Stat Graphics Final Project

Meher Kaky, Sreekar Kunaparaju, Alex Chen, Mike Zhou

2024-04-28

Introduction

Halloween is a beloved holiday that revolves around a simple yet effective social contract: homeowners provide candy to costumed trick-or-treaters, and in return, the trick-or-treaters refrain from engaging in mischievous late-night antics. To ensure a smooth and enjoyable Halloween experience for all parties involved, it is essential to determine the most preferable Halloween candy.

In an effort to answer this question, the popular website FiveThirtyEight conducted an online survey to gauge the internet's preference for various Halloween candies. The survey, which collected responses from 8,371 unique IP addresses, generated a total of 269,000 responses. Participants were presented with randomly generated matchups of different, popular Halloween candies and asked to choose their preferred option in each matchup.

By analyzing the data collected from this survey, we can gain valuable insights into the most favored Halloween candies among the online community. This information can help homeowners make informed decisions when purchasing candy for trick-or-treaters, ensuring that they offer treats that are likely to be well-received and appreciated by the costumed visitors.

Dataset

The candy dataset consists of attributes of 85 candies ranked on 13 different variable. Every column except for `winpercent`, `sugarpercent`, and `pricepercent` is binary and answers simple questions like: "Does the candy have chocolate?" or "Does the candy have caramel?" The `winpercent` column measures the percent of matchups won between a head-to-head matchup survey conducted between all of the candies. The `sugarpercent` column measures the percentile of sugar a given candy falls under relative to the rest of the dataset. Finally, the `pricepercent` column measures the percentile of price a given candy falls under relative to the rest of the dataset.

Research questions

The main focus of our research surrounded the question of what influences the win rate of a candy and thus makes it more appealing to consumers. In order to further explore this idea, we focused on the following questions.

Does price influence the preference of specific candies?

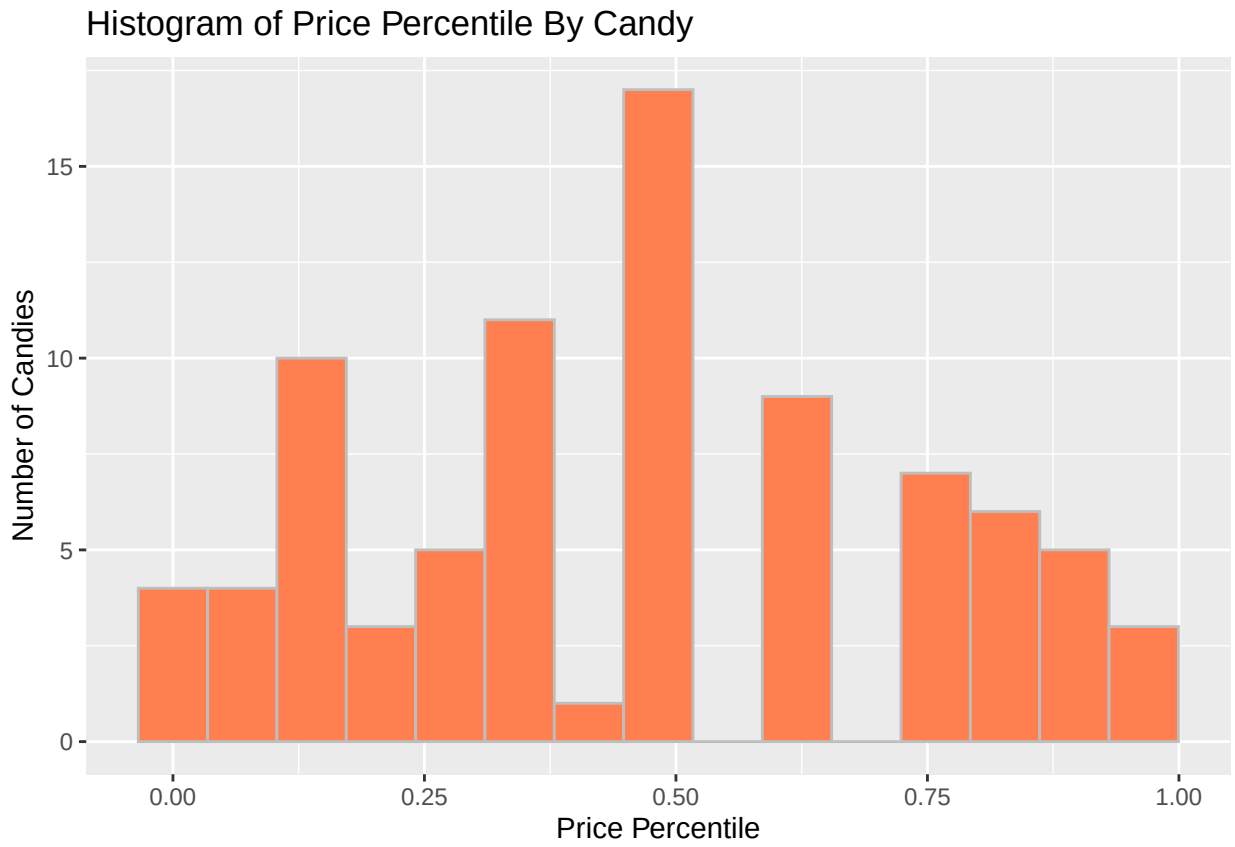
What effect does sugar percentage have on the win percentage of a candy?

