EXPLORATORY DATA ANALYSIS - STOCK MARKET OVER 10 YEARS. FROM START OF 2012 TO END OF 2021.

Exploring which sectors did best!

FEBRUARY 14, 2023

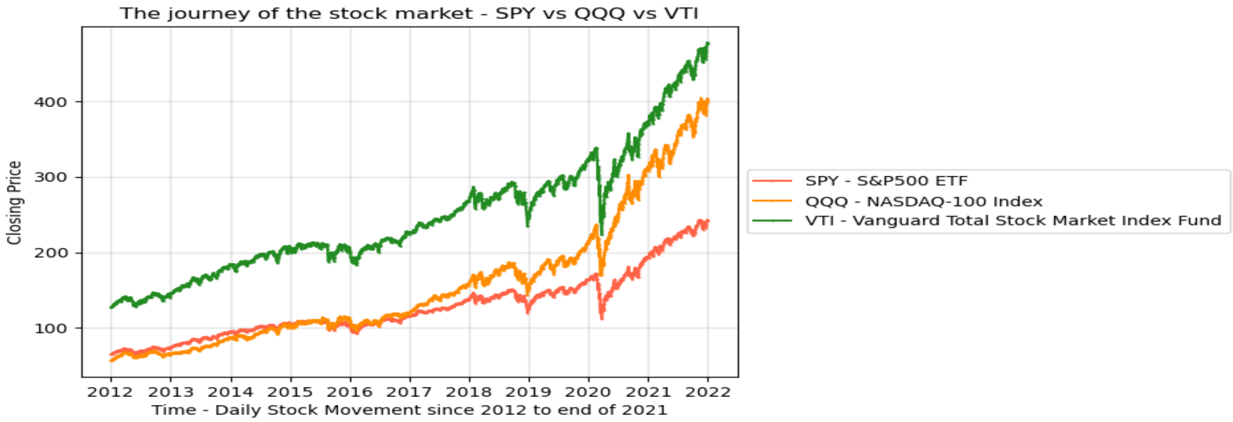
GROUP NAME: IT'S THE JOURNEY NOT THE DESTINATION
Ryszard (Richard) Malolepszy, Pratik Purohit, Mert Oztop, Isaac Rodriguez

1) Our first question was to find how the stock market performed over a 10-year period.

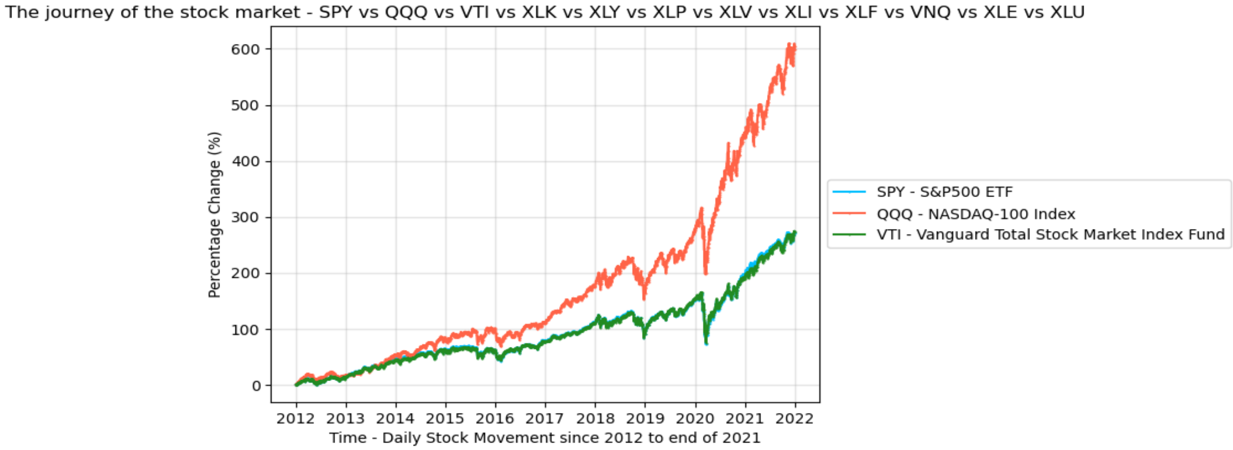
We chose a static time frame for the market from 2012-1-01 to 2022-01-01 so the data would be standardized. After quite a lot of searching for reliable data sources, we found a great source of data using a combination of Yfinance and Pandas Datareader which allowed us to ultimately source the data from Yahoo which has a good reputation for providing solid historical stock data.

