2017 U.S. County Data Correlations

By: Logan Chalifour, Gerry Crepeau & Riley Demanche

I. Executive Summary

After examining various demographic and economic statistics for each county within the United States from the 2017 census, it is clear that certain variables are correlated to one another. In an effort to determine what some of the most correlated demographic and economic variables are in general across the country, we utilized choropleth mapping, a colored or patterned map in proportion to a statistical variable, to find biases in our data and draw possible conclusions between them. Our main goal is to identify the most prominent biases within the data and learn more about the geographical patterns of them by leveraging geospatial data. Another objective behind creating this project is developing a dashboard to allow everyone to investigate the data the same way we are able to. Given the recent surge in social justice reform and call to action, these visualizations and correlations help give some evidence to different kinds of discrimination and bias that occur in the United States.

There are three main relationships that we discuss throughout this project. They are the correlation between Percent Black and Poverty Rate, Percent White and Percent Self-Employed, and Percent Asian and Average Income. Their relationships can be molded into geographic stories after looking at our visualizations as well. To explore these three examples and any more that you can think of, you can visit our dashboard here.

We'd also like to mention that while we provide a possible explanation for these trends they do not tell the full story, or indicate causation. The following sections include data collection, a briefing of our dashboard, and an overall look at important findings in greater detail. To see all aspects that go into the design of this project please visit our GitHub Repository here.

II. Data Collection

The data for this dashboard and project can be found on Kaggle. This data comes from the 2017 American Community Survey by the U.S. Census Bureau. The raw CSV file that was used for creating this dashboard is a subset of this Census data provided by the U.S. Census Demographic Dataset on Kaggle.

Each county observation contains 37 variables representing different kinds of demographic and economic statistics. For the purposes of our project, we reduce the number of variables within the data frame, and also exclude those counties in Puerto Rico. Some additional steps needed to take place to finalize our data such as converting some variables into a percentage to make it easier when comparing correlation. After this cleaning process, our final data includes 3,142 U.S. county observations with 18 variable columns. This included three identifying columns (FIPS, State, and County Name). The remaining 15 variables were then broken up into two different buckets in order to create the dashboard – demographic and economic. These demographic variables include Population, Percent Men, Percent Women, Percent White, Percent Black, Percent Hispanic, Percent Asian, Percent Pacific, and Percent Native. The economic variables consist of Average Income, Poverty Rate, Unemployment Rate, Percent Self-Employed, Percent Work at Home, and Mean Commute.

Our goal for this project was to look at the biases present in our country and attempt to draw conclusions based on their correlation and geographical patterns. This dataset helps achieve the entirety of that goal. First and foremost, it includes county FIPS codes to help us geographically plot the data with advanced Python packages. This allows us to create conclusions based on the geographical distribution of our variables. In addition, the dataset includes a plethora of demographic and economic variables. This is information that generally shows bias within communities, so having nationwide access to this data at a county level can help us find general biases present across the country.

III. Interactive Dashboard

With the dashboard that has been created, you will notice two different geospatial plots that outline each U.S. County on a map, along with a correlation calculation for the selected variables. These plots help represent the counties that have the highest concentration of each variable compared to the rest of the country using a color scale, with the lighter colored counties representing the low end of the spectrum and the darker colored counties representing the higher end of the spectrum. These plots also allow the user to zoom in and hover over each county individually to get more information on the selected variables and observe trends within their own specifications. On the left side of the dashboard, two drop-down menus allow users to select different kinds of demographic and economic variables. As these selected variables change, the maps and correlation values will update to show the new relationship between these demographic and economic factors. In addition, the dashboard will run a hypothesis test on the correlation between the user's two selected variables and show a scatterplot of their relationship below the maps.

IV. Correlations

Three specific correlations stood out from the others: Percent Black to Poverty Rate, Percent White to Percent Self-Employed, and Percent Asian to Average Income. These three appeared to have some of the highest correlation coefficients when comparing the relationship between all of the demographic and economic variables. All three of these examples were proven to be statistically significant after running a hypothesis test at the 5% significance level.

