Caso de estudio: Gourmet Analytics

El siguiente caso de estudio muestra el desarrollo de un trabajo de analítica para la empresa Chocolate and Tea, la cual busca saber qué países producen las barras de chocolate muy amargo (alto porcentaje de cacao) que tienen la mejor calificación, con el fin de crear un próximo menú de barras de chocolate.

Para llevar a cabo este análisis, se utiliza el dataset de *Chocolate Bar Ratings* disponible en Kaggle. Por medio de los cuadernos de R Markdown, es posible contar con una herramienta sólida para el procesamiento de datos con una alta reproducibilidad y formas de compartir los hallazgos; produciendo **visualizaciones de alta calidad**.

Para proseguir, cargamos el archivo .csv recuperado y lo almacenamos en un dataframe de R.

```
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.2     v readr     2.1.4
## v forcats   1.0.0     v stringr   1.5.0
## v ggplot2   3.4.2     v tibble    3.2.1
## v lubridate 1.9.2     v tidyr     1.3.0
## v purrr     1.0.1
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

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

```
## Rows: 1795 Columns: 9
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (6): Company
## (Maker-if known), Specific Bean Origin
## or Bar Name, Cocoa
## ...
## 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.
```

Add a new chunk by clicking the *Insert Chunk* button on the toolbar or by pressing *Ctrl+Alt+I*.

Ahora, llevaremos a cabo un breve Análisis Exploratorio de Datos para conocer las características básicas del dataframe. Para este, usamos str(), que nos permite conocer la dimensión del dataframe (1795x9), el nombre de las columnas y su tipo de dato:

```
str(flavors_df)
```

```
## spc_tbl_ [1,795 x 9] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
##  $ Company
## (Maker-if known)       : chr [1:1795] "A. Morin" "A. Morin" "A. Morin" "A. Morin" ...
##  $ Specific Bean Origin
```

```
## or 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] " " " " " " " " " " ...
##  $ Broad Bean
## Origin               : chr [1:1795] "Sao Tome" "Togo" "Togo" "Togo" ...
##  - attr(*, "spec")=
##   .. cols(
##   ..   `Company
##   .. (Maker-if known)` = col_character(),
##   ..   `Specific Bean Origin
##   .. or 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(),
##   ..   `Broad Bean
##   .. Origin` = col_character()
##   .. )
##  - attr(*, "problems")=<externalptr>
```

Con head() obtenemos una previsualización de las primeras 6 observaciones del dataframe:

```
head(flavors_df)
```

```
## # A tibble: 6 x 9
##   `Company \n(Maker-if known)` Specific Bean Origin\nor B~1   REF `Review\nDate`
##   <chr>                        <chr>                        <dbl>          <dbl>
## 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
## # i abbreviated name: 1: `Specific Bean Origin\nor Bar Name`
## # i 5 more variables: `Cocoa\nPercent` <chr>, `Company\nLocation` <chr>,
## #   Rating <dbl>, `Bean\nType` <chr>, `Broad Bean\nOrigin` <chr>
```

Se observa que el nombre de la compañía tiene la leyenda adicional (*Maker-if known*). Con motivos de claridad y consistencia, renombramos el nombre de la columna por únicamente *Company*. Adicionalmente, modificamos Cocoa Percent por *Cocoa.Percent*. Para esto, empleamos un pipeline como se muestra a continuación:

```
new_flavors_df <- flavors_df %>%
  rename(Company = `Company
(Maker-if known)`) %>%
  rename(Cocoa.Percent = `Cocoa
Percent`) %>%
  rename(Company.Location = `Company
Location`)
```

```
head(new_flavors_df)
```

```
## # A tibble: 6 x 9
##   Company  Specific Bean Origin\nor Bar Nam~1   REF `Review\nDate` Cocoa.Percent
##   <chr>    <chr>                              <dbl>          <dbl> <chr>
## 1 A. Morin Agua Grande                         1876           2016 63%
## 2 A. Morin Kpime                               1676           2015 70%
## 3 A. Morin Atsane                              1676           2015 70%
## 4 A. Morin Akata                               1680           2015 70%
## 5 A. Morin Quilla                              1704           2015 70%
## 6 A. Morin Carenero                            1315           2014 70%
## # i abbreviated name: 1: `Specific Bean Origin\nor Bar Name`
## # i 4 more variables: Company.Location <chr>, Rating <dbl>, `Bean\nType` <chr>,
## #   `Broad Bean\nOrigin` <chr>
```

Para nuestro análisis, nos enfocaremos en las variables de interés: *rating*, *Cocoa.Percent* y *Company*. Creamos un nuevo dataframe con esas tres variables únicamente:

```
trimmed_flavors_df <- new_flavors_df %>%
  select(Company, Company.Location, Rating, Cocoa.Percent)
```

```
head(trimmed_flavors_df)
```

```
## # A tibble: 6 x 4
##   Company  Company.Location Rating Cocoa.Percent
##   <chr>    <chr>             <dbl> <chr>
## 1 A. Morin France             3.75 63%
## 2 A. Morin France             2.75 70%
## 3 A. Morin France             3    70%
## 4 A. Morin France             3.5  70%
## 5 A. Morin France             3.5  70%
## 6 A. Morin France             2.75 70%
```

A continuación, se obtienen estadísticos básicos usando summarize() y sd() para obtener la desviación estándar en la calificación de los datos: es decir, qué tanta dispersión tienen las observaciones de los Ratings con respecto a la media.

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

```
## # A tibble: 1 x 1
##   RatingSD
##      <dbl>
## 1    0.478
```

A partir del caso de estudio, se determina que un Rating de 3.75 (alta calificación) y un porcentaje de Cacao superior al 80% se consideran como las condiciones objetivo del análisis. A continuación, se aplica un filter() para conocer qué observaciones cumplen con ambas condiciones:

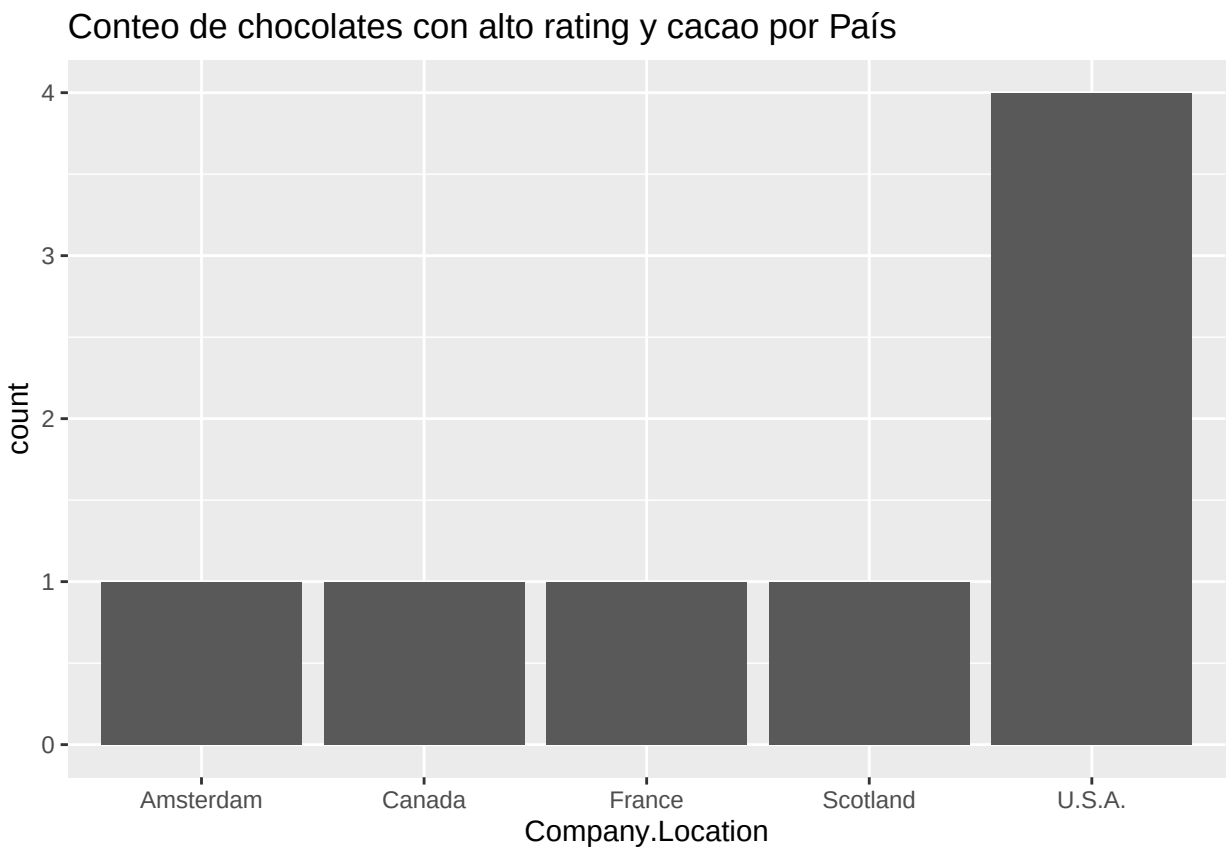
```r
best_trimmed_flavors_df <- trimmed_flavors_df %>%
  filter(Cocoa.Percent >= 80, Rating >= 3.75)
```

