Methods of Applied Statistics I (STA2101F): Project

Gold prediction dataset EDA

Abraham Morales

(a) Create a new `R project` for your final project. Create a new `R markdown` file to start recording the steps in your analysis. Write some code that reads your data into `R` from the original website where you obtained it, or from your own website that you create. (This is so I will be able to run your `.Rmd` file without actually storing your data on my computer.) The `R project` for this project is saved to the github repository: abraham-mv/gold_price_prediction.

(b) Load your data and do some quick quality checks – are there any missing values? If so, how many? How will you handle them in the analysis?

```
## [1] 1718   81
```

```
##  [1] "Date"          "Open"          "High"          "Low"          
##  [5] "Close"         "Adj.Close"     "Volume"        "SP_open"      
##  [9] "SP_high"       "SP_low"        "SP_close"      "SP_Ajclose"   
## [13] "SP_volume"     "DJ_open"       "DJ_high"       "DJ_low"       
## [17] "DJ_close"      "DJ_Ajclose"    "DJ_volume"     "EG_open"      
## [21] "EG_high"       "EG_low"        "EG_close"      "EG_Ajclose"   
## [25] "EG_volume"     "EU_Price"      "EU_open"       "EU_high"      
## [29] "EU_low"        "EU_Trend"      "OF_Price"      "OF_Open"      
## [33] "OF_High"       "OF_Low"        "OF_Volume"     "OF_Trend"     
## [37] "OS_Price"      "OS_Open"       "OS_High"       "OS_Low"       
## [41] "OS_Trend"      "SF_Price"      "SF_Open"       "SF_High"      
## [45] "SF_Low"        "SF_Volume"     "SF_Trend"      "USB_Price"    
## [49] "USB_Open"      "USB_High"      "USB_Low"       "USB_Trend"    
## [53] "PLT_Price"     "PLT_Open"      "PLT_High"      "PLT_Low"      
## [57] "PLT_Trend"     "PLD_Price"     "PLD_Open"      "PLD_High"     
## [61] "PLD_Low"       "PLD_Trend"     "RHO_PRICE"     "USDI_Price"   
## [65] "USDI_Open"     "USDI_High"     "USDI_Low"      "USDI_Volume"  
## [69] "USDI_Trend"    "GDX_Open"      "GDX_High"      "GDX_Low"      
## [73] "GDX_Close"     "GDX_Adj.Close" "GDX_Volume"    "USO_Open"     
## [77] "USO_High"      "USO_Low"       "USO_Close"     "USO_Adj.Close"
## [81] "USO_Volume"
```

We have 1718 observations and 81 columns. This columns correspond to different assets and indexes from the stock market. For example, the columns labeled as "SP" and "DJ" correspond to the Standard & Poor's and Dow Jones stock market indexes respectively, while "USO" refers to the United States Oil Fund. We know that the first columns: Open, High, Low, Close, Adj.Close and Volume are for gold, we'll take "Adj.Close" as our dependent variable. This data was collected from December 2011 to December 2018.
We convert the `Date` column to date format and check for null values.

```
## 
##  FALSE 
## 139158 
```

No values labeled as "NA" in the dataframe; however, there could still be labeled as zero. Since we are working with time series financial data, in doesn't make sense to have values at zero.

