M2.851_20241_May25 - Tipologia y Ciclo de Vida de los Datos

Manuel Muñoz Plá

Mayo 2025

Contents

1 Descripción del dataset

El dataset utilizado se ha construido mediante técnicas de scraping sobre la ***Sección V – A del Boletín Oficial del Estado (BOE)***, centrado exclusivamente en ***anuncios de licitación y formalización de contratos del sector público español***. Cubre un periodo comprendido entre el 1 de enero de 2014 y el 31 de diciembre de 2024, y ha sido

generado íntegramente mediante un flujo automatizado y reproducible descrito en la Práctica 1. La fuente es oficial, de dominio público y legalmente reutilizable conforme a la Ley 37/2007 - RISP[1]

En esta segunda fase se ha enriquecido el dataset original con cuatro nuevos atributos de valor económico y contractual, `Naturaleza`, `Valor_estimado_licitacion`, `Valor_oferta_adjudicada` y `Nombre_adjudicatario`. Esto permite abordar nuevas analíticas tanto con variables categoricas como cuantitativas, centradas en el gasto público y la estructura del mercado de contratación estatal[2]

El raspado se ha realizado modificando el codigo python original y es accesible en github

1.1 Objetivo y utilidad

Este conjunto de datos permite responder a cuestiones relacionadas con

- La distribución del gasto público entre organismos y sectores.
- El comportamiento contractual por regiones y comunidades autónomas
- El uso de diferentes procedimientos de adjudicación[3]
- La identificación de adjudicatarios dominantes
- La eficiencia comparada entre organismos públicos[4]
- El análisis económico de los contratos en función del presupuesto inicial y la adjudicación final[5]

1.2 Variables actuales

Atributos que conforman el dataset actualizado

- `Institucion`. Organismo principal bajo el cual se publica el anuncio
- `Organismo responsable`. Unidad o departamento que gestiona la contratación
- `Expediente`. Identificador único del procedimiento
- `Fecha`. fecha de publicación en formato `dd/mm/aaaa`
- `Tipo`. Tipo del anuncio `Licitación` o `Contratación`
- `Naturaleza`. Tipo general de contrato. `Servicios`, `Obras`, `Suministros`, etc… **(Nuevo atributo)**
- `Objeto`. Descripción textual del propósito contractual.
- `Procedimiento`. Tipo de proceso seguido. `Abierto`, `Negociado`, etc…
- `Ambito_geografico`. Localización del contrato. Nacional, Región, Comunidad, Localidad, (Cuando se aplica a varias zonas, habitualmente a nivel “Nacional” o “Parcial Nacional” suele quedar marcado como “Sin especificar”)
- `Materias_CPV`. Descripción temática del contrato, según el código CPV (Válido en toda la U.E.)
- `Codigos_CPV`. Códigos CPV concretos asociados al contrato. Suele ser una especificación de “Materias_CPV”
- `valor_estimado_licitacion`. Importe en euros presupuestado antes de la adjudicación. Se trata de una estimacion aproximada.**(Nuevo atributo)**
- `valor_oferta_adjudicada`. Importe en euros finalmente adjudicado. Valor real de la contratación. **(Nuevo atributo)**
- `nombre_adjudicatario`. Nombre del adjudicatario, contratista o empresa ganadora. **(Nuevo atributo)**
- `Enlace HTML`. URL al anuncio completo en la web del BOE

1.3 Tamaño y cobertura

El dataset contiene observaciones estructuradas en variables. El número de observaciones no se corresponde estrictamente con el número de expedientes, ya que algunos procesos de contratación pueden tener **adjudicación múltiple**. En estos casos, cuando un expediente se resuelve a favor de varios contratistas, por ejemplo, en contratos por lotes o adjudicaciones compartidas, se ha optado por **desdoblar el registro original**, generando una observación por cada adjudicatario asociado, de ahí que se pase de las **90.140 observaciones originales a 97.154 actuales**. Este enfoque garantiza una mayor granularidad en el análisis económico y de proveedores, permitiendo medir correctamente la concentración, distribución y repetición de adjudicatarios en el tiempo[6]

```r
# se carga el dataset
df <- readr::read_csv("CSV/licitaciones_contrataciones_BOE_2014_2024.csv", show_col_types = FALSE)

# número de filas y columnas del dataset
filas_columnas_csv <- data.frame(
  "Descripción" = c("Filas / Observaciones", "Columnas / Variables"),
  "Cantidad" = c(nrow(df), ncol(df))
)
```

```r
# kable
filas_columnas_csv %>%
  kable(col.names = c("Descripción", "Cantidad")) %>%
  kable_styling(full_width = FALSE, position = "center", font_size = 9)
```

Descripción	Cantidad
Filas / Observaciones	97154
Columnas / Variables	15

```r
# resumen 5 primeros elementos
df_resumen <- data.frame(

  Tipo = sapply(df, class),  # tipo de datos por variable
  Muestra = sapply(df, function(x) { # 5 elementos
      paste(head(x, 5), collapse = ", ")
    }
    )
  )

# kable
df_resumen %>%
  kable(col.names = c("Variable", "Tipo de Dato", "5 Primeras observaciones separadas por comas "), align = "
  kable_styling(full_width = FALSE, position = "center", font_size = 9,
                latex_options = "hold_position") %>%
  column_spec(1, width = "4cm") %>%
  column_spec(2, width = "2cm") %>%
  column_spec(3, width = "11cm")
```

Variable	Tipo de Dato	5 Primeras observaciones separadas por comas
Institucion	character	MINISTERIO DE DEFENSA, MINISTERIO DE DEFENSA, MINISTERIO DE DEFENSA, MINISTERIO DE HACIENDA Y ADMINISTRACIONES PÚBLICAS, MINISTERIO DE HACIENDA Y ADMINISTRACIONES PÚBLICAS
Organismo responsable	character	Dirección General del INTA, Dirección General del INTA, Jefatura de la Sección Económico Administrativa del Hospital Central de la Defensa Gómez Ulla., Delegación Especial de la Agencia Tributaria en Extremadura, Dirección del Servicio de Gestión Económica de la Agencia Estatal de la Administración Tributaria.
Expediente	character	500083225200., 500083225000., 441/2014., 13060116400., 13700181400.
Fecha	character	03/01/2014, 03/01/2014, 03/01/2014, 03/01/2014, 03/01/2014
Tipo	character	Contratación, Contratación, Contratación, Contratación, Contratación
Naturaleza	character	Suministros, Suministros, Suministros, Servicios, Obras
Objeto	character	LCVRS de calificación y vuelo de SO/PHI., Componentes electrónicos para SO/PHI., Pescado y Marisco Congelados TA2014., Servicio de Seguridad en los edificios de la Delegación Especial de Extremadura., Castilla y León-D.E. *Obras Complementarias pl. 3ª y Urbanización Nuevo Edificio.
Procedimiento	character	Negociado sin publicidad, Negociado sin publicidad, Abierto, Abierto, Negociado sin publicidad
Ambito_geografico	character	Comunidad de Madrid, Comunidad de Madrid, Comunidad de Madrid, Extremadura, Castilla y León
Materias_CPV	character	38000000 Equipo de laboratorio, óptico y de precisión (excepto gafas), 31000000 Máquinas, aparatos, equipo y productos consumibles eléctricos; iluminación, 15000000 Alimentos, bebidas, tabaco y productos afines, 79000000 Servicios a empresas: legislación, mercadotecnia, asesoría, selección de personal, imprenta y seguridad, 45000000 Trabajos de construcción

Codigos_CPV	character	38000000 (Equipo de laboratorio, óptico y de precisión (excepto gafas))., 31000000 (Máquinas, aparatos, equipo y productos consumibles eléctricos; iluminación)., 15000000 (Alimentos, bebidas, tabaco y productos afines)., 79710000 (Servicios de seguridad)., 45450000 (Otros trabajos de acabado de edificios).
valor_estimado_licitacion	character	120.000,00 euros, 321.860,00 euros, 181.818,18 euros, 453.081,90 euros, 527.216,51 euros
valor_oferta_adjudicada	character	120.000,00 euros, 321.860,00 euros, 200.000,00 euros, 410.602,24 euros, 527.216,51 euros
nombre_adjudicatario	character	Arcopix, S.A., Alter Technology Tüv Nord, S.A.U., Disblamar, S.L., Seguridad Integral Secoex, S.A., CORSAM CORVIAM CONSTRUCCION SA.
Enlace HTML	character	https://www.boe.es/diario_boe/txt.php?id=BOE-B-2014-201, https://www.boe.es/diario_boe/txt.php?id=BOE-B-2014-202, https://www.boe.es/diario_boe/txt.php?id=BOE-B-2014-203, https://www.boe.es/diario_boe/txt.php?id=BOE-B-2014-205, https://www.boe.es/diario_boe/txt.php?id=BOE-B-2014-206

Aunque todas las variables han sido leídas inicialmente como texto `character`, **ciertos campos requieren transformación posterior** para su correcto tratamiento analítico.

- La variable `Fecha` debe convertirse al tipo `Date`.
- Las variables `valor_estimado_licitacion` y `valor_oferta_adjudicada` contienen cantidades monetarias en euros y deben transformarse al tipo `numeric`, eliminando símbolos como "euros", comas o puntos de miles[6]^ En el análisis inicial se observa que los campos `valor_estimado_licitacion` y `valor_oferta_adjudicada` son **complementarios**
  - Si existe un valor en `valor_oferta_adjudicada`, el valor de licitación puede considerarse histórico o informativo.
  - Si sólo hay valor en `valor_estimado_licitacion` y `valor_oferta_adjudicada` está vacío, se trata de una licitación pendiente de adjudicación.
  - Muchos registros contienen el literal `"No disponible"` corresponden al contravalor de uno de los dos atributivos. Finalmente en la fase de limpieza, se puede optar por fundir ambos campos en un nuevo atributo Valor_en_Euros, o mantenerlos separados en dos campos como hasta ahora[4]

El resto de variables como `Institucion`, `Tipo`, `Naturaleza` o `Procedimiento`, serán tratadas como variables **categóricas nominales**, aunque presentan una gran variedad de formas, abreviaturas y errores ortográficos que requieren **normalización previa** (por ejemplo, "ADIF - Presidencia" vs "ADIF – Presidencia. vs Adif presidencia")[5]

También se ha detectado que la variable `Ambito_geografico` contiene con frecuencia valores como `"sin definir"`, que parecen corresponder a contratos de ámbito nacional (según se deduce del contenido de los enlaces HTML). Esta hipótesis deberá comprobarse tras cuantificar la frecuencia de aparición de dicho valor.

Todos estos ajustes, normalización de categorías, transformación de tipos y tratamiento de valores ausentes, serán abordados en el apartado de limpieza de datos.

2 Integración y selección de los datos

Se ha realizado un enriquecimiento del dataset original mediante una **integración adicional** a cada registro, complementando la información con atributos económicos y contractuales extraídos directamente del HTML oficial [2,6]

- `Naturaleza`. Categoría del contrato (Servicios, Obras, Suministros…)
- `valor_estimado_licitacion`. Importe presupuestado antes de la adjudicación, como estimación previa
- `valor_oferta_adjudicada`. Importe final adjudicado, es decir, el valor real de la adjudicación
- `nombre_adjudicatario`. Proveedor, contratista o empresa adjudicataria

Esta integración permite disponer de un esquema más completo para análisis comparativos entre lo licitado y lo adjudicado, así como para estudiar la concentración de adjudicatarios en sectores o regiones específicos[2]

A partir del apartado 4, se **focaliza el análisis exclusivamente en los registros de tipo `Contratación`**, ya que estos reflejan el resultado final del proceso administrativo. Las licitaciones muestran únicamente una **intención de gasto estimada**, mientras que las contrataciones recogen la **ejecución real del gasto público**, incluyendo importes firmes y adjudicatarios definidos[2,6]

Trabajar sólo con registros de tipo `Contratación` permite

- Evaluar la **concentración de adjudicaciones** en determinados adjudicatarios

- Analizar la **distribución del gasto** por **ámbito geográfico** (comunidades autónomas o regiones)
- Estudiar la **frecuencia de contratación** por sectores CPV (Clasificación Común de Productos por Actividades)
- Examinar la actividad contractual según la **naturaleza del contrato** (Servicios, Obras, Suministros), tanto de forma agregada como cruzada
- Realizar análisis **multidimensionales** combinando atributos como `Naturaleza × Ámbito_geografico`, `Procedimiento × Sector` o `Organismo responsable × Adjudicatario`, obteniendo una visión segmentada de las políticas de contratación pública

```r
df_contratacion <- df %>% filter(Tipo == "Contratación")

# número de filas y columnas sólo de contrataciones
filas_columnas_csv <- data.frame(
  "Descripción" = c("Filas / Observaciones", "Columnas / Variables"),
  "Cantidad" = c(nrow(df_contratacion), ncol(df_contratacion))
)

# tabla de tamaño
filas_columnas_csv %>%
  kable(
    col.names = c("Descripción", "Cantidad"),
    caption = "Numero de observaciones tras filtrar sólo registros de contratación"
    ) %>%
  kable_styling(full_width = FALSE, position = "center", font_size = 9)
```

Numero de observaciones tras filtrar sólo registros de contratación

Descripción	Cantidad
Filas / Observaciones	63101
Columnas / Variables	15

La limpieza y transformación de datos se realiza sobre el conjunto completo de registros (licitaciones y contrataciones) para maximizar la calidad de la información, minimizar la pérdida de casos potencialmente útiles y evitar la introducción de sesgos artificiales. Excluir las licitaciones en fases tempranas de limpieza podría ocultar patrones de error o inconsistencias presentes sólo en las fases iniciales del ciclo de vida del dato, o incluso provocar una subestimación de la frecuencia de valores ausentes en atributos relevantes[2,4,6] Sólo tras asegurar una depuración homogénea y reproducible se filtran los registros de tipo `Contratación` para el análisis principal, garantizando así la comparabilidad y robustez de los resultados.

3 Limpieza de los datos

En este apartado se aborda el proceso de transformación, depuración y codificación del dataset seleccionado, necesario para garantizar que los análisis posteriores se basen en datos fiables, comparables y consistentes. El tratamiento ha sido diseñado específicamente en función de la estructura y origen del dataset, así como del tipo de preguntas analíticas que se desean abordar. Aunque las tareas de limpieza se realiza sobre el dataset completo, una vez tratado sólo se seleccionan las observaciones del tipo "Contratación".

3.1 Conversión de atributos

Todas las variables han sido leídas inicialmente como texto `character`. Algunas de ellas requieren una **conversión explícita de tipo** para ser interpretadas correctamente

- `Fecha` se convierte al tipo `Date`, usando el formato `"dd/mm/yyyy"`.
- `valor_estimado_licitacion` y `valor_oferta_adjudicada` se transforman a tipo `numeric`, eliminando previamente símbolos no numéricos como "euros", puntos de miles o comas decimales. Este paso es necesario para poder realizar comparaciones, estadísticas agregadas y análisis económicos significativos[3]

El resto de variables serán tratadas como **categóricas nominales**, con posterior conversión a `factor` cuando se requiera una codificación explícita para modelos o visualizaciones.

```r
if (do_cleaning) {
  # conversión de fecha
  df <- df %>%
    mutate(Fecha = as.Date(Fecha, format = "%d/%m/%Y"))

  # limpieza de campos en euros
  limpiar_valor_euro_es <- function(x) {
    x <- trimws(x)
    x <- gsub("\\.", "", x)
    x <- gsub(",", ".", x)
    x <- gsub(" euros", "", x, ignore.case = TRUE)
    x <- ifelse(tolower(x) %in% c("no disponible", ""), NA, x)
    suppressWarnings(as.numeric(x))
  }

  df <- df %>%
    mutate(
      valor_estimado_licitacion = limpiar_valor_euro_es(valor_estimado_licitacion),
      valor_oferta_adjudicada   = limpiar_valor_euro_es(valor_oferta_adjudicada)
    )
}
```

3.2 Conversión a tipos base

En este apartado se realiza la conversión de cada variable al tipo base más adecuado: character, numeric, Date, etc… y se aplican transformaciones adicionales a las variables categóricas, como la normalización textual y la unificación de niveles similares.

- Eliminación de mayúsculas/minúsculas inconsistentes
- Unificación de nombres con diferencias por puntuación u ortografía Por ejemplo, `"ADIF - Presidencia"`, `"Adif Presidencia."`, y otras variantes... serán convertidos a `"ADIF PRESIDENCIA"` unificandolos en un unico `Organismo Responsable`
- Reagrupación manual en casos necesarios. Como unificar variantes de organismos bajo una categoría común

```r
if (do_cleaning) {
  df <- df %>%
    rename(Organismo_responsable = `Organismo responsable`) %>%
    mutate(Organismo_responsable = Organismo_responsable %>%
             toupper() %>%
             stringr::str_replace_all("[[:punct:]]", "") %>%
             stringr::str_squish())

  org_freq <- df %>%
    count(Organismo_responsable, name = "n") %>%
    arrange(desc(n))

  for (i in 1:(nrow(org_freq) - 1)) {
    for (j in (i + 1):nrow(org_freq)) {
      if (adist(org_freq$Organismo_responsable[i], org_freq$Organismo_responsable[j]) <= 2) {
        reemplazo <- if (org_freq$n[i] >= org_freq$n[j]) {
          org_freq$Organismo_responsable[i]
        } else {
          org_freq$Organismo_responsable[j]
        }
        df <- df %>%
          mutate(Organismo_responsable = ifelse(
            Organismo_responsable %in% c(org_freq$Organismo_responsable[i],
                                         org_freq$Organismo_responsable[j]),
            reemplazo,
```

```
            Organismo_responsable
          ))
      }
    }
  }
}
```

La misma normalizacion es aplicada a los Nombres de los adjudicatarios o Contratistas

```
if (do_cleaning) {
  df <- df %>%
    rename(nombre_adjudicatario = `nombre_adjudicatario`) %>%
    mutate(nombre_adjudicatario = nombre_adjudicatario %>%
             toupper() %>%
             stringr::str_replace_all("[[:punct:]]", "") %>%
             stringr::str_squish()) %>%
    mutate(nombre_adjudicatario = nombre_adjudicatario %>%
             stringr::str_replace_all("\\bSL\\b", "SLU") %>%
             stringr::str_replace_all("\\bSA\\b", "SAU"))

  adj_freq <- df %>%
    count(nombre_adjudicatario, name = "n") %>%
    arrange(desc(n))

  for (i in 1:(nrow(adj_freq) - 1)) {
    for (j in (i + 1):nrow(adj_freq)) {
      if (adist(adj_freq$nombre_adjudicatario[i], adj_freq$nombre_adjudicatario[j]) <= 2) {
        reemplazo <- if (adj_freq$n[i] >= adj_freq$n[j]) {
          adj_freq$nombre_adjudicatario[i]
        } else {
          adj_freq$nombre_adjudicatario[j]
        }
        df <- df %>%
          mutate(nombre_adjudicatario = ifelse(
            nombre_adjudicatario %in% c(adj_freq$nombre_adjudicatario[i],
                                        adj_freq$nombre_adjudicatario[j]),
            reemplazo,
            nombre_adjudicatario
          ))
      }
    }
  }
}
```

De la misma forma se reduce el valor de los atributos CPV a su codigo, quitando toda referencia textual. De ser necesario se extrae su correspondencia de datasets especificos de CPV

- En formato CSV
- En formato XLSX

```
# función que extrae todos los códigos de 8 dígitos como texto
extraer_cpv <- function(x) {
  codigos <- str_extract_all(x, "\\b\\d{8}\\b")  # lista de códigos
  sapply(codigos, function(vec) paste(vec, collapse = ", "))  # unir si hay más de uno
}

# aplicar a ambas columnas, asegurando tipo character
df <- df %>%
  mutate(
    Materias_CPV = as.character(extraer_cpv(Materias_CPV)),
```

```
    Codigos_CPV  = as.character(extraer_cpv(Codigos_CPV))
  )
df_export <- df %>%
  mutate(
    valor_estimado_licitacion = sprintf("%.2f", valor_estimado_licitacion),
    valor_oferta_adjudicada   = sprintf("%.2f", valor_oferta_adjudicada)
  )

# resumen 5 primeros elementos a partir de la segunda fila
df_resumen <- data.frame(

  Tipo = sapply(df, class),  # tipo de datos por variable
  Muestra = sapply(df, function(x) { # 5 elementos
      paste(head(x, 5), collapse = ", ")
    }
    )
  )

# kable
df_resumen %>%
  kable(col.names = c("Variable", "Tipo de Dato", "5 Primeras observaciones separadas por comas "), align = "
  kable_styling(full_width = FALSE, position = "center", font_size = 9,
                latex_options = "hold_position") %>%
  column_spec(1, width = "4cm") %>%
  column_spec(2, width = "2cm") %>%
  column_spec(3, width = "11cm")
```

Variable	Tipo de Dato	5 Primeras observaciones separadas por comas
Institucion	character	MINISTERIO DE DEFENSA, MINISTERIO DE DEFENSA, MINISTERIO DE DEFENSA, MINISTERIO DE HACIENDA Y ADMINISTRACIONES PÚBLICAS, MINISTERIO DE HACIENDA Y ADMINISTRACIONES PÚBLICAS
Organismo_responsable	character	DIRECCIÓN GENERAL DEL INTA, DIRECCIÓN GENERAL DEL INTA, JEFATURA DE LA SECCIÓN ECONÓMICO ADMINISTRATIVA DEL HOSPITAL CENTRAL DE LA DEFENSA GÓMEZ ULLA, DELEGACIÓN ESPECIAL DE LA AGENCIA TRIBUTARIA EN EXTREMADURA, DIRECCIÓN DEL SERVICIO DE GESTIÓN ECONÓMICA DE LA AGENCIA ESTATAL DE LA ADMINISTRACIÓN TRIBUTARIA
Expediente	character	500083225200., 500083225000., 441/2014., 13060116400., 13700181400.
Fecha	Date	2014-01-03, 2014-01-03, 2014-01-03, 2014-01-03, 2014-01-03
Tipo	character	Contratación, Contratación, Contratación, Contratación, Contratación
Naturaleza	character	Suministros, Suministros, Suministros, Servicios, Obras
Objeto	character	LCVRS de calificación y vuelo de SO/PHI., Componentes electrónicos para SO/PHI., Pescado y Marisco Congelados TA2014., Servicio de Seguridad en los edificios de la Delegación Especial de Extremadura., Castilla y León-D.E. *Obras Complementarias pl. 3ª y Urbanización Nuevo Edificio.
Procedimiento	character	Negociado sin publicidad, Negociado sin publicidad, Abierto, Abierto, Negociado sin publicidad
Ambito_geografico	character	Comunidad de Madrid, Comunidad de Madrid, Comunidad de Madrid, Extremadura, Castilla y León
Materias_CPV	character	38000000, 31000000, 15000000, 79000000, 45000000
Codigos_CPV	character	38000000, 31000000, 15000000, 79710000, 45450000
valor_estimado_licitacion	numeric	120000, 321860, 181818.18, 453081.9, 527216.51
valor_oferta_adjudicada	numeric	120000, 321860, 2e+05, 410602.24, 527216.51
nombre_adjudicatario	character	ARCOPIX SAU, ALTER TECHNOLOGY TÜV NORD SAU, DISBLAMAR SLU, SEGURIDAD INTEGRAL SECOEX SAU, CORSAM CORVIAM CONSTRUCCION SAU

Enlace HTML	character	https://www.boe.es/diario_boe/txt.php?id=BOE-B-2014-201, https://www.boe.es/diario_boe/txt.php?id=BOE-B-2014-202, https://www.boe.es/diario_boe/txt.php?id=BOE-B-2014-203, https://www.boe.es/diario_boe/txt.php?id=BOE-B-2014-205, https://www.boe.es/diario_boe/txt.php?id=BOE-B-2014-206

```
readr::write_csv(df_export, "CSV/licitaciones_contrataciones_BOE_2014_2024_clean.csv")
```

3.3 Valores ausentes, vacíos o nulos

El dataset presenta diversas situaciones de ausencia o irregularidad en los datos, valores `NA`, literales `"No disponible"`, y cadenas vacías `""`. Estos casos han sido estandarizados a `NA` para facilitar un tratamiento homogéneo[2]

El análisis cuantitativo muestra los siguientes porcentajes de valores ausentes o especiales por variable

