Candy Power Ranking

Lidl Analytics – Data Science Case Study

Sagarnil Dasgupta

dasgupta.sagarnil@gmail.com

# Brief summary of the candy case study

* Candy dataset is created based on 269,000 matchups, when 85 candies go head to head. Not only the ingredient but the packaging and the colour along with smell of the candy has a part to play in deciding the win percentage from a user perspective.
* Boolean values like 0 and 1 for chocolate and caramel says that if that ingredient is present or not but having a numeric value of number of grams of chocolate present in 100 grams of the candy could help to better analyse.

There are few candies which varies with win percentage in greater margin but has similar combination of ingredients(Below is an example):

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | competitorname | chocolate | fruity | caramel | peanutyalmondy | nougat | crispedricewafer | hard | bar | pluribus | sugarpercent | pricepercent | winpercent |
| 46 | Nik L Nip | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0.20 | 0.98 | 22.45 |
| 70 | Starburst | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0.15 | 0.22 | 67.04 |

Table 1

Nik L Nip and Starburst have all the ingredient present with almost same sugar percentage with a difference in price. However, there is a significant difference in win percentage which proves that there are some other attributes influencing the win percentage of a candy.

Few examples with similar features but significant difference in win percentage.

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | competitorname | chocolate | fruity | caramel | peanutyalmondy | nougat | crispedricewafer | hard | bar | pluribus | sugarpercent | pricepercent | winpercent |
| 33 | Lifesavers big ring gummies | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0.27 | 0.28 | 52.91 |
| 74 | Super Bubble | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0.16 | 0.12 | 27.30 |
| 35 | M&M’s | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0.82 | 0.65 | 66.57 |
| 61 | Sixlets | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0.22 | 0.08 | 34.72 |

Table 2

On further analysis, it is found that chocolate and fruity are highly co-related to each other as they are mutually exclusive except one- Hershey’s Special Dark which has both chocolate and fruity both as an ingredient. It seems that any candy manufacture finds it tough to add both chocolate and fruity flavour in a single candy.

When each ingredient is taken into account following trends are observed :

|  |  |
| --- | --- |
| CANDY TYPE | AVG. WIN SHARE |
| Nougat | 66% |
| Crispy | 64% |
| Chocolate | 61% |
| Candy bar | 61% |
| Caramel | 60% |
| Peanuts & nuts | 57% |
| Hard candy | 47% |
| Fruit | 44% |
| Multi-piece | 41% |

Fig 1 Table 3

The above chart (Fig 1) shows the average win percentage of each ingredient in candy. From this chart it seems that any candy having a nougat seems to be an important factor but when this is analysed using a regression model, Nougat seems to have not much significance in the win percentage.

When predicting the win percentage using a Regression model based on the 85 candies, the observation is that Nougat, Bar, Caramel and pluribus or multi-piece have no such significant influence in determining the win percentage where Chocolate, Fruit, Peanut & Nuts, Crispy and a high sugar percentage is liked by most of the user. Ideal candy should be soft not a jaw breaker and price should be at reasonable range.

|  |  |
| --- | --- |
| **CANDY TYPE** | **Value added to win percentage** |
| Chocolate | 19.9873 |
| Fruit | 8.6228 |
| Peanuts & nuts | 10.0435 |
| Crispy | 9.4243 |
| Hard Candy | -6.0456 |
| sugar percentage | 9.5396 |
| price percentage | -5.4628 |

Table 4

Fig 2

From the above table we can say that Chocolate and Peanut are a must in the ideal candy as its adds a value of (+20) and (+10) to the overall win percentage. Adding fruits will increase the win percentage but in the sample chocolate and Fruit combo are very rare and one taste could be supressing the other. From the regression model crispy normally add on the win percentage but chocolate + Peanuts + Crispy like (Snickers Crisper) doesn’t do as well as Reeves’s Peanut Butter cup and its spinoffs.

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | competitorname | chocolate | fruity | caramel | peanutyalmondy | nougat | crispedricewafer | hard | bar | pluribus | sugarpercent | pricepercent | winpercent |
| 53 | Reeves’s Miniatures | 1 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0.03 | 0.28 | 81.87 |
| 54 | Reeves’s Peanut Butter cup | 1 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0.72 | 0.65 | 84.18 |
| 56 | Reeves’s stuffed with pieces | 1 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0.99 | 0.65 | 72.89 |

Table 5

There are only 3 candies that are from same brand Reese’s which has same composition of only chocolate and Peanuts and have an average win percentage of 80% which says either this combination works out for most of the customers or the buyers are more connected with the Reese’s brand. As we also see 4 out of 10 most win percentage is also from Reese’s brand.

