CSC17104 – PROGRAMING FOR DATA SCIENCE

FINAL PROJECT

1. Introduction

Overview:

This project provides you with hands-on experience in conducting a complete data science workflow—from finding raw data to extracting meaningful insights. You will work in groups of **n people** to explore a real-world dataset of your choice and tell a compelling data story.

What you'll do:

- **As a group**, find and select a public dataset that interests your team
- Explore and understand the data through visualization and statistical analysis
- Identify **2 × n meaningful questions** that can be answered with your data (e.g., a group of 3 students must formulate 6 questions)
- Ensure each question is **substantial and challenging enough** to demonstrate deep analytical thinking
- Include **at least 1 question that requires building and evaluating a machine learning model** to solve
- Clean and preprocess the data to prepare it for analysis
- Conduct thorough analysis to answer each question using appropriate statistical methods and visualizations
- Draw conclusions and communicate your findings clearly

Dataset Requirements

Your dataset must meet the following criteria:

- **Source:** Publicly available from platforms such as:
  - Kaggle (https://www.kaggle.com/datasets)
  - UCI Machine Learning Repository (https://archive.ics.uci.edu/)
  - Data.gov (https://data.gov/)
  - Google Dataset Search (https://datasetsearch.research.google.com/)
  - Other reputable public data sources
- **Size:** Minimum 1,000 rows and 10 columns

- **Format:** CSV, Excel, JSON, or similar structured format
- **Quality:** Should have sufficient complexity for meaningful analysis (not too clean, but not impossibly messy)
- **Relevance:** Choose a topic that genuinely interests you or your group

Note:

- Avoid datasets that are too simple (e.g., Iris, Titanic) unless you can demonstrate advanced analysis techniques.
- Synthetic datasets are not permitted. You must use real-world data collected from actual observations, measurements, or events. Synthetic datasets lack the complexities, nuances, and real-world context that make data analysis meaningful and challenging.
- (Teacher will review your Dataset choice)

2. Project workflow

Your project will follow the standard data science workflow, consisting of five main phases:

1. Data Collection
2. Data Exploration *(often interleaved with preprocessing)*
3. Question Formulation
4. Data Analysis *(preprocessing + analysis per question)*
5. Conclusions & Reflection

Note: Phases 2, 3, and 4 are iterative—you may revisit earlier phases as you gain deeper understanding of your data.

The following guidelines are provided as a reference framework, not a rigid checklist. You should:

- Adapt these steps to fit your specific dataset and questions
- Add any additional exploration or analysis tasks you deem important
- Skip steps that are not relevant to your data
- Think critically about what your dataset needs

2.1 Data Collection

Answer the following questions in your notebook to document your dataset:

What subject is your data about?

- Describe the topic, domain, or phenomenon
- What real-world context does this represent?

What is the source of your data?

- Platform name (Kaggle, UCI, etc.) and full URL
- Original author(s) or organization
- Publication/collection date

Is this data licensed for your use?

- What license does the dataset have?
- Are you permitted to use it for educational purposes?
- Document any usage restrictions or attribution requirements

How was this data collected?

- Collection method (survey, sensors, administrative records, web scraping, etc.)
- Target population and sampling approach
- Time period of data collection
- Any known limitations or biases in collection

Why did you choose this dataset?

- What interests your group about this topic?
- What potential questions or insights could this data provide?

2.2 Data Exploration

Thoroughly investigate your dataset's structure, quality, and characteristics to understand what you're working with and identify potential issues.

Dataset overview

<u>Basic Information</u>

- How many rows does your dataset have?
- How many columns does your dataset have?
- What does each row represent? (e.g., one customer, one transaction, one day)
- What is the overall size of the dataset?

<u>Data Integrity</u>

- Are there any duplicated rows? If yes, how many?
- Should duplicates be kept or removed? (Justify your decision)
- Are all rows complete, or are some entirely empty?

Column Inventory

- What is the meaning/definition of each column?
- Which columns are relevant to potential analysis?
- Are there any columns that should be dropped? Why?

Data Types:

- What is the current data type of each column?
- Are there columns with inappropriate data types?
- Which columns need type conversion?

Numerical Columns Analysis

For each numerical column, investigate:

Distribution & Central Tendency:

- What is the distribution shape? (normal, skewed, bimodal, uniform)
- Create visualizations: histograms, box plots, density plots,…
- Calculate: mean, median, standard deviation

Range & Outliers:

- What are the minimum and maximum values?
- Are min/max values reasonable, or do they indicate errors?
- Identify outliers using box plots, IQR method, or z-scores
- Are outliers genuine extreme values or data entry errors?

Data Quality:

- What percentage of values are missing?
- Are there any impossible values? (e.g., negative ages, prices = 0)
- Are there placeholder values? (e.g., 999, -1, 0 used to indicate missing)

Categorical Columns Analysis

For each categorical column, investigate:

Value Distribution:

- How many unique/distinct values are there?
- What are the top 5-10 most frequent values?

- Create visualizations: bar charts, count plots
- Is the distribution balanced or highly imbalanced?

<u>Data Quality:</u>

- What percentage of values are missing?
- Are there inconsistencies in categories?
  - Example: "Male", "male", "M", "m" all meaning the same thing
  - Example: Typos or variations in spelling
- Are there unexpected or abnormal values?
- Are there categories with very few observations? Should they be grouped?

Missing Data Analysis

<u>Overall Assessment:</u>

- Create a missing values summary: column name, count, and percentage missing
- Visualize missing data patterns (heatmap or bar chart)
- Are missing values random, or is there a pattern?
  - Do certain rows or groups have more missing values?

