Exploration Report (Reporte)

- **Primeros descubrimientos**
    - Algunos atributos son los números de identificación.
        - **Número del animal**
        - **Nombre del grupo**
        - **Nombre(s) de selección de animal)**
    - Uno de los atributos es de clasificación.
        - **Estado de la reproducción**
    - Muchos de los atributos tienen observaciones con valores nulos > 50% de las observaciones totales.
    - Solo hay un atributo que podría ser considerado como variable dependiente (Y).
        - **Días en ordeño**

- **Hipótesis inicial y su impacto en el proyecto**
    - No existen correlaciones lo suficientemente fuertes para ser relevantes para el modelo, ni un mínimo de 5 atributos y un target que sirvan para el objetivo del proyecto.
        - Puede ser posible que este .csv no sea relevante para el proyecto y deba ser descartado.

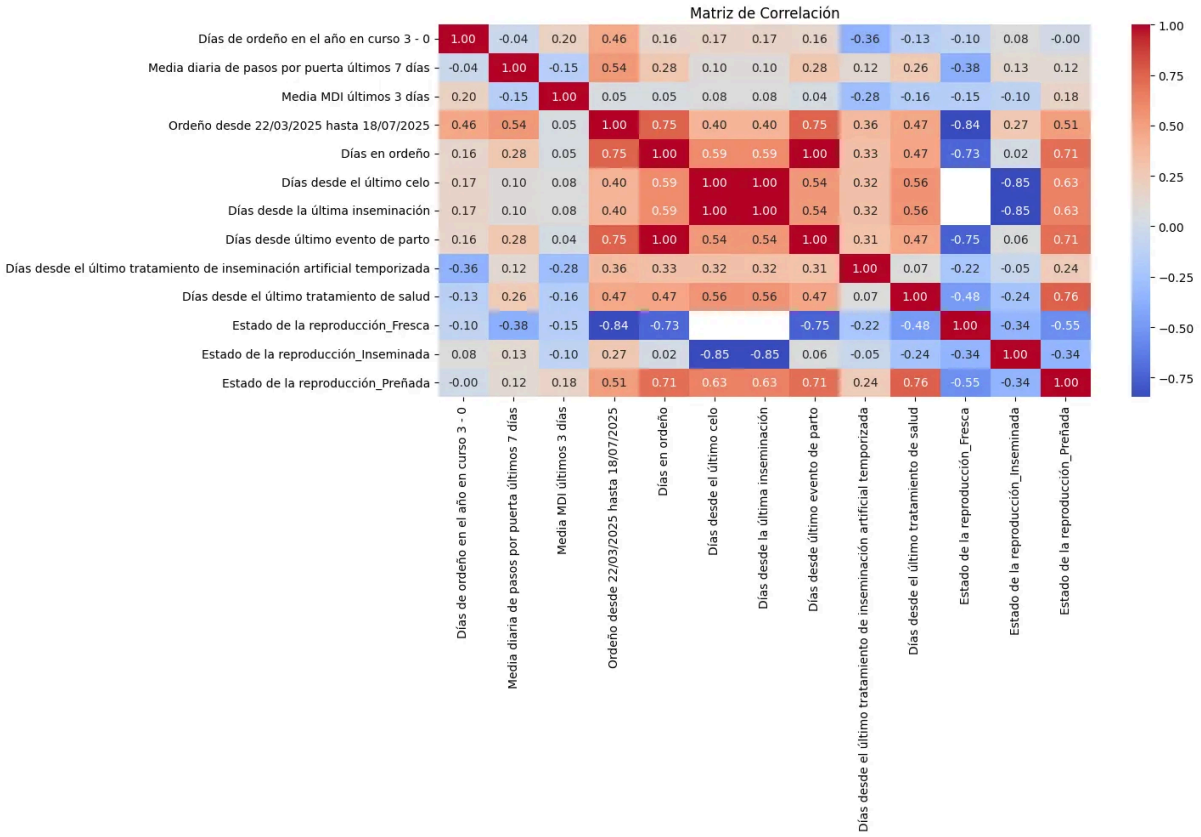
- **Gráficas y figuras**

Tipos de datos de todas las features

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 34 entries, 0 to 33
Data columns (total 21 columns):
 #   Column                                                                Non-Null Count  Dtype
---  ------                                                                --------------  -----
 0   #Pezones no encontrados en último ordeño                              3 non-null      float64
 1   Días de ordeño en el año en curso 3 - 0                               22 non-null     float64
 2   Media diaria de pasos por puerta últimos 7 días                       34 non-null     float64
 3   Media MDI últimos 3 días                                              33 non-null     float64
 4   Ordeño desde 22/03/2025 hasta 18/07/2025                              34 non-null     int64
