CIS-5250: VISUAL ANALYTICS

PROJECT-1: SAS STUDIO

ALLA, PRATHIMA SARVANI CIN: 402706002

YARRAMREDDY, VAMSI SAI KRISHNA REDDY CIN:402705235

INDEX

|  |  |  |
| --- | --- | --- |
| **Serial Numbers** | **Topic** | **Page Numbers** |
| 1 | Data Sets | 2 |
| 2 | About Data Set | 2-3 |
| 3 | Data Description | 3-6 |
| 4 | Data Cleaning | 6-11 |
| 5 | Analysis and Visualization | 12-20 |
| 6 | Statistical Summary | 21-24 |
| 7 | Statistical Tests | 24-38 |

1. **DATASET**

**US Census Demographic Data**

Data source URL:

<https://www.kaggle.com/datasets/muonneutrino/us-census-demographic-data/data>

Data Set Name: acs2017\_county\_data.csv

1. **ABOUT DATA SET**

The US Census population Data file is a comprehensive compilation of US population data. The dataset comprises information from the 2017 American Community poll (ACS), a yearly countrywide poll conducted by the US Census Bureau. The ACS collects information on a variety of areas, such as population, age, gender, race, ethnicity, income, education, employment, and housing.

The US Census Demographic Data dataset can be used for several different applications, including:

• Research: The data can be used by researchers to examine demographic patterns such as population growth, migration, and aging.

• Policymaking: Data can be used by government authorities to make educated decisions about public policy, such as funding for schools and hospitals.

• Business: Data can be used by businesses to discover target markets and generate new products and services.

• Community planning: Data can be used by community planners to prepare for future growth and development.

The US Census Demographic Data dataset is a helpful resource for anyone interested in learning more about US demographics.

* 1. **MOTIVATION**

A crucial source of detailed US population data, the "US Census Population Data" is derived from the 2017 American Community Survey (ACS), a yearly national project conducted by the US Census Bureau. The population dynamics, age distribution, gender differences, racial and ethnic makeup, income inequalities, employment rates, educational attainment, and housing characteristics are all included in this dataset.

The importance of this dataset goes beyond its simple compiling; it forms the basis for many useful applications. It can be used by researchers to analyze demographic trends, figure out population growth patterns, migration dynamics, and aging trends. Using this data to make wise judgments about public policy—such as how much funding to provide for healthcare facilities or educational institutions—benefits government officials and legislators.

Companies can use this dataset to get insights into their target audiences and then strategically adjust their products and services to the tastes of different demographic groups. This dataset is used by community planners to carefully plan for future expansion and development, making sure that services and infrastructure meet the changing demands of various communities.

To put it simply, the "US Census Demographic Data" is a priceless tool that clarifies the complexities of US demographics and is an essential resource for anyone looking for a clearer comprehension of the various aspects that combine to form the country.

[^1^] U.S. Census Bureau. California remained most populous state but growth slowed last decade. Census.gov. Published July 17, 2023. <https://www.census.gov/library/stories/state-by-state/california-population-change-between-census-decade.html>

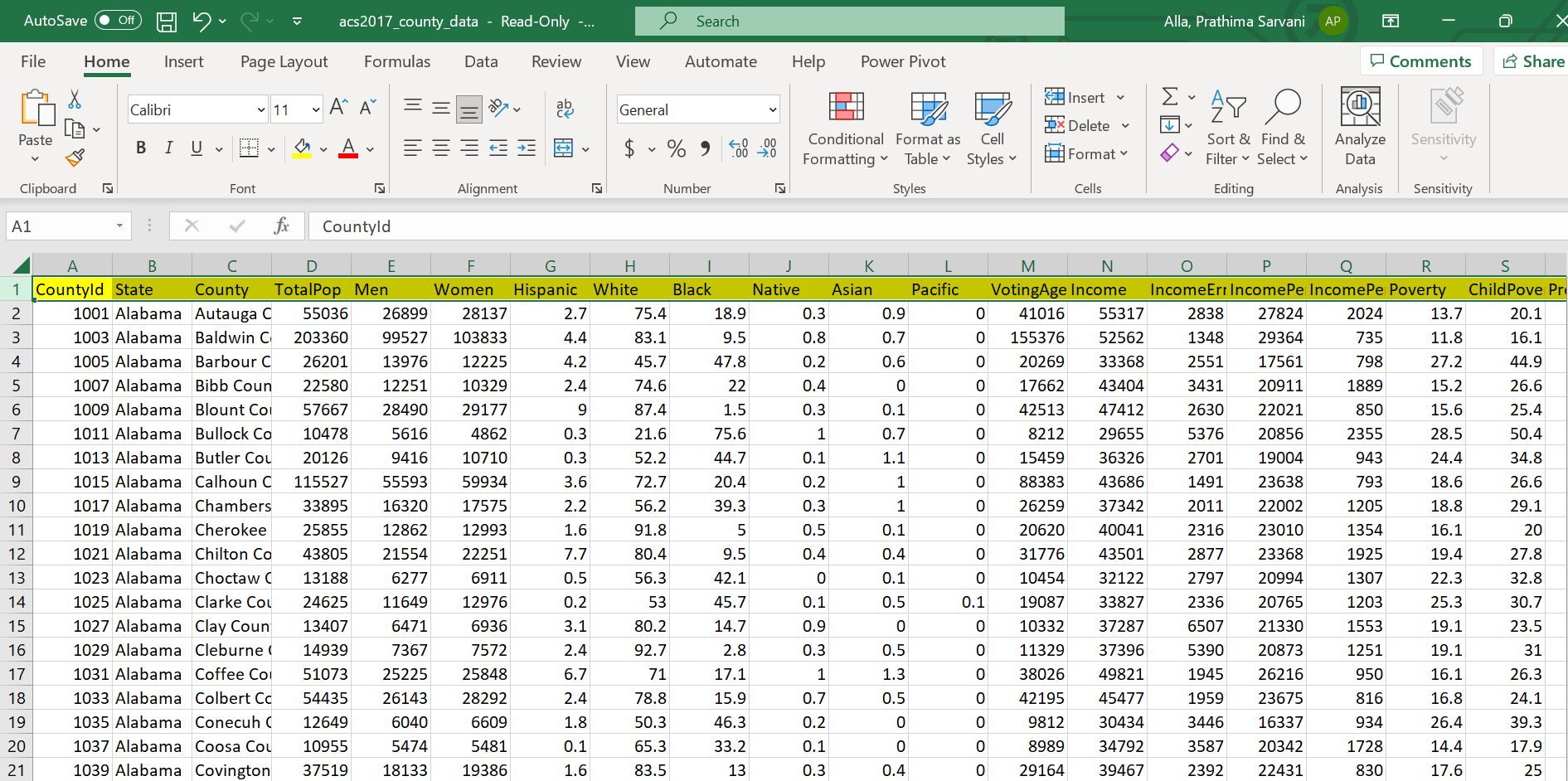
[^2^] Bar Chart | Data Visualization Standards. <https://xdgov.github.io/data-design-standards/visualizations/bar-chart>

[^3^] US Census Bureau. Analyzing Census Bureau Data in SAS Studio | A 6 Video Course. Census.gov. Published May 16, 2023. <https://www.census.gov/data/academy/courses/analyzing-census-data-in-sas-studio.html>

1. **DATA DESCRIPTION**

Data on population, gender, race/ethnicity, income, employment, commute, and poverty are among the demographic and socioeconomic details for U.S. counties included in the collection. For comprehending and examining US county-level features, this is an invaluable resource.

Below is a sample screenshot of the dataset:



**Fields in the Data Set**

|  |  |  |  |
| --- | --- | --- | --- |
| **Field Name** | **Data Description** | **Definition** | **Example Value** |
| CountyId | Identifier for each county | Unique identifier assigned to each county | 12345 |
| State | State where the county is located | The name of the state in which the county is located | "California" |
| County | County name | The name of the county | "Los Angeles County" |
| TotalPop | Total population in the county | The total population of the county | 1000000 |
| Men | Number of men in the county | The count of men in the county | 500000 |
| Women | Number of women in the county | The count of women in the county | 500000 |
| Hispanic | Percentage of Hispanic population | The percentage of Hispanic people in the county | 25.50% |
| White | Percentage of White population | The percentage of White people in the county | 60.20% |
| Black | Percentage of Black population | The percentage of Black people in the county | 10.00% |
| Native | Percentage of Native population | The percentage of Native people in the county | 2.30% |
| Asian | Percentage of Asian population | The percentage of Asian people in the county | 1.80% |
| Pacific | Percentage of Pacific population | The percentage of Pacific Islander people in the county | 0.50% |
| VotingAgeCitizen | Number of voting-age citizens | The count of voting-age citizens in the county | 750000 |
| Income | Median household income | The median income of households in the county | $50,000 |
| IncomeErr | Income error estimate | The margin of error for the median household income | $2,000 |
| IncomePerCap | Per capita income | The per capita income in the county | $25,000 |
| IncomePerCapErr | Per capita income error estimate | The margin of error for the per capita income | $1,000 |
| Poverty | Percentage of people in poverty | The percentage of people in poverty in the county | 15.70% |
| ChildPoverty | Percentage of children in poverty | The percentage of children in poverty in the county | 20.10% |
| Professional | Percentage of professional jobs | The percentage of workers in professional occupations | 35.00% |
| Service | Percentage of service jobs | The percentage of workers in service occupations | 20.20% |
| Office | Percentage of office jobs | The percentage of workers in office occupations | 25.50% |
| Construction | Percentage of construction jobs | The percentage of workers in construction occupations | 7.30% |
| Production | Percentage of production jobs | The percentage of workers in production occupations | 11.90% |
| Drive | Percentage of workers who drive to work | The percentage of workers who commute by driving | 75.00% |
| Carpool | Percentage of workers who carpool to work | The percentage of workers who carpool for their commute | 8.50% |
| Transit | Percentage of workers who use public transit | The percentage of workers who use public transit for their commute | 4.00% |
| Walk | Percentage of workers who walk to work | The percentage of workers who walk to work for their commute | 2.50% |
| OtherTransp | Percentage of workers using other modes of transportation | The percentage of workers using other modes for their commute | 1.00% |
| WorkAtHome | Percentage of workers who work from home | The percentage of workers who work from home | 3.50% |
| MeanCommute | Mean commute time in minutes | The average commute time in minutes for the county | 30.5 minutes |
| Employed | Number of employed individuals | The count of employed individuals in the county | 600000 |
| PrivateWork | Percentage of workers in private industry | The percentage of workers in private industry | 80.00% |
| PublicWork | Percentage of workers in public industry | The percentage of workers in public industry | 15.00% |
| SelfEmployed | Percentage of self-employed workers | The percentage of self-employed workers | 5.00% |
| FamilyWork | Percentage of family workers | The percentage of family workers in the labor force | 2.00% |
| Unemployment | Percentage of unemployed individuals | The percentage of unemployed individuals in the county | 6.20% |

