CORRELATIONS
RANKINGS
MACHINE LEARNING
LOGISTIC REGRESSION
REGRESSION TREES

PYTHON

EXPANDING OUR CANDY BRAND

CANDY POWER RANKING

DATA BY MIGUEL WWW.DATASCIENTIST.CZ

Executive summary

Problem → Design a new kind of bar for a supermarket chain in the Czech Republic (home brand)

- Supermarket's targets → **quality and price**
- Variables that are correlated with high perceived quality, are also correlated with high prices (good for quality, bad for price)

Solution?

Find the hidden gems → low price, high success.

How

Success correlates most with chocolate presence, bar shape and peanut-almond taste

Cheap prices correlate with fruity taste, hard texture or multi-packaging

Analysis

- Two candies are in the top quartiles for cheap price and success: Skittles original and Starbust
- The top 5 from rank analysis are: Reese's Miniatures, Starburst, Sour Patch Kids, Hershey's Kisses, Skittles original

Recommendations

- From the best ranked candies, there are only two that have chocolate (most correlated with success). Reese's Miniatures and Hershey's Kisses.
- Hershey's Kisses are culturally specific (since 1907) for the Czech market better Reese's

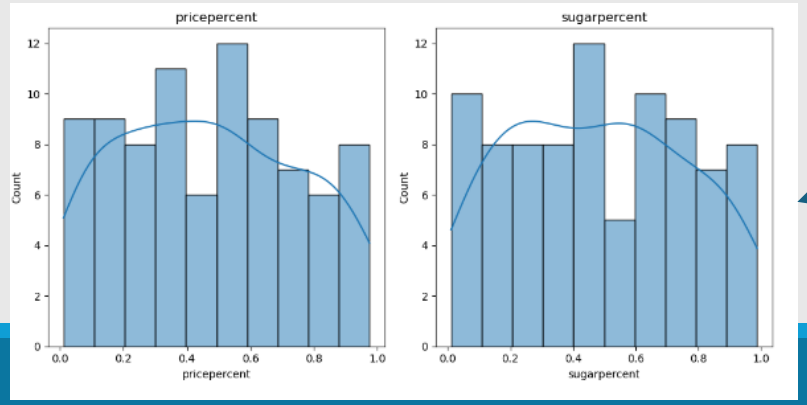
The dataset

Candy Power Ranking: a dataset widely used for projects on data analysis
https://github.com/fivethirtyeight/data/tree/master/candy-power-ranking

Provided by FiveThirtyEight under the Creative Commons Attribution 4.0 International license

- Data for the US market (opinions from US consummers)
- I assumed that the data are true and representative
- Winpercent interpreted as success (perceived quality) type „the more the better“.
- A new variable cheapness has ben created as cheapness=(1-pricepercent). To be also, „the more the better“
- Target variables: **cheapness** and **winpercent**
- Clean dataset no NaN, percentiles with uneven distribution → **probably some data were removed**

Header	Description
chocolate	Does it contain chocolate?
fruity	Is it fruit flavored?
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?
sugarpercent	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.



If sugarpercent and pricepercent are percentiles (as described), their distributions should be uniform and not like these

What can we do with the data?

Target variables for a supermarket:
winpercent (success)
cheapness (1-pricepercent)

Analysis

- Correlations analysis: Does a FAMD Analysis (similar to PCA analysis) make sense or not?
- Colinearity should be analyzed and values over 10 considered
- Classification through quartiles to see how many candies are performing well in both target variables
- Rank analysis can be used study ranks in each target variable and a joint rank