- **valor_estimado_licitacion**: 54,67% de NA; **valor_oferta_adjudicada**: 35,10% de NA. Estos dos atributos se comportan de forma complementaria, cuando uno está presente, el otro suele estar ausente. Se considera un comportamiento normal en el ciclo de vida de la contratación, si hay adjudicación, el estimado pasa a ser informativo; si no hay adjudicación, sólo se dispone del valor estimado. Por este motivo, **no se realiza imputación** sobre estos campos económicos, ya que no es posible deducir de forma fiable un valor monetario real a partir del otro.
- **Procedimiento**: 5,88% "No disponible". Este porcentaje refleja casos en los que el tipo de procedimiento no está informado en el anuncio. En los análisis descriptivos, estos valores se agruparán como "No especificado". No se procede a su imputación, para no introducir sesgos artificiales en el análisis de procedimientos[2,4]
- **Ambito_geografico**: 6,61% "sin definir". La mayoría de estos casos corresponden a contratos con posible ámbito nacional o para los que no se indica localización explícita. En los análisis territoriales, se agruparán bajo la categoría "Nacional/No especificado".
- **Codigos_CPV**: 8,95% de nulos. La ausencia de código CPV suele deberse a anuncios antiguos o a omisiones en la fuente. En los análisis sectoriales se considerarán como "Sin especificar".
- El resto de variables presentan porcentajes de NA, nulos o "No disponible" **muy inferiores al 1%**, sin impacto relevante sobre la validez global del dataset[3,4]

La incidencia de valores ausentes en variables económicas es inherente al propio ciclo administrativo, y la de los atributos categóricos se mantiene en niveles asumibles para el análisis. Se documenta explícitamente la presencia y tratamiento de estos casos, evitando imputaciones injustificadas y minimizando la introducción de sesgos[2,3]

```
library(dplyr)
library(knitr)
library(kableExtra)

# Función para calcular resumen de faltantes y valores especiales
resumen_valores <- function(x) {
  n <- length(x)
  c(
    NA_pct = round(100 * sum(is.na(x)) / n, 2),
    Nulos_pct = round(100 * sum(x == "" & !is.na(x)) / n, 2),
    Ceros_pct = if (is.numeric(x)) round(100 * sum(x == 0, na.rm = TRUE) / n, 2) else NA,
    No_disponible_pct = round(100 * sum(tolower(x) == "no disponible", na.rm = TRUE) / n, 2),
    Sin_definir_pct = round(100 * sum(tolower(x) == "sin definir", na.rm = TRUE) / n, 2)
  )
}

# Aplica a cada columna
resumen <- sapply(df, resumen_valores)
resumen_df <- as.data.frame(t(resumen))
resumen_df <- cbind(Atributo = rownames(resumen_df), resumen_df)
rownames(resumen_df) <- NULL

# Mostrar tabla ordenada
```

```r
resumen_df %>%
  kable(digits = 2, caption = "Porcentaje de NA, nulos, ceros, 'No disponible' y 'Sin definir' por variable")
  kable_styling(font_size = 9, position = "center", full_width = FALSE)
```

Porcentaje de NA, nulos, ceros, 'No disponible' y 'Sin definir' por variable

Atributo	NA_pct	Nulos_pct	Ceros_pct	No_disponible_pct	Sin_definir_pct
Institucion	0.00	0.00	NA	0.00	0.00
Organismo_responsable	0.00	0.00	NA	0.34	0.00
Expediente	0.00	0.00	NA	1.33	0.00
Fecha	0.00	NA	NA	0.00	0.00
Tipo	0.00	0.00	NA	0.00	0.00
Naturaleza	0.00	0.00	NA	0.03	0.00
Objeto	0.00	0.00	NA	1.06	0.00
Procedimiento	0.00	0.00	NA	5.88	0.00
Ambito_geografico	0.00	0.00	NA	0.04	6.61
Materias_CPV	0.00	0.12	NA	0.00	0.00
Codigos_CPV	0.00	8.95	NA	0.00	0.00
valor_estimado_licitacion	54.67	0.00	0.41	0.00	0.00
valor_oferta_adjudicada	35.10	0.00	1.15	0.00	0.00
nombre_adjudicatario	0.00	0.05	NA	35.10	0.00
Enlace HTML	0.00	0.00	NA	0.00	0.00

3.3.1 Tratamiento de valores ausentes, nulos y casos especiales

Tras analizar la tabla de valores ausentes y especiales, se ha decidido proceder del siguiente modo

- **Registros con `valor_oferta_adjudicada` igual a cero** serán eliminados del dataset, al considerarse contratos sin importe real. Esta medida elimina posibles errores de carga, expedientes anulados o anotaciones administrativas sin efecto económico. Su exclusión no introduce sesgo relevante en el análisis de totales[4,6]
- **Registros con NA en `Codigos_CPV`** se conservarán, ya que la materia contractual se encuentra recogida en `Materias_CPV` (donde los nulos apenas representan el 0,12%). En los análisis sectoriales, los contratos sin código CPV numérico se agruparán bajo "Sin especificar", permitiendo segmentaciones robustas sin pérdida informativa[2,4]
- El resto de variables con NA, "No disponible", nulos o "sin definir" están por debajo del 6% y no serán imputadas ni eliminarán registros, ya que no afectan a la representatividad global del dataset. Su presencia será tenida en cuenta al interpretar los resultados[2,4]
- **Exclusión adicional.** Sólo se eliminarán registros cuyo `valor_oferta_adjudicada` sea un outlier por debajo del Q1 (cuartil 25%) y en los que, además, tanto `Organismo_responsable` como `Expediente` tengan el valor "No disponible". Esto evita la presencia de filas residuales sin información relevante.

De este modo, se garantiza la integridad y la calidad del conjunto de datos, minimizando el sesgo y maximizando la utilidad para el análisis estadístico y económico.

```r
# Eliminar registros con valor_oferta_adjudicada igual a cero
df <- df %>%
  filter(is.na(valor_oferta_adjudicada) | valor_oferta_adjudicada != 0)

# Agrupar los NA de Codigos_CPV como "Sin especificar" (para análisis sectoriales)
df <- df %>%
  mutate(Codigos_CPV = ifelse(is.na(Codigos_CPV) | Codigos_CPV == "", "Sin especificar", Codigos_CPV))

# Calcular Q1
q1 <- quantile(df$valor_oferta_adjudicada, 0.25, na.rm = TRUE)

# Filtrado según criterio adicional
df <- df %>%
  filter(
    is.na(valor_oferta_adjudicada) |
      valor_oferta_adjudicada > q1 |
```

```
      !(tolower(Organismo_responsable) == "no disponible" & tolower(Expediente) == "no disponible")
  )
```

```
library(dplyr)
library(knitr)
library(kableExtra)

# Función para calcular resumen de faltantes y valores especiales
resumen_valores <- function(x) {
  n <- length(x)
  c(
    NA_pct = round(100 * sum(is.na(x)) / n, 2),
    Nulos_pct = round(100 * sum(x == "" & !is.na(x)) / n, 2),
    Ceros_pct = if (is.numeric(x)) round(100 * sum(x == 0, na.rm = TRUE) / n, 2) else NA,
    No_disponible_pct = round(100 * sum(tolower(x) == "no disponible", na.rm = TRUE) / n, 2),
    Sin_definir_pct = round(100 * sum(tolower(x) == "sin definir", na.rm = TRUE) / n, 2)
  )
}

# Aplica a cada columna
resumen <- sapply(df, resumen_valores)
resumen_df <- as.data.frame(t(resumen))
resumen_df <- cbind(Atributo = rownames(resumen_df), resumen_df)
rownames(resumen_df) <- NULL

# Mostrar tabla ordenada
resumen_df %>%
  kable(digits = 2, caption = "Porcentaje de NA, nulos, ceros, 'No disponible' y 'Sin definir' por variable") %>%
  kable_styling(font_size = 9, position = "center", full_width = FALSE)
```

Porcentaje de NA, nulos, ceros, 'No disponible' y 'Sin definir' por variable

Atributo	NA_pct	Nulos_pct	Ceros_pct	No_disponible_pct	Sin_definir_pct
Institucion	0.00	0.00	NA	0.00	0.00
Organismo_responsable	0.00	0.00	NA	0.34	0.00
Expediente	0.00	0.00	NA	1.34	0.00
Fecha	0.00	NA	NA	0.00	0.00
Tipo	0.00	0.00	NA	0.00	0.00
Naturaleza	0.00	0.00	NA	0.03	0.00
Objeto	0.00	0.00	NA	1.07	0.00
Procedimiento	0.00	0.00	NA	5.62	0.00
Ambito_geografico	0.00	0.00	NA	0.04	6.49
Materias_CPV	0.00	0.12	NA	0.00	0.00
Codigos_CPV	0.00	0.00	NA	0.00	0.00
valor_estimado_licitacion	54.48	0.00	0.32	0.00	0.00
valor_oferta_adjudicada	35.51	0.00	0.00	0.00	0.00
nombre_adjudicatario	0.00	0.03	NA	35.51	0.00
Enlace HTML	0.00	0.00	NA	0.00	0.00

```
# número de filas y columnas del dataset
filas_columnas_csv <- data.frame(
  "Descripción" = c("Filas / Observaciones", "Columnas / Variables"),
  "Cantidad" = c(nrow(df), ncol(df))
)

# kable
filas_columnas_csv %>%
  kable(col.names = c("Descripción", "Cantidad")) %>%
```

```
  kable_styling(full_width = FALSE, position = "center", font_size = 9)
```

Descripción	Cantidad
Filas / Observaciones	96032
Columnas / Variables	15

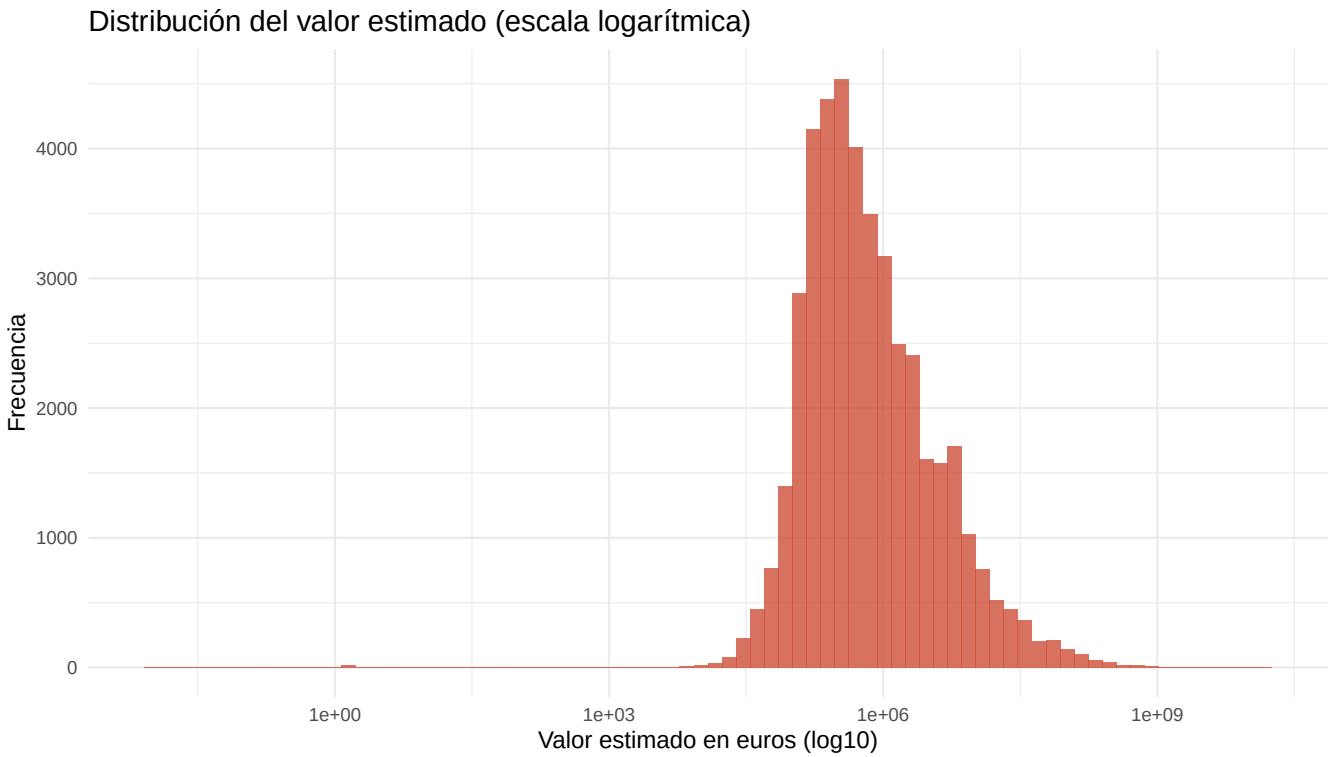
3.3.2 Histogramas y violin plot de valores económicos

Los histogramas muestran la **distribución del valor estimado** y del **valor adjudicado** de los contratos públicos, ambos representados en **escala logarítmica** para una mejor visualización de la dispersión y la asimetría de los datos. La forma de campana ligeramente sesgada hacia la derecha denota que la mayoría de los contratos se sitúan en un rango económico medio, con una concentración notable entre 105 y 106 euros, mientras que los valores extremadamente altos (macrocontratos) y bajos (microcontratos) aparecen como colas más dilatadas en ambos extremos de la distribución[6]

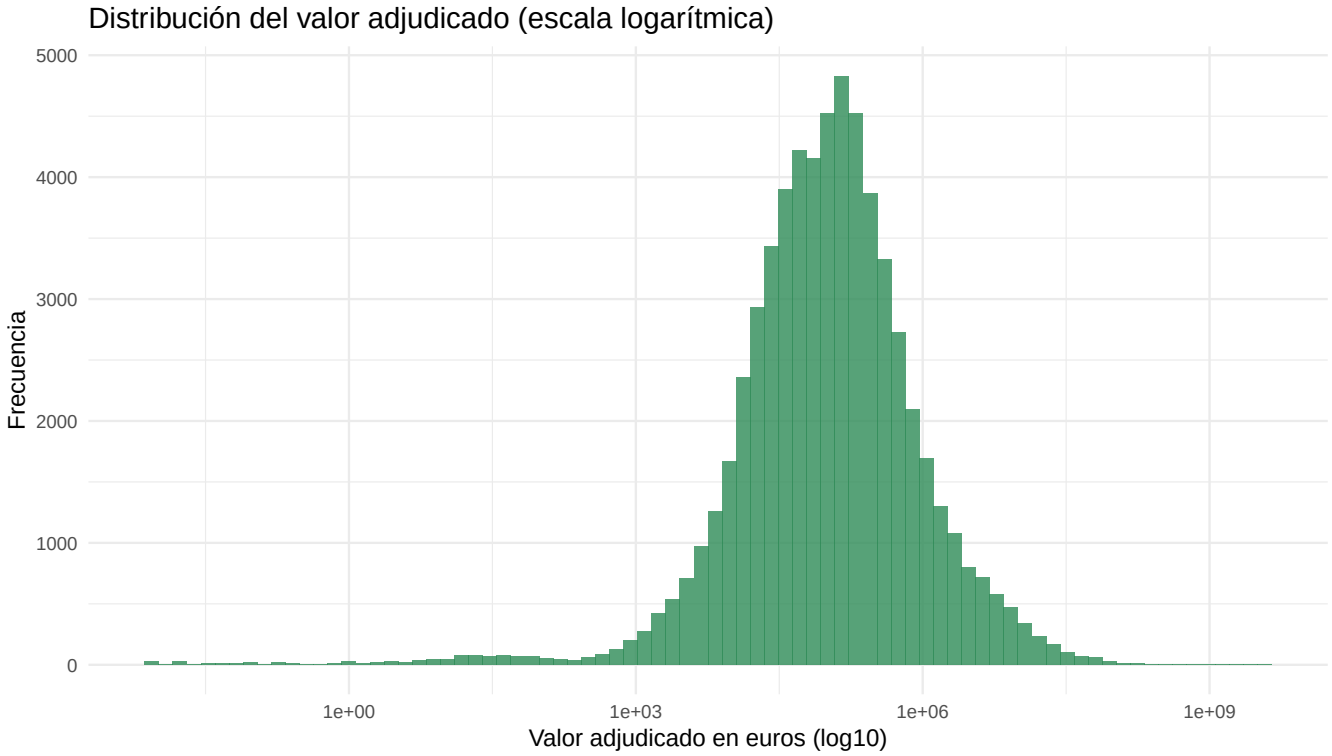
El **violin plot** proporciona una visión conjunta de la distribución de los valores estimados y adjudicados. Se observa que la mediana de los valores adjudicados es, en general, similar a la de los valores estimados, aunque la distribución de los valores adjudicados muestra una mayor concentración en torno a la mediana y colas algo menos pronunciadas que en el estimado. Esto sugiere que los importes finalmente adjudicados tienden a ajustarse a la estimación inicial, aunque persiste cierta variabilidad asociada a la competencia y a las circunstancias específicas de cada procedimiento contractual[4,6]

Estas visualizaciones denotan, además, la **ausencia de sesgos significativos** tras la limpieza y filtrado de registros con importe cero o nulo, validando la consistencia del dataset para posteriores análisis estadísticos y modelado económico[2,4,6]

```
library(ggplot2)

# filtrar valores positivos
df_plot <- df %>%
  filter(valor_estimado_licitacion > 0 | valor_oferta_adjudicada > 0)


# plot estimado
ggplot(df_plot, aes(x = valor_estimado_licitacion)) +
  geom_histogram(bins = 80, fill = "tomato3", alpha = 0.8) +
  scale_x_log10() +
  labs(title = "Distribución del valor estimado (escala logarítmica)",
       x = "Valor estimado en euros (log10)", y = "Frecuencia") +
  theme_minimal(base_size = 9)
```

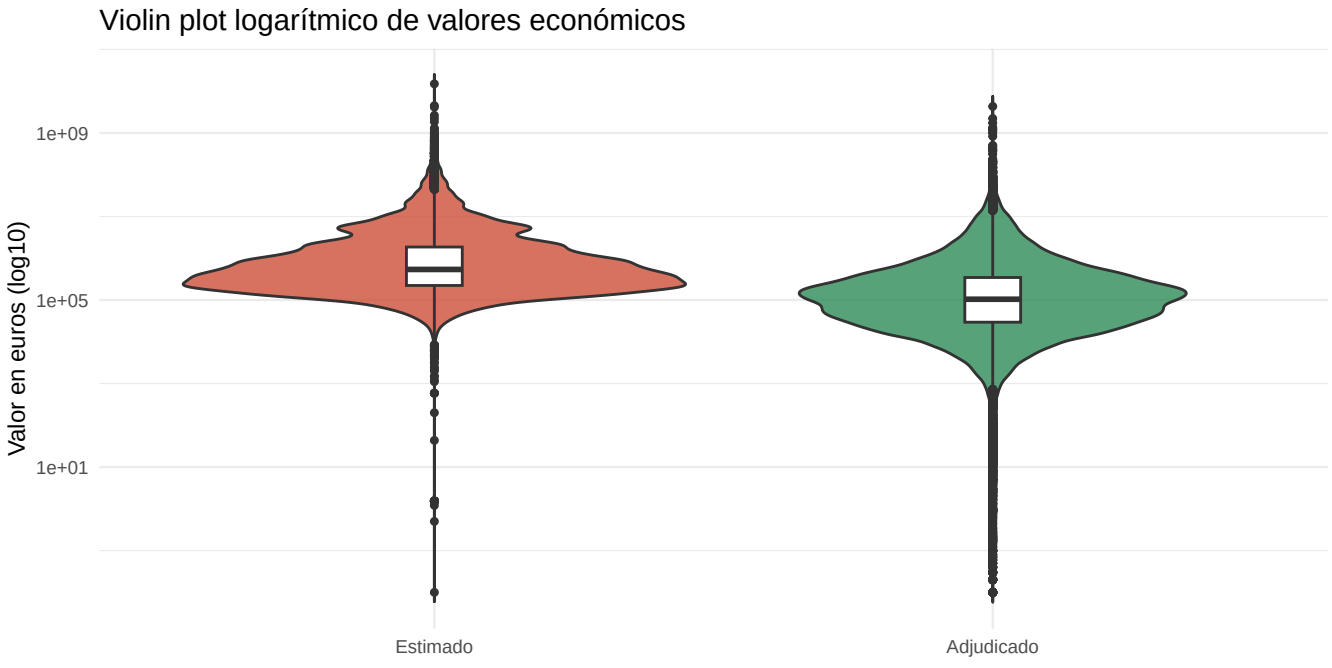



```r
# plot adjudicado
ggplot(df_plot, aes(x = valor_oferta_adjudicada)) +
  geom_histogram(bins = 80, fill = "seagreen", alpha = 0.8) +
  scale_x_log10() +
  labs(title = "Distribución del valor adjudicado (escala logarítmica)",
       x = "Valor adjudicado en euros (log10)", y = "Frecuencia") +
  theme_minimal(base_size = 9)
```




```r
library(ggplot2)
library(tidyr)
```

```r
# preparar datos en formato largo
df_violin <- df %>%
  select(valor_estimado_licitacion, valor_oferta_adjudicada) %>%
  pivot_longer(cols = everything(), names_to = "Tipo", values_to = "Euros") %>%
  filter(!is.na(Euros), Euros > 0)

# renombrar para mayor claridad
df_violin$Tipo <- recode(df_violin$Tipo,
                         valor_estimado_licitacion = "Estimado",
                         valor_oferta_adjudicada = "Adjudicado")

df_violin <- df_violin %>%
  mutate(
    Tipo = trimws(Tipo),  # elimina espacios extra
    Tipo = factor(Tipo, levels = c("Estimado", "Adjudicado"))
  )

mis_colores <- c("Estimado" = "tomato3", "Adjudicado" = "seagreen")


# violin plot ordenado
ggplot(df_violin, aes(x = Tipo, y = Euros, fill = Tipo)) +
  geom_violin(trim = FALSE, alpha = 0.8) +
  scale_fill_manual(values = mis_colores) +
  geom_boxplot(width = 0.1, outlier.size = 0.9, fill = "white") +
  scale_y_log10() +
  labs(title = "Violin plot logarítmico de valores económicos",
       x = NULL, y = "Valor en euros (log10)") +
  theme_minimal(base_size = 9) +
  theme(legend.position = "none")
```



3.4 Identificación y tratamiento de valores extremos

La detección y gestión de valores extremos o atípicos se centra principalmente en variables económicas, en particular el *valor adjudicado*[2,3,4]

Se ha optado por una **clasificación en tres segmentos**

- **Microcontratos.** Aquellos cuyo *valor adjudicado* es inferior al primer cuartil (Q1) de la distribución.
- **Contratos estándar.** Aquellos situados entre Q1 y Q3 (cuartiles 1 y 3).
- **Macrocontratos.** Aquellos cuyo *valor adjudicado* supera el tercer cuartil (Q3)[3,6]

Esta segmentación sigue los criterios robustos (cuartiles e IQR) recomendados en la literatura para el análisis de datos económicos y facilita la comparación y el estudio de patrones diferenciados según el tamaño de los contratos[3,4,6]

No se eliminarán automáticamente los valores extremos detectados, pero sí es necesario se etiquetarán y se analizarán por separado si distorsionan los resultados. Los valores igual a cero ya han sido eliminados previamente en el apartado anterior, y los importes superiores a umbrales poco plausibles se revisan individualmente[2,6]

```
# Filtrar sólo registros de contratación con valor adjudicado NO nulo
df_contratacion <- df %>%
  filter(Tipo == "Contratación" & !is.na(valor_oferta_adjudicada))

# Calcular cuartiles Q1 y Q3 de valor adjudicado (sobre las contrataciones)
q1 <- quantile(df_contratacion$valor_oferta_adjudicada, 0.25, na.rm = TRUE)
q3 <- quantile(df_contratacion$valor_oferta_adjudicada, 0.75, na.rm = TRUE)

# Clasificar según cuartiles
df_contratacion <- df_contratacion %>%
  mutate(
    tipo_contrato = case_when(
      valor_oferta_adjudicada < q1 ~ "microcontrato",
      valor_oferta_adjudicada >= q1 & valor_oferta_adjudicada <= q3 ~ "contrato estándar",
      valor_oferta_adjudicada > q3 ~ "macrocontrato",
      TRUE ~ NA_character_
    )
  )

# Tabla resumen por segmento (sin NA)
tabla_tipo_contrato <- df_contratacion %>%
  filter(!is.na(tipo_contrato)) %>%
  count(tipo_contrato, name = "N") %>%
  mutate(Porcentaje = round(100 * N / sum(N), 2))

tabla_tipo_contrato %>%
  kable(
    align = "c",
    caption = "Clasificación de contratos adjudicados por importe económico - cuartiles Q1 y Q3"
  ) %>%
  kable_styling(font_size = 9, full_width = FALSE, position = "center")
```

