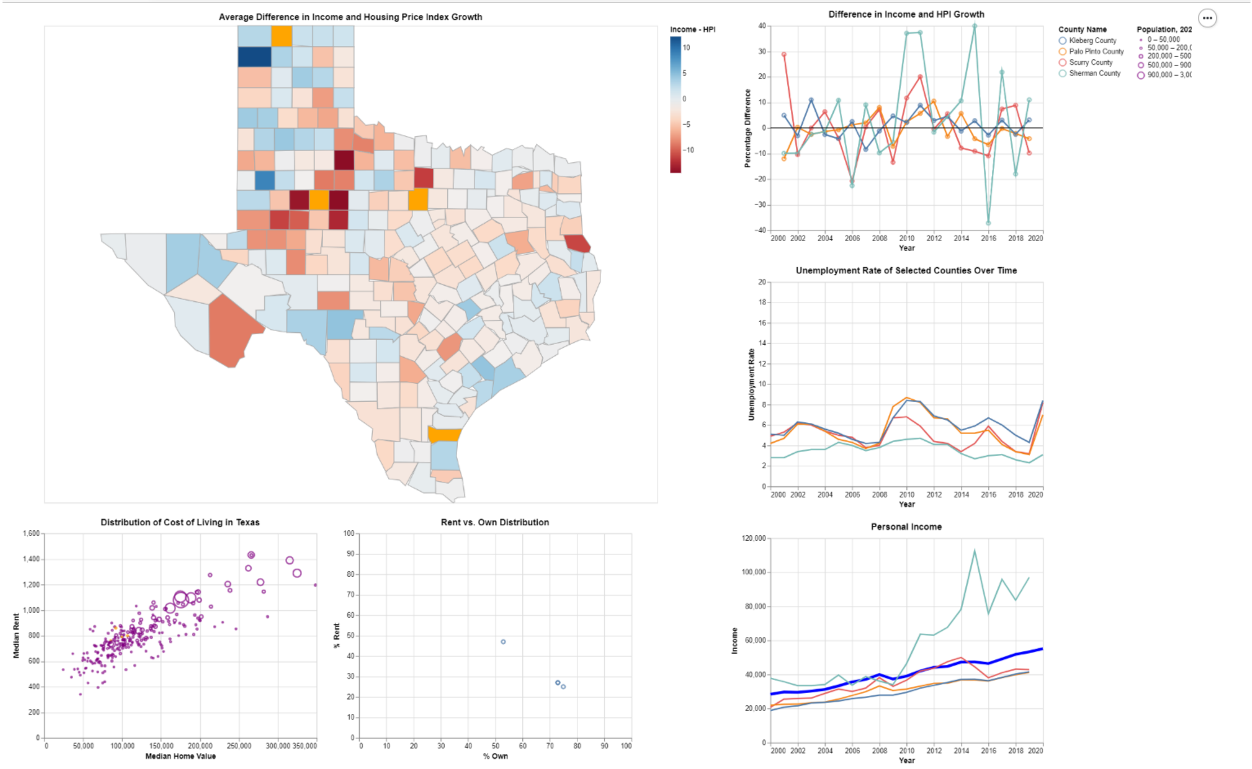
The Three Johns- Texas Counties Final Report



Concise Summary:

The user is any college graduate who is looking to move to a Texas county to start their career and settle for the long term. The goal of the visualization is to pick a county to live in based on comparisons of different metrics, such as the difference between personal income growth and housing price index, unemployment, cost of living, etc. To successfully visualize these comparisons, we use an interactive texas map where the user geographically views the counties as well as compare by color the difference in income and house price growth. Then, they can click multiple counties on the map to activate the other components of the visualization. The unemployment, personal income/HPI, and personal income line charts show the unemployment rate, personal income - HPI, and personal income growth time trend from 2000 to 2020. These are used for the user to economically distinguish the counties. Then, the cost of living scatterplot shows the distribution of cost of living in order for the user to study housing statistics of the counties to find their potential home. Through this project, we learned that data cleaning and transforming is extremely tedious. This led to us not being able to do everything we thought we could do in the proposal such as web hosting the visualization. Also, mean and standard deviation were not as readily used as we thought they would be because either they were already implemented in the data or they were not necessary.

