Shreya Khettry

Matthew Rattigan

CICS 397A

20th December, 2022

FINAL PROJECT

INTRODUCTION AND BACKGROUND

Social capital – the strength of our relationships and communities – has been shown to play an important role in outcomes ranging from income to health. Using privacy-protected data on 21 billion friendships from Facebook, Raj Chetty, Matthew 0. Jackson, Theresa Kuchler et. al, measure three types of social capital in each neighborhood, high school, and college in the United States:

- **Economic Connectedness**: The degree to which low-income and high-income people are friends with each other
- **Cohesiveness**: The degree to which social networks are fragmented into cliques
- **Civic Engagement**: Rates of volunteering and participation in community organizations.

The creators used data of Facebook users between the ages of 25 and 44, who reside in the United States, were active on the Facebook platform at least once in the prior 30 days, have at least 100 U.S.-based Facebook friends, and have a non-missing residential ZIP code as of May 28, 2022 to create their database and various other statistical columns that I will be talking about further.

DATA DESCRIPTION, DESIGN CHOICES, AND DATA PRE-PROCESSING

When I was going through the codebook support document, I noticed that everything in the document is divided on the basis of the three types of social capitals and I decided that I will also be structuring my analysis in a similar manner and explore my data based on the different social capitals at the county level. An exploration question that I would like to answer is, “Are the different variables for the different social capital types interconnected or not?” My initial thoughts are that most of them should be but I will explore this question throughout my project as my analysis is primarily based off of this.

For my analysis I decided to use the ‘County-Level Data’ file (I will upload the csv with my submission) . It has 26 columns and 3089 rows. The data is divided into 4 groups and I will be expanding on each group and will describe the variables that I will be using.

1) County Identifiers and Population Variables

   There are 4 variables and I use 3:

   - ‘county_name’: Name of the county and state.
   - ‘num_below_p50’: Number of children with below-national-median parental household income
   - ‘pop2018’: Population in 2018.

2) County Connectedness Statistics

There are 18 variables and I use 3:

- ‘ec_county’: Baseline definition of economic connectedness: two times the share of high-SES friends among low-SES individuals, averaged over all low-SES individuals in the county.
- ‘child_ec_county’: Childhood economic connectedness: two times the share of high parental-SES friends among low-parental-SES individuals averaged over all low-parental-SES individuals in the county, calculated using only individuals’ high school friends
- ‘ec_high_county’: Economic connectedness for high-SES individuals: two times the share of high-SES friends among high-SES individuals, averaged over all high-SES individuals in the county

3) County Cohesiveness Statistics

There are 2 variables and I use both:

- ‘clustering_county’: The average fraction of an individual’s friend pairs who are also friends with each other.
- ‘support_ratio_county’: The proportion of within-county friendships where the pair of friends share a third mutual friend within the same county.

4) County Civic Engagement Statistics

There are 2 variables and I use both:

- ‘volunteering_rate_county’: The percentage of Facebook users who are members of a group which is predicted to be about ‘volunteering’ or ‘activism’ based on group title and other group characteristics
- ‘civic_organizations_county’: The number of Facebook Pages predicted to be “Public Good” pages based on page title, category, and other page characteristics, per 1000 users in the county.

I decided to use these columns because based on their descriptions I felt that there could be some relationship between them that I could explore. Another factor that played into deciding which columns to use were the social capital type descriptions and I felt that the columns I chose to use in my exploration were a good representation of the definition. To reduce further complexities, I decided to narrow down my dataset to represent counties of 3 states: Massachusetts (east), Oklahoma (mid), and Washington (west). I decided to remove any rows that had null values so that I could limit the errors in my analysis. I also created two new columns by splitting the ‘county_name’ into ‘countyName’ and ‘state_name’ for ease of analysis.

I used JuPyter Notebooks for my exploration and the code can be ran by clicking the run button on each cell.

TASKS AND INSIGHTS

My initial idea was to perform 3 tasks: Visualization, Clustering, and Classification. I did include an Association Map however I am unsure if that would count as a complete task or not.