Clasificación de contratos adjudicados por importe económico - cuartiles Q1 y Q3

tipo_contrato	N	Porcentaje
contrato estándar	30965	50
macrocontrato	15483	25
microcontrato	15483	25

```
# Estadísticos descriptivos sólo sobre contrataciones válidas
df_contratacion %>%
  filter(!is.na(tipo_contrato)) %>%
  group_by(tipo_contrato) %>%
  summarise(
    N = n(),
    Mediana = median(valor_oferta_adjudicada, na.rm = TRUE),
    Mínimo = min(valor_oferta_adjudicada, na.rm = TRUE),
    Máximo = max(valor_oferta_adjudicada, na.rm = TRUE)
  ) %>%
```

```
  kable(
    caption = "Estadísticos descriptivos por segmento de contrato (sólo contrataciones adjudicadas)",
    digits = 0
  ) %>%
  kable_styling(font_size = 9, full_width = FALSE, position = "center")
```

Estadísticos descriptivos por segmento de contrato (sólo contrataciones adjudicadas)

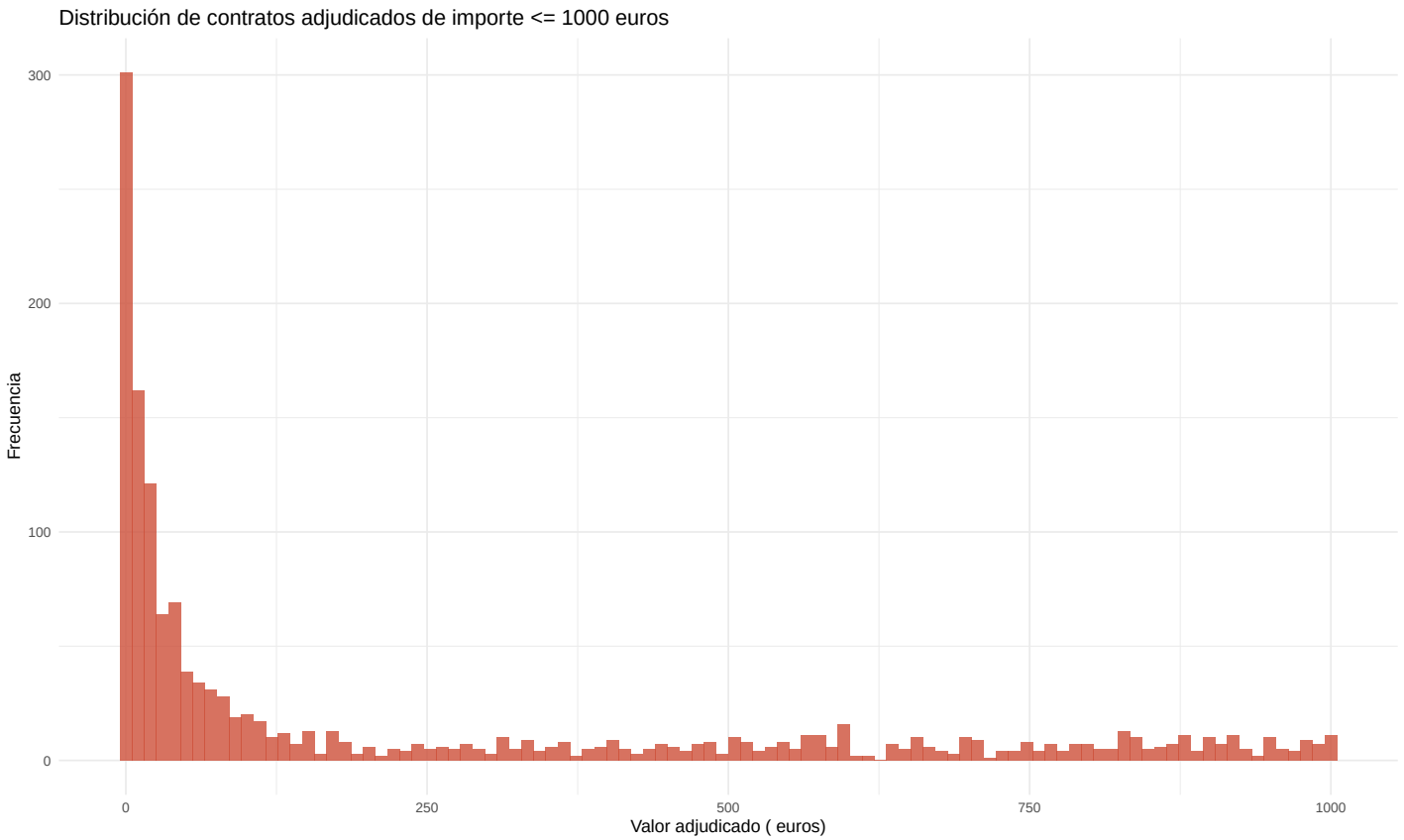
tipo_contrato	N	Mediana	Mínimo	Máximo
contrato estándar	30965	104447	29485	346400
macrocontrato	15483	912738	346450	4317800000
microcontrato	15483	12009	0	29470

```
# library(dplyr)
# library(ggplot2)
# library(knitr)
# library(kableExtra)

# sólo contratos adjudicados con valor > 0
df_contratacion <- df %>%
  filter(Tipo == "Contratación" & !is.na(valor_oferta_adjudicada) & valor_oferta_adjudicada > 0)

# Tabla de tramos (con tramo abierto hasta Inf)
tabla_pequenos <- df_contratacion %>%
  mutate(tramo = cut(
    valor_oferta_adjudicada,
    breaks = c(0, 1, 5, 10, 50, 100, 500, 1000, 5000, 10000, Inf),
    labels = c("0-1 euros", "1-5 euros", "5-10 euros", "10-50 euros", "50-100 euros",
               "100-500 euros", "500-1000 euros", "1000-5000 euros", "5000-10000 euros", ">10,000 euros"),
    right = TRUE,
    include.lowest = TRUE
  )) %>%
  group_by(tramo) %>%
  summarise(N = n()) %>%
  mutate(Porcentaje = round(100 * N / sum(N), 4))

# Mostrar tabla sin NA en los tramos
tabla_pequenos %>%
  kable(
    caption = "Distribución de contratos adjudicados por tramos de importe",
    align = "c"
  ) %>%
  kable_styling(font_size = 9, full_width = FALSE, position = "center")
```

Distribución de contratos adjudicados por tramos de importe

tramo	N	Porcentaje
0-1 euros	200	0.3229
1-5 euros	100	0.1615
5-10 euros	84	0.1356
10-50 euros	354	0.5716
50-100 euros	140	0.2261
100-500 euros	265	0.4279
500-1000 euros	334	0.5393
1000-5000 euros	2584	4.1724
5000-10000 euros	2755	4.4485
>10,000 euros	55115	88.9942

```r
# Pivotar los 20 contratos de menor valor adjudicado
tabla_pivot <- df_contratacion %>%
  arrange(valor_oferta_adjudicada) %>%
  select(Expediente, Objeto, valor_oferta_adjudicada) %>%
  head(20) %>%
  t() %>%
  as.data.frame()

# Ajustar nombres de filas para mayor claridad
rownames(tabla_pivot) <- c("Expediente", "Objeto", "Valor adjudicado (euros)")

# Histograma de valores bajos
ggplot(df_contratacion %>% filter(valor_oferta_adjudicada <= 1000),
       aes(x = valor_oferta_adjudicada)) +
  geom_histogram(bins = 100, fill = "tomato3", alpha = 0.8) +
  labs(title = "Distribución de contratos adjudicados de importe <= 1000 euros",
       x = "Valor adjudicado ( euros)", y = "Frecuencia") +
  theme_minimal(base_size = 9)
```




```r
# 1. Número de contratos antes de filtrar
n_total <- nrow(df_contratacion)

# 2. Filtrar contratos >= 1000  euros
df_contratacion_filtrado <- df_contratacion %>%
  filter(!is.na(valor_oferta_adjudicada) & valor_oferta_adjudicada >= 1000)

# 3. Número de contratos tras filtrar
n_filtrado <- nrow(df_contratacion_filtrado)

# 4. Porcentaje eliminado
porc_eliminado <- round(100 * (n_total - n_filtrado) / n_total, 4)
```

```r
# 5. Tabla resumen
data.frame(
  "Total original" = n_total,
  "Total tras filtrado" = n_filtrado,
  "Eliminados (<1000  euros)" = n_total - n_filtrado,
  "Porcentaje eliminado" = porc_eliminado
) %>%
  kable(caption = "Resumen del filtrado de contratos de bajo importe",
        align = "c") %>%
  kable_styling(font_size = 9, position = "center", full_width = FALSE)
```

Resumen del filtrado de contratos de bajo importe

Total.original	Total.tras.filtrado	Eliminados...1000..euros.	Porcentaje.eliminado
61931	60456	1475	2.3817

Se ha eliminado del conjunto de datos todos los contratos adjudicados cuyo valor es inferior a 1000 euros, ya que representan únicamente un 2,38 % del total de registros. Esta decisión se basa en

- La baja relevancia económica y estadística de estos contratos en el análisis global.
- La alta probabilidad de que valores extremadamente bajos correspondan a errores de extracción, microcontratos administrativos o anotaciones residuales no representativas[2,4,6]
- Al tratarse de un porcentaje pequeño, la exclusión no introduce un sesgo significativo en la estimación de sumatorios ni afecta a la representatividad del dataset.

Este enfoque es habitual en estudios de contratación pública y análisis económico, donde se busca evitar la distorsión provocada por valores atípicos irrelevantes en la "cola izquierda" de la distribución[2,4,6]

3.5 Conversión de variables categóricas

Todas las variables que representan categorías identificables (`Tipo`, `Naturaleza`, `Procedimiento`, `Institucion`, etc.) serán transformadas a tipo `factor`. Este paso facilita tanto la eficiencia de cálculo como la presentación visual, y permite preparar el dataset para análisis estadísticos y modelos supervisados[1,2]

Durante este proceso, se aplicará una **normalización textual** sobre los niveles de los factores[1,3]

- Eliminación de mayúsculas/minúsculas inconsistentes.
- Unificación de nombres con diferencias por puntuación u ortografía (`"ADIF - Presidencia"` vs `"Adif - Presidencia."`).
- Reagrupación manual en casos necesarios, como unificar variantes de organismos bajo una categoría común.

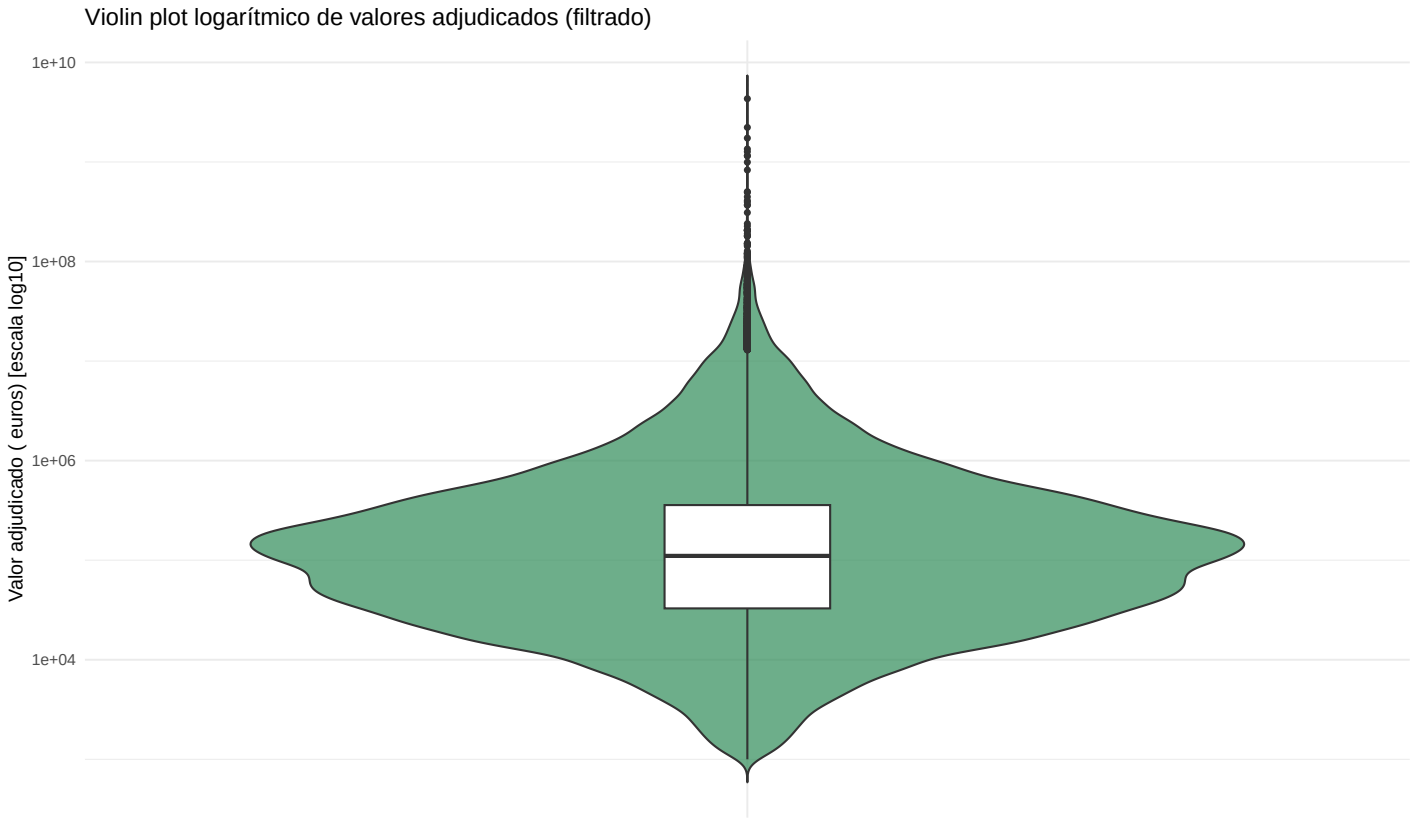
```r
library(knitr)
library(kableExtra)

# selección y transposición de muestra
tabla_muestra <- df_contratacion_filtrado %>%
  select(Fecha, Institucion, Organismo_responsable, Naturaleza, nombre_adjudicatario,valor_oferta_adjudicada)
  head(5) %>%
  t() %>%
  as.data.frame()

# eliminar encabezados V1, V2… y mostrar filas directamente
rownames(tabla_muestra) <- c("Fecha", "Institución", "Organismo", "Naturaleza", "Contratista", "Valor (euros)

# imprimir sin encabezado de columna (col.names = NULL)
kable(tabla_muestra, col.names = NULL, align = "l",
      caption = "Muestra de contratos adjudicados (tabla transpuesta)") %>%
  kable_styling(
    font_size = 9,
```

```
    latex_options = c("hold_position", "scale_down"),
    position = "center"
  ) %>%
  column_spec(1, bold = TRUE, width = "20mm") %>%
  column_spec(2:6, width = "25mm")
```

Muestra de contratos adjudicados (tabla transpuesta)

Fecha	2014-01-03	2014-01-03	2014-01-03	2014-01-03	2014-01-03
Institución	MINISTERIO DE DEFENSA	MINISTERIO DE DEFENSA	MINISTERIO DE DEFENSA	MINISTERIO DE HACIENDA Y ADMINISTRACIONES PÚBLICAS	MINISTERIO DE HACIENDA Y ADMINISTRACIONES PÚBLICAS
Organismo	DIRECCIÓN GENERAL DEL INTA	DIRECCIÓN GENERAL DEL INTA	JEFATURA DE LA SECCIÓN ECONÓMICO ADMINISTRATIVA DEL HOSPITAL CENTRAL DE LA DEFENSA GÓMEZ ULLA	DELEGACIÓN ESPECIAL DE LA AGENCIA TRIBUTARIA EN EXTREMADURA	DIRECCIÓN DEL SERVICIO DE GESTIÓN ECONÓMICA DE LA AGENCIA ESTATAL DE LA ADMINISTRACIÓN TRIBUTARIA
Naturaleza	Suministros	Suministros	Suministros	Servicios	Obras
Contratista	ARCOPIX SAU	ALTER TECHNOLOGY TÜV NORD SAU	DISBLAMAR SLU	SEGURIDAD INTEGRAL SECOEX SAU	CORSAM CORVIAM CONSTRUCCION SAU
Valor (euros)	120000.0	321860.0	200000.0	410602.2	527216.5

```
ggplot(df_contratacion_filtrado, aes(y = valor_oferta_adjudicada, x = "")) +
  geom_violin(fill = "seagreen", alpha = 0.7, trim = FALSE) +
  geom_boxplot(width = 0.15, fill = "white", outlier.size = 1) +
  scale_y_log10() +
  labs(
    title = "Violin plot logarítmico de valores adjudicados (filtrado)",
    y = "Valor adjudicado ( euros) [escala log10]",
    x = NULL
  ) +
  theme_minimal(base_size = 10)
```




```
df_contratacion_final <- df_contratacion_filtrado %>%
  filter(Tipo == "Contratación") %>%
  select(-Tipo, -valor_estimado_licitacion, -'Enlace HTML')

skim(df_contratacion_final)
```

Data summary

Name	df_contratacion_final
Number of rows	60456
Number of columns	12
Column type frequency:	
character	10
Date	1
numeric	1
Group variables	None

Variable type: character

skim_variable	n_missing	complete_rate	min	max	empty	n_unique	whitespace
Institucion	0	1	13	77	0	79	0
Organismo_responsable	0	1	1	381	0	1203	0
Expediente	0	1	4	178	0	47007	0
Naturaleza	0	1	5	29	0	10	0
Objeto	1	1	7	1144	0	46392	0
Procedimiento	0	1	7	24	0	7	0
Ambito_geografico	0	1	6	237	0	353	0
Materias_CPV	0	1	0	108	46	1125	0
Codigos_CPV	0	1	8	6078	0	10454	0

skim_variable	n_missing	complete_rate	min	max	empty	n_unique	whitespace
nombre_adjudicatario	0	1	0	714	29	16502	0

Variable type: Date

skim_variable	n_missing	complete_rate	min	max	median	n_unique
Fecha	0	1	2014-01-03	2024-12-31	2021-05-07	2994

Variable type: numeric

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100	hist
valor_oferta_adjudicada	0	1	1243788	24898186	1000	32786.08	110436.4	357786	4317800000	

```
# guardar como nuevo CSV limpio y final
readr::write_csv(df_contratacion_final, "CSV/contrataciones_BOE_2014_2024.csv")
```

Tras aplicar las fases de limpieza y depuración, el conjunto de datos final refleja una estructura más robusta y fiable para el análisis de la contratación pública. Se han eliminado los registros con valores adjudicados igual a cero, así como los contratos de importe inferior a 1.000 euros, permitiendo descartar posibles errores de extracción, anotaciones administrativas residuales y microcontratos de baja relevancia estadística[2,4,6]

El resultado es un dataset que conserva la integridad de la información relevante, minimiza la presencia de valores atípicos en la "cola izquierda" y mantiene la representatividad agregada, con la garantía de que no se introduce sesgo significativo en la muestra.

El **violin plot logarítmico** de los valores adjudicados (tras filtrado) denota claramente la distribución asimétrica de los importes de contratación

- La mayor densidad se concentra en un rango intermedio-alto (aproximadamente entre 10.000 euros y 1 millón de euros), mientras que los contratos de importes extremadamente altos (macrocontratos) forman una cola prolongada, tal y como es esperable en la contratación pública.
- El boxplot central confirma una mediana situada en la zona esperada, sin desplazamientos artificiales, y las colas estrechas hacia la parte inferior evidencian que los valores bajos han sido correctamente tratados.
- La eliminación de valores pequeños permite que la interpretación estadística y la segmentación por tamaño contractual sean más fiables, sin distorsiones provocadas por registros residuales.

En resumen, el dataset resultante está preparado para análisis económicos avanzados, comparativas sectoriales o regionales, y modelado predictivo, reflejando fielmente la estructura del gasto público estatal y la dinámica real de los procesos de adjudicación[2,4,6]

Tras completar la limpieza y depuración, se guarda el dataset resultante como **'contrataciones_BOE_2014_2024.csv'**. Esto permite garantizar que todos los análisis posteriores se basan únicamente en datos consistentes, depurados y relevantes para el estudio económico y estadístico de la contratación pública.

4 Análisis de los Datos

A partir del apartado 4, se trabaja únicamente con el dataset df_contratacion_final, que recoge las observaciones ya filtradas y depuradas para que incluyan sólo contratos adjudicados, excluyendo licitaciones sin adjudicación, registros de importe irrelevante y microcontratos por debajo de 1000 €. De esta forma se asegura que el análisis y los modelos aplicados se basen en datos homogéneos y comparables, evitando sesgos introducidos por expedientes no equiparables o por errores residuales de captura y transcripción[1]

4.1 Modelo supervisado. Regresión lineal múltiple

Se ha seleccionado la regresión lineal múltiple por

- Interpretabilidad. Permite cuantificar el efecto de cada variable categórica y temporal sobre el importe adjudicado, facilitando la explicación y comparación sectorial.
- Uso estándar en análisis económico-administrativo. Es el enfoque clásico en estudios de contratación pública y datasets administrativos, y favorece la reproducibilidad y comparación con literatura previa[3,4]
- Manejo eficiente de variables categóricas. El modelo gestiona de forma natural factores con múltiples niveles, algo imprescindible en este tipo de análisis.
- Robustez y claridad. Aunque otros modelos pueden superar en precisión a la regresión lineal múltiple en escenarios no lineales, este modelo es suficiente para captar tendencias agregadas y efectos principales, siempre que se interprete con la debida cautela ante outliers y alta heterogeneidad[2,6]

4.1.1 Preparación de variables

- Conversión del campo Fecha a año (Ano). Esto permite incorporar el efecto temporal y estudiar posibles tendencias o ciclos anuales en la adjudicación pública.
- Identificación y codificación de variables categóricas. Se convierten en factores (Institucion, Naturaleza, Ambito_geografico, Procedimiento, Organismo_responsable), imprescindible para su uso en modelos lineales en R, donde cada nivel se traduce en una variable dummy[2,4]

```
library(lubridate)

df_contratacion_final <- df_contratacion_final %>%
  mutate(Ano = year(Fecha))

set.seed(123)
n <- nrow(df_contratacion_final)
idx_train <- sample(seq_len(n), size = floor(0.7 * n))
train <- df_contratacion_final[idx_train, ]
test  <- df_contratacion_final[-idx_train, ]

factores <- c("Institucion", "Naturaleza", "Ambito_geografico", "Procedimiento", "Organismo_responsable")
for (f in factores) {
  train[[f]] <- droplevels(factor(train[[f]]))
  test[[f]]  <- factor(test[[f]], levels = levels(train[[f]]))
}

# ajusta el modelo
modelo_reg <- lm(valor_oferta_adjudicada ~ Institucion + Naturaleza + Ambito_geografico + Procedimiento + Ano

# predicción sobre test
pred <- predict(modelo_reg, newdata = test)

# resumen de métricas
library(Metrics)
mae <- mae(test$valor_oferta_adjudicada, pred)
rmse <- rmse(test$valor_oferta_adjudicada, pred)

# Coeficientes principales
coef <- summary(modelo_reg)$coefficients
rownames(coef)[1] <- "Intercepto"
coef_abs <- coef[order(abs(coef[,1]), decreasing = TRUE), ]
```

4.1.2 División en training y test

Se divide aleatoriamente el dataset en dos subconjuntos

- Entrenamiento (70%). Para ajustar el modelo.
- Test (30%). Para evaluar la capacidad predictiva y evitar el sobreajuste.

