PROJECT: HALLOWEEN CANDY POWER

MACHINE LEARNING –ALL MODULES

JIGNA THACKER

GCD - APRIL 2019 BATCH

PROJECT BRIEF

- **OBJECTIVE :** To identify the best (or most popular) Halloween candy attribute?
- **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

PRESENTATION FLOW

- Exploratory Data Analysis (EDA)
- Apply machine learning applications
  - Linear Regression
  - PCA
  - Clustering and K-Mean
- Comparison of various Machine Learning applications

EDA : Exploratory data analysis

PRELIMINARY OBSERVATIONS

- Binary candy features are captured as binary variables with

  1- “Yes” and 2- “No”

Feature	Description
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.

PROFILING USING PANDAS_PROFILING

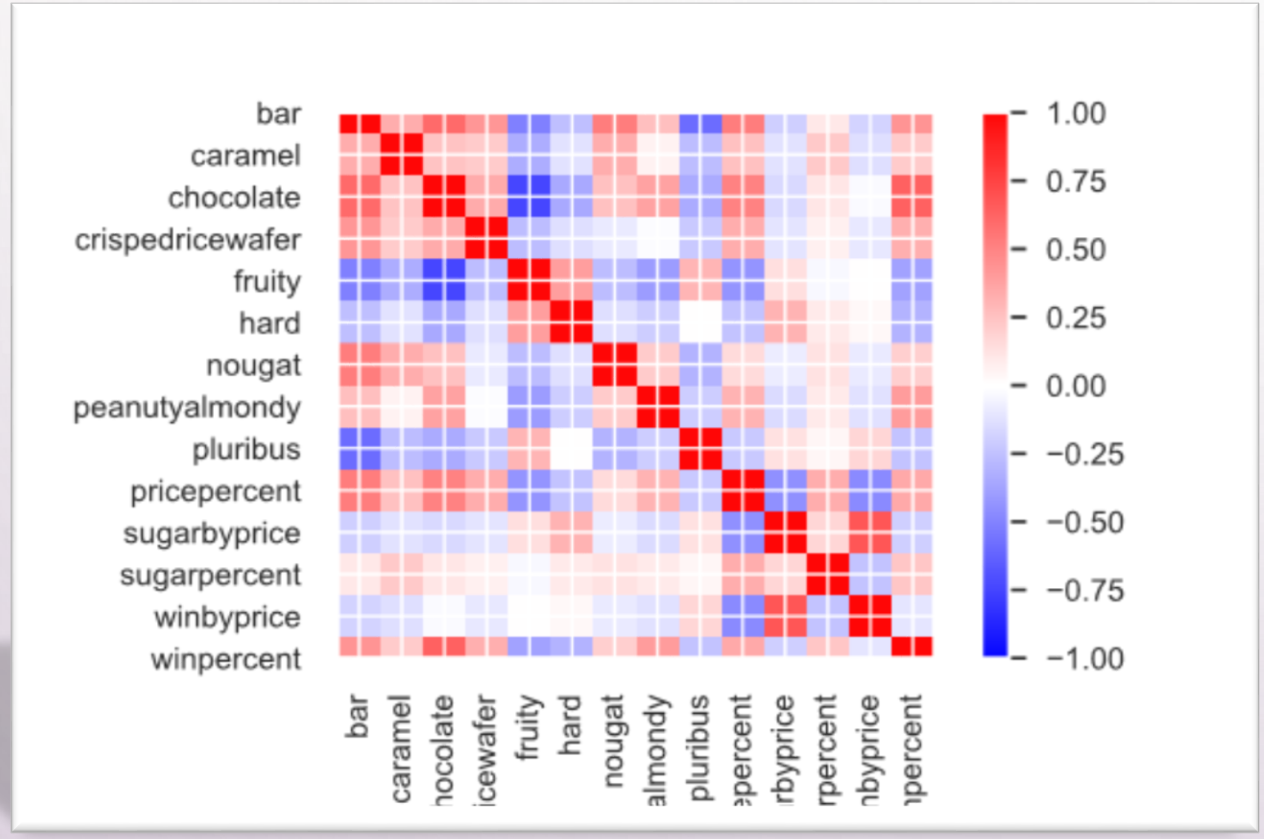
	Yes %	No %
Bar	24.7	75.3
Caramel	16.5	83.5
Chocolate 3	**43.5**	56.5
Crispedricewafer	8.2	91.8
Fruity 2	**44.7**	55.3
Hard	17.6	82.4
Nougat	8.2	91.8
Peanutyalmondy	16.5	83.5
Pluribus 1	**48.2**	51.8

Top three flavors are:

1. Pluribus
2. Fruity
3. Chocolate

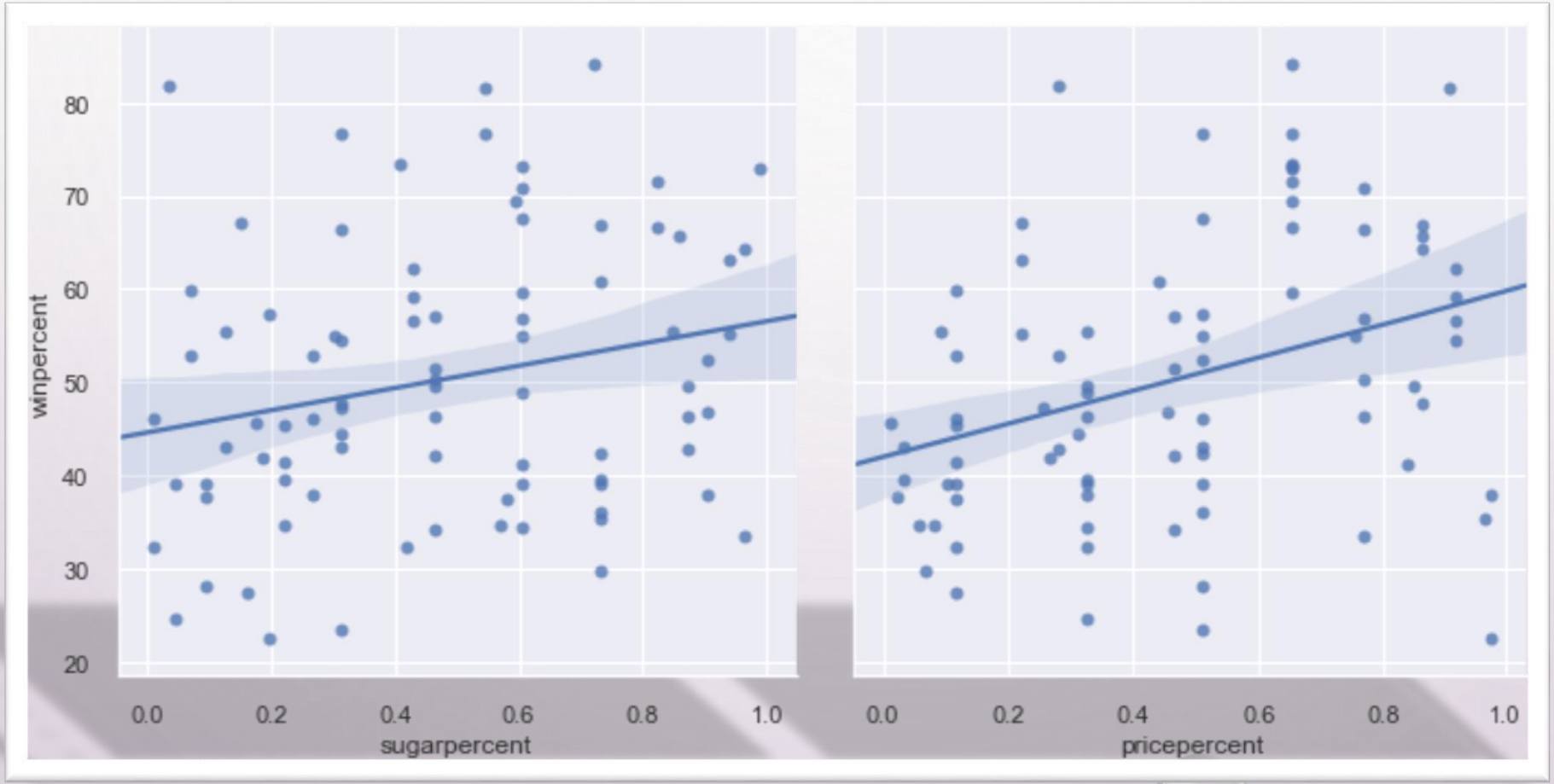
General assumption is Chocolate being top one – but here it is **NOT**