Visualization Tasks

My main objective with the Visualization Tasks were to do an Exploration of the data and see if I am able to obtain any insights in relation to my research question. My tasks:

Task 1:

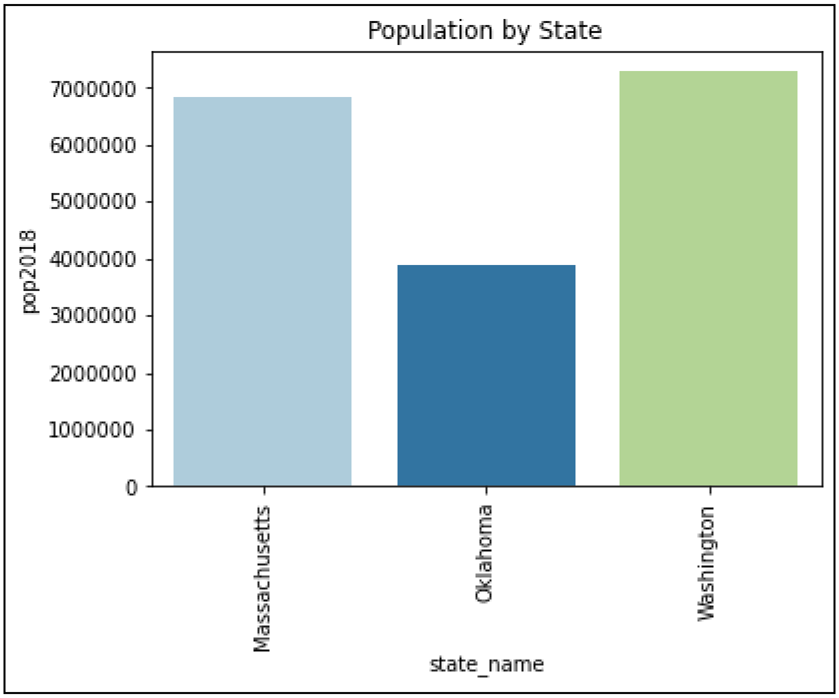

Figure 1: Population by State

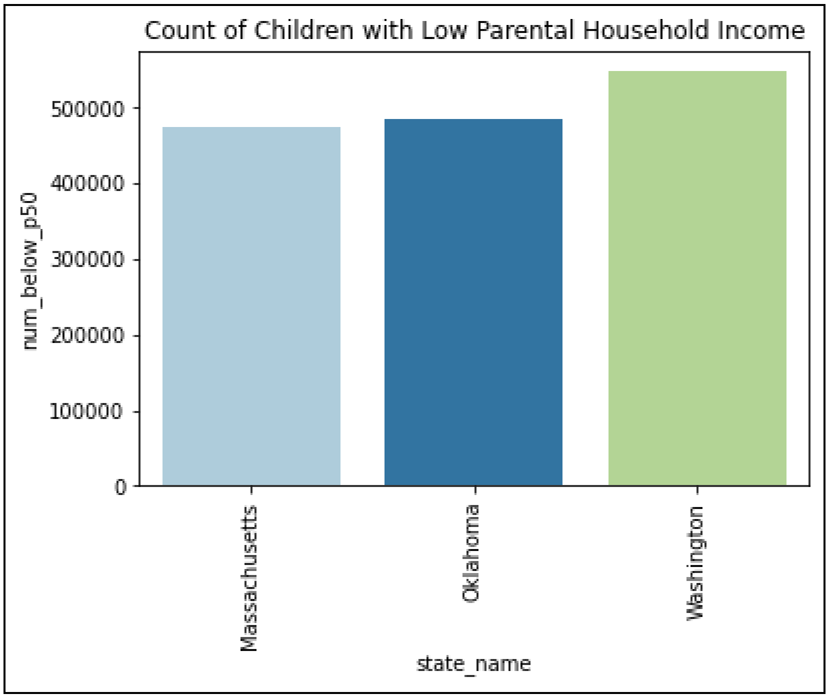

Figure 2: Number of Children with Low Parental Household Income

Description: In this task I wanted to see which states have the highest population and which states have the higher count of children with low parental household income, thereby exploring the County and Population Variables. From Figures 1 and 2, we can clearly see that Washington has the highest population in both regards, Massachusetts has the second highest population and the least count of children, and Oklahoma has a comparatively low population but a relatively high count of children in terms of its population. I can infer that Massachusetts has the lowest proportion of children with low parental household income as compared to the other two states.

