Limpieza y Validación de Datos

Practica 1

Descripción

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.

Miembros del equipo: Jennifer Samaniego

Contenido

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 datos 4

3.1 ¿Los datos contienen ceros o elementos vacíos? ¿Cómo gestionarías cada uno de estos casos? 5

3.1.1 Limpieza variable empresa 8

3.1.2 Limpieza variable Geo_region 8

3.1.3 Limpieza variable porcent_cacao 9

3.1.4 Limpieza variable locaci_empresa 9

3.2 Identificación y tratamiento de valores extremos 9

3.2.1 Discretización variable empresa 10

3.2.2 Discretización Geo_region 11

3.2.3 Discretización locaci_empresa 11

3.2.4 Transformación porcentaje cacao 12

4. Análisis de los datos. 16

4.1. Selección de los grupos de datos que se quieren analizar/comparar (planificación de los análisis a aplicar). 16

4.2. Comprobación de la normalidad y homogeneidad de la varianza. 18

4.2.1 Normalidad 18

4.2.2 Homogeneiedad 19

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. ... 19

4.3.1 Covarianza y correlación 19

4.3.2 Contraste gráfico 22

5. Representación de los resultados a partir de tablas y gráficas. 24

6. Resolución del problema. A partir de los resultados obtenidos, ¿cuáles son las conclusiones? ¿Los resultados permiten responder al problema? 26

7. Código: Hay que adjuntar el código, preferiblemente en R, con el que se ha realizado la limpieza, análisis y representación de los datos. Si lo preferís, también podéis trabajar en Python. 26

8. Bibliografía 26

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

El dataset a utilizar tiene datos sobre la barra de chocolate a nivel mundial, existen 1700 calificaciones de expertos de barras de chocolate. Contiene información sobre el nombre de la empresa que fabrica la barra, georegión origen de la barra, el porcentaje, variedad de chocolate.

Los rangos varían desde 1 (desagradable) hasta 5 (elite), a continuación, detallo el sistema de clasificación de sabores de cacao que se menciona en la guía de flavorsofcacacao (http://www.flavorsofcacao.com/review_guide.html):

Calificación	Clasificación	Descripción
5	Elite	Trancender más allá de los límites ordinarios
4	Premium	Desarrollo de sabor superior, carácter y estilo
3	Satisfactorio	(3.0) a loable (3.75) (bien hecho con cualidades especiales)
2	Decepcionante	Transitable, pero contiene al menos un defecto significativo
1	Desagradable	en su mayoría desagradable

El dataset se enfoca en el chocolate negro con el objetivo de apreciar los sabores del cacao cuando se transforma en chocolate.

Las calificaciones no tienen que ver con ventajas o desventajas para la salud. Para la calificación de la barra de chocolate existen varios parámetros que se toman en cuenta como: sabor, textura, aftermelt, opinión general. Se pueden observar 1795 registros con 9 atributos que se detallan a continuación:

#	Nombre	Descripción	Tipo
1	Company (Maker-if known) Company (Make	Nombre de la empresa que fabrica la barra	String
2	Specific Bean Origin or Bar Name	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	Review Date	Fecha de publicación de la revisión	Numeric
5	Cocoa Percent	Porcentaje de cacao (oscuridad) de la barra de chocolate que se revisa	Numeric
6	Company Location	País base del fabricante	String
7	Rating	Calificación de expertos	Numeric
8	Bean Type	La variedad de frijol utilizada, en el caso de que se proporcione	String
9	Broad Bean Origin	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? (Relación entre localización y calificación)
- ¿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.

Luego de haber realizado una revisión previa del dataset tomando en cuenta la preguntas que nos planteamos se concluyó que los atributos que debemos utilizar son:

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

3. Limpieza de datos

Antes de proceder con la limpieza de datos realice la lectura del archivo .CSV, además de una visualización general de los datos con el comando **head:**

```
> ##Lectura del DataSet y guardamos como un data frame para evitar errores a futuro
> cacao <- read.csv(file="C:/Users/jbsamaniego/Documents/2018/Maestria/TipologiaCicloDat
os/Prac2/flavors_of_cacao.csv",header=TRUE )
> cacao <- as.data.frame(cacao, stringsAsFactors=FALSE)
> # Revisión y lectura de los 10 primeros registros:
> head(cacao, 10)
   CompanyÂ...Maker.if.known. Specific.Bean.Origin.or.Bar.Name  REF Review.Date
1                   A. Morin                      Agua Grande 1876        2016
2                   A. Morin                            Kpime 1676        2015
3                   A. Morin                           Atsane 1676        2015
4                   A. Morin                            Akata 1680        2015
5                   A. Morin                           Quilla 1704        2015
6                   A. Morin                         Carenero 1315        2014
7                   A. Morin                             Cuba 1315        2014
8                   A. Morin                       Sur del Lago 1315        2014
9                   A. Morin                    Puerto Cabello 1319        2014
10                  A. Morin                           Pablino 1319        2014
   Cocoa.Percent Company.Location Rating Bean.Type Broad.Bean.Origin
1            63%           France   3.75        Â           Sao Tome
2            70%           France   2.75        Â               Togo
3            70%           France   3.00        Â               Togo
4            70%           France   3.50        Â               Togo
5            70%           France   3.50        Â               Peru
```

Luego de observar los nombres de las cabeceras originales observe que es necesari o realizar un cambió al nombre mediante el código, Además de verificar los cambios realizados revisamos los tipos de variables asignados a cada uno:

```
> colnames(cacao) <- c('Empresa','Geo_region','REF','fechaRevi','porcent_cocoa',
+                      'locaci_empresa','rating','tipo_fijol','origen_frijol')
> #Se realiza la revisión de tipo de variables asignado en cada atributo
> sapply(cacao, function(x) class(x))
       Empresa     Geo_region            REF      fechaRevi  porcent_cocoa
      "factor"       "factor"      "integer"      "integer"       "factor"
locaci_empresa         rating     tipo_fijol  origen_frijol
      "factor"      "numeric"       "factor"       "factor"
```

3.1 ¿Los datos contienen ceros o elementos vacíos? ¿Cómo gestionarías cada uno de estos casos?

Al haber leído anticipadamente el dataset pude observar caracteres extraños en la variable tipo_frijol:

```
> head(cacao, 10)
    Empresa     Geo_region  REF fechaRevi porcent_cocoa locaci_empresa rating
1  A. Morin    Agua Grande 1876      2016           63%         France   3.75
2  A. Morin          Kpime 1676      2015           70%         France   2.75
3  A. Morin         Atsane 1676      2015           70%         France   3.00
4  A. Morin          Akata 1680      2015           70%         France   3.50
5  A. Morin         Quilla 1704      2015           70%         France   3.50
6  A. Morin       Carenero 1315      2014           70%         France   2.75
7  A. Morin           Cuba 1315      2014           70%         France   3.50
8  A. Morin   Sur del Lago 1315      2014           70%         France   3.50
9  A. Morin Puerto Cabello 1319      2014           70%         France   3.75
10 A. Morin        Pablino 1319      2014           70%         France   4.00
   tipo_fijol origen_frijol
1           Â       Sao Tome
2           Â           Togo
3           Â           Togo
4           Â           Togo
5           Â           Peru
6     Criollo      Venezuela
7           Â           Cuba
8     Criollo      Venezuela
9     Criollo      Venezuela
10          Â           Peru
```

Por lo cual se hizo el cambio del carácter especial por una cadena vacía y luego fue remplazada por NA, el archivo fue guardado con un nuevo nombre "flavors_of_cacaoVacios":

```
> #A cotinuación cambio los caracteres especiales por NA
> '%!in%' <- function(x,y)!('%in%'(x,y))
> to_NA <- function(x, start){ #start is the index of the first non-NA piece of data i
n levels(x)
+   for (i in 1:length(x)){
+     if (x[i] %!in% levels(x)[start:length(levels(x))]){
+       x[i] <- NA
+     }
+   }
+   return(x)
+ }
> cacao$tipo_fijol <- to_NA(cacao$tipo_fijol, 3)
> cacao$origen_frijol <- to_NA(cacao$origen_frijol, 3)
> #Revisión de los cambios realizados
> head(cacao, 4)
   Empresa  Geo_region  REF fechaRevi porcent_cocoa locaci_empresa rating
1 A. Morin Agua Grande 1876      2016           63%         France   3.75
2 A. Morin       Kpime 1676      2015           70%         France   2.75
3 A. Morin      Atsane 1676      2015           70%         France   3.00
4 A. Morin       Akata 1680      2015           70%         France   3.50
  tipo_fijol origen_frijol
1       <NA>      Sao Tome
2       <NA>          Togo
3       <NA>          Togo
4       <NA>          Togo
> #Guardar un nuevo archivo con los cambios realizados
> write.csv(cacao[1:9],file="C:/Users/jbsamaniego/Documents/2018/Maestria/TipologiaCic
loDatos/Prac2/flavors_of_cacaoVacios.csv", row.names = FALSE)
>
```

Indagamos el número de elementos NA en cada variable

```
> sapply(cacao, function(x) sum(is.na(x)))
       Empresa     Geo_region            REF      fechaRevi  porcent_cocoa
             0              0              0              0              0
locaci_empresa         rating     tipo_fijol  origen_frijol
             0              0            888             74
```

