Variable Codebook

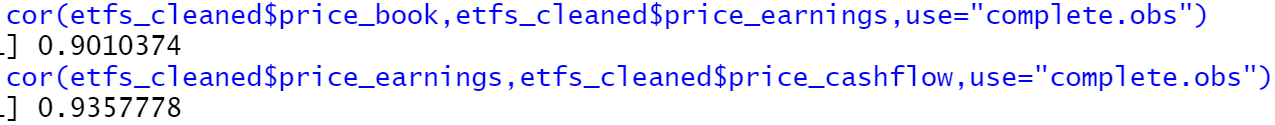
1. Net Assets - This variable represents the market value of all the fund’s financial assets (stocks, bonds, derivatives, and cash). It is quantitative and is measured in dollars ($). However, in our actual analysis, we may transform this variable to “millions of $” or “billions of $” for convenience and more friendly coefficients.
2. Investment type (Value, growth, blend, bond, commodity, currency) - This categorical variable indicates what type of investment strategy the fund uses when determining what stocks/bonds to buy. Value refers to investment in securities that trade at lower multiples and are associated with longevity and stability. Growth means investing in stocks whose prices are projected to rise rapidly due to the technologically advanced nature of the business (high risk, high reward). Blend refers to a mix of the 2 investment strategies. Bond indicates investment in bonds (can refer to any type of bond - municipal, treasury, corporate, sovereign). Commodity refers to investment in futures contracts of a certain commodity (which tend to be physical assets like precious metals, energy resources, and food items). Currency means investing in certain types of currencies, or even betting for or against one currency relative to another currency.
3. Investment Size (Large, medium, small) - This categorical variable refers to how big the fund is. Small usually means that the fund’s net assets are under $20 million. Large usually refers to funds whose net assets number in the billions, if not tens of billions. Medium, as you can guess, indicates somewhere in the middle. This variable is hard to define because its classification is determined by the fund issuer, rather than by a central governing authority.
4. Net annual expense ratio - This variable measures what proportion of a fund’s net assets are recouped by the fund managers every year as compensation or as a maintenance fee. It is given as a percent.
5. Fund Yield - For ETFs, fund yield is a measure of the annual cash flow to the shareholders resulting from bond interest payments, or dividend payments from stocks in the fund’s portfolio. This variable is quantitative and also given as a percent.
6. Percent of portfolio in stocks - As you can guess, this variable gives the percent of the fund’s assets invested in stocks as a whole (as opposed to bonds or derivatives).
7. Percent of portfolio in Financial industry - This quantitative variable gives the percent of the fund’s assets invested in stocks of companies in the finance/financial services industry.
8. Percent of portfolio in Technology industry - This quantitative variable gives the percent of the fund’s assets invested in stocks of companies in the technology industry.
9. Percent of portfolio in Energy industry - This quantitative variable gives the percent of the fund’s assets invested in stocks of companies in the energy (oil & gas, renewables) industry.
10. Percent of portfolio in Industrial industry - This quantitative variable gives the percent of the fund’s assets invested in stocks of companies in the manufacturing, construction, or defense industries.
11. Percent of portfolio in Healthcare industry - This quantitative variable gives the percent of the fund’s assets invested in stocks of companies in the health insurance, pharmaceutical, and biotech industries.
12. Percent of portfolio in Consumer Cyclical industry - This quantitative variable gives the percent of the fund’s assets invested in stocks of companies that make products affected by changes in the economic/business cycle.
13. Price/Earnings ratio - This variable divides the share price of the ETF by the “earnings” per share (usually the earnings of stocks held by the fund). It is quantitative, but does not have any units since it is a ratio of 2 numbers which both have $ as the unit. One can interpret this ratio as the number of dollars that investors are willing to pay (per share) for one dollar of earnings (per share).
14. Price/Cash flow ratio - Similar to the variable above, this one divides the price of one share of the ETF by the operating cash flow (OCF) per share (in this case, OCF refers to net income plus non-cash expenses, minus the change in non-cash assets; this number is calculated using the OCF of the individual stocks in the fund). This variable is quantitative and also does not have any units for the reason mentioned in the previous variable’s explanation. Interpret this variable as the number of dollars that investors are willing to pay (per share) for every dollar of OCF.
15. Price/Book ratio - This variable divides the price of one share of the ETF by the book value of one share of the ETF. Book value refers to the difference between net assets per share and net liabilities per share (more accurately, to the book value of each stock in the fund). Like the 2 preceding variables, this variable is unitless (for the same reason as well) and quantitative, and can be seen as the number of dollars investors are willing to pay (per share) for every dollar of equity (per share).
16. 5 year Fund beta - This variable measures how much the ETF moves relative to the market as a whole. Think of it as a “movement multiplier” that you can use to predict how much the share price went up or down based on the movement of the S&P 500. This quantitative variable is unitless.
17. 5 year Fund R^2 - Like its namesake in normal statistics, this variable measures how much of a security’s movement/performance can be accounted for by the performance/movement of the benchmark it tracks (in this case, the S&P 500). It is unitless and quantitative, just like the R^2 we use in class.
18. 5 year Sharpe ratio - This quantitative, unitless variable measures the fund’s rate of return relative to the rate of return of a “risk free asset” (Treasuries), while adjusting for risk. It’s formula is given by the difference between the fund’s return and the risk-free rate, divided by the volatility (standard deviation) of the fund. A high Sharpe ratio indicates a better investment.
19. 5 year Treynor ratio - Like the Sharpe ratio, the Treynor ratio also measures the fund’s return relative to an asset with zero or near-zero risk (Treasury security). However, it is different in that it divides the excess return by the beta of the fund, rather than the standard deviation of its returns. This difference only takes into account the systemic risk of the fund, rather than the total risk (systemic + non systemic). This variable is quantitative and unitless.
20. Inverse - This categorical variable indicates whether or not the ETF in question is an “inverse” ETF. An inverse ETF is an ETF whose performance is inversely related to the benchmark/commodity/index it is tracking. Theoretically, it will have a perfectly negative correlation with a “normal” ETF that tracks the same underlying asset. It has 2 levels, Yes and No (you can probably figure out what those levels mean). This would be a good variable to consider since we expect inverse ETFs to go up when most other ETFs go down and vice versa; this variable would be able to capture that relationship.

