Data Science Career Track

Capstone Two: Data Wrangling

Project Steps

Estimated Time: 4-8 Hours

You’re now in the data wrangling stage of your second capstone. Use the outline below as a reminder of what steps to follow, but be aware that your exact steps may change based on the issues you surface about your particular dataset and project.

Need a data wrangling refresher? We encourage you to review the data wrangling work you did during the guided capstone earlier in this course.

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.

Data Wrangling

All of the following steps should be performed in a Jupyter notebook with adequate notation and structure so that your mentor can understand the process you followed.

- Data Collection
    - Goal: Organize your data to streamline the next steps of your capstone
    - Time estimate: 1-2 hours

        - Data loading
        - Data joining
        - *Hint: Data Collection will require the use of the* pandas *library, and functions like* read_csv(), *depending on the type of data you want to read in!*
        - *Hint: when adding one dataset to another, make sure you use the right function: you might want to merge, join, or concatenate.*
- Data Organization
    - Goal: Create a file structure and add your work to the GitHub repository you've created for this project.
    - Time estimate: 1-2 hours
        - File structure
        - GitHub
        - *Hint: the* glob *library could come in handy here…*
        - *Remind yourself of why GitHub is useful. What are the main motivations for making a GitHub repository?*
- Data Definition
    - Goal: Gain an understanding of your data features to inform the next steps of your project.
    - Time estimate: 1-2 hours
        - Column names
        - Data types
        - Description of the columns
        - Counts and percents unique values
        - Ranges of values
    - *Hint: here are some useful questions to ask yourself during this process:*
        - *Do your column names correspond to what those columns store?*
        - *Check the data types of your columns. Are they sensible?*
        - *Calculate summary statistics for each of your columns, such as mean, median, mode, standard deviation, range, and number of unique values. What does this tell you about your data? What do you now need to investigate?*
- Data Cleaning
    - Goal: Clean up the data in order to prepare it for the next steps of your project.
    - Time estimate: 1-2 hours

        - NA or missing values
        - Duplicates
    - *Hint: don't forget about the following awesome Python functions for data cleaning, which make life a whole lot easier:*
        - *loc[] - filter your data by label*
        - *iloc[] - filter your data by indexes*
        - *apply() - execute a function across an axis of a DataFrame*
        - *drop() - drop columns from a DataFrame*
        - *is_unique() - check if a column is a unique identifier*
        - *Series methods, such as str.contains(), which can be used to check if a certain substring occurs in a string of a Series, and str.extract(), which can be used to extract capture groups with a certain regex (or* regular expression*) pattern*
        - numPy *methods like .where(), to clean columns. Recall that such methods have the structure:* np.where(condition, then, else)
        - *DataFrame methods to check for null values, such as df.isnull().values.any()*

- And don't forget that even the best data scientists and programmers Google things every day. Start with the idea: what you want to do with your data? Break it down into bitesize steps. Then, if you're not sure about a certain step, look through the data wrangling resources included in the course or execute a well-written Google Search, looking at trusted resources like StackOverflow.

Student Example

Want some inspiration? Check out this data wrangling notebook from Springboard student Abbas Zaidi:

Example 1: Data Wrangling for Traffic Capstone