Al observar el resultado en la variable tipo_frijol son 888 registro con valores NA se decide tratar estos datos de forma numérica, pero previamente se realiza una revisión bibliográfica del tipo_frijol encontrando un análisis por expertos quienes indican que las categorías generales son las más importantes: forastero, criollo, trinitario, blend y nacional en base a lo que indican [1], [2], [3]. Por ello se procedió a re categorizar la variable tipo_frijol con estas 4 categorías. Como se observa a continuación:

```
> ## Se observa en la variable tipo_frijol varias sub categorias de frijoles,se decide
 realizar una recategorización en base la
> ##categorización general en donde expertos indican que son más importantes las categ
orías generales
> cacao$tipo_fijol <- factor(cacao$tipo_fijol)
> cacao$tipo_fijol <- as.character(cacao$tipo_fijol)
> criollo <- c("Criollo", "Criollo (Ocumare)", "Criollo (Porcelana)","Criollo (Ocumare
 77)", "Criollo (Ocumare 61)", "Criollo (Amarru)", "Criollo (Wild)", "Criollo (Ocumare
 67)",  "Beniano", "EET" )
> cacao$tipo_fijol[which(cacao$tipo_fijol %in% criollo)] <- 'criollo'
> forastero <- c("Forastero", "Forastero (Amelonado)", "Forastero (Arriba)", "Foraster
o (Parazinho)","Forastero (Arriba) ASSS", "Forastero (Catongo)", "Forastero(Arriba, CC
N)", "Forastero (Arriba) ASS", "CCN51", "Matina")
> cacao$tipo_fijol[which(cacao$tipo_fijol %in% forastero)] <- 'forastero'
> nacional <- c("Forastero (Nacional)", "Nacional", "Nacional (Arriba)")
> cacao$tipo_fijol[which(cacao$tipo_fijol %in% nacional)] <- 'nacional'
> trinitario <- c("Trinitario", "Trinitario (Amelonado)", "Trinitario (Scavina)", "Tri
nitario, TCGA", "Trinitario (85% Criollo)")
> cacao$tipo_fijol[which(cacao$tipo_fijol %in% trinitario)] <- 'trinitario'
> blend <- c("Amazon, ICS", "Blend", "Criollo, +", "Criollo, Forastero",
+            "Criollo, Trinitario", "Forastero, Trinitario", "Trinitario, Criollo",
+            "Trinitario, Forastero", "Trinitario, Nacional", "Amazon", "Amazon mix",
"Blend-Forastero,Criollo")
> cacao$tipo_fijol[which(cacao$tipo_fijol %in% blend)] <- 'blend'
> cacao$tipo_fijol <- factor(cacao$tipo_fijol)
> ## Revisamos los cambios efectuados
> summary(cacao$tipo_fijol)
     blend    criollo  forastero   nacional trinitario       NA's
       102        177        147         57        424        888
>
```

Una vez con el resultado esperado, se procede dar un cambio de la variable tipo_frijol de "Factor" a "Numérico".

```
> ##Procedemos a discretizar la variable tipo_frijol a valores numéricos
> cacao$tipo_fijol <- recode(cacao$tipo_fijol,'"criollo"=1;"forastero"=2;"nacional"=3;
"trinitario"=4;"blend"=5;')
> ##Procedemos a discretizar la variable tipo_frijol a valores numéricos
> cacao$tipo_fijol <- recode(cacao$tipo_fijol,'"criollo"=1;"forastero"=2;"nacional"=3;
"trinitario"=4;"blend"=5;')
> ## Revisamos los cambios efectuados
> summary(cacao$tipo_fijol)
   1    2    3    4    5 NA's
 177  147   57  424  102  888
```

Al ver la cantidad de valores de NA en el dataset utilizo el método basado en k vecinos más próximos, porque eliminar estos registros no es conveniente para el análisis respectivo y la cantidad de registros que se perderían si eliminamos los mismos.

```
> cacao$tipo_fijol <- kNN(cacao)$tipo_fijol
> ## Revisamos los cambios efectuados
> summary(cacao$tipo_fijol)
   1    2    3    4    5
 277  224   83 1052  159
```

Las variables que se usarán son solo 6 por lo cual se procede a eliminar los 3 restantes:

```
> cacao <- cacao[,-(9)]
> cacao <- cacao[,-(3:4)]
> ## Revisamos los cambios efectuados
> head(cacao,4)
   Empresa  Geo_region porcent_cocoa locaci_empresa rating tipo_fijol
1 A. Morin Agua Grande           63%         France   3.75          4
2 A. Morin       Kpime           70%         France   2.75          4
3 A. Morin      Atsane           70%         France   3.00          4
4 A. Morin       Akata           70%         France   3.50          4
>
```

Ante de pasar a revisar y analizar los valores extremos realice un análisis del resto de datos en los cuales se encontró varias inconsistencias, por ello se realiza una limpieza previa como en nombres, caracteres especiales.

3.1.1 Limpieza variable empresa

Dentro de esta variable se encontró algunas inconsistencias como nombres incorrectos, caracteres especiales, etc. Por lo cual se aplicó el siguiente código con los resultados:

```
> #Ver el caracter especial
> print(cacao[1166,1])
[1] Naï¿½ve
416 Levels: A. Morin Acalli Adi Aequare (Gianduja) Ah Cacao ... Zotter
> #Reemplazo del nombre de la empresa Naive que se encuentra con caracteres es
> cacao$Empresa <- recode(cacao$Empresa,'"Naï¿½ve"="Naive";
+ "Cacao de Origin"="Cacao de Origen";"Shattell"="Shattel";')
> #se comprueba la linea 1166,1167,1168 que ya no tenga el caracter especial
> print(cacao[1166,1])
[1] Naive
413 Levels: A. Morin Acalli Adi Aequare (Gianduja) Ah Cacao ... Zotter
>
```

```
> #se verifica el caracter ' en la variable
> print(cacao[466,1])
[1] Cote d' Or (Kraft)
413 Levels: A. Morin Acalli Adi Aequare (Gianduja) Ah Cacao ... Zotter
> #se reemplaza los ' ya que provocarán problemas a futuro en la recodi
> cacao$Empresa <- gsub("'","",cacao$Empresa)
> #se comprueba las lineas 466,620, por ejemplo que ya no tenga el apos
> print(cacao[466,1])
[1] "Cote d Or (Kraft)"
>
```

3.1.2 Limpieza variable Geo_region

Dentro de esta variable se encontró algunas inconsistencias como nombres incorrectos, caracteres especiales, etc. Por lo cual se aplicó el siguiente código con los resultados:

```
> ##################Limpieza Geo_region#################
> #se reemplaza en la Geo_region el "
> cacao$Geo_region <- gsub('"',"",cacao$Geo_region)
> #se comprueba el reemplazo
> print(cacao[1608,2])
[1] "Ayacucho, El Guinacho"
> #se reemplaza en la Geo_region el '
> cacao$Geo_region <- gsub("'","",cacao$Geo_region)
> #se comprueba el reemplazo
> print(cacao[795,2])
[1] "Maunawili, Oahu, Agri Research C., 2014"
> #se reemplaza en la Geo-region el ;
> cacao$Geo_region <- gsub(';',",",cacao$Geo_region)
> #se comprueba el reemplazo
> print(cacao[293,2])
[1] "Macuare, Miranda, Chloe formula"
> #En la variable Geo_region existe nombres con * por lo cual procedí a quitarlos
> cacao$Geo_region<- recode(cacao$Geo_region,'"Concepcion*"="Concepcion";
+                           "Capistrano*"="Capistrano";"Equateur"="Ecuador";"Ambolika
piky P."="Ambolikapiky";
+                           "Ambolikapkly P."="Ambolikapiky";"Alto Beni, Palos Blanco
"="Alto Beni, Palos Blancos";
+                           "Chiapan"="Chiapas";"Brazilian"="Brazil";"Bolivian"="Boli
via";"Colombie"="Colombia";
+                           "Colombian"="Colombia";"Dominican Republicm, rustic"="Dom
inican Republic, rustic";
+                           "Fazenda Sempre Firme P., Bahia"="Fazenda Sempre Firme, B
ahia";
+                           "La Red, Guanconjeco"="La Red, Guaconejo";"Madagared"="Ma
dagascar";
+                           "Monte Alegre, Diego Badero"="Monte Alegre, D. Badero";"N
icaraqua"="Nicaragua";
+                           "Trinidad-Tobago"="Trinidad & Tobago";"Venzuela"="Venezue
la";
```

3.1.3 Limpieza variable porcent_cacao

Dentro de esta variable se encontró algunas inconsistencias como nombres incorrectos, caracteres especiales, etc. Por lo cual se aplicó el siguiente código con los resultados:

```
> #Procedo a cambiar el valor de porcentaje de cacao por valores enteros
> cacao$porcent_cocoa <- as.integer(sub("%", "", cacao$porcent_cocoa))
> #Verificamos el cambio realizado
> head(cacao$porcent_cocoa,10)
 [1] 63 70 70 70 70 70 70 70 70 70
```

