AMS572 Project Report
Group 01

Denis Sullivan & Willie Bedoya

1. Introduction

In this report, two non-trivial hypotheses of scientific interest in the field of quantitative finance are formulated and tested. The dataset for testing the two hypotheses contains 23 variables; for the first hypothesis there are 2 continuous and 1 categorical, while for the second hypothesis there are 15 continuous and 5 categorical. Additionally, the dataset for each has 333 samples.

These two hypotheses are as follows:

(1) The mean monthly return for the S&P 500 index is the same in periods of high inflation and low inflation.
   (a) H_0: $\mu_{high} = \mu_{low}$
   (b) H_a: $\mu_{high} \neq \mu_{low}$

(2) Variables such as commodity prices, US government treasury bonds, corporate bond yields, and categorical market descriptions are not statistically significantly related (linearly and/or non-linearly) to the monthly spread return between the Russell 2000 index (small-cap companies) and the S&P 500 index (large-cap companies).
   (a) H_{0j}: $\beta_j = 0$ for each parameter β_j
   (b) H_{1j}: $\beta_j \neq 0$ for each parameter β_j

First Hypothesis

The first hypothesis will be tested using a two-sample t-test approach. This widely used approach can be a powerful instrument for determining whether the means of two independent groups differ significantly. Macroeconomic factors, such as inflation, exert a significant and pervasive impact on the overall economy, influencing numerous aspects of economic activity across various sectors. Analyses like the one that will be performed to test our hypothesis are essential for understanding the impact these factors can have in the context of finance. With financial markets growing increasingly more complex, the ability to isolate the effects of inflation provides valuable insights into the behavior of indices under varying economic conditions.

This hypothesis seeks to explore whether the mean monthly return for the S&P 500 differs during periods of high inflation versus low inflation. Periods of high inflation will be defined as months having the inflation rates in the top 25% of our observations whereas periods of low inflation will be defined as months having inflation rates in the bottom 25% of observations. The necessary assumptions needed to conduct a t-test, namely the normality of returns within each group as well as the equality of variances, will be tested prior to conducting the analysis. If necessary, alternative methods of testing will be explored that better fit with our findings.

The hypothesis is of scientific interest for a number of reasons as it ties directly to the broader understanding of how inflation affects financial markets. Inflation has a significant effect on consumer behavior with periods of high inflation often characterized by increased uncertainty and less spending while periods of low inflation are often seen as times with increased spending and higher risk-taking.

Testing whether these differences have a statistically significant impact on the returns of the S&P 500 can shed light on how inflation affects market dynamics.

If a significant difference in the mean returns across the two groups is to be found it would have significant implications. The first being that inflation, a known macroeconomic indicator, may not be fully priced into the market indicating that the market is not efficient. This would indicate the potential for profitable investment strategies designed to take advantage of the inefficiency. A better understanding of the relationship between inflation and the market's returns would have the potential to drive investment and risk management strategies.

Second Hypothesis

The second hypothesis shall be tested via a multiple regression approach. This is a commonly-used approach for attempting to fit a relationship between a series of independent variables and a single response variable. In the field of quantitative finance particularly, relationships may not be straightforwardly linear, so having the ability to fit higher-order models is desirable. The three key assumptions of multiple regression will be tested: homoscedasticity of residuals, normality of residuals, and the absence of multicollinearity. After performing the fit, statistical inferences will be conducted on each of the fitted parameters β_j to determine which variables have statistically significant relationships with the monthly spread return between the Russell 2000 index and S&P 500 index. Since the null hypothesis being tested claims that these variables do not have a significant relationship (whether it be linear or quadratic) with the spread return, if β_j is not found to be significantly different from zero then the null hypothesis for that variable is not rejected.

This hypothesis is of scientific interest for a couple of reasons. The first is related to the Fama-French models. It is well-known that in the Fama-French three-factor model, one factor is the effect of small versus large market capitalization on the expected rate of return.[3] Historically, small-cap companies have had greater returns over large-cap companies. However, their model does not address the reason behind this. This hypothesis seeks to determine which variables may have significant relationships with the difference in monthly returns between small-cap stocks and large-cap stocks, which is a widely-accepted stylized fact.

Additionally, an answer to the second hypothesis may of course be practically useful. If it turns out that some commodities and yields have a statistically significant relationship with relative returns, this could be exploited and built upon by quantitative hedge funds. For these reasons, the second hypothesis is a scientific question of interest, as it may help explain the reasons why small-caps outperform large-caps. This hypothesis is also non-trivial, since although it is widely-known that there is an outperformance it is not clear what relationship this performance may have with commodities, yields, and other categorical variables of the market.

2. Data Description

First Hypothesis

The nature of the first hypothesis required a targeted selection of variables:

- SP_Return: The S&P 500 index monthly return %.
- Inflation_Rate: The monthly inflation rate %.
- Inflation_Group: categorical variable used to delineate periods of high/low inflation

        - If the observed inflation rate was below the 25th percentile of all inflation rates the month-long period was classified as Low.
        - If the observed inflation rate was above the 75th percentile of all inflation rates the month-long period was classified as High.
        - All other observations were placed into the medium group.
    - Month: identifier used to explore seasonal trends in inflation rates.

The data spans from January 1997 to October 2024, consistent with the second hypothesis. S&P 500 data was downloaded from while inflation data was calculated using CPI data provided by.[10,2]

Second Hypothesis

To test the second hypothesis, a series of United States Treasury yields and commodity prices from January 1997 to October 2024 was collected. This translates to 333 months, or samples, of data. Categorical data is also considered, such as whether the S&P 500 is in a bear market or not (defined as the index being in at least a 20% drawdown), the current financial quarter, and what political party occupies the White House. Fortunately, the sample period includes many outlier events, such as the Dot-Com Bubble, 2008 financial crisis, 2020 COVID-19 crisis, and the 2022 inflation bear market. With that being said, the dataset also includes a very long bull market ranging from around 2009 to 2020. Having a long enough time frame ensures that multiple market periods are accounted for in the multiple regression.

The fifteen continuous variables are:

1. Yields of corporate bonds rated CCC and lower
2. Yields of corporate bonds rated BBB
3. Yields of high yield corporate bonds (typically bonds rated BBB and lower)
4. Market price of copper
5. Market price of natural gas
6. Yields of corporate bonds rated AAA
7. Market price of gold
8. Market price of corn
9. Yields of US 1-year treasury bill
10. Yields of US 5-year treasury bond
11. Yields of US 30-year treasury bond
12. Yields of US 2-year treasury note
13. Market price of crude oil
14. Market price of the CBOE Volatility Index (VIX)
15. Yields of US 10-year treasury bond

The five categorical variables are:

1. Whether the S&P 500 is in a drawdown of greater than 20% (bull or bear market)
2. What the current financial quarter is (Q1, Q2, Q3, or Q4)
3. Whether the S&P 500 is in an uptrend, defined as the current price being greater than the average price over the last 10 months (uptrend or downtrend)
4. Political party of the current president (Democrat or Republican)
5. Whether the S&P 500 is overpriced, defined as the current price to earnings (P/E) ratio being greater than the average P/E ratio over the last 12 months (expensive or cheap)

All of the values in this dataset were downloaded from the market data provider TradingView.[9]

Note that for the two categorical variables which are dependent on a certain number of months having passed (i.e., uptrend/downtrend and expensive/cheap), for the first few initial months in the dataset the average is taken by going as far back as possible. For example, if only 4 months have passed then the average price and average P/E ratio is taken over the last 4 months. Once 10 and 12 months have passed, then the average price and average P/E ratio will be calculated using those timeframes, respectively.

In the bond market, companies that are less likely to be able to pay off future debts are given lower ratings by companies such as Moody's. For example, a company with a CCC corporate bond rating may offer a high yield relative to a AAA corporate bond, however they are also more risky investments and more likely to default on their debt. It is important to include these ranges of different bonds when determining how they may impact the spread of returns between small-cap and large-cap stocks because small-caps are less likely to have high corporate bond ratings, while the inverse is true for large-caps. Additionally, the prices of commodities used in the macroeconomy are important since a significant price increase in oil, for instance, may more negatively impact small companies who can't afford the significant spike in cost when compared to large companies.

Several transformations and adjustments were made to the data. First, before doing the initial regression the monthly prices of the indices and commodities, as well as the bond yields, the month-over-month percent change was calculated to get closer to stationary and normality. Typically, natural logarithms are used to transform the data. However, when this was done on the US treasury bond yields, there were some unusual results during periods of high volatility (i.e., yields decreasing by more than 100% which isn't possible unless the yield turns negative). So instead, monthly percent changes were opted for, and this generally does not significantly alter the results compared to using natural logarithms.

One of the key assumptions in multiple regression is assuming that multicollinearity is not present.[8] This means that the predictor variables x_i are related (linearly, quadratically, or cubically in the case for this model). Matrices may be used for multiple regression analysis, and if the columns of the matrix of predictor variables $\boldsymbol{X}$ are roughly dependent, then the matrix is singular. This means it's inverse can't be used to solve for the fitted parameters $\widehat{\beta} = (X'X)^{-1}X'y$, where $\widehat{\beta}$ is the vector of fitted parameters and y is the vector of response variables. This also leads to matrix $V = (X'X)^{-1}$ being very large, where each element of the matrix v_{jj} is related to the variance of the parameter $\widehat{\beta j}$ by Var($\widehat{\beta j}$)=$\sigma^2 v_{jj}$ and σ^2 is the error variance between the vector of response variables y and vector of fitted response variables $\widehat{y}$.

In short, multicollinearity leads to two main problems: the estimates for the parameters $\widehat{\beta j}$ are unreliable, and many of the coefficients may have large standard errors and consequently be statistically indistinguishable from 0. To remedy this potential issue, the multicollinearity of the variables can be measured via the determinant of the correlation matrix $R = (n - 1)^{-1}X'X$, where n is the total sample size. The determinant of this matrix will range between 0 if there is multicollinearity and 1 if all correlations are zero. If it is found that the determinant is near 0 (say below 0.1), then the diagonal elements of R will be inspected. Another method is to inspect elements called variance inflation factors (VIF), and if its value is greater than 10 then that predictor variable will be deemed as approximately dependent on other predictor variables. The VIFs shall be calculated in R after an initial fit of the multiple regression model. Any predictor variables with a VIF value greater than 10 will be discarded to ensure that multicollinearity is not a problem. Then, a new multiple regression model will be fit that doesn't include these multicollinear variables.

3. Exploratory Data Analysis

First Hypothesis

Once the dataset was loaded into R, preliminary exploratory data analysis was conducted to gain insights into the distribution of key variables, particularly SP_Returns and Inflation_Rate as they are central to the hypothesis at hand.

