DSE 2242– Fundamentals of Machine Learning Lab
Week 1 – Date: 6th January 2025

EXERCISE 1: Consider the "Titanic.csv" data set and note the following meta information:

PassengerId - Identifier for passenger
Survived - If passenger survived the ship, 1 if passenger survived, 0 otherwise
Pclass - Passenger Class (1 = 1st; 2 = 2nd; 3 = 3rd)
Name
Sex
Age
SibSp - Number of Siblings/Spouses Aboard
Parch - Number of Parents/Children Aboard
Ticket - Ticket number
Fare - Passenger Fare (in British pounds)
Cabin - Cabin number
Embarked - Port of Embarkation (C = Cherbourg; Q = Queenstown; S = Southampton)

Use Python and appropriate packages to answer the following questions:

1. Replace the index value of DataFrame with PassengerId
2. Identify the missing values of the columns Age, Embarked and Cabin columns
3. Get descriptive statistics on "object" and "number" datatypes separately.
4. Find how much % of the passengers were survived.
5. Find the % of male survival and female survival rate and check whether the gender has any relationship with the survival rates.
6. Find the age of passengers per passenger class, visualize using appropriate plot.
7. List out the passengers who had more than 2 family members on board.
8. Find and list all attributes that are related to survival using appropriate tests.
9. Determine the total number of male and female passengers in each class, categorized by whether they survived or did not survive.

EXERCISE 2: Descriptive Analytics and Visualization

The data file bollywood.csv contains box office collection and social media promotion information about movies released in 2013−2015 period. Following are the columns and their descriptions:

- SlNo
- Release Date
- MovieName – Name of the movie
- ReleaseTime – Mentions special time of release. LW (Long weekend), FS (Festive Season), HS (Holiday Season), N (Normal)
- Genre – Genre of the film such as Romance, Thriller, Action, Comedy, etc
- Budget – Movie creation budget
- BoxOfficeCollection – Box office collection

- YoutubeViews – Number of views of the YouTube trailers
- YoutubeLikes – Number of likes of the YouTube trailers
- YoutubeDislikes – Number of dislikes of the YouTube trailers

Use Python code to answer the following questions:

1. How many records are present in the dataset?
2. How many movies got released in each genre? Sort number of releases in each genre in descending order.
3. Which genre had highest number of releases?
4. How many movies in each genre got released in different release times like long weekends, festive season, etc. (Note: Do a cross tabulation between Genre and ReleaseTime.)
5. Which month of the year, maximum number movie releases are seen? (Note: Extract a new column called month from ReleaseDate column.)
6. Which month of the year typically sees most releases of high budgeted movies, that is, movies with budget of 25 crore or more?
7. Which are the top 10 movies with maximum return on investment (ROI)? Calculate return on investment (ROI) as (BoxOfficeCollection – Budget) / Budget.
8. Do the movies have higher ROI if they get released on festive seasons or long weekend? Calculate the average ROI for different release times.
9. Is there a correlation between box office collection and YouTube likes? Is the correlation positive or negative?
10. Which genre of movies typically sees more YouTube likes? Draw boxplots for each genre of movies to compare.
11. Which of the variables among Budget, BoxOfficeCollection, YoutubeView, YoutubeLikes, YoutubeDislikes are highly correlated? Note: Draw pair plot or heatmap.
12. During 2013−2015 period, highlight the genre of movies and their box office collection? Visualize with best fit graph.
13. Visualize the Budget and Box office collection based on Genre.
14. Find the distribution of movie budget for every Genre.
15. During 2013−2015, find the number of movies released in every year. Also, visualize with best fit graph.