TO: 040119 Data Science Cohort
FROM: DS Ed Team
DATE: May 5, 2019
SUBJECT: Revised Module 2 Project Instructions

PROJECT GOAL

The goal of this project is to test your ability to gather information from a real-world database and use your knowledge of statistical analysis and hypothesis testing to generate analytical insights that can be meaningful to the company/stakeholder.

USE OF ALTERNATIVE DATABASES

The original idea was to use the Northwind database, an example relational database created by Microsoft. After thinking about the goal of the project, we want to open up alternative datasets that can be used in lieu of the Northwind database to give students the opportunity to interact with real-world data.

SQL REQUIREMENTS

You are required to import your static files (i.e. .csv, .json, .txt) files into a PostgreSQL database. By working with a relational database, you'll get practice at crafting queries that pull out relevant data prior to performing statistical analysis.

STATISTICAL ANALYSIS REQUIREMENTS

The goal of your project is to query the database to get the data needed to perform a statistical analysis. In this statistical analysis, you'll need to perform a hypothesis test to answer at least one of the questions from the database you choose.

For each hypothesis, be sure to specify both the null hypothesis and the alternative hypothesis for your question. You should also specify if this is one-tail or a two-tail test.

In addition to answering this question with a hypothesis test, you will also need to come up with at least 3 other hypotheses to test on your own. These can by anything that you think could be important information for the company/stakeholder.

REVISED STAKEHOLDERS

The use of alternative datasets brings with it a question of who your audience is for this data science project. Much like the Module 1 project, picking an audience at the beginning of your project helps you define the scope of the project. Once a stakeholder is picked, keep them in mind as you're generating your statistical analysis. When translating statistics for a non-technical audience, be sure you are answering questions that are relevant to the stakeholder and being clear with the limitations of your findings.

DELIVERABLES

To complete this project, you will need to turn in the following 4 deliverables:

1. A *Jupyter Notebook* containing any code you've written for this project. This work will need to be pushed to your GitHub repository in order to submit your project.
    a. The notebook contains well-formatted, professional looking markdown cells explaining any substantial code. All functions have docstrings that act as professional-quality documentation.
    b. The notebook is written to technical audiences with a way to both understand your approach and reproduce your results. The target audience for this deliverable is other data scientists looking to validate your findings.
    c. The notebook should be well organized, easy to follow, and code is commented where appropriate.
    d. Your notebook should clearly show how you arrived at your results for each hypothesis test, including how you calculated your p-values.
2. A user-focused README.md file that explains your process, methodology and findings.
    a. Take the time to make sure that you craft your story well, and clearly explain your process and findings in a way that clearly shows both your technical expertise *and* your ability to communicate your results!
3. An *"Executive Summary" Keynote/PowerPoint/Google Slide presentation* (delivered as a PDF export) that explains the hypothesis tests you answered, your findings, and their relevance to the company/stakeholders.
    a. Make sure to also add and commit this pdf of your non-technical presentation to your repository with a file name of presentation.pdf
    b. Contain between 5-10 professional quality slides detailing:
        i. A high-level overview of your methodology
        ii. The results of your hypothesis tests
        iii. Any real-world recommendations you would like to make based on your findings (ask yourself--why should the executive team care about what you found? How can your findings help the company/stakeholder?)
        iv. Take no more than 5 minutes to present
        v. Avoid technical jargon and explain results in a clear, actionable way for non-technical audiences.

ALTERNATIVE DATABASES

- **Housing:** King County Assessor Data
  https://info.kingcounty.gov/assessor/DataDownload/default.aspx
  - We have a hypothesis that the housing market is hotter in spring and summer than in fall and winter. Suppose that you have just relocated to Seattle and you are in the market for a home. Would it be better to buy now, or would we expect to get a better deal if we wait until fall?
- **Grades:** University of Wisconsin, Madison
  https://www.kaggle.com/Madgrades/uw-madison-courses
  - Does your teacher have a statistically significant effect on the number of As earned in a course?
  - Does time of day have a statistically significant effect on the number of As earned in a course?
  - Do STEM fields have a statistically significantly difference in the number of As earned when compared to the humanities?
- **Music:** Pitchfork Reviews
  https://www.kaggle.com/nolanbconaway/pitchfork-data
  - Is there a statistical difference between the ratings of two different music genres?
  - Is there a difference between the ratings of {insert genre here} music and all other music?
  - Are the albums from one label rated differently than the wider population?
- **Football:** European Soccer Dataset
  https://www.kaggle.com/hugomathien/soccer
  - Is there a statistical difference in the odds of winning a game when a team is playing in front of their home crowd?