Big Data for Housing and Their Interaction with Market Dynamics

Jieun Lee
National University of Singapore

Kwan Ok Lee (✉️ kwan.lee@nus.edu.sg)
National University of Singapore

Research Article

Keywords: online listing, keywords, natural language processing, housing advertisement, COVID-19, PropTech

Posted Date: October 18th, 2022

DOI: https://doi.org/10.21203/rs.3.rs-2153935/v1

License: ☇️ ④ This work is licensed under a Creative Commons Attribution 4.0 International License. Read Full License

Additional Declarations: No competing interests reported.

Version of Record: A version of this preprint was published at Journal of Big Data on June 11th, 2023. See the published version at https://doi.org/10.1186/s40537-023-00786-5.
Abstract

This paper is the first to analyze the interactions between the keywords of online home listings and housing market dynamics. We consider the COVID-19 outbreak as a natural shock that brought a significant change to work modes and mobility and, in turn, consumer preference changes for home purchases. We link two types of big data: the universal transaction data of resale public housing and the database of more than 70,000 listings from the major online platform in Singapore. Using the Difference-in-Difference approach, we first find that housing units with a higher floor level and more rooms have experienced a significant increase in transaction prices while close proximity to public transportation and the central business district (CBD) led to a reduction in the price premium after COVID-19. Our text analysis results, using the natural language processing, suggest that the online listing keywords have consistently captured these trends and provide qualitative insights (e.g. view becoming increasingly popular) that could not be uncovered from the conventional database. Relevant keywords reveal trends earlier than transaction-based data, or at least in a timely manner.

1. Introduction

While the real estate sector has traditionally lacked the digitalization process, the emergence of big data and the Internet of Things has recently led to a boom in technological advancement, called Property Technology (PropTech). Online listing platforms have been widely used by potential home purchasers for searching and real estate agents for marketing. Consumers obtain useful information from the big data on these platforms and are able to save time and money by picking and choosing only the homes they are interested in. And to better attract the attention of consumers, real estate agents carefully choose keywords to advertise homes (Delmelle and Nilsson, 2021). If there are significant shocks to consumer preference changes for homes, real estate agents are likely to respond by adjusting their keywords for home listings.

In this paper, we attempt to analyze whether and how quickly keywords for online home listings reflect actual trends in consumers’ housing preferences. We use the COVID-19 outbreak as a natural shock to the local housing market in Singapore. As the COVID-19 pandemic has limited public mobility and transformed work arrangements, and in turn changed consumer preferences for their homes, it provides a good opportunity to explore how quickly keywords have captured changing preferences. For example, we would expect increasing popularity for more spacious homes with a nice view while accessibility to public transportation would be a less important factor for home-buying decisions in the post-COVID-19 period. Analyzing the big data from online listings is useful for a better understanding of market trends. For example, keywords may highlight the housing features, such as a view, which are not captured in the standard transaction database but important for homebuying decisions.

We use the transaction database of resale public housing provided by Singapore’s Housing and Development Board (HDB), together with more than 70,000 listings for the period of 2019 to 2021, from the major online listing company, PropertyGuru, which has 84% of the market share regarding agent
subscription revenue in Singapore (Business Wire, 2022). To understand housing features that become
more or less popular after the COVID-19 shock from actual transactions, we use the hedonic pricing
model. In doing so, the Difference-in-Difference (DID) modelling approach is applied to account for
temporal variation (before vs. after COVID-19) as well as variations in key housing features (i.e. floor level,
floor size, proximity to a subway station and the central business district [CBD]). Then, we explore the
keyword trend of online listings by performing an exploratory text analysis on keywords related to
housing features that are shown to have a significant change in their popularity in the housing market.
We also compare the temporal trends between housing price premiums and advertised keywords, using
the event study specifications.

Broadly, this research sits in the literature on big data analytics in online marketing (e.g. Guangting and
Junxuan, 2014; Akter and Wamba, 2016; Altaweel and Hadjitofi, 2020). To our knowledge, this is the first
paper to consider keywords for online advertisement as a means to reveal a dynamic change in
consumer preferences for homes. Online platforms for property listings provide a unique context where
consumers collect online information before viewing offline commodities which feature very
heterogenous attributes (Boeing, 2019; Su et al., 2021). Also, real estate transactions are recorded with
various characteristics and the relevant big data exist in some advanced economies like Singapore.
Hence, a focus on real estate listings provides us the excellent opportunity to examine how online
advertisement interacts with actual market trends, which could also be advertisement outcomes (i.e.
transactions). Real estate agents play an important role in this process as they are both the advertiser
and sales person on behalf of landowners. They spend a significant amount of time and effort to make
listings more attractive as they get commission generally based on the amount they sell (Jud and Frew,
1986; Yavas, 1994; Munneke and Yavas, 2001). Our analyses attempt to test how precisely and quickly
agents capture changes in consumer preferences and market trends.

