

Tipología y ciclo de vida de los datos

PRACTICA2:

En esta práctica se elabora un caso práctico orientado a aprender a identificar los datos relevantes para un proyecto analítico y usar las herramientas de integración, limpieza, validación y análisis de las mismas.

Integrantes: María Augusta Jimbo Granda
María Magdalena Romero Guzmán

Contenido

Resolución	3
1 Descripción del dataset. Por qué es importante y qué pregunta/problema pretende responder?	3
2 Integración y selección de los datos de interés a analizar.....	4
3 Limpieza de los datos.....	5
3.1 ¿Los datos contienen ceros o elementos vacíos?	5
3.1.1 Selección de los datos de interés	7
3.1.2 ¿Cómo gestionarías cada uno de estos casos?	8
3.2 Identificación y tratamiento de valores extremos.....	14
4 Análisis de los datos.....	26
4.1 Selección de los grupos de datos que se quieren analizar/comparar (planificación de los análisis a aplicar).....	26
4.2 Comprobación de la normalidad y homogeneidad de la varianza.....	29
4.3 Aplicación de pruebas estadísticas para comparar los grupos de datos. En función de los datos y el objetivo del estudio, aplicar pruebas de contraste de hipótesis, correlaciones, regresiones, etc.	31
5 Representación de los resultados a partir de tablas y gráficas.....	42
6 Resolución del problema. A partir de los resultados obtenidos, ¿cuáles son las conclusiones? ¿Los resultados permiten responder al problema?.....	45
7 Código.	46
8 Recursos.	47

Resolución

El objetivo de esta actividad será el tratamiento de un dataset:

1 Descripción del dataset. Por qué es importante y qué pregunta/problema pretende responder?

El dataset que vamos a utilizar en la presente práctica es el de clasificación de barras de chocolate a nivel mundial, es decir; el conjunto de datos llamado "flavors_of_cacao.csv"

Contexto:

El chocolate es uno de los dulces más populares en el mundo. Cada año, los residentes de los Estados Unidos colectivamente comen más de 2.8 billones de libras. Sin embargo, no todas las barras de chocolate son iguales! Este conjunto de datos contiene calificaciones de expertos de más de 1.700 barras de chocolate individuales, junto con información sobre su origen regional, el porcentaje de cacao, la variedad de alubia de chocolate utilizada y dónde se cultivaron los frijoles.

Sistema de calificación de sabores de cacao:

5 = Elite (que trasciende los límites normales)

4 = Premium (Desarrollo de sabor superior, carácter y estilo)

3 = Satisfactorio (3.0) a loable (3.75) (bien hecho con cualidades especiales)

2 = decepcionante (pasable, pero contiene al menos un defecto significativo)

1 = Desagradable (principalmente desagradable)

La base de datos se enfoca principalmente en el chocolate negro simple, con el objetivo de apreciar los sabores del cacao cuando se lo convierte en chocolate. Las calificaciones no reflejan los beneficios de salud, las misiones sociales o el estado orgánico.

La diversidad, el equilibrio, la intensidad y la pureza de los sabores son considerados. La genética, el terruño, las técnicas de poscosecha, el procesamiento y el almacenamiento pueden discutirse cuando se considera el componente de sabor.

La textura tiene un gran impacto en la experiencia general y también es posible que los problemas relacionados con la textura afecten el sabor. Es una buena forma de evaluar la visión del fabricante, la atención al detalle y el nivel de competencia.

El dataset contiene 1795 registros sin incluir el encabezado y 9 columnas o atributos que son:

#	Nombre	Descripción	Tipo
1	Empresa	Nombre de la empresa que fabrica la barra	String
2	Geo-región	La geo región de origen específica para la barra	String
3	REF	Un valor vinculado a cuando se ingresó la revisión en la base de datos. Más alto = más reciente	Numeric
4	Fecha de revisión	Fecha de publicación de la revisión	Numeric
5	Porcentaje de cacao	Porcentaje de cacao (oscuridad) de la barra de chocolate que se revisa	Numeric
6	Localización	País base del fabricante	String
7	Calificación	Calificación de expertos	Numeric
8	Tipo de frijol	La variedad de frijol utilizada, en el caso de que se proporcione	String
9	Origen de haba	La amplia geo región de origen para el haba	String

Las preguntas principales que se pueden revisar, analizar y evaluar son, por ejemplo:

- ❖ ¿Dónde se cultivan los mejores granos de cacao?
- ❖ ¿Qué países producen las barras mejor calificadas?
- ❖ ¿Cuál es la relación entre el porcentaje de sólidos de cacao y la calificación?
- ❖ ¿Existe alguna relación entre el porcentaje de sólidos de cacao y el tipo de frijol?

2 Integración y selección de los datos de interés a analizar.

Los datos que se analizarán son:

#	Nombre	Descripción	Tipo
1	Empresa	Nombre de la empresa que fabrica la barra	String
2	Geo-región	La geo región de origen específica para la barra	String
3	Porcentaje de cacao	Porcentaje de cacao (oscuridad) de la barra de chocolate que se revisa	Numeric
4	Localización	País base del fabricante	String
5	Calificación	Calificación de expertos	Numeric
6	Tipo de frijol	La variedad de frijol utilizada, en el caso de que se proporcione	String

Lo anterior, en razón de que parecen ser los más importantes al momento de obtener las conclusiones.

3 Limpieza de los datos.

Antes de realizar la limpieza, se cambió la cabecera del archivo original, con nombres más cortos para mejorar la comprensión de los mismos.

3.1 ¿Los datos contienen ceros o elementos vacíos?

Para revisar si existen elementos vacíos se realiza la lectura del archivo en R en formato csv con la siguiente línea de código, en el setwd se coloca la ruta donde se encuentra el archivo para realizar su lectura, con el read.csv se realiza la lectura y con head se revisa los primeros 10 registros:

```
> #Lectura del archivo de chocolate sin ningun cambio
> setwd("C:/")
> cacao <- read.csv(file = "flavors_of_cacao.csv", header=TRUE)
> head(cacao, 10)
  Empresa  Geo.region  REF Fecha.de.revision Porcentaje.de.cacao
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%
7 A. Morin       Cuba 1315        2014          70%
8 A. Morin  Sur del Lago 1315        2014          70%
9 A. Morin Puerto Cabello 1319        2014          70%
10 A. Morin     Pablino 1319       2014          70%
  Localizacion Calificacion Tipo.de.frijol Origen.de.haba
1 France           3.75      Á     Sao Tome
2 France           2.75      Á      Togo
3 France           3.00      Á      Togo
4 France           3.50      Á      Togo
5 France           3.50      Á      Peru
6 France          2.75  Criollo Venezuela
7 France           3.50      Á      Cuba
8 France           3.50  Criollo Venezuela
9 France           3.75  Criollo Venezuela
10 France          4.00      Á      Peru
> |
```

Como se puede observar los datos si contienen elementos vacíos. Se reemplazó los elementos vacíos del archivo flavors_of_cacao.csv, que se los representaba con Á por una cadena vacía para realizar un mejor tratamiento de los mismos. El archivo con el cambio es flavors_of_cacao1.csv

Ahora se realiza la lectura del siguiente archivo, donde se puede observar que los caracteres especiales ya están representados por una cadena vacía:

```
> cacao <- read.csv(file = "flavors_of_cacao1.csv", header=TRUE)
> head(cacao, 10)
  Empresa    Geo.region  REF Fecha.de.revision Porcentaje.de.cacao
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%
7 A. Morin      Cuba 1315      2014           70%
8 A. Morin  Sur del Lago 1315      2014           70%
9 A. Morin Puerto Cabello 1319      2014           70%
10 A. Morin     Pablino 1319     2014           70%
  Localizacion Calificacion Tipo.de.frijol Origen.de.haba
1       France          3.75            Sao Tome
2       France          2.75             Togo
3       France          3.00             Togo
4       France          3.50             Togo
5       France          3.50             Peru
6       France          2.75            Venezuela
7       France          3.50             Cuba
8       France          3.50            Venezuela
9       France          3.75            Venezuela
10      France          4.00             Peru
> |
```

Entonces se reemplaza los vacíos por NA (Not Available) para que R los reconozca como desconocidos y poder tratarlos:

```
> #ceros y elementos vacíos, se les coloca NA a todos los elementos vacíos
> for (i in 1:nrow(cacao)){
+ #print(cacao[i,8])
+   for(j in 1:9){
+     if (cacao[i,j] == ""){
+       cacao[i,j] <- NA
+     }
+   }
+ }
> head(cacao, 10)
  Empresa    Geo.region  REF Fecha.de.revision Porcentaje.de.cacao
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%
7 A. Morin      Cuba 1315      2014           70%
8 A. Morin  Sur del Lago 1315      2014           70%
9 A. Morin Puerto Cabello 1319      2014           70%
10 A. Morin     Pablino 1319     2014           70%
  Localizacion Calificacion Tipo.de.frijol Origen.de.haba
1       France          3.75            <NA>        Sao Tome
2       France          2.75            <NA>         Togo
3       France          3.00            <NA>        Togo
4       France          3.50            <NA>         Togo
5       France          3.50            <NA>         Peru
6       France          2.75            Criollo      Venezuela
7       France          3.50            <NA>         Cuba
8       France          3.50            Criollo      Venezuela
9       France          3.75            Criollo      Venezuela
10      France          4.00            <NA>         Peru
> |
```

Se verifica el tipo de dato asignado a cada variable, y el número de registros que son 1795:

```
> #Para ver el tipo de dato asignado a cada campo
> sapply(cacao, function(x) class(x))
  Empresa           Geo.region          REF  Fecha.de.revision
  "factor"         "factor"           "integer"        "integer"
Porcentaje.de.cacao    Localizacion      Calificacion     Tipo.de.frijol
  "factor"         "factor"           "numeric"        "factor"
  Origen.de.haba
  "factor"
> nrow(cacao)
[1] 1795
> |
```

3.1.1 Selección de los datos de interés

Se elimina las columnas 3, 4 y 7 correspondientes a REF (referencia de revisión de la barra de chocolate), Fecha de revisión y Origen de haba, porque las dos primeras no nos dan mucha información para el análisis que se desea hacer y la última se la elimina porque ya se tiene el otro campo que es Geo-region para determinar el origen específico de la barra de chocolate:

```
> #Selección de los datos a analizar
> #se elimina el campo 3 y 4 que no es de mucha utilidad para el análisis
> cacao <- cacao[,-(3:4)]
> #luego se elimina la columna 7 que en este caso en la primera carga era la
> #columna 9 de Origen del haba, se la elimina porque ya tenemos la columna 2
> # que es el origen de la geo-region de origen específica para la barra
> cacao <- cacao[,-7]
> head(cacao, 10)
  Empresa           Geo.region Porcentaje.de.cacao Localizacion Calificacion
1 A. Morin       Agua Grande      63%       France      3.75
2 A. Morin        Kpime        70%       France      2.75
3 A. Morin       Atsane        70%       France      3.00
4 A. Morin        Akata        70%       France      3.50
5 A. Morin        Quilla        70%       France      3.50
6 A. Morin      Carenero        70%       France      2.75
7 A. Morin        Cuba        70%       France      3.50
8 A. Morin   Sur del Lago      70%       France      3.50
9 A. Morin Puerto Cabello      70%       France      3.75
10 A. Morin      Pablino       70%       France      4.00
  Tipo.de.frijol
1 <NA>
2 <NA>
3 <NA>
4 <NA>
5 <NA>
6 Criollo
7 <NA>
8 Criollo
9 Criollo
10 <NA>
> |
```

3.1.2 ¿Cómo gestionarías cada uno de estos casos?

Como en el conjunto de datos que se está analizando no existen ceros, entonces se verifica cuantos casos existen con elementos desconocidos, que son 888 en el Tipo de frijol:

```
> #se comprueba los elementos vacíos en este caso, porque ceros no existen
> #se puede observar que existen 888 elementos vacíos
> sapply(cacao, function(x) sum(is.na(x)))
      Empresa          Geo.region Porcentaje.de.cacao      Localizacion
      0                  0                  0                  0
Calificacion        Tipo.de.frijol
      0                  888
> |
```

Para gestionar estos casos se va a utilizar el método basado en k vecinos más próximos (KNN-imputation), ya que se trata del tipo de frijol que se da en determinada región, entonces con el k vecinos obtiene los datos más cercanos, ya que eliminar estos registros no sería tan conveniente porque los necesitamos para realizar el análisis respectivo y se perderían bastantes registros, para ello se instala la librería VIM:

```
> install.packages("VIM")
Installing package into 'C:/Users/Magdalena/Documents/R/win-library/3.4'
(as 'lib' is unspecified)
--- Please select a CRAN mirror for use in this session ---
also installing the dependencies 'sp', 'laeken'

probando la URL 'https://cran.rediris.es/bin/windows/contrib/3.4/sp_1.2-7.zip'
Content type 'application/zip' length 1538524 bytes (1.5 MB)
downloaded 1.5 MB

probando la URL 'https://cran.rediris.es/bin/windows/contrib/3.4/laeken_0.4.6.zip'
Content type 'application/zip' length 2764908 bytes (2.6 MB)
```

Y se carga la librería:

```
> library(VIM)
Loading required package: colorspace
Loading required package: grid
Loading required package: data.table
data.table 1.10.4.3
  The fastest way to learn (by data.table authors): https://www.datacamp.com/co...
  Documentation: ?data.table, example(data.table) and browseVignettes("data.tab...
  Release notes, videos and slides: http://r-databasetable.com
VIM is ready to use.
  Since version 4.0.0 the GUI is in its own package VIMGUI.

  Please use the package to use the new (and old) GUI.

Suggestions and bug-reports can be submitted at: https://github.com/alexkowa/VIM

Attaching package: 'VIM'

The following object is masked from 'package:datasets':

  sleep

Warning message:
package 'VIM' was built under R version 3.4.4
> |
```

Y se reemplaza los valores desconocidos por el vecino más cercano y se comprueba que ya no existan:

```
> #vamos a reemplazar los elementos vacíos por el vecino más cercano
> #para trabajar con datos y no elementos vacíos
> cacao$Tipo.de.frijol <- kNN(cacao)$Tipo.de.frijol
> sapply(cacao, function(x) sum(is.na(x)))
      Empresa          Geo.region Porcentaje.de.cacao  Localizacion
1        0                   0                  0                 0
      Calificacion      Tipo.de.frijol
1        0                   0
> sapply(cacao, function(x) sum(is.nan(x)))
      Empresa          Geo.region Porcentaje.de.cacao  Localizacion
1        0                   0                  0                 0
      Calificacion      Tipo.de.frijol
1        0                   0
> |
```

Se verifica en los datos que se hayan reemplazado:

```
> head(cacao, 10)
      Empresa          Geo.region Porcentaje.de.cacao  Localizacion Calificacion
1 A. Morin       Agua Grande      63%     France    3.75
2 A. Morin           Kpime       70%     France    2.75
3 A. Morin         Atsane       70%     France    3.00
4 A. Morin         Akata        70%     France    3.50
5 A. Morin         Quilla       70%     France    3.50
6 A. Morin       Carenero      70%     France    2.75
7 A. Morin          Cuba        70%     France    3.50
8 A. Morin   Sur del Lago      70%     France    3.50
9 A. Morin     Puerto Cabello     70%     France    3.75
10 A. Morin       Pablino      70%     France    4.00
      Tipo.de.frijol
1           Criollo
2           Trinitario
3           Criollo
4 Criollo, Trinitario
5           Trinitario
6           Criollo
7           Criollo
8           Criollo
9           Criollo
10 Criollo, Trinitario
> |
```

Antes de identificar los valores extremos se realizó un análisis de los datos en los cuales se encontró algunas inconsistencias, por lo cual se realiza la limpieza de los mismos, en cuanto a nombres, caracteres especiales, etc.

