análisis exploratorio de datos EDA

Juan Camilo Suárez Soto

2025-03-24

I. Análisis descriptivo del consumo de recursos.

En este ejercicio se aboradará la importación y limpieza de datos, asegurando su calidad antes del análisis. Posteriormente, aplicaremos medidas estadísticas como la media, mediana, moda y coeficiente de variación para evaluar la dispersión y consistencia de los consumos registrados. Estos indicadores permitirán detectar posibles anomalías y tendencias significativas en el uso de recursos.

Importación de datos

Se utilizó la función `read.csv` de R para importar los datos y almacenarlos en la variable `data_factory.`:

```
data_f <- read.csv("./dataset/data_f.csv", stringsAsFactors = FALSE, check.names = FALSE, row.names = NU
```

Antes de realizar el análisis de datos, se aplicó una técnica sencilla de limpieza en tres variables clave que podrían afectar los resultados estadísticos. Se eliminaron las filas con valores vacíos en "Fecha", "Consumo" y "Concepto", por ser las más sensibles a la falta de datos. Finalmente, se ajustó el formato numérico en "Consumo", "ID" e "ID Unico", reemplazando las comas (,) por puntos (.) en la parte decimal.

```
## cantidad de filas eliminadas:  859
```

```
## cantidad de datos ajustados:  36864
```

Principales medidas de tendencia para cada tipo de recurso:

En el siguiente apartado se presentan las medidas de consumo para cada concepto, como los servicios de agua, luz, entre otros.

tipos de dato: En primer lugar, observamos que hay conceptos que tienen más de una unidad de medición, lo que indica que la variable debe ser normalizada a una única unidad.

Table 1: Tabla de unidades de cada concepto

concepto	unidades
A. COMPRIMIDO	[m3]
AGUA	[m3]
CDD	[°C]
Delta P	[inHg] [mBAR]

concepto	unidades
Energía Eléctrica	[kWh]
Energía Refrigeración	[TNh]
GAS	[m3]
HDD	[°C]
Running Time	[hrs]
VACIO	[m3]
VAPOR	[kg] [Tn]

Para analizar el comportamiento de los datos, se clasificaron las variables según su tipo. Las fechas son continuas, ya que siguen un orden. En contraste, "concepto", "línea", "equipo clave", "unidad" e "ID fabricación" son nominales, pues representan categorías sin jerarquía. Por último, "consumo" e "ID único" son numéricos pero no siguen un orden específico:

Table 2: Tabla de variables y tipos de datos

variables	tipos
FECHA INICIO	continua
FECHA TERMINO	continua
CONCEPTO	nominal
LINE	nominal
KEY Equipment	nominal
UNIDAD	nominal
CONSUMO	Continua
ID	Discreta
ID Unico	Discreta
ID FABRICACION	nominal

unidades de medida:

media: La media de aire comprimido y vacío destaca por su magnitud, sugiriendo un uso intensivo, posibles fugas o incluso errores de medición, mientras que agua y energía eléctrica, aunque menores, siguen siendo relevantes para los costos y la eficiencia de los procesos. El Delta P negativo indica una diferencia de presión particular, y los demás conceptos, con consumos más moderados.

```
library(knitr)
media_data  <- aggregate(CONSUMO ~ CONCEPTO, data = data_f, FUN = mean)
kable(media_data, caption = "Media de consumo por Concepto", digits = 2)
```

Table 3: Media de consumo por Concepto

CONCEPTO	CONSUMO
A. COMPRIMIDO	9565.22
AGUA	55.47
CDD	5.99
Delta P	-25.68
Energía Eléctrica	5144.79
Energía Refrigeracion	1647.49
GAS	2473.06

CONCEPTO	CONSUMO
HDD	27.93
Running Time	17.48
VACIO	47483.46
VAPOR	357.62

mediana: La mediana del consumo de aire comprimido y vacío es inflada por valores atípicos, mientras que la mediana refleja mejor el consumo típico. En los demás conceptos, la diferencia es menor, indicando una distribución más equilibrada y hace que la mediana sea más confiable que la media.

```
library(knitr)
mediana_data <- aggregate(CONSUMO ~ CONCEPTO, data = data_f, FUN = median)
kable(mediana_data, caption = "Mediana de consumo por Concepto", digits = 2)
```

