

Proyecto Análisis Exploratorio de Datos

ANÁLISIS DE DATOS PARA UNA CLÍNICA VETERINARIA
ESPECIALIZADA EN REHABILITACIÓN

Borja Zaragozá Serra

Análisis exploratorio de datos en Clínica Veterinaria de Rehabilitación

El proyecto EDA escogido para el Bootcamp consiste en analizar los datos obtenidos desde la apertura de una clínica veterinaria de rehabilitación enfocada en ofrecer tratamiento a todo tipo de mascotas que cuenten con lesiones que les impiden desarrollar su día a día con normalidad. La clínica cuenta con tres años de apertura y más de 400 pacientes.

A esta clínica vienen perros de cualquier edad, ya sean mestizos o de raza, y con diferentes lesiones, ya sea relacionada con su genética, debido a accidentes o por otras causas.

¿Por qué hemos querido hacer un Análisis Exploratorio de Datos sobre esta clínica?

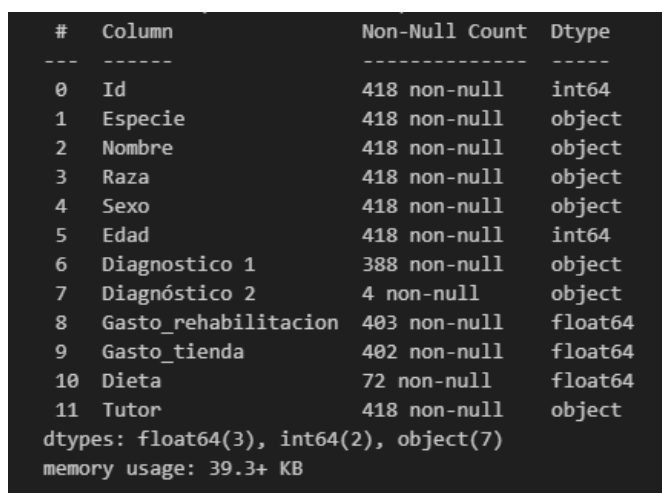
Bien es sabido que las razas puras suelen heredar patologías ligadas a la propia genética de la raza. Esta premisa, junto con un patrón observado por parte de las veterinarias ha hecho que con un objetivo principalmente científico-veterinario y en segundo lugar de asesoramiento/marketing ha llevado a hacer las siguientes hipótesis:

- ¿Qué podemos decir del paciente más común?
- ¿El tipo de lesión está relacionado con la raza?
- ¿Hay conexión entre la raza, edad y el sexo del perro?
- ¿El ingreso económico por paciente tiene que ver con la edad o con la raza?

1. LIMPIEZA DE DATOS

En primer lugar ordenamos, tratamos y limpiamos los datos. Los datos se han sacado del software veterinario que se utiliza en la clínica. Se han pasado a un Excel y se han abierto con la librería Pandas para su tratamiento.

La imagen 1 revela el contenido de los datos cuya tabla cuenta con 11 columnas, ["Especie", "Nombre", "Raza", "Sexo", "Edad", "Diagnóstico 1", "Diagnóstico 2", "Gasto en rehabilitación", "Gasto en la tienda", "Demanda de dieta" y "Tutor"].



| # | Column | Non-Null Count | Dtype |
|----|----------------------|----------------|---------|
| 0 | Id | 418 non-null | int64 |
| 1 | Especie | 418 non-null | object |
| 2 | Nombre | 418 non-null | object |
| 3 | Raza | 418 non-null | object |
| 4 | Sexo | 418 non-null | object |
| 5 | Edad | 418 non-null | int64 |
| 6 | Diagnostico 1 | 388 non-null | object |
| 7 | Diagnostico 2 | 4 non-null | object |
| 8 | Gasto_rehabilitacion | 403 non-null | float64 |
| 9 | Gasto_tienda | 402 non-null | float64 |
| 10 | Dieta | 72 non-null | float64 |
| 11 | Tutor | 418 non-null | object |

dtypes: float64(3), int64(2), object(7)
memory usage: 39.3+ KB

Imagen 1 Información sobre la distribución de los datos

El "Diagnóstico 2" se ha descartado porque la mayoría de las mascotas no cuentan con doble lesión y se ha preferido utilizar la columna de "Diagnóstico 1". Lo mismo ha ocurrido para la columna de "Dieta", donde son muy pocos los clientes que han demandado el servicio de dieta personalizada.

