Liana, Sonali, Tony

Extract Transform Load (ETL) Project

Data Science Bootcamp

For this ETL project we will use stock market data for ticker GLD from Yahoo Finance and the prices of gold from <https://datahub.io/core/gold-prices#resource-annual> to determine which is the better investment: physical gold or stock in GLD. The null hypothesis for this exercise is that due to the construction of the ETF, as a equity representation of physical gold, there should be no material difference. Although some differences in performance are evident. Those differences do not appear significant.

**Data Format**

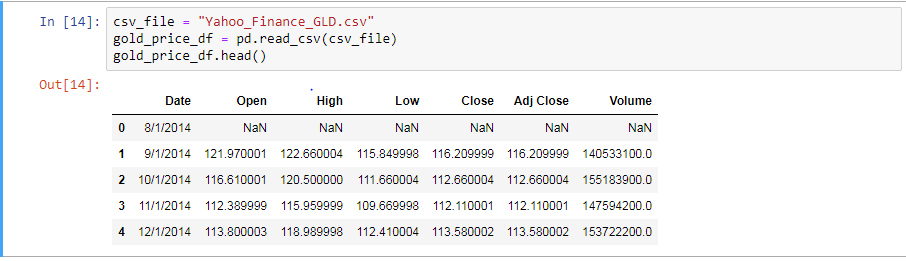
Stock market data from Yahoo Finance was provided as a CSV files. The file providing physical gold prices was provided as a JSON file.

**Data Extraction and Cleanup/Transformation**

We used pandas to extract and clean these csv files. Cleaning the data required to strip away any unnecessary variables and transforming dates as they were originally in different formats. We removed variables in the CSV file to only include date and closing prices every first of the month.

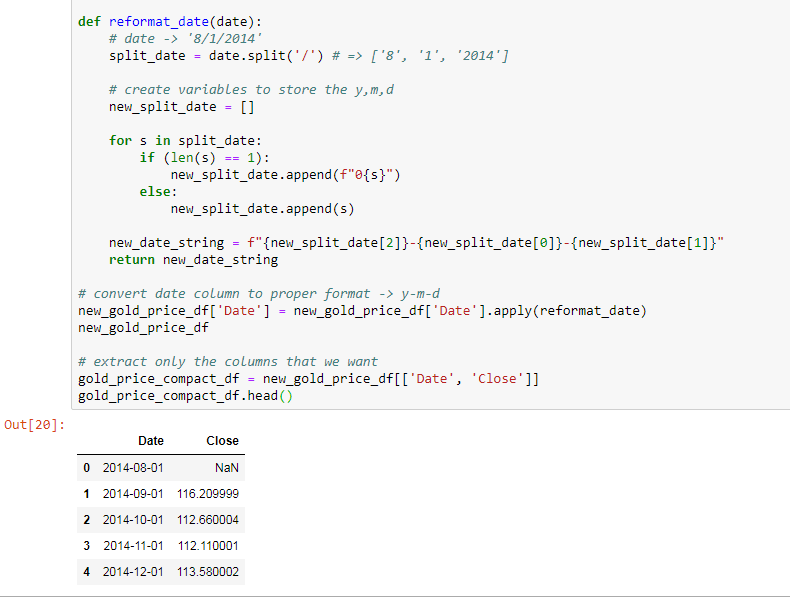
**Data Storage into a Database**

We created a connection using Jupyter Notebook. We created a database for both the CSV file and the JSON, pushed and joined the tables using Date as key for joining the files.

**Extraction: Screenshot 1 CSV file**

**Extraction: Screenshot 1 JSON file**

**Transformation- Screenshot 1: Transform the date CSV file to match format in JSON**

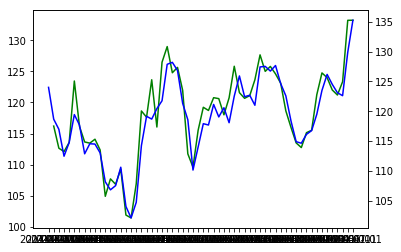


**Data Analysis**

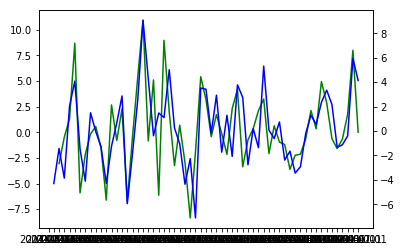
The data supported assertion that the ETF for gold (ticker symbol GLD) is virtually identical to the physical gold itself, identified in the data output by the currency ticker XAU.

