**Insurance Premium Prediction System**

**Data Wrangling**

Data wrangling refers to the process of cleaning, restructuring and enriching the raw data available into a more usable format.

**Steps followed for Data Wrangling:**

* Identify the vague values “NaN', 'N/A', '0', '0.01', '9999', '9999.99', '999999” and treat them as null value
* Drop duplicate rows in the dataset
* Replaces all the null values with 0
* Update Tobacco column values with numeric values such as “ No Preference “ with 0 and Tobacco User/Non-Tobacco User with 1
* “RatingAreaId” column has redundant value so remove the prefix “Rating Area” from all the rows using Regular Expression.
* Age is numeric field but it has non numeric value like 0-20 , Family Option and ’65 and over’ ,so replaced it with 20,101 and 65 respectively
* The name of the columns is very long and difficult to understand. So replaced the column names with more meaning full name
* Replaced state code with numeric value
* Merged 2 tables and created a third table so that we can identify which plan is dental or which is health plan only. This gives us clear idea to understand and decipher correct details for premium prediction.
* Create a new dataframe which contains only required column for data exploration

**Result Dataset (Sample) :**

