**Can Geographic Location be Utilized to Improve Stock Market Decisions?**

**Jacob Brecheisen (ID: 010945609)**

**Dr. Trent Krupa**

**ACCT 3543 (Section 002)**

**May 3, 2024**

**Executive Summary**

The effect of geographic location on stock market performance has long captivated investors and financial analysts. As data has become more accessible and analytical tools have progressed, we have gained the capacity to examine in-depth how regional economic dynamics can shape stock returns. In this project, we hypothesize that geographic location can significantly enhance stock market decision-making by correlating state-based economic activity with stock performance. By leveraging stock price data from *yahoo!finance* and detailed financial disclosures from *EDGAR*, we aim to uncover specific regional impacts on the financial markets.

Our methodology involves categorizing states into five distinct regions: Northeast, West, Midwest, Southeast, and Southwest. This regional classification allows us to analyze market performance across cultural and economic boundaries within the United States. Our initial findings show that while the Northeast and West experience higher average biweekly returns, these averages are heavily biased by a few high-performing stocks rather than the underlying regional market.

By employing both descriptive and diagnostic analytics, we found substantial differences in stock performance across these regions. Furthermore, our predictive analytics involved regression models to forecast stock performance based on geographic trends.

Based on this comprehensive analysis, we offer two specific investment strategies:

1. **Invest Predominantly in the Southwestern Region:** Our findings suggest that a portfolio heavily invested in the Southwest with a minor allocation to the West yields the best risk-return balance. As such, investors are advised to allocate much of their regional investment portfolio to the Southwest, complemented by a minor investment in the West, to maximize returns.
2. **Target High-Performing Stocks in the Northeast and West:** While general stock performance in the Northeast and West does not consistently outperform other regions, certain high-performing categories of stocks in these areas significantly skew average returns. Investors should focus on these specific outliers rather than the broader market within these regions. Targeting these stocks can provide exceptional returns without the broader regional risk.

These strategies are designed to optimize the geographical diversification of investment portfolios by focusing on areas and specific stocks within these regions that offer the most reliable returns.

**Report**

Our project makes use of financial data sourced from *yahoo!finance* as well as *EDGAR*. The *yahoo!finance* data includes comprehensive stock information such as date, open, high, low, close, adjusted close, and volume from 2015 to 2020. For our purposes, we discarded everything but the close price to enable a more constrained analysis of trends. Geographical information was obtained via *EDGAR* where we scraped the business address from each company’s 10-K filing. In this case, we once again narrowed our approach to utilize only state-level geographic information. We transformed and merged these two sources into a single table containing the state and closing prices of each stock for each biweekly period. Lastly, these closing prices were normalized into percentages by dividing each value by the initial price for that period.

Our exact analytical process is outlined as follows:

1. Descriptive Analytics:
   1. **Summarization:** Utilize summary statistics (mean, median, standard deviation, etc.) and other descriptive statistics to understand the current state of stock performance across different geographic locations.
   2. **Visualization:** Create charts and maps to visually summarize how stock performance correlates with geographic location.
2. Diagnostic Analytics:
   1. **Anomalies/Outlier Analysis:** Identify unusual stock performance and investigate potential causes in specific regions.
   2. **Variance Analysis (Mean):** Utilize ANOVA to determine if there are any statistically significant differences between the means of each region.
   3. **Variance Analysis (Median):** Utilize Kruskal-Wallis test to determine if there are any statistically significant differences between the medians of each region.
3. Predictive Analytics:
   1. **Regression Analysis:** Predict future stock performance with an integrated geographic component.
   2. **Cluster Analysis:** Group stocks based on similar performance trends across different geographic locations.
4. Prescriptive Analytics:
   1. **Optimization:** Develop models to find optimal stock picks based on geographic diversification.

**Descriptive Analytics**

Our descriptive analysis begins by grouping states into five regions: Northeast (NE), West (W), Midwest (MW), Southeast (SE), and Southwest (SW). While the distribution of data across these regions is relatively balanced, there is a noticeable discrepancy in the number of companies between NE/W and MW/SE/SW as is shown by the visible jump in *Figure 1*. In fact, most metrics reflect this same disparity.

