

Proyecto popularidad chocolates

Juan Mario



[Esta foto](#) de Autor desconocido está bajo licencia [CC BY-SA-NC](#)

Índice

Introducción2
Acerca del conjunto de datos proyecto libre2
Preparación de los datos2
Preguntas3
Limpieza4
Procesar5
Analizar6
Compartir11
Conclusiones12
Bibliografía13

Introducción

En el conjunto de datos de chocolates bar, vamos a clasificarlos por popularidad y porcentaje de cocoa, haciendo que los demás datos nos digan en que nivel de calificación están, en la más alta que es “Elite” o en la más baja que es “Decepcionante”.

Para ello utilizaremos herramientas de almacenamiento de datos como BigQuery, Excel, google sheets como intermediario entre hojas de cálculo, Rstudio y MySQL como otro repositorio de datos a usar, para la visualización de datos utilizaremos Tableau desktop con licencia de estudiante.

Haciendo estadísticas para averiguar si las tendencias están correctas, si hay alguna relación entre los datos cuantitativos que tenemos.

Limpiaremos los datos, dejando los datos null ya que no se pueden intercambiar entre columnas porque unos son datos cuantitativos y otros son datos cualitativos, solamente se eliminarán los datos de tipo String irreconocibles.

Acerca del conjunto de datos proyecto libre

“El chocolate es uno de los dulces más populares del mundo. Cada año, los residentes de los Estados Unidos comen colectivamente más de 2,800 millones de libras. Sin embargo, ¡no todas las barras de chocolate son iguales! Este conjunto de datos contiene calificaciones de expertos de más de 1700 barras de chocolate individuales, junto con información sobre su origen regional, porcentaje de cacao, la variedad de grano de chocolate utilizado y dónde se cultivaron los granos”¹

Preparación de los datos

Los datos fueron consultados y descargados de Kaggle, y almacenados en una base de datos en BigQuery y MySQL Workbench. Los datos varían en tipo de datos, tenemos String y double, de caracteres y de números decimales.

¹ Tatman, R. (2017). Chocolate Bar Ratings. Retrieved January 25, 2023, from Kaggle.com website: <https://www.kaggle.com/datasets/rtatman/chocolate-bar-ratings>

Sitio de datos: “Tatman, R. (2017). Chocolate Bar Ratings. Retrieved February 1, 2023, from Kaggle.com website: <https://www.kaggle.com/datasets/rtatman/chocolate-bar-ratings>”

Los datos están con la licencia de: “Creative Commons — CC0 1.0 Universal. (2023). Retrieved February 1, 2023, from Creativecommons.org website:

<https://creativecommons.org/publicdomain/zero/1.0/>”

Con los datos descargados podemos limpiarlos en SQL o Excel, como es un conjunto de datos pequeño optamos por limpiarlos en Excel.

Preguntas

Demostrar tendencias del conjunto de datos.

¿Como se relaciona la popularidad con el porcentaje de cacao?

R: No hay una relación en datos cuantitativos, la popularidad se mantiene alta si el porcentaje de cocoa esta entre 60% y 80%.

¿Qué país tiene la mejor popularidad?

R: U.S.A

¿Qué país exporta más semillas?

R: Perú

¿Qué empresa tiene mayor popularidad?

Fresco, Guittard y Arete, con popularidad mayor a tres.

