Laboratorio, Preprocesamiento de datos con R, Probabilidad y Estadística

Juan Camilo Narváez Tascón, 2140112-3743

Óscar David Cuaical, 2270657-3743

Fecha creación: 21-11-23; Última modificación: 3-12-23

Historia

Se desea caracterizar la huella hídrica de una institución de educación secundaria, se dispone de la base de datos `data/BD_huella.txt` , la cual contiene información de los estudiantes como: edad, género, zona, grado escolar, cantidad de HH directa e indirecta en m^3/año (`HHD` y `HHI`), el mayor componente de la HH directa e indirecta (`comp_HHD` y `comp_HHI`), y el número de personas que habitan en el hogar (`per.hog`).

Las variables cualitativas deben seguir la siguiente codificación:

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

1 Preprocesamiento y Limpieza de Datos

1.1 Verificación técnica de datos

Al revisar `BD_huella.txt` podremos notar que los titulares de los encabezados son correspondientes a los datos de cada propiedad, guardando una relación entre ellos, así que de manera general contamos con una base de datos correcta. También podremos notar que hace falta un preprocesamiento debido a datos faltantes o formatos no establecidos para mantener una consistencia.

```
##   ID edad   genero   zona grado HHD  HHI comp_HHD comp_HHI per.hog
## 1  1   18 femenino URBANO    10 152 1848 Uso.baño    Carne       3
## 2  2   11 femenino      1     6 117 1387 Uso_baño    Carne       2
## 3  3   11        1 Urbano sexto 276  567 USO.BAÑO    Carne       7
## 4  4   12 Femenino  Rural SEXTO 273 1356 Uso.baño    Carne       5
## 5  5   11 Femenino Urbano     6  92 1344 Uso.baño    Carne       1
## 6  6   14        2 Urbano     7  NA 1344 Uso.baño    Carne       7
```

1.2 Ecuaciones de consistencia

Empezamos aplicando la codificación definida para las variables cualitativas, y factorizándolas para futuras aplicaciones; la codificación será etiquetada para una mejor lectura. Consiguiente se define un formato general para las propiedades, estableciendo así una consistencia entre estas, se deben considerar cambios sutiles como tildes o mayúsculas. Las reglas matemáticas que permitirán esta consistencia están dadas por:

1. **Rango de Edad**: La edad de los estudiantes debe ser razonable para el nivel de educación secundaria: $\{\forall \texttt{edad} \in \mathbb{Z} \mid 10 \leqslant \texttt{edad} \leqslant 20\}$

2. **Consistencia de Género**: El género debe ser femenino (1) o masculino (2), tras la transformación realizada:

$$\{\forall \texttt{genero} \in \{\texttt{'1', '2'}\}$$
$$| \ \texttt{genero\_i} := \texttt{'1'} \Leftrightarrow (\texttt{genero\_i == 'femenino'} \vee \texttt{'1'})$$
$$\oplus \ \texttt{genero\_i} := \texttt{'2'} \Leftrightarrow (\texttt{genero\_i == 'masculino'} \vee \texttt{'2'})\}$$

3. **Consistencia de Zona**: La zona debe ser urbana (1) o rural (2), de acuerdo con la recodificación:

$$\{\forall \texttt{zona} \in \{\texttt{'1', '2'}\}$$
$$| \ \texttt{zona\_i} := \texttt{'1'} \Leftrightarrow (\texttt{zona\_i == 'urbano'} \vee \texttt{'1'})$$
$$\oplus \ \texttt{zona\_i} := \texttt{'2'} \Leftrightarrow (\texttt{zona\_i == 'rural'} \vee \texttt{'2'})\}$$

4. **Consistencia de Grado Escolar**: El grado escolar debe estar entre sexto (6) y once (11), según la recodificación.

$$\{\forall \texttt{grado} \in \{\texttt{'6', '7', '8', '9', '10', '11'}\}$$
$$| \ \texttt{grado\_i} := \texttt{'6'} \Leftrightarrow (\texttt{grado\_i == 'sexto'} \vee \texttt{'6'})$$
$$\oplus \ \texttt{grado\_i} := \texttt{'7'} \Leftrightarrow (\texttt{grado\_i == 'septimo'} \vee \texttt{'7'})$$
$$\oplus \ \texttt{grado\_i} := \texttt{'8'} \Leftrightarrow (\texttt{grado\_i == 'octavo'} \vee \texttt{'8'})$$
$$\oplus \ \texttt{grado\_i} := \texttt{'9'} \Leftrightarrow (\texttt{grado\_i == 'noveno'} \vee \texttt{'9'})$$
$$\oplus \ \texttt{grado\_i} := \texttt{'10'} \Leftrightarrow (\texttt{grado\_i == 'decimo'} \vee \texttt{'10'})$$
$$\oplus \ \texttt{grado\_i} := \texttt{'11'} \Leftrightarrow (\texttt{grado\_i == 'once'} \vee \texttt{'11'})\}$$

5. **Consistencia en `comp_HHD` y `comp_HHI`** ; no deben haber variaciones textuales de un mismo componente, por ende se usan minúsculas. Al tratarse de variables nominales no se hace uso de codificación y se respeta el formato establecido:
$\{\forall \texttt{comp\_HHD} \in \{\texttt{'uso\_baño','lavado\_ropa','uso\_cocina',...,}$
$\texttt{'}\langle\texttt{actividad}\rangle\langle\texttt{separador}\rangle\langle\texttt{lugar}\rangle\texttt{'}\} \ | \ \langle\texttt{separador}\rangle \ := \texttt{'\_'}\}$
$\{\forall \texttt{comp\_HHI} \in \{\texttt{'carne','fruta','cafe',...,}\langle\texttt{componente}\rangle\}\}$

6. **Validez de Huella Hídrica Directa e Indirecta (`HHD` y `HHI`)**: Estos valores deben ser positivos y lógicos. $Dominio = \{\texttt{HHD, HHI} \in \mathbb{R}^{+} \mid \texttt{HHD} > 0 \wedge \texttt{HHI} > 0\}$

   Además, si `HHI` , `HHD == ''` (o `NA`) se remplaza por una regresión lineal, esto es:

$$\texttt{HHD\_i} \notin Dominio \Rightarrow \texttt{HHD\_i} := f_{RL}(\texttt{HHD\_i})$$
$$\texttt{HHI\_i} \notin Dominio \Rightarrow \texttt{HHI\_i} := f_{RL}(\texttt{HHI\_i})$$

   $f_{RL}(x)$ representa la función de regresión lineal aplicada para estimar los valores faltantes de `HHD` o `HHI` , esto es:

$$f_{RL}(x) = a + (b_1 \times x_1) + (b_2 \times x_2) + \ldots + (b_n \times x_n) + \varepsilon$$

   Donde $a es el término de intercepción, que representa el valor esperado de f_{RL} cuando todas las x_i son cero. b_i son los coeficientes de regresión asociados con cada variable independiente, que representan el cambio esperado en f_{RL} por una unidad de cambio en x_i, manteniendo las demás x constantes. ε es el término de error, que representa la variación en f_{RL} no explicada por las variables independientes. Se asume que ε tiene una distribución normal con media cero y varianza constante (homocedasticidad).

