People’s Information Technology Program (PITP)

Affiliation: **MUET**

Institute: **Gextoninc**

Course: Data Science

Project: Gold Price Regression

Team members: Urooj and Shanza

Team leader: Urooj

Understanding the Gold Price Regression Dataset

Shape:

(2,290, 6) means we have 2,290 rows (records) and 6 columns (features).

Introduction:

Gold prices have historically been a strong economic indicator and are influenced by a variety of factors such as market indices, commodity prices, and currency exchange rates. Accurate forecasting of gold prices is valuable for financial analysts, investors, and policymakers. By using regression analysis, we aim to understand the relationship between the price of gold (the dependent variable) and several independent variables (predictors). These predictors could be market indicators, economic variables, commodity prices

Data Summary:

- The dataset spans financial information starting from 2008, covering significant commodity prices and exchange rates.
- All columns except for Date are of `float64` type, representing continuous numerical values.
- The Date column is currently stored as an object (string), which will be converted into integer for better data analysis.

This dataset contains daily financial data including major commodity prices and exchange rates from 2008 onward. It covers the following:

- SPX (S&P 500 Index): Daily closing values of this key U.S. stock market index.
- Gold: The price of gold per ounce in U.S. dollars.
- USO (U.S. Oil): U.S. oil prices, providing insight into the energy market.
- SLV (Silver): The price of silver per ounce in U.S. dollars.
- EUR/USD: The Euro to U.S. Dollar exchange rate.

Define the goal of the analysis or project:

- **Goal**: Predict the gold price based on related financial and economic indicators, such as SPX (S&P 500 index), USO (possibly crude oil), SLV (likely silver prices), and EUR/USD exchange rates.
- **Dataset Overview**: This dataset includes daily observations across several features, which may impact gold prices. This is a regression problem, where the target variable is **Gold**.

- **Usefulness**:
  - **Gold** is the target variable (dependent variable).
  - **SPX, USO, SLV, and EUR/USD** are related financial indicators that may influence or correlate with gold prices.
  - **Year, Month, and Day** provide time-based features that may reveal seasonal patterns or trends over time.

Data Sources and Preprocessing

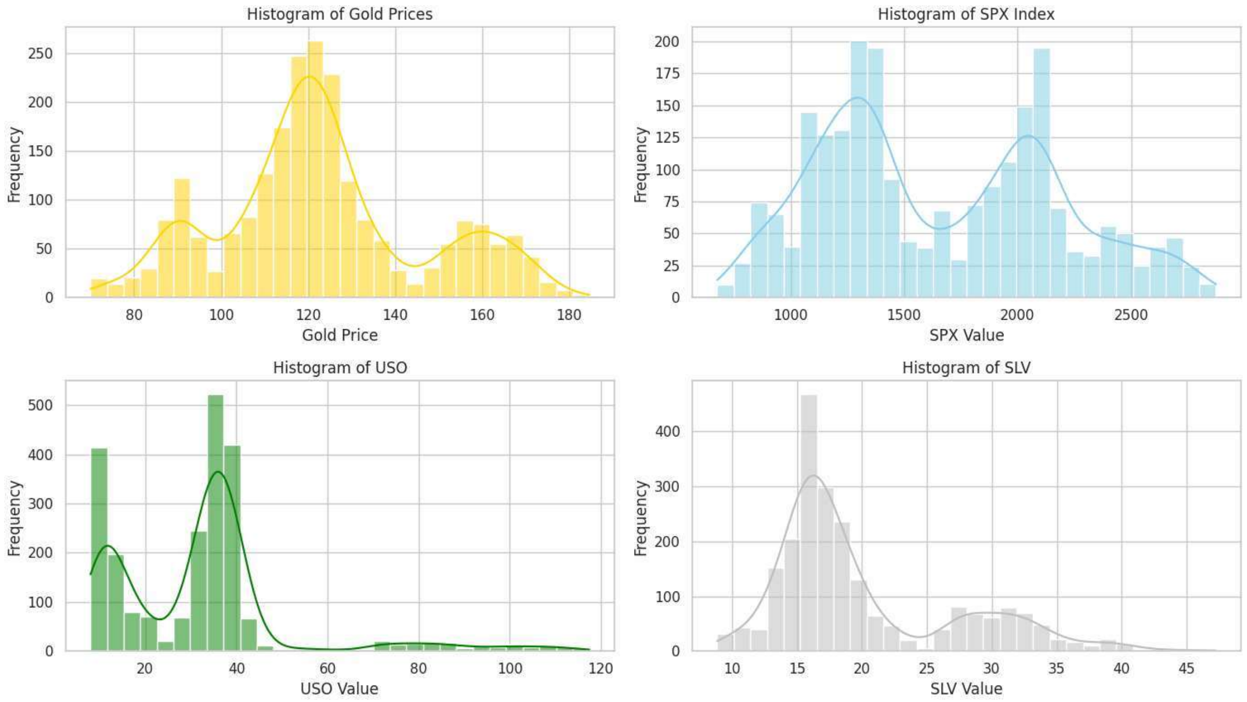
The dataset used for this project was obtained from historical records that include features such as the S&P 500 index (SPX), oil fund prices (USO), silver prices (`SLV`), and the EUR/USD exchange rate. The data spans from 2008 to 2019. The preprocessing steps included:

- **Loading and examining the dataset** to identify column names and types.
- **Renaming columns** for clarity.
- **Feature Engineering** Converting the `Date` column to datetime format and extracting `Year`, `Month`, and `Day` as separate features.
- **Handling missing values** by checking for null entries and ensuring data consistency.
- **Feature scaling** was applied using `StandardScaler` for numerical columns, ensuring the features were normalized for optimal model performance.

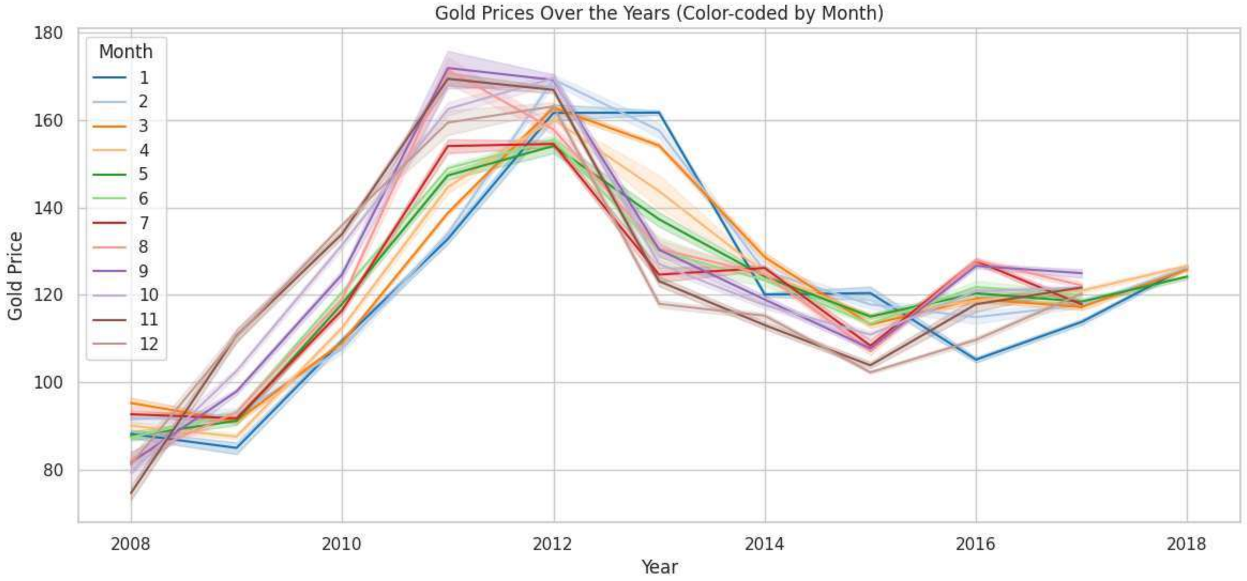
Exploratory Data Analysis (EDA)

Key findings from the EDA include:

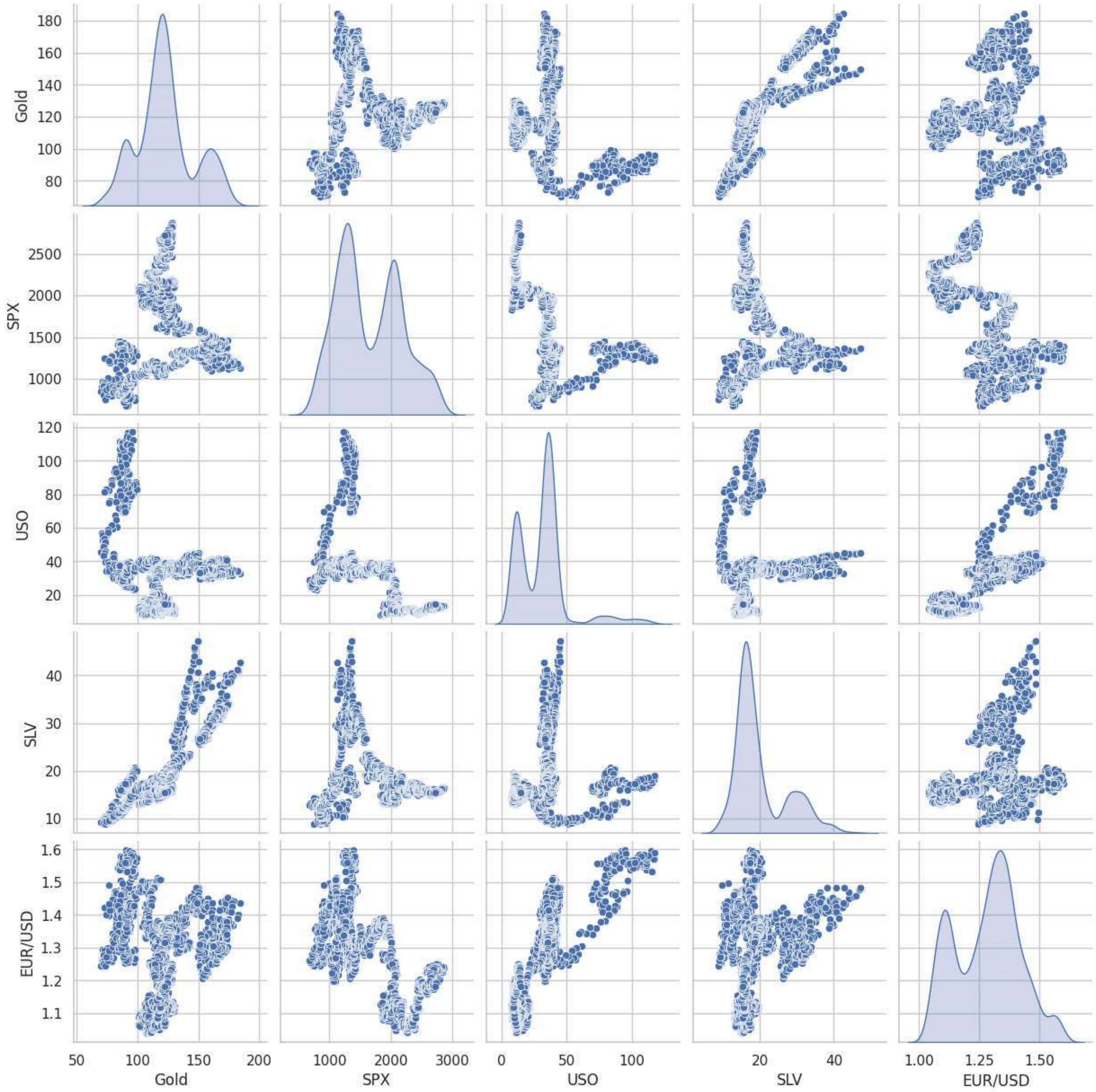
- **Distribution analysis**: The distribution of gold prices was visualized using histograms, revealing a right-skewed distribution.



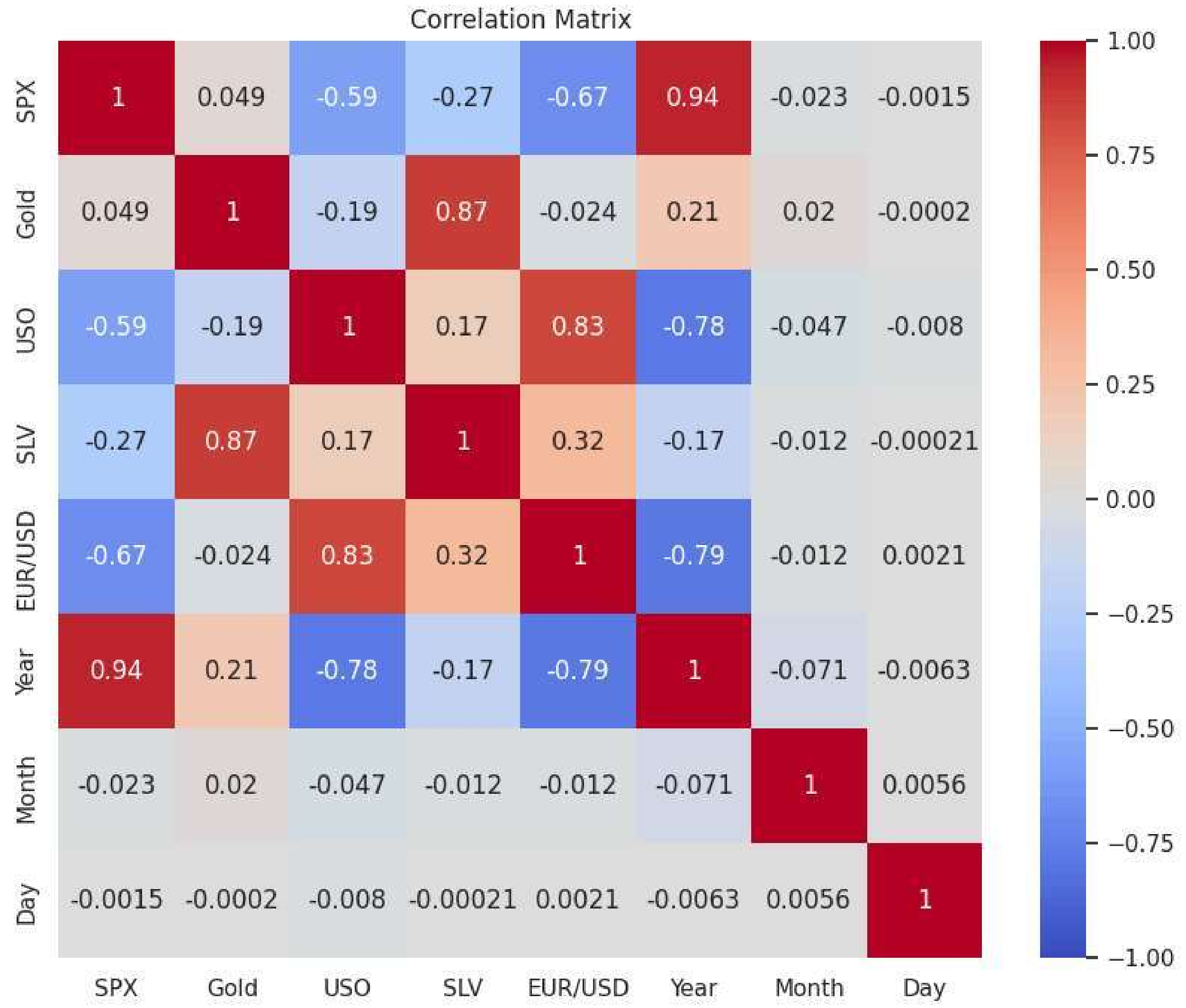
- **Trends over time**: A line plot showed trends in gold prices across the years, with a color-coded representation of months.



