Saemi Ramirez

011926418

D214 PA 2

Data Analytics & Executive Summary

12/22/2024

WGU

1. Summarize the original real-data research question

The research question is, “How accurately can the gold price be predicted?”   
 The price of gold reflects a complex interplay of economic, social, and financial factors. Gold is often considered a refuge during economic uncertainty, recession, or financial crises. Is rising price typically signals fear or instability in financial markets. Gold tends to retain its value over time, making it a popular choice for preserving purchasing power during periods of high inflation. Rising gold prices often indicate concerns about currency devaluation or inflationary pressures. Interest rates also influence gold prices. Lower interest rates reduce the opportunity cost of holding on-yielding assets such as gold or driving demand. The near future can be estimated by the price of the gold which can relate to the economy or financial situation by the gold price prediction. (Understanding)  
The null hypothesis of the research question is if the gold price can be predicted at 90% accuracy, and the alternate hypothesis is if the gold price can be predicted with less than 90% accuracy.

1. Report on the data-collection process by describing the collected data. One advantage and one disadvantage of data-gathering methodology. Discuss how I overcame any challenge.

The data was sourced from the official website of the World Gold Council. The dataset consisted of nine sheets, including general information about the data. The remaining sheets were organized by timeframe (yearly, quarterly, monthly, weekly, and daily) and categorized as either End of Play (EoP) or average values. The “Weekly\_EoP” sheet was used for this research, containing the data and gold prices in 19 different currencies, including USD. This sheet provided weekly gold prices (every Friday) from December 29, 1978, to November 22, 2024. One advantage of this dataset was its comprehensiveness, as all the information was consolidated into a single file. The sheet names were intuitive, and each sheet included the data and the corresponding gold price, clearly labeled as an average of EoP. The “Weekly\_EoP” sheet alone contained approximately 2,400 records, offering a robust dataset for analysis. Additionally, the data was updated in real-time, allowing access to the most recent values up to November 22, 2024 (the dataset was accessed on Nov 26, 2024).  
 However, one drawback was that the dataset was provided in an XLSX format with multiple sheets. It started the data from a certain row, which required additional steps to extract specific data. To address this, I used pandas.read\_excel instead of pandas.read\_csv with the parameters of sheet\_name and skiprows to retrieve the required sheet efficiently. Despite this minor inconvenience, the dataset fulfilled all the requirements for the analysis.

1. Data Extraction and Preparation

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Libraries** | **Function** | **Justification** | **Advantage** | **Disadvantage** |
| Matplotlib.dates | Provide sophisticated data plotting capabilities | I used it to set the dates as numbers for x. | Supports a wide variety of date and time formats, provides functions to customize date and time labels | Requires familiarity with additional classes; handling large datasets with many date points can slow down rendering and interaction |
| Matplotlib.pyplot | Make a data visualization such as graphs and plots | Created a graph to see the trend of the gold price. | Supports wide range plots including line, scatter, bar, histogram, pie, contour, 3D plots, and more | Advantage features can be challenging for beginners; creating complex plots often requires a significant amount of code |
| Pandas | Analyze and manipulate the dataset | Used read\_excel function with sheet\_name and skiprows parameters to read xlsx formatted dataset. checked the shape of the data (shape). See the first 5 rows of the dataset (head())  Changed the column names (columns) | Optimized for performance, allowing efficient manipulation of large datasets | High memory consumption for large datasets for in-memory processing, advanced features like multi-indexing or pivot tables can be challenging |
| Seasonal\_decompose | Examine of time series influenced by factors that vary cyclically or periodically over time | Shows the trend, seasonal, and residuals | It requires minimal code to perform easier to analyze and interpret the underlying patterns in the data by decomposing the time series into trend, seasonal, and residual | Assumes the seasonality is constant and does not change over time; only works with evenly spaced time series data |
| rcParams | Offer a method for moving some of the repetitive code into a configuration file | Set all the visualization sfigure.figsizes to 18, 8. | The simple dictionary-like syntax makes it easy to modify and retrieve settings, and supports a comprehensive list of customizable parameters (line styles, marker properties, legend, etc) | Modifying rcParams globally might lead to unintended side effects (in shared codebases or collaborative environments) |

