Case Study: Expanding our candy brand

Ermioni Kritsinioti

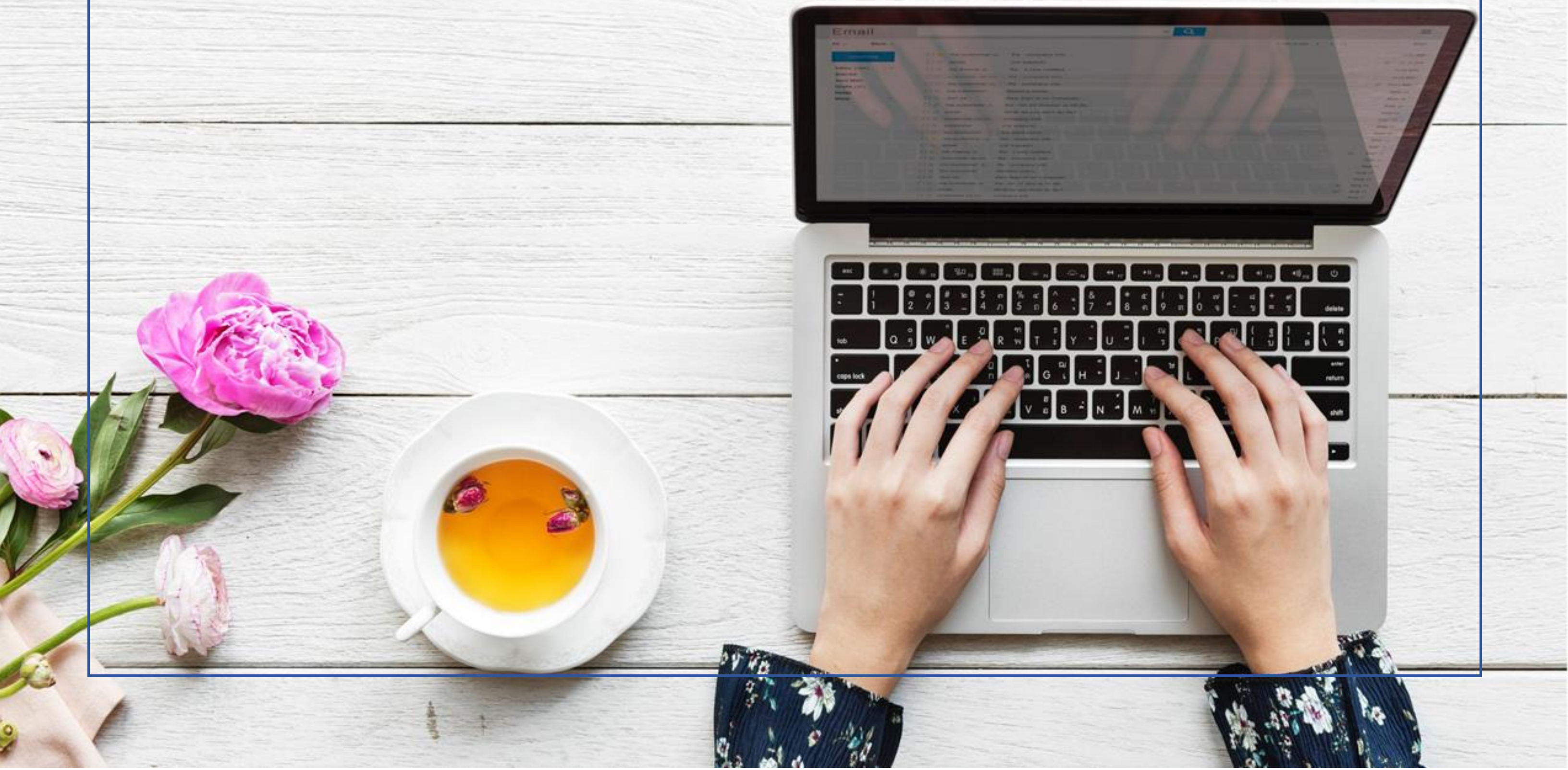


TABLE OF CONTENTS

SCENARIO

The Lidl purchasing group wants to expand our candy offering. These are store brand candies that we sell 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.