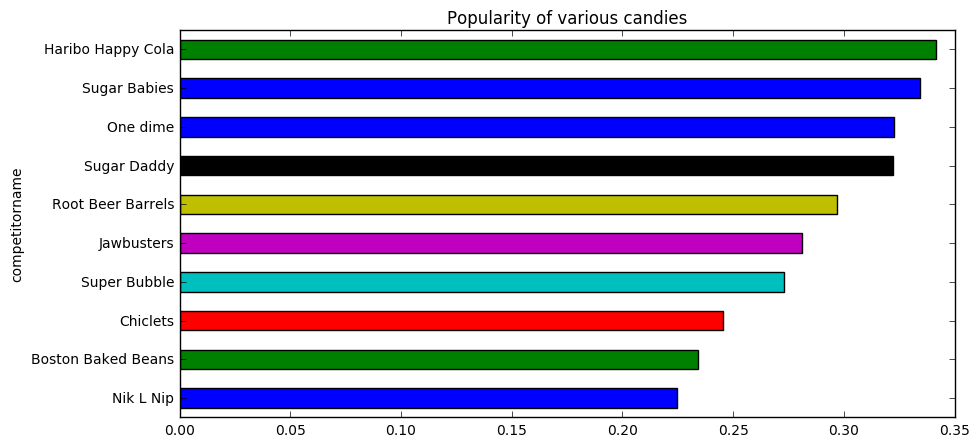


Fig 3

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | competitorname | chocolate | fruity | caramel | peanutyalmondy | nougat | crispedricewafer | hard | bar | pluribus | sugarpercent | pricepercent | winpercent |
| 52 | Reese's Peanut Butter cup | 1 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0.84 | 0.84 | 0.84 |
| 79 | Twix | 1 | 0 | 1 | 0 | 0 | 1 | 0 | 1 | 0 | 0.82 | 0.82 | 0.82 |
| 51 | Reese's Miniatures | 1 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0.82 | 0.82 | 0.82 |
| 28 | Kit Kat | 1 | 0 | 0 | 0 | 0 | 1 | 0 | 1 | 0 | 0.77 | 0.77 | 0.77 |
| 64 | Snickers | 1 | 0 | 1 | 1 | 1 | 0 | 0 | 1 | 0 | 0.77 | 0.77 | 0.77 |
| 54 | Reese's stuffed with pieces | 1 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0.73 | 0.73 | 0.73 |
| 53 | Reese's pieces | 1 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 1 | 0.73 | 0.73 | 0.73 |
| 36 | Milky Way | 1 | 0 | 1 | 0 | 1 | 0 | 0 | 1 | 0 | 0.73 | 0.73 | 0.73 |
| 42 | Nestle Butterfinger | 1 | 0 | 0 | 1 | 0 | 0 | 0 | 1 | 0 | 0.71 | 0.71 | 0.71 |
| 32 | Peanut butter M&M's | 1 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 1 | 0.71 | 0.71 | 0.71 |

Table 6

We have also seen that all the top 10 candies have chocolate as the common ingredient. In fact the worst 10 candies in term of win percentage do not have chocolate in it.

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | competitorname | chocolate | fruity | caramel | peanutyalmondy | nougat | crispedricewafer | hard | bar | pluribus | sugarpercent | pricepercent | winpercent |
| 44 | Nik L Nip | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0.22 | 0.22 | 0.22 |
| 7 | Boston Baked Beans | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 1 | 0.23 | 0.23 | 0.23 |
| 12 | Chiclets | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0.25 | 0.25 | 0.25 |
| 72 | Super Bubble | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0.27 | 0.27 | 0.27 |
| 26 | Jawbusters | 0 | 1 | 0 | 0 | 0 | 0 | 1 | 0 | 1 | 0.28 | 0.28 | 0.28 |
| 57 | Root Beer Barrels | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 1 | 0.30 | 0.30 | 0.30 |
| 2 | One dime | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0.32 | 0.32 | 0.32 |
| 71 | Sugar Daddy | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0.32 | 0.32 | 0.32 |
| 70 | Sugar Babies | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 1 | 0.33 | 0.33 | 0.33 |
| 19 | Haribo Happy Cola | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0.34 | 0.34 | 0.34 |

Table 7

## Perfect Candy

A perfect candy should have chocolate ,peanuts and non-jaw breaker like Reese’s Peanut Butter cup, fruity like starburst, Crispy like Twix as sugary as Reese’s stuffed with pieces and match the price with Tootsie Roll Midgies. Caramel, Nougat and bar are not mandatory ingredient but nothing harm in having any of these 3.

The above conclusion is based on multiple observations collected after doing multiple analytical solutions done on the data that are detailed below.

# List of Observations

[Observation 1: Reese’s Peanut Butter Cups and their spinoffs come out huge here, taking four of the top 10 spots and appearing pretty synonymous with the platonic ideal of Halloween candy. 11](#_Toc21297644)

[Observation 2: All the top 10 candies has chocolate attribute present. 11](#_Toc21297645)

[Observation 3: Reese's Miniatures is the cheapest when compared to top competitors and overall as well. 11](#_Toc21297646)

[Observation 4: The bottom 10 candies has chocolate attribute missing. 11](#_Toc21297647)

[Observation 5: Bottom 5 and top 5 candies have a big difference in win percentage thus giving confidence on the range of data. 12](#_Toc21297648)

[Observation 6: When non-chocolate candies are taken into consideration, Starburst top’s the chart. 13](#_Toc21297649)

[Observation 7: ReeseÕs stuffed with pieces is the top sugary candy. 14](#_Toc21297650)

[Observation 8: Bottom 10 candies in terms of Sugar lacks chocolate as well. 14](#_Toc21297651)

[Observation 9: Tootsie Roll Midgies seems to perform better when price and win percentages are considered. 16](#_Toc21297652)

[Observation 10: Dum Dums seems to perform better when price and sugar percentages are considered. 16](#_Toc21297653)

[Observation 11: A detail analysis is made regarding the chocolate and fruity attribute it has been observed that except 1(Tootsie Pop) which has chocolate as well as fruity. Other than this candy there is no other candy which has a chocolatey and fruity combo. Either it is chocolatey or fruity or none of these attributes. 17](#_Toc21297654)

[Observation 12: Chocolate is the more important feature in determining the winpercentage as if more of the candy eater think of chocolate when they think of candies. 18](#_Toc21297655)

[Observation 13: Chocolate, fruity, peanut-almond, crispy and sugar percentage has positive impact on win percentage but hardness should be avoided and as expected with any retail product less price is more liked by the customer. 21](#_Toc21297656)

[Observation 14: These 6 candies (Dum Dums, Fruit Chews, Pixie Sticks, Root Beer Barrels, Strawberry bon bons and Tootsie Roll Midgies) are different in some way from the rest. All these candies’ win percentages are on the bottom and may have a separate fan base for these candies. 22](#_Toc21297657)