Task 2:

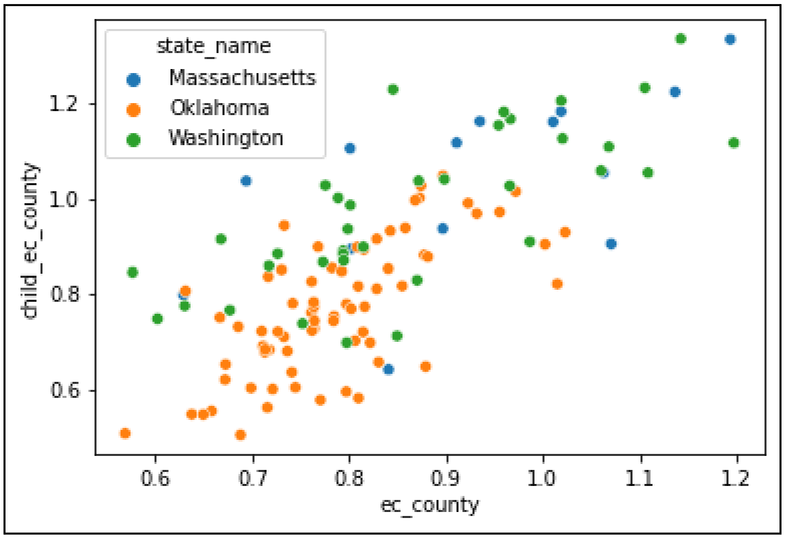

Figure 3: Comparison of two Connectedness Variables relative to the States

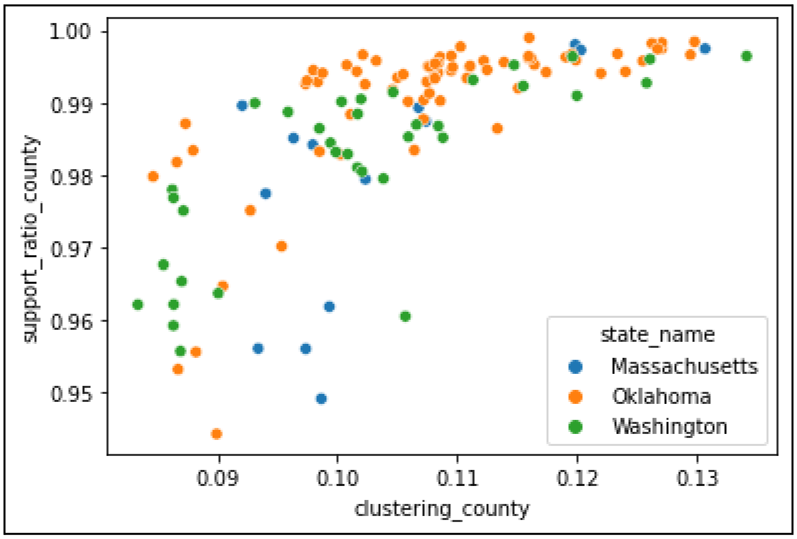

Figure 4: Comparison of two Cohesiveness variables relative to the States

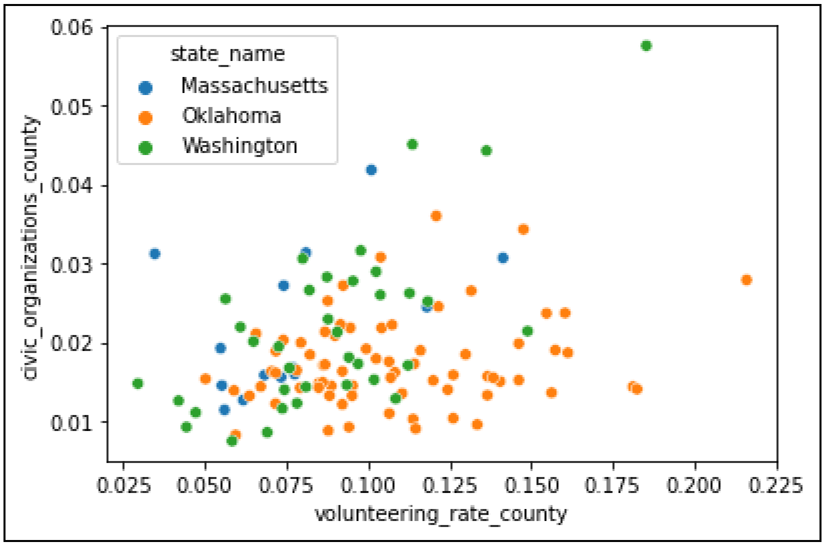

Figure 5: Comparison of two Civic Engagement variables relative to the States

Description: In this task I wanted to start exploring the data in relation to my exploration question, and decided to compare two variables from each of the three social capital types and see if there is any relationship.

- Figure 3: For this graph I decided to explore the Connectedness social capital type and decided to use the ‘ec_conunty’ and ‘child_ec_county’ variables. The graph is somewhat linear and is showing a strong positive relationship, which means that there is a high chance that if an individual with low-SES has high-SES friends then the children have a higher chance of having the same.
- Figure 4: For this graph I decided to explore the Cohesiveness social capital type and decided to use the ‘clustering_conunty’ and ‘support_ratio_county’ variables. I think that there is a negative exponential growth relationship however I am unable to put it into context.

- Figure 5: For this graph I decided to explore the Civic Engagement social capital type and decided to use the 'volunteering_rate_conunty' and 'civic_organizations_county' variables.There is no apparent relationship that I can observe between these two variables.

Task 3 (**Association Task**):

