

Clase_5

2024-10-04

COSO

Procesamiento y cargue de la data

```
rm(list = ls ())

library(ggplot2)
library(gridExtra)

datos = read.csv("/home/sebastian/Downloads/Saber.csv")

datos_limpios = na.omit(datos)
set.seed(123)
n=1000
muestra = datos_limpios[sample(1:nrow(datos_limpios),n),]
head(muestra)
```

```
##          ESTU_TIPODOCUMENTO ESTU_NACIONALIDAD ESTU_GENERO  ESTU_FECHANACIMIENTO
## 458052                TI          COLOMBIA          M 08/31/2003 12:00:00 AM
## 194856                CC          COLOMBIA          M 09/28/1999 12:00:00 AM
## 138374                TI          COLOMBIA          F 11/21/2003 12:00:00 AM
## 128043                CC          COLOMBIA          F 11/06/2001 12:00:00 AM
## 166121                TI          COLOMBIA          F 09/01/2003 12:00:00 AM
## 233260                TI          COLOMBIA          M 08/11/2003 12:00:00 AM
##          PERIODO ESTU_CONSECUTIVO ESTU_ESTUDIANTE ESTU_PAIS_RESIDE
## 458052    20204 SB11202040011238      ESTUDIANTE      COLOMBIA
## 194856    20204 SB11202040072865      ESTUDIANTE      COLOMBIA
## 138374    20204 SB11202040193044      ESTUDIANTE      COLOMBIA
## 128043    20204 SB11202040345222      ESTUDIANTE      COLOMBIA
## 166121    20204 SB11202040467031      ESTUDIANTE      COLOMBIA
## 233260    20204 SB11202040138481      ESTUDIANTE      COLOMBIA
##          ESTU_TIENEETNIA ESTU_DEPTO_RESIDE ESTU_COD_RESIDE_DEPTO
## 458052                No      CUNDINAMARCA                25
## 194856                No              CAUCA                19
## 138374                No              META                50
## 128043                No              VALLE                76
## 166121                No            BOYACA                15
## 233260                No            BOLIVAR                13
##          ESTU_MCPIO_RESIDE ESTU_COD_RESIDE_MCPIO FAMI_ESTRATOVIVIENDA
## 458052          LA CALERA                25377          Estrato 2
## 194856          CAJIBÍO                19130          Estrato 1
## 138374    VILLAVICENCIO                50001          Estrato 1
## 128043          CALI                76001          Estrato 3
## 166121          ALMEIDA                15022          Estrato 3
## 233260 CARTAGENA DE INDIAS                13001          Estrato 3
```

##	FAMI_PERSONASHOGAR	FAMI_CUARTOSHOGAR	FAMI_EDUCACIONPADRE
## 458052	1 a 2	Dos	Primaria incompleta
## 194856	5 a 6	Dos	Primaria incompleta
## 138374	3 a 4	Tres	No sabe
## 128043	3 a 4	Cuatro	Secundaria (Bachillerato) completa
## 166121	5 a 6	Dos	Secundaria (Bachillerato) completa
## 233260	5 a 6	Dos	No sabe
##	FAMI_EDUCACIONMADRE		
## 458052	Técnica o tecnológica completa		
## 194856	Primaria incompleta		
## 138374	Primaria completa		
## 128043	Educación profesional completa		
## 166121	Educación profesional completa		
## 233260	Secundaria (Bachillerato) incompleta		
##	FAMI_TRABAJOLABORPADRE		
## 458052	No sabe		
## 194856	Es agricultor, pesquero o jornalero		
## 138374	No sabe		
## 128043	Trabaja por cuenta propia (por ejemplo plomero, electricista)		
## 166121	Trabaja por cuenta propia (por ejemplo plomero, electricista)		
## 233260	Trabaja por cuenta propia (por ejemplo plomero, electricista)		
##	FAMI_TRABAJOLABORMADRE		
## 458052	Es agricultor, pesquero o jornalero		
## 194856	Es agricultor, pesquero o jornalero		
## 138374	Es dueño de un negocio pequeño (tiene pocos empleados o no tiene, por ejemplo tienda, papelería)		
## 128043	Trabaja como profesional (por ejemplo médico, abogado, ingeniero)		
## 166121	Trabaja como profesional (por ejemplo médico, abogado, ingeniero)		
## 233260	Es dueño de un negocio pequeño (tiene pocos empleados o no tiene, por ejemplo tienda, papelería)		
##	FAMI_TIENEINTERNET	FAMI_TIENESERVICIO	FAMI_TIENECOMPUTADOR
## 458052	Si	Si	Si
## 194856	No	No	No
## 138374	No	No	No
## 128043	Si	Si	Si
## 166121	Si	Si	Si
## 233260	Si	Si	Si
##	FAMI_TIENELAVADORA	FAMI_TIENEHORNOMICRO	FAMI_TIENEAUTOMOVIL
## 458052	Si	No	Si
## 194856	No	No	No
## 138374	No	Si	No
## 128043	Si	Si	No
## 166121	Si	No	No
## 233260	Si	Si	No
##	FAMI_TIENEMOTOCICLETA	FAMI_TIENECONSOLAVIDEOJUEGOS	FAMI_NUMLIBROS
## 458052	No	No	26 A 100 LIBROS
## 194856	Si	No	0 A 10 LIBROS
## 138374	No	No	0 A 10 LIBROS
## 128043	No	No	11 A 25 LIBROS
## 166121	Si	Si	0 A 10 LIBROS
## 233260	Si	No	0 A 10 LIBROS
##	FAMI_COMELECHEDERIVADOS	FAMI_COMECARNEPESCADO	FAMI_COMEHUEVO
## 458052	3 a 5 veces por semana	3 a 5 veces por semana	
## 194856	1 o 2 veces por semana	Todos o casi todos los días	
## 138374	1 o 2 veces por semana	1 o 2 veces por semana	
## 128043	Todos o casi todos los días	Todos o casi todos los días	

