Data Wrangling Final Project:

Comprehensive Data Wrangling and Analysis: Real-world Healthcare Dataset

Objective:

The purpose of this project is to provide students with a comprehensive, real-world dataset that necessitates extensive cleaning, transformation, and analysis. By the end of this project, students should be able to demonstrate proficiency in handling missing data, dealing with outliers, transforming variables, and conducting exploratory data analysis.

Project Description:

Students will work with a healthcare dataset, such as the Medical Expenditure Panel Survey (MEPS), which contains extensive information on healthcare utilization, expenditures, insurance coverage, and health status. This dataset is known for its complexity and is representative of the kinds of messy, real-world data that data scientists often encounter.

Deliverables:

1. **Cleaned and Transformed Dataset**
    - The final cleaned and transformed dataset should be uploaded to Kaggle, with appropriate documentation regarding the cleaning and transformation processes applied.
2. **Report**
    - A comprehensive report detailing the steps taken to clean and transform the dataset, including dealing with missing data, outliers, and variable transformations.
    - The report should also include exploratory data analysis conducted on the cleaned dataset, with appropriate visualizations and statistical analyses.
3. **Presentation**
    - A concise presentation summarizing the key findings and insights derived from the dataset.

Guidelines:

1. Dataset Acquisition:

- Download the Medical Expenditure Panel Survey (MEPS) dataset or a similar healthcare dataset.
- Review and adhere to the dataset's license and usage restrictions.

2. Data Wrangling:

- **Data Cleaning:**
  - Handle missing values appropriately.
  - Correct inconsistencies and inaccuracies in the dataset.
  - Handle outliers appropriately.
- **Data Transformation:**
  - Create new variables that are useful for analysis.
  - Transform variables to appropriate formats and types.
- **Data Reduction:**
  - Eliminate redundant variables.
  - Aggregate data where appropriate.
- **Data Integration:**
  - Integrate different parts of the dataset, if available in multiple files or tables.

3. Exploratory Data Analysis (EDA):

- Conduct a comprehensive exploratory data analysis to uncover patterns, relationships, anomalies, and trends in the dataset.
- Use appropriate visualizations to represent the findings.
- Perform statistical analysis to validate the findings. Univariate/Bivariate/Multivariate Analysi

4. Documentation:

- Document every step of the data wrangling and analysis process.
- Provide explanations and justifications for the choices made during the data wrangling process.

5. Kaggle Submission:

- Upload the cleaned and transformed dataset to Kaggle.
- Provide appropriate documentation and metadata for the uploaded dataset.

Evaluation Criteria:

1. **Data Wrangling (40 Points):**
    - Effectiveness of handling missing data.
    - Accuracy of data cleaning.
    - Appropriateness of data transformations.
    - Quality of data integration.
2. **Exploratory Data Analysis (30 Points):**
    - Comprehensiveness of the EDA.
    - Appropriateness of the statistical analyses.
    - Clarity and effectiveness of visualizations.
3. **Documentation and Reporting (20 Points):**
    - Quality of documentation.
    - Clarity and coherence of the report.
    - Relevance and conciseness of the presentation.
    - Write and publish a medium.com article and post on LinkedIn (Optional)
4. **Kaggle Submission (10 Points):**
    - Successful upload of the cleaned dataset.
    - Quality of documentation and metadata provided on Kaggle.

Timeline:

- **Due the last day of class**

Group Assignment:

- Students can form groups of up to 4 members.
- Each group must submit one copy of each deliverable.
- All group members are expected to contribute equally to the project.

Resources:

- Medical Expenditure Panel Survey (MEPS): MEPS Website
- Kaggle: Kaggle Website

Note:

Ensure to check the license and usage restrictions of the dataset before downloading and using it. The dataset must be publicly available and permissible for use in academic projects.

Submission:

- All deliverables should be submitted by [Submission Deadline].
- Late submissions will incur a penalty of [Specify Penalty] per day.

This project provides a comprehensive and challenging experience in data wrangling and analysis, allowing students to apply and demonstrate their skills in handling real-world, messy data.

*IMPORTANT NOTE: IF YOU WOULD LIKE TO USE A DIFFERENT DATASET, FEEL FREE TO DO SO BUT YOU MUST GET PERMISSION BEFORE TO ENSURE THE NEW DATASET IS "ROUGH" ENOUGH TO BENEFIT FROM DATA WRANGLING. You can consider things like: https://www.kaggle.com/competitions to find another dataset.