Sesión 3: Procesamiento y Visualización de Datos con R y Python

Mauricio Alejandro Mazo Lopera
Universidad Nacional de Colombia
Facultad de Ciencias
Escuela de Estadística
Medellín

Unión de bases de datos:

```
dat1 = data.frame(
  Nombre=c('Carlos','Juan','Sara','Pedro','Luis','Ana'),
  Edad=c(21,23,19,25,19,26))

dat2 = data.frame(
  Id=c(1029,2516,8437,9289,7373),
  Nombre=c('Ana','Carlos','Sara','Pedro','Luis'))

dat3 = data.frame(
  Genero=c('F','M','M','M','F','M'),
  Nombres=c('Ana','Carlos','Juan','Pedro','Sara','Luis'))

dat4 = data.frame(
  Nombre=c('Carlos','Juan','Sara','Pedro','Luis','Ana'),
  Programa=c('Estadistica','Matematicas','Ing. Sistemas',
             'Ing. Civil','Economia','Ing. Ambiental'))
```

Unión de bases de datos:

```
require(knitr)
kable(dat1)
```

Nombre	Edad
Carlos	21
Juan	23
Sara	19
Pedro	25
Luis	19
Ana	26

Unión de bases de datos:

```
kable(dat2)
```

Id	Nombre
1029	Ana
2516	Carlos
8437	Sara
9289	Pedro
7373	Luis

Unión de bases de datos:

```
kable(dat3)
```

Genero	Nombres
F	Ana
M	Carlos
M	Juan
M	Pedro
F	Sara
M	Luis

Unión de bases de datos:

```
kable(dat4)
```

Nombre	Programa
Carlos	Estadistica
Juan	Matematicas
Sara	Ing. Sistemas
Pedro	Ing. Civil
Luis	Economia
Ana	Ing. Ambiental

Unión de bases de datos:

```
datos_1<-merge(dat1,dat2,by='Nombre')
kable(datos_1)
```

Nombre	Edad	Id
Ana	26	1029
Carlos	21	2516
Luis	19	7373
Pedro	25	9289
Sara	19	8437

Unión de bases de datos:

```
datos_1<-merge(dat1,dat2,by='Nombre', sort=FALSE)
kable(datos_1)
```

Nombre	Edad	Id
Carlos	21	2516
Sara	19	8437
Pedro	25	9289
Luis	19	7373
Ana	26	1029

Unión de bases de datos:

```
datos_1<-merge(dat1,dat2,by='Nombre', all.x=TRUE)
kable(datos_1)
```

Nombre	Edad	Id
Ana	26	1029
Carlos	21	2516
Juan	23	NA
Luis	19	7373
Pedro	25	9289
Sara	19	8437

Unión de bases de datos:

```
datos_3<-merge(datos_1,dat4,by='Nombre', all=TRUE)
kable(datos_3)
```

Nombre	Edad	Id	Programa
Ana	26	1029	Ing. Ambiental
Carlos	21	2516	Estadistica
Juan	23	NA	Matematicas
Luis	19	7373	Economia
Pedro	25	9289	Ing. Civil
Sara	19	8437	Ing. Sistemas

Unión de bases de datos:

```
dat1 = data.frame(
  Nombre=c('Carlos','Juan','Sara','Pedro','Luis','Ana'),
  Edad=c(21,23,19,25,19,26))

dat2 = data.frame(
  Id=c(1029,2516,8437,9289,7373),
  Nombre=c('Ana','Carlos','Sara','Pedro','Luis'))

dat3 = data.frame(
  Genero=c('F','M','M','M','F','M'),
  Nombre=c('Ana','Carlos','Juan','Pedro','Sara','Luis'))

dat4 = data.frame(
  Nombre=c('Carlos','Juan','Sara','Pedro','Luis','Ana'),
  Programa=c('Estadistica','Matematicas','Ing. Sistemas',
             'Ing. Civil','Economia','Ing. Ambiental'))
```

Unión de bases de datos:

```
datos_4<-Reduce(merge, list(dat1,dat2,dat3,dat4))
kable(datos_4)
```

Nombre	Edad	Id	Genero	Programa
Ana	26	1029	F	Ing. Ambiental
Carlos	21	2516	M	Estadistica
Luis	19	7373	M	Economia
Pedro	25	9289	M	Ing. Civil
Sara	19	8437	F	Ing. Sistemas

Unión de bases de datos:

```
datos_4<-Reduce(function(...) merge (..., all=T),
                list(dat1,dat2,dat3,dat4))
kable(datos_4)
```

