Data Engineering Capstone Project – Week 3 Activities

Objectives

The target project activities should be:

1. *Continue With Selecting Dataset(s)*

As a team you should have at this point identified a broad topic for your project.

You may include further datasets as you go forward. However, at this point you should have decided on at least one dataset to analyse.

2. *Load and Explore your data with Python/Pandas*

Try to load your dataset or a subset of it into a Pandas DataFrame. Use Pandas to perform exploratory data analysis, while also using this as an opportunity to reinforce the Pandas/Python syntax covered during the module.

If you have stored your data in a database e.g. MongoDB, then you could try reading into a Pandas DataFrame directly from MongoDB.

Try to use Python/Pandas to perform any cleaning or shaping required on the data. This cleaned/shaped version of the data could then be loaded by Power BI for your final reporting.

Any cleaning, shaping or analysis done with Python/Pandas may form part of the project presentation, and so good use of markdown cells in notebooks is valuable.

You may also try to produce some simple plots/charts in the notebooks. This can be developed further as more Python plotting features are shown in later modules.

3. *Continue Hypothesis/question refinement*

As you continue with the project activity you can continue to refine and change the hypotheses and questions you are asking of the data. Commonly the Exploratory Data Anlysis done with Python and Power BI leads to further questions or changes to the original questions.

The final project presentation may include a description of the changes you made to these questions and why those changes happened. This is valuable insight into the EDA process and the learning process for the techs.