## 166121	3 a 5 veces por semana	3 a 5 veces por semana	
## 233260	3 a 5 veces por semana	Todos o casi todos los días	
##	FAMI_COMECEREALFRUTOSLEGUMBRE	FAMI_SITUACIONECONOMICA	
## 458052	1 o 2 veces por semana	Igual	
## 194856	Nunca o rara vez comemos eso	Igual	
## 138374	3 a 5 veces por semana	Igual	
## 128043	1 o 2 veces por semana	Igual	
## 166121	1 o 2 veces por semana	Mejor	
## 233260	1 o 2 veces por semana	Igual	
##	ESTU_DEDICACIONLECTURADIARIA	ESTU_DEDICACIONINTERNET	
## 458052	Entre 30 y 60 minutos	30 minutos o menos	
## 194856	30 minutos o menos	30 minutos o menos	
## 138374	30 minutos o menos	No Navega Internet	
## 128043	30 minutos o menos	Más de 3 horas	
## 166121	Entre 30 y 60 minutos	Entre 1 y 3 horas	
## 233260	Entre 1 y 2 horas	Más de 3 horas	
##	ESTU_HORASSEMANATRABAJA	ESTU_TIPOREMUNERACION	COLE_CODIGO_ICFES
## 458052	Menos de 10 horas	Si, en efectivo	8847
## 194856	Entre 11 y 20 horas	Si, en efectivo	120816
## 138374	0	No	81950
## 128043	0	No	267021
## 166121	0	No	745273
## 233260	0	No	146274
##	COLE_COD_DANE_ESTABLECIMIENTO	COLE_NOMBRE_ESTABLECIMIENTO	
## 458052	3.25377e+11	COLEGIO COOPERATIVO PAULO VI LA CALERA	
## 194856	2.19130e+11	I.E. EFRAIN OROZCO	
## 138374	1.50001e+11	COL DPTAL NARCISO JOSE MATUS TORRES	
## 128043	1.76001e+11	INSTITUCION EDUCATIVA REPUBLICA DE ISRAEL	
## 166121	3.15022e+11	I.E. ENRIQUE SUAREZ	
## 233260	1.13001e+11	INSTITUCION EDUCATIVA MERCEDES ABREGO	
##	COLE_GENERO	COLE_NATURALEZA	COLE_CALEDARIO
## 458052	MIXTO	NO OFICIAL	A N
## 194856	MIXTO	OFICIAL	A N
## 138374	MIXTO	OFICIAL	A N
## 128043	MIXTO	OFICIAL	A N
## 166121	MIXTO	OFICIAL	A N
## 233260	MIXTO	OFICIAL	A N
##	COLE_CARACTER	COLE_COD_DANE_SEDE	
## 458052	TÉCNICO	3.25377e+11	
## 194856	ACADÉMICO	2.19130e+11	
## 138374	TÉCNICO/ACADÉMICO	1.50001e+11	
## 128043	TÉCNICO/ACADÉMICO	1.76001e+11	
## 166121	TÉCNICO/ACADÉMICO	3.15022e+11	
## 233260	TÉCNICO/ACADÉMICO	1.13001e+11	
##	COLE_NOMBRE_SEDE	COLE_SEDE_PRINCIPAL	
## 458052	COLEGIO COOPERATIVO PAULO VI LA CALERA	S	
## 194856	EFRAIN OROZCO (sede principal)	S	
## 138374	IE COL DPTAL NARCISO JOSE MATUS TORRES	S	
## 128043	INSTITUCION EDUCATIVA REPUBLICA DE ISRAEL	S	
## 166121	I.E. ENRIQUE SUAREZ - SEDE PRINCIPAL	S	
## 233260	INSTITUCION EDUCATIVA MERCEDES ABREGO	S	
##	COLE_AREA_UBICACION	COLE_JORNADA	COLE_COD_MCPIO_UBICACION
## 458052	URBANO	MAÑANA	25377
## 194856	RURAL	MAÑANA	19130

##	138374	URBANO	TARDE	50001
##	128043	URBANO	UNICA	76001
##	166121	URBANO	UNICA	15022
##	233260	URBANO	MAÑANA	13001
##	COLE_MCPIO_UBICACION COLE_COD_DEPTO_UBICACION COLE_DEPTO_UBICACION			
##	458052	LA CALERA	25	CUNDINAMARCA
##	194856	CAJIBÍO	19	CAUCA
##	138374	VILLAVICENCIO	50	META
##	128043	CALI	76	VALLE
##	166121	ALMEIDA	15	BOYACA
##	233260	CARTAGENA DE INDIAS	13	BOLIVAR
##	ESTU_PRIVADO_LIBERTAD ESTU_COD_MCPIO_PRESENTACION			
##	458052	N	25377	
##	194856	N	19001	
##	138374	N	50001	
##	128043	N	76001	
##	166121	N	15299	
##	233260	N	13001	
##	ESTU_MCPIO_PRESENTACION ESTU_DEPTO_PRESENTACION			
##	458052	LA CALERA	CUNDINAMARCA	
##	194856	POPAYÁN	CAUCA	
##	138374	VILLAVICENCIO	META	
##	128043	CALI	VALLE	
##	166121	GARAGOA	BOYACA	
##	233260	CARTAGENA DE INDIAS	BOLIVAR	
##	ESTU_COD_DEPTO_PRESENTACION PUNT_LECTURA_CRITICA			
##	458052	25	55	
##	194856	19	43	
##	138374	50	52	
##	128043	76	48	
##	166121	15	70	
##	233260	13	56	
##	PERCENTIL_LECTURA_CRITICA DESEMP_LECTURA_CRITICA PUNT_MATEMATICAS			
##	458052	61	3	66
##	194856	20	2	47
##	138374	49	3	47
##	128043	35	2	55
##	166121	97	4	60
##	233260	62	3	59
##	PERCENTIL_MATEMATICAS DESEMP_MATEMATICAS PUNT_C_NATURALES			
##	458052	91	3	55
##	194856	36	2	45
##	138374	36	2	45
##	128043	62	3	43
##	166121	78	3	56
##	233260	74	3	55
##	PERCENTIL_C_NATURALES DESEMP_C_NATURALES PUNT_SOCIALES_CIUDADANAS			
##	458052	75	2	56
##	194856	42	2	49
##	138374	39	2	40
##	128043	34	2	40
##	166121	77	3	69
##	233260	73	2	50
##	PERCENTIL_SOCIALES_CIUDADANAS DESEMP_SOCIALES_CIUDADANAS PUNT_INGLES			

```
## 458052          74          3          49
## 194856          53          2          60
## 138374          28          1          42
## 128043          28          1          52
## 166121          96          3          52
## 233260          56          2          64
##      PERCENTIL_INGLES DESEMP_INGLES PUNT_GLOBAL PERCENTIL_GLOBAL
## 458052          67          A1          287          78
## 194856          88          A2          235          43
## 138374          36          A-          228          38
## 128043          74          A1          235          42
## 166121          74          A1          314          90
## 233260          91          A2          278          73
##      ESTU_INSE_INDIVIDUAL ESTU_NSE_INDIVIDUAL ESTU_NSE_ESTABLECIMIENTO
## 458052          51.89541          3          3
## 194856          27.71282          1          2
## 138374          40.41507          1          2
## 128043          60.74755          3          3
## 166121          55.94168          3          2
## 233260          53.27249          3          2
##      ESTU_ESTADAINVESTIGACION          ESTU_GENERACION.E
## 458052          PUBLICAR          NO
## 194856          PUBLICAR GENERACION E - GRATUIDAD
## 138374          PUBLICAR GENERACION E - GRATUIDAD
## 128043          PUBLICAR          NO
## 166121          PUBLICAR GENERACION E - GRATUIDAD
## 233260          PUBLICAR GENERACION E - GRATUIDAD
```

```
summary(muestra$PUNT_MATEMATICAS)
```

```
##      Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## 23.00  43.00  51.00  51.01  59.00 100.00
```

```
table(muestra$ESTU_GENERO)
```

```
##
##      F      M
## 540 460
```

```
#Histograma de frecuencias
```

```
histograma=ggplot(muestra, aes(x = PUNT_MATEMATICAS)) +
  geom_histogram(binwidth = 10, fill = "#861627", color = "white") +
  labs(title = "Histograma de Puntaje en Matemáticas",
        x = "Puntaje en Matemáticas",
        y = "Frecuencia") +
  stat_bin(binwidth = 10, geom = "text", aes(label = after_stat(count)), vjust = -0.5, color = "black")
theme_minimal()+
ylim(c(0,350))+
geom_density(aes(y = ..count.. * 10), color = "blue", linewidth = 1) # La multiplicación ajusta la es
histograma
```

```
## Warning: The dot-dot notation (`..count..`) was deprecated in ggplot2 3.4.0.
## i Please use `after_stat(count)` instead.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
## generated.
```