CORRELATION MATRIX
USING PANDAS_PROFILING



CORRELATION MATRIX

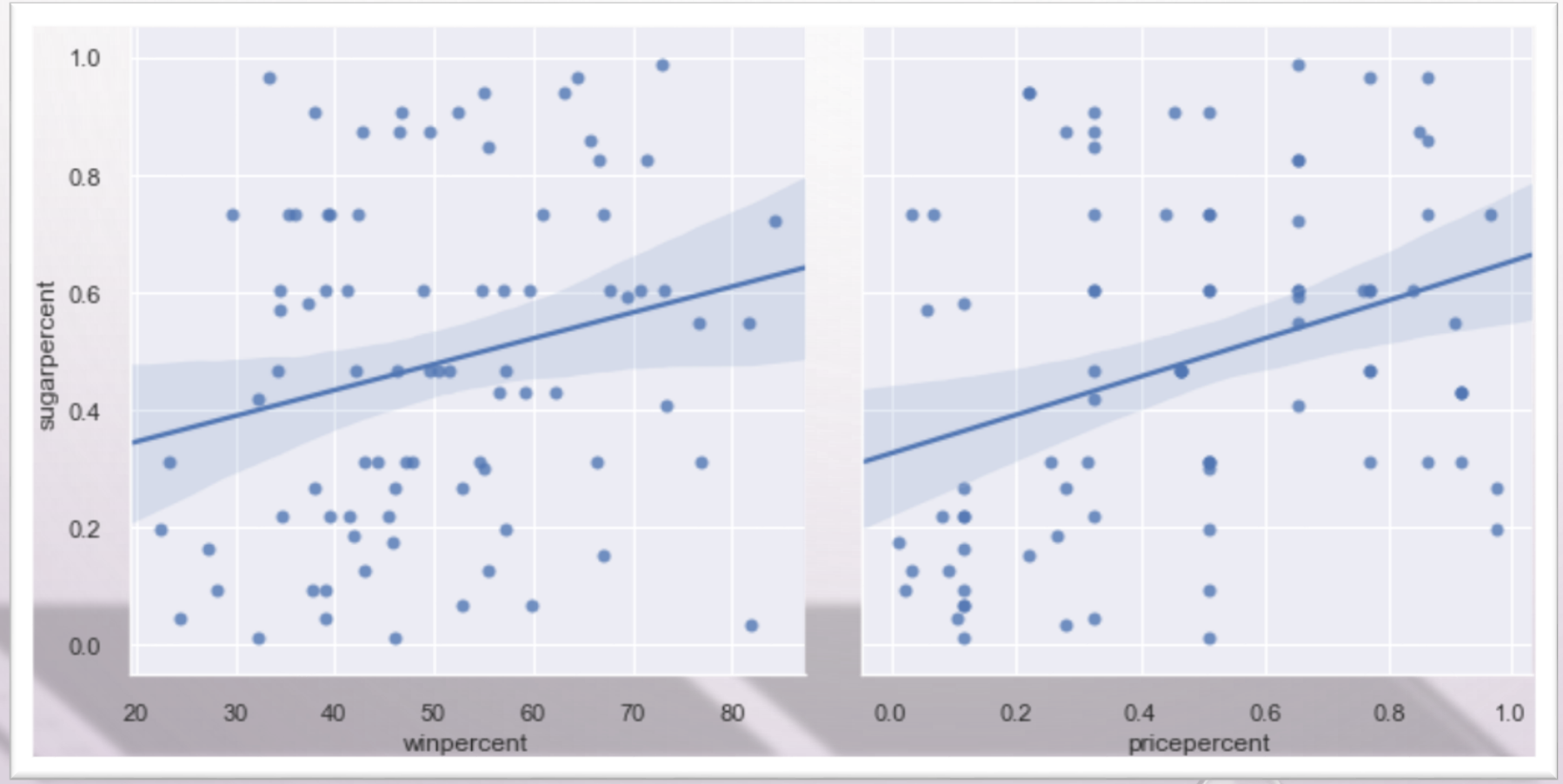
PAIRWISE: WIN% WITH SUGAR% AND PRICE%



No Strong linear relationship between Win% with Sugar% and Price%

CORRELATION MATRIX

PAIRWISE: SUGAR% WITH WIN% AND PRICE%



No Strong linear relationship between Sugar% with Win% and Price%

OVERALL WIN% BY CANDY

Top 10 brand basis win%

QUESTIONS AROUND :-

What made some candies more popular then others ??

Was it price?

Was it sugar contain?

Was it flavor?

Was it just winning (popularity) of the brand?

LET'S LOOK AT DATA WITH NEW ANGLE "PRICE"

- **SUGAR BY PRICE :** High value indicates candy is sweet and with low pricing
- **WIN BY PRICE :** High value indicates candy is liked more and with low pricing

FEW QUESTIONS AROUND THINS ANGLE:

- Which are top 10 brands by flavor
- Any non-chocolaty competitors are winner?
- Winner basis price
- What is the performance of Sugar candies
- Any candies having both chocolaty and fruity flavor?

LET'S TRY TO ANSWER

- **Q1: Which are top 10 brands**
  - REESEÕS PEANUT BUTTER CUP
  - REESEÕS MINIATURES
  - TWIX
  - KIT KAT
  - SNICKERS
  - REESEÕS PIECES
  - MILKY WAY
  - REESEÕS STUFFED WITH PIECES
  - PEANUT BUTTER M&MÕS
  - NESTLE BUTTERFINGER

With Chocolate being one of the ingredient

- **Q2: Any non-chocolaty competitors are winner?**
  - STARBURST
  - SKITTLES ORIGINAL
  - SOUR PATCH KIDS
  - HARIBO GOLD BEARS
  - NERDS
  - SKITTLES WILDBERRY
  - SWEDISH FISH
  - LIFESAVERS BIG RING GUMMIES
  - SOUR PATCH TRICKSTERS
  - AIR HEADS

