**BIG DATA ANALYSIS USING IBM CLOUD DATABASES**

|  |  |
| --- | --- |
| Date | 26-10-2023 |
| Team ID | 673 |
| Project Name | BIG DATA ANALYSIS WITH IBM CLOUD DATABASES |

**PHASE 4:**Development Part 2

**Problem Title: BIG DATA ANALYSIS**

**Problem Statement:** Dive into the world of big data analysis with IBM Cloud Databases. Uncover hidden insights from vast datasets, from climate trends to social patterns. Visualize your findings and derive valuable business intelligence. Embark on data-driven adventures, exploring the endless possibilities of big data!

**DATA ANALYSIS AND VISUALIZATION:**

**import pandas as pd**

**import matplotlib.pyplot as plt**

**import seaborn as sns**

**# Load the dataset**

**csv\_file = r'D:\anees\Documents\Phase\_03\rainfall in india 1901-2015\_ibm.csv' # Update the file path**

**df = pd.read\_csv(csv\_file)**

**# Simple Analysis**

**# 1. Basic Statistics**

**summary\_stats = df.describe()**

**# 2. Data Shape**

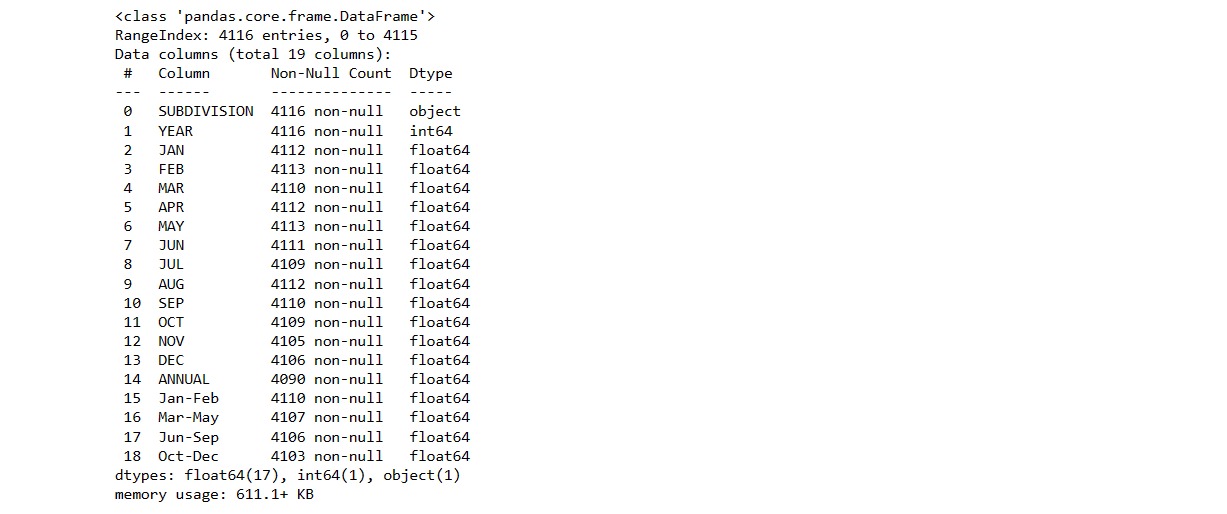
