



[DATA ARTICLE TEMPLATE V.18 (APRIL 2024)]

## ARTICLE INFORMATION

### Article title

From Archives to AI: Residential Property Data Across Three Decades in Brunei Darussalam

### Authors

Haziq Jamil\*,<sup>1</sup>, Amira Barizah Noorosmawie<sup>1</sup>, Hafeezul Waezz Rabu<sup>1</sup>, Lutfi Abdul Razak<sup>‡</sup>Razak<sup>2</sup>

### Affiliations

<sup>1</sup>Mathematical Sciences, Faculty of Science, Universiti Brunei Darussalam, Jalan Tungku Link, Bandar Seri Begawan BE1410, Brunei Darussalam

<sup>2</sup>UBD School of Business and Economics, Universiti Brunei Darussalam, Jalan Tungku Link, Bandar Seri Begawan BE1410, Brunei Darussalam

### Corresponding author's email address and Twitter handle

[haziq.jamil@ubd.edu.bn](mailto:haziq.jamil@ubd.edu.bn)

### Keywords

*Housing Market; Property Listings; Spatial Data; Web Scraping; Large Language Models; Brunei.*

### Abstract

This article introduces the first publicly available data set for ~~analyzing~~analysing the Brunei housing market, covering ~~31,495~~more than 30,000 property listings from 1993 to ~~2024~~early 2025. The data set, curated from property advertisements in newspapers and online platforms, includes key attributes such as price, location, property type, and physical characteristics, enriched with area-level spatial information. Comprehensive and historical, it complements the Brunei Darussalam Central Bank's Residential Property Price Index (RPPI), addressing the limitations of restricted access to raw RPPI data and its relatively short timeline since its inception in 2015. Data collection involved manual transcription from archival sources and automated web scraping using programmatic methods, supported by innovative ~~eleaning~~processing with Large Language Models (LLMs) to ~~structure~~unformatted~~codify unstructured~~ text. The data set enables spatial and temporal analysis, with potential applications in economics, urban planning, and real estate research. ~~While minor~~limitations exist, such as missing ~~Although listing prices are only a proxy for market~~ values and ~~spatial coverage bias toward the Brunei Muara district~~may deviate from actual sale prices due to negotiation dynamics and other factors, this data set still provides a ~~robust foundation for exploring~~valuable resource for quantitative analyses of housing market trends and for informing policy decisions.



## 33 SPECIFICATIONS TABLE

34

Subject	Real Estate Economics
Specific subject area	Residential property data across three decades in Brunei for spatial, temporal, and economic analysis.
Type of data	Table (Raw) in Comma Separated Values (CSV) format.
Data collection	The data were collected via manual transcription from newspaper advertisements and automated web scraping using R software (rvest package). Large Language Models ( <a href="#">Llama 3.1</a> <a href="#">DeepSeek R1</a> <a href="#">Distilled Qwen 14B</a> via <a href="#">tidychatmodels in R Ollama</a> ) were employed to <a href="#">clean</a> <a href="#">process</a> unstructured text into structured formats. Inclusion criteria focused on property listings with price, location, and type, while duplicates were removed. Spatial data were <a href="#">harmonized</a> <a href="#">harmonised</a> to match official administrative boundaries for consistency.
Data source location	The data were collected in Brunei Darussalam and are stored in a public GitHub repository.
Data accessibility	Repository name: GitHub/Bruneiverse Data identification number: N/A Direct URL to data: <a href="https://bruneiverse.github.io/house-data/data/hspbn_2024-12-12.csv">https://bruneiverse.github.io/house-data/data/hspbn_2024-12-12.csv</a>
Related research article	None.

35

## 36 VALUE OF THE DATA

- 37 • [First](#)[A first-of-its-kind](#) [data set](#) [for Brunei](#). This data set [enables comprehensive spatial and](#)  
38 [temporal analysis of](#)[is the](#) [Brunei housing market](#). [To our knowledge, no other](#)[first](#) publicly  
39 available [data set exists for this purpose, making it a significant contribution to housing](#)  
40 [market research](#). [compilation of Brunei's residential property data, covering 31,116 listings](#)  
41 [from 1993 to early 2025](#). Previous studies on Brunei's housing have [been limited to either](#)  
42 [relied on](#) recent or non-spatial data [1–3,2], or [have primarily employed](#) qualitative [in](#)  
43 [nature](#)[methods](#) due to the [lack](#)[absence](#) of a-structured data set [3,4,5]. This aligns with the  
44 [growing trend of using](#)[]. Researchers may use this data [analytics in the real estate industry](#)  
45 [\[6,7\]](#), [to explore temporal variations in property listings and to compare housing](#)  
46 [characteristics across different regions within Brunei](#).



- 47     • **Historical and spatial insights.** Covering data from 1993 to 2024 Spanning over three  
48        decades, this data set allows for the study enables analysis of long-term housing trends in  
49        Brunei. The spatial information available enables analysis at various administrative levels in  
50        Brunei, providing insights into local patterns and urban development patterns, regional  
51        differences, and the evolution of housing characteristics over time. This historical depth is  
52        particularly valuable given the lack of previous data before prior to the establishment of the  
53        Residential Property Price Index (RPPI) [85] in 2015.
- 54     • **Influence on economic and monetary policies.** Analysing the real estate market is crucial  
55        because the RPPI can potentially play a key role in shaping monetary policy and assessing  
56        economic stability. Changes in RPPI signal inflationary pressures, providing guidance to the  
57        central bank decisions. Developing a house price index using advertised prices, as explored  
58        by [9], aligns with practices in other countries, such as the UK's House Price Index developed  
59        by Rightmove PLC [10,11]. This data set demonstrates how computational methods can  
60        automate what is typically a time-consuming and labour-intensive process.
- 61     • **Methodological innovation in data curation.** The data collection process employs a unique  
62        combination of manual transcription from archival sources, automated web scraping, and AI-  
63        based data cleaning using Large Language Models (LLMs). This multi-method approach offers  
64        a reproducible framework for assembling complex data sets, which can be adapted for  
65        similar data collection efforts in other domains, thereby advancing best practices in the field.
- 66     • **Opportunities for handling of missing data.** While the data set has complete information for  
67        key variables such as price, date, and spatial variables, The data set includes instances of  
68        missing information on data, particularly in house characteristics creates, which present  
69        opportunities for further research. Evidently, house characteristics are inherently correlated  
70        (see Figure 6), so imputation from observed correlations seems highly promising. With the  
71        spatial information present, this opens avenues on data imputation techniques. Researchers  
72        may explore advanced statistical and machine learning methods to address missing values  
73        and improve data reliability. The structured format and detailed documentation of the data  
74        set support such methodological investigations, making it a valuable testbed for developing  
75        methodological and evaluating new approaches for handling missing data, encouraging  
76        innovation and exploration incomplete data in this area.
- 77     • **Application of Large Language Models (LLMs).** This project highlights an innovative use of  
78        Large Language Models (LLMs) for data cleaning. The LLM was employed to extract  
79        structured information from unstructured text descriptions, achieving an accuracy rate of  
80        93%. This method significantly reduced the time and effort required for data cleaning. It can  
81        also be applied to other data sets with unstructured text, such as social media or document  
82        archives, to extract valuable information for analysis real-world data sets.

## 83     BACKGROUND

84     The housing market is a key indicator of economic health and social well-being, yet comprehensive  
85        and publicly accessible data sets in Brunei remain limited. As far as we are aware To the best of our  
86        knowledge, this is the first data set of its kind in Brunei, motivated by the need to fill the gap in  
87        publicly available data on the local housing market data.



88 Currently, the Brunei Darussalam Central Bank (BDCB) produces a Residential Property Price Index  
89 (RPPI) [85] using data sourced from financial institutions, such as bank loan data sets. While the RPPI  
90 is published quarterly, the underlying raw data is not publicly available due to privacy restrictions.  
91 This limits research opportunities and transparency in understanding broader housing market trends.  
92 Furthermore, since the RPPI only began in 2015, historical housing data for Brunei is lacking.  
93 We address these challenges by providing a cost-effective and timely means to collect and analyse  
94 housing market data. Covering records from 1993 onward, it offers historical depth that  
95 complements—and extends beyond—the RPPI. It is valuable not only for tracking property price trends  
96 but also for advancing research in economics, urban planning, and real estate, supporting informed  
97 decision-making across sectors.

## 98 DATA DESCRIPTION

99 The data has been curated into a single Comma-Separated Values (CSV) file named hspbn\_2024-12-  
100 12.csv. The data set contains 31,495,116 property listing records which are enriched with area-level  
101 geotagged spatial information, spanning a period of 32 years from Mar 1993 to Dec 2024 Feb 2025.  
102 The 18 columns of this data set capture information for each property listing as detailed in [Table 1](#)  
103 below.

Table 1: Codebook for the house price data set.

Variable	Type	Details
1 id	Integer	Unique identifier for each property listing.
2 date	Date	Date when the property listing was collected.
3 quarter	Date	Quarter of the listing date in the format YYYY Qq (e.g., 2016 Q3).
4 kampong	Spatial Area	The village where the property is located.
5 mukim	Spatial Area	The sub-district administrative area where the property is located.
6 district	Spatial Area	The main district where the property is located.
7 price	Numeric	Listing price of the property in Brunei Dollars (BND).
8 type	Character	Type of property. One of "Detached", "Semi-Detached", "Terrace", "Apartment", or "Land".
9 tenure	Character	The land tenure for the property. One of "Freehold", "Leasehold", or "Strata".
10 status	Character	Current status of the listing. One of "Proposed", "Under Construction", "New", or "Resale".
11 plot_area	Numeric	Total area of the land plot in acres.
12 floor_area	Numeric	Built up floor area of the property in square feet.
13 storeys	Integer	Number of storeys or floors in the property.
14 beds	Integer	Number of bedrooms in the property.
15 baths	Integer	Number of bathrooms in the property.



16	agent	Character	Anonymised identifier of the real estate agent or agency handling the listing.
17	source	Character	Source of the listing.
18	method	Character	Method of data collection.

104

## 105 Property Characteristics

106 The data set includes a range of property characteristics suitable for exploring the relationship  
107 between property attributes and prices. This section clarifies and provides context for the key  
108 variables in the data set.

109 Brunei's private residential property market offers a variety of options, including detached houses,  
110 townhouses, and apartments [54]. Based on this [diversity](#), property types have been categorised into  
111 four main groups—Detached, Semi-Detached, Terrace, and Apartment—to accurately reflect the  
112 [diversityrange](#) of [property types](#) [residential properties](#) in Brunei. Additionally, there are a small  
113 number of records that reflect listings for land, which are categorised accordingly as "Land".

114 Property tenure refers to the legal terms under which a person holds ownership or occupancy rights  
115 to a property. In Brunei, property tenure can be classified into three main categories: Freehold (in  
116 perpetuity), Leasehold, and Strata. The latter two refer to a limited time-limited ownership, although  
117 details about the remaining duration of the tenure are almost never included in property listings.  
118 Strata titles differ from Leasehold titles in that they grant ownership of a specific portion of a  
119 property, such as an apartment, while sharing ownership of common areas.

120 The data set also includes information on the status of the property listing, indicating whether the  
121 advertisement refers to a proposed development, a newly completed development, or a property  
122 being resold. This categorical variable may be useful for analysing price differences across different  
123 types of listings. While the exact age of properties being resold would be invaluable for such  
124 analyses, this information is rarely included in advertisements. Instead, the listing status may serve as  
125 a useful proxy for property age.

126 The numerical variables in the data set are plot area, floor area, storeys, beds, and baths, each  
127 providing information on the physical attributes of the property. Note that plot area is measured in  
128 acres, while floor area in square feet, as these are the units most familiar and commonly used in  
129 Brunei. Users of this data set may choose to convert these units as needed for their analysis.

130 Finally, metadata about the property is included for transparency and informational purposes. The  
131 variable agent specifies the (anonymised) identifier of the real estate agent or agency responsible for  
132 the listing, while source identifies the platform or medium from which the listing was obtained, such  
133 as a newspaper, magazine, or website. The method variable details the data collection approach,  
134 which is further elaborated in the section below.

135



Table 2: Summary of housing data.