Looking at biweekly average returns, the NE and W regions experienced higher returns ranging from 1.9% to 3.5%, compared to notably lower returns in the MW, SE, and SW regions which posted average returns of 0.2%, 0.3%, and 0.8%, respectively. Below, *Figure 2* illustrates the national stock performance, highlighting that although the median return rate among the five regions is closely aligned – with variations not exceeding 0.1% from the breakeven point – the most significant discrepancies in performance tend to be driven by outlier companies in states that do not follow typical market cycles.

Volatility analysis reveals some of the same stark contrasts: NE and W exhibit remarkably high standard deviations, between 668.8% and 1109.9% which indicates extreme variability. In contrast, the MW, SE, and SW regions show much lower volatility, with standard deviations of 11.5%, 15.4%, and 135.5%, respectively. Given this high variability, particularly in the NE and W regions, it is crucial to adjust for risk when evaluating returns. By employing the Sharpe ratio, utilizing a risk-free return rate from a 3-month Treasury Bill of 5.26% (as of May), a more balanced picture emerges. After this adjustment, the MW region shows a negative ratio, whereas the other regions range between 0.3% and 0.9%, reflecting a more reasonable distribution of risk-adjusted returns.

**Diagnostic Analytics**

The first part of our diagnostic exploration focused on anomalies and outliers in stock performance. Just as before, our review confirmed significant discrepancies in the average biweekly returns across regions; notably, the Northeast (NE) and West (W) reported higher averages of 21.7% and 32.1% respectively, compared to more modest returns in the Midwest (MW), Southeast (SE), and Southwest (SW) shown in *Table 2*.

However, these averages were skewed by extreme values, as evidenced by the extraordinarily high standard deviations in NE (2189.9%) and W (3342.3%). Furthermore, despite the high averages, the median returns across all regions were tightly clustered, suggesting that typical regional stock performance might be more uniform than it first appears and largely independent of geographic location.

The Sharpe ratios further illustrated this point, with SE displaying the most favorable risk-adjusted return at 7.4%, contrasting sharply with the lower ratios in NE and W, both just under 1%. This analysis highlighted that the higher raw returns in NE and W were accompanied by significantly increased risks.

After identifying outliers in our dataset, we proceeded to assess the statistical differences between groups more rigorously. Initially, we utilized a conventional ANOVA test to explore potential discrepancies between group means. Despite its straightforward application, the results, as illustrated in *Table 3*, revealed no statistically significant variance among groups, with a P-value significantly exceeding the standard 5% threshold.

Still, given that our previous findings indicated the average return rates might not fully capture the nuances of regional performance, we employed the Kruskal-Wallis test. This non-parametric method focuses on medians rather than means, providing a more robust analysis given our dataset's characteristics. The results from the Kruskal-Wallis test confirmed a notable statistical difference, with a P-value well below the 5% threshold (*Table 4*). This outcome suggests that while overall average performance may appear consistent across regions, the performance of individual stocks can vary significantly, potentially excelling in specific areas.

One possible example of this, for instance, might be technology stocks showing enhanced performance in established tech hubs due to regional advantages in innovation and infrastructure. Similarly, agricultural stocks might outperform in regions like the Midwest, where agricultural activities are more concentrated and supported by local economies. This variance underscores the importance of considering regional strengths and specialties when analyzing stock performance.

**Predictive Analytics**

Before we could proceed with predictions, a primary challenge emerged: how to integrate our state-based geographical information into the model effectively. One solution was to one-hot encode the data, creating fifty binary columns for each state to indicate whether a company belongs to a particular state. However, this method is computationally intensive and could produce an overwhelming amount of data that might be difficult to interpret meaningfully. Alternatively, label encoding the state variable (e.g., ‘AL’ → ‘0’, ‘AK’ → ‘1’, ‘AR’ → ‘2’, etc.) presents a cleaner and more efficient approach, but it introduces an arbitrary numerical relationship that could mislead certain algorithms, such as linear regression, which interpret these numbers as ordinal data.

To address these issues, we employed target encoding. This technique assesses the relationship between our categorical geographic data and the target variable, transforming the categorical state information into a meaningful scalar value. This method not only maintains the dataset’s manageability but also has some level of interpretability, striking a balance ideal for our analysis.

With this encoding strategy in place, we executed three different linear regression models to explore the influence of geographic location on stock returns. The first model incorporated the newly encoded geographic variable alongside data from the preceding nine days to predict the return on the tenth day. The second model relied solely on the data from the preceding nine days, excluding the geographic factor. The third model utilized only the encoded geographic variable. The outcomes of these models are detailed in *Tables 5, 6, and 7*. Although the results were statistically insignificant, they provided a clear conclusion: geographic location, as encoded, does not serve as a reliable linear predictor of average stock returns. This insight, while subtle, underscores the complexity of market dynamics and suggests that additional factors or more sophisticated modeling techniques might be necessary to capture the nuanced impact of geography on stock performance.

