Exploration Report (ID Vaca(1204-8794))

- **Primeros descubrimientos**
    - La primera fila del dataset corresponde a diferentes grupos para las columnas. Así que, a la hora de la programación resultará útil usar como header la columna 2 para el nombre de las demás features, es decir, las demás columnas.
    - Algunas de las columnas corresponden a fechas o horas de ordeño.
    - Los datos tienen variables categóricas, tales como:
        - Acción
        - EO/PO
        - Destino Leche
    - Varias columnas, como Patada, Pezones Encontrados, Usuario y RCS no tienen registros útiles o la mayoría son nulos.
    - En la categoría de Sangre (ppm) tenemos un grupo de columnas que van desde DI-TD. Estas columnas nos indican la cantidad de sangre por partes de millón en la sangre. Como son pocos registros y en general el proyecto no se enfoca en estos rubros, se descartó este grupo de columnas.

- **Hipótesis inicial y su impacto en el proyecto**
    - Considerando el gran volumen de csv individuales de cada vaca, y su gran cantidad de datos, tras el limpiado, nos puede indicar que todos estos registros de lactancia y producción de las vacas serán sumamente importantes para el modelo. Y al tener registradas una gran cantidad de fechas, podremos determinar la fecha de finalización de la ordeña al conocer todos sus días de ordeña y cuando fue su día de secado en otro dataset

- **Gráficas y figuras**

Tipo de datos

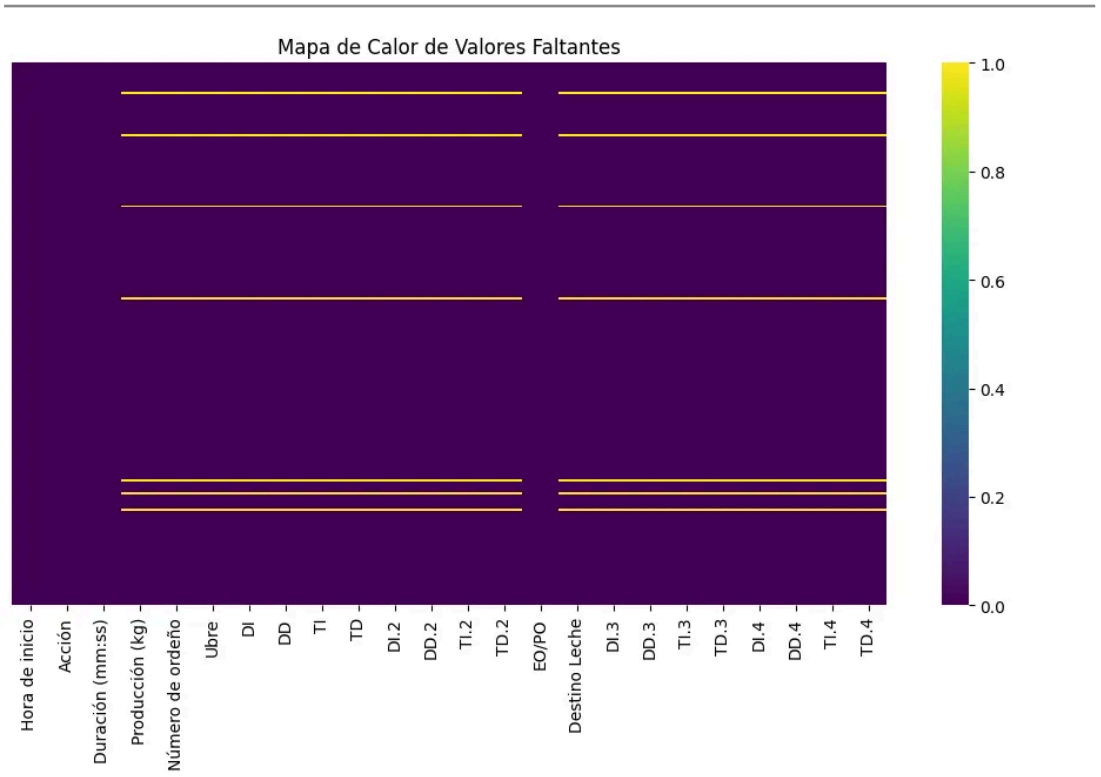
```
# Check the data types of each column
print("Tipo de dato de cada columna:")
print(df_dropped.dtypes)
```

```
Tipo de dato de cada columna:
Hora de inicio       object
Acción               object
Duración (mm:ss)     object
Producción (kg)     float64
Número de ordeño    float64
Ubre                float64
DI                  float64
DD                  float64
TI                  float64
TD                  float64
DI.1                float64
DD.1                float64
TI.1                float64
TD.1                float64
DI.2                float64
DD.2                float64
TI.2                float64
TD.2                float64
EO/PO                object
Destino Leche        object
DI.3                float64
DD.3                float64
TI.3                float64
TD.3                float64
DI.4                float64
DD.4                float64
TI.4                float64
TD.4                float64
dtype: object
```

Tipo de variables en el dataset vacas.csv

Como se puede ver la mayoría de los registros pertenecen a datos de tipo float64. Los restantes, de tipo object, como mencionamos anteriormente corresponden a variables categóricas o fechas.

