Análisis Exploratorio de Datos

Fast & Curious

1. Análisis Descriptivo Inicial

El objetivo del siguiente análisis es visualizar el conjunto de datos, identificar patrones y estudiar la estructura y calidad de la información.

1.1. Calidad, Estructura y Patrones de Datos

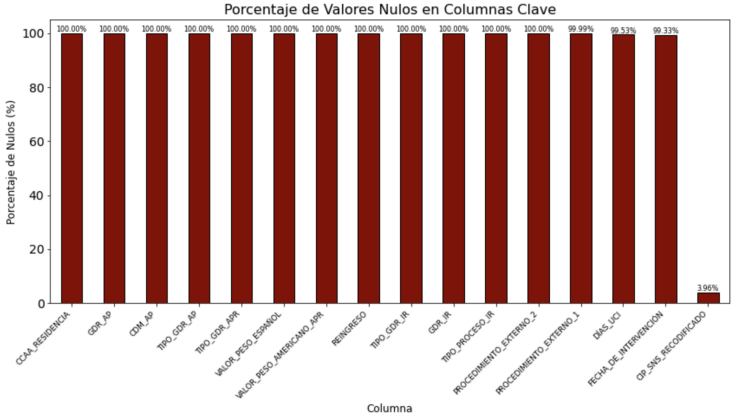
En primer lugar, se observa una gran variedad de tipos de datos dentro del conjunto además de una **alta completitud** en las **variables primarias** como *NOMBRE*, *FECHA_DE_NACIMIENTO, ESTANCIA_DÍAS* o *DIAGNÓSTICO_PRINCIPAL*, las cuales no presentan valores nulos o desconocidos ofreciendo robustez al análisis.

```
Tipos de Datos Clave (Tabla 1.1):
                 Columna Tipo Python          Tipo Lógico
0                 NOMBRE      object          Carácter/ID
1    FECHA_DE_NACIMIENTO      object                Fecha
2          ESTANCIA_DÍAS       int64              Numérico
3  DIAGNÓSTICO_PRINCIPAL      object  Categórico (Código)
```

Sin embargo, se identifican también patrones de **datos ausentes** o nulos:

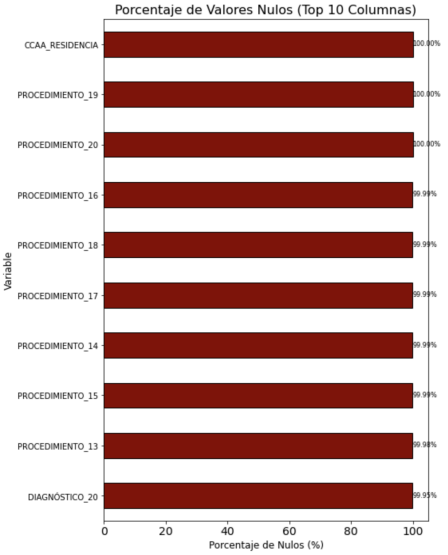
- **Variables de escasa utilidad:** desde el punto de vista de la calidad de los datos, se detecta un número elevado de columnas con **valores ausentes o nulos**. Algunas variables, como *CCAA_RESIDENCIA*, *GDR_AP*, *REINGRESO* o *DÍAS_UCI*, presentan una ausencia casi total de registros. Este hallazgo sugiere que dicha información no se recopila sistemáticamente o que resulta poco relevante en el contexto de este tipo de hospitalizaciones (por ejemplo, ingresos psiquiátricos con baja incidencia de estancias en UCI). Se propone su exclusión debido a la falta de información.

```
Resultados del Análisis de Nulos en Variables Críticas:
 CCAA_RESIDENCIA             100.00
GDR_AP                       100.00
CDM_AP                       100.00
TIPO_GDR_AP                  100.00
TIPO_GDR_APR                 100.00
VALOR_PESO_ESPAÑOL           100.00
VALOR_PESO_AMERICANO_APR     100.00
REINGRESO                    100.00
TIPO_GDR_IR                  100.00
GDR_IR                       100.00
TIPO_PROCESO_IR              100.00
PROCEDIMIENTO_EXTERNO_2      100.00
PROCEDIMIENTO_EXTERNO_1       99.99
DÍAS_UCI                      99.53
FECHA_DE_INTERVENCIÓN         99.33
CIP_SNS_RECODIFICADO           3.96
```



