PROJECT: Data Wrangling with Pandas and Regex: Learning Objectives

In this Project, you will be exposed to the following core Data Wrangling steps using Python's Pandas and Regex modules:

- Data Acquisition
- Data Cleansing
- Data Understanding: Basics
- Data Manipulation

Each step consists of objectives to be covered in the Project.

Data Acquisition Objectives

- How to import data in different formats? (read_excel, read_csv)
- How to import multiples files for storage and access? (store filenames in array)
- How are they read into by Pandas? (DataFrame)
- How to have a peek at the data after import? (head/tail)

Methods used:

- read_excel
- df.columns
- df.loc
- df.head

Data Cleansing Objectives

- Check attributes of each file
- Identify data types
- Apply coercion if applicable
- Check for corrupt/incorrect data
    - Check for NA/missing data
    - Check for data consistency (e.g., GPA cannot be less than 0)
- Remove/replace corrupt data
- Identify duplicate data
- Identifying and removing outliers

Methods used:

- df.dtypes
- df.isnull
- df.fillna
- df.drop
- df.drop_duplicates
- df.sort_values
- df.append

Basic Data Understanding Objectives

- Summary Statistics
- Dimensionality

Methods used:
- df.describe
- df.shape

Data Manipulation Objectives

- Merge/Concatenate DataFrame
- Filter to subset the data
- Mapping to create a new attribute
- Incorporate the use of multiple functions
- Discretize data

Methods used:
- pd.merge
- pd.cut
- unique

Regular Expressions

- Use regular expressions to find/match specific content
- String manipulation via. substring and replace methods
- Combine with data transformation methods to find, filter and manipulate data

Methods used:
- re.search
- re.sub