Project Report

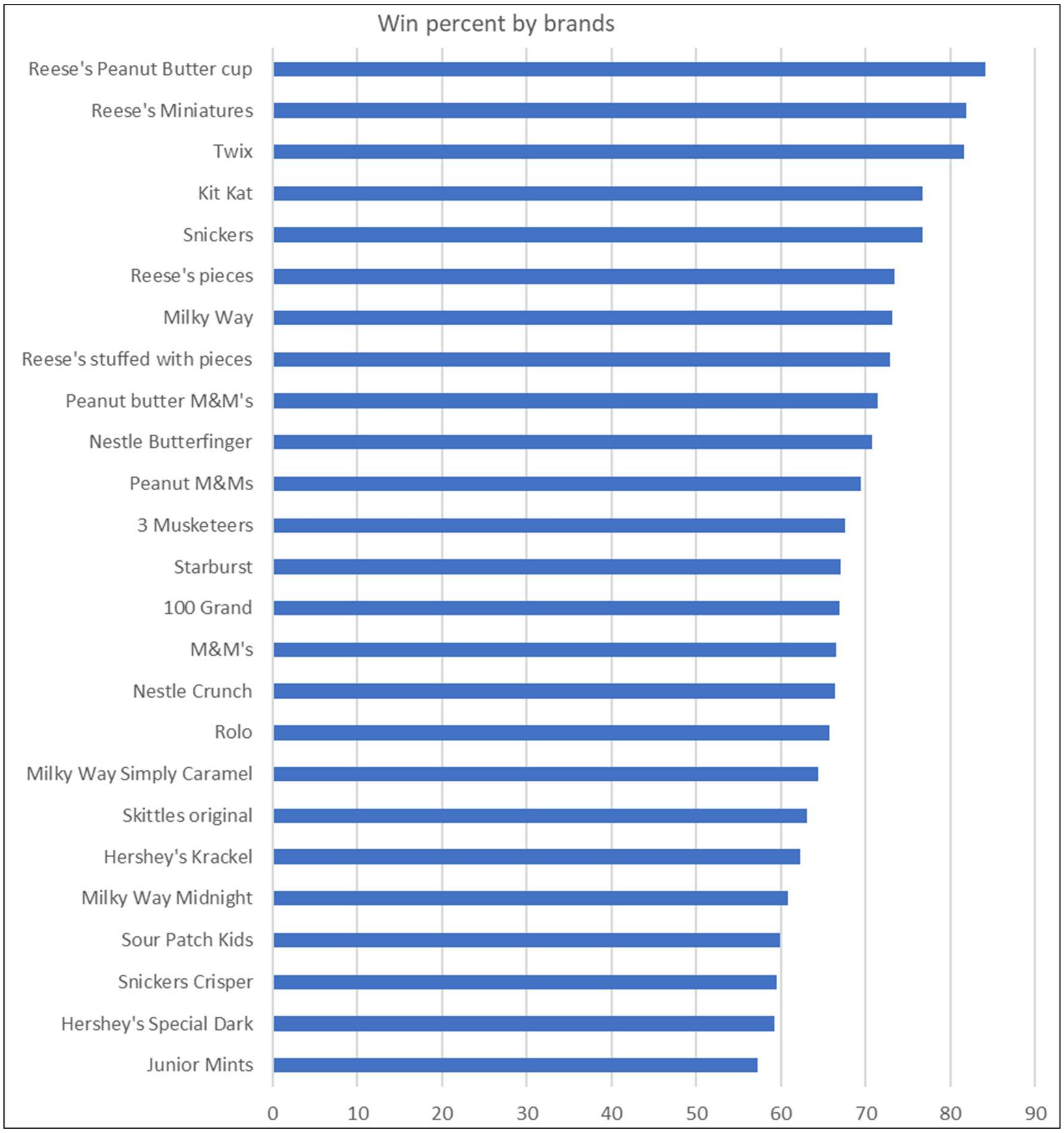
The report would have 4 sections: Data, Method, Analysis and Results.

Tools used: Python Programming and Microsoft Excel.

