Exploration Report (Inventario)

- **Primeros descubrimientos**
    - Los datos tienen algunas variables categórica, en este caso siendo:
        - Nombre de Grupo
        - Número(s) de selección de animal
        - Estado de la reproducción
    - Por lo que todas las columnas dicen lo mismo pues el dataset incluye las mismas categorías de las vacas.
    - Algo malo es que la mayoría de las columnas están vacías o la mayoría de los elementos, el problema de eso es que muchas de las columnas quedan inservibles porque no aportan nada de información

- **Hipótesis inicial y su impacto en el proyecto**
    - El momento del secado de las vacas es una parte importante de la producción, por ello importa cuánto tiempo tardará en volver a producir leche, sin embargo si no encontramos relación con otro .csv puede volverse irrelevante.
        - Este dataset puede ser irrelevante ante la información personal de las vacas.

- **Gráficas y figuras**

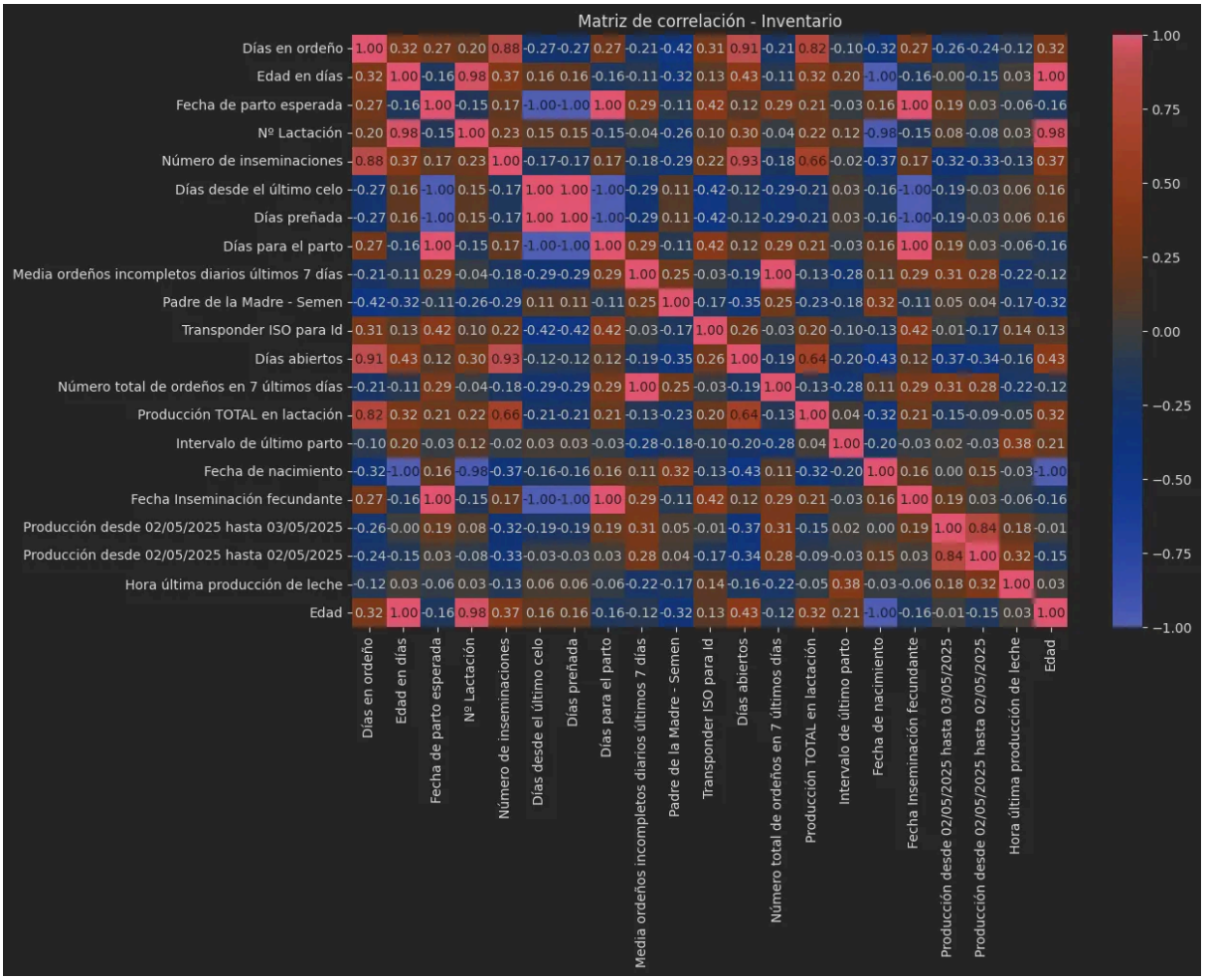
Tipos de datos de todas las features

#	Column	Non-Null Count	Dtype
0	Número del animal	33 non-null	int64
1	Nombre del grupo	33 non-null	object
2	Número(s) de selección de animal	33 non-null	int64
3	Estado de la reproducción	33 non-null	object
4	Días en ordeño	33 non-null	int64
5	Edad en días	33 non-null	int64
6	% de concentrado consumido ayer	33 non-null	int64
7	Fecha de parto esperada	33 non-null	object
8	N° Lactación	33 non-null	int64
9	Número de inseminaciones	33 non-null	int64
10	Días desde el último celo	33 non-null	int64
11	Días preñada	33 non-null	int64
