

# Project on “ Data Analysis with R ”

JIGNA THACKER

APRIL-2019 BATCH

# Agenda



## About the Project

- Dataset Information
- Data Collection Methodology
- Attribute Information
- Sample Details



## Exploratory Data Analysis (EDA)

- Scatter plots
- Identify Candy Features
- Ranking with Lollipop Chart



## Packages and Libraries

- Packages
- Libraries



## Closing Note

- Summary



## Data Details

- Sample
- Features
- %Points
- Data Reading

# About The Project



- **Dataset information** : “candy-data.csv” Information collected to get the most popular Halloween candy
- **Collection Methodology** : Online collection (<http://walthickey.com/2017/10/18/whats-the-best-halloween-candy> )
- **Attribute information** :
  - Various candy features: Chocolate , Fruity , Caramel , Peanutyalmondy, Nougat , Crispedricewafer , Hard , Bar and Pluribus
  - Various % points: Sugar % , Price % and Win %
- **Sample** : Overall 269,000 matchups were collected from 8,371 different IP addresses.
  - **“CANDY-DATA.CSV”** contains 85 records with 9 candy features and 3 % points

# Packages and Libraries of “R”

dplyr

ggplot2

ggalt

gridExtra

grid

png

# Data Details

---

- Sample: 85 Records
- Features : 9
- % Points : 3
- Data Reading





# Preliminary Observations

---

- Candy features are captured as binary variables with **1**- “Yes” and **0**- “No”

Feature	Description
Chocolate	Does it contain chocolate?
Fruity	Is it fruit flavoured?
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.

competitorname  
Length:85  
Class :character

# Data Reading (Data Summary)

	Min	1st Qu	Median	Mean	3rd Qu	Max
Features						
chocolate	0.0000	0.0000	0.0000	<b>0.4353</b>	1.0000	1.0000
fruity	0.0000	0.0000	0.0000	<b>0.4471</b>	1.0000	1.0000
caramel	0.0000	0.0000	0.0000	0.1647	0.0000	1.0000
peanutyalmondy	0.0000	0.0000	0.0000	0.1647	0.0000	1.0000
nougat	0.0000	0.0000	0.0000	0.0824	0.0000	1.0000
crispedricewafer	0.0000	0.0000	0.0000	0.0824	0.0000	1.0000
hard	0.0000	0.0000	0.0000	0.1765	0.0000	1.0000
bar	0.0000	0.0000	0.0000	0.2471	0.0000	1.0000
pluribus	0.0000	0.0000	1.0000	<b>0.5176</b>	1.0000	1.0000
% Points						
sugarpercent	0.011	0.22	0.465	0.4786	0.732	0.988
pricepercent	0.011	0.255	0.465	0.4689	0.651	0.976
winpercent	22.45	39.14	47.83	50.32	59.86	84.18

No Missing value found | Top 3 features with highest mean is highlighted in **bold**





