DAV 5400 Final Project Guidelines

***** You may work in small groups of no more than three (3) people to complete the Final Project *****

Throughout this course we explore a wide variety of data acquisition, data management, and data manipulation methods. An essential component of analytics and machine learning work is knowing when and how to apply a particular type of methodology, algorithm, or software tool when presented with a new challenge. For your **Final Project** of this course you are responsible for selecting your own data sources and defining your own research question(s). Once you've identified the data sources you are interested in working with, you should then define one or more research questions you will attempt to answer as part of your work for the Final Project. The data you select **must** serve to answer one or more formal research questions that you define for purposes of framing your **Final Project** work. This **Final Project** represents **20% of your final grade for the course**.

Your **Final Project** is comprised of three separate deliverables:

1. A formal Final Project Proposal;
2. Your Final Project writeup + Python code (in the form of a Jupyter Notebook);
3. A "live" presentation of your work during our final Live Session (Module 15).

A summary of the schedule and scoring for these deliverables is provided below.

Deliverables Schedule

Deliverable	***Date***	***Points***
Proposal	First Draft: Module 9; Final Draft: Module 11	**25**
Final Project	Module 15	**125**
Final Project Presentation	During Final Live Session (Module 15)	**50**

Final Project Checklist

To receive full credit for the Final Project, you will need to deliver on all of the items mentioned in the checklist shown below. **Please read carefully through this checklist before you make your project proposal**. You are (within these checklist constraints) strongly urged to limit scope and make the necessary simplifying assumptions so that you can deliver your work on time.

- ☐ Proposal describes your motivation for performing this analysis.
- ☐ Proposal describes from where you plan to source your data.
- ☐ Your project has a recognizable and reproducible "data science workflow." [Example: First the data is acquired and explored, then necessary transformations and clean-up are performed, then the analysis and presentation work is performed]
- ☐ Project includes data from at least ***two*** different *types* of data sources (e.g., two or more of these: (1) Downloadable files (e.g., CSV format), (2) scraped web page, (3) web API. ***AT LEAST ONE OF YOUR SOURCES MUST BE a SCRAPED WEB PAGE or WEB API. Relying solely on easily downloadable CSV files is NOT ACCEPTABLE*** for the Final Project.

- ☐ You are ***NOT*** allowed to use any dataset that that been provided to you for any DAV 5400 Assignment or Project. You are also ***NOT*** allowed to use any data set that is embedded within any Python library (e.g., any of the scikit-learn data sets).
- ☐ Project includes statistical analysis and graphics that describe and/or validate your data (e.g., EDA).
- ☐ Project includes at least one data reshaping operation. [Examples: transforming from wide to long; converting columns to date format]
- ☐ Project includes the use of data-appropriate data preparation and feature engineering methods.
- ☐ Project includes at least one grouping or aggregation.
- ☐ Project includes at least one graphic that supports your conclusion(s).
- ☐ Project includes at least one statistical analysis that supports your conclusion(s).
- ☐ Project includes at least one Python feature that we did not cover in class, e.g., something you discovered during your coursework or that you found to be necessary for completing your research.
- ☐ Presentation. Was the presentation delivered in the allotted time (8 to 10 minutes)?
- ☐ Presentation. Did you show (at least) one challenge you encountered in code and/or data, and what you did when you encountered that challenge? If you didn't encounter any challenges, your assignment was clearly too easy for you!
- ☐ Presentation. Did the audience come away with a clear understanding of your motivation for undertaking the project?
- ☐ Presentation. Did the audience come away with a clear understanding of at least one insight you gained or conclusion you reached or hypothesis you "confirmed" (rejected or failed to reject…)?
- ☐ Code and data. Have you delivered the submitted code and data where it is reproducible and self-contained—preferably in a Jupyter Notebook on GitHub? Can your results be fully reproduced with what you've delivered? You won't receive full credit if your code references data on your local machine.
- ☐ Code and data. Does all of the delivered code run without errors?
- ☐ Deadline management. Were your draft project proposal, project, and presentation delivered on time? Please turn in your work on time! You are of course welcome to deliver ahead of schedule.

Policy on Collaboration

You may work in teams of up to three people to complete the **Final Project**. Each project team member is responsible for understanding and being able to explain *all* of the submitted project work + Python code. Remember that you can take work that you find elsewhere as a base to build on, but you need to acknowledge the source, so that your grade is based upon what you actually contribute, ***not*** on what you start with.

Proposal Guidelines

Your first deliverable for this Project (25 Points) is the **Final Project Proposal**. The Proposal for the Final Project will be submitted in the form of a formal research proposal document. . Furthermore, you need to ensure that the Project you are proposing will satisfy all of the requirements specified in the **Final Project**

Checklist (see below). Your proposal must include each of the sections outlined below and must be submitted in the form of a Jupyter Notebook.

Introduction (5 Points)

This section should provide some context for the basis of the research questions you plan to answer without actually describing the research questions themselves. For example, if your research questions are focused on a health related issue, you might provide a brief summary of how many people are affected by that issue each year either regionally, nationally or globally, including any infection rate or mortality statistics you were able to gather. Basically, in the Introduction you are trying to make the reader understand why the research questions you are going to propose are relevant and should be of interest to them.