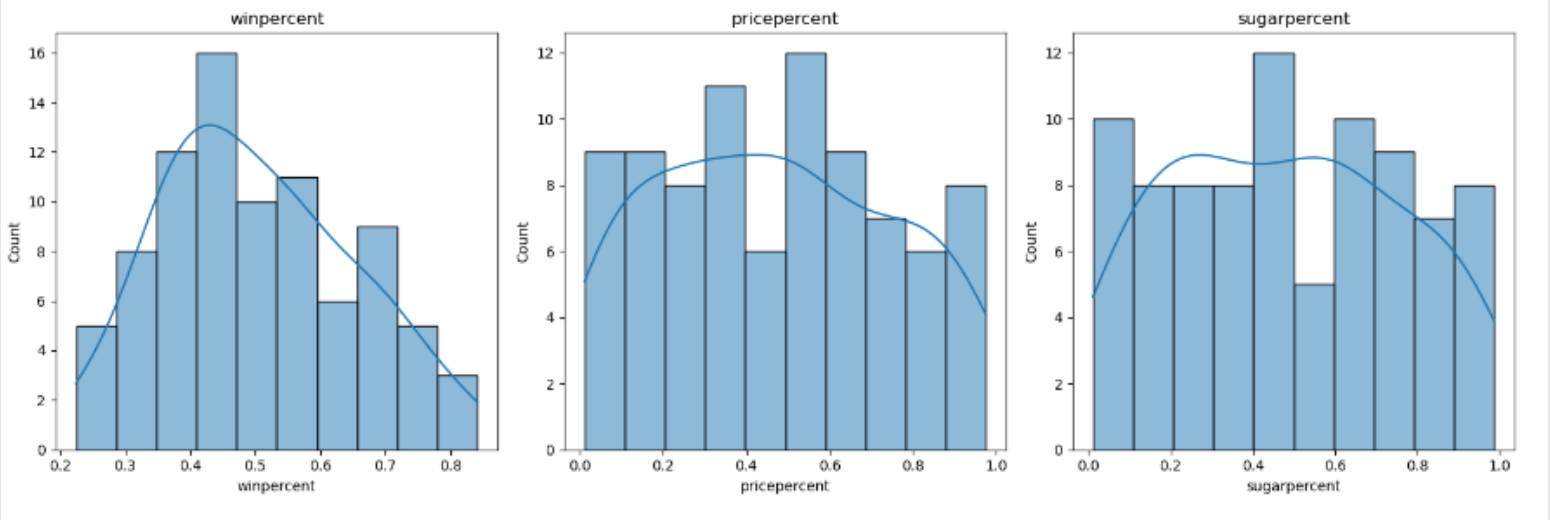
ML Models

- Since the independent variables are categorical (binary) and numerical, the techniques that seem adequate are **Logistic Regression** and **Regression Tree Analysis**.
- Study target variables: winpercent, cheapness and/or combination of both.
- Weighted combinations have not been studied, since there is no information on weights
- The models will be discarded if the accuracy is lower than 0.6 or if the depth is lower that 3
- Dataset: only 85 rows and most variables are Binary! Limited possibilities for calibration

