City Exploration #1: Tour of a Neighborhood

Jessica Tanumihardja

2023-02-08

Background

We are continuing our exploration of the Craigslist Housing Postings in Massachusetts (MA) dataset by Boston Area Research Initiative (BARI) [1]. According to the description, this dataset contains Craigslist housing listings (2020 - 2021) in the regions of Massachusetts as designated by Craigslist. We are continually exploring how this dataset explains or provides insights into the housing market in the Greater Boston area. It should be noted that this dataset mainly covers the pandemic time so further data scrapping is needed to get the "normal" (prior to 2020) and "new normal" (after 2022) trend.

This week, we are exploring new angles by creating new variables and making the data's content more interpretable. These new variables might also fix or shed new light on some issues in the data. This is a continuation of last week's discussion here: Pulse of the City [2].

R Studio is used to process the dataset throughout the semester to gain insights gradually until the final project submission. Following this week's module and Urban Informatics textbook reading on Chapter 5, R is used to create a new numerical variable, filter string from a text column, and produce multiple variable graphs.

New Variables

Location and price are the main determining factors for renters to pick which property to rent. Hence, any information that affects the location and price of the property will be helpful data for the analysis. The three new variables created from the dataset are:

- 'PRICE_PER_SQFT' = numerical variable by dividing the listed price ($) by property area (square foot)
- 'APT' and 'HOUSE' = categorical variables to distinguish whether the property posted is an apartment or house type
- 'STUDENT' = categorical variable to distinguish whether the post is student related

Prior to creating new variables, the dataset is preprocessed to get a subset of more usable data by removing the missing values (NA) and filtering cities to only include Greater Boston Area (41 cities and towns) [3]. Based on the Assignment 2 discussion, we know that there is a lot of address and location that is unusable (empty value, google map reference, etc.). This location sub-setting will enable us to look closer into the city rather than the state. The subset still contains 13,866 observations, which is still within a reasonable range of big data.

Price per Square Foot

Price per square foot is a numerical variable to generalize the price data and compare similar properties (based on location, size, style, etc.). It is definitely not the best valuation metric but one of the simplest before further diving into other complicated variables. It also can

give an overview of the demand of a particular location since a higher-demand location is expected to have a higher price per square foot value.

We discussed last week that some of the listed areas and prices are nonsensical. Therefore, these outliers are filtered before calculating the new variable by excluding the top and bottom quantiles (1%) of the price distribution. From the quantile numbers above we found that 1% of the observations have a listed price lower than $804.5 and higher than $5,500. Similarly, 1% of the observations have listed areas of less than 200 sq.ft. and higher than 2,560 sq.ft. Therefore, only observations with prices listed between $800 - $5,500 and areas between 200 - 3000 sq.ft. are included in the price per square foot analysis.

It should be noted that 1% is an arbitrary number and we can pick any quantile division desired (e.g., following the normal 25% quantile or 3-standard deviation). The purpose of eliminating only 1% of observations is to minimize data loss since we already filtered a lot of observations when cleaning the missing values and filtering the locations. In addition, the range of price and area are reasonable for a rental property listing.

The statistical summary of the PRICE_PER_SQFT variable is shown below (units are in $/sq.ft.). From the histogram (not shown), the distribution still follows a normal distribution (right-skewed). As with any other variable, we can see this new variable from renters, property owners, or cities' perspectives. Renters should expect to pay $2.25/ sq.ft. for a property. The higher the desired size of the property, the higher the rent will be.

Minimum	**1st Quantile**	**Median**	**Mean**	**3rd Quantile**	**Maximum**	**Standard deviation**
0.34	1.91	2.28	2.43	2.79	10.5	0.87

To serve as a comparison tool across different cities and towns in Greater Boston Area, it is useful to plot the median price per square foot as shown in Figure 1. Median values are chosen as it is less sensitive to outliers. The median value for each city is indicated by the dot and the bar is the range of the observation values. Overall, most of the neighborhoods has the same median price per square foot as the overall value ($2.28 /sq.ft.), except for Beacon Hill, East Boston, and Rochester. This shows that the Greater Boston Area has a similar "pulse" across the different neighborhood. These values seem to be more extreme than the rest, so it needs to be analyzed further for outlier or anomaly.

