ETFs, the Fama-French Model, and Market Factors - Data Analysis as an Explanatory Tool in Finance

Sebastian Reyes

May 3rd, 2025

Introduction

In the recent political climate, the stock market has been on the minds of many people, financial experts and everyday individuals alike. The massive price swings seen as of late—that is, the high volatility observed—can be mitigated through investments in exchange-traded funds (ETFs), which are funds that pool together shares of multiple companies within a given market index or sector, allowing convenient diversification of an investment portfolio. However, there are many different categories of ETFs, with each subtype carrying its own considerations surrounding projected performance.

These considerations are often highly complex and difficult to understand, but a large portion of them can be explained utilizing the general systematic market risk factors outlined in the Fama-French five-factor model. This project will analyze stock data from five major ETFs (SPY, QQQ, IWM, IWD, and IWF) and examine their behavior in relation to the economic theory behind the five-factor Fama-French model, along with a few additionall major market factors. In other words, this project aims to provide a comprehensive overview of the main considerations surrounding investment in the stock market using straight-forward, intuitive statistical analysis and visualizations—thus displaying the potential of statistics as an educational tool in finance.

Data was scraped from Yahoo Finance, Kenneth's French's (Finance Professor at Dartmouth) online data library, and from the Federal Reserve Bank of St. Louis using Python.

Data Variables

1. **Date** - The month in YYYY-MM-DD format. All days will be the first of the month.
2. **SPY** - the monthly return for the SPDR S&P 500 ETF as a proportion, calculated as $R_t = \frac{P_t - P_{t-1}}{P_{t-1}}$, with P_t representing the share price at the time the market closes on the first day of a given month t. All other ETF monthly return data was calculated in this same manner. This is a continuous variable.
3. **QQQ** - the monthly return for the Invesco QQQ Trust, Series 1 ETF as a proportion. This is a continuous variable.
4. **IWM** - the monthly return for the iShares Russell 2000 ETF as a proportion. This is a continuous variable.
5. **IWD** - the monthly return for the iShares Russell 1000 Value ETF as a proportion. This is a continuous variable.
6. **IWF** - the monthly return for the iShares Russell 1000 Growth ETF as a proportion. This is a continuous variable.
7. **Mkt.Rf** - Market Risk Premium. One of the five factors in the Fama-French model. It is the extra return (as a proportion) which may be expected from exposure to the overall market as opposed to risk-free investments (U.S. Treasury Bills). This is a continuous variable.

8. **SMB** - Small Minus Big: one of the five factors in the Fama-French model. It is the additional return (as a proportion) which can be expected from investments in companies with low market caps as opposed to companies with large market caps. This is a continuous variable.
9. **HML** - High Minus Low. One of the five factors in the Fama-French model. It is the additional return (as a proportion) which can be expected from investment in value stocks (high book-to-market ratio) as opposed to growth stocks (low book-to-market ratio). This is a continuous variable.
10. **RMW** - Robust Minus Weak. One of the five factors in the Fama-French model. It is the additional return (as a proportion) which can be expected from investment in companies with high profitability margins, as opposed to those with low profitability margins. This is a continuous variable.
11. **CMA** - Conservative Minus Aggressive. One of the five factors in the Fama-French model. It is the additional return (as a proportion) which can be expected from investment in companies which reinvest conservatively themselves as opposed to companies which reinvest aggressively. This is a continuous variable.
12. **RF** - Risk-Free Rate. The monthly return rate (as a proportion) which can be expected from a risk-free investment in U.S. Treasury Bills.
13. **VIX** - A categorical variable regarding VIX–a measure of market volatility. High VIX is defined as being 10% or more above the 6-month moving average, while low VIX is defined as being 10% or more below the 6-month moving average. Levels are `High`, `Low`, and `Neutral`.
14. **Bull_Bear -** A categorical variable describing whether a given month (S&P 500 acts as a proxy) was in a bull market, bear market, or neither. A bear market was defined as a month at least 10% below a recent peak, while a bull market was defined as a month having recovered from the most recent bear market. This is not completely accurate to the ± 20% definition used in finance, but it was a useful proxy considering sample size and technical limitations. Levels are `Bull`, `Bear`, and `Neutral`. In classic economics, there is admittedly no `Neutral` market, but this level was introduced for clearer differentiation.

Data Cleaning

Since I was in charge of the format of the data when importing it using Python, most of the data was fairly clean. The main issue I faced was that the ETF returns (`returns`) dataset and the Fama-French factors/market factors (`factors`) dataset each came from a different source, and they were thus formatted slightly differently. Datapoints regarding VIX and bull/bear markets were scraped from the Reserve Bank of St. Louis and calculated by hand, respectively. They were then appended onto the `factors` dataset.

To begin, I had to merge the two datasets. However, due to their differing sources, each dataset's date range differed. For instance, `factors` had information dating from January 1993 to December 2024, while `returns` had information dating from January 1993 to April 2025. Thus, when merging these two datasets, I had to ensure that I was doing so using the command `merge(returns, factors, by = "Date")`.