EDA

Basic

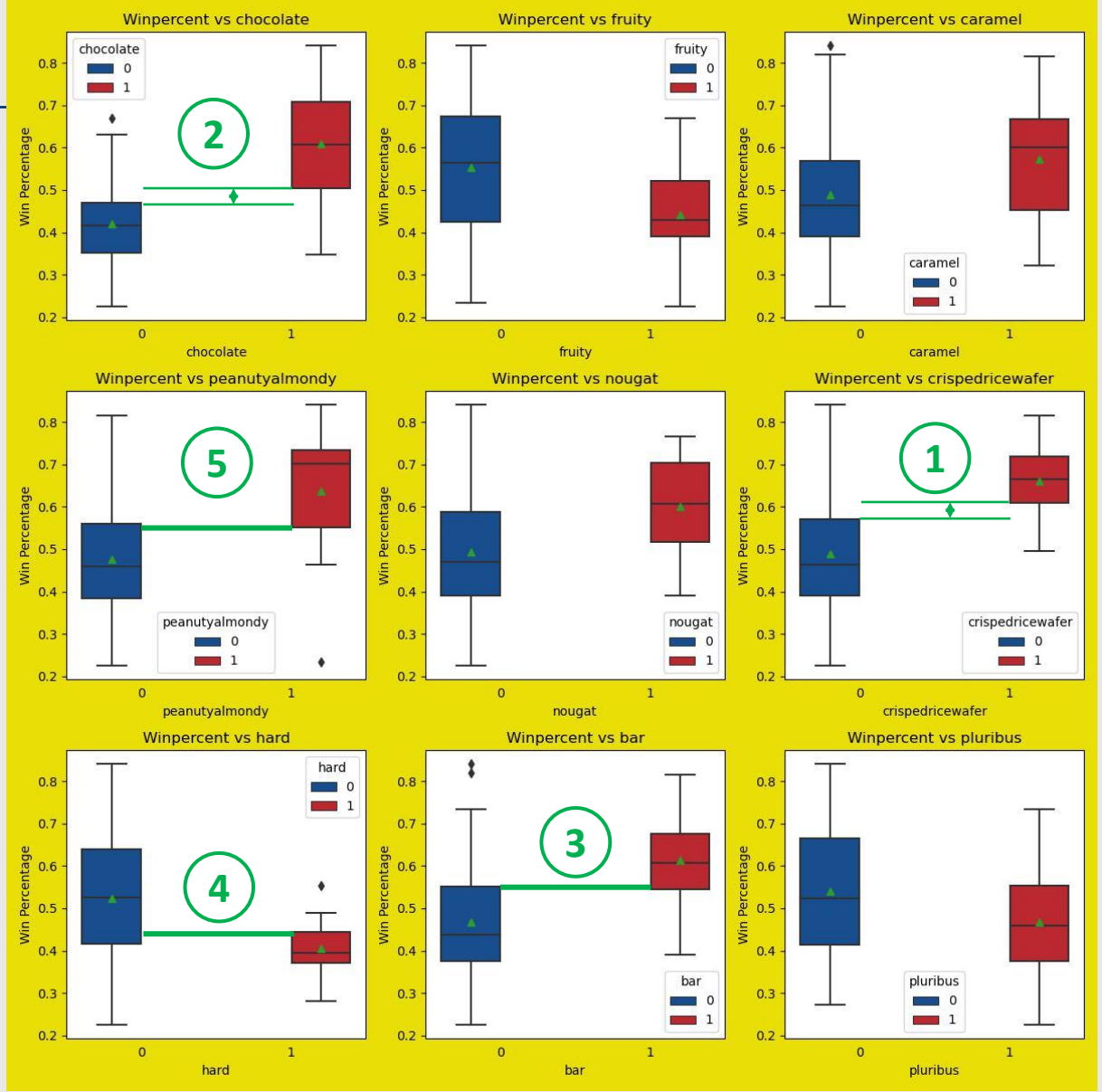
- No NaN
- Three numerical variables:
  - winpercent (percentage of success),
  - pricepercent (percentile of price)
  - sugarpercent (percentile of sugar)
- Pricepercent and sugarpercent are percentiles and should have uniform distribution → the dataset incomplete
- 9 Binary variables (0-1)
- No missing values



EDA

Box and whiskers Winpercent vs. Binary variables

- Variables that a priory are interesting for predictions are those whose distributions do not overlap or overlap less.
    1. Crispedricedwafer
    2. Chocolate
    3. Winpercent
    4. Caramel
    5. Peanutyalmondy



