Introduction to Data Science (DS 2001)

Fall 2024

Total marks : **100**

Due date: 09th October, 2024, 11:59 PM

Assignment 2

Instructions:

- All questions must be answered within a single notebook.
- Follow the file naming conventions: Name your submission file as RollNo.ipynb
- (e.g., i23_xxxx.ipynb, where xxxx is your Roll Number).
- Do NOT submit any dataset, only your .ipynb file.
- Use headings to distinguish each part in the notebook.
- Late submissions will not be accepted and will be given a zero.
- Any form of plagiarism will result in a zero for both parties involved.
- 5 marks will be deducted for not following the naming convention.
- To earn bonus marks, please use comments and adhere to correct PEP 8 coding conventions.

Data set description :

This dataset provides detailed information about applications available on the Mac App Store. It includes attributes such as app ID, artist name, track name, description, release date, price, average user rating, and supported devices etc. Each application is characterized by its features, genres, and content advisory ratings. This dataset is useful for analyzing trends, understanding user preferences, and evaluating app performance in the Mac App Store ecosystem.

TASK : You are requested to perform an analysis using Exploratory Data Analysis (EDA) techniques to derive insights from this dataset.

1. Data Loading and Exploration [10 marks]:

- Load the dataset and explore to understand its structure, size, and characteristics. **(10 marks)**

2. Data Cleaning and Preprocessing [20 marks]:

- Perform data cleaning and preprocessing according to your chosen methods. Use your creativity and judgment to decide how to handle various issues in the dataset. **(10 marks)**
- Add New Features: Create new columns based on your analysis. Think about what additional information could be valuable for understanding the dataset better. **(5 marks)**
- Remove Outliers: Identify and remove outliers from the dataset using methods of your choice. **(5 marks)**

3. Data Analysis [30 marks]:

- **Exploratory Data Analysis (EDA): (10 marks)**

1. Descriptive Statistics: Calculate descriptive statistics like mean, median, mode, and standard deviation for numerical columns like price, user ratings, etc.
2. Category Analysis: Analyze the distribution of apps across different genres, price ranges, and content advisory ratings.

- **User Ratings Analysis: (10 marks)**

1. Correlation Analysis: Examine the correlation between user ratings and other variables such as price or genre.
2. Top/Bottom Performers: Identify the apps with the highest and lowest user ratings and analyze their features.

- **Price Analysis: (10 marks)**

1. Distribution of App Prices: Analyze the distribution of app prices, including the proportion of free vs paid apps.
2. Price vs Rating Relationship: Investigate how the price of an app relates to its user rating and whether paid apps tend to receive higher ratings.

4. Data Visualization [30 marks]:

- Visualize your findings using Matplotlib. You have freedom to create any 4 unique plots to illustrate your analysis, ensuring that each visualization has appropriate labels, titles, and legends. **(15 marks)**
- Explain the significance of each visualization in your report**. (15 marks)**

5. Report [10 marks]:

Write a brief report summarizing your analysis and findings.

Include:

1. Insights gained from the data cleaning and preprocessing. **(5 marks)**
2. Key observations and interpretations based on your analysis and visualizations. **(5 marks)**

Note : make sure to implement the usage of all three libraries i.e. NumPy, Pandas, & matplotlib.

Happy Coding :)