PartePol

1. Introducción

La educación es ampliamente reconocida como uno de los pilares fundamentales para el desarrollo económico y social de un país. Sin embargo, el acceso y la permanencia en el sistema educativo conllevan una serie de costes directos e indirectos que son asumidos en gran medida por los hogares. Este gasto no se limita únicamente a las tasas y matrículas, sino que abarca un complejo ecosistema de bienes y servicios que incluye desde libros de texto y material escolar hasta servicios de comedor, transporte, actividades extraescolares y clases particulares.

Comprender la magnitud y la composición de este gasto es esencial para evaluar la equidad del sistema educativo y el esfuerzo económico que supone para las familias. El Instituto Nacional de Estadística (INE) proporciona una visión detallada de esta realidad a través de la Encuesta de Gasto de los Hogares en Educación (EGHE).

El presente proyecto tiene como objetivo realizar un Análisis Exploratorio de Datos (AED) sobre los microdatos de la EGHE 2019. Nuestro propósito es diseccionar la estructura del gasto en educación en España, identificando los patrones subyacentes, las diferencias entre distintos tipos de enseñanza y las características de los estudiantes que más gasto generan.

Para guiar nuestro análisis, planteamos las siguientes preguntas de investigación:

1. **¿Cómo se distribuye el gasto educativo entre los diferentes tipos de centros (públicos, concertados, privados) y qué factores explican estas diferencias?** Compararemos los patrones de gasto entre los tres regímenes de financiación, analizando no solo las diferencias en el gasto total sino también en su composición (matrículas, servicios complementarios, material).
2. **¿Qué características del hogar y del estudiante predicen un mayor gasto educativo? ¿Existen diferencias significativas por sexo, composición familiar o tamaño del municipio?** Identificaremos los principales predictores del gasto educativo mediante el análisis de variables demográficas y estructurales. Evaluaremos específicamente si existen brechas de gasto asociadas al sexo del estudiante y el efecto de la composición del hogar.
3. **¿Cómo evoluciona la estructura del gasto a lo largo del itinerario educativo, desde Infantil hasta la Universidad, y qué servicios adquieren mayor relevancia en cada etapa?** Compararemos el gasto medio en diferentes niveles, como Educación Infantil, Primaria, ESO, Bachillerato y Universidad.
4. **¿Qué peso real tienen los gastos complementarios (clases particulares, actividades extraescolares, material específico) en el presupuesto total y qué familias los asumen en mayor medida?** Cuantificaremos el impacto de partidas específicas como clases particulares, servicios de comedor, actividades extraescolares y uniformes escolares, que frecuentemente pasan desapercibidos en el debate público sobre el coste de la educación.
5.

Este informe detallará el proceso de limpieza y transformación de los datos, seguido del análisis univariante y bivariante para dar respuesta a estas preguntas, concluyendo con los principales hallazgos de nuestro estudio.

2. Carga y Preparación de Datos

El punto de partida de este análisis son los microdatos de la Encuesta de Gasto de los Hogares en Educación (EGHE) del año 2019, proporcionados por el INE. El fichero original, *EGHE_2019.csv*, es un conjunto de datos que, si bien es completo, presenta desafíos significativos para un análisis directo:

- **Codificación de Variables**: Los nombres de las variables (ej. GTT, C01, NEST2) y sus valores (ej. 1, 2, 3) siguen la codificación interna del INE, lo que los hace poco intuitivos.
- **Valores Nulos (NA)**: El dataset utiliza NA (Not Available) de manera ambigua. Un NA en un campo de importe (como importe_comedor) puede significar que el dato se desconoce o, más probablemente, que el estudiante no incurrió en dicho gasto (gasto de 0€).
- **Datos Irrelevantes**: El dataset original contiene numerosas columnas y filas que no son pertinentes para nuestras preguntas de investigación (ej. personas que no son estudiantes, columnas de control de la encuesta). Estos datos o variables son eliminados.

Para abordar estos retos, se ha implementado un proceso de limpieza y transformación estructurado en tres fases, ejecutadas secuencialmente mediante los scripts *limpieza_gastos.R*, *limpieza_hogar.R* y *limpieza_estudiantes.R*.

