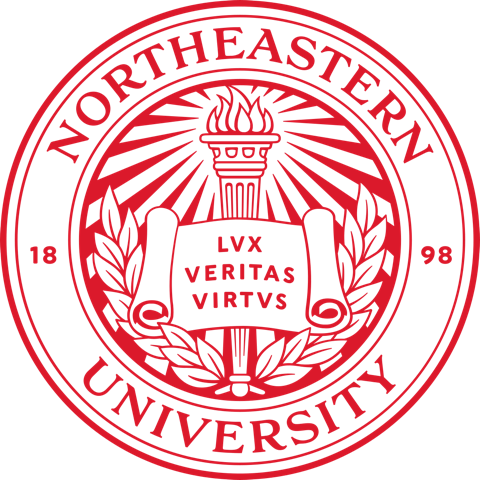
JG Foods - Halloween Candy Analysis

**Course:** ALY 6040

**Instructor’s Name:** Prof. Dr. Justin Grosz



**Submitted By:**

Sunil Raj Thota

Nissy Asish Gantepalli

**Date:** 10/17/2021

**Introduction:**

Every year, a large number of individuals celebrate Halloween. We choose to preserve so much of the candy bowl to ourselves when there are fewer trick-or-treaters. We are starting the year off well with a look at JG Foods Candy Business for this Halloween. Using the competitor's candy ingredients like chocolate, fruity, caramel, peanutalmondy, and bar are few examples.

We have then evaluated the data to assist JG Foods in making the most cost-effective confectionery purchases. Which Halloween candy is the finest (or at least the most famous)? Is it, for example, chocolaty? Is there any nougat here? How does it compare to other candies in terms of price? What percentage of individuals prefer this candy to another? As JG wants to make a new confection and wants to know what features it should contain. We'll look at this data to see what individuals are interested in and provide recommendations accordingly.

**Analysis:**

**Data Cleansing:**

We have imported the dataset into the Jupyter Notebook environment and performed our analysis. The dataset consists of 85 records with 13 attributes. We have done the column level datatype information and observed that Pluribus has 24 null values and pricepercent has 3 null values. As Pluribus is not providing any useful insights, we have decided to drop that column from the data frame. We have used data imputation techniques on the pricepercent and filled the null values with the median of the pricepercent attribute.

**Exploratory Data Analysis:**

1. **Distribution of categorical variables:**

Let’s plot some visualizations to have a look at 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.

1. **Distribution of Target variable:**

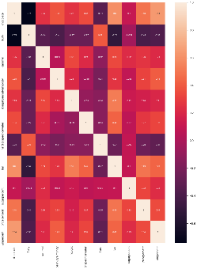
**Chart, histogram

Description automatically generated**

The percentage of winpercent varies 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.

1. **Correlation:**

A correlation matrix is generated to make a summary of a large amount of data to identify 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.



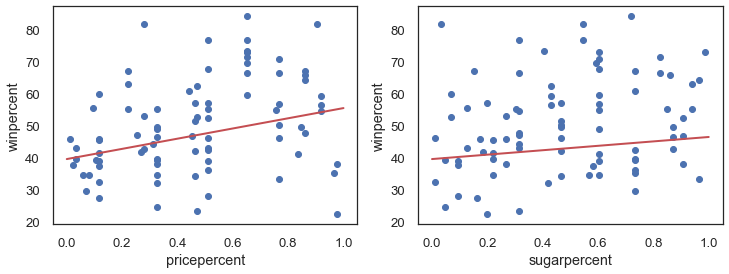
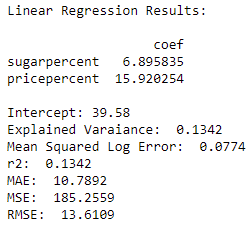
*Fig 2: Correlation Matrix*

A 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. The bar has a positive correlation coefficient of 0.52 with Nougat and 0.5 with pricepercent variables. Also, Chocolate and Fruity have 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 don’t go with bar type. Lastly, bar and fruity have a negative correlation coefficient of -0.52 which means that the fruity flavored candies may not go with bar type.

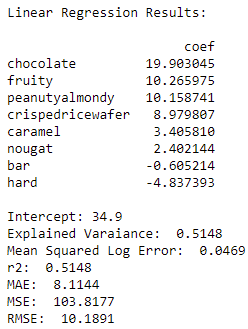
**Data Modelling:**

Let us see the relationship between the winpercent and other attributes

1. **Linear Regression:**

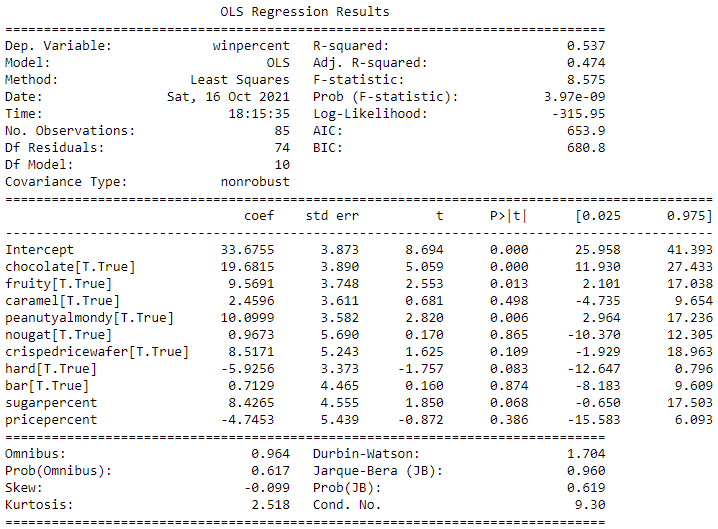
 

From the above results, we can clearly state that there is no link between winpercent and sugarpercent, nor between winpercent and pricepercent. Sugarpercent and pricepercent can be ignored.