The Three Johns- Texas Counties Final Report

Data Description

The first dataset that were used in the project is a geojson file which is obtained from https://github.com/glynnbird/usstatesgeojson/blob/master/texas.geojson. The data can be confirmed to be reliable when compared to a Texas map. The polygon data present is used to draw the Texas counties map. It also contains a “namelsad” data that represents the county name of the current polygon. We renamed “namelsad” to “County Name” to make the title more understandable in the tooltip. Eventually, “County Name” is the data used to build the county selection.

The second dataset that were used were constructed from two sources: https://fred.stlouisfed.org/release/tables?rid=175&eid=268680&od=2018-01-01# and https://fred.stlouisfed.org/tags/series?t=fhfa%3Btx&ob=pv&od=desc. These two sources provide us with the average income and housing price index in each Texas counties in 2000 – 2019. The data is provided by official government agencies, U.S. Federal Housing Finance Agency and U.S. Bureau of Economic Analysis, which should be reliable.

The data can be used to calculate the growth in each Texas counties. The two xlsx file downloaded will be combined into one file with “County Name”, and the years separated to their own columns. The resulting will be a wide-form dataset. To make it compatible with Altair, we converted the data into a long-form data with the columns “County Name”, “Time”, “Personal Income”, and “HPI”. This conversion is done by looping through the entire dataset and moving the data to a new dataframe. The resulting dataframe has around 7000 rows of data which is why we increased Altair’s 5000 rows limit.

Next, we added two new columns to represent the growth of Income and HPI each year. This transformation is done so the user can compare the growth instead of the raw number. Comparing the raw number would not make sense because the Housing Price data is indexed to their own specific counties, i.e., an HPI of 100 in one county will mean a different thing to an HPI of 100 in another county. So, it will be impossible to compare it to other counties as is. The transformation to growth per year is done using the shift(1) method. We would loop through each counties separately and subtract the current county’s series by the shifted series using shift(1) and then, dividing it by the shift(1) series. Then, we convert the result to percentage by multiplying it by 100. This is done on both the Personal Income and HPI data. The two resulting data will be combined by subtracting the HPI growth from the Personal Income growth. This is done for easier comparison throughout the counties, instead of 2 data per county each county will only have 1 data.

We added an “Unemployment” column in the resulting dataframe. Then unemployment data will be from the U.S. Bureau of Labor Statistics. The unemployment data is generated from their website (data tools) which should be reliable. Finally, we reformatted the “County Name” column to make it the same format as in the geojson data which will allow the selection interaction to work. We also added gid data to each county which reflects the gid in the geojson file which will be used as the key to combine the geojson and the dataframe using altair’s transform_lookup(). The resulting transformed data will be saved to a new xlsx file, “Map+Unployment”

The third dataset we used is a table consisting of the Cost of Living data of each county. The data is obtained from https://www.niche.com/places-to-live/search/counties-with-the-lowest-cost-of-living/s/texas/, sourced from the U.S. Census and BLS which are reliable sources. To make the data compatible with the previous dataset, we simply changed the “county” column to “County Name” which allows it to be used with our Altair selection.

Goals and Tasks

The general goal of this visualization will be to allow the user to compare different metrics, specifically across different counties in Texas. The user might have a specific set of counties that they are considering or they might be open to any counties in Texas.

When browsing through the visualization, the first task of the user, assuming they have no preference of any of the counties they are moving to, would be to compare the counties through the Texas map. The Texas map displays the difference in average growth of Personal Income and Housing Price Index. We used darker blue color to represent counties with higher Personal Income growth and darker red color to represent counties with higher Housing Price Index growth. Since the user is concerned with housing, they would search the map for the darker blue color and click those counties to display their metrics in the other visualizations. If they have a preference for any counties, they can pick those counties without considering the color of the Texas Map. In this example, the user selects Collin and El Paso since they have a relatively high Personal Income - HPI Growth compared to the other more urban counties. They also added Travis county because they are interested in it even though it has a worse growth difference.

