Final Report

Background

Exchange traded funds (or ETFs) are funds that track a benchmark or index (such as the S&P 500 or the price of oil in the near future). They usually contain shares or units of a single type of security that have something in common. They have become a safe, convenient way for retail investors to invest their savings without having to worry about diversification. This analysis will look at whether certain characteristics of an ETF can help predict its 5 year return, a metric that anybody looking to invest will care about.

The dataset we used for our analysis includes over 2,000 ETFs and was scraped from Yahoo Finance. Using this dataset, we intend to predict the 5 year fund return on an ETF. Some quantitative variables that might be useful in our analysis would be Net Assets and the proportion of investment allocated in each economic sector. Other quantitative variables that may help us would be ratios like price to book-value, cash-flow, and earnings. Categorical variables that could factor into any relationship would be investment type and fund size.

There have been a number of studies that have tried to answer questions similar to the one we are trying to answer. One study by Zhong & Enke used artificial neural networks to try and predict the daily return of the S&P 500. This study was similar to ours in that it looked at the return of an ETF and used some of the same predictor variables. However, it was also different from our analysis since it used AI and machine learning, as opposed to our multiple-regression model. Additionally, it also looked at economic indicators, instead of ETF-specific metrics.

Another similar study carried out at the University of Manchester used a Random Forest algorithm to determine whether or not a stock would appreciate by 10% in a year. This analysis was similar to ours in that it used many similar predictors such as price-to-book and market cap. However, it differs from ours in the model used (Random Forest vs. multiple regression) and the response variable (rate of return vs. yes/no question about stock appreciation).

A third similar study by researchers at Nicholls State and UTD sought to examine the relationship between a fund's past returns and its future returns. This study was similar to ours in that it tried to predict future returns. However, the fact that it only looked at past returns (as opposed to metrics like net assets and investment type) made it quite different from ours.

Methods

To get to the final model, we added information under the investment column for 450 rows with missing data for investment type. We used the following levels: Bond, Currency, Commodity. We also created a new variable that indicated whether or not the ETF was an inverse ETF.

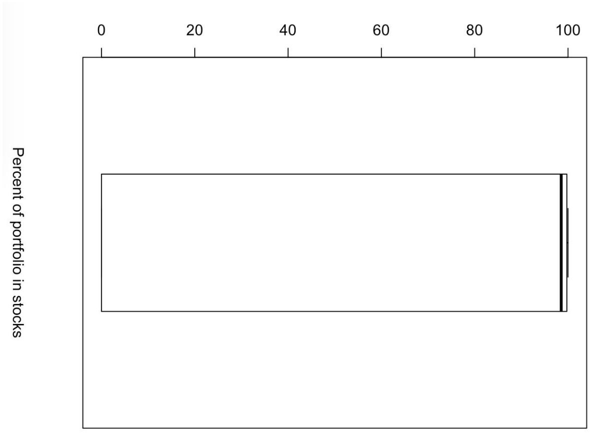
After doing this, we did the following things to clean the data:

1. Remove 30 ETFs which did not fit into any investment category
2. Remove 220 ETFs that had > 80% investment in one economic sector

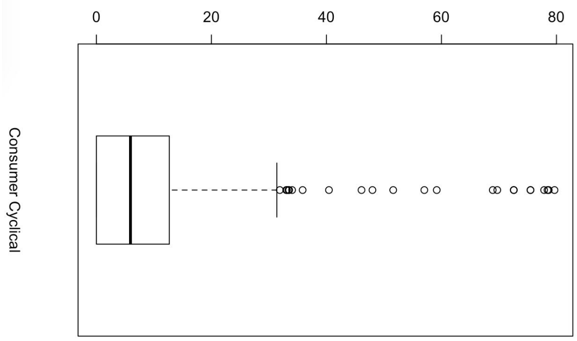
3. For 14 rows that had missing values for one column, use the mean of the column
4. Remove 3 ETFs that had NA for every column

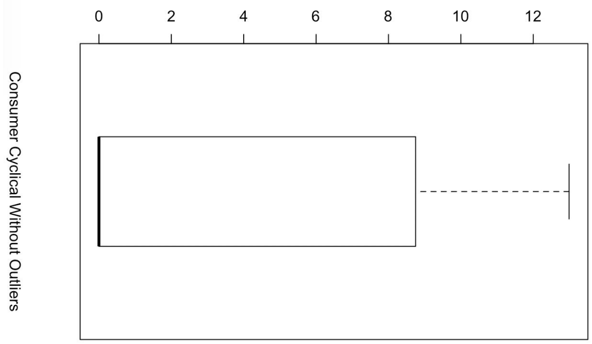
Since the collinearity between Price/Earnings and Price/Book & Price/Earnings and Price/Cash-flow is above 0.8 (considered an indicator of strong multicollinearity), we felt that Price/Book & Price/Cash-flow should be omitted from consideration.

