Executive Summary & Questions

Demographics are an important part to understanding how society functions. We are about to enter into society after the graduation of college and this topic of demographics has the potential to affect a lot of parts in our lives such as social norms, job recruitment statistics, and in general it is fascinating to wonder about how the USA will continue adapting to diversity.

Our project is investigating the change of the distributions of races across the states in the USA over 10 years, from 2009 to 2019. We are incorporating unsupervised learning to create clustering models for each race testing several K Means values to answer the question of "what are the similarities and differences of population proportions of races in all of the states from 2009 to 2019?" What number of clusters most accurately depict our scenario?

Data Collection & Cleaning

Our initial data set "race_df" (from https://www.statsamerica.org/downloads/default.aspx) contains census measurements from the Kelley School of Business Stats America at Indiana University of population estimates by race and ethnicity for the U.S., states, counties, metros, micros, and EDDs, from 2000 to 2019. Since they were not relevant to us, we dropped the columns "IBRC_Geo_ID", "Statefips", and "Countyfips", leaving us with the year, description of which state/territory/county it was, and the total population of every year for that row.

To make sure our data pertained only to states, we dropped any row with "county", a comma, "U.S.", "District of Columbia", or "Puerto Rico" as these values are not states. We also only kept the rows where the year value was equal to 2009 or 2019.

Since we need to work with proportions, we created two new data frames "race_df2009" and "race_df2019" that had 7 new columns for every race which contained the quotient of the specific race's population divided by the total population.

We needed to drop the "hispanic" column as according to the following article the term hispanic is not considered a race in the US Census, rather they are recognized as individuals that may fit into any race, and according to census data they are recognized as an ethnicity. Due to that, the hispanic column was not accounted for in the total population. https://www.verywellmind.com/difference-between-race-and-ethnicity-5074205

Finally we created a combined DF which accounts for both 2009 and 2019 - we accomplished this by merging the 2009 and 2019 DFs adding the suffixes "2009" and "2019" to each individual respective column.

Variable Selection

The following are the variables we are using from our data set:

1) White Prop 2009 & 2019 - the proportion of White people in the given year/state

2) Black Prop 2009 & 2019 - the proportion of Black people in the given year/state

3) Native Prop 2009 & 2019 - the proportion of Native Americans in the given year/state

4) Asian Prop 2009 & 2019 - the proportion of Asian people in the given year/state

5) Pacific Prop 2009 & 2019 - the proportion of Hawaiian or Pacific Islander people in the given year/state

6)Mixed Prop 2009 & 2019 - the proportion of mixed race people in the given year/state

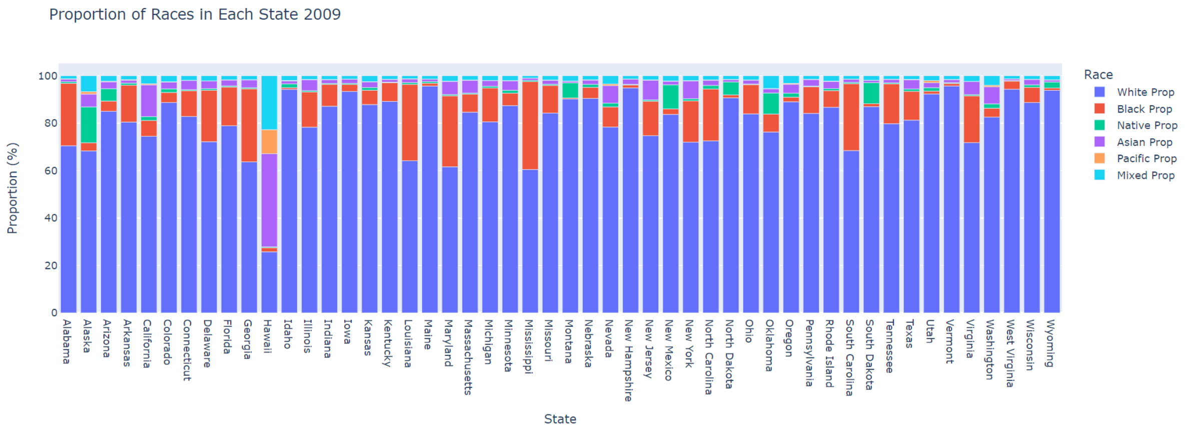
Our observational units are each State in the USA.

