TO: 100719 Data Science Cohort
DATE: November 4th, 2019
SUBJECT: Module 3 Project Instructions

YOUR MISSION

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.

Choosing your data
In this project, you are free to choose any data that you would like in order to conduct various hypothesis tests to answer questions that your company or stakeholder may be interested in. You should invest not more than 1 hour to find data. Your data source should be from an API but you may merge in data from another source such as a CSV file if you would like.

Stakeholders
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.

REQUIREMENTS:

Data Source (beware of GitHub limitations with data size)
For this project you are required to obtain data from:
- At least one API source
- Optional data from CSVs can be merged into your dataset

Statistical Analysis Requirements (each partner should test at least 1)
The goal of this project is to perform hypothesis testing on the collected data.
For the project you will be required to:
- Come up with 4 separate hypotheses to test (each test consisting of a clearly identified null and alternative hypothesis)
- Describe what statistical test you will use to test the hypothesis and why. (independent t-test, dependent t-test, ANOVA etc.).
- Be sure you are checking your test's assumptions (ex. equal variance, central limit theorem).

Visualization Requirements (each partner should make at least 1)
As a part of presenting your results to stakeholders you should include:

- At least 1 visualization per hypothesis test
- At least 2 visualizations from data exploration

And you should be able to justify how this is relevant to your presentation.

DELIVERABLES

Your team is expected to use git as a collaborative tool for this project to manage version control and history. All documents must be contained in a git repository that you create. You should use the templates provided by instructors here.

1. **A README.md file**
   a. Listing project members
   b. Goals
   c. Responsibilities
   d. Summary of the files in the repository

2. **Use of git**
   a. Multiple commits each day
   b. One push every day, minimum
   c. Must include short, descriptive commit messages
   d. Each project member should commit at least once
   e. Be sure to use branches to work individually and merge to master when complete

3. **Technical Narrative Jupyter Notebook**
This notebook is targeted to a technical audience and should contain the following:
   a. Documentation of where the data came from- API and any additional CSV sources
   b. **Clean and commented code** so an independent party can read your analysis and concur with your analytical choices
   c. Code should follow Pep8 standards
   d. Custom functions should be stored in a .py file and imported whenever possible

4. **Three Python files**
You should include these .py files using the templates provided in your GitHub repo and the functions in them in your technical notebook. The three files should be called:
   a. data_prep.py
   b. visualizations.py
   c. hypothesis_tests.py

5. **Slidedeck**
You should include a pdf of your slide deck targeted to the non-technical audience in your repo that includes:

a. The purpose of your analysis and why it matters
b. An abbreviated high-level overview of the methodology
c. 4 visualizations
d. Analysis of your hypothesis tests
e. Actionable insights based on the results of your hypothesis tests
f. No more than 10 slides

6. **Presentation**
Your team must prepare a 5-minute presentation that presents the results of your analysis.
Your presentation should use the template provided and include it.
Vocabulary targeted to a non-technical audience, avoid jargon.

SCHEDULE:
07/11 Thursday Afternoon
- Project Assignment

11/11 Monday Morning
- Review API and other data sources
- Review goals/questions
- Review hypothesis tests you plan to conduct
- Review work plan created for how teammates will approach and divide work

14/11 Thursday Afternoon
- Demo presentation with feedback from instructors
- Have a draft of deck completed
- Have a version of jupyter notebook completed

15/11 Friday Afternoon
- Presentations
- Afternoon project presentation to the class
- Science fair open to staff and fellow students

UNIT TEST
If you have any issues with these unit tests, please talk to a coach.
Test scripts and lint scores will be used to provide real-time feedback on project performance.

You can expect to see the following
Clean Data Tests:
- test_no_null_values
- test_no_duplicates
- test_cells_no_brackets
- test_column_name_lowercase

    - test_column_name_whitespace
    - test_if_dataframe

Visualization Tests:
    - test_if_matplotlib_object
    - test_title
    - test_xaxis
    - test_yaxis

If any requirements are missing or if significant gaps in **understanding** are uncovered, be prepared to do one or all of the following:
- Perform additional data cleanup, visualization, and/or feature selection
- Submit an improved version
- Meet again for another Project Presentation

What won't happen:
- You won't be yelled at, belittled, or scolded
- You won't be put on the spot without support
- There's nothing you can do to instantly fail or blow it