Exploratory Data Analysis Project: Dive Into the Data! 🔍📊

For the final project, you'll be putting everything you've learned in the course to work by doing an Exploratory Data Analysis (EDA) on a dataset that interests you.

You can choose to work independently or collaborate with a team of up to 4 people.

The goal is to clean, analyze, and visualize the data, and uncover meaningful insights that are relevant to you.

It's your chance to dive into a dataset, explore it from different angles, and see what patterns or trends you can discover!

Dataset Selection

Choose a dataset that interests you and is suitable for performing EDA.

The dataset should have at least 5 variables and at least 100 data points (rows of data).

Possible sources for datasets include platforms like Kaggle or any other open data repository you prefer. You may also use a personal dataset if you have one available. Please submit a brief description of the dataset, including its source and what each of the features/columns represents.

Proposals of Datasets

30,000 Spotify Songs - Kaggle

Olympic Data - Kaggle

Employee Dataset - Kaggle

HR Analytics - Kaggle

Data Cleaning and Preprocessing

Inspect the dataset for missing values and determine appropriate methods for handling them (e.g., imputation, deletion).

Identify and handle duplicates in the dataset.

Address outliers, if relevant, and ensure the data is ready for analysis (e.g., correct data types, normalization if needed).

Exploratory Data Analysis (EDA):

Provide a high-level summary of the dataset using descriptive statistics (e.g., mean, median, mode, standard deviation).

Perform visual analysis of the data using various visualization techniques, such as:

- Histograms and box plots to understand distributions of numerical variables.
- Scatter plots or pair plots to explore relationships between continuous variables.
- Correlation matrices and heatmaps to identify relationships between variables.
- Bar charts for categorical data analysis.

Include at least one hypothesis or insight derived from the data

Insights and Conclusion

Based on your analysis, highlight key patterns, trends, and any potential anomalies you discovered.

Provide a brief written conclusion summarizing the insights, including how they could be useful in a business, research, or personal context.

If relevant, suggest next steps for further analysis or potential areas for improving the dataset.

Documentation and Code:

Code must be well-commented to explain the logic and reasoning behind each step.

Visualizations should be clearly labeled and include appropriate titles, axis labels, and legends.

Ensure that your notebook is easy to follow and includes both the code and its output (e.g., visualizations, statistics).

Evaluation Criteria:

- **Understanding the Dataset**: How well do you explain what the dataset is about and why it’s interesting or important? Are you able to describe the key features and what they represent?
- **Data Cleaning**: How effectively did you handle issues like missing data, duplicates, or outliers? Did you clean the data in a way that makes it ready for analysis?
- **Depth of Analysis**: How detailed and thorough is your analysis? Did you explore the data from multiple angles, using different types of visualizations and statistical techniques? Did you go beyond basic analysis to really dig into the data?
- **Insight and Relevance**: How valuable are the insights you've drawn from the data? Are your conclusions meaningful and connected to the original goals of the analysis? Did you make any interesting observations that could lead to further questions or actions?
- **Code Quality**: Is your code easy to read and follow? Is it organized, well-commented, and free of unnecessary complexity? Does it follow good programming practices and make it clear why you're doing each step?
- **Presentation**: How well did you explain your process and results? Is your project easy to understand, with clear visualizations and a well-written narrative that ties everything together?

Submission Guidelines

- Project Due Date: **16 Nov 2024**
- The project should be submitted via Canvas platform as a Jupyter Notebook file (.ipynb). You can use Google Colab to download it.
- Don´t forget to include the names of the participants.

Make sure to include any necessary instructions for running your notebook (pip install) and any additional files required (e.g., dataset if not linked).

Good Luck with Your Project!

I’m excited to see what insights you uncover from your data!

Remember, this is your chance to explore, experiment, and really make the project your own.

Don’t be afraid to ask questions if you get stuck or need guidance—I’m here to help.

Most importantly, have fun with it! Data analysis is all about discovering something new, and I can’t wait to see where your curiosity takes you.

Good luck, and enjoy the process!