# Analysis of Influential Data on US Sector ETFs and the Potential for Forecasting Price Movements

Team 5

April 2018

## Members


- Preeti Bekal
- Wyatt Marciniak
- Neel Shah


## Project Idea

We will be analyzing influential relationships within the US stock market. To do this we will start by compiling index, sector Exchange Traded Funds (ETF), commodity prices, market/economic indicators and exchange rates and analyzing behaviors of certain relationships over time. The goal of project contains 2 parts:

- 1. Analyze influential relationships and their behaviour through and across time subsets
- 2. Employ methods for predicting 'up' and 'down' movements per period

To analyze the relationships, we will analyze correlations and simple regressions across multiple time periods and rank the influence of our factors based on these results. We will also use K-means and hierarchical clustering methods to analyze these relationships in a more dynamic way. This will provide us with a benchmark for building forecast models as well as allow us to discuss how market relationships have been maintained or changed over time. To predict price movements of the ETF securities, we will construct multivariate linear regression models and conduct testing using supervised decision trees as well. The factors we will use will be determined by the initial relationship analysis. Our goal here is to compare these prediction models and establish some process to which we can conclude with some level of certainty the future movements of these ETFs. By extension, we are trying to forecast sector price movements that could act as a complimentary analysis to macroeconomic analysis and asset allocation decisions. It should also be noted that the objective here is to predict direction of returns and not the magnitude as we believe directional predictions will be more accurate and useful in asset allocation decisions. We currently have a working script for data extraction and cleaning, which we will comment on in the following sections.

## Data Sources and Extraction Methodology

We will be using the 'quantmod' package in R to extract all of our data for this project. The sources used will be Yahoo finance (for indexes and equity data) and the FRED database (for economic, exchange rate and commodities data). Equity data sets are extracted using a set time frame ('to', 'from') and FRED (economic) data sets are extracted in their full time frames from the data sets inception date (Figure: 1). They will then be adjusted for the proper analysis time frame thereafter.

```
data_pull   <- function(type, tick, from, to)          # Pull stock/etf/indicator data
{
  if(type == "indicator")
  {
    cnames <- c("Data")
    getSymbols.FRED(tick, env = .GlobalEnv, return.class = "xts")
    temp <- data.frame(get(tick))
    colnames(temp) <- cnames
    #rm(get(tick))
    temp
  }
  else if(type == "stock/etf")
  {
    cnames <- c("Open","High","Low","Close","Volume","Adjusted")
    getSymbols(tick, from = from, to = to)
    if(tick == "^GSPC") { temp <- data.frame(GSPC) }
    else { temp <- data.frame(get(tick)) }
    colnames(temp) <- cnames
    temp
  }
  else { bad <- "Invalid data type to pull"; bad }
}
```

Figure 1: Function: Data Pull (Pulls data sets according to identifying code and source)

## Data Scrubbing

Scrubbing the data consists of 4 core steps:

- Step 1: Convert raw data sets into data frames of price/value and date
- Step 2: Subset the data into the proper periods (daily, weekly or monthly) (Figure: 2)
- Step 3: Combine the data, bound by the appropriate dates for accuracy, into combined data sets
- Step 4: Calculate the respective period returns (Figure: 3)

The objective here is to look at relationships that correspond to daily, weekly and monthly time periods. We are looking to see how relationships change over these time frames to see if certain subsets reveal stronger or weaker relationships. To complete this testing, we are also utilizing multiple factors that are published only weekly or monthly. We hope to conduct a full analysis to report as accurate results as possible.

```
comp.data   <- function(type, dlist, cnames, from, to)  # Put data sets together, match by date sequence
{
  # First check for time frame (days, months, years)
  if(type == "d")
  {
    dates <- seq(as.Date(from), as.Date(to), by = "day")      # Create a data list for subsetting
    dates <- as.character(dates)                              # Make data character type for compatibility
  }
  else if (type == "m")
  {
    dates <- seq(as.Date(from), as.Date(to), by = "month")    # Create a data list for subsetting
    dates <- as.character(dates)                              # Make data character type for compatibility
  }
  else
  {
    dates <- seq(as.Date(from), as.Date(to), by = "quarter")  # Create a data list for subsetting
    dates <- as.character(dates)                              # Make data character type for compatibility
  }

  for(i in 1:length(dlist))
  {
    print(noquote(paste("Data Frame:",i," - ",cnames[i])))
    if(i == 1)
    {
      temp.df             <- dlist[[i]]
      values              <- data.frame(temp.df[dates, ])
      rownames(values)    <- dates
      new.lst             <- data.frame(values)
    }
    else
    {
      temp.df             <- dlist[[i]]
      values              <- data.frame(temp.df[dates, ])
      rownames(values)    <- dates
      new.lst <- cbind(new.lst, values)
    }
  }

  colnames(new.lst) <- cnames
  new.lst
}
```