The wider price range can be seen in Boston, Cambridge, and Somerville. The higher values for Boston and Cambridge are expected since they are highly demanded locations. The higher value and wider range of prices for Somerville are unexpected and worth further investigation.

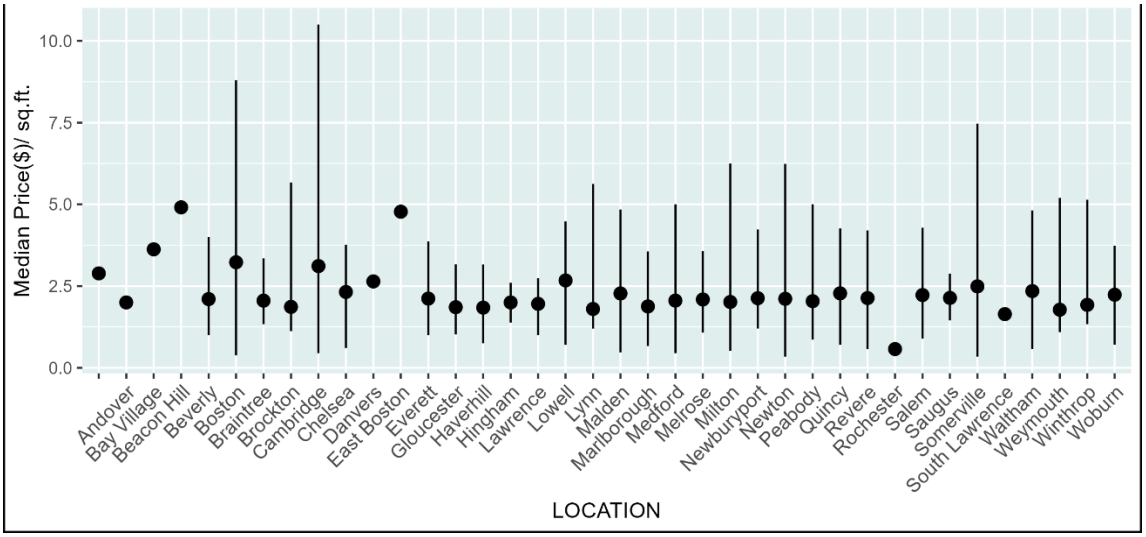

Figure 1 Median PRICE_PER_SQFT across all cities and towns in Greater Boston Area

Apartment or House

'APT' and 'HOUSE' variables are made by filtering the character variable 'BODY' of the Greater Boston Area dataset (without removing NA in LOCATION or ADDRESS variables). This variable contains description by the listing writer on any related information of the listing. It often contains important information such as number of bedrooms & bathrooms, convenient features of the property, and nearby places of interest. The next new variables are related as both indicate the type of property: apartment (multi-family) or house (single-family). This category is often an important distinction for renters as they might prefer one over another. The presence and amount of apartment or multi-family property listings might also indicate the population density of the city.

By filtering and counting the related word, there are 17,075 apartment-related ("apartment", "apt", "Apartment") postings among the 25,540 observations (66.9%). On the other hand, there are only 6,480 house-related ("house", "House") postings (25.4%). There are some gaps in the numbers, which means some posting does not clarify whether it is a house or an apartment type. This can be indirectly translated to higher population density in the Greater Boston area.

Figure 2 shows the distribution of the price according to the two new variables. The y-axis facet indicates whether it is a non-apartment (0) or an apartment-related (1) posting. The x-axis facet indicates the non-house (0) or house-related (1) posting. All of them follow the normal distribution curve, with a mean price of ~$2,000. The similar mean indicates that although the apartment size might be smaller than a house, it might be located closer to downtown Boston or Cambridge (presumably a highly sought area). Furthermore, most of the posts are apartment and non-house related, which aligns with the counts above. It is interesting that the mean price for apartments is higher in 2020 than in 2021, maybe due to a lack of demand in 2021 as also shown in last week's analysis.

