PRESENTATION
ON
CHOCOLATE BAR RATING

Presented by,
Divyani Vishwakarma

Before we get started, let's install and import all the relevant python packages which we would use for performing our analysis.

```
In [1]: import numpy as np
        import pandas as pd
        import matplotlib as mpl
        import matplotlib.pyplot as plt
        import seaborn as sns
```

use for Linear algebra

Use for Data processing

It is use to plot a graph

It is use to plot attractive graph

Problem Statement

- The dataset contain a detailed set of Chocolate Bars and the main problem statement here is to determine the dataset that should related to ratings of over individual chocolate bars.

- The data pre-processing is help to extract the useful information and Machine learning help to fit the best algorithm for model and fit best score.

- Let’s analyze the dataset and take a closer look at its content. The aim here to find the total number of data, which will easy to help us to understand the dataset.

```
data=pd.read_csv("flavors_of_cacao.csv")
```

read_csv function load the entire data file into Python environment.

```
data.head()
```

Head function returns the first 5 entries of the dataset.

	Company \n(Maker-if known)	Specific Bean Origin\nor Bar Name	REF	Review\nDate	Cocoa\nPercent	Company\nLocation	Rating	Bean\nType	Broad Bean\nOrigin
0	A. Morin	Agua Grande	1876	2016	63%	France	3.75		Sao Tome
1	A. Morin	Kpime	1676	2015	70%	France	2.75		Togo
2	A. Morin	Atsane	1676	2015	70%	France	3.00		Togo
3	A. Morin	Akata	1680	2015	70%	France	3.50		Togo
4	A. Morin	Quilla	1704	2015	70%	France	3.50		Peru

Introduction to Dataset- Chocolate Bar

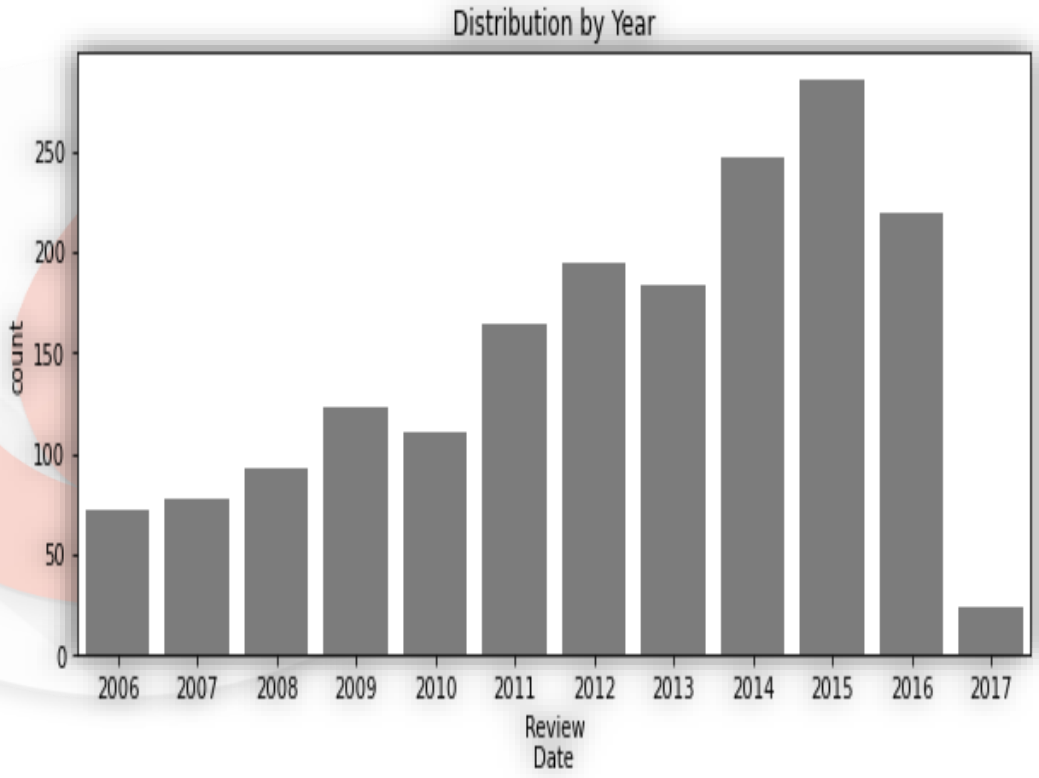
- Chocolate is one of the most popular candies in the world. Each year, residents of the United States collectively eat more than 2.8 billions pounds. However, not all chocolate bars are created equal!
- This dataset contains expert ratings of over 1,700 individual chocolate bars, along with information on their regional origin, percentage of cocoa, the variety of chocolate bean used and where the beans were grown.

