US Real Estate Trends Visualization

By Jacob Kim, Lawrence Kim, George Qi

Overview and Motivation

Provide an overview of the project goals and the motivation for it. Consider that this will be read by people who did not see your project proposal.

The United States housing market is one of the most relevant and impactful markets in the world. In particular, the housing bubble in 2006 and 2007 that strongly contributed to the 2007-2009 recession was one of the most significant economic events thus far in our lifetime. That being said, I believe very few people, especially at our age, have a strong understanding of the housing market and the discrepancies in simple metrics, even price. Although we may have general views (i.e. New York or California tend to be expensive, while rural areas remain cheap), we thought it would be interesting to take a closer look at the housing market and generate an interesting and effective way to view the big picture in a single location.

Questions

What questions are you trying to answer? How did these questions evolve over the course of the project? What new questions did you consider in the course of your analysis?

Initial questions include:

- Which geographic areas tend to be more expensive?
- Aside from inflation, when did housing prices tend to decrease/increase? Why is this the case?
- If we made a histogram on the house prices, what kind of distribution would it follow? How heavily is it skewed, if at all?

Other questions that we would like to think about incorporating into another visualization beyond just the analysis of the housing market would include linking this data with data on the US economy. It is intuitive that when the economy is growing, so should the average house price, but the real question is: to what degree? Perhaps although the country's GDP is rapidly growing, certain states or cities have yet to see their house prices increase very much. Or perhaps

there are certain odd time periods that we can point out from our trends that can be linked to an economic event, such as a law passed or whatnot. Bringing in other datasets would certainly be very interesting.

As for benefits, we believe our project would solve a common problem that we've noticed. When observing discrepancies in house prices, we have come across little to no visualization tools. We generally come across raw numbers for certain states, cities, etc., but in this form of presentation, it can be difficult to draw conclusions about the housing market (especially regional ones). This is exactly where visualization comes in and works best. Our project would enable newer, less informed people to learn a great deal of information in a limited amount of time. Its interactive nature would also attractive interest from people who would have otherwise been less interested.

Bringing in the additional economic data sets would also give the users an even better understanding of how the economy as a whole affects and is affected by the housing market.

While brainstorming what other views we could include in the visualization, we examined our data set to see what other questions we can answer. Our team also spent a considerable amount of time looking for data sets that might be interesting to add to our project. As the project developed, we came up with new questions that our visualization could possibly answer.

Some of these new questions include:

- During what years did the housing prices increase the most?
- How does a certain state's housing price trajectory compare to that of the average state's?
- What are some patterns involving the size of the home for certain states (1 bedroom, 2 bedroom, etc.)?
- How does a certain state's GDP trajectory compare to that of an average state's GDP?

Data Processing

Zillow House Value Data

Creating a JSON file to contain all the data was more challenging than expected. We created a Python script that scraped all the data from the various CSV files and combined them into one JSON file called ‘data.json’.

Initially, we decided to only scrape data for 50 major cities in the US. For missing values in the data, we stored the value to be -1.

We also added the cities’ latitude and longitude information to the data to help us place the nodes on the map. The latitude and longitude data was scraped from another data source found online and added to our current data set. This also required writing another Python script.

Here’s an example object from our JSON file:


```
{ "city" : "New York, NY",
   "latitude" : "-73.9865813",
   "longitude" : "40.7305993",
         "months" : [ { "1br" : -1,
        "2br" : 135800,
        "3br" : 171200,
        "4br" : 222000,
        "5br" : 276400,
        "allhomes" : 168500,
        "month" : "1996-04"
          },
          { "1br" : -1,
        "2br" : 136100,
        "3br" : 171800,
        "4br" : 222000,
          "5br" : 276100,
        "allhomes" : 169100,
        "month" : "1996-05"
          },…]
}
```


Once our visualization was more developed, we added 50 more major cities to our data set.

GDP Data

For the economic data, we decided that the most useful standards to measure against would be stock data and the United States GDP.

In particular, we found United States monthly GDP data for April 1996 to February 2015 from “YCharts”. It came in a CSV file, which we used to calculate percent gain or loss from month to month. From Yahoo Finance, we found monthly prices for two of the most important measurements: NASDAQ and S&P 500. Similarly, we calculated percent gain/loss, and then using our previous functions, we put this into a JSON.

An example object for the data “econdata.json” is:


```
[ { "gdp" : "17.69",
   "gdp_perc" : "0.000565611",
   "month" : "2015-02",
   "nasdaq_adj_close" : "4963.52979",
   "nasdaq_perc" : "0.070824713",
   "nasdaq_volume" : "1985720000",
   "spy_adj_close" : "2104.5",
   "spy_perc" : "0.054892511",
   "spy_volume" : "3806470500"
  }
]
```


However, after speaking with our TF Alain, we decided to look online for a data set that included state specific metrics. We also thought that it would be useful to find consumer price index data instead of GDP to compare our housing price data to. Unfortunately, we were unable to find a CPI data that was state specific. We also considered Disposable Income as well; however, this data was also not to be found.