3.1.4 Limpieza variable locaci_empresa

En esta variable se encontró inconsistencia en los nombres.

```
> ###############Limpieza locaci_empresa###################
> #se modifica los nombres incorrectos
> cacao$locaci_empresa <- recode(cacao$locaci_empresa,'"Eucador"="Ecuador";
+                            "Niacragua"="Nicaragua";"Domincan Republic"="Dominican
Republic";')
> #se verifica los cambios
> print(cacao[884,4])
[1] Dominican Republic
58 Levels: Amsterdam Argentina Australia Austria Belgium Bolivia Brazil ... Wales
```

Se elimina los siguientes registros ya que el nombre de la región no existe

cacao <- cacao[-246,]

cacao <- cacao[-779,]

cacao <- cacao[-1410,]

3.2 Identificación y tratamiento de valores extremos

Para poder evaluar los valores extremos es necesario que todas las variables tengan un mismo formato por ello vamos a trasformar a todas las variables en formato

numérico es decir debemos realizar una discretización a las variables, empezamos con el campo Empresa, para esto se carga la librería car para realizar la codificación de los valores NA encontrados, y se comprueba que se hayan reemplazado correctamente:

3.2.1 Discretización variable empresa

Tomando en cuenta que esta variable es un factor procedemos a la transformación numérica de la misma.

```
> cacao$Empresa <- recode(cacao$Empresa,'"A. Morin"=1;
+ "Acalli"=2;
+                         "Adi"=3;
+                         "Aequare (Gianduja)"=4;
+                         "Ah Cacao"=5;
+                         "Akessons (Pralus)"=6;
+                         "Alain Ducasse"=7;
+                         "Alexandre"=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;
+                         "Bahen & 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 procede a verificar la transformación realizada:

```
> ##Verificar cambios nnecesarios.
> summary.factor(cacao$Empresa)
  1   2   3   4   5   6   7   8   9  10  11  12  13  14  15  16  17  18  19  20
 23   2   4   2   1   3   5   4  10   9   4   2   6  13   5   1   1   4  22  16
 21  22  23  24  25  26  27  28  29  30  31  32  33  34  35  36  37  38  39  40
  1   6   5   2   5   1   3   2   4   4   3   3   6   1   8   4  14   3   1   2
 41  42  43  44  45  46  47  48  49  50  51  52  53  54  55  56  57  58  59  60
  6  26   2   2   9   6   1   4   1   1   5   1   4   5   5   7   1   4   9   3
 61  62  63  64  65  66  67  68  69  70  71  72  73  74  75  76  77  78  79  80
  1   4   2   7   5   6   2   5   5  14   4   1   1   1   2   1   1   1   2   2
 81  82  83  84  85  86  87  88  89  90  91  92  93  94  95  96  97  98  99 100
  3   2   2   7   1   3   6   1   1   1   8   2   5   1   1   1   1   1   4   1
101 102 103 104 105 106 107 108 109 110 111 112 113 114 115 116 117 118 119 120
  1   5   1   1  18   1   7   2   2   1   4  16   9   3   5   4   2   2   6   2
121 122 123 124 125 126 127 128 129 130 131 132 133 134 135 136 137 138 139 140
  1   3  13   4   1   2  22   4  13   1   1   5   3   2   1   1   2   7   2   2
141 142 143 144 145 146 147 148 149 150 151 152 153 154 155 156 157 158 159 160
  1   1   3   5   5   3   1   3   7   4   1   2   3   4   5  10  26   1  12   9
161 162 163 164 165 166 167 168 169 170 171 172 173 174 175 176 177 178 179 180
  1   5   2   3   1   1   1   3   5  22   9   1   1   1   4   1   1   1   4   1
181 182 183 184 185 186 187 188 189 190 191 192 193 194 195 196 197 198 199 200
  1   1   2   2   5   4   5   9   1   3  19   8  10   1   2   1   2   4   2   1
201 202 203 204 205 206 207 208 209 210 211 212 213 214 215 216 217 218 219 220
  1   3   4   2   8   2   1   1   1   7   4   4   9   4  10   1   1  10   1   1
221 222 223 224 225 226 227 228 229 230 231 232 233 234 235 236 237 238 239 240
  4   9   1   1   6   1   1   5   1   4   2   4  11   4   1   2   1   2   1   4
241 242 243 244 245 246 247 248 249 250 251 252 253 254 255 256 257 258 259 260
  1   7   7  10   3   1  10   4   2   3  18   4   4   1   5   2   6   2   1  10
```

3.2.2 Discretización Geo_region

Procedemos a discretizar de la misma forma que la variable empresa:

```
cacao$Geo_region <- recode(cacao$Geo_region,'"heirloom, Arriba Nacional"=1;
"2009 Hapa Nibby"=2;
                           "A case of the Xerces 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;
```

Verificamos los cambios efectuados:

```
> ##Verificar cambios necesarios.
> print(cacao[,2])
   [1]   14  476   64   15  789  166  275  899  781  707  712  547  129  321  229
  [16]  111  720  203  747  189  189  125  730  219  964  977  979  977  978  536
  [31]  536  906   84  550  638  954  986  547  203  755 1004  485  930  560  633
  [46]    8  225  989  387  899  237  125  125  730  641  313  400  203  634   85
  [61]  547  281  675  363  321  321  363  507  104   96  547  299  720  982  730
  [76]  752  760  659  203  321  442  383  982  547  950  944  944  944  185  636
  [91]  636  636  636  339   27  547  192  321  949  846  470  930  404  257  499
 [106]  815  490  263  741  339  603  505  352  645  784  689  609  220  352  399
 [121]  146  952  231  405  712  982  442  258   69   83  688  450  131  547  730
 [136]  299  246  702  543  922  256  292 1006  845  425  720  821   71  425   89
 [151]  426   71  582  316  192  821  879  303  129  321   61  870  132  304  325
 [166]  321  258  720  730  321  460   30  692   98   97   97   90  203  282  806
 [181]   87  887  650  547  574  434  321  167  768  451  684  656  840  841   96
 [196]  363   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  904
```

3.2.3 Discretización locaci_empresa

Procedemos a dicretizar a una variable numérica

```
> cacao$locaci_empresa <- recode(cacao$locaci_empresa,'"Amsterdam"=1;"Argentina"=2;
+                                "Australia"=3;"Austria"=4;"Belgium"=5;"Bolivia"=6;"Braz
il"=7;"Canada"=8;"Chile"=9;
+                                "Colombia"=10;"Costa Rica"=11;"Czech Republic"=12;"Denm
ark"=13;"Dominican Republic"=14;
+                                "Ecuador"=15;"Fiji"=16;"Finland"=17;"France"=18;"German
y"=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;"Madag
ascar"=32;"Martinique"=33;
+                                "Mexico"=34;"Netherlands"=35;"New Zealand"=36;"Nicaragu
a"=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;"Spai
n"=49;"St. Lucia"=50;"Suriname"=51;
+                                "Sweden"=52;"Switzerland"=53;"U.K."=54;"U.S.A."=55;"Ven
ezuela"=56;"Vietnam"=57;"Wales"=58;')
>
```

Verificamos los cambios realizados:

```
print(cacao[,4])
  [1] 18 18 18 18 18 18 18 18 18 18 18 18 18 18 18 18 18 18 18 18 18 18 18 55 55
 [26] 16 16 16 16 15 15 34 53 53 53 18 18 18 18 18 35 35 35 35 55 55 55 55 55 55
 [51] 55 55 55 55 55 55 55 55 55 55 55 55 55 55 49 49 49 49 38 38 8  8  8  8  8
 [76] 29 29 29 29 29 29 29 29 29 29 29 29 29 7  7  7  7  7  55 55 18 18 18 18 55
[101] 55 55 55 55 55 55 55 55 55 55 55 55 55 55 55 55 55 55 55 55 55 54 54 54 54
[126] 54 54 54 54 54 54 54 54 54 54 54 54 54 55 55 55 55 55 55 55 3  3  3  3  3  38
[151] 38 55 55 55 55 55 58 55 55 55 54 54 55 55 55 55 5  5  5  5  55 55 55 19 19
[176] 19 5  5  5  5  5  5  18 53 53 53 53 53 53 53 53 55 55 55 55 55 55 55 55 55
[201] 55 55 55 55 55 55 55 55 55 55 55 55 54 49 49 55 55 55 55 55 55 18 18 18 18
[226] 18 18 18 18 18 18 18 18 18 18 18 18 18 18 18 18 18 18 18 18 18 18 15 15 55
[251] 55 55 55 55 55 55 55 55 55 55 55 55 55 55 55 55 3  3  3  3  43 55 55 55
[276] 55 55 55 42 55 55 55 55 18 18 18 18 18 56 56 56 56 56 56 56 10 10 10 10 10
[301] 10 10 55 55 55 55 55 49 49 49 49 49 49 49 49 49 30 30 30 38 15 15 15 15 5
[326] 5  29 29 29 29 29 29 29 55 55 55 55 55 15 15 15 15 15 15 36 36 11 11 11 11
[351] 11 10 10 10 10 10 55 55 55 55 55 55 55 55 55 55 55 55 55 55 55 55 55 55 18
[376] 8  55 15 15 55 18 54 19 19 48 48 54 54 54 55 55 55 55 55 55 55 55 55 55 55
```

3.2.4 Transformación porcentaje cacao

Dividimos la variable porcentaje para 100.

