Projects - Sta102 / BME102 - Spring 2014

1 Background

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

Write as if you are explaining your results to whoever 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_Sp14/Project/custom.css",
              destfile = "custom.css")
download.file("http://stat.duke.edu/~cr173/Sta102_Sp14/Project/project1_1.Rmd",
              destfile = "project1_1.Rmd")
download.file("http://stat.duke.edu/~cr173/Sta102_Sp14/Project/project1_2.Rmd",
              destfile = "project1_2.Rmd")
download.file("http://stat.duke.edu/~cr173/Sta102_Sp14/Project/project2.Rmd",
              destfile = "project2.Rmd")
download.file("http://stat.duke.edu/~cr173/Sta102_Sp14/Project/project2_proposal.Rmd",
              destfile = "project2_proposal.Rmd")
```

3 Project 1 - due Friday, April 4th by 5 pm

Below is a list of severals datasets each with a brief description, *you are responsible for picking 2* and addressing the included research question.

- Rotten tomatoes - collection of movie ratings data from the Rotten tomatoes website.
- Course Evals - course evaluation data from UT Austin.
- Cereal - nutritional data from a variety of breakfast cereals.
- Eagles - foraging ecology of bald eagles.
- Heart Disease - retrospective study of coronary heart disease in South Africa.
- Mileage - mileage data on cars from 1999 to 2008.
- SIMS - data on mathematical achievement of middle school students.

Your proposal should be written using the markdown templates given (project1_1.Rmd and project1_2.Rmd), so that all R code, output, and plots will be automatically included in your write up. Your proposal should be at most 2 pages for each data set. Note that these Rmd files use a custom style file so that the text size will much smaller than with the weekly labs.

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 there is no such thing as 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, 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.

4.2 Proposal - due Friday, April 11th by 5 pm

On April 12 you will hand in a draft of the introduction section of your project. This introduction 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.). This discussion of the data should also include some preliminary exploratory data analysis (univariate descriptions of the variables relevant for your research question is sufficient).

The proposal should also include a short data analysis plan and a copy of the dataset with labeled variables (if your dataset is large, handing in a subset of observations is acceptable).

The data analysis plan should include

1. The comparison groups you will use if appropriate.
2. The outcome (dependent, response, Y) and predictor (independent, X) variables you will use to answer your question.
3. The statistical method(s) that you believe will be useful in answering your question(s).
4. What results from these specific statistical methods are needed to support your hypothesized answer?

The introduction should be no more than 2 pages (excluding figures) and the data analysis plan should be no more than 1 page. Your write up and all typesetting must be done with KnitR using Rmarkdown. You will be required to turn in both a written copy as well as electronic copies of the Rmd file and the knit html file.

It is not sufficient to just analyze the data, you must justify why your analysis is appropriate (including an explicit discussion of the necessary assumptions). You must also discuss your conclusions in the context of the research question and data.

4.3 Project - due Monday, April 28th by 9 am

After providing the description of your dataset and research question in the introduction you must apply what you have learned about descriptive statistics, graphical methods, normal approximations, correlation and regression, and hypothesis testing to your dataset. You must use RStudio for this part of your project and write up all results using KnitR. 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 begin to answer your research or policy questions. Also pay attention to your presentation. Neatness, coherency, and clarity count.

Your write up must 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 should also be included.

5 Submission

For each assignment you must turn in the *all* of the following:

- Hard copy turned in at my office, 223E Old Chem. Slide under the door if the door is closed.
- Online at Sakai under Assignments: You can upload the assignment multiple times without penalty until the deadline. See upload guide here.
  1. All markdown files (.Rmd)
  2. All knit output files (.html)

Late work policy applies (-10% per day) until both electronic and hard copy have been turned in.

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.