ព្រះរាជាណាចក្រកម្ពុជា

ជាតិ សាសនា ព្រះមហាក្សត្រ

Department of Applied Mathematics and Statistics

REPORT OF INTRODUCTION TO DATA SCIENCE

GROUP: I3_AMS_TPB

TEAM: 01

Name of Students	ID of Students	Score
1. PAV Limseng	e20211548	…….
2. KOUM Soknan	e20211754	…….
3. PEL Bunkhloem	e20201314	…….
4. MA Ousa	e20210359	…….
5. KHON Khengmeng	e20210176	…….

Lecturer: M. PEN Chentra(TP)

Dr. PHAUK Sokkhey (Course)

Academic Year 2023-2024

Table of Contents

I. Overview of project 1

1.1. Data Collection and Overview 1

1.1.1. Libraries Used 1

1.1.2. Data Loading 1

1.1.3. Data Exploration 1

1.2. Data Visualization 1

1.2.1. Time Series Plot 1

1.2.2. Check Outlier By using Boxplot 1

1.3. Machine Learning - Linear Regression 2

1.3.1. Feature Selection and Preprocessing 2

1.3.2. Train-Test Split 2

1.3.3. Linear Regression 2

1.3.4. Prediction 2

1.4. Evaluation Metrics and Results Visualization 2

1.4.1. Plotting Actual vs. Predicted 2

1.4.2. Evaluation Metrics 2

1.4.3 Additional Considerations 2

II. EDA (Exploratory Data Analysis) 2

2.1. Introduction 2

2.2. Data Overview 3

2.2.1. Dataset Information 3

2.2.2. Data Cleaning 3

2.3. Descriptive Statistics 3

2.3.1. Summary Statistics 3

2.3.2. Time Trends 3

2.4. Conclusion 3

III. Modeling and Evaluation 4

3.1. Regression By Using Machine Learning 4

3.1.1. Data Splitting 4

3.1.2. Build Model 4

3.1.3. Model Training 4

3.1.4. Prediction 4

IV. Conclusion 4

I. Overview of project

1.1. Data Collection and Overview

1.1.1. Libraries Used

pandas: For data manipulation and analysis.
numpy: For numerical operations.
matplotlib and seaborn: For data visualization.
scikit-learn: For machine learning tasks.

1.1.2. Data Loading

The code begins by importing the necessary libraries and loading exchange rate data from a CSV file into a pandas DataFrame (data). The file is named "KHR=X.csv," suggesting it contains exchange rate information.

1.1.3. Data Exploration

The head () function is used to display the first few rows of the DataFrame, providing an initial look at the structure and content of the data.

1.2. Data Visualization

1.2.1. Time Series Plot

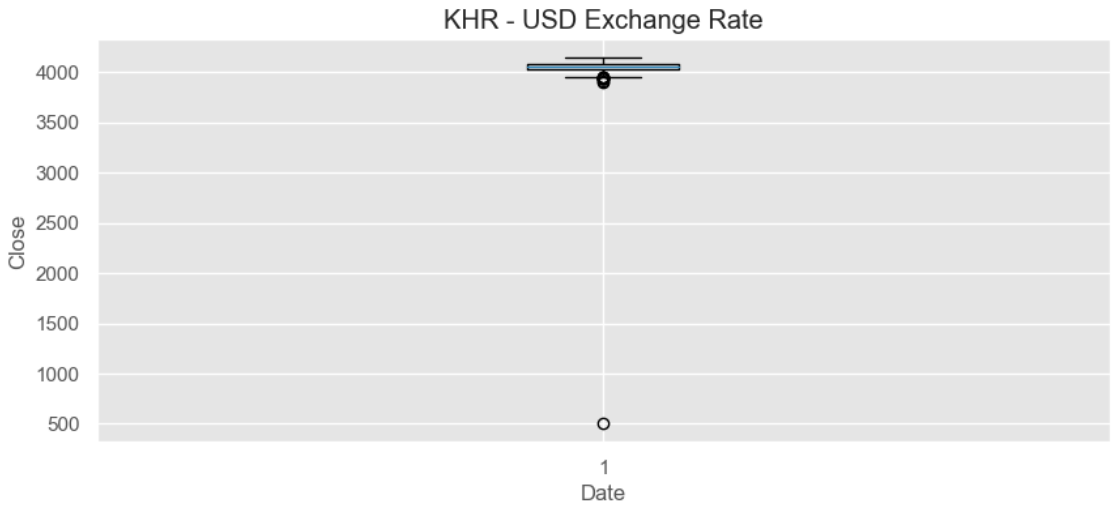
A time series plot is created using matplotlib. The x-axis represents dates, the y-axis represents the "Close" values, and the plot visualizes how the exchange rate changes over time.

Correlation Matrix and Heatmap:

A correlation matrix is calculated for numerical columns (excluding the first one) in the DataFrame. The seaborn library is then used to create a heatmap, offering insights into the relationships between different numerical features.

1.2.2. Check Outlier By using Boxplot

We create a box plot to identify outliers, focusing on dates and closing values. In this visualization, dates are represented on the X-axis as the independent variable, while closing values are on the Y-axis as the target variable.



