Halloween's Candy Ranking

22224546 SRICHAINAN JIRAWAN

22229519 CHAIORAWAN SIRIYAPORN

Introduction

WE WILL EXPLORE THE HALLOWEEN CANDY RANKINGS DATASET FROM 2017

THE ANALYSIS IS DIVIDED INTO **THREE** KEY PARTS:

1. **PROVING HYPOTHESES:** 3 HYPOTHESES
2. **DATA VISUALIZATION:** VISUAL INSIGHTS USING VARIOUS CHARTS TO ILLUSTRATE TRENDS AND PATTERNS
3. **STATISTICAL EVIDENCE:** USE T-STATISTICS AND P-VALUES TO VALIDATE OUR FINDINGS AND DEMONSTRATE STATISTICAL SIGNIFICANCE

Our hypothesis

REESE'S PEANUT BUTTERCUP IS THE MOST POPULAR CANDY DURING HALLOWEEN

CHOCOLATE IS THE MOST OFTEN USED INGREDIENT/FEATURE IN HALLOWEEN CANDY

CHOCOLATE AND CARAMEL IS THE COMBINATION OF INGREDIENT OR FEATURE WITH THE HIGHEST CORRELATION

Hypothesis 1

REESE'S IS THE MOST POPULAR CANDY DURING HALLOWEEN

```
[8]  # Hyp 1: Reese peanut buttercup is the most popular candy during Halloween

     # Check what is The most favorite Halloween's candy

     mydata = mydata[['competitorname','winpercent']]
     max_index = mydata['winpercent'].idxmax()
     most_favorite_candy = mydata.loc[max_index]

     most_favorite_candy
```

```
                          52
competitorname  ReeseÕs Peanut Butter cup
winpercent                    84.18029

dtype: object
```