Does the presence of chocolate in a candy's composition influence its chances of being chosen by consumers, thereby potentially boosting its win percentage?

Research Questions, Graphs, and Analysis

Does price influence the preference of specific candies?

People like getting expensive gifts – it makes them feel special and valued. Does the same dynamic work with trick-or-treaters? In order to answer this question, we started off by plotting the price percentile of the various candies in our dataset to determine the distribution of prices.



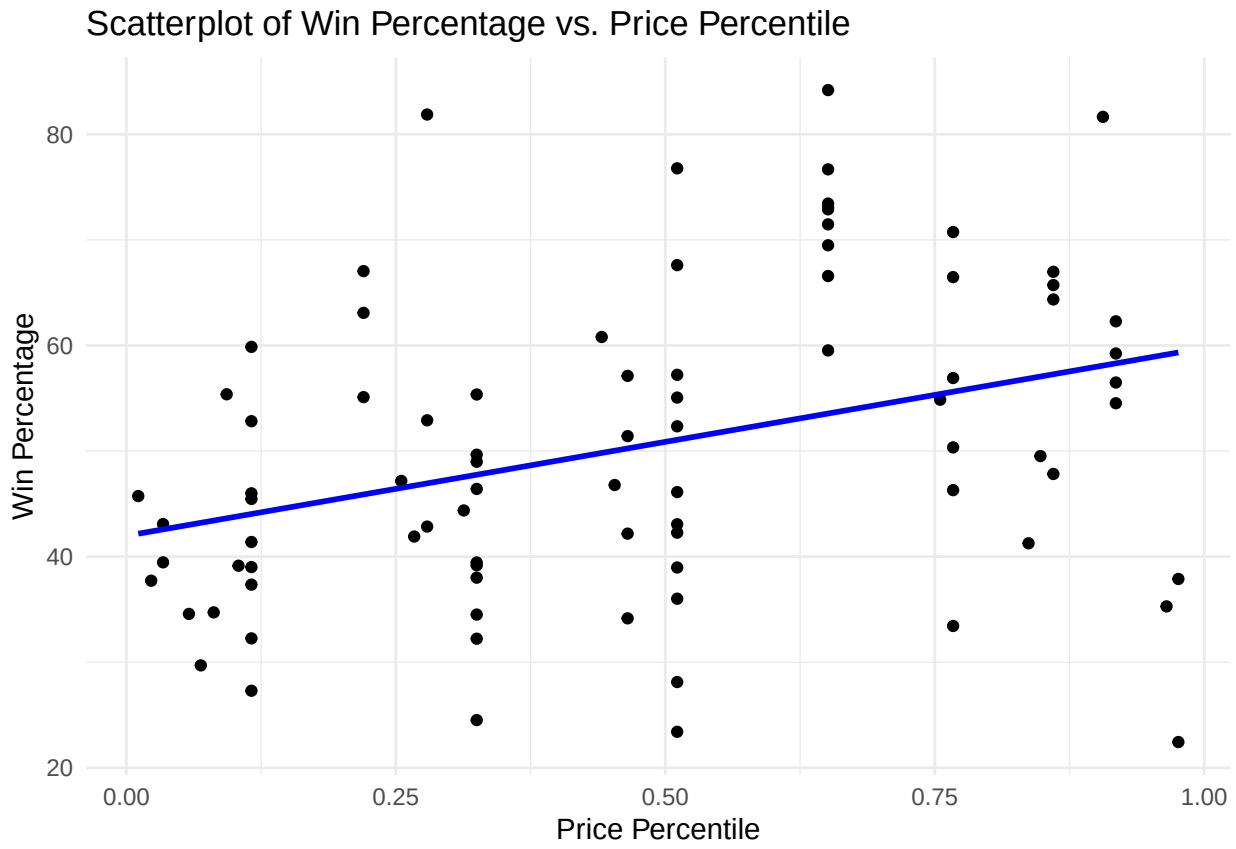
Looking at our histogram of prices, we see that our distribution roughly represents a normal distribution, peaking in the middle with roughly equally thick tails on either side. Most of the candies in our dataset fall near the median, with a significant amount of candies that are very cheap and very expensive. There aren't many in terms of informing our further analysis, however, this distribution will benefit us in that we will be able to observe the probability of winning a matchup across the entire price spectrum.