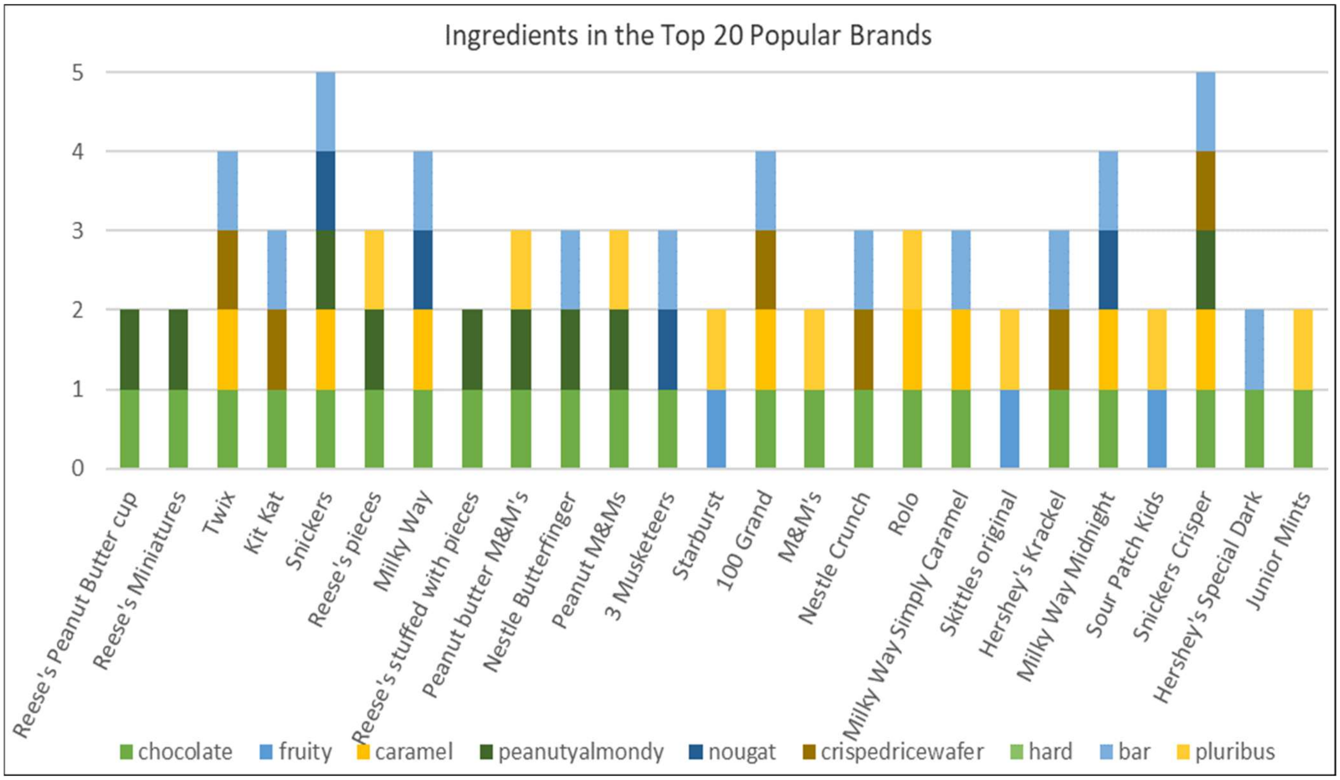
Data Section:

The dataset indicates the presence or absence of ingredients through binary values. The indicators for amount of sugar and pricing is indicated as percentiles of the entire dataset. The winning probability is depicted in percentage. The data is mostly clean but needed special characters to be replaced with punctuation.

From initial analysis in MS Excel, it is evident that certain brands enjoy more popularity over other brands. The chart below shows the Top 25 popular sweet treats:



But their popularity can be explained by combination of ingredients used in the sweet treats. Certain pairs/combinations of ingredients have a stronger correlation than others. The graph below shows the combination of ingredients in the Top 25 popular brands:



From initial stage, it can be observed that chocolate, peanut almond and caramel flavors in a bar form seem to be common occurrences.

