Time Series- Market Metric Forecasting

By: Syed Multani, Alvin Lee, Shamit Singh, Andrew Dugal

**Abstract:** The goal of this project is to build an app that allows the user to examine and forecast the values of gold, Federal reserve interest rates, and S&P 500 index- based on historical values. The app will allow users to examine relationships between these values to give insight into market trends in the future.

**Introduction and Discussion:**

For novice investors or just those interested in learning how financial markets work, it is difficult to know where to start. This project gives an overview of three aspects of the market that are important to understand for anyone involved in following or investing in the stock market. We have chosen to build an app based on historical values of gold, Federal reserve interest rates and S&P 500 index.

Each of these three values are important for different reasons. Gold is the ‘old faithful’ investment. Markets fluctuate, instability happens, but precious metals- gold in particular, are considered safe and stable long term investments. Federal reserve interest rate is an interesting metric to examine, not least for the fact that they are controlled by a centralized bank and not the government themselves. Many people do not realize the structure of the US financial system in that they do not print their own money or set their own interest rates. As such, the market can be manually manipulated by the banks that set the rates leading to a false sense of security in some cases or a false sense of insecurity in some cases. Regardless, the Federal reserve rates are and important metric to watch and understand on it’s own, and in relation to other market data. The S&P 500 is a stock market index of the 500 largest companies on the US stock exchange. It covers a huge range of the market cap. The rise and fall of S&P dividends is a good indicator of the overall health of the market, and can be used to dictate where to invest or divest.

In this time series project, we built a dataset from multiple individual datasets. Our first dataset is a monthly gold stock value showing the value of gold at the end of each month going back to the 1960s. Our second dataset is monthly Federal Reserve interest rate data going back monthly to the 1960s as well. Our third dataset is a monthly value of the S&P 500 going back to 1871. The data needs to be cleaned and merged before we can run any analysis. Once we have merged and compiled useful data from our datasets- we will run a Vector autoregression model to examine and plot relationships between the multiple variable that we are using (gold, interest rates, and S&P values). Our app will allow users to examine each metric individually or in any combination with one another, and forecast future values.

**Dataset:**

Our review mining will be completed using the dataset(s) accessed here:

Gold: <https://datahub.io/core/gold-prices>

This dataset contains 836 rows and two columns. This data is public domain and compiled by Deutsche Bundesbank- the central bank of Germany. The data is in USD per ounce of gold.

Federal Reserve Interest Rates:

<https://www.federalreserve.gov/datadownload/Download.aspx?rel=H15&series=40afb80a445c5903ca2c4888e40f3f1f&filetype=csv&label=include&layout=seriescolumn&from=01/01/1960&to=10/31/2019>

This dataset contains 717 rows and 2 columns. It was compiled manually through the Federal Reserve Data Download system on their website. This website allows you to build your own dataset based on your own parameters.

S&P 500: <https://datahub.io/core/s-and-p-500>

This dataset contains 1768 rows and 9 columns. For the purposes of this project we only need the date column and the S&P 500 value- so the other 7 columns will be removed. This data is open source and was made available and compiled based on the work of Yale Economist Robert Shiller <http://www.econ.yale.edu/~shiller/data.htm>

As stated earlier, we merged these datasets to create a consistent and usable dataset containing 592 rows and 4 columns (as you will see in our data preparation section).

**Ethical ML Framework:**

The goal of our report and app is to use historical data to forecast future values of gold, interest rates, and S&P values, as well as examine the relationship between the three as time goes on.

All data used is public domain, and free for use in our project. There are no ethical implications, or potential legal issues in the use of this data. It will simply be used for the purposes of examination, exploration, and educational purposes.

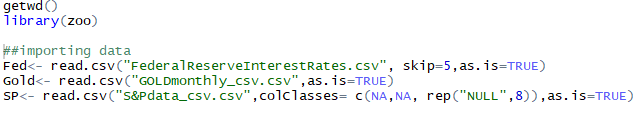
**Assumptions:**

We are assuming the data has been compiled and presented in an ethical and objective manner ie) market data that has not been manually manipulated from the original source. We have done all we can to ensure data was compiled from reputable sources, and as direct as possible from the source of collection.

One other main assumption we are making in this project is that the variables we are exploring have a relationship and influence between and amongst one another. This was the main reason for using Vector Autoregression modeling (VAR) in our project. VAR is used under the assumption that there are bi-directional relationships existing between the multiple variables used in analysis. Since we are using financial market data, and no aspect of the stock market or financial markets exist in a vacuum, it is a safe assumption to make that our variables will have a relationship with one another- therefore, it is assumed that VAR will provide us with the most accurate forecasting values.