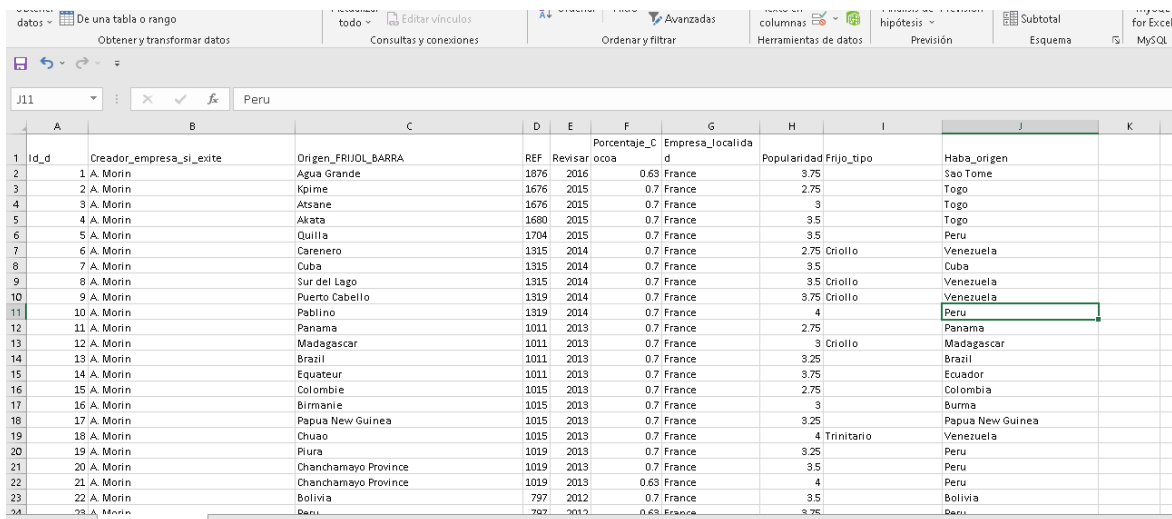
Limpieza

Los datos vienen con caracteres irreconocibles por Excel, BigQuery, MySQL, etc.

Así que la forma de limpieza se usaron las siguientes opciones en Excel, Filtro en datos, Ctrl+b para reemplazar en gran volumen de datos.

En la siguiente imagen, limpiamos el archivo ya que había caracteres irreconocibles por Excel y por gestores de bases de datos.

Usamos un filtro para eliminar esos caracteres.



	A	B	C	D	E	F	G	H	I	J	K
	Id_d	Creador_empresa_si_exite	Origen_FRIJOL_BARRA	REF	Revisar	Porcentaje_C	Empresa_localida	Popularidad	Frijo_tipo	Haba_origen	
1	1	A. Morin	Agua Grande	1876	2016	0.63	France	3.75		Sao Tome	
2	2	A. Morin	Kpime	1676	2015	0.7	France	2.75		Togo	
3	3	A. Morin	Atsane	1676	2015	0.7	France	3		Togo	
4	4	A. Morin	Akaka	1680	2015	0.7	France	3.5		Togo	
5	5	A. Morin	Quilla	1704	2015	0.7	France	3.5		Peru	
6	6	A. Morin	Carenero	1315	2014	0.7	France	2.75	Criollo	Venezuela	
7	7	A. Morin	Cuba	1315	2014	0.7	France	3.5		Cuba	
8	8	A. Morin	Sur del Lago	1315	2014	0.7	France	3.5	Criollo	Venezuela	
9	9	A. Morin	Puerto Cabello	1319	2014	0.7	France	3.75	Criollo	Venezuela	
10	10	A. Morin	Pablino	1319	2014	0.7	France	4		Peru	
11	11	A. Morin	Panama	1011	2013	0.7	France	2.75		Panama	
12	12	A. Morin	Madagascar	1011	2013	0.7	France	3	Criollo	Madagascar	
13	13	A. Morin	Brazil	1011	2013	0.7	France	3.25		Brazil	
14	14	A. Morin	Equateur	1011	2013	0.7	France	3.75		Ecuador	
15	15	A. Morin	Colombie	1015	2013	0.7	France	2.75		Colombia	
16	16	A. Morin	Birmanie	1015	2013	0.7	France	3		Burma	
17	17	A. Morin	Papua New Guinea	1015	2013	0.7	France	3.25		Papua New Guinea	
18	18	A. Morin	Chuao	1015	2013	0.7	France	4	Trinitario	Venezuela	
19	19	A. Morin	Piura	1019	2013	0.7	France	3.25		Peru	
20	20	A. Morin	Chanchamayo Province	1019	2013	0.7	France	3.5		Peru	
21	21	A. Morin	Chanchamayo Province	1019	2013	0.63	France	4		Peru	
22	22	A. Morin	Bolivia	797	2012	0.7	France	3.5		Bolivia	
23	23	A. Morin	Peru	787	2012	0.63	France	3.75		Peru	

Imagen 1: Propia

También usamos Ctrl + b para reemplazar el nombre de metadatos en que los cuales era imposible por el gestor de la base de datos leer.