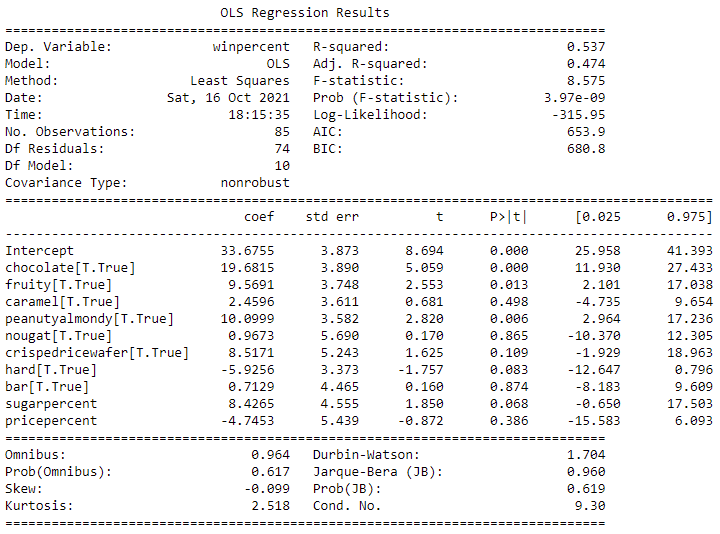
From the above results, all variables were subjected to linear regression, except for sugar and price. Overall, chocolate increases winpercentage by 20%. Because peanuts, caramel, nougat, and other candies get a beneficial impact, hybrid candy should be given a higher score. Fruit is chosen over chocolate. Hard type is selected over bars. Chocolate in the form of a bar, possibly containing peanuts/almonds, crisped rice/wafer, caramel, or nougat, is the greatest candy. Also, we have obtained a perfect candy score of 74.9% by adding the intercept and respective coefficient values.

1. **OLS Regression Analysis:**



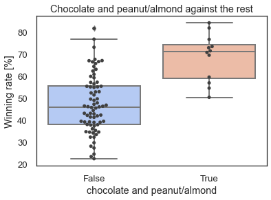
We have used statmodels API to perform the OLS regression analysis on the winpercent as target and the rest of the attributes without competitorname attributes. We were able to accomplish an R squared adjusted of 0.537. A few of the characteristics are incapable of contributing to the explanation of the dependent variable's variation. Caramel, nougat, and bar, for example, have p-values near 50 percent, whereas pricepercent has a p-value of 0.386. With a coefficient of around 20 and a standard error of 3.89, chocolate appears to be the most relevant component. Fruity, peanutyalmondy, crispedricewafer, hard, and sugarpercentage all appear to be significant. The discovery of a positive coefficient for the fruity feature was the biggest surprise. Fruity, in this case, also implies no chocolate from the correlation matrix.

However, both characteristics have a beneficial impact, the chocolate factor is twice as large, therefore we notice the significant impact of this attribute but not the significant improvement of chocolate if we choose fruity candies, and the overall impact turns unfavorable.



We experimented with several different regression equation parameters. It was accomplished by deleting different features and combining these to see if the magnitude and standard error of the coefficients changed significantly. ln consideration of adjusted R-squared, the optimum solution (best fit). Finally, we selected winpercent chocolate + fruity + chocolate peanutyalmondy + chocolate crispedricewafer + fruity hard + highsugar, which resulted in an adjusted R-squared of 0.557. To summarize, chocolate candy with peanuts, peanut butter, or almonds is likely to outperform the average. Finally, it appears that high sugar content is significant. However, when the three features of chocolate with peanut/almond incorporated and a high quantity of sugar is taken into account, the winning rate rises even higher.

If the product comprises chocolate and peanutyalmondy, count and averages the winning rate. If the product has all three of the most significant features, count and averages the winning rate.



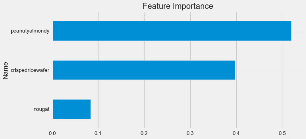
Various criteria of the top and the worst functioning candies' recurrence. The best-performing candy is typically in the shape of bars. Chocolate is much more commonly found in the best-selling confection. Avoid the hard type, which is more common in bottom sweets. Fruit is more commonly found in candies that people hate, so stay away from it.

1. **K-Means Clustering:**

To find the best chocolate mix and the combinations using the K-Means Clustering method.

