Data in Brief
Open access
Article template

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*,[1],~~Jamil[a,*], Amira Barizah ~~Noorosmawie[1]~~Noorosmawie[a], Hafeezul Waezz ~~Rabu[1]~~Rabu[a], Lutfi Abdul ~~Razak[1]~~Razak[b]

Affiliations

~~[1]Mathematical~~[a]Mathematical Sciences, Faculty of Science, Universiti Brunei Darussalam, Jalan Tungku Link, Bandar Seri Begawan BE1410, Brunei Darussalam

[b]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

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 ~~cleaning~~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.

SPECIFICATIONS TABLE

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 (~~Llama 3.1~~DeepSeek R1 Distilled Qwen 14B via ~~tidychatmodels in R~~Ollama) were employed to ~~clean~~process unstructured text into structured formats. Inclusion criteria focused on property listings with price, location, and type, while duplicates were removed. Spatial data were ~~harmonized~~harmonised to match official administrative boundaries for consistency.
Data source location	The data were collected in Brunei Darussalam and are stored in ~~a~~an online public ~~GitHub~~ repository.
Data accessibility	Repository name: ~~GitHub/Bruneiverse~~Zenodo Data identification number: ~~N/A~~14978544 ~~Direct URL to data: https://bruneiverse.github.io/house-data/data/hspbn_2024-12-12.csv~~ Direct URL to data: https://doi.org/10.5281/zenodo.14978544
Related research article	None.

VALUE OF THE DATA

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

- **Historical and spatial insights**. ~~Covering data from 1993 to 2024~~Spanning over three decades, this data set ~~allows for the study~~enables analysis of long-term housing trends ~~in Brunei. The spatial information available enables analysis~~ at various administrative levels in Brunei, providing insights into ~~local patterns and~~ urban development patterns, regional differences, and the evolution of housing characteristics over time. This historical depth is particularly valuable given the lack of ~~previous~~ data ~~before~~prior to the establishment of the Residential Property Price Index (RPPI) [~~8~~5] in 2015.

- ~~**Influence on economic and monetary policies**. Analysing the real estate market is crucial because the RPPI can potentially play a key role in shaping monetary policy and assessing economic stability. Changes in RPPI signal inflationary pressures, providing guidance to the central bank decisions. Developing a house price index using advertised prices, as explored by [9], aligns with practices in other countries, such as the UK's House Price Index developed by Rightmove PLC [10,11]. This data set demonstrates how computational methods can automate what is typically a time-consuming and labour-intensive process.~~

- **Methodological innovation in data curation**. The data collection process employs a unique combination of manual transcription from archival sources, automated web scraping, and AI-based data cleaning using Large Language Models (LLMs). This multi-method approach offers a reproducible framework for assembling complex data sets, which can be adapted for similar data collection efforts in other domains, thereby advancing best practices in the field.

- **Opportunities for handling of missing data**. ~~While the data set has complete information for key variables such as price, date, and spatial variables,~~ The data set includes instances of missing ~~information on~~data, particularly in house characteristics ~~creates~~, which present opportunities for further research~~. Evidently, house characteristics are inherently correlated (see Figure 6), so imputation from observed correlations seems highly promising. With the spatial information present, this opens avenues~~ on data imputation techniques. Researchers may explore advanced statistical and machine learning methods to address missing values and improve data reliability. The structured format and detailed documentation of the data set support such methodological investigations, making it a valuable testbed for developing ~~methodological~~ and evaluating new approaches ~~for~~to handling ~~missing data, encouraging innovation and exploration~~incomplete data in ~~this area.~~

- ~~**Application of Large Language Models (LLMs)**. This project highlights an innovative use of Large Language Models (LLMs) for data cleaning. The LLM was employed to extract structured information from unstructured text descriptions, achieving an accuracy rate of 93%. This method significantly reduced the time and effort required for data cleaning. It can also be applied to other data sets with unstructured text, such as social media or document archives, to extract valuable information for analysis~~real-world data sets.

BACKGROUND

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

Currently, the Brunei Darussalam Central Bank (BDCB) produces a Residential Property Price Index (RPPI) [~~8~~5] using data sourced from financial institutions, such as bank loan data sets. While the RPPI is published quarterly, the underlying raw data is not publicly available due to privacy restrictions. This limits research opportunities and transparency in understanding broader housing market trends. Furthermore, since the RPPI only began in 2015, historical housing data for Brunei is lacking.

We address these challenges by providing a cost-effective and timely means to collect and analyse housing market data. Covering records from 1993 onward, it offers historical depth that complements–and extends beyond–the RPPI. It is valuable not only for tracking property price trends but also for advancing research in economics, urban planning, and real estate, supporting informed decision-making across sectors.

DATA DESCRIPTION

The data has been curated into a single Comma-Separated Values (CSV) file named hspbn_2024-12-12.csv. The data set contains 31,~~495~~116 property listing records which are enriched with area-level geotagged spatial information, spanning a period of ~~32~~33 years from Mar 1993 to ~~Dec 2024~~Feb 2025. The 18 columns of this data set capture information for each property listing as detailed in Table 1 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.

Property Characteristics

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

Brunei's private residential property market offers a variety of options, including detached houses, townhouses, and apartments [~~5~~4]. Based on this diversity, property types have been categorised into four main groups–Detached, Semi-Detached, Terrace, and Apartment–to accurately reflect the ~~diversity~~range of ~~property types~~residential properties in Brunei. Additionally, there are a small number of records that reflect listings for land, which are categorised accordingly as "Land".

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

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

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

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

Table 2: Summary of housing data.

Variable	N	Overall N = 31,495[1]	Brunei-Muara N = 28,894[1]	Belait N = 1,513[1]	Tutong N = 790[1]	Temburong N = 298[1]
Price (BND 1,000)	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)
Property type	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%)
Land tenure	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%)
Development status	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%)
Plot area (acres)	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)

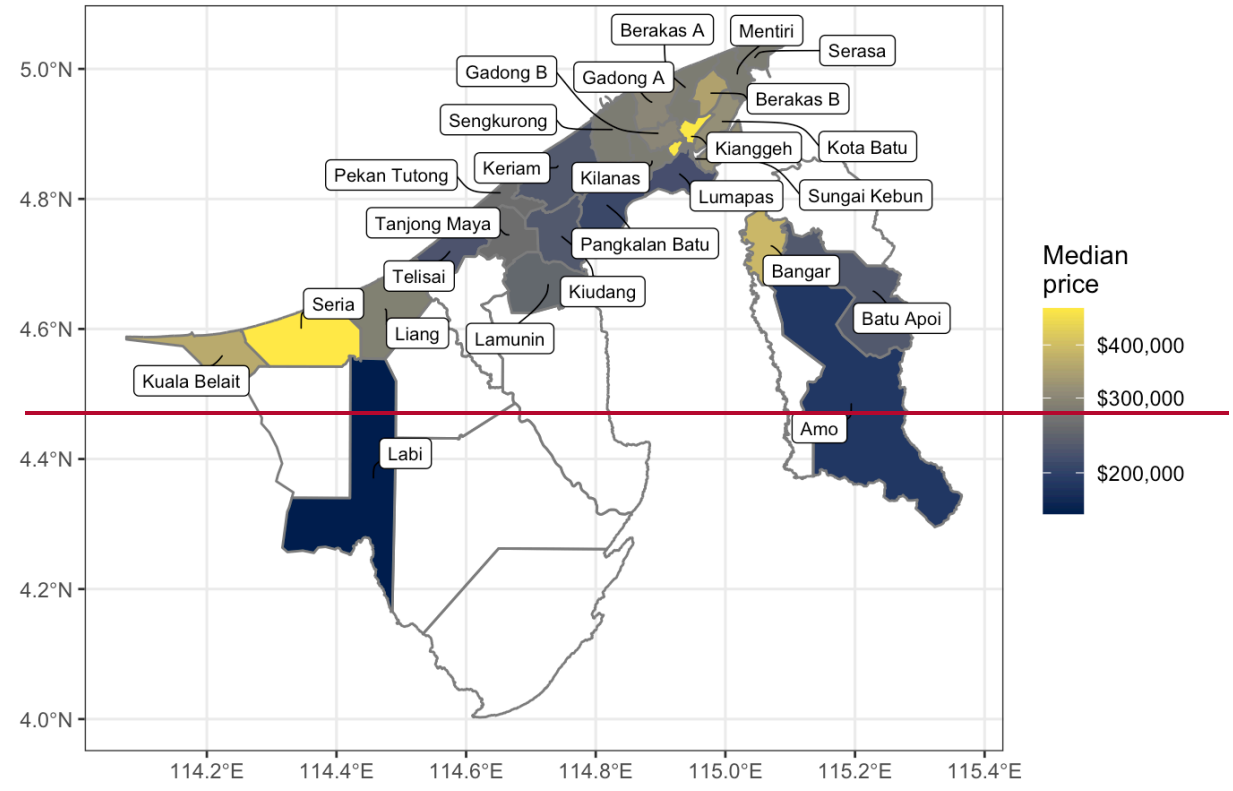
Floor area (sq. ft.)	16,863					
— Mean (SD)		2,590 (1,045)	2,616 (1,061)	2,418 (914)	2,130 (629)	2,796 (741)
— Min - Max		500 - 14,411	500 - 14,411	600 - 7,500	1,100 - 7,000	950 - 3,700
— Median (Q1, Q3)		2,410 (2,000, 3,000)	2,435 (2,000, 3,000)	2,207 (1,700, 2,800)	2,013 (1,826, 2,450)	3,031 (2,790, 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)