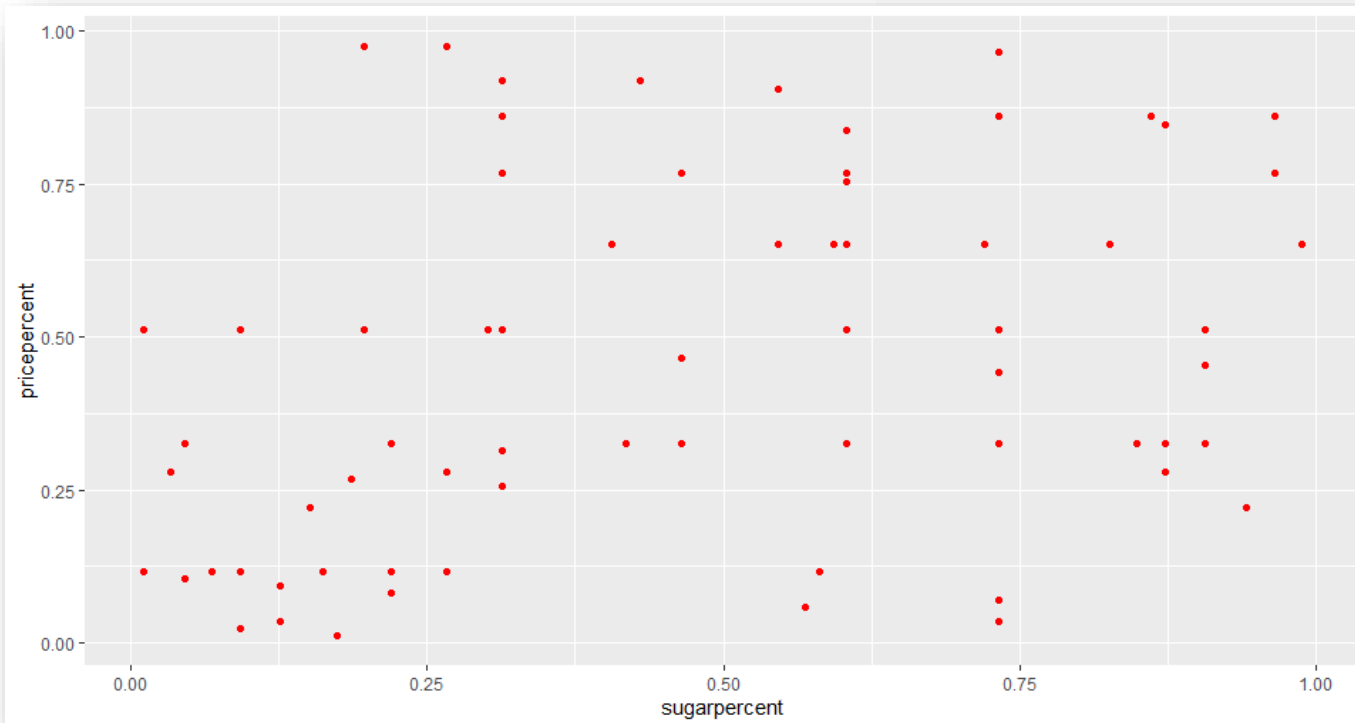
# Exploratory Data Analysis (EDA)

