COMP 4441 Final Project Guidelines

Deliverables

Your final project will include 3 deliverables:

- A proposal/plan for your final project
- A 10-15 minute presentation, which you will deliver during the Week 10 live session
- A final paper

General Guidelines

- You may work alone or in groups of 2 to complete your final project.
- You will choose a data set for your project and perform an analysis of this data. Your project will need to include:
    - exploratory data analysis
    - one or more research questions
    - application of 3 different statistical tests or methods we have covered in class
    - assessment of whether the assumptions of these statistical tests hold
    - conclusions about your research questions keeping in mind the ethical guidelines about statistical practice (such as presenting the limitations of analysis)
- You are expected to use R for your data visualization and data analysis.

Final Project Proposal Requirements

The proposal will be graded on completion. For the proposal, you should submit a document with the following information:

- Data set
    - Link to data set (or description of source if it is confidential)
    - Where or how it was collected
    - Who collected it
    - When it was collected
    - Number of observations
    - Names and descriptions of columns (or key columns, if large data set)
    - Any notes on privacy/confidentiality
- Context and motivation for your problem
- Research question(s)
    - Subject matter research question(s)
    - Translation to statistical research question(s)
- Methods
    - Specification of which 3 (or more) statistical tests will be used
    - How these methods will be used
    - Why you believe these methods are appropriate

Think of your proposal as a partial draft of the final paper. It can be a very short partial draft (with only the required elements listed above) or a long partial draft with much more of the paper completed. I will provide detailed feedback which you can incorporate into your final paper submission**.** The more you submit, the more you will be able to get feedback on and have the chance to improve before submitting for a final grade. **I highly recommended submitting a draft of the introduction, data understanding and preparation, and data exploration sections. Being able to get feedback on and revise these sections will likely improve your grade on the final paper.** Only the quality of your final draft will count towards your final grade in the class. The proposal is graded on completion.

Final Paper Requirements

There is no specific page count for this paper. However, there is a detailed checklist of information to include that is based on an excellent guide to best practices for statistical writing developed by Jennifer Van Mullekom. Your paper is considered complete when you have included all these items.

Your paper needs to include the following sections:

1. Executive Summary
2. Introduction
3. Data Understanding & Preparation
4. Data Exploration
5. Modeling & Analysis
6. Results, Interpretations & Recommendations
7. Limitations, Generalizability, and Future Work
8. Attached Code

Final Presentation Requirements

Your final presentation should be a summary of your final paper and should touch on each of the sections 2-7 in the final paper requirements. Plan for a presentation of around 10-15 minutes in length. I strongly recommend making this presentation in Quarto, but that is not a strict requirement, and you may use other presentation tools such as Google Slides or PowerPoint. However, if you use these other tools, you will need to copy and paste your data visualizations from R to the external tools.

Methods You May Want to Use

- Permutation test
- z-test
- t-test (one sample or two-sample)
- Wilcoxon Signed-Rank test
- Sign test
- F-test
- Mann-Whitney U test
- Chi-Square test

- Fisher's exact test
- Linear regression

Suggestions for Finding a Data Set

- Tidy Tuesday Project: https://github.com/rfordatascience/tidytuesday
- UCI Machine Learning Repository: https://archive.ics.uci.edu/
- Kaggle: https://www.kaggle.com/
- US Government: https://data.gov/
- Google data set search: https://datasetsearch.research.google.com/
- IPUMS: https://usa.ipums.org/usa/

Final Paper Checklist

1. Executive Summary

Write this last!

- ✓ Brief project background
- ✓ Project goal
- ✓ Brief description of data set and methods used
- ✓ Key findings and recommendations

No longer than 1-2 paragraphs

2. Introduction

- ✓ Context and motivation for your problem
- ✓ Subject matter research question(s)
- ✓ Translation to statistical research question(s)
- ✓ Brief summary of data source *(recommended: 1-2 paragraphs)*
- ✓ Brief preview of methods to be used *(recommended: 1 paragraph)*
- ✓ Brief answer to research question(s) *(recommended: a few sentences)*
- ✓ Intended audience of document and any notes on privacy/confidentiality

3. Data Understanding & Preparation

- ✓ Description of data set
    - Footnote or hyperlink with link to data set (or description of source if it is confidential)
    - Where or how it was collected
    - Who collected it
    - When it was collected
    - Number of observations
- ✓ Study design or protocol, if it was a designed study or randomized experiment

- ✓ Data definitions
  - Data dictionary with definition of each column in the data set (key columns described in paper but with full data dictionary in appendix if a data set with many columns)
  - Operational definitions of key terms or metrics
- ✓ Data munging steps, for example:
  - If using multiple data sets, how were these data sets merged? If any rows didn't match, how many, why, and how were these dealt with?
  - If any rows were removed from the data set before analysis (e.g. outliers, possible data errors), why and how many?
- ✓ Any data quality concerns
- ✓ Feature engineering
  - New columns created (e.g. combining info from multiple columns)
  - Columns modified (e.g. aggregating multiple categories, re-coding continuous variables to categorical ones)

4. Data Exploration

- ✓ Descriptive statistics
  - Numerical columns - min/max, mean, variance, etc.
  - Categorical columns - counts per value
- ✓ Exploratory data visualizations created using R
  - Single variable graphics (e.g. histograms for numeric variables, bar charts for categorical variables)
  - Multi-variable graphics (e.g. scatterplots for two numeric variables)
- ✓ Exploratory analysis - analysis of descriptive statistics and exploratory graphics
  - Commentary and decisions from data exploration
  - Statistical observations, such as data distributions
  - How do data exploration steps inform your overall modeling and analysis strategy?

5. Modeling & Analysis

- ✓ What didn't work and why? If you considered other statistical tests, why didn't you choose those?
- ✓ Choice of statistical test or model and reasoning for those choices
- ✓ Describe and verify assumptions - e.g. if the chosen test requires a normal distribution, show a qq-plot
- ✓ Relevant statistics related to final choice of statistical test or model

6. Results, Interpretations, Recommendations

- ✓ Restate research questions
- ✓ Discuss results in context of subject matter
- ✓ Interpret conclusions in the context of the project
- ✓ Advise on actions

7. Limitations, Generalizability & Future Work

- ✓ Call out caveats and limitations to work
- ✓ Describe any generalizability issues (why might the findings from this study differ from broader research in this area or not be able to generalize well to other similar problems?)
- ✓ Recommendations for future work
- ✓ Remember ATOM: accept uncertainty, be thoughtful, be open, be modest

8. Attachment With Code

- ✓ R code should be submitted as both a .Rmd file and as a knitted pdf file
- ✓ R code is well-commented and easy to follow
- ✓ R was used for all of the data visualizations, descriptive statistics, and statistical tests