Chcolates\_PROYECTO

Juan Mario

2023-01-28

##Cargamos las librerias Nota: las librerias siguientes se pueden instalar con “install.packages:

library(ggplot2) #paquete de gráficas

## Warning: package 'ggplot2' was built under R version 4.2.2

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

## Warning: package 'tidyverse' was built under R version 4.2.2

## ── Attaching packages ─────────────────────────────────────── tidyverse 1.3.2 ──  
## ✔ tibble 3.1.8 ✔ dplyr 1.0.10  
## ✔ tidyr 1.2.1 ✔ stringr 1.5.0   
## ✔ readr 2.1.3 ✔ forcats 0.5.2   
## ✔ purrr 0.3.5

## Warning: package 'tibble' was built under R version 4.2.2

## Warning: package 'tidyr' was built under R version 4.2.2

## Warning: package 'readr' was built under R version 4.2.2

## Warning: package 'purrr' was built under R version 4.2.2

## Warning: package 'dplyr' was built under R version 4.2.2

## Warning: package 'stringr' was built under R version 4.2.2

## Warning: package 'forcats' was built under R version 4.2.2

## ── Conflicts ────────────────────────────────────────── tidyverse\_conflicts() ──  
## ✖ dplyr::filter() masks stats::filter()  
## ✖ dplyr::lag() masks stats::lag()

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

## Warning: package 'rmarkdown' was built under R version 4.2.2

library(skimr) #para variables estadisticas

## Warning: package 'skimr' was built under R version 4.2.2

library(dplyr) #para editar los datos  
library(janitor) #funciones para la limpieza de datos

## Warning: package 'janitor' was built under R version 4.2.2

##   
## Attaching package: 'janitor'  
##   
## The following objects are masked from 'package:stats':  
##   
## chisq.test, fisher.test

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

## Warning: package 'here' was built under R version 4.2.2

## here() starts at C:/Users/moren/OneDrive/Documents/Google\_certifid

library(readr) #para leer datos

##Datos a analizar

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

flavors\_of\_cacao <- read\_csv("C:/Users/moren/OneDrive/Escritorio/Proyectos/Proyecto\_Chocolate/flavors\_of\_cacao.csv")

## Rows: 1795 Columns: 10  
## ── Column specification ────────────────────────────────────────────────────────  
## Delimiter: ","  
## chr (5): Creador\_empresa\_si\_exite, Origen\_FRIJOL\_BARRA, Empresa\_localidad, F...  
## dbl (5): Id\_d, REF, Revisar, Porcentaje\_Cocoa, Popularidad  
##   
## 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.

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")

## Rows: 1795 Columns: 11  
## ── Column specification ────────────────────────────────────────────────────────  
## Delimiter: ","  
## chr (6): Creador\_empresa\_si\_exite, Origen\_FRIJOL\_BARRA, Empresa\_localidad, F...  
## dbl (5): Id\_d, REF, Revisar, Porcentaje\_Cocoa, Popularidad  
##   
## 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.

View(flavors\_of\_cacao\_V3)

##Reporte de datos

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

skim\_without\_charts(flavors\_of\_cacao\_V3) #resumen detallado de los datos

Data summary

|  |  |
| --- | --- |
| Name | flavors\_of\_cacao\_V3 |
| Number of rows | 1795 |
| Number of columns | 11 |
| \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ |  |
| Column type frequency: |  |
| character | 6 |
| numeric | 5 |
| \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ |  |
| Group variables | None |

**Variable type: character**

| skim\_variable | n\_missing | complete\_rate | min | max | empty | n\_unique | whitespace |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Creador\_empresa\_si\_exite | 0 | 1.00 | 2 | 39 | 0 | 416 | 0 |
| Origen\_FRIJOL\_BARRA | 0 | 1.00 | 3 | 45 | 0 | 1039 | 0 |
| Empresa\_localidad | 0 | 1.00 | 4 | 17 | 0 | 60 | 0 |
| Frijo\_tipo | 888 | 0.51 | 3 | 23 | 0 | 39 | 0 |
| Haba\_origen | 74 | 0.96 | 4 | 29 | 0 | 99 | 0 |
| Popularidad\_Class | 0 | 1.00 | 5 | 14 | 0 | 5 | 0 |

