University of Chicago
Department of Computer Science

Final Project December 10, 2023

Sector Correlation: An Economic Sector's Response to External Stimuli

by

Caleb Derrickson

Abstract:

This paper explores the concept of sector correlation in the American housing market by investigating the interplay between individual companies within the sector and their response to external stimuli. The study focuses on macroeconomic indicators such as U.S. GDP, inflation, interest rates, and U6 unemployment, employing a vector represented by the U.S. housing sector for analysis. The chosen methodology involves constructing a complete graph based on the correlation between individual stocks, inspired by Graph Theory, and examining the spectral properties of the graph. The data collection utilizes ETFs tracking the American housing market and macroeconomic stimuli from various sources. Results indicate that correlation within the housing sector responds to economic uncertainty, as seen during the Covid-19 epidemic and the Russian invasion into Ukraine. The iterative removal of the least correlated indices shows different effects during these events, suggesting the sector's varied sensitivity to external factors. The study concludes by highlighting the potential for further refinement in correlation definitions and considerations for future research.

Introduction

Among a given sector of the American Economy, how does the success of each relate to each other, and to the sector as a whole? Is the overall health of the sector dictated by a small number of constituent companies, or is it guided by the interactions among a vast spread of the market? This concept, defined as “correlation” will be discussed in this paper, and we will seek to investigate it against several macroeconomic indicators, such as United States GDP, Inflation, and Unemployment. As a vector to investigate and establish this quantity, the US Housing sector has been targeted as a point of study.

This quantity, labelled as correlation, has several use cases, each of which contribute to the importance of this indicator. A highly correlated sector of the economy, significant once compared to several other sectors, could be a sign of over-speculation in that particular market. With respect to times of high economic volatility - i.e. the 2007 Housing Market crash, and the Coronavirus Epidemic of early 2020 - investigation into the behavior of the overall correlation of a given market may find subtle nuances into its structure. Intuitively, if a sector’s total correlation remains static in the face of relative economic crises, then macroeconomic concerns do not influence the sector as a whole, thus we would say it is highly insulated to the overall economy. On the other hand, if the total correlation responds following changes in the overall market, its health is then guided by the economy as a whole, thus the sector sees influence from others, and behaves as expected.

Another use case of this index can be seen among stock brokers and the distribution of stocks. Considering the hypothetical of a single individual who has previously invested in several companies of a given sector of the market. This person then wishes to invest in an additional company of that sector, and wants to diversify his investments as much as possible. Therefore, a natural choice in the recommended stock would be one with the least correlation with respect to their current holdings. This investigation can then be effectively done in the same vein as the one followed in this paper.

Finally, the last consideration of this index we will discuss is the prevalence of how many companies contribute to the total correlation of the sector. How well can we track the total correlation of a sector by picking only a select number of stocks? How does the total correlation change as we iteratively remove the least correlated indices among the

market? Does it retain its shape, or do we see new characteristics among the increasingly fewer considered indices? Invariance under the removal of the least correlated stocks would indicate the sector is largely guided by the few, higher correlated stocks, and as such see its health dictated by them. This is then characteristically similar to that of an oligopoly. In consideration of non-essential goods, such as fashion or luxury items, this is generally not the most desired structure.

Definitions

There are several definitions that this paper will employ that should be first mentioned and discussed before their use. We have already seen one such term, "correlation" that has been used. With respect to this paper, this term will capture the how effectively one time series tracks another. Therefore, if we are taking X_1 and X_2 as two time-series (stock prices), we will define correlation as

$$\text{Corr}(X_1, X_2) = \sum_{i<N} |X_1(t_i) - X_2(t_i)|^2, \tag{1}$$

where N is the length of time we are studying (five years), and t_i is the i-th day of that time span. This definition is more closely related to a chi-squared test; but since we are comparing two stocks instead of a one stock to a model, this definition is distinct. Furthermore, we are considering correlation as pair-wise phenomena - more sophisticated definitions of correlation, which could include the relative holdings of the given stocks in a given ETF (discussed later in 2) could be iterated on in the future. Intuition also holds that two stocks which are highly correlated should have high correlation, which is reflected in the *inverse* of our correlation function. Please take this into consideration when interpreting plots. Furthermore, the notion of "anti-correlation," where the correlation between stock is negative, is missing from our definition. We will take this idea of anti-correlation into consideration when two stocks have a high correlation value associated.