La columna “Diagnóstico 1” presenta valores nulos que se han corregido con la variable tipo “Acondicionamiento” la cual se percibe como más generalista y así no manipulamos y obtenemos falsos resultados sobre lesiones de mayor interés (*Imagen 2*). Para las visualizaciones en las gráficas, la lesión “Necrosis Avascular Cabeza femoral” se ha abreviado como “NAC”. El resto de lesiones son perfectamente entendibles.

```
Diagnostico 1
Hernia Discal      116
Artrosis           78
Acondicionamiento  62
Ligamento Cruzado 46
Mielopatía         37
Displasia Cadera   35
Luxación de Rotula 18
Displasia Codo     14
NAC                12
Name: count, dtype: int64
```

Imagen 2 Variables de la columna Diagnóstico 1

La columna de “Gasto en tienda” se ha tratado sustituyendo los nulos por el valor 0 que indica que no han consumido ningún producto. Esto se ha hecho para dejar la columna lista para quizás futuros estudios.

En cuanto a la columna “Gasto en rehabilitación”, los nulos se han sustituido por el valor 70, que es el coste mínimo de la primera consulta a la que se somete el paciente para evaluar su estado.

La columna “Raza” cuenta con un total de 58 razas diferentes, de las cuales solo van a ser sujetas a estudio las 7 primeras (*Imagen 3*) por interés veterinario de la clínica, que son las que más pacientes engloban bajo la misma raza. Los mestizos se van a descartar porque queremos conocer únicamente la relación entre mascotas de raza y sus lesiones más comunes.

```
df.Raza.nunique()

58

df.Raza.value_counts()

Raza
Mestizo      127
Teckel       28
Border Collie 22
Pastor Alemán 19
Labrador Retriever 19
Bulldog Francés 16
Golden Retriever 13
Bichón       13
```

Imagen 3 Razas objetivo seleccionadas para análisis bivalente

Los datos limpios se guardan en un “csv” del cual extraer los datos para los análisis tanto de una variane como de dos variantes y en algún caso de tres.

2. ANÁLISIS UNIVARIANTE

Con respecto a las columnas numéricas, en primer lugar podemos decir que el gasto medio de cliente es de 422 € mientras que la mediana es de 295 €, lo que demuestra que hay variabilidad entre los datos y que seguramente haya valores muy altos que estén subiendo la media de lo que gastan los clientes. Esto lo podemos confirmar con el coeficiente de variabilidad el cual devuelve un valor por encima del 100%.

Podemos observarlo a nivel visual (*Figura 1*) tanto en el histograma, donde encontramos que la mayor frecuencia de los datos están por debajo de 1.000 € y en la caja de dispersión encontramos valores outliers a partir de los 1.100 €.

