Project Milestone 1: Dataset Selection & Exploratory Data Analysis

I306 Statistics for Informatics

Overview

For your semester project, you will conduct a complete statistical analysis of a dataset of your choosing. This first milestone focuses on selecting an appropriate dataset and performing initial exploratory data analysis.

Due: End of Week 4 **Points:** 40

Dataset Requirements

Your dataset must meet the following criteria:

- At least 500 observations (rows)
- At least 5 variables
- A mix of numeric and categorical variables
- Publicly available (provide the source URL)

Suggested Data Sources

- Kaggle Datasets
- UCI Machine Learning Repository
- Data.gov
- FiveThirtyEight
- TidyTuesday
- Our World in Data

Deliverables

Submit a Quarto document (.qmd) and its rendered PDF containing:

1. Dataset Description (10 points)

- What is your dataset about?
- Where did it come from? Include a URL.

- How was the data collected?
- What are the observational units (what does each row represent)?

2. Variable Descriptions (10 points)

Create a table listing each variable:

Variable Name	Type (Numeric/Categorical)	Description
...	...	...

Identify which variables you plan to use as:

- Response variable(s)
- Explanatory variable(s)

*Note: If your dataset has more than 10 or so variables, just list 10 that you find interesting.

3. Summary Statistics (10 points)

For numeric variables:

- Mean, median, standard deviation, min, max
- Identify any obvious outliers (extreme values that don't seem to fit the distribution of the rest of the data).

For categorical variables:

- Frequency counts
- Proportions

4. Contingency Table (10 points)

Create at least one contingency table showing the relationship between two categorical variables. If your dataset doesn't have two suitable categorical variables, you may bin a numeric variable into categories.

Submission

Submit your `.qmd` source file and rendered output (PDF or HTML) to Canvas by the due date.

Grading Rubric

Component	Points	Criteria
Dataset Description	10	Complete, accurate description with source
Variable Descriptions	10	All variables documented with types
Summary Statistics	10	Appropriate statistics computed and interpreted
Contingency Table	10	Correctly constructed with meaningful interpretation

Tips

- Choose a dataset you find genuinely interesting—you'll be working with it all semester
- Make sure your dataset is rich enough to support the analyses we'll cover: visualization, hypothesis testing, confidence intervals, and regression
- If you're unsure whether your dataset is appropriate, ask during office hours