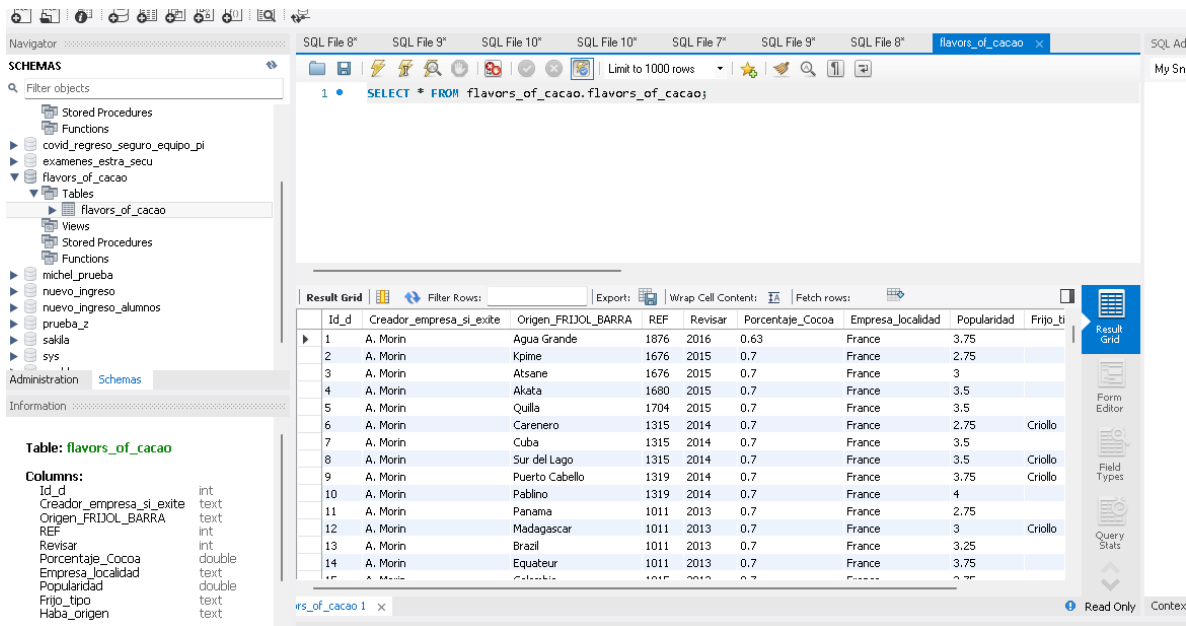


Imagen 2: Propia

Subimos el conjunto de datos BigQuery para tenerlos resguardados por si se llega a haber algún problema con otro gestor de datos. Ya cuando se nos pide eliminar todo lo haremos de forma definida, solo guardando los resultados del análisis.