Eventually, we came across a data set that included state specific GDP data for each year and decided this was the best data set available to us.

Here is an example object found in our state GDP data found in the file “state_gdp.json”:


```
{ "state" : "AL",
   "years" : [ { "gdp" : 23407,
        "year" : "1996"
      },
      { "gdp" : 31129,
```

```
      "year" : "1997"
    },
    { "gdp" : 32128,
      "year" : "1998"
    },...]
}
```

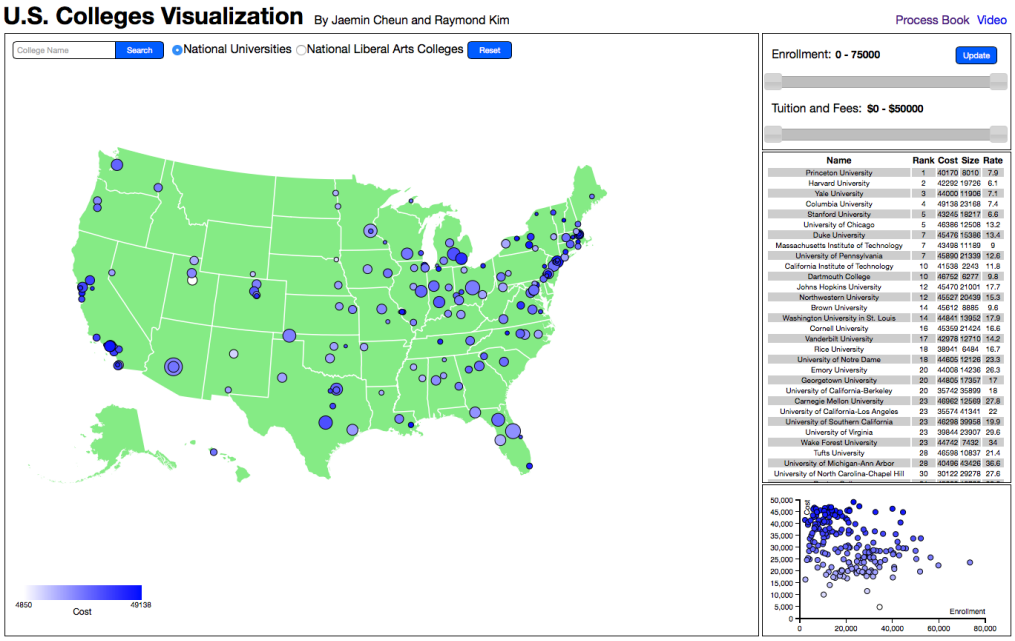

Note that the units on the National GDP and State GDP are different (the national GDP being in trillions of dollars while the State GDP is in GDP per Capita, so we had to adjust for this in our code in order to provide a normalized view incorporating both of them). The national GDP was also monthly whereas the State GDP was yearly, so I averaged the monthly national GDP values to yield a single annual national GDP value.

Related Work & Exploratory Data Analysis:

Anything that inspired you, such as a paper, a web site, visualizations we discussed in class, etc. What visualizations did you use to initially look at your data? What insights did you gain? How did these insights inform your design?

After finding our data set, we knew that we wanted to implement a map view that showed the locations of the cities. We came across one of last year's projects and used it as inspiration for our page design. This project provided a good example of a map view with multiple side views.

http://cheunjm.github.io/cs171-project-jcheun-rkim/



At first, our initial idea was a map with nodes on it; we indeed implemented this, and we will go into more detail with this later. Our logic was that we could use nodes for the most important cities, which would prevent the map from becoming too cluttered.

However, after feedback from our TF this past Monday (April 13), we realized that an alternate layout might also be very effective. A choropleth map, perhaps like the one in this visualization: http://bl.ocks.org/mbostock/4060606.

If we used the choropleth map, we could definitely use the counties information and have a much larger and more extensive data visualization. In particular, we can allow switching between the two modes, as sometimes having nodes for each of the major cities may be the more helpful visualization in certain circumstances.

However, after another meeting with our TF, we decided to scratch the idea of a choropleth map. Our data set had too many missing data values that a choropleth would not be the most effective way to display our data. Nodes on a map appeared to be the best option.

Design Evolution:

What are the different visualizations you considered? Justify the design decisions you made using the perceptual and design principles you learned in the course.

Milestone 1 (April 17)

After having processed the data, we then began to create visualizations of the map.