Se empieza con el campo Empresa, donde se carga la librería car para realizar la codificación de los valores encontrados, y se comprueba que se hayan reemplazado correctamente:

```
> library(car)
> #####reemplazos en campo Empresa
> #se reemplaza el nombre de la empresa Naive que en 3 registros se encuentra
> #con caracteres especiales, y dos mas con nombres mal
> cacao$Empresa <- recode(cacao$Empresa, "Naiive"="Naive";
+ "Cacao de Origin"="Cacao de Origen"; "Shattell"="Shattel"; )
> print(cacao[1166,1])
[1] Naive
413 Levels: A. Morin Acalli Adi Aequare (Gianduja) ... Zotter
> #se comprueba que se haya reemplazado Cacao de Origin
> print(cacao[296,1])
[1] Cacao de Origen
413 Levels: A. Morin Acalli Adi Aequare (Gianduja) ... Zotter
> #se comprueba que se haya reemplazado Shattell
> print(cacao[1456,1])
[1] Shattel
413 Levels: A. Morin Acalli Adi Aequare (Gianduja) ... Zotter
> |
```

Se reemplaza los ' porque da problemas para recodificar a los valores más adelante:

```
> #se reemplaza tambien los ' a los campos Empresa, porque da
> #problemas pra recodificar a los registros
> cacao$Empresa <- gsub("'", "", cacao$Empresa)
> #se comprueba las lineas 466,620, por ejemplo que ya no tenga el apostrofe
> print(cacao[466,1])
[1] "Cote d Or (Kraft)"
> print(cacao[620,1])
[1] "Emilys"
> |
```

También se reemplaza los ' en los valores de la variable Geo.region:

```
> #####se reemplaza en la Geo-region los '
> cacao$Geo.region <- gsub("'", "", cacao$Geo.region)
> #se comprueba el reemplazo
> print(cacao[795,2])
[1] "Maunawili, Oahu, Agri Research C., 2014"
> |
```

Se reemplaza los " en los valores de la variable Geo.region, porque también da problemas en la recodificación:

```
> #se reemplaza en la Geo-region el "
> cacao$Geo.region <- gsub('"', "", cacao$Geo.region)
> #se comprueba que se haya reemplazado
> print(cacao[474,])
      Empresa          Geo.region Porcentaje.de.cacao Localizacion
474    Creo heirloom, Arriba Nacional           85%       U.S.A.
      Calificacion      Tipo.de.frijol
474        3.25 Forastero (Nacional)
> |
```

Se reemplaza los ; por , en los valores de la variable Geo.region, porque también da problemas en la recodificación:

```
> #se reemplaza en la Geo-region el ;
> cacao$Geo.region <- gsub(';',",",cacao$Geo.region)
> #se comprueba que se haya reemplazado
> print(cacao[292,])
      Empresa          Geo.region Porcentaje.de.cacao Localizacion
292 Cacao de Origen Agua Fria, Sucre region           75%     Venezuela
      Calificacion Tipo.de.frijol
292                2.5      Trinitario
> |
```

Se reemplaza los % en los valores del campo Porcentaje.de.cacao, para poder tratar los datos y convertirlos a numéricos más adelante:

```
> #se reemplaza en la Porcentaje.de.cacao el %, para poder convertir
> #a numerico los valores
> cacao$Porcentaje.de.cacao <- gsub('%',"",cacao$Porcentaje.de.cacao)
> print(cacao[,3])
 [1] "63"   "70"   "70"   "70"   "70"   "70"   "70"   "70"   "70"   "70"
 [11] "70"   "70"   "70"   "70"   "70"   "70"   "70"   "70"   "70"   "70"
 [21] "63"   "70"   "63"   "70"   "70"   "60"   "80"   "88"   "72"   "55"
 [31] "70"   "70"   "75"   "75"   "65"   "75"   "75"   "75"   "75"   "75"
 [41] "70"   "70"   "70"   "60"   "60"   "60"   "60"   "60"   "60"   "60"
 [51] "60"   "80"   "60"   "60"   "70"   "70"   "70"   "70"   "70"   "70"
 [61] "70"   "70"   "70"   "70"   "85"   "85"   "72"   "73"   "64"
 [71] "66"   "75"   "63"   "70"   "68"   "70"   "75"   "70"   "70"   "70"
 [81] "70"   "70"   "70"   "70"   "70"   "63"   "70"   "66"   "75"   "85"
 [91] "50"   "75"   "60"   "75"   "75"   "75"   "72"   "75"   "75"   "70"
 [101] "70"   "73"   "70"   "70"   "70"   "70"   "70"   "70"   "70"   "70"
 [111] "70"   "73"   "70"   "68"   "70"   "70"   "70"   "70"   "75"   "70"
 [121] "75"   "72"   "72"   "72"   "100"  "72"   "72"   "72"   "72"   "72"
 [131] "75"   "72"   "72"   "80"   "75"   "72"   "72"   "68"   "72"   "72"
 [141] "70"   "77"   "75"   "70"   "80"   "70"   "70"   "70"   "70"   "70"
 [151] "70"   "70"   "70"   "70"   "70"   "70"   "80"   "65"   "70"   "65"
 [161] "73"   "72"   "80"   "70"   "70"   "90"   "64"   "64"   "64"   "71"
 [171] "70"   "70"   "70"   "83"   "78"   "83"   "74"   "74"   "74"   "73"
 [181] "72"   "72"   "55"   "64"   "88"   "72"   "72"   "70"   "74"   "64"
 [191] "72"   "76"   "76"   "78"   "86"   "72"   "75"   "70"   "65"   "70"
 [201] "78"   "75"   "65"   "75"   "65"   "71"   "75"   "68"   "70"   "70"
 [211] "70"   "70"   "70"   "82"   "72"   "82"   "75"   "75"   "75"   "70"
 [221] "70"   "75"   "75"   "65"   "75"   "75"   "75"   "75"   "75"   "75"
 [231] "75"   "75"   "75"   "75"   "75"   "75"   "75"   "75"   "75"   "75"
 [241] "75"   "75"   "75"   "75"   "75"   "100"  "75"   "75"   "77"   "100"
 [251] "70"   "70"   "70"   "70"   "68"   "70"   "70"   "72"   "70"   "75"
 [261] "85"   "60"   "80"   "70"   "80"   "80"   "60"   "70"   "72"   "70"
 [271] "72"   "68"   "70"   "68"   "72"   "72"   "72"   "72"   "60"   "70"
 [281] "75"   "75"   "75"   "75"   "65"   "70"   "75"   "72"   "66"   "77"
 [291] "75"   "75"   "74"   "75"   "70"   "74"   "71"   "74"   "72"   "64"
```

Se reemplaza unos nombres de la variable Geo.region que no se encuentran correctamente:

```
> #se reemplaza en la Geo-region el * y unos nombres que no se encuentran bien
> cacao$Geo.region <- recode(cacao$Geo.region, "Concepcion"="Concepcion";
+ "Capistrano"="Capistrano"; "Equateur"="Ecuador"; "Ambolikapiky P."="Ambolikapi$"
+ "Ambolikapiky P."="Ambolikapiky"; "Alto Beni, Palos Blanco"="Alto Beni, Palos $"
+ "Chiapan"="Chiapas"; "Brazilian"="Brazil"; "Bolivian"="Bolivia"; "Colombie"="Col$"
+ "Colombian"="Colombia"; "Dominican Republicm, rustic"="Dominican Republic, russ"
+ "Fazenda Sempre Firme P., Bahia"="Fazenda Sempre Firme, Bahia";
+ "La Red, Guanconjeco"="La Red, Guaconejo"; "Madagared"="Madagascar";
+ "Monte Alegre, Diego Badero"="Monte Alegre, D. Badero"; "Nicaraqua"="Nicaragua$"
+ "Trinidad-Tobago"="Trinidad & Tobago"; "Venezuela"="Venezuela";
+ "Wild Bolivian"="Wild Bolivia"; "Ba Lai"="Ba Ria"; "Caraibe"="Caribbean";
+ "Dominican"="Dominican Republic"; "Elvesia"="Elvesia P.";')
> #se comprueba que se haya reemplazado el *
> print(cacao[1642,])
  Empresa Geo.region Porcentaje.de.cacao Localizacion Calificacion
1642  Tejas Concepcion           80      U.S.A.          3
  Tipo.de.frijol
1642    Trinitario
> #comprobar que se haya reemplazado el Equateur
> print(cacao[376,])
  Empresa Geo.region Porcentaje.de.cacao Localizacion Calificacion
376  Cemoi   Ecuador            72      France        2.75
  Tipo.de.frijol
376    Trinitario
> |
```

Se elimina el registro 246 porque contiene un nombre de una región que no existe:

```
> #se elimina el registro 246 porque contiene una region con el nombre:
> #One Hundred que no existe
> print(cacao[246,])
  Empresa Geo.region Porcentaje.de.cacao Localizacion Calificacion
246  Bonnat One Hundred         100      France        1.5
  Tipo.de.frijol
246    Forastero
> cacao <- cacao[-246,]
> print(cacao[246,])
  Empresa Geo.region Porcentaje.de.cacao Localizacion Calificacion
247  Bonnat   Ceylan            75      France          3
  Tipo.de.frijol
247    Criollo
> |
```

También se elimina el registro 779 porque contiene un nombre de una región que no existe:

```
> #se elimina el registro 779 porque contiene una region con el nombre:
> #One Hundred que no existe
> print(cacao[779,])
  Empresa Geo.region Porcentaje.de.cacao Localizacion Calificacion
780 Habitual one hundred        100      Canada          2
  Tipo.de.frijol
780 Forastero (Arriba)
> cacao <- cacao[-779,]
> print(cacao[779,])
  Empresa Geo.region Porcentaje.de.cacao Localizacion Calificacion
781 Hachez   Arriba             77      Germany        2.5
  Tipo.de.frijol
781 Forastero (Arriba)
> |
```

También se elimina el registro 1410 porque contiene un nombre de una región que no existe:

```
> #se elimina el registro 1410 porque contiene una region con el nombre:  
> #100 percent que no existe  
> print(cacao[1410,])  
    Empresa Geo.region Porcentaje.de.cacao Localizacion Calificacion  
1412 S.A.I.D. 100 percent          100      Italy      1.5  
    Tipo.de.frijol  
1412     Forastero  
> cacao <- cacao[-1410,]  
> print(cacao[1410,])  
    Empresa Geo.region Porcentaje.de.cacao Localizacion Calificacion  
1413 S.A.I.D.     Samana          70      Italy      3  
    Tipo.de.frijol  
1413     Criollo  
> |
```

El número de registros ahora es de 1792:

```
> #revisa el numero de registros que quedaron: 1792  
> nrow(cacao)  
[1] 1792  
> |
```

Se reemplaza unos nombres en la variable Tipo.de.frijol:

```
> #####reemplazos en Tipo.de.frijol  
> cacao$Tipo.de.frijol <- recode(cacao$Tipo.de.frijol,"Criollo, +=""Criollo";  
+ "Forastero (Arriba) ASSS""Forastero (Arriba) ASS";')  
> #comprobar que se haya reemplazado Criollo, +  
> print(cacao[767,6])  
[1] Criollo  
38 Levels: Amazon Amazon mix Amazon, ICS Beniano ... Trinitario, TCGA  
> #comprobar que se haya reemplazado Forastero (Arriba) ASSS  
> print(cacao[1418,6])  
[1] Forastero (Arriba) ASS  
38 Levels: Amazon Amazon mix Amazon, ICS Beniano ... Trinitario, TCGA  
> |
```

Se reemplaza unos nombres en la variable Localización:

```
> #se cambia las localizaciones que estan mal el nombre  
> cacao$Localizacion <- recode(cacao$Localizacion,"Ecuador""Ecuador";  
+ "Niacragua""Nicaragua";"Dominican Republic""Dominican Republic";')  
> #se comprueba la de Ecuador que se encuentre correcto  
> print(cacao[1360,4])  
[1] Ecuador  
58 Levels: Amsterdam Argentina Australia Austria Belgium Bolivia ... Wales  
> #se comprueba la de Nicaragua que se encuentre correcto  
> print(cacao[1190,4])  
[1] Nicaragua  
58 Levels: Amsterdam Argentina Australia Austria Belgium Bolivia ... Wales  
> #se comprueba la de Dominican Republic que se encuentre correcto  
> print(cacao[884,4])  
[1] Dominican Republic  
58 Levels: Amsterdam Argentina Australia Austria Belgium Bolivia ... Wales  
> |
```

3.2 Identificación y tratamiento de valores extremos.

Los valores extremos son aquellos que parecen no ser congruentes si los comparamos con el resto de los datos.

Los valores extremos pueden causar serios problemas para los análisis estadísticos. Primero, generalmente sirven para aumentar la varianza del error y reducir el poder de las pruebas estadísticas. En segundo lugar, si no se distribuye aleatoriamente, pueden alterar sustancialmente las probabilidades de cometer errores tipo I y tipo II. En tercer lugar, pueden sesgar o influir seriamente en las estimaciones que pueden ser de interés sustancial porque pueden no ser generadas por la población de interés.

Para determinar los valores extremos se realiza la discretización de las variables tipo vector a numéricas:

Se discretiza la variable Empresa, no se captura todo porque son 413 Empresas que se las puede ver en el código del punto 7:

```
> #####DISCRETIZACION DE VARIABLES A NUMERICAS#####
> #se discretiza la variable Empresa a valores numéricos
> cacao$Empresa <- recode(cacao$Empresa, "A. Morin"=1;
+ "Acalli"=2;
+ "Adi"=3;
+ "Aequare (Gianduja)"=4;
+ "Ah Cacao"=5;
+ "Akessons (Pralus)"=6;
+ "Alain Ducasse"=7;
+ "Alexandrie"=8;
+ "Altus aka Cao Artisan"=9;
+ "Amano"=10;
+ "Amatller (Simon Coll)"=11;
+ "Amazona"=12;
+ "Ambrosia"=13;
+ "Amedei"=14;
+ "AMMA"=15;
+ "Anahata"=16;
+ "Animas"=17;
+ "Ara"=18;
+ "Arete"=19;
+ "Artisan du Chocolat"=20;
+ "Artisan du Chocolat (Casa Luker)"=21;
+ "Askinosie"=22;
+ "Bahan & Co."=23;
+ "Bakau"=24;
+ "Bar Au Chocolat"=25;
+ "Baravellis"=26;
+ "Batch"=27;
+ "Beau Cacao"=28;
+ "Beehive"=29;
+ "Belcolade"=30;
+ "Bellflower"=31;
+ "Belyzium"=32;
+ "Benoit Nihant"=33;
```

