Gaurang Sharma

Lidl Analytics – Data Science Case Study

Problem statement: *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. Some prefer cookie-based sweets while others think that it should be gummies. The market research data is now available, and it is your job to find out which product characteristics drive customer sentiment and subsequently make a recommendation on a new product.*

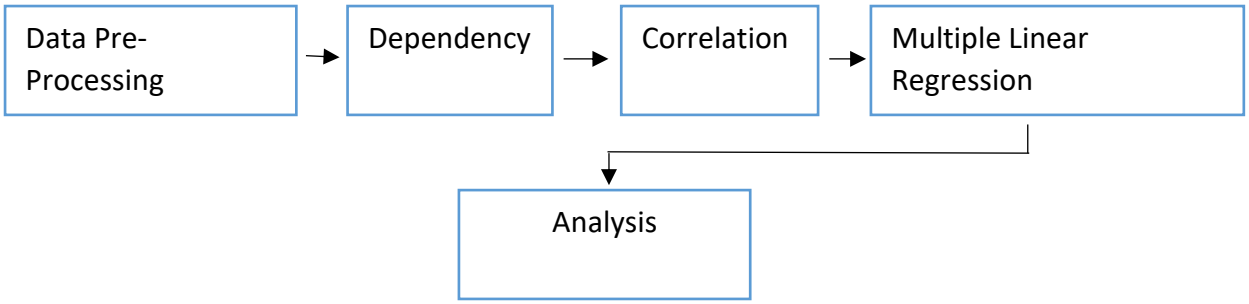
Programming language used: *Python*

Libraries implemented: *Pandas, Numpy, Scikit-learn, SciPy Bokeh, Matplotlib, Seaborn, Tkinter, Statsmodels.*

Platform: *Jupyter Notebook*

GitHub link for the code: *https://github.com/GaurangSharma44/Candy_Analysis*

Thought Process:

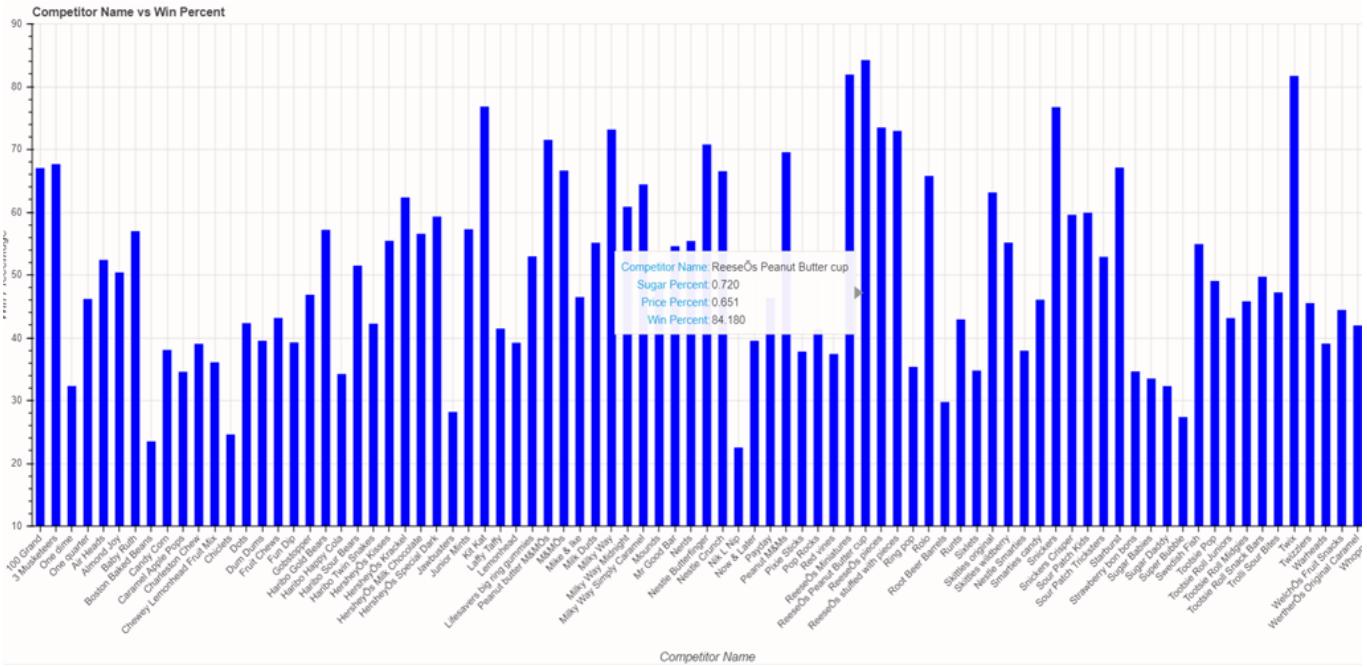


Thank you for this chance. Looking forward to the reply!