There were a couple things to consider before beginning with the map visualization

1. What would we want to display the map
2. What would be possible interactive features we wish to include?
3. How should different housing prices be scaled on the map?

Beginning with the first consideration, we knew that we would want to display the housing data based on for the different subclasses of houses ("1 bedroom," "2 bedrooms," "allhomes"). Ideally, we would wish to be able to filter between the different subclasses, and perhaps be able to combine them. In the case that more than 2 or more subclasses are selected, how would we visualize the data? There were two options that we had in mind: using the average of the subclasses and the sum. Because of the fact that we do not have weights for the proportion of houses that belong in each category, an average may be very misleading, simply due to the fact that perhaps there are much more houses of one subclasses in

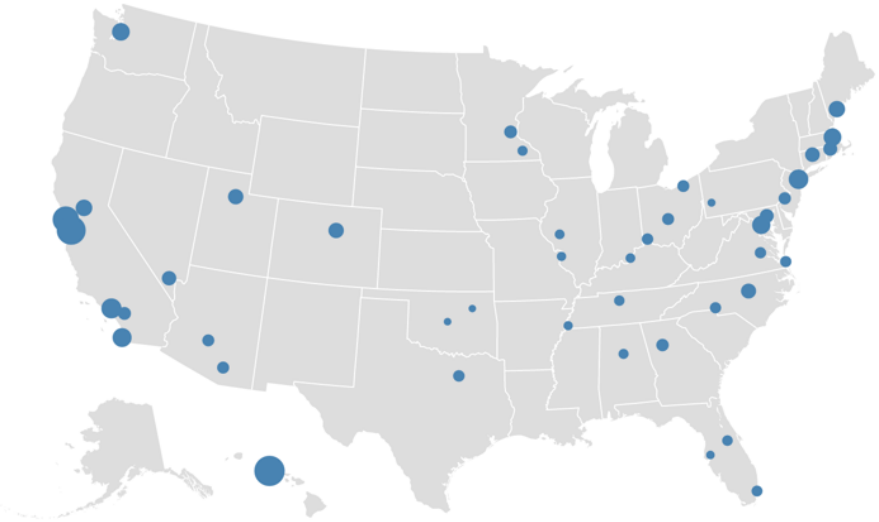
comparison to another. This could perhaps lead us to thinking that housing prices might be larger in one area when in fact they are not (think Simpson's Paradox).

In terms of the interactive features on the map, there were several ideas we considered. In addition, there were different levels of interaction to consider.

The first level of interaction would be shaping the map by changing the way that the map was represented in general. For this level of interaction, we implemented a time slider that allows users to alternate between different months to see the housing prices/housing relations between areas during that time period. In addition, we wish to implement a filter option to filter which subclass of housing prices is being displayed. This, currently, is not fully implemented on the prototype that we have included with the first milestone (currently only "allhomes" data is presented). However, we will definitely have this option present, and have a good solution to representing two subclasses at once (perhaps with shades of color?).

Another interaction level that we had to think of is a user's interactions with the map itself, without necessarily manipulating the data presented. Although this level of interaction is not implemented, we have several ideas wish we will implement. A primary example of this would be the interactivity that would trigger if a user hovered/clicked on a node. An obvious interaction feature would be to have a bubble of text to appear that allows you to know the city and home prices of the node that is hovered upon. Another feature we wish to implement, however, is to show an area's changes in housing prices over a time period, such as the 12 months before the current month that is selected. This would be visualized most likely with a line plot. In regards to how housing prices should be scaled on the map, we currently are utilizing nodes that are present on each city.

Our current visualization of the map is displayed below. As can be seen, the nodes are currently scaled based on "allhomes" prices.



Below, please find a picture of our slider as of now:

Currently, it is completely functional, and open moving the time, the nodes will indeed change sizes and the table will update. The filter also works and successfully only considers the certain bedroom type that we are interested in.

Next, please find a view of the table.

Real Estate Prices

City	State	All	1br	2br	3br	4br	5br
Atlanta	GA	$108,400	$82,600	$73,900	$96,800	$160,000	$228,200
Baltimore	MD	$119,700	$75,900	$86,000	$112,000	$172,700	$247,700
Birmingham	AL	$87,700	$44,200	$39,500	$73,400	$157,400	$240,100
Boston	MA	$152,600	$86,100	$116,900	$152,500	$221,500	$290,500
Buffalo	NY	No Data	No Data	No Data	No Data	No Data	No Data

Somewhat analogous to Homework 1, this filters, sorts, etc. When the time bar moves, everything updates, and sorting is maintained with respect to what the user had previously wanted to sort by.

Using this table, the user can very easily understand which cities and states tend to be more expensive, how much each type of bedroom costs, and so on.