[1] n (%)

Spatial Information

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

Importantly, the names of the kampongs, mukims, and districts have been harmonised with a standardised naming convention to ensure consistency across the data set. This also allows for ease of integration into Geographic Information Systems (GIS) software for spatial analysis and visualisation, namely the {`bruneimap`} R package [~~12~~6].



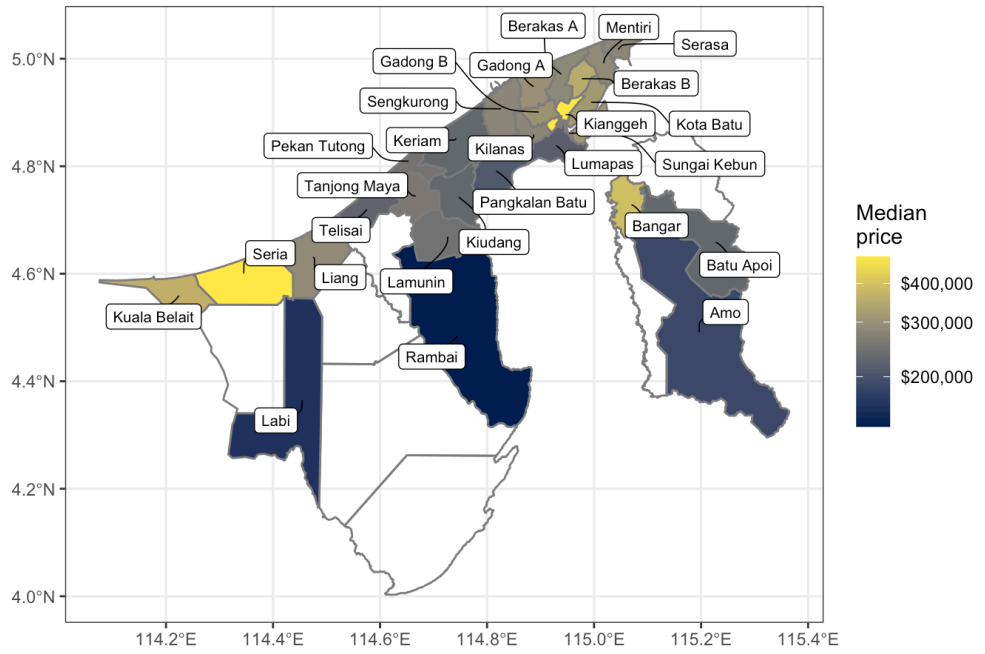


Figure 1: Spatial distribution of median property prices by mukim.

152

Table 2: Summary of housing data.

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

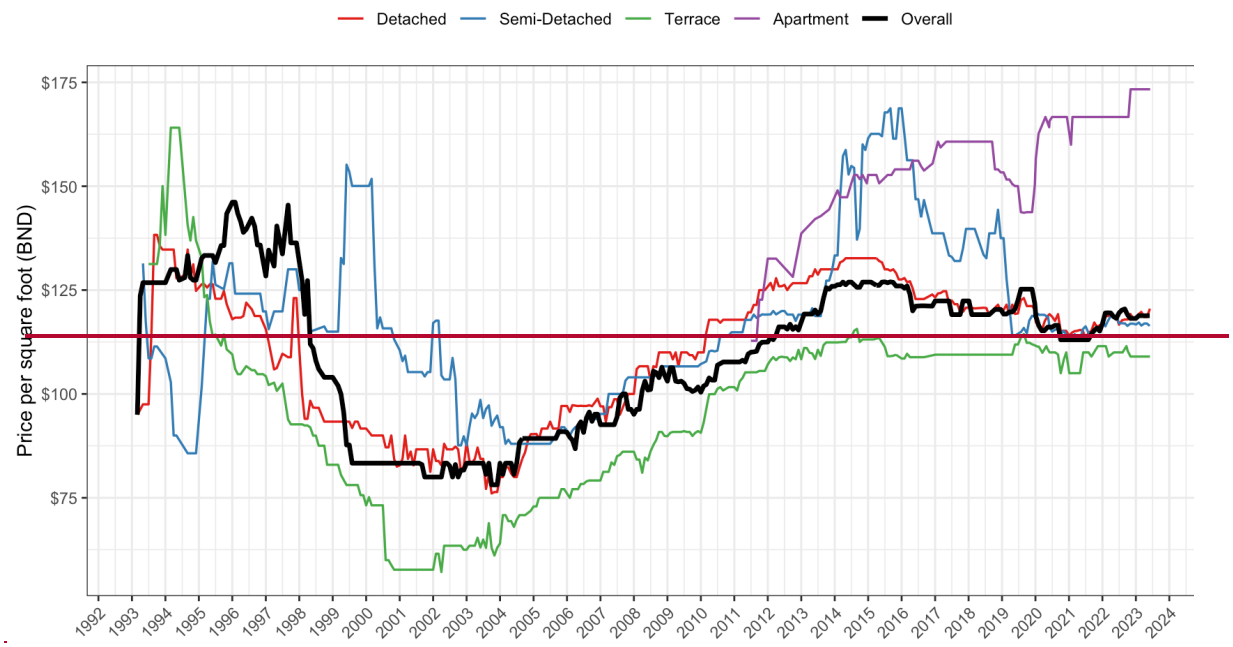
Plot area (acres)	23,368					
Mean (SD)		0.16 (0.12)	0.15 (0.11)	0.19 (0.15)	0.18 (0.17)	0.23 (0.21)
Min - Max		0.01 - 2.00	0.01 - 1.69	0.01 - 1.01	0.04 - 2.00	0.05 - 0.96
Median (Q1, Q3)		0.13 (0.08, 0.19)	0.13 (0.08, 0.19)	0.13 (0.06, 0.27)	0.14 (0.10, 0.21)	0.16 (0.13, 0.26)
Floor area (sq. ft.)	16,665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

[1]n (%)

