

BITÁCORA 2 - GRUPO 2

Integrantes

- Cristhian Jimenez Campos C33973
- Olman Camacho Jerez C31523
- Jose Manuel Alfaro Monge C30244

Bitácora 1

Pregunta de Investigación

¿Cuáles son los factores individuales, sociales y académicos que más inciden en la predicción de la deserción escolar en estudiantes de nivel secundaria?

Objeto de Estudio

El objeto de estudio es el fenómeno de la deserción escolar, entendido como el abandono prematuro del sistema educativo formal por parte de estudiantes, antes de completar el nivel educativo correspondiente.

Conceptos

Deserción escolar

Según Spady, citado por Floricely Dzay Chulim (2012), menciona dos definiciones operacionales acerca de la deserción universitaria: a) Incluye a cualquier persona que abandona la institución de educación superior donde se encuentra registrado y b) Se refiere a aquellas personas que reciben un título o grado de cualquier universidad. Esta segunda definición sostiene que el o la estudiante que haya empezado un proceso de aprendizaje superior, y en un cierto periodo de tiempo no ha obtenido su respectivo título o grado, se puede considerar un desertor.

Factores personales

Entendemos a los factores personales como todas aquellas características internas del estudiante como la motivación, las actitudes y las habilidades cognitivas, que influyen directamente su aprendizaje.

Factores sociales

Entendemos a los factores sociales como toda aquella influencia externa relacionada con el entorno del estudiante, que se presenta en la vida del mismo de manera inesperada, o al menos no planeada, que provenga directamente de su círculo personal (familiares, amigos cercanos, entre otros), de manera que su rendimiento académico se ve directamente afectado.

Teorías

Teoría de la Frustración

Esta teoría es un aporte del investigador Abram Amsel que, según Alejandro Baquero (2007) la Teoría de la Frustración de Abram Amsel expone la elaboración de una hipótesis acerca de la función de la omisión decepcionante de recompensa en circunstancias de recompensa no continua. De acuerdo a esa teoría, en la etapa de adquisición, el sujeto se instruye a prever la recompensa obtenida en el contexto experimental debido a la presencia de claves contextuales que las anuncian. Luego,

cuando sorprendentemente se omite la recompensa, el sujeto provoca una reacción emocional natural y aversiva denominada frustración que en la actualidad es precedida por las señales que previamente indicaban una recompensa. Esto genera un enfrentamiento al inicio del entrenamiento debido a que tanto la frustración como la recompensa son pronosticados por condicionamiento clásico debido a los mismos estímulos condicionados. A medida que avanza el entrenamiento, por efecto de un contracondicionamiento, el conflicto se resuelve a favor de contestar ya que el refuerzo no es predecible en una situación habitual de refuerzo incompleto, dado que en pruebas donde hay elementos que indican la falta de un refuerzo, se fortalece la respuesta instrumental. De esta manera, la respuesta continúa al introducir la extinción puesto que se ha condicionado a la previsión de falta de recompensa. Por otro lado, en los sujetos que reciben refuerzo constante, no existe nada que los estimule a contestar sin recibir recompensa.

Teoría de los Factores Múltiples

La teoría de factores múltiples mantiene que la deserción escolar tiene varios factores, como, individual, sociales y académico. Basándonos en la base de datos que vamos a trabajar, uno de estos factores son el desempeño académico, el nivel de educación, el factor de la familia, esto puede aumentar el riesgo de que un alumno abandone la universidad.

La deserción no se ve afectada por un único factor, generalmente se ven afectados por una interacción múltiple de factores y en ocasiones no se logran captar estas razones. Es importante detectar la deserción académica a tiempo para poder ayudar al estudiante con tutorías, apoyo económico y programas de acompañamiento psicológico, las cuales pueden ayudar para que el estudiante no deserte. Además de identificar estos factores o cuál factor tiene más peso para la deserción del estudiante.

Teoría de los Factores Económicos y Apoyo Familiar en la Deserción Universitaria

Esta teoría viene de varios documentos, en el cual se menciona que las dificultades económicas relacionadas con la estabilidad familia o individual, llegan a tener un impacto importante en los estudiantes haciendo que la probabilidad de que un estudiante esté en peligro de abandono de la universidad sea mayor. Cuando se tiene problemas económicos tanto individuales como en el grupo familiar, esto llega a significar una gran presión a los estudiantes, haciendo que traten de apoyar a sus familiares, generando que dejen de lado la universidad.

