Zihao Li, Jinghao Yan, Brody Vogel, and Ye Zhang

ANLY 501

Project 2

Sports and the Stock Market: What Do They Have to Do with Each Other?

Introduction

The data science question examined here is: *Does the U.S. stock market react to professional sports leagues and the performance of their teams, and, if it does, how?* American professional sports teams both spend and earn incredible amounts of money. As such, it seems reasonable to think that such a huge industry would affect the U.S. stock market. And, from a less intuitive perspective, professional sports' ubiquitous popularity could mean that the market is affected by the performance of certain teams, too. In the first part of the project, two sets of data were gathered and cleaned to start exploring this question: the open, high, low, and closing price of various stocks over the last 42 years, and the performance--game by game--of every sports team in the NFL, MLB, and NBA over that same span. In this part, the data was examined to see what--if any--relationships exist. To accomplish this, new variables were created that better suit the project goals; some basic statistical analyses and plotting were performed; the stocks were clustered in various ways to look for trends; the Apriori algorithm was used to mine association rules between the datasets; and finally, three hypotheses were tested regarding relationships between the sports and stock data.

New Variable Creation

To get started, a new set of variables was created in each dataset. There had been a lot of useful data accumulated, but it wasn't perfectly-suited for the analysis. To remedy this, three new variable types were built--one for the sports data and two for the stocks. For the sports data, pseudo-binned variables were created to classify the result of each game that every team played as -1 (a loss), 0 (no game), or 1 (a win). With the stock data, the first of the two new variable types is a normalization technique; a variable was added to each stock that tracks the daily percentage change in the closing price from the day before's. This will ensure that a price change of, say, $550 in the NASDAQ index isn't considered the same as an equal change in a much smaller stock, like the Vanguard Utilities Index. The second variable type added to the stock data uses percentiles to bin the just-mentioned percentage changes into five groups based on their size relative to other changes in that stock: BIG JUMP (>95% of the recorded percentage changes for that stock), JUMP (85%<X <95%), LITTLE MOVEMENT (35%<X<80%), DIP (5%<X<35%), and BIG DIP (<5%). So, for example, a daily change of +$30 in the NASDAQ Index price would fall between the 35th and 80th percentile, and so would be added to the "NASDAQ Change Category" column as "LITTLE MOVEMENT". The thinking behind binning these variables is that they're the type that are most important to the data science question being explored, and so it was very useful to have categorical bins for the association rule mining and some of the hypothesis testing that will be discussed later in this write up. With these three new variable types, some exploratory statistical analyses could be undertaken.

Beginning with Statistical Analyses

Next, to get a preliminary sense for the data, some basic statistical analyses were performed. Before this could begin, though, the data cleaning decisions that were made in the first project still had to make sense. Luckily, the raw data was very clean, and so there weren't any difficult decisions to be made. During Project 1, the only problems with the stock data were missing values from days when the market was closed, and a tricky formatting where dividend payouts were listed as days in the scraped data. These were thrown out; days when the market was closed (because there's no change to track) or dividend payouts were not useful for the analyses performed here. With the sports data, there were two easily-solved problems: ties and teams that moved cities. Games that ended in a tie were scrubbed because there were relatively few of them, and because they aren't related to the data science question under examination. For the teams that have moved, the columns were merged; For example, the Baltimore Colts and the Indianapolis Colts are now treated as the same team. This merging was the most difficult choice, but it is the best way to handle the problem; for one thing, there aren't many teams that have changed location, and, for another, a team's *performance* is more important to the analyses than its *location.* Thus, the cleaning decisions made in Project 1 still made sense here, and so it is safe to being with some exploratory statistical analyses.

For the most part, these analyses were done on the stocks data, since the sports data isn't conducive to such tests; the sports data is much more important to the association analysis and hypothesis testing below, since it mostly consists of binary "Outcomes" variables. With respect to the stocks data, then, each test was performed on the newly-built percentage change variables. These are the best variables to test because, ultimately, the changes in the market--which is all that is of interest here--are entirely contained in these new variables; for example, if the Dow Jones Index had an extreme daily high that was 20 percent higher than the day before's, or if it opened much higher than expected, or had a deep daily low, etc., these would be accounted for in the percentage change. These tests will hopefully indicate that the stock data is suitable for comparison with the sports data, and so provide the confidence necessary to move forward with association rule mining and hypothesis testing. The tests performed on the percentage change variables, accordingly, were: 1) mean, median, and standard deviation calculation, 2) outlier detection and handling, 3) frequency measurements, and 4) correlation calculation.

Mean, Median, and Standard Deviation

First, the results from the calculations of the mean, median, and standard deviation of each of the seventeen percentage change variables corresponding to the seventeen tracked stocks are below:

STOCK	**Mean (Percentage Change from Day Before)**	**Median**	**Standard Deviation**
NASDAQ Index	.00051	.00111	.01258
Dow Jones Index	.00041	.00055	.01097
S & P Index	.00039	.0005	.01068
Marathon Oil	.00031	0	.02222
Chevron Oil	.00045	0	.02229
Exxon Oil	.00044	0	.01445
Franklin Gold and Precious Metals Index	.00026	0	.01895
First Eagle Gold Fund	.00023	0	.01700
Sturm and Ruger Guns	.00073	0	.02641
American Outdoor Guns	.00197	0	.05445
Anheuser-Busch	.00066	.00027	.01456
VICEX (Sin Index)	.00035	.00070	.00986
RYDEX (Sin Index)	.00029	.00065	.01427
Utilities Index	.00019	0	.00872
Healthcare Index	.00038	.00069	.01033
IT Index	.00041	.00103	.01277

Treasury Bonds Index	0	0	.00306

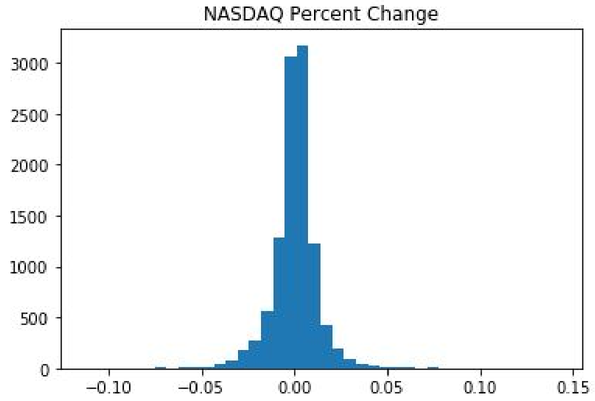
These metrics explain little more than that the data appears to behave the way it should, which is nice to see. Treasury Bonds excluded, the means and medians show that each stock averaged a tiny daily rise, which mirrors exactly the stock market as a whole since 1975. And, as would be expected, the stocks from sectors that are a bit more volatile (like oil funds and stocks for single companies like Sturm & Ruger or American Outdoor) have larger standard deviations than more stable funds (like Treasury Bonds). The mean, median, and standard deviation of the percentage change variables, then, indicated that it was time to move on to some more robust statistical tests under the assumption that there is nothing terribly off about the stocks data.

