Data Engineering Bootcamp
Open-ended Capstone Step 4: Data Exploration

Estimated Time: 6-9 Hours

Now that you've obtained the dataset, you need to clean, enrich, and transform it. Your dataset needs to be made ready for consumption by your clients or other members of the data team.

In this step, you will explore your data to understand how it is distributed. Use the lessons from this exploration to decide how you want to clean and split your data for efficient storage and querying. You must also create a data model that reduces data redundancy and is also efficient in terms of storage and querying.

Question To Answer In This Step:

1. Is the data homogenous in each column?
2. How do you anticipate this data will be used by data analysts and scientists downstream?
3. Does your answer to the last question give you an indication of how you can store the data for optimal querying speed and storage file compression?
4. What cleaning steps do you need to perform to make your dataset ready for consumption?
5. What wrangling steps do you need to perform to enrich your dataset with additional information?

Note: the work you do in this step will help you understand how you should perform the transformation and load step later in your ETL pipeline.

Deliverables:

1. A Jupyter Notebook demonstrating the steps that you performed in the exploration of the dataset
2. A write-up of the insights from data exploration
3. A write-up explaining how you plan to store the data for optimal query speed and compression of storage files
4. An Entity-Relationship diagram for your data model
5. Update the project's GitHub repo with the work you complete for this step of the project and update the README Markdown file

Slide deck creation:

Starting now, you'll need to create a slide deck and update it with the work you do in each step. The slide deck will be a part of your final submission. With this slide deck, you should be able to walk your mentor, or anyone else, through the development process of your project.

Your README should detail the *current state* of your capstone project and your rationale for it, while the slide deck should detail the *process* behind your project. The slide deck tells the story of how you got to the final state of your capstone, in a way that any engineer could follow.

From this step, you should include:

- Insights from your data exploration efforts and how they are influencing the direction/next steps of your project
- Any graphs, charts, or other visualizations you used in your exploration