This Python 3 environment comes with many helpful analytics libraries installed # It is defined by the kaggle/python Docker image: https://github.com/kaggle/docker-python # For example, here's several helpful packages to load import numpy as np # linear algebra import pandas as pd # data processing, CSV file I/O (e.g. pd.read_csv) # Input data files are available in the read-only "../input/" directory # For example, running this (by clicking run or pressing Shift+Enter) will list all files under the input directory import os for dirname, _, filenames in os.walk('/kaggle/input'): for filename in filenames: print(os.path.join(dirname, filename)) # You can write up to 20GB to the current directory (/kaggle/working/) that gets preserved as output when you create a version using "Save & Run All" # You can also write temporary files to /kaggle/temp/, but they won't be saved outside of the current session

import pandas as pd import numpy as np import matplotlib.pyplot as plt import seaborn as sns

df = pd.read_csv('/kaggle/input/student-performance-data-set/student-mat.csv')

df.info() df.describe() df.isnull().sum()

df = df.drop_duplicates() df = df.dropna() # si el dataset tiene valores faltantes

df.columns

for col in df.columns: print(repr(col))

df.columns = df.columns.str.strip() # quita espacios df.columns = df.columns.str.replace('\ufeff', '') # quita BOM df.columns = df.columns.str.replace('\r', '') # quita saltos df.columns = df.columns.str.replace('\n', '') # quita saltos invisibles

df.columns

df.columns = df.columns.str.strip() df.columns = df.columns.str.replace("\ufeff", "", regex=False)

for col in df.columns: print(repr(col))

df.columns = df.columns.str.replace(r'\s+', '', regex=True) df.columns = df.columns.str.replace('\ufeff', '', regex=False) df.columns = df.columns.str.replace('\r', '', regex=False) df.columns = df.columns.str.replace('\t', '', regex=False) df.columns = df.columns.str.strip()

"G3" in df.columns

for col in df.columns: print(repr(col))

df = pd.read_csv('/kaggle/input/student-performance-data-set/student-mat.csv', sep=';') df.head()

df.columns

plt.figure(figsize=(8,5)) sns.histplot(df['G3'], kde=True) plt.title('Distribución de la Nota Final (G3)') plt.xlabel('Nota Final') plt.ylabel('Frecuencia') plt.show()

df.describe()

sns.scatterplot(x=df['studytime'], y=df['G3']) plt.title('Horas de estudio vs Nota Final') plt.show()