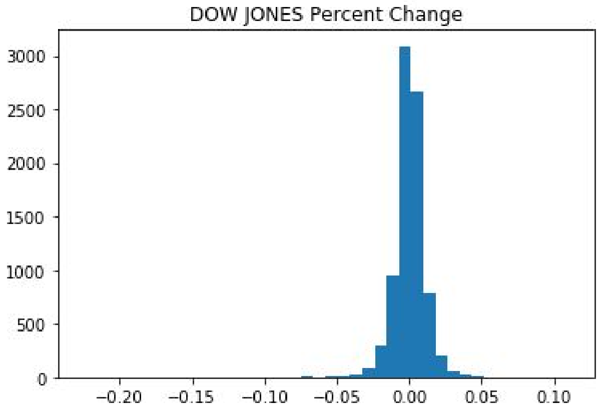
Outlier Detection and Handling

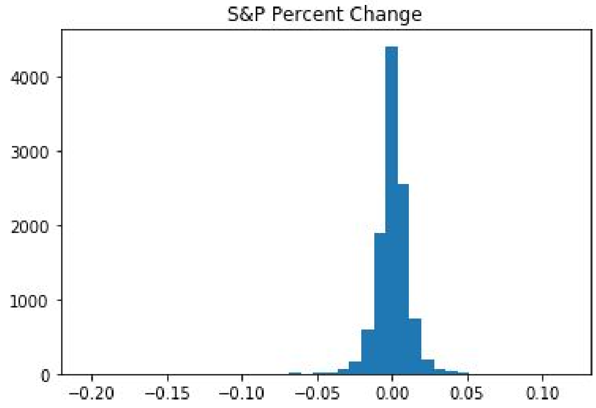
In Project1, each variable in the stock data was checked for outliers. No matter the method, it produced a high number of outliers, which makes sense because the stock market is incredibly volatile. No values were thrown out, though, because part of what's exciting about the stock market is its extreme values and the idea that they can be predicted. This time, any outliers in the seventeen new percentage change variables were points of interest. To find outliers, a modified Interquartile Range test was performed for each of these variables. Traditionally, an outlier is considered to be any value that is more than 1.5 times the interquartile range below the 25th percentile or above the 75th percentile. With data that fluctuates as much as that from the stock market, though, that process counts far too many data points as outliers (it counted 33,976 outliers). Thus, an outlier was counted as any price change that was *5* times the interquartile range below the 25th percentile or above the 75th percentile. This still found 348 outliers across the seventeen percentage change variables. After further investigation, they appeared to correspond to days when the market was especially turbulent (like Black Monday). There was no reason to get rid of this kind of data. If anything, the relatively large amount of extreme values suggests that there are trends to be discovered behind those large swings across the stock market.

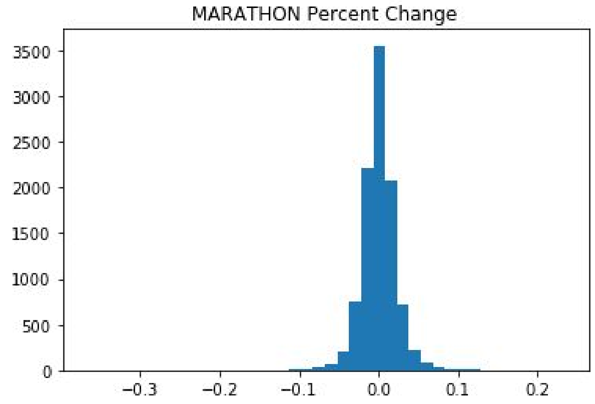
Frequency Measurements

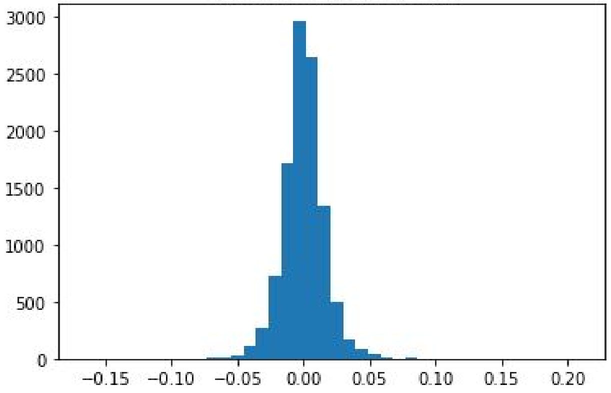
Next, for each daily stock price percentage change variable, a histogram was created to chart the frequency of swing sizes in that stock. This would shed more light on the percentage change variables than just finding the mean, median, and standard deviation could; for one thing, histograms give a better picture of the true spread and distribution of each variable. The histograms list the frequency in number of days on the y-axis and percentage change in the stock's closing price on the x-axis. These are below:

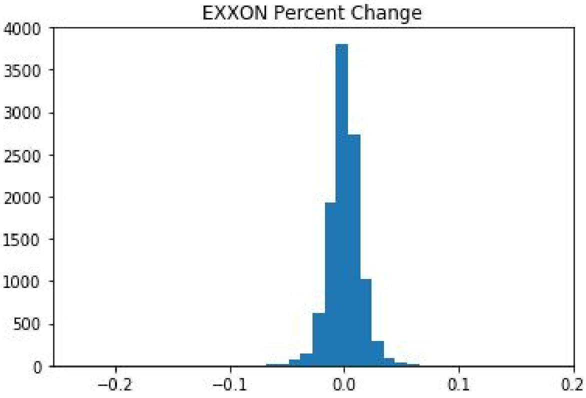


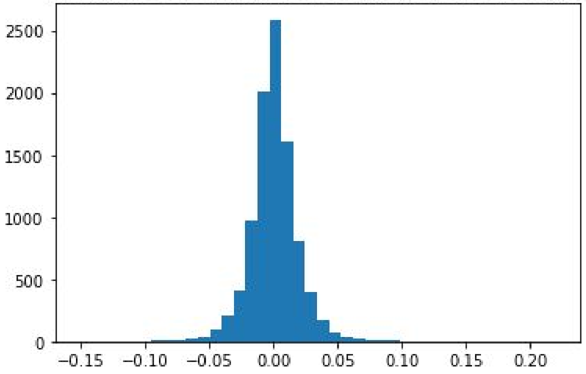


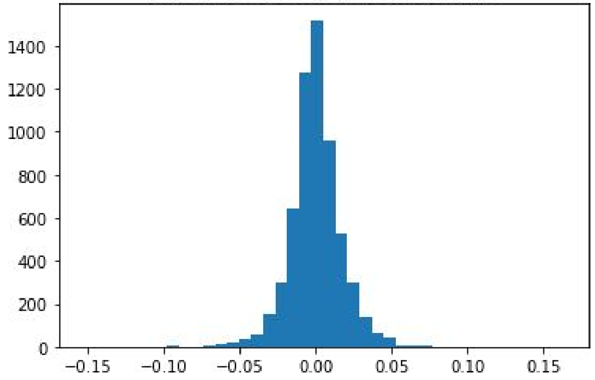


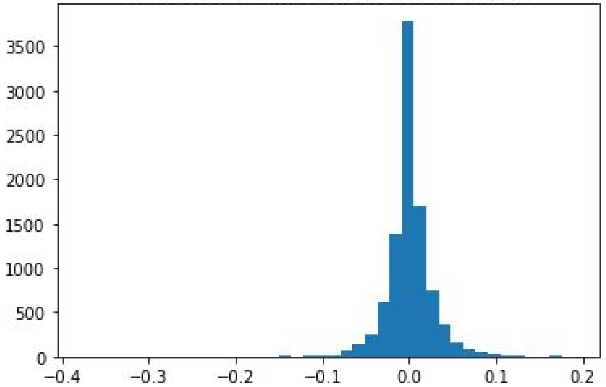


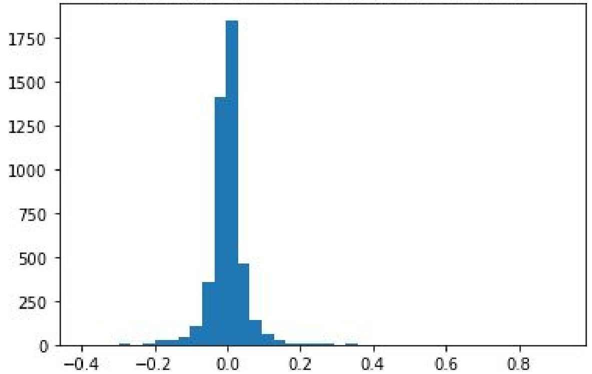

CHEVRON Percent Change

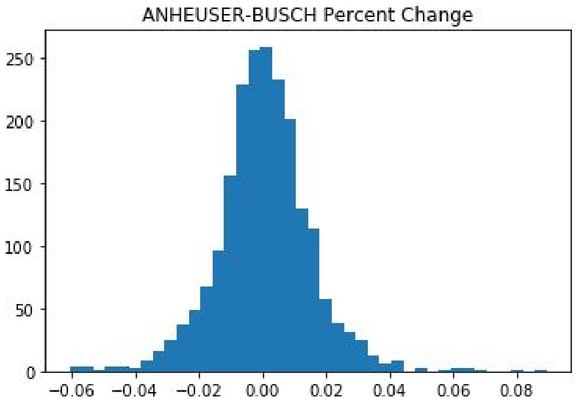

EXXON Percent Change


FRANKLIN GOLD AND_PRECIOUS METALS Percent Change

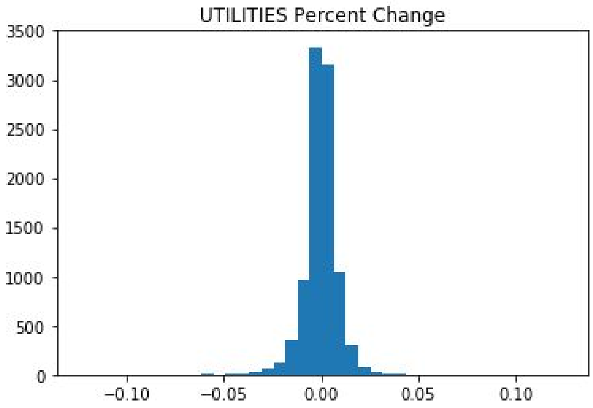

FIRST EAGLE GOLD Percent Change


STURM & RUGER (GUNS) Percent Change

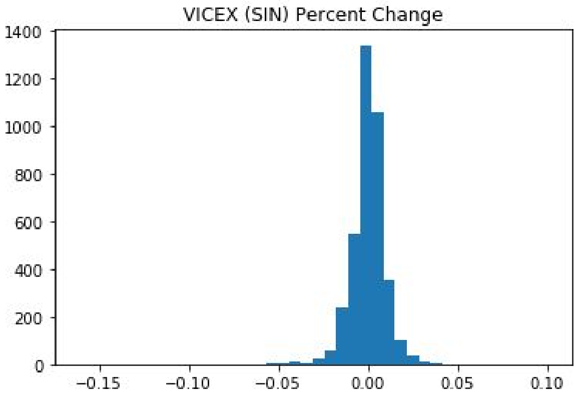

AMERICAN OUTDOOR (GUNS) Percent Change



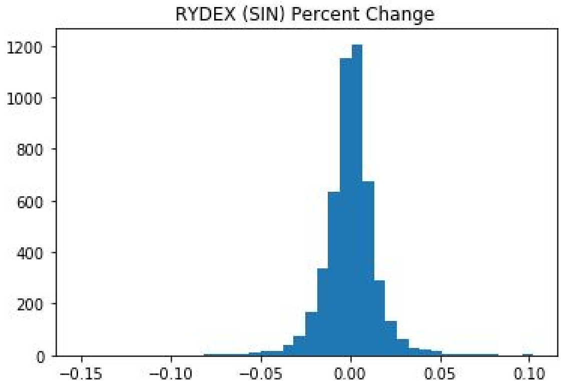
ANHEUSER-BUSCH Percent Change



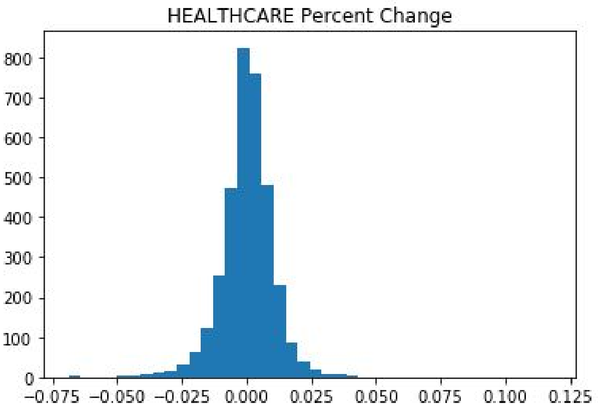
UTILITIES Percent Change



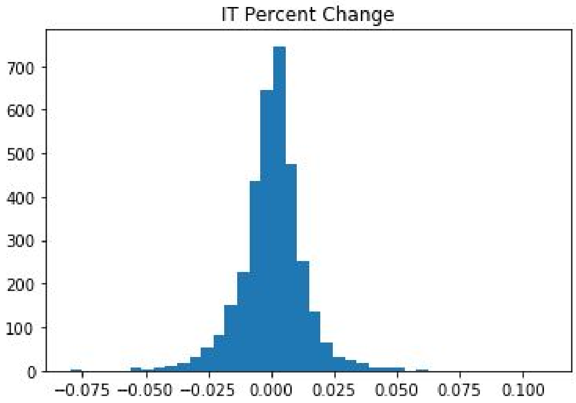
VICEX (SIN) Percent Change



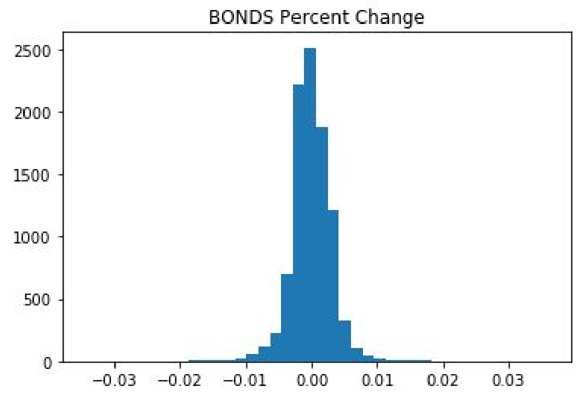
RYDEX (SIN) Percent Change



HEALTHCARE Percent Change



IT Percent Change



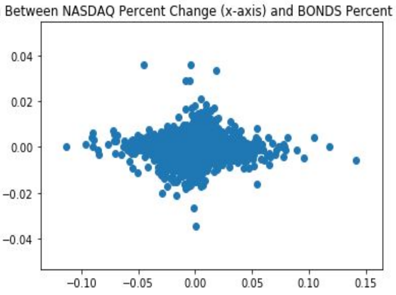
The plots above show that each percentage price change variable is roughly normally-distributed around a value very close to 0. This is exactly what was hoped for. If any of the stocks had appeared to cluster around, for instance, .02, that data would not have been usable because that stock would too often show positive correlation; for example, if a stock averaged a .02% daily rise for four decades, there would be a strong correlation between its rising and positive outcomes for any team that performed particularly well over that same timespan. The only plots that cause any of this type of concern are those for the Treasury Bond Index and Sturm & Ruger (Guns). For the former, a slightly higher than average amount of its daily percentage changes appear to be below 0; this could mean there was a high number of dips in the Bond Index. And, for the latter, it looks like an inordinately-high number of days saw a percentage change just below 0; this