TO: 080519 Data Science Cohort
DATE: September 10, 2019
SUBJECT: Module 3 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.

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.

Project Requirements:

Data Source

For this project you are required to obtain data from:

- At least one API source
- Optional data from a CSV can be merged into your dataset

Statistical Analysis Requirements

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).
- Explain what test (e.g. one-tailed t-test) you are using and why.

Visualization Requirements

As a part of presenting your results to stakeholders you should include:

- 4 meaningful visualizations related to your data exploration or hypothesis testing

Project 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 listing project members, goals, responsibilities, and a summary of the files in the repository.
2. At least 10 commits
    a. Must include short, descriptive commit messages
    b. Each project member should commit at least once
3. Technical 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 replicate your analysis and justify your analytical choices
    c. Code should follow Pep8 standards
    d. Custom functions should be stored in a .py file and imported whenever possible
4. Narrative Jupyter Notebook- This notebook is targeted to a non-technical audience and should contain the following:
    a. The purpose of your analysis and why it matters
    b. 4 well-annotated visualizations created using Matplotlib/Seaborn
    c. Results of your 4 hypothesis tests
    d. At least four actionable insights based on the results of your hypothesis tests

5. 3 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
6. Slidedeck- You should include a pdf of your slidedeck targeted to the non-technical audience in your repo that includes:
    a. Use of the template formatting
    b. An abbreviated high-level overview of methodology
    c. 4 visualizations
    d. Results of your hypothesis tests
    e. Exported visualizations from analysis
    f. Justification of at least 4 concrete recommendations
    g. No more than 10 slides
7. 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:
    a. Your project aims/questions
    b. The process you went through
    c. At least 4 meaningful data visualizations to help illustrate your findings
    d. Vocabulary targeted to a non-technical audience, avoid jargon
    e. No more than 10 slides

Project schedule

9/10 Tuesday - Project Assignment

- Schedule Wednesday check-in with coaches

9/11 Wednesday- Check in with coaches

- 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

9/12 Thursday - Demo presentation with feedback from instructors

- Have a draft of deck completed
- Have a version of jupyter notebook completed

9/13 Friday- Presentations

- Afternoon project presentation to the class
- Science fair open to staff and fellow students

Project Partners

1. Logan and Sebastian
2. Bailey and Ahmed
3. Kate, Aktan and Philip

Project Review

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