Urban Growth Visualization Process Book Project

By Tony Zhang and Cody Matthews

DS-4630

Table of Contents

Initial Project Proposal:

Background and Motivation

We chose this topic for a number of reasons:

1. Real estate investing:
    i. We are curious to see which places have grown in population/housing price the most/least over time and to then be able to piece together reasons for why this is and what to expect moving forward.
2. Home purchases and budgeting:
    i. The decision to either buy a home or rent is always a topic worthy of discussion. Seeing what housing prices are for various places over time can then be used to compare rental rates to be able to decide which path is best for each person.
3. Population growth and job opportunities:
    i. With urban growth comes opportunities for companies to relocate to fit the demand in each region or vice-a-versa. This allows for decision making of where to move that has a job that best fits your skill set. We also would like to see GDP by state to see if there are any key factors for movements.

Project Objectives

What would we like to learn and accomplish?

1. How Urban Growth has affected housing prices over time in different locations.
2. To understand which regions have experienced the fastest growth and which regions have seen the greatest increase in housing costs.
3. To better understand how market rates change based on population size.

Data

We plan on gathering the following data:

1. Urban growth data:
    i. This will come from population and urban density from the US census bureau, and other open-source data platforms. We also plan on gathering information related to the GDP of each state to better understand population shifts and the effects from them.
        1. https://www.census.gov/

Click for Table of Contents

            2. https://www.bea.gov/data/gdp/gdp-state

2. Housing price data:
    i. This will come from historical selling prices and current market rates through places such as Zillow, realtor.com, and other government related tax assessment records.
        1. https://www.zillow.com/
        2. https://www.realtor.com/
        3. https://www.bls.gov/eag/home.html
        4. https://www.redfin.com/news/data-center/
3. Employment Data:
    a.

Data Processing

We expect there to be data cleanup in order to correctly match up data that is gathered from different sources.

We plan to derive urban growth quantities and housing prices for the timeline for the regions we choose to highlight.

How will data processing be implemented?

Data will be scraped/read in using a python colab notebook. Inside the notebook is where all cleaning and formatting of the data will be done.

We will also be gathering data from the websites mentioned above as the majority of them have .csv files available for download that can easily be formatted to match each other.

Visualization Design

Develop three alternative prototype designs for your visualization

1. Have a main page with the US states and then sub plots off to the side that can be manipulated based on filters.
2. Have a main page with the US states that can be selected and then have plots off to the side reflect data from the highlighted states.
3. Have a heat map of the US states that reflects growth rate, population, etc that can be toggled.

And one final design that incorporates the best of your three designs

- A geographical map where different regions are color coded based on housing prices/population density. With multiple plots on the side.

Click for Table of Contents

Describe your designs and justify your choices of visual encodings.

Bubble chart for each region that represents the population size and its x/y orientation represents housing prices vs growth rate of population.

A heatmap of each region where darker colors represent higher prices/growth/density with an option to toggle between layers. Include a legend.

An urban growth map that toggles before/after views of housing prices/population density

Line graph (one for population growth and another for housing price trends

Must-Have Features

A map-based visual that shows each location's population and housing costs.

Line or bar graph that highlights each region's population and housing costs over time.

Optional Features

A map-based visual that shows a prediction of what each region will look like in 5 years, 10 years.

A comparative analysis that allows the user to compare two regions side by side.

When small regions are selected the counties/cities within the state(s) will be colored individually.

Project Schedule

- Week 3 (Sep 2-6)
    - Rough draft of project proposal
    - Create group GitHub repository
- Week 4 (Sep 9-13) Project Proposal Due
    - Complete and turn in Project Proposal
    - Schedule a time to meet with staff member
- Week 5 (Sep 16-20) Project Review w/ Staff (as a group!)