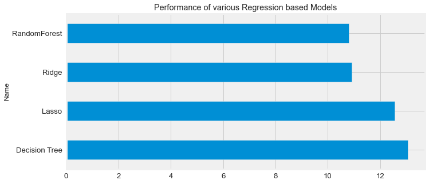
We have obtained a total of 9 clusters from the above graphs as the optimum number. According to this, the best candies are chocolate bar, caramel, and crispedricewafer.

1. **Decision Tree Regressor:**

****

We have obtained the features that are important using a Decision tree regressor model. Peanutyalmondy is one of the primary important features.

1. **Other Regression Models Comparison:**

****

Performed various regression models like Random Forest, Ridge, Lasso, and Decision Tree models

**Interpretations and Recommendations:**

So, Chocolate is the best and first choice to decide in a candy. As, the second choice we are choosing bar. We'll want chocolate in a bar shape because candies in this had a greater overall win rate than those in bits. We looked at a dataset to see what qualities define the most popular sweets. Chocolate, Bar, PeanutyAlmondy, are the most popular traits discovered. Yet, because we are left only with two items, drawing any firm conclusions is problematic, and it is likely more advisable to focus on the two features that contribute the most to effectiveness. There was no link discovered among winpercent, sugarpercent, and pricepercent.

**Conclusion:**

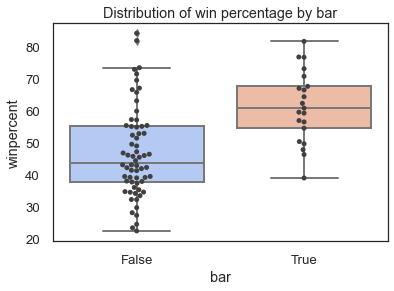
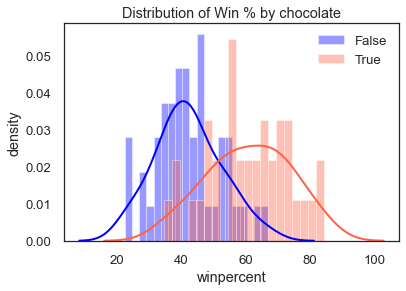
So, Chocolate is the best and first choice to decide in a candy. As, the second choice we are choosing bar. This dataset has provided us with a lot of critical thinking abilities and the ability to address problems promptly. We would like to do furthermore analysis on the customers' shopping habits. Are they concerned about their health? Do people buy what they want or settle for less expensive alternatives? Combine the results of this study with the results of the above steps to create candy that people would enjoy and purchase. Introduce new candies to the marketplace. Users may give greater attention to pricing and sugar content in a real-world situation. Different groups of people shop in different ways. Some health-conscious clients may pay extra for a sweet with less sugar, while others may choose the cheapest candy, regardless of sugar or composition.

**References:**

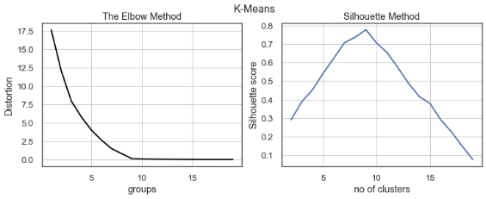
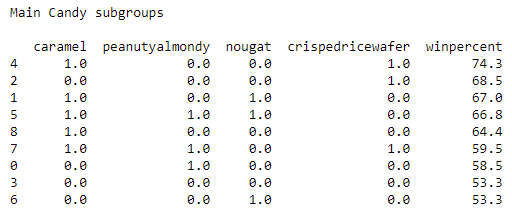
[1] Jason Brownlee. (October 5, 2020). Robust Regression for Machine Learning in Python. *Machine Learning Mastery*. Retrieved from https://machinelearningmastery.com/robust-regression-for-machine-learning-in-python/

[2] Valentina Alto. (Aug 17, 2019). Understanding the OLS method for Simple Linear Regression. *Towards Data Science Medium Article*. Retrieved from https://towardsdatascience.com/understanding-the-ols-method-for-simple-linear-regression-e0a4e8f692cc

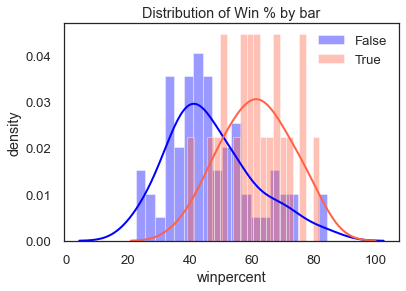
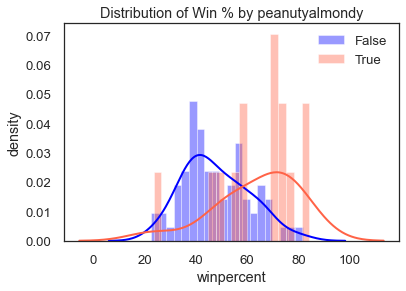
**Appendix:**

*Figure 1: Distribution of Win % By Bar Figure 2: Distribution of Win % By Bar*

*Figure 3: K-Means Elbow & Silhoutte Methods Figure 4: K-Means 9 Clusters*

*Figure 5: Distribution of Win % By Bar Figure 6: Distribution of Win % By peanutyalmondy*