CSC 401, Final Project, Fall 2024 Name _____

The final project is intentionally quite open-ended and flexible to be adapted to the students' individual interests. The guidelines and examples below will provide some guidance, and a list of the components you are required to submit, but if in doubt, consult with your instructor.

Description: The final project for this course is designed to give you some experience working with algorithms and applying them to real world data. There are two required elements to this project:

Element #1: you are required to write the code, in R, for an algorithm we have not directly covered in class.

Algorithms we are not coding from scratch in this course already include (but are not limited to):

- LOESS regression
- Smoothing splines
- K-Fold Cross Validation
- Leave-one-out Cross Validation
- Support Vector Machines
- Logistic Regression
- Divisive (Hierarchical) Clustering
- DBSCAN
- Full ARIMA or SARIMA models
- XGBoost
- Naïve Bayes
- Markov Chain Monte Carlo Methods

If you have another model in mind that you'd like to do, just let me know. I'll most likely approve it as long as we haven't covered it in detail in the course.

Element #2: You should apply it to a real-world dataset. For this, you may use one of the datasets I collected and posted in the course, or you can find a dataset of your own.

Components:

1. **Topic approval**. When you have selected one of the algorithms to code up, contact me so that we can discuss it (during our regular meetings is fine). Ideally, you should do this no later than mid-October or so. Let me know if you have trouble finding a suitable dataset.

2. **Rough Draft**. Students should submit a rough draft roughly two weeks before the final version is due (see Syllabus for date). While this is not expected to be your final, polished form, it should be far enough along in the analysis to see most of your graphs at least in a rough form, a discussion of the code and analysis conducted and your tentative conclusions.

3. **Presentation and Peer Review**: Students will present their results in class on the last regular class day. Peers will review their presentations and these responses will be returned so that any suggestions can be incorporated into the final report.

4. **Report**. The main deliverable is a detailed report of your analysis, describing what you did, what your sources were (factual, programming and data), and your conclusions based on your analysis. Your report should tell a clear story about your data. Include a discussion of any technology used, packages, pre-processing, data acquisition, etc. The report can be submitted in a document form or as an interactive webpage. A reasonable length for this document is about 10-12 pages (graphs and code samples can take up a significant amount of space). A little shorter is fine, a little longer is fine, but you don't need to write a dissertation.

5. **Code**. Include with your report submission any code you used to perform your analysis such as a Jupyter Notebook file, .r file, etc. If the file is too large to submit (such as an entire database), a document of your commands with or without screenshots can be used in replacement. (submit with report)

Your project should include most, if not all, of the following technical requirements:

- Use of the programming language R.

- Code up your algorithm with comments in the code, and apply it to a dataset. You may apply it to a built-in dataset first, to ensure that it runs properly, but should be applied to a real-world dataset by the end.

- Apply appropriate customizations to your algorithm.

- Importing data into R to perform analysis on it.

- Use of both standard and advanced statistical graph types. You should expect to visualize the outcome of your analysis appropriate to the type of data and the method you are applying.

- Use of statistical tools (graphs, descriptive statistics, summary tables, etc.) to analyze data.

- Cleaning data (such as imputing values or removing missing values).

- A sufficiently complex dataset, or more than one combined for analysis or comparison.

Write up your analysis. The written analysis should include (not necessarily in this order):

1. Use formal structure. Include an introduction that includes your research question, a section on any background and sources of the data and theory related to your analysis that a peer may not be familiar with including any limitations, a methods section that discusses how you cleaned and analyzed the data, graphics and tables should be numbered and captioned, results, discussion and conclusion, including any thoughts about where the research might go from this point or what you might have wanted to do but were not able to do here. Ideally, you should also include a short abstract of roughly 75-100 words.

2. Any supplemental research done to support your analysis (and citations/references) in a proper format (though the style is of your choice). Include sources for your data if known. If you use

code you find online, this is okay, but your source must be cited. You can customize it to make it your own.

3. Any equations should be formatted in an equation editor or in LaTeX. (You may or may not have any, though for this project, a discussion of the algorithm may require formulas. Word has a built-in equation editor you can use as needed.)

4. Discuss how the data and analysis connects to statistics discussed in class. Are there any ethical considerations with collecting or using this data? How was this data (likely) collected? Are there any ethical implications of the conclusions?

5. If you produce interactive graphs, you may include a link to where you've saved the graph in dynamic format. (If you are using an interactive website for your report, this is not necessary since it will be active on the page.)

6. While it is expected that a pdf submission is the most likely submission format, you may create your project in Quarto and post it live on a website. You can also create a Word document or pdf from Quarto. While Quarto is not required for this course, it is good practice for your final project for the program.

7. Keep your audience in mind. The technical write up should be more sophisticated mathematically than a presentation. Think of a presentation as directed at more of a lay audience, while the report should be directed at other analysts such as yourself. As this project is more code-based than the ones for other courses, it is expected that you will explain your code inside the report, however, do not include too much in the way of raw R output. Tell me what it means.