Data Pre-processing

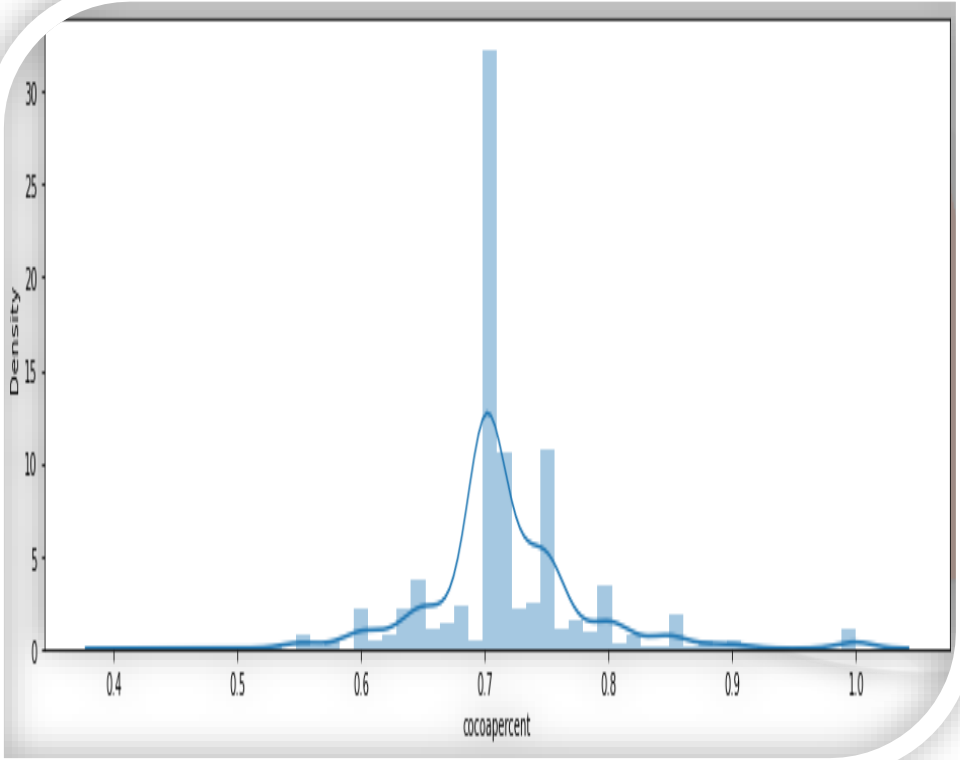
- We use isnull() function, it help to retrieve the total number of null values from the dataset. It is useful for data cleaning process.
- If we see the column names, there is some column names which difficult to spell, we'll try to rename columns.
- And also change the datatype of 'cocoa percent'.

Data Visualization

- Since 2006, when just 71 chocolate bars were rated, there was an annual upward trend until it peaked at 285 in 2015, but has since dipped significantly when just 105 bars were rated last year (2017).