El **chunk** de código asociado a la limpieza de los datos carga el conjunto de datos crudo y aplica estos tres scripts para generar el dataset limpio (**datos**) que se utilizará en el resto del informe. A continuación, se detallan las transformaciones clave realizadas durante este proceso.

2.1 Renombrado y Selección de Variables

El primer paso consistió en "traducir" el dataset. Se renombraron todas las variables de interés de sus códigos del INE a nombres descriptivos en español. Por ejemplo: *C01* se renombró a *Tipo_educacion*. Esto fué posible gracias al archivo de descripción del dataset proporcionado por el INE, donde se describe que tipo de nomenclatura se gasta en cada variable y como se codifican los datos.

Paralelamente, se eliminaron columnas innecesarias, variables con porcentajes demasiado elevados de NAs, como las relacionadas con información de becas, para simplificar la estructura del dataset y reducir el ruido.

2.2 Recodificación de Variables Categóricas

Las variables categóricas más importantes fueron transformadas a factores con etiquetas descriptivas para la correcta interpretación y visualización en los análisis.

Table 1: Ejemplos de Recodificación de Variables Categóricas

Variable	Código_INE	Etiqueta
Tipo_educacion	1, 2, 3	Pública, Concertada, Privada
SEXO	1, 2	Hombre, Mujer
Nacionalidad	1, 2, 3	Española, Extranjera, Doble nacionalidad

2.3 Imputación y Tratamiento de Valores Nulos

Partimos de la hipótesis de que, en un campo de gasto, un valor NA significa un gasto de 0€. Posteriormente, se realizó la comprobación de que, para todas las muestras, la suma de todos los tipos de gasto es igual al valor de la variable *gasto_total_educacion*:

$$\sum \textbf{importes individuales} = \textbf{gasto\_total\_educacion}$$

Tras comprobar que se cumplía esta condición, se aplicó la siguiente lógica de imputación:

- **Importes (*importe_*)**: Todos los valores NA en columnas de importes (ej. *importe_comedor*, *importe_clases_particulares*) se sustituyeron por 0, indicando un gasto nulo.
- **Servicios (*servicio_*)**: En las variables que indican si se usó un servicio (donde 1=Sí), los NA se imputaron como 2 (el código para "No").
- **Gastos Totales (*gasto_total_*)**: Del mismo modo, los NA en las columnas de gasto total se reemplazaron por 0.

2.4 Filtrado y Saneamiento de Datos

Finalmente, se aplicaron varios filtros para asegurar la coherencia y relevancia del dataset:

Se eliminaron registros con inconsistencias obvias, como un hogar con NHOGAR (número de personas en el hogar) inferior a 1.

Se eliminaron las filas donde variables fundamentales para nuestro análisis eran nulas, ya que no se podían imputar de forma fiable. Específicamente:

- Se eliminaron muestras con *gasto_total_educacion* nulo, pues es nuestra variable objetivo.
- Se eliminaron muestras con *Tipo_educacion* nulo, ya que esta es una variable explicativa central para la Pregunta 1.

Tras este proceso exhaustivo, el objeto datos queda listo para el análisis. Contiene únicamente las observaciones y variables relevantes, con tipos de datos correctos, sin valores nulos ambiguos y con etiquetas descriptivas.

3. Descripción del Conjunto de Datos Limpio: Análisis Univariante

Tras la ejecución de los scripts de limpieza, obtenemos el dataset `datos`. Este conjunto de datos contiene 4185 observaciones (cada una representando a un estudiante que reportó gastos) y 61 variables limpias y listas para el análisis.

El objetivo de esta sección es realizar un análisis univariante de las variables clave. Este paso es fundamental para entender la distribución y las características de cada variable de forma aislada, antes de buscar relaciones entre ellas. Nos centraremos en la variable objetivo (Gasto Total) y en las principales variables categóricas y numéricas que usaremos para responder a nuestras preguntas.

3.1 Variable Objetivo: Gasto Total Anual (gasto_total_educacion)

Nuestra principal variable de interés es el *gasto_total_educacion* anual por estudiante.

Table 2: Estadísticas Descriptivas del Gasto Total Anual por Estudiante

Métrica	Valor (€)
Media	1638.14

Mediana	972.00
Mínimo	0.00
Máximo	37853.00
1er Cuartil	422.00
3er Cuartil	1962.00
Desv. Estándar	2224.81

La tabla anterior revela un hallazgo clave: la media (1638.14 €) es significativamente más alta que la mediana (972 €). Esto indica una distribución asimétrica positiva (sesgada a la derecha).

