Empirical Analysis: Comparing demographic trends across US states

Group members: Dylan Weiss, Kofi Addae-Sakyi, Zachary Zhang

Hypothesis 1: States with more land area tend to have more electoral votes

Explanation:
More land area in a state means more opportunity for big cities to form that are not near each other which draws more people and in turn means that there is more population. Additionally, many big cities which have a lot of population tend to form near good natural geographical features such as rivers or bays and when a state has more land there is more probability for more of these good areas for population growth. More population means more electoral votes.

Method:
We did a correlation analysis to test this hypothesis. We first extracted the land area and electoral votes data for each state from the stateInfo array obtained from our DataScraper. We then calculated the Pearson correlation coefficient between land areas and electoral votes using the calculatePearsonCorrelation method. The Pearson correlation coefficient ranges from -1 to 1, where: 1 indicates a perfect positive linear relationship; 0 indicates no linear relationship; and -1 indicates a perfect negative linear relationship.

Conclusion:
Given that we had a Pearson correlation coefficient of 0.132, this suggests a weak positive relationship between land area and electoral votes. We can conclude that larger states tend to have slightly more electoral votes, but the impact is not that significant.

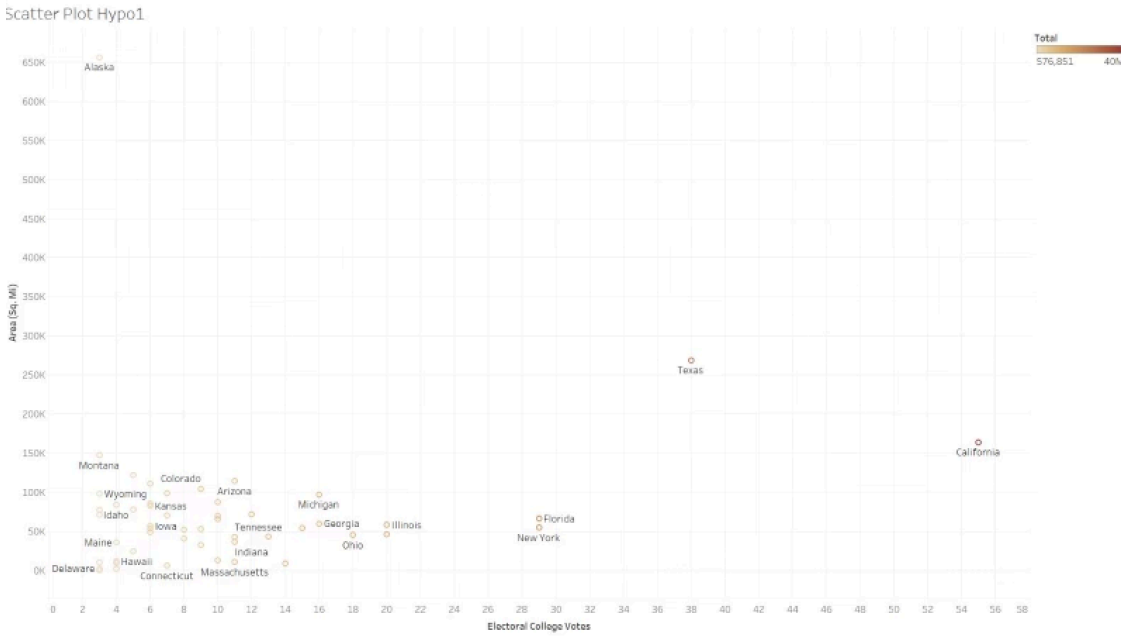

Figure 1: Scatter plot depicting state land area vs electoral college votes

Hypothesis 2: More populous states tend to have better education resources (reflected in the percentage of people with a high school degree and a bachelor's degree)

Explanation:
The areas that are more populated could collect more in taxes for education funding, which would make education more reachable for local residents. Furthermore, educational opportunities may also lead to a higher population in the area creating a positive feedback loop.