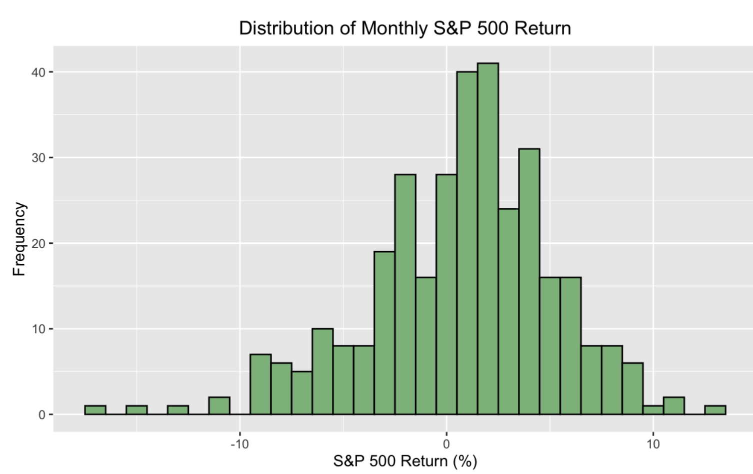
Table 1. Summary Statistics of Monthly Percent Changes of the S&P 500 Index and Inflation Rate

Summary Statistics of Monthly Percent Changes of The S&P 500 Index and Inflation Rate

	SP_Return	Inflation_Rate
Min	-16.9425	-1.7706
1st Quartile	-1.8023	0.0614
Median	1.1830	0.2011
Mean	0.7133	0.2056
3rd Quartile	3.5789	0.3423
Max	12.6844	1.3769
Std. Dev.	4.4799	0.2930

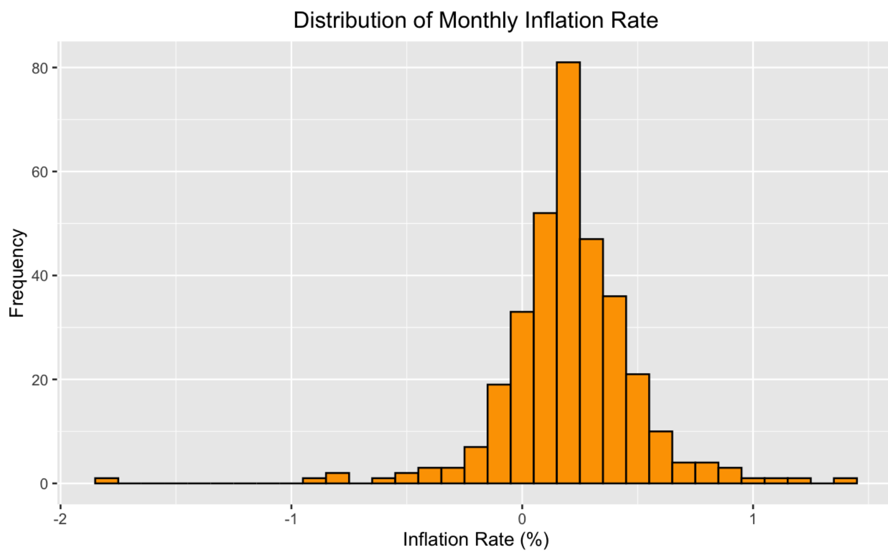
The summary statistics of the S&P 500 index returns and the inflation rate returns provide valuable insights into the two variables. The S&P 500 has experienced a wide range of monthly returns spanning from a substantial -16.94% decline to a significant increase of 12.68% while having a mean monthly return of just 0.71%. The positive skewness seen is consistent with expected typical market conditions, however the large minimum return underscores the potential for substantial market downturns.

Figure 1. Distribution of Monthly S&P 500 Returns



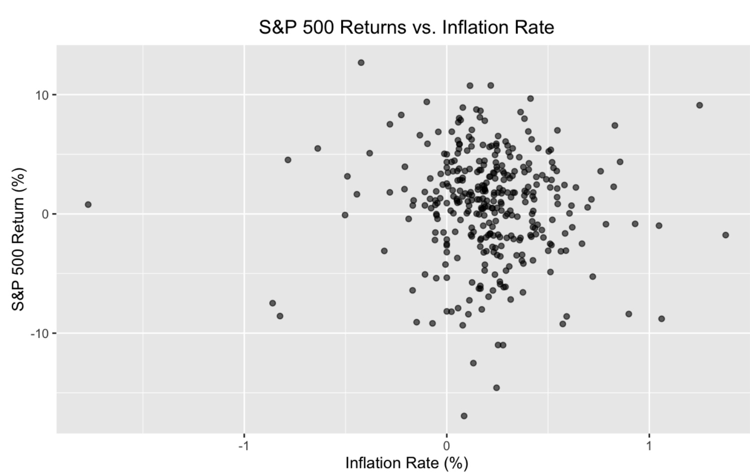
The distribution of monthly inflation rates also displays variability while having a narrower range of values when compared to the S&P 500 returns. Like the S&P 500 returns, a positive skew can be seen in the histogram of inflation rates aligning with the historical trend of prices increasing over time.

Figure 2. Distribution of Monthly Inflation Rate



Following our examination of the individual distributions, we shifted our focus to investigating the potential relationship between the S&P 500 monthly returns and the monthly inflation rates. A scatter plot, which can be seen in Figure 3, was created to further explore the relationship between the variables. The plot does not show any strong patterns between the two, suggesting a weak or non-existent linear relationship. These findings were further corroborated after the correlation coefficient between S&P Returns and Inflation Rate was calculated to be -0.0443. This value's proximity to 0 indicates that fluctuations in one of the variables has very little predictive power for the other.

Figure 3. Scatter plot of S&P 500 Returns vs. Inflation Rate

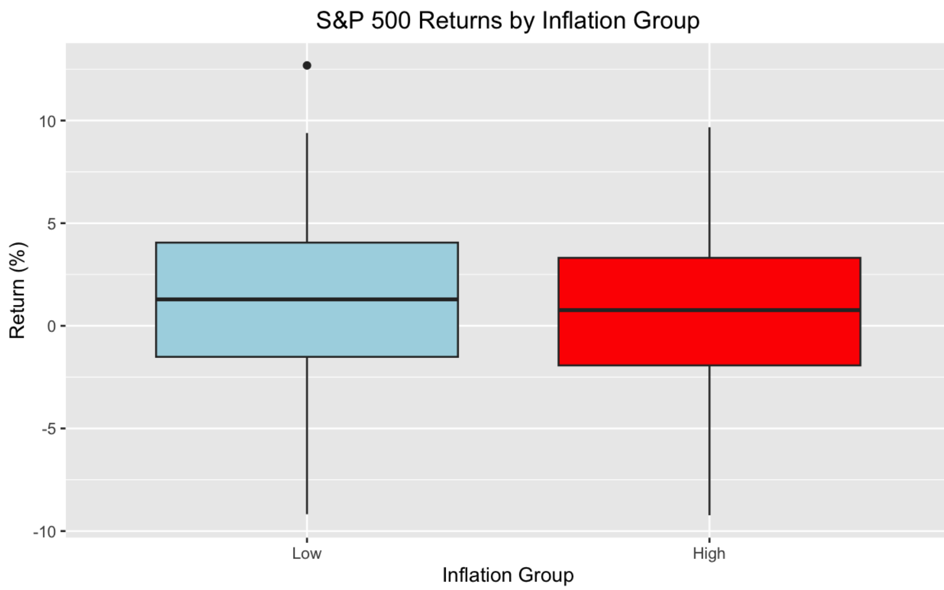


To prepare our data for the t-test, we needed to split it into two groups. In order to achieve this, a new variable "Inflation_Group" was created using the following criteria:

- If the date's inflation rate value was less than the 1st quartile inflation rate ~ "Low"
- If the date's inflation rate value was greater than the 3rd quartile inflation rate ~ "High"
- All other observations were assigned the value "Medium"

Using this new variable, we filtered our data to solely include observations falling into either the low or high inflation group. Figure 4 shows the distributions of the S&P 500 monthly returns across the groups.

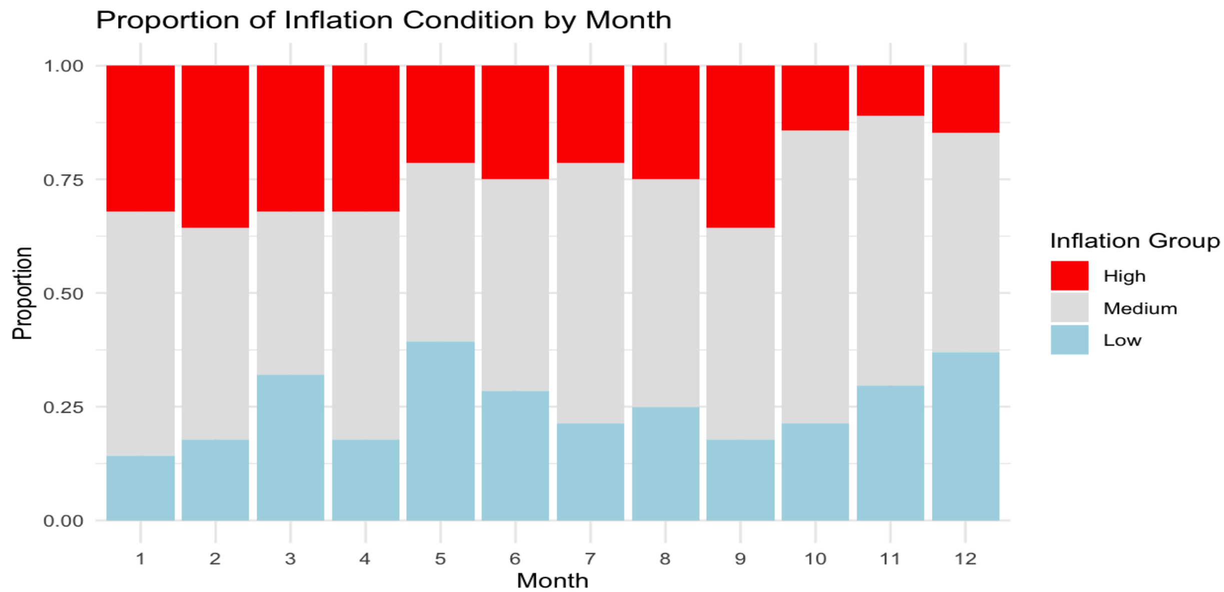
Figure 4. S&P 500 Returns by Inflation Group



The returns seen in the low inflation group had a higher 1st quartile, median, and 3rd quartile when compared to the high inflation group. While there is significant overlap between the two distributions, the shift upward of Low inflation group quartiles suggests that there may be a significant difference in the monthly S&P 500 returns between the groups.

Additionally, we explored any potential seasonal patterns in inflation by constructing a bar chart where each bar represented a month and was segmented into three colored sections corresponding to the proportion of observations falling within each inflation group. This chart, seen in *figure 6*, did not reveal any clear seasonal patterns in inflation however.