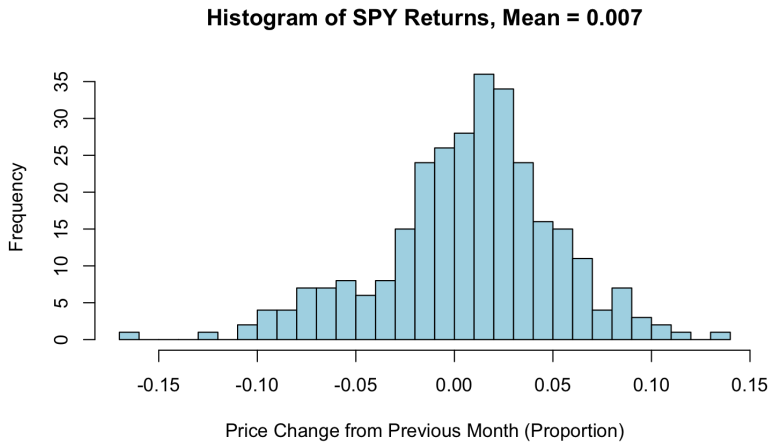
In relation to this, the **Date** variable was originally a character vector. I fixed this utilizing the `ymd()` function in the lubridate package. In this way, dates could be treated as continuous variables.

Additionally, there was one more consideration purely surrounding `returns`. More specifically, I had to import the information for each ETF individually. This became an issue due to different funds having varying date ranges for their available information. For instance, **SPY** had data available dating back to January 1993, while **IWF** only began displaying data in June of 2000. As a result, the months which did not have data for all 5 ETFs were removed via the command `na.omit()`. This gave me one final dataset which contained information from June 2000 to December 2024, which I simply named `df`.

Initial Descriptive Plots and Summary

Understanding ETF Monthly Returns

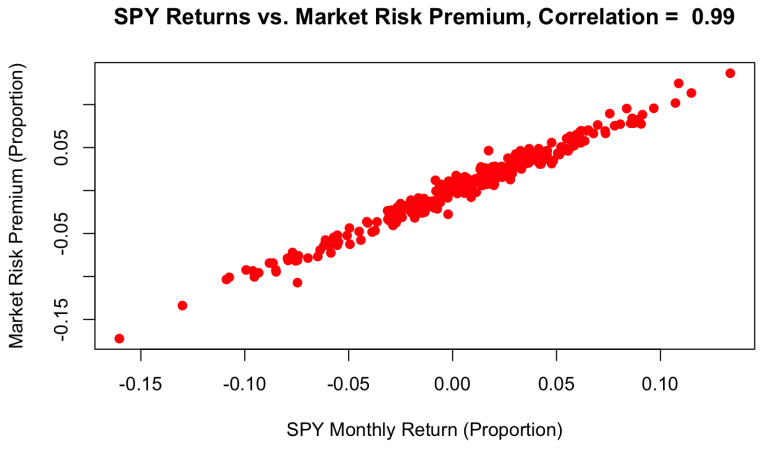
To understand the structure of the data better, it is helpful to generate some initial exploratory plots. For the sake of simplicity, constraining initial visualizations to using solely the **SPY** ETF can help explain the other variables, as **SPY** is often used as a proxy for the market as a whole.



The above histogram displays the monthly return values observed throughout this dataset. The mean of 0.007, when multiplied by 12, returns a value of 0.084. This indicates an annual return of 8.4%, which is approximately average for this time period.

Understanding Market Risk Premium (Mkt.Rf)

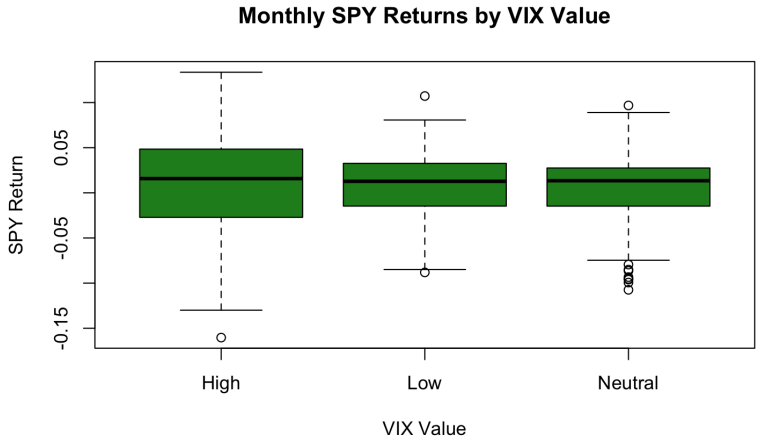
To explain the **Mkt.Rf** variable, it is helpful to use a scatterplot.



Market Risk Premium (**Mkt.Rf**) by definition, is the additional return which can be expected by bearing the risk associated with investment in the market, as opposed to putting one's capital in risk-free investments such as U.S. Treasury Bills. Since **SPY** is typically used as the main proxy for the market as a whole, it is no surprise that an extremely strong correlation can be observed market returns and **SPY** returns—the latter acts as an anchor for the former.