Nombre	Edad	Id	Genero	Programa
Ana	26	1029	F	Ing. Ambiental
Carlos	21	2516	M	Estadistica
Juan	23	NA	M	Matematicas
Luis	19	7373	M	Economia
Pedro	25	9289	M	Ing. Civil
Sara	19	8437	F	Ing. Sistemas

Ejemplo 6: Vehículos

Ejemplo 6: Vehículos

```
BD_1<-read.csv("DATOS/DB1.csv",header=TRUE)
names(BD_1)
```

```
## [1] "Codigo" "Marca"  "Peso"
```

```
BD_2<-read.csv("DATOS/DB2.csv",header=TRUE)
names(BD_2)
```

```
## [1] "Codigo" "Estado"
```

```
BD_3<-read.csv("DATOS/DB3.csv",header=TRUE)
names(BD_3)
```

```
## [1] "Importado"          "AireAcondicionado" "Codigo"
## [4] "Cilindraje"         "Potencia"
```

```
BD_4<-read.csv("DATOS/DB4.csv",header=TRUE)
names(BD_4)
```

```
## [1] "Estado"       "Nacionalidad" "Combustible"  "Puertas"
## [5] "Transmision"  "Codigo"
```

Ejemplo 6: Vehículos

```
Datos_6<-Reduce(function(...) merge (..., all=T),
                list(BD_1,BD_2,BD_3,BD_4))
names(Datos_6)
```

```
##  [1] "Codigo"          "Estado"          "Marca"
##  [4] "Peso"            "Importado"       "AireAcondicionado"
##  [7] "Cilindraje"      "Potencia"        "Nacionalidad"
## [10] "Combustible"     "Puertas"         "Transmision"
```

Ejemplo 6: Vehículos

```
str(Datos_6)
```

```
## 'data.frame':    12433 obs. of  12 variables:
##  $ Codigo           : int  10001001 10001002 10001003 10
##  $ Estado           : Factor w/ 1 level "Activo": 1 1 1
##  $ Marca            : Factor w/ 371 levels "ACB","ACURA"
##  $ Peso             : int  980 0 1005 1070 0 2225 1070 1
##  $ Importado        : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ AireAcondicionado: int  0 0 1 1 0 1 1 0 0 0 ...
##  $ Cilindraje       : int  1405 1400 1405 1405 1984 2956
##  $ Potencia         : int  43 0 75 84 0 113 83 170 160 1
##  $ Nacionalidad     : Factor w/ 29 levels "","ALE","ARG"
##  $ Combustible      : Factor w/ 6 levels "","DSL","ELT",
##  $ Puertas          : int  5 5 5 5 2 5 5 4 2 4 ...
##  $ Transmision      : Factor w/ 12 levels "","2X1","3X1"
```

