Exploratory Data Analysis with Chocolate Dataset

Anil Kumar Jha

23/09/2021

Introduction

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. This dataset was provided by [kaggle] (https://www.kaggle.com/rtatman/chocolate-bar-ratings)

Import required library

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --

## v ggplot2 3.3.5     v purrr   0.3.4
## v tibble  3.1.4     v dplyr   1.0.7
## v tidyr   1.1.3     v stringr 1.4.0
## v readr   2.0.1     v forcats 0.5.1

## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(dplyr)
library(ggplot2)
```

Load the dataset

```
chocolate_project<- read.csv("flavors_of_cacao.csv")
colnames(chocolate_project)
```

```
## [1] "CompanyÂ...Maker.if.known."       "Specific.Bean.Origin.or.Bar.Name"
## [3] "REF"                              "Review.Date"
## [5] "Cocoa.Percent"                    "Company.Location"
## [7] "Rating"                           "Bean.Type"
## [9] "Broad.Bean.Origin"
```

```
head(chocolate_project)
```

```
##   CompanyÂ...Maker.if.known. Specific.Bean.Origin.or.Bar.Name  REF Review.Date
## 1                  A. Morin                      Agua Grande 1876        2016
## 2                  A. Morin                            Kpime 1676        2015
## 3                  A. Morin                           Atsane 1676        2015
## 4                  A. Morin                            Akata 1680        2015
## 5                  A. Morin                           Quilla 1704        2015
## 6                  A. Morin                         Carenero 1315        2014
##   Cocoa.Percent Company.Location Rating Bean.Type Broad.Bean.Origin
## 1           63%           France   3.75        Â           Sao Tome
## 2           70%           France   2.75        Â               Togo
## 3           70%           France   3.00        Â               Togo
## 4           70%           France   3.50        Â               Togo
## 5           70%           France   3.50        Â               Peru
## 6           70%           France   2.75   Criollo         Venezuela
```

```
View(head(chocolate_project))
str(chocolate_project)
```

```
## 'data.frame':    1793 obs. of  9 variables:
##  $ CompanyÂ...Maker.if.known.      : chr  "A. Morin" "A. Morin" "A. Morin" "A. Morin" ...
##  $ Specific.Bean.Origin.or.Bar.Name: chr  "Agua Grande" "Kpime" "Atsane" "Akata" ...
##  $ REF                            : int  1876 1676 1676 1680 1704 1315 1315 1315 1319 1319 ...
##  $ Review.Date                    : int  2016 2015 2015 2015 2015 2014 2014 2014 2014 2014 ...
##  $ Cocoa.Percent                  : chr  "63%" "70%" "70%" "70%" ...
##  $ Company.Location               : chr  "France" "France" "France" "France" ...
##  $ Rating                         : num  3.75 2.75 3 3.5 3.5 2.75 3.5 3.5 3.75 4 ...
##  $ Bean.Type                      : chr  "Â " "Â " "Â " "Â " ...
##  $ Broad.Bean.Origin              : chr  "Sao Tome" "Togo" "Togo" "Togo" ...
```

Data Preparation

```
# Clean the column name
names(chocolate_project) <- gsub(x = names(chocolate_project), pattern = "\\.", replacement = "_")
str(chocolate_project)
```

```
## 'data.frame':    1793 obs. of  9 variables:
##  $ CompanyÂ___Maker_if_known_      : chr  "A. Morin" "A. Morin" "A. Morin" "A. Morin" ...
##  $ Specific_Bean_Origin_or_Bar_Name: chr  "Agua Grande" "Kpime" "Atsane" "Akata" ...
##  $ REF                            : int  1876 1676 1676 1680 1704 1315 1315 1315 1319 1319 ...
##  $ Review_Date                    : int  2016 2015 2015 2015 2015 2014 2014 2014 2014 2014 ...
##  $ Cocoa_Percent                  : chr  "63%" "70%" "70%" "70%" ...
##  $ Company_Location               : chr  "France" "France" "France" "France" ...
##  $ Rating                         : num  3.75 2.75 3 3.5 3.5 2.75 3.5 3.5 3.75 4 ...
##  $ Bean_Type                      : chr  "Â " "Â " "Â " "Â " ...
##  $ Broad_Bean_Origin              : chr  "Sao Tome" "Togo" "Togo" "Togo" ...
```

```
View(head(chocolate_project))

# Rename 2 column names
colnames(chocolate_project)[1]<- "Company_name"
colnames(chocolate_project)[2]<- "Bean_Origin"
View(head(chocolate_project))

# Find any null value in dataset
sapply(chocolate_project, function(x) sum(is.na(x)))
```

```
##      Company_name       Bean_Origin               REF       Review_Date
##                 0                 0                 0                 0
##     Cocoa_Percent  Company_Location            Rating         Bean_Type
##                 0                 0                 0                 0
## Broad_Bean_Origin
##                 0
```

```
summary(chocolate_project)
```

```
##  Company_name       Bean_Origin            REF        Review_Date
##  Length:1793        Length:1793        Min.   :   5   Min.   :2006
##  Class :character   Class :character   1st Qu.: 576   1st Qu.:2010
##  Mode  :character   Mode  :character   Median :1073   Median :2013
##                                        Mean   :1036   Mean   :2012
##                                        3rd Qu.:1502   3rd Qu.:2015
##                                        Max.   :1952   Max.   :2017
##  Cocoa_Percent      Company_Location       Rating       Bean_Type
##  Length:1793        Length:1793        Min.   :1.000   Length:1793
##  Class :character   Class :character   1st Qu.:3.000   Class :character
##  Mode  :character   Mode  :character   Median :3.250   Mode  :character
##                                        Mean   :3.186
##                                        3rd Qu.:3.500
##                                        Max.   :5.000
##  Broad_Bean_Origin
##  Length:1793
##  Class :character
##  Mode  :character
##
##
##
```

```
# Find the unquie value in a column
table(chocolate_project$Bean_Type)
```

```
##
##                      Ã                  Amazon              Amazon mix
##                    887                       1                       2
##             Amazon, ICS                 Beniano                   Blend
##                       2                       3                      41
##  Blend-Forastero,Criollo                   CCN51                 Criollo
##                       1                       1                     153
##         Criollo (Amarru)     Criollo (Ocumare 61)     Criollo (Ocumare 67)
```

```
##                         2                         2                         1
##     Criollo (Ocumare 77)        Criollo (Ocumare)      Criollo (Porcelana)
##                         1                         1                        10
##           Criollo (Wild)               Criollo, +       Criollo, Forastero
##                         1                         1                         2
##      Criollo, Trinitario                       EET                 Forastero
##                        39                         3                        87
##    Forastero (Amelonado)       Forastero (Arriba)   Forastero (Arriba) ASS
##                         1                        37                         6
##  Forastero (Arriba) ASSS      Forastero (Catongo)     Forastero (Nacional)
##                         1                         2                        52
##    Forastero (Parazinho)   Forastero(Arriba, CCN)    Forastero, Trinitario
##                         8                         1                         1
##                   Matina                  Nacional        Nacional (Arriba)
##                         3                         2                         3
##               Trinitario Trinitario (85% Criollo)   Trinitario (Amelonado)
##                       418                         2                         1
##     Trinitario (Scavina)      Trinitario, Criollo    Trinitario, Forastero
##                         1                         9                         2
##     Trinitario, Nacional         Trinitario, TCGA
##                         1                         1
```

```
# Convert % into decimal
chocolate_project$Cocoa_Percent <- as.numeric(sub("%", "",chocolate_project$Cocoa_Percent,fixed=TRUE))/
View(head(chocolate_project))
```

From the summary, we can find some information :

- The review of chocolate data was publicized from 2006 to 2017.
- The percentage of cocoa in chocolate was minimal 10% and maximal 99%.
- Some location of company which produced chocolate bars are USA, France, Canada, U.K., Italy, Ecuador, etc.
- The range of rating is 1 to 5.

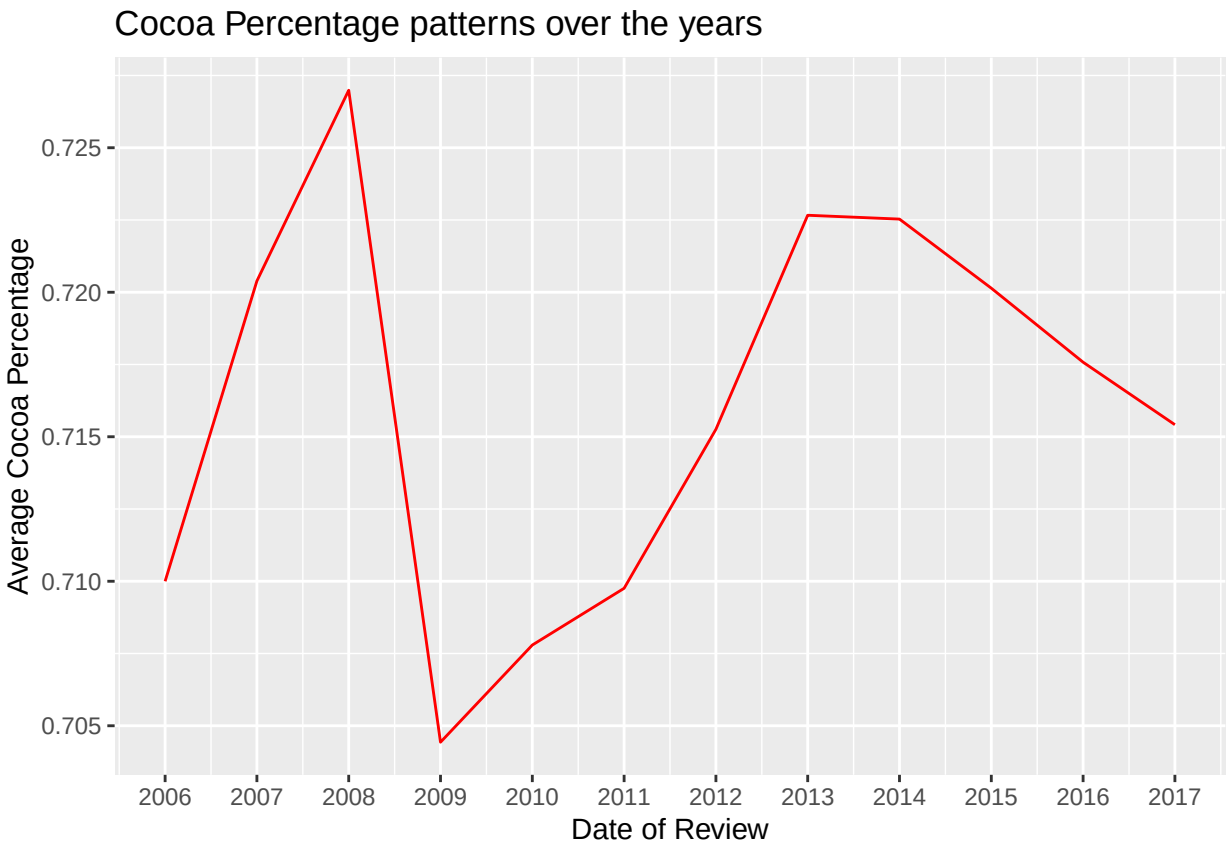
Cocoa Percentage patterns over the years

