## horizontal line



Gold Price Prediction

2025, October 12

**─**

Group 5:

Khanh Linh Le (Student ID: 126441237)

Michelle Cong (Student ID: 112582218 )

Course: BDA500 - Machine Learning

Group Project 1

# Overview

Gold is one of the most valuable and widely traded commodities in the world. Its price is influenced by many factors, including market demand, global economic conditions, currency fluctuations and investor sentiment. Accurate prediction of gold prices is essential for investors, policymakers, and analysts to make informed decisions. Gold is priced globally, a weakening dollar also makes holding dollar-denominated assets less attractive, pushing capital toward gold. Persistent inflation erodes purchasing power of flat currencies, so gold acts as a hedge, Conflicts (e.g. Eastern Europe, Middle East) supply chain disruption and trade tension, heighten risk perceptions and push investors into gold

This project aims to build a machine learning model that predicts gold prices based on historical trading data, using regression techniques to identify trends and relationships between key variables such as opening and closing price, highest and lowest price, and trading volume.

# Objective

The main objective of this project is to develop and evaluate predictive models that can accurately estimate the closing price of gold based on past market data.

### **Specific goals:**

* Perform **data cleaning and preparation** on historical gold price data.
* Conduct **Exploratory Data Analysis (EDA)** to understand trends, patterns, and correlations.
* Implement **Linear Regression** and **Ridge Regression** models to predict gold closing prices.
* Evaluate model performance using appropriate **metrics** such as MAE, MSE, RMSE, and R².
* Compare the performance of both models and discuss which provides better predictive accuracy.

# Method

* Dataset: Historical gold price dataset (Open, High, Low, Close, Volume) from Kaggle
* Splits: 80% training, 20% testing
* Features: Date (converted to datetime), Open, High, Low, Volume -> predict Close
* Pipeline: Preprocessing (missing value handing, scaling) -> model fitting -> evaluation
* Models:
  + Linear Regression
  + Ridge Regression
  + Polynominal Regression (degree = 2)
  + Decision Tree & Random Forest (with 5-fold CV)
* Tuning Strategy: Grid search over regularization strength Ridge and tree depth

# Results

|  |  |  |  |
| --- | --- | --- | --- |
| Model | MAE | RMSE | R2 |
| Linear Regression | 3.726056 | 5.699283 | 0.999892 |
| Ridge Regression | 3.726057 | 5.699281 | 0.999892 |
| Decision Treee | 5.896121 | 9.386801 | 0.999706 |
| Random Forest | 5.031080 | 7.497811 | 0.999812 |
| Polynominal  (deg = 2) | 3.7365 | 5.8258 | 0.9999 |

# Discussion:

The Linear and Ridge Regression models slightly outperformed the others, achieving the lowest error metrics and highest R2, suggests that the relationship between features and gold price is primarily linear. Ridge Regression offered minimal improvement over linear regression due to low overfitting risk in this dataset. Polynomial and decision Tree models introduced slight overfitting without accuracy gains.

Error Analysis: Minor deviations likely result from short-term market volatility not captured in historical features.

Risks/Bias: the dataset lacks macroeconomic indicators (e.g. inflation, USD strength)