We should then discuss the medium by which we determine a sector here. This will be discussed further into the paper, but its definition will be stated here. We will be taking a sector as the top number of highest traded ETFs (Exchange-Traded Funds) which only tracks the American housing market. The definition of an ETF with respect to this paper is simply a collection of stocks, invested in such a way to track some quantity or index.

In reality, ETFs can be a collection of any traded commodity, but in order to simplify this study, we will only be focusing on stocks.

One final definition to mention is that of Macroeconomic stimuli. This is just a “catch-all” term to describe factors external to the housing sector as a whole. The list of stimuli which would affect the sector is more than this paper will investigate, but again for the purposes of simplification, we will only be taking macroeconomic factors to be the following: Inflation, GDP, new daily Coronavirus cases [1], U6 unemployment, and the average interest on new 30 year fixed housing mortgages.

Methodology

There could several methods by which to investigate the correlation in a market. These could include a different method of defining a market, a different definition of correlation, a different lens by which we take the total sector correlation, etc. We will discuss why the methods chosen adequately describe the sector’s true correlation.

The method to create the structure of the overall housing sector will be to collect all unique indices within the top considered ETF’s which track the housing sector. The reason to define the sector as such is severalfold, but to do so exploits the behavior of the chosen ETFs. An ETF should retain some idea of attractiveness for potential investors, and as such will divest as many companies within that sector, as well as potential complementary sectors as well. Examples of this would be an ETF investing in the production of new homes, not just their sale. As such we would see investment in companies that provide the necessary resources, such as lumber, labor, concrete, glass, steel, etc. These complementary sectors do play a role in the success of the housing market, however their inclusion is skeptical. In my data collection, I did not take into account the individual companies invested - just the summary of the ETF as a whole. Their inclusion necessarily implies a reliance on other markets of the American economy, and we would expect to see some semblance of market insulation be dulled. This could be an aspect upon which other studies could iterate, but for the purposes of this paper, their inclusion is unimportant.

[1] The span of time in which we are investigating (five years) is strongly dependent on the number of coronavirus cases. And as such we will be labelling this as a macroeconomic factor.

We take inspiration from Graph Theory for the total market's correlation. We will construct a complete graph, with vertices indicating individual stocks with their relevant information, and edges conveying their correlation between the two considered nodes. As such, an edge will connect only two nodes, edges are undirected, and there will be an edge between each node (stock) in the graph (sector). the total correlation of the sector will be the volume of the graph, or the sum over all all edge weights. Note that by our definition of correlation (1), this will be a non-negative quantity. As such, correlation will be treated as a positive number, where a value closer to zero indicates a higher correlation and greater values with less correlation.

The reason in choosing the housing sector to investigate is its importance to the average American, as well as its susceptibility to (irrational) consumer sentiment [4, 5]. In investigation of this market, we hope to see its total correlation be tracked by the considered macroeconomic factors. Since (almost) every American citizen interacts with this market, we hope to see the removal of the least correlated indices reflect very little in the total correlation of the market.

Data Collection

In choosing the ETFs by which to represent the Housing sector, the website etfdb.com [3] was consulted. This website houses an extensive list of the number of publicly traded ETFs, and can also be used to isolate the ETFs which track the American Housing market. As such, the chosen ETFs are as follows:

- REM (iShares Mortgage Real Estate ETF)
- REZ (iShares Residential and Multisector Real Estate ETF)
- IYR (iShares U.S. Real Estate ETF)
- ICF (iShares Cohen & Steers REIT ETF)
- REET (iShares Global REIT ETF (REET))[2]
- USRT (iShares Core U.S. REIT ETF)

[2]Only American stocks from this index were considered.

The relative stock holdings of each ETF is publicly known and can be accessed from [3]. Collecting all unique indices, we see that we can pick among 417 companies to represent the housing market. Among these indices, roughly 60% of these can be pulled using Yahoo Finance's Python API [3], and when normalizing the data, we lose another 20% due to stock prices not reporting daily closing prices. As such, we have a working dataset of roughly 177 companies to represent the housing market.