Understanding Volatility (VIX)

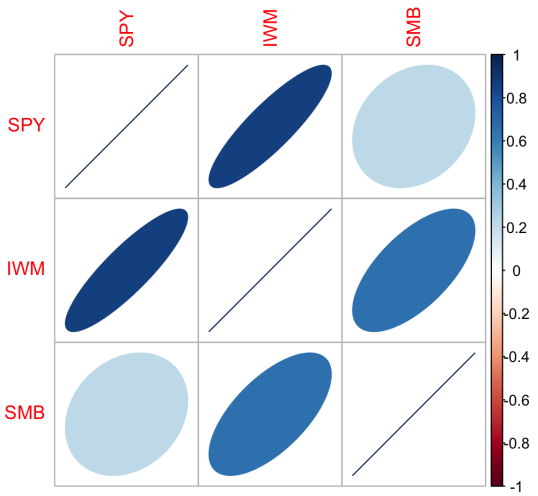
To better understand the **VIX** variable in the dataset, a box plot can be drawn comparing **SPY** returns across the different levels of **VIX**.



Volatility is, by definition, a measurement of the degree of price fluctuations in the market. **VIX** quantifies this volatility, with a higher **VIX** value indicating larger price swings. As a result, it is no surprise that the spread of returns in the above box plot is observed to be much higher when the **VIX** metric is large.

Understanding Small Minus Big (SMB)

"Small Minus Big", by definition, is a measure of the additional return which can be expected from investment in companies with small market caps ($< \$2$ billion) as opposed to large market caps ($> \$10$ billion). This is an important market factor, as since small-cap companies generally entail a greater investment risk, shareholders demand greater returns. **SPY** contains a mix of small-cap and large-cap companies, with a bias towards the latter, while **IWD** focuses on small-cap companies.

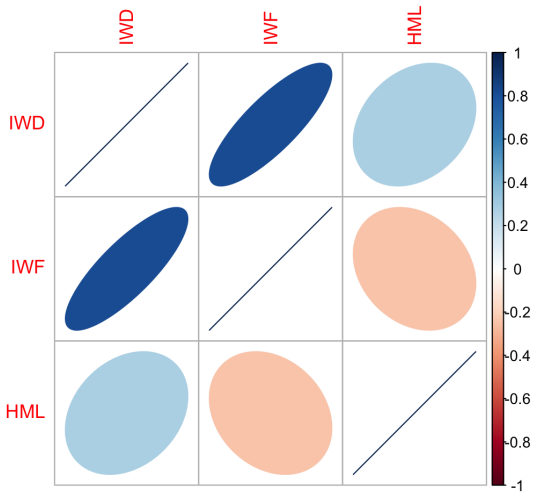


As observed, **IWM** has a much stronger correlation with the **SMB** market factor than **SPY** does. Once more, this is because **SMB** quantifies the bonus returns from small-cap companies in comparison to large-cap companies, and **IWM** is a small-cap fund, while **SPY** contains broad a mix of large-cap and small-cap companies. However, **SPY** is also mildly positively correlated with **SMB**, as small-cap performance remains

tied to the market as a whole, which is proxied by **SPY.** It is for this same reason that it is unsurprising to see returns from **IWM** and **SPY** are positively correlated directly with one another.

Understanding High Minus Low (HML)

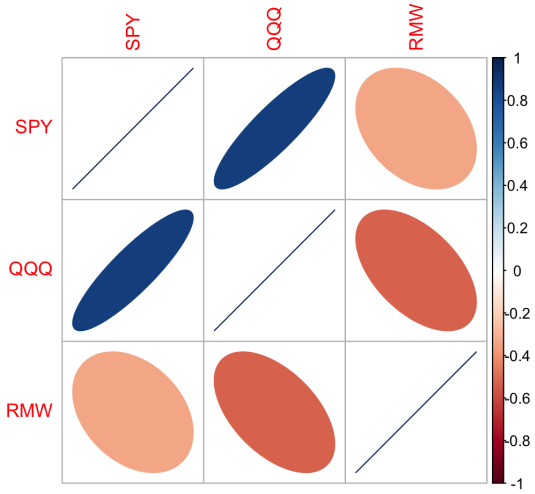
High Minus Low (**HML**) is a measure of the additional returns which can be expected from investment in stocks with a high book-to-market ratio (value stocks) as opposed to low book-to-market ratio (growth stocks). The two primary ETFs which will model this market factor well are **IWD** and **IWF**, the large-cap value and growth Russell 1000 funds, respectively.



As expected, the value fund **IWD** is positively correlated with the **HML** market factor. It is of note that as opposed to the previous example comparing **IWM**, **SPY**, and **SMB**, the "counterexample" fund in this case—**IWF**–is actually *negatively* correlated with this market factor of interest. This is because **IWD** and **IWF** are polar opposites, focusing solely on value or growth funds, respectively. They are on opposite sides of the spectrum. This was not the case for the previous example of **SPY** vs. **IWM**. Instead, **IWM** acted as a fund on an extreme end of the small-cap/large-cap investment spectrum, while **SPY** acted as a reference point closer to the middle.

