Mapping Economic Disparities

(COMP3125 Individual Project)

Daren Yun  
*Data Science Fundamentals COMP3125*

*Abstract*— The purpose of this research is to examine the income inequality amongst geographic terrains and its involvement in employment, economic development, and overall living costs across nationwide regions. Through quantitative analysis of geography and economy, it will reveal disparities and trends that can be studied to infer predictions as well as address any potential critical issues.

Keywords— Geographic inequality, Household income, Economic growth, Cost of living

# Introduction

Inequality in income stems from one of many variables, geography, and this is key to understanding the socioeconomic dynamics of our regions. I would like to funnel research primarily into analyzing these disparities and trends in overall household income within and across a few regions that have been influenced on its geographic attributes to address five key questions: (1) What kind of income inequality exists within a region due to geography? (2) Why does the median household income vary by region? (3) How is employment related to household income compared to other regions? (4) Which regions show the most significant economic growth or economic decline over the years? (5) How does cost of living and household income compare across localities?

These questions come across as vital to me because of the experiences and discussions amongst peers and friends at different parts of the country that experience their own economic policies, resource allocation, and social welfare programs, different than my own. It leaves a big hole in understanding just how much even one physically different attribute can play such an immense influence over a region.

My approach involves quantitative methods through matplotlib or pandas to analyze the income data [1], employment rates, and key economic indicators [2] over a set period of time. It also involves using geographic information systems to conclude a visual pattern to better identify any disparities or trends.

# Datasets

## Source of dataset

The datasets used in my analysis were sourced exclusively from official .gov websites to ensure a high base level of reliability as well as credibility with two sources coming from .org websites. The excel sheets containing the data were downloaded from agencies ranging from the U.S. Census Bureau, Bureau of Economic Analysis (BEA), to the World Bank Group to establish credibility for their comprehensive and accurate data collection methodologies from government-administered surveys and records to economic modeling such as nationwide household surveys, labor force surveys, as well as standardized economic models. Most of the sources I collected data from are from recent years (2023, 2024) apart from the World Bank Group source being collective from 1963 until 2022 to better calculate the Gini Index. Nonetheless, these sources are verified for accuracy and maintained by their respective agencies, which allows me to ensure consistency in data analysis.

## Character of the datasets

|  |  |  |
| --- | --- | --- |
| Dataset Types | Parameters | Units |
| Census | Median income | USD, Ages, Percentage |
| Employment | Region, Weekly Wages | USD |
| GDP | Region, GDP | USD, Percentage |
| Cost of Living Index | Region, Living Index | Index Value (0-200) |
| Region Classification | Region, Classification | Categorical |
| Income Inequality | Region, Gini Coefficient | Index Value (0-50) |

# Methodology

## Dual Metric Visualization across States

The goal is to obtain visualization of regional disparities involving income and more specifically, its weekly wages using GeoPandas, and Matplotlib

My primary assumption for going with this method is that it’ll be able to express a key data point to be used later for calculations depending on concluding results but also for the advantages of highlighting the economic inequalities across the U.S. and the trends associated with it. However, through this method, it doesn’t require too much of in-depth statistical analysis.

## Clustering

The purpose is to segment the regions into their respective clusters using K-Means, by finding similar regions based on shared economic traits or rather come to the solution each region can ultimately be grouped altogether as one.

The main advantage will provide me with objective groupings for further analysis, however the disadvantage relies on if there are enough number of clusters (k) to comfortably continue.

The main Python tool to utilize this method will be coming from “*sklearn.cluster.KMeans*”.

# Results

## Income to Wage Affordability Ratio Visualization

A blue and black graph

Description automatically generated with medium confidence

*Bar graph generated from calculated analysis of income and weekly wages throughout the U.S states*

This graph comes from the calculated formula of normalized cost of living to the normalized mean weekly wages to determine a ratio that would define each state’s ratio of being affordable for the average citizen in their respective state. It’s important to note that the states with scores closer to 0.0 would mean that they’re more affordable as their wages better resemble their cost of living. This means the latter when looking at states with a much higher calculated value of affordability ratio, that are less affordable, indicative of living expenses that cannot be kept up with given their weekly wages.

## Clustering Results

To first understand the methodology of the clustering scatter plot, it is crucial to understand where the number of optimal clusters to sort the states into, comes from. There is an Elbow Method [3] that identifies a point where the rate of decrease within a cluster’s sum of squares changes sharply indicating that any more clusters after that point, would be ineffective and yield diminishing results. This “elbow” point is what determines the most effective clusters to utilize for a dataset.

A graph with a dotted line

Description automatically generated

*Graph generated from calculated analysis of sampling a range of 1-10 clusters*

Given the answer of “3” being the most effective number to divide the states amongst, all that is left is to make a scatter plot visualizing the clusters with black centroids as “X”.

A screen shot of a graph

Description automatically generated

*Scatter plot generated from taking the states and dividing up amongst 3 clusters*

With this graph, we can see the keys indicating Clusters 0, 1, and 2 with centroids representing the average position of states within each cluster. Cluster 0 is indicative of states with lower affordability ratios and moderate to high weekly wages, boasting the “good” states. Cluster 1 is indicative of high affordability ratios but with mediocre weekly wages, revealing states with high costs of living relative to their wages. Lastly, Cluster 2 is indicative of state(s) with very low wages but is balanced out by lower costs of living as well to offset the low wages.

