Reporte Cambios a `BD_huella.txt`

Universidad del Valle - Probabilidad y Estadística 761001C-F02 (6)

Juan Camilo Narváez Tascón - 2140112-3743

Óscar David Cuaical, 2270657-3743

Inicialización y Carga de Datos:

Se cargaron los paquetes necesarios en R para el manejo de datos y visualizaciones. Luego, se leyó la base de datos `BD_huella.txt` usando `read.csv`, lo que implicó la importación de la base de datos de la huella hídrica de estudiantes en una institución educativa.

Transformación y Estandarización de Datos Cualitativos:

Las variables categóricas `genero`, `zona` y `grado` fueron procesadas para estandarizar su formato. Esto incluyó la conversión de caracteres a codificación ASCII para evitar problemas con caracteres especiales o acentos, y la transformación de todos los textos a minúsculas para mantener la consistencia. Se reasignaron los valores textuales a codificaciones numéricas predefinidas para su uso en análisis estadísticos, donde cualquier dato no reconocido fue marcado como `NA`, representando un valor faltante o no disponible. Esto por medio de la tabla

Atributo	Valor
genero	1=femenino; 2=masculino
zona	1=urbano; 2=rural
grado	6=sexto; 7=séptimo; 8=octavo, 9=noveno, 10=decimo, 11=once

Codificación y Factores, normalización de Componentes de Huella Hídrica:

Estas variables cualitativas fueron convertidas a factores, que en R son variables categóricas utilizadas en modelos estadísticos. Se asignaron etiquetas legibles para `genero`, `zona` y `grado`, proporcionando una representación clara de cada categoría. Se normalizaron los nombres de los componentes de la huella hídrica (`comp_HHD` y `comp_HHI`), asegurando que no hubiera inconsistencias debidas a mayúsculas, minúsculas o caracteres especiales.

Este paso se implementó teniendo en cuenta las ecuaciones matemáticas establecidas en `consistencia_matemática.txt`, que definen el dominio de cada propiedad. El código respectivo es:

```
datos <- datos %>%
  mutate(
    genero = iconv(as.character(genero), to = "ASCII//TRANSLIT"),
    zona = iconv(as.character(zona), to = "ASCII//TRANSLIT"),
    grado = iconv(as.character(grado), to = "ASCII//TRANSLIT"),
    genero = tolower(genero),
    zona = tolower(zona),
    grado = tolower(grado),
    genero = case_when(genero %in% c("femenino", "1") ~ "1",
                       genero %in% c("masculino", "2") ~ "2",
                       TRUE ~ NA_character_),
    zona = case_when(zona %in% c("urbano", "1") ~ "1",
                     zona %in% c("rural", "2") ~ "2",
                     TRUE ~ NA_character_),
    grado = case_when(grado %in% c("6", "sexto") ~ "6",
                      grado %in% c("7", "septimo") ~ "7",
                      grado %in% c("8", "octavo") ~ "8",
                      grado %in% c("9", "noveno") ~ "9",
                      grado %in% c("10", "decimo") ~ "10",
                      grado %in% c("11", "once") ~ "11",
                      TRUE ~ NA_character_),
    genero = factor(genero, levels = c("1", "2"), labels = c("femenino", "masculino")),
    zona = factor(zona, levels = c("1", "2"), labels = c("urbano", "rural")),
    grado = factor(grado, levels = c("6", "7", "8", "9", "10", "11"), labels = c("sexto", "séptimo", "octavo", "noveno", "décimo", "once"))
    comp_HHD = tolower(gsub("[._]", "_", comp_HHD)),
    comp_HHI = tolower(comp_HHI)
  )
```

Manejo de Datos Atípicos:

Se aplicó una función para identificar y manejar datos atípicos en las variables `HHD` y `per.hog`. Los datos atípicos son valores extremos que se desvían significativamente del resto de los datos y pueden sesgar el análisis. Se calculó un límite, conocido como el cerco superior, que es una medida estadística que ayuda a identificar estos valores extremos. Los valores que excedían este límite fueron reemplazados por `NA`, indicando que son atípicos y deben ser tratados con precaución en análisis subsiguientes.

```
# Función para calcular el cerco superior y contar datos atípicos, y luego imprimir el resultado
calcular_cerco_superior_y_conteo <- function(data, variable) {
    Q3 <- quantile(data[[variable]], 0.75, na.rm = TRUE)
    IQR <- IQR(data[[variable]], na.rm = TRUE)
    upper_bound <- Q3 + 1.5 * IQR
    conteo_atipicos <- sum(data[[variable]] > upper_bound, na.rm = TRUE)

    mensaje <- paste("Cerco superior de", variable, ":", upper_bound, "; Número de datos atípicos:", conteo_atipicos)
    print(mensaje)
}

# Aplicar la función a las variables de interés
calcular_cerco_superior_y_conteo(datos, "HHD")
calcular_cerco_superior_y_conteo(datos, "HHI")
calcular_cerco_superior_y_conteo(datos, "per.hog")
```