Listing Dates

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

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



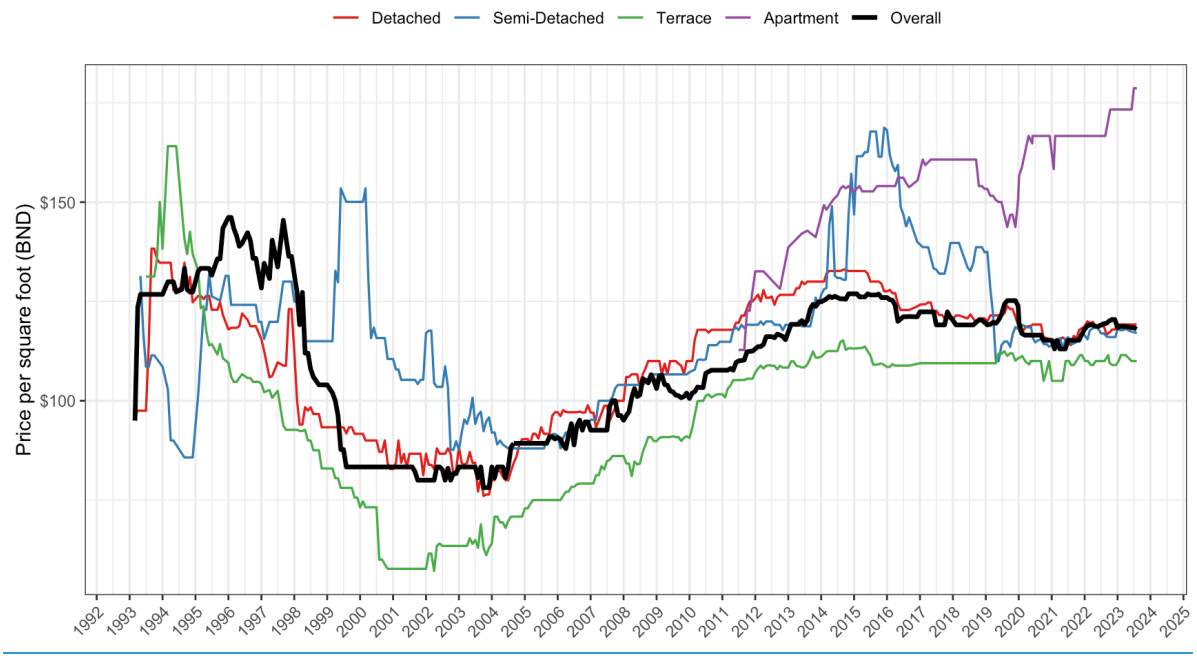


Figure 2: Median smoothed prices per square foot by property type using a 24-month (8-quarter) rolling window.

Missing Values

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

Missing values are represented by blank cells in the CSV file, and the severity of ~~missing values~~missingness is ~~summarised~~summarized in Table 3. In ~~summary~~total, 10.~~5~~1% of the records contain missing values for all key house characteristics (i.e. plot area, floor area, beds, and baths), which, depending on the research question, may ~~necessitate~~require imputation or the exclusion of these records.

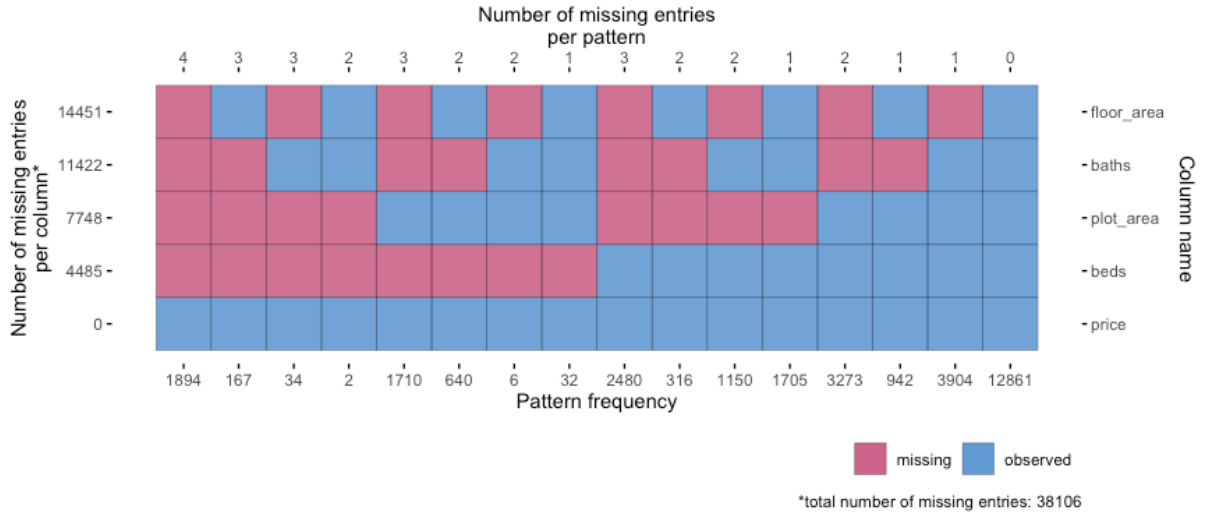
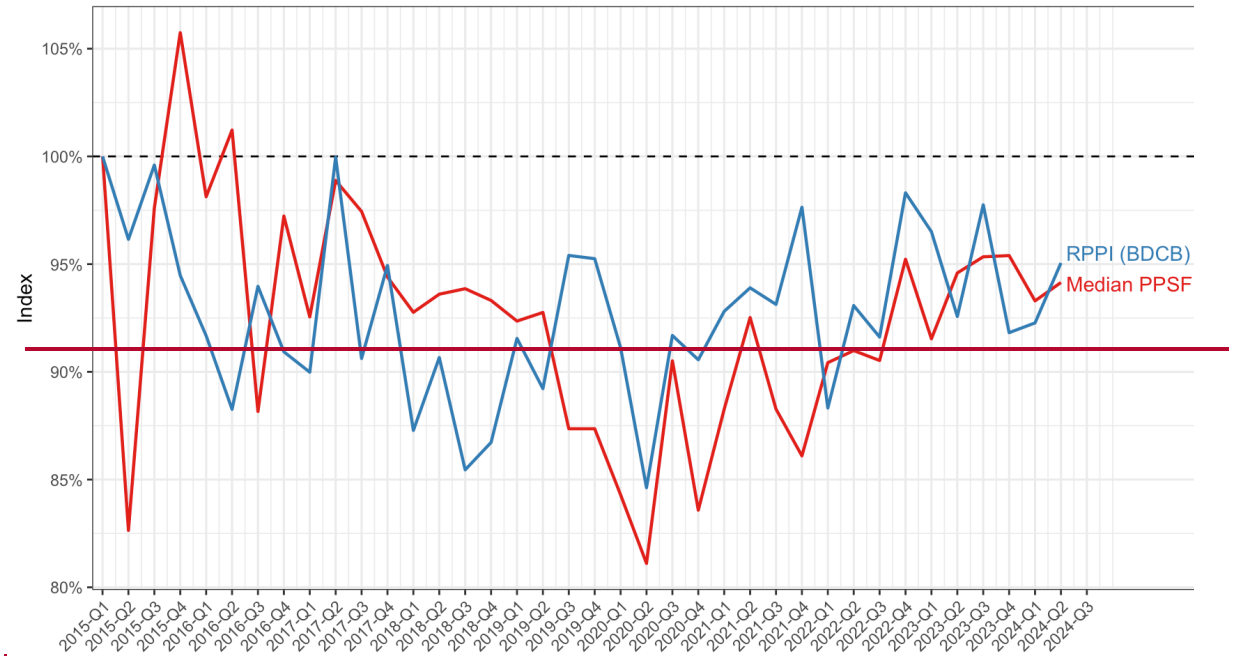


Figure 3: Missing data patterns for key house characteristics.

