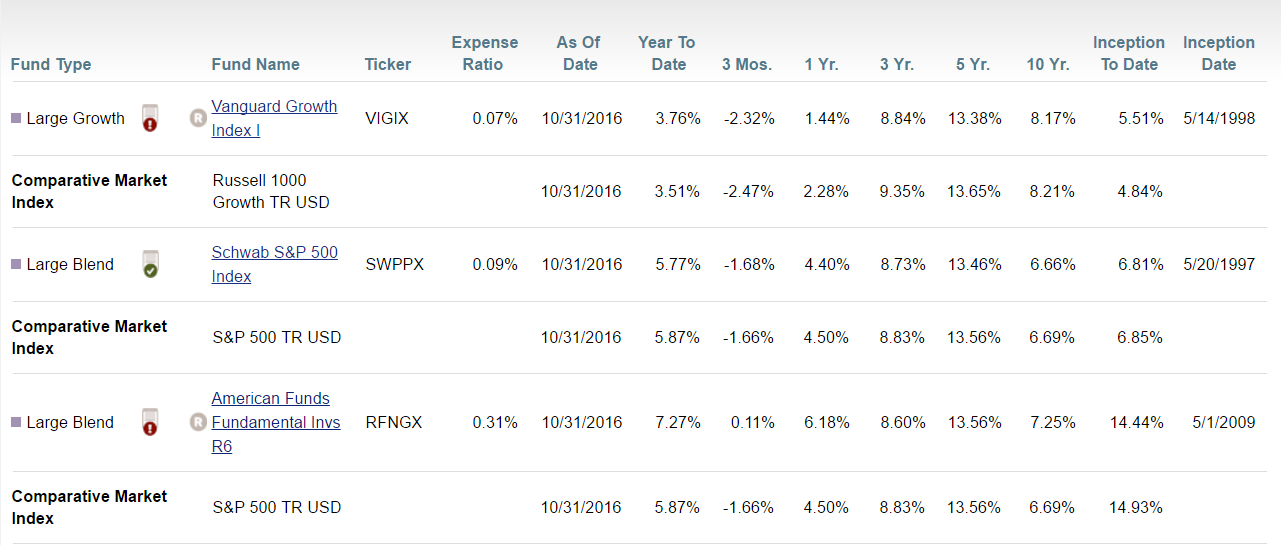
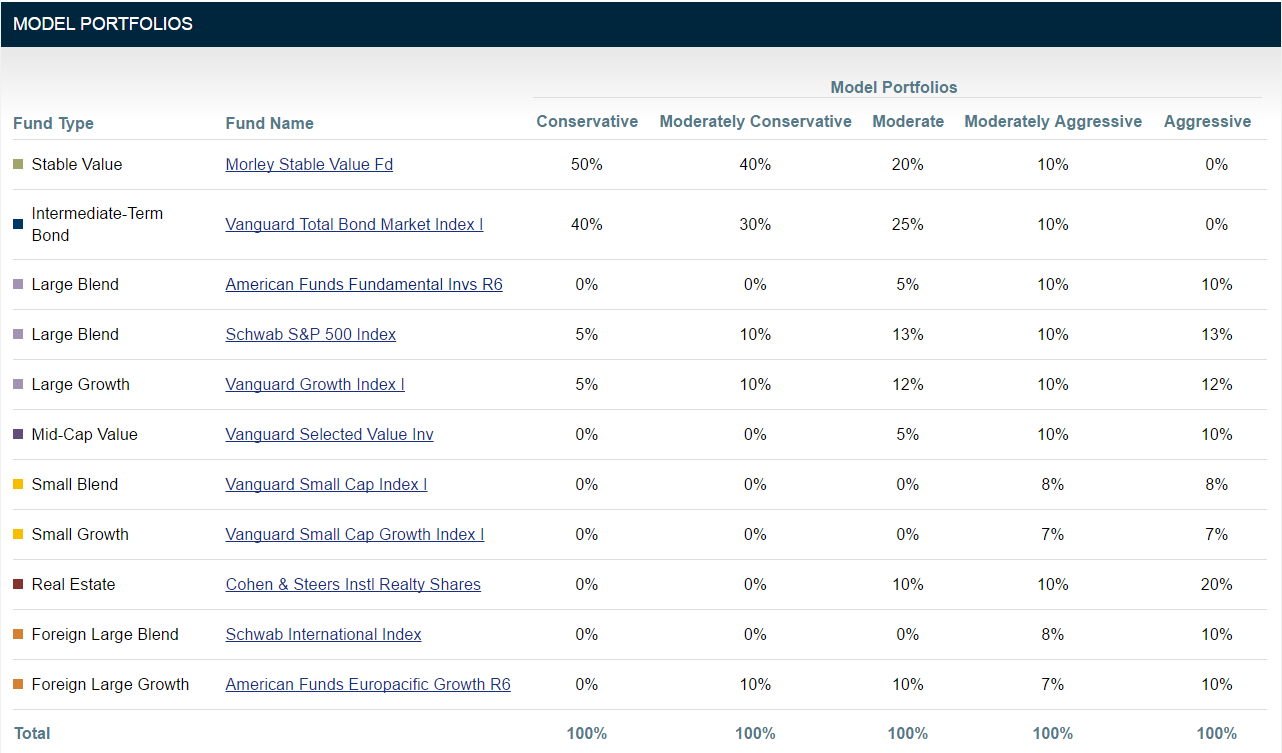
# Problem Definition

