Capstone 1: Evaluating the Five-Year Performance of the S&P 500

1.0 Introduction

There are a large number of index funds (a type of investment fund based on a preset collection of stocks) in the American stock market for an investor to consider. The most notable of these indexes include the Dow Jones Industrial Average, the Standard & Poor's (S&P) 500, and the Nasdaq 100. While all three indexes can be similar, the S&P 500 contains the largest 500 companies in the United States, compared to only 30 and 100 companies in the Dow Industrial and Nasdaq, respectively (Royal, 2019).

Therefore, the S&P 500 is considered to be a diversified, valubale representation for overall market performace. The index includes companies from a wide range of sectors, as compared to other market indexes. For reference, the breakdown by sectors was as follows in 2017: information technology (24.9%), financials (14.7%), health care (13.7%), consumer discretionary (12.7 %), industrials (10.2%), consumer staples (7.7%), energy (5.7%), utilities (2.9%), materials (2.9%), real estate (2.8%), and telecom services (1.9%)(Amadeo, 2019).

In this analytical research report, the performance of the S&P 500 over the past five years will be evaluated to examine any exisitng trends, and whether or not investing in S&P 500 is an advisable strategy for investors.

1.1 Why is this data significant?

According to indsutry experts, including the renowned Warren Buffet, the S&P 500 represents a simple solution for individual investors looking for an edge (Royal, 2019). While it hasn't occured every year over the past ten, the S&P 500 generally returns approximately 10 percent a year (Amadeo, 2019). Therefore, buying and holding an index fund based on the S&P 500 is a safe and effective strategy for investors, according to many financial advisors.

Given its potential as an investing strategy for individuals, data that provides insight into the trends of S&P 500 index funds can be extremely valuable. Moreover, providing context to recent trends based on recent historical data can help investors make decisons, and understand if recent performance can be considered normal.

Lastly, the S&P 500 is regarded as a valuable indicator for overall market performance and movement. Better understanding the S&P 500 through this dataset will also help provide information for the market in general.

2.0 Exploratory Analysis

The .head() method is used to ensure the CSV file was properly read into a dataframe (df), as well as obtain a quick visual of the data frame without having to see every row.

In [1]:
```
import pandas as pd
df = pd.read_csv('cs-1.csv')
df.head()
```

Out[1]:

	date	open	high	low	close	volume	Name
0	2013-02-08	15.07	15.12	14.63	14.75	8407500	AAL
1	2013-02-11	14.89	15.01	14.26	14.46	8882000	AAL
2	2013-02-12	14.45	14.51	14.10	14.27	8126000	AAL
3	2013-02-13	14.30	14.94	14.25	14.66	10259500	AAL
4	2013-02-14	14.94	14.96	13.16	13.99	31879900	AAL

Using the .shape method, we can get a better sense of the size of the dataframe. The output shows that there are seven columns of data, which matches what was shown earlier; additionally, there are 260 data points for each column.

In [2]:
```
df.shape
```

Out[2]:
```
(619040, 7)
```

In order to check for any potential missing values, the isnull() and .sum() methods can be used. As demonstrated below, there are no missing values in this data set.

In [3]:
```
df.isnull().sum()
```

Out[3]:
```
date       0
open      11
high       8
low        8
close      0
volume     0
Name       0
dtype: int64
```

The .describe() method can be used to calculate multiple summary statistics all at once. The standard deviation, mean, maximum and minimum values may provide useful information about each of the categories above. For instance, the highest opening value for the S&P 500 over the last five years was 3024.47 USD. Evaluating the distribution of the data will provide more insight into the usefullness of other summary statistics.

In addition, the "count" row is another way to check for any missing values. As expected, there are 260 values for each column.

In [4]:
```
df.describe()
```

Out[4]:

	open	high	low	close	volume
count	619029.000000	619032.000000	619032.000000	619040.000000	6.190400e+05
mean	83.023334	83.778311	82.256096	83.043763	4.321823e+06
std	97.378769	98.207519	96.507421	97.389748	8.693610e+06
min	1.620000	1.690000	1.500000	1.590000	0.000000e+00
25%	40.220000	40.620000	39.830000	40.245000	1.070320e+06
50%	62.590000	63.150000	62.020000	62.620000	2.082094e+06
75%	94.370000	95.180000	93.540000	94.410000	4.284509e+06
max	2044.000000	2067.990000	2035.110000	2049.000000	6.182376e+08

Here, the index of the dataframe is switched to the date, in order to make the data more readable. This can be confirmed by using the .head() method, again.

