Introduction: Why Gold?

For thousands of years, gold has stood as one of the most sought-after metals on the planet. The metal holds an immense number of applications in the modern world, from jewelry to electronics to coins, and its influence on historical cultural, economic, and technological developments is undeniably substantial. Gold also represents a timeless and universal store of value, as its stability throughout periods of economic volatility has made it one of the most common "safe haven" assets throughout history.[1] Because of its ability to retain its value, investors tend to purchase gold to hedge against market downturns or high-inflation environments, making it one of the most traded assets in the world.[2] As well, gold has historically experienced low or negative correlation with other common assets such as equities, bonds, and real estate.[3] This makes it an attractive investment for portfolio diversification, but consequently, a difficult asset to predict.

As a result, understanding the factors that influence gold prices is critical for any investor looking to better understand the dynamics of financial markets and the global economy. Understanding the relationships gold holds with other assets can help investors make better investment decisions and optimally diversify their portfolios. Additionally, the ability to use movements within other assets to predict gold returns can also benefit gold investors and help provide insight into the future state of the economy.

Therefore, the goal of this project is not only to understand what relationships exist between the movements of other assets and gold, but to also seek an effective predictive model for modelling future gold price movements.

Dataset Description

The dataset chosen is the Gold Price Prediction Dataset sourced from Kaggle. The various trading prices and metrics of gold, proxied by the exchange-traded fund (ETF) SPDR Gold Shares (NYSEARCA: GLD), and 14 other different assets were tracked between December 15, 2011, to December 31, 2018. There are 1718 observations (rows) and 81 variables (columns).

The full list of variables listing the specific asset name and their respective metrics tracked, as well as descriptions of the different price levels tracked for each asset are located in Appendix 1.

Our response variable would therefore be the Adjusted Close price of the SPDR Gold Shares ETF. The adjusted closing price is used because it is typically considered as the best way to measure the price of an asset over a long period of time, given that it adjusts for corporate actions that may distort the ETF's price. As a result, the adjusted close provides a relatively accurate representation of the asset's price over time compared to the other price levels.

[1] https://www.cbsnews.com/news/why-gold-is-a-safe-haven-asset/

[2] https://www.gold.org/goldhub/data/gold-trading-volumes

[3] https://www.forbes.com/2010/03/30/gold-dollar-correlation-intelligent-investing-asset-allocation.html?sh=4fc2d81e22b7

The SPDR Gold Shares ETF is also chosen as a proxy to represent gold prices, because commodities are typically purchased through investing in derivatives and funds rather than bought outright in their physical form. This ETF represents the largest physically backed gold ETF in the world, and analyzing the prices of this fund can provide a good representation of the factors that impact the value of a gold investment in an average investor's portfolio.

Methods

Data Pre-processing

To prepare the data for analysis, the data was imported and the variables were specified to represent their appropriate variable types. As listed in the previous table, all the Open, High, Low, Close, Adjusted Close, and Price columns across all the assets were specified to be numeric, all Volume columns were specified to be of integer type, and the Trend columns were imported as factor types. The clean_names() function from the *janitor* package was also used to reformat all variable names due to the high number of spaces and capital letters present in many variable names. As well, given that the response variable is solely the adjusted closing price of gold, and since the goal is to evaluate the relationships that other assets have with gold, the other characteristics of gold prices (i.e. Open, High, Low, Close, Volume) were excluded.

A correlation analysis also identified the Close and Adjusted Close prices of the USO (United States Oil ETF) and DJI (Dow Jones Index) variables to be identical. The Adjusted Close columns for those two assets were also excluded accordingly.

In asset price modelling, logarithmic returns are typically the preferred measure used to assess the performance of an asset. This is due to its additive properties across time periods and its adherence to normality.[4] Converting the prices of the assets into logarithmic returns will also eliminate the issue of varying scales across all the different assets. For example, the Dow Jones Adjusted Close ranges from 11,766 to 26,828, while the EURUSD exchange rate ranges from 1.039 to 1.393. This difference in scale does not imply a difference in intrinsic value, and therefore, converting the prices to represent asset returns will help standardize the variables into a common scale. Overall, since the project goal is to evaluate the *movements/returns* in the price of gold rather than the price itself, a log differencing transformation was applied to all the numeric price-level variables in the dataset, including the response variable, in order to transform all the relevant variables into log returns. The response variable is therefore the log returns of gold prices (SPDR Gold ETF).

No missing values were identified. The import and pre-processing code is found in Appendix 2.

Exploratory Analysis Methods

To gain a quick overview into the numeric distributions and price ranges for each variable, identify potential outliers, and assess the balance within each category (0 and 1) for the binary Trend variables, the **summary()** function was used. **Histograms** of each asset's adjusted close price were also plotted to verify lognormality. Under the assumption of efficient markets, it is

[4] https://quantivity.wordpress.com/2011/02/21/why-log-returns/

theorized that asset returns should generally follow a normal distribution, because asset prices are hypothesized to behave like a random walk and the resulting movements of assets tend to approximate a symmetric distribution around mean 0.[5] Ensuring lognormal distributions also helps fulfill assumptions required for some parametric models.

To further check the distributional properties of the response variable, characteristics such as **skewness and kurtosis** were also calculated for the transformed response. A **spectral density plot** was also fit to analyze the underlying periodic trends of the original, non-transformed price levels of gold.

Afterwards, **time series plots** were fit to the original, non-transformed adjusted closing prices of each asset, as this would provide an overview into how the different asset prices changed over time and which assets may exhibit similar patterns with gold over time.

Main Data Analysis Methods

As the project goal is to develop a model that is both interpretable and holds high predictive power, the data will first be split into training and testing datasets based on an 80/20 split (Appendix 3).

Linear regression will serve as the base model for comparison, which will help initially identify what variables may be most significant in the dataset. However, given the high number of predictors and the high likelihood of multicollinearity, **LASSO regression** with a cross-validated penalty term will help conduct variable selection and reduce the impacts of collinear variables, which can overall aid in increasing interpretability of the model. In the case that the relationships between the predictors and the response is non-linear, **Generalized Additive Models (GAM)** will be fit with all the adjusted close prices of each asset. A **random forest model** will also be considered for modelling non-linear relationships, although this comes at the cost of weaker interpretability.

To evaluate the performance of each model, the mean-squared error (MSE) will be used as the primary evaluation metric.

Additionally, due to the time series characteristics of the dataset, a time series model can be fit to assess whether the response variable itself can explain its future values. While this does not directly align with the goal of observing how the movements in other asset classes impact movements within the response, it can help provide insight into whether logarithmic returns in gold exhibit any predictive patterns. An ARMA(p, q) model can help show the relationship between current gold return values and its past movements.

[5] https://www.investopedia.com/terms/r/randomwalktheory.asp#:~:text=Random%20walk%20theory%20claims%20that,efficient%2C%20reflecting%20all%20available%20information/

Results

Exploratory Data Analysis

The numeric summary and histograms of each variable showed all numeric variables, including the response variable, adhering strongly to normal distributions following the log differencing transformation. Some variables showed very similar distributions with one another given the investment type / asset class. For example, the S&P 500 Index and Dow Jones Index were distributed very similarly, given that both their prices are based on the prices of the most traded equities in the U.S. Some assets had a much wider spread in their distributions, such as NYSE:EGO, while others had very narrow spreads, such as the EURUSD exchange rate. Assets with a wider distributional spread indicate higher historical daily volatility, while those with narrower spreads indicate less fluctuation. The remaining assets showed consistent spreads of around (-0.1, 0.1), meaning that most assets experience daily fluctuations in value between -10% and 10% over the 7-year period, and all variables had a center at around 0. No notable outliers were identified. This is further reflected by the equal balance of observations across both classes for the binary Trend variables, indicating that there were a relatively equal number of days where the asset moved upwards compared to downwards.