We were able to find outliers in quantitative variables by constructing boxplots. For the percentage of assets invested in stocks, we noticed that there are no outliers.

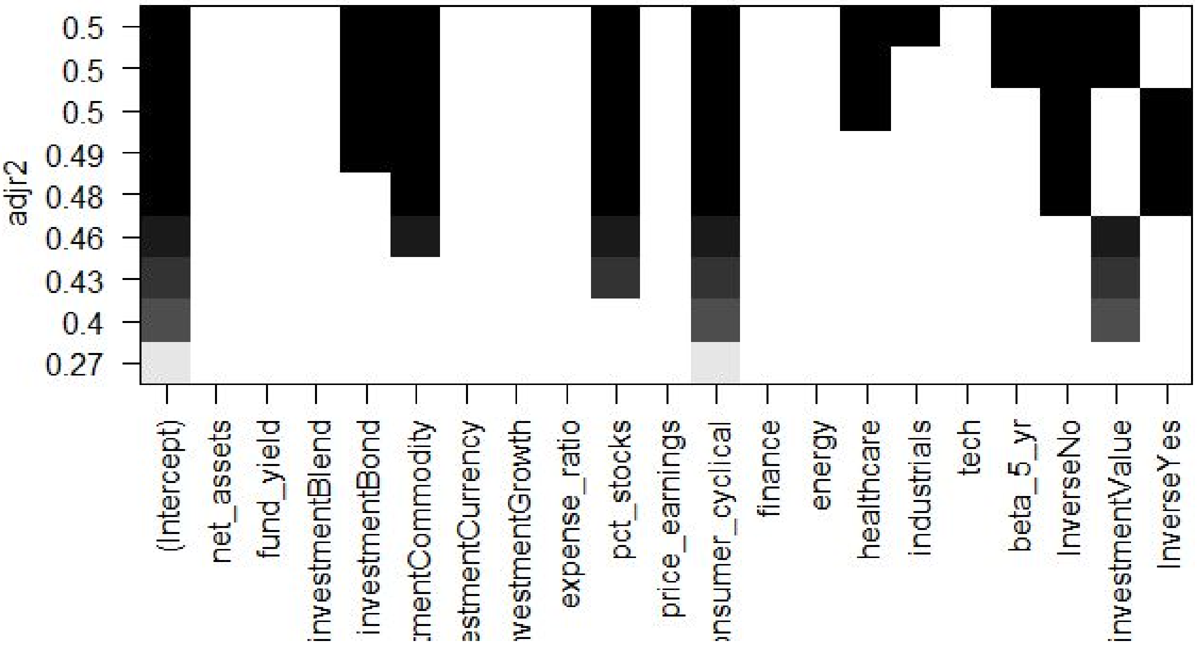


The other quantitative variable is the percent of portfolios in the consumer cyclical industry. In a box plot (left side image below), we noticed that there are many outliers above the third quarter, which heavily distort the median and mean percentages. These outliers can misdirect the users into making the wrong decision, therefore it is necessary to compare and contrast the variable with and without outliers.





We used regsubsets to help us determine which variables would be worth looking at.