Mapa de calor



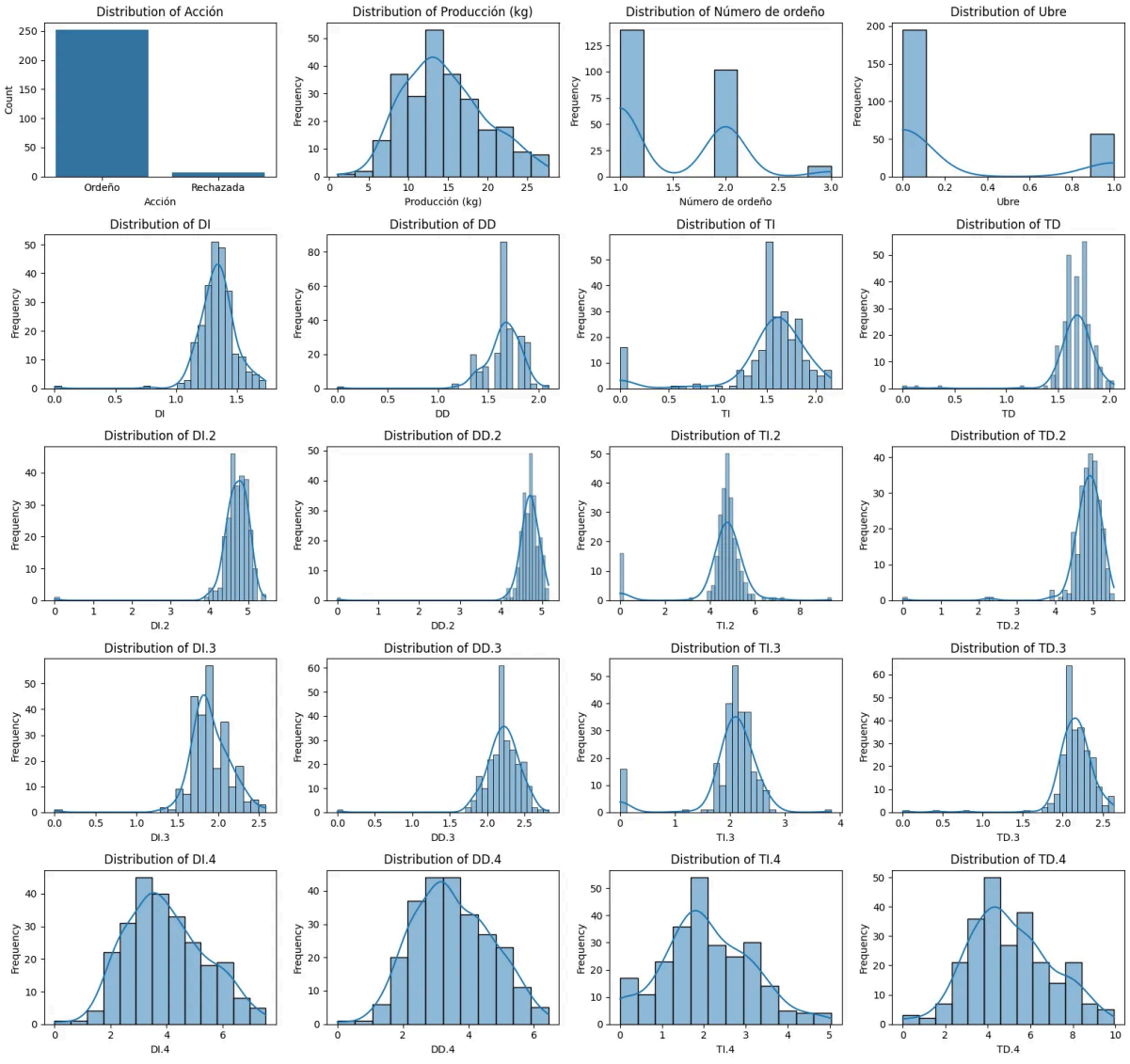
Mapa de calor con valores faltantes del archivo vaca.csv

Nótese que parece que los registros donde hay valores faltantes pertenecen a una misma fila, es decir, pertenecen a una misma fecha donde NO SE ORDEÑO A LA VACA ESE DÍA, fue rechazada.

Esto será importante de considerar y tener en mente ya que estos días varían dependiendo de la vaca que estemos analizando.