Variable	N	Overall N = 31,495 <sup>‡</sup>	Brunei Muara N = 28,894 <sup>‡</sup>	Belait N = 1,513 <sup>‡</sup>	Tutong N = 790 <sup>‡</sup>	Temburong N = 298 <sup>‡</sup>
<b>Price (BND 1,000)</b>	31,495					
— Mean (SD)	340 (380)	339 (392)	372 (209)	260 (87)	419 (323)	
— Min – Max	70 – 13,800	70 – 13,800	98 – 2,800	116 – 680	118 – 1,800	
— Median (Q1, Q3)	285 (230, 380)	285 (230, 380)	320 (268, 400)	245 (198, 310)	390 (250, 430)	
<b>Property type</b>	27,592					
— Detached	17,685 (64%)	16,548 (65%)	524 (41%)	532 (75%)	81 (56%)	
— Semi-Detached	3,808 (14%)	3,574 (14%)	97 (7.6%)	130 (18%)	7 (4.8%)	
— Terrace	4,502 (16%)	4,183 (16%)	219 (17%)	46 (6.5%)	54 (37%)	
— Apartment	1,582 (5.7%)	1,151 (4.5%)	424 (33%)	4 (0.6%)	3 (2.1%)	
— Land	15 (<0.1%)	10 (<0.1%)	4 (0.3%)	1 (0.1%)	0 (0%)	
<b>Land tenure</b>	13,064					
— Freehold	9,398 (72%)	8,477 (75%)	381 (33%)	396 (80%)	144 (97%)	
— Leasehold	2,850 (22%)	2,273 (20%)	477 (41%)	96 (19%)	4 (2.7%)	
— Strata	816 (6.2%)	516 (4.6%)	297 (26%)	3 (0.6%)	0 (0%)	
<b>Development status</b>	22,831					
— Proposed	3,902 (17%)	3,562 (17%)	101 (8.4%)	195 (32%)	44 (30%)	
— Under Construction	9,715 (43%)	8,856 (42%)	553 (46%)	264 (43%)	42 (29%)	
— New	8,011 (35%)	7,389 (35%)	428 (36%)	135 (22%)	59 (40%)	
— Resale	1,203 (5.3%)	1,061 (5.1%)	116 (9.7%)	25 (4.0%)	1 (0.7%)	
<b>Plot area (acres)</b>	23,581					
— Mean (SD)	0.16 (0.12)	0.15 (0.11)	0.19 (0.15)	0.18 (0.17)	0.23 (0.21)	
— Min – Max	0.01 – 2.00	0.01 – 1.63	0.01 – 1.01	0.04 – 2.00	0.05 – 0.96	
— Median (Q1, Q3)	0.13 (0.08, 0.19)	0.13 (0.08, 0.19)	0.13 (0.06, 0.27)	0.14 (0.10, 0.20)	0.16 (0.13, 0.26)	



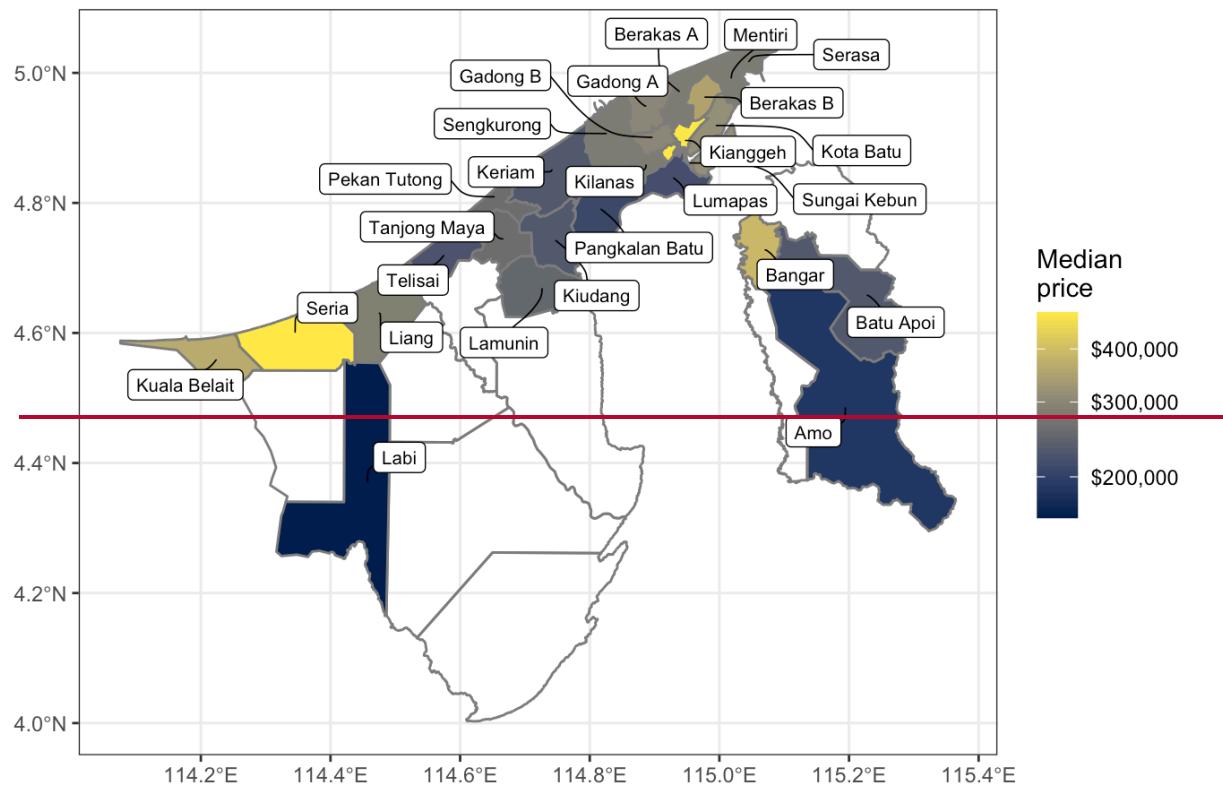
Floor area (sq. ft.)	16,863				
— Mean (SD)	2,590 (1,045)	2,616 (1,061)	2,418 (914)	2,130 (629)	2,796 (741)
— Min — Max	500 — 14,411	500 — 14,411	600 — 7,500	1,100 — 7,000	950 — 3,700
— Median (Q1, Q3)	2,410 (2,000, 3,000)	2,435 (2,000, 3,000)	2,207 (1,700, 2,800)	2,013 (1,826, 2,450)	3,031 (2,790, 3,229)
Number of storeys	13,797				
— 1	1,709 (12%)	1,472 (12%)	160 (35%)	70 (16%)	7 (4.2%)
— 2	11,420 (83%)	10,630 (83%)	280 (61%)	365 (84%)	145 (87%)
— 3+	668 (4.8%)	631 (5.0%)	20 (4.3%)	2 (0.5%)	15 (9.0%)
Number of bedrooms	26,968				
— Mean (SD)	4.2 (0.9)	4.2 (0.9)	4.0 (1.1)	3.9 (0.7)	4.7 (1.0)
— Min — Max	0.0 — 12.0	0.0 — 12.0	1.0 — 10.0	2.0 — 7.0	2.0 — 7.0
— Median (Q1, Q3)	4.0 (4.0, 5.0)	4.0 (4.0, 5.0)	4.0 (3.0, 4.0)	4.0 (3.0, 4.0)	5.0 (4.0, 5.0)
Number of bathrooms	19,957				
— Mean (SD)	3.7 (1.2)	3.7 (1.2)	3.3 (1.1)	3.3 (1.0)	3.2 (1.5)
— Min — Max	1.0 — 11.0	1.0 — 11.0	1.0 — 8.0	1.0 — 7.0	1.0 — 5.0
— Median (Q1, Q3)	3.0 (3.0, 4.0)	3.0 (3.0, 4.0)	3.0 (3.0, 4.0)	3.0 (2.0, 4.0)	2.0 (2.0, 5.0)

<sup>a</sup>n (%)

## 138 Spatial Information

139 In Brunei Darussalam, the administrative areas are organised hierarchically into three levels. At the  
140 smallest level is the *kampong*, the Malay word for village. While a typical village refers to a traditional  
141 rural settlement, it is also used to describe an urbanised area located within or near the capital city  
142 or a town. It may even refer to a part of public housing estates. Several kampongs grouped together  
143 form a *mukim*, which serves as a sub-district administrative area. Finally, multiple mukims are nested  
144 within a *district*, the largest administrative unit, of which Brunei has four: Brunei-Muara, Belait,  
145 Tutong, and Temburong. In our data set, each property listing is associated with a specific kampong,  
146 mukim, and district, allowing for spatial analysis at different scales.

147 Importantly, the names of the kampongs, mukims, and districts have been harmonised with a  
148 standardised naming convention to ensure consistency across the data set. This also allows for ease  
149 of integration into Geographic Information Systems (GIS) software for spatial analysis and  
150 visualisation, namely the {bruneimap} R package [426].



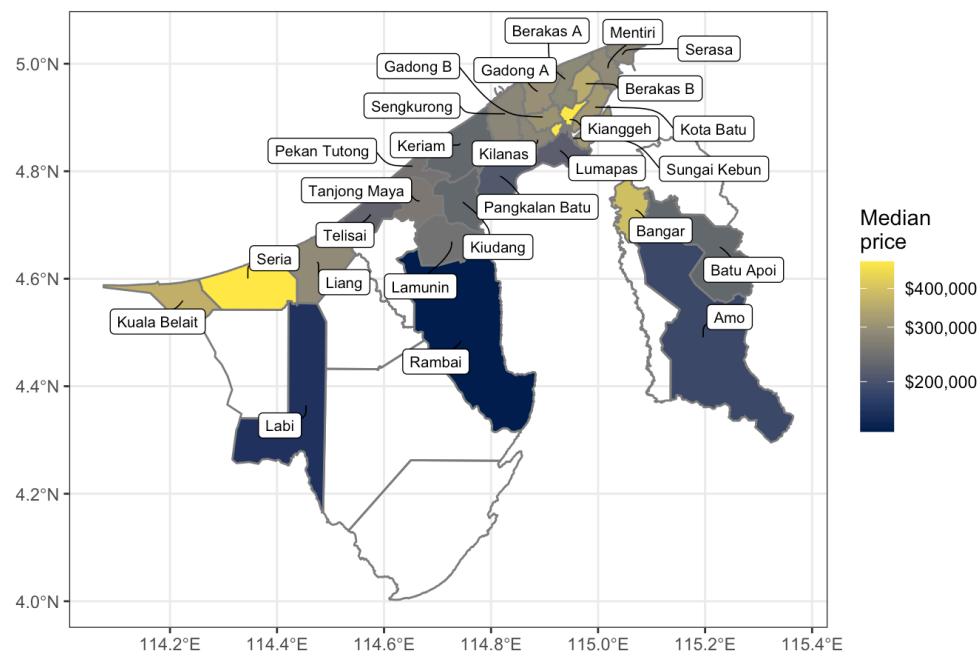


Figure 1: Spatial distribution of median property prices by mukim.



Table 2: Summary of housing data.

Variable	N	Overall N = 31,116 <sup>1</sup>	Brunei-Muara N = 28,570 <sup>1</sup>	Belait N = 1,484 <sup>1</sup>	Tutong N = 767 <sup>1</sup>	Temburong N = 295 <sup>1</sup>
<u>Price (BND 1,000)</u>	<u>31,116</u>					
Mean (SD)		<u>340 (381)</u>	<u>340 (393)</u>	<u>372 (208)</u>	<u>259 (87)</u>	<u>421 (324)</u>
Min - Max		<u>70 - 13,800</u>	<u>70 - 13,800</u>	<u>98 - 2,800</u>	<u>116 - 680</u>	<u>118 - 1,800</u>
Median (Q1, Q3)		<u>288 (230, 380)</u>	<u>285 (230, 380)</u>	<u>320 (268, 400)</u>	<u>245 (198, 310)</u>	<u>390 (250, 430)</u>
<u>Property type</u>	<u>27,231</u>					
Detached		<u>17,416 (64%)</u>	<u>16,307 (65%)</u>	<u>520 (42%)</u>	<u>509 (74%)</u>	<u>80 (56%)</u>
Semi-Detached		<u>3,823 (14%)</u>	<u>3,591 (14%)</u>	<u>97 (7.8%)</u>	<u>128 (19%)</u>	<u>7 (4.9%)</u>
Terrace		<u>4,449 (16%)</u>	<u>4,134 (16%)</u>	<u>213 (17%)</u>	<u>48 (7.0%)</u>	<u>54 (38%)</u>
Apartment		<u>1,527 (5.6%)</u>	<u>1,106 (4.4%)</u>	<u>414 (33%)</u>	<u>4 (0.6%)</u>	<u>3 (2.1%)</u>
Land		<u>16 (&lt;0.1%)</u>	<u>11 (&lt;0.1%)</u>	<u>4 (0.3%)</u>	<u>1 (0.1%)</u>	<u>0 (0%)</u>
<u>Land tenure</u>	<u>12,877</u>					
Freehold		<u>9,296 (72%)</u>	<u>8,405 (76%)</u>	<u>368 (33%)</u>	<u>381 (80%)</u>	<u>142 (97%)</u>
Leasehold		<u>2,783 (22%)</u>	<u>2,221 (20%)</u>	<u>467 (41%)</u>	<u>91 (19%)</u>	<u>4 (2.7%)</u>
Strata		<u>798 (6.2%)</u>	<u>504 (4.5%)</u>	<u>291 (26%)</u>	<u>3 (0.6%)</u>	<u>0 (0%)</u>
<u>Development status</u>	<u>22,481</u>					
Proposed		<u>4,004 (18%)</u>	<u>3,660 (18%)</u>	<u>103 (8.8%)</u>	<u>197 (33%)</u>	<u>44 (31%)</u>
Under Construction		<u>9,420 (42%)</u>	<u>8,600 (42%)</u>	<u>535 (46%)</u>	<u>244 (41%)</u>	<u>41 (29%)</u>
New		<u>7,724 (34%)</u>	<u>7,122 (35%)</u>	<u>413 (35%)</u>	<u>132 (22%)</u>	<u>57 (40%)</u>
Resale		<u>1,333 (5.9%)</u>	<u>1,186 (5.8%)</u>	<u>120 (10%)</u>	<u>26 (4.3%)</u>	<u>1 (0.7%)</u>