~~Comparison to RPPI Data~~

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

EXPERIMENTAL DESIGN, MATERIALS AND METHODS

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

All analyses were conducted using the R programming language [~~13~~7], with specific packages used described in each subsection below.

Table 3: Data availability by year.

	Count	Spatial coverage (mukim)[1]	Missing data severity		Data source			LLM post-processing
			Property Type[2]	Property Characteristics[3]	National Archive	Online Archive	Web Scraping	
1993	~~417~~400	~~23.1~~33.3%	0.0%	19.~~2~~0%	✓			
1994	~~654~~653	~~35~~51.9%	65.~~7~~8%	~~28.0~~27.9%	✓			
1995	~~669~~668	~~48.7~~70.4%	66.8%	21.~~2~~3%	✓			
1996	~~563~~561	~~35~~51.9%	69.~~8~~7%	12.1%	✓			
1997	385	~~35~~51.9%	38.4%	26.8%	✓			
1998	345	~~33.3~~48.1%	36.8%	28.7%	✓			
1999	~~322~~317	~~35~~51.9%	31.~~4~~9%	~~27.3~~26.2%	✓			
2000	~~379~~378	~~43.6~~63.0%	0.8%	4.2%	✓			
2001	~~344~~342	~~43.6~~63.0%	0.3%	2.3%	✓			
2002	~~443~~437	~~43.6~~63.0%	0.0%	20.~~1~~4%	✓			
2003	~~454~~449	~~46.2~~66.7%	0.0%	13.~~2~~4%	✓			
2004	~~442~~440	~~43.6~~63.0%	0.0%	19.~~0~~1%	✓			
2005	~~496~~493	~~46.2~~66.7%	0.0%	13.~~3~~2%	✓			
2006	~~661~~653	~~41.0~~59.3%	0.2%	11.3%	✓			
2007	~~644~~638	~~38.5~~55.6%	0.0%	~~13.0~~12.9%	✓			
2008	~~699~~687	~~41.0~~59.3%	0.3%	5.~~7~~8%	✓			
2009	~~542~~531	~~35~~51.9%	0.2%	~~3.9~~4.0%	✓			
2010	~~578~~571	~~38.5~~55.6%	0.0%	2.1%	✓			
2011	~~605~~594	~~38.5~~55.6%	0.2%	~~9.9~~10.1%	✓			
2012	~~937~~934	~~43.6~~63.0%	8.0%	4.3%	✓		✓	
2013	~~888~~882	~~41.0~~59.3%	2.9%	26.~~6~~4%	✓	✓	✓	
2014	~~710~~709	~~46.2~~66.7%	10.6%	8.0%		✓	✓	
2015	~~1121~~868	~~46.2~~66.7%	~~14.4~~12.6%	~~4.9~~6.3%		✓	✓	
2016	~~1462~~1461	~~48.7~~70.4%	13.6%	4.6%		✓	✓	
2017	~~1643~~1638	~~48.7~~70.4%	14.~~4~~5%	4.~~6~~5%		✓	✓	
2018	~~2880~~2646	~~46.2~~66.7%	17.~~8~~6%	0.0%			✓	
2019	~~3596~~3586	~~43.6~~63.0%	15.9%	0.0%			✓	
2020	~~1397~~1363	~~46.2~~66.7%	10.~~2~~4%	0.0%			✓	✓
2021	~~1163~~1115	~~53~~77.8%	~~1.9~~2.0%	0.0%			✓	✓
2022	~~1324~~1235	~~53~~77.8%	3.~~4~~6%	0.0%			✓	✓
2023	~~1635~~1593	~~53~~77.8%	2.~~7~~8%	0.0%			✓	✓
2024	~~3097~~2972	~~53~~77.8%	4.~~4~~6%	0.1%			✓	✓
2025	572	59.3%	14.7%	0.0%			✓	✓
~~mean~~Mean	~~984~~943	~~42.9~~62.0%	13.5%	10.~~5~~1%	—	—	—	—

~~[1]Total number of mukims in Brunei = 39.~~[1]Of Brunei's 39 mukims, only 27 are considered transactable—excluding water villages and remote, non-developable areas.

[2]Unknown property type.

[3]Missing all of plot area, floor area, beds, and baths variables.

Manual Data Collection

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

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

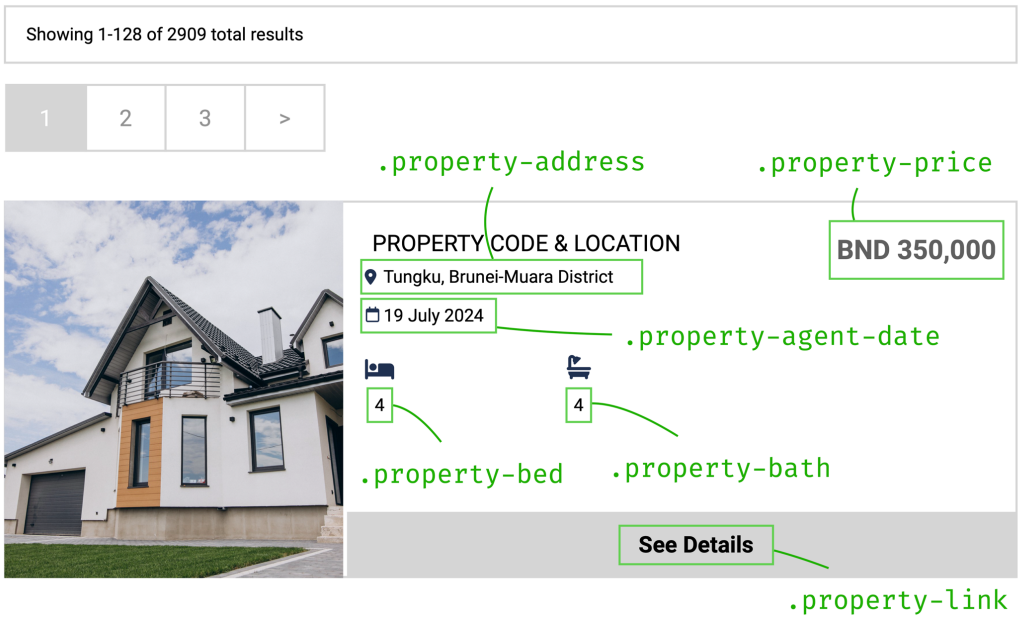
Although daily newspapers from 1993 onward were available at the National Archive, the classified sections were not always present. From 1993 to 1999, property advertisements were found only in Friday editions, and occasionally on Saturdays. Thus, newspapers from both these days were reviewed weekly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