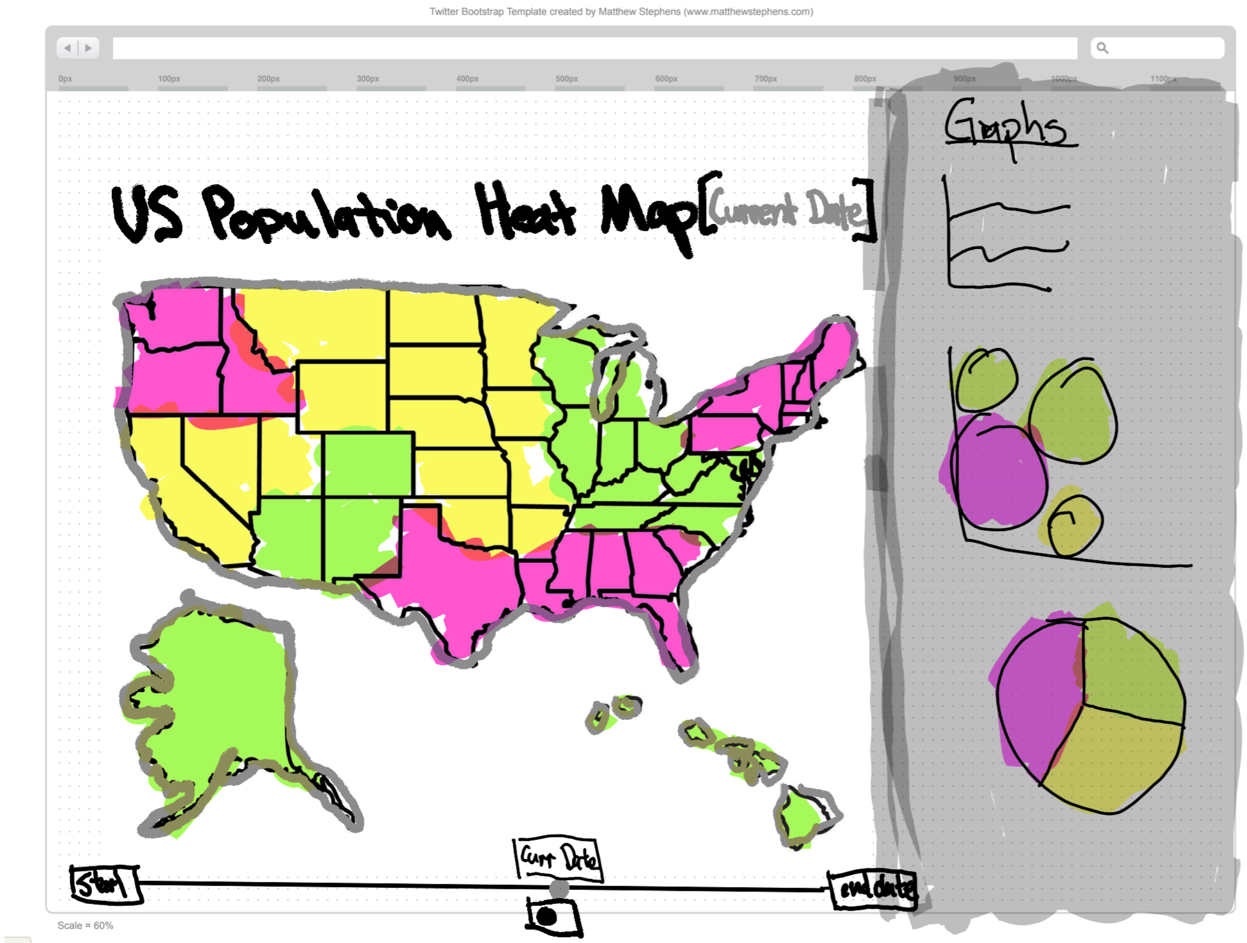
Click for Table of Contents

    - Meet with TA for project review

- Week 6 (Sep 23-27) Project Review w/ Staff (as a group!)
    - Gather data needed to implement project
- Week 7 (Sep 30-Oct 4) Midterm Exam
- Week 8 (Oct 7-11) Fall Break
- Week 9 (Oct 14-18)
    - Finish compiling data for project milestone
- Week 10 (Oct 21-25) Project Milestone Due
    - Submit project milestone for review
    - Start coding website
- Week 11 (Oct 28-Nov 1)
    - Code website
- Week 12 (Nov 4-8)
    - Code website/fix bugs
- Week 13 (Nov 11-15)
    - Finish coding website/fix bugs
- Week 14 (Nov 18-22) Project Screencast Submission Due
    - Record and submit Project Screencast
- Week 15 (Nov 25-29)
- Week 16 (Dec 2-6) Final Project Submission / Group Member Feedback Due
- Week 17 (Dec 9-13) Final Exam

