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 ─────────────────────────────────────── tidyvers
e 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_conf
licts() ──
## ✖ 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/Proyecto
s/Proyecto_Chocolate/flavors_of_cacao.csv")

## Rows: 1795 Columns: 10
## ── Column specification ────────────────────────────────────────────────
─────────
## Delimiter: ","
## chr (5): Creador_empresa_si_exite, Origen_FRIJOL_BARRA, Empresa_locali
dad, 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 th
is message.

View(flavors_of_cacao)

##Datros con clasificación de popularidad

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

## Rows: 1795 Columns: 11
## ── Column specification ────────────────────────────────────────────────
─────────
## Delimiter: ","
## chr (6): Creador_empresa_si_exite, Origen_FRIJOL_BARRA, Empresa_locali
```

```
dad, 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 th
is 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

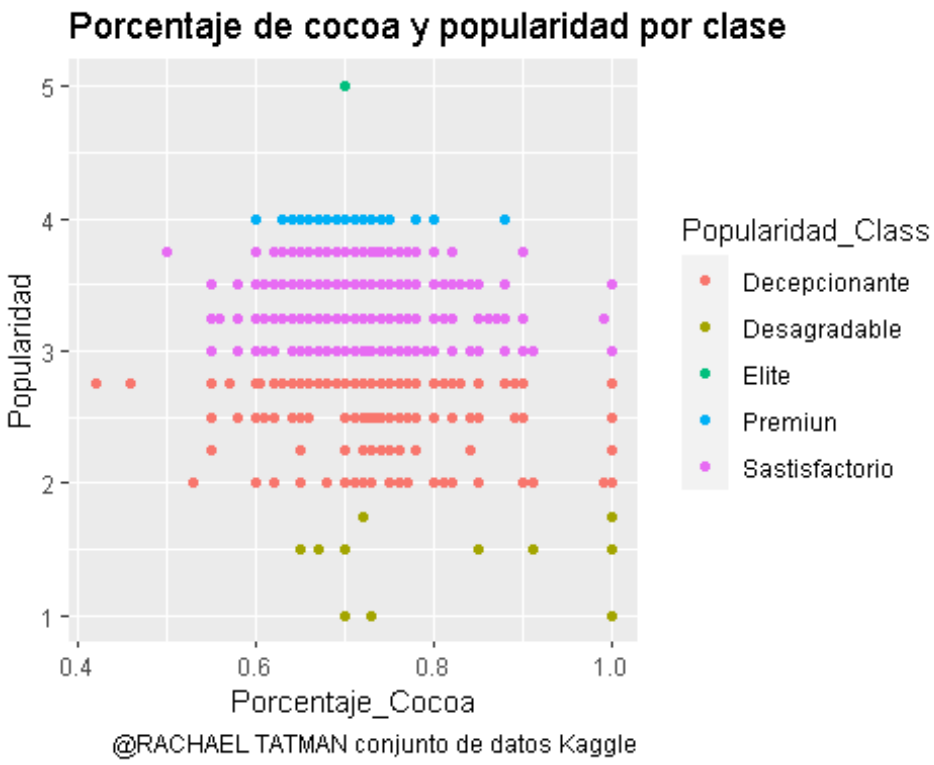
skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100
Id_d	0	1	898.00	518.32	1.00	449.50	898.00	1346.50	1795
REF	0	1	1035.90	552.89	5.00	576.00	1069.00	1502.00	1952
Revisar	0	1	2012.33	2.93	2006.00	2010.00	2013.00	2015.00	2017
Porcentaje_Cocoa	0	1	0.72	0.06	0.42	0.70	0.70	0.75	1
Popularidad	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", "Ak
ata", "Q…
## $ REF                      <dbl> 1876, 1676, 1676, 1680, 1704, 1315, 1
315, 131…
## $ Revisar                  <dbl> 2016, 2015, 2015, 2015, 2015, 2014, 2
014, 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, "C
riollo",…
## $ Haba_origen              <chr> "Sao Tome", "Togo", "Togo", "Togo", "
Peru", "…
## $ Popularidad_Class        <chr> "Sastisfactorio", "Decepcionante", "S
astisfac…