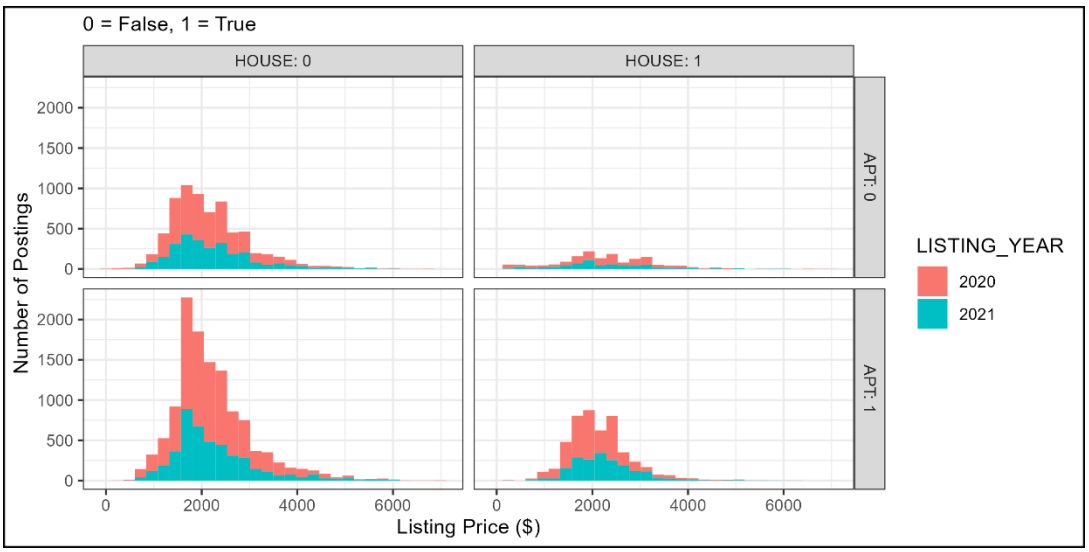

Figure 2 Histogram of the price of Apartment or House related postings

Performing the same analysis as the first variable across different locations, Figure 3 shows the bar chart of the two new variables. The top figures represent the apartment distribution across the Greater Boston area. For both graphs, Cambridge, Boston, and Somerville are still the top three posting counts respectively. Most of them are apartment-related posting. This might indicate that in high-density places, there is limited availability for single-family housing. It might also indicate more multi-family residency than single-family housing. One possibility is the owner or renter that lives in the bigger cities might like the location more and don't want to move out. Another possibility is more developers build multi-family housing in the bigger city due to limited space. The highest house postings are in Saugus (11 miles from Boston), followed by Lawrence and Lowell.

Student

Similar to the previous variable, the next new variable also obtained by filtering the 'BODY' variable. The postings that have words related to student ("student", "Student", "campus", "Campus", "University", "university", "school", "School") are lumped together. Surprisingly, only 3,074 postings among 25,540 postings (12%) are related to student in the Greater Boston Area. This number might increase if we limit the location to the Boston neighborhood only. The lower student-related postings indicate that although there are multiple universities in Greater Boston Area, the majority of the population still consists of families and workers. We can match this with the census data of age range in the location or the occupation data discussed in the next section.

We have created new variables: 'PRICE_PER_SQFT', 'APARTMENT', 'HOUSE', and 'STUDENT'. The first variable is numeric while the remaining are binary (categorical). We can compare this number with other cities and compare whether this is a high number or a low number. Likewise, we can create a new data frame of the cities NOT in the greater Boston area to compare this result.

