**SUMMARY**

After visualization we start with data reduction and data dimension, luckily, we don’t have missing values but have many categorical variables. We went through various articles to see what factors/determinants to look for based on which a customer decides to buy an insurance. The idea was to move in a similar direction with our data set based.

Furthermore, we are assuming that certain variables would play an important role in explaining why one group of customers to buy an insurance over the other. We had 10 main categories and 44 sub-categories. The idea was to blend 44 sub-categories in to main categories so that when we look at a certain main category, the sub-category would act as a characteristic. This will help us know which group of people are more likely to buy insurance as compared to others.

Finally, we did a correlation test do make sure they are not highly correlated with Response variable following are results.

Graphical user interface

Description automatically generated with medium confidence

Correlation gives us a good indication which we could use to assess the strength of relationship. Had they been correlated, it would have been an issue. Since we have several categorical data, we will be converting the variables which we need into dummies.

For number of houses we created 4 dummies

1. HasOneHouse
2. hasTwoHouses
3. has4Or5Houses
4. moreThan5Houses

For age we divided into

1. earlyAge
2. middleAge
3. oldAge.

We are still working on income variable.

A histogram of categorization of subcategories into 1 main category

Chart, bar chart

Description automatically generated

Our data set targets most to the grownups.