Projects - Sta102 / BME102 - Spring 2015

1 Background

The project(s) in this class represent an opportunity for you to tackle open ended statistical analyses on a novel datasets in order to address a specific research questions. The goal of these projects is for you to demonstrate proficiency in the techniques we have covered in this class and apply them to a dataset in a meaningful and appropriate way. All analyses must be done in RStudio and written up using rmarkdown.

You should write as if you are explaining your results to someone who would be interested in your research question, whether this is other scholars in your field or peers sharing your interest in the topic. Keep in mind that this audience may or may not have taken statistics, but you must be statistically accurate and use correct statistical terminology, but must also explain your conclusions in a way that anyone can understand.

2 Template Files

To download the template files for Project 1 and Project 2 run the following code in RStudio:

```
download.file("http://stat.duke.edu/~cr173/Sta102_Sp15/Proj/custom.css", method="wget", destfile="custom.css")
download.file("http://stat.duke.edu/~cr173/Sta102_Sp15/Proj/project1.Rmd", method="wget", destfile="project1.Rmd")
download.file("http://stat.duke.edu/~cr173/Sta102_Sp15/Proj/project2.Rmd", method="wget", destfile="project2.Rmd")
download.file("http://stat.duke.edu/~cr173/Sta102_Sp15/Proj/proposal.Rmd", method="wget", destfile="proposal.Rmd")
```

3 Project 1 - due Friday, April 3rd by 8 pm

Below is a list of severals datasets each with a brief description, *you are responsible for picking one of the following* and addressing the included research question. Each name below is a link to a description of the data and code for downloading and loading the data.

- Eagles - foraging ecology of bald eagles.
- Heart Disease - retrospective study of coronary heart disease in South Africa.
- SIMS - data on mathematical achievement of middle school students.
- Rotten tomatoes - collection of movie ratings data from the Rotten tomatoes website.
- Course Evals - course evaluation data from UT Austin.

Your write ups should be created using the provided markdown template (‘project1_1.Rmd‘), so that all R code, output, and plots will be automatically included in your write up. Your write ups should be at around 4 pages, including figures. Note that these Rmd files use a custom style file so that the text size will much smaller than the weekly labs. Formating and organization of your analysis will count so be sure to only include results that contribute to your overall conclusions. For example, remove any extraneous exploratory data analysis figures that are not relevant to your analysis.

4 Project 2

4.1 Data set

In order for you to have the greatest chance of success with this project it is important that you choose a manageable dataset. This means that the data should be readily accessible and large enough that multiple relationships can be explored. As such, your dataset must have at least 30 observations and between 5 to 20 variables (exceptions can be made but you must speak with me first). Additionally, your data must represent a sample, and not a population as it is no possible to perform inference with population data. The dataset's variables should include categorical variables (e.g. political party affiliation, gender), discrete numerical variables (e.g. years of education, number of foreign languages spoken fluently), and continuous numerical variables (e.g. height, weight).

All analyses must be done in RStudio using the template file provided. Make sure that you are able to load your data into RStudio as this can be tricky depending on the source. If you are having trouble ask for help before it is too late. Also remember that you must include the code to load your data in the Rmd document as well as any supplementary code or tools (e.g. the inference function).

4.2 Proposal - due Friday, April 10th by 8 pm

On April 10th you will hand in a project proposal. This consists of completing the provided template ('proposal.Rmd') and answering the included questions. This should introduce your general research question (this should include your hypothesized answer) and your data (where it came from, how it was collected, what are the cases, what are the variables, etc.). You will also include some preliminary exploratory data analysis (univariate descriptions of the variables relevant for your research question is sufficient) in order to prove the data is imported into Rstudio and is correctly formated. You will be provided with feedback on the quality of your research question and data so that you will be able to address any issues before completing the final project.

4.3 Project - due Wednesday, April 29th by 8 pm

Project 2 has the same format as the first project, you should provide an introduction that describes your dataset and research question. The bulk of the assignment will consist of a detailed analysis of your data which addresses your research question. You must use RStudio for your analysis and write up all results using knitr using the provided template. This does not mean handing in formulas, but rather an interpretation of what you have found. The goal is not to do an exhaustive data analysis i.e., do not calculate every statistic and procedure you have learned for every variable, but rather let me know that you are proficient at using a statistical package at a basic level, and that you are proficient at interpreting and presenting the results. Focus on methods that help you to answer your specific research questions. Also pay attention to the presentation of your write up - neatness, organization, coherency, and clarity count. Simply reporting summary statistics or p-values is not enough, you must describe what these values mean in context.

Your write up should also include a one to two page conclusion and discussion. This will require a summary of what you have learned about your research question along with statistical arguments supporting your conclusions. Also critique your own methods and provide suggestions for improving your analysis. Issues pertaining to the reliability and validity of your data, and appropriateness of the statistical analysis should be discussed here. A paragraph on what you would do differently if you were able to start over with the project or what you would do next if you were going to continue work on the project can also be included.

5 Submission

For each assignment you must turn in your write up using Sakai's Assignments tool, you will be allowed to upload the assignment(s) multiple times without penalty until the deadline.

Your submission must include:

1. All markdown files (.Rmd)
2. All knit output files (.html)

You do not need to include your datafiles.

Late work policy applies (-10% per day) until all files are submitted in working format. It is your responsibility to confirm that any file uploaded to Sakai are working properly (i.e. corrupted files are not an excuse for late work).

6 Grading

Grading of the project by the professor and TAs will take into account the following:

- Content - What is the quality of research and/or policy question and relevancy of data to those questions?
- Correctness - Are statistical procedures carried out and explained correctly?
- Writing and Presentation - What is the quality of the statistical presentation, writing and explanations?

- Creativity and Critical Thought - Is the project carefully thought out? Are the limitations carefully considered? Does it appear that time and effort went into the planning and implementation of the project?

A general breakdown of scoring is as follows:

> 90%-100% - Outstanding effort. Student understands how to apply all statistical concepts, can put the results into a cogent argument, can identify weaknesses in the argument, and can clearly communicate the results to others.
>
> 80%-89% - Good effort. Student understands most of the concepts, puts together an adequate argument, identifies some weaknesses of their argument, and communicates most results clearly to others.
>
> 70%-79% - Passing effort. Student has misunderstanding of concepts in several areas, has some trouble putting results together in a cogent argument, and communication of results is sometimes unclear.
>
> 60%-69% - Struggling effort. Student is making some effort, but has misunderstanding of many concepts and is unable to put together a cogent argument. Communication of results is unclear.
>
> Below 60% - Student is not making a sufficient effort.

Remember that if you score less 30% on the project you cannot pass this course and that late projects are assessed a 10% per day penalty.