Process Implementation (Rough Draft)

Data Scraping:

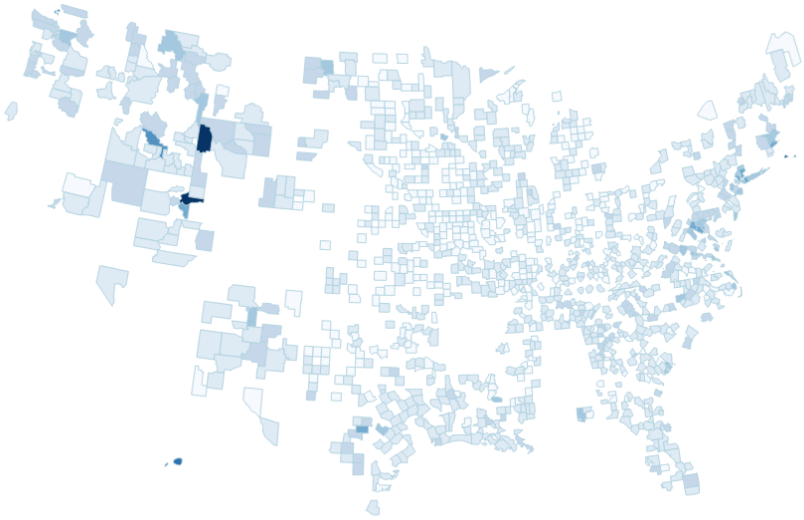
Data Scraping and Munging composed the majority of efforts leading up to Mile Stone 1. Using the Quandl API, every available housing dataset and associate metric (allHomes, Median Listing Price, 1 Room, 2 Room, etc) for each US County was downloaded. However there were some significant hurdles during this stage. First, there are approximately 3200 US counties and more than 32 associate metrics for each county, which increased the number of requests to Quandl. Quandl however has a limitation of approximately 2000 requests per 10 minutes, which limited the ability to scrape the dataset. Also there are different types of requests that must be made to access certain types of datasets, which further complicated the process. Originally the data was saved as a .json file, however this file size proved to be too large (almost 1GB) and the smaller .csv format was chosen instead.

All data was scraped using Python. All relevant data scraping code can be located in the folder titled /scrape.

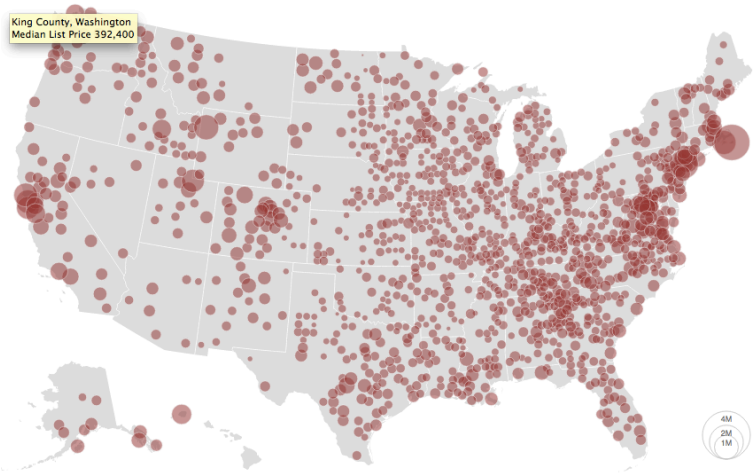
Initial Visualization:

We experimented with different visualizations after discovering the data set provided by Zillow was actually quite sparse in geographical coverage. For example, we were not pleased to see this as the resulting choropleth map in our initial attempt, using the metric “Median Selling Price”:

We then realized, certain metrics are more lacking than others. We eventually choose to go with “Median List Price” as it seemed like the most populated metric among counties. Yet, the map still looks rather incomplete:

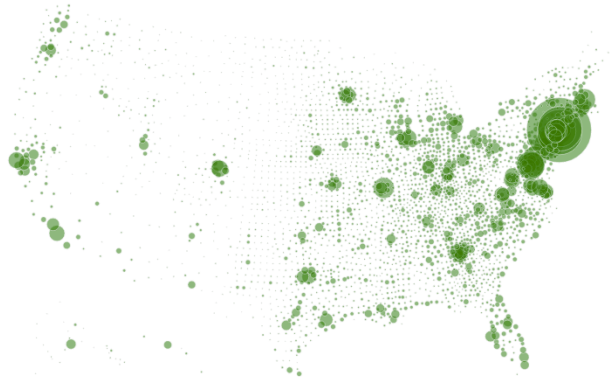


Also, we decided that the physical geography was not the focus of this visualization, rather, the housing prices themselves. With this new focus in mind, we opted to use the following graduated symbol map. We decided the most appropriate graduated symbol is a circle proportional to the associated Median List Price:



This initial visualization made sense on multiple levels, as compared to our initial cholorpleth maps. First, it allows discovery of housing prices in certain regions. On the cholorpleth map it would go unnoticed that Nantucket County in MA would have one of the highest median list prices in the US due to the county's small size. Secondly, Zillow's data scarcity made sense as not every inch of the United States is populated. Our visualization should aim, instead of for geographical completeness, for hotspots or important cities and regions that people actually live in.

Next we wanted to get a sense of what factors could affect such housing price discrepancies. This is not an easy task and may go outside the scope of our visualization. Nevertheless we plotted population densities from the National Census Bureau:



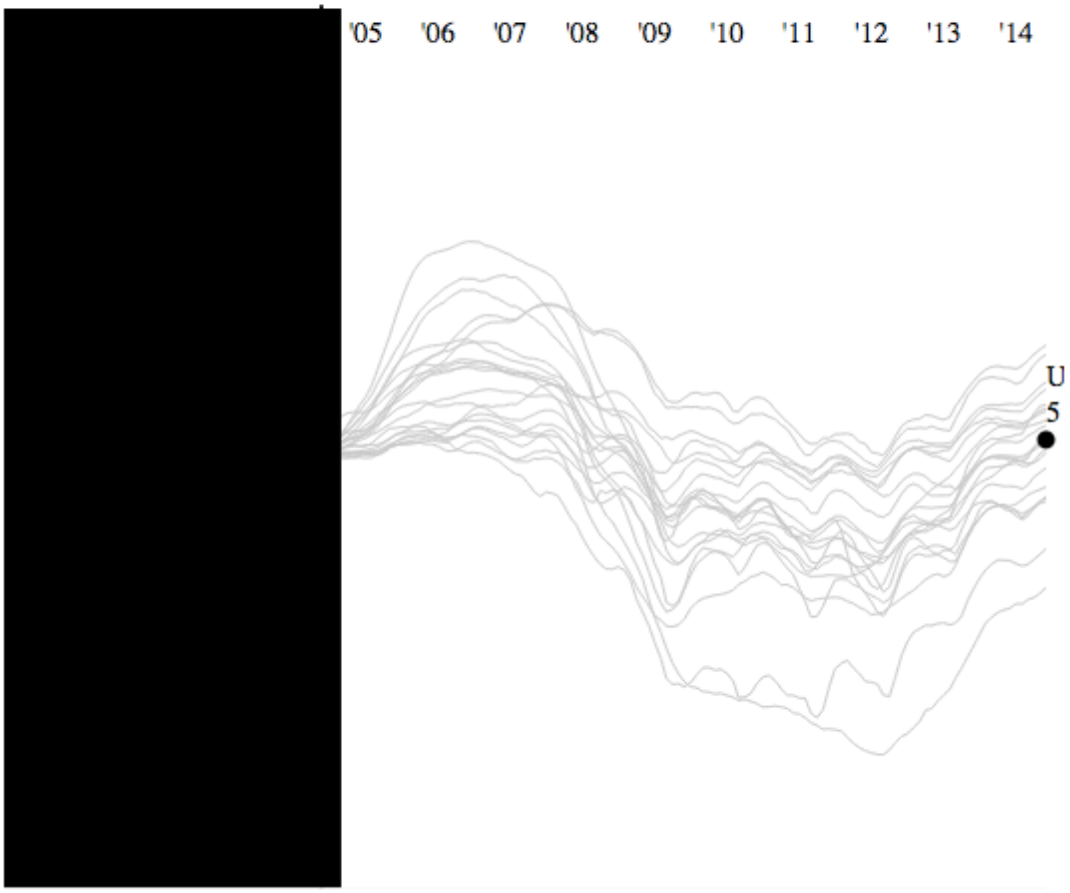
It paints an interesting picture, but the housing market is much more nuanced than population densities alone. It is interesting to note, populations are dense near major railway intersections, but this does not determine housing price as many hubs are actually cheaper than places that are not hubs.

