APREF

Australian Property Research and Education Fund

Research Paper

Towards understanding the dynamics of the regional property markets of Australia

Spatio-temporal analysis, modelling, and visualisation of sales in regional SA and NSW

Executive Summary

Housing one third of the population and generating two thirds of the export economy, regional towns and rural areas are a significant part of the Australian socio-economic structure. Although there is substantial policy focus on regional development and improving the well-being of its residents at federal and state levels, much of the detailed, large scale, longitudinal studies on the built environment, real-estate and property markets are often focused on the capital city regions. To tackle this issue, this study uses the detailed, comprehensive, large scale, long-term longitudinal data on property sales made available by the Valuer General offices of New South Wales and South Australia to understand the dynamics of the property markets in regional Australia and the factors affecting them.

The study examines the volume, price, quality, and other characteristics of the regional property market in these states over the last decade to identify the spatio-temporal patterns in their behaviour. The transaction level sales dataset, when aggregated and visualised, proves to be an incredibly useful tool which can enable practitioners in property industry to profile and understand the regional markets clearly and succinctly while providing them opportunity for a more nuanced discussion around valuation and pricing.

The study also attempts to create a simple, clear, and understandable categorisation of the regional markets in terms of their temporal price trends using hierarchical clustering. The study also makes use of the publicly available datasets such as census on population and demography and links it to the associated regional property markets. Finally, the study visualises these results and insights as an interactive, map-based visualisation that can be used by stakeholders of property industry to better understand, manage, and work in these regional markets.

The primary research output is a dynamic, interactive, geo-spatial visualisation showing the spatio-temporal patterns of property sales in regional NSW and SA along with results of the analysis in the form of summaries and KPIs across various statistical areas. The visualisation also showcases the outputs of the clustering showing the distinct types based on the price trends in these regional property markets. The secondary output is an aggregated index dataset which shows the trends in price and volume of sales, clustering results and a brief census profile for each regional area at multiple boundary levels.

Research Highlights

The research used the transaction level data on sales across New South Wales and South Australia to create a longitudinal, geo-coded, harmonized dataset of sales for regional areas from 2010 – 2020.

The transaction level dataset was aggregated at various ABS boundaries and linked with corresponding demographic data from Census 2021 to create a representative property market profile for regional areas in NSW and SA.

The 10-year trend of median prices in the regional areas were hierarchically clustered to create 6 distinct classification which describe the overall movement of prices in corresponding markets.

The research outputs are visualized as an interactive, geospatial visualisation at https://dashboard.regionalpropertymarkets.com/ and the aggregated outputs has been released as 'geo-packages' which combine geospatial data along with other research outputs – https://agg.regionalpropertymarkets.com/

Budget Acquittal

The funds requested from APREF were spent **as budgeted without any changes**. The project received additional support for AWS Cloud infrastructure from UNSW Research Technology (ResTech) grant which was used to lease high performance computing resources.

Payment Type	Details	Amount Budgeted	Amount Spent	Covered by
Equipment	AWS instances for hosting the compute node and dashboard	AUD 1,000	AUD 12,000	ResTech Grant (UNSW)
Staff	• Dr Balamurugan S (0.10 FTE) • Mr. Mohammed H RA (0.10 FTE) • Dr Chyi Lin L (0.02 FTE) • Prof Ali S (0.02 FTE) • Prof Chris P (0.02 FTE) • Mr. Peter R (0.02 FTE)	AUD 10,500 AUD 9,500 AUD 4,000 AUD 5,000 AUD 5,000 AUD 4,000	AUD 10,500 AUD 9,500 AUD 4,000 AUD 5,000 AUD 5,000 AUD 4,000	**APREF Grant** **APREF Grant** In-kind UNSW In-kind UNSW In-kind UNSW In-kind UNSW
TOTAL		AUD 39,000	AUD 50,000	

Article

A step towards understanding the dynamics of the regional property markets of Australia.

By UNSW's Dr Balamurugan Soundararaj, Dr Chyi Lin Lee, Prof Chris Pettit and UniSA's Dr Ali Soltani, Peter Rossini, and Mohammed Heydari.
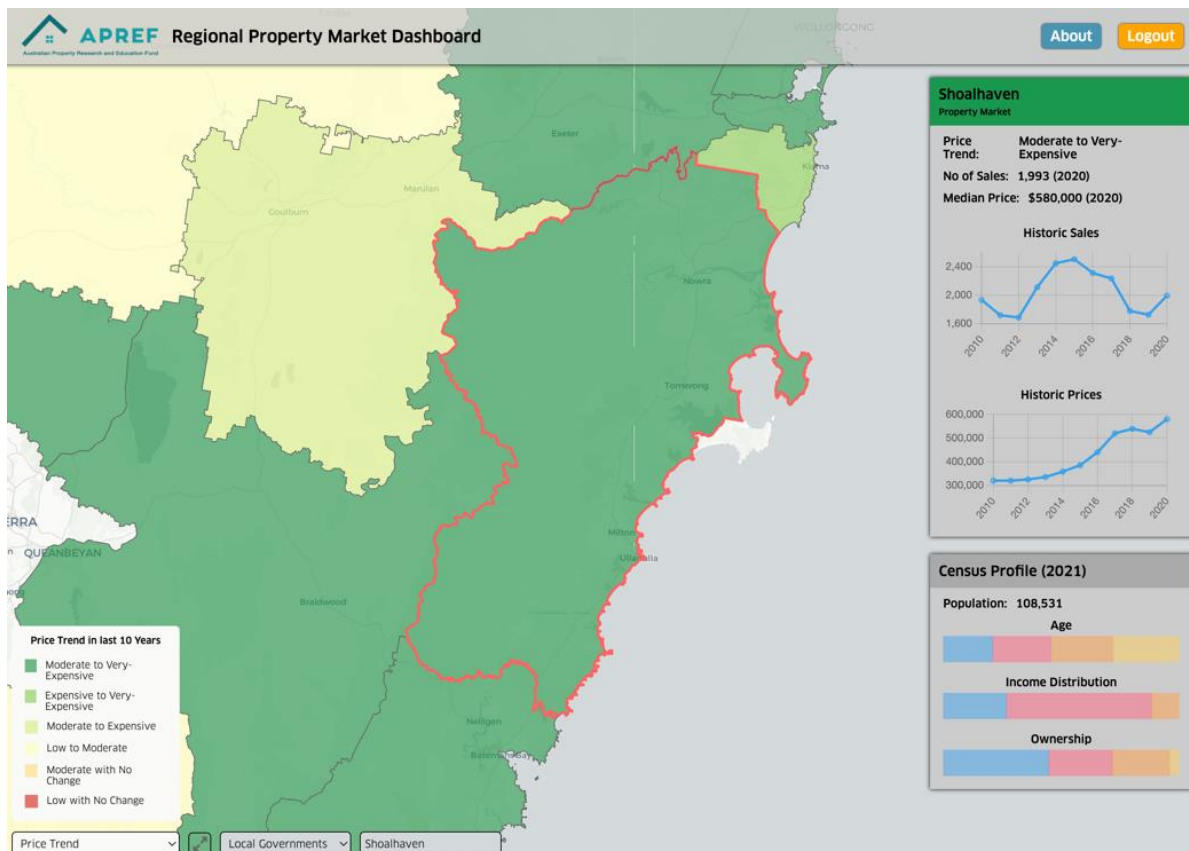
New research uses transaction level data on sales across New South Wales and South Australia to create a harmonized spatio-temporal dataset of property market profiles for regional areas to help practitioners in understanding their dynamics.

Housing one third of the population and generating two thirds of the export economy, regional towns and rural areas are a significant part of the Australian socio-economic structure. Although there is substantial policy focus on regional development and improving the well-being of its residents at federal and state levels, much of the detailed, large scale, longitudinal studies on the built environment, real-estate and property markets are often

focused on the capital city regions. To tackle this issue, researchers have used the detailed, comprehensive data on property sales made available by the Valuer General offices of New South Wales and South Australia to profile and understand the state and dynamics of the property markets across regional areas in these states.

In this study researchers have examined the sales volume, median price, demography, and other characteristics of the regional property market in NSW and SA over the last decade to identify the spatio-temporal patterns in their behaviour. The transaction level sales dataset, when aggregated and visualised, proves to be an incredibly useful tool which can enable practitioners in property industry to profile and understand the regional markets clearly and succinctly while providing them opportunity for a more nuanced discussion around valuation and pricing.

The research team have also created a simple, clear, and understandable categorisation of the regional markets in terms of their temporal price trends using hierarchical clustering. The study also uses the publicly available 2021 Census datasets on population and demography by linking it to the corresponding regional property markets. Most importantly, the results and insights were visualised as an easy to use, interactive, map-based visualisation that can be used by practitioners and broader community to better understand, manage, and work in these regional markets.

What are the main outputs?

The primary output of the project - a dynamic, interactive, multi-level, geo-spatial visualisation which is available at https://dashboard.regionalpropertymarkets.com showing the spatio-temporal patterns of property markets in regional NSW and SA along with results of the analysis in the form of summary profiles such as the median price trend clusters, demographic profiles etc. across various statistical areas. For example, the visualisation immediately shows that the 17 regional LGA which have experienced rapid growth in their median price in the last decade were all located in coastal NSW. On a more detailed inspection, it can show how much of an outlier Kiama and Byron are in terms median price and their change compared to the rest of the state. The secondary output is an aggregated index published at https://agg.regionalpropertymarkets.com which shows the trends in price and volume of sales, clustering results and a brief census profile for each regional area at multiple boundary levels.

