Predictive Analysis of United States Presidential Elections Using K-Prototype Clustering

Final Report

Submitted to:

Rachel Armstrong-Ceron, Senior Academic Advisor, Engineering Honors

Submitted by:

Honors Capstone Project

Sebastian Munoz, Senior, Industrial & Operations Engineering

Date Submitted:

April 25, 2022

TABLE OF CONTENTS

1. INTRODUCTION

The intended outcome for this work is to create a predictive modeling tool that can be used in order to forecast future major United States political elections. As a baseline goal, this project will manage to identify key clusters of voter classifications as well as determine relevant identifiers that strongly influence the outcome of election results. As a stretch goal, this project will manage to forecast future elections based on random datasets with embedded machine learning principles. I will assess whether this project is successful based on if it meets the Honors College of Engineering standards and is able to identify relevant factors to consider in predicting United States elections.

The significance of this project is that it will contribute political insights to the broader community. It will also show UM Engineering students that they can apply the technical analysis skills learned in courses to non-STEM subjects. This is important since a well-rounded engineer must be able to work with all kinds of data as well as have a general understanding of political scenes within their respective country. My target audience is other College of Engineering students who will be present at the Honors Capstone Design Expo during April.

1.1 Background

Political awareness, although being a crucial skill in the modern world, is something that I personally believe that not a lot of engineering students possess. I only became aware of my own shortcoming in this subject after working on various political science research projects related to other countries and realized that I had no knowledge of the politics that go on within the United States. This project seeks to teach other engineering students some of the processes that are involved in creating a predictive model as well as key classifiers within political elections. In order to do so, this project will create a K-Prototype clustering model. Based on partitioning, this algorithm is an improvement of the K-Means and K-Mode clustering models by handling mixed data types. In other words, it is capable of analyzing both numerical and categorical features to arrive at a more thorough classification of clusters.

That being said, this project aims to create a predictive modeling tool that can be used in order to forecast future major United States political elections. In other words, I will create a K-Prototype clustering model that will predict which way a state will vote during an election. I will optimize the results of this clustering model utilizing various cost estimation methods which will grant me unique insights into the United States political sphere. In order to do so, I will obtain relevant and reliable data from various reputable sources and use data manipulation techniques in order to clean the gathered information. From this, I plan on creating a predictive modeling tool that will forecast future elections based on randomized data. At a minimum, this project will identify key clusters of voter classifications as well as determine relevant identifiers that strongly influence the outcome of election results.

1.2 Key Questions

The primary objective of the project is to create a predictive modeling tool that can be used in order to forecast future major United States political elections. In doing so, this project seeks to expand my understanding of the US political sphere and challenge my technical skills. Some of the research questions that this project seeks to explore are listed on the following page:

Honors Capstone Project Questions

- Is it possible to predict which way a state will vote?
- What are some key factors that affect state election results?
- What population indicators separate clustering groups?

2. METHODS

Various methods of data collection and analysis were executed throughout the course of this project. The following sections describe the collection and analysis process for each method:

2.1 Data Collection

The dataset, which was extracted from United States government databases, included detailed 2020 census data, historical election counts for every election from 1976-2020, and historical election results for every state from 1976-2020. In total, there were a total of 116 unique data variables for each of the 50 states as well as the District of Columbia.

2.2 Data Analysis

Based on partitioning, the K-Prototype clustering algorithm is an improvement of the K-Means and K-Mode clustering models by handling mixed data types. In other words, it is capable of analyzing both numerical and categorical features to arrive at a more thorough classification of clusters. The relevant steps for performing this algorithm are listed below:

1. Read in feature selected key population indicators, comprised of both categorical and numerical data variables
2. Initialize prototype selection by selecting k points as the initial prototypes for k clusters at random, considering the dissimilarity measure
3. Initialize clustering allocation by assigning each object to a cluster which has the minimum difference with its prototype with the previous method
4. Iterate through the algorithm and reallocate data points until the different moves is unchanged, indicating a best result

The Elbow Method was then utilized in order to determine the optimal number of clusters that should be implemented. This is reflected by the K value corresponding to the point where the graph starts to move almost parallel to the X-axis. The initial clustering results, which are depicted on the following page, found the optimal number of clusters to be K = 3.