Given that we know the our distribution of candies follows a rough normal distribution, we can fit a regression to our data and determine the effect of price on how often a given candy won.

```
##
## Call:
## lm(formula = winpercent ~ pricepercent, data = candy)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -36.889  -8.573  -0.544   8.784  34.926
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)    41.979      2.908  14.435  < 2e-16 ***
```

```
## pricepercent   17.783      5.305   3.352  0.00121 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 13.89 on 83 degrees of freedom
## Multiple R-squared:  0.1192, Adjusted R-squared:  0.1086
## F-statistic: 11.24 on 1 and 83 DF,  p-value: 0.001209

## `geom_smooth()` using formula = 'y ~ x'
```

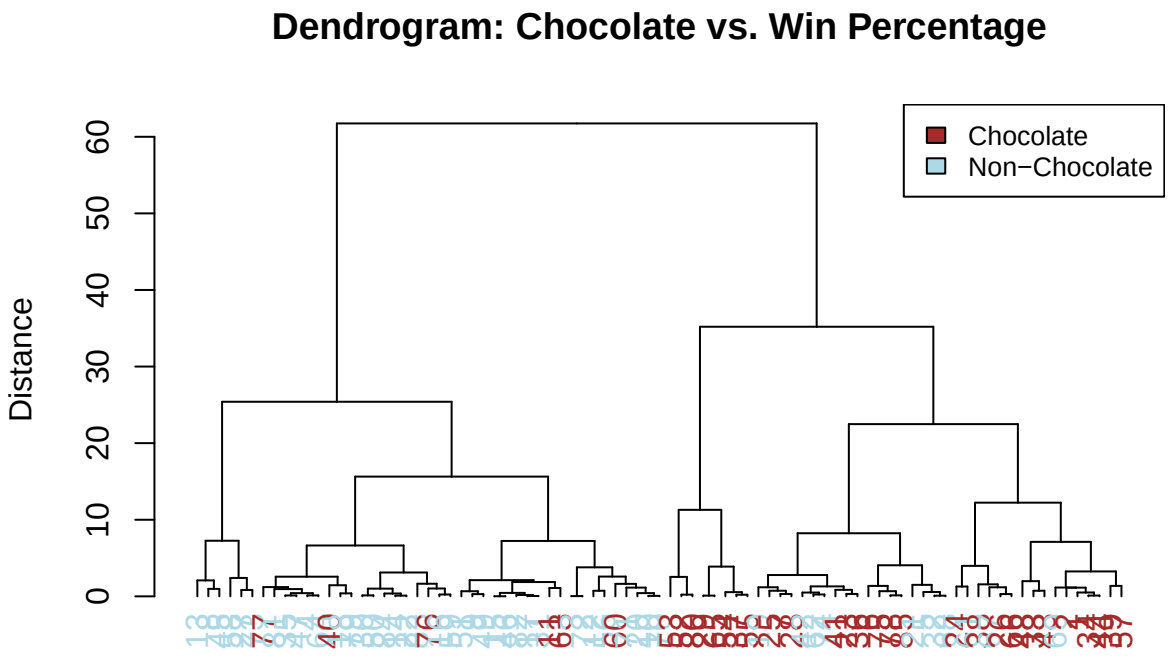


Looking at the graph and the trend of the plotted regression line, we see that the price of candies and their chance of winning a given matchup are positively correlated. In other words, as the price of a given candy goes up, it has a higher chance of winning any given matchup with another candy. This is