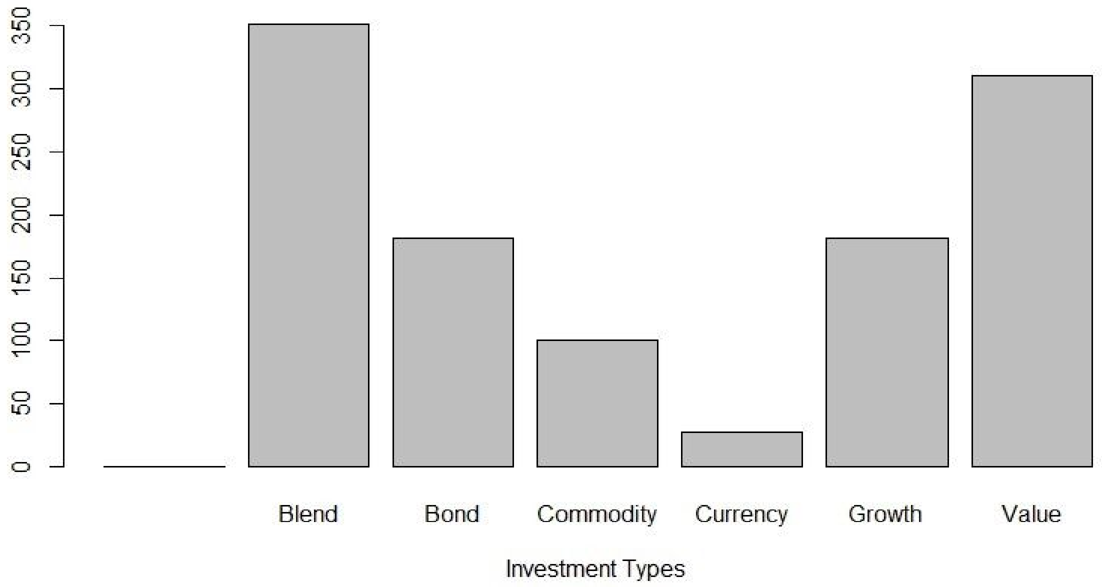
Looking at the output, the 4 variables we chose were investment type, pct_stocks, consumer_cyclical, and Inverse. Our regression model is modeled to predict the 5 year return on ETFs from these four predictor variables, which we can also use to compare ETFs across several industries.

All regression assumptions have been met. Independence is not violated as the 5-year return of one fund is not dependent on another. Looking at the residuals vs. fitted model, we see that neither linearity and homoscedasticity are violated as there is an equal spread along the linear relationship of the plot. Looking at the qqnorm plot, we see the normality assumption is not violated since the relationship is relatively linear and diagonal.

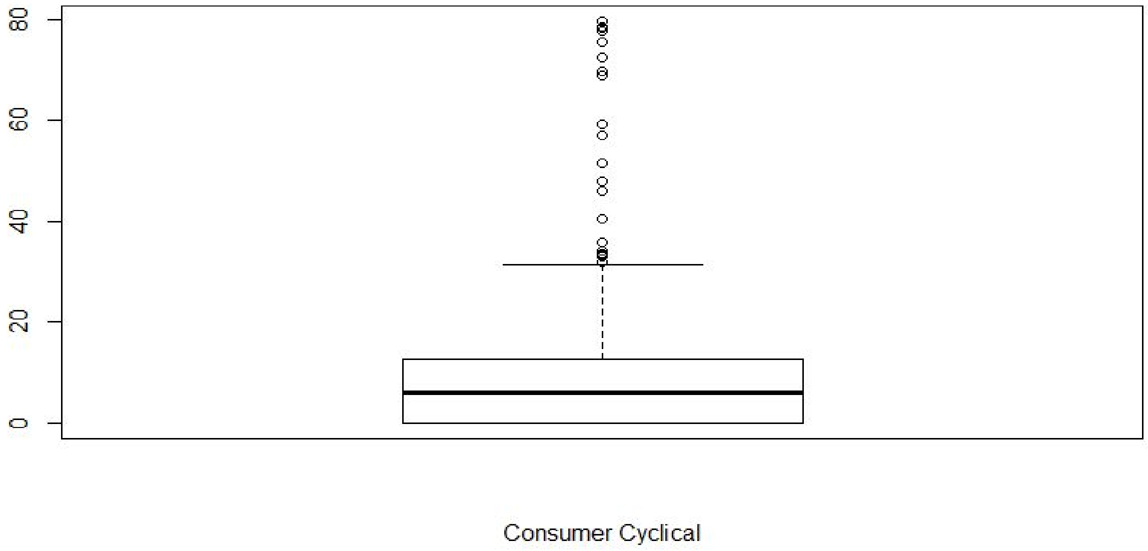
Descriptive Statistics

Predictor variables:

Investment Types

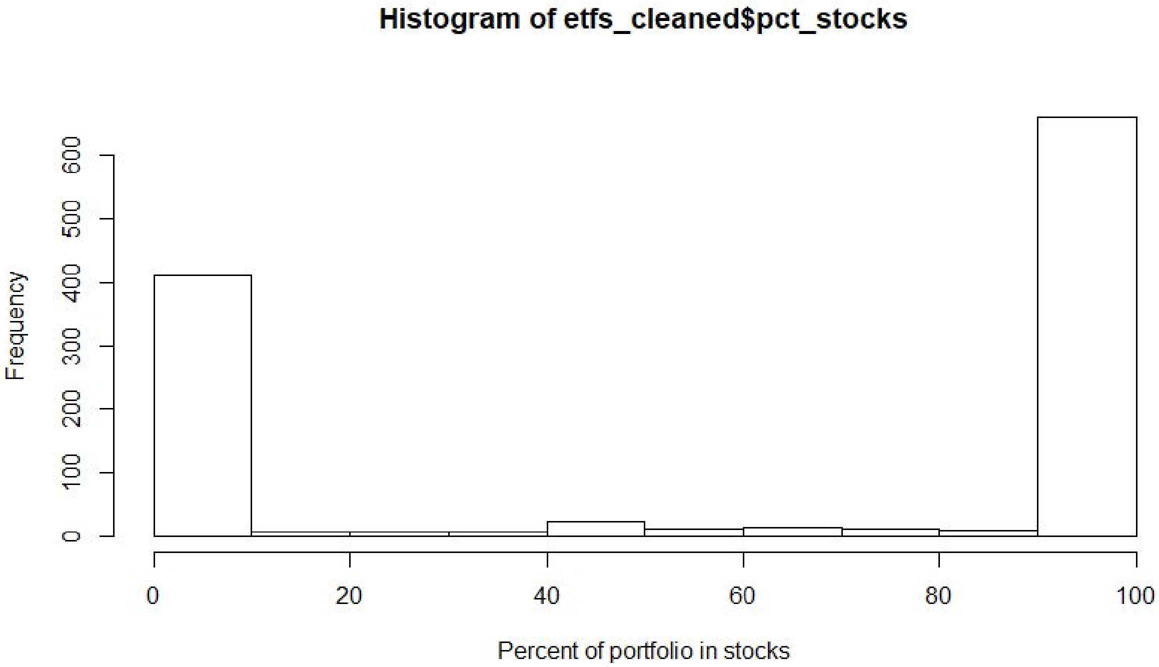


Percent of Portfolio in Consumer Cyclical Industry



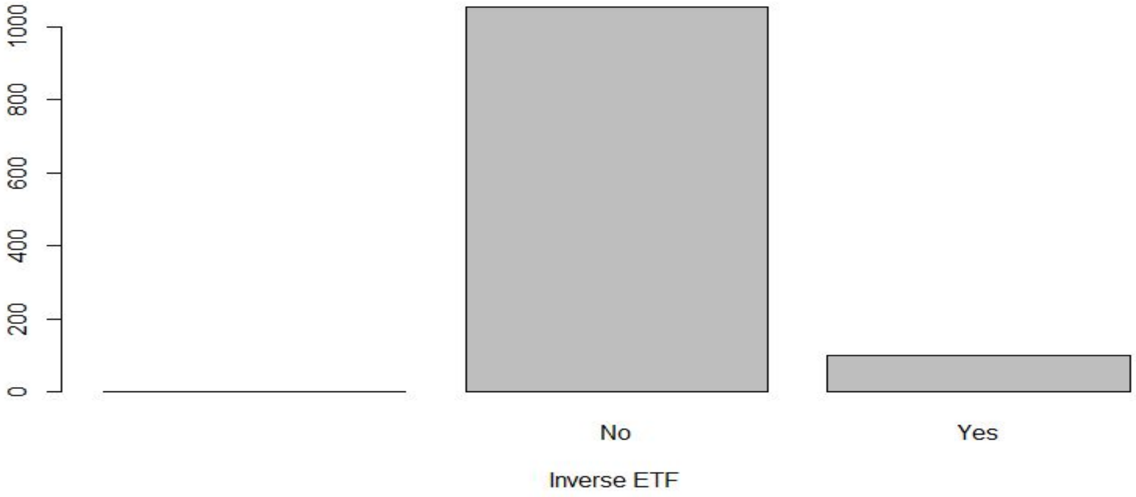
Mean: 7.95%.
SD: 10.35%

Percentage of Portfolio in Stocks