Para llegar a liberar presión sobre los estudiantes, el acceso a becas y ayudas financieras se identifican como un factor que ayuda a los estudiantes para evitar que estén en peligro de abandonar la universidad, dándoles un respiro y apoyo, teniendo un impacto en el rendimiento.

Bibliografía

https://admissionfen.cl/wp-content/uploads/2024/03/Revista-N%C2%B024_jaime_miranda.pdf

http://www.scielo.org.bo/scielo.php?script=sci_arttext&pid=S2521-27372019000200007

https://www.scielo.sa.cr/scielo.php?script=sci_arttext&pid=S2215-34702021000100019

http://www.scielo.edu.uy/scielo.php?script=sci_arttext&pid=S1688-93042023000301206#B56

https://admissionfen.cl/wp-content/uploads/2024/03/Revista-N%C2%B024_jaime_miranda.pdf

http://www.scielo.org.co/scielo.php?script=sci_arttext&pid=S2145-94442014000200010

<https://revie.gob.do/index.php/revie/article/view/193/365>

<https://www.ride.org.mx/index.php/RIDE/article/view/1923/5020>

Dzay Chulim, F., Narváez Trejo, O. M., Universidad de Quintana Roo, & Universidad Veracruzana. (2012). *La deserción escolar desde la perspectiva estudiantil* [Book]. La Editorial Manda. <https://www.uv.mx/personal/onarvaez/files/2013/02/la-desercion-escolar.pdf>

Bitácora 2

Datos

Características de la tabla:

Esta base de datos contiene registros de 4424 estudiantes, quienes serán clasificados de distintas maneras, desde estado civil hasta los cursos que están cursando, entre otros. La base de datos fue publicada en 2021 y presenta diversos factores, incluyendo variables relacionadas con los padres, para analizar si influyen en la vida académica del estudiante. Fue creada por Valentim Realinho, Mónica Vieira Martins, Jorge Machado y Luís Baptista, investigadores del Instituto Politécnico de Portalegre en Portugal, y descargada desde el enlace [link](#). Los datos corresponden al segundo semestre, aunque no se especifica el año.

Las variables están distribuidas en distintas categorías: variables relacionadas con la trayectoria académica, variables demográficas y variables socioeconómicas. Los tipos de datos incluyen variables reales, categóricas y enteras.

Poblacion de estudio:

Estudiantes matriculados en diferentes carreras de pregrado de una institucion de educacion superior.

Muestra observada:

4,424 estudiantes.

Unidad estadística o individuos:

Cada uno de los 4,424 estudiantes de educación superior durante determinados semestres.

Identificación de las variables de estudio:

Las variables de estudio incluyen información sobre la trayectoria académica, datos demográficos y factores socioeconómicos de los estudiantes, así como su rendimiento académico al final del primer y segundo semestre. El problema se plantea como una tarea de clasificación en tres categorías: abandono, matriculado y graduado.

Primeas 5 filas de la tabla de datos:

```
library(dplyr)
library(ggplot2)
datos <- read.csv2("data.csv", sep = ";", header = TRUE, stringsAsFactors = FALSE)
head(datos, 5)
```

	Marital.status	Application.mode	Application.order	Course
1	1	17	5	171
2	1	15	1	9254
3	1	1	5	9070
4	1	17	2	9773
5	2	39	1	8014

	Daytime.evening.attendance.	Previous.qualification
1	1	1
2	1	1
3	1	1
4	1	1
5	0	1

	Previous.qualification..grade.	Nacionality	Mother.s.qualification
1	122.0	1	19
2	160.0	1	1
3	122.0	1	37

4	122.0	1	38
5	100.0	1	37

	Father.s.qualification	Mother.s.occupation	Father.s.occupation
1	12	5	9
2	3	3	3
3	37	9	9
4	37	5	3
5	38	9	9

	Admission.grade	Displaced	Educational.special.needs	Debtor
1	127.3	1	0	0
2	142.5	1	0	0
3	124.8	1	0	0
4	119.6	1	0	0
5	141.5	0	0	0

	Tuition.fees.up.to.date	Gender	Scholarship.holder	Age.at.enrollment
1	1	1	0	20
2	0	1	0	19
3	0	1	0	19
4	1	0	0	20
5	1	0	0	45

	International	Curricular.units.1st.sem..credited.
1	0	0
2	0	0
3	0	0
4	0	0
5	0	0

	Curricular.units.1st.sem..enrolled.	Curricular.units.1st.sem..evaluations.
1	0	0
2	6	6
3	6	0

4	6	8
5	6	9

Curricular.units.1st.sem..approved. Curricular.units.1st.sem..grade.