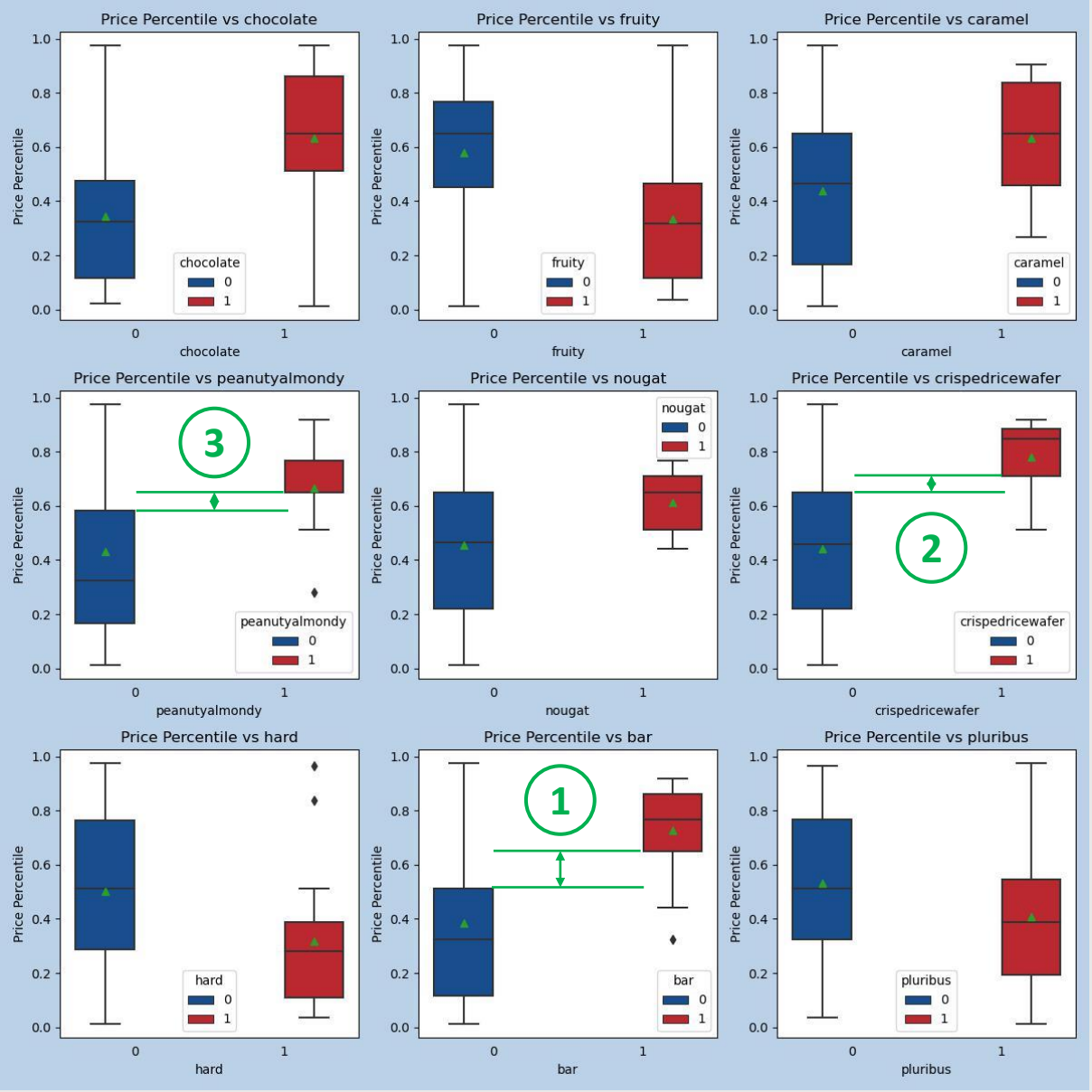
EDA

Box and whiskers pricepercent vs. Binary variables

- In the case of price prediction, the variables a priori more interesting are:
  1. Bar
  2. Crispedricewafer
  3. Peanutyalmondy

  Chocolate and fruity have very wide distributions over price: a candy can have any price regardless of the presence of chocolate or fruity taste.