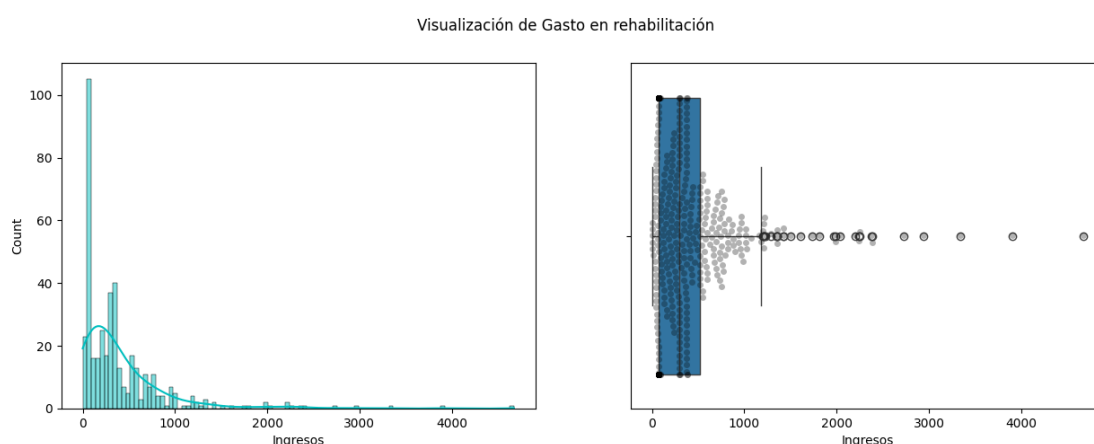


Figura 1 Visualización univariante de la variable Gasto en rehabilitación

En contraposición a la dispersión del gasto en rehabilitación, los datos de la edad (*Figura 2*) tienen valores sin outliers. La edad media junto con el 50% del valor de los datos está en torno a 9 años, encontrando un cliente al que hacer mención debido a su edad de 20 años. En el histograma se destaca la presencia mayoritaria de pacientes con edades entre los 10 y 12 años. También se aprecia un pico entre los 5 y 7,5 años.

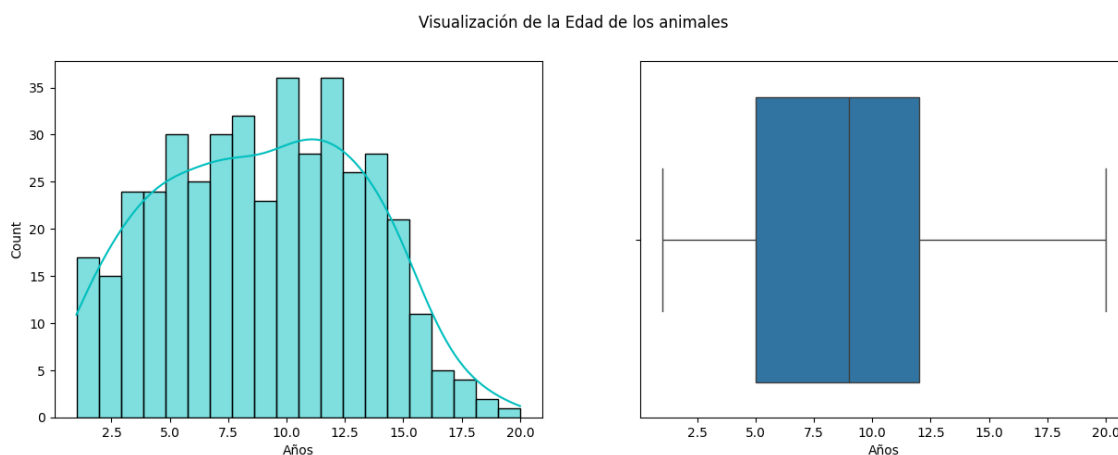


Figura 2 Visualización univariante de la variable Edad

A título informativo, un perro se considera geriátrico a partir de los 7-8 años, por lo que podríamos definir que la media de los pacientes de la clínica de rehabilitación es de edad avanzada entrada ya la vejez.

Analizando los datos para las variables categóricas (*Figura 3*), observamos que el sexo principal del tutor es de mujeres, existiendo una diferencia notable y representando alrededor del 70% de la clientela. Con respecto al sexo del perro, la proporción de machos y hembras es prácticamente igual en el conjunto de los datos, habrá que analizar posteriormente si hay relación del sexo entre razas.

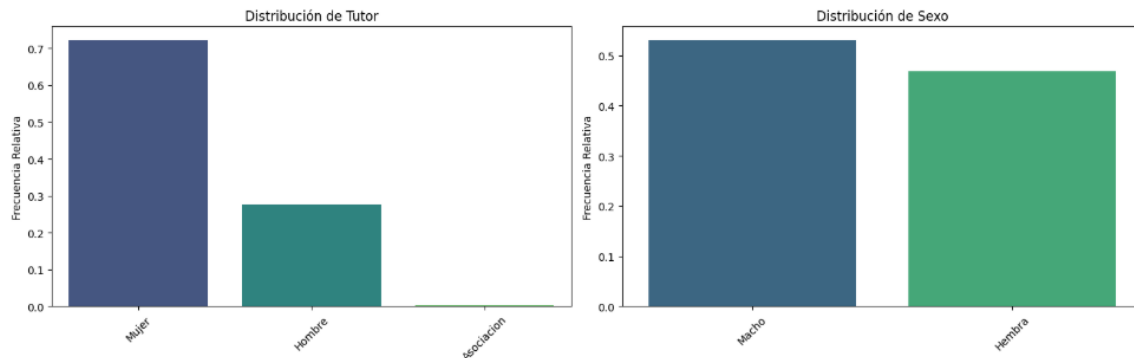


Figura 3 Visualización univariante para las variables Tutor (izq.) y Sexo (dcha.)

