Zillow Housing Data

Team Members:
Iris Lin
Miranda Zhong
Xing Chen

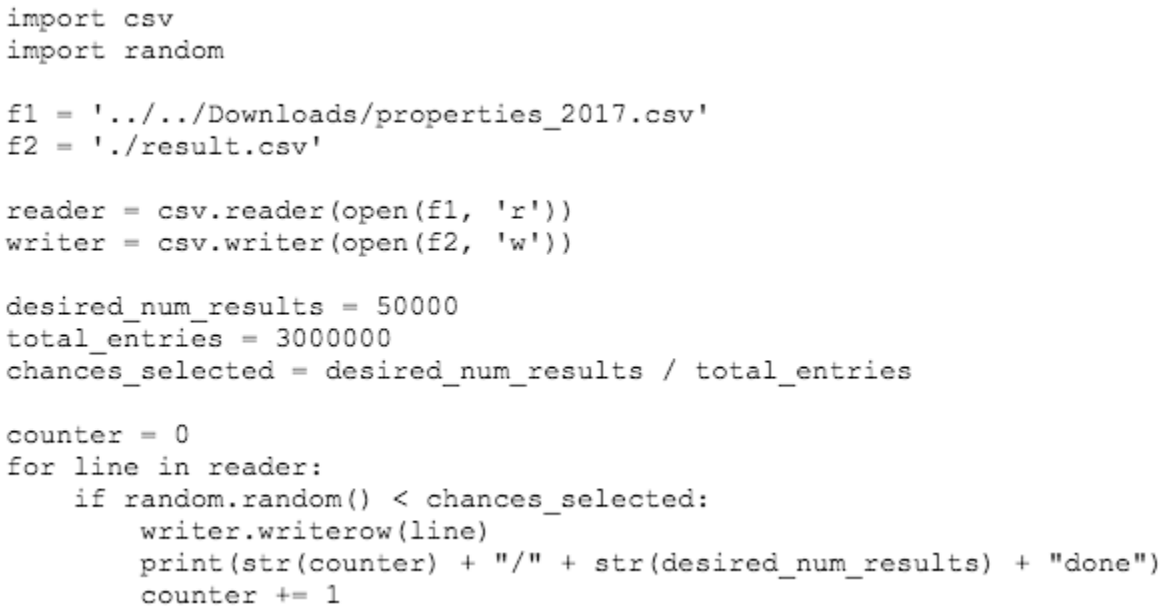
I. Goal of the Report

This report provides a data visualization for house buyers in Los Angeles to help them understand the **characteristics of houses**, the **price seasonality**, and the **related tax issues**. Our approaches to this visualization include correlation matrix, map view, and line chart. Results of data visualization showed that some characteristics are related to others and the year built is a vital characteristic that can guide the buyer in purchasing behavior. Moreover, a seasonal price pattern is displayed in the report, and the relationship between tax and location is revealed.

II. Exploration of the Data

Python: Data Cleansing and Preparation:

In the report, we extracted random sample data from the original data - properties_2017 (Zillow Housing Data) and utilized Kepler and Tableau to create data visualizations. Firstly, We used Python to build up our sample dataset from dataset properties_2017. Meanwhile, in order to create maps on Kepler, we created two new columns (trans_lat and trans_long) that are the results of original latitude and longitude divided by 1 million. The following is our code:

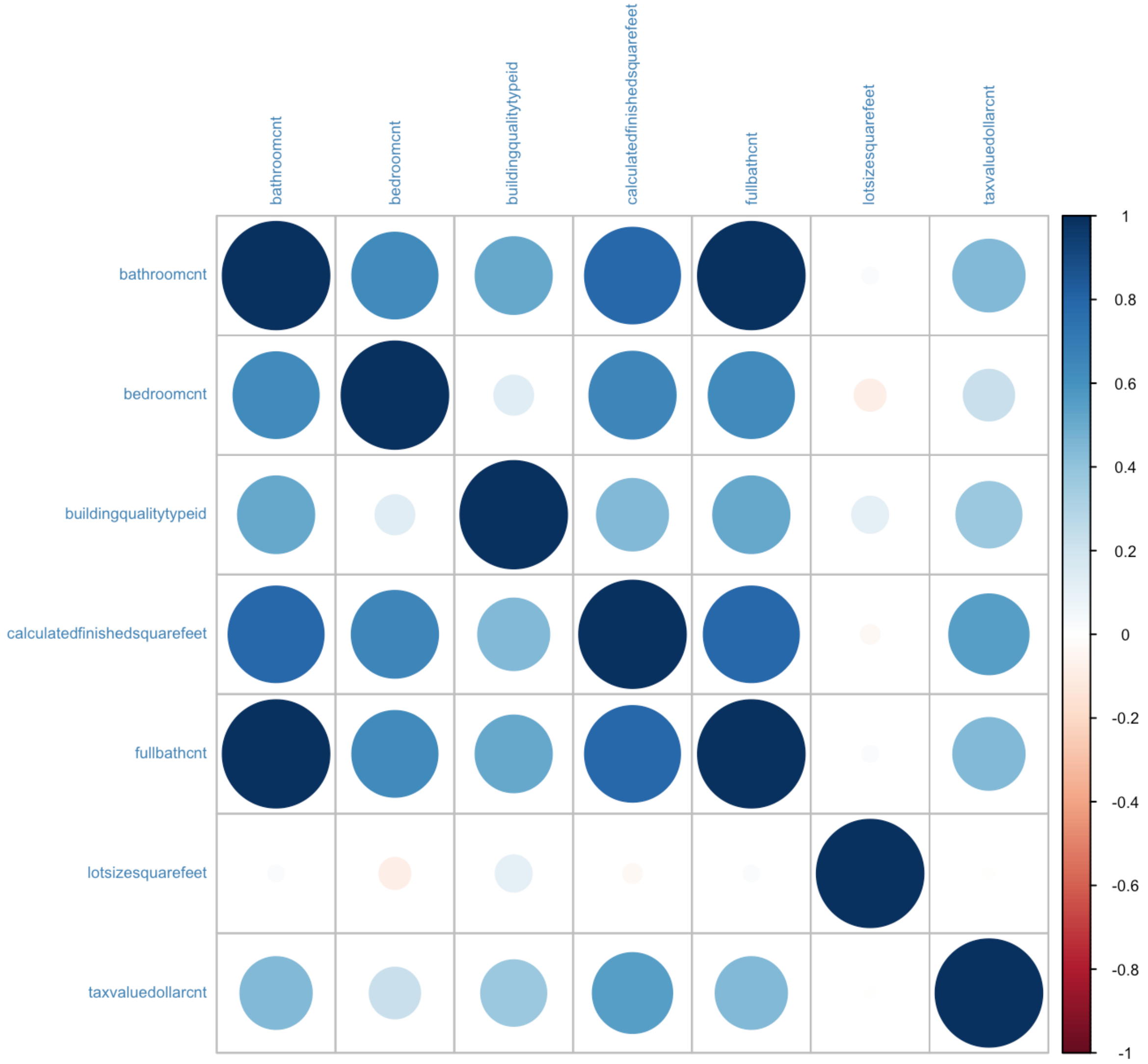


```
import csv
import random

f1 = '../../Downloads/properties_2017.csv'
f2 = './result.csv'

reader = csv.reader(open(f1, 'r'))
writer = csv.writer(open(f2, 'w'))

desired_num_results = 50000
total_entries = 3000000
chances_selected = desired_num_results / total_entries

counter = 0
for line in reader:
    if random.random() < chances_selected:
        writer.writerow(line)
        print(str(counter) + "/" + str(desired_num_results) + "done")
        counter += 1
```

We did not randomly sample data from datasets such as train2016 and train2017. The visualization of these two datasets contains the whole data points.

R: What Characteristics of Houses Are Correlated?

In the correlation matrix below, we tried to show the correlation among several variables. Blue indicates positive correlation and orange indicates a negative correlation. The size of the dots represents the magnitude of the strengths. The variable BuildingqualityTypeID represents the overall assessment of the house. The higher the number, the worse the quality of the building. From the chart, we can see that bedroom number, bathroom number, total square feet of the house, total tax value are positively correlated with each other. It is reasonable since people generally are paying more money for an extra room or a bigger house. However, the size of the lot is irrelevant to other variables. Another interesting thing is that the quality of the house negatively correlates with the size of the entire house and total counts of bedrooms or bathrooms.