In terms of visualization, we always considered using a map for our visualization. However, how exactly we would display the data was the main thing that changed as we considered different visualizations. At first, we considered using nodes to represent the different cities, with the size of the nodes changing based on the housing prices for each city. This visualization, we have decided to stick with. However, there are other alternatives that we began to consider. The choropleth alternative, for example, that we have mentioned above is something that we began to consider.

The biggest evolutions to visualization occurred in regards to interactive visualization and the other views that we would include. At first, we were planning on having a pie chart to represent different ratios of housing prices for the different sub-classes, upon hovering on nodes. Although this is something that we are still considering, we have currently decided to go with a bar chart because it easily presents the scale of housing prices visually, due to the presence of the axes. Both bar charts and pie charts are equally good for being able to compare the different sub-classes against each other visually.

We also did not initially consider putting in a line plot to represent change over time of housing prices (this has not been implemented yet, but will definitely be done). A reason for deciding this was because although the map visualization and table visualization provides a good measure to represent housing prices over the current month, our web page didn't have a way for comparison over time. Therefore, we thought that such a feature would be absolutely necessary.

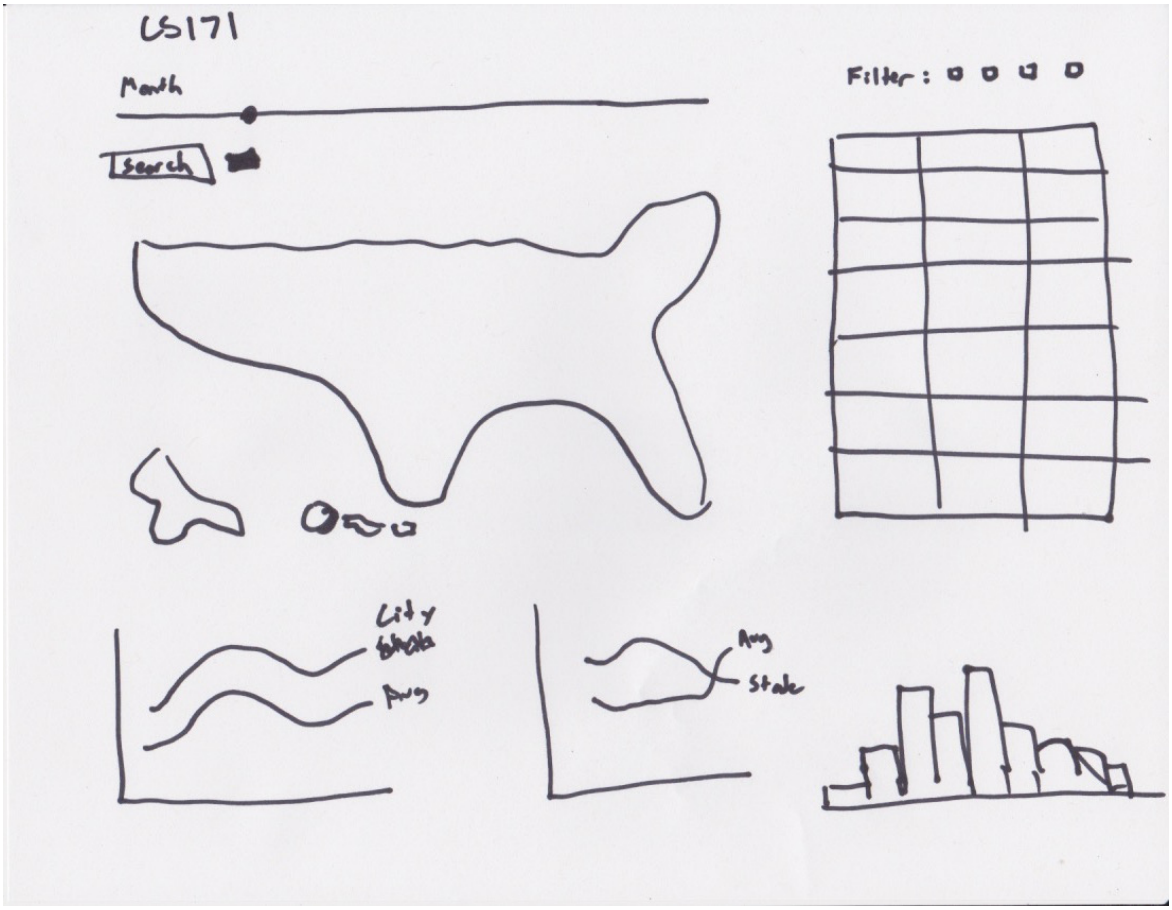
Finally, one of the overarching decisions for the implementation of things like the datatable that we implemented for the milestone is the realization that interactivity is one of the most important things that we should stress (which plays in with the decisions for many of our visualizations discussed above). In particular, we considered having things a line graph of the way house prices changed over the entire time period, but with a time slider, this is a static view rather than dynamic and does not engage with the user as well. A table, on the other hand, is interactive and can show how certain cities go up/down with respect to 1 bedroom houses, for example. Histograms like the distribution of house prices for a given filter can be interactive and useful to understanding phenomenons like whether 1 bedroom houses are becoming more or less sought after; we can notice this by observing that their distribution is becoming less normal and instead more skewed, for example.

