Tzu-Chieh Kurt Hong
PhD Candidate
School of Architecture
Georgia Institute of Technology
khong@gatech.edu
723 Cherry St NW, Atlanta, GA 30332

10/18/2017
Introduction

An overview of Zillow data set and its purpose:

This data set is called ZHVI (Zillow Home Value Index) which is a historic “estimated” value table of all median homes (not most expensive ones, not cheapest ones). It lists out the estimated value index of real estate within different geographic regions (city, county, etc) in US, and the data set is basically from 1996 April to current time (2017 July). The unit of ZHVI is US dollar and each value index is derived with a method called ZHVI Methodology which calculates the value not only for those homes sold but also for those home not sold. This value index is generated within seven different geographic levels: neighborhood, ZIP code, city, congressional district, county, metropolitan area, state and the nation. Also, Zillow provides users to download the data with different categories such as one-bedroom, or 2-bedroom. According to the information from Zillow, the purpose of this data set is to be a reference for real estate market. Furthermore, there are forecast data available in the website of Zillow Research. Those forecast data might provide investors a reference estimate the home value in the future.

Zillow website
https://www.zillow.com/
The data set “ZHVI” (Zillow Home Value Index) is provided by Zillow company since 2006, and the file format is .csv file. The data sets they provide have different geographical levels such as nation, state, city, metro, region and so forth. Also, they provide different data sets for different types of homes such as 1-bedroom home, 2-bedroom home, single family home, etc. In the data set, the first column is “Region Name” which is defined by Zillow, each region is derived by subdividing the entire country with certain zoning rules defined by Zillow as well. The second column to the fifth column are the geographical levels. By sorting the name of these columns, user can retrieve different subsets with the same geographical level. For example, if user wants the data from Atlanta only, he/she can simply sort the entire data set with the city column. The sixth column to the last column are the ZHVI (in US dollar) calculated by Zillow with an algorithm.

In the data set, Zillow lists out all the ZHVI data from April 1996 to current time (July 2017) and keeps updating them. The algorithm to calculate the ZHVI will be described in the later sections.

Dataset from Zillow

Source: https://www.zillow.com/research/zhvi-methodology-6032/

Zillow Data website https://www.zillow.com/research/data/#median-home-value

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>6181</td>
<td>New York</td>
<td>NY</td>
<td>New York</td>
<td>1</td>
<td>127800</td>
<td>127400</td>
<td>126900</td>
<td>126300</td>
<td>126200</td>
<td>126400</td>
</tr>
<tr>
<td>12447</td>
<td>Los Angeles</td>
<td>CA</td>
<td>Los Angeles</td>
<td>2</td>
<td>133100</td>
<td>133300</td>
<td>133200</td>
<td>133200</td>
<td>133000</td>
<td>133000</td>
</tr>
<tr>
<td>17426</td>
<td>Chicago</td>
<td>IL</td>
<td>Chicago</td>
<td>3</td>
<td>50800</td>
<td>50100</td>
<td>49400</td>
<td>49000</td>
<td>48900</td>
<td>48800</td>
</tr>
<tr>
<td>13271</td>
<td>Philadelphia</td>
<td>PA</td>
<td>Philadelphia</td>
<td>4</td>
<td>66600</td>
<td>66900</td>
<td>67200</td>
<td>67400</td>
<td>67600</td>
<td>67900</td>
</tr>
<tr>
<td>40326</td>
<td>Phoenix</td>
<td>AZ</td>
<td>Phoenix</td>
<td>5</td>
<td>89900</td>
<td>90000</td>
<td>89800</td>
<td>89400</td>
<td>89500</td>
<td>89600</td>
</tr>
<tr>
<td>18959</td>
<td>Las Vegas</td>
<td>NV</td>
<td>Las Vegas</td>
<td>6</td>
<td>124700</td>
<td>124800</td>
<td>125200</td>
<td>125300</td>
<td>124900</td>
<td>124500</td>
</tr>
</tbody>
</table>
For Zillow Data interview, the interviewee is Yongsung Lee, who is a senior PhD candidate in School of City and Regional Planning. Yongsung Lee is a researcher focusing on the relationship between urban form and travel behavior, and its implications to energy consumption and greenhouse gas emission. Recently, Yongsung is using the data from Zillow to dig more into the research. By using Zillow data, Yongsung is trying to find the correlations among real estate, energy consumption and travel behavior.

In the interview, Yongsung described himself as a user of Zillow dataset. Most of researchers in his lab have some experience on Zillow data, however, there is a researcher who is specifically responsible for cleaning all the data. Therefore, he is not the person who is dealing the data from scratch, his task is to analyze the data. The relationship between Yongsung and the data is not direct, so he called himself as a second-handed data user. But, he basically works with his colleagues as a team and more or less participate the clean process.