Research Questions (6 Points)

Provide a single succinct sentence describing each of your research questions. Then provide a paragraph or two explaining how the results of your research might be used/implemented in the "real world".

Data to be Used (4 Points)

Clearly identify the sources of your data and explain the methods you will use to collect the data from those sources, e.g., "Data will need to be collected from this source via scraping of a web page..", etc.

Approach (10 Points)

Explain how you plan to manage the data you are collecting: e.g., will you be storing it within some sort of database, etc.? Explain what types of statistical analysis you plan to utilize to help answer your research questions. Explain any graphics you plan to generate to help answer your research questions. The reader should come away with a clear understanding of how you plan to proceed with your work. Keep in mind that since this is a proposal, you need to be able to convince the reader that your proposed project is: a) realistic; and b) feasible within the time allotted for the project. Also, be sure to clearly articulate the roles and responsibilities of each team member for your Project work.

Your Final Project Proposal Jupyter Notebook deliverable should be similar to that of a publication-quality / professional caliber document and should include clearly labeled graphics, high-quality formatting, clearly defined section and sub-section headers, and be free of spelling and grammar errors.

Upload / submit your Jupyter Notebook within the provided Final Project Proposal Canvas submission portal. Be sure to save your Notebook using the following nomenclature: **first initial_last name_FinalProjectProposal**" (e.g., J_Smith_FinalProjectProposal). ***Small groups should identity all group members at the start of the Jupyter Notebook and each team member MUST submit their own copy of the team's work within Canvas.***

Proposal Approval

Once you've submitted your proposal its content will be reviewed for purposes of determining whether or not what you have proposed is acceptable as a Final Project for this course. If so, you will be conditionally approved to start work on your Final Project. If not, you will receive detailed feedback regarding any issues that need to be addressed before you can receive approval for your Project. You will be able to re-submit your Proposal as many times as necessary to achieve the required approval. Once you receive the conditional approval you will have earned the full 25 points possible for the Proposal component of the Final Project (assuming you had originally submitted the Proposal no later than the due date specified in Module 11).

<u>Your Second deliverable for this Project (125 Points)</u> is your Final Project Jupyter Notebook. It should contain a combination of Python code cells and explanatory narratives contained within **properly formatted Markdown cells**. The Notebook should contain (at a minimum) the following sections (including the relevant Python code for each section):

1) **Abstract (10 Points):** Use 250 words or less to summarize your problem, methodology, and major outcomes.

2) **Introduction (10 Points):** Describe your project, including the scientific or business motivation for the research question you have chosen to answer. This section should summarize the content of your Final Project Proposal, so be sure to explain your research question, describe the source and content of the data set you have chosen to work with, and summarize your approach to meeting the requirements for the Project.

3) **Research Approach (10 Points):** Explain + present the end-to-end methodology you made use of for all aspects of your Final Project work, including your EDA, data preparation, and investigative analysis work. Be sure to include a description of your data management strategy as part of your narrative.

4) **Exploratory Data Analysis (25 Points)**: Explain + present your EDA work including any conclusions you draw from your analysis, including any preliminary predictive inferences. This section should include any Python code used for the EDA.

5) **Data Preparation (15 Points)**: Describe + show the steps you have taken to address the data integrity + usability issues you identified in your EDA, including any feature engineering techniques you have applied to the data set. This section should include any Python code used for Data Preparation.

6) **Prepped Data Review (5 Points)**: Explain + present your post-Data Prep EDA analysis. This section should include any Python code used for re-running your EDA on the variables adjusted during your Data Preparation work.

7) Investigative Analysis & Results (40 Points): Explain + present your investigative analysis work, including any Python code used as part of that process. Provide and explain your answers to your research questions.

8) Conclusions (10 Points): Summarize your work and clearly state the conclusions of your research. Were you able to answer the research questions you originally posed in your Proposal? Comment on any potential future extensions of the work you've completed for the Project.

Your Jupyter Notebook deliverable should be similar to that of a publication-quality / professional caliber document and should include clearly labeled graphics, high-quality formatting, clearly defined section and sub-section headers, and be free of spelling and grammar errors. Furthermore, your Pythion code should include succinct explanatory comments.

Upload / submit your Jupyter Notebook within the provided Final Project Canvas submission portal. Be sure to save your Notebook using the following nomenclature: **first initial_last name_FInalProject**" (e.g., J_Smith_FinalProject). ***Small groups should identity all group members at the start of the Jupyter Notebook and each team member <u>MUST</u> submit their own copy of the team's work within Canvas.***

<u>Your third deliverable for this Project</u> (50 Points) is an approximately 8-10 minute live presentation of your work. Your presentation should include a brief overview of your research questions and the data you selected to work with, your EDA work, a high-level explanation of your data preparation + feature engineering

process, a discussion of your approach to answering your research questions, the results of your research, and your conclusory statements.