Method:
We used the data from the web scraping algorithm about each state and its respective population as well as a csv file with information on each state's percentage of people with a high school degree and a bachelor's degree. We then used hypothesis testing on the given data (ANOVA, t-test and correlation coefficient) for population vs high school degree and population vs bachelor degree to see the statistical difference between the groups. The results of these tests will help determine if the hypothesis is correct or not. Additionally, we visualised the data by making a bar chart of each state and its population as well as two scatter plots showing population vs high school degree as well as population vs bachelor degree in two graphs Tableau.
The education data used is obtained from US Census tables, using the data cleaned keeping the state name, percentage of 25+ aged people high school graduate or higher and the percentage of 25+ aged people with bachelor's degree or higher.
Calculation is done using the ttest_ind and f_oneway in scipy library, I calculated the the t-test score, ANOVA score and correlation coefficient between population and high school graduation rate / bachelor's degree rate, and here is the printed results:

Conclusion:
T-test Result:
Population and high school graduate or higher:
TtestResult(statistic=6.316806047334502, pvalue=7.065823672498182e-09)
Population and bachelors degree or higher:
TtestResult(statistic=6.316806594991512, pvalue=7.065805688846119e-09)

ANOVA Result:
Population and high school graduate or higher:
F_onewayResult(statistic=39.90203863964173, pvalue=7.065823672498275e-09)
Population and bachelors degree or higher:
F_onewayResult(statistic=39.902045558528265, pvalue=7.065805688846224e-09)

Correlation coefficient:
Population and high school graduate or higher: -0.47788784868743467
Population and bachelors degree or higher: 0.05724716936392127

We could see from the t-test and ANOVA results that the p-value is very small (less than threshold 0.05), which means that by a high probability that population has a relationship with the percentage of high school graduates or higher and bachelor's degree or higher.

However, our hypothesis is not correct. We predicted a positive correlation between both the percentage of high school or higher and bachelor or higher. But the correlation coefficient of population and percentage of high school graduates or higher indicated a rather significant negative relationship; And the low value of population and percentage of bachelor's degree or higher indicated a very weak positive influence on each other. These facts helped prove our hypothesis incorrect.

We had surmised that a state with a higher population would probably have a higher percentage of people with bachelor degrees because higher population usually indicates presence of big cities which draws in people with more education. But our data indicated that this is not necessarily the case and that in fact the opposite is true, and states such as Connecticut with lower populations have a higher percentage of people with bachelor degrees. It was very interesting to note how the graphs we plotted clearly indicated this negative correlation between size of population and percent of people with bachelor degrees which can be clearly seen. It was also interesting to see how our hypothesis testing further proved this point which disproved our hypothesis.

Additionally, the data on the high school graduates seemed to contradict the data on the bachelor degrees. Just going off of high school graduates showed that the states with a higher population tend to have a higher percentage of people who have graduated high school. The results of the hypothesis testing indicated a slight positive correlation between this trend, although not as strong as the negative correlation between the bachelor degrees and state populations. This was also very apparent as there wasn't that strong of a correlation on the visual graph we plotted as compared to the other graph. All in all it was interesting to see the results and compare it to our hypothesis and prove our original theory wrong. But it was also very interesting to see how states with large populations tend to have a lot of percentage of people who graduate high school and this may indicate good basic education in those states, btu that states with smaller population tend to attract more percentage of people compared to their overall percentage with bachelor degrees to live there.

Below are some figures created by Tableau for better understanding in our results:

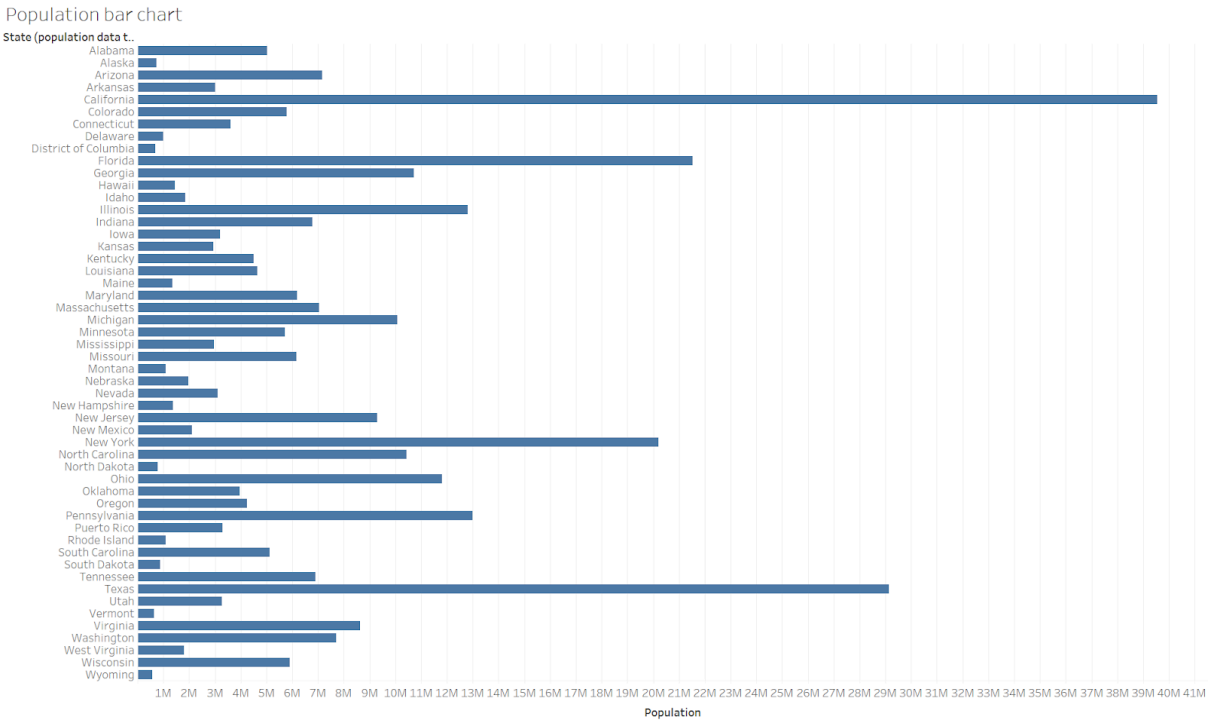

Figure 1: Bar chart depicting each state and its population

Figure 2: Scatter plot depicting state population vs % of adult population bachelor degree or higher

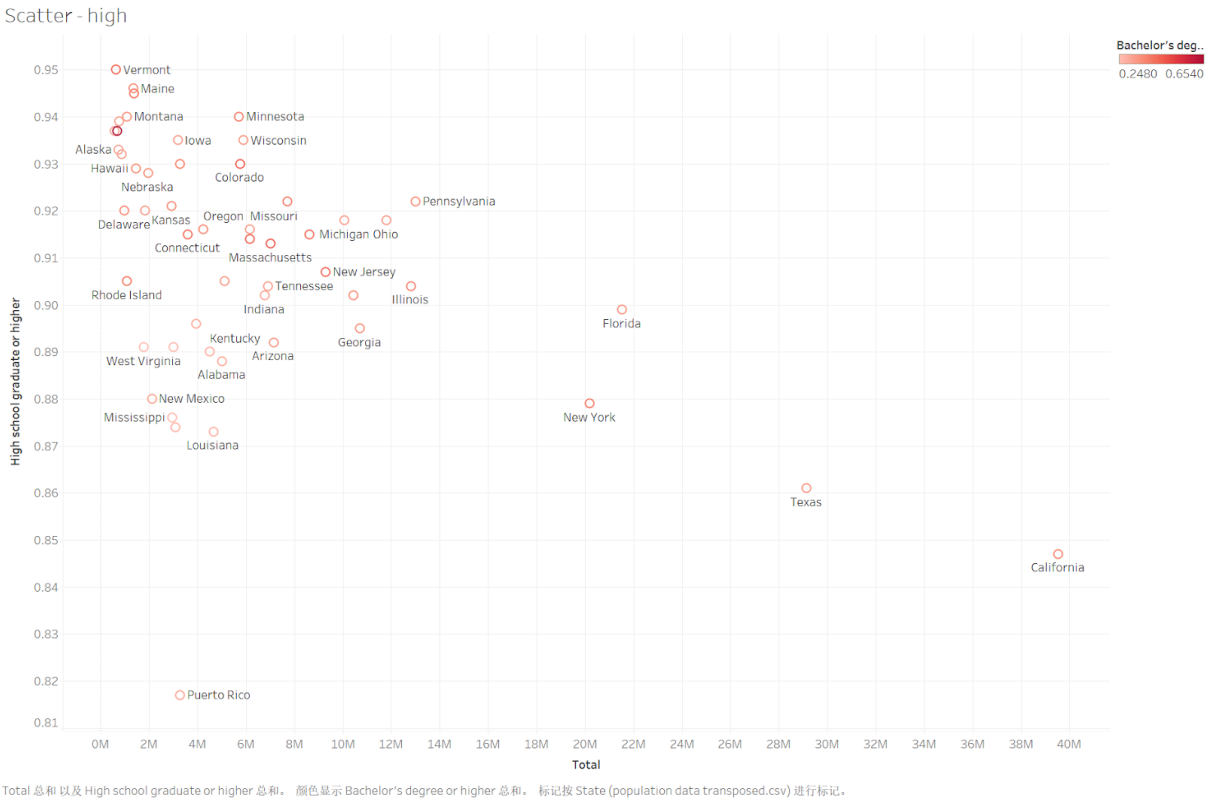

Figure 3: Scatter plot depicting state population vs % of adult population with high school degree or higher

<u>Hypothesis 3: Number of counties in a state and the population is proportional and steady across all the states</u>

Explanation:
As population grows and more people start to spread across the state, new counties may be created to properly serve the needs of these people. This could mean the larger the population in a state, the more counties there are likely to be.

