CSIT 356/556 Introduction to Data Science

Final Project

Instructions: In the project, you need to prepare an idea and a data set from real world. Convert them to Pandas and apply multiple techniques for data analysis.

Group work: Both individual and group work are allowed in this project. Each group can include at most 3 students. All the names of group members should be indicated in the project design report.

About the data set:

You could find the data by your self or select from the following resources:

Stanford Large Network Dataset Collection	https://snap.stanford.edu/data/
Dataverse Network	https://dataverse.org/
Reddit Open Data	https://www.reddit.com/r/opendata/
CDC Data	https://www.cdc.gov/nchs/tools/index.htm?CDC_AA_refVal=https%3A%2F%2Fwww.cdc.gov%2Fnchs%2Fdata_access%2Fdata_tools.htm
World Bank Catalog	https://datacatalog.worldbank.org/
Metor Boston Data Common	https://datacommon.mapc.org/
COVID-19 Data Repository by Johns Hopkins University	https://github.com/CSSEGISandData/COVID-19

Don'ts

- Don't use a standard machine learning dataset (Kaggle, UCI ML Repository). These are pre-processed and only suitable for analysis, not for the whole DS process
- Don't pick a dataset where structured data is hard to extract, E.g.,
  - text-only, relying on advanced NLP,
  - extracting data from collection of PDFs,
  - running your own survey (it's hard to run a good survey)

Project Milestone

Project Progress	Deadline
Announce your team and project title	Friday, Oct. 27

Submit your project proposal	Friday, Nov. 3
Get feedback from instructor	Friday, Nov. 10
Submit your final project	Tuesday, Dec. 5
Project Presentation	Thursday, Dec. 7 and Monday, Dec. 11

Project Requirements

The project MUST includes the following techniques:

- Data Loading
- Data Cleaning
- Data Analysis using Descriptive Statistics
- Data Visualization

At least one technique in data manipulating from:

- Data Wrangling
- Data Aggregation
- Time Series

Final Project Submission

A final submission should include project report, all the source code, data set and slides for the presentation.

CSIT 356/556: Introduction to Data Science

Rubric of the Final Project

Project Title: _____

Student Names: _____

	Points out of Total
1. Submission of the title and team members with	_____/ 2
a. Title of the project (1 points) _____	
b. Names of all the team members (1 points) _____	
2. Project proposal submitted on time with	_____/ 8
a. Basic information (1 points) _____	
b. Project objectives (1 points) _____	
c. Description of the data set (2 points) _____	
d. Analysis methodology (2 points) _____	
e. Project schedule (2 points) _____	
3. Final Project	_____/ 30
a. Data Loading (6 points)	_____
b. Data Cleaning (6 points)	_____
c. Data Analysis using descriptive statistics (6 points)	_____
d. Data Visualization to show the results (6 points)	_____
e. One technique in data manipulating (6 points)	_____
4. Project Presentation	_____/ 10
a. Presenters are well-prepared (2 points)	_____
b. Slides should present material in an informative manner (2 points)	_____
c. Presentation is logically organized and presenters appear to be fluid (2 points)	_____
d. There is a balance between high-level motivational material & technical detail (2 points)	_____
e. Presenters should respond well to questions and critique (2 points)	_____
Total Score	_____/ 50

Graders Comments:

CSIT 356/556: Introduction Data Science

Template of the project proposal

Project Title:

Team Members: *List the names of all the members*

Basic Information

Provide the basic information about the motivation of the project.

Project Objectives

Provide the primary questions you are trying to answer in your project.

Description of the Data Set

Provide a brief description of the data set selected.

Analysis Methodology

Provide the basic methodology and what data science techniques are planned to be applied.

Project Schedule

Provide the tasks assigned to each team member and the deadline for each milestone.

CSIT 356/556: Introduction to Data Science

Template of the Project Report

Project Title:

Team Members: *List the names of all the members*

---

1. Description

1.1 Basic Information

Provide the basic information about the motivation of the project.

1.2 Project Objectives

Provide the primary questions you are trying to answer in your project.

1.3 Description of the Data Set

Provide a brief description of the data set selected.

2. Exploration of Data Analysis

2.1 Data Preparation

Describe the process to load and clean the data, especially how to select the necessary data. List the key codes if necessary.

2.2 Data Analysis using Descriptive Statistics

Describe the measurements of descriptive statistics to answer the questions in your objectives. List the key codes if necessary.

2.3 Other Techniques

Describe all the other techniques of data manipulating applied in the project. List the key codes if necessary.

3. Data Visualization

Describe how to visualize the results. List the plots and key codes if necessary.

4. Conclusion

Based on the test results, describe your observation.