Table 4: Mediana de consumo por Concepto

CONCEPTO	CONSUMO
A. COMPRIMIDO	2055.71
AGUA	3.50
CDD	2.70
Delta P	-15.16
Energía Eléctrica	1030.12
Energía Refrigeracion	596.43
GAS	566.22
HDD	21.30
Running Time	18.12
VACIO	39115.10
VAPOR	31.97

moda: En el dataset, los conceptos como aire comprimido, agua, CDD, energía eléctrica, gas y vapor tienen 0.00 como el valor más frecuente, mientras que en los demás conceptos predominan otros valores.

```
library(knitr)
# Función para calcular la moda
getmode <- function(v) {
  uniqv <- unique(v)
  uniqv[which.max(tabulate(match(v, uniqv)))]
}
moda_data <- aggregate(CONSUMO ~ CONCEPTO, data = data_f, FUN = getmode)
kable(moda_data, caption = "moda de consumo por Concepto", digits = 2)
```

Table 5: moda de consumo por Concepto

CONCEPTO	CONSUMO
A. COMPRIMIDO	0.00
AGUA	0.00
CDD	0.00
Delta P	-41.91

CONCEPTO	CONSUMO
Energía Eléctrica	0.00
Energía Refrigeracion	9.00
GAS	0.00
HDD	62.10
Running Time	18.75
VACIO	62075.11
VAPOR	0.00

varianza muestral: En el consumo por concepto, destaca la alta variabilidad en vacío y aire comprimido, lo que sugiere valores atípicos. En contraste, AGUA, CDD, HDD, Running Time y VAPOR presentan dispersión aceptable.

```
library(knitr)
sd_data  <- aggregate(CONSUMO ~ CONCEPTO, data = data_f, FUN = sd)
kable(sd_data, caption = "Desv. estándar de consumo por Concepto", digits = 2)
```

Table 6: Desv. estándar de consumo por Concepto

CONCEPTO	CONSUMO
A. COMPRIMIDO	33875.87
AGUA	1047.48
CDD	7.26
Delta P	28.22
Energía Eléctrica	19951.97
Energía Refrigeracion	1751.14
GAS	3703.74
HDD	23.33
Running Time	9.76
VACIO	42473.31
VAPOR	962.45

coeficiente de variación: En el consumo por concepto, destaca la alta variabilidad en vacío y aire comprimido, lo que sugiere valores atípicos. En contraste, AGUA, CDD, HDD, Running Time y VAPOR presentan menor dispersión. La desviación negativa de Delta P tiene una notoria y negativa desviación, algo muy particular en esta variable

```
library(knitr)
cv_data <- aggregate(CONSUMO ~ CONCEPTO, data = data_f, FUN = function(x) sd(x) / mean(x))
kable(sd_data, caption = "Coeficiente de variación de consumo por Concepto", digits = 2)
```

Table 7: Coeficiente de variación de consumo por Concepto

CONCEPTO	CONSUMO
A. COMPRIMIDO	33875.87
AGUA	1047.48
CDD	7.26
Delta P	28.22

CONCEPTO	CONSUMO
Energía Eléctrica	19951.97
Energía Refrigeracion	1751.14
GAS	3703.74
HDD	23.33
Running Time	9.76
VACIO	42473.31
VAPOR	962.45

Resumen estadístico

Tras calcular la media y la mediana se puede concluir que, la media muestral del consumo promedio está sesgada por valores atípicos, mientras que la mediana refleja el consumo típico sin ser afectada por extremos, lo que la hace más fiable en este caso; y la gran diferencia entre ambas indica la presencia de outliers, como en el caso del aire comprimido y el vacío. Además, el coeficiente de variación confirma la hipótesis de la poca consistencia en el consumo de cada concepto.

Table 8: Resumen estadístico de consumo por concepto

CONCEPTO	Mediana	Media	Moda	SD	CV
A. COMPRIMIDO	2055.71	9565.22	0.00	33875.87	3.54
AGUA	3.50	55.47	0.00	1047.48	18.89
CDD	2.70	5.99	0.00	7.26	1.21
Delta P	-15.16	-25.68	-41.91	28.22	-1.10
Energía Eléctrica	1030.12	5144.79	0.00	19951.97	3.88
Energía Refrigeracion	596.43	1647.49	9.00	1751.14	1.06
GAS	566.22	2473.06	0.00	3703.74	1.50
HDD	21.30	27.93	62.10	23.33	0.84
Running Time	18.12	17.48	18.75	9.76	0.56
VACIO	39115.10	47483.46	62075.11	42473.31	0.89
VAPOR	31.97	357.62	0.00	962.45	2.69

II. Visualización y distribución de consumo

En el siguiente apartado se presenta un histograma de densidad por tipo de consumo. En el eje y se muestra la cantidad de registros correspondientes a cada nivel de consumo específico.

