Data Wrangling with Python

CONCATENATING AND MERGING DATA FROM DIFFERENT SOURCES



Pratheerth Padman FREELANCE DATA SCIENTIST

What Will We Focus On?



Data collection/gathering



Data preparation



Data analytics



Module Overview



Concatenating datasets

Merging datasets

Merge keys

"How" parameter

Disambiguating merged columns and updating dataframes



Software and Course Prerequisites





Basic to intermediate level of Python

Fundamental knowledge of the machine learning pipeline

Pandas

Jupyter Notebook



Concatenating Datasets



Data Wrangling

Data Wrangling is the process of transforming data from its raw form to another format with the purpose of making it more valuable and appropriate for data analytics and machine learning.

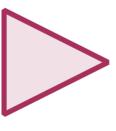


Concatenating Datasets

Concatenating datasets refers to the linking together of any two or more sets of data



The data need not be related



pd.concat()



pd.df.append()



Merging Datasets



Merging and Joining Datasets



Merge - combines dataframes by aligning on columns

The process of joining two dataframes together based on some common columns



Many types of joins/merges - inner, left, right and outer



pd.merge()



Categories of Joins / Merges

One-to-one joins

Many-to-one joins

Many-to-many joins



Merge Keys



Types of Merge Keys



On parameter – explicitly specify the name of the key column



Left_on, right_on parameters – for merging two datasets with different column names, but similar data



Left_index, right_index parameters – for merging on an index rather than a column

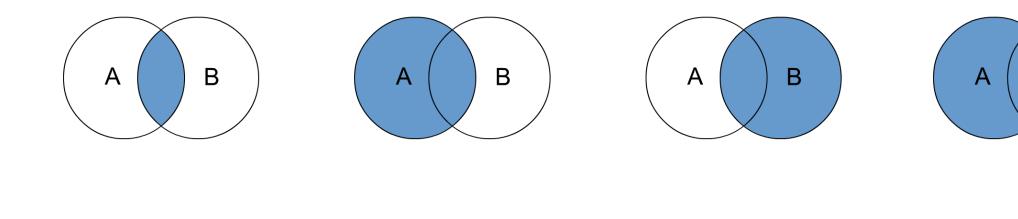


The "How" Parameter



How Parameter Values

Right join



Left join

Inner join

В

Outer join

Demo



Disambiguating merged columns and updating dataframes



Summary



Concatenating datasets - concat and append methods

Merging datasets - one to one, one to many and many to one joins

Merge keys - on, (left_on, right_on), (left_index, right_index) parameters

How parameter – inner, left, right and outer join

Disambiguation of merged columns and updating dataframes