In [6]:
```
df = df.set_index('date')
df.head()
```

Out[6]:

date	open	high	low	close	volume	Name
2013-02-08	15.07	15.12	14.63	14.75	8407500	AAL
2013-02-11	14.89	15.01	14.26	14.46	8882000	AAL
2013-02-12	14.45	14.51	14.10	14.27	8126000	AAL
2013-02-13	14.30	14.94	14.25	14.66	10259500	AAL
2013-02-14	14.94	14.96	13.16	13.99	31879900	AAL

A new column, "% Change", is added to the dataframe using the above code. This statistic will be useful for obtaining a metric for the change in stock value over a given week.

In [8]:
```
df['% Change'] = ((df['close'] - df['open']) / df['open']) * 100
df.head()
```

Out[8]:

date	open	high	low	close	volume	Name	% Change
2013-02-08	15.07	15.12	14.63	14.75	8407500	AAL	-2.123424
2013-02-11	14.89	15.01	14.26	14.46	8882000	AAL	-2.887844
2013-02-12	14.45	14.51	14.10	14.27	8126000	AAL	-1.245675
2013-02-13	14.30	14.94	14.25	14.66	10259500	AAL	2.517483
2013-02-14	14.94	14.96	13.16	13.99	31879900	AAL	-6.358768

3.0 Research Questions

To answer the following research questions, the python libraries below are imported.

In [9]:
```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
from scipy import stats
%matplotlib inline
```

3.1.1 Does the weekly value of the S&P 500 represent a normal distribution?

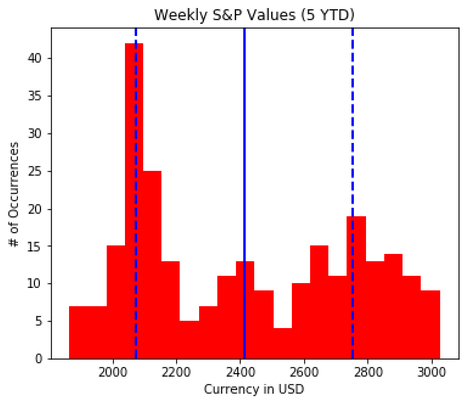
In order to evaluate the weekly value distribution as normal, a histogram is generated to reveal the overall shape of the distribution. In addition, a normal test is performed to evaluate the probability that the null hypothesis, which is that the distribution is normal, is correct.

In [8]:
```
find_mean = df['Adj Close'].mean()
find_std = df['Adj Close'].std()

#plotting the histogram
plt.figure(figsize=(6, 5))
plt.hist(df['Adj Close'], bins=20, color='r')
plt.title('Weekly S&P Values (5 YTD)')
plt.ylabel('# of Occurrences')
plt.xlabel('Currency in USD')
plt.axvline(find_mean, color='b', linestyle='solid', linewidth=2)
plt.axvline(find_mean + find_std, color='b', linestyle='dashed', linewidth=2)
plt.axvline(find_mean - find_std, color='b', linestyle='dashed', linewidth=2)
plt.show()

print('Mean = {}'.format(find_mean))

#performing normal test
print(stats.normaltest(df['Adj Close']))
```




```
Mean = 2412.0688504423083
NormaltestResult(statistic=1620.609612347861, pvalue=0.0)
```

The distribution of the weekly S&P 500 values is not normal at all; the majority of values are not centered around the mean. This is confirmed by the normal test that generates a p-value of 0.0, according to the normal test from the SciPy library. A possible explanation for the non-normal distribution is that the value of the S&P 500 is not an independent variable in this scenario. The value is likely to be closer to its value of the previous week.

3.1.2 Does the weekly percent change of the S&P 500 represent a normal distribution?

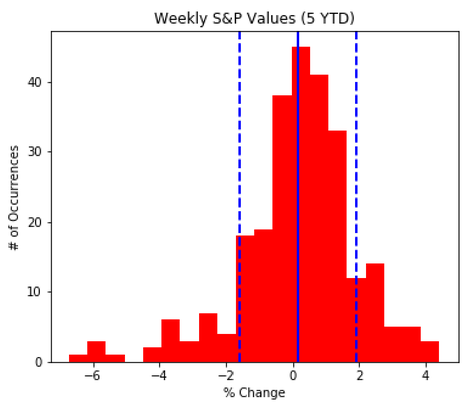
The weekly percent change in value for the S&P 500 is more likely to show a normal distribution because the percent change in weekly values does represent an independent variable. In order for a distribution to be considered normal, the data must follow the "empirical rule" (defined below). Again, this can be evaluated by creating a histogram of the data set and executing a normal test.

```
-The "empirical rule": For a variable to be classified as normally-distributed, the following criteria must be met: ~68% of values are within 1 standard deviation of the mean, 95% of values are within two standard deviations of the mean, and 9
9.7% of values are within 3 standard deviations of the mean
```

