**Descriptive/Exploratory Data Analysis of Chocolate Bar Preference**

**Data Origin Description**

The data I chose was an open access dataset available on Kaggle, used by a Kaggle data scientist to demonstrate coding in Python. It can be found here: <https://www.kaggle.com/rtatman/chocolate-bar-ratings/version/1#flavors_of_cacao.csv>

I tried to stay away from data already analyzed with R in order to demonstrate my understanding and original work. The original data was compiled by Brady Brelinski of the Manhattan Chocolate Society. The data includes information on chocolate bars tasted, catalogued, and rated by the author of “Flavors of Cacao” <http://flavorsofcacao.com/index.html>.

I selected the csv. dataset from the Kaggle website that I previously noted and imported the entire dataset into R Studio software for data analysis. In R Studio, I imported the Lubridate, Dyplr, ggplot, and Tidyverse packages to help with data clean-up, analysis, and visualization.

**Variable Description**

Data includes information that can be categorized into library visits, circulation, collection size, operational such as hours, revenue and staffing, resource format, and service outlet information including location, and type.

The original dataset contained 1,795 rows and 9 columns of information. Each row represents a unique chocolate bar “batch” tasted and ranked by the author. The dataset is focused on plain dark chocolate.

The original column headings are included below and represent the survey questions. Batch numbers, vintages and review dates are included in the database when known.   
This survey is a universe survey and not a sample.

Column Headings

[1] "CompanyÂ...Maker.if.known."

[2] "Specific.Bean.Origin.or.Bar.Name"

[3] "REF"

[4] "Review.Date"

[5] "Cocoa.Percent"

[6] "Company.Location"

[7] "Rating"

[8] "Bean.Type"

[9] "Broad.Bean.Origin"

Ratings are ranked on the following scale. Integers are not all whole and can be represented from 1 – 5 in 0.25 intervals.

5= Elite (Transcending beyond the ordinary limits)

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

3= Satisfactory (3.0) to praise-worthy (3.75) (well made with special qualities)

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

1= Unpleasant (mostly unpalatable)

**Dataset Limitations/Missing Data**

Column 2, “"Specific.Bean.Origin.or.Bar.Name" seems to be a mix of variables. As “bean origin” is in column 9, I will assume that column 2 acts as a unique identifier.

The “Rating” According to the “Flavors of Cacao” website, each chocolate is evaluated from a combination of both objective qualities and subjective interpretation from a single reviewer.

Cocoa.Percent is listed as a factor. I removed the “%” sign and changed them into integers.

“Bean Type” has 887 instances of missing data. “Broad Bean Origin” has 73 instances of missing data.

**Central Tendency, Range, and distribution by Variable**

The variables that I focused on were “Rating” and “Cacao Percent”. The distribution, along with visuals are below. I additionally used R to calculate the frequency of the company name and location variables.

Both Rating and Cacao Percent had fairly standard distributions, with the average Rating being about “3 – Satisfactory”, which trends with the scale. The mean of the Cacao Percent is 71 with a standard deviation of 6.32.