  Crispedricewafer and Peanutyalmondy were also interesting for winpercent prediction



EDA

Top 10 beloved candies

```
: 1 print(data.sort_values("winpercent", ascending=False).head(10))
```

	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	0.720	0.651	0.841803
51	Reese's Miniatures	1	0	0	1	0	0	0	0	0	0.034	0.279	0.818663
79	Twix	1	0	1	0	0	1	0	1	0	0.546	0.906	0.816429
28	Kit Kat	1	0	0	0	0	1	0	1	0	0.313	0.511	0.767686
64	Snickers	1	0	1	1	1	0	0	1	0	0.546	0.651	0.766738
53	Reese's pieces	1	0	0	1	0	0	0	0	1	0.406	0.651	0.734350
36	Milky Way	1	0	1	0	1	0	0	1	0	0.604	0.651	0.730996
54	Reese's stuffed with pieces	1	0	0	1	0	0	0	0	0	0.988	0.651	0.728879
32	Peanut butter M&M's	1	0	0	1	0	0	0	0	1	0.825	0.651	0.714651
42	Nestle Butterfinger	1	0	0	1	0	0	0	1	0	0.604	0.767	0.707356

from the top 10 with highest winpercent
only Reeses miniatures has low price
cheap and successful!

Correlations

- Only significant correlation is chocolate, correlated with **winpercent**
- Everything correlated with winpercent is inversely correlated with cheapness and the other way round

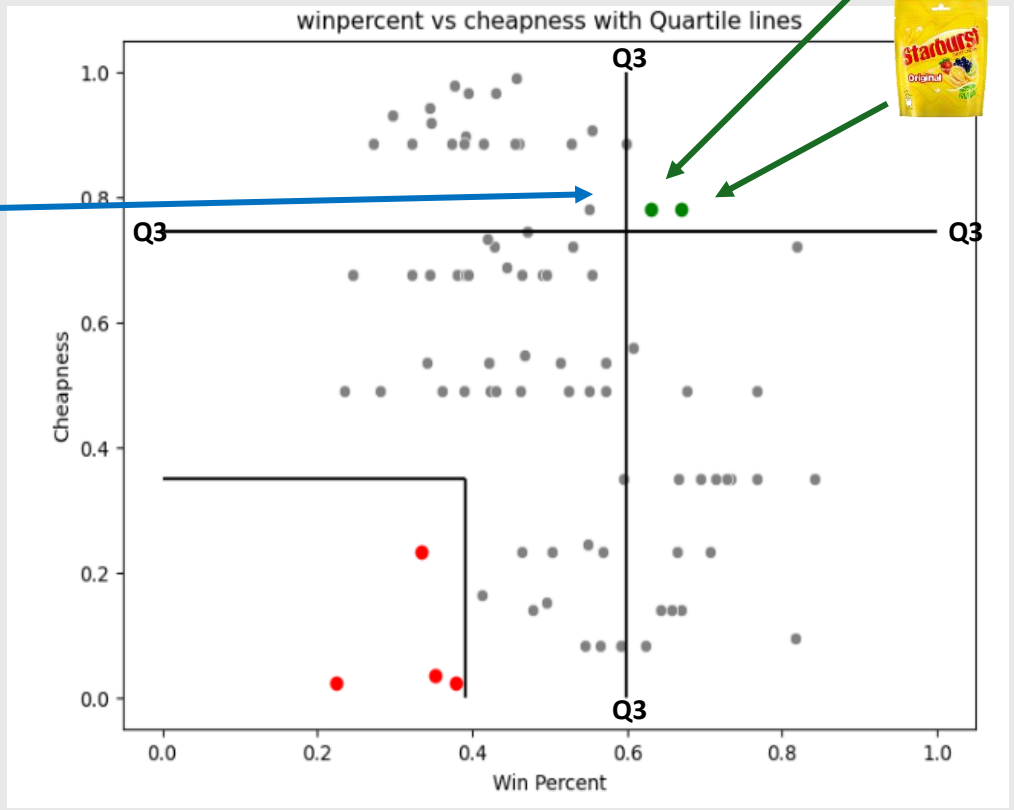
What is good for winpercent is bad for cheapness!