**Data Preparation:**

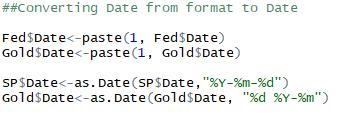
Import data:



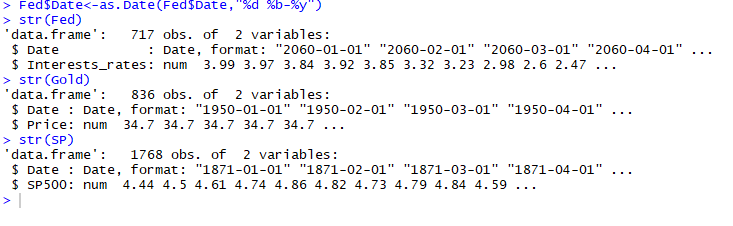
Rename the column name in the Fed dataset:



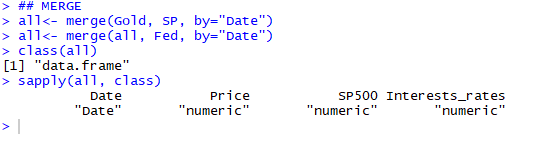
Convert date from existing format to date:



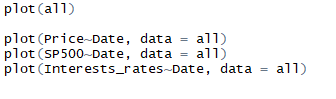
Examine to confirm date format consistency:



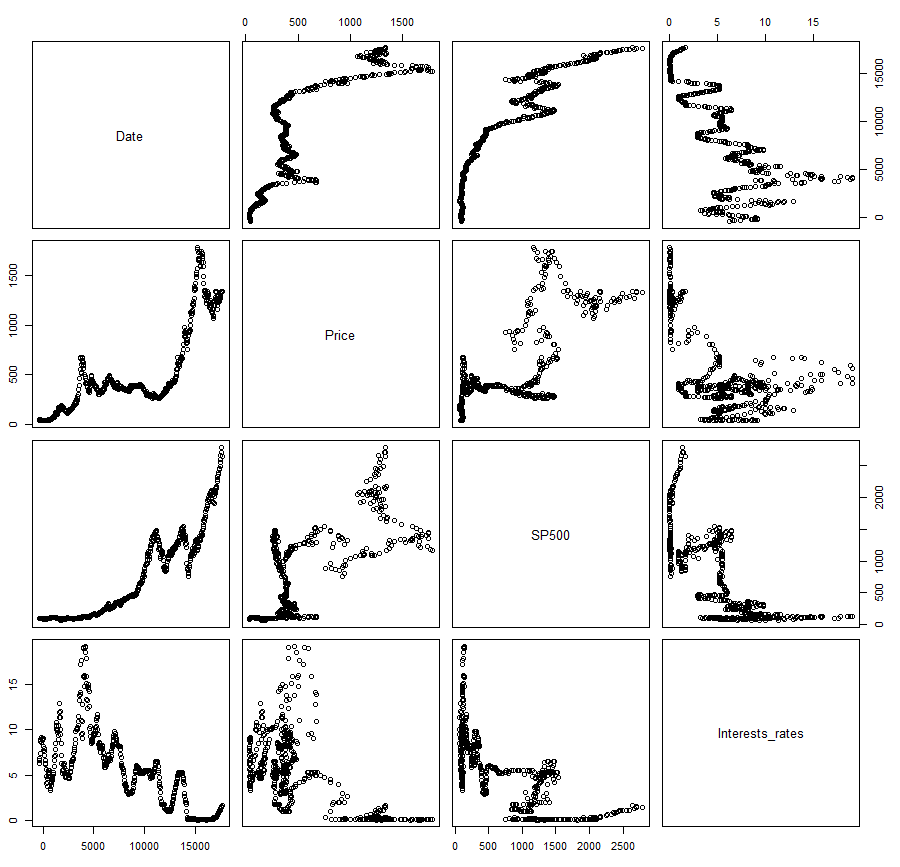
Merge datasets, check column format:



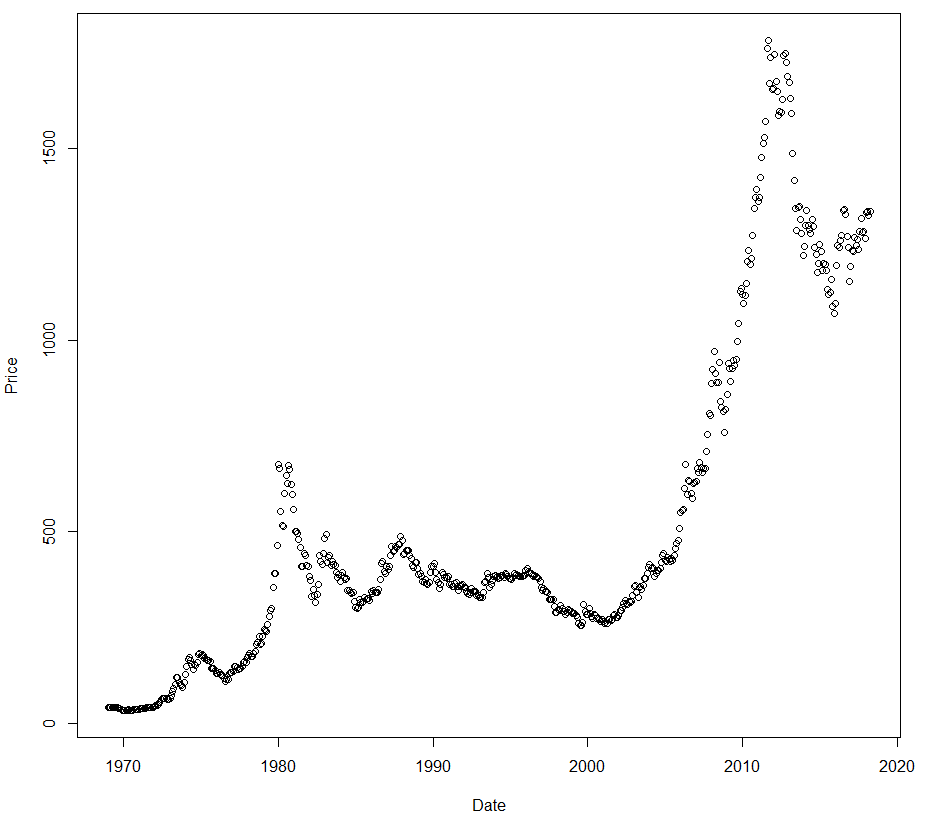
Visualizations for all variables together and then individually:



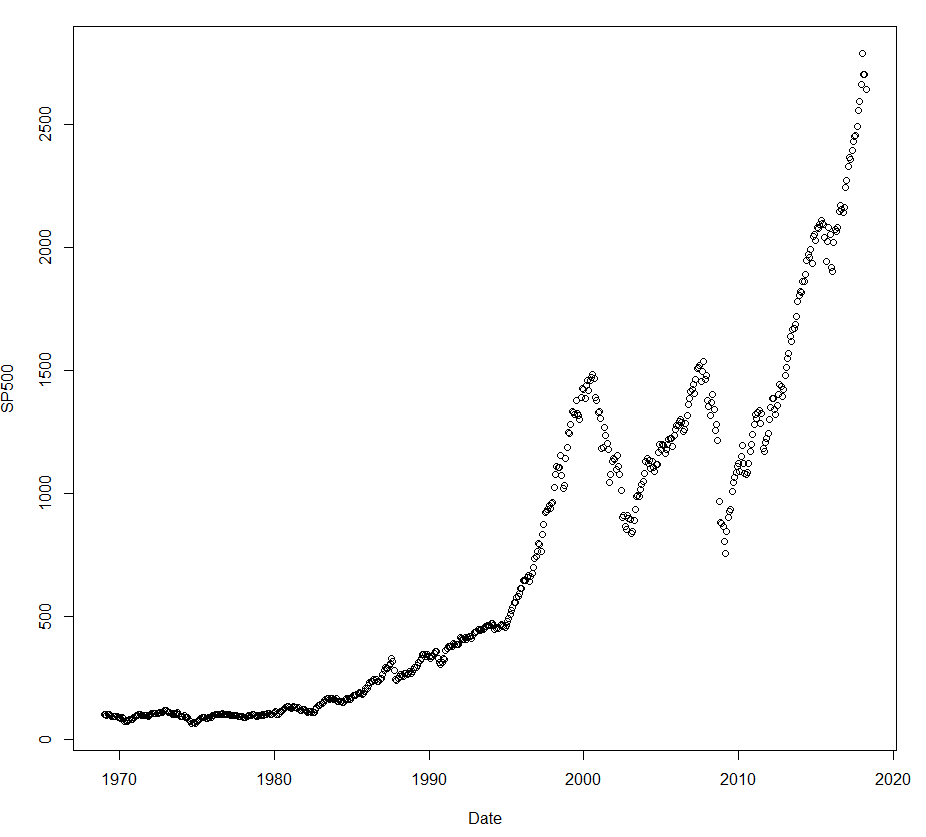
All variable together:



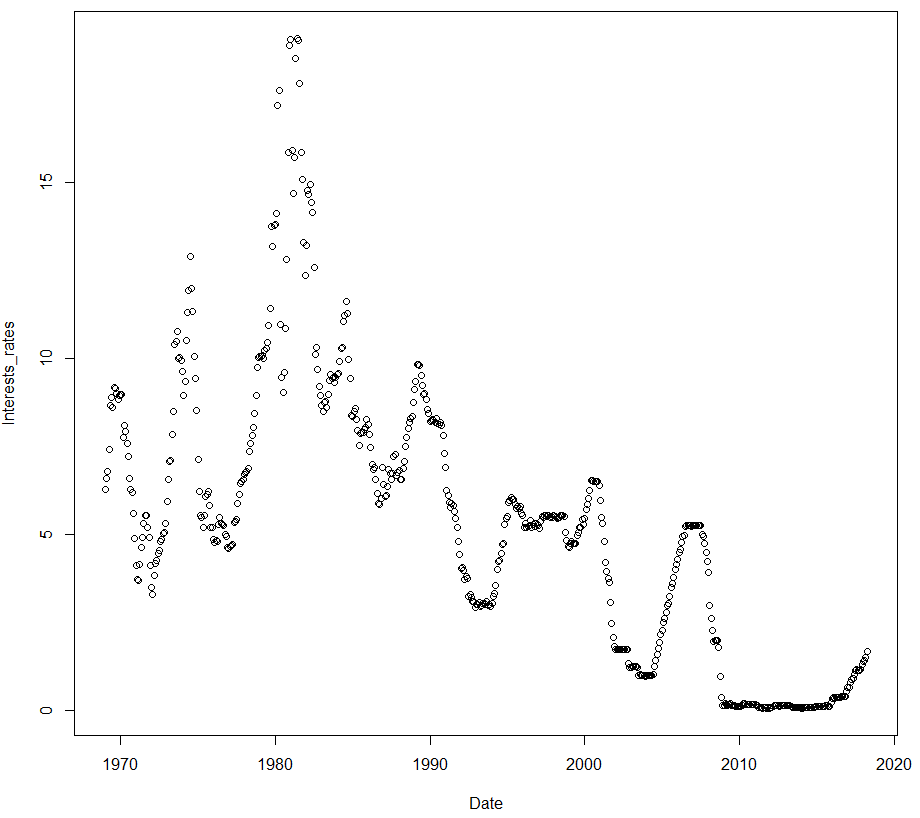
Gold prices by date:



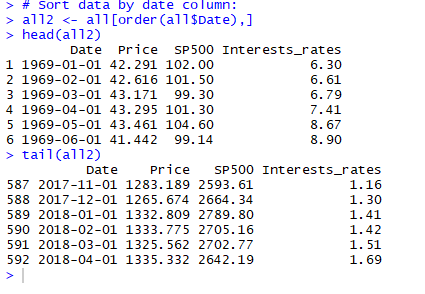
S&P 500 by date:



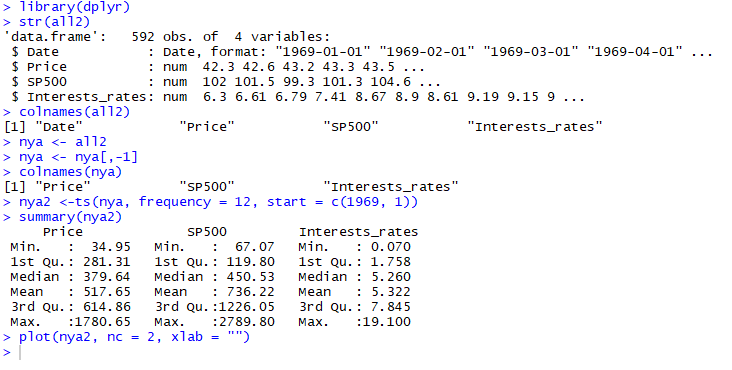
Federal Reserve interest rates by date:

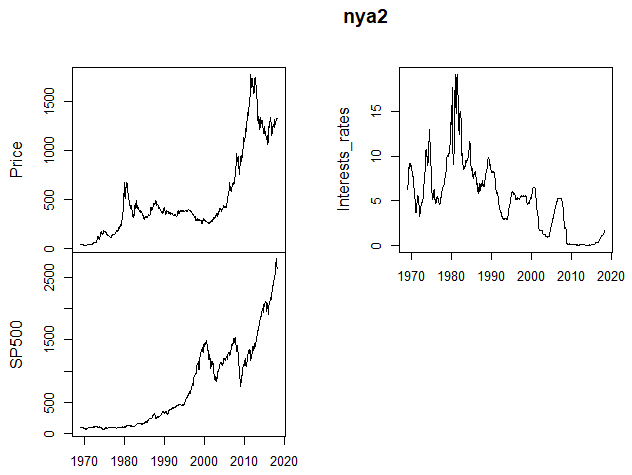


Sort data by date column and examine head and tail:

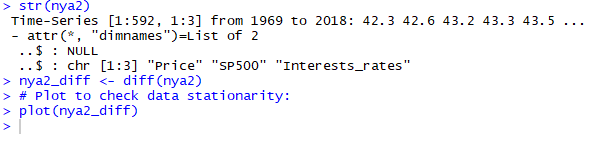


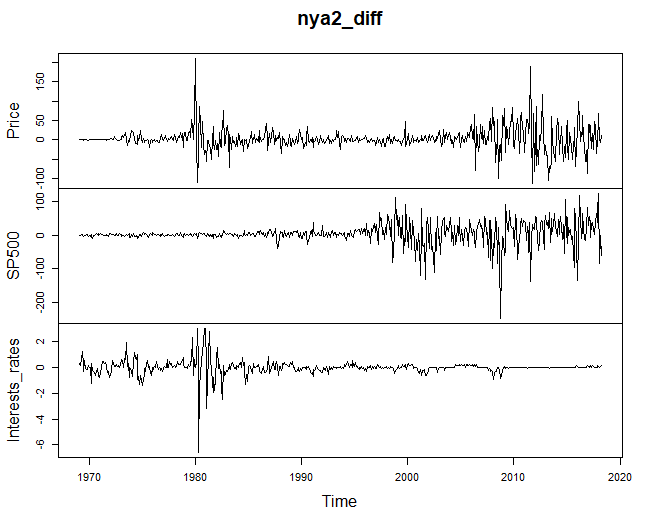
Convert to time series, view summary, and plot:





Transform non-stationery data into a stationery series for analysis. To address the trend component, take difference of the time series. Below, we have also included a plot of the data to give a visual representation that the data is now stationary:

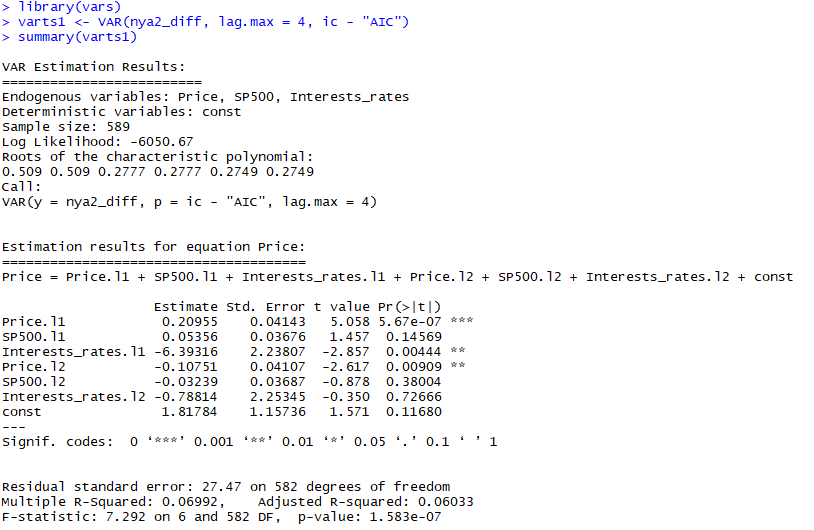


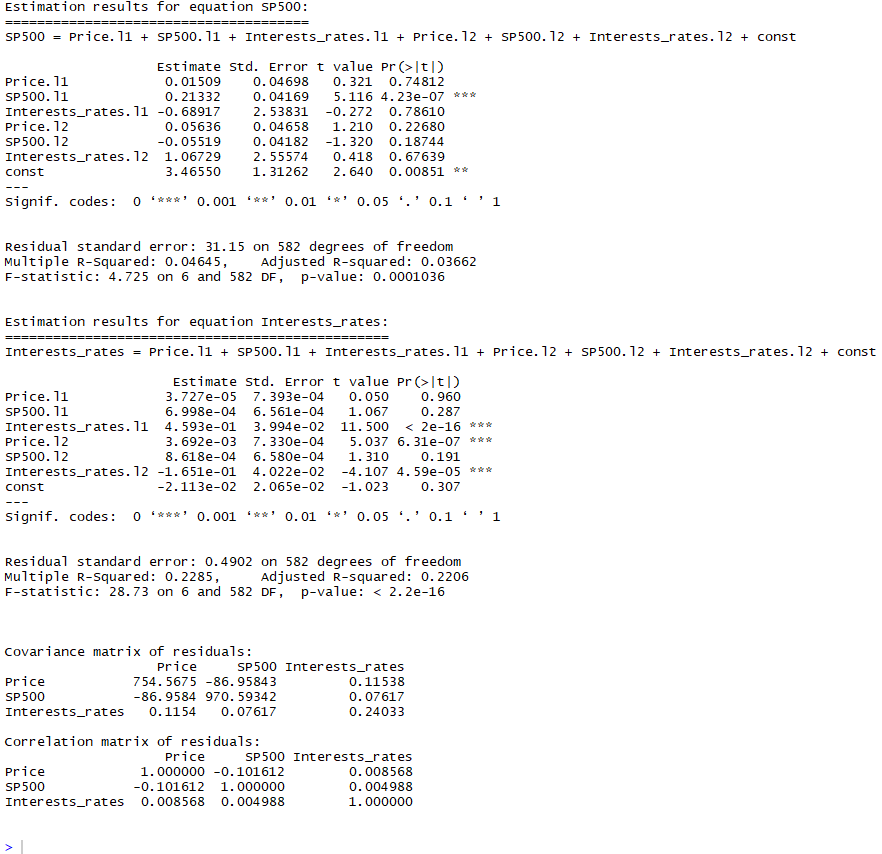


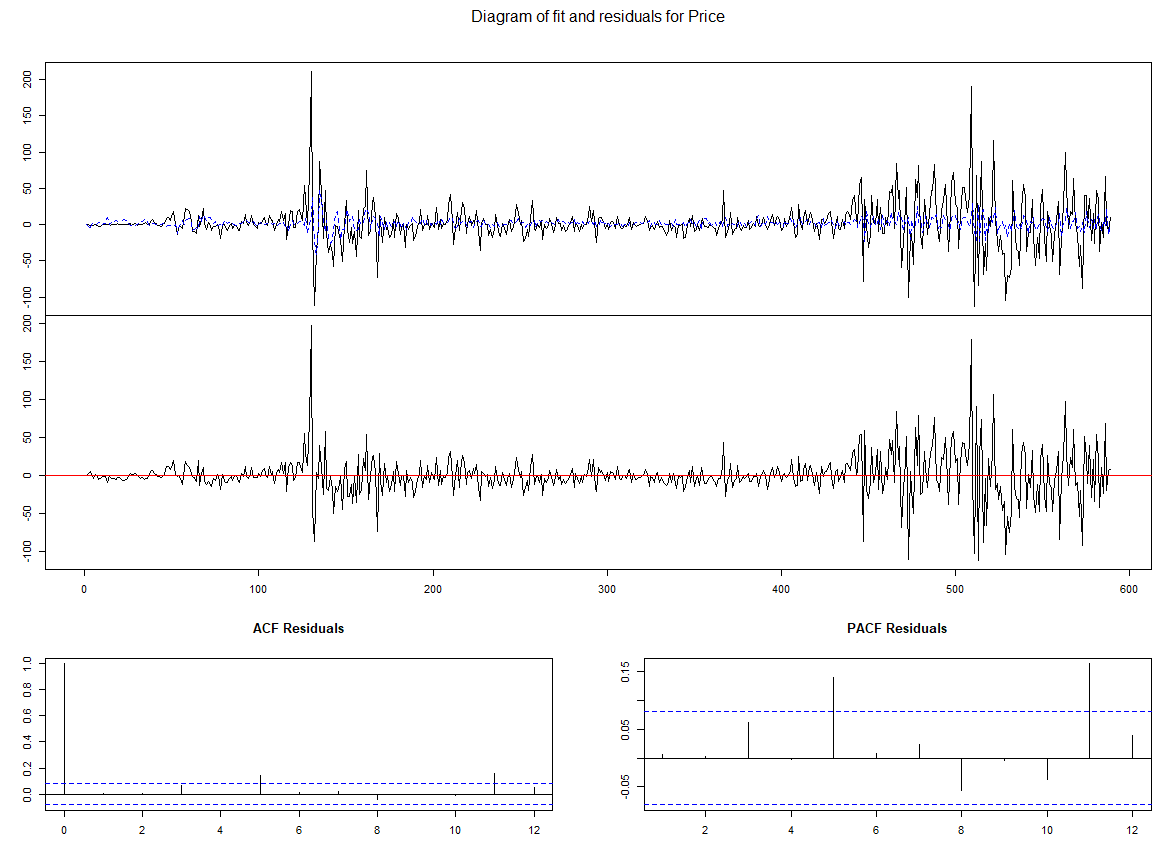
**Modelling with Vector Autoregulation (VAR):**

