Housing Prices in Los Angeles

Authors: Louis Banegas, Marc Calvillo, Christian, Momdjian, Joshua Ng, Adrian Soriano
Department of Information Systems, California State University Los Angeles
CIS 4560-02 Introduction to Big Data
lbanega@calstatela.edu, mcalvi14@calstatela.edu, cmomdji@calstatela.edu, jng32@calstatela.edu, asoria55@calstatela.edu

Abstract: The source for this project is a collection of property prices across the United States. The data contains the address of each property, the type of property it is (Residential, Mobile Home, Apartment, etc.), the year the house was built, and the measurement of the building and property. Stored at more than 50 GB, the dataset will be reduced to only include pricing within the state of California. Once reduced, the data will be uploaded to Hadoop and analyzed by creating tables and using extensive queries. The data retrieved will be used to create a visualization displaying and comparing average pricing and number of properties for sale throughout cities in the greater Los Angeles area.

1. Introduction

Our project uses HDFS, Hive, and PowerBI. Its purpose is to filter and visualize housing market data that has been sold in California. The data used in this project contains homes currently on the market at the time the dataset was downloaded

This specific dataset was chosen because it has been regularly updated and has a comprehensive collection of housing data from areas across the US. It allows us to compare today's housing market prices and explore the wide gaps in property value between counties in Los Angeles, our specific focus for this project. It also allows us to view and visualize the specific areas where property value is most affected and which areas have the most properties on the market.

2. Related Work

The data used for this project will place a sharp focus on Southern California, particularly the Los Angeles surrounding areas. In recent years, the housing market has faced a sharp decline in affordability for single-family homes. Statista, a globally recognized German company that focuses on analyzing consumer and market data, compiled a 2022 housing summary report comparing the median sales prices for single-family homes throughout the US in the fourth quarter of 2020 to the same prices one year later. The differences were substantial. One significant takeaway is that, despite showing all US cities, five CA cities appeared in the top six cities and remained in their spot one year later, telling us what may already be obvious: CA is one of the most expensive places to live.
Figure 1 shows these findings.

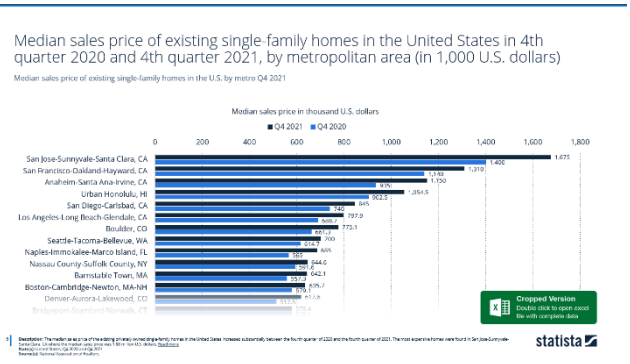


Figure 1

Taking a closer look at the data, Figure 2 shows the percentage change in median price for single-family homes in CA. Unremarkably, the bay area remains at the top with the highest increase in price comparatively, with Santa Cruz showing a 45.45% increase in sales prices in just one year. These figures represent the change from March of 2021 to March of 2022, while figure 1 examined the year prior. Both show that prices are continually rising with no change in sight.

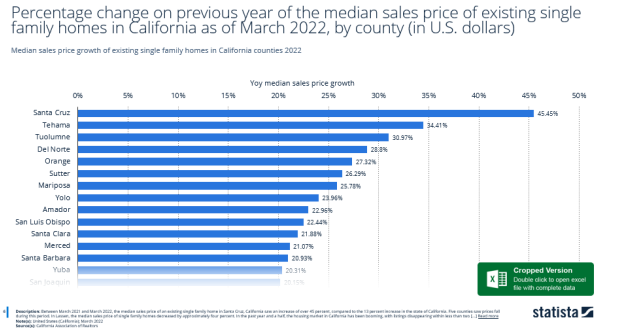
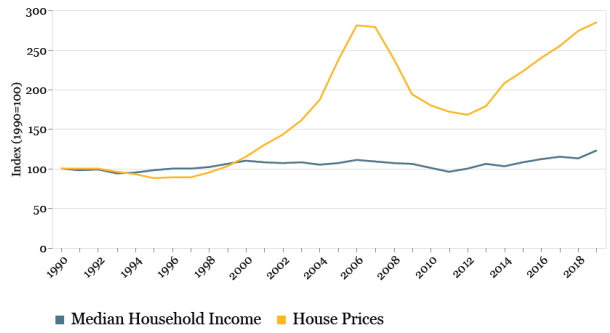