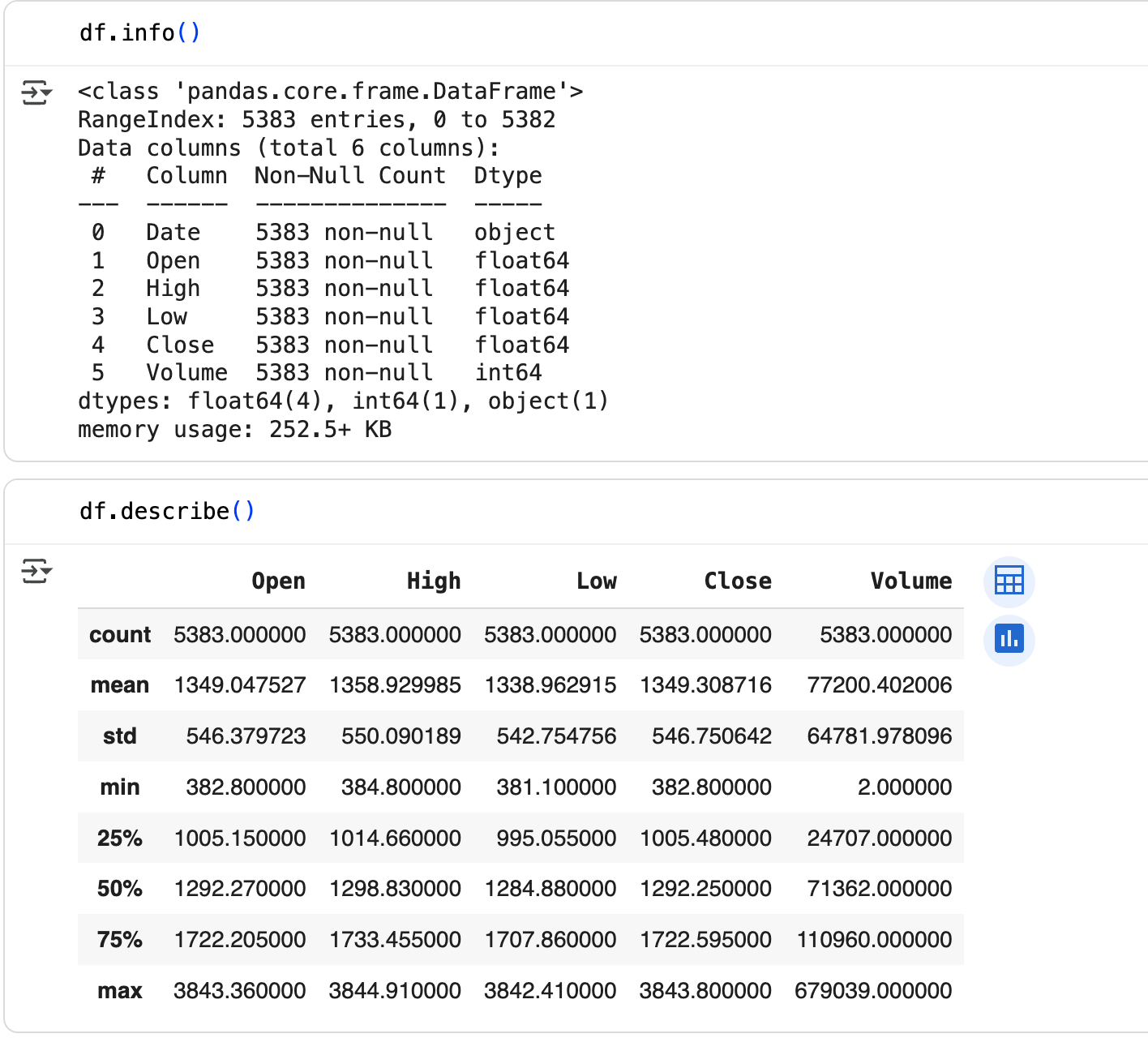
Assumptions: Prices depend solely on past trading variables, external shocks are ignored

# Responsibility Table

|  |  |  |
| --- | --- | --- |
| **Topic / Task** | **What You Did (1–2 lines)** | **Group Member Responsible** |
| **Coding Tasks** |  |  |
| – Specific Models | Implemented and trained Linear Regression and Ridge Regression models to predict gold prices. | Khanh Linh Le |
| – Pipelines | Built the preprocessing and model pipeline for reproducibility and scaling. | Khanh Linh Le |
| – EDA | Performed exploratory data analysis (EDA) with summary statistics and visualizations of price trends. | Khanh Linh Le |
| – Feature Engineering | Converted Date into datetime format, extracted time-based features, and handled missing values. | Khanh Linh Le |
| **Report Writing** |  |  |
| – Methods | Described dataset preprocessing, modeling approach, and evaluation metrics. | Khanh Linh Le |
| – Results | Reported and compared Linear vs Ridge model performance (R², MAE, MSE, RMSE). | Khanh Linh Le |
| – Discussion | Discussed the strengths and weaknesses of both models and possible improvements. | Michelle Cong |
| – Limitations | Mentioned potential issues such as limited features and lack of external factors (e.g., inflation). | Michelle Cong |
| **Plots & Visuals** |  |  |
| – EDA Plots | Created scatter and line plots to visualize price movement over time | Khanh Linh Le |
| – Comparison Bar Chart | Plotted comparison of evaluation metrics between Linear and Ridge models. | Michelle Cong |
| – Residuals | Generated residual plots to assess model fit and variance. | Michelle Cong |