Esta proporción sigue recomendaciones estándar en ciencia de datos y asegura que la evaluación se realice sobre datos no vistos durante el entrenamiento[3,2]

Homogeneización de niveles en los factores

Para evitar errores en la predicción, como niveles de factor presentes en test pero no en train, se fuerzan los mismos niveles de factores en ambos subconjuntos. Se eliminan niveles no usados en el entrenamiento (droplevels), evitando advertencias y problemas en la generación de matrices de diseño[1,2]

```
df_contratacion_final <- df_contratacion_final %>%
  mutate(Ano = year(Fecha))

# Selecciona sólo columnas relevantes
factores <- c("Institucion", "Naturaleza", "Ambito_geografico", "Procedimiento", "Organismo_responsable")
modelo_vars <- c(factores, "Ano", "valor_oferta_adjudicada")
df_modelo <- df_contratacion_final %>%
  select(all_of(modelo_vars)) %>%
  drop_na()

set.seed(123)
n <- nrow(df_modelo)
idx_train <- sample(seq_len(n), size = floor(0.7 * n))
train <- df_modelo[idx_train, ]
test  <- df_modelo[-idx_train, ]

for (f in factores) {
  train[[f]] <- droplevels(factor(train[[f]]))
  test[[f]]  <- factor(test[[f]], levels = levels(train[[f]]))
}
test <- test %>% filter(complete.cases(.))

modelo_reg <- lm(valor_oferta_adjudicada ~ Institucion + Naturaleza + Ambito_geografico + Procedimiento + Ano
pred <- predict(modelo_reg, newdata = test)


# Frecuencia por año
tabla_ano <- df_contratacion_final %>%
  mutate(Ano = year(Fecha)) %>%
  count(Ano) %>%
  arrange(Ano)
kable(tabla_ano, caption = "Frecuencia de contratos adjudicados por año") %>%
  kable_styling(font_size = 9, position = "center")
```

Frecuencia de contratos adjudicados por año

Ano	n
2014	2338
2015	2455
2016	2790
2017	3786
2018	4622
2019	5892
2020	6223
2021	7440
2022	7769
2023	8237
2024	8904

4.1.3 Tablas y Métricas

Se ajusta un modelo de regresión lineal múltiple con la variable respuesta valor_oferta_adjudicada y como predictores todas las variables categóricas clave y el año (Ano). Se trata de un modelo clásico para análisis explicativos y para cuantificar el peso de las diferentes variables sobre el importe final adjudicado[2,3,4]

La evaluación se realiza sobre el conjunto de test, extrayendo métricas de error MAE (error absoluto medio) y RMSE (raíz del error cuadrático medio), que permiten cuantificar la precisión y dispersión del modelo[2,4]

```
# Dataframe para visualizaciones
df_pred <- test %>%
  mutate(prediccion = pred)

# Frecuencia por ámbito geográfico
tabla_ambito <- df_contratacion_final %>%
  count(Ambito_geografico) %>%
  arrange(desc(n)) %>%
  head(15) # top 15 para no desbordar
kable(tabla_ambito, caption = "Top 15 ámbitos geográficos por número de contratos") %>%
  kable_styling(font_size = 9, position = "center")
```

Top 15 ámbitos geográficos por número de contratos

Ambito_geografico	n
Comunidad de Madrid	22138
Andalucía	5792
Nacional	4439
Sin definir	3782
Galicia	2773
Comunidad Valenciana	2770
Castilla y León	2324
Cataluña	2166
Castilla-La Mancha	1930
Aragón	1581
Región de Murcia	1463
Ciudad de Ceuta	1110
Canarias	1083
Extremadura	1041
Illes Balears	770

```
tabla_cruzada_naturaleza_ano <- df_contratacion_final %>%
  mutate(Ano = year(Fecha)) %>%
  count(Naturaleza, Ano) %>%
  pivot_wider(names_from = Ano, values_from = n, values_fill = 0)

kable(tabla_cruzada_naturaleza_ano, caption = "Contratos por naturaleza y año") %>%
  kable_styling()
```

Contratos por naturaleza y año

Naturaleza	2014	2015	2016	2017	2018	2019	2020	2021	2022	2023	2024
Administrativo especial	51	66	32	40	1	5	1	0	0	0	0
Gestión de Servicios Públicos	3	2	41	8	5	2	1	0	0	0	0
Mixto - Obras	0	0	0	0	4	3	11	10	29	17	18
Mixto - Servicios	0	0	0	1	1	2	1	12	11	17	4
Mixto - Suministros	0	1	0	1	7	19	45	68	68	46	47
No disponible	0	0	1	0	0	0	0	1	0	0	0
Obras	310	315	368	327	351	489	482	756	844	900	935

Privado	4	20	10	20	30	20	1	0	0	0	0
Servicios	903	975	1303	1847	2370	3452	3537	4201	4064	4313	4789
Suministros	1067	1076	1035	1542	1853	1900	2144	2392	2753	2944	3111

```r
# Estadísticos descriptivos de valores adjudicados
stats_adjudicado <- df_contratacion_final %>%
  summarise(
    media = mean(valor_oferta_adjudicada, na.rm = TRUE),
    mediana = median(valor_oferta_adjudicada, na.rm = TRUE),
    minimo = min(valor_oferta_adjudicada, na.rm = TRUE),
    maximo = max(valor_oferta_adjudicada, na.rm = TRUE)
  )
kable(stats_adjudicado, caption = "Estadísticos descriptivos del valor adjudicado") %>%
  kable_styling(font_size = 9, position = "center")
```

Estadísticos descriptivos del valor adjudicado

media	mediana	minimo	maximo
1243788	110436.4	1000	4317800000

```r
# Calcula todas las métricas
mae_val <- Metrics::mae(test$valor_oferta_adjudicada, pred)
rmse_val <- Metrics::rmse(test$valor_oferta_adjudicada, pred)
summary_reg <- summary(modelo_reg)
r2 <- summary_reg$r.squared
r2_adj <- summary_reg$adj.r.squared

# Junta todo en una tabla pivotada (métrica como fila)
tabla_metricas <- data.frame(
  Métrica = c("MAE", "RMSE", "R2", "R2 ajustado"),
  Valor = c(round(mae_val, 2), round(rmse_val, 2), round(r2, 3), round(r2_adj, 3))
)

# Imprime tabla compacta y centrada
kable(tabla_metricas, caption = "Métricas de rendimiento del modelo supervisado") %>%
  kable_styling(font_size = 9, position = "center", full_width = FALSE)
```

Métricas de rendimiento del modelo supervisado

Métrica	Valor
MAE	1911532.770
RMSE	11066601.100
R2	0.014
R2 ajustado	0.005

Estas tablas denotan que la **contratación pública en España está concentrada geográficamente y por sectores**, presenta una alta asimetría en los importes y que los modelos lineales clásicos tienen un poder explicativo limitado, aunque resultan útiles para captar tendencias globales y orientar análisis más detallados

Top 15 ámbitos geográficos por número de contratos

Esta tabla recoge las **15 principales comunidades autónomas y áreas geográficas** por volumen de contratos adjudicados. Se observa una clara concentración en la Comunidad de Madrid y Andalucía, seguidas de la categoría "Nacional" y otras regiones, denotando la centralización del gasto público en grandes polos administrativos y económicos. El resto de ámbitos muestra valores significativamente inferiores, apuntando a una distribución asimétrica del número de contratos entre territorios[2,6]

Contratos por naturaleza y año

En esta tabla se desglosa la **evolución anual de los contratos adjudicados según la naturaleza del contrato** (Obras, Servicios, Suministros, etc.). Se aprecia un predominio claro de los servicios y suministros a lo largo del periodo, siendo "Obras" el tercer gran grupo. Se detecta un crecimiento progresivo de la contratación pública desde 2014, con máximos en los años más recientes, pudiendo estar relacionado tanto con factores económicos como administrativos o coyunturales, como los fondos europeos Next Generation en 2021-2023[2,6]

Estadísticos descriptivos del valor adjudicado

Esta tabla muestra los **principales estadísticos de los importes adjudicados**, media, mediana, mínimo y máximo. Se observa una gran diferencia entre la media y la mediana, mostrando una fuerte asimetría positiva típica de este tipo de datos, donde la mayoría de contratos son de importe medio/bajo, pero existen unos pocos macrocontratos de cuantía muy elevada. El rango de importes adjudicados abarca desde microcontratos (1.000 €) hasta grandes contratos por encima de los 4.000 millones de euros[2,3,6]

Métricas de rendimiento del modelo supervisado

En la tabla final se recogen las **principales métricas de ajuste del modelo de regresión**

- **MAE** (Error absoluto medio) y **RMSE** (raíz del error cuadrático medio). Ambos valores elevados en términos absolutos, interpretandose como reflejo del rango y dispersión real de los importes en la contratación pública
- **R^2** y **R^2 ajustado**. Ambos bajos, denotando que las variables seleccionadas (institución, naturaleza, ámbito, procedimiento y año) **explican sólo una pequeña parte de la variabilidad total**. Este resultado es habitual en modelos aplicados a datos heterogéneos y sectoriales, donde la ausencia de atributos adicionales relevantes, como pudiera ser la duración del contrato, el número de contratistas, la complejidad técnica o el tipo de proyecto, limita la capacidad explicativa de la regresión lineal. La inclusión de variables financieras o técnicas específicas habría permitido, previsiblemente, mejorar de forma significativa los valores de R^2 y R^2 ajustado, alineando el modelo con las mejores prácticas en análisis económico y administración pública[3,4]

4.1.4 Principales coeficientes del modelo

Top 10 coeficientes absolutos del modelo

Variable	Estimate	Std. Error	t value	Pr > t
Ambito_geografico: Illes Balears, Ciudad de Ceuta, Ciudad de Melilla, Canarias	368565766	28483765	12.9395033	0.0000000
Ambito_geografico: Nacional, Illes Balears, Ciudad de Ceuta, Ciudad de Melilla, Canarias	254090684	20154109	12.6073886	0.0000000
Intercepto	81042114	148746011	0.5448355	0.5858695
Naturaleza: Gestión de Servicios Públicos	52702737	5192482	10.1498166	0.0000000
Ambito_geografico: Comunidad Foral de Navarra, Aragón, Comunidad Valenciana	34627415	28478665	1.2159072	0.2240271
Ambito_geografico: Galicia, Cantabria, País Vasco, La Rioja, Aragón	17141820	28519573	0.6010546	0.5478069
Ambito_geografico: Illes Balears, Canarias	16475503	11685432	1.4099182	0.1585713
InstitucionMINISTERI: O D: E POL: ÍTIC: A TERRITORIA: L Y FUNCI: ÓN P: ÚBLIC: A	15226619	2605957	5.8430051	0.0000000
Ambito_geografico: Comunidad de Madrid, Ciudad de Melilla	10930284	28476012	0.3838418	0.7010976
Procedimiento: Diálogo competitivo	8081893	20155869	0.4009697	0.6884445

La tabla evidencia cómo determinadas áreas territoriales, tipos de contrato y algunos organismos concentran la mayor parte de la variabilidad explicada por el modelo, alineándose con los patrones detectados previamente en los análisis exploratorios. El resultado refuerza la idea de que el importe adjudicado en la contratación pública depende de una combinación de factores estructurales, sobre todo geográficos y funcionales, y subraya la conveniencia de emplear modelos explicativos para interpretar y cuantificar este impacto[2,3,4,6]

La tabla presenta los **10 coeficientes absolutos más relevantes** del modelo de regresión lineal múltiple ajustado sobre los contratos adjudicados. Cada fila de la tabla corresponde a un término del modelo, ya sea un nivel específico de una variable categórica, el intercepto o una combinación de categorías agrupadas por R en la codificación de factores

- **Variable.** Denota el nombre de la categoría o nivel que impacta significativamente en el importe adjudicado
- **Estimate.** Es el valor estimado del coeficiente, es decir, cuánto se incrementa o reduce el importe esperado al estar presente esa categoría, manteniendo el resto constante

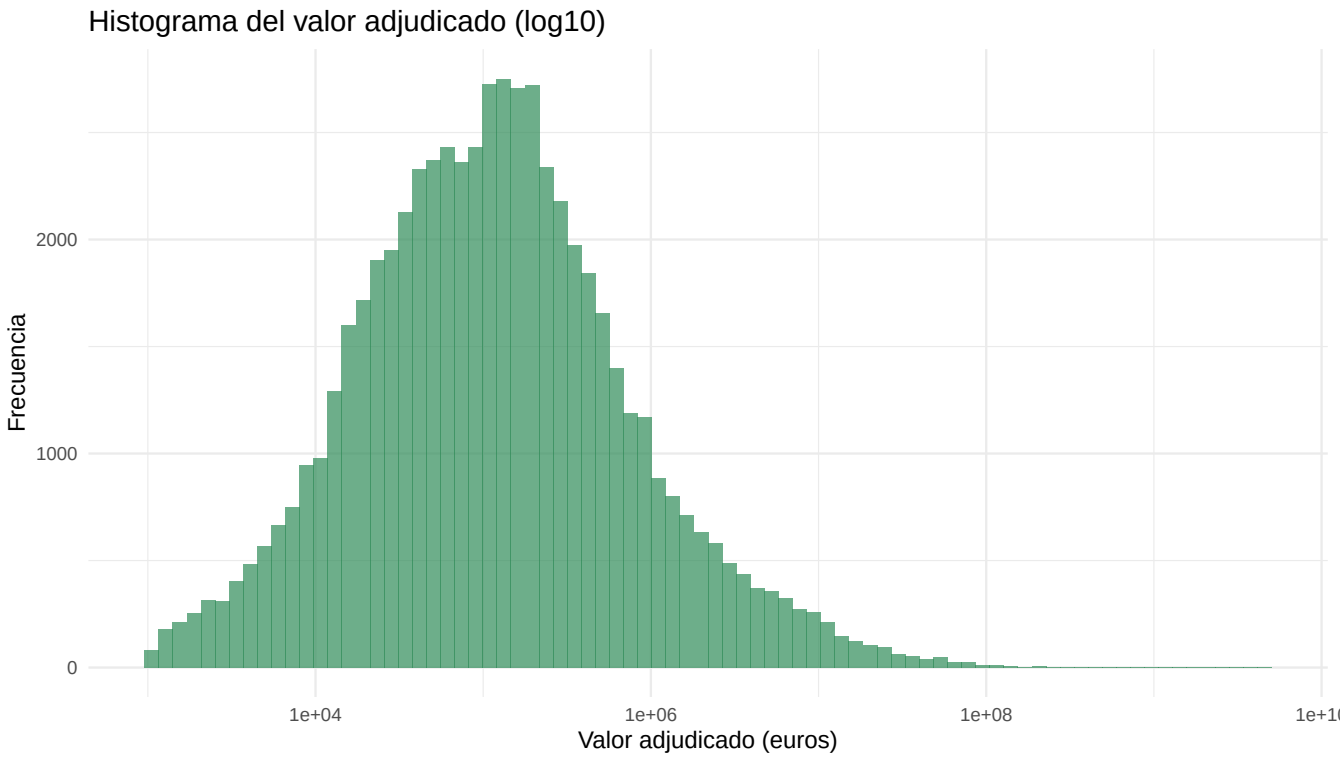
- **Std. Error.** Indica la variabilidad o incertidumbre de la estimación
- **t value.** Estadístico t para contrastar la hipótesis nula de que ese coeficiente es igual a cero
- **Pr > |t|.** Probabilidad asociada al test estadístico; valores pequeños (<0.05) denotan efecto estadísticamente significativo
- Se observa que los coeficientes de mayor magnitud corresponden, necesariamente, a **determinados ámbitos geográficos** y categorías administrativas concretas como, *Illes Balears, Ceuta, Melilla, Canarias*, indicando que el territorio de adjudicación influye fuertemente en la variabilidad del importe
- El coeficiente de **Naturaleza. Gestión de Servicios Públicos** aparece también entre los más destacados, denotando que la tipología funcional del contrato tiene impacto relevante sobre el volumen económico gestionado
- Otros coeficientes relevantes corresponden a instituciones concretas como, el **MINISTERIO DE POLÍTICA TERRITORIAL Y FUNCIÓN PÚBLICA** y procedimientos como el **diálogo competitivo**, aunque su significación estadística debe valorarse según el $Pr > |t|$
- El **intercepto** representa el valor base del modelo, sobre el que se suman los efectos de las distintas categorías
- Se denota que la **geografía** y la **naturaleza administrativa** son los ejes explicativos principales del modelo, mientras que el impacto de otras variables es menor o menos consistente a nivel estadístico
- Las categorías con $Pr > |t|$ bajo son las que verdaderamente aportan explicación robusta, mientras que otras pueden aparecer por el proceso de ajuste pero carecer de significado real
- La dispersión y el rango de los coeficientes reflejan la diversidad inherente al sector público y la existencia de conglomerados administrativos con capacidad para gestionar grandes volúmenes contractuales

4.1.5 Uso de escala logarítmica en la visualización

En general, se escoge la **escala logarítmica** porque permite **visualizar patrones, concentraciones y dispersiones** que serían invisibles en escala lineal. La representación logarítmica es imprescindible en todos aquellos contextos donde los datos presentan rangos de varios órdenes de magnitud, especialmente en economía y administración pública[3,5,6]

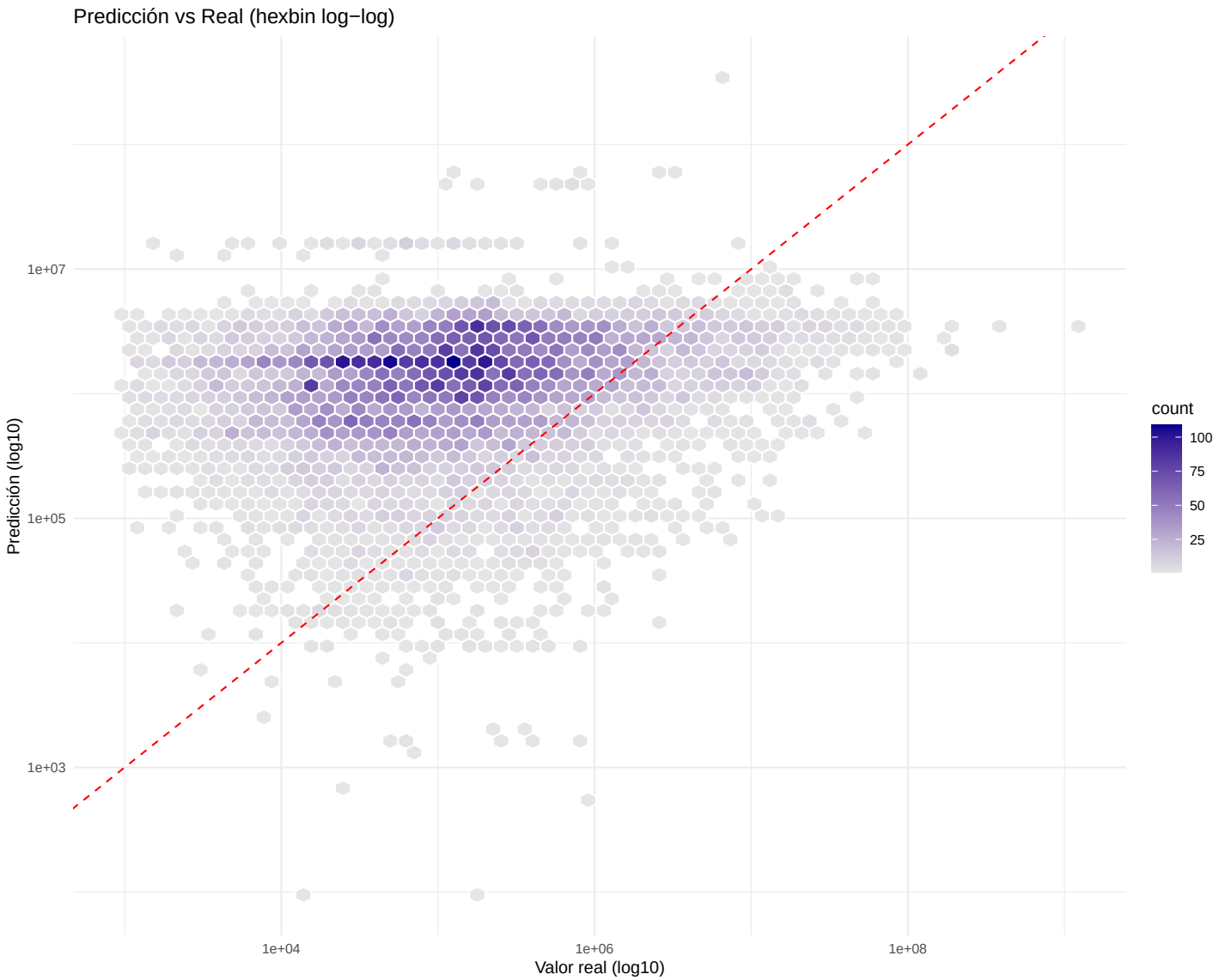
1. **Amplio rango de valores.** Los importes adjudicados en contratos públicos pueden variar desde microcontratos de apenas unos miles de euros hasta macroproyectos que superan los mil millones de euros. Representar estos datos en escala lineal haría que la mayor parte de los contratos (los más pequeños y medianos) quedasen "aplastados" en la parte inferior del gráfico, impidiendo apreciar cualquier patrón o diferencia relevante.
2. **Distribución asimétrica y sesgada.** El dataset presenta una **asimetría positiva** (cola larga a la derecha), típica de variables económicas y financieras, en la que la mayoría de los contratos se concentran en importes medios, pero existen unos pocos casos excepcionales de gran magnitud. La escala logarítmica corrige visualmente esta asimetría, facilitando la interpretación de tendencias y agrupaciones reales[2,4,6]
3. **Comparabilidad y robustez.** Al usar escala logarítmica, se denota mejor la dispersión relativa entre contratos y se evitan los efectos distorsionadores de los valores atípicos. Permite comparar de manera más justa la variabilidad dentro de cada grupo y entre categorías, tanto en histogramas como en boxplots, violin plots, gráficos de dispersión y análisis de errores
4. **Facilita la interpretación de los modelos.** Tanto en la regresión como en el análisis no supervisado, la visualización logarítmica ayuda a detectar si el modelo funciona igual de bien para los distintos órdenes de magnitud del dato, o si hay sesgos en los extremos (micro vs. macrocontratos)

Histograma del valor adjudicado (log10)



Se visualiza la distribución logarítmica de los importes adjudicados. La mayor parte de los contratos se sitúa entre cien mil euros (10^5€) y un millon de euros (10^6€). Las colas a ambos extremos representan la presencia tanto de microcontratos como de macrocontratos, habituales en la contratación pública. El uso de la escala logarítmica es necesaria para visualizar adecuadamente este rango tan amplio de importes y para evitar que los valores extremos eclipsen la estructura central de la distribución[2,4,6]

Dispersión Hexbin predicción vs real - Escala Log-Log



El gráfico de dispersión hexbin en escala log-log representa la relación entre los valores adjudicados reales y las predicciones del modelo de regresión lineal múltiple, permitiendo visualizar la densidad de registros en cada rango de importe

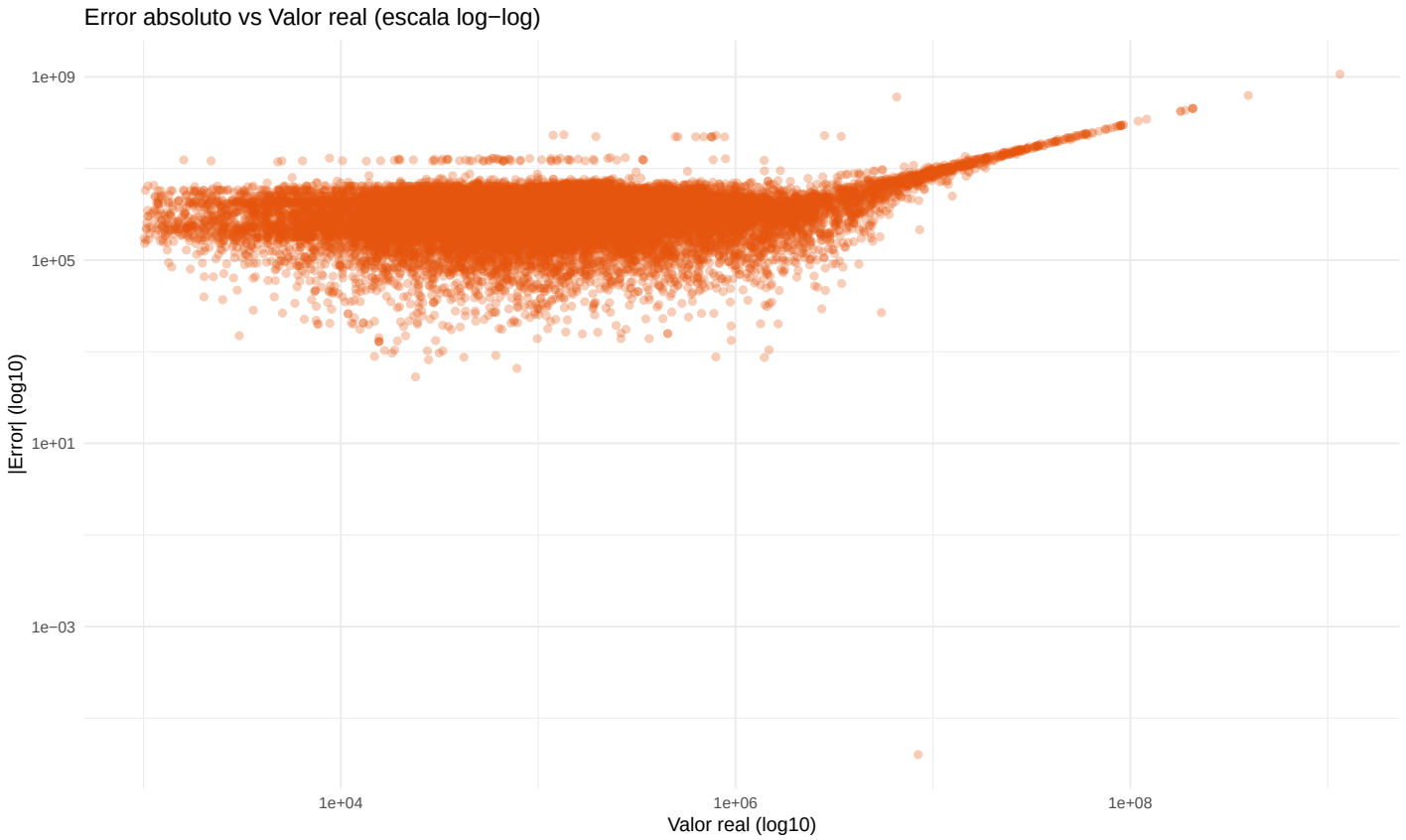
Se observa una **concentración de hexágonos azules en torno a la diagonal**, visualizando que el modelo logra predecir con mayor precisión la mayoría de los contratos estándar, con importes intermedios. Esta concentración central indica que, para estos contratos, el error absoluto es bajo en proporción al rango total de importes. La barra lateral "count" indica el número de contratos que caen dentro de cada rango de importe previsto y real, evidenciando que la mayor parte de la contratación pública se produce en importes de rango medio