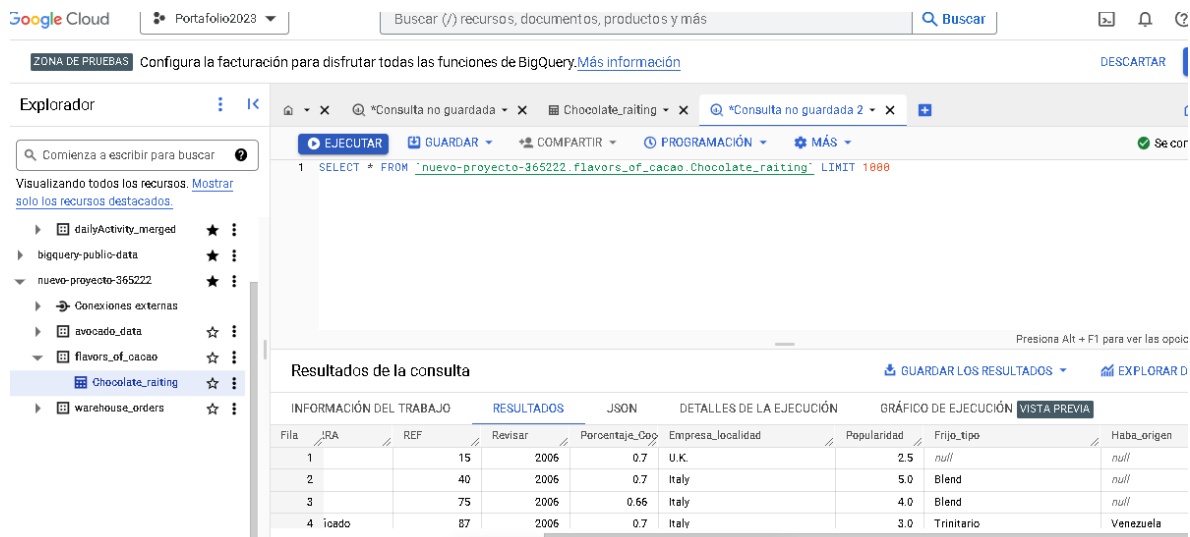


Imagen 3: Propia

Procesar

En R studio necesitábamos de una nueva columna que nos analizara por calidad de los chocolates en un rating de 0 a 5, lo cual nos permitiría hacer más gráficos.

Las herramientas que estamos usando son SQL en BigQuery, Excel, R Studio y Tableau.

Hemos almacenado en una base de datos para futuros registros de datos ya sea en BigQuery o Mysql

Los datos los limpiamos quitando los caracteres irreconocibles.

Analizar

Hemos visto las graficas en R nos demuestra los procesos y las calificaciones por popularidad y el porcentaje de cocoa que tienen estos chocolates.

Los datos de chocolates, tienen en más porcentaje y popularidad los países U.S.A, Venezuela, U.K, Spain, y en azul con poca Nicaragua. Con el tiempo han aumentado las reseñas de los chocolates y la actualización del conjunto de datos.

También tenemos la popularidad y la relación con las reseñas por año, que igual manera han ido aumentando. Los que tienen un porcentaje de Cocoa mayor a %70 y 3 esto quiere decir que el porcentaje de cocoa es bueno, pero falta más producción o reseñas. Podemos decir que el conjunto de datos está bien, pero faltan datos más cuantitativos, para un análisis más completo.

Todo este proceso analítico puede revisarse en: “RPubs - Chocolates_Projetc. (2023, February 2).

Retrieved February 2, 2023, from Rpubs.com website: <https://rpubs.com/Marioouo/998280>“

En R estudio, pusimos los siguientes scritps para realizar gráficas y estadísticas.

##Cargamos las librerias

Nota: las librerias siguientes se pueden instalar con "install.packages:

```
```{r}
```

```
library(ggplot2) #paquete de gráficas
```

```
library(tidyverse) #Paquete que nos ayuda a ocnectar con más paquetes
```

```
library(rmarkdown) #paquete que nos ayuda a cargar un informrte en HTML, word, etc
```

```
library(skimr) #para variables estadisticas
```

```
library(dplyr) #para editar los datos
```

```
library(janitor) #funciones para la limpieza de datos
```

```
library("here") #Este paquete facilita la consulta de los archivos
```

```
library(readr) #para leer datos
```

```
```
```

```
##Datos a analizar
```

para poder cargar un documentos cvs usamos la siguiente función de R

```
```{r}

flavors_of_cacao <-
read_csv("C:/Users/moren/OneDrive/Escritorio/Proyectos/Proyecto_Chocolate/flavors_of_cacao.csv")

View(flavors_of_cacao)
```

```
##Datros con clasificación de popularidad
```

```
flavors_of_cacao_V3 <-
read_csv("C:/Users/moren/OneDrive/Escritorio/Proyectos/Proyecto_Chocolate/flavors_of_cacao_V3.csv")