---

- Scatter Plots
- Identify Candy Features
- Ranking with Lollipop Chart



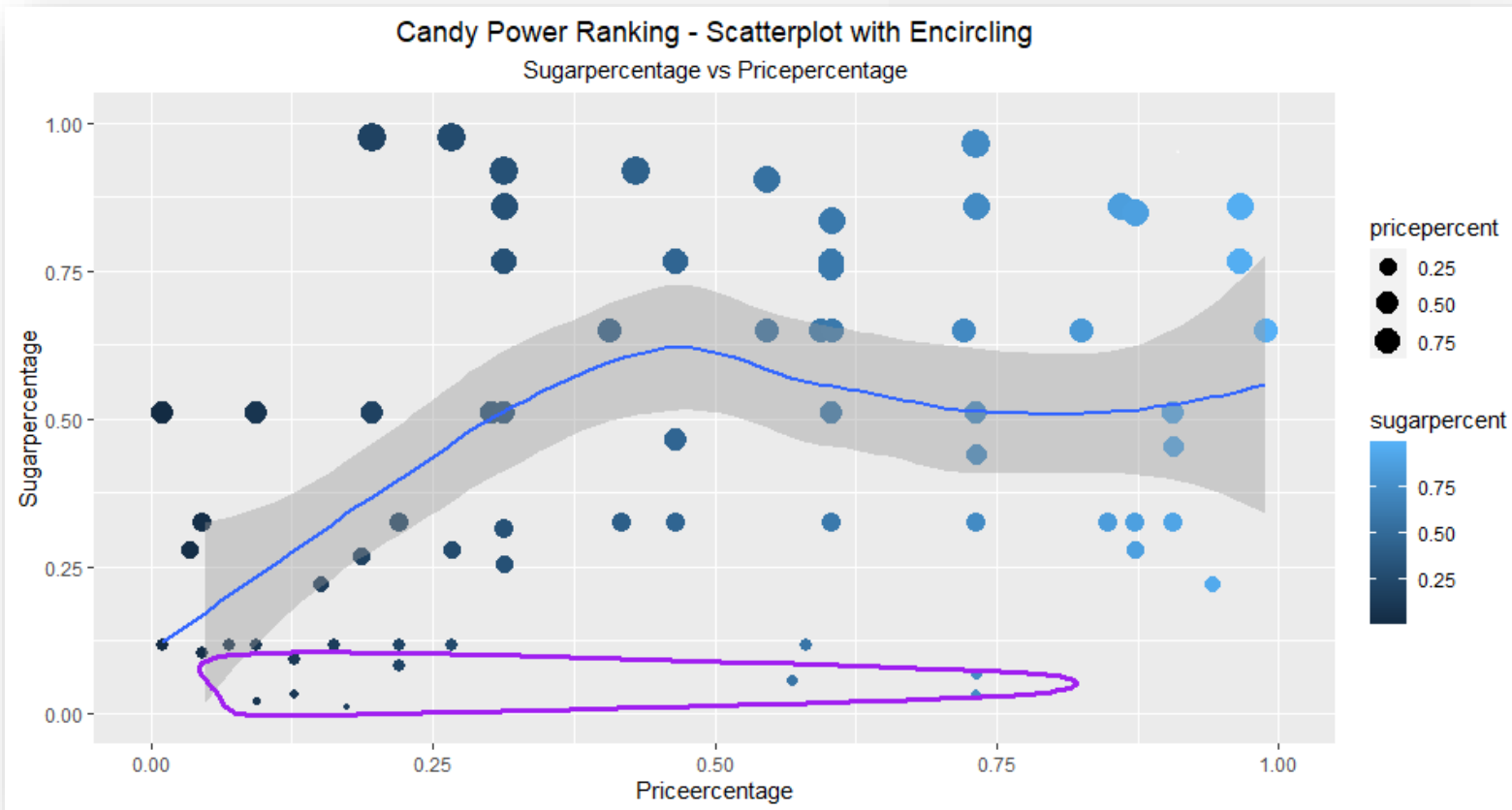
# Scatter Plot



**Sugar% vs Price%**

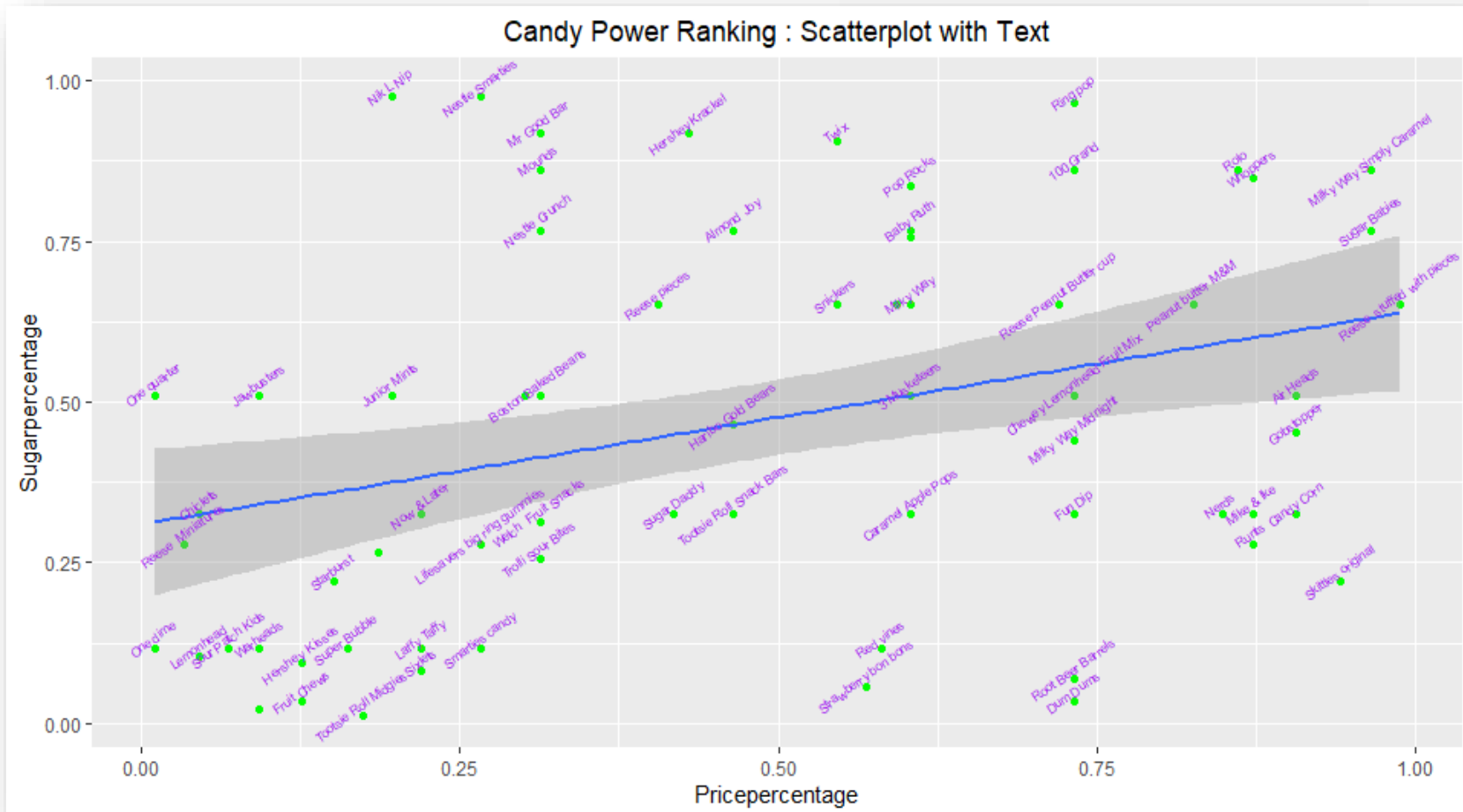
- This scatterplot is created to see how the amount of sugar has on the price of candy.
- Each point in the plot is determined by the value of the variable on the x-axis (sugar percentage) and on the y-axis (price percentage).

