ISOM 5160 – Python Programming for Business Applications
2025-26 Fall Term Group Project

General Information

- This is a group project, with **5-6** members per group

- The project accounts for **30%** of the course final mark

- Task: leverage Python coding to conduct exploratory data analysis on a real business dataset and present the findings

- Deadlines:

  - Self-sign in the groups on Canvas by **Sep19 (Tue)**
  - Email the group topic or the link of your dataset to TA (cc the Instructor) by **Oct 3 (Tue)**
  - Demo and present in class on Oct 13 (L1) / Oct 14 (L2)
  - Submit the required files onto Canvas by Oct 18 (Sat)
  - Complete peer review by Oct 18 (Sat)

Requirements

- Each group may choose a dataset from the list below. The files can be downloaded from Canvas, under Group Project page in Assignment Module.

	File Name	No. of Rows	No. of Columns
1	amazon_food_review.csv	10,828	10
2	bank_marketing_response.csv	32,590	21
3	bigcities_healthdata.csv	13,512	11
4	customer_churn.csv	10,000	13
5	employee retention.csv	14,999	10
6	netflix_shows.csv	7,788	12
7	online_shoppers_intention.csv	12,330	18
8	ted_talks.csv	2,550	17
9	udemy_courses.csv	3,678	12
10	youtube_trending_videos.csv	9,600	16

- In none of the above datasets interests your group, you can search and download a dataset online. However, the dataset must satisfy the following requirements:
  - Contains at least 5,000 rows
  - Contains at least 8 columns with meaningful column names
  - Must be from a real business scenario (not synthetic data)

- Each group has **15 minutes** to present the Python code and analysis findings. All members must participate in the presentation. After the presentation, the group may answer 1-2 questions from the instructor.

- After the presentation, one group member needs to submit **one single .zip file** onto Canvas, including a **. py** or **.jpynb** file for the Python code, a **.pdf** file for the presentation slides, and a **.csv** file for the dataset. At the beginning of the code file, team members must clearly document their individual contributions by specifying which sections of the code they developed (e.g., "Member A: data cleaning and preprocessing," "Member B: visualization implementation").

- All group members must complete the peer review online (*details will be announced on Canvas). Penalty may be imposed to those who do not complete the peer reviews.

Marking Scheme

- Coding (20%)
    - Data cleaning (e.g., missing value handling)
    - Data transformation (e.g., transforming format, normalization, encoding, etc.)
    - Data summarization (e.g., primary statistics, pivot_table, groupby, merge, etc.)
    - Exploratory analysis (e.g., visualization, correlation, outlier, moving trend, etc.)
    - Each member may be assigned with different individual score according to the code quality (correctness, effectiveness, readability, etc.)

- Presentation (5%)
    - Steps taken for the analysis
    - Findings of the analysis results
    - Potential business value of the findings
    - Future work that can further be done using Python
    - All members of the same group will be assigned with the same presentation score

- Peer Evaluation (5%)
    - Each member may with assigned with different individual score based on the reviews

Data Set Sources

- Google Cloud Public Datasets
  https://cloud.google.com/solutions/datasets

- Kaggle: Your Home for Data Science (registration needed)
  https://www.kaggle.com/

- World Bank Open Data
  http://data.worldbank.org/

- Datasets for Data Mining and Data Science
  http://www.kdnuggets.com/datasets/index.html

- UCI Machine Learning Repository
  https://archive.ics.uci.edu/datasets

- data.world Community (registration needed)
  https://data.world/community/open-community/