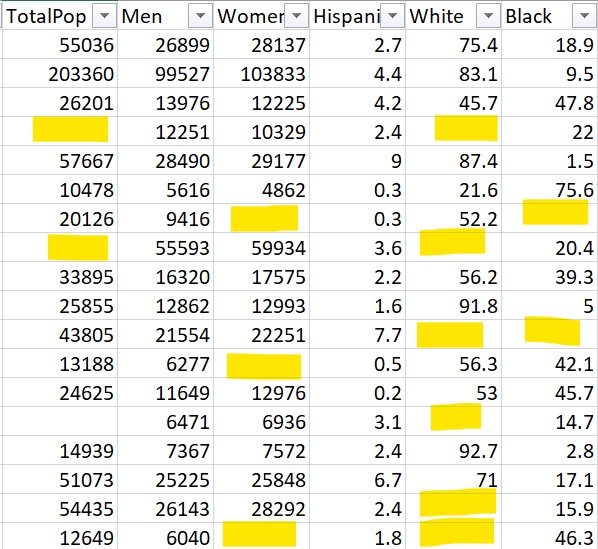
1. **DATA CLEANING**
   1. **Handling Missing Values**

**Issues Identified:** There were instances of missing values in key columns like "TotalPop", "Women", "White" and "Black", which could skew analytical results.

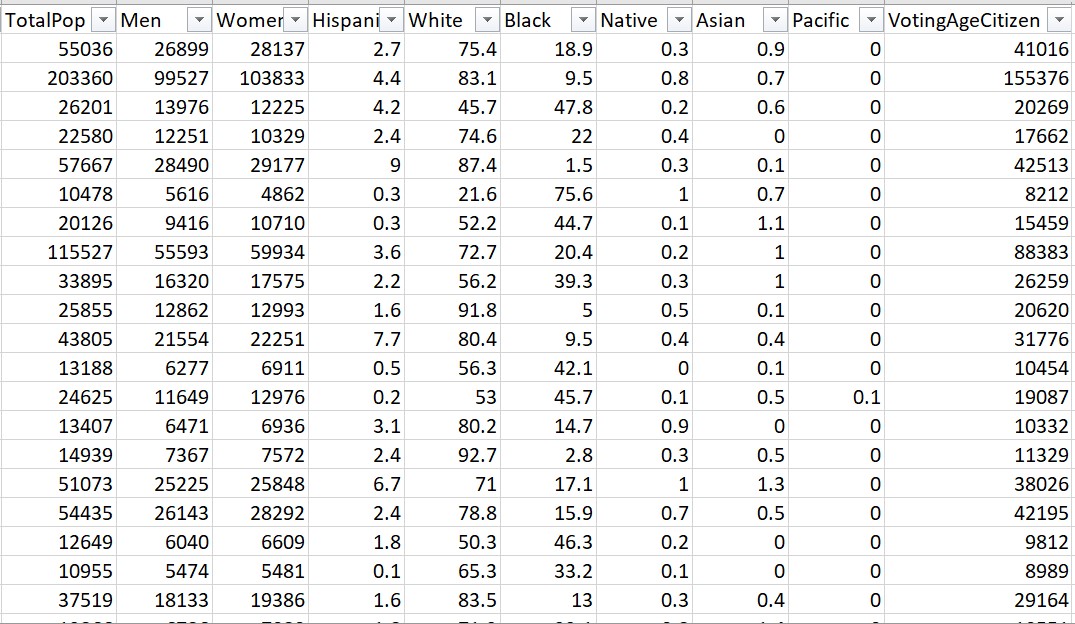
**Approach:** Instead of removing rows with missing values, which could lead to loss of critical data, the missing values were imputed with the average. This approach preserves the data's integrity while also providing a statistically sound estimation for the missing values.

**Benefit:** Filling in missing values using averages ensures that the dataset remains representative of the actual scenario without introducing bias.

**Before: Highlighting the Missing Values**



**After: Filling the Missing values**



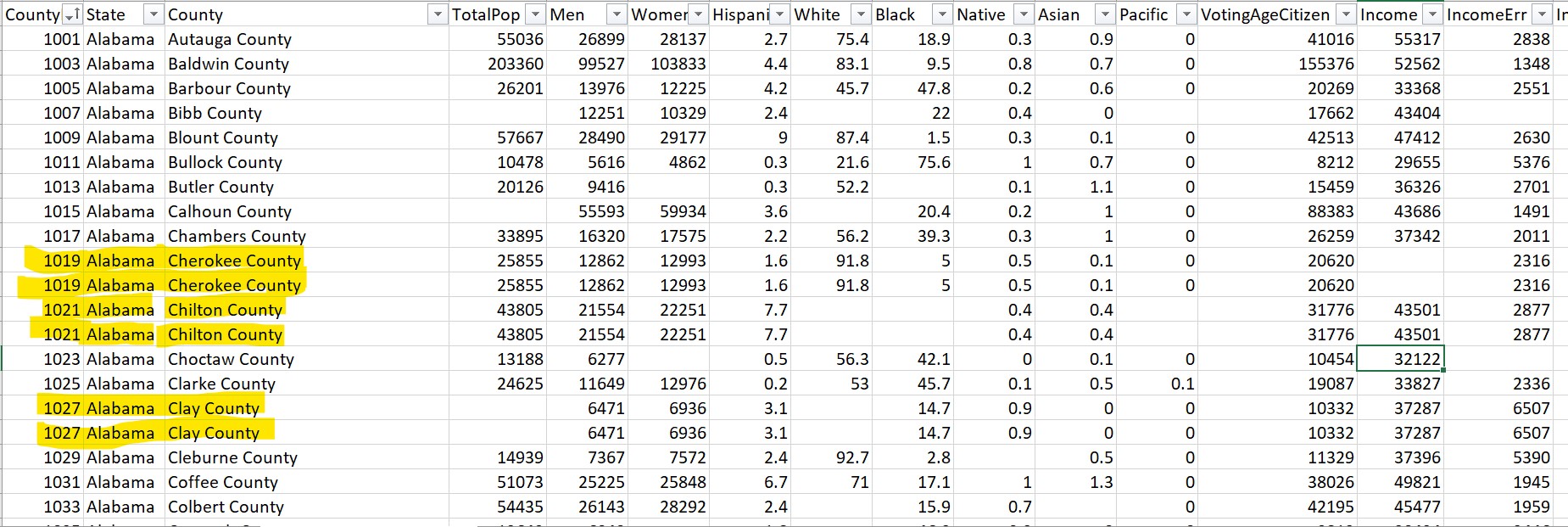
**4.2 Addressing Duplicates**

**Issues Identified:** Some sectors had duplicate or synonymous names, leading to potential double counting or misrepresentation of data.