**Variable type: numeric**

| skim\_variable | n\_missing | complete\_rate | mean | sd | p0 | p25 | p50 | p75 | p100 |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Id\_d | 0 | 1 | 898.00 | 518.32 | 1.00 | 449.50 | 898.00 | 1346.50 | 1795 |
| REF | 0 | 1 | 1035.90 | 552.89 | 5.00 | 576.00 | 1069.00 | 1502.00 | 1952 |
| Revisar | 0 | 1 | 2012.33 | 2.93 | 2006.00 | 2010.00 | 2013.00 | 2015.00 | 2017 |
| Porcentaje\_Cocoa | 0 | 1 | 0.72 | 0.06 | 0.42 | 0.70 | 0.70 | 0.75 | 1 |
| Popularidad | 0 | 1 | 3.19 | 0.48 | 1.00 | 2.88 | 3.25 | 3.50 | 5 |

glimpse(flavors\_of\_cacao\_V3) #resumen de las columnas

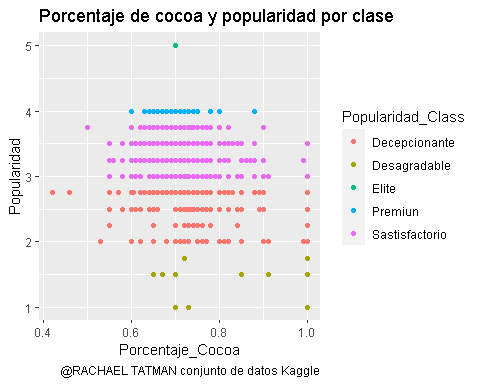
## Rows: 1,795  
## Columns: 11  
## $ Id\_d <dbl> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14…  
## $ Creador\_empresa\_si\_exite <chr> "A. Morin", "A. Morin", "A. Morin", "A. Morin…  
## $ Origen\_FRIJOL\_BARRA <chr> "Agua Grande", "Kpime", "Atsane", "Akata", "Q…  
## $ REF <dbl> 1876, 1676, 1676, 1680, 1704, 1315, 1315, 131…  
## $ Revisar <dbl> 2016, 2015, 2015, 2015, 2015, 2014, 2014, 201…  
## $ Porcentaje\_Cocoa <dbl> 0.63, 0.70, 0.70, 0.70, 0.70, 0.70, 0.70, 0.7…  
## $ Empresa\_localidad <chr> "France", "France", "France", "France", "Fran…  
## $ Popularidad <dbl> 3.75, 2.75, 3.00, 3.50, 3.50, 2.75, 3.50, 3.5…  
## $ Frijo\_tipo <chr> NA, NA, NA, NA, NA, "Criollo", NA, "Criollo",…  
## $ Haba\_origen <chr> "Sao Tome", "Togo", "Togo", "Togo", "Peru", "…  
## $ Popularidad\_Class <chr> "Sastisfactorio", "Decepcionante", "Sastisfac…

head(flavors\_of\_cacao\_V3)

## # A tibble: 6 × 11  
## Id\_d Creador\_…¹ Orige…² REF Revisar Porce…³ Empre…⁴ Popul…⁵ Frijo…⁶ Haba\_…⁷  
## <dbl> <chr> <chr> <dbl> <dbl> <dbl> <chr> <dbl> <chr> <chr>   
## 1 1 A. Morin Agua G… 1876 2016 0.63 France 3.75 <NA> Sao To…  
## 2 2 A. Morin Kpime 1676 2015 0.7 France 2.75 <NA> Togo   
## 3 3 A. Morin Atsane 1676 2015 0.7 France 3 <NA> Togo   
## 4 4 A. Morin Akata 1680 2015 0.7 France 3.5 <NA> Togo   
## 5 5 A. Morin Quilla 1704 2015 0.7 France 3.5 <NA> Peru   
## 6 6 A. Morin Carene… 1315 2014 0.7 France 2.75 Criollo Venezu…  
## # … with 1 more variable: Popularidad\_Class <chr>, and abbreviated variable  
## # names ¹​Creador\_empresa\_si\_exite, ²​Origen\_FRIJOL\_BARRA, ³​Porcentaje\_Cocoa,  
## # ⁴​Empresa\_localidad, ⁵​Popularidad, ⁶​Frijo\_tipo, ⁷​Haba\_origen

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

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

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")