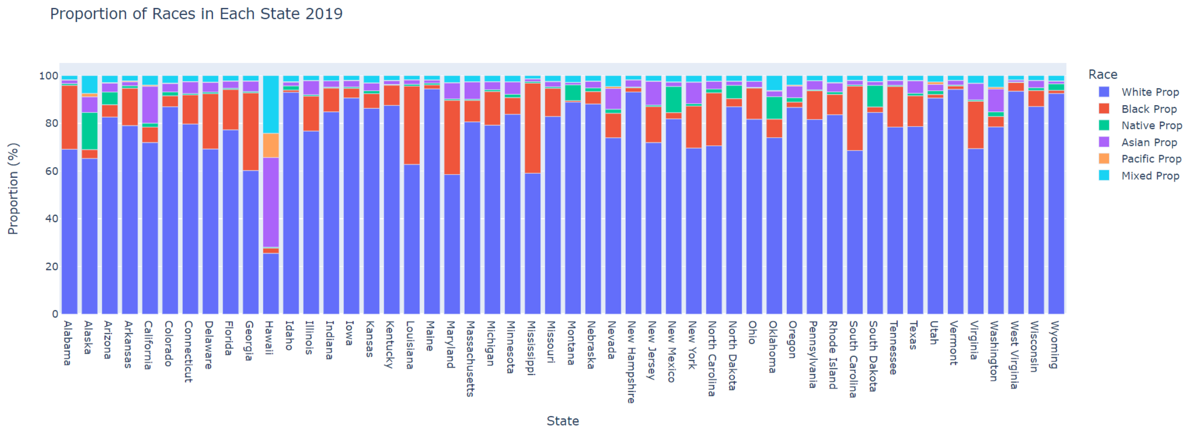
These variables are fairly self explanatory and are calculated numeric values from each state/year's census. Using these variables we will be able to fit an unsupervised machine learning model to cluster the proportions of every race per state to see if the proportion clusters drastically change by states depending on the years.

Visualizations

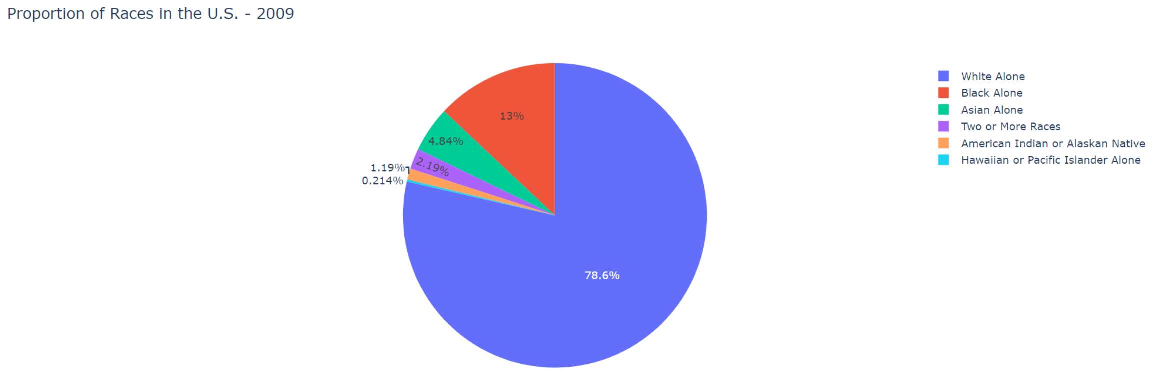
We have created stacked bar charts for the proportion of races per state in 2009 and 2019 respectively as shown. As seen here, it is very difficult to note a major difference in the proportions per state - but the difference is there. If you hover over every different color in the two separate graphs in the code notebook (these are screenshots,will not be able to here) it is evident that the percentage has changed slightly over the years. The majority of states with the exception of Hawaii have the majority race proportion being White and the minorities being mixed race, Pacific, Asian, Native, and Black.

