flavors of cacao

Shir ohayon 322590274 & Ye'ela granot 209133107 & Reut lev 207385741

2023-04-04

In this exercise we perform EDA on the chocolate bar rating Data-set.
First we load the data and explore its structure.

```
cacao <-read.csv("flavors_of_cacao.csv")
str(cacao)
```

```
## 'data.frame':    1795 obs. of  9 variables:
##  $ Company...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  " " " " " " " " " " ...
##  $ Broad.Bean.Origin               : chr  "Sao Tome" "Togo" "Togo" "Togo" ...
```

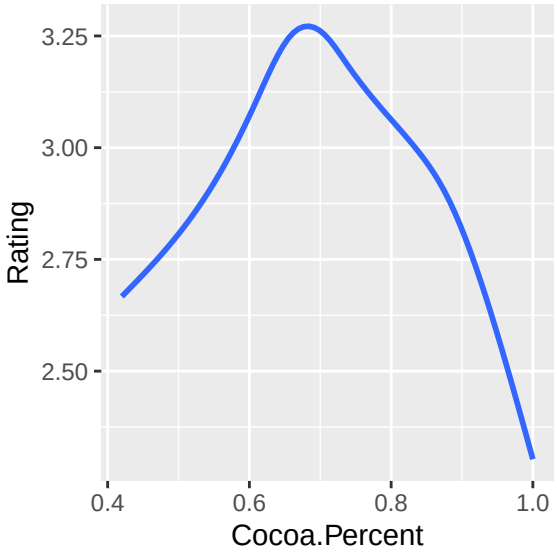
For cleaning the data we choose not to use features selection since most of our data is not numeric. We would like to convert the percentages in the 'cacao percent' field to numerical values in order to be able to perform analyzes and comparisons on the data. In addition we would like to clean the 'REF' column, since it's not informative enough.

```
# transform values to numeric
cacao$Cocoa.Percent <- as.numeric(gsub('%', '', cacao$Cocoa.Percent)) / 100
#remove REF column
caClean <- subset(cacao,select = -c(REF))
```

We would like to analyze the data and learn about what are the factors that affect the most on the chocolate rating, in order to find the features of the best chocolate. The first question that interests us is what's the correlation between rating and cocoa percent.

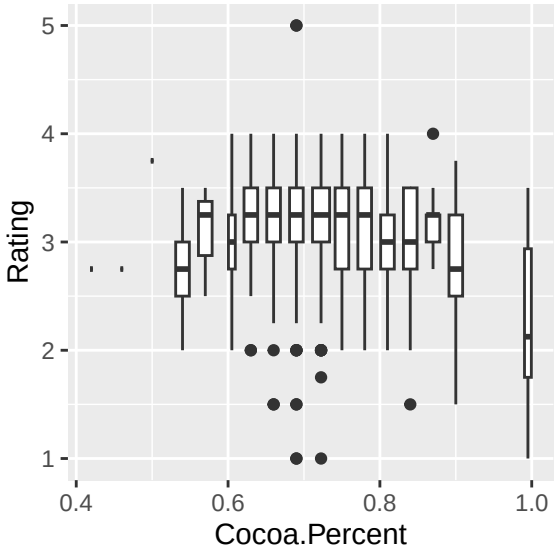
```
library(ggplot2)
# plotting a scatter plot (with smoothing curve)
ggplot(caClean, aes(x=Cocoa.Percent, y=Rating)) + geom_smooth(se=FALSE)
```

```
## `geom_smooth()` using method = 'gam' and formula = 'y ~ s(x, bs = "cs")'
```



As shown in the above graph the chocolate with the highest average rating has 70% cocoa percent. Since average is statistical measure that can be biased by edge values, we will also introduce boxplot graph. This kind of graph can display the data distribution and isolate outliers.

```
# creating boxplot
ggplot(data = caClean, mapping = aes(x = Cocoa.Percent, y = Rating)) +
geom_boxplot(mapping = aes(group = cut_width(Cocoa.Percent, 0.03)))
```



As we saw, the previous graph was highly affected by outliers. If we look on the distribution of the data by the median, the range that gets the highest rating is from 0.63-0.77 cocoa percent. the cocoa percent that gets the lowest rating is 100% cocoa, standing on rating of about 2.25. Therefore, according to the findings from the above data, we infer that in order to make a good chocolate, its cocoa percentage should be between 0.63-0.77%.