1	0	0.0
2	6	14.0
3	0	0.0
4	6	13.428571428571429
5	5	12.333333333333334

Curricular.units.1st.sem..without.evaluations.

1	0
2	0
3	0
4	0
5	0

Curricular.units.2nd.sem..credited. Curricular.units.2nd.sem..enrolled.

1	0	0
2	0	6
3	0	6
4	0	6
5	0	6

Curricular.units.2nd.sem..evaluations. Curricular.units.2nd.sem..approved.

1	0	0
2	6	6
3	0	0
4	10	5
5	6	6

Curricular.units.2nd.sem..grade.

1	0.0
2	13.666666666666666
3	0.0

4	12.4		
5	13.0		
Curricular.units.2nd.sem...without.evaluations. Unemployment.rate			
1	0	10.8	
2	0	13.9	
3	0	10.8	
4	0	9.4	
5	0	13.9	
Inflation.rate GDP Target			
1	1.4	1.74	Dropout
2	-0.3	0.79	Graduate
3	1.4	1.74	Dropout
4	-0.8	-3.12	Graduate
5	-0.3	0.79	Graduate

La tabla se encuentra en formato tabular, esto se puede ver y tambien se comenta en la pagina de descarga

Resumen de 5 números de las variables cuantitativas y analizar el mismo:

```
library(dplyr)
# Se selecciona las variables cuantitativas
variables_cuantitativas <- select_if(datos, is.numeric)
# Calcular resumen de 5 números para cada variable
#'Vamos a usar sapply para aplicar la funcion fivenum a la base
#'el firenum es una funcion que nos ayuda a calcular el minimo y maximo, los Q1 y Q3, ad
resumen_5_numeros <- sapply(variables_cuantitativas, fivenum)
#para no tener problema trasponemos a resumen_5_numeros
```

```

resumen_5_numeros <- t(resumen_5_numeros)
#'Para facilitar la lectura vamos a ponerle nombres claros a las columnas
colnames(resumen_5_numeros) <- c("Minimo","Q1","Mediana","Q3","Máximo")
print(resumen_5_numeros)

```

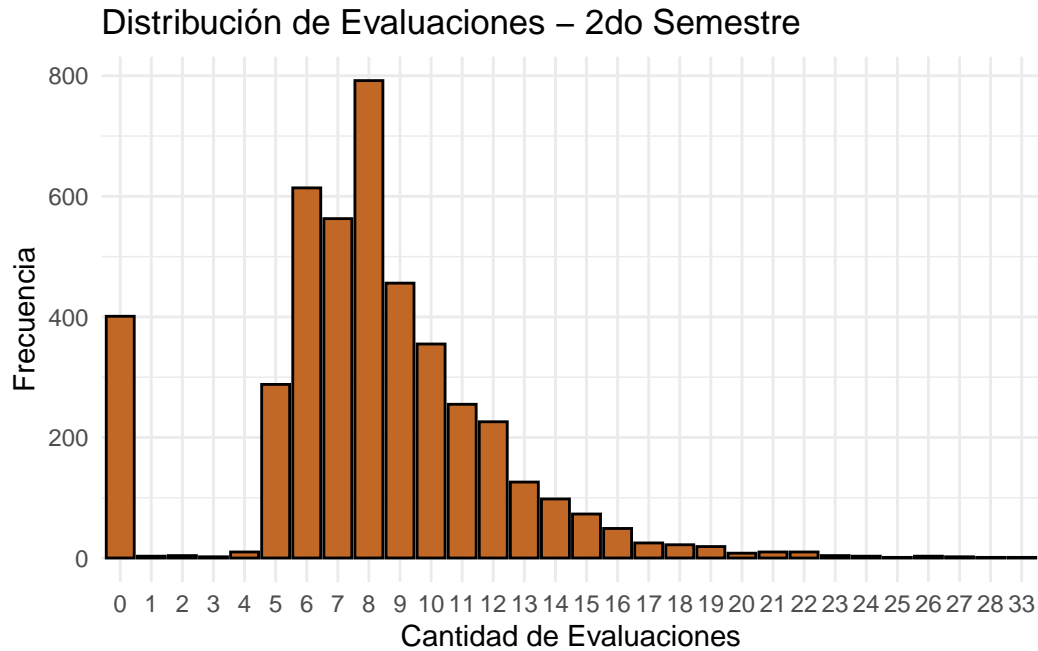
	Minimo	Q1	Mediana	Q3	Máximo
Marital.status	1	1	1	1	6
Application.mode	1	1	17	39	57
Application.order	0	1	1	2	9
Course	33	9085	9238	9556	9991
Daytime.evening.attendance.	0	1	1	1	1
Previous.qualification	1	1	1	1	43
Nacionality	1	1	1	1	109
Mother.s.qualification	1	2	19	37	44
Father.s.qualification	1	3	19	37	44
Mother.s.occupation	0	4	5	9	194
Father.s.occupation	0	4	7	9	195
Displaced	0	0	1	1	1
Educational.special.needs	0	0	0	0	1
Debtor	0	0	0	0	1
Tuition.fees.up.to.date	0	1	1	1	1
Gender	0	0	0	1	1
Scholarship.holder	0	0	0	0	1
Age.at.enrollment	17	19	20	25	70
International	0	0	0	0	1
Curricular.units.1st.sem..credited.	0	0	0	0	20
Curricular.units.1st.sem..enrolled.	0	5	6	7	26
Curricular.units.1st.sem..evaluations.	0	6	8	10	45
Curricular.units.1st.sem..approved.	0	3	5	6	26
Curricular.units.1st.sem..without.evaluations.	0	0	0	0	12