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

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

STEP 1: IDENTIFY THE BUISNESS PROBLEM

The idea is to create a brand-new product.

Which product would be preferred by most customers? Combination of desserts/candies/flavors?

- break down sales into quantity and price

STEP 2: ANALYTICS OBJECTS

Supervised Learning

Random forest is used in e-commerce to determine whether a customer will actually like the product or not. The candy with the highest number of votes will be your final choice for the trip.

Unsupervised Learning

Association rule algorithms, such as apriori principle and Markov Chain, can uncover hidden patterns and relationships in our data. For example, we can know which products are often viewed and bought together

- sequence or relationship between the data points as output

“Can we promote new products that customers are likely to buy?

What makes some candies more desirable than others?

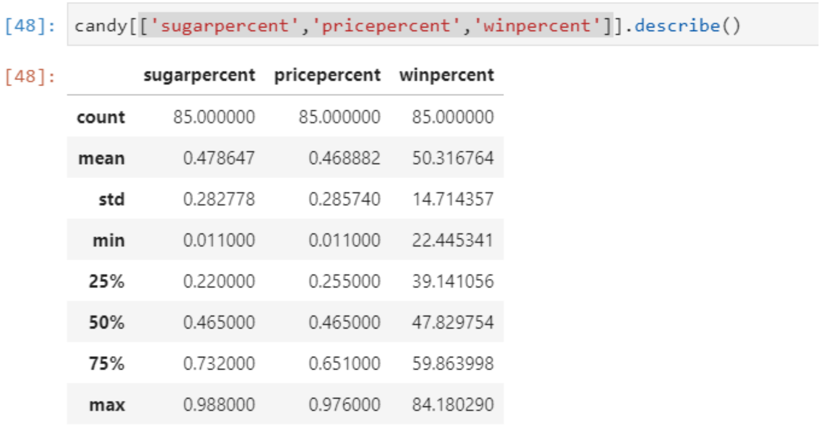
STEP 3: DATA PREPARATION

Sources and Extraction

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/)

Cleansing and Transformation

- checking the head of the data: the data types
- shape (85, 13) : gives an idea of how to handle the small dataset
- describe(): getting statistical data from the numerical data



```
[48]: candy[['sugarpercent','pricepercent','winpercent']].describe()
```

[48]:

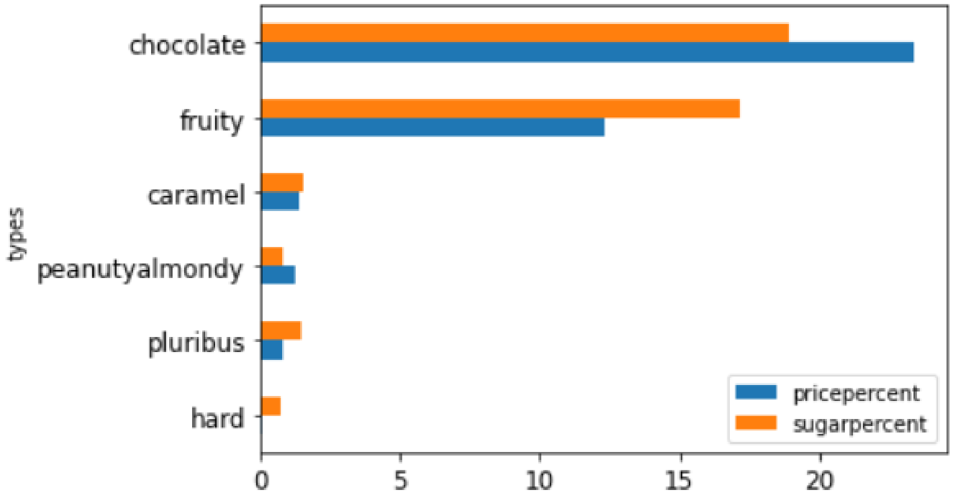
	sugarpercent	pricepercent	winpercent
count	85.000000	85.000000	85.000000
mean	0.478647	0.468882	50.316764
std	0.282778	0.285740	14.714357
min	0.011000	0.011000	22.445341
25%	0.220000	0.255000	39.141056
50%	0.465000	0.465000	47.829754
75%	0.732000	0.651000	59.863998
max	0.988000	0.976000	84.180290