12	Días para el parto	33 non-null	int64
13	Fecha de parto esperada.1	33 non-null	object
14	Edad (a:mm)	33 non-null	object
15	Media ordeños incompletos diarios últimos 7 días	33 non-null	float64
16	Padre de la Madre - Semen	25 non-null	float64
17	Transponder ISO para Id	33 non-null	int64
18	Media ordeños incompletos diarios últimas 48 h.	33 non-null	float64
19	Días abiertos	33 non-null	int64
20	Número total de ordeños en 7 últimos días	33 non-null	float64
21	Producción TOTAL en lactación	33 non-null	float64
22	Intervalo de último parto	21 non-null	float64
23	Edad (a:mm).1	33 non-null	object
24	Número de inseminaciones.1	33 non-null	int64
25	Fecha de nacimiento	33 non-null	object
26	Fecha Inseminación fecundante	33 non-null	object
27	Producción desde 02/05/2025 hasta 03/05/2025	25 non-null	float64
28	Producción desde 02/05/2025 hasta 02/05/2025	25 non-null	float64
29	Hora última producción de leche	33 non-null	object

Al hacer un análisis de los elementos vimos que muchos de ellos son valores enteros o flotantes, por lo que ayudan para el análisis, además algunos son objetos, lo cual pueden ser fechas que con una transformación podrían ayudar a el análisis correcto.

En esta tabla se puede ver el tipo de dato que es cada atributo y el número de valores nulos que contienen los 29 atributos de las observaciones, sin embargo algunos de ellos son objetos por lo tanto no serán evaluados ya que no se cuenta con información suficientes para ser observados.

Mapa de correlaciones:



Se hizo una matriz de correlación con las variables y con las transformaciones necesarias obtenidas.