Figure 2

Many would argue that prices are as high as they are because CA also has higher wages than in other states. This is true. However, over the past two decades, home values have risen increasingly faster than household incomes. A study conducted by the Terner Center Report tells us that despite the negative effects of the great recession and foreclosure crises, house prices have increased 180 percent between the years 2000 and 2019. In comparison, the household median income in CA has only increased about 23 percent in the same 19 years [3]. Figure 3 details the growth in House Prices as well as the growth in Median Household Income. Prices took a notable dip after the recession of 2008 but began a steady rise around 2014. Median Household Income had a much less drastic change.

Figure 2. Growth in House Prices Compared to Median Household Incomes



Source: Federal Reserve Bank of St. Louis, Federal Reserve Economic Data (FRED), Economic Research Division, 1990-2020.
Note: All-Transactions House Price Index for California, Index Q1 1990=100, Quarterly, Not Seasonally Adjusted. Median household income inflation adjusted using 2019 CPI-U-RS adjusted dollars.

Figure 3

The country is currently in a state of recovery. The global pandemic took a heavy toll on the country's economy, and the repercussions are not truly over. As Jennifer Boehm, Associate Director for finance reports at Mintel stated at the end of 2020: "Unemployment is improving, but the economic ramifications will be felt long after the vaccine becomes widely available" [5]. Two years after this quote was said, it has turned out to be true. Inflation tends to be the hot topic of conversation nowadays and shows very little sign of improving. The housing market is no better. Lending has become quite an issue with some borrowers facing a real struggle trying to pay back their loans in these times of crisis. The data used for this project allowed us to see the changes in pricing over the past three years. In doing so, we saw some of the changes that came about from residual effects of the pandemic. Figure 4 shows a frozen still from the temporal visualization created through this project. The still is from the beginning of 2021 and shows a heat map of the average housing prices in each city. The entirety of the temporal map ranges from 2019-2021, but this is the most affected quarter in those years. Compared to other visualizations created for this project (figures 8, 9, 11), this quarter (first quarter of 2021) shows the lowest average prices in the range.

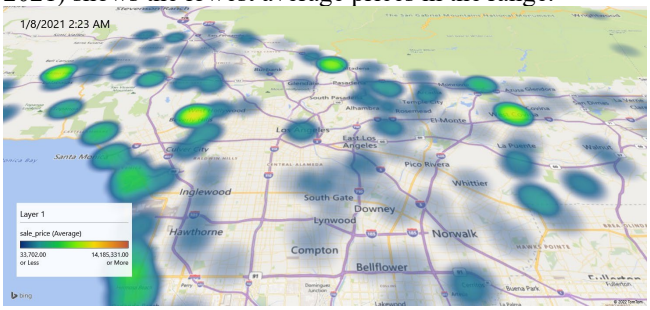


Figure 4

The volatile nature of the housing market is the foremost motivation for the analysis in this project. We wanted to examine the current state of housing in CA. More specifically, the state of housing around the LA area, the region with which we are all most familiar.

3. Preparing the Data

The data we used for this project was available through an open-source website called Dolthub. Dolthub is very similar to GitHub in that it allows users to upload entire databases to a repository and make it freely available for anyone to download the data. However, the entire dataset could not be downloaded using tools like *wget*. The data we needed had to be downloaded using the graphical user interface of the site. Even after downloading the raw data from Dolthub, the entire dataset's size was almost 50GB.

Since our team knew we only needed to analyze data from within California, we only needed to download a portion of the data. Luckily, the data would be downloaded in a way that sorts the data alphabetically by state, with California being near the beginning of the part file. The data ended up being just over 2GB in total and included all our California data along with some other data from other states.

Once we had our housing prices dataset downloaded, we wanted to make it easier to download the data without going through the manual process we went through using Dolthub initially. We decided to store the file in a Dropbox location so people who went through the tutorial could download the housing data using *wget*. Figure 5 shows a workflow diagram visualizing the work done in processing and visualizing the data.