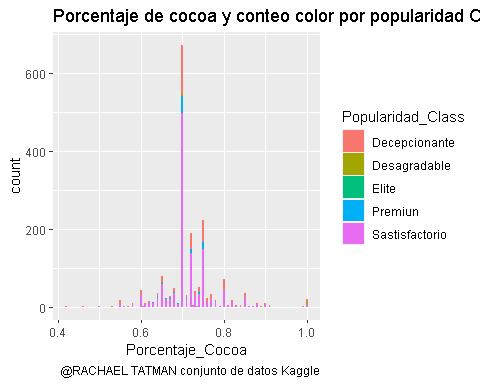
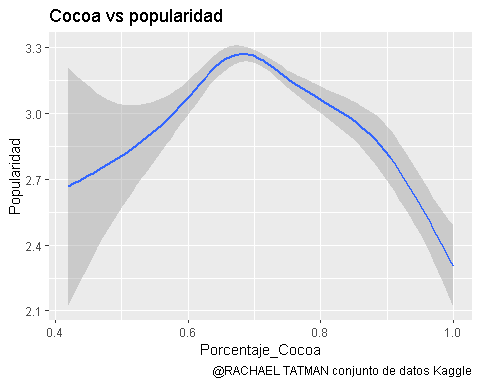


Grafico de porcentaje de cocoa vs popularidad el porcentaje de cocoa en 0.7 la popularidad es la más alta.

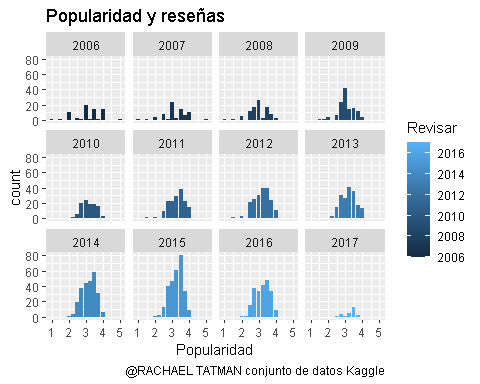
ggplot(data = flavors\_of\_cacao) + geom\_smooth((mapping =   
 aes(x= Porcentaje\_Cocoa,   
 y= Popularidad)))+  
 labs(title="Cocoa vs popularidad",  
 caption= "@RACHAEL TATMAN conjunto de datos Kaggle")

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



Popularidad y sus reseñas en cuestión del tiempo por gráficos

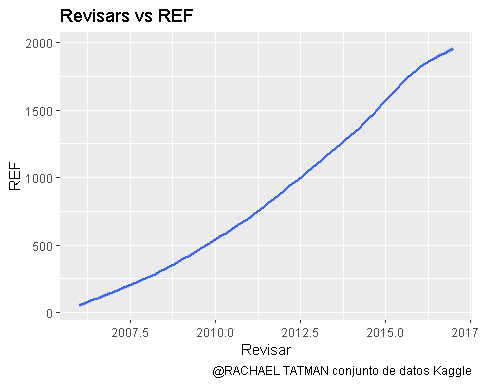
ggplot(data = flavors\_of\_cacao)+  
 geom\_bar(mapping=aes(x= Popularidad, fill=Revisar))+  
 facet\_wrap(~Revisar)+  
 labs(title="Popularidad y reseñas",  
 caption= "@RACHAEL TATMAN conjunto de datos Kaggle")



Valor revisado y su aumento REF

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")

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



Popularidad y porcentaje de Cocoa por localidad

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")

## `geom\_smooth()` using method = 'loess' and formula = 'y ~ x'

## Warning in simpleLoess(y, x, w, span, degree = degree, parametric =  
## parametric, : span too small. fewer data values than degrees of freedom.

## Warning in simpleLoess(y, x, w, span, degree = degree, parametric =  
## parametric, : pseudoinverse used at 3.2475

## Warning in simpleLoess(y, x, w, span, degree = degree, parametric =  
## parametric, : neighborhood radius 0.2525

## Warning in simpleLoess(y, x, w, span, degree = degree, parametric =  
## parametric, : reciprocal condition number 0

## Warning in simpleLoess(y, x, w, span, degree = degree, parametric =  
## parametric, : There are other near singularities as well. 0.063756

## Warning in predLoess(object$y, object$x, newx = if  
## (is.null(newdata)) object$x else if (is.data.frame(newdata))  
## as.matrix(model.frame(delete.response(terms(object)), : span too small. fewer  
## data values than degrees of freedom.