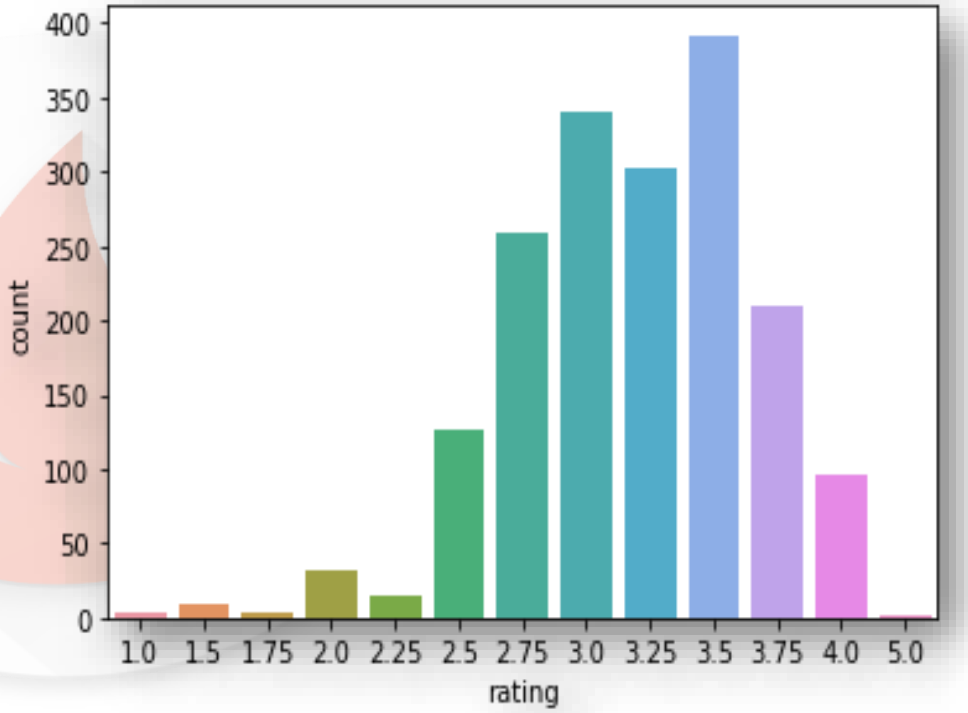
Data Visualization



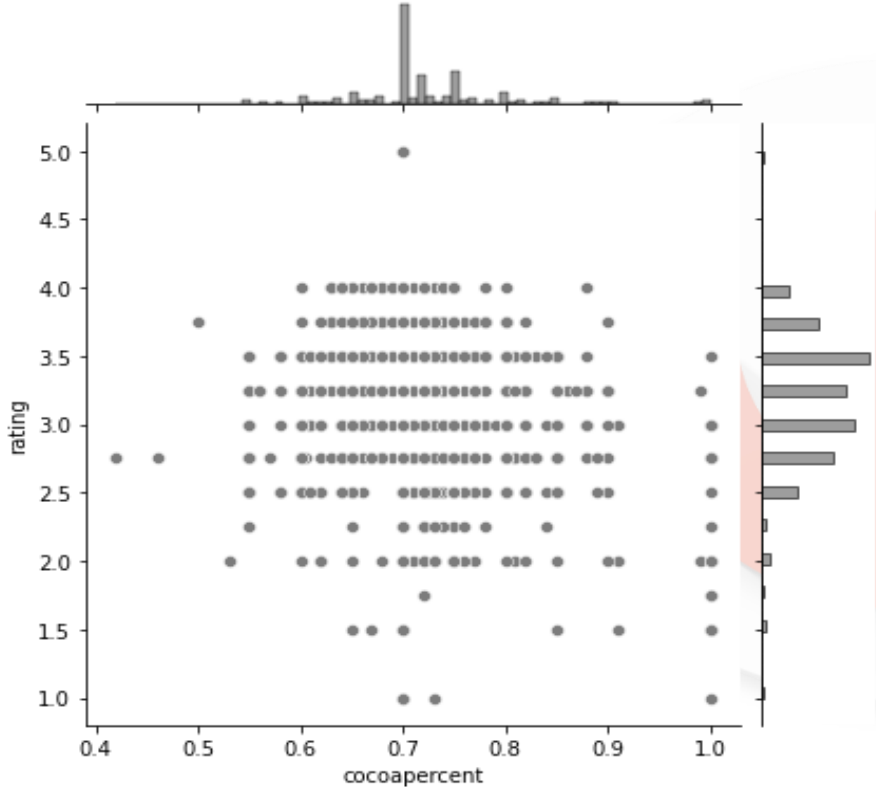
- In the graph, x-axis represent the ‘cocoa percent’ and y-axis represent the ‘density’.
- As we can see visualization, Most of the chocolate has 70% of cocoa in them.

Data Visualization

- The most number of ratings that were given was between 3.0 to 3.5, with the highest being 3.5 with a number of around 380 ratings. This shows us that most individuals are giving chocolate bars a rating of a little bit more than satisfactory.



