**Dataset:**

This dataset contains attributes of candy along with its ranking with binary variables, 1 means yes, 0 means no. This dataset has all sorts of different data about the kind of candies. All the data in attributes are distributed in binary categorical variables.

* 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.

[Candy - Data.csv](Candy%20-%20Data.csv)

**Research question:**

What is the best Halloween Candy? This was the question this dataset was collected to answer. To predict which candy is best at the time of Halloween and also helps the candy factory to increase the production and improve the quality and maintain the same obsession of that candy and sustain in the top of the charts and gain benefits from the sales. We can also see if that best candy is more chocolaty, more caramel, what will be difference in their costs compared to other candies, and how many customers prefer one over the other candy.

**Structure models:**

I am planning to do my analysis on **Linear Regression** model, we can estimate the better candy by plotting all the data and detecting the mean, and median values of that data. And defining a line in that plot by using least square distance between the data and the line drawn by using the mean or median values. **Logistic Regression** model, with this we can predict the true/false, 0/1, yes/no values. This predicts the future events that might occur by fitting the data into logit function. With this model we can answer the probabilities of choosing that candy over others by investigating the factors.

**Logistic regression model** predicts whether something is true or false. This logistic regression fits an ‘S’ shaped “logistic function”. This curve goes from 0 to 1, this curve also means the probability of occurrence in future. This curve is drawn based on maximum likelihood of occurrence or non-occurrence of prediction based on data. This can work with continuous data (dependent variable) and discrete data (independent variable) which is helps the prediction to meet the accuracy. The ability of using continuous and discrete measurements makes this regression as the popular one in machine learning methods.

* Generating inputs using .CSV file.
* Importing the libraries.
* Applying regression.
* Validating the model.

Let us understand Why, What and How in **Linear Regression** **model**. Firstly, Why Linear Regression? This regression model draws the relationship between the dependent and independent variables, where dependent variable is otherwise known as predictor variable and independent variable is known as response variable. Now let us investigate What is Linear Regression? This is a statistical analysis that attempts to show the relationship between two variables. How Linear Regression works? With the data provided by the dataset, we draw a plot between independent and dependent variable and spot a mean value for those values. It selects the best fit line that passes through the mean of the data. But as we know there will be multiple lines that passes through that mean point. To solve this scenario, we move the line until we have least square distance from all the data points. To analyze this technique there few steps, that need to be followed:

* Generating inputs using .CSV file.
* Importing the libraries.
* Applying regression
* Validating the model.

The main difference in linear and logistic regressions is how the line is fit in data.

**Complication:**

For this dataset there is no noisy data that need to be prepared and processed. But drawing relationship between the attributes is a very interesting task to do, to obtain accurate predictions. **Evaluation of dataset in R:**

#importing dataset into R  
candy<-read.csv("Candy - Data.csv", header = TRUE)  
#display the top rows in the data  
head(candy)

## competitorname chocolate fruity caramel peanutyalmondy nougat  
## 1 100 Grand 1 0 1 0 0  
## 2 3 Musketeers 1 0 0 0 1  
## 3 One dime 0 0 0 0 0  
## 4 One quarter 0 0 0 0 0  
## 5 Air Heads 0 1 0 0 0  
## 6 Almond Joy 1 0 0 1 0  
## crispedricewafer hard bar pluribus sugarpercent pricepercent winpercent  
## 1 1 0 1 0 0.732 0.860 66.97173  
## 2 0 0 1 0 0.604 0.511 67.60294  
## 3 0 0 0 0 0.011 0.116 32.26109  
## 4 0 0 0 0 0.011 0.511 46.11650  
## 5 0 0 0 0 0.906 0.511 52.34146  
## 6 0 0 1 0 0.465 0.767 50.34755

#to know if there is any noisy data in the data set.  
summary(candy)

