**Major Data Issues:**

* First, there are three datasets which are not easy to analyze the whole information.
* Second, there are some issues in the datasets, such as wrong data type, missing values, outliers, and duplicated rows.
* Third, some of the variables must connect to other related variables to give some useful ideas to the analysis.

**Solutions:**

* **Corrected the data types of each variables when we imported the dataset into the SAS Enterprise**

There are some variables in the dataset with wrong variable type, such as “Service\_From\_Date”, “Service\_To\_Date” should be in Date format but in String. We matched them in the correct data types when we import the dataset into the SAS Enterprise.

* **Detected & deleted outliers**
  + Deleted the negative values of the “AvgWeeklyWage”, and made the range between 0-1202.
  + There are lots of values of “ClaimantAge\_at\_DOI” out the rage of the 18-78, which doesn’t make sense, and should be outliers. Deleted the values of Age which are less than 18 or higher than 78.
* **Deleted or replaced missing values**
  + Replaced the missing values with the mean of the “AvgWeeklyWage”, “ClaimantAge\_at\_DOI”.
  + Take out the missing values from “ClaimantClosedDate” since they are treated as the claims are still processing, which are not considered in our analysis.
  + The small missing values are taken from “Serve\_to\_Date” and “Serve\_from\_Date”.
* **Appended the Transactions 1 & Transactions 2**

Transactions 1 and Transactions 2 with the same variables are recorded to show the same information but for different transactions. To get a whole idea about the relationship between the claims and the transactions, we combined them into a new data set – Transactions.

* **Calculated new columns which can be used to measure the processing time.**
  + **“Total\_days\_service” (“ServiceToDate – ServiceFromDate)**
* **Aggregated the PaymentAmount and Total\_days\_service by the ClaimIdentifier to get two new columns which can be joined to the CliamsData.**
* **Deleted the duplication rows by the ClaimIdentifiers in the Transactions**
* **Got the extended Claim dataset:**

Finally merged the two new variables (Total\_Payment\_Amount & Total\_Days\_Service) and the “IsLitigated” variable to the cleaned claim dataset.

* The reason we add those three variables is that total days of service can tell the efficiency of the service and transactions. And sum of the payment amount can show the total amount of transaction for each person. Whether is litigated or not, can be compared with the amount not litigated or other amount of payment being paid.