# Datamap for cleaning data downloaded from internet:

**Input Folder**. Change names of files to batch\_\* for data from api, for processing in python. net\_data\_\* for data from website. \* stands for nos 1, 2, 3, n so on.

Data downloaded from api or web scraping or website.Original raw files folder.

c

Columns:

date

Year

Month

Day

SVI

batch\_number

**Output Folder**. The files processed from python code will be saved in this folder. Data structure of files saved here is as follows.

Concat all the datasets. Concat datasets initially have 2 parts.

Datasets 1 and 2 has to be handled differently as data structure of both the datasets are different.

Clean snp500 files starting from 31/12/2003 till today and create time\_series\_period\_week in the dataset

Select distinct year, month and day and create time\_series\_period\_week and week\_number file

Groupby keyword\_name, year, month and day and drop all duplicate columns to get cleaned dataset

Concat both the datasets to get final\_dataset file

Merge both the files with final\_dataset as and when required.

**Datasets** 1) net\_data.csv..this is combined datasets downloaded from websites.

**Datasets** 2) apidata.csv..this is combined datasets downloaded from api.