Yangdi Shen-Q12

January 28, 2018

1 Exploratory Data Analysis

Here you'll bring together some of the individual skills that you demonstrated above and create a Jupyter notebook based blog post on data analysis.

1. Find a dataset that interests you and relates to a question or problem that you find intriguing
2. Using a Jupyter notebook, describe the dataset, the source of the data, and the reason the dataset was of interest.
3. Check the data and see if they need to be cleaned: are there missing values? Are there clearly erroneous values? Do two tables need to be merged together? Clean the data so it can be visualized.
4. Plot the data, demonstrating interesting features that you discover. Are there any relationships between variables that were surprising or patterns that emerged? Please exercise creativity and curiosity in your plots.
5. What insights are you able to take away from exploring the data? Is there a reason why analyzing the dataset you chose is particularly interesting or important? Summarize this as if your target audience was the readership of a major news organization - boil down your findings in a way that is accessible, but still accurate.
6. Create a public repository on your github account titled "machine-learning-course". In it, create a readme file that contains the heading "ECE590: Introductory Machine Learning for Data Science". Add, commit, and push that Jupyter notebook to the master branch. Provide the link to the that post here.

This dataset is about the ratings of chocolate bars. The attributes include location of production, time, cocoapercent, bean type and so on. The interesting point of this dataset is that many people consume chocolate bars every day and they rate them online, but we hardly discuss the statitical feature of our ratings. What affect our judgement on Chocolate bar? Where could we find the most popular chocolate bars? The question is simple but many of us want to know the answers.

This dataset consists of over 1700 rows of data and it is downloaded from kaggle as one of the hottest datasets. All the data have been completed, therefore we don't need to clean it. The rating scale is 5 points. Rating System could be described as the following:

5= Elite (Transcending beyond the ordinary limits)

4= Premium (Superior flavor development, character and style)

3= Satisfactory(3.0) to praiseworthy(3.75) (well made with special qualities)

2= Disappointing (Passable but contains at least one significant flaw)

1= Unpleasant (mostly unpalatable)

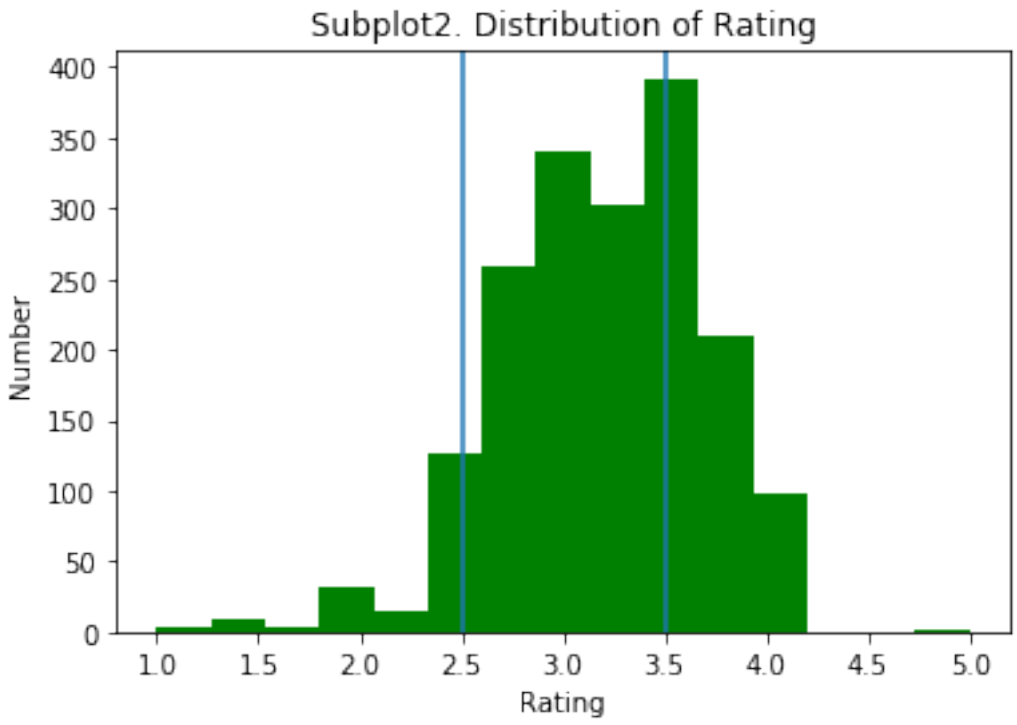
In general, I attempt to know what is the relationship between bar ratings and cocoa percent and where could we find best chocolate.