Now, we want to see what is the correlation between Broad Bean feature and the chocolate Rating - Which are the countries that grow the best cocoa beans? Since there are lot of origins, we will look for those who appears only a few times.

```r
# Load required libraries
suppressPackageStartupMessages(library(dplyr))

# Calculate the average rating
avg_ratings1 <- caClean %>%
  group_by(Broad.Bean.Origin) %>%
  summarize(mean_rating = mean(Rating, na.rm = TRUE))

highest_ratings <- filter(avg_ratings1, mean_rating == max(mean_rating))
print(highest_ratings)
```

```
## # A tibble: 6 x 2
##   Broad.Bean.Origin            mean_rating
##   <chr>                              <dbl>
## 1 Dom. Rep., Madagascar                  4
## 2 Gre., PNG, Haw., Haiti, Mad            4
## 3 Guat., D.R., Peru, Mad., PNG           4
## 4 Peru, Dom. Rep                         4
## 5 Ven, Bolivia, D.R.                     4
## 6 Venezuela, Java                        4
```

The top origin countries are - Madagascar, Haiti, Peru, Dominican Republic, Bolivia, Venezuela.. Apparently these results are not reliable and are affected by noise, whether it is due to an inaccurate formulation of the name of the area or the value that appeareed only once in the data, so we will filter only the 5 most common origin countries and explore their rating.

```r
#find the top 5 values
top_5_values <- caClean %>%
  count(Broad.Bean.Origin, sort = TRUE) %>%
  slice_head(n = 5) %>%
  pull(Broad.Bean.Origin)
# Calculate the average rating for each of the top 5 values
avg_ratings2 <- caClean %>%
  filter(Broad.Bean.Origin %in% top_5_values) %>%
  group_by(Broad.Bean.Origin) %>%
  summarize(mean_rating = mean(Rating, na.rm = TRUE))%>%
  arrange(desc(mean_rating))

# Print the top 5 values and their corresponding mean ratings
print(avg_ratings2)
```

```
## # A tibble: 5 x 2
##   Broad.Bean.Origin  mean_rating
##   <chr>                    <dbl>
## 1 Madagascar                3.27
## 2 Venezuela                 3.25
## 3 Dominican Republic        3.21
## 4 Peru                      3.14
## 5 Ecuador                   3.13
```

As it can be seen, the top 5 origions of cocoa are Madagascar, Venezuela, Dominican Republic, Peru, Ecuador. They all have similiar rating average of about 3.13-3.26. It is important to note that certain values in the

