**#¿Qué es data?**

type(data)

**#Dimension de los datos**

data.shape

**#Vistazo de datos**

data.head(10)

**#Tipo de cada dato**

data.dtypes

**#Valores nulos**

missing = pd.DataFrame(data.isnull().sum(),columns=['Valores\_Nulos'])

missing

**#valores únicos**

unival = pd.DataFrame(data.nunique(), columns=['Valores\_Unicos'])

Unival

**#Detalles estadísticos**

data.describe(include='all')

**# Variable -> Tipo de Vidrio**

#cuántos datos hay de cada tipo de vidrio

sns.countplot(x=data["Tipo\_Vidrio"])

**Conclusión 1:** no hay datos sobre el tipo de vidrio 4 (ventanas de vehículos no procesadas por flotación), por lo tanto no se puede predecir nada sobre este tipo de vidrio.

Conclusión 2: el tipo de vidrio 2 (2 ventanas de edificios no procesadas por flotación) es el que tiene el más alto índice de refracción

Conclusión 3: a menor potasio, mayor calcio en los vidrios

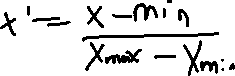
Conclusión 4: es más probable que un vidrio sea del tipo 5 ó 7 cuando tienen mayor cantidad de potasio.

**ESCALAMIENTO**

|  |  |  |  |
| --- | --- | --- | --- |
| **Empleado** | **Salario** | **Calificaciones escuela** | **¿buen empleado?** |
| 1 | 500,000 | 9 | Sí |
| 2 | 2’000,000 | 10 | NO |
| 3 | 600,000 | 5 | NO |
| 4 | 200,000 | 6 | NO |
| 5 | 300,000 | 7 | SI |
| 6 | 1’000,000 | 6 | SI |

Poner valores en el mismo rango o escala para que ninguna variable sea dominada por otra

**Normalizacion**



* Valores cercanos a su media
* La distribución de las variables no cambia
* Cuando queremos tener datos positivos
* Modelos basados en distancias (KNN, K-means, Redes neuronales)
* Datos con distribuciones desconocidas
* Cuando el rango de los datos es más importante

**Estandarizacion**



* Media 0 y desviación estándar de 1. Distribución normal
* Cuando admitimos datos negativos
* Cuando hay atípicos que son importantes y no queremos perder su impacto
* Modelos lineales: modelos que asumen que las variables tienen una distribución normal o gaussiana como la regresión lineal, regresión logística, maquina de vector soporte
* Técnicas de reducción de dimensionalidad (PCA)
* Datos con distribuciones normales.

**#Datos escalados**

sns.boxplot(y=standard['Indice\_Refraccion'])

**\*\*¿Escalar o quitar datos atípicos primero?\*\***

Depende.... pero se recomienda primero quitar los outliers (si es que estamos seguros que son errores) y luego escalar los datos.