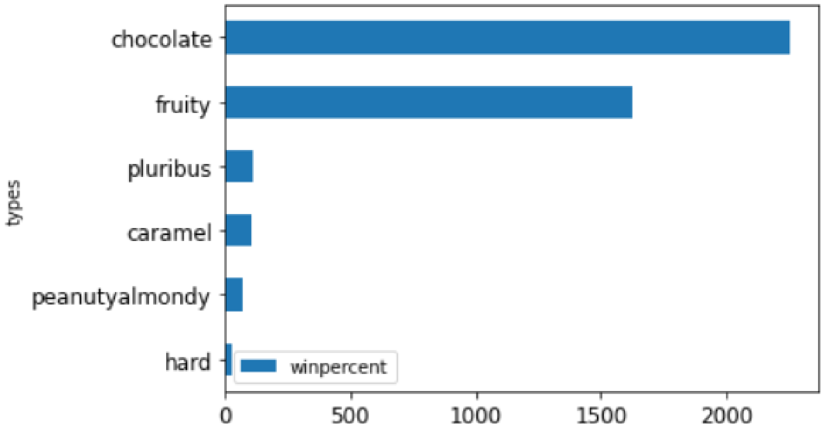
- null/NaN/None or missing values don’t exist

  For necessary visualizations the One-hot Encoding was reversed by adding a column as a flavor type.