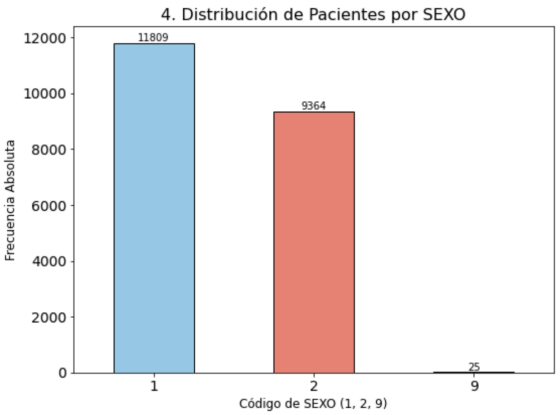
- **Patrón en cascada:** en cuanto a las variables *DIAGNÓSTICO_N* y *PROCEDIMIENTO_N*, se observa un patrón en cascada de **valores nulos**. Por ejemplo, *DIAGNÓSTICO_2* presenta 2.602 valores nulos, mientras que *DIAGNÓSTICO_3* alcanza los 6.144. Este comportamiento es coherente con la práctica clínica: la mayoría de los pacientes cuentan con uno o dos diagnósticos principales, y solo un porcentaje reducido acumula un número mayor. Por tanto, no se trata de un error de calidad de datos, sino de un **patrón estructural** esperable

```
Resultados del Análisis de Nulos (Porcentaje):
 CCAA_RESIDENCIA     100.00
PROCEDIMIENTO_19     100.00
PROCEDIMIENTO_20     100.00
PROCEDIMIENTO_16      99.99
PROCEDIMIENTO_18      99.99
PROCEDIMIENTO_17      99.99
PROCEDIMIENTO_14      99.99
PROCEDIMIENTO_15      99.99
PROCEDIMIENTO_13      99.98
DIAGNÓSTICO_20        99.95
PROCEDIMIENTO_12      99.95
PROCEDIMIENTO_11      99.92
DIAGNÓSTICO_19        99.91
PROCEDIMIENTO_10      99.89
DIAGNÓSTICO_18        99.88
```



- **Codificación categórica:** la variable *SEXO* contiene valores 1, 2 y 9. Si bien los dos primeros se corresponden presumiblemente con “Hombre” y “Mujer”, respectivamente, el valor 9 —presente en 25 registros— probablemente indica **“No especificado”** o **“Desconocido”**. Este aspecto debe tenerse en cuenta en la fase de limpieza y codificación de variables categóricas. Se propone analizar con cautela si dichos registros difieren significativamente en otras variables antes de excluirlos o recodificarlos.

```
--- Análisis de Frecuencias de SEXO ---
SEXO
1    11809
2     9364
9       25
Name: count, dtype: int64
```



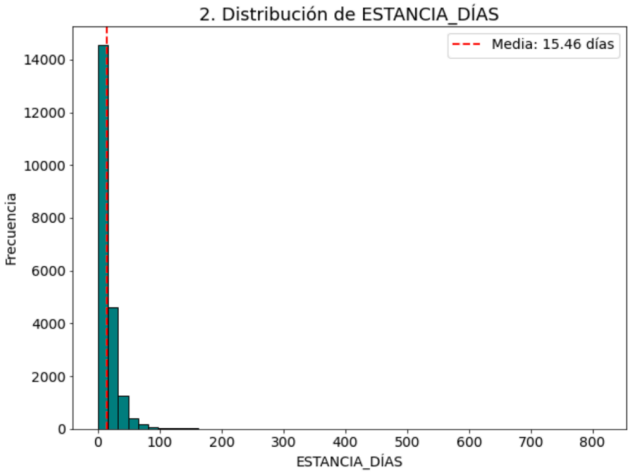
1.2. Perfil del Paciente y Estancia Hospitalaria

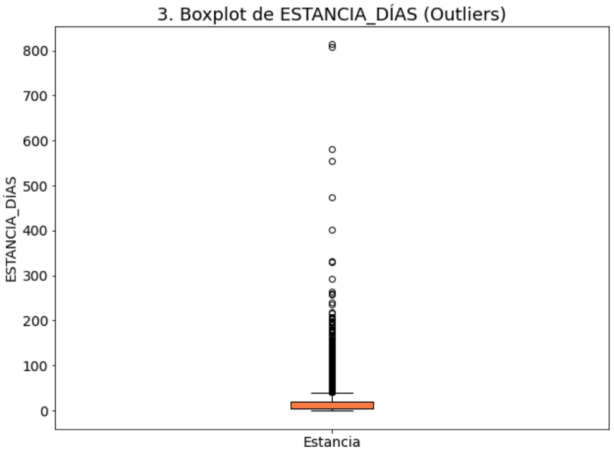
El análisis de las variables clave define el perfil de la hospitalización psiquiátrica:

- **Estadísticas de Estancia:** la variable estancia hospitalaria (*ESTANCIA_DÍAS*) revela una **media de 15,5 días** y una **mediana de 11 días**, lo que sugiere una **distribución asimétrica** hacia la derecha. Este comportamiento indica que, aunque la mayoría de los pacientes son dados de alta antes de los 11 días, existen casos con estancias significativamente prolongadas que elevan el promedio general.

```
--- 1.2 Estudio Estadístico de ESTANCIA_DÍAS ---
count    21198.00
mean        15.46
std         19.88
min          0.00
25%          5.00
50%         11.00
75%         19.00
max        814.00
Name: ESTANCIA_DÍAS, dtype: float64
```

- **Outliers y riesgo:** se destaca un valor máximo de **814 días** de hospitalización, equivalente a más de dos años continuos. Este registro constituye un **valor atípico extremo (outlier)** que podría corresponder a un caso clínico de alta complejidad o, alternativamente, a un error de registro. En cualquier caso, debe considerarse con cautela durante los análisis estadísticos posteriores.





