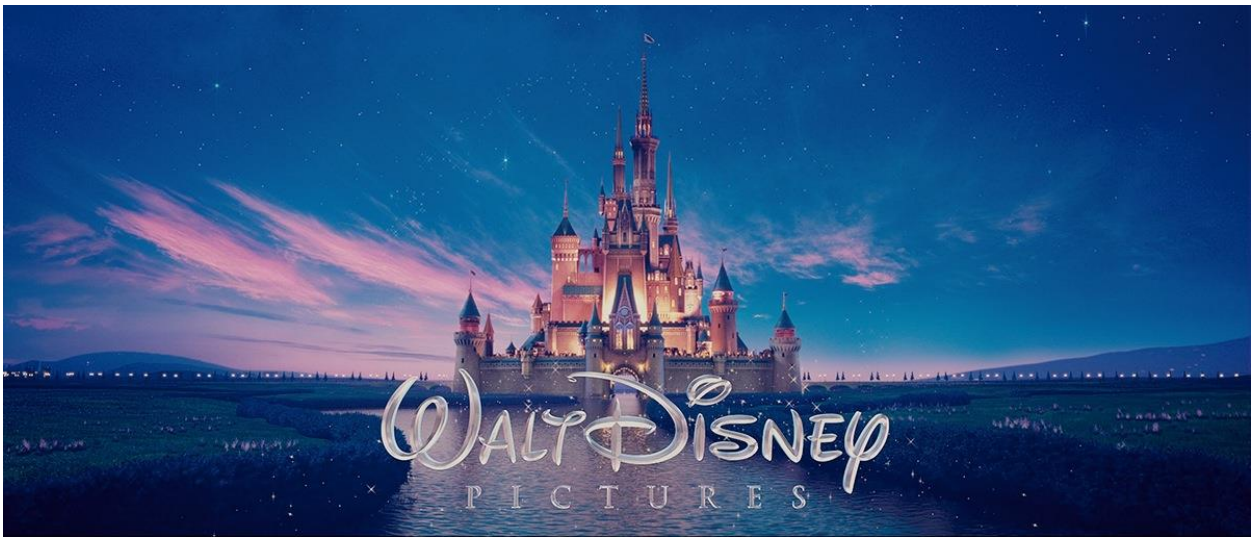
Disney Movies & Box Office Success Project

Applied Data Analysis with Python Libraries

Self-Guided Coding Project



General Instructions

This optional project is designed to complement the session on pandas and data manipulation. It requires the use of Python, specifically focusing on libraries such as **`pandas`**, **`numpy`**, **`matplotlib`**, and **`seaborn`**, to analyze the provided Disney dataset.

Dataset: There's only one dataset to be used, although you could explore external sources for further context.
Portfolio: The assignment can count as a project that you can showcase in your portfolio and display on GitHub.
Due date: There is no due date for this project.

The solution has been made available in the same project folder via **GitHub.**

For any questions regarding the project, please contact our teaching staff through the MMA office using the subject line [*Disney Movies – Workshop Project*] at mma@mcgill.ca.

Disney Movies & Box Office Success Project

Applied Data Analysis & Modeling with Python Libraries

Introduction: Since the 1930s, Walt Disney Studios has produced over 600 films, captivating audiences worldwide. With a rich history of cinematic achievements, the studio's box office success varies greatly across different films. This project aims to explore and analyze Disney movie data to understand the factors contributing to box office success. By building a linear regression model, we will predict the financial performance of Disney movies, providing insights into what drives a movie's success.

Disney Movies Dataset: The dataset provides an extensive look at Disney movies released between **1937** and **2016**. The dataset can be accessed on data.world, where it is publicly available for exploration and analysis.

This dataset includes detailed information on 579 Disney movies, covering various attributes that are key to analyzing the success of each film. The dataset's main features include:

- `Movie Title:` The name of the Disney movie.
- `Release Date:` The specific date when the movie was released.
- `Genre:` The genre classification of the movie, such as Animation, Action, Adventure, etc.
- `MPAA Rating:` The movie's rating according to the Motion Picture Association of America, indicating its suitability for different audiences (e.g., G, PG, PG-13).
- `Total Gross:` The total revenue generated by the movie at the box office.
- `Inflation-Adjusted Gross:` The movie's total gross revenue adjusted for inflation, allowing for a fair comparison of earnings across different time periods.

This dataset provides a comprehensive basis for analyzing trends in Disney movie success, examining how various factors such as genre, release year, and rating influence financial performance.