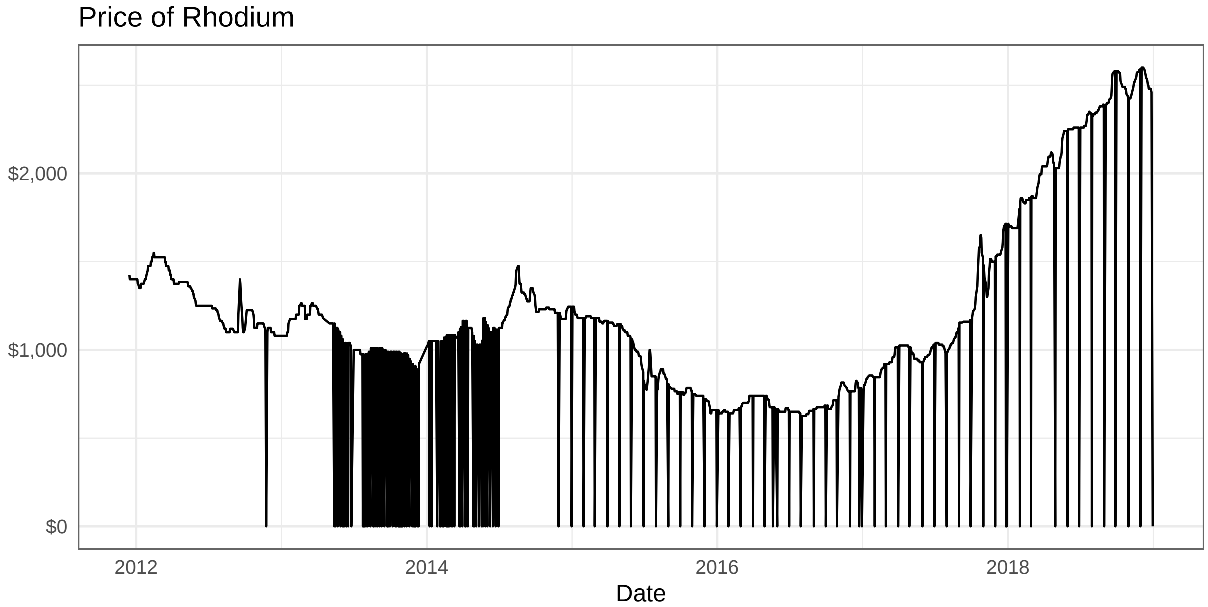
We can run a quick summary of some of the columns, just to show some inconsistencies in the data.

```
##       Date              Adj.Close       SP_Ajclose       EU_Trend
##  Min.   :2011-12-15   Min.   :100.5   Min.   :104.5   Min.   :0.0000
##  1st Qu.:2013-10-03   1st Qu.:116.1   1st Qu.:153.0   1st Qu.:0.0000
##  Median :2015-07-18   Median :121.8   Median :191.7   Median :0.0000
##  Mean   :2015-07-06   Mean   :127.3   Mean   :192.2   Mean   :0.4948
##  3rd Qu.:2017-04-09   3rd Qu.:128.5   3rd Qu.:228.7   3rd Qu.:1.0000
##  Max.   :2018-12-31   Max.   :173.6   Max.   :290.6   Max.   :1.0000
##     OF_Trend          SF_Price       RHO_PRICE     USO_Adj.Close
##  Min.   :0.0000   Min.   :33170   Min.   :   0   Min.   : 7.96
##  1st Qu.:0.0000   1st Qu.:38019   1st Qu.: 785   1st Qu.:11.39
##  Median :0.0000   Median :40522   Median :1100   Median :16.34
##  Mean   :0.4988   Mean   :43284   Mean   :1130   Mean   :22.11
##  3rd Qu.:1.0000   3rd Qu.:46581   3rd Qu.:1308   3rd Qu.:34.42
##  Max.   :1.0000   Max.   :65292   Max.   :2600   Max.   :42.01
```