Se comprueba que se hayan transformado todos los valores a numéricos:

```
> #se comprueba que se hayan transformado todos los valores a numericos
> #de la variable Empresa
> print(cacao[,1])
 [1]  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1
[19]  1  1  1  1  1  2  2  3  3  3  3  3  4  4  5  6  6  6  6  7
[37]  7  7  7  7  8  8  8  9  9  9  9  9  9  9  9  9  9  9  9  9  9  9  9
[55] 10 10 10 10 10 10 10 10 11 11 11 11 11 12 12 13 13 13
[73] 13 13 13 14 14 14 14 14 14 14 14 14 14 14 14 14 14 14 15 15
[91] 15 15 15 16 17 18 18 18 19 19 19 19 19 19 19 19 19 19 19 19
[109] 19 19 19 19 19 19 19 19 19 19 19 19 19 19 19 19 20 20 20 20
[127] 20 20 20 20 20 20 20 20 20 20 20 20 20 21 22 22 22 22 22 22
[145] 23 23 23 23 23 24 24 25 25 25 25 25 26 27 27 27 27 28 28
[163] 29 29 29 29 30 30 30 30 31 31 31 31 32 32 32 33 33 33 33
[181] 33 33 34 35 35 35 35 35 35 35 35 36 36 36 36 36 37 37 37
[199] 37 37 37 37 37 37 37 37 37 37 37 38 38 38 39 40 40 41
[217] 41 41 41 41 41 42 42 42 42 42 42 42 42 42 42 42 42 42 42
[235] 42 42 42 42 42 42 42 42 42 42 42 42 42 43 43 43 44 44 45
[253] 45 45 45 45 45 45 45 45 46 46 46 46 46 46 46 47 47 48 48
[271] 48 49 50 51 51 51 51 52 53 53 53 53 53 54 54 54 54 54 54
[289] 55 55 55 55 55 55 55 56 56 56 56 56 56 56 56 57 58 58 58
[307] 58 59 59 59 59 59 59 59 59 59 60 60 60 61 62 62 62 62
[325] 63 63 64 64 64 64 64 64 65 65 65 65 65 66 66 66 66 66
[343] 66 66 67 67 68 68 68 68 69 69 69 69 69 70 70 70 70 70
[361] 70 70 70 70 70 70 70 70 71 71 71 71 72 73 74 75
[379] 75 76 77 78 79 79 80 80 81 81 81 82 82 83 83 84 84 84
[397] 84 84 84 84 85 86 86 86 87 87 87 87 87 88 88 89 90 91
[415] 91 91 91 91 91 91 92 92 93 93 93 93 93 94 95 96 97
[433] 98 99 99 99 100 101 102 102 102 102 102 103 104 105 105 105 105
[451] 105 105 105 105 105 105 105 105 105 105 105 105 105 106 107 107 107
[469] 107 107 107 107 108 108 109 109 110 111 111 111 112 112 112 112 112
[487] 112 112 112 112 112 112 112 112 112 113 113 113 113 113 113 113 113
[505] 113 113 114 114 114 115 115 115 115 116 116 116 116 117 117 117 118 118
[523] 119 119 119 119 119 119 120 120 120 121 122 122 122 123 123 123 123 123
[541] 123 123 123 123 123 123 123 124 124 124 124 125 126 126 127 127 127 127
[559] 127 127 127 127 127 127 127 127 127 127 127 127 127 127 127 127 127 127
```

Se discretiza la variable Geo.region, no se captura todo porque son 1012 Geo regiones que se las puede ver en el código del punto 7:

```
|> #se discretiza la variable Geo.region a valores n mericos
|> cacao$Geo.region <- recode(cacao$Geo.region, "heirloom, Arriba Nacional"=1;
|+ "2009 Hapa Nibby"=2;
|+ "A case of the Xerxes Blues, triple roast"=3;
|+ "Abinao"=4;
|+ "ABOCFA Coop"=5;
|+ "Abstract S. w/ Jamaica nibs,batch abs60323.0"=6;
|+ "Acarigua, w/ nibs"=7;
|+ "Acopagro"=8;
|+ "Acul-du-Nord, 2015"=9;
|+ "Africa"=10;
|+ "Africa meets Latina"=11;
|+ "AgroCriso Plantation"=12;
|+ "Agua Fria, Sucre region"=13;
|+ "Agua Grande"=14;
|+ "Akata"=15;
|+ "Akesson Estate"=16;
|+ "Akesson P."=17;
|+ "Akessons E., Sambirano V."=18;
|+ "Akessons Estate"=19;
|+ "Akessons Estate, Sambirano, 2013"=20;
|+ "Akessons Estate, Sambirano, Ambanja"=21;
|+ "Akessons, batch 4411"=22;
|+ "Akosombo"=23;
|+ "Almendra Blanca, batch 1004"=24;
|+ "Alpaco"=25;
|+ "Alta Verapaz, 2014"=26;
|+ "Alto Beni"=27;
|+ "Alto Beni, Covendo Region"=28;
|+ "Alto Beni, Cru Savage"=29;
|+ "Alto Beni, Palos Blancos"=30;
|+ "Alto Beni, Upper Rio Beni, 2014"=31;
|+ "Alto Beni, Upper Rio Beni, 2015"=32;
|+ "Alto Beni, Wild Bolivian"=33;
|+ "Alto Beni, Wild Harvest, Itenez R. 24hr c."=34;
```

Se comprueba que se hayan transformado todos los valores:

```
> #se comprueba que se hayan transformado todos los valores a numerosicos
> #de la variable Geo.region
> print(cacao[,2])
 [1] 14 476 64 15 789 166 275 899 781 707 712 547 129 321
[15] 229 111 720 203 747 189 189 125 730 219 964 977 979 977
[29] 978 536 536 906 84 550 638 954 986 547 203 755 1004 485
[43] 930 560 633 8 225 989 387 899 237 125 125 730 641 313
[57] 400 203 634 85 547 281 675 363 321 321 363 507 104 96
[71] 547 299 720 982 730 752 760 659 203 321 442 383 982 547
[85] 950 944 944 944 185 636 636 636 339 27 547 192 321
[99] 949 846 470 930 404 257 499 815 490 263 741 339 603 505
[113] 352 645 784 689 609 220 352 399 146 952 231 405 712 982
[127] 442 258 69 83 688 450 131 547 730 299 246 702 543 922
[141] 256 292 1006 845 425 720 821 71 425 89 426 71 582 316
[155] 192 821 879 303 129 321 61 870 132 304 325 321 258 720
[169] 730 321 460 30 692 98 97 97 90 203 282 806 87 887
[183] 650 547 574 434 321 167 768 451 684 656 840 841 96 363
[197] 86 2 86 832 685 783 783 821 821 821 119 783 48 492
[211] 168 604 118 249 700 18 499 470 1010 399 547 866 462 524
[225] 904 579 361 275 280 748 516 405 549 457 449 440 761 775
[239] 1007 777 321 954 686 576 547 188 203 337 337 609 237 397
[253] 699 237 609 203 252 157 363 214 339 339 26 45 609 45
[267] 442 884 573 155 375 321 299 950 627 96 115 326 785 958
[281] 302 301 724 377 858 914 982 621 203 797 13 546 209
[295] 402 545 729 59 873 957 128 57 458 269 309 299 309 559
[309] 547 321 718 383 982 982 773 1008 547 730 986 747 41 346
[323] 122 744 385 82 275 299 504 321 982 859 950 760 914 762
[337] 125 129 346 565 526 565 526 346 818 477 50 909 49 258
[351] 258 57 957 254 427 187 615 927 317 743 873 57 339 125
[365] 237 851 392 655 958 391 321 730 982 125 321 967 612 60
[379] 526 481 116 321 610 609 620 919 730 731 299 547 982 960
[393] 742 1000 967 299 299 675 321 655 730 947 248 734 581 168
[407] 730 321 730 547 1006 724 194 1006 821 465 918 675 124 401
[421] 578 906 906 675 771 129 129 675 389 194 583 201 203 924
[435] 925 924 780 154 821 899 168 644 675 346 146 931 330 365
```

Se discretiza la variable Localización:

```

> #se discretiza la variable Localizacion a valores n mericos
> cacao$Localizacion <- recode(cacao$Localizacion, "Amsterdam"=1;"Argentina"=2;
+ "Australia"=3;"Austria"=4;"Belgium"=5;"Bolivia"=6;"Brazil"=7;"Canada"=8;"Chile"=9;
+ "Colombia"=10;"Costa Rica"=11;"Czech Republic"=12;"Denmark"=13;"Dominican Republic"=14;
+ "Ecuador"=15;"Fiji"=16;"Finland"=17;"France"=18;"Germany"=19;"Ghana"=20;"Grenada"=21;
+ "Guatemala"=22;"Honduras"=23;"Hungary"=24;"Iceland"=25;"India"=26;"Ireland"=27;
+ "Israel"=28;"Italy"=29;"Japan"=30;"Lithuania"=31;"Madagascar"=32;"Martinique"=33;
+ "Mexico"=34;"Netherlands"=35;"New Zealand"=36;"Nicaragua"=37;"Peru"=38;"Philippines"=39;
+ "Poland"=40;"Portugal"=41;"Puerto Rico"=42;"Russia"=43;"Sao Tome"=44;"Scotland"=45;
+ "Singapore"=46;"South Africa"=47;"South Korea"=48;"Spain"=49;"St. Lucia"=50;"Suriname"=51;
+ "Sweden"=52;"Switzerland"=53;"U.K."=54;"U.S.A."=55;"Venezuela"=56;"Vietnam"=57;"Wales"=58';
+ as.factor.result=FALSE)
> |

```

Se comprueba que se hayan transformado todos los valores:

Se discretiza la variable Tipo.de.frijol:

```
> #se discretiza la variable Tipo.de.frijol a valores numéricos
> cacao$Tipo.de.frijol <- recode(cacao$Tipo.de.frijol, "Amazon"=1;
+ "Amazon mix"=2;
+ "Amazon, ICS"=3;
+ "Beniano"=4;
+ "Blend"=5;
+ "Blend-Forastero,Criollo"=6;
+ "CCN51"=7;
+ "Criollo"=8;
+ "Criollo (Amarru)"=9;
+ "Criollo (Ocumare 61)"=10;
+ "Criollo (Ocumare 67)"=11;
+ "Criollo (Ocumare 77)"=12;
+ "Criollo (Ocumare)"=13;
+ "Criollo (Porcelana)"=14;
+ "Criollo (Wild)"=15;
+ "Criollo, Forastero"=16;
+ "Criollo, Trinitario"=17;
+ "EET"=18;
+ "Forastero"=19;
+ "Forastero (Amelonado)"=20;
+ "Forastero (Arriba)"=21;
+ "Forastero (Arriba) ASS"=22;
+ "Forastero (Catongo)"=23;
+ "Forastero (Nacional)"=24;
+ "Forastero (Parazinho)"=25;
+ "Forastero(Arriba, CCN)"=26;
+ "Forastero, Trinitario"=27;
+ "Matina"=28;
+ "Nacional"=29;
+ "Nacional (Arriba)"=30;
+ "Trinitario"=31;
+ "Trinitario (85% Criollo)"=32;
+ "Trinitario (Amelonado)"=33;
+ "Trinitario (Scavina)"=34;
```

Se comprueba que se hayan transformado todos los valores:

```
> #se comprueba que se hayan transformado todos los valores a numeros
> #de la variable Tipo.de.frijol
> print(cacao[,6])
[1] 8 31 8 17 31 8 8 8 8 17 31 8 31 8 31 8 31 31 31 19 31 19 8 8 8 31
[27] 31 31 31 21 21 8 31 8 19 31 31 31 31 24 17 31 19 31 8 31 31 31 31 31
[53] 31 31 32 8 31 31 8 31 31 8 19 31 24 19 31 31 31 31 31 31 31 8 8 14 5
[79] 31 31 31 31 32 32 31 5 5 23 25 25 25 25 8 19 31 31 8 8 31 8 31 31 31
[105] 31 8 31 31 24 8 31 31 31 24 24 31 31 17 8 31 31 31 31 31 31 21 31 31 31 31
[131] 17 31 31 17 8 31 19 31 31 31 31 31 21 31 31 31 8 5 31 19 8 24 31 31 19 31
[157] 31 31 8 31 31 31 31 31 31 8 31 31 17 31 17 31 31 31 31 8 8 31 31 31
[183] 21 31 8 31 31 31 14 31 31 31 31 31 19 31 5 31 31 8 31 31 31 31 31 19 31
[209] 19 31 31 8 31 19 31 31 8 31 8 31 31 31 8 31 31 19 31 31 19 24 31 31 8 31
[235] 31 8 8 8 14 21 31 31 8 31 8 31 21 21 31 31 8 8 31 31 31 31 31 31 19 31
[261] 31 31 31 31 31 31 31 31 31 31 31 31 31 31 31 31 5 31 8 8 31 8 8 35 8 8
[287] 31 8 31 31 31 31 31 31 8 8 31 31 24 8 31 31 8 31 8 31 8 31 31 31 31
[313] 8 31 8 8 31 31 31 31 31 31 31 31 31 31 31 31 5 31 31 19 31 8 8 8 8 31
[339] 21 21 21 21 21 31 8 31 31 31 19 31 31 31 31 7 31 31 17 31 17 31 17 31 31
[365] 31 8 31 31 8 31 24 31 31 8 31 19 31 21 21 19 17 31 31 31 31 8 8 8 31
[391] 31 31 31 8 31 31 8 13 31 31 31 19 8 24 31 24 21 8 8 8 31 8 31 31 31
[417] 31 8 21 31 5 31 31 8 8 8 13 8 8 31 24 8 31 19 19 19 19 31 31 31 31 31
[443] 8 24 31 31 24 19 19 31 31 31 29 31 31 10 8 8 17 17 8 14 31 31 19 31 19
[469] 31 31 31 19 24 24 31 31 31 31 8 31 31 8 31 31 31 8 31 31 31 8 31 8
[495] 31 31 31 31 31 8 31 8 31 31 8 8 31 31 31 31 8 31 19 31 31 19 31 31
[521] 19 8 31 31 31 19 31 31 8 31 31 31 8 31 31 31 8 31 31 31 8 31 31 8 31
[547] 31 8 31 31 17 8 19 21 8 12 8 8 8 31 31 31 5 5 14 10 31 31 8 8 21
[573] 11 31 31 31 31 31 17 17 8 8 8 31 8 31 31 8 31 8 8 31 31 24 31 31
[599] 31 14 31 31 8 31 31 24 8 31 8 8 31 31 8 8 8 31 31 8 8 31 19 31 31
[625] 31 31 8 8 31 8 31 31 14 8 31 31 31 8 8 31 31 31 31 31 31 35 8 8 21 15 31
[651] 31 31 31 8 31 31 31 19 31 8 8 8 31 24 31 31 19 31 24 31 31 31 31 8 31 21
[677] 24 31 24 24 24 24 31 31 8 31 24 24 31 31 31 19 31 17 31 31 31 31 31 31 31
[703] 31 8 8 31 17 17 17 17 17 8 17 17 17 4 2 31 24 8 8 8 31 8 5 31
[729] 19 31 31 31 17 36 31 31 31 8 31 31 31 22 19 31 8 31 8 31 31 31 31 31
[755] 31 31 31 31 5 31 31 31 5 31 35 8 21 5 5 31 31 31 31 8 8 31 31 21 31
[781] 31 8 31 8 31 8 8 1 2 3 4 18 31 31 18 18 3 4 19 8 31 31 8 31 31 31
[807] 31 31 8 22 22 22 22 31 31 31 31 31 31 31 31 31 31 31 31 31 31 31 24
```