Variables To Consider in Our Analysis

Before we start our preliminary analysis, let’s first define what the purpose of our model should be: the model should be able to take inputs that are known at the moment to predict 5 year returns for the past and the future; this means that we don’t want to use variables that are derived from calculations made in the future using the response variable.

We will be leaving out the following variables out of our preliminary analysis:

1. Investment Size (variable #3)
   1. Reason: The definition for this variable is not consistent and is dependent on the financial institution issuing this security, which could cause misleading results.
2. Price/Cash-flow ratio (variable #14)
   1. Reason: Highly correlated with price-to-earnings ratio variable
3. Price/Book ratio (variable #15)
   1. Reason: Highly correlated with price-to-earnings ratio variable



Since the collinearity between P/E and P/B & P/E and P/CF is above 0.8 (considered an indicator of strong multicollinearity), we feel that the high values necessitate the omission of the aforementioned variables.

1. Sharpe ratio (variable #18)
   1. Reason: The n-year Sharpe ratio is calculated using the return over the last n years. Therefore, we feel it is misleading to include this variable in our analysis since it would be akin to predicting 5 year return from 3 year return.
2. Treynor ratio (variable #19)
   1. Reason: Same as Sharpe ratio. Also, this variable would be highly correlated with the Sharpe ratio, which is something we don’t want.
3. 5 Year Fund R^2 (variable #17)
   1. Reason: This variable is calculated after a time period. Since our analysis seeks to predict future returns as well as past ones, the inclusion of this variable does not seem appropriate since we cannot use a metric that is calculated in the future to predict the future.

Cleaning the Data

The first thing we had to do was go into the csv file and add information under the investment column. Out of the original 1400 (approximate) cases, there were roughly 450 with missing data for investment type. We manually went through and filled in each cell with one of the following levels: Bond, Currency, Commodity (there were a few ETFs which didn’t belong to any of the buckets, so we decided to leave them blank).

The next thing we did was create a new variable that indicated whether or not the ETF was an inverse ETF. The general methodology for doing this goes as follows

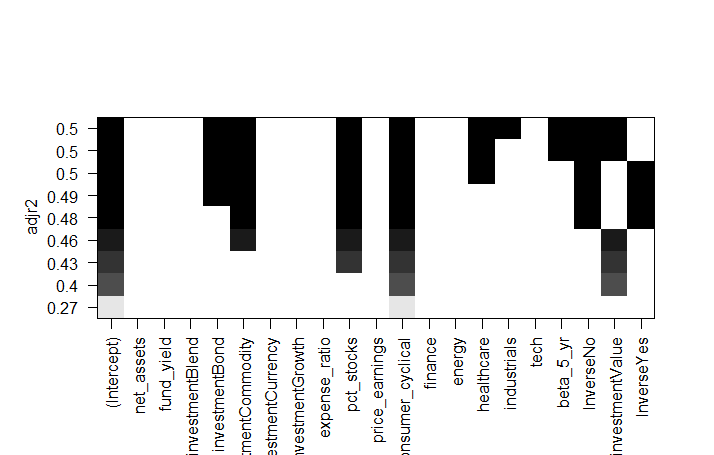
* Inverse = “Yes” if the ETF name has any of the following phrases
  + “Bear”
  + “Short” (but not “Short-term” or “short-duration” or any similar phrases)
  + “Inverse”
  + “-1x”, “-2x”, “-3x”

After doing this, we imported the dataset into R Studio and did the following things to clean the data:

1. Remove ETFs which did not fit into any of the investment categories (~30 rows)
2. Remove ETFs that had more than 80% of their capital invested in a single economic sector (tech, industrials, energy, etc.) (~220 rows)
3. Get rid of extraneous columns
4. For columns that had missing values for one column, replace fill in those fields with the mean of the column (14 rows)
5. Remove ETFs that had NA for every column (3 rows)
6. Rename some columns for convenience