# Scatter Plot with Encircling



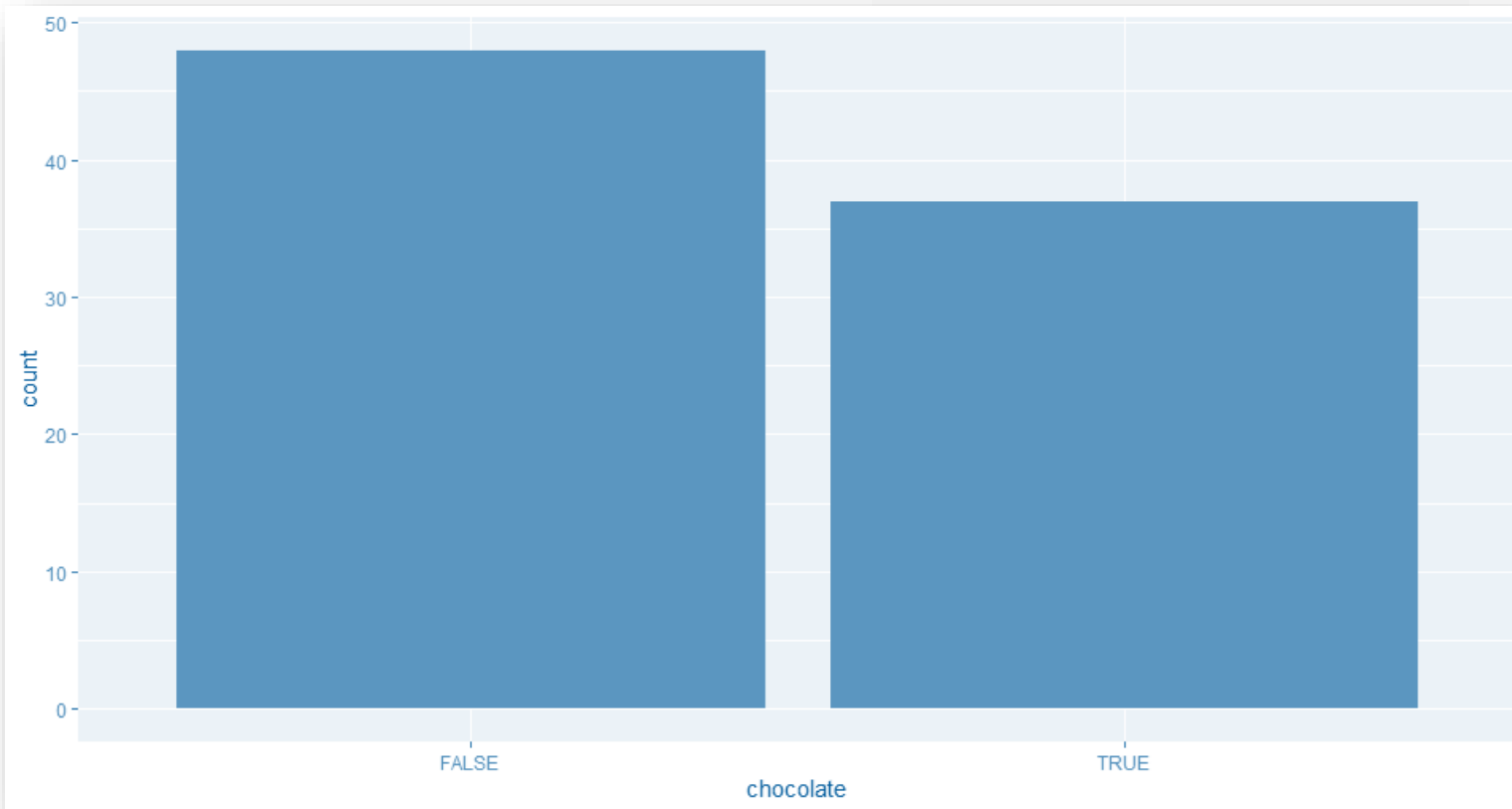
- This graph is created to encircle group of points in the scattered plot to draw attention with respect to Sugarpercentage and Pricepercentage.

