Programming Languages for Data Science Project Fall 2020

Andy Alverson and John Tipton

Final project due 12 noon on Thursday December 17

Final Project

> My crush rejected me for being a math nerd. I was 2^2 to ask her out...
>
> — Unknown.

> Let us have the serenity to embrace the variation we cannot reduce, the courage to reduce the variation we cannot embrace, and the wisdom to distinguish one from the other .
>
> — Andrew Gelman.

The objective of this project is to apply what you have learned in the course to interesting data. By the end of the project, you will have developed a data science blog hosted through gitHub. For the project, you will choose two different data sets and write up a blog post that tells a story about each of the datasets using the programming methods learned from this course. For data scientists, a blog and a gitHub presence is important for marketing yourself and finding internships and employment post-graduation. The blog will allow you to demonstrate your knowledge, skills, and abilities in a way that makes it clear that you have mastery over data science.

1 Deliverables

1) **Project Proposal (10% of course grade)**
   a) **Initial proposal: due 11:59PM on Thursday December 3**
   b) **Rough draft: due 5PM on Thusrday December 10**
2) **Final report (20% of course grade): due 12 noon on Thursday December 17** (Note: Assigned final exam period is Monday December 14 in the afternoon so this is additional time)

1.1 Project proposal

The project proposal will consist of two components: the initial data description and a rough draft of the project.

1.1.1 Initial data description

The initial data description must be uploaded to gradescope as a pdf. The proposal will consist of the proposed datasets for the analysis, the questions of interest, and a general outline of the project. The proposal will consist of some (very) preliminary analyses and graphics using the proposed datasets, and the proposed questions to be investigate. The initial proposal will be **approximately one-half to one page** that introduces the data, describes the number of observations and variables in the data, and proposes some questions. In addition, the submitted pdf must contain an html link to your blog skeleton that will be used for the final project. An html link to the blogdown site can be created using the syntax (outside a code chunk and in the regular text) `[html link to the blogdown site](https://bookdown.org/yihui/blogdown/)`

The main purpose of the initial data description is to choose the two datasets for the project and to demonstrate that the dataset has enough detail for the project. **Keep this brief and short: We have the power to veto any proposed project at this point - don't put in too much effort until the project is approved!**

1.1.2 Rough Draft

The rough draft must be uploaded to gitHub by **5PM on Thusrday December 10**. The rough draft should, at a minimum, consist of the blog site structure including the about me page, resume page, and links to a blog with at least two posts. Each blog post should include most of the graphics and analyses while the writing and interpretation might still be in a rough draft. The rough draft is an opportunity to get feedback on the project. The rough draft should include enough of the elements of the final report so that we can give feedback on whether the final blog will be sufficient.

1.2 Final Blog

The final report must be submitted by **12 noon on Thursday December 17**. The final blog will consist of

- An about me page where you describe a little about who you are, what your interests are, and why you are interested in data science. This doesn't need to be much.
- A page that links to your resume.
- Two blog posts (one for each dataset). Each blog post will consist of 3-4 graphics that tell the story of the data and 3-4 distinct questions that can be answered using the data. You must have variety in these questions (i.e., don't just repeat the same analysis and plots for three different variables). The final project will be an online blog that will be hosted on gitHub.

Make sure your project appears clean, clear, and professional. The project can be used in the future as a demonstration of knowledge, skills, and abilities in potential job opportunities/graduate school applications/etc. Use good grammar and sentence structure. Future employers will evaluate you based on your analyses and, importantly, whether you can write cogently about what you did and found.

At a minimum, the data analysis must include the following sections:

1.2.1 Introduction (typically 1 paragraph)

Describe the dataset, how the data was collected (if possible) and give a big-picture overview of the data and the questions that you are going to ask/answer with the data. Talk about the source of the data and identify the hypotheses that you wish to test and the questions you seek to answer in clear, concise English in the introduction.

1.2.2 Exploratory Data Analysis (typically 2-3 paragraphs and 1-2 figures)

Exploratory data analysis (EDA) consists of many things that might include calculating summary statistics, visualizing the data to explore relationships in the data, etc. Much of the EDA work in real data analyses is not going to be included in the final blog but this is an important step in the process. Make sure you only include the EDA results that are relevant to the questions you want to answer with the data.

1.2.3 Data analyses (3-4 paragraphs, with included figures)

State the questions that you are wanting to solve with the data. Make sure you describe what the questions are clearly and concisely and describe what steps you are taking to answer the questions.

An important part of any data analysis is to understand the weaknesses/challenges in the data. Think carefully and critically about the data and what kinds of issues might there be. If you were to be able to make changes to the data collection, what would these be? Are there any issues in the data that might make you question your conclusions? If you don't find any weaknesses, state why.

1.2.4 Conclusion (1-2 paragraphs)

Clearly and concisely describe the results and interpret the results in a broader context. Make sure you summarize your results and state what was learned.

2 Datasets

You can use data that are interesting and relevant to you for the project. If you have your own dataset, feel free to propose using that data in the initial proposal. If you don't have access to your own data, there are many different online datasets available. Below I include a list of many available datasets.

Many of the datasets available below have had articles written about the data. You are free to use the articles to understand the data; however, **the project and the analysis must be entirely your own work.**

1) Tidy Tuesday: Tidy Tuesday aggregates a new dataset each Tuesday from a variety of sites. A tutorial on downloading the data from gitHub is available here

2) FiveThirtyEight: FiveThirtyEight is a political and sports analytics website that is focused on data journalism. There is also a *fivethirtyeight R* package

    ```
    library(fivethirtyeight)
    ```

    ```
    ## Some larger datasets need to be installed separately, like
    ## senators and house_district_forecast. To install these, we
    ## recommend you install the fivethirtyeightdata package by
    ## running: install.packages('fivethirtyeightdata', repos =
    ## 'https://fivethirtyeightdata.github.io/drat/', type =
    ## 'source')

    ##
    ## Attaching package: 'fivethirtyeight'

    ## The following object is masked from 'package:openintro':
    ##
    ##     drug_use
    ```

3) BuzzFeed News: Buzzfeed is another data journalism site. There are datasets on a variety of news-worthy topics here.

4) Kaggle: Kaggle is a machine learning competition website. There are many different datasets available on Kaggle ranging from computer vision to sports analytics.

5) UCI Machine Learning Repository: The University of California Irvine website has many classic statistics and machine learning datasets available (we have even used some in this class).

6) Data.gov: A portal to US Government data. Sometimes the data are easy to access, sometimes it can be difficult to find what you are looking for.

7) World Bank Open Data: Data from the World Bank on global development.

8) NYC Open Data: Data from New York City.

9) Reddit Datasets: Reddit datasets subreddit. Who knows what you might find here.

3 Grading

The final project score will be broken down into two components: a proposal grade for the data description and rough draft and a project grade. Each grade will be assigned a score out of 100.

3.1 Initial proposal grading

The initial data description and rough draft will be assigned a completion grade (as long as the work is submitted in good faith and is sufficient progress towards the final draft). This means that as long as you do a reasonable job making progress on the project, the project should get these points; however, we get to make final decisions on what is reasonable progress.

Deliverable	Points
data description	30 pts
rough draft	70 pts

3.2 Final Draft grading

The final draft will be grading according to the following rubric:

Deliverable	Points
Functioning blog	10 pts
About me page	10 pts
resume	10 pts
analysis for data set 1	35 pts
analysis for data set 2	35 pts