Given the unexpected results from our linear regression models, we explored an alternative approach by clustering stocks into groups and then comparing these groups across different regions, as shown in *Figure 3*. To perform this analysis, we employed Dynamic Time Warping (DTW), a method particularly adept at identifying similarities in time series data that may experience shifts or distortions over time. DTW is more forgiving of temporal discrepancies in patterns than standard techniques, making it suitable for financial data, which often exhibit non-linear behaviors and subtle idiosyncrasies.

Due to computational constraints, we limited our sample to 1,000 observations for the DTW analysis. Despite this limitation, the findings clearly demonstrated that significant stock groups span across regions, with the most notable differences attributed to the prevalence of these groups within each region. For example, the NE/W regions showed a much larger number of stocks from group 4. While this group's presence didn't notably influence median earnings, it did lead to a pronounced skew in the average returns. This skewness in the averages primarily arose from the concentration of high-performing stocks within these groups, which, although they do not represent the majority, have a disproportionate impact on the financial metrics.

This insight sheds light on the seemingly contradictory observations present in our initial analysis. While geographic location alone is not a definitive predictor of stock returns, the regional distribution of specific stock groups can significantly influence overall financial outcomes.

**Prescriptive Analytics**

Finally, we wanted something actionable to take away from these insights. So, we applied the Markowitz portfolio selection method to our regional data. The Markowitz model, foundational in modern portfolio theory, aims to optimize the allocation of assets in a portfolio by maximizing returns for a given level of risk, or conversely, minimizing risk for a desired level of return. This method uses the variance of asset returns as a measure of risk and computes the covariance between pairs of assets to assess diversification benefits.

In our application of the Markowitz model, the Southwestern (SW) region emerged as the predominant choice. The model allocated a substantial 97.9% of the total investment to the SW region and a mere 2.1% to the West (W) region, completely excluding other regions. This allocation suggests that, from a macro perspective, investing heavily in the SW region might offer the best risk-return trade-off according to the model's calculations. It's important to note, however, that this analysis did not delve into specific stocks within each region. There might be individual stocks in other regions that offer attractive investment opportunities, even though they did not surface in our region-wide analysis. This indicates that while the SW region may be perceived as a 'safer bet' overall, potentially lucrative investments could still exist elsewhere. This is a potential avenue for future analysis.

Regardless, the results of this specific portfolio strategy, illustrated in *Figure 4*, forecast an expected annual return of just above 8%. Despite the high level of volatility, pegged at 135.4%, the portfolio achieved a desirable Sharpe Ratio of 5%. This ratio, which measures the risk-adjusted return, was the most favorable compared to all other configurations. This underlines the efficacy of the Markowitz model in crafting a portfolio that balances risk and reward, providing a valuable, actionable strategy based on our regional stock performance data.

