INFO 248 Project Guidelines and Requirements

This project is an opportunity for you to explore an aspect of a topic or domain using data science techniques. You will work with one other student during the semester on developing this project. There are several “milestones” during the semester where you and your partner will report on the progress you are making. The project will be due at the end of the semester, see Moodle for dates.

The most important aspect of this project is that you choose a topic that interests you (and your partner). Formulate a question or questions you would like to investigate about that topic to guide your work.

Data science is a process that involves these main steps:

1- Obtaining and "wrangling" data.
2- Exploring, modeling and analyzing data.
3- Presenting conclusions in reports and visualizations.

An outline of topics/techniques covered in this class that you can use in your project.
Note that we require that you use at least two different type of models, and that you use R.

- Describing how a data field is distributed:
    - descriptive stats
    - plots (also useful presenting modeling results)
- Comparing distributions (numeric and categorical)
    - hypothesis tests
    - ANOVA
    - contingency tables
    - chi-square
- Models:
    - linear regression (numeric outcome)
    - logistic regression (categorical outcome)
    - KNN (categorical outcome)
    - decision tree (numeric or categorical outcome)
    - time series (forecasting the future of a distribution)
    - clustering
- Data sourcing:
    - webAPI
    - relational databases
    - web scraping

1.1 Scope of the Project Topic

It is not enough to say that you “will explore income disparity”. You must state more specific questions or aspects of income disparity that you will explore. The more specific your focus the better. Why? Because a narrowly focused question is easier to explore and analyze, and draw conclusions about.

There are many examples of R code for all kinds of applications and domains. You are encouraged to look for examples that relate to your project topic. You are welcome to use any R code you find as a starter for your project. If you do this, you must include an attribution and a url for the site. You are expected to make ***significant modifications*** to any code you download.

1.2 Project Structure

The goal of this project is to learn how to gather data, formulate a problem statement, explore data through statistics and visualization, apply several models to analyze your main question(s), and learn to communicate data in both a technical and non-technical way. The skills learned in this project will be useful for going out into industry and research labs as well as learning to communicate results to a non-technical audience. Finally, this project could be used on students’ resumés and portfolios.

1.3 Format and Submission

The final project submission should include a minimum of 2 documents: the project report, a pdf file- no other format accepted, and the R code you wrote for this project.

Document 1: Project report: A document in pdf format. This should be a knitted Rmd file. This document must include the following sections:

1. The title of the project.
2. Current Date.
3. Your names.
4. A description of what your project is about and an overview of what you did. Include the main research questions.
5. Exploratory analysis: show us the relevant aspects of the data set/sets you are working with: how are the data distributed? Perform any modeling or other statistics that show any interactions between the observations, provide visualizations of the **relevant data**.
6. Analysis: You must use at least three different analysis techniques presented in the course in your project. For example, you may fit a linear regression model and a decision tree model to the same data and compare their results. You may perform

clustering on your data, but be sure to use two different types of clustering to compare results.
7. Summary: This is a summary of your process. State your conclusions, and or a summary of what you found, what was challenging or what worked well. Mentioned any surprises if any.
8. A description of resources you used in the project. This includes any libraries, data sources, other software that you used. Provide links to datasets, but not actual files. Also include any resources you used as a reference or example. - Screen captures of visual output- unless you produce these in the form of a knitted html or pdf file. Note: Your data sources may not be those used in this class.

Document 2: The Rmd file, as well as any other code you write.

1.4 Recommended Steps in the Project

1. You will need to communicate closely with your partner.
2. Choose a domain and topic; what is the focus of your project?
3. Write a paragraph that describes your project. Include as much information as you can about your idea, data sources and other resources you think you will use.
4. Immediately look for one or more data sets to use, download some data and explore it to see if it will be useful.
5. With your partner, make a plan on how you will develop your project. This includes how you will communicate: email, slack, instagram, etc.
6. Set regular meeting dates and times- it is good to meet often, either face to face or remotely. Who will do what tasks? Have your partner check over your work for typos and mistakes. There will be class and lab time as well as office hours to check in with the instructor and staff.
7. Expect that the project work will be done mostly outside of class time.
8. Make prototype/ tests. Install and test any libraries or software that you plan to use and test that it will work for your project. There are usually one or more simple examples that come with a library/software module. Run it!
9. Develop your project in small steps, making sure you have everything working at each step. Make sure you can explain what you are doing as that will be included in your project description.

1.5 Evaluation

You will receive the same grade as your partner. I assume that all students will put forth a good effort on the project and support their partner. If you have any issues with how your project is

progressing, please let the instructor know as soon as possible. There will be check-in times during the rest of the semester as well.