Quartiles

Which candies are in the upper quartile for cheapness AND winpercent?

- Only two candies within both fourth quartile interval: Starburst and Skittles
- Both are atypical: lacking chocolate, lacking peanuty-almondy taste → **brand and history** (Skittles since 1974)

Starburst and Skittles forbidden in the EU!!
(TiO_2 – E171) since 2022, reputational concern now

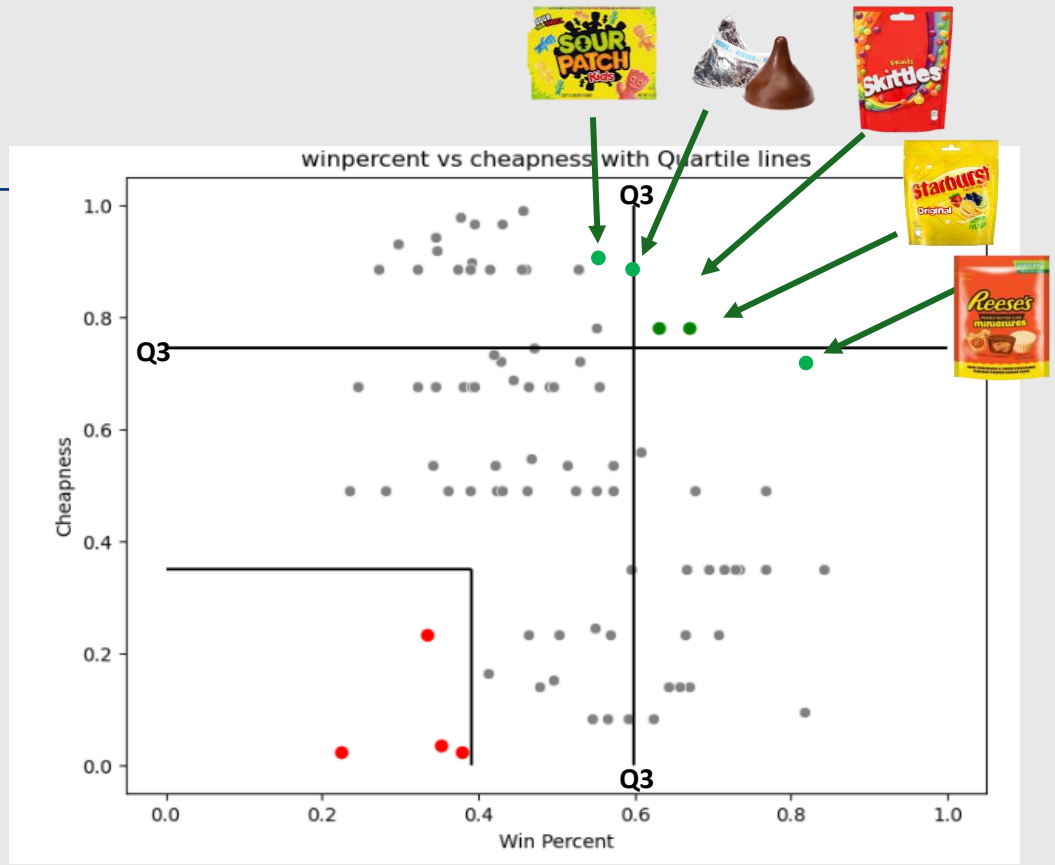


	chocolate	fruity	caramel	peanuty	nougat	crispy	hard	bar	pluribus	sugar
Skittles	X	✓	X	X	X	X	X	X	✓	0.941
Starbust	X	✓	X	X	X	X	X	X	✓	0.151

