National University of Computer and Emerging Sciences, Lahore Campus

Course Name:	**Data Analysis and Visualization**	**Course Code:**	**DL3001**
Program:	**BS Data Sciences**	**Semester:**	**Fall 2023**
Section	**BS DS 5A, 5B, 5C, 5D**	**Total Marks:**	**50**
Due Date:	**21st September 2023**	**Weightage:**	
Exam Type:	**Assignment 1**	**Page(s):**	**3**

Instructions:

- Submit the solution in a zipped file named as your roll number., i.e., 21L-1234.zip
- You are not allowed to copy solutions from other students (same/cross section). Your code will be checked for plagiarism. If any sort of cheating is found, heavy penalties will be given to all students involved.
- Late submission of your solution is not allowed.

Question 1: Data Scraping

Part 1: Selenium [15 Marks]

You are provided with the URL of a YouTube channel. Your task is to write Python code using Selenium to scrape to extract relevant information and gain insights from the collected data.

URL: https://www.youtube.com/@UnfoldDataScience

- Use Selenium to access the provided YouTube channel URL. Scrape the videos uploaded between Sep 10, 2019 and Sep 10, 2023.
- Extract the following information for each video on the channel's page:
    - Video Title
    - Views Count
    - Likes Count
    - Upload Date
    - Number of Comments
- Store the extracted data in a structured format.
- Create functions for the following tasks on the scraped data:
    - Calculate the average views count per video for videos uploaded in the last 30 days.
    - Identify the video with the highest likes-to-views ratio.

    - Find the correlation between the number of likes and the number of dislikes for the videos.
    - Determine the most common day of the week for video uploads.
    - Detect any outliers in the views count.

Part 2: Beautiful Soup [15 Marks]

You are provided with the URL of a website that lists the top-rated movies. Your task is to write Python code using Beautiful Soup to scrape to perform following tasks.

URL: https://www.imdb.com/

- Scrape the movies released between 2013 and 2023.
- Write a Python script using BeautifulSoup to scrape the following information for each movie:
    - Movie Title
    - Release Year
    - IMDb Rating
    - Director
    - Genre
- Store the scraped data in a structured format, such as a CSV file.
- Create functions for each of the tasks listed below:
  - Average IMDb rating for the top-rated movies.
  - The most common genre among the top-rated movies.
  - Identify the director with the highest average IMDb rating.
  - Determine the year with the highest number of top-rated movies.

Question 2: Data Wrangling [20 Marks]

You are provided with a dataset that contains the information about the housing prices. Use the given dataset to perform the following tasks:

- Load the dataset into a Pandas dataframe and display the first few rows.
- Discretize the "age" variable into three bins: 'Young', 'Middle-aged', and 'Old'.
- Create a binary variable "is_charles_river" based on the "chas" column.
- Detect and remove outliers for each numerical column in the dataset using the Interquartile Range (IQR) method. (**Don't use any in-built library for this part.)**
- Identify and remove noisy data points from the dataset.
- Apply smoothing to the "rm" column and create a new smoothed column.
- Normalize the "tax" and "lstat" columns using Min-Max normalization.
- Perform a simple linear regression to predict the median value of "medv" based on the "rm" variable.
- After the regression analysis, explain if you observe any relationship between "medv" and "rm," providing interpretations based on the regression results.

NOTE: You can find the description of the dataset from the link below for better understanding, https://www.cs.toronto.edu/~delve/data/boston/bostonDetail.html.

Happy Coding!