PAX_EDA_database

Jose Luis Rivas Calduch

8/4/2021

Descripción del trabajo:

El presente *R Markdown* forma parte del trabajo realizado por José Luis Rivas Calduch como alumno de **Master en Ciencia de Datos** de la **UOC** para la asignatura **Visualización de Datos** en concreto para la PEC 2 del curso 20-21 en su segundo semestre.

En concreto se desarrolla la fase de análisis y exploración de los datos (*EDA*) objeto del trabajo anteriormente mencionado.

Cabe destacar que la base de datos (*Database*) *pax_all_agreements_data* ha sido elaborada por la profesora Christine Bell de la Universidad de Edimburgo y que su uso se encuentra regido por los siguientes términos contenidos en el enlace de a continuación: https://www.peaceagreements.org/files/Terms_of_Use.pdf

Carga de los datos:

Como ya hemos indicado en el análisis preliminar de los datos inicialmente para este trabajo nos vamos a centrar en los primeros 26 atributos de la base de datos para su exploración.

```
## tibble[,26] [1,868 x 26] (S3: tbl_df/tbl/data.frame)
##  $ Con         : chr [1:1868] "Afghanistan" "Afghanistan" "Afghanistan" "Afghanistan" ...
##  $ Contp       : chr [1:1868] "Government/territory" "Government" "Government" "Government" ...
##  $ PP          : num [1:1868] 2 2 2 2 2 2 2 2 2 2 ...
##  $ PPName      : chr [1:1868] "Afghanistan: 2000s Post-intervention process" "Afghanistan: 2000s Pos
##  $ Reg         : chr [1:1868] "Europe and Eurasia" "Europe and Eurasia" "Europe and Eurasia" "Europe
##  $ AgtId       : num [1:1868] 2232 1739 1923 864 848 ...
##  $ Ver         : num [1:1868] 3 2 2 1 1 1 1 1 1 1 ...
##  $ Agt         : chr [1:1868] "Resolution of Intra Afghan Peace Conference in Doha, Qatar (Doha Roadm
##  $ Dat         : chr [1:1868] "2019-07-08" "2016-09-22" "2014-09-21" "2012-07-08" ...
##  $ Status      : chr [1:1868] "Multiparty signed/agreed" "Multiparty signed/agreed" "Multiparty signe
##  $ Lgt         : num [1:1868] 2 4 4 14 7 10 9 4 7 4 ...
##  $ N_characters: num [1:1868] 5235 10203 11110 39746 24106 ...
##  $ Agtp        : chr [1:1868] "Intra" "Intra" "Intra" "InterIntra" ...
##  $ Stage       : chr [1:1868] "Pre" "SubPar" "Imp" "Imp" ...
##  $ StageSub    : chr [1:1868] "PreMix" "MultIss" "ExtSub" "ExtSub" ...
##  $ Part        : chr [1:1868] "Stated to be Participants to the Afghan Peace Conference.\n(secondary
##  $ ThrdPart    : chr [1:1868] "Qatar\nGerman Government \nUnited Nations\nUSA\nCountries in the regio
##  $ OthAgr      : chr [1:1868] "Page 2, 8:\n8. We acknowledge and approve the recent resolution of int
##  $ Loc1ISO     : chr [1:1868] "AFG" "AFG" "AFG" "AFG" ...
##  $ Loc2ISO     : chr [1:1868] NA NA NA NA ...
##  $ Loc1GWNO    : num [1:1868] 700 700 700 700 700 700 700 700 700 700 ...
##  $ Loc2GWNO    : num [1:1868] NA NA NA NA NA NA NA NA NA NA ...
##  $ UcdpCon     : chr [1:1868] "333" "333" "333" "333" ...
```

```
##  $ UcdpAgr     : chr [1:1868] "NA" "1488" "NA" "NA" ...
##  $ PamAgr      : chr [1:1868] "NA" NA "NA" "NA" ...
##  $ CowWar      : chr [1:1868] "225" "851" "225" "225" ...
```

se han identificado dentro del grupo seleccionado una serie de atributos categóricos que también se va a considerar su dejarlos fuera del análisis dado que cada registro es diferente al ser campos de texto libre ya que son descripciones. Dichos campos son:

- *Agt*: Nombre del acuerdo.
- *Part*: Firmantes del acuerdo.
- *ThrdPart*: Terceros firmantes del acuerdo.
- *OthAgr*: Otros acuerdos incluidos en el documento.

```
datos <- select(datos,-Agt, -Part, -ThrdPart, -OthAgr)
```

Limpieza de los datos:

Vamos a identificar la existencia de valores nulos (NA´s) o vacíos en el datos set.

Estadísticas de los valores nulos

```
colSums(is.na(datos))
```

```
##          Con        Contp           PP       PPName          Reg        AgtId
##            0            0            0            0            0            0
##          Ver          Dat       Status          Lgt N_characters         Agtp
##            0            0            0            0            0            0
##        Stage     StageSub      Loc1ISO      Loc2ISO     Loc1GWNO     Loc2GWNO
##            0           25           35         1622           11         1622
##      UcdpCon      UcdpAgr       PamAgr       CowWar
##           14         1221         1470          434
```

Se han identificado 9 atributos que contienen valores nulos por lo que habrá que decidir como tratarlos.

Estadísticas de los valores cero o vacios

```
colSums(datos=="")
```

```
##          Con        Contp           PP       PPName          Reg        AgtId
##            0            0            0            0            0            0
##          Ver          Dat       Status          Lgt N_characters         Agtp
##            0            0            0            0            0            0
##        Stage     StageSub      Loc1ISO      Loc2ISO     Loc1GWNO     Loc2GWNO
##            0           NA           NA           NA           NA           NA
##      UcdpCon      UcdpAgr       PamAgr       CowWar
##           NA           NA           NA           NA
```

```
colSums(datos==0)
```

```
##          Con        Contp           PP       PPName          Reg        AgtId
##            0            0            0            0            0            0
```

```
##          Ver          Dat       Status          Lgt N_characters         Agtp
##            0            0            0            0            0            0
##        Stage     StageSub      Loc1ISO      Loc2ISO     Loc1GWNO     Loc2GWNO
##            0           NA           NA           NA           NA           NA
##      UcdpCon      UcdpAgr       PamAgr       CowWar
##           NA           NA           NA           NA
```

No se identifican valores cero o vacíos.

Atributos categóricos:

***Con* - Pais/Entidad:** Se observa que el atributo contiene gran cantidad de valores únicos.

- Número de valores únicos:

```
length(unique(datos$Con))
```

```
## [1] 170
```

Analizar el listado no se aprecia que podamos utilizar la columna para alguna visual dado que muchos de los valores esta compuesto por concatenación de paises o entidades y sería necesario tratar el atributo para su uso. Por lo que por el momento la descartamos.