The VAR model implies that everything depends on everything. Here you will see we fit a VAR model and decide optimal lag length as per AIC. An AR(p) model is an autoregressive model where specific lagged values of yt are used as predictor variables. Lags are where results from one time period affect following periods.

The value for "p" is called the order. For example, an AR(1) would be a "first order autoregressive process". The outcome variable in a first order AR process at some point in time “t” is related only to time periods that are one period apart (i.e. the value of the variable at t - 1). A second or third order AR process would be related to data two or three periods apart. The autocorrelation is the average correlation between all pairs of years that differ by 1, 2 years, and so on. The differences are referred to as lags.







Analysis:

Positive correlation is a relationship between two variables where if one variable increases, the other one also increases. A positive correlation also exists if one decreases and the other also decreases.

Negative correlation is a relationship between two variables in which one variable increases as the other decreases, and vice versa. In statistics, a perfect negative correlation is represented by the value -1, a 0 indicates no correlation, and a +1 indicates a perfect positive correlation

From the model summary you can see Gold price is highly correlated with its own value at both lag 1 and lag2. Gold price is also statistically significant for Interest rate at lag 1. Similarly, SP500 is high correlated with its own value for lag 1 only. No other correlations are statistically significant. Finally, interest rate has significant impact because of Interest rate at lag 1 and lag 2. Also, it is significantly correlated with gold price at lag 2.

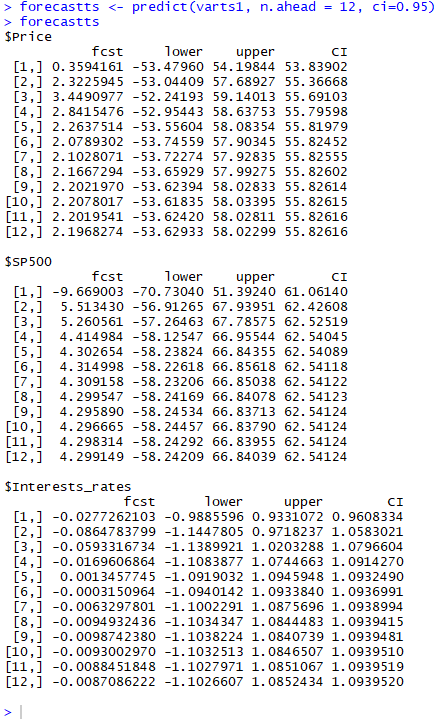
Covariance is a measure of the joint variability of two random variables. The sign of the covariance therefore shows the tendency in the linear relationship between the variables. The magnitude of the covariance is not easy to interpret because it is not normalized and hence depends on the magnitudes of the variables. The normalized version of the covariance, the correlation coefficient, however, shows by its magnitude the strength of the linear relation.

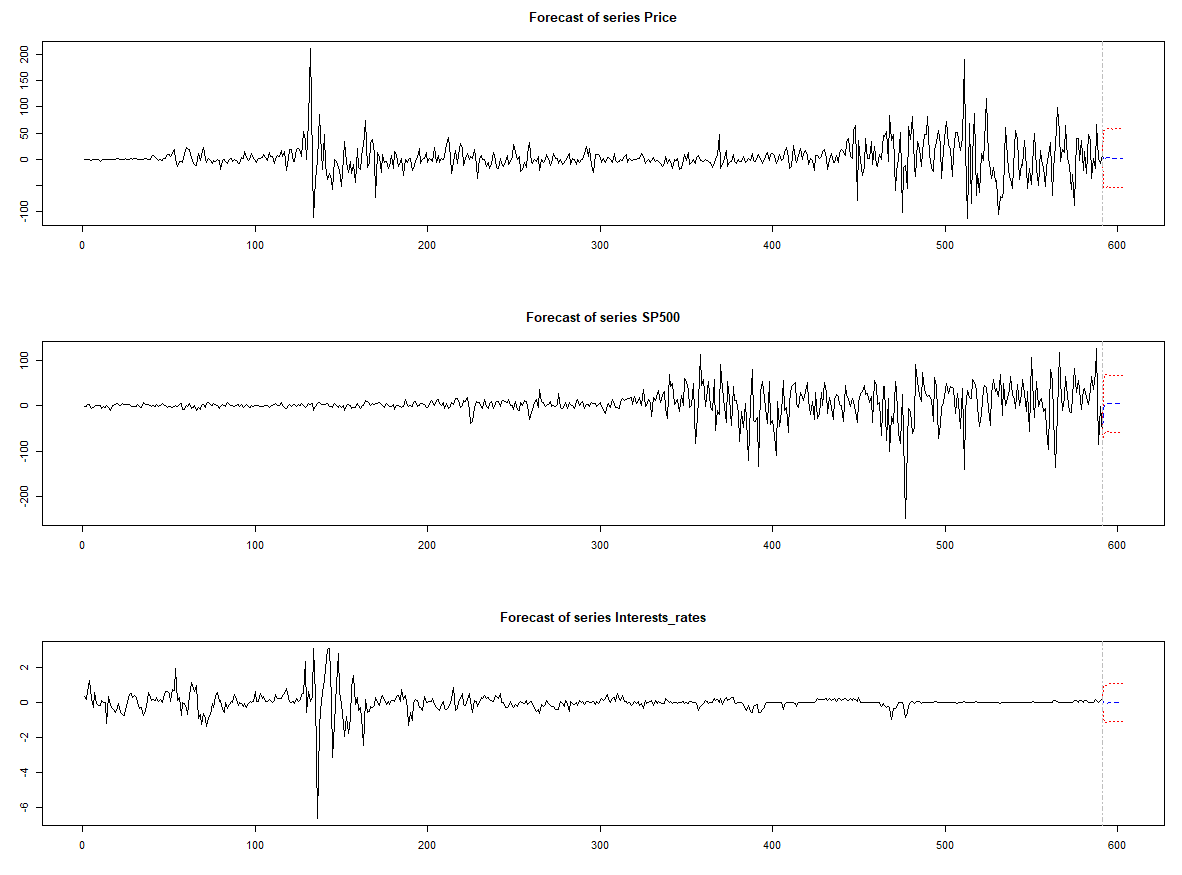
Our model's Covariance matrix shows positive covariance between gold price and interest rates, whereas negative covariance exists between gold price and SP500. SP500 shows positive covariance with interest rates. Interest rate shows positive covariance with itself, gold price and SP500