```r
head(best_trimmed_flavors_df)
```

```
## # A tibble: 6 x 4
##   Company             Company.Location Rating Cocoa.Percent
##   <chr>               <chr>             <dbl> <chr>
## 1 Chocolate Makers    Amsterdam          3.75 80%
## 2 Chocolate Tree, The Scotland           3.75 80%
## 3 Ethereal            U.S.A.             3.75 80%
## 4 Potomac             U.S.A.             3.75 82%
## 5 Pralus              France             4    80%
## 6 Rogue               U.S.A.             3.75 80%
```

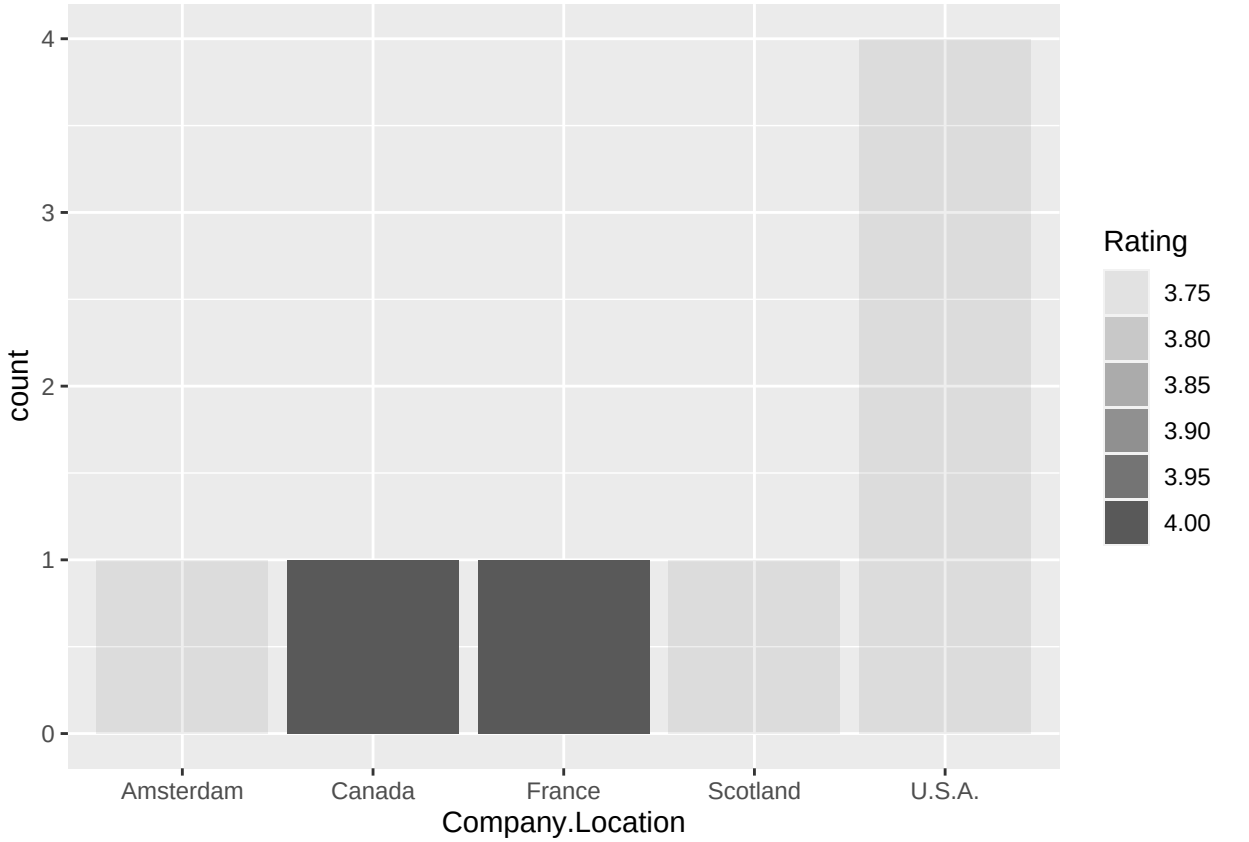
Con los datos limpios, iniciamos a crear visualizaciones para analizarlos usando ggplot2. A continuación, creamos un *barplot* para analizar de donde provienen mayormente los chocolates que cumplen con ambas condiciones (alta calificación y alto porcentaje de cacao).