Acudiendo a la columna de diagnóstico (*Figura 4*), encontramos como hay una predominancia de la lesión tipo “Hernia discal”, siendo la que más se trata con un 27,75% de aparición, seguida de “Artrosis”, “Acondicionamiento” y “Ligamento cruzado” con un 18,66%, 14,83% y 11% de presencia respectivamente. El resto de lesiones suman conjuntamente alrededor del 25%.

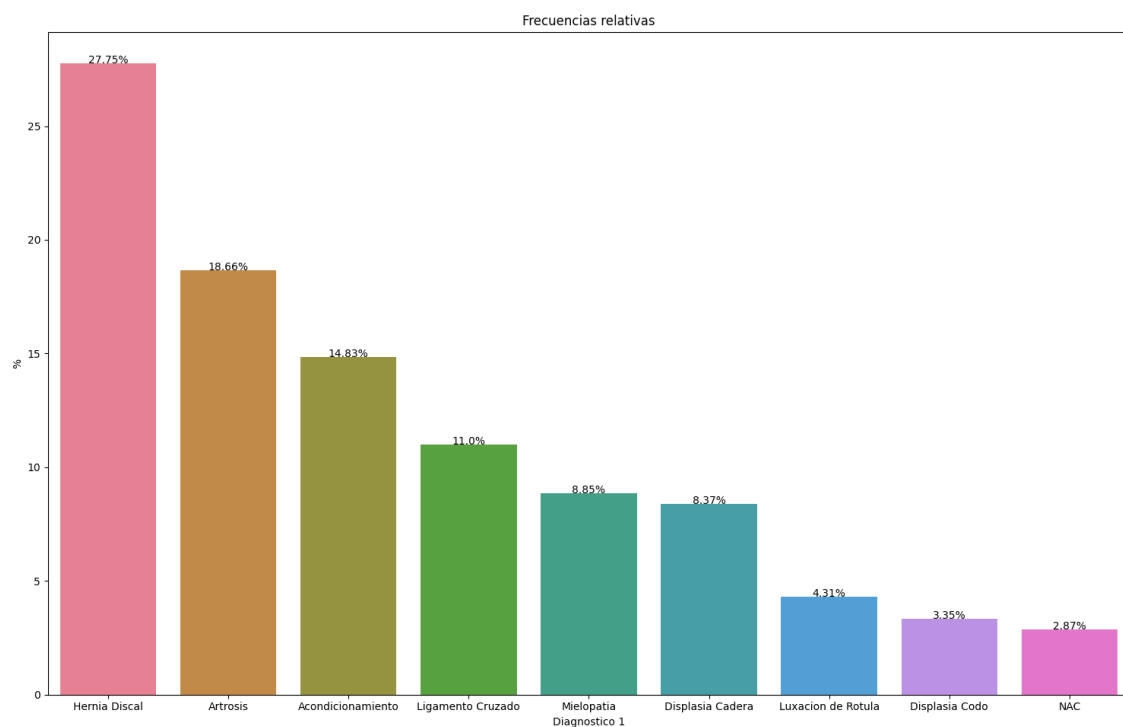


Figura 4 Frecuencias relativas de las lesiones de los pacientes

Tras la observación de las razas mayoritarias, las escogidas para el siguiente análisis y con objetivos meramente veterinarios son la raza “Teckel”, “Border Collie”, “Pastor Alemán”, “Labrador Retriever”, “Bulldog Francés”, “Golden Retriever” y “Bichón”.

3. ANÁLISIS BIVARIANTE

Se crea un nuevo *dataframe* con las 7 razas descritas previamente para hacer los análisis pertinentes y de interés para la clínica.

