<u>Visualization #1</u>: Analyzing Mean Commute Time by County

Link to Visual:
https://public.tableau.com/app/profile/william.beck.russell/viz/BusinessAnalytics-Project4Visual1/Dashboard1?publish=yes

<u>Insight:</u>

Map:

Creating a data map measuring by 'County' first, allowed a deep look into the data from the very beginning, particularly with the insights it offered into the diverse data outcomes within what might otherwise be judged as homogenous states.

The measures I used for the county map included: total population, income per capita, unemployment rate, and mean commute time (in minutes). These were beneficial measures to start with as they gave me a basis for understanding how the state demographic data might interact in later models (especially with how certain variables might correlate later on).

Bar Chart:

The county bar chart measures the 'Mean Commute Time' by county and was created to get a helpful visual as to how counties within the same state could vary greatly in data. Texas alone has 254 counties, with the lowest mean commute time to work being 9.7 minutes (Kent County), and the highest being more than 30 minutes longer at 40.9 minutes (San Jacinto County). There were actually rather large differences between longest/shortest commute times across many of the states, so this chart allowed for a relatively easy look into each state's county statistics, especially when using a filter to look at states and their counties one at a time.

Scatter Plots:

With the goal with this visualization being to analyze the Mean Commute Time across counties, these scatter plots were created to find out what other variables might be correlated with commute time. Plotting the 2 'Commute type' variables of "% Who Walk" and "% Work At Home" (i.e. the % of the employed population who commute to work by walking and the % of the population who don't require commuting for work, respectively) and using a trend line, the results were 2 negative relationships.

"Walk" had a correlation coefficient of 0.27457, indicating a weak (almost moderate) negative relationship with commute time. "Work At Home" however, had a correlation coefficient of 0.435582, which is a moderate negative relationship with our commute time variable. What this means is that we can expect a lower commute time to work (on average), whenever we observe higher walking commuters, and even more so when we observe higher percentages of employees who work from home (and vice versa). A logical reason for this might be that when more people walk or work from home, there are less cars on the road, and subsequently less traffic slowing down other work commuters.

Histogram:
The final chart made for this section is a histogram with information about certain groupings of commute times, how many counties fall within those groupings, and what geographical location those counties belong to.

The commute time data in this chart is “binned” so that the number of counties observed within 2 minute intervals of commute time are grouped together. For example, if we zoom in to “8” on the x-axis, we can see that there are a total of 12 counties that have average commute times between 8 minutes and 10 minutes. Comparing this to the largest group of average commute times exemplifies just how uncommon a commute time of <10min is when compared to counties across the United States. The largest group of commute times is between 22 minutes and 24 minutes, and there is an unbelievable total of 479 counties that fall into that work commute range. What’s also unbelievable is that can also observe that nearly half of these counties (212) are in the South region of the US, with the Midwest region not too far behind with 169 of their own counties falling in this range too.

This chart allows us to clearly see that most of our data’s average commute times are between 14 to 30 minutes, and that the majority of these counties in this range are in the South and Midwest regions of the United States. From surface looks alone, it appears that the South and Midwest have the most counties of the U.S. regions. Just comparing Texas to California, we see that although California has roughly 12 million more in population (38,421,464 vs TX’s 26,538,614), California is also made up of almost 200 less counties (only 58 compared to 254 in TX). Noticing these differences in how some states choose to group/classify themselves gives much more insight into how our data might behave in other scenarios.

Design:
Most of the chart designs used stay within a ‘blue to orange’ color palette so that data is easier for anyone to distinguish, including any individuals who may be colorblind. Filters were used in almost every chart so that there would only be a smaller amount of data on the dashboard at a time, reducing clutter, and making it easier for observers to find information they’re looking for. The only charts not color-coded are the scatter plots, and this was done in order to emphasize the trend lines as the most important aspects of these plots.