Why are they important?

This study is important to the Australian property industry since there is a considerable lack of usable publicly available data and resources regarding regional property markets compared to the Capital Cities. Although few portals and datasets do exist, they are compiled at a state/local government level and distributed in such formats which require some level of expertise in data analysis for generating usable insights. This research is also important to specific aspects of the property industry such as valuation, where understanding the regional context of the property's location is one of the key steps in the process. Outputs from this research will enable property professionals to form a quick, baseline understanding of regional property markets in comparison to others before they go on to investigate them in detail at a suburb or property level. This research is also an important step in building

pathways for translating research into an easily understandable format for the community of public and professionals to create maximum impact in practice. The research team hopes the regional property market dashboard will be widely used and will act as an example showcasing the value in building and expanding such tools and resources at a national level.

What do your findings mean for the Australian property buyers/sellers and/or professionals and how might your findings impact these groups?

Although the scope of this research does not delve into the prescriptive findings for the stakeholders of the property industry in their decision-making process, it forms a base for such insights to be derived on top of it. In this regard, the output aggregated data generated from this research is crucial. This dataset will form a part of and distributed via the National Housing Data Exchange (AHDE) and the technology/platform that powers the dashboard will form a part of the Australian Property Market Explorer (APME) – initiatives from the Australian Housing Data Analytics Platform (AHDAP). These components will be leveraged with the broader resources available to the AHDAP to develop a more comprehensive, national level data infrastructure and ecosystem of tools built on top of it. When developed comprehensively, the outputs of the research outputs can be a key resource for multiple stakeholders in various ways. For example, it can help policy makers in understanding housing affordability and stress, financial institutions in understanding risk, market participants in making better transaction decisions, government in optimising planning, service provision and revenue generation etc.

What are your recommendations and potential for future work?

This project and outputs demonstrate the immense value in utilising the data on sales collected by the revenue departments of state governments when combined with right aggregation, analysis, and visualisation. The authors recommend that datasets be collected across Australia for all states and a nationally harmonised dataset for regional areas be created. Such dataset with a standardised and regularly updated profiles on property market and demography will be invaluable for the Australian property industry. There are multiple avenues in which the research can be taken forward. From the data side, we could look at including more closely related property related datasets such as rental data from housing bonds that are lodged with state government, mortgage related datasets from the central bank etc. From the analysis side we could look in to further strengthening the demographic profiling by using more data points related to property market along with using a more advanced clustering algorithm such as k-means or db-scan. On the visualisation and distribution side, we need to look for ways to sustainably host the outputs for the property industry in the future along with a plan for updating the outputs at regular intervals. When implemented in a right way this can grow an ecosystem around these outputs which will create significant impact for the property industry and regional areas of Australia by promoting a data-driven discussion and decision making in the property market.

Dashboard available at: https://dashboard.regionalpropertymarkets.com/

Aggregated outputs available at: https://agg.regionalpropertymarkets.com/

Research Report

Towards understanding the dynamics of the regional property markets of Australia

Spatio-temporal analysis, modelling, and visualisation of sales in regional SA and NSW

INTRODUCTION

Housing one third of the population and generating two thirds of the export economy, regional towns and rural areas are a significant part of the Australian socio-economic structure. Although there is substantial policy focus on regional development and improving the well-being of its residents at federal and state levels, much of the detailed, large scale, longitudinal studies on the built environment, real-estate and property markets are often focused on the capital city regions. Disruptive global & national economic events, disasters such as bushfires, sea level rise and most recently, the COVID19 pandemic have all had an impact on Australia's property markets. These events along with government's responses to them, are expected to have significant impact to the regional markets because of the long-term changes in people's risk perception, work from home preferences, and so on. In this context, it is critical to understand the dynamics of the property markets in regional Australia and the factors affecting them. This study uses the detailed, comprehensive, large scale, long-term longitudinal data on property sales made available by the Valuer General offices of New South Wales and South Australia to understand the dynamics of the property markets in regional Australia and the factors affecting them. The study examines the volume, price, demographic mix, and other characteristics of the regional property market in these states over the last decade to identify the spatio-temporal patterns in their behaviour.

DATA

A set of comprehensive data on location, contract date and sale price are collected for every sale of property in Australia by the relevant revenue authority at state level for stamp duty purposes. This dataset is often accompanied with further data on the attributes of the property which are collected to assist with accurately valuing it. Although collected by public institutions, these datasets are often not accessible to the public and property practitioners for use because of their scale, format, and method of distribution. This project looks at the sales datasets from two the state government departments from New South Wales (NSW) and South Australia (SA) for understanding the dynamics of regional property markets. The data for all sales in NSW since 1990 has been made available by the Valuer General (VG) office publicly as a bulk download from their website - https://valuation.property.nsw.gov.au/embed/propertySalesInformation. The bulk data extracts of property sales for SA is made available by the VG office through Land Services

SA and subsequent downstream vendors via data request/ procurement process - https://www.sa.gov.au/topics/housing/planning-and--=property/researching-a-property#two. These two datasets form the core of this research.

Being distributed through different channels and procedures, the volume and detail available in these two datasets varied widely. The NSW datasets was comprehensive in terms of coverage but lacked detail in terms of attributes of the property themselves. NSW dataset contained approximately 5,000,000 sales between 1990 – 2020. The SA dataset was much more limited in terms of coverage but detailed in terms of the attributes of the property being sold. SA dataset consisted of approx. 200,000 sales between 2010 – 2020. The NSW dataset lacked any explicit geographic data but included the address of the property in plain text. The SA dataset was geocoded and provided geographic co-ordinates for each property. The data fields covered by each property is shown in Table 1.

Table 1: Data fields available in sales datasets from NSW & SA.

	SOUTH AUSTRALIA	NEW SOUTH WALES
Location	Geometry (Coordinates)	Address (Text)
Sale Details	Sale Price	Sale Price
	Sale Date	Sale Date (contract, settlement)
	Vendor, Vendor Type	
	Purchaser, Purchaser Type	
Land Details	Allotment ID, Plan Reference	Parcel ID
	Land use, Previous Land use	Land Use
	Zone	Zone
	Land Area	Land Area
	Frontage	
	Clearance	
Building Details	Rooms	Property ID
	Storeys	Property Type (house/apt)
	Built Area	
	Condition	
	Roof & Wall Material	
	Year Built	
	Architectural Style	
	Ensuite	
	Improvements	

These raw datasets were then cleaned, filtered, geo-coded, and harmonised to create one consistent sales dataset for regional NSW and SA. The first challenge was to geo-code the NSW data points from the text address field. This was done using a simple, bespoke address matching algorithm which matches the address text to Geocoded National Address File (GNAF) database. The steps of this algorithm included cleaning the text for special characters, standardising abbreviations, removing specialised property names etc. A

summary of this process has been illustrated in Figure 1 and detailed explanation is given in Appendix 1. This matching algorithm resulted in a 96% match with errors most errors occurring in urban areas with changing addresses and large high-rise developments. Then this dataset was filtered for residential properties using a combination of land use, zoning, and property type fields. SA data didn't require similar cleaning as it already geocoded and was filtered for residential properties. Then the dataset was cleaned for invalid or incomplete data, duplicated records, and outliers in terms of price and land size. Finally, the dataset was filtered for sales between 2010 – 2020 to make them consistent between states resulting in a dataset which had approximately 2.3 million properties for NSW and 70,000 sales for SA which forms a median representative spatio-temporal sales data for both the states. It is notable that by having more details on properties, the SA data had more invalid/incomplete records while filtering. A detailed explanation of this filtering process is provided in Appendix 2.
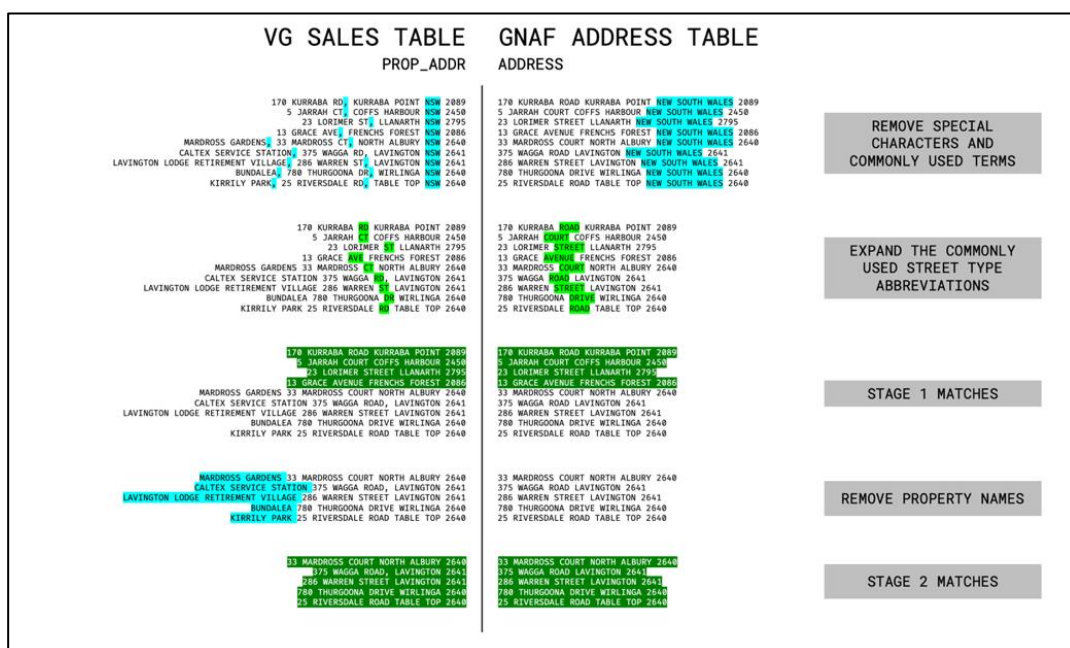

Figure 1: Summary of the algorithm which matches address to geo-code sales data for NSW using GNAF.