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 Criol
lo Venezu…
## # … with 1 more variable: Popularidad_Class <chr>, and abbreviated var
iable
## #   names ¹Creador_empresa_si_exite, ²Origen_FRIJOL_BARRA, ³Porcentaje
_Cocoa,
## #   ⁴Empresa_localidad, ⁵Popularidad, ⁶Frijo_tipo, ⁷Haba_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_Coco
a,
                                                       y = Popularidad, co
lor =
                                                         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_C
lass
                                                 )))+
  labs(title="Porcentaje de cocoa y conteo color por popularidad Clase",
       caption= "@RACHAEL TATMAN conjunto de datos Kaggle")
```

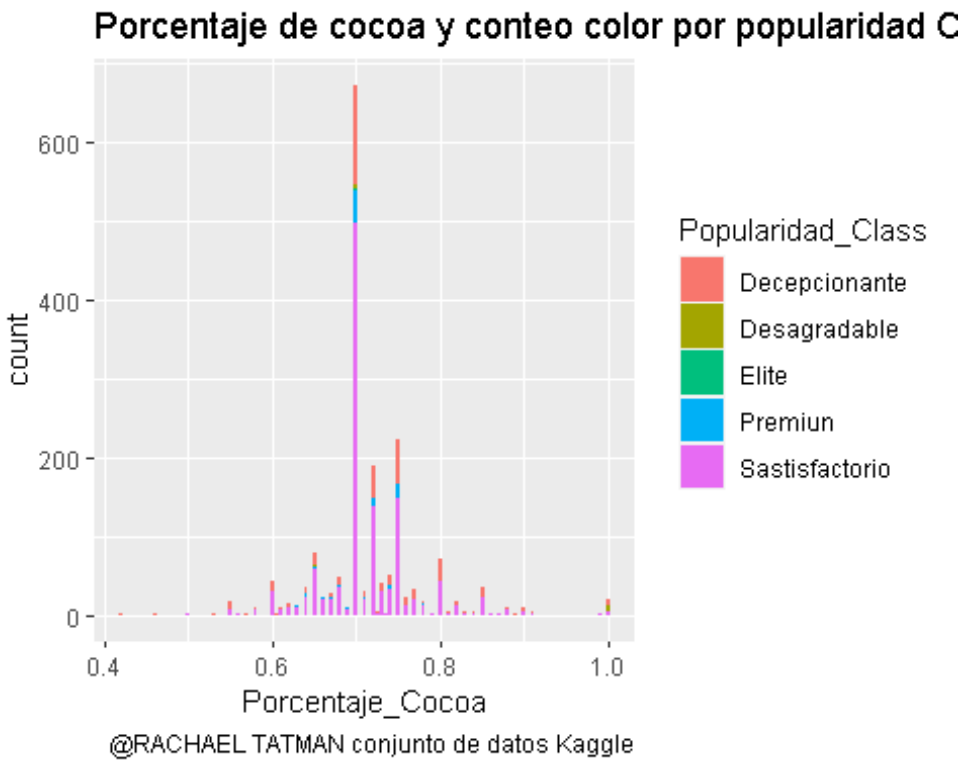
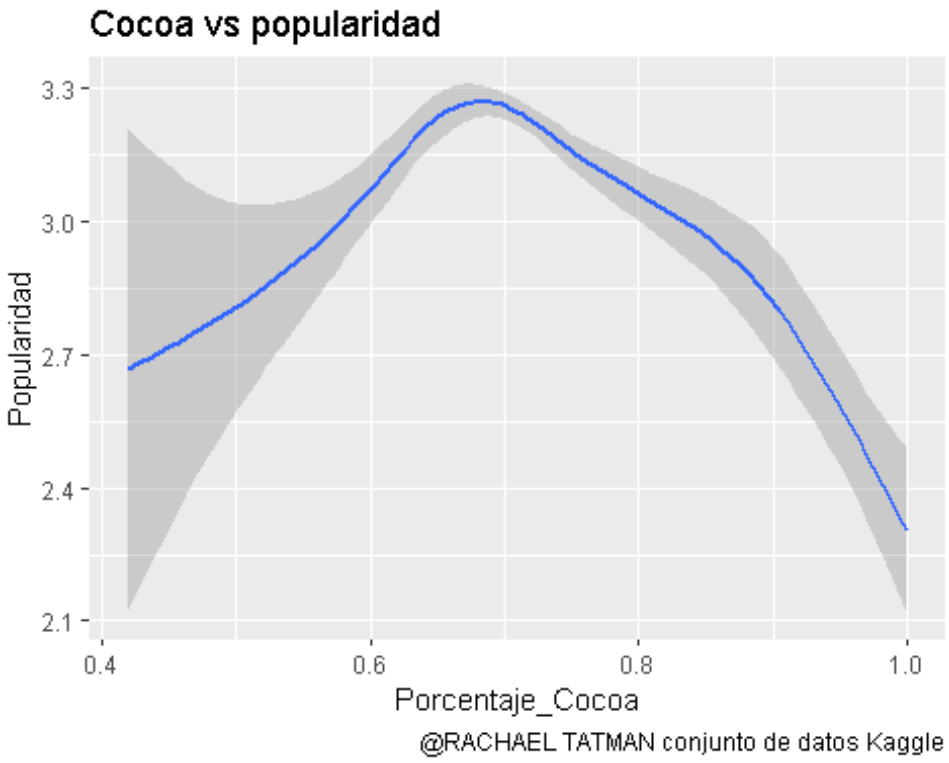


Grafico de porcentaje de cocoa vs popularidad el porcentaje de cocoa en 0.7 la popularidad es la más alta.

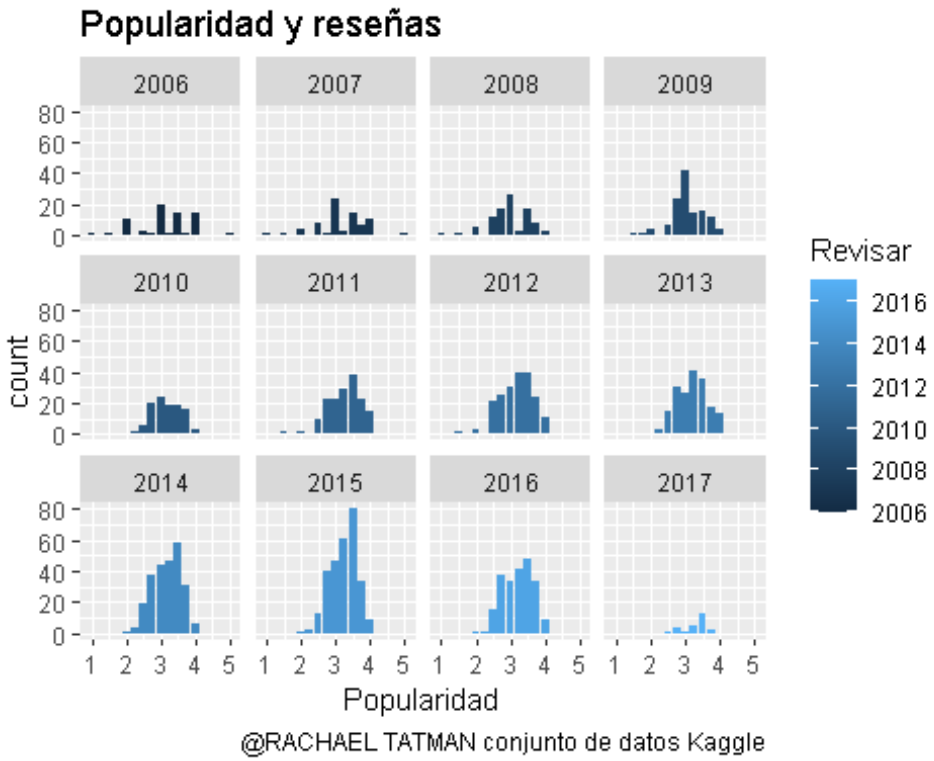
```
ggplot(data = flavors_of_cacao) + geom_smooth((mapping =
                                                 aes(x= Porcentaje_Cocoa,
                                                     y= Popularidad)))+
  labs(title="Cocoa vs popularidad",
       caption= "@RACHAEL TATMAN conjunto de datos Kaggle")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