Resources:

- https://www.kaggle.com/datasets/muonneutrino/us-census-demographic-data
- https://help.tableau.com/current/pro/desktop/en-us/trendlines_add.htm
- https://www2.census.gov/geo/pdfs/maps-data/maps/reference/us_regdiv.pdf
- https://www.youtube.com/watch?v=VwDPBWuHu3Q

Visualization #2: Analyzing the Determinants of Income at the State Level

Link to Visual:
https://public.tableau.com/app/profile/william.beck.russell/viz/BusinessAnalytics-Project4Visual2_16545104753350/Dashboard1?publish=yes

Insight:
Map:
In this data map, variables were measured on the 'State' level, and those variables include: average income, average % of worker population in construction, avg. % in office work, avg. % in production work, avg. % in service work, and the avg. % in professional work. The reason for these variables is that I wanted to analyze the types of work available/popular/beneficial in the United States, and the level of effect each of them may have on the average income variable. The map's color gradient reflects the average income in each state, with darker orange meaning a lower avg. income, and blue being indicating a higher one.

Scatter Plots:
Several scatter plots were created in order to see if any variables correlated to Average Income Per Capita to a significant degree. Overall, 5 plots were created, one for each type of work presented in the data source. Those 5 types of work were:

1.) professional (management, business, science, arts)
2.) production (production, transportation, material movement)
3.) construction (natural resources, construction, maintenance)
4.) office (sales & office jobs)
5.) service (service work)

Each one of these variables was measured against average income per capita, and then a trend line was plotted in order to find the correlation coefficient. It should be noted that since Puerto Rico was an obvious outlier in this dataset, it was excluded from these calculations. The only strong relationship was found with the professional variable, which had a positive correlation coefficient of 0.795, indicating that we can expect the average income per capita to be higher whenever a population has a higher percentage of population doing 'professional work' (on average). The other coefficients were as follows:

Production: |-0.447| -> negative relationship, moderate correlation
Construction: |-0.356| ->negative relationship, moderate correlation
Office: |-0.048| ->negative relationship, weak correlation
Service: |-0.029| ->negative relationship, weak correlation

In other words, we could also expect that whenever workforce populations have a higher percent of 'production' or 'construction' type work, average income per capita can be expected

to be lower (on average), however this change wouldn't be as large as a change from the 'professional' work type. Although not all of these variables have a significant relationship with average income per capita, plotting all of these charts helped us figure out that there are some that do have a relationship. After this, I concluded that % of employed population doing 'professional' work is one of the bigger factors that impact Income, and I will subsequently focus on this variable more in the data as a result.

Comparative Bar Charts:

These bar charts were plotted mainly to demonstrate in a more comprehensive way that a relationship exists between the Income and 'professional work' variable. The left chart measures the 'professional work population %' for each state, while the right side is lined up so that each state's average income is shown as well. It is arranged so that the states with the highest 'professional %' are at the top and descend in order accordingly. As you look down the graph, one can observe that, on average, as the 'professional %' of each state decreases, a similar decrease is taking place in the average income as well. Knowing this information could provide both individuals and groups important insights into the types of work to focus on, if going after that higher average income is an important factor to them.

Sub-Visualization Link:

https://public.tableau.com/app/profile/william.beck.russell/viz/BusinessAnalytics-Project4Sub-Visuals/Dashboard2?publish=yes

The purpose of these extra scatter plots is to show that focusing on the % of professional work might be an important factor to tackle other variables it has a relationship with. In this case it was measured against both poverty and childhood poverty, and, interestingly enough, when Puerto Rico's data was not excluded, the resulting correlation coefficients were weak (poverty: |-0.116| & child poverty: |-0.130|). However, after excluding Puerto Rico, the correlation coefficients both indicated moderate, negative relationships with 'professional work %' (poverty: |-0.477| & child poverty: |-0.500|), indicating that excluding Puerto Rico's data for fear of it being an outlier was probably the correct thing to do. The updated correlation indicates a a moderate decrease in poverty can be expected, on average, as 'professional work %' becomes higher.