indicates that Sturm & Ruger (Guns) may have had a rough decade that skewed its 42-year performance. These will be things to watch out for, but, thankfully, each stock's histogram appears to show a normal distribution centered around a number very close to 0, and so, just as before, it can be safely assumed that none of the results will be horribly skewed by a stock that out- or underperformed its peers over the last 42 years.

Correlation Calculations

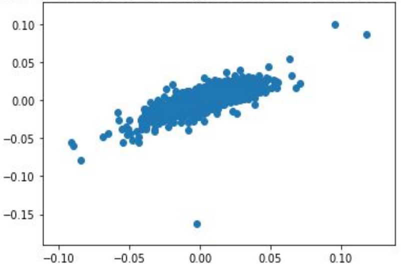
To round out the basic statistical analyses, the correlations between the percentage change variables for what seemed like five disparate stocks were calculated: the NASDAQ Index, VICEX (Sin) Index, Vanguard Utilities Index, American Outdoor (Guns), and the Treasury Bonds Index. This was performed to confirm that the data isn't affected by trends that took over the whole stock market. For instance, an event like the bursting of the tech bubble, which pulled the entire market down by over 15 percent, would skew the results. Since the stock market has consistently gone up over the last 42 years, a broad positive correlation--but nothing too significant--was expected. This expectation was confirmed by low correlation numbers, as shown in the table of correlation coefficients and scatter plots below:

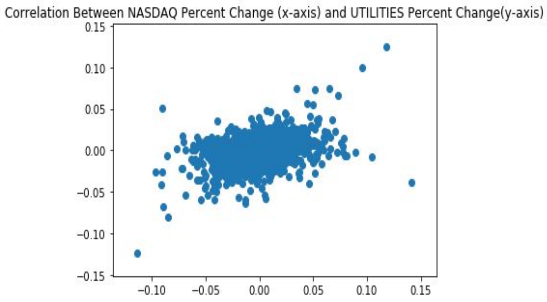
STOCK	NASDAQ	VICEX	Utilities	American Outdoor	Treasury Bonds
NASDAQ	1	.382	.460	.641	-.0047
VICEX	.382	1	.242	.371	-.0147
Utilities	.460	.242	1	.244	.038
American Outdoor	.641	.371	.244	1	-.025
Treasury Bonds	-.0047	-.0147	.038	-.025	1

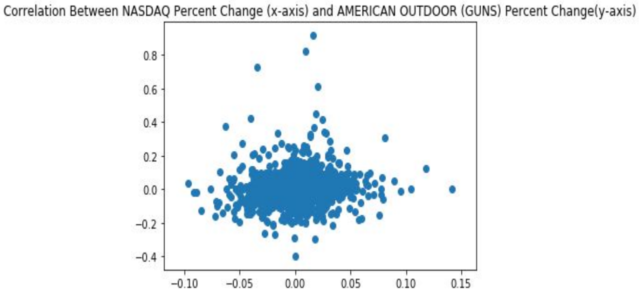
Correlation Between NASDAQ Percent Change (x-axis) and BONDS Percent Change(y-axis)

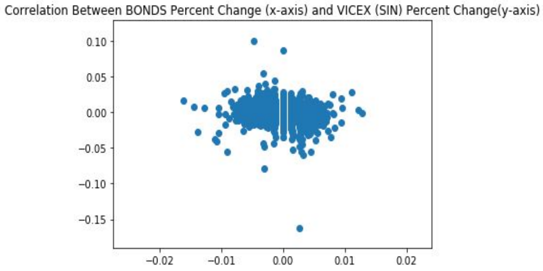


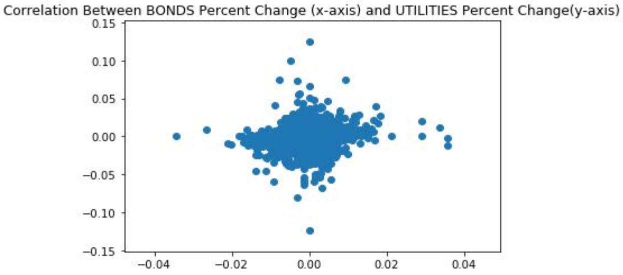
Correlation Between NASDAQ Percent Change (x-axis) and VICEX (SIN) Percent Change(y-axis)