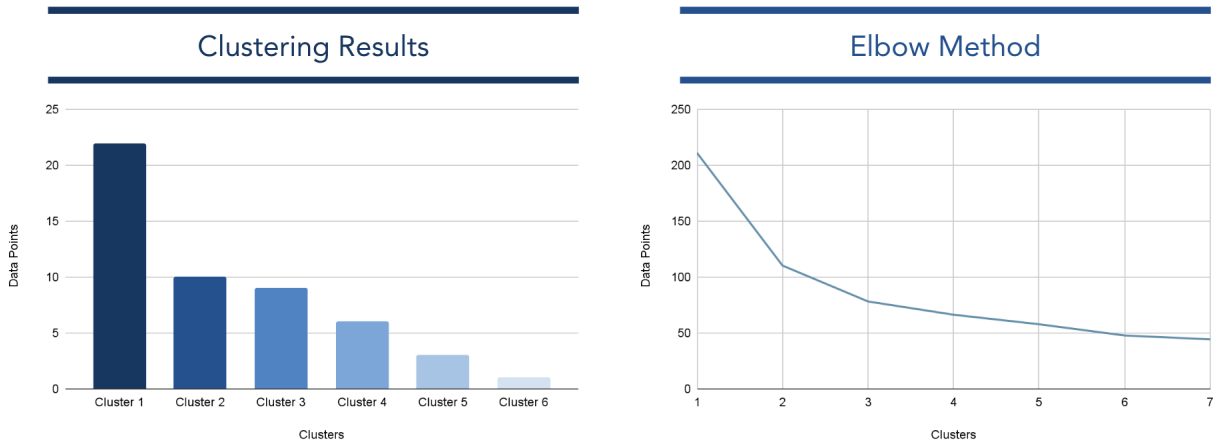

Figure 1. Initial result optimization with cost estimation methods

From this, a weighted forecasting model was then adopted in order to predict the 2024 United States presidential election. In addition to this, a comprehensive analysis of the associated parameters was then performed in order to better understand the clustering results and identify key population indicators.

3. RESULTS

The finalized clustering results along with the elbow method are depicted below:

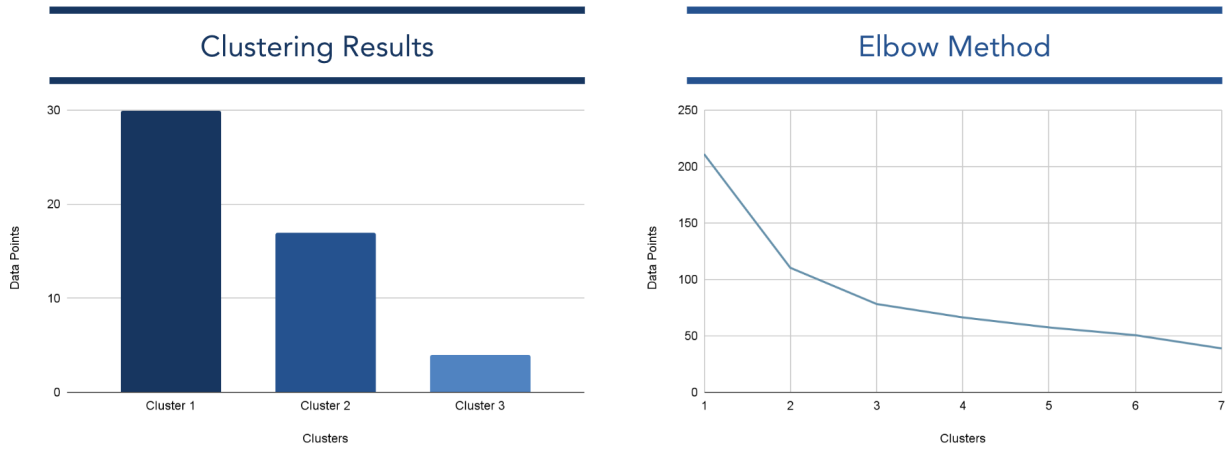

Figure 2. Final result optimization with cost estimation methods

The related state classifications, in which each state is assigned to a unique clustering identifier based on the dissimilarity measure, are shown on the following page:

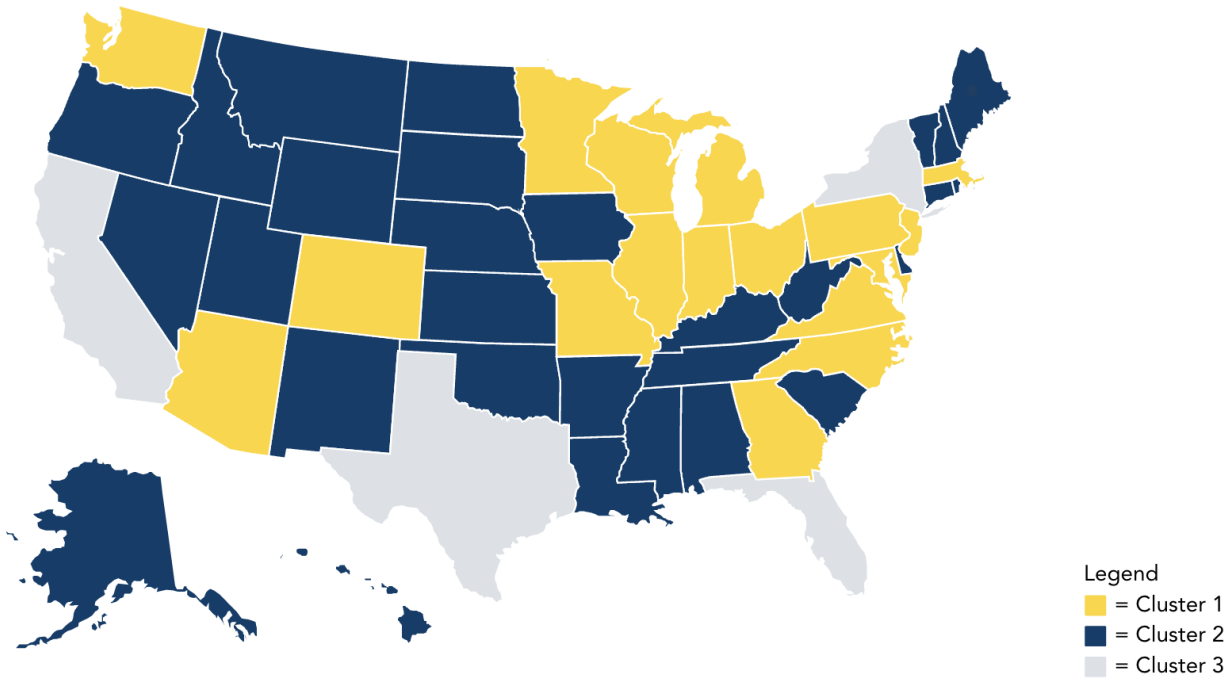

Figure 3. Clustering breakdown by state

The subsequent K-Prototype clustering-based, forecasting model prediction for the 2024 United States presidential election, broken down by state election results for a particular voting party, is depicted in the following graphic:

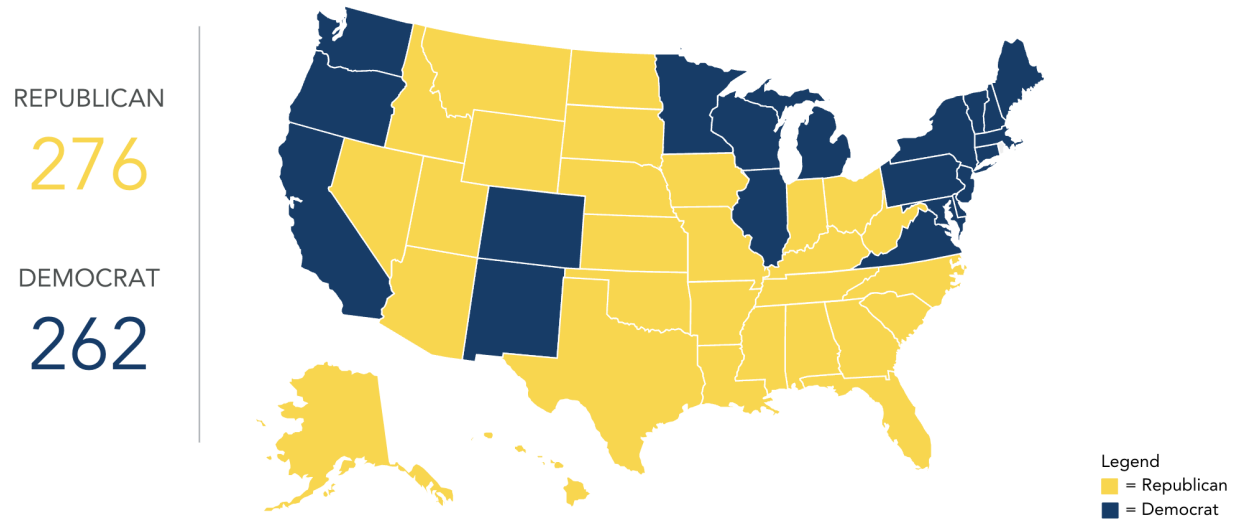

Figure 4. Forecasted 2024 Electoral College results

The numbers shown on the previous page reflect the final predicted outcome of the Electoral College process, which consists of 538 electors divided among all 50 states and the District of Columbia based on the same number of electors as they have in its Congressional delegation. A majority of 270 votes is required to win the election. Therefore, it is possible to conclude that the Republican Party will win the 2024 United States presidential election.

4. ASSESSMENT

Although this project was able to successfully develop a predictive modeling tool that can be used in order to forecast future US presidential elections, it is also important to understand and analyze the various characteristics that comprise the different clustering groups. That being said, a detailed breakdown of several key factors is provided within the following sections.

4.1 Voting Patterns

The results of the cluster-based, weighted forecasting model for the 2024 United States presidential election is summarized in the table below. It is interesting to note that cluster 1 constitutes the majority of Republican-voting states whereas cluster 2 represents the majority of Democrat-voting states.