Final Product (May 5)

Choropleth

We decided not to implement the choropleth map due to several reasons. We did not have enough data to provide a useful choropleth map. For example, we did not have home value data for every city in America–in fact, we only had data for a small fraction of all the cities. Also, many of these cities had missing values. Thus, much of the map would be empty and not be a useful visualization.

Page Layout



Final Design Sketch

For a visualization that had multiple views that interacted with each other, the best idea was to place all views viewable on one page. With this feature, the user can see all at once how certain interactions, such as using the time slider, affects all the views. We used Twitter bootstrap to help put our views in place.

All interactive elements (time slider, filters, search, map, table) were placed together on the top of the page to encourage the user to explore the interactivity of the visualization. We also do not want the user to have to scroll down to interact with some of the visualizations.

The filter and search bar placed on the left side directly below the time slider interacts with the three views under it (map and two line graphs). Similarly the filter on the right interacts with the two views under it (table and the histogram). As discussed in class, the location of certain elements is very important creating an intuitive and helpful visualization.

The map and the table are the two biggest views because they require the most user interactivity. Both views also contain the most data.

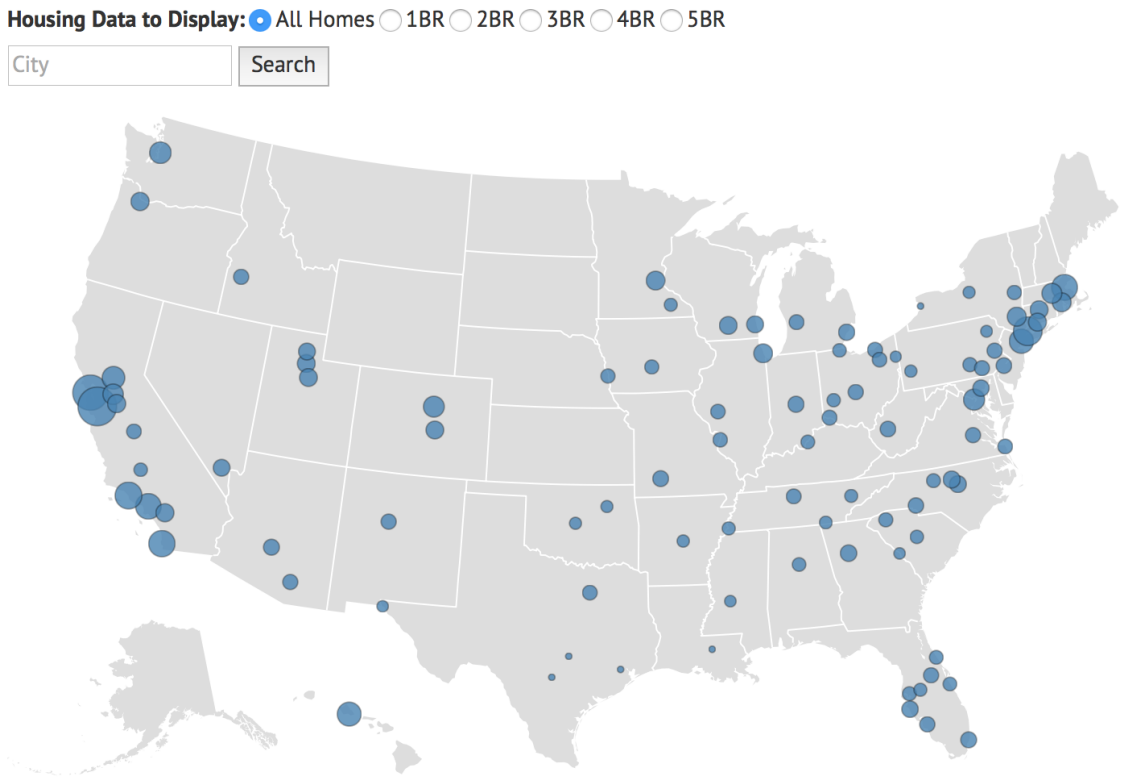
Implementation

Describe the intent and functionality of the interactive visualizations you implemented. Provide clear and well-referenced images showing the key design and interaction elements.

