Title: Documentation for Data Analysis and Cleaning of Forex Dataset

Author: Richard Kabiru Reg No: 150684

Table of Contents

Introduction

The code provided aims at analyzing a dataset containing Forex exchange rates over a period of time. The primary steps undertaken include importing necessary libraries, creating the dataset, data exploration, data cleaning, and feature selection to ensure data quality and integrity which are vital for accurate analysis.

Importing Libraries

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

The necessary libraries for data manipulation, analysis, and visualization are imported.

Creating the Dataset

```
# Data
data = {
    'month' : ["apr", "may", "jun", "jul", "aug", "sep", "oct", "nov", "dec"],
    '1 usd' : [107.95,107.43,107.81,108.14, 1092.4, 110.13, 110.86,112.92, 112.91],
    "1 Pound Sterling" :
[149.3,151.06,151.45,149.37,150.87,151.51,151.58,150.97,150.95],
    "1 Euro" : [129.13, 130.41,130.07,127.98,128.59,129.8,128.6,127.94, 127.64],
    "100 Japanese Yen" : [99.03,98.43,97.99,98,99.47,100.04,98.03,98.12,99.2],
    "1 SA Rand" : [7.49,7.63,7.76, NA, 7.39,7.57,7.44,7.23,7.1],
    "USHS/KES" : [33.59, 33.09, 32.84, 32.84, 32.39, 32.07, 32.28,31.74, 3.149],
    "TSHS/KES" : [21.48,21.59,21.51,21.44,21.23,21.04,20.78,20.57,20.4]
}

# Creating a DataFrame
forex = pd.DataFrame(data)
```

A dictionary is created to hold the data, which is then converted into a DataFrame using pandas for ease of manipulation and analysis.

Data Exploration

```
# Viewing the first and last few rows of the dataset
forex.head()
forex.tail()

# Checking column names, shape, and brief statistics of the dataset
forex.columns
forex.shape
forex.describe()

# Checking for unique values and null values in the dataset
forex.nunique()
forex.isna().values.any()
forex.isnull().any()
```

Basic exploratory data analysis is performed to understand the structure, completeness, and basic statistics of the dataset.

Data Cleaning

```
# Filling missing value in '1 SA Rand' column with the column mean
forex['1 SA Rand'].fillna((forex['1 SA Rand'].mean()), inplace=True)

# Verifying that the missing value has been replaced
forex.isna().any()
```

The missing value in the '1 SA Rand' column is identified and filled with the mean of the column to maintain data continuity.

Feature Selection

```
# Renaming the dataset
ken_forex = forex

# Calculating and visualizing Pearson correlation among features
plt.figure(figsize=(12,10))
cor = ken_forex.corr()
sns.heatmap(cor, annot=True, cmap=plt.cm.Blues)
plt.show()
```

The dataset is renamed, and a correlation matrix is computed and visualized using a heatmap to identify relationships between features, which is crucial for feature selection in model building.

By adhering to the steps of data preparation including exploration, cleaning, and feature selection, we ensure that the data is of high quality and ready for subsequent analysis or modeling. This process is fundamental to derive accurate insights and build robust predictive models.