Empezamos observando la relación entre Raza y Edad (*Figura 5*). Respaldándonos en la estadística descriptiva, podemos apreciar como las razas más jóvenes en acudir a tratamiento son el “Border Collie”, los “Teckel”, y el “Bulldog Francés” mientras que los perros más viejos son los de raza “Labrador Retriever”, “Golden Retriever” y “Pastor Alemán” seguidos del “Bichón”.

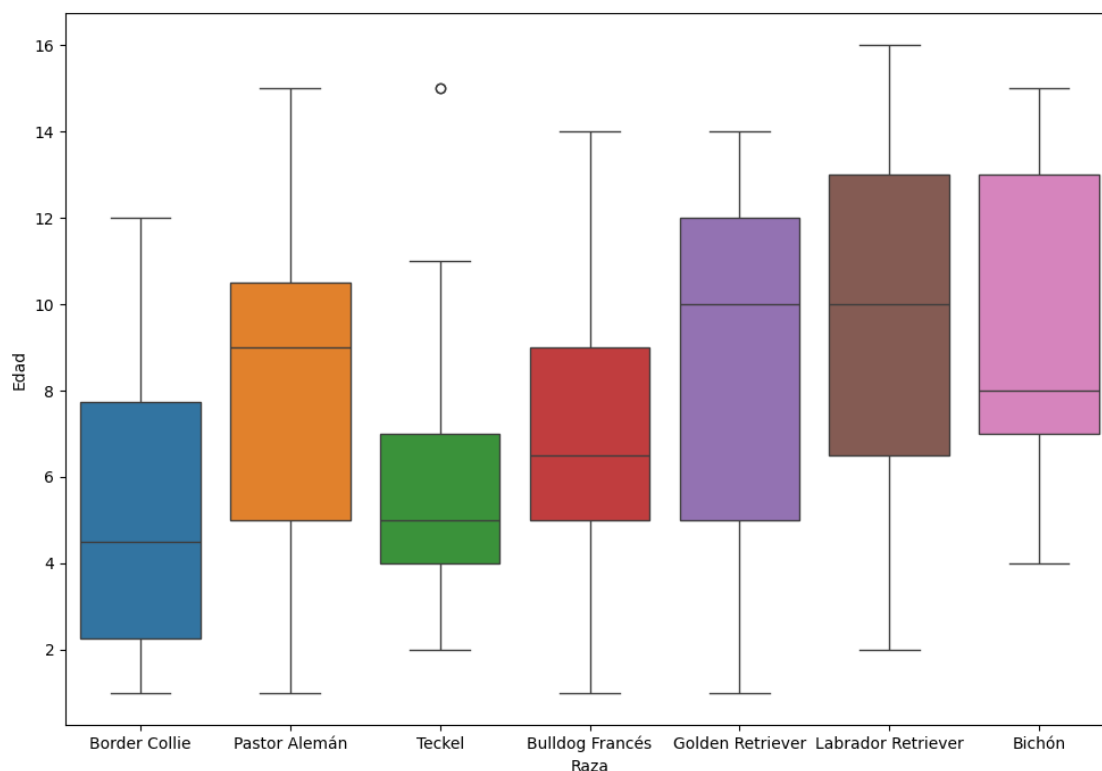


Figura 5 Visualización relación entre Raza y Edad

Acudiendo al análisis ANOVA de un factor, el valor F indica que hay cierta variabilidad entre las razas que no puede explicarse por el azar solamente y el p-valor < 0,01 significa que hay evidencia estadísticamente significativa para rechazar la hipótesis nula de que las edades medias son iguales entre las razas escogidas. Podríamos decir que la media de edad difiere significativamente entre al menos algunas razas.

En cuanto a las lesiones por raza, podemos concluir, apoyándonos de manera visual en el gráfico de barras (*Figura 6*), que los perros jóvenes padecen mayoritariamente de “Hernia Discal” siendo de un 89% y 87,5% para “Teckel” y “Bulldog Francés” respectivamente. El “Border Collie” cuenta principalmente con “Acondicionamiento” y “Ligamento Cruzado”, esto se debe a que es una raza que participa en la disciplina de *Agility* y suelen ir a la clínica a mantenimiento, siendo una lesión frecuente la de ligamento en las rodillas.