OUTCOME

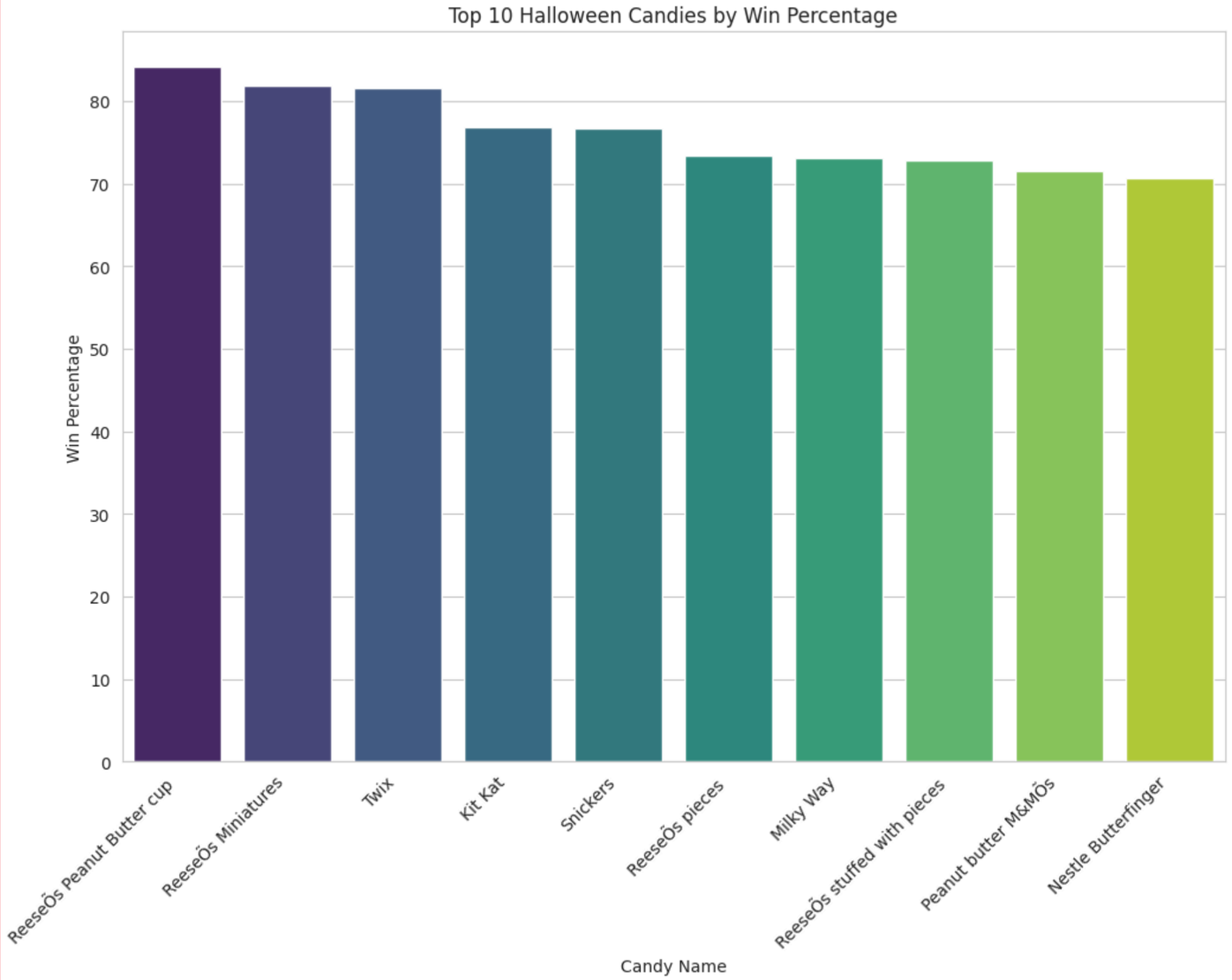
1ST STEP

WE CHECK WHICH CANDY IS THE MOST POPULAR BY USING IDXMAX FUNCTION TO PROVE OUR HYPOTHESIS

NOTED: ACCESS A SPECIFIC ROW (TO RETRIEVE THE ENTIRE ROW CORRESPONDING TO THAT INDEX)

THE TABLE PROVE THAT OUR HYPOTHESIS IS CORRECT WHICH WIN PERCENTAGE IS APPROXIMATELY 84%

2ND STEP DATA VISUALIZATION




```
import matplotlib.pyplot as plt
import seaborn as sns

# sort the data to see the top 10 candies
top_10 = mydata.sort_values(by='winpercent', ascending=False).head(10)

sns.set_style('whitegrid')

plt.figure(figsize=(12, 8))
sns.barplot(x='competitorname', y='winpercent', data= top_10, palette='viridis')

plt.xticks(rotation=45, ha='right')

plt.xlabel('Candy Name')
plt.ylabel('Win Percentage')
plt.title('Top 10 Halloween Candies by Win Percentage')

plt.show()
```

ANALYSIS

THE BAR CHART SHOWS THAT REESE PEANUT BUTTERCUP IS THE MOST POPULAR CANDY DURING HALLOWEEN BASED ON THE WIN PERCENTAGE, AS IT RANKS FIRST IN THE CHART

Hypothesis 1

3RD STEP

NULL HYPOTHESIS : THE WIN PERCENTAGE OF REESE'S PEANUT BUTTER CUP IS NOT DIFFERENT FROM THE WIN PERCENTAGE OF OTHER HALLOWEEN CANDIES.

ALTERNATIVE HYPOTHESIS: REESE PEANUT BUTTERCUP IS THE MOST POPULAR CANDY DURING HALLOWEEN

```
import pandas as pd
from scipy import stats

# Assuming 'mydata' is your DataFrame
reese_winpercent = mydata.loc[mydata['competitorname'] == 'ReeseÕs Peanut Butter cup', 'winpercent'].values[0]
other_candies_winpercent = mydata.loc[mydata['competitorname'] != 'ReeseÕs Peanut Butter cup', 'winpercent']
```

USE LOC FUNCTION TO EXTRACT ONLY THE WIN PERCENTAGE OF REESE'S PEANUT BUTTER CUP FROM THE DATAFRAME, ROW(0)

USE LOC FUNCTION TO EXTRACT THE WIN PERCENTAGE OF CANDIES OTHERS THAN REESE'S PEANUT BUTTER (!= –> NOT EQUAL) CUP FROM THE DATAFRAME