Histograma de todas las columnas para la visualización de distribuciones

---

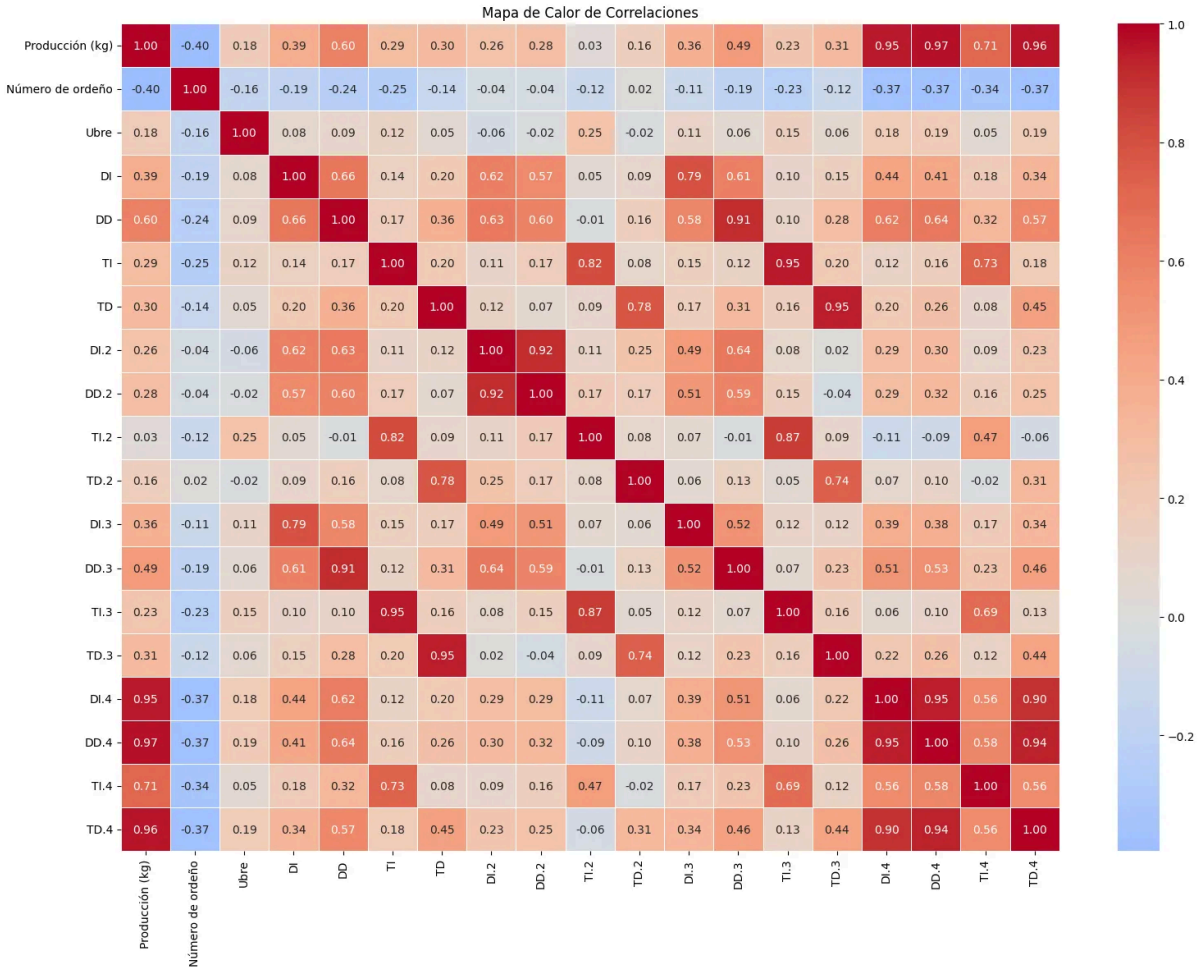


Es importante destacar que la producción de leche por lactancia en cada uno de los cuartos de la vaca parece seguir una distribución normal. No se observan valores atípicos (outliers), y la cantidad de litros producidos por lactancia muestran una variabilidad relativamente baja entre los diferentes cuartos. Esto sugiere una consistencia en el patrón de producción, sin desviaciones significativas.

- Aunado a esto, nótese que la cantidad de leche por lactancia va AUMENTANDO. En la primera y segunda lactancia se puede observar que la producción tiene valores más pequeños en el eje de las X´s, donde el número de litros generados en promedio es de 1.5 lts. Que, al comparar con la tercera y cuarta lactancia, se puede observar que la media de producción es de 2 o incluso 3.
- Este incremento puede deberse a un mayor desarrollo y tamaño de la ubre, así como al aumento del tamaño corporal en comparación con la primera lactancia. Considerar estos resultados y comparar con otras vacas resulta algo interesante a tomar en cuenta.

La frecuencia de veces que esta vaca se va a ordeñar es de 1. Sin embargo, parece que 2 veces al día también tiene una gran magnitud en el histograma. Dependiendo de este valor, y de la cantidad de leche producida, se podría determinar si la vaca es una alta productora dentro de CAETEC.

Mapa de correlaciones de todas las variables del dataframe



Para validar un poco lo que vimos en el histograma, la producción denota una mayor correlación con DI4-TD4, es decir, con la 4ta lactancia.