***Contp* - Tipo conflicto:**

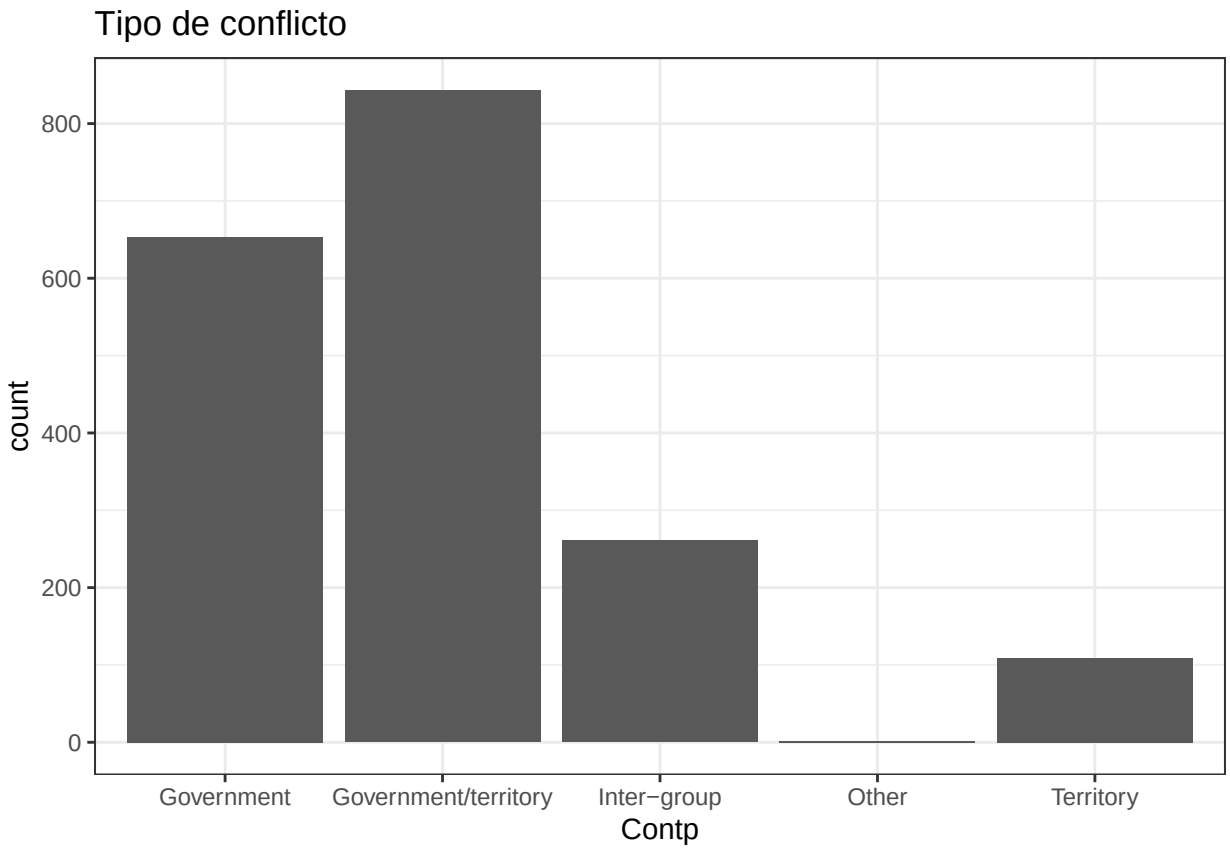
- Número de valores únicos:

```
length(unique(datos$Contp))
```

```
## [1] 5
```

- Diagrama de distribución:

```
ggplot(data = datos, aes(x = Contp, y = ..count..)) +
  geom_bar() +
  labs(title = "Tipo de conflicto") +
  theme_bw() +
  theme(legend.position = "bottom")
```



- Tabla de frecuencias (#):

```
table(datos$Contp)
```

```
##
##            Government Government/territory          Inter-group
##                   653                   843                   261
##                 Other             Territory
##                     2                   109
```

- Tabla de frecuencias (%):

```
prop.table(table(datos$Contp)) %>% round(digits = 2)
```

```
##
##            Government Government/territory          Inter-group
##                  0.35                  0.45                  0.14
##                 Other             Territory
##                  0.00                  0.06
```

Se observa que la mayoría de los acuerdos están relacionados con conflictos gubernamentales (80%).

***PPName* - Nombre proceso:** Se observa que el atributo contiene gran cantidad de valores únicos.

- Número de valores únicos:

```
length(unique(datos$PPName))
```

```
## [1] 157
```

Analizar el listado no se aprecia que podamos utilizar la columna para alguna visual dado que muchos de los valores esta compuesto por concatenación de paises o entidades y sería necesario tratar el atributo para su uso. Por lo que por el momento la descartamos.

***Reg* - Región:**

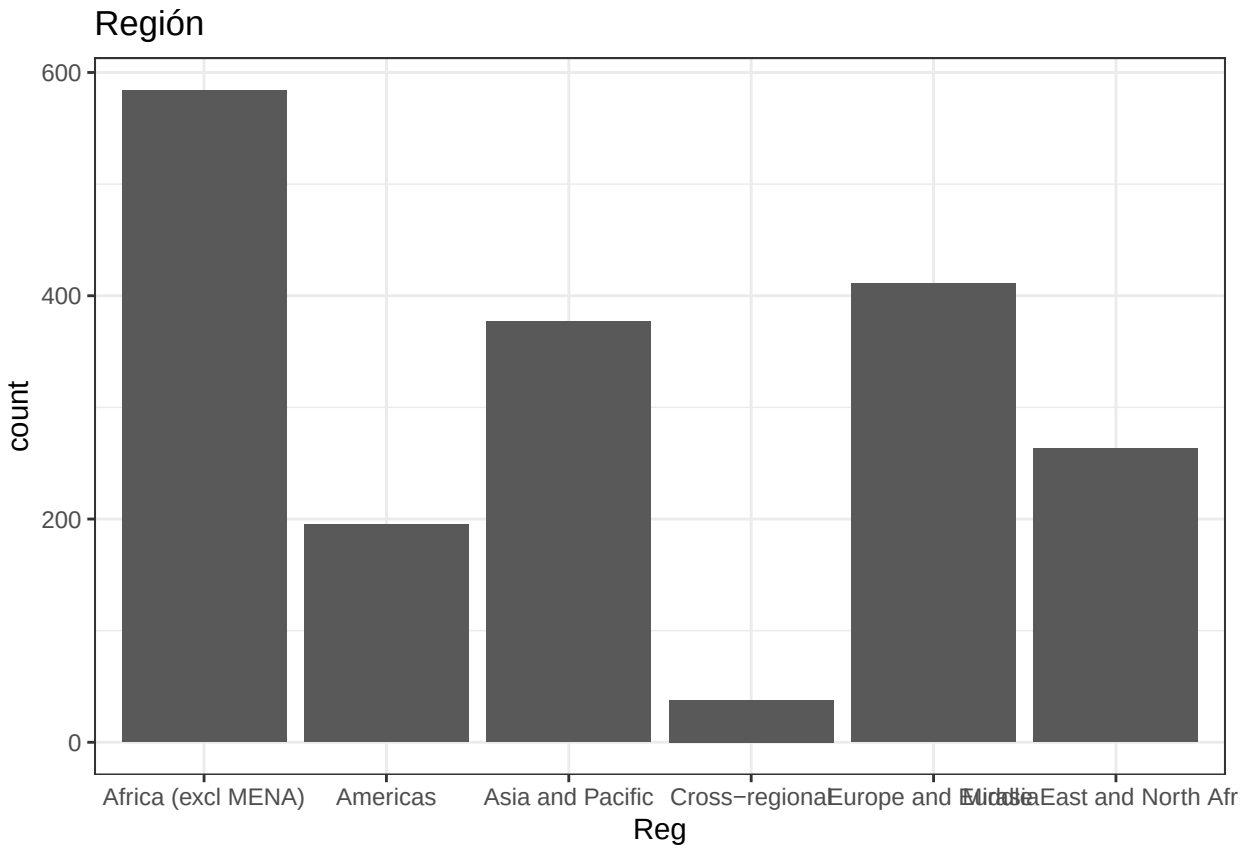
- Número de valores únicos:

```
length(unique(datos$Reg))
```

```
## [1] 6
```

- Diagrama de distribución:

```
ggplot(data = datos, aes(x = Reg, y = ..count..)) +
  geom_bar() +
  labs(title = "Región") +
  theme_bw() +
  theme(legend.position = "bottom")
```



- Tabla de frecuencias (#):

```
table(datos$Reg)
```

```
##
##           Africa (excl MENA)                     Americas
##                          584                          195
##             Asia and Pacific                Cross-regional
##                          377                           38
##           Europe and Eurasia Middle East and North Africa
##                          411                          263
```

- Tabla de frecuencias (%):

```
prop.table(table(datos$Reg)) %>% round(digits = 2)
```

```
##
##           Africa (excl MENA)                     Americas
##                         0.31                         0.10
##             Asia and Pacific                Cross-regional
##                         0.20                         0.02
##           Europe and Eurasia Middle East and North Africa
##                         0.22                         0.14
```

Se observa que la mayoría de los acuerdos se concentran en Africa (45%).