View(flavors_of_cacao_V3)

```
```

```
##Reporte de datos
```

Usamos las siguientes funciones para que nos de un resumen de los datos que estamos usando.

```
```{r}

skim_without_charts(flavors_of_cacao_V3) #resumen detallado de los datos

glimpse(flavors_of_cacao_V3) #resumen de las columnas

head(flavors_of_cacao_V3)

```
```


##Gráficas

Vemos que en el diagrama de dispersión tenemos la popularidad de Desagradable a Elite y sus niveles y como es que se comportan.

```
```{r}

ggplot(data = flavors_of_cacao_V3) + geom_point((mapping =
 aes(x = Porcentaje_Cocoa,
 y = Popularidad, color =
 Popularidad_Class
))) +
 labs(title="Porcentaje de cocoa y popularidad por clase",
 caption= "@RACHAEL TATMAN conjunto de datos Kaggle")
```
```

Ahora tenemos Tenemos que el porcentaje de Cocoa en mayor numero de conteo es en nivel satisfactorio

```
```{r}

ggplot(data = flavors_of_cacao_V3) + geom_bar((mapping =
 aes(x = Porcentaje_Cocoa
 , fill= Popularidad_Class
))) +
 labs(title="Porcentaje de cocoa y conteo color por popularidad Clase",
 caption= "@RACHAEL TATMAN conjunto de datos Kaggle")
```
```

Grafico de porcentaje de cocoa vs popularidad

el porcentaje de cocoa en 0.7 la popularidad es la más alta.

```
```{r}
```

```
ggplot(data = flavors_of_cacao) + geom_smooth(mapping =
 aes(x= Porcentaje_Cocoa,
 y= Popularidad)))+
labs(title="Cocoa vs popularidad",
 caption= "@RACHAEL TATMAN conjunto de datos Kaggle")
```

```

Popularidad y sus reseñas en cuestión del tiempo por gráficos

```
```{r}
ggplot(data = flavors_of_cacao)+
 geom_bar(mapping=aes(x= Popularidad, fill=Revisar))+
 facet_wrap(~Revisar)+
 labs(title="Popularidad y reseñas",
 caption= "@RACHAEL TATMAN conjunto de datos Kaggle")
```

```

Valor revisado y su aumento REF

```
```{r}
ggplot(data = flavors_of_cacao) +geom_smooth(mapping =
 aes(x = Revisar,
 y = REF)) +
labs(title="Revisars vs REF",
 caption= "@RACHAEL TATMAN conjunto de datos Kaggle")
```

```

Popularidad y porcentaje de Cocoa por localidad

```
```{r}
```

```
ggplot(data = flavors_of_cacao) +geom_jitter(mapping =
 aes(x = Popularidad,
 y = Porcentaje_Cocoa, color = Empresa_localidad))+
geom_smooth(mapping = aes(x = Popularidad,
 y = Porcentaje_Cocoa, color = Empresa_localidad))+
labs(title="Popularidad y porcentaje de cocoa por Localidad",
 caption= "@RACHAEL TATMAN conjunto de datos Kaggle")
```

```

Popularidad clase y conteo, numero de datos que más hay por clase

```
```{r}
ggplot(data = flavors_of_cacao_V3)+
 geom_bar(mapping=aes(x=Popularidad_Class, fill = Popularidad_Class))+
 labs(title="Clase y conteo",
 caption= "@RACHAEL TATMAN conjunto de datos Kaggle")
```

```

##Estadísticas

Teneiendo los datos de los chocolates, en cuestión de estadísticas, las columnas de Porcentaje cococa y popularidad no tienen relación alguna, podemos verlos en las siguientes estadísticas y gráficas.

```
```{r}
flavors_of_cacao_V3 %>%
 group_by(Popularidad_Class) %>%
 summarise(mean(Popularidad), sd(Porcentaje_Cocoa), mean(Porcentaje_Cocoa), sd(Popularidad),
 cor(Popularidad, Porcentaje_Cocoa))
```

```

Gráfica de la estadísticas anterior.

```
```{r}
```

```
ggplot(flavors_of_cacao_V3, aes(Porcentaje_Cocoa, Popularidad)) +
```

```
 geom_point() + geom_smooth(method = lm, se=FALSE)
```

```
```
```

Compartir

Creamos unos Dashboards en tableau para dar a entender las tendencias que nos dicen los datos:

Podemos ver que tenemos la popularidad como filtro de datos, donde los datos se miden por categorías, y de ahí dependen los países que tienen una calidad elite o decepcionante.