Understanding Robust Minus Weak (RMW)

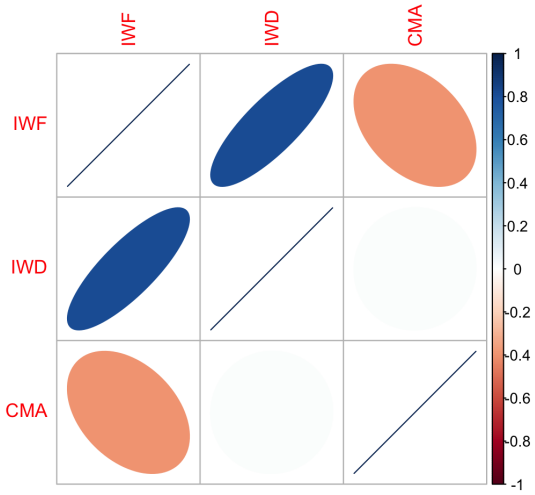
Robust Minus Weak (**RMW**) is a measure of the additional returns which can be expected from investment in companies with high profitability margins, as opposed to those with low profitability margins. **QQQ** is an excellent ETF to demonstrate this, as **QQQ** is comprised largely of growth and technology companies, which reinvest heavily, leading to small profit margins in comparison to their book equity.



As observed above, **QQQ** is affected more heavily by the **RMW** factor than the overall market (as proxied by **SPY**). There exists a strong negative correlation, as **QQQ**'s investment in companies with comparatively weak profit margins results in lagging performance when robust profit companies do well. **SPY** is also heavily biased towards the technology sector (which typically reports comparatively weaker profits), but its diversification results in a negative correlation of smaller magnitude.

Understanding Conservative Minus Aggressive (CMA)

Conservative Minus Aggressive (**CMA**) is a measure of the additional return which can be expected from investment in companies which reinvest conservatively in themselves as opposed to companies which reinvest aggressively. The **IWD** and **IWF** ETFs prove useful in this regard once more, as value-tilt companies (**IWD**) typically reinvest more conservatively, while growth-tilt companies (**IWF**) typically reinvest more aggressively.



As observed above, there is a negative correlation between **CMA** and **IWF**. This is to be expected, as **IWF** is an ETF focused largely on growth-tilt companies, and **CMA** is a measure of the degree to which investments in companies which conservatively reinvest (value-tilt) outperform investments in those who aggressively reinvest (growth-tilt). As such, when the **CMA** factor is high, **IWF** returns tend to lag behind the returns of investments in value-tilt companies (**IWD**).

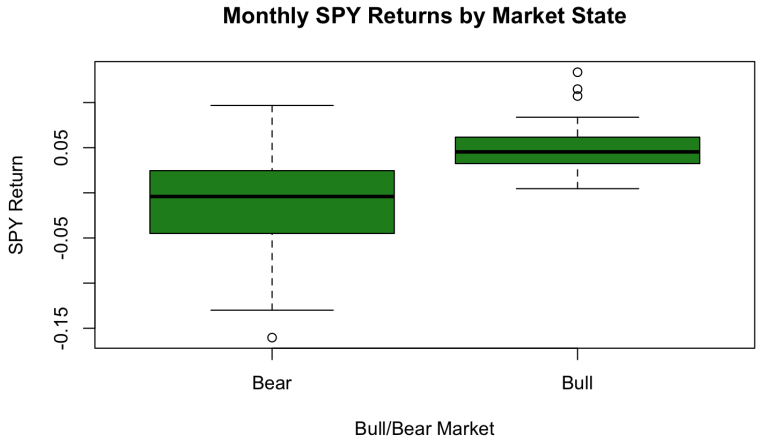
It is of note that there is a correlation of nearly zero between **IWD** returns and the **CMA** factor. This is largely due to the different weighting of stocks between **IWF** and **IWD**. That is, **IWF** contains a greater

mix of both conservative and aggressive investors than **IWD**. Despite each ETF having approximately equal weighting in their respective growth-tilt and value-tilt, this characteristic is only *associated* with investor aggression—there are exceptions, and it is not synonymous.

It is also unsurprising that **IWF** and **IWD** returns are positively correlated with one another, as their performances remain tied to the same general market. Their differing relationship with the **CMA** factor is indicative only of smaller distinctions in their investment strategies, not of opposing overall market trends.

Understanding Bull and Bear Markets (Bull_Bear)

Bull and bear markets are descriptors used to represent sustained upward or downward price movements in the market, respectively. As a way of better understanding these patterns, a box plot can be of use.

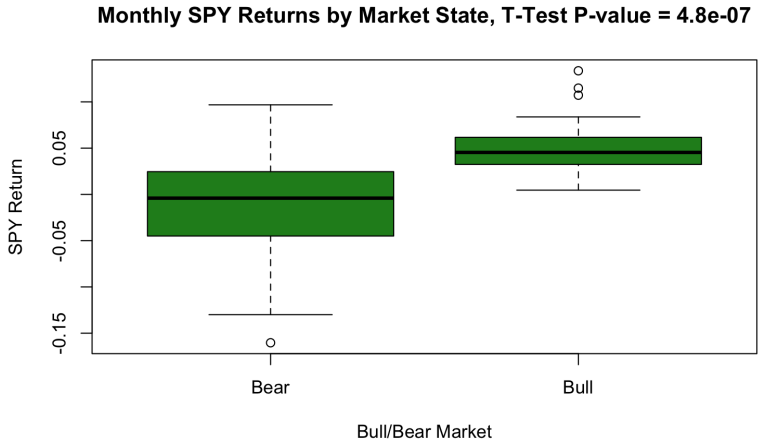


As expected, there is a large difference in market returns depending on the state of the market. A bull market is associated with rising stock prices, while a bear market is associated with sharp economic decline and stock price drops. As such, **SPY** returns are generally higher during bull markets.