Model Selection

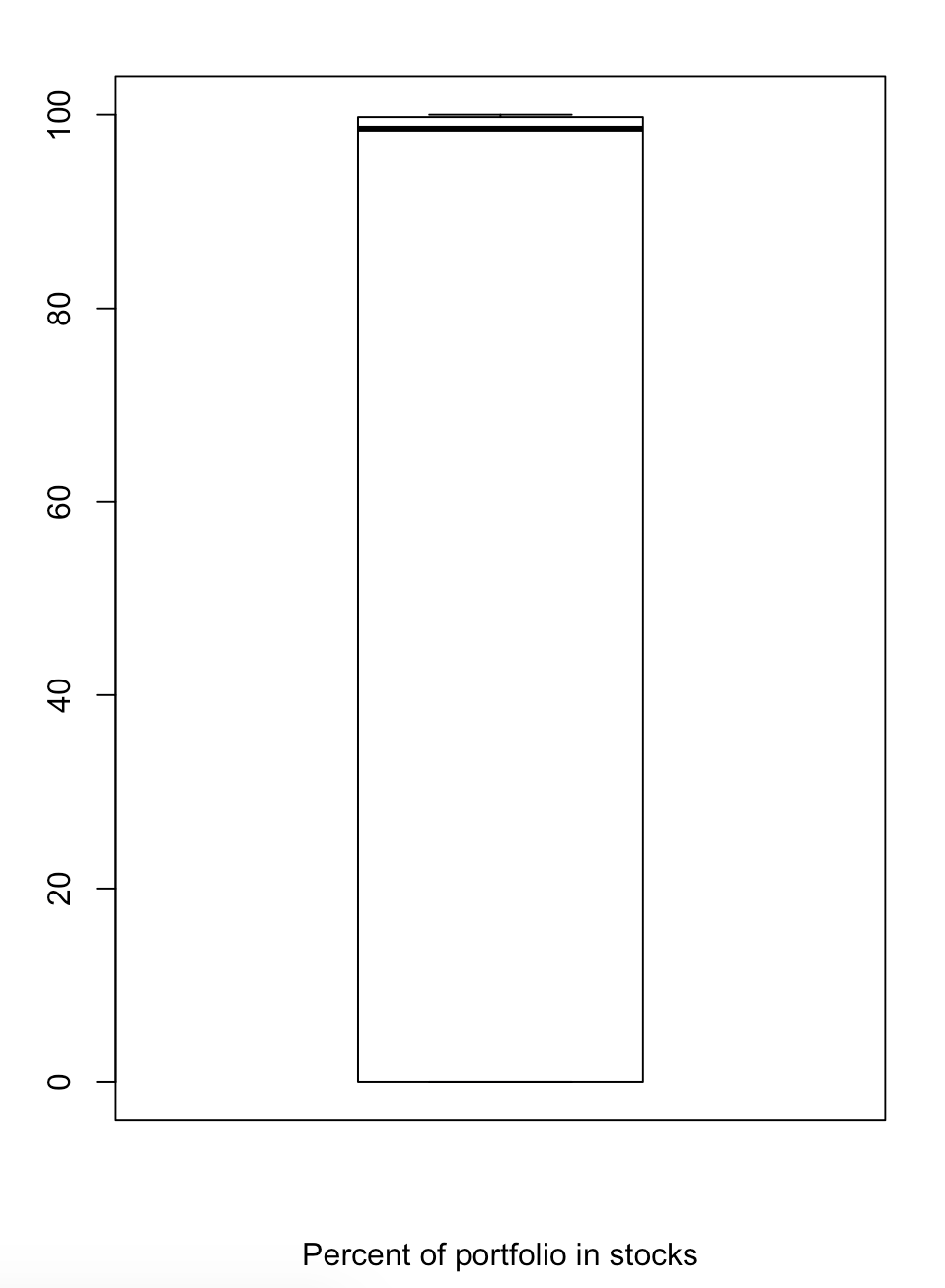
We used the regsubsets function to help us determine which variables would be worth looking at for further analysis. Here is the output of the regsubsets plot:



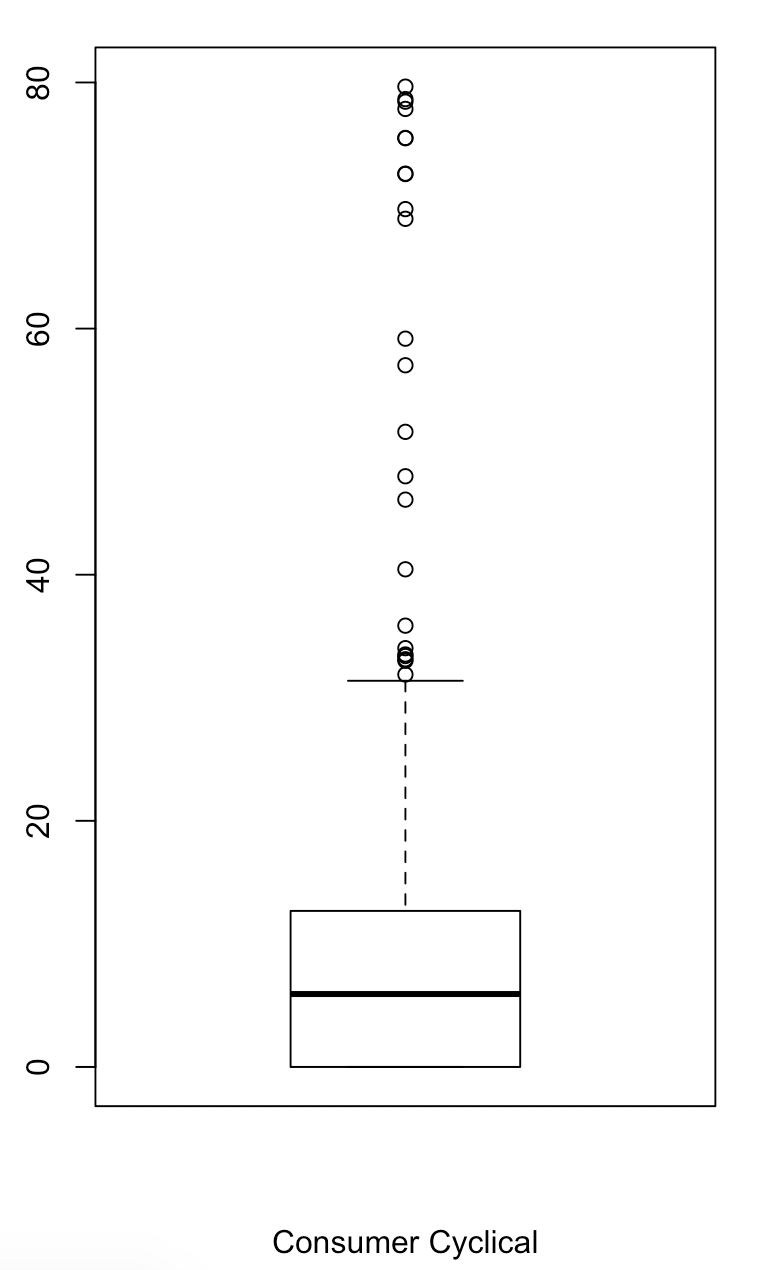
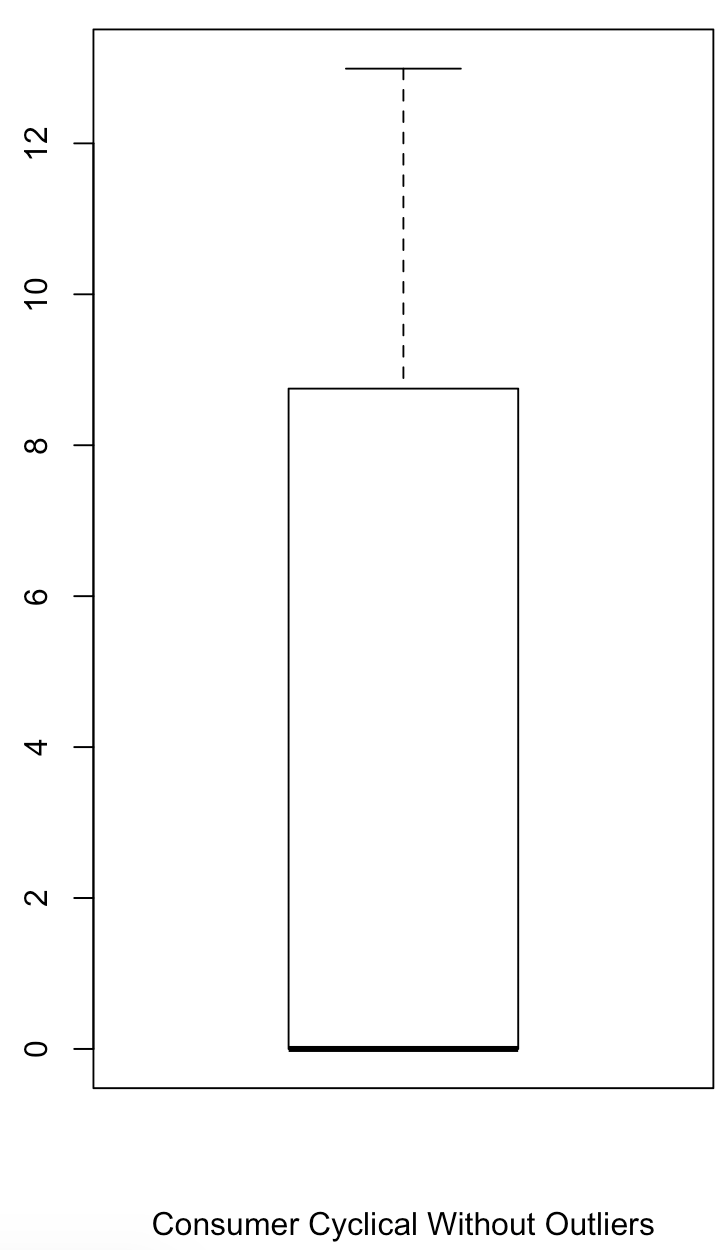
After looking at the output, the 4 variables we have chosen are investment, pct\_stocks, consumer\_cyclical (% of invested capital in the consumer\_cyclical industry), and Inverse. The plot suggests that a model with these 4 variables could have an adjusted R2 as high as 0.49. While we could have eliminated Inverse and still had an adjusted R2 of 0.46, we felt that the extra 0.03 in predictive power was worth the inclusion of Inverse since we also expect inverse ETFs to generally perform worse than their non-inverse counterparts (since our data is from the time period 2013-18 during which there was a bull market).

