Taller Análisis Exploratorio Estadístico del ICFES

Minería de Datos

Jonatan Camilo Igua Contreras
ID: 808919 | NRC: 73466

2025-04-20

Análisis Exploratorio resultados de la prueba saber 11

En el siguiente taller se realizo un análisis estadístico completo de los resultados del ICFES 2012, en donde se limpiaron los datos con errores tipográficos o con valores faltantes, también se organizaron y modificaron algunos nombres de las variables para facilitar los análisis. Los análisis se realizaron para comprender el desempeño de los estudiantes durante la prueba, en donde se tomaron en cuenta variables como la edad, colegios de estudio y las encuestas socioeconomicas, todo este análisis se realizo con el fin de comprender si a la falta de acceso a Internet, contar con herramientas tecnológicas o tener buenos puntajes en ciertas materias pudieron haber generado un ventaja a la hora de presentar el examen.

1. Carga y exploración de los datos

Codificación dataset de ICFES:
Para poder importar los datos del dataset del ICFES primero se tiene que mirar que tipo de codificación tiene el archivo .csv para saber como importarlo. Para esto se usa la función guess_encoding() del paquete readr, permite detectar la codificación de caracteres de un archivo CSV.

```
library(knitr)
library(dplyr)
library(readr)
guess_encoding("SB11-20121-RGSTRO-CLFCCN-V1-0-txt-csv.csv") %>%
```

```
  kable(col.names = c("Codificación", "Confianza"),
  caption = "Codificaciones detectadas por `guess_encoding`")
```

Table 1: Codificaciones detectadas por `guess_encoding`

Codificación	Confianza
UTF-8	1.00
windows-1252	0.48
windows-1250	0.27

Como se puede apreciar se obtuvo las codificaciones sugeridas, la confidence indica la confianza del tipo de codificación en este caso el valor es 1.00 el cual es la máxima confianza, por ende el archivo se debe leer con la codificación UTF-8

Importar dataset de ICFES:
Para importar los datos del archivo csv al proyecto de R se uso read_csv del paquete readr, para leer los datos y convertirlos en un data frame, se declaro que el tipo de codificación de los caracteres es UTF-8.

```
library(readr)
datos_icfes <- read_csv("SB11-20121-RGSTRO-CLFCCN-V1-0-txt-csv.csv",
locale = locale(encoding = "UTF-8"), show_col_types = FALSE)
```

Después de visualizar los primeros datos importados en R se obtuvo un error Delimiter: "," indicando que si bien los datos se importaron bien algunos caracteres se importaron con un formato erróneo por ese motivo se creo una segunda variable con el delimitador sep y se elimino el error.Para los análisis se tomo datos_icfes_delimiter para evitar errores con los datos.

```
datos_icfes_delimiter <- read.csv("SB11-20121-RGSTRO-CLFCCN-V1-0-txt-csv.csv",
sep = "," , encoding = "UTF-8")
```

¿Cuantas observaciones y variables hay?
Para poder ver el numero de observaciones las filas de los datos se usa la función nrow().

Table 2: Cantidad total de filas en el dataset

Número de filas
45390

Para poder ver el numero de variables las columnas se usa la función ncol().

Table 3: Cantidad total de columnas en el dataset

Número de columnas
112

¿Que tipo de variables predominan: Numéricas o categóricas?
Para poder determinar que tipos de variables predominan en el conjunto de datos de las notas del ICFES, se uso la función str() para poder ver la estructura del data frame, esto permite ver el tipo de cada variable y su contenido. Se uso Kable para poder organizar los datos en formato de tabla.

```
#Primeros 20 elementos de los datos
estructura <- capture.output(str(datos_icfes_delimiter[ , 1:20]))
#Se guarda en la variable
estructura_df <- data.frame(estructura)

# Mostrar los datos
estructura_df %>%
  kable(col.names = c("Estructura de Datos"),
  caption = "Estructura del Dataset (primeras 20 columnas)")
```

Table 4: Estructura del Dataset (primeras 20 columnas)

Estructura de Datos
‘data.frame’: 45390 obs. of 20 variables:
$ PERIODO : int 20121 20121 20121 20121 20121 20121 20121 20121 20121 20121 . . .
$ ESTU_TIPO_DOCUMENTO : chr “C” “C” “R” “C” . . .
$ ESTU_PAIS_RESIDE : chr “CO” “CO” “CO” “CO” . . .
$ ESTU_GENERO : chr “M” “M” “F” “M” . . .

Estructura de Datos
$ ESTU_NACIMIENTO_DIA : int 25 5 3 5 9 21 30 22 29 21 ...
$ ESTU_NACIMIENTO_MES : int 8 7 1 9 5 10 10 4 10 12 ...
$ ESTU_NACIMIENTO_ANNO : int 1992 1970 1994 1974 1972 1990 1984 1986 1994 1987 ...
$ ESTU_LIMITA_BAJAVISION : logi NA NA NA NA NA NA ...
$ ESTU_LIMITA_SORDOCEGUERA : chr " " " " " " " " ...
$ ESTU_LIMITA_COGNITIVA : chr " " " " " " " " ...
$ ESTU_LIMITA_INVIDENTE : chr " " " " " " " " ...
$ ESTU_LIMITA_MOTRIZ : chr " " " " " " " " ...
$ ESTU_LIMITA_SORDOINTERPRETE : chr " " " " " " " " ...
$ ESTU_LIMITA_SORDONOINTERPRETE: chr " " " " " " " " ...
$ ESTU_ETNIA : int NA NA 1 NA NA NA NA NA NA NA ...
$ ESTU_COD_RESIDE_MCPIO : int 5837 5837 5837 5837 5837 23001 5837 5837 5837 5837 ...
$ ESTU_RESIDE_MCPIO : chr "TURBO" "TURBO" "TURBO" "TURBO" ...
$ ESTU_RESIDE_DEPTO : chr "ANTIOQUIA" "ANTIOQUIA" "ANTIOQUIA" "ANTIOQUIA" ...
$ ESTU_ZONA_RESIDE : int 10 10 10 10 10 10 10 10 10 10 ...
$ ESTU_AREA_RESIDE : int 1 1 1 1 1 1 1 2 1 1 ...

Utilizando la función str() sin limitar la cantidad de datos a mostrar se obtuvieron 45390 filas y 112 columnas .

Ver nombres y tipo de datos de las variables

Con la función sapply podemos visualizar el nombre de las variables y su tipo de dato, se creo una tabla con dos columnas una con los nombres de las variables y otra con su tipo de dato para poder visualizar mejor los datos.

```
library(knitr)

tabla_tipos <- tibble::tibble(
  Variable = names(datos_icfes_delimiter),
  Tipo = sapply(datos_icfes_delimiter, class)
)

kable(tabla_tipos, caption = "Tipos de Variables en el DataFrame")
```

Table 5: Tipos de Variables en el DataFrame

Variable	Tipo
PERIODO	integer
ESTU_TIPO_DOCUMENTO	character
ESTU_PAIS_RESIDE	character
ESTU_GENERO	character
ESTU_NACIMIENTO_DIA	integer
ESTU_NACIMIENTO_MES	integer
ESTU_NACIMIENTO_ANNO	integer
ESTU_LIMITA_BAJAVISION	logical
ESTU_LIMITA_SORDOCEGUERA	character
ESTU_LIMITA_COGNITIVA	character
ESTU_LIMITA_INVIDENTE	character
ESTU_LIMITA_MOTRIZ	character
ESTU_LIMITA_SORDOINTERPRETE	character
ESTU_LIMITA_SORDONOINTERPRETE	character
ESTU_ETNIA	integer
ESTU_COD_RESIDE_MCPIO	integer
ESTU_RESIDE_MCPIO	character
ESTU_RESIDE_DEPTO	character
ESTU_ZONA_RESIDE	integer
ESTU_AREA_RESIDE	integer
IND_COD_ICFES_TERMINO	integer
COLE_COD_ICFES	integer
COLE_COD_DANE_INSTITUCION	numeric
COLE_NOMBRE_SEDE	character
COLE_CALENDARIO	character
COLE_GENERO	character
COLE_NATURALEZA	character
COLE_BILINGUE	character
COLE_JORNADA	character
COLE_CARACTER	character
COLE_VALOR_PENSION	character
ESTU_VECES_ESTADO	character
ESTU_CARRDESEADA_TIPO	integer
ESTU_IES_COD_DESEADA	integer
ESTU_IES_COD_MPIO_DESEADA	integer
ESTU_IES_DEPT_DESEADA	character
ESTU_IES_DESEADA_NOMBRE	character

Variable	Tipo
ESTU_IES_MPIO_DESEADA	character
ESTU_TOTAL_ALUMNOS_CURSO	character
ESTU_ANO_MATRICULA_PRIMERO	integer
ESTU_ANO_TERMINO_QUINTO	integer
ESTU_ANOS_COLEGIO_ACTUAL	integer
ESTU_ANO_MATRICULA_SEXTO	integer
ESTU_ANOS_PREESCOLAR	integer
ESTU_CUANTOS_COLE_ESTUDIO	integer
ESTU_REPROBO_CUARTO	integer
ESTU_REPROBO_DECIMO	integer
ESTU_REPROBO_NOVENO	integer
ESTU_REPROBO_OCTAVO	integer
ESTU_REPROBO_PRIMERO	integer
ESTU_REPROBO_QUINTO	integer
ESTU_REPROBO_SEGUNDO	integer
ESTU_REPROBO_SEPTIMO	integer
ESTU_REPROBO_SEXTO	integer
ESTU_REPROBO_TERCERO	integer
ESTU_REPROBO_ONCE_MAS	integer
ESTU_POR_MEJORARPOSICIONSOCIAL	character
ESTU_POR_COLOMBIAAPRENDE	character
ESTU_POR_INFLUENCIAALGUIEN	character
ESTU_POR_INTERESPERSONAL	character
ESTU_POR_BUSCANDOCARRERA	character
ESTU_POR_TRADICIONFAMILIAR	character
ESTU_POR_ORIENTACIONVOCACIONAL	character
ESTU_RAZON_RETIRO	integer
ESTU_POR_AMIGOSESTUDIANDO	integer
ESTU_POR_COSTOMATRICULA	character
ESTU_POR_OPORTUNIDADES	character
ESTU_POR_OTRARAZON	character
ESTU_PRESTIGIOINSTITUCION	character
ESTU_POR_UBICACION	integer
ESTU_POR_UNICAQUEOFRECE	integer
ESTU_RETIRARSE_COLEGIO	character
ESTU_COD_MCPIO_PRESENTACION	character
ESTU_MCPIO_PRESENTACION	character
ESTU_DEPTO_PRESENTACION	character
ESTU_EXAM_NOMBREEXAMEN	character