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

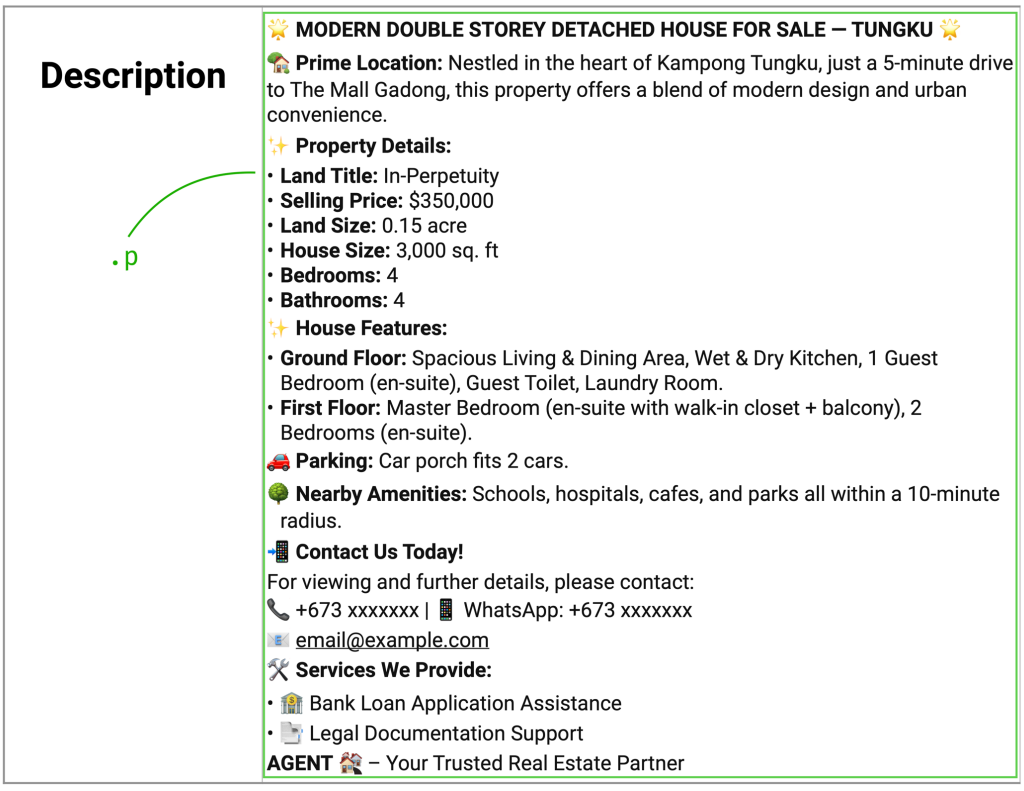
Web Scraping

To compile additional property data for the study beyond manual data collection, web scraping was employed using the R programming language, making use of the {rvest} package [~~14~~8]. This method enabled the systematic extraction of structured information from various local property listing websites such as panvilla.com (now defunct), bruhome.com, and bruneiproperty.com.bn. Such websites provide extensive details on properties listed as “for sale” in Brunei, aggregating advertisements from real estate agents and property developers.

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



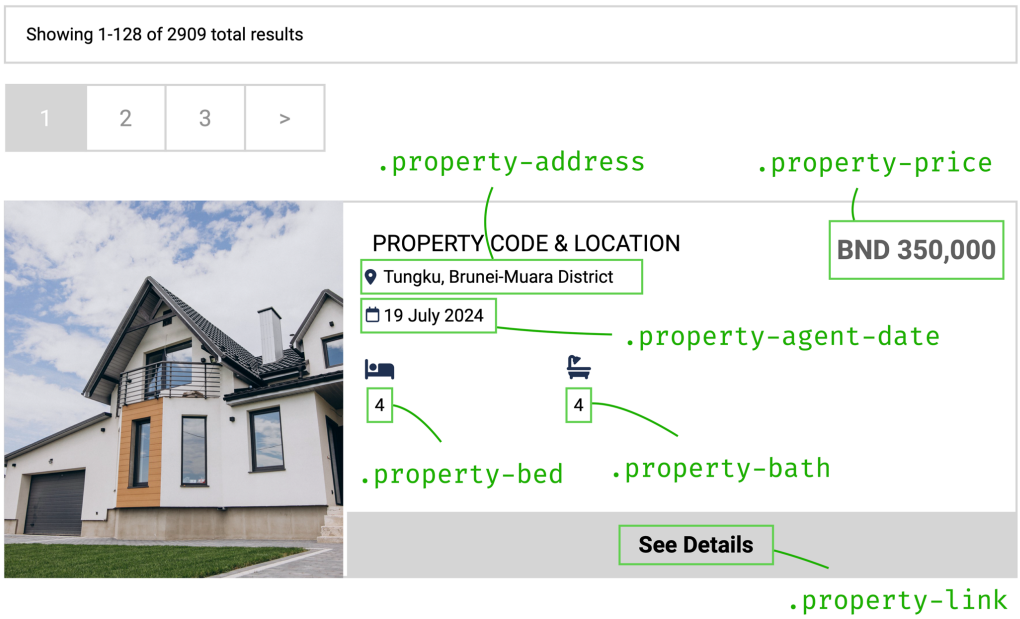
(a)



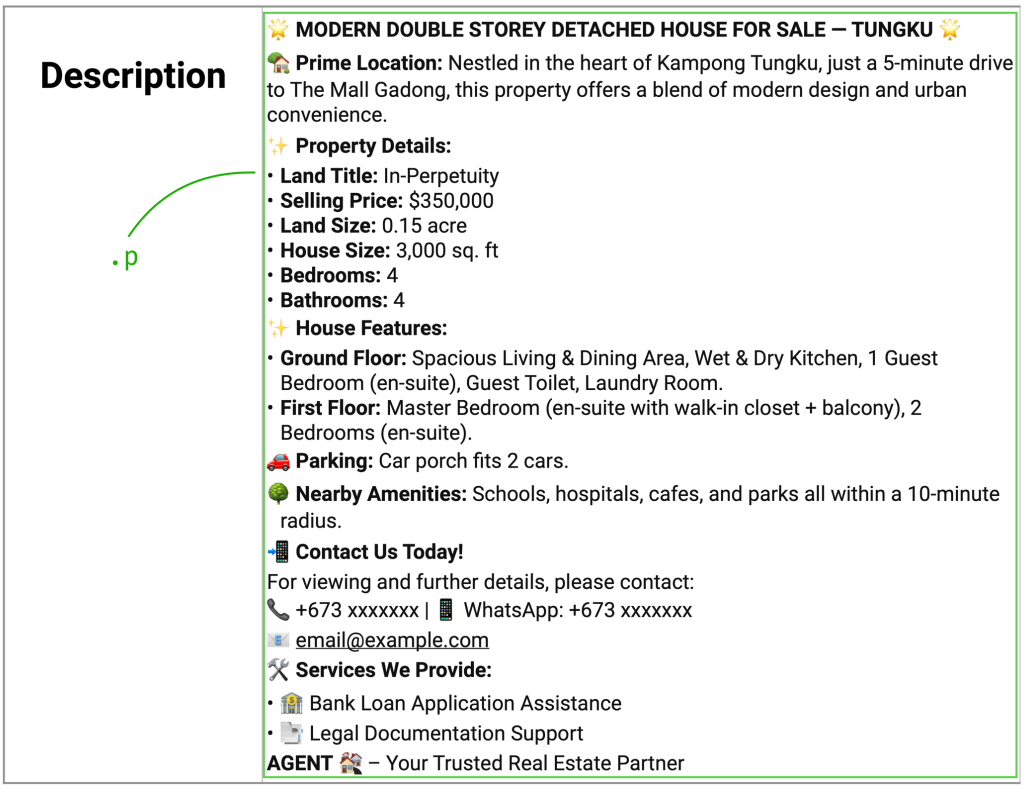
(b)

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

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



~~(a)~~



~~(b)~~

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

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

- `price` variables were cleaned by removing non-numeric characters and converted to integers.
- `beds` and `baths` were converted to integers.
- `date` variables were formatted properly as Date objects.
- Locations were stored as text strings. See the subsection below on spatial data harmonisation ~~for more details~~.
- Any additional information was extracted from the property descriptions and saved as a character vector. This very often contained valuable insights not captured in the primary fields.

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

LLM Data ~~Cleaning~~Extraction

As previously mentioned, the web scraping process also captured unstructured information from the property descriptions, which often contained valuable details not captured in the primary fields. In this subsection, we detail the data ~~cleaning~~extraction process using a pre-trained ~~large language model~~Large Language Model (LLM) to extract structured information from the unstructured text. The LLM used was ~~Llama 3.1 with 8B parameters [15~~the DeepSeek R1 distilled Qwen 14B [9,10], accessed using the {~~tidychatmodels~~ellmer} R package [~~16~~11] via the Ollama[1] API, a local interface platform to the LLM.