Next, our paper contributes to the thin literature on the role of big data in real estate markets. Existing
research suggests that online keyword search volume or consumer sentiment from major news articles
help forecast the prices or returns for real estate properties from several months to four quarters in
advance (Yang et al., 2013; Venkataraman et al., 2018; Beracha et al., 2019). Different from these papers,
however, our focus is on the changes in advertised keywords driven by the COVID-19 shock, which can
suggest qualitative changes in consumer preferences. Our focus on keywords builds upon the existing
research. Nowak and Smith (2017) and Shen and Ross (2021) use online listing descriptions to measure
the uniqueness of housing properties while Delmelle and Nilsson (2021) use them to identify
neighborhood characteristics. Because keywords highlight the housing features that are not captured in
the standard transaction database but are important for homebuying decisions, our analyses can provide
a better understanding of the market trend. If we reveal that online listings capture market dynamics more
quickly than ex-post transaction data, this result has an important implication for housing market
forecasting.

Finally, our paper adds insights into how increasing work-from-home (WFH) arrangements influences
housing choices. A series of research suggest that the COVID-19 pandemic has influenced housing
demand especially with the substantial increase in WFH along with reduced mobility because of restriction measures and infection concerns (e.g. Di Renzo et al., 2020; Muhyi and Adianto, 2021; Zarrabi et al., 2021; Bottero et al., 2021). Researchers report that residential demand has increased in neighborhoods with lower density and crowdedness (Liu and Su, 2021; Balemi et al., 2021) and in housing units with more space for the home office and terrace (Bottero et al., 2021; Boesel et al., 2021; Zarrabi et al., 2021). Based on existing evidence, we encompass both physical and locational features of housing units for our analysis. One advantage of using the resale public housing data in Singapore is that units are quite homogeneous in terms of design and community facilities (Lee, 2021), making it easier to disentangle the price premium changes associated with specific features like a higher floor level or proximity to a subway station. Our research also features the Singapore’s unique context for residential choices within one city, which differs from most other countries that have more than one metropolitan area with urban and suburban locations. Also, unlike existing research, we focus on multifamily housing units in the high-density environment. We attempt to test the external validity of existing evidence from other countries and report how Singapore’s unique contexts lead to similar or different results.

The rest of this paper is organized as follows. The next section provides brief scholarly and institutional backgrounds for our research followed by the section that describes data and methodology. Next, the main results of this research are presented with the COVID-19 effects on changes in housing price premiums with respect to different housing features, and the exploratory text analysis shows how keywords reflect actual market dynamics. Finally, we conclude and provide implications.

2. Background

2.1. The COVID-19 Outbreak as a Shock to Consumer Preference Change

The COVID-19 pandemic has influenced household housing demand, especially with the dramatic increase in work-from-home (WFH) arrangements along with restricted mobility and infection concerns (Di Renzo et al., 2020; Muhyi and Adianto, 2021; Lee and Lee, 2021; Zarrabi et al., 2021; Bottero et al., 2021). At the same time, the COVID-19 shock has influenced preferences for residential locations. A decline in demand was reported to be strongest in more densely populated neighborhoods and central cities (Liu and Su, 2021). In the pre-COVID-19 period, homes in densely populated areas in cities had been valued for their access to amenities such as shopping malls and railway stations as well as proximity to workplaces. However, after the COVID-19 outbreak, visitors to and the expenditures at crowded places such as shops, restaurants, and gyms have decreased (Allcott et al., 2021), and these amenities may have been less attractive to potential homebuyers (Liu and Su, 2021). Also, the perceived risk for infection tends to be higher in high-density areas, and in turn, decreases the demand for housing in these areas (Balemi et al., 2021).

Simultaneously, research has reported a shift in consumer demand for larger space, both in the home's interior and exterior (Boesel et al., 2021). National lockdowns and social distancing measures to fight
against the COVID-19 pandemic have required people to conduct daily activities at home (Kim, 2021). Hence, homes with larger interior living space and larger lots have experienced a higher increase in their housing prices. This is to accommodate the increased demand for separate spaces for working individuals who need a home office and school-age children who take in-home classes (Boesel et al., 2021). Also evident is that consumer demand has increased for homes with natural light, visibility, the better acoustics of interior spaces, and open or semi-open space (e.g. terraces), which is associated with efforts to prevent potential psychological damage caused by staying at home (Zarrabi et al., 2021).