Variable	Tipo
FAMI_EDUCA_PADRE	character
FAMI_EDUCA_MADRE	character
FAMI_OCUPA_PADRE	integer
FAMI_OCUPA_MADRE	integer
FAMI_ESTRATO_VIVIENDA	integer
FAMI_NIVEL_SISBEN	integer
FAMI_PERSONAS_HOGAR	integer
FAMI_CUARTOS_HOGAR	integer
FAMI_PISOSHOGAR	integer
FAMI_TELEFONO_FIJO	integer
FAMI_CELULAR	integer
FAMI_INTERNET	integer
FAMI_SERVICIO_TELEVISION	integer
FAMI_COMPUTADOR	integer
FAMI_LAVADORA	integer
FAMI_NEVERA	integer
FAMI_HORNO	integer
FAMI_DVD	integer
FAMI_MICROONDAS	integer
FAMI_AUTOMOVIL	integer
FAMI_INGRESO_FMILIAR_MENSUAL	integer
ESTU_TRABAJA	integer
ESTU_HORAS_TRABAJA	integer
PUNT_LENGUAJE	integer
PUNT_MATEMATICAS	integer
PUNT_C_SOCIALES	integer
PUNT_FILOSOFIA	integer
PUNT_BIOLOGIA	integer
PUNT_QUIMICA	integer
PUNT_FISICA	integer
PUNT_INGLES	integer
DESEMP_INGLES	character
NOMBRE_COMP_FLEXIBLE	character
PUNT_COMP_FLEXIBLE	character
DESEMP_COMP_FLEXIBLE	character
ESTU_PUESTO	character

Análisis de las variables del dataset de ICFES Saber 11

Los tipos de datos encontrados en el dataset de las notas del ICFES son:

Character: Es el tipo de dato usado para texto o cadenas de caracteres.

Integer: Se usa para los números enteros sin decimales.

numeric: Se usa para los números decimales se puede usan para números enteros también.

logical: Se usa para valores booleanos TRUE o FALSE.

Los variables del conjunto de datos son sobre datos del estudiante, residencia del estudiante, información sobre el colegio, información sobre los grados cursados, información familiar e ingresos.

Cantidad de cada tipo de variable

Para poder contar cada tipo de variable se uso la función sapply() la cual aplica una función a cada columna del data frame, se le aplico la función class para que tome el tipo de datos de cada variable, lo que genera es un vector con los datos, con la función table() se cuenta cuantas veces aparece un valor del vector creado con sapply() y asi obtener la cantidad de cada tipo de variable.

```
tabla_tipos <- table(sapply(datos_icfes_delimiter, class))
tabla_tipos
```

```
##
## character   integer   logical   numeric
##        47        63         1         1
```

Los datos obtenidos son con el dataset original, sin haberle hecho ninguna técnica de transformación y limpieza de datos, por ende se van a tomar como datos cualitativos el tipo character y como datos cuantitativos los tipos: integer y numeric. Con este análisis hay en total 47 variables categóricas o cualitativas y 64 variables cuantitativas o numéricas. Pero en el conjunto hay una variable con un tipo de dato logical la cual es la siguiente:

```
variables_logicas <- names(datos_icfes_delimiter)[sapply(datos_icfes_delimiter, is
variables_logicas
```

```
## [1] "ESTU_LIMITA_BAJAVISION"
```

Para poder ver que datos en su mayoría tiene a variable se uso unique(), los datos son los siguientes datos:

```
unique(datos_icfes_delimiter$ESTU_LIMITA_BAJAVISION)
```

```
## [1] NA
```

Con la función all() se verifica si todo los datos son de tipo NA, esto arroja un valor de TRUE

```
all(is.na(datos_icfes_delimiter$ESTU_LIMITA_BAJAVISION))
```

```
## [1] TRUE
```

Con la función unique permite ver cuantos valores distintos hay el resultado fue NA osea esta variable no contiene datos y se uso all para verificar si en toda la variable soy hay datos NA y arrojo TRUE quiere decir que esta variable este campo nunca tuvo datos.

Conclusión ¿Que tipo de variables predominan?

Después de conocer y analizar cada variable se puede concluir que hay mas variables de tipo numéricas con 63 variables y 47 de tipo categóricas por ahora se toman como categóricas, la variable logical se descarta ya que al no contar con información no es importante en este análisis de tipo de variables.

2. Limpieza y preparación de datos

Verificar valores faltantes En el anterior análisis se descubrió que la variable ESTU_LIMITA_BAJAVISION tiene valores faltantes, si se usa la función any con is.na se debe obtener un valor True.

```
any(is.na(datos_icfes_delimiter))
```

```
## [1] TRUE
```

Para contar cuantos Na hay en todo el conjunto de datos se usa la función sum().

```
sum(is.na(datos_icfes_delimiter))
```

```
## [1] 1318206
```

Se encontró un valor considerable de valores NA

Variables con valores NA
Para ver los valores faltantes por columna se usa colSums() para poder ver de manera sencilla la cantidad de valores NA de cada variable.

```
#Por columna
na_por_columna <- colSums(is.na(datos_icfes_delimiter))

# columnas con NA
na_filtrados <- na_por_columna[na_por_columna > 0]

# Convertir a data frame
naKable_df <- data.frame(
  Variable = names(na_filtrados),
  Cantidad_NA = as.numeric(na_filtrados)
)

knitr::kable(naKable_df,
caption = "Cantidad de valores faltantes (NA) por columna", align = "lc")
```

Table 6: Cantidad de valores faltantes (NA) por columna

Variable	Cantidad_NA
ESTU_NACIMIENTO_DIA	436
ESTU_NACIMIENTO_MES	436
ESTU_NACIMIENTO_ANNO	436
ESTU_LIMITA_BAJAVISION	45390
ESTU_ETNIA	42344
ESTU_COD_RESIDE_MCPIO	336
ESTU_ZONA_RESIDE	318
ESTU_AREA_RESIDE	39
IND_COD_ICFES_TERMINO	2660

Variable	Cantidad_NA
COLE_COD_ICFES	4
COLE_COD_DANE_INSTITUCION	9526
ESTU_CARRDESEADA_TIPO	45268
ESTU_IES_COD_DESEADA	45281
ESTU_IES_COD_MPIO_DESEADA	45285
ESTU_ANO_MATRICULA_PRIMERO	45349
ESTU_ANO_TERMINO_QUINTO	45349
ESTU_ANOS_COLEGIO_ACTUAL	45349
ESTU_ANO_MATRICULA_SEXTO	45349
ESTU_ANOS_PREESCOLAR	45349
ESTU_CUANTOS_COLE_ESTUDIO	45355
ESTU_REPROBO_CUARTO	45349
ESTU_REPROBO_DECIMO	45349
ESTU_REPROBO_NOVENO	45349
ESTU_REPROBO_OCTAVO	45349
ESTU_REPROBO_PRIMERO	45349
ESTU_REPROBO_QUINTO	45349
ESTU_REPROBO_SEGUNDO	45349
ESTU_REPROBO_SEPTIMO	45349
ESTU_REPROBO_SEXTO	45349
ESTU_REPROBO_TERCERO	45349
ESTU_REPROBO_ONCE_MAS	45349
ESTU_RAZON_RETIRO	45380
ESTU_POR_AMIGOSESTUDIANDO	45350
ESTU_POR_UBICACION	45350
ESTU_POR_UNICAQUEOFRECE	45350
FAMI_OCUPA_PADRE	283
FAMI_OCUPA_MADRE	283
FAMI_ESTRATO_VIVIENDA	2380
FAMI_NIVEL_SISBEN	289
FAMI_PERSONAS_HOGAR	297
FAMI_CUARTOS_HOGAR	41423
FAMI_PISOSHOGAR	354
FAMI_TELEFONO_FIJO	286
FAMI_CELULAR	45
FAMI_INTERNET	280
FAMI_SERVICIO_TELEVISION	282
FAMI_COMPUTADOR	282
FAMI_LAVADORA	282

Variable	Cantidad_NA
FAMI_NEVERA	45
FAMI_HORNO	280
FAMI_DVD	282
FAMI_MICROONDAS	282
FAMI_AUTOMOVIL	282
FAMI_INGRESO_FMILIAR_MENSUAL	283
ESTU_TRABAJA	38100
ESTU_HORAS_TRABAJA	41676
PUNT_LENGUAJE	78
PUNT_MATEMATICAS	4

Análisis de variables irrelevantes o vacías
Después de mirar que variables contienen valores de tipo Na se analizo un conjunto de variables que para los siguientes análisis del taller no afectan a los resultados del análisis, estas variables son:
Las variables de reprobación de grados: Se eliminan porque es una variable que no es relevante para los próximos análisis.
Variables a eliminar:

ESTU_REPROBO_CUARTO
ESTU_REPROBO_DECIMO
ESTU_REPROBO_NOVENO
ESTU_REPROBO_OCTAVO
ESTU_REPROBO_PRIMERO
ESTU_REPROBO_QUINTO
ESTU_REPROBO_SEGUNDO
ESTU_REPROBO_SEPTIMO
ESTU_REPROBO_SEXTO
ESTU_REPROBO_TERCERO
ESTU_REPROBO_ONCE_MAS