1.3. Machine Learning - Linear Regression

1.3.1. Feature Selection and Preprocessing

Features ("Open," "High," "Low") and the target variable ("Close") are selected for machine learning. These features are then converted to numpy arrays and reshaped as needed.

1.3.2. Train-Test Split

The data is split into training and testing sets using the train_test_split function from scikit-learn. This step is crucial for assessing the model's performance on unseen data.

1.3.3. Linear Regression

A Linear Regression model from scikit-learn is instantiated (model). The model is trained on the training set (xtrain, ytrain).

1.3.4. Prediction

The trained model is used to make predictions on the test set (xtest). Predictions are stored in the ypred variable.

1.4. Evaluation Metrics and Results Visualization

1.4.1. Plotting Actual vs. Predicted

A matplotlib plot is created to visualize the actual vs. predicted exchange rates. This allows for a visual assessment of how well the model performs.

1.4.2. Evaluation Metrics

Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and Accuracy based on Mean Absolute Percentage Error (MAPE) are calculated and printed. These metrics provide quantitative measures of the model's performance.

1.4.3 Additional Considerations

The code assumes that the necessary libraries are installed, and the CSV file is present in the specified location.

Ensure that data preprocessing steps, such as handling missing values, are performed before training the machine learning model.

Model evaluation and selection could be further refined based on specific project requirements and goals.

This code provides a comprehensive overview of data exploration, visualization, and machine learning using a Decision Tree Regressor to predict exchange rates

II. EDA (Exploratory Data Analysis)

2.1. Introduction

This report aims to analyze and explore the historical exchange rates between Cambodian Riels (KHR) and US Dollars (USD). The dataset used for this analysis spans from October 11, 2022, to October 11, 2023.

2.2. Data Overview

2.2.1. Dataset Information

The dataset consists of daily exchange rates, recorded over the specified timeframe. It contains 262 rows, 7 columns and 7 variables include the date, open, high, low and close, adj close, and volume. The variables are all numeric.

2.2.2. Data Cleaning

Prior to analysis, the dataset underwent cleaning procedures to handle missing values and outliers. There are no missing values and cleaned dataset contains 262 data points.

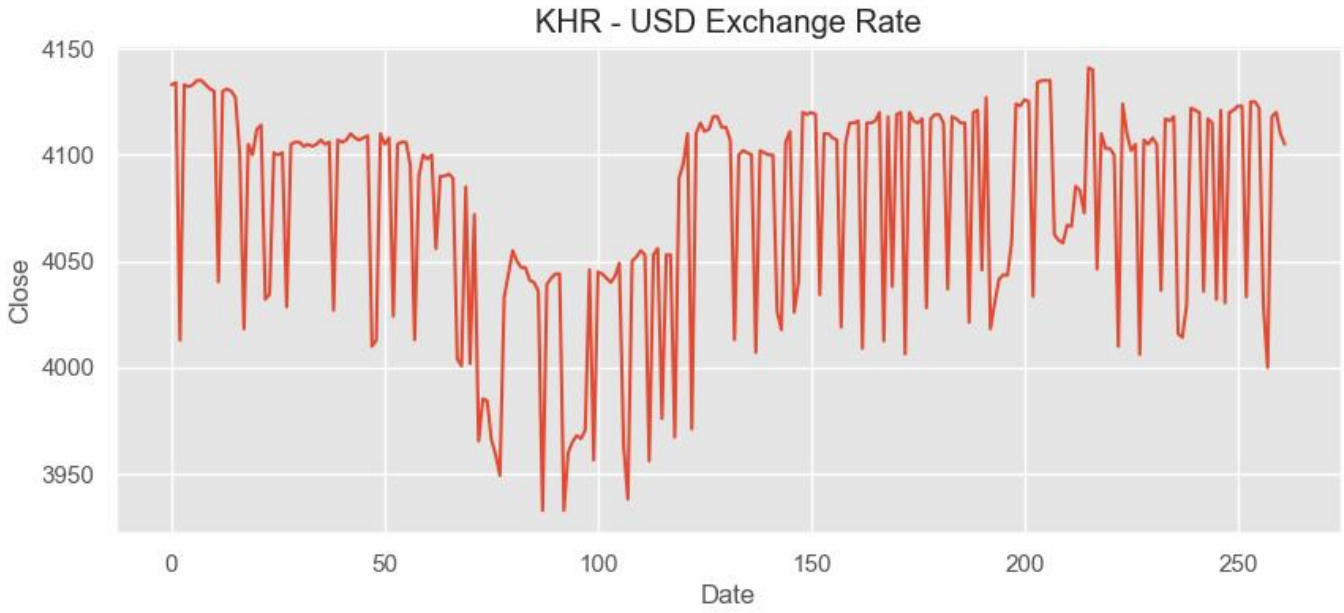
2.3. Descriptive Statistics

2.3.1. Summary Statistics

KHR to USD Exchange Rate

- Mean : 4075.864231
- Standard Deviation: 50.607815
- Minimum : 3932.716309
- Maximum : 4141.000000

2.3.2. Time Trends



The line graph illustrates the overall trend in KHR to USD exchange rates over the analyzed period. There are a number of factors that can affect the exchange rate between two currencies, including economic growth, inflation, and interest rates. In the case of the Cambodian riel and the US dollar, the exchange rate is also influenced by the flow of foreign investment and tourism.
The graph suggests that the Cambodian riel is a relatively stable currency. However, it is important to be aware that the exchange rate can fluctuate in the short term.