Steps:

- ### Data Pre-processing:

1. Understanding the data and figuring out how to tackle the problem was my first step. I visualised some of the data using Bokeh plots and Pandas to get familiar with it and move forward with the analysis.



2. The table shows 10 ‘Candy’ with the highest ‘Win Percent’

	competitorname	chocolate	fruity	caramel	peanutyalmondy	nougat	crispedricewafer	hard	bar	pluribus	sugarpercent	pricepercent	winpercent
52	ReeseÕs Peanut Butter cup	1	0	0	1	0	0	0	0	0	0.720	0.651	84.180290
51	ReeseÕs Miniatures	1	0	0	1	0	0	0	0	0	0.034	0.279	81.866257
79	Twix	1	0	1	0	0	1	0	1	0	0.546	0.906	81.642914
28	Kit Kat	1	0	0	0	0	1	0	1	0	0.313	0.511	76.768600
64	Snickers	1	0	1	1	1	0	0	1	0	0.546	0.651	76.673782
53	ReeseÕs pieces	1	0	0	1	0	0	0	0	1	0.406	0.651	73.434990
36	Milky Way	1	0	1	0	1	0	0	1	0	0.604	0.651	73.099556
54	ReeseÕs stuffed with pieces	1	0	0	1	0	0	0	0	0	0.988	0.651	72.887901
32	Peanut butter M&MÕs	1	0	0	1	0	0	0	0	1	0.825	0.651	71.465050
42	Nestle Butterfinger	1	0	0	1	0	0	0	1	0	0.604	0.767	70.735641

3. Reeso’s candy is quite popular, it seems, and every candy has chocolate in it.
4. Sorting through the highest ‘Win Percent’ and the lowest ‘Price Percentile’ gave a similar result, which I believe is the goal for every business.

Thank you for this chance. Looking forward to the reply!

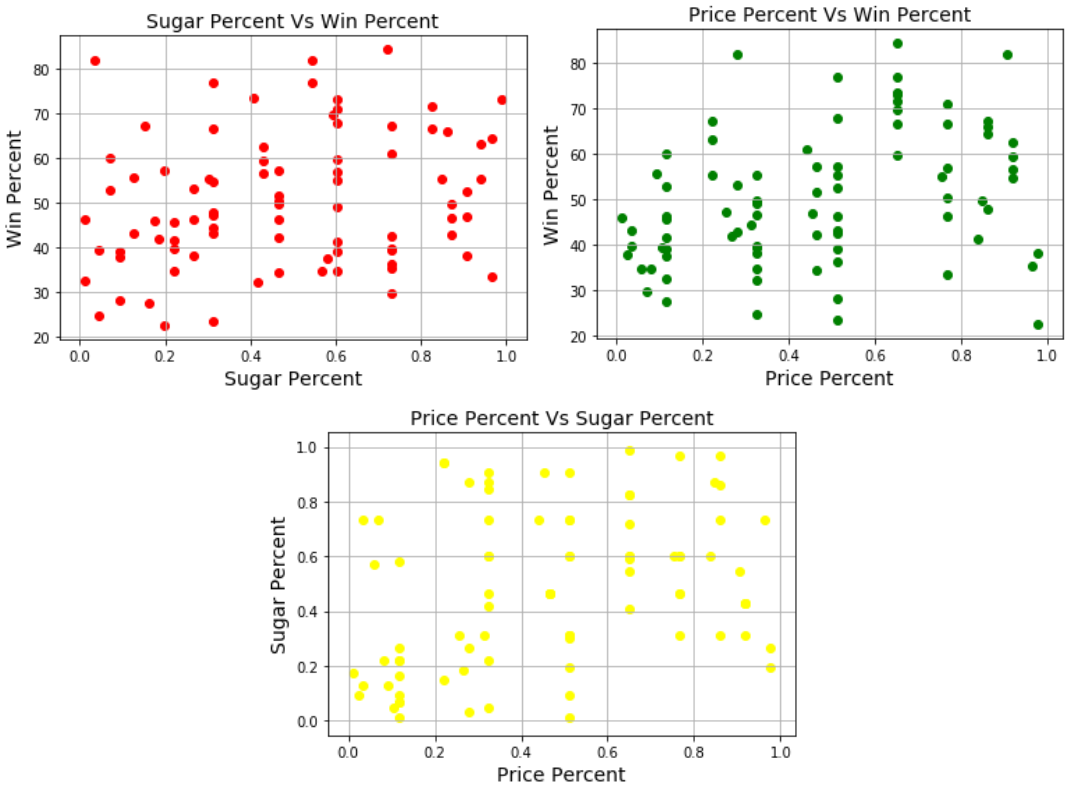
	competitorname	chocolate	fruity	caramel	peanutyalmondy	nougat	crispedricewafer	hard	bar	pluribus	sugarpercent	pricepercent	winpercent
52	ReeseÕs Peanut Butter cup	1	0	0	1	0	0	0	0	0	0.720	0.651	84.180290
51	ReeseÕs Miniatures	1	0	0	1	0	0	0	0	0	0.034	0.279	81.866257
79	Twix	1	0	1	0	0	1	0	1	0	0.546	0.906	81.642914
28	Kit Kat	1	0	0	0	0	1	0	1	0	0.313	0.511	76.768600
64	Snickers	1	0	1	1	1	0	0	1	0	0.546	0.651	76.673782
53	ReeseÕs pieces	1	0	0	1	0	0	0	0	1	0.406	0.651	73.434990
36	Milky Way	1	0	1	0	1	0	0	1	0	0.604	0.651	73.099556
54	ReeseÕs stuffed with pieces	1	0	0	1	0	0	0	0	0	0.988	0.651	72.887901
32	Peanut butter M&MÕs	1	0	0	1	0	0	0	0	1	0.825	0.651	71.465050
42	Nestle Butterfinger	1	0	0	1	0	0	0	1	0	0.604	0.767	70.735641