corroborated by our regression data, which shows that for every 1 percentile increase in unit price relative to other candies, the percentage of matchups won by that specific candy is expected to increase by 17.78% on average. However, while this data is encouraging, it is important to note the high level of uncertainty concerning this estimate. Looking at R-squared, we see that the model explains only 12% of the variance in the data, and the standard errors of the coefficient is 5.305, which is almost 30% of the estimate itself, a high degree of uncertainty. Thus, while there is an observed increase in the percent of matchups won given an increase in price percentile, this increase is fairly volatile.

Does the presence of chocolate in a candy's composition influence its chances of being chosen by consumers, thereby potentially boosting its win percentage?

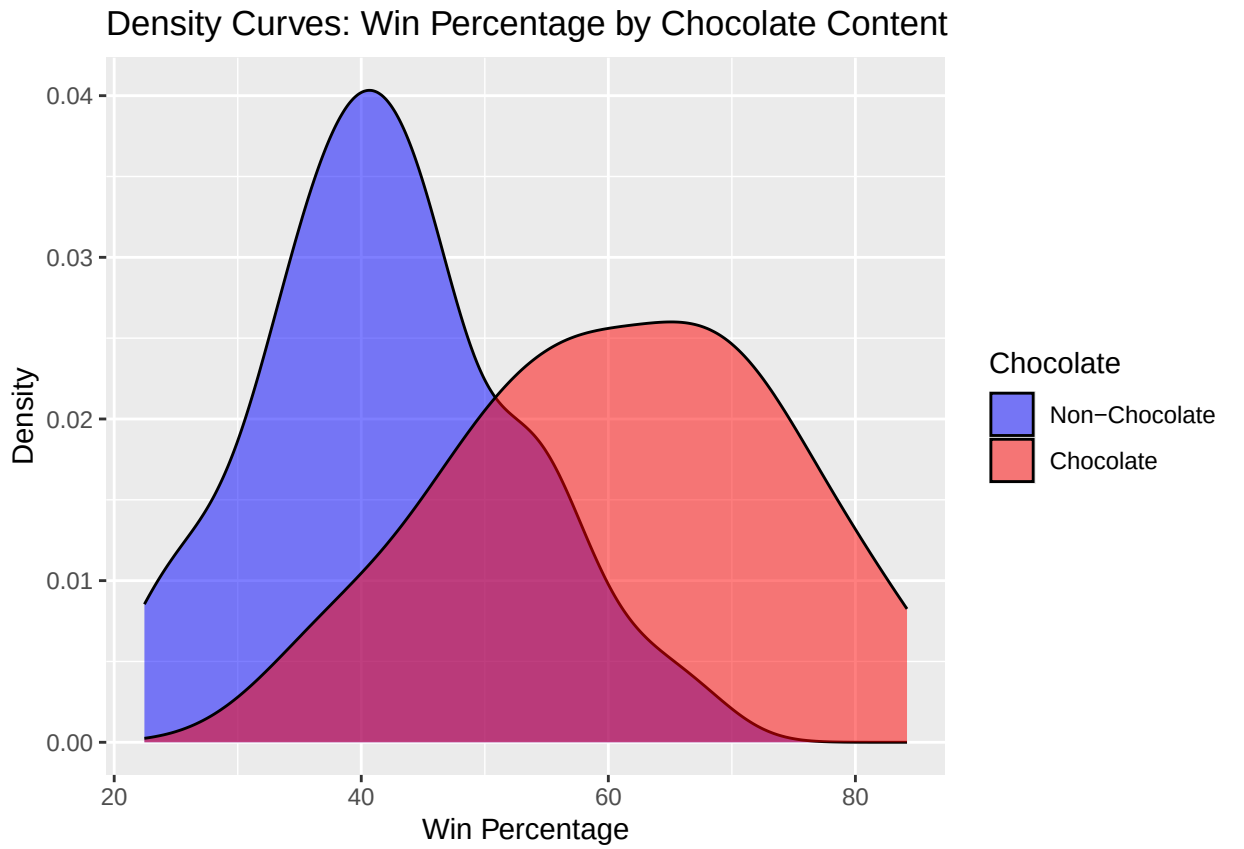
One of the most popular types of candy is undeniably chocolate. Considering its immense popularity, we were curious to investigate whether the presence of chocolate in a candy had any influence on its likelihood of