```
> cacao$porcent_cocoa <- cacao$porcent_cocoa/100
> #comprueba el resultado de Porcentaje.de.cacao
> print(cacao[,3])
  [1] 0.63 0.70 0.70 0.70 0.70 0.70 0.70 0.70 0.70 0.70 0.70 0.70 0.70 0.70 0.70 0.70
 [16] 0.70 0.70 0.70 0.70 0.70 0.63 0.70 0.63 0.70 0.70 0.60 0.80 0.88 0.72 0.55
 [31] 0.70 0.70 0.75 0.75 0.75 0.65 0.75 0.75 0.75 0.75 0.70 0.70 0.70 0.70 0.60
 [46] 0.60 0.60 0.60 0.60 0.60 0.60 0.80 0.60 0.60 0.70 0.70 0.70 0.70 0.70 0.70
 [61] 0.70 0.70 0.70 0.70 0.70 0.85 0.85 0.72 0.73 0.64 0.66 0.75 0.63 0.70 0.68
 [76] 0.70 0.70 0.75 0.70 0.70 0.70 0.70 0.70 0.70 0.70 0.63 0.70 0.66 0.75 0.85
 [91] 0.50 0.75 0.60 0.75 0.75 0.75 0.72 0.75 0.75 0.70 0.70 0.73 0.70 0.70 0.70
```

Procedemos a verificar los tipos asignados a cada variable:

```
> sapply(cacao, function(x) class(x))
      Empresa    Geo_region  porcent_cocoa locaci_empresa         rating
    "numeric"     "numeric"     "numeric"     "numeric"     "numeric"
   tipo_fijol
    "numeric"
```

3.2.4 Revisión de valores máximos

Con la discretización realizada vamos a emplear el algoritmo kmeans para poder evaluar si existen valores extremos. Por ello se fija una semilla es decir se fija el punto inicial del grupo. En este caso cree la semilla con un valor de 80. Tomando en cuenta que el algoritmo Kmeans selecciona las observaciones de forma aleatoria:

```
> #Kmeans algoritmo para revisar valores extremos
> set.seed(80)
> cacaoEx <- kmeans(cacao,centers=4)
> print(cacaoEx)#se presenta todo con el names también
K-means clustering with 4 clusters of sizes 509, 565, 350, 368

Cluster means:
    Empresa Geo_region porcent_cocoa locaci_empresa   rating tipo_fijol
1 209.22200   547.0943     0.7122986       39.16110 3.229862   3.337917
2 206.92035   847.4000     0.7144956       38.28850 3.160177   3.178761
3 320.99143   212.3486     0.7224571       40.62000 3.177143   3.391429
4  92.29348   214.5027     0.7196739       38.64946 3.185462   3.472826

Clustering vector:
  1   2   3   4   5   6   7   8   9  10  11  12  13  14  15  16
  4   1   4   4   2   4   4   2   2   2   2   1   4   4   4   4
 17  18  19  20  21  22  23  24  25  26  27  28  29  30  31  32
  2   4   2   4   4   4   2   4   2   2   2   2   2   1   1   2
 33  34  35  36  37  38  39  40  41  42  43  44  45  46  47  48
  4   1   1   2   2   1   4   2   2   1   2   1   1   4   4   2
 49  50  51  52  53  54  55  56  57  58  59  60  61  62  63  64
  4   2   4   4   4   2   1   4   4   4   1   4   1   4   1   4
```

Revisamos a la opción de información de la asignación de las observaciones de los clusters a la que se puede acceder además de poder determinar sus diferentes medidas:

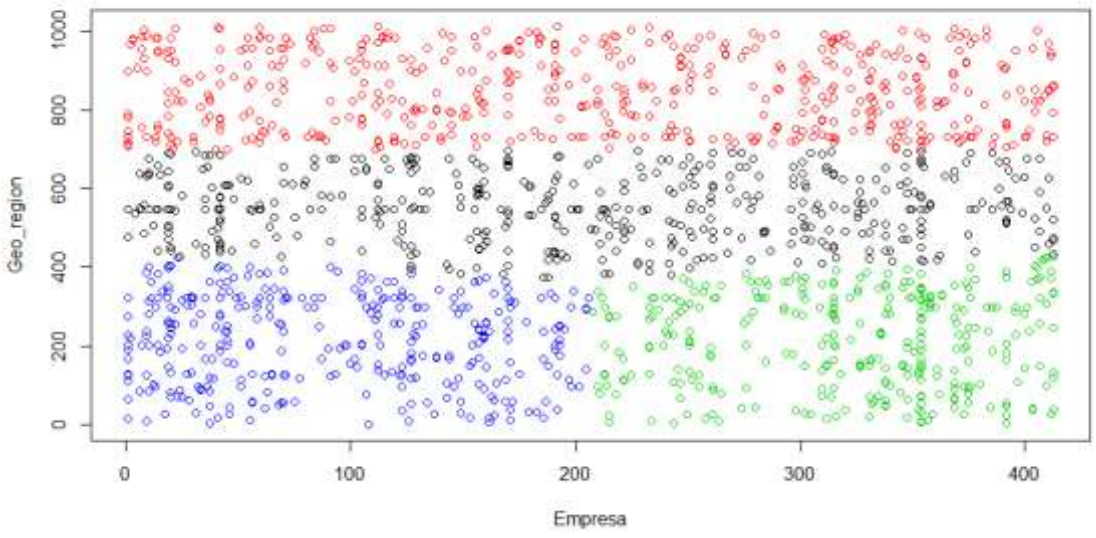
```
> names(cacaoEx) #contenido del cluster
[1] "cluster"      "centers"      "totss"        "withinss"     "tot.withinss"
[6] "betweenss"    "size"         "iter"         "ifault"
```

Podemos revisar los centros de los 4 grupos por cada una de las 6 variables:

```
> cacaoEx$centers
    Empresa Geo_region porcent_cocoa locaci_empresa   rating tipo_fijol
1 209.22200   547.0943     0.7122986       39.16110 3.229862   3.337917
2 206.92035   847.4000     0.7144956       38.28850 3.160177   3.178761
3 320.99143   212.3486     0.7224571       40.62000 3.177143   3.391429
4  92.29348   214.5027     0.7196739       38.64946 3.185462   3.472826
```

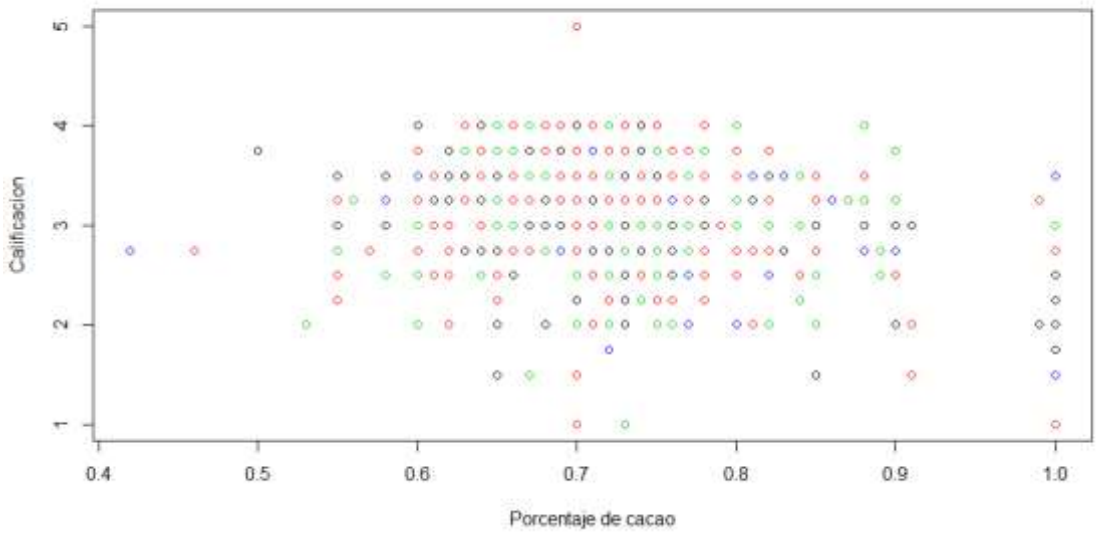
Procedemos a realizar una revisión de valores extremos por grupos de variables

3.2.4.1 Valor extremo empresa y Geo_region



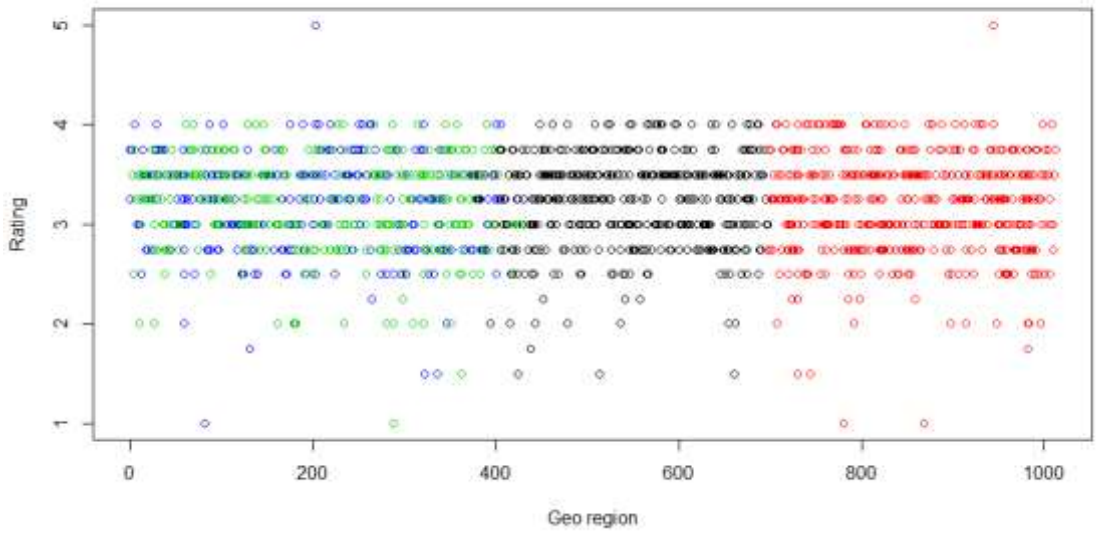
No se encuentran valores máximos entre estas variables.

3.2.4.2 Valor extremo porcent_cocoa – Rating



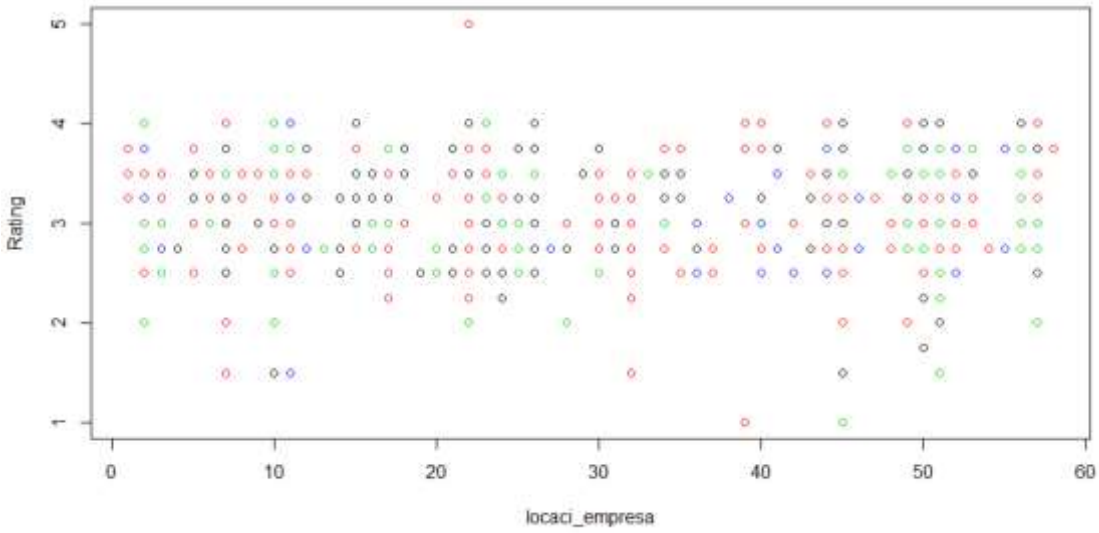
Se observan valores extremos cuando la calificación es 1 y 5 entre los valores de porcentaje de cacao de 0.7 a 0.75

3.2.4.3 Valor extremo Geo_region – Rating



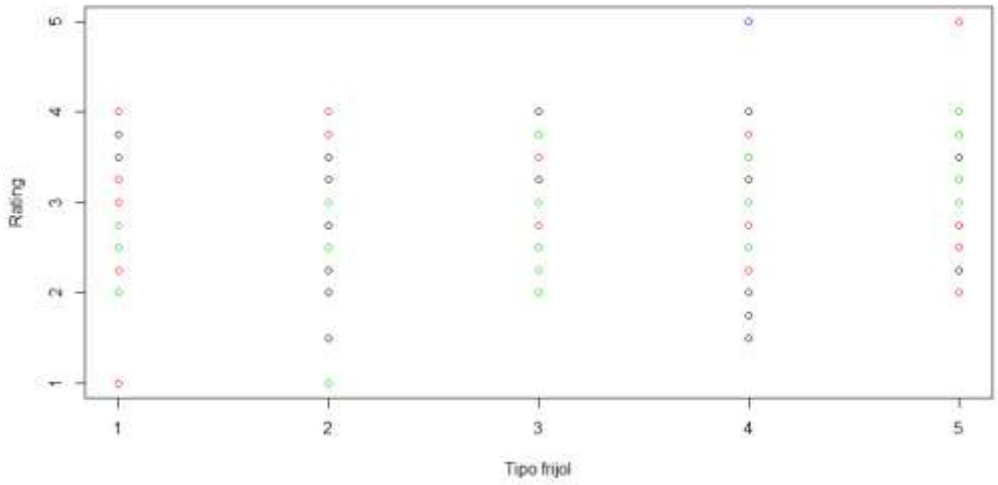
Existen 6 valores extremos de los cuales 2 tienen el rating más alto y 4 el rating más bajo estos valores los mantenemos para un estudio posterior.

3.2.4.4 Valor extremo locaci_empresa– Rating



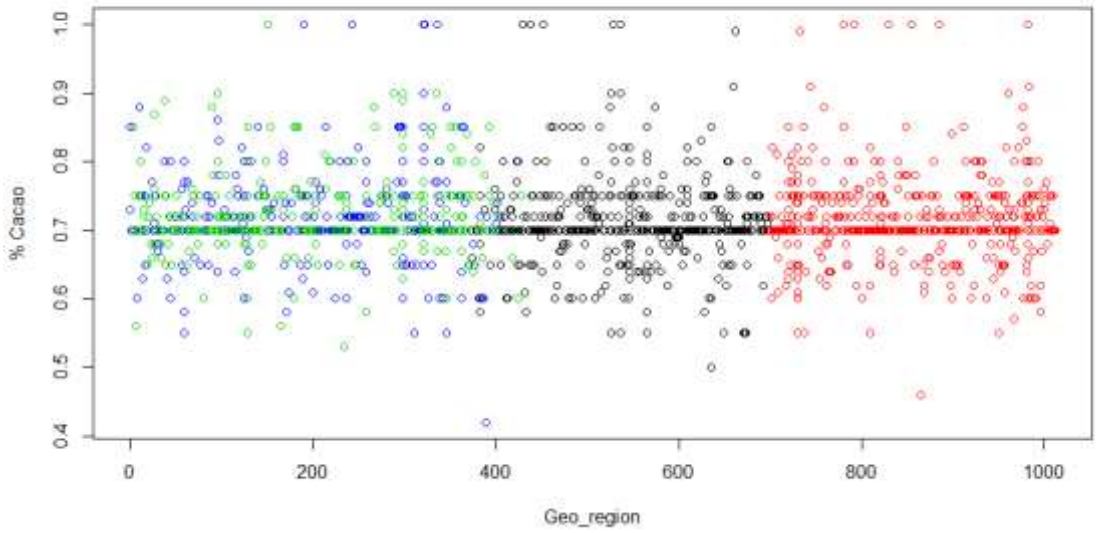
Los valores extremos los encontramos en la calificación 5 y 1 los dejé para su posterior análisis.

3.2.4.5 Valor extremo tipo_frijol-rating



Los valores extremos los vemos en el rating 1 y 5.

3.2.4.6 Valor extremo porcent_cocoa-Geo_region



Los valores extremos que visualizamos mayormente es cuando el cacao el 100%.

Como ya hemos culminado con la limpieza de datos procedemos a guardar un archivo con todos los cambios.

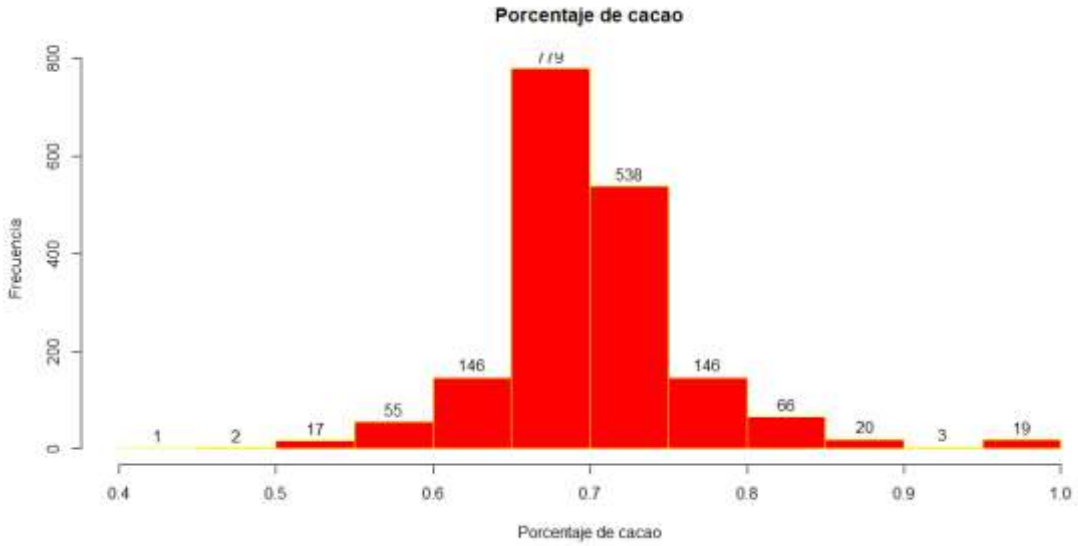
```
> write.csv(cacao[1:6],file="C:/Users/jbsamaniego/Documents/2018/MaestrioDatos/Prac2/flavors_of_cacaoClean.csv", row.names = FALSE)
> #Reviso los cambios realizados
> head(cacao, 4)
  Empresa Geo_region porcent_cocoa locaci_empresa rating tipo_fijol
1       1         14          0.63             10   3.75          4
2       1        476          0.70             10   2.75          4
3       1         64          0.70             10   3.00          4
4       1         15          0.70             10   3.50          4
```

head(x, ...)

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 el análisis de datos voy a utilizar histogramas que me permiten ver a nivel general el comportamiento de los datos.

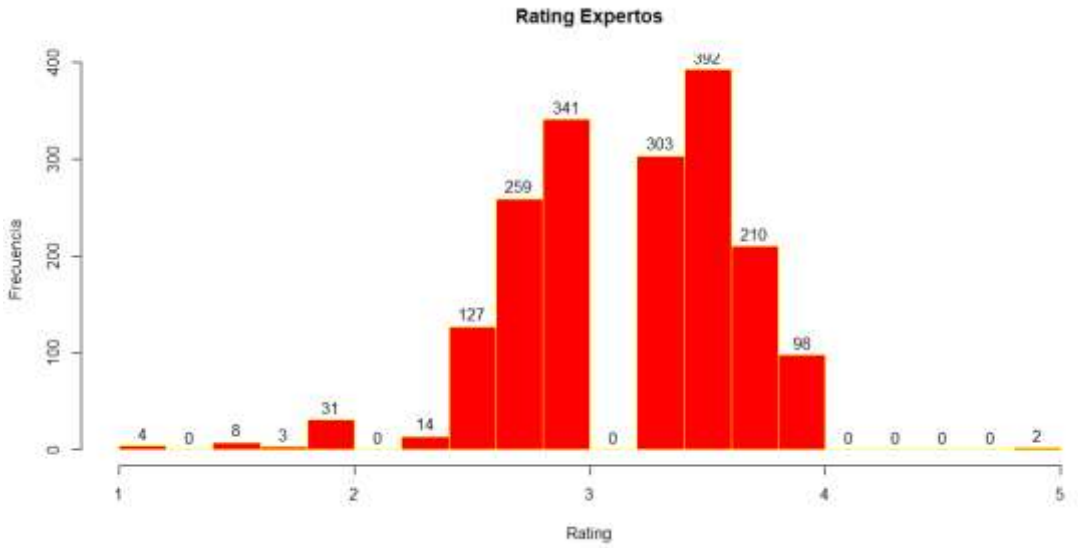


En el histograma se puede diferenciar que los conjuntos de datos de mayor interés son los de mayor frecuencia, que en este caso son: 0.65 0.70 Y 0.75

Tomando en cuenta el resultado hago agrupaciones de acuerdo a los porcentajes de mayor frecuencia, en este caso se analizarán 1807 registros:

```
> #Agrupo por los porcentajes 0.65, 0.70 y 0.75 debido a su cantidad de datos
> cacao.porcen1 <- cacao[cacao$porcent_cocoa == 0.70,]
> cacao.porcen2 <- cacao[cacao$porcent_cocoa == 0.72,]
> cacao.porcen3 <- cacao[cacao$porcent_cocoa == 0.75,]
> #total del conjunto de datos de acuerdo al porcentaje 972 registros
> nrow(cacao.porcen1)+nrow(cacao.porcen2)+nrow(cacao.porcen3)
[1] 1087
```

Ahora graficaremos la calificación o rating de expertos



Se observa que el mayor consumo es de cacao entre satisfactorio y Premium es decir entre 2.50,2.75, 3, 3.25, 3.50, 3.65. Se agrupan en base a estos valores y se observa q se analizarán 1422 registros.

```
> cacao.ratin1 <- cacao[cacao$rating == 2.50,]
> cacao.ratin2 <- cacao[cacao$rating == 2.75,]
> cacao.ratin3 <- cacao[cacao$rating == 3,]
> cacao.ratin4 <- cacao[cacao$rating == 3.25,]
> cacao.ratin5 <- cacao[cacao$rating == 3.50,]
> cacao.ratin6 <- cacao[cacao$rating == 3.65,]
> #total de registros de los conjuntos de datos en cuanto a rating de expertos es de 1
422
> nrow(cacao.ratin1)+nrow(cacao.ratin2)+nrow(cacao.ratin3)+nrow(cacao.ratin4)+nrow(cac
ao.ratin5)+nrow(cacao.ratin6)
[1] 1422
```

4.2. Comprobación de la normalidad y homogeneidad de la varianza.

4.2.1 Normalidad

Para comprobar la normalidad de la varianza existe varios métodos uno de ellos es el de Anderson Darling como lo menciona [4] Esta prueba se aplica para evaluar el ajuste a cualquier distribución de probabilidades, está basada en la comparación de distribución de probabilidades acumulada empírica con la distribución de probabilidades acumulada teórica. Por ellos planteamos la siguiente hipótesis.

- ✓ H0: Los datos siguen una distribución especificada.
- ✓ H1: Los datos no siguen una distribución especificada.

Para aceptar o rechazar esta hipótesis debemos tomar en cuenta la siguiente tabla:

α	0.1	0.05	0.025	0.01
A_T^2	0.631	0.752	0.873	1.035

Es necesario llamar a la librería nortest y dar el valor a alfa en nuestro caso de 0.05:

```
> library(nortest)
> alpha = 0.05 #asignación del valor alfa
> col.name= colnames(cacao)
> print (col.name)
[1] "Empresa"        "Geo_region"     "porcent_cocoa"  "locaci_empresa"
[5] "rating"         "tipo_fijol"
> for (i in 1:ncol(cacao)) {
+   if (i == 1) cat("Los siguientes atributos no siguen una distribución normal:\n")
+   if (is.integer(cacao[,i]) | is.numeric(cacao[,i])) {
+     p_val = ad.test(cacao[,i])$p.value
+     print(p_val)
+     if (p_val < alpha) {
+       cat(col.name[i])
+       # Format output
+       if (i < ncol(cacao) - 1) cat(", ")
+       if (i %% 3 == 0) cat("\n")
+     }
+   }
+ }
Los siguientes atributos no siguen una distribución normal:
[1] 3.7e-24
Empresa, [1] 3.7e-24
Geo_region, [1] 3.7e-24
porcent_cocoa,
[1] 3.7e-24
locaci_empresa, [1] 3.7e-24
rating[1] 3.7e-24
tipo_fijol
```

Vemos que el resultado indica que ningún atributo tiene una distribución normal.

4.2.2 Homogeneiedad

Al igual que para el cálculo de normalidad existen algunos métodos en mi caso voy a implementar el test de Fligner-Killeen, [5] el mismo que es un test no paramétrico que compara varianzas, basándose en la mediana. Tomando en cuenta que es una alternativa cuando no se cumple la condición de normalidad en las muestras.

Para implementar esta muestra vamos a tomar dos variables: Rating y porcent_cocoa, tomando en cuenta que son los campos más importantes.

La hipótesis planteada es que las varianzas de ambas muestras son homogéneas:

```
> #Test de Fligner-KilleenSe trata de un test no paramétrico que compara las variar
 basándose en la mediana
> fligner.test(cacao$rating~ cacao$porcent_cocoa)

        Fligner-Killeen test of homogeneity of variances

data:  cacao$rating by cacao$porcent_cocoa
Fligner-Killeen:med chi-squared = 53.069, df = 41, p-value = 0.09805
```

Luego de aplicar el test se puede decir que se confirma que la hipótesis es correcta, tomando en cuenta que es resultado es mayor a 0.05.

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.

Para este punto vamos a leer el archivo final que se obtuvo luego de la limpieza de datos, se verifica que todas las variables sean numéricas.

4.3.1 Covarianza y correlación

```
> ##Aplicación de pruebas estadísticas para comparar los grupos de datos.
> #### Covarianza y Correlación:
> cacao <- read.csv(file="C:/Users/jbsamaniego/Documents/2018/Maestria/TipologiaCicloD
atos/Prac2/flavors_of_cacaoClean.csv",header=TRUE )
> print (cacao)
     Empresa Geo_region porcent_cocoa locaci_empresa rating tipo_fijol
1          1         14          0.63             10   3.75          4
2          1        476          0.70             10   2.75          4
3          1         64          0.70             10   3.00          4
4          1         15          0.70             10   3.50          4
5          1        789          0.70             10   3.50          4
6          1        166          0.70             10   2.75          1
7          1        275          0.70             10   3.50          1
8          1        899          0.70             10   3.50          1
9          1        781          0.70             10   3.75          1
```

Planteo el algoritmo para evaluar la covarianza y correlación

```
> cat("1: Empresa                \r\n");
1: Empresa
> cat("2: Geo.region             \r\n");
2: Geo.region
> cat("3: Porcent_cocoa          \r\n");
3: Porcent_cocoa
> cat("4: Locaci_empresa         \r\n");
4: Locaci_empresa
> cat("5: Rating                 \r\n");
5: Rating
> cat("6: tipo_frijol            \r\n");
```

```
6: tipo_frijol
> cat("                          \r\n");