Producción y tiempo en ordeño

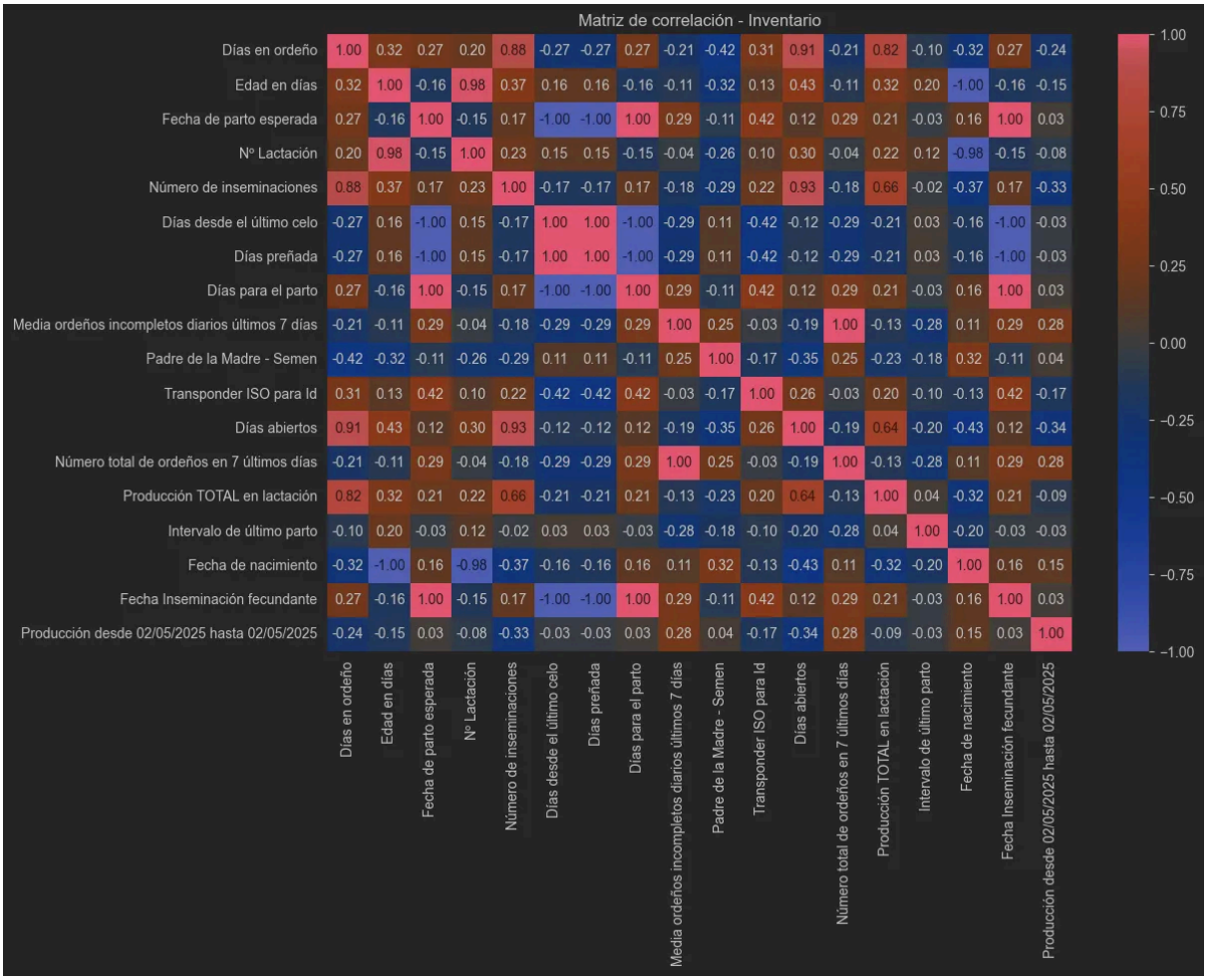
- **Producción total en lactación** tiene alta correlación con días en ordeño (≈ 0.82).
- También se correlaciona con **número de ordeños en 7 días** (≈ 0.64) y con **número de inseminaciones** (≈ 0.66), indicando que animales con más control reproductivo suelen mantener producción más alta.
- **Fecha de inseminación fecundante** y **fecha de parto esperada** (≈ 1.00): relación perfecta — coherente, pues una determina la otra.
- **Días preñada** y **días para el parto** (≈ −1.00): inversa perfecta
- **Días desde el último celo** y **días preñada** (≈ 0.29): leve relación, pues el celo antecede la concepción.

Observaciones:

- Reducir redundancia: excluir variables casi idénticas (r > 0.95) al modelar.

