Brief Project Review

•••

CS 217

Topics

You submitted proposals on everything from:

- Examining the relationship between autism and vaccinations
- Exploring the distributions of Wi-Fi hotspots in different boroughs
- Looking at which types of landscapes are more likely to lead to forest fires
- Asking whether the 1996 Bulls or 2016 Warriors were a better team

Data Over Everything

- Good data first, narrative second
- Don’t spend hours figuring out how to pull data for the perfect story
- A successful regression is better than an ambitious topic with no results

Defining Your Problem

- Be specific about the question you are trying to answer with your analysis
- You can reverse engineer this once you figure that out :)
- Good: What is the relationship between GDP and IQ per country?
- **Better**: Does a country’s higher GDP lead to a higher IQ for the population?

Defining Your Variables

- Good: Does the United States have more expensive healthcare per capita than other developed countries?
- **Better:** Use a clear definition of what a “developed country” is
- Good: Has the median temperature gone up in NYC over time?
- **Better:** Is the median temperature of NYC over the past ten years different than the ninety years before it?

Problem Solving

- Hypothesis testing (one thing is greater than the other), or regression (this thing is caused by this other thing or series of things) is recommended for your analysis, but not required
- You can weave a narrative based on Exploratory Data Analysis, but the burden is on you to tell a story from end-to-end
- Example: Say you are looking at suicide rates across different countries
    - Which countries have disproportionately high suicide rates? Or disproportionately low suicide rates? Are they formal outliers, as per a boxplot?
    - Which countries have disproportionately high suicide rates for women? Or for men? Which countries have a disproportionately high gap between men and women in suicide rates?
    - What are the characteristics (GDP, location, etc...) of countries that stand out in one way or another?

Timeline Moving Forward

- Project is due on **May 8**
- That sounds like it’s in a long time for now. **It is not.**
- March:
  - Finalize dataset, ensure you can pull data successfully
  - Cleaning and exploring data
- April:
  - Exploring data
  - Visualizing data
  - Modelling data
  - Hypothesis testing, ANOVA Testing, Regression

Timeline Moving Forward

- Next deliverable: Due **March 27**
- 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
    - Histograms of two different variables, for example
  - Any concerns you are having about the project at-large