Milestone 2

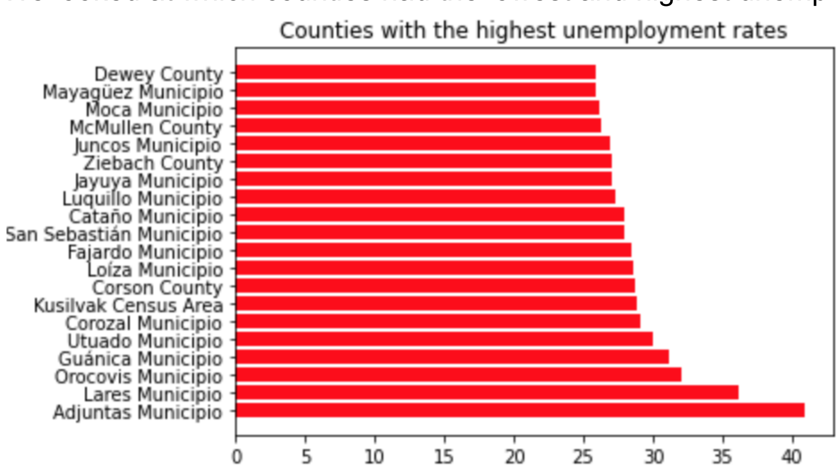
Github Link: https://github.com/Surabhi7602/Data-Warriors

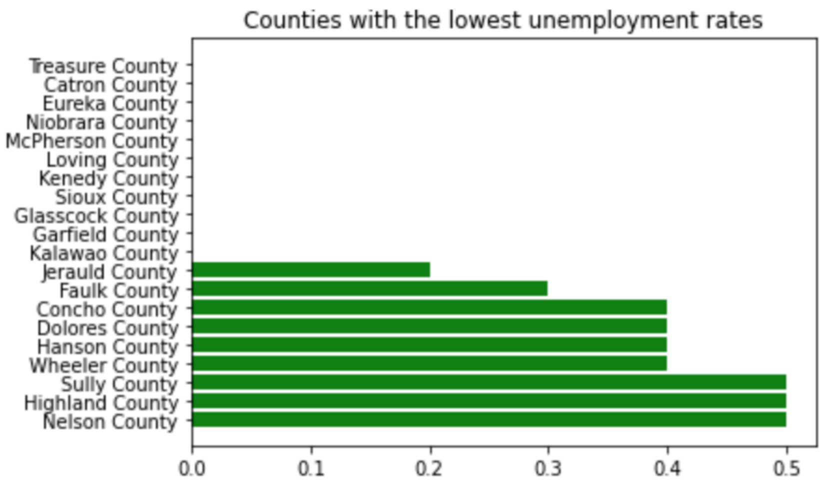
1. **Writeup:** The writeup should consist of an explanation of how your team's dataset was explored and why it's useful to your team project (1-2 pages). Some questions to consider in your writeup include (but not limited to):
    a. What was the most interesting insight you uncovered from exploring your dataset? (ie: most prominent aspect you noticed from your dataset)
    b. How did the exploration of the dataset influence your direction with your team project?

Dataset 1: https://www.kaggle.com/muonneutrino/us-census-demographic-data
Unemployment vs. Segregation
- How segregation is related to unemployment
    - Create visualization on unemployment
- The pattern between segregation and low capital income

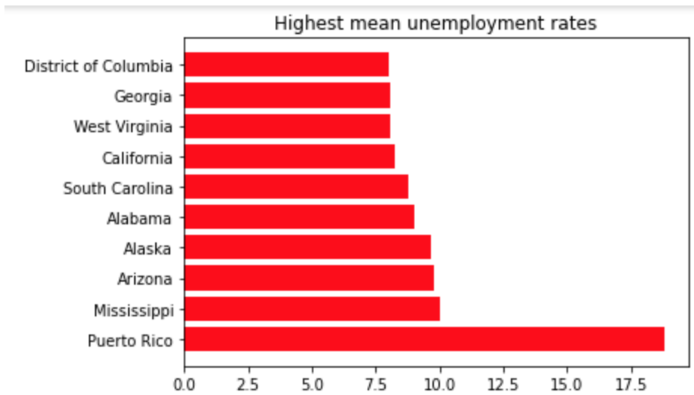
For now, we mainly looked into the per capita incomes and unemployment levels of the 3220 United States counties in the dataset. Our intention was to get some insight into these metrics that may have an impact on segregation levels in these counties.

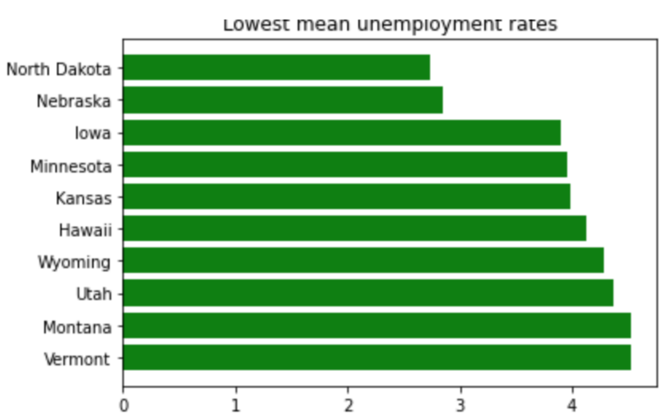
We looked at which counties had the lowest and highest unemployment levels.





