DS 3000 – Final Project Deliverables

The final project is intended for you to become an independent data scientist and work on a full DS project from conceptualization to reporting. The project is designed to give you hands-on experience applying the skills from this class to a life-like DS project in which you identify a problem and dataset; clean, process, and prepare the dataset for analysis; analyze and visualize the data using Python libraries; and report your findings and insights for decision making. You will work in groups of 2 or 3 students throughout the semester and submit multiple deliverables as specified in **Course Schedule**. The project will culminate in a DS report and project presentation. Before starting to work on the project, you will need to participate in online discussions to identify potential teammates based on common interests.

1. FP1: Topic Proposals

The first deliverable of the project is a topic proposal for your DS project. You should start by thinking about a real-world problem that interests you and also has clear connections to an area to which data science can contribute. Each group will submit 2-3 topic proposals (depending on whether your group consists of two or three members), and your TAs will help evaluate which topic seems most promising for this project.

Here are some ideas for the project. You are by no means limited to these ideas and may propose anything else that would be similar to these examples in scope (requiring approximately 30-60 hours of work).

- ✓ Collect visual data from medical diagnostic tools (MRI, CT Scans, ultrasound, etc.) store image files in a file format, and retrieve the information in a manner useful for physician/patient. Retrieve data for descriptive and predictive analysis.

- ✓ Retrieve data from social media web APIs, e.g. Twitter, Facebook, Instagram, etc., clean the data, and create an alternative data schema for a particular file format and retrieve data for descriptive and predictive analysis.

- ✓ Collect publicly available economic data via web APIs offered by Yahoo Finance, US Bureau of Economic Analysis, and Quandl Resource Hub, etc., filter it for a particular industry/task, store in a file, and retrieve data for descriptive and predictive analysis.

- ✓ Collect publicly available data produced by manufacturing and production systems with the aim of increasing efficiency for a certain manufacturing process. Try Amazon Public Datasets, DBpedia, or World Bank for dataset sources.

A well-designed project is one that takes a relatively complex dataset from some source (or scrapes it), cleans it, stores it in a file, and retrieves the data from the file. Your project should provide a description of the dataset both statistically and visually. All visualizations must be interpreted. **At least one machine learning/modeling technique needs to be applied to the data for a classification or a prediction**. A collection of models needs to be created using the technique as well as an evaluation of the models.

For each of the proposed ideas, the deliverable should include a description of the problem you would like to tackle, why you think it matters, and whether you have identified any potential datasets that might be useful for the project. A template will be made available on Canvas.

2. FP2: Dataset

After your proposal is confirmed, you will identify your dataset (if you haven't already) and retrieve your data, which may involve you downloading a data file or (even better) scraping your own dataset from an online source. As part of this deliverable, you will submit your dataset along with a Jupyter Notebook.

3. FP3: Data Analysis Plan

Each group will write a short summary (approximately three/four paragraphs) of their DS project, including a brief introduction to the problem and its significance, a statement of research questions/hypotheses and a description of the dataset and data analysis plan (i.e., the techniques you are planning to use to answer your research questions and/or test your hypotheses). Groups will then share this summary with classmates (on Canvas) and provide each other with constructive feedback.

4. FP4: DS Report – Jupyter Notebook

The end-product of the project will be a DS report in the form of a Jupyter Notebook. In this report, you will include the main sections of a general DS report, i.e., executive summary, introduction, method, results, and discussion, as described below:

- ✓ The introduction section will orient the reader to your DS project, providing sufficient background information on your project, the problem you attempted to tackle, and the dataset you analyzed.
- ✓ The method section will include a detailed description of your dataset, variables, models, data analysis techniques, and procedures.
- ✓ The results section will present the results of your data analysis, along with the corresponding code and visualizations in your Jupyter Notebook.
- ✓ The discussion section will include a thorough interpretation of your results with respect to your research questions and/or hypotheses and a discussion of insights you gained from the data and future work.

A template Jupyter Notebook for your DS report will be provided on Canvas.

5. FP5: Project Presentation

The final deliverable of the project will be a project presentation (video recording) in which you will describe your project and explain your overall data analysis and interpretation process.

6. FP6: Peer Evaluation

Because the final project involves you working in groups, you will submit peer evaluations of your team members' contribution to the project throughout the semester. Given the scope of the project, peer evaluations will have a substantial impact on your grade for the final project; your final grade will be determined based on the peer-evaluation (up to 50% of the total final project grade could be docked on the basis of peer evaluations).

If you repeatedly fail to make substantial contributions to the deliverables of the final project, your team members may decide to kick you of the team. In this case, your final project grade will immediately evaluate to zero, which usually means you automatically failing the class. Simply, don't do this to yourself! It is not fun for anyone (trust me). If you are having some team dynamics issues, please do **NOT** let them escalate to this level of seriousness. Instead, inform me in a timely manner.