Method of Approach:

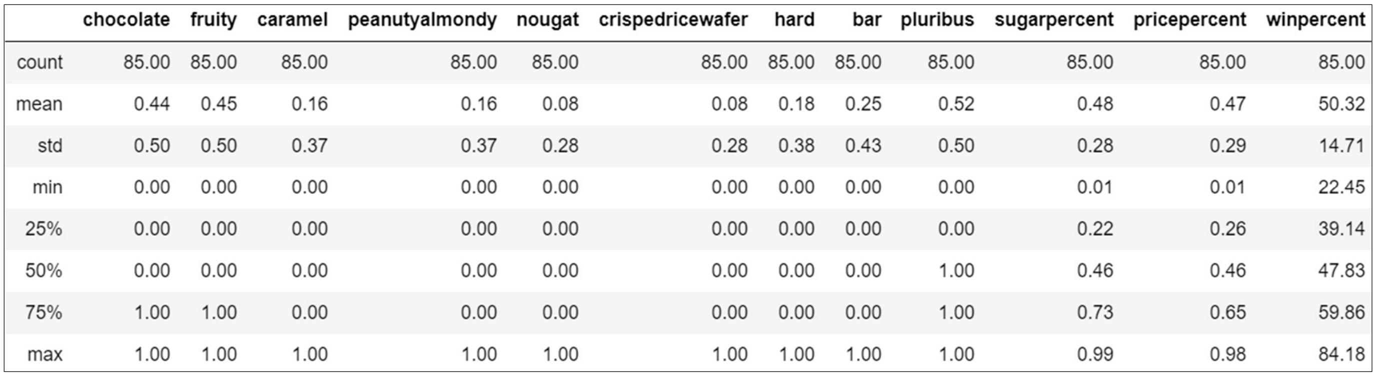
The dataset was further analyzed by utilizing Python programming. Following libraries were used.

File operations: Pandas

Visualization: matplotlib, seaborn, altair, vega plot

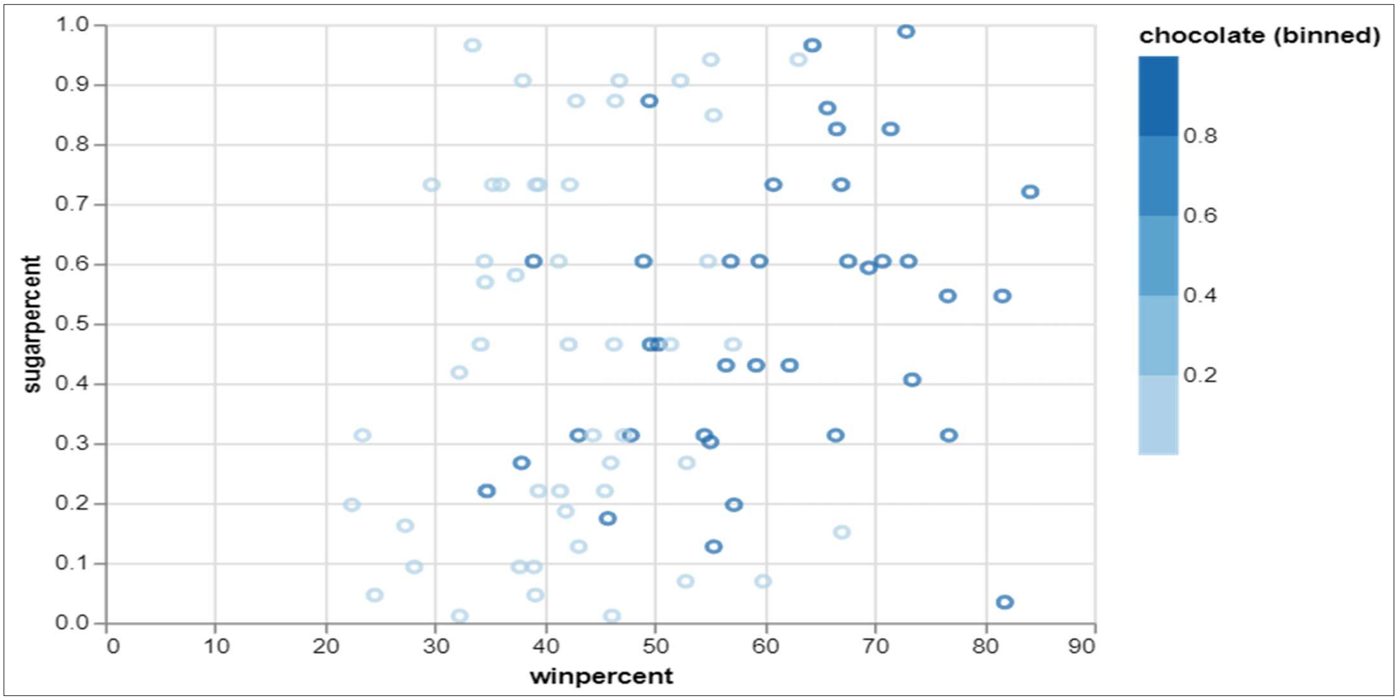
Statistical operations: Seaborn and Scikit Learn