AGGREGATION AND ANALYSIS

With the datasets from both SA and NSW cleaned and geo-coded they were compiled in to one consistent dataset with three main data fields – location, sale price and sale date from 2010 – 2020. This point dataset was then linked with 5 levels of ABS boundaries (GCCSA, SA4, SA3, SA2, and LGA) and then aggerated into volume and median price for every year. Although aggregates for further granular boundaries such as SA1 and Suburbs are possible, they were not used to avoid the possibility of reverse engineering of the property level data back from these granular aggregations. This process resulted in 5 spatio-temporal datasets – one for each boundary for all regional areas in both NSW and SA. There are few key things to note in this aggregation process. Firstly, the volume parameter should be treated as a relative indicator rather than absolute number as a very aggressive cleaning of the

dataset was carried out to make sure the data used was clean and consistent. This is particularly significant since significant number of sales were cleaned out in SA for being incomplete or duplicated. Secondly, because of the above-mentioned cleaning and removal of outliers in terms of both land area and price, the median prices reported might differ significantly for few areas. This data when visualised shows us a heat map of property prices in space and how it evolved over the past decade. Figure 2 shows choropleth visualisations of the median price at LGA level in NSW and SA in 2010 and 2020 highlighting the significant change in the median prices in the east coast of NSW.
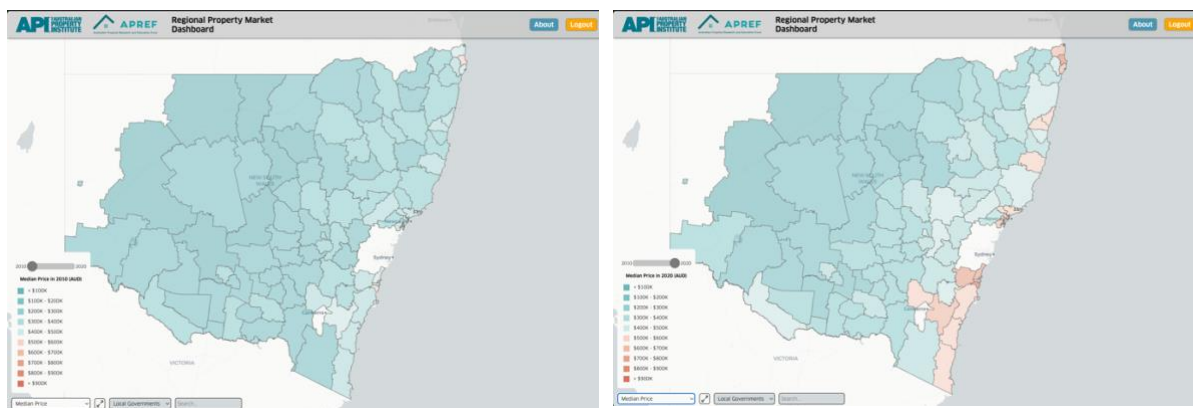

Figure 2: Median price across LGAs in NSW and SA in 2010 (left) and 2020 (right)
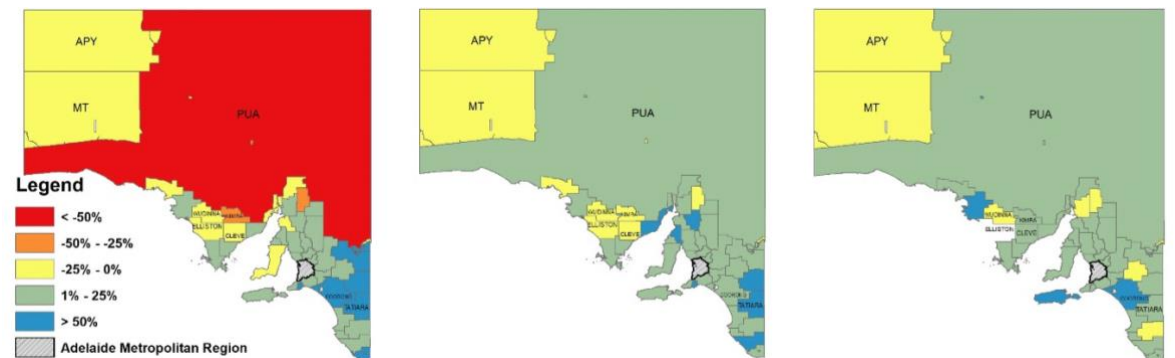

Figure 3: The percentage change in median price South Australia in 10 years, 3 years and 1 year respectively.

As a first step to describe and understand the velocity of the median price in these markets, the temporal change in median price for each level the ABS boundaries were clustered in to 6 main categories. This was done using a hierarchical clustering algorithm - Dynamic Time Warping (DTW) where similarity/distance between two time series is measured and they are grouped together based on this distance. We clustered all the boundary levels except GCCSA, into 5 distinct clusters. The number of clusters (5) was arrived through examining the resulting dendrogram of the analysis of all the boundary levels and describing them manually into clear, easily understandable labels. 5 clusters that emerged in common to all levels are Moderate to Very expensive where median prices started at moderate level and reached very high levels, Expensive to Very expensive, Moderate to Expensive, Low to Moderate, Moderate with No significant change, and Low with no significant change. These

clusters show the overall trend of median prices in these regional markets effortlessly. For example, the visualisation immediately shows that the 17 regional LGA which have experienced rapid growth in their median price in the last decade were all located in coastal NSW. Figure 3 shows the evolution of median price in these 5 clusters across all four boundary levels. The key thing to notice here is that because of the automated process used for clustering some artifacts do appear in the final outputs which can be counter-intuitive to a manual process. This issue can be solved in further research using a different algorithm or using an average of multiple clustering methods. Further explanation for this clustering methodology is provided at Appendix 3.
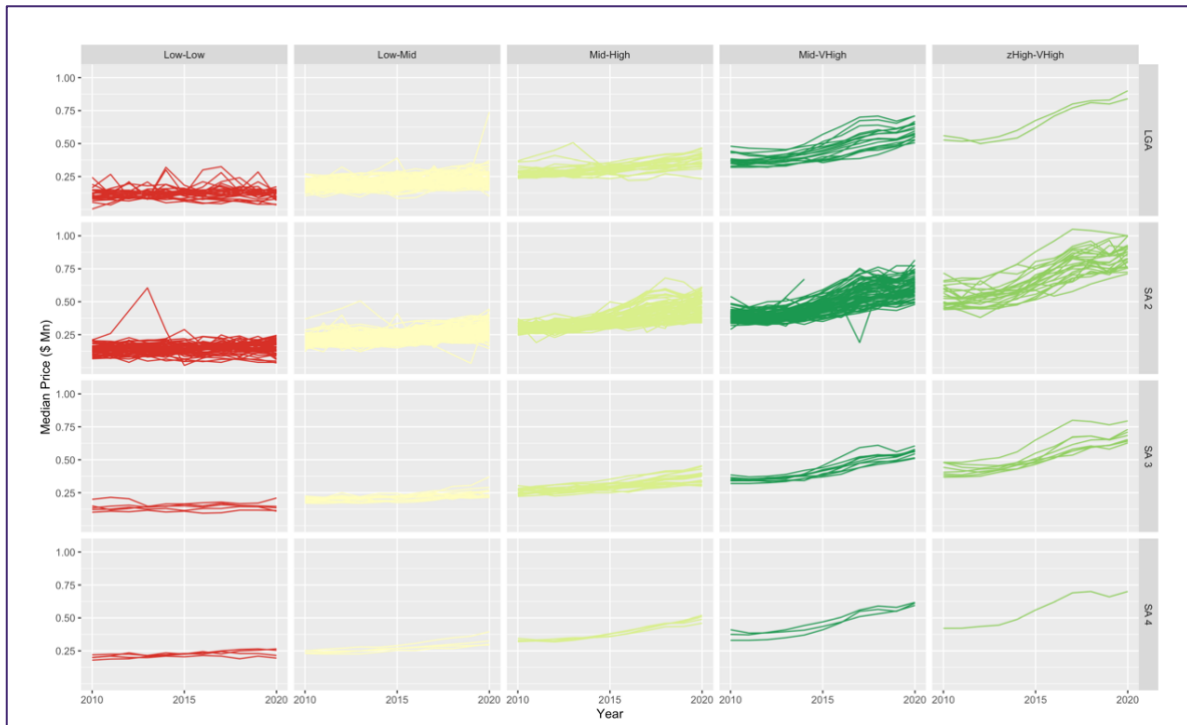

Figure 4: The trend of median price from 2010 – 2020 clustered into 5 categories across LGA, SA2, SA3, and SA4.

