Exploration Report (Registro Ordeño)

- **Primeros descubrimiento**
    - Se tienen registros de ordeño de 44 vacas.
        - Las vacas analizadas son del corral 6 del CAETEC.
    - Cada archivo `.csv` de ordeño tiene información de solo una vaca. Para el entrenamiento del modelo y creación del dataset, se recomienda hacer un merge de todos ellos para trabajar con todo el volumen de datos y encontrar una mayor correlación entre variables.
    - Varias de las columnas tienen registros vacíos, por lo que resulta SUMAMENTE importante aplicar técnicas de imputación.
        - Como los valores y registros en la parte de los cuartos son independientes, optaría por usar técnicas de imputación como llenado de 0´s para estas columnas.
        - Para otras variables que tiene valores categóricos optaría por usar técnicas como la creación de dummies en lo que se descartan o no.
        - Usar técnicas como one-hot encoding para las variables categóricas como Destino de la leche.

- **Hipótesis inicial y su impacto en el proyecto**
    - Los reportes de ordeño tendrán un peso en el proyecto, ya que el conocimiento de las lactancias, y su producción, son variables que podrán determinar un gran peso en nuestro modelo.

- **Gráficas y figuras**

Tipos de datos de todas las features

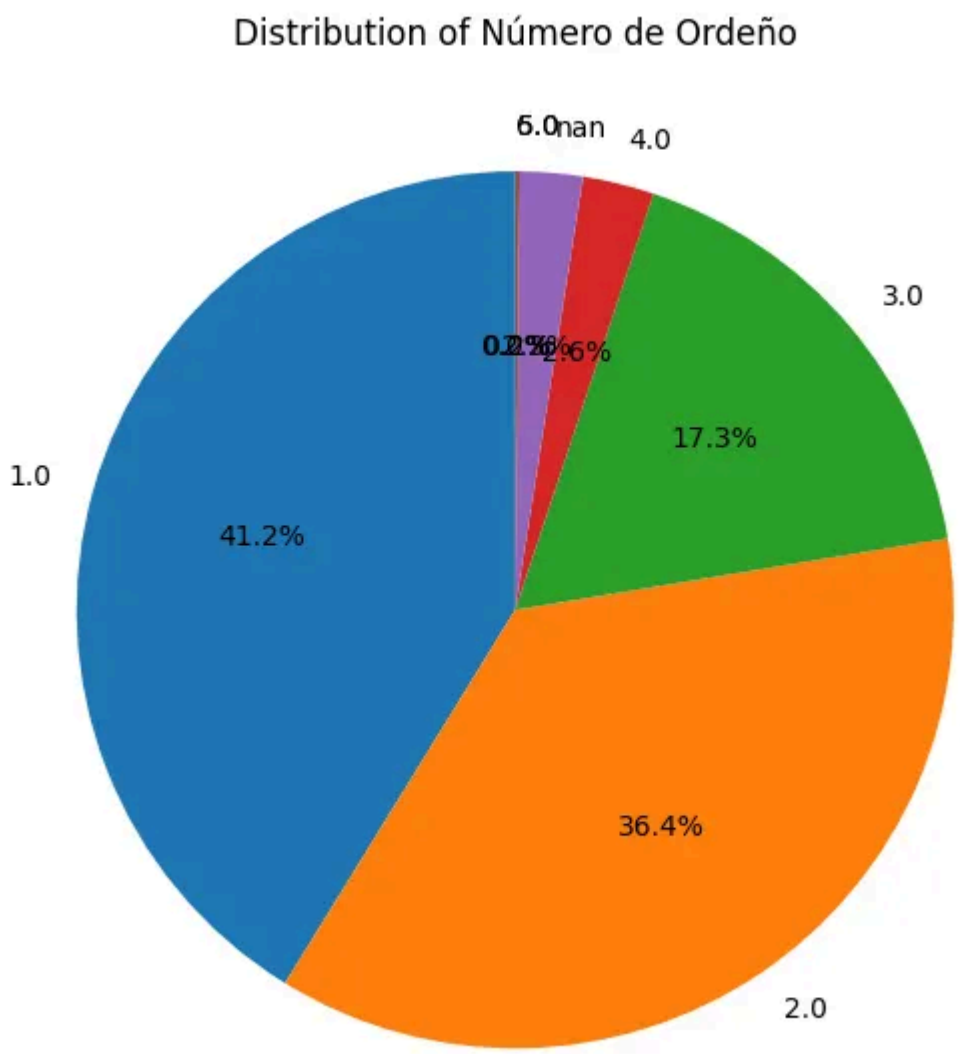
Acción	object
Duración (mm:ss)	object
Producción (kg)	float64
Número de ordeño	float64
Patada	object
Incompleto	object
Pezones no encontrados	object
Ubre	float64
Pezón	object
DI - media de flujos	float64
DD - media de flujos	float64
TI - media de flujos	float64
TT - media de flujos	float64
DI - sangre	float64
DD - sangre	float64
TI - sangre	float64
TT - sangre	float64
DI - conductividad	float64
DD - conductividad	float64
TI - conductividad	float64
TT - conductividad	float64
EO/PO	object
Destino Leche	object
DI - flujos	float64
DD - flujos	float64
TI - flujos	float64
TT - flujos	float64
DI - producciones	float64
DD - producciones	float64
TI - producciones	float64
TT - producciones	float64

En la imagen de arriba vemos todos los tipos de datos que hay en nuestros registros. Viendo que la mayoría de los datos son de tipo `float64` . Donde, adicionado a esto, tenemos datos de tipo `Object` . Siendo:

- Patada, Incompleto y Pezones son variables con registros de los cuartos. Pueden ser útiles así que no vale la pena descartarlas.
- Duración y hora de inicio son fechas: Estos registros son de suma importancia para conocer cuando inicio y concluyo cada lactancia.

