Capstone I

Unknown user • Jul 8, 2022 (Edited Dec 12, 2023)

5 points

For this assignment, you will be able to choose a set of data on Kaggle or similar and do your own analysis of it! This project is meant to highlight the **analysis process** as you use multiple languages and platforms to analyze the same set of data.

Choosing a Data Set
Below, I have added a few common resources for finding data. Some of these websites change regularly. Some of them are not formatted ideally for what we do. Before you choose a set, try and make sure that it is a data set you have the tools to analyze. For example, some soccer data is arranged into graph data, which is not a normal data-analysis format; if there's a tabular or noSQL format, that's more common and will be a better resume piece anyways; so don't try and reinvent the wheel just because you found a cool data set. I recommend you find a topic you like and search elsewhere for similar data sets.

Imagine you are working for a business and your boss **doesn't understand data**. Your mission is to create a report explaining what is going on with the data and what conclusions you have made from the data. (The data set you choose doesn't have to be related to a business at all).

Before you start, I want you to **write out a hypothesis** about the data, even if you have no clue what the data means. You may also add hypotheses as you go and test them - often you will have many, many questions about your data that you'll want to answer. Then, you'll be using R and Python to analyze whatever data you find. I recommend going through the entire process with one language and then trying it with the other. Googling things is totally encouraged if you need to find specific tools for Python or R, particularly if there are specific tools from the tidyverse to help out; this won't be needed for every project. This project is intentionally open-ended, as an important part of data analysis is **knowing your data**.

After you've looked at a lot of the data, you will write a report detailing your process and showing the results of your analyses. For each major hypothesis you look into, you should cover that in the report and include charts, graphs, anything necessary to communicate the data. I am much less concerned with things like grammar and styling and **very interested in seeing how you got the conclusions you made in a way that you can explain to someone who doesn't speak data.**

Have fun with this project! There is no length requirement on the report; I just want to see you communicate well about what you've found and what the data means. I anticipate the project takes around 10 hours; if it takes more or less, that's okay, but you should be very familiar with your data and writing a good report on it.

The submission should include **all your technical work, and the report itself**. Put them all in a GitHub repo and post the repo here.
Please book here after submission: https://calendly.com/d/48z-p4c-5kb/data-capstone-presentation

Sports Datasets ...
https://sports-statistic

GitHub - statsbo...
https://github.com/sta

Datasets - Spotif...
https://research.atspo

Project datasets ...
https://perso.telecom-

Free Public Data ...
https://www.tableau.c

Our World in Data
https://ourworldindata

Data.gov Home -...
https://data.gov/

Class comments

Add a class comment

Your work Assigned

+ Add or create

Mark as done

Private comments

Add comment to Unknown user