CS 397A Final Project
Exploring Economic Mobility Data from Opportunity Insights

Team Members: Arnav Jain and Arham Choraria

Contribution: Arnav performed clustering and geoplotting on the code while editing and formatting the data for better understanding while also writing the overview of our code in the written section. Arham performed visualizations and predictive modeling while writing the tasks section, dataset description, and future exploration.

Dataset: Economic Mobility Data from Opportunity Insights

https://data.humdata.org/dataset/85ee8e10-0c66-4635-b997-79b6fad44c71/resource/ec896b64-c922-4737-b759-e4bd7f73b8cc/download/social_capital_county.csv
https://data.humdata.org/dataset/85ee8e10-0c66-4635-b997-79b6fad44c71/resource/ab878625-279b-4bef-a2b3-c132168d536e/download/social_capital_zip.csv

This data was obtained in a csv format which included around 3090 rows and 26 columns. The data also came with a code book from the link that was attached to the pre-approved datasets section in the final project instructions. By reading the codebook, we were able to understand the data and get meaning out of it. This dataset includes a number of social capital statistics aggregated to the level of U.S. counties, ZIP codes, high schools, and colleges. The data is inspired from the Social Capital Atlas and measures social capital in the USA in three major ways. The first being connectedness which is the extent to which people with different characteristics (e.g., low vs. high socioeconomic status) are friends with each other. The second is Cohesiveness which is the degree to which friendship networks are clustered into cliques and whether friendships tend to be supported by mutual friends. And the third one is Civic Engagement which is an index of participation in civic organizations or volunteering groups. The dataset contains mostly numerical data with a few columns having string type that hold the information about the county and state of the USA.

Not much preprocessing was required for this dataset as it was simple to read the csv file in and use the dataframe formed. Empty values were switched to NA in the dataframe using regex which were later dropped for a better understanding of the dataset. For our specific tasks, we copied the dataset into other variable names (like bar_data, line_data, etc) to avoid changing the original data values for exploration. Moreover, we also grouped the data by state names after splitting the county_name column as this would make analysis easier because the dataset was very large compared to every country's data. Specific to the situation, we used the sum or mean function to group the data which is explained below.

Description of Tasks:

In total, our team performed 4 major tasks on this dataset which matched the requirements mentioned in the instructions.

Our first task was to explore the data, understand it and visualize it. With the help of the codebook provided, we explored the columns and tried to understand what value the attributes held and symbolized. For example, the 'num_below_p50' represents the number of children with below-national-median parental household income in every county. We thought of starting our exploration by analyzing the population across the states mentioned in the dataset. In order to check the population for every state, we had to split the 'county_name' column into the state name and county name. After splitting, we had to group the data by the name of the states which would help us calculate the population of every state rather than every county which was initially provided. As we had to perform further tasks on the given dataset, we decided to copy the data to a temporary dataframe to perform the group by and split tasks as this wouldn't affect our actual data. After grouping it by taking the sum, we used sns and matplotlib to visualize the population by every state in a bar plot. By adjusting the tick parameters and label sizes, we were able to make the label names easy to understand as there were a lot of them. We chose y-axis to show the population and x-axis to show the names of the states. With the help of visualization, it was easy to understand and conclude that the state of California had the largest population in the US according to the dataset followed by Texas.. While Wyoming had the lowest population. Next, we wanted to see how economic connectedness varied for each state by creating a line plot. According to the code book, the baseline definition of economic connectedness (ec_county variable) is two times the share of high socioeconomic status (SES) friends among low-SES individuals, averaged over all low-SES individuals in the county. We used the same dataframe formed for the previous visualization but this time instead of finding the sum over the counties for a state we used the mean (as this would give us the average scores throughout the counties for every state) to group the data by for our line plot. Based on the result of the line plot we were able to see that New Hampshire state had the highest economic connectedness over the USA and South Carolina had the lowest economic connectedness. This line plot gave us an idea about how the economic connectedness varied across the states.

For our second task, we went further in the lines of visualization by using geopandas and plotly express. We wanted to analyze and visualize the 'num_below_p50' variable, which is the number of children with below-national-median parental household income, and we wanted to visualize it with every county on a map similar to what we did in a lab assignment during our course. With the help of the county_name column provided in the dataset, we were able to determine the exact location of the counties and states thanks to the geolocator function. Moving on, we added the latitude and longitude of each county to another column in our dataframe. With this, we were able to implement mapbox to visualize the number of children below the parental household

income across the counties in the USA. This helped us understand and visualize the number of children with below-national-median parental household income on a map of the USA which made it easier to take a broader look at the situation in different states.

Moving on, we thought of implementing clustering on the dataset to identify groups of similar objects. We thought of choosing KMeans clustering as we wanted to cluster our data in k-sets simultaneously which agglomerative clustering does not enable us to do. We chose to explore the volunteering rate, the number of civic organizations, and economic connectedness for the clustering model (volunteering_rate_county, civic_organizations_county, and ec_county variables respectively). By choosing the parameters for the model through trial and error and checking the output for numerous values of n_clusters, we chose 10 as the number of clusters which provided us with accurate results. We got a successful output where the states were clustered into similar groups when their volunteering rate, economic connectedness, and civic organizations were compared. This helped us catch similarities between states based on the variables and we received a set of interesting results with Massachusetts, Connecticut, and Vermont being in the same cluster showing that their volunteering rates, economic connectedness, and the number of civic organizations were similar while based on prior knowledge, they are also states that are very close to each other