7. **Consistencia en Número de Personas en el Hogar (`per.hog`)**: Debe ser un número positivo y lógico.

$$Dominio = \{\texttt{per.hog} \in \mathbb{Z} \mid \texttt{per.hog} > 0\}$$

Por ende, en caso de `per.hog == ''` (o `NA`) se remplaza por la media, esto es:
$\texttt{per.hog\_i} \notin Dominio \Rightarrow \texttt{per.hog\_i} := \bar{x}(textttper.\,hog)$ Donde
$\bar{x}(\texttt{per.hog}) = (\Sigma \texttt{per.hog\_i})/N$

Todas estas reglas son implementadas en `consistencia.txt`, adaptadas a la librería `editrules`, la cual se usa para verificar que los datos cumplan con las reglas establecidas.

1.3 Aplicación de reglas de consistencia

Si omitimos parte de la regla 6 y 7 de momento, las cuales corresponden a llenar datos faltantes y solo tenemos en cuenta el dominio establecido, al aplicar estas reglas tenemos:

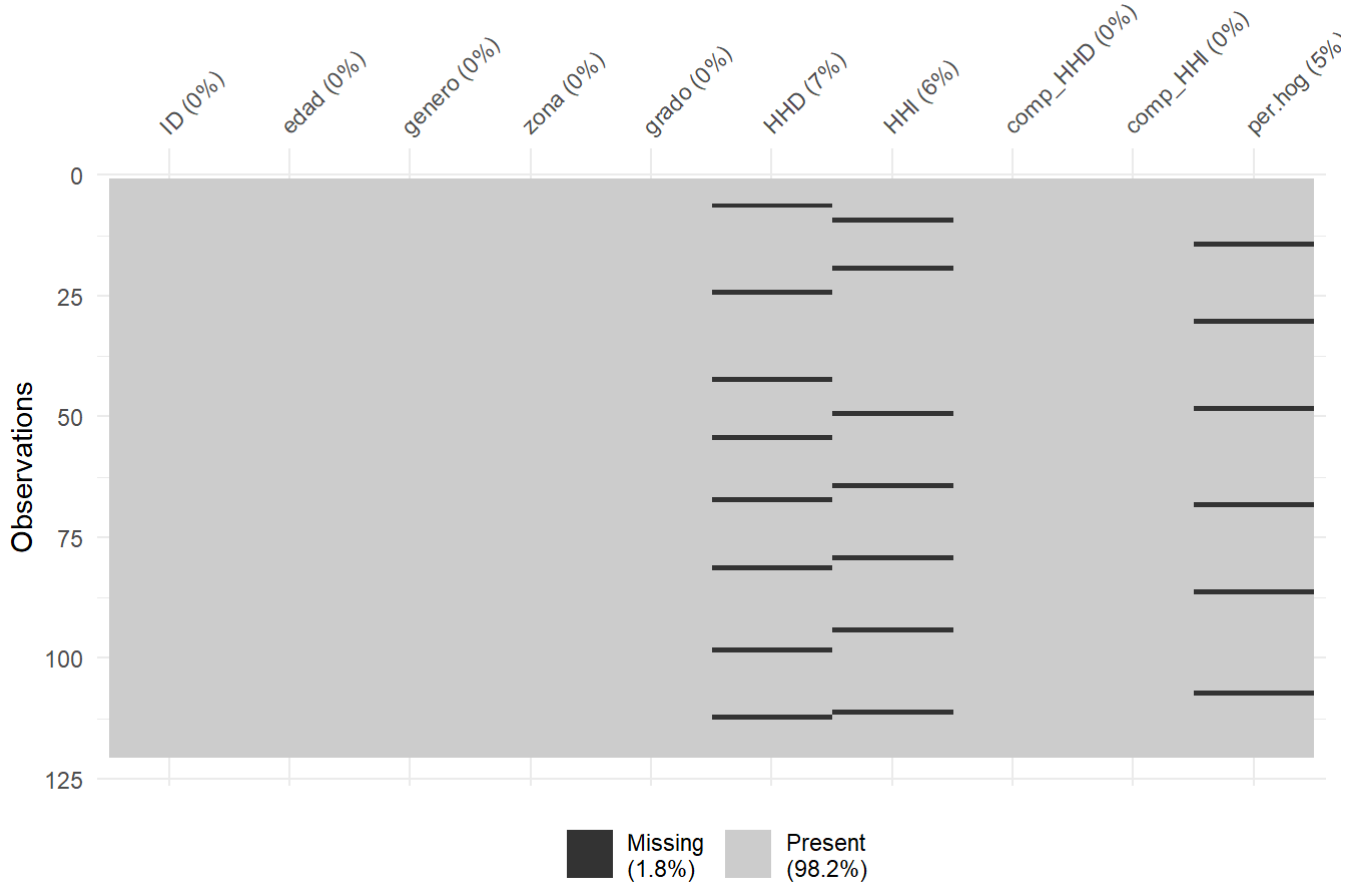
```
##   ID edad    genero   zona   grado HHD  HHI comp_HHD comp_HHI per.hog
## 1  1   18  femenino urbano  décimo 152 1848 uso_baño    carne       3
## 2  2   11  femenino urbano   sexto 117 1387 uso_baño    carne       2
## 3  3   11  femenino urbano   sexto 276  567 uso_baño    carne       7
## 4  4   12  femenino  rural   sexto 273 1356 uso_baño    carne       5
## 5  5   11  femenino urbano   sexto  92 1344 uso_baño    carne       1
## 6  6   14 masculino urbano séptimo  NA 1344 uso_baño    carne       7
```

De momento se omite la muestra de la tabla completa debido a su tamaño, sin embargo con esta muestra podemos observar una mayor consistencia con respecto a la muestra anterior, en el `ID` 6 todavía se observa un dato faltante `NA`, pero podemos notar el correcto etiquetado de cada propiedad. Si queremos ver el resumen de los datos validos usando `editrules` tenemos:

```
## Edit violations, 120 observations, 0 completely missing (0%):
##
##  editname freq  rel
##      num4    1 0.8%
##
## Edit violations per record:
##
##  errors freq   rel
##       0   98 81.7%
##       1   22 18.3%
```