## Warning in predLoess(object$y, object$x, newx = if  
## (is.null(newdata)) object$x else if (is.data.frame(newdata))  
## as.matrix(model.frame(delete.response(terms(object)), : pseudoinverse used at  
## 3.2475

## Warning in predLoess(object$y, object$x, newx = if  
## (is.null(newdata)) object$x else if (is.data.frame(newdata))  
## as.matrix(model.frame(delete.response(terms(object)), : neighborhood radius  
## 0.2525

## Warning in predLoess(object$y, object$x, newx = if  
## (is.null(newdata)) object$x else if (is.data.frame(newdata))  
## as.matrix(model.frame(delete.response(terms(object)), : reciprocal condition  
## number 0

## Warning in predLoess(object$y, object$x, newx = if  
## (is.null(newdata)) object$x else if (is.data.frame(newdata))  
## as.matrix(model.frame(delete.response(terms(object)), : There are other near  
## singularities as well. 0.063756

## Warning in simpleLoess(y, x, w, span, degree = degree, parametric =  
## parametric, : pseudoinverse used at 2.745

## Warning in simpleLoess(y, x, w, span, degree = degree, parametric =  
## parametric, : neighborhood radius 0.755

## Warning in simpleLoess(y, x, w, span, degree = degree, parametric =  
## parametric, : reciprocal condition number 1.4316e-16

## Warning in simpleLoess(y, x, w, span, degree = degree, parametric =  
## parametric, : There are other near singularities as well. 0.25502

## Warning in predLoess(object$y, object$x, newx = if  
## (is.null(newdata)) object$x else if (is.data.frame(newdata))  
## as.matrix(model.frame(delete.response(terms(object)), : pseudoinverse used at  
## 2.745

## Warning in predLoess(object$y, object$x, newx = if  
## (is.null(newdata)) object$x else if (is.data.frame(newdata))  
## as.matrix(model.frame(delete.response(terms(object)), : neighborhood radius  
## 0.755

## Warning in predLoess(object$y, object$x, newx = if  
## (is.null(newdata)) object$x else if (is.data.frame(newdata))  
## as.matrix(model.frame(delete.response(terms(object)), : reciprocal condition  
## number 1.4316e-16

## Warning in predLoess(object$y, object$x, newx = if  
## (is.null(newdata)) object$x else if (is.data.frame(newdata))  
## as.matrix(model.frame(delete.response(terms(object)), : There are other near  
## singularities as well. 0.25502

## Warning in simpleLoess(y, x, w, span, degree = degree, parametric =  
## parametric, : pseudoinverse used at 3.25

## Warning in simpleLoess(y, x, w, span, degree = degree, parametric =  
## parametric, : neighborhood radius 0.25

## Warning in simpleLoess(y, x, w, span, degree = degree, parametric =  
## parametric, : reciprocal condition number 0

## Warning in predLoess(object$y, object$x, newx = if  
## (is.null(newdata)) object$x else if (is.data.frame(newdata))  
## as.matrix(model.frame(delete.response(terms(object)), : pseudoinverse used at  
## 3.25

## Warning in predLoess(object$y, object$x, newx = if  
## (is.null(newdata)) object$x else if (is.data.frame(newdata))  
## as.matrix(model.frame(delete.response(terms(object)), : neighborhood radius 0.25

## Warning in predLoess(object$y, object$x, newx = if  
## (is.null(newdata)) object$x else if (is.data.frame(newdata))  
## as.matrix(model.frame(delete.response(terms(object)), : reciprocal condition  
## number 0

## Warning in simpleLoess(y, x, w, span, degree = degree, parametric =  
## parametric, : span too small. fewer data values than degrees of freedom.

## Warning in simpleLoess(y, x, w, span, degree = degree, parametric =  
## parametric, : at 2.745

## Warning in simpleLoess(y, x, w, span, degree = degree, parametric =  
## parametric, : radius 2.5e-05

## Warning in simpleLoess(y, x, w, span, degree = degree, parametric =  
## parametric, : all data on boundary of neighborhood. make span bigger

## Warning in simpleLoess(y, x, w, span, degree = degree, parametric =  
## parametric, : pseudoinverse used at 2.745

## Warning in simpleLoess(y, x, w, span, degree = degree, parametric =  
## parametric, : neighborhood radius 0.005

## Warning in simpleLoess(y, x, w, span, degree = degree, parametric =  
## parametric, : reciprocal condition number 1

## Warning in simpleLoess(y, x, w, span, degree = degree, parametric =  
## parametric, : at 3.755

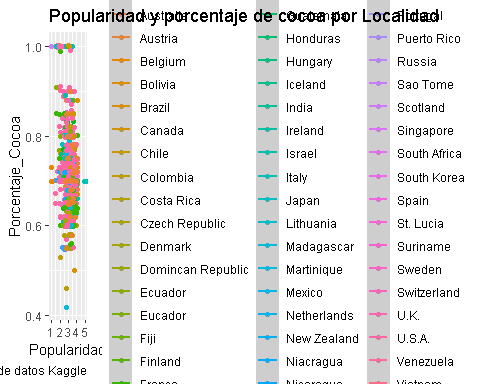
## Warning in simpleLoess(y, x, w, span, degree = degree, parametric =  
## parametric, : radius 2.5e-05