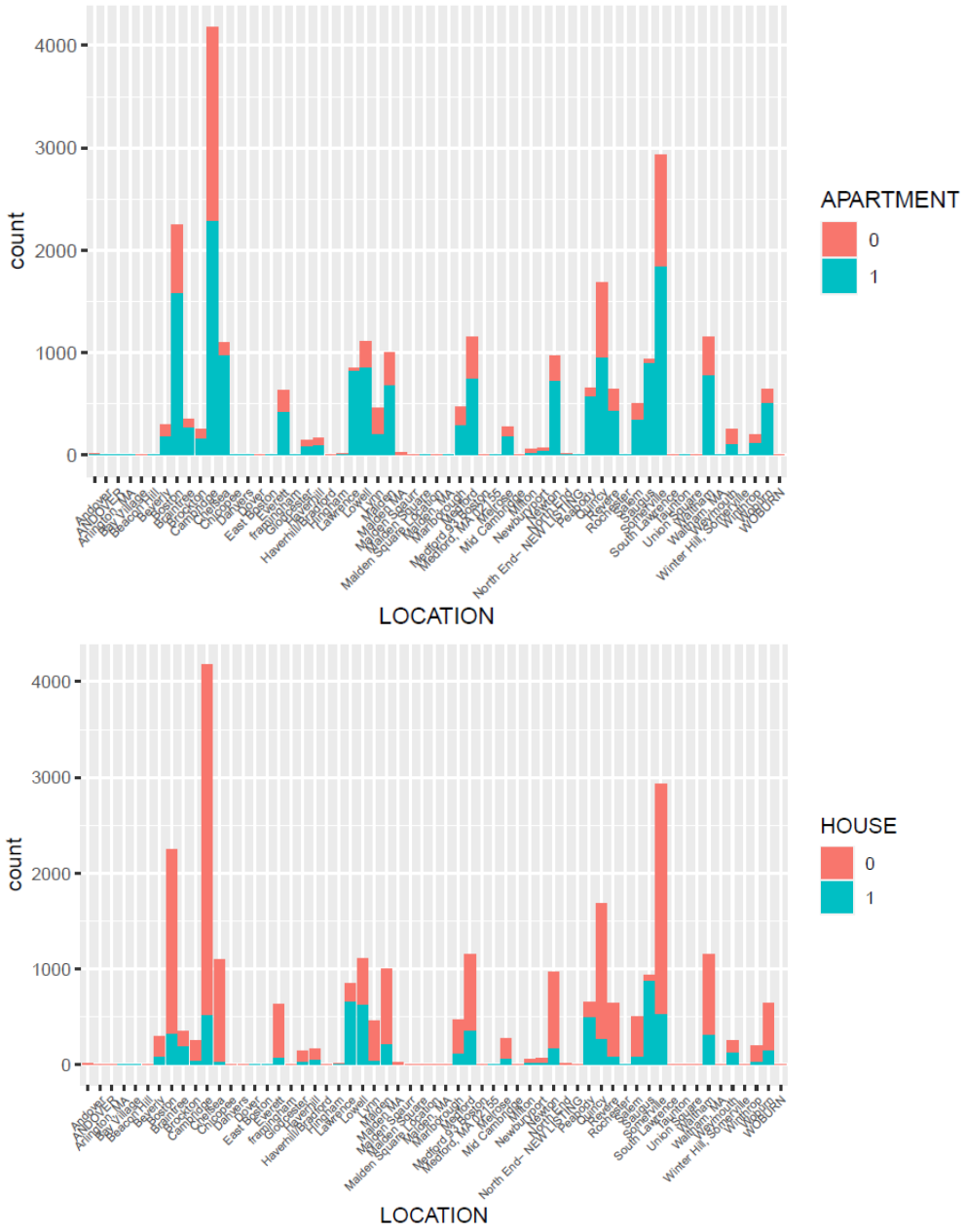

Figure 3 Property type distribution per neighborhood

Neighborhood exploration: Somerville, MA

From the new variable analysis above, Cambridge, Boston, and Somerville frequently appear in the top three posting counts. Since Boston and Cambridge are popular locations, Somerville is chosen for this city exploration since it is less popular, at least for an outsider, but relatively close to downtown Boston (2 miles northeast). If the population continues to grow in Boston, this area might be the next popular location to live. Somerville also has a balance of apartment and housing posts, with surprisingly higher student-related posting counts than Boston.

According to the City of Somerville website, the city is home to 79,815 people with diverse occupations and ethnicity within ~4 square miles of the city boundaries [4]. Somerville has a long history as an incorporated town since 1630 and became a town in 1842. The city was also involved in the American Revolutionary War. With its proximity to Boston, the city has been growing to serve the increasing population growth. This is apparent as Somerville is the first community to activate 311. The city serves various services and amenities as it is paving its way to being one of the metropolitan cities [3]. Somerville has an active social media account (Facebook, Twitter, Instagram, YouTube, and Vimeo) and its own independent newspaper, the Somerville Times [5].

Several articles voted Somerville as the top 10 places to live in Massachusetts for its historical and cultural roots, quality of life, and affordability [6, 7, 8]. It is ranked 7 in the 2022 top 50 Best Places to Live in the US, with Milton as the only other city in Massachusetts that made it to the rank [8]. A quick internet search shows numerous restaurants and pubs (especially East Somerville), multiple dog parks, and various grocery stores (local and chain).