	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

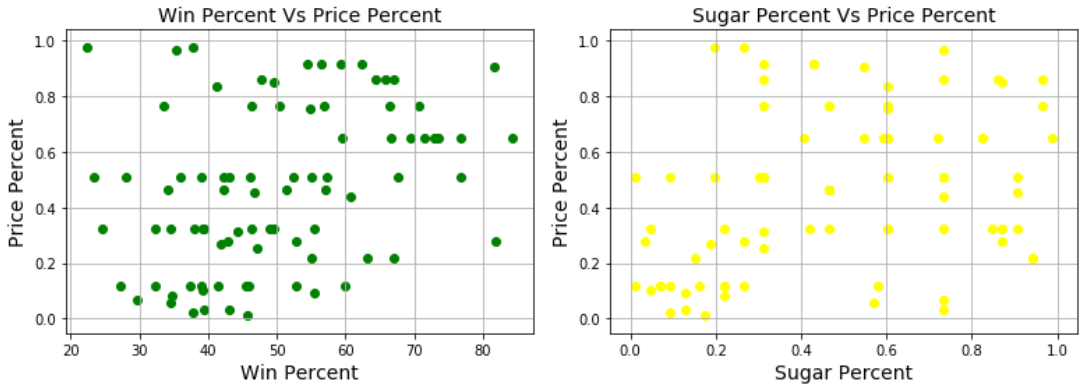
Thank you for this chance. Looking forward to the reply!

- **Dependency:**

1. I wanted to see if there was any dependency between Price Percent, Sugar Percent and Win Percent.
2. Plotted a Scatter plot for the following:



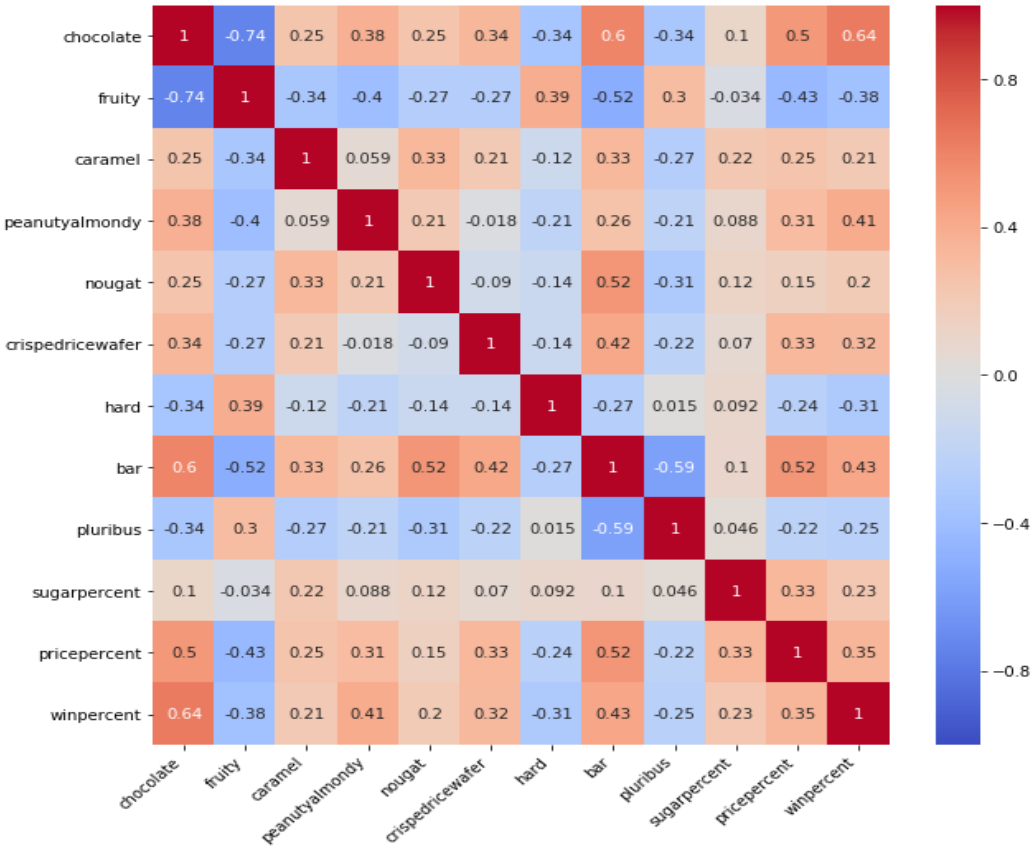
3. There was no real dependency on any of these factors, but a very slight trend was noticed; that Price could play an important role with the Win Percent.
4. Also, that Sugar affects Price.



Thank you for this chance. Looking forward to the reply!

- ## Correlation:

1. I plotted a heat map for all the columns to see which ingredient, which type of candy and if sugar & price affect win percent the most.
2. To show if there is any correlation between multiple factors.

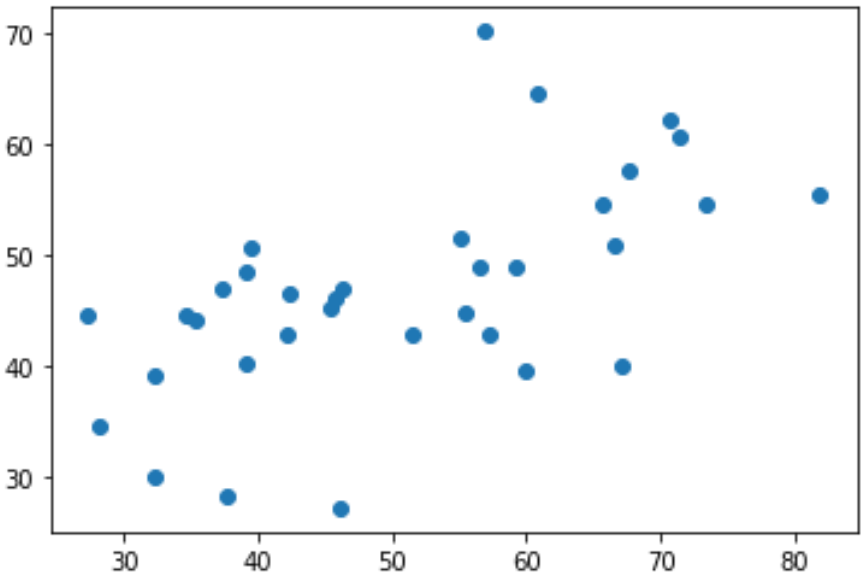


3. There can be seen that ingredient ‘Chocolate’ & ‘Peanut & Almond, type ‘Bar’ and ‘Sugar Percent’ affect ‘Win Percent’ in positive the most.
4. Also, ingredient ‘Fruity’, type ‘Bar’ and ‘Pluribus’ has a negative effect on ‘Win Percent’.

