

# Chocolate and Tea Project

Olumide Olaoye

11/15/2021

## Setting up R environment by installing and loading ‘tidyverse’ packages

```
install.packages("tidyverse")

## Installing package into '/cloud/lib/x86_64-pc-linux-gnu-library/4.1'
## (as 'lib' is unspecified)

library(tidyverse)

## -- Attaching packages ----- tidyverse 1.3.1 --
## v ggplot2 3.3.5      v purrr   0.3.4
## v tibble  3.1.6      v dplyr  1.0.7
## v tidyr   1.1.4      v stringr 1.4.0
## v readr   2.1.0      v forcats 0.5.1

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

## Read the ‘flavors\_of\_cacao’ dataset into R and assigned it to a dataframe with variable ‘flavors\_df’

```
flavors_df <- read_csv("flavors_of_cacao.csv")

## Rows: 1795 Columns: 9

## -- Column specification -----
## Delimiter: ","
## chr (6): Brand, Bar_Name, Cocoa_Percent, Company_Location, Bean_Type, Bean_0...
## dbl (3): REF, Review
## Date, Rating

##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

## Getting closer to the datasets

```
str(flavors_df)

## spec_tbl_df [1,795 x 9] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
## $ Brand      : chr [1:1795] "A. Morin" "A. Morin" "A. Morin" "A. Morin" ...
## $ Bar_Name    : chr [1:1795] "Agua Grande" "Kpime" "Atsane" "Akata" ...
## $ REF         : num [1:1795] 1876 1676 1676 1680 1704 ...
```

```
## $ Review
## Date      : num [1:1795] 2016 2015 2015 2015 2015 ...
## $ Cocoa_Percent : chr [1:1795] "63%" "70%" "70%" "70%" ...
## $ Company_Location: chr [1:1795] "France" "France" "France" "France" ...
## $ Rating        : num [1:1795] 3.75 2.75 3 3.5 3.5 2.75 3.5 3.5 3.75 4 ...
## $ Bean_Type      : chr [1:1795] " " " " " " " " " " ...
## $ Bean_Origin    : chr [1:1795] "Sao Tome" "Togo" "Togo" "Togo" ...
## - attr(*, "spec")=
## .. cols(
## ..   Brand = col_character(),
## ..   Bar_Name = col_character(),
## ..   REF = col_double(),
## ..   `Review
## .. Date` = col_double(),
## ..   Cocoa_Percent = col_character(),
## ..   Company_Location = col_character(),
## ..   Rating = col_double(),
## ..   Bean_Type = col_character(),
## ..   Bean_Origin = col_character()
## .. )
## - attr(*, "problems")=<externalptr>
```

## Check the columns in the dataframe

```
colnames(flavors_df)
```

```
## [1] "Brand"          "Bar_Name"        "REF"             "Review\nDate"
## [5] "Cocoa_Percent"  "Company_Location" "Rating"          "Bean_Type"
## [9] "Bean_Origin"
```

```
#Take a glimpse of the dataframe 'flavors_df'
```

```
glimpse(flavors_df)
```

```
## Rows: 1,795
## Columns: 9
## $ Brand      <chr> "A. Morin", "A. Morin", "A. Morin", "A. Morin", "A. M~
## $ Bar_Name   <chr> "Agua Grande", "Kpime", "Atsane", "Akata", "Quilla", ~
## $ REF        <dbl> 1876, 1676, 1676, 1680, 1704, 1315, 1315, 1315, 1319,~
## $ `Review\nDate` <dbl> 2016, 2015, 2015, 2015, 2015, 2014, 2014, 2014, 2014,~
## $ Cocoa_Percent <chr> "63%", "70%", "70%", "70%", "70%", "70%", "70%", "70%~
## $ Company_Location <chr> "France", "France", "France", "France", "France", "Fr~
## $ Rating     <dbl> 3.75, 2.75, 3.00, 3.50, 3.50, 2.75, 3.50, 3.50, 3.75,~
## $ Bean_Type   <chr> " ", " ", " ", " ", " ", " ", " ", "Criollo", " ", "Criollo", "~
## $ Bean_Origin <chr> "Sao Tome", "Togo", "Togo", "Togo", "Peru", "Venezuel~
```