100%



- In this graph adding the candy names to the plot will provide even more benefit to the analysis to see name of the candy wrt Sugarpercentage and Pricepercentage.

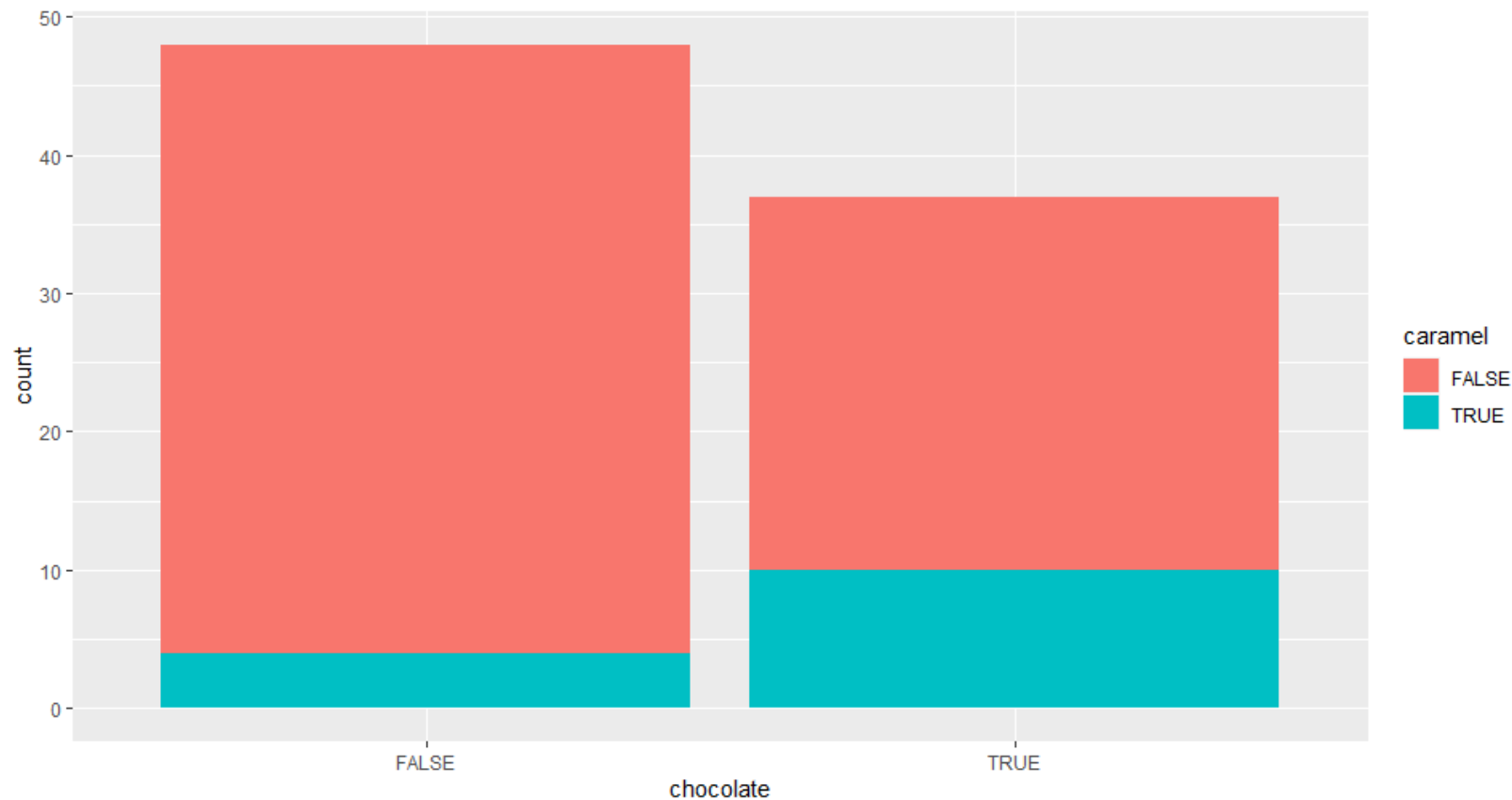
# Identify Candy Features



Chocolate Bar

- Here in this graph we can see
  - 37 brands has chocolate feature
  - 48 brands are non-chocolate feature
- Like chocolate , all feature has such distribution
- Few Candies has more than one feature. For eg: **Tootsie Pop** has chocolate , fruity and hard features.

