Práctica 2: Limpieza y Validación de los Datos

Tipología y Ciclo de Vida de los Datos, Universitat Oberta de Catalunya

Abel Romero Búrdalo y Paula Muñoz Lago

27 diciembre 2020

Contents

1. Descripción del Dataset **2**

1.1. Importancia y objetivo del análisis 4

2. Integración y selección de los datos de interés a analizar **4**

3. Limpieza de los datos **5**

3.1. Tipos de Variables 5

3.2. Gestión de datos inválidos 6

3.3. Identificación y tratamiento de valores extremos 7

4. Análisis de los datos **7**

4.1. Selección de los grupos de datos a analizar 8

4.2. Comprobación de la normalidad y homogeneidad de la varianza 8

4.3. Aplicación de pruebas estadísticas 8

5. Representación de los resultados a partir de tablas y gráficas **8**

6. Conclusiones **8**

1. Descripción del Dataset

El dataset con el que vamos a realizar la práctica está relacionado con el número de votos por estado de EEUU en las recientes elecciones. Se ha obtenido de Kaggle: https://www.kaggle.com/imoore/2020-us-general-election-turnout-rates. Cabe destacar que los datos presentes en el dataset tratan únicamente la población con derecho a voto, es decir, únicamente los estadounidenses mayores de 18 años. La primera fila del dataset incluye información de la totalidad del país.

El conjunto dispone de 15 columnas:

- **State**: Indica el estado del que trata la fila de datos.
- **Source**: Fuente de datos (url).
- **Official/Unofficial**: Esta columna indica si los datos reportados son una vez el conteo ha alcanzado el 100% (oficial), o si aún no se ha terminado el conteo (unofficial).
- **Total Ballots Counted (Estimate)**: número total de votos en dicho estado.
- **Vote for Highest Office**: Votos a la presidencia.
- **VEP Turnout Rate**: Porcentaje de votantes. VEP, en inglés: Voting Elegible Population
- **Voting-Eligible Population**: Población con derecho a voto.
- **Voting-Age Population (VAP)**: Población total de estados unidos con 18 años o más, incluyendo a personas sin derecho a voto por razones diferentes a la edad, como personas sin la nacionalidad o criminales de ciertos estados, donde la ley se lo prohibe. fuente
- **% Non-citizen**: Porcentaje de personas con derecho a voto que no son ciudadanos estadounidenses.
- **Prision**: Número de votantes desde la carcel.
- **Probation**: Número de criminales con el tercer grado. Es decir, disfrutan de un periodo fuera de la carcel bajo supervisión.
- **Parole**: Personas con permiso de permanencia temporal en EEUU.
- **Total Ineligible Felon**: Número de personas en dicho estado que no tienen derecho a voto por criminalidad.
- **Overseas Eligible**: Número de estadounidenses viviendo fuera del país, independientemente del estado.
- **State abv**: Abreviatura del estado.