2.4. Conclusion

This exploratory data analysis provides valuable insights into the historical trends of KHR to USD exchange rates. It is very useful to keep tracking on the fluctuation of the trend. It might not the perfect prediction

because it is our first project in the data science field but it will lead us to make a better outcome for the next project.

III. Modeling and Evaluation

3.1. Regression By Using Machine Learning

3.1.1. Data Splitting

Train_test_split in **scikit-learn s**plits data into random train and test subsets. It takes arrays as inputs and you can specify the size of the test and train datasets. It also allows for reproducible output and stratified splitting. The split is stratified, with 80% of the data reserved for training the model and 20% for evaluating its performance.

3.1.2. Build Model

This prediction was made using the **LinearRegression** model, which is a technique that builds a tree-like model for predicting continuous variables.

3.1.3. Model Training

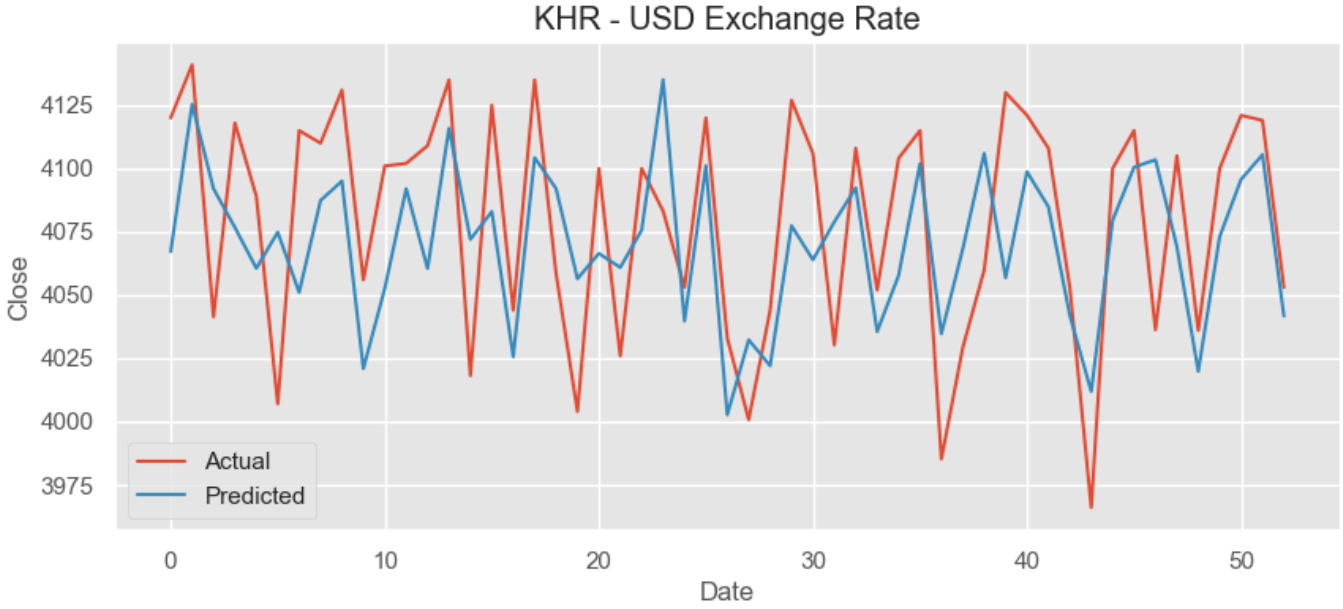
The next step involves training the **LinearRegression** using the training data (**xtrain** and **ytrain**)and the **fit** method is used to allow the model to learn the patterns and relationships within the data.

3.1.4. Prediction

After training, the model is used to make predictions on the test set (**xtest**). The predicted values are stored in the variable **ypred**.

IV. Conclusion

To concludes predictive capabilities of the **LinearRegression**, the actual vs. predicted exchange rates were visualized using Matplotlib to offer the understanding of how well the predicted model captures the historical trends in the actual data. Where the predicted line closely follows the actual line, it concludes that the model has successfully captured the historical trends present in the data. This tells us that the model has learned the underlying patterns in the exchange rate movements.



It's important to note that the actual and predicted lines may not overlap perfectly, Due to several ignored factors contributing the exchanging rate that has the unpredicted nature (supply and demand, Economic Recession, Inflation…….). So, it's important to know the limitations of the model's predictive

accuracy, by introducing the descriptive value such as mean absolute error and other several error indicator and accuracy percentage, etc.

```
MAE: 34.05611763717342
MSE: 1431.5542850975658
RMSE: 37.835886207376795
R2: 0.2827642346807775
Adjusted R2: 0.274386033043288
```