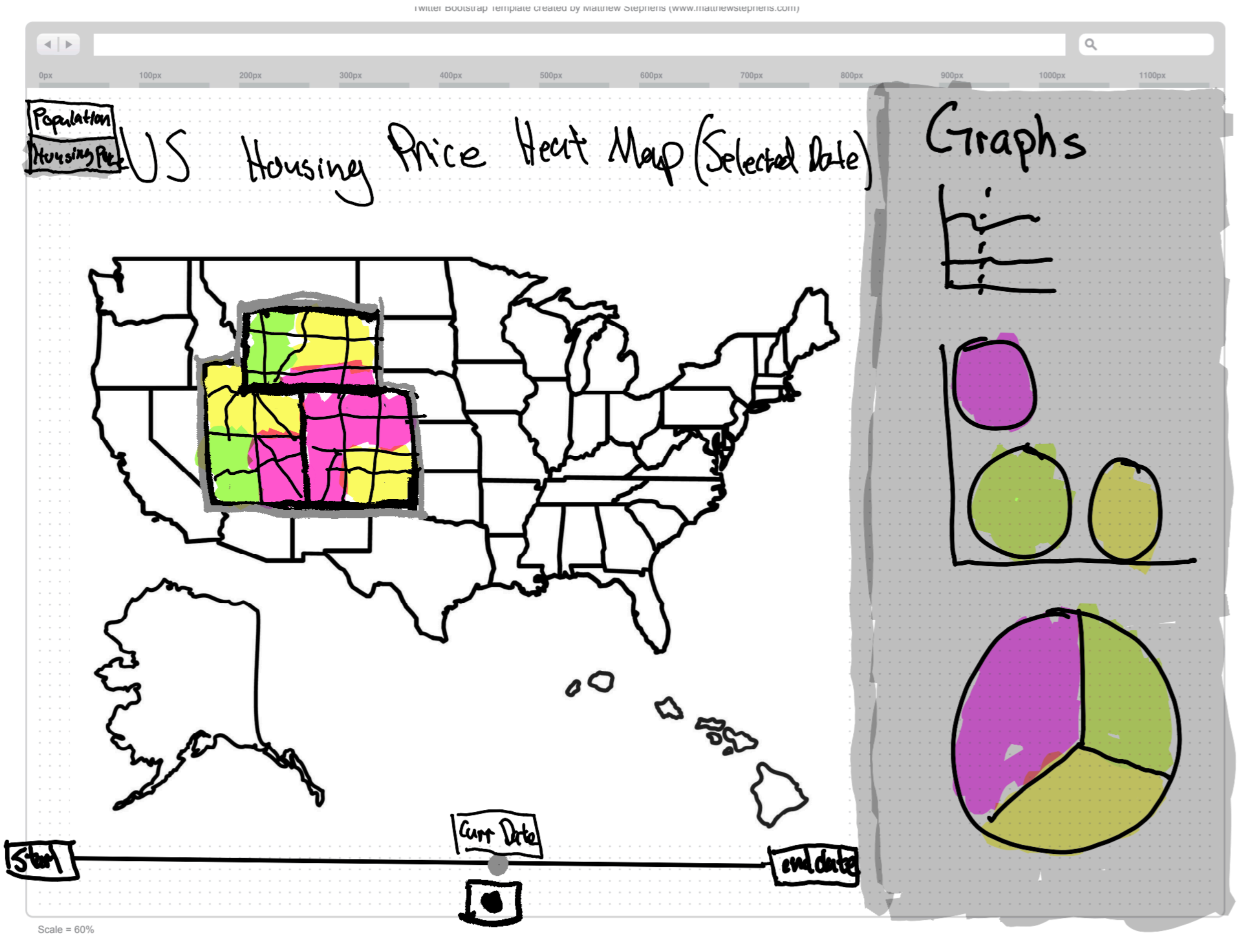
Click for Table of Contents

Opening View

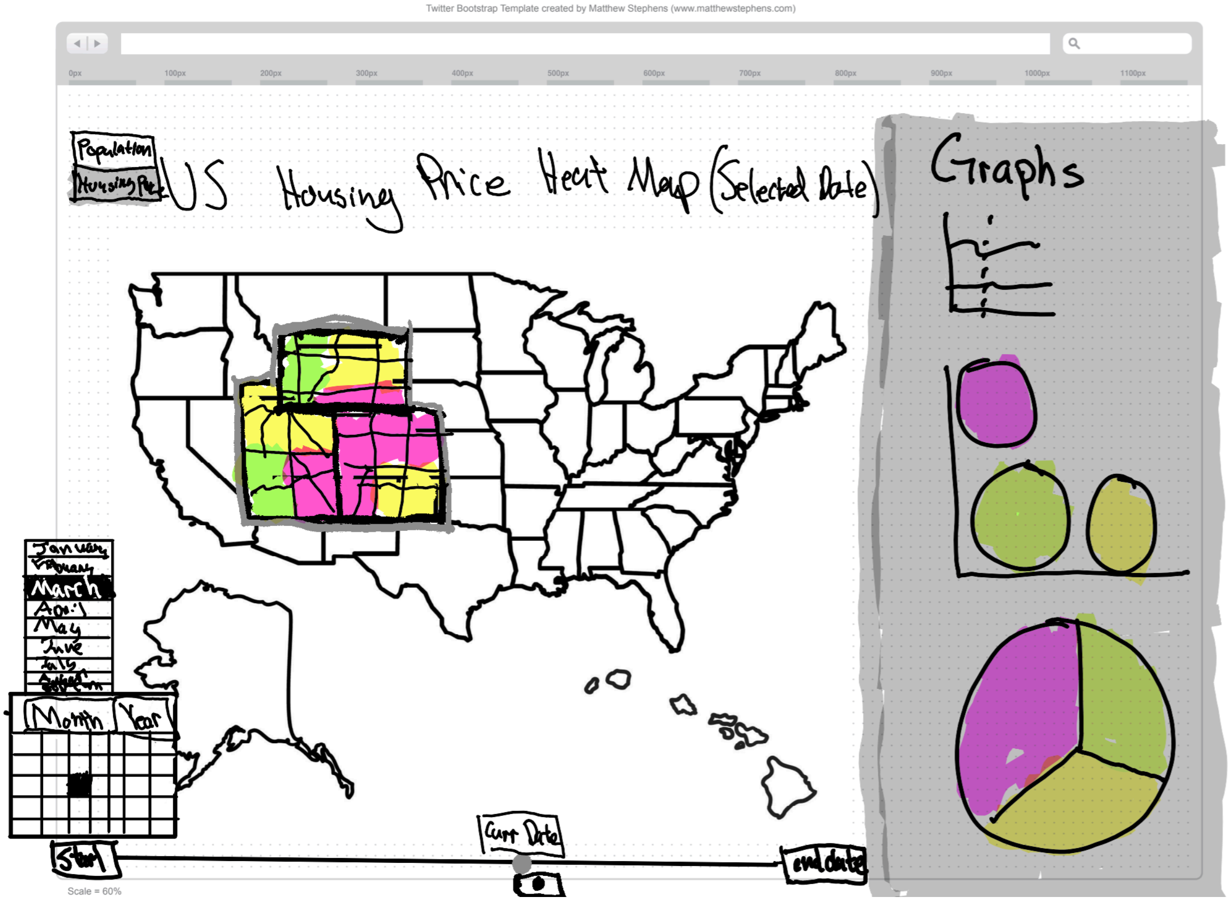

Twitter Bootstrap Template created by Matthew Stephens (www.matthewstephens.com)
0px
100px
200px
300px
400px
500px
600px
700px
800px
900px
1000px
1100px
US Population Heat Map[Current Date]
Graphs
Start
Curr Date
end date
Scale = 60%

