Case study: *gummy-like* or *cookie-like* candies?

Problem statement

- We want to create a brand-new candy.
- We have a **market survey** on **customer sentiment** for competitor brands of candies.
- We want our analysis to be driven by the **maximization of customer preference**.
- We need to choose what the new candy will be: a **gummy** or a **cookie-based sweet**?

Recommendation

Results from data analysis suggest that we should opt for a **cookie-based sweet**.

Assumptions and procedure

The market survey does not specify which candies are gummy-like or cookie-like.

So, first of all we need to create two clusters (based on the properties listed in the marked survey).
We assume that gummies and cookies can be classified as following:

gummy-like	*cookie-like*
• **Not hard**: they are very different from hard candies. • **Fruity**: often, gummies have fruity flavours. • **Pluribus**: often, gummies come in packages which contains many of the same type.	• **Not hard**: they are very different from hard candies. • **Crispedricewafer**: of course they must have a crispy texture, similar to cookies.

Results of the analysis

First observations

So far, the likeability of a candy is indicated by a continuous value (*winpercent*). We want to simplify things:

We say that a candy is a *winner* if its winpercent is > 50

This means that, in a random match with another candy, a winner candy has more than 50% the chance of being preferred by a customer.

What percentage of *cookie-like* candies in our dataset are winners?
What about *gummy-like* candies?

```
Percentage of cookie-like candies that are winners:  85.71428571428571
Percentage of gummmy-like candies that are winners:  42.10526315789473
```

Results of the analysis

Correlation matrix

Let's now look at flavours and physical features of candies.

	chocolate	fruity	caramel	peanutyalmondy	nougat	crispedricewafer	hard	bar	pluribus	sugarpercent	pricepercent	winpercent
chocolate	1.0000	-0.7417	0.2499	0.3778	0.2549	0.3412	-0.3442	0.5974	-0.3397	0.1042	0.5047	0.6365
fruity	-0.7417	1.0000	-0.3355	-0.3993	-0.2694	-0.2694	0.3907	-0.5151	0.2997	-0.0344	-0.4310	-0.3809
caramel	0.2499	-0.3355	1.0000	0.0594	0.3285	0.2131	-0.1224	0.3340	-0.2696	0.2219	0.2543	0.2134
peanutyalmondy	0.3778	-0.3993	0.0594	1.0000	0.2131	-0.0176	-0.2056	0.2604	-0.2061	0.0879	0.3092	0.4062
nougat	0.2549	-0.2694	0.3285	0.2131	1.0000	-0.0897	-0.1387	0.5230	-0.3103	0.1231	0.1532	0.1994
crispedricewafer	0.3412	-0.2694	0.2131	-0.0176	-0.0897	1.0000	-0.1387	0.4238	-0.2247	0.0699	0.3283	0.3247
hard	-0.3442	0.3907	-0.1224	-0.2056	-0.1387	-0.1387	1.0000	-0.2652	0.0145	0.0918	-0.2444	-0.3104
bar	0.5974	-0.5151	0.3340	0.2604	0.5230	0.4238	-0.2652	1.0000	-0.5934	0.1000	0.5184	0.4299
pluribus	-0.3397	0.2997	-0.2696	-0.2061	-0.3103	-0.2247	0.0145	-0.5934	1.0000	0.0455	-0.2208	-0.2474
sugarpercent	0.1042	-0.0344	0.2219	0.0879	0.1231	0.0699	0.0918	0.1000	0.0455	1.0000	0.3297	0.2292
pricepercent	0.5047	-0.4310	0.2543	0.3092	0.1532	0.3283	-0.2444	0.5184	-0.2208	0.3297	1.0000	0.3453
winpercent	0.6365	-0.3809	0.2134	0.4062	0.1994	0.3247	-0.3104	0.4299	-0.2474	0.2292	0.3453	1.0000

Some features are heavily correlated (see for example chocolate and fruity).

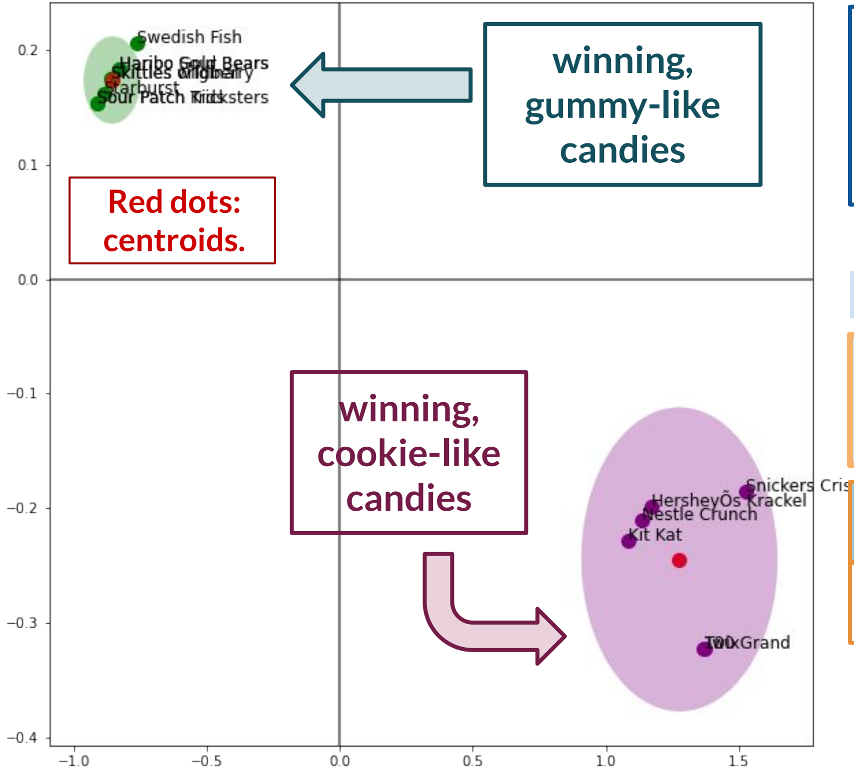
There most likely are **redundant variables.**

This suggest an approach of **dimensionality reduction** before trying to extract value from the data.

Results of the analysis

Principal Components Analysis

We reduce the dimensionality of the dataset by applying **Principal Components Analysis**. Now every candy can be described by **only two variables**, so we can plot them on a cartesian grid.



We can expect with 95% confidence that a new cookie-like/gummy-like candy will fall inside the purple/green ellipse.

Linear regression

Q: What is the expected likeability for the **mean** gummy-like or cookie-like candy?

gummy-like	*cookie-like*
42.52%	*62.02%*