Por otro lado, los perros más viejos cuentan con lesiones más características propias de la edad como la “Artrosis” en “Labrador (21%)” y “Pastor Alemán (26%)”, “Displasia de Cadera” la cual es lesión genética para los pastores alemanes suponiendo un 36,84% y también la vemos en los “Golden Retriever” con un 38,46%. El “Ligamento Cruzado” es otra lesión predominante en “Labrador (15%)” y “Golden Retriever (30,70%)”. Aparece para el “Bichón” como lesión más frecuente la “Hernia Discal”, suponiendo más del 50% de frecuencia de lesión.

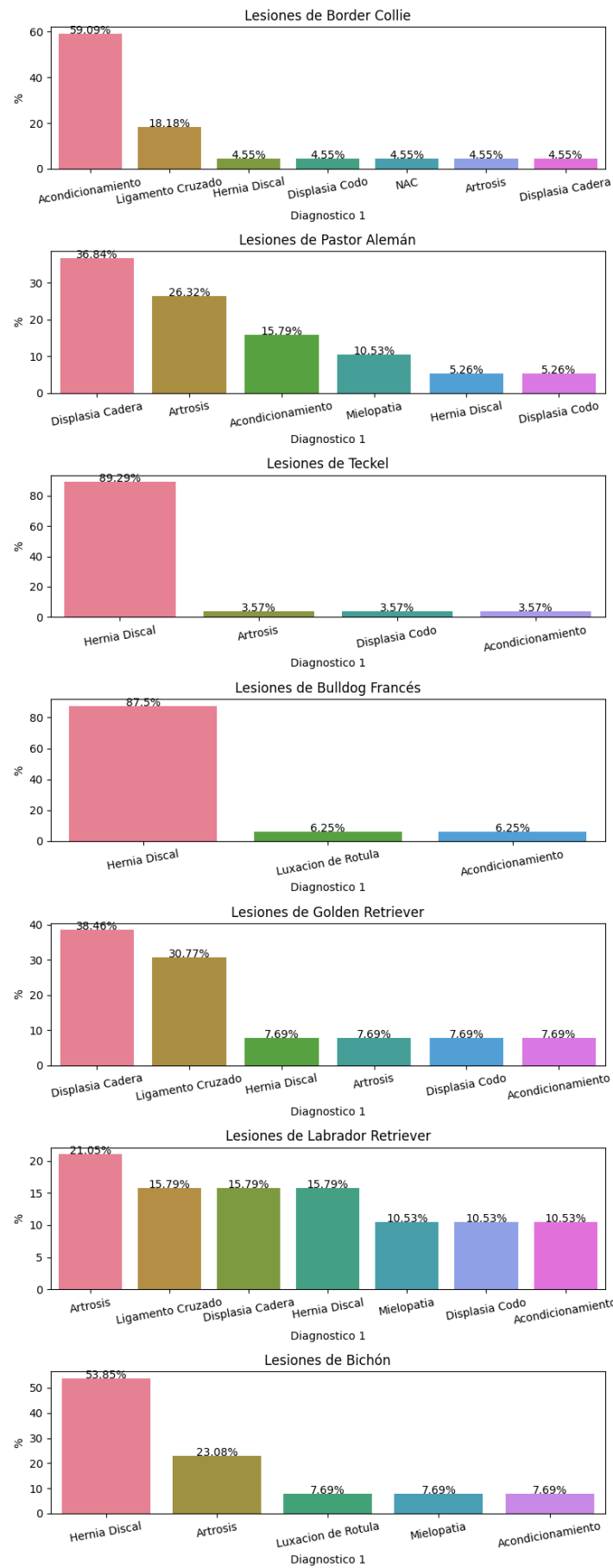


Figura 6 Relación Raza y Diagnóstico

Confirmando lo que hemos visto al analizar de manera univariante el sexo de las mascotas, no hay distinción notable entre macho y hembra como podemos apreciar visualmente al relacionar la raza, el sexo y la edad de las mascotas (*Figura 7*). Si acaso encontramos cierta notoriedad para las razas “Golden Retriever” y “Teckel” habiendo más machos que hembras y al revés para “Labrador Retriever”.