Time Slider

This is a standard time slider that controls the time period of our visualization. The text of the "Current Month" changes as you move the slider indicating which month we are looking at. This slider is a very important part of our visualization since it interacts with all 5 views on the page. Because of its importance, we decided that the most intuitive place to place it was at top of our web page.

Map View



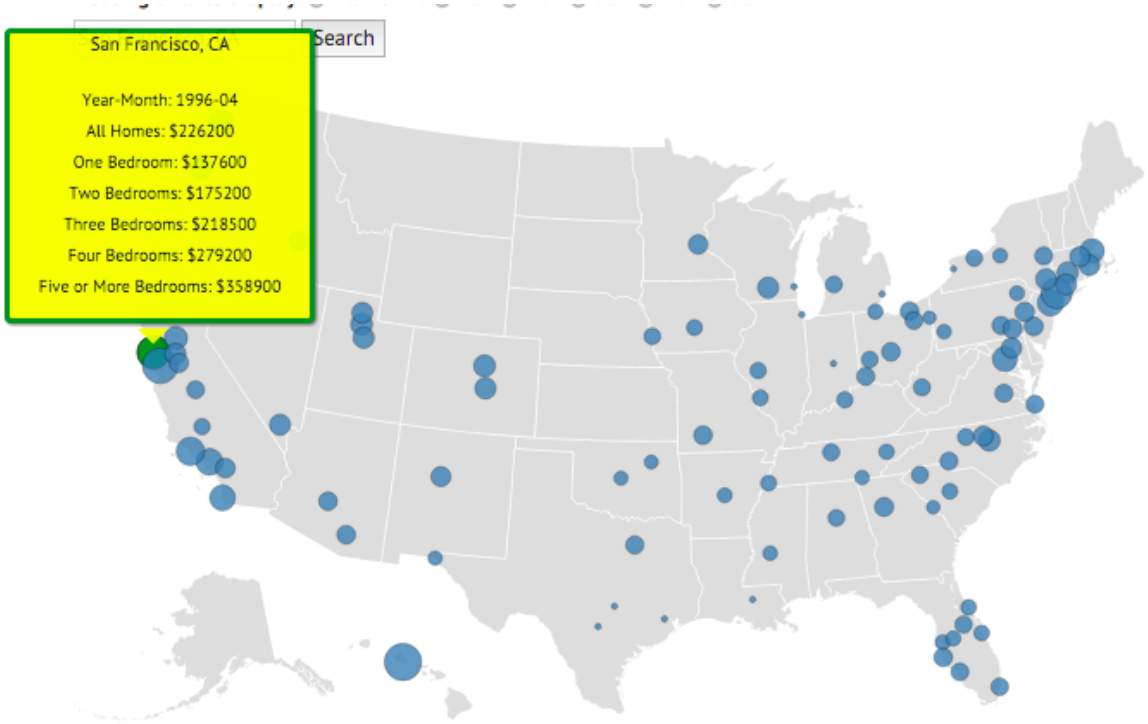
You will notice a few main differences from our milestone view.

Home Type Filter

We added a radio filter at the top to visualization the home prices based on home type. We did this because home values vary largely due to the size of the space. Thus, it is important to compare prices that should be controlled for size. As you click on the radio, the sizes of the nodes will change depending on the data.

Search Bar

It is important to include a search bar in the case that a user wants to see the individual stats for a specific city of interest. For example, a graduating Harvard student might want to move to San Francisco, CA and would like to see a price estimate of a 1 bedroom apartment.



Wow San Francisco is expensive!

Also one nice detail is that the search bar will auto-complete the city name for the user to help the user find the city he/she is looking for.

Nodes

When we added 50 more cities, the nodes on our map appeared to be very cluttered and difficult to distinguish. Thus, we lowered the opacity of the nodes and added a border to help visualize

the borders between cities. This makes it much easier for the user to distinguish cities in the Bay Area of California and in New England.

We also added the feature of displaying a city's data when hovering the mouse over it. This way a user can explore data just by moving the mouse on the map. The node changes color as when hovered over to help the user distinguish which city he/she is looking at.

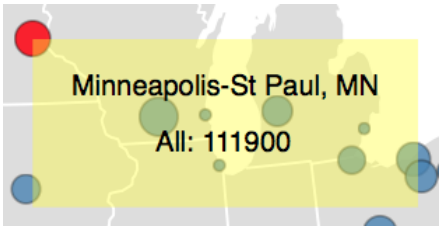


Table View

Filter By: All 1br 2br 3br 4br 5br

Real Estate Values

City	State	All	1br	2br	3br	4br	5br
Akron	OH	$98,600	$54,400	$71,300	$96,200	$164,100	$159,100
Albany	NY	$98,400	$62,600	$76,300	$95,500	$130,100	$133,900
Albuquerque	NM	$116,500	$82,600	$94,300	$119,400	$157,300	$201,400
Allentown	PA	$107,300	$76,400	$82,800	$97,800	$162,200	$98,300
Atlanta	GA	$110,200	$82,900	$74,300	$98,600	$162,000	$232,700
Augusta	GA	$64,200	No Data	$50,900	$74,300	$121,000	$171,000
Austin	TX	No Data	No Data	No Data	No Data	No Data	No Data
Bakersfield	CA	$85,400	$42,200	$56,700	$82,800	$121,200	$170,100
Baltimore	MD	$119,100	$72,800	$85,000	$111,400	$172,300	$246,300
Baton Rouge	LA	No Data	No Data	No Data	$115,100	No Data	No Data