Se convierte la variable Porcentaje.de.cacao de carácter a numeric, para tratar los datos, también se transforma la variable dividiendo para 100 para obtener un rango de valores más manejable:

```
> #para convertir la variable Porcentaje.de.cacao de caracter a numeric
> cacao$Porcentaje.de.cacao <- as.numeric(cacao$Porcentaje.de.cacao)
> #Transformacion de variable Porcentaje.de.cacao, dividiendo para 100
> cacao$Porcentaje.de.cacao <- cacao$Porcentaje.de.cacao/100
> #comprueba el resultado de Porcentaje.de.cacao
> print(cacao[,3])
[1] 0.630 0.700 0.700 0.700 0.700 0.700 0.700 0.700 0.700 0.700 0.700 0.700 0.700 0.700
[14] 0.700 0.700 0.700 0.700 0.700 0.700 0.630 0.700 0.630 0.700 0.700 0.700 0.600
[27] 0.800 0.880 0.720 0.550 0.700 0.700 0.750 0.750 0.750 0.650 0.750 0.750 0.750
[40] 0.750 0.700 0.700 0.700 0.700 0.600 0.600 0.600 0.600 0.600 0.600 0.600 0.800
[53] 0.600 0.600 0.700 0.700 0.700 0.700 0.700 0.700 0.700 0.700 0.700 0.700 0.700
[66] 0.850 0.850 0.720 0.730 0.640 0.660 0.750 0.630 0.700 0.680 0.700 0.700 0.750
[79] 0.700 0.700 0.700 0.700 0.700 0.700 0.630 0.700 0.660 0.750 0.850 0.500
[92] 0.750 0.600 0.750 0.750 0.720 0.750 0.750 0.700 0.700 0.730 0.700 0.700 0.700
[105] 0.700 0.700 0.700 0.700 0.700 0.700 0.700 0.730 0.700 0.680 0.700 0.700 0.700
[118] 0.700 0.750 0.700 0.750 0.720 0.720 0.720 0.720 1.000 0.720 0.720 0.720 0.720
[131] 0.750 0.720 0.720 0.800 0.750 0.720 0.720 0.720 0.680 0.720 0.700 0.770 0.750
[144] 0.700 0.800 0.700 0.700 0.700 0.700 0.700 0.700 0.700 0.700 0.700 0.700 0.700
[157] 0.800 0.650 0.700 0.650 0.730 0.720 0.800 0.700 0.700 0.900 0.640 0.640 0.640
[170] 0.710 0.700 0.700 0.700 0.830 0.780 0.830 0.740 0.740 0.740 0.730 0.720 0.720
[183] 0.550 0.640 0.880 0.720 0.720 0.700 0.740 0.640 0.720 0.760 0.760 0.780 0.860
[196] 0.720 0.750 0.700 0.650 0.700 0.780 0.750 0.650 0.750 0.650 0.710 0.750 0.680
[209] 0.700 0.700 0.700 0.700 0.700 0.820 0.720 0.820 0.750 0.750 0.750 0.700 0.700
[222] 0.750 0.750 0.750 0.650 0.750 0.750 0.750 0.750 0.750 0.750 0.750 0.750 0.750
[235] 0.750 0.750 0.750 0.750 0.750 0.750 0.750 0.750 0.750 0.750 0.750 0.750 0.750
[248] 0.770 1.000 0.700 0.700 0.700 0.700 0.680 0.700 0.700 0.720 0.700 0.750 0.850
[261] 0.600 0.800 0.700 0.800 0.800 0.600 0.700 0.720 0.700 0.720 0.680 0.700 0.680
[274] 0.720 0.720 0.720 0.720 0.600 0.700 0.750 0.750 0.750 0.750 0.650 0.700 0.750
[287] 0.720 0.660 0.770 0.750 0.750 0.740 0.750 0.700 0.740 0.710 0.740 0.720 0.640
[300] 0.700 0.690 0.700 0.700 0.720 0.720 0.660 0.650 0.700 0.710 0.710 0.710 0.710
[313] 0.710 0.770 0.700 0.700 0.700 0.700 0.700 0.700 0.630 0.710 0.820 0.910 0.600
[326] 0.700 0.750 0.720 0.700 1.000 0.800 0.780 0.700 0.750 0.730 0.700 0.750 0.700
[339] 0.770 0.770 0.770 0.550 0.550 0.550 0.700 0.700 0.720 0.720 0.720 0.720 0.800
[352] 0.650 0.650 0.650 0.650 0.650 0.720 0.720 0.700 0.720 0.720 0.760 0.700 0.700
[365] 0.730 0.700 0.700 0.720 0.700 0.720 0.700 0.700 0.700 0.700 0.720 0.700 0.700
```

Se verifica el tipo de datos de cada variable, que todos son numeric:

```
> #Para ver el tipo de dato asignado a cada campo
> sapply(cacao, function(x) class(x))
      Empresa          Geo.region  Porcentaje.de.cacao       Localizacion
    "numeric"        "numeric"      "numeric"           "numeric"
Calificacion      Tipo.de.frijol
    "numeric"        "numeric"
```

Ahora si con los valores numéricos se va utilizar el algoritmo kmeans de los modelos de agregación, para crear los clusters, y poder ver si existen valores extremos en los grupos, para ello se debe determinar una semilla, es decir, se fija un punto inicial del espacio como centro de un grupo potencial. Esta semilla puede ser tanto un objeto seleccionado detalladamente entre el conjunto de objetos inicial como una combinación de valores creada de manera artificial y que corresponda al resumen de las características de varios objetos. En este caso se creará una semilla con valor 80, para que al reproducir el mismo ejercicio nos dé el mismo resultado, caso contrario nos daría un resultado diferente, ya que k-means,

selecciona aleatoriamente las observaciones. La semilla se la fija con la siguiente línea de instrucción:

```
> #clustering para revisar valores extremos
> set.seed(80)
> |
```

Para generar el modelo de agregación con el método kmeans, se realiza el clustering con 4 clusters que es valor de k =4, y se imprime los clusters, que se encuentran un poco homogéneos:

```
> agrkm <- kmeans(cacao,centers=4)
> print(grmk) #se presenta todo con el names tambien
K-means clustering with 4 clusters of sizes 510, 565, 350, 367

Cluster means:
  Empresa Geo.region Porcentaje.de.cacao Localizacion Calificacion Tipo.de.frijol
1 210.11569   546.7765      0.7124216    38.14314    3.227451    23.73137
2 206.92035   847.4000      0.7145221    37.41593    3.160177    22.29735
3 320.19429   210.8314      0.7228714    35.42286    3.177143    22.70000
4 91.49319    215.4850      0.7191826    39.13079    3.188692    24.15531

Clustering vector:
  1   2   3   4   5   6   7   8   9   10  11  12  13  14  15  16  17 
  4   1   4   4   2   4   4   2   2   2   2   1   4   4   4   4   2 
  18  19  20  21  22  23  24  25  26  27  28  29  30  31  32  33  34 
  4   2   4   4   4   2   4   2   2   2   2   1   1   2   2   4   1 
  35  36  37  38  39  40  41  42  43  44  45  46  47  48  49  50  51 
  1   2   2   1   4   2   2   1   2   1   1   4   4   4   2   4   2 
  52  53  54  55  56  57  58  59  60  61  62  63  64  65  66  67  68 
  4   4   2   1   4   4   4   1   4   1   4   1   4   4   4   4   1 
  69  70  71  72  73  74  75  76  77  78  79  80  81  82  83  84  85 
  4   4   1   4   2   2   2   2   2   1   4   4   1   4   2   1   2 
  86  87  88  89  90  91  92  93  94  95  96  97  98  99  100 101 102 
  2   2   2   4   1   1   1   1   4   4   1   4   4   2   2   1   2 
  103 104 105 106 107 108 109 110 111 112 113 114 115 116 117 118 119 
  4   4   1   2   1   4   2   4   1   1   4   1   2   1   1   4   4 
  120 121 122 123 124 125 126 127 128 129 130 131 132 133 134 135 136 
  4   4   2   4   4   2   2   1   4   4   4   1   1   4   1   2   4 
  137 138 139 140 141 142 143 144 145 146 147 148 149 150 151 152 153 
  4   2   1   2   4   4   2   2   4   2   2   4   4   4   4   4   1 
  154 155 156 157 158 159 160 161 162 163 164 165 166 167 168 169 170 
  4   4   2   2   4   4   4   4   2   4   4   4   4   4   2   2   4 
  171 172 173 174 175 176 177 178 179 180 181 182 183 184 185 186 187 
  1   4   1   4   4   4   4   4   2   4   2   1   1   1   1   1   4 
  188 189 190 191 192 193 194 195 196 197 198 199 200 201 202 203 204
```

Con el names, se puede ver el conjunto del objeto con k-means creado, lo cual indica que se puede acceder a la información de la asignación de las observaciones a los clusters y las distintas inercias para determinar sus medidas, como se muestra en la siguiente pantalla:

```
> names(grmk) #contenido del cluster
[1] "cluster"      "centers"       "totss"        "withinss"     "tot.withinss"
[6] "betweenss"    "size"         "iter"         "ifault"
> |
```

El objetivo del algoritmo k-means es maximizar la variación inter-cluster entre los propios grupos y minimizar la intra-cluster dentro de cada grupo.

Con totss, se determina la inercia total, con betweenss la inercia inter grupos la cual interesa que sea lo más alta posible, con withinss la inercia intra grupos, con tot.withinss la inercia intra grupos(total) la cual interesa que sea lo menor posible, en este caso si se cumple:

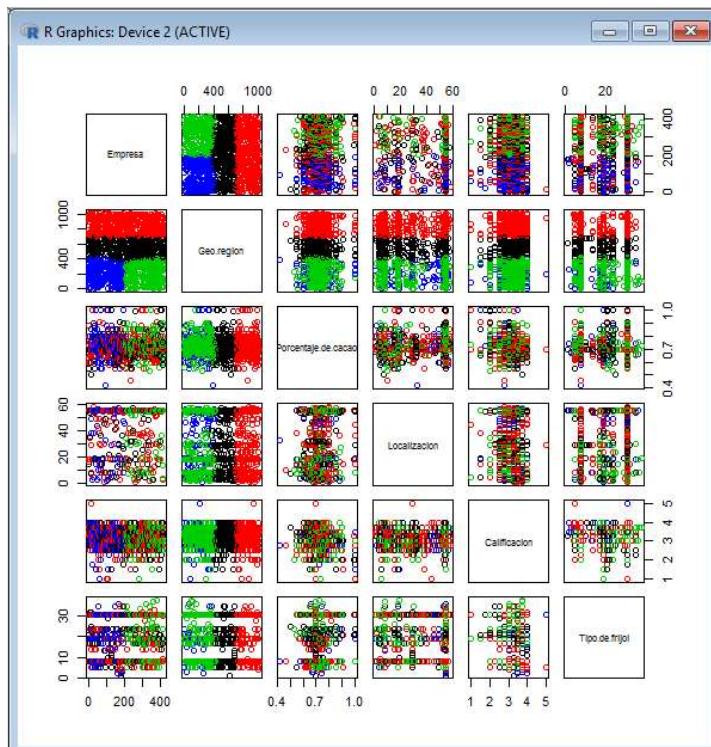
```
> agrkm$totss #inercia total
[1] 174671121
> agrkm$betweenss #inercia inter grupos, interesa que sea lo mas alta posible
[1] 137549883
> agrkm$withinss #inercia intra grupos
[1] 10870374 14044027 6058420 6148417
> agrkm$tot.withinss #inercia intra grupos(total), interesa que sea lo menor posible
[1] 37121238
> |
```

Con centers, se muestra los 4 grupos, como se han distribuido, de acuerdo a las 6 variables:

```
> agrkm$centers #muestras los centros de los grupos es igual al aggregate
   Empresa Geo.region Porcentaje.de.cacao Localizacion Calificacion Tipo.de.frijol
1 210.11569    546.7765      0.7124216    38.14314     3.227451    23.73137
2 206.92035    847.4000      0.7145221    37.41593     3.160177    22.29735
3 320.19429    210.8314      0.7228714    35.42286     3.177143    22.70000
4  91.49319    215.4850      0.7191826    39.13079     3.188692    24.15531
> |
```

Con plot se visualiza los clusters entre las 6 variables, donde se encuentran algunos valores extremos:

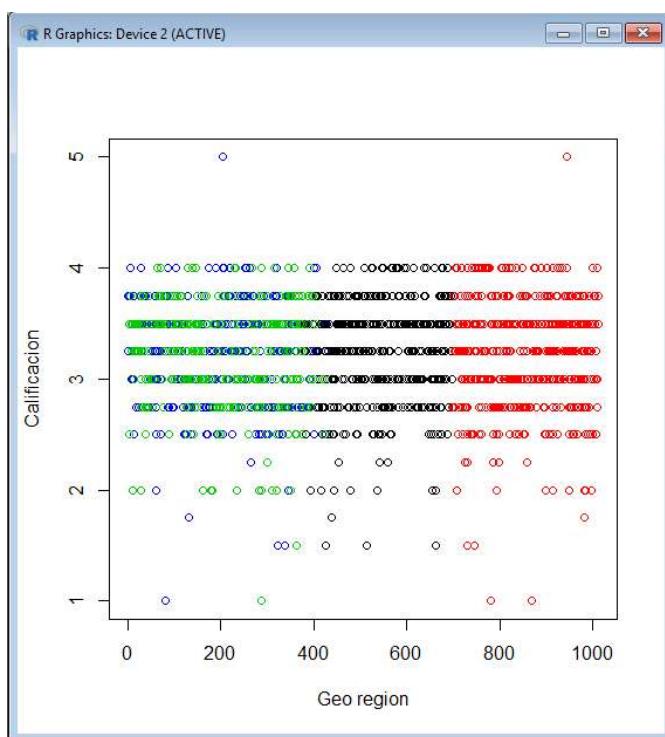
```
> plot(cacao,col=agrkm$cluster)
> |
```



Como en la figura anterior no se diferencia muy bien, se va a revisar si existen valores extremos entre las variables Geo.region y Calificacion:

```
> #revisar los resultados de acuerdo a dos variables
> plot(cacao$Geo.region,cacao$Calificacion,col=agrkm$cluster,xlab="Geo region",ylab="Calificacion")
```

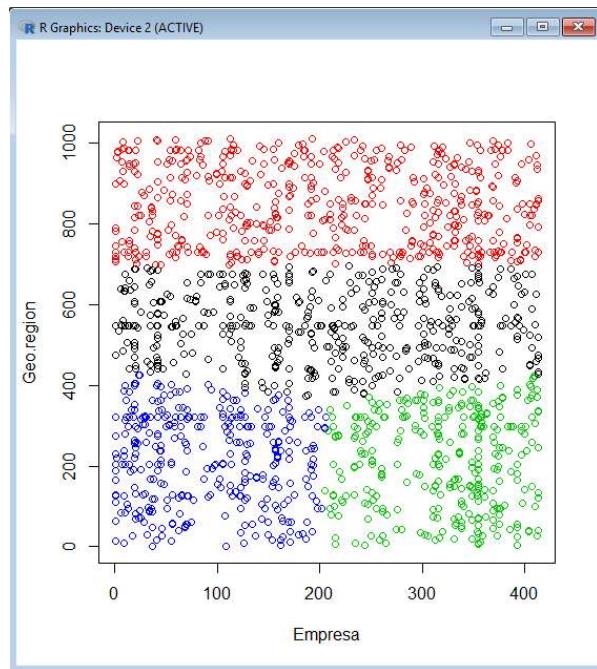
Se puede observar que existen 2 Geo.regiones donde la calificación es 5 y estos pueden ser valores extremos, pero en este caso no se los elimina porque se los necesitará para realizar el análisis de los datos, ya que son los únicos valores con calificación alta y en el caso de los 4 registros que tienen calificación 1 tampoco se los elimina para su posterior análisis:



Se revisa los valores extremos entre las variables Empresa y Geo.region:

```
> #aqui no existen valores extremos
> plot(cacao$Empresa,cacao$Geo.region,col=agrkm$cluster,xlab="Empresa",ylab="Geo.region")
```

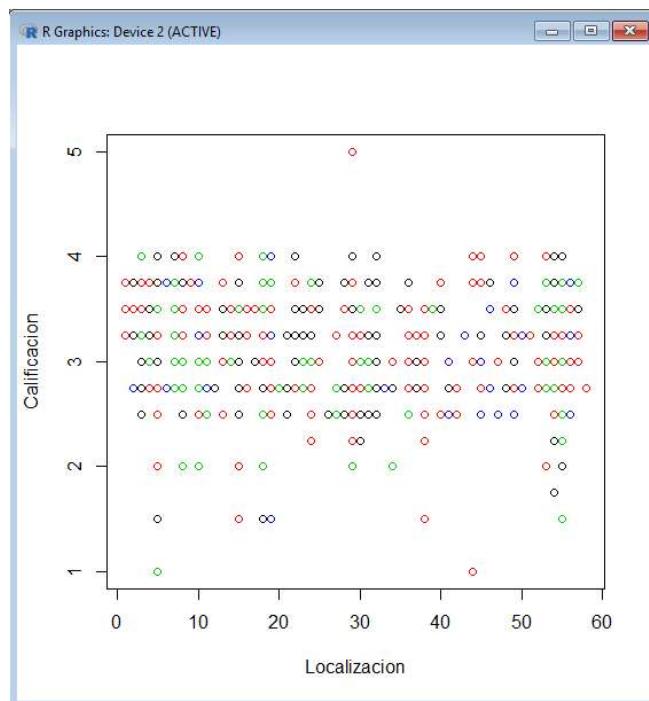
En este caso no existen valores extremos, ya que todos los datos se los ve conjuntamente:



Se revisa los valores extremos entre las variables Localizacion y Calificacion:

```
> plot(cacao$Localizacion,cacao$Calificacion,col=agrkm$cluster,xlab="Localizacion",ylab="Calificacion")  
>
```

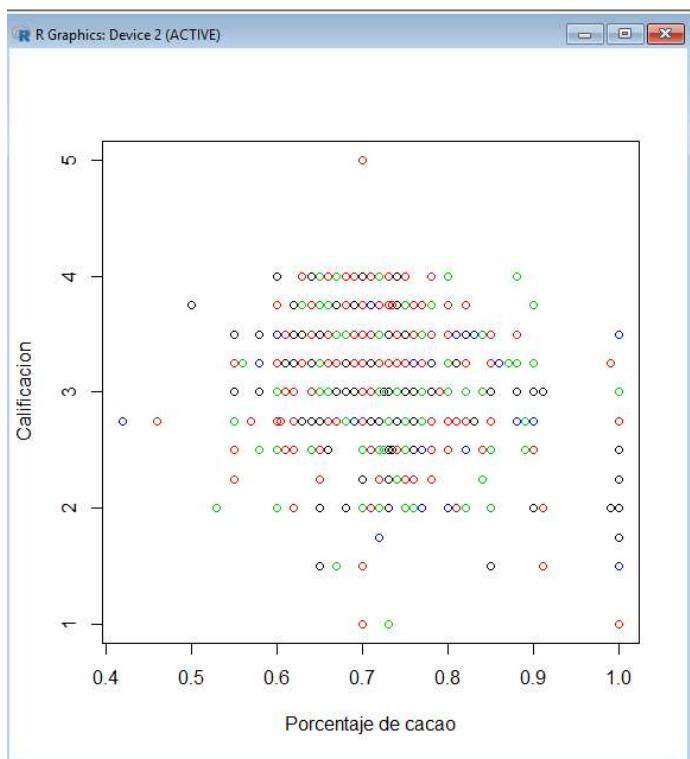
En este caso se puede ver que existen valores extremos cuando la calificación es 5 y cuando es 1, pero no se los elimina para su posterior análisis:



Se revisa los valores extremos entre las variables Porcentaje de cacao y Calificación:

```
> plot(cacao$Porcentaje.de.cacao,cacao$Calificacion,col=agrkm$cluster,xlab="Porcentaje de cacao",ylab="Calificacion")  
> |
```

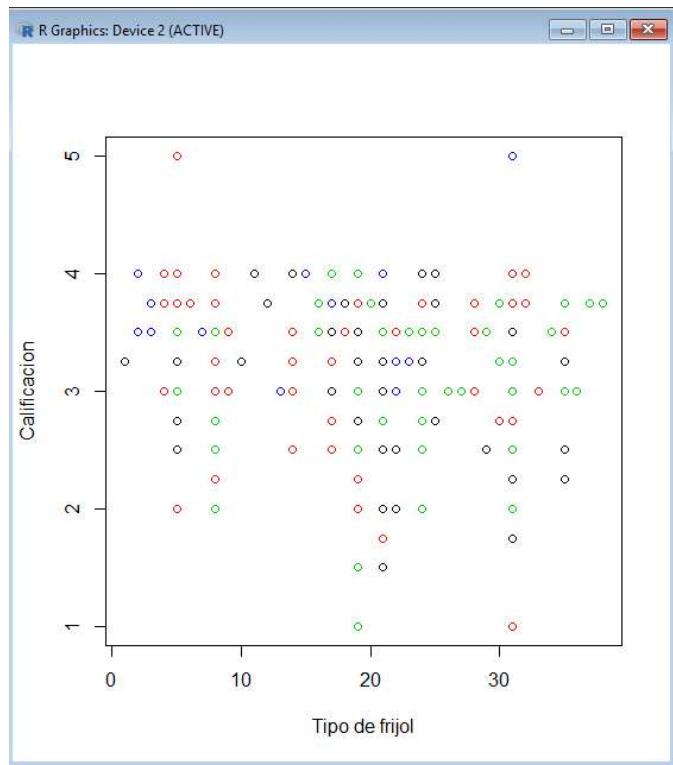
También en este caso se puede ver que existen valores extremos cuando la calificación es 5 o 1 y cuando el Porcentaje de cacao se encuentra entre 0.4 y 0.5, pero no se los elimina para su posterior análisis:



Se revisa los valores extremos entre las variables Tipo de frijol y Calificación:

```
> plot(cacao$Tipo.de.frijol,cacao$Calificacion,col=agrkm$cluster,xlab="Tipo de frijol",ylab="Calificacion")  
> |
```

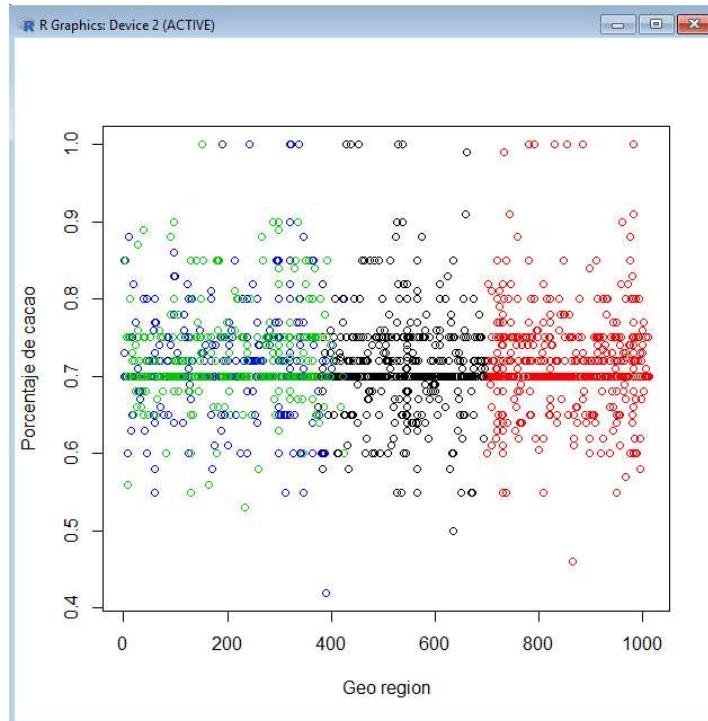
También en este caso se puede ver que existen valores extremos cuando la calificación es 5 o 1, pero no se los elimina para su posterior análisis:



Se revisa los valores extremos entre las variables Geo region y Porcentaje de cacao:

```
> plot(cacao$Geo.region,cacao$Porcentaje.de.cacao,col=agrkm$cluster,xlab="Geo region",ylab="Porcentaje de cacao")  
> |
```

En este caso se puede ver que existen valores extremos cuando el porcentaje de cacao se encuentra en 40, 50 y 100%, pero no se los elimina para su posterior análisis:



Como se pudo ver en los clusters las variables que tienen valores extremos son: Porcentaje de cacao y Calificación, los cuales se los muestra a través de la función boxplot.stats():

```
> #con el boxplot tambien se puede diferenciar los valores extremos de las
> #variables Calificación y Porcentaje.de.cacao
> boxplot.stats(cacao$Porcentaje.de.cacao)$out
[1] 0.600 0.880 0.550 0.600 0.600 0.600 0.600 0.600 0.600 0.600 0.600 0.600 0.850
[14] 0.850 0.850 0.500 0.600 1.000 0.900 0.830 0.830 0.550 0.880 0.860 1.000 0.850
[27] 0.600 0.600 0.600 0.910 0.600 1.000 0.550 0.550 0.550 0.900 0.600 0.600 0.600
[40] 0.420 0.610 1.000 0.880 0.850 1.000 0.620 0.830 0.850 0.600 0.600 0.580 0.600
[53] 0.850 0.880 1.000 0.600 0.605 0.580 0.610 0.600 0.550 0.550 0.620 0.620 0.580
[66] 0.620 0.850 0.600 0.600 0.610 0.600 0.850 0.850 0.600 0.610 0.550 0.910 0.610
[79] 0.610 0.550 0.850 0.570 0.580 0.580 1.000 0.900 1.000 1.000 1.000 0.610 0.550
[92] 0.620 0.850 0.850 0.850 0.600 0.900 0.600 0.560 0.600 0.600 0.600 0.600 0.850
[105] 0.460 0.600 0.580 0.550 0.550 0.550 0.850 0.620 0.850 0.890 0.600 0.850 0.990
[118] 0.850 0.580 0.850 1.000 0.880 0.850 0.850 0.600 0.600 0.850 0.600 1.000 0.550
[131] 0.850 0.850 0.620 1.000 0.850 0.850 0.550 0.600 0.850 0.850 0.840 0.830 0.530
[144] 0.620 0.620 0.600 1.000 1.000 0.850 0.620 0.880 0.850 0.840 0.840 0.610 0.600
[157] 0.880 0.620 0.600 0.870 0.620 0.990 0.600 0.840 0.910 0.600 0.560 0.850 0.850
[170] 0.900 0.900 1.000 0.900 0.890 0.880 1.000 0.600 0.850 0.850 0.620 0.900 0.580
[183] 0.620
> boxplot.stats(cacao$Calificación)$out
[1] 5.00 5.00 1.75 1.75 1.50 2.00 2.00 1.50 1.00 2.00 1.00 1.50 1.00 1.50 2.00 2.00
[17] 2.00 2.00 1.75 2.00 2.00 2.00 1.50 2.00 2.00 2.00 2.00 1.50 2.00 1.00 2.00
[33] 2.00 2.00 2.00 2.00 2.00 2.00 2.00 1.50 2.00 2.00 2.00 1.50 2.00 2.00 2.00
> |
```

Por lo antes indicado, se visualiza que estos son consistentes y están dentro de rangos normales por cada uno de los atributos o variables, en el caso de las calificaciones de expertos igual a 5, solo existen 2 registros, pero son los únicos que existen, es por ello que el manejo de estos valores extremos consistirá en simplemente dejarlos como actualmente están levantados, es decir; no sufren ningún tipo de alteración ya que se los necesitará para su posterior análisis.

Una vez que se ha realizado la limpieza de los datos, se procede a guardarlos en un nuevo archivo: flavors_of_cacao_clean.csv

```
> #exportacion de los datos limpios en un nuevo archivo, el row.names=F es para
> #que no grabe la primera columna que se genera con la secuencia de los datos
> write.csv(cacao, "D:/flavors_of_cacao_clean.csv", row.names=F)
> |
```

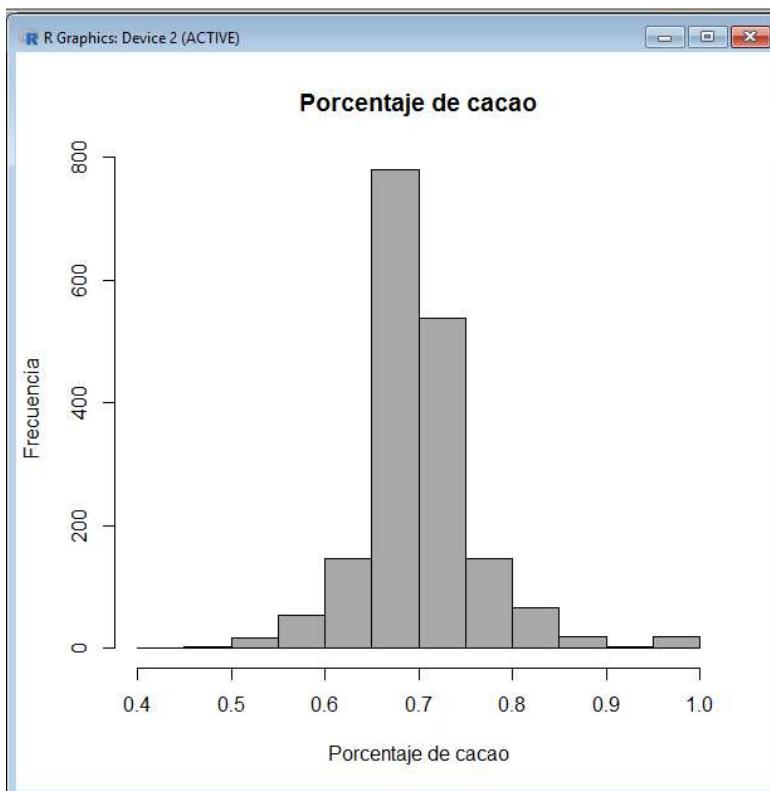
4 Análisis de los datos.

4.1 Selección de los grupos de datos que se quieren analizar/comparar (planificación de los análisis a aplicar).

Para determinar el conjunto de datos que se va a analizar y comparar, se realiza un histograma de acuerdo al Porcentaje de cacao que es una de las variables que más se la necesita para analizar sus datos, de acuerdo a las preguntas planteadas al inicio.

```
> #se realiza un histograma para determinar que conjunto de datos son los de
> #mayor interes del Porcentaje.de.cacao
> hist(cacao$Porcentaje.de.cacao, breaks="Sturges", col="darkgray", main="Porcentaje de cacao",
+ xlab="Porcentaje de cacao",ylab="Frecuencia")
> |
```

En el histograma se puede diferenciar que los conjuntos de datos de mayor interés son los de mayor frecuencia, que en este caso son: 0.70, 0.72 y 0.75:



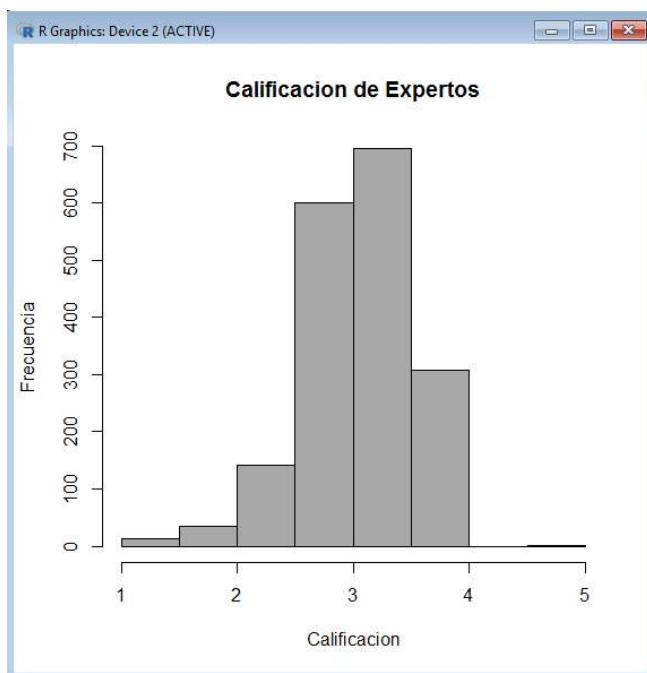
Se realiza la agrupación de acuerdo a los porcentajes de mayor frecuencia, en este caso se analizarán 1083 registros:

```
> #Agrupacion por los porcentajes 0.70, 0.72 0.75 que son los que tienen
> #mayor cantidad de datos para analizar
> cacao.porcentaje1 <- cacao[cacao$Porcentaje.de.cacao == 0.70,]
> cacao.porcentaje2 <- cacao[cacao$Porcentaje.de.cacao == 0.72,]
> cacao.porcentaje3 <- cacao[cacao$Porcentaje.de.cacao == 0.75,]
>
> #se verifica la cantidad de datos
> nrow(cacao.porcentaje1)
[1] 672
> nrow(cacao.porcentaje2)
[1] 189
> nrow(cacao.porcentaje3)
[1] 222
> #total del conjunto de datos de acuerdo al porcentaje 1083 registros
> nrow(cacao.porcentaje1)+nrow(cacao.porcentaje2)+nrow(cacao.porcentaje3)
[1] 1083
> |
```

