2-2 Assignment: Merging Data with Python

Sarah Steinbaum

Southern New Hampshire University

DAT430: Leverage Data for Org Results

Dr. Matthew Keane

May 14, 2023

Python is a programming language that many data analysts use due to its many functions and abilities. Python can be utilized for many reasons, including merging data sets. The merge() function is similar to the ‘join’ function in SQL and can be applied in four different ways; inner, outer, left, and right. Left refers to the data frame that was imported first and right refers to the data frame that was imported second. This function allows separate datasets to come together to form one singular dataset. Python defaults to inner merge which only displays the common data between the two data sets on the column where the ‘on’ value exists. Left merge keeps all the data from the left data frame and the data between the two data frames on the column that has the ‘on’ value. Right merge keeps all the data from the right data frame and the data between the two data frames on the column that has the ‘on’ value. Outer merge displays all the data between the two data frames. The ‘NaaN’ value is displayed where there is no value.

In order to merge the datasets, they must first be imported. The data can be imported through ‘pandas’ as displayed below. Pandas is a library that is used for working on datasets and provides functions for data normalization, cleansing, visualization, merges, etc. I entitled the first dataset as df1 to represent data frame 1 and the second set as df2 to represent data frame 2. The files can be imported as a csv file. Be sure to use pd.read\_csv() and type the file name in quotation marks to ensure it will be properly imported.

A screenshot of a computer

Description automatically generated with medium confidence

Once the datasets are imported, the head() function can be used to display the first five rows. The head() function can display a particular number of rows if it is specified. For example, head(10) would display the first 10 rows. This portion of the two datasets is displayed below.

A screenshot of a computer

Description automatically generated with medium confidence

A screenshot of a computer

Description automatically generated with medium confidence

In order for the two datasets to join, a new name needs to be created as it is essentially a new dataset. Below, I have labeled the merged table as df3 to represent data frame 3. The column names for the dataset need to be displayed in quotation marks to ensure the data will display properly. There must also be a column for the datasets to merge on. In this scenario, I chose ‘id’ to be the column the datasets merge on. Then, I used the head() function to see the first five rows of the merged dataset.

A screenshot of a computer

Description automatically generated with medium confidence

Since the merge function uses the inner merge as default, the dataset currently only shows the data that both data sets have in common. We need to specify which merge type is needed. We do not need to lose any information, so we will use outer merge. ‘How’ can be used to signify which type of merge is needed as displayed below. By default, the letters x and y are used to differentiate the separate data frames. ‘Suffixes’ can be used to change x and y to a more desired title. I changed them to \_df1 to represent that the data is from data frame 1 and \_df2 to present data frame 2. ‘Indicator’ can be used to signify whether the data is from both data sets, right only, or left only.

A screenshot of a computer

Description automatically generated with medium confidence

References

S, S. (2022, August 9). *What is Pandas in python? everything you need to know*. ActiveState. https://www.activestate.com/resources/quick-reads/what-is-pandas-in-python-everything-you-need-to-know/

Lynn, S. (n.d.). Shane Lynn. https://www.shanelynn.ie/merge-join-dataframes-python-pandas-index-1/