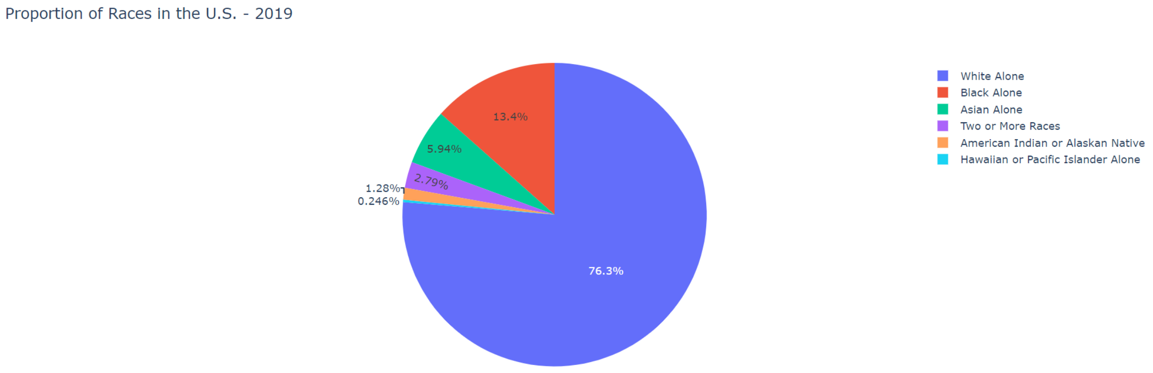
Without a machine learning model, it is fairly difficult to see how the race populations have changed in 10 years.





Below, we have created two pie charts that depict the race proportions for the USA as a whole for 2009 and 2019. As shown, the percentages have not drastically changed and the majority/minority groups have remained the same, but a fluctuation of approximately 1% is evident for most.



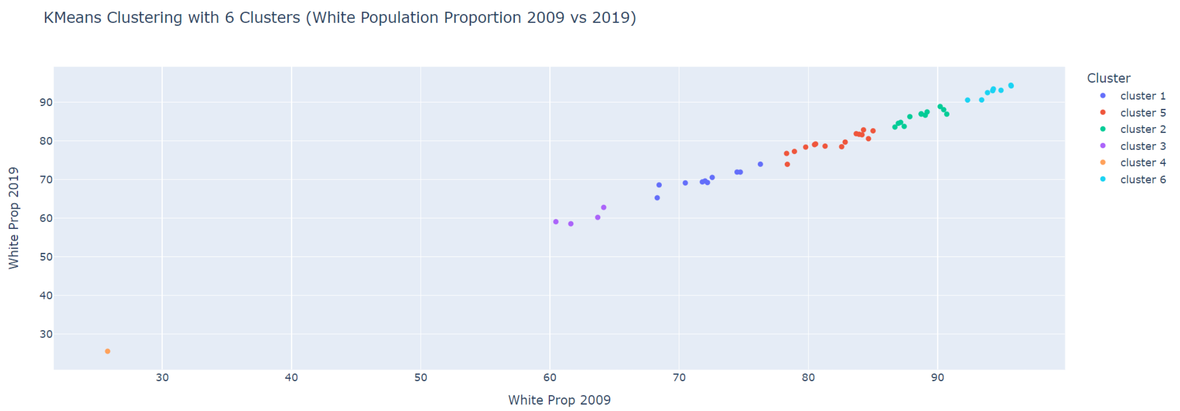


Model Selection - Machine Learning

Since our topic of proportion of races per states was to see similarities and differences of them and not necessarily predict any values, we decided to incorporate unsupervized learning models and we ended up selecting KMeans clustering to display which states are similar and different based on the race population proportion between 2009 and 2019. We can not do cross validation because this is unsupervized learning so there is no real correct answer for the clustering that we can compare.

