TO: 010620 Data Science Cohort DATE: February 27, 2020

SUBJECT: Module 4 Project Guidelines

PROJECT GOAL

The goal of this project is to be able to utilize Regression modeling 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 stakeholder.

Choosing your data

In this project, you are free to choose any dataset you like that would enable you to build a predictive model regarding information your company/stakeholder is interested in.

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)

Modeling Requirements

The goal of this project is to have you complete a very common real-world predictive task in regard to Regression. However, real world problems often come with a significant degree of ambiguity, which requires you to use your knowledge of statistics and data science to think critically about and answer. You are required to use either regression or classification to build your model, however, you may try to answer more than one overarching question which would require you to use both modeling tools. The following must be included:

- Assumptions of the model checked
- Model validation
- Model comparisons to the other models used
- Interpretation of model coefficients and parameters
- Predictions

Model/Metric Visualizations

Regression is an area of data science that lends well to intuitive data visualizations. Any findings worth mentioning in this problem are probably also worth visualizing. Your notebook should make use of data visualizations as appropriate to make your findings obvious to any readers.

Also, remember that if a visualization is worth creating, then it's also worth taking the extra few minutes to make sure that it is easily understandable and well-formatted. When creating visualizations, make sure that they have:

- A title
- Clearly labeled X and Y axes, with appropriate scale for each
- A legend, when necessary
- No overlapping text that makes it hard to read
- An intelligent use of color--multiple lines should have different colors and/or symbols to make them easily differentiable to the eye (please, no rainbow color scheme), color should be used to represent something!
- An appropriate amount of information--avoid creating graphs that are "too busy"--for instance, don't create a line graph with 25 different lines on it

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.