Click for Table of Contents

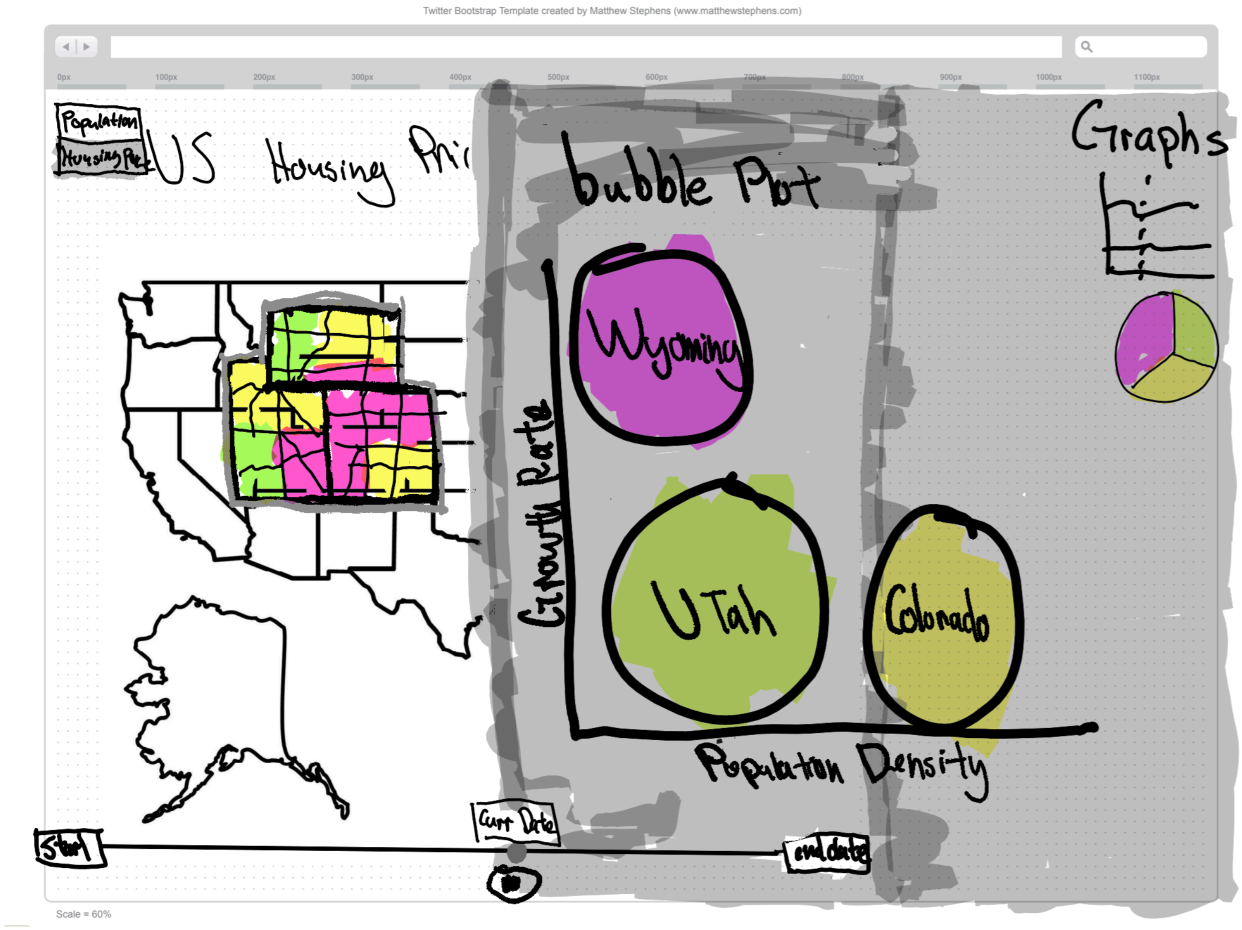
Region Selection



Choose starting dates for animation

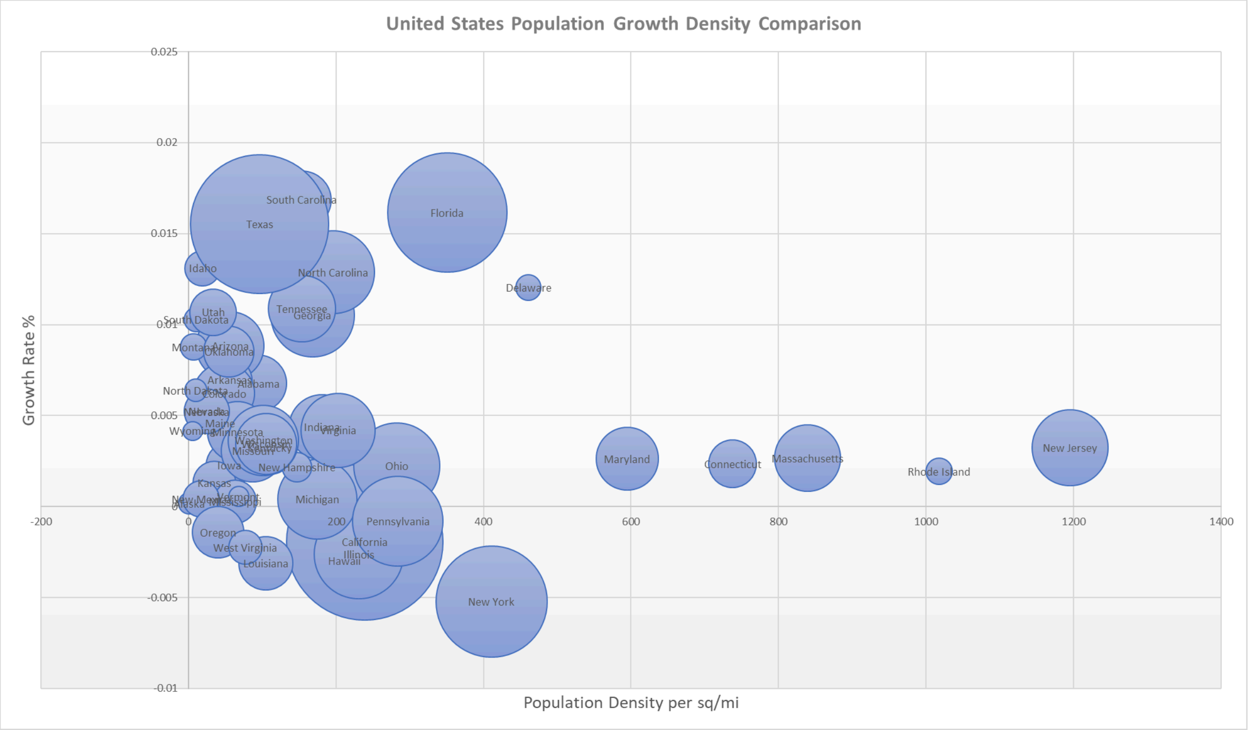


Enlarging a graph



Click for Table of Contents

Bubble plot example of the states with population vs growth rate vs density



[Click for Table of Contents]

Feedback on Initial Meeting with TA Rifat Ara Proma

[Entry by Tony Zhang and Cody Mathews, 9-19]

During our meeting with Rifat, her main critique was that our proposal visualization sketches weren't very clear and didn't seem to directly accomplish our objectives in our proposal. She suggested that we look for other visualizations to get inspiration and come up with some clearer relationships that we wanted to investigate/demonstrate within our visualization.

Click for Table of Contents

1st Data Retrieval/Cleaning

[Entry by Tony Zhang, 10-19]

After our very brief overview of the potential sources for data for the proposal it was time to start looking into the sources we mentioned in the proposal more carefully. Something that was quickly noticed was that the initial data seemed to cover different periods of time as the initial median income dataset went from 1984-2018 and the population dataset went from 2019-2023. It also did not have county level data. However, I decided to clean it so my teammate could have some data to start working with for the visualization.

