Second Assignment

Data Science Course

In this assignment, you will improve your real-world data skills through more emphasis on Exploratory Data Analysis. The purpose of this project is to develop your ability to extract meaningful insights from complex datasets, apply statistical methods to real-world problems, and communicate your findings effectively through visualization and storytelling.

The project contains three sections overall:

You are required to submit a notebook (.ipynb) file which has the code related to the practical section with proper and clear classification of each subsection. Your submission should also include two PDF files: one containing a full report on the code, methodology, analysis, visualization, storytelling; and another one for the answers to the theoretical section of this project. As mentioned before, **do not include any code in your report.**

1 Theoretical Questions

Question 1:

Explain the difference between correlation and causation with an example from a real-world dataset.

Question 2:

(a) What are the major types of issues found in raw data, and how do they affect analysis?

(b) List the four major tasks of data preprocessing.

(c) What are some methods for handling missing values in a dataset?

Question 3:

Describe the 'binning' method for managing noisy data. Give an example.

Question 4:

(a) Discuss the importance of data quality in EDA, including common issues like outliers and inconsistencies.

(b) Describe a scenario where data quality issues might lead to misleading conclusions in a real-world analysis.

(c) Explain how EDA techniques can be used to identify and address these issues.

Question 5:

What is normalization, and why is it important in EDA? Name three methods.

Question 6:

What is the goal of data reduction, and what techniques are commonly used

Question 7:

(a) Why is data visualization considered a powerful storytelling tool in data science?

(b) Compare traditional data presentation with storytelling visualizations using an example. What design elements transform a chart into a compelling story?

Question 8:

(a) What factors determine the best type of chart to use for a dataset?

(b) How can distribution charts help in EDA, and what questions do they answer?

(c) Discuss how a heatmap of a correlation matrix can help identify patterns in a multivariate dataset.

Question 9:

Compare bar charts, line charts, and pie charts in terms of the types of insights they provide and the nature of the data they are best suited for.

2 Practical Questions

2.1 YouTube Videos Analysis

Uncovering the Secrets of YouTube Trending Videos

Your mission is to explore the **YouTube trending videos dataset** and uncover the key factors behind viral content. Through **exploratory data analysis (EDA)** and **data visualization**, you will analyze engagement patterns, trends, and content characteristics. However, this is not just about answering a set of predefined questions—you need to craft your own **data-driven** story and develop meaningful insights based on it.

What You Need to Do

- **Prepare the Data** – Clean and preprocess the dataset to ensure it is ready for analysis.
- **Analyze Key Questions** – Use EDA and visualization to explore engagement trends and content patterns.
- **Develop Your Own Storyline** – Identify an interesting theme or question about YouTube trends and build your analysis around it.
- **Create Additional Meaningful Questions** – Based on your storyline, develop ***at least seven additional purposeful questions*** and answer them with EDA and data visualization.
- **Go Beyond the Dataset** – Web scraping to collect extra data from YouTube will earn you extra points.

Required Questions to Answer

Regardless of the story you choose, your analysis must provide insights into these key questions:

1. How are engagement metrics (views, likes and dislikes) distributed overall and across different video categories?
2. Which YouTube channels and video categories trend the most in each country and globally?
3. Are there seasonal or day-of-week patterns in trending videos? How does the upload day and time impact video engagement?
4. Do controversial videos, defined by a high dislike ratio, receive more engagement than universally liked ones?
5. How do video tags influence engagement, and which tags are most commonly used in trending videos?
6. How does the length of a video title impact engagement levels?
7. Is there a relationship between video title sentiment, whether positive, neutral, or negative, and engagement levels? (extra point)
8. Do clickbait-style titles, such as those containing words like ”shocking” or ”must watch,” result in higher engagement? (extra point)

Your Challenge

Beyond these questions, you must develop at least seven additional questions that align with the story you are telling. Your questions should be well-thought-out, data-driven, and provide deeper insights into your chosen theme. **Your entire analysis should flow as a cohesive narrative rather than just a list of separate analyses.**

Extra Points For

- A **strong and unique story line** that frames your analysis.
- **Well-structured** and meaningful additional questions that fit within your story.
- **Advanced and insightful visualizations** that clearly communicate your findings.
- Collecting extra data through **web scraping** to enhance your analysis.

Final Goal

This project is an opportunity to develop **real-world data analysis skills**. Use EDA and visualization to uncover trends, explain patterns, and tell a compelling story about YouTube trending videos. The best analyses will not just answer questions but reveal surprising, valuable, or thought-provoking insights about what drives video popularity.

Be analytical, be creative, and let your **data-driven** story shine!

2.2 Statistical Questions

1. Is there a significant association between the day of the week a video is published and its likelihood of trending?

2. Is there a significant difference in viewer engagement (likes-to-views ratio) across different video categories?

Hint: Use the **Kruskal-Wallis H test** to compare the distribution of engagement metrics across multiple video categories. This non-parametric test is appropriate when the data may not follow a normal distribution and when comparing more than two groups.

$$H = (N-1)\frac{\sum_{i=1}^{g} n_i(\bar{r}_i - \bar{r})^2}{\sum_{i=1}^{g}\sum_{j=1}^{n_i}(r_{ij} - \bar{r})^2}$$

where g is the number of categories, n_i is the number of observations in category i, r_{ij} is the rank of observation j from category i, N is the total number of observations, $\bar{r}_i$ is the mean rank of category i, and $\bar{r}$ is the mean of all ranks.