Antes de proseguir, cargaremos los datos y realizaremos una breve inspección sobre los mismos (excepto sobre la columna sources, que contiene urls), para estudiar los valores contenidos en cada columna.

```
fileDirectory <- getwd()
csv_usa <- file.path(fileDirectory, '2020 November General Election - Turnout Rates.csv')
usa_elections <- read.csv(csv_usa)
```

```
library(Hmisc)
Hmisc::describe(usa_elections[, -2])
```

```
## usa_elections[, -2]
##
##  14  Variables      52  Observations
## ---------------------------------------------------------------------------
## State
##        n  missing distinct
##       52        0       52
##
## lowest : Alabama       Alaska        Arizona       Arkansas      California
## highest: Virginia      Washington    West Virginia Wisconsin     Wyoming
## ---------------------------------------------------------------------------
```

```
## Official.Unofficial
##        n  missing distinct
##       52        0        3
##
## Value                   OFFICIAL Unofficial
## Frequency          27          2         23
## Proportion      0.519      0.038      0.442
## ------------------------------------------------------------------------------
## Total.Ballots.Counted..Estimate.
##        n  missing distinct
##       52        0       50
##
## lowest : 1,212,030 1,330,000 1,340,000 1,370,000 1,450,000
## highest: 814,092   860,000   875,000   923,612   948,852
## ------------------------------------------------------------------------------
## Vote.for.Highest.Office..President.
##        n  missing distinct
##       52        0       25
##
## lowest :            1,206,697  1,333,513  1,560,699  11,231,799
## highest: 603,635    803,833    867,258    919,377    935,232
## ------------------------------------------------------------------------------
## VEP.Turnout.Rate
##        n  missing distinct
##       52        0       48
##
## lowest : 55.0% 55.5% 57.0% 57.5% 59.7%, highest: 76.1% 76.4% 78.6% 79.2% 79.9%
## ------------------------------------------------------------------------------
## Voting.Eligible.Population..VEP.
##        n  missing distinct
##       52        0       52
##
## lowest : 1,007,920 1,079,434 1,085,285 1,292,701 1,383,551
## highest: 799,642   8,859,167 837,298   9,027,082 9,781,976
## ------------------------------------------------------------------------------
## Voting.Age.Population..VAP.
##        n  missing distinct
##       52        0       52
##
## lowest : 1,104,489 1,114,466 1,115,916 1,384,683 1,422,098
## highest: 8,328,642 851,663   857,507   9,144,626 9,832,749
## ------------------------------------------------------------------------------
## X..Non.citizen
##        n  missing distinct
##       52        0       37
##
## lowest : 0.9%  1.2%  1.4%  1.7%  10.1%, highest: 7.8%  8.4%  8.7%  8.9%  9.1%
## ------------------------------------------------------------------------------
## Prison
##        n  missing distinct
##       52        0       50
##
## lowest : 0         1,461,074 1,679     104,730   12,399
## highest: 8,378     9,216     9,712     9,882     91,674
```

```
## ------------------------------------------------------------------------------
## Probation
##        n  missing distinct
##       52        0       30
##
## lowest : 0         1,962,811 100,076   12,090    14,176
## highest: 61,253    63,111    76,672    76,844    80,068
## ------------------------------------------------------------------------------
## Parole
##        n  missing distinct
##       52        0       33
##
## lowest : 0      1,348  1,780  1,860  10,266, highest: 7,381  7,536  9,866  934    958
## ------------------------------------------------------------------------------
## Total.Ineligible.Felon
##        n  missing distinct
##       52        0       50
##
## lowest : 0      1,679  10,781 12,399 13,795, highest: 79,140 83,304 87,600 93,699 97,497
## ------------------------------------------------------------------------------
## Overseas.Eligible
##        n  missing distinct
##       52        0        2
##
## Value                4,971,025
## Frequency               51         1
## Proportion           0.981     0.019
## ------------------------------------------------------------------------------
## State.Abv
##        n  missing distinct
##       52        0       52
##
## lowest :    AK AL AR AZ, highest: VT WA WI WV WY
## ------------------------------------------------------------------------------
```

1.1. Importancia y objetivo del análisis

Gracias a este dataset podemos estudiar cual ha sido el porcentaje de votantes, ya sea a favor de Trump o Biden. Así como plantear algunas clonclusiones sobre las diferencias por estados en cuanto a votos republicanos o demócratas.

Estos análisis son de gran relevancia a la hora de establecer patrones de voto en grupos poblacionales en función de ciertas características.

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

En vistas a la descripción de las columnas observamos que disponemos de columnas repetidas, como es el caso de la abreviatura del estado y el nombre del mismo. Por ello, la columna relativa a la abreviatura del estado será la primera que eliminemos, con el fin de evitar redundancia en los datos.

```
usa_elections <- usa_elections[, -ncol(usa_elections)]
```

Proseguiremos con la nueva última columna, "Overseas Eligible", que se refiere al número de estadounidenses viviendo fuera del país. Ésta columna solo tiene un valor diferente a null, y está relacionado con el dato en la primera fila, correspondiente con la totalidad de estados. Es por ello, que a continuación retiraremos la priemra fila y la guardaremos en una variable, para así poder estuadiar los datos por estado, pero manteniendo la información del total por si nos hiciese falta a continuación. Finalmente eliminaremos la columna "Overseas Eligible", dado que todos sus valores son null.

```
usa_total <- usa_elections[1,]
usa_elections <- usa_elections[2:nrow(usa_elections),-ncol(usa_elections)]
```

La carencia de utilidad de las columnas relativas a la fuente de datos y si se trata de una fuente oficial o no, hacen que también procedamos a eliminarlas del dataset.

```
usa_elections <- usa_elections[,-c(grep("Source", colnames(usa_elections)), grep("Official.Unofficial",
```

3. Limpieza de los datos

En este apartado llevaremos a cabo un proceso de limpieza de datos, comenzando por establecer los tipos de datos correctos para cada variable (columna), y gestionando los valores nulos (casillas vacías). Finalmente, estudiaremos la presencia de valores extremos y cómo tratarlos.