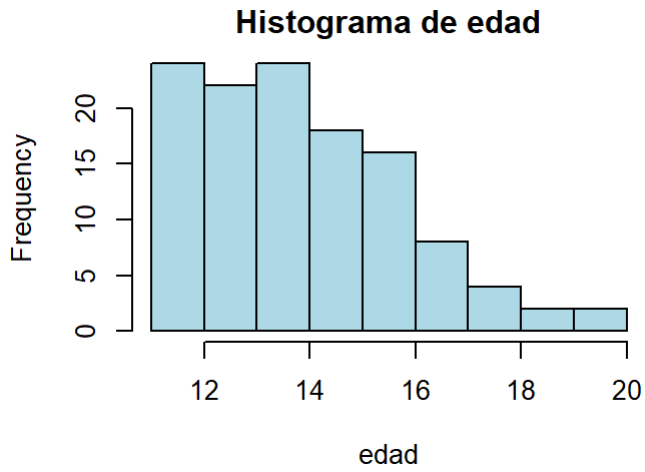
1.4 Estudio de datos faltantes

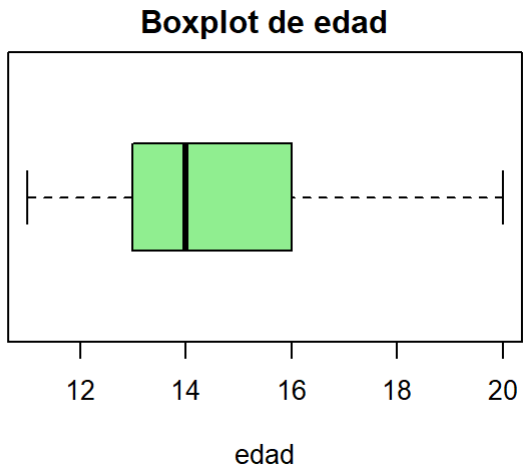
En el punto 1.3 observamos que las reglas establecidas aún no se cumplen, esto puede deberse a los datos faltantes de `HHD`, `HHI` y `per.hogar`, que constituyen el 1.8% de la totalidad de nuestros datos:

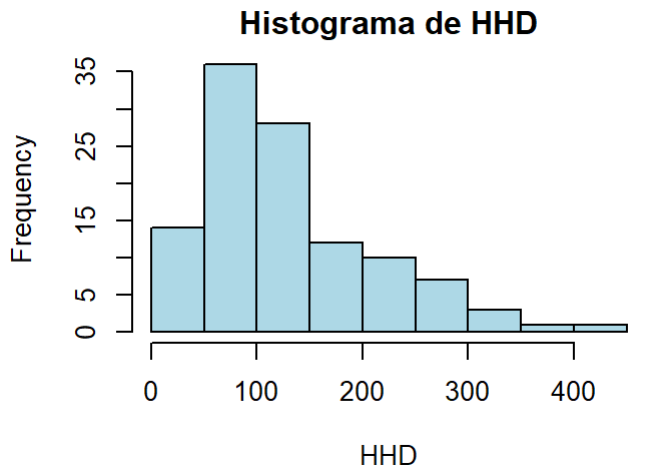


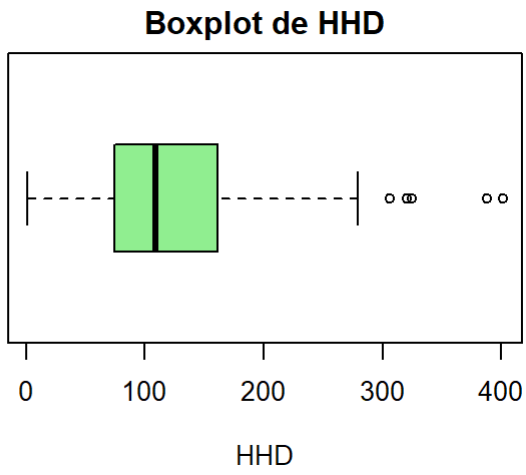
1.5 Estudio de datos atípicos

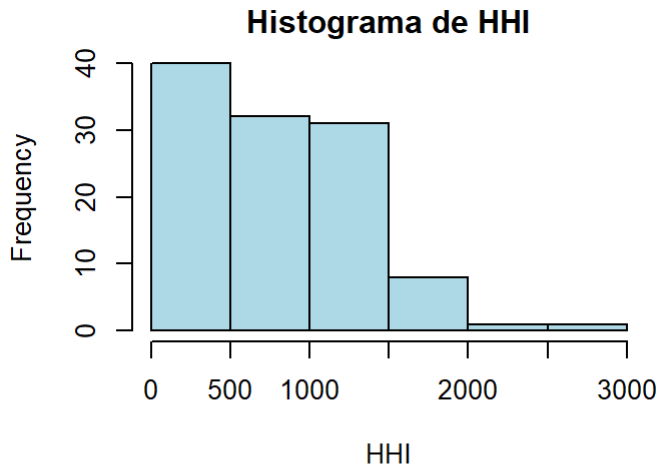
Es útil identificar los datos atípicos, de manera que al momento de remplazar los datos faltantes estos sean más confiables. Al visualizar las variables cuantitativas de manera genera podremos notar comportamientos relativos a datos atípicos y de esta manera se podrá estudiar dicha propiedad de manera más específica:

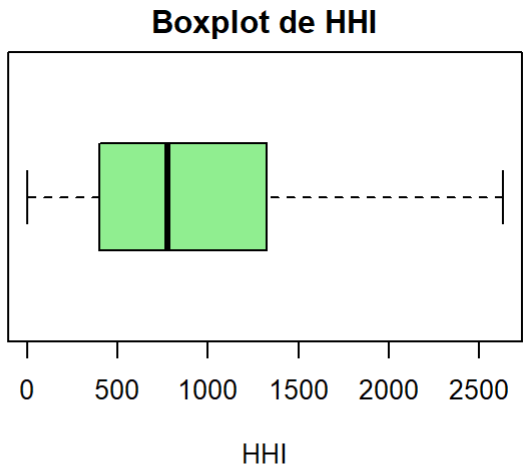


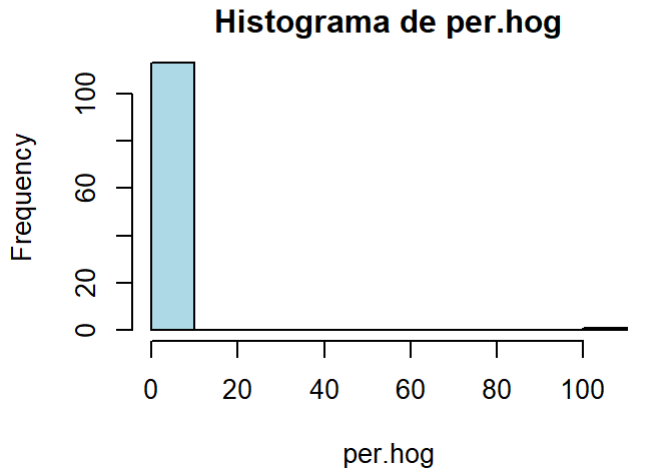


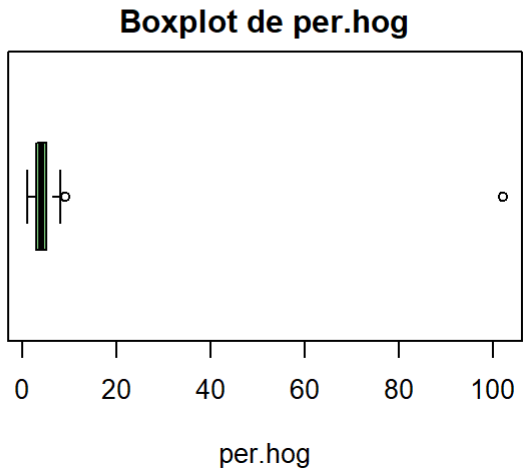












Podemos observar que la edad cumple con nuestras reglas establecidas, y no hay datos atípicos. Para `HHD`, `HHI` y `per.hogar` podemos intuir comportamientos respectivos a datos atípicos. Dado que todos nuestros datos son positivos podemos calcular solo los cercos superiores (estos pueden visualizarse en los boxplot) y

hacer un conteo de estos datos atípicos, permitiéndonos una mejor comprensión de los mismos:

```
## [1] "Cerco superior de HHD : 281.875 ; Número de datos atípicos: 5 ; Valor máximo: 401"
```

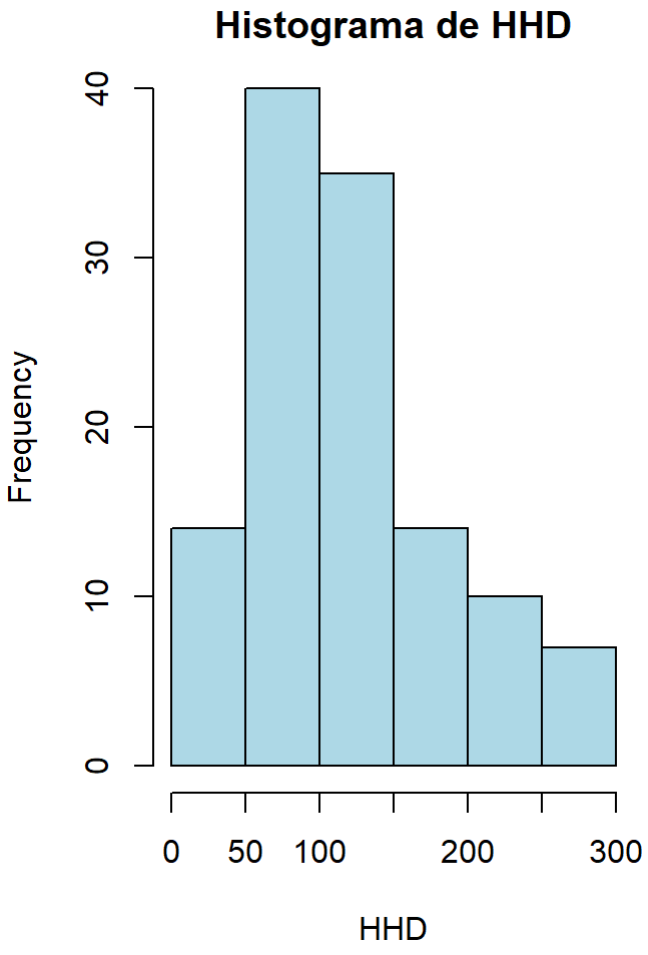
```
## [1] "Cerco superior de HHI : 2711 ; Número de datos atípicos: 0 ; Valor máximo: 2631"
```

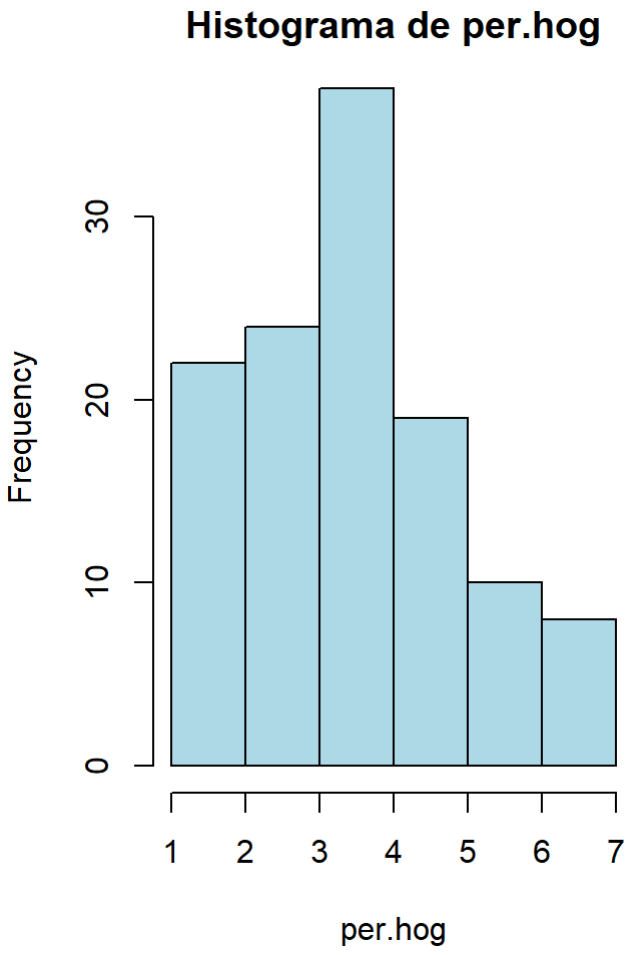
```
## [1] "Cerco superior de per.hog : 8 ; Número de datos atípicos: 2 ; Valor máximo: 102"
```

1.6 Corrección de datos

Identificados los datos atípicos procederemos a eliminarlos dejándolos como faltantes, para consiguiente implementar las ecuación de relleno establecidas en el punto 1.2.

