**Preprocessing – Statistics for Attributes**

By utilizing WEKA Explorer, we can use the preprocessing tool to develop insights for sample statistics on the 26 attributes from the dataset. The basic statistic for each attribute is provided with statistics such as minimum, maximum, mean, standard deviation, distinct values, and unique values. The basic statistics can be used for various reasons in cleaning and understanding each variable in the data. For instance, we can use the mean value for the attribute “COMMUTE\_DISTANCE” if the vehicle use is for commuting. Distinct and unique values will also assist us in creating insights as distinct values tell us how many different values there are in an attribute and unique means the number of records having a value for an attribute that no other record has in the data. As it is important to distinguish the types of attributes we are working with in the dataset, WEKA simplifies this process as it showcases the type of attribute in the preprocess tab in the basic statistics. There are 14 attributes that are of numeric type (represent a floating-point number), 11 attributes that are of nominal type (represent a fixed set of nominal values), and 1 string attribute (represent a dynamically expanding set of nominal values). From the basic statistics developed for each attribute, a few insights we could develop were for driver information with the average customer for CGIC being the age of 42, would have 20 years of being licensed and 7 years of being a principal driver. This allows us to understand that majority of the customers who have products insured with CGIC are older than younger customers. Furthermore, there is a higher demand for their products from customers who have been licensed for a longer time. In addition, younger drivers have a higher chance of driving recklessly than older drivers as the mean for CONVICTION\_COUNT\_MINOR\_3YRS is 0.165 whereas, the mean for CONVICTION\_COUNT\_MAJOR\_3YRS is only 0.004.

As we continue our data exploration, we utilize the visualize feature in WEKA to develop a plot matrix. As we examine the plot matrix, we can notice that class colors (blue and orange) showcase the output attributes. As the target attribute is IS\_BOUND, we are comparing each attribute with IS\_BOUND to determine any patterns between the attributes. The class color scale goes from 0 to 1, defining whether a quote will become bound if orange and not bound if blue. In addition, for the purpose of visually seeing the data more clearly, we use the jitter function, however, it does create artificial random noise to the coordinates of the plotted points. Although there is random noise, it does not affect the data quality as the jitter function can be reduced to display the exact plot points. From the 52 plot matrixes provided on WEKA, our first analysis will be between the attributes “AGE” and “ANNUAL\_KM” as this can help us determine a pattern or any outliers. Figure 1.1 showcases that majority of the sample population from ages 26 to 58 have annual kms from 6,696 to 23,156. This was calculated using the standard deviation and mean of the two attributes, as the standard deviation for Age is 16, we can say that it is 1 point away from the mean (42). By analyzing the plot matrix, we can assume that any plotted points over or near 250,000 kms are outliers and they can be discarded from our data as it is highly unrealistic for a vehicle to drive that much in a year.

![Graphical user interface, application

Description automatically generated]()