We have gotten a large amount of practice over the last two weeks. We’ve been working through data sets, and we’ve been working through the Data Science Libraries to understand how to apply what we’ve learned about Stats to our Data Sets. For the Final Project, I’ll be asking you to go through a Data Set on your own. We are taking the training wheels off so to speak. You will be expected to submit your assignment next Monday. You are expected to work on your Data Set alone.

I will outline the deliverables that will be expected of you.

1. Project Proposal - This should be a formal written proposal that outlines what your data is, and the questions/understanding you wish to gain from your data set. It should also include a reason as to why you chose this data set. It’s suggested that you approach Data from a question/answer standpoint. What questions do I have from my data set, and how can I extract answers, or a greater understanding from my Data Set? Do not start your project until your proposal has been approved by a TKH Staff member. Aim to get this in by tomorrow.
2. Jupyter Notebook File - This should contain all your code as you work through your project. Your file should be easy to read and work through. The expectation now is that your code is CLEARLY COMMENTED AND SEGMENTED. Each cell should be self contained, and each line/cell should be commented( If applicable) If I cannot easily read through your code, you will have points deducted. The only requirement here is that you at least include 3 visualizations that have been created with PURPOSE.
3. Written Project Wrap Up - This should be a written summary of what you learned from the project, and what you learned from your data set. Were your questions answered? Were you able to find some extra information about your data set? Was there a new function or method you discovered that proved helpful for your project? Where did you struggle? This should be a 1 page minimum.
4. Written Script - This should be a script that you should be able to present from. You should be able to use the script to explain to a potential client or employer about your project. It should cover the entirety of your project, and it should be easy and clear to follow.
5. In class presentations. On the 21st you will all be presenting. Maximum of 10 minutes to talk. The presentation will be recorded.

Below will be a high level overview of approximately how long each task should take and an overview of how to spend your time/days. Be prepared to spend longer than the expected time on certain tasks if you are not comfortable with them yet.

**Block 1**

**Task : Finding a data set and writing a proposal**

*Expected Time to Complete. 2-3 Hours.*

Find A Data Set that interests you.

Write a proposal as to why you choose this data set.

In the proposal, write down questions you want answered or explored by your dataset.

Explain why these three questions are important to the dataset, or what implications that they have.

**Block 2**

**Task : Explore the data and build out a checklist.**

*Expected Time to Complete. 2-3 Hours.*

Working through the data set and making sure we are using high quality EFFECTIVE code to sift through our Data Set.

Ensure we are commenting and that we are writing down our thoughts as we work through the set.

Ensure that we are building a CHECKLIST as to what we should be cleaning from our data set, what subframes we are going to need, and what the order of the next steps will be.

**Block 3**

**Task : Cleaning the data.**

*Expected Time to Complete. 5 Hours.*

Cleaning our Data Set.

Ensure that we are using high quality effective code, and that we are commenting along the way.

**Block 4**

**Task : Creating visualizations and interpreting the visuals and exploring new data.**

*Expected Time to Complete. 5 Hours.*

Building out visualizations that help us understand the data, and get a feel for what it means.

Ensure that we are writing down our thoughts and our analysis of the Data/Visualizations.

**Block 5**

**Task : Writing a formal paper of what we learned in terms of Data Science/Python/Pandas and what we learned about our data.**

*Expected Time to Complete. 4 Hours.*

Writing a formal explanation of what new data we got, and how we got it, as well as explaining what this new data suggests.

Was it able to answer our previous questions?

If it did, explain and work through the questions and the new understanding gained from the project.

If not, explain why the data set wasn’t appropriate, or the mistakes you made along the way.

**Block 6**

**Task : Writing a script that helps us present our project.**

*Expected Time to Complete. 3 Hours.*

Write a script that will help you present your project. This project is intended to go on your resume, and you should be prepared to speak about this project to a potential employer. While there isn’t a minimum or maximum length, ensure you are covering topics such as, what methods you used, any additional libraries you used, any additional information you needed to pull, and new methods or functions you learned and how you implemented them, as well as how your understanding of the data grew.

**Block 7**

**Task : Reviewing your project and practicing your presentation.**

*Expected Time to Complete. N/A*

Go over your assignment and make sure that you’ve completed all the requirements, and that everything has been polished.

Expected Time spent working on the Project.

Minimum: 25 hours

Make sure you can set aside at least this much time this week. If you feel like this project might take you longer than this, ensure you are setting up a plan, giving yourself enough time each day to complete the assignment. Ensure that you are planning for office hours if you need, Tutoring hours if needed, group work time if needed ( I encourage you to cooperate with your classmates to share understanding of the libraries).

This is a large assignment, but we will give you as much time during the week in the classroom to work on it. You will BE REQUIRED TO SPEND ADDITIONAL TIME OUTSIDE OF CLASS ON THIS PROJECT. PLAN FOR THIS. If you cannot spend this time over the next week on this, please message me NOW.

**Notes for project approval.**

The Data Set should not be 100% Clean.

The Data Set should have at least 10 columns and 100 rows.

Data set should be raw. It should not have extrapolated columns already created.

This should be a data set you have not touched before.

A maximum of 2 people will be able to work on the same data set. Not together.

Data set should be a topic that you care about or are interested in. You will be working for hours on this project, make sure you can keep it exciting.

Full sentences and full ideas should be expressed in your proposal, not a series of run-on sentences or bullet points.

I will have open office hours to accommodate this project all of this week from 2-5PM Tuesday and Wednesday. From 12-5PM on Thursday, and 12-6PM on Friday.

Please schedule time at least 1 day beforehand. (Except for tomorrow).

You can sign up for multiple time slots, but only 1 hour continuously.

Please come with specific questions or specific techniques you need help with, do not come with general questions.