> 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:",colnam
es(cacao)[j],"la covarianza es: ",covarianza,"\n\r");
+       cat("Entre el campo: ",colnames(cacao)[i],"y el campo:",colnam
es(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:  -17
1.8169
Entre el campo:  Empresa y el campo: Geo_region la correlación es:  -0
.004843315

--------------
Entre el campo:  Empresa y el campo: porcent_cocoa la covarianza es:
0.2674853
Entre el campo:  Empresa y el campo: porcent_cocoa la correlación es:
0.03462111

--------------
Entre el campo:  Empresa y el campo: locaci_empresa la covarianza es:
161.3009
Entre el campo:  Empresa y el campo: locaci_empresa la correlación es:
0.07408409

--------------
Entre el campo:  Empresa y el campo: rating la covarianza es:  -1.0705
54
Entre el campo:  Empresa y el campo: rating la correlación es:  -0.018
17596

--------------
Entre el campo:  Empresa y el campo: tipo_fijol la covarianza es:  1.8
19945
Entre el campo:  Empresa y el campo: tipo_fijol la correlación es:  0.
01248544

--------------
```

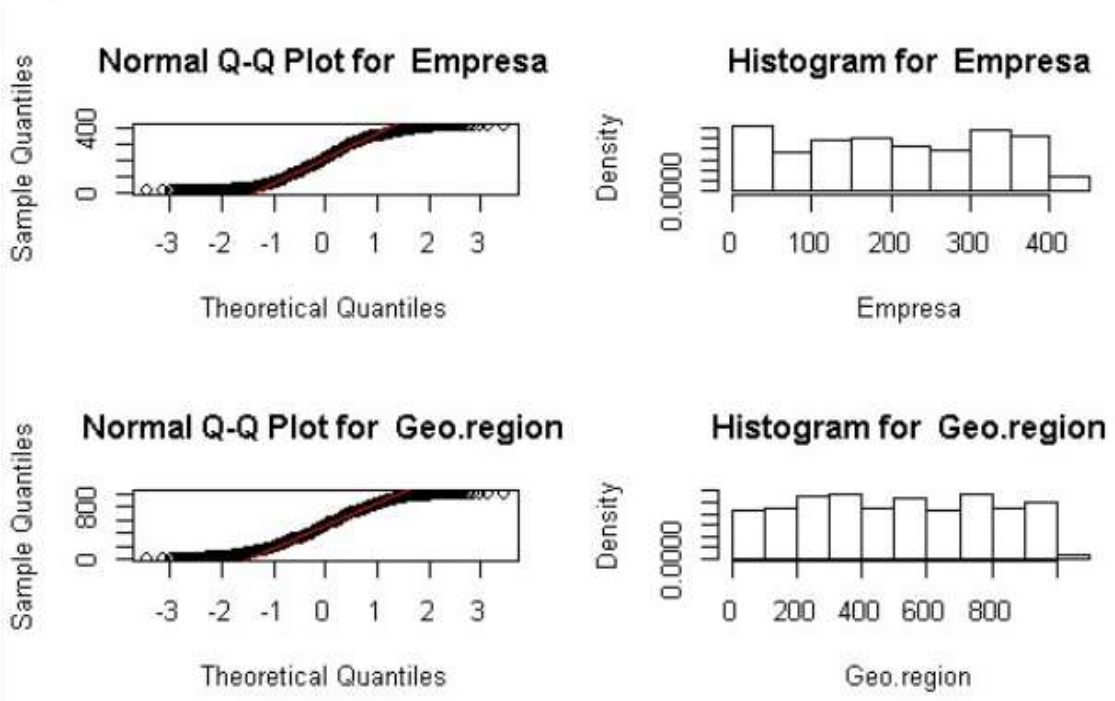
```
Entre el campo:  Geo_region y el campo: porcent_cocoa la covarianza es
:  -0.7213153
Entre el campo:  Geo_region y el campo: porcent_cocoa la correlación e
s:  -0.0405844

--------------
Entre el campo:  Geo_region y el campo: locaci_empresa la covarianza e
s:  -66.56981
Entre el campo:  Geo_region y el campo: locaci_empresa la correlación
es:  -0.01329102

--------------
Entre el campo:  Geo_region y el campo: rating la covarianza es:  -2.7
44128
Entre el campo:  Geo_region y el campo: rating la correlación es:  -0.
02025284

--------------
Entre el campo:  Geo_region y el campo: tipo_fijol la covarianza es:
-27.39309
Entre el campo:  Geo_region y el campo: tipo_fijol la correlación es:
-0.081692

--------------
Entre el campo:  porcent_cocoa y el campo: locaci_empresa la covarianz
a es:  0.04368185
Entre el campo:  porcent_cocoa y el campo: locaci_empresa la correlaci
ón es:  0.0400448

--------------
Entre el campo:  porcent_cocoa y el campo: rating la covarianza es:  -
0.004262693
Entre el campo:  porcent_cocoa y el campo: rating la correlación es:
-0.1444541

--------------
Entre el campo:  porcent_cocoa y el campo: tipo_fijol la covarianza es
:  -1.124172e-05
Entre el campo:  porcent_cocoa y el campo: tipo_fijol la correlación e
s:  -0.0001539343

--------------
Entre el campo:  locaci_empresa y el campo: rating la covarianza es:
-0.1676383
Entre el campo:  locaci_empresa y el campo: rating la correlación es:
-0.02015887

--------------
Entre el campo:  locaci_empresa y el campo: tipo_fijol la covarianza e
s:  1.360438
Entre el campo:  locaci_empresa y el campo: tipo_fijol la correlación
es:  0.06610411

--------------
Entre el campo:  rating y el campo: tipo_fijol la covarianza es:  0.05
216621
Entre el campo:  rating y el campo: tipo_fijol la correlación es:  0.0
9369998

--------------
```

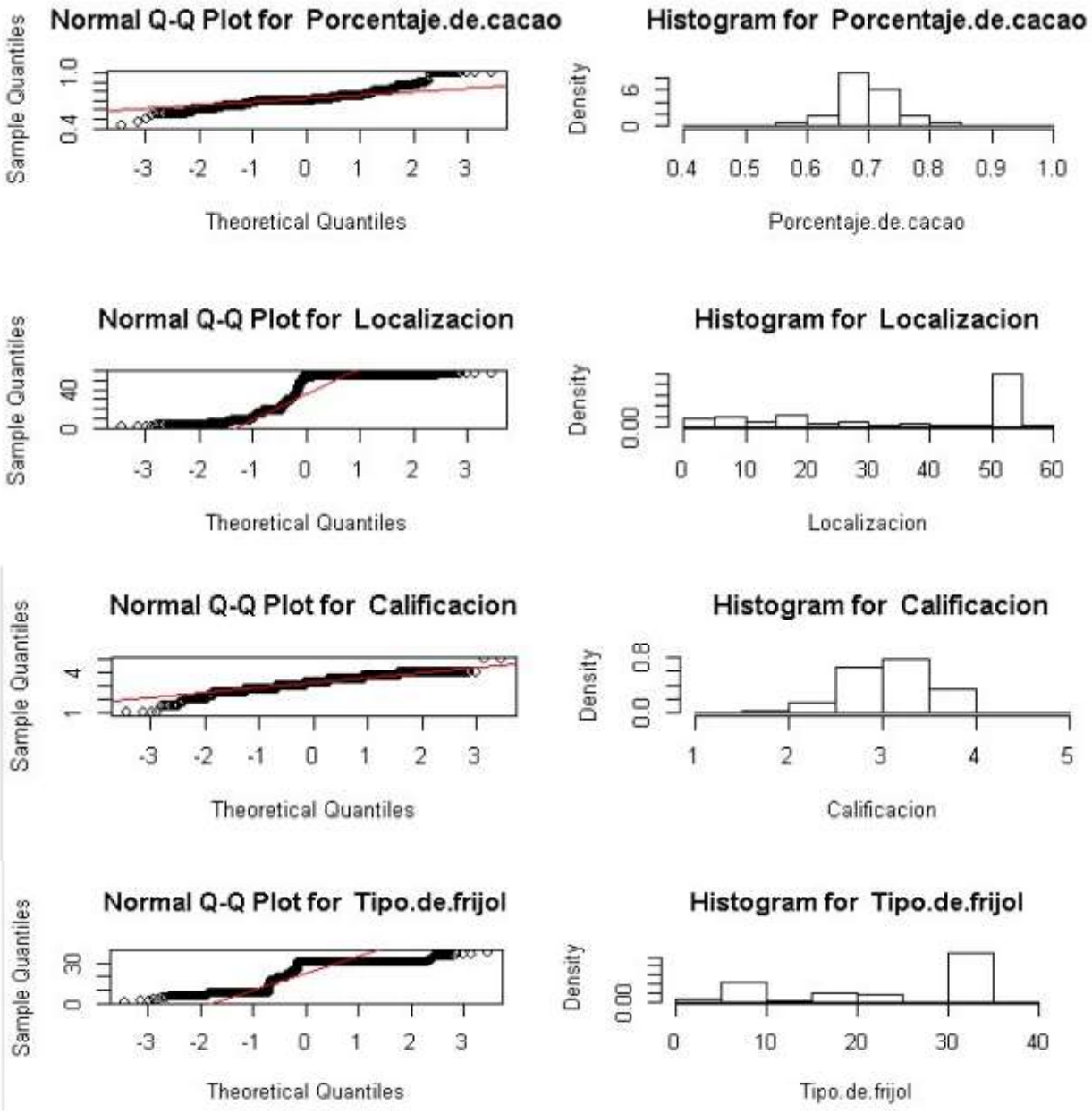
El total de combinaciones fueron 15. De los cuales los atributos que tiene mayor correlación son Porcentaje de cacao y rating, pues su correlación es de -0.1446795, que es el más próximo a -1.

4.3.2 Contraste gráfico

Para realizar el contraste gráfico uso gráficas de quantile-quantile plot y el histograma.