The skewness of the response was identified to be -0.34157, and kurtosis was calculated to be 9.07283. The negative skewness value indicates a slightly longer left tail and some asymmetry, and the large kurtosis value indicates heavier-than-normal tails. Overall, these metrics mean that daily log returns have tended to be trend more negatively between 2011 – 2018, and there exist more extreme return values than a typical normal distribution. The smoothed periodogram also shows a cyclical pattern with a frequency of

In the time series plots, assets of similar classes showed similar patterns (Crude Oil WTI/USD and Brent Oil Futures, S&P 500 and Dow Jones, etc.). However, two notable assets that showed very similar patterns to gold prices were NYSE:EGO and Platinum Futures.

All plots included in this exploratory data analysis section are included in Appendix 3.

Main Data Analysis

Linear Regression

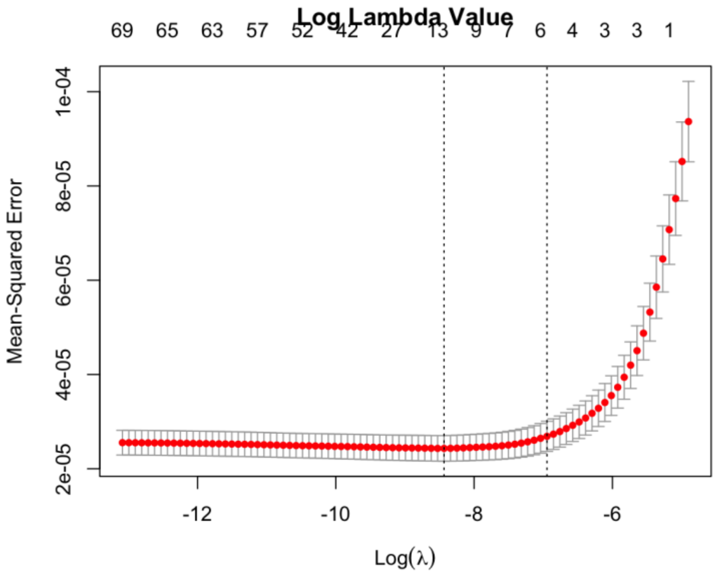
In the fitted linear regression model, most variables (62 / 72) were not significant at the $\alpha = 0.05$ significance level, but the overall model had an Adjusted R-squared value of 0.7591 and was statistically significant. This indicates relatively high multicollinearity present, which is expected given the high number of similar predictors in the model. The resulting MSE was $2.3643 * 10^{-5}$.

At first glance, the variables with the lowest p-values were *plt_price* and *sf_price*, indicating that movements within platinum future prices and silver futures prices may have the greatest statistically significant effects in explaining the response. However, more models and analysis would need to be completed to prove this.

LASSO

A LASSO (L1 regularization) model was fit to solve the multicollinearity issues identified in the previous linear regression model and reduce the risk of overfitting through shrinking the coefficients of less important variables towards zero. 10-fold cross-validation was also applied to search for the optimal shrinkage parameter based on the MSE criterion, which is the value that provides the best trade-off between model complexity and fit.

The optimal parameter calculated was 0.00021717. As this value is extremely small and close to zero, very little shrinkage was applied to the coefficients, meaning that the coefficients do not need to penalized from their original values by much.



At the α = 0.05 significance level, the 12 following coefficients were selected by the algorithm:

Variable Name	**LASSO Model Coefficients**
(Intercept)	0.00035292
dj_high	-0.09345237
dj_close	-0.02901735
of_high	-0.00327185
sf_price	0.1809317
sf_open	-0.01396292
sf_volume	-8.776219 x 10^{-9}
usb_price	-0.03119
plt_price	0.1821865
plt_low	0.02122521
usdi_price	-0.2035144
gdx_high	0.001055361
gdx_close	0.1658254

The sparsity of the model appeared to benefit the predictive performance of the model, as MSE was reduced slightly to 2.2871 x 10^{-5}. The largest coefficients in the model are *sf_price*, *plt_price*, *usdi_price (US Dollar Index)*, and *gdx_close (Gold Miners ETF)*. We see further evidence that changes in silver futures prices and platinum futures prices have significant effects on gold price returns. The log returns of the US Dollar Index had the largest absolute coefficient of -0.2035, meaning that for each percent increase in log returns of the USDI, the log returns of gold prices are expected to decrease by 0.2035% on average, assuming all other predictors are held constant.

Generalized Additive Models (GAM)

To assess whether a non-linear model would better suit the relationships between the predictors and the response, a GAM was fitted by applying smoothing functions to the adjusted close prices of the 14 asset classes.

The model summary only lists the intercept as a parametric coefficient, meaning that no parametric terms are included for any of the other predictors and smooth functions are exclusively used for modelling the relationships between the predictors and the response. The model further identifies 7 significant smoothed predictors in the model with the following effective degrees of freedom and relationship direction:

Smoothed Predictors	**Effective Degrees of Freedom (EDF)**	**Direction**
s(of_price)	1.494	Negative
s(sf_price)	1.811	Positive
s(usb_price)	1.758	Negative
s(plt_price)	8.943	Positive
s(usdi_price)	5.839	Negative
s(gdx_adj_close)	1.000	Positive
s(uso_close)	3.387	Positive

The effective degrees of freedom of each variable indicates the complexity of the smooth. The higher the EDF, the more complex the function fit. We see that the log returns of the adjusted close prices of GDX (Gold Miners ETF) is completely linear (1 degree of freedom), while platinum prices have 8 degrees of freedom, showing a highly polynomial relationship. Additionally, the degrees of freedom for returns of Brent crude oil futures, silver futures, and the U.S. 10-Year Treasury note are small, indicating a relationship that is relatively straight, while the log returns within the U.S. Oil ETF and the U.S. Dollar Index are relatively cubic and quintic/sextic respectively.

Plots of the pairwise polynomial relationships between the significant predictors and the response also show that the confidence interval bands are most narrow for silver futures, platinum futures, and GDX adjusted close. This indicates a more precise estimate of the smooth function and lower variability within the predictors. Therefore, these three variables provide more reliable estimates of gold log returns.