For the first relationship, which can be seen in Part A of the Appendix below, you will notice a correlation coefficient of 0.467 between Percent Black and Poverty Rate. This value represents a positive correlation between these two variables, so as the Percent Black in a county increases, the Poverty Rate would be expected to increase as well. This can also be seen when comparing the visuals from the dashboard, as those areas within the U.S. (primarily in the Deep South) that have higher concentrations of African-Americans tend to have a higher Poverty Rate compared to the rest of the map. Together, these maps provide a clear picture of that correlation. It is no mystery that minorities are oppressed in this country, but these maps contextualize the correlation further. Along with the streak of counties in the southeast with a high poverty rate, there are a few other regions in America with a similar trend. These regions in South Dakota, southern Texas, and Arizona/New Mexico, are home to many Native American reservations and immigrants. We are interested in the black communities for this specific investigation, but the map provides significant context to our data. This context introduces a clear bias in our nation and should be explored further by policymakers with the ability to change the trajectories of the communities affected by this inequality.

The next relationship that we examined more closely was that between Percent White and Percent Self-Employed within counties (Part B of Appendix). This correlation value of 0.234 again suggests a positive effect showing that as Percent White increases, the expected Percent of Self-Employed individuals increases as well. This correlation is a little harder to visualize compared to the last example because most parts of the country have a high concentration of white residents. However, as you can see on the bottom map, the Midwest has a much higher concentration of self-employed individuals. This region is characterized by its plains and bountiful farmland, so it could make sense that farms in these areas are family-owned workplaces. The scatter plot in Part B of the Appendix reveals there aren't too many counties with a high self-employed percentage, but most that do have one of these higher values are in a county with a greater concentration of white individuals. From this section, one of the most important conclusions we can draw is the discrepancy of lifestyle between the coastal areas and midsection of our country. The map of Percent Self-Employed shows that people in different areas of the country could have different interests and political views. This trend is one to explore further with future data exploration to see if it could reveal any distinct biases.

One last example to discuss is the relationship between Percent Asian and Average Income of counties (Part C of Appendix). With a correlation value of 0.451, this would suggest yet another positive relationship between the selected variables; as Percent Asian within a county increases, we expect Average Income to follow. From the first map in Part C of the Appendix, it is evident that most counties have a minimal Asian representation. However, some counties have robust Asian populations which most often appear in coastal urban areas such as Los Angeles, San Francisco, Washington D.C., New York, and Boston. Interestingly, there is a large percentage of Asian individuals in Hawaiian communities as well. In the bottom map, we see many locations with a high average income. Most of these locations appear to be the large cities that drive business across the nation. While not every high-income area has a large Asian presence, it appears that a majority of the high percentage Asian communities are in high-income areas.

As we mentioned in our Executive Summary, there is no way to determine how or why all of these relationships occur, but perhaps this can be explained by the fact that some of the highest Asian populated counties tend to be more urban cities. In general, cities tend to be more economically successful due to the combination of big businesses and an incredible amount of high-level employment opportunities.

V. Conclusion

Now that we have taken the time to explore various relationships between demographic and economic variables and identified certain trends, it can be difficult to digest all of this information. The year of 2020 has shed some light on the amount of disparity and inequality that is present within our country. Different groups of minorities have been discriminated against in this country for centuries, and our visualizations and findings reinforce these unfortunate themes.

It is clear that there are certain demographic and economic disparities that the United States needs to improve upon. If there is any sort of change that is going to occur across the country, patterns, and correlations related to race and economics need to be addressed. Our project sheds light on a few of these correlations but more importantly offers an easy way to identify them through the use of geospatial visuals.