## Warning in simpleLoess(y, x, w, span, degree = degree, parametric =  
## parametric, : all data on boundary of neighborhood. make span bigger

## Warning in simpleLoess(y, x, w, span, degree = degree, parametric =  
## parametric, : There are other near singularities as well. 2.5e-05

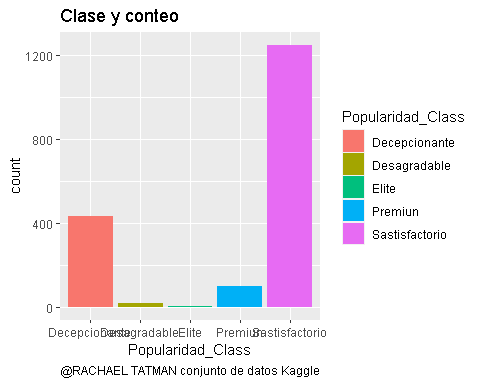
## Warning in simpleLoess(y, x, w, span, degree = degree, parametric =  
## parametric, : zero-width neighborhood. make span bigger  
  
## Warning in simpleLoess(y, x, w, span, degree = degree, parametric =  
## parametric, : zero-width neighborhood. make span bigger

## Warning: Computation failed in `stat\_smooth()`  
## Caused by error in `predLoess()`:  
## ! NA/NaN/Inf en llamada a una función externa (arg 5)



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

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 estadisticas, las columnas de Porcentaje cococa y popularidad no tienen relación alguna, podemos verlos en las siguientes estadísticas y gráficas.

flavors\_of\_cacao\_V3 %>%   
 group\_by(Popularidad\_Class) %>%   
 summarise(mean(Popularidad), sd(Porcentaje\_Cocoa), mean(Porcentaje\_Cocoa), sd(Popularidad),   
 cor(Popularidad, Porcentaje\_Cocoa))

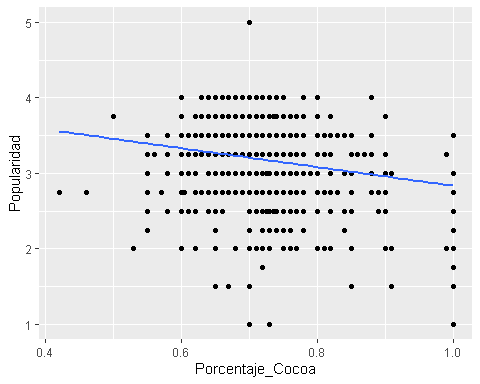
## Warning in cor(Popularidad, Porcentaje\_Cocoa): the standard deviation is zero  
  
## Warning in cor(Popularidad, Porcentaje\_Cocoa): the standard deviation is zero

## # A tibble: 5 × 6  
## Popularidad\_Class `mean(Popularidad)` sd(Porcentaje\_…¹ mean(…² sd(Po…³ cor(P…⁴  
## <chr> <dbl> <dbl> <dbl> <dbl> <dbl>  
## 1 Decepcionante 2.60 0.0777 0.725 0.217 -0.124   
## 2 Desagradable 1.43 0.149 0.843 0.262 0.271   
## 3 Elite 5 0 0.7 0 NA   
## 4 Premiun 4 0.0388 0.708 0 NA   
## 5 Sastisfactorio 3.34 0.0549 0.713 0.265 -0.0585  
## # … with abbreviated variable names ¹​`sd(Porcentaje\_Cocoa)`,  
## # ²​`mean(Porcentaje\_Cocoa)`, ³​`sd(Popularidad)`,  
## # ⁴​`cor(Popularidad, Porcentaje\_Cocoa)`

Gráfica de la estadísticas anterior.

ggplot(flavors\_of\_cacao\_V3, aes(Porcentaje\_Cocoa, Popularidad)) +  
 geom\_point() + geom\_smooth(method = lm, se=FALSE)

## `geom\_smooth()` using formula = 'y ~ x'



##Conclusión Los datos de chocolates, tienen en más porcentaje y popularidad los paises 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. Tambien 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 produción o reseñas. Podemos decir que el conjunto de datos esta bien pero faltan datos más cuantitativos, para un análisis más completo.