The primary goal was to extract the house characteristics of interest, specifically variables 7 to 15 as per Table 1, from the unstructured verbose descriptions scraped from property listing websites. Each description was processed with a carefully designed prompt (Figure 5) to ensure consistent output. This prompt instructed the model to return only the required information in a semicolon-separated format, while handling edge cases such as missing descriptions~~,~~ or non-residential (commercial) properties~~, or rental property advertisements~~.

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

```
"The following is the description from a property sale listing in Brunei. This de
scription will contain the information about the property, including its characte
ristics, 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 yo
u 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, A
partment, 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 re
ntal 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.

The verbose descriptions were fed into the model one at a time using a loop, with the LLM extracting and returning the relevant details. Note that this loop was not parallelised, as the LLM already utilises multiple cores, and further parallelisation would not yield significant efficiency gains due to ~~the computational resources required by the LLM~~resource constraints. The cleaned results were then parsed and stored in a data frame, which was then subjected to manual data-type validation to ensure conformity with the existing data set~~.~~ (see the last subsection). It takes, on average, ~~2.12~~81.7 seconds to process a single description ~~running on Apple MacBook Air M2 Silicon on Chip (SoC) and System in Chip (SiP) processors~~using a Mac Pro 3.2GHz 16 core Intel Xeon W with ~~16GB~~48GB of DDR4

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

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

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

Spatial Data Harmonisation

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

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

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

2. Unknown entries, where the correct kampong could sometimes be inferred from the geographical context; otherwise, these were set to NA.

3. Multiple matches, occurring when two or more kampongs shared the same name (e.g., Kampong Panchor in Mukim Mentiri and Kampong Panchor in Mukim Lumapas). Additional information from the listing was used to determine the correct match, but where this was not possible, these entries were also set to NA.

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

Data Validation

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

1. **Outlier detection.** Summary statistics analyses to identify and flag anomalous values. For instance, a built-up area recorded as 0.1 square feet or an implausibly high number of beds and baths (e.g., >20) would highly likely indicate a possible error, prompting manual review. Such anomalies were flagged and manually reviewed for correction.

2. **Internal consistency checks.** This made use of substantive knowledge about Brunei's housing market [13]. An example is using the price per square foot indicator, whose value typically falls within a known range. Therefore any deviations were scrutinised to ensure that variables such as price and floor area were correctly recorded. This was similarly applied to other variables such as plot area, beds, and baths.

3. **Duplicate records detection.** We also performed duplicate record detection to identify any repeated entries that might have arisen from overlapping data sources or transcription errors. Any duplicates that were identified were carefully reviewed and removed to ensure that each property listing was uniquely represented in the span of one calendar month.