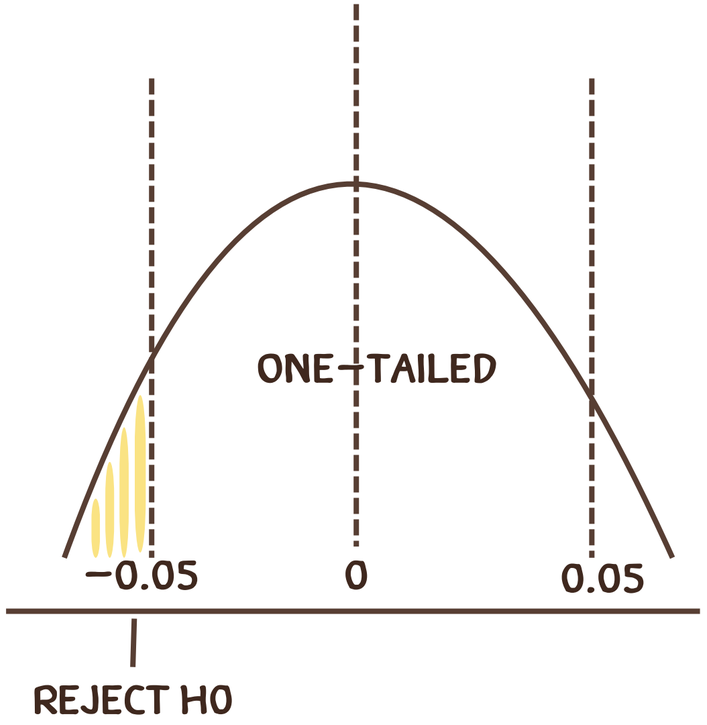
Hypothesis 1

CONDUCT T-STATS /P-VALUE

```
t_statistic, p_value = stats.ttest_1samp(other_candies_winpercent, reese_winpercent, alternative='less')

print(f"T-statistic: {t_statistic}")
print(f"P-value: {p_value}")

alpha = 0.05  # Significance level
if p_value < alpha:
    print("Reject the null hypothesis.")
    print("There is evidence to suggest that Reese's win percentage is higher than other candies.")
else:
    print("Fail to reject the null hypothesis.")
    print("There is not enough evidence to suggest that Reese's win percentage is higher than other candies.")
```

```
T-statistic: -21.58601292220484
P-value: 4.2236538607275814e-36
Reject the null hypothesis.
There is evidence to suggest that Reese's win percentage is higher than other candies.
```

THE FOCUS IS ON THE LEFT TAIL OF THE CURVE, REPRESENTING VALUES THAT ARE SIGNIFICANTLY LOWER THAN REESE'S WIN PERCENTAGE



THE OUTPUT SHOWS THAT IT REJECTS THE NULL HYPOTHESIS (P-VALUE=4.22E-36) WHICH MEANS THAT THE DATA SHOW ENOUGH A STRONG CORRELATION, CONCLUDING THAT REESE'S PEANUT BUTTER CUP IS STATISTICALLY MORE POPULAR THAN OTHER HALLOWEEN CANDIES.

Hypothesis 2

CHOCOLATE IS THE MOST OFTEN USED INGREDIENT/FEATURE IN HALLOWEEN CANDY.

Hypothesis 2

Chocolate is the most popular ingredient in Halloween's candy

```
mydata2 = mydata.drop(['competitorname','pluribus','pricepercent','winpercent','sugarpercent'], axis=1)
mydata_sum = mydata2.sum()

mydata_sum
```

	0
chocolate	37
fruity	38
caramel	14
peanutyalmondy	14
nougat	7
crispedricewafer	7
hard	15
bar	21

