First Project

Project Overview

The goal of this project is for you to demonstrate proficiency in the material we have covered in this class (and beyond, if you like) and apply them to a 'novel' dataset of your choosing in a meaningful way. You are not asked to do an exhaustive data analysis by generating every statistic/visualizations that we have learned of in the course. Rather, I expect you to demonstrate skills in asking meaningful questions and answering them using carefully selected data analysis techniques and tools. These analyses must be done in R in a reproducible manner (quarto).

Multiple data frames are available but your group will choose just one to use for this project. After choosing the data frame, your will need to brainstorm on several issues about the data such as the sample and sampling methods that might have been used to gather the data, a possible target population, variables and their types (numerical vs categorical), among others. Your team will then come up with a research topic/goal and two research questions (directly related to the topic) that you will answer by conducting the analyses. Your analyses should include data visualizations, numerical statistics, and regression models. All these must be appropriately linked to help tell a coherent story.

Research Topic and Questions

Your research topic is a short statement that describes the general area of interest that your research questions will be addressing. It should be broad enough to allow for multiple research questions to be asked, but specific enough to give a clear idea of what you are interested in. For example, if you are interested in the relationship between income and happiness, your research topic could be "**The relationship between income and happiness in the United States**." This topic could then be used to generate multiple research questions, such as

- ***Question 1:*** Is there a relationship between income and happiness in the United States?"

- ***Question 2:*** How does the relationship between income and happiness vary by age group in the United States?

At least one question should involve multiple predictor variables. Ideally, you want to perform model selection to identify the "best" predictors to include in the final model.

The Data

Multiple data sets are available online for you to choose from. If your team decides that you want to collect your own data, please contact me so we can talk about that. The following are possible sources of datasets:

- Openintro website: Most of these data sets can be loaded into your work space as long as the openintro package is installed and running.
- Kaggle: This is a platform for predictive modeling and analytics competitions in which companies and researchers post data and statisticians and data miners from all over the world compete to produce the best models.
- TidyTuesday: Tidy Tuesday is a public and free platform where data science enthusiasts share data and analyses. A new dataset is posted every Monday morning to a public data repository. The data is in a standardized format to help participants analyze and visualize it. The data can be downloaded from their GitHub repository or accessed through the tidytuesdayR package.
- FiveThirtyEight: This is an American website that does opinion poll analysis on politics, economics, and sports in the United States. 538 shares their datasets, related articles, and code.

If your team decides to use existing data set, it is important to understand that the data may have been collected for a different purpose and may not be perfect for your research questions. You will need to examine the data to see if it is appropriate for your research questions. If need be, you can modify your research questions appropriately for the purposes of this project.

Your chosen data set should meet the following criteria:

- Must have at least 200 observations (cases) and 8 variables (columns)
- At least 4 of the variable must be useful and unique predictor variables:
  - Variables such as Property address, Social security number, name, among others may not be useful predictors.
  - If you have multiple columns with the same information (e.g. "state abbreviation" and "state name"), then they are not unique predictors.

- Have both numerical and categorical variables that can be used as predictors.
- Your data set must have at least one variable that can be identified as a unique response variable. This can be numerical or categorical.
- You may not use data that has previously been used in any course materials, such as labs or class activities.

If you are unsure whether your chosen data meets the above criteria, please check with me or the course TA.

Project Timeline

The four milestones for this project are as follows:

1. Milestone 1 - Project proposal.
2. Milestone 2 - A first draft of your project write-up.
3. Milestone 3 - Peer review, on another team's first project.
4. Milestone 4 - Completed project report with write up of your analysis.
5. Milestone 5 - Presentation with slides summarizing key points and findings of your project.

Milestone 1 - Proposal

There are two main purposes of the project proposal:

- To help you think about the project early, so you can get a head start on finding data, reading relevant literature, thinking about the questions you wish to answer, etc.
- To ensure that the data you wish to analyze, methods you plan to use, and the scope of your analysis are feasible and will allow you to be successful for this project.

The project proposal will be due roughly one week after the project is assigned. It must contain the following details:

1. **Team members and their roles**: List the names of all team members and their roles in the project. Roles can include Data cleaning/wrangling, Data analysis (performing the analyses in RStudio), Writing (e.g., abstract, intro, literature review, findings, conclusion), Proof reading and typo fixing, Team leader (e.g., setting up meetings, follow up, communicate with professor), etc.
2. **Research topic and Questions**: As indicated earlier, this is a short statement that describes the general area of interest that your research questions will be addressing. You should have at least two research questions that you will answer by conducting the analyses.

