STAT 228: Introduction to Data Science

Final Project: The Data Science Pipeline

Contents

Project Guidelines

This project incorporates the entire data science pipeline, from start to finish.

The goal of this mini-project 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 demonstrate that you are proficient at asking meaningful questions and answering them with results of data analysis, and that you can interpret and present the results in a meaningful way. You do not have to apply every statistical procedure covered during the semester. 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 can also be discussed here. In short: **Tell a compelling story with the data!**

You should create some kind of compelling visualization(s) of your chosen data, and describe the steps you take in your analysis without being overly technical/jargony. There is no limit on what tools or packages you may use (as long as it is in R). You do not need to visualize all of the data at once. A single high-quality visualization is much more effective than a large number of poor-quality visualizations. Also pay attention to your presentation. Neatness, coherency, and clarity will count. I also recommend including "**Introduction**" and "**Conclusion**" sections to improve the flow of the story you're trying to tell. Also, try to give your post a catchy or engaging title!

You will compose this post as an **R Studio project** (just in the last 2 projects) You should write your blog post in R Markdown and create your data graphics using `ggplot2`. You will upload:

1. **A .Rmd source file**
2. **The knitted .html file**
3. **Any external (e.g., .csv) dataset files (if necessary)**

   Make sure you disable any unnecessary messages that appear when loading packages, warning messages from plots, etc. Do this by using appropriate chunk options (see Project Tips slides). Pretend you're submitting this post as part of a job application!

In addition, you will be giving a short *five-minute* **speed presentation** of your mini-project on *one of the final three days of class* (**May 1**, **May 3**, or **May 6**). Use your presentation to give a **brief overview** of your project idea.

- *What is the most intriguing thing about your work that will make others curious about it?*
- If you have a figure that summarizes your project well, use that to drive your presentation.

Your zipped project file must be turned into moodle by May 6 at 11:59PM !

Late Work

Late submissions for the project will incur a penalty of 10% per day. After 3 days (May 9), no submissions will be accepted.

Data

You are free to use whatever data you want, *with a few exceptions* (scroll down for more on this). There are perfectly good data sets available through R packages and online that are already well-curated:

R Packages

- `Lahman`: comprehensive historical archive of major league baseball data. Read more about this package here.
- `fueleconomy`: fuel economy data from the EPA, 1985–2015
- `fivethirtyeight`: provides access to data sets that drive many articles on FiveThirtyEight. Read more about this package here.

Other Data Sources

- TidyTuesday – A lot of **great** and interesting datasets from a wide range of topics
- Kaggle Datasets – Need to set up an account to download datasets, but it's simple and free
- Awesome Public Datasets
- Any other data source that you might have in mind…

Do not use these datasets

There are a few datasets that we have covered so much during class, that we have probably exhausted most of their possibilities. They are:

- The `babynames` data
- Any datasets from the `nycflights13` package
- Ames Housing

Grading Rubric

There are 20 possible points for this project.

Note: This rubric is likely not perfect. Please don't be shy to ask for clarifications!

Baseline

- +2 for a project that compiles without errors. The R Markdown file must "knit" correctly *when someone else knits it.* This is known as producing **reproducible research**. This is why it is essential to use an R Studio project when creating your blog post.
- +2 for clearly explaining any data wrangling steps that you take (this can be in a few sentences in text surrounding the R code chunks)
- +1 for readable, well-documented code (i.e. appropriate comments, breaking code up into logical chunks, clean code)
- +1 for including a problem statement/research question (what are you trying to figure out?)
- +1 for describing the dataset including a citation for your data (either as a link to a website or stating the R package it is from)
- +1 unnecessary messages from R are hidden from being displayed in the HTML
- +2 for clearly presenting your project idea via a *speed presentation* to the class (need to be understandable and professional)

Average

- +1 for explaining coherently what we can learn from these data
- +2 for including at least one well-labeled and informative data graphic
- +1 for providing context or background useful in interpreting the graphic
- +3 for including modeling methods in your post where you *train*, *test*,and *evaluate* the model.
  - Possible models: *logistic regression*, *decision trees*, *bagging/random forests*, and *boosting models*, but you're welcome to explore other predictive modeling/machine learning techniques if you'd like!

Advanced

- +1 Blog post provides context, background, and/or motivation useful in analyzing the chosen dataset
- +0–2 reflects the professor's judgment of the overall quality of your submission
  - 0: Post is relatively simple, confusing, sloppily executed, or not well thought out; graphics are relatively simple, confusing, sloppily executed, or not well thought out; analysis is muddled or uninformed.
  - 1: A solid all-around effort, each section is appropriate and well-motivated; topics are well-studied and communicated effectively
  - 2: Post is special, showing creativity, an original design, and/or exceptional attention to detail (e.g., sharing the post on your blog website, graphical techniques not specifically covered in class, comparing multiple predictive models); clear in the sense that a novice in R and in the data subject matter could learn something from your post.