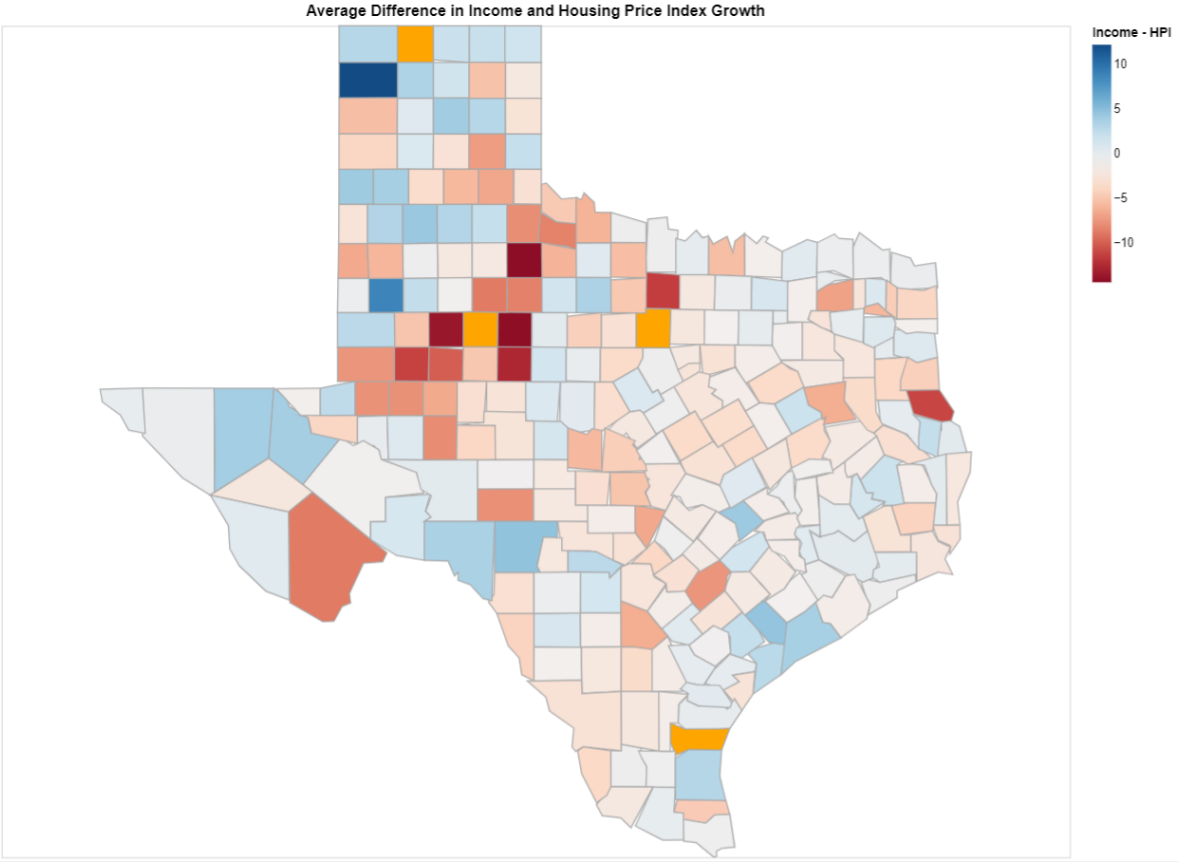
The second task of the user would be to compare the other statistics provided. Ideally, the user would find a county with a historically high Personal Income and low unemployment which can be compared using the line graphs. The user would also prefer high Income - HPI Growth in recent years, instead of high Income - HPI historically which can be seen on the line graph. They would also prefer a lower median house price and rent which can be seen in the cost of the living scatterplot.

In the example, the first thing the user notices, in this case, is the lower average personal income in El Paso compared to the other two counties which can be seen in the Personal Income line graph. The personal income in El Paso also seems to be lower than the Texas average which the user does not prefer. But this is somewhat balanced out by the fact that the median home value and rent are also a lot lower in El Paso compared to the other two counties which can be seen in the Cost of Living scatterplot. The user also sees that El Paso and Collin have higher own % which the user prefers. But the unemployment rate in El Paso also seems to be higher historically which might suggest that the user will have more difficulty finding a job, but the user decided that the unemployment rate in recent years does not differ by much and El Paso almost caught up to the other counties. Finally, the user sees that the Income - HPI growth in Travis and Collin is significantly worse compared to El Paso and even going to the negatives. In El Paso, the growth has been positive or equal which is preferable. In the end, the user decided to pick El Paso because of their fear of Travis and Collin county’s housing price growth compared to their personal income growth as it reached almost -7 % as recent as 2016.