Curricular.units.2nd.sem..credited.	0	0	0	0	19
Curricular.units.2nd.sem..enrolled.	0	5	6	7	23
Curricular.units.2nd.sem..evaluations.	0	6	8	10	33
Curricular.units.2nd.sem..approved.	0	2	5	6	20
Curricular.units.2nd.sem..without.evaluations.	0	0	0	0	12

La tabla muestra el resumen de 5 numeros de las variables cuantitativas de nuestra base de datos. en la tabla se puede ver que la mayoria de las variables tienen un minimo de en 0, la forma en la que se evualua nuestra variables es de una manera difereente ya que estas van con diferentes rangos par asignarles puede que sea continio o que se llegue a saltar numeros, pasa de 33 o 53 y cosas así.

Hacer al menos un gráfico que describa la distribución para cada una de las variables cuantitativas:

```
ggplot(datos, aes(x = factor(Curricular.units.2nd.sem..evaluations.))) +
  geom_bar(fill = "#C26827", color = "#000000") +
  labs(
    title = "Distribución de Evaluaciones - 2do Semestre",
    x = "Cantidad de Evaluaciones",
    y = "Frecuencia"
  ) +
  theme_minimal()
```



Hacer al menos dos gráficos que describan la relación entre las variables:

```
grafico_internacional_genero <- function(datos, genero, internacional, pais) {
  df_plot <- datos %>%
    mutate(
      genero = factor({{genero}}, levels = c(1, 0), labels = c("Hombres", "Mujeres")),
      tipo = factor({{internacional}}, levels = c(0, 1), labels = c("Nacional", "Internacional"))
    ) %>%
    group_by(genero, tipo) %>%
    summarise(Frecuencia = n(), .groups = "drop")

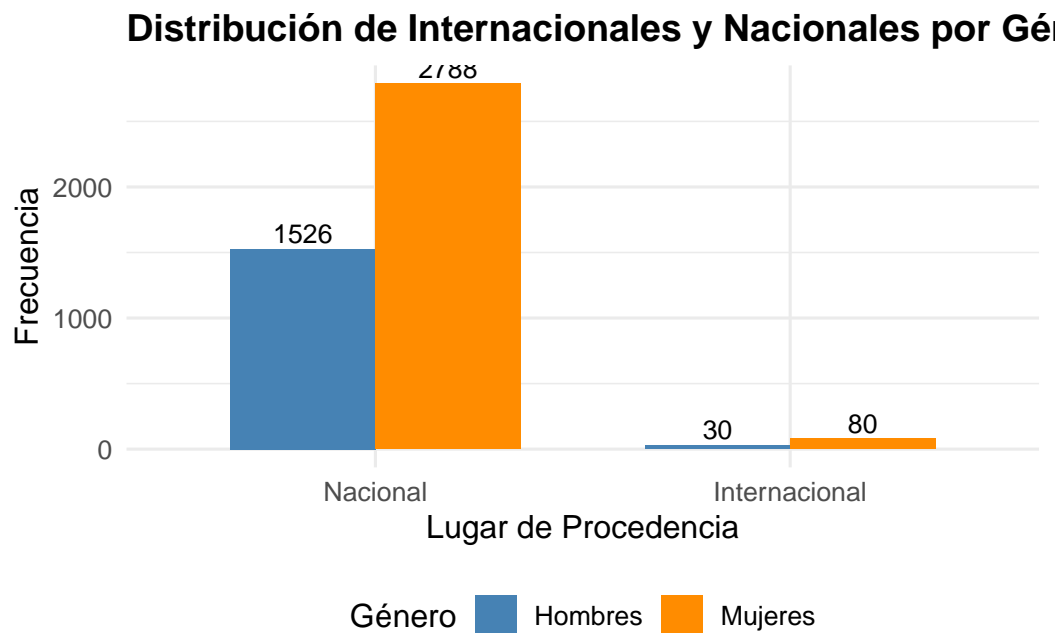
  ggplot(df_plot, aes(x = tipo, y = Frecuencia, fill = genero)) +
    geom_col(position = "dodge", width = 0.7) +
    geom_text(aes(label = Frecuencia),
              position = position_dodge(width = 0.7),
```

```

      vjust = -0.3, size = 3.5) +
scale_fill_manual(values = c("steelblue", "darkorange")) +
labs(
  title = "Distribución de Internacionales y Nacionales por Género",
  x = "Lugar de Procedencia",
  y = "Frecuencia",
  fill = "Género"
) +
theme_minimal(base_size = 12) +
theme(
  legend.position = "bottom",
  plot.title = element_text(face = "bold"),
  axis.text.x = element_text(angle = 0, hjust = 0.5)
)
}

grafico_internacional_genero(datos, datos$Gender, datos$International, datos$Nationality

