CSC 369 Project Report

Team GPT: Alexander Nunn, Nicholas Hotelling, Ethan Trantalis, Justin Kopcinski

For our project, we wanted to analyze economic activity at the zip code level across the United States using publicly available demographic and business datasets. Our objective was to identify areas with high potential for new business development by examining economic density and purchasing power. While our primary goal was to explore these regions in a general sense, we hope to apply this same methodology to specific business industries in the future.

Our analysis involved combining two major datasets found online. The first came from the U.S. Census Bureau's County Business Patterns (CBP), which lists the number of business establishments within each zip code, categorized by a NAICS (North American Industry Classification System) code. Although the NAICS code identifies specific industries, we had decided to exclude that from our analysis as we felt it was too specific to include in our first experiment. Instead, we used the total number of establishments per zip code as a general indicator of business density.

The second dataset was obtained from the Inter-university Consortium for Political and Social Research (ICPSR), which includes demographic information by zip code. From this, we extracted the total population and median income values for each zip code. These two variables were chosen because they represent key dimensions of a region's economic environment. This includes how many people live in an area, and what the average purchasing power might be.

We created a simple but meaningful metric called establishments per capita, calculated by dividing the total number of establishments by the total population in each zip code. This provided a normalized measure of business density, accounting for variations in population size.

We then plotted this value against the median income per zip code, forming a two-dimensional dataset that represents economic activity potential.

To analyze this data, we applied k-means clustering to segment zip codes into distinct groups based on their economic profiles. This approach allowed us to visualize clusters of zip codes with similar combinations of business density and income level. By selecting a relatively high value for k, we were able to capture more nuanced economic patterns across the country. One possible use case for this clustering is to identify optimal locations for opening new establishments. This experiment can be further improved if we were to filter by industry type using NAICS codes, where we would plan to complete this in future iterations.

Ultimately, our methodology provides a replicable framework for identifying zip codes with strong economic potential. By leveraging publicly available data and simple statistical techniques, we can help inform business expansion strategies or community investment planning.

For our results, we were able to run a smaller subset of our dataset. Our Scala code that we had created seemed to be inefficient and would take way too long with our original dataset with over 25 thousands individual lines of data. We cut the data down to 50 data points and three centroids so we were able to get a result. We had also created a version of our k-means clustering in Python with 3, 5 and 8 clusters, to compare to our results in Scala, and even though we ran the full dataset in Python compared to only 50 datapoints and 3 centroids, the results look very similar to one another. The position coordinates that were returned from the algorithms were very similar in geolocation, which means even with a limited dataset in the Scala code, the results would be extremely similar.