The main analysis in his research is to find out the pattern of the price transition with a regression model. By using Zillow data, he approximately can plot a home value model with some factors such as built year, distance to bus/train station, size, number of rooms, number of bathrooms, population of the area, etc.

In the interview, he suggested people use the actual price transition data provided by Zillow, because the actual price transition data are the real records collected by people which are very hard to get. Zillow provides two types of data, one is actual price transition data and the other is estimated price transition data. The estimated price data are calculated by an algorithm, which is affected by algorithm designer, but the actual price data are more objective.

When I asked him if he found any errors in the dataset, he mentioned that Zillow data do have some strange numbers such as 300 year-old house which is impossible and some houses cannot be found on Google map. So they either remove the unreasonable data or use Google map to fix the data. And when I ask him if Zillow obscures some important information, his answer is the information of buyers and sellers. He described that buyers’ and sellers’ information is actually important when they are modeling the price transition. For example, the young buyers’ house preferences are very different from the preferences of older buyers. Also, sellers’ information is a factor impacting the price transition, for example, a real estate investor’s’ preferences are different from a general person who wants to sell his/her house.

When I ask Yongsung if the data also collected in other places for other purposes, he answered that home price data are commonly used in tax calculation. Therefore, governments have been collecting this kind of data for a long time, but typically the data are not available for general public. Also, the data are usually not updated. However, Zillow, as a private company, is willing to share the data with general public, maybe for commercial benefits, maybe for advertisement profit, maybe for other purposes. But, as an outsider, he doesn’t know much about why Zillow open the data for us.
**Introduction**

This visualization practice is based on the data from Zillow. In this assignment, the geography level of the data is neighborhood, and the region is Atlanta Fulton county. The data are the price indices calculated by Zillow for all the neighborhoods in Atlanta (ex: midtown, downtown, Buckhead, etc). In the visualization, each circle is representing a neighborhood, and the radius represents the price index (bigger circle means higher price), opacity represents the population of the neighborhood (deeper color means more population). On the left, there is a time slider for users to change the data from 1996 to 2017.

**Visualization Tool | Processing**

Data Description | From Zillow ([https://www.zillow.com/research/data/](https://www.zillow.com/research/data/))

Data Type: ZVHI (price index for buying a home)

Geography Level: Neighborhood level

Home Type: 2-bedroom homes (median homes)

Region: Atlanta

Time: From 1996 to 2017

**Legend**

X-Axis: Normalizes Latitude

Y-Axis: Normalized Longitude

Each Circle: Represents a neighborhood

Radius: Price index of 2-bedroom apartment

Opacity: Population (deeper color represents more population)

Timeline: Select data from 1996 to 2017

*Data from 2015*
7 Steps of Visualization

**Acquire:**

The data is acquired from Zillow Data (https://www.zillow.com/research/data/) called ZHVI, Zillow Home Value Index. The data are available for downloading with different geographical levels such as state level, metro level, neighborhood level, etc. The data acquired are neighborhood level.

**Parse:**

The data format on Zillow Data is .xlsx file, so I use Excel to change the format to .csv file for Processing. The data are originally sorted based on the population rankings which can not let me focus on the data of Atlanta only, so I sorted the data based on metro name to extract out the data of Atlanta.

**Filter:**

First of all, I only use the data of Atlanta instead of the entire national data. Secondly, since the data are converged to the range of Atlanta, there are some common columns which can be removed such as “state name” and “metro name” (all the data are in “Georgia state” and “Atlanta”). Thirdly, the data are monthly recorded from 1996 to 2017 and the amount of data is too much for this practice now, so I selected the data in July for each year to get a smaller dataset.

**Mine:**

Before using Processing, I put the data into Raw graph to understand the data set. Also, I download another .csv file which is a table for Zip Code and position (longitude and latitude) transformation to furtherly locate the data on the relative positions.

**Represent:**

The representation I choosed is circle diagram to represent the dataset. Each circle represents the data of a neighborhood including population, ZHVI (Zillow Home Value Index), relative geographical position. I think it will be good to see the “price” with relative location, for example, some circles in southwestern area are not increasing the size (the price is not increasing) after 2011. There seems a boundary between these circles and the circles in other area.

**Refine:**

In order to make these circles more understandable, I normalize the longitude and latitude to get relative position for each circle. Without normalization, the circles are distributed in a wide range and hard to observe the relationships among these circles. I also normalize the ZHVI and population to make color and size more understandable.

