* Match the data characterizations
* Take out the negative values of the “AvgWeeklyWage”, and make the range between 0-1202, because anything above is the outlier. And also replaced the missing values with the mean, because there are a lot of null values.
* Make the range of “ClaimantAge\_at\_DOI” as 18-78, since 78 above is outlier, and age of 18-78 of make sense for working age, anything out of this range, might be mistyped. Missing values are replaced by mean values as well, because of the huge amount of null values.
* Take out the missing values from “ClaimantClosedDate” since they are treated as the claims are still processing.
* The Transaction\_1 and Transaction\_2 datasets are appended.
* The small missing values are taken from “Serve\_to\_Date” and “Serve\_from\_Date”.
* We created two new variables that are “Total\_days\_service” (“ServiceToDate – ServiceFromDate and “Sum\_of\_PaymentAmount” sum them both by group ClaimantIdentifier.
* Then reduce the duplicate by the ClaimantIdentifier.
* Finally add the two new variables 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.