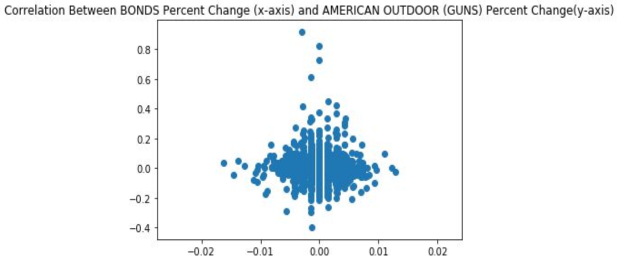


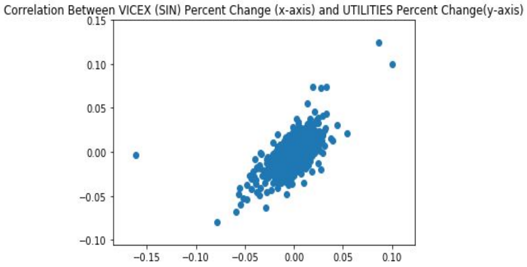


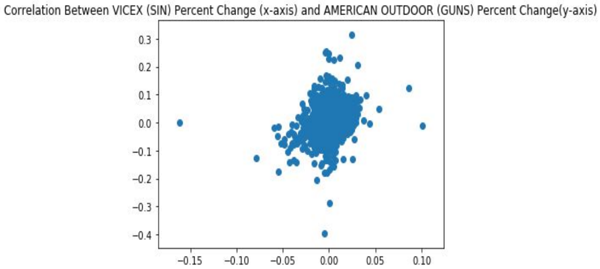


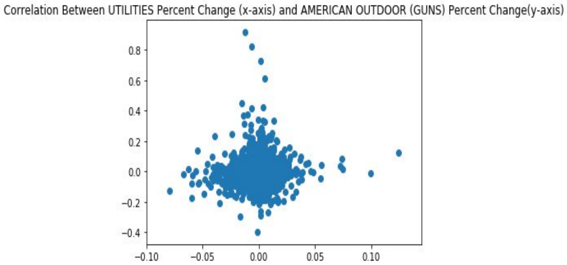












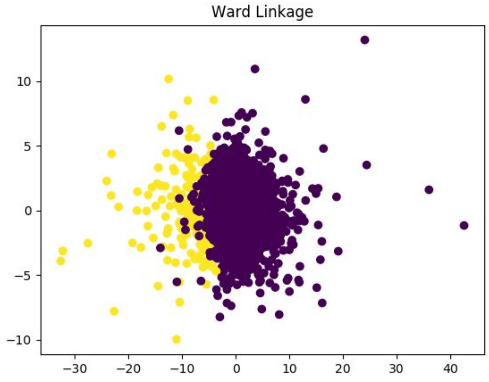
The NASDAQ Index has semi-alarming correlations with American Outdoor (Guns) and the Utilities Index, but neither is large enough (.641 and .460) to cause real concern; and, considering the NASDAQ is an total-market index fund, it's unsurprising to find it somewhat correlates with individual stocks, since the market as a whole rose over the period we're testing. Based on these results, additional analyses can be performed under the assumption that none of the percentage change in stock price variables are highly-correlated with any other, and so shouldn't undercut any relationships with the sports data.

The basic statistical analyses, therefore, suggest that the stock data is suitable for comparison with the sports data. Before starting to look for possible relationships between the two, though, the stock data was put through various clustering algorithms to look for any so-far-undiscovered trends or similarities.

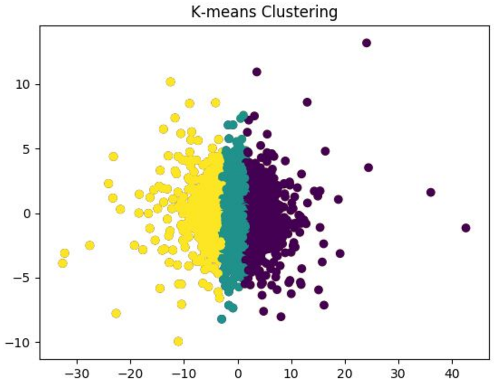
Clustering Analysis

Three cluster analyses was conducted on the stocks data. Using these three different clustering methods, the overall percent change columns for the seventeen different stocks were looked at to determine what--if any--clusters would present themselves as interesting. Since some stocks don't have data for the percentage change columns for certain days (because those stocks did not yet exist), the null values were replaced with the mean of non-null percentage change values for each respective stock. These were then compiled into a copy of the stocks dataframe for clustering analysis.

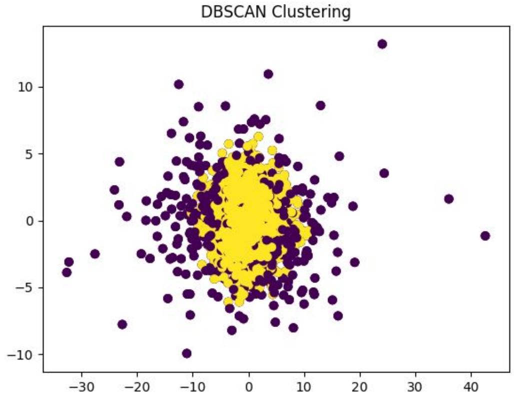
Using hierarchical clustering with Ward linkage, and plotting a PCA projection of the clusters, there appears to be 2 clusters, as shown below, for the percentage change data. This also gives the best Silhouette average score of 0.465, which is a decent indicator of the clustering performance.



Next, k-means clustering was performed, which produced the best Silhouette average score with 3 clusters: - 0.2823. This indicates that k-means may not be the best method for clustering the data; one possible explanation is that the points were just too close together to create meaningful clusters. Another reasonable explanation could be that the sizes and shapes of the clusters differ too much to produce any interesting results.



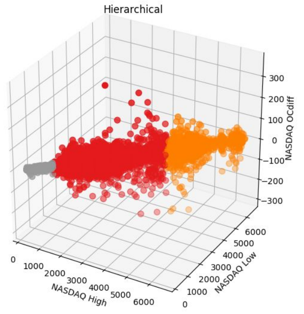
Finally, DBSCAN clustering was conducted, which works better with arbitrarily-shaped clusters. With an eps of 3.5, DBSCAN produced a Silhouette average score of 0.5571.