Design:

For the map design, I kept the orange-blue divergent color scale about the same, as it effectively grabs attention towards data to highlight, along with being a color-scale still distinguishable to those with potential color-blindness, or anyone else. Because I did not want to overwhelm observers with color for each state, I went ahead and kept the bar charts blue. I felt the length of each of the bars portrays the information well, especially considering that both bar charts are lined up. As for the scatter plots, I began with shapes representing the different regions, but all of the data points just jumbled together even worse, so I kept all the points blue and homogenous so that the main focus would be the trend line outputs. In the next visualization, I will represent a scatterplot by region in a more graspable way than what would be possible here.

Resources:

- https://www.kaggle.com/datasets/muonneutrino/us-census-demographic-data
- https://help.tableau.com/current/pro/desktop/en-us/trendlines_add.htm
- https://www2.census.gov/geo/pdfs/maps-data/maps/reference/us_regdiv.pdf
- https://www.youtube.com/watch?v=bDdNQRatR5w&t=207s

Visualization #3: Analyzing Unemployment, Average Income, and the Working Professional % by Regional Data

Link to Visual:
https://public.tableau.com/app/profile/william.beck.russell/viz/BusinessAnalytics-Project4Visual3/Dashboard1?publish=yes

Insight:

Map:
For this data map, variables were represented to the "regional" level, where each region was made up of a number of states. In order to determine the region groups, I used the information for how the U.S. census defines the United States regions, like so:

- *The West:* 13 territories - AZ, CA, WA, OR, CO, NM, UT, NV, ID, WY, MT, AK, HI
- *The Midwest:* 12 territories - ND, SD, NE, KS, MO, IA, MN, WI, IL, IN, MI, OH
- *The South:* 17 territories - District of Columbia, TX, OK, LA, AR, MS, AL, TN, NC, SC, VA, WV, MD, DE, FL, GA, KY
- *The Northeast*: 9 territories - NY, ME, VT, RI, PA, NH, MA, CT, NJ
- *Other:* Puerto Rico

Along with grouping the states as regions, it was color-coded for the variable average income. Average income, the region name, and the average unemployment % are displayed, by label, on the map. The map is also keeping track of the total population, which can be seen from hovering over any region and reading the ToolTip.

My reasoning for again using average income for the map color-gradient, is to test whether similar conclusions can be made about the Income variable when aggregating/grouping such diverse states and populations into large regional groups.

,

Side-by-Side Bar Charts:
In my last visualization, I concluded that Income's relationship with the 'production work %' and 'construction work %' were moderately negative relationships, while 'professional work %' was a strong positive relationship. In order to begin checking these assumptions on regional data, I constructed 4 bar charts to represent each region grouping, and then included the average % of all the work type variables for each region. They were set up in this way so as to more easily measure each region's worker type percentages against others, side-by-side.

Just by observing the chart, we can see that the highest Professional Work % belongs to the Northeast region at 35.17%, while the lowest belongs to the South region, which has 29.53%. Just by going off our previous assumptions, we might expect that the Northeast region has the highest average income, while the South has the lowest average income. Not only is this true, but the rankings of each region by 'Professional Work %' matches up exactly to the average income ranking of each region:

1.) *Northeast:* 35.17% professional work —---------- $56,385 avg. household income
2.) *West:* 32.80% professional work —---------- $50,215 avg. household income
3.) *Midwest:* 31.48% professional work —---------- $48,585 avg. household income
4.) *South:* 29.53% professional work —---------- $43,002 avg. household income

So it would appear that our assumptions derived from state data were able to hold firm for these regional groups. But, those assumptions were derived from trends containing ALL United States data points. What would happen if we attempted to calculate trend lines for individual regions?

Professional Work % Scatter Plot (by region):

Upon first glance, all the trend lines (and variable relationships) seem positive, which is what was expected. However, when inspecting each of the separate trend lines by region, it appears as though our assumptions don't hold true for all regions. Ranked, in order of greatest correlation to least, the coefficients for our trend lines are:

Original Income per cap. VS Professional %Correlation Calculation

1.) ***The South:*** **|0.970| —-> strong, positive relationship**
2.) *The Northeast:* |0.869| —-> strong, positive relationship
3.) *The Midwest:* |0.366| —--> moderate, positive relationship
4.) ***The West:*** **|0.051| —-> very weak, positive relationship**

Seeing such a wide range of correlation coefficients that are only different based on region alone is very interesting and suggests either an issue with the grouping together of such diverse demographics or another/other underlying variable(s) that could have unknown relationships with certain variables in our data pool, especially with the West region.