## View 'flavors\_df'

```
flavors_df
```

```
## # A tibble: 1,795 x 9
##   Brand      Bar_Name    REF `Review\nDate` Cocoa_Percent Company_Location Rating
##   <chr>      <chr>      <dbl>      <dbl> <chr>          <chr>          <dbl>
## 1 A. Morin Agua Gra~  1876        2016 63%          France         3.75
## 2 A. Morin Kpime    1676        2015 70%          France         2.75
```

```
## 3 A. Morin Atsane      1676      2015 70%      France      3
## 4 A. Morin Akata      1680      2015 70%      France      3.5
## 5 A. Morin Quilla     1704      2015 70%      France      3.5
## 6 A. Morin Carenero   1315      2014 70%      France      2.75
## 7 A. Morin Cuba       1315      2014 70%      France      3.5
## 8 A. Morin Sur del ~  1315      2014 70%      France      3.5
## 9 A. Morin Puerto C~  1319      2014 70%      France      3.75
## 10 A. Morin Pablino    1319      2014 70%      France      4
## # ... with 1,785 more rows, and 2 more variables: Bean_Type <chr>,
## #   Bean_Origin <chr>
```

Select the columns needed for our analysis in the dataframe and assign to a new variable ‘trimmed\_flavors\_df’

```
trimmed_flavors_df <- flavors_df %>%
  select(Rating, Cocoa_Percent, Brand, Company_Location)
```

View ‘trimmed\_flavors\_df’

```
trimmed_flavors_df

## # A tibble: 1,795 x 4
##   Rating Cocoa_Percent Brand   Company_Location
##   <dbl> <chr>      <chr>      <chr>
## 1  3.75 63%      A. Morin France
## 2  2.75 70%      A. Morin France
## 3  3    70%      A. Morin France
## 4  3.5 70%      A. Morin France
## 5  3.5 70%      A. Morin France
## 6  2.75 70%      A. Morin France
## 7  3.5 70%      A. Morin France
## 8  3.5 70%      A. Morin France
## 9  3.75 70%      A. Morin France
## 10 4    70%      A. Morin France
## # ... with 1,785 more rows
```

Get the standard deviation of ‘trimmed\_flavors\_df’ from the column ‘Rating’

```
trimmed_flavors_df %>%
  summarize(sd(Rating))
```

```
## # A tibble: 1 x 1
##   `sd(Rating)`
##   <dbl>
## 1      0.478
```

Clean up the dataframe with filter() function for chocolates with percentage of cocoa greater than or equal to 75 and rating score greater than or equal to 3.9. Assign to a new variable ‘best\_trimmed\_flavors\_df’

```
best_trimmed_flavors_df <- trimmed_flavors_df %>%
  filter(Cocoa_Percent >= 75, Rating >= 3.9)
```

