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

Summary:

Data set consists of 86 variables, from which 43 variables are demographic and rest are product usage related i.e number of car, delivery van, motorcycle policies. Our interested variable is number of mobile home policies which indicates no of people who bought the car insurance. Dataset has 5822 train observations, and 4000 test observations, we will be randomizing training set again and again so we can come up with best set that doesn’t have biasness in it and we will plot ggplot, boxplot, histogram to analyze distribution and get action able insights as mentioned in the class. Our goal would be to identify variables which directs effect our response variable (Y).

**Questions:**

* We have 86 variables do we have to use all of them? Can we group similar variables into one?
* How can we select good predictors from above all variables?
* How would we determine which supervised learning algorithm we have to implement? ( what are the factors we take into consideration )
* How to rank our predictor variables is it something like the one who is highly correlated with dependent variable has highest rank?

**Article we went through:**

<https://www.agilehealthinsurance.com/health-insurance-learning-center/what-should-i-look-for-when-buying-health-insurance>

https://www.forbes.com/advisor/life-insurance/best-tips-first-time-buyers/

What have we learned?

* Always look for deductible and network coverage before buying.
* Compare provided rates with other insurance
* It’s cheaper to stick with one insurance plan that offers combine coverages i.e ( car and rental )
* Asses your budget are you able to stick with insurance for 1 year at least?

We started working on R code with Data Visualization and exploration as first step

Table

Description automatically generated

Chart, bar chart

Description automatically generated