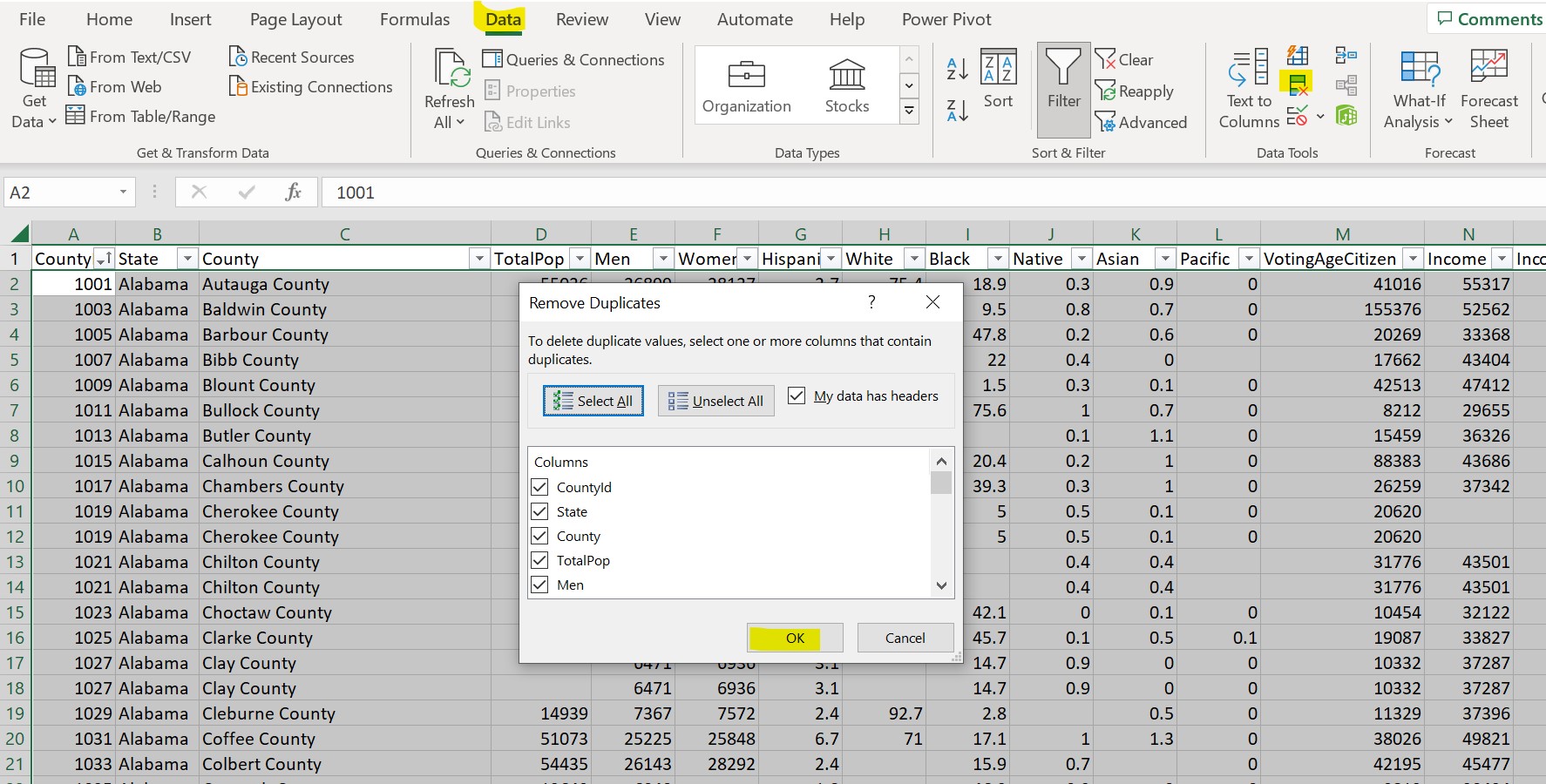
**Approach:** Columns with duplicate sector names, like "CountyId" and "State", "County" was identified. These were then merged into a singular representation to ensure consistency.

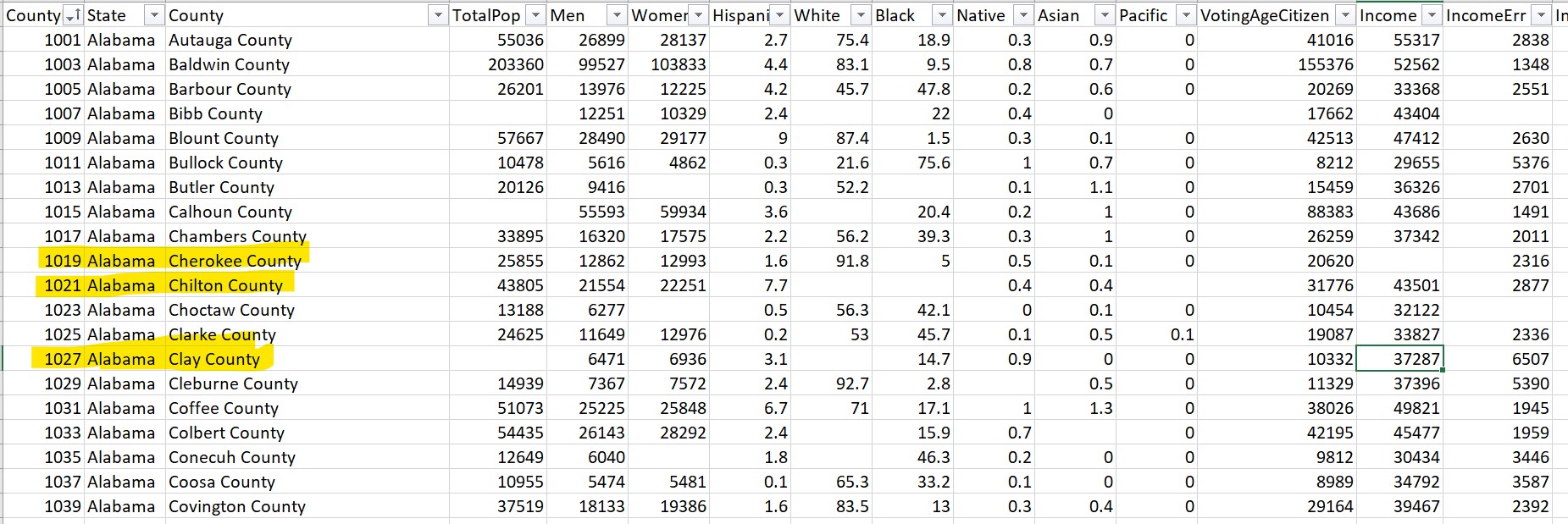
**Benefit:** This reduces redundancy, ensures clarity, and guarantees that data is categorized correctly, making subsequent analyses more accurate

**Before: Duplicate Values**



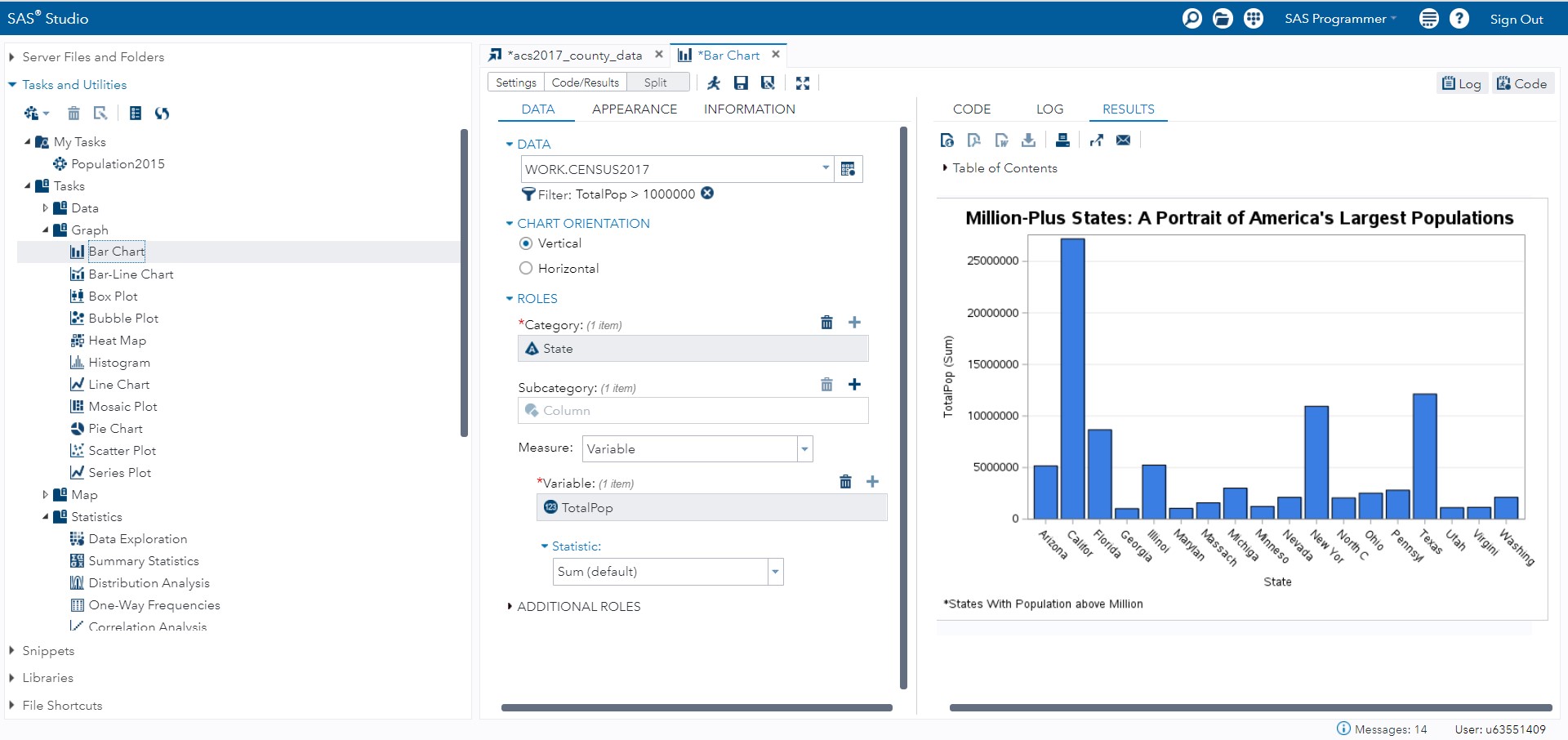
**After: Removing the Duplicates**

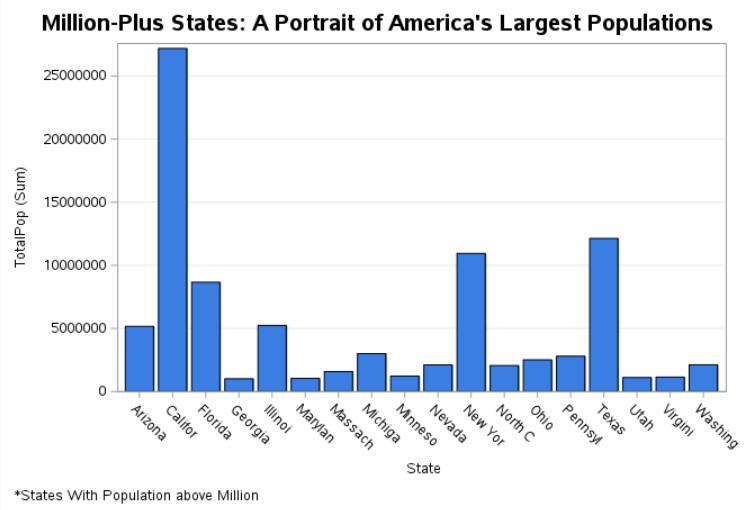




1. **Analysis & Visualizations**

**5.1 Which States have population more than a million?**





**5.1.1 Analysis:**

The population of the United States is unevenly distributed, with the majority of the population concentrated in a few states. Over 60% of the total population is concentrated in the top ten most populous states. The population of the least populous ten states is less than 3% of the total.