<u>Per Column Strategy:</u>

- For each column with missing values:
  - Why might values be missing? (random, not applicable, data collection issue)
  - What is your plan to handle them? (remove, impute, keep as separate category)

Relationships & Correlations

<u>Preliminary Patterns:</u>

- Calculate correlation matrix for numerical variables
- Create correlation heatmap
- Identify strongly correlated pairs (positive or negative)
- Are there any surprising relationships?

<u>Cross-tabulations:</u>

- For important categorical × categorical combinations, create frequency tables

- For numerical × categorical combinations, create grouped summary statistics

Initial Observations & Insights

<u>Summary:</u>

- What are 3-5 key observations from your exploration?
- What data quality issues did you identify?
- What preprocessing steps will be necessary?
- What interesting patterns emerged that could lead to research questions?

<u>Red Flags:</u>

- List any serious data quality concerns
- Note any limitations that might affect your analysis

2.3 Question formulation

Develop meaningful, challenging research questions that will drive your analysis.

Quantity:

- **2 × n meaningful questions** that can be answered with your data (e.g., a group of 3 students must formulate 6 questions)
- **At least 1 question must require a machine learning model** to answer

Quality Criteria:

Each question should be:

Meaningful:

- Has clear practical or theoretical value
- Provides actionable insights or deeper understanding
- Relates to real-world applications or decision-making

Challenging:

- Requires substantial analysis, not just a simple calculation
- Cannot be answered with a single line of code or basic function
- Involves multiple steps: data preparation, analysis, visualization, and interpretation
- Demonstrates analytical depth and critical thinking

Important: Focus on **quality over quantity**. It's better to have fewer well-crafted, insightful questions than many superficial ones. Each question should lead to meaningful discoveries about your data.

Documentation for Each Question:

In your notebook, present each question with the following structure:

1. The Question

- State your research question clearly and specifically
- Make it precise enough to be answerable with your data

2. Motivation & Benefits

- Why is this question worth investigating?
- What benefits or insights would be answering this question provide?
- Who would care about the answer? (stakeholders, decision-makers, researchers, etc.)
- What real-world problem or decision does this inform?

2.4 Data Analysis

For each research question, complete the following:

A. Preprocessing (if needed)

Written Explanation:

- Describe preprocessing steps clearly in markdown
- Sketch the workflow so readers understand **without reading code**
- Explain the logic and reasoning behind each step
- Use numbered steps or bullet points

Code Implementation:

- Implement each step with clean, readable code
- Use **meaningful variable names**
- Add **comments for non-obvious logic** (explain WHY, not just WHAT)
- Keep **lines concise** (< 100 characters; break long chains across lines)
- Show intermediate results when helpful

B. Analysis

Written Explanation:

- Describe your analytical approach in markdown
- Explain what methods you'll use and why
- Outline expected outputs (statistics, visualizations, models)
- Write so readers understand methodology **without reading code**

Code Implementation:

- Implement analysis following code quality standards:
  - Meaningful variable names
  - Strategic comments
  - Concise, readable lines
  - Proper visualization labels (title, axes, legend)

For ML Questions specifically:

- Explain: problem setup, data split, models chosen, evaluation metrics
- Code: train ≥ 2 models, evaluate thoroughly, compare performance, interpret features

C. Results & Interpretation

Visualizations:

- Create 2+ relevant, well-labelled plots per question
- Proper titles, axis labels, legends, units

Written Analysis:

- Answer the question explicitly with evidence
- Cite specific numbers, statistics, patterns from your analysis
- Discuss practical meaning and implications
- Note surprises or unexpected findings
- Acknowledge limitations

2.5 Project Summary

Key Findings:

- List 3-5 most important insights from your analysis

- Highlight the most interesting or surprising discovery

Limitations:

- Dataset limitations (sample size, biases, missing data)
- Analysis limitations (methodology constraints, unanswered aspects)
- Scope limitations (what you couldn't address)

Future Directions (If You Had More Time)

- What additional questions would you explore?
- What deeper analysis would you conduct?
- What alternative methods or approaches would you try?
- What additional data would you seek?
- How could this work be expanded or improved?

Individual Reflections

Each group member should write a personal reflection covering:

Challenges & Difficulties Encountered:

- What specific obstacles did you face? (technical, analytical, conceptual)
- How did you overcome them?
- What was most challenging and why?

Learning & Growth:

- What have you learned? (technical skills, analytical approaches, domain knowledge)
- What surprised you most?
- How has this project shaped your understanding of data science?

3 Project Deliverable

You must submit the following items:

1. Team Plan and Work Distribution

- Team member information
- Work breakdown by member (tasks, contributions, percentage)
- Collaboration process description

- Project planning and timeline

2. Jupyter Notebook(s)

- **Single notebook preferred** (if manageable size)
- **Multiple notebooks allowed** if project is too large (>5000 lines)
    - Include logical organization and clear navigation
- Must include all phases: data collection, exploration, questions, preprocessing, analysis, conclusions
- All code, visualizations, and documentation
- Runs without errors from top to bottom

3. README File (README.md)

- Project overview and team info
- Dataset source and description
- Research questions list
- Key findings summary
- File structure explanation (if multiple notebooks)
- How to run instructions
- Dependencies list

4. Additional Source Code (if applicable)

- Custom Python modules (utils.py, models.py, etc.)
- Helper scripts
- Well-documented and commented