- Segmentar por lactación: la relación entre producción y edad podría variar según el número de partos.

Para reducir las variables redundantes se eliminaron ciertas columnas dejando unas seleccionadas, con eso se volvió a realizar la matriz de correlación:



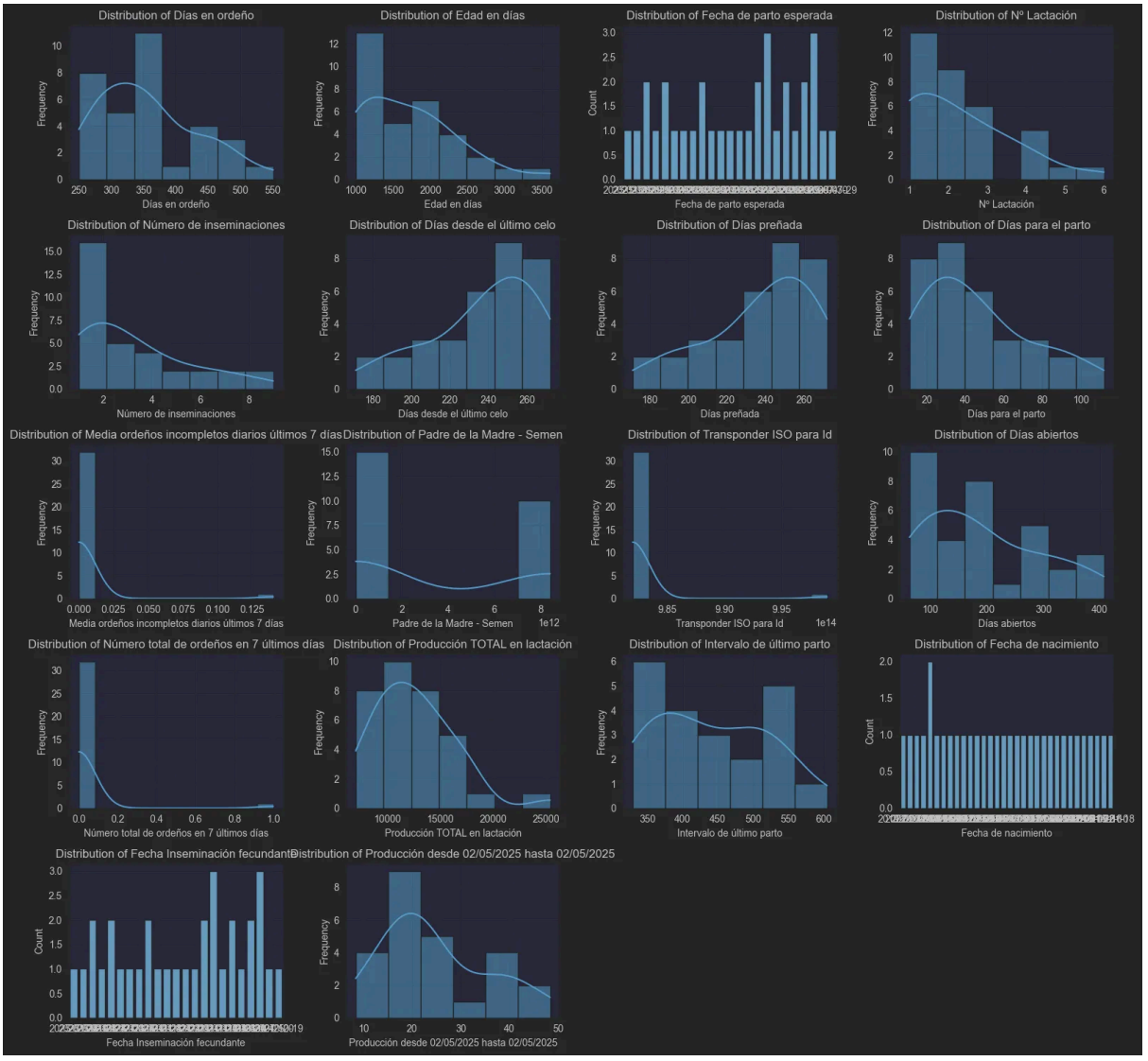
Correlaciones más relevantes

- **Días en ordeño ↔ Días abiertos**: 0.91. Las vacas con más días en ordeño también permanecen más tiempo abiertas (sin preñez).
- **Días en ordeño ↔ Producción total en lactación**: 0.82. Cuanto más tiempo están en ordeño, mayor es la producción acumulada.
- **Días en ordeño ↔ Nº de inseminaciones**: 0.88. Las vacas con más tiempo productivo suelen tener más intentos de inseminación.
- **Producción total en lactación ↔ Nº de inseminaciones**: 0.66. A mayor número de inseminaciones, tiende a observarse mayor producción total.
- **Edad en días ↔ Nº de lactación**: 0.98. Las vacas mayores tienen más ciclos de lactancia acumulados.
- **Fecha de inseminación fecundante ↔ Fecha de parto esperada**: 1.00. Correlación perfecta, ya que una determina directamente la otra.
- **Días preñada ↔ Días para el parto**: −1.00. Relación inversa perfecta: conforme avanzan los días de gestación, disminuyen los días restantes al parto.

- **Edad en días ↔ Fecha de nacimiento:** −0.96. Las vacas nacidas antes son naturalmente las más viejas.
- **Días desde el último celo ↔ Días preñada**: −1.00. Una vez preñada, el conteo de días desde el último celo se reinicia.

Histograma de todas las columnas para la visualización de distribuciones

---



Para tener un poco de información de lo que incluyen los dataset, se hizo un histograma de la distribución de las columnas.

El conjunto de histogramas refleja la **distribución de frecuencia de las variables reproductivas, productivas y fisiológicas** del hato.

Algunas presentan **asimetrías marcadas (sesgo)**, mientras que otras tienden a una forma **más normal (campana)**. Esto es importante para identificar valores atípicos y planificar modelos predictivos.