We also undertook exploratory analysis on the relationship between the price and other variables of the sales such as building age, area etc. This is particularly relevant for South Australia where data on the property being sold were available along with the sale price. As shown in Figure 5, there are significant relationships between the property features (explanatory variables) and the corresponding sale prices. It is important to note that most of these relationships are non-linear i.e., the price don't scale up consistently along with increase/decrease in property features. For example, from the first chart in Figure 5 we can see that building age affects the price of the property, but this relationship is not linear. New built houses are generally less expensive (per sqm) than house which are 15 to 20 years old while houses older than 25 years are less expensive as they get older. This trend changes around the 100 years mark where the houses move into historic category. At this stage, these old properties could be either very large, has historic value or in some of the best possible locations which could explain the prices growing exponentially with age.

While looking at categorical variable such as building conditions the relationship shows a possible sigmoid curve where changes in building quality at the lower end and higher end doesn't affect the price significantly but has an exponential effect in the middle. This inference is significantly constrained as very small number of categories defined for building quality. Among other categorical variables, we found that building style had a significant effect on the price where buildings with "Contemporary"/"Architectural" style are 20-25% more expensive per sqm than "Conventional" styles. These relationships show that modelling property prices based on their features is not a linear process. It requires much more complex models where the unique effect of each parameter on property needs to be considered independently while taking into consideration how they interact with each other. Models proposed by Lee et al. (2023), Gao et al. (2022), and Soltani et al. (2022) could be good starting points for developing such models in the future with a focus on regional areas.
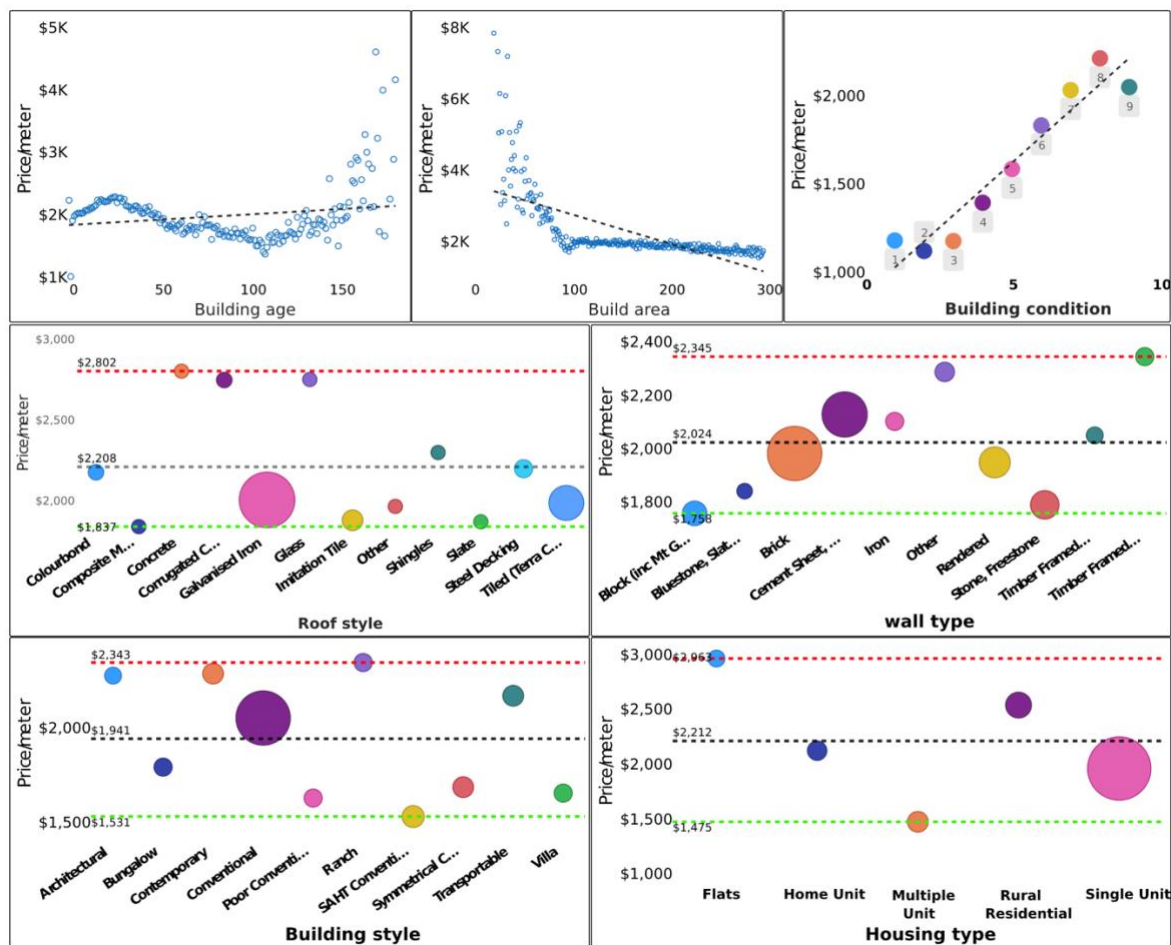

Figure 5: Relationship between property features and price per square metre for residential sales across South Australia.

Finally, the property sales data was augmented with information on the population in the corresponding boundaries by using the census data from Australian Bureau of Statistics. The data points chosen for creating such profile are population count, age profile, home ownership profile and income profile. These demographic profiles are simplified from the original ABS classifications to make them clearer and more understandable. Table 2 shows

the various fields that were combined to create the population profile for the regional areas. This gives us an overall snapshot of the type of population living in the regional area. For example, the LGA of Byron NSW has a younger population profile compared to the nearby Richmond valley but at the same time it has a much higher proportion of high-income population than Richmond valley.

Table 2: Data sets used from Census 2022 and their aggregation to a simpler profile.

Age	**Children**	Less than 20 years old
	Young	20 – 45 years old
	Middle	45 – 65 years old
	Old	Greater than 65 years old
Income	**Low**	Less than $800 per week
	Middle	$800 to $3000 per week
	High	Greater than $3000 per week
Home Ownership	Owned Outright, Mortgaged, Rented and Other	

VISUALISATION

The final step of the research was to communicate the outputs in a clear and effective manner to the stakeholders of the property industry. The outputs were visualised using a bespoke, interactive, map-based dashboard. The basic structure and architecture of the dashboard was developed for the Australian Property Market Explorer under the Australian Housing Data Analytics Platform project. The dashboard consists of various components at the backend, which were developed and deployed as micro-services. This provides the platform with clean development standard and potential scalability in the future. A simple schematic diagram of the architecture is shown in Figure 6.
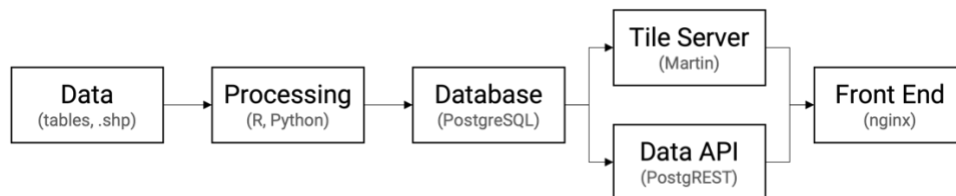

Figure 6: Simple Schematic diagram showing the architecture of the visualisation tool.

The front-end user interface of the dashboard has three major components - the full screen map, information panel on the right-hand side and control panel at the bottom. The map shows the study area – NSW and SA and plots a choropleth or thematic map of the boundaries chosen in the control panel as layers and can be panned and zoomed using the mouse. The control panel allows the user to change the data displayed on the layer – Clusters or the median price, change the year for the median price shown, reset the zoom level, select the boundary level, and search for an area in the chosen zoom level. When an area is selected, the associated information panel is displayed which shows detailed information on its property market. The information includes the volume, price and clustering result for the area, interactive charts on the historic volume and median price in the area and few interactive charts on the demographic profile of the area. Thus, users can start from a

general overview of the regional markets and dynamically able to select various areas of interest and investigate them in detail. A comprehensive user manual for the dashboard is provided in Appendix 4.
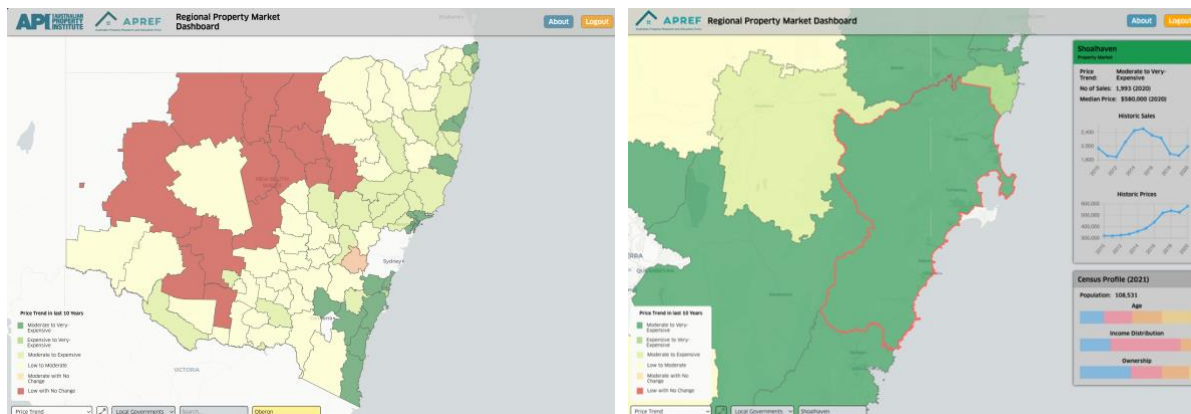

Figure 7: Screenshots showing the capabilities of the regional property market dashboard.