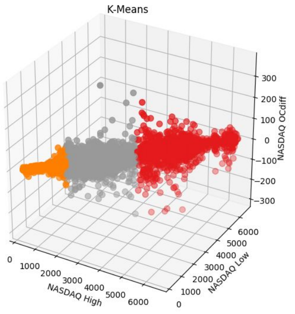


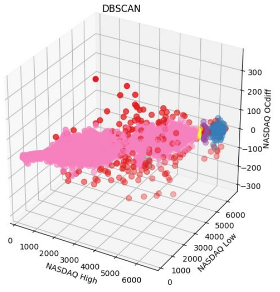
From the different clustering algorithms, the seventeen percentage change columns seem to be best grouped by two clusters. Digging a bit deeper into the plots above, it looks like the data groups into clusters where there was little change in the market and days when there were either dips or jumps. This might be an indicator that the association rule mining to follow will produce nice results.

In addition, one of the individual stocks--the NASDAQ Index--was looked at in more depth to see if its analysis would mirror that in the previous categories. The original NASDAQ data contains: “NASDAQ Open”, “NASDAQ Close”, “NASDAQ High”, “NASDAQ Low” and “NASDAQ Volume”. This high dimensionality could reduce the accuracy of the clustering, so the number of variables used was cut down. Since the value of the stock volume variable is generally greater than the others, it would largely affect the result of clustering in terms of skewness, and so it was dropped from the analysis, along with the stock’s open and close. A new variable was also created: “NASDAQ OCdiff”, which is the difference between the opening price and closing price on a given day. In sum, then, the “NASDAQ Low”, “NASDAQ High” and “NASDAQ OCdiff” were the three final attributes used for clustering.

Next, the same methods outlined above in the percentage change clustering were conducted with several values of parameters for fitting the models. For k-means clustering, k = 3, or a predetermined number of clusters of three, was used. For DBSCAN, eps = 100 and min_samples = 30 were set for the best fit. And finally, for hierarchical clustering with ward linkage, n=3 was the chosen parameter. Through each method, the clustering results were generated in 3D for better visualization, as shown below.







K-means with three clusters and Hierarchical clustering produced fairly nice accuracy scores of .6693 and .5899, respectively; DBSCAN only produced a score of 0.41175. Since this tracked one broad index stock, it may be the case that it was best suited by three clusters: small changes in the market, moderate changes, and large changes. The graphs above appear to agree with this idea. Another thing to note is that, unlike the clustering performed on the percent change columns, k-means clustering performed best here. One possible explanation is that there is less shape or density variation in a single stock or index, such as the NASDAQ, compared to the other stocks tested. This could also produce something of interest in the association rule mining discussed in the next section.

Association Rule Mining

At this stage, it made sense to look at the relationships between teams' performance and the stock market. To best do this, the newly-created variables described earlier were examined: W/L variables for the teams, and the daily percentage change from the stocks; as has been mentioned, these are most closely-related to the data science question of interest. The Apriori algorithm was run on test data consisting of three large-market sports teams and a representative stock from each sector: 1) the Dallas Cowboys, New York Yankees, and Los Angeles Lakers, and 2) the NASDAQ Index, Exxon Oil, Franklin Gold and Precious Metals, Sturm & Ruger (Guns), Anheuser-Busch, VICEX (Sin), Utilities Index, Healthcare Index, IT Index, and the Treasury Bonds Index. For each team, there were two possible values for every data entry: 1 (the team won) or -1 (lost)--days when the teams didn't have a game were not considered. And, for each stock, there were five possible values for each data entry: BIG JUMP, JUMP, LITTLE MOVEMENT, DIP, and BIG DIP. To be more specific, the Apriori algorithm was run on three datasets--one for each team. So, for example, the New York Yankees dataset consisted of every day on which the Yankees played a game between 1972 and 2017 and the way the seven test stocks reacted the following day. This means there were 2 * 7 * 5 = 70 possible association rules for each run through the Apriori algorithm (70 * 3 = 210 total), none of which produced a particularly-high support or confidence value. The results were so poor, in fact, that the use of very low minimum support values was necessary to complete the tests. Setting the minimum support to .2 trimmed the 210 potential association rules down to 40; .25 to 15 rules; and .28 to 2 rules. A summary of these results follows in the table below (all rules with support > .25 are listed).

Event A	**Event B**	**Support**	**Confidence**
Dallas Cowboys WIN	Exxon LITTLE	.2762	.6210

	MOVEMENT		
Dallas Cowboys WIN	Sturm & Ruger (Guns) LITTLE MOVEMENT	.2754	.6151
Dallas Cowboys WIN	Anheuser-Busch LITTLE MOVEMENT	.2593	.4795
Dallas Cowboys WIN	VICEX (SIN) LITTLE MOVEMENT	.2720	.5812
Dallas Cowboys WIN	BONDS LITTLE MOVEMENT	.2609	.4751
New York Yankees WIN	Exxon LITTLE MOVEMENT	.2578	.4570
New York Yankees WIN	Franklin Gold and Precious Metals LITTLE MOVEMENT	.2559	.4563
New York Yankees WIN	Sturm & Ruger (Guns) LITTLE MOVEMENT	.2511	.5662
New York Yankees WIN	IT LITTLE MOVEMENT	.2549	.4469
LA Lakers WIN	NASDAQ LITTLE MOVEMENT	.2606	.6006
LA Lakers WIN	Exxon LITTLE MOVEMENT	.2647	.4338
LA Lakers WIN	Franklin Gold and Precious Metals LITTLE MOVEMENT	.2799	.4525
LA Lakers WIN	Sturm & Ruger (Guns) Little Movement	.2974	.6383
LA Lakers WIN	Utilities LITTLE MOVEMENT	.2759	.6150
LA Lakers WIN	BONDS LITTLE MOVEMENT	.2819	.45582

To put this in context, the most promising association rule was between the Lakers winning and the Bonds Index seeing little movement, with a support value of .2819. Relative to the way things were set up, this means that, on ~28 percent of the days that the Lakers played a game between 1972 and 2017, both the Lakers won and the Bond Index experienced little movement.

This, essentially, is meaningless: the Lakers had the highest winning percentage of all the teams tested, and the Bond Index--as mentioned above with its mean percentage change of 0--was a very consistent stock, so it's not surprising that the highest support is between the winningest team and a highly-consistent stock. Considering these two things--that the highest support was both relatively small and very much uninteresting--the association rule mining produced little about which to be optimistic.

To continue, though, the highest confidence result was .6383 between the Lakers winning and Sturm & Ruger (Guns) seeing little movement. This means that, if the Lakers lost, Sturm & Ruger (Guns) experienced little movement ~64 percent of the time. As could be predicted, this result isn't exciting either: as was alluded to in the last paragraph, the best results came from stocks that consistently saw little movement, and so it's not surprising that all of the highest confidence results included Sturm & Ruger (Guns) seeing little movement and a team's winning; Sturm & Ruger (Guns) rarely did anything but see little movement, and all three of the tested teams had higher winning than losing percentages. And, as for the Lakers winning, that's not telling either: they could only win or lose, and they did the former more often than the latter. So, really, the best confidence numbers are basically approximations of the teams' winning percentages, since the corresponding stocks were just those that most consistently saw little movement.