Figure 2: Function: Compile Data (Subset data by respective dates for select periods)

```r
calc.ret    <- function(data) # Calculate returns for 1 column data frame
{
  rnames <- rownames(data); rnames <- rnames[-1]
  temp <- c()

  for(i in 1:length(data[,1]))
  {
    temp <- c(temp, log(data[i,1]/data[i-1,1]))
  }
  new.df <- data.frame(temp)
  colnames(new.df) <- c("Data")
  rownames(new.df) <- rnames
  new.df
}

calc.ret.df <- function(data) # Calculate return for 2+ column data frame
{
  rnames <- rownames(data); rnames <- rnames[-1]
  cnames <- colnames(data)
  new.df <- data.frame(matrix(nrow = length(data[,1])-1))

  for(i in 1:length(data))
  {
    temp <- c()
    for(j in 1:length(data[,i]))
    {
      temp <- c(temp, log(data[j,i]/data[j-1,i]))
    }
    new.df <- cbind(new.df, temp)
  }
  new.df <- new.df[-1]
  colnames(new.df) <- cnames
  rownames(new.df) <- rnames
  new.df
}
```

Figure 3: Function: Calculate Return (Calculates period return for 1 or more column data sets)

## Data Sets

The current data sets we have compiled and will use are as follows:

| Data Set | Description |
|---|---|
| S&P 500 Index | Broad Market Index for top 500 US companies |
| Dow Jones Industrial Average | Core 30 companies that represent the US economy |
| Nasdaq | US Technology Sector Index |
| Financial Sector SPDR ETF | ETF tracking US Financial Sector |
| Technology Sector SPDR ETF | ETF tracking US Technology Sector |
| Health Care Sector SPDR ETF | ETF tracking US Health Care Sector |
| Energy Sector SPDR ETF | ETF tracking US Energy Sector |
| Industrial Sector SPDR ETF | ETF tracking US Industrial Sector |
| Consumer Discretionary Sector SPDR ETF | ETF tracking US Consumer Discretionary Sector |
| Consumer Staples Sector SPDR ETF | ETF tracking US Consumer Staples Sector |
| Utilities Sector SPDR ETF | ETF tracking US Utilities Sector |
| Materials Sector SPDR ETF | ETF tracking US Materials Sector |
| Biotechnology Sector SPDR ETF | ETF tracking US Biotechnology Sector |
| US/Euro Exchange rate | US dollar to Euro zone Euro exchange rate |
| US/Euro Exchange rate | US Dollar to Euro zone Euro exchange rate |
| Yen/US Exchange rate | Japanese Yen to US Dollar exchange rate |
| Yuan/Euro Exchange rate | Chinese Yuan to US Dollar exchange rate |
| CD/US Exchange rate | Canadian Dollar to US Dollar exchange rate |
| Peso/Euro Exchange rate | Mexican Peso to US Dollar exchange rate |
| Won/US Exchange rate | South Korean Won to US Dollar exchange rate |
| Real/US Exchange rate | Brazilian Real to US Dollar exchange rate |
| Rupee/US Exchange rate | Indian Rupee to US Dollar exchange rate |
| Franc/US Exchange rate | Swiss Franc to US Dollar exchange rate |
| Federal Funds Rate | US Federal Funds Rate (Core US Interest Rate Benchmark) |
| Vix Index | US Stock Market Volatility Index |
| University of Michigan Sentiment | US Consumer Sentiment Survey Index |
| Oil Price | US Crude oil price per barrel |
| Case Schiller Index | US Home Price Index |
| US Dollar Index | US Dollar value against a basket of currencies |
| Federal Debt | US Public Debt as percentage of GDP |
| US Unemployment Index | US unemployment rate |
| Gold Price | Spot Gold Price Index |

The 3 indexes we are analyzing cover the behavior of the US economy and stock markets as a whole. We will use these indexes to monitor the broad market movements. We expect, however, that these indexes will not be used heavily as they encompass too many components to specific relationship analysis. The 10 sector ETFs are the dependant or observation variables we will be testing the relationships for. We will identify, using remaining data sets, what factors are most influential for determining sector value movements and attempt to predict future movements. The 10 exchange rate data sets will be used to analyze the relationships of the world's top economies with the US and how global trade and investor expectation drive sector movements. The 9 economic indicators will be used to assess the influence of US economic output, consumer behaviour, debt market and stock market behaviors on individual US sectors. We will be compiling additional data sets pertaining to commodity price indexes after vetting a few options from the FRED database. Commodity prices are key drivers of manufacturing, consumer sales and business operations. We have included our compiled data sets with the proposal submission as they are large and previews of the data sets shown here would be too small to review.

## Methodologies for Analysis

- Correlation and Simple Linear Regression Analysis (for relationship analysis)
- K-means and Hierarchical Clustering (for relationship analysis)
- Multivariate Linear Models (for forecasting movements)
- Supervised Decision Trees (for forecasting movements)

We will be building functions for automating analysis as well as utilizing an array of R packages. A complete list will be given with corresponding functions utilized when we have verified the final functionalities to use. Currently, we are using:

- quantmod - Data retrieval
- corrplot - Correlation Matrix Visualization
- fbasics/fTrading - Time series and Return Data Analysis
- broom - Summary Object Manipulation (extraction of Summary components)

## Next Steps

We will finalize our data sets and begin analysis testing to pool the most influential data parameters for forecasting price movements. We expect that by the first update we will have these sets selected and have begun conducting forecast testing for out final submission.