Analysis

Two-Sample T-Test

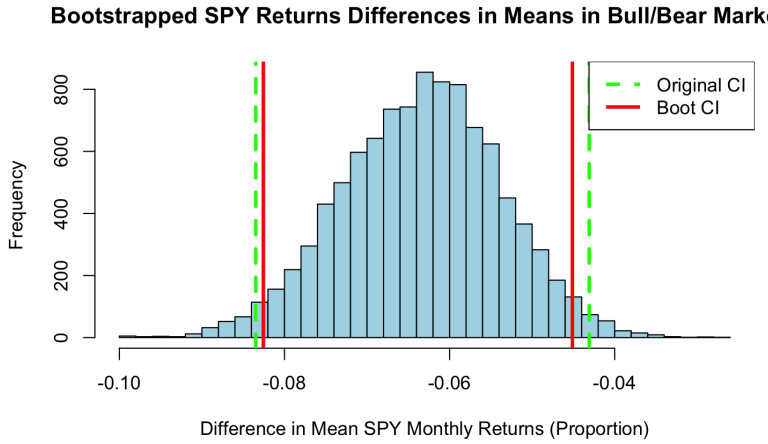
Considering the large difference previously observed between **SPY** returns depending on whether the current month is in a bull or bear market, a two-sample t-test could prove useful. For the sake of computation, the sample was subsetted to only include months within the `Bull` and `Bear` levels (i.e. not `Neutral`).



There appears to be a significant difference of means in **SPY** monthly returns depending on the state of the market. The two-sample t-test confirms this, outputting a p-value of $4.8 * 10^{-7}$. The confidence interval output is (-0.083, -0.043). This confidence interval does not include zero, and thus is not surprising in context.

Two-Sample Bootstrapped Confidence Interval

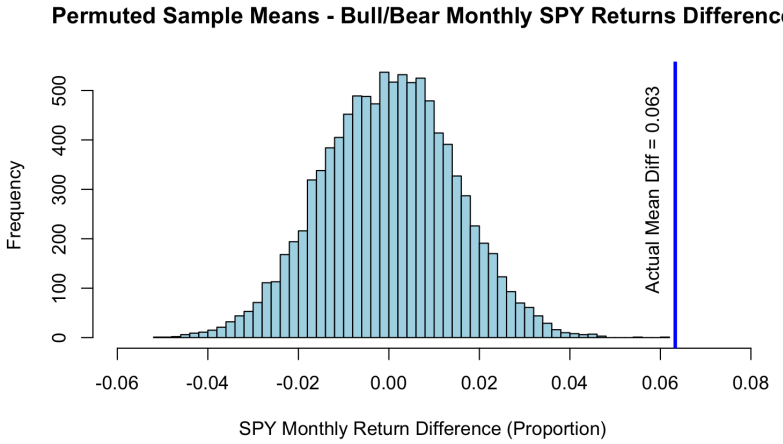
However, this Bull market sample size is rather small. As a result of this, it could be beneficial to perform a two-sample confidence interval bootstrap in order to confirm these results.



As observed above, despite the small sample size of this months assigned to the bull market category, the original two-sample t-test worked well. The bootstrapped confidence interval for the difference in means is very similar—albeit slightly narrower—and zero does not form part of it. This means that, as expected, there is strong statistical evidence of a difference in mean SPY monthly returns when comparing bull markets and bear markets.

Permutation Test

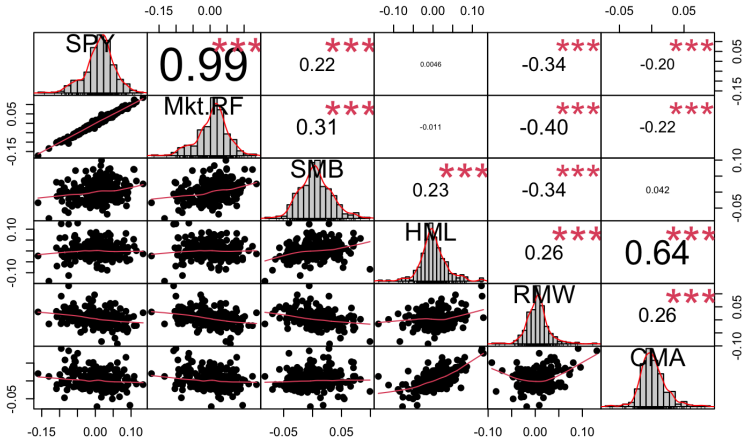
To further strengthen the robustness of statistical tests, it could prove useful to utilize a permutation test as well.



As seen above, the original sample difference in means is a massive outlier. Using the command `diffvals >= 0.63`, the returned p-value is 0. As such, there is extremely strong evidence of a difference in mean monthly **SPY** returns in bull vs. bear markets.