Click for Table of Contents

2nd Data Retrieval

[Entry by Tony Zhang, 10-22]

I spent a decent amount of time looking for new data sources so that all the datasets can overlap allowing for visualization that will allow for comparisons between the datasets. It seems that the data on housing prices will be the limiting dataset in terms of the years that can be analyzed as it only goes from 2012-2024. Unemployment data did not cover this time period and could not find more data on this so found a source for job growth which should still get the desired effect that this visualization is hoping for. The median and population datasets are the datasets that I cannot find monthly data for. Hopefully we will be eventually able to find the necessary datasets to find the monthly data for allowing for a more in depth visualization but we should be able to implement all of the desired features with this data. We are probably going to give up on visualizing the data on a county level as I could not find any trace of data that we could use for it.

Sources for initial data:

- **Median Income Data Sources:**
    - 2019-2023: https://fred.stlouisfed.org/release/tables?eid=259515&rid=249
    - 1984-2018: https://www2.census.gov/programs-surveys/cps/tables/time-series/historical-income-households/h08.xls
    - 2024: https://worldpopulationreview.com/state-rankings/median-household-income-by-state
    -
- **Housing Data Sources:**
    - 2012-2024: https://www.redfin.com/news/data-center/

- **Population Data Sources:**
    - 2010-2019: https://www.census.gov/data/datasets/time-series/demo/popest/2010s-state-total.html
    - 2020-2023: https://www.census.gov/data/tables/time-series/demo/popest/2020s-state-total.html
    - 2024:https://worldpopulationreview.com/states

- **Employment Data Sources:**
    - **Unemployment:** https://www.kaggle.com/datasets/justin2028/unemployment-in-america-per-us-state?resource=download
    - **Job Growth:** https://seidmaninstitute.com/job-growth/state/

Click for Table of Contents

2nd Data Cleaning

[Entry by Tony Zhang, 10-23]

For this task, I intended to use the techniques used in the data wrangling course I took (DS-2500). After uploading these datasets to a shared google drive for our team, these techniques did prove useful and the data was cleaned and formatted within a Google Colab Document (see Colab Document here) to be used in the D3 data visualization. The bulk of the cleaning was removing unneeded columns and renaming the needed columns into a shared common one. A majority of the datasets had years listed as columns so they were all melted down into a single year column. Some of the columns had numbers listed with commas and dollar signs so those were removed as well. Finally, the 2020-2024 median income and job growth data sources didn't have data files that I could download so I simply copied the data from the website, pasted it into ChatGPT to format it to be pasted into an excel file, and used those files. Finally, for those datasets that covered different time periods I simply concatenated them together and exported them to CSV files. The Colab document is definitely messy but I am happy with the final cleaned datasets.

Click for Table of Contents

1st Code Entry

[Entry by Cody Mathews, 10-20]

For this task, I started out by creating the US map. I first searched D3's website for graphics and interactions. I came across this map:
https://observablehq.com/@d3/zoom-to-bounding-box

I used the file us.json and chart.svg from this site to draw and render the US map.

I then tied the cleaned data to the us.json. I did this by using Promise.all(). Within this, I linked the paths to the feature selected and tied it to the purple color scheme that adjusted it's saturation fill based on the population in relation to other states.
ttps://d3js.org/d3-scale-chromatic/sequential#interpolateRdPu

I used tooltip to create a hover description tied to each state that displays the state name and data associated with it (population) based on the hover feature that was used in assignment 3 for the bar graph..

Click for Table of Contents

2nd Code Entry

[Entry by Cody Mathews, 10-20]

I created a drop down menu that was linked to each year (2019-2024). Based on the selected year that the user selected, the states would then adjust saturation in relation to each other and correctly display the new population for that year when hovered.

A click feature was added that highlighted the state red when clicked.

Click for Table of Contents

3rd Code Entry

[Entry by Cody Mathews, 10-21]

I created three separate containers to resemble the three graphs that we had shown in our original design thought.

I then implemented the first chart which was a line graph. I used the same theory from assignment 3 to draw and scale this graph. I used red for the line to match the highlighted state.

4th Code Entry

[Entry by Cody Mathews, 10-21]

I implemented the bar graph with the same theory to that of assignment 3. I realized there was a lot of color to the chart, so I changed it to the class standard of steelblue to make it easier on the eyes.

I added a transition to the states for the hover feature. Each state would transition to the highlighted color over 300ms instead of instantly to make it look smoother.

Click for Table of Contents

5th Code Entry

[Entry by Cody Mathews, 10-21]

There were a lot of colors still and wanting to make it simpler, I changed the color scheme of the map to grayscale and the highlights to steelblue.