Figure 5. Proportion of Inflation Condition by Month



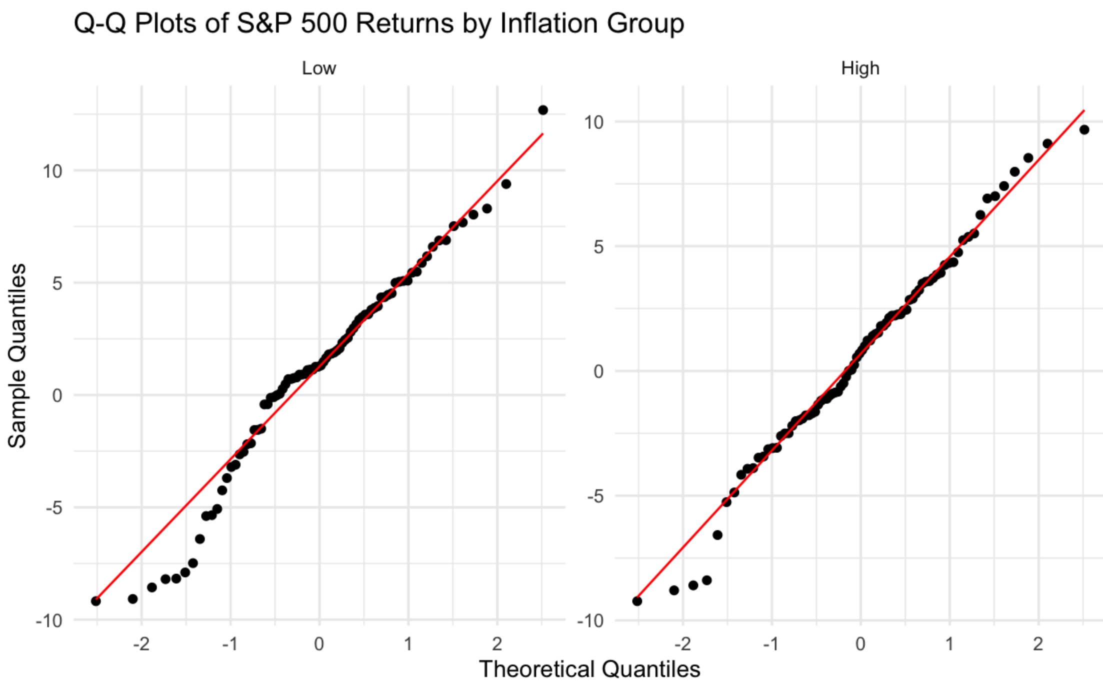
Before proceeding with the t-test, we conducted preliminary analysis on our data to ensure that the necessary assumptions were met. We first assessed the normality assumption by performing Shapiro-Wilks tests on the returns of both the high and low inflation groups. The results of the tests, as seen in Table 2 allowed us to conclude that our data is likely normally distributed at the $\alpha = 0.05$ level.

Table 2. Shapiro-Wilk Test Results

	Low Inflation	High Inflation
W	0.9709	0.0540
P-value	0.9874	0.5929

To further confirm that our data was normally distributed, we created Q-Q plots of the S&P 500 returns for both inflation groups. These plots, as seen in Figure 6, roughly follow a straight line providing further confirmation that our data is normally distributed.

Figure 6. Q-Q Plot of S&P 500 Returns by Inflation Group



To verify that the equal variances assumption was met, an F test was performed under the following hypotheses:

- H_0: $\sigma^2_{high} = \sigma^2_{low}$
- H_a: $\sigma^2_{high} \neq \sigma^2_{low}$

Because our calculated F-statistic (F = 1.2504) is less than the critical F-value ($F_{0.025}$ = 1.43787) at the α = 0.05 level for 83 degrees of freedom in both the numerator and denominator, we do not have sufficient evidence to reject the null hypothesis that the variances for the returns of the low and high inflation groups differ.

Given that the assumptions of normality and homogeneity of variances were met, we then proceeded with conducting our t-test.

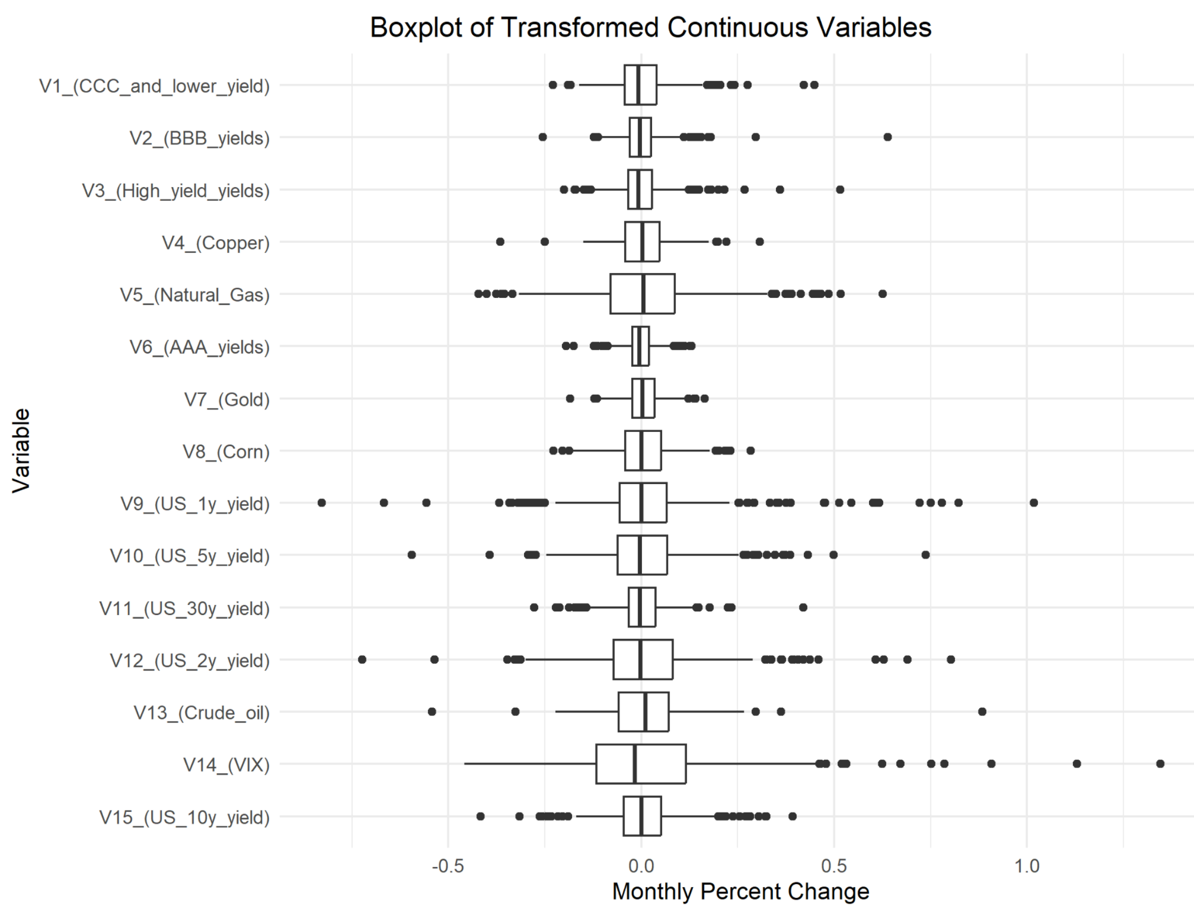
Second Hypothesis

Several approaches will be taken to explore the dataset. First, summary statistics for each of the continuous variables will be computed via R. Second, boxplots of the continuous variables will also be generated. This will help in detecting any potential outliers in the variables. To assess the categorical variables, frequency plots which count how many times (months) each category occurs will be generated for each of them. Third, a scatter plot for each predictor variable will be made to compare the bivariate relationships between the response variable and each of the continuous predictor variables. These scatter plots also include a linear model fit using the least squares method. Lastly, the correlation matrix of continuous variables and VIF values will be computed to determine if there is any potential issue of multicollinearity.

Table 3. Summary Statistics of Monthly Percent Changes of Continuous Variables

Variable	N	Mean	Std. Dev.	Min	Pctl. 25	Median	Pctl. 75	Max
V1_(CCC_and_lower_yield)	333	0.0025	0.08	-0.23	-0.043	-0.0077	0.04	0.45
V2_(BBB_yields)	333	0.00099	0.065	-0.25	-0.029	-0.0037	0.025	0.64
V3_(High_yield_yields)	333	0.0015	0.071	-0.2	-0.034	-0.0078	0.028	0.52
V4_(Copper)	333	0.0069	0.072	-0.36	-0.041	0.0026	0.048	0.31
V5_(Natural_Gas)	333	0.013	0.16	-0.42	-0.079	0.0054	0.088	0.63
V6_(AAA_yields)	333	-0.00035	0.042	-0.2	-0.022	-0.0047	0.02	0.13
V7_(Gold)	333	0.0073	0.046	-0.18	-0.023	0.0036	0.034	0.16
V8_(Corn)	333	0.0043	0.078	-0.23	-0.042	0.00094	0.051	0.28
V9_(US_1y_yield)	333	0.018	0.19	-0.83	-0.056	0	0.067	1
V10_(US_5y_yield)	333	0.0076	0.13	-0.59	-0.061	-0.0028	0.068	0.74
V11_(US_30y_yield)	333	0.0011	0.07	-0.28	-0.033	-0.0028	0.038	0.42
V12_(US_2y_yield)	333	0.013	0.17	-0.72	-0.071	-0.0025	0.082	0.8
V13_(Crude_oil)	333	0.009	0.11	-0.54	-0.058	0.012	0.072	0.88
V14_(VIX)	333	0.023	0.23	-0.46	-0.12	-0.016	0.12	1.3
V15_(US_10y_yield)	333	0.0034	0.097	-0.42	-0.045	0.00067	0.051	0.39

Figure 7. Boxplot of Monthly Percent Changes of Continuous Variables



From Table 3 and Figure 7 shown above, there are a number of conclusions that can be made about the continuous variables. First, the variable with greatest mean month-to-month change was the VIX with 2.3%, although it surprisingly also had the lowest median at -1.6%. These values indicate that the VIX data is skewed towards the higher end, with a few very large values pulling the mean upwards while more than 50% of the data points are near the lower-end. This makes sense, as it is well-documented how volatility in the market can skyrocket during times of crisis and rapidly go from a level of 20 to over 60. With that being said, most of the variables have a mean and median close to zero, with the greatest absolute median being 1.2% when not considering the VIX. It is visually clear from the boxplots and also numerically shown in the summary statistics that this is the case. This suggests that while there is great variability in the month-to-month changes in these variables, overall there is little trend in them and there is no significant directional bias over the observed period. This falls in line with expectations, which is that most traded assets have on average daily returns of nearly 0%, while monthly returns may average to be at most roughly 1%.

Second, some variables were found to have very high standard deviations (i.e., high volatility). Once again, the VIX had the highest standard deviation of 23%, and this is not unexpected as occasional market shocks result in large spikes in volatility. Other variables with large standard deviations include natural gas, crude oil, and the 1-year, 2-year, and 5-year US treasury yields. Commodities like crude oil are known for being very volatile assets, so again this result was not unexpected and aligns with their sensitivity to the macroeconomy as well as geopolitical issues.