Another question our visualization attempts to answer is how housing prices have varied over time as opposed to region. Eventually in further iterations we hope to add more dynamic control over the graduated symbol map (such as selecting for certain metrics such as single family home, etc. in our current dataset as well as have an automatically increasing slider to demonstrate changing housing prices).

We have been experimenting with different graph and chart combinations, one showing only major US cities and the percentage change over time. We will need to better integrate this particular visualization however as it currently uses manually constructed from data scraped from Quandl:

If you bought an average home in a major city **around** Jun. 2004 **it would be worth 5 percent more today.**

Price changes from **June 2004**...
...to **June 2014**



City	Year-over-year change	Since peak	Since June 2004
Las Vegas	+15%	-42%	-27%
San Francisco	+13%	-11%	+14%
Miami	+12%	-34%	+4%

A challenge for our future design iterations will be how to handle the large amounts of data, without making a slow and burdensome visualization, as well as realize our initial sketches for the map options contained in proposal.pdf.

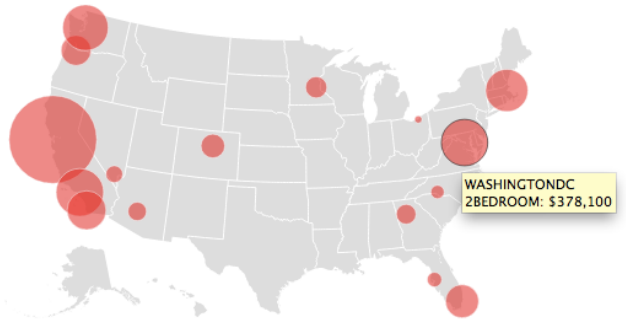
Post-Milestone 1

After our meeting with Alain, we took measures to expand both the interactivity of our initial visualizations and the visualizations required to further our story of the post-mortgage crisis housing market. Alain suggested that for our graduated symbol

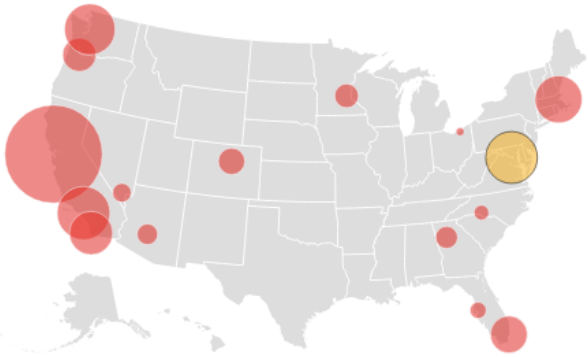
map that we decrease the number of counties to focus on or implement some kind of zoom and pan ability to our map.

After some discussion we felt that the visualization should focus on an index of 20 cities commonly used to represent the US housing market. In this way, the visualization could run faster (wouldn't be bogged down by updating data for 3200+ counties) and be more interactive while still telling a strong story. Also, per Alain's suggestion that we integrate the time slider with the map as well as have actions on the map highlight metrics on both the graph and table, we decided to implement dual click-hover functionality for own map.

When hovering, a user can see the specifics of a particular city:



Upon a click, the city is highlighted:

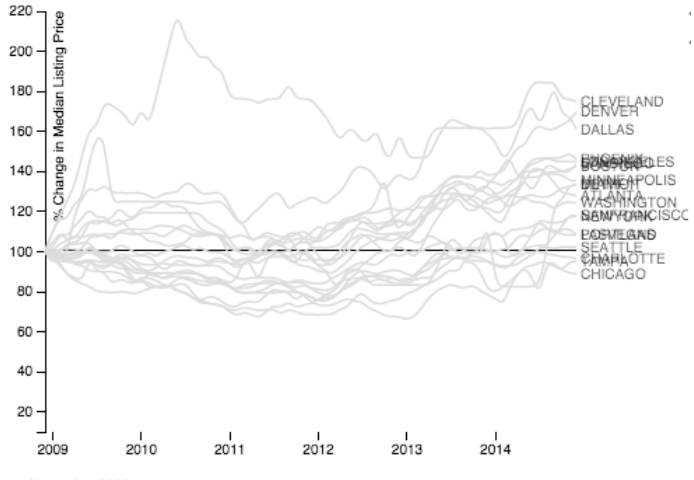


We felt the graduated symbol map combined with cities provided users both the ability to see the relative price difference between cities (as a chloropleth map would not be effective in communicating price differences on a city level (ie. Not enough area per city to demonstrate a color gradient)).