The table is a great way to compare cities and their home values.

<u>Filters</u>

While the default table shows all the data, the filter allows us to look at only certain home values based on the home type. The filters are important because it removes unnecessary values that a user might not be interested in. For example, a graduating senior might only be interesting in comparing home values of 1 bedroom apartments and has no interested in knowing the values of a 4 bedroom place. Hiding unnecessary data is helpful to the user.

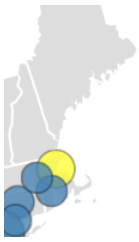
<u>Headers</u>

The table headers are clickable and will change the ordering of the table depending on which header was clicked. If you are wondering why the header of the state column is not lined up perfect,

Scrollable

Also, another important detail is that the table is scrollable. The table is very large in size; thus, making it scrollable helps make the entire visualization fit in one page. Also, it is important to note that the headers are static while the rest of the table body is scrollable. This is helpful for the user to know which columns he/she is looking when looking at the bottom of the table.
*Side note: The column of "State" is not perfectly lined up because of this feature. The headers are included in a separate table element to make it static. We could not figure out how to perfectly line them up with this feature.

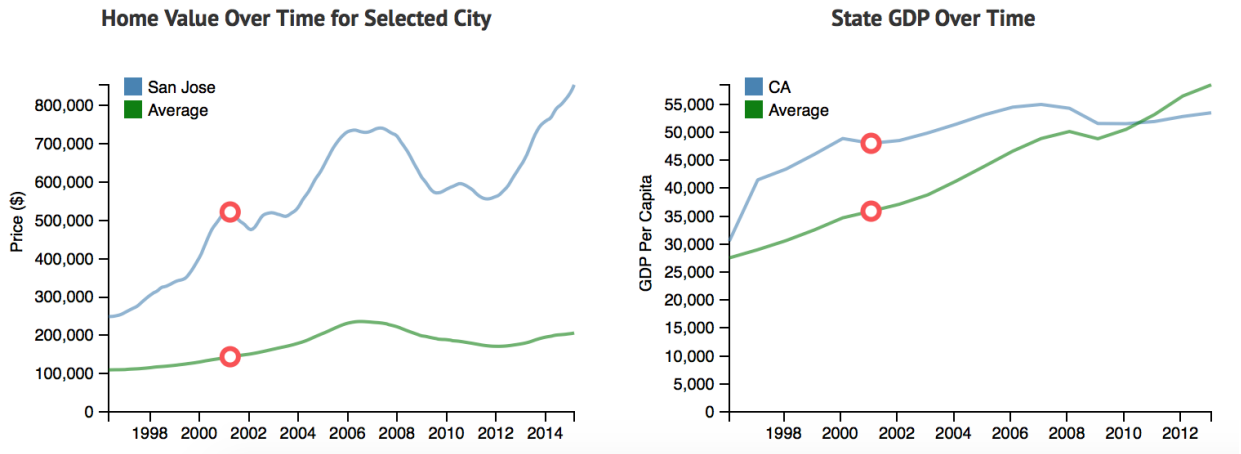
Highlighting



Birmingham	AL	$87,400	$40,800	$39,800	$75,100	$159,500	$232,600
Boise City	ID	$112,200	No Data	$79,400	$103,300	$142,200	$184,100
Boston	MA	$157,800	$87,700	$121,900	$158,400	$227,100	$307,400
Buffalo	NY	No Data	No Data	No Data	No Data	No Data	No Data
Charleston	SC	$87,000	$69,100	$66,100	$89,000	$159,400	$257,000

When the mouse hovers over a row of the data, the row will be highlighted. This helps the user know what he/she is looking at. Another cool feature is that the city that is being hovered over will also be highlighted as yellow on the map. The user will now able to see the geographical location of the city that they are looking at.

