Mathematics For Computer Science

UE23MA242A

Teaching Assistants: Shruti C and Sragvi Anil Shetty

Problem Set: Streaming Platform Analysis

Introduction

In this data science exercise, we will explore a dataset that includes detailed information on popular TV shows across various streaming platforms. By analyzing this data, we can gain insights into genre popularity, rating trends, and platform availability, and we can examine viewer demographics. This dataset enables streaming services and media analysts to make data-driven decisions for content development, viewer engagement strategies, and platform performance analysis.

Background

The dataset provided includes features such as title, year, age rating, IMDb ratings, Rotten Tomatoes scores, and availability across streaming platforms. Analyzing this data can reveal trends in streaming content, uncover viewer preferences, and inform content decisions for entertainment providers.development.

Case Study

A researcher is analyzing TV shows across multiple streaming platforms, focusing on viewer age demographics, ratings, and platform distribution. The aim is to explore patterns in viewer age suitability, correlate ratings across different platforms, and identify trends in streaming preferences.

Dataset Description

1. **Title:** The name of the TV show.
2. **Year:** The release year of the show.
3. **Age:** The age rating of the show.
4. **IMDb:** IMDb rating out of 10.
5. **Rotten Tomatoes:** Rotten Tomatoes rating out of 100%.
6. **Netflix, Hulu, Prime Video, Disney+:** Availability of the show on each platform (binary: 1 if available, 0 if not).

Problem Set

Solve the below given questions. Bonus questions, if any can be solved if you have additional time left but are not mandatory. Questions that aren’t applicable to the dataset provided, or questions that yield irregular output maybe omitted/the irregular output will be accepted.

Unit 1: Data Understanding and Preprocessing

1. **Feature Classification**:
   - Classify each variable in the dataset into the appropriate data types (nominal, ordinal, interval, or ratio). Explain your reasoning behind each classification.
2. **Data Quality Issues**:
   - Identify any data quality issues in the dataset, including missing values, duplicates, or inconsistent data formats. Explain the preprocessing steps you would take to clean and correct the data.
3. **Summary Statistics**:
   - Provide summary statistics for the numerical features, focusing on central tendency and dispersion. Calculate measures such as mean, median, and standard deviation for IMDb ratings and Rotten Tomatoes scores.
4. **Visualizations**:
   - Create histograms and box plots to explore the distributions of IMDb ratings and Rotten Tomatoes scores in the TV show dataset.
     - **i)** Describe the distribution shape (e.g., normal, skewed) for both IMDb and Rotten Tomatoes ratings.
     - **ii)** Calculate the number of outliers present in each rating column.
     - **iii)** *Bonus*: Adjust the visual scale, if necessary, to enhance the clarity of the distributions and outlier representation
5. **Outlier Handling**:
   - Explain how you would address any outliers identified in IMDb and Rotten Tomatoes ratings. Show the results before and after handling outliers using visualizations.
6. **Normal Probability Plot (Q-Q Plot)**:
   - Create and interpret a Q-Q plot for IMDb ratings. Based on the plot shape, describe any conclusions you can draw regarding the distribution of ratings.

7. **Correlation Analysis**:
    - Perform a correlation analysis between IMDb and Rotten Tomatoes ratings. Identify any other variables with strong correlations and interpret these relationships.
8. **Pair Plot Analysis**:
    - Generate a pair plot for a random sample of 100 shows to analyze relationships among IMDb, Rotten Tomatoes, and Year of release. Describe any noticeable patterns.

Unit 2: Hypothesis Testing

9. **Hypothesis Testing**:
    - Test whether there is a significant difference in mean IMDb ratings between shows available on Netflix and shows available on Hulu. Formulate the null and alternative hypotheses and perform an independent two-sample t-test to evaluate the difference.
10. **Margin of Error**:
    - Calculate the margin of error to assess the reliability of your analysis in the previous question. Interpret what this margin of error suggests about the accuracy of your findings.

Unit 3: Prediction and Feature Engineering

11. **Linear Regression Analysis**:
    - Use linear regression to predict IMDb ratings based on Rotten Tomatoes scores. Plot the predicted vs. actual ratings and calculate the MSE and RMSE.
12. **Feature Engineering**:
    - Propose two new features that could improve predictions of IMDb ratings, considering the existing features in the dataset.