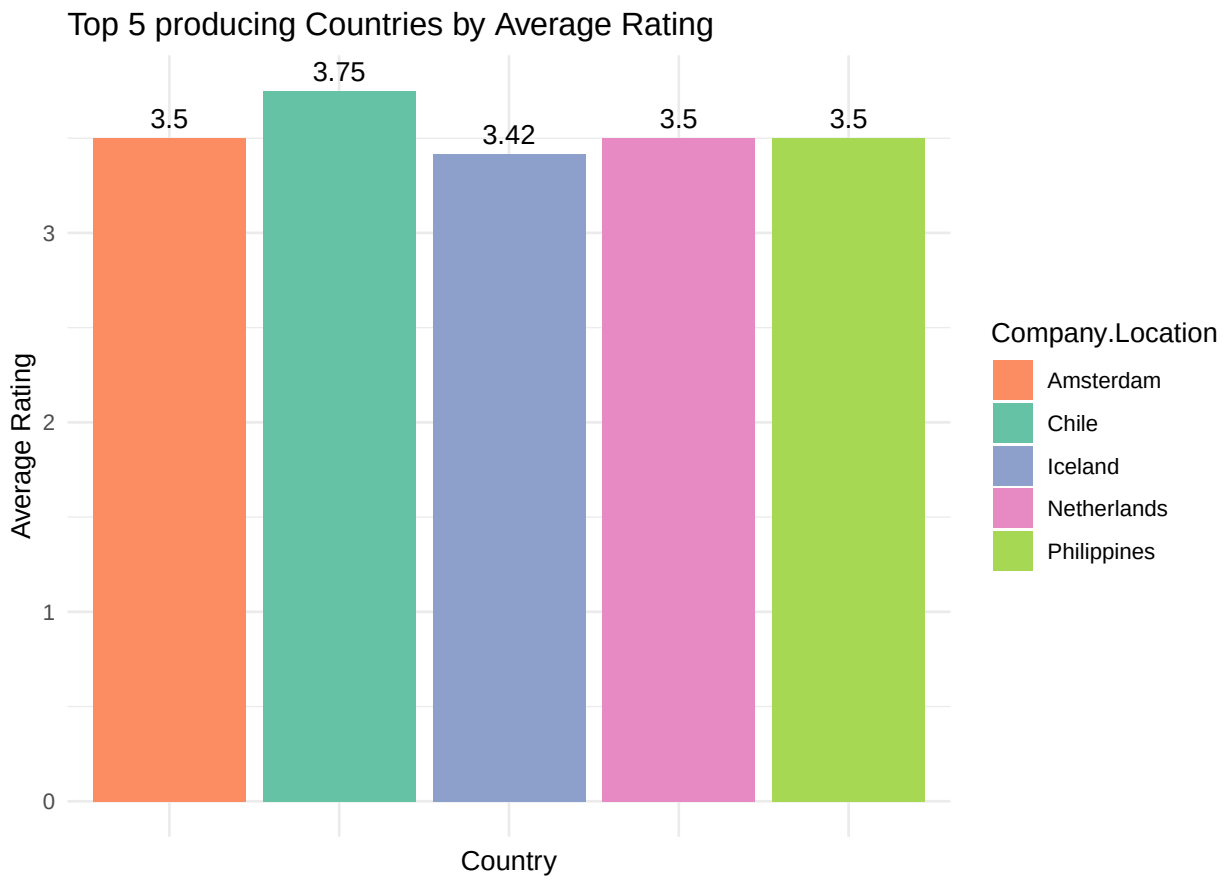
data may not have been taken into account due to anomalies in their name and we decided to ignore such noise.

In this context, we would like to find what is the correlation between company location feature and the chocolate Rating - Which are the countries that produce the best chocolate bars?

```
# group data by country and calculate average rating
df_summary <- caClean %>%
  group_by(Company.Location) %>%
  summarize(avg_rating = mean(Rating))

# rank countries by average rating
df_ranked <- df_summary %>%
  arrange(desc(avg_rating))

top_5 <- df_ranked$Company.Location[1:5]
cat(paste0("The top 5 countries by rating are: ", paste(top_5, collapse = ", ")))
```

```
## The top 5 countries by rating are: Chile, Amsterdam, Netherlands, Philippines, Iceland
```