Variables con distribuciones normales o casi simétricas

- **Días en ordeño**: distribución concentrada entre 200 y 450 días, con ligera tendencia a la derecha. La mayoría de las vacas están en su periodo medio de lactancia.
- **Edad en días**: muestra una forma suavemente decreciente, lo que sugiere predominio de vacas adultas jóvenes (2–6 años).
- **Producción total en lactación**: campana asimétrica hacia la derecha (sesgo positivo). La mayoría produce entre 10 000 y 20 000 L, con pocos casos excepcionales de alta producción.
- **Número de ordeños en 7 días**: tiende a concentrarse entre 0.6 y 1.0, indicando regularidad en la rutina de ordeño.

Variables con sesgo positivo

- **Número de inseminaciones:** la mayoría de vacas tiene 1–3 inseminaciones, pero unas pocas llegan a 6 o más → posibles casos de fertilidad baja.
- **Días desde el último celo:** incremento gradual hasta ~260 días, reflejando animales ya preñados o próximos a parto.
- **Días preñada**: crece hacia los 250 días, como se espera en un grupo con diferentes etapas de gestación.
- **Días abiertos**: la mayoría entre 50 y 150, pero algunos casos llegan a 400 → vacas que tardan mucho en concebir.
- **Intervalo de último parto:** forma irregular con algunos casos muy altos (>500 días), indicando períodos de descanso prolongado o problemas reproductivos.

Variables categóricas numéricas con patrones atípicos

- **Fecha de parto esperada, fecha de inseminación fecundante y fecha de nacimiento**: muestran barras separadas (distribución discreta), reflejando que los datos son fechas convertidas a valores numéricos, no continuos.
- **Padre de la madre – semen y Transponder ISO:** distribuciones sesgadas o agrupadas en valores únicos —identificadores más que métricas útiles para análisis estadístico.
- **Media ordeños incompletos diarios últimos 7 días**: muy concentrada cerca de 0 → indica ordeños completos casi siempre.