***Company/Maker*** **(top instances)**

Soma : 47

Bonnat : 27

Fresco : 26

Pralus : 25

A. Morin: 23

Arete : 22

(Other) :1625

***Company Location* (top instances)**

U.S.A. :764

France :156

Canada :125

U.K. : 96

Italy : 63

Ecuador: 54

(Other):537

***Rating***

Min. :1.000

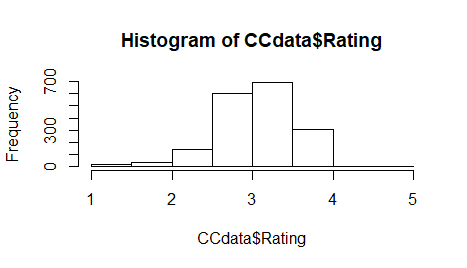
1st Qu.:2.875

Median :3.250

Mean :3.186

3rd Qu.:3.500

Max. :5.000



***Cocoa Percent***

Min. : 42.0

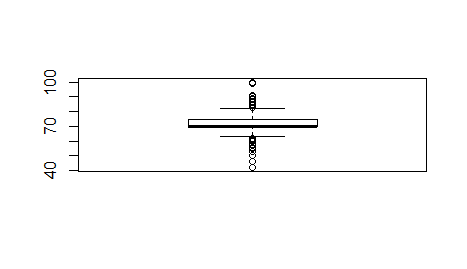
1st Qu.: 70.0

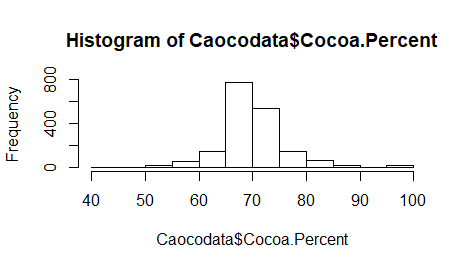
Median : 70.0

Mean : 71.7

3rd Qu.: 75.0

Max. :100.0

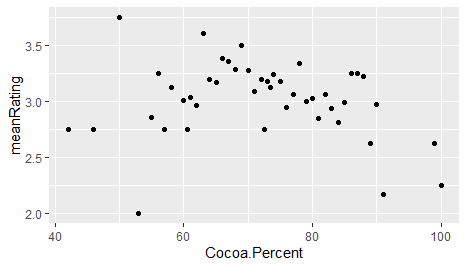


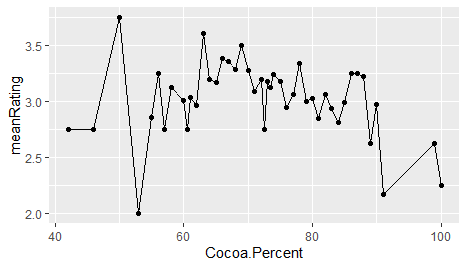


**Further Analysis**

After attempts visually analyze the data several ways (which can be found in the “Bloopers and Outtakes” section in the Appendix!), I attempted to explore the Rating as it related to the Percent of Cocoa of the chocolate bars.

The scatterplots below could indicate that the reviewer likes chocolate bars made with about 50% Cocoa the best. However, a further look into the distribution and the individual variables seem to show that the reviewer may enjoy 87% Cocoa, due to the clustering around those variables.