The 401K portfolio is an investment that any investor needs to be continuously monitoring to ensure that his/her hard earned money is being put to good use. Unfortunately, in many cases a lack of understanding on how to analyze fund data along with schedule constraints makes the management of this task cumbersome which results in missed opportunities and loss of money. In addition, the 401K management interface, does not have a robust suggestion feature that users can avail to make investment decisions. Investors are presented with one of two options:

* Make their own investment decisions using the fund information feature



* Use the model portfolio feature that looks like below to allocate funds by default



In both cases above, the crucial element missing is factoring in the actual performance of funds in the economy which is not ideal. The problem that this project is attempting to solve is to provide this information to an investor in a timely fashion, so that he/she can continuously visit their portfolio and make informed investment decisions.

# Data Set

The day end prices for all funds can be obtained from Quandl. The data elements available through the API were recorded into data frames with following information.

* Date – Date of record
* Adjusted Close – Day end price of the 401K fund
* Ticker – Ticker symbol

Additionally, a few derived variables have been added to add more meaning to the dataset. The approach is to create subsets for each ticker.

* Index – Increase/decrease from previous day
* Indicator – 1 if increase; 0 if decrease

Once above subsets have been created, they are summarized/aggregated together into a single data frame based on which the final model will be created. The fields in the new data frame are:

* Ticker
* Year
* Quarter
* Month
* No\_Growth\_Days – Count of days with positive growth
* No\_No\_Growth\_Days – Count of days with negative growth
* No\_Positive\_Runs – Highest continuous streak of positive growth
* No\_Negative\_Runs – Highest continuous streak of negative growth
* Month\_Begin\_Price – Ticker price at the beginning of the month
* Month\_End\_Price – Ticker price at the end of the month
* Range – Month end price – Month begin price
* Growth\_Ratio – Percentage of growth that was continuous
* No\_Growth\_Ratio – Percentage of negative growth that was continuous

# Analysis

A summary table with the findings are as below with information on the recommended portfolio strategies and percentage allocations. Data range from 2013 – 2016. It is reliable to assume after the analysis that the suggested portfolios do result in good returns if invested over a longer duration. But the analysis was required to validate the data.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | % Months Positive Growth | Ticker Price Improvement | Conservative | Moderately Conservative | Moderate | Moderately Aggressive | Aggressive |
| RFNGX | 68.75 | 19.35 |  |  | 5% | 10% | 10% |
| VSCIX | 64.58 | 25.02 |  |  |  | 8% | 8% |
| VSGIX | 64.58 | 13.00 |  |  |  | 7% | 7% |
| SWPPX | 62.50 | 12.44 | 5% | 10% | 13% | 10% | 13% |
| VASVX | 60.42 | 10.23 |  |  | 5% | 10% | 10% |
| VIGIX | 58.33 | 19.38 | 5% | 10% | 12% | 10% | 12% |
| CSRIX | 58.33 | 7.50 |  |  | 10% | 10% | 20% |
| VBTIX | 56.25 | 0.66 | 40% | 30% | 25% | 10% |  |
| SWISX | 54.17 | 1.90 |  |  |  | 8% | 10% |
| LSBDX | 54.17 | 1.32 |  |  |  |  |  |
| RERGX | 50.00 | 6.00 |  | 10% | 10% | 7% | 10% |
| VIPSX | 50.00 | -0.52 |  |  |  |  |  |
| DFEMX | 45.83 | -2.31 |  |  |  |  |  |
| DEMSX | 41.67 | -1.02 |  |  |  |  |  |
| Morley\_Stable |  |  | 50% | 40% | 20% | 10% |  |

The generic trend on most funds has been an upward movement in the ticker price as shown below.