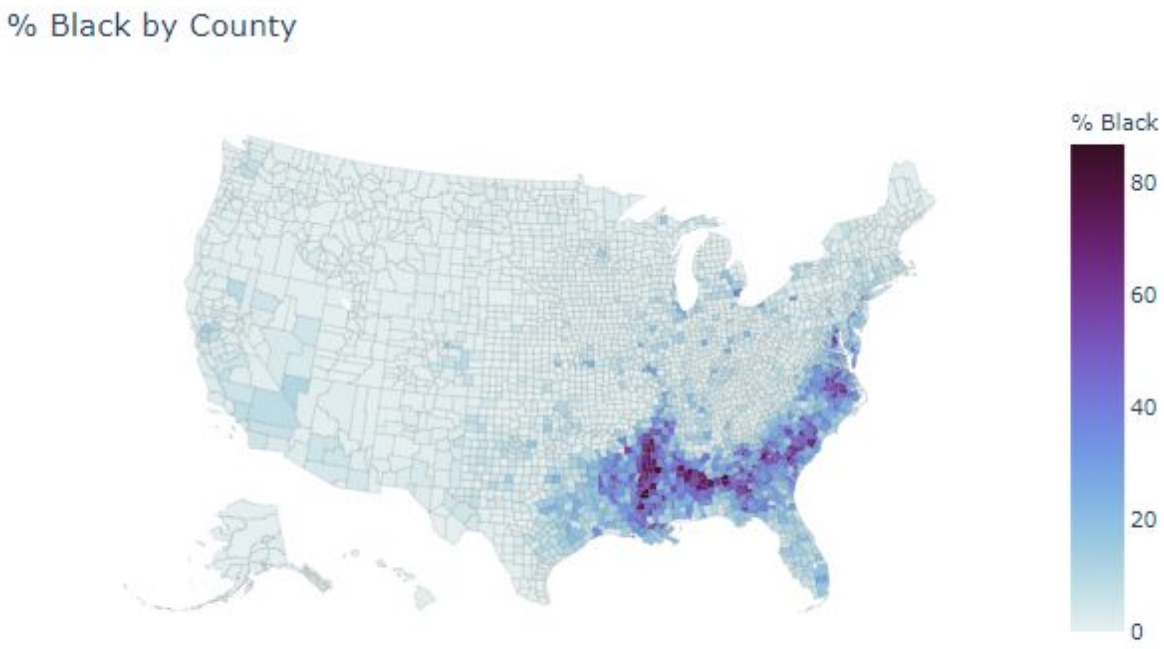
Moving forward, different policies and procedures must increase equality and break these trends that have been shown in the past. Hopefully, our project findings and dashboard development can help

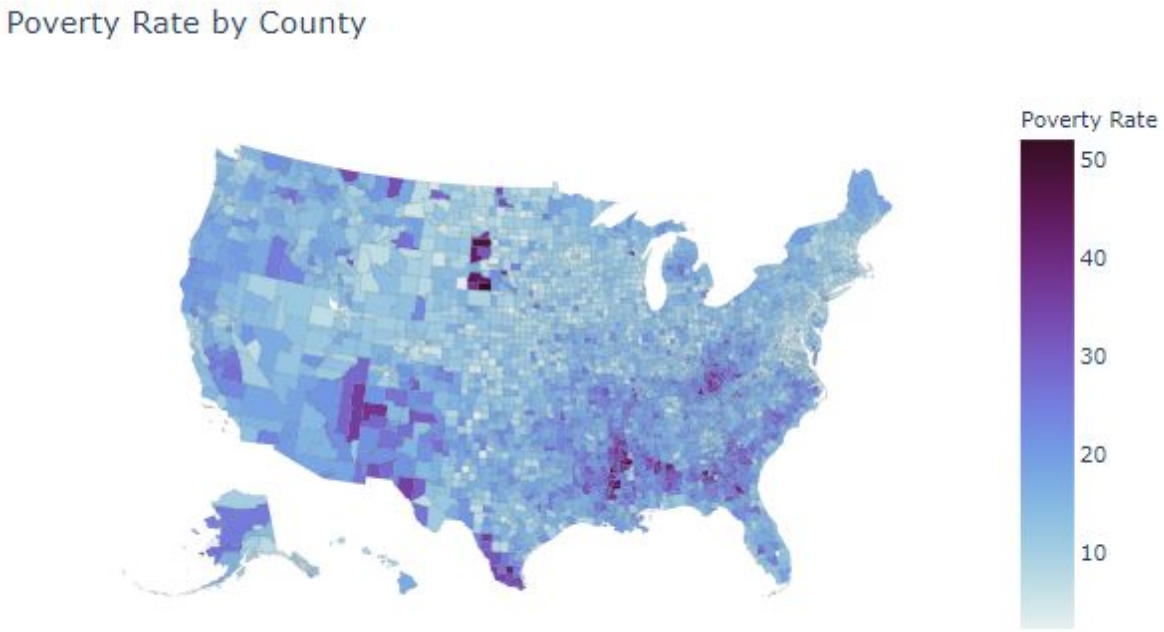
others realize and identify certain patterns of bias. A similar process could be repeated in future years, and it would be interesting to see if any significant changes occur over time. The first step in order to enact change is proving to people that these problems do exist, and we believe this project is useful in that regard.