dtype: int64

1ST STEP

WE CHECK IF OUR HYPOTHESIS CORRECT OR NOT BY USING SUM FUNCTION

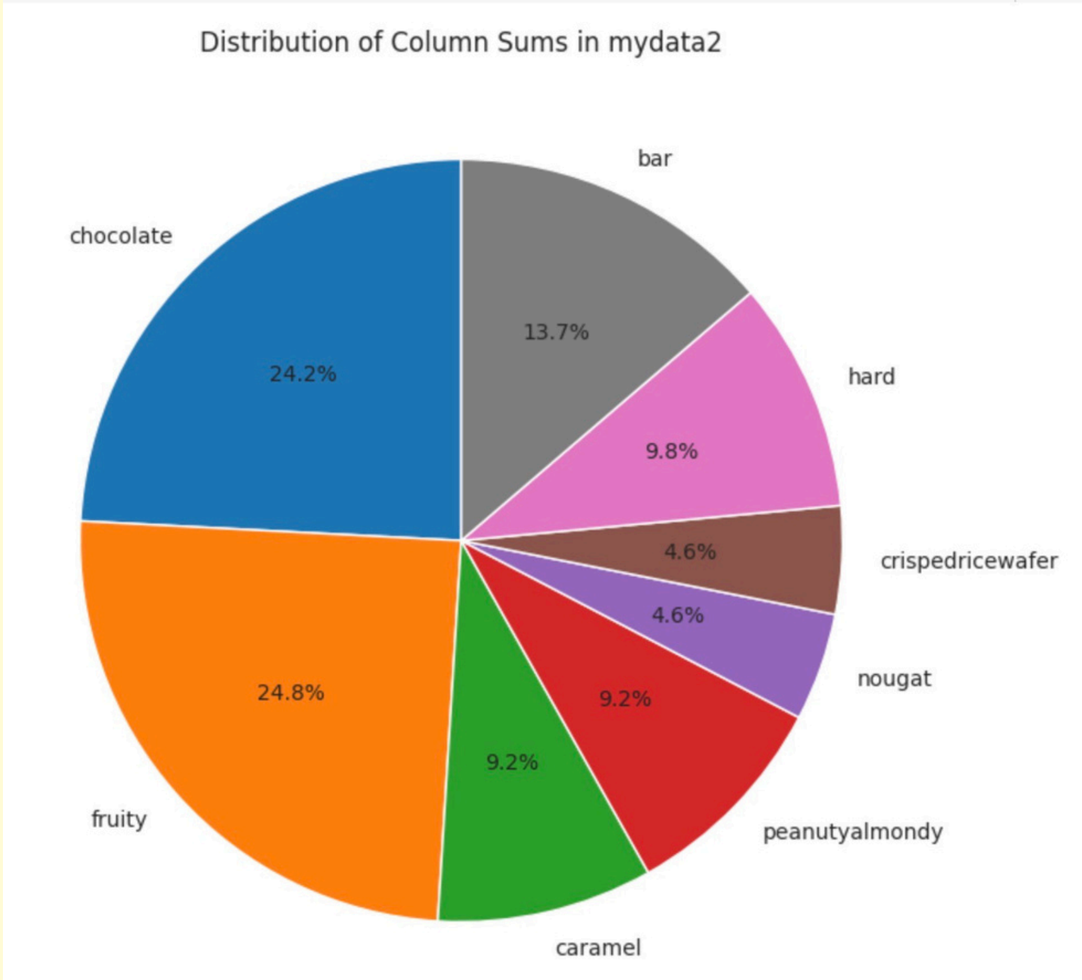
NOTED: USE DROP FUNCTION TO REMOVE AN IRRELEVANT COLUMN (AXIS=1)

THE TABLE SHOWS THAT THE MOST COMMON INGREDIENT FOUND IN HALLOWEEN CANDY WAS FRUITY, NOT CHOCOLATE

Hypothesis 2

```
import matplotlib.pyplot as plt

mydata_sum_df = pd.DataFrame(mydata_sum, columns=['Sum'])

# Create the pie chart
plt.figure(figsize=(8, 8))  # Adjust figure size if needed
plt.pie(mydata_sum.values, labels=mydata_sum.index, autopct='%1.1f%%', startangle=90)
plt.title('Distribution of Column Sums in mydata2')
plt.show()
```

2ND STEP DATA VISUALIZATION



ANALYSIS

FROM THE PIE CHART ABOVE, IT SHOWS THAT THE MOST POPULAR ATTRIBUTE IN HALLOWEEN'S CANDY IS NOT CHOCOLATE (24.2%. FRUITY HAS THE HIGHEST PERCENTAGE OF POPULARITY (24.8%), FOLLOWED BY CHOCOLATE IN SECOND PLACE AND BARS IN THIRD PLACE (13.7%).