3. **Data**: Describe the data you will be using. This should include the source of the data, the number of observations and variables, and a brief description of the variables and their types (e.g., numerical, categorical). You should also include a brief description of the data collection process. Note that the data collection process may not be available for all data sets. In this case, you are allowed to speculate (reasonably) on the data collection process and be sure to mention that in your proposal.
4. **Analysis plan**: Describe the analyses you plan to conduct. This should include the types of visualizations you plan to create, the numerical summaries you plan to calculate, and the regression models you plan to fit. You should also describe how you plan to link these analyses together to tell a coherent story.

Instructions and grading criteria for this milestone are outlined in the grading rubric for Milestone 1.

Milestone 2 - First Draft

Your project draft is a completed project, that we shall use to obtain feedback from the professor and peers in class. The draft should include the following sections:

1. **Abstract**: This section should provide a brief overview of the research topic, questions, data, and analyses you are conducting. It should also provide a brief overview of the results you have obtained. Think of the abstract as an executive summary of your work that someone can read and understand without reading the rest of the report. You should write the abstract last, after you have completed the rest of the report.
2. **Introduction**: This section should introduce the research topic and questions you are addressing. It should also provide a brief overview of the data you are using and the analyses you plan to conduct. Cite a few sources that are relevant to your research topic.
3. **Literature Review**: This section should provide a brief overview of the literature related to your research topic. You should cite at least 5 sources and synthesize their methods and findings in relation to your own research. Guidelines for writing an effective literature review can be found here.
4. **Methods**: This section should describe the data you are using and the analyses you plan to conduct. It should include the following subsections:
    - **Data Description**: This section should describe the data you are using. This should include the source of the data, the number of observations and variables, and a brief description of the variables and their types (e.g., numerical, categorical). You should also include a brief description of the data collection process. Note that the data collection process may not be available for all data sets. In this case, you are allowed to speculate (reasonably) on the data collection process and be sure to mention that in your proposal.

- **Research Questions and Analyses**: This section should list the research questions you are addressing. You should also describe the variables of interest in your data set that will help you answer these questions. Discuus the methids you plan to use for each question and why you think they are appropriate.
- **Data Cleaning**: This section should describe the steps you took to clean the data. This may include removing missing values, transforming variables, etc. Depending on your data set, you may not need to do much data cleaning. If this is the case, you should still describe the steps you took to clean the data, even if they are minimal.

5. **Data Analysis and Results**: This is one of the most important heavily weighted parts of your project. The analyses should be sound and justified appropriately (under methods). Here, you just need to implement the analyses described under the methodology section of your paper. Your analyses should demonstrate a thorough understanding of statistical concepts learned in this course (and beyond, if you like). Your data analyses must include both numerical statistics and data visualizations. You must also use regression modelling (linear or logistic) techniques as part of your analyses. Interpretations of visualizations created as well as the models created must be made in a coherent manner and in the context of the research questions. These are the results/findings of your project.
6. **Conclusion**: This section should summarize the research questions you addressed, the data you used, the analyses you conducted, and the results you obtained. Findings of your study must be tied back to those reported in the literature review; for example, do your findings concur with previous studies or do they contradict them? It should also discuss the implications of your results and suggest directions for future research. Be sure to state any limitations of your study and what would be done differently if you were to do the study again.
7. **References**: This section should list all the sources you cited in your report. Use APA format for your references.

Instructions and grading criteria for this milestone are outlined in Milestone 2: Proposal.

Milestone 3 - Peer review

Critically reviewing others' work is a crucial part of the scientific process, and MATH 246 is no exception. Your team will be assigned at least one project from another team to review. Team members should read individually before meeting during which the team will come up with collective feedback for the other team. The peer review process will be double blinded, meaning that you will not know who is reviewing your project, and you will not know whose project you are reviewing. This feedback is intended to help you create a high quality final project, as well as give you experience reading and constructively critiquing the work of others.

Instructions and grading criteria for this milestone are outlined in Milestone 3: Peer review.

Milestone 4 - Final Write up

The final write-up is the culmination of your project. It should include all the sections of the first draft, but with revisions based on the feedback you received from your peers and the professor. The final write-up should be a polished document that clearly communicates your research topic, questions, data, analyses, and results. The final write-up should be written in a clear, concise, and professional manner. It should be free of grammatical errors and typos. The final write-up should also include a title page with the title of your project, the names of all team members, and the date. A quarto template for the final write-up will be provided.

Instructions and grading criteria for this milestone are outlined in Milestone 4: Writeup + presentation.

Milestone 5 - Presentation

The final milestone is a presentation of your project. The presentation should be about 10 minutes long and should include slides that summarize the key points and findings of your project. The presentation should be clear, concise, and professional. The presentation will be graded on the clarity of the slides, the clarity of the presentation, and the ability to communicate your research topic, questions, data, analyses, and results to an audience. The presentation will be followed by a question and answer session with the professor and your peers.