Overall, we have identified three specific correlations that showed biases and their geographical features. We statistically tested these relationships, and through the introduction of choropleth maps, we drew some possible conclusions as to why these trends occur based on our own knowledge of the country. There are so many other trends to be explored within this data, as our visualization tools provide a distinct way of analyzing data. We encourage you to continue exploring this dataset, finding new significant correlations, and drawing your own possible conclusions through the use of our dashboard. You can find that dashboard HERE!

VI. Appendix

Part A: Comparison of Percent Black and Poverty Rate by County

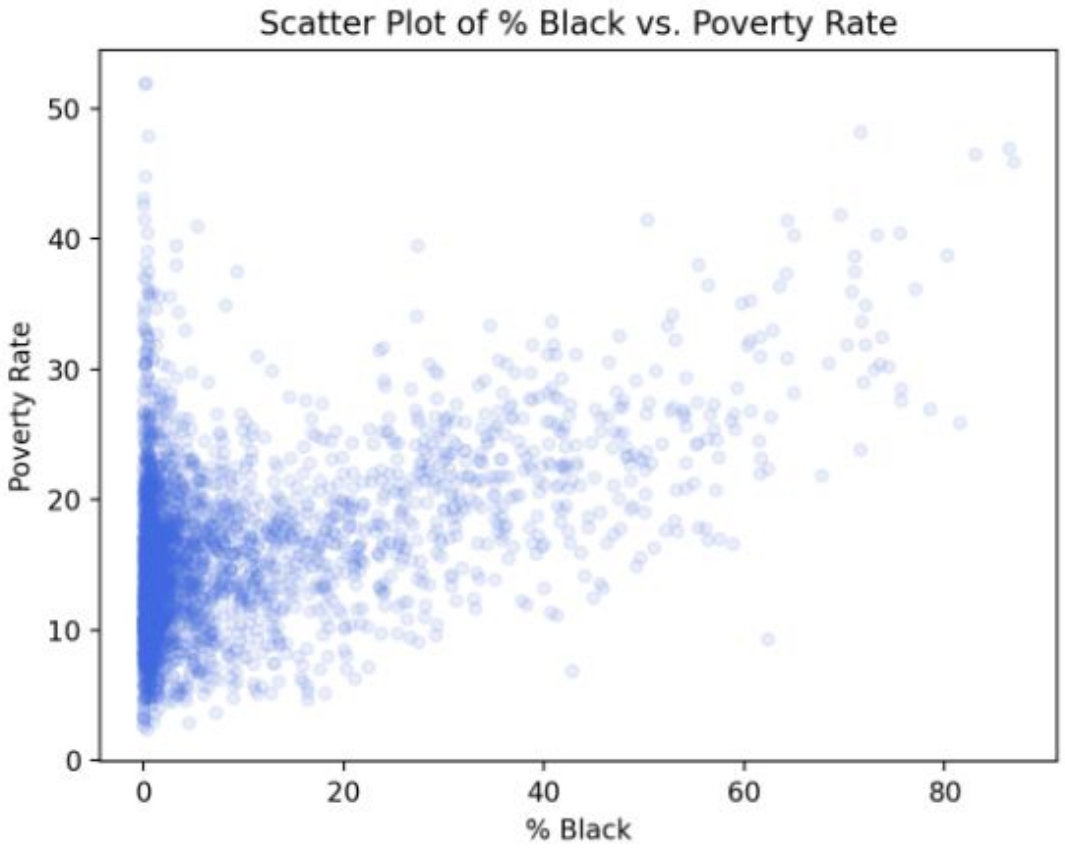




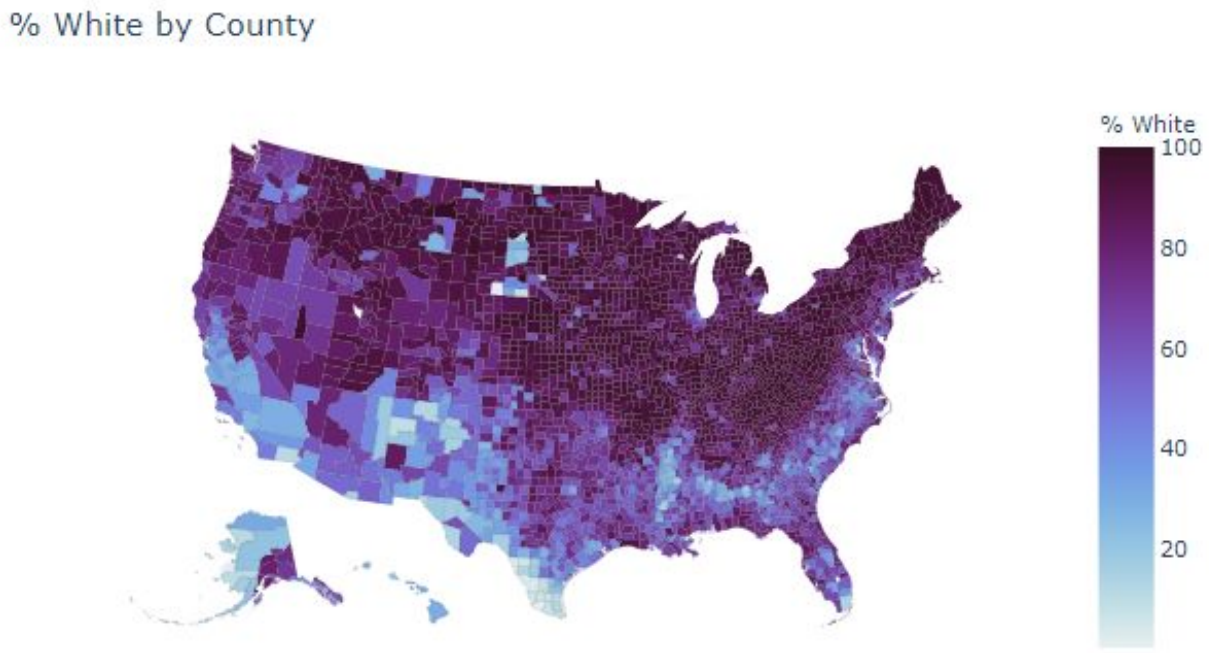
There is a statistically significant correlation between % Black and Poverty Rate for counties in the USA.

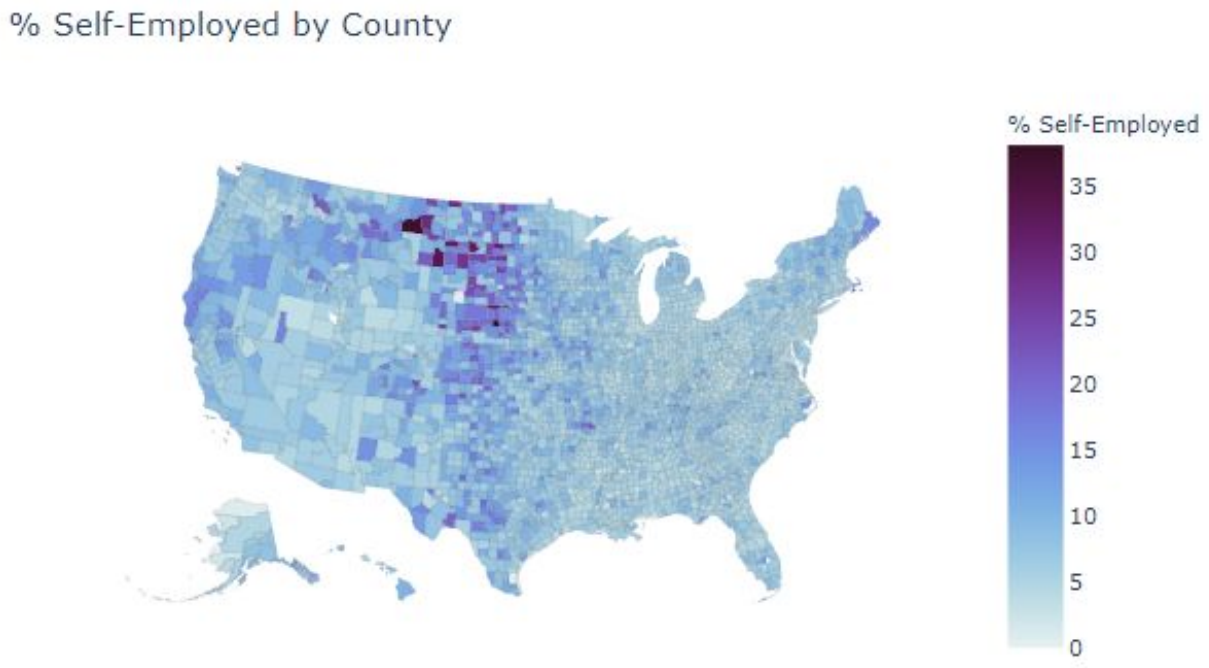
The correlation coefficient between % Black and Poverty Rate is 0.467.