In [9]:
```
find_mean = df['% Change'].mean()
find_std = df['% Change'].std()

#plotting the histogram
plt.figure(figsize=(6, 5))
plt.hist(df['% Change'], bins=20, color='r')
plt.title('S&P Values (5 YTD)')
plt.ylabel('# of Occurrences')
plt.xlabel('% Change')
plt.axvline(find_mean, color='b', linestyle='solid', linewidth=2)
plt.axvline(find_mean + find_std, color='b', linestyle='dashed', linewidth=2)
plt.axvline(find_mean - find_std, color='b', linestyle='dashed', linewidth=2)
plt.show()

#Mean and Std Dev Calculations
print('Mean = {}'.format(find_mean))
print('Standard deviation= {}'.format(find_std))
```




```
Mean = 0.15277656019636715
Standard deviation= 1.7589542324484895
```

Based on the shape of the histogram, the '% Change' distribution can be considered normal. The majority of values are centered around the mean, and the values form a bell curve. Given that the shape of the histogram is approximately normal, summary statistics (i.e. mean) hold meaningful value.

**See Appenix A for the results of the normal test using the SciPy library.*

3.2 How has the value of the S&P 500 changed over the past five years?

The answer to this question can be graphed by plotting the weekly value of the S&P 500 over five years. In addition, the total percentage change in value from the beginning of the five-year period until the final closing value can be calculated to provide an overall result.

In [12]:
```
#Creating a plot
plt.figure(figsize=(20, 8))
plt.plot(df['close'], color = 'g')
plt.tick_params(axis ='both', which = 'both', length = 10)
plt.tick_params(axis ='x', rotation = 90)
plt.xticks(np.arange(0, 260, 12))
plt.title('S&P 500 five-year growth')
plt.ylabel('Currency in USD')
plt.show()

#Calculating overall percent change over five years
start = df.open[0]
end = df['close'][-1]
percent_change_ovr = start / end * 100
print('5 Year Change: {}'.format(percent_change_ovr))
```


```
5 Year Change: 20.403466016788517
```

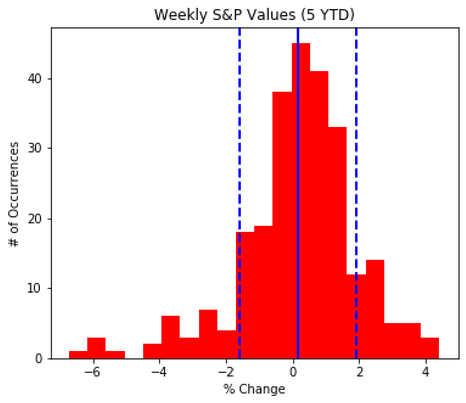
As seen in the plot above, the value of the stock has clearly risen over this time period, despite periods of volatility. That equates to a +20.40% increase over 5 years.

3.3 Based on this data set, how does the S&P 500 perform over an average week?

By showing the percentage change in value for the stock every week, the same histogram used earlier helps to visualize the range that can be expected of the S&P 500, based on the past five years. Calculating the mean will help provide a statistical answer to the question.

In [11]:
```
find_mean = df['% Change'].mean()
find_std = df['% Change'].std()

#plotting histogram
plt.figure(figsize=(6, 5))
plt.hist(df['% Change'], bins=20, color='r')
plt.title('Weekly S&P Values (5 YTD)')
plt.ylabel('# of Occurrences')
plt.xlabel('% Change')
plt.axvline(find_mean, color='b', linestyle='solid', linewidth=2)
plt.axvline(find_mean + find_std, color='b', linestyle='dashed', linewidth=2)
plt.axvline(find_mean - find_std, color='b', linestyle='dashed', linewidth=2)
plt.show()
```



Using the .desribe() method for percent change will also prove useful in calculating summary statistics that better summarize the expected range of values for an average week.

In [12]:
```
df['% Change'].describe()
```

Out[12]:
```
count    260.000000
mean       0.152777
std        1.758954
min       -6.721215
25%       -0.576872
50%        0.268516
75%        1.188246
max        4.426994
Name: % Change, dtype: float64
```

Based visually on the histogram, the value of the S&P 500 generally has a percent change between -2 and +2.5%, and has a net postitive percent change more often than not, with a mean value of 0.15. This means that during average week, the stock increased by 0.15%. The quartile ranges suggest that 50% of the time, the percent change for the week will fall between -0.58% and +1.19%.

