Candy Power Ranking

# Management Summary & Recommendations

## We need more data

FiveThirtyEight.com claims it aggregated the winpercent over 269.000 matchups of the 83 candy types in our data set. While this is a good number of matchups that provides us with trustworthy information on how well a certain candy scores in the market. However, when developing a new candy, we do not look to just copy the characteristics of the best scoring candy, but we look to understand the effect of the different features on the performance in the market so we can develop our candy with the best combination of features in mind. In this case, we only have 83 possible combinations of features (85 candies) and their win percentage. This means that we only have 83 datapoints to work with in this analysis. Some features are underrepresented like nougat and crisped rice/wafer/cookie and some features do not exist without another, like nearly all candybars are chocolate tasting. This makes it difficult to formulate conclusive advice on these feats.

## Chocolate candy with peanuts and/or almonds should be developed

From our multivariate analyses it seems that there is a limited number of features of which we can confidently say that they influence the winpercent significantly. For our 'perfect candy', we can only say that the candy must be chocolate tasting with some peanuts/almonds. From the corresponding branch in our decision tree we learn that there is some price sensitivity involved. Above the the 71st percentile, the winpercent significantly decreases and is 16 percent lower. It is therefore a good idea to make a chocolate candy, with peanuts/almonds in it and is priced underneath the 71st percentile.

For the values of other features, we limited the sample to the candy types that had chocolate and peanuts/almonds in them. Our ‘perfect’ candy will not be fruity nor hard, as there are also no candy types that have these characteristics together with chocolate and peanuts/almonds. Furthermore, we will also not add crisped rice/wafer/cookie as only 1 candy has this feature together with having chocolate and having peanuts/almonds, so it’s safer to go with the majority.

It seems that for chocolate and peanut/almond candy, bars are significantly scoring less then, so we will let our candy not be in the form of a bar. For caramel, nougat and pluribus, there is no significant difference, so we will leave the choice for these to the product developers. Our ‘perfect’ candy looks like this:

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Chocolate | Fruity | Caramel | PeanutyAlmondy | Nougat | CrispedRicewafer | Hard | Bar | Pluribus | Sugar | Price |
| 1 | 0 | 0/1 | 1 | 0/1 | 0 | 0 | 0 | 0/1 | 0%-100% | <71% |

Based on our data available and the above configuration, we should be expecting a winpercent of 75.6%.

## Next steps: collecting more data

Since our dataset now only contained 85 datapoints there are still some questions that are unanswered. We can only be completely sure about the overall effect of chocolate and peanuts and almonds. After fixing these feature levels we saw some new strong correlations rise, but the truth is that we are still left with some questions about some effects, especially those that are underrepresented or those that only occur together; what if they occur separately form one another.

We suggest further analyses using the conjoint analysis method. The most common methodology here is the choice-based conjoint analysis (CBC), where the concept is surprisingly similar to the analysis carried out by FiveThirtyEight.com. Three to five candy profiles with different factor levels are given as a choice to a test panel. The candy profile that is chosen gets a win (1) and the other ones get a loss (0). After enough matchups with different feature levels for the different candy types (hopefully more then 83), we can do a logistic regression and analyse the factor coefficients. This allows us to collect more data and test the conclusions made in this summary. The bigger amount of data also allows us to do a train-test-split so we can cross-validate the data. As we had limited data, we chose to use all the data in the model on purpose to not lose information, which might cause overfitting.

# Procedure

## Description data & preliminary assumptions

FiveThirtyEight.com collected the data by randomly pitting 2 candy types from a sample of 85 candy types against each other. 269.000 matches were made and the winning percentage of the matchups for each specific candy type were recorded.

While 269.000 matchups have been registered to define the winpercent, only 85 possible combinations of features have been researched, which makes it quite hard to make conclusive statements on the effects of certain characteristics.

Our first assumption is therefore that further research is needed. As we are dealing with limited datapoints (85 candytypes and their winpercent) we will focus on the descriptive part of our analysis and not on the prescriptive. Our hypotheses can be tested in further analyses down the line. We will therefore not make any train-test-split of the data in our model and not do cross-validation. This way, we maintain our 85 datapoints and make use of the full data, which is preferred as we don’t have too much datapoints to base our models on in the first place.