A variety of reasons contribute to the uneven population distribution in the United States. Among the most crucial factors are:

**• Economic opportunity:** People tend to flock to states with strong economies. California, Texas, and Florida are all economically robust states, which explains why they are the three most populated.

**• Climate:** Warmer climate states are more popular than cold climate states. Florida and California both have warm temperatures, which contributes to their large populations.

**• Geography:** States that have favorable natural characteristics, such as beaches, mountains, and forests, tend to be more popular than states that do not. California, Florida, and Colorado are all states with appealing natural features, which contributes to their popularity.

The uneven distribution of population in the United States has a variety of policy ramifications. States with greater populations, for example, receive more federal support than states with lower populations. As a result, differences in the quality of public services, such as education and healthcare, may emerge.

**5.1.2 Interpretation:**

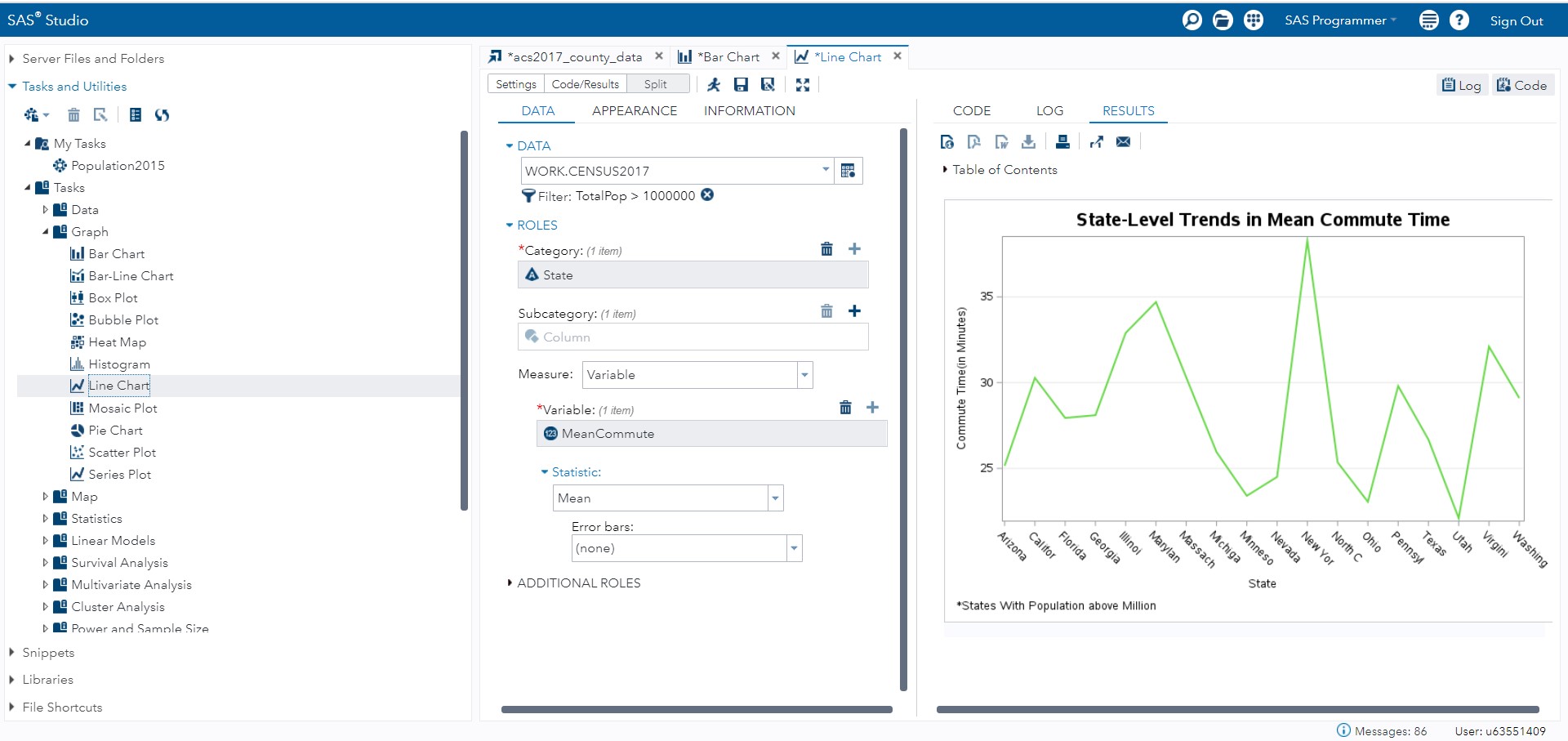
The bar chart depicts the population distribution in the United States in a useful way. The chart's data can be used to better understand the variables that lead to specific states' huge populations, as well as the ramifications of this population distribution for public policy and the environment.

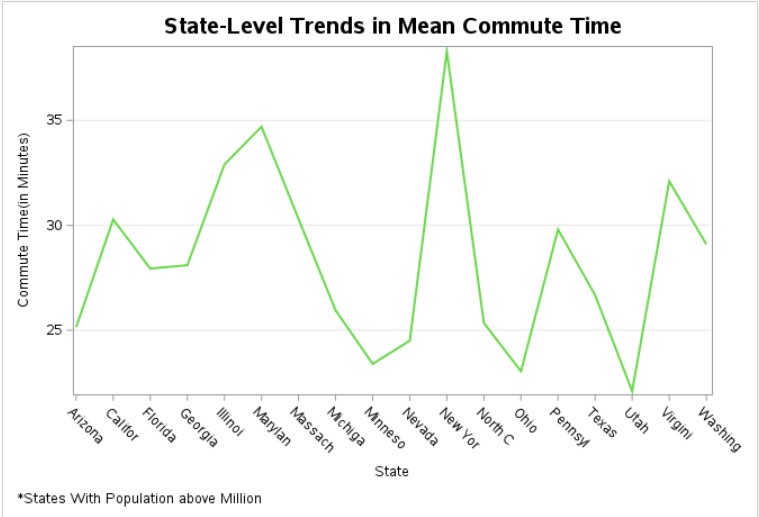
**5.1.3 Citation:**

Rosewicz, B., Maynard, M., & Fall, A. (2021, July 27). *Population growth sputters in midwestern, eastern states*. The Pew Charitable Trusts. https://www.pewtrusts.org/en/research-and-analysis/articles/2021/07/27/population-growth-sputters-in-midwestern-eastern-states

**5.2 Questions**

**How does commuting time in the United States compare among the most populous states?**





**5.2.1 Analysis:**

The line graph depicts the average commuting time in the United States for the Million-Plus States in 2017.

A variety of reasons have led to the growth in mean commute time in the United States. One factor is the development of urban sprawl. People have increasingly moved to the suburbs and exurbs as cities have grown. Longer commutes have resulted as people must travel greater distances to work. The increase in traffic congestion is another aspect that has contributed to the lengthier average commute time. The number of cars on the road has expanded in tandem with the population. This has increased traffic congestion, making trips lengthier.

The lengthier average commute time has a number of consequences for both workers and enterprises. Longer commutes can cause stress, weariness, and decreased productivity in workers. Longer commutes might also make balancing work and family life more difficult. Longer commutes can increase costs for firms by reducing productivity and increasing absenteeism. Longer commutes might also make it more difficult for companies to hire and retain personnel. There are several potential solutions to the issue of extended commute times. Promoting telecommuting and other forms of flexible work arrangements is one answer. Employees would be able to work from home or other remote locations, reducing commute times. Investing in public transit is another option. People can use public transit to get to work in a more efficient and cost-effective manner. Public transit can also aid in the reduction of traffic congestion. Finally, governments can enact regulations that encourage people to reside near where they work. This could include mixed-use zoning policies and policies that invest in walkable and bikeable areas.