3.4 How does the S&P 500 perform on a year to year basis?

In [13]:
```
#creating seperate dataframes
df_five_ydt = df.iloc[0:52, 4:5]
df_four_ydt = df.iloc[52:104, 4:5]
df_three_ydt = df.iloc[104:156, 4:5]
df_two_ydt = df.iloc[156:208, 4:5]
df_one_ydt = df.iloc[208:260, 4:5]
```

In order to answer this question, the code above is used to create a seperate dataframe for each year, and the code below is used to graph each year for an indivual visual analysis.

3.5 How has the S&P 500 performed over the past month, compared to the year as a whole (in terms of volatility and percent change in value)?

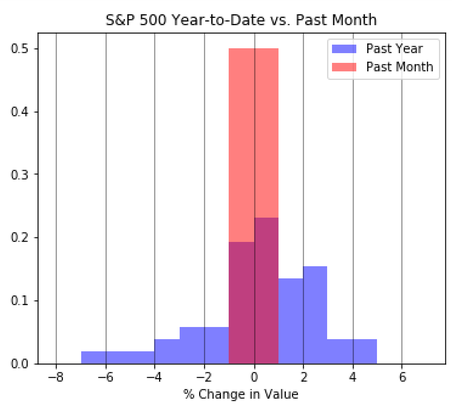
To answer this question, a histogram overlapping the past year with a histrogram of the past month can be used. In addition, a t-test can be performed to evaluate whether or not the difference between the two time periods is significant. The null hypothesis is that the past month and the past year share have performed the same.

In [15]:
```
#Plotting the two histograms
df_one_ydt = np.array(df.iloc[208:260, 6:7])
df_past_mo = np.array(df.iloc[256:260, 6:7])
plt.figure(figsize=(6,5))

plt.hist(df_one_ydt, density=True, color = 'blue', bins=np.arange(-8, 8), alpha = .5, label= "Past Year")
plt.hist(df_past_mo, density=True, color = 'red', bins=np.arange(-8, 8), alpha = .5, label= "Past Month")
plt.title('S&P 500 Year-to-Date vs. Past Month')
plt.xlabel('% Change in Value')
plt.legend(loc='best')

plt.grid(axis = 'x', c='black', alpha=0.5)
plt.show()

#T-test
import scipy
from scipy.stats import ttest_ind
print(ttest_ind(df_one_ydt, df_past_mo, equal_var=True))
```




```
Ttest_indResult(statistic=array([0.10471053]), pvalue=array([0.91699347]))
```

The histogram above shows the volatility and percent change of the S&P 500 during the past month (red), versus the rest of the past year (blue). From this visual representaiton, it is clear that this past month has experienced less volatility than the rest of year, to this date.

The results of the t-test are less rigorous than the visual data. Because the p-value is .92, the null hypothesis fails to be rejected (.92 > .05; if alpha = .05). While visually there is less volatility over the past month in the histogram, the t-test from the SciPy library does not indicate statistical significance. Meaning, the t-test failed to prove a difference between the two time periods exists. In addition, any difference in the mean for percent change cannot be considered statistically significant, either.

4.0 Conclusion

In this analytical report, the performance of the S&P 500 over the past five years was evaluated using visual and statistical methods. The distribution of the raw data (S&P 500 values) was proven not to be normal. In contrast, the visual data from the figure in Section 3.1.2 suggests the distribution for '% Change' in the S&P 500 weekly values is normal. It was graphically determined that the value of the S&P 500 stock has risen greatly over the past five years. The discussed strategy from the introduction, buying and holding an index fund based on the S&P 500, did therefore prove to be a an effective strategy for investors (there was a 65.73% increase in value over the five-year period). When the average week for S&P 500 was analyzed, it was found that the average percent change in stock value was +0.15%. In addition, the average percent change for a week appeared to be between plus/minus two percent, based on the histogram. Furthermore, the visual data from the figure in Section 3.4 helped identify a potential trend. The plots in the figure indicate a tendency for the value of the S&P 500 to drop from the last quarter to the first quarter of the calendar year. Lastly, the histogram in Section 3.5 visually indicated that the past month has been less volatile than the rest of the year. However, this finding was not found to be statistically signficant, using a t-test from the SciPy library.

