Dataset I am using is from Kaggle challenge Porto Seguro’s Safe Driver Prediction (<https://www.kaggle.com/c/porto-seguro-safe-driver-prediction)>. The competition provides a training dataset and test dataset.

#1 Read in training dataset through pd.read\_csv() and extract preliminary information, for example using head() to print out the first 5 columns, using shape to understand dataset size, and using info() to see what type of data each column holds and if there are any missing field.

#2 Double check for missing fields by running dropna with threshold of 1, and as expected, from looking at results of info() from #1, there are no missing fields

#3 Try to understand what each of the feature mean. The dataset on purpose did not give detailed column names, try to make sense of what the dataset is. First, pick out all the integer typed features, because they are either categorical or binary. Create a dictionary that maps column name to Counter dictionary, where the Counter dictionary will count how many times each value of a column appeared. Now iterate over each column, and pull out elements of each of column from training DataFrame and start counting how many times each value occurred.

#4 Now that the values have been counted, plot these values out. Iterate over each Counter and plot out its keys as x and values (count) as y in scatter plot. By looking at the charts, we can pick out what is binary feature and categorical feature.

#5 Notice that some charts have some sort of distribution. Mark them down, because these might be useful for future calculations.

#6 We want to do the same analysis for Test dataset, instead of rewriting everything, create a function that reuses code used for Training dataset. Run function on Test dataset and observe result.

#7 Compare scatter plots, it appears that all the plots seems to follow similar shape.