```
> ##Prueba de contraste gráfico##
> cacao <- read.csv(file="C:/Users/jbsamaniego/Documents/2018/Maestria/TipologiaCiclol
atos/Prac2/flavors_of_cacaoClean.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)
+   }
+ }
```





Se puede concluir que los atributos empresa y geo_ region tiene una distribución similar.

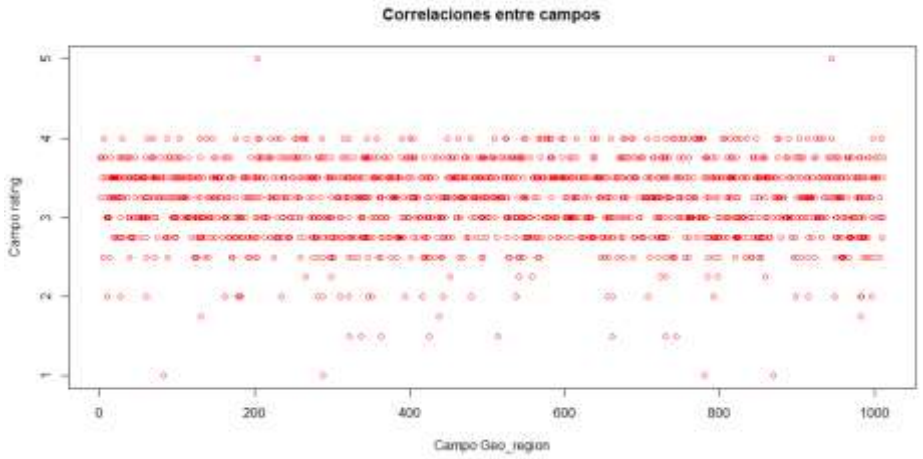
5. Representación de los resultados a partir de tablas y gráficas.

```
> ####Representación gráfica
> cacao <- read.csv(file="C:/Users/jbsamaniego/Documents/2018/Maestria/TipologiaCicloD
atos/Prac2/flavors_of_cacaoClean.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
 correlacion 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
+ }
Entre el campo:  Empresa y el campo: Geo_region la correlación es:  -0.004843315

--------------
Entre el campo:  Empresa y el campo: porcent_cocoa la correlación es:  0.03462111
```

Los resultados se basan en Rating como campo principal contestamos las preguntas a continuación:

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

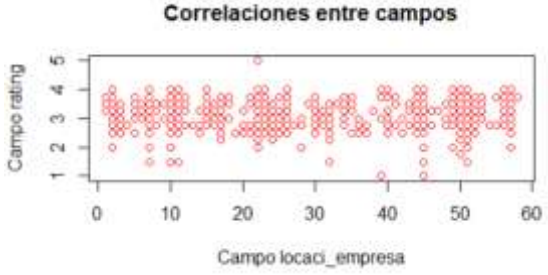
```
Entre el campo:Geo_region y el campo: rating la correlación es:-0.0202
5284
```



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?

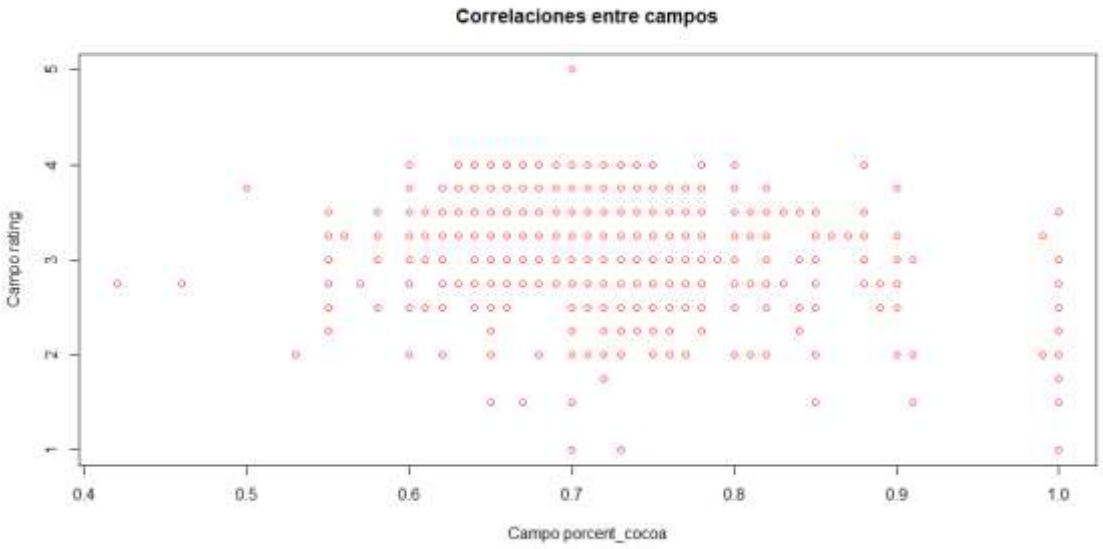
```
Entre el campo:  locaci_empresa y el campo: rating la correlación es:
-0.02015887
```



Los países que producen las mejores barras de chocolate son las localizaciones 22, 29, 18, 8,50 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?

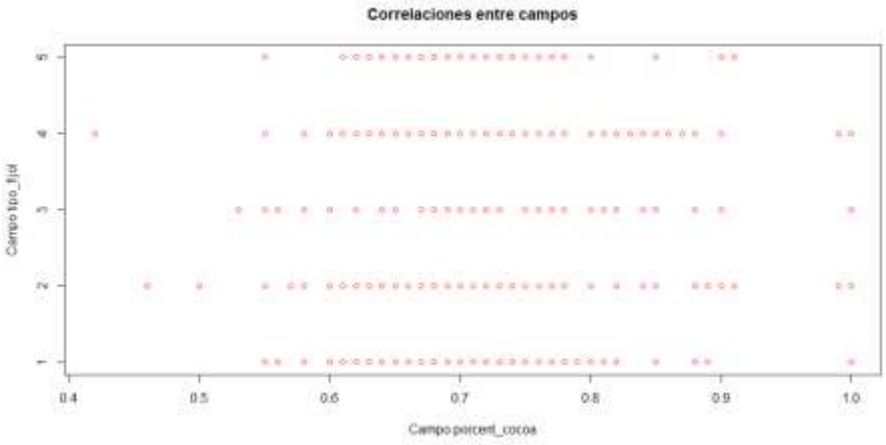
Entre el campo: porcent_cocoa y el campo: rating la correlación es: -0.1444541



Los resultados indican que cuando el porcentaje de cacao está en 70%b se alcanza la mayor calificación que es 5 en la calidad de la barra de chocolate.

¿Existe alguna relación entre el porcentaje de sólidos de cacao y el tipo de frijol?

Entre el campo: porcent_cocoa y el campo: tipo_fijol la correlación es: -0.0001539343



Se puede decir que cuando el porcentaje de cacao esta entre 0.6 a 0.90 el tipo de frijol más usado es el 5 y 4 que son los tipos Trinitario y Blend.

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

- Es mejor discretizar todas las variables para poder trabajar de mejor forma con los datos.
- La técnica del vecino más cercano permitió remplazar los elementos "NA".
- Para determinar que una distribución es mejor el estadístico Anderson-Darling debe ser menor que los demás, nuestro resultado fue que los datos no siguen una distribución normal.
- Se determinó que los campos con mayor correlación entre sí son Porcentaje de cacao y Rating.
- El test de Fligner-Killeen compara varianzas, basándose en la mediana. Siendo una alternativa cuando no se cumple la condición de normalidad en las muestras.
- Se pudo responder a las preguntas planteada correlacionado los diferentes atributos.

7. Código: Hay que adjuntar el código, preferiblemente en R, con el que se ha realizado la limpieza, análisis y representación de los datos. Si lo preferís, también podéis trabajar en Python.

El código en R consta dentro del fichero JSAMANIEGOF_TCDPRAC2.R que se puede descargar en GitHub, desde la carpeta código.

El archivo final se encuentra como flavors_of_cacaoClean.csv.

Todo el proyecto se encuentra en el repositorio:

https://github.com/jenn1991/barraChocolateTipologia

8. Bibliografía

[1] https://blog.barandcocoa.com/about-chocolate/varieties-of-cocoa-beans/

[2] https://thechocolatejournalist.com/single-origin-vs-blend/

[3] http://www.espae.espol.edu.ec/wp-content/uploads/2016/12/industriacacao.pdf

[4]http://www.estadisticacondago.com/images/estadistica_inferencial/pruebas%20de%20normalidad.pdf

[5] https://rpubs.com/Joaquin_AR/218466

[6] https://www.youtube.com/watch?v=w04NxLX6IfM