***Status* - Estado del acuerdo:**

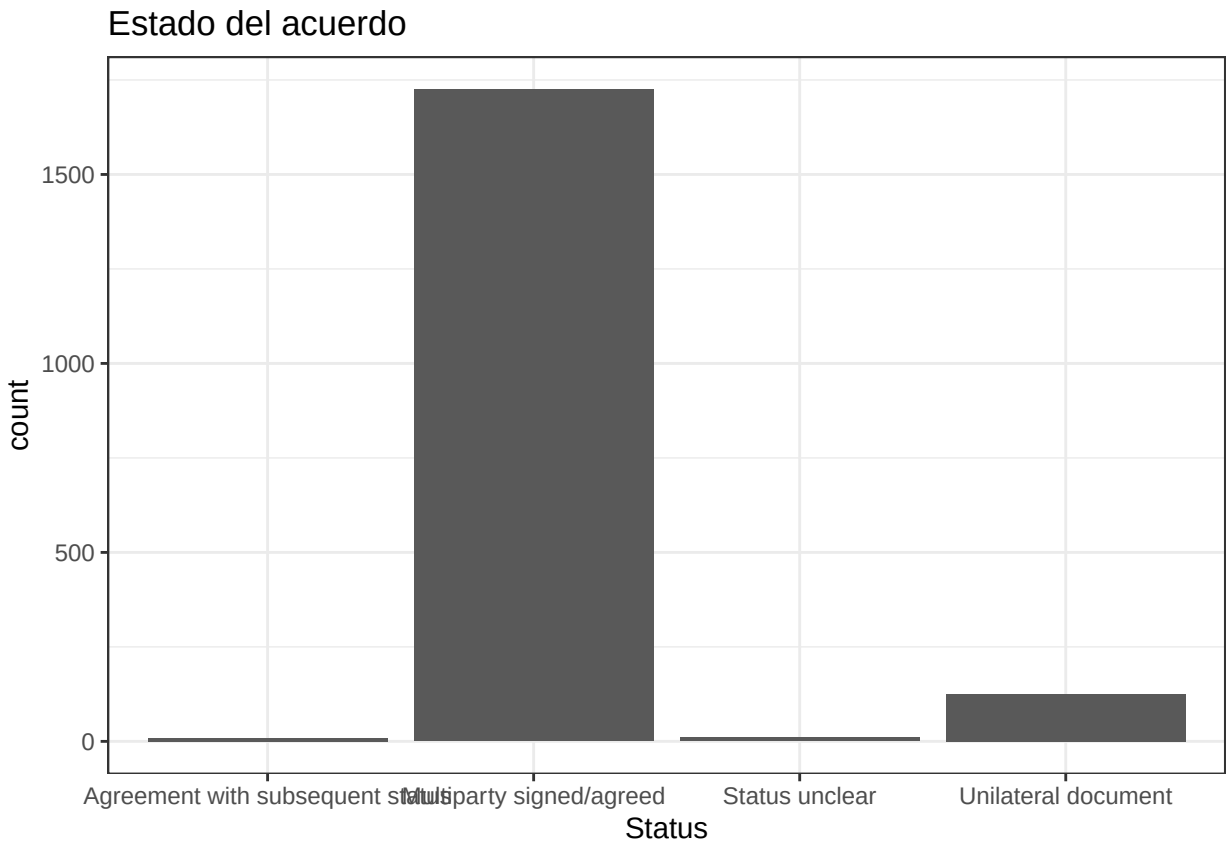
- Número de valores únicos:

```
length(unique(datos$Status))
```

```
## [1] 4
```

- Diagrama de distribución:

```
ggplot(data = datos, aes(x = Status, y = ..count..)) +
  geom_bar() +
  labs(title = "Estado del acuerdo") +
  theme_bw() +
  theme(legend.position = "bottom")
```



- Tabla de frecuencias (#):

```
table(datos$Status)
```

```
##
## Agreement with subsequent status         Multiparty signed/agreed
##                                8                             1725
##                    Status unclear              Unilateral document
##                               10                              125
```

- Tabla de frecuencias (%):

```
prop.table(table(datos$Status)) %>% round(digits = 2)
```

```
##
## Agreement with subsequent status         Multiparty signed/agreed
##                                0.00                             0.92
##                    Status unclear              Unilateral document
##                                0.01                             0.07
```

Se observa que la mayoría de los acuerdos se concentran se encuentran firmado por las diferentes partes (92%).

Agtp - Tipo de acuerdo:

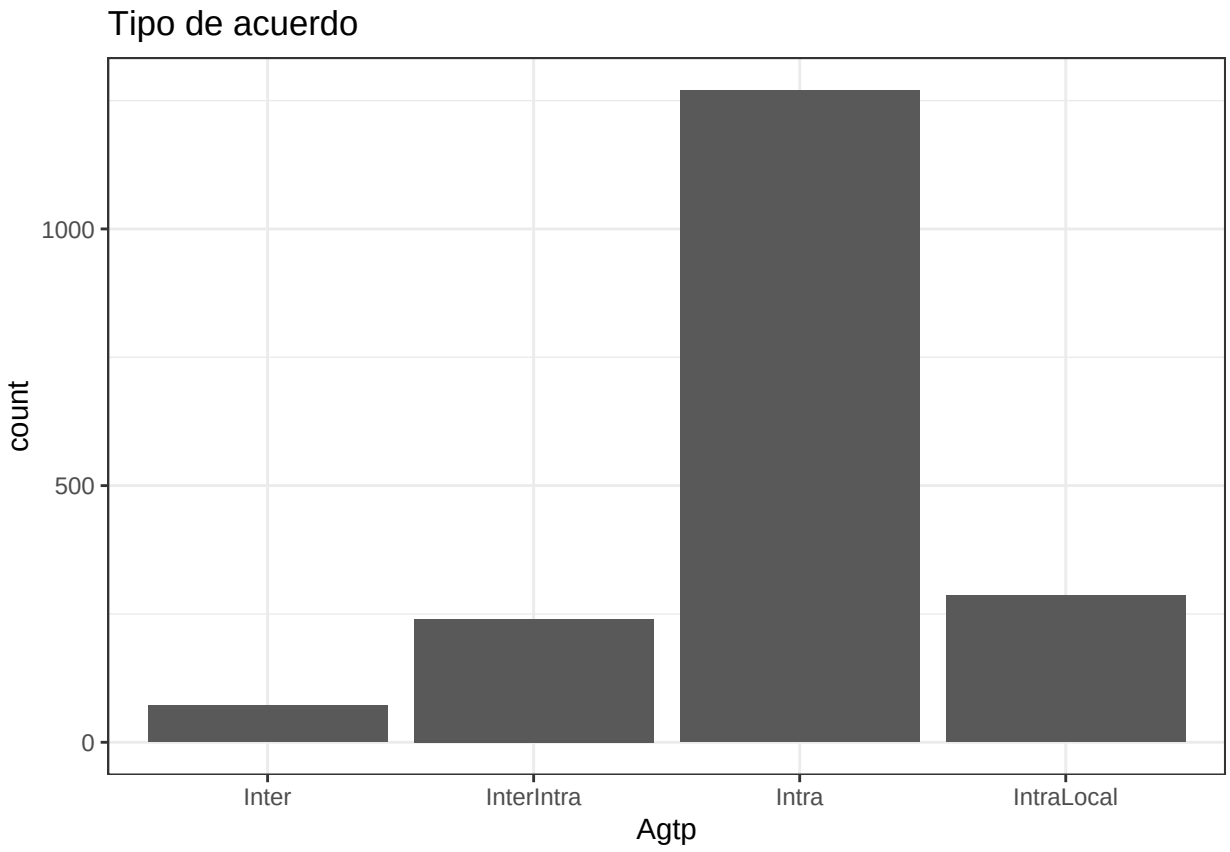
- Número de valores únicos:

```
length(unique(datos$Agtp))
```

```
## [1] 4
```

- Diagrama de distribución:

```
ggplot(data = datos, aes(x = Agtp, y = ..count..)) +
  geom_bar() +
  labs(title = "Tipo de acuerdo") +
  theme_bw() +
  theme(legend.position = "bottom")
```



- Tabla de frecuencias (#):

```
table(datos$Agtp)
```

```
##
##      Inter InterIntra      Intra IntraLocal
##         72        240       1270        286
```

- Tabla de frecuencias (%):

```
prop.table(table(datos$Agtp)) %>% round(digits = 2)
```

```
##
##      Inter InterIntra      Intra IntraLocal
##       0.04       0.13       0.68       0.15
```

Se observa que la mayoría de los acuerdos se concentran en conflictos dentro de los propios estados (83%).

***Stage* - Etapa:**

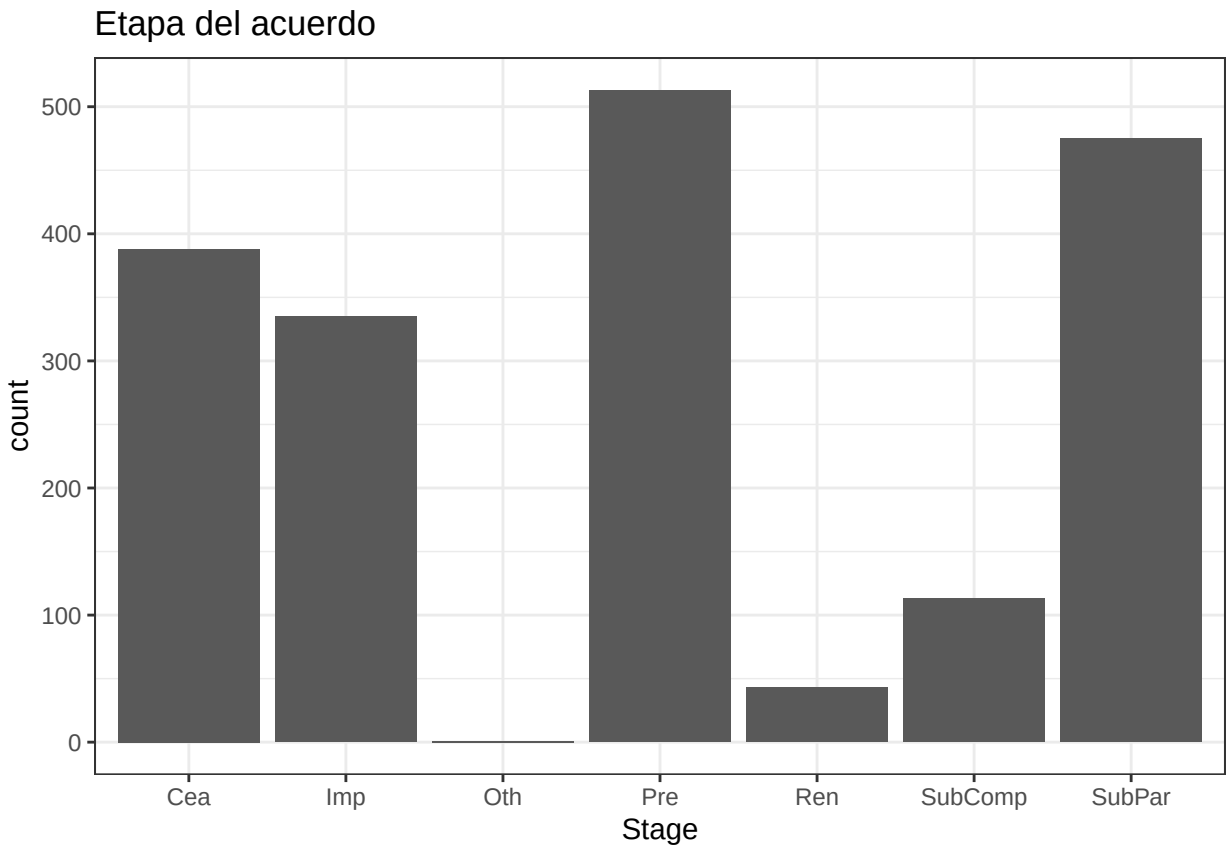
- Número de valores únicos:

```
length(unique(datos$Stage))
```

```
## [1] 7
```

- Diagrama de distribución:

```
ggplot(data = datos, aes(x = Stage, y = ..count..)) +
  geom_bar() +
  labs(title = "Etapa del acuerdo") +
  theme_bw() +
  theme(legend.position = "bottom")
```



- Tabla de frecuencias (#):

```
table(datos$Stage)
```

```
##
##     Cea     Imp     Oth     Pre     Ren SubComp  SubPar
##     388     335       1     513      43     113     475
```

- Tabla de frecuencias (%):

```
prop.table(table(datos$Stage)) %>% round(digits = 2)
```

```
##
##     Cea     Imp     Oth     Pre     Ren SubComp  SubPar
##    0.21    0.18    0.00    0.27    0.02    0.06    0.25
```

***StageSub* - Subetapa del acuerdo:** Dado que se considera que los atributos *Stage* y *StageSub* se encuentran ligados se cree apropiado que se analicen ambos en combinación creando un atributo sintético con su concatenado. También de esta manera tratamos los valores NAs existentes en el atributo.

```
datos <- unite(datos, stage_sintetico, c('Stage','StageSub'), sep = "&", remove=FALSE)
```

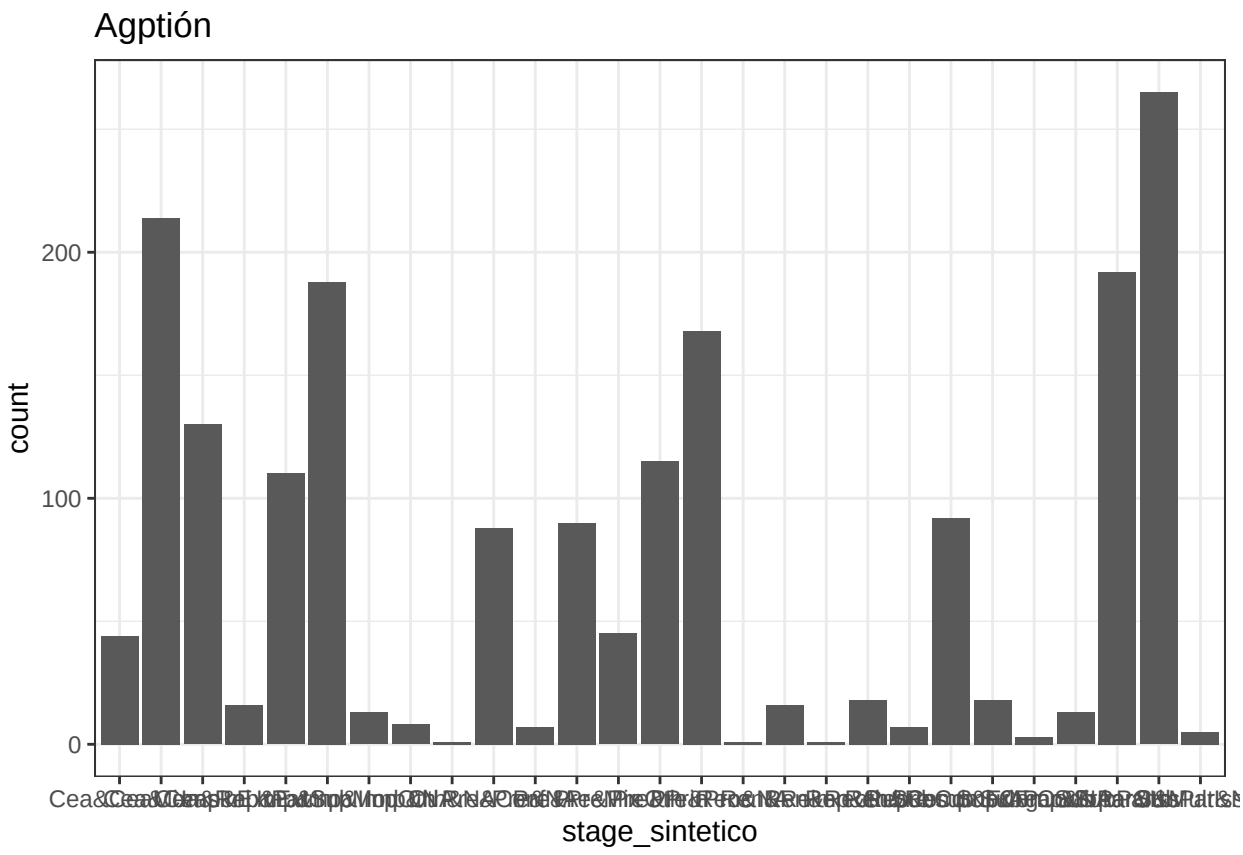
- Número de valores únicos:

```
length(unique(datos$stage_sintetico))
```

```
## [1] 27
```

- Diagrama de distribución:

```
ggplot(data = datos, aes(x = stage_sintetico, y = ..count..)) +
  geom_bar() +
  labs(title = "Agptión") +
  theme_bw() +
  theme(legend.position = "bottom")
```



- Tabla de frecuencias (#):

```
table(datos$stage_sintetico)
```

```
##
##      Cea&CeaMix        Cea&Ceas         Cea&Rel      Imp&ExtPar      Imp&ExtSub
##              44             214             130              16             110
##      Imp&ImpMod      Imp&ImpOth          Imp&NA          Oth&NA        Pre&Conf
##             188              13               8               1              88
##          Pre&NA      Pre&PreMix      Pre&PreOth        Pre&Prin        Pre&Proc
##               7              90              45             115             168
##          Ren&NA       Ren&Reimp       Ren&Reoth       Ren&Repre       Ren&Resub
##               1              16               1              18               7
##    SubComp&FrAg  SubComp&FrCons      SubComp&NA SubPar&FrparOth      SubPar&Iss
##              92              18               3              13             192
##  SubPar&MultIss       SubPar&NA
##             265               5
```

- Tabla de frecuencias (%):

```
prop.table(table(datos$stage_sintetico)) %>% round(digits = 2)
```

```
##
##      Cea&CeaMix        Cea&Ceas         Cea&Rel      Imp&ExtPar      Imp&ExtSub
##            0.02            0.11            0.07            0.01            0.06
##      Imp&ImpMod      Imp&ImpOth          Imp&NA          Oth&NA        Pre&Conf
##            0.10            0.01            0.00            0.00            0.05
##          Pre&NA      Pre&PreMix      Pre&PreOth        Pre&Prin        Pre&Proc
##            0.00            0.05            0.02            0.06            0.09
##          Ren&NA       Ren&Reimp       Ren&Reoth       Ren&Repre       Ren&Resub
##            0.00            0.01            0.00            0.01            0.00
##    SubComp&FrAg  SubComp&FrCons      SubComp&NA SubPar&FrparOth      SubPar&Iss
##            0.05            0.01            0.00            0.01            0.10
##  SubPar&MultIss       SubPar&NA
##            0.14            0.00
```

Se observa que el número de niveles es muy elevado por lo que por el momento se decide que no se va a incorporar al análisis.

***Loc1ISO* - Localización ISO principal:**

- Número de valores únicos:

```
length(unique(datos$Loc1ISO))
```

```
## [1] 83
```

Se observa que existen 83 localizaciones diferentes. Esta información se puede utilizar para crear un mapa en el que podamos visualizar la distribución de los acuerdos. Cabe destacar que se han identificado 35 valores nulos en el atributo que deben ser tratados.

***Loc2ISO* - Localización ISO secundaria:** Este atributo nos indica en el caso de que el acuerdo abarque varios países un segundo país según lo que se indica en la especificaciones de la base de datos la selección es aleatoria. No obstante se decide para simplificar el trabajo que solo se va a trabajar con la localización principal. Por otro lado indicar que el atributo cuenta con alto porcentaje de valores nulos dado que la mayoría de los conflictos solo afectan a un país.

Atributos numéricos:

***PP* - ID proceso:** Es el identificador numérico del proceso de paz y está vinculado con el *PPName*. No lo vamos a utilizar en el análisis.

***AgtId* - ID acuerdo:** Es el identificador numérico del acuerdo y está vinculado con el *Agt*. No lo vamos a considerar para el análisis.

***Ver* - Versión del acuerdo:**

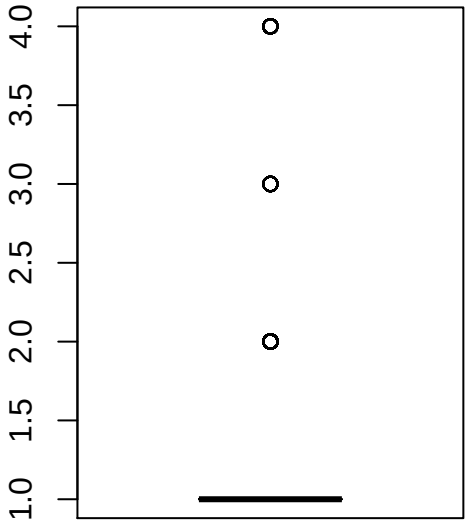
- Medidas de tendencia central:

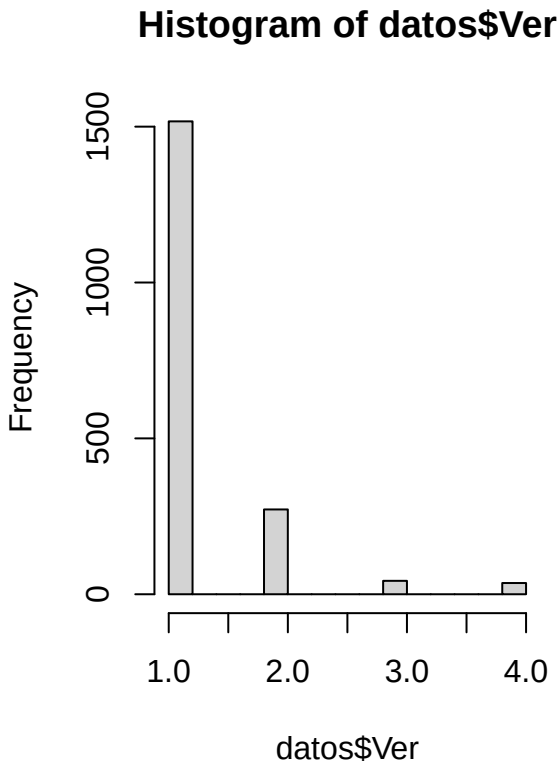
```
summary(datos$Ver)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   1.000   1.000   1.000   1.249   1.000   4.000
```

- Boxplot e histograma:

```
par(mfrow = c(1,2))
boxplot(datos$Ver)
hist(datos$Ver)
```





***Lgt* - Número de páginas:**

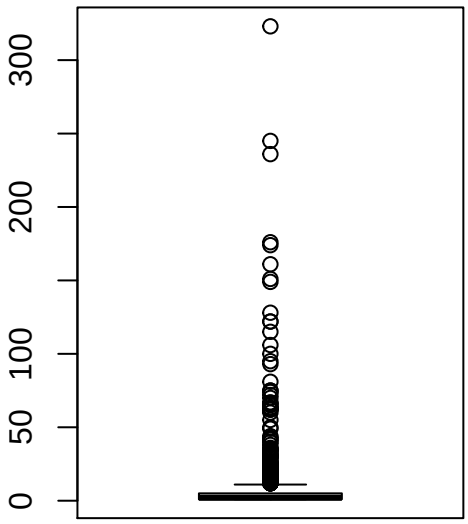
- Medidas de tendencia central:

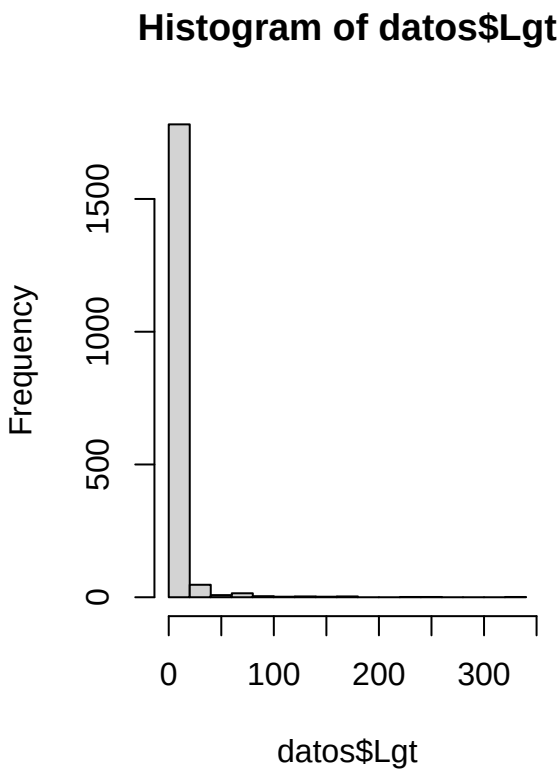
```
summary(datos$Lgt)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   1.000   1.000   2.000   6.275   5.000 323.000
```

- Boxplot e histograma:

```
par(mfrow = c(1,2))
boxplot(datos$Lgt)
hist(datos$Lgt)
```





***N_characters* - Número de caracteres:**

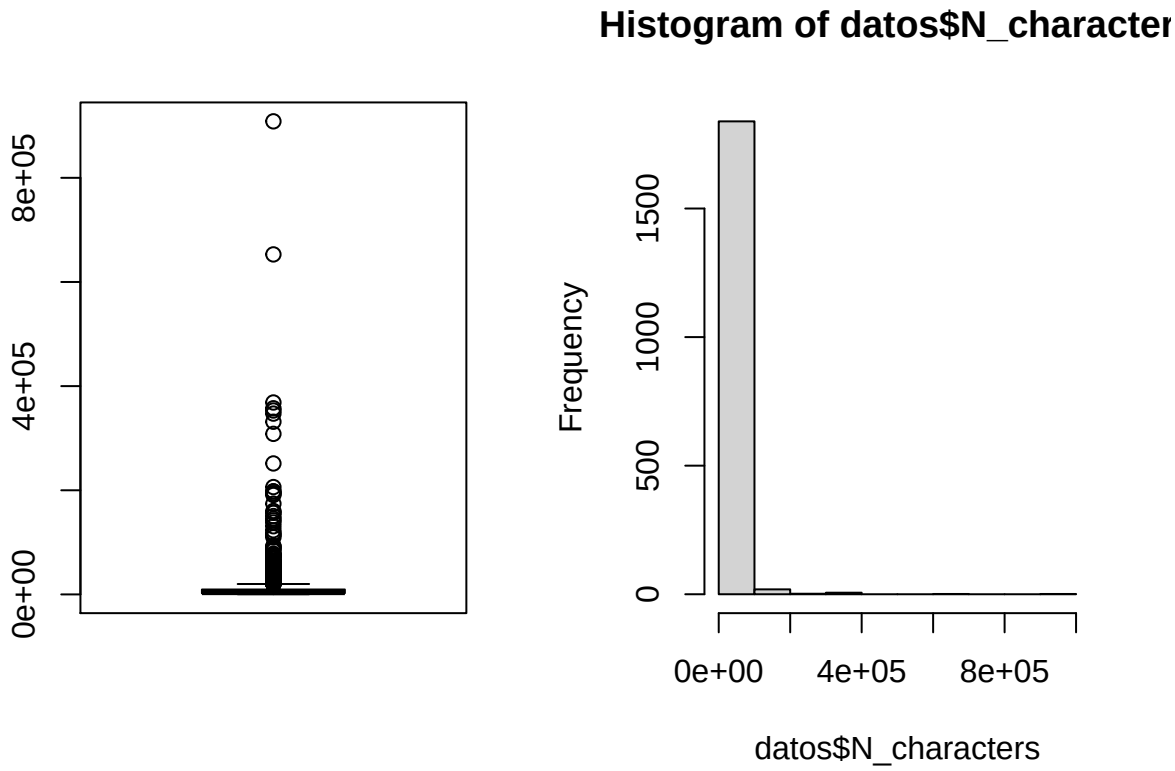
- Medidas de tendencia central:

```
summary(datos$N_characters)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     329    2397    4354   11952    9420  908459
```

- Boxplot e histograma:

```
par(mfrow = c(1,2))
boxplot(datos$N_characters)
hist(datos$N_characters)
```



***Loc1GWNO* - Localización GWC principal.** El atributo es similar al al *Loc1ISO* ya que es un identificador de la localización. Se decide utilizar el ISO.

Loc2GWNO - Localización GWC secundaria. El atributo es similar al al *Loc2ISO* ya que es un identificador de la localización. Se decide utilizar el ISO.

UcdpCon - ID conflicto Uppsala No se va a utilizar el atributo para el análisis.

UcdpAgr - ID acuerdo Uppsala No se va a utilizar el atributo para el análisis.

PamAgr - ID matriz No se va a utilizar el atributo para el análisis. No obstante este atributo tiene muchos registros nulos (NAs) hecho que ya se advierte en el documento de las especificaciones.

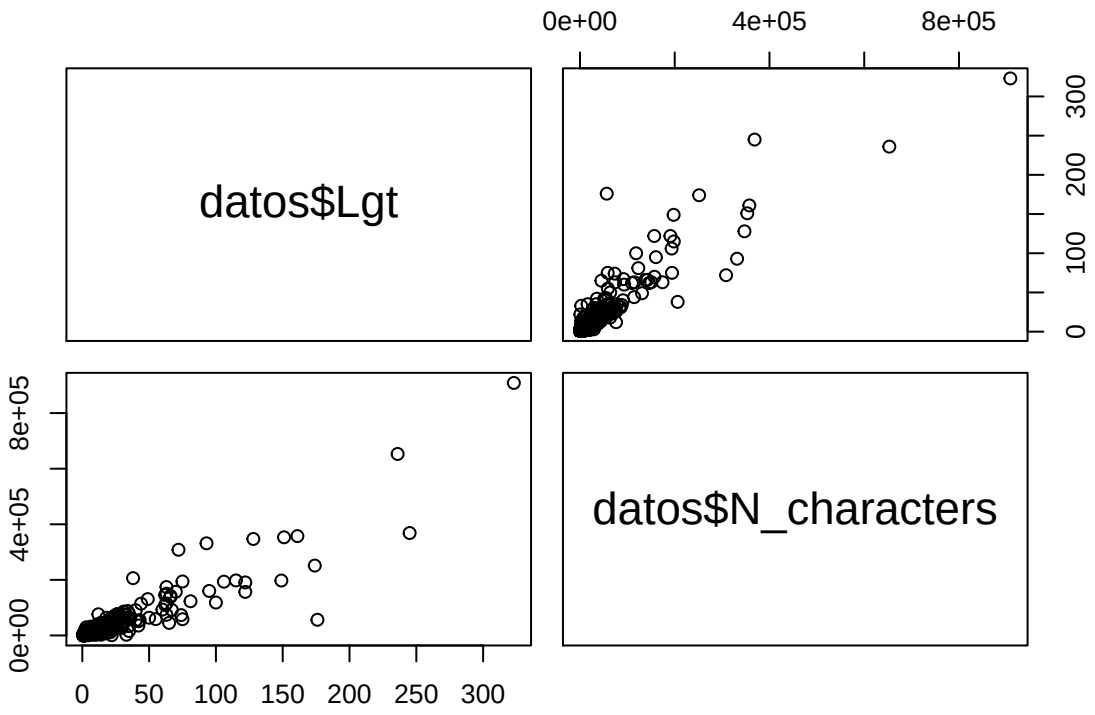
CowWar - ID guerra No se va a utilizar el atributo para el análisis. No obstante este atributo tiene muchos registros nulos (NAs) hecho que ya se advierte en el documento de las especificaciones.

```
cor(datos$Lgt,datos$N_characters)
```

Análisis de correlación entre *Lgt* y *N_characters*:

```
## [1] 0.9190978
```

```
pairs(datos$Lgt ~ datos$N_characters)
```



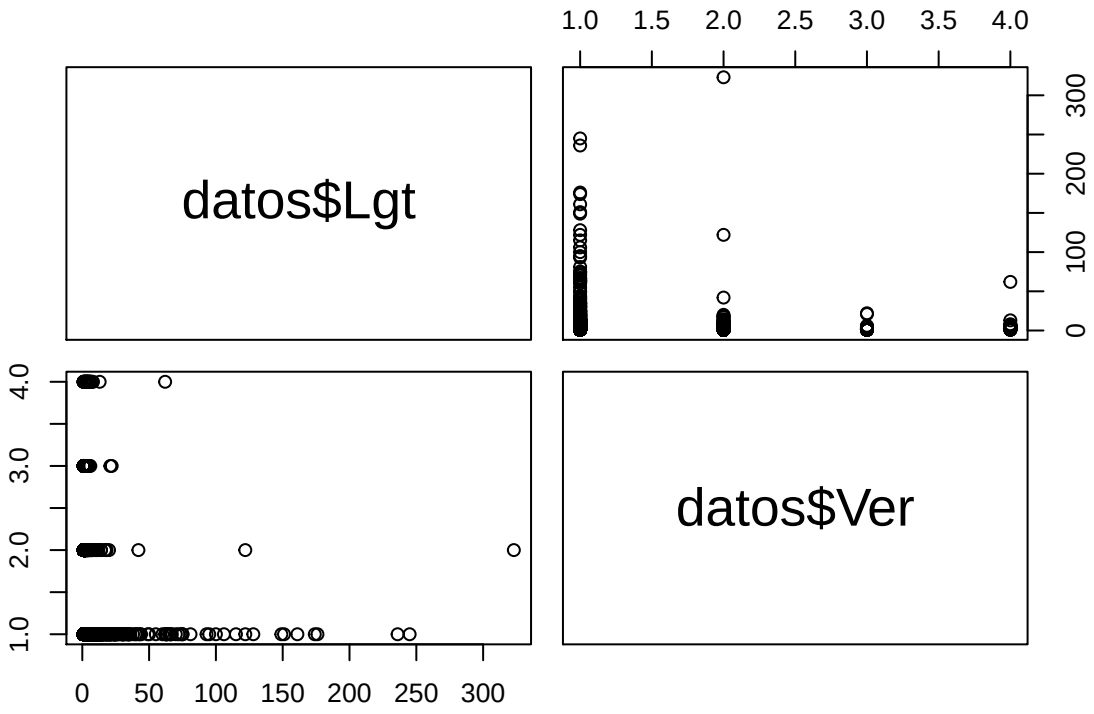
Se observa que existe una fuerte correlación entre el número de páginas y palabras contenidas en el documento.

```
cor(datos$Lgt,datos$Ver)
```

Análisis de correlación entre *Lgt* y *Ver*:

```
## [1] -0.0457094
```

```
pairs(datos$Lgt ~ datos$Ver)
```



No existe correlación.

Atributo fecha:

***Dat* - Fecha de firma del acuerdo:** Se va a descomponer el atributo en año, mes y fecha para realizar su análisis.

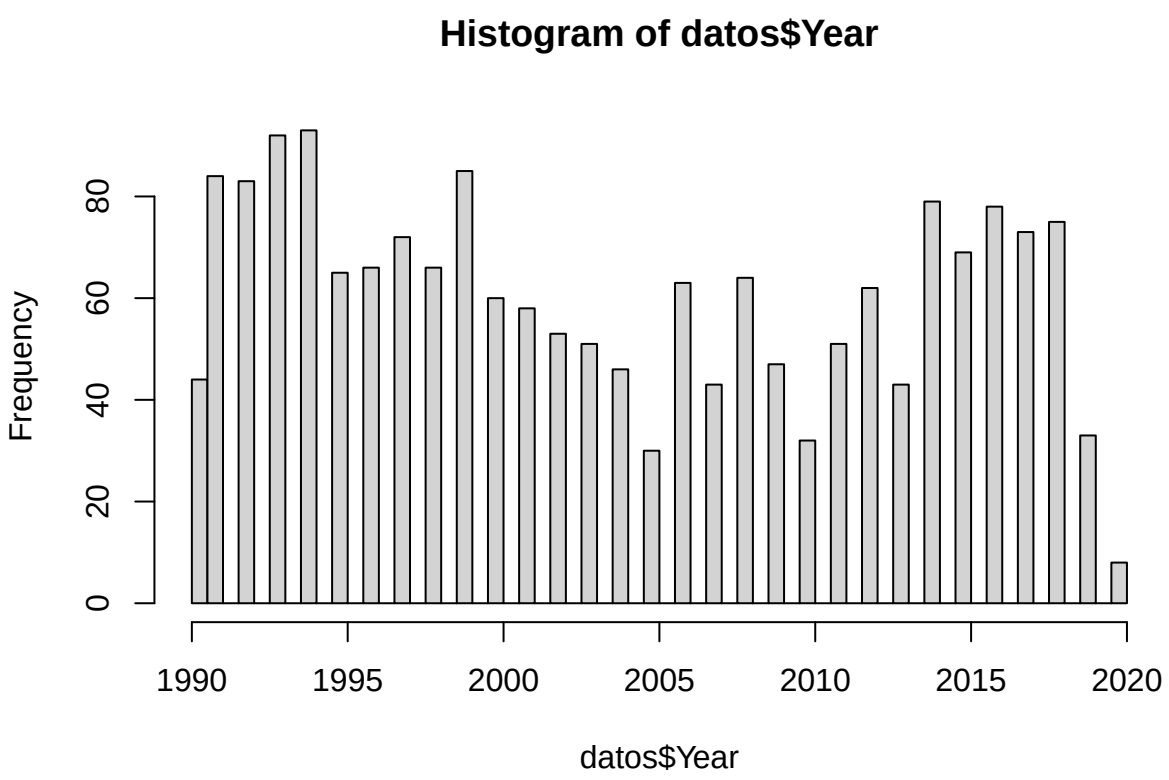
```
datos$Year <- substr(datos$Dat,1,4)
datos$Year <- as.integer(datos$Year)

datos$Month <- substr(datos$Dat,6,7)
datos$Month <- as.integer(datos$Month)

datos$Day <- substr(datos$Dat,9,10)
datos$Day <- as.integer(datos$Day)
```

- Año:

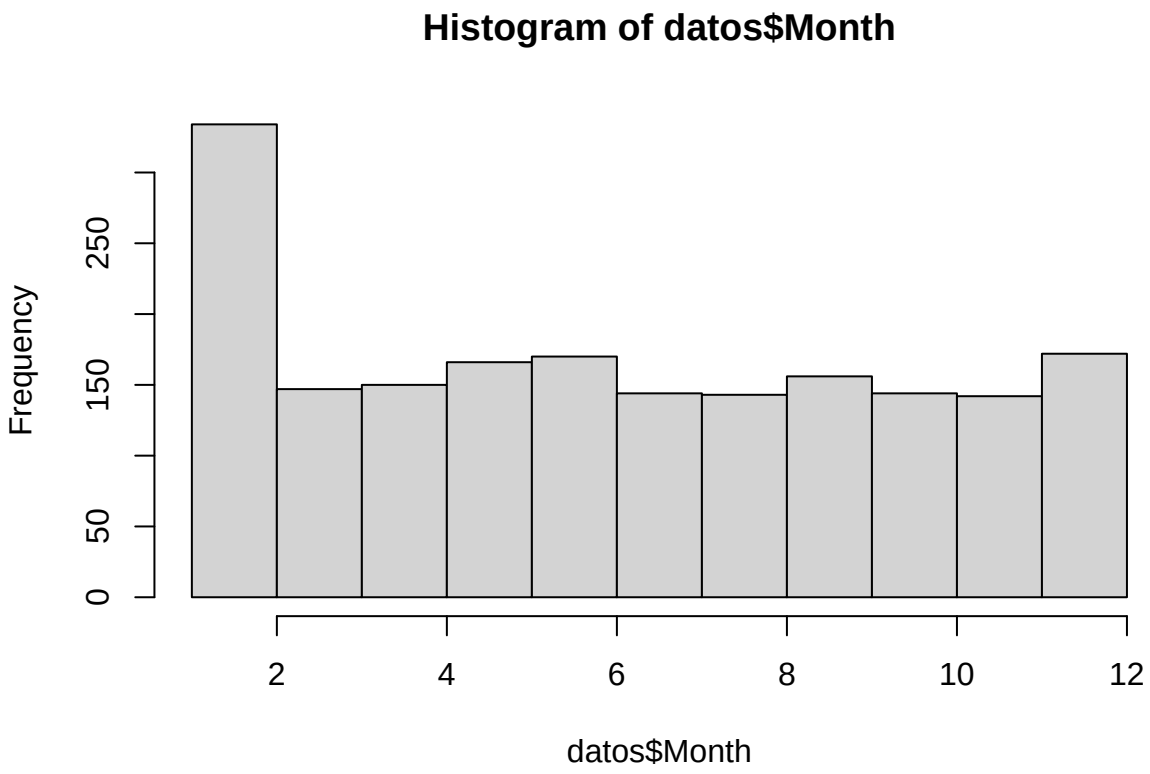
```
hist(datos$Year, breaks = 50)
```



Del la observación del histograma se puede interpretar que el número de acuerdos tiene una tendencia descendente aunque con un pico en la década 10.

- Mes:

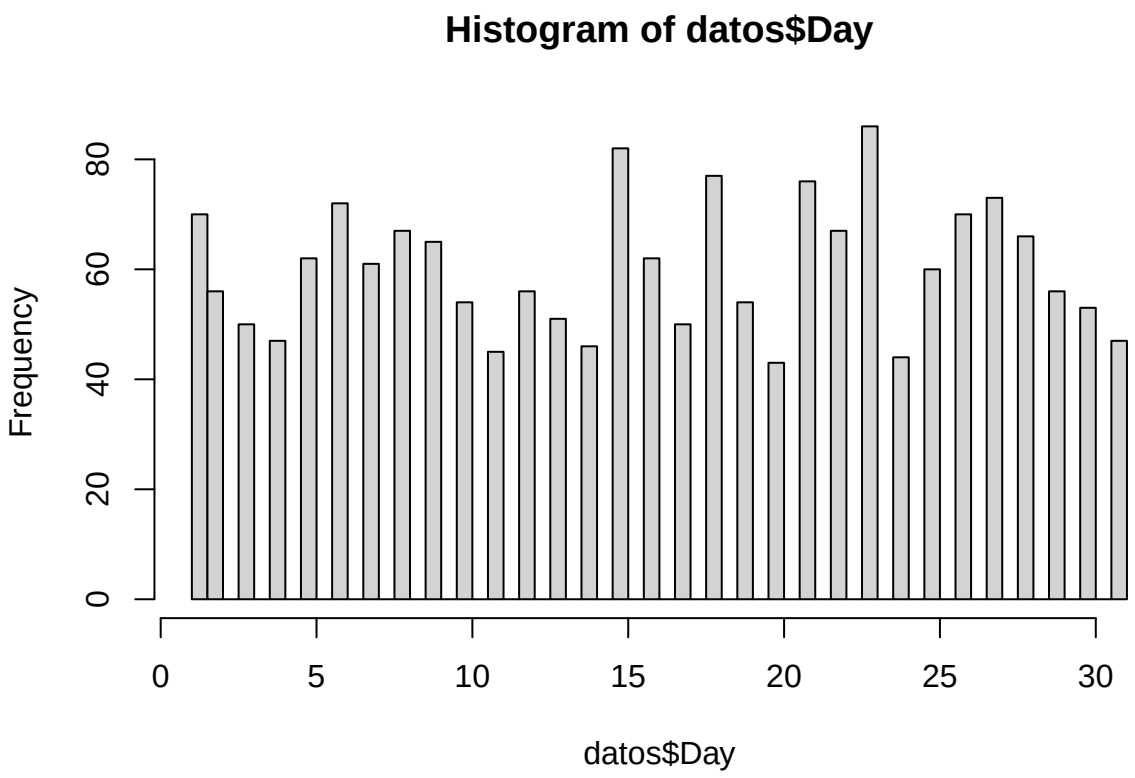
```
hist(datos$Month)
```



Se observa un patrón por el cual la firma de acuerdos más del doble en el mes de enero mientras que en el resto de meses se mantiene estable.

- Día:

```
hist(datos$Day, breaks = 50)
```



Respecto al día de firma de los acuerdos no se observa un patrón claro.

Una vez analizados se decide que se va a emplear el nuevo atributo sintético con el año con el objeto de analizar las tendencias.

Dataset de salida para el análisis visual:

```
#Creamos un dataset con los datos a exportar

salida <- data.frame(datos$Contp,datos$Reg,datos$Year,datos$Status, datos$Agtp, datos$Stage, datos$Loc1
```

Incorporacion de variables externas Dado que solo tenemos la codificación de los países codificada en caracteres vamos a incorporar a nuestro data set para su posterior uso.

Para ello hemos obtenido una tabla con la codificación del siguiente reporsitorio de *GitHub*: https://gist.github.com/brenes/1095110, en el cual se proporciona el mapeo de la codificación con el nombre de los paises.

```
##
## -- Column specification --------------------------------------------------------
## cols(
##   nombre = col_character(),
##   name = col_character(),
##   nom = col_character(),
##   iso2 = col_character(),
```

```
##   iso3 = col_character(),
##   phone_code = col_character()
## )
```

```
## spec_tbl_df[,6] [248 x 6] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
##  $ nombre    : chr [1:248] "Afganistán" "Albania" "Alemania" "Andorra" ...
##  $ name      : chr [1:248] "Afghanistan" "Albania" "Germany" "Andorra" ...
##  $ nom       : chr [1:248] "Afghanistan" "Albanie" "Allemagne" "Andorra" ...
##  $ iso2      : chr [1:248] "AF" "AL" "DE" "AD" ...
##  $ iso3      : chr [1:248] "AFG" "ALB" "DEU" "AND" ...
##  $ phone_code: chr [1:248] "93" "355" "49" "376" ...
##  - attr(*, "spec")=
##   .. cols(
##   ..   nombre = col_character(),
##   ..   name = col_character(),
##   ..   nom = col_character(),
##   ..   iso2 = col_character(),
##   ..   iso3 = col_character(),
##   ..   phone_code = col_character()
##   .. )
```

```
paises_corto <- select(paises, iso3, nombre)

str(paises_corto)
```

```
## tibble[,2] [248 x 2] (S3: tbl_df/tbl/data.frame)
##  $ iso3  : chr [1:248] "AFG" "ALB" "DEU" "AND" ...
##  $ nombre: chr [1:248] "Afganistán" "Albania" "Alemania" "Andorra" ...
```

```
#Renombrados las cabeceras para que sean más legibles
names(salida) = c("Tipo_conflicto","Region","Año","Estado_acuerdo","Tipo_acuerdo","Etapa","Localizacion_
names(paises_corto) = c("Localizacion_ISO", "Nombre_ISO")
```

```
#Fusionamos los data frames
salida_final <- merge(salida, paises_corto, by = "Localizacion_ISO", all.x = TRUE, incomparables = NULL)
```

Exportamos: Creamos un fichero en excel que será la entrada de los datos para Tableau.

```
#Guardamos la salida en excel
write.xlsx(salida_final, "../data/pax_all_agreements_data_proc.xlsx")
```