3.1. Tipos de Variables

Cada columna de nuestro dataframe `usa_elections` es un Factor, conteniendo diferentes niveles. En el siguiente resumen de nuestro dataframe podemos observar el tipo de datos correspondiente con cada uno de ellos.

```
str(usa_elections)
```

```
## 'data.frame':    51 obs. of  11 variables:
##  $ State                              : Factor w/ 52 levels "Alabama","Alaska",..: 1 2 3 4 5 6 7 8 9
##  $ Total.Ballots.Counted..Estimate.   : Factor w/ 50 levels "1,212,030","1,330,000",..: 15 29 25 1 1
##  $ Vote.for.Highest.Office..President.: Factor w/ 25 levels "","1,206,697",..: 8 1 1 2 1 1 1 19 1 1
##  $ VEP.Turnout.Rate                   : Factor w/ 48 levels "55.0%","55.5%",..: 10 30 20 2 16 45 33 3
##  $ Voting.Eligible.Population..VEP.   : Factor w/ 52 levels "1,007,920","1,079,434",..: 24 38 36 13 1
##  $ Voting.Age.Population..VAP.        : Factor w/ 52 levels "1,104,489","1,114,466",..: 24 38 36 13 1
##  $ X..Non.citizen                     : Factor w/ 37 levels "0.9%","1.2%",..: 12 18 36 20 10 25 32 2
##  $ Prison                             : Factor w/ 50 levels "0","1,461,074",..: 21 34 33 9 4 14 6 40
##  $ Probation                          : Factor w/ 30 levels "0","1,962,811",..: 24 8 29 15 1 1 1 5 1
##  $ Parole                             : Factor w/ 33 levels "0","1,348","1,780",..: 5 2 30 16 7 1 22
##  $ Total.Ineligible.Felon             : Factor w/ 50 levels "0","1,679","10,781",..: 38 36 49 37 16
```

Como se aprecia, los datos relativos a números se han cargado como strings, por lo que podemos concluir que la única columna que se encuentra en un estado definitivo es la primera, State. Para el resto de columnas debemos aplicar una limpieza, quitando los caracteres no numéricos y así poder convertirlos al tipo de datos correcto. Además, dichas columnas no queremos que sean de tipo Factor, dado que son variables contínuas, la única que mantendremos como tipo Factor será State, dado que es una variable discreta.

Para ello definiremos dos funciones, que aplicaremos a las columnas que lo necesiten. Estas funciones eliminarán los caracteres necesarios para a continuación poder convertir los datos a números.

```
# Definición de las funciones
remove_comma <- function(x) gsub(',', '', x)
remove_percent <- function(x) gsub('%', '', x)

# Aplicación de las mismas sobre las columnas apropiadas
usa_elections[,2] <- sapply(usa_elections[,2], remove_comma)
usa_elections[,3] <- sapply(usa_elections[,3], remove_comma)
usa_elections[,4] <- sapply(usa_elections[,4], remove_percent)
usa_elections[,5] <- sapply(usa_elections[,5], remove_comma)
usa_elections[,6] <- sapply(usa_elections[,6], remove_comma)
usa_elections[,7] <- sapply(usa_elections[,7], remove_percent)
usa_elections[,8] <- sapply(usa_elections[,8], remove_comma)
usa_elections[,9] <- sapply(usa_elections[,9], remove_comma)
usa_elections[,10] <- sapply(usa_elections[,10], remove_comma)
usa_elections[,11] <- sapply(usa_elections[,11], remove_comma)
```

Una vez obtenido el resultado necesario para poder convertir al tipo deseado, ejecutamos las siguientes líneas:

```
usa_elections[,2] <- as.numeric(usa_elections[,2])
usa_elections[,3] <- as.numeric(usa_elections[,3])
usa_elections[,4] <- as.numeric(usa_elections[,4])
usa_elections[,5] <- as.numeric(usa_elections[,5])
usa_elections[,6] <- as.numeric(usa_elections[,6])
usa_elections[,7] <- as.numeric(usa_elections[,7])
usa_elections[,8] <- as.numeric(usa_elections[,8])
usa_elections[,9] <- as.numeric(usa_elections[,9])
usa_elections[,10] <- as.numeric(usa_elections[,10])
usa_elections[,11] <- as.numeric(usa_elections[,11])
```