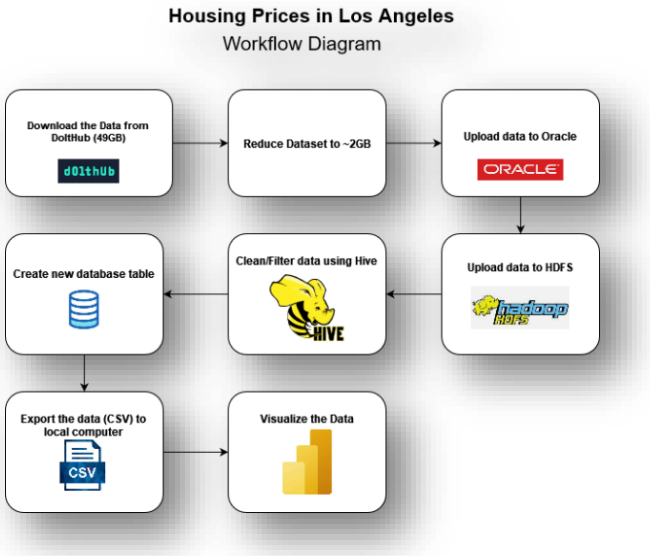


Figure 5

4. Using HDFS

Most of the filtering and cleaning of the data was done through HDFS and Hive. The Hadoop Distributed File System (HDFS) and Hive functions are major components of Hadoop, a software created for the purpose of storing and processing large data distributed across many commodity servers. Hadoop stores data using HDFS and uses MapReduce, a programming model, to process/query the data within the Hadoop framework. Additionally, Hadoop also uses Hive, an application that runs over the Hadoop framework, that provides an SQL-like interface for also processing/querying data. After uploading the entire 2GB dataset to HDFS, we noticed that many of the fields were NULL, so we needed to filter it down to only the fields we thought were necessary to gather useful information from the dataset. We also noticed that the sale_price field, which represents how much the property sold for, was extremely

low for several records based on historical U.S. housing costs in California. Some records even had a sale_price of zero. Due to this, we filtered the Hive query to only pull records with a sale_price of $100,000 or greater.

Another pattern we noticed is that the dataset included properties that are not necessarily related to housing for families or individuals. It included property data for vacant lots and commercial properties. To further filter the data down, we added a WHERE clause that searches for only properties with keywords that are closely related to residential properties. Below is the snippet we used to do this:

```
AND (property_type LIKE '%RESIDENCE%'
OR property_type LIKE '%CONDO%');
```

All the filtered data was set to a new Hive table based on the original dataset in its entirety and eventually exported to a new csv file with only the fields we needed for visualization.

The downloaded file, called *US housing prices*, was uploaded, and stored in Oracle, but before we could do anything, we needed to rename it. With it renamed, we could create a new directory in HDFS to hold the data. After the directory was created, we uploaded the .CSV file data into the HDFS "housing data" directory we just created and then loaded it into tables using Beeline. Using HDFS allowed us to store it properly and prevent space from being taken in the Oracle cluster.

5. Excel and Power BI

Microsoft Excel is a powerful spreadsheet data analytics tool for creating pivot tables, calculating large data sets, and displaying graphical visualizations. It adopts a spreadsheet model that can organize data into rows and columns and perform mathematical formulas and calculations quickly and easily. Microsoft Excel is ubiquitously used by small businesses and large corporations worldwide. It also can visualize data with charts, graphs, and a 3D map function tool under the Insert tab. We loaded the US Housing.CSV file and graphically visualized the data onto a 3D map, which provides a better understanding of the data. An alternative to Excel is Microsoft Power BI. Power BI is specifically designed to help businesses visualize small and large amounts of data. After using both Excel and Power BI, our group came to a consensus that Power BI would help us achieve the project's goals more effectively. We will cover our experiences with Microsoft Excel and Power BI and list the advantages and disadvantages.

Two of the advantages of Excel are displaying and labeling the columns of data; it was made simple, and the data is well organized and structured, which makes it easy to interpret. The 3D map function tool was also relatively straightforward to use and configure. Creating charts and graphs is made simple by the powerful tools built in and importing/exporting spreadsheets is uncomplicated. However, Excel struggled to handle file sizes larger than 2 GBs and 1.5 million rows of records.