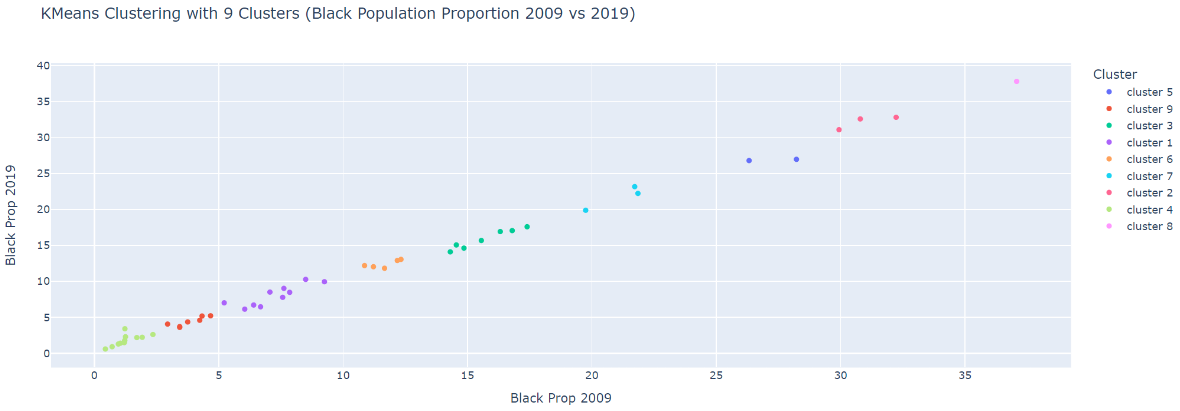
White Proportions

Starting off with analyzing the White proportions - we settled on a Kmeans value of 6 by implementing a for loop to diagram Kmeans values 3-8. We saw that although values before 6 were plausible, the states were clearly divided into other clusters as some states were closer to others and further from the rest but still in the same clusters. After value 6, the clusters started overlapping so we settled on value 6 as this showed the cleanest distribution of clusters. In the code notebook you are also able to hover over each point to see which states belong in which clusters, and see every iteration.



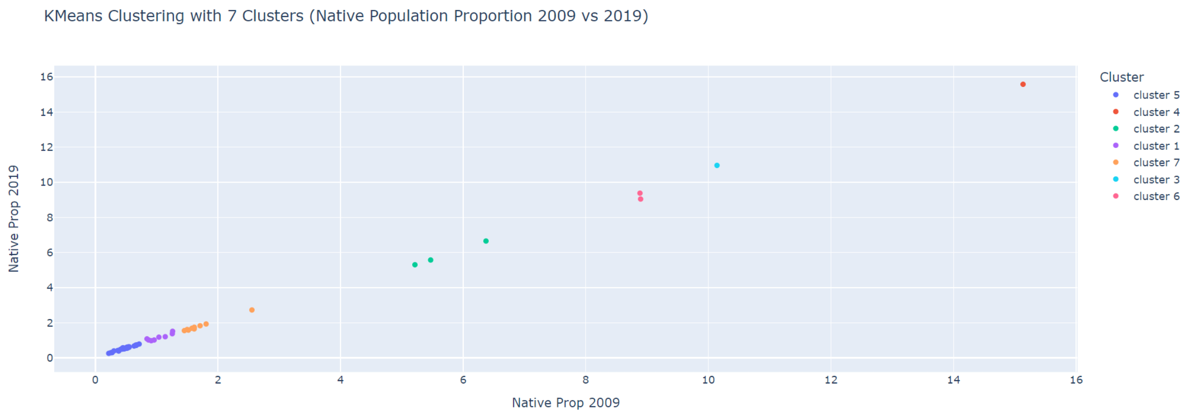
Black Proportions

Using a for loop we discovered Kmeans clustering value 9 worked best for the Black proportions 2009 vs 2019



