LSUHSC NO - Introduction to Data Science: Homework 01

Questions

September 13, 2024

# Part 1: Conceptual Questions

**1. Define Data Science**

Provide a brief definition of data science. What are its key components?

## 2. Structured vs. Unstructured Data

Explain the difference between structured and unstructured data. Give two examples of each.

**3. Data Quality Dimensions**

Discuss three critical dimensions of data quality. Why is each important?

## 4. Common Data Quality Challenges

List and describe three common challenges associated with ensuring data quality in data science.

## 5. Exploratory Data Analysis (EDA)

What is the purpose of EDA in the data science process? List three key techniques used in EDA.

## 6. Key Steps in the Data Science Process

Outline the key steps involved in the data science process, from data collection to deploying a machine learning model.

# Part 2: Practical Exercise with Python

Using the pandas and NumPy libraries, complete the following data handling exercises. Use the dataset provided as movie ratings.csv. If you do not have this file, create a small CSV file with similar columns based on the examples in the lecture notes.

## 1. Reading Data

Read the movie ratings.csv file into a pandas DataFrame and display the first 5 rows of the data.

## 2. Data Overview

Use pandas functions to find out:

* The number of rows and columns in the dataset.
* The column names and their corresponding data types.

## 3. Handling Missing Data

Identify any missing values in the dataset. If found, fill them with appropriate values or handle them using the appropriate method (e.g., fill with mean or drop rows).

## 4. Data Type Conversion

Convert the ’Release Date’ column to a datetime type and create a new column ’Release Year’ that extracts the year from the ’Release Date’.

## 5. Calculating Statistics

Compute the mean, median, and standard deviation of the ’Worldwide Gross’ and ’Production Budget’ columns.

## 6. Creating New Columns

Create a new column named Profit that calculates the difference between ’Worldwide Gross’ and ’Production Budget’.

## 7. Filtering and Subsetting Data

Filter the DataFrame to include only movies released after the year 2000. Save this subset to a new DataFrame.

## 8. Data Visualization (Optional)

Create a simple visualization (e.g., a histogram or bar plot) showing the distribution of the ’IMDB Rating’ column using matplotlib or seaborn.

# Part 3: Data Quality Exercise

## 1. Data Quality Analysis

Analyze the dataset provided and check for any inconsistencies, duplicates, or anomalies. Provide a brief report summarizing your findings and suggesting ways to improve the data quality.

# Submission Instructions

* Submit your code file (e.g., Homework 01.ipynb if you are using a Jupyter Notebook).
* Include a text document or PDF summarizing your answers to Part 1 and the data quality analysis from Part 3.
* Ensure your submission is clear, concise, and formatted properly for readability.