Se crea una nueva variable llamada datos_icfes_reprobo que guardara los datos sin tener en cuenta los que se van a descartar para los análisis.

```r
library(dplyr)

datos_icfes_reprobo <- datos_icfes_delimiter %>% select(-c(
  ESTU_REPROBO_CUARTO,
    ESTU_REPROBO_DECIMO,
    ESTU_REPROBO_NOVENO,
    ESTU_REPROBO_OCTAVO,
    ESTU_REPROBO_PRIMERO,
    ESTU_REPROBO_QUINTO,
    ESTU_REPROBO_SEGUNDO,
    ESTU_REPROBO_SEPTIMO,
    ESTU_REPROBO_SEXTO,
    ESTU_REPROBO_TERCERO,
    ESTU_REPROBO_ONCE_MAS
))
```

Variables de Incapacidad: En el conjunto de datos se elimino la variable ESTU_LIMITA_BAJAVISION ya que no es relevante para el análisis del taller se eliminara de la variable datos_icfes_reprobo

```r
datos_icfes_incapa <- datos_icfes_reprobo %>% select(-c(
  ESTU_LIMITA_BAJAVISION))
```

Valores inconsistentes o faltantes

ESTU_CUANTOS_COLE_ESTUDIO: Se analizo si hay datos inconsistentes en la cantidad de colegios de un estudiante se valida que si un estudiante tiene registrado mas de 10 colegios ya es un dato dudoso.Se encontraron ceros inconsistencias.

```r
library(dplyr)

datos_icfes_incapa %>%
  filter(ESTU_CUANTOS_COLE_ESTUDIO > 10) %>%
  select(ESTU_CUANTOS_COLE_ESTUDIO) %>%
  distinct() %>%
  arrange(desc(ESTU_CUANTOS_COLE_ESTUDIO))
```

```
## [1] ESTU_CUANTOS_COLE_ESTUDIO
## <0 rows> (or 0-length row.names)
```

FAMI_PERSONAS_HOGAR: Se analiza esta variable en donde si hay 0 personas del hogar o mas de 20 seria un dato inconsistente.Se encontraron ceros inconsistencias.

```
datos_icfes_incapa %>%
  filter(FAMI_PERSONAS_HOGAR == 0 | FAMI_PERSONAS_HOGAR > 20) %>%
  select(FAMI_PERSONAS_HOGAR) %>%
  distinct() %>%
  arrange(FAMI_PERSONAS_HOGAR)
```

```
## [1] FAMI_PERSONAS_HOGAR
## <0 rows> (or 0-length row.names)
```

FAMI_ESTRATO_VIVIENDA: A la hora de analizar el estrado se encontraron inconsistencias con los datos ya que se encontraron estratos mayores del estrato 6

```
library(dplyr)
library(knitr)

# Tabla de frecuencias del estrato de vivienda (incluye NA)
datos_icfes_incapa %>%
  group_by(FAMI_ESTRATO_VIVIENDA) %>%
  summarise(Frecuencia = n()) %>%
  arrange(desc(Frecuencia)) %>%
  kable(caption = "Distribución de la variable FAMI_ESTRATO_VIVIENDA")
```

Table 7: Distribución de la variable FAMI_ESTRATO_VIVIENDA

FAMI_ESTRATO_VIVIENDA	Frecuencia
2	17500
1	12225
3	10337
NA	2380
4	2040
5	548
6	237

FAMI_ESTRATO_VIVIENDA	Frecuencia
22	46
8	42
26	7
14	5
10	3
17	3
18	3
19	3
21	3
16	2
20	2
7	1
12	1
15	1
23	1

Como solución se decidió por filtrar y limpiar los valores mayores del estrato 6.

```
#  Solo estratos válidos
datos_icfes_estrato <- datos_icfes_incapa %>%
mutate(FAMI_ESTRATO_VIVIENDA = as.character(FAMI_ESTRATO_VIVIENDA)) %>%
filter(FAMI_ESTRATO_VIVIENDA %in% c("1", "2", "3", "4", "5", "6") | is.na(FAMI_EST
mutate(FAMI_ESTRATO_VIVIENDA = as.numeric(FAMI_ESTRATO_VIVIENDA))

#Estratos
library(dplyr)
library(knitr)

datos_icfes_estrato %>%
  group_by(FAMI_ESTRATO_VIVIENDA) %>%
  summarise(Frecuencia = n()) %>%
  arrange(desc(Frecuencia)) %>%
  kable(caption = "Distribución nueva de la variable FAMI_ESTRATO_VIVIENDA")
```

Table 8: Distribución nueva de la variable FAMI_ESTRATO_VIVIENDA

FAMI_ESTRATO_VIVIENDA	Frecuencia
2	17500
1	12225
3	10337
NA	2380
4	2040
5	548
6	237

Estandarizar columnas con errores estructurales

Se estandarizo los datos de los documentos de los estudiantes, ya que los valores que contiene esta variable son:

```
## [1] "C" "R" "T" "E" "Q" "V" "P" ""
```

Para mejorar los datos se crea un vector con los nombres completos de los documentos y se usa dlpyr::recode para remplazar los valores.

```
library(dplyr)

datos_icfes_limpios <- datos_icfes_estrato %>%
  mutate(ESTU_TIPO_DOCUMENTO = case_when(
    ESTU_TIPO_DOCUMENTO == "C" ~ "Cédula",
    ESTU_TIPO_DOCUMENTO == "P" ~ "Pasaporte",
    ESTU_TIPO_DOCUMENTO == "T" ~ "Tarjeta de Identidad",
    ESTU_TIPO_DOCUMENTO == "R" ~ "Registro Civil",
    ESTU_TIPO_DOCUMENTO == "E" ~ "Cédula de Extranjería",
    ESTU_TIPO_DOCUMENTO == "Q" ~ "Permiso Especial de Permanencia",
    ESTU_TIPO_DOCUMENTO == "V" ~ "Visa",
    ESTU_TIPO_DOCUMENTO == ""  ~ "Sin Información",
    TRUE ~ ESTU_TIPO_DOCUMENTO  # valores atípicos
  ))

unique(datos_icfes_limpios$ESTU_TIPO_DOCUMENTO)
```

```
## [1] "Cédula"                           "Registro Civil"
## [3] "Tarjeta de Identidad"             "Cédula de Extranjería"
## [5] "Permiso Especial de Permanencia" "Visa"
## [7] "Pasaporte"                        "Sin Información"
```

Imputar datos faltantes

Primero se usa la función colSums con is.na para ver las columnas que contiene valores NA del dataset limpiado.

```
library(knitr)
library(dplyr)

colSums(is.na(datos_icfes_limpios)) %>%
  as.data.frame() %>%
  tibble::rownames_to_column(var = "Variable") %>%
  rename(Valores_Faltantes = ".") %>%
  filter(Valores_Faltantes > 0) %>%
  kable(caption = "Cantidad de valores faltantes por variable")
```

Table 9: Cantidad de valores faltantes por variable

Variable	Valores_Faltantes
ESTU_NACIMIENTO_DIA	434
ESTU_NACIMIENTO_MES	434
ESTU_NACIMIENTO_ANNO	434
ESTU_ETNIA	42228
ESTU_COD_RESIDE_MCPIO	333
ESTU_ZONA_RESIDE	317
ESTU_AREA_RESIDE	38
IND_COD_ICFES_TERMINO	2622
COLE_COD_ICFES	3
COLE_COD_DANE_INSTITUCION	9508
ESTU_CARRDESEADA_TIPO	45226
ESTU_IES_COD_DESEADA	45163
ESTU_IES_COD_MPIO_DESEADA	45163
ESTU_ANO_MATRICULA_PRIMERO	45226
ESTU_ANO_TERMINO_QUINTO	45226
ESTU_ANOS_COLEGIO_ACTUAL	45226
ESTU_ANO_MATRICULA_SEXTO	45226

Variable	Valores_Faltantes
ESTU_ANOS_PREESCOLAR	45226
ESTU_CUANTOS_COLE_ESTUDIO	45232
ESTU_RAZON_RETIRO	45257
ESTU_POR_AMIGOSESTUDIANDO	45227
ESTU_POR_UBICACION	45227
ESTU_POR_UNICAQUEOFRECE	45227
FAMI_OCUPA_PADRE	282
FAMI_OCUPA_MADRE	282
FAMI_ESTRATO_VIVIENDA	2380
FAMI_NIVEL_SISBEN	283
FAMI_PERSONAS_HOGAR	296
FAMI_CUARTOS_HOGAR	41382
FAMI_PISOSHOGAR	282
FAMI_TELEFONO_FIJO	281
FAMI_CELULAR	44
FAMI_INTERNET	279
FAMI_SERVICIO_TELEVISION	281
FAMI_COMPUTADOR	281
FAMI_LAVADORA	281
FAMI_NEVERA	44
FAMI_HORNO	279
FAMI_DVD	281
FAMI_MICROONDAS	281
FAMI_AUTOMOVIL	281
FAMI_INGRESO_FMILIAR_MENSUAL	282
ESTU_TRABAJA	38076
ESTU_HORAS_TRABAJA	41588
PUNT_LENGUAJE	4

Con las variables con datos NA, se procederá a escoger algunas variables para realizar la imputación de datos, la cual consiste en llenar los valores faltantes (NA) usando una técnica lógica en este caso usando la media, moda o diferentes técnicas. Las variables que cuenten con una valor elevado de NA no serán imputadas por su dificultad, como lo son: Estu_ETNIA, ESTU_CARRDESEADA_TIPO, ESTU_IES-COD-DESEADA etc Por otro lado hay variables como lo son: FAMI_OCUPA_PADRE, FAMI_OCUPA_MADRE, ESTU_TRABAJA, FAMI_CELULAR y FAMI_NEVERA que tienen pocos valores NA, se pueden imputar