Hypothesis 2

3RD STEP

NULL HYPOTHESIS : THE FREQUENCY OF CHOCOLATE AS A FEATURE IN HALLOWEEN CANDIES IS NOT DIFFERENT THAN THE FREQUENCY OF OTHER INGREDIENTS

ALTERNATIVE HYPOTHESIS: CHOCOLATE IS THE MOST OFTEN USED INGREDIENT/FEATURE IN HALLOWEEN CANDY.

index	competitorname	chocolate	fruity	caramel	peanutyalmondy	nougat	crispedricewafer	hard	bar	pluribus
52	ReeseÕs Peanut Butter cup	1	0	0	1	0	0	0	0	0
51	ReeseÕs Miniatures	1	0	0	1	0	0	0	0	0
79	Twix	1	0	1	0	0	1	0	1	0
28	Kit Kat	1	0	0	0	0	1	0	1	0
64	Snickers	1	0	1	1	1	0	0	1	0
53	ReeseÕs pieces	1	0	0	1	0	0	0	0	1
36	Milky Way	1	0	1	0	1	0	0	1	0
54	ReeseÕs stuffed with pieces	1	0	0	1	0	0	0	0	0
32	Peanut butter M&MÕs	1	0	0	1	0	0	0	0	1
42	Nestle Butterfinger	1	0	0	1	0	0	0	1	0
47	Peanut M&Ms	1	0	0	1	0	0	0	0	1
1	3 Musketeers	1	0	0	0	1	0	0	1	0
68	Starburst	0	1	0	0	0	0	0	0	1
0	100 Grand	1	0	1	0	0	1	0	1	0
33	M&MÕs	1	0	0	0	0	0	0	0	1
43	Nestle Crunch	1	0	0	0	0	1	0	1	0
56	Rolo	1	0	1	0	0	0	0	0	1

1 = YES (THERE A FOLLOWING INGREDIENT IN THE CANDY)

0= NO (THERE IS NO FOLLOWING INGREDIENT IN THE CANDY)

Hypothesis 2

```
#Drop chocolate column (Other candies)
hypo2 = mydata2.drop(['chocolate'], axis=1)
```

	fruity	caramel	peanutyalmondy	nougat	crispedricewafer	hard	bar
0	0	1	0	0	1	0	1
1	0	0	0	1	0	0	1
2	0	0	0	0	0	0	0
3	0	0	0	0	0	0	0
4	1	0	0	0	0	0	0
...	...	...	...	...	...	...	...
80	1	0	0	0	0	0	0
81	1	0	0	0	0	1	0
82	1	0	0	0	0	0	0
83	0	1	0	0	0	1	0
84	0	0	0	0	1	0	0

CREATE A DATASET (HYPO2) FOR CANDIES THAT DO NOT CONTAIN CHOCOLATE —> COMPARE THE PRESENCE OF CHOCOLATE (CHOCOLATE CANDIES) VS OTHER ATTRIBUTES (NON-CHOCOLATE CANDIES) TO BE USE IN A T-TEST

```
[101] hypo2.mean()
```

	0
fruity	0.447059
caramel	0.164706
peanutyalmondy	0.164706
nougat	0.082353
crispedricewafer	0.082353
hard	0.176471
bar	0.247059

dtype: float64

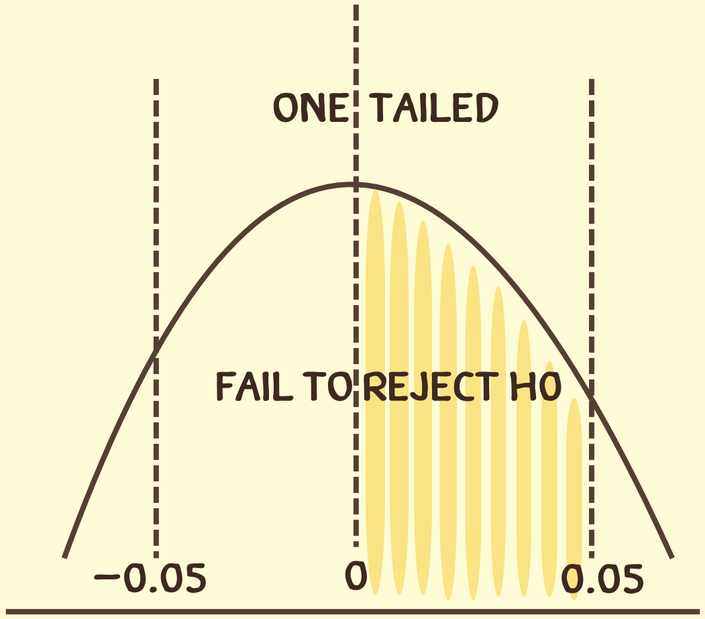
```
mydata['chocolate'].mean()
```

```
0.43529411764705883
```