```



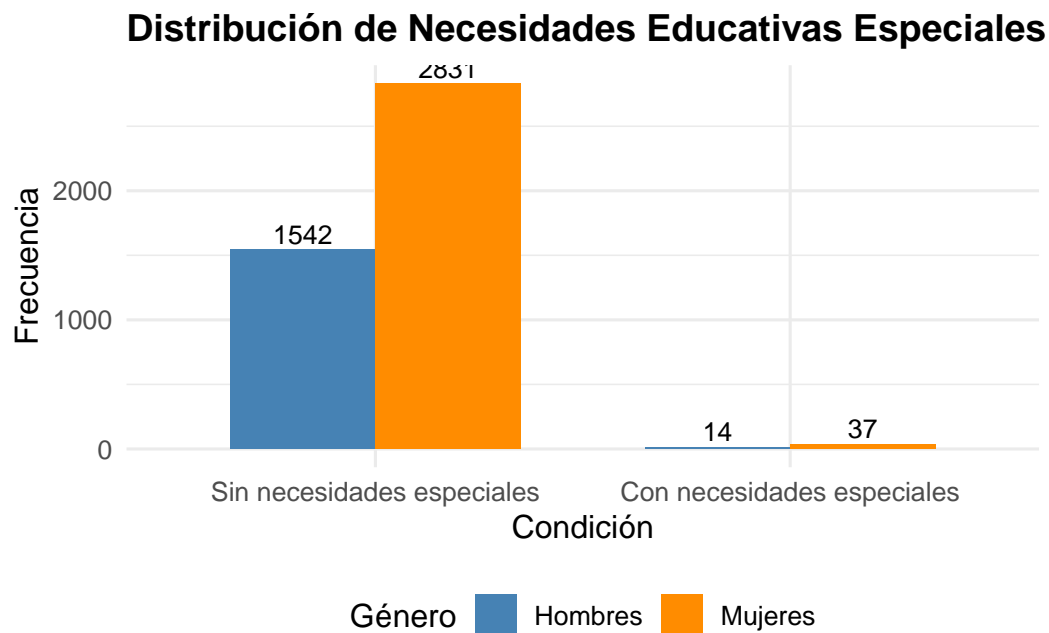
```

gráfico_necesidad_genero <- function(df, genero, needs){
  df_plot <- datos %>%
    mutate(
      genero = factor({{genero}}, levels = c(1, 0), labels = c("Hombres", "Mujeres")),
      necesidades = factor({{needs}}, levels = c(0, 1),
        labels = c("Sin necesidades especiales", "Con necesidades especiales"))
    ) %>%
    group_by(genero, necesidades) %>%
    summarise(Frecuencia = n(), .groups = "drop")

  ggplot(df_plot, aes(x = necesidades, y = Frecuencia, fill = genero)) +
    geom_col(position = "dodge", width = 0.7) +
    geom_text(aes(label = Frecuencia),
      position = position_dodge(width = 0.7),
      vjust = -0.3, size = 3.5) +
    scale_fill_manual(values = c("steelblue", "darkorange")) +
    labs(
      title = "Distribución de Necesidades Educativas Especiales por Género",
      x = "Condición",
      y = "Frecuencia",
      fill = "Género"
    ) +
    theme_minimal(base_size = 12) +
    theme(
      legend.position = "bottom",
      plot.title = element_text(face = "bold"),
      axis.text.x = element_text(angle = 0, hjust = 0.5)
    )
}

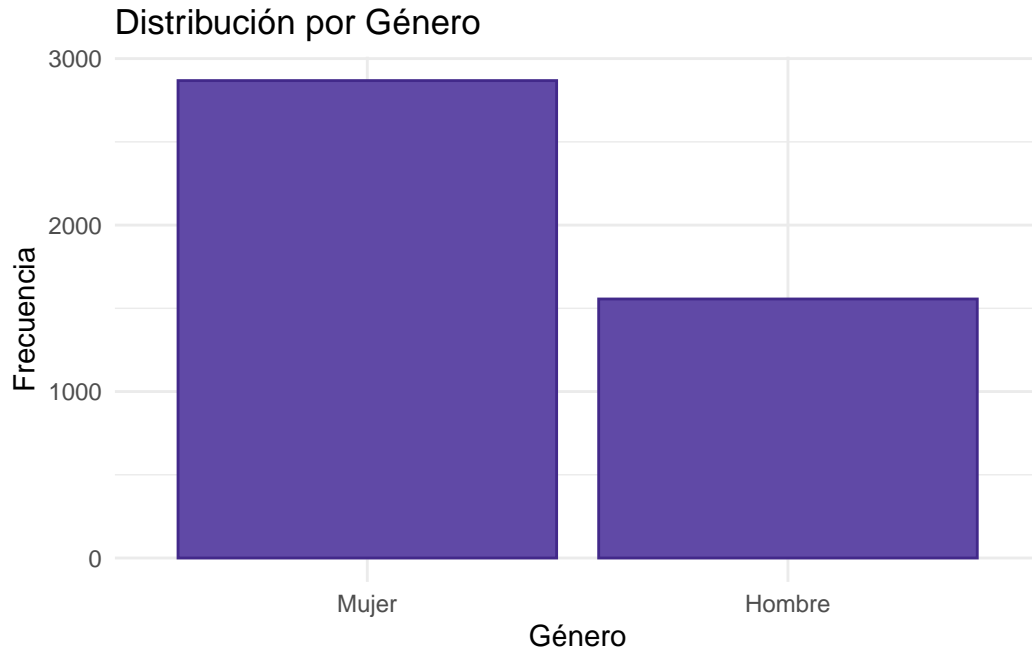
gráfico_necesidad_genero(datos, datos$Gender, datos$Educational.special.needs)

