COURSE WORK I: PROGRAMMING FOR

DATA SCIENCE

MSc Data Science: Coventry University UK (**2024.2 BATCH**)

Name: P D Lolitha Lakshan Weerasinghe

Index Number: COMScDS242P-003

**QUESTION 1**

## a)

* + Implemented "Student", "Teacher," and "Staff" classes derived from the parent "Person" class.
  + Defined "name," "age," and "address" in the "Person" class.
  + Created an "address" class which encapsulates attributes such as street, city, state and zip code. Added “Address” to the "Person" class which act as a compound attribute.
  + Also, initializers are created for the classes
  + Class and methods comments added.

## b)

* + Defined “assign\_grades” method in the "Student" class where the method accepts “subject\_grade\_dict “as a parameter which holds the subjects and the marks.
  + The method first checks for an empty dictionary, if not empty then proceeds with the calculation.
  + The method outputs the average grade.
  + Method comments are included.
  + Unit tests are added to verify the functionality of the method for the scenarios such as validity of the calculation, handling empty dictionary and how method behaves for a single value.

## c)

* + Added the “ssn" attribute to the "Person" class.
  + Created a getter and setter to access the “ssn" attribute, however on “Person” class this would raise an “AttributeError”.
  + This would prevent from any subclasses setting / accessing the “ssn” property inherently, if any customized implementation is required for “Student” and “Teacher”; A customized implementation is provided to set and return the “ssn”.
  + “ssn” is exposed through the initializer as well, therefore only way to set the “ssn” is through the setter.
  + Also string validation is done in the setter to only accepts valid “ssn”.
  + Further improvements can be done to incorporate regex to validate “ssn”.

## d)

* + Implemented “role\_duties(self)” method in the “Person” class which returns all the duties supported as an array. So multiple duties are supported by each “Person”.
  + The duties are declared as an enum.
  + Each sub class overrides “role\_duties(self)” method and returns suitable duty for each subclass.

## e)

* + A “Subject” class is declared to hold subject related information such as identifier and name.
  + A “ScheduleClass” class declared to hold class information such as “Subject”, day of the week, start time and end time.
  + The day of the week is declared as an enum with days of the week.
  + A “Subjects” array is declared in the subclass “Teacher” to hold the list of subjects taught by a “Teacher”.
  + The “Subjects” can be assigned to a “Teacher” by using the “add\_subject()” method in the “Teacher” subclass.
  + A “class\_schedule” dictionary is declared “Teacher” subclass to hold class schedule information.
  + Day of the week will act as the key in the dictionary and values will be a list of the class schedules.
  + A class schedule can be added using by introduced “schedule\_classes” method which checks the day exists on the dictionary, if exists appends the schedule else adds a new list with the schedule.

## f)

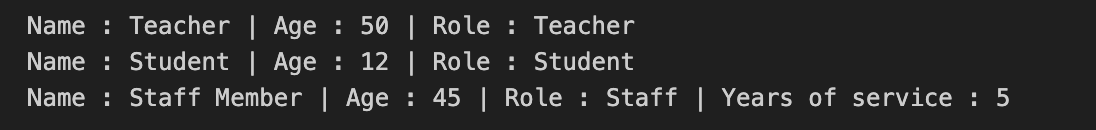
* + To keep track of student attendance a “AttendanceRecord” class is declared. This class holds information about “Subject”, attendance status and date.
  + A list named “attendance\_records” is declared in the “Student” class which holds the “AttendanceRecord” instances.
  + An “attendance(self, date: datetime, subject: Subject, attended: bool)” method creates a new “AttendanceRecord” and adds to the “attendance\_records” list.
  + display\_attendance(self) method iterates through “attendance\_records” list and calculates numbers of attended classes, number of missed classes and attended percentage.

## g)

* + Declared “StaffRole” enum which declares staff roles such as MANAGER, ADMINISTRATIVE and OTHER along with their base salary.
  + Modified Staff subclass with new attributes “role”, “years\_of\_service” and “salary”.
  + The “role” and “years\_of\_service” is initialized upon instance creation using the constructor and salary is initially initialized to the value of None.
  + The “calculate\_salary(self)” method will calculate the Staff instance salary using the “role” and “years\_of\_service”, then updates “salary” variable.
  + The “get\_salary()” will return the calculated salary , if the salary is not present then salary is calculated prior to returning the value.
  + Also, unit tests are added to verify calculations are correct for each role.

