Trabajo final de curso de Ciencia de Datos con R

Martín S. Armoa
Docente:
Leyla Scheli

12 de agosto de 2021

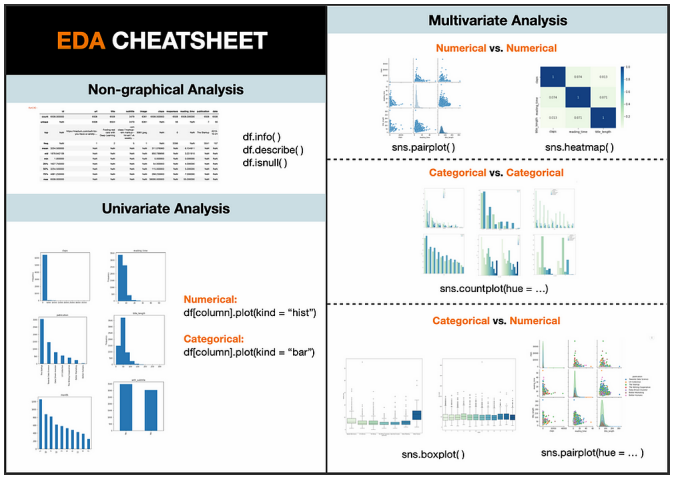
1. Introducción

Antes de empezar un proyecto de análisis de datos, siempre con el fin de mejorar la toma de decisión o para sacar conclusiones sobre un hecho del mundo en el caso de un científico, necesitamos recolectar datos para luego intentar responder las preguntas de interés o las hipótesis planteadas. Puede pasar que los datos sean de terceros y no sepamos las características de ellos, como su origen, formato, limitaciones, etc. Para lograr tener información acerca de los datos que tenemos a disposición el estadístico estadounidense John Wilder Tukey desarrollo un conjunto de técnicas para poder comprender de manera rápida la naturaleza de una colección de datos o dataset. A ese conjunto de técnicas de lo denomina por sus siglas en ingles Exploratory Data Analysis (EDA) con el cual se genera una primera impresión de los datos, se determina que tipo de información contiene, ya sea números, categorías, etc. se buscan patrones y anomalías y se hacen resúmenes estadísticos y gráficas para una mejor visualización como se puede ver en la Figura 1.

Se presentan a continuación algunas de la técnicas implementadas en el EDA realizado en la sección 2:

-Resúmenes Estadísticos -Percentiles y BoxPlot -Correlaciones

Figura 1. Resumen gráfico de técnicas de EDA



2. Desarrollo

Se parte de un conjunto de datos dados por terceros relacionados con la Educación Sexual. Lo primero que se realizo fue una visualización rápida de los datos para saber de que información dispongo, como cantidad de datos, la clase, las categorías y de que tipo son.

Se puede ver en la Figura 2 que el "Data Frame.^{es}ta formado por 15157 filas con 6 columnas las cuales indican distintas categorías como edad, años de educación, numero de hijos etc. referidos a distintas personas (id) los cuales son de tipo Real (dbl) (ver código en R)

Figura 2. Visualización rápida del dataset

	id	edad	anios_educ	en_pareja	num_hijos	bajo_socioecon
3	3	15	9	0	0	1
4	4	17	9	1	0	0
5	5	18	9	1	0	0
6	6	17	9	0	0	0
7	7	18	9	0	0	1
8	8	16	12	0	1	1
9	9	16	7	0	0	0
10	10	17	7	1	1	0
11	11	19	12	0	0	0
12	12	17	8	0	0	1
13	13	16	8	0	0	1

Showing 2 to 14 of 15,157 entries, 6 total columns

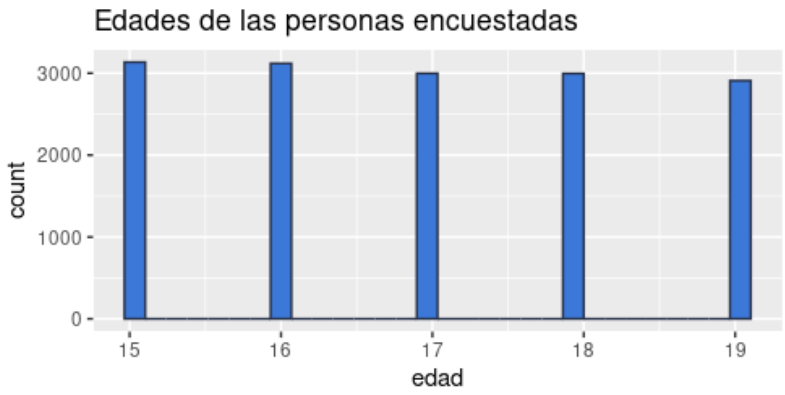
Luego se realizo un resumen estadístico en el cual se obtuvo información sobre la media, la mediana, la desviación típica o el mínimo y el máximo de cada una de las variables categóricas como se puede ver en la Figura 3. Rápidamente se puede ver que el rango de edad de las 15157 personas va de 15 a 19 años con una franja de años de educación entre 6 y 12 años de los cuales mu