```



Hacer al menos un gráfico que muestre la distribución de las variables categóricas:

```
ggplot(datos, aes(x = factor(Gender,
                             levels = c(0, 1),
                             labels = c("Mujer", "Hombre")))) +
  geom_bar(fill = "#6049A6", color = "#42298A") +
  labs(
    title = "Distribución por Género",
    x = "Género",
    y = "Frecuencia"
  ) +
  theme_minimal()
```



Identificar valores faltantes y posibles outliers:

```
datos_faltantes <- datos %>%  
  filter(if_any(everything(), is.na))  
head(datos_faltantes)
```

```
[1] Marital.status  
[2] Application.mode  
[3] Application.order  
[4] Course  
[5] Daytime.evening.attendance.  
[6] Previous.qualification  
[7] Previous.qualification..grade.  
[8] Nationality  
[9] Mother.s.qualification  
[10] Father.s.qualification
```



```

[11] Mother.s.occupation
[12] Father.s.occupation
[13] Admission.grade
[14] Displaced
[15] Educational.special.needs
[16] Debtor
[17] Tuition.fees.up.to.date
[18] Gender
[19] Scholarship.holder
[20] Age.at.enrollment
[21] International
[22] Curricular.units.1st.sem..credited.
[23] Curricular.units.1st.sem..enrolled.
[24] Curricular.units.1st.sem..evaluations.
[25] Curricular.units.1st.sem..approved.
[26] Curricular.units.1st.sem..grade.
[27] Curricular.units.1st.sem..without.evaluations.
[28] Curricular.units.2nd.sem..credited.
[29] Curricular.units.2nd.sem..enrolled.
[30] Curricular.units.2nd.sem..evaluations.
[31] Curricular.units.2nd.sem..approved.
[32] Curricular.units.2nd.sem..grade.
[33] Curricular.units.2nd.sem..without.evaluations.
[34] Unemployment.rate
[35] Inflation.rate
[36] GDP
[37] Target
<0 rows> (o 0- extensión row.names)

