**Project:** Extract Transform Load (ETL)

**Submitted by:** Liana Murray, Sonali Fulzele, Tony Simonelli

**Course:** Data Science Bootcamp

**Synopsis**

For this ETL project we are using 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 an equity representation of physical gold, there should be no material difference. Although some differences in performance are evident. Those differences do not appear significant.

**Extraction Process**

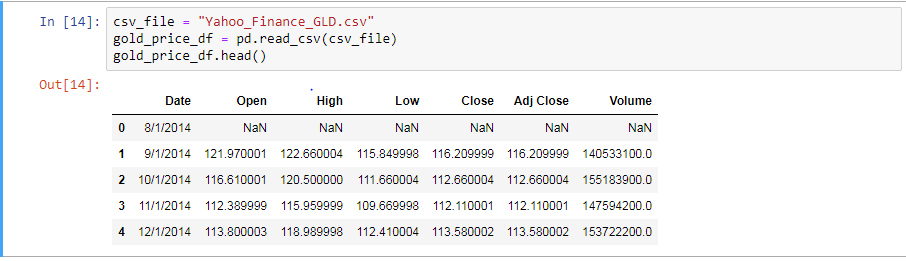
The data was extracted from two different sources. The first source used was Yahoo Finances for stock market data (SPDR GLD) ticker. The second source used was <https://datahub.io/core/gold-prices#resource-annual> to get the historic Gold prices year over year.

The data formats used for this project are as listed below

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

We used Pandas to extract the CSV and JSON files and converted them as Data Frames.

**Extraction: Screenshot 1 CSV file**



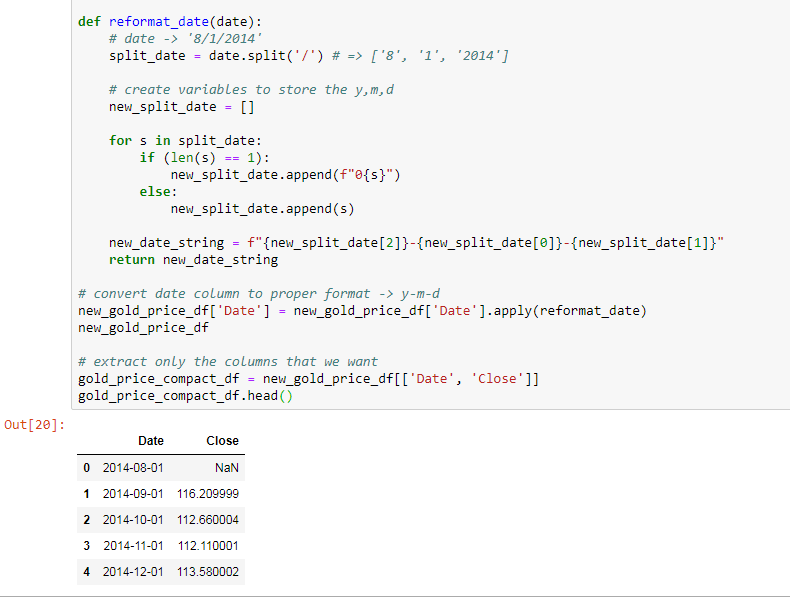
**Extraction: Screenshot 1 JSON file**



**Transformation Process**

The transformation process involved cleaning the data by stripping away any unnecessary piece of data and transforming dates as they were originally in different formats. We removed records from the CSV file to only include date and Closing Prices every first of the month. The cleansed and formatted Data Frames from these two formats were then merged using Date as a joining key.

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



**Loading Data Process**

We used MongoDB to load our data. The steps followed in this process were

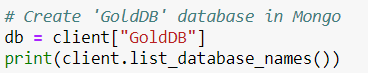
1. Create a connection to MongoDB
2. Create a Database named GoldDB
3. Create three collections as below
   * Gold\_Data
   * Stock\_Price\_Data
   * Reporting\_Data
4. Load Documents in the collection

The gold price data was loaded as documents in the Gold\_Data collection. The Stock price data was loaded in Stock\_Price\_Data collection and lastly the merged data was loaded in Reporting\_Data Collection.

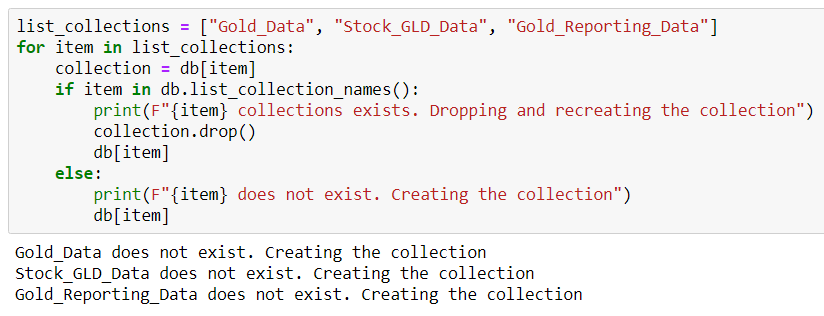
**Loading - Screenshot 1: Connection to MongoDB**