Grading

Overall grading

The overall grade breakdown by milestone (M_i) is as follows:

Total	100 pts
M1: Project Proposal	10 pts
M2: First Draft	20 pts
M3: Peer Review	10 pts
M4: Final Writeup	40 pts
M5: Slides + Presentation	10 pts
Collaboration & Teamwork	10 pts

Grading Milestone details

M_1 - Project Proposal

- ***Completion (2pts)*** - Did the team complete all parts of the proposal as required?
- ***Content (2pts)*** - What is the quality of research and/or policy question and relevancy of data to those questions?
- ***Correctness (2pts)*** - Are the statistical procedures sound/reasonable for the stated questions?
- ***Communication & Writing (2pts)*** - What is the quality of the statistical presentation, writing, and explanations?
- ***Creativity and Critical Thought (2pts)*** - Is the project carefully thought out? Does it appear that time and effort went into the planning of the project?

M_2 - First Draft

- ***Abstract (2 pts)*** - Does the abstract provide a brief overview of the research topic, questions, data, and analyses you are conducting?
- ***Introduction (2 pts)*** - Does the introduction articulate the research topic and questions you are addressing? Does it provide a brief overview of the data? Does it provide a couple of citations to help you make the case for your chosen topic of study?
- ***Literature Review (3 pts)*** - Does the literature review provide relevant in-text citations? Does it examine and synthesize the sources in context of the current study?
- ***Methods (3 pts)*** - Does the methods section describe the sample, the data, and the analysis methods used? Does it justify the choice of analysis methods? Does it describe the data cleaning processes?
- ***Data Analysis and Results (6 pts)*** - Are the analyses implemented correctly? Are the interpretations for both research questions done correctly and in context of the study? Do the analyses demonstrate a thorough understanding of statistical concepts learned in the course?
- ***Conclusion (2 pts)*** - Does the conclusion summarize the research questions you addressed, the data you used, the analyses you conducted, and the results you obtained? Does it discuss the implications of your results and suggest directions for future research?
- ***References (2 pts)*** - Are all the sources you cited in your report listed in the references section? Are they in APA format?

M_3 - Peer Review

Peer reviews will be graded on the extent to which it ***comprehensively and constructively*** addresses the components of the reviewee's team's report.

Only the team members participating in the review process are eligible for points for the peer review. Teams may choose to meet in-person or virtually. The team should submit a single review document that includes the following:

- ***Quality of feedback (3pts)*** - Did the team provide ***constructive and actionable*** feedback? Please note that actionable does not mean telling the reviewee team what to do, but rather providing feedback that the reviewee team can use to improve its work.
- ***Correctness (2pts)*** - Are the critiques provided sound and reasonable?
- ***Communication & Writing (2pts)*** - Is the feedback report understandable and free from grammatical errors and typos?
- ***Completion (2pts)*** - Did the team complete the review as required?

M_4 - Final Writeup

The grading of the final writeup follows the same grading criteria as Milestone 2 but gets additional points. To get maximum points on milestone 4, you should address the feedback given to you on your project draft by your peers and/or professor. If there is any feedback that you think does not need to be addressed, please provide a justification for why you think so.

- ***Abstract (4 pts)*** - Does the abstract provide a brief overview of the research topic, questions, data, and analyses you are conducting?
- ***Introduction (4 pts)*** - Does the introduction articulate the research topic and questions you are addressing? Does it provide a brief overview of the data? Does it provide a couple of citations to help you make the case for your chosen topic of study?
- ***Literature Review (6 pts)*** - Does the literature review provide relevant in-text citations? Does it examine and synthesize the sources in context of the current study?
- ***Methods (6 pts)*** - Does the methods section describe the sample, the data, and the analysis methods used? Does it justify the choice of analysis methods? Does it describe the data cleaning processes?
- ***Data Analysis and Results (14 pts)*** - Are the analyses implemented correctly? Are the interpretations for both research questions done correctly and in context of the study? Do the analyses demonstrate a thorough understanding of statistical concepts learned in the course?
- ***Conclusion (4 pts)*** - Does the conclusion summarize the research questions you addressed, the data you used, the analyses you conducted, and the results you obtained? Does it discuss the implications of your results and suggest directions for future research?
- ***References (2 pts)*** - Are all the sources you cited in your report listed in the references section? Are they in APA format?

M_5 - Presentation

- ***Slides (3 pts)*** - Are the slides clear, well-organized, and concise? Do they summarize the key points and findings of the project?
- ***Presentation (4 pts)*** - Is the presentation clear and does it show a thorough understanding of the material learned? Are speakers audible enough and facing the audience? Are they engaging and do they maintain eye contact with the audience?
- ***Question and Answer (Q&A) (3 pts)*** - How well does the team answer questions about their project? Are they able to communicate their research topic, questions, data, analyses, and results to an audience?

Submission

Submissions for all milestones should be made through Canvas. You will see the assignments created in Canvas for each milestone with more detailed instructions.