Imputación Variables

Para las variables de tipo Integer como FAMI_OCUPA_PADRE se realiza la imputación con la mediana la cual remplaza los valores de NA por la media de los valores existentes, no se eliminan solo se remplazan.

```
#Copia del dataframe
datos_icfes_limpios2 <- datos_icfes_limpios
#FAMI_OCUPA_PADRE con la mediana
datos_icfes_limpios2 $FAMI_OCUPA_PADRE[is.na(datos_icfes_limpios2$FAMI_OCUPA_PADRE
  median(datos_icfes_limpios2$FAMI_OCUPA_PADRE, na.rm = TRUE)

#La imputación se realizó correctamente
cat("Variable FAMI_OCUPA_PADRE imputada con mediana:\n")
```

```
## Variable FAMI_OCUPA_PADRE imputada con mediana:
```

```
summary(datos_icfes_limpios2$FAMI_OCUPA_PADRE)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1.00   17.00   21.00   19.27   21.00   26.00
```

```
#FAMI_OCUPA_MADRE
datos_icfes_limpios2$FAMI_OCUPA_MADRE[is.na(datos_icfes_limpios2$FAMI_OCUPA_MADRE)
  median(datos_icfes_limpios2$FAMI_OCUPA_MADRE, na.rm = TRUE)

cat("Variable FAMI_OCUPA_MADRE imputada con mediana:\n")
```

```
## Variable FAMI_OCUPA_MADRE imputada con mediana:
```

```
summary(datos_icfes_limpios2$FAMI_OCUPA_MADRE)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1.00   19.00   22.00   19.98   22.00   26.00
```

```
#ESTU_TRABAJA integer
datos_icfes_limpios2$ESTU_TRABAJA[is.na(datos_icfes_limpios2$ESTU_TRABAJA)] <-
  median(datos_icfes_limpios2$ESTU_TRABAJA, na.rm = TRUE)

# La imputación se realizó correctamente
cat("Variable ESTU_TRABAJA imputada con mediana:\n")
```

```
## Variable ESTU_TRABAJA imputada con mediana:
```

```
summary(datos_icfes_limpios2$ESTU_TRABAJA)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.000   1.000   1.000   1.133   1.000   7.000
```

```
#FAMI_CELULAR integer

datos_icfes_limpios2$FAMI_CELULAR[is.na(datos_icfes_limpios2$FAMI_CELULAR)] <-
  median(datos_icfes_limpios2$FAMI_CELULAR, na.rm = TRUE)

# La imputación se realizó correctamente
cat("Variable FAMI_CELULAR imputada con mediana:\n")
```

```
## Variable FAMI_CELULAR imputada con mediana:
```

```
summary(datos_icfes_limpios2$FAMI_CELULAR)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.0000  1.0000  1.0000  0.9437  1.0000  1.0000
```

```
#FAMI_NEVERA integer

datos_icfes_limpios2$FAMI_NEVERA[is.na(datos_icfes_limpios2$FAMI_NEVERA)] <-
  median(datos_icfes_limpios2$FAMI_NEVERA, na.rm = TRUE)

# La imputación se realizó correctamente
cat("Variable FAMI_NEVERA imputada con mediana:\n")
```

```
## Variable FAMI_NEVERA imputada con mediana:
```

```
summary(datos_icfes_limpios2$FAMI_NEVERA)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.0000  1.0000  1.0000  0.8971  1.0000  1.0000
```

Las variables seleccionadas se imputaron correctamente por esta razon si se mira el data frame datos_icfes_limpios2 estas no aparecen cuando se filtra por valores NA

```
library(knitr)
library(dplyr)

colSums(is.na(datos_icfes_limpios2)) %>%
  as.data.frame() %>%
  tibble::rownames_to_column(var = "Variable") %>%
  rename(Valores_Faltantes = ".") %>%
  filter(Valores_Faltantes > 0) %>%
  kable(caption = "Cantidad de valores faltantes por variable Imputado")
```

Table 10: Cantidad de valores faltantes por variable Imputado

Variable	Valores_Faltantes
ESTU_NACIMIENTO_DIA	434
ESTU_NACIMIENTO_MES	434
ESTU_NACIMIENTO_ANNO	434
ESTU_ETNIA	42228
ESTU_COD_RESIDE_MCPIO	333
ESTU_ZONA_RESIDE	317
ESTU_AREA_RESIDE	38
IND_COD_ICFES_TERMINO	2622
COLE_COD_ICFES	3
COLE_COD_DANE_INSTITUCION	9508
ESTU_CARRDESEADA_TIPO	45226
ESTU_IES_COD_DESEADA	45163
ESTU_IES_COD_MPIO_DESEADA	45163
ESTU_ANO_MATRICULA_PRIMERO	45226
ESTU_ANO_TERMINO_QUINTO	45226
ESTU_ANOS_COLEGIO_ACTUAL	45226
ESTU_ANO_MATRICULA_SEXTO	45226
ESTU_ANOS_PREESCOLAR	45226
ESTU_CUANTOS_COLE_ESTUDIO	45232
ESTU_RAZON_RETIRO	45257
ESTU_POR_AMIGOSESTUDIANDO	45227

Variable	Valores_Faltantes
ESTU_POR_UBICACION	45227
ESTU_POR_UNICAQUEOFRECE	45227
FAMI_ESTRATO_VIVIENDA	2380
FAMI_NIVEL_SISBEN	283
FAMI_PERSONAS_HOGAR	296
FAMI_CUARTOS_HOGAR	41382
FAMI_PISOSHOGAR	282
FAMI_TELEFONO_FIJO	281
FAMI_INTERNET	279
FAMI_SERVICIO_TELEVISION	281
FAMI_COMPUTADOR	281
FAMI_LAVADORA	281
FAMI_HORNO	279
FAMI_DVD	281
FAMI_MICROONDAS	281
FAMI_AUTOMOVIL	281
FAMI_INGRESO_FMILIAR_MENSUAL	282
ESTU_HORAS_TRABAJA	41588
PUNT_LENGUAJE	4

2.1 Nuevas transformaciones

Calcular edad estudiante

Para poder calcular edad del estudiante a partir de su fecha de nacimiento y la fecha del examen que es del 2012 se usan las variables: ESTU_NACIMIENTO_DIA, ESTU_NACIMIENTO_MES, ESTU_NACIMIENTO_ANNO.

Se creo una nueva columna con la fecha completa del estudiante llamada FECHA_NACIMIENTO

```
#Nueva columna con fecha completa
datos_icfes_limpios2$FECHA_NACIMIENTO <- as.Date(paste
format = "%Y-%m-%d")

#Se toma una fecha del año 2012 ya que no esta la fecha exacta en los datos
fecha_examen <- as.Date("2012-01-01")

# Calcular la edad se restan la fecha de nacimiento de la fecha del examen
```

```r
datos_icfes_limpios2$EDAD <- as.numeric(
difftime(fecha_examen, datos_icfes_limpios2$FECHA_NACIMIENTO,
units = "weeks")) / 52.25
# Redondear las edades a enteros
datos_icfes_limpios2$EDAD <- round(datos_icfes_limpios2$EDAD)
library(knitr)

# Mostrar solo las primeras 15 filas con la columna "EDAD"
kable(head(datos_icfes_limpios2 %>% select(EDAD), 15),
      col.names = c("Edad"),
      caption = "Primeras 15 Edades")
```

Table 11: Primeras 15 Edades

Edad
19
41
18
37
40
21
27
26
17
24
23
20
17
18
21

Nueva columna edad categorizada

Se creo una nueva columna llamada CATEGORIA_EDAD, para guardar las categorías de las edades, se uso la función mutate(), para poder agregar los valores en las columnas, y con el case_when() se crean las categorías de la edad.

```r
# Crear la nueva columna "CATEGORIA_EDAD"
datos_icfes_limpios2 <- datos_icfes_limpios2 %>%
  mutate(CATEGORIA_EDAD = case_when(
```

```
    EDAD >= 12 & EDAD <= 17 ~ "Adolescente",
    EDAD >= 18 & EDAD <= 26 ~ "Joven",
    EDAD >= 27 & EDAD <= 59 ~ "Adulto",
    EDAD >= 60 ~ "Adulto mayor",
    TRUE ~ "Desconocido"  # Para edades fuera de los rangos
  ))

library(knitr)

kable(head(datos_icfes_limpios2 %>% select(EDAD, CATEGORIA_EDAD), 15),
      col.names = c("Edad", "Categoría de Edad"),
      caption = "Primeras 15 Edades y Categorías de Edad")
```

Table 12: Primeras 15 Edades y Categorías de Edad

Edad	Categoría de Edad
19	Joven
41	Adulto
18	Joven
37	Adulto
40	Adulto
21	Joven
27	Adulto
26	Joven
17	Adolescente
24	Joven
23	Joven
20	Joven
17	Adolescente
18	Joven
21	Joven

3. Analisis Exploratorio y Univariado

En esta sección del taller se van a analizar algunas variables del dataset limpio el cual es datos_icfes_limpios2, se realizara un análisis univariado para comprender el comportamiento de los datos. Para obtener los valores generales estadísticos de las variables: Edad_Estudiantes, PUNT_LENGUAJE,

PUNT_MATEMATICAS, PUNT_C_SOCIALES, PUNT_FILOSOFIA, PUNT_BIOLOGIA, PUNT_QUIMICA, PUNT_FISICA, PUNT_INGLES Se creara un data frame en el cual se calculara cada valor como lo es: minimo, Q1, mediana, promedio, Q3, maximo y desviación estándar para poder llamar los datos de cada variable y poder analizar de mejor manera los datos.

```
obtener_estadisticos <- function(variable, nombre_variable) {
  data.frame(
    Variable = nombre_variable,
    Mínimo = min(variable, na.rm = TRUE),
    Q1 = quantile(variable, 0.25, na.rm = TRUE),
    Mediana = median(variable, na.rm = TRUE),
    Promedio = mean(variable, na.rm = TRUE),
    Q3 = quantile(variable, 0.75, na.rm = TRUE),
    Máximo = max(variable, na.rm = TRUE),
    Desviación_Estándar = sd(variable, na.rm = TRUE)
  )
}
```

Datos estadisticos Edad_estudiantes

```
kable(obtener_estadisticos(datos_icfes_limpios2$EDAD, "EDAD"),
      caption = "Estadísticos Descriptivos: Edad", digits = 2)
```

Table 13: Estadísticos Descriptivos: Edad

	Variable	Mínimo	Q1	Mediana	Promedio	Q3	Máximo	Desviación_Estándar
25%	EDAD	6	18	20	22.9	25	88	7.4

La edad de los estudiantes cuenta con un rango muy amplio con un valor mínimo de 6 años y un máximo de 88 años, el promedio de la edad es de 22.9 años en comparación con la mediana se encuentra en 20 años, esto indica que más de la mitad de los estudiantes están en los 20 años, el cuartil Q1 está en 18 años, el Q3 está en 25 años esto indica que la mayoría de los estudiantes están en el intervalo de entre los 18 y 25 años. La desviación estándar es de 7.4 esto es una dispersión moderada, pero puede ser debida a los datos atípicos como los de 6 y 88 años, pero en general la población de estudiantes es mayormente joven y los datos no están tan dispersos.