### 2.2. Online Platforms and Keywords for Home Advertisement

According to a survey conducted by the National Association of Realtors (NAR; 2022), 95% of homebuyers use online platforms to search for homes, and over half of them end up purchasing the housing units they found on these platforms. The first step most homebuyers across all age groups take during their home purchase process is searching the internet for available properties (NAR, 2022). Keywords attract people’s attention to connect sellers’ target audience with a detailed webpage of each listing. When the keywords effectively catch the attention of target customers, they could bring more potential buyers to the product’s webpage and increase the probability of successful sales (Ghose and Yang, 2009; Yang and Ghose, 2010; Rutz, Bucklin and Sonnier, 2012).

The existing evidence suggests that online listing descriptions affect the transaction prices of advertised housing units. Haag et al. (2000) suggest that while both objectively verifiable and subjective keywords have a significant price effect, some keywords falling in the former category (e.g. good location) have a negative association with selling prices potentially because they are hype. Goodwin et al. (2014) argue that the price effect of online keywords is more significant in heterogeneous housing markets (i.e. detached housing markets). Sing and Zou (2021) suggest that online listing keywords could play an important role in the price premium for advertised housing units if they deliver essential information and provide easy-to-read and mood-independent descriptions.

### 2.3. The Singapore Context

Singapore’s government has responded to the COVID-19 pandemic with a range of measures such as strict border controls; contact tracing; home isolation with the closure of schools, universities, and workplaces; social distancing; and allowing only essential businesses, such as grocery stores and banks, to open. The government imposed a circuit breaker, a stay-at-home order that mandates the closing of non-essential businesses and transitioning to home-based learning, on 3 April in 2020. After the end of the circuit breaker on June 1, three phases of gradual reopening were imposed. In Phase 1 (2–18 June, 2020), offices re-opened, but work-from-home was mandated to the maximum extent. In Phase 2 (19 June–27 December, 2020), work-from-home was a default arrangement, and the group size of gatherings was restricted to five people. In Phase 3 (28 December 2020 – July 2021), the government changed safety measures based on domestic COVID-19 cases, ranging from work-from-home as a default to a requirement of only 50% or 75% of workers in offices.
These strict measures lasting over one year have significantly affected not only the national economy but also individuals’ lives in Singapore. Work-from-home (WFH) and home-based learning became a default lifestyle for most households while their social gatherings were largely limited. These have significantly decreased public mobility and changed household housing demand in Singapore. Survey results show that most respondents consider WFH as a longer-term trend and they are willing to trade commuting time and other preferences for housing with more space and a better view (Deng, 2020; Soo, 2022). In addition, anecdotal evidence suggests that as the whole family spent most of their time at home, more households wanted larger spaces (Lin, 2021). Some households decided to move away from the central area to the less expensive, suburban areas to larger housing units.

Given that the COVID-19 outbreak and associated measures mainly affected locals’ lives and changed their housing demand, our analyses focus on online listings and public housing transactions in Singapore. Public housing in Singapore is developed and sold by a government agency called the Housing and Development Board (HDB). According to the 2021 Singapore Statistics, 78.3% of Singapore resident households reside in public housing. While foreign investors can purchase private housing and drive its market trends, public housing units can be purchased only by Singapore citizens or permanent residents. Public housing units can be purchased at a subsidized price when directly buying from the HDB, but after a 5-year minimum occupation period, homeowners of public housing are allowed to sell their units in on the open market without price control. The online listing and transaction database used for all analyses are for resale public housing. Hence, the changes in keywords for online listings and price premiums observed from actual transactions can capture the trends in the local public housing market.

3. Data And Methodology

3.1. Data

We use two type of big data for real estate properties. The first is a dataset on universal resale transactions of public housing from January 2017 to December 2021 in Singapore. The data are collected by the Housing Development Board (HDB) as the HDB administers public housing transactions and all resale transactions should be reported to the HDB. Therefore, this data has full coverage of all resale market transactions and include transaction price, floor level, floor area, number of rooms, location, and age of each residential building.

The second data source is propertyguru.com.sg, the biggest online house listing platform, which has 84% of the market share regarding agent subscription revenue in Singapore. We collected public housing listings for resale at 10 distinct time points over the period from September 2019 to July 2021 using a python-based web-scraping tool. After removing duplicate listings, our sample has more than 70,000 listings with various information including unit characteristics (i.e. the number of rooms, floor area, floor level category [high/middle/low]), location characteristics (i.e. latitude and longitude, neighborhood town), building age, and the asking price. The data also include the text headline used to advertise each home listing, which is important for our analyses.
Table 1 provides descriptive statistics of the resale transactions of public housing used for our empirical analyses. In our first analysis of actual housing demand change, our sample covers the data from January 2017 to December 2021 to test longer-term price trends associated with key housing features. Compared to transactions before the COVID-19 outbreak (column 2), housing units traded after COVID-19 (column 3) show higher transaction prices. Results also suggest that units with more rooms were transacted more frequently after COVID-19 while the average floor area of transacted units did not change significantly. In terms of public transport accessibility, units closer to the nearest subway station were transacted less after COVID-19. These may suggest potential changes in transaction trends in resale public housing markets.
Table 1
Summary Statistics of Public Housing Resale Transaction Data