**data\_shape = df.shape**

**# 3. Column Information**

**column\_info = df.info()**

**# 4. Data Head**

**data\_head = df.head()**

****

**# Complex Analysis**

**# 5. Data Distribution Visualization**

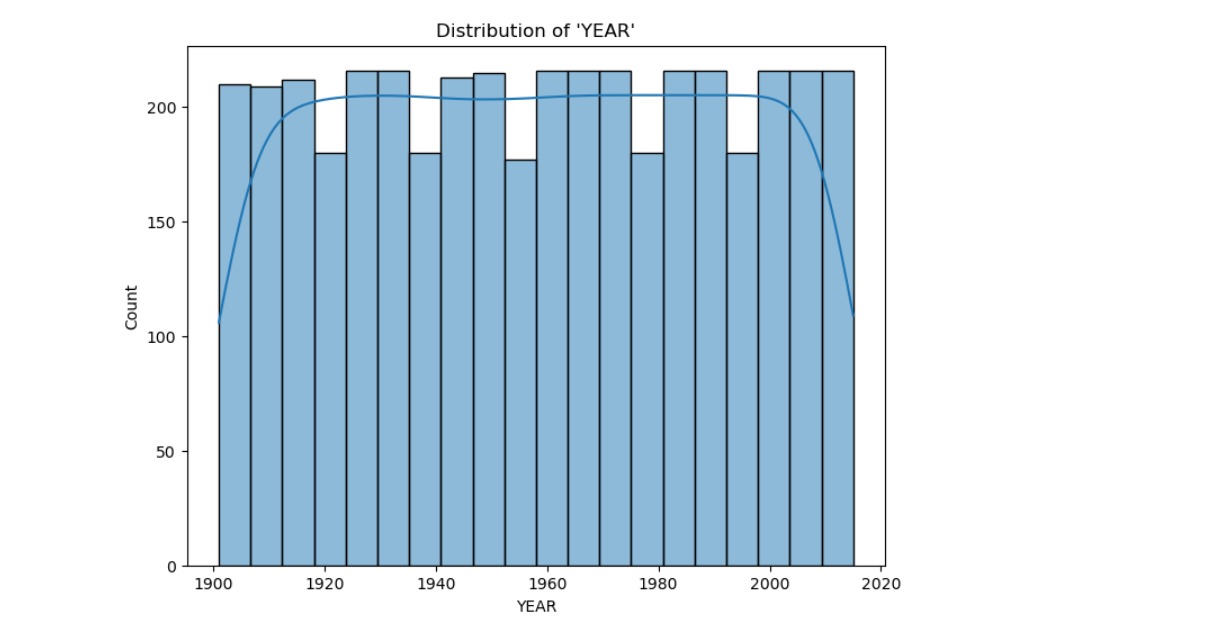
**# Example: Histogram of a numeric column**

**plt.figure(figsize=(8, 6))**

**sns.histplot(data=df, x='YEAR', bins=20, kde=True)**

**plt.title("Distribution of 'YEAR'")**

**plt.show()**

****

**# 6. Correlation Analysis**

**# Example: Correlation heatmap of numeric columns**

**df\_numeric = df.select\_dtypes(include=['number'])**

**# Correlation Analysis**

**# Calculate the correlation matrix**

**correlation\_matrix = df\_numeric.corr()**

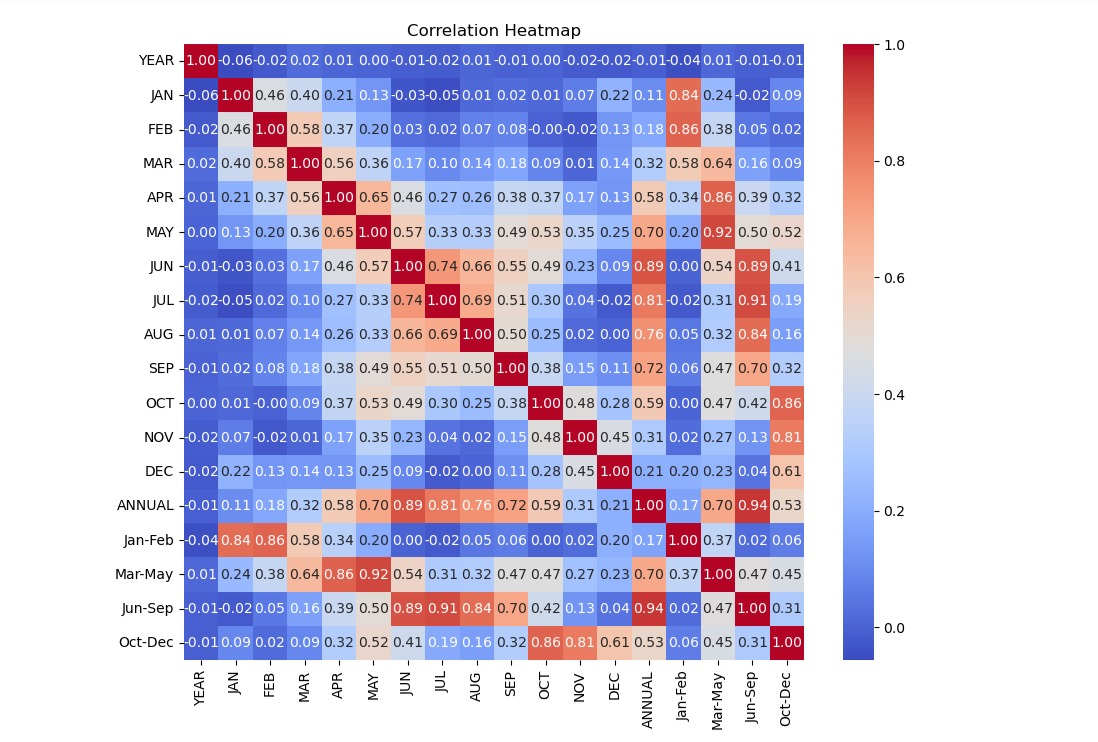
**# Create a heatmap of the correlation matrix**