The final dataset is a csv table of the average Texas Personal Income data from the U.S. Bureau of Labor Statistics. We also renamed and reformatted one of the columns of this file to make it comparable with the selection.

Visualization:

1. Interactive Texas Map



Description: This interactive map of Texas counties communicates by color the difference between each county’s personal income index and housing price index, which helps the user figure out which county he/she wants to live in based on overall income versus housing expenses. Then, the user can click specific counties on the map to discover even more information about them.

Encoding: Using geopandas, Texas is geographically split into its counties where color encodes the difference between the overall personal income and the housing prices of each county. This allows the user to see geographically where they might want to live while learning about the county's growth over time.

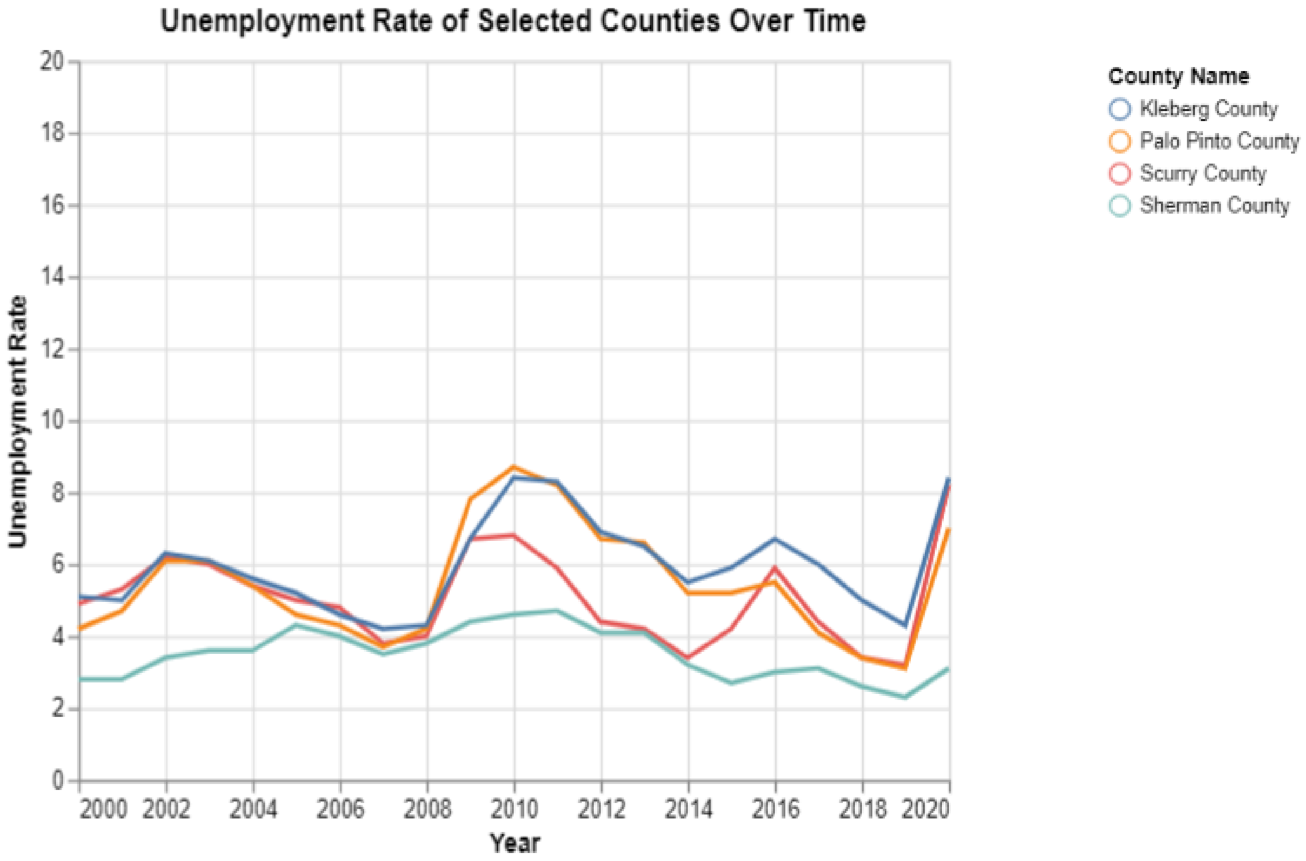
Interactions:

- By hovering over a specific county, a tooltip pops up, giving the county name, the population of the county in 2020, and the exact difference between personal income and housing price index for 2020.
- Using multi-selection, the user can shift+click on multiple counties to find out more information about them.

Links to Other Views:

- After clicking on one or more counties, a line is added to the unemployment, Income/Housing Price Index, and personal income line graphs, one for each county.
- After clicking on one or more counties, a point is added to the rent vs. own scatterplot and the point for that county is highlighted orange on the cost of living scatterplot

2. Unemployment Line Chart



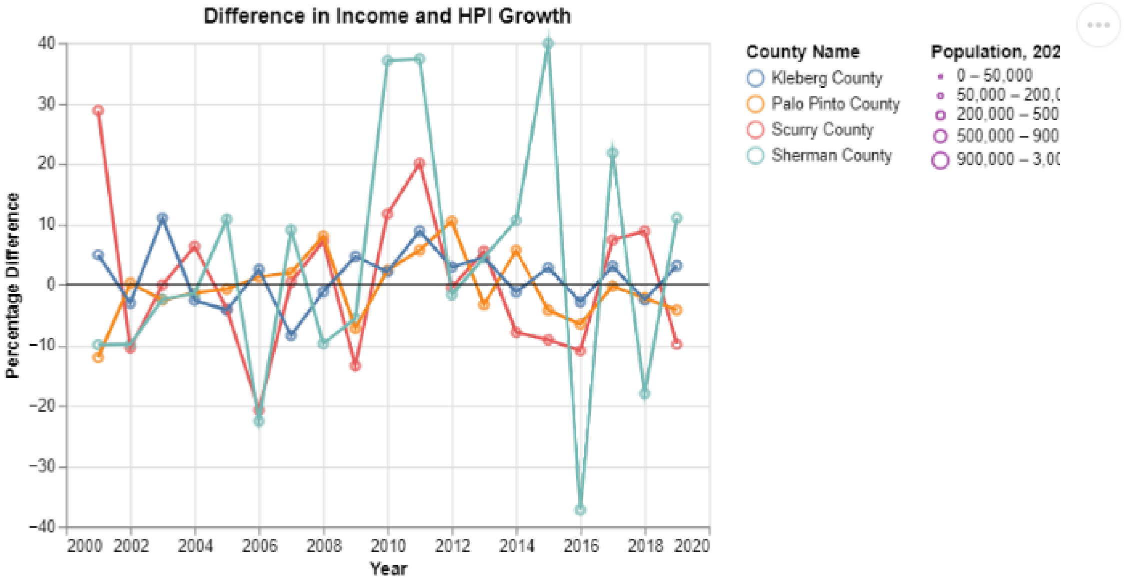
Description: Once a county is selected on the Texas map, a line for that county is added to the unemployment line graph, which shows the unemployment rate trend from 2000 to 2020 in that county the user is interested in learning more about.

Encoding: A line graph for each selected county shows the trend of unemployment in that county over time while color distinguishes the counties. This allows the user to compare counties based on employment, as this user is looking for work and job security in the county they wish to live in.