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of observations</td>
<td>116,397</td>
<td>63,971</td>
<td>52,426</td>
</tr>
<tr>
<td>Variable</td>
<td>Mean</td>
<td>Std. Dev.</td>
<td>Mean</td>
</tr>
<tr>
<td><strong>Unit characteristics</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Real price (S$)</td>
<td>463,404</td>
<td>160,717</td>
<td>443,391</td>
</tr>
<tr>
<td>Price per sqm</td>
<td>4,771.4</td>
<td>1,298.0</td>
<td>4,568.1</td>
</tr>
<tr>
<td>% 10th or higher floor</td>
<td>0.38</td>
<td>0.49</td>
<td>0.38</td>
</tr>
<tr>
<td>% four or more rooms</td>
<td>0.75</td>
<td>0.43</td>
<td>0.59</td>
</tr>
<tr>
<td>Floor area (sqm)</td>
<td>97.83</td>
<td>24.10</td>
<td>97.72</td>
</tr>
<tr>
<td><strong>Location characteristics</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Distance to the closest subway station (m)</td>
<td>775.17</td>
<td>445.94</td>
<td>757.16</td>
</tr>
<tr>
<td>% near subway (within 400 m)</td>
<td>0.19</td>
<td>0.40</td>
<td>0.20</td>
</tr>
<tr>
<td>Distance to CBD (km)</td>
<td>10.40</td>
<td>4.53</td>
<td>10.51</td>
</tr>
<tr>
<td>% near CBD (within 2 km)</td>
<td>0.04</td>
<td>0.19</td>
<td>0.04</td>
</tr>
<tr>
<td>Distance to the closest park</td>
<td>1,319.94</td>
<td>686.82</td>
<td>1,333.46</td>
</tr>
<tr>
<td>Distance to the closest mall</td>
<td>625.67</td>
<td>360.56</td>
<td>627.78</td>
</tr>
<tr>
<td>Distance to the closest primary school</td>
<td>598.91</td>
<td>722.76</td>
<td>595.50</td>
</tr>
<tr>
<td><strong>Temporal characteristics</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>% Post-COVID-19</td>
<td>0.45</td>
<td>0.50</td>
<td>0.00</td>
</tr>
</tbody>
</table>

Note: All monetary values are in CPI-adjusted real terms (S$2019). S$1 = US$0.74 as of December 2019.

3.2. Methodology
### 3.2.1. Empirical Model to Examine COVID-19 Effects on Actual Housing Price Premiums

To explore how housing price premiums have changed with respect to key housing features after COVID-19 based on actual resale transactions, we employ difference in differences (DID) as follows:

\[
y_{it} = \alpha + \delta D_i + \gamma T_i + \beta D_i T_i + \theta X_i + \phi F_{tj} + e_i
\]

where \( y_{it} \) is the transaction price per square meter of the housing unit \( i \) in month \( t \), \( D_i \) is the binary treatment indicator of whether unit \( i \) has a certain housing feature (i.e. 10th or higher floor, four or more rooms including the living room, within 400m to a subway station, and within 2km to the CBD or not), and \( T_i \) is the binary indicator of whether unit \( i \) was transacted before or after the COVID-19 outbreak. The coefficient of the interaction term of \( D_i \) and \( T_i \), \( \beta \), reflects the change in the price premium of a housing feature of our interest after the COVID-19 outbreak. \( X_i \) is a control vector of unit- and location-specific characteristics. Unit characteristics include the floor level and area of the unit, building age, and their square terms. Location characteristics include the distances to the closest subway station, park, primary school, and mall and the distance to the CBD. \( F_{tj} \) is the vector of stringent fixed effect terms regarding both geographical (street) and temporal (year-month) variations to increase the probability that the treatment effect is as exogenous as possible.

### 3.2.2. Text Analysis to Explore Online Housing Advertisement Trends after COVID-19

To understand whether and how online advertisement captures the trends from actual housing markets shown above, we focus on description headline texts obtained from online listings and perform natural language processing (NLP) such as word cloud, unigram, and bigram analyses. Research has frequently used NLP to categorize unstructured texts into certain patterns and generalize the common trends from the texts (Konkol et al., 2015; Altaweel and Hadjitofi, 2020). Real estate literature on online advertisement has also used NLP to understand soft features which could be obtained only from unstructured text data, in addition to hard attributes of properties (Hill et al., 1997; Pryce and Oates, 2008; Nowak and Smith, 2016; Goodwin et al., 2018; Alfano and Guarino, 2022).