Ahora se realiza un histograma de acuerdo a la Calificación de los expertos que es otra de las variables que más se la necesita para analizar sus datos, de acuerdo a las preguntas planteadas al inicio:

```
> #se realiza otro histograma para determinar que conjuntos de datos son los de
> #mayor interes de la Calificación de expertos
> hist(cacao$Calificacion, breaks="Sturges", col="darkgray", main="Calificacion de Expertos",
+ xlab="Calificacion",ylab="Frecuencia")
> |
```

En el histograma se puede diferenciar que los conjuntos de datos de mayor interés son los de mayor frecuencia, que en este caso son: 2.50, 2.75, 3, 3.25. 3.50, 3.75:



Se realiza la agrupación de acuerdo a las calificaciones de mayor frecuencia, en este caso se analizarán 1632 registros:

```
> #Agrupacion por calificacion, entre los cuales los de mayor interes son:
> #2.50, 2.75, 3, 3.25. 3.50, 3.75, como se puede verificar en el histograma
> cacao.calificacion1 <- cacao[cacao$Calificacion == 2.50,]
> cacao.calificacion2 <- cacao[cacao$Calificacion == 2.75,]
> cacao.calificacion3 <- cacao[cacao$Calificacion == 3,]
> cacao.calificacion4 <- cacao[cacao$Calificacion == 3.25,]
> cacao.calificacion5 <- cacao[cacao$Calificacion == 3.50,]
> cacao.calificacion6 <- cacao[cacao$Calificacion == 3.75,]
>
> #total de registros del conjunto de datos de acuerdo a la Calificacion de Expertos
> #1632 registros
> nrow(cacao.calificacion1)+nrow(cacao.calificacion2)+nrow(cacao.calificacion3)+nrow(cacao.calificacion4)+nrow(cacao.calificacion5)+nrow(cacao.calificacion6)
[1] 1632
> |
```

```
> #subconjuntos
> nrow(cacao.calificacion1)
[1] 127
> nrow(cacao.calificacion2)
[1] 259
> nrow(cacao.calificacion3)
[1] 341
> nrow(cacao.calificacion4)
[1] 303
> nrow(cacao.calificacion5)
[1] 392
> nrow(cacao.calificacion6)
[1] 210
> |
```

4.2 Comprobación de la normalidad y homogeneidad de la varianza.

Para realizar la comprobación de la normalidad y homogeneidad de la varianza, se realizó la prueba de normalidad, aplicando el método de Anderson-Darling, el cual consiste en medir qué tan bien siguen los datos una distribución específica. Para un conjunto de datos y distribución en particular, mientras mejor se ajuste la distribución a los datos, menor será este estadístico. Por ejemplo, usted puede utilizar el estadístico de Anderson-Darling para determinar si los datos cumplen el supuesto de normalidad para una prueba t.

Las hipótesis para la prueba de Anderson-Darling son:

- H_0 : Los datos siguen una distribución especificada
- H_1 : Los datos no siguen una distribución especificada

Utilice el valor p correspondiente (si está disponible) para probar si los datos provienen de la distribución elegida. Si el valor p es menor que un nivel de significancia elegido (por lo general 0.05 o 0.10), entonces rechace la hipótesis nula de que los datos provienen de esa distribución.

Para nuestro dataset, se carga la librería nortest y se establece el valor alpha (nivel de significancia) a 0.05.

```
> library(nortest)
> alpha = 0.05
> col.names = colnames(cacao)
> print (col.names)
[1] "Empresa"           "Geo.region"        "Porcentaje.de.cacao"
[4] "Localizacion"       "Calificacion"      "Tipo.de.frijol"
> for (i in 1:ncol(cacao)) {
+   if (i == 1) cat("Variables que no siguen una distribución normal:\n")
+   if (is.integer(cacao[,i]) | is.numeric(cacao[,i])) {
+     p_val = ad.test(cacao[,i])$p.value
+     if (p_val < alpha) {
+       cat(col.names[i])
+       # Format output
+       if (i < ncol(cacao) - 1) cat(", ")
+       if (i %% 3 == 0) cat("\n")
+     }
+   }
+ }
Variables que no siguen una distribución normal:
Empresa, Geo.region, Porcentaje.de.cacao,
Localizacion, Calificacion, Tipo.de.frijol
> |
```

Según los cálculos realizados, todas las variables que intervienen no siguen una distribución normal o especificado, es decir; se cumple la hipótesis H_1 .

En cuanto a la homogeneidad de varianzas, se usó el test de Fligner-Killeen, el mismo que es un test no paramétrico que compara varianzas, basándose en la mediana.

Para nuestro dataset, se consideran 2 variables numéricas que son la calificación y porcentaje de cacao, esto debido a que son campos importantes que se requieren para el análisis.

En el siguiente test, la hipótesis nula consiste en que ambas varianzas son iguales.

```
> #Homogeneidad de varianzas mediante el test de Fligner-Killeen:
> #Se trata de un test no paramétrico que compara las varianzas basándose en la mediana
> #De las variables obtenidas en el paso anterior, se toman 2 de ellas que son
> #numéricas: Calificacion y Porcentaje.de.cacao
> fligner.test(Calificacion ~ Porcentaje.de.cacao, data = cacao)

Fligner-Killeen test of homogeneity of variances

data: Calificacion by Porcentaje.de.cacao
Fligner-Killeen:med chi-squared = 57.15, df = 44, p-value = 0.08821
```

Puesto que obtenemos un p-valor superior a 0,05, aceptamos la hipótesis de que las varianzas de ambas muestras son homogéneas.

4.3 Aplicación de pruebas estadísticas para comparar los grupos de datos. En función de los datos y el objetivo del estudio, aplicar pruebas de contraste de hipótesis, correlaciones, regresiones, etc.

4.3.1 Covarianza y Correlación

Se lee el directorio y nombre del archivo obtenido, llamado flavors_of_cacao_clean.csv

Se imprime el contenido del archivo. La información se encuentra transformada como se ha dicho anteriormente, es decir; a valores numéricos.

```
> #4.3.1 Covarianza y Correlación:  
> setwd("D:/")  
> cacao <- read.csv(file = "flavors_of_cacao_clean.csv", header=TRUE)  
> print (cacao)
```

	Empresa	Geo.region	Porcentaje.de.cacao	Localizacion	Calificacion	Tipo.de.frijol
1	1	14	0.630	18	3.75	8
2	1	476	0.700	18	2.75	31
3	1	64	0.700	18	3.00	8
4	1	15	0.700	18	3.50	31
5	1	789	0.700	18	3.50	19
6	1	166	0.700	18	2.75	8
7	1	275	0.700	18	3.50	19
8	1	899	0.700	18	3.50	8
9	1	781	0.700	18	3.75	8
10	1	707	0.700	18	4.00	24
11	1	712	0.700	18	2.75	31
12	1	547	0.700	18	3.00	8
13	1	129	0.700	18	3.25	31
14	1	321	0.700	18	3.75	8
15	1	229	0.700	18	2.75	31
16	1	111	0.700	18	3.00	8
17	1	720	0.700	18	3.25	31
18	1	203	0.700	18	4.00	31
19	1	747	0.700	18	3.25	31

Se presentan los nombres de los campos o variables.

```
165     29      325          0.700      55      2.75      31  
166     29      321          0.900      55      2.75      8  
[ reached getoption("max.print") -- omitted 1626 rows ]  
> cat("1: Empresa\n                           \r\n");  
1: Empresa  
> cat("2: Geo.region\n                           \r\n");  
2: Geo.region  
> cat("3: Porcentaje.de.cacao\n                           \r\n");  
3: Porcentaje.de.cacao  
> cat("4: Localizacion\n                           \r\n");  
4: Localizacion  
> cat("5: Calificacion\n                           \r\n");  
5: Calificacion  
> cat("6: Tipo.de.frijol\n                           \r\n");  
6: Tipo.de.frijol  
> cat("                           \r\n");
```

Se construye el algoritmo correspondiente para determinar la covarianza y correlación entre los campos del archivo.

```
> num <- 0
> cont <- 2
> for (i in 1:ncol(cacao)){
+   if (cont < ncol(cacao)+1){
+     for (j in cont:ncol(cacao)){
+       covarianza <- cov(cacao[[i]],cacao[[j]])
+       correlacion <- cor(cacao[[i]],cacao[[j]])
+       cat("Entre el campo: ",colnames(cacao)[i],"y el campo:",colnames(cacao)[j],"la
covarianza es: ",covarianza,"\n\r");
+       cat("Entre el campo: ",colnames(cacao)[i],"y el campo:",colnames(cacao)[j],"la
correlación es: ",correlacion,"\n\r");
+
+       plot(cacao[[i]],cacao[[j]],
+             main = "Dispersión",
+             ylab = paste("Campo",colnames(cacao)[j]),
+             xlab = paste("Campo",colnames(cacao)[i]),
+             col = "red")
+
+       cat ("\n\r-----\n\r");
+       num <- num + 1
+     }
+   }
+   cont <- cont + 1
+ }
```

Los resultados son:

```
Entre el campo: Empresa y el campo: Geo.region la covarianza es: -171.8169
Entre el campo: Empresa y el campo: Geo.region la correlación es: -0.004843315
-----
Entre el campo: Empresa y el campo: Porcentaje.de.cacao la covarianza es: 0.2672586
Entre el campo: Empresa y el campo: Porcentaje.de.cacao la correlación es: 0.03459321
-----
Entre el campo: Empresa y el campo: Localizacion la covarianza es: -192.0783
Entre el campo: Empresa y el campo: Localizacion la correlación es: -0.07736848
-----
Entre el campo: Empresa y el campo: Calificacion la covarianza es: -1.070554
Entre el campo: Empresa y el campo: Calificacion la correlación es: -0.01817596
-----
Entre el campo: Empresa y el campo: Tipo.de.frijol la covarianza es: -20.60954
Entre el campo: Empresa y el campo: Tipo.de.frijol la correlación es: -0.01677114
-----
Entre el campo: Geo.region y el campo: Porcentaje.de.cacao la covarianza es: -0.71861
48
Entre el campo: Geo.region y el campo: Porcentaje.de.cacao la correlación es: -0.0404
3415
-----
Entre el campo: Geo.region y el campo: Localizacion la covarianza es: 144.8602
Entre el campo: Geo.region y el campo: Localizacion la correlación es: 0.02536459
```

```
Entre el campo: Geo.region y el campo: Calificacion la covarianza es: -2.744128
Entre el campo: Geo.region y el campo: Calificacion la correlación es: -0.02025284

-----
Entre el campo: Geo.region y el campo: Tipo.de.frijol la covarianza es: -127.2098
Entre el campo: Geo.region y el campo: Tipo.de.frijol la correlación es: -0.04499952

-----
Entre el campo: Porcentaje.de.cacao y el campo: Localizacion la covarianza es: 0.03514129
Entre el campo: Porcentaje.de.cacao y el campo: Localizacion la correlación es: 0.02825392

-----
Entre el campo: Porcentaje.de.cacao y el campo: Calificacion la covarianza es: -0.00426916
8
Entre el campo: Porcentaje.de.cacao y el campo: Calificacion la correlación es: -0.1446795

-----
Entre el campo: Porcentaje.de.cacao y el campo: Tipo.de.frijol la covarianza es: -0.005002
958
Entre el campo: Porcentaje.de.cacao y el campo: Tipo.de.frijol la correlación es: -0.00812
6366

-----
Entre el campo: Localizacion y el campo: Calificacion la covarianza es: -0.8904374
Entre el campo: Localizacion y el campo: Calificacion la correlación es: -0.09390611

-----
Entre el campo: Localizacion y el campo: Tipo.de.frijol la covarianza es: 7.823829
Entre el campo: Localizacion y el campo: Tipo.de.frijol la correlación es: 0.03954716

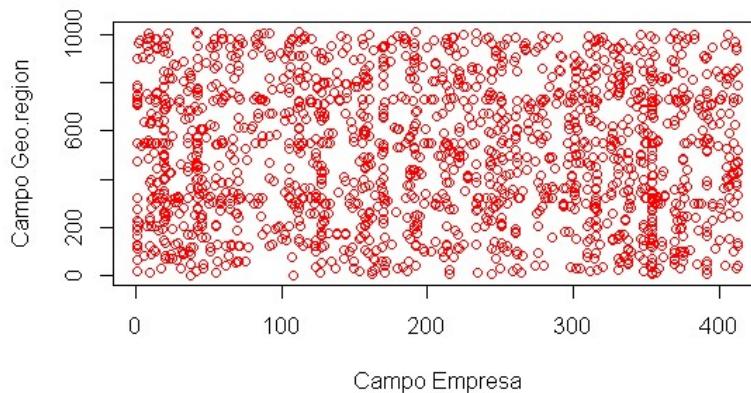
-----
Entre el campo: Calificacion y el campo: Tipo.de.frijol la covarianza es: -0.2534062
Entre el campo: Calificacion y el campo: Tipo.de.frijol la correlación es: -0.05399037

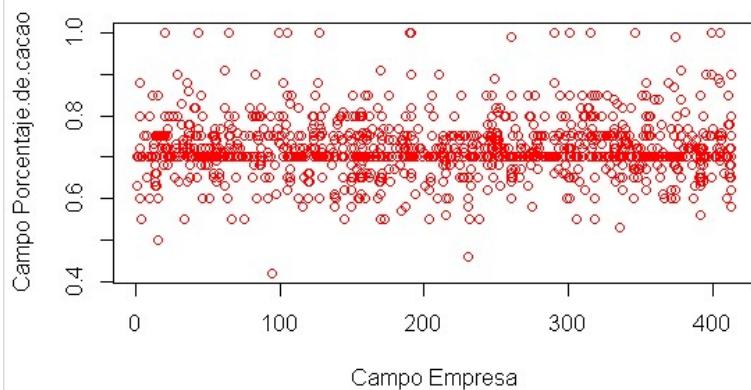
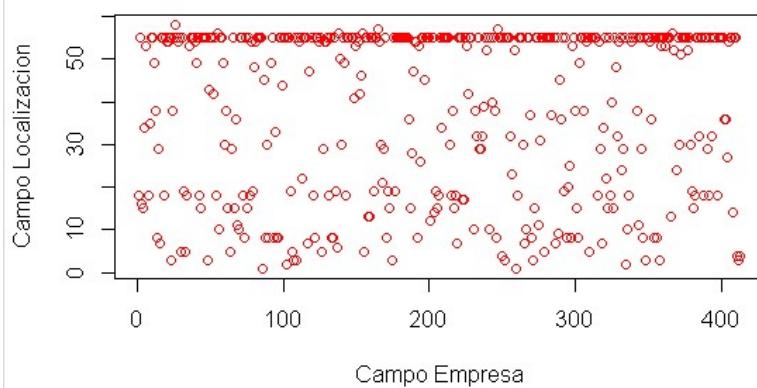
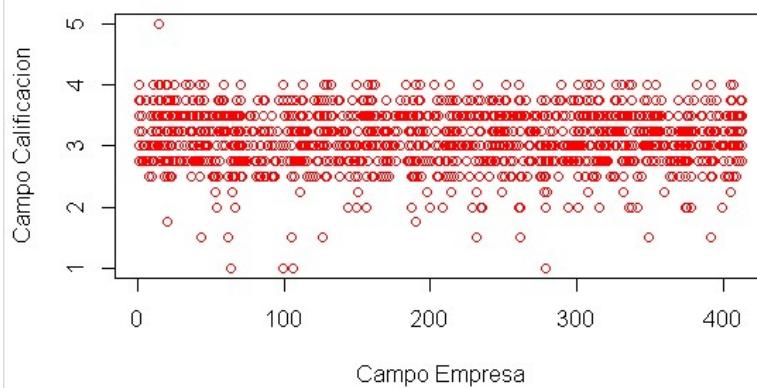
> cat("El número total de combinaciones es:", num)
El número total de combinaciones es: 15
```