Missing Data

There are 17 rows with missing data. Out of these, there are 3 rows that have NA for every single column. We have deleted these rows from our dataset. Out of the remaining 14, 12 are missing data for net\_assets while the remaining 2 are missing data for fund\_yield. Since the 14 rows comprise 1% of all rows in the dataset (uncleaned), we feel it is okay to replace the missing value in each row with the mean of the respective column.

Outliers 

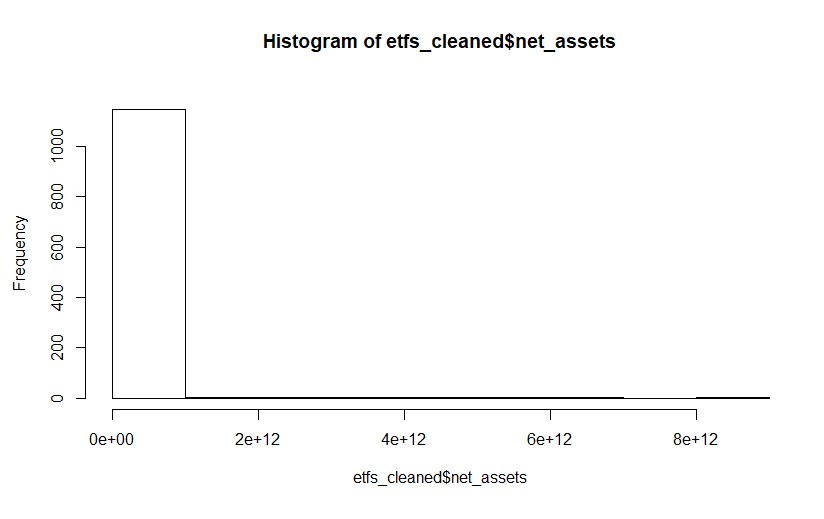
While it is difficult to find outliers in categorical variables, such as investment types and inverse variables, we were able to find outliers in quantitative variables by constructing boxplots. For the percentage of assets invested in stocks, we noticed that there are no outliers. In the summary, the median is 98.55 and the mean is 60.58. While the 1st quarter is 0, the third quarter is 99.75.

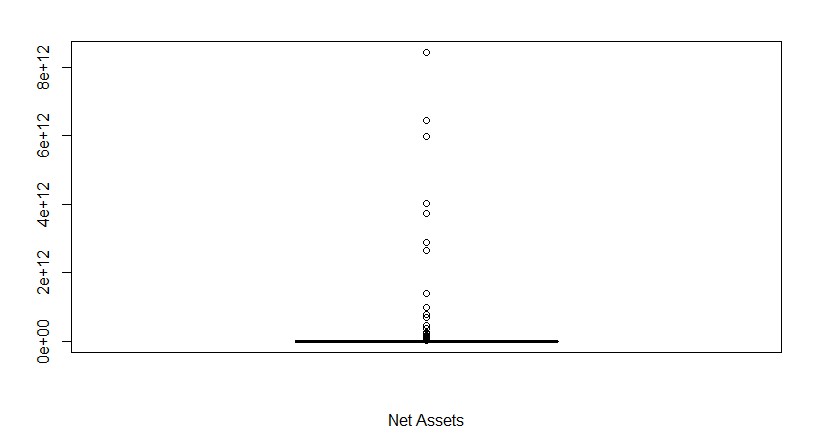
The other quantitative variable is the percent of portfolios in the consumer cyclical industry. In a box plot (left side image), we noticed that there are many outliers above the third quarter. These outliers are heavily impacting the median and mean percentages. For example, the median is 5.9 points greater than the median without the outliers and the mean is 4 points greater than the mean without outliers. 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. 