- **Edad y correlación:** en cuanto a la edad de los pacientes, se observa una media de **43,6 años** con un rango amplio, desde recién nacidos (0 años) hasta pacientes de edad avanzada (96 años). Esto confirma que el conjunto de datos abarca un espectro completo de edades. Además, se calculó el coeficiente de correlación de Pearson entre **Estancia** y **Edad**, resultando en un 0.0794 y demostrando que **no**

existe una relación lineal entre la edad de los pacientes y la duración de su ingreso. No obstante, más adelante se mostrará que el análisis concluye en una mayor estancia media para los grupos de edad más avanzados.

1.3. Panorama Clínico y Limitaciones Geográficas

```
Top 5 COMUNIDAD_AUTÓNOMA (%
 COMUNIDAD_AUTÓNOMA
ANDALUCÍA    94.45
LA RIOJA      5.55
Name: proportion, dtype: float64

Top 5 DIAGNÓSTICO_PRINCIPAL (%
 DIAGNÓSTICO_PRINCIPAL
F20.0    21.57
F60.3     6.47
F29       4.27
F31.2     4.13
F25.0     3.86
Name: proportion, dtype: float64
```

Por un lado, se destaca que el **diagnóstico principal** más frecuente es la **esquizofrenia (*F20.0*)**, con 9.120 casos, casi el doble del segundo grupo más común, correspondiente a los **trastornos del humor** (5.219 casos). Este patrón refleja que la población analizada está fuertemente concentrada en trastornos mentales graves, lo que orienta las políticas de gestión clínica hacia esta patología.

Además, por otro lado se evidencia un **desequilibrio geográfico** considerable: la comunidad de **Andalucía** concentra la gran mayoría de los registros (94.5%), frente al 5.6% de pacientes procedentes de **La Rioja**. Esta disparidad debe destacarse como una **limitación analítica**, ya que las conclusiones extraídas representarán de manera predominante la realidad asistencial de Andalucía.

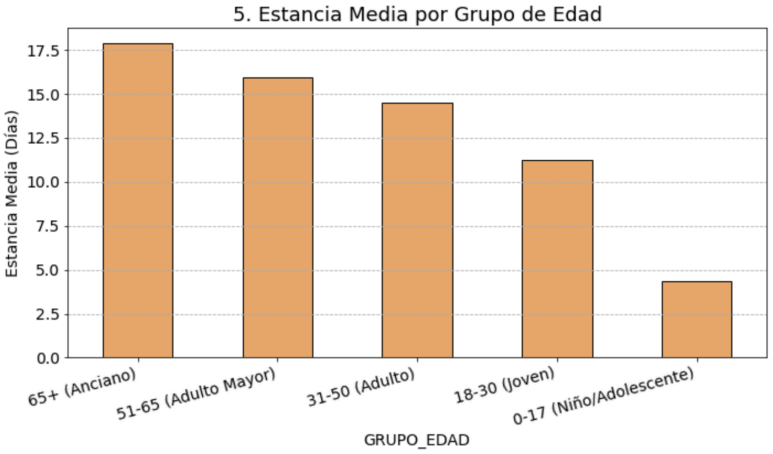
2. Ingeniería de Características

Esta sección detalla la creación de nuevas variables que permiten un mayor análisis de los datos y mayor profundidad de su estudio.

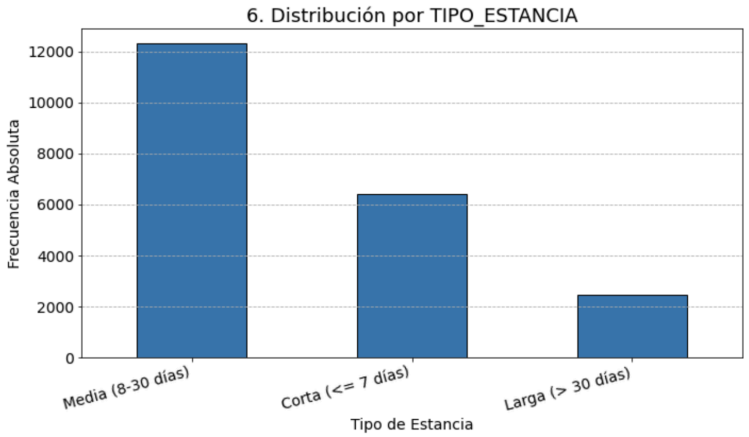
2.1. Nuevas Variables Útiles

Las siguientes variables son creadas con el objetivo de potenciar los límites de las variables originales:

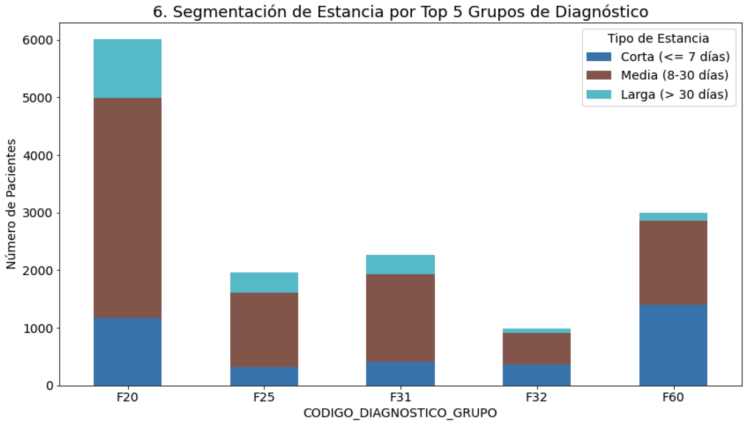
- **EDAD_CALCULADA y GRUPO_EDAD:** distinguiendo a los pacientes según las distintas franjas de edad: 0-17 (Niño/Adolescente), 18-30 (Joven), 31-50 (Adulto), 51-65 (Adulto mayor), 65+ (Anciano). Asimismo, permite investigar cómo la estancia media de hospitalización aumenta en los grupos de Adulto mayor y Anciano.



- **TIPO_ESTANCIA:** creada a partir de la variable *ESTANCIA_DÍAS* y dividida en Corta (<= 7 días), Media (8-30 días) y Larga (> 30 días) por decisión administrativa de los datos . Se observa que alrededor del 52.2% de los casos tienen una estancia media.



- **CÓDIGO_DIAGNÓSTICO_GRUPO:** se extrae el código de la variable *DIAGNÓSTICO_PRINCIPAL* con el fin de agrupar por categorías diagnósticas. De esta manera, se pueden realizar análisis más profundos como el número de pacientes según su diagnóstico y su respectiva duración.



2.2. Segmentación y Descubrimiento de Patrones

La implementación de estas nuevas variables permite obtener análisis de segmentación y descubrir nuevos patrones en los datos como se observa en las gráficas anteriores.

En primer lugar, el análisis de la **estancia media por grupo de edad** destaca en la necesidad de asignar recursos específicos para pacientes de mayor edad, donde la complejidad clínica prolonga su hospitalización, siendo este grupo el mayoritario en estancias largas.

Por otra parte, la **distribución de la estancia según el diagnóstico** proporciona una vista de la demanda de habitaciones o recursos según la patología del paciente y cuánto contribuyen desproporcionadamente a estancias largas, como el *F.20* o esquizofrenia y el *F.31* o trastorno bipolar).

2.2. Transformación y Codificación

Además, se observa la implementación de transformaciones en las variables nuevas analizadas y la codificación realizada en variables anteriores.

La variable *FECHA_DE_NACIMIENTO* requirió un **proceso de transformación** debido al formato de dos dígitos que conlleva a la asignación incorrecta del siglo a las fechas anteriores al año 2000, resultando por ejemplo en la interpretación del año 2050 en lugar de 1950. La transformación garantiza la exactitud de la variable *EDAD_CALCULADA*: se identifica primeramente si la fecha de nacimiento calculada es posterior a la fecha actual y en caso de serlo, se aplica una reducción de 100 años a dicha fecha; finalmente, la edad es calculada siendo la diferencia entre la fecha actual y la de nacimiento.

Aunque no se ha realizado **codificación** en las variables nuevas, se observa un ejemplo de este proceso en variables como *SEXO* donde los códigos 1 y 2 corresponderían a “Hombre” y “Mujer”, mientras que el código 9 se interpreta que representaría “No especificado” o “Desconocido” debido al bajo número de casos.

3. Conclusión Final

El análisis descriptivo confirma la excelente calidad de las variables clave y expone las irregularidades estructurales vinculadas a la duración de hospitalización de los pacientes. La diferencia entre la media y la mediana en la estancia confirma que el sistema se ve influido por una minoría de casos de larga estancia. La fase de ingeniería de características propone nuevas variables que permiten obtener una mayor visión de este problema, entre otros. La creación de dichas variables supone una ayuda a la planificación y el control de riesgo de los diferentes grupos según edad o diagnóstico.