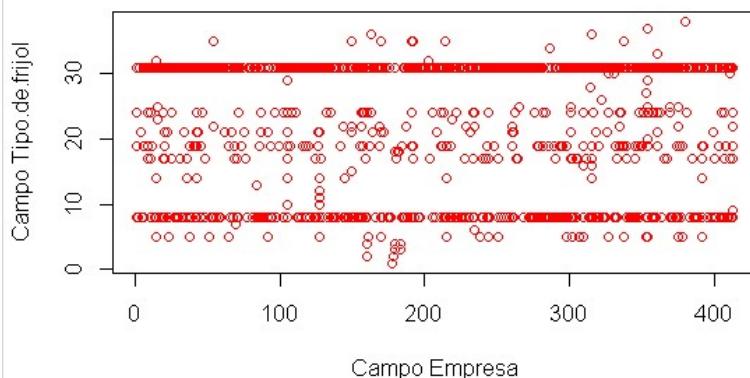
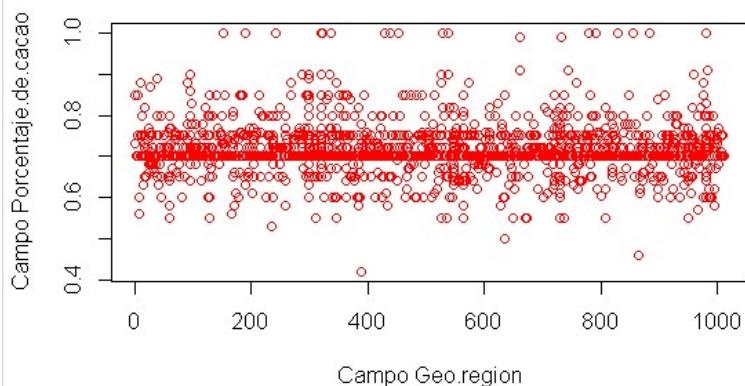
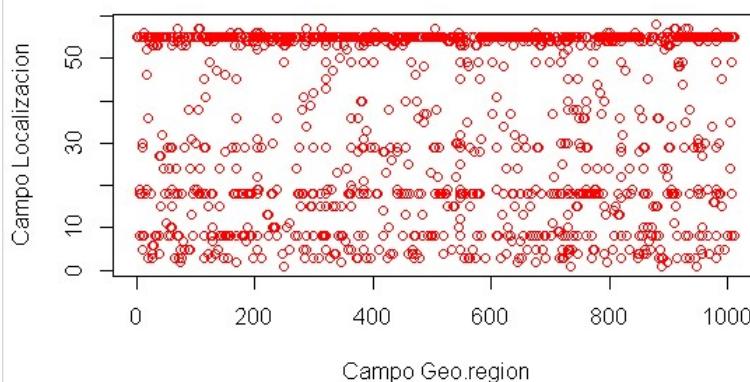
A nivel de gráficas, las dispersiones son las que se presentan a continuación, según la combinación. Así:

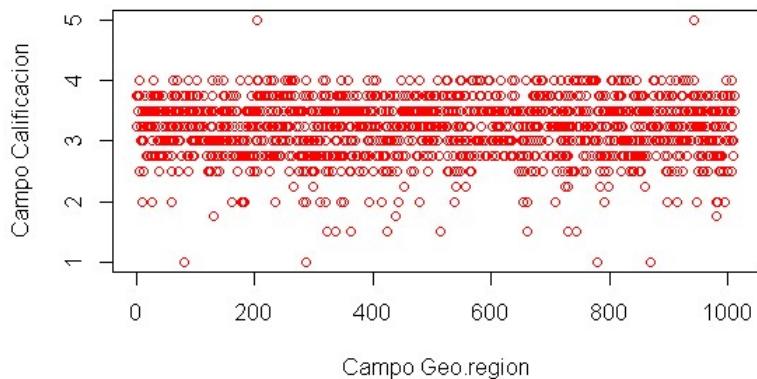
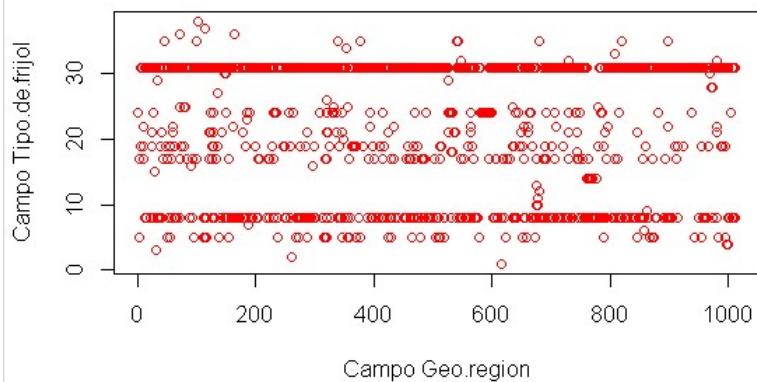
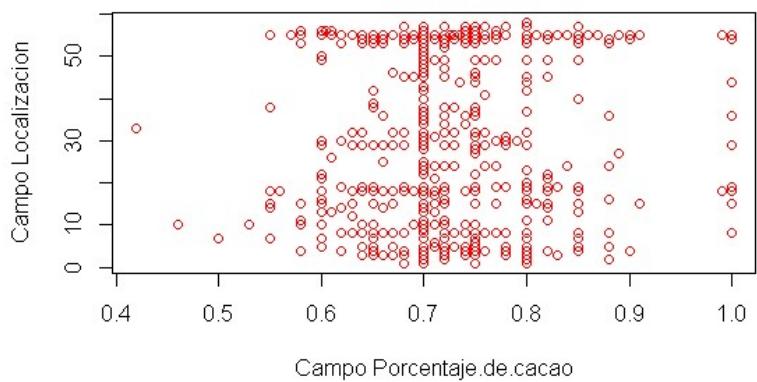
Combinación 1

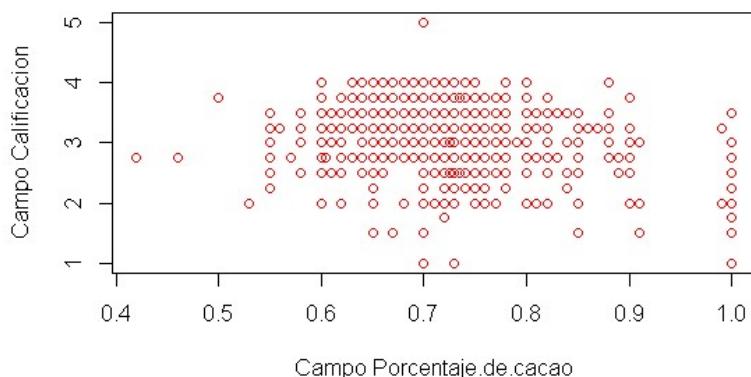
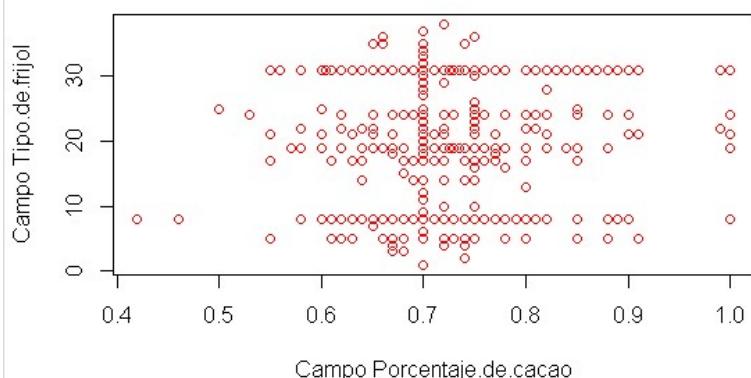
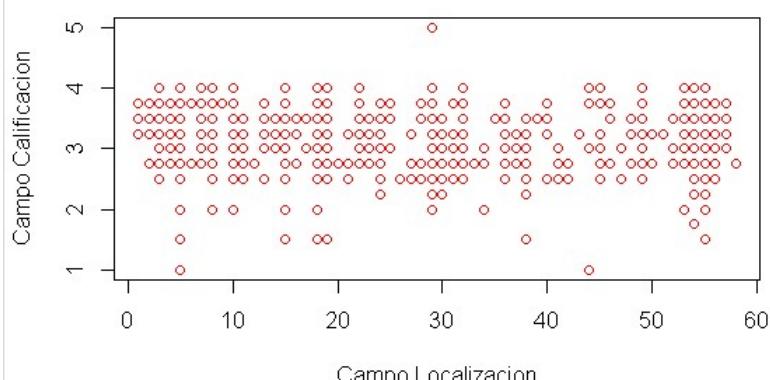
Dispersión

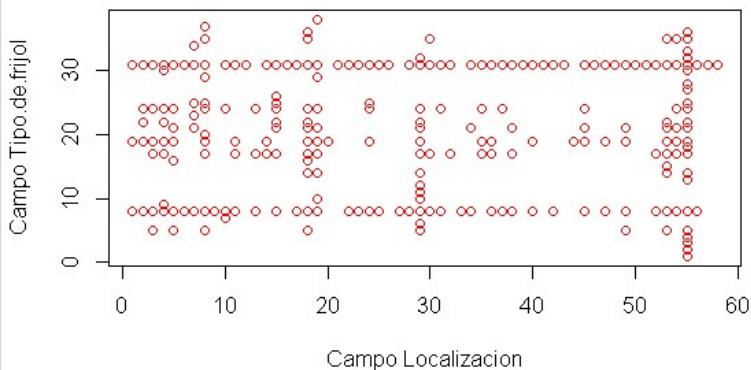
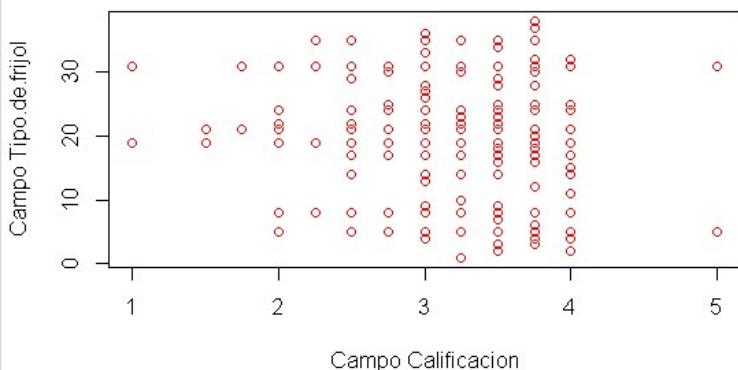


Combinación 2**Dispersión****Combinación 3****Dispersión****Combinación 4****Dispersión**

Combinación 5**Dispersión****Combinación 6****Dispersión****Combinación 7****Dispersión**

Combinación 8**Dispersión****Combinación 9****Dispersión****Combinación 10****Dispersión**

Combinación 11**Dispersión****Combinación 12****Dispersión****Combinación 13****Dispersión**

Combinación 14**Dispersión****Combinación 15****Dispersión**

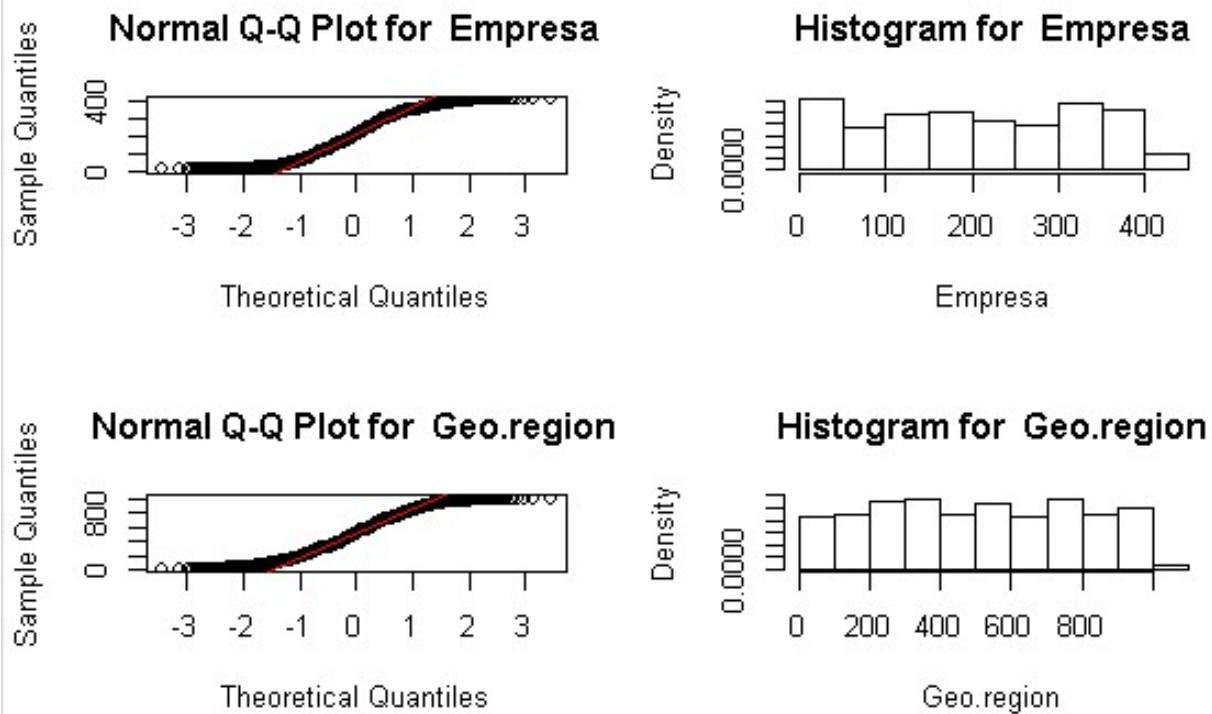
Los 2 campos que tienen mayor correlación entre sí son Porcentaje de cacao y Calificacion, pues su correlación es de -0.1446795, que es el más próximo a -1.