```
chocolate_review_date<-chocolate_project %>%
  group_by(Review_Date) %>%
  summarise(Cocoa_Percent = mean(Cocoa_Percent))

chocolate_review_date
```

```
## # A tibble: 12 x 2
##    Review_Date Cocoa_Percent
##          <int>         <dbl>
##  1        2006         0.71
##  2        2007         0.720
##  3        2008         0.727
##  4        2009         0.704
##  5        2010         0.708
```

```
##  6        2011         0.710
##  7        2012         0.715
##  8        2013         0.723
##  9        2014         0.723
## 10        2015         0.720
## 11        2016         0.718
## 12        2017         0.715
```

```
ggplot(data=chocolate_review_date, mapping=aes(x=Review_Date, y=Cocoa_Percent)) +
  geom_line( color="red")+
  scale_x_continuous(breaks = seq(2006, 2017, by = 1))+
  xlab("Date of Review") +
  ylab("Average Cocoa Percentage") +
  ggtitle("Cocoa Percentage patterns over the years")
```



Percentage of Cocoa over the years (Taking the average amounts per year)

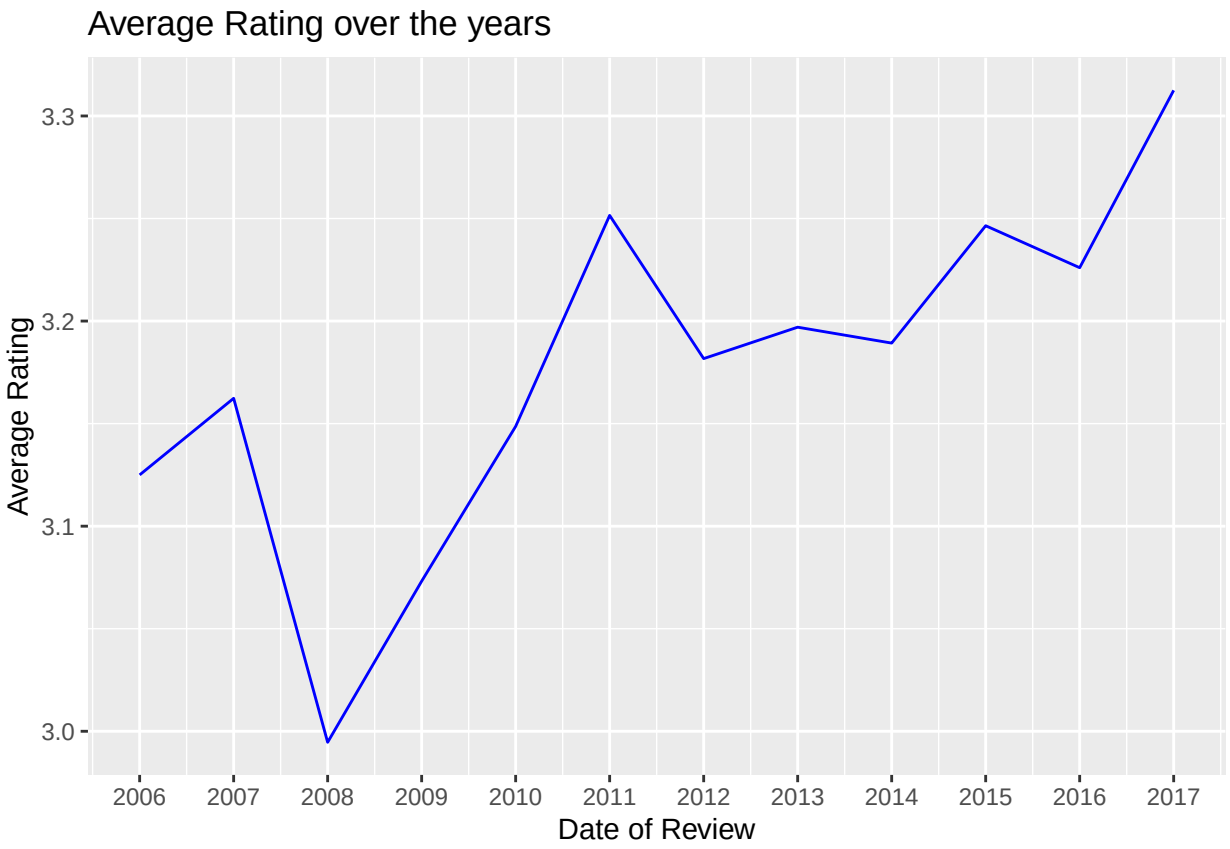
- The highest percentage of cocoa in a chocolate bar came in 2008 and was about 73%.
- The lowest percentage of cocoa followed in the very next year, 2009 and hit 69%.
- There was a steep rise in the amount of cocoa in chocolate from 2009 to 2013 where it rose to about 72.2% from 69%.
- From 2014, a steady decline in cocoa percentage in chocolate bars have been noticed and in 2017, it stands at just above 71.5%

Rating Patterns over the year

```
rating_review_date<- chocolate_project %>%
  group_by(Review_Date) %>%
  summarise(Rating = mean(Rating))

rating_review_date
```

```
## # A tibble: 12 x 2
##    Review_Date Rating
##          <int>  <dbl>
##  1        2006   3.12
##  2        2007   3.16
##  3        2008   2.99
##  4        2009   3.07
##  5        2010   3.15
##  6        2011   3.25
##  7        2012   3.18
##  8        2013   3.20
##  9        2014   3.19
## 10        2015   3.25
## 11        2016   3.23
## 12        2017   3.31
```

```
ggplot(data=rating_review_date, mapping=aes(x=Review_Date, y=Rating))+
  geom_line(color="blue")+
  scale_x_continuous(breaks = seq(2006, 2017, by = 1))+
  xlab("Date of Review")+
  ylab("Average Rating")+
  ggtitle("Average Rating over the years")
```



Rating over the years (Taking the average amounts per year)

- The lowest ever average rating was around 3 and it came in 2008.
- Since then to 2011, there was a steady increase in average ratings and in 2011 it was at 3.26.
- From 2011 to 2017, there have been several fluctuations in the ratings, and in 2017 the rating lies at its apex at around 3.31.

Following trends found in year 2008:

- The highest average cocoa percent was in 2008
- The lowest average ratings came in 2008

The next year 2009 saw two major changes from the previous year :

- There was a drastic reduction in cocoa content on an average
- The average rating across the world had an increase from 3.00 to 3.08 in 2009.

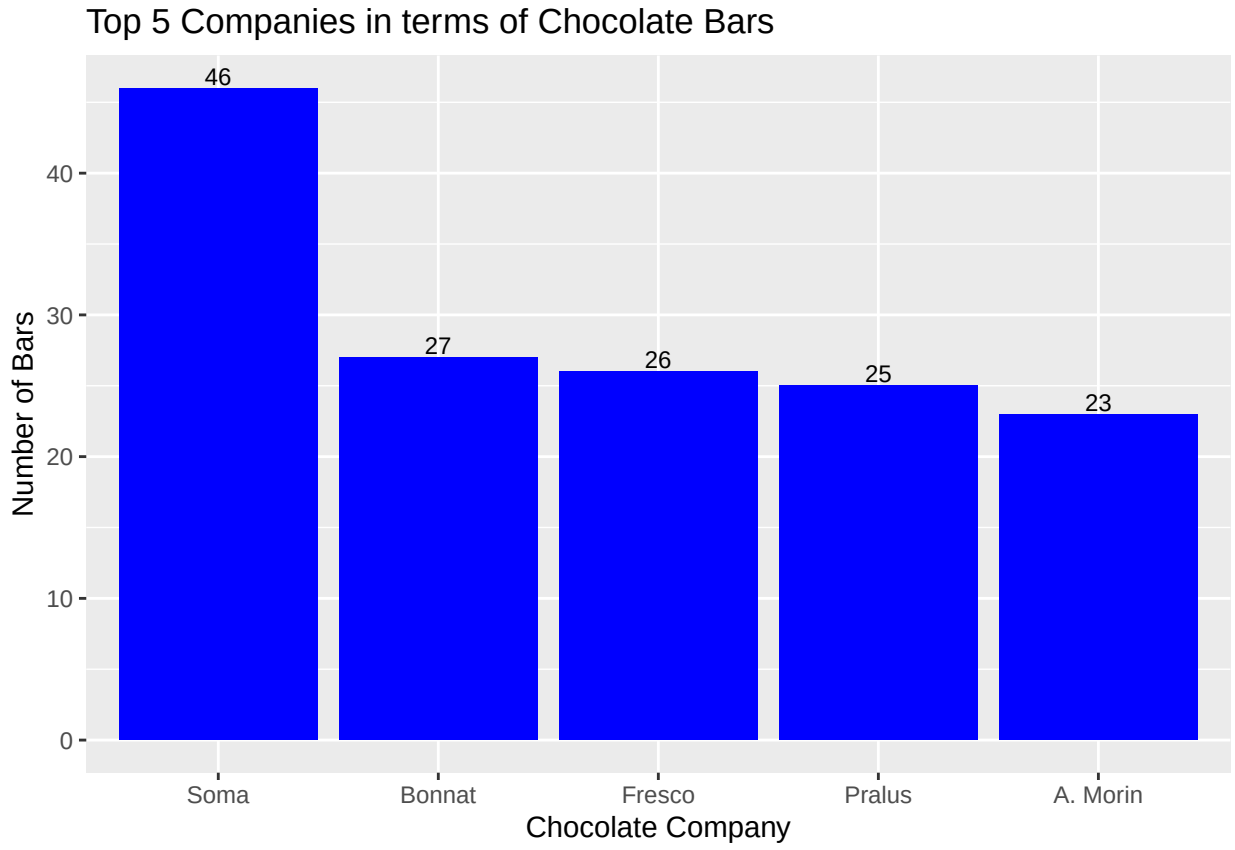
Analysing the best pattern for the Chocolate companies

```
# Top 5 Companies in terms of Chocolate Bars"
top5_company <- chocolate_project %>%
```

```
  count(Company_name, sort = TRUE) %>%
  slice(1:5)

top5_company
```

```
##   Company_name  n
## 1         Soma 46
## 2       Bonnat 27
## 3       Fresco 26
## 4       Pralus 25
## 5     A. Morin 23
```