El gráfico **no es perfectamente simétrico respecto a la diagonal**. En la parte superior izquierda, donde el valor predicho supera al real, se detecta una mayor dispersión y presencia de hexágonos, denotando que el modelo tiende a **sobreestimar** el valor de algunos contratos de importe especialmente bajo. Este fenómeno puede deberse a la menor frecuencia y mayor variabilidad de los **macrocontratos** y microcontratos, dificultando el ajuste del modelo en estos extremos de la distribución

Por otro lado, pueden existir **efectos no capturados** por las variables del modelo, tales como características específicas de contratos singulares, adjudicaciones con múltiples contratistas, por lotes o Uniones Temporales de Empresas, o la ausencia de información relevante en el dataset, como pueda ser la duración del contrato u otros criterios técnicos, etc… Estas limitaciones explican que la dispersión sea mayor en los extremos y que el modelo presente errores más elevados tanto en la predicción de microcontratos como de macrocontratos

La **asimetría respecto a la diagonal** refleja tanto la heterogeneidad de los datos de contratación pública como la dificultad inherente para ajustar modelos predictivos en presencia de importes muy dispares y contratos singulares[2,4,6]

Histograma de residuos - Error absoluto vs Valor Real - Log-Log

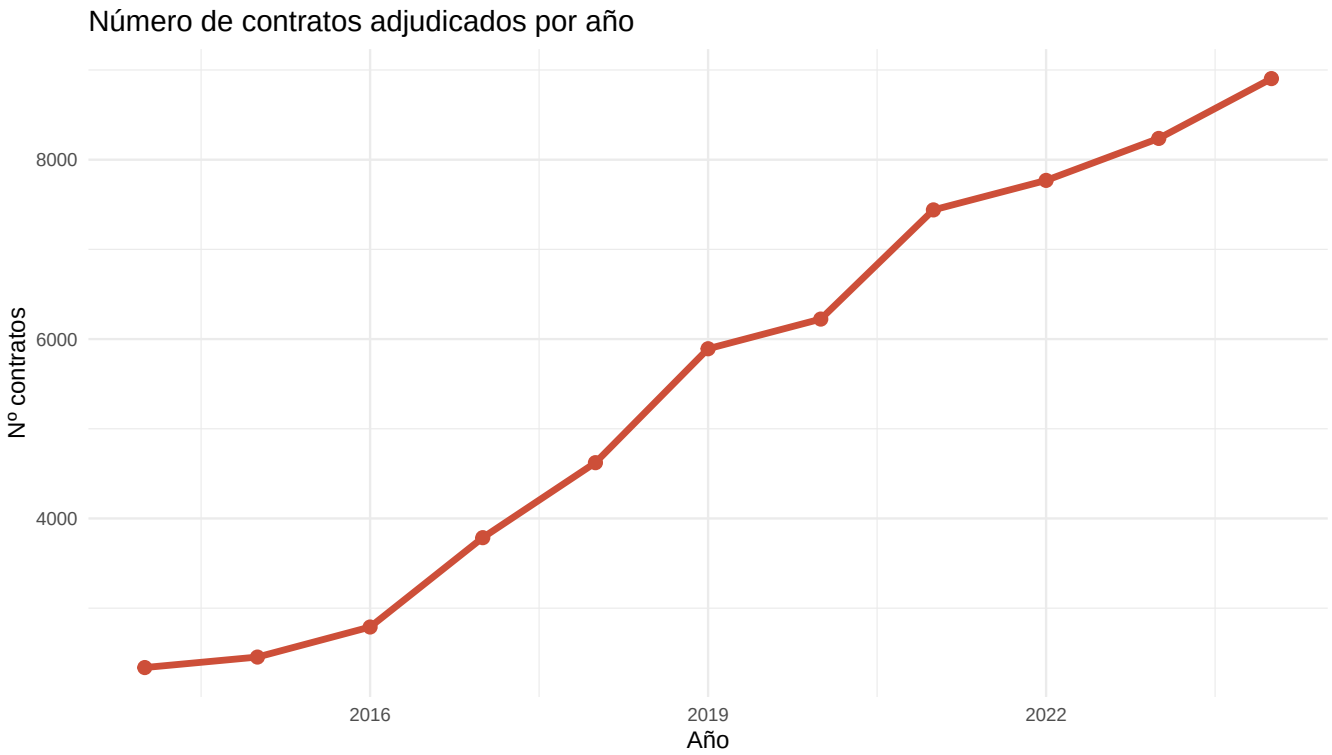


En el histograma de residuos se representa la distribución de los errores absolutos de predicción frente al valor adjudicado real, ambos en escala logarítmica (log-log). A diferencia de una gráfica lineal no logarítmica, en este contexto los ejes nunca toman valor cero, ya que el logaritmo de cero no está definido. Se observa que la mayor parte de los puntos se sitúa en la zona inferior y central del gráfico, mostrando que el error absoluto es bajo en relación al rango total de importes manejados

Esta concentración indica que, para la mayoría de los contratos estándar, aquellos de importes intermedios y que representan el grueso de la contratación pública, el modelo logra una predicción razonablemente precisa. Por el contrario, en la parte superior y derecha del gráfico se detectan errores más elevados, especialmente asociados a los contratos de importe muy alto - macrocontratos. Esta dispersión creciente en los extremos es habitual en datasets económicos y administrativos, donde los valores atípicos y la menor frecuencia de casos limitan la capacidad de ajuste del modelo lineal[2,4,6]

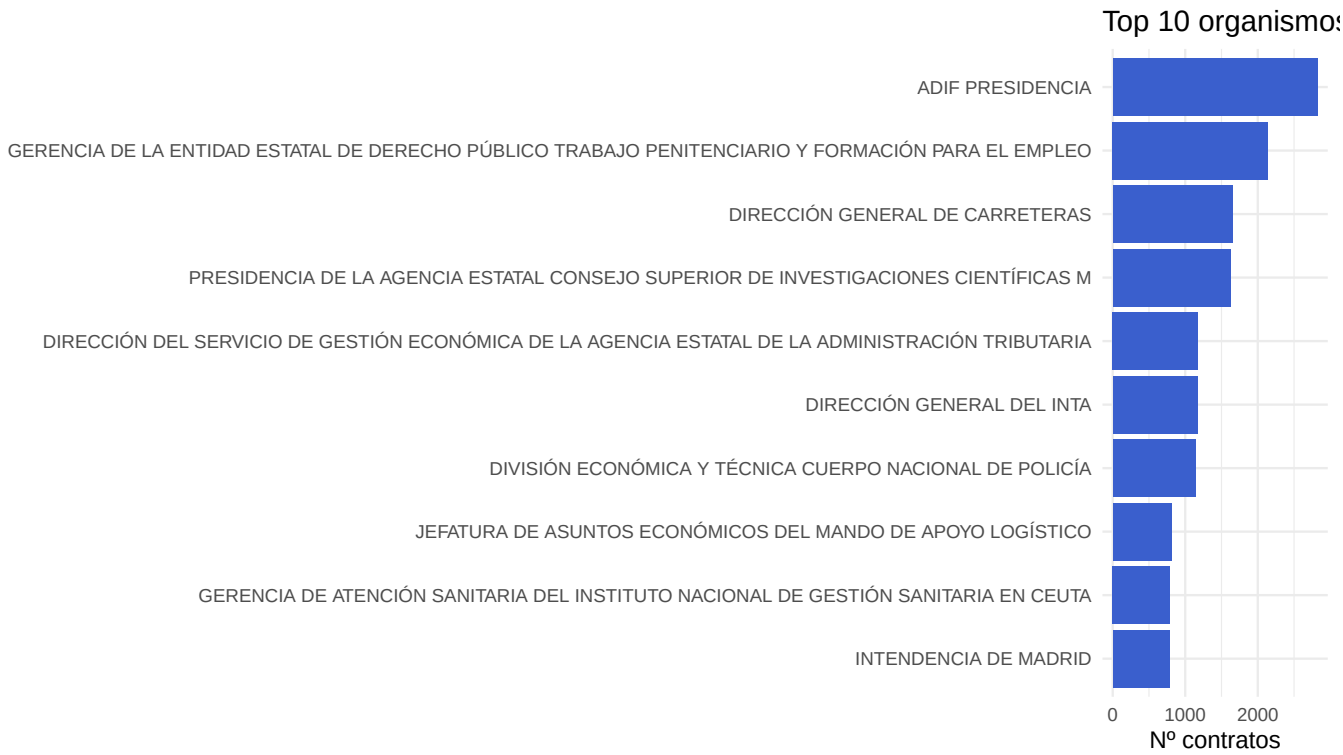
Debe señalarse que cuando se indica que “la mayoría de los residuos se concentra cerca de cero”, esto se interpreta, en escala logarítmica, como “errores pequeños en proporción al rango global de errores posibles”, aunque nunca exactamente cero en valor absoluto. La escala log-log refuerza la legibilidad y permite denotar patrones de precisión y dispersión en todo el rango de importes, evitando la pérdida de matices tanto en los contratos pequeños como en los grandes

Serie temporal del número de contratos por año



Se observa la evolución temporal de la contratación pública en el periodo 2014–2024. Se detectan fluctuaciones probablemente a ciclos presupuestarios, cambios normativos o situaciones excepcionales, como la pandemia de 2020

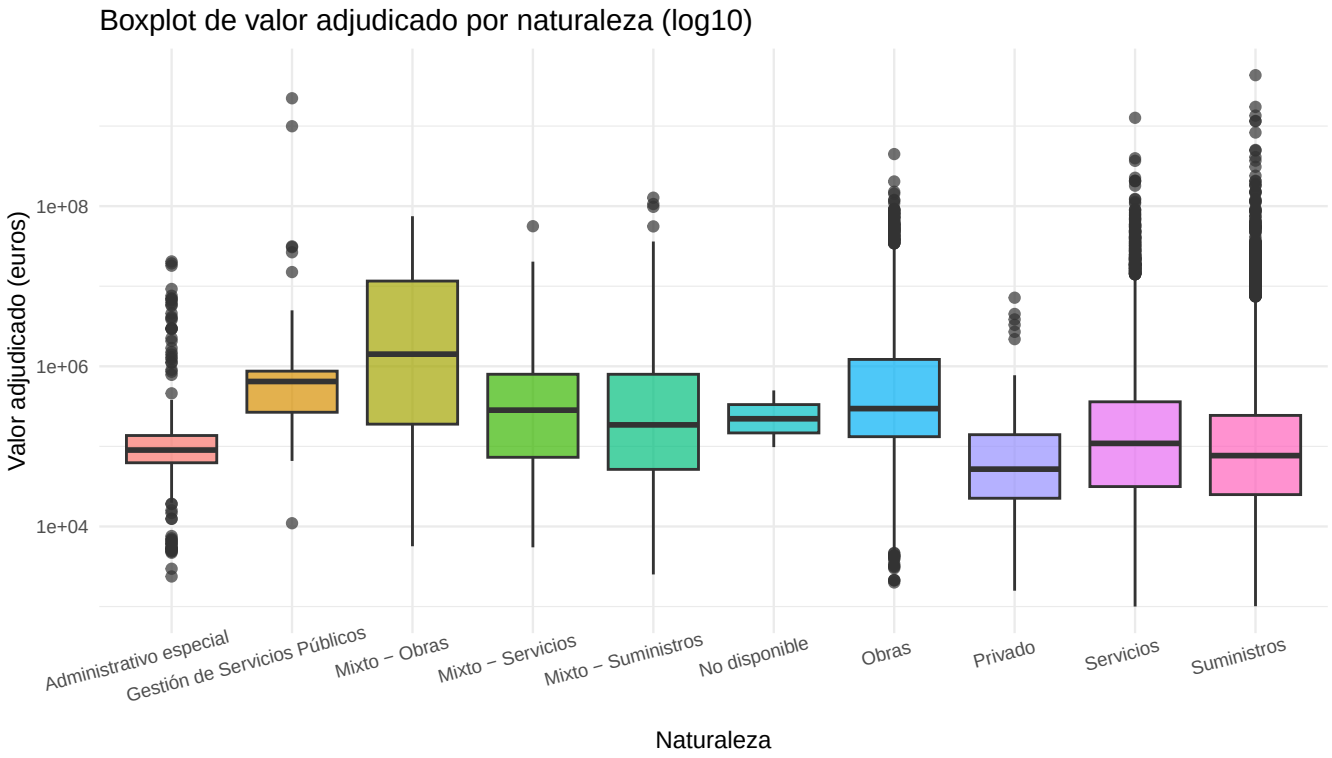
Top 10 organismos responsables



Se representa el ranking de los diez organismos públicos con mayor número de contratos adjudicados en el periodo analizado. Se observa que la **ADIF PRESIDENCIA** y la **Gerencia de la Entidad Estatal de Derecho Público Trabajo Penitenciario y Formación para el Empleo** encabezan el listado, mostrando una elevada actividad contractual asociada tanto a infraestructuras ferroviarias como a servicios penitenciarios y sociales

Aparecen también en posiciones destacadas la Dirección General de Carreteras, organismos de investigación y agencias tributarias, confirmando la diversidad sectorial del gasto público

Boxplot de valor adjudicado por naturaleza

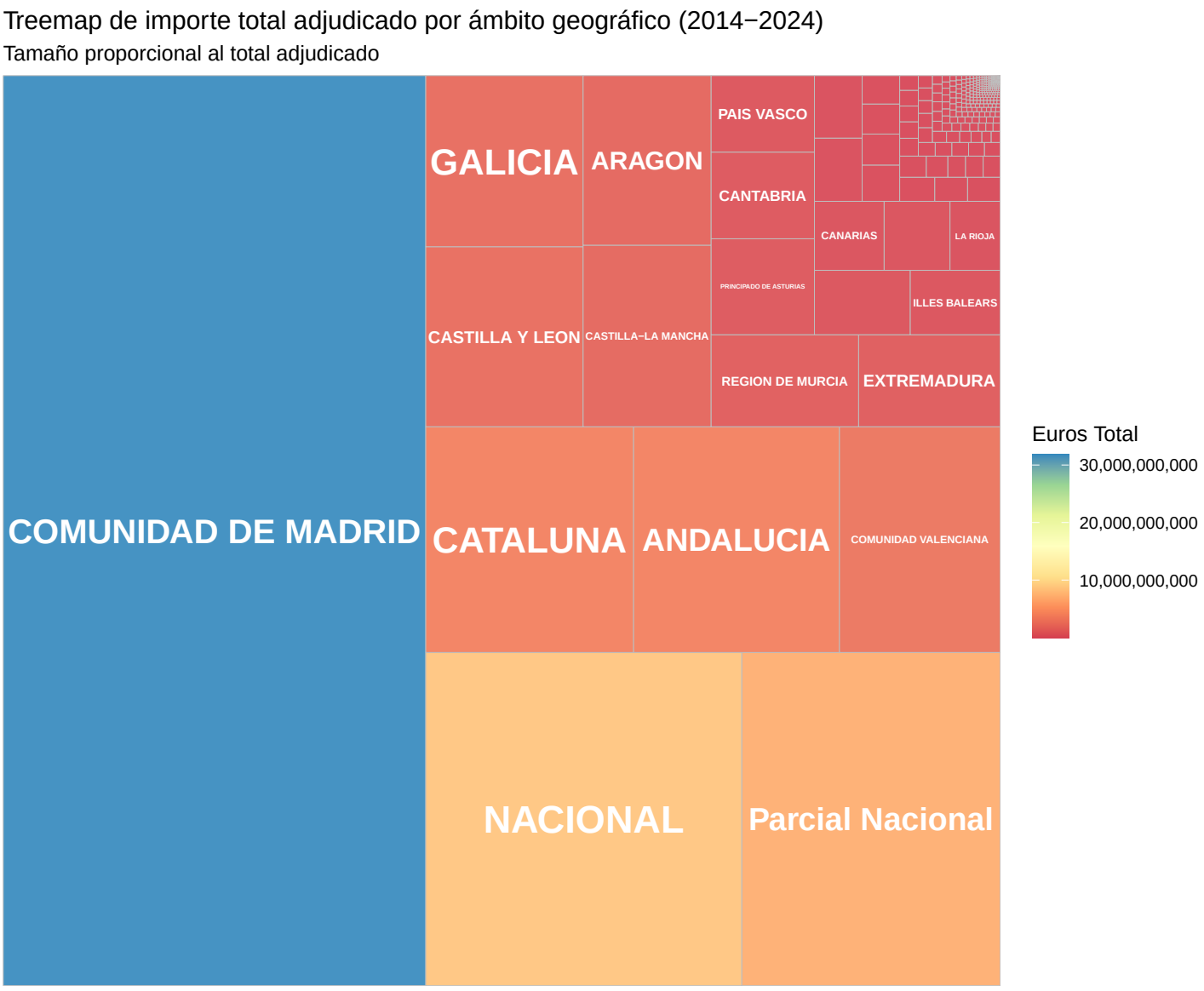


En este boxplot se visualiza la **distribución del importe adjudicado** según la naturaleza del contrato, utilizando escala logarítmica para reflejar la amplitud de importes

Se aprecia cómo los contratos de Obras presentan una mayor dispersión y valores máximos más elevados, denotando la existencia de grandes proyectos de inversión pública en infraestructuras. Los Servicios y Suministros muestran una distribución más concentrada, aunque también con presencia de valores atípicos de considerable importe

```
library(treemapify)
# Prepara los datos, suma total adjudicado por ámbito geográfico
df_treemap <- df_contratacion_final %>%
  mutate(Ambito_geografico = toupper(Ambito_geografico),
         Ambito_geografico = stringi::stri_trans_general(Ambito_geografico, "Latin-ASCII")) %>%
  group_by(Ambito_geografico) %>%
  summarise(Importe_total = sum(valor_oferta_adjudicada, na.rm = TRUE)) %>%
  filter(Importe_total > 0) %>%
  arrange(desc(Importe_total))

# Para evitar solapamientos, recodifica los valores residuales si quieres
df_treemap$Ambito_geografico[df_treemap$Ambito_geografico %in% c("SIN DEFINIR")] <- "Parcial Nacional"

ggplot(df_treemap, aes(
  area = Importe_total,
  fill = Importe_total,
  label = Ambito_geografico
)) +
  geom_treemap(alpha = 0.9) +
  geom_treemap_text(
    fontface = "bold",
    colour = "white",
    place = "centre",
    min.size = 2,
    grow = FALSE
  ) +
  # Puedes elegir entre viridis, distiller, etc.
```

```
  scale_fill_distiller(
    palette = "Spectral", direction = 1,
    trans = "identity", label = scales::comma
  ) +
  labs(
    title = "Treemap de importe total adjudicado por ámbito geográfico (2014-2024)",
    subtitle = "Tamaño proporcional al total adjudicado",
    fill = "Euros Total"
  ) +
  theme_minimal(base_size = 10) +
  theme(legend.position = "right")
```



Se representa la distribución geográfica del importe total adjudicado en contratos públicos entre 2014 y 2024, agregando los datos por comunidad autónoma. Se normalizan los nombres de territorio para asegurar una integración correcta con el shapefile oficial. La escala logarítmica en el color permite visualizar tanto grandes diferencias entre territorios como la actividad de las comunidades con menor volumen de contratación

Este mapa facilita detectar rápidamente los centros de mayor actividad económica pública y permite comparar el peso relativo de cada territorio, eliminando distorsiones debidas a diferencias de escala bruta. Se aprecia cómo regiones con grandes capitales o núcleos administrativos concentran el mayor importe adjudicado, mientras que otras comunidades presentan una menor densidad contractual

La visualización denota posibles desequilibrios o patrones de concentración en el reparto de fondos público[2,4,6]

```r
# top 10 adjudicatarios por importe total adjudicado
df_treemap_adj <- df_contratacion_final %>%
  mutate(nombre_adjudicatario = toupper(nombre_adjudicatario),
         nombre_adjudicatario = stringi::stri_trans_general(nombre_adjudicatario, "Latin-ASCII")) %>%
  group_by(nombre_adjudicatario) %>%
  summarise(Importe_total = sum(valor_oferta_adjudicada, na.rm = TRUE)) %>%
  filter(Importe_total > 0) %>%
  arrange(desc(Importe_total)) %>%
  slice_head(n = 10) %>%
  ungroup()

# Mejorar legibilidad de nombres largos
df_treemap_adj <- df_treemap_adj %>%
  mutate(etiqueta = str_wrap(nombre_adjudicatario, width = 16))

ggplot(df_treemap_adj, aes(
  area = Importe_total,
  fill = Importe_total,
  label = etiqueta
)) +
  geom_treemap(alpha = 0.9) +
  geom_treemap_text(
    fontface = "bold",
    colour = "white",
    place = "centre",
    min.size = 2.5,
    grow = FALSE
  ) +
  scale_fill_distiller(
    palette = "Spectral", direction = 1,
    trans = "identity", label = scales::comma
  ) +
  labs(
    title = "Top 10 adjudicatarios por importe total adjudicado (2014-2024)",
    subtitle = "Tamaño proporcional al total adjudicado. Nombres adaptados a varias líneas.",
    fill = "Euros Total"
  ) +
  theme_minimal(base_size = 10) +
  theme(legend.position = "right")
```

Top 10 adjudicatarios por importe total adjudicado (2014−2024)

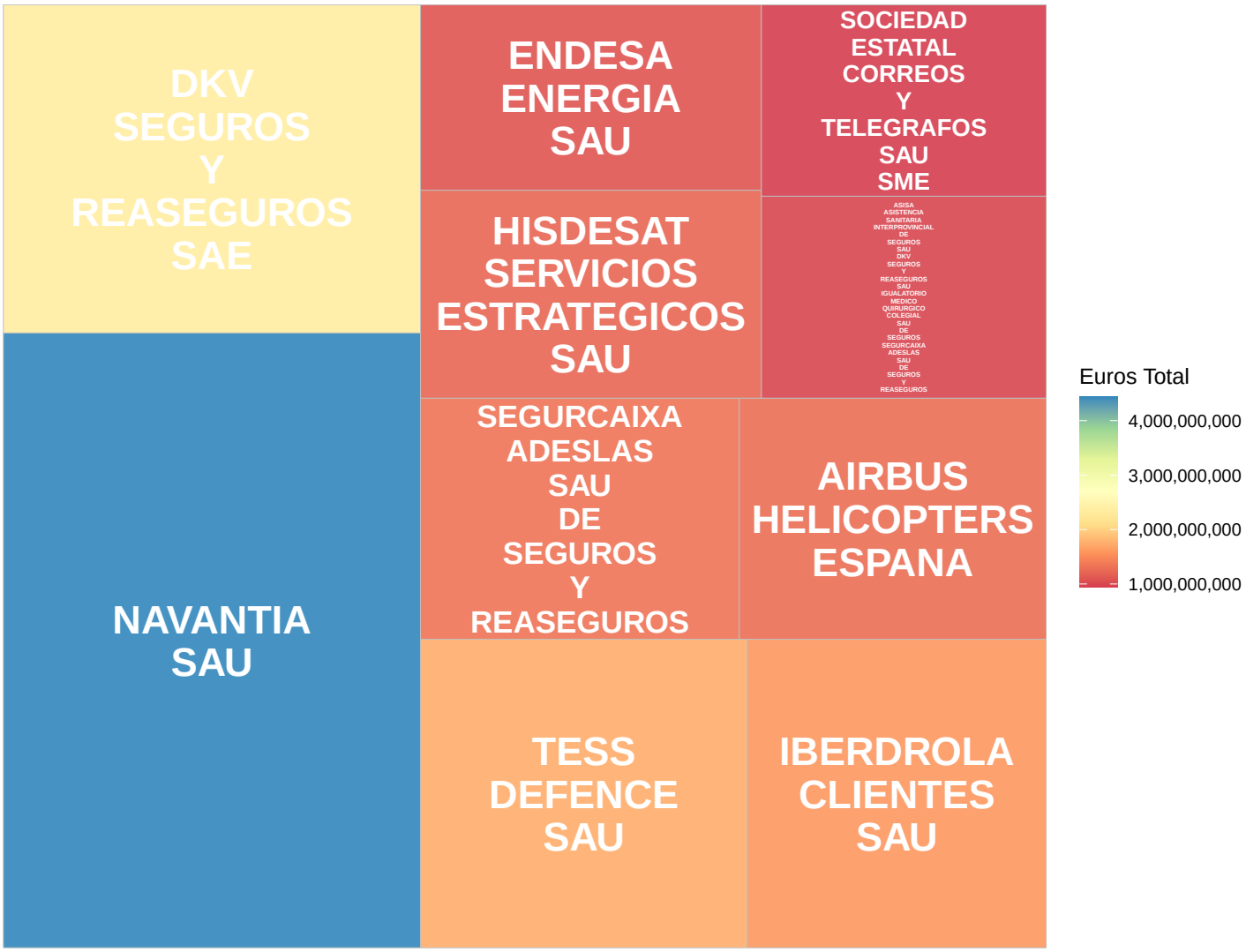
Tamaño proporcional al total adjudicado. Nombres adaptados a varias líneas.


```r
# Ajustar nombre a mayúsculas y ASCII para evitar duplicados por acentos
df_treemap <- df_contratacion_final %>%
  mutate(
    nombre_adjudicatario = toupper(nombre_adjudicatario),
    nombre_adjudicatario = stringi::stri_trans_general(nombre_adjudicatario, "Latin-ASCII")
  ) %>%
  group_by(nombre_adjudicatario) %>%
  summarise(Importe_total = sum(valor_oferta_adjudicada, na.rm = TRUE)) %>%
  filter(Importe_total > 0) %>%
  arrange(desc(Importe_total))

# Excluir explícitamente al TACRC
df_treemap <- df_treemap %>%
  filter(!grepl("TRIBUNAL ADMINISTRATIVO CENTRAL DE RECURSOS CONTRACTUALES", nombre_adjudicatario))

# Seleccionar el top 10 adjudicatarios reales
df_top10 <- df_treemap %>% head(10)

# dividir los nombres largos en varias líneas
df_top10$nombre_adjudicatario <- gsub(" ", "\n", df_top10$nombre_adjudicatario, fixed = TRUE)

# Graficar treemap
```

```r
ggplot(df_top10, aes(
  area = Importe_total,
  fill = Importe_total,
  label = nombre_adjudicatario
)) +
  geom_treemap(alpha = 0.9) +
  geom_treemap_text(
    fontface = "bold",
    colour = "white",
    place = "centre",
    min.size = 2,
    grow = FALSE
  ) +
  scale_fill_distiller(
    palette = "Spectral", direction = 1,
    trans = "identity", label = scales::comma
  ) +
  labs(
    title = "Top 10 adjudicatarios reales por importe total adjudicado (2014-2024)",
    subtitle = "Tamaño proporcional al total adjudicado. TACRC excluido.",
    fill = "Euros Total"
  ) +
  theme_minimal(base_size = 10) +
  theme(legend.position = "right")
```

Top 10 adjudicatarios reales por importe total adjudicado (2014−2024)

Tamaño proporcional al total adjudicado. TACRC excluido.