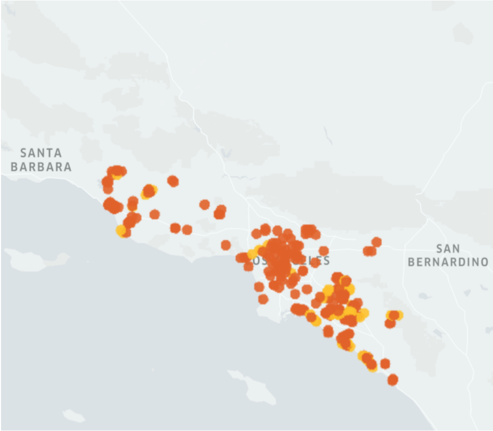


It seems that house buyers cannot achieve quality and house size at the same time from this chart. Another tradeoff is between the price (implied by the tax value) and house size; yet quality, house size, and price are the most important elements people consider when buying a house. The bottom line here is that, to understand exactly what they are paying for, people should do more researches on what characteristics correlates with price and quality and the strength of the correlation. Also, by asking the question "how much I am willing to pay for this characteristic', people could prioritize on the factors of a house.

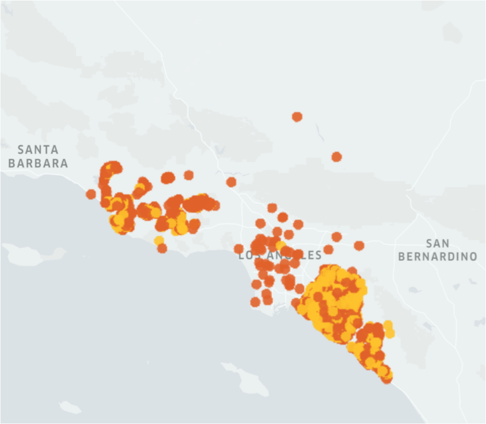
Tableau: How Year Built Related to House Characteristics?

According to the dataset we have, we visualized the ***number of stories, building quality*** and ***calculated finished square feet*** of houses, which were built over the past century. We created one video for each of catachrestic, and three screenshots from each video visualization are as follow.

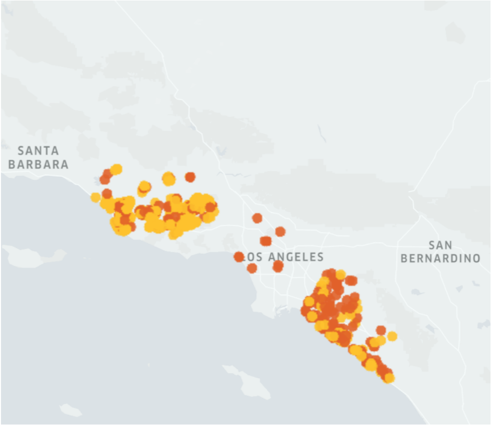
1. Number of Stories
Video Link: https://youtu.be/uO28hMIGk2E



