TO: 010620 Data Science Cohort DATE: February 13, 2019

SUBJECT: Module 3 Project Guidelines

PROJECT GOAL

The goal of this project is to be able to utilize statistical analysis and hypothesis testing to answer questions that your company/ stakeholder may be interested in. You will be tested on your ability to gather information from a real-world database and generate analytical insights that will be meaningful to the company/stakeholder.

Choosing your data

In this project, you are free to choose any data that you would like that enable you to test out your various hypotheses regarding information your company/stakeholder is interested in.. You should invest not more than 1 hour to find data. If you're having trouble coming up with ideas, we recommend googling APIs for a subject of interest to you. Maybe you can merge it with different datasets.

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 are conducting 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, it is recommended that you gather your own data using API sources. You are required to obtain data **NOT** from any pre-cleaned data sources. Ensure that your dataset contains information that would help you test your hypothesis.

(Be aware of GitHub limitations with data size)

Statistical Analysis Requirements

The goal of this project is to follow all the necessary steps of hypothesis testing on the collected data. For the project you will be required to:

- Come up with at least 3 separate hypotheses to test (each test consisting of a clearly identified null and alternative hypothesis).
- Explain why you are using those specific tests (e.g. one-tailed t-test). Be sure to check that your test's assumptions have been met (ex. equal variance, central limit theorem).
- Provide at least 2 confidence intervals to support your insights.

Make sure to report all test results along with effect sizes in your technical notebook.

Visualization Requirements
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.

You should also be able to justify how these visualizations are relevant to your presentation.

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. This summary should also include a guide to navigate your notebook.

2) **Multiple commits and at least one push every day**.
   a) Must include short, descriptive commit messages.
   b) Each project member should commit at least once.
   c) Be sure to use branches to work individually and merge to master when complete.

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

Although this notebook is called “technical” it should be well-commented and should include proper reasoning for each subsequent step taken.

4) **Python files** - You should include .py files using the templates provided in your GitHub repo and the functions in them in your technical notebook. Example files may be:
   a) data_prep.py
   b) visualizations.py
   c) utils.py (for extraneous functions)

5) **Slidedeck -** You should include a PDF of your slide deck targeted at the non-technical audience in your repo. It should include:
   a) The purpose of your analysis and why it matters.
   b) A high-level overview of your data sources.
   c) Analysis of your test results.
   d) All visualizations from your analysis.
   e) Actionable insights based on the results of your hypothesis tests.
   f) Conclusions and possible future actions
   g) No more than 10 slides.

h) No python screenshots

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

Project schedule:

02/13 Thursday Afternoon - Project Assignment

- Start brainstorming potential questions
- Begin researching data sources of interest

02/14 Friday Afternoon - Check in with coaches and leads to review the agenda:

- Hypothesis tests you plan to conduct
- Data sources
- Plan for how the team will divide the work.

02/18 Tuesday Morning - Check in with coaches and leads on agenda:

- Sample of data you obtained
- Discuss again what hypothesis tests you plan to conduct

02/19 Wednesday Morning - Check in with coaches:

- Talk through any assumptions made
- Answer any lingering questions/concerns

02/19 Wednesday Afternoon - Demo presentation with feedback from instructors based on agenda:

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

02/20 Thursday Afternoon - Presentations

- Project presentation to the class

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