**PRE-ANÁLISIS DESCRIPTIVO DE LOS DATOS CRUDOS**

**PROYECTO: Optimización de la captura y visualización de datos del funcionamiento que usa el esquema del tren de evaporación en una empresa del sector azucarero.**

**Evaluación inicial de los datos**

1. **Cargue y visualización:**

Se realiza con el uso de la biblioteca Pandas en Python y SQL.

Tomando los archivos entregados desde las dos fuentes PCS7 y Sigind, en el batch inicial los datos se conforman como se mostrará a continuación, con el fin de calcular y obtener el indicador 1 (Eficiencia vapor) e indicador 2 (Eficiencia evaporación):

|  |  |  |  |
| --- | --- | --- | --- |
| **Variables PCS7** | | | |
| **Item** | **Tipo de dato** | **Uso** | **Indicador** |
| Flujo cond Klb/hr Time | Date | Si | 1 |
| Flujo cond Klb/hr ValueY | Float | Si | 1 |
| Flujo cana Ton Time | Date | Si | 1 |
| Flujo cana Ton ValueY | Float | Si | 1 |
| Jcla C1 m3/hr Time | Date | Si | 2 |
| Jcla C1 m3/hr ValueY | Date | Si | 2 |
| Jcla C2 m3/hr Time | Date | Si | 2 |
| Jcla C2 m3/hr ValueY | Float | Si | 2 |
| Jcla C3 m3/hr Time | Date | Si | 2 |
| Jcla C3 m3/hr ValueY | Float | Si | 2 |
| Jcla C4 m3/hr Time | Date | Si | 2 |
| Jcla C4 m3/hr ValueY | Date | Si | 2 |
| Cond C1 m3/hr Time | Date | No | NaN |
| Cond C1 m3/hr ValueY | Float | No | NaN |
| Cond C2 m3/hr Time | Date | No | NaN |
| Cond C2 m3/hr ValueY | Float | No | NaN |
| Cond C3 m3/hr Time | Date | No | NaN |
| Cond C3 m3/hr ValueY | Date | No | NaN |
| Cond C4 m3/hr Time | Date | No | NaN |
| Cond C4 m3/hr ValueY | Float | No | NaN |
| Brix mel Time | Date | Si | 2 |
| Brix mel ValueY | Float | Si | 2 |

*Tabla No.1 Variables de archivo de fuente PCS7*

|  |  |  |  |
| --- | --- | --- | --- |
| **Variables Sigind** | | | |
| **Item** | **Tipo de dato** | **Uso** | **Indicador** |
| FECHA/HORA | Date | Si | 2 |
| JCLARBX | Float | Si | 2 |
| MSCBXTA | Float | Si | 2 |
| JALCPH | Float | No | NaN |
| JCLPHCL | Float | No | NaN |
| MSCPHTA | Float | No | NaN |
| JCLARPZ | Float | No | NaN |
| MSCPZTA | Float | No | NaN |

*Tabla No.2 Variables de archivo de fuente Sigind*

1. **Identificación de valores nulos:**

En el proceso de carga y transformación; los datos provenientes de las fuentes PCS7 y Sigind, las variables fueron convertidas por el tipo de formato “object”, debido a los datos nulos que presentan algunos registros. En este proceso se identificó un estimado de 14.400 datos nulos para la mayoría de las variables, a excepción de las variables Brix mel Time y Brix mel ValueY que presentaron 23.040 datos nulos aproximadamente.

Tabla

El contenido generado por IA puede ser incorrecto.

*Figura No.1 Visualización de cantidad de nulos*

1. **Comprensión del origen de los datos nulos:**

Es importante tener en cuenta que la generación de datos nulos se da porque cuando se cargan los datos en el DataFrame, esta toma la mayor cantidad de registros por columna para construir el número de filas, para las columnas con menos registros las llena con nulos, por esto toma la totalidad de los datos de flujo caña, como el tamaño de las filas, es decir, 28.800. Para el resto de las variables se expande a esa cantidad, y teniendo en cuenta que algunas columnas solo tienen 14.400 registros, en la formación del DataFrame llena los campos adicionales y faltantes con dato nulo para así completar las 28.800 filas.