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?

## Data science is a multi-disciplinary field that uses scientific methods, processes, algorithms, and systems to extract knowledge, and insights from data in various forms, both structured and unstructured, like data mining. It contains components like visualization, statistical computing, statistical modeling, data technology, data research, data consulting, real-world applications, and scientific methods

## 2. Structured vs. Unstructured Data

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

Structured data consists of data that has been categorized or divided into a readable or organized structure where it becomes easy to clean, analyze, query, and visualize the data. E.g.: Spreadsheets, financial bank account information, etc.

Unstructured data consists of abstract data that cannot directly be analyzed and requires more effort as it is not organized in a very readable format. E.g.: Images, text-based answers to survey questions, etc.

**3. Data Quality Dimensions**

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

* Timeliness: We need to make sure the data meets the requirements for the time interval from collection to processing to analysis and it has been procured on time.
* Readability: We need to make sure that the data is readable and easy to understand for further use.
* Credibility: We have to make sure the data is taken from a credible source and will be accurate when utilized.

## 4. Common Data Quality Challenges

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

* The costs and time of the process of acquiring, storing, cleaning, retrieving, and processing unstructured data can add up to quite an investment before we can start reaping value from this process.
* The process of making sure that the data is of the highest quality can be quite time-consuming and cumbersome sometimes.
* Missing or inaccurate data can reduce the data quality. For instance, someone typed 4 instead of 44 and this inaccuracy through human error might affect the quality of that data.

## 5. Exploratory Data Analysis (EDA)

## Exploratory Data Analysis (EDA) is a crucial step in the data science process that involves understanding the underlying patterns, characteristics, and relationships within a dataset before formal analysis begins. It is a general overview of the data.

* Descriptive tables: we use this to understand the range and variety of data better
* Plots/graphs: data visualization which can summarize the type of observations
* Correlation: to find the variables that might need to be analyzed together on a larger scale

## 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.

Data collection – Data Preparation – Data Analysis – Visualization of data – Get results – Interpretations and deployment

# 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.

**DATA QUALITY REPORT**

After checking for some key data quality issues, performing the 8 tasks and making a quality check, here is a summary:

* No duplicate rows were found in the dataset.
* Production Budget: Ranges from approximately $1.3 million to $198.7 million, with a mean of $100.5 million.
* Worldwide Gross: Ranges from $2 million to nearly $1 billion, with a mean of $477.8 million.
* User Rating: Ranges from 1.0 to 5.0, with an average rating of 2.95.
* User Votes: Varies from 100 to 4997 votes, with a mean of 2569 votes.
* No obvious anomalies in the statistics (e.g., no extremely high or negative values).
* No anomalies (such as negative values) were found in these columns.
* Genres: The unique genres in the dataset are: Comedy, Drama, Thriller, Horror, Sci-Fi, Romance, Fantasy, and Action. No inconsistencies in genre naming were found.
* No missing data was found in any of the columns.

Suggestions:

So far, the data looks consistent and of high quality. Sources of data can be confirmed for accuracy. The movie names seem odd and different than usual movie names, so a source check for accuracy would ensure higher quality.