Treemap de adjudicatarios principales

El treemap mostrado representa el **top 10 de adjudicatarios por importe total adjudicado** en el periodo 2014–2024, visualizando de forma proporcional la concentración del gasto público en grandes contratos. Al analizar la distribución, se detecta que la mayor parte de la contratación pública se reparte entre grandes corporaciones del sector sanitario (DKV, SegurCaixa Adeslas, Asisa, etc.), recursos estratégicos (Navantia, Hisdesat, Airbus Helicopters, Endesa, Iberdrola) y defensa (Tess Defence)

Debe señalarse la **presencia anómala del Tribunal Administrativo Central de Recursos Contractuales (TACRC)**, un organismo que no es en realidad un adjudicatario, sino un órgano encargado de la resolución de recursos y conflictos en contratación pública. Su aparición en el ranking se debe a errores en la extracción o clasificación de datos desde el BOE, que el scraper interpreta erróneamente como contratos asignados al tribunal. Esta circunstancia es habitual en datasets administrativos y denota la necesidad de revisión manual y refinamiento de las rutinas de procesamiento para evitar este tipo de interpretaciones erróneas[2,4,6]

Una vez eliminada la **presencia anómala del Tribunal Administrativo Central de Recursos Contractuales (TACRC)**, se observa que el grueso de la contratación pública en España entre 2014 y 2024 se concentra en grandes empresas del sector sanitario, industrial, energético y de defensa

- **DKV SEGUROS Y REASEGUROS SAE, SEGURCAIXA ADESLAS SAU DE SEGUROS Y REASEGUROS y otros operadores de seguros** copan una parte muy significativa del total, denotando la importancia del gasto sanitario y de pólizas para la administración pública.
- **NAVANTIA SAU, TESS DEFENCE SAU y AIRBUS HELICOPTERS ESPAÑA** aparecen como adjudicatarios principales, reflejando la elevada inversión en sectores estratégicos como defensa, construcción naval y

aeroespacial

- El sector energético también ocupa un papel relevante, representado por **ENDESA ENERGÍA SAU e IBERDROLA CLIENTES SAU**
- **HISDESAT SERVICIOS ESTRATÉGICOS SAU** destaca por la adjudicación de contratos estratégicos vinculados a servicios de comunicaciones seguras para la administración y defensa
- La **SOCIEDAD ESTATAL CORREOS Y TELÉGRAFOS SAU SME** muestra el peso que tienen los servicios logísticos y de correo público en el gasto estatal

Esta visualización denota una fuerte concentración del importe adjudicado en pocos adjudicatarios, la mayoría grandes empresas multinacionales o participadas estatalmente, que cubren áreas consideradas críticas o esenciales, como salud, energía, defensa, logística y servicios estratégicos

4.2 Análisis no supervisado. Segmentación de adjudicatarios por volumen y cuantía

En este apartado se aplica un método no supervisado de *clustering k-means* sobre el conjunto de adjudicatarios, utilizando dos variables numéricas relevantes, el **número de contratos adjudicados** y el **importe total adjudicado** en el periodo 2014–2024. Este análisis permite denotar patrones de concentración, identificar operadores dominantes y caracterizar la estructura competitiva del mercado de contratación pública[3,4,6]

4.2.1 Preparación de los datos

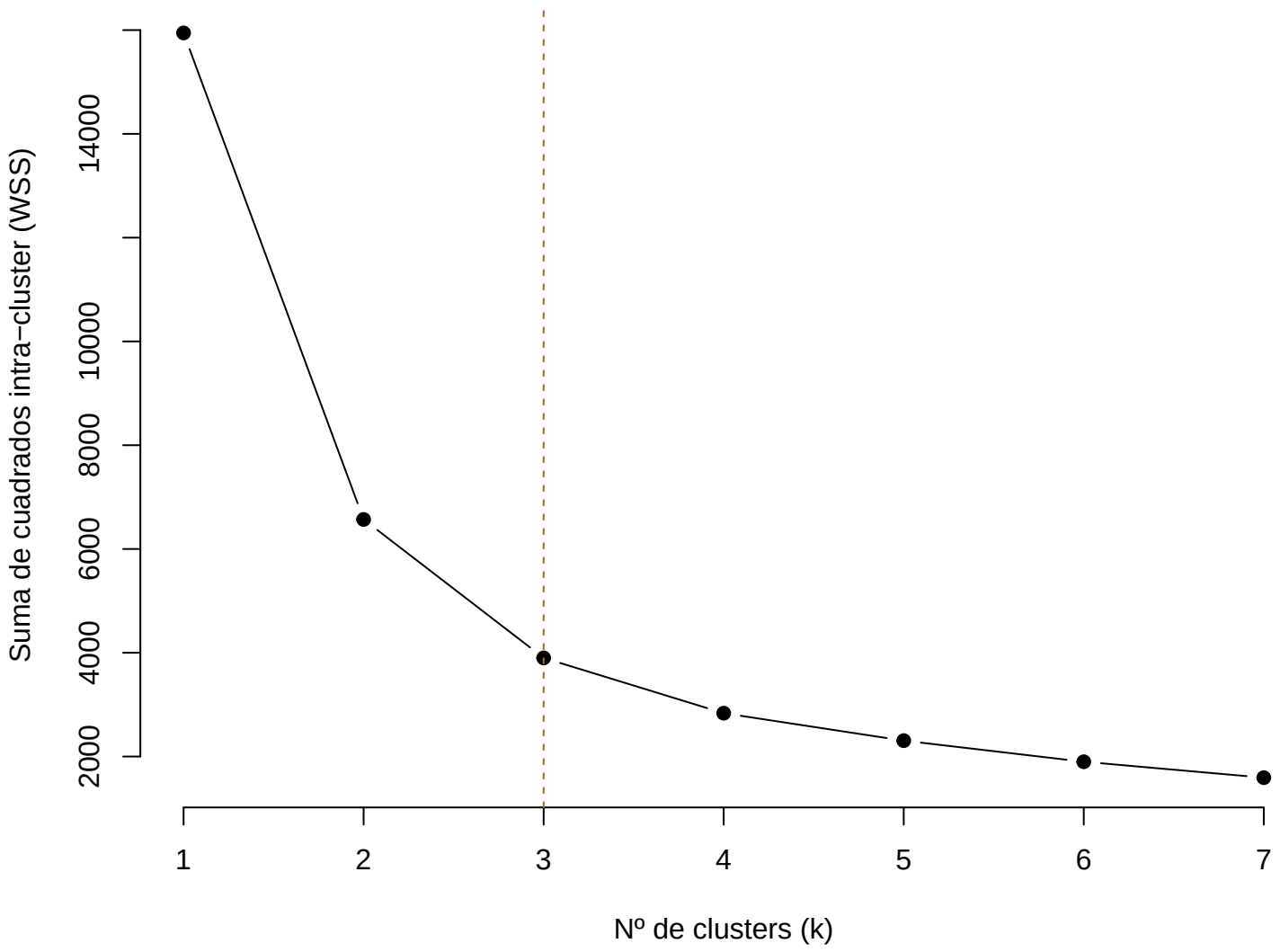
Se agregan los datos por adjudicatario, computando tanto el número de contratos como la suma total adjudicada. Se aplica una transformación logarítmica para mejorar la visualización y homogeneizar las escalas

```
library(dplyr)

df_adj <- df_contratacion_final %>%
  filter(!is.na(nombre_adjudicatario), valor_oferta_adjudicada > 0) %>%
  group_by(nombre_adjudicatario) %>%
  summarise(
    n_contratos = n(),
    importe_total = sum(valor_oferta_adjudicada, na.rm = TRUE)
  ) %>%
  ungroup() %>%
  mutate(
    log_n = log10(n_contratos + 1),
    log_importe = log10(importe_total + 1)
  )
```

4.2.2 Selección del número óptimo de clústeres

Se utiliza el método del codo para seleccionar el valor óptimo de k

```
set.seed(42)
wss_adj <- sapply(1:7, function(k) {
  kmeans(df_adj[, c("log_n", "log_importe")], centers = k, nstart = 20)$tot.withinss
})
plot(1:7, wss_adj, type = "b", pch = 19, frame = FALSE,
     xlab = "Nº de clusters (k)", ylab = "Suma de cuadrados intra-cluster (WSS)",
     main = "Método del codo para adjudicatarios")
abline(v = 3, lty = 2, col = "tomato3")
```



4.2.3 Ajuste del modelo y asignación de clústeres

Se aplica k-means con $k = 3$, agrupando adjudicatarios según su volumen y cuantía total

```
k_adj <- 3
modelo_kmeans_adj <- kmeans(df_adj[, c("log_n", "log_importe")], centers = k_adj, nstart = 25)
df_adj$cluster <- factor(modelo_kmeans_adj$cluster)
```

Tabla por clúster

Se presenta una tabla que resume el número de adjudicatarios, la media de contratos y el importe medio por clúster

```
library(knitr)
library(kableExtra)

tabla_clusters_adj <- df_adj %>%
  group_by(cluster) %>%
  summarise(
    N_adjudicatarios = n(),
    Media_n_contratos = mean(n_contratos),
    Mediana_n_contratos = median(n_contratos),
    Media_importe = mean(importe_total),
    Mediana_importe = median(importe_total)
  )
```

```r
kable(tabla_clusters_adj, caption = "Resumen descriptivo de cada clúster de adjudicatarios") %>%
  kable_styling(font_size = 9, position = "center")
```

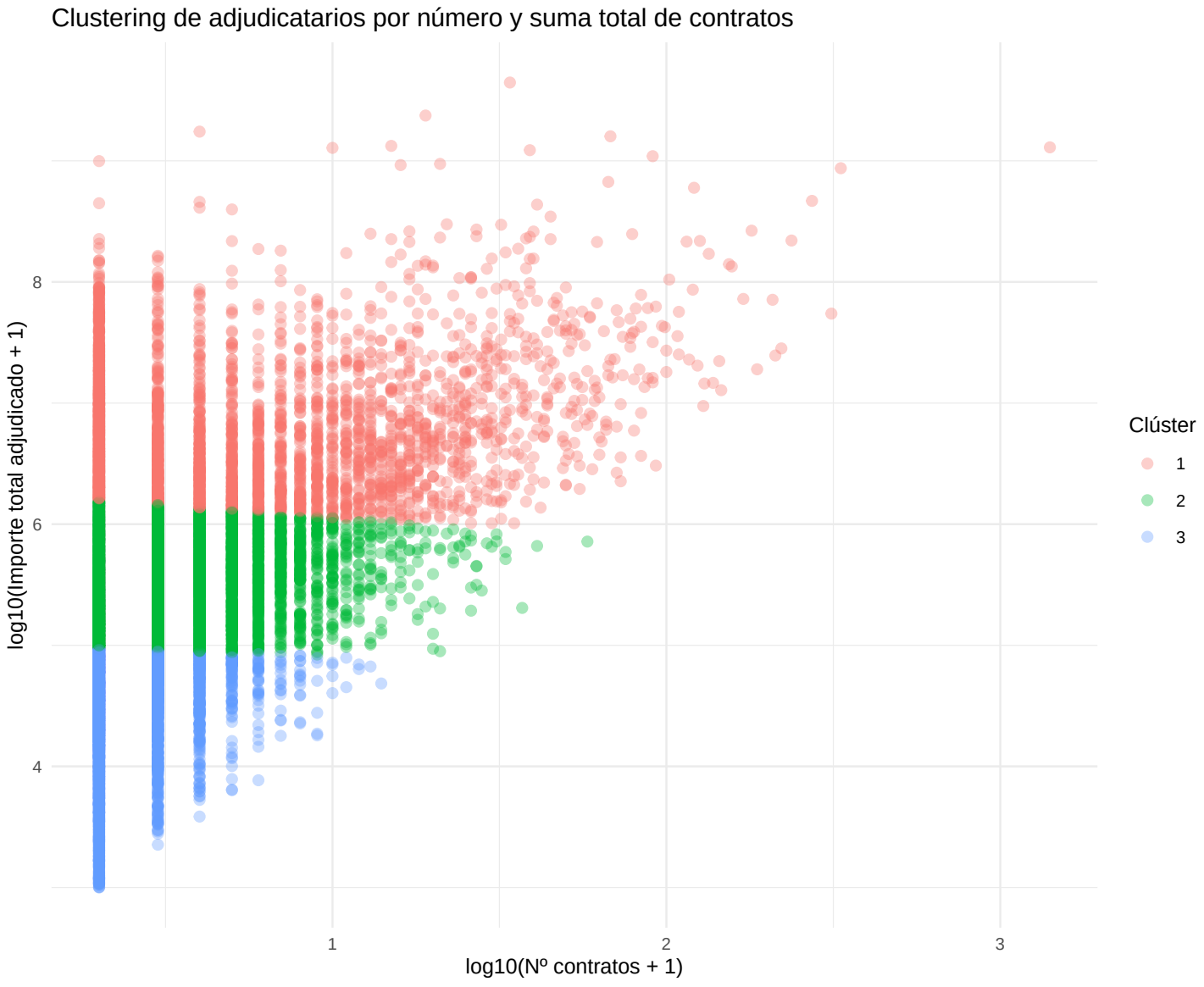
Resumen descriptivo de cada clúster de adjudicatarios

cluster	N_adjudicatarios	Media_n_contratos	Mediana_n_contratos	Media_importe	Mediana_importe
1	3710	9.544744	3	19309559.8	4428633.9
2	7770	2.337580	1	433724.4	314048.7
3	5022	1.370370	1	37022.0	32258.2

4.2.4 Clustering de adjudicatarios

Se visualiza el resultado del clustering con los ejes en escala logarítmica, distinguiendo los diferentes perfiles de adjudicatario

```r
library(ggplot2)
ggplot(df_adj, aes(x = log_n, y = log_importe, color = cluster)) +
  geom_point(alpha = 0.35, size = 2) +
  scale_x_continuous(name = "log10(Nº contratos + 1)") +
  scale_y_continuous(name = "log10(Importe total adjudicado + 1)", labels = scales::comma) +
  labs(title = "Clustering de adjudicatarios por número y suma total de contratos",
       color = "Clúster") +
  theme_minimal(base_size = 10)
```

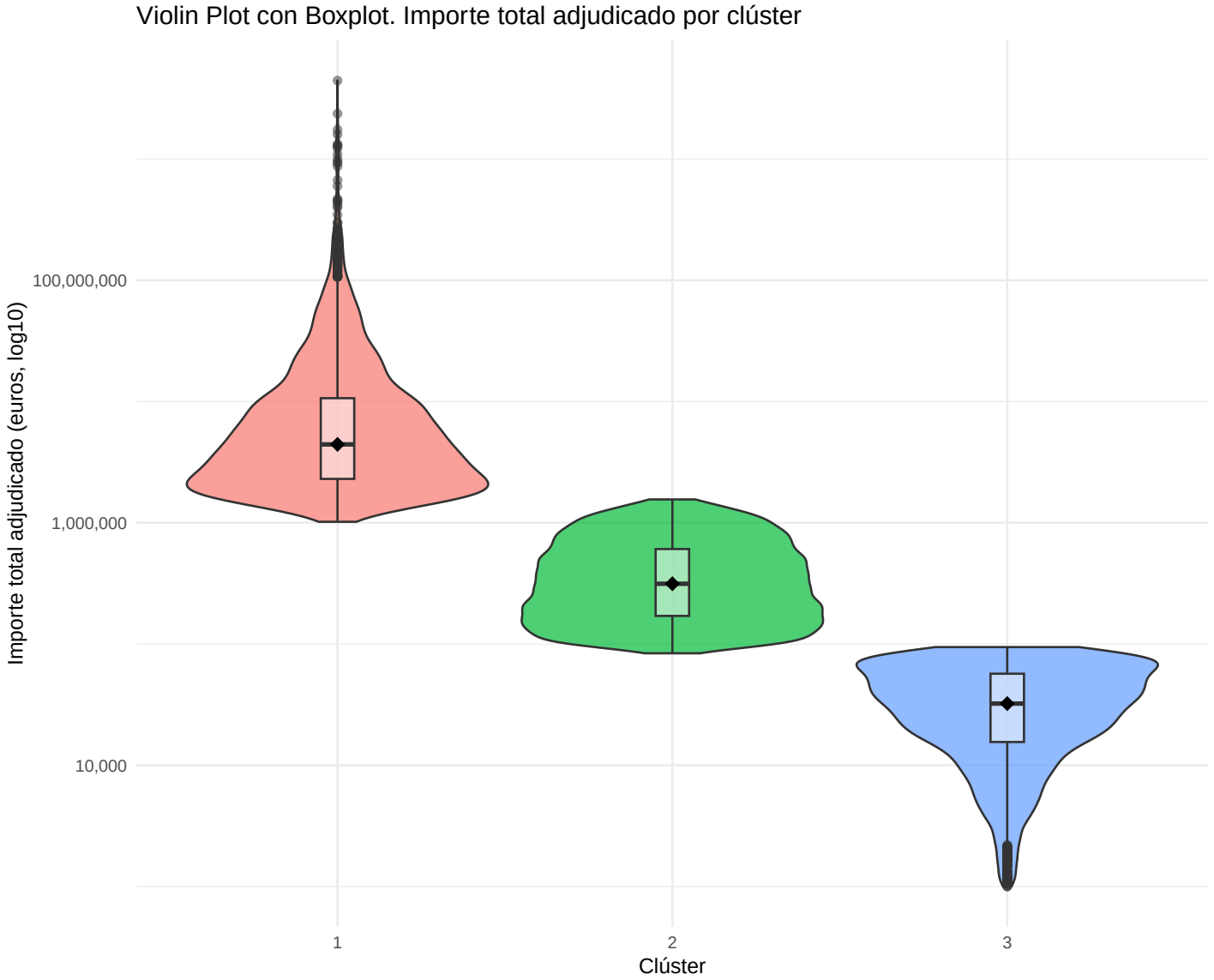


El scatterplot confirma la utilidad del clustering no supervisado en la detección de nichos, operadores dominantes y patrones de comportamiento en la contratación pública española. Su interpretación refuerza la evidencia obtenida en las visualizaciones previas y proporciona una base objetiva para la formulación de hipótesis o recomendaciones de política sectorial[3,4,6]

El gráfico representa el **clustering de adjudicatarios según el número de contratos adjudicados y la suma total de importes obtenidos**. Cada punto denota un adjudicatario, situándolo en el plano log-log donde el eje X corresponde al logaritmo del número de contratos (+1) y el eje Y al logaritmo del importe total adjudicado (+1). Los colores representan los clústeres identificados mediante k-means, distinguiendo claramente **tres segmentos principales** de actores en el mercado de la contratación pública

- **Estructura de los segmentos**
  - **Microoperadores** (azul). Adjudicatarios con muy pocos contratos y bajos importes totales, situados en la parte inferior izquierda del gráfico. Constituyen la base de la pirámide, mostrando participación puntual y escasa relevancia presupuestaria.
  - **Adjudicatarios estándar** (verde). Entidades que gestionan un número intermedio de contratos y acumulan una suma total relevante, localizándose en la parte central del gráfico. Representan el tejido habitual y recurrente de la contratación pública, proporcionando estabilidad al sistema.
  - **Macrooperadores** (rojo). Adjudicatarios que concentran numerosos contratos y elevados importes agregados, ubicados en la zona superior derecha. Su presencia refleja posiciones dominantes, ya sea por especialización, tamaño o ventajas competitivas estructurales[3,4,6]
- **Patrones visuales y segmentación.** El gráfico pone de manifiesto la marcada **asimetría** en la distribución de contratos y fondos, la gran mayoría de adjudicatarios pertenece a los segmentos bajos e intermedios, mientras que sólo unos pocos acceden a los volúmenes más altos de contratación. La segmentación generada por k-means permite distinguir no sólo el nivel absoluto de participación, sino también las estrategias de cada operador, unos optan por volumen (muchos contratos de importe medio-bajo), otros por pocos contratos pero de gran cuantía, y los más relevantes por ambas dimensiones simultáneamente
- **Aplicación práctica.** Este tipo de análisis, basado en agregados económicos y no en atributos individuales de los contratos, **revela la estructura interna y la desigualdad de la competencia** en el sector público. Es una herramienta imprescindible para auditores, responsables de transparencia y diseñadores de políticas de diversificación del gasto, permitiendo detectar posiciones de dominio y orientar intervenciones regulatorias o competitivas[3,4,6]

```
ggplot(df_adj, aes(x = cluster, y = importe_total, fill = cluster)) +
  geom_violin(alpha = 0.7, scale = "width") +
  geom_boxplot(width = 0.1, fill = "white", alpha = 0.5) +
  stat_summary(fun = "median", geom = "point", shape = 18, size = 3, color = "black") +
  scale_y_log10(labels = scales::comma) +
  labs(title = "Violin Plot con Boxplot. Importe total adjudicado por clúster",
       x = "Clúster", y = "Importe total adjudicado (euros, log10)") +
  theme_minimal(base_size = 10) +
  theme(legend.position = "none")
```



El gráfico denota cómo el modelo de clustering ha segmentado el universo contractual en tres perfiles diferenciados, microcontratos (muchos, de bajo valor), contratos estándar (moderados en número y cuantía), y macrocontratos (pocos, pero de altísimo importe), permitiendo abordar el análisis posterior con un enfoque adaptado a la realidad de los datos[2,3,4,6]