1917-1942



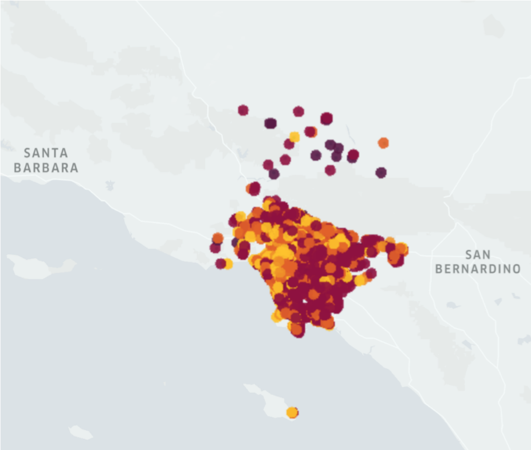
1950-1975



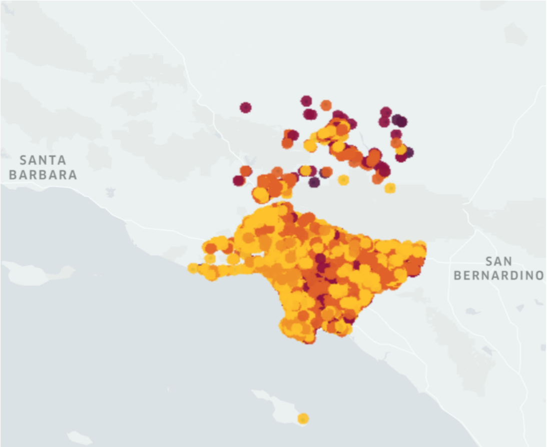
1990-2015

For the number of building stories, we developed a visualization to find the relationship between built year and number of building stories. In the viewing, orange means buildings with one story and yellow means buildings with two stories. As we can see above or in the video, one story buildings were more popular at the beginning of the 20th century, but two stories building started to prevail after that. Especially for buildings in suburban of LA, two stories building exceeded the one-story ones. Moreover, we can see that with the time going, there are fewer buildings built in the center of LA and there are more buildings built in the suburban of the city. For a house buyer, he/she can refer to this visualization to guide their searching behavior. For example, if a person plans to buy a two stories house, he/she should better search in the suburban area and look for houses built after the middle of the 20th century.

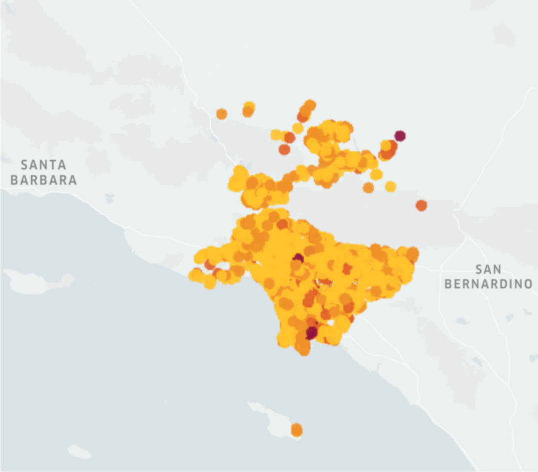
2. Building Quality Type
Video Link: https://youtu.be/pL3_gQf0PRs



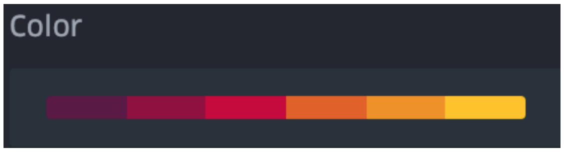
1917-1942



1950-1975

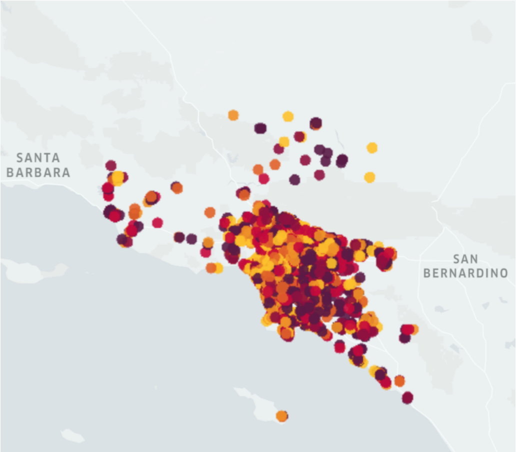


1988-2013

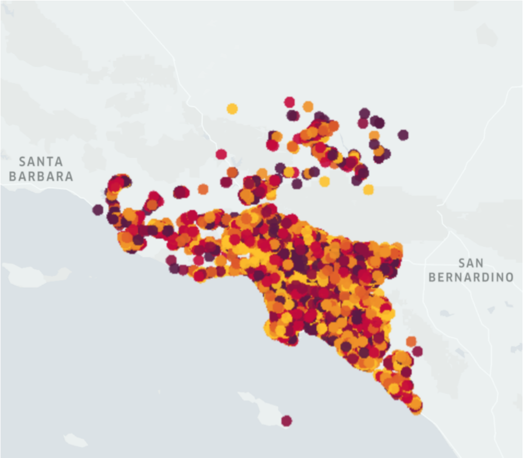


In the visualization of Building Quality Type, the color of points represents the quality of the building: the brighter the color, the higher the quality of the construction. As we can see from the video and graphs above, the overall color becomes brighter over time, which means the quality of the building is generally increasing over the past century. The change is consistent in both the center of LA and the east north of the city. Hence, a house buyer may take this insight into account when he/she would like to find a building with the quality he/she needs.