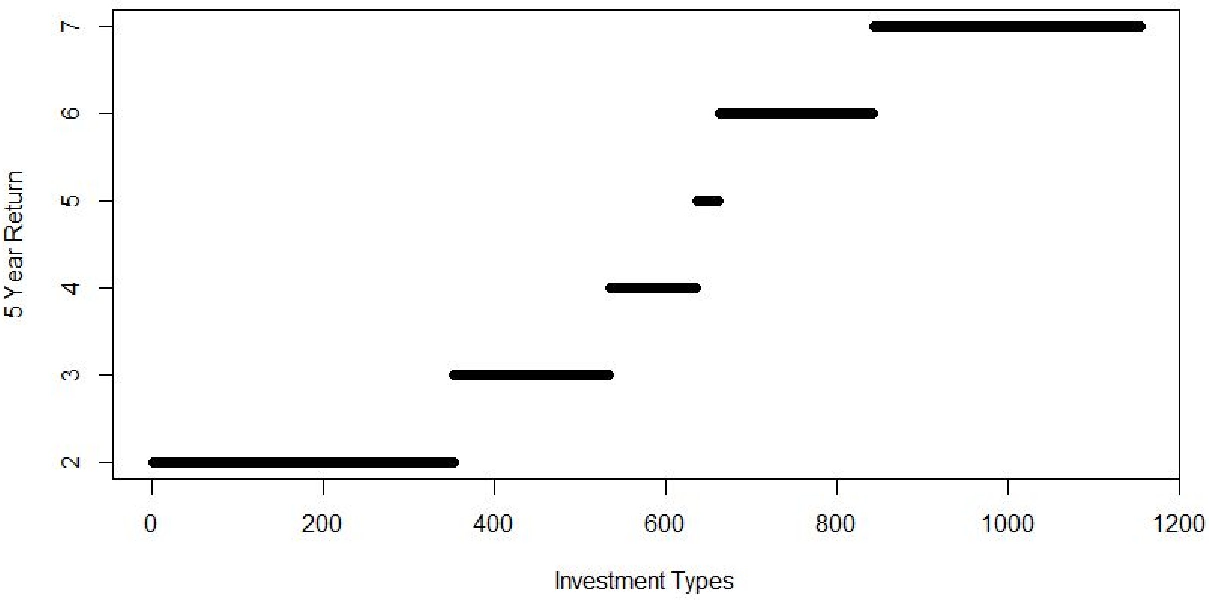
Mean: 60.58%.
SD: 46.88%

Inverse

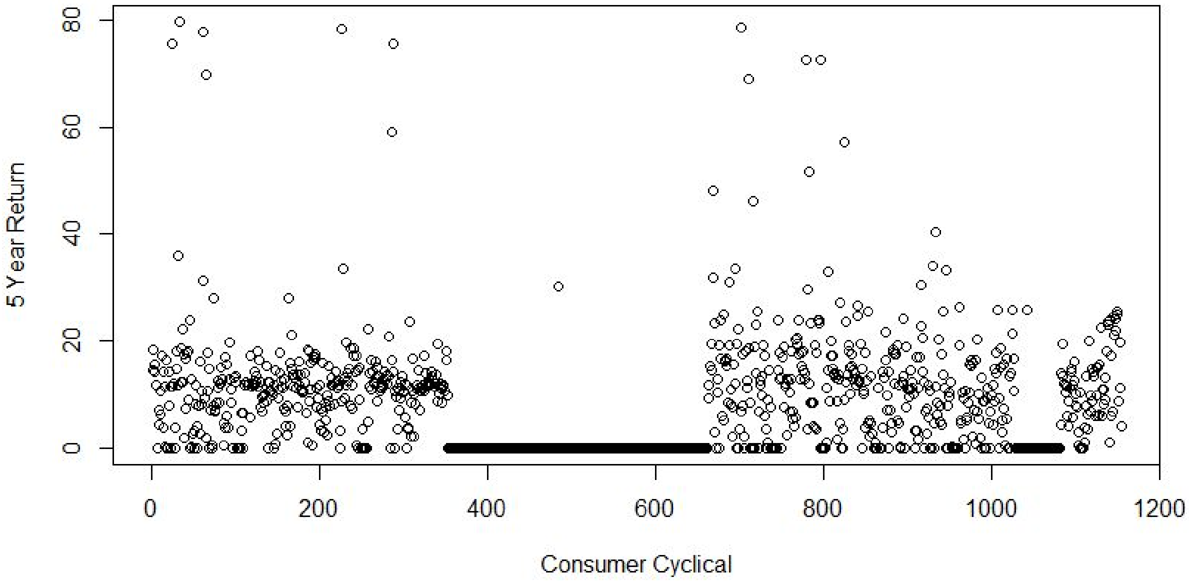


Relationship between response variable and each predictor variable:

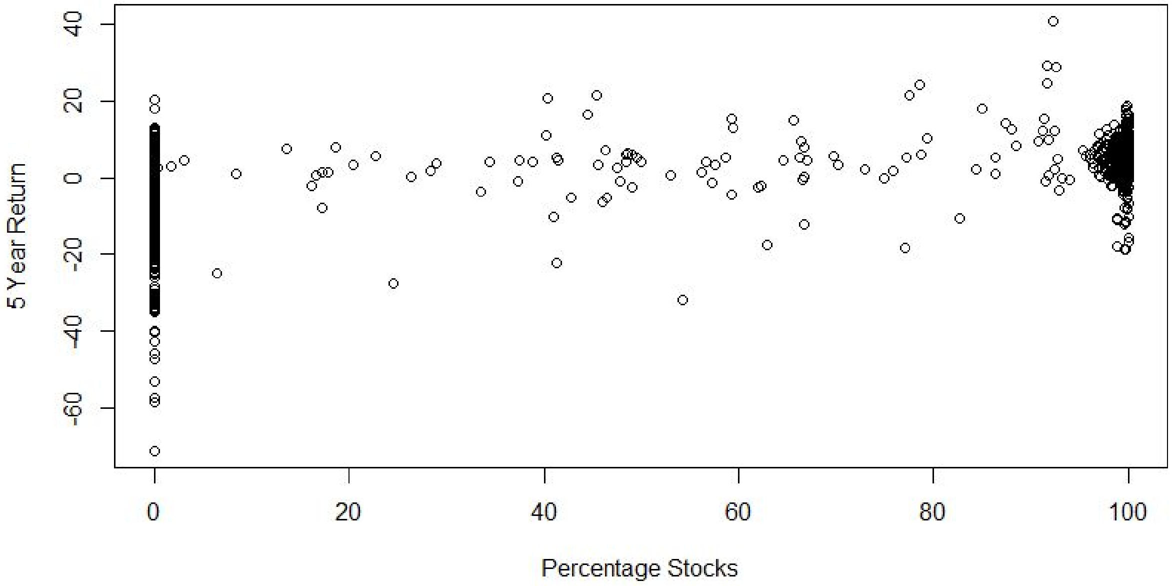
Investment Types



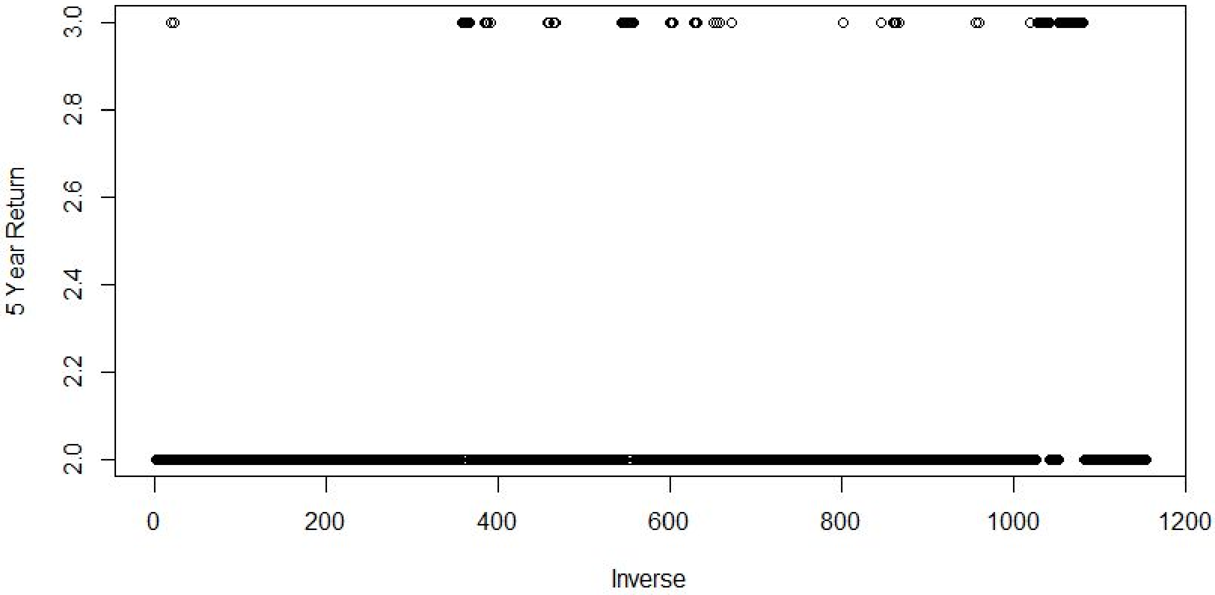
Percent of Portfolio in Consumer Cyclical Industry