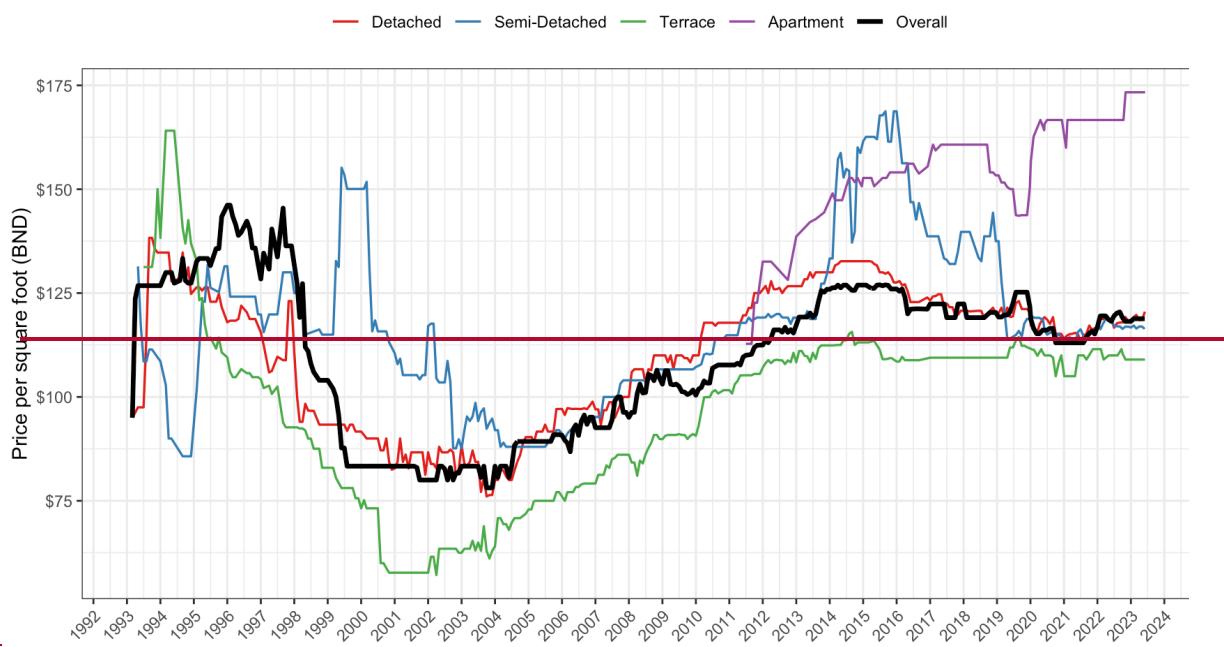
<u>Plot area (acres)</u>	<u>23,368</u>					
<u>Mean (SD)</u>		<u>0.16 (0.12)</u>	<u>0.15 (0.11)</u>	<u>0.19 (0.15)</u>	<u>0.18 (0.17)</u>	<u>0.23 (0.21)</u>
<u>Min - Max</u>		<u>0.01 - 2.00</u>	<u>0.01 - 1.69</u>	<u>0.01 - 1.01</u>	<u>0.04 - 2.00</u>	<u>0.05 - 0.96</u>
<u>Median (Q1, Q3)</u>		<u>0.13 (0.08, 0.19)</u>	<u>0.13 (0.08, 0.19)</u>	<u>0.13 (0.06, 0.27)</u>	<u>0.14 (0.10, 0.21)</u>	<u>0.16 (0.13, 0.26)</u>
<u>Floor area (sq. ft.)</u>	<u>16,665</u>					
<u>Mean (SD)</u>		<u>2,602 (1,047)</u>	<u>2,629 (1,062)</u>	<u>2,423 (913)</u>	<u>2,133 (651)</u>	<u>2,786 (751)</u>
<u>Min - Max</u>		<u>500 - 14,411</u>	<u>500 - 14,411</u>	<u>600 - 7,500</u>	<u>1,093 - 7,000</u>	<u>950 - 3,700</u>
<u>Median (Q1, Q3)</u>		<u>2,427 (2,000, 3,000)</u>	<u>2,465 (2,000, 3,000)</u>	<u>2,218 (1,800, 2,800)</u>	<u>2,013 (1,826, 2,450)</u>	<u>3,016 (2,790, 3,229)</u>
<u>Number of storeys</u>	<u>13,644</u>					
<u>1</u>		<u>1,700 (12%)</u>	<u>1,462 (12%)</u>	<u>160 (35%)</u>	<u>71 (17%)</u>	<u>7 (4.2%)</u>
<u>2</u>		<u>11,266 (83%)</u>	<u>10,493 (83%)</u>	<u>280 (61%)</u>	<u>348 (83%)</u>	<u>145 (87%)</u>
<u>3+</u>		<u>678 (5.0%)</u>	<u>642 (5.1%)</u>	<u>19 (4.1%)</u>	<u>2 (0.5%)</u>	<u>15 (9.0%)</u>
<u>Number of bedrooms</u>	<u>26,631</u>					
<u>Mean (SD)</u>		<u>4.2 (0.9)</u>	<u>4.2 (0.9)</u>	<u>4.0 (1.1)</u>	<u>3.9 (0.7)</u>	<u>4.7 (1.0)</u>
<u>Min - Max</u>		<u>0.0 - 12.0</u>	<u>0.0 - 12.0</u>	<u>1.0 - 10.0</u>	<u>2.0 - 7.0</u>	<u>2.0 - 7.0</u>
<u>Median (Q1, Q3)</u>		<u>4.0 (4.0, 5.0)</u>	<u>4.0 (4.0, 5.0)</u>	<u>4.0 (3.0, 4.0)</u>	<u>4.0 (3.0, 4.0)</u>	<u>5.0 (4.0, 5.0)</u>
<u>Number of bathrooms</u>	<u>19,694</u>					
<u>Mean (SD)</u>		<u>3.7 (1.2)</u>	<u>3.7 (1.2)</u>	<u>3.3 (1.1)</u>	<u>3.3 (1.0)</u>	<u>3.2 (1.5)</u>
<u>Min - Max</u>		<u>1.0 - 11.0</u>	<u>1.0 - 11.0</u>	<u>1.0 - 8.0</u>	<u>1.0 - 7.0</u>	<u>1.0 - 5.0</u>
<u>Median (Q1, Q3)</u>		<u>3.0 (3.0, 4.0)</u>	<u>3.0 (3.0, 4.0)</u>	<u>3.0 (3.0, 4.0)</u>	<u>3.0 (2.0, 4.0)</u>	<u>2.0 (2.0, 5.0)</u>

<sup>1</sup>n (%)

## 154 Listing Dates

155 The date variable ~~in the data set refers to~~represents the date on which the property listing was  
156 obtained. ~~It is important to note that this is not, rather than~~ the date ~~the property was sold, nor does~~  
157 ~~it necessarily reflect the precise timing of other transactions related to the property. Users~~sale or  
158 transaction, and should set their expectations accordingly, be interpreted as the primary purpose  
159 a snapshot of the date is to capture the state of the housing market conditions at a specific given point  
160 in time.

161 For analysis, we recommend aggregating data by quarters, as represented by the quarter variable.  
162 This aggregation helps address potential issues like missing data (see subsection below) and provides  
163 a more stable and robust representation of market trends, making it suitable for temporal analysis of  
164 the housing market.



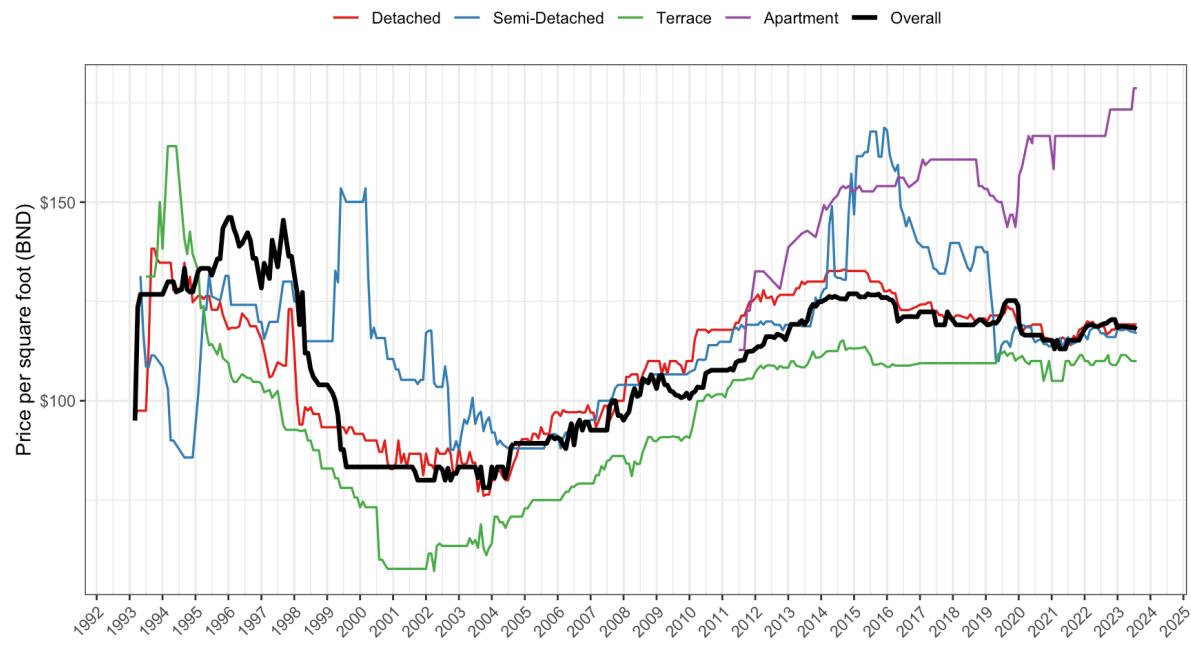


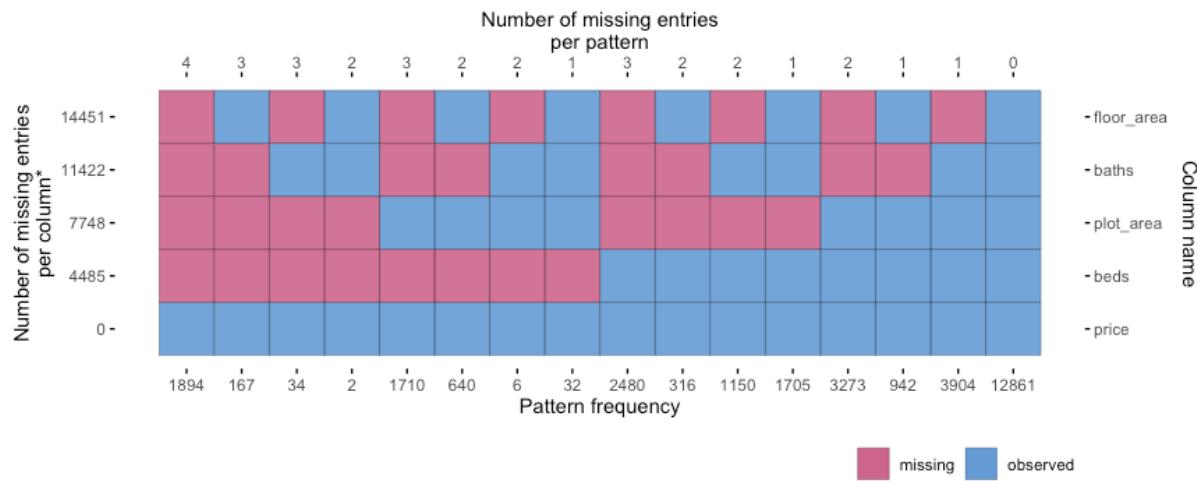
Figure 2: Median smoothed prices per square foot by property type using a 24-month (8-quarter) rolling window.