Party	Cluster 1	Cluster 2	Cluster 3	Total
Democrat	7	11	2	20
Republican	23	6	2	31
Total	30	17	4	51

Figure 5. Clustering election results

These results can be attributed to the clustering algorithm which determines a states allocation by considering the dissimilarity measure for each associated variable. With the initial analysis complete, this project then began to asses the key population indicators that differentiate clusters.

4.2 Unique Variables

Various analysis methods were then completed across several relevant population indicators, granting many unique insights. Some of the interesting variable deep-dives include the native born population rate, bilingual rate, poverty rate, median household income, occupational breakdown, and cluster industry breakdown. These are further explored below.

Population Breakdown

An in-depth analysis was performed regarding various characteristics related to population amongst the various clustering groups. The summarized results are shown below:

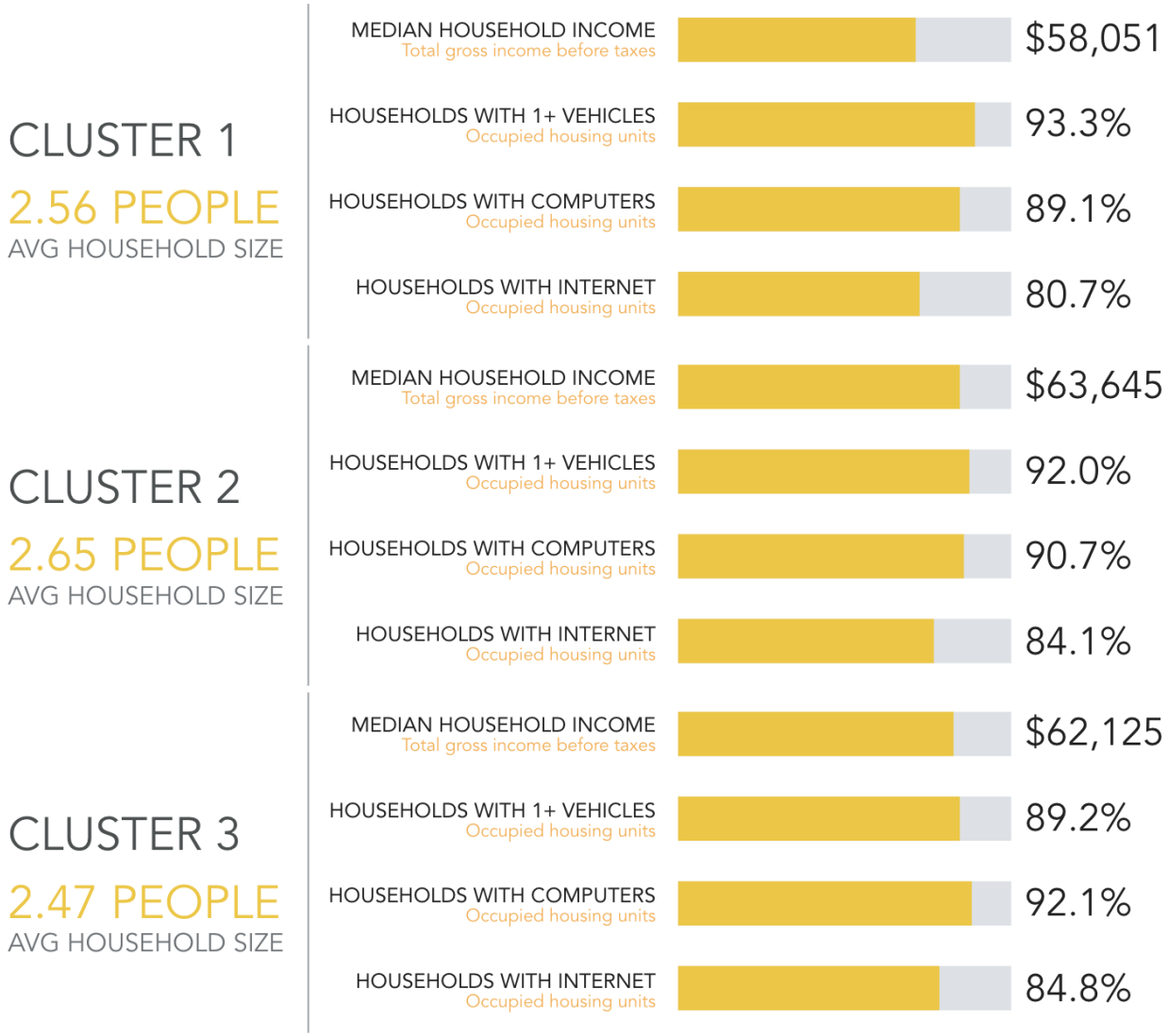

Figure 6. Population breakdown analysis results

The interesting aspect about this analysis is that it informs us that Cluster 1 tends to have the lowest percentage of households with internet access; Cluster 2 tends to have the highest median household income; and Cluster 3 tends to have the lowest average household size.