**Loading - Screenshot 2: Create the database GoldDB**



**Loading – Screenshot 3: Creating the collections**

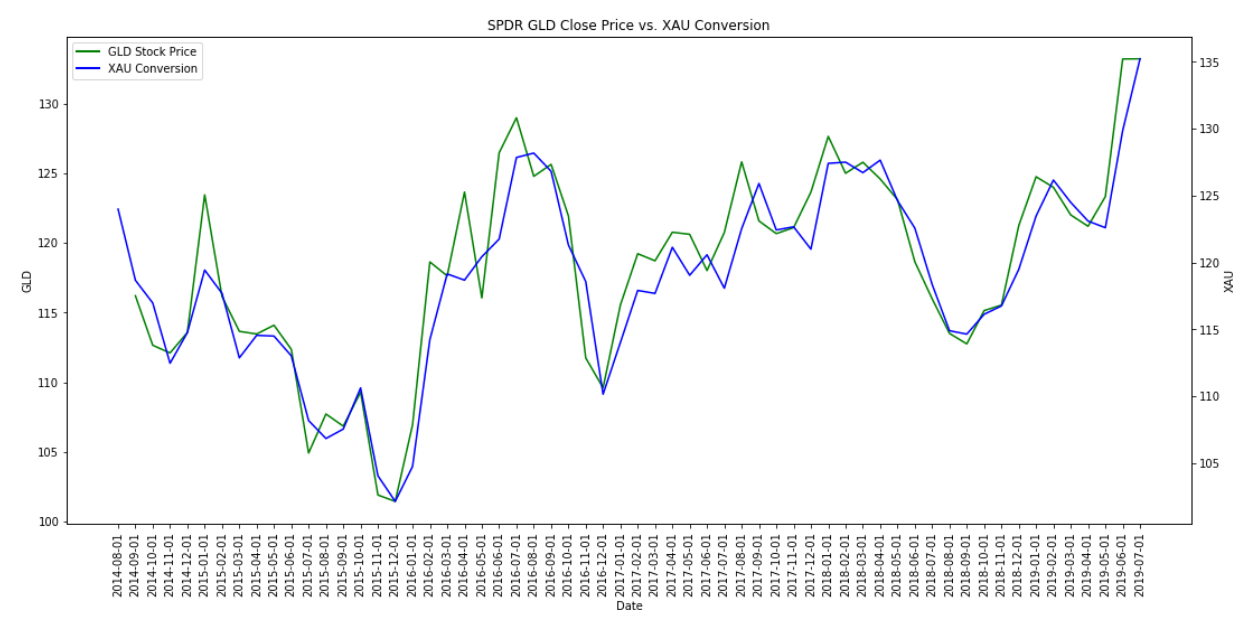


**Data Analysis**

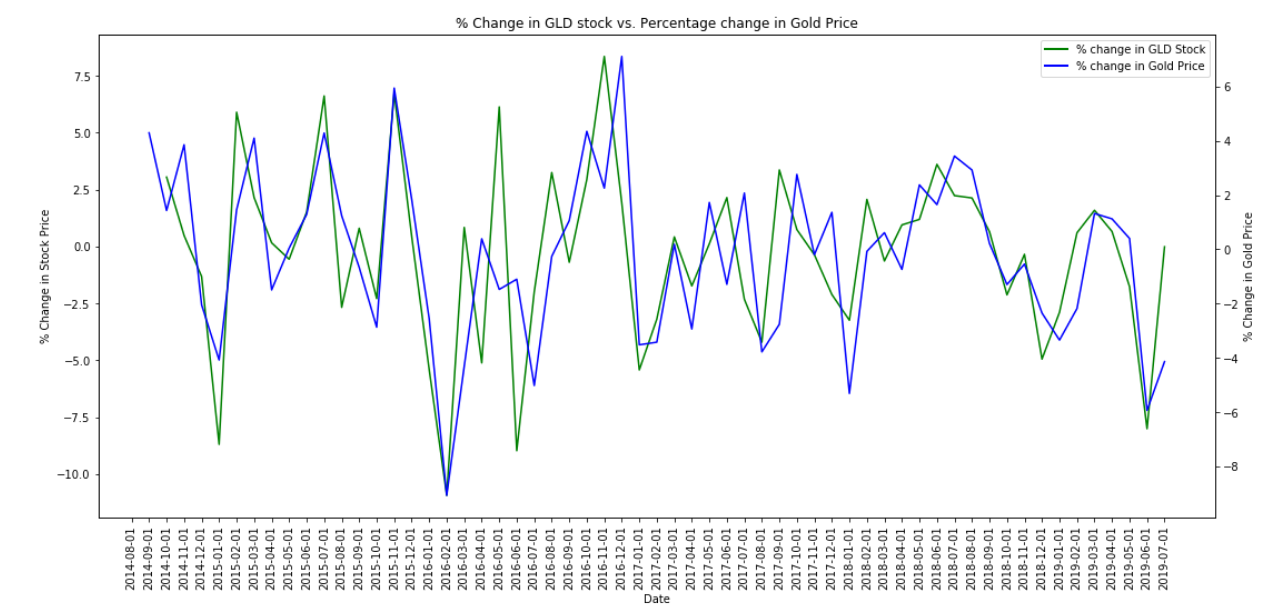
The data supported assertion that the ETF for gold (ticker symbol GLD) is virtually identical to the physical gold itself, identified in the data output by the currency ticker XAU.

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