DA 6223 Course Project

Description: For the course project, each team of 4-5 students will be evaluated based on a 15-minute presentation and a final project report.

Project Teams

In this part, you will be working with other students in your team. Consider teaming up with students who have the same intention. Each team can have 4 or 5 students.

If you need me to find a group for you, please notify me as soon as possible.

Project Conceptualization and Data Collection

One important aspect of this part is to identify the problem and collect the data sets that you are going to use for the project. Each team may use **only publicly available data sources** subject to and in compliance with applicable license terms that allow full usage in this course and are not subject to any other restrictions on use or access. Data sets must be anonymized and not include personally identifiable or confidential information. Make sure to include the link to the data or data sets. Teams must give appropriate attribution as applicable.

I suggest you find a messy dataset that you can clean, summarize, and prepare for further analysis. Try government resources or the original (uncleaned) source if going with a Kaggle dataset.

Here are some potential websites to collect data:

https://www.google.com/publicdata/directory
https://www.data.gov/
https://github.com/awesomedata/awesome-public-datasets
https://registry.opendata.aws/
https://www.kaggle.com/datasets

Overview

For the project, teams will work on improving the data quality of public datasets using SAS Enterprise Guide and SQL. The project involves a comprehensive data-cleaning workflow, covering aspects such as data quality, validity, accuracy, completeness, consistency, and uniformity.

This document provides you with a guideline for the project. It includes the tasks I expect each team to complete, potential quality check examples, a rubric for assessment, and guidelines for peer- and team-evaluations.

Project Deliverables

Each team is expected to define the analytical problem, conduct data preparation, and provide a reasonable summary of the data. We specifically focus on the following aspects:

- Problem Statement
- Data
- Data Cleaning/Validation (see the additional pdf on data cleaning)
- Visualization (graphs, charts, etc.)
- Generalization and Suggestions

1 - SAS Enterprise Guide Project Flow

A SAS Enterprise Guide project file (.egp) needs to be submitted through Blackboard by each team before the presentation. This project file will be able to provide all the results (data sets, tables, figures, etc.) on your presentation slides.

2 - Presentation

Each team will be delivering a 15-minute presentation on the project. During the presentation, each of the team members is required to speak in terms of presenting the project or answering questions. The presentation will be evaluated by the audience in class (your classmates and the instructor). Presentations by the teams will be judged according to the following criteria and considerations:

- Problem definition and explanation
- Data source, explanation, and verification
- Use of appropriate methodology
- Presentation of results for a general non-technical audience
- Conclusions and recommendations

The presentation file also needs to be submitted through Canvas.

3 - Report

Each team is expected to conduct data analysis and write a project report that defines the problem, describes the analysis performed, and presents the results in such a manner as to be useful in business, science, government, education, health care, etc. Each team will be judged on the broad accessibility of results, so papers should be appropriately technical but still understandable to non-analytics audiences. The report must be two to five (2-5) pages long (with single space), excluding the appendix. The written report must contain the following components:

- Introduction - Is the introduction appropriate and provide a concise lead-in to the report given?
- Data - Is the source of the data adequately documented?
- Problem - Is the problem(s) clearly defined and the objectives of the study precisely given?
- Data cleaning/validation - Is the data cleaning process adequately explained?

- Analysis - Were appropriate analytical methods used and adequately explained?
- Visualization - Are appropriate graphics and visuals provided?
- Generalization - Are results provided for a general, non-technical audience?
- Suggestions for future studies - Are issues for future studies/different problem approaches given?
- Conclusions - Are conclusions concisely given?

4- Peer and Team Evaluation

Each student is required to complete the evaluation form in order to avoid the deduction of points for this project. Please fill out the evaluation form posted on Canvas. In the evaluation form, the maximum points to be given are written under each criterion, and the total column is the sum of four columns.

Each student should evaluate the other teams and students fairly and reasonably. Please use the presentation evaluation rubric given at the end of this document.