# Part A - Data Preparation, EDA, Categorical Handling

## Check basic information

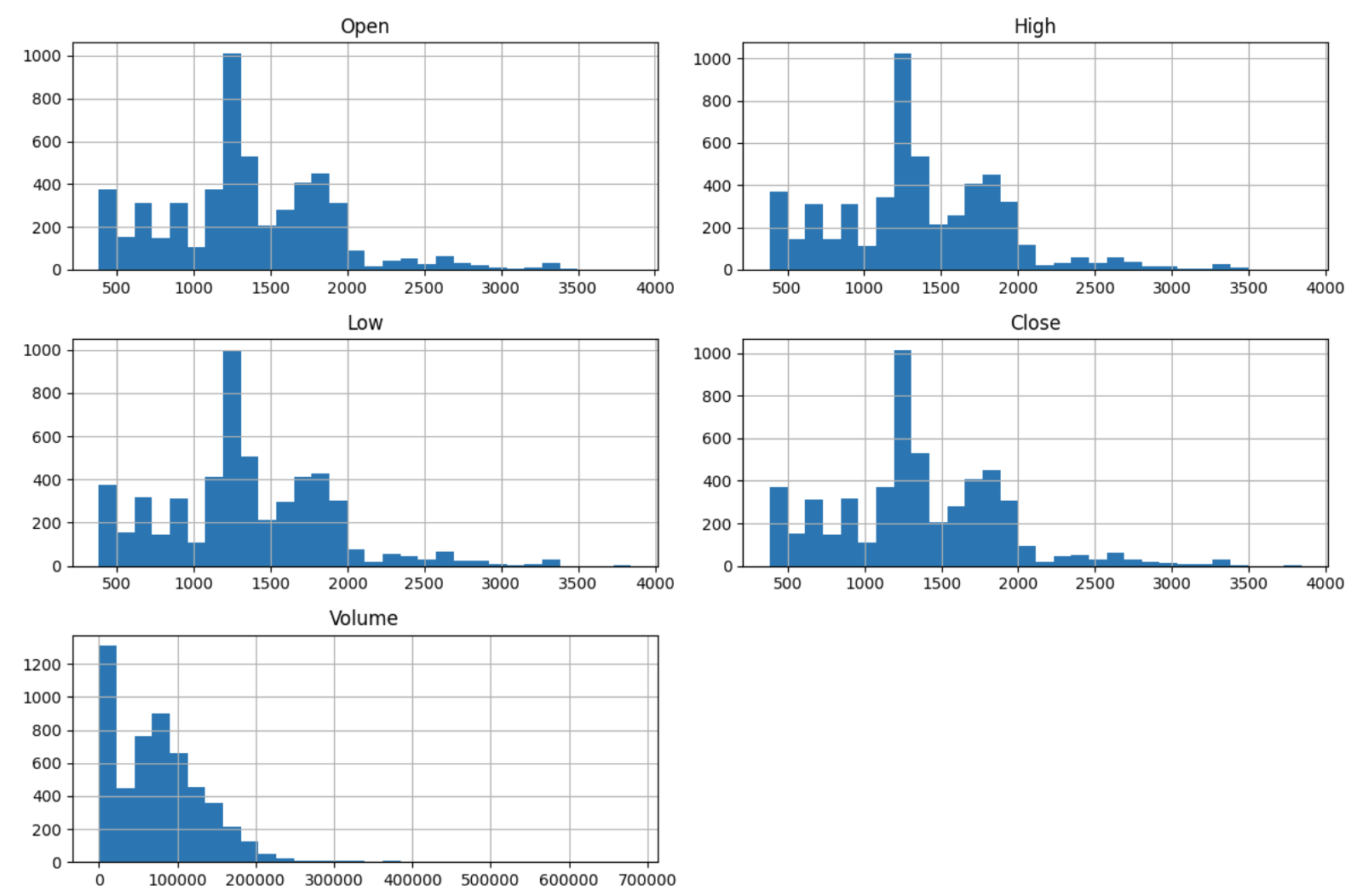


