# Challenges of Categorical Datasets

## Introduction

In addition to the numerical datasets used, a modeling process was performed where all of the variables were converted to the categorical datatypes. However, while this conversion is conceptually accurate, it caused major problems for the analysis. We will delve into these issues later.

First, the decision to convert every variable is due to the fact that, from a business and conceptual standpoint, the variables in this dataset are all nominal or ordinal in nature. This concept is very clear for some variables such as STYPE as well as everything starting with the letter “M” or “P”. However, even for the variables starting with the letter “A”, such as AWAPART (number of 3rd party insurance), it is conceptually better to convert these to an ordinal datatype. Looking at these variables, it’s not certain that there is a linear pattern. For example, we cannot say that having 2 third-party insurances is twice as impactful as only having 1 third-party insurance.

While this conversion makes sense, it made analysis and modeling very difficult by aggravating 3 major negative characteristics. These major characteristics in the dataset are:

1. Skew, or a serious imbalance in the target class variable, where the CARAVAN=T observations make up only 6% of the overall data. This characteristic is inherent in the original dataset, but it becomes even more glaring when we approach the modeling with categorical variables.

## Challenges

### Measuring Correlations

During the Exploratory Data Analysis of the modeling process it is common to check for correlations between the explanatory variables. When the dataset is numerical then this check will be looking for collinearity between the variables, or in other words, looking for a linear relationship between two numeric variables. However, if we have categorical variables then the concept of “linearity” doesn’t exist anymore and thus it doesn’t make sense to look for a linear relationship.

There are analogous techniques that we can use for categorical variables. The most common is to simply look at a contingency table between the two variables.