CONCLUSION

The primary output of the project - a dynamic, interactive, multi-level, geo-spatial visualisation which is available at https://dashboard.regionalpropertymarkets.com showing the spatio-temporal patterns of property markets in regional NSW and SA along with results of the analysis in the form of summary profiles such as the median price trend clusters, demographic profiles etc. across various statistical areas. For example, the visualisation immediately shows that the 17 regional LGA which have experienced rapid growth in their median price in the last decade were all located in coastal NSW. On a more detailed inspection, it can show how much of an outlier Kiama and Byron are in terms median price and their change compared to the rest of the state. The secondary output is an aggregated index published at https://agg.regionalpropertymarkets.com which shows the trends in price and volume of sales, clustering results and a brief census profile for each regional area at multiple boundary levels.

This study is important to the Australian property industry since there is a considerable lack of usable publicly available data and resources regarding regional property markets compared to the Capital Cities. Although few portals and datasets do exist, they are compiled at a state/local government level and distributed in such formats which require some level of expertise in data analysis for generating usable insights. This research is also important to specific aspects of the property industry such as valuation, where understanding the regional context of the property's location is one of the key steps in the process. Outputs from this research will enable property professionals to form a quick, baseline understanding of regional property markets in comparison to others before they go on to investigate them in detail at a suburb or property level. This research is also an important step in building pathways for translating research into an easily understandable format for the community of public and professionals to create maximum impact in practice. The research team hopes

the regional property market dashboard will be widely used and will act as an example showcasing the value in building and expanding such tools and resources at a national level.

Although the scope of this research does not delve into the prescriptive findings for the vstakeholders of the property industry in their decision-making process, it forms a base for such insights to be derived on top of it. In this regard, the output aggregated data generated from this research is crucial. This dataset will form a part of and distributed via the National Housing Data Exchange (AHDE) and the technology/platform that powers the dashboard will form a part of the Australian Property Market Explorer (APME) – initiatives from the Australian Housing Data Analytics Platform (AHDAP). These components will be leveraged with the broader resources available to the AHDAP to develop a more comprehensive, national level data infrastructure and ecosystem of tools built on top of it. When developed comprehensively, the outputs of the research outputs can be a key resource for multiple stakeholders in various ways. For example, it can help policy makers in understanding housing affordability and stress, financial institutions in understanding risk, market participants in making better transaction decisions, government in optimising planning, service provision and revenue generation etc. It is important to contrast the effort by this research to other efforts such as property portals such as domain, CoreLogic and research data portals such as AURIN. The commercial portals usually focus on the end users in terms of property sellers and buyers and tailor their offering for individual property level. The AURIN portal makes a similar dataset available but analysing and producing visualisations & insights from it requires significant time and resources. This project demonstrates that it is possible to build a system which is quick to update and easy to use.

This project and outputs demonstrate the immense value in utilising the data on sales collected by the revenue departments of state governments when combined with right aggregation, analysis, and visualisation. We recommend that datasets be collected across Australia for all states and a nationally harmonised dataset for regional areas be created. Such dataset with a standardised and regularly updated profiles on property market and demography will be invaluable for the Australian property industry. There are multiple avenues in which the research can be taken forward. From the data side, we could look at including more closely related property related datasets such as rental data from housing bonds that are lodged with state government, mortgage related datasets from the central bank etc. From the analysis side we could look in to further strengthening the demographic profiling by using more data points related to property market along with using a more advanced clustering algorithm such as k-means or db-scan. On the visualisation and distribution side, we need to look for ways to sustainably host the outputs for the property industry in the future along with a plan for updating the outputs at regular intervals. When implemented in a right way this can grow an ecosystem around these outputs which will create significant impact for the property industry and regional areas of Australia by promoting a data-driven discussion and decision making in the property market.

Appendices

APPENDIX 1: ADDRESS MATCHING ALGORITHM FOR GEO-CODING NSW DATA FROM GNAF.

The geocoding of data for NSW is done using a text-based address matching algorithm using data from GNAF – Geocoded national address file. The matching was done in six steps,

Step 1: Creating an address and geographic point table from GNAF.

Step 2: Cleaning the property use classification into simple categories and identify houses and units. Cleaning the land use zoning classes to simpler categories.

Step 3: Remove special characters from the address strings and perform a direct match.

Step 4: Remove property names before the house number and perform the match again.

Step 5: Use only the locality (suburb) and postcode to look for a match to give us an approximate location of the property.

This process resulted in a 96% match for all the records in the property sales database. Although more reliable geocoding processes exist, considering the size of the property sales database, ease of use of the text-based address matching and the high hit rate of the results we used the above address matching method.

The dictionaries used to clean the data and the relevant R programming code is given below.

PROPERTY USE CLASSIFICATION

Property Use	Class	Sub Cls	Property Use	Class	Sub Cls
Cropping	Agriculture	Other	Health	Public Utils	Other
Dairies	Agriculture	Other	Kindergartens	Public Utils	Other
Farmlet	Agriculture	Other	Leisure/Resort	Public Utils	Other
Grazing	Agriculture	Other	Open Space	Public Utils	Other
Horticulture	Agriculture	Other	Public Utilities	Public Utils	Other
Irrigation Watered	Agriculture	Other	Schools, Education	Public Utils	Other
Grazing & Cropping	Agriculture	Other	Regional Shopping Centre	Shops & Utils	Other
Oyster Farms	Agriculture	Other	Restaurants	Shops & Utils	Other
Specialist Production	Agriculture	Other	Shop And Dwelling	Shops & Utils	Other
Vineyards/Wineries	Agriculture	Other	Shops	Shops & Utils	Other
Board Houses, Hostels	Hotels	Other	Shops & Offices	Shops & Utils	Other
Hotels	Hotels	Other	Strata Shop	Shops & Utils	Other
Motels	Hotels	Other	Wholesale	Shops & Utils	Other
Timeshare	Hotels	Other	Car Yards	Transport	Other
Tourist Resorts	Hotels	Other	Caravan Parks	Transport	Other
Cottage	Residential	Houses	Transport	Transport	Other
Improved Rural Res.	Residential	Houses	Other Transp.	Transport	Other

Coal	Industrial	Other	Service Station, Garages	Transport	Other
Factory/Workshop	Industrial	Other	Strata Car Space	Transport	Other
Industrial Strata	Industrial	Other	Block Units/Inv. Flats	Residential	Units
Metallic	Industrial	Other	Rt. Villages/Units	Residential	Units
Non-Metallic	Industrial	Other	Single Res. Strata	Residential	Units
Timber/Forestry	Industrial	Other	Englobo Lands	Vacant	Other
Warehouse	Industrial	Other	Vacant (Rural, Non Comm)	Vacant	Other
Offices	Offices	Other	Vacant	Vacant	Other
Strata Office	Offices	Other	Vacant (Density Sites)	Vacant	Other
Churches, Public Assm.	Public Utils	Other	Vacant Land	Vacant	Other
Clubs	Public Utils	Other	Vacant (Single Res)	Vacant	Other

LAND USE ZONING CATEGORIES

Code	Zone	Sub Zone	Code	Zone	Sub Zone
A	Residential	Residential	B1	Business	Neighbourhood Centre
B	Business	Business	B2	Business	Local Centre
C	Sydney Commercial	Sydney Commercial	B3	Business	Commercial Core
D	Sustainable Mixed	Sustainable Mixed Use	B4	Business	Mixed Use
E	Employment	Employment	B5	Business	Business Development
I	Industrial	Industrial	B6	Business	Enterprise Corridor
M	Mixed Residential	Mixed Residential	B7	Business	Business Park
N	National Parks	National Parks	B8	Business	Metropolitan Centre
O	Open Space	Open Space	IN1	Industrial	General Industrial
P	Protection	Protection	IN2	Industrial	Light Industrial
R	Non-Urban	Non-Urban	IN3	Industrial	Heavy Industrial
S	Special Uses	Special Uses	IN4	Industrial	Working Waterfront
T	North Sydney Commer.	North Sydney Commer.	SP1	Special	Special Activities
U	Community Uses	Community Uses	SP2	Special	Infrastructure
V	Comprehensive Centre	Comprehensive Centre	SP3	Special	Tourist
W	Reserve Open Space	Reserve Open Space	RE1	Open Space	Public Recreation
X	Reserved Roads	Reserved Roads	RE2	Open Space	Private Recreation
Y	Reserved Special	Reserved Special Uses	E1	Non-Urban	National Parks & Reserves
Z	Undetermined/Village	Undetermined/Village	E2	Non-Urban	Environmental Conserv.
RU1	Non-Urban	Primary Production	E3	Non-Urban	Environmental Management
RU2	Non-Urban	Rural Landscape	E4	Non-Urban	Environmental Living
RU3	Non-Urban	Forestry	W1	Non-Urban	Natural Waterways
RU4	Non-Urban	Rural Small Holdings	W2	Open Space	Recreational Waterways
RU5	Residential	Village	W3	Industrial	Working Waterways
RU6	Non-Urban	Transition			
R1	Residential	General Residential			
R2	Residential	Low Density Res			
R3	Residential	Medium Density Res			
R4	Residential	High Density Res			
R5	Residential	Large Lot Res			