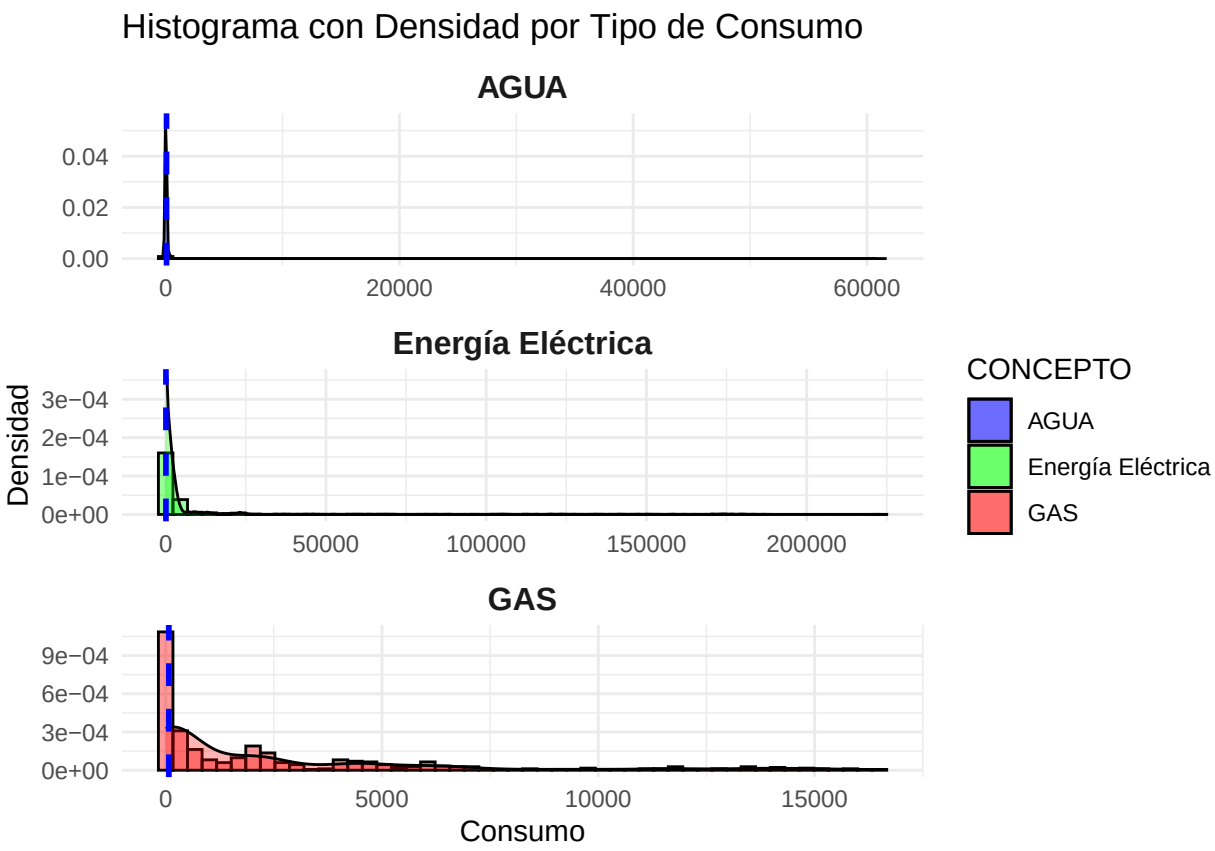
El código filtra los datos para seleccionar los principales recursos (“GAS”, “AGUA” y “Energía Eléctrica”) y crea un histograma con densidad para cada uno. Se utiliza ggplot para trazar la distribución de consumo con 50 bins, se aplican colores personalizados y se añade una línea vertical que indica la mediana. Además, los gráficos se organizan en una sola columna para facilitar la comparación entre conceptos.

```
library(ggplot2)
library(dplyr)

# Filtrar los principales recursos
recursos_principales <- c("GAS", "AGUA", "Energía Eléctrica")
datos_principales <- data_f %>% filter(CONCEPTO %in% recursos_principales)

# Crear el histograma con densidad en formato vertical
```