Interactions: None

Links to Other Views: None

3. Income and Housing Price Growth Line Graph



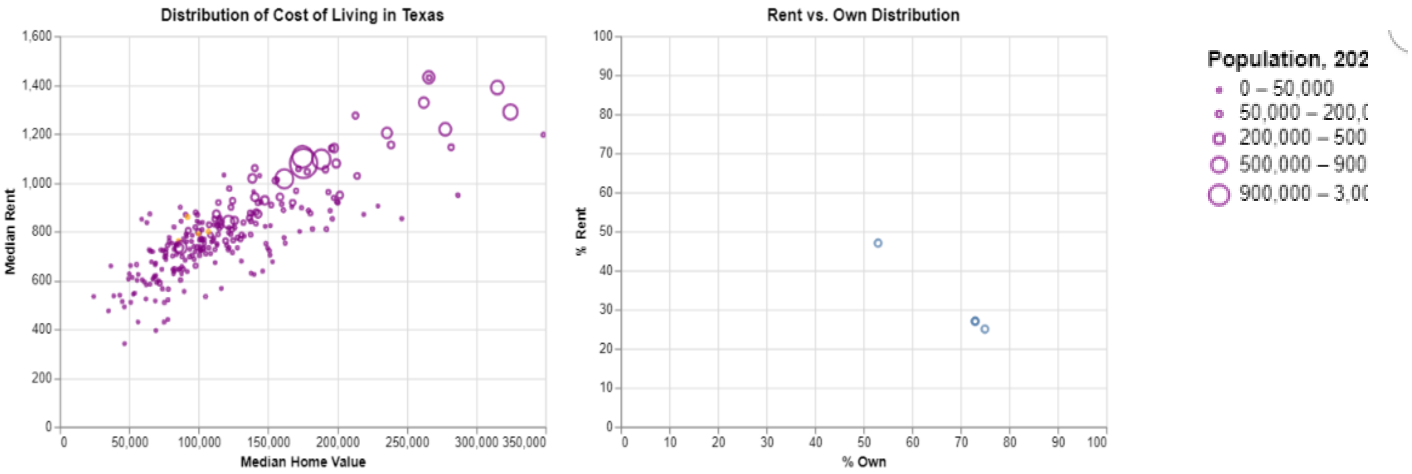
Description: Once one or more counties are selected on the map, a line is added to this graph to show the trend of income growth vs. housing price growth from 2000 to 2020. The texas map just shows the difference for 2020 so this graph aims to see trends over time as well as compare the growth for different counties.

Encoding: Similar to the unemployment line graph, a line for each county selected appears, showing the trend over time while coding color to distinguish between the counties. Similar to the Texas map, this allows the user to now see the time trend of income growth vs. housing price growth to plan for the future instead of just looking at 2020.

Interactions: By hovering over a specific county, a tooltip pops up, giving the county name and the exact difference between personal income and housing price index for specific years.

Links to Other Views: None

4. Cost of Living Scatterplots



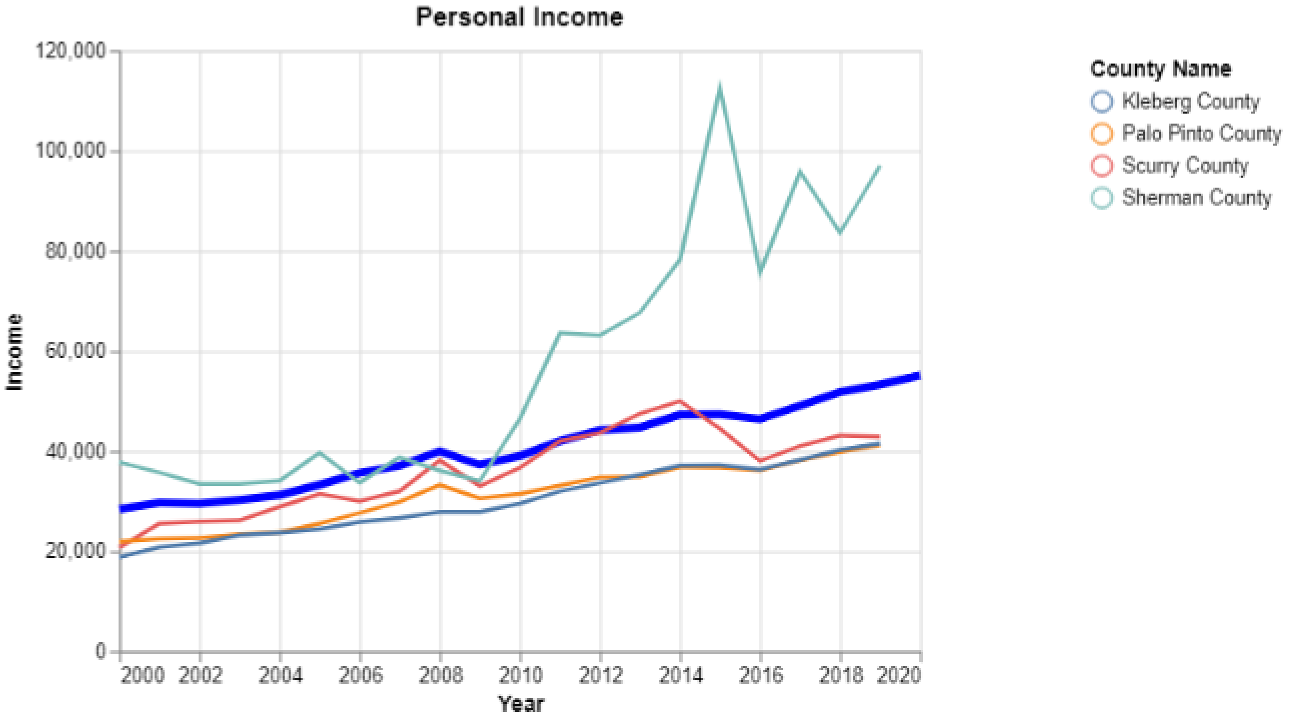
Description: This scatterplot shows the distribution of counties by their overall cost of living in 2021 as well as graphing the percentage of residents in each county that rent vs. own houses to live.