STREET TYPE ABBREVIATIONS

SHORT	LONG	SHORT	LONG	SHORT	LONG	SHORT	LONG	SHORT	LONG
Accs	Access	Pde	Parade	Cove	Cove	Stps	Steps	Gate	Gate

Ambl	Amble	Psge	Passage	Crss	Cross	Sbwy	Subway	Gra	Gra
Artl	Arterial	Piaz	Piazza	Cutt	Cutting	Thor	Thoroughfare	Hts	Hts
Ba	Banan	Pkt	Pocket	Dip	Dip	Tor	Tor	Hill	Hill
Bvd	Boulevard	Prom	Promenade	Drwy	Driveway	Turn	Turn	Key	Key
Brk	Break	Qys	Quays	End	End	Viad	Viaduct	Line	Line
Bypa	Bypass	Rest	Rest	Exp	Expressway	Walk	Walk	Loop	Loop
Ctr	Centre	Ring	Ring	Ftrk	Firetrack	Whrf	Wharf	Mews	Mews
Clt	Circlet	Rds	Roads	Ftwy	Footway	Alwy	Alleyway	Otlk	Otlk
Cl	Close	Row	Row	Frtg	Frontage	Arc	Arcade	Pass	Pass
Cps	Copse	Shun	Shunt	Gdns	Gardens	Ave	Avenue	Phwy	Phwy
Ctyd	Courtyard	Strs	Stairs	Glen	Glen	Bwlk	Boardwalk	Plza	Plza
Crst	Crest	Strp	Strip	Gr	Grove	Brae	Brae	Port	Port
Csac	Cul-de-sac	Tce	Terrace	Hwy	Highway	Brow	Brow	Qy	Qy
Devn	Deviation	Top	Top	Jnc	Junction	Caus	Causeway	Ramp	Ramp
Dr	Drive	Trl	Trail	Lnwy	Laneway	Cir	Circle	Rdge	Rdge
Elb	Elbow	Vale	Vale	Lkt	Lookout	Crcs	Circus	Rd	Rd
Esp	Esplanade	Vsta	Vista	Mndr	Meander	Con	Concourse	Rte	Rte
Fawy	Fairway	Way	Way	Nook	Nook	Ct	Court	Svwy	Svwy
Folw	Follow	Ally	Alley	Pkwy	Parkway	Cr	Crescent	Sq	Sq
Fwy	Freeway	App	Approach	Path	Path	Cres	Crescent	St	St
Gdn	Garden	Art	Artery	Pl	Place	Crsg	Crossing	Tarn	Tarn
Gld	Glade	Bend	Bend	Pnt	Point	Dale	Dale	Tlwy	Tlwy
Grn	Green	Brce	Brace	Qdrt	Quadrant	Dstr	Distributor	Trk	Trk
Hrd	Highroad	Bdwy	Broadway	Rmbl	Ramble	Edge	Edge	Upas	Upas
Intg	Interchange	Bywy	Byway	Rtt	Retreat	Ent	Entrance	View	View
Lane	Lane	Ch	Chase	Rise	Rise	Extn	Extension	Wkwy	Wkwy
Link	Link	Cct	Circuit	Rty	Rotary	Fitr	Firetrail	Wynd	Wynd
Mall	Mall	Cmmn	Common	Rue	Rue	Form	Formation		
Mwy	Motorway	Cnr	Corner	Spur	Spur	Gap	Gap		

ADDRESS MATCHING ALGORITHM (R)

```
#! /usr/bin/env Rscript

library(tidyverse)
library(lubridate)
library(fuzzyjoin)
library(magrittr)
library(gnaf.r)

#---------------------------------------------------- Setting up Global Variables
GNAF_FOLDER ← "Data/gnaf_202105/G-NAF/G-NAF MAY 2021/"
SALES_FILE ← "Data/valuer_general/202106/20210714 - Sales.txt"
ABBREVIATIONS_FILE ← "Data/valuer_general/street_type_abbreviations.csv"
SALE_CODES_FILE ← "Data/valuer_general/202106/sales_code_description.csv"
OUTPUT_FILE ← "Data/valuer_general/202106/202106_sales_gnaf_matched.csv"

#---------------------------------------------------- Building GNAF Lookup Table
setup(dir = GNAF_FOLDER, states = "nsw")
```

```
gnaf ← build_gnaf() ▷
  filter(!is.na(LATITUDE), !is.na(LONGITUDE), LATITUDE ≠ '', LONGITUDE ≠ '') ▷
  mutate(LATITUDE = as.numeric(LATITUDE), LONGITUDE = as.numeric(LONGITUDE),
         address = str_replace(address, ' NEW SOUTH WALES ', ' '))

gnaf_duplicates_removed ← gnaf ▷
  select(address, ADDRESS_DETAIL_PID, STREET_NAME, STREET_TYPE,
         LOCALITY_NAME, POSTCODE, LONGITUDE, LATITUDE, LEGAL_PARCEL_ID) ▷
  group_by(address) ▷
  sample_n(1)

#--------------------------------------------------------- Reading the sales data
sales ← read_delim(SALES_FILE, delim = ";", quote = "@")
street_type_abbreviations ← read_csv(ABBREVIATIONS_FILE) ▷
  mutate(ABBR = paste0(" ",ABBR,", ")) ▷
  mutate(FULL = paste0(" ",FULL,", ")) %$%
  set_names(x = FULL, nm = ABBR)
sale_codes ← read_csv(SALE_CODES_FILE)

sales_coded ← sales ▷ left_join(sale_codes)

#--------------------------------------------------------- Matching full address
stage_1 ← sales_coded ▷
  mutate(address = PROP_ADDR ▷
     str_replace_all(street_type_abbreviations) ▷
     str_replace(' / ', ' ') ▷
     str_replace_all(',', '') ▷
     str_replace(' NSW ', ' ')) ▷
  left_join(gnaf_duplicates_removed, by = "address")

#------------------------------------------------- Matching without property name
stage_2 ← stage_1 ▷
  filter(is.na(LATITUDE) | is.na(LONGITUDE) ) ▷
  select(-ADDRESS_DETAIL_PID, -STREET_NAME, -STREET_TYPE, -LOCALITY_NAME,
         -POSTCODE, -LONGITUDE, -LATITUDE, -LEGAL_PARCEL_ID) ▷
  mutate(address = PROP_ADDR ▷
           str_replace_all(street_type_abbreviations) ▷
           str_replace(' / ', ' ') ▷
           str_replace(' NSW ', ' ')
         ) ▷
  separate(address, c("a_01","a_02","a_03","a_04","a_05",
                      "a_06","a_07","a_08","a_09","a_10"),
           sep=", ", fill = "left")  ▷
  unite(address, c(a_09, a_10), sep = " ", na.rm = TRUE) ▷
  select(-starts_with("a_")) ▷
  left_join(gnaf_duplicates_removed, by = "address")

#------------------------------------------------- Matching with locality & postcode

stage_3 ← stage_2 ▷
  filter(is.na(LATITUDE) | is.na(LONGITUDE) ) ▷
  select(-LONGITUDE, -LATITUDE, -address) ▷
  mutate(address = PROP_ADDR ▷
           str_replace_all(street_type_abbreviations) ▷
           str_replace(' / ', ' ') ▷
```

```
                str_replace(' NSW ', ' ') ) ▷
  separate(address, c("a_01","a_02","a_03","a_04","a_05",
                      "a_06","a_07","a_08","a_09","a_10"),
           sep=", ", fill = "left") ▷
  mutate(address = a_10) ▷
  select(-starts_with("a_") ) ▷
  left_join(gnaf_duplicates_removed ▷
              unite(address, c(LOCALITY_NAME, POSTCODE),sep = " ", na.rm = TRUE) ▷
              group_by(address) ▷
              summarize(LATITUDE = mean(LATITUDE), LONGITUDE = mean(LONGITUDE) ) ) )


#---------------------------------------------Clean and combine all matches
sales_gnaf_matched ← rbind(stage_3,
    stage_2 ▷ filter(!is.na(LATITUDE) & !is.na(LONGITUDE)),
    stage_1 ▷ filter(!is.na(LATITUDE) & !is.na(LONGITUDE))) ▷
  select(
    gnaf_property_id = ADDRESS_DETAIL_PID,
    legal_parcel_id = LEGAL_PARCEL_ID,
    prop_id = PROP_ID,
    deal_number = DEAL_NO,
    property_type = PROP_TYPE,
    property_use = sale_code_class,
    property_use_subclass = sale_code_subclass,
    area = AREA,
    area_type = AREA_TYPE,
    contract_date = CONT_DATE,
    settlement_date = SETT_DATE,
    purchase_price = PURCH_PRICE,
    zone_code = ZONE_CODE,
    address = PROP_ADDR,
    gnaf_matched_address = address,
    street_name = STREET_NAME,
    suburb = LOCALITY_NAME,
    postcode = POSTCODE,
    council_name = COUNCIL_NAME,
    latitude = LATITUDE,
    longitude = LONGITUDE ) ▷
  mutate(
    contract_date = as_date(contract_date,format="%d/%m/%Y"),
    settlement_date = as_date(settlement_date,format="%d/%m/%Y") )

#---------------------------------------------------- Write the output file
sales_gnaf_matched ▷ write_csv(OUTPUT_FILE)
```

APPENDIX 2: DATA CLEANING AND FILTERING PROCESS

The data used for the analysis and clustering was cleaned extensively so that the results are not affected by the input data disproportionately. This improves the reliability and completeness of the data which is especially important for the South Australian dataset which has multiple fields. The pre-processing was done in three steps,

- **Step 1 –** Identify records which has incomplete data, are non-residential sales, are not linked to a geographic location and have invalid data in columns.