On the other hand, Power BI is a data visualization software that mainly focuses on helping businesses interpret small or large amounts of data. Power BI is capable of handling larger datasets compared to Excel. In addition, we preferred Power BI's visualization tools as they could cater to our project's end goal with fewer complications. For example, the filtering tool offers a detailed approach to rearranging our data into smaller, more specific sets. Furthermore, Power BI has a wide variety of visualizations compared to Excel, and there is also a marketplace where we can download other charts and graphs.

6. Visualization

Using HDFS, the data needed for this project will need to be queried from the tables previously created. We will be using data that only pertains to CA and specific rows from the tables created. Creating a query that defines these parameters and filters out unusable data, we create a CSV (Comma Separated Values) file to move this specific data to another platform that will help us further analyze, Microsoft PowerBI. The created CSV file is then downloaded from Hadoop to the Linux system where it can be sent to the local computer and opened in the installed PowerBI application.

6.1 Average Property Price per City in LA County

Using PowerBI, the CSV file is opened, and the data is loaded to a table that is structured in the manner specified by the file (following header naming conventions and placing data in positions based on comma separation). In the creation of the CSV file, columns that contained little to no entries or information that did not pertain to sales numbers or location were removed to make the data more legible. The remaining information included the state, street_address, city, county, sale_price, property_type, and sale_datetime. These parameters make way for a visualization based on sales prices and location to examine where the most expensive houses are located.

To begin this visualization, we use the map icon to give us the appropriate template to set our parameters in. One option to consider would be to include every single address on the map and use a colored circle to represent the intensity of the value compared to the others. Unfortunately, this creates a very messy visual because there are far too many values that need to be represented! Therefore, to improve on this idea and still evaluate which areas are currently selling the most expensive and least expensive homes, the data will be filtered by county.

Setting the City column under the location parameters will help us place values at every county (County names are placed in the City entry). To help adjust the data and further improve it visually, entries that hold Los Angeles as the city of residence are filtered out. This is due to the large number of entries in this criterion. They hold such heavy weight that they make the other counties less significant and overpower the chart visual. If needed, the data can be evaluated separately. The sales price column is then used to calculate the average sales price for homes in each county. The average price value for each county will be represented by the size and color of the bubble surrounding it. The bigger the bubble, the higher the value. In terms of color, the higher the intensity of the color red filling the bubble, the higher the value. Conversely, smaller bubbles with lighter shades of red represent smaller values. Figure 4 shows the results of the

adjusted data. The chart shows us an idea of how much we'd likely spend on a home depending on what specific area of Los Angeles we wanted to move to.

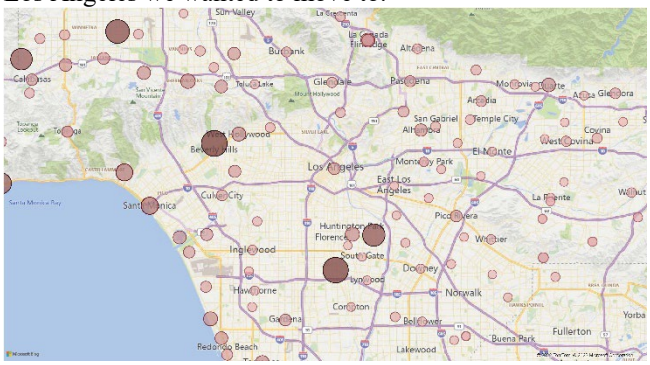


Figure 6

The area with the highest cost for housing is Beverly Hills. The area is marked with a large dark red circle signifying its high value compared to the rest. There are other areas that come close to similarly large numbers, but the amount of homes currently up for sale, coupled with the large values is specific to Beverly hills. The average home in Beverly Hills costs more than $10,000,000. A close second to these values is Reseda where home prices average over $8,000,000. The following section will explore the houses available for purchase across Los Angeles County.

6.2 Number of Properties for Sale in Each City

Using similar techniques in finding the average pricing across multiple counties in Los Angeles, the data for the number of homes for sale in these same counties can also be visualized. To create the visualization in figure 5, the previous example is taken, duplicated, and altered to show the alternative perspective below. This view has a bubble representing the number of homes available for purchase in each county. The size and color of the bubbles are the biggest difference here. The choice to change the colors was made to distinguish both views. In this case, the red bubbles represent a low number while green bubbles represent the highest values.