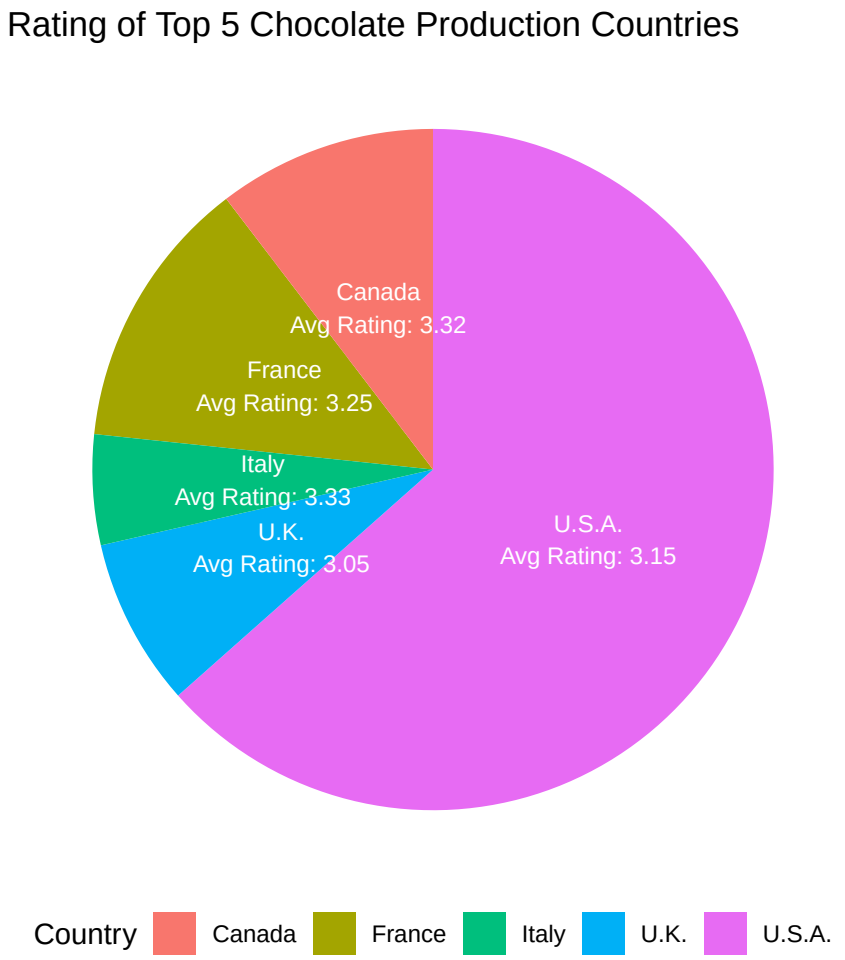
```
# plot bar chart
ggplot(df_ranked[1:5,], aes(x = Company.Location, y = avg_rating, fill = Company.Location)) +
  geom_bar(stat = "identity") +
  scale_fill_manual(values = c("#FC8D62", "#66C2A5", "#8DA0CB", "#E78AC3", "#A6D854"))+
  geom_text(aes(label = round(avg_rating, 2)), vjust = -0.5) +
  labs(title = "Top 5 producing Countries by Average Rating") +
  xlab("Country") +
  ylab("Average Rating") +
  theme_minimal()+
  theme(axis.text.x = element_blank())
```



The top 5 producing countries by rating are: Chile, Amsterdam, Netherlands, Philippines, Iceland. Again, We would like to look only at countries that have passed a certain threshold, so we will take the 5 countries that appear the most, that is, that produce the largest amount of chocolate.

```
# load necessary packages
library(dplyr)
library(ggplot2)

# group the data by country and calculate the average rating for each country
country_ratings <- caClean %>%
  group_by(Company.Location) %>%
  summarize(avg_rating = mean(Rating), count = n()) %>%
  arrange(desc(count)) %>%
  top_n(5, count)

# add labels with the average rating for each country
ggplot(country_ratings, aes(x = "", y = count, fill = Company.Location)) +
  geom_bar(stat = "identity", width = 1) +
  coord_polar(theta = "y") +
  labs(title = "Rating of Top 5 Chocolate Production Countries",
       x = NULL,
       y = NULL,
       fill = "Country") +
  theme_void() +
  theme(legend.position = "bottom") +
```

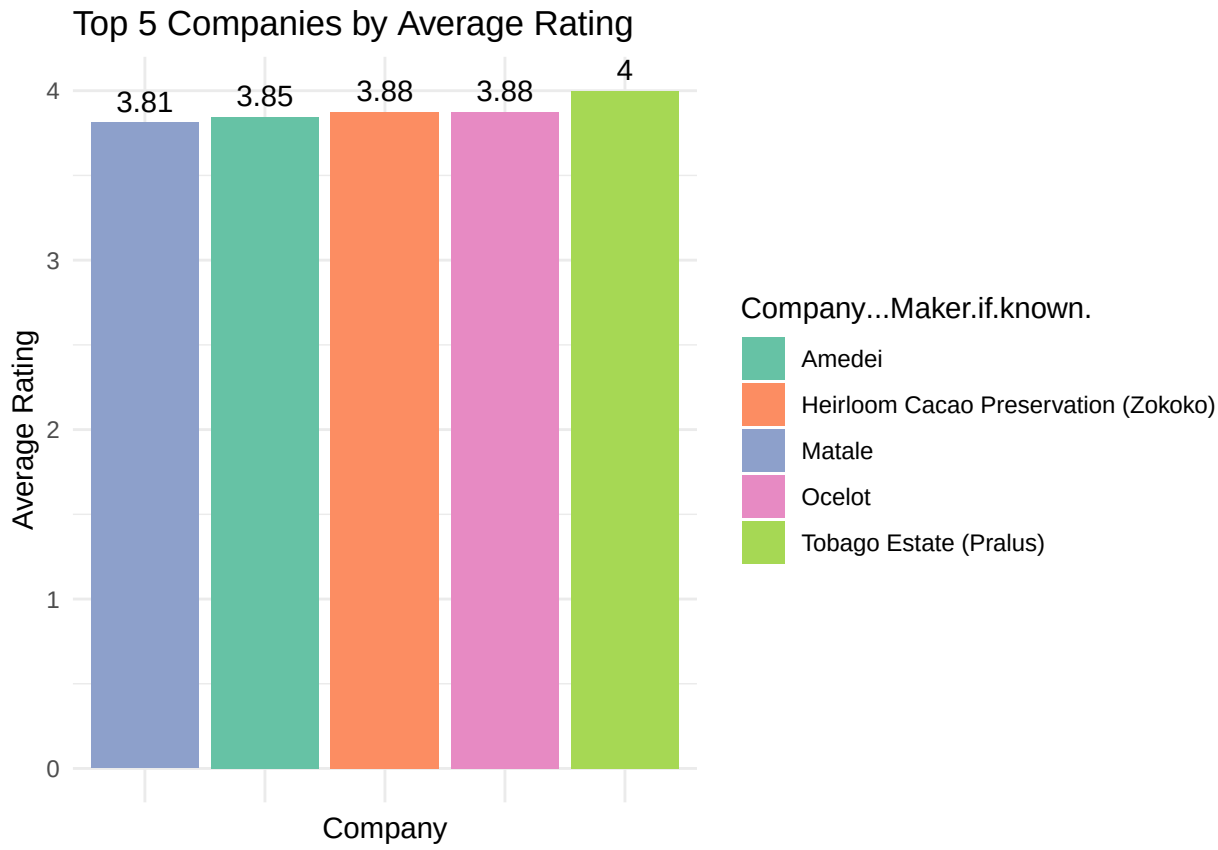
```
  geom_text(aes(label = paste0(Company.Location, "\nAvg Rating: ", round(avg_rating, 2))),
            position = position_stack(vjust = 0.5),
            size = 3,
            color = "white")
```



Next, we want to know what is the correlation between company maker feature and the chocolate Rating - Which are the companies that has the best chocolates bar?

```
# Compute the mean ratings for each company
mean_ratings <- aggregate(Rating ~ Company...Maker.if.known., caClean, mean)

# Sort the companies by mean rating
mean_ratings <- mean_ratings[order(mean_ratings$Rating, decreasing = TRUE), ]

# Select the top 5 companies
top_companies <- mean_ratings[1:5, ]

# Create a bar plot with labels and color
ggplot(data = top_companies, aes(x = reorder(Company...Maker.if.known., Rating), y = Rating, fill = Comp
  geom_bar(stat = "identity") +
  labs(x = "Company", y = "Average Rating") +
  ggtitle("Top 5 Companies by Average Rating") +
  theme_minimal() +
```

```
  scale_fill_brewer(palette = "Set2") +
  geom_text(aes(label = round(Rating, 2)), vjust = -0.5) +
  theme(axis.text.x = element_blank(), axis.ticks.x = element_blank())
```



As you can see, the ranking of the companies with the highest rating average is relatively close, so the information of the rating average is not significant enough.