```

```

datos %>%
  summarise(
    across(
      where(is.numeric),
      ~sum(
        .<quantile(.,0.25,na.rm=TRUE)-1.5*IQR(.)|
        .>quantile(.,0.75,na.rm=TRUE)+1.5*IQR(.),na.rm = TRUE
      )
    )
  )
)#cantidad de outliers por variable

```

	Marital.status	Application.mode	Application.order	Course
1	505	0	541	442
	Daytime.evening.attendance.	Previous.qualification	Nacionality	
1	483	707	110	
	Mother.s.qualification	Father.s.qualification	Mother.s.occupation	
1	0	0	182	
	Father.s.occupation	Displaced	Educational.special.needs	Debtor
1	177	0	51	503
	Tuition.fees.up.to.date	Gender	Scholarship.holder	Age.at.enrollment
1	528	0	1099	441
	International	Curricular.units.1st.sem..credited.		
1	110	577		
	Curricular.units.1st.sem..enrolled.	Curricular.units.1st.sem..evaluations.		
1	424	158		
	Curricular.units.1st.sem..approved.			
1	180			
	Curricular.units.1st.sem..without.evaluations.			
1	294			

	Curricular.units.2nd.sem..credited.	Curricular.units.2nd.sem..enrolled.
1	530	369
	Curricular.units.2nd.sem..evaluations.	Curricular.units.2nd.sem..approved.
1	109	44
	Curricular.units.2nd.sem..without.evaluations.	
1	282	

```

es_outlier <- function(x) {

  if(!is.numeric(x)) return(rep(FALSE,length(x)))
  q1 <- quantile(x,0.25,na.rm = TRUE)
  q3 <- quantile(x,0.75,na.rm = TRUE)

  resultado <- x<(q1-1.5*IQR(x,na.rm = TRUE))|
              x>(q3+1.5*IQR(x,na.rm = TRUE))
  return(resultado)
}

outliers <- as.data.frame(sapply(datos,es_outlier))
head(outliers)

```

	Marital.status	Application.mode	Application.order	Course
1	FALSE	FALSE	TRUE	TRUE
2	FALSE	FALSE	FALSE	FALSE
3	FALSE	FALSE	TRUE	FALSE
4	FALSE	FALSE	FALSE	FALSE
5	TRUE	FALSE	FALSE	TRUE
6	TRUE	FALSE	FALSE	FALSE
	Daytime.evening.attendance.	Previous.qualification		
1	FALSE	FALSE		

2	FALSE	FALSE
3	FALSE	FALSE
4	FALSE	FALSE
5	TRUE	FALSE
6	TRUE	TRUE

Previous.qualification..grade. Nacionality Mother.s.qualification

1	FALSE	FALSE	FALSE
2	FALSE	FALSE	FALSE
3	FALSE	FALSE	FALSE
4	FALSE	FALSE	FALSE
5	FALSE	FALSE	FALSE
6	FALSE	FALSE	FALSE

Father.s.qualification Mother.s.occupation Father.s.occupation

1	FALSE	FALSE	FALSE
2	FALSE	FALSE	FALSE
3	FALSE	FALSE	FALSE
4	FALSE	FALSE	FALSE
5	FALSE	FALSE	FALSE
6	FALSE	FALSE	FALSE

Admission.grade Displaced Educational.special.needs Debtor

1	FALSE	FALSE	FALSE	FALSE
2	FALSE	FALSE	FALSE	FALSE
3	FALSE	FALSE	FALSE	FALSE
4	FALSE	FALSE	FALSE	FALSE
5	FALSE	FALSE	FALSE	FALSE
6	FALSE	FALSE	FALSE	TRUE

Tuition.fees.up.to.date Gender Scholarship.holder Age.at.enrollment

1	FALSE	FALSE	FALSE	FALSE
2	TRUE	FALSE	FALSE	FALSE
3	TRUE	FALSE	FALSE	FALSE

4	FALSE	FALSE	FALSE	FALSE
5	FALSE	FALSE	FALSE	TRUE
6	FALSE	FALSE	FALSE	TRUE

International Curricular.units.1st.sem..credited.

1	FALSE	FALSE
2	FALSE	FALSE
3	FALSE	FALSE
4	FALSE	FALSE
5	FALSE	FALSE
6	FALSE	FALSE

Curricular.units.1st.sem..enrolled. Curricular.units.1st.sem..evaluations.

1	TRUE	FALSE
2	FALSE	FALSE
3	FALSE	FALSE
4	FALSE	FALSE
5	FALSE	FALSE
6	FALSE	FALSE

Curricular.units.1st.sem..approved. Curricular.units.1st.sem..grade.