[Observation 15. 'Dum Dums' and 'Tootsie Roll Midgies' are sort of opposite of each other. The first one is fruity and the second one chocolaty. 23](#_Toc21297658)

[Observation 16: Cluster ID 0 contains competitors which are mostly chocolaty, sugary and more favourable. Cluster ID 2, although being chocolaty has a low sugar percentile. 23](#_Toc21297659)

[Observation 17: All the chocolates which don't belong to Cluster ID 0 have made to the top 10 list of `winbyprice`. They are all cheap. 23](#_Toc21297660)

# Table of contents

[1. Brief Conclusion of the candy case study 1](#_Toc21290813)

[2. Perfect Candy 5](#_Toc21290814)

[3. List of Observation 6](#_Toc21290815)

[4. Table of contents 7](#_Toc21290816)

[5. Introduction 9](#_Toc21290817)

[5.1. Scenario 9](#_Toc21290818)

[5.2. Data 9](#_Toc21290819)

[5.3. Data Description 9](#_Toc21290820)

[6. Data Manipulation 10](#_Toc21290821)

[6.1. Missing value check 10](#_Toc21290822)

[6.2. Rectifying scale 10](#_Toc21290823)

[6.3. Rounding Numeric variable 11](#_Toc21290824)

[6.4. Removing Special Character 11](#_Toc21290825)

[7. Data Analysis 11](#_Toc21290826)

[7.1. Top 10 candy by win percentage 11](#_Toc21290827)

[7.2. Bottom 10 top candy by Win percentage 12](#_Toc21290828)

[7.3. Top 5 and bottom 5 candies by win percentage 13](#_Toc21290829)

[7.4. Top candy without chocolate 14](#_Toc21290830)

[7.5. Top 10 sugary candy 14](#_Toc21290831)

[7.6. Bottom 10 sugary candy 15](#_Toc21290832)

[7.7. Top 5 and bottom 5 candies by sugar percentage 16](#_Toc21290833)

[7.8. Impact of price on win and sugar 16](#_Toc21290834)

[7.9. Correlation 17](#_Toc21290835)

[7.10. Feature Importance 19](#_Toc21290836)

[7.11. Best Candy with ingredient 19](#_Toc21290837)

[8. Linear Regression 20](#_Toc21290838)

[8.1. Predicting Win percentage 20](#_Toc21290839)

[9. Clustering 22](#_Toc21290840)

[9.1. Average Attribute description of each cluster: 23](#_Toc21290841)

# Introduction

This document details about the approach taken to analyse the candy database and

to find out which product characteristics drive customer sentiment and subsequently make a recommendation on a new product.

## Scenario

The Lidl purchasing group wants to expand their candy offering. These are store brand candies that are sold along the brand offerings. The idea is to create a brand-new product. The team is discussing various options at the moment.

Some prefer cookie-based sweets while others think that it should be gummies. The Divisional Director responsible for purchasing has decided to use a more data-driven approach. He contracted with a market research group to collect data on products in the market and their characteristics and customer sentiment.

The market research data is now available and it is being used to find out which product characteristics drive customer sentiment and subsequently make a recommendation on a new product.

## Data

The data set is located (incl. a short description) here:

<https://github.com/fivethirtyeight/data/tree/master/candy-power-ranking>

The data set is provided by FiveThirtyEight under the Creative Commons Attribution 4.0 International license (<https://creativecommons.org/licenses/by/4.0/> )

## Data Description

| **Header** | **Description** |
| --- | --- |
| chocolate | Does it contain chocolate? |
| fruity | Is it fruit flavoured? |
| caramel | Is there caramel in the candy? |
| peanutalmondy | Does it contain peanuts, peanut butter or almonds? |
| nougat | Does it contain nougat? |
| crispedricewafer | Does it contain crisped rice, wafers, or a cookie component? |
| hard | Is it a hard candy? |
| bar | Is it a candy bar? |
| pluribus | Is it one of many candies in a bag or box? |
| sugar percent | The percentile of sugar it falls under within the data set. |
| pricepercent | The unit price percentile compared to the rest of the set. |
| winpercent | The overall win percentage according to 269,000 matchups. |

# Data Manipulation

## Missing value check

There are no missing values in the dataset.

competitorname 0

chocolate 0

fruity 0

caramel 0

peanutyalmondy 0

nougat 0

crispedricewafer 0

hard 0

bar 0

pluribus 0

sugarpercent 0

pricepercent 0

winpercent 0

## Rectifying scale

Sugar percentage and price percentage are in the scale from 0 to 1, but win percentage are in the scale of 0 to 100. When we are predicting the win percentage in a regression model its ok to keep win percentage in different scale. But we require win percentage to be used in creating derived variable and using the data in clustering algorithm where the data need to be in same scale.

## Rounding Numeric variable

Numeric Values are rounded off to 2 decimal place for the 3 numeric variables as the data will look clean.

## Removing Special Character

Competitor name as Õ character which is replaced by ‘.

# Data Analysis

Candy-data.csv file consist of 85 rows which are 85 candy types and each one of them is defined based on attributes like chocolate, fruity etc… and the target variable- win percentage determines the likeliness of the user.

## Top 10 candy by win percentage

As win percentage is the attribute which defines whether user likes the particular candy compared to other candy brand. It is assumed a higher win percentage means that candy is liked better than other candies.