Ejemplo 6: Vehículos

```
summary(Datos_6)
```

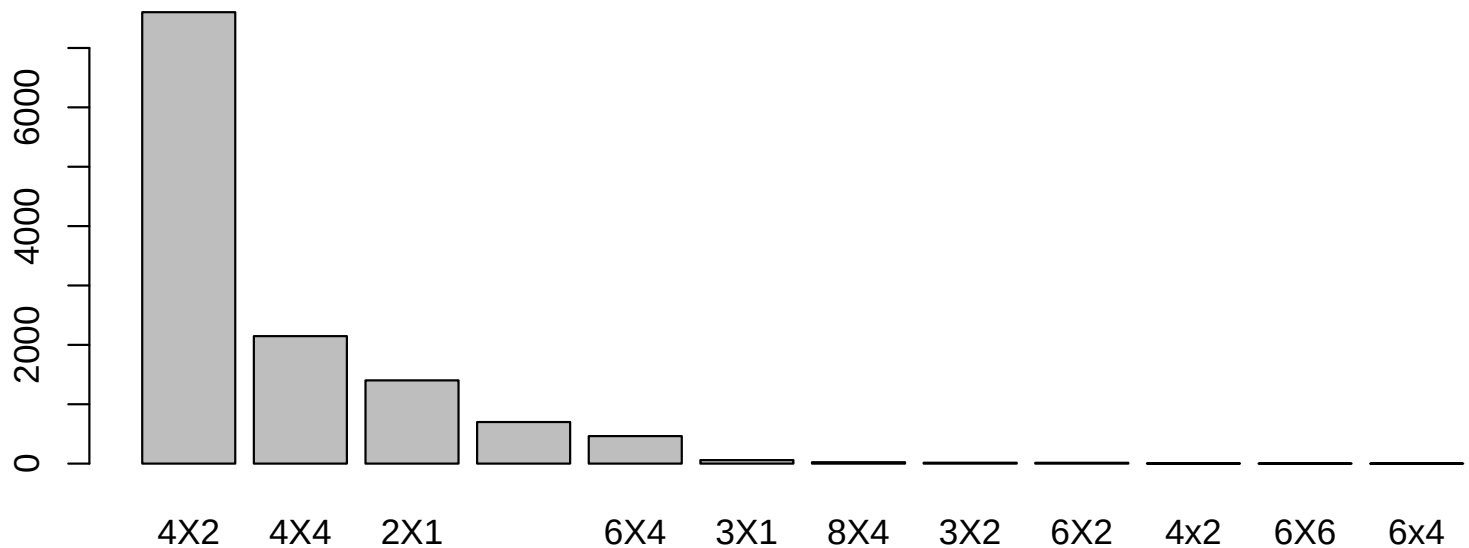
```
##      Codigo            Estado                  Marca           Peso
##  Min.   :  101001   Activo:12433   CHEVROLET     : 953   Min.   :    0
##  1st Qu.: 3201240                  MERCEDES BENZ: 505   1st Qu.:  268
##  Median : 6201016                  HYUNDAI       : 494   Median : 1355
##  Mean   : 8406562                  VOLKSWAGEN    : 493   Mean   : 1778
##  3rd Qu.: 9401013                  NISSAN        : 492   3rd Qu.: 2005
##  Max.   :40301001                  BMW           : 481   Max.   :41000
##                                    (Other)       :9015
##    Importado      AireAcondicionado   Cilindraje       Potencia
##  Min.   :0.0000   Min.   :0.0000    Min.   :    0   Min.   :  0.0
##  1st Qu.:1.0000   1st Qu.:0.0000    1st Qu.: 1339   1st Qu.: 80.0
##  Median :1.0000   Median :1.0000    Median : 1998   Median :122.0
##  Mean   :0.8407   Mean   :0.5501    Mean   : 2585   Mean   :134.7
##  3rd Qu.:1.0000   3rd Qu.:1.0000    3rd Qu.: 3246   3rd Qu.:175.0
##  Max.   :1.0000   Max.   :1.0000    Max.   :15950   Max.   :662.0
##
##   Nacionalidad  Combustible    Puertas       Transmision
##  COL    :1982       : 726    Min.   :0.000   4X2    :7603
##  CHI    :1810   DSL:3495     1st Qu.:2.000   4X4    :2147
##  JAP    :1535   ELT:  16     Median :3.000   2X1    :1402
##  ALE    :1384   GAS:  25     Mean   :2.952          : 700
##  KOR    :1123   GSL:8161     3rd Qu.:5.000   6X4    : 463
##  USA    : 913   HBD:  10     Max.   :6.000   3X1    :  59
##  (Other):3686                                (Other):  59
```

Ejemplo 6: Vehículos

```
summary(Datos_6$Transmision)
```

```
##          2X1   3X1   3X2   4x2   4X2   4X4   6X2   6x4   6X4   6X6   8X4
##  700    1402    59    15     4  7603  2147    14     1   463     2    23
```

```
barplot(sort(summary(Datos_6$Transmision),
             decreasing=TRUE))
```



Ejemplo 6: Vehículos

```
require(editrules)

Cond1<-editset(c("Transmision %in%
c('2X1','3X1', '3X2','4x2','4X2','4X4','6X2',
'6x4','6X4','6X6','8X4')", "Peso>0", "Potencia>0",
"Cilindraje>0", "Combustible %in%
c('DSL','ELT','GAS','GSL','HBD')"))
Cond1
```

```
##
## Data model:
## dat1 : Combustible %in% c('DSL', 'ELT', 'GAS', 'GSL', 'HBD')
## dat2 : Transmision %in% c('2X1', '3X1', '3X2', '4x2', '4X2',
##
## Edit set:
## num1 : 0 < Peso
## num2 : 0 < Potencia
## num3 : 0 < Cilindraje
```