# Identify Candy Features



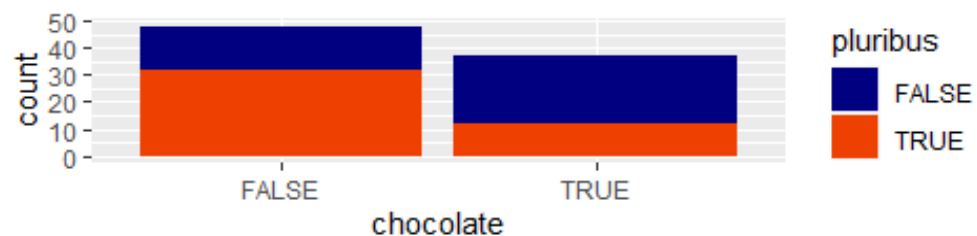
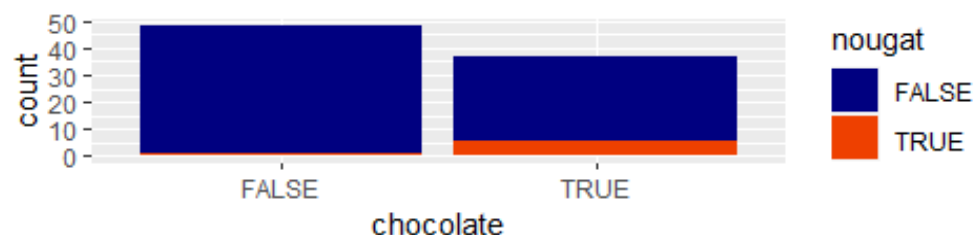
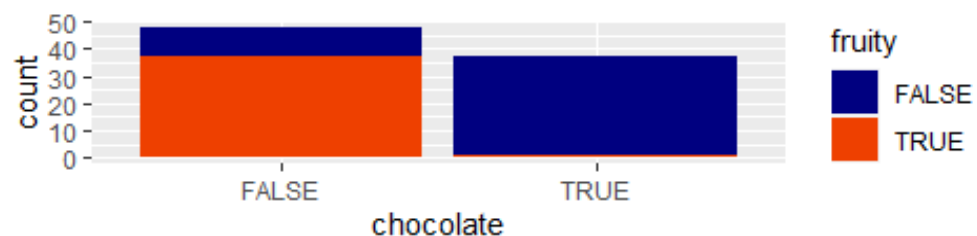
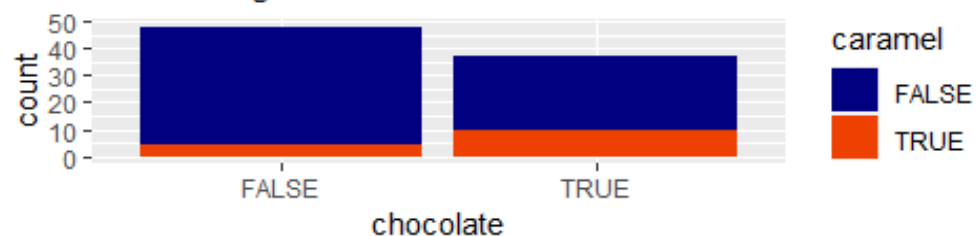
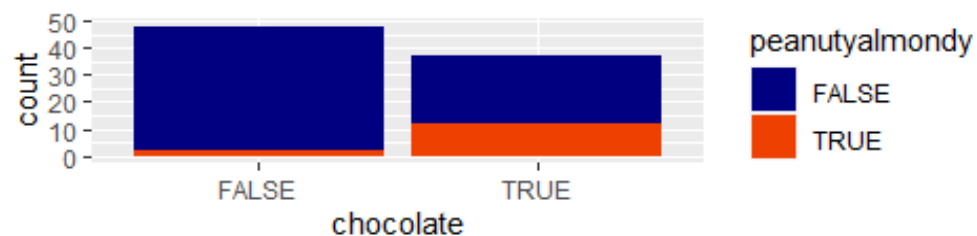
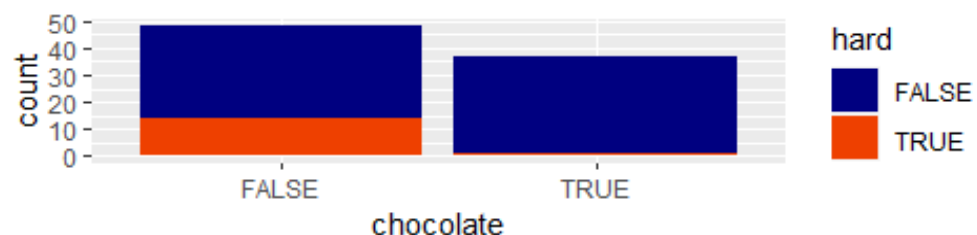
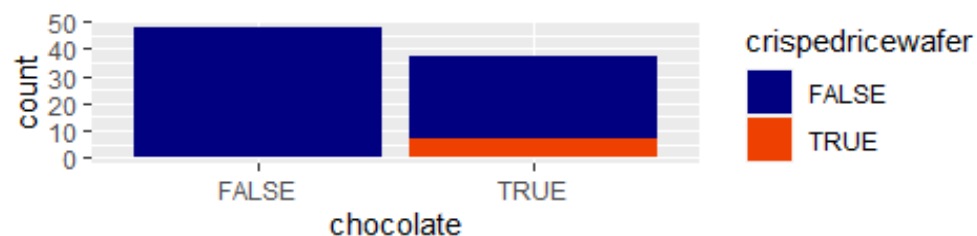
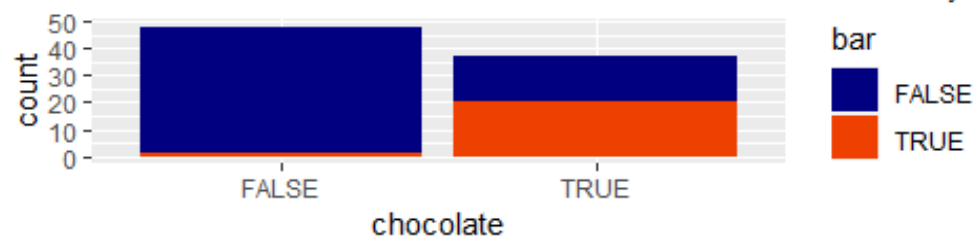
Chocolate and Caramel Bar

- Here in this graph we can see:
  - **10** brands has chocolate and Caramel combination
  - **27** brands has non-Caramel and chocolate feature
  - **4** brands has Caramel and non-chocolate feature
  - **44** brands has non-Caramel and non-chocolate feature