* 1. Install pmdarima  
     A screenshot of a computer

     Description automatically generated
  2. Import the necessary libraries and read the dataset  
     A screenshot of a computer program

     Description automatically generated
  3. Check the head  
     A screenshot of a computer screen

     Description automatically generated
  4. Give the new column names  
     A screenshot of a computer screen

     Description automatically generated
  5. Create another DataFrame with ‘Date’ and ‘USD’ columns only  
     A screenshot of a computer

     Description automatically generated
  6. Check for any null values and information  
     A screenshot of a computer program

     Description automatically generated  
     A screenshot of a computer

     Description automatically generated
  7. Set the “Date” column as an index  
     A screenshot of a computer

     Description automatically generated
  8. Check the info and save the cleaned dataset as CSV  
     A screenshot of a computer program

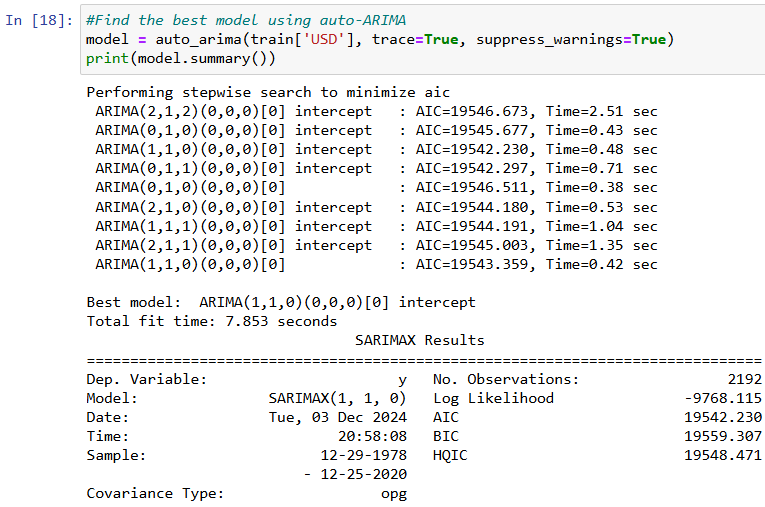
     Description automatically generated
  9. Visualize the cleaned\_data  
     A screenshot of a computer code

     Description automatically generated  
     A graph with a line going up

     Description automatically generated
  10. Graph the seasonal\_decompose to see the trend, seasonal and residual  
      A screenshot of a computer screen

      Description automatically generated
  11. Divide the data into training and test sets, then save  
      A screenshot of a computer program

      Description automatically generated

1. Analysis
   1. Report the data analysis process by describing the analysis technique(s) used. Include the calculations that are performed and the outputs.   
        
      Conduct auto\_arima function to find the best model for ARIMA  
        
      A screenshot of a computer

      Description automatically generated  
        
      Conduct ARIMA function with the best model order 1, 1, 0  
      A screenshot of a computer

      Description automatically generated  
      A screenshot of a computer

      Description automatically generated  
        
      Plot the diagnostics to visualize the standardized residual for “U,” histogram plus estimated density, normal Q-Q, and correlogram  
      A screenshot of a computer screen

      Description automatically generated  
        
      Predict the gold price based on the training set  
      A screenshot of a computer

      Description automatically generated  
        
      Set the date as an index  
      A screenshot of a computer

      Description automatically generated  
        
      A screenshot of a computer program

      Description automatically generated  
      A screenshot of a computer

      Description automatically generated  
        
      Trim the original cleaned\_df to start from Jan 7, 2000, instead of Dec 29, 1978, for the visualization purpose  
      A screenshot of a computer