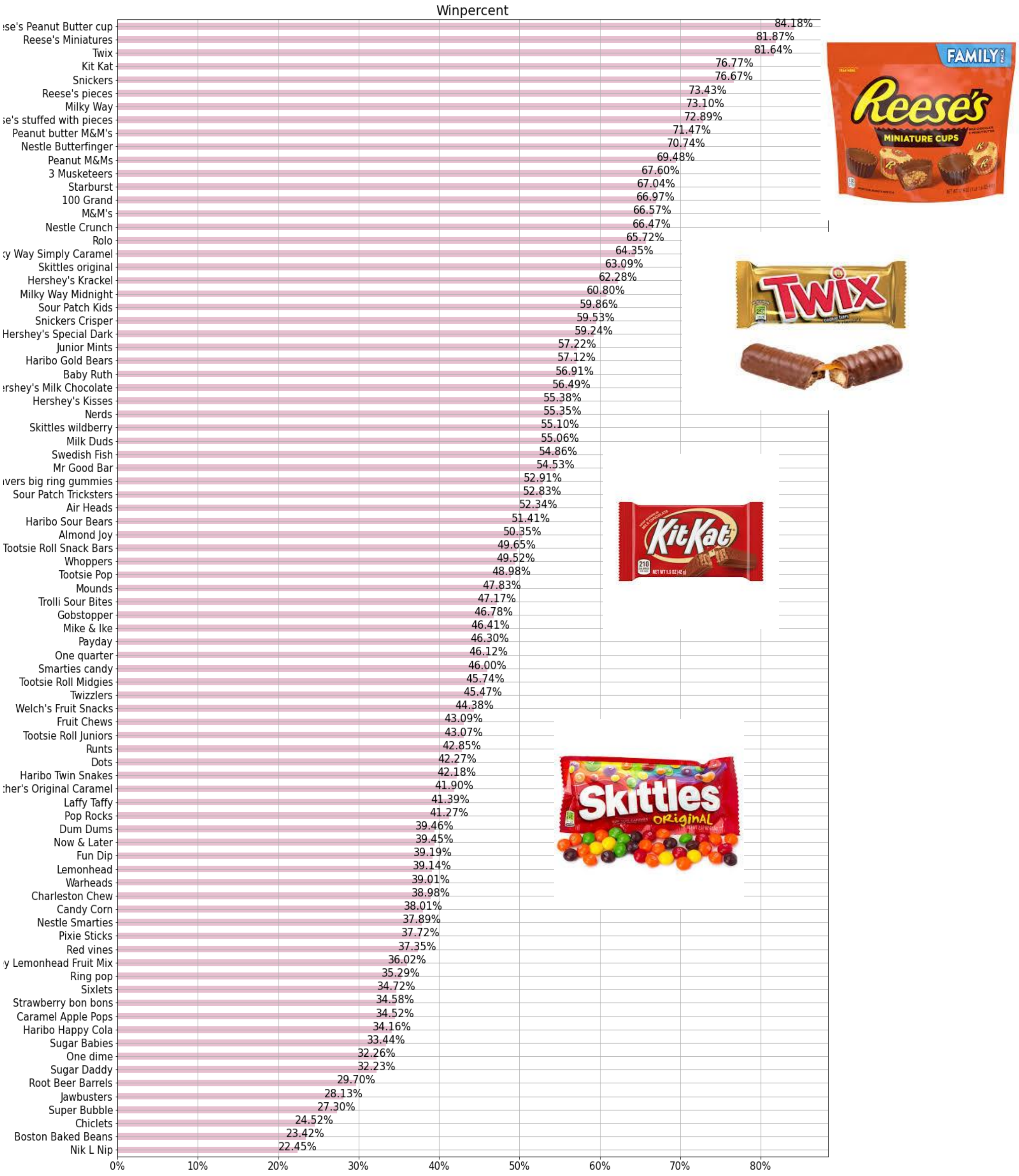
Exploration and Visualization

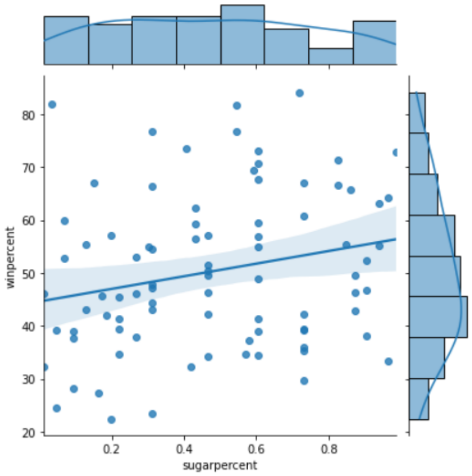
By extracting the summation of the flavor types and plotting it with 'winpercent', I answered to the question, "**Which flavor seems to be the most preferred by the clients?"**

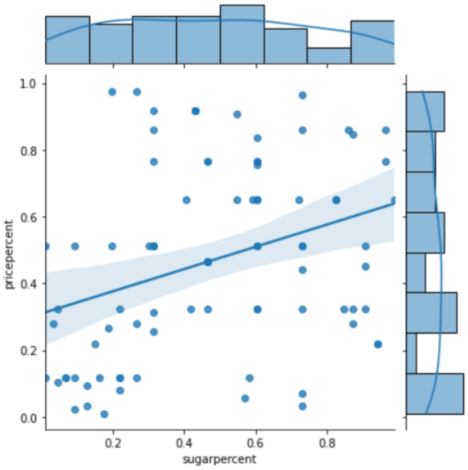




```
Clearly chocolate plays a big Role here
```