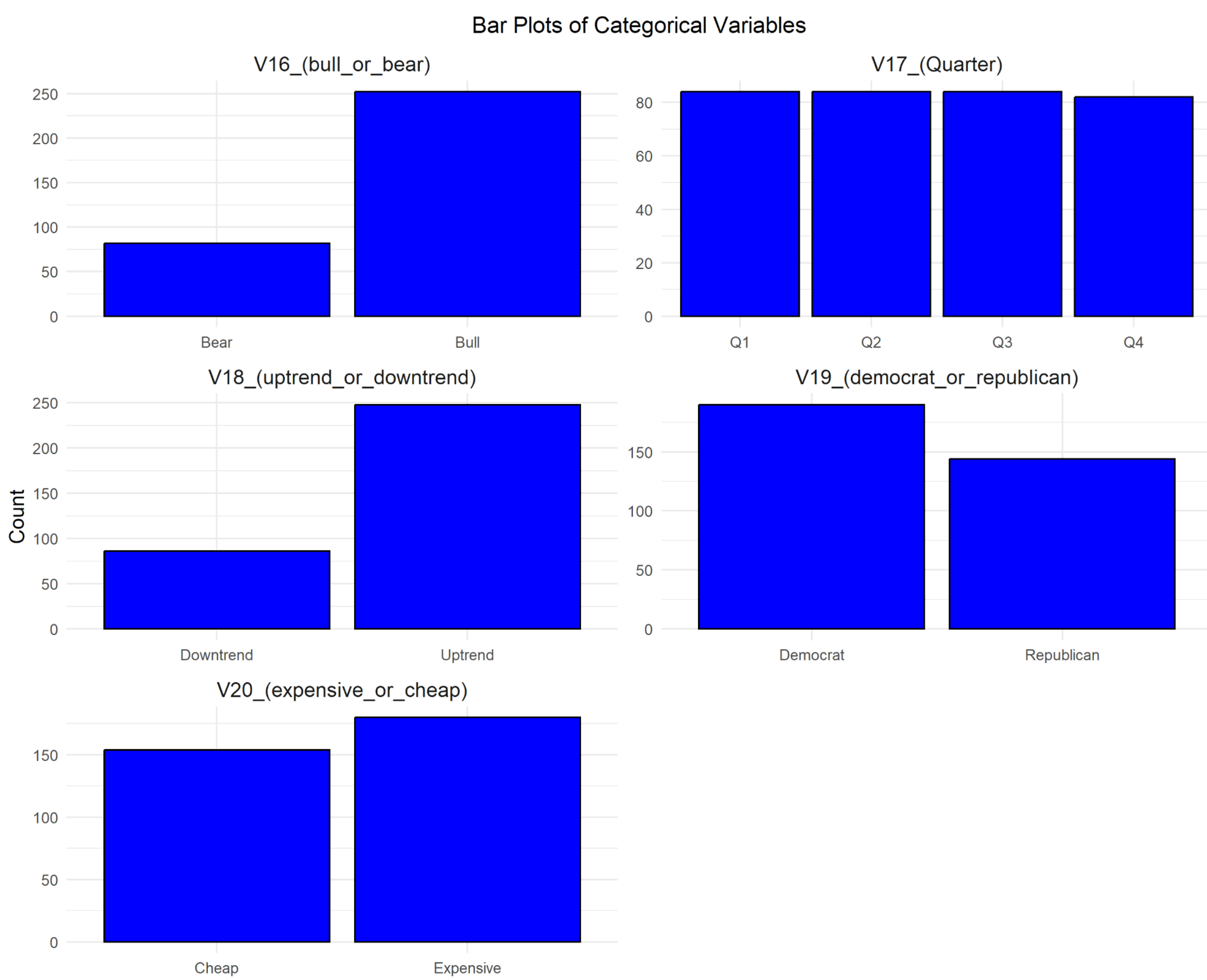
Third, at the maximum and minimum extremes the VIX is once again found to be an outlier amongst the other variables. It experienced the largest maximum change at a 130% increase in one month. After that comes the 1-year US treasury yield and crude oil. The variables that experienced the lowest minimum change were the 1-year, 2-year, and 5-year US treasury yields. Another trend appearing here is that the short-term US treasury yields (i.e., 5 years or less) are experiencing large extremes, in terms of standard deviations and minimum/maximums. This is also not entirely unexpected; while bonds are generally low-volatile assets, yields themselves may experience more volatility because they are typically more constrained in what values they can take. For instance, short-term yields may range between 0% to 2%. In such a case, a change from 2% to 1% would represent a 50% decrease, or a 100% increase in the other direction. Additionally, short-term yields for most of the 2010s were extremely low, typically below 1% or even 0.5%. Thus, even very small changes like 0.25% could represent monthly percent changes of 50% to 100%. Short-term yields are also more influenced by changes in the federal funds rate, which is directly controlled by the Federal Reserve and are hence more susceptible to large rises/falls. For these reasons, it is unsurprising that short-term yields have been found to have large standard deviations and minimum/maximums.

On the other end of the spectrum, variables like corporate bond yields (AAA, BBB, CCC, and high yield), longer-term US treasury yields like the 30-year, and commodities such as gold and corn were shown to have lower standard deviations/volatility. These results are also not unusual. Indices that track corporate bonds typically have higher durations, so they are less susceptible to changes in the federal funds rate and not as volatile, which was similarly found for long-term US treasury yields like the 30-year. Additionally, it can be seen from Table 3 that corporate bonds with lower ratings (i.e., CCC or high yield) had higher standard deviations than those with higher ratings (AAA or BBB). This makes sense, as lower-rated corporate bonds are associated with riskier companies and greater uncertainty/volatility. Lastly, commodities like gold and corn are not energy-related variables and not as sensitive to geopolitical shocks, so their lower standard deviations seem sensible.

A last observation is that there are quite a number of outliers shown in Figure 7. Any values that are not within the first quartile minus 1.5 times the interquartile range and the third quartile plus 1.5 times the interquartile range are considered outliers. It is a well-known stylized fact that asset or market returns are very heavy-tailed, with extreme values occurring more frequently than would be predicted by a normal distribution. On top of that, there are a large number of samples, so even if an outlier is taken to mean outside of 95% of the values, this would leave approximately 16 outliers per variable. These are a couple of reasons for why there may be a noticeable amount of outliers in Figure 7.

Overall, a few general conclusions can be made about the continuous variables. First, variables tied to energy (i.e., natural gas and crude oil) have higher variability and this reflects their sensitivity to global economic events. Second, short-term US treasury yields also experienced high standard deviations, as opposed to their long-term counterparts like the 30-year US treasury yield and corporate bond yields. Third, traded assets have heavy-tailed distribution, so the variables tend to have quite a number of outliers. Lastly, some assets such as gold, high-grade corporate bond yields (AAA), and corn tend to be very stable variables with low variations over time.

Figure 8. Bar Plot of Categorical Variables



From the five bar plots of categorical variables in Figure 8, there are a few notable observations. Firstly, the uptrend/downtrend and bull/bear market categories are remarkably similar. There were 248 months of uptrend and 252 months of bull markets, and 86 months of downtrend and 82 months of bear markets. This suggests that these two categorical variables represent nearly the same thing (positive trending market or negative trending market), and there may be an issue of collinearity here. VIF values will be calculated in the future to confirm if this is the case or not. A second observation is that the expensive/cheap categorical variable does not appear to be highly related to the bull/bear or uptrend/downtrend variables. This is a bit surprising, since one might suspect that an expensive market suggests the market is in an uptrend and vice versa. One reason this may not be the case is because the expensive/cheap category incorporates earnings, while the bull/bear and uptrend/downtrend both only incorporate price. Lastly, Democrats occupied the White House more frequently than Republicans over the sample period. While it is not expected that the spread between small-cap stocks and large-cap stocks benefits from one political party over another, it is good to know that there is a discrepancy over the sample period in the event that a statistically significant relationship between political party and spread returns is found.

Figure 9. Scatter Plots of Monthly Percent Changes in Continuous Variables 1 – 4

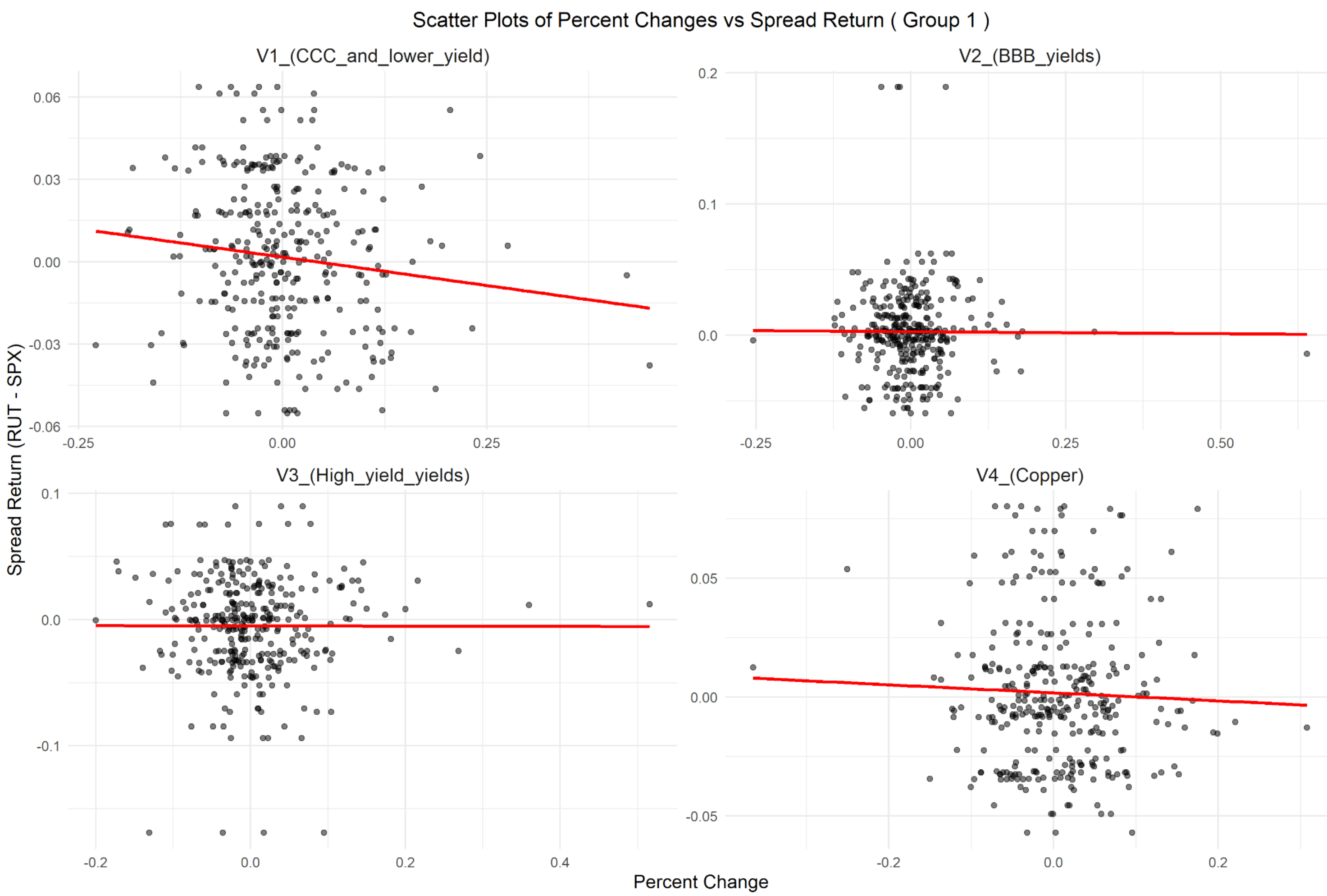


Figure 10. Scatter Plots of Monthly Percent Changes in Continuous Variables 5 – 8