Reporte de Chocolates

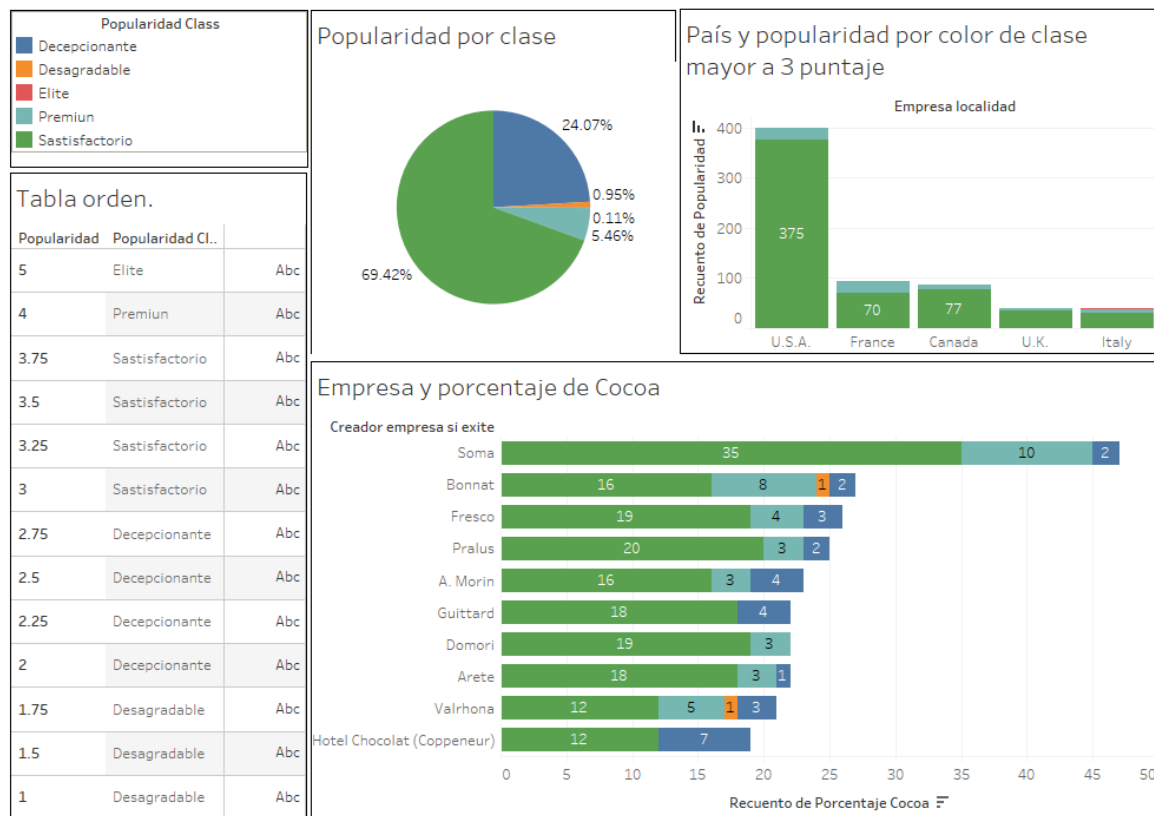


Imagen 4: propia

De igual forma podemos filtrar por semillas de cocoa y los países que más las usan, con ello el tiempo de revisión de los datos y su aumento en popularidad o disminución. Más el promedio de la popularidad por país.