Datos estadisticos PUNT_LENGUAJE

```
kable(obtener_estadisticos(datos_icfes_limpios2$PUNT_LENGUAJE, "LENG"),
      caption = "Estadísticos Descriptivos: Lenguaje", digits = 2)
```

Table 14: Estadísticos Descriptivos: Lenguaje

	Variable	Mínimo	Q1	Mediana	Promedio	Q3	Máximo	Desviación_Estándar
25%	LENG	0	46	50	49.89	54	90	8.3

El puntaje de la materia de lenguaje muestra un promedio de 49.89 puntos tiene una media de 50 esto indica que los datos están muy centrados, el puntaje mínimo fue de 0 y el máximo fue de 90 puntos el Q1 es de 46 y Q3 es de 54, esto dice que 50% de los estudiantes están en este intervalo la desviación fue de 8.3 da una dispersión normal, en general los estudiantes tuvieron un buen desempeño en esta asignatura.

Datos estadisticos PUNT_MATEMATICAS

```
kable(obtener_estadisticos(datos_icfes_limpios2$PUNT_MATEMATICAS, "MATE"),
      caption = "Estadísticos Descriptivos: Matemáticas", digits = 2)
```

Table 15: Estadísticos Descriptivos: Matemáticas

	Variable	Mínimo	Q1	Mediana	Promedio	Q3	Máximo	Desviación_Estándar
25%	MATE	0	40	46	46.55	52	117	11.47

En la materia de matemáticas los estudiantes obtuvieron promedio de 46.55 con una mediana de 46, el mínimo puntaje fue 0 y el máximo 117 el Q1 fue de 40 y Q3 es de 52. La desviación estándar fue de 11.47 es muy alta lo que indica que los puntajes de los estudiantes fueron muy distintos entre si.

Datos estadisticos PUNT_C_SOCIALES

```
kable(obtener_estadisticos(datos_icfes_limpios2$PUNT_C_SOCIALES, "SOCIA"),
      caption = "Estadísticos Descriptivos: Ciencias Sociales", digits = 2)
```

Table 16: Estadísticos Descriptivos: Ciencias Sociales

	Variable	Mínimo	Q1	Mediana	Promedio	Q3	Máximo	Desviación_Estándar
25%	SOCIA	0	41	46	46.01	51	82	8.23

En la materia de sociales los estudiantes tuvieron un puntaje de 46.01 con una mediana de 46 esto indica una distribución con simetría, el puntaje mínimo fue 0 y el máximo 82, el Q1 es de 41 y el Q3 51 esto indica que el 50% de los estudiantes tuvieron puntajes entre 41 y 51. La desviación fue de 8.23 es muy baja ya que como se vio anterior mente la distribución es muy simétrica.

Datos estadisticos PUNT_FILOSOFIA

```
kable(obtener_estadisticos(datos_icfes_limpios2$PUNT_FILOSOFIA, "FILO"),
      caption = "Estadísticos Descriptivos: Filosofía", digits = 2)
```

Table 17: Estadísticos Descriptivos: Filosofía

	Variable	Mínimo	Q1	Mediana	Promedio	Q3	Máximo	Desviación_Estándar
25%	FILO	0	34	40	40.23	46	103	10.01

En la materia de filosofía el promedio fue de 40.23 la mediana de 40 la distribución es muy simétrica, el menor puntaje fue 0 y el mayor fue 103. El Q1 es de 34 y el Q3 de 46 los puntajes de los estudiantes están entre 34 y 46, la desviación fue de 10.01 esto es una dispersión moderada.

Datos estadisticos PUNT_BIOLOGIA

```
kable(obtener_estadisticos(datos_icfes_limpios2$PUNT_BIOLOGIA, "BIOLO"),
      caption = "Estadísticos Descriptivos: Biología", digits = 2)
```

Table 18: Estadísticos Descriptivos: Biología

	Variable	Mínimo	Q1	Mediana	Promedio	Q3	Máximo	Desviación_Estándar
25%	BIOLO	0	42	46	46.03	51	100	8.22

En bilogía los puntajes estuvieron en promedio de 46.03 mediana de 46, mínimo de 0 máximo de 100 con un Q1 de 42 Y Q3 de 53 la desviación fue de 8.22 esta materia cuenta con unos puntajes moderados en dispersión.

Datos estadisticos PUNT_QUIMICA

```
kable(obtener_estadisticos(datos_icfes_limpios2$PUNT_QUIMICA, "QUIMI"),
      caption = "Estadísticos Descriptivos: Química", digits = 2)
```

Table 19: Estadísticos Descriptivos: Química

	Variable	Mínimo	Q1	Mediana	Promedio	Q3	Máximo	Desviación_Estándar
25%	QUIMI	0	40.5	45	45.76	50	92	7.91

En la materia de química los puntajes estuvieron en promedio de 45.76, mediana de 45, mínimo de 0, máximo de 92, con un Q1 de 40.5 y Q3 de 50. La desviación fue de 7.91. Esta materia cuenta con unos puntajes moderados en dispersión y una distribución centrada en valores del rango.

Datos estadisticos PUNT_FISICA

```
kable(obtener_estadisticos(datos_icfes_limpios2$PUNT_FISICA, "FISI"),
      caption = "Estadísticos Descriptivos: Física", digits = 2)
```

Table 20: Estadísticos Descriptivos: Física

	Variable	Mínimo	Q1	Mediana	Promedio	Q3	Máximo	Desviación_Estándar
25%	FISI	0	37	43	44.2	51	123	11.03

En la materia de física los puntajes promedios fueron de 44.2, se registró una mediana de 43, el puntaje mínimo fue de 0 y el máximo de 123, se contó con un Q1 de 37 y Q3 de 51 siendo los puntajes de física de 37 y 51. La desviación fue de 11.03. Esta materia presenta una dispersión alta en los puntajes, lo que indica que hay una mayor variabilidad en el desempeño de los estudiantes durante la presentación del examen.

Datos estadisticos PUNT_INGLES

```
kable(obtener_estadisticos(datos_icfes_limpios2$PUNT_INGLES, "ING"),
      caption = "Estadísticos Descriptivos: Inglés", digits = 2)
```

Table 21: Estadísticos Descriptivos: Inglés

	Variable	Mínimo	Q1	Mediana	Promedio	Q3	Máximo	Desviación_Estándar
25%	ING	-1	37	42	42.91	46	100	11.89

En la materia de inglés los puntajes promedios fueron de 42.91, se registró una mediana de 42, el puntaje mínimo fue de -1 y el máximo de 100, se contó con un Q1 de 37 y Q3 de 46, siendo el rango de los puntajes de inglés de 37 y 46. La desviación fue de 11.89. Esta materia presenta una alta dispersión en los puntajes, lo cual indica una gran variabilidad en los resultados de los estudiantes, posiblemente gracias a los diferentes niveles de dominio del idioma.

¿Cuál es la materia con mayor puntaje promedio?
Para obtener este valor se calcula el promedio de cada materia y se usa la función max(), para saber cual es el mayor promedio.

```
# Promedios
promedios_materias <- data.frame(
  Materia = c("Lenguaje", "Matemáticas", "Ciencias Sociales", "Filosofía",
              "Biología", "Química", "Física", "Inglés"),
  Promedio = c(mean(datos_icfes_limpios2$PUNT_LENGUAJE, na.rm = TRUE),
               mean(datos_icfes_limpios2$PUNT_MATEMATICAS, na.rm = TRUE),
               mean(datos_icfes_limpios2$PUNT_C_SOCIALES, na.rm = TRUE),
               mean(datos_icfes_limpios2$PUNT_FILOSOFIA, na.rm = TRUE),
               mean(datos_icfes_limpios2$PUNT_BIOLOGIA, na.rm = TRUE),
               mean(datos_icfes_limpios2$PUNT_QUIMICA, na.rm = TRUE),
               mean(datos_icfes_limpios2$PUNT_FISICA, na.rm = TRUE),
               mean(datos_icfes_limpios2$PUNT_INGLES, na.rm = TRUE))
)

# Mostrar la materia con el mayor puntaje promedio
materia_mayor_promedio <- promedios_materias[which.max(promedios_materias$Promedio
materia_mayor_promedio
```

```
##    Materia Promedio
## 1 Lenguaje   49.893
```

El resultado obtenido fue que la materia con mayor puntaje promedio fue la materia de Lenguaje con un promedio de 49.893, esto indica que en términos generales

los estudiantes obtuvieron mejores resultados de calificación en esta área evaluada, se puede interpretar que el grupo de estudiantes estudiado tiene fortalezas relacionadas con el área del lenguaje como la lectura crítica o compresión lectora.

