TO: 100719 Data Science Cohort
DATE: November 14, 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. What we **don't** want to see is data imported from sources like Kaggle or UCI. 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 .csv files.

**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** (beware of GitHub limitations with data size)
For this project you are required to obtain data **NOT** from:

- Kaggle
- UCI
- Pre-cleaned data sources

**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 2 separate hypotheses to test (each test consisting of a clearly identified null and alternative hypothesis).
- Explain why you're using one test over the other (e.g. one-tailed t-test). Be sure you are proving your test's assumptions have been met (ex. equal variance, central limit theorem).

**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.

And you should be able to justify how this is 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.

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. 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. 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.

## Project schedule:

**11/14 Thursday Afternoon** - Project Assignment
- Begin researching data sources of interest
- Start brainstorming potential questions
- Schedule Monday check-in with coaches

**11/18 Monday Morning** - Check in with coaches to review:
- Data sources
- Hypothesis tests you plan to conduct
- Plan for how the team will divide the work.

**11/19 Tuesday Afternoon** - Demo presentation with feedback from instructors
- Have a draft of deck completed
- Have a version of master notebook completed

**11/20 Wednesday Afternoon** - Presentations
- Afternoon project presentation to the class
- Science fair open to staff and fellow students

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