## h)

* + The “Person” super class is declared with two new methods get\_role (self) and display\_info(self).
  + The get\_role (self) method returns the role for the class, subclasses such as “Teacher”, “Student” and “Staff” overrides this method to return their own “role”.
  + The display\_info(self) method prints the name, age and role for the class and declared in the super class.
  + A custom implementation for the “The display\_info(self) method is provided in the ““Staff” subclass.
  + This custom implementation prints the number of years of service apart from the name, age and role which is specific to the “Staff” class implementation.
  + Polymorphism is demonstrated by creating instances of the “Student”, “Teacher” and “Staff” classes, added to a list and invoking the “display\_info” method in each instance which prints the class specific implantations.



**QUESTION 2**

## 1.Data Loading and Preprocessing:

The data is scraped using BeautfulSoup library and loaded into a CSV to for the subsequent reading. The top 250 movies are extracted from json data in the first <script> of the IMDB page. This returns the movie name, url, rating and genre. Then url is used to extract the data such as released year, box office earnings, actors and director list from details page and combine that data into the previously extracted dataset.

Then dataset is ermined for any null values and outliers.

It was found out there four movies were missing box office earnings, and one movie is missing the release year. It was decided to drop these movies to clean the datasets.

Also, year column is converted to integer values from a float value, also box office values were converted from string to a float.

## 2. Descriptive and Statistical Analysis

First data set was analyzed using the describe method for the numeric columns. Histograms were created to check distributions for ratings and movie release year.

And Box office field was checked for outliers.

A screen shot of a black and white screen

AI-generated content may be incorrect.

A graph of a bar graph

AI-generated content may be incorrect.

A diagram of a box

AI-generated content may be incorrect.

Following are the findings,

1. Rating

- Average rating 8.3 which tells us majority movies are well liked by the viewers.

- Standard deviation is 0.233, which is low.

- Indicates most movies are near the mean value.

- Minimum rating is 8.0 which means, even the lowest rated movie is well received.

- Maximum rating is 9.3 which means highest rated movie has almost a perfect rating.

- The middle 50% ratings are between 8.1 - 8.4.

2. Release Year

- Min is 1921 and max is 2024, which means data set contain movies throughout the century which is a large span.

- Standard deviation is 25 which is quite high. Higher variability of years.

- Q1 is 1972 and Q3 is 2009 which indicates majority of movies were released between this time span since dataset skewed to more recent movies according to the plots.

3. Box Office

- Mean is 230.64M which means movies earned that much on average.

- Standard deviation is 371.53M which is high, which indicates high variation.

- Min and Max confirm that because min is 399.00 and max is 2.8 billon which indicates two extreme end values.

## 2. Data Visualizations

Matlab plot was used uncover patterns related check there are any correlations between ratings vs box office earnings and box office earnings vs released year.

And following things were uncovered.

Movie Rating vs Box office

* + - Seems like movies rated between 8.1 - 8.4 have consistently given good box office records.
    - There's no indication that highest rated movies earned more than lower rated movies.
    - Also, there are few outliers related to earnings.

Box office vs Revenue

* + - Box office earnings have been steadily increasing over time.
    - After 1990 there has been a significant increase of box office earnings.
    - During 1920 - 1960 earnings have been stagnating.
    - Spikes may indicate blockbuster releases.

A comparison of a graph

AI-generated content may be incorrect.

## 4. Movie and Director Trends

4.1 Most recurrent directors and actors over the years were found using Panda’s library.

A screen shot of a computer

AI-generated content may be incorrect.A screen shot of a movie

AI-generated content may be incorrect.

## 4.2 & 4.3 Identify trends in the number of movies directed by prominent directors, and how often they appear in the top-grossing movies and analyze the influence of certain directors or actors on movie success.

From the 250 highest rated movies 50 most box office grossing were selected to identify the top directors and relation between directors and earnings. Directors who have more than two movies in the 50 list were given priority and earnings more than one billion dollars .

A graph of a number of movies

AI-generated content may be incorrect.

Number of Top 50 Grossing Movies by Director:

* As shown in the charts Christopher Nolan has directed the most movies in the top 50 highest-grossing films.
* While Steven Spielberg, Lee Unkrich, Peter Jackson, and Pete Docter also have multiple movies in the top-grossing category.
* Other directors like Anthony Russo, Andrew Stanton, Brad Bird, and Joe Russo appear at least twice. Christopher Nolan stands out more clearly as he has much as five movies in the top 50 grossing movies.

A graph of a chart

AI-generated content may be incorrect.

