**Distribution of categorical variables:**

Let’s plot some visualizations to have a look of each of these binary and categorical variables. From the dataset we have the main ingredients of different candies from different manufacturers of candies. They are chocolate, fruity, caramel, nougat, peanutyalmondy and crispedricewafer. This JG Foods dataset contains 85rows and 12 columns in total in which column pluribus has 24 Null values and pricepercent has 3 Null values in total.

**A picture containing text, clipart

Description automatically generated**

*Fig 1: Distribution of Variables*

From Figure 1: Distribution of variables, by performing some summary statistics using functions value counts, info and describe on the dataset we can notice that 37 candy manufacturing companies out of 85 have chocolate as their main ingredient and 38 companies out of 85 has fruity as the ingredient in the candy. Also, 14 candy manufacturing companies have caramel and peanutyalmondy as their ingredient, while only 7 has nougat and crispedricewafer as their ingredient. From this we can interpret that chocolate and fruity are the two main ingredients that are used by candy manufacturing companies.

**Distribution of Target variable:**

**Chart, histogram

Description automatically generated**

The percentage of winpercent vary amongst brands, ranging from roughly 20 percent to 80 percent. Ideally, we need to know which qualities make a candy more likely to improve the winpercent to its competitors in a matchup.

**Correlation:**

A correlation matrix is generated to make a summary of a large amount of data with the purpose of identifying trends. Each cell in the table shows the correlation between two variables. This is important in fitting a model to the data, as it gives us information about the variables which might be useful in model-fitting.

*Chart, treemap chart

Description automatically generated*

*Fig 2: Correlation Matrix*

Correlation plot is plotted to visualize the strongest and weakest correlation coefficients between variables. From figure 2, Chocolate has a positive correlation coefficient of 0.6 to the bar, 0.64 to the winpercent, and 0.49 to the pricepercent. Bar has a positive correlation coefficient of 0.52 with Nougat and 0.5 with pricepercent variables. Also, Chocolate and Fruity has the highest negative correlation coefficient of -0.74 which means that the candy will have either chocolate or fruity as its main ingredient and not both of them. Also, pluribus has the second highest negative correlation coefficient of -0.59 with bar which means that many candies in a bag or box doesn’t go with bar type. Lastly, bar and fruity has a negative correlation coefficient of -0.52 which means that the fruity flavored candies may not go with bar type.

Further observations from the Correlation matrix are Sugar content is evenly distributed throughout all categories, with the exception of price (cheaper candies also have lower sugar levels). Chocolate is much more expensive than the average, but fruit-flavored products are less expensive. Chocolate is frequently available in bar form, with peanut/almond flavoring, a crisp texture, and nougat or caramel. Fruit-flavored candies, on the other hand, frequently fall outside of those categories.