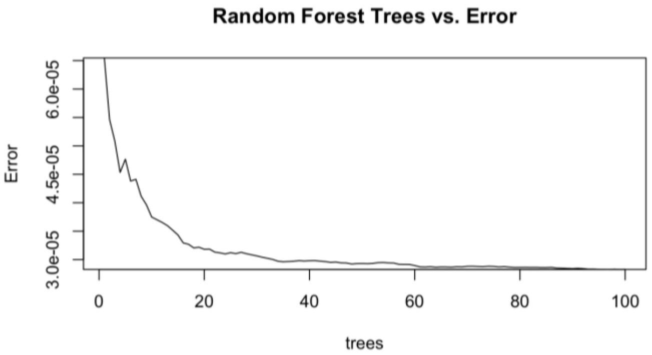
In terms of the direction, the Brent crude futures U.S. Dollar Index, and U.S. Treasury bonds hold negative linear relationships with the response. This means that as the log returns of these three predictors increase, the log returns of gold prices will tend to decrease as a result. On the other hand, when the log returns of silver futures, Gold Miners ETF, U.S. Oil ETF, and platinum futures increase, the log returns of gold increases on average. For example, the silver futures and GDX ETF adjusted close smooth plots illustrate a linear relationship, where both function look to be around $y = \frac{1}{5}x$. For each 0.01% increase in the log returns of silver futures or GDX ETFs, gold log returns will increase by around 0.002%. This relationship is reasonable, as there exists historical statistical evidence that gold and silver tend to move in tandem over time, with silver prices being more volatile than gold.[6] As well, the GDX ETF is directly based on the performance of gold, as the fund seeks to replicate companies within the gold mining industry.[7] Overall, this directly aligns with our conclusions for the smooth plot.

An interesting plot to note is the relationship between gold and platinum futures. The graph shows that logarithmic gold returns do not change much if the daily log returns for platinum futures vary between -0.03% and 0.03%. However, the response starts to increase drastically when platinum returns begin to exceed 0.03% and decreases rapidly in the opposite direction when platinum returns fall below -0.03%. This indicates the existence of a relationship between gold returns and the volatility of platinum futures; the higher the platinum futures volatility, the more likely gold will similarly move in an extreme way.

The MSE of this model was 2.2007 x 10^{-5}, an improvement from linear regression and LASSO.

Random Forest Model

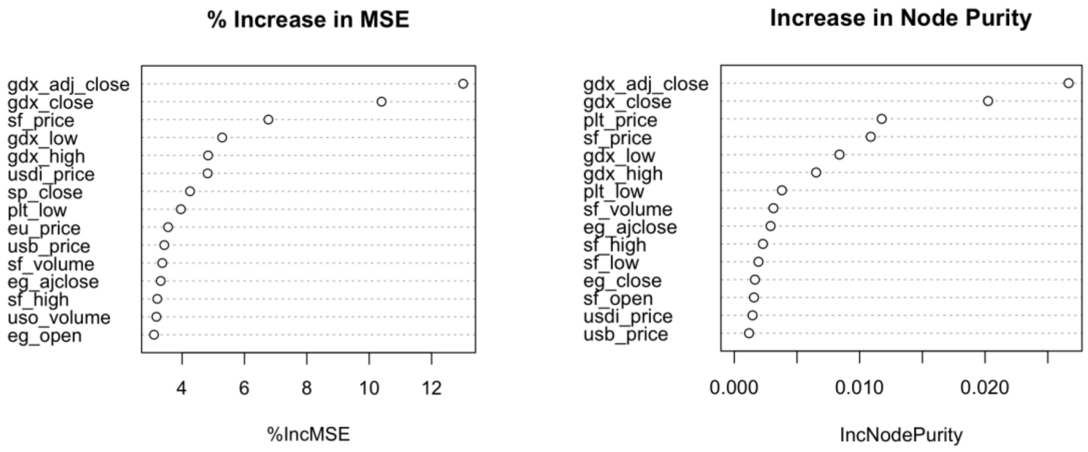
Given the strong predictive performance of the GAM, a random forest model was also fit to assess whether modelling non-linear relationships between the predictors and the response in the dataset would yield improved performance. 100 trees were used to fit the forest, with the errors plateauing at around the 50-tree mark.



Variable importance plots show that the logarithmic returns based on the adjusted close and close price of GDX (Gold Miners ETF) contributed most to improving model performance and predictive power. The left-side plot shows which variables contribute most to improving the

[6] https://www.morganstanley.com/articles/investing-gold-silver-decision-guide
[7] https://www.vaneck.com/us/en/investments/gold-miners-etf-gdx/overview/

model's predictive accuracy, while the right plot shows the variables improve node purity the most, which leads to more accurate model predictions.



All GDX price levels are included in both variable importance plots other than gdx_open (open price of Gold Miners ETF), indicating that the ETF overall represents an extremely important predictor in the model for effective predictions.

Other important variables are the various price levels of silver futures and platinum futures. Although interpretability is limited with a random forest model, the variable importance plots help provide additional evidence which aligns with the previously drawn conclusions that there exists a strong relationship between silver and platinum with gold. The log returns of the usdi_price (U.S. Dollar Index) and usb_price (U.S. 10-year Treasury notes) are also included in both plots, however the importance of these two variables in increasing the predictive accuracy of the model is slightly less.

Model Performance Summary	
Model	**Mean-Squared Error**
Linear Regression	0.000023643
LASSO	0.000022871
GAM	0.000022007
Random Forest Model	0.000027174

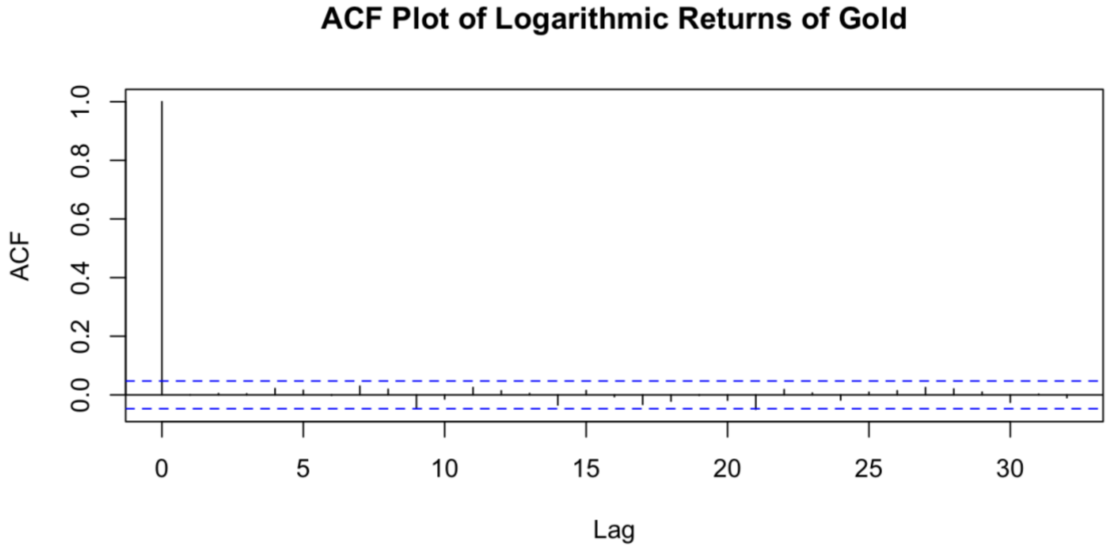
Autoregressive Moving-Average Model (ARMA)

Since the dataset is in the format of a time series, as it records the daily price levels of different assets, a time series model can also be considered to assess whether logarithmic returns of gold can be explained by its own past movements. The ARMA model was chosen for its simple implementation and its effectiveness in modelling stationary time series, which is the case with our logarithmic returns.