Total Box Office Earnings of Directors in Top 50 Grossing Movies

* When considering gross amount more than one billion earnings Anthony Russo and Joe Russo have the highest total box office earnings.
* Christopher Nolan, Lee Unkrich, Peter Jackson, and Steven Spielberg also have high cumulative earnings.
* Pete Docter, Jon Watts, and Joseph Kosinski follow with significant earnings but slightly lower than the top-tier directors.
* It seems directors with multiple blockbuster movies tend to have the highest cumulative box office earnings. Christopher Nolan is not only the most frequent director in the top 50 but also one of the highest-grossing ones. The Joe Russo dominates in earnings despite directing fewer movies, indicating his films are exceptionally high-grossing.

A graph with colored lines and dots

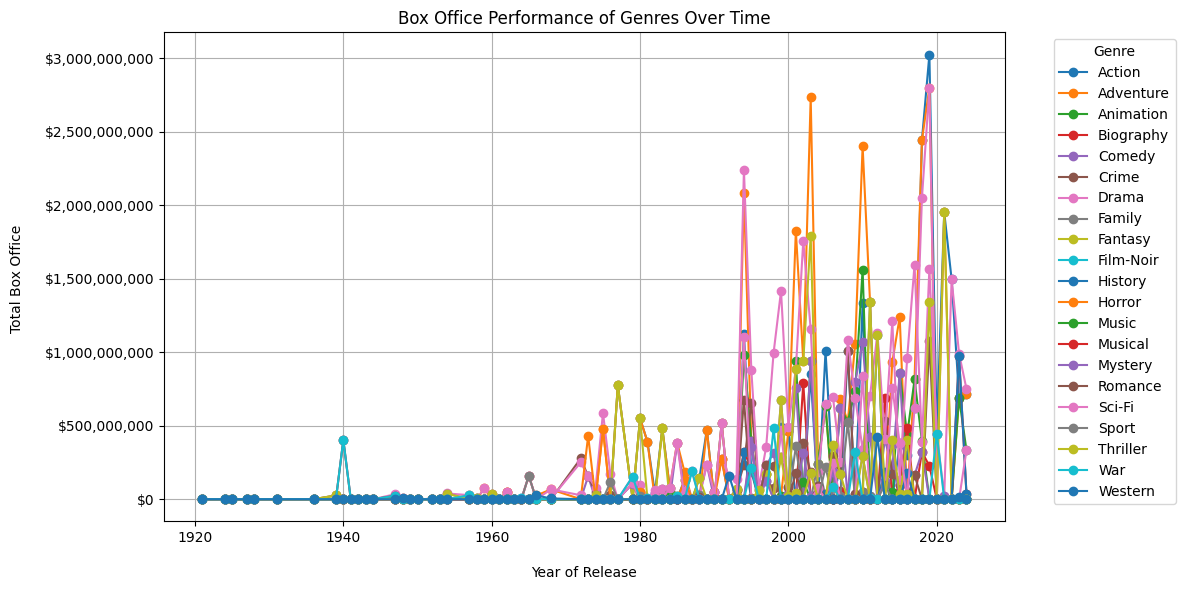
AI-generated content may be incorrect.

Director Influence and trends over the years

* Christopher Nolan stands out in both frequency and earnings, showcasing consistent success.
* The Russos (Anthony & Joe Russo) have the highest total box office earnings despite directing fewer films, suggesting their movies are blockbusters. Steven Spielberg & Peter Jackson maintain strong positions, assuring their legendary status in the industry.
* Joe Russo’s is shown on dramatic rise in earnings exceeding $2.5 billion in the 2010 era at a very short period of time.
* Chirstian Nolan shows consistent earnings over time; the earnings are consistent and does not show spikes like Russos.
* Spielberg seems to be prominent in 70's, 80's and 90's, his earnings seem to be high compared to the era.
* Early 2000s Peter Jackson is shown a sudden rise in a short period of time and then fades away.
* Unkrich, Pete Docter, and Brad Bird have steady earnings as well.
* Overall, 1990s & Early 2000s dominated by Spielberg, Peter Jackson, and animated directors dominate in 2010s. 2020s Christopher Nolan and Russo dominates the earnings.

## 4.4 Highlight the evolution of film styles (e.g., changes in genre preferences, special effects, or trends in specific types of movies like superhero movies).

Genres rise and fall was plotted throughout the years to find the trends of genres.



* 1920s - 1960s Had Low Box earnings, most popular Genres seems to be Western, War, and Musical. The few spikes can be attributes major hits.
* 1970s - 1990s Fantasy, Adventure, and Sci-Fi start growing and previously popular Musicals and Westerns declines, and earning have relatively increased.
* 2000s - 2010s shows Action, Sci-Fi, Adventure, Fantasy drastic rise in earnings. Drama and Crime start stagnating although still popular.
* Post-2010s: Shows huge spikes of earnings of Action and Sci-Fi movies. Western, Musical, and War genres have further stagnated.