# Discussion

The weakness that I’m unsatisfied with is personally myself. I hadn’t checked the submission date for 2 weeks, confident in hearing a classmate say that it was due on the 11th. I had a regression model planned to better show efforts of the learnings of data science and wanted to seriously apply it to this to show off the fruits of my labor, but it didn’t go as planned. For future work, I will be keeping an active reminder in front of my face at all times of due dates. As for the data, I would’ve also enjoyed figuring out a way to implement a heatmap that colormaps the states according to whatever index range I’d customize it for. I’d have some kind of basis heatmap to tinker with as opposed to being unprepared. As for the methodologies, while it was fun to implement the graphs I had, I wanted more time to be able to implement more key factors into them rather than stick with three of my sources and see how that possibly plays a role in affecting the discoveries of economic inequalities such as the Gini coefficient or the GDP of each state.

# Conclusion

The important results that I gathered from this relate to possibly State policy regulations and the possible negative effects on individual citizens. The Affordability Ratio that I had determined revealed many states with a low ratio, most likely indicating a smoother life experience without dramatic highs and lows of the economic policies, but it also determined the states with high ratios, pointing to the conclusion that quality of life isn’t as best as it can be as the wages earned still do not mean anything to the costs of living associated within their state. This could mean several wide variables or explanations, however one that it does entail is that the average citizen would have to work two jobs to better combat the living costs in their respective state based on 40-hour work weeks, which through that, can cause ripple effects of large-scale reactions. Further real-world effects can be large-scale migrations to states with much more manageable living costs relative to their paid wages or higher affordability ratio states having to dramatically emphasize workforce shortages or wealth inequality relative to housing costs and stagnant wages.

Addressing the five initial questions that sparked my interest in this data science project, is as follows accordingly. Income inequality varied significantly due to both geographic features, cost of living, and urbanization which through the clustering methodology revealed clear disparities that high-income urban areas such as California are much less affordable for the average citizen whereas rural states boasted a comfortable lifestyle that struggles with wage growth however is much more affordable given their costs of living.

The median household income varies from different economic structures for either rural or urban counties and associated living costs. Regions with robust industries such as Massachusetts or California will have reported higher incomes but fail at offsetting them with proportional affordability. Regions without robust industries and rather agricultural such as Oklahoma, have lower median household incomes but lower costs of living keep their ratio very balanced.

Employment relating to household income does not signify affordability and it is evident as such in Hawaii or New York that despite their employment rates being very high, their high affordability ratios and disproportionate living costs often negate that incredibly.

Regions showing the most significant economic growth can be seen within states like Florida which benefits from population growth, living costs, and multiple industries. However regions showing decline or stagnation such as Illinois indicate that their economic hubs struggle with the challenges of affordability due to its uneven income distribution.

Lastly, the costs of living and household income compared to those across localities are very distinct and sharp. New York boasting high incomes are among the states of least affordability due to policies and living conditions that are exorbitant whereas Georgia’s wages reflect a better ratio along with living expenses.

Overall, it signifies that essentially half of the U.S is encroaching on the limit of the supposed affordability ratio that if they cross, they’ll be living in a reality that their working wages can no longer support their costs of living, which unfortunately is already the case for a few states that deal with its own wide variety of variables and policies along with overpopulation.

##### References

1. “State Minimum Wage Laws,” *DOL*. <https://www.dol.gov/agencies/whd/minimum-wage/state#content>
2. The World Bank, “United States | data,” *Worldbank.org*, 2022. <https://data.worldbank.org/country/united-states?view=chart>
3. A. Gupta, “Elbow Method for optimal value of k in KMeans,” *GeeksforGeeks*, Jun. 06, 2019. <https://www.geeksforgeeks.org/elbow-method-for-optimal-value-of-k-in-kmeans/>
4. *Census.gov*, 2024. <https://www2.census.gov/programs-surveys/demo/tables/p60/282/tableA1.xlsx>
5. *Bea.gov*, 2024. <https://www.bea.gov/sites/default/files/2024-09/stgdppi2q24.xlsx> (accessed Dec. 08, 2024)
6. Bureau of Economic Analysis, “GDP by State | U.S. Bureau of Economic Analysis (BEA),” *Bea.gov*, 2024. <https://www.bea.gov/data/gdp/gdp-state>
7. “Cost Of Living By State 2024,” *Data Pandas*, 2024. <https://www.datapandas.org/ranking/cost-of-living-by-state#table> (accessed Dec. 08, 2024).
8. “USDA ERS - Rural-Urban Continuum Codes,” *www.ers.usda.gov*. <https://www.ers.usda.gov/data-products/rural-urban-continuum-codes/>
9. “World Bank Open Data,” *World Bank Open Data*, 2017. <https://data.worldbank.org/indicator/SI.POV.GINI?end=2022&locations=US&start=1963&view=chart>