on.

*Can be one of the following:*

Original Screenplay

Based on Fiction Book/Short Story

Based on Comic/Graphic Novel

Remake

Based on Real Life Events

Based on TV

Based on Factual Book/Article

Spin-Off

Based on Folk Tale/Legend/Fairytale

Based on Game

Based on Play

Based on Theme Park Ride

Based on Toy

Based on Religious Text

Based on Short Film

Based on Musical or Opera

Based on Movie

Compilation

Based on Song

Based on Musical Group

Based on Web Series

Based on Ballet

Based on Radio

**Production Method:**

Data Type: string (categorical)

The method by which the movie is made. For example, if using humans as actors, then the production method is live action.

*Can be one of the following:*

Live Action

Animation/Live Action

Digital Animation

Hand Animation

Stop-Motion Animation

Multiple Production Methods

Rotoscoping

**Creative Type:**

Data Type: string (categorical)

Categories where a movie can be placed. (similar to genre but descriptive in a different way)

*Can be one of the following:*

Contemporary Fiction

Kids Fiction

Science Fiction

Super Hero

Fantasy

Historical Fiction

Dramatization

Factual

Multiple Creative Types

**Production Budget:**

Data Type: int

The amount of money it cost to make the movie including pre-production, film and post-production, but excluding distribution costs.

**Domestic Box Office:**

Data Type: int

Total money spent on tickets by moviegoers in the “domestic market”, which is defined as the United States, Canada, Puerto Rico and Guam.

**Inflation Adjusted Domestic Box Office:**

Data Type: int

The domestic box office revenue but adjusted to the current inflation rate. So, the revenue a movie released in 1995 would’ve actually earned if it was released today.

**International Box Office:**

Data Type: int

Total money spent on tickets by moviegoers in the “international market”, which is defined as everywhere outside the United States, Canada, Puerto Rico and Guam.

**Worldwide Box Office:**

Data Type: int

Sum of Domestic Box Office and International Box Office

Here is what my project process will look like:

1. Use requests to download the 27 html files for all the tables from 1995 to 2022. This will require using a for loop 27 times and changing the year values for the URL in every iteration.
2. Parse the HTML source code using etree and xpath to get all the table values.
3. Extract all the required column values and store them in lists.
4. Make a DoL and create a DataFrame for one table using pandas.
5. Repeat steps 3-4 for all the 27 tables.
6. Combine the 27 tables as 1 dataset using concatenation.
7. Add a profit column to the dataset and populate using calculations.
8. Exporting the dataset and storing it in a database.
9. Explore the dataset to answer the central questions.
10. Make graphs, charts, and plots using matplotlib library to further provide evidence for answering the questions.
11. Summarize the results.

Once I have the main dataset, to answer my central questions, I can use groupby for the categorical columns like Genre, Theatrical Distributor, Source, Production Method, Creative Type and use aggregate to see the total revenue/profit for each genre/distributor. I can also observe how that has changed over the years from 1995 to 2022. Or I can observe all-time statistics for them.

To get the top 10 most expensive movies ever made, I can sort the whole dataset on the Budget column (descending order) and get just the head of 10 rows.

**Functions:**

I can include the following functions in the code:

urlRequest(string URL) -> returns the root of the HTML

createDF(dictionary DoL) -> returns a pandas dataframe from the DoL

createDataBase(dataframe) -> creates and stores the data frame into a sqlite database

I can also make utility functions like below to get all the respective columns from the HTML root.

**def getMovieTitle(header):**

try:

return header[0].find("a").getText()

except:

return 'NA'

**Fields/Columns I will be calculating and adding:**

**Profit:** difference between production budget and worldwide box office revenue

computeProfit() -> int

finds the profit for a given movie

**At the end, the dataset is gonna have the following entities for each table:**

|  |
| --- |
| **Movies Database** |
| Rank  Release Year  Release Date - primary key  Title - primary key  Theatrical Distributor - primary key  Genre  Source  Production Method  Creative Type  Production Budget  Domestic Box Office  Inflation Adjusted Domestic Box Office  International Box Office  Worldwide Box Office  Profit |

I can use Release Date, Title, and Theatrical Distributor as primary keys to uniquely identify every movie. This should work because a theatrical distributor wouldn’t release multiple movies with the same title on the same release date. So, they are the independent variables of the dataset.