Finalmente, imprimiremos el resumen de las columnas de nuestro dataset para comprobar que todo se ha transformado correctamente.

```
str(usa_elections)
```

```
## 'data.frame':    51 obs. of  11 variables:
##  $ State                              : Factor w/ 52 levels "Alabama","Alaska",..: 1 2 3 4 5 6 7 8 9
##  $ Total.Ballots.Counted..Estimate.   : num  2306587 367000 3400000 1212030 16800000 ...
##  $ Vote.for.Highest.Office..President.: num  2297295 NA NA 1206697 NA ...
##  $ VEP.Turnout.Rate                   : num  62.6 69.8 65.5 55.5 64.7 76.4 71.1 70.5 64.7 71.7 ...
##  $ Voting.Eligible.Population..VEP.   : num  3683055 525568 5189000 2182375 25962648 ...
##  $ Voting.Age.Population..VAP.        : num  3837540 551117 5798473 2331171 30783255 ...
##  $ X..Non.citizen                     : num  2.3 3.4 8.9 3.6 15 5.7 7.7 5.8 7.1 10.1 ...
##  $ Prison                             : num  25898 4293 38520 17510 104730 ...
##  $ Probation                          : num  50997 2074 76844 36719 0 ...
##  $ Parole                             : num  10266 1348 7536 24698 102586 ...
##  $ Total.Ineligible.Felon             : num  67782 6927 93699 64974 207316 ...
```

3.2. Gestión de datos inválidos

Para comprobar qué columnas contienen datos 'vacíos' y poder proceder a trabajar con ellas, utilizaremos la función colSums, que aplica una función a todas las columnas de un dataframe y después aplica una suma.

```
colSums(is.na(usa_elections))
```

```
##                                State    Total.Ballots.Counted..Estimate.
##                                    0                                   0
## Vote.for.Highest.Office..President.                    VEP.Turnout.Rate
##                                   27                                   0
##    Voting.Eligible.Population..VEP.         Voting.Age.Population..VAP.
##                                    0                                   0
##                       X..Non.citizen                              Prison
##                                    0                                   0
##                            Probation                              Parole
##                                    0                                   0
##              Total.Ineligible.Felon
##                                    0
```

Como vemos, únicamente disponemos de una columna con datos vacíos, Vote for Highest Office President. Aquellos campos que contentan el valor NA, los cambbiaremos por 0, dando a entender que en ese estado no ha ganado el presidente. REVISAR: no le veo mucho sentido, quizás no significa lo que creo?

3.3. Identificación y tratamiento de valores extremos

4. Análisis de los datos

Gracias al tratamiento de los datos como numéricos en el punto 3.1, podemos ejecutar pequeños análisis estadísticos, en los que observar la distribución de los datos.

```
summary(usa_elections)
```

```
##         State    Total.Ballots.Counted..Estimate.
##  Alabama   : 1   Min.   :  278503
##  Alaska    : 1   1st Qu.:  867500
##  Arizona   : 1   Median : 2155000
##  Arkansas  : 1   Mean   : 3114412
##  California: 1   3rd Qu.: 3912500
##  Colorado  : 1   Max.   :16800000
##  (Other)   :45
##  Vote.for.Highest.Office..President. VEP.Turnout.Rate
##  Min.   :  276765                    Min.   :55.00
##  1st Qu.:  753784                    1st Qu.:64.35
##  Median : 1447106                    Median :67.60
##  Mean   : 2178698                    Mean   :67.77
##  3rd Qu.: 2638094                    3rd Qu.:72.60
##  Max.   :11231799                    Max.   :79.90
##  NA's   :27
##  Voting.Eligible.Population..VEP. Voting.Age.Population..VAP. X..Non.citizen
##  Min.   :  431364                 Min.   :  447915            Min.   : 0.900
##  1st Qu.: 1338126                 1st Qu.: 1403390            1st Qu.: 3.000
##  Median : 3312250                 Median : 3479257            Median : 4.300
##  Mean   : 4600645                 Mean   : 5051080            Mean   : 5.443
##  3rd Qu.: 5313422                 3rd Qu.: 5934260            3rd Qu.: 7.400
##  Max.   :25962648                 Max.   :30783255            Max.   :15.000
```

```
##
##      Prison          Probation          Parole       Total.Ineligible.Felon
##  Min.   :     0   Min.   :     0   Min.   :     0   Min.   :     0
##  1st Qu.:  6080   1st Qu.:     0   1st Qu.:     0   1st Qu.: 11590
##  Median : 18099   Median :  5989   Median :  1860   Median : 33933
##  Mean   : 23719   Mean   : 38167   Mean   :  9972   Mean   : 57354
##  3rd Qu.: 31208   3rd Qu.: 42236   3rd Qu.: 10066   3rd Qu.: 73060
##  Max.   :154913   Max.   :416771   Max.   :109213   Max.   :492390
##
```

4.1. Selección de los grupos de datos a analizar

4.2. Comprobación de la normalidad y homogeneidad de la varianza

4.3. Aplicación de pruebas estadísticas

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

Mostrar los datos sobre el mapa de eeuu

6. Conclusiones