We were able to grab our first piece of information from extracting the ETFs representing the base stock performance of the market indexes in the US through the 3 most popular ETFs. SPY represents the SP&500, QQQ which is the NASDAQ and VTI which is the TOTAL Stock Market index.

After making the chart the prices showed up over the span of 10 years but the information wasn't standardized so we came up with the method of finding the Rate of Return for the stocks so they would then be able to be compared to one another.



We used the method of Rate of Return that resulted in this chart:

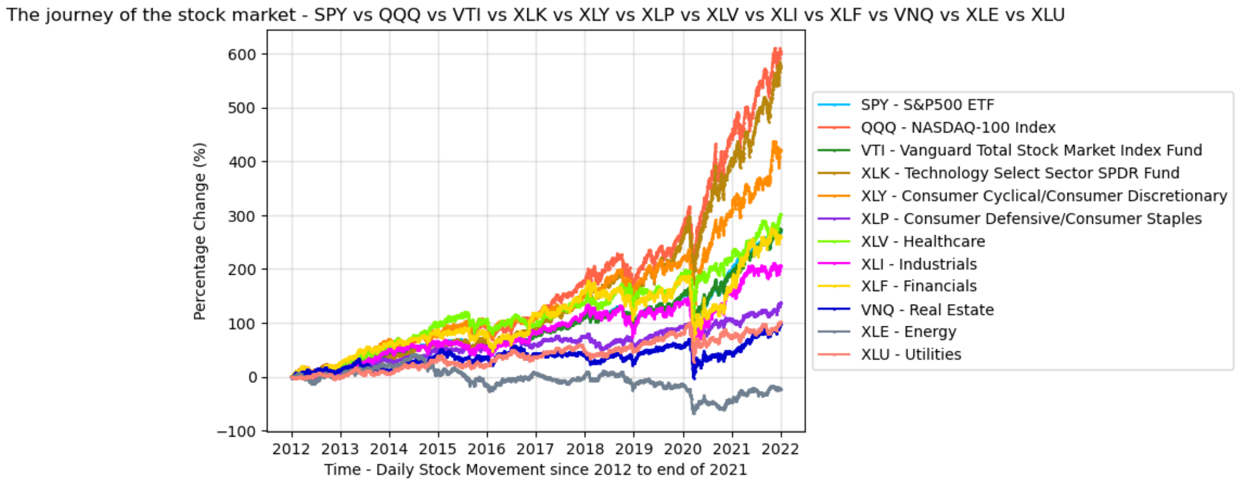


This made us realize that standardization of data helps uncover items which were initially not seen. In this case both the SPY – S&P500 and VTI Total Stock Index both covered almost identical, yet not fully so areas of the market. So, we decided to use SPY going forward as our benchmark, The QQQ was the Nasdaq and very Tech heavy and does not represent the full idea of the stock market.

2) Add in the individual industries and visually see how all of them did.

We then added in the remainder of the ETFs representing the indexes of the market. In total, we have: 'SPY - S&P500 ETF', 'QQQ - NASDAQ-100 Index', 'VTI - Vanguard Total Stock Market Index Fund', 'XLK - Technology Select Sector SPDR Fund', 'XLY - Consumer Cyclical/Consumer Discretionary', 'XLP - Consumer Defensive/Consumer Staples', 'XLV - Healthcare', 'XLI - Industrials', 'XLF - Financials', 'VNQ - Real Estate', 'XLE - Energy', 'XLU - Utilities'

This helped us generate this chart which graphically showed us the performance of ALL the indexes against one another over 10 years:



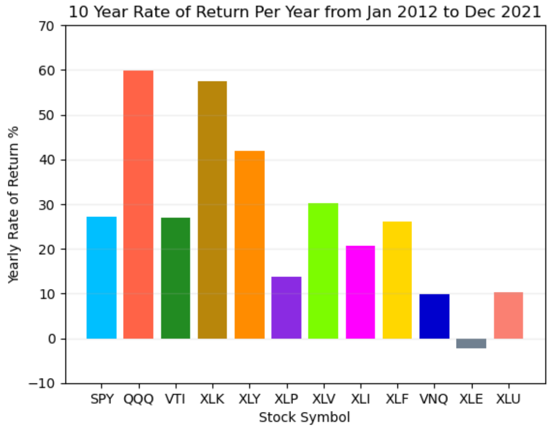
But we still needed to find the mathematical proof of which industry did best. So, it led us to our next question relating to the performance that we could visually and definitively prove was higher than the rest.