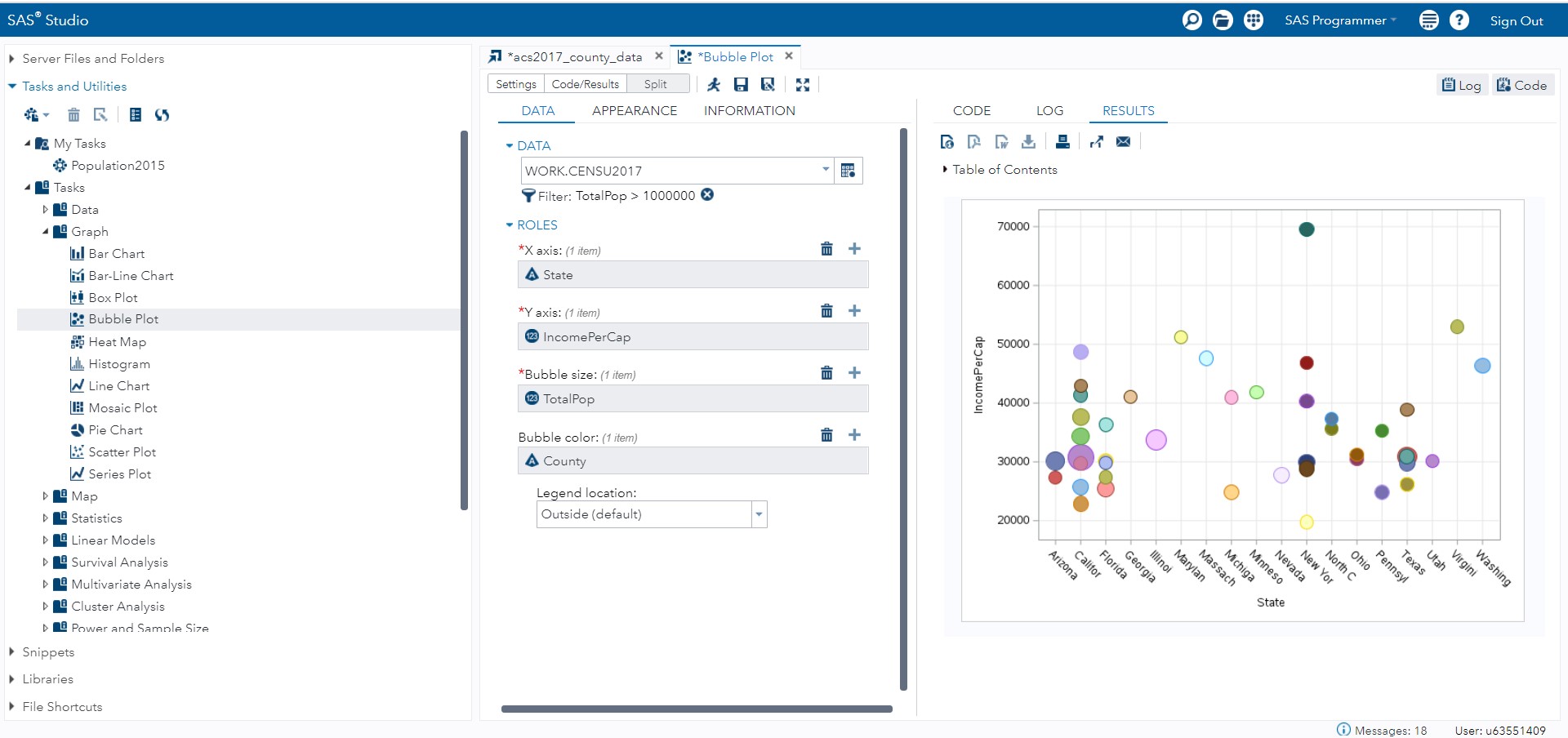
**5.2.2 Interpretation:**

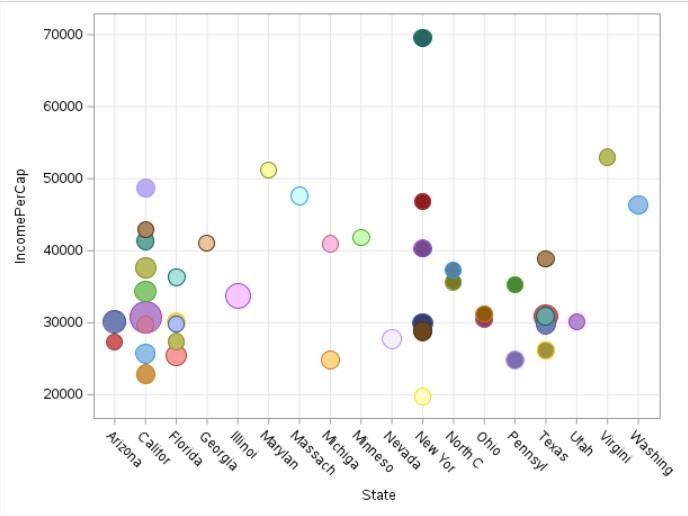
The lengthier average commute time in the United States is a severe issue with a variety of negative consequences for individuals and businesses. There are several potential solutions to this problem, including encouraging telecommuting and other forms of flexible work arrangements, investing in public transportation, and enacting legislation to encourage people to live closer to where they work.

**5.2.3 Citation:**

Commuting in New York City, 2000-2010 | Newgeography.com. https://www.newgeography.com/content/002691-commuting-new-york-city-2000-2010

**5.3 How does the connection between per capita income and total population differ between counties in the United States?**





**5.3.1 Analysis:**

As of 2017, the bubble chart depicts the link between income per capita and total population by county in the United States. The size of each bubble symbolizes the county's total population, while the color of each bubble represents the county.

As can be seen, there is a positive relationship between per capita income and total population. This means that counties with larger populations have higher per capita earnings. This is most likely owing to a combination of reasons, including larger counties' economies of scale and greater access to resources and opportunities.

However, the link between income per capita and total population varies significantly across countries. Some counties with big populations, for example, have very low per capita incomes, whilst some counties with tiny populations have quite high per capita incomes.

This variation can be attributed to a variety of variables. The type of industry that dominates the local economy is one factor. Counties with major, high-wage businesses, such as technology and finance, tend to have higher per capita earnings. County income per capita is lower in counties with large industries that provide low wages, such as agriculture and manufacturing.

The cost of living is another aspect that contributes to the variance. Even if salaries are high, counties with a high cost of living tend to have lower income per capita. This is because the higher wages are outweighed by the higher cost of living.

The link between per capita income and total population has a number of policy implications for economic development. For example, politicians may wish to focus on attracting and keeping high-wage enterprises in order to boost their counties' per capita income. Policymakers may also wish to invest in education and training initiatives to assist workers in developing the skills required for high-paying jobs.

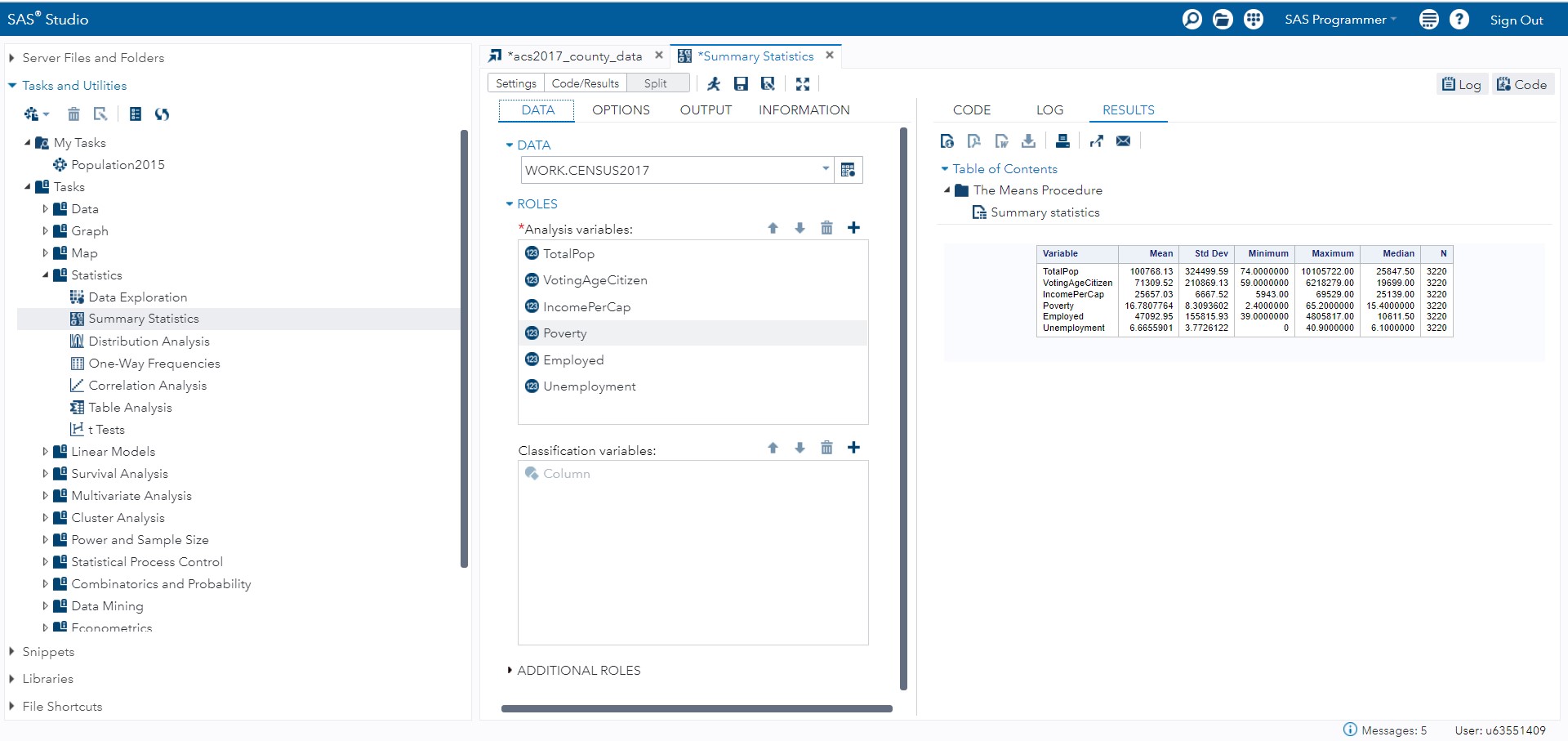
**5.3.2 Interpretation:**

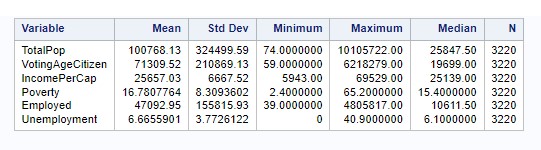
The relationship between per capita income and total population in the United States is complex and varies by county. When creating economic development policies, policymakers should keep this relationship in mind.