**Appendix**

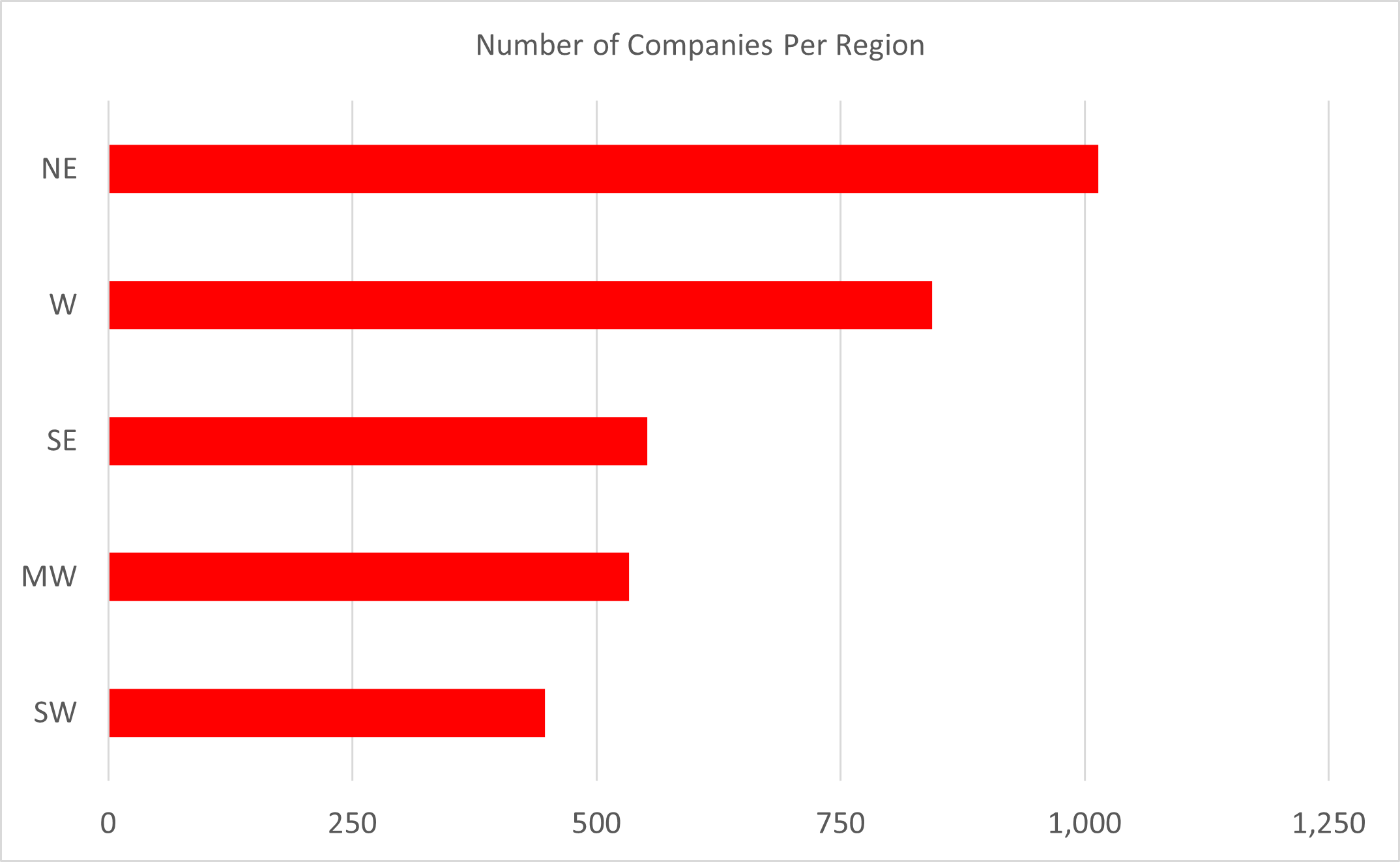


Figure 1 – Frequency distribution showing the number of listed companies per region.