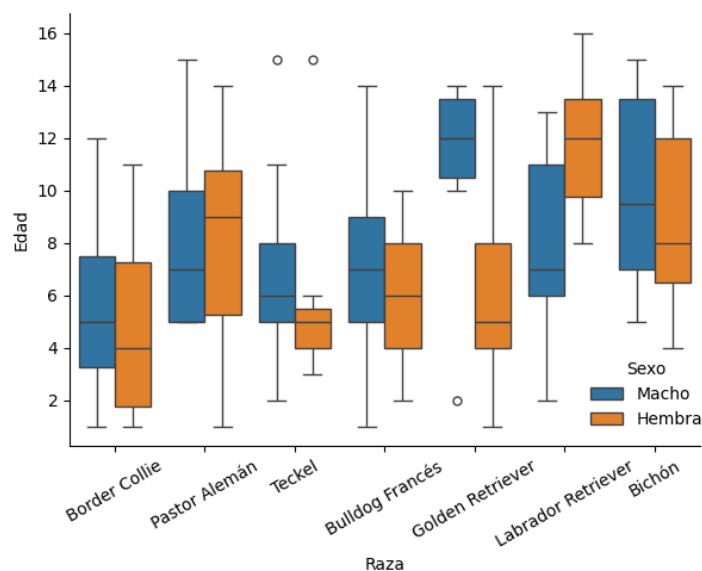


Figura 7 Relación Raza, Edad y Sexo

Vistas las relaciones para las variables biológicas, pasemos a las económicas. Analicemos entonces si existe distinción entre el gasto económico y la raza de cada cliente o si la dependencia del gasto económico se encuentra relacionada con la edad. Seguimos utilizando las razas objetivo, aunque para esta parte última también utilizaremos los datos de todos los pacientes.

Como podemos observar en la figura de cajas de dispersión (*Figura 8*), aparentemente no parece haber ninguna correlación entre la raza y el gasto económico en la clínica. Se ven, en comparación con las demás, como en los “teckel” y “Pastor Alemán” la mediana de los datos esta ligeramente por encima del resto de las razas y cabria destacar los outliers que se ven para varias razas donde sobrepasan los 1000 € llegando a alcanzar hasta 3.900 € en la raza de “Labrador”.

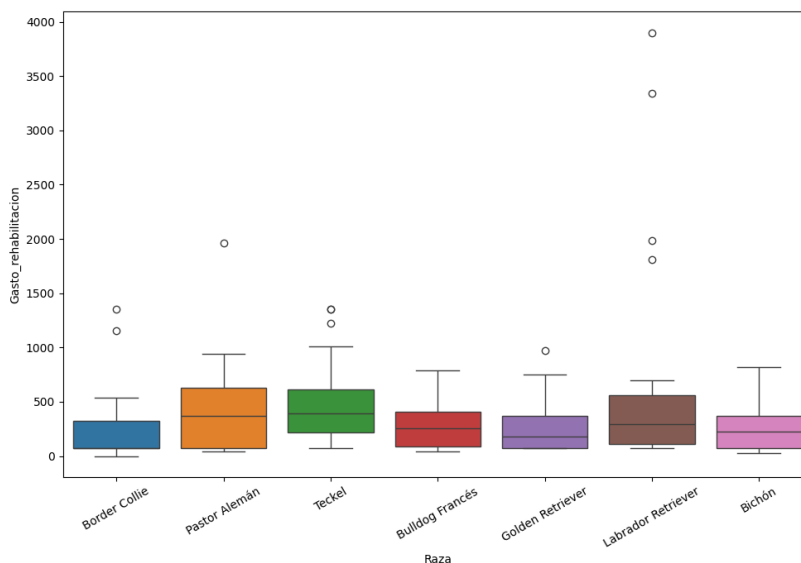


Figura 8 Relación Gasto en rehabilitación y Raza

Sin embargo, acudiendo a un análisis estadístico de ANOVA, el p-valor (0,06) está ligeramente por encima del umbral típico de significancia de 0,05. Esto significa que no se puede rechazar la hipótesis nula con seguridad, aunque se está cerca de poder hacerlo. No hay evidencia estadística significativa de que el gasto varíe por raza, pero casi. Aunque el resultado es marginal y podría considerarse tendencialmente significativo si se usara un umbral más laxo como 0,10.