In [153]:
```
#import data and show the head of dataframe
import pandas as pd
import matplotlib.pyplot as plt
import numpy as np

d=pd.read_csv(r'D:\Machine\flavors_of_cacao1.csv')
d.head()
```

Out[153]:

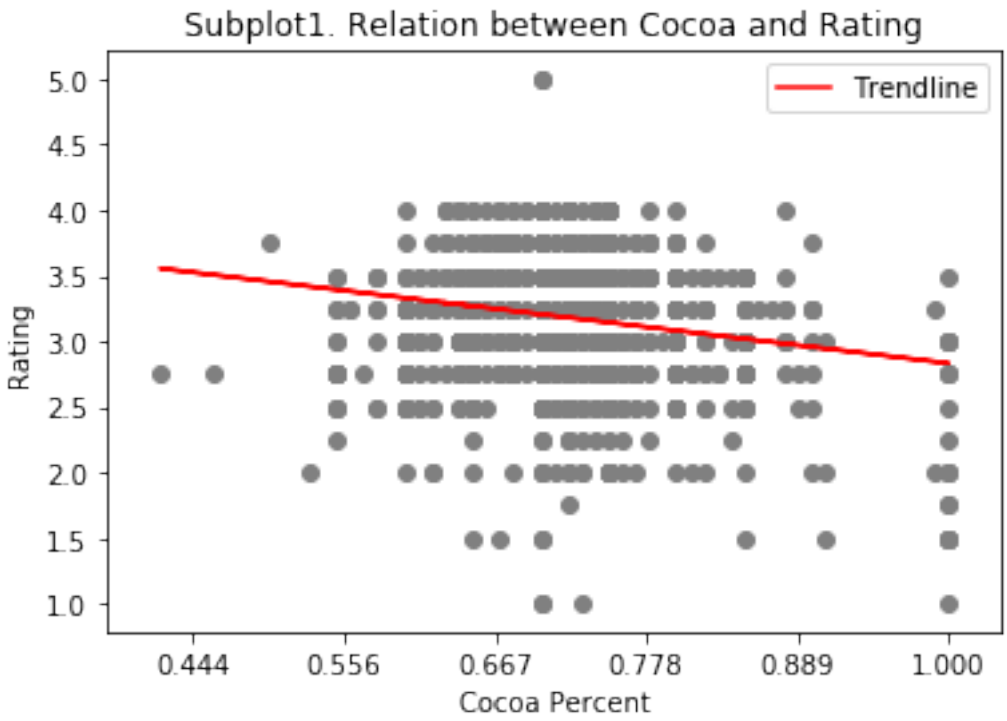
	Companyă\n(Maker-if known)	Specific Bean Origin\nor Bar Name	REF	Review\nDate	CocoaPercent	location	Rating	Bean\nType	Broad Bean\nOrigin
0	A. Morin	Agua Grande	1876	2016	0.63	France	3.75	ă	Sao Tome
1	A. Morin	Kpime	1676	2015	0.70	France	2.75	ă	Togo
2	A. Morin	Atsane	1676	2015	0.70	France	3.00	ă	Togo
3	A. Morin	Akata	1680	2015	0.70	France	3.50	ă	Togo
4	A. Morin	Quilla	1704	2015	0.70	France	3.50	ă	Peru

In [176]:
```
#Distribution of Rating

plt.hist(d['Rating'],15,color="green")
plt.title('Subplot2. Distribution of Rating')
plt.axvline(2.5)
plt.axvline(3.5)
plt.xlabel('Rating')
plt.ylabel('Number')
plt.show()
```