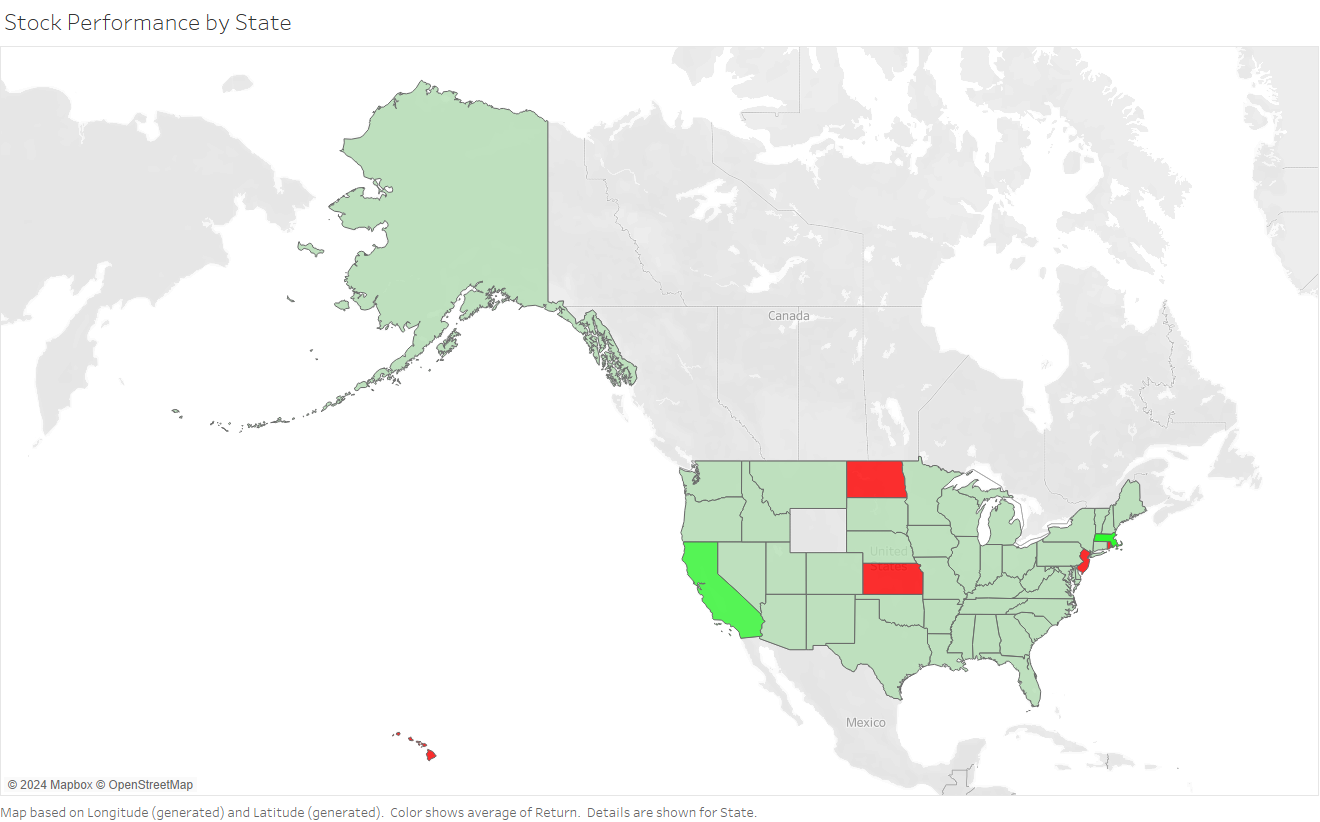


Figure 2 – Heat map showing average ROI by state based on historical data.