## 165 Missing Values

166 ThisIn any data collection effort, it is unsurprising to encounter missing data. Likewise in our data set  
167 endeavours to provide complete information regarding the listing date, spatial information,  
168 advertised price, and metadata. Nonetheless, we hereby report that missing values are  
169 presentoccur across variousseveral property characteristics, including type, tenure, status, such as  
170 plot area, floor area, storeys, beds, and baths. The reason for this is due to the nature, and others. A  
171 preliminary analysis of the property advertisements, which may missing data patterns indicates that  
172 the missingness is not alwayscompletely random, with certain variables displaying dependencies on  
173 others. Advertisers often include complete information when advertised byonly the information they  
174 deem most marketable or necessary, while other details may be omitted if they are considered  
175 standard or implied. For instance, a listing might specify the real estate agents, square footage and  
176 price but leave out the number of bedrooms and bathrooms, assuming that prospective buyers are  
177 able to infer these details.

178 Missing values are represented by blank cells in the CSV file, and the severity of missing  
179 valuesmissingness is summarisedsummarized in Table 3. In summarytotal, 10.51% of the records  
180 contain missing values for all key house characteristics (i.e. plot area, floor area, beds, and baths),  
181 which, depending on the research question, may necessitaterequire imputation or the exclusion of  
182 these records.

183



184

185 [Figure 3: Missing data patterns for key house characteristics.](#)186 [Comparison to RPPI Data](#)

187 To demonstrate the quality of the data set, we compared it with the Residential Property Price Index  
188 (RPPI) [8] published by the Brunei Darussalam Central Bank (BDCB). A simple median price per  
189 square foot (PPSF) index can be calculated by aggregating the data by quarters. This approach  
190 minimises the impact of missing values, as the index is based on aggregated data. [Figure 3](#) shows the  
191 comparison between the RPPI and the PPSF index calculated from our data set. The mean absolute  
192 error (MAE) between the two indices is calculated to be 4.66%, indicating a good level of agreement  
193 between the two data sets.

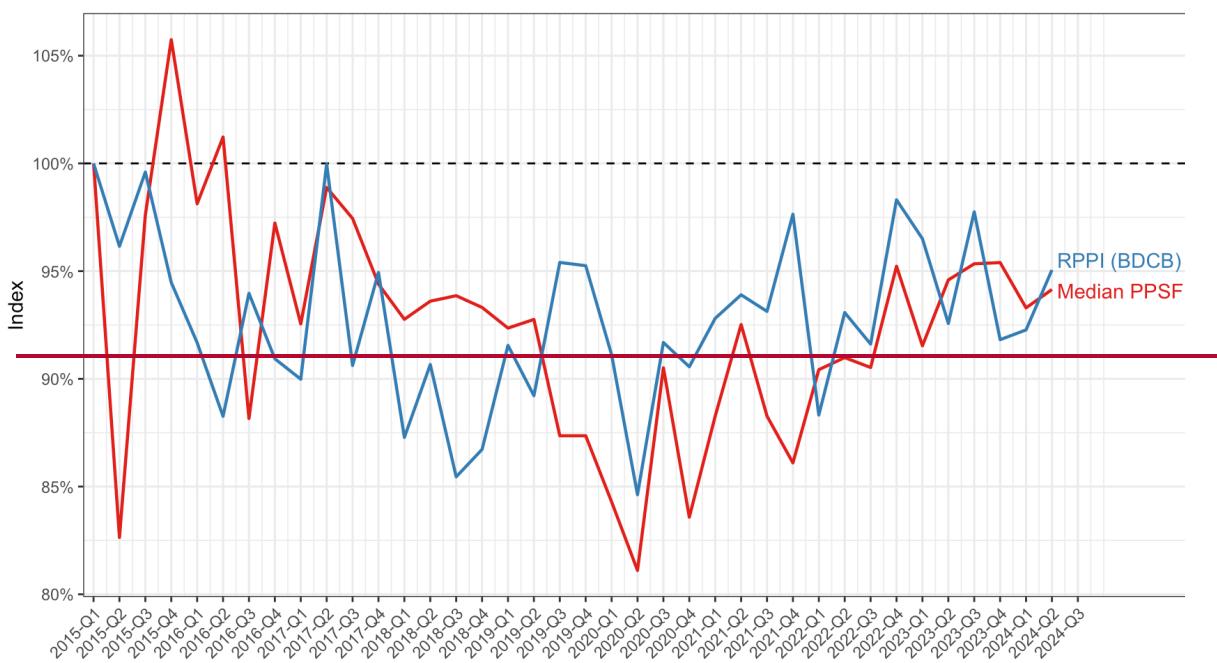




Figure 3: Comparison of quarterly median price per square foot indices (Median PPSF) and the official Residential Property Price Index (RPPI) from Brunei Darussalam Central Bank (BDCB).

194

## 195 EXPERIMENTAL DESIGN, MATERIALS AND METHODS

196 In this section we describe the data collection process, which involved either a manual transcription  
197 of property listings from newspapers, or web scraping of online property agent listings. The data  
198 collection method varied over the years due to the availability of data sources and the evolution of  
199 technology. For the later years, a large language model (LLM) was also employed to perform data  
200 cleaning on the web scraped data. [Table 3](#) details which method was used for each year in the data  
201 set, and whether the data was subjected to LLM post-processing.

202 All analyses were conducted using the R programming language [[137](#)], with specific packages used  
203 described in each subsection below.

204



Data  
in Brief

*Data in Brief*

*Open access*



Article template

Table 3: Data availability by year.

	Count	Missing data severity			Data source				LLM post-processing
		Spatial coverage (mukim) <sup>1</sup>	Property Type <sup>2</sup>	Property Characteristics <sup>3</sup>	National Archive	Online Archive	Web Scraping		
1993	<a href="#">417400</a>	<a href="#">23.1</a> <a href="#">33.3</a> %	0.0%	<a href="#">19.2</a> <a href="#">0</a> %	✓				
1994	<a href="#">654653</a>	<a href="#">35</a> <a href="#">51</a> .9%	65. <a href="#">7</a> <a href="#">8</a> %	<a href="#">28.0</a> <a href="#">27.9</a> %	✓				
1995	<a href="#">669668</a>	<a href="#">48.7</a> <a href="#">70.4</a> %	66.8%	<a href="#">21.2</a> <a href="#">3</a> %	✓				
1996	<a href="#">563561</a>	<a href="#">35</a> <a href="#">51</a> .9%	69. <a href="#">8</a> <a href="#">7</a> %	12.1%	✓				
1997	385	<a href="#">35</a> <a href="#">51</a> .9%	38.4%	26.8%	✓				
1998	345	<a href="#">33.3</a> <a href="#">48.1</a> %	36.8%	28.7%	✓				
1999	<a href="#">322317</a>	<a href="#">35</a> <a href="#">51</a> .9%	31. <a href="#">4</a> <a href="#">9</a> %	<a href="#">27.3</a> <a href="#">26.2</a> %	✓				
2000	<a href="#">379378</a>	<a href="#">43.6</a> <a href="#">63.0</a> %	0.8%	4.2%	✓				
2001	<a href="#">344342</a>	<a href="#">43.6</a> <a href="#">63.0</a> %	0.3%	2.3%	✓				
2002	<a href="#">443437</a>	<a href="#">43.6</a> <a href="#">63.0</a> %	0.0%	20. <a href="#">1</a> <a href="#">4</a> %	✓				
2003	<a href="#">454449</a>	<a href="#">46.2</a> <a href="#">66.7</a> %	0.0%	13. <a href="#">2</a> <a href="#">4</a> %	✓				
2004	<a href="#">442440</a>	<a href="#">43.6</a> <a href="#">63.0</a> %	0.0%	19. <a href="#">0</a> <a href="#">1</a> %	✓				
2005	<a href="#">496493</a>	<a href="#">46.2</a> <a href="#">66.7</a> %	0.0%	13. <a href="#">3</a> <a href="#">2</a> %	✓				
2006	<a href="#">661653</a>	<a href="#">41.0</a> <a href="#">59.3</a> %	0.2%	11.3%	✓				
2007	<a href="#">644638</a>	<a href="#">38.5</a> <a href="#">55.6</a> %	0.0%	<a href="#">13.0</a> <a href="#">12.9</a> %	✓				
2008	<a href="#">699687</a>	<a href="#">41.0</a> <a href="#">59.3</a> %	0.3%	5. <a href="#">7</a> <a href="#">8</a> %	✓				
2009	<a href="#">542531</a>	<a href="#">35</a> <a href="#">51</a> .9%	0.2%	<a href="#">3.9</a> <a href="#">4.0</a> %	✓				
2010	<a href="#">578571</a>	<a href="#">38.5</a> <a href="#">55.6</a> %	0.0%	2.1%	✓				
2011	<a href="#">605594</a>	<a href="#">38.5</a> <a href="#">55.6</a> %	0.2%	<a href="#">9.9</a> <a href="#">10.1</a> %	✓				
2012	<a href="#">937934</a>	<a href="#">43.6</a> <a href="#">63.0</a> %	8.0%	4.3%	✓		✓		
2013	<a href="#">888882</a>	<a href="#">41.0</a> <a href="#">59.3</a> %	2.9%	26. <a href="#">6</a> <a href="#">4</a> %	✓	✓	✓		
2014	<a href="#">710709</a>	<a href="#">46.2</a> <a href="#">66.7</a> %	10.6%	8.0%		✓	✓		
2015	<a href="#">1121868</a>	<a href="#">46.2</a> <a href="#">66.7</a> %	<a href="#">14.4</a> <a href="#">12.6</a> %	<a href="#">4.9</a> <a href="#">6.3</a> %		✓	✓		
2016	<a href="#">14621461</a>	<a href="#">48.7</a> <a href="#">70.4</a> %	13.6%	4.6%		✓	✓		
2017	<a href="#">16431638</a>	<a href="#">48.7</a> <a href="#">70.4</a> %	<a href="#">14.4</a> <a href="#">45</a> %	<a href="#">4.6</a> <a href="#">5</a> %		✓	✓		
2018	<a href="#">28802646</a>	<a href="#">46.2</a> <a href="#">66.7</a> %	<a href="#">17.8</a> <a href="#">6</a> %	0.0%			✓		
2019	<a href="#">35963586</a>	<a href="#">43.6</a> <a href="#">63.0</a> %	15.9%	0.0%			✓		
2020	<a href="#">13971363</a>	<a href="#">46.2</a> <a href="#">66.7</a> %	<a href="#">10.2</a> <a href="#">4</a> %	0.0%			✓		✓
2021	<a href="#">11631115</a>	<a href="#">53</a> <a href="#">77</a> .8%	<a href="#">4.9</a> <a href="#">2.0</a> %	0.0%			✓		✓
2022	<a href="#">13241235</a>	<a href="#">53</a> <a href="#">77</a> .8%	<a href="#">3.4</a> <a href="#">6</a> %	0.0%			✓		✓
2023	<a href="#">16351593</a>	<a href="#">53</a> <a href="#">77</a> .8%	<a href="#">2.7</a> <a href="#">8</a> %	0.0%			✓		✓
2024	<a href="#">30972972</a>	<a href="#">53</a> <a href="#">77</a> .8%	<a href="#">4.4</a> <a href="#">6</a> %	0.1%			✓		✓
2025	572	<a href="#">59.3</a> %	<a href="#">14.7</a> %	<a href="#">0.0</a> %			✓		✓
mean	<a href="#">984943</a>	<a href="#">42.9</a> <a href="#">62.0</a> %	13.5%	<a href="#">10.5</a> <a href="#">1</a> %	—	—	—	—	—

<sup>1</sup>Total number of mukims in Brunei = 39. <sup>2</sup>Of Brunei's 39 mukims, only 27 are considered transactable—excluding water villages and remote, non-developable areas.

<sup>2</sup>Unknown property type.

<sup>3</sup>Missing all of plot area, floor area, beds, and baths variables.



206 **Manual Data Collection**

207 Early years data collection was conducted manually, involving the transcription of property listing  
208 details from advertisements into a digital tabular format. This process was carried out by two of the  
209 authors over a period of nine months, from October 2023 to July 2024—[working at a manageable  
210 pace. A total of 12,092 data points were collected in this manner, which translates to processing  
211 approximately 150 entries per week per person. Spreading the task over such an extended period  
212 ensured that transcribers were never under undue time pressure, which could have led to errors due  
213 to fatigue.](#)

214 The primary sources of the property listings were local newspapers and magazines. Physical copies  
215 were accessed through the National Archive of Brunei Darussalam, while digital versions, which are  
216 digitised replicas of the physical newspapers, were obtained online. These digital formats could not  
217 be scraped due to their lack of structured data, necessitating manual transcription.