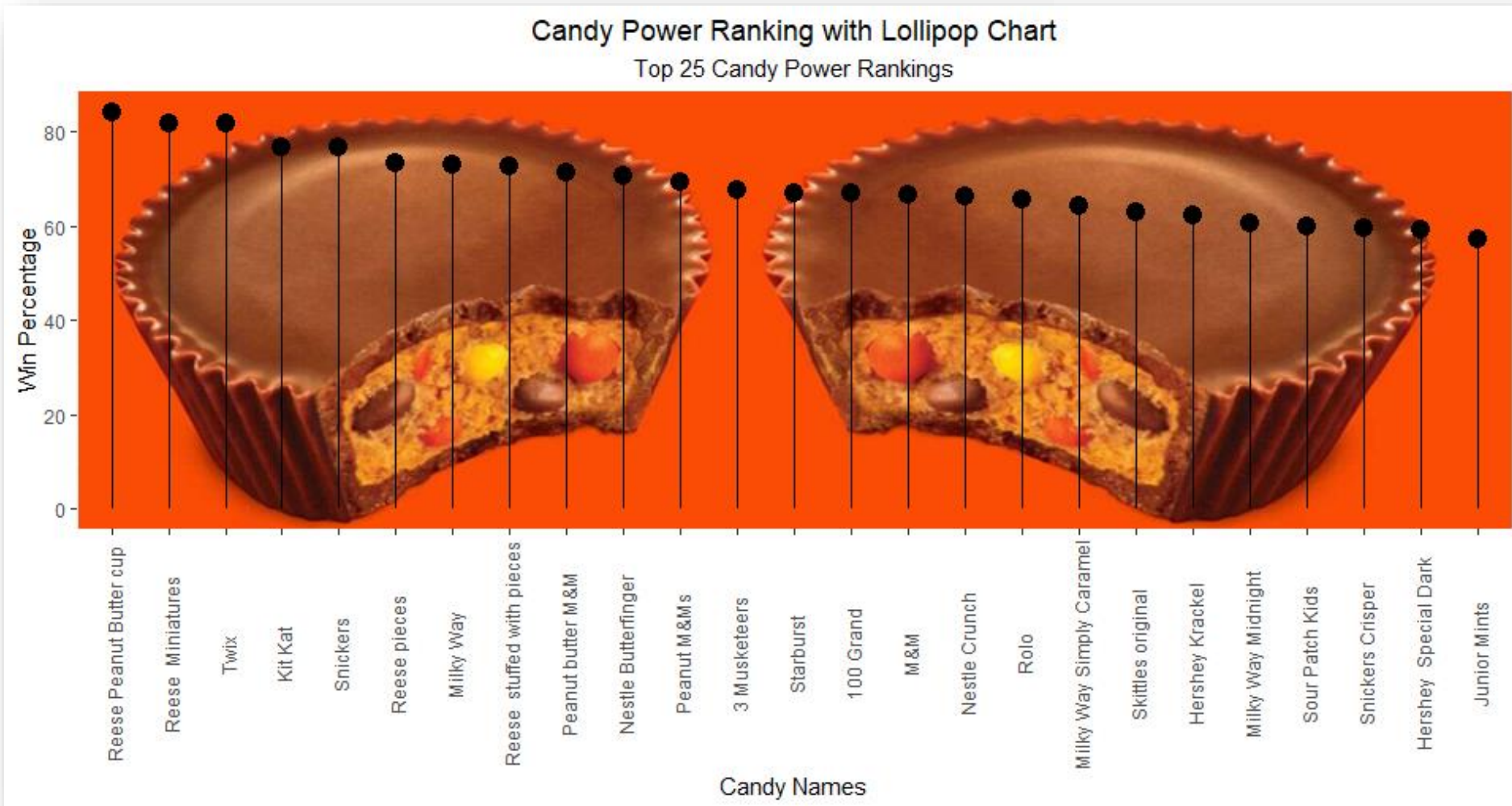


# Identify Candy Features: All Features

Chocolate Candy Features Grid Arrange



# Ranking with Lollipop Chart : Top 25 candy brands



# Summary



- Data Summary:
  - Top 3 features : Pluribus, Fruity and Chocolate has highest mean
- EDA:
  - Scatter plot is created with
    - Encircling and text format
- Identify Candy features:
  - Chocolate
  - Chocolate and Caramel
  - All features with Chocolate
- Lollipop chart with top 25 candy brands

# Thank you



For further information please contact:



Jigna Thacker



jignazt@yahoo.com



[https://github.com/jmps967/INSAID-DAR\\_Jigna-Thacker](https://github.com/jmps967/INSAID-DAR_Jigna-Thacker)