ALSO, THE MEAN PROVIDES A CENTRAL TENDENCY OF THE DATA — > TO COMPARE THE AVERAGE VALUE OF CANDIES WITH CHOCOLATE AGAINST CANDIES WITHOUT CHOCOLATE

Hypothesis 2

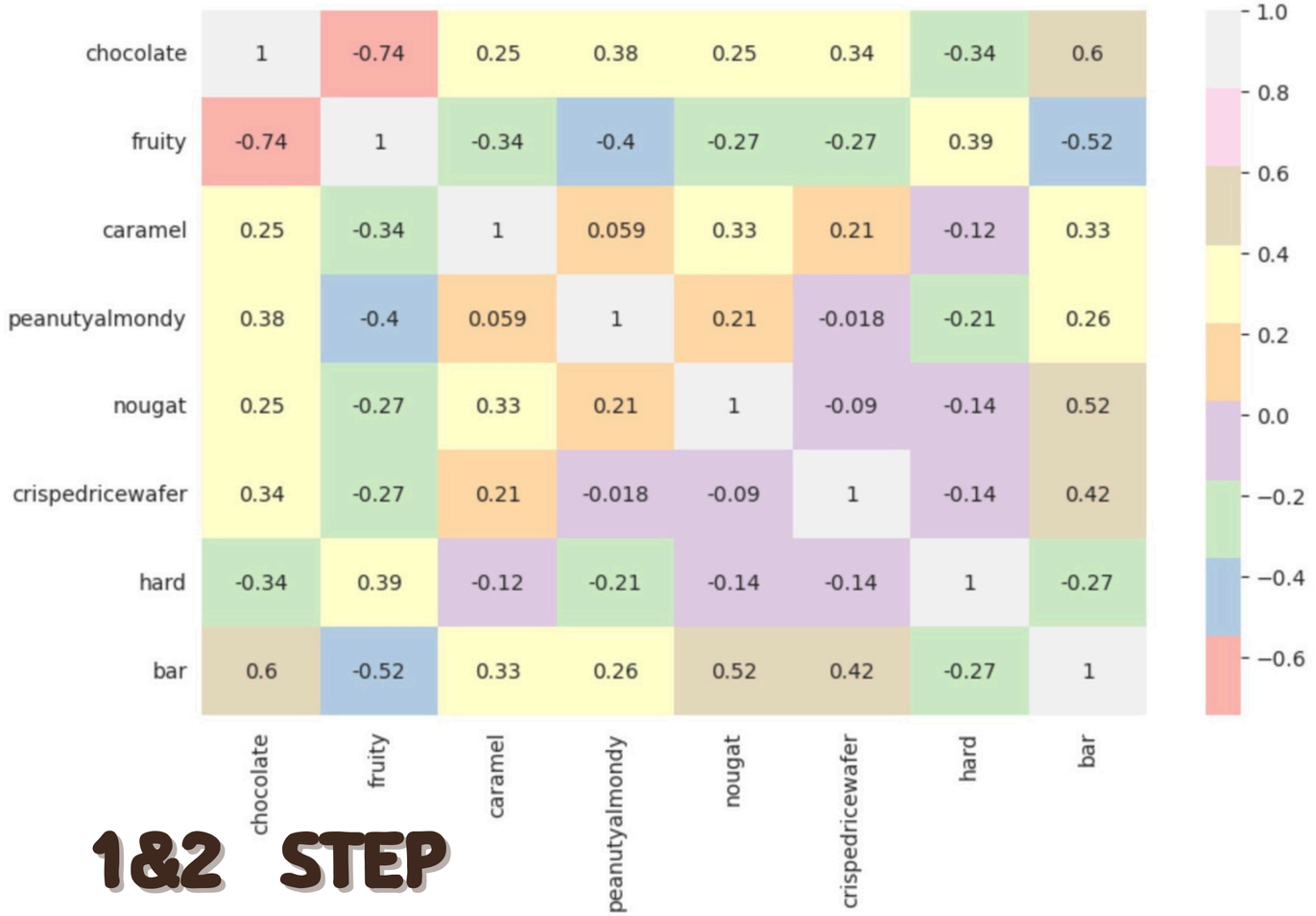
```
[153] from scipy import stats

      # Perform one-tailed t-test (alternative='greater')
      t_statistic, p_value = stats.ttest_ind(mydata['chocolate'].mean(), hypo2.mean(), alternative='greater')

      print(f"T-statistic: {t_statistic}")
      print(f"P-value: {p_value}")

      #Interpret the results
      alpha = 0.05 #Significant level
      if p_value < alpha:
        print("Reject the null hypothesis.")
        print("There is evidence to suggest that chocolate is the most popular attribute in Halloween's candy.")
      else:
        print("Fail to reject the null hypothesis.")
        print("There is not enough evidence to suggest that chocolate is the most popular attribute in Halloween's candy.")

      T-statistic: 1.7980192164898514
      P-value: 0.06114370372125492
      Fail to reject the null hypothesis.
      There is not enough evidence to suggest that chocolate is the most popular attribute in Halloween's candy.
```