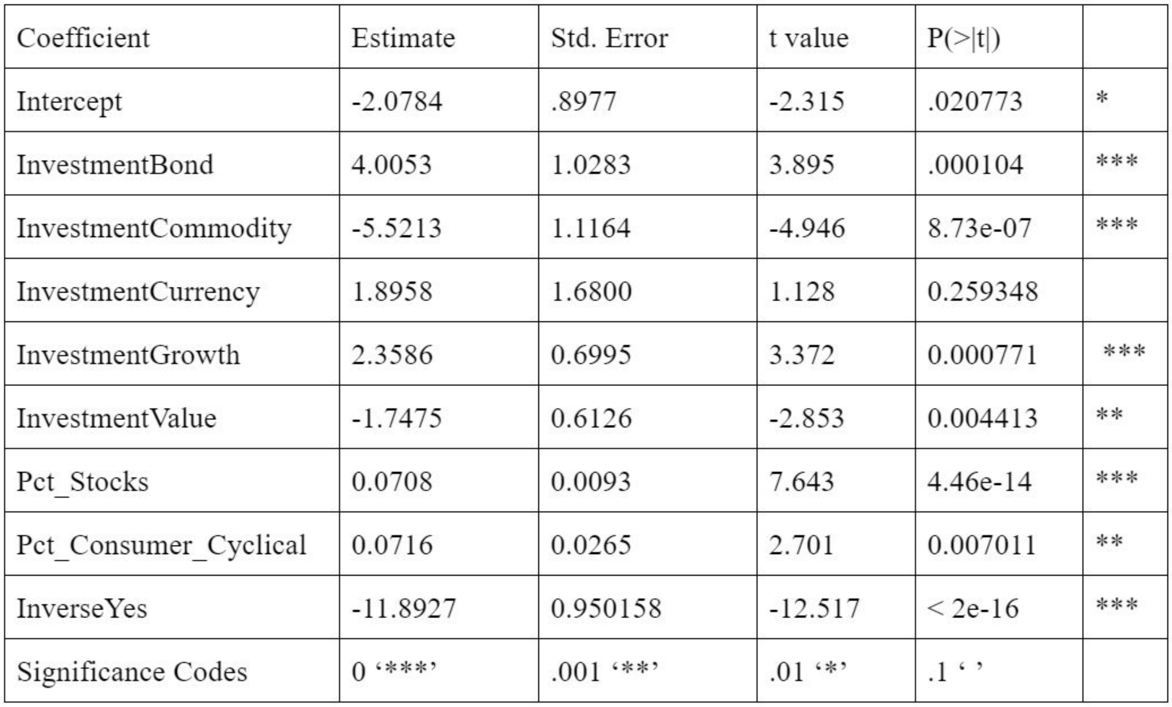
Percentage of Portfolio in Stocks



Inverse



Results



Coefficient	Estimate	Std. Error	t value	P(>|t|)	
Intercept	-2.0784	.8977	-2.315	.020773	*
InvestmentBond	4.0053	1.0283	3.895	.000104	***
InvestmentCommodity	-5.5213	1.1164	-4.946	8.73e-07	***
InvestmentCurrency	1.8958	1.6800	1.128	0.259348	
InvestmentGrowth	2.3586	0.6995	3.372	0.000771	***
InvestmentValue	-1.7475	0.6126	-2.853	0.004413	**
Pct_Stocks	0.0708	0.0093	7.643	4.46e-14	***
Pct_Consumer_Cyclical	0.0716	0.0265	2.701	0.007011	**
InverseYes	-11.8927	0.950158	-12.517	< 2e-16	***
Significance Codes	0 '***'	.001 '**'	.01 '*'	.1 ' '	

