**1.Introduction:**

We found the data related to “US Supply Chain Information for COVID19” at this link: https://www.kaggle.com/skeller/us-supply-chain-information-for-covid19

Format: Datasets are available in the .txt and .csv file format

**Dataset 1:**

Name: cfs-2012-pumf-csv

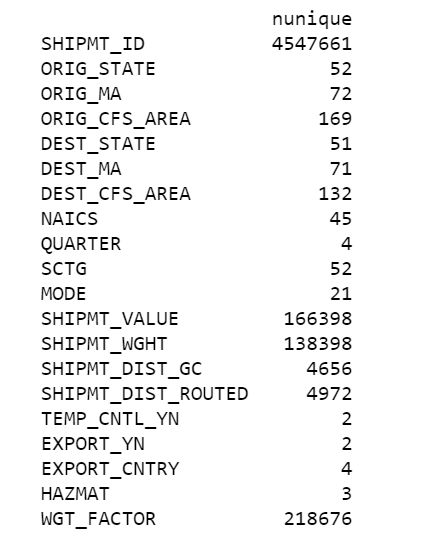
Fields: "'SHIPMT\_ID', 'ORIG\_STATE', 'ORIG\_MA', 'ORIG\_CFS\_AREA', 'DEST\_STATE',

'DEST\_MA', 'DEST\_CFS\_AREA', 'NAICS', 'QUARTER', 'SCTG', 'MODE',

'SHIPMT\_VALUE', 'SHIPMT\_WGHT', 'SHIPMT\_DIST\_GC', 'SHIPMT\_DIST\_ROUTED',

'TEMP\_CNTL\_YN', 'EXPORT\_YN', 'EXPORT\_CNTRY', 'HAZMAT', 'WGT\_FACTOR'

# Records: 4547661



Link: <https://www.kaggle.com/skeller/us-supply-chain-information-for-covid19?select=cfs-2012-pumf-csv>

**Data Exploration:**

The main transactional dataset consists of 20 columns and 4.5million records in total.

Out of 20 columns, 13 are quantitative types and rest 7 are character types.

There is a variable ‘SHIPMT\_VALUE’ which depicts about the value of a shipment i.e. its price

The first variable SHIPMT\_ID does not have any duplicate records so it can be treated as an index and each id can be related to a shipment

There are no missing values in any of the columns in the dataset

There are extreme values in the some of the quantitative variables such as SHIPMT\_VALUE

All the data combined and analysed well tell us regarding the shipment value of a good or a commodity. We will create dummy variables for certain columns

Yes, it can be considered as a big dataset as the total records are around 4.5 million.

Our data set is both huge and capable enough to give us insights for the supply chain information of critical goods.