Figura 3

```
> summary(edusex)
       id             edad         anios_educ      en_pareja        num_hijos      bajo_socioecon
 Min.   :    1   Min.   :15.00   Min.   : 6.000   Min.   :0.0000   Min.   :0.0000   Min.   :0.00
 1st Qu.: 3790   1st Qu.:16.00   1st Qu.: 8.000   1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.00
 Median : 7579   Median :17.00   Median : 9.000   Median :0.0000   Median :0.0000   Median :0.00
 Mean   : 7579   Mean   :16.96   Mean   : 8.507   Mean   :0.1617   Mean   :0.1509   Mean   :0.25
 3rd Qu.:11368   3rd Qu.:18.00   3rd Qu.: 9.000   3rd Qu.:0.0000   3rd Qu.:0.0000   3rd Qu.:1.00
 Max.   :15157   Max.   :19.00   Max.   :12.000   Max.   :1.0000   Max.   :3.0000   Max.   :1.00
```

Luego se realiza histograma con cada variable para tener una mejor visualización de las distribuciones de los datos:

Figura 4



Se realizo un diagrama Figura 8 de cajas para comparar la edad con la cantidad de hijos en el cual se puede ver datos outliers en el Boxplot de Nro de hijos.

Figura 5

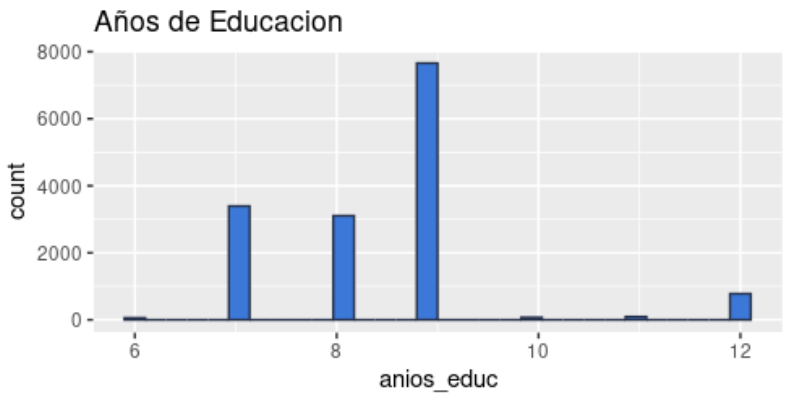


Figura 6

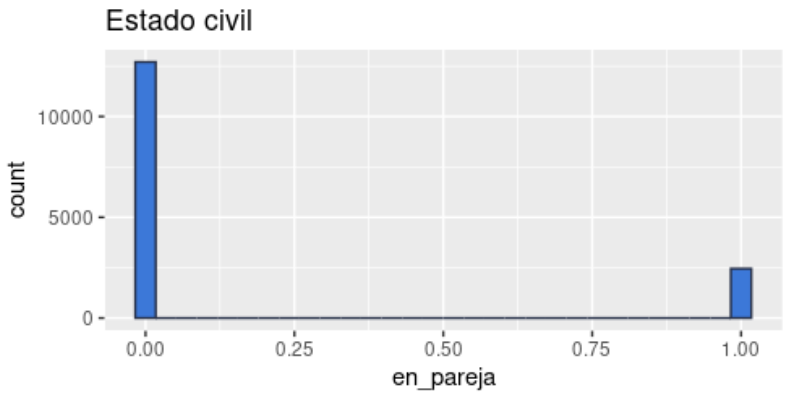
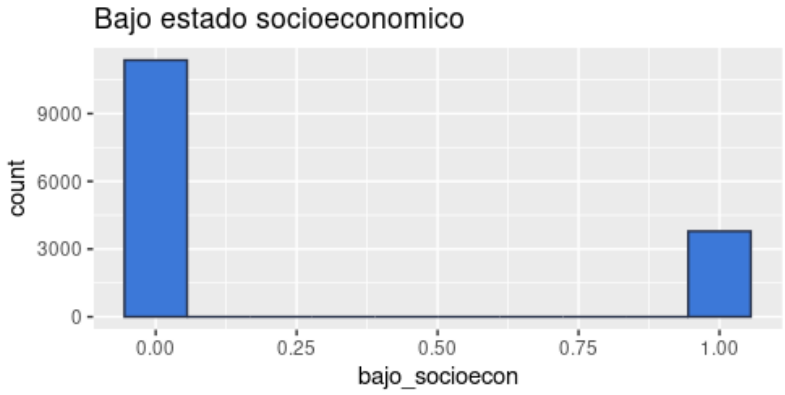


Figura 7. Visualización rápida del dataset



A continuación se hace un estudio de la correlación entre las distintas variables.