Top 10 candies when terms of win percentage are as follows:

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | competitorname | chocolate | fruity | caramel | peanutyalmondy | nougat | crispedricewafer | hard | bar | pluribus | sugarpercent | pricepercent | winpercent |
| 52 | Reese's Peanut Butter cup | 1 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0.72 | 0.65 | 0.84 |
| 79 | Twix | 1 | 0 | 1 | 0 | 0 | 1 | 0 | 1 | 0 | 0.55 | 0.91 | 0.82 |
| 51 | Reese's Miniatures | 1 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0.03 | 0.28 | 0.82 |
| 28 | Kit Kat | 1 | 0 | 0 | 0 | 0 | 1 | 0 | 1 | 0 | 0.31 | 0.51 | 0.77 |
| 64 | Snickers | 1 | 0 | 1 | 1 | 1 | 0 | 0 | 1 | 0 | 0.55 | 0.65 | 0.77 |
| 54 | Reese's stuffed with pieces | 1 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0.99 | 0.65 | 0.73 |
| 53 | Reese's pieces | 1 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 1 | 0.41 | 0.65 | 0.73 |
| 36 | Milky Way | 1 | 0 | 1 | 0 | 1 | 0 | 0 | 1 | 0 | 0.60 | 0.65 | 0.73 |
| 42 | Nestle Butterfinger | 1 | 0 | 0 | 1 | 0 | 0 | 0 | 1 | 0 | 0.60 | 0.77 | 0.71 |
| 32 | Peanut butter M&M's | 1 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 1 | 0.82 | 0.65 | 0.71 |

Table 8

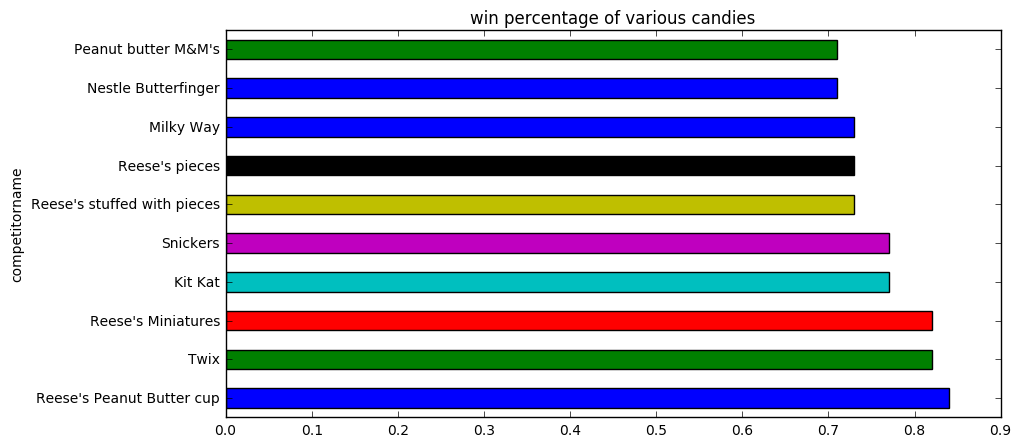


Fig 4

Observation 1: Reese’s Peanut Butter Cups and their spinoffs come out huge here, taking four of the top 10 spots and appearing pretty synonymous with the platonic ideal of Halloween candy.

Observation 2: All the top 10 candies has chocolate attribute present.

Observation 3: Reese's Miniatures is the cheapest when compared to top competitors and overall as well.

## Bottom 10 top candy by Win percentage

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | competitor name | chocolate | fruity | caramel | peanutyalmondy | nougat | crispedricewafer | hard | bar | pluribus | sugarpercent | pricepercent | winpercent |
| 44 | Nik L Nip | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0.20 | 0.98 | 0.22 |
| 7 | Boston Baked Beans | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 1 | 0.31 | 0.51 | 0.23 |
| 12 | Chiclets | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0.05 | 0.32 | 0.25 |
| 72 | Super Bubble | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0.16 | 0.12 | 0.27 |
| 26 | Jawbusters | 0 | 1 | 0 | 0 | 0 | 0 | 1 | 0 | 1 | 0.09 | 0.51 | 0.28 |
| 57 | Root Beer Barrels | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 1 | 0.73 | 0.07 | 0.30 |
| 2 | One dime | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0.01 | 0.12 | 0.32 |
| 71 | Sugar Daddy | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0.42 | 0.32 | 0.32 |
| 70 | Sugar Babies | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 1 | 0.96 | 0.77 | 0.33 |
| 19 | Haribo Happy Cola | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0.46 | 0.46 | 0.34 |

Table 9

Observation 4: The bottom 10 candies has chocolate attribute missing.