Multiple Regression

Using multiple regression, the five-factor model can be investigated more in-depth. First, when choosing the dependent variable, **SPY** appears to be a safe choice. One important consideration, however, is that all of the factors in the Fama-French model are themselves excess returns over the risk-free rate—**RF**. As such, in order to be consistent in how the values in the model are calculated, the dependent variable must become the *excess* return of the S&P 500 **(SPY-RF).** This said, the dependent variables of this model then become the five factors of the Fama-French model. That is, **Mkt.RF, SMB, HML, RMW,** and **CMA**.



As observed above, **SPY** appears to have a linear relationship with all predictors. There also appear to be no issues of heteroscedasticity. However, there are issues of multicollinearity (**HML/CMA**) which need to be kept in mind moving forward.

This established, backwards stepwise regression can begin. The model `lm((SPY-RF) ~ Mkt.RF + SMB + HML + RMW + CMA, data = df)` is fit. Summary statistics indicate an adjusted R^2 value of 0.9862, along with a non-significant p-value for the **CMA** predictor. This could be due to multiple reasons. Firstly, **SPY** as an ETF includes a broad mix of companies with conservative and aggressive investors, decreasing the impact of the **CMA** predictor. Additionally, the previously seen multicollinearity may have had an effect.

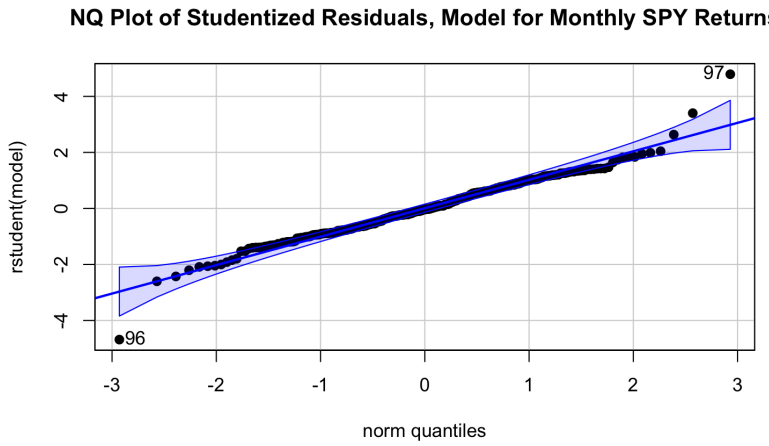
Following this, the newly fit model becomes `lm((SPY-RF) ~ Mkt.RF + SMB + HML + RMW, data = df)`. All predictors are now significant at $\alpha = 0.001$. The results of this final model are displayed below.

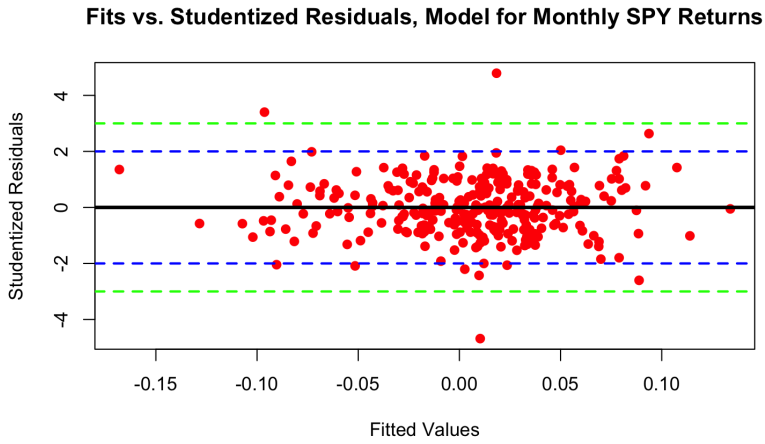
```
Call:
lm(formula = (SPY - RF) ~ Mkt.RF + SMB + HML + RMW, data = df)

Residuals:
       Min         1Q     Median         3Q        Max
-0.0235417 -0.0035530 -0.0001764  0.0036098  0.0239109

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.0003743  0.0003212  -1.165 0.244807
Mkt.RF       1.0068954  0.0074752 134.698  < 2e-16 ***
SMB         -0.1562268  0.0124965 -12.502  < 2e-16 ***
HML          0.0384772  0.0099607   3.863 0.000138 ***
RMW          0.0485598  0.0145459   3.338 0.000953 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.005261 on 290 degrees of freedom
Multiple R-squared:  0.9863,    Adjusted R-squared:  0.9862
F-statistic:  5237 on 4 and 290 DF,  p-value: < 2.2e-16
```

The final observed R^2 statistic is 0.9863. This is not surprising, as since **SPY** acts as a proxy for the market as a whole, it should be expected that the the factors of the Fama-French model—which is specifically designed to model the market—will be able to predict its excess returns well. As for the individual predictors, they are outlined below.