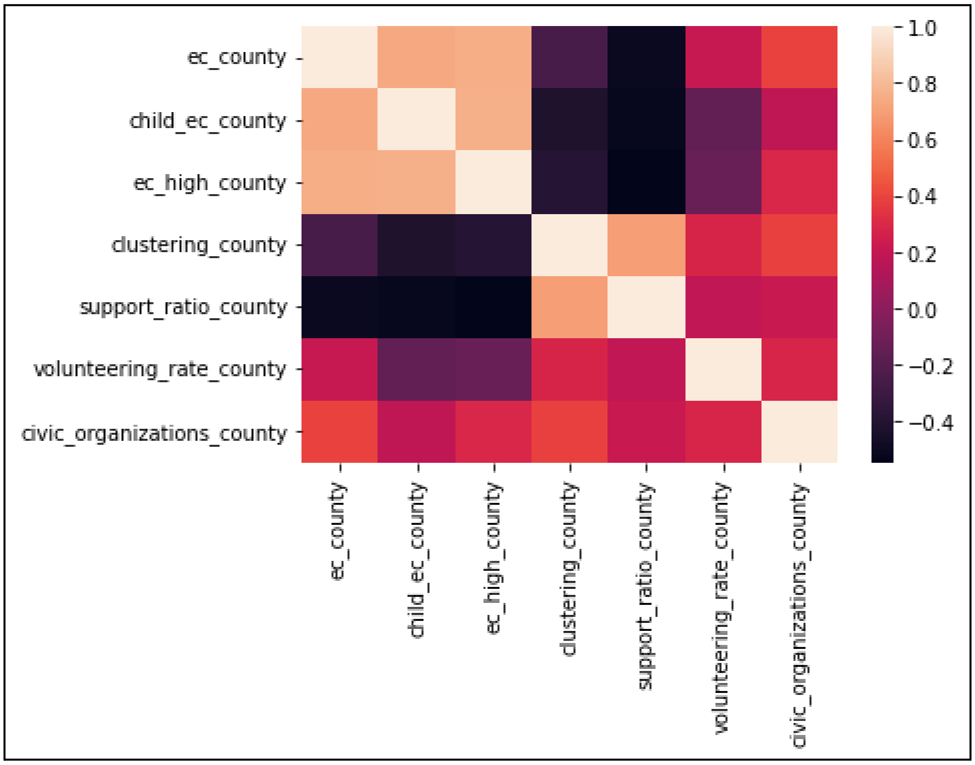

Figure 6: Correlation Matrix between all the Social Capital Types Variables

Description: To further understand and answer my exploration question I thought it would be interesting to look at a correlation matrix between all the social capital types variables and explore if there are any correlations between them. From the graph I can clearly see that the correlation between all the variables of the Connectedness Type is quite strong and positive with each other indicating that my choice of variables to analyze this social capital type was quite correct. There seems to be a strong positive correlation between the Cohesiveness Type as well as their variables also have a high positive value. There definitely is a positive relation between the Civic Engagement variables however it is moderately strong. An interesting relationship that I noticed was between the variables of the Connectedness and Cohesiveness Types as the relation seems to be negatively strong which indicates that there actually is no connection between these two types.

Clustering Task

Task 4

- I decided to cluster all the variables of the three Social Capital Types to see what are the various types of clusters that are getting formed and explore if there are any relations between the groupings.
- For this task I wanted to use K-means clustering however I was getting an error that I was not able to resolve so I tried out the Agglomerative Clustering Method and that seemed to work pretty well.
- I used the linkage method to be ‘ward’ and the number of clusters to be 5 as these were producing the best results in terms of dividing up the counties. Using a number higher or lower was resulting in clusters having either too few entities or too many entities.

- From the clusters I noticed that most of them were divided fairly as most values of most of the variables for most of the counties were in similar ranges, hence the clustering made sense.
- There was one cluster that I was quite surprised to see had only counties of Oklahoma and no other state. This was probably due to the fact that Oklahoma's Counties had the minimum values for most of the variables as compared to Massachusetts and Washington.

Classification Task

Task 5

- I decided to use the 'pop2018' column from the dataset and my class label is called 'Population'. I decided to have three levels: "Big Sized County", "Medium Sized County", and "Small Sized County".
- They are classified as follows:
    - Big Sized County: Population greater than 1500000
    - Medium Sized County: Population less than 1500000 and more than 250000
    - Small Sized County: Population less than 250000