      Description automatically generated  
        
      Graph the original data (from 1/7/2000 – 11/22/2024) with blue and predicted with green  
      A screen shot of a computer

      Description automatically generated  
        
      Calculate the RMSE and the accuracy of this time series  
      A screenshot of a computer program

      Description automatically generated
   2. Justify how the selected analysis technique(s) used to include one advantage and one disadvantage of these techniques

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Libary** | **Function** | **Used** | **Advantage** | **Disadvantage** |
| Numpy | Enable efficient processing of large volumes of numerical data | Used to calculate the square root on the MSE to calculate the rmse on the time series model | It uses less memory and storage space, making it easier and more convenient to use | Less flexibility than other storage array systems (Bigelow) |
| Auto\_arima | Identify the optimal set for the ARIMA model | Used to identify the best order for the ARIMA function, which was 1, 1, 0 | It saves the time enormously, and remove the requirement to grasp the statistics and theory underlying the model selection (Prasad) | It is hard to tell how other models perform, makes only one order is the best option(Prasad) |
| Mean\_squared\_error | Evaluate the accuracy of the model’s predictions by determining how closely they match the ground truth | Used MSE function to calculate the RMSE with numpy sqrt function on this time series model | The most robust and uses the same unit as the output variable (Aishwarya) | Larger error terms are not adequately penalized (Aishwarya) |
| ARIMA | Track the occurrences taking place over a specific time period | Output with SARIMAX result | Used for both short-term and the long-term trends, does not rely on any specific data distribution (Purwar) | Struggles with complex data patterns requires significantly more time to train on large datasets compared to other time series models, and is sensitive to outliers and extreme values (Purwar) |

1. Data Summary
   1. Summarize the implications of the data analysis by discussing the results, including one limitation of the analysis.  
       The time series model achieved an accuracy of 88.5%, effectively identifying underlying trends, seasonality, and cyclic behaviors in gold prices. This capability can provide valuable insights for broader economic strategies, as gold prices are often influenced by factors such as inflation rates, currency fluctuations, and geopolitical events. However, a limitation of this analysis is that the dataset included only dates and weekly gold prices, without additional variables that could impact gold prices. The screenshot below illustrates the model’s prediction of gold prices starting from 2011-03-04, using a 7:3 train-test split. The training set showed a steady upward trend with minimal fluctuations, contributing to the model achieving an 80% accuracy for this prediction.
   2. Recommend a course of action based on the results. Propose two directions or approaches for future study of the data set  
       The time series model’s predictions can help individuals capitalize on gold price fluctuations by purchasing during dips and selling during peaks. Gold also serves as a reliable safe-haven asset during economic uncertainty or high inflation.
2. In-text citations and references
   1. Aishwarya. *Regression Metrics – Of all metrics why MSE?* LinkedIn. (February 19, 2022). <https://www.linkedin.com/pulse/regression-metrics-all-why-mse-aishwarya-b/>
   2. Bigelow, Stephen. *What is NumPy? Explaining how it works in Python*. TechTarget. <https://www.techtarget.com/whatis/definition/What-is-NumPy-Explaining-how-it-works-in-Python#:~:text=NumPy%20is%20an%20open%20source,arrays%2C%20masked%20arrays%20and%20matrices>.
   3. Prasad, Eswara. *What is Auto-ARIMA?* Medium. (May 20, 2021). <https://medium.com/featurepreneur/what-is-auto-arima-b8025c6d732d>.
   4. Purwar, Siddhartha. *ARIMA Model*. Medium. (September 23, 2023). <https://siddp6.medium.com/arima-model-424a1b0f0970>.
   5. *Understanding the Price of Gold*. Gold Avenue. <https://www.goldavenue.com/en/precious-metals-guide/a-beginner-s-guide-to-gold/understanding-the-price-of-gold>.