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



Popularidad y sus reseñas en cuestión del tiempo por gráficos

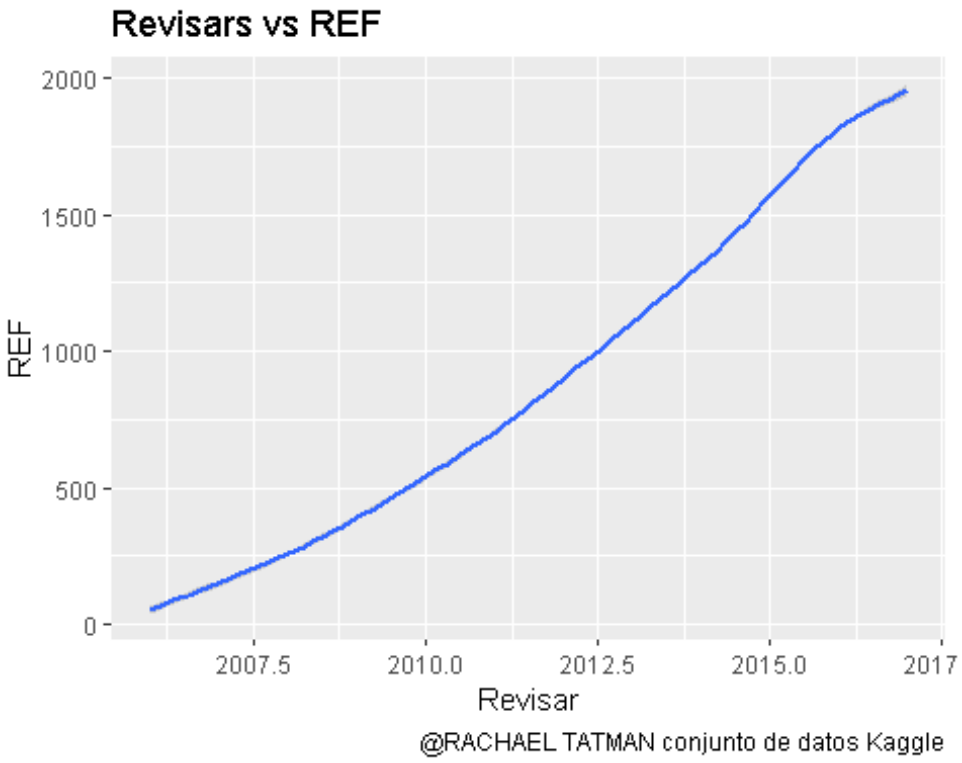
```
ggplot(data = flavors_of_cacao)+
  geom_bar(mapping=aes(x= Popularidad, fill=Revisar))+
  facet_wrap(~Revisar)+
  labs(title="Popularidad y reseñas",
       caption= "@RACHAEL TATMAN conjunto de datos Kaggle")
```