## View ‘best\_trimmed\_flavors\_df’

```
best_trimmed_flavors_df
```

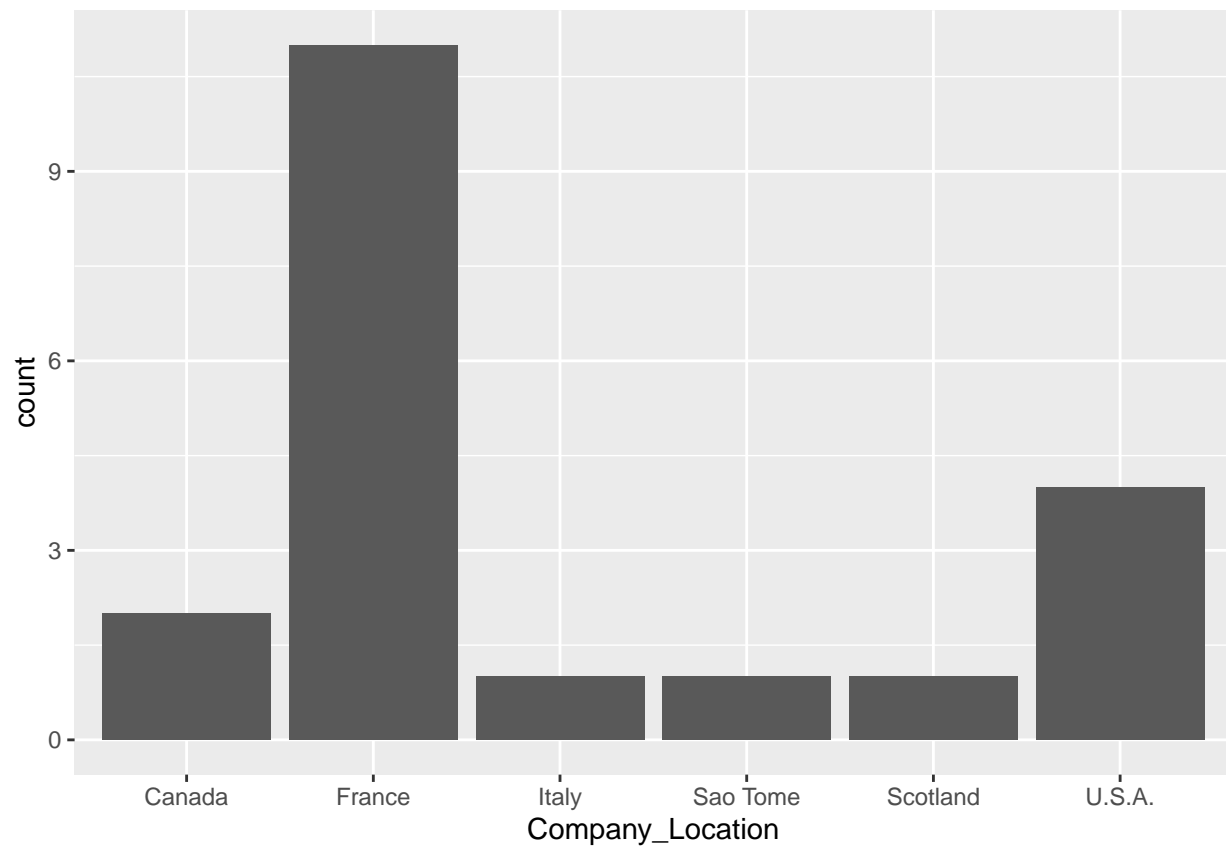
```
## # A tibble: 20 x 4
##   Rating Cocoa_Percent Brand      Company_Location
##   <dbl> <chr>      <chr>      <chr>
## 1     4 75%      Amedei      Italy
## 2     4 75%      Bonnat      France
## 3     4 75%      Bonnat      France
## 4     4 75%      Bonnat      France
## 5     4 75%      Bonnat      France
## 6     4 75%      Bonnat      France
## 7     4 75%      Bonnat      France
## 8     4 75%      Bonnat      France
## 9     4 75%      Bonnat      France
## 10    4 75%      Claudio Corallo Sao Tome
## 11    4 75%      Ocelot      Scotland
## 12    4 75%      Patric      U.S.A.
## 13    4 75%      Pralus      France
## 14    4 75%      Pralus      France
## 15    4 80%      Pralus      France
## 16    4 75%      Ritual      U.S.A.
## 17    4 75%      Rogue       U.S.A.
## 18    4 78%      Scharffen Berger U.S.A.
## 19    4 75%      Soma        Canada
## 20    4 88%      Soma        Canada
```

## Visualize our analysis

Based on:

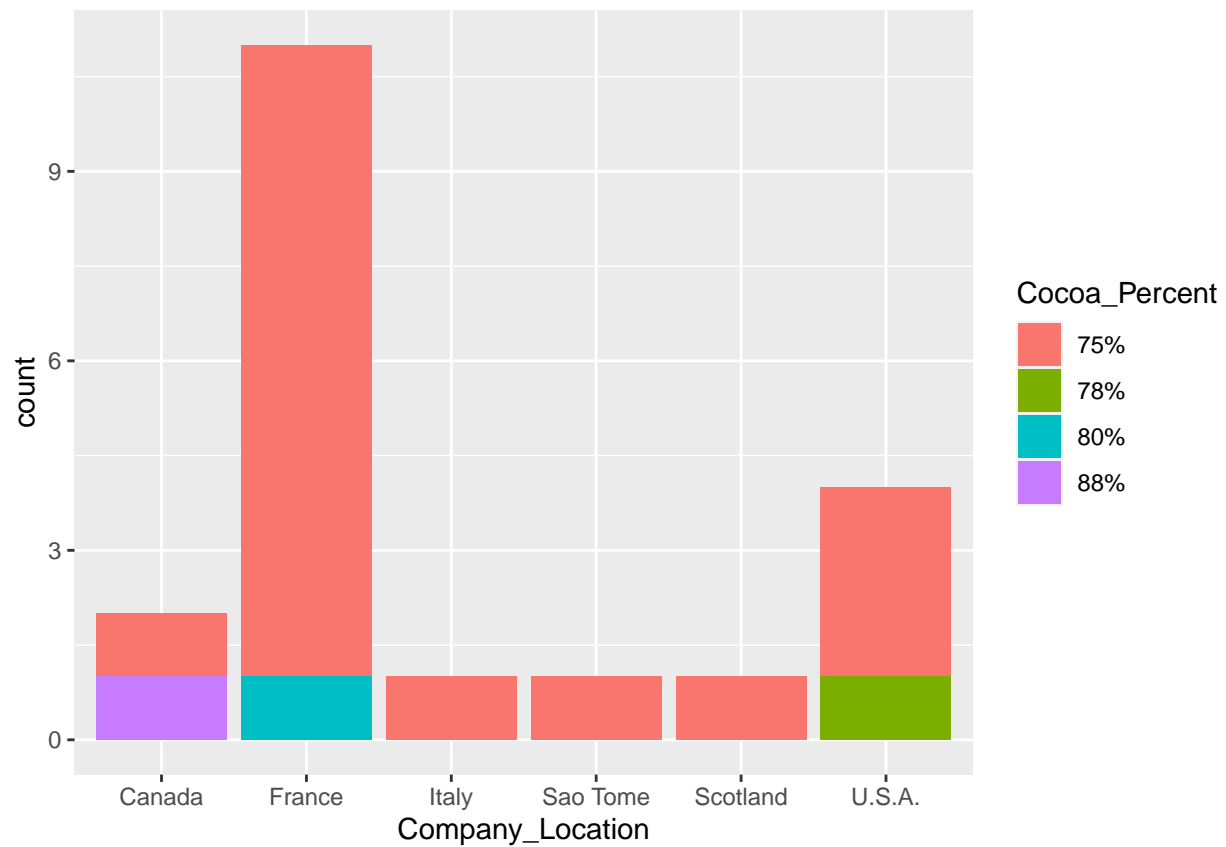
- Company Location

```
ggplot(data=best_trimmed_flavors_df)+
  geom_bar(mapping=aes(x=Company_Location))
```