being chosen by consumers, thereby potentially boosting its win percentage and making it more preferrable during Halloween. To further explore this idea, we first explored the dendrogram between the win percentage and whether a candy is made of chocolate or not.



Right away in the dendrogram, we see an overall strong separation between the chocolate was non-chocolate win percentage. We notice that there are two separate clusters prevalent in the dendrogram where one cluster appears to be for more non-chocolate candies and the other cluster appears to be for chocolate candies. The length of the branches for the two main clusters also tells us that there is a good amount of distance, or dissimilarity, between the chocolate and non-chocolate branches. So, from here, we have evidence suggesting that whether or not a candy has chocolate has a pretty strong affect on the win percentage.

Let us further look into this idea by examining the density curves of the win percentage for both the chocolate and non-chocolate candies.



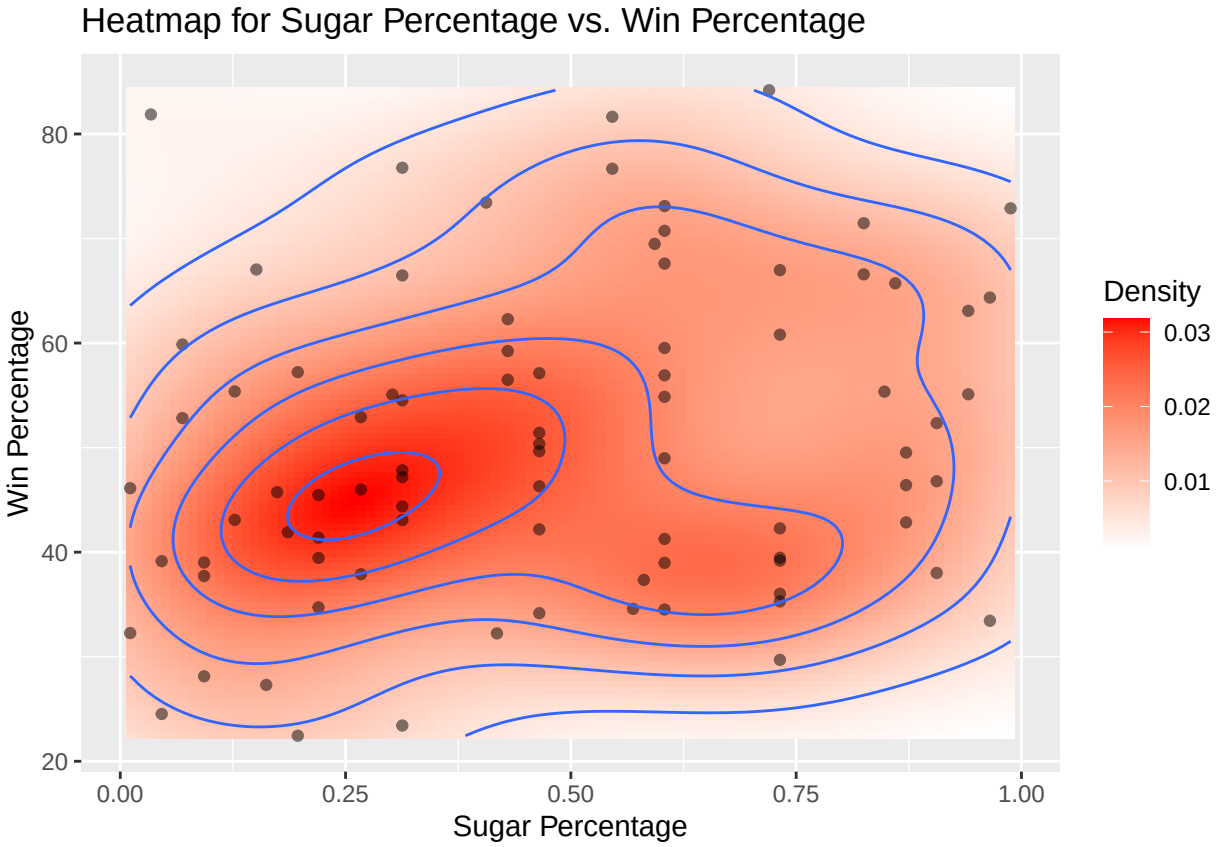
First, we can utilize density plots to observe the shape of the win percentage based on whether the candy was made of chocolate or not made of chocolate. Initially, we see that both curves seem to be approximately normally distributed. But, we see right away that the chocolate candies have a win percentage distribution shifted to the right of that of the non-chocolate candies. This indicates to us that the means of the distributions are different and mean win percentage for the chocolate candies is higher. Apart from the center, we notice a pretty evident difference in the shape of the density curves as well, with the non-chocolate density curve appearing to have more kurtosis than the chocolate density curve. From this plot, we seem to be able to see a trend where the chocolate candies have a higher win percentage overall than the non-chocolate ones, which tells us that chocolate as an impact of how likely a candy is to win. We can confirm this idea by formally testing it with a KS-test.

```
##
##  Exact two-sample Kolmogorov-Smirnov test
##
## data:  win_percent_choc and win_percent_non_choc
## D = 0.6357, p-value = 1.789e-08
## alternative hypothesis: two-sided
```