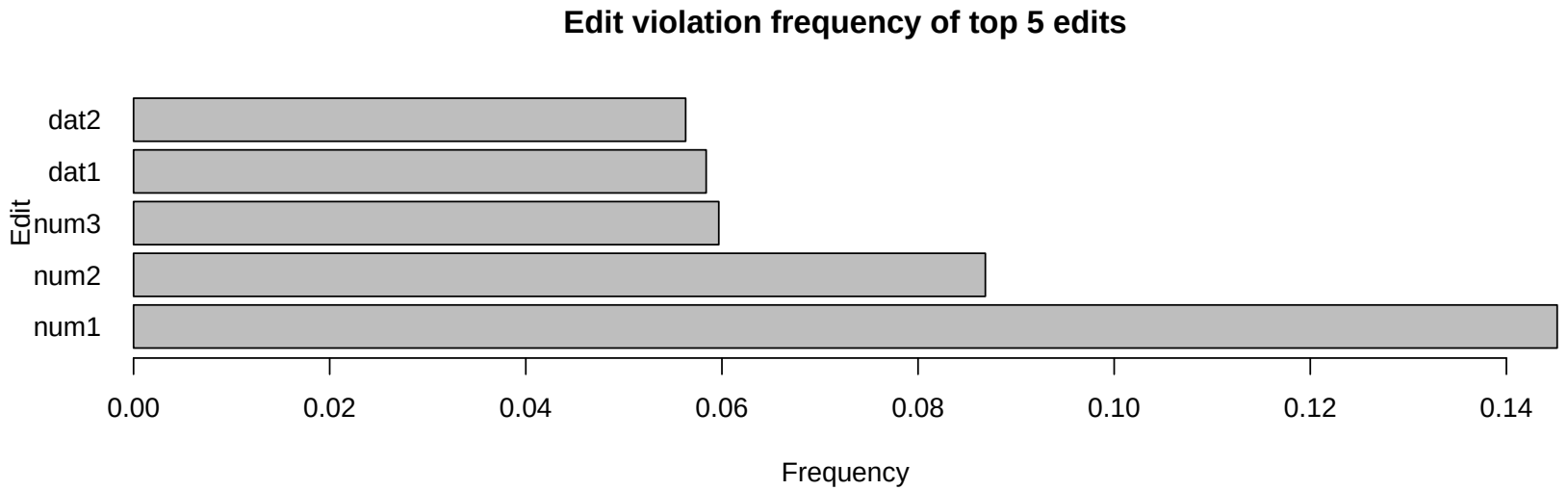
Ejemplo 6: Vehículos

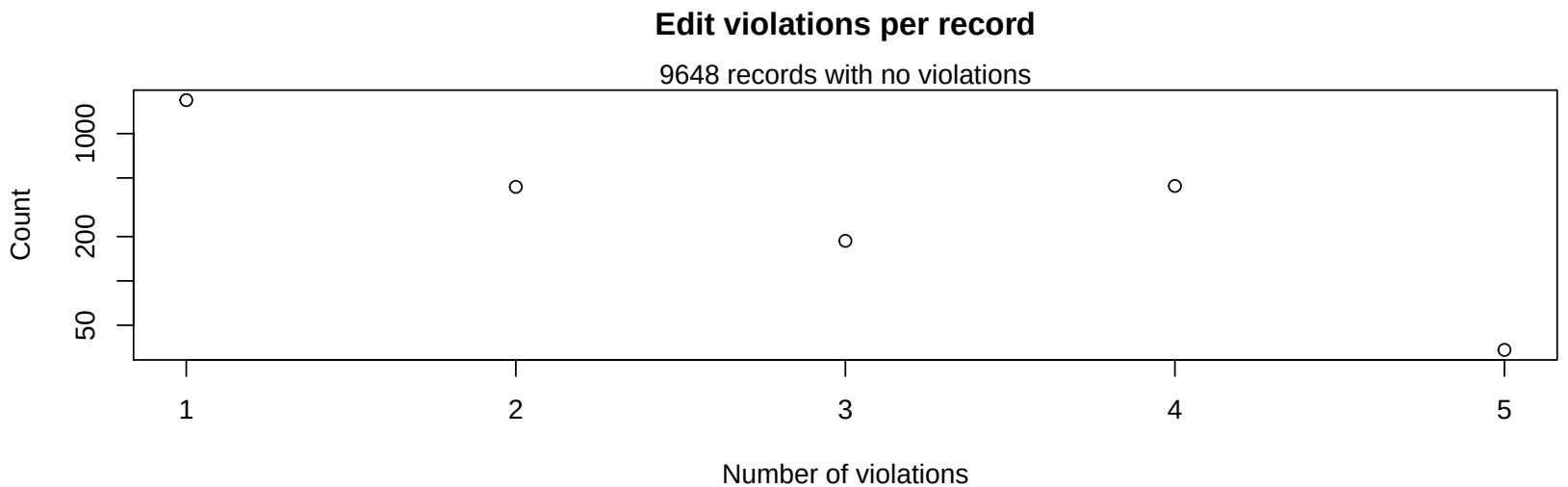
```
Errores1<-violatedEdits(c(Cond1), Datos_6)

Errores1[1:10,]
```

```
##        edit
## record  num1  num2  num3  dat1  dat2
##      1  FALSE FALSE FALSE FALSE FALSE
##      2   TRUE  TRUE FALSE FALSE FALSE
##      3  FALSE FALSE FALSE  TRUE  TRUE
##      4  FALSE FALSE FALSE  TRUE  TRUE
##      5   TRUE  TRUE FALSE  TRUE  TRUE
##      6  FALSE FALSE FALSE  TRUE FALSE
##      7  FALSE FALSE FALSE  TRUE  TRUE
##      8  FALSE FALSE FALSE FALSE FALSE
##      9  FALSE FALSE FALSE FALSE FALSE
##      10 FALSE FALSE FALSE FALSE FALSE
```