Special notes from the analysis:

Tickers VASVX, VSCIX and VSGIX showed positive results in the month of February for all years

Tickers DEMSX, LSBDX, SWISX showed positive results in the month of February for all years

Tickers RFNGX, SWPPX, VIGIX, VSCIX, VSGIX showed positive results in the month of May for all years

Tickers VIPSX showed positive results in the month of July for all years

Tickers VSCIX, VSGIX showed positive results in the month of November for all years

The approach to better understand data was to conduct a linear regression and the results were as follows:

Call:

lm(formula = Month\_End\_Price ~ No\_Growth\_Days + No\_No\_Growth\_Days +

No\_Positive\_Runs + No\_Negative\_Runs + Ticker + as.factor(Month),

data = summary\_funds\_df, subset = Year == 2013 | Year ==

2014 | Year == 2015 | Year == 2016)

Residuals:

Min 1Q Median 3Q Max

-12.179 -1.153 0.181 1.802 9.890

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 40.514731 3.919584 10.336 < 2e-16 \*\*\*

No\_Growth\_Days 0.008633 0.199753 0.043 0.965541

No\_No\_Growth\_Days -0.095325 0.194167 -0.491 0.623635

No\_Positive\_Runs -0.081839 0.119711 -0.684 0.494448

No\_Negative\_Runs -0.034348 0.126505 -0.272 0.786078

TickerDEMSX -21.816533 0.716826 -30.435 < 2e-16 \*\*\*

TickerDFEMX -17.004421 0.711973 -23.884 < 2e-16 \*\*\*

TickerLSBDX -27.668558 0.718996 -38.482 < 2e-16 \*\*\*

TickerRERGX 4.305813 0.707481 6.086 1.99e-09 \*\*\*

TickerRFNGX 5.534063 0.706293 7.835 1.94e-14 \*\*\*

TickerSWISX -23.238023 0.708030 -32.821 < 2e-16 \*\*\*

TickerSWPPX -11.575618 0.707363 -16.364 < 2e-16 \*\*\*

TickerVASVX -15.571401 0.707061 -22.023 < 2e-16 \*\*\*

TickerVBTIX -30.687745 0.715969 -42.862 < 2e-16 \*\*\*

TickerVIGIX 8.984572 0.706822 12.711 < 2e-16 \*\*\*

TickerVIPSX -27.846904 0.714582 -38.969 < 2e-16 \*\*\*

TickerVSCIX 11.072119 0.707595 15.648 < 2e-16 \*\*\*

TickerVSGIX -7.411023 0.707886 -10.469 < 2e-16 \*\*\*

as.factor(Month)2 0.720255 0.685955 1.050 0.294111

as.factor(Month)3 1.483569 0.685866 2.163 0.030904 \*

as.factor(Month)4 1.751903 0.687796 2.547 0.011093 \*

as.factor(Month)5 1.904342 0.669730 2.843 0.004605 \*\*

as.factor(Month)6 1.929973 0.685350 2.816 0.005011 \*\*

as.factor(Month)7 2.427354 0.701631 3.460 0.000577 \*\*\*

as.factor(Month)8 1.954140 0.716502 2.727 0.006559 \*\*

as.factor(Month)9 1.833611 0.664794 2.758 0.005978 \*\*

as.factor(Month)10 2.562745 0.756838 3.386 0.000752 \*\*\*

as.factor(Month)11 2.786380 0.658866 4.229 2.69e-05 \*\*\*

as.factor(Month)12 2.673894 0.699507 3.823 0.000145 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.46 on 643 degrees of freedom

Multiple R-squared: 0.9458, Adjusted R-squared: 0.9434

F-statistic: 400.7 on 28 and 643 DF, p-value: < 2.2e-16

While testing for assumptions of linear regression, it was found that while the data fit well into the model, the premise for using a multiple linear regression would not be accurate as explained by below diagnostics:

