Data Science Career Track

Capstone Three: Data Wrangling & Exploratory Data Analysis

Data Wrangling

Estimated Time: 6-12 Hours

You're now in the data wrangling and EDA stage of your third capstone. In addition to the data wrangling steps applied in your previous capstone projects, you now need to address some unique characteristics related to the advanced nature of your third capstone project. The exact steps depend heavily on the type of data you're working with for this capstone project. Some examples by project type are listed below:

NLP: stemming, lemmatization, tokenization, stop word removal, frequency analysis.

Image processing: scale thresholding, applying filters, transformations, and segmentation.

Recommendation System: calculate sparsity, create Implicit rating data.

Network Analysis: create matrix representations of graphs.

Use the resources in the unit related to your project topic to guide you and ask your mentor if you need help determining which steps to take to get your data ready for EDA and modeling.

Exploratory Data Analysis

Estimated Time: 6-12 Hours

Remember the goal of EDA is to investigate the relationships between features and the relationship between the response variable and the features. Be thoughtful and creative about how to deepen your understanding of the data and inform the overall project goal with your findings.

Things you may want to evaluate include:

- Is the response variable unbalanced?
- What is the distribution of each of the features?
- Are there features are correlated with a particular response value?
- Are there collinear features in the data?
- Are there outliers?
- What are the seasonal or linear trends?

Ways to evaluate these include:

- Histograms or Distribution plots
- Histograms, Word Count- bigram, trigrams
- Scatter plots or Bi-plots
- Pearson correlation coefficients heat map
- Box plots
- Line plots

Please note that the time estimates associated with this step of your capstone are approximated — you may take more or less time based on the complexity of your data.