Graphical and Numerical Summaries

Summaries for the Y variable:

*Net Assets*

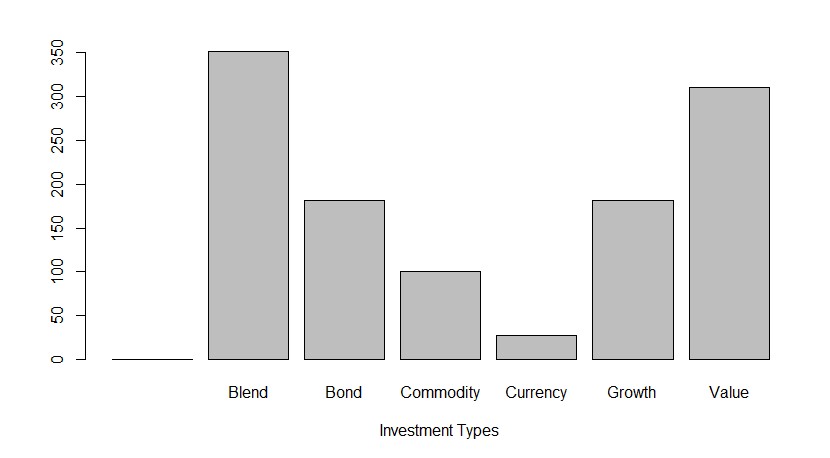
**



As we can see in the above histogram and boxplot, each data point represents one fund’s market value of all its financial assets. There is a vast range of market values ranging from a minimum of $196 thousand to a maximum of $8 trillion, with a mean of $37 billion.

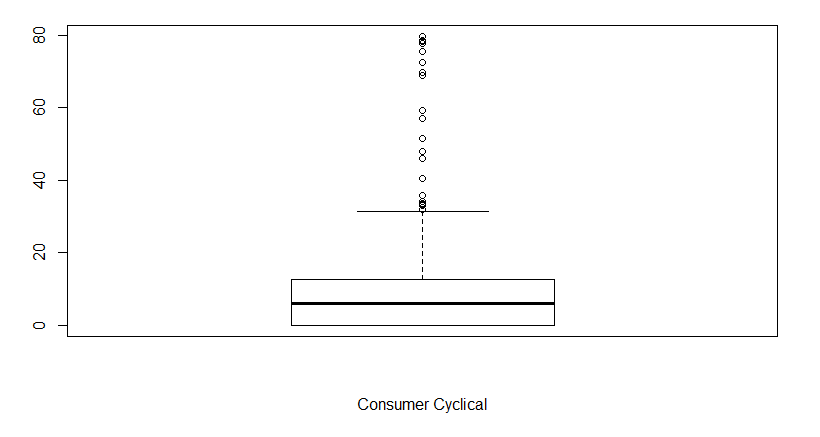
Summaries for X variables:

*Investment Types*



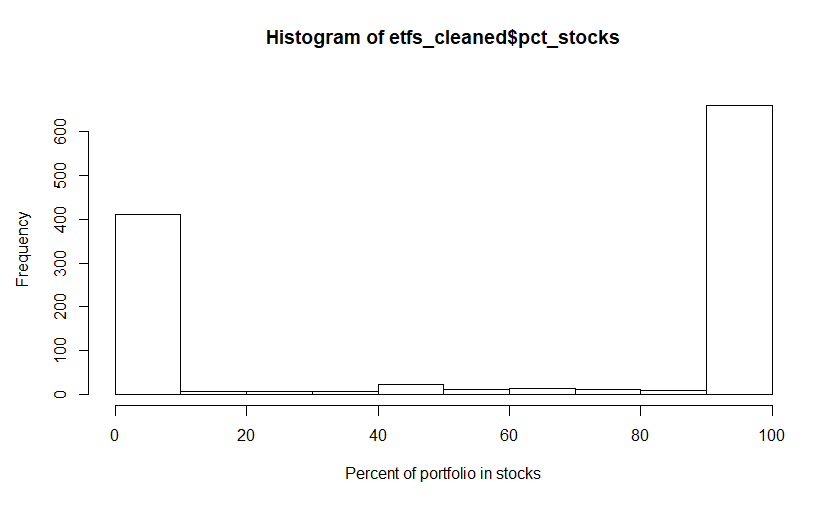
The above bar graph shows us the number of funds that have a specific investment type. Blend and Value were the most common types of investments in our dataset with over 300 ETFs.

*Percent of Portfolio in Consumer Cyclical Industry*



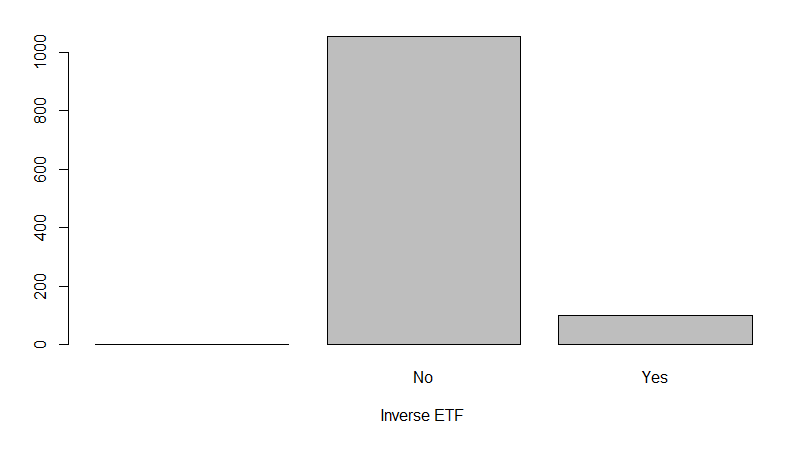
The above boxplot represents the percentage of portfolios that have assets in the consumer cyclical industry. The measure of spread ranges from a minimum of 0 (representing a portfolio with 0% assets in this industry) and a maximum of 79.66%. The mean is 7.95%.