Al volver a graficar `HHD` y `per.hogar` notaremos que ahora los datos se encuentran dentro del límite identificado por el cerco superior. En este punto los valores de los cercos pueden variar si se vuelven a cualcular debido a la eliminación de extremos, sin embargo podemos estar seguros de contar con datos menos variados:

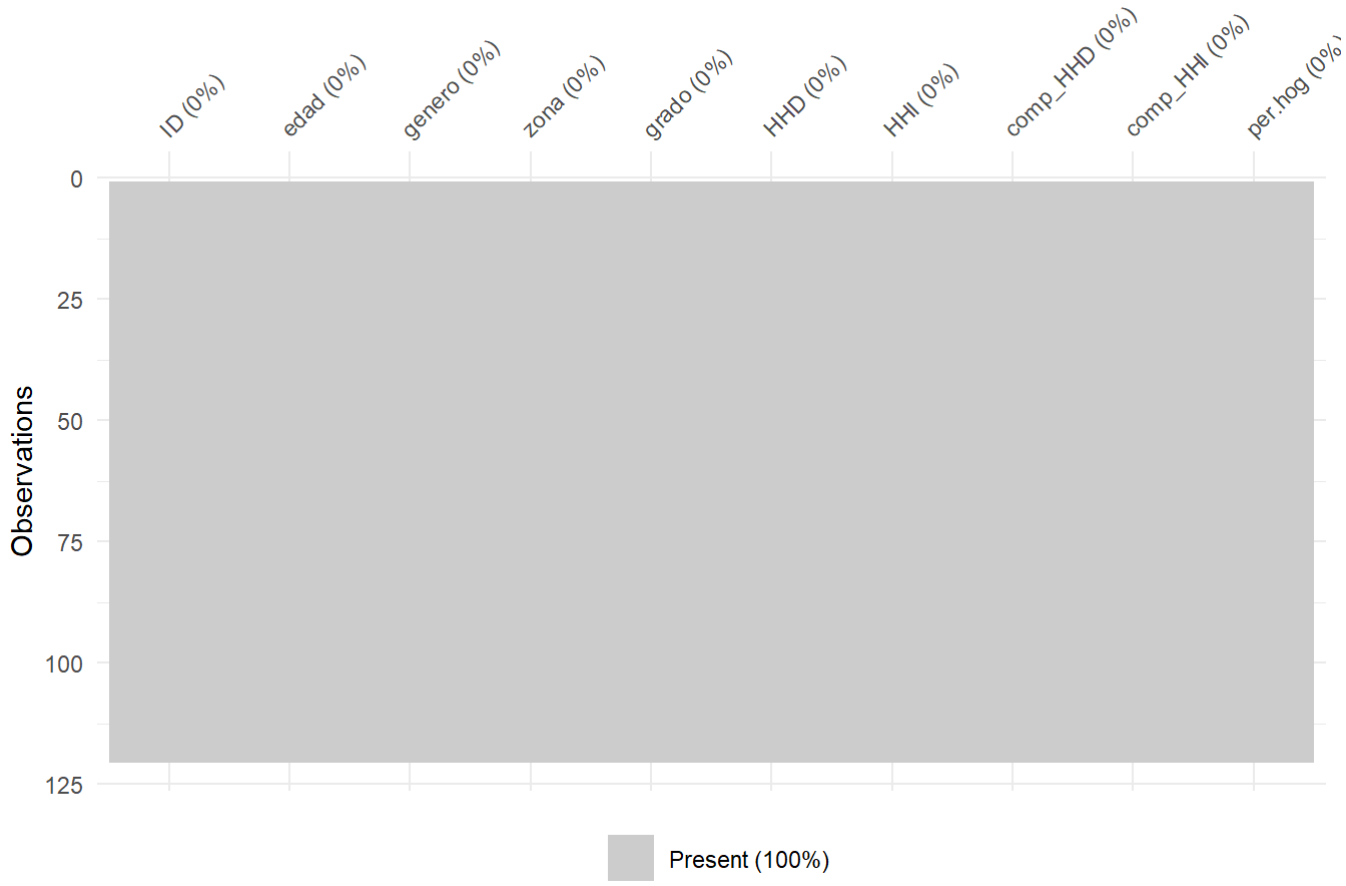





```
## [1] "Valor máximo de HHD: 279"
```

```
## [1] "Valor máximo de per.hogar: 7"
```

Hemos decidido remplazar los datos faltantes de `HHD` y `HHI` con una regresión lineal debido a que la naturaleza de estos valores suele variar entre números muy grandes, y su varianza suele ser mayor. Lo contrario sucede con `per.hogar` la cual puede ser remplazada simplemente por la media.



Podemos observar entonces que ya no tenemos datos faltantes, y si ahora volvemos a validar las reglas tendremos

```
## No violations detected, 0 checks evaluated to NA
```

```
## NULL
```

Hemos limpiado los datos de manera que cumplan con las reglas establecidas en el punto 1.2, eliminando datos atípicos y faltantes, y factorizando las variables cualitativas que lo solicitaban.

```
##   ID edad    genero   zona   grado HHD  HHI comp_HHD comp_HHI per.hog
## 1  1   18  femenino urbano  décimo 152 1848 uso_baño    carne       3
## 2  2   11  femenino urbano   sexto 117 1387 uso_baño    carne       2
## 3  3   11  femenino urbano   sexto 276  567 uso_baño    carne       7
## 4  4   12  femenino  rural   sexto 273 1356 uso_baño    carne       5
## 5  5   11  femenino urbano   sexto  92 1344 uso_baño    carne       1
## 6  6   14 masculino urbano séptimo  90 1344 uso_baño    carne       7
```

En esta muestra no se alcanzan a visualizar todos los cambios, esto por efectos prácticos y que el documento no quede tan largo, puede encontrar la base de datos limpia en `clean_huella.csv` .

Cabe mencionar que Excel puede tener problemas para la decodificación de archivos UTF-8 en csv, mostrando caracteres hispanos como símbolos faltantes, aun así hemos decidido mantener estos caracteres hispanos por fidelidad a la lengua española. R interpreta bien la decodificación de estos por medio del `fileEncoding="UTF-8"` , y si se quiere ver el archivo con toda su integridad se puede abrir `clean_huella.csv` desde un editor de código, o especificar en Excel la decodificación `65001: Unicode (UTF-8)` al importar la hoja, esta no se debe guardar ya que puede afectar la tabla original. Mencionamos todo esto para evitar problemas, y al mismo tiempo dejamos `clean_huella.txt` como alternativa.

1.8 Creación de varible de Huella Hídrica Total

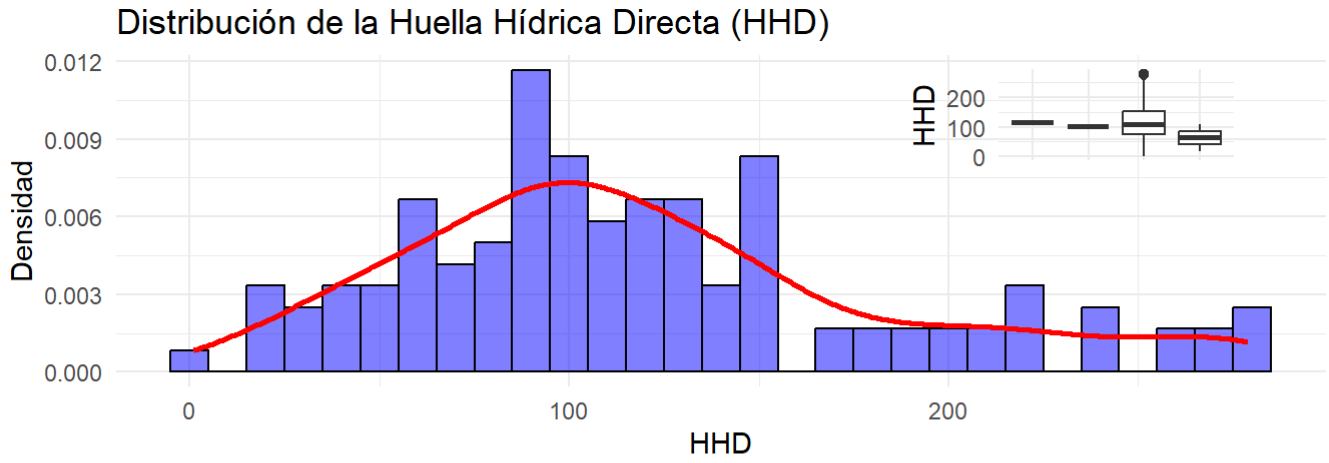
Procedemos a crear la variable correspondiente a la huella hídrica total `HHT` siendo la suma de la huella hídrica directa e indirecta, y una variable de clasificación `HHT_clas` con base siguientes condiciones:

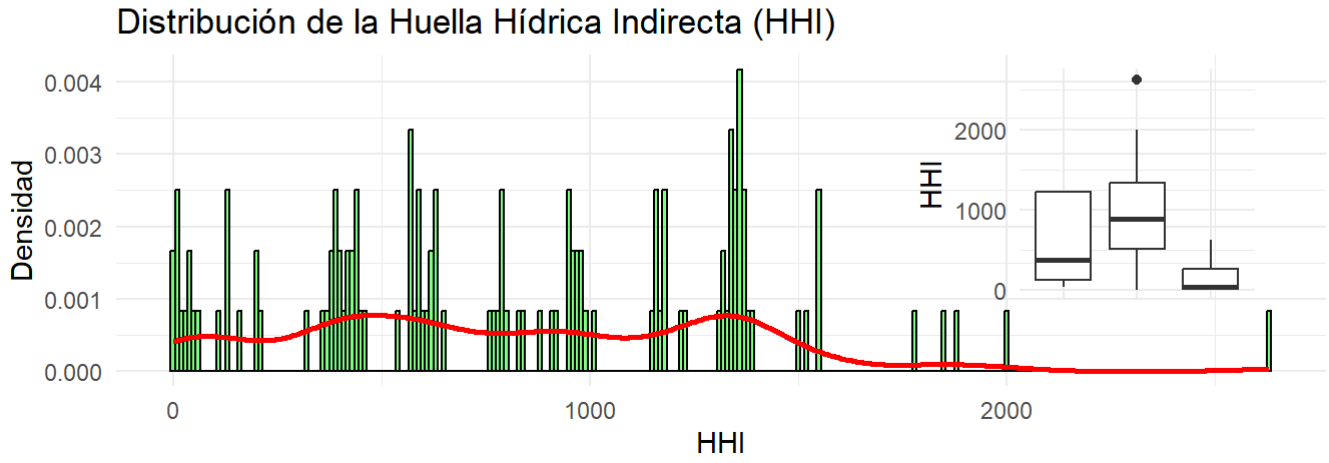
Grupo	Rango de clasificación
Bajo	si HHT ≤ 1789
Medio	si 1789 < HHT ≤ 1887
Alto	si HHT > 1887

2 Visualización de los datos

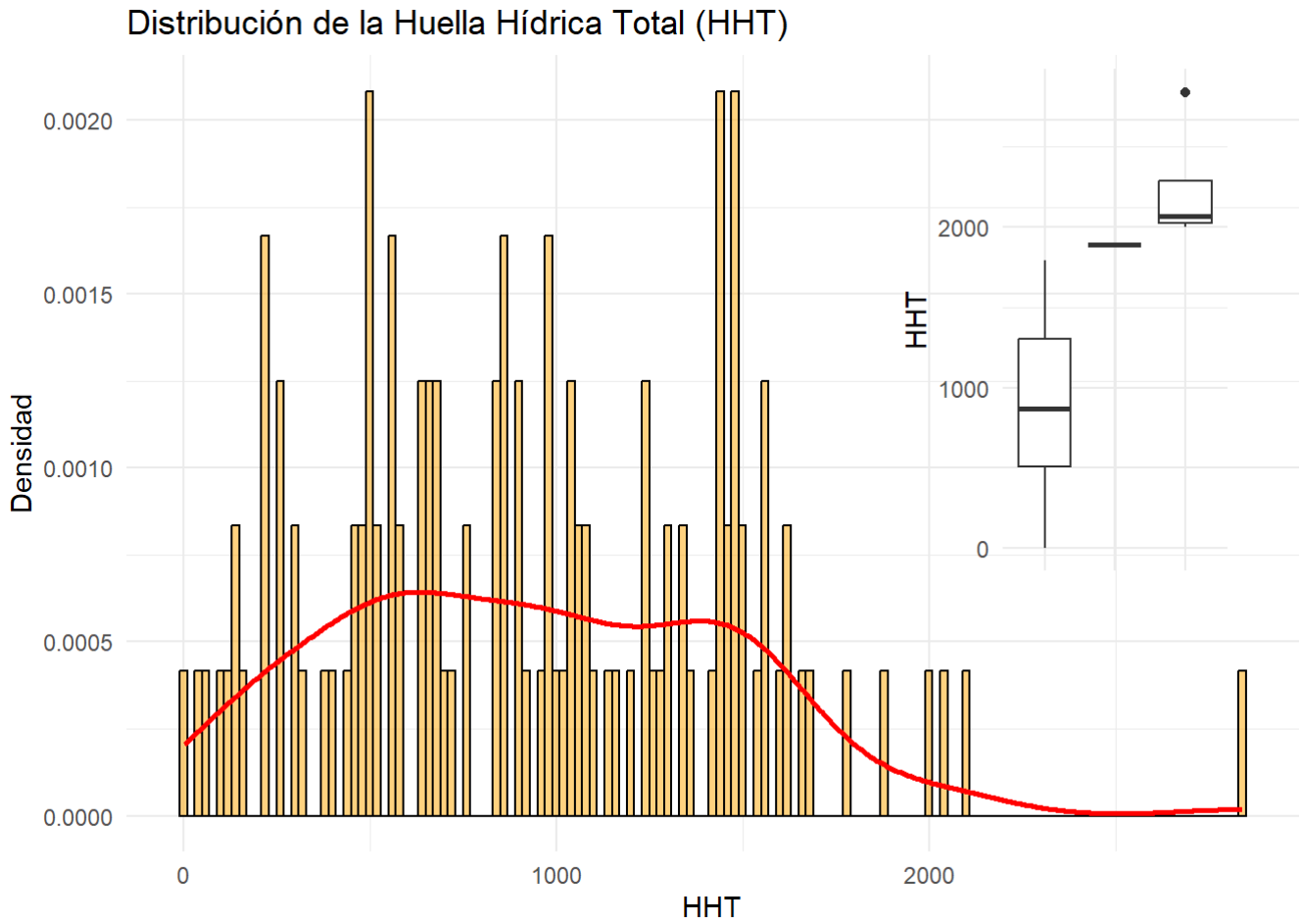
2.1 Gráficas de distribuciones de Huella Hídrica

La gráfica que representa la Huella Hídrica Directa (HHD) muestra una distribución que parece inclinarse hacia un único pico principal, lo que sugiere que la mayoría de los estudiantes tiene un nivel de consumo de agua directo relativamente homogéneo. La presencia de un pico central y dominante puede indicar que la mayoría de los estudiantes comparten prácticas similares en cuanto al uso directo de agua. En la esquina superior derecha podemos observar un boxplot respectivo al comp_HHD en el oden `lavado_ropa` , `riego_jardin` , `uso_baño` , `uso_cocina` respectivamente. Notamos diferencias notables entre los componentes del uso del agua, como el lavado de ropa, riego de jardín, uso de baño y uso de cocina. El uso de cocina y el uso del baño parecen tener una mediana más alta y una variabilidad mayor en el consumo de agua, indicada por el rango intercuartílico más amplio y los valores atípicos (para las nuevas medidas) presentes.





Por otro lado, la gráfica de la Huella Hídrica Indirecta (HHI) muestra claramente dos picos, lo que indica una distribución bimodal. Esto sugiere que hay dos patrones principales distintos de consumo de agua indirecto entre los estudiantes. El boxplot superior derecho con respecto a comp_HHI tiene el orden `café`, `carne`, `fruta` respectivamente. La carne, en particular, tiene una mediana considerablemente más alta y un rango más extenso (similar a la del café, pero con menos variación), lo que refleja un consumo de agua indirecto significativamente mayor asociado con este componente. Esto puede deberse al hecho de que la producción de carne generalmente requiere más recursos hídricos que otros alimentos. Los estudiantes que consumen más carne, por lo tanto, contribuyen a una huella hídrica indirecta más elevada.



La distribución de la Huella Hídrica Total (HHT) parece tener un rango amplio de valores con varias concentraciones a lo largo del espectro. La presencia de múltiples picos en el histograma sugiere que hay diferentes grupos de estudiantes con distintos niveles de consumo total de agua. Esto puede reflejar la diversidad en las prácticas de consumo y producción que afectan tanto la huella hídrica directa como la indirecta de los estudiantes.

El boxplot superior a la derecha para HHT es con respecto a los grupos `Bajo`, `Medio` y `Alto`, muestra una clara distinción entre los tres grupos. Los estudiantes clasificados en el grupo `Bajo` tienen una mediana inferior a los de los grupos `Medio` y `Alto`, lo que indica que sus prácticas generales de consumo de agua son menos intensivas. El grupo `Medio` presenta una variabilidad relativamente menor comparado con el grupo `Alto`, lo que podría sugerir que los estudiantes en esta categoría tienen hábitos de consumo de agua más consistentes entre ellos.

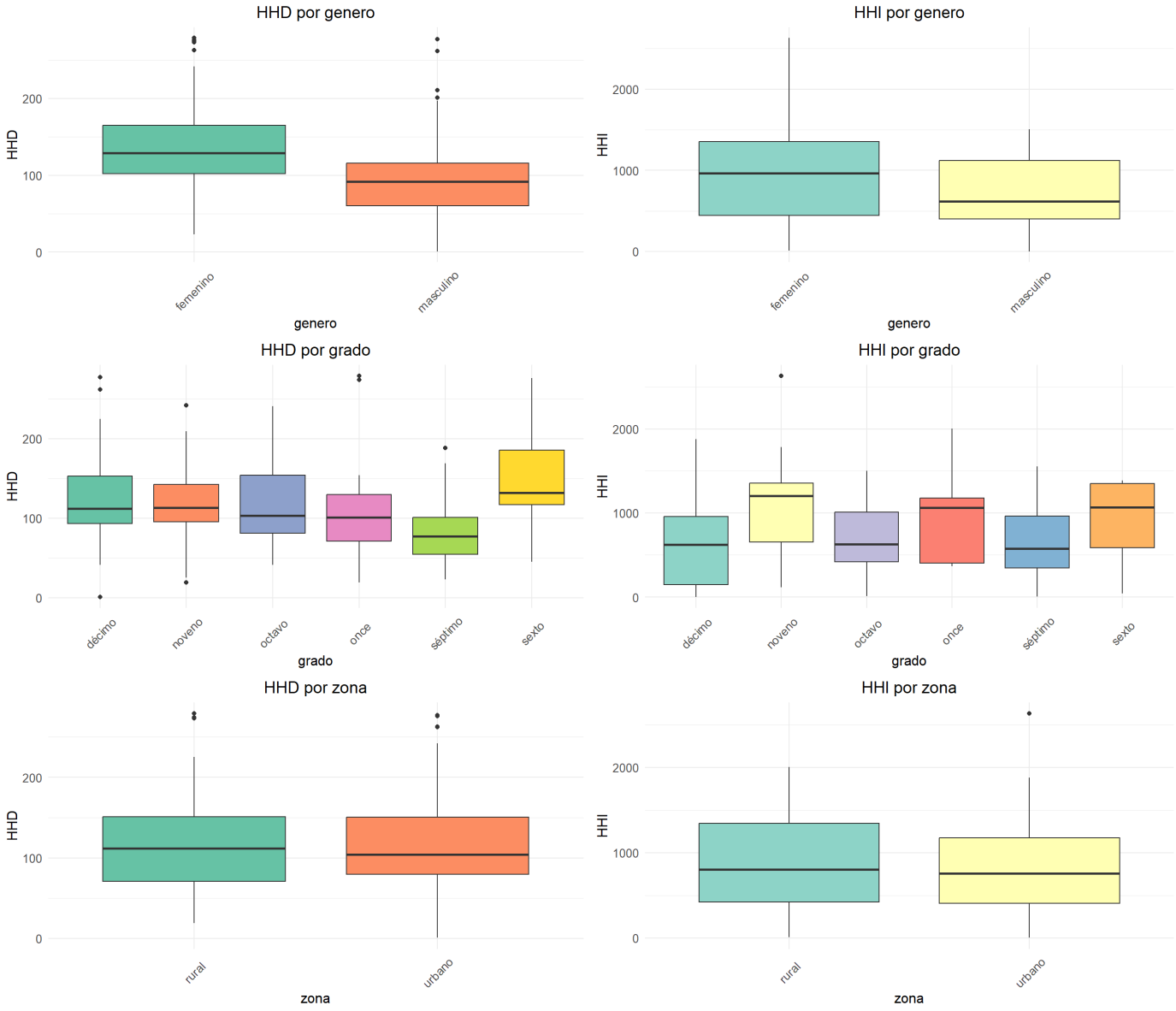
El grupo `Alto` muestra no solo una mediana más elevada, sino también la presencia de valores atípicos. Esto indica que hay una variación significativa en el consumo de agua dentro de este grupo, con algunos estudiantes que tienen un consumo de agua particularmente alto. Estos valores atípicos pueden deberse a prácticas específicas o a circunstancias individuales que resultan en un consumo mucho mayor que el promedio de sus pares.