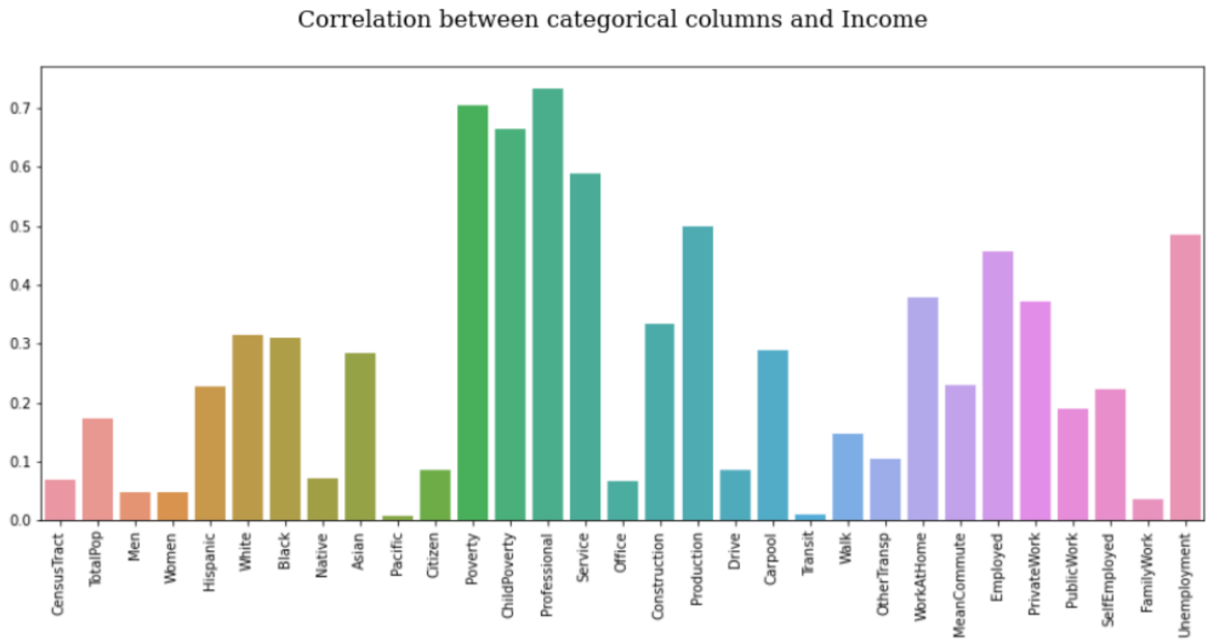
Then, we grouped counties by State, calculated the mean unemployment levels, and identified which states had high and low mean unemployment levels.





We also performed K means clustering on counties on the basis of their per capita incomes. From the elbow method, we identified that 5 would be the ideal number of clusters. As we work more on the project, we could try to find links between such clusters of counties and their segregation indices.

We will also look at the correlation between the categorical columns and incom. We found that poverty, child, and professional have the highest relation to income.This will be important to understand the reasoning behind unemployment for some groups.



Dataset 2: Diversity and Disparities Database
https://s4.ad.brown.edu/projects/diversity/Data/data.htm

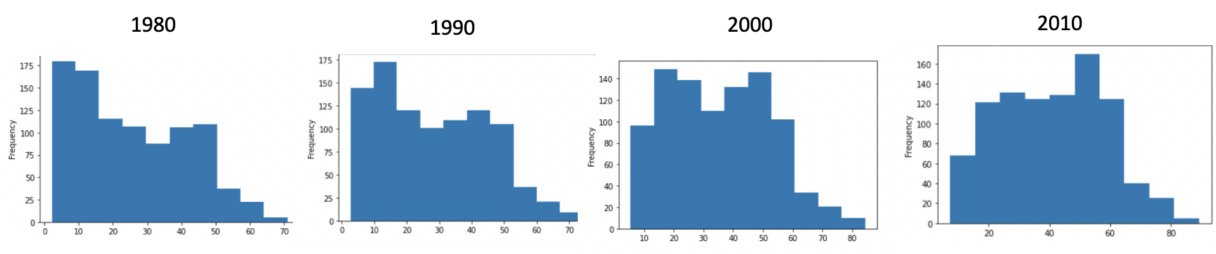
This database has many sub-datasets that can be explored in future work. For now, we looked at cities diversity indices (DIs) just to see how the nation looks as a whole.

This sub-dataset consists of 938 metropolitan areas and their DIs labeled as entropy scores from census years 1980-2010. The same data can be linked to many other features from the database such as ethnic breakdowns of these cities. Additionally, the data can be viewed more microscopically, such as at the county level, or macroscopically up to the state level.