218 Although daily newspapers from 1993 onward were available at the National Archive, the classified  
219 sections were not always present. From 1993 to 1999, property advertisements were found only in  
220 Friday editions, and occasionally on Saturdays. Thus, newspapers from both these days were  
221 reviewed weekly to capture the listings data. This yielded roughly between 300 and 700 listings per  
222 year.

223 From the year 2000 onwards, property advertisements were published daily in the classifieds  
224 section. However, reviewing every single daily edition was not practical and would increase the  
225 likelihood of recording duplicate listings, thus necessitating a sampling strategy. The sampling was  
226 done as follows. Three newspaper editions per week were selected, and the classifieds section was  
227 reviewed for property listings. When a listing was found, it was recorded after careful filtering to  
228 ensure it was unique. This manual filtering process involved cross-checking based on the real estate  
229 agent, house characteristics, price, location, and date proximity. To avoid duplication, the same  
230 house listing was not recorded more than once within a quarter. This process yielded roughly the  
231 same number of listings per year as the earlier years.

232 **Web Scraping**

233 To compile additional property data for the study beyond manual data collection, web scraping was  
234 employed using the R programming language, making use of the {rvest} package [148]. This method  
235 enabled the systematic extraction of structured information from various local property listing  
236 websites such as panvilla.com (now defunct), bruhome.com, and bruneiproPERTY.com.bn. Such  
237 websites provide extensive details on properties listed as “for sale” in Brunei, aggregating  
238 advertisements from real estate agents and property developers.

239 The process began by identifying the structure of the target websites, focusing on the HTML tags and  
240 classes containing the relevant information. The goal here is to programmatically pinpoint and collect  
241 specific information like text, links, or attributes. For example, elements such as property prices,  
242 number of bedrooms, bathrooms, location, and other features were enclosed within specific HTML  
243 elements, which {rvest} functions like `html_elements()` and `html_text2()` could target and  
244 extract efficiently. [Figure 4 \(a\)](#) illustrates the structure of a typical property listing showing the  
245 various HTML elements to target. Example code to perform this task is available in the repository.

246

Showing 1-128 of 2909 total results

1 2 3 &gt;

.property-address                    .property-price



**PROPERTY CODE & LOCATION**

Tungku, Brunei-Muara District

19 July 2024

**BND 350,000**

.property-agent-date

 4

.property-bed

 4

.property-bath

**See Details**

.property-link

(a)

**Description**

.p

**MODERN DOUBLE STOREY DETACHED HOUSE FOR SALE – TUNGKU**

 **Prime Location:** Nestled in the heart of Kampong Tungku, just a 5-minute drive to The Mall Gadong, this property offers a blend of modern design and urban convenience.

 **Property Details:**

- Land Title: In-Perpetuity
- Selling Price: \$350,000
- Land Size: 0.15 acre
- House Size: 3,000 sq. ft
- Bedrooms: 4
- Bathrooms: 4

 **House Features:**

- **Ground Floor:** Spacious Living & Dining Area, Wet & Dry Kitchen, 1 Guest Bedroom (en-suite), Guest Toilet, Laundry Room.
- **First Floor:** Master Bedroom (en-suite with walk-in closet + balcony), 2 Bedrooms (en-suite).

 **Parking:** Car porch fits 2 cars.

 **Nearby Amenities:** Schools, hospitals, cafes, and parks all within a 10-minute radius.

 **Contact Us Today!**

For viewing and further details, please contact:

 +673 xxxxxxxx |  WhatsApp: +673 xxxxxxxx

 email@example.com

 **Services We Provide:**

-  Bank Loan Application Assistance
-  Legal Documentation Support

**AGENT** (b)

247

Figure 4: Illustration of a property listing from a typical Bruneian property portal. Attribution: Freepik.



248 Each webpage displayed a fixed number of listings (e.g. 128 per page). To scrape all pages, a loop was  
249 created to iterate through each page by modifying the URL (such as with an &offset=<number>  
250 parameter, where <number> represents the cumulative number of listings already scraped).

251

Showing 1-128 of 2909 total results

1 2 3 >

The screenshot shows a property listing for a modern two-story house. The listing includes the following details:

- PROPERTY ADDRESS:** Tungku, Brunei-Muara District
- PROPERTY PRICE:** BND 350,000
- PROPERTY AGENT DATE:** 19 July 2024
- PROPERTY BEDS:** 4
- PROPERTY BATHS:** 4
- See Details**

Annotations with green arrows and labels point to specific elements:

- A green arrow points from the label ".property-address" to the address text.
- A green arrow points from the label ".property-price" to the price text.
- A green arrow points from the label ".property-agent-date" to the agent date text.
- A green arrow points from the label ".property-link" to the "See Details" button.
- A green arrow points from the label ".property-bed" to the beds count.
- A green arrow points from the label ".property-bath" to the baths count.

{e}

**Description**

• p

**MODERN DOUBLE STOREY DETACHED HOUSE FOR SALE – TUNGKU**

**Prime Location:** Nestled in the heart of Kampong Tungku, just a 5-minute drive to The Mall Gadong, this property offers a blend of modern design and urban convenience.

**Property Details:**

- Land Title: In-Perpetuity
- Selling Price: \$350,000
- Land Size: 0.15 acre
- House Size: 3,000 sq. ft
- Bedrooms: 4
- Bathrooms: 4

**House Features:**

- **Ground Floor:** Spacious Living & Dining Area, Wet & Dry Kitchen, 1 Guest Bedroom (en-suite), Guest Toilet, Laundry Room.
- **First Floor:** Master Bedroom (en-suite with walk-in closet + balcony), 2 Bedrooms (en-suite).

**Parking:** Car porch fits 2 cars.

**Nearby Amenities:** Schools, hospitals, cafes, and parks all within a 10-minute radius.

**Contact Us Today!**  
For viewing and further details, please contact:  
 +673 xxxxxxxx |  WhatsApp: +673 xxxxxxxx  
 [email@example.com](mailto:email@example.com)

**Services We Provide:**

-  Bank Loan Application Assistance
-  Legal Documentation Support

**AGENT**  – Your Trusted Real Estate Partner

(b)

252 **Figure 4:** Illustration of a property listing from a typical Bruneian property portal. Attribution: Freepik.

253 Extracted data required some cleaning to standardise formats for analysis. Specifically:

- 254 • price variables were cleaned by removing non-numeric characters and converted to  
255 integers.
- 256 • beds and baths were converted to integers.
- 257 • date variables were formatted properly as Date objects.
- 258 • Locations were stored as text strings. See the subsection below on spatial data  
259 harmonisation [for more details](#).
- 260 • Any additional information was extracted from the property descriptions and saved as a  
261 character vector. This very often contained valuable insights not captured in the primary  
262 fields.

263 Data from 2012 up until the present ([December 2024](#)[January 2025](#)) were managed to be collected  
264 using this method, averaging around 1,500 listings per year. While highly efficient, this process relied  
265 heavily on the consistency of the site structure of the source webpages. Changes to website layouts  
266 or closures over time required significant updates to the scraping scripts. To overcome these issues,  
267 alternative approaches, such as using Large Language Models (LLMs) were considered. This is  
268 explained in the next subsection.



269 LLM Data Cleaning Extraction  
270 As previously mentioned, the web scraping process also captured unstructured information from the  
271 property descriptions, which often contained valuable details not captured in the primary fields. In  
272 this subsection, we detail the data cleaning extraction process using a pre-trained large language  
273 model Large Language Model (LLM) to extract structured information from the unstructured text.  
274 The LLM used was Llama 3.1 with 8B parameters [15] the DeepSeek R1 distilled Qwen 14B [9,10],  
275 accessed using the {tidychatmodelsellmer} R package [4611] via the Ollama<sup>1</sup> API, a local  
276 interface platform to the LLM.  
277 The primary goal was to extract the house characteristics of interest, specifically variables 7 to 15 as  
278 per Table 1, from the unstructured verbose descriptions scraped from property listing websites. Each  
279 description was processed with a carefully designed prompt (Figure 5) to ensure consistent output.  
280 This prompt instructed the model to return only the required information in a semicolon-separated  
281 format, while handling edge cases such as missing descriptions, or non-residential (commercial)  
282 properties, or rental property advertisements.  
283

---

1 <https://github.com/ollama/ollama>

"The following is the description from a property sale listing in Brunei. This description will contain the information about the property, including its characteristics, price, and location. However, some of these descriptions may not contain property listings, and instead contain other or no information at all.

In the case where this description is in fact a property listing, I would like you to extract the following information:

1. Location / area of the property in Brunei, CHARACTER.
2. Price of the property in Brunei Dollars, NUMERIC.
3. Type of property, CHARACTER -- select from Detached, Semi-Detached, Terrace, Apartment, or Land.
4. Land tenure, CHARACTER -- select from Freehold, Leasehold, or Strata. If other than this, return 'NA'.
5. Status of the property, CHARACTER -- select from Proposed, Under Construction, New, or Resale.
6. Land area in acres, NUMERIC.
7. Built up area in square feet, NUMERIC.
8. Number of storeys, INTEGER.
9. Number of bedrooms, INTEGER.
10. Number of bathrooms, INTEGER.

Further instructions:

- Please return \*\*semicolon\*\* separated values like this:

Kg Tanah Jambu; 250000; Detached ; Freehold ; New ; 0.3 ; 2500; 2; 3; 3  
Kg Tungku ; 300000; Terrace ; Leasehold; Resale ; 0.25; 1700; 2; 3; 2  
Kg Kiarong ; 200000; Apartment; Strata ; Proposed; 0.1 ; 1000; NA; 2; 2  
etc.

NUMBERS SHOULD NOT CONTAIN comma (,) for thousands separator

- If any of the 10 values are missing, please return 'NA' for that value.
- If the description does not contain a property listing (for example, it is a rental property advertisement), return 'NA' for all 10 values.
- DO NOT RESPOND WITH ANYTHING ELSE OTHER THAN THE REQUIRED INFORMATION."

Figure 5: The LLM prompt to clean descriptions obtained from web scraping.

284 The verbose descriptions were fed into the model one at a time using a loop, with the LLM extracting  
285 and returning the relevant details. Note that this loop was not parallelised, [as the LLM already](#)  
286 [utilises multiple cores, and further parallelisation would not yield significant efficiency gains](#) due to  
287 [the computational resources required by the LLM resource constraints](#). The cleaned results were then  
288 parsed and stored in a data frame, which was then subjected to manual data-type validation to  
289 ensure conformity with the existing data set. [\(see the last subsection\)](#). It takes, on average, [2.1281.7](#)  
290 seconds to process a single description ~~running on Apple MacBook Air M2 Silicon on Chip (SoC) and~~  
291 ~~System in Chip (SiP) processors using a Mac Pro 3.2GHz 16 core Intel Xeon W with 16GB48GB of DDR4~~



292 RAM. We processed 4,820~~5,055~~ descriptions ~~in total~~ from 2020 to 2024~~2025~~ using this method, with  
293 a ~~runtime~~run time of approximately 20 hours in total (spread over three ~~hours.~~machines).

294 ~~To test the accuracy of the LLM, a random sample of 329 descriptions was selected and manually~~  
295 ~~verified for correctness. Of these, 306 were deemed correct (correct entries or identified “non-~~  
296 ~~listings” correctly), resulting in an accuracy rate of 93.0%. Rare errors were spotted due to model~~  
297 ~~hallucinations (despite setting the LLM temperature to the lowest setting), but these are typically~~  
298 ~~minor and unlikely to significantly impact the overall analysis. Large errors on the other hand were~~  
299 ~~corrected manually, by filtering for outliers (values exceeding three standard deviations from the~~  
300 ~~mean in magnitude) or inconsistencies in the variables. To evaluate the accuracy of the LLM data~~  
301 ~~extraction, a test data set of 100 artificially generated descriptions was created, with each~~  
302 ~~description written in the style of a typical property listing. For each house characteristic, the~~  
303 ~~accuracy of the LLM was calculated by comparing the extracted value to the ground truth. For~~  
304 ~~numeric variables, values were considered accurate if they fell within 1% of the corresponding~~  
305 ~~ground truth value. For character variables, accuracy was determined using the normalised~~  
306 ~~Levenshtein distance, ensuring differences remained within a set threshold. The test data also~~  
307 ~~contained missing information in certain property characteristics, which the LLM was expected to~~  
308 ~~handle correctly. A correct handling was counted if both the extracted and ground truth values were~~  
309 ~~missing. Accuracy was then averaged per characteristic as well as across all characteristics.~~

310 Our experiments show that the LLM data extraction process with the deepseek-r1:14b model  
311 achieved an overall accuracy rate of 96.9% (see Appendix for complete results). Errors mostly  
312 stemmed from incorrect status classification, likely due to the vagueness of the advertisement  
313 listing. As for the rest of the variables, obvious errors were flagged and corrected manually during  
314 our data validation process (described in the subsection following the next one). Overall, the LLM  
315 was found to be a valuable tool for extracting structured information from unstructured text,  
316 significantly reducing the time and effort required for data cleaning. Users may wish to exclude these  
317 records from their analysis if they are concerned about the accuracy of the extracted data.extraction.