## Data import & cleaning

The data was first imported, and the different features were checked. 11 characteristics were defined of which 8 were related to flavour (chocolate, fruity, peanut/almondy, nougat, crisped rice/wafer/cookie, hard, sugarpercent) and 3 that were related to form or presentation (bar, pluribus, pricepercent). We did not miss any values and but did have to delete 2 datapoints that were added as a joke and do not carry any informational value about candy (‘One dime’, ‘One quarter’). This left us with 83 datapoints

The numerical features were plotted (sugarpercent and pricepercent) and seemed quite normally distributed and not very skewed (mean and median are very close). Winpercent (dependent variable, numerical) was also plotted and is a bit skewed but still ok for our regression purpose (skew = 0.30).

The categorical variables were also plotted as a barplot and we saw immediately that some of the features were underrepresented in our sample (nougat, crispedricewafer).

## Correlation

After this step, we constructed the correlation matrix. We saw that fruity and chocolate are heavily correlated (-0.78), bar and chocolate (0.59) and bar and pluribus (-0.62). It already gave us an insight that chocolate and fruity flavours might not combine, bars come a lot in the form of chocolate and candy bars do not come in plural.

## Univariate analysis

Next, we checked the direct influence of these features on the dependent variable (winpercent). For the categorical features, we used boxplots and the student t-test to see what the direct effect was. For the numerical variables, we used scatter plots and the correlation to assess the relationship.

## Multivariate analysis

### 4.1 Regression Analysis

The first step in the multivariate analysis was a regression analysis. We tested all the different features and some higher order effects of the numerical features (,). We eliminated the features without significant contributions (including the higher order ones since they did not contribute) and tested again with a feature that summed the amount of ‘flavour’ features next to the ones that were already in the model (maybe the number of different flavours in a candy was important). This effect also proved insignificant, and we were left with chocolate, fruity and peanutyalmondy. It was confirmed to us that chocolate is the way to go, and best combined with peanutyalmondy. If the candy did not contain chocolate, it would still be a good idea to make it fruity, which was already suspected because of the univariate analysis (fruity and chocolate do not coexist in our sample).

### 4.3 Decision Tree

To test for some nonlinear, hierarchic effects we also developed a decision tree. Chocolate was once again confirmed to be the most important effect together with peanutyalmondy. As expected, fruity did also play a role (as the opposite of chocolate). Furthermore, we observed some effects in pricepercent, sugarpercent and caramel that we did not really assess before.

## Univariate analysis after fixing chocolate and peanutyalmondy

After the multivariate analyses we saw that both chocolate and peanutyalmondy were the only significant ‘drivers’ for our winning candy. Therefore, we fixed the factor levels so that both chocolat and peanuts/almonds were present to see what the values were for the other features to maximize our winpercentage and therefore the consumer sentiment of our ‘perfect candy’.

# Detailed results from the analyses

## Univariate

* **Chocolate:** Significant difference between candies that are tasting like chocolate and candies that are not (p = 0.000000000143). Chocolate candy tastes on average 19% better.
* **Fruity:** Initially, there is a significant difference for fruity (p = 0.0144362732) and candy that is fruity scores 12 percent worse than candy that is not. However, once we look at the cross table with chocolate, we see that chocolate and fruity are mutually exclusive (only 1 candy type that is both chocolate and fruity). We therefore fixed the effect of chocolate on 0, and saw another significant effect arising: this time fruity candy was on average 8 percent higher (p = 0.020356076647). This means that people do prefer chocolate candy, but when it’s not chocolate candy, they still prefer fruity candy over non-fruity candy.
* **Peanutyalmondy:** There is a significant difference for peanutyalmondy (p = 0.000159221981). Candy with peanuts/almonds scores on average 16 percent better in winpercent.
* **Crispedricewafer:** Initially there seems to be a significant difference between candy that has crisped rice/wafer/cookie and candy that is not. However, after making the crosstable with chocolate, we see that all of the candies that contain crispedricewafer are also chocolate. When we fix the chocolate feature level on 1, the significance disappears.
* **Hard:** Initially there seems to be a significant difference between candy that is hard. However, after making the crosstable with fruity, we see that the vast majority of the hard candies that are hard are also fruity. When we fix the fruity feature level on 1, the significance disappears.
* **Bar:** Initially there seems to be a significant difference between candy that comes in a bar and candy that does not. However, after making the crosstable with chocolate, we see that the vast majority of candies that come in bar are also chocolate. When we fix the chocolate feature level on 1, the significance disappears.
* **Pluribus:** Candies that come in pluribus seem to score on average 8 percent lower than their non-pluribus counterpart (p = 0.012663239057).

Caramel and nougat did not produce significant differences (p > 0.05) and for our numerical variables, no strong correlation/linear trend was observed. We can therefore conclude that in this first analysis, the main influencing factors are chocolate, fruity, peanutyalmondy and pluribus.