Distribución de datos - Histogramas y Gráficas de Pie

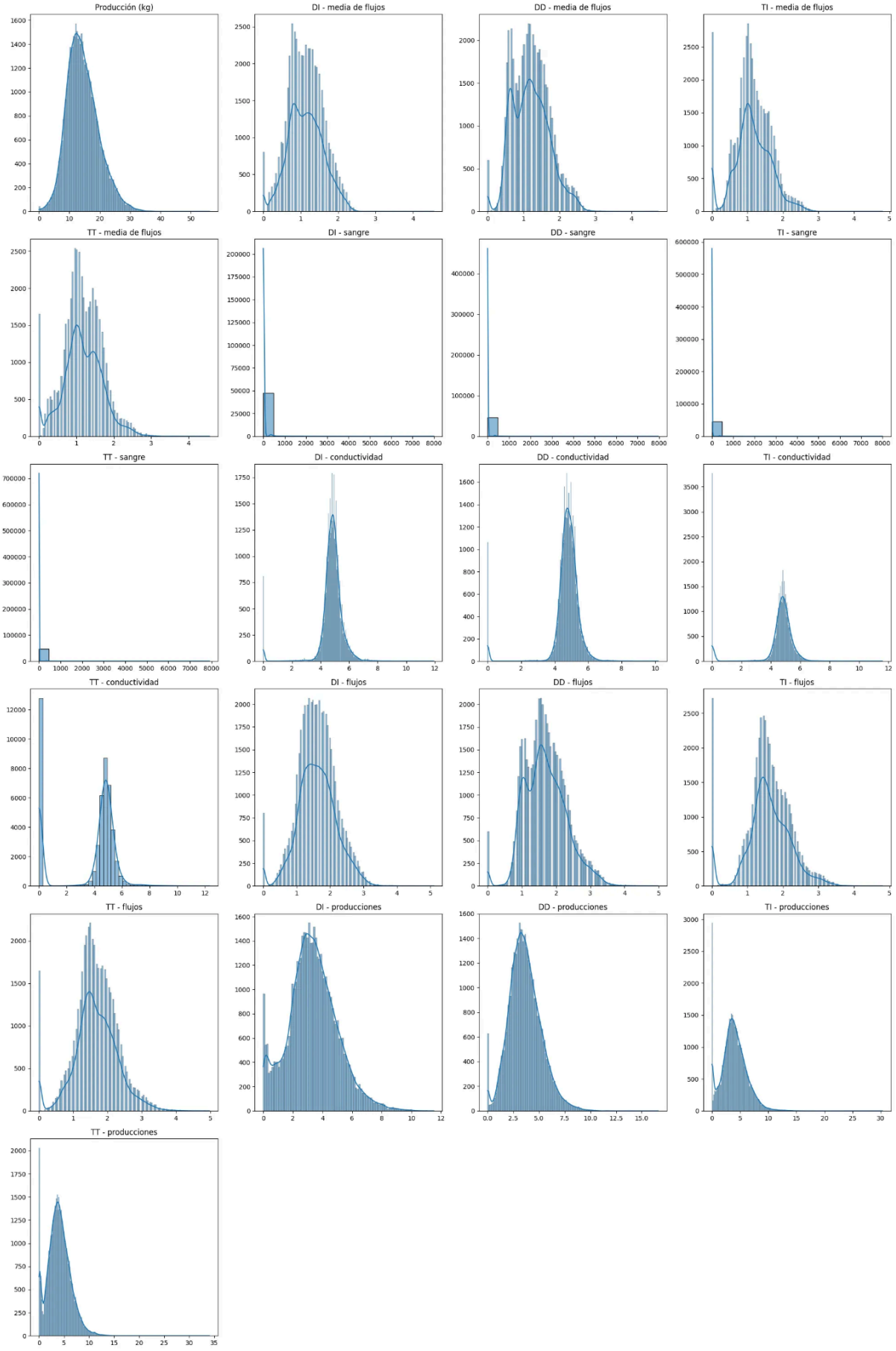
Para variables categóricas, como Número de Ordeño, decidimos usar otro tipo de interpretación de datos como lo puede ser una gráfica de pastel. En este caso, viendo la gráfica podemos rescatar lo siguiente:



Para Num de Ordeño:

- Las vacas normalmente solo se ordeñan una vez al día (41.2%)
- Aproximadamente un 3% de los registros son de vacas que van a ordeñarse de hasta 4-6 veces. Estos valores deben tratarse como outliers.
- Casi no hay registros nulos, pero resulta interesante saber que estos días las vacas no se ordeñaron.
    - La imputación de datos puede marcarse como 0.
    - Investigar por qué no se ordeñaron estos días.