Click for Table of Contents

6th Code Entry

[Entry by Cody Mathews, 10-22]

Now that everything is working smoothly and updates with the correct features, I decided to add transitions to the two charts as well. I was able to make the entry work, but the exit was still disappearing instantly without the smooth transition that was done in assignment 3.

Click for Table of Contents

7th Code Entry

[Entry by Cody Mathews, 10-24]

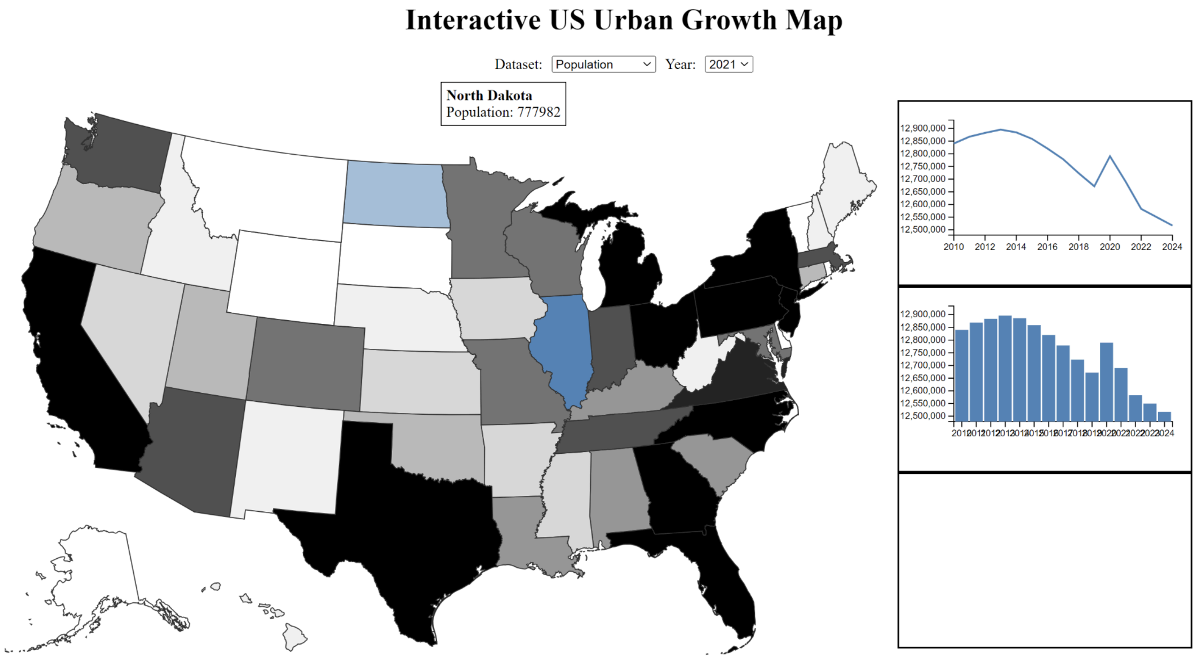
I removed the current data sets and replaced them with updated ones. The reason for this is the current population dataset was missing the year 2022. Also, we felt 5 years of information was relatively small for when we want to do our comparison later on between data sets to find correlations and patterns for growth over time in each state.

I updated both the line graph and bar chart to populate the new data with the extended years.

I was able to implement the exit transition as I found out I was removing the data at the beginning of the method in addition to and before the exit method at the end.

I added another drop down menu to choose from the datasets. I implemented the switch function to swap between datasets. The next task will be to update information in the US map, and graphs to change and reflect the different data.

The image below shows the graphs and click feature tied to Illinois. It shows the population density amongst the states.   It also depicts the hover feature for North Dakota as it's in transition to turning steel blue. The drop down menus to choose between datasets and years are also shown.