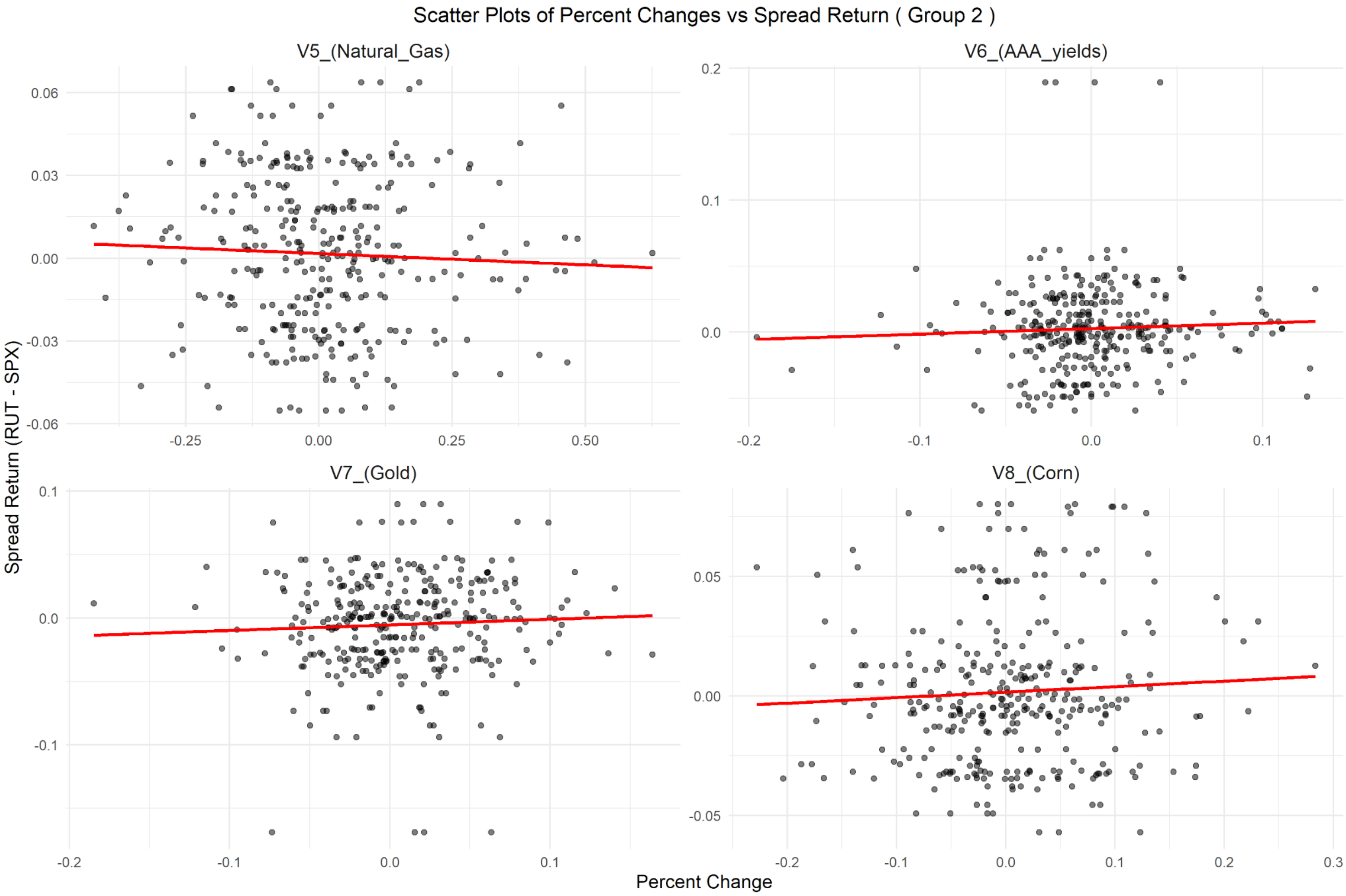


Figure 11. Scatter Plots of Monthly Percent Changes in Continuous Variables 9 – 12

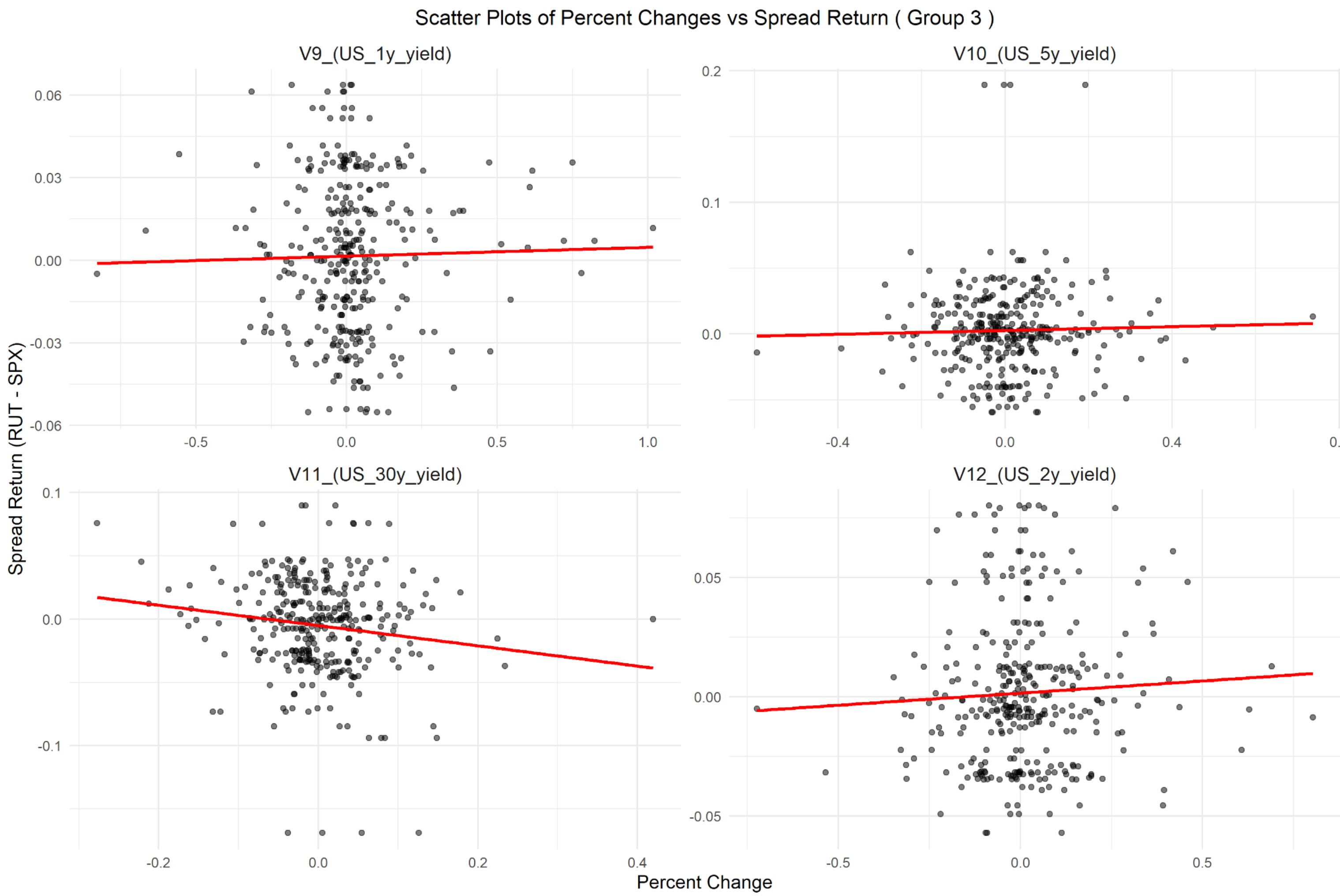
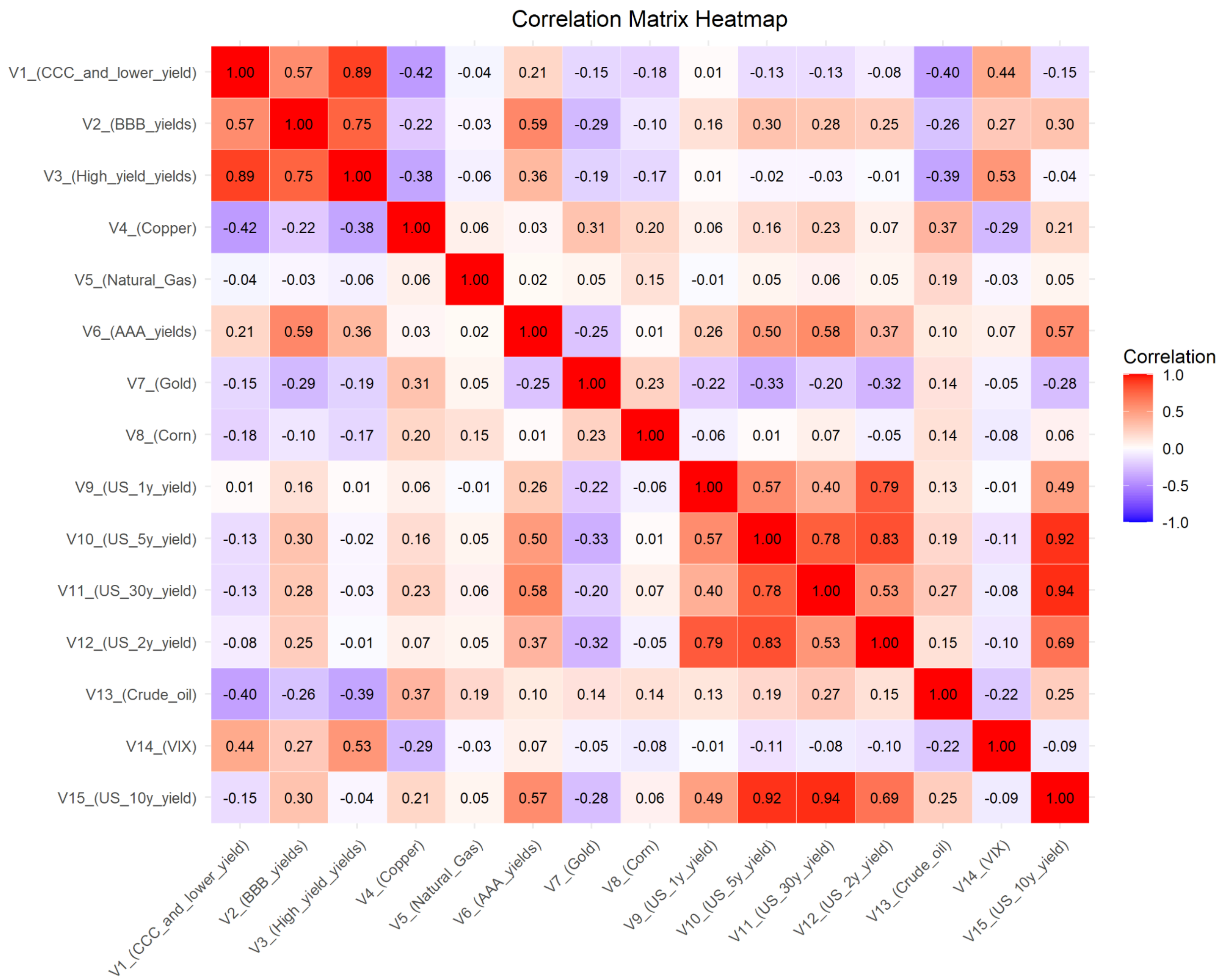


Figure 12. Scatter Plots of Monthly Percent Changes in Continuous Variables 13 – 15

From the scatter plots comparing the spread return to the predictor variables in Figures 9 – 12, it is clear that some bivariate relationships appear to have little dependency while others have more positive or negative sloping trends. In the first group of 4 scatter plots, the BBB yields and high yield corporate bond yields appear to have little relationship with the spread return between the Russell 2000 and S&P 500. However, the CCC yields appear to have a considerable negative relationship with the spread, while copper has a slightly and possibly insignificant negative relationship. In the next group of 4 variables, natural gas appears to have a slightly negative relationship with the spread, while AAA bond yields, gold, and corn all have slight positive relationships. In the third group, the 1-year, 2-year, and 5-year US treasury yields have slight positive relationships with the spread, while the longer-duration 30-year has a more significant positive relationship. For the last three continuous variables, the VIX and crude oil seem to have virtually no relationship with the spread, while the 10-year yield has a slight positive relationship.