Looking at Beverly Hills once more, we see a green bubble representing the County. Hovering over the bubble gives us more detail and shows the reasoning behind the assigned color. There are 534 homes for sale in Beverly Hills with an average asking price of more than $10,000,000 (the figure taken from figure 4). The county coming up behind, Reseda, has only 36 homes for sale. This tells us that although Reseda has a high price average, it's only divided among a small number of properties (small relative to other local counties). Figure 5 points to a larger issue that was referenced in the Related Work section. There are far too many expensive homes and not enough people with the wealth to purchase them.

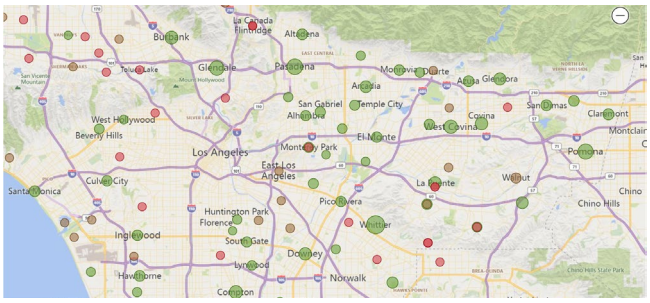


Figure 7

6.3 Temporal Visualization

To create a temporal visualization for the data, we used the 3D Maps tools found in Excel. We began by using similar techniques to the visuals created in PowerBI. The data file used for the previous two visuals was opened and loaded into a table. From here, the 3D Maps button was selected opening a world map view with a selection of options for visualizing the data to the right. Much like the first visual, we will use these tools to compare the **average price per city** but will include the **sale_datetime** field to compare the changes year to year in these averages. This will show us in which cities the property prices increased the most and when prices dipped to their lowest. The temporal graph created will be set to include quarterly changes, in the form of a heat map, between the years of 2019 and the end of 2021. Figure 8 shows the beginning of this range, quarter one of 2019. The data in this period shows us what the data looked like prior to the pandemic.

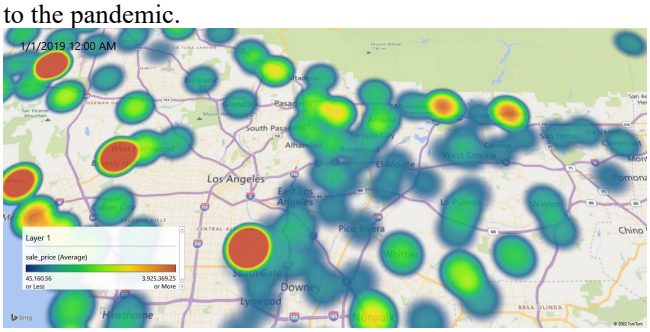


Figure 8

Moving into the following year, we see that the data has not changed much. In May of 2020 (Figure 9), the outlook was still optimistic. The early part of the pandemic had a positive effect on the housing market and lending as whole. As Boehm pointed out in her analytical report on the housing market back in late 2020, the COVID-19 crisis surprisingly resulted in consumers paying down their debt. The total household debt, nationally, declined for the first time in years. Much of this, especially in lending for credit cards, resulted from less spending due to shutdowns and travel restrictions keeping people home. Factors like the government stimulus, increased unemployment benefits, and generous offers for deferment and forbearance on financial responsibilities like mortgage and student loans gave citizens more disposable income than they may have had prior to the pandemic [5]. These forms of relief were intended to solve a temporary problem, but the length of the pandemic was impossible to predict, and this assistance could only do so much.

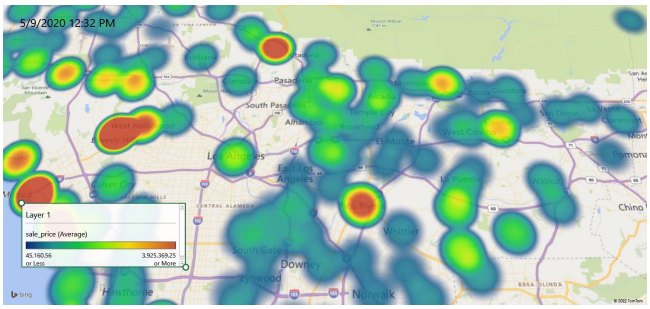


Figure 9

The most detrimental effects of the pandemic can be seen when we compare these figures to the first quarter of 2021 where, as mentioned previously in the Related Works section of this project, the average prices for housing in the surrounding areas of Los Angeles were at their lowest. Shown again here in Figure 10, we can see just how much prices changed.