Lastly, we thought of exploring our data with predictive modeling. We wanted to analyze and predict the economic connectedness (explained above) across the US. We ranked economic connectedness in 3 segments, high, average, and low and we separated these segments by the ec_county values. If the value was greater than 1.02, we ranked it as high, if it was between 1.02 and 0.68 then we ranked it as average, and in other cases, we ranked it as low. We were able to find these boundaries to split by checking the maximum and the minimum values of economic connectedness and taking averages. This method helped us convert the column to a categorical variable. We ran a KNN classifier on the dataset after splitting the data into 80% training and 20% testing. The task was performed successfully and our model achieved an accuracy of about 68.69% which shows that our model performed fairly well and did not overfit. We also calculated the number of misclassifications which was 170 in our case and that was fairly small compared to the large dataset. To get better results, we also ran a 10 fold cross validation which gave us the number of occurrences of high connectedness, average and low. We saw that average connectedness had a very high number of occurrences which shows that in the USA, people are more likely to have an average social connectedness.

Overview of the Code:

We created a Jupyter notebook file for the coding part of the project. In the beginning, we imported all the libraries that were required for our tasks and all of them were also libraries that we used during the course. As we mentioned before, we first read in the csv file with the dataset,

stored it in a pandas dataframe, and cleaned the data by converting the blank values to nan and then dropping the NA values since they would give us no meaningful information.

The next few chunks were made for visualizations where sns.barplot (from the seasborn) and plt.plot (from the matplotlib.pyplot) libraries were used to get a barplot and a lineplot respectively. Most importantly, we split the column ‘county_name’ into states and counties and grouped the data by the states as we felt analyzing states as a whole would make analysis faster and easier. We used the sum and mean functions to group by. For example, as we wanted to explore the population to see how populated the states are, we used the groupby.sum to find the total population across counties per state. However, when it came to visualizing economic connectedness over the USA for the line plot, we used groupby.mean to get an average of the values as summing that would not make sense. Various ticks parameters such as axis, rotation, style, and labelsize were adjusted to get appropriate axis width and axis labels for the barplot and the lineplot like setting the rotation to 90 to view the labels properly.

Moving onto the geocoding section we used the .map() function to map each row of the dataframe index (states) to the geolocator.geocode function and get the necessary latitudes and longitudes for each state. Thanks to splitting the data, we ran the function on the total number of states and not counties as that would take a lot of time. The latitudes and longitudes were then used to plot the states across the USA and our program depicted the num_below_p50 column for each state on a map using the mapbox. To access the mapbox we had to first get access TOKEN. The parameter carto-position was chosen for the mapbox_style as this helped the map appear in the most understandable way.

For clustering of the data, we first selected the columns we wanted to cluster. Since we had a larger dataset, we decided to group it by states to make it easier to analyze like we explained above. After trial and error along with analyzing the results, we were able to decide 10 as a good number for clusters as we noticed better results with that value with clusters being formed with high similarity. We used StandardScaler and the KMeans libraries to Scale and get the clustering output respectively and analyze the results. We chose the columns 'ec_county', 'volunteering_rate_county' and 'civic_organizations_county' as we wanted to explore how similar states are based on these variables.

The chunks after clustering are for the KNN classification modeling. To find the appropriate split for the economic connectedness column (ec_county) into three groups, we found the difference between the maximum and minimum values of ec_county and divided the result by 3 to get three equal splits. Using our self-defined function convert_to_categorical, we converted the continuous values of the ec_county column to, high connectedness, average connectedness and low connectedness based on the splits of greater than 1.02 for high, between 1.02 and 0.68 for average and less than 0.68 for low. Next, we created the training and the testing sets with 80% of

the data being a part of the training set and 20% in the testing set. On multiple trials and errors, we ended up with 40 as the number of neighbors for our KNN model. Since we are using a large dataset, a fairly large number of neighbors ensured that noise has lesser influence on the result. We also received a decent accuracy as mentioned above which suggested our model was more realistic and did not overfit the data. Lastly, we also ran k-fold cross validation with 7 folds and we chose this by taking a square root of the number of samples to get more accurate results. We saw that most of the USA had average connectedness based on the results.

Our code has comments mentioned for each part and does not require any specific method to run it. Just that the code should be run sequentially given in the format provided. The chunks are separated for each task for easier understanding and require the libraries to be imported and the dataset to be read before running.

Challenges Faced:

The challenge we faced was the length of the dataset. Information for every county was given and that data was getting very difficult to visualize. It took 25 minutes to visualize the data in the geoplot section with more than 3000 rows and the labels for the barplots were extremely long. Grouping the data by every state helped us analyze and visualize the code easily.

What we learned from the Dataset:

We found out that the entire data was based on a study of economic connectedness and it took time to understand every part of the data through the code book. A lot of statistical data was provided with formulas given about how economic connectedness was calculated and how much error it had. We learned a lot about how socially connected the USA is which we never knew before.

Future Exploration:

We explored the population across states, the number of children between median parental household income, the volunteering rates, and other such variables to see the economic connectedness across the USA. It helped us answer questions like which states had the highest social economic connectedness, volunteering rates, etc. For future exploration, we would like to find out which variables affect economic connectedness more. Furthermore, we would also like to find the correlation between certain variables like volunteering rate and the number of civic organizations to see whether states have higher rates of volunteerism due to more organizations or not.