MSIS 502 Group Project
Total: 50 points

1. Project Overview

The aim of this project is to integrate Python programming, data visualization techniques, and business decision-making to derive actionable insights from a real-world dataset. Data analytics are best learned by **doing**. Students will work in groups to analyze data, create visualizations, and present their findings in a comprehensive report and presentation, highlighting the business implications of their analysis.

2. Evaluation

You will be graded on your ability to demonstrate your grasp of basic python programming and understanding of different types of visualizations and how they can be applied in a business context. The projects will be evaluated on:

- **Data Preparation (5%)**: Quality and thoroughness of data cleaning and preprocessing.
- **Visualizations (35%)**: The quality and originality of the ideas and the extent of analyses
  - Quality and effort of exploratory analyses and visualization
  - Clarity, relevance, and informativeness of visualizations
  - The difficulty of the project (from both data and visualization perspective)
- **Predictive Modelling (30%)**: The quality and originality of the ideas and the extent of analyses
  - Quality and effort of predictive analyses
  - You need to include at least one linear regression model
  - You are encouraged to try out other data mining models
- **Business Insights (20%):** Depth and applicability of business insights derived from the analysis
  - The evidence to support
  - The appropriateness of the conclusions and business insights
- **Report and Presentation (10%):** Clarity, professionalism, and completeness of the report and presentation.
  - The quality of the presentation
  - The quality of the written report (organization, effectiveness of the communication)
  - Time management of the presentation

3. Project Requirements

This group project is a great opportunity for you to experience data analysis and show off your python skills. Go and hunt some interesting datasets from some business scenario and extract valuable information and good insights out of it.

Some external examples can be found at:

MSIS 502
Group Project

https://python-public-policy.afeld.me/en/nyu/final_project/examples.html

The primary objective of the group project is to provide students an opportunity to explore and think about potential applications of the techniques they learn in this class in a real-world business environment.

Data is everywhere and I encourage you to select a topic that is of interest to you! Topics could include sports, healthcare, GDP, Beatles' songs, airline delays, etc. It you have an interesting problem at work that can potentially be answered with data, that may make a great project. Your team will identify a business or an industry which is benefiting or can potentially benefit from data exploration, apply in-depth analyses, and write up a report.

I listed useful sources of datasets below but you don't need to restrict your search to them.

- Useful sources of datasets
  - https://www.kaggle.com/
  - https://public.tableau.com/s/resources?qt-overview_resources=1
  - http://archive.ics.uci.edu/ml/index.php
  - https://www.data.gov/
  - https://datacatalog.worldbank.org/

4. Project Contents

You should follow the steps in data mining processes, and you are recommended to carry out (but not limited to) the following analyses.

A. Develop an understanding of your data and the purpose of the data analytics project.

B. Data exploration and visualization.
   You need to use Python libraries to showcase a variety of data exploration and visualization techniques learned in this course. For example, you can use scatter plots to select variables that might be useful in predicting your target. Boxplot can be used to identify outliers and you can study whether including the outliers in your data will lead to biased conclusions. You are also encouraged to use AI-tools to go beyond the knowledge learnt in class.

C. Data cleaning and preprocessing. For example, if your dataset is large, you can random sample a subset for training your model. How should missing data be handled? How should outliers be handled? For data cleaning and preprocessing, please feel free to use AI tools.

D. Choose the appropriate business goals that can be approached through multiple learning regression. Evaluate your model. If the error of your model is high, can you try other models and discover possible reasons for the poor performance of your models?

E. Interpret the results and provide business insights. Connect the insights to specific business questions or problems (e.g., customer segmentation, sales trends, operational efficiency). Will the data be useful to improve the business objectives of the company? How will you communicate the results to the management?

5. Deliverables

Note that a large dataset is usually hard to deal with computationally. Yet, if a dataset is too small, there is really not much to mine from it. A dataset that is around 50KB ~ 500MB is a good choice for gaining useful insights without getting into computational troubles.

A. On **Aug. 11**, your group needs to present to the class the information and insights that you have extracted from your datasets. Presentation order will be randomly assigned and will be announced at the beginning of the class meeting. **All group members are required to be involved in the presentation.** The presentation should last around **13 minutes with additional 2 minutes of Q&A**.

For all groups, **presentation file** and a **pdf report** are due on **Aug. 11 before class**.

B. In your presentation, you should describe the highlights on the project. The **final report** should be in 12-point font, double-spaced, and between 5-15 pages in length including all appendices and exhibits**. You are not required to submit your python codes**. Each project is different, but the suggested contents of the report and presentation are (but not limited to)

- The goal of the project
- Description of your dataset
- Data visualization and any exploratory analysis you did
- Predictive modeling
- The results and your insights. You should support your results and conclusions with exhibits as needed
  - Who can benefit from the data?
  - How would the data help them to make better decisions?
  - What other data would be useful to have?
  - What are your key insights and recommendations?
- A discussion about what you would do differently or how the project could be improved in the future
- Any special obstacles that you overcame
- Reflections on how AI tools assisted you in the group project. You can mention specific questions you asked and how the responses helped you.

C. In order to avoid free-riding, there will be a **peer evaluation** for this group project.

Again, it's your chance to show off your Python and data analytical skills. Be CREATIVE in finding and mining data!