Data Analysis – You’re Now the Hacker  
Position Paper 1

DSE6003 – SU1 2025  
Nick Lagoni  
6/5/2025

# *Summary of Findings*

# Introduction

The scenario I selected to drive the structure of this exploratory data analysis is centered around a political science/campaigning firm looking for a data-driven analysis examining what characteristics contribute to political involvement/rates of civic engagement. This firm is particularly interested in how consumer characteristics may be correlated with or influence their political involvement. They would like to look broadly at available factors, ranging from data on debt, income, and consumption metrics to seemingly inconsequential lifestyle choices. The hypothetical firm is particularly interested in seeing which factors are related to an actual voting history, compared to which factors are correlated with a person “picking a side,” or choosing a political affiliation.

The end goal of this analysis is to provide a cursory exploration of the available data in the customer data set provided in order to inform further, more in-depth analysis to more comprehensively address the needs of the hypothetical political firm. More specific goals will be set after the exploratory analysis is complete and the “client” (myself in the future for Assignment #3) has had a chance to review it.

# *Technical Report*

# Data Processing

## Feature Creation and Consolidation – Tableau Prep

In the previous portion of this project, I conducted a series of data preparation steps in Tableau Prep to enable exploratory data visualization and, in this portion, segmentation. I created indices to coerce multiple relevant features into one, consolidating columns. I created the Tech Literacy Score to assess each individual's technological understanding based on their use of appliances and services; the Media Consumption Score to quantify exposure to both politically relevant and consumeristic media sources; and the Socioeconomic Agility Score to classify each customer’s economic mobility and resilience. These scores were then bucketed into qualitative categories to support exploration and profiling. I also calculated comparison averages and integrated external regional mapping data to support geospatial analysis.

## K-Means Clustering – R

Using the cleaned data set produced as an end product of Assignment 3 as an input, I utilized R to do unsupervised clustering of the data based on multiple features that are most relevant to the business question at hand. I chose an unsupervised method to help me sift through the large number of dense, numeric features that felt relevant to the business question at hand. I selected K-means as it is relatively simple to implement and interpret and produces clusters that are clear and distinct from each other, My step-by-step process was:

1. I created a new data frame by selecting only the variables most relevant to segmentation for the political firm: *Age, Education years, Socioeconomic Agility Score, Media Consumption Score, Tech Literacy Score, and Commute time*. These variables were chosen based on their perceived relevance to voter engagement, while also providing some context on demographic, behavioral, and lifestyle factors.
2. I standardized the features above, ensuring that each feature contributes equally to the clustering process. This step was necessary because the selected variables were measured on different scales (e.g., years vs. arbitrary index scores). These skewed scales would throw off the Euclidian distance based k-means clustering method if they were not normalized.
3. I visualized the within-cluster sum of squares (WSS) across different values of k using the fviz\_nbclust() function from the factoextra package. This produced a typical elbow method plot, which identified the point at which increasing the number of clusters no longer significantly reduces within-cluster variation. The optimal number of clusters was determined to be 3.
4. I visualized the determined clusters using principal component analysis and plotted them in ordination space to see their separation. I used the PCA loadings to determine what most strongly contributed to the clustering.

This K-means clustering resulted in 3 distinct clusters, with a small amount of overlap between them. Their positions in PCA-ordination space are shown in Figure 1. The primary features contributing to the clustering were age, socioeconomic agility, media consumption, and tech literacy.

Cluster 1 is comprised of younger (average age of 32) customers with the lowest average income, home ownership, short employment tenure, and education; these individuals are associated with middling tech literacy and media consumption scores. Cluster 1, henceforth dubbed the Nascent Navigators, seems to be a group that is lacking in, most of all, *time*. They haven’t had the time to build the wealth, or the education, or the engagement, that the other clusters have. They’re less well established and have less of an understanding and interest in the technology and media available to them, possibly because they haven’t had the exposure yet, or because they don’t have the means or time to focus on anything else but climbing. There could be any number of reasons behind this, but the end product is clear: a younger cluster that is financially rocky and tepid towards technology and media, still navigating their formative years and yet to establish their place.