The following Macroeconomic stimuli were chosen, as well as their sources:

- US GDP [8]
- Inflation (Shelter)[4][7]
- Fixed 30-Year Housing Interest Rates [6]
- US Daily new Coronavirus cases [2]
- U6 Unemployment [9]

This is by no means an exhaustive list of external sources which influence the housing market. These are just a few which can be useful to gauge the market's insulation.

Results

Figure 1 is a representation of the complete graph of the Housing sector. The coloring of each stock is a binary flag, which represents the ETFs within which a stock is found. Admittedly, the coloring is somewhat difficult to discern which stocks are found in which ETFs; due to the nature of this variable (it lies in n-dimensional space, where n is the number of ETFs), coloring can only go so far in describing it. The Graph is drawn using spring layout, which spreads nodes apart as a function of the edge weight between them. Because of this, we would expect to see the most correlated stocks with respect to the entire graph closer to the center, as this minimizes the distance between all nodes. Figure 2 is then the Spectrum of the laplacian of this graph. Note by the nature of its shape, as it slightly increases up until roughly the last 40 eigenvalues, it seems that we

[3] This is a known issue with the API, see [1].

[4] I am only looking at inflation with respect to "shelter" as it is defined.

can consider only the top most 40 correlated indices to retain the structure of the total correlation.

The Volume of this graph is a static number, if we are just summing over the edge weights. Instead, we can take sub windows of the total interval, and calculate the total correlation within that interval. This is demonstrated in Figure 3. Note that the dates listed are the ending dates of the window considered, so the value listed at November 22, 2023 considers the stock prices of only the past year.

Figures 4 shows how the difference in total correlation is related to the given Macroeconomic factors. Note that the beginning of the Coronavirus epidemic was March 2020 (in the United States), and the Russian invasion into Ukraine was February 2022. These dates are important when interpreting these plots. Furthermore, the dates listed for the correlation difference will lag behind roughly one year by the nature of the correlation plot.

Lastly, Figures 5 shows the removal of the worst k indices with respect to the overall correlation plot. Note what is being plotted in these figures is the *inverted* correlation definition of 1, so a higher inverted correlation aligns with the standard intuition. It seems that we the conjecture made when analyzing Figure 2 is manifested here.

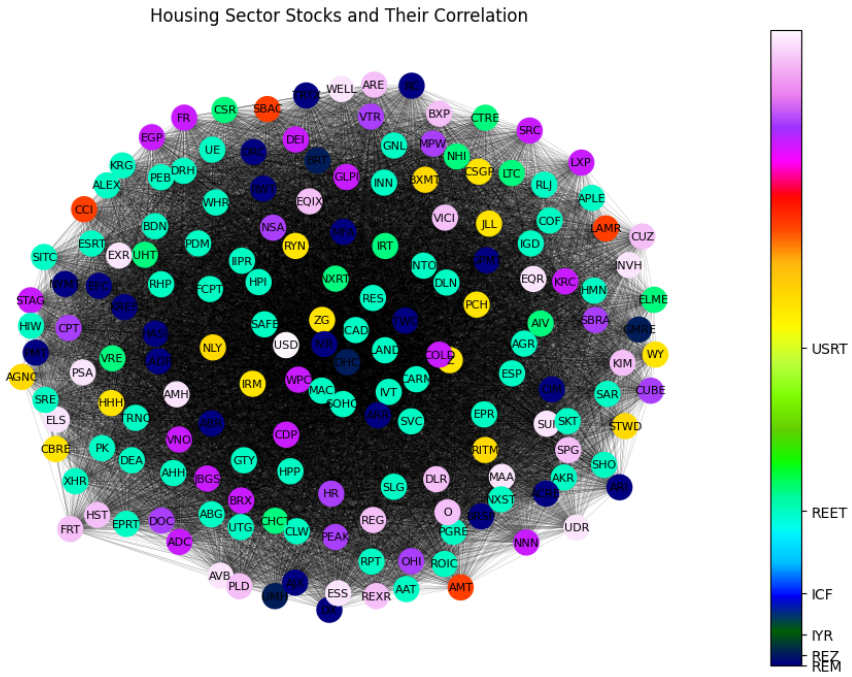

Figure 1: A complete Graph representing the Housing Market, where each stock is represented as its Ticker name. The coloring is a binary flag used to represent the ETFs within which a stock is found.