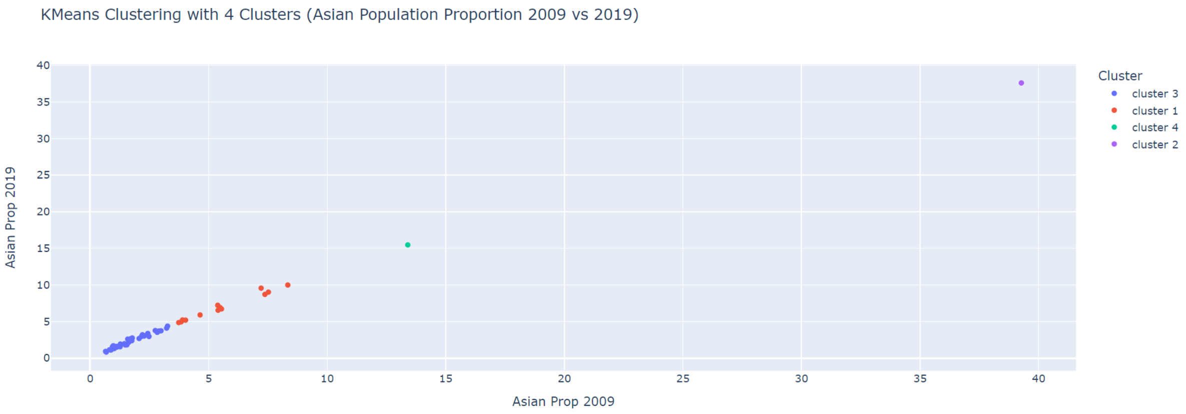
Native Proportions

Using a for loop we discovered Kmeans clustering value 7 worked best for the Native proportions 2009 vs 2019



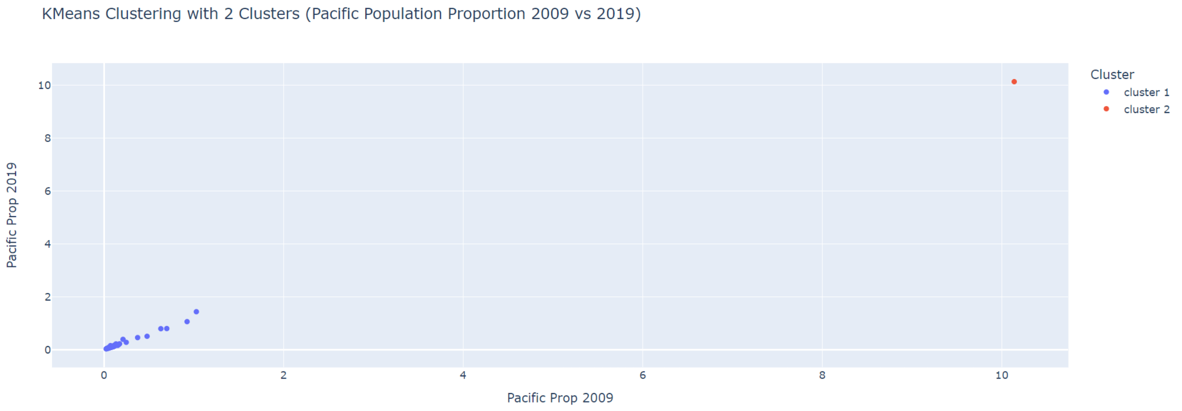
Asian Proportions

Using a for loop we discovered Kmeans clustering value 4 worked best for the Asian proportions 2009 vs 2019