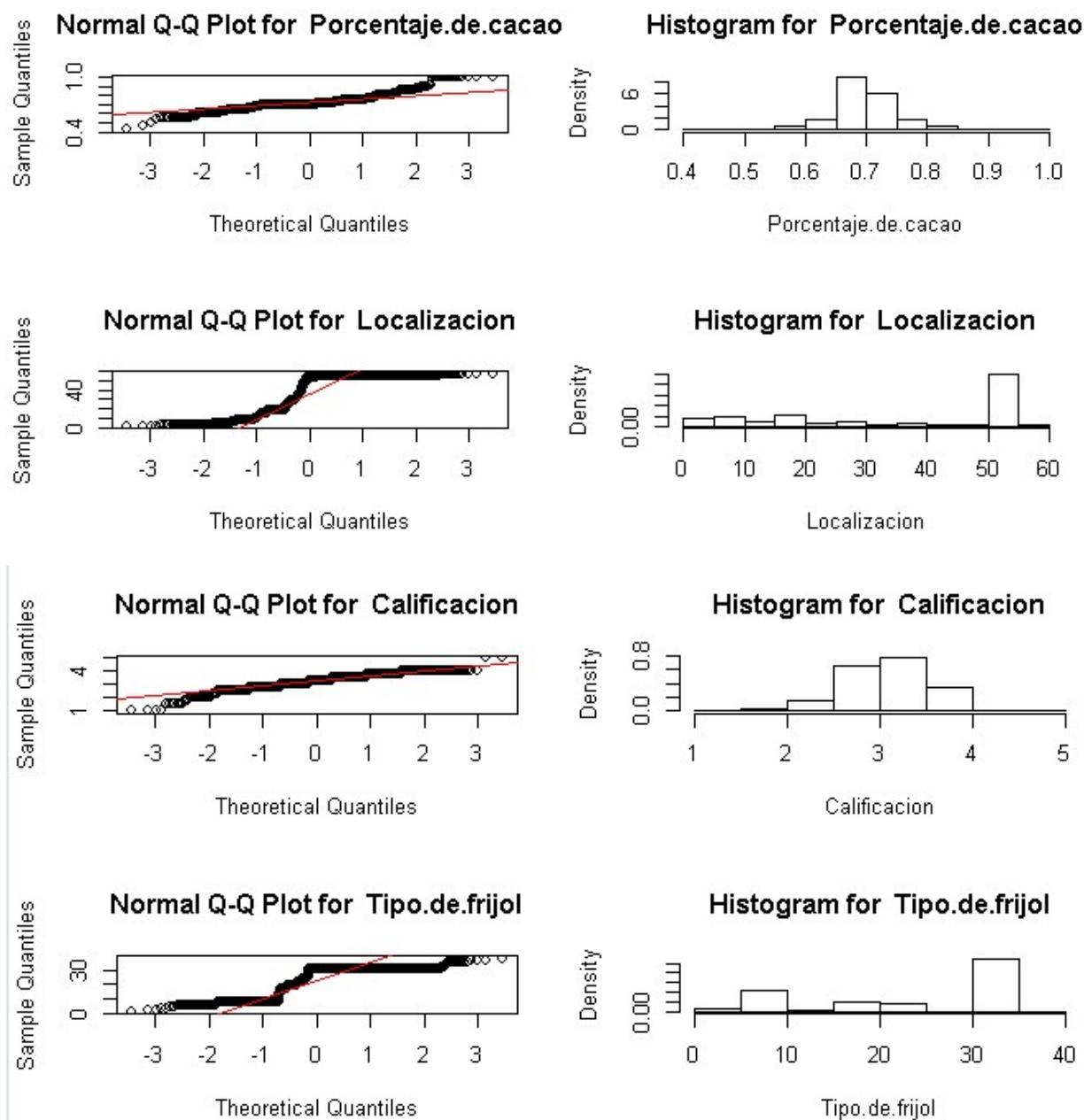
4.3.2 Prueba de contraste gráfico

Se lee el directorio y archivo con el que se trabaja y se aplica el algoritmo correspondiente para realizar la prueba de contraste (usando gráficas de quantile-quantile plot y el histograma).

```
> setwd("D:/")
> cacao <- read.csv(file = "flavors_of_cacao_clean.csv", header=TRUE)
> par(mfrow=c(2,2))
> for(i in 1:ncol(cacao)) {
+   if (is.numeric(cacao[,i])){
+     qqnorm(cacao[,i],main = paste("Normal Q-Q Plot for ",colnames(cacao)[i]))
+     qqline(cacao[,i],col="red")
+     hist(cacao[,i],
+           main=paste("Histogram for ", colnames(cacao)[i]),
+           xlab=colnames(cacao)[i], freq = FALSE)
+   }
+ }
```

Las gráficas obtenidas son:





Según se observa en la gráfica, en los campos Empresa y Geo region es donde se encuentra la distribución de los datos en forma similar, porque la línea en rojo tiende a ser lineal.

4.3.3 Regresión lineal

El modelo pertenece a los algoritmos supervisados, los cuales tienen claro el objetivo que se persigue y generan un resultado concreto.

Se realiza la lectura del directorio y archivo.

Se usa regresores cuantitativos para poder realizar predicciones de las calificaciones, según características propias del chocolate.

Se obtiene algunos modelos de regresión (5), para lo cual se utilizará aquellas variables que estén más correlacionadas con respecto a la calificación o rating y de esta manera obtener un modelo de regresión lineal eficiente. El modelo que será más útil es el que presente mayor coeficiente de determinación (R²). Así:

```
> setwd("D:/")
> cacao <- read.csv(file = "flavors_of_cacao_clean.csv", header=TRUE)
> # Regresores cuantitativos
> emp = cacao$Empresa
> reg = cacao$Geo.region
> por = cacao$Porcentaje.de.cacao
> loc = cacao$Localizacion
> tip = cacao$Tipo.de.frijol
> # Variable a predecir
> cal = cacao$Calificacion
> # Generación de varios modelos
> modelo1 <- lm(cal ~ emp + reg + tip + loc, data = cacao)
> modelo2 <- lm(cal ~ reg + por + loc, data = cacao)
> modelo3 <- lm(cal ~ reg + loc + tip, data = cacao)
> modelo4 <- lm(cal ~ tip + loc + por, data = cacao)
> modelo5 <- lm(cal ~ por + reg, data = cacao)

> # Tabla con coeficientes de determinación de cada modelo
> tabla.coeficientes <- matrix(c(1, summary(modelo1)$r.squared,
+                                 2, summary(modelo2)$r.squared,
+                                 3, summary(modelo3)$r.squared,
+                                 4, summary(modelo4)$r.squared,
+                                 5, summary(modelo5)$r.squared),
+                                 ncol = 2, byrow = TRUE)
> colnames(tabla.coeficientes) <- c("Modelo", "R^2")
> tabla.coeficientes
  Modelo      R^2
[1,] 1 0.01244958
[2,] 2 0.02956991
[3,] 3 0.01175872
[4,] 4 0.03167193
[5,] 5 0.02161464
> |
```

El modelo que resulta ser más útil es el cuarto porque con este se obtiene un mayor coeficiente de determinación.

Entonces si se usa el modelo 4, ingresando correspondientes valores, se predice su calificación.

Para el ejemplo, se ha colocado a tipo de frejol = 8, localizacion = 18 y porcentaje de cacao = 0.9; la calificación que se predijo es de 3.06734

```
> nuevodato <- data.frame(  
+   tip = 8,  
+   loc = 18,  
+   por = 0.9  
+ )  
> # Predecir la calificación  
> predict(modelo4, nuevodato)  
    1  
3.06734  
> |
```

5 Representación de los resultados a partir de tablas y gráficas.

El algoritmo usado para determinar la correlación que existe entre los diferentes campos del archivo flavors_of_cacao_clean.csv es el que se presenta a continuación.

```
> setwd("D:/")  
> cacao <- read.csv(file = "flavors_of_cacao_clean.csv", header=TRUE)  
> num <- 0  
> cont <- 2  
> for (i in 1:ncol(cacao)){  
+   if (cont < ncol(cacao)+1){  
+     for (j in cont:ncol(cacao)){  
+       correlacion <- cor(cacao[[i]],cacao[[j]])  
+       cat("Entre el campo: ",colnames(cacao)[i],"y el campo:",colnames(cacao)[j],  
"la correlación es: ",correlacion,"\\n\\r");  
+       plot(cacao[[i]],cacao[[j]],  
+             main = "Correlaciones entre campos",  
+             ylab = paste("Campo",colnames(cacao)[j]),  
+             xlab = paste("Campo",colnames(cacao)[i]),  
+             col = "red")  
+       cat ("\\n\\r-----\\n\\r");  
+       num <- num + 1  
+     }  
+   }  
+   cont <- cont + 1  
+ }
```

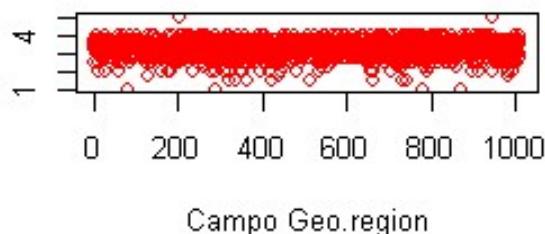
Los resultados que se obtuvieron, teniendo presente como campo principal de relación al de **Calificacion**, son los que se detallan en las interrogantes planteadas. Así se tiene:

- **¿Dónde se cultivan los mejores granos de cacao? (Relación entre geo-region y calificación).**

Entre el campo: Geo.region y el campo: Calificacion, la correlación es:
-0.02025284

Campo Calificacion

Correlaciones entre campos



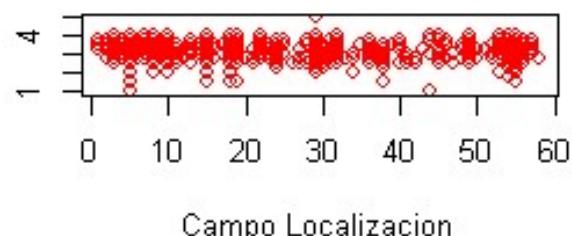
Tal y como se aprecia en la gráfica, los mejores granos de cacao se cultivan en las regiones comprendidas entre los valores 200 a 220, los más relevantes corresponden a Chuao y los que están entre 940 a 945, los más relevantes correspondientes a Toscano Black.

- **¿Qué países producen las barras mejor calificadas? (Relación entre localización y calificación)**

Entre el campo: Localizacion y el campo: Calificacion la correlación es:
-0.09390611

Campo Calificacion

Correlaciones entre campos



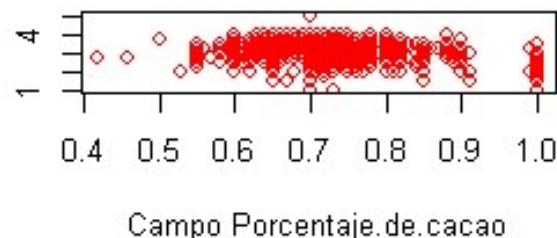
Los países que producen las mejores barras de chocolate son las localizaciones 29, 18, 8 49 y 55 correspondiente a Italy, France, Canada, Spain y U.S.A.

- **¿Cuál es la relación entre el porcentaje de sólidos de cacao y la calificación?**

Entre el campo: Porcentaje.de.cacao y el campo: Calificacion la correlación es: -0.1446795

Campo Calificación

Correlaciones entre campos

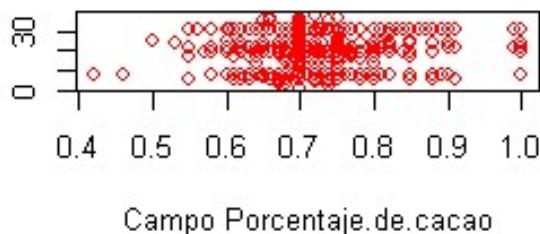


Según la gráfica, cuando el porcentaje de cacao está en 0.7, que corresponde al 70%, se alcanza la mayor calificación que es 5.

- **¿Existe alguna relación entre el porcentaje de sólidos de cacao y el tipo de frijol?**
Entre el campo: Porcentaje.de.cacao y el campo: Tipo.de.frijol la correlación es: -0.008126366

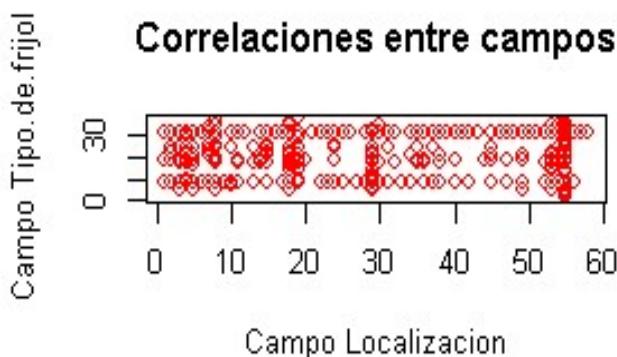
Campo Tipo.de.frijol

Correlaciones entre campos



Se determina que cuando el porcentaje de cacao está en 0.7 que es el 70% (con mayor distribución, según se observa en la gráfica), el tipo de frijol más usado es el 31 y 8 que son los tipos Trinitario y Criollo respectivamente.

- **¿Existe alguna relación entre localización y tipo de frijol?**
Entre el campo: Localizacion y el campo: Tipo.de.frijol la correlación es: 0.03954716



Según se aprecia, donde existe una distribución mayor entre localización y tipo de frijol es cuando la localización toma el valor de 55, es decir; en U.S.A. y tipo de frijol está entre 31 y 35, es decir; es trinitario.

6 Resolución del problema. A partir de los resultados obtenidos, ¿cuáles son las conclusiones? ¿Los resultados permiten responder al problema?

Las conclusiones que se obtiene son:

- ✓ Todo el tratamiento de depuración de datos se lo ejecutó desde la herramienta estadística R studio.
- ✓ Para trabajar de mejor manera con los datos, a estos se los discretizó y configuró con tipo de dato numérico.
- ✓ Con la opción de los vecinos más cercanos se completó los valores vacíos existentes en las variables del archivo.
- ✓ Dentro de esta práctica, se usó el algoritmo k-means para crear los clusters, y poder determinar los valores extremos en los grupos, el objetivo principal de este algoritmo es el de maximizar la variación inter-cluster entre los propios grupos y minimizar la intra-cluster dentro de cada grupo.
- ✓ Se usaron histogramas para realizar el análisis de la selección de las variables.
- ✓ Las variables que se comprobaron y las únicas que tienen valores extremos son: Porcentaje de cacao y Calificacion; para realizar dicha comprobación, se usó la función boxplot.stats()
- ✓ Para concluir que una distribución es la mejor, el estadístico de Anderson-Darling debe ser sustancialmente menor que los demás, es decir; en nuestro caso, los datos no siguen una distribución especificada.

- ✓ La homogeneidad de varianzas se la determinó, mediante el test de Fligner-Killeen, el cual se trata un test no paramétrico que compara las varianzas basándose en la mediana.
- ✓ Según lo planteado al inicio de la práctica, la variable Calificacion es la que más se consideró o predijo al momento de realizar las diferentes pruebas estadísticas sobre un conjunto de datos.
- ✓ Los dos campos que tienen mayor correlación entre sí son Porcentaje de cacao y Calificacion, este valor es de -0.1446795
- ✓ Se obtiene la regresión lineal simple a partir de las correlaciones que puedan existir entre los campos.
- ✓ Se generó un nuevo dataset llamado flavors_of_cacao_clean.csv como resultado de la limpieza de caracteres especiales, de valores vacíos, valores mal digitados, etc.
- ✓ Se realizaron 3 tipos de pruebas estadísticas que son:
 - a. Covarianza y correlación
 - b. Prueba de contraste gráfico
 - c. Regresión lineal.

7 Código.

El código en R consta dentro del fichero chocolateCocoa.R que se puede descargar en GitHub, desde la carpeta código.

Los datos de salida, se exportaron mediante el siguiente comando: write.csv(cacao, "D:/flavors_of_cacao_clean.csv", row.names=F)

Y también se puede descargar en GitHub, desde la carpeta csv, en el siguiente enlace:

<https://github.com/mmagdarom/LimpiezaDatos>

8 Recursos.

Covarianza y correlación

- ✓ <https://www.youtube.com/watch?v=nCnscXRG8Ws>

Homogeneidad de varianza

- ✓ https://rpubs.com/Joaquin_AR/218466

El estadístico de Anderson-Darling

- ✓ <https://support.minitab.com/es-mx/minitab/18/help-and-how-to/statistics/basic-statistics/supporting-topics/normality/the-anderson-darling-statistic/>

Prueba de Contraste Gráfico

- ✓ <https://www.youtube.com/watch?v=52Q2gq7tgI0>

Test for homogeneity of variances

- ✓ <https://biostats.w.uib.no/test-for-homogeneity-of-variances-levenes-test/>

Data Cleaning Basics

- ✓ Jason W. Osborne (2010). *Data Cleaning Basics: Best Practices in Dealing with Extreme Scores*. Newborn and Infant Nursing Reviews