Methods:
We are gathering the results from the web scraper and using Tableau to plot a line chart to show the ratio of population against county number (Population / County Number) across all states. If the line is rather horizontal, this could show that the ratio of population and number of counties are likely being proportional. Otherwise, it would prove that this hypothesis is incorrect.

Conclusion:
Given the line chart below, we could see that the population - county number ratio varies greatly across the states, and only a few are staying around the average line. Most are being distributed far away from the average line. This means that our hypothesis is incorrect and there are likely other more significant factors that would contribute to the number of counties, or even be due to the decision makers of each state, historical issues and more.

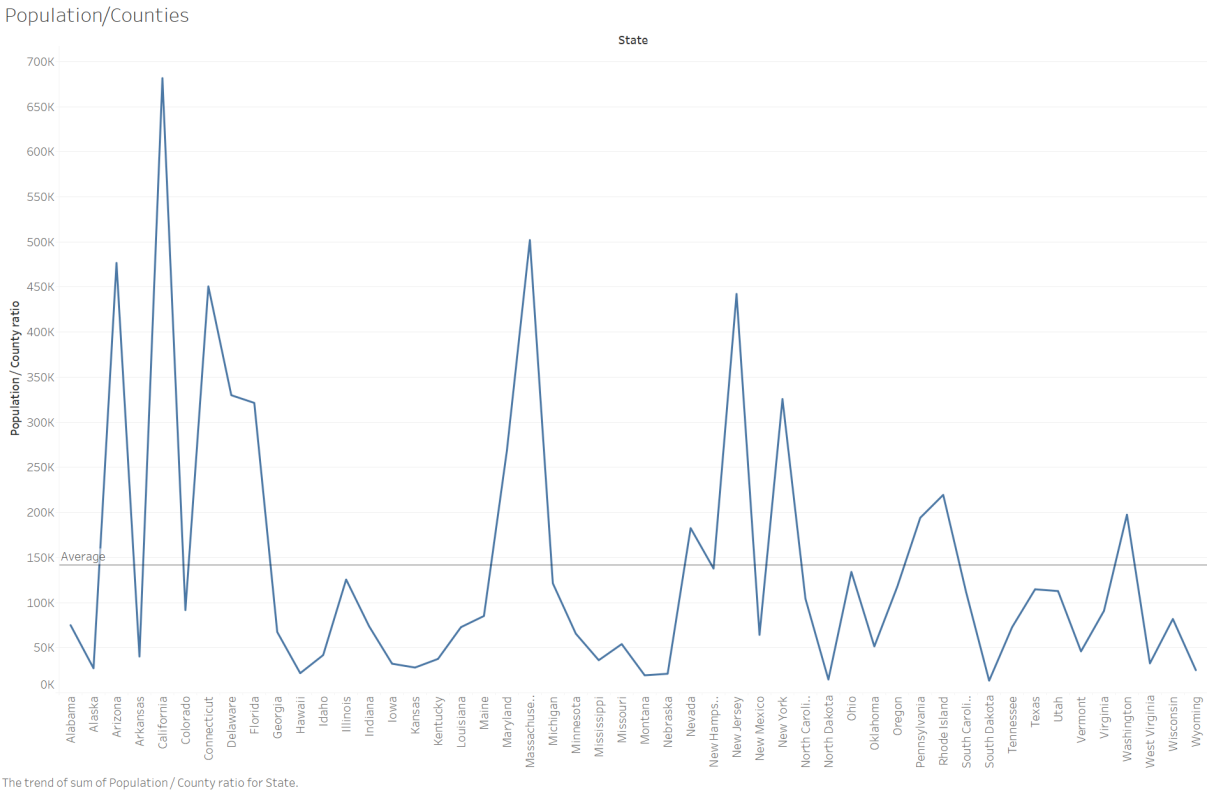

Figure 1: Line chart of the population - county number ratio (population / number of counties) of each state