Housing Breakdown

An in-depth analysis was performed regarding various characteristics related to housing amongst the various clustering groups. The summarized results are shown below:

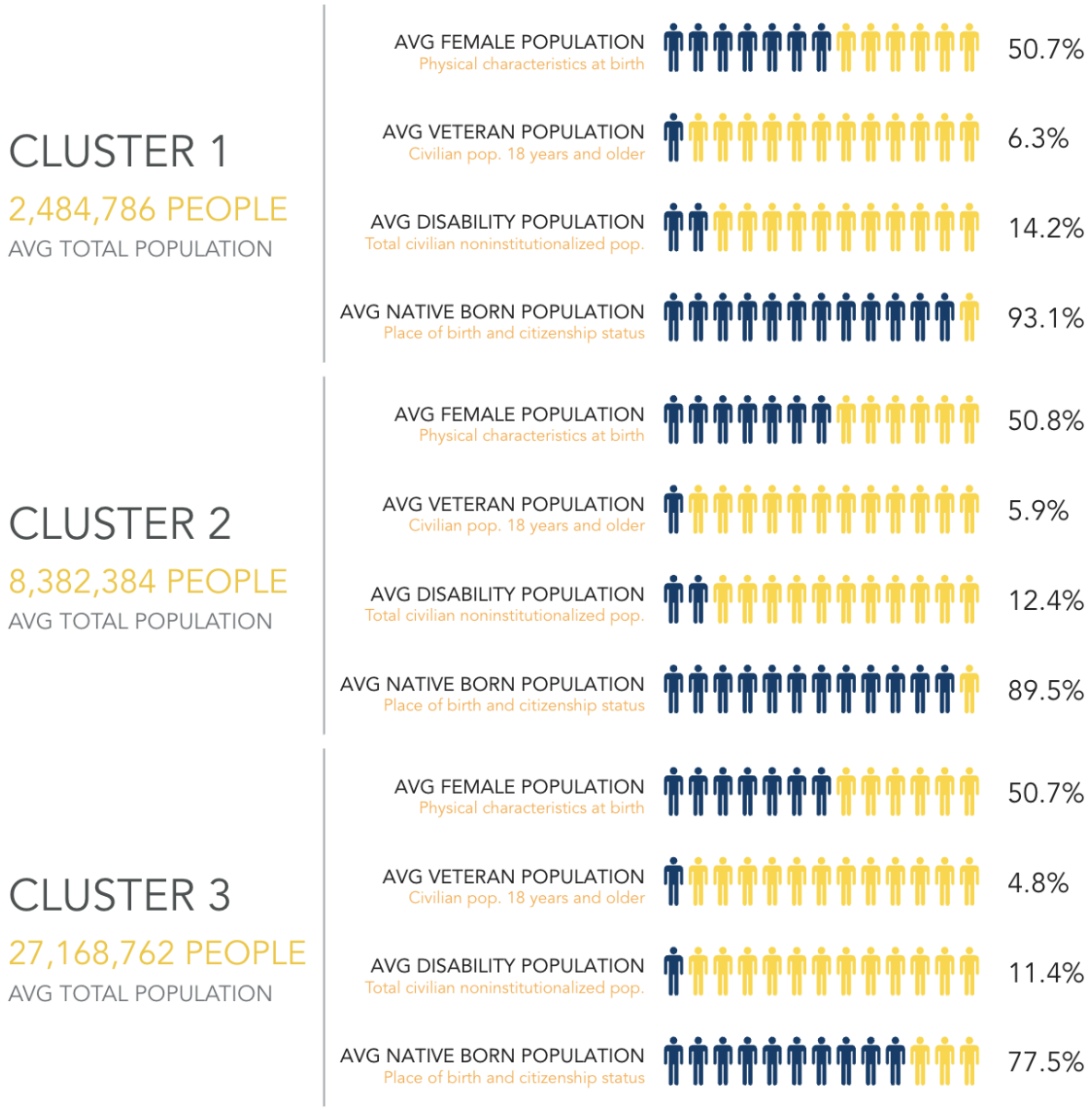

Figure 7. Housing breakdown analysis results

The interesting aspect about this analysis is that it informs us that Cluster 1 tends to have the highest average native born population; Cluster 2 tends to have the largest female population; and Cluster 3 tends to have the largest average total population.