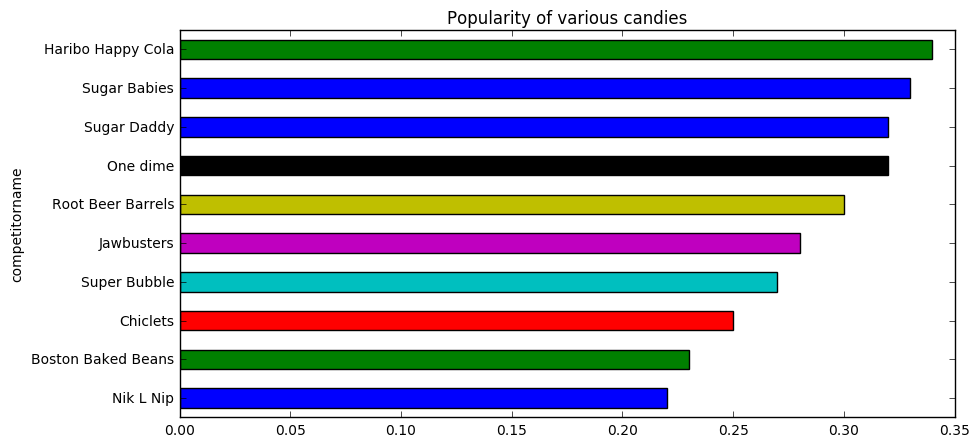


Fig 5

## Top 5 and bottom 5 candies by win percentage

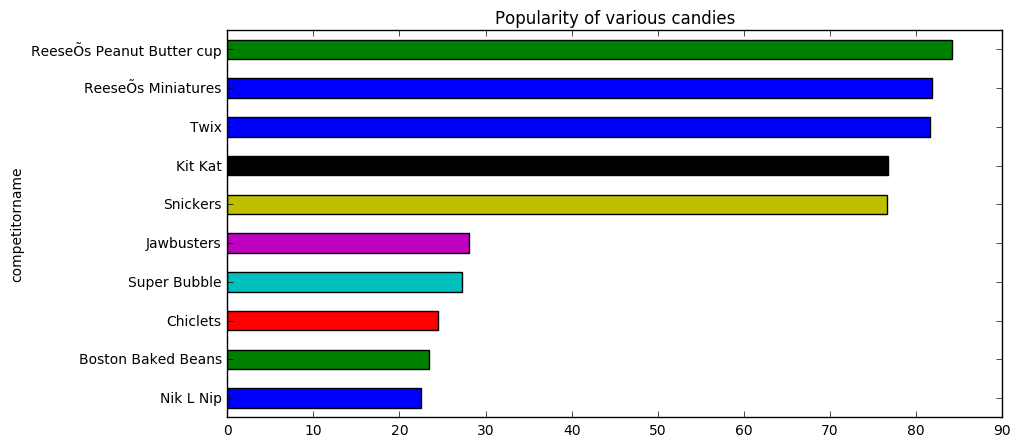


Fig 6

Observation 5: Bottom 5 and top 5 candies have a big difference in win percentage thus giving confidence on the range of data.

## Top candy without chocolate

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | competitorname | chocolate | fruity | caramel | peanutyalmondy | nougat | crispedricewafer | hard | bar | pluribus | sugarpercent | pricepercent | winpercent |
| 68 | Starburst | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0.15 | 0.22 | 67.04 |
| 60 | Skittles original | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0.94 | 0.22 | 63.09 |
| 66 | Sour Patch Kids | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0.07 | 0.12 | 59.86 |

Table 10

Observation 6: When non-chocolate candies are taken into consideration, Starburst top’s the chart.

## Top 10 sugary candy

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | competitorname | chocolate | fruity | caramel | peanutyalmondy | nougat | crispedricewafer | hard | bar | pluribus | sugarpercent | pricepercent | winpercent |
| 54 | Reese's stuffed with pieces | 1 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0.99 | 0.65 | 0.73 |
| 70 | Sugar Babies | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 1 | 0.96 | 0.77 | 0.33 |
| 38 | Milky Way Simply Caramel | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 1 | 0 | 0.96 | 0.86 | 0.64 |
| 61 | Skittles wildberry | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0.94 | 0.22 | 0.55 |
| 60 | Skittles original | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0.94 | 0.22 | 0.63 |
| 17 | Gobstopper | 0 | 1 | 0 | 0 | 0 | 0 | 1 | 0 | 1 | 0.91 | 0.45 | 0.47 |
| 4 | Air Heads | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0.91 | 0.51 | 0.52 |
| 8 | Candy Corn | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0.91 | 0.32 | 0.38 |
| 34 | Mike & Ike | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0.87 | 0.32 | 0.46 |
| 84 | Whoppers | 1 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 1 | 0.87 | 0.85 | 0.50 |

Table 11

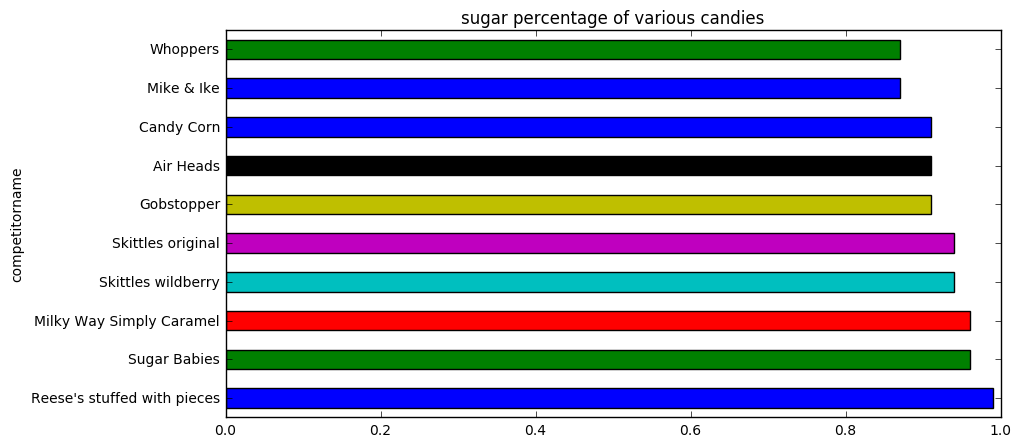


Fig 7

Observation 7: ReeseÕs stuffed with pieces is the top sugary candy.

## Bottom 10 sugary candy

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | competitorname | chocolate | fruity | caramel | peanutyalmondy | nougat | crispedricewafer | hard | bar | pluribus | sugarpercent | pricepercent | winpercent |
| 2 | One dime | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0.01 | 0.12 | 0.32 |
| 3 | One quarter | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0.01 | 0.51 | 0.46 |
| 51 | Reese's Miniatures | 1 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0.03 | 0.28 | 0.82 |
| 30 | Lemonhead | 0 | 1 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0.05 | 0.10 | 0.39 |
| 12 | Chiclets | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0.05 | 0.32 | 0.25 |
| 66 | Sour Patch Kids | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0.07 | 0.12 | 0.60 |
| 67 | Sour Patch Tricksters | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0.07 | 0.12 | 0.53 |
| 48 | Pixie Sticks | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0.09 | 0.02 | 0.38 |
| 26 | Jawbusters | 0 | 1 | 0 | 0 | 0 | 0 | 1 | 0 | 1 | 0.09 | 0.51 | 0.28 |
| 81 | Warheads | 0 | 1 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0.09 | 0.12 | 0.39 |