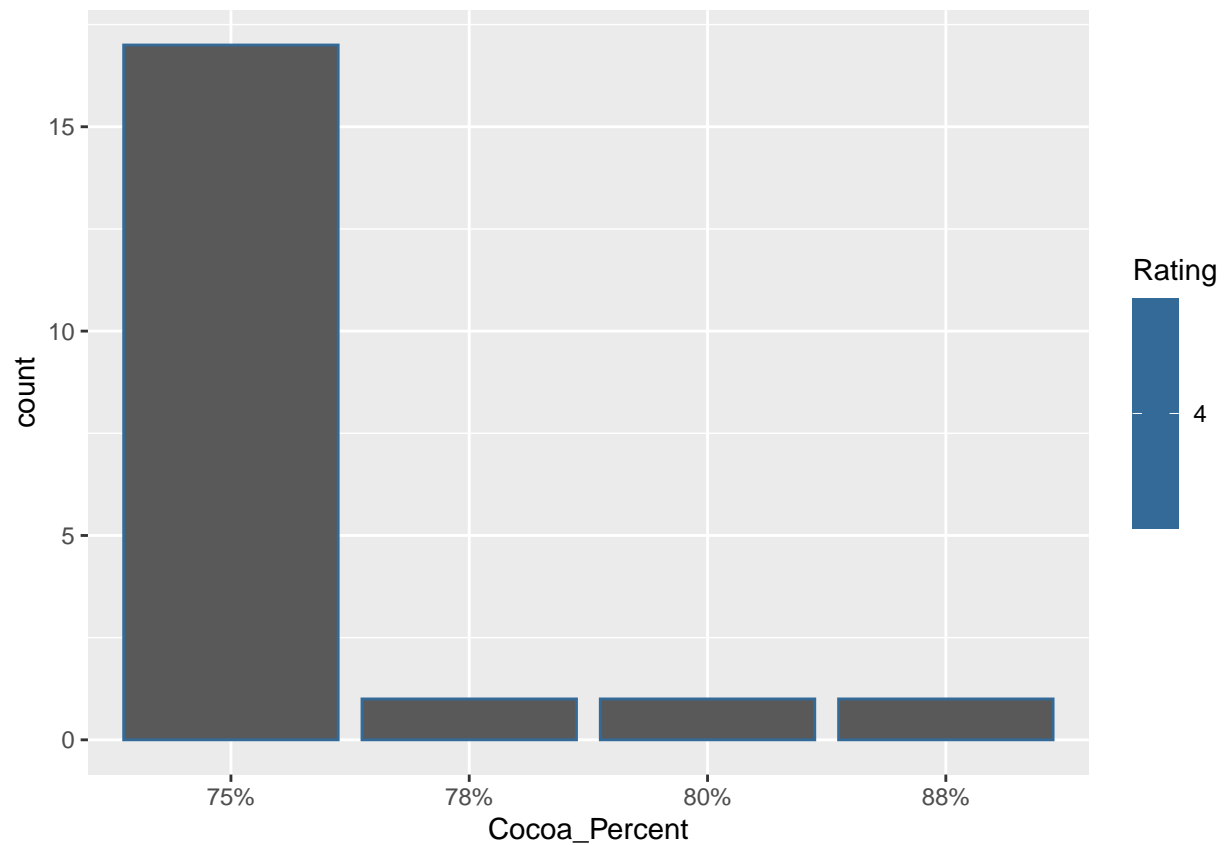
- Company Location Vs Percentage of Cocoa

```
ggplot(data=best_trimmed_flavors_df)+  
  geom_bar(mapping=aes(x=Company_Location, fill=Cocoa_Percent))
```