Imputación de Datos Faltantes:

Se calculó la media de `HHD` y `per.hog`, excluyendo los valores `NA`, y luego se utilizó este promedio para reemplazar los valores faltantes. Este paso es crucial porque los modelos estadísticos generalmente no pueden manejar valores `NA`, y la imputación con la media es una técnica estándar para preservar la estructura general de los datos.

```
# Función para reemplazar datos atípicos con NA
reemplazar_atipicos_con_NA <- function(data, variable) {
    Q3 <- quantile(data[[variable]], 0.75, na.rm = TRUE)
    IQR <- IQR(data[[variable]], na.rm = TRUE)
    upper_bound <- Q3 + 1.5 * IQR

    # Reemplazar datos que exceden el cerco superior con NA
    data[[variable]][data[[variable]] >= upper_bound] <- NA
    return(data)
}

# Aplicar la función a HHD y per.hog
datos <- reemplazar_atipicos_con_NA(datos, "HHD")
datos <- reemplazar_atipicos_con_NA(datos, "per.hog")

# Imputar valores faltantes para HHD y per.hog con la media de cada variable
mean_HHD <- mean(datos$HHD, na.rm = TRUE)
mean_per_hog <- mean(datos$per.hog, na.rm = TRUE)

datos$HHD[is.na(datos$HHD)] <- mean_HHD
datos$per.hog[is.na(datos$per.hog)] <- mean_per_hog
```

Modelado y Estimación:

Se ajustaron modelos de regresión lineal para `HHD` y `HHI` utilizando las variables `edad`, `genero`, `zona`, `grado` y `per.hog` como predictores. Estos modelos se usaron para predecir valores para `HHD` y `HHI` cuando los datos originales eran faltantes o no válidos (menores o iguales a cero). Los valores predichos fueron redondeados para mantener la coherencia con la naturaleza de los datos originales, que presumiblemente son enteros o cuentas discretas.

```
# Modelos de regresión lineal para HHD y HHI
modelo_HHD <- lm(HHD ~ edad + genero + zona + grado + per.hog, data = datos)
modelo_HHI <- lm(HHI ~ edad + genero + zona + grado + per.hog, data = datos)

# Aplicación de los modelos para estimar HHD y HHI
datos <- datos %>%
  mutate(
    HHD = ifelse(is.na(HHD) | HHD <= 0, predict(modelo_HHD, newdata = datos), HHD),
    HHI = ifelse(is.na(HHI) | HHI <= 0, predict(modelo_HHI, newdata = datos), HHI)
```

```
  ) %>%
  mutate(
    HHD = round(HHD),
    HHI = round(HHI)
  )
```

Limpieza Adicional y Redondeo:

Además, se realizó una limpieza adicional de la variable `per.hog`, reemplazando los valores no válidos o faltantes con la media calculada y redondeando todos los valores para garantizar que los datos sean discretos y manejables.

```
# Limpieza de per.hog
datos$per.hog <- ifelse(is.na(datos$per.hog) | datos$per.hog <= 0, mean_per_hog, datos$per.hog)
datos$per.hog <- round(datos$per.hog)
```

Validación de Datos:

Se aplicaron reglas de validación externas desde `consistencia.txt` para identificar cualquier otra violación de las reglas de datos que pudieran haber quedado después de la limpieza. Esto ayuda a garantizar que los datos estén en un formato adecuado y sean consistentes con las expectativas del análisis.

```
# Reglas de validación
rules <- editrules::editfile("consistencia.txt")
Valid_Data <- editrules::violatedEdits(rules, datos)
summary(Valid_Data)
```

Almacenamiento de Datos Limpios:

Finalmente, se preparó el código para guardar el conjunto de datos limpio en un nuevo archivo `clean_huella.txt`, lo que permite una fácil reutilización y análisis en el futuro. Este paso concluye el proceso de limpieza y preparación de datos, asegurando que el conjunto de datos esté listo para análisis más avanzados.

```
ruta <- "../Data/clean_huella.txt"
write.table(datos, file = ruta, sep = "\t", row.names = FALSE, na = "")
```