```r
ggplot(data = best_trimmed_flavors_df) +
  geom_bar(mapping = aes(x = Company.Location)) +
  labs(title = "Conteo de chocolates con alto rating y cacao por País")
```



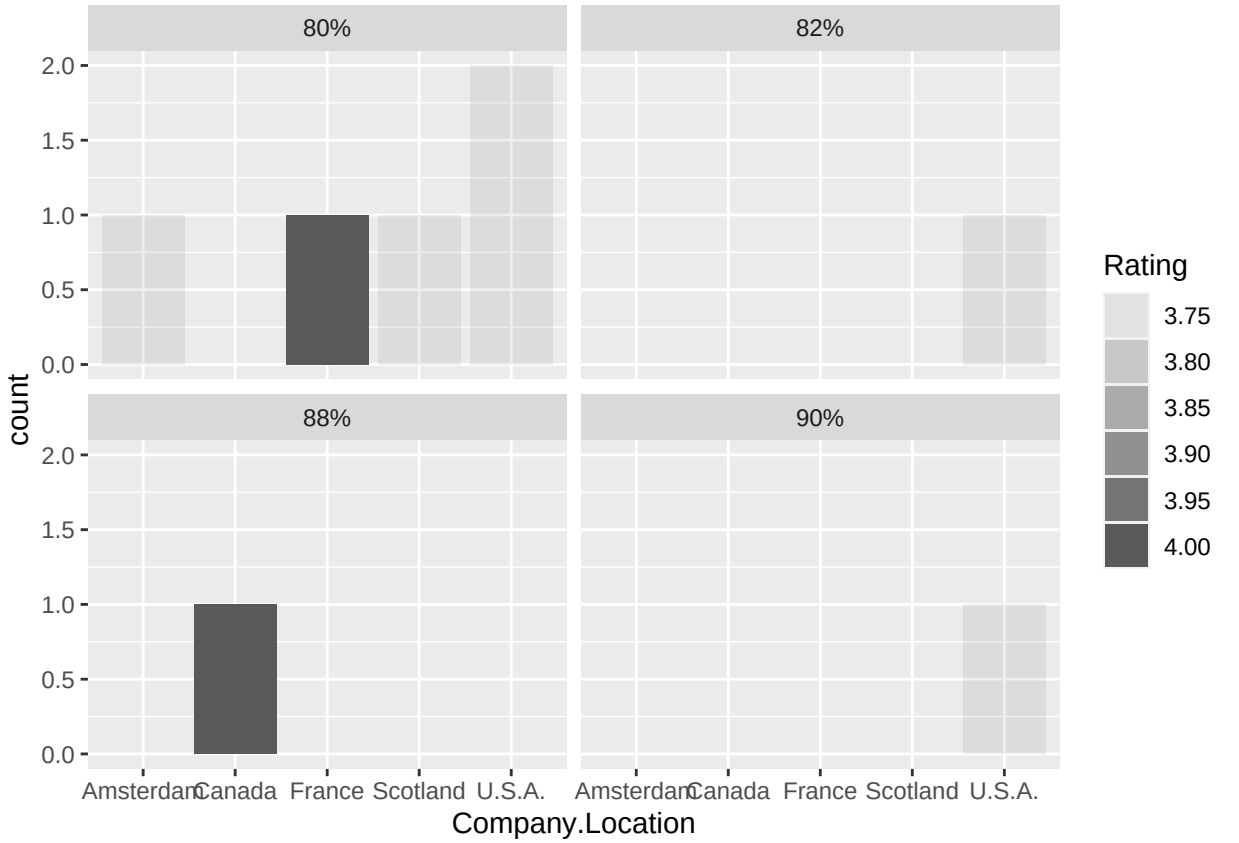
El gráfico de barras muestras las ubicaciones donde se producen las barras de chocolate mejor calificadas. Para comprender mejor la calificación específica de cada ubicación, resaltamos cada barra para ubicar exactamente qué país contiene los ratings más altos.

