An Outline of the Final Project

1) Create a Data Science Problem/ or formulate objectives

Your Introduction will fully explain and describe the problem (or question or issue), its history, who it affects, why it matters, etc.

For example, start with a clearly defined problem: A government contractor is experiencing rising costs and is no longer able to submit competitive contract proposals. One of many questions to solve this business problem might include: Can the company reduce its staff without compromising quality?

More General Areas

1. Bio and health
2. Climate
3. Science
4. Finance and Economics
5. Social and public policy
6. Transportation
7. Education
8. Crime and Punishment
9. Politics and government
10. Zoology and Botany
11. Sports

Remember - a data science question will allow you to use all of the analyses that we study in this class, and can be broken down into at least 10 different supporting questions.

Here are some Data Science questions;

1. Does smoking cause lung cancer? Do teenagers who smoke marijuana tend to move on to harder drugs?

2.the amount of money being spent on the nation’s education system was too little, too much, or the right amount?

3. Is there a relationship between the gender of an individual and whom they voted for in the 2000 presidential election, are people who live in certain regions happier, are their educational differences in support for the death penalty? (GSS data)

4. Gender Inequality.

5. Who is going to win the presidential election next year? What region support more on the main parties? What predictions can we make on voting?

6. What are the factors that mostly affect the environmental pollution/Global warming? What will happen in the future?

7. What will happened to the oil prices in the mere future? How will it affect certain companies?

8. The abundance (population estimation) of a certain rare/extinction species? What causes their low population growth?

9. Clinical Trial experiments: for certain diseases; Diabetes, Alzheimers ,.. etc

2. **Collect the Data**
   - Collect multiple datasets.
   - You can use any data set or multiple data sets you want
   - You need to know it is the right data for answering your question;
   - You need to draw accurate conclusions from that data; and
   - You need data that informs your decision-making process.

3. **Clean the Data (As done in ANLY 501)**

   If there is no need to clean your dataset, you can omit this step.

   1. Missing values
   2. Incorrect values
   3. Values with improper formatting
   4. General data consistency
   5. Data formatting
   6. Outliers
   7. Normalization
   8. Transformation (as needed)
   9. Removing unnecessary columns

Questions you should be asking as you clean:

1. Is this cleaning affecting the integrity of the data?
2. Does the cleaning create an imbalance in the data?
3. If you replace a missing value with a measure (such as mean or median), how does this affect the variance? How does it affect the information contained in the data?

4. Analyze Data

After you've collected the right data to answer your question from Step 1, it's time for deeper data analysis. Begin by manipulating your data in a number of different ways, such as,

Doing a descriptive statistical analysis to get a general idea of your data set.

- Use graphical representation (Histograms, boxplots or any type of graphs)
  - Tableau, ggplot is great for Data Viz.
- Use summary statistics

As you manipulate data, you may find you have the exact data you need, but more likely, you might need to revise your original question or collect more data. Either way, this initial analysis of trends, correlations, variations and outliers helps you FOCUS YOUR DATA ANALYSIS ON BETTER ANSWERING YOUR QUESTION.

Then you can decide what are the Statistical tools/methods that you need to use for your analysis: (mostly use methods we cover in class, but if you are going to use more methods that you have learned outside of class, please let me know. Please keep the focus on ANLY511 materials.)

Topics we cover in class:

- Probabilistic models for analysis and simulation.
- Probability distributions and their use in R.
- Expected value, moments, moment generating functions, elements of Monte Carlo simulation.
- Conditional probabilities and distributions, joint distributions, elements of Bayesian network and Markov chains.

- Law of Large Numbers, Central Limit Theorem, random samples, sampling distributions.
- Estimation methods: method of moments, maximum likelihood, confidence intervals.
- Hypothesis tests, permutation tests, likelihood ratio tests, power and error probabilities, Bayesian approach.

- Bootstrap methods for confidence intervals, bias removal and variance estimation.
- Linear regression, prediction, confidence.

During this step, data analysis tools and software are extremely helpful. However, in this class, we will use R. If you use any other statistical software, please mention it in your report.

5. Interpret Results

After analyzing your data and possibly conducting further research, it's finally time to interpret your results. As you interpret your analysis, keep in mind that you cannot ever prove a hypothesis true: rather, you can only fail to reject the hypothesis. Meaning that no matter how much data you collect, chance could always interfere with your results.

As you interpret the results of your data, ask yourself these key questions:

- Does the data answer your original question? How?
- Does the data help you defend against any objections? How?
- Are there any limitation on your conclusions, any angles you haven't considered?
- Is there a new result you found that you didn't expect to observe?

If your interpretation of the data holds up under all of these questions and considerations, then you likely have come to a productive conclusion. The only remaining step is to use the results of your data analysis process to decide your best course of action.

By following these five steps in your data analysis process, you make better decisions because your choices are backed by data that has been robustly collected and analyzed. With practice, your data analysis gets faster and more accurate – meaning you make better, more informed decisions most effectively.

UNDERSTANDING REQUIRED SECTIONS IN ALL PROJECT REPORT SUBMISSIONS

The Introduction: (4 - 6 paragraphs)

- An Introduction is about the data science question. It is about the topics you plan to explore.
- The Introduction is not about the datasets, variables, methods or models.
- The introduction helps the reader to understand what the data science question is, what the supporting topics and issues are, and what the overall area is about.
- An introduction allows the reader to "get to know" the data science question and related areas of interest.
- Ideally, an introduction helps the reader to *care* about the topics and to want to read more.
- The Intro should not contain any information about the dataset or the data cleaning, prep, processing, etc. Everything about the dataset goes into the Analysis section under the "About the Data" subsection.
- Introductions can and should include basis, background, history, the state-of-the-art, images, references, etc.
- An introduction will also help the reader to understand who the topics affect and why the topics matter.

The Analysis/Statistical methods Section

The Analysis section will contain many **subsections.** You may name these subsections in any way that makes sense. However, the first subsection will focus on the datasets, the background, cleaning and preparation, formatting, and other processing.

The next subsections will be about the data **analysis using statistical methods/models. Give an explanation about why you used each statistical analysis method, and a brief theoretical introduction. Also, describe how you used these approaches when analyzing your datasets.**

The Results Section

The Results section will also have many subsections – one for each model or method – as well as many tables, visualizations (graphically), and technical explanations.

Results are technical and they explore what each model or method revealed. Results also discuss and compare parameters.

The Results section offers technical information about what was found in the analysis.

Note: Results can be messy. Make sure your paper has a good flow and offers clarity and ease of reading. Tables, figures, etc. are helpful.

The Conclusions Section

This area is not technical at all.

This area explains what was actually found in a way that would make sense to anyone. The Conclusions are one of the most important parts of any paper. They will not be technical. All technical results should have been described and discussed in the Results section.

The Conclusions should focus on key and important findings and how these findings affect real-life and real people.

Some say that the Conclusions are the most difficult to write. If you do not understand what you really did, how can you explain it to others? Being able to make technical results and complex models useable to normal humans (like managers, CEOs, Deans, clients, etc.) is critical in data science. The Conclusions area is important and if it is not good, many points can be lost.

The Appendix Section

Include an rmd. Code in here that you have used to analyze your data. This is basically include your R codes.