*Percentage of Portfolio in Stocks*



The above bar plot tells us how much of the entire portfolio’s assets are invested in stocks rather than another asset, such as bonds or derivatives. We have two peaks, one smaller peak in the 0-10% range and one larger peak in the 90-100% range. The mean is 60.58%.

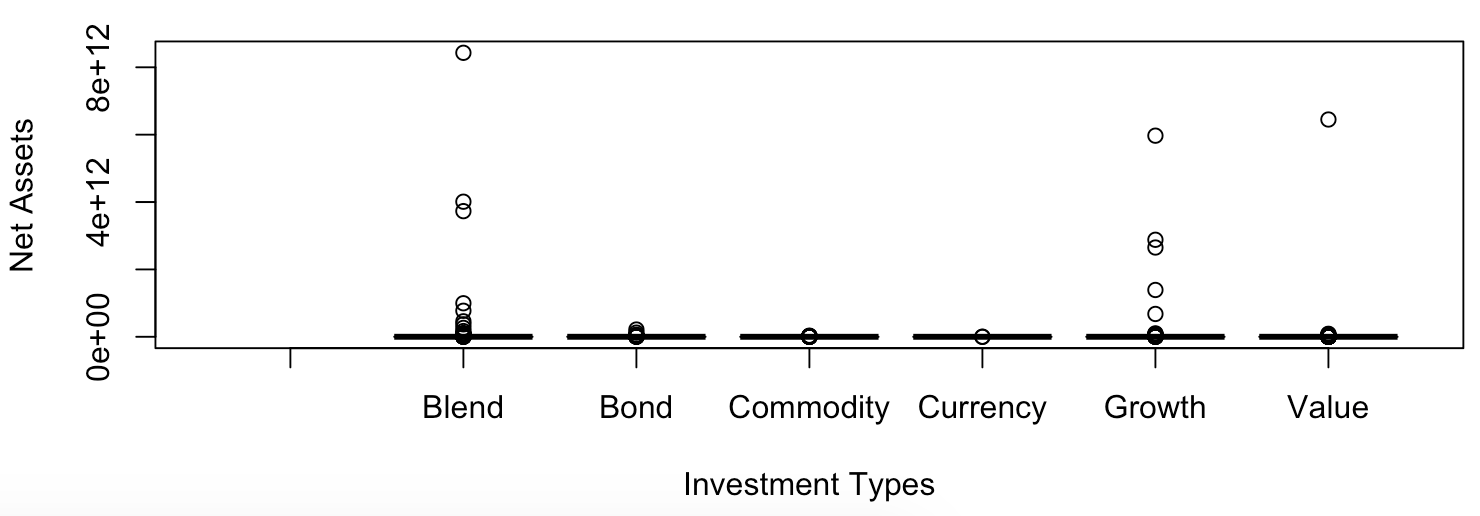
*Inverse*

**

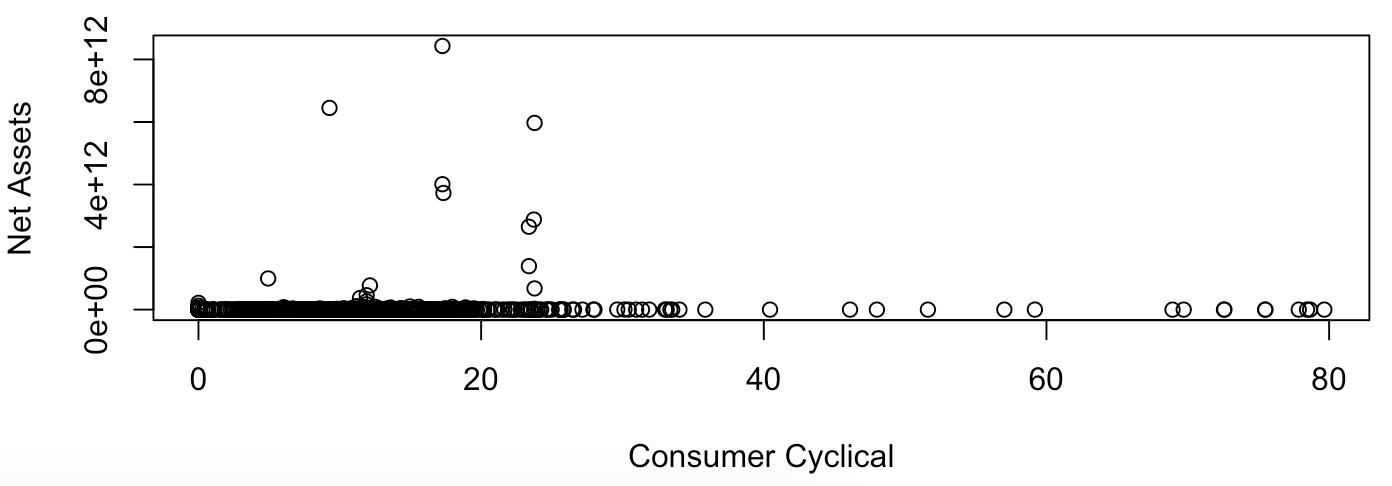
The “inverse” variable is a categorical variable, which indicates whether or not the ETF in question is an “inverse” ETF. As we can see in the above bar graph, most of the ETFs in our data set are not inverse ETFs, meaning they are normal ETFs. It’s about a 1:10 ratio of inverse to normal.