En la práctica, esto significa que la mayoría de los hogares reportan un gasto relativamente bajo, pero un pequeño número de hogares (probablemente aquellos en centros privados de élite o con altos gastos en actividades complementarias) tienen gastos muy elevados, "tirando" de la media hacia arriba. Por esta razón, y tal como se planteaba en nuestras preguntas de investigación, la mediana será una métrica más robusta para describir el gasto del hogar "típico".

El siguiente histograma confirma visualmente esta fuerte asimetría. La gran mayoría de las observaciones se concentran en la parte izquierda del gráfico, con una larga cola de outliers hacia la derecha.

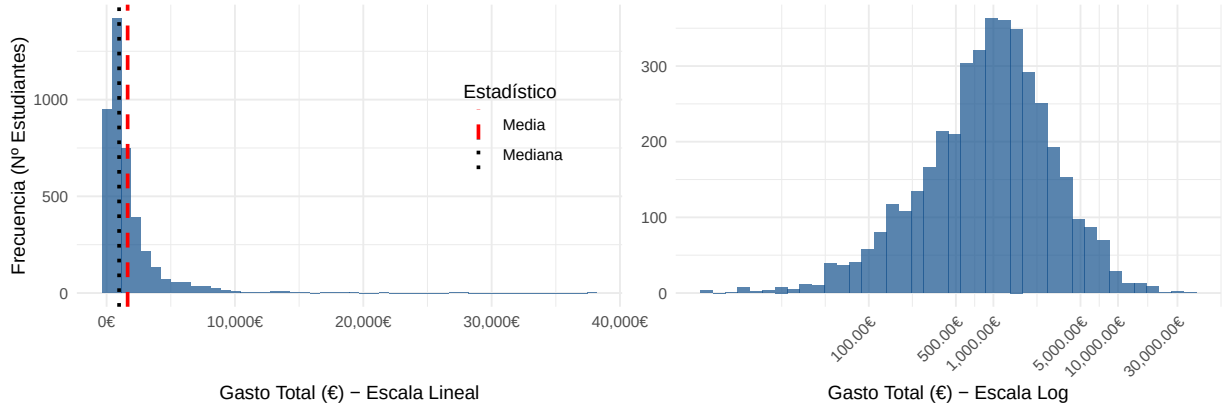


Figure 1: Histograma del Gasto Total Anual (Escala Lineal y Logarítmica).

La Figura 1 combina dos vistas de la misma variable:

- El gráfico de la izquierda (escala lineal) nos permite observar la fuerte asimetría positiva y el impacto de los *outliers*, que elevan la media (1638.14 €) muy por encima de la mediana (972 €).
- El gráfico de la derecha (escala logarítmica) complementa al primero. Al comprimir los valores altos y expandir los bajos, nos permite visualizar la forma de la distribución del grueso de los hogares. En esta vista, se puede apreciar mejor dónde se concentran los gastos más comunes.

3.2 Variables Categóricas Principales

A continuación, exploramos la composición de nuestra muestra en función de las variables categóricas clave que guían nuestras preguntas de investigación.

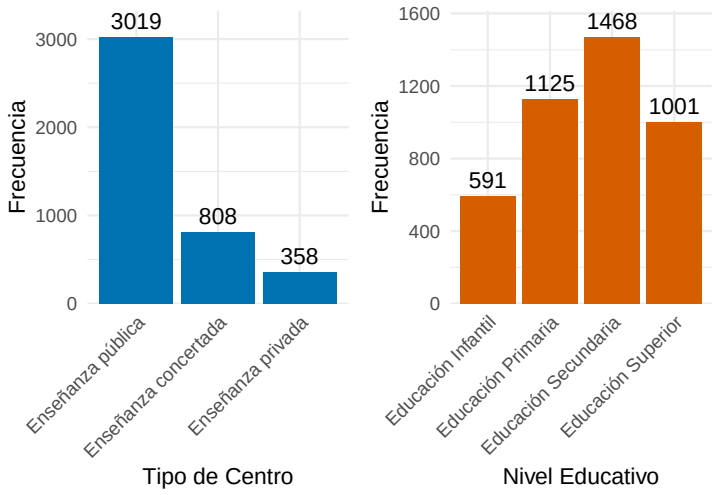


Figure 2: Nº de Estudiantes por Tipo de Centro y por Nivel Educativo

De estos gráficos, observamos que:

- **Tipo de Centro**: La mayoría de los estudiantes de la muestra se encuentran en la enseñanza pública, seguida de la concertada y, finalmente, la privada. Esta distribución es coherente con la estructura del sistema educativo español.
- **Nivel Educativo**: La muestra tiene una representación significativa en todas las etapas, con los picos de frecuencia en Educación Primaria y Secundaria. Es importante notar que si un grupo está poco representado (ej. "Otros estudios"), deberemos ser cautelosos al generalizar los resultados para ese grupo.

3.3 Componentes del Gasto (Bienes vs. Servicios)

Nuestra Pregunta 2 busca identificar cuál de los dos grandes componentes del gasto (Servicios vs. Bienes) tiene un mayor peso.

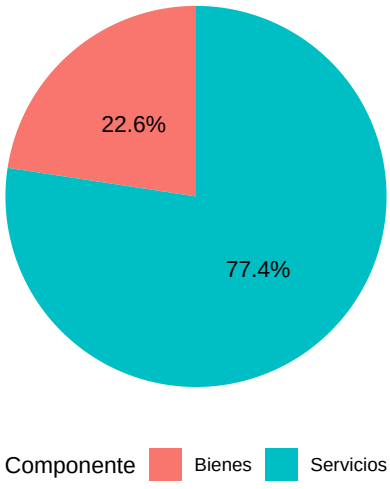


Figure 3: Desglose del Gasto Educativo Total

El gráfico circular muestra de forma concluyente que, a nivel agregado, el gasto en Servicios (matrículas, comedor, transporte, extraescolares, etc.) representa la mayor parte del desembolso total de los hogares, superando ampliamente al gasto en Bienes (libros, material, uniformes).

3.4 Gastos Complementarios

Finalmente, el análisis debe responder a dos preguntas planteadas inicialmente en la Pregunta 4:

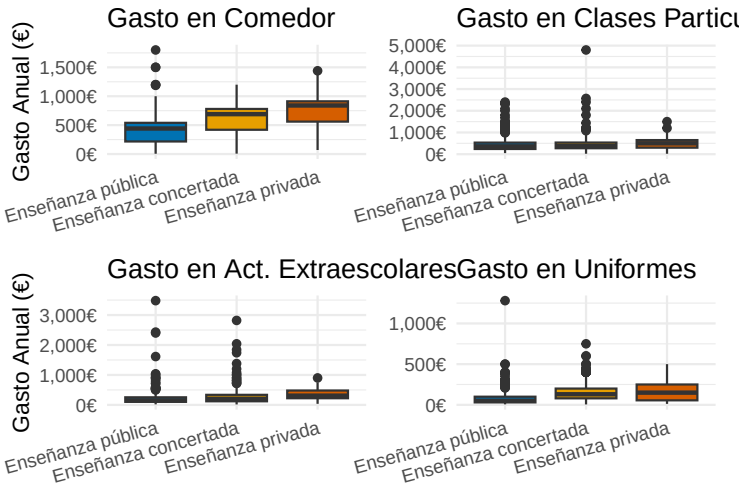
- ¿Qué porcentaje de familias incurre realmente en los gastos complementarios?
- Entre las que sí gastan, ¿cuál es el gasto medio?

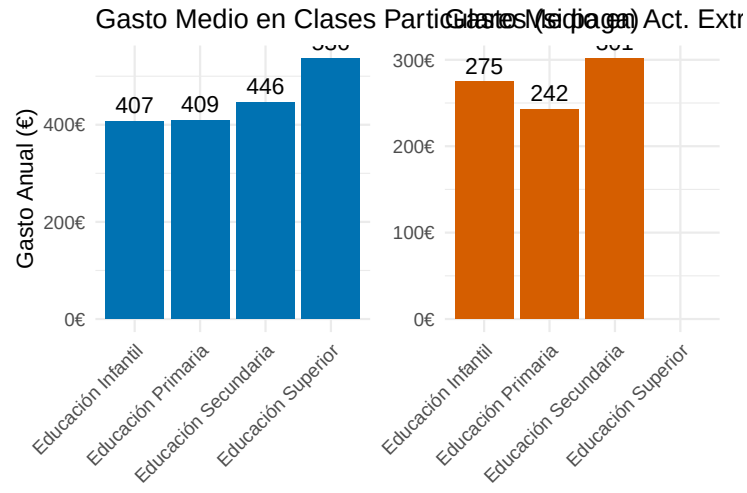
Table 3: Análisis de Gastos Complementarios Específicos