Furthermore, we expanded the interaction between visualizations for our main presentation. We decided that there should be five main components that would be important in enhancing our storytelling ability as well as allowing users to explore.

First, as discussed in length, a map of the United States would be the main centerpiece to let users understand where growth, stagnation, etc are occurring.

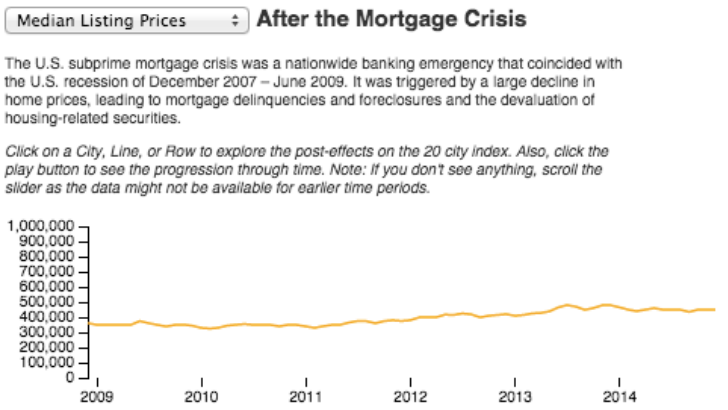
Second, a graph showing the % change of the metric considered. This gives a clear point in our story of recovery post mortgage crisis since it easily allow users to see which cities have recovered and which have not based on if they fall above 100% (performing better than 2009) or below (performing worse than 2009):



Third, a table that shows the actual values of the time period considered, as well as calculations that compare to 2009 and compare to the peak of the metric.

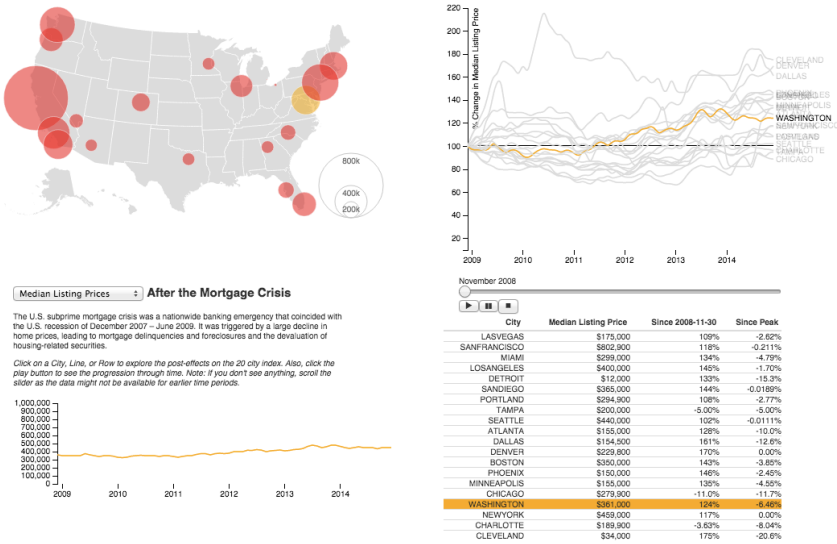
City	Median Listing Price	Since 2008-11-30	Since Peak
LASVEGAS	$175,000	109%	-2.62%
SANFRANCISCO	$802,900	118%	-0.211%
MIAMI	$299,000	134%	-4.79%
LOSANGELES	$400,000	145%	-1.70%
DETROIT	$12,000	133%	-15.3%
SANDIEGO	$365,000	144%	-0.0189%
PORTLAND	$294,900	108%	-2.77%
TAMPA	$200,000	-5.00%	-5.00%
SEATTLE	$440,000	102%	-0.0111%
ATLANTA	$155,000	128%	-10.0%
DALLAS	$154,500	161%	-12.6%
DENVER	$229,800	170%	0.00%
BOSTON	$350,000	143%	-3.85%
PHOENIX	$150,000	146%	-2.45%
MINNEAPOLIS	$155,000	135%	-4.55%
CHICAGO	$279,900	-11.0%	-11.7%
WASHINGTON	$361,000	124%	-6.46%
NEWYORK	$459,000	117%	0.00%
CHARLOTTE	$189,900	-3.63%	-8.04%
CLEVELAND	$34,000	175%	-20.6%