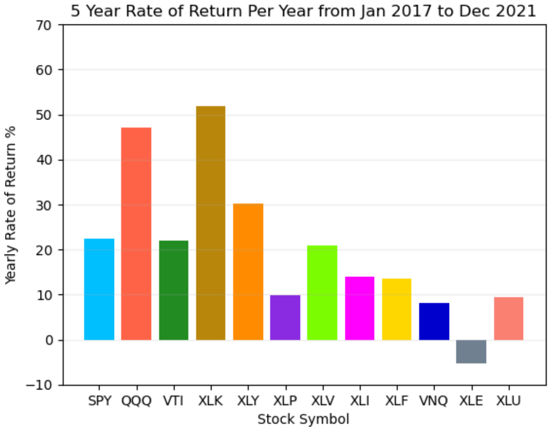
3) Which industries did the best historically:

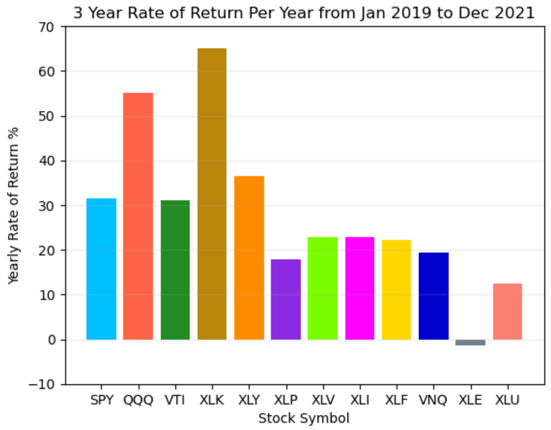
After creating a Rate of Return Calculation for all of our Stock Indexes, we used this information to create a Data Frame which gave us an easy-to-read grid of returns for all ETF tickers that represent the respective industries and Prove mathematically which industry did best over 10, 5, 3, 1 year.

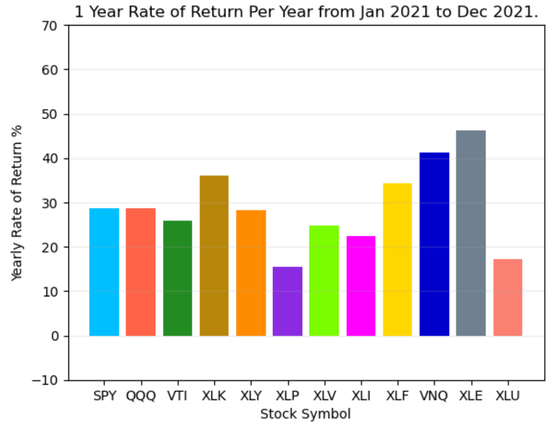
	10 Year Yearly Rate of Return	**5 Year Yearly Rate of Return**	**3 Year Yearly Rate of Return**	**1 Year Yearly Rate of Return**
SPY	27.251764	22.496309	31.496114	28.788736
QQQ	59.920913	47.159013	55.183996	28.625007
VTI	27.024997	21.873049	31.169354	25.835205
XLK	57.365362	51.906531	65.081789	35.942137
XLY	41.954259	30.230958	36.489073	28.239873
XLP	13.755392	9.824019	17.878729	15.572541
XLV	30.093911	20.873221	22.930393	24.736612
XLI	20.678455	14.01157	22.855927	22.451097
XLF	26.034896	13.591397	22.103918	34.28473
VNQ	9.834159	8.113414	19.307561	41.182919
XLE	-2.1886	-5.262878	-1.381693	46.206537
XLU	10.254669	9.474985	12.551283	17.113879

The Data Frame allowed to graphically represent the data in Bar Charts later as well:


10 Year Rate of Return Per Year from Jan 2012 to Dec 2021


5 Year Rate of Return Per Year from Jan 2017 to Dec 2021

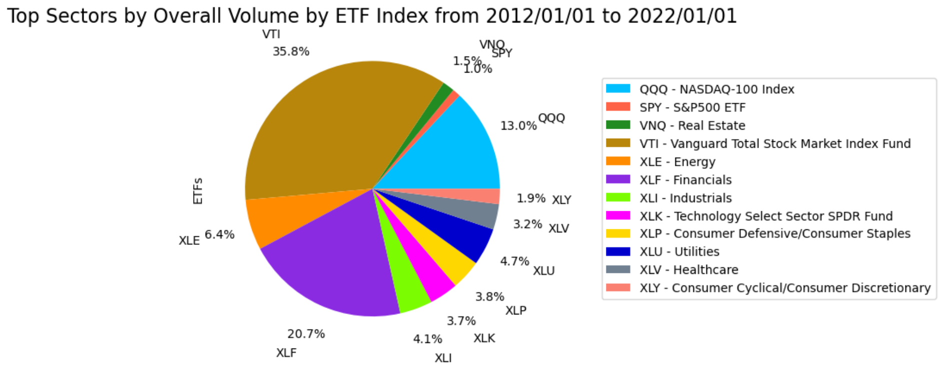

3 Year Rate of Return Per Year from Jan 2019 to Dec 2021


1 Year Rate of Return Per Year from Jan 2021 to Dec 2021.

The QQQ did the best out of all the indexes over 10 years. The XLK did best for the 5 year period. The XLK did best over the 3 year period. And then lastly XLE did the best over the 1 year period. The big takeaway is that High Technology firms as a part of this index would have made you the most money over the 10 years, meanwhile even though XLE had 1 good year at the end you would have lost money over a 10 year time frame.

4) In this question we wanted to find out more about the Industries by Volume and Market Cap and more information about these sectors' dividend yields. Here are the results:

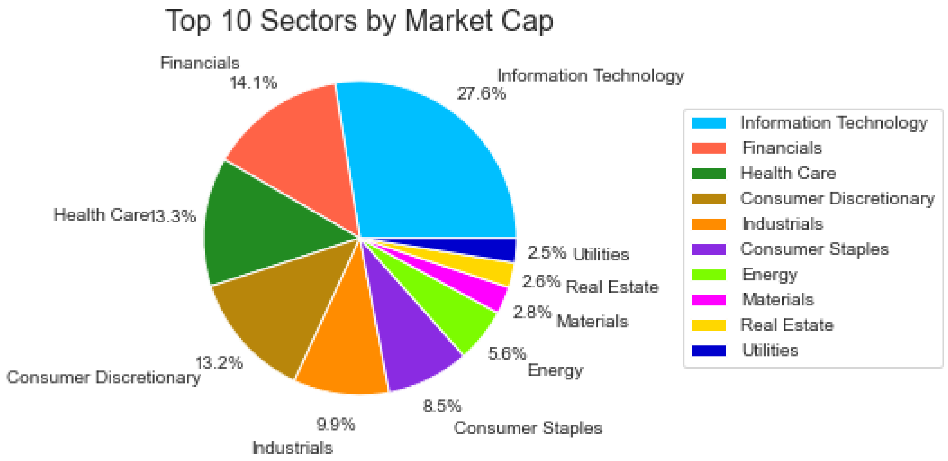
Previously, we looked at the volume traded over the past 10 years on all ETF indexes, to see which had the most trade volume.



This showed us that the most liquidly traded indexes were the VTI – World Trade Index at 35.8% followed by the Consumer Defensive/Consumer Staples, Followed by the main diversified SPY index.

Top 10 Sectors by Market Cap:

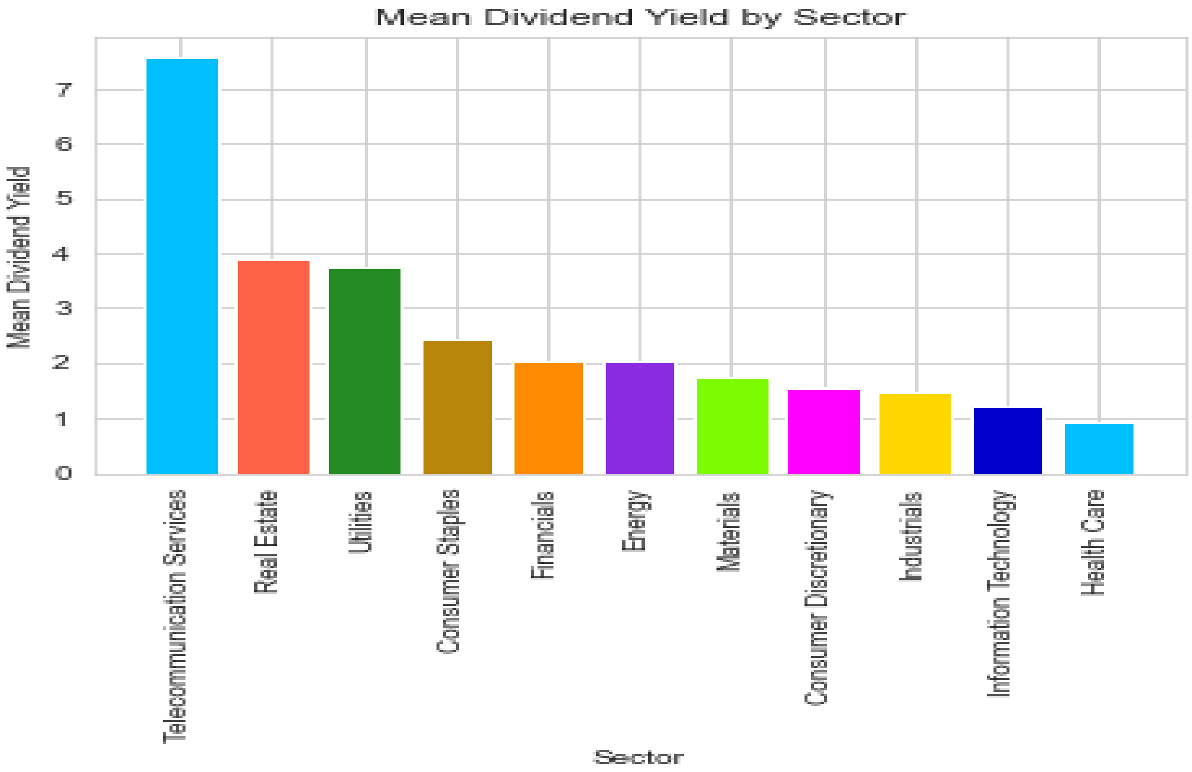
Since the market cap was not part of the data pulled from Yahoo, we chose to use a secondary set of data for market cap from Data Hub. The following pie chart shows the breakdown by industry at the end of 2021.



As you can see, the tech industries in the S&P are clearly outperforming every sector by a large margin at 27.6%, followed by Financials at 14.1% and in third place Health Care at 13.3%.

Mean Dividend Yield by Sector.

Also using our outside data source from Datahub we pulled the data for Dividend yields. The mean dividend yield by sector is visualized. By doing this simple analysis, we can see that it is the Telecommunication Services sector that actually paid the biggest dividend out of the 500 companies in S&P 500.



5) Question five was a bonus idea trying to find if these indexes have a correlation to one another, if we were to pursue this question, we will have to research the methodology and code math to do so which was out of our technical abilities at this moment.

Bonus Project Takeaways:

- A huge component of this project was finding Data to answer our questions. If we would have picked poor Data we would have been spent a lot of time fixing it instead of working on trying to look for meaningful takeaways from it. We took the time to search through the internet through various sources and materials and we discarded a lot of sources just because of how incomplete or fragmented their data was.

- There was a lot of trial and error with making parts of the code to work, and not only work but work correctly. Just because your code works, doesn't mean it provides correct output. One area where we had a learning moment was when we found a Rate of Return formula initially based on Daily Change but it provided a slightly off calculation output which compounded over a 10 year period and added up to significant deviation from reality. That code has to be scrapped and a new method for calculating the Rate of Return had to be checked and cross referenced to other websites and even manually calculated using a calculator just to make sure the results were accurate.

- Communication is Key. Without it a project greatly suffers and when instructions are not followed setbacks occur. One learning moment example of this was GitHub errors.

- Making the presentation look attractive was not something fully taught and a lot of self research was need.

- Same goes with finding Data sources. A lot of self research and trial and error had to occur.