Analysis of ranks

Ranking

- Winpercent rank, cheapness rank and combined rank.
- Similar results as quartile analysis
- Two candies with chocolate (most correlated to winpercent)
- Reese's miniature is among the cheapests and low procent of sugar
- All of them seem interesting as potential ideas for Lidl's home brand



Top 5 in the ranking

	chocolate	fruity	caramel	peanuty	nougat	crispy	hard	bar	pluribus	sugar
Reese‘s miniature	✓	X	X	✓	X	X	X	X	X	0.034
Starbust	X	✓	X	X	X	X	X	X	✓	0.151
Sour Patch Kids	X	✓	X	X	X	X	X	X	✓	0.069
Hershey‘s kisses	✓	X	X	X	X	X	X	X	✓	0.127
Skittles	X	✓	X	X	X	X	X	X	✓	0.941

Discarded models

Logistic Regression

- Target variable was:
  $$target = winpercent + cheapness$$
- Target converted into quartiles (must be categorical)
- Only 85 rows, split in 80% for callibration and 20% for testing
- Low accuracy → 0,4706
- Coefficients not consistent with analysis of correlations: Supposed to be not fruity and not chocolate…

Logistic regression accepts both categorical and numerical variables as independent variables, but the target must be categorical (split into quartiles…)

Coefficients of the model

```
# Get the coefficients and intercept of the model
coefficients = model.coef_[0]
intercept = model.intercept_

# Create a DataFrame to display the coefficients
coefficients_df = pd.DataFrame({"Feature": x_variables, "Coefficient": coefficients})

# Add the intercept to the DataFrame
coefficients_df = coefficients_df.append({"Feature": "Intercept", "Coefficient": intercept[0]}, ignore_index=True)

print("Coefficients:")
print(coefficients_df)
```

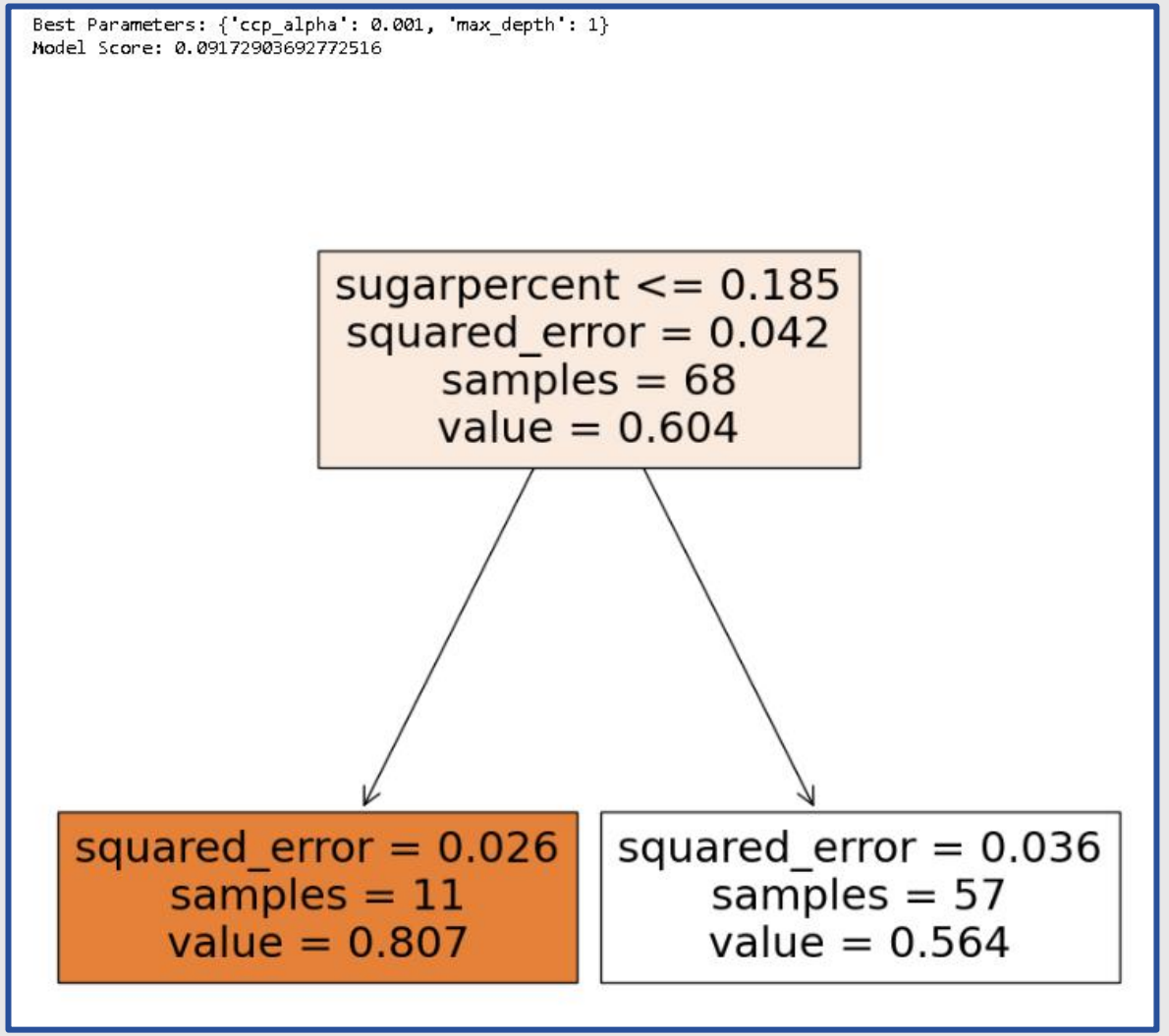
```
Coefficients:
              Feature  Coefficient
0           chocolate    -0.505709
1              fruity    -0.585758
2             caramel    -0.006973
3      peanutyalmondy     0.145830
4              nougat     0.008482
5   crispedricewafer    -0.096124
6                hard     0.333883
7                 bar     1.219606
8             pluribus     0.442362
9        sugarpercent     0.060675
10          Intercept    -0.050955
```