To assess what parameters *p* (used in modelling the autoregressive component of the model) and *q* (used to model the moving average component) should be used to best model the response, a function was constructed to fit all potential combinations of the model (i.e. ARMA[1, 1], ARMA[1, 2], …, ARMA[4, 4]) and identify which model achieved the lowest AIC value. The minimum AIC value achieved was -11076.95 when p = 1 and q = 2. Therefore, the time series model ARMA(1, 2) would achieve the best balance between model complexity and fit.

	MA0	**MA1**	**MA2**	**MA3**	**MA4**
AR0	-11,068.53	-11,074.74	-11,073.44	-11,072.81	-11,070.81
AR1	-11,074.34	-11,073.12	-11,076.95	-11,070.81	-11,068.81
AR2	-11,073.73	-11,072.97	-11,070.26	-11,073.52	-11,071.89
AR3	-11072.61	-11,070.61	-11,073.27	-11,072.12	-11,073.31
AR4	-11,070.61	-11,068.62	-11,071.74	-11,074.82	-11,070.48

An autocorrelation function plot was also plotted to examine whether ARMA(1, 2) adequately captures the autocorrelation structure of the dataset.



The plot shows that all lags stay within the interval, indicating strong fit of the ARMA(1, 2) model. Finally, a Ljung-Box test was conducted to quantify the degree of goodness-of-fit of the model. The model returned a p-value of 0.8082, meaning that the residuals of the model are independent and represent random white noise. As a result, no additional autocorrelation remains unexplained by the model, and the overall model serves as a good fit for the data.

The ARMA(1, 2) model can be broken down in two parts: the AR(p) term and the MA(q) term. In this model, the first AR(1) part indicates that the current value of the series is linearly dependent on the previous observed value. The second MA(2) term implies that the current value is linearly dependent on the previous two error terms.

A summary of the ARMA model shows that logarithmic gold returns from 2011 to 2018 is modelled by the following equation:

$$X_t = -0.8857X_{t-1} + \varepsilon_t + 0.8203\varepsilon_{t-1} - 0.0878\varepsilon_{t-2}$$

The AR(1) coefficient is -0.8857, which represents the degree of the autoregressive relationship between the current value X_t and the previous value X_{t-1}. The MA(2) coefficients are 0.8203 and -0.0878, which show the influence that the past two errors ε_{t-1} and ε_{t-2} have on X_t. The white noise term $\varepsilon_t \sim N(-0.0001, 0.00009188)$.

Overall, this concludes that the value of future logarithmic gold returns may be explained by the behaviour of short-term historical observations. This provides the opportunity to conduct future time series analyses with the response variable, where Bayesian time series models and Vector Error Correction Models (VECM) can be considered, as they take into consideration the influence of predictor variables over time to model the response variable.

Conclusion

Gold has represented an indispensable resource in global history, culture, and economies, and its attractiveness as a stable investment option is predicted to remain strong in the long-term future. After modelling the relationships between the logarithmic returns of gold (SPDR Gold ETF) and the logarithmic returns of 14 other assets using linear regression, LASSO regression, a Generalized Additive Model, and a random forest model, it was found that the logarithmic returns of plt_price (platinum futures), sf_price (silver futures), and gdx_adj_close (adjusted close price of Gold Miners ETF (NYSEARCA: GDX)) were most statistically significant in modelling and predicting log gold returns across all the generated models. The variables usdi_price (U.S. Dollar Index) and usb_price (U.S. 10-year Treasury Note) had slightly weaker influence on the response, however they remained statistically significant across all the models.

The relationship between log returns of silver futures and the GDX ETF on the response was very similar, being positive and linear. Generally, for every 0.01% increase in the log returns of either variable, the log return of the gold ETF tends to increase by 0.002%. With platinum futures, the relationship is highly polynomial, where more volatile changes in the asset contribute to more drastic changes in the response.

Overall, these results suggest that investors can look to analyzing the behaviour of these five main assets in order to predict future gold prices and understand the factors that influence the price of gold. These relationships also suggest that to achieve greater portfolio diversification, certain assets should not be included in the same portfolio with gold, such as silver futures or platinum futures.

Appendix

Appendix 1: Dataset Variables Description.

Variable List		
Variables in dataset	**Variable Prefix**	**Metrics tracked in the dataset**
Date	Date	Daily trading days
Gold (SPDR Gold Shares) – Response		Open, High, Low, Close, Adjusted Close, Volume
S&P 500 Index	SP	Open, High, Low, Close, Adjusted Close, Volume
Dow Jones Index	DJ	Open, High, Low, Close, Adjusted Close, Volume
Eldorado Gold Corporation (NYSE: EGO)	EG	Open, High, Low, Close, Adjusted Close, Volume
EURUSD Exchange Rate	EU	Open, High, Low, Price, Trend
Brent Crude Oil Futures	OF	Open, High, Low, Price, Volume, Trend
WTI/USD (Crude Oil WTI Spot US Dollar)	OS	Open, High, Low, Price, Trend
Silver Futures	SF	Open, High, Low, Price, Volume, Trend
US 10-Year Treasury Note	USB	Open, High, Low, Price, Trend
Platinum Futures	PLT	Open, High, Low, Price, Trend
Palladium Futures	PLD	Open, High, Low, Price, Trend
Rhodium Spot Price	RHO	Price
US Dollar Index	USDI	Open, High, Low, Price, Volume, Trend
Gold Miners ETF (NYSEARCA: GDX)	GDX	Open, High, Low, Close, Adjusted Close, Volume
United States Oil ETF (NYSEARCA: USO)	USO	Open, High, Low, Close, Adjusted Close, Volume

Price Level Descriptions		
Price Level	**Description**	**Variable Type**
Open	Price of the asset when the market first opens	Numeric
High	Maximum/Highest price attained by the asset during the trading day	Numeric
Low	Minimum/Lowest price attained by the asset during the trading day	Numeric
Close	Price of the asset at the end of the trading day when the market closes	Numeric
Adjusted Close (Response)	**The closing price of an asset, adjusted for corporate actions such as dividends, stock splits, and new stock offerings**	**Numeric**
Volume	Total number of shares or contracts traded during the trading day	Integer
Price	For some variables, Price is equivalent to Adjusted Close	Numeric
Trend	A binary variable indicating an increase in the asset's price from the previous trading day (1) or a decrease (0)	Factor

Appendix 2: Data Import Code

```
data = read_csv("gold_prices.csv",
            na = c("", NA))

# specify column types
data = data |>
            clean_names() |>
            mutate_at(vars(7, 13, 19, 25, 35, 46, 68, 75, 81), as.integer) |>
            mutate(across(c(30, 36, 41, 47, 52, 57, 62, 69), as.factor)) |>
            select(-c(2:5, 7, 18, 80)) |> # since we're not evaluating any of the other characteristics
          of gold
            rename("gold_adj_close" = "adj_close") |>
            mutate_at(vars(2:7, 9:12, 14:18, 20:23, 25:28, 31:34, 36:39, 42:45, 47:50, 52:55, 57:61,
          64:68,
                    70:73), ~c(0, diff(log(.)))) |> # get log returns of gold
            slice(-1) # remove the first row since we don't have the log returns
```