The initial description table is as below:



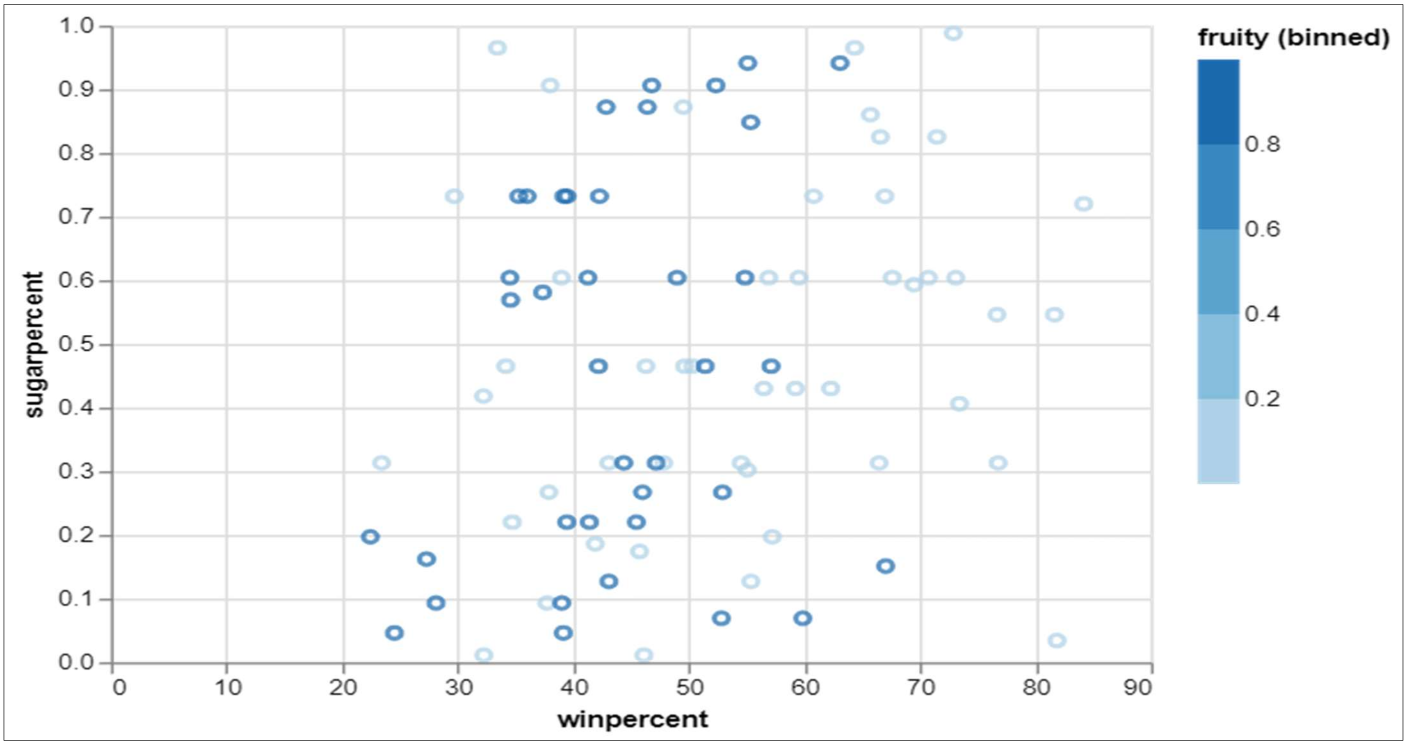
	chocolate	fruity	caramel	peanutyalmondy	nougat	crispedricewafer	hard	bar	pluribus	sugarpercent	pricepercent	winpercent
count	85.00	85.00	85.00	85.00	85.00	85.00	85.00	85.00	85.00	85.00	85.00	85.00
mean	0.44	0.45	0.16	0.16	0.08	0.08	0.18	0.25	0.52	0.48	0.47	50.32
std	0.50	0.50	0.37	0.37	0.28	0.28	0.38	0.43	0.50	0.28	0.29	14.71
min	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.01	22.45
25%	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.22	0.26	39.14
50%	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.46	0.46	47.83
75%	1.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.73	0.65	59.86
max	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.99	0.98	84.18