Summaries of the relationship between X and Y:

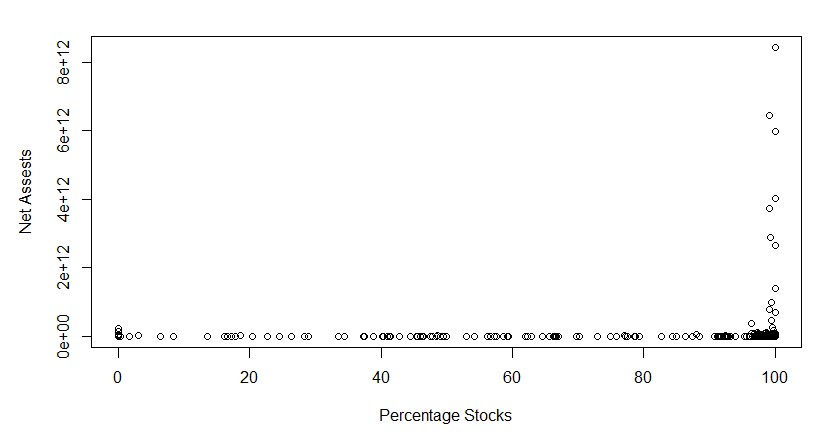
*Investment Types*

When observing the investment variable, we noticed that a blend and growth of investment types tend to generate more net assets compared to other categorical variables that are strictly one investment type. When building a regression to predict net assets with investment types, we observed that the variable is statistically significant, because one of its dummy variable’s (investment bonds) p-value was less than .05. 

*Percent of Portfolio in Consumer Cyclical Industry*

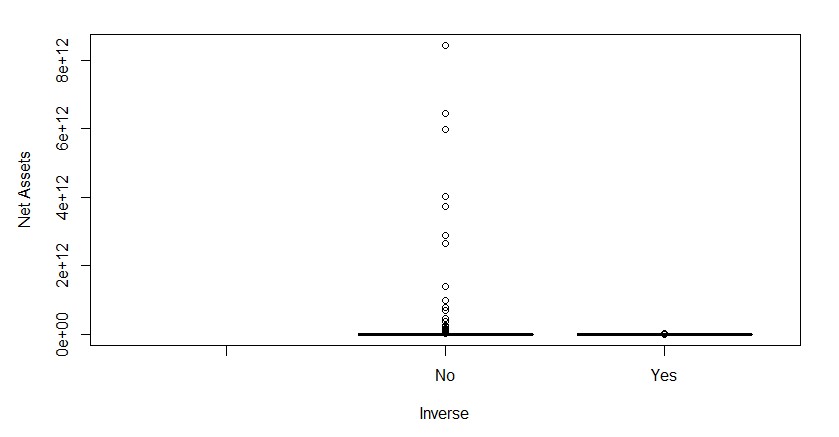
The percentage of portfolios in the Consumer Cyclical Industry was one of the more influential variables in the regsubset analysis. After plotting and generating a regression model to predict net assets with the stocks of companies that make products affected by changes in the economic cycle, we notice there is more movement between 1% to 25% range. When analyzing the summary, we observed that the variable is statistically significant with a p-value less than .05. In terms of practical significance, we notice that the variable has a R^2 of .5% which may be a bit lower than we desire for a more accurate representation. 

*Percentage of Portfolio in Stocks*



The “percentage of portfolios in stocks” variable tells us how much of the entire portfolio’s assets are invested in stocks rather than another asset, such as bonds or derivatives. Through our regression model, we observed that the portfolios whose assets were 90% to 100% invested in stocks generally generated greater net assets than those portfolios that were made up of less than 90% stocks. After running a summary analysis, we observed that the variable is statistically significant with a p-value less than .05. In terms of practical significance, we notice that the variable has a R^2 of .004 which is lower than desired.

*Inverse*



The “inverse” variable is a categorical variable, which indicates whether or not the ETF in question is an “inverse” ETF. An inverse ETF is an ETF whose performance is inversely related to the benchmark/commodity/index it is tracking. This variable captures that relationship between inverse ETFs and ETFs. Through our regression model, we observed that most were not inverse ETFs, meaning they were “normal” ETFs. The normal ETFs generally generated greater net assets than inverse ETFs. When analyzing the summary, we observed that the variable is not statistically significant with a p-value greater than .05. In terms of practical significance, we notice that the variable has a R^2 of .0007 which is lower than we desire for a more accurate representation.