Ahora, en la parte de los histogramas, obtuvimos lo siguiente:

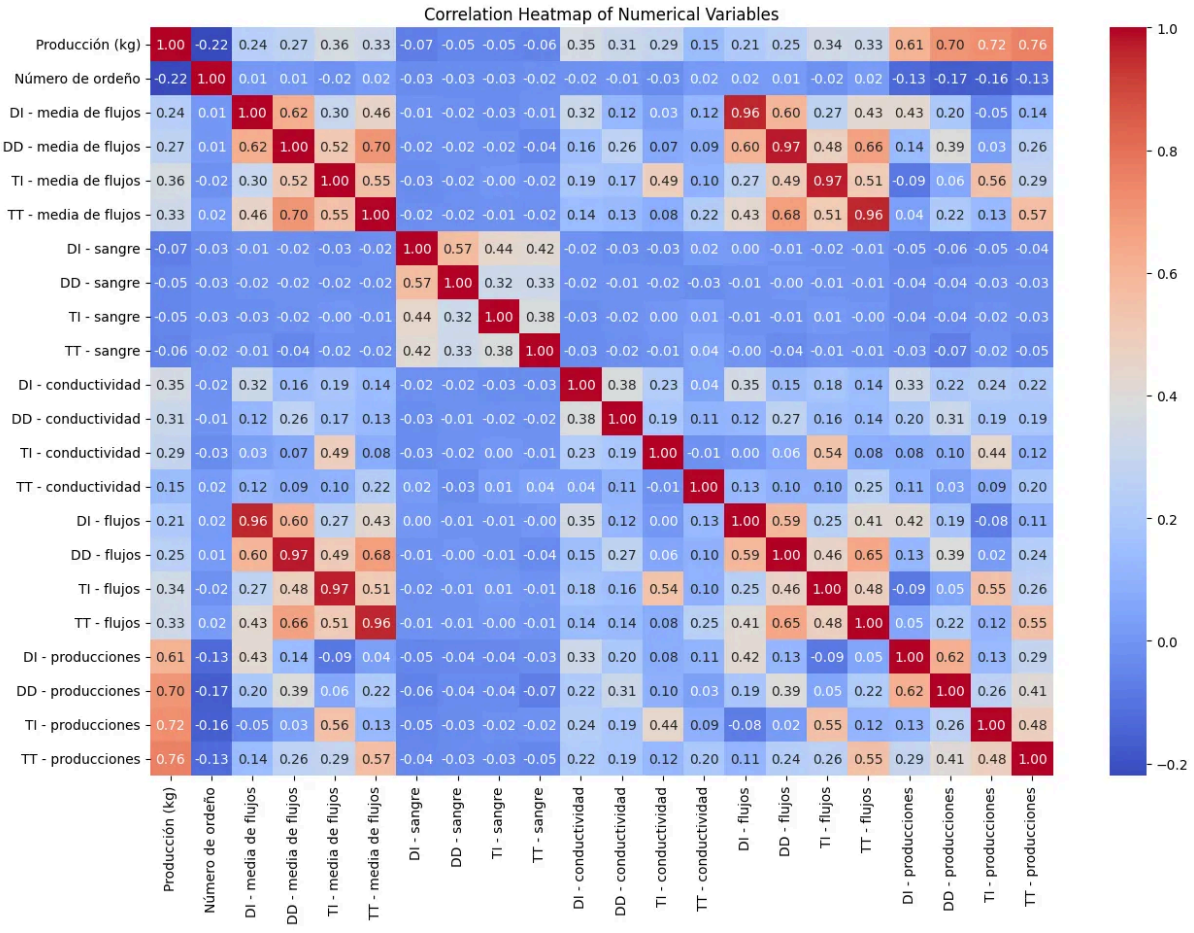


Viendo los histogramas podemos ver que en general todos tienen una distribución normal.

- Algunos como DI-Sangre y TI-Sangre se ve que tienen distribuciones schewed to the right, es decir, sus datos están sesgados en cantidades bajas por lo que casi NO HAY SANGRADO en alguno de los cuartos de la vaca.
  - Por la baja distribución y que no hay un objetivo claro en el proyecto con estas columnas se propone descartarlas.
- Usar medidas de estadística descriptiva para las otras columnas podría resultar en un acierto clave para conocer promedios de producción. Que hasta el momento, parece ser una variable que tendrá un gran peso en nuestra variable dependiente del modelo.

Mapa de correlaciones

Realizando eliminación de columnas con datos nulos como `Usuarios`, hicimos una correlación con la fecha de evento y el ID dando la siguiente matriz de correlación:



Con esta información se puede apreciar que correlaciones tendrán más peso unas con otras (independientemente de si son negativas o positivas) para ver cuales son las que podrían tener un mayor poder predictivo en nuestro modelo.