**Final Project Part 1**

**Research Proposal and Data Introduction**

**Due: October 22, 2021 by 6PM ET**

**Goal of the Assessment:**

The purpose of this assessment is to give you a head start with your final project by finding an area of interest to study, real-world data to work with, and to research a little into your area of interest to see what has been accomplished surrounding your question. This is the general process in proposing a research question and will form the basis for a solid introduction section for your final project report. It will also give you the chance to think about the appropriateness of linear regression as a tool for statisticians. Lastly, it provides an opportunity to get some feedback on your writing and research question that can be used to improve your final report.

**Instructions:**

1. Decide on one (or a few possible) areas of interest that you may want to explore. This can be anything that matters or is of interest to you. Some examples could be (but are certainly not limited to) sports, medicine, public health, economics, video games, literature, etc. Pick something that you really care about.
2. Next, think about possible research questions you may want to study in these areas. What do you want to know in this area? You want to make sure that your question is able to be answered/studied using linear regression models. So you’ll want to frame your question to be something related to modelling a relationship or predicting a value based on this relationship.
3. After coming up with a research question, you will need to find some open-source data that you may use in your data analysis. You want to make sure that the data you find has your response variable of interest (or has variables that could be used to create that variable), as well as any other variable you may want to use as predictors. By looking for data online, you may realize you need to modify your research question slightly or pick another one if you can’t quite find the data you’re looking for. Alternatively, you can stick with your research question but be sure to mention that you expect there to be many limitations to the dataset because it doesn’t quite meet your needs. Step 4 can also help you decide what predictors might be needed for you to answer your question.
4. Once you’ve found your dataset and have decided on your research question (or you can work on steps 2-4 simultaneously and use what you find in all of them to finalize your research question), you need to look at what others have studied in relation to your research question. Do a quick search on the University of Toronto library website to learn about anything related to your area of interest and research question. Look for papers that studied the same question, or something related, that tells you a bit more about what you may need to consider in your analysis and why your research question is important.
   * Focus on giving your reader a rough idea of how many papers have studied this topic (or related to this topic) – this tells us how popular the area of research is and how much research has been done.
   * Give examples from a few important papers about what has been found/discovered to be important in relation to your question (this can be important variables, important results, surprising results, etc.) – this tells us that you are aware of prior results and that you will be using these to plan your analysis.
   * Think about how your research question fits into the area of research. Is it different or new (e.g. nobody has studied this, or maybe it hasn’t been done in this way or this population, etc.)? – this tells us that you see the importance of what you are researching and can frame it against what has already been done.
5. Lastly, perform a short exploratory data analysis of your chosen dataset. You’ll want to focus on identifying anything that you may need to consider moving forward. This includes identifying skews, statistical outliers, variables with high spread or observations that don’t make sense, and missing data, or a dataset that you think doesn’t quite have what you need. You’ll need to present numerical and/or graphical summaries describing the variables. Choose the options that highlight the features of the data that you want to point out but will also let your reader clearly understand the data that you’ll be working with.
   * You want to make sure you specifically mention the presence of any of the above characteristics (or lack thereof) and what this means for the analysis you will eventually perform (i.e. how this might cause problems with the results of linear regression).

**Submission Requirements:**

Content Specific:

Your proposal should be written to satisfy the following requirements:

* Should be organized clearly (consider using headings or sections) and include the following information:
  + Your research question, why you chose it (i.e. why it’s of interest to you), and some background information related to your question/area
  + Justification for why regression is appropriate to use to answer your question and whether you will be aiming for a simpler model that has fewer predictors but is more easily interpretable, or whether you will be aiming for a model with prediction in mind and thus should have strong statistical properties at the expense of interpretability.
  + Details and summaries on your chosen dataset including the variables collected, the number of observations and anything that stands out in the data that would need to be addressed/investigated further in your analysis.
  + References for where you located the data, and your background research on your topic
* Should be written for an audience that has some statistics background but is not necessarily familiar with the area of your research question or linear regression models.
* Should contain figures and/or tables with captions and proper labels/titles as appropriate in your Exploratory Data Analysis section
* Should have references listed in proper APA format
* Should not contain R codes in the proposal itself

Technical Requirements:

Your submission should include the following:

1. Your proposal should be either a Word document or a PDF (both can be created using RMarkdown)
   * This document should be no longer than 3 pages in length (excluding references)
     + About 1 page for the data and research question proposal, and up to 2 pages for your exploratory data analysis and figures/tables
   * Should be typed using no less than 10 point font and be single-spaced.
2. You will also upload a separate Rmd file (or file containing all your R code) in addition to your proposal
3. You will also upload the dataset you are proposing to use in your final project (as a csv or excel file)

**Resources:**

Examples of open data sources:

* <https://open.toronto.ca> for open data from Toronto
* <https://data.ontario.ca> for open data from Ontario
* <https://www150.statcan.gc.ca/n1/en/type/data?MM=1> for data collected by Statistics Canada
* <https://sports-statistics.com/sports-data/sports-data-sets-for-data-modeling-visualization-predictions-machine-learning/> for various sports-related datasets
* <https://data.oecd.org> for data on various country-level variables
* <https://mdl.library.utoronto.ca> for links to many other data portals through the UofT library