El gráfico mostrado combina un **violin plot** con un **boxplot** superpuesto, representando la distribución del **importe total adjudicado** (en escala logarítmica) para cada uno de los clústeres identificados mediante *k-means*

En primer lugar, la **elección del violin plot** permite visualizar simultáneamente la densidad y la dispersión de los valores adjudicados dentro de cada grupo, una opción imprescindible cuando se trabaja con datos fuertemente asimétricos y de rango amplio, como es habitual en contratación pública[2,4,6]

Forma de los violines

- El **clúster 1** (izquierda, color salmón) muestra una clara **asimetría positiva**, con una base ancha a valores intermedios y una "cola" que se extiende hacia importes extremadamente altos. Esto denota que este grupo concentra la mayor parte de los **macrocontratos**, es decir, adjudicaciones de cuantía excepcional, aunque la mayoría de sus miembros presenta importes más moderados
- El **clúster 2** (centro, verde) aparece con un cuerpo central más compacto y menor dispersión vertical. Aquí predominan los **contratos estándar**. La mayoría de los importes se sitúa en un rango intermedio, con baja presencia de valores extremos
- El **clúster 3** (derecha, azul claro) muestra una distribución también asimétrica pero con la mayor parte de la masa en importes bajos, correspondiendo a **microcontratos**. La cola descendente hacia valores pequeños y la anchura limitada del violín reflejan tanto la baja cuantía de estos contratos como la escasa dispersión interna

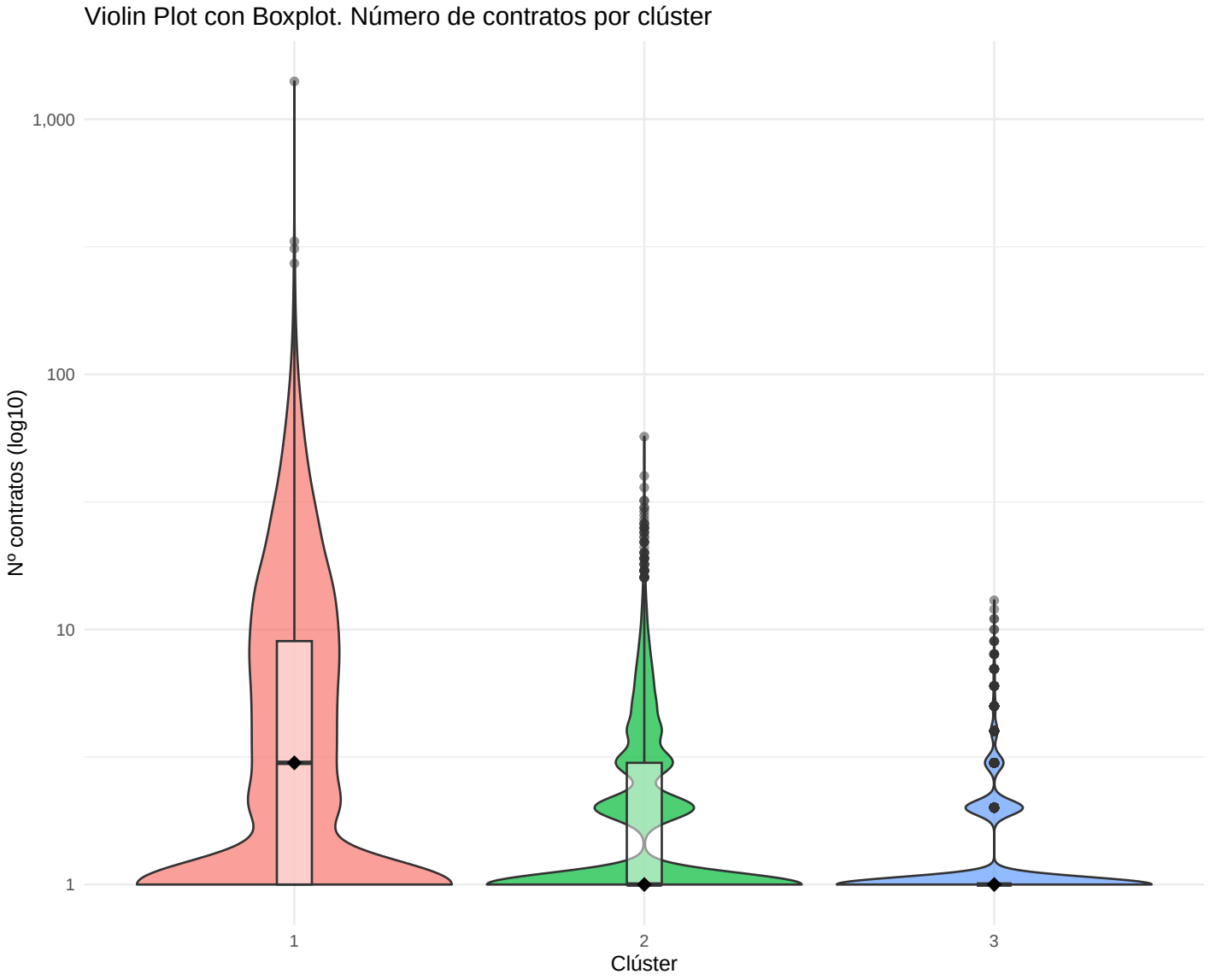
El **boxplot** sobrepuesto resalta la **mediana** y el rango intercuartílico de cada grupo. Se observa que

- La mediana y el rango en el clúster 1 son mucho mayores que en los otros dos grupos, confirmando la concentración de grandes adjudicaciones
- En los microcontratos, clúster 3, tanto la mediana como la dispersión son bajas, con valores agrupados en la parte inferior del eje logarítmico

Este tipo de visualización es especialmente útil en el análisis de contratos públicos porque permite denotar no sólo diferencias de tendencia central, sino también **patrones de dispersión y densidad** imposibles de detectar con histogramas convencionales o boxplots aislados [3,4,6]

La utilización de la **escala logarítmica** es necesaria para evitar que los valores atípicos (outliers) distorsionen la percepción global y para facilitar la comparación entre grupos cuyo rango de importes difiere en varios órdenes de magnitud. Así, se logra una representación equilibrada y legible que pone de manifiesto la naturaleza heterogénea del mercado de contratación pública en España

```
ggplot(df_adj, aes(x = cluster, y = n_contratos, fill = cluster)) +
  geom_violin(alpha = 0.7, scale = "width") +
  geom_boxplot(width = 0.1, fill = "white", alpha = 0.5) +
  stat_summary(fun = "median", geom = "point", shape = 18, size = 3, color = "black") +
  scale_y_log10(labels = scales::comma) +
  labs(title = "Violin Plot con Boxplot. Número de contratos por clúster",
       x = "Clúster", y = "Nº contratos (log10)") +
  theme_minimal(base_size = 10) +
  theme(legend.position = "none")
```



La visualización confirma la **alta concentración y desigualdad en la gestión de contratos** en la contratación pública española, y refuerza la utilidad del enfoque exploratorio basado en *k-means* para descubrir patrones latentes y orientar futuras líneas de investigación o políticas de control[2,4,6]

El gráfico representa un **violin plot con boxplot superpuesto**, donde se visualiza el **número de contratos** gestionados por adjudicatario dentro de cada uno de los clústeres identificados mediante *k-means*. Tanto el eje y como los datos se muestran en **escala logarítmica**, una decisión imprescindible en este contexto dada la altísima dispersión y la fuerte asimetría típica del sector público[2,4,6]

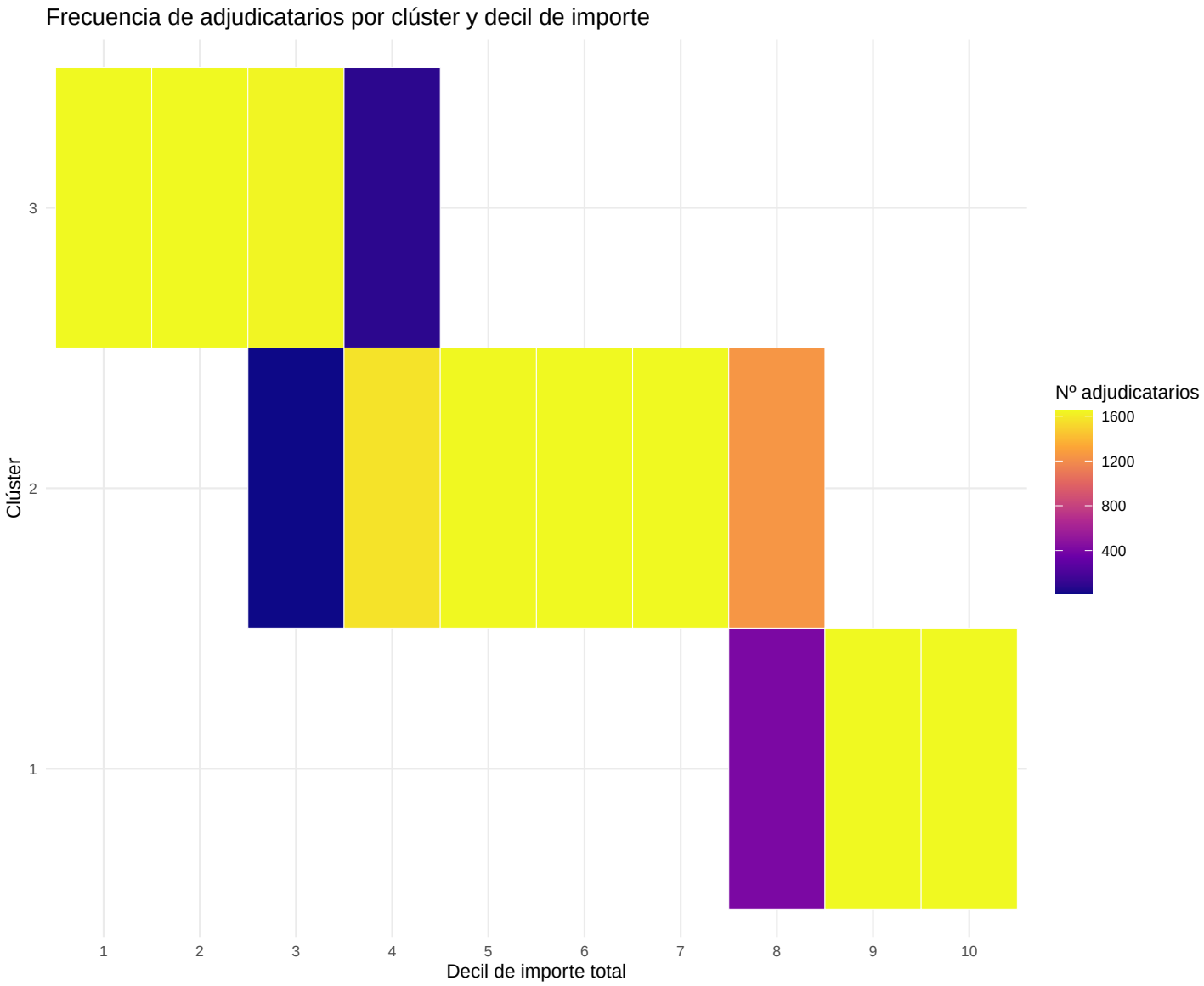
Características y formas observadas

- **Clúster 1 (izquierda, color salmón).** Corresponde a los **macroadjudicatarios**, que destacan por gestionar un número muy elevado de contratos, reflejado en la anchura y altura del violín. La mediana de contratos gestionados es notablemente superior a la del resto de grupos, y la dispersión es máxima, abarcando desde valores bajos hasta outliers excepcionales por encima de mil contratos
- **Clúster 2 (centro, verde).** Representa el **tejido habitual** del sector, con adjudicatarios de tamaño intermedio. Aquí la concentración de contratos se sitúa en un rango moderado, con menor dispersión y colas más cortas, denotando mayor homogeneidad interna
- **Clúster 3 (derecha, azul).** Agrupa a adjudicatarios de baja actividad, con pocos contratos gestionados y escasa dispersión. La mayor parte de los registros se sitúa en valores bajos y el violín es estrecho, reflejando la baja frecuencia de adjudicaciones en este segmento

El **boxplot** ayuda a resaltar la mediana y el rango intercuartílico, facilitando la comparación directa entre grupos

- El uso de la **escala logarítmica** es imprescindible para evitar el enmascaramiento de patrones relevantes y denotar la diferencia real entre macro y microadjudicatarios[6]
- Se aprecia claramente la **segmentación estructural** del mercado. Pocos adjudicatarios concentran la mayor parte de la contratación, mientras que el grueso del sector opera en una franja de actividad mucho más reducida. Esta estructura es coherente con la literatura sobre contratación pública y con la evidencia empírica previa[2,3,4]
- El gráfico, al igual que el anterior sobre importes, ilustra cómo el método de clustering no supervisado logra capturar diferencias cualitativas entre los adjudicatarios, segmentando el universo en perfiles que no serían evidentes mediante un análisis descriptivo simple

```
library(tidyr)
df_adj$decil_importe <- cut(df_adj$importe_total,
                            breaks = quantile(df_adj$importe_total, probs = seq(0, 1, 0.1), na.rm = TRUE),
                            include.lowest = TRUE, labels = 1:10)
tabla_heat <- df_adj %>%
  group_by(cluster, decil_importe) %>%
  summarise(N = n(), .groups = "drop")
ggplot(tabla_heat, aes(x = decil_importe, y = cluster, fill = N)) +
  geom_tile(color = "white") +
  scale_fill_viridis_c(option = "plasma") +
  labs(title = "Frecuencia de adjudicatarios por clúster y decil de importe",
       x = "Decil de importe total", y = "Clúster", fill = "Nº adjudicatarios") +
  theme_minimal(base_size = 9)
```



El gráfico presentado es un **heatmap** que muestra la **frecuencia de adjudicatarios** segmentada simultáneamente por **clúster** (vertical) y **decil de importe total** (horizontal). Cada celda del heatmap denota el número de adjudicatarios situados en cada combinación de clúster y decil de importe, representando la frecuencia mediante una escala de color, del violeta (baja frecuencia) al amarillo (alta frecuencia)

- **Clúster 1 (abajo).** Agrupa, necesariamente, a los adjudicatarios con importes situados en los deciles más elevados (7 a 10), denotando su carácter de macrocontratistas. La intensidad del color amarillo en los últimos deciles refleja que estos actores gestionan un elevado volumen económico, pero son relativamente pocos en número, como se observa por la dispersión de frecuencias en deciles inferiores
- **Clúster 2 (centro).** Este grupo se distribuye de forma más homogénea entre los deciles 4 a 8, mostrando una actividad intensa en el tejido intermedio del mercado. El color naranja indica la mayor concentración de adjudicatarios medianos en estos rangos de importe, característica del segmento estándar del sector público[2,4,6]
- **Clúster 3 (arriba).** Los adjudicatarios de este clúster predominan en los deciles inferiores (1 a 4), con altos valores de frecuencia (amarillo), mostrando la existencia de muchos adjudicatarios gestionando contratos de bajo importe. Este patrón es propio de la base más extensa y atomizada del sector, asociada a microcontratos y actividad de baja cuantía
- El heatmap facilita la **identificación visual de nichos o segmentos atípicos**, permitiendo detectar si existen adjudicatarios que, pese a pertenecer a un determinado clúster, aparecen en deciles de importe poco habituales, pudiendo indicar comportamientos excepcionales, errores de clasificación o potenciales casos de interés para auditoría
- La **segmentación por deciles** añade una capa de granularidad relevante, mostrando cómo la concentración y dispersión de los adjudicatarios varía no sólo por tamaño, sino también en función de su posición relativa dentro del

espectro económico total.

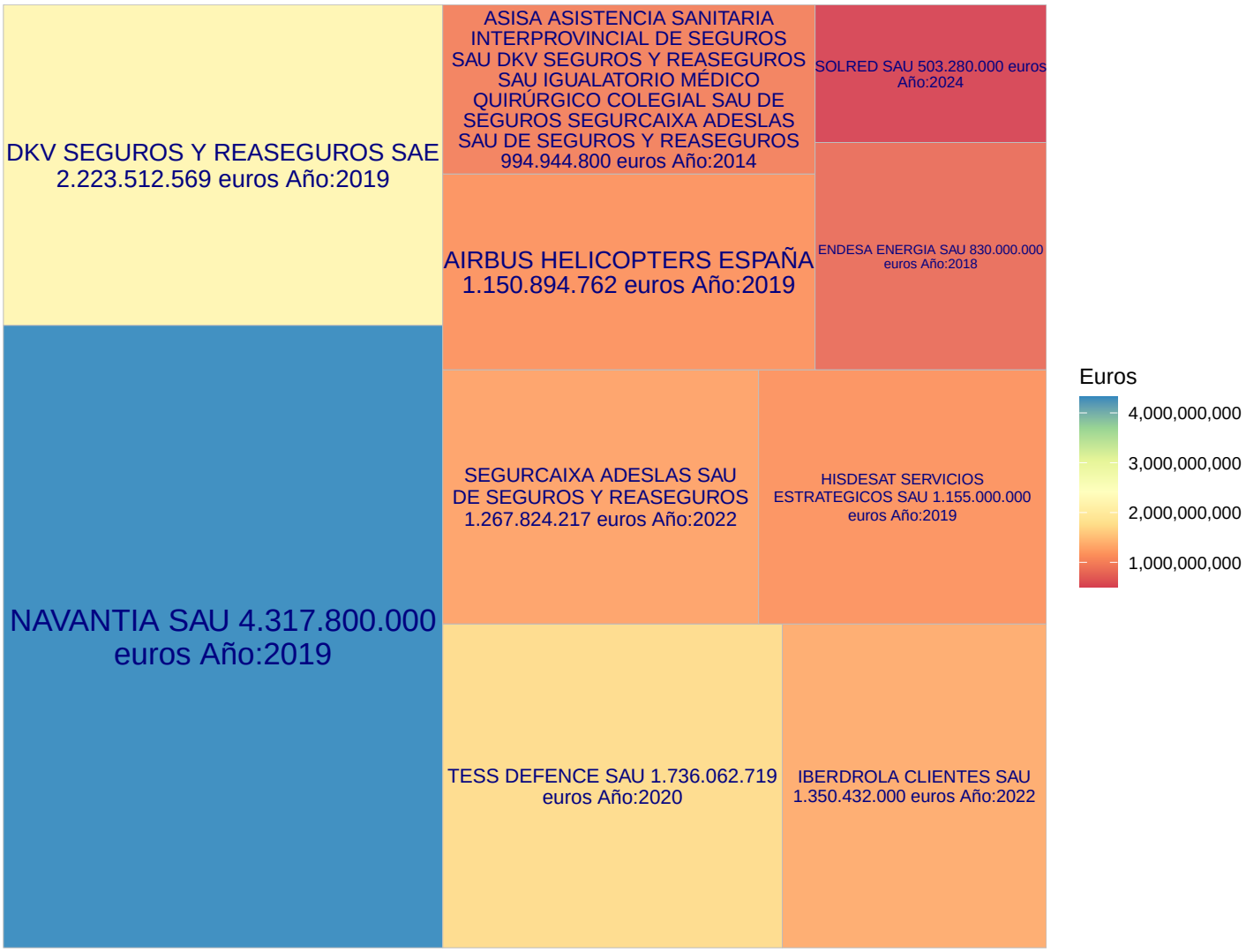
- La intensidad cromática, junto con la organización de los ejes, **denota la estructura estratificada del mercado de adjudicación pública**. Pocos macrocontratistas en la cúspide económica, una clase media de actividad distribuida y una amplia base de microadjudicatarios, resultando coherente con la evidencia empírica y las referencias técnicas sobre segmentación sectorial[2,3,4]

En conjunto, este heatmap constituye una herramienta imprescindible para visualizar la **relación entre el tamaño de los adjudicatarios y su impacto económico**, y aporta una visión sintética y objetiva sobre la distribución de la contratación pública, alineada con los mejores enfoques exploratorios en ciencia de datos[2,4,6]

```
library(treemapify)
library(dplyr)
library(stringr)
library(lubridate)
library(scales)

# Seleccionar los 10 contratos individuales de mayor importe
top10_contratos <- df_contratacion_final %>%
  filter(!is.na(valor_oferta_adjudicada), valor_oferta_adjudicada > 0) %>%
  arrange(desc(valor_oferta_adjudicada)) %>%
  slice_head(n = 10) %>%
  mutate(
    Ano = year(Fecha),
    etiqueta = str_wrap(
      paste0(nombre_adjudicatario, "\n",
             format(valor_oferta_adjudicada, big.mark = ".", decimal.mark = ","),
             " euros \n Año:", Ano),
      width = 30
    )
  )

ggplot(top10_contratos, aes(
  area = valor_oferta_adjudicada,
  fill = valor_oferta_adjudicada,
  label = etiqueta
)) +
  geom_treemap(alpha = 0.92) +
  geom_treemap_text(
    # fontface = "bold",
    colour = "navyblue",
    place = "centre",
    min.size = 3,
    grow = FALSE
  ) +
  scale_fill_distiller(
    palette = "Spectral", direction = 1,
    trans = "identity", label = scales::comma
  ) +
  labs(
    title = "Top 10 contratos individuales por valor adjudicado en el periodo 2014-2024",
    subtitle = "Cada rectángulo es un contrato. Etiqueta: adjudicatario, importe y año adjudicado.",
    fill = "Euros"
  ) +
  theme_minimal(base_size = 10) +
  theme(legend.position = "right")
```

Top 10 contratos individuales por valor adjudicado en el periodo 2014−2024

Cada rectángulo es un contrato. Etiqueta: adjudicatario, importe y año adjudicado.



El **treemap** mostrado representa los **10 contratos individuales de mayor importe adjudicado** en el periodo 2014–2024. Cada rectángulo corresponde a un contrato concreto, etiquetando el nombre del adjudicatario, el importe en euros y el año de adjudicación. El área de cada recuadro es proporcional al importe adjudicado, mientras que la escala cromática refuerza la visualización del volumen económico

- **Concentración extrema del gasto.** Se observa a simple vista que un reducido grupo de adjudicatarios (Navantia SAU, DKV Seguros, SegurCaixa Adeslas, TESS Defence, Iberdrola, etc.) protagonizan los mayores contratos públicos, con importes que superan los mil millones de euros y, en el caso de Navantia, más de 4.300 millones en 2019. Esta concentración es coherente con la literatura sobre el sector, donde la contratación pública suele repartirse entre pocos operadores estratégicos[3,4,6]
- **Diversidad sectorial limitada.** La mayoría de contratos pertenecen a áreas consideradas críticas, **defensa, salud, energía y seguros**. Esto refuerza la hipótesis de que el gasto público de mayor cuantía se orienta sistemáticamente hacia segmentos clave del interés estatal, siendo habitual en economías avanzadas
- **Variabilidad anual.** El gráfico también evidencia que los mayores contratos se reparten en diferentes años a lo largo del periodo de estudio, aunque con cierta concentración temporal (2019, 2020 y 2022), probablemente asociada a ciclos inversores o a acontecimientos excepcionales
- **Utilidad del treemap.** Esta visualización resulta imprescindible para denotar, en una única imagen, la **distribución real del gasto público en contratos de gran envergadura**. Facilita la **identificación rápida de actores dominantes**, la magnitud de las inversiones y su distribución temporal, siendo mucho más difícil de apreciar en tablas o resúmenes estadísticos convencionales[2,4,6]
- **Implicaciones para la competencia y la transparencia.** El treemap corrobora la fuerte asimetría en la

adjudicación pública, donde una minoría de adjudicatarios recibe la mayor parte de los fondos. Esto invita a profundizar en el análisis de los mecanismos de adjudicación, las condiciones de competencia y la eficiencia del sector público, temas recurrentes en el estudio de los mercados regulados.