- **Step 2 –** Remove duplicates and outliers in Building Area, Land Area, and Sales Price

- **Step 3 –** Removing records with missing data and interpolating them if possible.

Removing outliers is crucial to make the outputs of the analyses using this dataset more stable and meaningful. The outliers are removed using the Inter-Quartile Range (IQR) method at the columns – Building Area, Price, and Land Area. Based on this IQR method, an outlier x can be detected if it satisfies one of the following conditions,

$$x < Q_1 - (1.5 \times IQR) \ \text{(or)} \ x < Q_3 + (1.5 \times IQR)$$

where Q_1 is the value of the 25th percentile, Q_3 is the value of the 75th percentile and $IQR = Q_3 - Q_1$

The missing data is filled with estimated/interpolated values wherever possible. The missing data is replaced with the dominant value where the data type is categorical or by the average of values when it is numerical data type.

Column name	Handling policy
Condition	Filled with average value
Number of Rooms	Filled with average value
Ensuite	Filled with dominant value
Number of Storeys	Filled with dominant value
Property Roof type	Filled with dominant value
Property Style	Filled with dominant value
Property wall type	Filled with dominant value
Purchaser Type	Filled with dominant value
Vendor Type	Filled with dominant value

SA

Non-residential property	954,445	73738
Incomplete cases	1,128,759	14603
Duplicated data	4,972	2924
Not located data	41	338
Outlier sale price	277,209	39144
Outlier build area	622,906	3507

NSW Invalid Data 2449333 Valid Data 2301513

Invalid Data 106444 Valid Data 70797

Summary of the results from the data cleaning operations for both and NSW

DATA CLEANING OPERATIONS (SQL)