Line Views

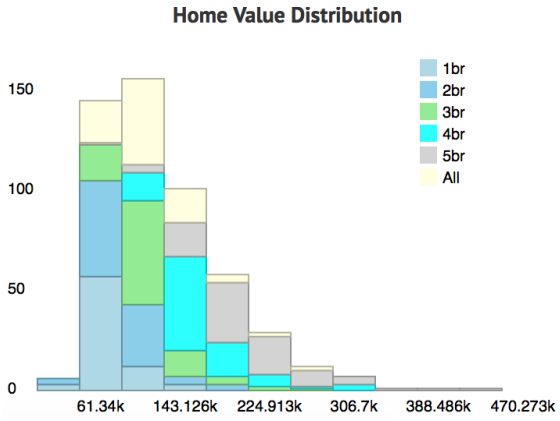


These line plots allow us to take a closer look at a certain cities home value trajectories over time. They also allow us to make comparisons. For example, from the screenshot above, it appears that San Jose home values has skyrocketed over the last few years and is much higher than the values of an average state.

The GDP view provides another view that can help the viewer gain new insight. The plots above show that California per capita GDP has increased over the years which can be associated with the increase in home values.

There is a red circle over the lines that correspond to the time slider. This helps the viewer connect these line plot views to the other visualizations in the page and identify the time period we are looking at.

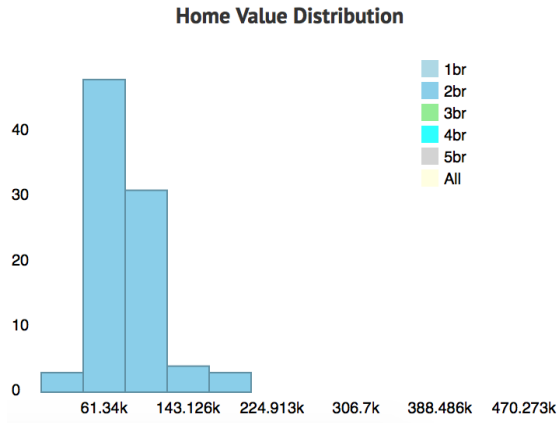
Histogram View



This histogram view serves the basic purpose of visualizing the distribution of the home values. It is no surprised that the prices are right skewed.

We added an element of color to help the user see how exactly home size contributes to the distribution. For example, it appears that many of the homes that are 61.34k are 1 bedroom. There are no states with 1 bedroom apartments that are worth more than 307k.

This histogram also interacts with the filters that are found above the table view. The filter allows the user to look at the distribution of values for a particular type of home. Here is an example of a histogram for 2 bedroom homes.



Evaluation

What did you learn about the data by using your visualizations? How did you answer your questions? How well does your visualization work, and how could you further improve it?

Milestone 1 (April 17)

As of the Project Milestone, we have finished all of the requirements; that is, we have all of our data that we could possibly want (the CSV files, and more importantly, the code to scrape any additional code that we may need), as well as a functional map prototype that demonstrates the direction that we plan on taking. Our structure is also completely set up such that we can simply create new JS files and directly integrate them. By separating the JS files and the "index.html" file, it makes each of the views shorter and more readable.

In particular, we plan on still taking on the views that we showed in our Proposal, as well as a version of the choropleth map as well. One strong feature that we also want to add is, whenever we hover over a county, it would display its historic house prices for a certain house type. You could then have another line for the average house price for the entire country for comparison. One weakness would be that there could potentially be a nontrivial number of missing data points for a given county, but that does not detract from the power that this visualization would yield. This is strongly linked with the discovery of the importance of interactivity as mentioned in the previous section.

Therefore, as of now, we are confident in our progress and our future direction. We have a lot of ideas and have demonstrated our ability to gather and process all of our necessary data as well as create some difficult visualizations (namely, the map).

Final Product (May 5)

What did you learn about the data by using your visualizations?

Some interesting facts that we learned:

- California, Honolulu, and New York are very expensive.
- State GDP is slightly positively correlated with state home value prices
- Home values are heavily right skewed.
- In 1996 Honolulu was the most expensive. Today, San Francisco and San Jose are the most expensive.
- The cities on the coasts tend to be more expensive that cities in inland US.

How did you answer your questions?

We answered these questions by examining geographic location of certain expensive cities. We also compared housing price trajectories over time and their relationship with state GDP but examining the two line graphs displayed. The time-slider and the table allows us to see how the rankings of the city change over time. The histogram shows us the distributions of the home values and how those change over time.

How well does your visualization work, and how could you further improve it?

Our visualization works fairly well. The design of the page appears to be very clean and organized. The interactive elements are also intuitive and easy to use. Interaction with the visualization performs very smoothly and the speed of the visualization remains fine as the data is quickly accessed and filtered.

One thing that we could improve is the search functionality. The search functionality seems to rely on the auto-complete and does not work if you push enter. You have to click the search button to make it work.

It would also be nice if the red marker on the line graphs moved more smoothly as the user uses the time slider.