Discarded models

Regression Tree Model

- Only 85 rows, split in 80% for callibration and 20% for testing.
- Low accuracy
- Best parameters: max_depth=1, based in sugarpercent… Not useful for decision making on choice of candy bars

Regression tree accepts both categorical and numerical variables as independent variables, and as target (no need to convert into categorical)



Conclussions

- All variables correlated with successful candies are correlated also with expensive prices.
- Chocolate is the most relevant variable concerning correlation with success (winpercent)
- From the analysis of rankings, 5 candies appear as most interesting:
  - Cocolate is present in 2 of them, which are also relatively low in sugar: **Reese's** and **Hershey's**
  - **Hershey's kisses** are kind of a cultural tradition in the US (since 1907). Problem of adaptation to the Czech market
  - **Sour patch kids** might be an option but similar to existing products (*kyselé žížalky* and *kyselé rybičky*)
  - **Skittles** and **Starbust** have the problem of TiO2. Image and legal issues in the EU and Skittles with high sugar.

		chocolate
1	Reese‘s miniature	✓
2	Starbust	X
3	Sour Patch Kids	X
4	Hershey‘s kisses	✓
5	Skittles	X

Reese‘s miniature seems to be the winner
Best ranking overall
With chocolate
Relatively low sugar

Possible discussion

- In this model, the target variables to optimize have been cheapness and winpercent
- No weights have been given to the variables: a panel of experts or the management strategy might determine different weights for each variable:

$$target = W_{win} * winpercent + W_{cheap} * cheapness$$

- The management of the supermarket might be interested in the opposite price strategy: identify very expensive and very successful candies so that the home brand copies can be cheaper

Thank you!

DATA BY MIGUEL
WWW.DATASCIENTIST.CZ
MIGUEL@DATASCIENTIST.CZ