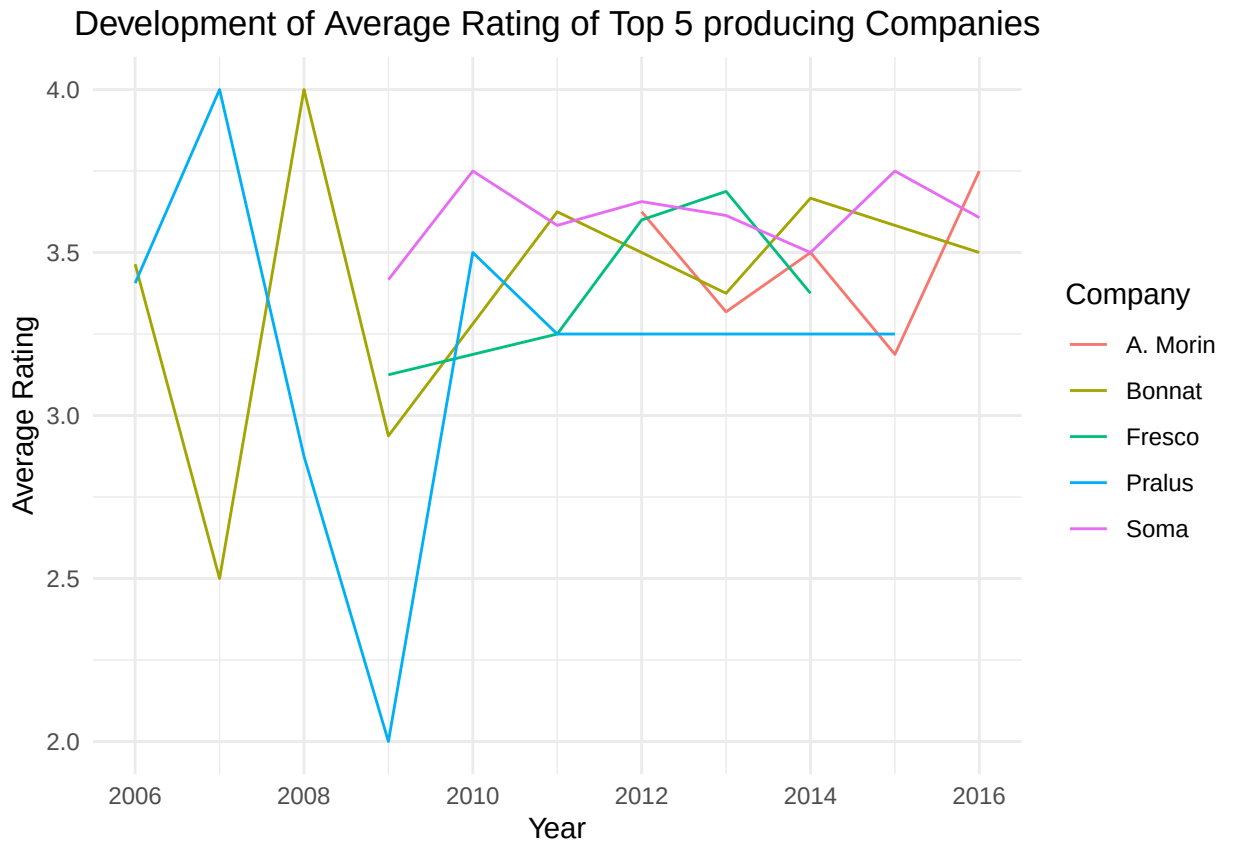
Another interesting question we would like to see is how the average rating values, of the companies who produce the highest number of products, developed over the years.

```
# extract the top 5 companies that appear the most in the "Company...Maker.if.known." column
top_companies <- caClean %>%
  count(`Company...Maker.if.known.`) %>%
  arrange(desc(n)) %>%
  head(5) %>%
  select(`Company...Maker.if.known.`)

# filter the dataset to include only the reviews for the top 5 companies
caClean_top <- caClean %>%
  filter(`Company...Maker.if.known.` %in% top_companies$`Company...Maker.if.known.`)

# extract the average rating of the top 5 companies by year
avg_rating <- caClean_top %>%
  group_by(`Company...Maker.if.known.`, Review.Date) %>%
  summarise(avg_rating = mean(Rating))
```

```
## `summarise()` has grouped output by 'Company...Maker.if.known.'. You can
```

```
## override using the `.groups` argument.
```

```
# plot the development of the average rating of the top 5 companies by year
ggplot(avg_rating, aes(x = Review.Date, y = avg_rating, color = `Company...Maker.if.known.`)) +
  geom_line() +
  labs(title = "Development of Average Rating of Top 5 producing Companies",
       x = "Year",
       y = "Average Rating",
       color = "Company") +
  theme_minimal() +
  theme(plot.title = element_text(hjust = 0.5))
```