Educational Attainment

An in-depth analysis was performed regarding various characteristics related to educational attainment for the population over 25 years of age amongst the various clustering groups. The summarized results are shown below:

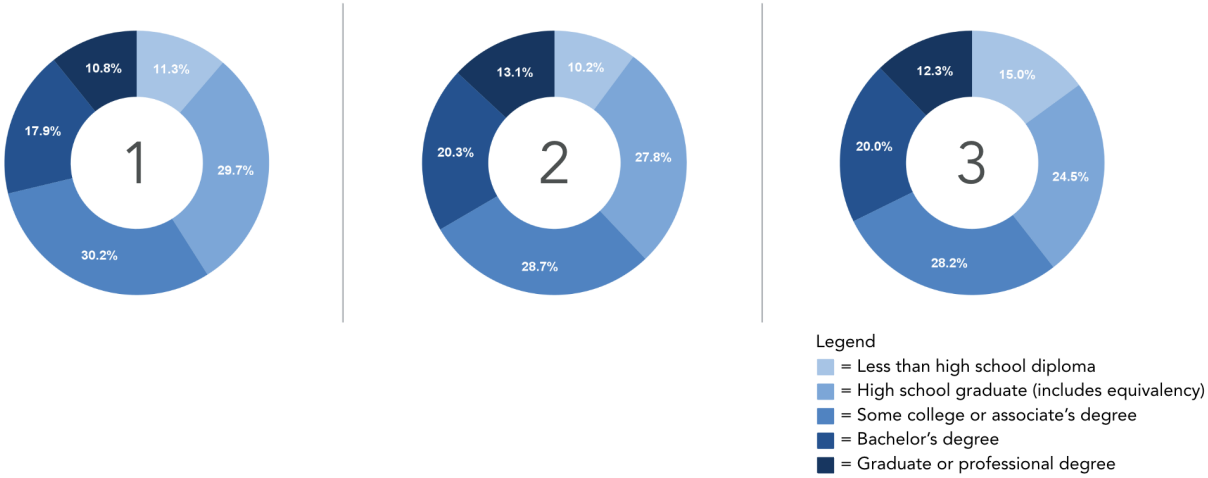

Figure 8. Educational attainment analysis results

The interesting aspect about this analysis is that it informs us that Cluster 1 tends to have the lowest rates of graduate attainment; Cluster 2 tends to have the highest rates of some college and above; and Cluster 3 tends to have the highest rates of less than a high school diploma.

Occupational Breakdown

An in-depth analysis was performed regarding various characteristics related to occupational position for the population over 16 years of age amongst the various clustering groups. The summarized results are shown on the following page:

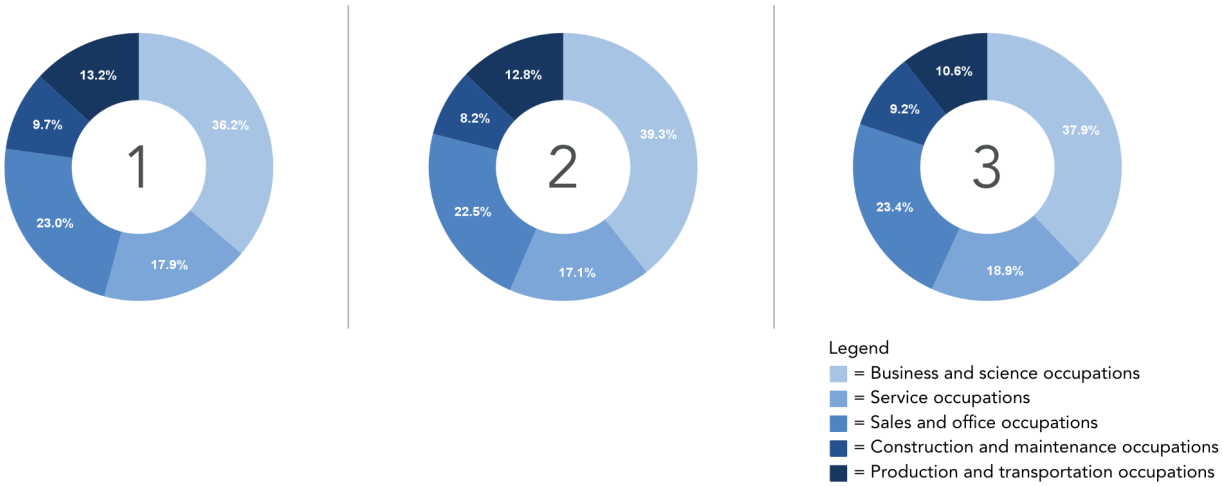

Figure 9. Occupational breakdown analysis results

The interesting aspect about this analysis is that it informs us that Cluster 1 tends to have the largest construction and production related roles; Cluster 2 tends to have the largest business and science related roles; and Cluster 3 tends to have the largest sales and office roles.