If we run a KS-test where the null hypothesis is that the win percentage distribution for the chocolate and non-chocolate candies is the same and the alternative is that they differ, we can reject the null hypothesis since the p-value of 1.79e-08 is less than 0.05. Thus, we have evidence that chocolate is a popular flavor of candy for Halloween and has a statistically significant effect on the win percentage.

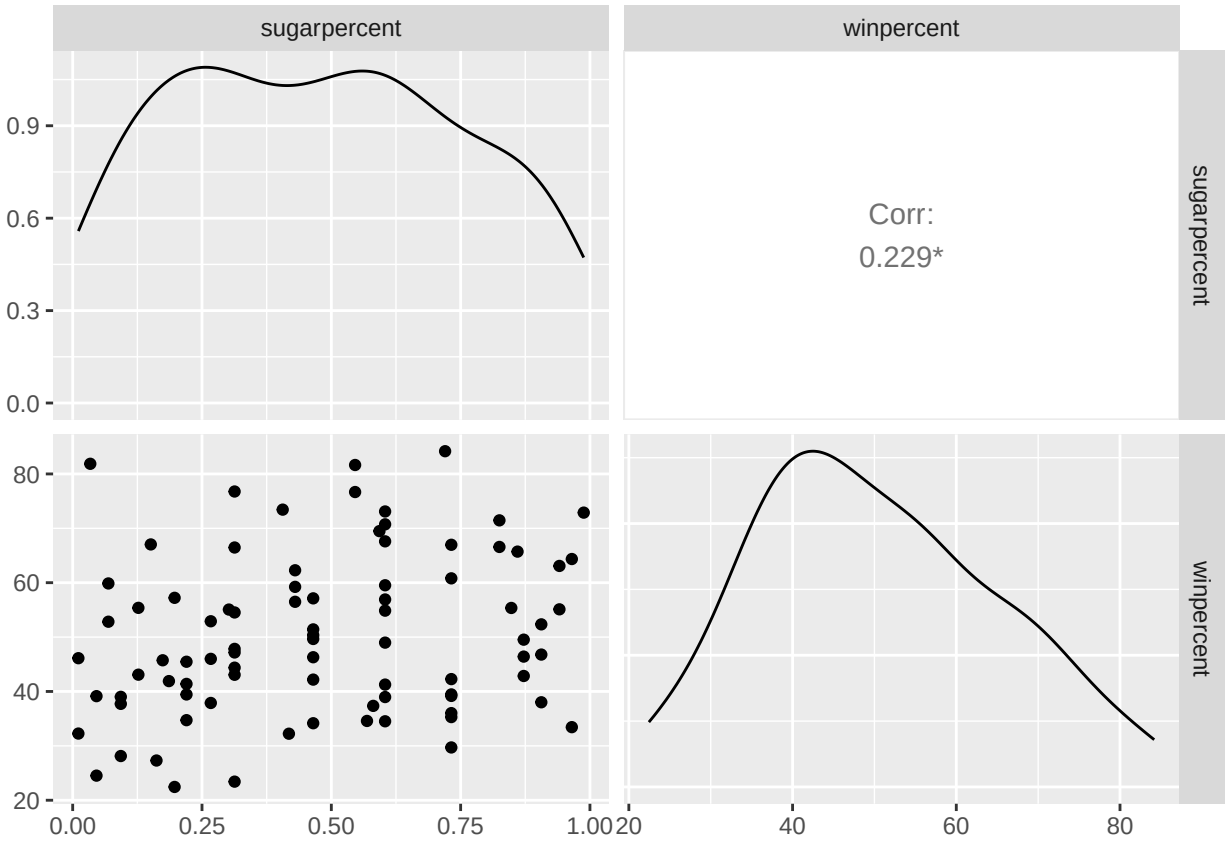
What effect does sugar percentage have on the win percentage of a candy?