THE OUTPUT SHOWS THAT IT FAILS TO REJECT THE NULL HYPOTHESIS (P-VALUE=0.06) WHICH MEANS THAT THE DATA DON'T SHOW A STRONG ENOUGH CORRELATION, INDICATING CHOCOLATE MAY NOT BE THE MOST INFLUENTIAL FACTOR IN DETERMINING A CANDY'S POPULARITY DURING HALLOWEEN.

Hypothesis 3

CHOCOLATE AND CARAMEL IS THE COMBINATION OF INGREDIENT OR FEATURE WITH THE HIGHEST CORRELATION



1&2 STEP

```
import matplotlib.pyplot as plt
import seaborn as sns

# Assuming 'competitorname' is the only non-numeric column you want to exclude
mydata3 = mydata2.select_dtypes(include=['number']) # Select only numeric columns

plt.figure(figsize=(10, 6))

sns.heatmap(mydata3.corr(), annot=True, cmap='Pastel1')  # Annotation (with numbers)
plt.show()
```

ANALYSIS

FROM THE HEAT MAP ABOVE, IT SHOWS THE CORRELATION BETWEEN EACH VARIABLE. THE COMBINATION OF ATTRIBUTE WITH THE MOST CORRELATION IS NOT CHOCOLATE AND CAREMEL (0.25), BUT CHOCOLATE AND BAR (0.6) BECAUSE THE CORRELATION COEFFICIENT IS CLOSER TO 1, WHICH MEANS THERE IS A STRONG POSITIVE CORRELATION BETWEEN THE CHOCOLATE AND BAR

Hypothesis 3

3RD STEP

NULL HYPOTHESIS: THERE IS NO SIGNIFICANT CORRELATION BETWEEN THE PRESENCE OF CHOCOLATE AND CARAMEL IN HALLOWEEN CANDIES

ALTERNATIVE HYPOTHESIS: CHOCOLATE AND CARAMEL IS THE COMBINATION OF INGREDIENT OR FEATURE WITH THE HIGHEST CORRELATION

```
[22] choco_caramel = mydata['chocolate'].corr(mydata['caramel'])
```

```
0.24987534731992408
```

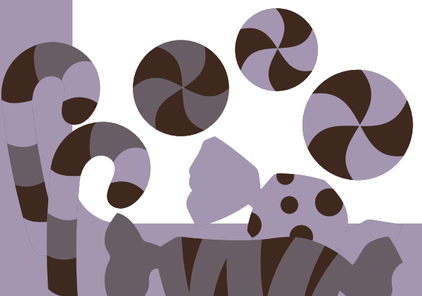
THE CORRELATION VALUE OF CHOCOLATE AND CARAMEL

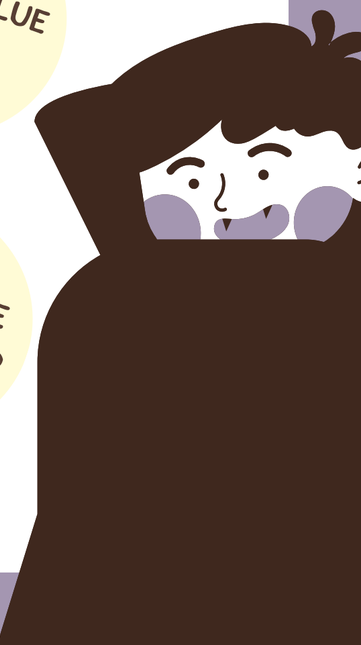
```
[23] drop3 = mydata2.drop(['caramel','chocolate'], axis=1)
     hypo3 = drop3.corrwith(mydata['chocolate'])
```

