IS 451 Group Project
Due March 9 at the beginning of the class
Total: 20 points

1. Project Overview

The goal of the project is provide realistic experience in data analysis. Data analytics are best learned by **doing**. Based on their collective work experience, each group should identify, and to the extent possible, execute a business intelligence project that relies on the data mining techniques we covered in the class.

2. Evaluation

You will be graded on your ability to demonstrate your grasp of basic data mining concepts and how they can be applied in a business context queries. The projects will be evaluated on:

- The difficulty of the project (from both data and methodological perspective)
- The relevance of the topic
- The quality and originality of the ideas and the extent of analyses presented
- The evidence to support
- The appropriateness of the conclusions and business insights
- The quality of the presentation
- The quality of the written report (organization, effectiveness of the communication)

3. Project Requirements

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

You can find many (external) **examples** of the final project at:
http://www.galitshmueli.com/student-projects

There are two options you can choose to work on.

Option 1

Choose **one** of the following two real world cases which are described in details in Chapter 21 of the textbook. The datasets are provided on Canvas.

Case	**Description**	**Data**
Charles Book Club	Page 499 - 505	CharlesBookClub.csv
Predicting Bankruptcy	Page 525 - 527	Bankruptcy.csv

Your team will use the data to conduct in-depth analyses, provide detailed interpretations of the results, and make useful suggestions. It is advised that you should use these cases as the business contexts, and go beyond what have been suggested in the case description such as "assignments."

Option 2

The primary objective of this option 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. If 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 mining applications or methods, 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.

- Good Overview of Potential Sources
  https://visual.ly/blog/data-sources/

- 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

No matter whether you choose **Option 1** or **Option 2**, you should follow the steps in data mining processes, and you are recommended to carry out the following (but not limited to) analyses.

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

B. Data exploration and visualization.
You may want to use 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.

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?

D. Choose the data mining model to be used and use R to implement it. 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. Will the data be useful to improve the business objectives of the company? How will you communicate the results to the management? You are also encouraged to visualize the extracted information (e.g., graphs, charts). For example, a histogram of the RMSE of a multiple linear regression.

5. Deliverables

A. Each team should submit a project proposal before **11:59pm, Feb 26**, which includes the choice of project option, a description of the data, the size of the data file, the number of records it contained and the number of variables, and a brief summary of your objective.

Note that a large dataset is usually hard to deal with computationally. A dataset that is around 100KB ~ 500MB is a good choice for gaining useful insights without getting into computational troubles.

B. On **March 9** and **March 11**, your group needs to present to the class the information and insights that you have extracted from your datasets. Presentation date and order will be randomly assigned and will be announced later. The presentation should last **10 minutes with additional 2 minutes of Q&A**.

For all groups, **presentation file** and a **pdf report** is due on **March 9 before class**.
On your presentation day, make two hardcopies of your presentation for me and your TA.

C. 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 don't need to submit your R 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
- Data modeling, and the methodology
- 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?
- A discussion about what you would do differently or how the project could be improved in the future
- Any special obstacles that you overcame

D. 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 data analytical skills. Be CREATIVE in finding and mining data!