Not only will you evaluate the other teams, but you will also evaluate your team members. **Please DO NOT evaluate your own team and yourself.**

Project Tasks Checklist

1. **Data Inspection:**
    - Import data
    - Conduct data profiling using the SAS Enterprise Guide tasks (e.g., Data Set Attributes and Characterize Data) or PROC SQL.
    - Utilize summary statistics to describe data characteristics.
2. **Data Cleaning:**
    - Clean irrelevant data and remove duplicates (e.g., Query Builder with "select distinct rows only.")
    - Make necessary type conversions.
    - Correct syntax errors (extra white space, typos).
    - Standardize data (uppercase, lowercase, same measure unit).
    - Scale data if necessary.
    - Drop missing values with appropriate justification.
    - Address outliers and document removal decisions.
3. **Advanced PROC SQL Queries:**
    - Utilize advanced PROC SQL queries to demonstrate a deeper understanding of SQL capabilities.
    - Summarize data utilizing subqueries, in-line views, and/or complex joins (e.g., reflexive joins)
4. **Data Verification:**
    - Inspect the results after cleaning to verify correctness.
5. **Reporting:**
    - Create visualizations (tables, figures, graphs) to summarize data.
    - Create a detailed report documenting the changes made during the cleaning process.
    - Assess and report on the quality of the currently stored data.
6. **Customized Reporting:**
    - Explore advanced techniques for generating customized reports using SAS EG.
    - Use prompts in tasks and queries to enhance user interactivity.
7. **Additional Task (Choose One):**
    - **Handling Categorical Data:**
        - Addressing missing values in categorical variables.
        - Encoding categorical variables appropriately.
        - Exploring techniques for handling imbalanced classes.
    - **Advanced Data Transformation:**
        - Feature engineering: Creating new features based on existing ones.
        - Exploring and applying advanced data transformation techniques.
    - **Outlier Detection and Treatment:**
        - Utilizing statistical methods or machine learning algorithms for outlier detection.
        - Applying appropriate techniques to handle outliers.
    - **Data Imputation:**
        - Implementing strategies for imputing missing data, especially in numerical variables.
    - **Temporal Analysis:**
        - Handling time-series data if applicable.

  - Exploring trends, seasonality, and patterns over time.
- **Data Security Considerations:**
  - Ensuring the project adheres to data security and privacy best practices.
  - Discussing any potential ethical considerations related to the dataset.
- **Documentation:**
  - Providing clear and comprehensive documentation for the entire data cleaning and transformation process.
  - Including comments in the code to explain reasoning and decisions made during the project.
- **Collaboration and Communication:**
  - Encouraging collaboration within the team (if applicable) and effective communication of findings.
  - Presenting key insights and improvements in a concise and understandable manner.
- **Performance Metrics:**
  - Defining and measuring performance metrics for the data cleaning process.
  - Evaluating the impact of the cleaning process on subsequent analytics tasks.
- **Exploratory Data Analysis (EDA) Enhancement:**
  - Expanding EDA techniques beyond the basics, incorporating more advanced statistical methods.
  - Developing insights into relationships and patterns within the data.

Rubric for Project Assessment (INSTRUCTOR USE ONLY)

The project will be assessed based on the following criteria:

1. **Data Inspection (20 points):**
   - Effective use of data profiling techniques.
   - Clear summary statistics and data descriptions.
2. **Data Cleaning (30 points):**
   - Proper handling of irrelevant data and duplicates.
   - Accurate type conversions and syntax error corrections.
   - Consistent data standardization.
   - Justification for dropping missing values and addressing outliers.
3. **Data Verification (20 points):**
   - Thorough inspection and verification of cleaned data.
   - Evidence of correctness after the cleaning process.
4. **Reporting (20 points):**
   - Comprehensive visualizations to summarize data.
   - Clear and detailed report documenting changes made during cleaning.
   - Assessment of the quality of the currently stored data.
5. **Advanced PROC SQL Queries (20 Points):**
   - Demonstrates proficiency in using advanced SQL features.
   - Applies subqueries, in-line views, and/or complex joins effectively.
6. **At least one additional task (20 Points):**
   - To enhance the complexity and comprehensiveness of your capstone project, perform.
7. **15-Minute Presentation (25 points from audience):**
   - Clear articulation of the project objectives, methods, and findings.
   - Engaging presentation style.
   - Addressing questions effectively during the Q&A session.
8. **5-6 Page Project Report (15 points):**
   - Well-organized report covering project goals, methodology, results, and conclusions.
   - Clear explanations of the data cleaning and analysis process.
   - Proper use of citations and references if external sources are used.
9. **SAS Enterprise Guide Project Flow (20 points):**
   - Demonstrates a well-organized and logical flow in the SAS Enterprise Guide project.
   - Efficient use of tasks and steps to achieve the project objectives.
   - Effective use of SAS Enterprise Guide functionalities.
10. **Peer-Review Evaluation (10 points):**
    - Contributes constructively to peer reviews.
    - Provide thoughtful and insightful feedback to peers.
    - Demonstrates professionalism and collaboration in the peer-review process.