| Popularidad Class | |
|-------------------|--|
| Decepcionante | |
| Desagradable | |
| Elite | |
| Premiun | |
| Sastisfactorio | |

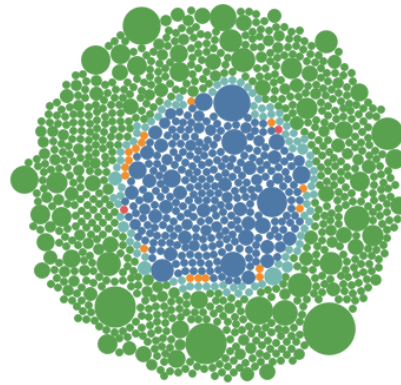
Recuento de Popularidad
1 a 764
y valores NULL

Prom. Popularidad
2.5 a 3.75
y valores NULL

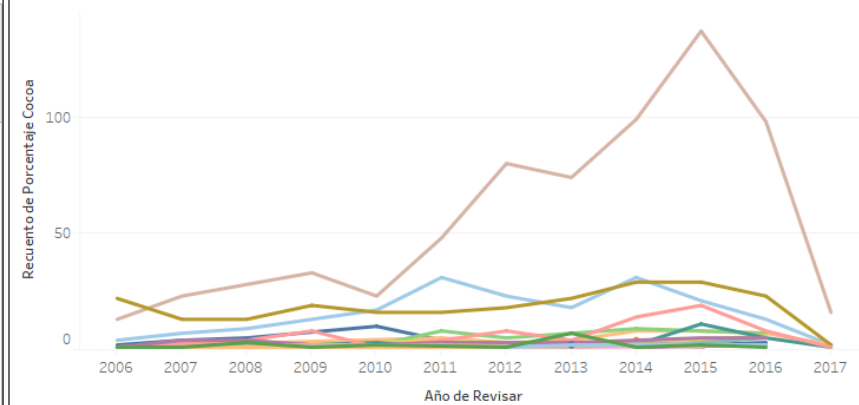
Tabla_Empresa_PRO

| Empresa locali... | Prom. ... | Rec.. |
|-------------------|-----------|-------|
| U.S.A. | 3.2 | 764.0 |
| France | 3.3 | 156.0 |
| Canada | 3.3 | 125.0 |
| U.K. | 3.1 | 96.0 |
| Italy | 3.3 | 63.0 |
| Ecuador | 3.0 | 54.0 |
| Australia | 3.4 | 49.0 |
| Belgium | 3.1 | 40.0 |
| Switzerland | 3.3 | 38.0 |
| Germany | 3.2 | 35.0 |
| Austria | 3.2 | 26.0 |
| Spain | 3.3 | 25.0 |
| Colombia | 3.2 | 23.0 |
| Hungary | 3.2 | 22.0 |
| Venezuela | 3.2 | 20.0 |
| Peru | 2.9 | 17.0 |
| New Zealand | 3.2 | 17.0 |

Popularidad por semilla



Tiempo de revisión y porcentaje de cocoa



Conclusiones

Los datos de chocolates, tienen en más porcentaje y popularidad los países U.S.A, Venezuela, U.K, Spain, y en azul con poca Nicaragua.

Con el tiempo han aumentado las reseñas de los chocolates y la actualización del conjunto de datos.

También tenemos la popularidad y la relación con las reseñas por año, que igual manera han ido aumentando.

Los que tienen un porcentaje de Cocoa mayor a %70 y <80% tienen una popularidad >3

esto quiere decir que el porcentaje de cocoa es bueno, pero falta más producción o reseñas.

Podemos decir que el conjunto de datos está bien, pero faltan datos más cuantitativos, para un análisis más completo.

Bibliografía

1. Tatman, R. (2017). Chocolate Bar Ratings. Retrieved February 1, 2023, from Kaggle.com website: <https://www.kaggle.com/datasets/rtatman/chocolate-bar-ratings>
2. Creative Commons — CC0 1.0 Universal. (2023). Retrieved February 1, 2023, from Creativecommons.org website: <https://creativecommons.org/publicdomain/zero/1.0/>