The Apriori algorithm applications didn't produce great results, then, but further analysis was still performed, as described in the coming sections.

Testing Hypotheses

The association rule mining above gave hints about where to begin with the formal hypothesis testing. Based on the just-mentioned association rule results, three hypotheses were put through a variety of tests: 1. The stock market reacts to the Yankees' performance, 2. The stock market reacts to the Lakers' performance, and 3. The stock market reacts to the number of games played.

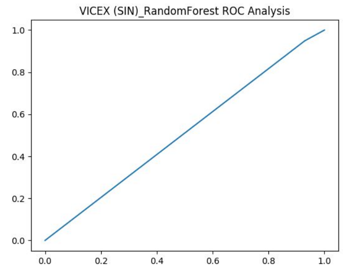
Hypothesis 1: The Stock Market Reacts to the Way the Yankees Perform.

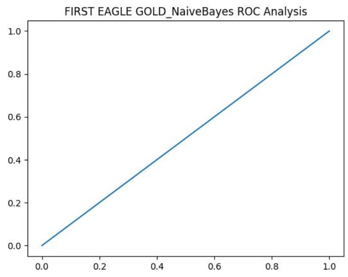
The first tested hypothesis examined the relationship between one of the largest-markets sports teams (the New York Yankees) and movement in the stock market. Six stocks were tested: the S&P Index, Anheuser-Busch, VICEX (Sin), Exxon, First Eagle Gold, and the Treasury Bond Index. Specifically, a relationship was looked for between the Yankees' game results on a given day and the performance of the stock on the next available day. To prepare the data, a separate dataset was created containing the game day date, a categorical variable representing a win or loss (1 if won, 0 if lost), the score difference (positive if won, negative if lost), the next available stock date after the game, and the aforementioned binned stock performance variable based on the percentage change column, if available.

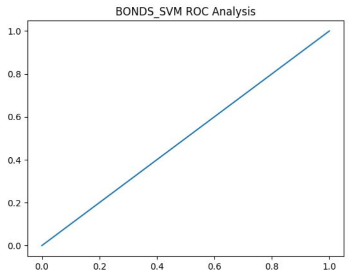
Using Random Forest classification, Naive Baye's, and SVM, a model was trained on the categorical, win/loss variable and the score difference, and then used to predict whether the stock value would go up or down. Almost all of the prediction accuracy scores, as shown below, were slightly better than 0.50, which indicates that the trained model was slightly better than being 50% correct in predicting the stock performance based on the game results. One interesting result was that modeling ofthe Bonds index led to an accuracy score of approximately 0.616, which is higher than the rest of the stocks.

Stock/Index	Random Forest	Naive Bayes	SVM
S&P	0.50375	0.51649	0.51649
ANHEUSER-BUSCH	0.5	0.49265	0.48897
VICEX (SIN)	0.53416	0.52795	0.52795
EXXON	0.51874	0.52024	0.51574
First Eagle Gold	0.55482	0.56143	0.56143
BONDS	0.61591	0.6176	0.6176

Furthermore, most of the generated ROC curves are similar to the line represented by random chance, with some kinks in certain cases. Several examples are shown below, which give the ROC curves for the VICEX (Sin) stock from Random Forest, First Eagle Gold from Naive Baye's, and BONDS index from the SVM method.







These results, combined with the accuracy scores from the prediction methods, produced uncertainty about the connection between the Yankees' performance and the stock market. It looks like, for the most part, there is no relationship; but, as was just outlined, some of the prediction models give reason to hesitate before fully concluding that to be the case. Thus, the null hypothesis that the Yankees' performance does not affect the stock market in any systematic way cannot be confidently rejected.

Hypothesis 2: The Stock Market Reacts to the Way the Lakers Perform

Because the association rule mining somewhat implied that the stock market reacts to Lakers games (that is, Lakers' associations accounted for the best of a bad bunch of results), this idea was put through a series of formal tests. Five stocks were used--the Treasury Bonds Index, NASDAQ Index, Chevron Oil, Franklin Gold and Precious Metals, and Sturm & Ruger (Guns)--and tested to see if their activity differed based on the way the Lakers performed. Specifically, the outcome of every Lakers game since 1972 was tested against the next day's percentage change of these five stocks. First an ANOVA calculation was performed for a difference in the mean percentage change in stock prices on days after the Lakers won, lost, or didn't play. Then, because the way the market performs when teams don't play was not of interest, t-tests were run to determine if there is a difference in market activity based on whether the Lakers win or lose. The table below lists the p-values for these tests, sorted by the stock the Lakers' performance was being tested against:

Stock Tracked	**ANOVA p-value**	**T-Test p-value**
Bonds Index	.4677	.7058
NASDAQ Index	.6417	.3434
Chevron Oil	.0393	.0237
Franklin Gold and Precious Metals	.7698	.5083
Sturm & Ruger (Guns)	.0425	.0441

As can be seen, most of the stocks produced very high p-values for both the ANOVA and t-test calculations described above. This means that neither of their tests resulted in statistical significance; that is, the difference in the means of percentage changes in those stocks the days after the Lakers won, lost, or didn't play cannot be confidently significant or telling of an underlying trend.

But, Chevron and Sturm & Ruger (Guns) *did* seem to react to the Lakers' performance; they both produced very low p-values. Tackling the former first, the mean percentage change in Chevron Oil when the Lakers lost was +.0000180 and +.0012 when the Lakers won. That *is* a significant difference, especially when individual shares of Chevron currently cost ~$117. What about Sturm & Ruger (Guns)? When the Lakers lost the average change was -.0011 and +.00077 when they won, with a current stock price of ~$56 per share. That's an even bigger difference. This seems strange, and so it was checked if these stocks' activity was the result of industry-wide phenomena. The results are below, presented in the same format as those above:

Stock Tracked	**ANOVA p-value**	**T-Test p-value**
Chevron Oil	.0393	.0237
Exxon Oil	.2027	.2434
Marathon Oil	.5198	.7221
Sturm & Ruger (Guns)	.0425	.0441
American Outdoor (Guns)	.0536	.3656

In a nutshell, *no*, the trends weren't industry-wide. The American Outdoors (Guns) fund has a low p-value for the ANOVA test for a difference in means, but the t-test that only considers the difference in mean percentage change on the days after the Lakers won versus those when they lost shows much poorer results. These findings were positive, if only because they were the first to show anything exciting. The difference in these two stocks on days after the Lakers win versus days after they lose *is statistically significant at the p = .05 level.*

Additionally, this hypothesis was put through further testing: k-nearest neighbors and decision tree classifiers. For this, the Lakers W/L data--days when they didn't play were not considered--and the binned stock percentage change categories described earlier were used. The goal here was to see if, based on the outcome of the Lakers' games, the classifiers could predict whether a stock would fall into the category: BIG JUMP, JUMP, LITTLE MOVEMENT, DIP, or BIG DIP.