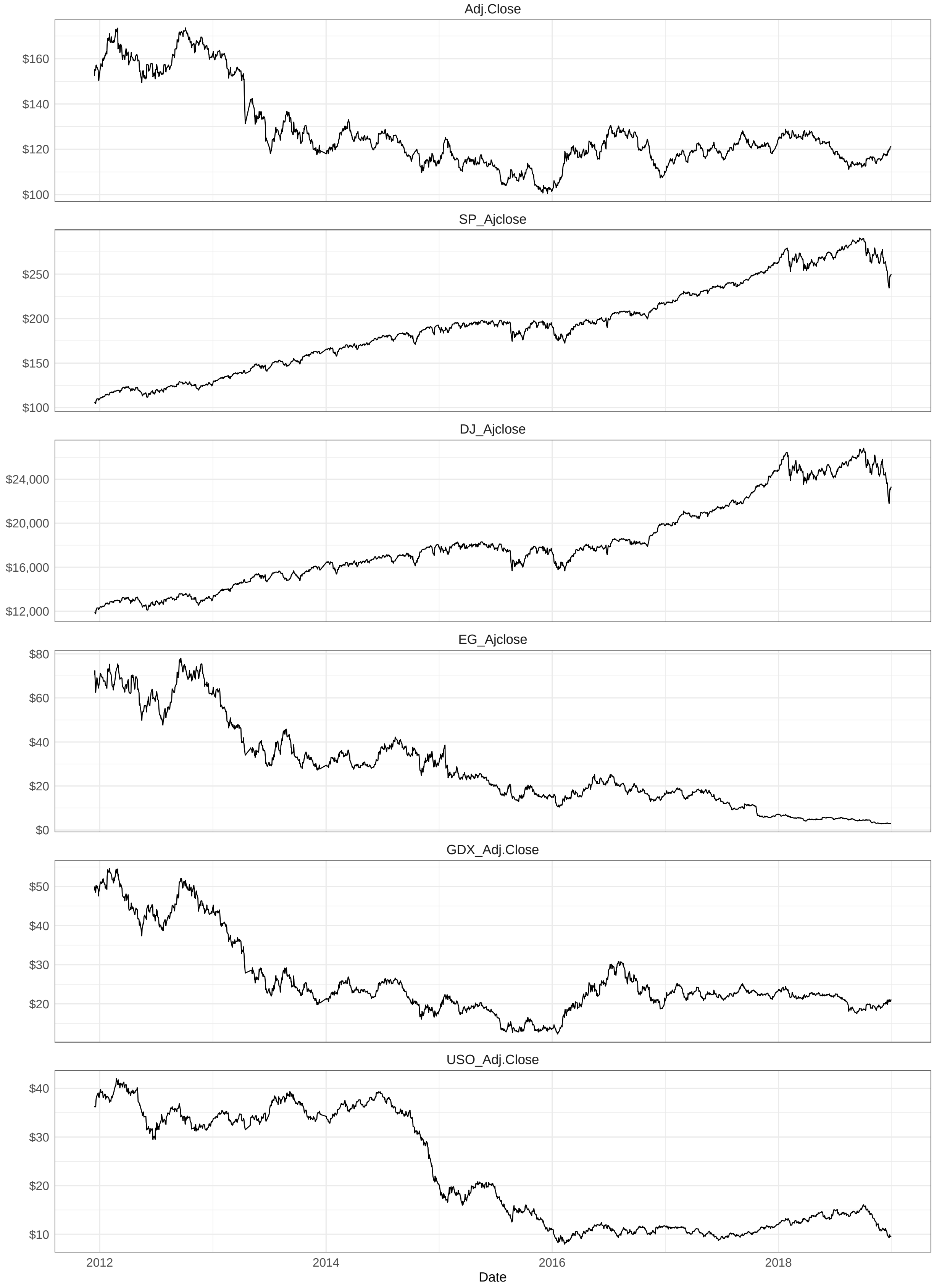
It appears that the variables with suffix Trend are categorical, coded as 0 and 1. We should confirm this as follows:

```
##  EU_Trend OF_Trend OS_Trend SF_Trend USB_Trend PLT_Trend PLD_Trend USDI_Trend
##  0:868    0:861    0:853    0:892    0:876     0:886     0:806     0:837
##  1:850    1:857    1:865    1:826    1:842     1:832     1:912     1:881
```

