Applied Data Science – Final Group Project

Class Project

Each student will join a group that will undertake a data science project in an area of interest. The project should provide a comprehensive investigation of some issue relevant to the field of study.

Some example topics could include any of the following:

- How do the prices of airline tickets change as you get closer to the date of the flight?
- Are surgeries more successful on weekdays versus weekends?
- Which factors influence the volume of inventory for retail products?

Your team may select any reasonable question and source of data to study. The examples that are presented in class are all worthy of additional exploration, or you may choose a topic that is interesting to you. The teaching staff will be available to help you think about these issues and the approach.

The focus of the project can range from exploring different methods of analysis to building innovative applications. But, overall the project must answer some questions about what is contained in the data and how this information can be applied in a real setting.

Data Sets

- Your team is welcome to select a data set of its choice.
- The complexity and dimension of the data can vary. You must be able to work with the information provided, and the data does not have to be enormous to be effective. However, the information should be of a sufficient quality to address the questions you'd like to answer.
- However, it is important that the data you choose should be real in some meaningful sense. There should be some practical challenges to investigate. Ideally your project would have a degree of uniqueness to it. For this reason, it is recommended that you consider data sources apart from the typical data sets often used in data science courses and competitions. These issues can be discussed in greater detail during the checkpoints.

Working in Groups

- Each group will include approximately 4 members.
- The groups will be randomly assigned. This ensures a greater degree of fairness in the work and grading. It's also more reflective of projects outside of the classroom, in which teams often have to coalesce without prior working relationships.
- Each team may decide how to organize its efforts. However, each member of the team must take on a reasonable degree of responsibility.

- For each section of the report, the team will report on the contributions of each member in terms of percentages. The team will collectively fill out a Division of Responsibilities form that shows how the work was conducted.
- Each team member must have the primary responsibility (greater than 50%) for at least 2 sections of the final report.

The Final Presentation

Each team will prepare a final presentation to be delivered at the end of the semester. These presentations will take place during the last two classes of the semester, with the schedule to be randomly assigned.

Each presentation should be approximately 10 minutes in length. It should highlight:

- The problem you are trying to solve.
- The relevant data.
- An overview of your analysis.
- The results.
- A demonstration of the application you built for the project.

The slides for the report should be **written in RMarkdown** in a reproducible format.

The Final Report

Your report should carefully explain the nature of the work you performed. This must include the following sections:

- **Introduction**: Which problem are you trying to solve? Why is it important to the field you are working in?
- **Sources of Data**: Where did you obtain the information that you made use of? How do you know it is reliable, accurate, and representative of the population you are studying?
- **Examination of the Data**: How did you check the quality of the data? Were there any problems that you found? Did you restructure the data, and if so, why was that necessary?
- **Your Investigation**: What work did you do to investigate your questions? What did you build? Which methods did you use?
- **The Results**: What answers did you obtain? Provide any relevant tables or visualizations that can help to explain what you found.
- **Interpretation**: What conclusions can you draw from the results? How should your findings impact the overall work in the field? Will it change how decisions are made, products are designed, or services are provided?
- **Assumptions**: Which judgments and assumptions did you make to produce your results? Why are these assumptions reasonable?
- **Limitations and Uncertainties**: What are some reasons that might suggest caution in how your results are used? What are the limitations and uncertainties?

- **Areas of Future Investigation**: Now that you have some results, what else can you do to make additional contributions?
- **References**: Which background materials informed your work? Please provide a list of references to any resources that gave you information about the field of study, the methods you used, or the technical problems you solved.

The final report should be in the ballpark of 3000-5000 words. This will not be strictly enforced, but please don't write a report that is much too long or much too short.

Many projects lead to working on answering a wide variety of questions, some of which may be loosely connected to the broader themes. For this project, it is better to provide a high-quality investigation of a smaller number of questions rather than trying to include every fact or figure.

Additionally, the focus of the report should be at a high level. It should be easy for an interested person without specific experience in data science to read and understand your work. While it is important to explain the choice of the methods, the report should not go too far into the details of the more technical steps (such as data cleaning). Instead, you are encouraged to focus on **telling a story** about the problem you are trying to solve, the investigation, and how the results can help us better understand the field. The tables and figures should be readable and succinct. Long sections of coding output may be useful for your internal understanding, but please ensure that the inclusion of code and output in the report adds to the narrative.

Building an Application

Your team will also be required to build some kind of application related to the project. The application must be built using R (e.g. with the shiny or flexdashboard package). The application can focus on data visualizations, summarization of the data, segmentation of the relevant information over time or in subgroups, or otherwise show relevant information. The application must be interactive in some way, which will enable the user to make selections about the content to view and dynamically generate the results.

Submission Requirements

Each final project will include the following components:

- A copy of or link to your data.
- All relevant R code used in your work.
- A written report of 8-12 pages (approximately 3000-5000 words). The original RMarkdown file must be provided along with output in HTML, PDF, or Word format.
- The R code and associated files for the application.
- Slides for the Final Presentation, along with the source code used to generate them.
- A **Division of Responsibilities** form. For each section of the report, the team will state in percentage terms how each member contributed to the development of the section. Each team member must have the primary responsibility (with a stake greater than 50%) of at least 2 sections.
- A **Summary of Collaboration** form. Each student will separately provide a short report on the team's collaboration and the work of each other team member. This form will be kept confidential.

Supplementary materials may be considered but are not necessary. These may include:

- Short videos

- Data dictionaries
- Other relevant material.

Reproducibility

The final reports, presentations, and applications must be written in a **reproducible** format. Each group should write this content using RMarkdown and submit this code along with its output (a file in .docx, .pdf, or .html format). As much as possible, the numbers, tables, and plots provided in the report should be connected to calculations within the source code. All input files (data, code, documentation) should be provided in a directory structure that makes the code easy to run from any machine. In short, those reading your report should be able to run your code and reproduce the results.

Submissions and Checkpoints

- **Checkpoint 1**: Each team must have completed at least one introductory meeting. The team must select a unique name for itself. The team should provide some information on some potential ideas for topics and sources of data to explore.

- **Checkpoint 2**. All groups must provide a progress report. Please include the following details:

    - Finalized Topic for the Project
    - A copy of or link to the data you plan to use.
    - A short description (one paragraph) of the data.
    - A brief overview (2-3 paragraphs) of your plans.
    - A summary (2-3 paragraphs) of your progress so far.

- **Checkpoint 3**. All groups must provide an additional progress report. In this case, we aim to see some preliminary results and discuss steps to improve the overall quality of the report. Please include the following details:

    - Topic for the Project
    - A copy of or link to the data you are using.
    - A short description (one paragraph) of the current state of the project.
    - A brief overview (2-3 paragraphs) of the most important models, including some details about the variables and code used so far.
    - A short list of questions about how to improve the project in the remaining time.
- **Final Presentation**: Your team must send a copy of its slideshow, application, and any files needed to run the application.

- **Final Report** All groups must provide a final report. This must include all of the files for the application and written report. **No late projects will be accepted.**