```
ggplot(data=top5_company, aes(x= reorder(Company_name, -n),y=n))+
  geom_bar(stat="identity", fill="blue")+
  geom_text(aes(label = n), vjust = -0.2, size = 3,position = position_dodge(0.9))+
labs(x="Chocolate Company", y="Number of Bars", title="Top 5 Companies in terms of Chocolate Bars")
```



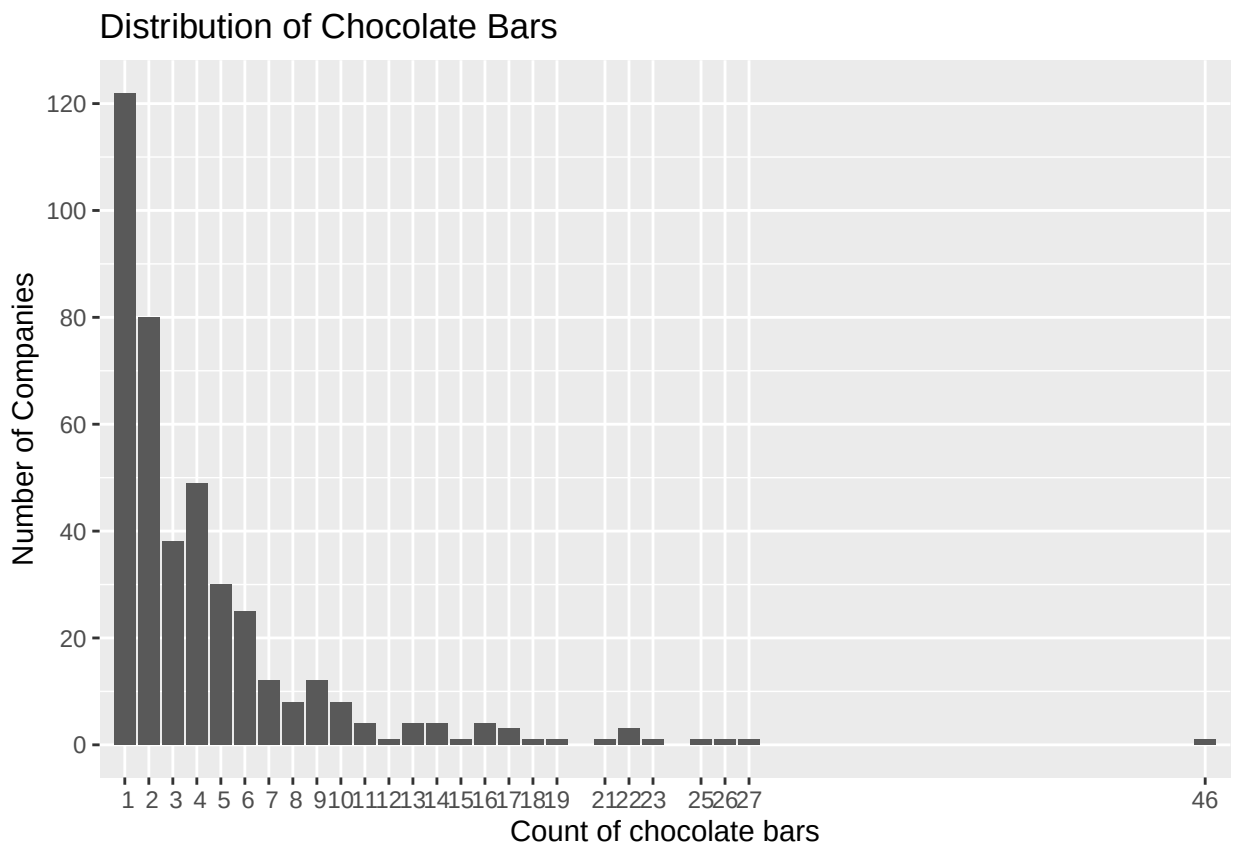
- Soma has the highest number of chocolate bars in this dataset with 46.

```
# Distribution of Chocolate Bars
company_count_chocolate_bars<-chocolate_project %>%
  group_by(Company_name) %>%
  count(Company_name, sort = TRUE)

company_count_chocolate_bars
```

```
## # A tibble: 416 x 2
## # Groups:   Company_name [416]
##    Company_name                    n
##    <chr>                       <int>
##  1 Soma                           46
##  2 Bonnat                         27
##  3 Fresco                         26
##  4 Pralus                         25
##  5 A. Morin                       23
##  6 Arete                          22
##  7 Domori                         22
##  8 Guittard                       22
##  9 Valrhona                       21
## 10 Hotel Chocolat (Coppeneur)     19
## # ... with 406 more rows
```

```
ggplot(data=company_count_chocolate_bars, aes(x= company_count_chocolate_bars$n))+
  geom_bar(stat="count")+
  scale_y_continuous(breaks = seq(0, 120, by = 20))+
  scale_x_discrete(limits = company_count_chocolate_bars$n, breaks = company_count_chocolate_bars$n)+
  labs(x="Count of chocolate bars", y="Number of Companies", title="Distribution of Chocolate Bars")
```

```
## Warning: Continuous limits supplied to discrete scale.
## Did you mean `limits = factor(...)` or `scale_*_continuous()`?
```

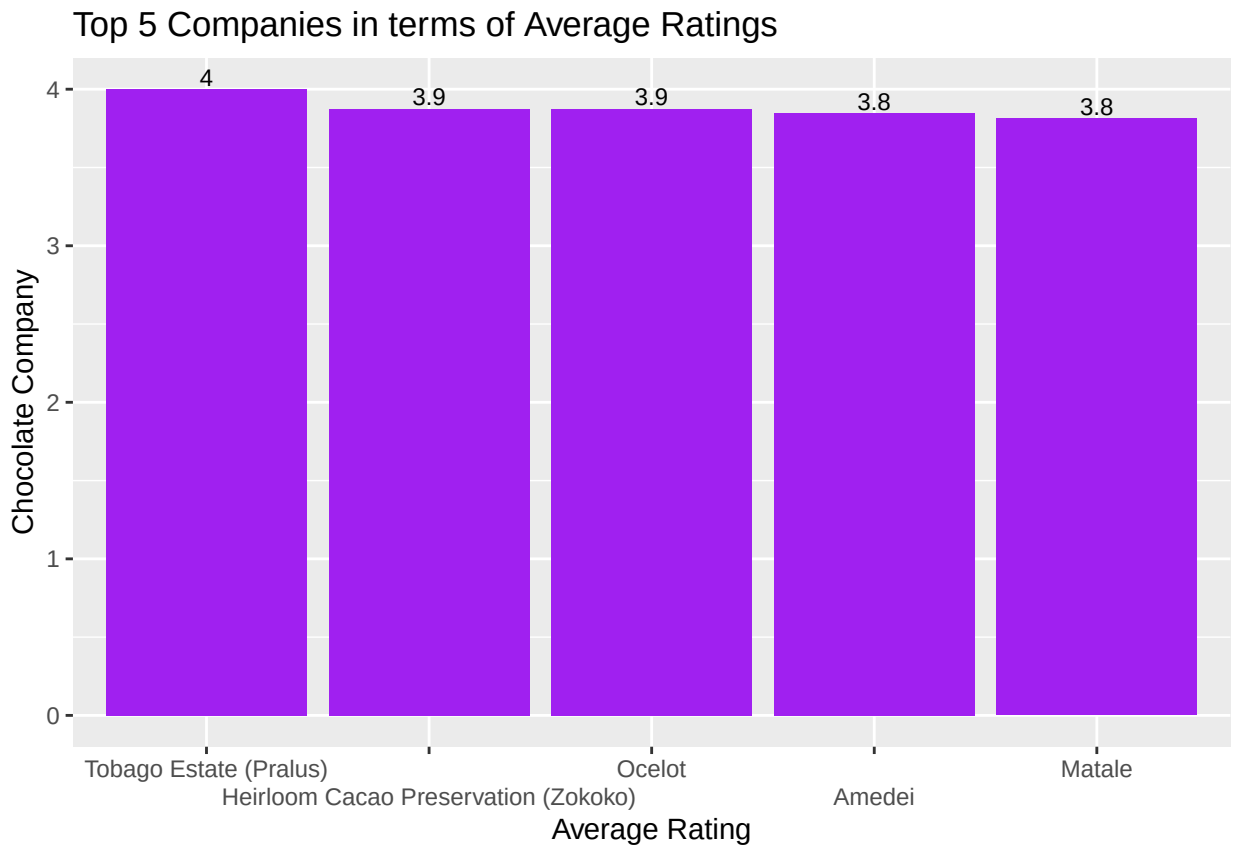


- 120+ companies have just one entry in this dataset.

```
# Top 5 companies in terms of average ratings

average_rating_company <- aggregate(Rating ~ Company_name, data = chocolate_project, FUN = mean)
top5_average_rating_company<- head(average_rating_company[order(-average_rating_company$Rating),],5)

top5_average_rating_company
```

```
##                              Company_name   Rating
## 385               Tobago Estate (Pralus) 4.000000
## 184 Heirloom Cacao Preservation (Zokoko) 3.875000
## 291                                Ocelot 3.875000
## 14                                 Amedei 3.846154
## 253                                Matale 3.812500
```

```
ggplot(data=top5_average_rating_company, aes(x= reorder(Company_name, -Rating),y=Rating))+
  geom_bar(stat="identity", fill="purple")+
  scale_x_discrete(guide = guide_axis(n.dodge = 2))+
  geom_text(aes(label = round(Rating, digits = 1)), vjust = -0.2, size = 3,position = position_dodge(0.
  labs(x="Average Rating", y="Chocolate Company", title="Top 5 Companies in terms of Average Ratings")
```

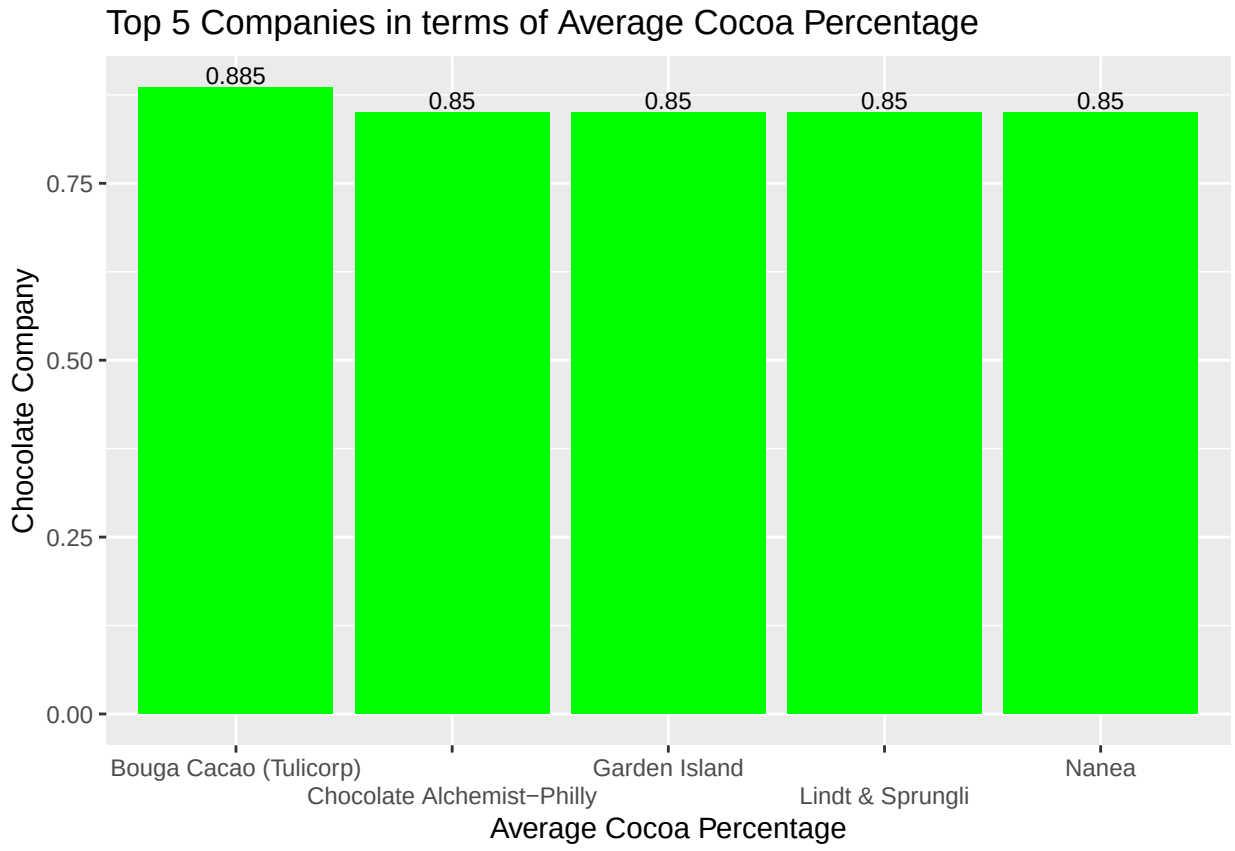


- These top 5 companies have very high ratings, however they have very low chocolate bars in the dataset.