## Exploratory Data Analysis (EDA)

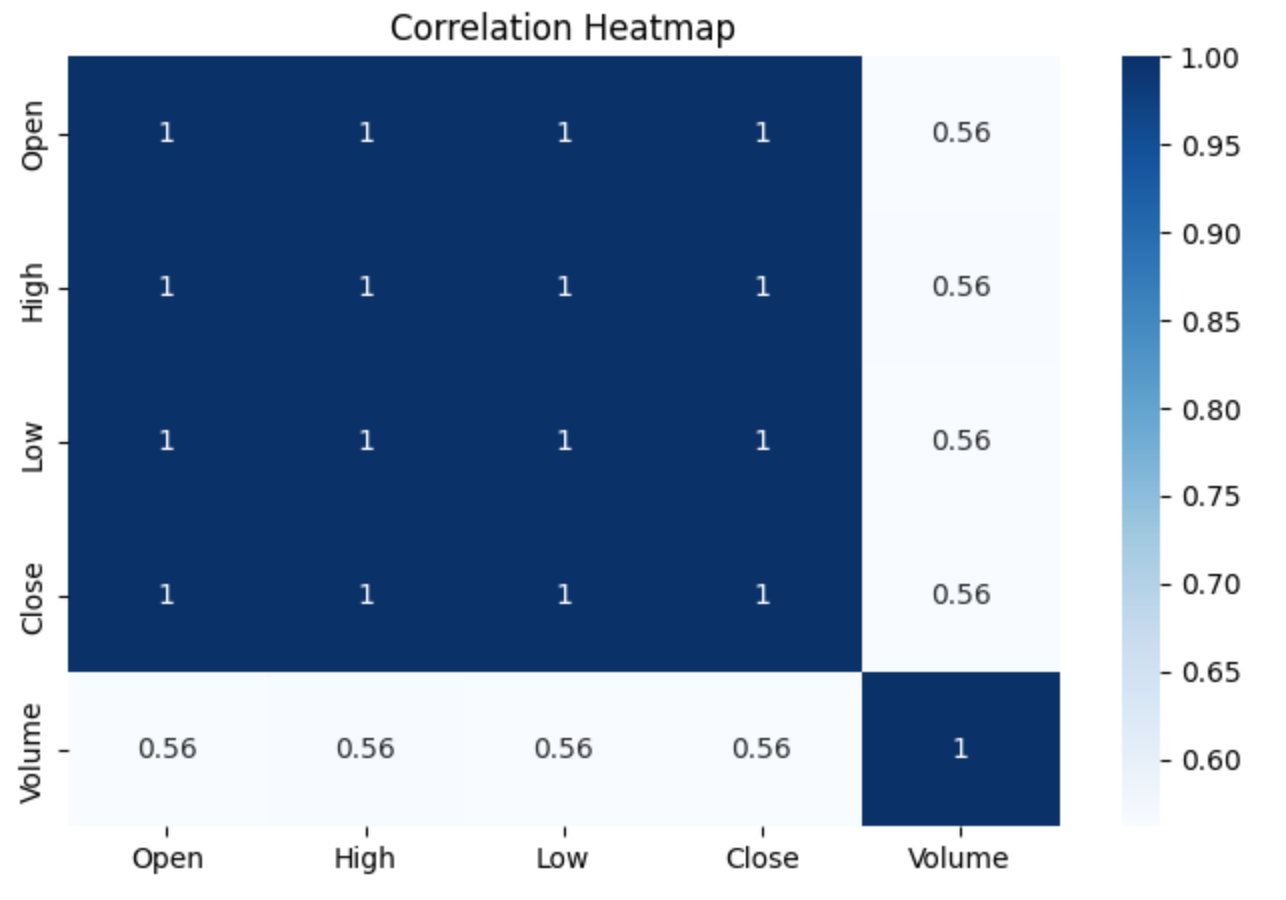
### High Prices Over Time



### Histograms of numeric features

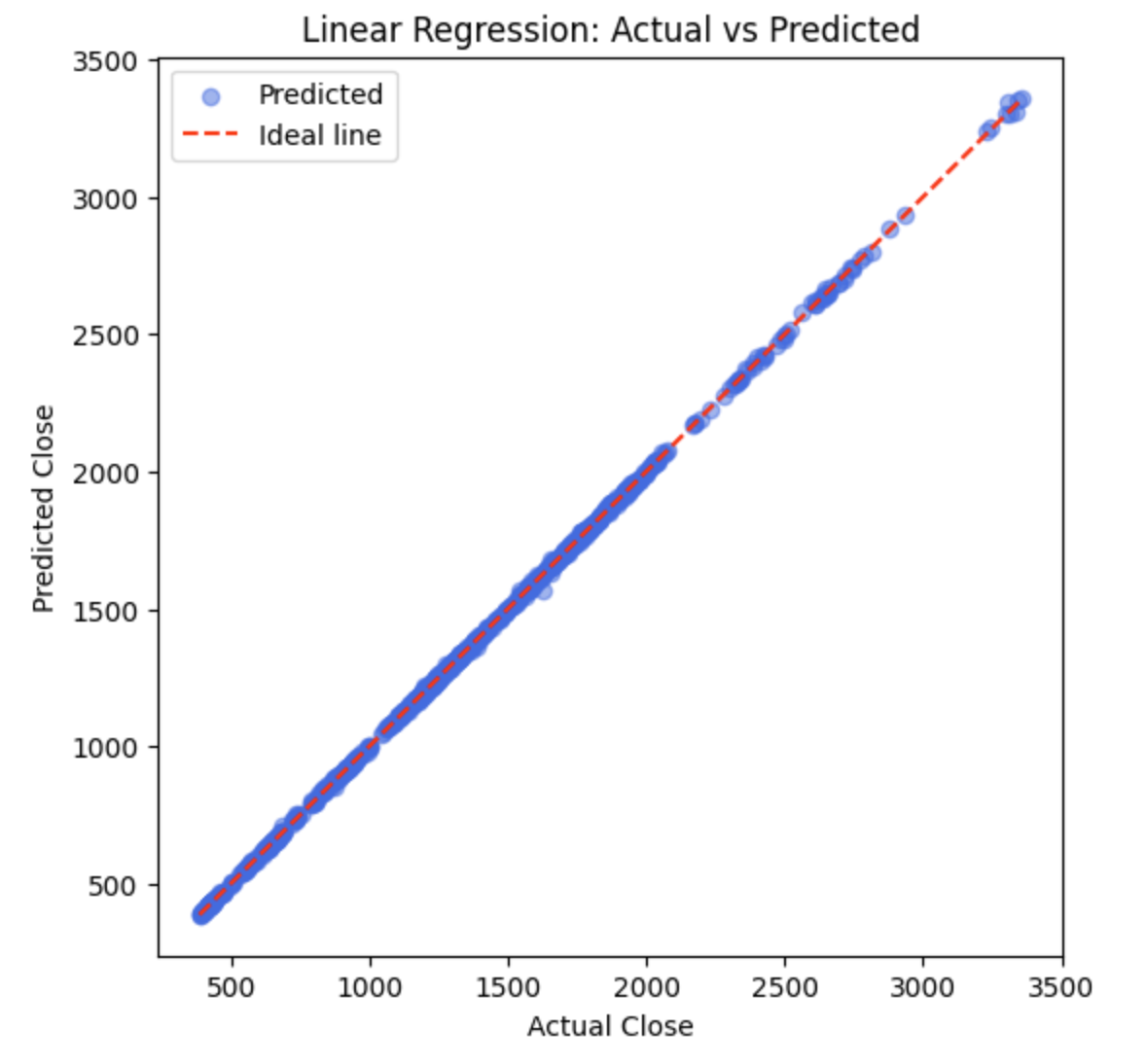


### Correlation between numeric features



# Part B - Modeling

### Linear Regression Model



Linear Regression Metrics:

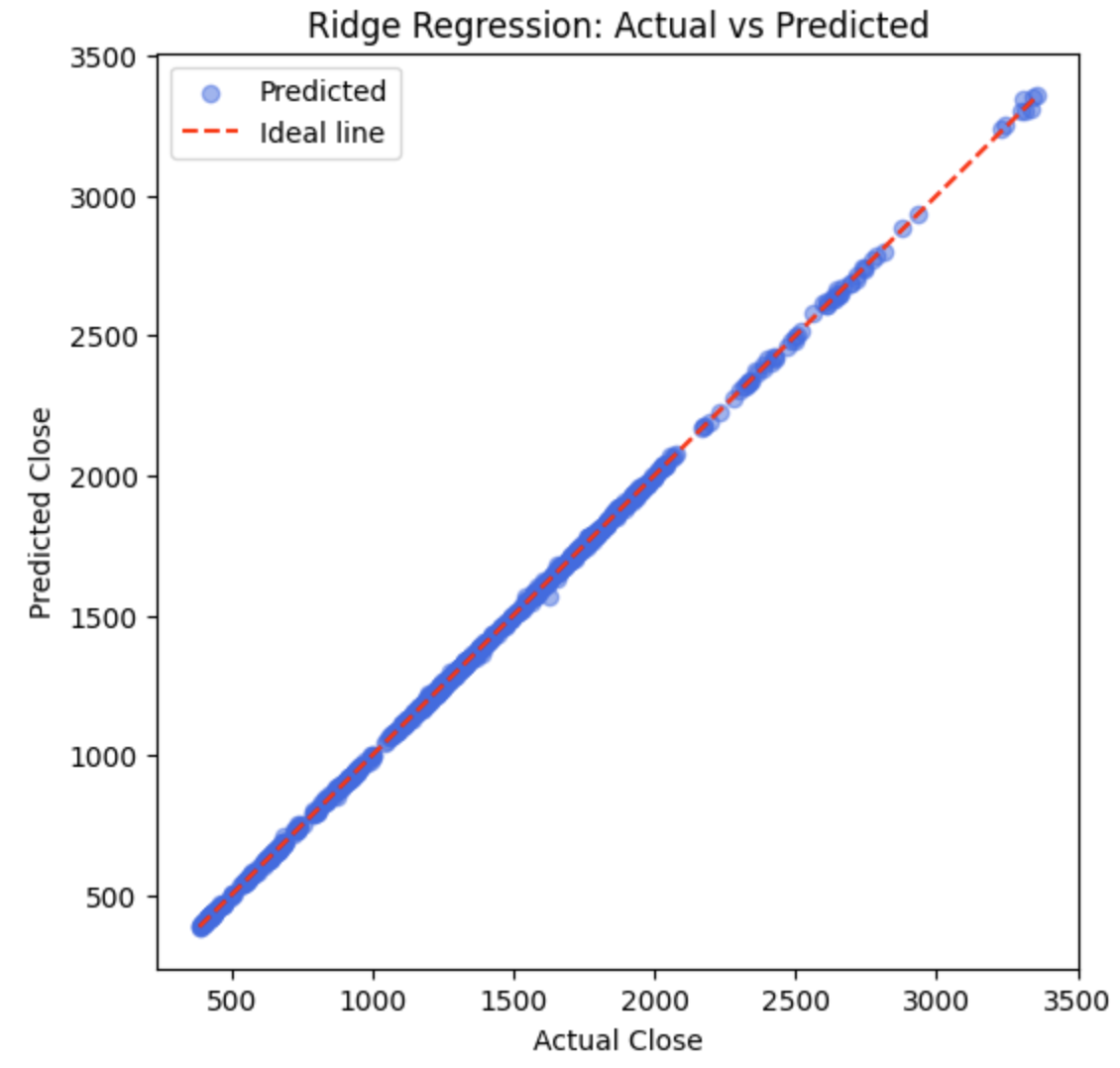
MAE: 3.7261

MSE: 32.4818

RMSE: 5.6993

R²: 0.9999

### Ridge Regression Models



Ridge Regression Metrics:

MAE: 3.7261

MSE: 32.4818

RMSE: 5.6993

R²: 0.9999

### Polynomial Regression (degree = 2)

A graph of a graph

AI-generated content may be incorrect.

Polynomial degree=2 Regression Metrics

MAE: 3.7365

RMSE: 5.8258

R²: 0.9999

### Residuals

A chart with blue dots

AI-generated content may be incorrect.

# Part C -

1. Decision Tree: K-Fold CV

A diagram of a company structure

AI-generated content may be incorrect.

1.2 Predicted vs Actual Plot

A graph with blue dots and red dots

AI-generated content may be incorrect.

A graph with blue dots

AI-generated content may be incorrect.

1.4 RandomForestRegressor

A graph with a line

AI-generated content may be incorrect.

Best Random Forest parameters: {'n\_estimators': 100, 'min\_samples\_split': 2, 'min\_samples\_leaf': 2, 'max\_features': 0.8, 'max\_depth': 10}

MAE: 5.0311

RMSE: 7.4978

R²: 0.9998

1.5 Model comparision

A graph showing a number of different colored bars

AI-generated content may be incorrect.

A screenshot of a graph

AI-generated content may be incorrect.