```
ggplot(data = best_trimmed_flavors_df) +
  geom_bar(mapping = aes(x = Company.Location, alpha = Rating))
```



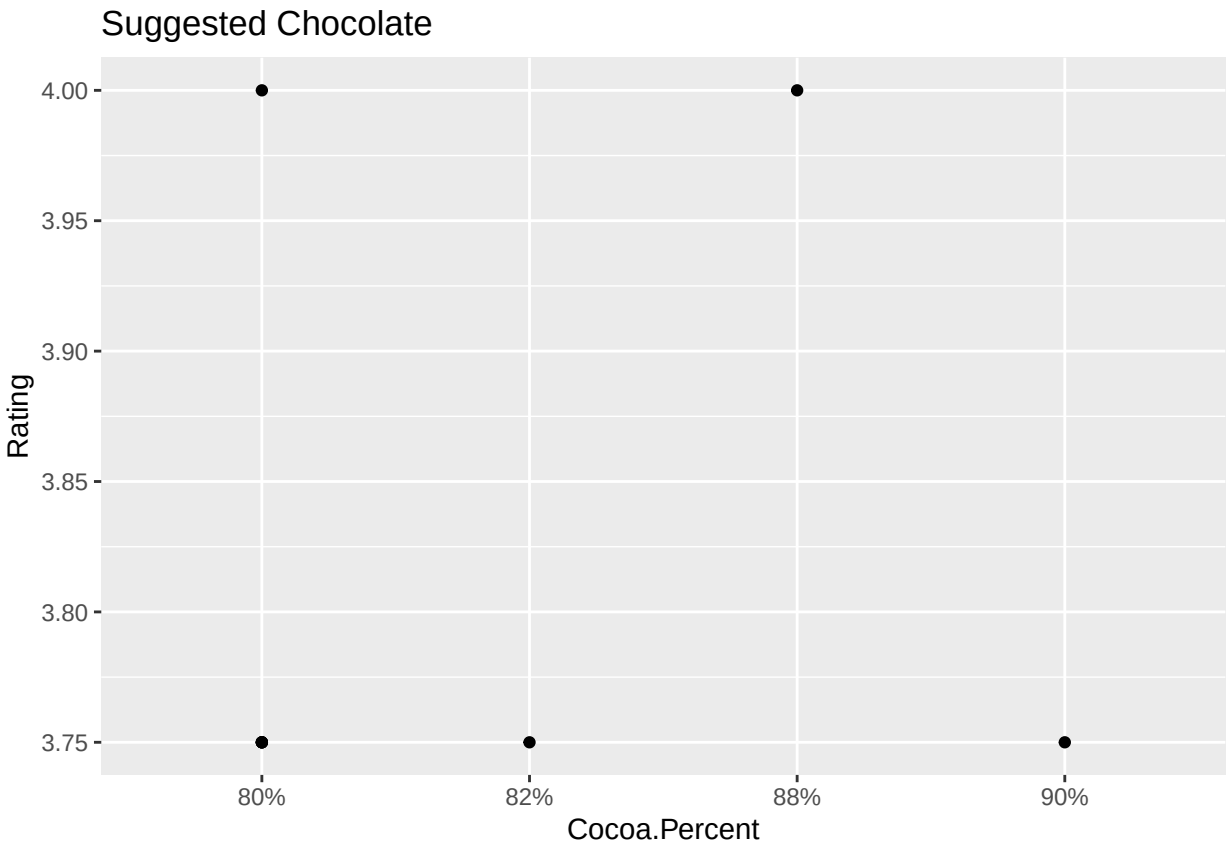
Nuevamente, creamos esta visualización, pero usando un faucet en su lugar para obtener gráficas individuales por cada país:

```
ggplot(data = best_trimmed_flavors_df) +
  geom_bar(mapping = aes(x = Company.Location, alpha = Rating)) +
  facet_wrap(~Cocoa.Percent)
```



Luego, creamos un diagrama de dispersión para observar la relación entre el porcentaje de Cocoa y el rating de cada chocolate.

```
ggplot(data = best_trimmed_flavors_df) +
  geom_point(mapping = aes(x = Cocoa.Percent, y = Rating)) +
  labs(title = "Suggested Chocolate")
```



Guardmos esta gráfica con:

```
ggsave("chocolate.png")
```

```
## Saving 6.5 x 4.5 in image
```