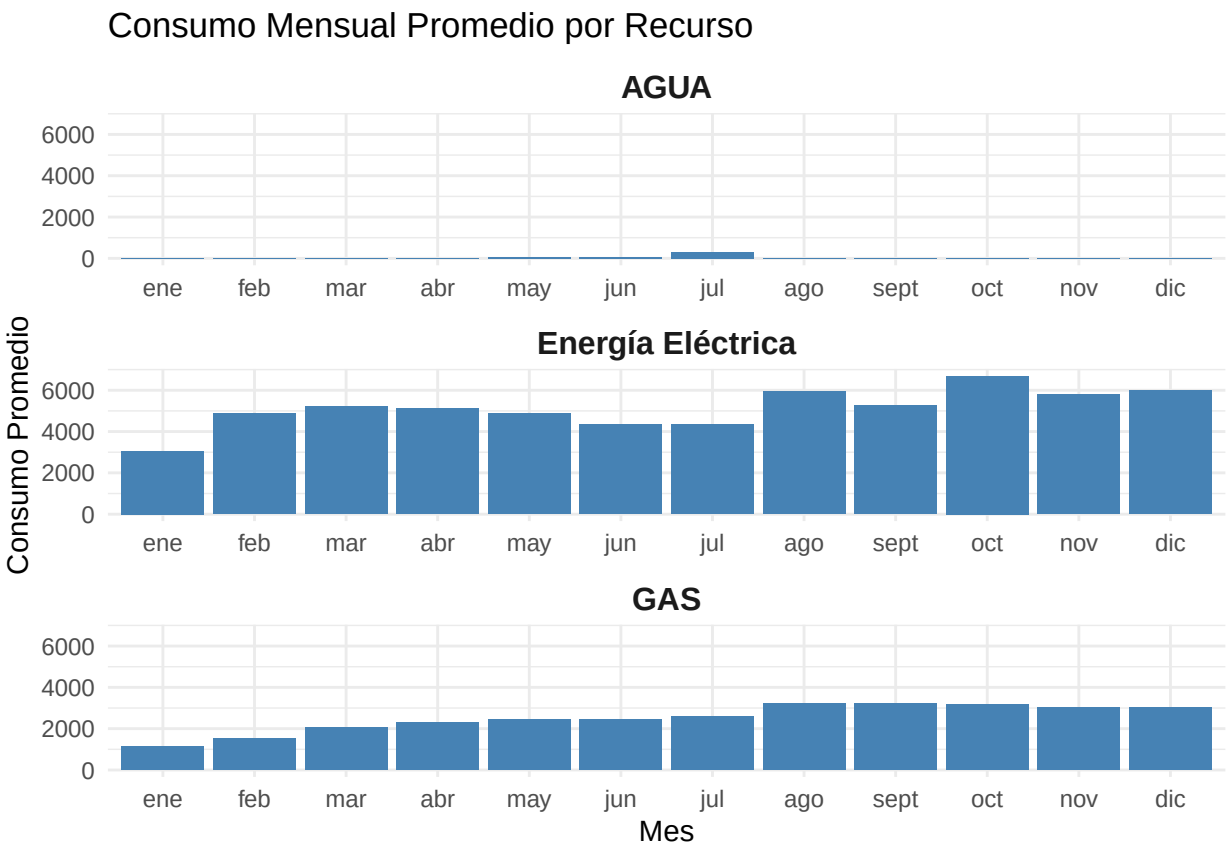
```r
ggplot(datos_principales, aes(x = CONSUMO, fill = CONCEPTO)) +
  geom_histogram(aes(y = ..density..), bins = 50, color = "black", alpha = 0.4, position = "identity") +
  geom_density(alpha = 0.3) +
  # Cambiamos ncol = 1 para tener una sola columna (gráficos en fila)
  facet_wrap(~CONCEPTO, scales = "free", ncol = 1) +
  scale_fill_manual(values = c("blue", "green", "red")) +
  geom_vline(aes(xintercept = median(CONSUMO, na.rm = TRUE)),
             color = "blue", linetype = "dashed", size = 1) +
  labs(title = "Histograma con Densidad por Tipo de Consumo",
       x = "Consumo",
       y = "Densidad",
       fill = "CONCEPTO") +
  theme_minimal() +
  theme(strip.text = element_text(face = "bold", size = 12))
```



El histograma con densidad por tipo de consumo muestra una fuerte concentración de datos hacia valores bajos, lo que sugiere un sesgo hacia la izquierda. Algunos valores más altos pueden considerarse atípicos, indicando consumos inusuales que conviene analizar con detalle. Estas observaciones son clave al decidir si el modelo estadístico debe asumir una distribución paramétrica o no, y resaltan la importancia de anotar y monitorear posibles outliers para comprender mejor la variabilidad de cada recurso.