## 5. Analyze how different movie genres have evolved over the years.

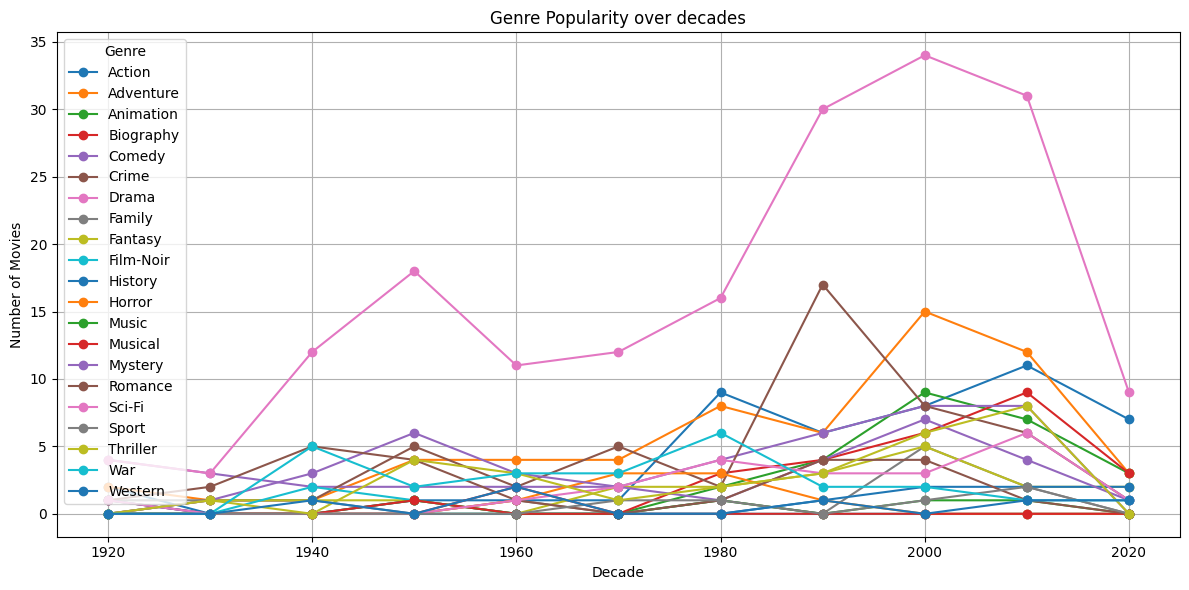
## 5.1 Investigate the most successful genres in terms of box office revenue.

A graph of a number of people

AI-generated content may be incorrect.

* Adventure, Drama and Action movies generate the highest total box office revenues. The revenues are significantly higher than other genres.
* Animation, Sci-Fi, Fantasy and Comedy, Crime and Thrillers have done well too.
* Film-Noir, Musical, and Sport generate the least revenue, however this could be to lower number of movies on the top 250 movies.

## 5.2 Identify changes in genre popularity over the decades (e.g., the rise of action or sci-fi in the 2000s).

****

* The Sci-Fi genre experienced significant growth over the years, peaked around the 2000s with over 30 movies released and then experienced a sharp decline in the 2020s.
* Adventure and Action genres have been steadily increasing since the 1970s and 1990s, peaking in the 2000s and continue to do well.
* Drama and Romance have been doing well over the decade with sporadic spikes and continue to do well.
* Film-Noir, Western genres seem to be on decline.
* 2000 seems to be peak period for multiple genres.

### 5.3 Examine whether audience genre preferences have shifted, particularly in relation to

### emerging movie technologies (e.g., CGI, IMAX, 3D).

A graph of different colored lines

AI-generated content may be incorrect.

* The rise in adventure movies can be assumed to CGI with the year 2000.
* The rise in action movies can be assumed to IMAX and 3D technologies around 2010.
* The rise in sci-fi can be attributed both IMAX and CGI.
* The Animation rise in early 2000 can be also attributed to 3D technologies.

A chart with many colored dots

AI-generated content may be incorrect.

Large blue and orange bubbles (Action & Sci-Fi) after 2010 indicate movies such as The Dark Knight Rises, Avengers, Spider-Man these have thrived with CGI advancements, IMAX. Proves above points.

Large purple bubbles, Toy Story 3, Frozen, Inside Out show how 3D animation movies remain popular.

## 6. Correlations and Discoveries

A red and blue squares

AI-generated content may be incorrect.

There seems to be a moderate positive correlation (0.41) between release year and box office earnings, indicating newer movies tend to earn more. There is also a weak positive correlation (0.17) between rating and box office, indicating higher-rated movies may perform slightly better financially.