```r
# Top 5 companies in terms of average Cocoa Percentage

average_cocoa_company <- aggregate(Cocoa_Percent ~ Company_name, data = chocolate_project, FUN = mean)
top5_average_cocoa_company<- head(average_cocoa_company[order(-average_cocoa_company$Cocoa_Percent),],5)

top5_average_cocoa_company
```

```
##                   Company_name Cocoa_Percent
## 43      Bouga Cacao (Tulicorp)         0.885
## 84  Chocolate Alchemist-Philly         0.850
## 162              Garden Island         0.850
## 227           Lindt & Sprungli         0.850
## 279                     Nanea         0.850
```

```r
ggplot(data=top5_average_cocoa_company, aes(x= reorder(Company_name, -Cocoa_Percent),y=Cocoa_Percent))+
  geom_bar(stat="identity", fill="green")+
  scale_x_discrete(guide = guide_axis(n.dodge = 2))+
  geom_text(aes(label = Cocoa_Percent), vjust = -0.2, size = 3,position = position_dodge(0.9))+
  labs(x="Average Cocoa Percentage", y="Chocolate Company", title="Top 5 Companies in terms of Average C
```



- All these companies produce chocolate with very high cocoa percentage (more than 80%)

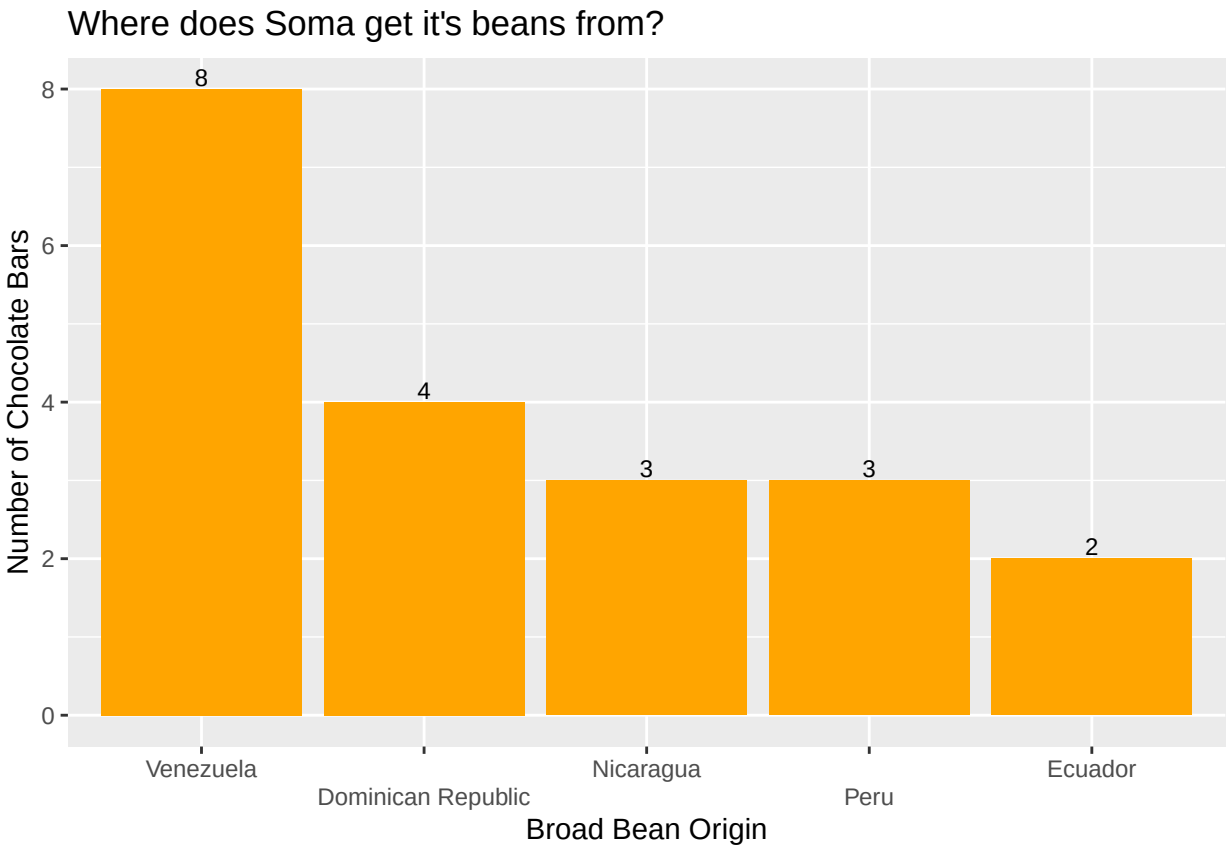
In terms of quantity Soma is the Largest Chocolate Bar Producer

```
# From where Soma get's their Beans?
# Select Company_name & Broad_Bean_origin from dataset

company_Bean_origin<-select(chocolate_project, Company_name, Broad_Bean_Origin)
```

```
# Filter Soma, groupby broad_bean_origin, sort and select top 5

top5_soma_bean_origin<- filter(company_Bean_origin, Company_name == "Soma") %>%
  group_by(Broad_Bean_Origin) %>%
  tally(sort = T) %>%
  arrange(desc(n)) %>% slice(1:5)

top5_soma_bean_origin
```

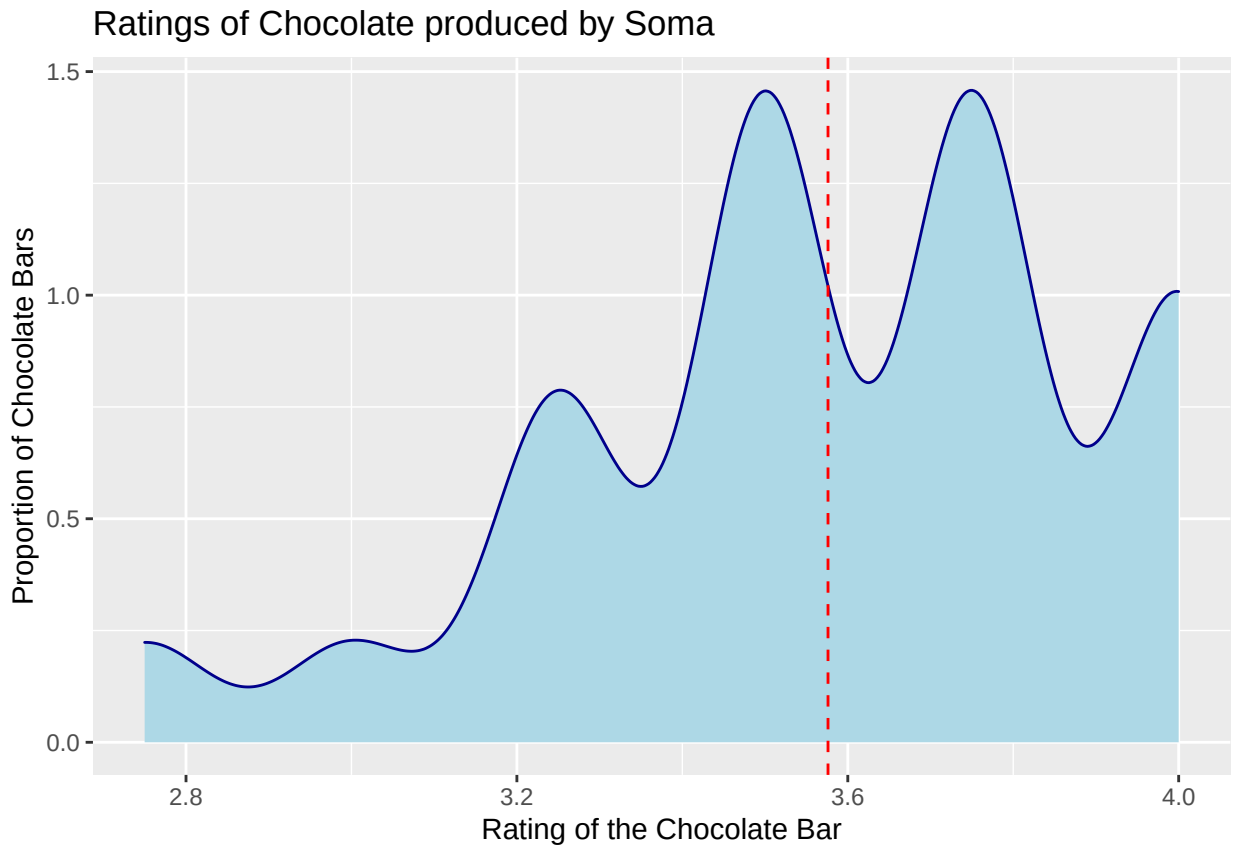
```
## # A tibble: 5 x 2
##   Broad_Bean_Origin      n
##   <chr>              <int>
## 1 Venezuela              8
## 2 Dominican Republic     4
## 3 Nicaragua              3
## 4 Peru                   3
## 5 Ecuador                2
```

```
ggplot(data=top5_soma_bean_origin, aes(x= reorder( Broad_Bean_Origin, -n),y=n))+
  geom_bar(stat="identity", fill="orange")+
  scale_x_discrete(guide = guide_axis(n.dodge = 2))+
  geom_text(aes(label = n), vjust = -0.2, size = 3,position = position_dodge(0.9))+
  labs(x="Broad Bean Origin", y="Number of Chocolate Bars", title="Where does Soma get it's beans from?
```



- Venezuela is the largest provider of Soma's beans.