Several google map pictures show a general view of the city in Figure 4 below. Most of the single-family homes are two-story with minimal yard area. Federal, Greek Revival, Italianate, and Queen Anne are the most common house style in Somerville [4]. A few of the buildings still carry the historical façade as the City of Somerville is working on historical preservation. Most apartment complexes are not high-rise and are usually located nearby bigger streets. The commercial area has a wide street, street parking, bike lanes, and wide pedestrian walkways.

Figure 4 Google Maps Street View of a sample of Sommerville's Apartment (top left), commercial street (top right), residential (bottom left), and mix-used street (bottom right) areas [9]

Population and housing characteristics of Somerville can be obtained from the BARI's ArcGIS dataset. This website map each neighborhood characteristics according to 2015-2019 census for each census tracts. For a sample census tract shown in the figure, the age distribution of Sommerville's resident is [10]:

- Age <18 years old: 11%
- Age 18 – 34 years old: 45%
- Age 35 – 64 years old: 31%
- Age over 65 years old: 13%

The higher young adult distribution explains the higher property listing related to student in Somerville. The city is neighboring 4 Universities: Tufts University (Medford), Harvard University (Cambridge), Massachusetts Institute of Technology (Cambridge), and Lesley University (Cambridge). There are also numerous colleges and university in Boston that is within 5 miles distance of the city.

Most of the population still chooses to commute by car (~39% in the sample census tract) with public transportation coming in a close second (~27%). The majority of the population has at least a bachelor's degree (~35%), which shows that this city has a strong educational background. Most of the neighborhoods also have more female populations compared to males. Also, the city has a relatively low unemployment rate of ~2 – 3 %.

The median household income for Somerville is shown in Figure 5. For one sample of the neighborhood, the median household income is $97,833 [10]. For reference, the estimated nationwide median household income is $70,784 in 2021 [11]. There are more areas with higher median income, especially in West Somerville. East Somerville has more commercial areas so it might be the reason for the lower median income. It is interesting that the Craigslist housing map from the BARI map for the median rent (2021) coincides with the median household income map below; the areas with lower median household income also have lower median rent.