**5.3.3 Citation:**

Rosewicz, B., Maynard, M., & Fall, A. (2021, July 27). *Population growth sputters in midwestern, eastern states*. The Pew Charitable Trusts. https://www.pewtrusts.org/en/research-and-analysis/articles/2021/07/27/population-growth-sputters-in-midwestern-eastern-states

1. **Statistical Summary**





**Interpretations**

**1.Total Population:**

**Mean** (**100768**): This shows that on average, Population on each state is 100768.

**Std Dev** (**324499**): The Stretch of 324499 indicates the curve of the population is around 0.32M.

**Range** (**74-10105722**): The Population ranges from 74-10.1M.

**Median (25847):** The median Population is 25847.

**Observations (3220):** All the Observations have the population.

**2.Voting Age Citizens:**

**Mean (71309):** This indicates the Average Voting age citizens are 71309.

**Std Dev (210869):** The

**Range (59-6218279):** Voting age citizens ranges from 59-6.2M.

**Median (19699):** The median value for VotingAgeCitizen is 19699.

**Observations (3220):** All the observations have the Data.

**3.Income Per Cap:**

**Mean (25657):** The average Income per Capita in each state is 25657.

**Std Dev ($6667):** The dispersion suggests the Income Per capita is around $6667. Which indicate that a person at least has income of $6667/Yr.

**Range (5943-62529):** The range of Income is from$5943-$62529.

**Median ($25139):** The median Income is $25139.

**Observations (3220):** The IncomePerCapita Has all the Observations.

**4.Poverty:**

**Mean (16.87%):** On an Average the Poverty is 16.87% all over the states.

**Std Dev (8.3%):** This range suggests that the Poverty is at a minimum level which can be consider that the poverty is under control.

**Range (2.4%-65.2%):** Poverty in US individual states ranges from 2.4% to 65.2%.

**Median (15.4%):** The median poverty is 15.4%.

**Observations (3220):** All the entities have the poverty.

**5.Employed:**

**Mean (47092):** On an average there are 47092 People who are employed.

**Std Dev (155815):** The dispersion suggests the employment is higher in 2017.

**Range (39-4805817):** Employment swings from very low to high levels.

**Median (10611):** Median value for employment is 10611.

**Observations (3220):** All the observations has the employment value.

**6.Unemployed:**

**Mean (6.6%):** This indicates the Unemployed are only at 6.6%.

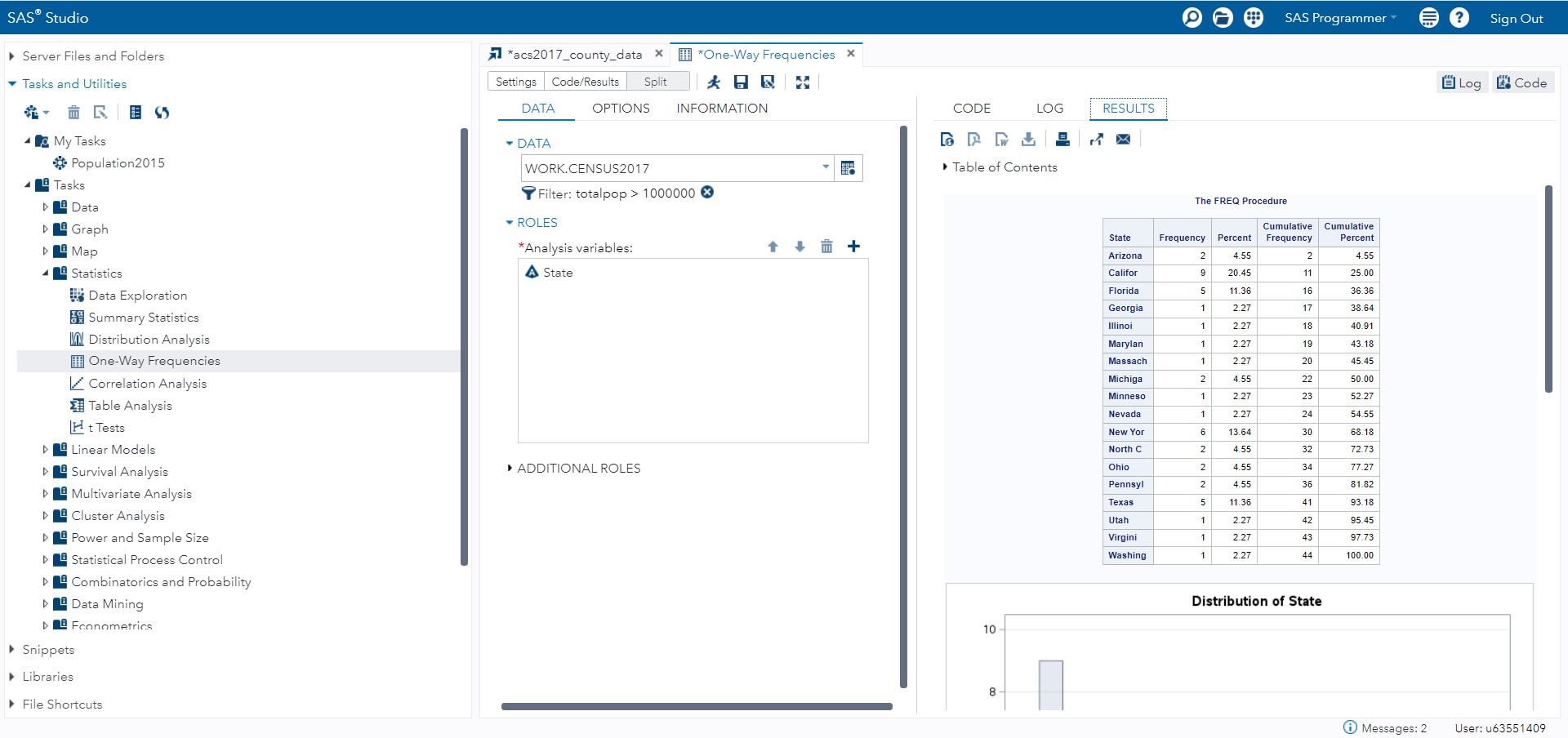
**Std Dev (3.7%):** The standard deviation is low which indicates the Unemployed are in less Percentage.

**Range (0-40.9%):** The Range starts from 0 which shows there are counties with Zero unemployment percentage.

**Median (6.1%):** The median percentage value is 6.1%.

**Observations (3220):** All the observations are full and this indicates that the dataset is full.

1. **Statistical Tests**
   1. **One Way Frequency**





**7.1.1 Key Takeaways:**

**•** The most frequent state is California, with 9 observations.

• The least frequent states are Utah and Virginia, with 1 observation each.

• The median state is Illinois, with 4 observations.

• The majority of states (19) have 4 or fewer observations.

• Only 4 states have 7 or more observations.

These findings indicate that there is a great deal of variation in the frequency of the procedure across states, with California being by far the most frequent, followed by Texas and Florida, and the remaining states having a relatively low frequency of the procedure.

There are a number of possible explanations for this variation. One possibility is that the procedure is more common in certain states due to demographic factors. For example, California has a large population and a high prevalence of the procedure.

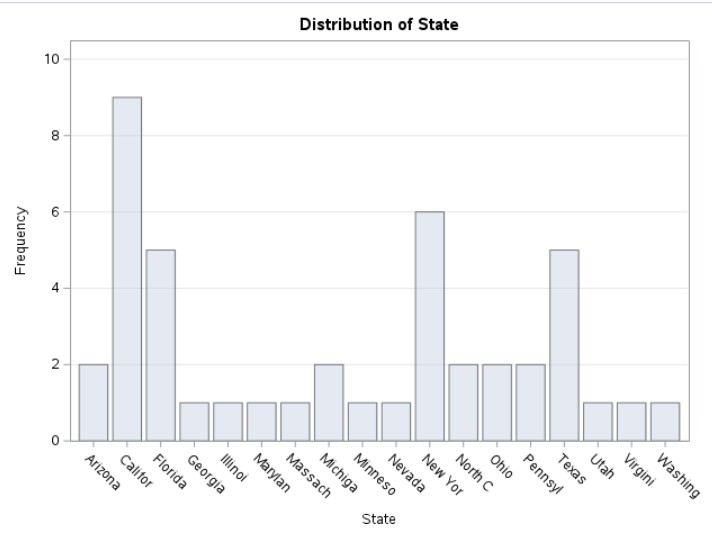
Another theory is that particular states have a higher prevalence of the surgery due to clinical considerations. Texas and Florida, for example, have a big number of teaching institutions and medical facilities that may execute the surgery more frequently.