This means that % Black and Poverty Rate are positively correlated.



Part B: Comparison of Percent White and Percent Self-Employed by County

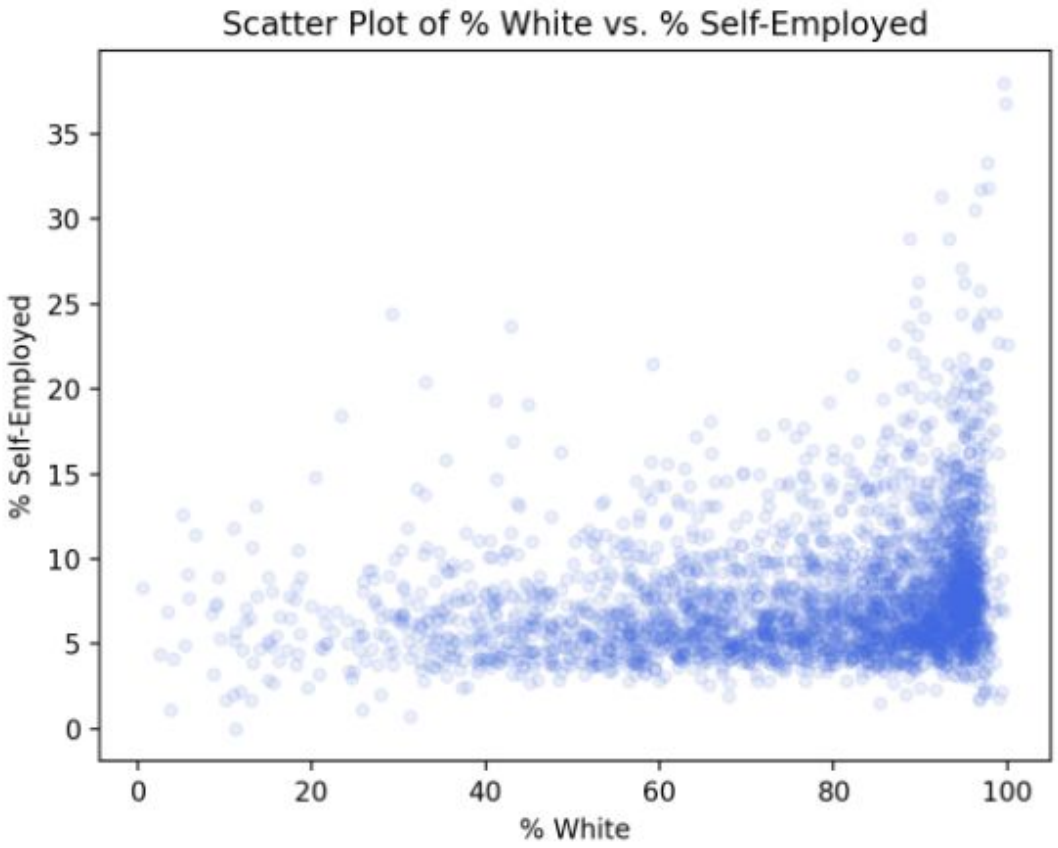




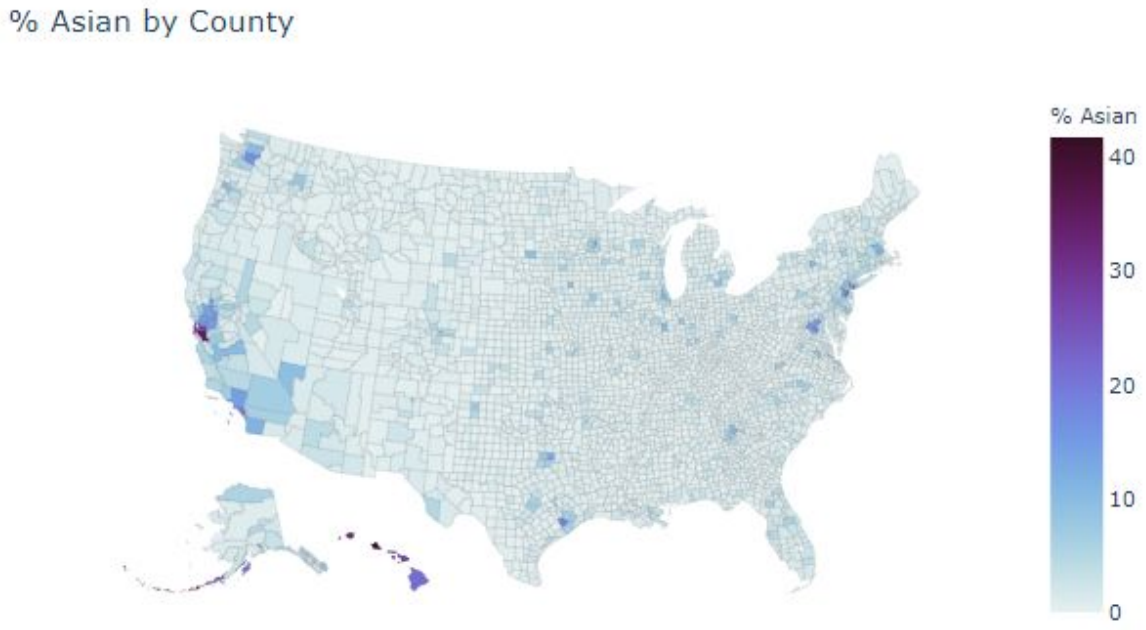
There is a statistically significant correlation between % White and % Self-Employed for counties in the USA.

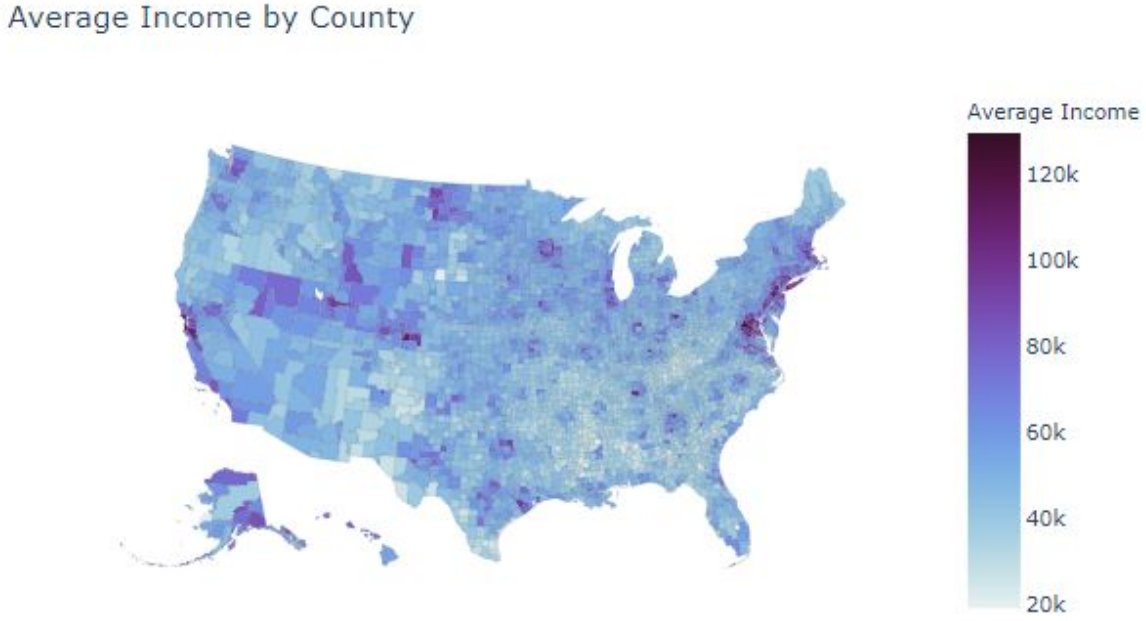
The correlation coefficient between % White and % Self-Employed is 0.234.

This means that % White and % Self-Employed are positively correlated.



Part C: Comparison of Percent Asian and Average Income by County





There is a statistically significant correlation between % Asian and Average Income for counties in the USA.

The correlation coefficient between % Asian and Average Income is 0.451.

This means that % Asian and Average Income are positively correlated.