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 = E
mpresa_localidad))+
  geom_smooth(mapping = aes(x = Popularidad,
                                                  y = Porcentaje_Cocoa, col
or = 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 freedo
m.
```

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

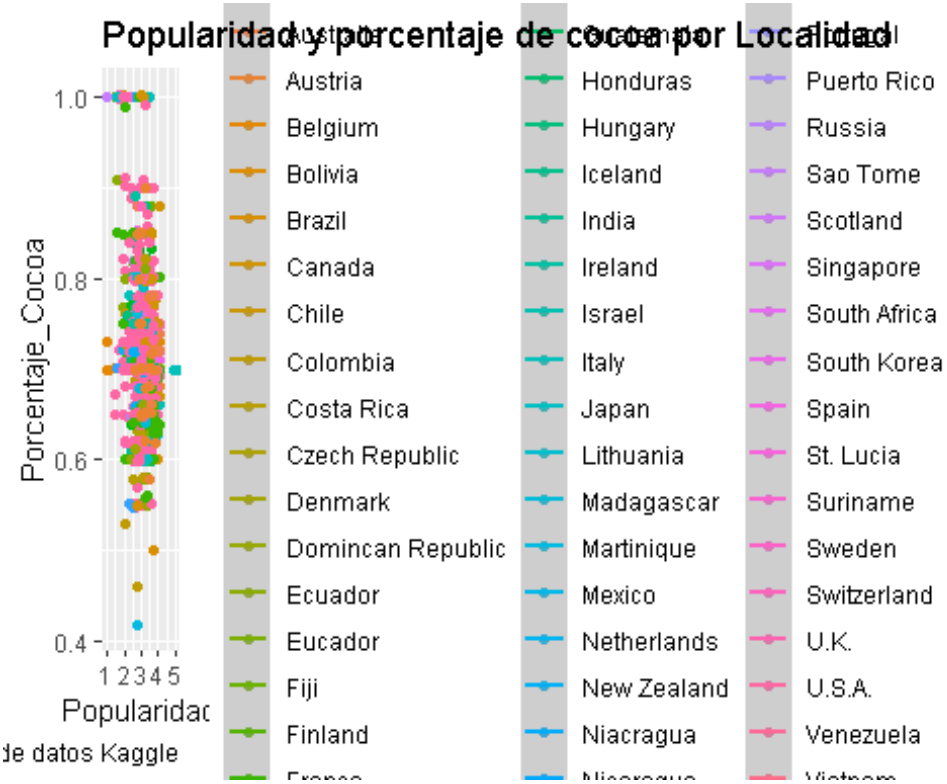
```
## Warning in simpleLoess(y, x, w, span, degree = degree, parametric =
## parametric, : neighborhood radius 0.2525
```

```
## 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 r
adius
## 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 con
dition
## 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 othe
r 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
```

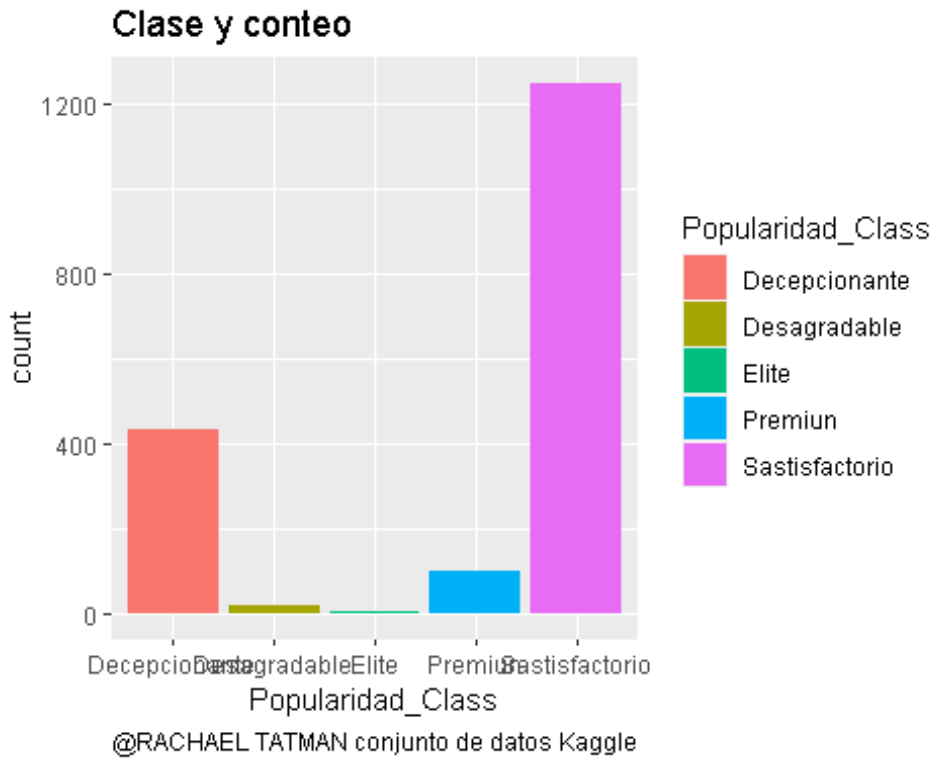
```
## Warning in predLoess(object$y, object$x, newx = if
## (is.null(newdata)) object$x else if (is.data.frame(newdata))
## as.matrix(model.frame(delete.response(terms(object)), : neighborhood r
adius
## 0.755

