**Will the Sweetest Candy Taste Victory or will the Costliest Confection Reign Supreme? Data Analysis of Halloween Candy using Linear Regression Methods**

Rachel Cutlan and Timothy Goodwin

We will be performing analysis on the data set “The Ultimate Halloween Candy Power Ranking” from the FiveThirtyEight data repository. For the final project, we will be doing the Data Analysis project, using regression methods used in class. We want to determine how price and sugar content affects the popularity of a candy, and whether certain characteristics of the candy, such as chocolate vs fruity, affects this popularity.

The data set involves 85 different candy types, from 100 Grand to Whoppers. These candies each have sugar percentage, price percentage, and win percentage. The sugar percentage ranges from 1.1% (One dime and One quarter) to 98.8% (Reese’s stuffed with pieces) and is the percentile of sugar it falls in within the data set. The price percentage ranges from 1.1% (Tootsie Roll Midgies) to 97.6% (Nik L Nip) and is the unit price percentile compared to the rest of the data set. The win percentage ranges from 22.4% (Nik L Nip) to 84.18% (Reese’s Peanut Butter Cup) and is the overall win percentage according to 269,000 matchups.

The attributes of each candy are also provided. The candies are given binary variables of 0 and 1 for nine different attributes, including containing chocolate, caramel, peanuts, fruit flavor, or if it comes as a single or multiple candies. Each candy can have multiple attributes. Thirty-seven of the candies contain chocolate and fourteen of the candies contain caramel. Thirty-eight of the candies are fruit flavored. Fifteen of the candies are considered hard candies and twenty-one of the candies are considered candy bars.

For this analysis, we will treat all variables as independent variables except for the win percentage, which will be the dependent, or outcome, variable. We will start by plotting the price percentage and sugar percentage variables to ensure that they meet the assumptions. This will involve using tools such as QQ plots. Any necessary transformations of the data will be completed. We will also run the necessary tests to check for outliers and influential points. This will involve computing values such as R-student residuals and Cook’s distance.

Once we have adequately shown that the final model has no violation of assumptions, we will use it to help us determine what factors influence the popularity of a candy. We will create models for the complete data set, and for the binary data. We will use confidence intervals to support our results, as p-values alone will not answer the question sufficiently. All results will be shown either using figures or tables to adequately explain the results. We will also explain any issues with our model and ways we might improve upon this analysis in the future.