Sub-Visualization Link:
https://public.tableau.com/app/profile/william.beck.russell/viz/Project4Visual3Sub-Viz/Dashboard1?publish=yes

Production + Construction % Scatter Plot (by region):

This scatter plot was made in order to challenge the other assumptions made regarding the variables that Income appeared to have moderate and negative relationships with: Production Work % and Construction Work %. This plot is also analyzing trends by region instead of trying to trend the entire dataset at once. Ranked, from greatest correlation coefficient to lowest, these are the results of the new trend lines calculated:

Original Income per cap. VS Production & Construction % Calculation

1.) ***The South:*** **|-0.924| —-> strong, negative relationship**
2.) *The Northeast:* |-0.878| —-> strong, negative relationship
3.) *The Midwest:* |-0.068| —--> very weak, negative relationship
4.) ***The West:*** **|-0.013| —-> very weak, negative relationship**

At this point in the calculation, I realized the problem in my correlation calculations was that I needed to add more granularity to my plots. Whereas in these previous 2 plots, I was calculating trends based on single data points for each state, representing the average of all of a state's counties. As I spoke on earlier, even counties within a state can be vastly different, so after making 2 new plots and calculating their trend lines, here are the more granular results:

New Income per capita VS Professional %Correlation Calculation (all county points)

1.) ***The Northeast:*** **|0.677| —-> moderately strong, positive relationship**
2.) *The South:* |0.565| —-> moderate, positive relationship
3.) *Puerto Rico:* |0.366| ---> moderate, positive relationship
4.) *The Midwest:* |0.301| —-> moderately weak, positive relationship
5.) ***The West:*** **|0.272| —-> moderately weak, positive relationship**

New Income per cap. VS Production & Construction % Calculation (all county points)

1.) ***The Northeast:*** **|-0.485| —--> moderate, negative relationship**
2.) *Puerto Rico:* |-0.471| —--> moderate, negative relationship
3.) *The South:* |-0.326| —--> moderate, negative relationship
4.) *The Midwest:* |-0.120| —--> weak, negative relationship
5.) ***The West:*** **|-0.107| —--> weak, negative relationship**

Now that the plots have more granularity, the data results make much more sense. The addition of so many more data points even allowed the plot to include proper calculations for Puerto Rico. Even though the West and the Midwest region don't have as large of a correlation between Income and Professional Work % as I had assumed in previous visualizations, a relationship does appear to exist, even if it's fairly weak. Although it's weak, we can still expect at least a little change on average when one of our variables increases/decreases.

As for the relationships between Income & both Production % and Construction %, it's interesting to note that the West and Midwest regions Income variable still appears to have a very weak relationship to construction % and production %. However, these variables are based on percentage of populations (i.e. if one variable increases, another must decrease b/c you can't exceed 100% of a population), so their weak relationships to 'professional %' logically explains the lack of relationship with these other related variables.

Design: For the map coloration, I stuck to the similar blue-orange color scale to maintain consistency between visuals, and, as always, that gradient maintains its ability to effectively point out important information without being displeasing to the eye. For the bar charts, color was a necessary tool this time to help anyone looking at the data make easy connections from bar graph to bar graph. As for the scatter plots, the addition of color made it easier to select from the multitude of trend lines to find the information I was looking for. While the updated scatter plots include every county as a data point, those points themselves don't lend well to color, but having the legend made it simple to isolate data being searched for. Overall the design of my visual avoided unnecessary clutter and the use of color was appropriate for the tasks at hand.

Resources:

- **https://www.kaggle.com/datasets/muonneutrino/us-census-demographic-data**
- **https://help.tableau.com/current/pro/desktop/en-us/trendlines_add.htm**
- **https://www2.census.gov/geo/pdfs/maps-data/maps/reference/us_regdiv.pdf**