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

26 diciembre 2020

Contents

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

Todas aquellas que son porcentajes (string), cambiar a numérico. Quitar las comas que separan los números. https://www.kaggle.com/thomaskonstantin/2020-us-election-turnout-rates-eda

3.1. Gestión de datos inválidos

3.2. Identificación y tratamiento de valores extremos

4. Análisis de los datos

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

6. Conclusiones