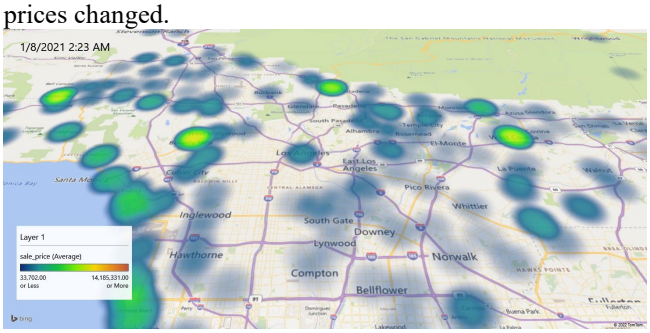


Figure 10

The end of the range defined in our data, the last quarter of 2021, shows a promising change. That is, for the housing market. Figure 11 shows us the average pricing per city during this time. It seems like prices are back to where they were pre pandemic, perhaps at even higher prices than before.

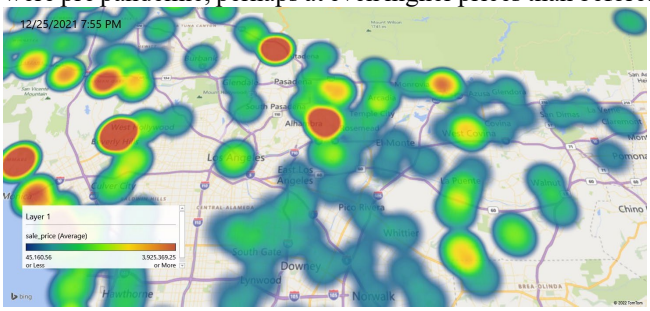


Figure 11

7. Conclusion

Using HDFS, Hive, Excel, and Power BI, we could conclude that the areas with the highest property values in Los Angeles County were Beverly Hills, Reseda, and Calabasas. In addition, the areas with the highest concentration of properties for sale were Long Beach (7124), Whittier (3554), Torrance (2674), Pasadena (2630) and Pomona (2462). Unfortunately, we could not analyze, process, and clean the initial large dataset of 50 GBs from Dolthub. However, extracting and analyzing the larger dataset is possible if a more robust database were accessible. For more information about the project, please visit the GitHub link provided.

We find our work interesting because utilizing and visualizing data from housing in California is fascinating. On a personal level, we are interested in the changes in price in our surrounding area, being that we are all from the Los Angeles area. Our personal connection with the data creates a sense of intrigue for us when pursuing this project and allows us to get more insight into the economic changes in our area. It is also interesting to observe the changes in prices in different mediums. Not only do we experience the changes in terms financially, but we have the opportunity to view these changes in a cohesive and organized way thanks to the usage of the 3D Map function in Microsoft Power BI.

References

[1] L. Liu, *Mortgage Loan and Housing Market*, Elsevier, 2022. Retrieved from: https://doi.org/10.1016/j.iref.2022.10.012

[2] Statista, *Residential Real Estate in California*, Statista, 2022. Retrieved from: https://www-statista-com.mimas.calstatela.edu/study/56661/california-residential-real-estate/

[3] D. Garcia, S. Manji, Q. Underriner, and C. Reid, *The Landscape of Middle-Income Housing Affordability in Calfornia*, Terner Center for Housing Innovation, 2022. Retrieved from: https://ternercenter.berkeley.edu/wp-content/uploads/2022/04/Landscape-of-Middle-Income-Housing-Affordability-April-2022.pdf

[4] S. Malpezzi, *Housing Affordability and Responses During Times of Stress: A Preliminary Look During the COVID-19 Pandemic*, Contemporary Economic Policy, 2020. Retrieved from: https://doi.org/10.1111/coep.12563

[5] J.W. Boehm, *Lending: INCL Impact of COVID-19,* Mintel*,* 2020. Retrieved from: *https://reports-mintel-com.mimas.calstatela.edu/display/987162/?fromSearch=%3Ffilters.category%3D116%26freetext%3Dincl%2520impact%26last_filter%3Dcategory%26resultPosition%3D1*

[6]*US Housing Prices*. (n.d.). Retrieved from: https://www.dolthub.com/repositories/dolthub/us-housing-prices/data/main

[7] Github Link: https://github.com/cmomdji/CIS-4560-Team-2