¿Cuál presenta la mayor desviación estándar?
Se calcula la mayor desviación estándar mirando la desviación de cada una.

```
desviacion_estandar_materias <- data.frame(
  Materia = c("Lenguaje", "Matemáticas", "Ciencias Sociales", "Filosofía",
              "Biología", "Química", "Física", "Inglés"),
  Desviacion_Estandar = c(sd(datos_icfes_limpios2$PUNT_LENGUAJE, na.rm = TRUE),
                          sd(datos_icfes_limpios2$PUNT_MATEMATICAS, na.rm = TRUE),
                          sd(datos_icfes_limpios2$PUNT_C_SOCIALES, na.rm = TRUE),
                          sd(datos_icfes_limpios2$PUNT_FILOSOFIA, na.rm = TRUE),
                          sd(datos_icfes_limpios2$PUNT_BIOLOGIA, na.rm = TRUE),
                          sd(datos_icfes_limpios2$PUNT_QUIMICA, na.rm = TRUE),
                          sd(datos_icfes_limpios2$PUNT_FISICA, na.rm = TRUE),
                          sd(datos_icfes_limpios2$PUNT_INGLES, na.rm = TRUE))
)

materia_mayor_desviacion <-desviacion_estandar_materias[which.max(desviacion_estan
materia_mayor_desviacion
```

```
##   Materia Desviacion_Estandar
## 8  Inglés            11.88893
```

La materia que presento mayor desviación estándar fue ingles con un valor 11.88893, esto indica que los puntajes individuales de cada estudiante no fueron uniformes no hay uniformidad de datos, esto indica que diferentes estudiantes sacaron datos bajos o altos pero no hay una uniformidad que diga que la mayoría saco un puntaje en general. Esto puede ser debido a que durante el examen los estudiantes no contaban un buen nivel de inglés y otros al contrario si tenían bases sólidas en el idioma generando la diferencia de los datos.

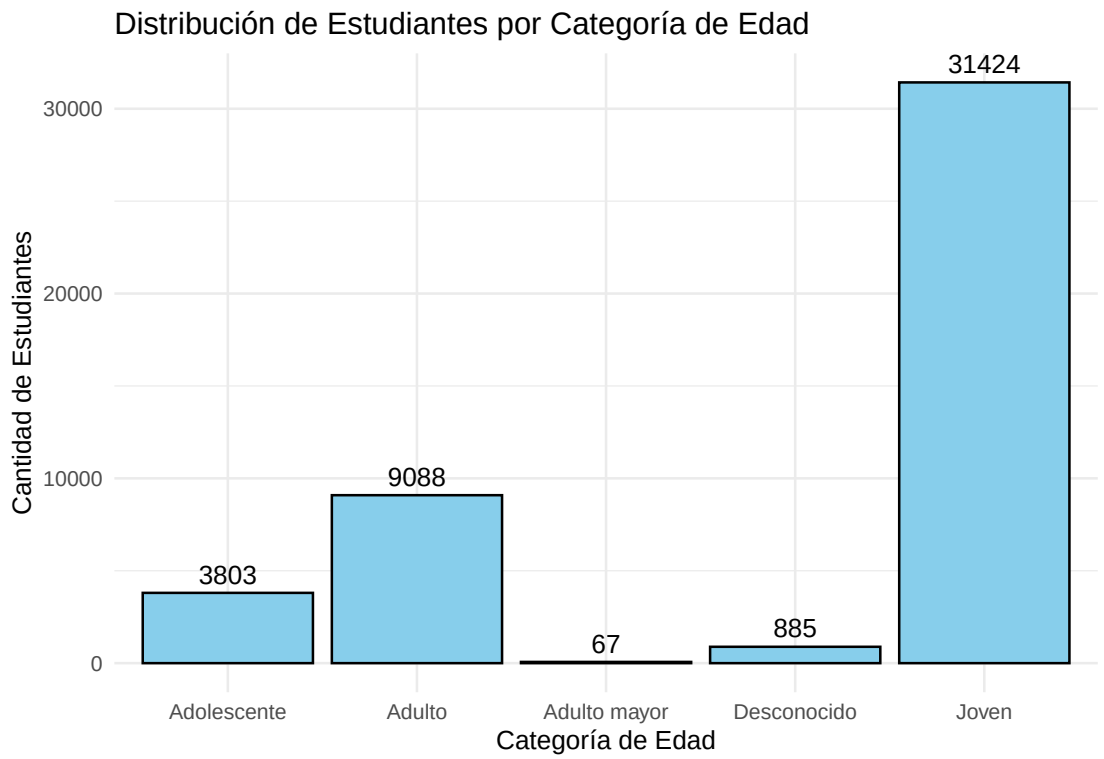
¿Cual es la distribución de edades entre los estudiantes

Para poder ver la distribución de edades de los estudiantes de hace uso de la variable categoria_edad creada anteriormente, los datos se mostraron con un diagrama de barras para ver cómo se distribuyen las categorías de las edades.

```r
library(ggplot2)
library(dplyr)

# Contar la cantidad de estudiantes por categoría
conteo_edades <- datos_icfes_limpios2 %>%
  count(CATEGORIA_EDAD)

# Gráfico de barras
ggplot(conteo_edades, aes(x = CATEGORIA_EDAD, y = n)) +
  geom_bar(stat = "identity", fill = "skyblue", color = "black") +
  geom_text(aes(label = n), vjust = -0.5, size = 4) +  # valores
  labs(title = "Distribución de Estudiantes por Categoría de Edad",
       x = "Categoría de Edad",
       y = "Cantidad de Estudiantes") +
  theme_minimal()
```



Con los datos obtenidos del diagrama se puede decir que la categorización de las edades es de la siguiente manera:
Jóvenes (18 - 26 años): 31.424 estudiantes
Adultos (27 - 59 años): 9.088 estudiantes

Adolescentes (12 - 17 años): 3.803 estudiantes
Adulto mayor (60+ años): 67 estudiantes
Desconocido: 885 estudiantes

La variable desconocida se creó para guardar los datos faltantes o erróneos de la variable edad. La mayoría de los estudiantes se encuentran en la categoría de Jóvenes, este dato es coherente ya que en la mayoría de los casos los estudiantes que presentan el ICFES están a punto de salir del colegio y sus edades son de 18 para arriba.

En el análisis se obtuvo que participo una gran proporción de adultos mayores de 26 años, esto se debe posiblemente a que este grupo pudo estar compuesto por personas que retomaron los estudios o quieren estudiar en una Universidad y necesitan la nota del examen para aplicar a los programas universitarios.

Los adolescentes representan una proporción menor de la muestra de la población de estudiantes, esto se debe a que a pesar que en el grado 11 los estudiantes tienen o son mayores a 18 años hay excepciones en donde están a punto de cumplirlos cuando presentan el examen. Los adultos mayores son una gran minoría ya que es un dato esperado en el análisis pero la presencia de los valores desconocidos que fueron 885 esto indica que existen registros que están incorrectos o ausentes, esto es un limitante en el análisis ya que no se puede tomar en cuenta toda la muestra de los estudiantes que presentaron el examen de ICFES en el 2012.

4. Análisis Bivariado

Teniendo en cuenta que hay que analizar las comparaciones con: ciudad, La edad (Categorizada), tipo y caracterización del colegio y nivel de ingles se va realizar un análisis bivariado teniendo en cuenta la variable y el genero para analizar si los hombres o las mujeres son mejores en matemáticas. Se tomaron como variables independientes las comparaciones sugeridas y el genero (hombre o mujer) están se relacionan con una variable dependiente los puntajes de matemáticas o PUNT_MATEMATICAS.

Análisis Bivariado

En el ejercicio se solicita las comparaciones por ciudad pero a la falta de esta información se opto por usar la clasificación por departamento. Se tomaron cada variable con el genero y se realiza un group_by() para poder generar el análisis correctamente.

Departamento y género

```
library(dplyr)
library(ggplot2)
library(knitr)

# Promedio por género y departamento
mat_ciudad <- datos_icfes_limpios2 %>%
  group_by(ESTU_RESIDE_DEPTO , ESTU_GENERO) %>%
  summarise(Prom_MATE = mean(PUNT_MATEMATICAS, na.rm = TRUE), .groups = "drop")

kable(head(mat_ciudad, 10), caption = "Promedio de Matemáticas por Género y Depart
```

Table 22: Promedio de Matemáticas por Género y Departamento

ESTU_RESIDE_DEPTO	ESTU_GENERO	Prom_MATE
	X	45.58559
AMAZONAS	F	43.50000
AMAZONAS	M	45.14286
ANTIOQUIA	F	41.87516
ANTIOQUIA	M	45.10169
ANTIOQUIA	X	44.06667
ARAUCA	F	40.95789
ARAUCA	M	45.46154
ATLANTICO	F	41.56995
ATLANTICO	M	42.88889

En el departamento de Amazonas las mujeres (F) tienen un puntaje promedio de 43.50 en matemáticas por otro lado los hombres (M) tiene un puntaje promedio de 45.14, los hombres en este departamento tienen mayor puntaje.

En el departamento de Antioquia las mujeres (F) tienen un puntaje promedio de 41.88 en matematicas por otro lado los hombres (M) tiene un puntaje promedio de 45.10, los hombres en este departamento tienen mayor puntaje.