Appendix 3: Train/Test Split Code

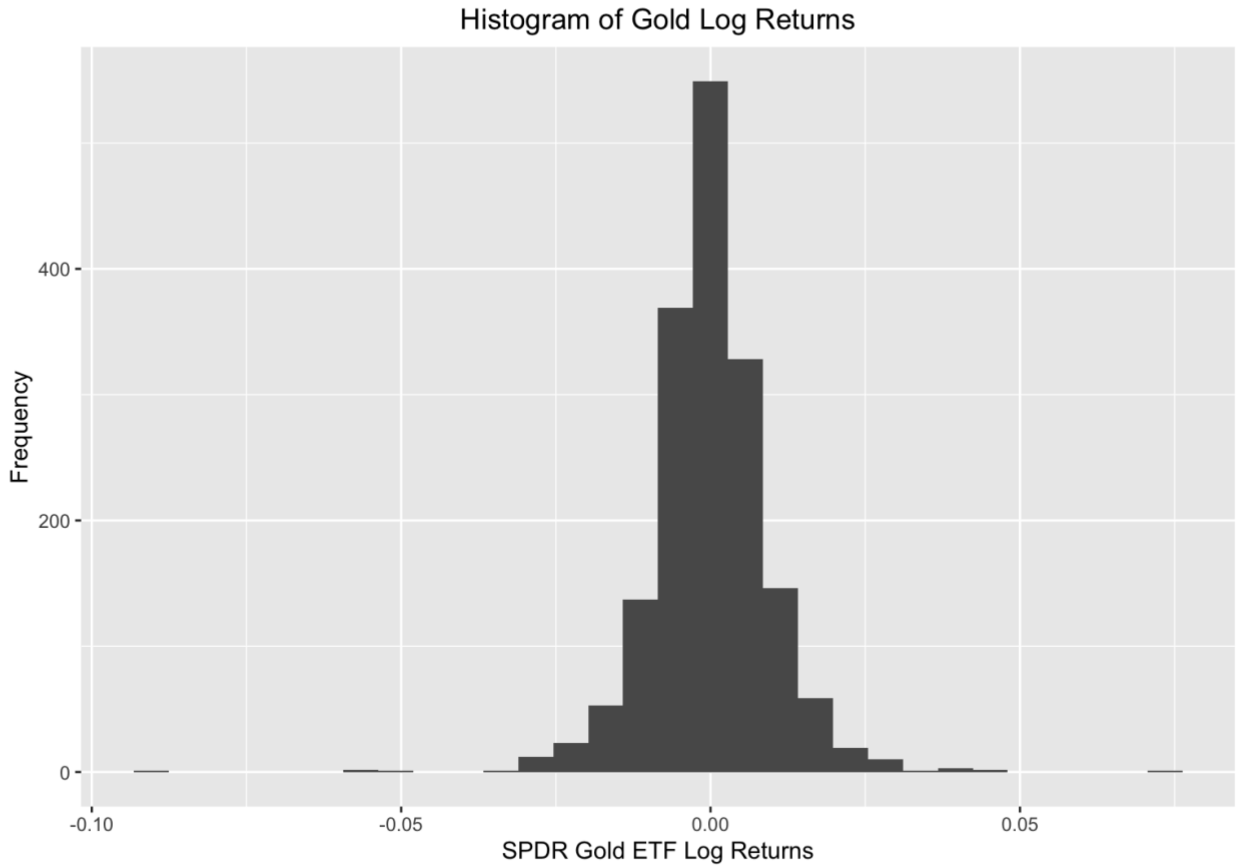
```
library(caret)
index = createDataPartition(data$gold_adj_close, p = 0.8, list = FALSE) # 80/20 split
train_data = data[index,] # training set
test_data = data[-index,] # testing set

x_train = as.matrix(select(train_data, !gold_adj_close))
y_train = train_data$gold_adj_close
x_test = as.matrix(select(test_data, !gold_adj_close))
y_test = test_data$gold_adj_close
```

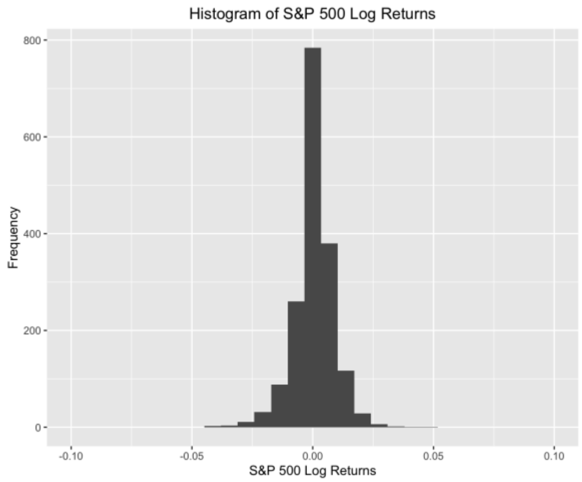
Appendix 4: Exploratory Data Analysis

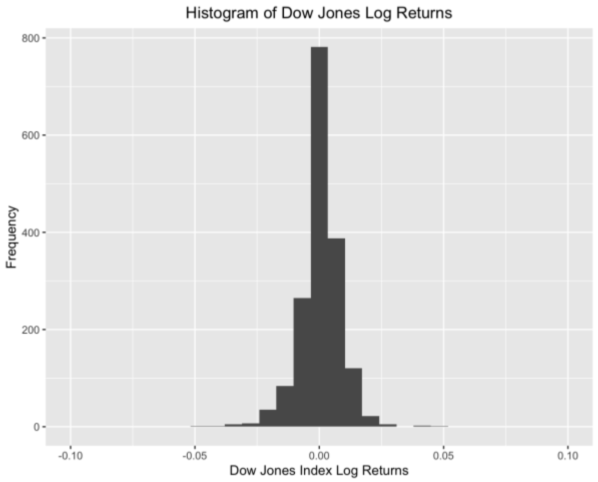
Histogram Code Sample (S&P 500)

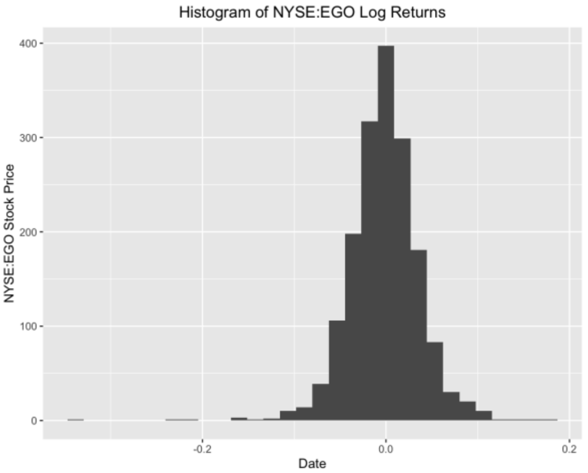
```
ggplot(data = data, mapping = aes(x = sp_ajclose)) +
        geom_histogram() +
        labs(x = “S&P 500 Log Returns”, y = “Frequency”, title = “Histogram of S&P 500 Log
                Returns”) +
        theme(plot.title = element_text(hjust = 0.5)) +
        xlim(-0.1, 0.1)
```