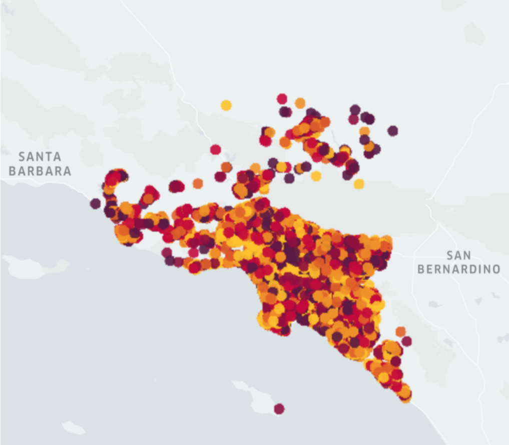
3. Calculated Finished Square feet
Video Link: https://youtu.be/UDf8NlV-NK8



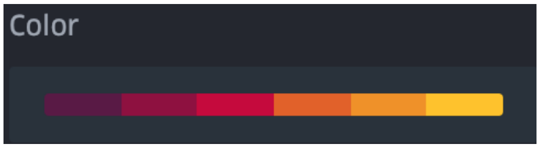
1917-1942



1950-1975



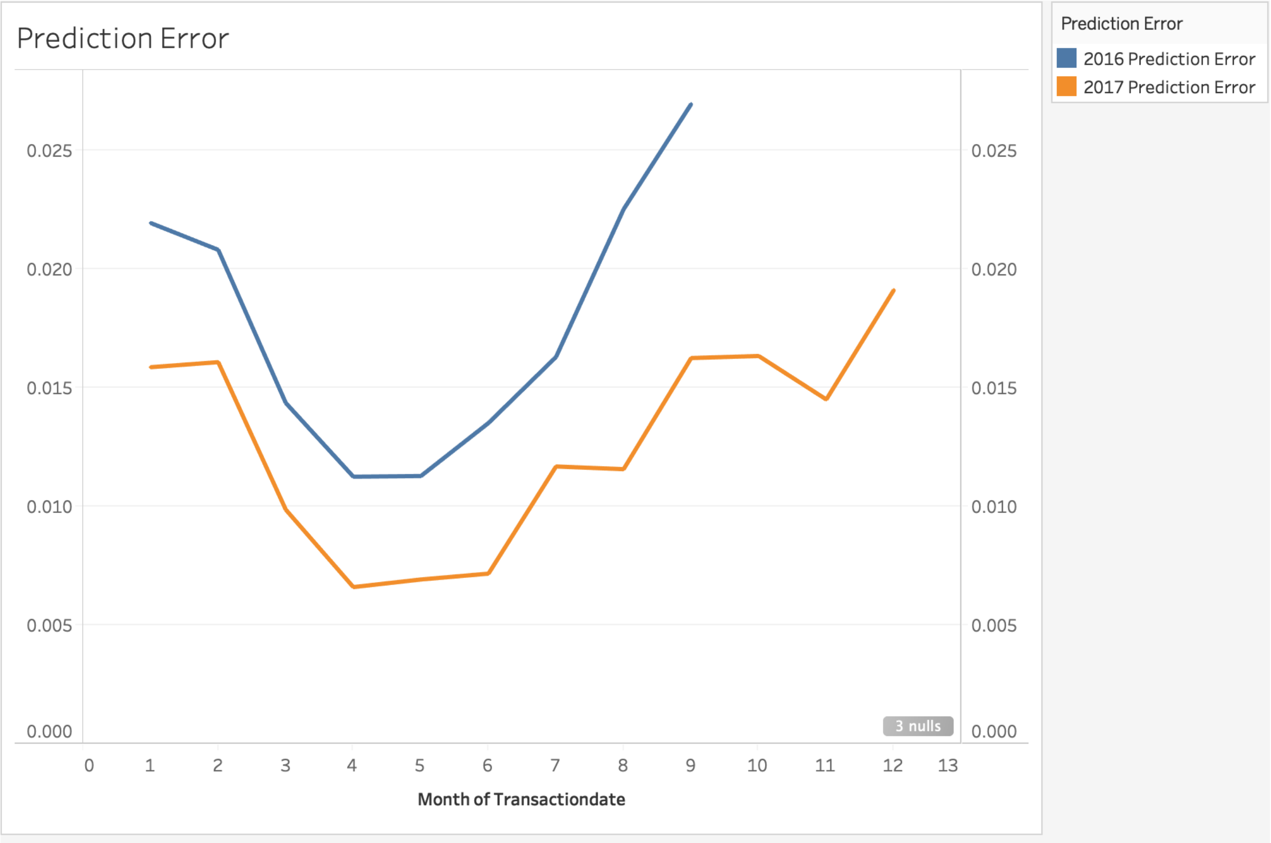
1990-2015



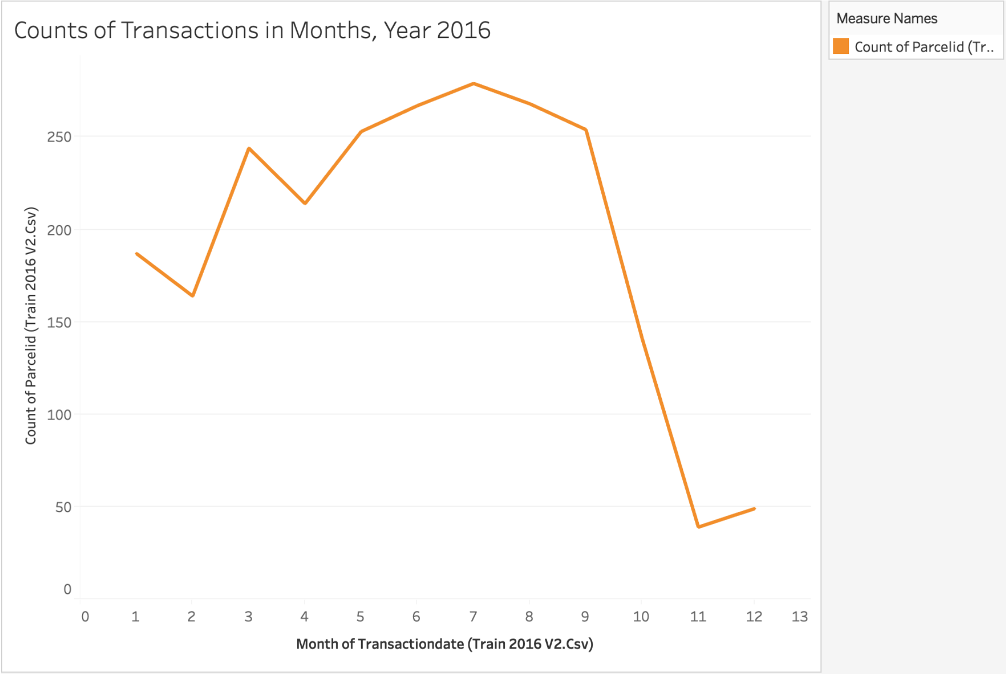
In the visualization about Finished Square feet, the color represents the size of the house: the brighter the color, the larger the house is. Overall, building sizes are varied even though they were built in the same period. However, there is a trend that the overall size of buildings increases over time and the newer buildings usually have larger calculated Finished Square feet. A house buyer can refer this to find the house with the appropriate size for them.

Tableau: Seasonality Matters in Housing Price

This chart shows the average prediction error in the year 2016 and year 2017 data. We can see that across two years, the prediction error decreases in summer seasons but is prominent in winter. For potential customers, it is essential to identify seasonality trends in the housing market. One research indicates that for families, they are reluctant to move in holidays especially the weather is unpredictable. For families with children, they are more likely to move after school days (Boykin, 2017). Besides, a less competitive house market in winter seasons results in a promising discount of houses and reducing the price of houses.