- It would help in predicting if the county is big, medium, or small in size based on the total population.
- I was quite surprised to see that the performance of the model was quite accurate and almost perfect. I did expect my classification of the columns to work as I had taken out the maximum and minimum population and then decided the

different thresholds, but I did not expect it to be this accurate even after the cross validation.

- I tried various different parameters such as precision_score, recall_score, f1_score with different average values such as macro, micro, weighted etc. Most of them seemed to work and gave me similar accurate results. A few combinations did not work such as precision and macro as they are more suited for binary operations.
- Overall the effect of the various different parameter and average attribute combinations that did work, were quite similar in relation to one another and there was no significant performance change in terms of the results I got.

CHALLENGES ENCOUNTERED

- The biggest challenge I faced was to get KMeans Clustering to work. I was subsetting my original dataframe and using the subsetted data frame for my analysis and this was the primary reason for this task not working. I tried to use the original data frame as well however after the first two-three steps it failed again and I was unable to resolve this
- I did have a little difficulty in interpreting the dataset as I was having problems in understanding what each variable and their values represented and how they relate to each other, and this took up most of my time. Had I understood it better I would have been able to perform the tasks in a better manner than I did.

FUTURE EXPLORATION AND CONCLUSION

I believe that I was able to answer my exploration question to some extent as I did discover various relationships between the various social capital variable types. This dataset has so much more to offer and in the future I could further analyze more states, compare relationships between various states, analyze more columns, connect to the other data layers such as zip, high school and colleges and analyze them in relation to one another and the social capital types.

REFERENCES

- https://s3.us-east-1.amazonaws.com/hdx-production-filestore/resources/fbe5b0b9-e81c-41c7-a9f2-3ebf8212cf64/data_release_readme_31_07_2022_nomatrix.pdf?AWSAccessKeyId=AKIAXYC32WNARK756OUG&Signature=iESyZ2P1spflzzKolyoEXF1zlsM%3D&Expires=1671586363
- https://socialcapital.org/?dimension=EconomicConnectednessIndividual&dim1=EconomicConnectednessIndividual&dim2=CohesivenessClustering&dim3=CivicEngagementVolunteeringRates&geoLevel=county&selectedId=06037
- My Previous Lab and Homework Assignments