Disney Movies Dataset Overview

	A	B	C	D	E	F
1	movie_title	release_date	genre	mpaa_rating	total_gross	inflation_adjusted_gross
2	Snow White and the Seven Dwarfs	1937-12-21	Musical	G	184925485	5228953251
3	Pinocchio	1940-02-09	Adventure	G	84300000	2188229052
4	Fantasia	1940-11-13	Musical	G	83320000	2187090808
5	Song of the South	1946-11-12	Adventure	G	65000000	1078510579
6	Cinderella	1950-02-15	Drama	G	85000000	920608730
7	20,000 Leagues Under the Sea	1954-12-23	Adventure		28200000	528279994
8	Lady and the Tramp	1955-06-22	Drama	G	93600000	1236035515
9	Sleeping Beauty	1959-01-29	Drama		9464608	21505832
10	101 Dalmatians	1961-01-25	Comedy	G	153000000	1362870985
11	The Absent Minded Professor	1961-03-16	Comedy		25381407	310094574
12	Babes in Toyland	1961-12-14	Musical	G	10218316	124841160
13	Bon Voyage!	1962-05-17	Comedy	Not Rated	9230769	109581646
14	The Sword in the Stone	1963-12-25	Adventure		22182353	153870834
15	The Jungle Book	1967-10-18	Musical	Not Rated	141843000	789612346
16	Blackbeard's Ghost	1968-02-08	Comedy		21540050	138612686
17	The Aristocats	1970-04-24	Musical	G	55675257	255161499
18	The Boatniks	1970-07-01	Comedy		18607492	101200742
19	Bedknobs and Broomsticks	1971-10-07	Musical		17871174	91305448
20	The Apple Dumpling Gang	1975-07-01	Comedy		31916500	131246872
21	Freaky Friday	1977-01-21	Comedy		25942000	98067733
22	The Many Adventures of Winnie the Pooh	1977-03-11			0	0
23	The Rescuers	1977-06-22	Adventure		48775599	159743914
24	Herbie Goes to Monte Carlo	1977-06-24			28000000	105847527
25	The Black Hole	1979-12-21			35841901	120377374
26	Midnight Madness	1980-02-08			2900000	9088096
27	The Last Flight of Noahâ€™s Ark	1980-06-25			11000000	34472116
28	The Devil and Max Devlin	1981-01-01			16000000	48517980

disney_movies_total_gross

This dataset is in .csv format, which makes it easy to access and work with for this project.

Disney Movies & Box Office Success Project

Applied Data Analysis & Modeling with Python Libraries

Analytical tasks: In this project, you will undertake a series of analytical tasks designed to explore and understand the factors contributing to the success of Disney movies. The tasks will guide you through data exploration, visualization, hypothesis testing, and predictive modeling. By completing these tasks, you'll gain insights into trends in Disney movie performance and develop skills in data manipulation, statistical analysis, and machine learning using Python. Each task builds on the previous one which provides a comprehensive approach to analyzing the dataset and drawing meaningful conclusions about what drives box office success.

Part 1: Descriptive analytics

In this section, you'll begin by exploring the Disney dataset to identify the top-grossing movies. The goal is to determine which Disney films have earned the most at the box office when adjusted for inflation. This task involves sorting and analyzing the dataset to reveal insights into the movies that have been most successful over the years.

Steps to solution:

- Load the Disney movies dataset into your Python environment.
- Sort the movies by their inflation-adjusted gross revenue in descending order.
- Display the top 10 movies to identify the biggest financial successes.

This analysis will allow you to observe trends among the top performers, such as the most popular genres or common release periods for successful movies.

- **Visualization choice**: Use a bar graph or similar visual to represent the top 10 movies. This will help you clearly see the differences in revenue among the leading films and make it easier to identify standout performers.
- **Business insight**: Understanding the financial success of these movies can inform Disney's future production and marketing strategies by highlighting the characteristics of high-grossing films, such as genre preferences or optimal release timings.

Part 2: Genre popularity

The second part of the analysis focuses on examining trends in Disney movie genres over time. This task will help you understand which genres have gained or lost popularity, providing valuable insights into audience preferences and shifts in the film industry.

Steps to solution:

- Extract the release year from each movie's release date to facilitate year-by-year analysis.
- Group the data by genre and release year to explore how each genre's performance has changed over time.
- Compute the average inflation-adjusted gross revenue for each genre by year to identify patterns in genre popularity.