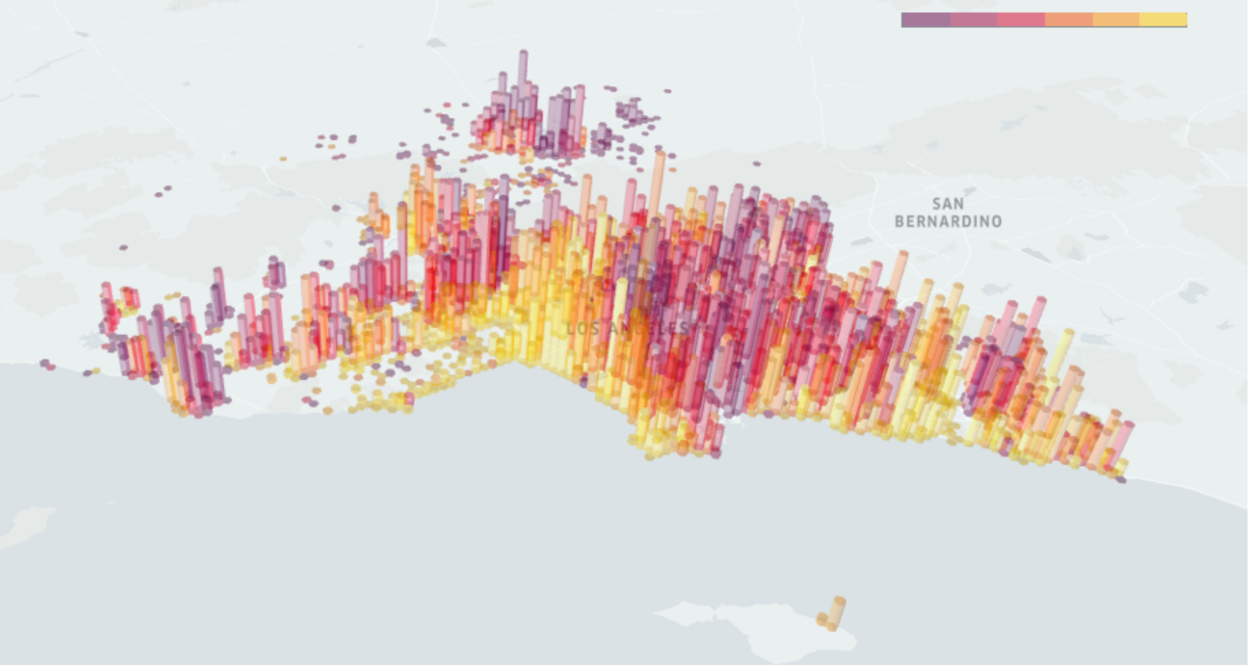
To validate this research, we explored the dataset and identified a trend. In the chart below, we can see the trend of house transactions in the Los Angeles area. It is indicating that the summer break is a busy moving period and a bull market of housing.