- **Pairwise relationships**: A pair plot was created to identify correlations between key variables such as `Gold`, `SPX`, `USO`, `SLV`, and `EUR/USD`.



- **Correlation matrix**: A heatmap was used to visualize correlations, showing that `SLV` and `USO` had stronger relationships with `Gold` than the S&P 500 index or EUR/USD.



Modeling Process

Sr.no	Models name	Best parameteres	Mean Squared Error	R² Score
1.	Linear Regression	-----	45.33531404842033	0.9173165736364479
2.	Ridge Regression	{'alpha': 0.1, 'max_iter': 1000,	45.33224098188889	0.9173221783548798

		'solver': 'sparse_cg', 'tol': 0.001}		
3.	Support Vector Regression (SVR)	{'C': 10, 'gamma': 'auto', 'kernel': 'rbf'}	23.20632129265635	0.9576758604623112
4.	Decision Tree Regression	{'max_depth': 10, 'min_samples_leaf': 1, 'min_samples_split': 2}	5.065377207381185	0.9907616666582966
5.	Random Forest Regression	{'max_depth': None, 'min_samples_leaf': 1, 'min_samples_split': 2, 'n_estimators': 200}	2.1418154872657116	0.9960937153112802

Multiple models were evaluated in this project:

1. **Linear Regression**: This baseline model showed moderate predictive performance.
2. **Ridge Regression**: Used to handle potential multicollinearity, with hyperparameters tuned using Grid Search. The best-performing model had parameters optimizing the balance between bias and variance.
3. **Support Vector Regression (SVR)**: Applied to capture nonlinear relationships. Grid Search was used for hyperparameter tuning.
4. **Decision Tree Regression**: Explored for its simplicity and ability to capture non-linear patterns, with hyperparameters like `max_depth` and `min_samples_split` tuned.
5. **Random Forest Regression**: Used for its ensemble learning properties, enhancing model robustness. Hyperparameters such as `n_estimators` and `max_depth` were optimized using Grid Search.

Performance Metrics:

- Mean Squared Error (MSE) and R-squared (R^2) scores were used to evaluate models.
- The best-performing model was determined based on the highest R^2 score and the lowest MSE.

According to results Random Forest Regressor is used in our project.Because it contains higher accuracy.

Conclusions

The project successfully demonstrated the use of various regression models for predicting gold prices. Key findings include:

- **Ridge Regression and Random Forest Regression** showed the most promise due to their balance of bias and variance.
- **SVR** captured some non-linear patterns but required more computational resources.
- **Feature importance** analysis from Random Forest indicated that silver prices (SLV) and oil prices (USO) were significant predictors of gold prices.

Future Work:

- Incorporate more recent data to enhance model reliability.
- Explore other economic indicators such as geopolitical events or inflation rates.
- Apply deep learning models to compare their performance with traditional machine learning approaches.