Winpercent
ese's Peanut Butter cup 84.18%
Reese's Miniatures 81.87%
Twix 81.64%
Kit Kat 76.77%
Snickers 76.67%
Reese's pieces 73.43%
Milky Way 73.10%
se's stuffed with pieces 72.89%
Peanut butter M&M's 71.47%
Nestle Butterfinger 70.74%
Peanut M&Ms 69.48%
3 Musketeers 67.60%
Starburst 67.04%
100 Grand 66.97%
M&M's 66.57%
Nestle Crunch 66.47%
Rolo 65.72%
ky Way Simply Caramel 64.35%
Skittles original 63.09%
Hershey's Krackel 62.28%
Milky Way Midnight 60.80%
Sour Patch Kids 59.86%
Snickers Crisper 59.53%
Hershey's Special Dark 59.24%
Junior Mints 57.22%
Haribo Gold Bears 57.12%
Baby Ruth 56.91%
ershey's Milk Chocolate 56.49%
Hershey's Kisses 55.38%
Nerds 55.35%
Skittles wildberry 55.10%
Milk Duds 55.06%
Swedish Fish 54.86%
Mr Good Bar 54.53%
ivers big ring gummies 52.91%
Sour Patch Tricksters 52.83%
Air Heads 52.34%
Haribo Sour Bears 51.41%
Almond Joy 50.35%
Tootsie Roll Snack Bars 49.65%
Whoppers 49.52%
Tootsie Pop 48.98%
Mounds 47.83%
Trolli Sour Bites 47.17%
Gobstopper 46.78%
Mike & Ike 46.41%
Payday 46.30%
One quarter 46.12%
Smarties candy 46.00%
Tootsie Roll Midgies 45.74%
Twizzlers 45.47%
Welch's Fruit Snacks 44.38%
Fruit Chews 43.09%
Tootsie Roll Juniors 43.07%
Runts 42.85%
Dots 42.27%
Haribo Twin Snakes 42.18%
ther's Original Caramel 41.90%
Laffy Taffy 41.39%
Pop Rocks 41.27%
Dum Dums 39.46%
Now & Later 39.45%
Fun Dip 39.19%
Lemonhead 39.14%
Warheads 39.01%
Charleston Chew 38.98%
Candy Corn 38.01%
Nestle Smarties 37.89%
Pixie Sticks 37.72%
Red vines 37.35%
y Lemonhead Fruit Mix 36.02%
Ring pop 35.29%
Sixlets 34.72%
Strawberry bon bons 34.58%
Caramel Apple Pops 34.52%
Haribo Happy Cola 34.16%
Sugar Babies 33.44%
One dime 32.26%
Sugar Daddy 32.23%
Root Beer Barrels 29.70%
Jawbusters 28.13%
Super Bubble 27.30%
Chiclets 24.52%
Boston Baked Beans 23.42%
Nik L Nip 22.45%
0%
10%
20%
30%
40%
50%
60%
70%
80%
FAMILY
Reese's
MINIATURE CUPS
TWIX
KitKat
Skittles
ORIGINAL