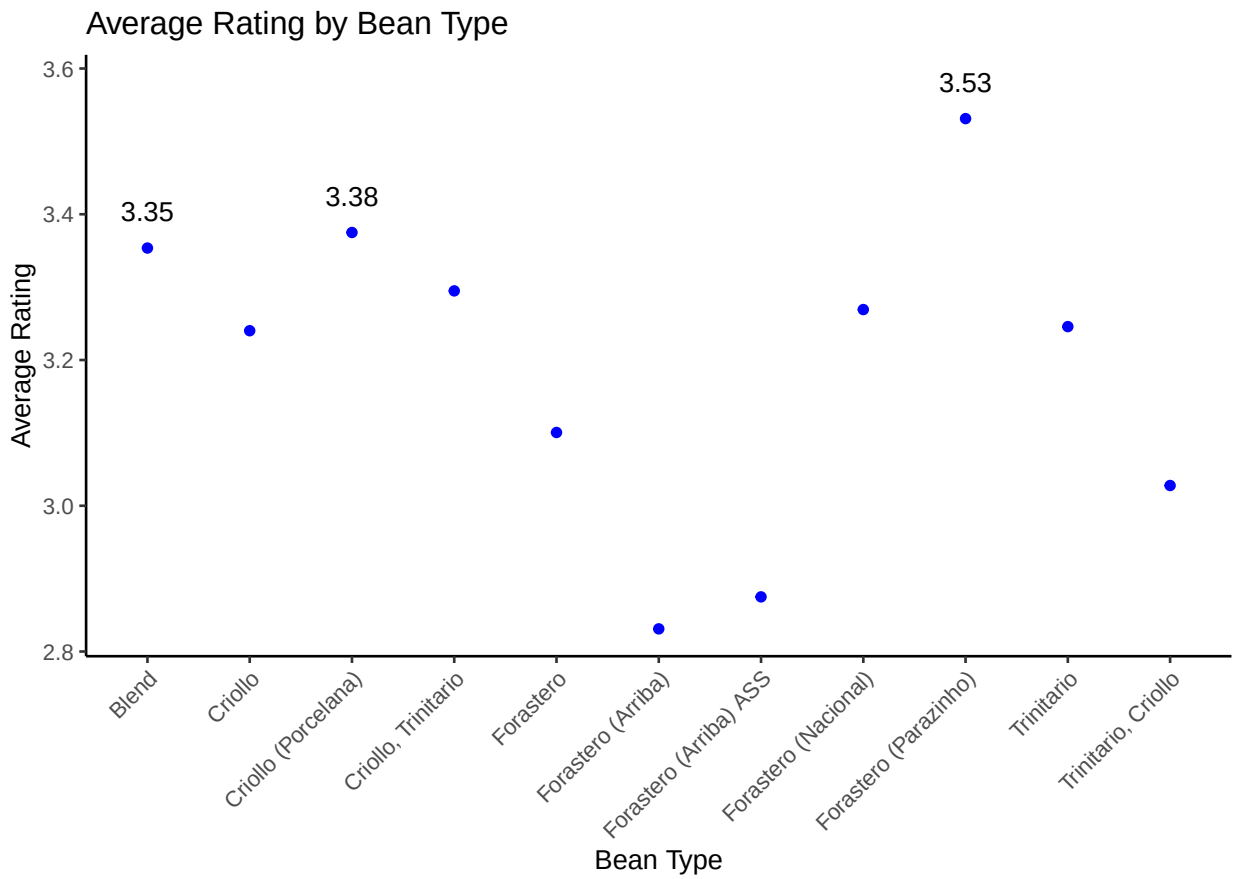


In the previous graph these companies are not mentioned at all, apparently they do not have many products and therefore they do not appear in the most common companies. When you look in this graph at the most common companies, you can see that the company Soma maintains a relatively high and stable rating over the years. In contrast, Pralus and Bonnat have high peaks, but after that a decrease was observed.

The last question we would like to ask is what is the correlation between the rating and the type of cocoa - which type of cocoa bean is the best?

```
# Filter out rows with empty or non-letter values in Bean.Type column
caClean <- caClean %>%
  filter(grepl("[a-zA-Z]+", Bean.Type))

# Get the count of each bean type
bean_counts <- caClean %>% count(Bean.Type)
```

```r
# Filter to only include bean types that show up at least 5 times
bean_counts <- bean_counts %>% filter(n >= 5)

# Filter the original dataset to only include these bean types
caClean <- caClean %>% filter(Bean.Type %in% bean_counts$Bean.Type)

# Calculate the average rating for each bean type
bean_ratings <- caClean %>% group_by(Bean.Type) %>% summarise(avg_rating = mean(Rating))

# Get the top three bean types by average rating
top_three <- bean_ratings %>% top_n(3, avg_rating)

# Create a scatter plot of the data
ggplot(bean_ratings, aes(x = Bean.Type, y = avg_rating)) +
  geom_point(color = "blue") +
  theme_classic() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1)) +
  labs(x = "Bean Type", y = "Average Rating", title = "Average Rating by Bean Type") +
  theme(legend.position = "none") + # Remove the legend
  # Add labels for top three bean types
  geom_text(data = top_three, aes(label = round(avg_rating, 2), x = Bean.Type, y = avg_rating + 0.05),
            size = 4, color = "black")
```



In this analysis, we made a filter on unique values that appeared at least 5 times for avoiding noise. The cocoa beans that received the best ratings are: Forastero(Parazinho), Criollo(Porcelana) and Blend. In contrast,