## 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 con
dition
## 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 othe
r 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 r
adius 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 con
dition
## number 0

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

## 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

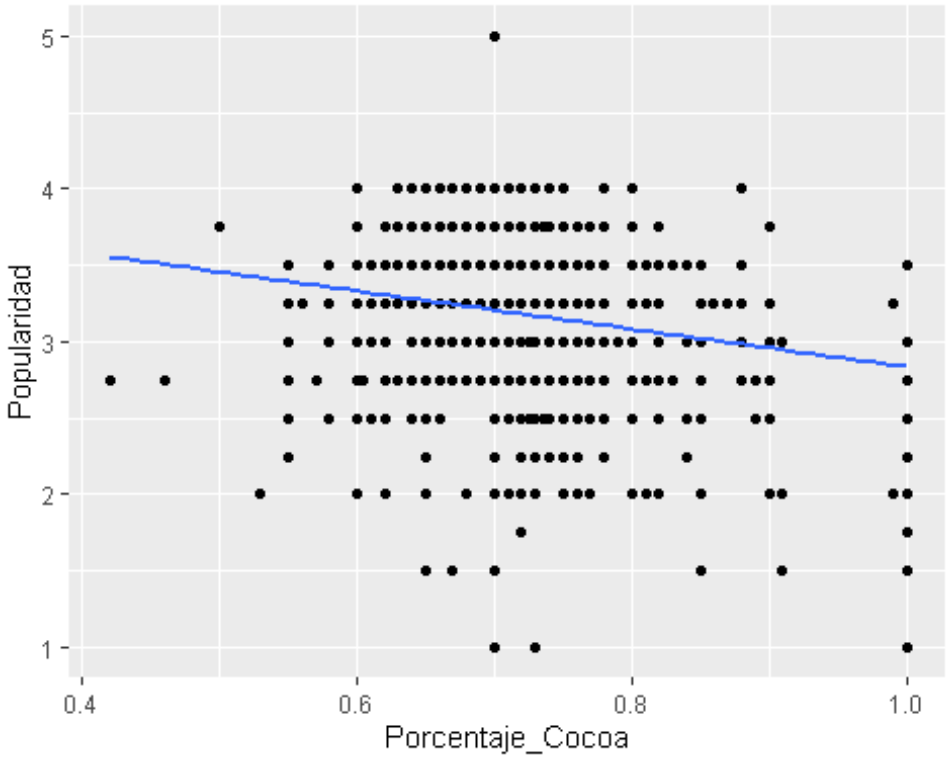
Teneiendo los datos de los chocolates, en cuestión de estadisticas, las columnas de Porcentaje cococa y popularidad no tienen relación alguna, podemos verlos en las siguientes estadísticas y gráficas.

```
flavors_of_cacao_V3 %>%
  group_by(Popularidad_Class) %>%
  summarise(mean(Popularidad), sd(Porcentaje_Cocoa), mean(Porcentaje_Coco
a), 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>   <db
l>   <dbl>
## 1 Decepcionante                    2.60           0.0777   0.725   0.2
17 -0.124
## 2 Desagradable                     1.43           0.149    0.843   0.2
62  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   0.2
65 -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.