1. **Have you chosen your variables? Any challenges in data dimension?**

As our data set has 86 variables, 85 of those are categorical and 2 of them are numeric. Our first instinct was to apply PCA but much to our dismay we could not do it because our dataset was not numeric. We thought about going the longer route by converting variables of interest into binary and then applying PCA but the problem we encountered was that we would little room to explain our findings as the scope of binary was limiting us.

So then the challenge, was to select a distinct number of variables that could explain or shed some light on our research question, \*who is more like to buy insurance?\*. The problem we encountered was that not only did we have main categories of customers but rather sub-categories as well. So instead of taking all of those categories into account we grouped sub-categories into main categories.

In order to accomplish this we used the help of excel and pivot tables to determine which sub-category belonged to which main category. To our surprise, our grouping was surprisingly accurate. Hence, we now have 10 main categories which consists of their respective subcategories. This helps us with our 2nd research question, \*what are the characteristics of those likely to buy insurance?\*.

In order to find answers to these questions, we have identified the following variables of interest:

1. One
2. Two
3. Three
4. Four
5. Five
6. Six
7. Seven
8. Eight
9. Nine
10. Ten
11. **Have you chosen the technique (data mining task)?**

As mentioned above since our dataset is categorical, applying data mining techniques related to numerical data would be a total disaster. Hence, we have applied the basic rules of data mining which are as follows:

1. **Descriptive Analytics**
2. **ELT [Extract, Load, Transform] / Cleaning**
   1. Duplicate identification
   2. Missing values identification
   3. Visualization
      1. Bar Charts – because categorical data is best explained by it
      2. Histograms
      3. Boxplots
   4. Outlier identification
3. **Regression Analysis – Logit/Probit????**
4. **Have you partitioned the data?**

Yes, we have partitioned our data with a 60% in testing and 40% in validation phase. We have utilized the library “**CaTools”** to perform the sample split quickly.

We have run the partitioned with the approach that it takes into account outliers and the other doesn’t. That way we would compare the drift between two and see whether we could explain our findings with the outliers or without

1. **Have you started running the model?**
2. **Remember you need to need to apply TWO techniques.**
3. **Any questions?**
4. **Write a paragraph of the advances you have made for your project.**