Fourth, another graph that demonstrates the actual values of the metric considered over the entire time period of available data since 2009:



Finally, fifth to tie all the visualizations together an autoplay function that gives user an ability to effortlessly see the dynamic change over time as the time line progresses.

Ultimately, the disparate elements should all be highlighted when either city, row, or line is clicked. Also upon click, the entire available metric data should be included. This also solves multiple issues with clutter in the visualization. If we consider the percentage change chart above it is very difficult to a) distinguish between line b) see which line is associated with which city label. Likewise, the table is somewhat overwhelming without some sort of guidance towards a main topic. We settled on the intial map selection color (yellow) to clearly distinguish a selected city and all its data from the others. We also increased the opacity of the title for the percentage graph to solve for issue b) above:

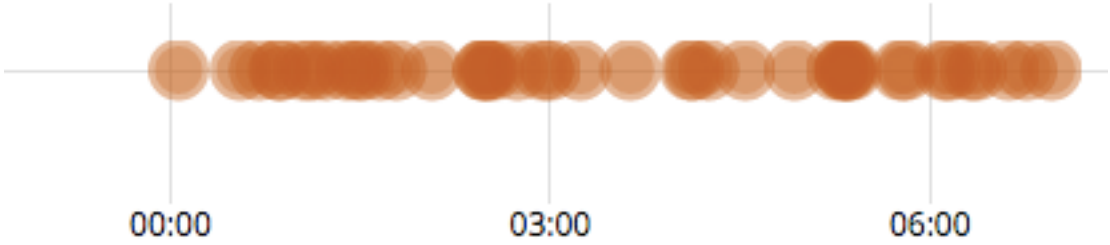


With the added autoplay functionality, a user is able to follow a particular city's recovery or further decline over the period after the mortgage crisis. Give it a try by pressing the play button on the website!

Also, as per Alain's suggestion that we should focus on some key metrics as opposed to every possible metric our dataset contained, we chose the following: Median Listing Prices, 2 Bedroom Prices, Rental Prices, and Foreclosure per 10000 homes.

⇒ Median Listing Prices have largely increased post-crisis
⇒ Two Bedroom House Prices have exploded, especially in the San Francisco Bay Area post-crisis.
⇒ Rental Prices have fallen somewhat, but most have returned to post-crisis levels.
⇒ Finally, the level foreclosures have decreased significantly since the crisis

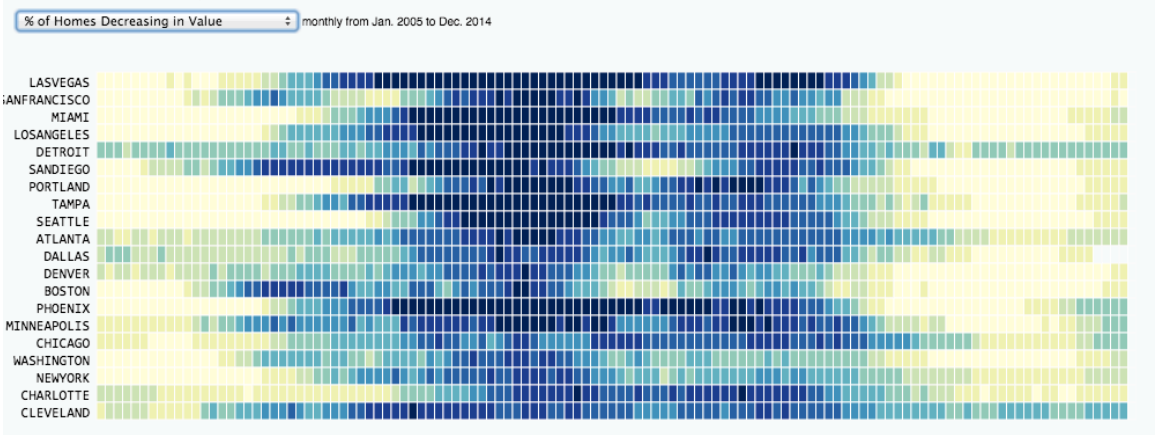
We also decided a more comprehensive view of crisis and its aftermath should be examined between all cities. We first considered a timeline type of visualization that would be repeated for each city:



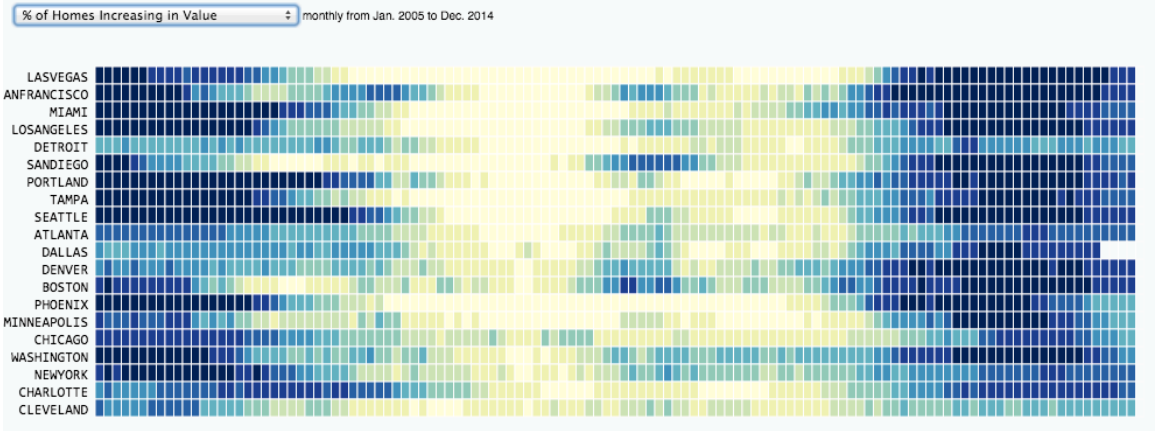
However, we felt using the area to represent the magnitude of the decrease or increase of a certain metric wouldn't be fitting since areas are hard to judge. In retrospect, we had to make a compromise with the map and using area circles to represent magnitude, since we felt we wanted to focus on cities for our story and cities do not allow for chloropleths due to their small size (a chlorpleth map would be more appropriate for areas as large, as say, states).

Without the need for such a compromise we opted for a visualization that incorporated graduated colors of the same gradient to represent the percentage of Homes decreasing in value, the percentage of Homes increasing in value, the decrease in the average price of a 2 bedroom home, foreclosures per 10,000 homes, and the median rent price over a time period that spans both the crisis and its aftermath (Jan. 2005 to Dec. 2014).

The results were shocking. The mortgage crisis of 2007-2009 significant decreased home values all across the United States, from empirically price-safe cities to exacerbating cities whose home prices were already on decline.



Most places recovered, however in particular places recovery has been slow, such as in Detroit.



Unfortunately there was not enough rent data available to see what happened to rent prices before and during the crisis.

Finally, we wanted to answer the question: is the rent too damn high since the mortgage crisis?

First, we had to determine some measure for “too high”. We settled on a relative measure against monthly mortgage payments. If the average monthly rent was greater than the mortgage payment for an average home in a city then the rent is too high (ie. Its probably better to buy a house than the rent one). We assumed a 2.5% interest rate over a period of 30 years.

We also wanted to show users how much the difference was between the average monthly rent and the average mortgage payment. Therefore we plotted the average monthly rent and if rent was higher than mortgage payment then we marked the difference as red. Likewise, if the mortgage payment was higher than the rent we marked the difference as green. Therefore, user can see both the increase or

decrease of the average price of rent in a city as well as its relative price to monthly mortgage payments.

The results show the rent is too damn high in many of our 20 city index in the post-crisis period. Our visualization shows Los Angeles, San Diego, Portland, Seattle, and Denver as places where one should rent.