

Exploration Report (Patadas)

- **Primeros descubrimientos**

- Alrededor de 12 columnas tienen registros nulos o con 0.
- De aproximadamente 1/3 de las columnas restantes no se tiene información del significado de cada feature.
- Las columnas **DI4-TD4** cuentan con datos de tipo int y float.
- La columna **Hora de inicio del ordeño** y **Media intervalo de ordeño** son las únicas columnas con objetos como fechas.

- **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 .csv.

- **Gráficas y figuras**

Tipos de datos de todas las features

```

Tipo de dato de cada columna:
Número del animal      int64
DEL                    int64
Hora Inicio Ordeño     object
Ubre                   int64
DI                     int64
DD                     int64
TI                     int64
TD                     int64
MDI                    float64
MI>12                  int64
Media Interv. Ordeño   object
Éxito IO (h)           float64
DI.1                   int64
DD.1                   int64
TI.1                   int64
TD.1                   int64
DI.2                   int64
DD.2                   int64
TI.2                   int64
TD.2                   int64
DI.3                   int64
DD.3                   int64
TI.3                   int64
TD.3                   int64
DI.4                   float64
DD.4                   float64
TI.4                   float64
TD.4                   float64
DI.5                   int64
DD.5                   int64
TI.5                   float64
TD.5                   float64
RCS (x 1000)           int64
dtype: object

```

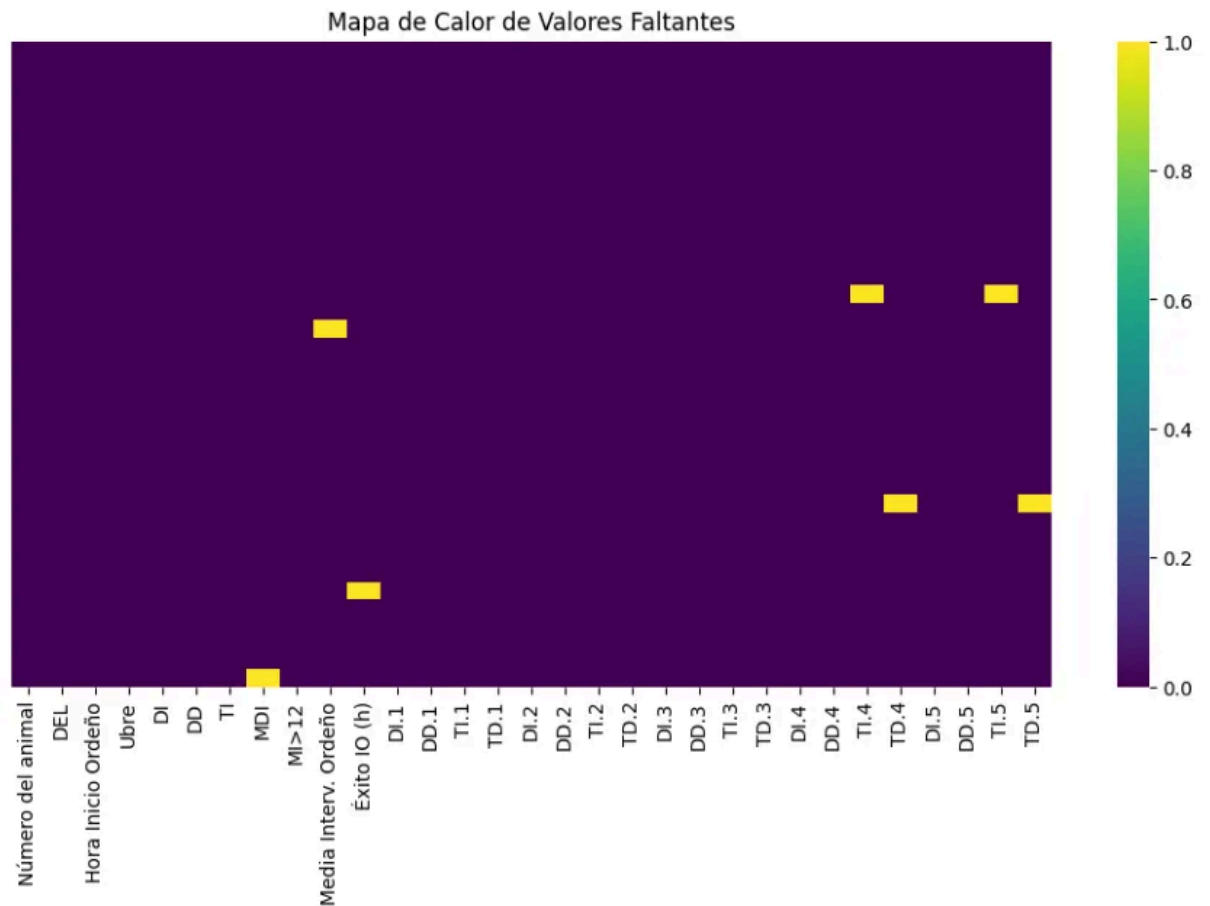
Tipo de datos de patadas.csv

Nótese que la mayoría de datos del .csv corresponden a valores numéricos. Ya sean datos de tipo float o tipo int. A futuro, para la parte de cada cuarto, puede que convenga estandarizar todos estos números a un solo tipo de dato, como de tipo float.