Sugar is often an important factor when deciding whether a certain candy may taste good or not. A candy's sugar percentage might affect the consumers perception of the candy which in turn may affect its demand. Thus, we wanted to better understand how the sugar content in a candy may affect the win percentage overall of the candy. So, we proceeded to explore the relationship between the sugar percentage and the win rate.



To begin, we plot a heatmap of the sugar percentage vs the win percentage of the various candies in the dataset. Right away, we see a density of points around the sugar percentage of 0.25 and a win percentage of approxiatemly 0.45. This tells us that there is a cluster of points where the sugar percentage is lower and the win percentage is not too high either. This provides some evidence of the theory that a higher sugar percentage might influence the win rate of a candy. However, it is not definitively clear just based on the plot above whether a higher sugar percentage may lead to a better win percentage.

To better test whether there is some relationship between the two variables, let us take a look at the correlations between the variables.



Although initially we appear to see some trend where the sugar percentage and the win percentage rise together by looking at the bivariate distribution of the variables, the uni variate distributions do not provide the same strong evidence. Although both the win percentage and the sugar percentage seem to rise together, the winpercentage appears to drop much faster. We see further evidence for this weak relationship when we observe the correlation of 0.229, which is relatively very low. Thus, we do not have too much strong evidence that an increase in sugar percentage effects the win percentage of a candy.So, we can not say definitively that a sweeter candy will be preferred on Halloween.

Conclusion

From the above analyses, we were not able to definitively determine whether sugar percentage or price of a candy makes it a more appealing choice for Halloween. Although we observed some weak trends when doing this analysis, the correlation and R^2 values were too low to get any significant results. This does not mean that the price and sugar content do not influence the win percentage or popularity of a candy. Rather, it suggests there are other variables which may be playing an important role which are not obvious.

On the contrary, we observe that a candy which contains chocolate has a significant effect on the popularity of the candy overall. From the dendrogram's cluster, we saw a pretty strong separation for the win percentage between the two categories. Furthermore, when we tested to see if the distributions for the win percentages are different for the categories, we obtained significant results and saw in our density plots that the curve chocolate candies had a higher mean win percentage, indicating chocolate candies are a more popular choice.

In the future, more work will be done to determine whether other categorical variables such as whether the candy has caramel, is a bar, has peanuts etc.. have any affect on the win percentage and rather if any combinations of these variables may be significant. Additionally, we were not able to get very significant or strong results when we tried to find a relationship between price and sugar percentage on the win percentage. Since price and sugar seem to be important variables in determining popularity, more complex relationships

between the variables might be needed. Thus, more complex methods, such as possibly using the principle components of the price and sugar percentage to predict the win percentage might be required to reveal underlying relationships between those variables.