Sour Patch Kids has highest win price "516.068948"

- **Q3: Winner basis price**
  - TOOTSIE ROLL MIDGIES
  - PIXIE STICKS
  - FRUIT CHEWS

Best Valued brand
Tootsie Roll Midgies has heist win price "4157.886182"

LET'S TRY TO ANSWER

- **Q4: What is the performance of Sugar candies**
  - REESE STUFFED WITH PIECES
  - MILKY WAY SIMPLY CARAMEL
  - SUGAR BABIES

Top 3 brands have lesser win price compare to other brands

- **Q2: Any candies having both chocolaty and fruity flavor?**
  - TOOTSIE POP

Having win price "150.715854"

OVERALL WIN BY PRICE IS…

Question		Top Brand	Winbyprice	Flavor
Which are top 10 brands	3	ReeseÕs Miniatures	293.427434	Chocolate + PeanutyAlmondy
Any non-chocolaty competitors are winner?	2	Sour Patch Kids	516.068948	Fruity + Pluribus
Winner basis price	1	Tootsie Roll Midgies	4157.886182	Chocolate + Pluribus
What is the performance of Sugar candies		Skittles original	286.750636	Fruity + Pluribus
Any candies having both chocolaty and fruity flavor?		Tootsie Pop	150.715854	Chocolate + Fruity

Top 3 Winbyprice shows popularity towards Pluribus, Chocolate & Fruity

EDA- FINDINGS

- Based on the EDA we can see below are the attributes preferred :
  - Chocolate
  - Fruity
  - Pluribus
- This is not giving us clear idea specific attribute
- Lets use machine learning application to deep dive on the objective

Apply machine learning applications

"To get the most important feature attribute?"

APPLY MACHINE LEARNING APPLICATIONS

APPLICATIONS USED:

- Linear Regression
- PCA
- Clustering and K-Mean

APPLICATION: 1 : LINEAR REGRESSION
(WITH 80:20 SPLIT)

- Performing 80:20 split on the data
- Result for coefficient by each flavor:

Chocolate :17.949723899964514	1	Crispedricewafer 14.279451610532584	2
Fruity 10.284564315506424		Hard -7.406472613965985	
Caramel 0.6481223157790419		Bar -5.1951597383610935	
Peanutyalmondy 10.09750660356215 4		Pluribus -2.392981549050286	
Nougat 13.185773462822908	3		

Linear Equation with 80:20 (Train and Test) data:

y = 37.94 + 17.95 * chocolate + 10.28 * fruity + 0.65 * caramel + 10.1 * peanutyalmondy + 13.18 * nougat + 14.27 * crispedricewafer - 7.41 * hard - 5.19 * bar - 2.39 * pluribus

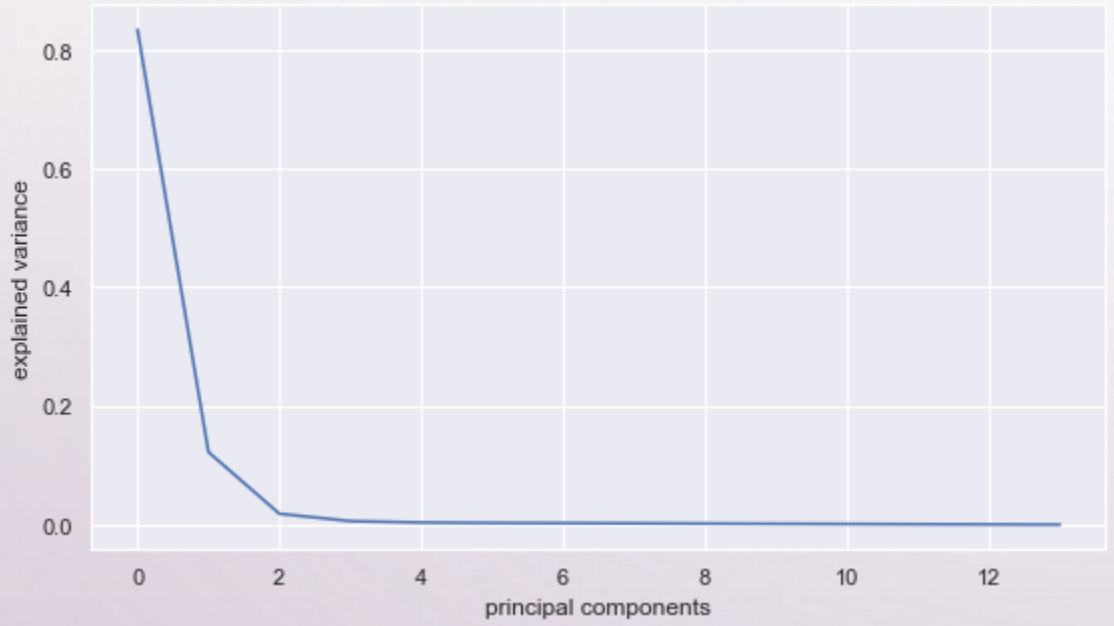
APPLICATION: 1 : LINEAR REGRESSION
(WITH 80:20 SPLIT)