Las únicas variables de tipo categóricas corresponden a los objetos que se visualizan en la imagen anterior siendo:

- Hora Inicio Ordeño (fecha en formato DD-MM-YYYY HH:MM)
- Media Interv. Ordeño (Hora en formato HH:MM)

Mapa de calor de todas las features



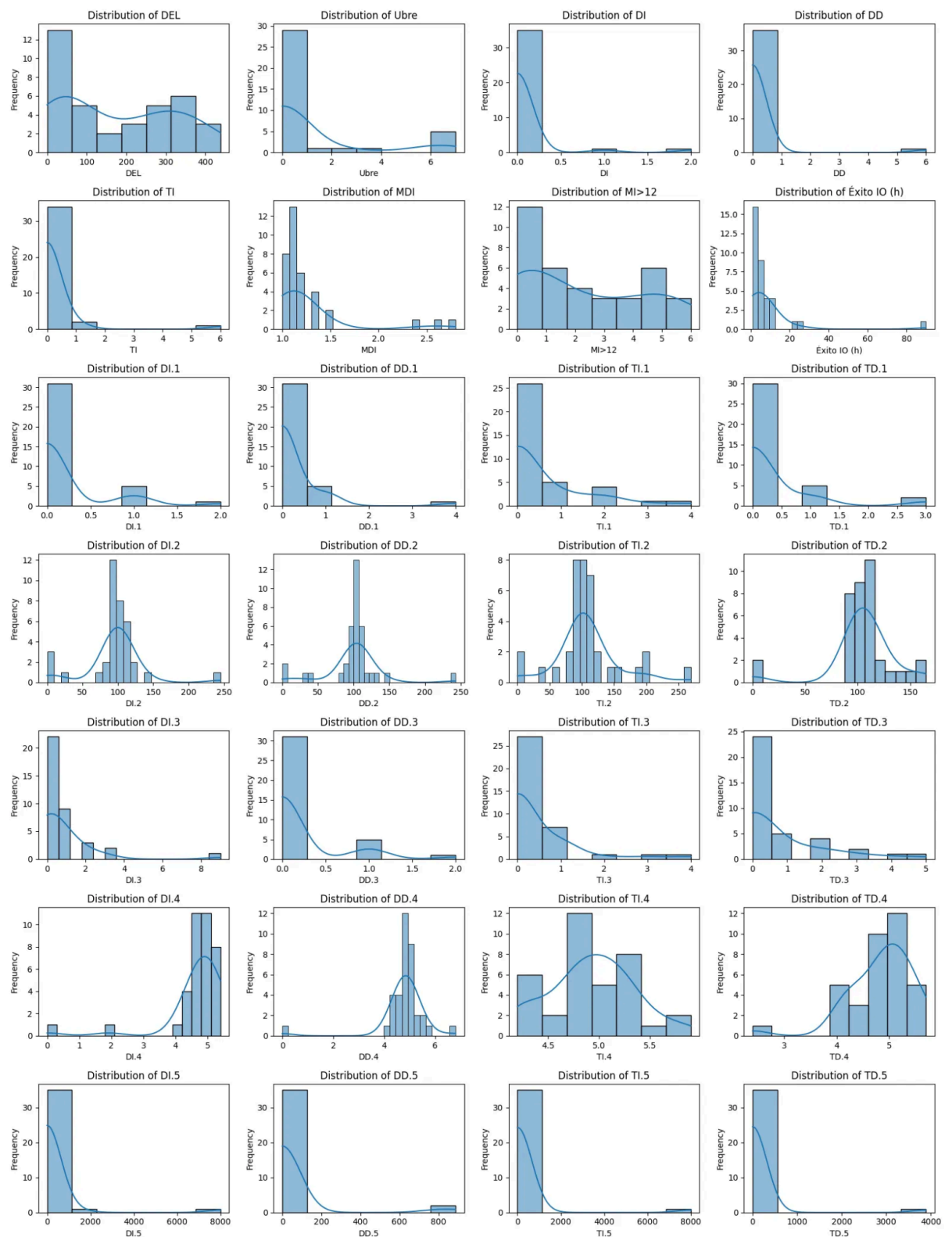
Mapa de calor de todas las columnas de patadas.csv

Los datos faltantes en las columnas de los cuartos de la vaca nos puede indicar que ese cuarto no se ordeño por alguna razón como:

- El robot no ordeño el cuarto por que puede que sea un “cuarto seco”.
- Lesiones o problemas de mastitis en alguno de los cuartos.
- Atrofia glandular.

Se desconoce por qué hay registros faltantes en las features de MDI y Éxito IO.