### 318 Spatial Data Harmonisation

319 Whether the data was collected manually, through web scraping, or cleaned using the LLM, the  
320 spatial information extracted was often inconsistent in terms of naming conventions and granularity.  
321 To address this, a spatial data harmonisation process was conducted to standardise the names of the  
322 kampongs (villages) in the data set to the format used by Department of Economic Planning and  
323 Statistics (DEPS), Ministry of Finance and Economy, Brunei Darussalam as per the most recent census  
324 [[17](#)[12](#)]. This is the same format used by the R package {bruneimap} [[12](#)[6](#)]. The CSV file  
325 bn\_kpg\_level\_data.csv obtained from this package was used as a reference to standardise the  
326 kampong names in the data set, which conveniently also ~~included~~includes the mukim and district  
327 names for each kampung.

328 The majority of house listings in Brunei specify the property location using the kampong name, the  
329 smallest administrative unit in the country. The task in hand was then to match these kampong  
330 names in the data set with the standardised names in the reference file. Several challenges were  
331 encountered during this process, including:

332 1. Spelling variations or misspellings, though these were relatively straightforward to correct.



- 333        2. Unknown entries, where the correct kampong could sometimes be inferred from the  
334        geographical context; otherwise, these were set to NA.  
  
335        3. Multiple matches, occurring when two or more kampongs shared the same name (e.g.,  
336        Kampong Panchor in Mukim Mentiri and Kampong Panchor in Mukim Lumapas). Additional  
337        information from the listing was used to determine the correct match, but where this was  
338        not possible, these entries were also set to NA.

339 This process was carried out manually using data filtering features in Microsoft Excel. Once  
340 completed, all entries marked as NA were removed so as to provide complete spatial information for  
341 each listing.

### 342 Data Validation

343 To ensure data quality, a series of consistency and validity checks were performed on the data set,  
344 especially after manual transcription and LLM data cleaning. These checks include

- 345        1. **Outlier detection.** Summary statistics analyses to identify and flag anomalous values. For  
346        instance, a built-up area recorded as 0.1 square feet or an implausibly high number of beds  
347        and baths (e.g., >20) would highly likely indicate a possible error, prompting manual review.  
348        Such anomalies were flagged and manually reviewed for correction.
- 349        2. **Internal consistency checks.** This made use of substantive knowledge about Brunei's housing  
350        market [13]. An example is using the price per square foot indicator, whose value typically  
351        falls within a known range. Therefore any deviations were scrutinised to ensure that  
352        variables such as price and floor area were correctly recorded. This was similarly applied to  
353        other variables such as plot area, beds, and baths.
- 354        3. **Duplicate records detection.** We also performed duplicate record detection to identify any  
355        repeated entries that might have arisen from overlapping data sources or transcription  
356        errors. Any duplicates that were identified were carefully reviewed and removed to ensure  
357        that each property listing was uniquely represented in the span of one calendar month.

358 These data validation procedures collectively contribute to a robust and reliable data set, providing a  
359 solid foundation for subsequent analysis.

## 360 LIMITATIONS

361 Listing prices in our data set serve as a proxy for market values, capturing advertised trends rather  
362 than final sale outcomes. This enables timely analysis of market sentiment, even though factors such  
363 as negotiation dynamics, seller strategies, and market conditions may cause deviations from actual  
364 transaction prices.

365 Despite significant efforts to ensure data quality, some limitations remain. DuplicateFirst, integrating  
366 historical data from manually transcribed sources with later web-scraped data may introduce  
367 inconsistencies affecting comparability over time. Second, although duplicate listings were carefully  
368 reviewed and removed, though there remains a slight possibility that duplicated records still exist of  
369 residual duplicates in the data set. The spatial data coverage is also heavily skewed towards the  
370 Brunei-Muara district, which accounts for 91.7% of the listings. This is a reflection of the district's  
371 higher population and greater volume of property transactions [18], which may bias analyses toward



372 this region. Furthermore [Thirdly](#), while we have confidence in the data quality from 2015 to 2024,  
373 property price trends between 1993 and 2014 cannot be fully verified. Nonetheless, this study serves  
374 as a valuable starting point. Future research could benefit greatly from access to administrative  
375 transaction data, which would allow for more comprehensive and accurate analyses.

376 ~~Missing data in house characteristics is another limitation, although this should be seen as an  
377 opportunity for further research to develop imputation methods or alternative analytical  
378 approaches. Minor inaccuracies were also observed in the LLM-based data cleaning process, but  
379 these issues are infrequent and can be mitigated by subsetting or refining the affected entries.~~

380 Finally, while significant effort was made to harmonise spatial data, the matching of kampong names  
381 to standardised references may not be entirely error-free. However, aggregation to the mukim level  
382 provides a reliable alternative for spatial analyses, ensuring that the data set remains valuable for  
383 research and analysis.

## 384 ETHICS STATEMENT

385 The authors confirm that the current work does not involve human subjects, animal experiments, or  
386 data collected from social media platforms. The data described in this article were obtained from  
387 publicly available, non-personal, and factual sources, including physical and digital newspapers and  
388 magazines.

389 Web scraping from local property listings websites was conducted in compliance with ethical and  
390 legal considerations. Specifically, data were not collected from behind login barriers, and the terms of  
391 service (ToS) for the websites did not explicitly prohibit web scraping. Furthermore, the robots.txt  
392 files for the websites were reviewed, and any policies outlined there were adhered to.

393 The data collected consisted exclusively of non-copyrightable factual information, such as property  
394 characteristics and spatial locations, and excluded any potentially copyrighted content such as  
395 images. To ensure privacy, no personally identifiable information, including specific property  
396 addresses, was scraped nor included in the data set. To this end, data from the description fields  
397 processed by the LLM are not included in the data set, as they may contain sensitive information  
398 such as contact details and names of companies. Furthermore, we have anonymised the names of  
399 the real estate agents and companies in the data set.

## 400 CRediT AUTHOR STATEMENT

- 401 • **Haziq Jamil:** Conceptualisation, Methodology, Software, Formal analysis, Data curation,  
402 Writing-Original Draft, Visualisation, Supervision, Project administration, Funding acquisition.
- 403 • **Amira Barizah Noorosmawie:** Software, Data curation, Writing-Original Draft.
- 404 • **Hafeezul Waezz Rabu:** Software, Data curation, Writing-Original Draft.
- 405 • **Lutfi Abdul Razak:** Conceptualisation, Validation, Supervision, Funding acquisition.

## 406 ACKNOWLEDGEMENTS

407 The authors gratefully acknowledge the contributions of Atikah Farhain Yahya, Nurulhanisah Abdul  
408 Manan, and Nina Zuhairi towards the collection and processing of the data contained within. [We also](#)



409 [thank the Brunei Darussalam Central Bank \(BDCB\) for the engaging discussions and support, which](#)  
410 [were instrumental in the initiation of this project.](#)

## 411 DECLARATION OF COMPETING INTERESTS

412 The authors declare that they have no known competing financial interests or personal relationships  
413 that could have appeared to influence the work reported in this paper.

## 414 REFERENCES

- 415 [1] V. Shabunko, C.M. Lim, S. Brahim, S. Mathew, [Developing building benchmarking for Brunei](#)  
416 [Darussalam, Energy and Buildings 85 \(2014\) 79–85. <https://doi.org/10.1016/j.enbuild.2014.08.047>](#)
- 417 [2][1] M.K.M. Ng, Z. Shabrina, B. Buyuklieva, [Characterising Land Cover Change in Brunei](#)  
418 [Darussalam's Capital District, Applied Spatial Analysis and Policy 15 \(2022\) 919–946.](#)  
419 <https://doi.org/10.1007/s12061-021-09429-9>.
- 420 [32] H. Jamil, F. Usop, H.M. Ramli, [Leveraging Sparse Gaussian Processes for Property Price](#)  
421 [Modelling and Sustainable Urban Planning, in: S.A. Abdul Karim, A. Baharum \(Eds.\), Intelligent](#)  
422 [Systems of Computing and Informatics in Sustainable Urban Development, Taylor and Francis/CRC](#)  
423 [Press, 2025.](#)
- 424 [43] N.H. Hassan, I. Azrein, K. Ibrahim, G. Yong, [Cultural Consideration in Vertical Living in Brunei](#)  
425 [Darussalam Cultural Consideration in Vertical Living in Brunei Darussalam, in: Managing Urban](#)  
426 [Growth: Challenges for Small Cities, 2011.](#)
- 427 [54] N.H. Hassan, [The Sociocultural Significance of Homeownership in Brunei Darussalam, in: L.](#)  
428 [Kwen Fee, P.J. Carnegie, N.H. Hassan \(Eds.\), \(Re\)presenting Brunei Darussalam: A Sociology of the](#)  
429 [Everyday, Springer Nature, Singapore, 2023: pp. 185–206. \[https://doi.org/10.1007/978-981-19-6059-8\\\_11\]\(https://doi.org/10.1007/978-981-19-6059-8\_11\)](#)
- 431 [6] F. Braesemann, A. Baum, [PropTech: Turning Real Estate Into a Data-Driven Market?, SSRN](#)  
432 [Electronic Journal \(2020\). <https://doi.org/10.2139/ssrn.3607238>](#)
- 433 [7] J.R. DeLisle, B. Never, T.V. Grissom, [The big data regime shift in real estate, Journal of](#)  
434 [Property Investment & Finance 38 \(2020\) 363–395. <https://doi.org/10.1108/JPIF-10-2019-0134>](#)
- 435 [8][5] BDCB, [Technical Notes for Residential Property Price Index \(RPPI\), Brunei Darussalam Central](#)  
436 [Bank, 2021.](#)
- 437 [9] A.N. Rachman, [Residential property price index for Indonesia using big data: The case of](#)  
438 [Jakarta, in: International Conference on Real Estate Statistics, 2019.](#)
- 439 [10] HM Land Registry, [Comparing house price indices in the UK, HM Land Registry, 2023.](#)
- 440 [11] European Commission, Eurostat, [Handbook on residential property prices indices \(RPPIs\),](#)  
441 [Publications Office, EU, 2013.](#)
- 442 [12][6] H. Jamil, [Bruneimap: Maps and Spatial Data of Brunei \(R package version 0.3.1.9001\), 2024.](#)
- 443 [13][7] R Core Team, [R: A language and environment for statistical computing, R Foundation for](#)  
444 [Statistical Computing, Vienna, Austria, 2024.](#)