```
# How are ratings of Chocolate bars by Soma ?

company_name_soma<-filter(chocolate_project, Company_name == "Soma")

ggplot(company_name_soma, aes(x=Rating)) +
  geom_density(color="darkblue", fill="lightblue")+
  geom_vline(aes(xintercept=mean(Rating)),
             color="red", linetype="dashed", size=.5)+
  labs(x="Rating of the Chocolate Bar", y="Proportion of Chocolate Bars", title="Ratings of Chocolate p
```

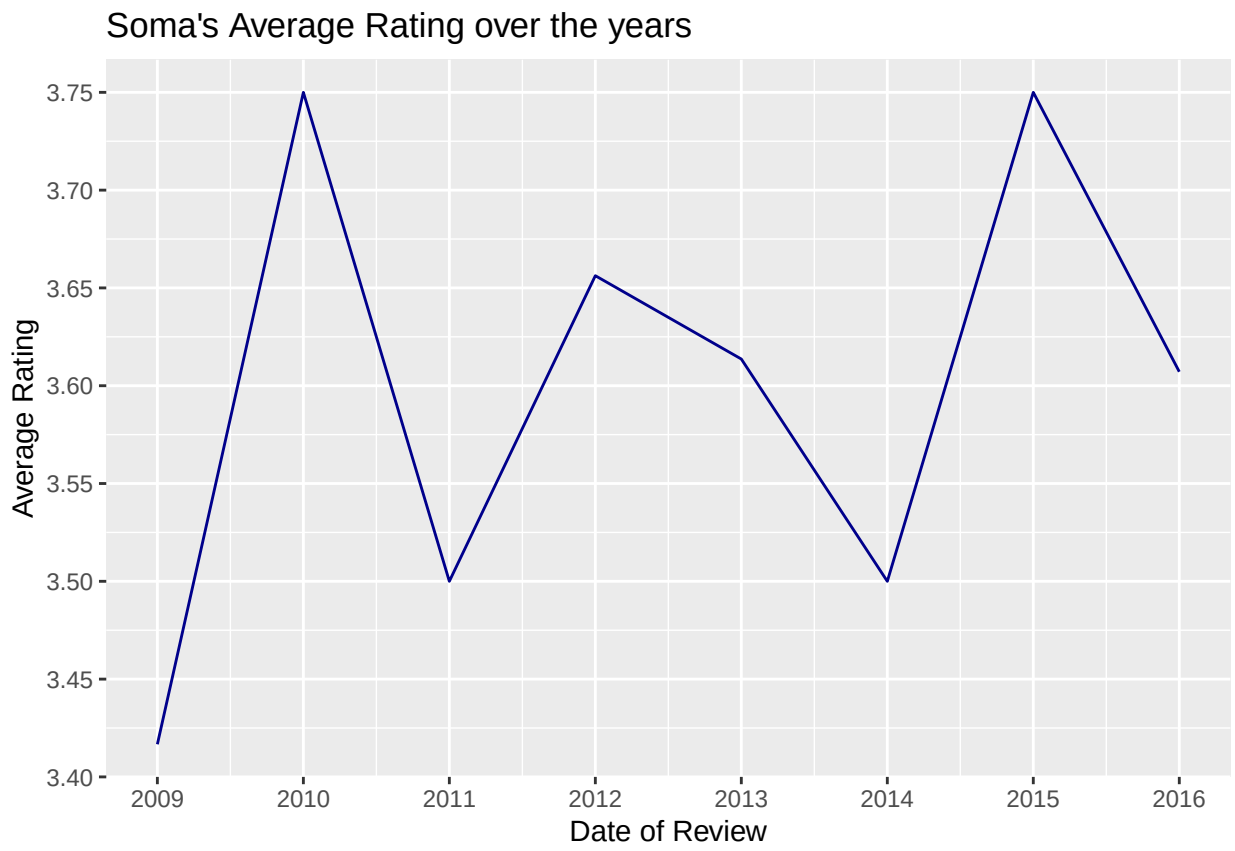


- As majority of chocolate bars produced by Soma has a rating above overall mean rating. So, they produce really some **good** chocolate

```
# Soma's performance over the years
soma_performance<- aggregate(Rating ~ Review_Date, data = company_name_soma, FUN = mean)

soma_performance
```

```
##   Review_Date   Rating
## 1        2009 3.416667
## 2        2010 3.750000
## 3        2011 3.500000
## 4        2012 3.656250
## 5        2013 3.613636
## 6        2014 3.500000
## 7        2015 3.750000
## 8        2016 3.607143
```

```
ggplot(data=soma_performance, mapping=aes(x=Review_Date, y=Rating))+
  geom_line(color="darkblue")+
  scale_x_continuous(breaks = seq(2009, 2016, by = 1))+
  scale_y_continuous(breaks = seq(3.40, 3.75, by = .05))+
  xlab("Date of Review")+
  ylab("Average Rating")+
  ggtitle("Soma's Average Rating over the years")
```



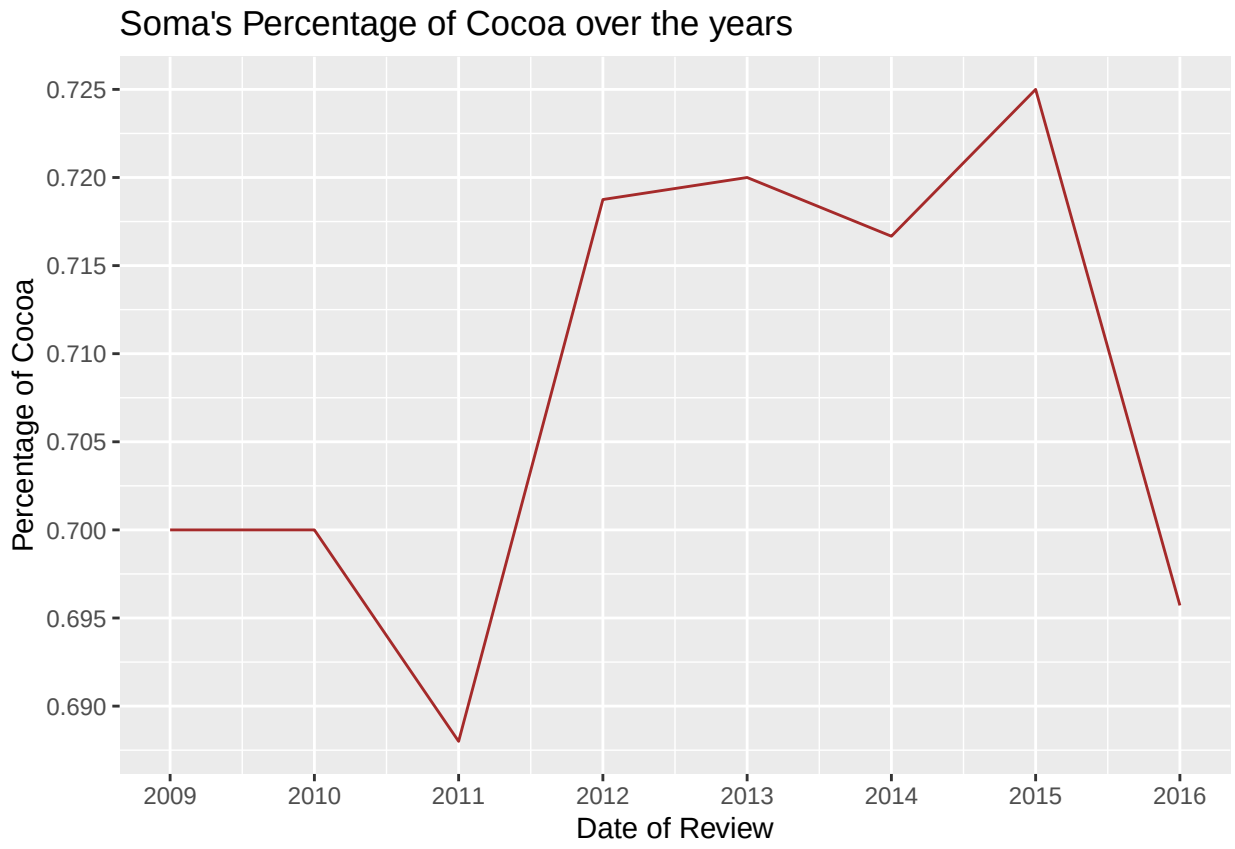
Analysing Soma's rating over period of time

- The worst average rating Soma ever got came in the year 2009 at 3.42, when it was first reviewed
- The highest average rating achieved came in 2010 at 3.75 (a significant rise from the previous year)
- Between 2012 and 2014, Soma's average rating saw a slump which revived after 3.75 was achieved in 2015 again; it again goes down to 3.61 in 2016

```
# Soma's Percentage of Cocoa over the years

soma_performance_percentage_cocoa<-aggregate(Cocoa_Percent ~ Review_Date, data = company_name_soma, FUN
soma_performance_percentage_cocoa
```

```
##   Review_Date Cocoa_Percent
## 1        2009     0.7000000
## 2        2010     0.7000000
## 3        2011     0.6880000
## 4        2012     0.7187500
## 5        2013     0.7200000
## 6        2014     0.7166667
## 7        2015     0.7250000
## 8        2016     0.6957143
```

```
ggplot(data=soma_performance_percentage_cocoa, mapping=aes(x=Review_Date, y=Cocoa_Percent))+
  geom_line(color="brown")+
  scale_x_continuous(breaks = seq(2009, 2016, by = 1))+
  scale_y_continuous(breaks = seq(.690, .725, by = .005))+
  xlab("Date of Review")+
  ylab("Percentage of Cocoa")+
  ggtitle("Soma's Percentage of Cocoa over the years")
```



Cocoa percent in Soma chocolates over Time

- First review in 2009 showed 70% cocoa
- The lowest percentage of cocoa in a Soma bar was in 2011 at 69%
- In 2015, Soma had the highest ever cocoa percent in their chocolate bar at 72.5%
- Latest review in 2016 discloses 69.6% cocoa in Soma's chocolate bars

Categorizing Chocolate based on Ratings

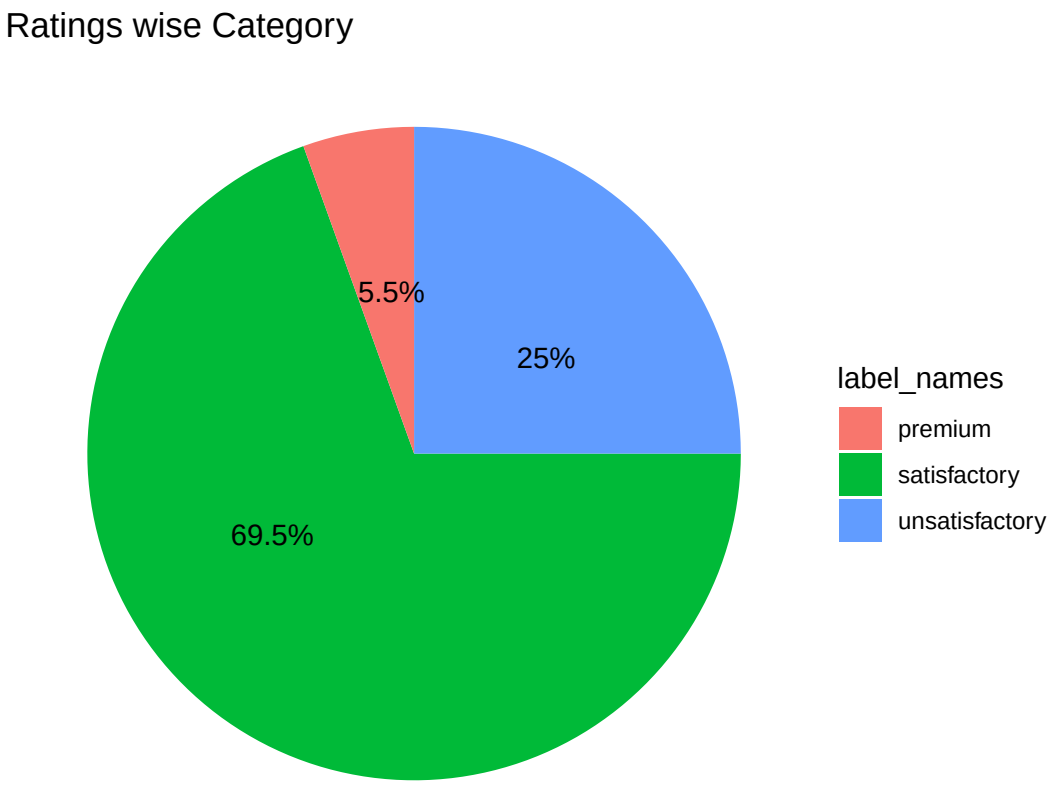
How many Chocolate bars are above or below 'Satisfactory levels'?

