CS171 Project Process Book

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 many of our households. 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?

Primary 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?
- Which features tend to affect price most heavily?
- If we made a histogram on the house prices, what kind of distribution would it follow? How heavily is it skewed, if at all? Does this vary by region?

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.

Data Processing

It took a decent amount of time to create the JSON file containing all the data. We created a script that scraped all the data from the various CSV files and combined them into one JSON file called 'data.json'.
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.
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"
         },…]
}
```


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"
 }
]
```


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?

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.

Zooming in could also give a greater degree of interactivity; a simple example of this is showed in http://bl.ocks.org/mbostock/2206590.

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.

Creating the Map

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).

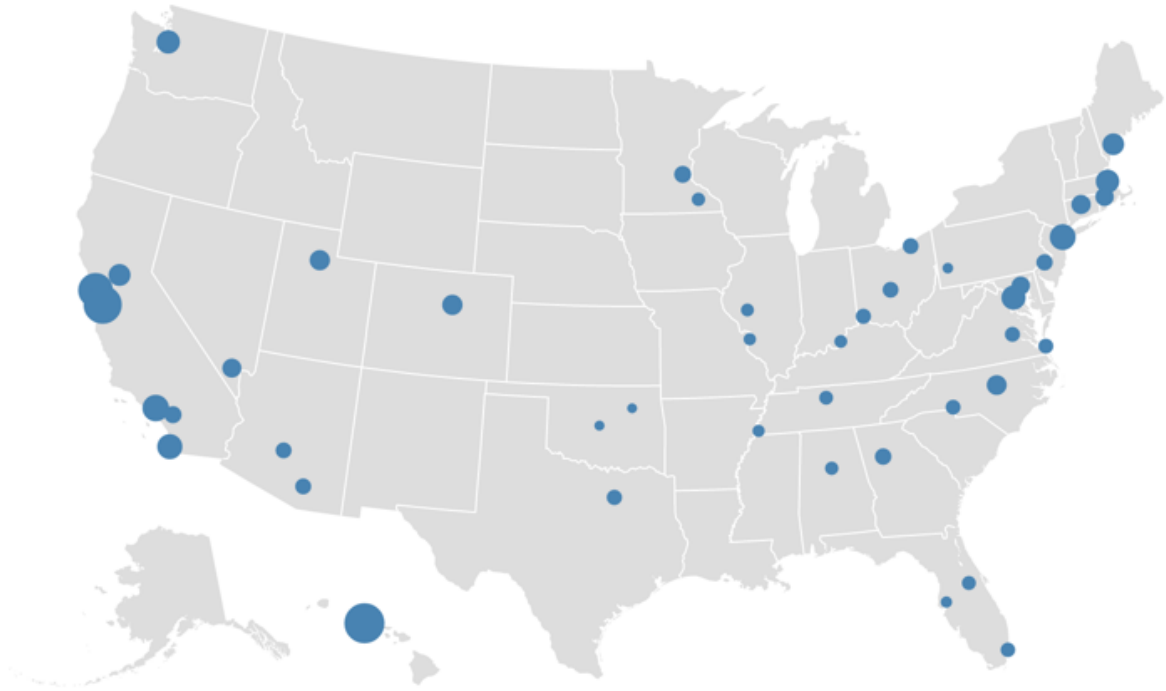
In terms of the interactive features on the map, there were several ideas we considered. In addition, there were different levels of interaction to consider.

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 scatterplot. Perhaps, rather than showing the change in the area's housing prices over time, we can have a bar chart to show how the different subclasses of housing prices differ against each other. Finally, we would also like to be able to zoom into specific places on a map, perhaps upon holding onto the map. These are definitely interactions that we will display for our project.

In regards to how housing prices should be scaled on the map, we currently are utilizing nodes that are present on each city. Another option, however, we are considering is using chloropleth maps. A realistic method of using choropleth maps would be to perhaps average all of the housing prices within a state, and then use the choropleth visualization to represent which states/areas of the United States have the highest/lowest housing prices. This would most likely be triggered as another mode, since the node visualization method is most likely still valuable.

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.

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.

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 scatterplot 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.

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?

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).