1	FALSE	FALSE
2	FALSE	FALSE
3	FALSE	FALSE
4	FALSE	FALSE
5	FALSE	FALSE
6	FALSE	FALSE

Curricular.units.1st.sem..without.evaluations.

1	FALSE
2	FALSE
3	FALSE
4	FALSE
5	FALSE

6	FALSE	
	Curricular.units.2nd.sem..credited.	Curricular.units.2nd.sem..enrolled.
1	FALSE	TRUE
2	FALSE	FALSE
3	FALSE	FALSE
4	FALSE	FALSE
5	FALSE	FALSE
6	FALSE	FALSE
	Curricular.units.2nd.sem..evaluations.	Curricular.units.2nd.sem..approved.
1	FALSE	FALSE
2	FALSE	FALSE
3	FALSE	FALSE
4	FALSE	FALSE
5	FALSE	FALSE
6	TRUE	FALSE
	Curricular.units.2nd.sem..grade.	
1	FALSE	
2	FALSE	
3	FALSE	
4	FALSE	
5	FALSE	
6	FALSE	
	Curricular.units.2nd.sem..without.evaluations.	Unemployment.rate
1	FALSE	FALSE
2	FALSE	FALSE
3	FALSE	FALSE
4	FALSE	FALSE
5	FALSE	FALSE
6	TRUE	FALSE
	Inflation.rate	GDP Target

1	FALSE	FALSE	FALSE
2	FALSE	FALSE	FALSE
3	FALSE	FALSE	FALSE
4	FALSE	FALSE	FALSE
5	FALSE	FALSE	FALSE
6	FALSE	FALSE	FALSE

Investigar técnicas que permitan subsanar los valores perdidos y outliers:

Manejo de valores faltantes y outliers

Valores faltantes:

Los valores faltantes pueden ser tratados mediante diferentes métodos que modifican los datos de distinta manera. Por eso, a la hora de utilizar estos valores, sus resultados se verán afectados de mayor o menor manera forma dependiendo de la técnica utilizada. Existen técnicas como la eliminación de los datos faltantes, sin embargo, esta práctica puede perjudicar al análisis de los datos, por lo que nos centraremos en los métodos que buscan modificarlos de manera que aporten información. Se pueden utilizar métodos simples como la imputación por la media, mediana o moda, que son métodos fáciles de implementar, sin embargo, estos datos pueden verse distorsionados por la presencia de valores atípicos. Por eso existen métodos más robustos como la imputación por vecinos más cercanos, imputación por regresión, imputación múltiple o por series de tiempo. Estas técnicas conservan de mejor manera la estructura de los datos.

Outliers:

Hay tres opciones posibles a la hora de enfrentar un outlier, tanto univariable como multivariable, estas opciones son eliminarlo, mantenerlo o reemplazarlo. Al igual que en los valores faltantes, se pueden eliminar todos los outliers, pero este método provoca pérdida de información que es normalmente valiosa y aporta a la investigación. También se tiene la opción de utilizar métodos simples como reemplazarlos por la media o moda, aunque el uso de estas técnicas no es recomendado porque

pueden generar resultados artificiales y sesgados. Por lo tanto, se recomienda utilizar técnicas más robustas que reduzcan la influencia y no los eliminen por completo.

Ahora veremos algunas formas de enfrentar los outliers dependiendo de que sean univariantes o multivariantes.

Outliers univariantes

Detectar y clasificar: Se realiza una revisión manual para determinar si son errores, casos aleatorios o datos de interés que se deberían analizar por separado. Luego se decide individualmente qué hacer con cada uno.

Winsorización: Se sustituye el dato por el valor del percentil más cercano. Esto conserva la estructura de los datos y limita los efectos sobre las estadísticas principales.

Outliers multivariantes

Revisión contextual: Se verifica que las variables que se estén comparando no tengan una relación incoherente o anómala que genere valores extraños o fuera de contexto.

Análisis por separado: Se analizan los casos atípicos como un conjunto separado a los datos comunes para identificar patrones y no alterar los datos principales.

Regresión: Son métodos que consisten en modelar la relación de una variable dependiente con otras independientes, para evaluar la influencia de los outliers sobre las relaciones entre variables.

Bibliografía:

<https://www.maximaformacion.es/blog-dat/como-describir-tus-datos-en-r-paso-1/>

https://rpubs.com/Elyn1017/Aunivariado_Vcuantitativas_CasoMedicos

<https://www.uca.edu.sv/mpe/wp-content/uploads/2020/09/61.-Hernandez-W-y-Montano-Y.-2020-Analisis-de-la-desercion-escolar.pdf>