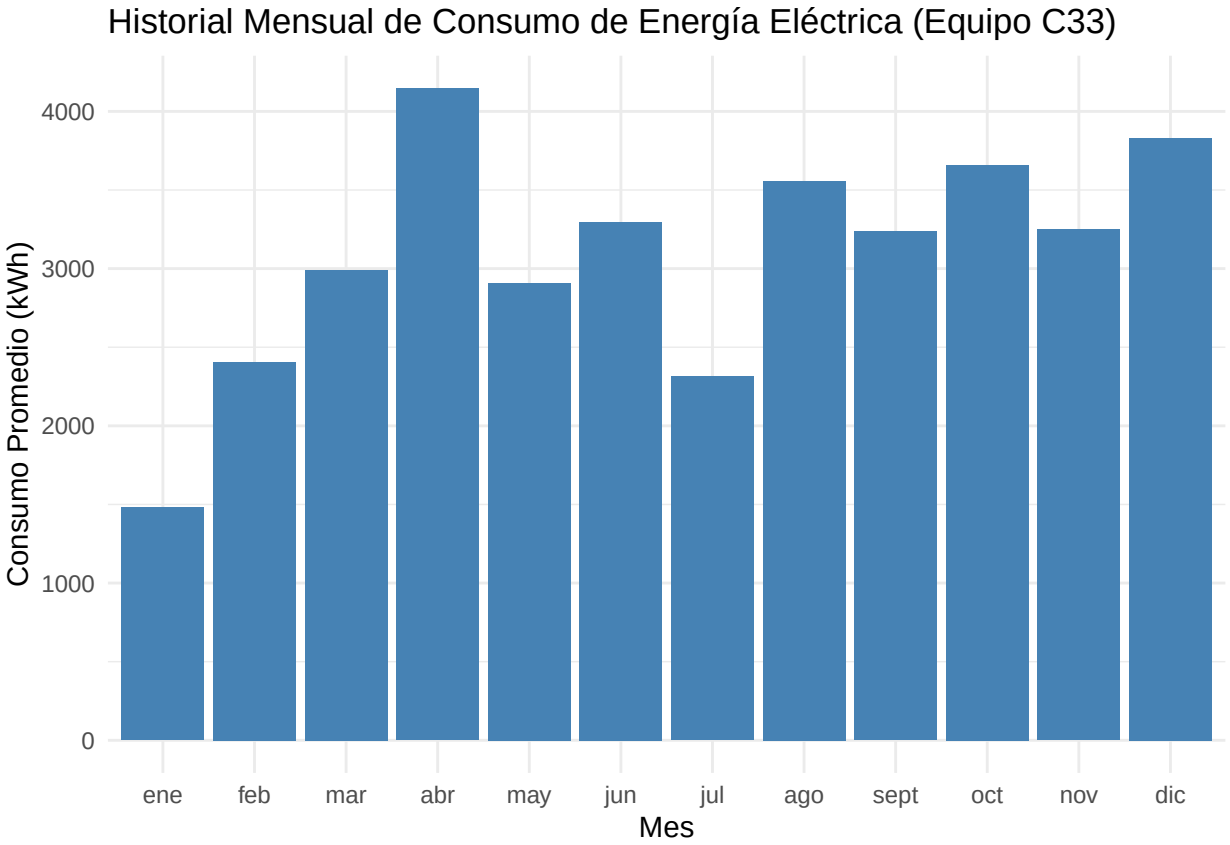
III. Análisis de patrones temporales de consumo

En el siguiente apartado se presenta el historial de consumo mensual promedio del agua, gas y luz.



De este modo podemos darnos cuenta que el concepto menos consumido es el agua, seguido del gas y la energía.

Ahora podrá observar el consumo eléctrico mensual de la máquina C33, la cual presenta un alto consumo energético en promedio.



El segundo gráfico muestra el consumo mensual de energía eléctrica del equipo C33, mientras que el primer gráfico presenta el consumo promedio mensual de tres recursos: agua, electricidad y gas.

Observando ambos gráficos, se puede notar que el consumo eléctrico del equipo C33 sigue un patrón similar al del consumo total de energía eléctrica en el primer gráfico. Esto sugiere que el equipo C33 podría tener una contribución significativa al consumo general de electricidad. Sin embargo, dado que el primer gráfico representa un promedio de todos los equipos y consumos, no se puede afirmar con certeza que el equipo C33 sea el principal responsable sin un análisis más detallado.

En cuanto a la relación con los otros recursos (agua y gas), no se observa una correlación directa con el comportamiento del consumo del equipo C33, ya que su gráfico solo muestra el consumo de electricidad. Para determinar si su funcionamiento influye en el consumo de agua o gas, sería necesario contar con datos más detallados sobre los procesos en los que se involucra.