For the last portion of the exploratory data analysis, the correlation between continuous variables will be analyzed.

Figure 13. Correlation Matrix Heatmap of the 15 Continuous Variables



The heatmap shown above in Figure 13 illustrates that most of the continuous variables are not significantly correlated with one another. The highest correlation is between the 10-year and 30-year US treasury yields, which has a correlation of 0.94. The 10-year and 5-year US treasury yields follow right behind that with a correlation of 0.92, which is then followed by the high yield and CCC corporate bonds with a correlation of 0.89. Therefore, it can be concluded that some of these variables appear on the surface to suffer from multicollinearity. In the next section, an initial multiple regression model will be fit along with VIFs calculated. For any VIFs above 10, those corresponding variables will be discarded and a new model will be fit.

4. Answering Hypotheses

Hypothesis 1

Our first hypothesis sought to determine whether the mean monthly returns of the S&P 500 significantly differs between periods of high inflation and low inflation. In order to carry this test out, a two-sample t-test was conducted with the inflation levels being categorized into “High” and “Low” groups based on thresholds from the data.

The results of the t-test did not indicate any significant differences in the mean monthly returns of the S&P 500 across the two inflation conditions. While the test revealed that there was a spread between mean monthly return during low inflation periods and high inflation periods (1.106% vs. 0.602% respectively), the difference did not prove to be statistically significant. Our calculated t-value, 0.75752, fell well short of the critical value of 1.97436. This sentiment was echoed in our p-value of 0.4498 which is much greater than the 0.05 significance level. Additionally, our 95% confidence interval for the mean difference in S&P 500 monthly returns was [-0.811, 1.821]. Because 0 falls within the defined range, the possibility of the mean difference in S&P 500 monthly returns between inflation groups being 0 cannot be ruled out. These findings suggest that we do not have enough evidence to reject the null hypothesis that the mean monthly return of the S&P 500 is equal in periods of high inflation and low inflation.

Hypothesis 2

To test the second hypothesis that continuous (commodity prices, US government treasury bonds, etc.) and categorical variables (market uptrend, expensive, etc.) have statistically significant effects (linear and/or quadratic) on the monthly spread return between the Russell 2000 index and the S&P 500 index, a second-order multiple regression model was created. This means that not only were linear relationships considered, but quadratic relationships were analyzed as well. However, it was found that the second-order terms yielded no statistically significant relationships whatsoever. The largest absolute t-value was 1.118, which is far below the threshold required for 2-sided t-tests at a statistical significance level of $\alpha = 0.05$ (corresponding to a t-value of 1.96). Therefore, to streamline the model quadratic terms were removed and a new model was generated using multiple linear regression (as opposed to general multiple regression).

Next, after fitting this new first-order model, the VIF values for all 20 variables were calculated. It was found that the following variables had VIFs above 10:

1. Yields of high yield corporate bonds (VIF = 11.4)
2. Yields of US 5-year treasury bond (VIF = 16.9)
3. Yields of US 30-year treasury bond (VIF = 14.2)
4. Yields of US 10-year treasury bond (VIF = 35.2)

To determine which of these variables should be kept and which should be discarded, the variable with the greatest VIF was removed first. In this case, that was the yield of the US 10-year treasury. After removing this variable, only one variable with a high VIF remained: yields of high yield corporate bonds (VIF = 11.3). Clearly, the US 10-year treasury yield had multicollinear relationships with US treasury yields, like the 5-year and 30-year. Removing this one variable helped reduce the VIF of these variables from 16.9 and 14.2 to 7.5 and 3.8, respectively. The last step was to remove the high yield corporate bond variable, as this still had a high VIF.

From this new model of linear terms only, and discluding the 10-year treasury yield and the high yield corporate bond yields, it was found that only five variables were statistically significant ($\alpha \leq 0.05$):

1. Yields of corporate bonds rated CCC and lower (t-value = -5.091, $\widehat{\beta}_1 = -0.1654$)

2. Yields of US 5-year treasury bond (t-value = 3.024, $\widehat{\beta}_2 = 0.1057$)
3. Yields of US 30-year treasury bond (t-value = -2.026, $\widehat{\beta}_3 = -0.0952$)
4. Yields of US 2-year treasury note (t-value = -2.257, $\widehat{\beta}_4 = -0.0626$)
5. Market price of crude oil (t-value = 2.061, $\widehat{\beta}_5 = 0.0392$)

Some variables which were not statistically significant but were marginally significant ($\alpha \leq 0.10$) include the US 1-year treasury note yield (t-value = 1.601) and the uptrend categorical variable (t-value = -1.598). The coefficient for the intercept was estimated to be $\widehat{\beta}_0 = 0.0042551$, but it was not statistically significant (t-value = 0.720). This represents the expected spread return (Russell 2000 minus S&P 500) when all the predictor variables are equal to 0.

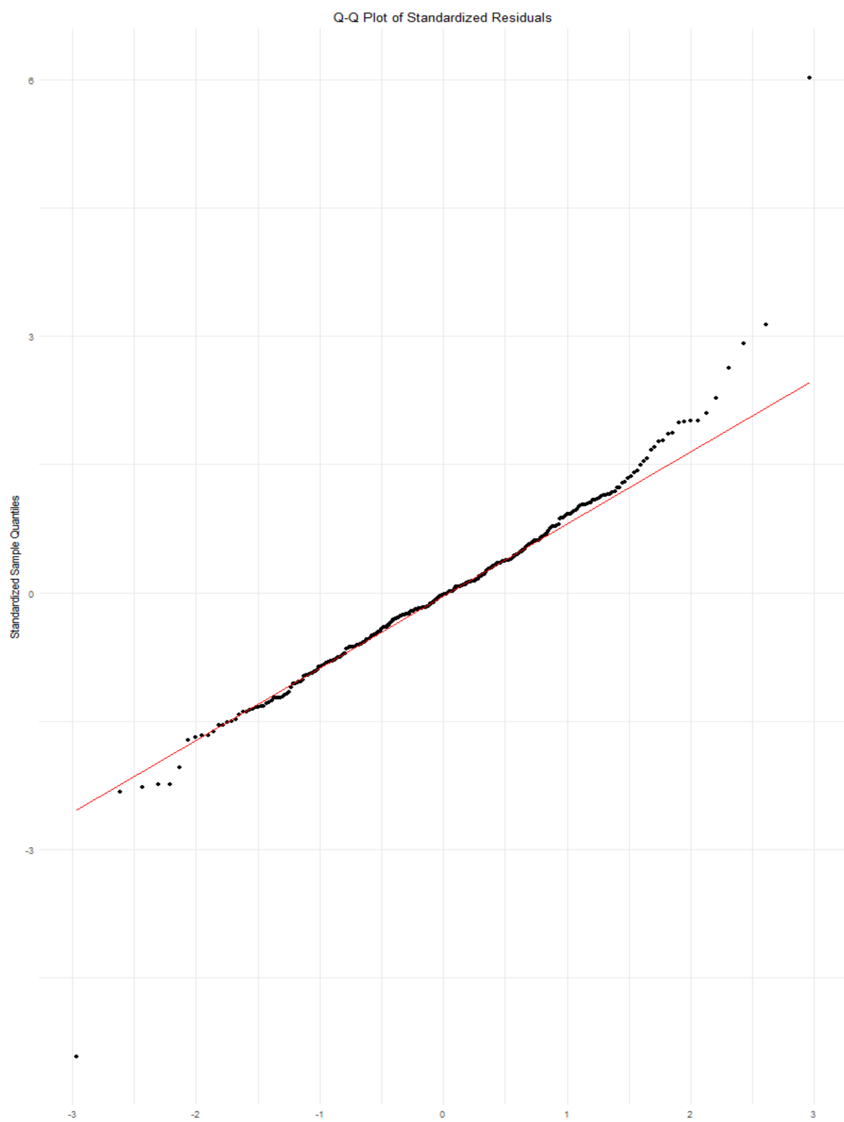
These results provide clarity on what variables are statistically significant in their relationship to the monthly spread return between small-caps and large-caps. It is clear that yields of bonds are some of the most important variables when it comes to the relationship between different assets and the relative spread between small-cap and large-cap stocks. This includes not just short-term yields like the US 2-year, but also long-term yields like the US 30-year or CCC corporate bonds. The low-grade corporate bond portion of this is also not surprising, since one would expect increases in low-grade yield (typically associated with smaller companies) to signify greater risk in small-cap companies and a related decline in small-caps relative to large-caps. The signs of the coefficients $\widehat{\beta}$ suggest that the relative return has a negative relationship with most yields, like the short-term US 2-year, the long-term US 30-year, and the low-grade CCC corporates. However, for a more intermediate-term yield like that associated with the US 5-year bond, the relative return has a positive relationship. These results indicate that an increase in short- or long-term bond yields is associated with relative underperformance of the Russell 2000 compared to the S&P 500, while an increase in intermediate-term bond yields is associated with relative overperformance of the Russell 2000. It is also clear that multiple durations are significant variables in the relative performance. Future work therefore might involve modeling the entire yield curve and using that as a predictor variable.

Besides US treasury and low-grade corporate bond yields, the only other statistically significant variable was crude oil. This had a positive relationship with the spread return, indicating that small-caps tend to outperform large-caps when the price of crude oil rises. This is contrary to what was assumed, as small-caps are typically more volatile and so one might anticipate them to underperform when oil prices increase (which may indicate a geopolitical shock). However, it was instead found that an increase in oil price is associated with an overperformance of small-caps relative to large-caps. This may be the case because of how index sector weightings differ. As of this month, the Russell 2000 has almost double the sector weighting of energy compared to the S&P 500.[6,7] As a consequence, if oil prices rise this could mean increased profits for energy-related companies, which would favor the index with more energy-related stocks (i.e., the Russell 2000). Further analysis of how the sector weightings between the two indices has changed over the course of the sample period would be needed to make a more definitive conclusion.