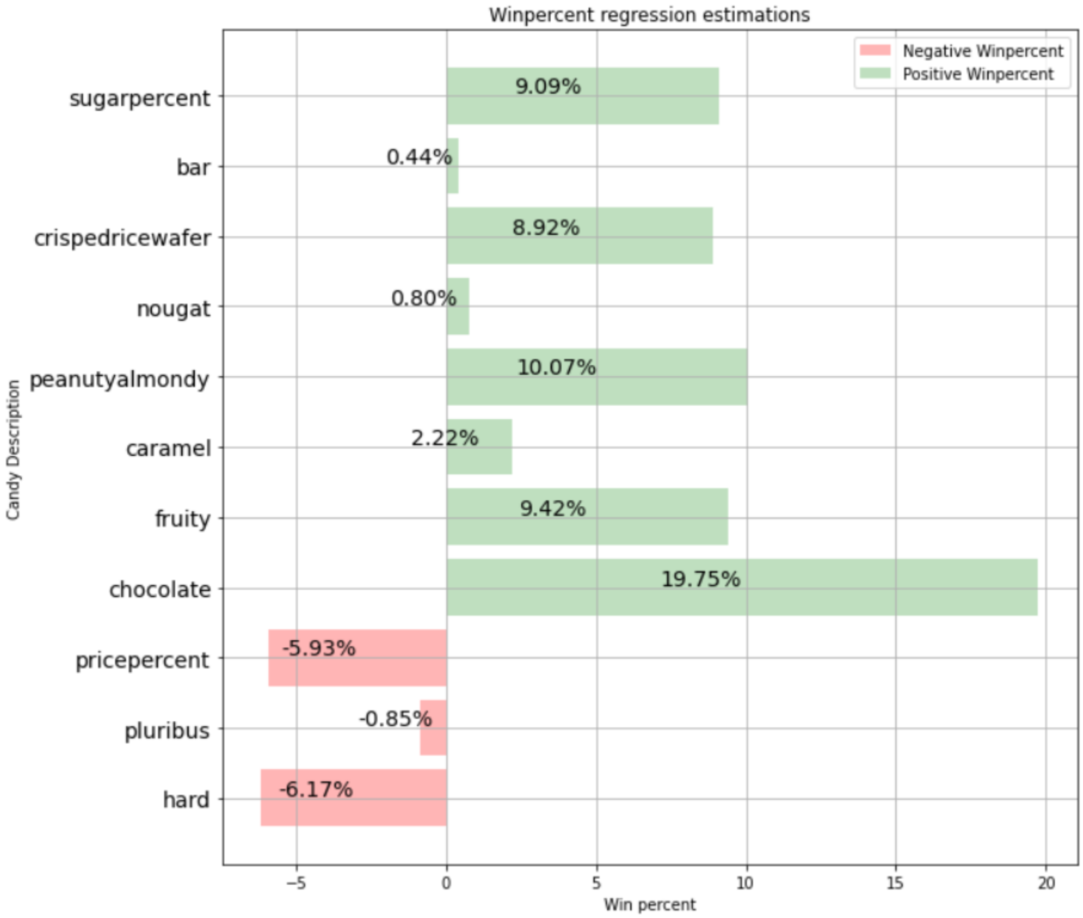




Winpercent, sugarpercent and *pricepercent* seem to be really sparse in relation. But how can we see if it is a confounding bias error?

STEP 4: MODEL DEVELOPMENT

Linear Regression



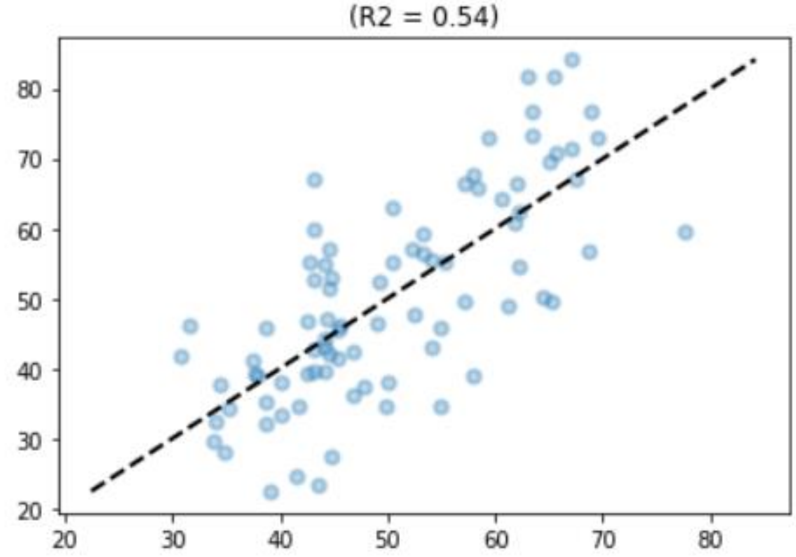
With a simple Linear Regression model we can predict how well each feature can do regarding the winpercent.

The coefficient is a factor that describes the relationship with an unknown variable.

- Having chocolate in the candy, increases the average winpercent by 19,75 %
- Having peanutyalmondy flavor, increases the winpercent by 10,1 %
- Increasing the sugarpercent by 1 unit, the winpercent increases by 9.1%
- Increasing the pricepercent by 1 unit, the winpercent decreases by 5.9%

`I see chocolate and peanutalmondy is taking the lead here.`

How well is the model fitted?