Interactive US Urban Growth Map

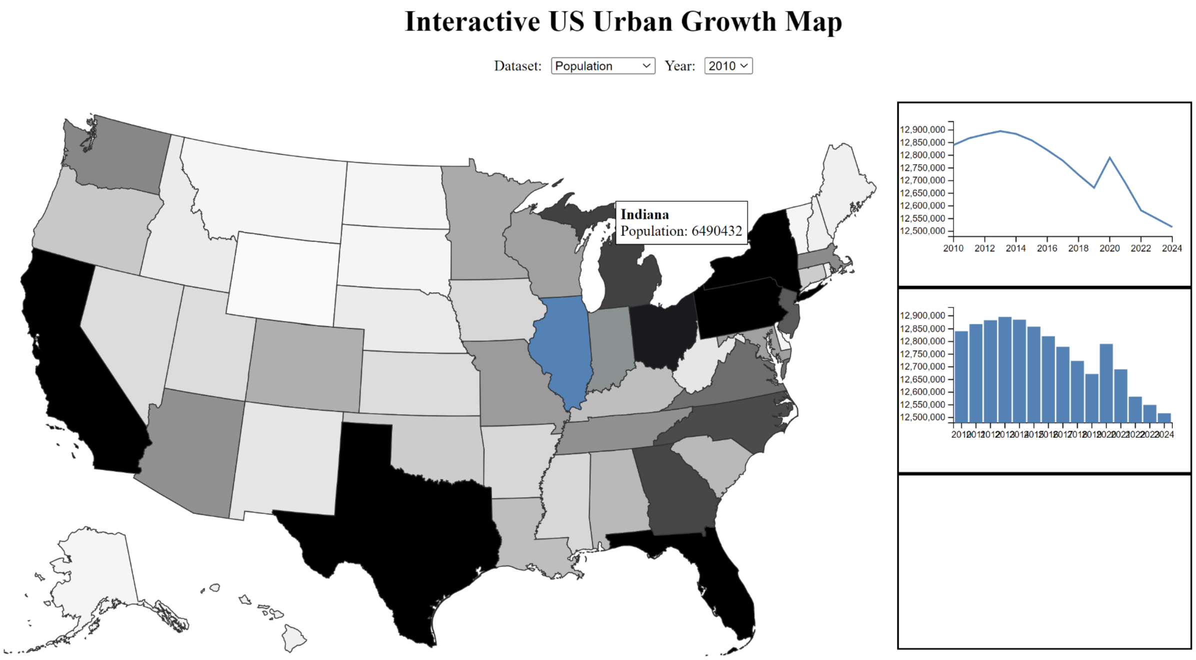
Dataset: Population Year: 2021

North Dakota
Population: 777982

8th Code Entry

[Entry by Cody Mathews, 10-24]

I changed the color scheme from schemeGreys[k] to interpolateGreys(t) and set the max at 2,000,000 for the population at 80%. This change allowed the dataset to go from only 9 gradients to match the scale of the population which allowed for more states such as Georgia, North Carolina and Michigan to distinguish themselves from the states that have more than double their population to give a better depiction of the overall population of each state in respect to each other.



Interactive US Urban Growth Map

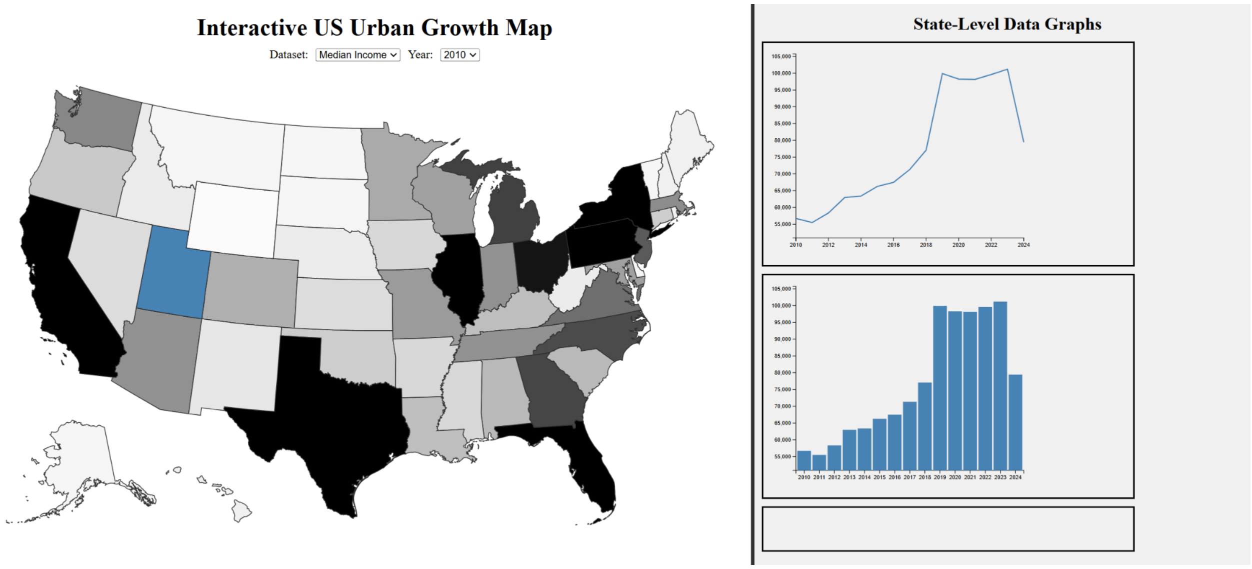
Dataset: Population Year: 2010

Indiana
Population: 6490432

Generalizing Code To Other Datasets

[Entry by Tony Zhang and Cody Matthews, 10-24]

Made changes to the layout of the page so graphs and the map are separated from each other on the page:



Interactive US Urban Growth Map

Dataset: Median Income Year: 2010

State-Level Data Graphs

State level graphs now correctly display for other datasets but heat map does not yet currently, tried many things but did not work, will take some more time to figure out.