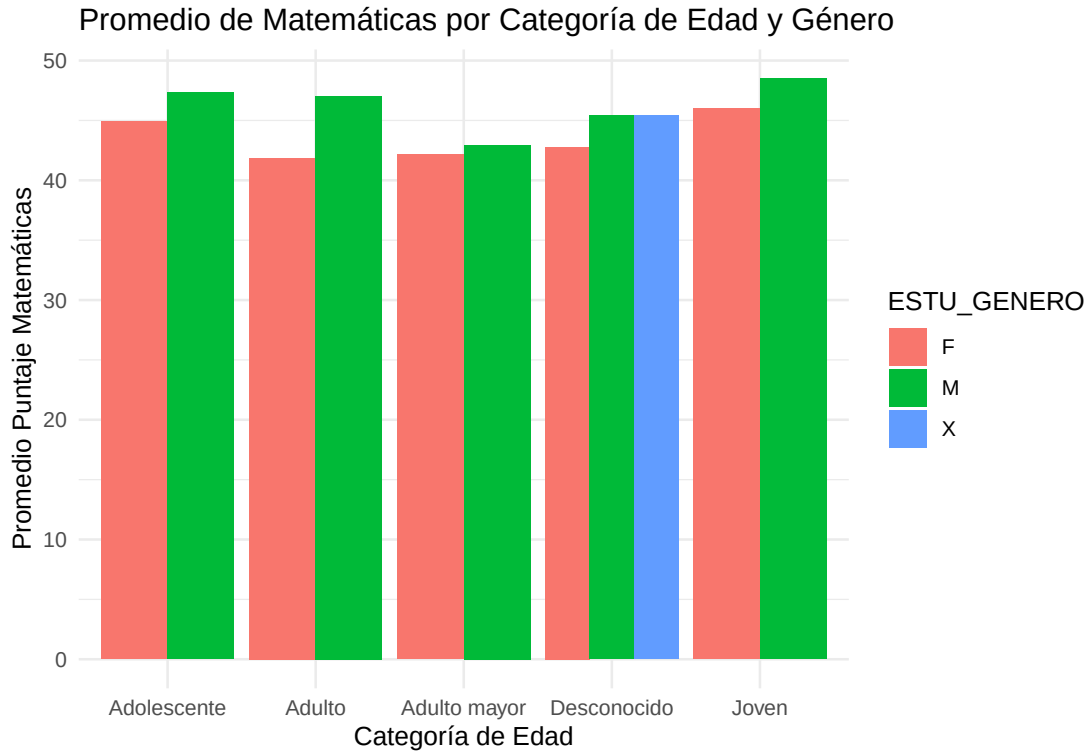
En el departamento de Arauca las mujeres (F) tienen un puntaje promedio de 40.96 en matematicas por otro lado los hombres (M) tiene un puntaje promedio de 45.46, los hombres en este departamento tienen mayor puntaje.

En el departamento de Atlantico las mujeres (F) tienen un puntaje promedio de 41.88 en matematicas por otro lado los hombres (M) tiene un puntaje promedio de 45.10, los hombres en este departamento tienen mayor puntaje.

Por categoria edad y genero

```
mat_edad <- datos_icfes_limpios2 %>%
  group_by(CATEGORIA_EDAD, ESTU_GENERO) %>%
  summarise(Prom_MATE = mean(PUNT_MATEMATICAS, na.rm = TRUE), .groups = "drop")

ggplot(mat_edad, aes(x = CATEGORIA_EDAD, y = Prom_MATE, fill = ESTU_GENERO)) +
  geom_col(position = "dodge") +
  labs(title = "Promedio de Matemáticas por Categoría de Edad y Género",
       x = "Categoría de Edad", y = "Promedio Puntaje Matemáticas") +
  theme_minimal()
```



Adolescentes: Las mujeres (F) tuvieron un puntaje de 45 y los hombres (M) de 48 los hombres tienen mejor puntaje que las mujeres.

Adulto: Las mujeres (F) tuvieron un puntaje de 43 y los hombres (M) de 48 la diferencia fue de 5 puntos con respecto a las mujeres.

Adulto Mayor: Las mujeres (F) tuvieron un puntaje de 43 y los hombres (M) de 44 la diferencia fue muy pequeña.

Joven: Las mujeres (F) obtuvieron 47 y los hombres (M) un puntaje de 49

Se puede concluir que en todas las categorías de edad los hombres (M) obtienen mejores promedios en matemáticas que las mujeres (F).

Tipo y caracterización del colegio y género

```
library(dplyr)
library(ggplot2)
library(knitr)

# Agrupamos por variables del colegio y género
mat_colegio <- datos_icfes_limpios2 %>%
  group_by(ESTU_GENERO, COLE_CARACTER, COLE_GENERO) %>%
  summarise(Prom_MATEMATICAS = mean(PUNT_MATEMATICAS, na.rm = TRUE), .groups = "dr

# Primeros resultados
kable(head(mat_colegio, 10), caption = "Promedio de Matemáticas según Género y Car
```

Table 23: Promedio de Matemáticas según Género y Características del Colegio

ESTU_GENERO	COLE_CARACTER	COLE_GENERO	Prom_MATEMATICAS
F			42.91
F		X	43.11
F	ACADEMICO	F	52.85
F	ACADEMICO	M	46.42
F	ACADEMICO	X	43.95
F	ACADEMICO Y TECNICO	F	49.30
F	ACADEMICO Y TECNICO	M	42.31
F	ACADEMICO Y TECNICO	X	46.50
F	DESCONOCIDO	F	43.00
F	DESCONOCIDO	M	37.50

Con base a los resultados las mujeres que estudian en un colegios femeninos tienen puntajes más altos con 52.85, por otro lado los colegios de género masculino tienen un promedio más bajo con 37.50, se puede concluir que las mujeres tienen mejores resultados si el colegio es puramente académico y femenino.

Nivel de inglés y género

```
# Agrupamos por género y nivel de inglés
mat_ingles <- datos_icfes_limpios2 %>%
group_by(ESTU_GENERO, DESEMP_INGLES) %>%
summarise(Prom_MATEMATICAS = mean(PUNT_MATEMATICAS, na.rm = TRUE),
  .groups = "drop")

# Tabla
kable(mat_ingles,
caption = "Promedio de Matemáticas por Género y Nivel de Inglés", digits = 2)
```

Table 24: Promedio de Matemáticas por Género y Nivel de Inglés

ESTU_GENERO	DESEMP_INGLES	Prom_MATEMATICAS
F		46.00
F	42	50.00
F	43	57.00
F	75	54.00
F	A-	41.96
F	A1	47.83
F	A2	53.86
F	B+	61.09
F	B1	57.20
M		39.50
M	57	61.00
M	A-	44.11
M	A1	50.55
M	A2	56.86
M	B+	65.79
M	B1	60.70
X	35	44.00
X	A-	42.36
X	A1	48.50
X	A2	56.19
X	B+	59.00
X	B1	60.05

Según los datos de nivel de inglés los mayores niveles B1 y B+ tuvieron un mayor promedio de matemáticas, esto es para mujeres y hombres, los hombres tienen los mayores promedios en ingles en B+ es de 65.79 y B1 de 60.70, las mujeres tienen buenos puntajes también B+ de 61.09 y B1 de 57.20. Los hombres son los que tienen mayor puntaje en ingles por ende la relación es proporcional con matemáticas.

¿Quien se destaca más en matemáticas hombres o mujeres?

Según los datos analizados de cada variable los hombres destacan más en el área de matemáticas que las mujeres, ya que mostraron mayores promedios en casi todos los casos de estudio como en edad, la ubicación geográfica, el tipo de colegio y en el nivel de inglés

¿Existe diferencia significativa en el puntaje global por género?
Si existe una diferencia significativa en el puntaje global de género, según los análisis realizados sobre las variables de departamento, edad, tipo de colegio y nivel de inglés, se puedo encontrar que en la variable por genero los hombres obtuvieron promedios más altos que las mujeres, en la clasificación por departamentos los hombres también tuvieron los mejores puntajes por ejemplo en Antioquia los hombres tuvieron 45.10 y las mujeres 41.87. En la clasificación de edad los hombres también lideraron los puntajes se puede concluir que si existe una gran diferencia de puntajes de matemáticas entre las mujeres y los hombres.

¿Qué tan fuerte es la relación entre matemáticas y lectura crítica?

Para poder mirar la relación entre matemáticas y lectura critica primero se tiene que calcular la correlación entre estas dos materias esto se hace con la función cor().

```
# Correlación entre Matemáticas y Lectura Crítica
correlacion_mat_lectura <- cor(
  datos_icfes_limpios2$PUNT_MATEMATICAS,
  datos_icfes_limpios2$PUNT_LENGUAJE,
  use = "complete.obs"
)

correlacion_mat_lectura
```

```
## [1] 0.472713
```

El coeficiente de correlación de Pearson se caracteriza por +1 esto indica que la correlación es perfecta positiva, las dos variables lenguaje y matemáticas suben

juntas, si es 0 no hay relación lineal y si es -1 hay una correlación negativa esto es cuando ambas variables bajan. En este caso el valor de correlación es de 0.472713 esto indica que al ser positivo entre mayor puntaje de lectura mayor es el puntaje de matemáticas. Por ese motivo se puede decir que existe una correlación positiva entre las materias de matemáticas y lenguaje, esto permite entender que si un estudiante tiene un buen rendimiento en un área también tiende a hacer bueno en la otra área.

5. Visualización socioeconómica

verificar tipo de respuestas

Antes de crear los gráficos se va a analizar que tipos de respuestas tienen estas variables con la función unique().

```
unique(datos_icfes_limpios2$FAMI_INTERNET)
```

```
## [1]  0  1 NA
```

```
unique(datos_icfes_limpios2$FAMI_COMPUTADOR)
```

```
## [1]  3  0  1 NA
```

```
unique(datos_icfes_limpios2$FAMI_CELULAR)
```

```
## [1] 1 0
```

```
unique(datos_icfes_limpios2$FAMI_SERVICIO_TELEVISION)
```

```
## [1]  1  0 NA
```

```
unique(datos_icfes_limpios2$FAMI_TELEFONO_FIJO)
```