Chart 1 shows the historical price changes of the two instruments on a monthly basis.

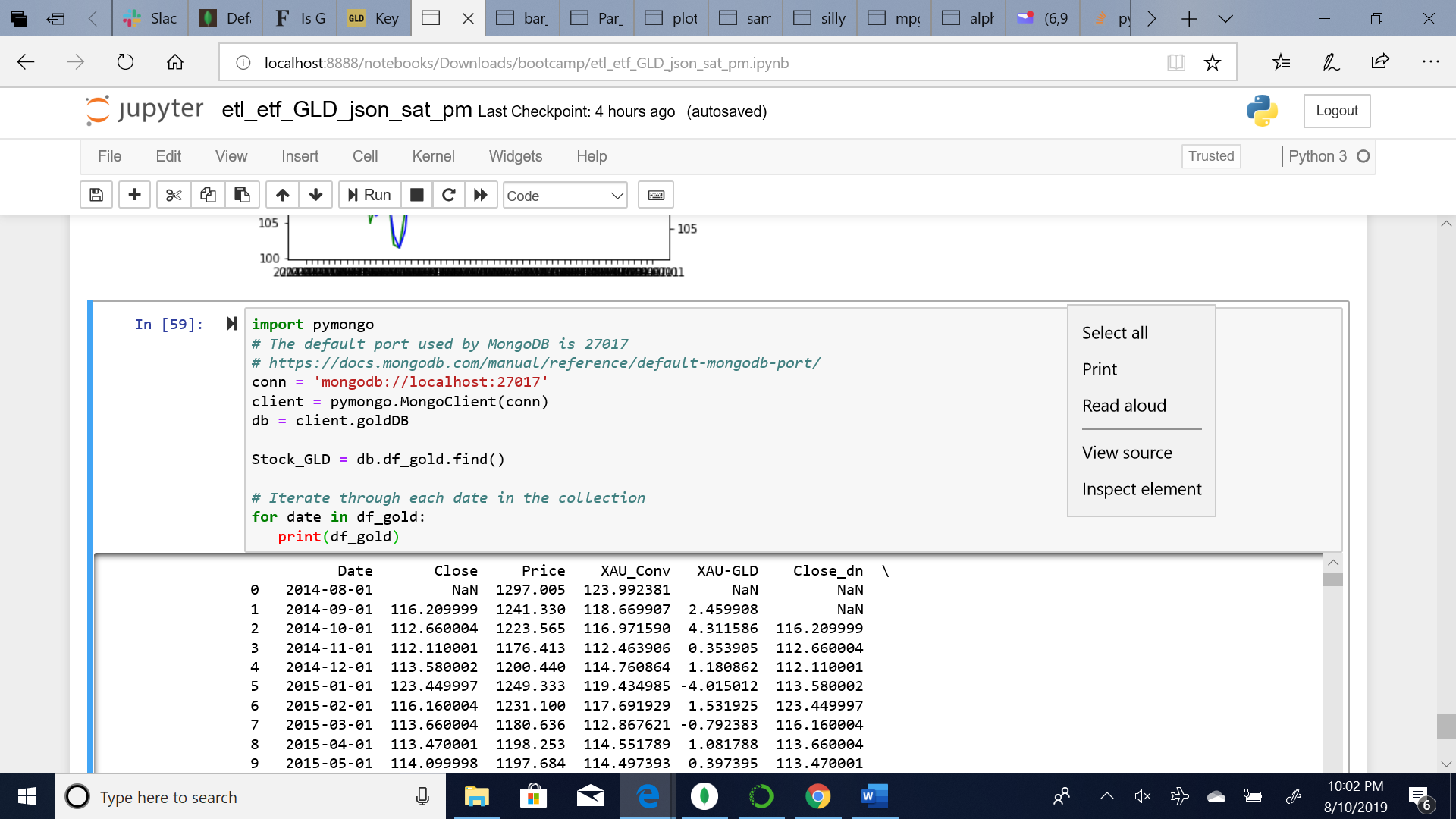
**Chart 1: Monthly Returns – GLD v. XAU**



**Chart 2 shows the historical returns**

****

The final output was loaded to MongoDB using the following code:



The Mongo DB database was identified as goldDB and the database contains three collections: CSV file input, JSON file input, and Python/Pandas Notebook dataframe output (df\_gold).

The null hypothesis that the group considered at outset was this: physical gold and the SPDR gold ETF should have nearly identical monthly returnes. This was in fact the case as seen from the charts for price history and the monthly returns.