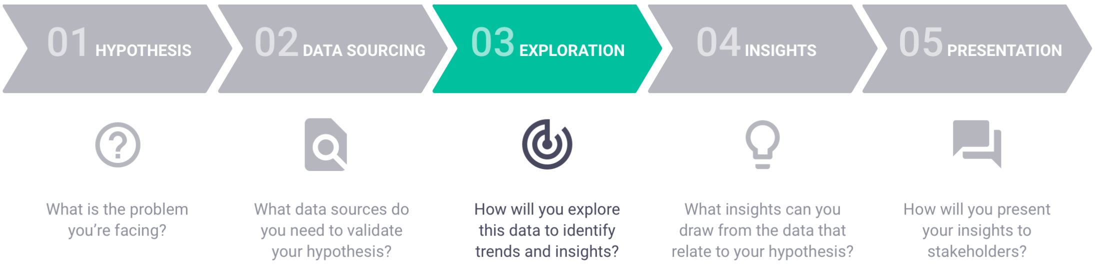
Data Analytics Career Track

Capstone Project Two: Data Exploration

Estimated time: **15 - 30 Hours**



01 HYPOTHESIS
What is the problem you're facing?
02 DATA SOURCING
What data sources do you need to validate your hypothesis?
03 EXPLORATION
How will you explore this data to identify trends and insights?
04 INSIGHTS
What insights can you draw from the data that relate to your hypothesis?
05 PRESENTATION
How will you present your insights to stakeholders?

Project Steps

Step One: Set up your analysis and Jupyter Notebook

1-2 Hour

Using the dataset that's been approved by your mentor,

a. Import the Python libraries needed to complete your analysis
b. Import your dataset into a Jupyter Notebook
c. Create a mark-down statement called **Data Cleansing**. This will be the section that includes any data cleansing and data preparation activities you complete for your analysis. You don't need to do the data cleansing now — we will cover that in Step Two.
d. Create multiple mark-down statements in your Jupyter Notebook that identify the issues you are analyzing. These issues should be derived from your issue tree.
e. Create a section in your Jupyter Notebook titled **"Insights"** – this will be where your insights will be shown

Step Two: Data cleansing

3-6 hours

It is likely your dataset isn't completely clean. By clean, we mean the dataset has formatting or potentially other issues that prevent you from using the dataset immediately without 'cleaning' it for use. If this is the case, you will have to make use of Python's data cleansing capabilities by creating functions, or using existing Python functions to help you cleanse your data.

Since every project is different, we recommend that you discuss the best way to clean your data during your next mentor call. The follow functions can be used for cleaning data:

1. Loops (Iterators)
2. Panda Operations (i.e. pd.rolling functions)
3. Lambda Functions
4. Conditional Logic (i.e. IF statements)

Step Three: Analyse your data (Exploratory Data Analysis)

10 - 20 Hours

Once you've set up your Jupyter Notebook and cleaned your data, start the analysis of your data in Python.

Important note: Be sure to analyze your data based on the issues you identified when creating your issue tree. All of your analysis should relate back to the issues and hypotheses you've identified.

Some tips to keep in mind:

- It's important that you get into the habit of leaving comments in your code to ensure that others can follow along with the analysis you've done.
- If you find yourself repeating operations, consider writing or finding a function that can do this operation for you.
  - The Python function declaration is shown below:
    i.e. Def FunctionA (Variable A):
    ….#Do Something
    To refresh your understanding of function definition, check out [this tutorial.]

- **Use Matplotlib or Seaborn to complete your work**
  While we do not require you to use specific descriptive or inferential statistics methodologies, please make sure to use Matplotlib or Seaborn to create your visualizations.

- Use these questions to guide your analysis
  1. What are the main issues I'm exploring?
  2. What is the best way to visualize this information? *(For example, if you are analyzing a large amount of time-series data, consider using a box plot to show the movements in variables instead of a line plot.)*

Step Four: Discuss your analysis

1 - 2 Hours

Show your analysis to your mentor during your call and ask for suggestions about what you can improve or continue doing. Please submit a link to your Jupyter Notebook once you have completed your analysis.