Cluster 2 is comprised of primarily middle-aged customers (an average age of 46), highly educated, high-income, tech literate, homeowners and vehicle owners who consume a large amount of media. Cluster 2, dubbed the Established Earners, represents a more well established, wealthier cluster that engages much more heavily with technology and media. They represent an important voter bloc of keen interest to the political consulting firm, as they have the resources to motivate change as well as the education and social consciousness—learned from their engagement with technology and media—to be invested in that change.

Cluster 3 is the oldest of the three clusters, with an average age of 63. Cluster 3, dubbed the Stable Silvers, strikes the middle ground in a lot of areas—they are moderately educated, with just over 12 years of education on average, suggesting that most are nearly high school graduates or just beyond. They have moderate income and low debt, suggesting that their finances are calm, but that they are likely retired or transitioned to lower-intensity careers. They have the highest socioeconomic agility of all three clusters, the longest card tenure, a small household size, and a high rate of home ownership, suggesting that they have a lot of financial flexibility. The small household size and highest average number of pets suggests that they have moved beyond the stage of their lives where they raise a family, either they never did or are empty nesters. The Stable Silvers have the lowest tech literacy of the three groups but a moderate media consumption score, higher than the Nascent Navigators but lower than the Established Earners. This suggests that the Stable Silvers may be particularly prone to misinformation or inflammatory campaigns, as they are less equipped to access free, unbiased information, and may have built less critical thinking skills when it comes to digital media, while still staking in a substantial amount of media.

A graph of a diagram

AI-generated content may be incorrect.

Figure 1 – Principal Component Analysis of K-Means clustering with 3 resulting clusters. Dimension 1, Media and Technology Engagement, represents each individual’s interaction with media and technology and is primarily informed by Media Consumption and Tech Literacy scores. Dimension 2, Level of Establishment, represents how established an individual is in terms of finances and life-stage and is primarily informed by Age and Cluster 1 represents the Nascent Navigators, a young demographic with little financial stability, education, or wealth. Cluster 2 represents the Established Earners, a middle-aged demographic with high income, education, and media and tech literacy. Cluster 3 represents the Stable Silvers, the eldest demographic with moderate income, well-established wealth and assets and low tech literacy.

## Rules Based Segmentation – Tableau Prep

This rules-based segmentation approach created clusters rooted in geography, job type, town size, and union affiliation. These features were selected as significant determinants of a person’s social and cultural standing, while still being easily recognizable and part of the general public conscience. The goal was to do the exact opposite of the unsupervised K-means clustering: instead of building clusters purely off of the shape of the data, I built clusters based purely off of a general interpretation of different archetypes of American geopolitics and culture. The details of how these clusters were built in Tableau Prep as a calculated field using a set of rules are detailed in Table 1.

The Coastal Elites cluster captures highly educated professionals and salespeople in mega-cities or extra-large towns on the Northeast and West Coast. These regions were selected as they are commonly associated in the general public consciousness with cultural influence, higher income brackets, and more progressive values. This group reflects the archetypal affluent, urban, socioeconomically powerful voter.

The Rural Blue-Collar cluster instead focuses on individuals in small or large towns across the South and Midwest working in labor or agriculture jobs who are not part of a union. These selections represent a traditional working-class demographic, often in economically vulnerable areas, and likely to have distinct political and cultural attitudes. They are meant to represent the individualistic, geographically isolated, traditionalist rural worker, as contrasted by workers in a union who are inherently more collectivist.

The Urban Blue-Collar cluster includes service and labor workers in major cities or large metros in the Midwest and on both coasts. These regions were selected as they house the most populous cities in the country. These individuals often work in physically demanding or low-wage jobs but reside in denser urban environments, setting them apart from their rural counterparts. They may view their current employment as temporary, whereas rural blue-collar workers may be more ingrained in their positions.

Skilled Union Tradespeople were isolated based on job category (crafts or labor) and union membership, regardless of location. This group was designed to capture the politically important segment of blue-collar workers with organized labor protections and often strong collective identities. This archetype is designed to capture the more established and less individualistic blue-collar workers.

The Suburban Middle-Class cluster was created to fill the gap of individuals in medium or large towns who hold mainstream jobs in service, sales, or professional sectors, outside of the coastal elite regions. These individuals represent the middle of the economic and geographic spectrum and might not be strongly aligned with either extreme but represent a large, moderate segment of the population. This archetype was created to fill a large gap that I identified after creating the initial 4 archetypes, when I recognized that the only place I was pulling professional or middle-of-the-road workers from were the coasts.