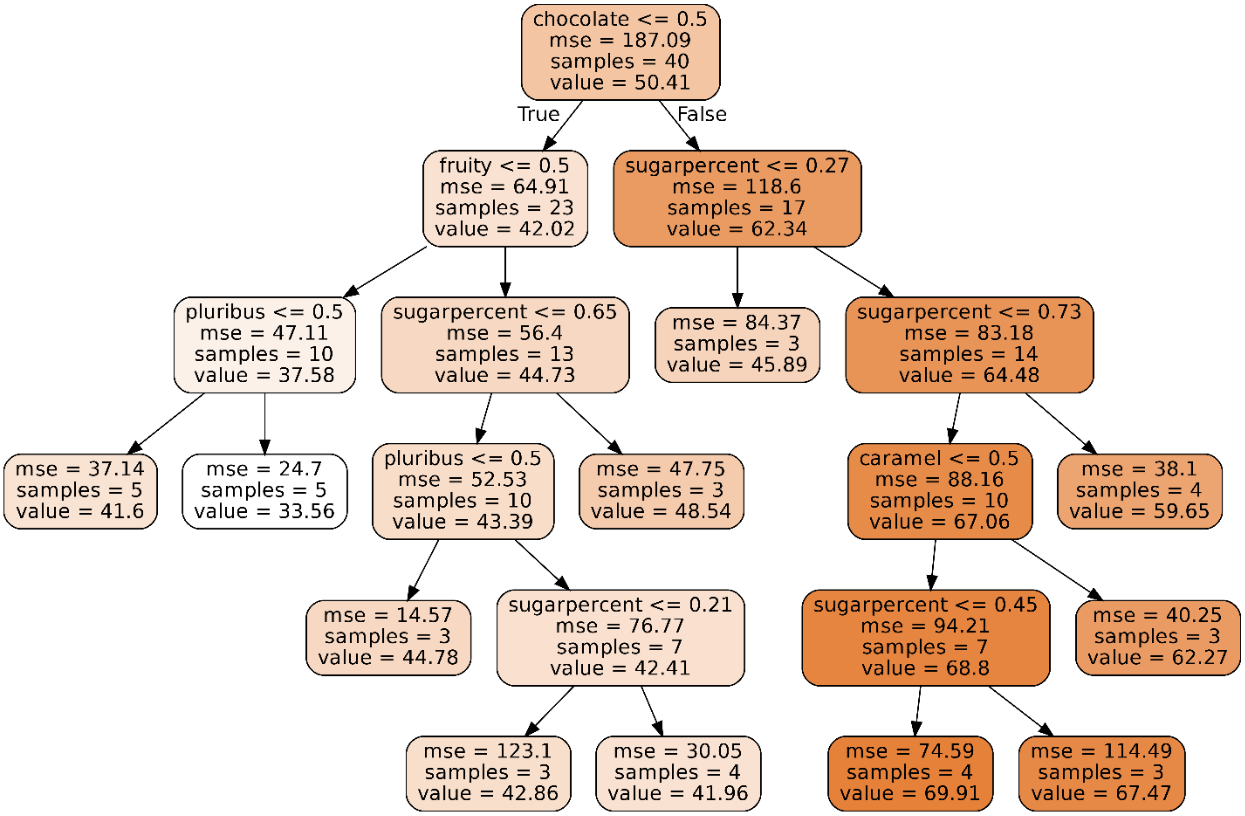


The line seems to be fitted well with an R = 0.54

However, this seems not to help with making a decision on choosing a new product.

Random Forest

This model will focus on the path that a customer may take on choosing a candy to buy. It may try to simulate how well a new product is going, regarding the *winperecent* target.



What we deduct from this tree is that following the path of chocolate 0.55% -> sugarpercent 28% -> sugarpercent 70% -> caramel 50% -> sugarpercent 40% -> **value 69.91**

CONCLUSION

What is the conclusion? Concerning the data explored I can suggest first of all a larger dataset.

Deducting from the models analysed ***a caramel, chocolate with sugarpercent around 40%*** would be preferred in the market with an accuracy of 81%.

However, additional data would be required:

- Age, careers and season of the year would be helpful additional features
- A dataset above 150 entries.

*Jupyter Notebook is attached supporting the data and conclusions.