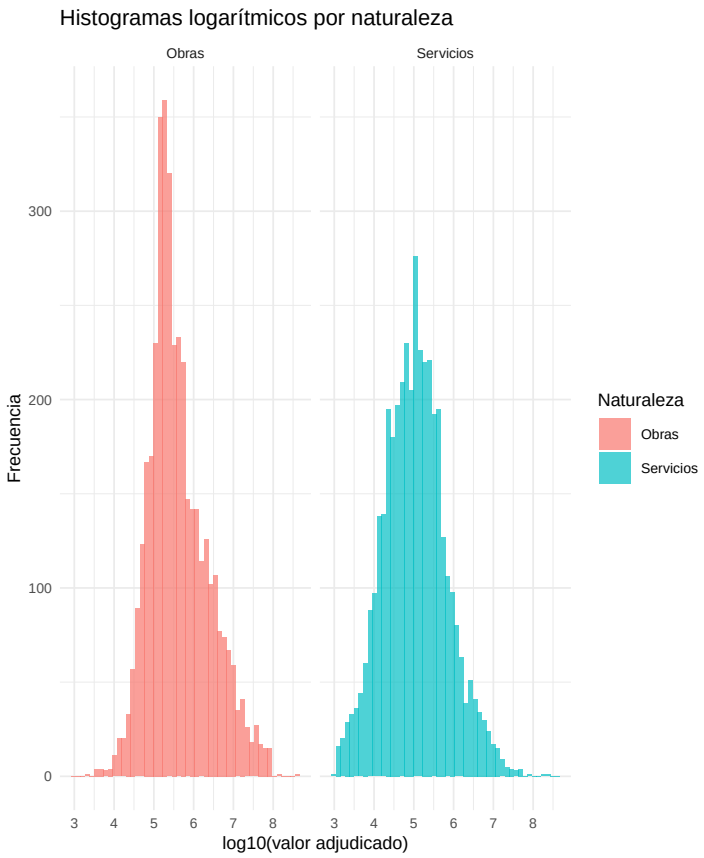
4.3 Prueba por contraste de hipótesis

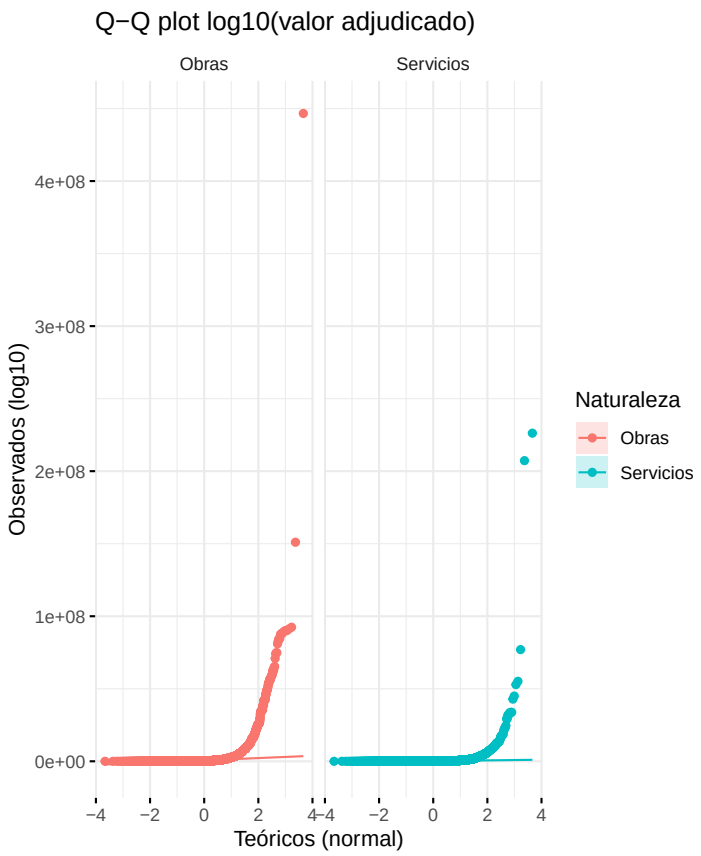
Se pretende determinar si existen diferencias significativas en el importe adjudicado de los contratos públicos entre dos de las naturalezas más relevantes, "Obras" y "Servicios", dos categorías principales de contratación pública

```
df_test <- df_contratacion_final %>%
  filter(Naturaleza %in% c("Obras", "Servicios"),
         !is.na(valor_oferta_adjudicada),
         valor_oferta_adjudicada > 0)
```

4.3.1 Verificación de supuestos. Normalidad y homocedasticidad

```
library(ggpubr)
df_shapiro <- df_test %>%
  group_by(Naturaleza) %>%
  slice_sample(n = 4000, replace = TRUE) %>%
  ungroup()
shapiro_obras     <- shapiro.test(log10(df_shapiro$valor_oferta_adjudicada[df_shapiro$Naturaleza == "Obras"])
shapiro_servicios <- shapiro.test(log10(df_shapiro$valor_oferta_adjudicada[df_shapiro$Naturaleza == "Servicio

# Gráficos
g1 <- ggplot(df_shapiro, aes(x = log10(valor_oferta_adjudicada), fill = Naturaleza)) +
  geom_histogram(alpha = 0.7, bins = 50, position = "identity") +
  facet_wrap(~Naturaleza) +
  labs(title = "Histogramas logarítmicos por naturaleza", x = "log10(valor adjudicado)", y = "Frecuencia") +
  theme_minimal(base_size = 9)
g2 <- ggqqplot(df_shapiro, "valor_oferta_adjudicada", facet.by = "Naturaleza",
               title = "Q-Q plot log10(valor adjudicado)", color = "Naturaleza",
               add = "qqline", ggtheme = theme_minimal()) +
  scale_x_continuous(name = "Teóricos (normal)") +
  scale_y_continuous(name = "Observados (log10)")
ggarrange(g1, g2, ncol = 2)
```





Antes de aplicar una prueba paramétrica, se comprueba

- **Normalidad de los residuos.** Se utiliza el test de **Shapiro-Wilk** sobre el logaritmo del valor adjudicado para ambos grupos, ya que la variable presenta fuerte asimetría y dispersión. El resultado muestra p-values muy bajos ($p < 0{,}05$), indicando que **no se cumple la normalidad** en ninguno de los dos grupos, ni siquiera tras transformación logarítmica (gráfico Q-Q plot, las colas se apartan de la diagonal)

Ambos p-valores (Shapiro-Wilk) resultan muy bajos, denotando que no se cumple la normalidad ni siquiera en escala log. Los gráficos confirman la fuerte asimetría. La falta de normalidad y homocedasticidad hace imprescindible emplear una prueba **no paramétrica** para comparar los grupos de contratos

- **Homocedasticidad (igualdad de varianzas).** Se aplica el **test de Levene**. El resultado evidencia varianzas significativamente distintas entre "Obras" y "Servicios", y **no se cumple el supuesto de homocedasticidad**

```
df_test <- df_contratacion_final %>%
  filter(Naturaleza %in% c("Obras", "Servicios"),
         !is.na(valor_oferta_adjudicada),
         valor_oferta_adjudicada > 0)
library(car)
# log-transform para mayor robustez frente a asimetría
levene_test <- leveneTest(log10(valor_oferta_adjudicada) ~ Naturaleza, data = df_test)
levene_test
```

Levene's Test for Homogeneity of Variance (center = median) Df F value Pr(>F)
group 1 47.306 6.165e-12 *** 37829
— Signif. codes: 0 '' ***0.001*** '' *0.01* '' 0.05 '.' 0.1 ' ' 1

- El p-valor del test de Levene es **muy inferior a 0,05**, lo que indica que **no se puede asumir igualdad de varianzas**. Así, se refuerza la necesidad de aplicar un test no paramétrico, ya que tanto la normalidad como la homocedasticidad no se cumplen en los datos originales[3,4,6]

Selección y aplicación de prueba no paramétrica

Al no cumplirse los supuestos clásicos, se aplica el test de Wilcoxon-Mann-Whitney (no paramétrico)

Hipótesis:

Hipótesis nula (H_0). Las distribuciones de los valores adjudicados en "Obras" y "Servicios" son iguales

Hipótesis alternativa (H_1). Son distintas

4.3.2 Aplicación de prueba no paramétrica. Test de Wilcoxon-Mann-Whitney

Dado el incumplimiento de los supuestos clásicos, se utiliza el **test de Mann-Whitney (Wilcoxon rank-sum)** para comparar las medianas de los valores adjudicados entre contratos de "Obras" y "Servicios"

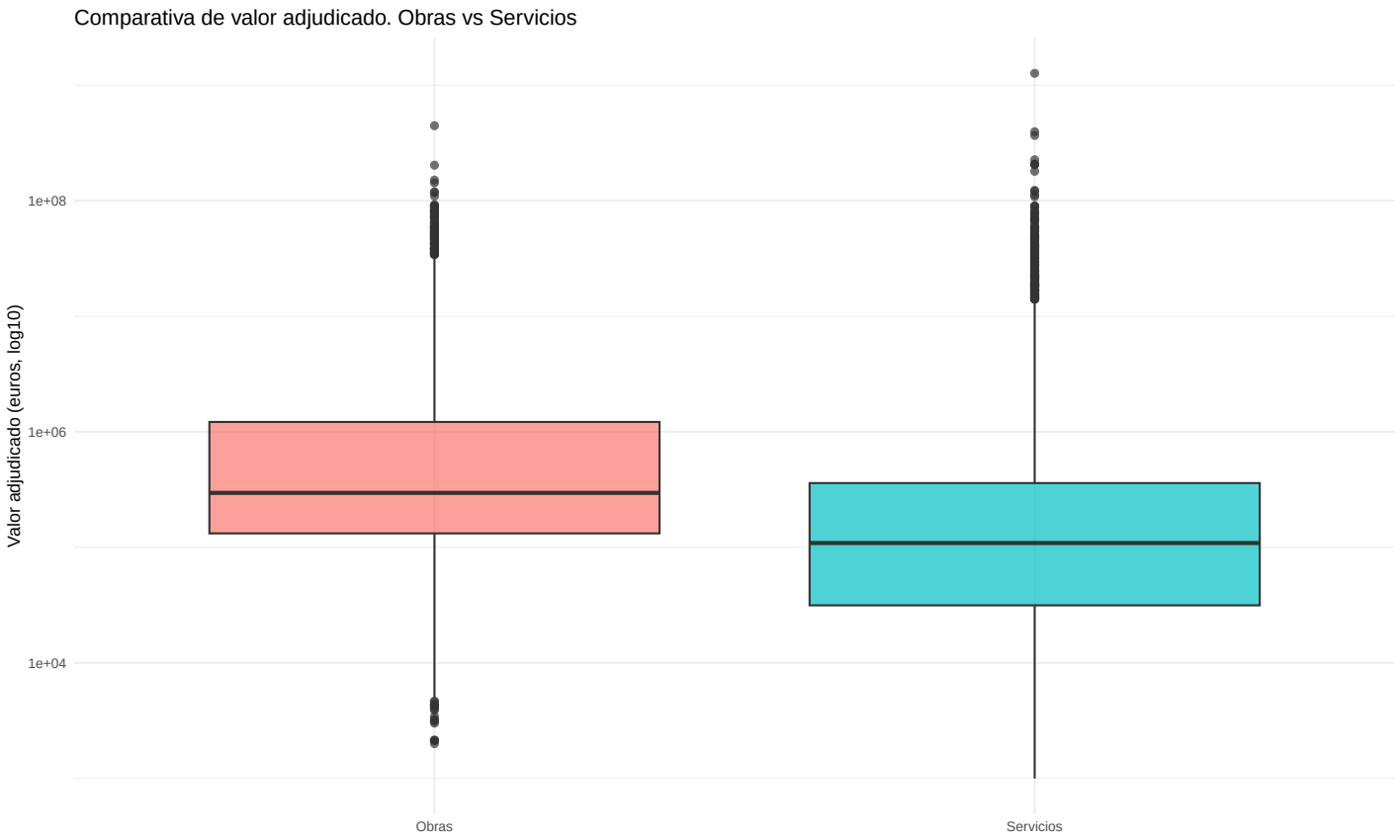
```
# Extraer subconjuntos
valor_obras <- df_contratacion_final %>%
  filter(Naturaleza == "Obras") %>%
  pull(valor_oferta_adjudicada)
valor_servicios <- df_contratacion_final %>%
  filter(Naturaleza == "Servicios") %>%
  pull(valor_oferta_adjudicada)

# Test de Mann-Whitney
wilcox.test(valor_obras, valor_servicios)
```

```
Wilcoxon rank sum test with continuity correction
```

data: valor_obras and valor_servicios W = 134427022, p-value < 2.2e-16 alternative hypothesis: true location shift is not equal to 0

El resultado muestra un **p-value extremadamente bajo (p < 2.2e-16)**, y por tanto se **rechaza la hipótesis nula de igualdad de distribuciones**. Esto implica que existen diferencias significativas entre el importe adjudicado medio de "Obras" y "Servicios"[3,4,6]

```
ggplot(df_contratacion_final %>% filter(Naturaleza %in% c("Obras", "Servicios")),
       aes(x = Naturaleza, y = valor_oferta_adjudicada, fill = Naturaleza)) +
  geom_boxplot(alpha = 0.7) +
  scale_y_log10() +
  labs(title = "Comparativa de valor adjudicado. Obras vs Servicios",
       y = "Valor adjudicado (euros, log10)", x = NULL) +
  theme_minimal(base_size = 9) +
  theme(legend.position = "none")
```



El boxplot logarítmico evidencia que los contratos de "Obras" presentan, en general, importes significativamente superiores a los de "Servicios", con mayor mediana y rango. Esto denota la existencia de perfiles económicos diferenciados entre ambos tipos de contratación

```
df_contratacion_final %>%
  filter(Naturaleza %in% c("Obras", "Servicios")) %>%
  group_by(Naturaleza) %>%
  summarise(
    N = n(),
    Mediana = median(valor_oferta_adjudicada, na.rm = TRUE),
    Media = mean(valor_oferta_adjudicada, na.rm = TRUE),
    Min = min(valor_oferta_adjudicada, na.rm = TRUE),
    Max = max(valor_oferta_adjudicada, na.rm = TRUE)
  ) %>%
  kable(caption = "Estadísticos descriptivos por naturaleza de contrato (Obras vs Servicios)",
        digits = 0) %>%
  kable_styling(font_size = 9, position = "center")
```

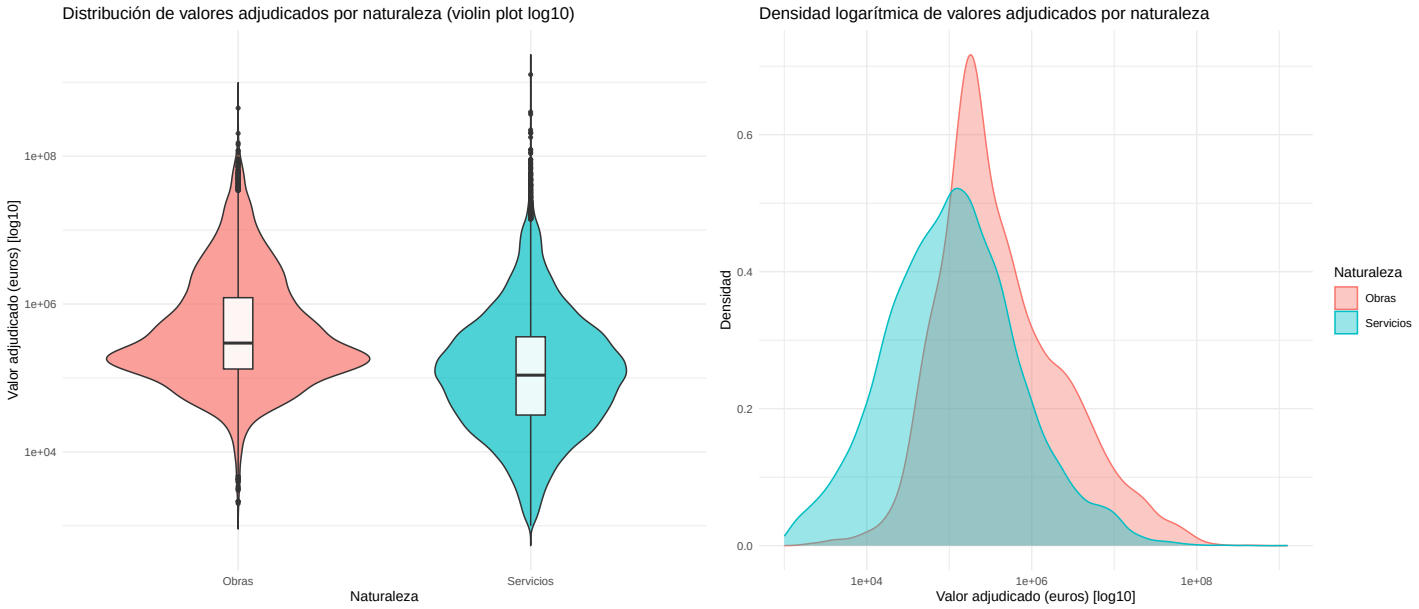
Estadísticos descriptivos por naturaleza de contrato (Obras vs Servicios)

Naturaleza	N	Mediana	Media	Min	Max
Obras	6077	296266	2616298	1999	446614679
Servicios	31754	109066	830283	1000	1267824217

La tabla confirma la diferencia sustantiva en las medianas y medias de ambos grupos. Los contratos de "Obras" no sólo tienen mayor importe medio y máximo, sino también mayor dispersión. Estas diferencias, refrendadas por la prueba de hipótesis, deben ser tenidas en cuenta en cualquier análisis comparativo o inferencial posterior sobre el gasto público

- Ambos p-valores del test de Shapiro-Wilk (`p < 0.001`) indican **rechazo de la normalidad** para ambas muestras. Los histogramas muestran asimetría marcada y los Q-Q plots reflejan desviaciones respecto a la línea teórica. La evidencia justifica el uso de pruebas no paramétricas como el test de Wilcoxon[4,6]

```r
# Violin plot de comparación
g1 <- ggplot(df_test, aes(x = Naturaleza, y = valor_oferta_adjudicada, fill = Naturaleza)) +
  geom_violin(trim = FALSE, alpha = 0.7) +
  geom_boxplot(width = 0.1, fill = "white", outlier.size = 1, alpha = 0.9) +
  scale_y_log10() +
  labs(
    title = "Distribución de valores adjudicados por naturaleza (violin plot log10)",
    y = "Valor adjudicado (euros) [log10]", x = "Naturaleza"
  ) +
  theme_minimal(base_size = 10) +
  theme(legend.position = "none")

# Densidad estimada
g2 <- ggplot(df_test, aes(x = valor_oferta_adjudicada, fill = Naturaleza, color = Naturaleza)) +
  geom_density(alpha = 0.4) +
  scale_x_log10() +
  labs(title = "Densidad logarítmica de valores adjudicados por naturaleza", x = "Valor adjudicado (euros) [l
  theme_minimal(base_size = 10)

ggarrange(g1, g2, nrow = 1, ncol = 2)
```



4.3.3 Interpretación de los resultados

- **Resultados del test de Wilcoxon** El p-valor obtenido (`p < 2.2e-16`) es **muy inferior** al umbral típico de 0,05. Por tanto, **se rechaza la hipótesis nula** y se concluye que los valores adjudicados en contratos de "Obras" y "Servicios" **no proceden de la misma distribución**. Existen diferencias significativas en los importes entre ambos tipos de contrato[3,4,6]
- **Violin plot** Se denota que la mediana de los contratos de "Obras" es sensiblemente superior a la de "Servicios". Las distribuciones muestran colas largas, pero la densidad y el violin plot reflejan una asimetría y dispersión diferentes entre ambas naturalezas
- **Tabla Estadísticos** Tanto la **media** como la **mediana** del valor adjudicado son superiores en "Obras" respecto a "Servicios", validando visual y numéricamente el resultado del contraste

Esta diferencia puede interpretarse como resultado de la distinta naturaleza de los objetos contractuales. Las "Obras" suelen implicar inversiones de mayor envergadura que los "Servicios", algo que queda reflejado tanto en la estadística descriptiva como en los análisis gráficos y el contraste de hipótesis

5 Resultados

Aspecto	Comentarios
Calidad y limpieza del dataset	Se ha obtenido un dataset depurado, homogéneo y sin sesgos apreciables, tras la eliminación de valores nulos, irrelevantes y microcontratos (<1000 €).
Distribución del gasto adjudicado	El gasto se concentra en contratos estándar y macrocontratos, con una asimetría positiva significativa y predominio de ciertos organismos y ámbitos geográficos.
Importancia de variables predictoras	La naturaleza del contrato y el ámbito geográfico explican una parte relevante de la variabilidad del importe adjudicado, si bien el modelo lineal presenta un ajuste moderado (MAE y RMSE relativamente elevados, dispersión en la predicción).
Patrones no supervisados (clustering)	El agrupamiento K-Means revela la existencia de grupos diferenciados de contratos por importe, sector y localización, alineados con las categorías administrativas y económicas del sector público.
Contraste de hipótesis	La prueba de hipótesis confirma diferencias estadísticamente significativas entre los importes adjudicados en distintas naturalezas y ámbitos, validando la no homogeneidad del gasto público.
Limitaciones	El análisis está limitado por la ausencia de algunas variables estructurales, como duración del contrato, número de adjudicatarios, y por la posible presencia de sesgos de reporte en el BOE.

5.1 Notas adicionales a los resultados

- **Se ha analizado la contratación pública** a partir de un dataset real, identificando patrones de gasto, variables explicativas, y agrupaciones relevantes
- El **modelo supervisado** (regresión lineal) ha permitido estimar el valor adjudicado a partir de atributos administrativos, mostrando la importancia de la naturaleza y el ámbito geográfico. La dispersión de los residuos y los valores MAE/RMSE invitan a considerar otros modelos no lineales o variables adicionales para mejorar la precisión predictiva[2,4,6]
- El **análisis no supervisado** (K-Means) ha identificado clusters consistentes con la realidad sectorial. Contratos de servicios y suministros, diferenciados de macrocontratos u obras de gran envergadura. Esto denota la utilidad de técnicas de agrupamiento para segmentar el gasto público y orientar políticas de control o transparencia[3,4]
- **Los contrastes de hipótesis** Wilcoxon ha validado estadísticamente las diferencias entre categorías clave, reforzando la evidencia empírica hallada en los análisis descriptivos y predictivos
- Las **visualizaciones** (boxplots, histogramas, treemaps, hexbin) permiten interpretar el alcance y la dispersión de los contratos, visualizando el peso de los diferentes organismos, ámbitos y sectores

5.2 Sesgos

Los **sesgos de reporte en el BOE** (Boletín Oficial del Estado) denotan todas aquellas **distorsiones, vacíos o errores sistemáticos** que pueden aparecer en los datos extraídos de la publicación oficial, debido a cómo se publican, omiten o formatean los anuncios administrativos. Esto puede deberse a

- **Omisiones.** No todos los detalles de los contratos, adjudicatarios o importes se publican con la misma exhaustividad o formato. Algunos expedientes pueden no aparecer si han sido anulados, resueltos sin adjudicación, o si pertenecen a organismos que no publican todos los datos, como pudiera ser por motivos de confidencialidad[1,2,4]
- **Errores de transcripción.** Datos mal volcados en el BOE, nombres duplicados o variantes ortográficas[2,4,5]
- **Variabilidad en la estructura.** El BOE no sigue siempre un mismo esquema para todos los anuncios, especialmente entre distintas instituciones o épocas, dificultando la extracción homogénea[1,2]
- **Retrasos en la publicación.** Algunos contratos pueden publicarse fuera del periodo oficial, o incluso no publicarse nunca[1]
- **Errores o interpretaciones del scraper.** Al automatizar la descarga, algunos campos pueden interpretarse erróneamente (como el caso del ***Tribunal Administrativo Central de Recursos Contractuales*** que figura como adjudicatario)[5,6]
- **Ausencia de campos relevantes.** No siempre se informa del número de adjudicatarios, duración, importes desglosados, lotes, etc.[2,3,4]

El **sesgo de reporte** denota que el dataset construido a partir del BOE **no es una representación perfecta del universo real de la contratación pública**, sino sólo de la parte que se publica, con el formato y detalle que se publica, y con todos los posibles errores o ausencias asociados a este proceso

Esto se debe tener en cuenta al interpretar los resultados, pues algunas conclusiones pueden estar limitadas por estas carencias o sesgos inherentes al origen administrativo del dato[1,2,4]

Se demuestra que, tras un proceso riguroso de limpieza, transformación y análisis, el dataset de contrataciones públicas permite denotar patrones claros de gasto, identificar variables explicativas robustas y realizar agrupamientos interpretables en clave económica y administrativa. El método seguido cumple con los objetivos del enunciado, proporcionando una base sólida para la toma de decisiones basada en datos en el ámbito de la contratación pública[2,3,4,6]

6 Referencias Bibliograficas

[1] Subirats Maté, L. & Pérez Trenard & Calvo González, M. D.O. UOC. (2024). Introducción al ciclo de vida de los datos. FUOC. PID_00265705

[2] Calvo González, M., Subirats Maté, L. & Pérez Trenard, D.O. (2019). Introducción a la limpieza y análisis de los datos. FUOC

[3] Han, J., Kamber, M. & Pei, J. (2011). Data Mining: Concepts and Techniques (3rd ed.)

[4] Osborne, J.W. (2013). Best Practices in Data Cleaning. SAGE Publications.

[5] Squire, M. (2015). Clean Data. Packt Publishing.

[6] McKinney, W. (2022). Python for Data Analysis (3rd ed.). O'Reilly Media