$s_e = 7.586$ | Multiple $R^2 = 0.4336$ | Adjusted $R^2 = 0.4297$ | p-value: < 2.2e-16

$$\text{5 yr return} = \hat{Y} = -2.0784 + 4.0053 * InvestmentBond - 5.5213 * InvestmentCommodity + 1.8958 * InvestmentCurrency + 2.3586 * Investment\ Growth - 1.7475 * InvestmentValue + 0.0708 * Pct\_Stocks + 0.0716 * Pct\_Consumer\_Cyclical - 11.8927 * InverseYes$$

Prediction: A non-Inverse, Growth ETF with 8% of its capital invested in the consumer-cyclical industry and 90% invested in stocks will have a 5 year return of 7.228%.

Interpretation

InvestmentBond: 4.0053; Among etfs with the same percent allocation in stocks and the consumer cyclical industry, a Bond investment strategy results in a 4.0053% higher 5 year return compared to ETFs with a Blend strategy.

InvestmentCommodity: -5.5213; Among etfs with the same percent allocation in stocks and the consumer cyclical industry, a Commodity investment strategy results in a 5.5213% lower 5 year return compared to ETFs with a Blend strategy.

InvestmentCurrency: 1.8958; Among etfs with the same percent allocation in stocks and the consumer cyclical industry, a Currency investment strategy results in a 1.8958% higher 5 year return compared to ETFs with a Blend strategy.

InvestmentGrowth: 2.3586; Among etfs with the same percent allocation in stocks and the consumer cyclical industry, a Growth investment strategy results in a 2.3856% higher 5 year return compared to ETFs with a Blend strategy.

InvestmentValue: -1.7475; Among etfs with the same percent allocation in stocks and the consumer cyclical industry, a Value investment strategy results in a 1.7475% lower 5 year return compared to ETFs with a Blend strategy.

Pct_Stocks: 0.0708; Among etfs with the same investment strategy and the same percent of their capital invested in the consumer cyclical industry, each additional percent of an etfs capital invested in stocks results in a 0.0708% increase in 5 year return.

Pct_Consumer_Cyclical: .0716; Among etfs with the same investment strategy and the same percentage of their capital invested in stocks, each additional percent of an etfs capital invested in the consumer cyclical industry results in a 0.0716% increase in 5 year return.

InverseYes: -11.8927; Among etfs with the same investment strategy and the same percentage of their capital invested in stocks and the consumer cyclical industry, Inverse etfs have an 11.8927% lower 5 year return relative to their non-inverse counterparts.

R^2: 0.4336; 43.36% of the variation in 5 year return can be accounted for by the variation in the explanatory variables.

Residual Standard Error: 7.586; on average, the predicted 5 year return will be off by 7.586%

p-value: <2.2e-16; Since the p-value is less than 0.05, we can reject the null hypothesis that there is no relationship between the x variables and 5 year return.

Confidence Intervals:

We are 95% sure that the coefficient is between these values for each predictor variable.

	2.5%	97.5%
Pct_stocks	.053	0.089
Consumer_Cyclical	0.020	0.124
InverseYes	-13.840	-0.317
InvestmentBond	1.988	6.023
InvestmentCommodity	-7.712	-3.331
InvestmentCurrency	-1.400	5.192
InvestmentGrowth	0.986	3.731
InvestmentValue	-2.950	-0.546

Conclusion

In conclusion, the model that we have created in our analysis should be used as a starting point for analyzing the future performance of any ETF. While the R^2 of 0.4336 and the significance of the coefficients seem to indicate that there is some predictive power in our model, it is not high enough to where the model alone should be used to determine whether or not to invest in a certain ETF.

There are also many limitations associated with this dataset. The 5 year returns span 2013-2018, which was during the longest bull run in American stock market history. Therefore, when it comes to investing and making predictions during a bear market or a neutral/flat market (something that we are going through right now), there is no guarantee that the model will be useful for predicting future returns. For this reason, analysts using this model should tread cautiously when economic conditions are not similar to the ones between 2013 and 2018.

In the future, it would be interesting to see similar analyses done on data that includes years where the stock market as a whole went down. That way, analysts would be able to more accurately pinpoint any relationships between the characteristics of an ETF (capital allocation, net assets, investment type/size, etc.) and its returns.