2.2 Comportamiento de HHD y HHI con cada factor

Se observa que la variable HHD, que representa el consumo de agua directa, es más alto en el grupo femenino. Esta diferencia es evidente en la gráfica, donde la media es relativamente superior a la del grupo masculino. Además, se destaca una mayor variabilidad en este grupo. Lo mismo sucede con el consumo de agua indirecta, el cual es considerablemente mayor en el grupo femenino.

En cuanto al comportamiento de HHD por grado, notamos que la gráfica muestra un mayor consumo para el sexto grado. Sin embargo, se aprecia una variabilidad significativa entre cada grado. Por otro lado, el consumo indirecto de agua es más alto para el noveno grado, con un rango intercuartílico más estrecho y tiende a ser mayor en el 50% de las mediciones.

Finalmente, se evidencia que el consumo de agua directo e indirecto, medido en cada zona, tiene un comportamiento similar. La media y los rangos intercuartílicos están muy cercanos. A partir de esto, podemos concluir que el consumo de agua fue similar para ambas zonas, aunque es importante destacar que para la zona urbana en el consumo directo de agua hay más variabilidad, mientras que en las mediciones del consumo indirecto hay mas variabilidad en la zona rural.



2.3 Resumen de principales indicadores descriptivos por factor

En el resumen general, se destaca que el consumo promedio de agua directa (HHD) para la institución de educación secundaria es de aproximadamente 120 metros cúbicos por año, con una mediana de 109 metros cúbicos. La variabilidad en estos datos es moderada, como se refleja en la desviación estándar de alrededor

de 64.43 metros cúbicos. En cuanto al consumo de agua indirecta (HHI), el promedio es más alto, alrededor de 819.53 metros cúbicos por año, con una mediana de 789 metros cúbicos. La variabilidad en HHI es mayor que en HHD, con una desviación estándar de aproximadamente 523.27 metros cúbicos.

Resumen General de HHD

count_HHD	mean_HHD	median_HHD	sd_HHD	min_HHD	max_HHD	mode_HHD
120	120.1667	109	64.43293	1	279	154

Resumen General de HHI

count_HHI	mean_HHI	median_HHI	sd_HHI	min_HHI	max_HHI	mode_HHI
120	819.525	789	523.2737	2	2631	1356

En términos de género, se evidencia que el grupo femenino tiende a tener un mayor consumo tanto en HHD como en HHI en comparación con el grupo masculino. El consumo promedio y la mediana son significativamente más altos para el grupo femenino en ambas categorías. Además, la variabilidad es mayor en el grupo femenino, lo que sugiere una diversidad de patrones de consumo entre las estudiantes.

Resumen de HHD por Género

genero_HHD	count_HHD	mean_HHD	median_HHD	sd_HHD	min_HHD	max_HHD	mode_HHD
femenino	62	139.58065	129	64.61045	23	279	120
masculino	58	99.41379	92	57.89405	1	277	93

Resumen de HHI por Género

genero_HHI	count_HHI	mean_HHI	median_HHI	sd_HHI	min_HHI	max_HHI	mode_HHI
femenino	62	929.0806	961.5	580.3872	12	2631	1356
masculino	58	702.4138	614.0	429.1280	2	1503	1177

En el análisis por grado escolar, se destaca que el sexto grado presenta el mayor consumo promedio de HHD, mientras que el noveno grado lidera en el consumo de HHI. La variabilidad entre los diferentes grados es notable, indicando que hay variaciones significativas en los patrones de consumo de agua entre los distintos niveles escolares.

Resumen de HHD por Grado Escolar

grado_HHD	count_HHD	mean_HHD	median_HHD	sd_HHD	min_HHD	max_HHD	mode_HHD
décimo	23	125.00000	112	66.27628	1	277	95
noveno	18	118.55556	113	61.54503	19	242	109
octavo	25	126.00000	103	62.75083	41	241	241
once	14	118.78571	101	74.81174	19	279	279
sexto	18	153.22222	132	67.16753	45	276	117
séptimo	22	83.63636	77	42.85473	23	188	120

Resumen de HHI por Grado Escolar

grado_HHI	count_HHI	mean_HHI	median_HHI	sd_HHI	min_HHI	max_HHI	mode_HHI
décimo	23	674.5217	623.0	556.0561	2	1879	1848
noveno	18	1121.0000	1204.5	589.0774	115	2631	1356
octavo	25	707.8800	629.0	442.8529	12	1503	629
once	14	922.5000	1060.0	507.6441	368	2004	1165
sexto	18	946.0556	1066.5	455.1466	44	1387	1344
séptimo	22	682.2727	578.0	488.9994	6	1553	961

Al analizar por zona, se observa que el consumo promedio de HHD es similar tanto en zonas rurales como urbanas, aunque con ligeras diferencias. La variabilidad es un poco más pronunciada en la zona urbana. En cuanto a HHI, la zona rural muestra un consumo promedio ligeramente más alto que la zona urbana, y nuevamente se observa una mayor variabilidad en la zona rural. Esto podría indicar diferencias en los patrones de consumo de agua entre áreas geográficas.

Resumen de HHD por Zona

zona_HHD	count_HHD	mean_HHD	median_HHD	sd_HHD	min_HHD	max_HHD	mode_HHD
rural	49	118.8980	112	64.16133	19	279	120
urbano	71	121.0423	104	65.06117	1	277	154

Resumen de HHI por Zona

zona_HHI	count_HHI	mean_HHI	median_HHI	sd_HHI	min_HHI	max_HHI	mode_HHI
rural	49	862.1020	802	499.8421	6	2004	1356
urbano	71	790.1408	758	540.3898	2	2631	1344