Following the standard text pre-processing methods (Vijayarani et al., 2015; Potrawa and Tetreva, 2022), we removed unmeaningful words including prepositions, articles, and ‘be’ verbs for the keywords of each listing headline and converted all words to a simpler form and lower case. Then, we performed a word-cloud analysis before and after COVID-19 for exploratory purpose. After depicting a brief overview of keyword trends, we perform N-gram analyses with unigrams and bigrams. With the pre-processed headline texts, we ranked the top 1,000 most frequently appearing keywords in each year and calculate the share of the frequency of each keyword \( p_{it} \) out of the total frequency sum of each year. Then, we calculate and visualize the percent changes of the share of each keyword in the years 2020 and 2021.
compared to 2019 to understand their trends after the COVID-19 outbreak. We also perform the bigram text analysis in the same manner but with a pair of words. In doing so, we categorize keywords into different dimensions such as view-related, space-related, and location-related, and display results by each category. We determine these categories to be related with housing features that we focus on in our DID model introduced above. For example, we compare results for view- and space-related keywords with DID results for the floor level and size, respectively, while associating location-related keywords with the proximity to the subway station and CBD.

3.2.3. Empirical Model to Compare the Trends of Price Premiums and Advertised Keywords

In addition to the above DID model, we investigate the quarterly movement of price premiums compared to the first quarter of 2019 with an event study identification, which is commonly used to address the parallel trends issues for DID approaches (Lee, 2021; Lee and Phang, 2022). This allows us to relax our identification with multiple time windows and observe the dynamic changes in price premiums for each housing feature. By overlaying these changes with the temporal trends of relevant keywords from online listings, we attempt to find out how quickly the online listing keywords respond to, or even lead, the changing trends in housing markets. We note that housing transaction data are usually available one month later. This means that there may be a slight lag time for the real estate agents to adjust keywords to market trends.

4. Results

4.1. COVID-19’s Effects on Housing Demand Change based on Actual Transactions

Table 2 displays estimates of the Difference-in-Difference (DID) model presented in Eq. (1) with price per square meter as a dependent variable. The average transaction prices increased after the COVID-19 outbreak in Singapore. Other countries have shown heterogeneous housing market outcomes in the post-COVID-19 period. For example, studies in China have reported a reduction or minor change in housing transaction volume and prices, especially in the regions where infection rates were high and the government maintained anti-speculation policies (Tian et al., 2021; Qian et al., 2021; Cheung et al., 2021, Zeng and Yi, 2022). However, in many other countries, housing prices have significantly increased due to low interest rates, a desire for more space as people work from home, and accumulated savings during the lockdown (Romei and Giles, 2021). In the case of Singapore, the main reason for the price boom was the limited supply because of delayed construction coupled with strong investment demand from overseas because of the stability of the Singapore housing market as well as domestic demand rising from higher liquidity (Ting, 2021).
Table 2
COVID-19 Effects on Price Premiums Associated with Actual Housing Features (Difference-in-Difference Regression Results)

<table>
<thead>
<tr>
<th>Dependent Variable: resale price per square meter</th>
<th>(1) 10th + floor</th>
<th>(2) 4 + rooms (3 + bedrooms with a living room)</th>
<th>(3) Within 400m to subway</th>
<th>(4) Within 2km to CBD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Post-COVID-19</td>
<td>319.45</td>
<td>288.95</td>
<td>354.38</td>
<td>348.15</td>
</tr>
<tr>
<td></td>
<td>(4.62)</td>
<td>(6.52)</td>
<td>(4.16)</td>
<td>(3.97)</td>
</tr>
<tr>
<td>Post-COVID-19 × 10th or higher floor</td>
<td>54.72</td>
<td>***</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(5.25)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10th or higher floor</td>
<td>296.66</td>
<td>***</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(3.48)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Post-COVID-19 × four or more rooms</td>
<td>82.55</td>
<td>***</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(6.24)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>four or more rooms</td>
<td>-151.33</td>
<td>***</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(4.62)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Post-COVID-19 × near subway</td>
<td>-49.46</td>
<td>***</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(6.54)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>near subway</td>
<td>252.29</td>
<td>***</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(5.26)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Post-COVID-19 × near CBD</td>
<td>-42.62</td>
<td>***</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(15.97)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>near CBD</td>
<td>66.29</td>
<td>***</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(25.02)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Location characteristics</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
</tr>
</tbody>
</table>

* p < 0.1; ** p < 0.05; *** p < 0.01. Robust standard errors in parentheses.