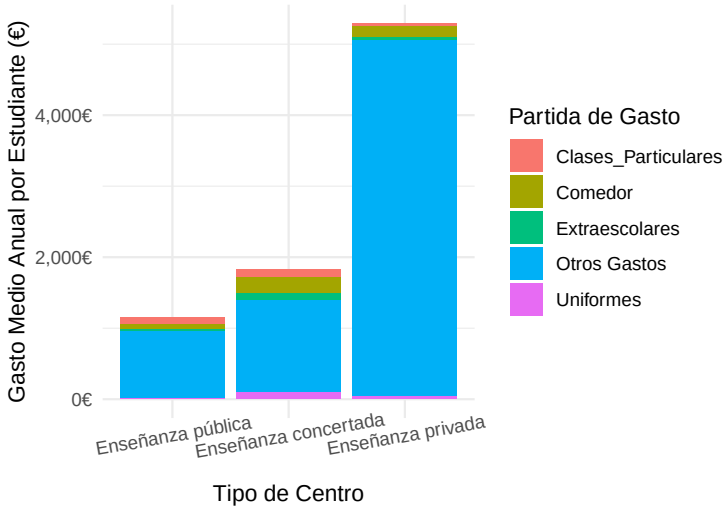
Tipo de Gasto	% Estudiantes que pagan	Gasto Medio (si paga) (€)	Gasto Medio (Total) (€)
Matrícula/Tasas	50.70	1294.65	656.45
Clases Particulares	18.73	449.11	84.13
Comedor	22.56	508.14	114.62
Act. Extraescolares	15.60	262.47	40.95
Uniformes	38.54	109.82	42.33

En esta tabla se observa que el gasto en Uniformes es el más frecuente (asumido por un 38.5% de los hogares), pero su desembolso medio es el más bajo (109.82€). En el extremo opuesto, el Comedor y las Clases Particulares, aunque son menos frecuentes (22.6% y 18.7% respectivamente), suponen el mayor coste medio para las familias que sí los pagan, con 508.14€ y 449.11€ de media.

4.1








```r
# 1. Sumar el gasto total de todas las matrículas (El
# dataset 'datos' ya tiene los NA=0 gracias a la limpieza)
total_matriculas <- sum(datos$importe_matricula_clases, na.rm = TRUE)

# 2. Sumar el gasto total educativo
total_gasto <- sum(datos$gasto_total_educacion, na.rm = TRUE)

# 3. Calcular el porcentaje
porcentaje_matriculas <- (total_matriculas/total_gasto) * 100

# 4. Mostrar el resultado
print(paste0("El gasto en matrículas (importe_matricula_clases) es un ",
    round(porcentaje_matriculas, 2), "% del gasto total."))
```

```
Cálculo Porcentaje Matrículas [1] "El gasto en matrículas (importe_matricula_clases) es un 40.07% del ga
```

```r
# 1. Contar cuántos alumnos tienen un gasto en matrícula >
# 0 (Los NAs ya fueron imputados a 0 en la limpieza)
alumnos_que_pagan <- sum(datos$importe_matricula_clases > 0,
    na.rm = TRUE)

# 2. Obtener el número total de alumnos en el dataset
total_alumnos <- nrow(datos)

# 3. Calcular el porcentaje
porcentaje_pagan <- (alumnos_que_pagan/total_alumnos) * 100

# 4. Mostrar el resultado
print(paste0("Un ", round(porcentaje_pagan, 2), "% de los alumnos ('",
    alumnos_que_pagan, "' de '", total_alumnos, "') gasta en matrículas (importe_matricula_clases)."))
```

```
Cálculo Porcentaje Alumnos que pagan Matrícula [1] "Un 50.7% de los alumnos ('2122' de '4185') gasta en
```