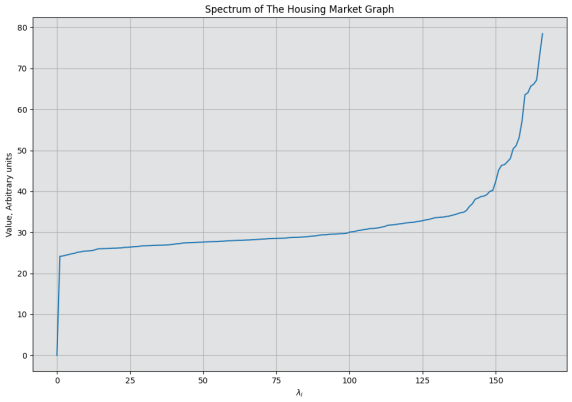

Figure 2: The Spectrum of the Laplacian of the shown in Figure 1.

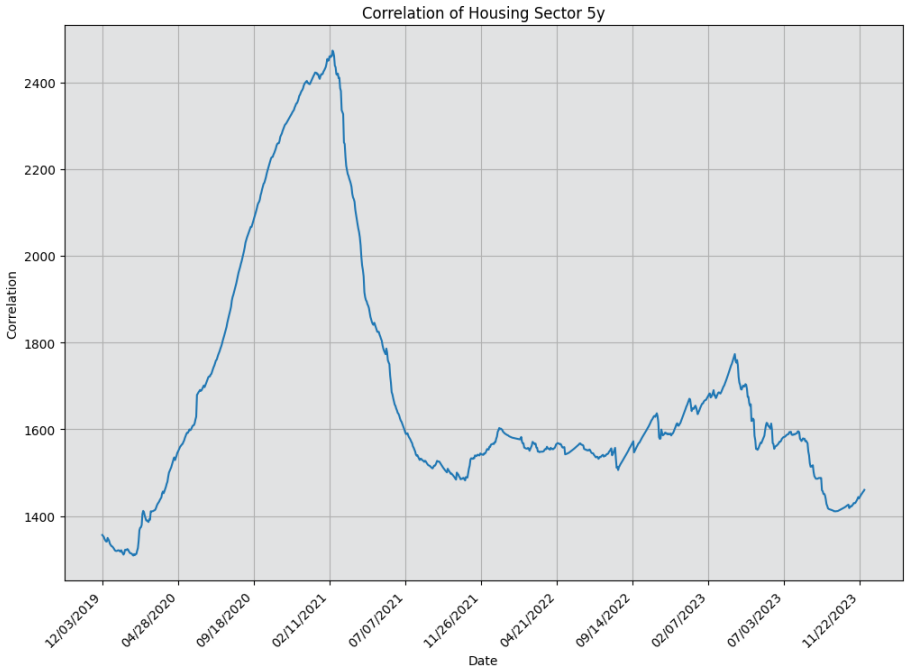

Figure 3: Overall correlation of the American Housing Sector for the past 5 years. Note the date listed according to each value is the ending date of a year-long window.