IV. Análisis de correlaciones entre consumos

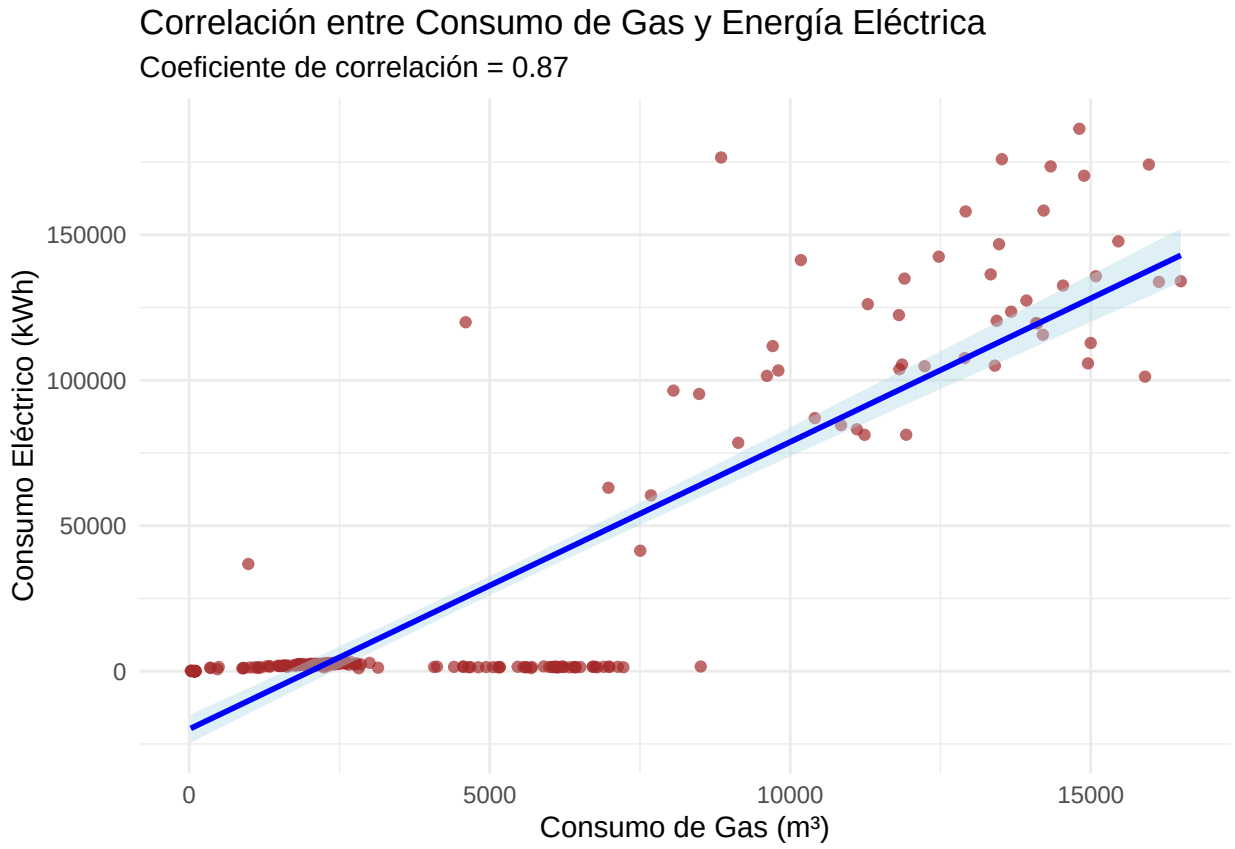
Preparación de datos:

A continuación se hará una correlación entre el consumo de energía eléctrica y gas, agrupando primero el tipo de equipo y la fecha de terminación:

Un coeficiente de correlación de 0.86 indica una relación fuerte y positiva entre dos variables, lo que significa que cuando una aumenta, la otra tiende a hacerlo también, y cuando una disminuye, la otra sigue el mismo patrón. Este valor, cercano a 1, sugiere que los datos se alinean bastante con una tendencia lineal ascendente, lo que permite una buena capacidad de predicción entre ambas variables. Sin embargo, es importante

recordar que una correlación alta no implica causalidad; la relación observada puede deberse a otros factores o simplemente a una coincidencia.

