Group 8.

Caravan Insurance plan:

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 ggplot 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 )

Article we went through <https://wallethub.com/edu/ci/factors-that-affect-car-insurance-rates/69302>

What have we learned?

* Age, education and gender are key factor to calculate the insurance rate, upon calculation of rate we determine if user would buy insurance or not so we will focus on these variables from our dataset as well.

Please just write a paragraph on:

"Topic of the project and checking out datasets"

Your update should include:

- You need to post at least two questions (more questions will be beneficial in the future).

- Check one or two articles (if you want more) about the topic. Please provide the references.

- Check your dataset, start with the first steps in data exploration: missing values? outliers?  Visualization will help you.

This is the time you ask questions, please include them in your update.