## Multivariate

### Regression Analysis

As discussed in the methodology, we made a regression (ordinary least-squares) model. Next to the given 11 features, we also tried to include higher order effects of the numerical variables (, ) to check whether there might be some nonlinear effects. After initial elimination of the nonsignificant variables, we also added a ‘sum’ variable of the eliminated ones to check whether the number of features would have an effect.

These were all not significant and we ended up with the model as described in [Appendix 1](#_Regression_result_table). Chocolate has a coefficient of 0.226 (p=0.000), peanutyalmondy has a coefficient of 0.0923 (p=0.011) and fruity has a coefficient of 0.0846 (p=0.034). This means that chocolate has the biggest positive influence, peanutyalmondy the second biggest and fruity comes in on the third place. We have to be aware that there is a large negative correlation between chocolate and fruity, and the data shows that they are mostly mutually exclusive. In a real-world testing scenario, we will probably choose one of the 2, and since chocolate has the biggest positive effect, it seems logic to make our candy chocolaty and leave fruity behind. We can then combine chocolate with peanuts/almonds.

If we decide to collect more data after this experiment, it would be interesting to get more datapoints that are both fruity and chocolaty and to see what the winpercent for those candytypes is.

### Decision tree

We finally included a decision tree to assess some of the nonlinear, hierarchic effects for winpercent. The decision tree and graph with the importance of the different features can be consulted in [Appendix 2](#_Decision_tree). We chose a max depth of 3 to really see the most significant splits. We see that chocolate is the feature with the most importance (0.65). The difference between the samples for this split is 19%.

If we look at the most successful group (chocolate = 1) we see that the next split is made with peanutyalmondy (importance = 0.09374217). The difference between the samples is now 11%. When we go further down the winning path, we see some price sensitivity occurring. It seems that chocolate candy with peanut/almond taste that score lower than 71% in the price percent have a win percent that is on average 16 percent higher. This is our winning candy.

Looking at our decision tree, when a candy does not have chocolate, it should be fruity. The difference between candies that do not have chocolate but are fruity and those that do not contain chocolate but are not fruity is on average 9 percent. After that, the sugar percent becomes the split criteria for fruity candy and the amount of nougat for non-fruity candy.

We can conclude that some of our earlier ideas from the univariate analysis and the regression analysis are confirmed here. Chocolate is the most important feature with chocolates being more popular than non-chocolates (60.9% on average vs 42.3%). A chocolate-tasting candy should also contain some peanuts/almonds (68.5% vs 57.3%). If it is not a chocolate candy, it can still score quite high if it is fruity (44% vs 35.2%). An interesting new insight is that in the winning path, there is also some price sensitivity and that if a candy type is both chocolaty and contains peanuts/almonds, it will score significantly higher if it is under the 71st percentile in price. In the case it’s higher than that, people seem to turn to candies that don’t contain peanuts/almonds but contain caramel.

## Univariate after fixing chocolate and peanutyalmondy

The main (positive) effects that we saw coming back in both our initial univariate analysis and our multivariate analysis was chocolate and peanutyalmondy. We therefore looked at the candy types that had both these features present and did some more univariate analyses to know which other features our new ‘perfect’ ‘winning’ candy should contain. This analysis is therefore based on a small subsample.

* **Hard & crispedricewafer**: both these variables will be fixed on 0 for the simple reason that all or nearly all of our candies in our subsample do not have these features. We could try to make a unique candy here by including them but for the mere purpose of defining a ‘winning’ candy we will go with the majority here.
* **Caramel, Nougat, pluribus and sugarpercent**: For these features there is no significant difference between factor levels or significant correlation when it comes to winpercent in our subsample. We therefore leave the choice to the product developer.
* **Bar & pricepercent**: For these 2 features we do see some significant influence on winpercent. Bar candy within our subsample of chocolaty candy with almonds/peanuts seem to score much lower in winning percentage then non-bar candy. We will therefore fix the factor level on 0. We do see a negative price correlation of -0.66 with winpercent in our subsample. We already noticed this price sensitivity in our decision tree and it seems like the cutoff value is 71%. We will therefore make sure that the pricepercent is lower then the 71st percentile. This leads to the following ‘winning’ candy:

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Chocolate | Fruity | Caramel | PeanutyAlmondy | Nougat | CrispedRicewafer | Hard | Bar | Pluribus | Sugar | Price |
| 1 | 0 | 0/1 | 1 | 0/1 | 0 | 0 | 0 | 0/1 | 0%-100% | <71% |

We can expect an average winning percentage of 75.6%.

# Appendix

## Regression result table

Table

Description automatically generated

## Decision tree

Chart

Description automatically generated