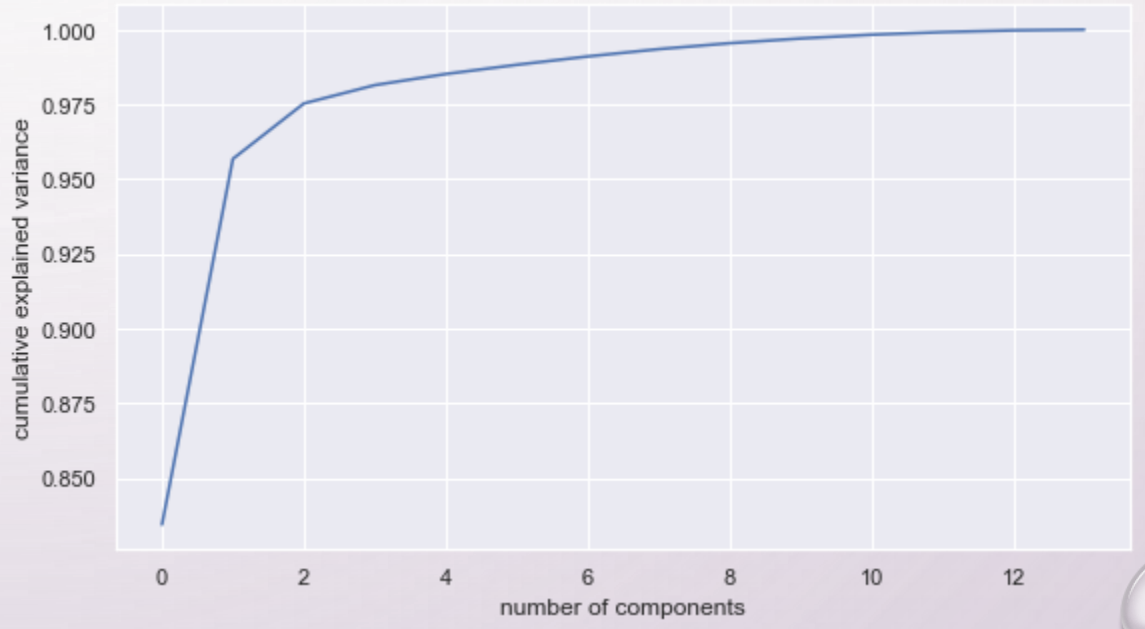
Observations:

- We can see here candies which contains chocolate is 17.95 % points higher in terms of winpercent compared to candies with none chocolate.
- Also fruity taste has a relatively high positive coefficient which contradicts our correlation heatmap. Also from the correlation heatmap we can see chocolate and fruity has a strong negative correlation.
- We got the RMSE for test data is 11.9 and RSquared value is 0.5268, i.e 52.68% of the variance of winpercent can be explained by the factors we have used.
- For VIF is 2.11 , we can say that there is no multicolinearity present in the model.

