TO: 0402219 Data Science Cohort
DATE: June 4, 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.

Project pairs

Tim (driver) + Jon (navigator) + Manisha (navigator)
Brahm (driver) + Adam (navigator)
Filis (driver) + Angel (navigator)
Helen (driver) + Llewellyn (navigator)
Mark (driver) + Ryan (navigator) + Bassel (navigator)
Pablo (driver) + Kelly (navigator)

Notice : the driver is expected to write most of the final code (~70%).

Project planning

Before starting the project please write down a work-plan with priorities and how do you intend to split the work. You should spend ~1 hour to devise this plan.

Project schedule

Wednesday:

09:00-10:00 - Find your dataset/use existing datasets (bottom of the document)
10:00-11:00 - Import dataset to an SQLite3 file
11:00-13:00 - Form your Hypotheses (3-4) and the outline of what they will require from you to read/code
13:00-14:00 - Devise a work plan of how you intend to split the work.

Choosing your data

In this project you are free to choose any data that you would like in order to conduct various hypothesis tests. You should invest not more than 1 hour to find data, and not more than 2 hours to convert the data to a .db file that can be accessed by sqlite. If you don't have a dataset in mind, use one of the datasets provided at the bottom of this document - you must answer one of the questions associated with the chosen dataset.

SQL REQUIREMENTS

You are required to import your static files (i.e. .csv, .json, .txt) files into a sqlite3 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 this project is to query a SQL database and perform hypothesis testing on the collected data. You will come up with 3-4 separate hypotheses to test (each test consisting of a clearly identified null and alternative hypothesis). Be sure to explain what test (e.g. one-tailed t-test) you are using and why.

If you use one of the provided datasets, then you will have to answer 1 of the provided questions, then come up with 3 more tests of your own. If you use your own dataset, then you will have to come up with 4 hypothesis tests.

During the day 1 coach check in at 2PM, be prepared to share and discuss your chosen hypothesis tests. Have clear null and alternative hypotheses for each test.

STAKEHOLDERS

The use of any dataset brings with it a question of who your audience is for this data science 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 3 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. 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.
    b. The notebook should be well organized, easy to follow, and code is commented where appropriate.
    c. 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

- **Grades:** University of Wisconsin, Madison
  https://www.kaggle.com/Madgrades/uw-madison-courses
    - Do STEM fields have a statistically significant 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?