the cocoa beans the got the lowest average rating are Forastero(Aribba)ASS and Forestero(Arriba).

In conclusion, we wanted to check the factors that affect the chocolate rating. We saw that the most recommended percentage of cocoa is in the range of 0.63-0.77%, and the lowest rating is for 100% cocoa. In addition, we saw that the most successful types of cocoa are: Forastero(Parazinho), Criollo(Porcelana), And the least loved are: Forastero(Aribba)ASS and Forestero(Arriba). Furthermore, we saw that the origin countries of the cocoa beans that received the highest rating average of 4 are: Madagascar, Haiti, Peru, Dominican Republic, Bolivia and Venezuela. While the 5 countries that export the most cocoa are: Madagascar, Venezuela, Dominican Republic, Peru and Ecuador. Also, The producing countries that got the highest rating are: Chile, Amsterdam, Netherlands, Philippines, Iceland. And the country that produce the largest amount of chocolate is by far U.S.A. Another finding is that the company which had the highest average rating is Tobago Estate (Pralus), but the differences between this company and the other companies were not really significant. In addition, we discovered that the 5 companies that produced chocolate the most over the years are: A.morin, Bonnat, Fresco, Pralus and Soma - When the company Soma maintained a high and stable rating the most.

Relevent follow-up research questions:

-What makes soma so stable and high-rated? does it uses the preferred cocoa range of 0.63-0.77%? what type of cocoa beans it uses and from which origin country?

-What are the reasons for the instability found in companies that produce chocolate in large quantities- “Bonnat” and “Pralus”? is it from real reasone or just an outliers in the data?

-what is the correlation between rating and the year of the review? What are the factors that caused changes in the rating over the years?