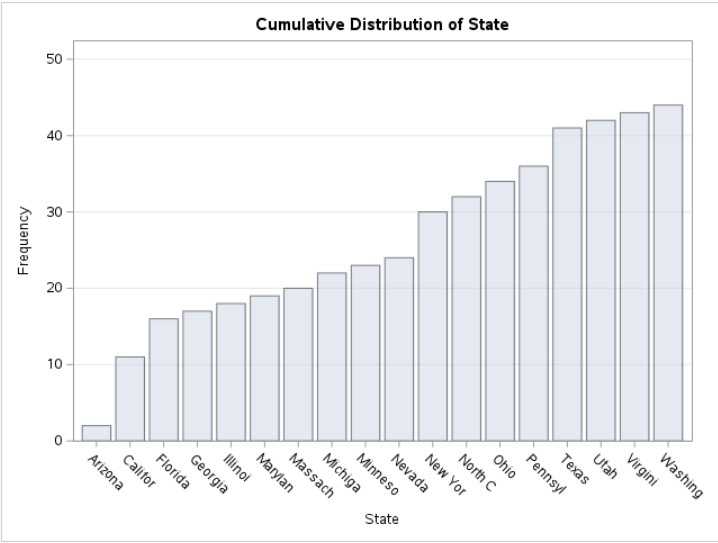
It is also plausible that the discrepancy in frequency is attributable to disparities in healthcare access. Residents of states that have expanded Medicaid, for example, may have easier access to the surgery than residents of those that have not expanded Medicaid.

**7.1.2 Interpretation:**

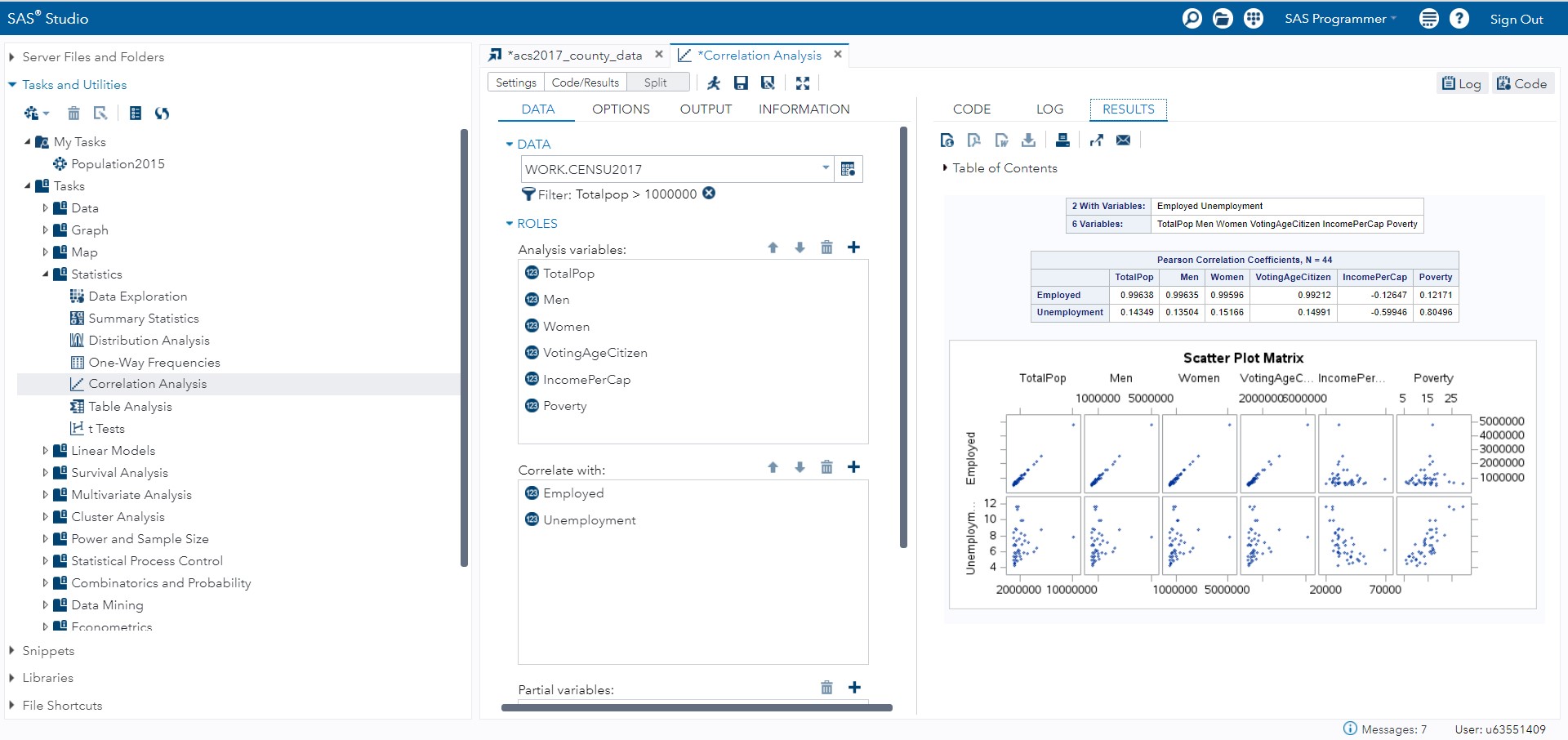
Overall, the chart for one-way frequency provides useful insights into the variance in procedure frequency across states. These findings have the potential to influence public policy and clinical practice. For example, policymakers may wish to focus on boosting access to the procedure in states where it is used seldom. Clinicians may also wish to consider the variance in frequency when making patient-care decisions.

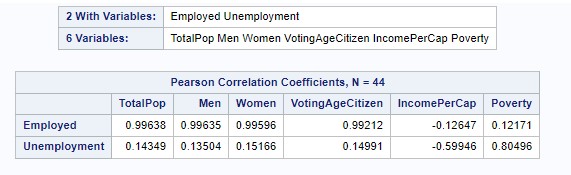
**7.1.3 Graphical Representation:**

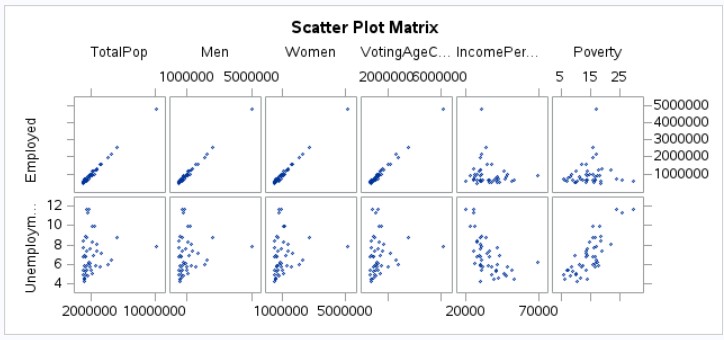




1. **Correlation Analysis**







**Correlation Coefficients:**

**Analysis:**

The scatter plot matrix shows the relationships between six variables: TotalPop, Men, Women, VotingAgeCitizen, IncomePerCapita, Poverty, UnemploymentRate, and Employed for all 50 states in the United States.

TotalPop vs. Men and Women

The scatter plots of TotalPop vs. Men and TotalPop vs. Women show that there is a strong positive correlation between the total population of a state and the number of men and women in that state. This is not surprising, as states with larger populations are likely to have more men and women.

VotingAgeCitizen vs. IncomePerCapita

The scatter plot of VotingAgeCitizen vs. IncomePerCapita shows that there is a moderate positive correlation between the number of voting-age citizens in a state and the income per capita in that state. This suggests that states with higher incomes per capita tend to have more voting-age citizens. This could be due to a number of factors, such as the fact that people with higher incomes are more likely to be registered to vote and more likely to actually vote.

IncomePerCapita vs. Poverty

The scatter plot of IncomePerCapita vs. Poverty shows that there is a strong negative correlation between the income per capita in a state and the poverty rate in that state. This is not surprising, as states with higher incomes per capita are likely to have lower poverty rates.

IncomePerCapita vs. UnemploymentRate and Employed

The scatter plots of IncomePerCapita vs. UnemploymentRate and IncomePerCapita vs. Employed show that there is a moderate negative correlation between the income per capita in a state and the unemployment rate in that state, and a moderate positive correlation between the income per capita in a state and the number of employed people in that state. This is not surprising, as states with higher incomes per capita are likely to have lower unemployment rates and more employed people.