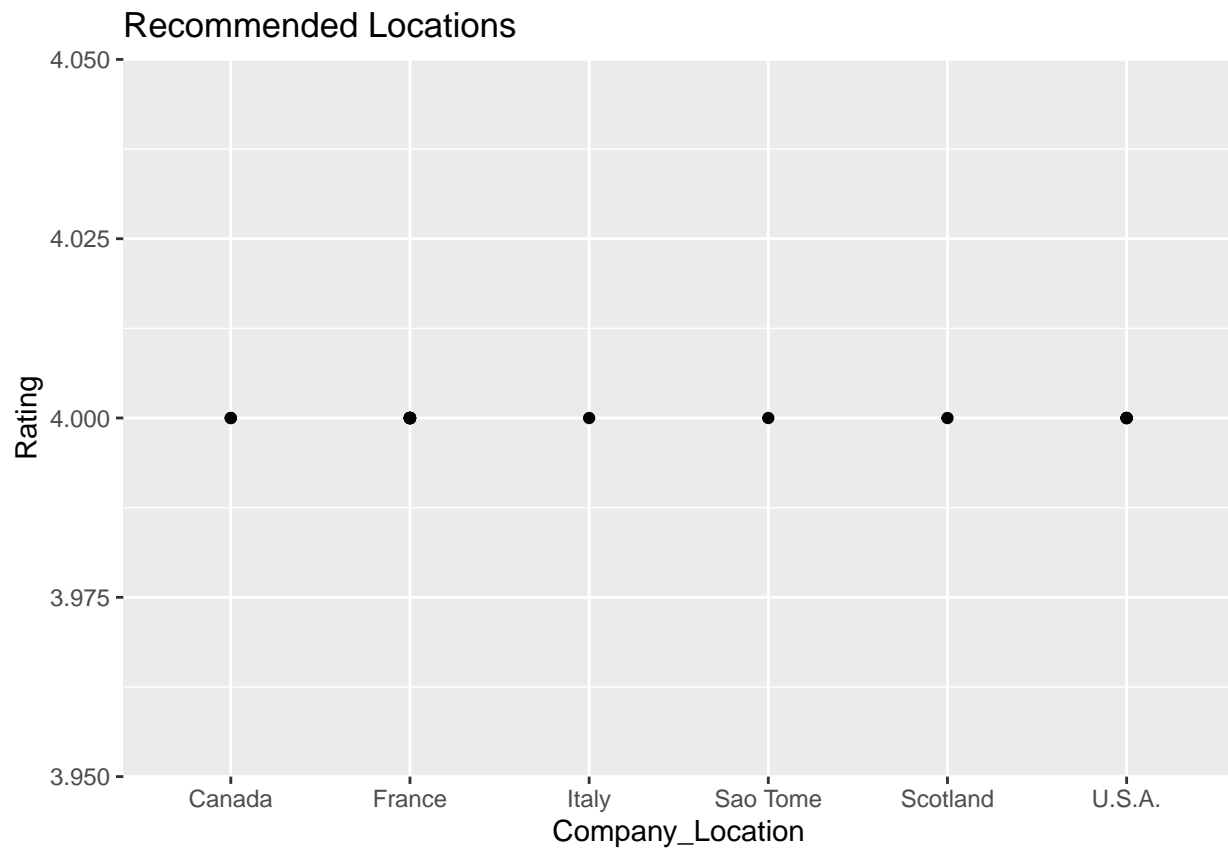
Check the rating score of the chocolate bars based on the percentage of cocoa

```
ggplot(data=best_trimmed_flavors_df)+  
  geom_bar(mapping=aes(x=Cocoa_Percent))+  
  geom_bar(mapping=aes(x=Cocoa_Percent, color=Rating))
```



Recoomended Locations to concentrate chocolate production based on Rating score

```
ggplot(data=best_trimmed_flavors_df)+  
  geom_point(mapping=aes(x=Company_Location, y=Rating))+  
  labs(title="Recommended Locations")
```



## Conclusion

Based on the questions asked by the stakeholders and the result of the analysis, the business is advised to continue productions in all the six locations where rating is greater than or equal to 3.9 and cocoa percentage is greater than or equal to 75%.

But the country with the highest count based on company location is France. So we can target more campaign activities towards France and share equal budgets to the other 5 locations.

Thank you!