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

LIMITATIONS

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

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

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

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

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

ETHICS STATEMENT

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

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

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

CRediT AUTHOR STATEMENT

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

ACKNOWLEDGEMENTS

The authors gratefully acknowledge the contributions of Atikah Farhain Yahya, Nurulhanisah Abdul Manan, and Nina Zuhairi towards the collection and processing of the data contained within. We also

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

DECLARATION OF COMPETING INTERESTS

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

REFERENCES

~~[1] V. Shabunko, C.M. Lim, S. Brahim, S. Mathew, Developing building benchmarking for Brunei Darussalam, Energy and Buildings 85 (2014) 79–85. https://doi.org/10.1016/j.enbuild.2014.08.047.~~

[~~2~~1] M.K.M. Ng, Z. Shabrina, B. Buyuklieva, Characterising Land Cover Change in Brunei Darussalam's Capital District, Applied Spatial Analysis and Policy 15 (2022) 919–946. https://doi.org/10.1007/s12061-021-09429-9.

[~~3~~2] H. Jamil, F. Usop, H.M. Ramli, Leveraging Sparse Gaussian Processes for Property Price Modelling and Sustainable Urban Planning, in: S.A. Abdul Karim, A. Baharum (Eds.), Intelligent Systems of Computing and Informatics in Sustainable Urban Development, Taylor and Francis/CRC Press, 2025.

[~~4~~3] N.H. Hassan, I. Azrein, K. Ibrahim, G. Yong, Cultural Consideration in Vertical Living in Brunei Darussalam Cultural Consideration in Vertical Living in Brunei Darussalam, in: Managing Urban Growth: Challenges for Small Cities, 2011.

[~~5~~4] N.H. Hassan, The Sociocultural Significance of Homeownership in Brunei Darussalam, in: L. Kwen Fee, P.J. Carnegie, N.H. Hassan (Eds.), (Re)presenting Brunei Darussalam: A Sociology of the Everyday, Springer Nature, Singapore, 2023: pp. 185–206. https://doi.org/10.1007/978-981-19-6059-8_11.

~~[6] F. Braesemann, A. Baum, PropTech: Turning Real Estate Into a Data-Driven Market?, SSRN Electronic Journal (2020). https://doi.org/10.2139/ssrn.3607238.~~

~~[7] J.R. DeLisle, B. Never, T.V. Grissom, The big data regime shift in real estate, Journal of Property Investment & Finance 38 (2020) 363–395. https://doi.org/10.1108/JPIF-10-2019-0134.~~

[~~8~~5] BDCB, Technical Notes for Residential Property Price Index (RPPI), Brunei Darussalam Central Bank, 2021.

~~[9] A.N. Rachman, Residential property price index for Indonesia using big data: The case of Jakarta, in: International Conference on Real Estate Statistics, 2019.~~

~~[10] HM Land Registry, Comparing house price indices in the UK, HM Land Registry, 2023.~~

~~[11] European Commission. Eurostat, Handbook on residential property prices indices (RPPIs)., Publications Office, LU, 2013.~~

[~~12~~6] H. Jamil, Bruneimap: Maps and Spatial Data of Brunei (R package version 0.3.1.9001), 2024.

[~~13~~7] R Core Team, R: A language and environment for statistical computing, R Foundation for Statistical Computing, Vienna, Austria, 2024.

[148] H. Wickham, Rvest: Easily harvest (scrape) web pages, 2024.

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