 5   Días en ordeño                                                        33 non-null     float64
 6   Días preñada                                                          12 non-null     float64
 7   Días desde el último celo                                             19 non-null     float64
 8   Días desde el último control de gest.                                 15 non-null     float64
 9   Días desde el último secado                                           0 non-null      float64
 10  Días desde la última inseminación                                     19 non-null     float64
 11  Días desde la última inseminación fecundante                          13 non-null     float64
 12  Días desde próximo celo                                               6 non-null      float64
 13  Días Desde Último Aborto                                              6 non-null      float64
 14  Días desde último evento de parto                                     32 non-null     float64
 15  Días desde el último tratamiento de inseminación artificial temporizada  19 non-null  object
 16  Días desde el último tratamiento de salud                             34 non-null     int64
 17  Días desde el último tratamiento de vacunación                        4 non-null      object
 18  Estado de la reproducción_Fresca                                      34 non-null     bool
 19  Estado de la reproducción_Inseminada                                  34 non-null     bool
 20  Estado de la reproducción_Preñada                                     34 non-null     bool
dtypes: bool(3), float64(14), int64(2), object(2)
memory usage: 5.0+ KB
```

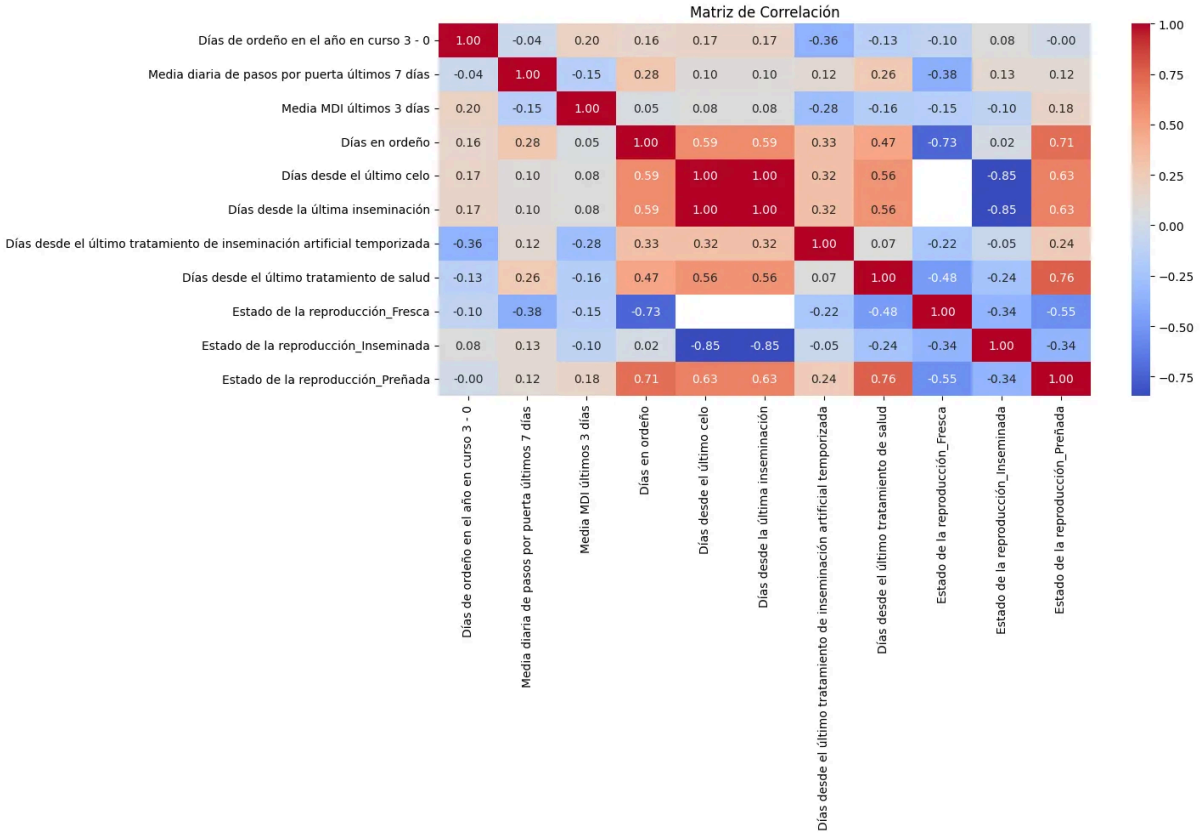
En esta tabla se puede ver el tipo de dato que es cada atributo y el número de valores nulos que contienen. 8 de los 19 atributos tienen > 50% de las observaciones nulas por lo tanto no serán evaluadas ya que no se cuenta con las observaciones suficientes para utilizar esos atributos.

Mapa de correlaciones:



Se realizó una matriz de correlación para saber cuál era la relación entre cada uno de los atributos, la dirección de la relación y la magnitud. En este .csv viendo los diferentes atributos que hay se determinó de manera temporal que la variable dependiente será: **Días en ordeño**. Algunas de las correlaciones significativas son:

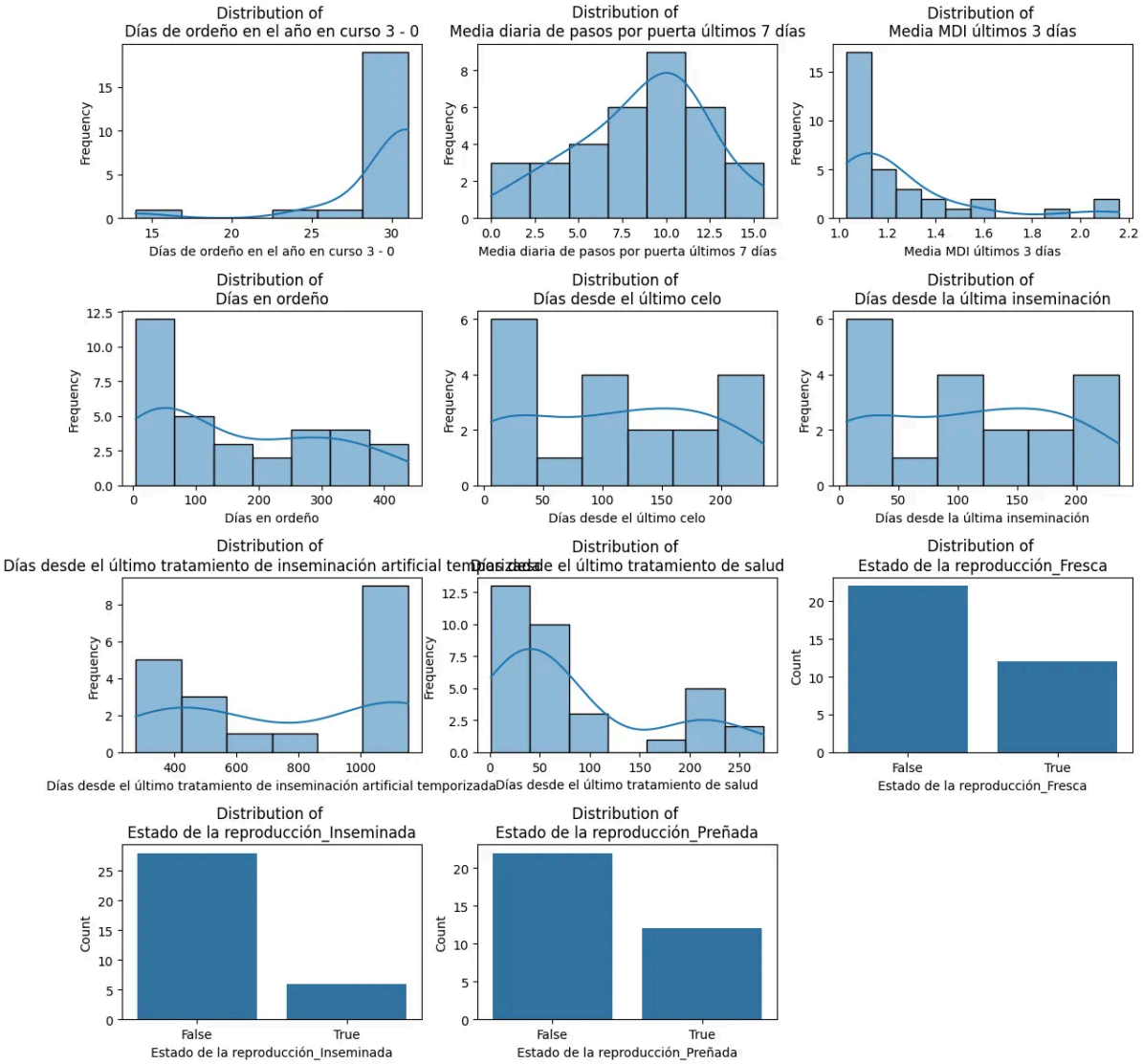
- **Días desde último evento de parto** = 1.00: Debido a que se están contando los mismo días que el target, este atributo no puede ser considerado para el modelo.
- **Ordeño desde 22/03/2025 hasta 18/07/2025** = 0.75: Debido a que se están contando los días de ordeño pero en un periodo determinado, este atributo no puede ser considerado para el modelo.
- **Estado de la reproducción_Fresca** = -0.73: Debido a que esto es una dummy de un atributo de clasificación no puede ser considerado para el modelo.
- **Estado de la reproducción_Preñada** = 0.71: Debido a que esto es una dummy de un atributo de clasificación no puede ser considerado para el modelo.



Quitando los 4 atributos mencionados previamente, las nuevas correlaciones significativas son:

- **Días desde el último celo** = 0.59: Debido a que tiene una correlación con mayor magnitud con otro atributo (**Días desde la última inseminación**), se debe eliminar una de ellas para evitar multicolinealidad.
- **Días desde la última inseminación** = 0.59: Debido a que tiene una correlación con mayor magnitud con otro atributo (**Días desde el último celo**), se debe eliminar una de ellas para evitar multicolinealidad.
- **Días desde el último tratamiento de salud** = 0.47: Debido a que tiene una correlación con mayor magnitud con otros atributos (**Días desde el último celo, Días desde la última inseminación**), se debe eliminar una de ellas para evitar multicolinealidad.
- **Días desde el último tratamiento de inseminación artificial temporizada** = 0.33: Debido a que tiene una correlación con mayor magnitud con otro atributo (**Días de ordeño en el año en curso 3 - 0**), se debe eliminar una de ellas para evitar multicolinealidad.
- **Media diaria de pasos por puerta últimos 7 días** = 0.28: Este atributo cuenta con una magnitud en la correlación con el target baja, lo que significa que no representa mucho del comportamiento del target.

Histograma de todas las columnas para la visualización de distribuciones



- El histograma de Media diaria de pasos por puerta últimos 7 días es el atributo que tiene la distribución más parecida a una distribución normal, sin embargo los extremos de igual forma son muy comunes lo que significa que la media no es tan común como se esperaría.
- El histograma de la Media MDI últimos 3 días tiene una distribución de schewed to the right lo que significa que la media de los datos se encuentra desplazada hacia la izquierda y que existen algunos valores muy desplazados hacia la derecha que mueven la tendencia de los valores.
- Los histogramas en general tienen distribuciones particulares esto por una parte es debido a que para la mayoría de ellos se tuvieron que eliminar y por eso muchos valores tienen tendencias separadas, por lo que es difícil realmente saber la tendencia real.