**plt.figure(figsize=(10, 8))**

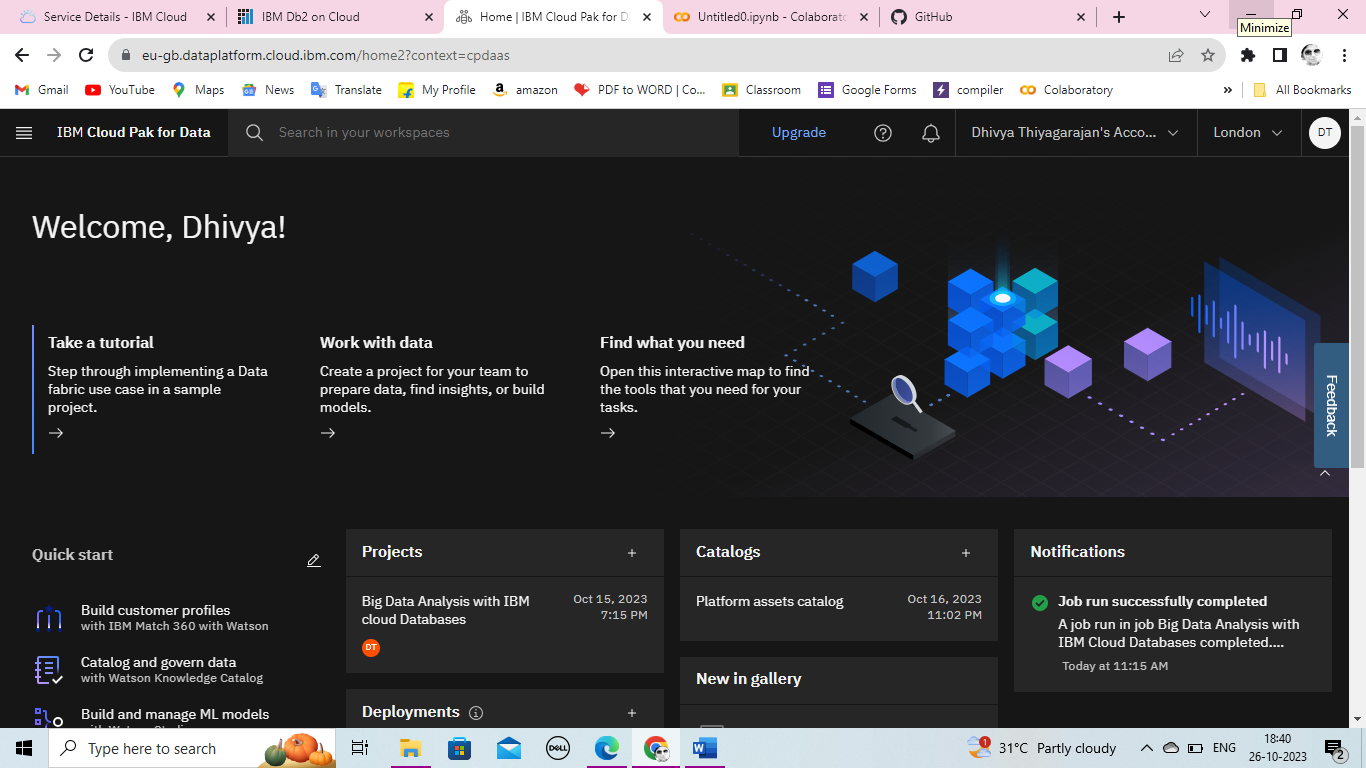
**sns.heatmap(correlation\_matrix, annot=True, cmap="coolwarm", fmt=".2f")**