~~Zhang, L. Guo, L. Yu, L. Moshkovich, L. Wehrstedt, M. Khabsa, M. Avalani, M. Bhatt, M. Mankus, M. Hasson, M. Lennie, M. Reso, M. Groshev, M. Naumov, M. Lathi, M. Keneally, M. Liu, M.L. Seltzer, M. Valko, M. Restrepo, M. Patel, M. Vyatskov, M. Samvelyan, M. Clark, M. Macey, M. Wang, M.J. Hermoso, M. Metanat, M. Rastegari, M. Bansal, N. Santhanam, N. Parks, N. White, N. Bawa, N. Singhal, N. Egebo, N. Usunier, N. Mehta, N.P. Laptev, N. Dong, N. Cheng, O. Chernoguz, O. Hart, O. Salpekar, O. Kalinli, P. Kent, P. Parekh, P. Saab, P. Balaji, P. Rittner, P. Bontrager, P. Roux, P. Dollar, P. Zvyagina, P. Ratanchandani, P. Yuvraj, Q. Liang, R. Alao, R. Rodriguez, R. Ayub, R. Murthy, R. Nayani, R. Mitra, R. Parthasarathy, R. Li, R. Hogan, R. Battey, R. Wang, R. Howes, R. Rinott, S. Mehta, S. Siby, S.J. Bondu, S. Datta, S. Chugh, S. Hunt, S. Dhillon, S. Sidorov, S. Pan, S. Mahajan, S. Verma, S. Yamamoto, S. Ramaswamy, S. Lindsay, S. Lindsay, S. Feng, S. Lin, S.C. Zha, S. Patil, S. Shankar, S. Zhang, S. Zhang, S. Wang, S. Agarwal, S. Sajuyigbe, S. Chintala, S. Max, S. Chen, S. Kehoe, S. Satterfield, S. Govindaprasad, S. Gupta, S. Deng, S. Cho, S. Virk, S. Subramanian, S. Choudhury, S. Goldman, T. Remez, T. Glaser, T. Best, T. Koehler, T. Robinson, T. Li, T. Zhang, T. Matthews, T. Chou, T. Shaked, V. Vontimitta, V. Ajayi, V. Montanez, V. Mohan, V.S. Kumar, V. Mangla, V. Ionescu, V. Poenaru, V.T. Mihailescu, V. Ivanov, W. Li, W. Wang, W. Jiang, W. Bouaziz, W. Constable, X. Tang, X. Wu, X. Wang, X. 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. Zhao, Y. Hao, Y. Qian, Y. Li, Y. He, Z. Rait, Z. DeVito, Z. Rosnbrick, Z. Wen, Z. Yang, Z. Zhao, Z. Ma, The Llama 3 Herd of Models, (2024). https://doi.org/10.48550/arXiv.2407.21783.~~

~~[16] A. Rapp, Tidychatmodels~~[9] DeepSeek-AI, DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning, (2025). https://doi.org/10.48550/arXiv.2501.12948.

[10] Qwen, Qwen2.5 Technical Report, (2025). https://doi.org/10.48550/arXiv.2412.15115.

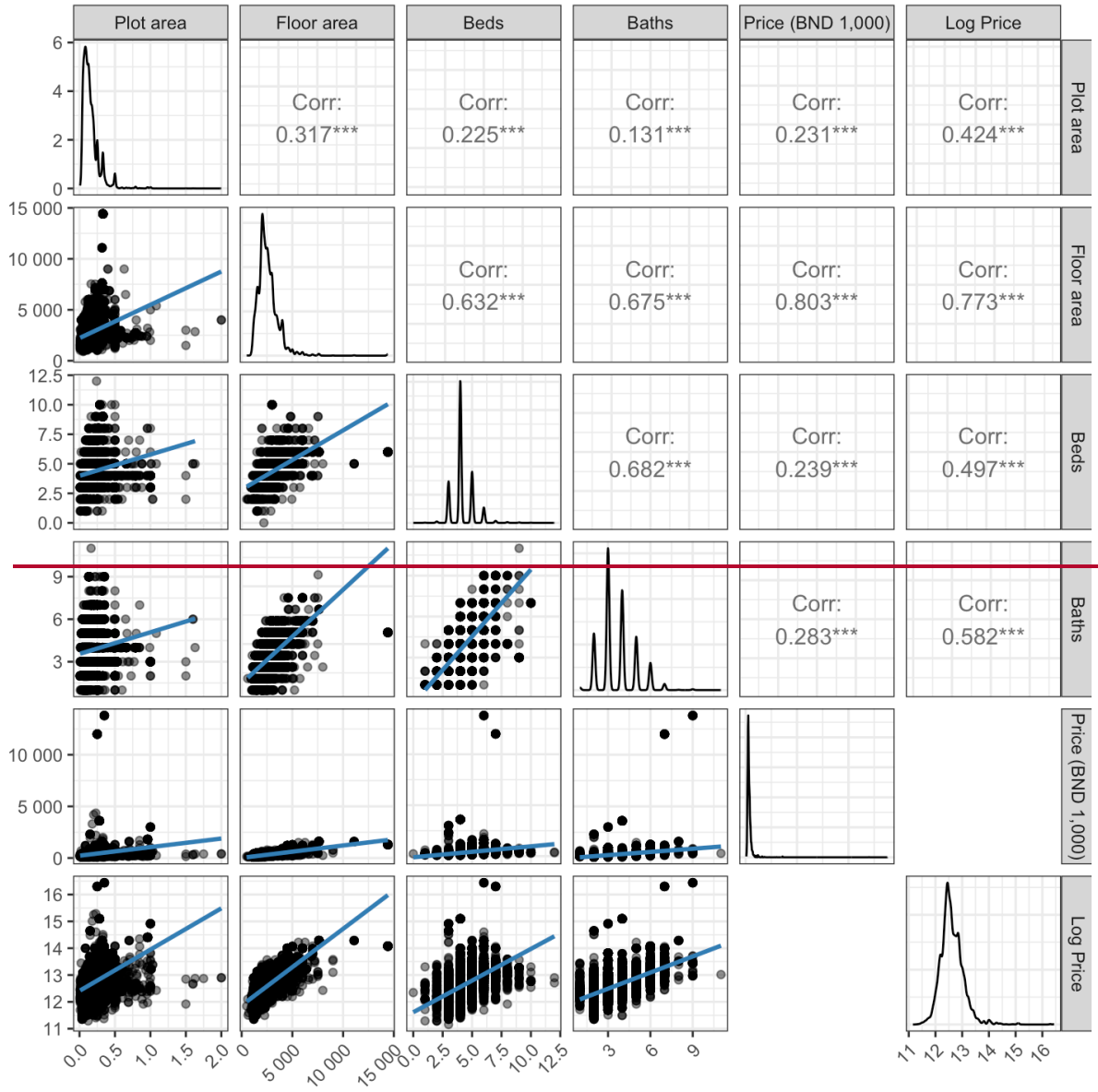
[11] H. Wickham, J. Cheng, A. Jacobs, Ellmer: Chat with ~~all kinds of AI~~large language models ~~through a common interface, 2024~~, 2025.

[~~17~~12] DEPS, The Population and Housing Census Report (BPP) 2021: Demographic, Household and Housing Characteristics, Department of Economic Planning and Statistics, Ministry of Finance and Economy, Brunei Darussalam, 2022.

[~~18~~13] H. Jamil, A spatio-temporal analysis of ~~property~~house prices in Brunei Darussalam, (~~2024~~2025). https://doi.org/10.13140/RG.2.2.32533.74720.

Appendix

Pairwise Correlation Plot of Numeric Variables



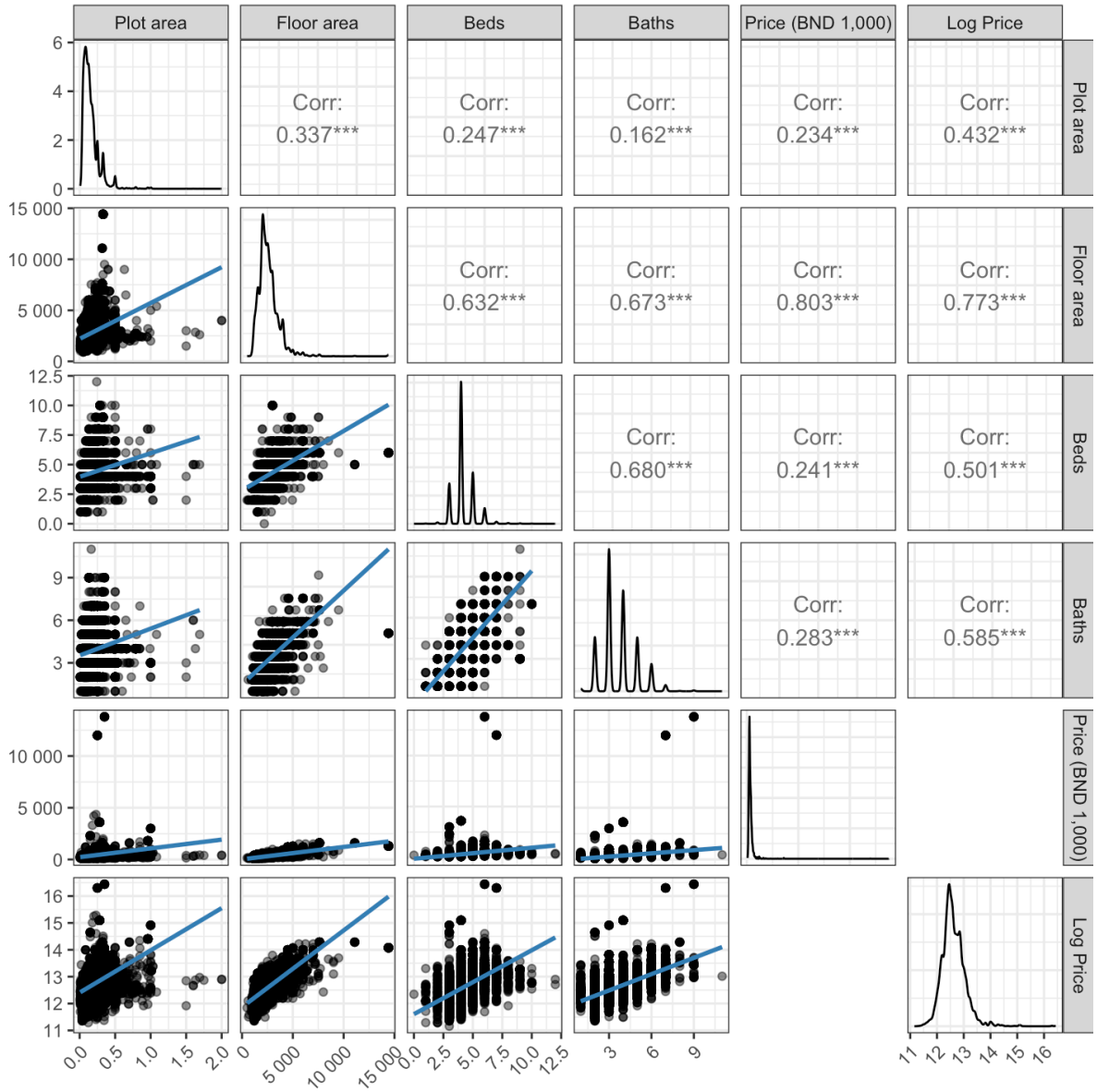


Figure 6: Pairwise correlation plot of continuous variables.

Comparison to RPPI Data

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

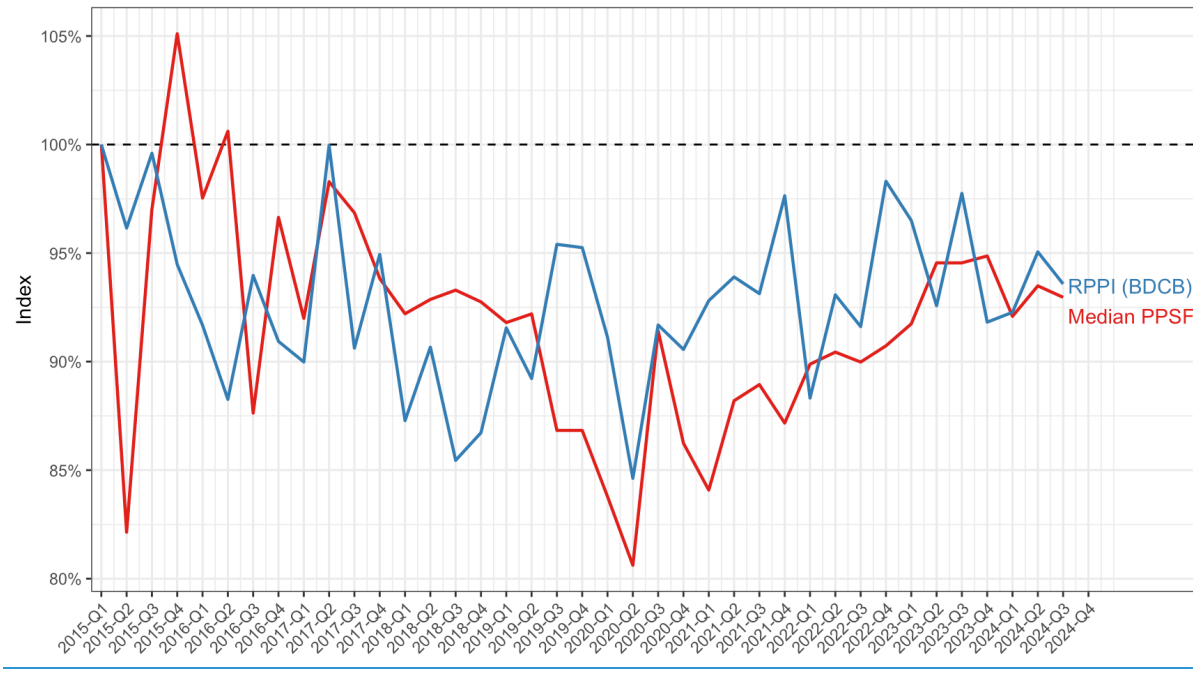


Figure 7: Comparison of quarterly median price per square foot indices (Median PPSF) and the official Residential Property Price Index (RPPI) from Brunei Darussalam Central Bank (BDCB).

LLM Accuracy Test

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

OpenAI's models topped the accuracy charts, with the o1-mini and gpt-4o models achieving the accuracy scores of 99.2% and 98.9% respectively. The reasoning model deepseek-r1:14b was the best performing locally run model, scoring 96.9% accuracy. Evidently, the smaller the model, the less accurate the extraction process (cf. llama3.2 70.8%). Most inaccuracies occurred in the status variable, where models struggled to parse the correct build status from vague advertisement descriptions. However, key variables such as `price`, `type`, `plot_area`, `floor_area`, `beds`, and `baths` are generally reliable (>90% accuracy).

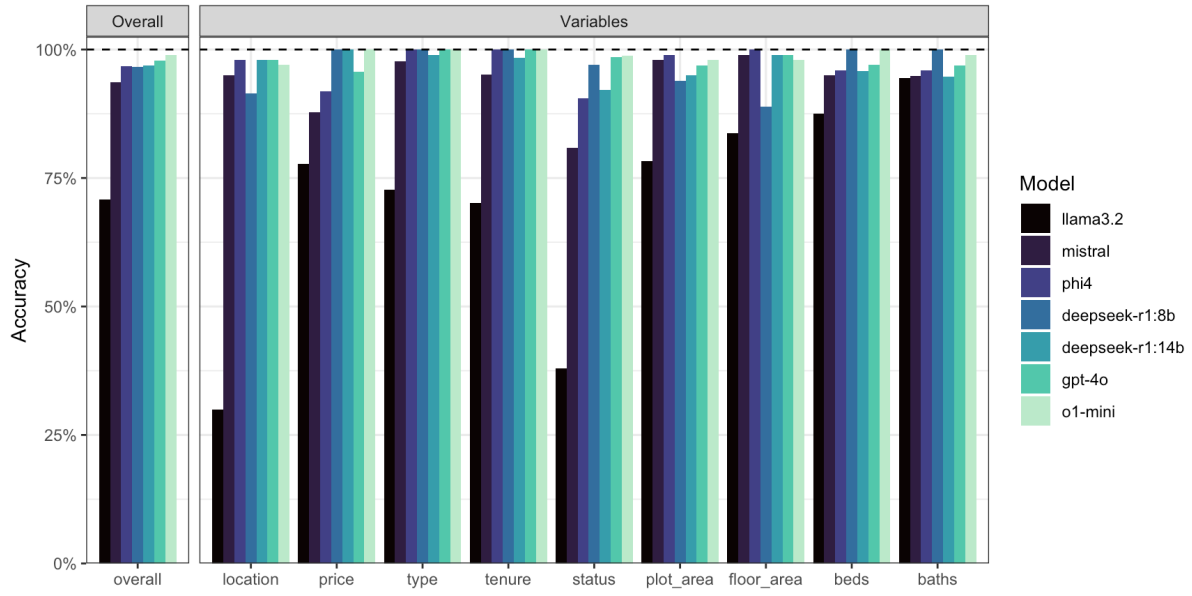


Figure 8: Comparison of data extraction accuracy across multiple LLM models on the test dataset. Each bar represents the percentage of correctly extracted fields for a given model.

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

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

	Time (seconds)			
Model	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