Histograma de todas las columnas para la visualización de distribuciones



Histograma de distribuciones para todas las columnas de patadas.csv

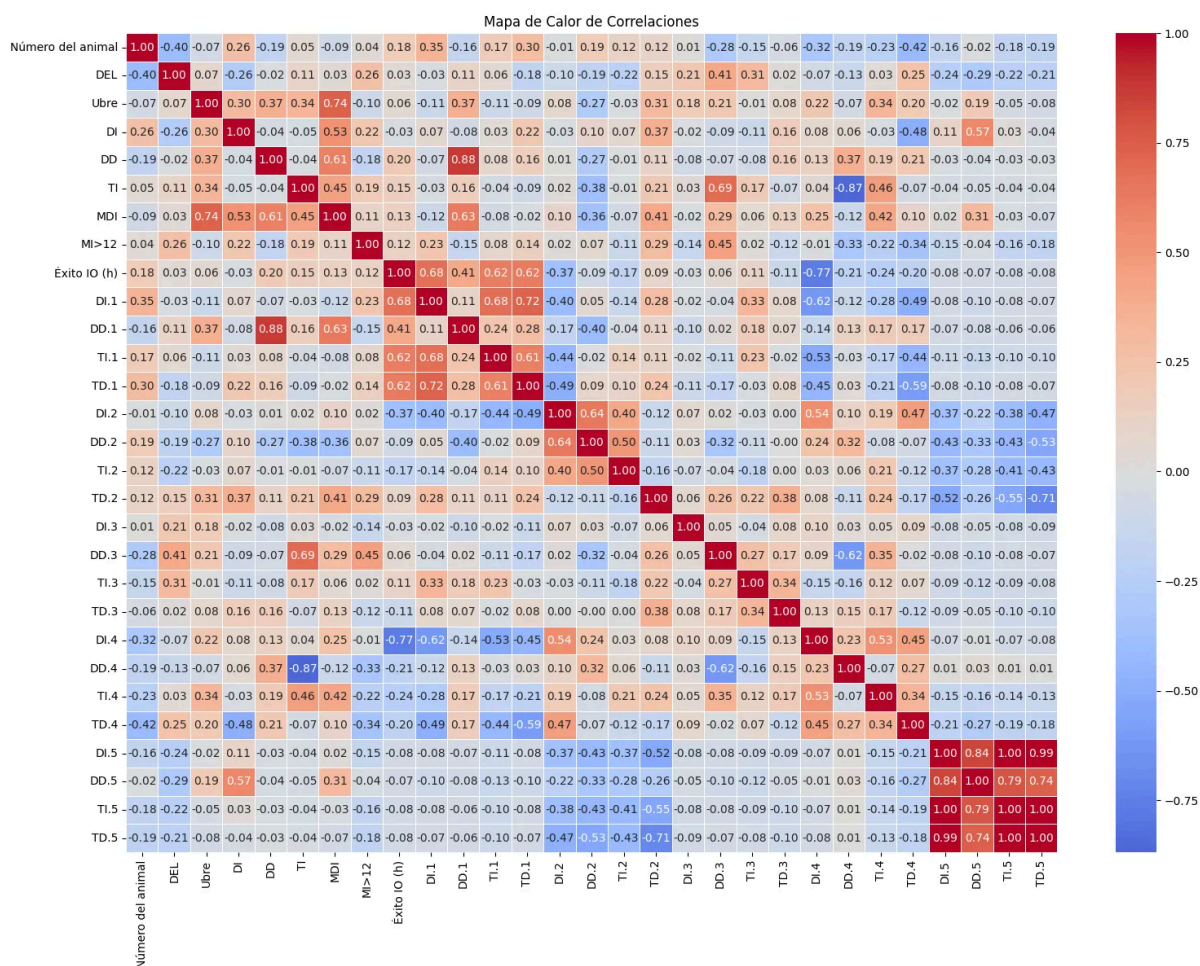
El histograma de DEL tiene una distribución muy particular, que al calcular su media nos da un valor de ~176 (Días en leche, cuántos días lleva produciendo leche). Esta información puede ser una pieza muy importante considerando los propósitos del proyecto.

Del histograma Ubre y MDI no se tiene información relevante para evaluar ni interpretar el resultado obtenido.

Los histogramas de los diferentes cuartos de ordeño, desde DI0 a TD5 son muy variados. No tienen un patrón establecido entre diferentes lactancias. Por lo que es probable que esto no tenga relación con la producción de leche en cada lactancia.

- Parece que los histogramas de la 6ta lactancia (DI5-TD5) no son significativos, ya que viendo el dataset y todos sus registros, solo hay dos registros que presentan los valores más altos del histograma. El valor con más frecuencia que se visualiza en el histograma es 0.
- Se necesita verificar a que hace referencia cada lactancia, ya que no tienen que ver con producción (para esto último se tiene cada dataset individual de cada vaca para corroborar este dato).

Mapa de correlaciones:



Mapa de calor con las correlaciones de todas las features de patadas.csv

- Nótese que las correlaciones más grandes que presenta DEL (Días en leche, cuántos días lleva produciendo leche) parece que se dan con DI3-TD3. Es importante verificar que significa esto último para entender por qué se da esta relación y cómo es que se puede interpretar.