TO: 111819 Data Science Cohort
FROM: DS Ed Team
DATE: December 30, 2019
PRESENTATIONS : 1pm, January 3rd, 2020
SUBJECT: Module 3 Project Guidelines

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.

STATISTICAL ANALYSIS REQUIREMENTS

For this project, you will need data to perform a statistical analysis. In the statistical analysis, you'll need to perform hypothesis tests to answer at least three questions from the data source you choose.

For each hypothesis, be sure to specify both the null hypothesis and the alternative hypothesis for your question. Also, describe what statistical test you will use to test the hypothesis. (independent t-test, dependent t-test, ANOVA etc.).

Example:
Data Source : King County Assessor Data

https://info.kingcounty.gov/assessor/DataDownload/default.aspx

- H_0 = Housing market is not “hotter” during the spring and summer than the fall and winter

  H_a = There is a difference between the spring/summer market and fall/winter market
- Suppose that you have just relocated to King County and you are in the market for a home. Would it be better to buy now, or would we expect to get a better deal if we wait until fall?
- How would you go about testing this?

DATA/DATABASE REQUIREMENTS

You will identify a data source from which you will collect data to formulate questions about. Using web-scraping or calling an API (or a combination) you will get your data and store it in a SQL database. Try to clean your data **before** it goes into your database - this will make your life easier. Once your data is in your database, you can query it into Pandas to perform your analysis.

FALLBACK

If you are unable to successfully retrieve your own data 48 hours after project launch, we have provided a few datasets (at the end of this document) that you can use instead of finding your own. We highly encourage you to find your own data as it makes for a more interesting analysis and therefore a strong project for your portfolio.

STAKEHOLDERS

The use of alternative datasets brings with it a question of who your audience is for this data science project. Much like the Module 1 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. All functions have docstrings that act as professional-quality documentation.
    b. 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.
    c. The notebook should be well organized, easy to follow, and code is commented where appropriate.
    d. 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* that explains the hypothesis tests you answered, your findings, and their relevance to the company/stakeholders.
    a. Make sure to also add and commit a pdf copy 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.

Please add your projects to our cohort's project spreadsheet .

PAST PROJECTS

https://docs.google.com/spreadsheets/d/1E6yV7UK6pJ3NpPXX_3bywWVDuK4exkEjzErzXRolFeU/edit#gid=393293709

ALTERNATIVE DATABASES

- **Housing:** King County Assessor Data
  https://info.kingcounty.gov/assessor/DataDownload/default.aspx
    - We have a hypothesis that the housing market is hotter in spring and summer than in fall and winter. Suppose that you have just relocated to Seattle and you are in the market for a home. Would it be better to buy now, or would we expect to get a better deal if we wait until fall?
- **Grades:** University of Wisconsin, Madison
  https://www.kaggle.com/Madgrades/uw-madison-courses
    - Does your teacher have a statistically significant correlation with the number of As earned in a course?
    - Does time of day have a statistically significant correlation with the number of As earned in a course?
    - 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?