Note: For public housing in Singapore, the living room is included in the number of rooms. So we essentially capture whether the unit has three or more bedrooms.
Our main interest lies in the extent to which the average price premium has changed with respect to key housing features, including the floor level, spaciousness, proximity to subway stations, and proximity to the CBD, after the COVID-19 outbreak. We first find that the units located on a higher floor level have experienced an increase in their transaction price by $54.72 per square meter after COVID-19 (column 1, Table 2). Similarly, the transaction price for units with three or more rooms has significantly increased after COVID-19 and the premium appears to be $82.55 per square meter (column 2). Existing evidence suggests that semi-open spaces, like balconies in apartment units, have become more important for residents who have been forced to stay at home during COVID-19 (Aydin and Sayar, 2020; Molaei et al., 2021). Hence, a potential explanation for our result is that mobility restrictions and staying home longer during the lockdown period and beyond have increased demand for open air and better views from homes located on a higher floor level. At the same time, more activities done at home, such as WFH and home-based learning, have likely increased families’ desire for more separate spaces (Boesel et al., 2021; Kim, 2021).

On the other hand, while the proximity to subway stations generally has a positive effect on price changes, units located within 400 meters of the nearest subway station have experienced a reduction of $49.46 per square meter after COVID-19 (column 3). Similarly, units located within 2 kilometers of the CBD enjoyed higher premiums before the pandemic, but these premiums have almost disappeared afterward (column 4). This reduction in price premiums associated with locational attributes appears to come mainly from trading these attributes for other features. Our additional analysis (not shown) suggests that units with decreased price premiums after COVID-19 are likely to have fewer rooms or be on a lower floor level while being located closer to subway stations or the CBD. In other words, homebuyers have transferred some portion of their premium payment from locational advantages to higher floors and more rooms after COVID-19.

These results are consistent with anecdotal evidence that lifestyle changes such as WFH as a default mode and Singapore’s strict mobility restrictions reduced the desire to live closer to public transportation and the CBD (Soo, 2022). Singapore primarily consists of high-density urban areas. However, our findings

### Table 2: Regression Results

<table>
<thead>
<tr>
<th>Dependent Variable: resale price per square meter</th>
</tr>
</thead>
<tbody>
<tr>
<td>Unit characteristics</td>
</tr>
<tr>
<td>Year &amp; Month FE</td>
</tr>
<tr>
<td>street name FE</td>
</tr>
<tr>
<td>N</td>
</tr>
<tr>
<td>adjusted-R²</td>
</tr>
</tbody>
</table>

* p < 0.1; ** p < 0.05; *** p < 0.01. Robust standard errors in parentheses.

**Note:** For public housing in Singapore, the living room is included in the number of rooms. So we essentially capture whether the unit has three or more bedrooms.
suggest that changes in housing market trends are similar to other large countries like the US, where many households moved to suburban areas and home prices near transit fell in the post-COVID period (Boesel et al., 2021; Liu and Su, 2021; Allcott et al., 2021; Rosenthal et al., 2022).

4.2. How Online Listing Keywords Capture Consumer Preferences and Market Trends

We now test whether the advertised keywords on the online listing platform captured the housing market trend and how these keywords uncover changes in other dimensions that we were unable to find with conventional transaction data. We begin with the word-cloud analysis that allowed us to explore frequently appearing groups of words before and after the COVID-19 outbreak (Fig. 1). In line with the classic hedonic model results, the number of rooms, floor level (e.g. high floor), and proximity to public transportation have been frequently advertised. A new finding is that qualitative features like view and maintenance are also popular keywords, and they tend to be shown more frequently in advertisements in the post-COVID-19 period. We were also able to find how the units that are almost identical in terms of address, floor level, and floor area are advertised with different keywords before and after the COVID-19 outbreak. As shown in Fig. 2, while units in the same building, on the same floor, and with the same size were advertised for their convenient location and short distance to subway (i.e. Mass Rapid Transit, MRT) stations in 2019, they were advertised for their high floor with unblocked view and spaciousness in the post-COVID-19 period.

To provide more concrete analysis results, we next perform unigram and bigram text analyses and report the percent change in the relative frequency of advertised keywords between pre- and post-COVID-19 periods. First, both Fig. 3 and Fig. 4 suggest that view-related keywords appeared in online listings more frequently after the COVID-19 outbreak. For example, a single word of “view” and dual words directly indicating a good view such as “panoramic view”, “river view”, “view (is) beautiful”, “greenery view”, and “open view” have significantly increased their appearance in listings in 2020 and 2021 compared to 2019. In addition, less direct words such as “top”, “unblock”, and “balcony” also increasingly appeared in listing headlines after the COVID-19 outbreak (Fig. 3). Because units on the higher floor level are more likely to feature a better view than lower-level units, this result is consistent with the previous finding that higher-level units experienced increased price premiums after COVID-19 (Table 2, column 1). Figure 4 further shows that the probability of being advertised with “top floor” increased by more than 30 percent after COVID-19, whereas the keywords of “low”, “ground”, and “mid-floor” appeared much less frequently.