Table 12

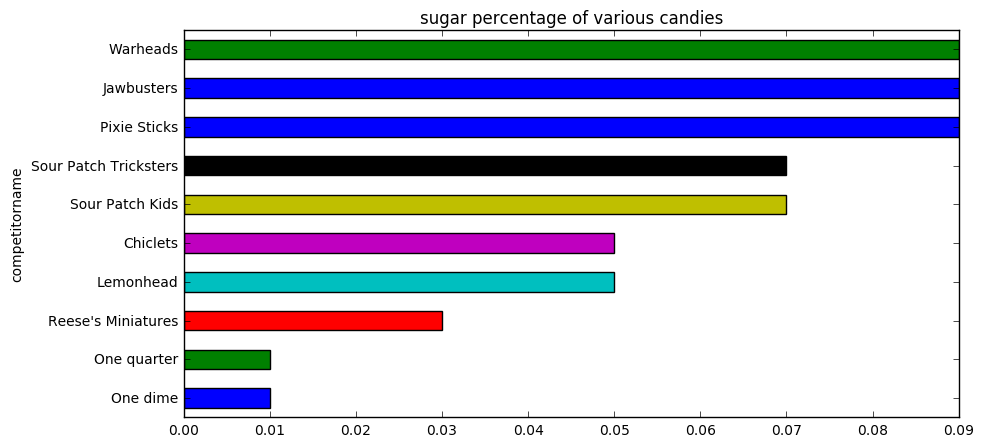


Fig 8

Observation 8: Bottom 10 candies in terms of Sugar lacks chocolate as well.

## Top 5 and bottom 5 candies by sugar percentage

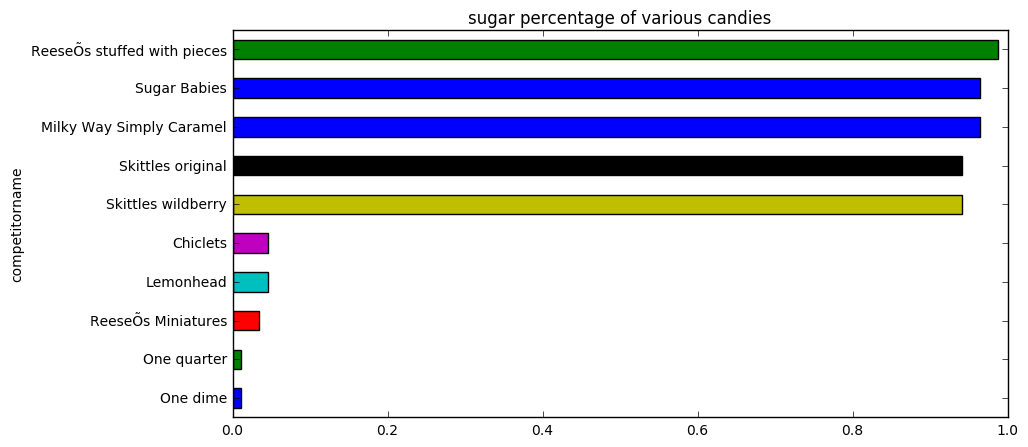


Fig 9

## Impact of price on win and sugar

Two new derived attributes are used

sugarbyprice = sugar percentage / price

winbyprice = win percentage / price

Higher sugarbyprice value means the candy is sweet as well as cheap.

Higher winbyprice value means the candy is more liked as well as cheap.

Top 10 candies are identified by sorting on the basis of winbyprice.

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | competitorname | chocolate | fruity | caramel | peanutyalmondy | nougat | crispedricewafer | hard | bar | pluribus | sugarpercent | pricepercent | winpercent | sugarbyprice | winbyprice |
| 76 | Tootsie Roll Midgies | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0.17 | 0.01 | 0.46 | 17.00 | 46.00 |
| 48 | Pixie Sticks | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0.09 | 0.02 | 0.38 | 4.50 | 19.00 |
| 15 | Fruit Chews | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0.13 | 0.03 | 0.43 | 4.33 | 14.33 |
| 14 | Dum Dums | 0 | 1 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0.73 | 0.03 | 0.39 | 24.33 | 13.00 |
| 22 | Hershey's Kisses | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0.13 | 0.09 | 0.55 | 1.44 | 6.11 |
| 69 | Strawberry bon bons | 0 | 1 | 0 | 0 | 0 | 0 | 1 | 0 | 1 | 0.57 | 0.06 | 0.35 | 9.50 | 5.83 |
| 66 | Sour Patch Kids | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0.07 | 0.12 | 0.60 | 0.58 | 5.00 |
| 67 | Sour Patch Tricksters | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0.07 | 0.12 | 0.53 | 0.58 | 4.42 |
| 59 | Sixlets | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0.22 | 0.08 | 0.35 | 2.75 | 4.38 |
| 57 | Root Beer Barrels | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 1 | 0.73 | 0.07 | 0.30 | 10.43 | 4.29 |

Table 13

Observation 9: Tootsie Roll Midgies seems to perform better when price and win percentages are considered.

Top 10 candies are identified by sorting on the basis of sugarbyprice.