Table - Rules-based clustering archetypes. Each archetype represents a culturally and politically relevant type of individual, described by their geography, community size, profession, and union status. Each column represents a cluster or archetype and each row represents a feature used to inform the rules-based clustering by creating a calculated field in Tableau Prep Builder.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Archetype | **Coastal Elites** | **Suburban Middle-Class** | **Rural Blue-Collar** | **Urban Blue-Collar** | **Skilled Union Tradespeople** |
| Regions | Northeast/  West Coast | Midwest/  Southeast/  Southwest | Southwest/  Southeast/  Midwest | Midwest/  Northwest/  Northeast | Any |
| Town size | Extra-Large/  Mega-city | Small/Large | Small/Large | Extra-Large/  Mega-City | Any |
| Job category | Professional/Sales | Service/  Sales/  Professional | Labor/  Agriculture | Service/  Labor | Crafts/Labor |
| Union member | Either | Either | No | Either | Yes |

# Qualitative and Graphical Analysis

## Unsupervised Clusters

I felt that the most valuable single-variable plots I could create were explorations of the two most prominent political features, *Political Affiliation* and *Voting History*, as well as a cursory glance at the major calculated fields I created, which are focal points of the analysis: *Tech Literacy*, *Media Consumption*, and *Socioeconomic Agility*. These five plots are all contained under a single dashboard, displayed in Figure 1. Interestingly, just over half of customers have a voting history, yet the majority are politically unaffiliated. Tech literacy and media consumption are relatively low, with most customers falling under the low classification. Socioeconomic agility is dominated by the moderate bucket, suggesting that most customers have some degree of economic mobility and resilience, but not a significant amount. Very few customers fall under the high classification for socioeconomic agility, suggesting that, within this group, high amounts of wealth are uncommon.

Figure 2 - Single-variable plots examining the number of individual customers associated with each level of Political Affiliation, Voting History, Tech Literacy, Media Consumption, and Socioeconomic Agility. The X-axis represents the number of customers associated with each of the qualitative levels/labels on the Y-axis.

## Two-Variable Analyses

### Voting History Distribution

Figure 2 illustrates the distribution of voting history compared against tech literacy and socioeconomic agility scores. The overall distribution is mostly clustered around the middle of the socioeconomic agility range and the bottom of the tech literacy range, which is in line with what the single-variable graphs in Figure 1 suggested. No trend immediately jumps out when comparing customers with a voting history to those without a voting history, although there do appear to be more clusters of customers with a voting history on the fringes of the tech literacy and socioeconomic ranges, compared to those without a voting history who seem to be clustered around the low center.