1. Market Risk Premium (**Mkt.RF**) - $\beta = 1.007$. This corresponds to a positive change in excess **SPY** returns of 1.007 for every single unit change in **Mkt.RF**. This is unsurprising, as by definition, the market risk premium is a measure of the excess return from exposure to the market, why **SPY** acts as a proxy for. That is, **Mkt.RF** and **SPY-RF** are measuring the exact same metric. However, the predictor had to be included in the model nonetheless, as it is it a pillar of the Fama-French model.
2. Small Minus Big (**SMB**) - β = -0.156. This corresponds to a negative change in excess **SPY** returns of 0.156 for every single unit change in **SMB**. This indicates that **SPY** as an ETF leans away from small-cap companies, which is perfectly accurate considering that **SPY** is mostly comprised of large-cap stocks.
3. High Minus Low (**HML**) - $\beta = 0.0385$. This corresponds to a positive change of 0.0385 in excess **SPY** returns for every singular unit change in **HML**. This indicates that **SPY** leans slightly towards value stocks as opposed to growth stocks, which is accurate considering that **SPY** is comprised of the largest corporations in the U.S., and that value stocks are typically associated with established, stable companies.
4. Robust Minus Weak (**RMW**) - $\beta = 0.0486$. This corresponds to a positive change of 0.0486 in excess **SPY** returns for every singular unit change in **RMW**. This indicates that **SPY** leans slightly towards companies with large profit margins as opposed to lower profit margins, which is accurate considering that lower profit margins are typically associated with growing companies which are aggressively expanding, while larger profit margins are typically associated with larger, well-established companies (such as those which **SPY** tracks).

Finally, checking assumptions surrounding heteroscedasticity and normally distributed errors, the plots below are generated.



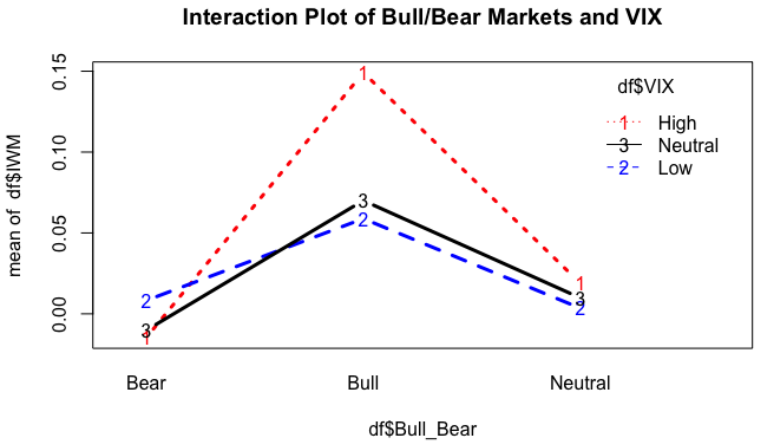


Thus, the errors are found to be normally distributed and without evidence of heteroscedasticity. There is also no evidence of outliers or influential points. The multiple regression is successful.

ANOVA

Seeing as the two categorical variables in this dataset revolve around bear/bull markets (**Bull_Bear**) and volatility (**VIX**), when conducting a two-way analysis of variance test, it be ideal to pick an ETF which is most likely to be most impacted by these factors, as it will result in a cleaner test. **IWM** immediately stands out, as its tilt towards small-cap stocks makes it more likely to be impacted by the market factors described by **Bull_Bear** and **VIX**.

First, it should be checked whether there is potential for an interaction.



The generated interaction plot displays non-parallel lines, indicating a potential interaction. As such, the ANOVA model should include an interaction term, so as to check its significance. That is, the model fit should be `aov(IWM ~ Bull_Bear + VIX + Bull_Bear*VIX, data = df)`. The output of this model is displayed below.

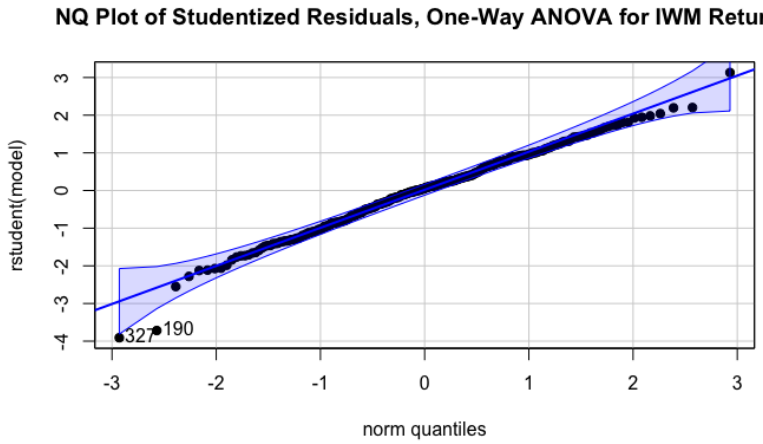
```
               Df Sum Sq Mean Sq F value   Pr(>F)
Bull_Bear       2 0.0885 0.04423  14.244 1.27e-06 ***
VIX             2 0.0014 0.00072   0.231   0.7940
Bull_Bear:VIX   4 0.0286 0.00714   2.299   0.0591 .
Residuals     286 0.8881 0.00311
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

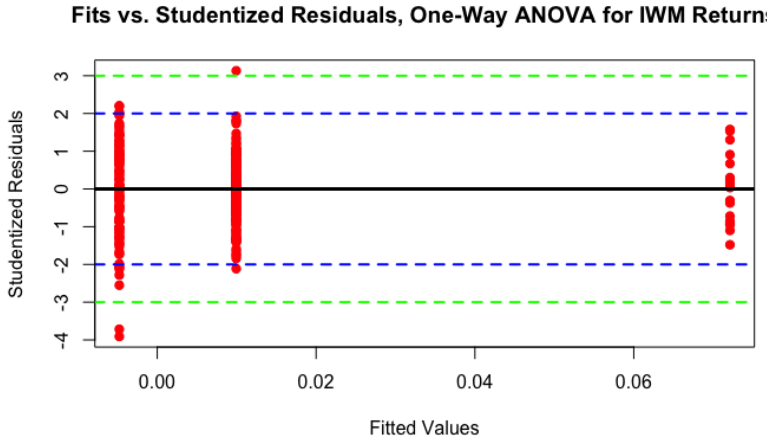
As observed, neither the **Bull_Bear**/**VIX** interaction nor **VIX** itself is found to be significant. As such, this model then becomes a standard one-way ANOVA. That is, this model becomes `aov(IWM ~ Bull_Bear, data = df)`. The output of this model is shown below.