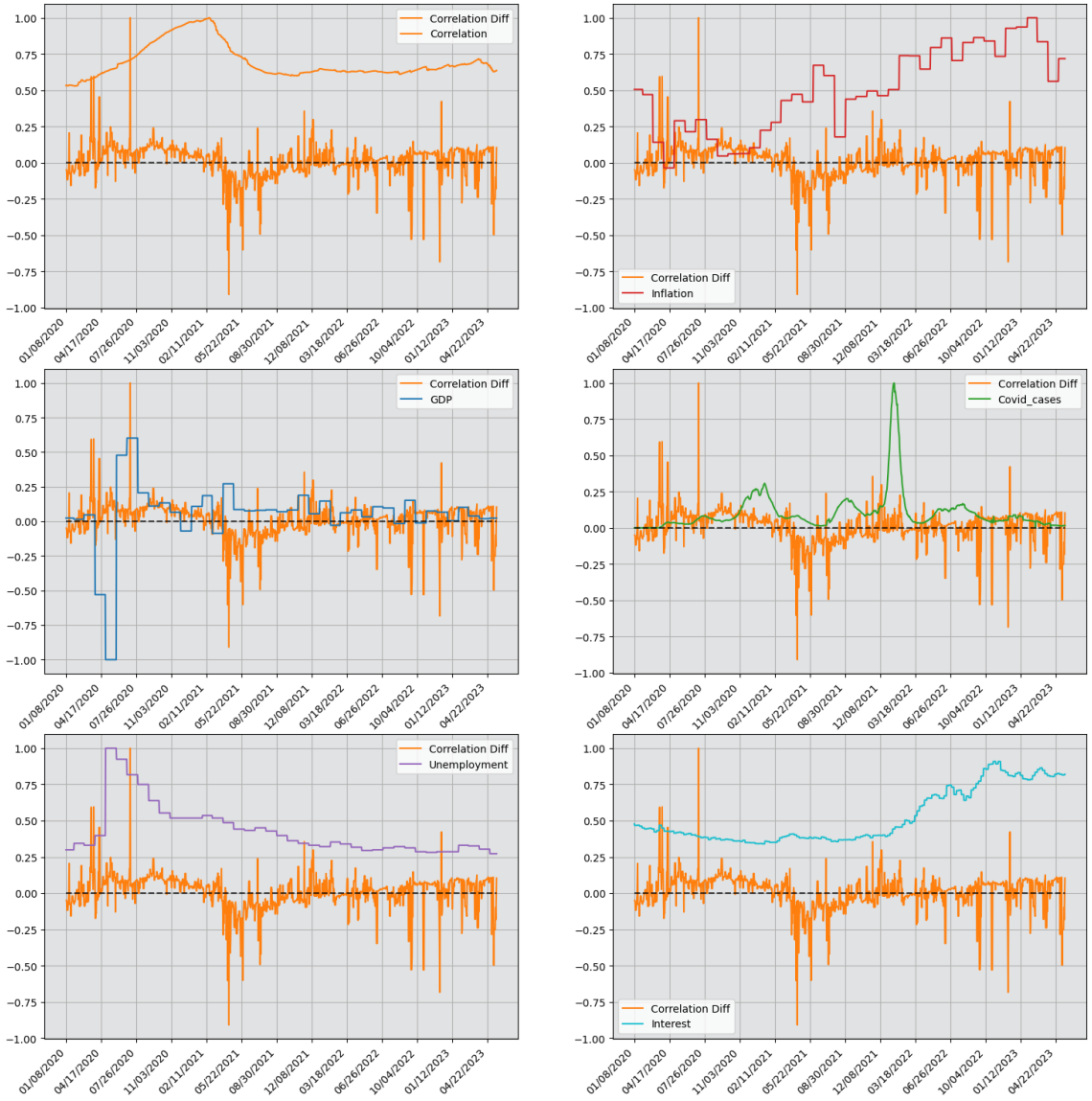

Figure 4: Rate of change in the Figure 3 when comparing to the considered Macroeconomic influences. The first figure just shows the change with respect to the correlation plot itself. Each time series has been normalized for easier viewing.

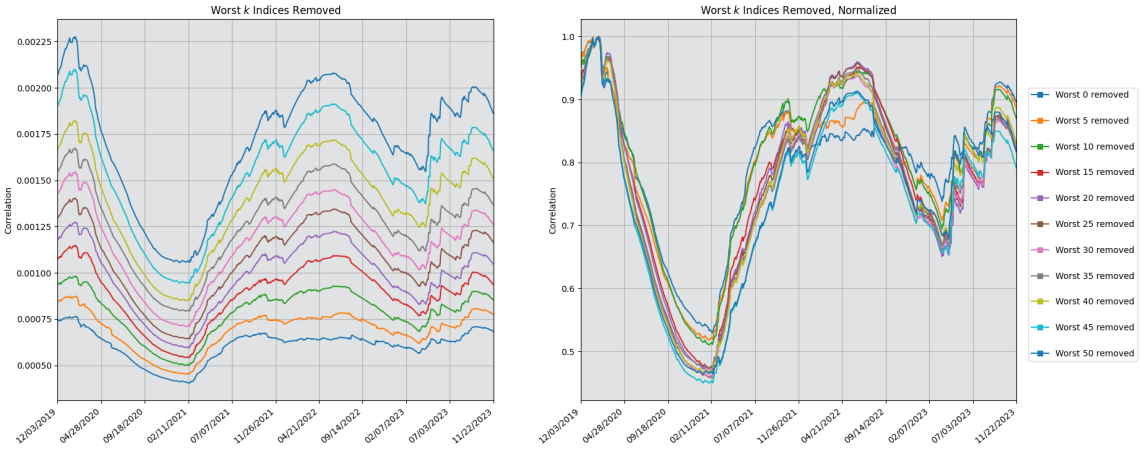

Figure 5: Inverse Correlation plots representing the least correlated k indices removed. (Left) raw values, (right) normalized.