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | competitorname | chocolate | fruity | caramel | peanutyalmondy | nougat | crispedricewafer | hard | bar | pluribus | sugarpercent | pricepercent | winpercent | sugarbyprice | winbyprice |
| 14 | Dum Dums | 0 | 1 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0.73 | 0.03 | 0.39 | 24.33 | 13.00 |
| 76 | Tootsie Roll Midgies | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0.17 | 0.01 | 0.46 | 17.00 | 46.00 |
| 57 | Root Beer Barrels | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 1 | 0.73 | 0.07 | 0.30 | 10.43 | 4.29 |
| 69 | Strawberry bon bons | 0 | 1 | 0 | 0 | 0 | 0 | 1 | 0 | 1 | 0.57 | 0.06 | 0.35 | 9.50 | 5.83 |
| 50 | Red vines | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0.58 | 0.12 | 0.37 | 4.83 | 3.08 |
| 48 | Pixie Sticks | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0.09 | 0.02 | 0.38 | 4.50 | 19.00 |
| 15 | Fruit Chews | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0.13 | 0.03 | 0.43 | 4.33 | 14.33 |
| 60 | Skittles original | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0.94 | 0.22 | 0.63 | 4.27 | 2.86 |
| 61 | Skittles wildberry | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0.94 | 0.22 | 0.55 | 4.27 | 2.50 |
| 58 | Runts | 0 | 1 | 0 | 0 | 0 | 0 | 1 | 0 | 1 | 0.87 | 0.28 | 0.43 | 3.11 | 1.54 |

Table 14

Observation 10: Dum Dums seems to perform better when price and sugar percentages are considered.

## Correlation



Fig 10

Top 10 correlation relationship:

|  |  |  |  |
| --- | --- | --- | --- |
|  | Feature1 | Feature2 | corr |
| 0 | chocolate | fruity | 0.741721 |
| 1 | sugarbyprice | winbyprice | 0.675094 |
| 2 | chocolate | win percent | 0.636517 |
| 3 | chocolate | bar | 0.597421 |
| 4 | bar | pluribus | 0.593409 |
| 5 | nougat | bar | 0.522976 |
| 6 | bar | pricepercent | 0.518407 |
| 7 | fruity | bar | 0.515066 |
| 8 | chocolate | pricepercent | 0.504675 |
| 9 | pricepercent | winbyprice | 0.471809 |

Table 15

Observation 11: A detail analysis is made regarding the chocolate and fruity attribute it has been observed that except 1(Tootsie Pop) which has chocolate as well as fruity. Other than this candy there is no other candy which has a chocolatey and fruity combo. Either it is chocolatey or fruity or none of these attributes.

## Feature Importance

Decision Tree Regressor is used to find the more important feature while predicting win percentage.

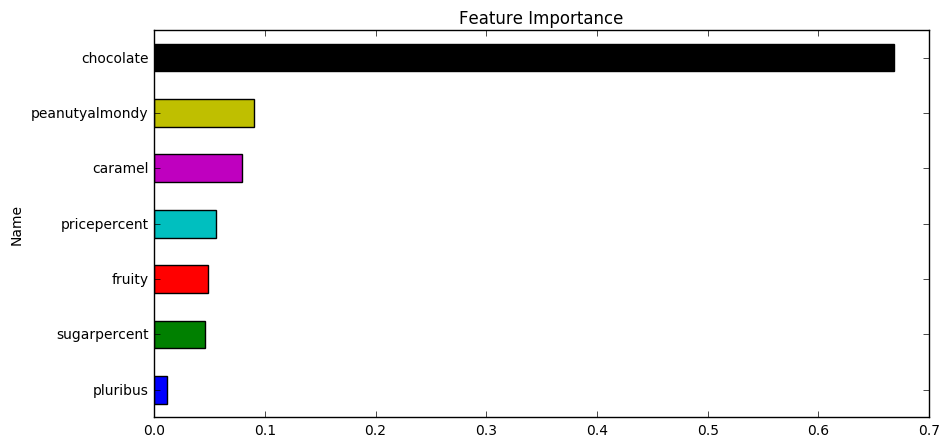


Fig 11

Observation 12: Chocolate is the more important feature in determining the winpercentage as if more of the candy eater think of chocolate when they think of candies.

## Best Candy with ingredient

* Best candy with chocolate- Reese’s Peanut Butter cup
* Best candy with fruit- Starburst
* Best candy with nuts- Reese’s Peanut Butter cup
* Best Candy with crispedricewafer- Twix
* Best Candy which is not hard - Reese’s Peanut Butter cup
* Candy with more sugar- Reese’s stuffed with pieces
* Cheapest Candy- Tootsie Roll Midgies

# Linear Regression

## Predicting Win percentage

Win percentage can be dependent on 11 IDV variable:

| **Header** | **Description** |
| --- | --- |
| chocolate | Does it contain chocolate? |
| fruity | Is it fruit flavoured? |
| caramel | Is there caramel in the candy? |
| peanutyalmondy | Does it contain peanuts, peanut butter or almonds? |
| nougat | Does it contain nougat? |
| crispedricewafer | Does it contain crisped rice, wafers, or a cookie component? |
| hard | Is it a hard candy? |
| bar | Is it a candy bar? |
| pluribus | Is it one of many candies in a bag or box? |
| sugarpercent | The percentile of sugar it falls under within the data set. |
| pricepercent | The unit price percentile compared to the rest of the set. |

Linear Regression can used to predict the win percentage and the influence of each variable

After Splitting dataset into 80% train and 20% test below table shows the implementation of OLS Regression when 11 IDV are used.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | **attribute** | **coef** | **std err** | **t** | **P>|t|** | **[95.0% Conf. Int.]** |
| **const** |  | 34.534 | 4.32 | 7.994 | 0 | 25.924 43.144 |
| x1 | chocolate | 19.7481 | 3.899 | 5.065 | 0 | 11.978 27.518 |
| x2 | fruity | 9.4223 | 3.763 | 2.504 | 0.015 | 1.923 16.922 |
| x3 | caramel | 2.2245 | 3.657 | 0.608 | 0.545 | -5.065 9.514 |
| x4 | peanutyalmondy | 10.0707 | 3.616 | 2.785 | 0.007 | 2.864 17.277 |
| x5 | nougat | 0.8043 | 5.716 | 0.141 | 0.888 | -10.588 12.197 |
| x6 | crispedricewafer | 8.919 | 5.268 | 1.693 | 0.095 | -1.580 19.418 |
| x7 | hard | -6.1653 | 3.455 | -1.784 | 0.079 | -13.051 0.721 |
| x8 | bar | 0.4415 | 5.061 | 0.087 | 0.931 | -9.645 10.528 |
| x9 | pluribus | -0.8545 | 3.04 | -0.281 | 0.779 | -6.913 5.204 |
| x10 | sugarpercent | 9.0868 | 4.659 | 1.95 | 0.055 | -0.200 18.373 |
| x11 | pricepercent | -5.9284 | 5.513 | -1.075 | 0.286 | -16.916 5.060 |

Table 16

Clearly Carmel, nougat, bar and pluribus seems to have very little impact on win percentage as there p-value is more than 0.05 when industry standard of 5% significant level is considered.