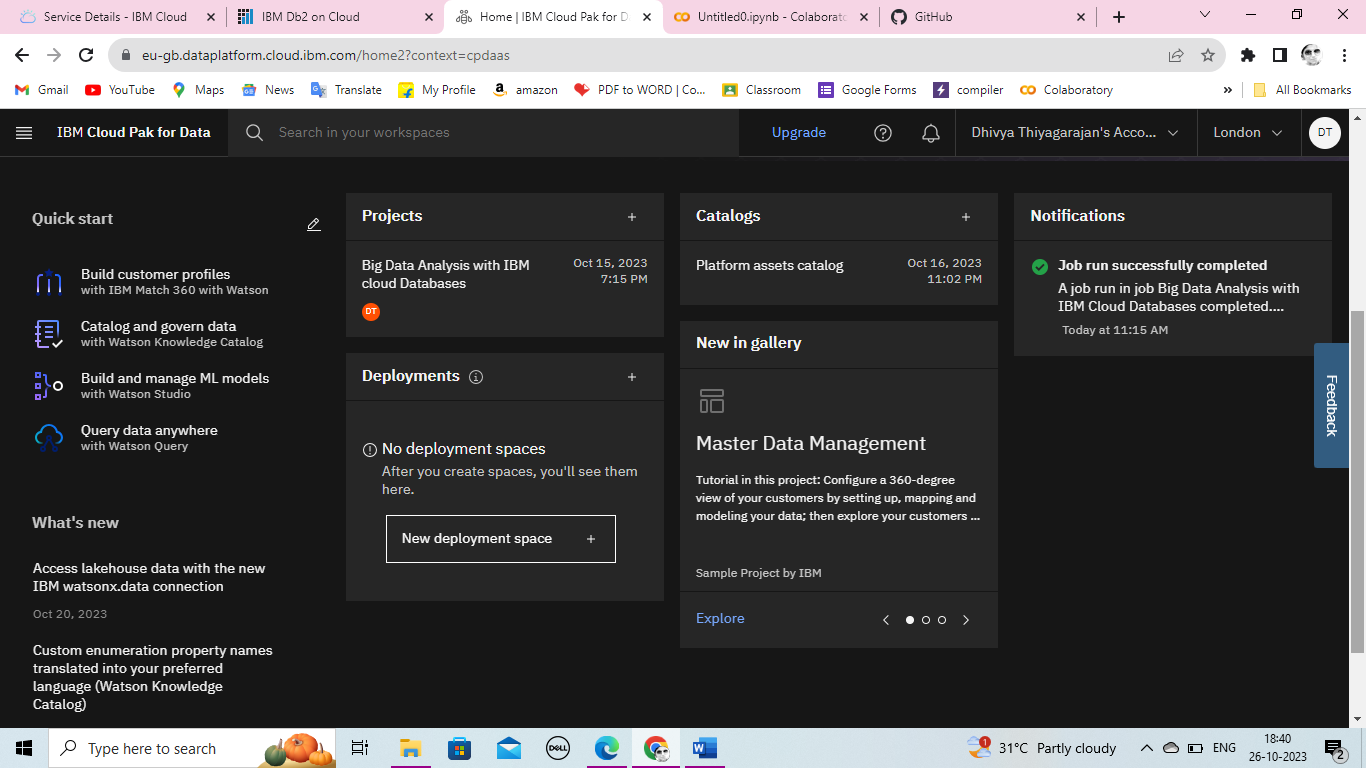
**plt.title("Correlation Heatmap")**

**plt.show()**

****

**IBM CLOUD PAK FOR DATA:**

****

****

**# 7. Categorical Data Analysis**

**# Example: Countplot of a categorical column**