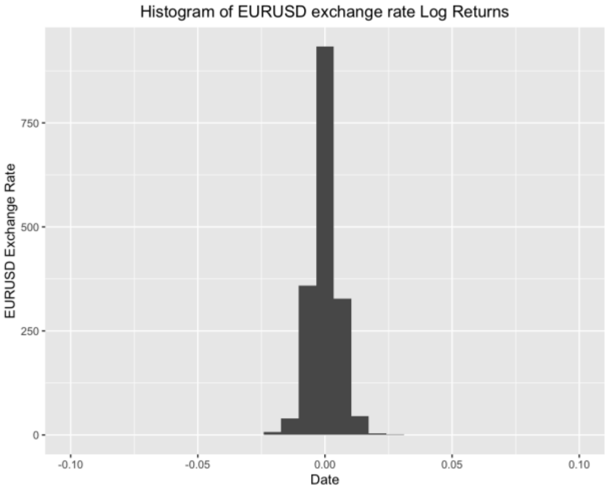
Histograms of Asset Log Returns



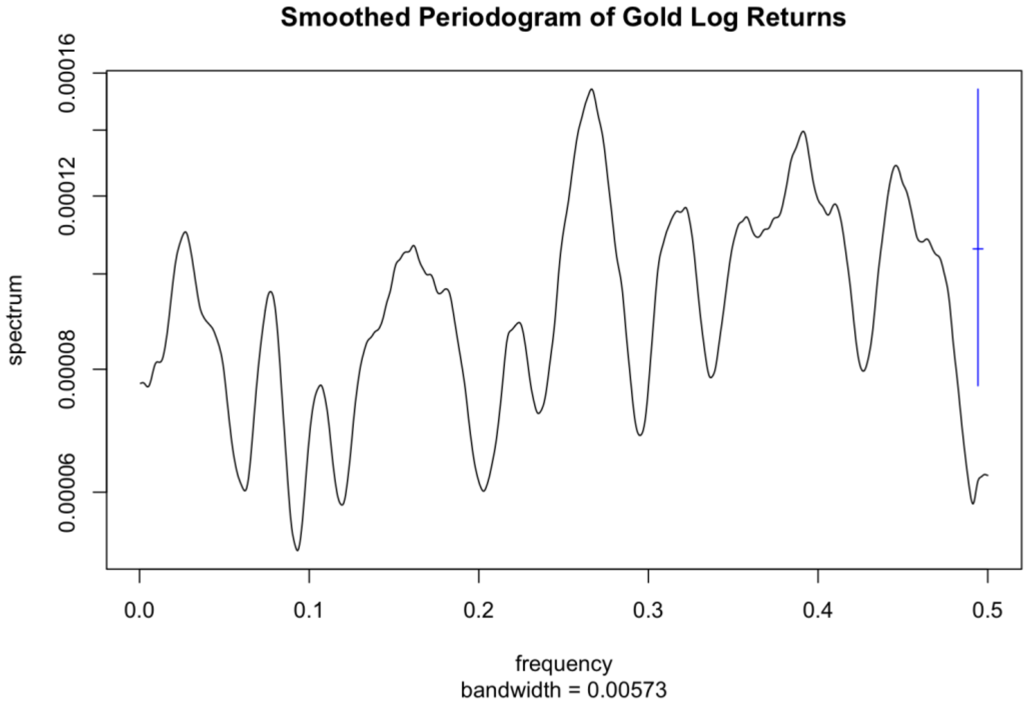








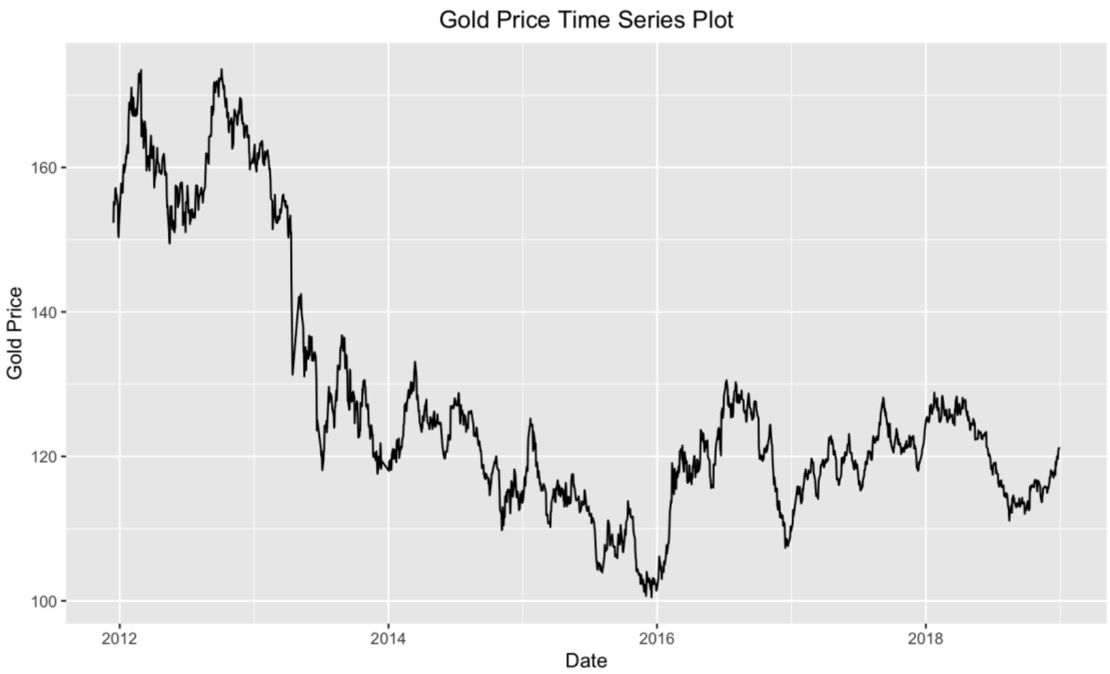
Distribution of Binary Trend Variables		
Variable Name	**Count of “0”**	**Count of “1”**
eu_trend (EURUSD Exchange Rate)	868	849
of_trend (Brent Crude Oil Futures)	861	856
os_trend (WTI/USD)	852	865
sf_trend (Silver Futures)	892	825
usb_trend (US 10-Year Treasury Note)	876	841
plt_trend (Platinum Futures)	885	832
pld_trend (Palladium Futures)	806	911
usdi_trend (US Dollar Index)	836	881

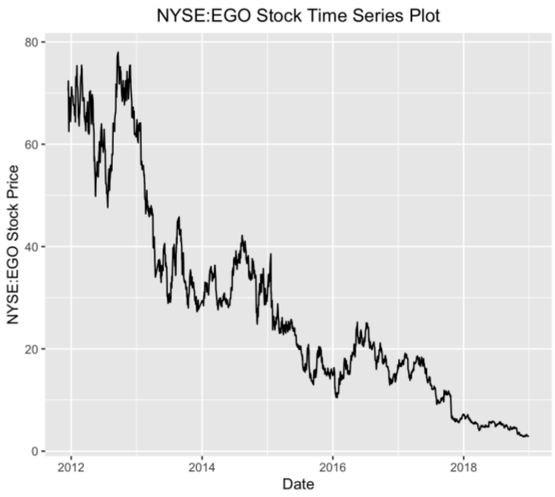


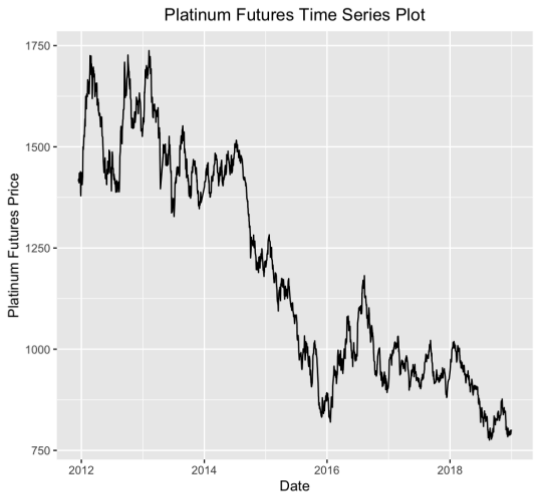
Time Series Code Sample (Gold Prices)

```
ggplot(data, aes(x = date, y = gold_adj_close)) +
        geom_line() +
        labs(x = "Date", y = "Gold Price", title = "Gold Price Time Series Plot") +
        theme(plot.title = element_text(hjust = 0.5))
```