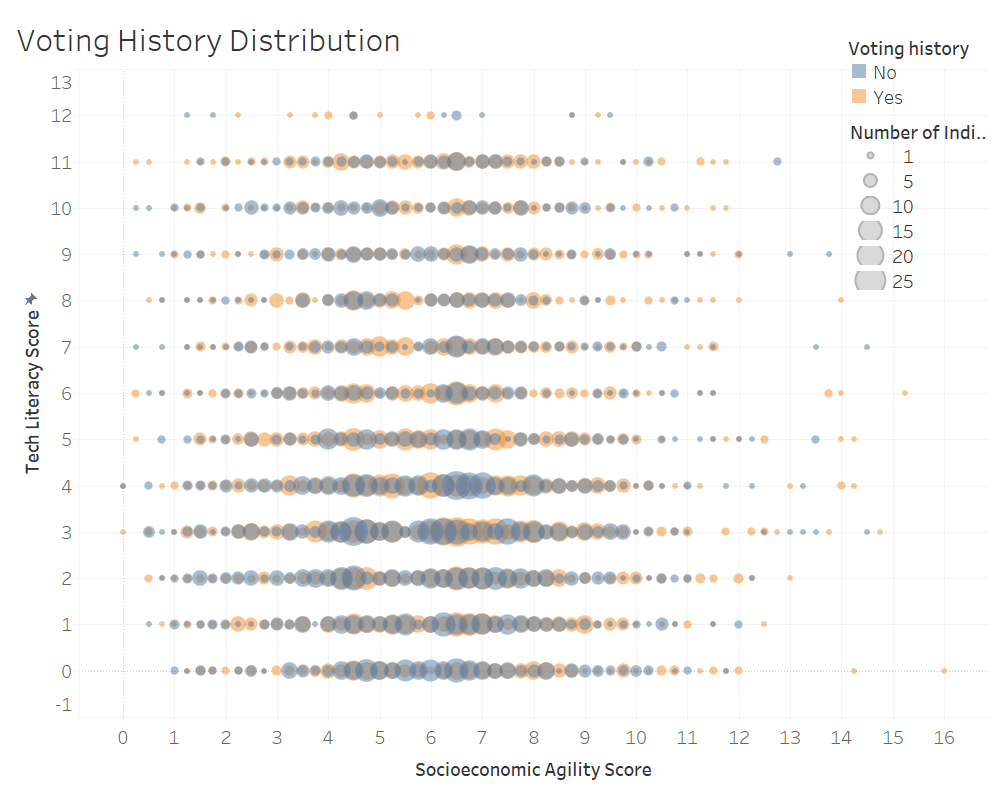


Figure - Distribution of Voting History by Socioeconomic Agility and Tech Literacy Scores. Each bubble represents a cluster of individual customers, with larger bubbles containing more individuals. Blue bubbles represent clusters that do not have a voting history, while yellow/orange represent clusters that do have a voting history.

### Voting History and Socioeconomic Agility

Figure 3 digs deeper than Figure 2 to explore specifically the differences in voting history across the socioeconomic agility spectrum. Like in Figure 2, there are no dramatic trends. There is a slight trend where the number of customers that have a voting history overtake the customers without a voting history at the high end of the socioeconomic agility scores. This may suggest that customers with a higher socioeconomic agility score are more likely to be active voters than those with lower scores, though this is only speculative based on the limited qualitative information available from this analysis.