The F-statistic was equal to 3.396, with 20 and 312 degrees of freedom. This is equivalent to a very small p-value of 2.258e-06, indicating that the multiple linear regression model provides a significantly better fit compared to a model with no predictors (i.e., rejecting the null hypothesis $\beta_1 = \beta_2 = \ldots = \beta_k = 0$). However, the R^2 was equal to 0.1788 while its adjusted version was equal to 0.1262. These results indicate that the model only explains 17.88% of the variability in the response variable (i.e., the spread return) and only explains 12.62% of the variability when adjusting for the number of predictor variables. While the five variables listed previously have statistically significant relationships with the relative return between small-caps and large-caps, the low R^2 means that the variables only explain a small portion of the returns variability.

There are several assumptions that need to be checked before concluding that the variables found to be statistically significant are so. First, it is assumed that there is a linear relationship between the predictor variables and the response variable (Russell 2000 and S&P 500 spread return). From all of the scatter plots shown in the exploratory data analysis section above, it is clear that these variables have linear relationships and are not quadratic or cubic. This was also shown to be the case when statistical significance tests were conducted and none of the quadratic terms were found to be significant, while multiple linear terms were found to be significant. Second, it is assumed that the residual values are normally distributed.

Figure 14. Q-Q Plot of Residuals



The Q-Q plot above in Figure 14 shows that the residuals appear to be somewhat normal. The upper tail appears to be a bit heavy, which is not surprising given that asset returns themselves are heavy-tailed and the residuals may subsequently tend to have some heavier tails. There are also two very significant outliers, one at the bottom left and one at the top right of the plot. These are likely the result of some extreme outlier month, such as a variable like the VIX doubling or a short-term US treasury yield falling by 50% or more. It is known that these outlier events exist because of the box plots and summary statistics shown earlier, and it is likely this is where these two extreme residual outliers arise from.

Even though the Q-Q plot looks to be approximately normal, there are these two issues as mentioned above. Therefore, to assist with further analyzing whether the residuals are normally distributed, the Kolmogorov-Smirnov normality test was applied. Initially, consideration was given to using the Shapiro-Wilk test. However, given that this test is more suitable for smaller sample sizes and the sample size being worked with is over 300, a 2-sided Kolmogorov-Smirnov test was opted for instead.[5] The null hypothesis for this test is that the sample is drawn from a normal distribution, with a p-value less than 0.05 indicating that this null hypothesis is rejected.[4] The p-value calculated is 0.2443, indicating that the null hypothesis is not rejected and the distribution of the residuals is assumed to be normal.

The third assumption is that there is no multicollinearity amongst the predictor variables. Tests for multicollinearity were already done previously in this section, and variables with VIFs greater than 10 were removed until no such variables existed. As a consequence, the 10-year US treasury bond yield was removed as well as the high yield corporate bond yields. Since none of the variables left have a VIF above 10 (the highest is 5-year US treasury bond yields with VIF = 7.5), the multicollinearity assumption is satisfied.

The fourth and final assumption when conducting multiple linear regression is that the residuals are homoscedastic, which means that the variance of the errors is similar across independent predictor variables. The test used to determine this was the Breusch-Pagan test, of which the null hypothesis is that homoscedasticity is present.[1] Like with the Kolmogorov-Smirnov test, if the p-value is less than 0.05 then the null hypothesis is rejected. The value calculated for the residuals was p-value = 0.792, which indicates that the null hypothesis is not rejected and the residuals are homoscedastic. This concludes the fourth and final assumption required for linear regression.

To cap off this section, it was found that 5 of the 20 variables are statistically significant in their relationship to monthly spread returns between the Russell 2000 and S&P 500 indices. The null hypothesis was that there is not a statistically significant relationship between these predictor variables (yields, commodities, etc.) and the response variable (spread return). In other words, $\beta_j = 0$ for each unknown parameter β_j. In this case, there are $j = 20$ variables. The alternative hypothesis was that there is a statistically significant relationship between predictor variables and the response variable ($\beta_j \neq 0$ for each β_j). After first fitting a quadratic model but dropping all of the second-order terms since none of them were close to being statistically significant, and then removing two variables that created issues of multicollinearity, it was found that the parameters β_j for crude oil, CCC and lower corporate bond yields, and the US 2-year, 5-year, and 30-year treasury yields all had statistically significant parameters at the α = 0.05 level. These parameters were equal to 0.0392, -0.1654, -0.0626, 0.1057, and -0.0952, respectively.

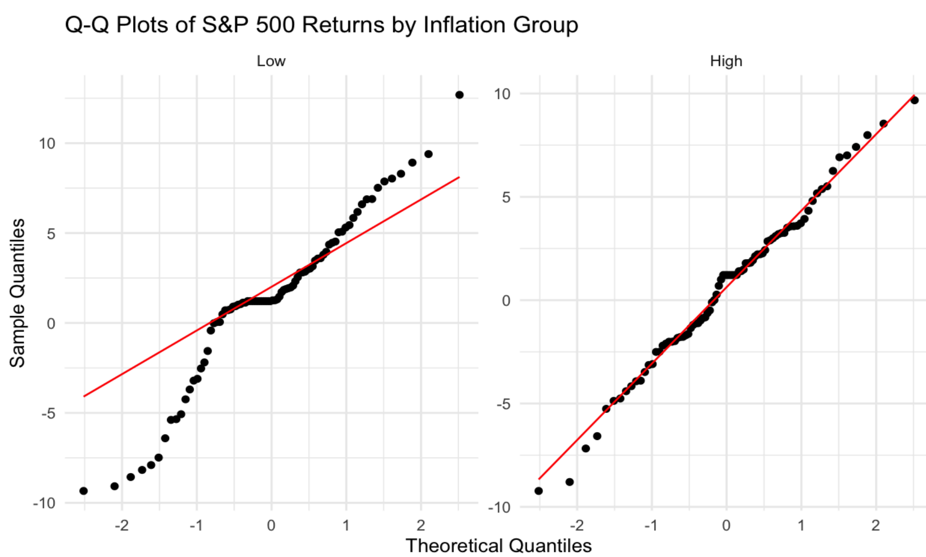
5. Effects of Missing Values

First Hypothesis

The initial type of missing data encountered was Missing Completely at Random (MCAR). To simulate this we decided to randomly sample and remove 10% S&P 500 monthly returns and 20% of the monthly inflation rate observations. To repair our data we decided to replace missing values with the median of their respective columns. This method was selected over others that were considered such as K-Nearest Neighbors imputation due to the fact that no predictive relationship had been uncovered between the S&P 500 returns and the inflation rate.

While imputing the column mean was effective in maintaining the central tendency of the data, it had unexpected drawbacks which forced us to reconsider how we would conduct the t-test. After repairing the data, recalculating the inflation group thresholds, assigning observations to their respective groups and then splitting the data into high and low groups we began re-testing the necessary assumptions for a two sample t-test and were surprised to discover that our data was no longer normally distributed. The p-value of the Shapiro-Wilk test for normality for the low inflation group was a very low 0.002 indicating a lack of normality. The below Q-Q plot for the data confirmed this as well as the data did not adhere to the straight line.

Figure 15. Q-Q Plots of S&P 500 Returns by Inflation Group



Seeing that our normality assumption had been violated, we decided to use the Wilcoxon Rank-Sum Test to test our hypothesis as it does not depend on the assumption of normality. This test yielded a W value of 3957 and a p-value of .1738 leaving us unable to reject the null hypothesis that the median monthly return of the S&P 500 varies in periods of high inflation vs. low inflation.

For the second type of missing data, Missing Not at Random (MNAR) we decided to simulate market crashes by assigning a 70% chance that any observation having a S&P 500 Return less than the 5th percentile of all returns in the data set (95th percentile of losses) would be deleted. Our thought process was that during periods of extreme market downturns we may see increased data reporting issues or market closures leading to missing data. Because of the extreme threshold we set coupled by the probabilistic nature that it would get deleted we were only faced with 14 missing values. Because this represents a small proportion of our dataset, we decided to remove the rows having missing return values. We were able to conduct our analysis in an identical manner to our original attempt, likely due to the dataset remaining largely unchanged. Our t-test yielded the exact same p-value (0.4498) once again leaving us unable to reject the null hypothesis that the mean monthly returns of the S&P 500 varies across low and high inflation time periods. If we set a larger threshold we would likely lose variance in our data, making it more difficult to find statistically significant evidence for the difference of returns across the inflation groups.

Second Hypothesis

The approach taken for handling missing values completely at random (MCAR) in the multiple regression model was to replace any missing values with the historical median (i.e., up until the point in time where missingness occurred) of that variable. This approach is immune from forward-looking biases that may be present in other methods, such as using the median of the entire dataset to replace missing values. In the context of finance particularly, engaging in such bias is incredibly damaging to the model. Another approach would be to simply replace missing values with 0, since financial assets tend to average near 0% returns day-over-day, however this may be less rigorous compared to using the historical median.

Instead of beginning with a second-order multiple regression model, the initial model will be linear only. This is because from the previous section it was obvious that the second-order terms add nothing of value that is statistically significant. The three variables selected to experience 20% missingness were the CCC yields, the price of copper, and the price of gold.

First, it was discovered that not all the same variables as before had VIFs greater than 10. As a result, the only variable that needed to be removed was the US 10-year treasury yield. One of the first conclusions drawn here for MCAR is that the multicollinear variables may not be the same as when there is no missingness.

Then, the following variables were found to be statistically significant after replacing the missing values with their historical medians:

1. Yields of corporate bonds rated BBB (t-value = 2.356, $\widehat{\beta}_1 = 0.1198$)
2. Yields of high yield corporate bonds (t-value = -2.834, $\widehat{\beta}_2 = -0.1771$)
3. Yields of US 5-year treasury bond (t-value = 3.249, $\widehat{\beta}_3 = 0.1137$)
4. Yields of US 30-year treasury bond (t-value = -2.326, $\widehat{\beta}_4 = -0.1111$)
5. Yields of US 2-year treasury note (t-value = -2.154, $\widehat{\beta}_5 = -0.05981$)

The intercept once again was not statistically significant. Variables which were marginally significant ($\alpha \leq 0.10$) were crude oil (t-value = 1.945) and gold (t-value = 1.839). Based on the Breusch-Pagan test for homoscedasticity, p-value = 0.849 so the null hypothesis of homoscedasticity is not rejected and constant variance of residuals is assumed. Additionally, the Kolmogorov-Smirnov test for normality was p-value = 0.1913, so again the null hypothesis of normality is not rejected and a normal distribution for residuals is assumed.

