Mini-Project - Week 6 & 7

Note: This project is due at the end of week 7. There is no weekly assignment for week 6 but there will be a weekly assignment for week 7.

In this mini-project, you will work with a dataset and go through the data science process using the tools and techniques learnt from Week 1 - 5. The aim is to potentially use these tools together to achieve the objective of data exploration. The samples are for your reference but you may choose to do things differently.

1. **High Level View [2 pts]**
    a. Describe the dataset in words (50 words). Look at the data samples and describe what they represent and how they could be useful in a variety of data science tasks.

2. **Preliminary Exploration [4 pts]**
    a. In this step, you should explore what is present in the data and how the data is organized.

    You are expected to answer the following questions using the pandas library and markdown cells to describe your actions:

- Are there quality issues in the dataset (noisy, missing data, etc.)?
- What will you need to do to clean and/or transform the raw data for analysis?
- What are trends in the dataset using descriptive statistics (mean, median etc) and distribution of numerical data (eg. histograms)?

You are expected to show a minimum of 2 preliminary exploration tasks that you performed with justification. Typically, preliminary exploration helps us in identifying specific objectives for data analysis tasks (Step 3).

Sample :

a. Checking for null values (df.isnull().sum())
b. Histogram of distribution of happiness scores

3. **Defining objectives - [3 pts]**

Now that you have a better understanding of the data, you will want to form a research question which is interesting to you. The research question should be broad enough to be of interest to a reader but narrow enough that the question can be answered with the data. Some examples:

- Too Narrow: What is the GDP of the U.S. for 2011? This is just asking for a fact or a single data point.

- Too Broad: What is the primary reason for global poverty? This could be a Ph.D. thesis and would still be way too broad. What data will you use to answer this question? Even if a single dataset offered an answer, would it be defendable given the variety of datasets out there?
- Good: Can you use movie duration in a movie database to analyze viewer behavior over the years? If you have, or can obtain, data on a variety of movies and you have their box office earnings, this is a question which you can potentially answer well.

You are expected to define a minimum of 3 objectives for the mini-project.

Sample objectives -

a. How closely are GDP per capita and Happiness scores related? Or slightly more general, how are different factors correlated in determining happiness scores.
b. What are some of the geographies that have higher/lower happiness scores? Can they be aggregated or categorized by region?

4. Present Your Findings [9 pts]

This step involves using the libraries like numpy and pandas to extract data from the main dataset to forms that help answer the objectives - common applications would be filtering, aggregation, data modification, augmentation etc. The data analysis should allow you to create visualizations that make the report informative and easy to read.

1. You are required to present a minimum of 3 data analysis tasks and accompanying visualizations (one for each question) but any supporting visualizations can also be added. Visualizations should include -
    a. Justification for choice of plot
    b. Plot (with appropriate details and aesthetics)
    c. Inference/Conclusion from the visualization
2. The data analysis and visualization may not be done strictly in order. You may choose to report findings in a way that is easy to understand and read.

Sample visualizations -

a. Heatmaps & Scatterplots for correlation trends
b. Bar graphs/Pie Charts for categorical data

5. Ethics [2 pts]

Describe in words, or supporting visualization minimum 1 ethical concern you observe in the dataset.

Sample : Can we be confident that some of the features like ‘freedom’ can be represented numerically? Can there be bias?