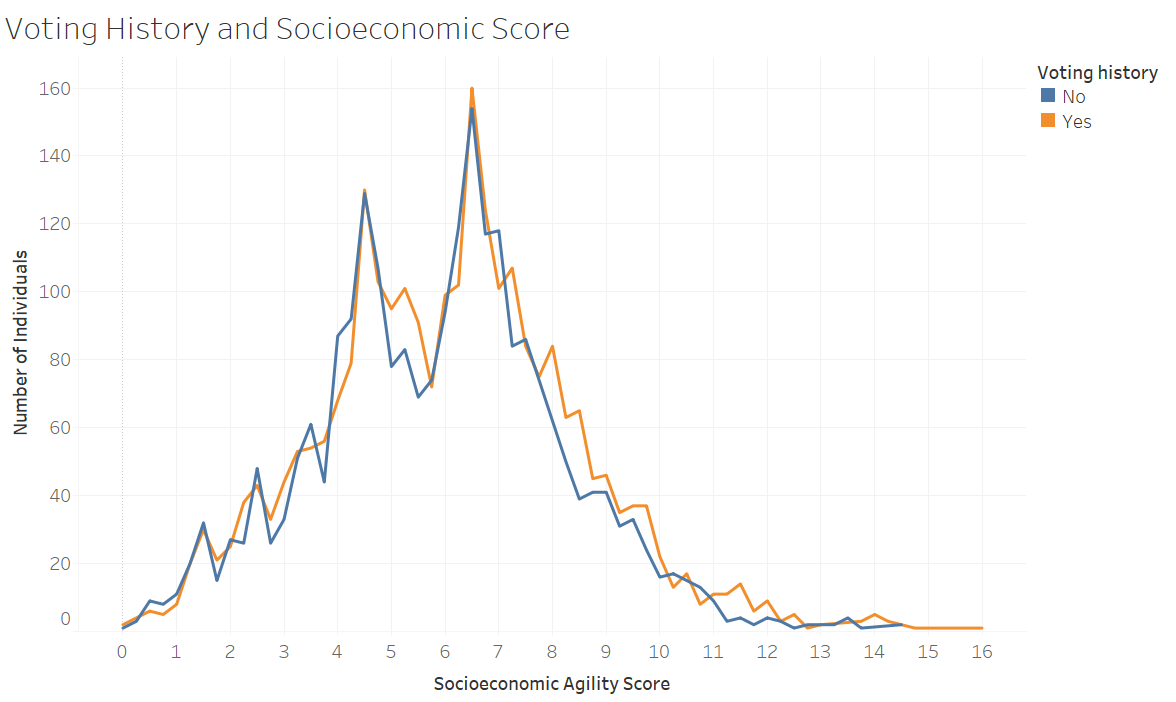


Figure 4 - Relationship Between Voting History and Socioeconomic Agility Score. This line chart illustrates the number of individuals across varying Socioeconomic Agility Scores, comparing those with a voting history (orange) to those without (blue). Both groups peak around scores 4.5 and 6.5, with generally similar distributions, though individuals with voting history appear slightly more concentrated at higher scores.

### Lifestyle Impacts on Voting History

One goal set out by the political consulting firm in the introduction is to examine how lifestyle behaviors that are not obviously related to politics may influence voter behavior. Figure 4 explores that question, examining how the number of pets an individual owns and their level of physical activity relates to their voting history. The first graph in Figure 4 shows nearly no change in the proportion of customers with or without a voting history as the number of pets increases. The second graph in Figure 4 also shows no difference in the proportion of customers with or without a voting history between those with an inactive and active lifestyle. This suggests that neither of these lifestyle behaviors are related to voter engagement, though it is not a conclusive analysis. There are other lifestyle behaviors included in the customer dataset that may prove to be more relevant in future analyses.

### Relationship Between Voting History, Tech Literacy, Media Consumption, and Socioeconomic Agility

Figure 5 examines the relationship between voting history and the average score of the three derived indices, tech literacy, media consumption, and socioeconomic agility. Average media consumption and socioeconomic agility scores do not vary much between customers who have and do not have a voting history, suggesting that these indices are not strongly related to voter engagement. Average tech literacy scores, however, are slightly higher in customers with a voting history, suggesting a possible connection between technological understanding and voter engagement.