Ejemplo 6: Vehículos

```
plot(Errores1)
```





Ejemplo 6: Vehículos

```
Locali1<-localizeErrors(Cond1,Datos_6)$adapt
apply(X=Locali1, MARGIN = 2, FUN=function(x)
  which(x==TRUE))
```

```
## $Codigo
## named integer(0)
##
## $Estado
## named integer(0)
##
## $Marca
## named integer(0)
##
## $Peso
##     2     5    17    19    51    53    55    56    59
##     2     5    17    19    51    53    55    56    59
##    67    68    69    70    72    74    79    83    84
##    67    68    69    70    72    74    79    83    84
```

Ejemplo 6: Vehículos

```
Pesos_prob<-apply(X=Locali1, MARGIN = 2,
                  FUN=function(x)
  which(x==TRUE))$Peso
Pesos_prob[1:5]
```

```
##  2  5 17 19 51
##  2  5 17 19 51
```

```
Datos_6[Pesos_prob[1:5],1:4]
```

```
##       Codigo Estado Marca Peso
## 2   10001002 Activo  TATA    0
## 5   10004001 Activo  TATA    0
## 17  10012001 Activo  TATA    0
## 19  10021001 Activo  TATA    0
## 51  10203001 Activo IVECO    0
```

Ejemplo 6: Vehículos

```
Trans_prob<-apply(X=Locali1, MARGIN = 2,
                  FUN=function(x)
  which(x==TRUE))$Transmision
Trans_prob[1:5]
```

```
##  3  4  5  7 17
##  3  4  5  7 17
```

```
Datos_6[Trans_prob[1:5],10:12]
```

```
##    Combustible Puertas Transmision
## 3                    5
## 4                    5
## 5                    2
## 7                    5
## 17                   0
```

Ejemplo 6: Vehículos

```
Pote_prob<-apply(X=Locali1, MARGIN = 2,
                 FUN=function(x)
  which(x==TRUE))$Potencia
Pote_prob[1:5]
```

```
##  2  5 17 19 74
##  2  5 17 19 74
```

```
Datos_6[Pote_prob[1:5],6:9]
```

```
##    AireAcondicionado Cilindraje Potencia Nacionalidad
## 2                  0       1400        0         IND
## 5                  0       1984        0         IND
## 17                 0       2000        0         IND
## 19                 0       2000        0         IND
## 74                 0       3600        0
```

Ejemplo 6: Vehículos

```
Cilin_prob<-apply(X=Locali1, MARGIN = 2,
                  FUN=function(x)
  which(x==TRUE))$Cilindraje
Cilin_prob[1:5]
```

```
##  72  89 107 110 125
##  72  89 107 110 125
```

```
Datos_6[Cilin_prob[1:5],6:9]
```

```
##     AireAcondicionado Cilindraje Potencia Nacionalidad
## 72                  0          0      180         ARG
## 89                  0          0      180         ARG
## 107                 0          0      180         ARG
## 110                 0          0        0         COL
## 125                 0          0        0         ARG
```

Ejemplo 6: Vehículos

```
Comb_prob<-apply(X=Locali1, MARGIN = 2,
                 FUN=function(x)
  which(x==TRUE))$Combustible
Comb_prob[1:5]
```

```
## 3 4 5 6 7
## 3 4 5 6 7
```

```
Datos_6[Comb_prob[1:5],9:11]
```

```
##   Nacionalidad Combustible Puertas
## 3          IND                    5
## 4          IND                    5
## 5          IND                    2
## 6          IND                    5
## 7          IND                    5
```

Ejemplo 6: Vehículos

```
Problemas_1<-Reduce(union, list(Pesos_prob,Trans_prob,
                    Pote_prob,Cilin_prob,Comb_prob))

length(Problemas_1)
```

```
## [1] 2785
```

```
nrow(Datos_6)
```

```
## [1] 12433
```