4.1 Further Research

After analyzing the performance of the S&P 500 over the past five years, there are several areas of research that would be worth exploring further.

As mentioned in the introduction, the S&P 500 is one of the most important indexes in the American stock market. Because it contains 500 of the largest US companies, it is considered to be one of the best indicators of overall market performance. However, both the Dow Jones Industrial Average and the Nasdaq 100 are also very popular indexes, much in the same way. It would be worth exploring how these indexes compare to the S&P 500 in terms of long-term performance and market stability. Which of these three indexes represents the best choice for investors looking for long-term gains? All three of those indexes are comprised of large-cap companies, how would an index comprised of small-cap companies (i.e. Russell 2000 index) differ (Chen, 2019)? The answers to these questions would help guide long-term investing strategies.

Another area worth exploring is the trend for a regression in the S&P 500's value over the last and first quarters of the calendar year. Is this occurence more specific to the past five years, or is it representative of a larger trend? Is there a time period of regression that is statistically significant? Given the discussed significance of the S&P 500, the answer to these questions has large implications for investors and the American economy as a whole.

5.0 References

Amadeo, K. (2019, June 25). What the S&P 500 Tells You About America's Health. Retrieved November 14, 2019, from https://www.thebalance.com/what-is-the-sandp-500-3305888 (https://www.thebalance.com/what-is-the-sandp-500-3305888).

Chen, J. (2019, November 18). Russell 2000 Index Definition. Retrieved November 15, 2019, from https://www.investopedia.com/terms/r/russell2000.asp (https://www.investopedia.com/terms/r/russell2000.asp).

Kenton, W. (2019, October 23). S&P 500 Index – Standard & Poor's 500 Index Definition. Retrieved November 15, 2019, from https://www.investopedia.com/terms/s/sp500.asp (https://www.investopedia.com/terms/s/sp500.asp).

Royal, J. (2019, October 22). How To Buy An S&P 500 Index Fund. Retrieved November 14, 2019, from https://www.bankrate.com/investing/how-to-buy-sp-500-index-fund/ (https://www.bankrate.com/investing/how-to-buy-sp-500-index-fund/).

S&P 500 (^GSPC) Historical Data. (2019, October 18). Retrieved October 18, 2019, from https://finance.yahoo.com/quote/^GSPC/history?period1=1541998800&period2=1573534800&interval=1wk&filter=history&frequency=1wk (https://finance.yahoo.com/quote/%5EGSPC/history?period1=1541998800&period2=1573534800&interval=1wk&filter=history&frequency=1wk).

Appendix A

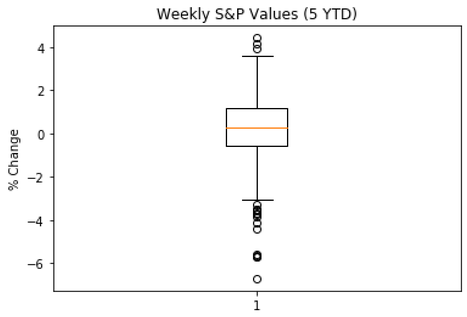
The results of the SciPy normal test mentioned in Section 3.1.2 are below:

In [16]:
```
#Normal Test
print(stats.normaltest(df['% Change']))
k2, p_value = stats.normaltest(df['% Change'])
alpha = 1e-3
print("p_value = {:g}".format(p_value))
if p_value < alpha:  # null hypothesis: x comes from a normal distribution
     print("The null hypothesis can be rejected (reject H0)")
else:
    print("The null hypothesis cannot be rejected (fail to reject H0)")
```

```
NormaltestResult(statistic=39.422758532946624, pvalue=2.7507856738971925e-09)
p_value = 2.75079e-09
The null hypothesis can be rejected (reject H0)
```

According to the SciPy normal test, the p-value is less than alpha (.001). Based on this specfic SciPy test with an aplha value of .001, this dirstibution is not normal enough (null hypothesis is rejected). However, this result contradicts the normal distribution visually displayed in the figure from 3.1.2. It is possible that using a different library, or processing the data set differently (i.e. taking the log of the data) would achieve more normal results. In addition, a boxplot may prove useful in visualizing the impact of potential outliers on the distribution and resulting normal test.

In [17]:
```
#Generating a boxplot
plt.boxplot(df['% Change'])
plt.title('Weekly S&P Values (5 YTD)')
plt.ylabel('% Change')
plt.show()
```



The boxplot above shows potential outliers, a possible explanation for the failed normal test using the SciPy library. Further investigation beyond the scope of this report would be required to make a determination.