Course Project

The course project is a team-based work that aims to offer you an experience to discover collaborators, cooperate with your colleagues, and apply the tools and concepts you have learned in the course to a practical data-driven decision.

Each team should find a data-driven decision setting where you can get some data and obtain some insights by exploring the data. Here are some example projects from prior class. You may choose some topics similar to (but not restricted to) one of the following:

- Creating a Diversified Stock Portfolio Using Clustering Analysis
  Data from Kaggle: S&P500 Index, S&P500 Companies, and S&P500 Stocks
- Detecting the need for ICU admission among COVID-19 patients. Data available from Kaggle:
  https://www.kaggle.com/S%C3%ADrio-Libanes/covid19
- Use the inventory and sales data to improve the inventory turnover in Drug store (private data).
- Predict flight delays and cancellations. Data available from Kaggle:
  https://www.kaggle.com/usdot/flight-delays
- Predicting Texas automobile accidents. Data available from Kaggle:
  https://www.kaggle.com/sobhanmoosavi/us-accidents
- HR analytics: job change of data scientists. Data available from Kaggle:
  https://www.kaggle.com/arashnic/hr-analytics-job-change-of-data-scientists
- Analysis of H1B visa approvals. Data available from Department of Labor:
  https://www.dol.gov/agencies/eta/foreign-labor/performance
- Airbnb - predicting country of travel. Data available from Kaggle:
  https://www.kaggle.com/c/airbnb-recruiting-new-user-bookings
- Predicting loan default. Data from Kaggle: https://www.kaggle.com/c/loan-default-prediction
- Predicting the presence of heart disease in patients. Data from Kaggle:
  https://www.kaggle.com/ronitf/heart-disease-uci

You and your team will explore the data using R. You may find your own data source or reuse some of the datasets linked above in your project. I also provide a list of open databases as your reference. Please get in touch with me if your team is having difficulty coming up with a project. We can coordinate the time over email/MS Teams chat.

Deliverables:

1. Project Proposal (Due 3/24 at 11:59 pm). The proposal WILL NOT be graded. It mainly serves as a milestone for your group project. I will evaluate the feasibility of the project proposal and contact your group should there be any questions.

2. Final Project (Due 4/21 at 11:59 pm). The final project WILL be graded. Each group shall submit an R script, the slides for presentation, and a report. The report should address the following:
   - Setting – explain the business context and the problem.
   - Data – briefly explain the data that you used, reference to the data source, and the challenges that you have encountered while working with the data.
   - Analysis and discussion – summarize your analyses, findings, and implications.

List of open databases

- https://www.data.gov/
- https://data.gov.uk/
- https://data.europa.eu/euodp/en/data/
- https://datahub.io/dataset
- https://docs.aws.amazon.com/data-exchange/
- https://www.tableau.com/learn/articles/free-public-data-sets
- https://github.com/awesomedata/awesome-public-datasets
- https://www.kaggle.com/datasets
- https://data.cms.gov/