In order to verify the above, a scatterplot was drawn which compares sugar percent with win percent by using chocolate.



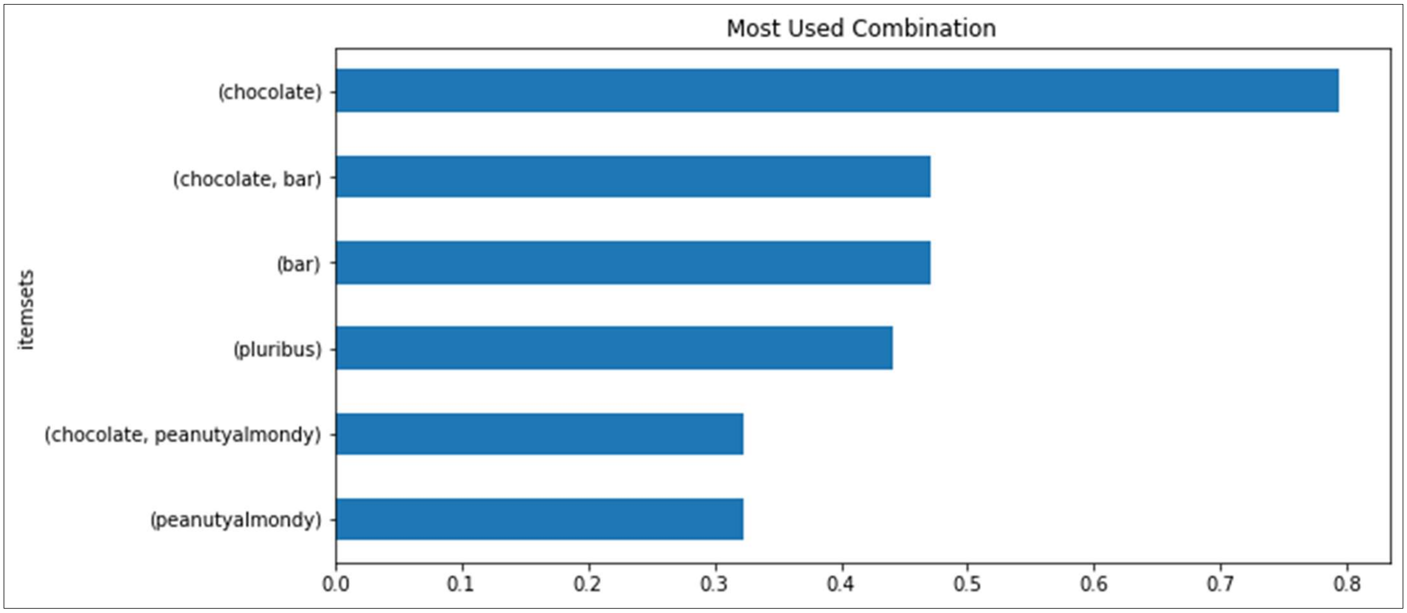
The plot gives us an idea that using chocolate with medium to high amount of sugar would give us a good win percentage.