APPLICATION: 2 : PCA

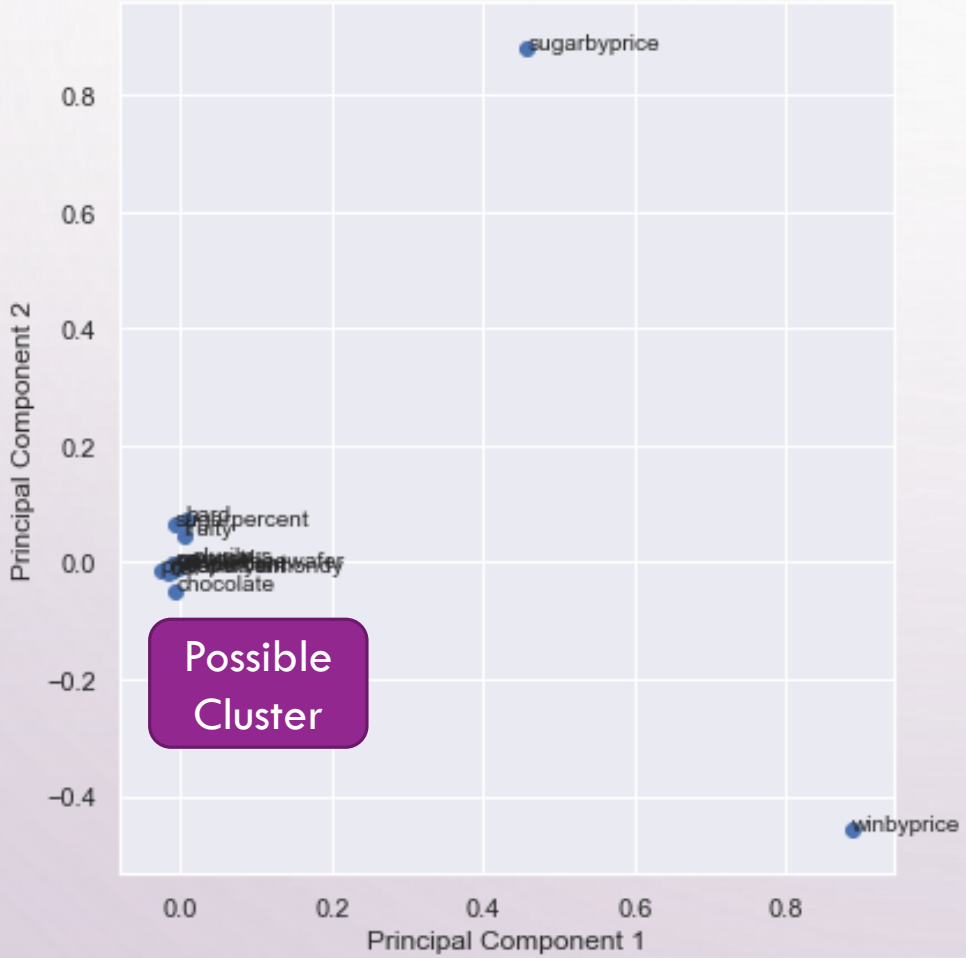
- First 2 or 3 components are suggested as per elbow method:





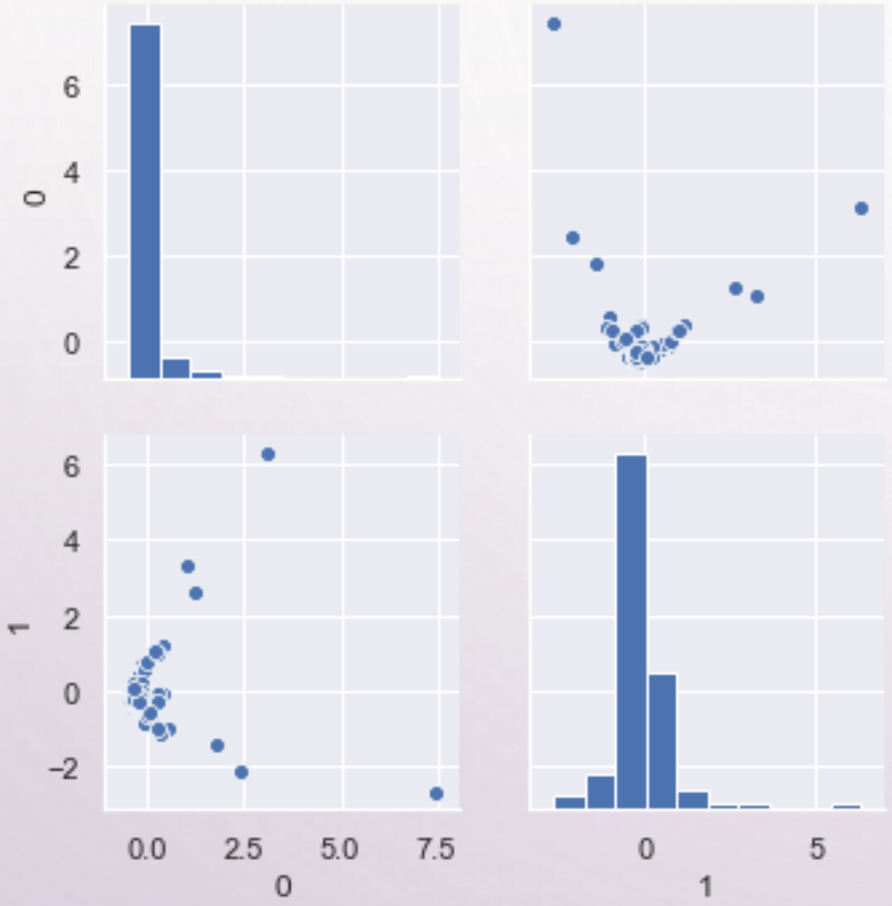
- Variance explained by first 2,3,4 is : 0.957, 0.975, 0.981

APPLICATION: 2 : PCA



- Except Sugarbyprice and Winbyprice other features seems to be clustered in a group.