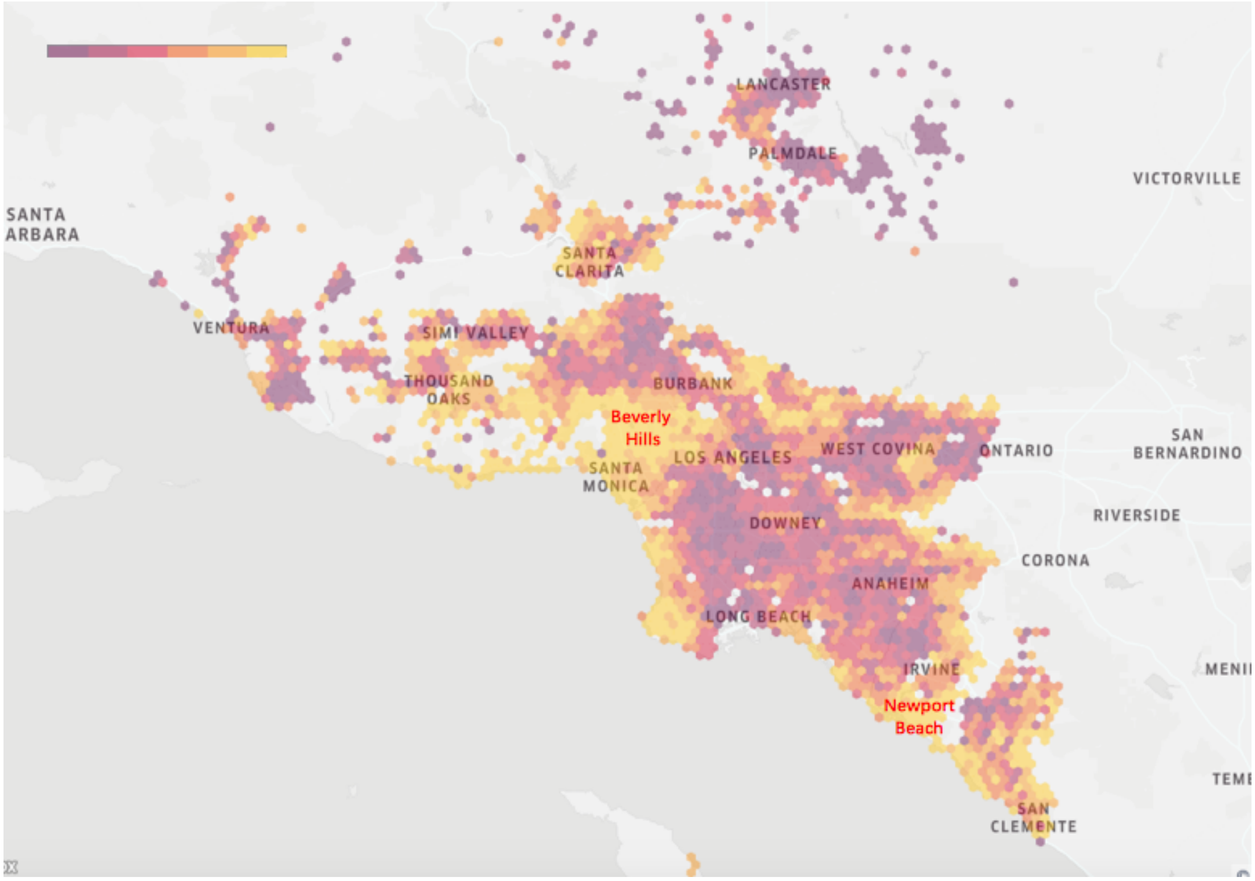


It could be also possible that purchase decisions in winter seasons are more irrational than in summer seasons. The population which purchases houses during summer seasons take more into account the characteristic and quality of house itself than the population which purchases houses during winter seasons. However, we still highly recommend house buyers taking the advantage of winter seasons to purchase houses. Go against your 'reluctance', you will have a better deal.

Kepler.gl: The Relationship between Tax and Location

In these two graphs, each hexagon means the average value of tax amount in that location, and with the amount of tax increased, the color change from purple to light yellow. Meanwhile, in the first graph, the heights of hexagonal prisms are based on the number of properties in this location. According to the first graph, we can know that a large proportion of properties are located in the central part of LA, and with the extending out from the center, the amount of housing is decreasing. Based on the second graph, we can clearly figure out that with extending out from the downtown of LA, the tax amount is gradually increased. Especially in the Beverly Hills part, it shows the light yellow, which means the highest amount of the tax. In Los Angeles County, the average residents pay for these combined property taxes is 1.16%, or $11.60 for every $1,000 of assessed value (Matthew 2014). Hence, the Beverly Hills properties are the highest in Los Angelos County. As is known to all, a lot of celebrities have their property that located in Beverly Hills, including Johnny Weissmuller, Nicolas Cage, and etc. And almost all of these properties are built up with the large floor area and fancy architectural decoration styles, which lead the price of these properties higher than others in LA.





Under this case, when housing buyer who does not have a lot of budgets, they should avoid some locations, including Beverly Hills, Santa Monica, and Newport Beach. And they can choose properties that locate at downtown of LA or outside of LA County (which with the purple color shown on the graphs) because of the large amount and low price. For purchasers who want the property nearby the beach with low price, they can choose parcels which locate at Long Beach, which price is lower than other beachside properties. For buyers who have enough budget and want to integrate into brownstone, Beverly Hills is the best choice for them.

III. Design Choice

In this report, we integrated geospatial analysis and time series animation to depict the trends in house market and take the location into account at the same time. Besides, in order to visualize the correlation between certain variables, we use correlation matrix. Correlation matrix with different colors to show the direction of correlation and size of dots to show the strength of correlation helps us clearly visualize a large number of relationships. However, we did consider parallel coordinate plot to depict the correlation. Yet, since the size of data is substantial, a large number of lines between columns make the plot messy. To depict the seasonality trend, we choose to apply a line chart. Because we only consider an aggregate level of data in this case, a line chart is more straightforward and there is no need to use Kepler.

Appendix:

Boykin, Ryan (2017). Seasons Impact Real Estate More Than You Think. Retrieved from https://www.investopedia.com/articles/investing/010717/seasons-impact-real-estate-more-you-think.asp.

Matthew Gaskill (2014). Los Angeles Property Tax: Which Cities Pay the Least and the Most? http://www.luxuryhomeslosangeles.com/blog/los-angeles-property-tax-which-cities-pay-the-least-and-the-most.html