Library resources:

* <https://guides.library.utoronto.ca/librarysearchtips/gettingstarted> for more details about searching for articles related to your question
* <https://guides.library.utoronto.ca/citing> for details about why and how to cite your references
* <https://guides.library.utoronto.ca/c.php?g=251103&p=1673071> for help getting the correct citation format

Grading Rubric

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Characteristics** | **Excellent**  **(3 points)** | **Good Effort**  **(2 points)** | **Needs Improvement**  **(1 point)** | **Missing or Incomplete**  **(0 points)** |
| **Clarity and originality of research question**  *(i.e. what are you studying and why does it matter to you)* | The proposed research question is clear and well-defined (i.e. obvious what you are going to study)  AND  The area/topic is creative and is clear why it is of interest to the student | The proposed research question is reasonably clear and well-defined  OR  The area/topic is reasonably creative and is mostly clear why it is of interest to the student | The proposed research question makes it a bit difficult to understand exactly what you will study  AND  The area/topic lacks creativity (e.g. a common/well-known dataset) and/or it is not clear why it is of interest to the student | A clearly stated research question is not provided  AND  No reasons provided as to why this area is of interest to the student |
| **Knowledge of Background Area/Research**  *(i.e. what has already been done, how is what you’re doing different, and why does it matter to the field)* | Student has comprehensively summarized the existing research by providing details on the extent of existing research,  AND  has referenced at least 3 papers directly, with commentary of the importance of the results in relation to their research question  AND  Highlights how the proposed research question fits into the literature and its relevance | Student has summarized well the existing research by providing details on the extent of existing research  AND  Has referenced only 2 papers directly with commentary of the results in relation to their research question (or has referenced more than 2 papers but commentary is lacking or unclear)  OR  It is not entirely clear how the proposed research question fits into the literature or how it is relevant | Student has not provided details on the extent of existing research in their proposed area  AND  Has referenced only 1 paper with commentary of the results in relation to their research question (or has referenced more than 1 paper but commentary is lacking or unclear)  AND  It is difficult to understand how the proposed research question fits into the literature and how it is relevant. | No summary of existing research in relation to the proposed research question is provided  AND  It has not been discussed how the proposed question fits into the literature and its relevance |
| **Appropriateness of Linear Models**  *(i.e. why are linear models are the right tool)* | Student has clearly described in words how linear models will be used to model a relationship and why it is an appropriate statistical tool to address the research question (e.g. correct variable types, assumptions likely satisfied or at least discussed, meaningful to use a model) | Student has clearly described in words how linear models will be used but it is not entirely clear whether it would be an appropriate statistical tool for the data (e.g. incorrect variable types, assumptions no mention of anticipating assumption violations, is it meaningful to answer question with model) | Student has not clearly described in words how linear models will be used to model a relationship to address the research question and so it is not clear whether it would be an appropriate statistical tool (e.g. variable types not specified, no discussion about assumptions, etc.) | No justification provided for why linear models would be appropriate nor how they would be used to address the proposed question |
| **Elements of the Exploratory Data Analysis (EDA)**  *(i.e. is it clear what your data looks like and what might be important)* | EDA is performed and reported in a clear and easily understandable format with all variables of interest defined  AND  Highlights features of the data that may impact/be relevant for the later data analysis | EDA provides some summaries of chosen dataset but not all are presented or done so in an easily understandable way or some variables of interest are not defined  OR  Some features of the data that may impact/be relevant for the later data analysis not clearly highlighted | Minimal EDA is performed and/or summaries of the dataset are not clearly presented, or some are missing, and/or variables of interest are not clearly defined  AND  Some features of the data are highlighted but no reason given for why they are relevant to later data analysis steps | No EDA is performed or reported so the reader does not understand the dataset  AND/OR  Important features for future analysis are not highlighted or discussed in any way |
| **Organization and Clarity of Proposal**  *(i.e. was it properly formatted and organized for easy reading)* | Page length and font size respected  AND  Reference format appropriate and correct  AND  Tables and figures properly labelled and captioned  AND  Information organized in a way that made reading easy and message clear  AND  Minimal spelling or grammar mistakes | Page length and font size respected  AND/OR  Reference format appropriate with minor edits needed  AND/OR  Tables and figures labelled but captions could be more informative  AND/OR  Information organized reasonably clearly but additional structure could improve the flow  AND/OR  Some spelling or grammar mistakes made reading slightly difficult | Page or font requirements not followed  AND/OR  References not formatted correctly or not appropriate  AND/OR  Tables and figures need more appropriate labels and more informative captions  AND/OR  Proposal not structured well and could benefit from more structure to improve readability  AND/OR  Many spelling or grammar mistakes made reading difficult | Page or font requirements not followed  AND  References not formatted correctly  AND  Tables and figures not labelled or captioned  AND  Proposal lacks structure making it difficult to read  AND  Many spelling mistakes made for difficult reading |