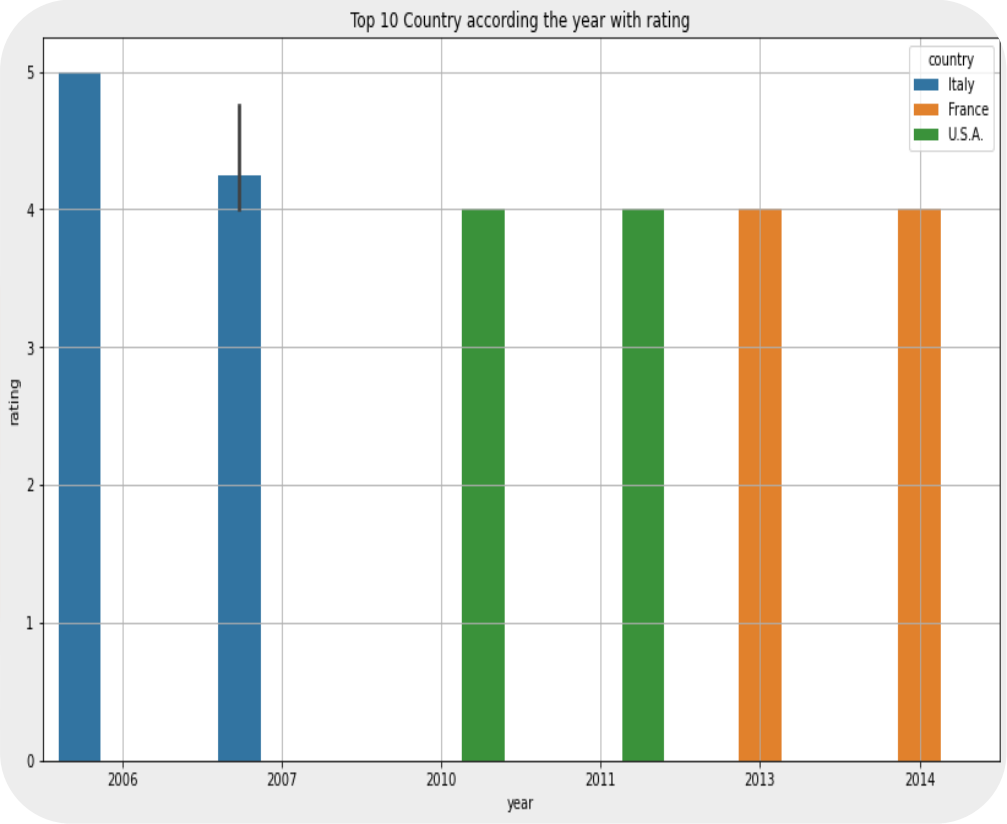
Data Visualization



- In the plot below, we see that cocoa percentage and ratings have a weak relationship, though it does lean slightly negative where a higher cocoa percentage corresponds to a lower rating.

Data Visualization

- In the graph, x-axis represent the ‘year’ and y-axis represent the ‘rating’.
- As per visualization, in 2016 the rating of cocoa beans is highly sale in Italy, and then France and USA



Machine Learning Modelling

Define dependent and independent variable

In [27]:
```
y=data1['rating']
x=data1.drop('rating',axis=1)
```

In [28]:
```
x
```

Out[28]:

	REF	Review\nDate	cocoapercent	A. Morin	AMMA	Acalli	Adi	Aequare (Gianduja)	Ah Cacao	Akesson's (Pralus)	...	Venezuela	Venezuela, Carribean	Venezuela, Dom. Rep.	Venezuela, Ghana
0	1876	2016	0.63	1	0	0	0	0	0	0	...	0	0	0	0
1	1676	2015	0.70	1	0	0	0	0	0	0	...	0	0	0	0
2	1676	2015	0.70	1	0	0	0	0	0	0	...	0	0	0	0
3	1680	2015	0.70	1	0	0	0	0	0	0	...	0	0	0	0
4	1704	2015	0.70	1	0	0	0	0	0	0	...	0	0	0	0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
1790	647	2011	0.70	0	0	0	0	0	0	0	...	0	0	0	0
1791	749	2011	0.65	0	0	0	0	0	0	0	...	0	0	0	0
1792	749	2011	0.65	0	0	0	0	0	0	0	...	0	0	0	0
1793	781	2011	0.62	0	0	0	0	0	0	0	...	0	0	0	0

Scaling

In [29]:
```
from sklearn.preprocessing import MinMaxScaler
```

In [30]:
```
scaler=MinMaxScaler()
x= scaler.fit_transform(x)
```

Machine Learning Modelling

Training Model

In [31]:
```
from sklearn.model_selection import train_test_split
x_train, x_test, y_train, y_test=train_test_split(x,y,test_size=0.2,random_state=42)
```

In [32]:
```
from sklearn.ensemble import RandomForestRegressor
reg_rf= RandomForestRegressor()
reg_rf.fit(x_train,y_train)
```

Out[32]:
```
RandomForestRegressor()
```

In [33]:
```
y_pred=reg_rf.predict(x_test)
```

In [34]:
```
reg_rf.score(x_train,y_train)
```

Out[34]:
```
0.8885184851503938
```

In [35]:
```
reg_rf.score(x_test,y_test)
```

Out[35]:
```
0.16213179486722062
```

Advantage

i. It reduces overfitting in decision trees and helps to improve the accuracy
ii. It is flexible to both classification and regression problems
iii. It works well with both categorical and continuous values
iv. It automates missing values present in the data
v. Normalizing of data is not required as it uses a rule-based approach.

Disadvantage

I. It requires much computational power as well as resources as it builds numerous trees to combine their outputs.
II. It also requires much time for training as it combines a lot of decision trees to determine the class.
III. Due to the ensemble of decision trees, it also suffers interpretability and fails to determine the significance of each variable.

Conclusion

- The chocolates of 70% coco were in maximum production.
- Chocolates with rating of 3.5 were around 380 i.e. maximum units.
- The “Random Forest” model gave us accuracy of 88.85% as in predicting the ratings

Future Scope

- ✓ Chocolates with 70% cocoa will se increase in production , as it balances the bitterness of dark chocolate(95%) and sweetness of milk chocolate (40%).

- ✓ With prediction of 88.85 % its possible that chocolate with rating 3.5 will be used for production.