## competitorname chocolate fruity caramel   
## 100 Grand : 1 Min. :0.0000 Min. :0.0000 Min. :0.0000   
## 3 Musketeers : 1 1st Qu.:0.0000 1st Qu.:0.0000 1st Qu.:0.0000   
## Air Heads : 1 Median :0.0000 Median :0.0000 Median :0.0000   
## Almond Joy : 1 Mean :0.4353 Mean :0.4471 Mean :0.1647   
## Baby Ruth : 1 3rd Qu.:1.0000 3rd Qu.:1.0000 3rd Qu.:0.0000   
## Boston Baked Beans: 1 Max. :1.0000 Max. :1.0000 Max. :1.0000   
## (Other) :79   
## peanutyalmondy nougat crispedricewafer hard   
## Min. :0.0000 Min. :0.00000 Min. :0.00000 Min. :0.0000   
## 1st Qu.:0.0000 1st Qu.:0.00000 1st Qu.:0.00000 1st Qu.:0.0000   
## Median :0.0000 Median :0.00000 Median :0.00000 Median :0.0000   
## Mean :0.1647 Mean :0.08235 Mean :0.08235 Mean :0.1765   
## 3rd Qu.:0.0000 3rd Qu.:0.00000 3rd Qu.:0.00000 3rd Qu.:0.0000   
## Max. :1.0000 Max. :1.00000 Max. :1.00000 Max. :1.0000   
##   
## bar pluribus sugarpercent pricepercent   
## Min. :0.0000 Min. :0.0000 Min. :0.0110 Min. :0.0110   
## 1st Qu.:0.0000 1st Qu.:0.0000 1st Qu.:0.2200 1st Qu.:0.2550   
## Median :0.0000 Median :1.0000 Median :0.4650 Median :0.4650   
## Mean :0.2471 Mean :0.5176 Mean :0.4786 Mean :0.4689   
## 3rd Qu.:0.0000 3rd Qu.:1.0000 3rd Qu.:0.7320 3rd Qu.:0.6510   
## Max. :1.0000 Max. :1.0000 Max. :0.9880 Max. :0.9760   
##   
## winpercent   
## Min. :22.45   
## 1st Qu.:39.14   
## Median :47.83   
## Mean :50.32   
## 3rd Qu.:59.86   
## Max. :84.18   
##

#attribute and its data type.  
str(candy)

## 'data.frame': 85 obs. of 13 variables:  
## $ competitorname : Factor w/ 85 levels "100 Grand","3 Musketeers",..: 1 2 45 46 3 4 5 6 7 8 ...  
## $ chocolate : int 1 1 0 0 0 1 1 0 0 0 ...  
## $ fruity : int 0 0 0 0 1 0 0 0 0 1 ...  
## $ caramel : int 1 0 0 0 0 0 1 0 0 1 ...  
## $ peanutyalmondy : int 0 0 0 0 0 1 1 1 0 0 ...  
## $ nougat : int 0 1 0 0 0 0 1 0 0 0 ...  
## $ crispedricewafer: int 1 0 0 0 0 0 0 0 0 0 ...  
## $ hard : int 0 0 0 0 0 0 0 0 0 0 ...  
## $ bar : int 1 1 0 0 0 1 1 0 0 0 ...  
## $ pluribus : int 0 0 0 0 0 0 0 1 1 0 ...  
## $ sugarpercent : num 0.732 0.604 0.011 0.011 0.906 ...  
## $ pricepercent : num 0.86 0.511 0.116 0.511 0.511 ...  
## $ winpercent : num 67 67.6 32.3 46.1 52.3 ...

References:

1. <https://www.youtube.com/watch?v=yIYKR4sgzI8>
2. <https://www.youtube.com/watch?v=nk2CQITm_eo>
3. I have taken this dataset from [https://www.kaggle.com/fivethirtyeight/the-ultimate halloween-candy-power-ranking/](https://www.kaggle.com/fivethirtyeight/the-ultimate%20halloween-candy-power-ranking/).