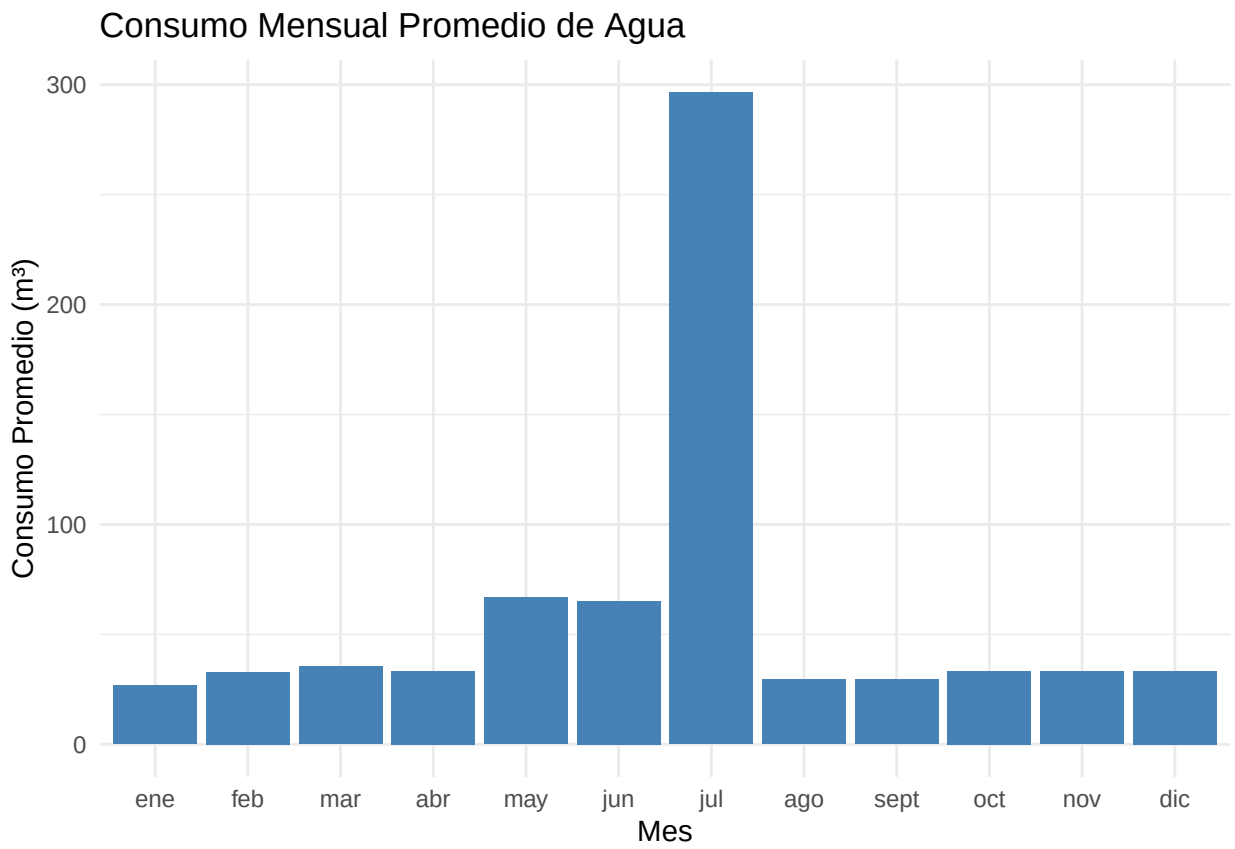
```
## [1] 0.8739639
```



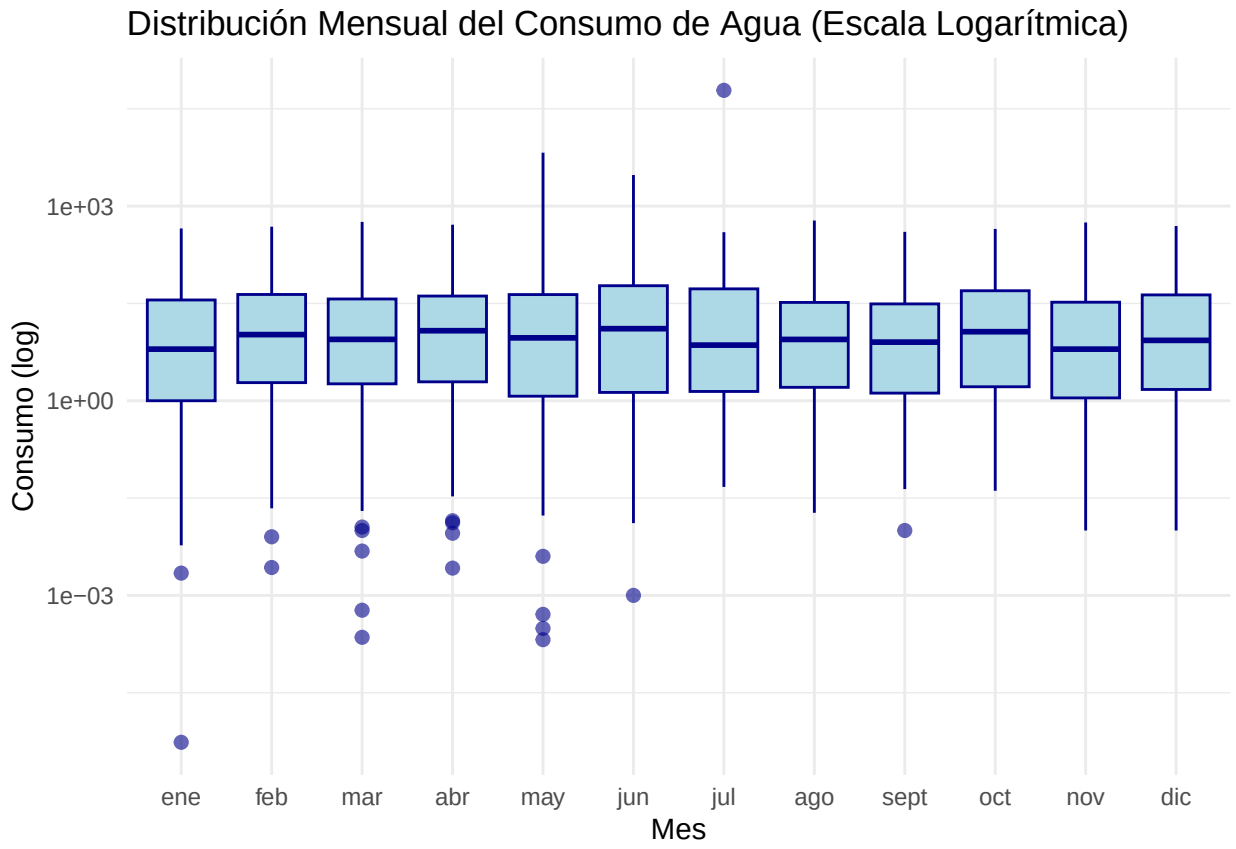
Los valores atípicos en torno a 0 en consumo eléctrico, pero con alto consumo de gas, podrían indicar procesos donde el gas es la fuente principal de energía sin necesidad de apoyo eléctrico, como en sistemas de calefacción o producción térmica independiente. Por otro lado, el grupo de valores en el límite superior sugiere puntos donde ambos consumos son extremadamente altos, posiblemente reflejando industrias o equipos que utilizan simultáneamente grandes cantidades de gas y electricidad, como plantas de cogeneración o sistemas con alta demanda energética. Estos patrones podrían señalar ineficiencias o comportamientos específicos de consumo que merecen un análisis más detallado.

