**Caravan Insurance (Group 8 ) Karim and Mustafa**

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 article and see what factors to look for customer to see if he would buy a insurance or not then we identify similar variables in our dataset, for now

we are assuming following variable would play important roles, and explain, number of houses, customer type, we had 40 sub type of customer we categorized them using pivot table in excel that categorized sub types into 1 customer type this will indicate help us to know which group of people are more likely to buy insurance as compared to other, average household how many people living in house on average, age, income. then we did correlation test do make sure they are not highly correlated with Response variable following are results.

#P3 = MGEMOMV (Avg size household) 0.05

#P4 = MGEMLEEF (Average age) 0.00

#P5 = MOSHOOFD (Customer Main Type) -0.06

#P2 = MAANTHUI (Number of houses) 0.00

Correlation gives us good indication that we could use these variables, then we start creating dummies of variables, for number of houses we created 4 dummies, hasOneHouse, hasTwoHouses, has4Or5Houses, moreThan5Houses, for age we divided into earlyAge, middleAge, 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.