Spring 2019

Project: Probability and Statistics - Course Project

Evan Agovino
CUNY City College

NYC Tech-in-Residence Corps

The goal of the CSC 217 Course Project is to give students the opportunity to explore and analyze one or more data sets of their choosing with the goal of telling a compelling narrative using data. Projects should incorporate all of the following, as per the official course objectives:

Single Variable Explorations: Examine a single variable, understand its underlying distribution, and choose the appropriate summary statistics for it.

Pair-Wise Exploration: Identify possible relationships between variables and compute correlations and linear fits.

Estimation and Hypothesis Testing: Understand the following three questions when reporting statistical results: 1) How big is the effect? 2) How much variability should we expect if we run the same measurement again? 3) Is it possible that the apparent effect is due to chance?

Visualization: Use data visualization as a tool for examining data and communicating results

Think of something that interests you and work backwards from there - feel free to get creative with the data you're pulling. What is the economic difference between red and blue states? Who is the better basketball player? Is the American economy better today or in the mid-90s? Have the winters in New York been getting warmer? Are the subways getting worse? You don't need to reinvent the wheel! I would much rather have you focus on analysis than extensive data collection.

Think of a hypothesis you want to prove or disprove, and then think of the data you need to retrieve to answer it and the tools you need to properly answer your question. Perhaps most importantly, think and define what metrics will be relevant in answering your question. How will you determine which basketball player is better? Why is that your metric of choice?

Your project should state your objective clearly, provide the reader with clear context of the data using descriptive statistics, and then answer your question by clearly defining the metrics needed to answer it and then leveraging the statistical frameworks discussed in class. Specifically you are encouraged to either do a hypothesis test or a regression, but exploratory data analysis as a means of answering your question will be allowed if your work is thorough, convincing, and helps answer a question.

Your work should be something you're proud of! Students are encouraged to post the project on their Github to add to their portfolio as something they can show to potential employers.

Project Deliverables

Due *March 13th*:

A **project proposal** that includes:

- A brief paragraph identifying the major theme the project aims to cover
- A list of at least three potential questions the analysis will address
- A data set that addresses those questions
- Four or more variables to be included in the project
- Any concerns you are having about the data pull

Due *March 27th*:

A **project update** that includes:

- A brief section addressing feedback from the project proposal
- A 200+ word summary of findings from the data exploration
- At least two visuals from exploratory data analysis
- Any concerns you are having about the project at-large

Due *April 10th*::

A **project update** that includes:

- A brief section addressing feedback from the last update
- Any concerns you are having about the project at-large

Due *May 8th*:

The **final project** is due. Deliverables include:

- A write-up of 500+ words containing your findings
- A Jupyter Notebook containing your workflow, along with any supporting comments in Markdown cells. This code should be clean and follow your narrative. This code needs to be reproducible.
- A self-assessment of at least 200 words talking about your experience with the project: what you enjoyed, what you didn't enjoy and what you found challenging and/or limiting in getting you to your goal

Grading

This project makes up 25% of the grade for CS 217. Each student's grade will be determined by the rubric below.

Clarity

- Is the hypothesis/problem stated clearly?
- Is it clear why the data chosen should help answer the question being asked?
- Is the clear why this problem is important and/or interesting? Is it presented in an engaging manner?

Data Exploration

- Did you examine the distributions of all of your relevant features?
- Did you check for metrics of central tendency, metrics of variability, and presence of outliers in each of your features?
- Did you examine the relationships between each of your variables?
- Did you provide relevant visuals to support your findings?
- Are the visuals properly labeled, with titles and labels on axes?

Statistics

- Is your hypothesis correctly stated?
- Is your process and conclusion statistically sound?

Write-Up

- Do you tell a coherent story with a beginning, middle and end?
- Is you writing free of spelling errors and grammatical mistakes?

Timeliness

- Were all of your deliverables submitted on time?
- Did you address all of my feedback?

Code

- Is your code clean?

- Does your Jupyter Notebook tell your story, complete with Markdown code?
- Is your code reproducible?