Health Insurance Coverage

An in-depth analysis was performed regarding various characteristics related to health insurance coverage amongst the various clustering groups. The summarized results are shown below:

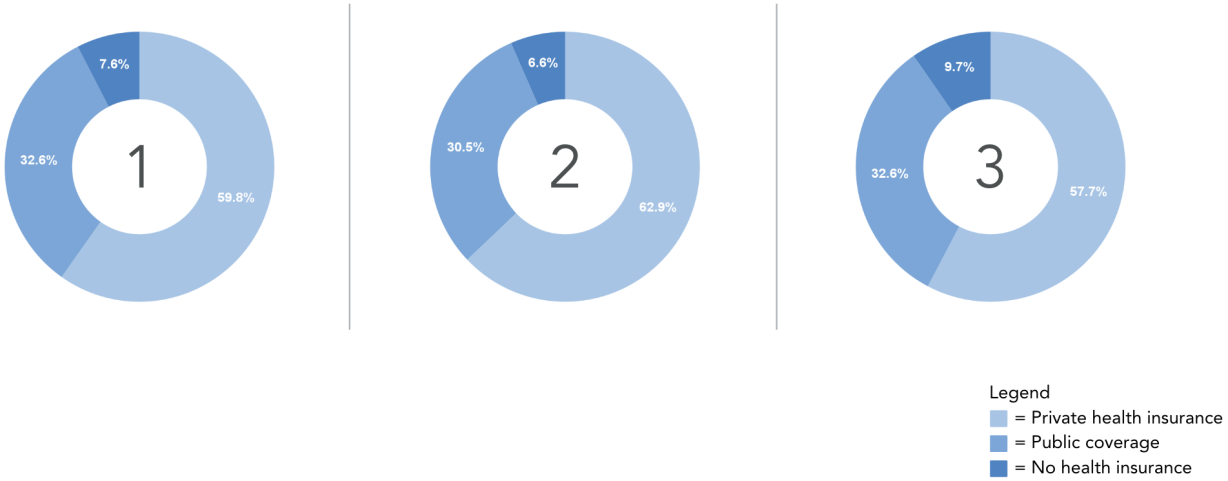

Figure 10. Health insurance coverage analysis results

The interesting aspect about this analysis is that it informs us that Cluster 1 tends to have one of the largest public health insured population; Cluster 2 tends to have the most insured population; and Cluster 3 tends to have the lowets private health insured population.

Worker Class

An in-depth analysis was performed regarding various characteristics related to worker class amongst the various clustering groups. The summarized results are shown below:

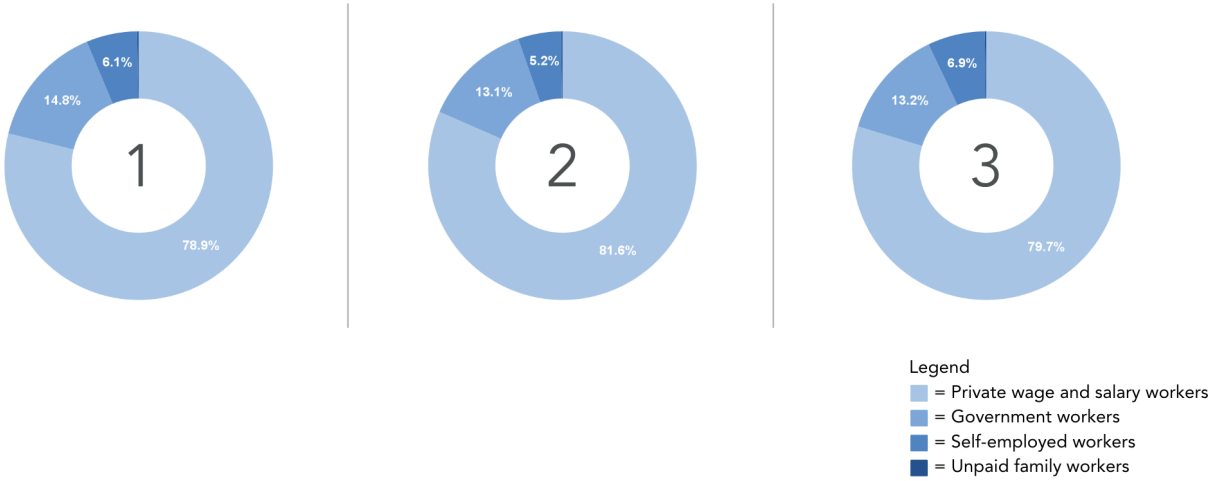

Figure 11. Worker class analysis results

The interesting aspect about this analysis is that it informs us that Cluster 1 tends to have the highest percentage of government related workers; Cluster 2 tends to have the highest percentage of private wage related workers; and Cluster 3 tends to have the highest percentage of self-employed workers.