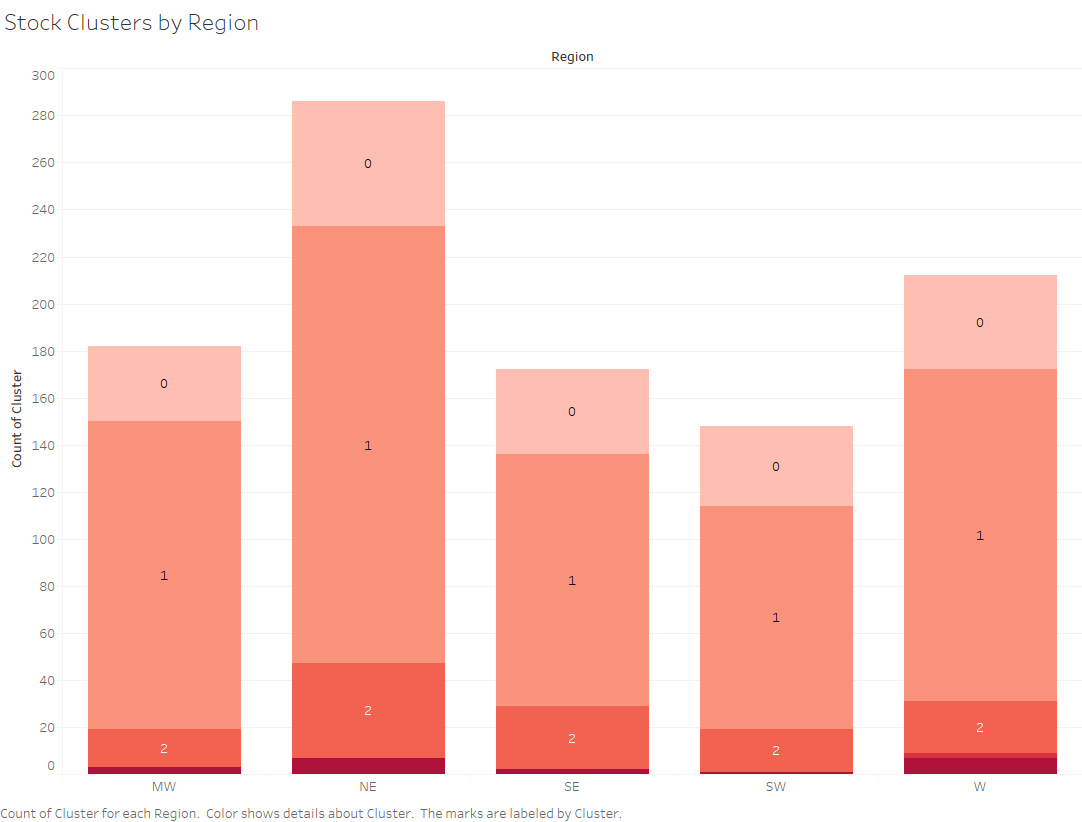


Figure 3 – Regional Distribution of Stock Clusters Using Dynamic Time Warping Analysis.

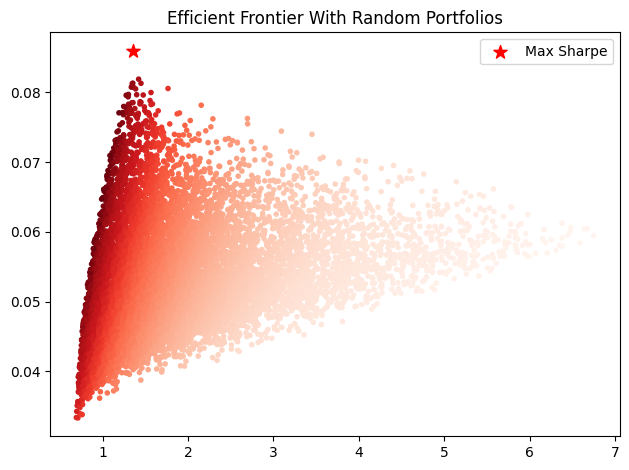


Figure 4 – Efficient Frontier with Random Portfolios Illustrating the Application of the Markowitz Model.

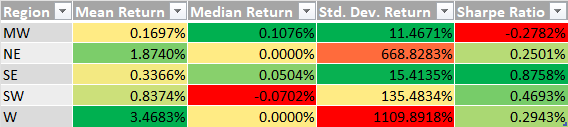


Table 1 – Stock performance by region.

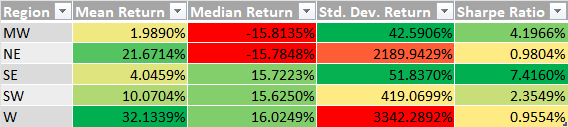


Table 2 – Outlier stock performance by region.

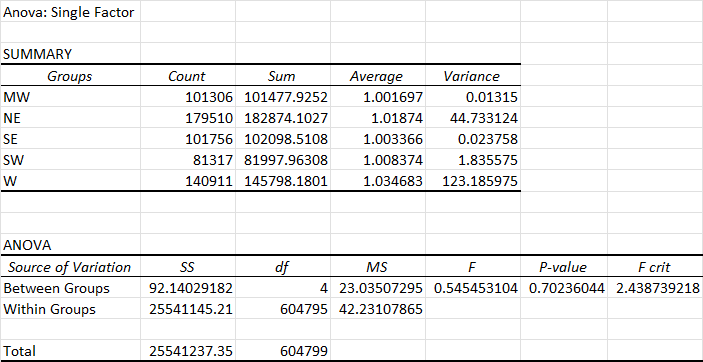


Table 3 – ANOVA on the mean return of each region.

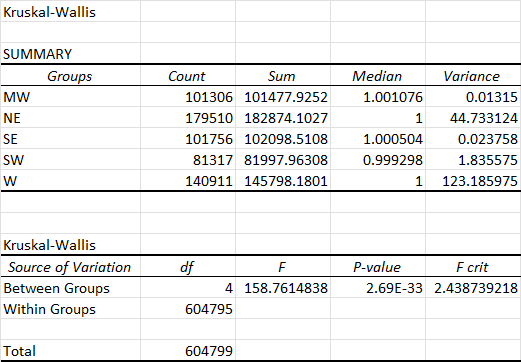


Table 4 – Kruskal-Wallis test on the median return of each region.