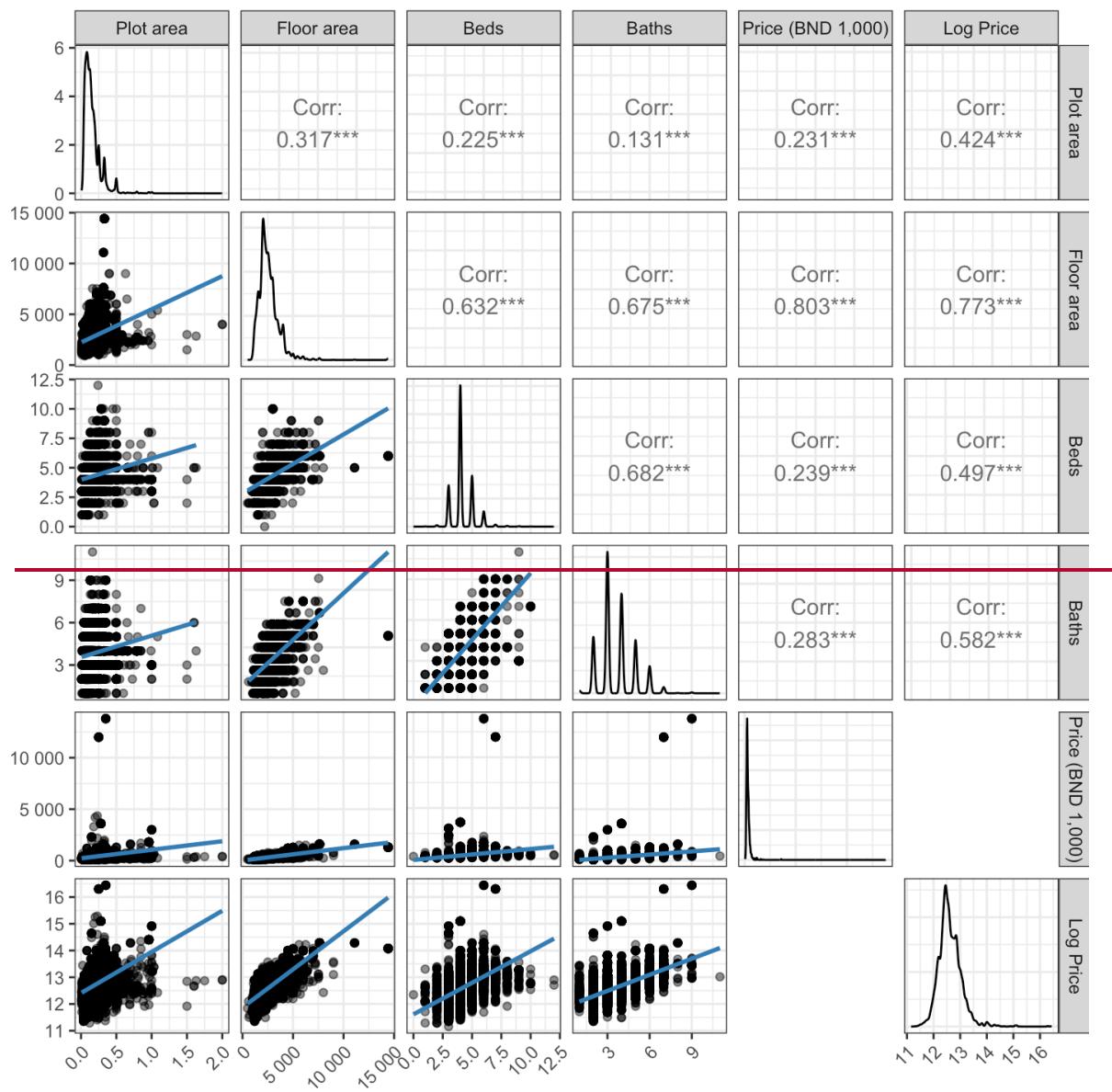
- 445 [148] H. Wickham, Rvest: Easily harvest (scrape) web pages, 2024.
- 446 [15] A. Grattafiori, A. Dubey, A. Jauhri, A. Pandey, A. Kadian, A. Al Dahle, A. Letman, A. Mathur, A.  
447 Schelten, A. Vaughan, A. Yang, A. Fan, A. Goyal, A. Hartshorn, A. Yang, A. Mitra, A. Sravankumar, A.  
448 Korenev, A. Hinsvark, A. Rao, A. Zhang, A. Rodriguez, A. Gregerson, A. Spataru, B. Roziere, B. Biron, B.  
449 Tang, B. Chern, C. Caucheteux, C. Nayak, C. Bi, C. Marra, C. McConnell, C. Keller, C. Touret, C. Wu, C.  
450 Wong, C.C. Ferrer, C. Nikolaidis, D. Allonsius, D. Song, D. Pintz, D. Livshits, D. Wyatt, D. Esiobu, D.  
451 Choudhary, D. Mahajan, D. Garcia Olano, D. Perino, D. Hupkes, E. Lakomkin, E. AlBadawy, E.  
452 Leabanova, E. Dinan, E.M. Smith, F. Radenovic, F. Guzmán, F. Zhang, G. Synnaeve, G. Lee, G.L.  
453 Anderson, G. Thattai, G. Nail, G. Mialon, G. Pang, G. Cucurell, H. Nguyen, H. Korevaar, H. Xu, H.  
454 Tøttrup, I. Zarov, I.A. Ibarra, I. Klöumann, I. Misra, I. Evtimov, J. Zhang, J. Copet, J. Lee, J. Geffert, J.  
455 Vranes, J. Park, J. Mahadeokar, J. Shah, J. van der Linde, J. Billock, J. Hong, J. Lee, J. Fu, J. Chi, J.  
456 Huang, J. Liu, J. Wang, J. Yu, J. Bitton, J. Spisak, J. Park, J. Rocca, J. Johnstun, J. Saxe, J. Jia, K.V. Alwala,  
457 K. Prasad, K. Upasani, K. Plawiak, K. Li, K. Heafield, K. Stone, K. El Arini, K. Iyer, K. Malik, K. Chiu, K.  
458 Bhalla, K. Lakhota, L. Rantala Yeary, L. van der Maaten, L. Chen, L. Tan, L. Jenkins, L. Martin, L.  
459 Madaan, L. Malo, L. Blecher, L. Landzaat, L. de Oliveira, M. Muzzi, M. Pasupuleti, M. Singh, M. Paluri,  
460 M. Kardas, M. Tsimpoukelli, M. Oldham, M. Rita, M. Pavlova, M. Kambadur, M. Lewis, M. Si, M.K.  
461 Singh, M. Hassan, N. Goyal, N. Terabi, N. Bashlykov, N. Bögoychev, N. Chatterji, N. Zhang, O.  
462 Duchenne, O. Çelebi, P. Alrassy, P. Zhang, P. Li, P. Vasic, P. Weng, P. Bhargava, P. Dubal, P. Krishnan, P.S.  
463 Koura, P. Xu, Q. He, Q. Dong, R. Srinivasan, R. Ganapathy, R. Calderer, R.S. Cabral, R. Stojnic, R.  
464 Raileanu, R. Maheswari, R. Girdhar, R. Patel, R. Sauvestre, R. Polidoro, R. Sumbaly, R. Taylor, R. Silva,  
465 R. Hou, R. Wang, S. Hosseini, S. Chennabasappa, S. Singh, S. Bell, S.S. Kim, S. Edunov, S. Nie, S.  
466 Narang, S. Raparthys, S. Shen, S. Wan, S. Bhosale, S. Zhang, S. Vandenhende, S. Batra, S. Whitman, S.  
467 Sootla, S. Collot, S. Gururangan, S. Borodinsky, T. Herman, T. Fowler, T. Sheasha, T. Georgiou, T.  
468 Seialom, T. Speckbacher, T. Mihaylov, T. Xiao, U. Karn, V. Goswami, V. Gupta, V. Ramanathan, V.  
469 Kerkez, V. Gonguet, V. De, V. Vogeti, V. Albiero, V. Petrovic, W. Chu, W. Xiong, W. Fu, W. Meers, X.  
470 Martinet, X. Wang, X. Wang, X.E. Tan, X. Xia, X. Xie, X. Jia, X. Wang, Y. Goldschlag, Y. Gaur, Y. Babaei, Y.  
471 Wen, Y. Song, Y. Zhang, Y. Li, Y. Mao, Z.D. Coudert, Z. Yan, Z. Chen, Z. Papakipos, A. Singh, A.  
472 Srivastava, A. Jain, A. Kelsey, A. Shajnfeld, A. Gangidi, A. Victoria, A. Goldstand, A. Menon, A. Sharma,  
473 A. Boesenberg, A. Baevski, A. Feinstein, A. Kallet, A. Sangani, A. Teo, A. Yunus, A. Lupu, A. Alvarado, A.  
474 Caples, A. Gu, A. Ho, A. Poulton, A. Ryan, A. Ramchandani, A. Dong, A. Franco, A. Goyal, A. Saraf, A.  
475 Chowdhury, A. Gabriel, A. Bharambe, A. Eisenman, A. Yazdan, B. James, B. Maurer, B. Leonhardi, B.  
476 Huang, B. Loyd, B.D. Paola, B. Paranjape, B. Liu, B. Wu, B. Ni, B. Hancock, B. Wasti, B. Spence, B.  
477 Stojkovic, B. Gamido, B. Montalvo, C. Parker, C. Burton, C. Mejia, C. Liu, C. Wang, C. Kim, C. Zhou, C.  
478 Hu, C. H. Chu, C. Cai, C. Tindal, C. Feichtenhofer, C. Gao, D. Civin, D. Beaty, D. Kreymer, D. Li, D.  
479 Adkins, D. Xu, D. Testuggine, D. David, D. Parikh, D. Liskovich, D. Foss, D. Wang, D. Le, D. Holland, E.  
480 Dowling, E. Jamil, E. Montgomery, E. Presani, E. Hahn, E. Wood, E. T. Le, E. Brinkman, E. Arcaute, E.  
481 Dunbar, E. Smothers, F. Sun, F. Kreuk, F. Tian, F. Kokkinos, F. Ozgenel, F. Caggioni, F. Kanayet, F. Seide,  
482 G.M. Florez, G. Schwarz, G. Badeer, G. Swee, G. Halpern, G. Herman, G. Sizov, Guangyi, Zhang, G.  
483 Lakshminarayanan, H. Inan, H. Shojanazeri, H. Zou, H. Wang, H. Zha, H. Habeeb, H. Rudolph, H. Suk,  
484 H. Aspegren, H. Goldman, H. Zhan, I. Damlaj, I. Molybog, I. Tufanov, I. Leontiadis, I. E. Veliche, I. Gat,  
485 J. Weissman, J. Geboski, J. Kohli, J. Lam, J. Asher, J. B. Gaya, J. Marcus, J. Tang, J. Chan, J. Zhen, J.  
486 Reizenstein, J. Teboul, J. Zhong, J. Jin, J. Yang, J. Cummings, J. Carvill, J. Shepard, J. McPhie, J. Torres, J.  
487 Ginsburg, J. Wang, K. Wu, K.H. U, K. Saxena, K. Khandelwal, K. Zand, K. Matosich, K. Veeraraghavan, K.  
488 Michelena, K. Li, K. Jagadeesh, K. Huang, K. Chawla, K. Huang, L. Chen, L. Garg, L. A, L. Silva, L. Bell, L.



- 489 Zhang, L. Guo, L. Yu, L. Moshkovich, L. Wehrstedt, M. Khabsa, M. Avalani, M. Bhatt, M. Mankus, M.  
490 Hasson, M. Lennie, M. Reso, M. Groshev, M. Naumov, M. Lathi, M. Keneally, M. Liu, M.L. Seltzer, M.  
491 Valko, M. Restrepo, M. Patel, M. Vyatskov, M. Samvelyan, M. Clark, M. Macey, M. Wang, M.J.  
492 Hermoso, M. Metanat, M. Rastegari, M. Bansal, N. Santhanam, N. Parks, N. White, N. Bawa, N.  
493 Singhal, N. Egebo, N. Usunier, N. Mehta, N.P. Laptev, N. Dong, N. Cheng, O. Chernoguz, O. Hart, O.  
494 Salpekar, O. Kalinli, P. Kent, P. Parekh, P. Saab, P. Balaji, P. Rittner, P. Bontrager, P. Roux, P. Dollar, P.  
495 Zvyagina, P. Ratanchandani, P. Yuvraj, Q. Liang, R. Alao, R. Rodriguez, R. Ayub, R. Murthy, R. Nayani, R.  
496 Mitra, R. Parthasarathy, R. Li, R. Hogan, R. Battey, R. Wang, R. Howes, R. Rinott, S. Mehta, S. Siby, S.J.  
497 Bondu, S. Datta, S. Chugh, S. Hunt, S. Dhillon, S. Sidorov, S. Pan, S. Mahajan, S. Verma, S. Yamamoto,  
498 S. Ramaswamy, S. Lindsay, S. Lindsay, S. Feng, S. Lin, S.C. Zha, S. Patil, S. Shankar, S. Zhang, S. Zhang, S.  
499 Wang, S. Agarwal, S. Sajuyigbe, S. Chintala, S. Max, S. Chen, S. Kehoe, S. Satterfield, S.  
500 Govindaprasad, S. Gupta, S. Deng, S. Cho, S. Virk, S. Subramanian, S. Choudhury, S. Goldman, T.  
501 Remez, T. Glaser, T. Best, T. Kochler, T. Robinson, T. Li, T. Zhang, T. Matthews, T. Chou, T. Shaked, V.  
502 Ventimitta, V. Ajayi, V. Montanez, V. Mohan, V.S. Kumar, V. Mangla, V. Ionescu, V. Poenaru, V.T.  
503 Mihailescu, V. Ivanov, W. Li, W. Wang, W. Jiang, W. Bouaziz, W. Constable, X. Tang, X. Wu, X. Wang, X.  
504 Wu, X. Gao, Y. Kleinman, Y. Chen, Y. Hu, Y. Jia, Y. Qi, Y. Li, Y. Zhang, Y. Zhang, Y. Adi, Y. Nam, Yu, Wang, Y.  
505 Zhao, Y. Hao, Y. Qian, Y. Li, Y. He, Z. Rait, Z. DeVito, Z. Rosnbrick, Z. Wen, Z. Yang, Z. Zhao, Z. Ma, The  
506 Llama 3 Herd of Models, (2024). <https://doi.org/10.48550/arXiv.2407.21783>.
- 507 [16] A. Rapp, [Tidychatmodels](#)[9] DeepSeek-AI, DeepSeek-R1: Incentivizing Reasoning  
508 Capability in LLMs via Reinforcement Learning, (2025). <https://doi.org/10.48550/arXiv.2501.12948>.
- 509 [10] Qwen, [Qwen2.5 Technical Report](#), (2025). <https://doi.org/10.48550/arXiv.2412.15115>.
- 510 [11] H. Wickham, J. Cheng, A. Jacobs, [Ellmer](#): Chat with all kinds of AI large language models  
511 through a common interface, 2024, 2025.
- 512 [1712] DEPS, The Population and Housing Census Report (BPP) 2021: Demographic, Household and  
513 Housing Characteristics, Department of Economic Planning and Statistics, Ministry of Finance and  
514 Economy, Brunei Darussalam, 2022.
- 515 [4813] H. Jamil, A spatio-temporal analysis of [property](#)[house](#) prices in Brunei Darussalam,  
516 (20242025). <https://doi.org/10.13140/RG.2.2.32533.74720>.

