Midterm Exam: Data Analysis Project

Overview

This midterm exam is a comprehensive analytical project that demonstrates your ability to work with real-world data using Python. You will independently analyze a dataset of your choice, applying the statistical concepts and programming skills covered in Weeks 1-5.

Project Components

Your submission must include:

1. **Jupyter Notebook** (.ipynb file) containing:
    - Well-commented Python code
    - Clear section headers for each analysis stage
    - Inline markdown explanations of your methodology and findings
2. **Analytical Report** (9-10 pages) summarizing:
    - Dataset description and source
    - Key insights and findings
    - Interpretation of results
    - Limitations and potential areas for further analysis

Required Analysis Tasks

1. Data Preparation & Exploration (20%)

- Load your dataset using Pandas
- Display basic information about the dataset (dimensions, data types, missing values)
- Perform necessary data cleaning (handle missing values, remove duplicates, correct data types)
- Document any transformations or decisions made during cleaning

2. Descriptive Statistics (25%)

- Calculate and present measures of central tendency (mean, median, mode) for key numerical variables
- Compute measures of dispersion (variance, standard deviation, range, IQR)
- Identify and analyze different data types in your dataset (categorical, numeric, continuous)

- Create summary tables showing descriptive statistics grouped by categorical variables
- Identify and discuss any outliers

3. Data Visualization (25%)

- Create **at least 5 different visualizations** that include:
    - Histogram showing distribution of numerical variables
    - Boxplot comparing distributions across groups or identifying outliers
    - Bar chart for categorical data analysis
    - Scatterplot exploring relationships between variables
    - At least one additional chart type relevant to your data
- Ensure all visualizations have proper labels, titles, and legends
- Use appropriate chart types based on data characteristics

4. Relationship Analysis (20%)

- Compute correlation coefficients (Pearson and/or Spearman) between numerical variables
- Create and interpret a correlation matrix or heatmap
- Perform grouped analysis to explore how relationships vary across categories
- Visualize relationships using scatterplots with regression lines or grouped visualizations
- Discuss the strength and direction of observed relationships

5. Probability & Distribution Analysis (10%)

- Identify variables that appear to follow common probability distributions (Normal, etc.)
- Test for normality using visualizations (Q-Q plots, histograms with normal curve overlay)
- If applicable, simulate random samples from theoretical distributions to compare with your data
- Calculate and interpret probabilities for specific events in your dataset

Technical Requirements

- Use **NumPy** and **Pandas** for data manipulation and statistical calculations
- Use **Matplotlib** and/or **Seaborn** for all visualizations
- Code must be well-organized, readable, and properly commented
- Notebook should run without errors from top to bottom

Dataset Selection

You may:

- Choose your own dataset from sources like Kaggle, UCI Machine Learning Repository, government open data portals, or other reputable sources
- Select one of the suggested datasets below
- Use a dataset provided by the instructor

Dataset requirements:

- Minimum 500 observations
- At least 5 variables (mix of numerical and categorical)
- Real-world data (not synthetic examples)
- Clearly cite your data source

Suggested Datasets

1. Airbnb Listings Dataset

- **Source:** Kaggle (https://www.kaggle.com/datasets/joyshil0599/airbnb-listing-data-for-data-science)
- **Description:** Detailed listings data for Airbnb rentals in various cities worldwide

2. World Happiness Report

- **Source:** Kaggle (https://www.kaggle.com/datasets/unsdsn/world-happiness)
- **Description:** Country-level data on happiness scores and contributing factors

3. Student Performance Dataset

- **Source:** UCI Machine Learning Repository or Kaggle (https://archive.ics.uci.edu/dataset/320/student+performance)
- **Description:** Student achievement in secondary education with demographic and social factors

4. Global Video Game Sales

- **Source:** Kaggle (https://www.kaggle.com/datasets/gregorut/videogamesales)
- **Description:** Sales data for video games across different platforms and regions

5. Employee Attrition and Performance

- **Source:** Kaggle (https://www.kaggle.com/datasets/pavansubhasht/ibm-hr-analytics-attrition-dataset/data)
- **Description:** Fictional employee data including demographics, job satisfaction, and attrition status

Submission Guidelines

Submit the following files:

1. name_surname_midterm_analysis.ipynb - Your Jupyter Notebook
2. name_surname_midterm_report.pdf - Your analytical report
3. dataset.csv (or appropriate format) - Your dataset file
4. README.txt - Brief description of dataset source and any setup instructions