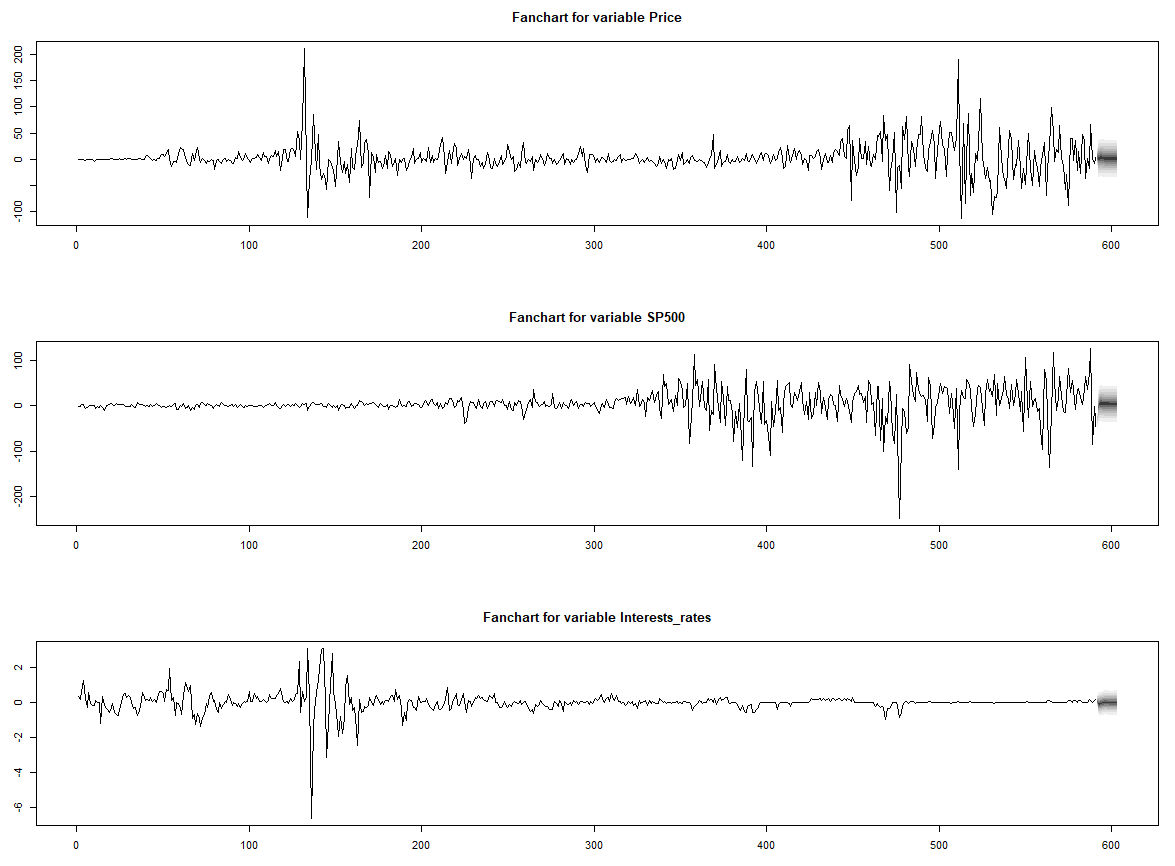
Our model's correlation matrix shows positive correlation between gold price and interest rate but negative correlation between gold price and SP500. Similarly SP500 shows positive correlation for interest rate. Interest rate shows positive correlation not only with itself but also with gold price and Interest rate.