Realizando el ANOVA para el conjunto total de razas, el p-valor resultante está por debajo de 0,01 lo que rechazaría la hipótesis de que el gasto es el mismo por raza, reforzando lo comentado previamente de que hay una variación entre el gasto y la raza.

Al analizar la correlación entre el gasto económico y la edad de los pacientes de la muestra objetivo (Figura 9), en la gráfica se aprecia como no existe una correlación positiva fuerte. La correlación de Pearson de 0.31 indica que existe una correlación positiva pero moderada, lo que significa que se mueven en la misma dirección aunque la relación no es muy fuerte. El p-valor de 0,0002 confirma que la relación no se debe al azar y se puede concluir que existe una relación estadísticamente significativa.

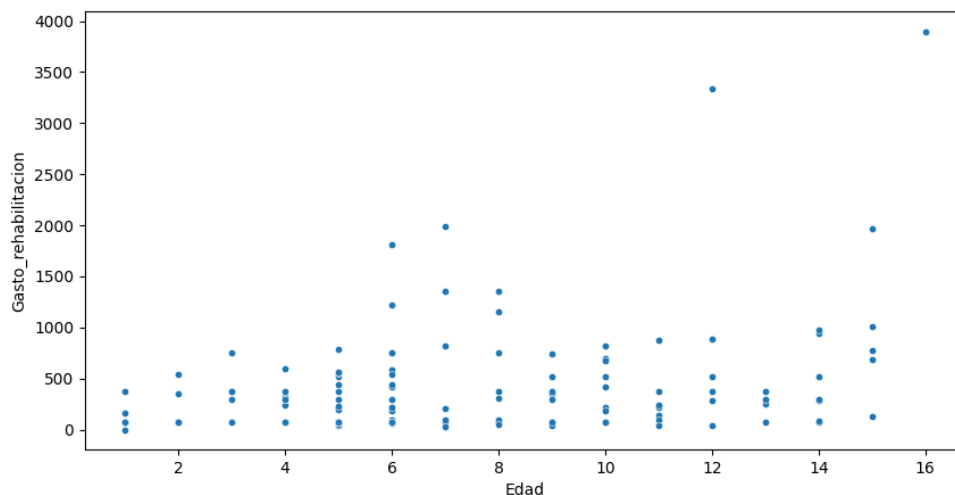


Figura 9 Correlación Gasto en rehabilitación y Edad de razas objetivo

Calculando la correlación de Pearson para el conjunto de los pacientes totales, la correlación vuelve a aparecer positiva pero de carácter débil (0,22) con un p-valor de 5,15, lo que confirma que no hay suficiente evidencia para concluir que en el conjunto de los datos, exista una relación significativa entre las dos variables aunque hay una ligera tendencia a aumentar juntas.

4. CONCLUSIONES

- El principal cliente que acude a la clínica veterinaria tiene por tutor a una mujer, siendo la lesión más común la hernia discal y el sexo del perro es macho, aunque las frecuencias porcentuales de ambos sexos están cercanas al 50%.
- Principalmente acuden perros mestizos, siendo la raza “Teckel” la mayoritaria seguida de “Border Collie” y “Pastor Alemán”.
- Las razas objetivo de menor edad tienen muy marcada la lesión de “Hernia discal”, tanto para la raza “Teckel” como para el “Bulldog francés”, representando en ambas porcentajes muy cercanos al 90%. La raza “Bichón” con un 53% de representación también por “Hernia discal”.
- Entre los perros más viejos, encontramos lesiones más relacionadas con la edad, sobrepeso, actividad, como es el caso de “Artrosis”, “Displasia de Cadera” y “Ligamento Cruzado” en razas como “Labrador Retriever”, “Golden Retriever” y “Pastor Alemán”. Ésta última raza sufre de “Displasia de Cadera” por advenimiento genético.
- El sexo no parece ser una variable importante, mientras que la raza y la edad sí parecen estar ligadas entre las razas objetivo.
- Los ingresos por paciente aparecen estar más relacionados con respecto a la raza que con la edad, aunque ésta última presenta una correlación positiva pero moderada.