APPLICATION: 2 : PCA



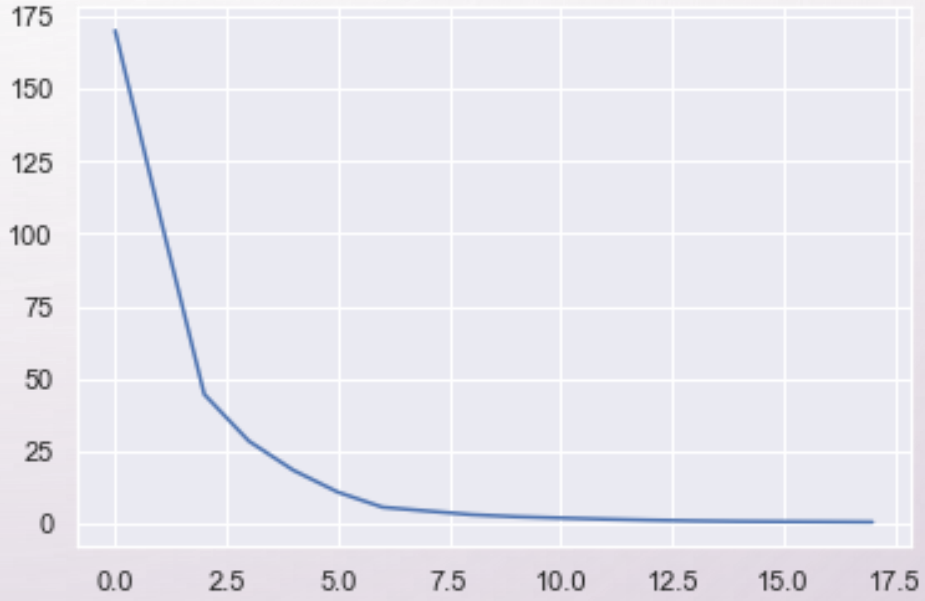
- Post transforming data one cluster is clearly visible. There is possibility of second cluster.

Lets apply clustering and K-Mean

APPLICATION: 3 : CLUSTERING AND K-MEAN

- According to Hopkins statistics there is a possibility of clustering. Hopkins value is 0.978274355139793 7