Encodings: Each point on the graph represents a single county and the size of the county (by population in 2021) is represented by a larger circle on the scatterplot. The circles are colored purple and when a county is selected on the map that corresponding circle on the scatterplot turns orange. This allows the user to compare cost of living prices between counties (based on median rent and home values) and see if they are more likely to rent their first home or own it.

Interactions: By hovering over a specific county, a tooltip pops up, giving the county name, the cost of living ranking, the median home value, and the median rent value. You can also zoom in and out on the scatterplots to see a specific county in comparison to the other counties or better see the overall distribution.

Links to Other Views: None

5. Personal Income Growth Line Graph



Description: This line graph shows the growth of personal income specifically over time for selected counties via the texas map. A line for the state average is also included (in dark blue) to compare the specific counties. Overall, you can compare income growth with other counties and the state average to investigate job success.

Encodings: Similar to the unemployment line graph, a line for each county selected appears, showing the trend over time while coding color to distinguish between the counties. This allows the user to look at income growth directly and compare it between counties and the state average to find a job where they are more likely to gain income over time and make a life for themselves.

Interactions: None

Links to Other Views: None

The Three Johns- Texas Counties Final Report

Reflection:

We have changed our target user base from people wanting to move to Texas to college graduates looking for a place to live in Texas. Although people looking to move could be interested in this, college graduates would be most interested in this visualization because they typically have debt, and are looking for a start to their career. As the project progressed from the alpha implementation, we have noticed the rigor of data cleaning, especially from government sources. Sometimes column labels would take multiple cells and would mess up conversion into a dataframe. We would have to clean the data inside Jupyter Notebooks and Excel. Some of our data upon further analysis was deemed useless for our purpose, so we had to find new data, which was hard to find. We have dropped some sub-goals of the project, such as web-hosting and statistics. Web-hosting is useless in this scenario because a simple html file would fulfill our purpose of a demo. Secondly, statistics such as standard deviation and mean values are already given in the graph, or are useless. We did not think giving a standard deviation would be useful here since it could be not that useful for our users. Mean values of incomes are already graphed out, so putting it in a tooltip would be useless. Our original proposal was realistic - we already had experience making projects in past classes, so this was easy. All of our ideas were easily implemented using existing python libraries.

Team Assessment:

a. Jonathan made the county map using external libraries. He managed to make it work in altair using geopandas.
b. John did the Personal Income graph. He cleaned up and managed the data for TXCPI.csv, cost_of_living1.xlsx, and Map+Unemployment.xlsx.
c. Evan did the other graphs. He cleaned up data for cost_of_living1.xlsx. He made the Texas county map interactive with the other graphs.