```
# Chocolate Bar levels
```

```r
rating_pie<-chocolate_project %>%
  select(Rating) %>%
  mutate(label_names = case_when(Rating < 3.0 ~ "unsatisfactory",
                                 Rating < 4.0 & Rating >= 3.0 ~ "satisfactory",
                                 Rating >= 4.0 ~ "premium")) %>%
  mutate(count = n()) %>%
  select(label_names,count)

rating_count<- count(rating_pie, label_names,sort = TRUE)

rating_count_percent<- rating_count %>%
  mutate(percent = n / sum(n) * 100) %>%
  mutate_if(is.numeric, round, 1)

rating_count_percent
```

```
##      label_names    n percent
## 1   satisfactory 1246    69.5
## 2 unsatisfactory  448    25.0
## 3        premium   99     5.5
```

```r
ggplot(rating_count_percent, aes(x="", y=percent, fill=label_names))+
  geom_bar(width = 1, stat = "identity")+
  coord_polar("y", start=0)+
  geom_text(aes(label = paste0(percent, "%")), position = position_stack( vjust = 0.6))+
  labs(title="Ratings wise Category")+
  theme_void()
```



- This pi chat affirms that premium chocolate is very rare, at only 5.5%.
- 69.5% of the chocolate bars in the study belong to 'Satisfactory'('premium' are also a part of this category).
- And, 25% of the chocolate bars that have been rated have ratings under 3.0.

Rating Distributions

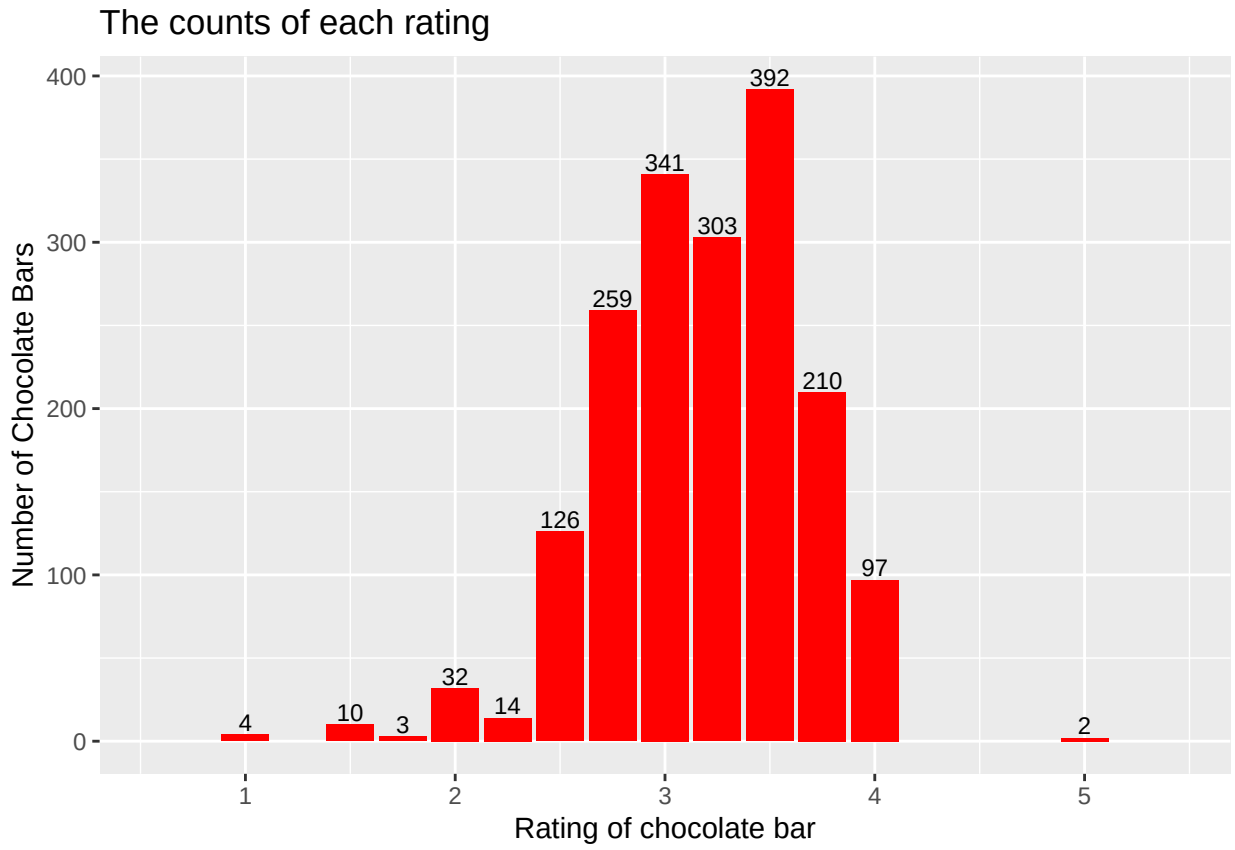
```
# The counts of each rating

rating_count<-count(chocolate_project, Rating,sort = TRUE)

rating_count
```

```
##    Rating   n
## 1    3.50 392
## 2    3.00 341
## 3    3.25 303
## 4    2.75 259
## 5    3.75 210
## 6    2.50 126
## 7    4.00  97
## 8    2.00  32
```

```
## 9    2.25  14
## 10   1.50  10
## 11   1.00   4
## 12   1.75   3
## 13   5.00   2
```

```
rating_count %>%
  ggplot(aes(x=Rating, y = n)) +
  geom_bar(stat = "identity", fill="red")+
  geom_text(aes(label = n), vjust = -0.2, size = 3,position = position_dodge(0.9))+
  labs(x="Rating of chocolate bar", y="Number of Chocolate Bars", title = "The counts of each rating")
```

```
## Warning: position_dodge requires non-overlapping x intervals
```



- Most bars have been rated at 3.5.
- Only 2 bars are rated at 5.0 (elite).

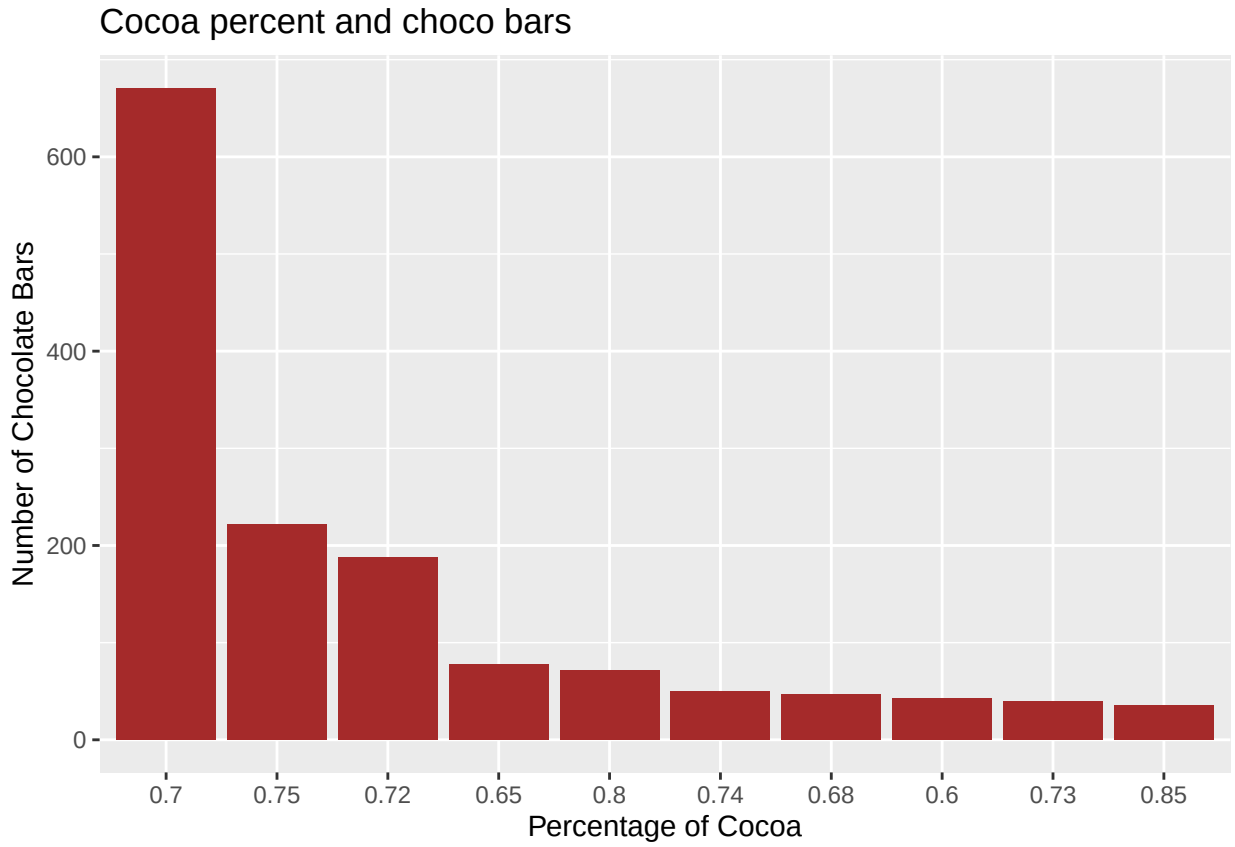
Number of Chocolate bars per percentage of Cocoa

```
# Cocoa percent and choco bars

cocoa_percentage_chocolate_bars<- count(chocolate_project, Cocoa_Percent,sort = TRUE) %>%
```

```
  slice(1:10)

cocoa_percentage_chocolate_bars
```

```
##    Cocoa_Percent   n
## 1           0.70 671
## 2           0.75 222
## 3           0.72 188
## 4           0.65  78
## 5           0.80  72
## 6           0.74  50
## 7           0.68  47
## 8           0.60  43
## 9           0.73  40
## 10          0.85  36
```