By using fruit as a factor, the win percent is not too high. The scatterplot below confirms this:



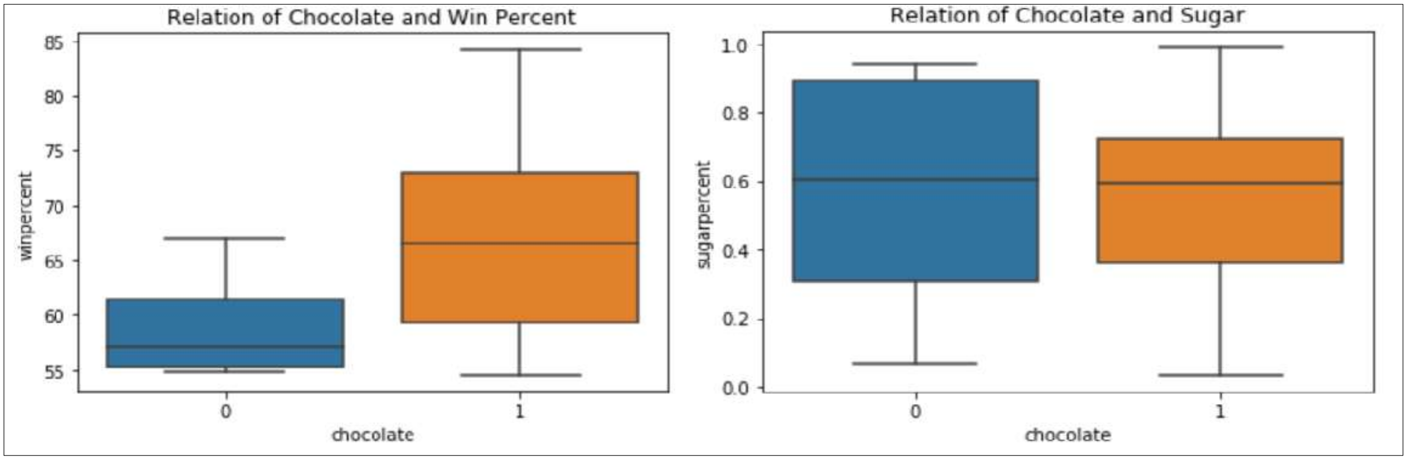
The above scatterplots simply indicate certain trends but do not give a very clear view of what really pushes up the win percentage. So, further analysis was done by including more than one ingredient.

The below bar graph shows us the actual picture of what combinations work the best.



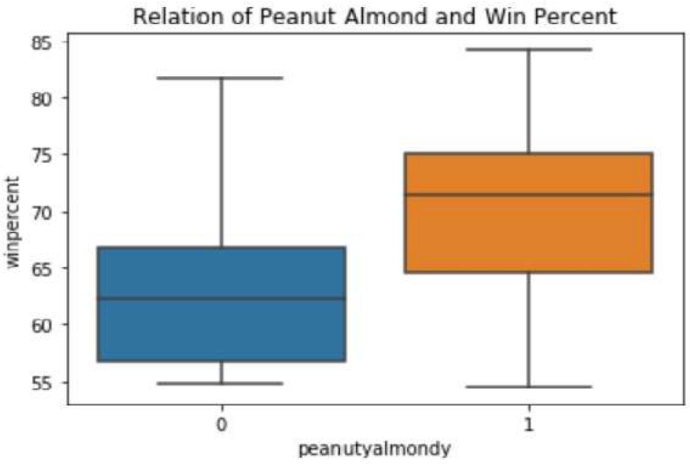
As can be seen, chocolate is the best ingredient to fetch maximum returns. If it is sold as a bar then the chances increase even further. Also, if they can be sold in minibar shape as pluribus, then it may fetch higher returns too. But in this plot, sugar and price percent have not been added yet. We get to it next.