Time Series Plots


Gold Price Time Series Plot


NYSE:EGO Stock Time Series Plot


Platinum Futures Time Series Plot

Appendix 5: Main Data Analysis

Linear Regression Code

```
lmod = lm(gold_adj_close ~ . , data = train_data)
summary(lmod)

lmod_preds = predict(lmod, test_data, type = "response")
mse_lmod = mean((y_test - lmod_preds)^2)
mse_lmod
```

LASSO Regression and Cross-Validated Lambda Code

```
library(glmnet)
lambda_value = cv.glmnet(x = x_train,
                         y = y_train,
                         alpha = 1, # LASSO regression
                         nfolds = 10, # 10 fold CV, performs a grid search to find best lambda value
                         family = "gaussian")

plot(lambda_value, main = "Log Lambda Value")
lambda = lambda_value$lambda.min


lasso_mod = glmnet(x = x_train,
                   y = y_train,
                   alpha = 1,
                   lambda = lambda,
                   family = "gaussian")

lasso_coefs = coef.glmnet(lasso_mod)

lasso_preds = predict(lasso_mod, newx = x_test, s = lambda, type = "response")
mse_lasso = mean((y_test - lasso_preds)^2)
mse_lasso # slightly less than lmod
```

Generalized Additive Models Code

```
library(mgcv)

gam_mod = gam(gold_adj_close ~ s(sp_ajclose) + s(dj_close) + s(eg_ajclose) + s(eu_price) +
            s(of_price) + s(os_price) + s(sf_price) + s(usb_price) + s(plt_price) +
            s(pld_price) + s(rho_price) + s(usdi_price) + s(gdx_adj_close) + s(uso_close), data =
                  train_data)

summary(gam_mod)

library(ggeffects)
gam_plots = plot(ggpredict(gam_mod))

gam_preds = predict(gam_mod, newdata = test_data, type = "response")
mse_gam = mean((y_test - gam_preds)^2)
mse_gam
```

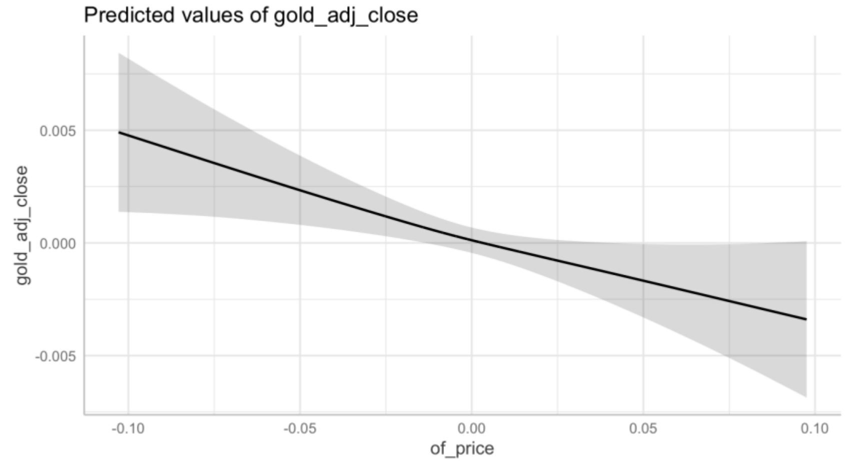
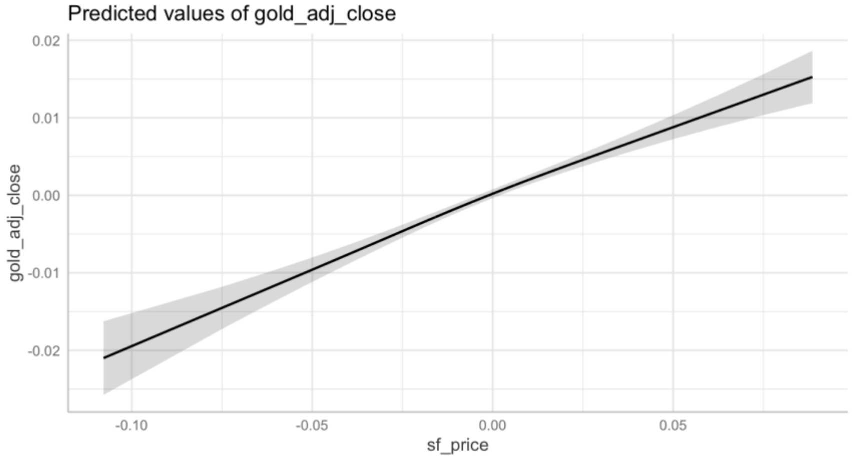
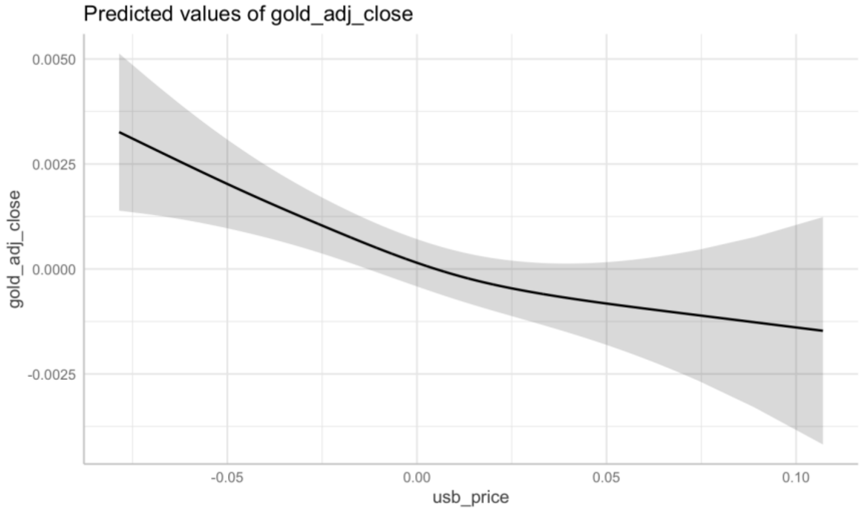
Random Forest Model Code