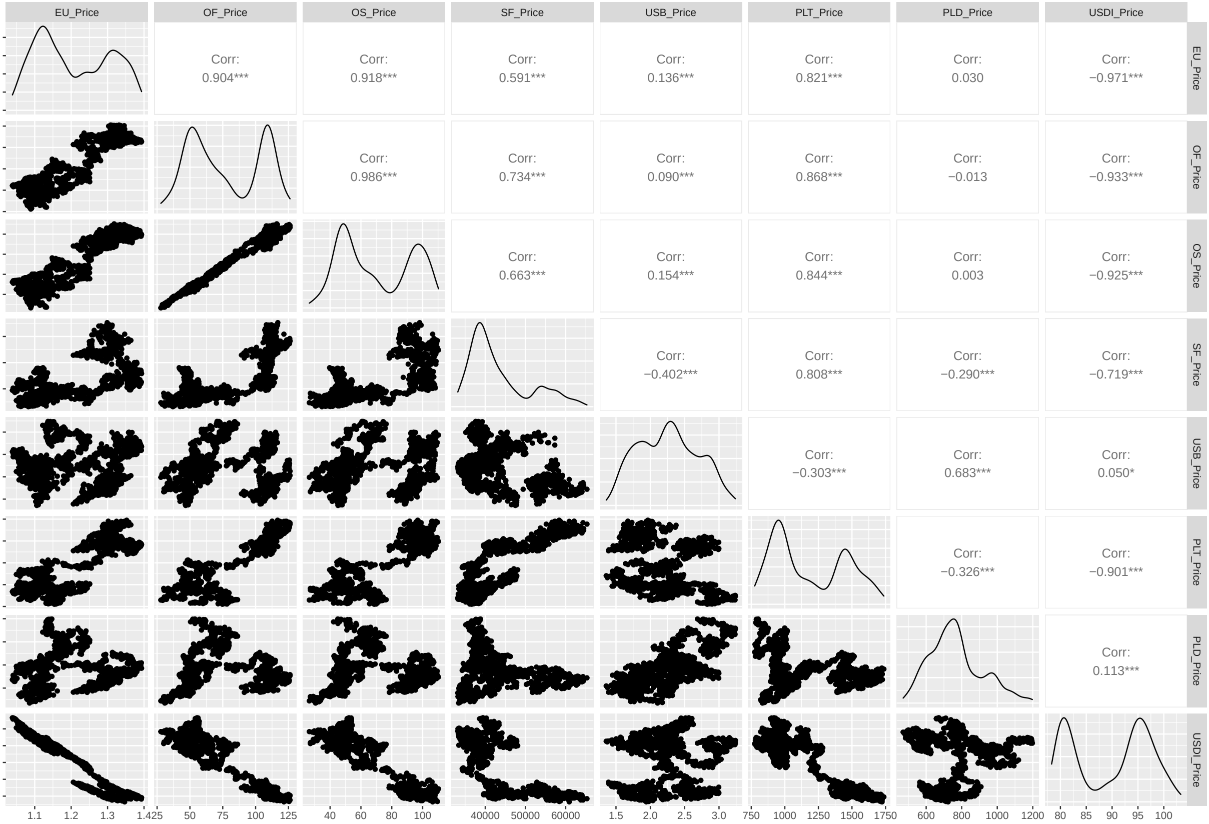
(c) Construct some preliminary plots of the data, for example histograms, boxplots, and/or scatterplots, and comment on any anomalies.
We can see that the variable `RHO_price` has minimum value of zero, which doesn't make a lot of sense.



`RHO_Price` is the only variable with zeros, other than the binary ones. We could exclude this covariate from the analysis, however, its trend is easily spottable, so we could estimate missing values with simple interpolation, or use smoothing techniques such as splines, kernel or a moving average filter. Let's take a few plots from our time series data.


Adj.Close
$160
$140
$120
$100
SP_Ajclose
$250
$200
$150
$100
DJ_Ajclose
$24,000
$20,000
$16,000
$12,000
EG_Ajclose
$80
$60
$40
$20
$0
GDX_Adj.Close
$50
$40
$30
$20
USO_Adj.Close
$40
$30
$20
$10
2012
2014
2016
2018
Date

Let's take correlation plots for the price covariates.



We can see that there are a few covariates that are highly correlated; for example, the price of USDI and EU show a correlation of -0.971, the prices of OS and OF show a correlation of 0.986. To deal with this issue, we could ignore one of those two variables that show high correlation, since they wouldn't add any valuable information to the model.