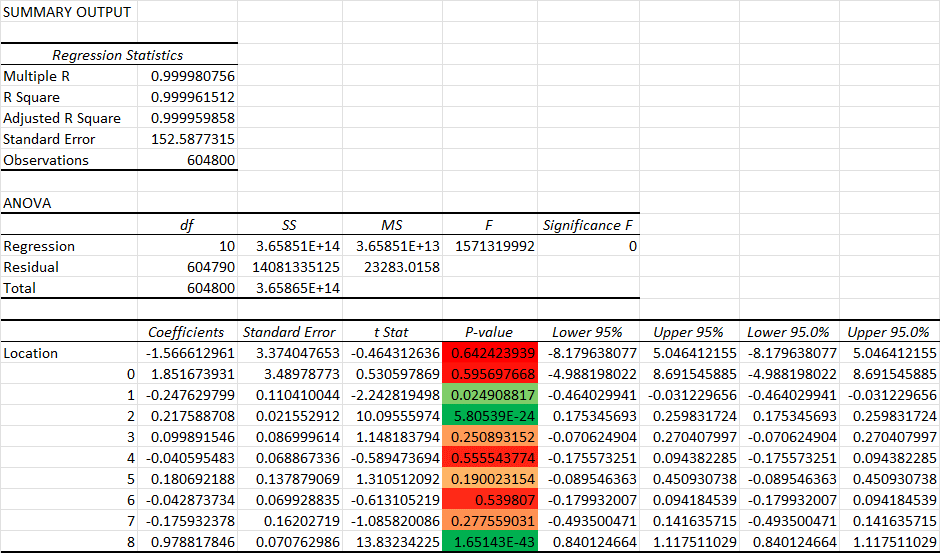


Table 5 – Return regression with location variable included.

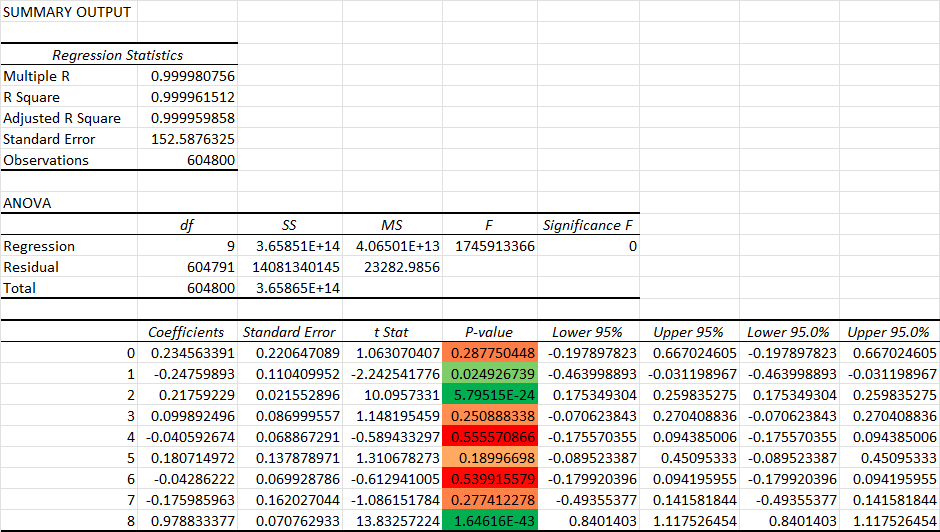


Table 6 – Return regression with location variable excluded.

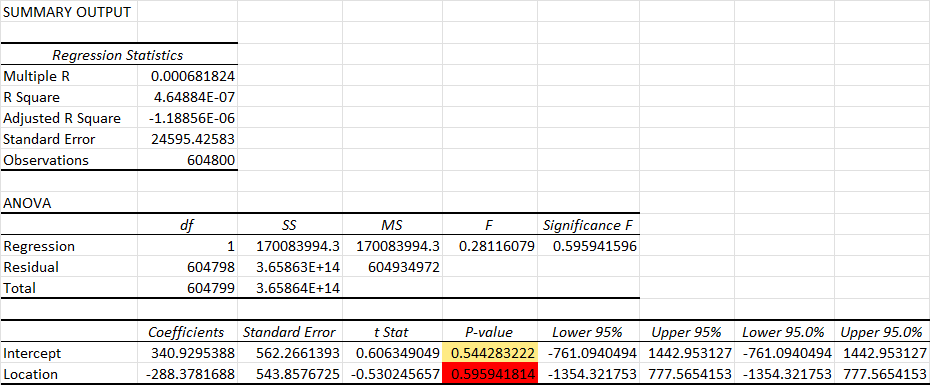


Table 7 – Return regression with only location variable.