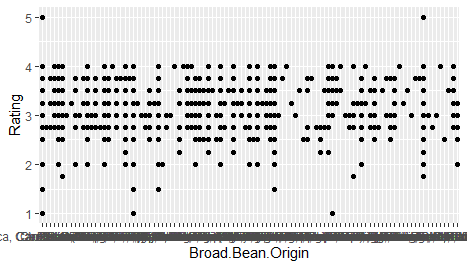




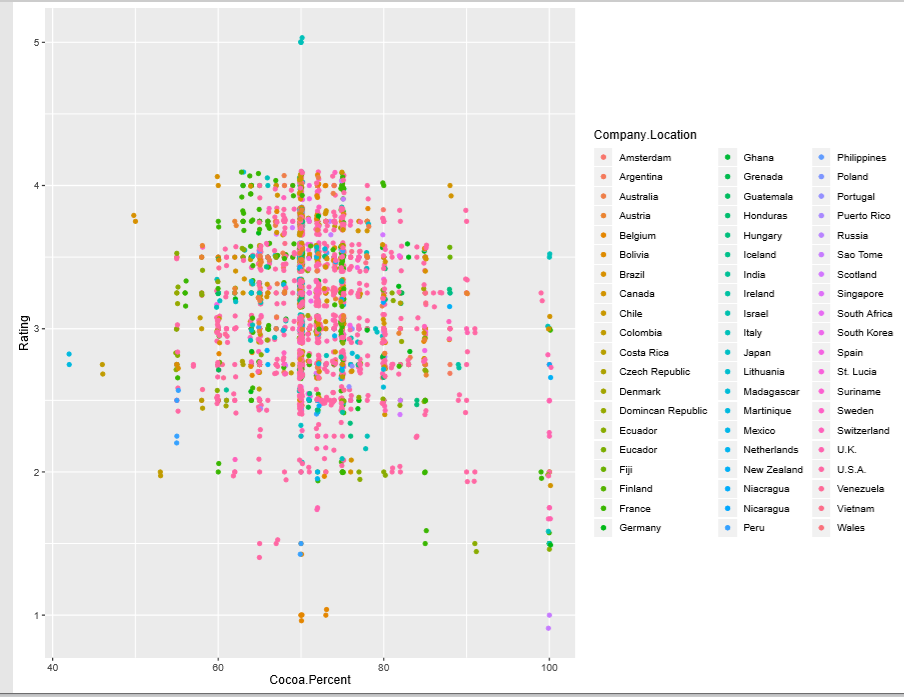
**Appendix**

***Bloopers and outtakes***

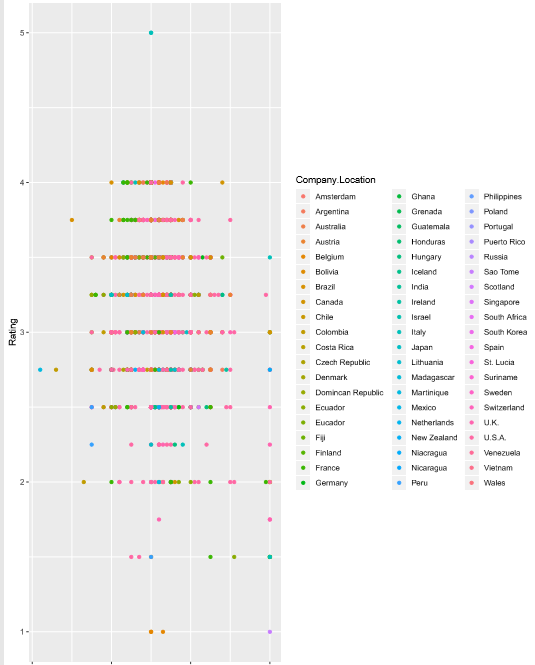
**Rating by Bean Origin** – Not the best way to analyze it, as the variables aren’t standardized



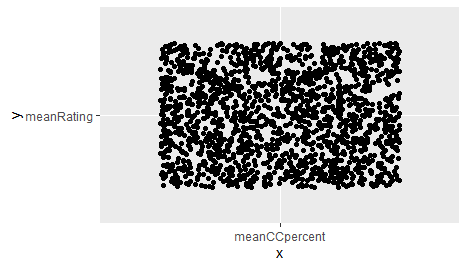
**Rating v % Cocoa by Company Location** – Way too much data!!!!



**Rating v % Cocoa by Company Location Grouped** – Still pretty weird and tells me little



**Plotting Averages** – I can’t even tell you what happened here, but I know I should never do this again



**Appendix**

**Code**

install.packages("ggthemes")

library(lubridate)

library(dplyr)

library(tidyverse)

library(ggplot2)

library(ggthemes)

Caocodata <- read.csv("School/Data Analysis/ExplorDataAssign/flavors\_of\_cacao.csv", header = TRUE)

class(Caocodata)

dim(Caocodata)

names(Caocodata)

str(Caocodata)

glimpse(Caocodata)

summary(Caocodata)

head(Caocodata)

tail(Caocodata)

Caocodata$Cocoa.Percent <- gsub("\\%","",Caocodata$Cocoa.Percent)

head(Caocodata)

Caocodata$Cocoa.Percent <- as.numeric(Caocodata$Cocoa.Percent)

hist(Caocodata$Cocoa.Percent)

boxplot(Caocodata$Cocoa.Percent)

hist(CCdata$Rating)

boxplot(CCdata$Rating)

Caocodata$Bean.Type <- gsub("\\Â","",Caocodata$Bean.Type)

CCbean <- Caocodata$Bean.Type [Caocodata$Bean.Type==""]<-"NA"

Caocodata$Broad.Bean.Origin <- gsub("\\Â","",Caocodata$Broad.Bean.Origin)

CCorigin <- Caocodata$Broad.Bean.Origin [Caocodata$Broad.Bean.Origin==""]<-"NA"

sum(is.na(Caocodata))

CCdata <- Caocodata[,]

head(CCdata)

mean(CCdata$Rating)

median(CCdata$Rating)

maxrating <- max(CCdata$Rating)

print(maxrating)

minrating <- min(CCdata$Rating)

print(minrating)

glimpse(CCdata)

hist(CCdata$Rating)

unique(CCdata$Broad.Bean.Origin)

unique(CCdata$Bean.Type)

ggplot(CCdata,aes(x=Broad.Bean.Origin, y=Rating)) + geom\_point()

summary(CCdata$Rating)

CCdata %>%

summarise(meanRating = mean(Rating),

sdevRating = sd(Rating, na.rm=FALSE))

CCdata %>%

summarise(meanCCpercent = mean(Cocoa.Percent),

sdevRating = sd(Cocoa.Percent, na.rm=FALSE))

ggplot(CCdata, aes(x= Cocoa.Percent, y = Rating)) +

geom\_point() +

geom\_jitter()

ggplot(CCdata, aes(x= Cocoa.Percent, y = Rating, color = Company.Location)) +

geom\_point() +

geom\_jitter()

ggplot(CCdata, aes(x= "meanCCpercent", y = "meanRating")) +

geom\_point() +

geom\_jitter()

CoP\_by\_Rating <- CCdata %>%

group\_by(Cocoa.Percent) %>%

summarise(meanRating = mean(Rating))

ggplot(CoP\_by\_Rating, aes(y= meanRating, x = Cocoa.Percent)) +

geom\_point()+

geom\_line()

summary(CCdata)