**Final Project Part 3**

**Final Data Analysis Report**

**Due: December 17, 2021 by 11:59PM ET on Quercus**

**No late submissions will be accepted**

**Goal of the Assessment:**

Part 3 of the Final Project is your opportunity to demonstrate all that you have learned throughout the course, by showing us that you can use the methods and techniques learned in the course appropriately. You can use the feedback that you have received in Part 1 and 2, as well as in the video project to write a report that is in a common research paper format (IMRD: Introduction, Methods, Results, Discussion). This is likely something that, as a graduate student or a statistician working in industry, you will find yourself doing occasionally.

Since this is used to assess how familiar you are with the use of the tools and methods from this course only, you should NOT use materials that were not covered in this course. Instead, focus on showing us how much you know about everything we have discussed throughout the term.

**Instructions:**

Using any methods and techniques presented in the lecture slides and videos throughout the term, you are tasked with answering your proposed research question by creating the ‘best’ linear regression model that meets the requirements of the research question. You will then need to write a report (details below) that introduces the research question and presents some background, outlines the steps in your analysis that you took to reach the ‘best’ model, presents the results of your analysis and describes and justifies the decisions you made, and finally discusses the final model, its interpretation and its limitations in terms of its ability to meet your research goals. It should be made clear whether you are aiming for a model that makes good predictions, or a model that is more descriptive and easier to interpret, or some combination of both.

The feedback and work you have put into Part 1 of the final project should help you structure your report in a professional and easy-to-read fashion, as well as provide you with a good beginning to your introduction section. You may want to consider adding some additional background research or more discussion about how your research question is important and different from the background you present. The EDA portion of part 1 should be helpful in writing the beginning of the results section, where you describe the data that will be used, combined with the methods you outline, to answer your question.

The feedback and work you have put into Part 2 of the final project should help you structure your methods section of your report, where you will outline the process you took/tools and methods you used to answer your research question. It should also help you with how you approach your data analysis itself.

The video project and feedback should give you some practice into displaying and summarizing your results and interpreting the final model along with a discussion of the limitations that may arise from your final model.

**How to present your results:**

Once you have decided upon the ‘best’ model to fulfill the goal of the project, you must write up a short scientific report. There should be 4 main sections of your report:

* **Introduction section**: where you introduce the purpose and relevance/importance of the project, and provide some relevant background information on the topic (no results or data should be presented here).
* **Methods section**: where you describe and explain the methods, tools and techniques used to arrive at your final model (no results or data should be presented here).
* **Results section**: where you present a description of your study sample, important results that led you to make crucial decisions in building your model (following the methods you outline in the earlier section), and the final model and any other important results
* **Discussion section**: where you interpret your final model and describe why it answers the research question and why it is important, as well as discuss any limitations that still exist based on your results.

You may use tables and plots to help present your results, but they must be relevant and well-thought out so as to convey as much information as possible without being too overwhelming or confusing. When explaining your methods and results, try to avoid just stating that you used a specific method, but add an explanation for why it is the correct tool for the job at hand. See the rubric for more information regarding the various report components.

If you want more information about how to structure your report and what should be contained in each section, see this [cheat sheet](https://www.cmu.edu/gcc/handouts-and-resources/handouts/imrd.pdf) and this [outline for reports](http://ssbcs.smu.ca/webfiles/LabReportsIMRAD.pdf) (you may ignore the abstract portion since you do not need one). Note that not all the elements in these resources need to be included in your report. But you can use these to better understand how to structure your submission.

Finally, if you use any external resources outside of the lecture slides, e.g. to provide background on your topic, you should include a reference section at the end of your report. You may follow APA citation styles to help format your references. For some resources on how to cite, see the library page on [citations](https://guides.library.utoronto.ca/c.php?g=251103&p=1673071).

**What to do if you want to change your dataset or research question:**

If you wish to change your dataset or research question from what was originally proposed in Part 1, you are allowed to do so. However, you will need to write an additional piece that will need to be included with your final report submission. In order to change your dataset or research question, you will need to submit a 1-page document (to be submitted with your final report) that answers the following two questions:

1. Why are you changing your topic or dataset? Elaborate on what made your original dataset or topic not appropriate for the final project.
2. What makes your new topic and/or dataset more appropriate than the previous one? Be sure to clearly state your research question, provide a short, written description of where you located your dataset and what information it contains.

Students who are changing their dataset CANNOT use data that are part of educational resources (e.g. openIntro), that are available on Kaggle or UCI Machine Learning Repository, or are found in an R package or in a textbook. If you are uncertain whether your dataset falls under one of these situations, feel free to ask during office hours.

**Technical Requirements of the Final Report:**

Your report should be typed using whatever software you prefer but must be saved and submitted as a PDF or .docx file on Quercus. Your report must meet the following requirements:

* **Font:** 12-point font in a style similar to Times New Roman
* **Spacing:** single-spaced
* **Word count**: up to a maximum of 1500 words in total (this does not include captions on figures and tables, however, you should also not make captions excessively long or contain information that isn’t mentioned in the main text)
* **Number of tables/figures in main report**: 5 in total, but you may use any combination of tables and figures
* **Figures and table captions**: all figures and tables included should include a caption that describes what is being presented (caption not included in work count).
  + Captions should not contain information that is not also discussed in the main report
* **Figure properties**:
  + All plots should have appropriate title and axis labels, avoiding the use of variables names as they appear in the dataset
  + A figure may include multiple individual plots but they should be related to each other and make sense as to why they are being presented together
    - Avoid having too many plots in the same figure to ensure that they are legible and clear.
* **Reference list** or bibliography at the end of the report (will not count towards word count), using appropriate citation style
* **Appendix:** you may add an appendix at the end of your report to include some additional tables or figures that were not important enough to be part of the main report, but still relevant to your analysis:
  + up to 3 additional tables/figures but they should only be included if they are relevant to the analysis and are referred to in the main text.
* **R code**: In a separate file, you should upload your cleaned and complete version of the R code that was used to conduct your analysis. It should be well-organized and commented appropriately to indicate what each line/section of code is doing.
* **(if appropriate) Explanation for changing data**: If you need to change your dataset or research question, you will also need to upload the 1-page explanation document as a separate file.

**Checklist for submitting final project part 3:**

1. Your final written report, following the requirements above.
2. Your R code that shows your complete analysis
3. (if applicable) your explanation for changing your dataset.

**Things to keep in mind while writing your final report:**

* You do not need to write out the results of every step you took in your analysis as this will make your report too long.
  + Instead, focus on summarizing the most important results, especially where a big decision was made and you need to justify it.
  + For the rest of your results, very short mentions of the process with a brief piece of evidence provided is enough to allow your reader to follow your analysis and understand how you arrived at the final model.
* Rather than presenting results of each step separately (if doing so through a table), consider putting together one larger table that you can refer to in your discussion of many steps in your analysis so that you don’t use too much space
  + For example, if you are selecting between a few different models, you could consider presenting a table that includes many different summaries of the fit of each model and refer to each part as needed in the text, instead of making individual tables for each component.
* Avoid using R output taken directly from R/RStudio. Instead create your own tables where you select only the relevant pieces of the output to display.
* Generally, the methods and results sections tend to be the longest sections, while the introduction and discussion tend to be shorter.
  + Keep this in mind when deciding how much background to provide in your introduction. Often just a paragraph or two is plenty, given the word limits in this project.
  + However, make sure you leave yourself enough space for a solid discussion where you can discuss the impact of the limitations that may exist in your model.