**Interact:**

Since the data are historical records ranging from 1996 to 2017, I used time slider for users to change the sub-dataset. By positioning cursor on the time bars, users can see data from different period.
To derive the estimation of the price, Zillow designs an algorithm to calculate Zestimates and ZHVIs. In the first step, Zillow collects the price records of actual sold homes. Second, the collected data are subdivided into “micro-regions” according to the geographical/physical features such as location, environmental condition, physical features, and so forth. The third step is to look at the last price transaction of each home and find out the relationship between price transaction and selected features. The selected features can be the “distance to train station”, “the ranking of the closest school”, “the distance to the closest park”, “built year”, “size”, etc. Zillow uses machine learning method to figure out the effective factor of each selected feature and plot a model for estimating the value of a home by inputting new selected features. This estimated price is called Zestimate and Zillow collects these Zestimates to calculate ZHVIs. For the fourth step, once Zillow collects enough Zestimates, Zillow will derive the Zestimate of the “median home” by dividing all the homes into three tiers: expensive, middle and cheap. Zillow discards the expensive homes and cheap homes to filter out the extreme cases. The fifth step is to fix the systematic error which is the error from the model. When Zillow collect enough historic data and the model will be adjusted to fit the actual data again to eliminate the gap between estimated value and actual value. The sixth step is to use Henderson Filter (mathematical filter) to filter out the noises and unreasonable cases and get a smooth estimation curve for each micro region. The seventh step is to equalize the seasonal bias which means people prefer to buy homes or not to buy homes in certain seasons, and this will affect the values of homes and need to be equalized. The last step is to refine the entire model again by increasing the number of the Zestimates or input more features.


**Introduction:**

The interface of accessing Zillow data is through Zillow website designed by Zillow. The target users of this website are the people who are looking for homes to buy/sell, rent/lease or invest. This website is also for the real estate companies who want to post their properties on the website. The Zillow website is made with a Bing Map where users can “see” the data locating in specific positions. By searching zip code, address, city or any other geographic features, users can quickly focus on the results they are looking for. Also, the website provide seller to attach pictures, descriptions and information of homes to give users more details.

**Analytical context:**

In Zillow website, the data are placed in a digital map which is provided by Bing Map. Besides placing ZHVI (Zillow Home Value Index) on the map, Zillow also show the pictures and descriptions of the homes which are on the market. These pictures and the descriptions can be viewed as a detailed view of the digital map, users can thus link the price and the real house. I think the most valuable actions that Zillow does including exterior pictures and interior pictures, these pictures visualize the price in some level. In Zillow website, users can search a specific area with address, zip code, state or other geographic information, after searching, Zillow shows two different results, red dots and blue dots. Red dots are the homes in the market now, and the blue ones are the potential objects which may be put in the market soon.

**Discursive context:**

The home price data were typically unavailable to public, however, Zillow collected these data and made them accessible. It is surprising that Zillow doesn’t have ad banners on the webpage, the website stay clean. It means Zillow’s business model is not selling ad banners, instead, they sell the “information boxes” which look like the introductions of objects, but actually they are alternative ads. For the real estate agents, they can buy these boxes to make their objects more visible and promote their companies. The more you pay, the more oftenly your objects will show up in Zillow. The business model is similar to Google Search. In other words, if Zillow has more people visiting their website, the information boxes can have higher value. So, Zillow’s goal is to keep people visiting the website everyday (even every hour) to increase the amount of visitings. The way they achieve this goal is to update the ZHVI very frequently, and people will be curious about the price variations of the homes they want to buy or sell, just like the stock market.

**Operational context: Algorithmic element**

To derive the estimation of the price, Zillow designs an algorithm to calculate Zestimates and ZHVIs. In the first step, Zillow collects the price records of actual sold homes. Second, the collected data are subdivided into “micro-regions” according to the geographical/physical features such as location, environmental condition, physical features, and so forth. The third step is to look at the last price transaction of each home and find out the relationship between price transaction and selected features. The selected features can be the “distance to train station”, “the ranking of the closest school”, “the distance to the closest park”, “built year”, “size”, etc. Zillow uses machine learning method to figure out the effective factor of each selected feature and plot a model for estimating the value of a home by inputting new selected features. This estimated price is called Zestimate and Zillow collects these Zestimate to calculate ZHVIs.

**Ethical issue:**

Shoud home value be public information? What if I want to sell my home at a higher price because I am the one who really understand the “value” of my home. However, Zillow suggests us an “estimated value” they calculate despite some important facts. For example, a house designed by Frank Lloyd Wright will be the same as the house next to it. The spatial quality, inspiration, aesthetics given by the architect are not included in the value system. This example shows the ridiculous situation of this system and the estimated price truly offend the owner of the home. To be brief, what Zillow does is tagging the prices on our homes without asking us.
Sources

Websites


References

**Tzu-Chieh Kurt Hong** is a current PhD student in School of Architecture at Georgia Institute of Technology focusing on Design Computation. The current research is on implementation of shape grammar interpreter, especially on shape detection. Kurt received a MS (Master of Science in Architecture and Design) from University of Michigan focusing on digital fabrication and computational design. Kurt’s undergrad is in electronics engineering and holds a master’s degree in digital circuit design from National Chiao Tung University, mainly working on high speed transmission chip design such as HDMI, and USB 3.0 from 2008 to 2012. He received a MArch degree from National Chiao Tung University in 2015.