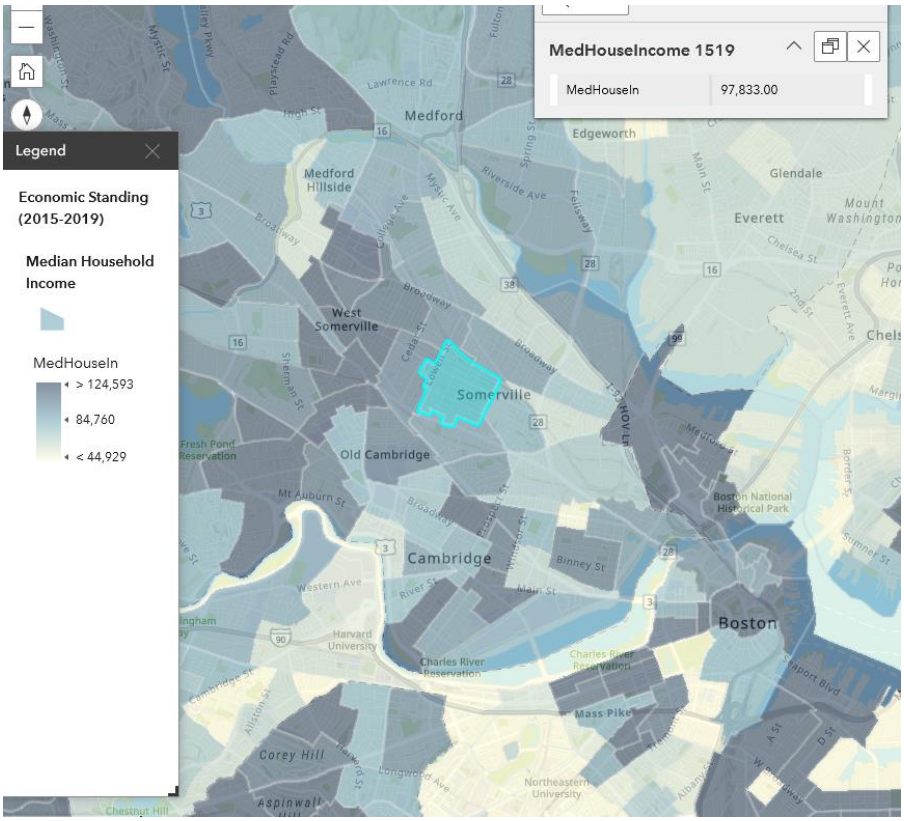

Figure 5 Sommerville's median household income distribution based on 2015-2019 Census (darker color means higher income) [10]

Re-evaluation and Summary

All things considered, Somerville is an appealing city to live in for people who like city life and convenience but also appreciate the peacefulness within Greater Boston. This explains the higher rental posting counts and median rental price in Somerville. The median household income for Somerville is also higher than average Massachusetts or US nationwide, hence, the higher listing price is justifiable. Somerville is also rich in cultural and historical values which might increase the quality of life for the residents.

Although there are more apartment listings for Somerville based on the Craigslist dataset, it was quite rare to find apartment complex through Google street view. Most of them are located in bigger intersection or bigger street. As described above, most of the apartment is also not high-rise buildings. It might strengthen the hypothesis that the fewer house rental listing might be caused by most of the houses being owned by the family instead of rented. The higher apartment listing counts might correlate with the higher student-related listings as most students likely prefer to rent apartments to houses.

Somerville also has convenient public transportation routes. It might be good to create a new variable related to public transportation (buses and trains) to see if it is listed as an advantage when listing a rental property. Analysis of Craigslist observation per census tract ID might also be useful since most of the population characteristics are divided per census tract instead of relying on the location listed in the 'BODY' variable.

References

[1] E. Castro, M. Amiri, R. Tucker and D. T. O'Brien, "Harvard Dataverse," Boston Area Research Initiative's Boston Data Portal, 24 08 2020. [Online]. Available: https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/52WSPT. [Accessed 23 01 2023].

[2] J. Tanumihardja, "SPPUA5262 Big Data for Cities," Wordpress, 01 2023. [Online]. Available: https://sppua5262.wordpress.com/2023/01/30/pulse-of-the-city-craigslist-housing/.

[3] "Greater Boston Regions," Meet Boston, [Online]. Available: https://www.meetboston.com/plan/greater-boston-regions/. [Accessed 07 02 2023].

[4] "About Somerville," City of Somerville, 2023. [Online]. Available: https://www.somervillema.gov/about. [Accessed 08 07 2023].

[5] "The Somerville Times," The Somerville Times, 2021. [Online]. Available: https://www.thesomervilletimes.com/. [Accessed 08 02 2023].

[6] "The 9 Best Neighborhoods to Live in Boston in 2022," Bungalow, 2023. [Online]. Available: https://bungalow.com/articles/best-neighborhoods-in-boston-massachusetts. [Accessed 07 02 2023].

[7] M. Thielking, "Why Somerville, Massachusetts Is One of the Coolest Suburbs in America," Apartment Therapy, 22 05 2019. [Online]. Available: https://www.apartmenttherapy.com/somerville-massachusetts-guide-32253445. [Accessed 08 02 2023].

[8] A. H. A. J. Y. A. A. J. J. K. M. L. C. J. G. M. M. M. E. C. a. S. H. Kristen Bahler, "The 50 Best Places to Live in the U.S. | 2022—2023," Money, 2021. [Online]. Available: https://money.com/best-places-to-live/somerville-massachusetts/. [Accessed 08 02 2023].

[9] "Somerville, Massachusetts," Google Maps, 2019 - 2020. [Online]. Available: https://www.google.com/maps/place/Somerville,+MA/. [Accessed 08 02 2023].

[10] "Boston Area Research Map," Northeastern University, 2021. [Online]. Available: https://boston-area-research-map-nu.hub.arcgis.com/ . [Accessed 08 02 2023].

[11] J. S. a. M. Kollar, "Income in the United States: 2021," U.S. Census Bureau, 13 09 2022. [Online]. Available: https://www.census.gov/library/publications/2022/demo/p60-276.html. [Accessed 08 02 2023].