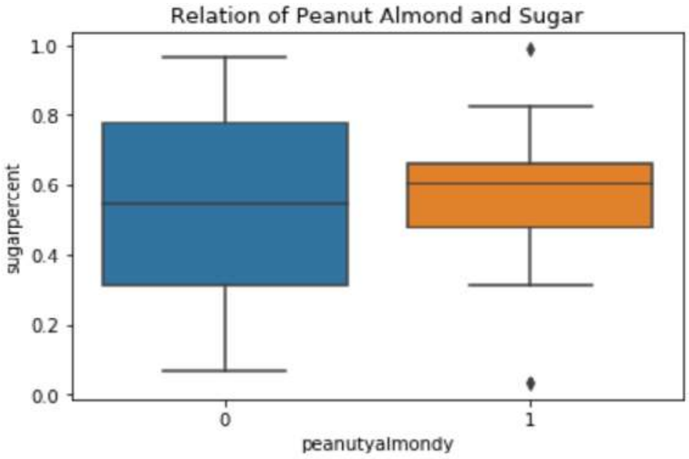
The below box plot merely confirm that chocolate is the best item to have in the candy offering.



The plots suggest that more than 65% people would prefer chocolate with less sugar in it.

The box plot below states that more than 70% people would like to have peanut almond in their chocolate with less sugar in it. The outliers are due to certain brands having peanut almond but with extreme sides on sugar scale, either too much, or too less.





Analysis:

So, finally, scikit learn package was utilized to create a linear regression model to determine and predict the win percent using various ingredients with different levels of sugar and price points.

The pandas dataframe was setup to have elements for X and Y to create a model object for the regression. Then the binary values as given in the dataset were taken into consideration for calculating the win percentage. The model run returned the below results with consisting of regression coefficients in a table for each item/ingredient:

```
Win percent with chocolate and peanutalmondy in a nougat bar: [70.49667027]
win percent with chocolate, caramel, peanut almond with crisp rice wafer in nougat
bar: [78.7220202]
win percent with fruit, peanut almond in a nougat bar: [60.17092536]
Win percent with fruit, peanut almond in a hard bar: [54.40636262]
```

The regression coefficient table also shows the effect of each ongredient on the overall composition of the sweet candy.

Ingredient Reg Coeff	Reg Coeff	Proportionality	Effect
chocolate	19.748067	Direct	High
fruity	9.42232207	Direct	Medium
caramel	2.22448136	Direct	Low
peanutyalmondy	10.0706885	Direct	High
nougat	0.8043306	Direct	Low
crispedricewafer	8.91896981	Direct	Medium
hard	-6.1653265	Inverse	Medium
bar	0.44154009	Direct	Low
pluribus	-0.8544995	Inverse	Low
sugarpercent	9.08676286	Direct	Medium
pricepercent	-5.9283614	Inverse	Medium

Results:

As can be seen from above code results, a combination of chocolate, caramel, peanut almond with crisp rice wafer in a medium sized nougat bar with moderate amount of sugar, with price being above 60th percentile, would fetch the most returns. The regression coefficient results also confirm that chocolate has the strongest attraction as an ingredient whereas.

The reason for this conclusion is, realistically the ingredients listed above are certainly not cheap, hence, the price can not be kept too low. But because people all over are more health conscious, the amount of sugar can be reduced to increase the returns, although the win percent may dip slightly. But, even with just 2 ingredients(chocolate and peanut almond), the win percentage could be decent enough to get good margins on the sale of the sweet treat.

Code Link: https://github.com/chinmoysarangi/Masters/blob/master/CandyProject.ipynb

References:

https://pbpython.com/pdvega.html
https://altair-viz.github.io/user_guide/troubleshooting.html#notebook-vega-lite-3-object
https://www.youtube.com/watch?v=J_LnPL3Qg70&list=PLeo1K3hjS3uvCeTYTeyfe0-rN5r8zn9rw&index=3
https://estongsy.wordpress.com/2019/03/07/valueerror-unknown-label-type-continuous/
https://laptrinhx.com/intro-to-pdvega-plotting-for-pandas-using-vega-lite-1039741255/
https://www.kaggle.com/grosvenpaul/what-do-people-look-for-in-a-candy
https://www.kaggle.com/pratik2901/halloween-candy-power-ranking