## 517 Appendix

## 518 Pairwise Correlation Plot of Numeric Variables



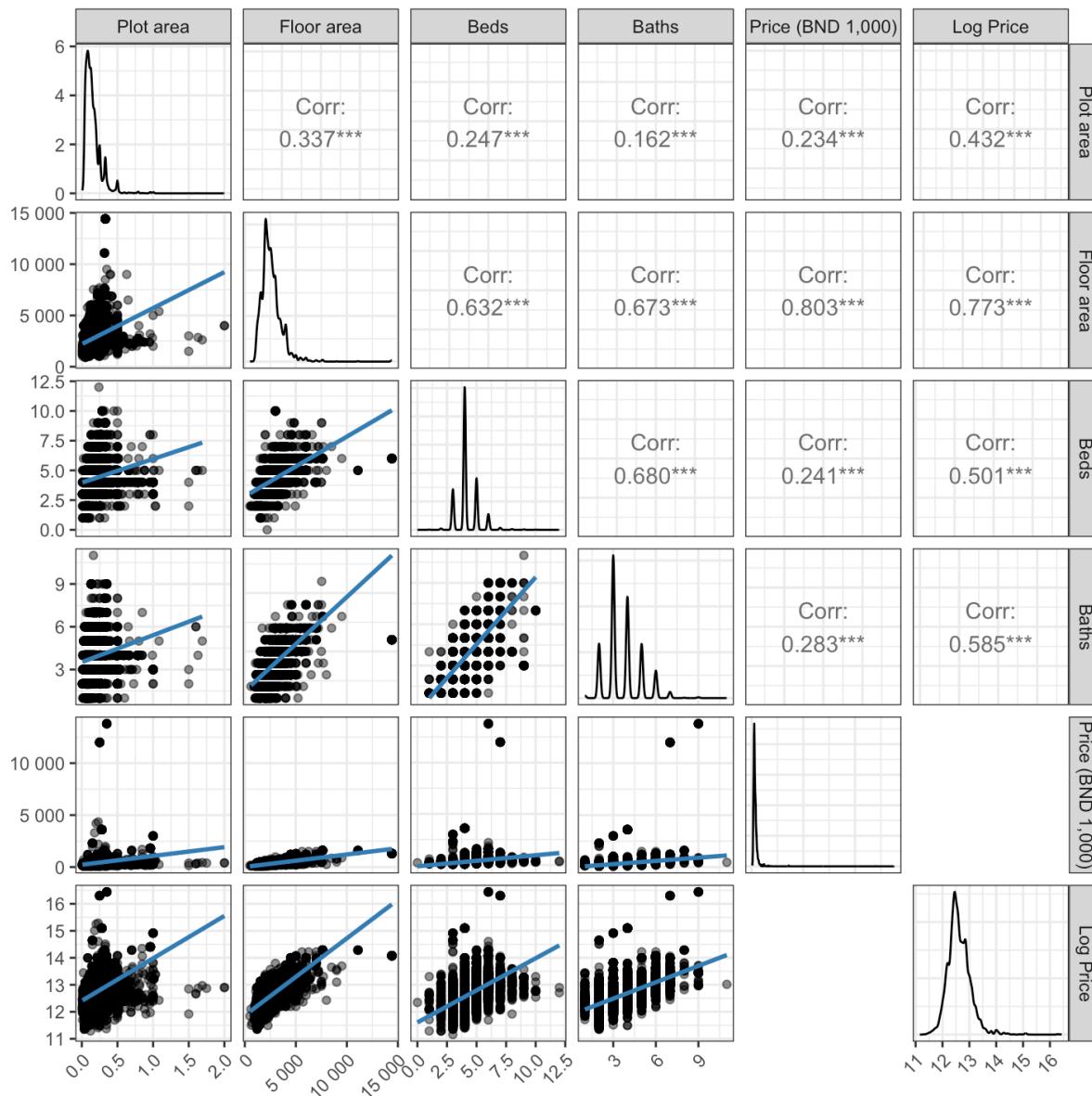
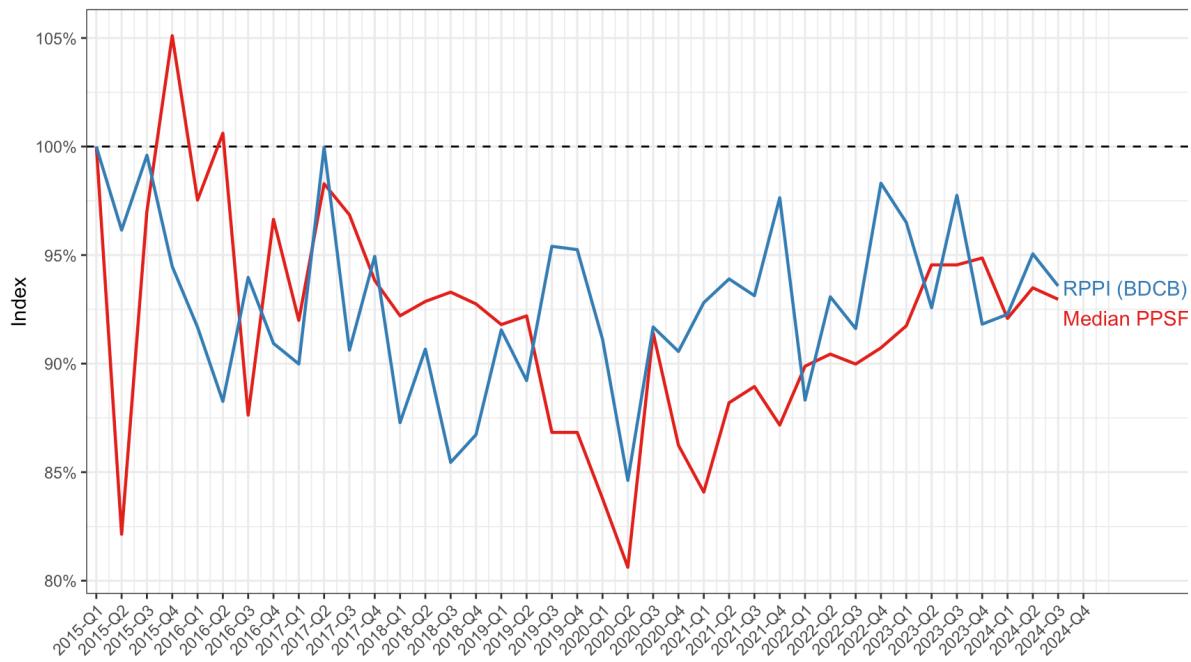


Figure 6: Pairwise correlation plot of continuous variables.

519 [Comparison to RPPI Data](#)  
520 To demonstrate the quality of the data set, we compared it with the Residential Property Price Index  
521 (RPPI) [5] published by the Brunei Darussalam Central Bank (BDCB). A simple median price per  
522 square foot (PPSF) index can be calculated by aggregating the data by quarters. This approach  
523 minimises the impact of missing values, as the index is based on aggregated data. Figure 7 shows the  
524 comparison between the RPPI and the PPSF index calculated from our data set. The mean absolute  
525 error (MAE) between the two indices is calculated to be 4.71%, indicating a good level of agreement  
526 between the two data sets.

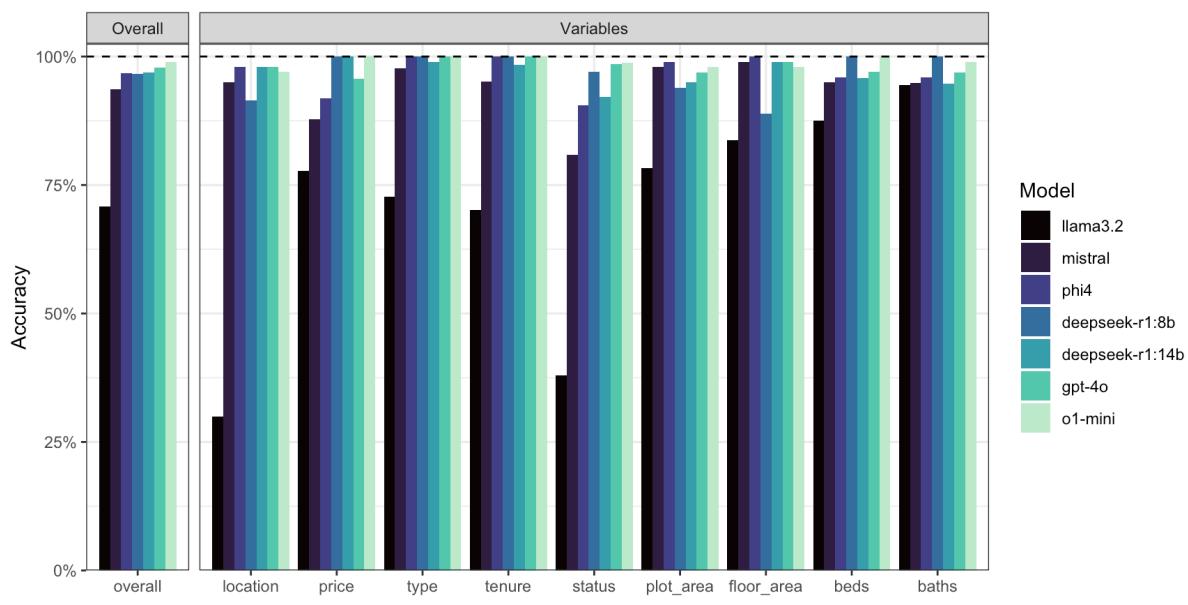


[Figure 7: Comparison of quarterly median price per square foot indices \(Median PPSF\) and the official Residential Property Price Index \(RPPI\) from Brunei Darussalam Central Bank \(BDCB\).](#)

## 527 LLM Accuracy Test

528 To test the accuracy of the LLM data extraction, we created a test data set of 100 house  
529 advertisements in the style of the web scraped data. Several popular models from Ollama were used,  
530 namely the Llama3.2 (3B), Mistral (7B), Phi 4 (14B), and DeepSeek-R1 distilled reasoning models  
531 based on Llama (8B) and Qwen (14B). For the locally run Ollama models, the settings were set to the  
532 default values, with the exception of a lowered temperature setting: `temperature = 0.1`, `top-`  
533 `p=0.9` (nucleus sampling), `top-k=40` (top-k sampling), `max-tokens=128`, and `repeat-`  
534 `penalty=1.1`. Additionally, two models from OpenAI were included for comparison. These were  
535 the GPT-4o and the o1-mini, with the latter being a reasoning model.

536 OpenAI's models topped the accuracy charts, with the o1-mini and gpt-4o models achieving the  
537 accuracy scores of 99.2% and 98.9% respectively. The reasoning model deepseek-r1:14b was the best  
538 performing locally run model, scoring 96.9% accuracy. Evidently, the smaller the model, the less  
539 accurate the extraction process (cf. Llama3.2 70.8%). Most inaccuracies occurred in the status  
540 variable, where models struggled to parse the correct build status from vague advertisement  
541 descriptions. However, key variables such as price, type, plot\_area, floor\_area, beds, and baths  
542 are generally reliable (>90% accuracy).



**Figure 8: Comparison of data extraction accuracy across multiple LLM models on the test dataset.**  
Each bar represents the percentage of correctly extracted fields for a given model.

543 The run time statistics for each model is shown in the table below. The computer used was a Mac Pro  
 544 3.2GHz 16 core Intel Xeon W with 48GB of DDR4 RAM. At the time of running the tests, graphics card  
 545 support was not available for the LLM models, which would have significantly reduced the run time.  
 546 Certainly cloud-based models such as OpenAI's models are much faster than running models locally,  
 547 though a paid API key is required.

**Table 4: Single run time statistics for LLM data extraction for each model in seconds.**

Model	Time (seconds)			
	Minimum	Mean	Median	Maximum
llama3.2	1.08	6.40	9.60	10.16
mistral	2.66	10.44	2.91	22.21
phi4	3.72	14.11	4.27	38.01
deepseek-r1:8b	36.83	65.97	54.61	112.78
deepseek-r1:14b	40.89	117.67	81.71	151.20
gpt-4o	1.28	2.46	1.73	5.38
o1-mini	5.97	7.95	7.51	9.94

548

549