For both the decision tree and k-nearest neighbors classifiers, cross validation was used to train and then test the model. Because this hypothesis is just testing one explanatory variable against one categorical response variable, these tests were chosen somewhat arbitrarily. However, both were suited for this hypothesis; both decision trees and k-nearest neighbors analysis works well when there are more observations than variables, which was certainly the case here. k = 3 was set for the k-nearest neighbors classifier; this means that the model will look for the three "most similar" data points for each data point in the test set and take their average to predict the unknown point's category. For the decision tree, the model will use the training data to search for patterns and then implement them in a model that filters the test data points into predicted categories; in this case, it won't be a very large tree--it'll have a root node that asks whether the Lakers won or lost that then splits into the six possible stock change categories. Here are the subsequent accuracy scores from the tests, again sorted by the stock being tracked against the Lakers' performance:

Stock Tracked	**KNN Accuracy Score**	**Decision Tree Accuracy Score**
Bonds Index	.2284	.3847
NASDAQ Index	.2499	.4407
Chevron Oil	.4468	.2847
Franklin Gold and Precious Metals	.2569	.3789
Sturm & Ruger (Guns)	.2892	.3824

On the surface, some of these results look promising, but further exploration shows they aren’t. In each case, the model just predicted the change category in the stock that most often occurred in the training data. This is evidenced by one of the decision tree confusion matrices, that for Sturm & Ruger:

```
                [LITTLE MOVEMENT]
|0      0       156     0       0       0|
|0      0       805     0       0       0|
|0      0       1226    0       0       0|
|0      0       393     0       0       0|
|0      0       129     0       0       0|
```

The models just predicted that the stock would remain about the same--regardless of what the Lakers did--because it most often stayed consistent in the training data. And, when the accuracy score (.3824) is tested against the number of days that Sturm & Ruger (Guns) actually did see little movement over the test timespan, it can be found that the percentages are identical. So the best outcomes were really just those that corresponded to prolonged trends in one stock in the market, and thus had nothing to do with the Lakers.

Thus, these further tests produced little of interest. Really, all they suggested were which stocks were most consistent. But, the hypothesis that the stock market reacts to the Lakers’ performance cannot be rejected, because the ANOVA and t-test results seem to imply that Chevron and Sturm & Ruger (Guns) *really do* react to Lakers game. Is this likely, or even believable? No--the Lakers probably just happened to be better than usual in years when those stocks saw nice jumps. But, this cannot be solidified without further research, and so this second hypothesis cannot be confidently rejected. That is, it is *not certain* that the Lakers' performance *does not* affect the stock market.

Hypothesis 3: The Stock Market Reacts to the Number of Professional Sports Games Played the Day Before

To test this final hypothesis, the percentage change in daily stock prices the day after varying numbers of professional sports games were played were measured. For example, does the stock market see a bump the day after NFL Sundays, when there are usually 13 games played? There were a lot of possible combinations that could be tested, but the analyses below measure the NASDAQ Index against the number of MLB games, NBA games, NFL games, and total games across the three professional leagues. This was the only variant of the hypothesis that seemed plausible--that the market as a whole (which is the reason for using a total market index) reacts to the number of games played. Furthermore, the NASDAQ Index is the oldest of the tracked stocks, and so has the best chance of producing relevant results. To test these 4 possible relationships, the number of games were binned into 5 buckets, and then an ANOVA test was conducted for a difference in the mean percentage change in the NASDAQ following days with numbers of games that fell in those buckets. The results are below:

NASDAQ Percent Change Based On:	**ANOVA p-value**
Number of MLB Games	.1753
Number of NBA Games	.4924
Number of NFL Games	.7064
Total Number of Professional Sports Games	.9850

None of these p-values were high enough to be statistically significant, and this final hypothesis' validity thus seemed unlikely. (The p-value for MLB Games is kind of interesting, but if the bins are altered at all it rockets up).

A regression line was fit to the data, though, so as to be certain that this hypothesis should be rejected. In each case, the regression analysis produced insignificant p-values (the likelihood that the coefficient related to the numbers of games played is equal to 0 and so irrelevant to the NASDAQ's movement). The results are below:

NASDAQ Percent Change Based On:	**Linear Regression Analysis p-value (Likelihood that Number of Games Coefficient is Meaningless)**
Number of MLB Games	.113
Number of NBA Games	.120
Number of NFL Games	.149
Total Number of Professional Sports Games	.928

The p-values are somewhat close to significant, but if the linear model's complexity is increased at all the p-values shoot up. For example, if the NASDAQ's percentage change was modeled according to three variables--the number of MLB games, NBA games, and NFL games, separately--the p-values reach: .160 (NFL), .307 (NBA), and .506 (MLB). The NFL p-value stays relatively low, but when put in the context of the corresponding high ANOVA value above, it seems to be accountable to coincidence. One possible explanation is that the market usually has the highest change from Friday - Monday, since the market is closed on weekends; and, since most NFL games are played on Sundays, it's nothing more than a coincidence that the market is volatile following days with NFL games.

Thus, from these tests, the third hypothesis is **rejected**: it can be confidently concluded that the market does not react to the number of professional sports games played the day before.

Conclusion and Potential Further Analyses

In the end, one hypothesis was outright rejected (the stock market reacts to the number of professional sports games played the day before), another gave us uncertainty (the market reacts to the Yankees' performance), and another gave puzzling results (the market reacts to the Lakers' performance). As mentioned in Hypothesis 1, there is some evidence that models can be trained to predict the movement of the market based on Yankees games. The evidence is muddled, though, and so it would require more complex testing to get any concrete results. From Hypothesis 2, though, it certainly looks like Chevron Oil and Sturm & Ruger (Guns) react to Lakers games; if the Lakers win, they shoot up. Thinking rationally, this has to be coincidental, but there's no way to tell from the data. In order to be sure, digging into the stocks' histories and the Lakers' past performance would be necessary. So, maybe, load up on Chevron and Sturm & Ruger (Guns) if the Lakers go on a winning streak? Probably not great portfolio advice, but that isn't what the the analysis above suggests.

Regarding further tests, it would be interesting to explore the way the market reacts to certain trends or big events in the tracked sports leagues. What happens when *all* the large market teams are good? What about during the playoffs? Maybe the non-negligible jump in the market the Mondays after NFL Sundays isn't just a coincidence? All of these are possible subjects of future research.