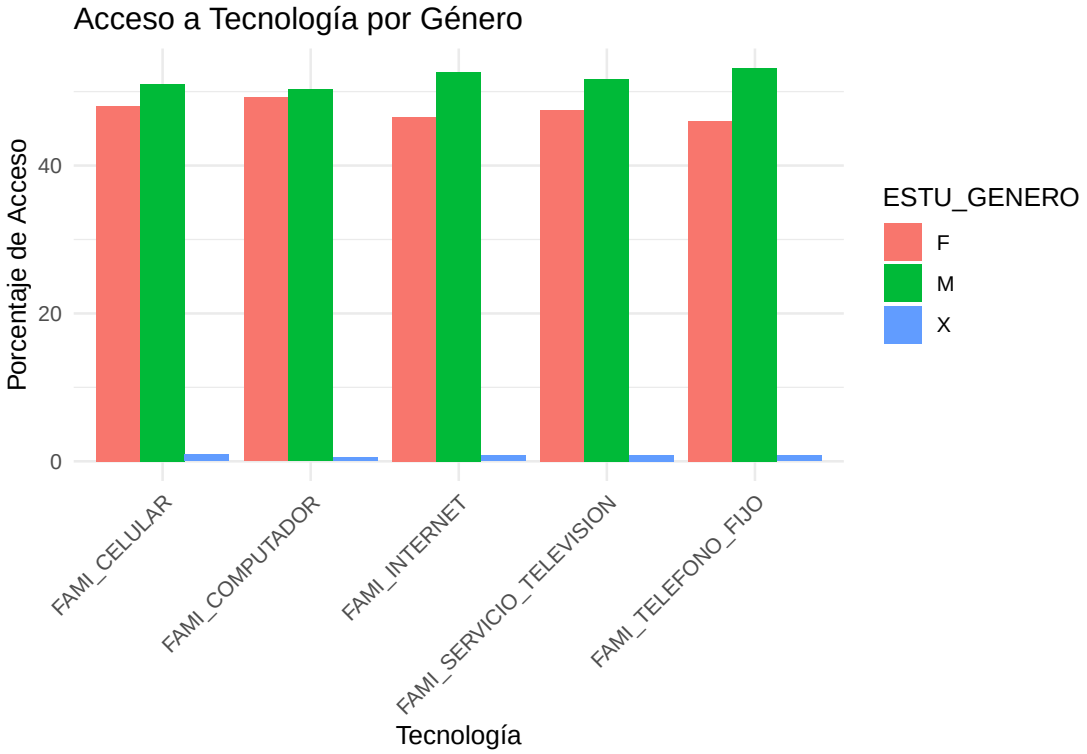
```
## [1]  0  1 NA  2  4
```

Los resultados obtenidos indican que las respuestas están configuradas de manera numérica, se analizó y se concluyó que los valores 1 son "Si" y las variables con 0 son "No", las demás variables con 2,3 y 4 tal vez hacen referencia a la cantidad, pero para el análisis se van a descartar y solo tomar en cuenta los valores con 1.

Gráfico de acceso tecnológico por genero

```
library(tidyr)
tecnologia_genero <- datos_icfes_limpios2 %>%
  select(ESTU_GENERO, FAMI_INTERNET, FAMI_COMPUTADOR, FAMI_CELULAR,
         FAMI_SERVICIO_TELEVISION, FAMI_TELEFONO_FIJO) %>%
  pivot_longer(cols = -ESTU_GENERO, names_to = "Tecnologia",
  values_to = "Acceso") %>%
  filter(Acceso == 1) %>%
  group_by(ESTU_GENERO, Tecnologia) %>%
  summarise(Acceso_Porcentaje = n(), .groups = "drop") %>%
  group_by(Tecnologia) %>%
  mutate(Porcentaje = 100 * Acceso_Porcentaje / sum(Acceso_Porcentaje))

ggplot(tecnologia_genero, aes(x = Tecnologia, y = Porcentaje, fill = ESTU_GENERO))
  geom_col(position = "dodge") +
  labs(title = "Acceso a Tecnología por Género",
       x = "Tecnología", y = "Porcentaje de Acceso") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```

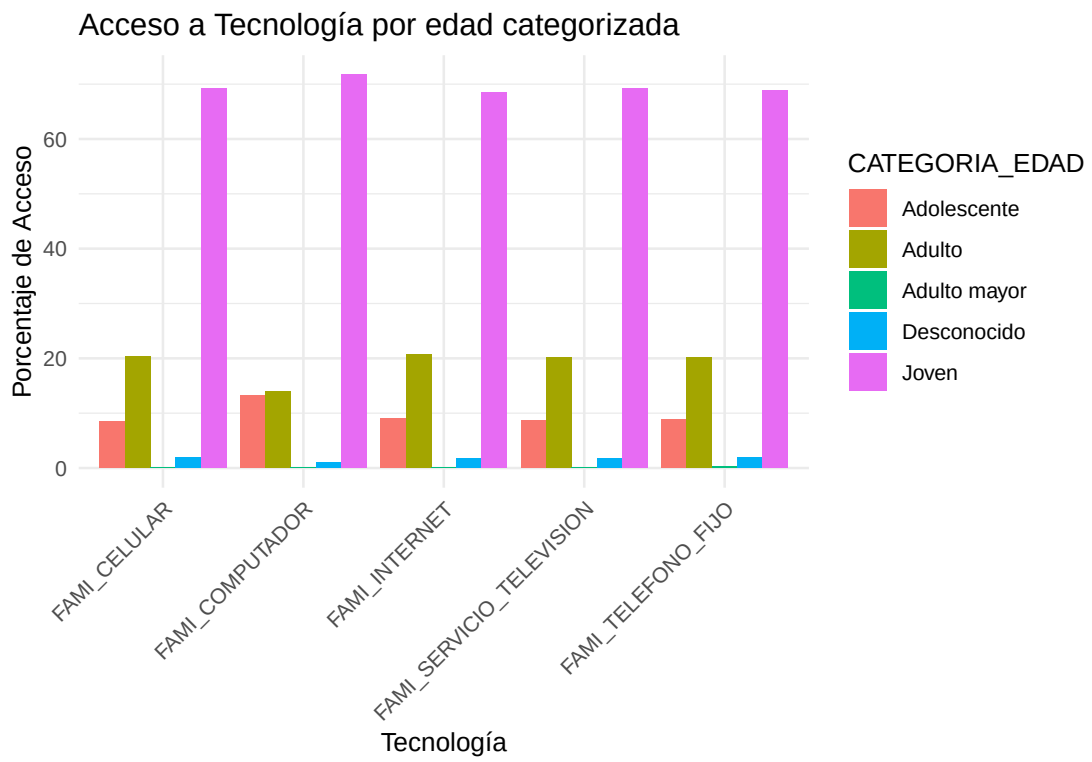


Los resultados obtenidos al analizar las variables por genero de:
FAMI_CELULAR
FAMI_INTERNET
FAMI_COMPUTADOR
FAMI_SERVICIO_TELEVISION
FAMI_TELEFONO_FIJO
El gráfico arrojo que los hombres (M)con el color verde, tienen acceso a todas las herramientas tecnológicas, las herramientas que más se destacan son las de acceso a Internet y a teléfono fijo. Por otro lado, las mujeres (F) con el color rosado, también tienen acceso a todas las herramientas como los hombres, pero en comparación hay una brecha con respecto a las mujeres. El grupo X de color azul es una población muy pequeña por ese motivo no es representativa para este análisis. Del grafico se puede concluir que el celular es el medio con mayor acceso en todos los géneros, el computador y el internet también tiene alto nivel de acceso pero una pequeña brecha de géneros y el teléfono fijo es el de menor acceso por la preferencia del uso del celular.

Gráfico de acceso tecnologico por edad categorizada

```
tecnologia_edad <- datos_icfes_limpios2 %>%
  select(CATEGORIA_EDAD, FAMI_INTERNET, FAMI_COMPUTADOR, FAMI_CELULAR,
         FAMI_SERVICIO_TELEVISION, FAMI_TELEFONO_FIJO) %>%
  pivot_longer(cols = -CATEGORIA_EDAD, names_to = "Tecnologia",
  values_to = "Acceso") %>%
  filter(Acceso == 1) %>%
  group_by(CATEGORIA_EDAD, Tecnologia) %>%
  summarise(Acceso_Porcentaje = n(), .groups = "drop") %>%
  group_by(Tecnologia) %>%
  mutate(Porcentaje = 100 * Acceso_Porcentaje / sum(Acceso_Porcentaje))

ggplot(tecnologia_edad, aes(x = Tecnologia, y = Porcentaje, fill = CATEGORIA_EDAD)
  geom_col(position = "dodge") +
  labs(title = "Acceso a Tecnología por edad categorizada",
       x = "Tecnología", y = "Porcentaje de Acceso") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```



La variable creada en el apartado 2 denominada como CATEGORIA_EDAD se agrupo de la siguiente manera:
15 años o menos de color rojo
16 - 17 años de color verde
18 - 19 años de color turquesa
20 o más años de color morado

De los resultados se puede decir que de la variable FAMI_CELULAR el acceso aumenta con la edad y que el grupo de 15 años o menos son los que tienen el menor acceso. De la variable FAMI_COMPUTADOR se puede deducir que la categoría azul de 18 a 19 tienen más acceso como con el grupo morado de 20 o más años

Los resultados de FAMI_INTERNET indican que el grupo morado tiene el mayor acceso y los grupos menores de 15 o menos y de 16 y 17 tienen menos acceso.

La variable de televisión y de teléfono fijo casi tiene la misma distribución de datos el grupo de 20 o más tiene el mayor acceso y los demás siguen el patrón de menor a mayor con la edad.

Se puede decir que la edad es un factor determinante en el acceso a la tecnología ya que las personas con 20 o más años tienen más acceso a todas las tecnologías analizadas, los estudiantes del grupo de 15 años o menos tienen un acceso muy bajo a las tecnologías siendo determinante en sus resultados de la prueba del ICFES.

¿Qué relación puede existir entre acceso a servicios y el rendimiento académico?

Como se pudo apreciar en los análisis de las variables socioeconómicas el acceso a servicios tecnológicos es muy variado y si se analizan los resultados los hombres (M) tienen mayor acceso y tal vez por este motivo se puede dar razón a los anteriores análisis donde los hombres tienen mayores puntajes en las áreas de matemáticas que las mujeres ya que si se analizan que al tener mayor facilidades para usar las herramientas tecnológicas se puede aprender y pulir más las habilidades para el examen, así que la diferencia de acceso pudo ser un factor influyente en el rendimiento del examen.

Por otro lado, si se analiza la disponibilidad de acceso tecnológico por edad categorizada los mayores de 20 años tienen más acceso que los menores de 15 años y los de 18 y 19 años tiene un acceso mayor pero no tanto como los de 20 años. En los anteriores análisis se detectó que la mayoría de estudiantes son mayores de 18 años entonces es la población que no tiene tantas facilidades de acceso a la tecnología por ende no tienen fácil acceso a información para estudiar para el examen por este motivo se puede concluir que el acceso a los servicios tecnológicos afectan al rendimiento de los estudiantes a la hora de presentar el examen del ICFES.