	CBSA FIPS	CBSA Name	Number of Counties	Entropy Score, 1980	Entropy Score, 1990	Entropy Score, 2000	Entropy Score, 2010
0	10020	Abbeville, LA	1	31.582374	34.000301	38.338343	42.835399
1	10100	Aberdeen, SD	2	8.988803	10.503277	13.967252	20.410427
2	10140	Aberdeen, WA	1	16.800876	20.674586	33.130631	42.475171
3	10180	Abilene, TX	3	39.006235	42.550281	51.544556	57.032636
4	10220	Ada, OK	1	27.834359	37.096560	45.495782	52.680073
...	...	...	...	...	...	...	...
934	49620	York-Hanover, PA	1	13.675730	16.486384	24.475778	35.443066
935	49660	Youngstown-Warren-Boardman, OH-PA	3	24.951226	26.495257	32.059443	36.212160
936	49700	Yuba City, CA	2	46.584687	55.688489	67.662412	72.969935
937	49740	Yuma, AZ	1	56.393014	55.828014	56.724528	54.599220
938	49780	Zanesville, OH	1	14.092794	14.294799	18.848991	21.670218

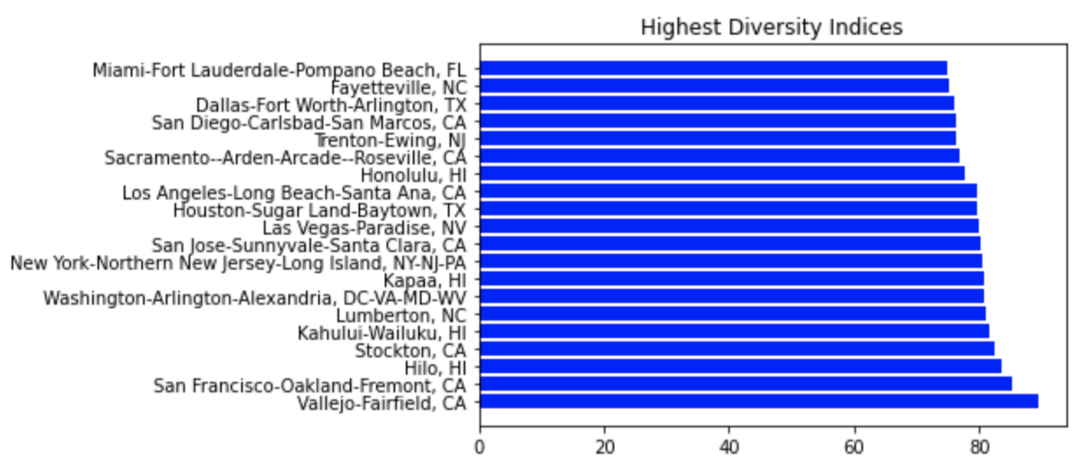
939 rows × 7 columns

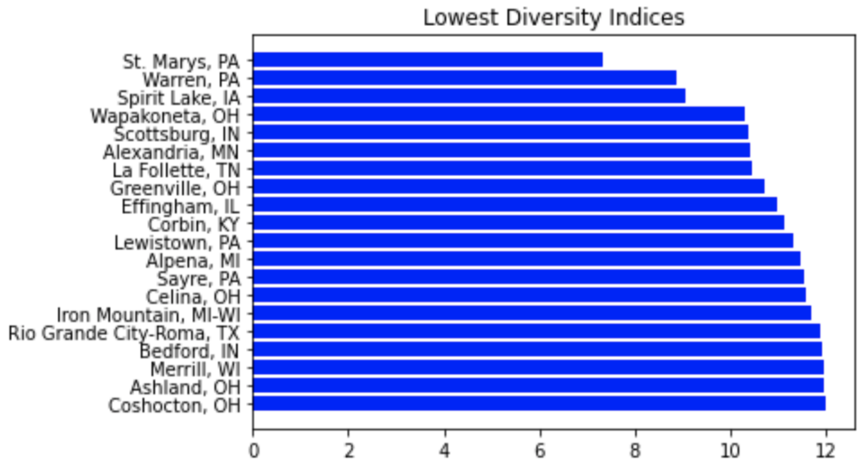
First, we wanted to see in general what the trend of diversity indices were overall. Plotting histograms of all these scores through each census year shows that since 1980, there has been a shift in diversity overall in Metropolitan cities:



In 1980, the histogram was left-skewed, meaning more cities had lower DIs. As time progresses, the shape of the histogram has become more normally distributed. The median DI shifts higher, meaning there has been an increase in overall diversity.

Next, we looked at what particular cities had the highest and lowest DIs:





Many of the cities with high diversity indices are in California. Many of the cities with the lowest diversity indices are in midwestern states or the east coast such as Ohio and Pennsylvania.

This analysis allows us to get an overview of the DI score and understand how the U.S.A looks at a national level. For future work, specific models can be made to cluster cities based on their diversity scores using metrics such as income level, ethnic breakdown, etc. This will require more data preprocessing to link these diversity scores to these predictors.