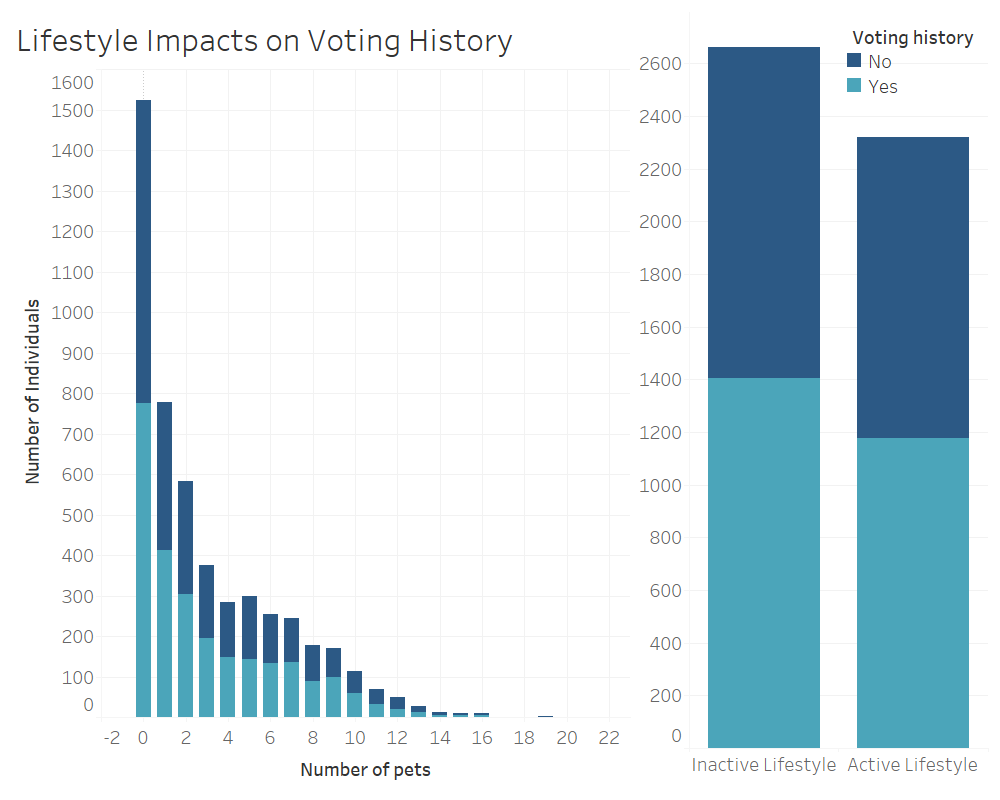


Figure 5 - Impact of Lifestyle Factors on Voting History. The left chart shows the distribution of individuals by number of pets and voting history, with those not voting (dark blue) and those voting (light blue). Voting rates appear to remain mostly consistent across the range of pet ownership. The right chart compares voting history across lifestyle types, showing nearly no difference in the proportion of voters between the inactive and active lifestyle groups.

### Regional Differences in Socioeconomic Agility

Figure 6 examines the differences in socioeconomic agility between the five regions that customers are categorized into. There is a slight difference in the average socioeconomic agility scores between the five regions, with the West Coast and Midwest representing the top end, the Northeast and Southwest representing the middle of the range, and the Southeast representing the bottom. While there is some clear stratification in the map in Figure 6, the actual range is quite tight, ranging only from an average score of about 5.9 to about 6.15. For an index with a range of 16, this range of about 0.25 is quite small, suggesting that there is little difference between regions. Figure 6 serves more as a proof of concept than anything, a demonstration to the political consulting firm that regional differences can be explored. This is an important benchmark as politics in the US are inherently closely linked to geography.

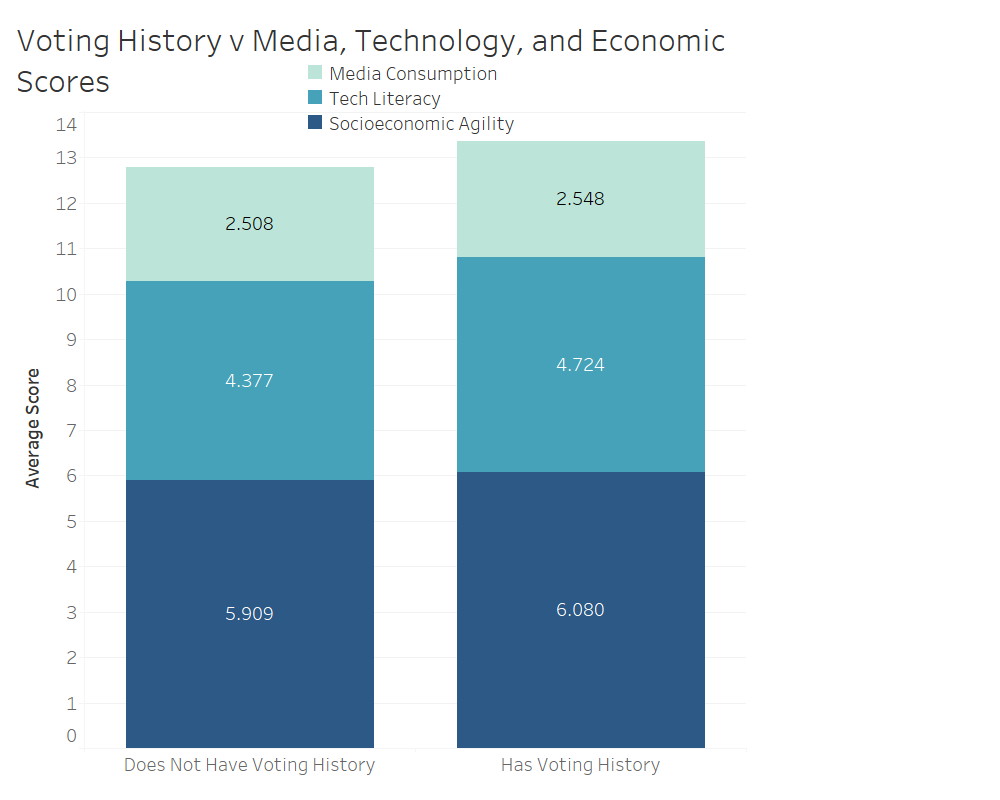


Figure 6 - Relationship between Voting History, Tech Literacy, Media Consumption, and Socioeconomic Agility. The left stacked-bar represents the average scores for each index in customers without a voting history, while the right stacked-bar represents the same in customers without a voting history. Media consumption (sea-foam) and socioeconomic agility (navy) vary only slightly between the two groups of customers. Tech literacy varies more than the other two indices between the two groups of customers, with a slightly higher average tech literacy score in customers with a voting history.