Conclusions

It seems that, in times of economic uncertainty - such as the Covid-19 Epidemic and the Russian Invasion into Ukraine - cause our definition of correlation to spike, meaning that the closing prices between stock vary highly within the 1 year time span considered in each window. We can see in the Macroeconomic analysis, Figure 4, that the correlation difference does seem to be tracked by the factors considered, especially in the fist wave of Covid-19, and GDP. Analyzing these factors raw themselves is somewhat challenging, since the macroeconomic concerns are influenced by themselves.

The test into iterative removal of the least correlated indices, Figure 5, seemed to result in little change when considering the start of the Coronavirus epidemic. When looking at February 2022 however, we do notice an interesting difference when removing more indices. It seems that these two factors, influenced our plots in different ways, causing low variance in March 2020 and higher variance in February 2022. One interpretation into this is the American housing sector was more tangibly affected by the Coronavirus outbreak, so each company roughly suffered the same in this instance. The Russian Invasion however, provides a knock on effect to housing, affecting other markets to be hist around it instead of housing itself.

Overall, we see our definition of correlation to provide an interesting investigation into the health of an individual sector of the American economy. Further studies into this matter should keep in mind the restrictions imposed on this study, as well as other insights into other markets. As an example, the loss of roughy 20% of our stock data can be recovered via interpolation of the data. This was done in investigating Macroeconomic data, in Figures 4, so a similar manner can be performed here. Another aspect which can be improved upon is the definition of our correlation function. Either taking it as the inverse to match intuitions, or adding considerations as to what ETFs are in common between the two stocks, as well as their relative percentage holdings, could provide a more robust definition and give a more accurate picture into the overall market correlation.

Acknowledgements

I would like to thank Lorenzo Orecchia and Konstantinos Ameranis in the Department of Computer Science for conducting lectures and providing insight into Spectral Graph Theory as a whole. Although the finer aspects of the subject were merely glanced in this study, great insight was taken from the course overall.

References

[1] Ran Aroussi. *Issues - yfinance*. GitHub. 2023. URL: https://github.com/ranaroussi/yfinance/issues/359.

[2] Our World in Data. *Coronavirus (COVID-19) Vaccinations - Statistics and Research*. Our World in Data. 2023. URL: https://ourworldindata.org/coronavirus/country/united-states.

[3] ETF Database. *Real Estate ETFs*. ETF Database. 2023. URL: https://etfdb.com/etfdb-category/real-estate/.

[4] Changha Jin, Gokce Soydemir, and Alan Tidwell. "The U.S. Housing Market and the Pricing of Risk: Fundamental Analysis and Market Sentiment". In: *Journal of Real Estate Research* 36.2 (2014), pp. 187–220. URL: https://ideas.repec.org/a/jre/issued/v36n22014p187-220.html.

[5] Cindy Soo. "Quantifying Animal Spirits: News Media and Sentiment in the Housing Market". In: (Oct. 2015). Ross School of Business Paper No. 1200. DOI: 10.2139/ssrn.2330392. URL: https://ssrn.com/abstract=2330392.

[6] Federal Reserve Bank of St. Louis. *30-Year Fixed Rate Mortgage Average in the United States*. Federal Reserve Bank of St. Louis. 2023. URL: https://fred.stlouisfed.org/series/MORTGAGE30US.

[7] Federal Reserve Bank of St. Louis. *Consumer Price Index for All Urban Consumers: All Items*. Federal Reserve Bank of St. Louis. 2023. URL: https://fred.stlouisfed.org/series/CUUR0000SAH1.

[8] YCharts. *US Monthly GDP*. YCharts. 2023. URL: https://ycharts.com/indicators/us_monthly_gdp.

[9] YCharts. *US U-6 Unemployment Rate (Unadjusted)*. YCharts. 2023. URL: https://ycharts.com/indicators/us_u_6_unemployment_rate_unadjusted.