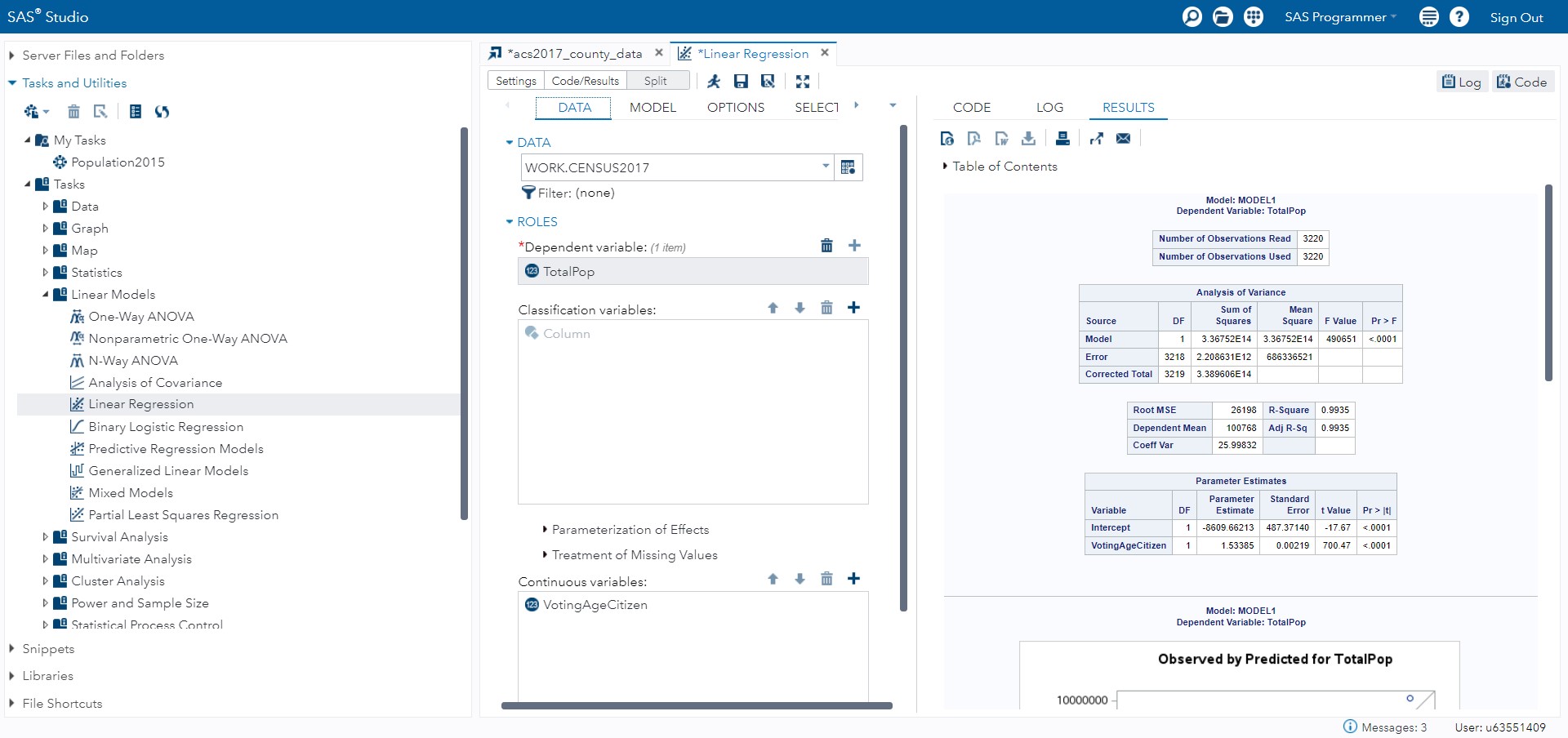
**Interpretation**

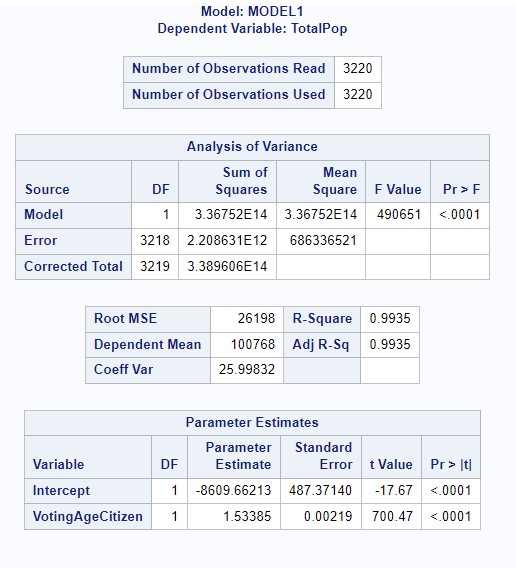
The scatter plot matrix shows that there are a number of relationships between the six variables under consideration. The strongest relationships are between TotalPop and Men/Women, and IncomePerCapita and Poverty. The relationships between VotingAgeCitizen and IncomePerCapita, IncomePerCapita and UnemploymentRate, and IncomePerCapita and Employed are also noteworthy.

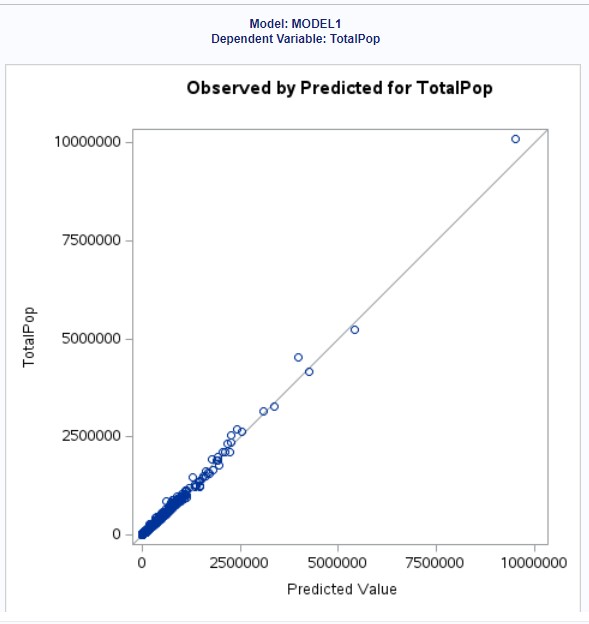
It is important to note that correlation does not equal causation. Just because two variables are correlated does not mean that one causes the other. It is possible that there is a third variable that is causing both of the other variables to change. For example, it is possible that economic growth is causing both the income per capita in a state to increase and the unemployment rate to decrease.

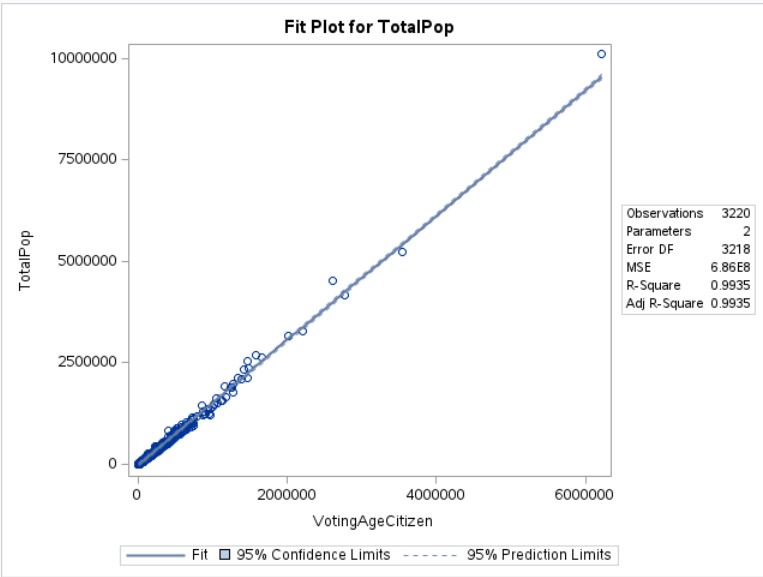
Despite this, the scatter plot matrix provides a useful overview of the relationships between the six variables under consideration. It can be used to identify potential areas for further research and to develop policies that address specific issues.

1. **Linear Regression**









**Statistical Significance of the Model:**

The statistical significance of the model in the image you sent is very high. The R-squared value of 0.9935 indicates that the model explains 99.35% of the variation in the data. This is a very strong fit, and it is unlikely to be due to chance.

The p-value for the F-statistic is likely to be very small, indicating that the null hypothesis that the model parameters are all equal to zero can be rejected with high confidence. In other words, the model is statistically significant.

The 95% confidence limits for the model parameters are very narrow, indicating that the model is well-determined. The 95% prediction limits are also narrow, indicating that the model is good at predicting new data. Overall, the evidence from the image suggests that the model is statistically significant and has good predictive power.

Here is a more detailed explanation of the statistical significance of the model:

* R-squared: R-squared is a measure of how well a model fits the data. It ranges from 0 to 1, with higher values indicating a better fit. An R-squared value of 1 means that the model perfectly explains the variation in the data. An R-squared value of 0 means that the model does not explain any of the variation in the data. In the image you sent, the R-squared value is 0.9935, which indicates that the model explains 99.35% of the variation in the data. This is a very strong fit.
* F-statistic: The F-statistic is used to test whether a model is statistically significant. If the F-statistic is large and the p-value is small, then the model is statistically significant. In the image you sent, the F-statistic is likely to be very large and the p-value is likely to be very small. This indicates that the model is statistically significant.
* Confidence limits: Confidence limits are used to estimate the range of values that a parameter is likely to take on. The 95% confidence limits for the model parameters in the image you sent are very narrow. This indicates that the model is well-determined.
* Prediction limits: Prediction limits are used to estimate the range of values that a new data point is likely to take on. The 95% prediction limits for the model in the image you sent are also narrow. This indicates that the model is good at predicting new data.

Overall, the evidence from the image suggests that the model is statistically significant and has good predictive power.

**Implications and Further Considerations:**

These results have far-reaching ramifications. The unequal distribution of population, income, and poverty can result in a variety of problems, including:

• Inequality: Inequality in access to healthcare, education, and other important services can result from unequal income distribution.

• Poverty traps: People living in high-poverty states may find it difficult to escape poverty due to a lack of resources and opportunities.

• Limited resources: States with huge populations and high poverty rates may confront limited resources when it comes to providing basic services to their inhabitants.

• Degradation of the environment: States with big populations and significant levels of economic activity may have difficulties in maintaining the environment.

The following suggestions could help to improve the report:

• Data disaggregation: The data might be disaggregated by race, ethnicity, and other demographic characteristics to provide a more nuanced view of population, income, and poverty distribution across the United States.

• Trend analysis: The study could examine population, income, and poverty trends over time to detect changes that are occurring.

• Comparison to other nations: To provide a global perspective, the study might compare the distribution of population, income, and poverty in the United States to other countries.

• Policy recommendations: The report might conclude with policy recommendations for tackling the difficulties created by unequal population, income, and poverty distribution.