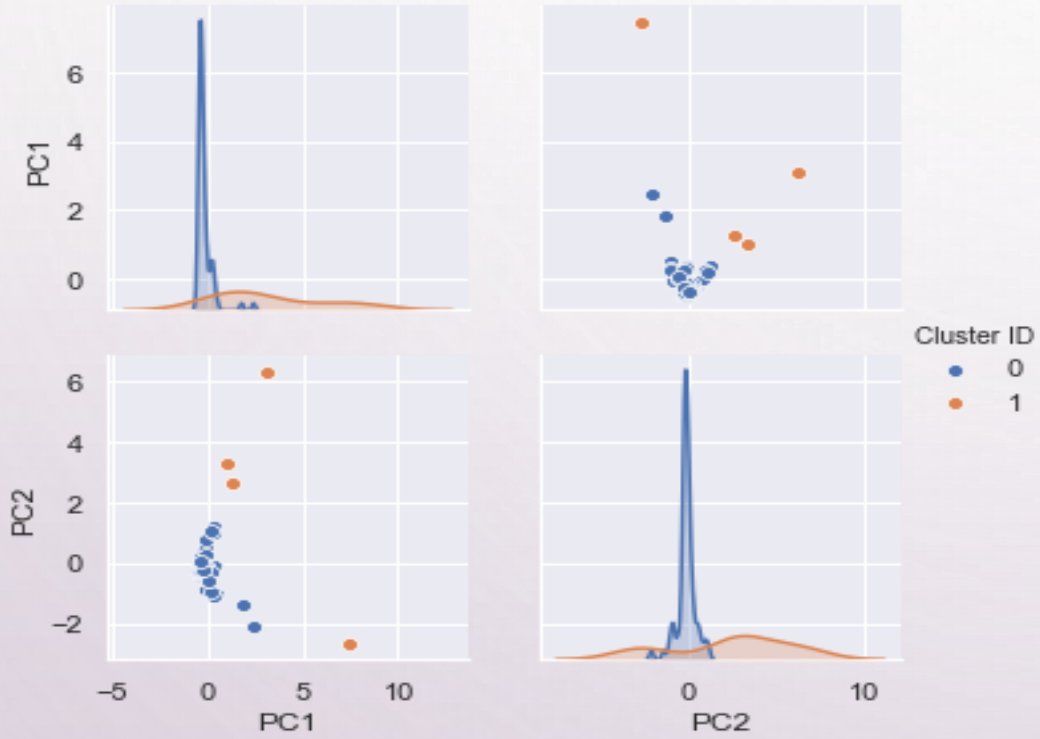
Maximum silhouette score at k=2



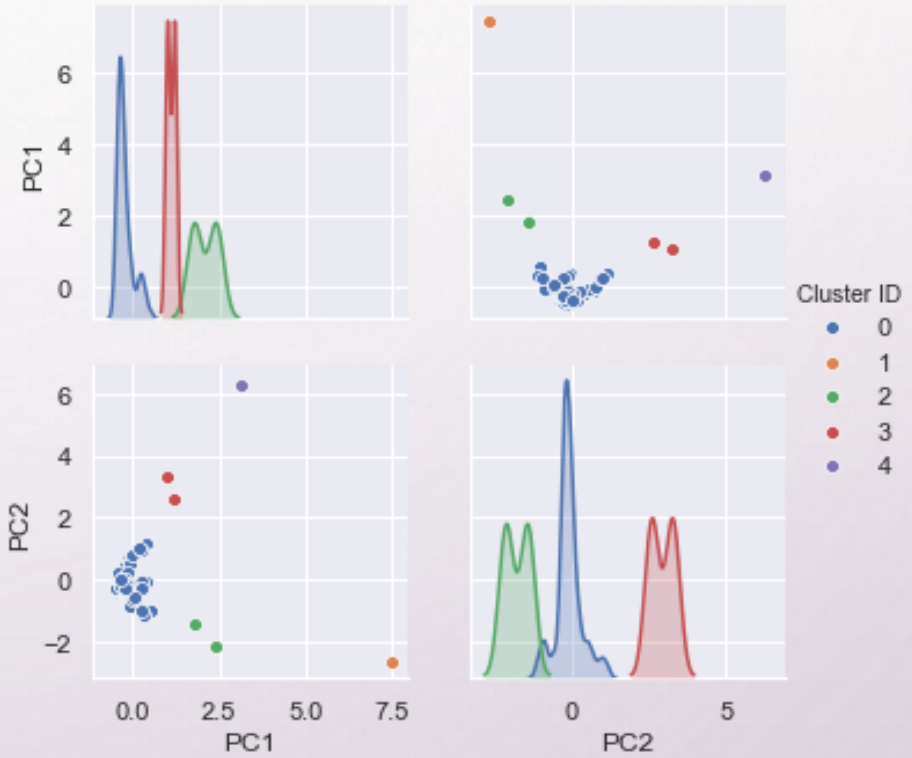
Elbow seems to form at 2

APPLICATION: 3 : CLUSTERING AND K-MEAN

- PCA 1 and PCA 2 visual with cluster 2 IDs



- PCA 1 and PCA 2 visual with cluster 5 IDs



APPLICATION: 3 : CLUSTERING K-MEAN
(POST PCA)

Observations: Analysis of cluster 0 to see how it is differing from rest:

- It is noted that only Cluster ID 4 (Dum Dums) and 1 (Tootsie Roll Midgies) are far away from Cluster ID 0
- 'Dum Dums' is fruity and 'Tootsie Roll Midgies' is chocolaty. Both are sort of opposite of each other.
- Cluster ID 0 contains competitors which are mostly chocolaty, sugary and more favorable. Cluster ID 1, although being chocolaty has a low sugar percentile.
- All the chocolates which don't belong to Cluster ID 0 have made the top 10 list of `winbyprice`. They are all cheap.
- Cluster ID 0 contains competitors which are more chocolaty and more pricey.

APPLICATION: 3 : CLUSTERING K-MEAN
(WITH RIDGE LINEAR REGRESSION)

- Performing Ridge Linear Regression.
- Ridge intercept is 0.3727885990571671
- Result for coefficient by each flavor:

Chocolate :17.395113450849347	1	Crispedricewafer 7.301821467840015
Fruity 7.72063968640181	3	Hard -4.44249782763022
Caramel 2.907233687069857 6		Bar 0.730959088206818 8
Peanutyalmondy 9.09428184646823	2	Pluribus -0.14235867560122606
Nougat 1.376495256219395		

Linear Equation with Ridge method:

y = 32.27 + 17.39 * chocolate + 7.72 * fruity + 2.91 * caramel + 9.09 * peanutyalmondy + 1.38 * nougat + 7.30 * crispedricewafer - 4.44 * hard + 0.73 * bar - 0.14 * pluribus

Comparison of various Machine Learning applications

COMPARISON OF RESULT

ALL THREE APPLICATIONS

Linear Regression

- As per the evaluation with (80:20) training and testing data Chocolate is the most important attribute of a followed by Crispedricewafer, Peanutyalmondy and Nougat.
- As per the evaluation with ridge method Chocolate is the most important attribute of a candy followed by peanutyalmondy, Fruity and crispedricewafer.

PCA

- Other than sugarbyprice rest all are forming single cluster
- 2) After transformation, its observed possibility of one more cluster

Cluster & K-Mean

- Mainly cluser contains Chocolaty brands. Cluster 0 has Sugar contain whereas other brands which are not belong to cluster 0 who are part of top 10 winbyprice are all cheaper.

Overall recommendation is in favor of Chocolate attribute.

Thank you

For further information please contact:
Jigna Thacker
jignazt@yahoo.com
https://github.com/jmps967/INSAID-ML-All-Modules_Jigna-Thacker