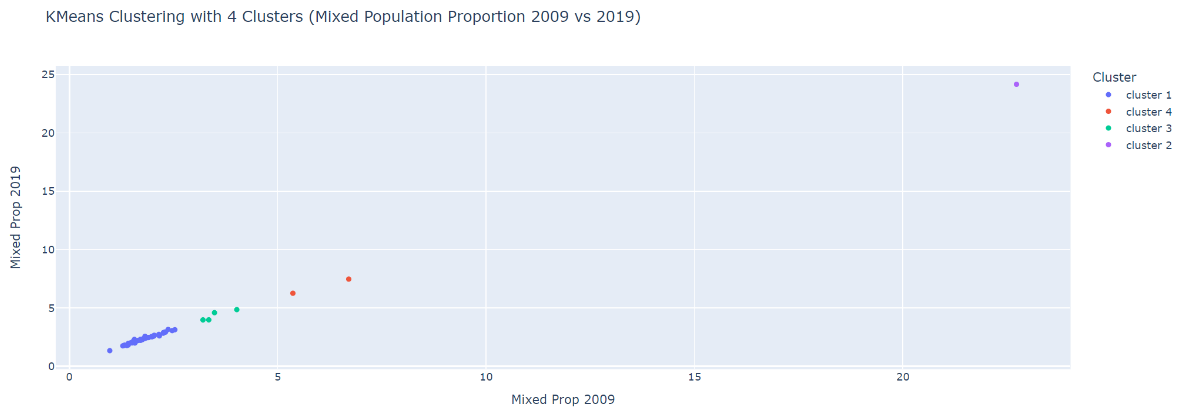
Pacific Islander Proportions

Using a for loop we discovered Kmeans clustering value 2 worked best for the Pacific Islander proportions 2009 vs 2019



Mixed Race Proportions

Using a for loop we discovered Kmeans clustering value 4 worked best for the Mixed Race proportions 2009 vs 2019



Conclusion

These clustering results highlight the complex and varied demographic changes in racial proportions across the United States between the years 2009 and 2019. The number of clusters for each racial group suggests that some groups experienced more uniform trends nationwide (like Asian and Pacific Islanders which had clusters of 4 and 2 respectively), while others saw more varied patterns (like Black and Native American populations, 9 and 7 respectively). This information can be vital for understanding social dynamics between these groups and planning for future demographic shifts. It also highlights the diversity of the American population and how it continues to shift and evolve differently across regions of the country. The variation in cluster numbers also suggests that different factors and influences may be at play for each racial group.

Additionally, as the clustering appeared mainly linear across the scatter plots for every race, it is plausible to conclude that from the years 2009 - 2019, the proportions of the races in each state did not increase substantially. The clusters are established to help us understand how varied the proportions for every race is.

These clustering results not only highlight the demographic changes across the United States from 2009 to 2019 but also show distinct patterns of racial concentration in certain states that could be driving the observed trends:

Mixed Race: Hawaii stands out for its mixed-race population. The clustering model likely detected unique patterns in Hawaii, which is known for its racially diverse population, with a significant number of individuals identifying as mixed race.

Pacific Islander: Hawaii again shows a significant proportion of the Pacific Islander population. This significant proportion likely resulted in Hawaii being placed in a distinct cluster for the Pacific Islander group.

Asian: Both Hawaii and California are noted for their higher proportions of the Asian population. These states could be the anchors of their clusters, characterized by high Asian proportions, influenced by historical immigration patterns.

Native American: Alaska is distinguished by its Native American population, which is higher than in most other states. This higher proportion could lead to Alaska forming its cluster or being part of a cluster with other states that have significant Native American populations.

Black: Mississippi is highlighted for its high proportion of the Black population. The clustering for this group likely reflects the historical and cultural significance of the Black community in Mississippi, which may differ from trends seen in other states.

White: Hawaii is indicated to have the lowest proportion of the White population compared to other states, which may place it in a unique cluster in the analysis of White population proportions.

By incorporating these state-specific insights, we can infer that the demographic shifts in racial proportions are not just a matter of general trends but are also significantly influenced by the historical, cultural, and social contexts of individual states.

The optimal number of clusters for each racial group should take into account these state-specific peculiarities to accurately depict the scenario. The clustering should be fine-grained enough to differentiate states with such distinct demographic features while still general enough to capture broader nationwide trends.

To refer back to our research question - the similarities between the years and the linear behavior of the cluster models for every race indicate no great change between the years 2009 and 2019 for the proportion of races in the US per state. Each proportion of race in majority of states increased or decreased proportionally due to the linearity.

For the differences: Black and Native American populations had a larger number of clusters (9 and 7, respectively), indicating greater variability in how their population proportions have changed across states. This variability points to diverse local factors or conditions affecting these groups differently from state to state.The variation in the number of clusters for each racial group suggests that different influences may be at work. These could include economic opportunities, social mobility, migration patterns, or policy changes that have different impacts on each racial group.