Industry Breakdown

An in-depth analysis was performed regarding various characteristics related to industry amongst the various clustering groups. The summarized results are shown on the following page:

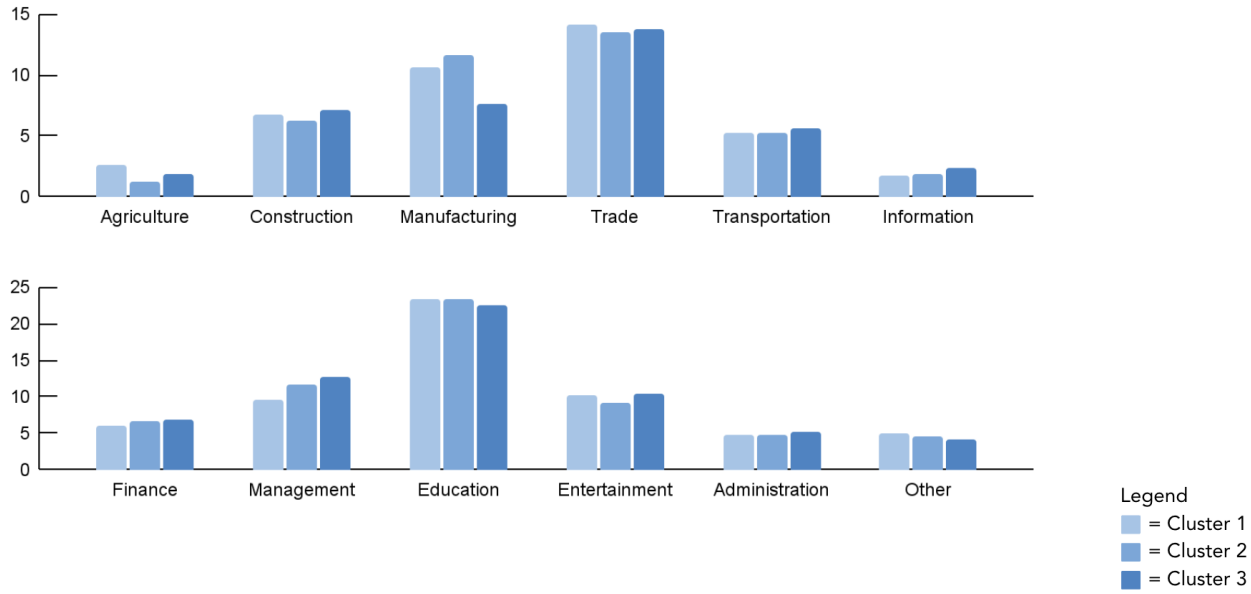

Figure 12. Industry breakdown analysis results

The interesting aspect about this analysis is that it informs us that Cluster 1 tends to have the highest percentage of agriculture and trade related industries; Cluster 2 tends to have the highest percentage manufacturing related industries; and Cluster 3 tends to have the highest percentage of management and information related industries.

Interesting Indicators

An individual analysis was performed regarding several interesting characteristics amongst the various clustering groups. The summarized results are shown on the following page.

The interesting aspect about this analysis is that it informs us that Cluster 1 tends to have the lowest bilingual rates; Cluster 2 tends to have the lowest poverty rate; and Cluster 3 tends to have the youngest median age.

	Cluster 1	Cluster 2	Cluster 3
Median Age	38.3 years	38.56 years	37.96 years
Bilingual Rate	11.33 %	15.79 %	36.79 %
Jobless Rate	5.07 %	4.98 %	5.50 %
Poverty Rate	13.54 %	12.19 %	14.03 %

Figure 13. Interesting indicators analysis results

It is interesting to note that the descriptions of the various clustering groups tend to align very similarly with the current population perception of the respective electoral parties. Cluster 1 tended to be Republican-leaning states, constituted by majority American-born citizens that have the lowest median household income. Cluster 2 tended to be Democratic-leaning states, constituted by the largest rates of educational attainment and smallest poverty rate. Cluster 3 is arguably the most interesting group, as the K-Prototype clustering method managed to identify the states of California, Florida, New York, and Texas as deserving of a separate cluster. These states, although divided politically, tended to be the youngest, most foreign-born population that are typical of the stature of these globally important states. The summarized synthesis of the various indicator variables is further explored in the conclusion section.

5. CONCLUSIONS

Ultimately, the aforementioned analysis helps in understanding the forecasted presidential election results and confirms the validity of the K-Prototype clustering-based predictive modeling tool. The intended outcome was successful in producing a weighted forecasting model that relied on the results of the clustering algorithm to forecast the 2024 United States

presidential election. Although there are many other predictive models that attempt to solve this fascinating and relevant problem, my original goal was simply to create my own model. In doing so, I was also successful in creating a tool that could predict future elections, as the results are compounding and forecasted if required.

Every state was partitioned into clustering groups based on unique characteristics that aligned with those of other states, and these clusters were then reflective of these population indicators. In doing so, this project was able to successfully answer all of the initial questions. It is possible to predict which way a state will vote; there are various key factors that affect state election results; and there are interesting population indicators that separate clustering groups.

On a personal note, I am grateful for the opportunity to have worked on this Honors Capstone project thanks to the tireless efforts of Edgar Franco-Vicanco, Assistant Professor of Political Science, and the endless aid of Rachel Armstrong-Ceron, Senior Academic advisor. Go Blue!