**plt.figure(figsize=(8, 6)) # Set the figure size**

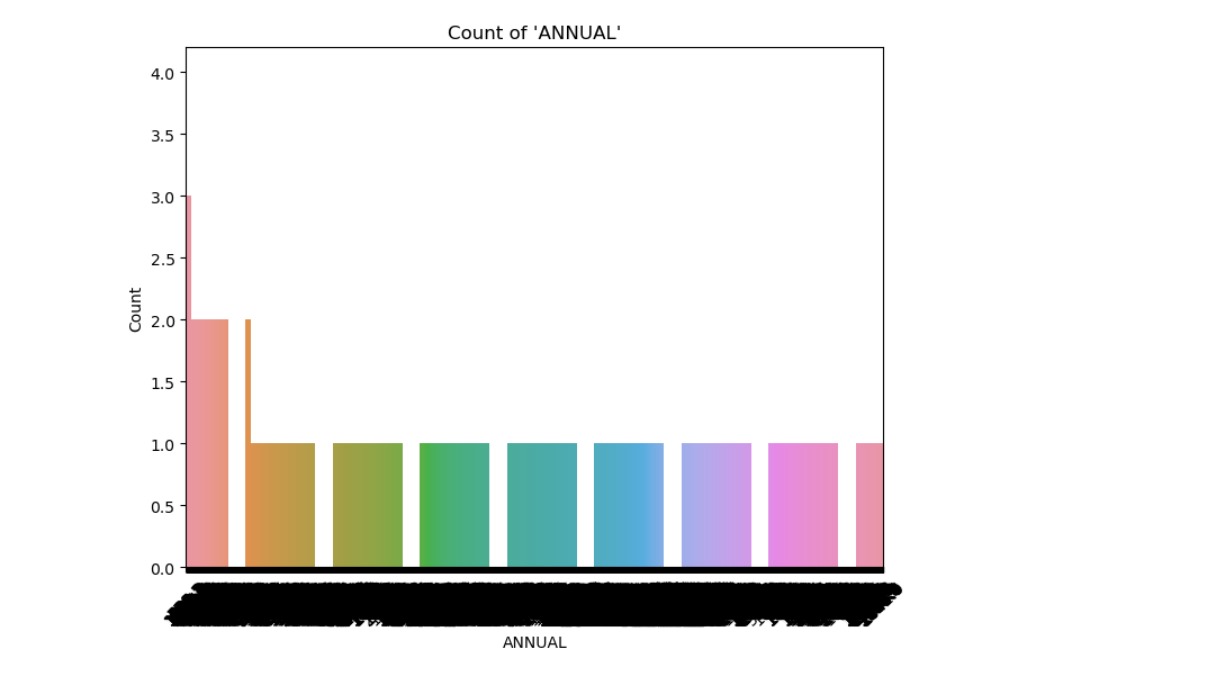
**ax = sns.countplot(data=df, x='ANNUAL', order=df['ANNUAL'].value\_counts().index) # Use 'order' to set the value range**

