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

**Summary:**

Our selected data is meticulous and we expect to run a few research questions to get there answers. The data set consists of 86 variables. Our interested variable is number of mobile home policies which indicates number of people who bought the car insurance. Dataset has roughly 9820+ observations.

The way the observations have been recorded are through indexing, which means that each index number represents a certain percentage of a variable or a value.

Our goal would be to identify variables and their characteristics that determines if person would buy insurance or not. In order to find this, we will be randomizing training set again and again so we can come up with best training dataset that doesn’t have biasness in it and we will plot ggplot, boxplot, histogram to analyze distribution and get actionable insights as mentioned in class.

**Base 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?

**Research Questions**

* Who is more likely to buy insurance?
* What are the characteristics of those buyers?

**Optional Research Question**

* Which type of buyer is likely to buy which type of insurance?

**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?

**Missing Values?**

Our dataset does not contain any missing values.

At places we see 0 which represents a certain value. It does not mean that we do not have data for that particular record.

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

Table

Description automatically generated

Chart, bar chart

Description automatically generated

Chart, histogram

Description automatically generated

Graphical user interface, application

Description automatically generated with medium confidence

Code can be found here for better traceability https://github.com/karimkhamwani/data-mining