```
cocoa_percentage_chocolate_bars$Cocoa_Percent <- factor(cocoa_percentage_chocolate_bars$Cocoa_Percent,

ggplot(data=cocoa_percentage_chocolate_bars,aes(x= Cocoa_Percent, y=n))+
  geom_bar(stat="identity", fill="brown")+
  scale_x_discrete(limits=cocoa_percentage_chocolate_bars$Cocoa_Percent)+
  labs(x="Percentage of Cocoa", y="Number of Chocolate Bars", title="Cocoa percent and choco bars")
```



- The plot shows top 10 cocoa percentages in terms of number of chocolate bars.
- The vast majority of bars have 70% cocoa, followed by 75% and 72%.

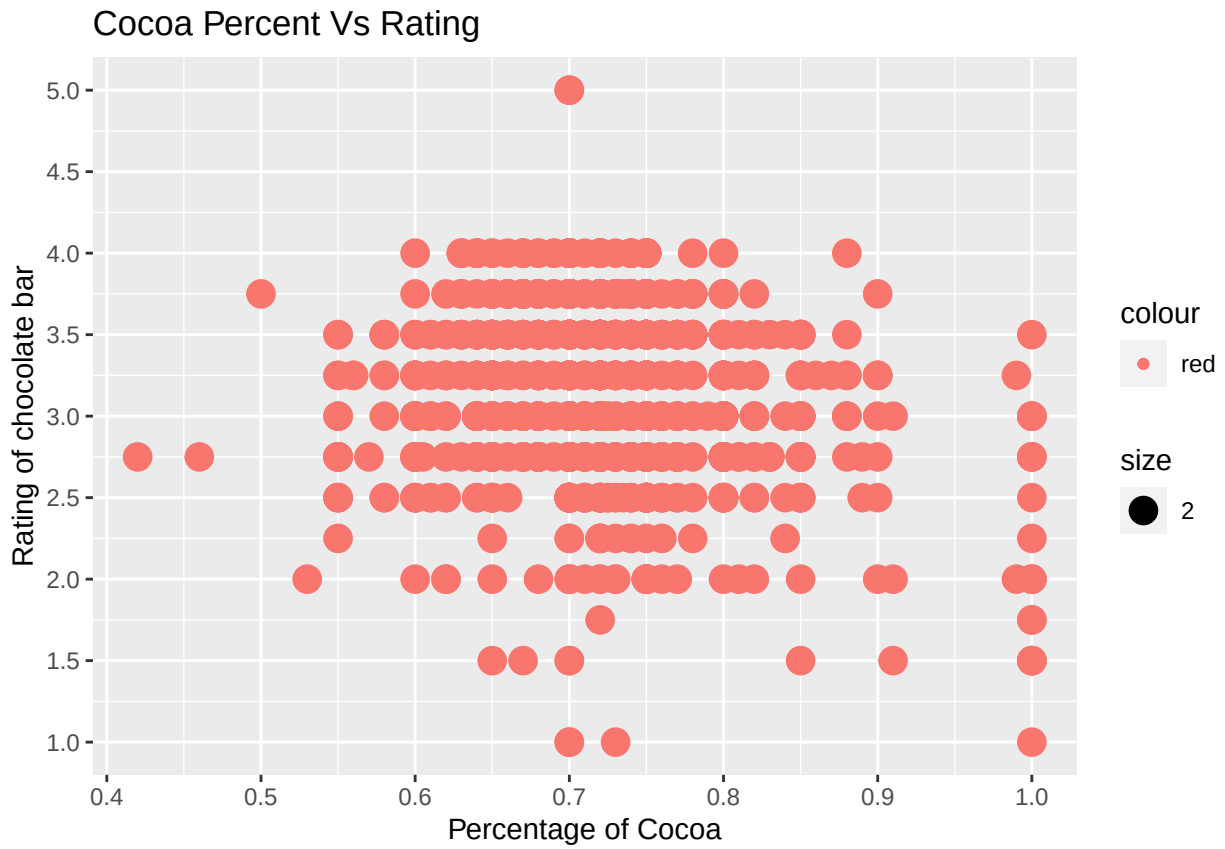
What is the relation between 'Cocoa Percent' and 'Rating'?

- Is there any correlation between Cocoa Percent and Rating of the bar?
- If it is, is that a positive correlation or a negative one?
- Can we predict rating of a bar given it's cocoa percentage?

```
# Cocoa Percent Vs Rating

cor(chocolate_project$Cocoa_Percent,chocolate_project$Rating)
```

```
## [1] -0.1647583
```

```
ggplot(chocolate_project, aes(x=Cocoa_Percent, y=Rating)) +
  geom_point(aes(colour = "red", size=2))+
  scale_x_continuous(breaks = seq(0.4, 1.0, by = 0.1))+
  scale_y_continuous(breaks = seq(1, 5, by = .5))+
  xlab("Percentage of Cocoa")+
  ylab("Rating of chocolate bar")+
  ggtitle("Cocoa Percent Vs Rating")
```



From the Scatterplot above, we conclude that:

- No evident correlation. A numerical correlation gives a weak negative correlation coefficient of -0.16
- The density of the graph is highest between 65% and 80% of cocoa Chocolate bars.

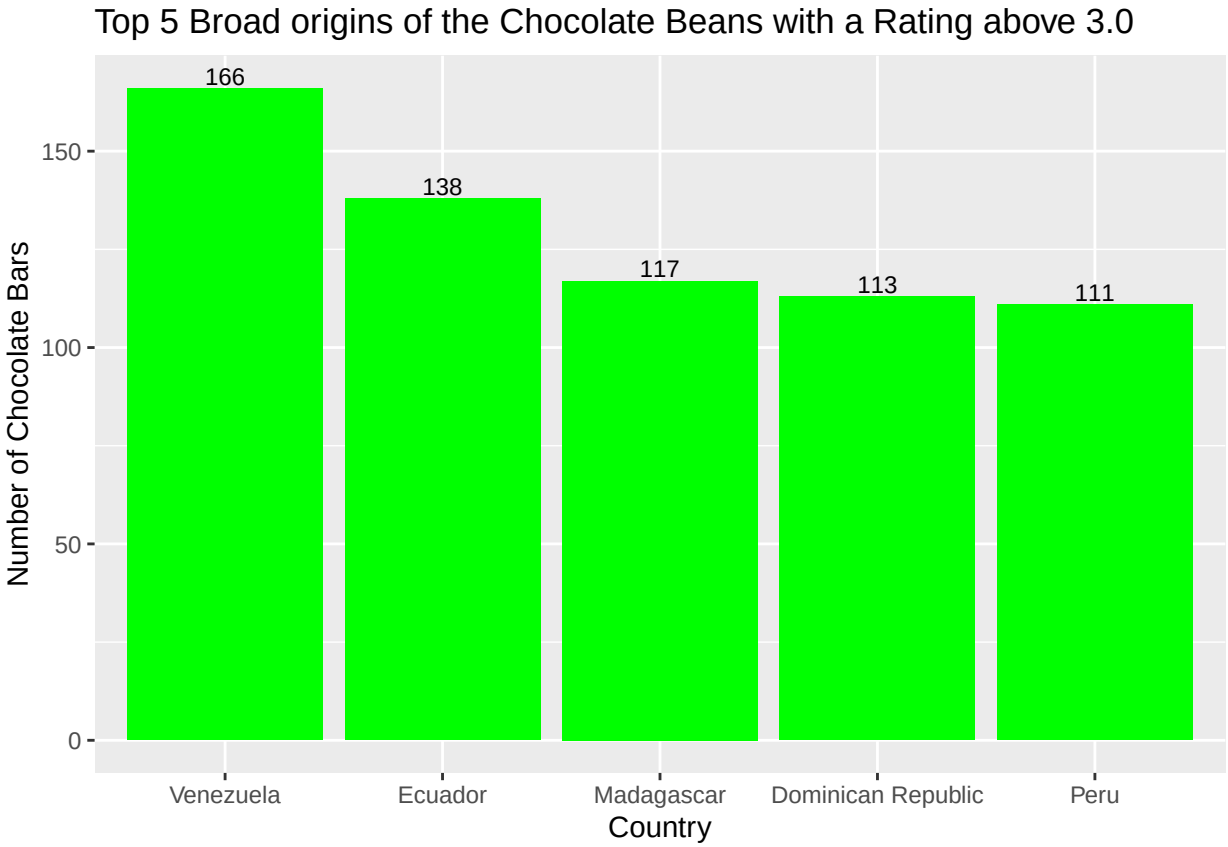
- With low cocoa percentage(less than 50%) and high cocoa percentage(above 90%) are less in number.
- The most important fact is that most of these chocolate bars have a rating of less than 3,i.e they have been 'Unsatisfactory'
- Seems like people do not prefer very low or very high cocoa percentages in their chocolate!
- From the scatter plot above, we can infer that it would not be a good idea to guess a chocolate's rating based on its Cocoa Percentage.

Where are the Best Cocoa Beans grown?

```
# Top 5 countries producing most number of satisfactory rating chocolate Beans
satisfactory_rating_bean_origin<- filter(chocolate_project, Rating >= 3) %>%
  group_by(Broad_Bean_Origin) %>%
  tally(sort = T) %>%
  arrange(desc(n)) %>% slice(1:5)

satisfactory_rating_bean_origin
```

```
## # A tibble: 5 x 2
##   Broad_Bean_Origin      n
##   <chr>              <int>
## 1 Venezuela            166
## 2 Ecuador              138
## 3 Madagascar           117
## 4 Dominican Republic   113
## 5 Peru                 111
```

```
satisfactory_rating_bean_origin %>%
  ggplot(aes(reorder(x=Broad_Bean_Origin,-n),y=n))+
  geom_bar(stat = "identity", fill="green")+
  geom_text(aes(label = n), vjust = -0.2, size = 3,position = position_dodge(0.9))+
  labs(x="Country", y="Number of Chocolate Bars", title = "Top 5 Broad origins of the Chocolate Beans w
```

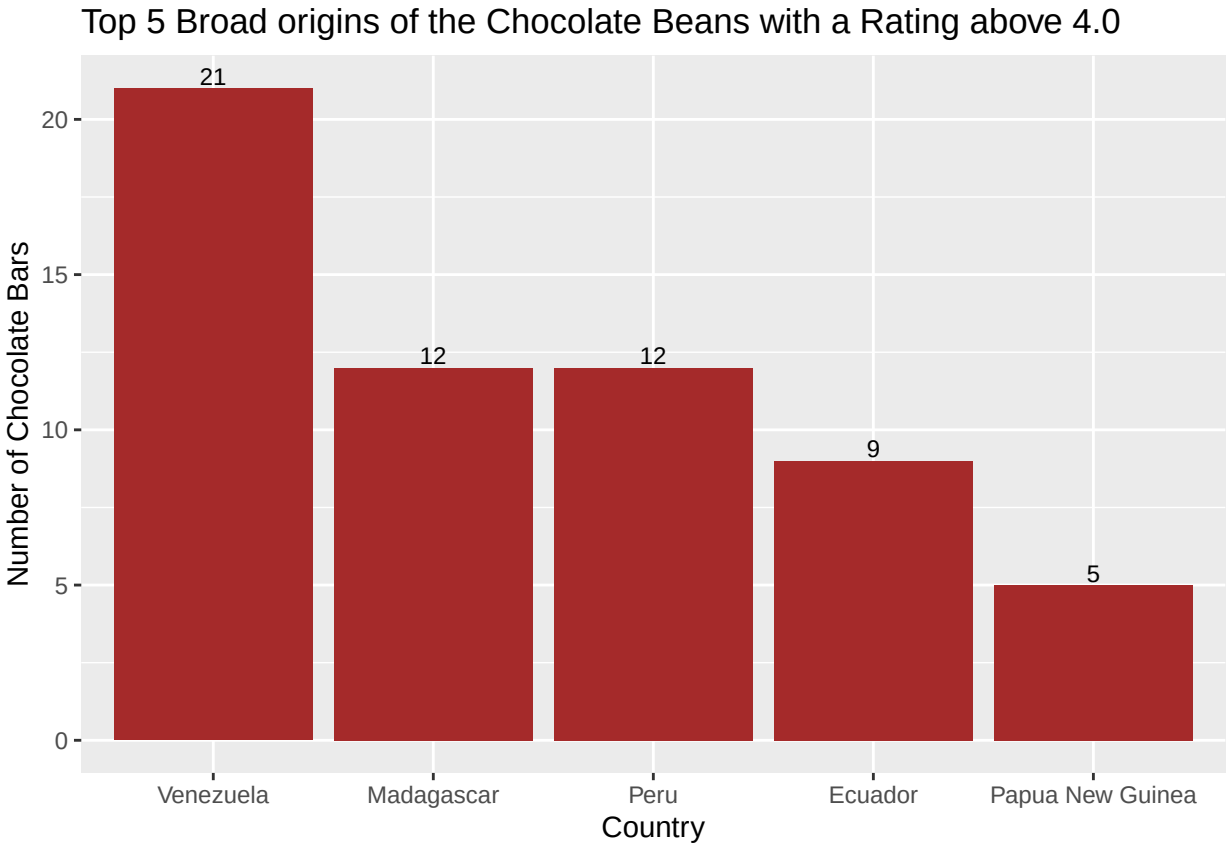


- Venezuela has the largest number of chocolate bars rated above 3.0