Figura 8. Diagrama de cajas Edad vs Cant de hijos

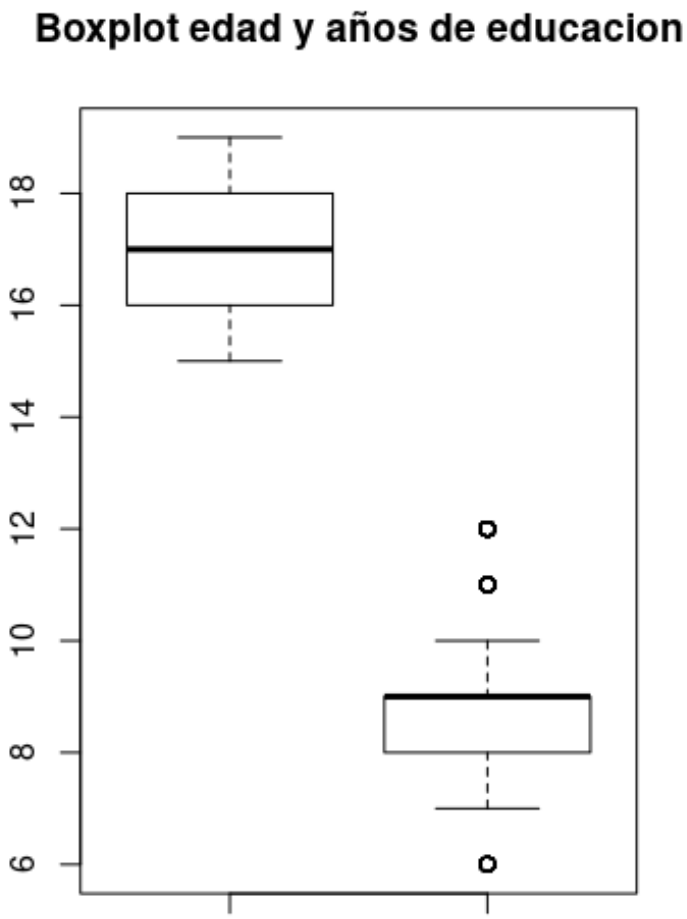


Tabla 1. Correlacion entre las distintas variables

Variable 1	variable 2	coef. de correlación
Edad	Años de educación	0.04
Años de educación	Educación sexual	-0.0009
En pareja	años de educación	0.15
Numero de hijos	Estado socioecon.	0.019

3. Conclusión

Como podemos ver en un análisis exploratorio de datos podemos obtener información acerca de los datos que tenemos a disposición, aunque sea en una primera aproximación en el que se pueden generar preguntas acercas de los datos y intentar buscar respuestas visualizando, trasformando y modelando los datos.

En nuestro caso podemos ver una dataset formado de 15157 personas(id) las cuales se agrupan por edad, años de educación, si están en pareja, cuantos hijos tienen y su situación económica si es mala o no.

Sabes que la cantidad de personas encuestadas tiene una edad entre 15 y 17 años las cuales se encuestaron aproximadamente 3000 para cada edad como muestra la Figura 4 las cuales mas de la mitad, aproximadamente 8000 personas tienen 9 años de educación y gran parte de nuestra muestra no esta en pareja Figura 6 con un estado socio económico bueno (interpretando a 1 como bajo y cero como bueno).

Haciendo un análisis de correlación podemos ver que las variables no se encuentran correlacionadas, dando en coeficiente mas alto 0.15 para Años de educación vs En parejas.

Para ser mas concluyentes hace falta un análisis mas detallado y un mayor conocimiento de las herramientas que nos brinda el lenguaje R.

Referencias

[1] http://www.stat.cmu.edu/ hseltman/309/Book/chapter4.pdf

[2] https://es.wikipedia.org/wiki/$John_{W.T}ukey$

[3] https://online.datasciencedojo.com/blogs/data-science-roadmap-a-comprehensive-career-guide