Boston University

Department of Mathematics and Statistics

# MA213 - Fall 2025

Basic Statistics and Probability

Project 1 : Data Exploration

## Overview

Your group will select a dataset of your interest and conduct exploratory data analysis. This should include :

- One numerical exploratory data anlysis

- One categorical exploratory data analysis

You will present your analyses through either:

- An in-class presentation

- A pre-reorded video presentation

## Suggested Outline

- Introduction

  1. Introduce the dataset and its source

  2. Tell the story of your reasoning and possible hypothesis

  3. Define your variables of interest

- Data Analysis

  1. Numerical Exploratory Data Analysis Part

    (a) Summary statistics

    (b) The relationship between two numerical variables

    (c) Distribution of variable(s) of interest

    (d) Discuss shape, central tendency, spreadity and outliers.

2. Categorical Exploratory Data Analysis Part

    (a) Summarize categorical variables

    (b) Contingency Table (two categorical varaibles)

    (c) Visualizations: bar plots, pie charts, etc.

- Conclusion

1. Summarize key insights

2. Suggest future analysis or potential applications

---

# Project requirements

- Source of data must be cited

- R code should be included (Rscript or Rmd files)

- References should be provided at the end

---

# Deadlines

| Item | Description | Due |
|---|---|---|
| Deliverable | Brief introduction of the analysis and roles of group members | Week 5 (after Lab 3) |
| R Script or RMarkdown File | Include data preparation, analysis, and visualizations | Week 7 |
| Slide File | 4–5 slides (excluding Title and Reference slides); 5-minute presentation | Week 7 |

Table 1: Deliverable and Deadlines

## Data Sources to Consider

- OpenIntro Datasets

- fivethirtyeight R Package

- datsets R package

- Kaggle datasets

- US government's open data

# Evaluation Criteria

| Pass | Almost Pass | Not Yet |
|------|-------------|---------|

*Each item is marked with ✔, ✔−, or ✘:*

✔ = *satisfied*        ✔− = *mostly satisfied*        ✘ = *not yet*

✔  ✔−  ✘  Introduction section correctly summarize motivation, data, result and takeaways

✔  ✔−  ✘  Code correctly implemented (Results can be easily reproduced)

✔  ✔−  ✘  Conclusion section is coherent and the direction for future work is interesting and plausible

✔  ✔−  ✘  Discussed shape, central tendency, spreadity in a correct manner

✔  ✔−  ✘  Plots and summaries support your explanation

✔  ✔−  ✘  Well participated in the project as a group work

✔  ✔−  ✘  Answered one question from the presentation

✔  ✔−  ✘  Topic and the results are interesting and easy to understand

✔  ✔−  ✘  Asked a relevent/valid question to other group

# Grading Criteria

- **Pass:** ✔ on 7 of 9 rubric items, no ✘ marks

- **Almost Pass:** ✔ on 3 of 9 rubric items, no ✘ marks