Total: 200 points

You can use the following rubric to evaluate other teams' presentations. (To be used by the audience)

Criteria	Excellent (100%)	Good (80%)	Fair (60%)	Needs Improvement (30%)	Unsatisfactory (10%)	Total Points
Content and Structure (8 points)	- Clearly articulates project objectives, methods, and findings. - Presentation follows a logical structure with a clear introduction, detailed explanation of data cleaning steps, and a conclusion summarizing key insights.	- Articulates project objectives, methods, and findings effectively. - Presentation follows a generally logical structure with an introduction, an overview of data cleaning, and a conclusion summarizing key insights.	- Some articulation of project objectives, methods, and findings. - Presentation lacks a consistent logical structure and may not cover all data cleaning steps adequately.	- Articulation of project objectives, methods, and findings is unclear. - Presentation lacks a coherent structure, and some key data cleaning steps are not covered.	- Incoherent articulation of project objectives, methods, and findings. - No clear structure in the presentation, and major data cleaning steps are omitted.	8
Engagement and Delivery (8 points)	- Engages the audience effectively with a confident and enthusiastic delivery. - Maintains eye contact, uses appropriate gestures, and speaks clearly. - Demonstrates passion for the project.	- Engages the audience with a generally confident delivery. - Maintains good eye contact, gestures appropriately, and speaks clearly. - Conveys enthusiasm for the project.	- Engagement with the audience is somewhat effective. - Delivery lacks consistency in confidence. - Occasional lapses in eye contact and clarity. - Limited expression of enthusiasm.	- Limited engagement with the audience. - Delivery lacks confidence, with frequent lapses in eye contact and clarity. - Minimal expression of enthusiasm.	- Minimal engagement with the audience. - Lack of confidence and frequent lapses in eye contact and clarity. - No expression of enthusiasm.	8
Q&A Session (4 points)	- Responds to questions with clarity and demonstrates a deep understanding of the project. - Handles questions confidently and provides additional insights.	- Responds to questions with clarity and demonstrates a good understanding of the project. - Handles questions confidently most of the time. - Provides additional insights when appropriate.	- Provides adequate responses to questions but may lack depth in understanding. - Some difficulty in handling questions confidently. - Limited additional insights provided.	- Struggles to provide clear and adequate responses to questions. - Demonstrates a limited understanding of the project. - Minimal additional insights provided.	- Unable to answer questions effectively. - Demonstrates a lack of understanding of the project.	4
Visual Aids (5 points)	- Visual aids are clear, well-designed, and enhance the presentation. - Images, charts, and graphs effectively illustrate key data-cleaning steps and insights.	- Visual aids are generally clear and well-designed. - Most images, charts, and graphs effectively support key data cleaning points. - Visual aids contribute to the overall understanding of the project.	- Visual aids are somewhat clear and may lack consistency in design. - Some images, charts, or graphs could be improved for clarity. - Visual aids provide limited support to key data cleaning points.	- Visual aids are unclear or inconsistent in design. - Images, charts, or graphs may not effectively support key data cleaning points. - Visual aids may confuse rather than clarify.	- Visual aids are confusing or irrelevant. - Detract from rather than support key data cleaning points.	5

A printed version of this form needs to be submitted after the presentations end. **Please DO NOT evaluate your own team and yourself**.

Name:_______________________________ **Team**:______________________________

Peer-Evaluation

Name	Contribution to propose and define business problems	Contribution to data analysis	Contribution to preparing presentation	Contribution to writing the report	Overall Participation	Total
Member 1						
Member 2						
Member 3						
Member 4						

Team Evaluation

Aspects	Quality of Slides	Organization	Presentation Skills	Technical Analysis	Impact of Results	Total
Team 1						
Team 2						
Team 3						
Team 4						
Team 5						
Team 6						

Required comments about other teams' performances:

____

____

____

____

____

____

____

____

____

____

____

____

____