V. Análisis de distribución de consumo y dimensionamiento de recursos.

Dustrubución del consumo de agua: Al analizar nuevamente el histograma del consumo mensual de agua, observamos que la distribución es asimétrica. En el primer semestre (enero - junio), el consumo promedio se sitúa ligeramente por encima de los 50 m^3. Sin embargo, en el segundo semestre (julio - diciembre), el consumo aumenta significativamente, alcanzando 300 m^3 en julio. Este pico de consumo provoca un sesgo en la distribución, lo que confirma que no es simétrica.



Por otro lado, se generó un diagrama **boxplot** con escala logarítmica, ya que en una escala lineal no se visualiza correctamente debido a la gran variabilidad de los datos. Para interpretar mejor este gráfico, según la Tabla 8, la media del consumo de agua es de 55,47 m^3, mientras que la mediana es de 3,50 m^3. En comparación con el histograma anterior, podemos concluir que los valores atípicos más altos corresponden al mes de julio.



determinación de datos normales: El script calcula un rango "normal" de consumo asumiendo que, en una distribución normal, aproximadamente el 68% de los datos se ubican dentro de la media ± una desviación estándar. En este caso, se utiliza la media (55.47) y la desviación estándar (1047.48) para establecer los límites. Dado que el consumo no puede ser negativo, se ajusta el límite inferior a 0 si el cálculo arroja un valor negativo. Aunque la gran diferencia entre la media y la mediana (3.50) sugiere que la distribución está fuertemente sesgada por valores extremos, este método proporciona un punto de partida objetivo para identificar el rango de consumo "normal".

```
# Calcular los cuartiles
# Valores resumen obtenidos:
media <- 55.47
mediana <- 3.50
sd_val <- 1047.48

# Establecer límites usando la regla de la desviación estándar (mean ± 1 SD)
# Dado que el consumo no puede ser negativo, se establece el límite inferior en 0 si corresponde.
limite_inferior <- max(0, media - sd_val)
limite_superior <- media + sd_val

cat("Rango normal de consumo (mean ± 1 SD): [", limite_inferior, ",", limite_superior, "]\n")
```

```
## Rango normal de consumo (mean ± 1 SD): [ 0 , 1102.95 ]
```

También se hizo otro script que determina el porcentaje de valores que están sobre y bajo 1000 m^3

```r
library(dplyr)

porcentajes_agua <- data_f %>%
  filter(CONCEPTO == "AGUA") %>%
  summarise(
    total = n(),
    debajo_1000 = sum(CONSUMO < 1000, na.rm = TRUE),
    porcentaje_debajo = 100 * debajo_1000 / total,
    encima_1000 = sum(CONSUMO >= 1000, na.rm = TRUE),
    porcentaje_encima = 100 * encima_1000 / total
  )

print(porcentajes_agua)
```

```
##   total debajo_1000 porcentaje_debajo encima_1000 porcentaje_encima
## 1  3444        3438          99.82578           6          0.174216
```