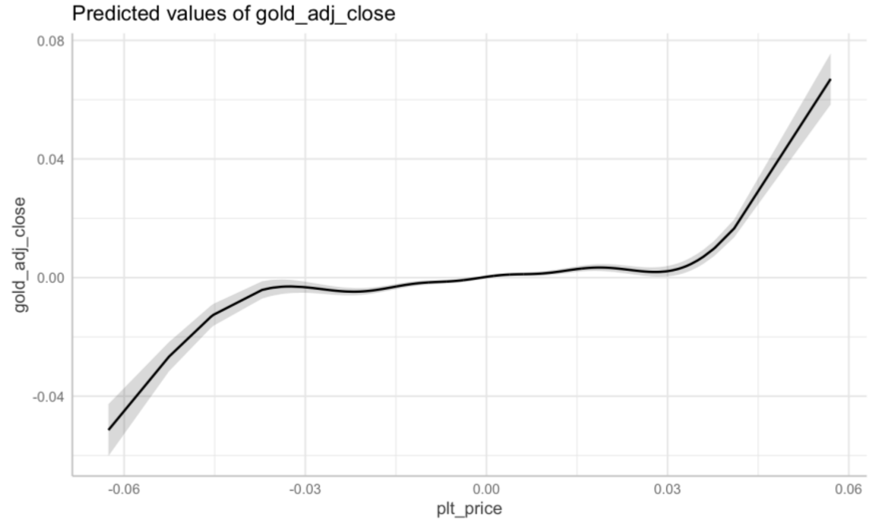
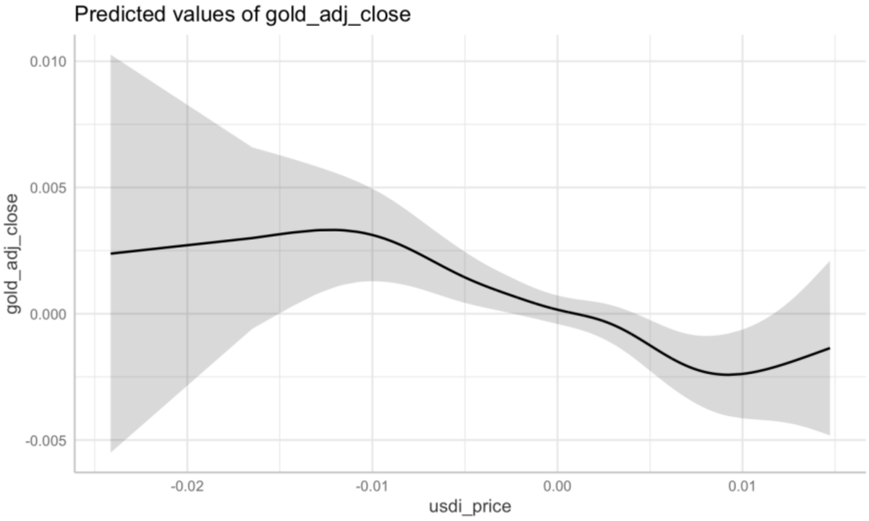
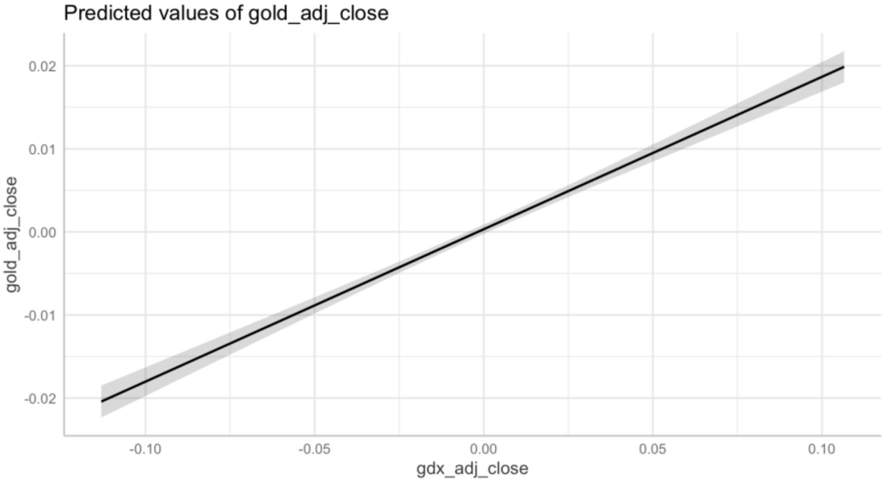
```
library(randomForest)
rf_mod = randomForest(gold_adj_close ~ ., data = train_data, importance = TRUE, ntree = 100)
plot(rf_mod, main = "Random Forest Trees vs. Error")

varImpPlot(rf_mod, type = 1, n.var = 15, main = "% Increase in MSE")
varImpPlot(rf_mod, type = 2, n.var = 15, main = "Increase in Node Purity")

rf_preds = predict(rf_mod, newdata = test_data, type = "response")
mse_rf = mean((y_test - rf_preds)^2)
mse_rf
```

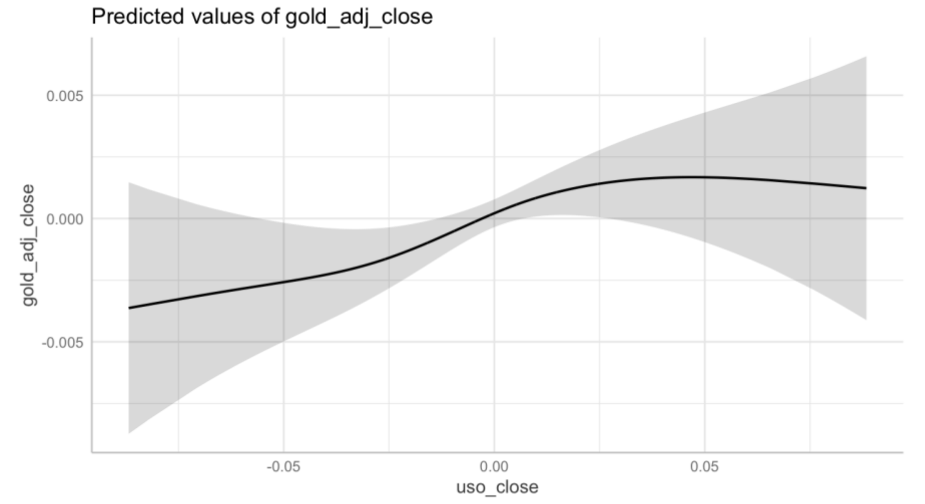
ARMA Model Code

```
library(xts)

gold_ts = as.ts(xts(data$gold_adj_close, order.by = as.Date(data$date)))

aic_table = function(dataset, p, q){
  table = matrix(NA, (p + 1), (q + 1))
  for (i in 0:p){
    for (j in 0:q){
      table[i + 1, j + 1] = arima(dataset, order = c(i, 0, j))$aic
    }
  }
  dimnames(table) = list(paste("AR",0:p,sep=""),paste("MA",0:q,sep=""))
  table
}

p_select = aic_table(gold_ts,4,4)
min(p_select) # lowest aic at ARMA(1, 2)

arma_ts = arima(gold_ts, order = c(1, 0, 2)) # arima model
acf(resid(arma_ts), main = "ACF Plot of Logarithmic Returns of Gold") # acf plot

Box.test(resid(arma_ts), lag=20, type="Ljung-Box") # ljung box test
```

GAM Pairwise Plots

Smoothed Predictors	Plots
s(of_price)	
s(sf_price)	
s(usb_price)	

s(plt_price)	
s(usdi_price)	
s(gdx_adj_close)	

s(uso_close)	Predicted values of gold_adj_close gold_adj_close 0.005 0.000 -0.005 -0.05 0.00 0.05 uso_close