Back propagation is used to get rid of non-significant variable one by one after considering the change in R-square and adjusted R-square the below model best suits the purpose.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | **attribute** | **coef** | **std err** | **t** | **P>|t|** | **[95.0% Conf. Int.]** |
| **const** |  | 34.3934 | 3.769 | 9.125 | 0 | 26.888 41.898 |
| x1 | chocolate | 19.9873 | 3.669 | 5.447 | 0 | 12.681 27.294 |
| x2 | fruity | 8.6228 | 3.566 | 2.418 | 0.018 | 1.523 15.723 |
| x3 | peanutyalmondy | 10.0435 | 3.483 | 2.883 | 0.005 | 3.108 16.979 |
| x4 | crispedricewafer | 9.4243 | 4.585 | 2.055 | 0.043 | 0.294 18.554 |
| x5 | hard | -6.0456 | 3.305 | -1.829 | 0.071 | -12.626 0.535 |
| x6 | sugar percentage | 9.5396 | 4.377 | 2.18 | 0.032 | 0.824 18.255 |
| x7 | price percentage | -5.4628 | 5.12 | -1.067 | 0.289 | -15.658 4.733 |

Table 17

Better display of win percentage:

|  |  |
| --- | --- |
| **CANDY TYPE** | **Value added to win percentage** |
| chocolate | 19.9873 |
| fruity | 8.6228 |
| peanutyalmondy | 10.0435 |
| crispedricewafer | 9.4243 |
| hard | -6.0456 |
| sugar percentage | 9.5396 |
| price percentage | -5.4628 |



Table 18 Table 19

Observation 13: Chocolate, fruity, peanut-almond, crispy and sugar percentage has positive impact on win percentage but hardness should be avoided and as expected with any retail product less price is more liked by the customer.

# Clustering

Plotting elbow curve using K-Means algorithm the data seems to form 5 major clustering:

|  |  |
| --- | --- |
| Cluster No | No of candies |
| 0 | 79 |
| 1 | 2 |
| 2 | 1 |
| 3 | 1 |
| 4 | 2 |

Table 20

## Average Attribute description of each cluster:

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Cluster ID | chocolate | fruity | caramel | peanutyalmondy | nougat | crispedricewafer | hard | bar | pluribus | sugarpercent | pricepercent | winpercent | sugarbyprice | winbyprice |
| 0 | 0.46 | 0.44 | 0.18 | 0.18 | 0.09 | 0.09 | 0.15 | 0.27 | 0.49 | 0.48 | 0.50 | 0.51 | 1.19 | 1.51 |
| 1 | 0.00 | 0.50 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 1.00 | 0.11 | 0.02 | 0.40 | 4.42 | 16.66 |
| 2 | 1.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 1.00 | 0.17 | 0.01 | 0.46 | 17.00 | 46.00 |
| 3 | 0.00 | 1.00 | 0.00 | 0.00 | 0.00 | 0.00 | 1.00 | 0.00 | 0.00 | 0.73 | 0.03 | 0.39 | 24.33 | 13.00 |
| 4 | 0.00 | 0.50 | 0.00 | 0.00 | 0.00 | 0.00 | 1.00 | 0.00 | 1.00 | 0.65 | 0.06 | 0.32 | 9.96 | 5.06 |

Table 21

This table says that these below 6 candies are very different from rest of the bunch:

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | competitorname | chocolate | fruity | caramel | peanutyalmondy | nougat | crispedricewafer | hard | bar | pluribus | sugarpercent | pricepercent | winpercent | sugarbyprice | winbyprice | Cluster ID |
| 14 | Dum Dums | 0 | 1 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0.73 | 0.03 | 0.39 | 24.33 | 13.00 | 3 |
| 15 | Fruit Chews | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0.13 | 0.03 | 0.43 | 4.33 | 14.33 | 1 |
| 48 | Pixie Sticks | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0.09 | 0.02 | 0.38 | 4.50 | 19.00 | 1 |
| 57 | Root Beer Barrels | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 1 | 0.73 | 0.07 | 0.30 | 10.43 | 4.29 | 4 |
| 69 | Strawberry bon bons | 0 | 1 | 0 | 0 | 0 | 0 | 1 | 0 | 1 | 0.57 | 0.06 | 0.35 | 9.50 | 5.83 | 4 |
| 76 | Tootsie Roll Midgies | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0.17 | 0.01 | 0.46 | 17.00 | 46.00 | 2 |

Table 22

Observation 14: These 6 candies (Dum Dums, Fruit Chews, Pixie Sticks, Root Beer Barrels, Strawberry bon bons and Tootsie Roll Midgies) are different in some way from the rest. All these candies’ win percentages are on the bottom and may have a separate fan base for these candies.

Observation 15. 'Dum Dums' and 'Tootsie Roll Midgies' are sort of opposite of each other. The first one is fruity and the second one chocolaty.

Observation 16: Cluster ID 0 contains competitors which are mostly chocolaty, sugary and more favourable. Cluster ID 2, although being chocolaty has a low sugar percentile.

Observation 17: All the chocolates which don't belong to Cluster ID 0 have made to the top 10 list of `winbyprice`. They are all cheap.