```
Df Sum Sq Mean Sq F value   Pr(>F)
Bull_Bear     2 0.0885 0.04423   14.07 1.47e-06 ***
Residuals   292 0.9181 0.00314
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

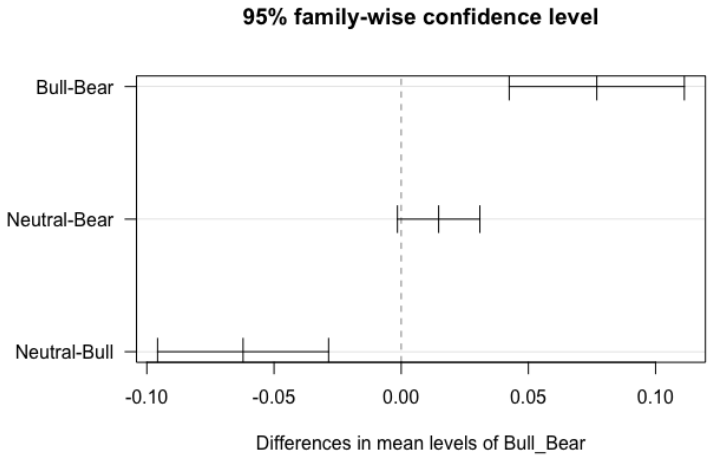
The sample standard deviation of each **Bull_Bear** group is also calculated via the function `by(df$IWM, df$Bull_Bear, sd)`, and no major difference between standard deviations is found, indicating roughly equal variance.

As such, **Bull_Bear** is observed to be the only significant categorical factor. This established, the residuals of the model are examined next.





As observed above, the residuals of this ANOVA model are normally distributed and, for the most part, homoscedastic. As such, this model is successful. It is found that at between the categories of **Bull_Bear**, (`Bull`, `Bear`, `Neutral`), at least one group has a statistically significant mean in **IWM** returns. Tukey comparisons can be used to further extend this claim.



As displayed above, Tukey's HSD displays significant differences in mean **IWM** monthly returns between

the `Bull/Bear` and `Bull/Neutral` levels. There is also very nearly evidence of a difference in means between the `Neutral/Bear` levels, but zero is just barely within the confidence interval of differences in means, so we cannot say there is evidence.

As a final confirmation, it could prove fruitful to also confirm results using Welch's Anova, as it could be argued that there was a slight degree of heteroscedasticity visible in the previous residual plot. The output of this is displayed below.

```
One-way analysis of means (not assuming equal variances)

data:  IWM and Bull_Bear
F = 14.438, num df = 2.000, denom df = 43.265, p-value = 1.573e-05
```

As visible, Welch's ANOVA still displays a highly significant difference in mean **IWM** returns based on **Bull_Bear**. As such, there is evidence that the results from the previously conducted ANOVA can in fact be trusted.

In context, this is to be expected. **IWM** is a small-cap fund, which makes it much more susceptible to large swings in the overall market (i.e. the state of the market as either a `Bull` or `Bear` market), and as such, it is no surprise that its mean returns are confirmed to be influenced by the **Bull_Bear** factor in this analysis.

Conclusions and Summary

Through an analysis of five ETFs covering a majority of U.S. equities, an explanation of the five Fama-French factors (with the help of histograms, scatterplots, box plots, and correlation plots), and an explanation of how bull and bear markets interact with volatility indexes in tangent with these factors (using two-sample t-tests, bootstrapping, permutation tests, multiple regression, and ANOVA), My hope is that this project provided a thorough and comprehensive overview of basic stock market investment considerations.

The finance industry has a reputation of being exclusive, convoluted, and elitist. The aim of this project was to display that the major ideas behind the stock market are rather accessible using basic statistics. At first glance, a term such as "the Fama-French five-factor model" may seem abstract and intimidating to the average American, but a comprehensive overview of its components accompanied by intuitive data visualizations has the potential to massively demystify the stock market, empowering prospective investors to conduct further research and take rein of their personal finances.