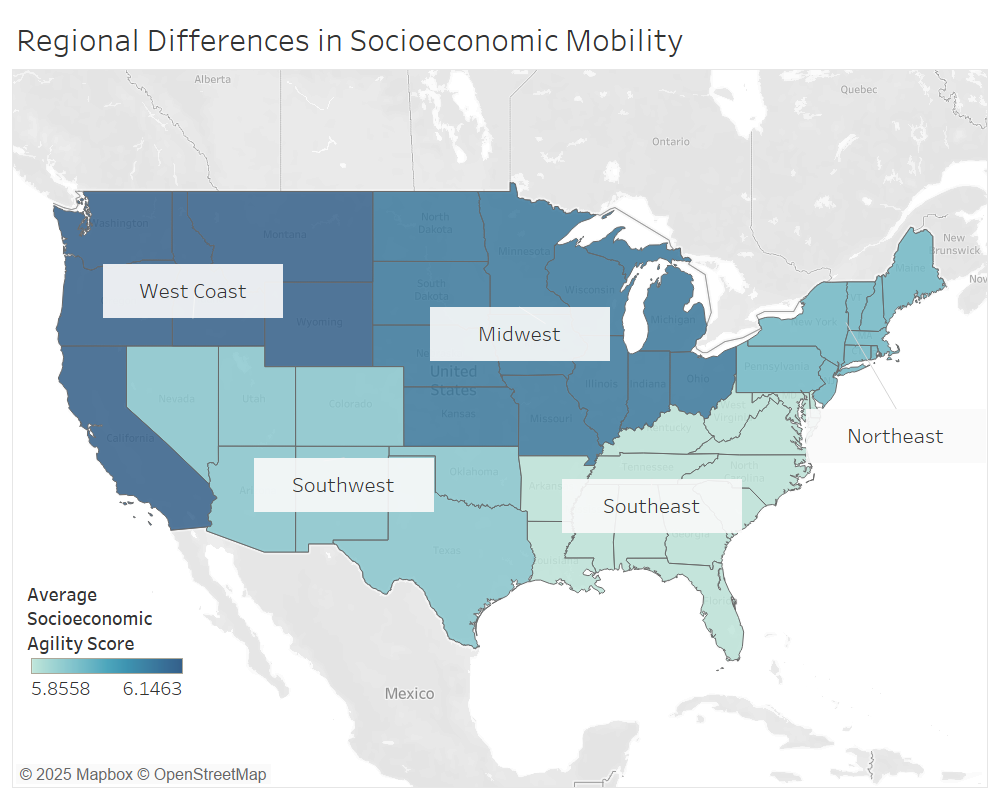


Figure 7 - Regional variation in average Socioeconomic Agility Scores across the contiguous U.S. The West Coast and Midwest show the highest averages, followed by the Northeast and Southwest, with the Southeast lowest. Despite visible regional differences, the overall range is narrow, from 5.8558 to 6.1463. The overall range of this index is 0 to 16, so a range of approximately 0.25 to 0.3 is quite small, suggesting little variation between regions.

# Future Steps

## Further Feature Reduction

The final cleaned and culled dataset is still quite large, as there are multiple features that were involved in the early exploration stages that did not make it to the “final” product (i.e., *Union status, Education, Age*). I anticipate these features being valuable, so they will remain for now, but as the final portion of this work gets underway, I will remove them as I go along to improve efficiency.

## Feature Creation and Tuning

There are still a lot of unused yet likely relevant features (as mentioned above) that remain in the dataset. I intend to look critically at what remains and try to draw connections between them to coerce them down into a score/calculated variable like I did with the derived indices in this analysis. I’d also like to look deeper at the existing indices and see if the weightings need to be tweaked and if any features need to be added or removed.