**ax.set(ylabel="Count") # Customize the y-axis label**

**plt.title("Count of 'ANNUAL'")**

**plt.xticks(rotation=45)**

**plt.show()**

****

**# 8. Time Series Analysis (if applicable)**

**# Example: Line plot of a time series column**

**df['YEAR'] = pd.to\_datetime(df['YEAR'])**

**df.set\_index('YEAR', inplace=True)**

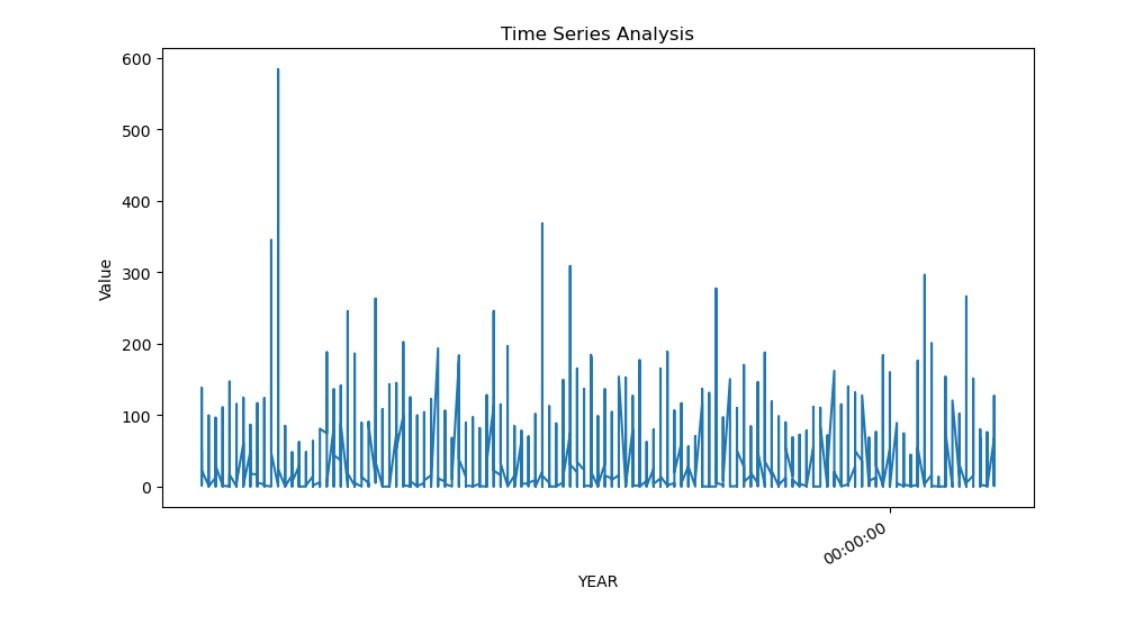
**df['JAN'].plot(figsize=(10, 6))**

**plt.title("Time Series Analysis")**

**plt.xlabel('YEAR')**

**plt.ylabel('Value')**

**plt.show()**

****

**# Display Results**

**print("\nSummary Statistics:\n")**

**print(summary\_stats)**

**print("\nData Shape:\n", data\_shape)**

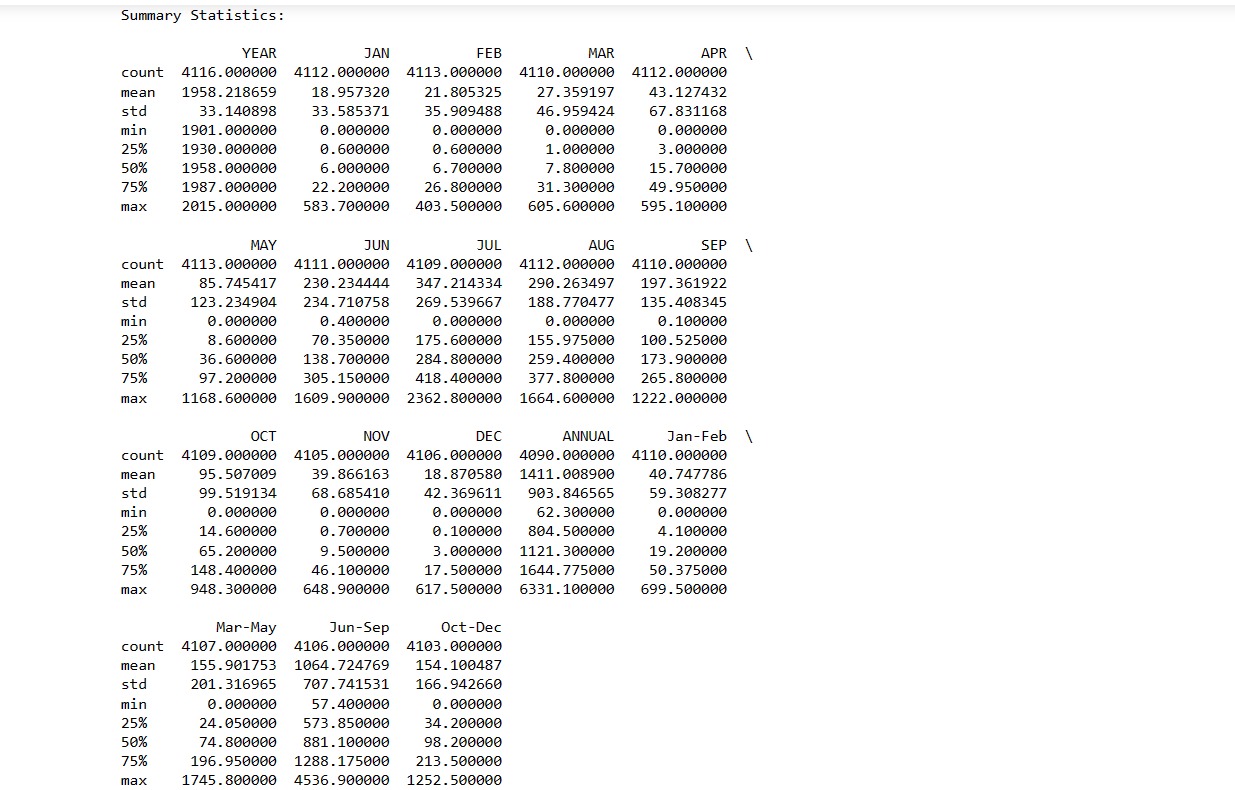
**print("\nColumn Information:\n")**

**print(column\_info)**

**print("\nData Head:\n")**

**print(data\_head)**

**print(f"\nAccuracy of the model: {accuracy:.2f}")**

****