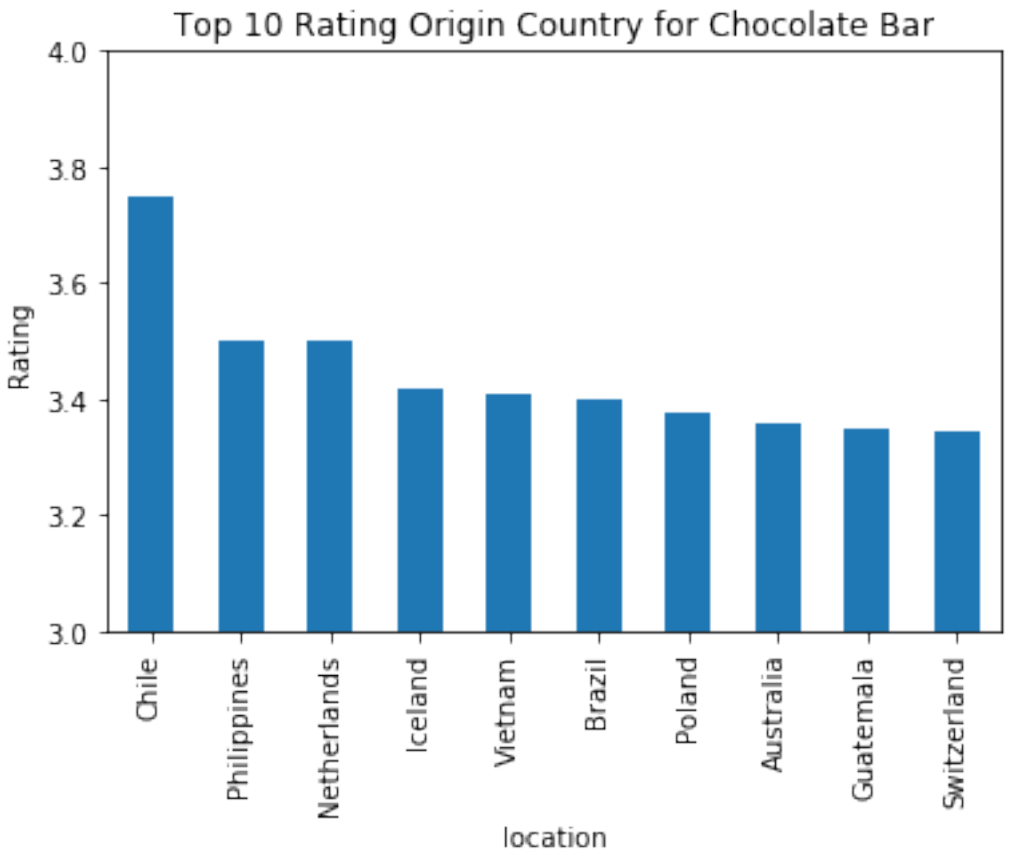
Initially, we should know the overall situation here. Plotting the distribution of ratings, we could see most of the ratings are located in the interval from 2.5 to 3.5. The distribution is a little bit negative-skewed.

In [155]:
```
#Relationship between cocoapercent and rating

tick1=np.linspace(0,1,10)
plt.xticks(tick1)
plt.scatter(d['CocoaPercent'],d['Rating'],c="grey")
z = np.polyfit(d['CocoaPercent'], d['Rating'], 1)
p = np.poly1d(z)
plt.plot(d['CocoaPercent'],p(d['CocoaPercent']),c="red")

plt.xlabel("Cocoa Percent")
plt.ylabel("Rating")
plt.title('Subplot1. Relation between Cocoa and Rating')
plt.legend(['Trendline'])
plt.show()
```



Then, we could use a scatter plot and a trendline to show the relationship between cocoa percents and ratings. As is shown in the graph, these two factors have a negative correlation. The higher cocoa percent the chocolate has, the lower rating it would receive.

```
In [178]: #Top 10 country on chocolate bar
          group=d.groupby("location")
          geo=group['Rating'].agg('mean')
          geo=geo.sort_values(ascending=False)
          geo1=geo[0:10]
          geo1.plot(kind="bar")

          plt.ylim(3,4)
          plt.title('Top 10 Rating Origin Country for Chocolate Bar')
          plt.ylabel('Rating')
          plt.show()
```



We could use groupby tools to get the mean values for each country. I chose ten countries which have the highest ratings for its chocolate bars. Chile has the highest ratings, followed by philippines and netherland. Most of the countries are located in the central-south america and Europe. Surpringly, France and Italy, two countries which is famous for its chocolate, are not in this "elites list".