The project must include all sections described in 1.3 Format and Submission:

- Title, Current Date, Your names
- Description
- Exploratory analysis
- Analysis
- Summary
- Resources

Each section will be graded for completeness and correctness, as well as the following criteria:

- Professional presentation: Typos, vague or unreadable language, all images and tables captioned and numbered.
- Complete and correct project description: You give a complete description of what your project was and what you did.
- Plots/tables clearly presented with a title, axes labeled. Raw plot or table output will be downgraded.
- Unformatted output of data, summary, and other raw output will be downgraded.
- Complete project summary;
- Working code- so that someone else can repeat what you did.

Project Milestones

In order to facilitate your project development, there will be a series of three project milestones during the semester where you will submit materials. Note that there are no time extensions for these milestone assignments. The purpose of the milestones is to keep you on track and for us to provide any feedback and guidance so your project is successful.

1.6 Notes on Possible Phases of Project

Phase 1 - Data Gathering

Gathering data. Potential methods of gathering data could include:

- Data repositories: https://archive.ics.uci.edu/, https://towardsdatascience.com/top-sources-for-machine-learning-datasets-bb6d0dc3378b, https://en.wikipedia.org/wiki/List_of_datasets_for_machine-learning_research, https://guides.library.cmu.edu/machine-learning
- Web APIs (eg: NASA, spotify, etc)
- Websites of government agencies (NOAA, CDC<etc)
- Finding data on websites such as Kaggle or data [dot] org, etc with rows > X data points NOTE: most of the data on a site like kaggle has been processed from an original source to suit the needs of the people who are posting it. They often provide their analysis of this data as well. While you may use another analysis as an inspiration for your work, using someone else's work as your own is not acceptable for this project. Your work must be sufficiently different from anyone else.

Phase 2 - Ideating and Questioning Phase

The questioning phase helps you to understand your data and decide on the type of analysis. In this section, students will propose a problem statement or a problem they are trying to solve using data. This section will determine their goal for the project and what they are interested in or trying to achieve

Phase 3 - Exploratory Data Analysis

This section is where students check if the data they have is suitable to answer their questions They will start to develop a sketch of the solution. This can be done without any formal predictive modeling. Ideally they would generate summary statistics, clean and manipulate their data to make it legible for their solution, and make plots to discover their data.

Phase 4 - Statistical Testing and Predictive Modeling

In this phase, students can perform statistical tests learned in class such as t-tests, chi-squared tests (using p-values etc) and will start to experiment with models learned in class such as Regression, Classification models. They will have to clearly outline their data features/predictors and target variables (if any). They will also showcase their predictions and results from their models and do a technical write up (involving p-values, R^2 values, coefficients, etc).

Phase 5 - Results and Interpretation

Students will clearly write up their results from their analyses and modeling and communicate it in a non-technical way. For example, how did their features determine a change in their target variable. For example: When there was a large rise in Cholesterol, there was a large increase in

BMI. The result could also be showcased on a shiny application (This is a good way to present to an audience to interact with it, used at an industry level!)

Example Project Ideas:

You may expand upon any topic we have or will cover in this course, or choose a topic we have not covered.

Social science topics.
- some areas include psychology, economics, sociology

Educational data
- https://cran.r-project.org/web/packages/eeptools/vignettes/intro.html
- https://www.data.gov/education/
- https://github.com/UrbanInstitute/education-data-package-r

Sentiment analysis
- https://uc-r.github.io/sentiment_analysis

Social media data, one example being Twitter Data- text analysis: capture and analyze tweets.

Deep learning and neural networks
- Tensorflow, keras libraries
- https://towardsdatascience.com/how-to-implement-deep-learning-in-r-using-keras-and-tensorflow-82d135ae4889

Propensity scoring on observational data (causality).
- Used a lot by social scientists.

Bioinformatics
- gene finding, microarray analysis
- protein analysis
- biostatistics data

Health Informatics
- epidemiological data
- pathology, such as cancer diagnosis, etc.

Sports Analytics
- https://www.user2017.brussels/uploads/kovalchik_sports_analytics.html - https://blog.revolutionanalytics.com/sports-1/

Financial analysis
- Quantmod library

Climate Data
- Text: http://scrippsscholars.ucsd.edu/s4shen/files/r-textbysamshenjune2017.pdf - https://rclimate.wordpress.com/

- Ocean temperatures
- Seismic events

Computational Archeology

Database:

- A cool site: https://rebrickable.com/downloads/

Sustainable Building Practices

- Passive house heating and cooling data analysis

The list is endless!

Libraries

https://www.computerworld.com/article/3109890/these-r-packages-import-sports-weather-stock data-and-more.html