```
# Top 5 countries producing most number of best rating chocolate Beans

best_rating_bean_origin<- filter(chocolate_project, Rating >= 4) %>%
  group_by(Broad_Bean_Origin) %>%
  tally(sort = T) %>%
  arrange(desc(n)) %>% slice(1:5)

best_rating_bean_origin
```

```
## # A tibble: 5 x 2
##   Broad_Bean_Origin     n
##   <chr>             <int>
## 1 Venezuela            21
## 2 Madagascar           12
## 3 Peru                 12
## 4 Ecuador               9
## 5 Papua New Guinea      5
```

```
best_rating_bean_origin %>%
  ggplot(aes(reorder(x=Broad_Bean_Origin,-n),y=n))+
  geom_bar(stat = "identity", fill="brown")+
  geom_text(aes(label = n), vjust = -0.2, size = 3,position = position_dodge(0.9))+
  labs(x="Country", y="Number of Chocolate Bars", title = "Top 5 Broad origins of the Chocolate Beans wi
```



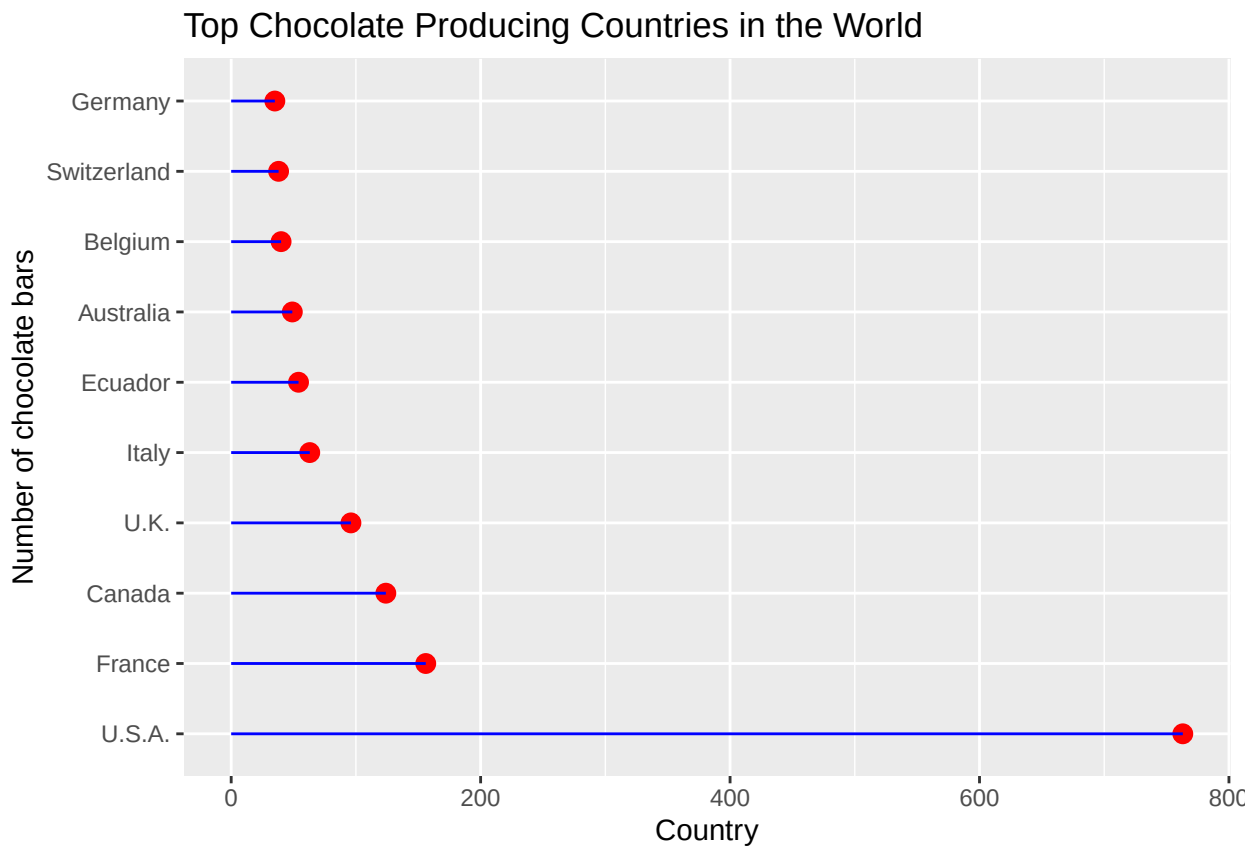
- So, we conclude that the best cocoa beans are also grown in Venezuela.
- There are 21 bars from Venezuela that have a rating of 4 and above.

Analysis of the Producing Countries

```
# Top Chocolate Producing Countries in the World

top10_chocolate_producing_Country<-count(chocolate_project, Company_Location,sort = TRUE) %>%
  slice(1:10)
top10_chocolate_producing_Country
```

```
##    Company_Location   n
## 1           U.S.A. 763
## 2            France 156
## 3            Canada 124
## 4              U.K.  96
## 5             Italy  63
## 6           Ecuador  54
## 7         Australia  49
## 8           Belgium  40
## 9       Switzerland  38
## 10          Germany  35
```

```
top10_chocolate_producing_Country %>%
  ggplot(aes(x=reorder(Company_Location,-n),y=n)) +
  geom_point(size = 3, colour = "red") +
  geom_segment( aes(x=Company_Location, xend=Company_Location, y=0, yend=n), colour = "blue")+
  coord_flip()+
  labs(x= "Number of chocolate bars", y="Country", title = "Top Chocolate Producing Countries in the Wo
```



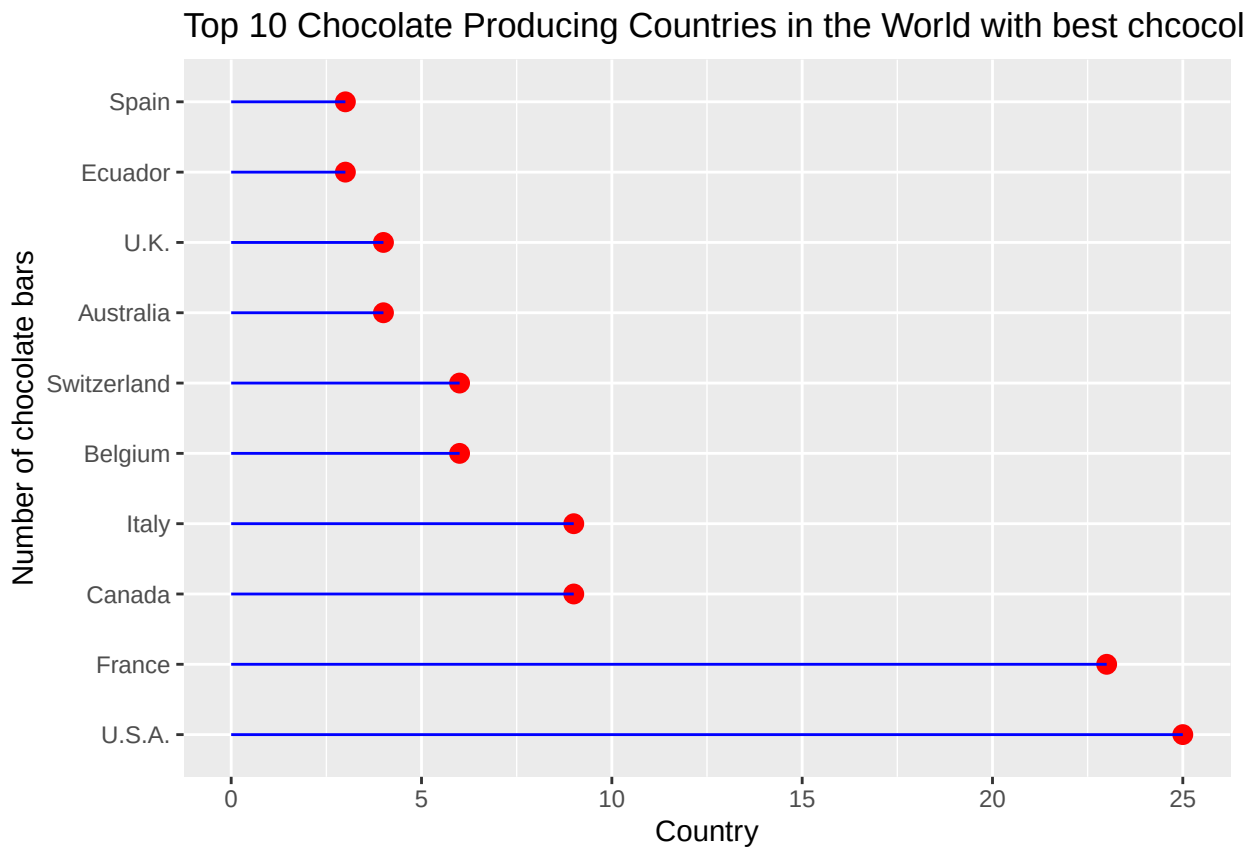
- U.S.A produces much more chocolate companies than any other country has according to this data.

```
# Top Chocolate Producing Countries in the World (Ratings above 4.0)

top10_best_rating_company_location<- filter(chocolate_project, Rating >= 4) %>%
  group_by(Company_Location) %>%
  tally(sort = T) %>%
  arrange(desc(n)) %>% slice(1:10)

top10_best_rating_company_location
```

```
## # A tibble: 10 x 2
##    Company_Location     n
##    <chr>            <int>
##  1 U.S.A.              25
##  2 France              23
##  3 Canada               9
```

```
##  4 Italy                9
##  5 Belgium              6
##  6 Switzerland          6
##  7 Australia            4
##  8 U.K.                 4
##  9 Ecuador              3
## 10 Spain                3
```

```
top10_best_rating_company_location %>%
  ggplot(aes(x=reorder(Company_Location,-n),y=n)) +
  geom_point(size = 3, colour = "red") +
  geom_segment( aes(x=Company_Location, xend=Company_Location, y=0, yend=n), colour = "blue")+
  coord_flip()+
  labs(x= "Number of chocolate bars", y="Country", title = "Top 10 Chocolate Producing Countries in the
```



- {'U.S.A.': 25, 'France': 23, 'Canada': 9, 'Italy': 9, 'Belgium': 6, 'Switzerland': 6, 'Australia': 4, 'U.K.': 4, 'Ecuador': 3, 'Spain': 3}
- USA produces the highest number of 4 and above rated choco bars