Several observations can be made from these results. First, BBB corporate bond yields are statistically significant now whereas before they were not. Second, high yield corporate bond yields are also statistically significant. In the previous model, this variable had a multicollinear relationship so it had to be removed from the dataset and no statistical significance testing was done on it. It is interesting to see that it is now significant, and it may very well have been significant in the original model too were it not for multicollinearity. Third, the US treasury yields again showed that they are dominant predictor variables. The 2-year, 5-year, and 30-year were also all significant in the original model. Fourth, while crude oil was not statistically significant it was nearly so, with a p-value of 0.05264. Lastly, gold is also now nearly significant, with a p-value of 0.06692. In the original model, it had a p-value of 0.4158, so this is a significant change. From these results, it is clear that having missing values at random considerably altered the conclusion of the hypothesis. Previously, the parameters for high yield and BBB yields did not reject the null hypothesis, where in this case they do. Additionally, the parameter for crude oil which was previously significant no longer rejects the null hypothesis. The parameter for CCC and lower yields also no longer rejects the null hypothesis, as its p-value is only 0.2288 when missingness is present. The parameters for US 2-year, 5-year, and 30-year treasury yields remain significant and reject the null hypothesis.

The approach for handling missingness which is non-ignorable (MNAR) is slightly different. In a theoretical world, MNAR for traded assets may occur during times of significant volatility and crisis. Consider a situation such as the COVID-19 crisis, where some stocks lost half their value in a single month. Markets might halt trading and the release of market data might be delayed. Similarly, imagine that platforms or services that provide prices are overwhelmed with traders attempting to access their accounts so that they can sell off their assets. The platforms may go down and price data would no longer be available for a period of time. Therefore, a non-ignorable type of missingness would likely occur when assets are experiencing considerable volatility and a steep decline in price.

To simulate this, the same three variables as in MCAR (CCC yields, price of copper, and price of gold) were set to experience 10% missingness. The difference will be that instead of using the historical median to fill in missing values, the historical 5th percentile will be used. This will represent a lower value and relate to a mechanism where the missingness means that some shock event has happened where such returns only occur 5% of the time.

Once again, to ensure multicollinearity is not an issue the US 10-year yield had to be removed from the data. The following variables were found to be statistically significant after replacing the missing values with their historical 5th percentile:

1. Yields of high yield corporate bonds (t-value = -2.126, $\widehat{\beta}_1$ = -0.1383)
2. Yields of US 5-year treasury bond (t-value = 2.968, $\widehat{\beta}_2$ = 0.1039)
3. Yields of US 30-year treasury bond (t-value = -2.120, $\widehat{\beta}_3$ = -0.1016)
4. Yields of US 2-year treasury note (t-value = -2.264, $\widehat{\beta}_4$ = -0.06314)
5. Market price of crude oil (t-value = 2.050, $\widehat{\beta}_5$ = 0.03912)

The intercept once again was not statistically significant. Variables which were marginally significant ($\alpha \leq 0.10$) were BBB yields (t-value = 1.918) and CCC and lower yields (t-value = -1.862). Based on the Breusch-Pagan test for homoscedasticity, p-value = 0.867 so the null hypothesis of homoscedasticity is not rejected and constant variance of residuals is assumed. Additionally, the Kolmogorov-Smirnov test for normality had p-value = 0.1401, so again the null hypothesis of normality is not rejected and a normal distribution for residuals is assumed.

Several observations can be made from these results, compared to both the MCAR and original model. First, high yield corporate bond yields are statistically significant like they were in the MCAR model. Once again, it is interesting that this value appears significant when it was not tested in the original model due to multicollinearity. Second, for a third consecutive time the 2-year, 5-year, and 30-year treasury yields are all significant. This repeated result shows that changes in yield across multiple maturities really has a significant relationship with spread returns. Third, the market price of crude oil variable has returned to being significant. It was significant in the original model, but only marginally so in MCAR. Fourth, while BBB corporate bond yields were statistically significant in MCAR they are only marginally significant in MNAR. Similarly, the yield for corporate bonds rated CCC and lower is only marginally significant, compared to significant in MCAR and the original model. Lastly, while gold was marginally significant in MCAR it is nowhere near so in MNAR as it has a t-value of 0.054. From these results, it is clear that not only does having missing values impact the conclusions of the hypothesis, but there are differences between the different kinds of missingness as well.

When considering MNAR, the parameter for high yield corporate bond yields was statistically significant and rejects the null hypothesis. This differs from the original model where it was not even considered due to multicollinearity. Additionally, the parameter for crude oil is significant and rejects the null hypothesis, as was the case in the original model. The only parameter that no longer rejects the null hypothesis when comparing to the original model is that of the CCC and lower yields variable. Although, it is marginally significant. The parameters for US 2-year, 5-year, and 30-year treasury yields continue to remain significant and reject the null hypothesis.

6. Conclusions

First Hypothesis

This study sought to explore the relationship between inflation and the returns of the S&P 500, specifically examining whether the monthly returns of the S&P 500 would differ significantly between periods of high and low inflation. While periods of high inflation are often associated with economic uncertainty and subdued market activity, the results of this analysis did not reveal a statistically significant difference in S&P 500 returns across inflation groups. Despite observing a spread between mean returns in low inflation periods (1.106%) and high inflation periods (0.602%), the p-value of 0.4498 and the inclusion of 0 in the 95% confidence interval for the mean difference indicate that the data does not provide sufficient evidence to reject the null hypothesis of equal means.

Additionally, missing data simulations shined a light on the challenges of real-world analysis. The study demonstrated that handling missing values using imputation or row removal can alter data distribution which in turn can impact the validity of assumptions required for statistical tests. Imputing missing values under the MCAR assumption led to a violation of the normality assumption, necessitating a shift to nonparametric testing methods. Similarly, simulating MNAR missing data during market crashes revealed the potential for bias in extreme scenarios, although the low proportion of missing values in this case minimized its impact on results.

This analysis reinforced the idea of an efficient market, showing that drastic changes in macroeconomic factors such as inflation are instantaneously built into the market. These findings suggest that inflation alone may not be a strong enough predictor for the market's returns. Future research could explore a wider array of macroeconomic indicators as well as investigating their nonlinear relationships with market returns.

Second Hypothesis

The second hypothesis claimed that variables such as commodities, yields, and categorical market descriptors are not statistically significant related to the relative return between the Russell 2000 and the S&P 500. This is mathematically equivalent to H_{0j}: $\beta_j = 0$ versus H_{1j}: $\beta_j \neq 0$ for each parameter β_j. For each variable, a parameter statistically significant from zero would reject the null hypothesis and indicate that the variable actually does have a significant relationship with the spread return. Fifteen continuous and five categorical variables were used, spanning a sample of 333 (January 1997 – October 2024).

This hypothesis is non trivial and of scientific interest for several reasons. First, it is well-documented that small caps typically outperform large caps, however it is not obvious why this is the case. Famously, one of the factors in the Fama-French three-factor model is this excess return of small-caps. Analyzing this relationship helps to provide further context to this model. Second, for a fund manager who has an outlook on how commodities or yields might behave, knowing what variables are significantly tied to the small-cap/large-cap spread would be informative. For instance, if a manager wants to bet on short-term yields rising, how should they anticipate this impacting their portfolio if they have a mix of small-cap and large-cap assets? Knowing relationships between assets is critical and helps inform decision-making.

Initially, a second-order multiple regression model was considered, which included both linear and quadratic terms. However, it was quickly discovered that none of the second-order variables were significant. This was also clear from scatter plots of each variable against the spread return. Implementing a simple multiple linear regression model resulted in five variables with statistically significant parameters: yields of corporate bonds rated CCC and lower, the market price of crude oil, and yields of the US 2-year, 5-year, and 30-year treasuries. For these particular parameters β_j, the null hypothesis was rejected.

Before obtaining these results, several assumptions were satisfied. First, the scatter plots showed the relationship between predictor variables and the response variable were all linear and not of any higher-orders. Second, some variables exhibited multicollinearity, and those that did were removed before fitting the final model. Third, the residuals were calculated and tested for normality and homoscedasticity. The Kolmogorov-Smirnov test confirmed they followed a normal distribution, while the Breusch-Pagan test confirmed they were homoscedastic.

After fitting the original multiple linear regression model, missing values were simulated and analyzed for impact. First, completely random missingness was applied to three variables at a rate of 20%. When a value was missing, it was replaced with its historical median up to that point in time. The results showed that while the 2-year, 5-year, and 30-year treasury yield variables remained statistically significant, other variables like the price of crude oil and corporate bonds rated CCC and lower no longer met the $\alpha \leq 0.05$ threshold. Next, non-ignorable or non-random missing values were simulated by considering instances where providers of market data were overwhelmed with customer demand to sell their assets. It assumed that in such cases, data was missing/unavailable and prices of assets were in volatile decline. Therefore, missing values were replaced with the historical 5th percentile of the variable as opposed to its historical median. This simulation was applied to the same three variables as before, however at a rate of only 10%. The results indicated that once again the US treasury yields were dominant in their statistical significance. It also resulted in crude oil reentering the space of statistically significant variables and corporate bonds rated CCC and lower remaining insignificant.

Overall, the results of the multiple regression test is that yields are very significant in the spread of returns between the Russell 2000 and S&P 500. The only commodity tested that proved to be significant was crude oil. Future work may include creating sophisticated models of the yield curve, which if capable of predicting yields out into the future could be used in combination with the regression model found to create a profitable trading strategy or better-inform fund managers.

7. References

1. *Breusch-Pagan Test*. The R Project for Statistical Computing. (n.d.). https://search.r-project.org/CRAN/refmans/lmtest/html/bptest.html

2. *Consumer Price Index for All Urban Consumers: All Items in U.S. City Average*. FRED. (2024, November 13). https://fred.stlouisfed.org/series/CPIAUCSL

3. *Fama-French Three-Factor Model*. Corporate Finance Institute. (2023, May 9). https://corporatefinanceinstitute.com/resources/valuation/fama-french-three-factor-model/#:~:text=The%20Fama%2DFrench%20model%20aims,%2Dto%2Dmarket%20value%20companies.

4. *Kolmogorov-Smirnov Test*. Kolmogorov-Smirnov Test - An Overview | ScienceDirect Topics. (n.d.). https://www.sciencedirect.com/topics/earth-and-planetary-sciences/kolmogorov-smirnov-test

5. Mishra, P., Pandey, C. M., Singh, U., Gupta, A., Sahu, C., & Keshri, A. (2019). *Descriptive statistics and normality tests for statistical data*. Descriptive Statistics and Normality Tests for Statistical Data. https://pmc.ncbi.nlm.nih.gov/articles/PMC6350423/#:~:text=The%20Shapiro%E2%80%93Wilk%20test%20is%20more%20appropriate%20method,data%20are%20taken%20from%20normal%20distributed%20population.

6. *Russell 2000 Sector Weightings*. Siblis Research. (2024, January 19). https://siblisresearch.com/data/russell-sector-weightings/

7. SPDR S&P 500 ETF Trust (SPY) Holdings - Yahoo Finance. (n.d.). https://finance.yahoo.com/quote/SPY/holdings/

8. Tamhane, A. C., & Dunlop, D. D. (2007). *Statistics and Data Analysis from Elementary to Intermediate*. TPB.

9. TradingView. (n.d.). https://www.tradingview.com/

10. Yahoo! (2024, December 4). *S&P 500 (^GSPC) stock price, news, Quote & History*. Yahoo! Finance. https://finance.yahoo.com/quote/%5EGSPC/