	0
fruity	-0.741721
peanutyalmondy	0.377824
nougat	0.254892
crispedricewafer	0.341210
hard	-0.344177
bar	0.597421

dtype: float64

THE CORRELATION VALUE OF CHOCOLATE AND OTHER FEATURES

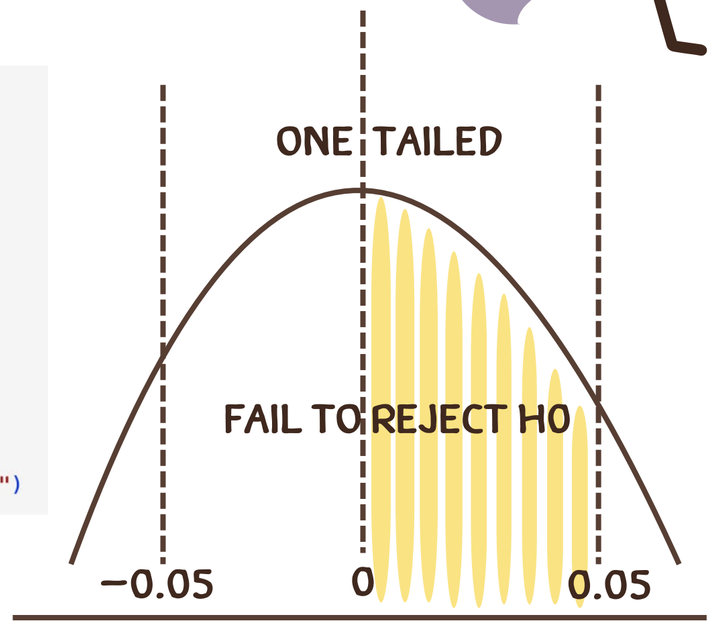




Hypothesis 3

```
[214] from scipy import stats

      # Perform one-tailed t-test (alternative='greater')
      t_statistic, p_value = stats.ttest_ind(choco_caramel, hypo3 , alternative='greater')

      print(f"T-statistic: {t_statistic}")
      print(f"P-value: {p_value}")

      #Interpret the results
      if p_value < alpha:
        print("Reject the null hypothesis.")
        print("There is evidence to suggest that chocolate and caramel is the most popular combination in Halloween's candy.")
      else:
        print("Fail to reject the null hypothesis.")
        print("There is not enough evidence to suggest that chocolate and caramel is the most popular combination in Halloween's candy.")
```

```
T-statistic: 0.30557253505433846
P-value: 0.38612137362771903
Fail to reject the null hypothesis.
There is not enough evidence to suggest that chocolate and caramel is the most popular combination in Halloween's candy.
```



THE OUTPUT SHOWS THAT IT FAILS TO REJECT THE NULL HYPOTHESIS (P-VALUE=0.38) WHICH MEANS THAT THE DATA DON'T SHOW A STRONG ENOUGH CORRELATION, INDICATING THE COMBINATION OF CHOCOLATE AND CARAMEL MAY NOT BE THE INGREDIENT OR FEATURE WITH THE HIGHEST CORRELATION

Conclusion

FROM OUR THREE HYPOTHESES, AFTER WE USE T-TEST TO PROVE, WE CAN CONCLUDE THAT:

FIRST HYPOTHESIS – REESE'S PEANUT BUTTERCUP IS THE MOST POPULAR CANDY DURING HALLOWEEN WAS PROVEN TO BE TRUE

SECOND HYPOTHESIS – CHOCOLATE IS THE MOST OFTEN USED INGREDIENT/FEATURE IN HALLOWEEN CANDY WAS PROVEN TO BE FALSE (FRUITY IS THE MOST USED)

THIRD HYPOTHESIS – CHOCOLATE AND CARAMEL IS THE COMBINATION OF INGREDIENT OR FEATURE WITH THE HIGHEST CORRELATION WAS PROVEN TO BE FALSE (CHOCOLATE AND BAR HAVE THE HIGHEST CORRELATION)

Citation

dataset

APA CITATION

THE ULTIMATE HALLOWEEN CANDY POWER RANKING. (2017, OCTOBER 31). KAGGLE. HTTPS://WWW.KAGGLE.COM/DATASETS/FIVETHIRTYEIGHT/THE-ULTIMATE-HALLOWEEN-CANDY-POWER-RANKING