To test the results and to check the overall fit of the model, we can analyse the value of R-squared. R-squared shows the proportion of variation explained by the estimated regression line. The smaller the variation, the better the regression model fits the sample data. R-squared always takes on a value between 0 and 1. The closer R-squared is to 1, the better the estimated regression equation fits or explain the relationship b/w variables. Adjusted R-squared is used in case of multiple regression analysis. R-squared increases as new independent variables are added to regression equation. For our model, we can see the Adjusted R-squared is higher than zero for all three estimations.

Prediction using the VAR model:



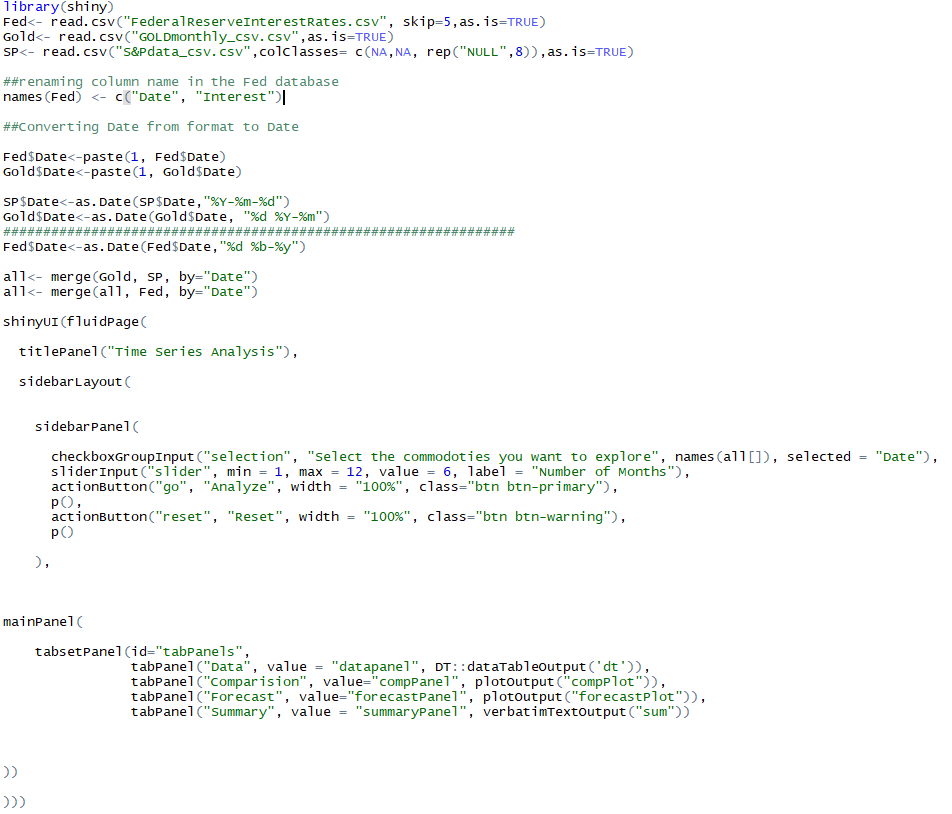




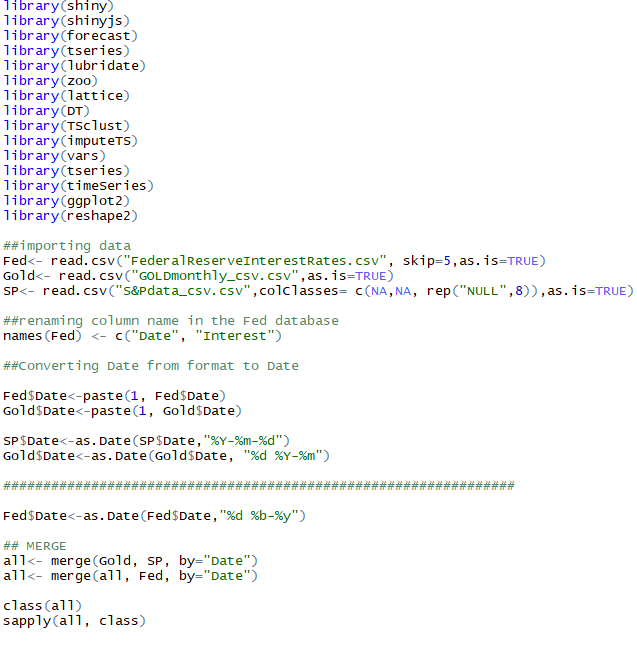
**Prepare Shiny App for Deployment:**

Link to app can be found here:  <https://shamit-singh.shinyapps.io/TimeSeriesAnalaysis/>

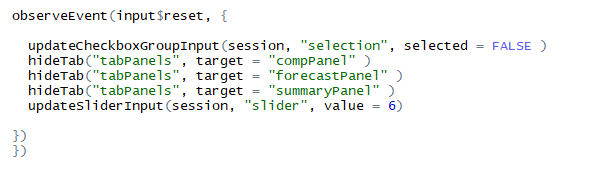
Build and prepare ui.R (user interface):



Build and prepare server.R file:







**App Deployment Discussion:**

As you will see from running our Ui, and Server files- our app offers a look at comparison and forecasting of our variables. Future iterations of this app would include increasing the number of variables to give a deeper look into how the stock market reacts to global financial trends.

Resources:

“Applied Multivariate Statistics with R”

by Daniel Zelterman

“Time Series Analysis and it’s Applications with R Examples 4th Ed”

by Robert H. Shumway and David S. Stoffer

“VAR, SVAR and SVEC Models: Implementation

Within R Package vars”

by Bernhard Pfaff