Next, in terms of space-related features, unigram analysis results show that 4- and 5-rooms became popular keywords for units in 2020 and 2021 compared to 2019, while having 2 rooms is advertised a lot less (Fig. 3). In addition, keywords suggesting a better living environment, such as “work” and “quiet”, appear more frequently after COVID-19. Bigram analysis results similarly show that keywords indicating more separate spaces within the housing unit with better housing quality—such as “plus study”, “renovated 5room”, and “spacious 4room”—appeared a lot more frequently in the post-COVID-19 periods (Fig. 4). These keywords directly reflect the increasing popularity of public housing units with more rooms
shown in Table 2 (column 2). As mentioned above, we believe that this is mainly caused by the increasing WFH trend and higher demand for homes to accommodate diverse activities like home-schooling or teleworking.

Finally, both unigram and bigram analyses demonstrate that location-related words appear less often in the headline of online listings after the COVID-19 outbreak. For example, single words presenting the distance or certain location such as “km”, “connectivity”, “proximity”, “station”, “mrt”, and “central”, appear less frequently after COVID-19 (Fig. 3). We also discover that dual words presenting public transportation accessibility, such as “mrt walk”, “opposite mrt”, “distance (to) mrt”, and “mrt station”, appear less frequently (Fig. 4). This downward trend is directly linked to a significant reduction in the price premium for proximity to public transportation and the CBD after COVID-19 shown in Table 2 (columns 3 and 4). While highlighting the convenient and accessible locations used to be critical for online advertisement, real estate agents have likely changed their strategies to adapt to new market trends.

These results suggest that trends in online listing keywords are very consistent with our earlier findings based on actual transactions. In addition, keywords from the online listings allow us to understand qualitative features associated with housing market changes that cannot be uncovered by standard hedonic pricing models using the transaction data. For example, although we find increased premiums for units on higher floor levels in Table 2, we do not fully understand the reasons. Our text analyses add insights that households have a growing desire for an open and unobstructed view, which is closely associated with the floor level, after the COVID-19 outbreak.

4.3. Temporal Trends of Price Premiums and Advertised Keywords

How quickly do these keywords capture housing market changes? As shown in Panel A of Fig. 5, the price premiums for the units with more rooms and on a higher floor level have significantly increased from the 2nd quarter of 2020. We believe that this is related to strict COVID-19 measures in Singapore, because the 2nd and 3rd quarters of 2020 were during and right after the lockdown (7 April–1 June, 2020). Anecdotal evidence from several real estate agents suggests that many customers were looking for new units with more space and a better view even during the lockdown period. These home-seekers have likely acted quickly and changes in their preferences would have been reflected in changes in price premiums for associated housing features.

At the same time, the real estate agents should have learned consumer preference changes from their interaction with customers as well as the housing transaction database, and reflected them when they put up the online advertisement. In this regard, we find that the shares of online listings highlighted with space- and view-related keywords increased significantly in the 3rd and 4th quarter of 2020, respectively (Panel B, Fig. 5). We note that we do not have the listing data for the 2nd quarter so our observation can only begin from the 3rd quarter. Considering the slightly lagged release of transaction data and a substantially smaller number of transactions in the 2nd quarter of 2020 (3,225 in the 2nd quarter vs. 7,365 in the 3rd quarter), we believe that agents respond to market trends in a timely manner.
For the location characteristics, the price premiums for units that are proximate to a subway station and the CBD show significantly decreasing trends from the 3rd and 4th quarter of 2020, respectively (Panel A, Fig. 5). The reduction on the share of listings highlighted with public transport–related keywords became significant from the 3rd quarter of 2020 (Panel B), which coincides with the time when the price premium for the proximity to the closest subway station became negative. On the other hand, central-location keywords show a less vivid change in their presence in the listing headlines over time, compared with other keywords that fluctuate more substantially after the COVID-19 outbreak. When we look deeper into the data, we find a substantial increase in the share of listings that are proximate to the CBD, from 4.7% in the 4th quarter of 2019 to 7.4% in the 3rd quarter of 2020. This suggests that even with an increased supply of resale public housing units in the central locations, the share of relevant keywords remained at a similar level. In other words, the probability of presenting central-location keywords decreased from the 3rd quarter of 2020. Overall, we claim these results are suggestive evidence that online listing keywords capture dynamic changes in housing market trends earlier than the transaction outcomes or at least in a timely manner.