This step-by-step approach will reveal which genres have been on the rise and which have seen a decline in recent years.

- **Visualization choice**: A line plot is recommended for visualizing these trends, as it effectively displays changes in genre performance over time, making it easy to spot increasing or decreasing popularity.

- **Business insight**: By analyzing these trends, Disney can make data-driven decisions about which genres to invest in for future projects, aligning their content with evolving audience tastes and maximizing potential revenue.

Part 3: Predictive modeling

Disclaimer: This section is intended for more advanced students who are comfortable with machine learning concepts and linear regression modeling.

In this part, you will build a predictive model to forecast the box office success of Disney movies based on features such as genre and release year. This task aims to provide a deeper understanding of how various factors contribute to a movie's financial performance and allows you to apply data science techniques to real-world scenarios.

Steps to solution:

- Perform one-hot encoding on categorical variables, such as genre, to convert them into a numerical format suitable for modeling.
- Split the dataset into training and testing sets to evaluate the model's performance accurately.
- Build and evaluate the linear regression model, using the training set to fit the model and the testing set to assess its predictive accuracy.

This approach enables you to leverage historical data to make informed predictions about future movie revenues.

- **Model explanation**: Linear regression is chosen for this task because it is a straightforward method to model the relationship between independent variables (like genre and release year) and the dependent variable (box office gross). It is particularly useful in understanding the weight and impact of each feature on movie success.

- **Business insight**: This model can be used by Disney to forecast future movie revenues, helping decision-makers identify which types of movies are likely to perform well financially. By understanding these predictive patterns, Disney can strategize their production choices, allocate resources more effectively, and optimize marketing efforts.

***

Remember, you're not required to complete all tasks – this project is all about learning, exploring, and enjoying the data! Try tackling the assignment without relying on GenAI applications; it's a great way to deepen your understanding of the material and boost your coding confidence. There's no single right way to approach this, so be original, get creative, and dive into the data as much as you can. Programming can be incredibly fun, especially when you're working with the enchanting world of Disney! Take your time, enjoy the process, and don't stress about perfection: just have fun uncovering the patterns behind these movies. You might even stumble upon some surprising insights along the way! Keep an open mind, stay curious, and embrace the journey – you've got this!

Appendix

For those new to programming, we've got you covered! Below, you'll find a selection of helpful navigation resources to guide you through the notebook and support your learning journey.

Overview of notebook structure & navigation

The notebook for the Disney Movies & Box Office Success project is organized into several key sections, each focusing on specific tasks that guide you through the analysis and modeling process. Below is an overview of how the notebook is structured and how to navigate it effectively:

Introduction/Dataset Overview

This section introduces the project, providing context on Disney's extensive history in film production and the purpose of the analysis. It includes a brief description of the dataset used, detailing its key features such as movie title, release date, genre, MPAA rating, total gross, and inflation-adjusted gross.

Data Loading/Initial exploration

This section covers the steps to load the dataset into the notebook using pandas. Basic data exploration functions like head(), describe(), and info() are used to gain initial insights into the data structure, identify missing values, and understand the distribution of key variables.

Data cleaning & preparation

Here, you will find code dedicated to cleaning the dataset, including handling missing values, correcting data types, and removing any inconsistencies. The section also includes preparation steps like extracting the release year from the release date and converting categorical variables into numerical formats through techniques like one-hot encoding.

Descriptive analytics

This part of the notebook dives into exploratory data analysis (EDA), where visualizations such as bar charts and line plots are used to explore trends in the data. It includes tasks like identifying the top-grossing movies, analyzing genre popularity trends over time, and examining how movie attributes correlate with box office success.

Hypothesis testing

In this section, statistical inference techniques are employed to test hypotheses related to movie success. Two-sample bootstrap hypothesis tests are used to compare means and assess whether differences in revenue between groups (e.g., genres, ratings) are statistically significant.

Predictive modeling

This section covers the construction of a linear regression model to predict box office success based on movie attributes. It includes steps for splitting the data into training and testing sets, fitting the model, and evaluating its performance using metrics like R-squared and mean squared error.

***

How to navigate the notebook

- Each section is clearly titled, allowing you to scroll through the notebook and locate specific tasks easily.
- Code cells are interspersed with markdown cells that provide explanations, making the notebook easy to follow even for beginners.
- Use the table of contents or section headers to jump directly to areas of interest, especially when reviewing or revising specific analyses.
- To modify or add to the analysis, locate the relevant section, make your changes in the code cells, and re-run the cells to see updated outputs.