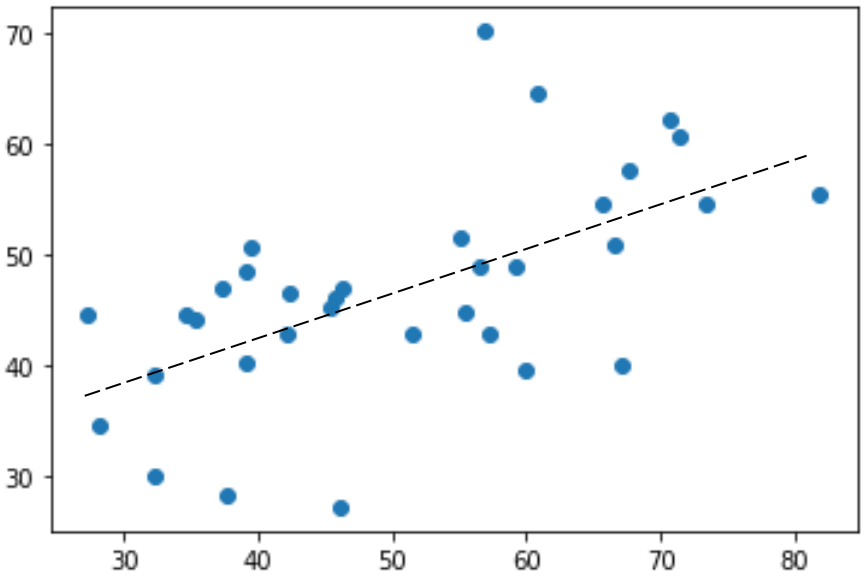
Thank you for this chance. Looking forward to the reply!

- ## Multiple Regression:

1. I, then applied Multiple Regression on the data.
2. First split up our data into an X array that contains the features to train on, and a y with the target variable, in the ‘Win Percent’ column. I tossed out the ‘Competitor Name’ column because it only has text info that the linear regression model can’t use.
3. Split the data in two parts. The test set serves as a proxy for new data.Trained data is the data on which we apply the linear regression algorithm.



4. A somewhat linear model was achieved, with this prediction model, I believe a new candy with a good amount of ‘Win Percent’ can be created.



Thank you for this chance. Looking forward to the reply!

5. A candy with the ingredients ‘Chocolate’, ‘Caramel’, ‘Peanut & Almond’ and of the type ‘Crisped Rice Wafer’ and ‘Bar’ achieves a ‘Win Percent’ of 82.

```
#chocolate','fruity','caramel','peanutyalmondy','nougat','crispedricewafer','hard','bar','pluribus','sugarpercent','pricepercent
print(lm.predict([[1,0,1,1,0,1,0,1,0,0.89,0.80]]))
```

```
[82.79164522]
```

Thank you for this chance. Looking forward to the reply!

- ## Final Analysis (Conclusion):

1. Alone ‘Sugar’ or ‘Price’ has a very low dependency on ‘Win Percent’.
2. Combination of highly rated ingredients, type, etc. with low price and high sugar can be used to create a new competitor candy.
3. The ‘Win Percent’ can be predicted by a Multiple Regression Model or any Predictive Analysis.
4. Every customer would have his personal choice of taste, but from the data available, the high rated ingredients were noticed.
5. A survey of customers and their choices can also be used to derive a more accurate prediction.
6. Not taking other aspects in regard such as, Marketing, Quality, Truthfulness of Brand, I think this is the best prediction which can be made of the given data.
7. A model / code can be written which takes in account all the features and gives out the highest achievable Win Percent, but it would take more time.

References:

- https://programminghistorian.org/en/lessons/visualizing-with-bokeh
- https://bokeh.pydata.org/en/latest/docs/user_guide/tools.html
- https://stackoverflow.com/questions/4150171/how-to-create-a-density-plot-in-matplotlib
- https://medium.com/@chrisshaw982/seaborn-correlation-heatmaps-customized-10246f4f7f4b
- https://datatofish.com/multiple-linear-regression-python/
- https://datascience.stackexchange.com/questions/39034/how-to-train-ml-model-with-multiple-variables
- https://medium.com/analytics-vidhya/linear-regression-using-python-ce21aa90ade6

Thank you for this chance. Looking forward to the reply!