5. Conclusions

The dynamic interrelationship between online listing keywords and consumer behaviors manifested in actual transactions has important implications on housing markets, especially when there is a strong shock and dynamic changes in the markets. Our analyses linking the keywords of online listings with housing price premiums from the transaction database provide a promising opportunity to estimate such interrelationships more accurately. Using the difference-in-difference model and big data of resale public housing transactions in Singapore, we first find that housing units on a higher floor level and with more rooms have experienced a significant increase in transaction prices while proximity to public transport and the CBD leads to the lower prices after COVID-19. Hence, our results confirm the external validity beyond the lower-density, Western context of a shifting trend of housing demand from smaller units in the urban area to larger units in the suburban area (e.g. Liu and Su, 2021; Allcott et al., 2021; Rosenthal et al., 2022).

Our text analysis results suggest that the online listing keywords have quickly captured the above trends based on big data from the major online platform in Singapore. Consistent with the results on transaction price premiums, both unigram and bigram analyses suggest a significant increase in the frequency of keywords related to view and more space while showing a decrease in the frequency of keywords highlighting the proximity to public transportation and a central location. These keywords show similar trends earlier than actual transaction outcomes, or at least in a timely manner. Moreover, our text analysis results provide a comprehensive understanding that could not be learned from the conventional database. For example, our evidence that view-related keywords increasingly appear on online listing headlines after COVID-19 gives an additional explanation of why units with a higher floor level have become more expensive among post-COVID-19 transactions.
These findings have important industry and policy implications. On the one hand, the big data for housing, especially online listing platforms, could be utilized to forecast changes in household housing demand and evaluate the dynamic trend of real estate markets. These data could easily be accessed on a real-time basis so they become publicly available a lot earlier than administrative data that government agencies collect and report several quarters or years later. Therefore, the cycle of capturing market trends could be much shorter, and the real estate industry could adjust in a timely manner. For example, real estate developers and brokers could be better tailored to the preferences of consumers. In particular, the capacity of real estate agents to quickly capture and highlight these features is likely to play an important role in their performance in a dynamically changing environment. Likewise, urban planners and policymakers could utilize the data to understand changes in citizens’ housing demand, especially for public housing. In countries like Singapore where the majority of households reside in public housing or public policies drive private housing markets, the government’s plan has a strong impact on the size, location, and density of future housing development. If planners do not take into account market trends and customers’ preferences, it would have detrimental effects on the everyday lives of citizens.

**Abbreviations**

CBD  
Central Business District  
HDB  
Housing and Development Board  
DID  
Difference-in-Difference  
WFH  
Work-from-home  
NAR  
National Association of Realtors  
CPI  
Consumer Price Index  
NLP  
Natural Language Processing  
MRT  
Mass Rapid Transit

**Declarations**

**Ethics approval and consent to participate**

Not applicable.

**Consent for publication**
Not applicable.

**Availability of data and materials**

The datasets generated and analyzed during the current study are available from the corresponding author on reasonable request.

**Competing interests**

The authors declare that they have no competing interests.

**Funding**

Funding is provided by the Singapore Ministry of Education (R-297-000-146-115).

**Authors’ contributions**

Jieun Lee prepared the data and performed analyses. Kwan Ok Lee shaped the research framework and wrote the main manuscript text. All the authors read and approved the final manuscript.

**Acknowledgements**

We deeply thank Dr. Filip Biljecki for the help to collect online listing data.

**References**


Figures
Figure 1

Word Clouds with Listing Keywords based on before (a) and after (b) the COVID-19
Figure 2

Example Cases for Keywords Changes for Similar Units
Figure 3

Unigram Analysis Results
Figure 4

Bigram Analysis Results
Figure 5

Trends for Price Premiums (**A**) and Share of listing used related Keywords (**B**)
Note 1: Panel A shows the trend of price premiums with the event study analyses focusing on how the price premium for each feature changed in each quarter compared to the 1st quarter of 2019 as a baseline. Panel B shows the trend of the share of listings advertised with each keyword category at each data collection point.

Note 2: The keywords we used for each category in Panel B are as follows: (1) ‘view’ keywords include river view, seaside view, greenery view, open view, beautiful view, garden view, superb view, park view, nice view, unblock view, panoramic view, view beautiful, view good, great view, high floor, balcony; (2) ‘rooms/space’ keywords include plus study, sized bedroom, study room, bedroom plus, spacious bedroom, bedroom spacious, 4 full bedroom, 5 full bedroom, 5room, 4room, 5bedroom, 4bedroom, 3bedroom, 5i, 5a, 4a; (3) ‘public transport’ keywords include mrt, lrt, bus, train, station; (4) ‘central location’ keywords include central, centre, cbd, central district names(e.g. raffles), and central district numbers(e.g. district 1, district 2, ..., district 14).

Note 3: For the online listings, we collected data at 10 distinct time points as follows: October 28, 2019; September 29, 2020; November 2, 2020; December 4, 2020; January 14, 2021; February 21 2021; March 26, 2021; March 29, 2021; May 27, 2021; and July 19, 2021.