The SQL code below shows the data cleaning operations done on the SA data in detail. A similar cleaning was carried out on NSW data as well (not included to avoid duplication)

```sql
-- Identify Incomplete Cases, cases with non-residential, unlocated and invalid year cases
------------------------------------------------------------------------------------------
ALTER TABLE public."SA_JointData" ADD COLUMN incompleteCases INT;

UPDATE public."SA_JointData" SET incompleteCases = CASE WHEN saleprice = '$0' AND buildarea IS
NULL AND yearbuilt IS  NULL THEN 1 ELSE 0 END;

ALTER TABLE public."SA_JointData" ADD COLUMN notResidential INT;

UPDATE public."SA_JointData"  AS dat SET notresidential = 1 FROM   public."SA_LandUseCodes" AS
codes WHERE dat.landusecode  = codes.landusecode AND "LandClass" <> 'RESIDENTIAL';

UPDATE  public."SA_JointData"  AS dat SET notresidential = 0 FROM  public."SA_LandUseCodes"
AS codes WHERE dat.landusecode  = codes.landusecode AND "LandClass" = 'RESIDENTIAL';

ALTER TABLE public."SA_JointData" ADD COLUMN unlocatedData INT;

UPDATE public."SA_JointData" SET unlocatedData = CASE WHEN geom IS NULL THEN 1 ELSE 0 END;

ALTER TABLE public."SA_JointData" ADD COLUMN Invalidyearbuit INT;

UPDATE public."SA_JointData" SET Invalidyearbuit = CASE WHEN yearbuilt IS NULL THEN 1 ELSE 0
END;

SELECT Invalidyearbuit, COUNT(Invalidyearbuit) FROM public."SA_JointData" GROUP BY
Invalidyearbuit;

-- Step 2: Remove duplicates and outliers in Building Area, Land Area, and Sales Price
------------------------------------------------------------------------------------------
ALTER TABLE public."SA_JointData" ADD COLUMN OTL_IQR_buildArea INT;

CREATE OR REPLACE VIEW notnullbuildarea AS SELECT * FROM public."SA_JointData" WHERE buildarea
IS NOT NULL;

WITH orderedList AS ( SELECT buildarea, ROW_NUMBER() OVER (ORDER BY buildarea) AS row_n FROM
notnullbuildarea), iqr AS ( SELECT buildarea, ( SELECT buildarea AS quartile_break FROM
orderedList WHERE row_n = FLOOR((SELECT COUNT(*) FROM notnullbuildarea)*0.75)) AS q_three, (
SELECT buildarea AS quartile_break FROM orderedList WHERE row_n = FLOOR((SELECT COUNT(*) FROM
notnullbuildarea)*0.25)) AS q_one, 1.5 * (( SELECT buildarea AS quartile_break FROM
orderedList WHERE row_n = FLOOR((SELECT COUNT(*) FROM notnullbuildarea)*0.75)) - ( SELECT
buildarea AS quartile_break FROM orderedList WHERE row_n = FLOOR((SELECT COUNT(*) FROM
notnullbuildarea)*0.25))) AS outlier_range FROM orderedList) UPDATE public."SA_JointData" SET
OTL_IQR_buildArea = CASE WHEN buildarea >= ((SELECT MAX(q_three) FROM iqr) + (SELECT
MAX(outlier_range) FROM iqr)) OR buildarea <= ((SELECT MAX(q_one) FROM iqr) - (SELECT
MAX(outlier_range) FROM iqr)) THEN 1 ELSE 0 END;

UPDATE public."SA_JointData" SET OTL_IQR_buildArea = CASE WHEN buildarea IS NULL THEN 2 ELSE
OTL_IQR_buildArea END;

SELECT OTL_IQR_buildArea, COUNT(OTL_IQR_buildArea) FROM public."SA_JointData" GROUP BY
OTL_IQR_buildArea;

ALTER TABLE public."SA_JointData" ADD COLUMN OTL_IQR_landarea INT;

WITH orderedList AS ( SELECT landarea, ROW_NUMBER() OVER (ORDER BY landarea) AS row_n FROM
public."SA_JointData"), iqr AS ( SELECT landarea, ( SELECT landarea AS quartile_break FROM
orderedList WHERE row_n = FLOOR((SELECT COUNT(*) FROM public."SA_JointData")*0.75)) AS
q_three, ( SELECT landarea AS quartile_break FROM orderedList WHERE row_n = FLOOR((SELECT
COUNT(*) FROM public."SA_JointData")*0.25)) AS q_one, 1.5 * (( SELECT landarea AS
quartile_break FROM orderedList WHERE row_n = FLOOR((SELECT COUNT(*) FROM
public."SA_JointData")*0.75)) - ( SELECT landarea AS quartile_break FROM orderedList WHERE
row_n = FLOOR((SELECT COUNT(*) FROM public."SA_JointData")*0.25))) AS outlier_range FROM
orderedList);
```

```sql
UPDATE public."SA_JointData" SET OTL_IQR_landarea = CASE WHEN landarea >= ((SELECT
MAX(q_three) FROM iqr) + (SELECT MAX(outlier_range) FROM iqr)) OR landarea <= ((SELECT
MAX(q_one) FROM iqr) - (SELECT MAX(outlier_range) FROM iqr)) THEN 1 ELSE 0 END

UPDATE    public."SA_JointData" SET OTL_IQR_landarea = CASE WHEN landarea IS NULL THEN 2 ELSE
OTL_IQR_landarea END;

SELECT OTL_IQR_landarea, COUNT(OTL_IQR_landarea) FROM public."SA_JointData" GROUP BY
OTL_IQR_landareaalter TABLE public."SA_JointData" ADD COLUMN OTL_IQR_SalePrice INT;

WITH orderedList AS ( SELECT saleprice, ROW_NUMBER() OVER (ORDER BY saleprice) AS row_n FROM
public."SA_JointData"), iqr AS ( SELECT saleprice, ( SELECT saleprice AS quartile_break FROM
orderedList WHERE row_n = FLOOR((SELECT COUNT(*) FROM public."SA_JointData")*0.75)) AS
q_three, ( SELECT saleprice AS quartile_break FROM orderedList WHERE row_n = FLOOR((SELECT
COUNT(*) FROM public."SA_JointData")*0.25)) AS q_one, 1.5 * (( SELECT saleprice AS
quartile_break FROM orderedList WHERE row_n = FLOOR((SELECT COUNT(*) FROM
public."SA_JointData")*0.75)) - ( SELECT saleprice AS quartile_break FROM orderedList WHERE
row_n = FLOOR((SELECT COUNT(*) FROM public."SA_JointData")*0.25))) AS outlier_range FROM
orderedList) UPDATE public."SA_JointData" SET OTL_IQR_SalePrice = CASE WHEN saleprice >=
((SELECT MAX(q_three) FROM iqr) + (SELECT MAX(outlier_range) FROM iqr)) OR saleprice <=
((SELECT MAX(q_one) FROM iqr) - (SELECT MAX(outlier_range) FROM iqr)) THEN 1 ELSE 0 END;

UPDATE public."SA_JointData" SET OTL_IQR_SalePrice  = CASE WHEN saleprice :: NUMERIC < 10000
THEN 1 ELSE OTL_IQR_SalePrice END;

SELECT OTL_IQR_SalePrice, COUNT(OTL_IQR_SalePrice) FROM public."SA_JointData" GROUP BY
OTL_IQR_SalePricealter TABLE "SA_JointData" ADD COLUMN duplicateddata INT;

UPDATE public."SA_JointData" tt SET duplicateddata = 1 FROM  (SELECT id, valuationnumber  FROM
( SELECT id, valuationnumber, saledate, saleprice, COUNT(*) OVER (PARTITION BY
valuationnumber,saledate,saleprice) AS counta FROM public."SA_JointData" WHERE unlocateddata =
0 AND incompletecases = 0 AND notresidential = 0 AND otl_iqr_saleprice = 0 AND
otl_iqr_buildarea = 0 AND OTL_IQR_landarea = 0 AND invalidyearbuit = 0 ORDER BY counta DESC)
AS tableWithCount WHERE tableWithCount.counta > 1) dd WHERE tt.id = dd.id;

UPDATE public."SA_JointData" tt SET duplicateddata = 0 FROM  (SELECT valuationnumber, MIN(id)
AS id FROM ( SELECT id, valuationnumber, saledate, saleprice, COUNT(*) OVER (PARTITION BY
valuationnumber,saledate,saleprice) AS counta FROM public."SA_JointData" WHERE unlocateddata =
0 AND incompletecases = 0 AND notresidential = 0 AND otl_iqr_saleprice = 0 AND
otl_iqr_buildarea = 0 AND OTL_IQR_landarea = 0 AND invalidyearbuit = 0 ORDER BY
valuationnumber DESC) AS tableWithCount WHERE tableWithCount.counta > 1 GROUP BY
distinct(valuationnumber)) ss WHERE tt.id = ss.id;

UPDATE public."SA_JointData" tt SET duplicateddata = CASE WHEN duplicateddata IS NULL THEN 0
ELSE duplicateddata END;

UPDATE public."SA_JointData" tt SET duplicateddata = 1 FROM  (SELECT id FROM ( SELECT id,
valuationnumber, saledate,saleprice, COUNT(*) OVER (PARTITION BY valuationnumber,saledate) AS
counta FROM public."SA_JointData" WHERE unlocateddata = 0 AND incompletecases = 0 AND
notresidential = 0 AND otl_iqr_saleprice = 0 AND otl_iqr_buildarea = 0 AND OTL_IQR_landarea =
0 AND invalidyearbuit = 0 AND duplicateddata = 0 ORDER BY counta DESC) AS tableWithCount WHERE
tableWithCount.counta > 1 ORDER BY valuationnumber ) AS ss WHERE tt.id = ss.id;

SELECT duplicateddata, COUNT(*) FROM public."SA_JointData" GROUP BY duplicateddata;

-- Step 3: Removing records with missing data and interpolating them if possible
------------------------------------------------------------------------------------------
UPDATE public."SA_JointData" SET condition = ( SELECT ROUND(AVG(condition))  FROM
public."SA_JointData" WHERE condition IS NOT NULL AND unlocateddata = 0 AND incompletecases =
0 AND notresidential = 0 AND otl_iqr_saleprice = 0 AND otl_iqr_buildarea = 0 AND
OTL_IQR_landarea = 0 AND invalidyearbuit = 0and duplicateddata = 0) WHERE condition IS NULL
AND unlocateddata = 0 AND incompletecases = 0 AND notresidential = 0 AND otl_iqr_saleprice = 0
```

```sql
AND otl_iqr_buildarea = 0 AND OTL_IQR_landarea = 0 AND invalidyearbuit = 0and duplicateddata =
0;

SELECT vendortype,COUNT(*) FROM public."SA_JointData" WHERE unlocateddata = 0 AND
incompletecases = 0 AND notresidential = 0 AND otl_iqr_saleprice = 0 AND otl_iqr_buildarea = 0
AND OTL_IQR_landarea = 0 AND invalidyearbuit = 0 AND duplicateddata = 0 GROUP BY vendortype
ORDER BY COUNT(vendortype) DESC;

UPDATE public."SA_JointData" SET vendortype = ( SELECT rf FROM ( SELECT   vendortype AS rf,
COUNT( vendortype) AS co_rf FROM public."SA_JointData" WHERE  vendortype <> '0' AND
unlocateddata = 0 AND incompletecases = 0 AND notresidential = 0 AND otl_iqr_saleprice = 0 AND
otl_iqr_buildarea = 0 AND OTL_IQR_landarea = 0 AND invalidyearbuit = 0 AND duplicateddata = 0
GROUP BY (vendortype) ORDER BY co_rf DESC) AS aaa LIMIT 1) WHERE  vendortype = '0' AND
unlocateddata = 0 AND incompletecases = 0 AND notresidential = 0 AND otl_iqr_saleprice = 0 AND
otl_iqr_buildarea = 0 AND OTL_IQR_landarea = 0 AND invalidyearbuit = 0 AND duplicateddata =
0UPDATE public."SA_JointData" SET ensuite = 0 WHERE  ensuite IS NULL AND unlocateddata = 0 AND
incompletecases = 0 AND notresidential = 0 AND otl_iqr_saleprice = 0 AND otl_iqr_buildarea = 0
AND OTL_IQR_landarea = 0 AND invalidyearbuit = 0and duplicateddata = 0;

UPDATE public."SA_JointData" SET purchasertype = ( SELECT rf FROM ( SELECT  purchasertype AS
rf, COUNT(purchasertype) AS co_rf FROM public."SA_JointData" WHERE purchasertype <> '0' AND
unlocateddata = 0 AND incompletecases = 0 AND notresidential = 0 AND otl_iqr_saleprice = 0 AND
otl_iqr_buildarea = 0 AND OTL_IQR_landarea = 0 AND invalidyearbuit = 0 AND duplicateddata = 0
GROUP BY (purchasertype) ORDER BY co_rf DESC) AS aaa LIMIT 1) WHERE purchasertype = '0' AND
unlocateddata = 0 AND incompletecases = 0 AND notresidential = 0 AND otl_iqr_saleprice = 0 AND
otl_iqr_buildarea = 0 AND OTL_IQR_landarea = 0 AND invalidyearbuit = 0 AND duplicateddata =
0UPDATE public."SA_JointData" SET roof = ( SELECT rf FROM ( SELECT  roof AS rf, COUNT(roof) AS
co_rf FROM public."SA_JointData" WHERE roof <> ' ' AND unlocateddata = 0 AND incompletecases =
0 AND notresidential = 0 AND otl_iqr_saleprice = 0 AND otl_iqr_buildarea = 0 AND
OTL_IQR_landarea = 0 AND invalidyearbuit = 0 AND duplicateddata = 0 GROUP BY (roof) ORDER BY
co_rf DESC) AS aaa LIMIT 1) WHERE roof = ' ' AND unlocateddata = 0 AND incompletecases = 0 AND
notresidential = 0 AND otl_iqr_saleprice = 0 AND otl_iqr_buildarea = 0 AND OTL_IQR_landarea =
0 AND invalidyearbuit = 0 AND duplicateddata = 0;

UPDATE public."SA_JointData" SET rooms = ( SELECT ROUND(AVG(rooms))  FROM
public."SA_JointData" WHERE rooms IS NOT NULL AND unlocateddata = 0 AND incompletecases = 0
AND notresidential = 0 AND otl_iqr_saleprice = 0 AND otl_iqr_buildarea = 0 AND
OTL_IQR_landarea = 0 AND invalidyearbuit = 0 AND duplicateddata = 0 AND duplicateddata = 0)
WHERE rooms IS NULL AND unlocateddata = 0 AND incompletecases = 0 AND notresidential = 0 AND
otl_iqr_saleprice = 0 AND otl_iqr_buildarea = 0 AND OTL_IQR_landarea = 0 AND invalidyearbuit =
0 AND duplicateddata = 0 UPDATE public."SA_JointData" SET storeys = 1 WHERE  storeys IS NULL
AND unlocateddata = 0 AND incompletecases = 0 AND notresidential = 0 AND otl_iqr_saleprice = 0
AND otl_iqr_buildarea = 0 AND OTL_IQR_landarea = 0 AND invalidyearbuit = 0 AND duplicateddata
= 0;

UPDATE public."SA_JointData" SET style = ( SELECT rf FROM ( SELECT   style AS rf, COUNT(
style) AS co_rf FROM public."SA_JointData" WHERE style <> ' ' AND unlocateddata = 0 AND
incompletecases = 0 AND notresidential = 0 AND otl_iqr_saleprice = 0 AND otl_iqr_buildarea = 0
AND OTL_IQR_landarea = 0 AND invalidyearbuit = 0 AND duplicateddata = 0 GROUP BY ( style)
ORDER BY co_rf DESC) AS aaa LIMIT 1) WHERE  style = ' ' AND unlocateddata = 0 AND
incompletecases = 0 AND notresidential = 0 AND otl_iqr_saleprice = 0 AND otl_iqr_buildarea = 0
AND OTL_IQR_landarea = 0 AND invalidyearbuit = 0 AND duplicateddata = 0UPDATE
public."SA_JointData" SET wall = ( SELECT rf FROM ( SELECT   wall AS rf, COUNT( wall) AS co_rf
FROM public."SA_JointData" WHERE wall <> ' ' AND unlocateddata = 0 AND incompletecases = 0 AND
notresidential = 0 AND otl_iqr_saleprice = 0 AND otl_iqr_buildarea = 0 AND OTL_IQR_landarea =
0 AND invalidyearbuit = 0 AND duplicateddata = 0 GROUP BY ( wall) ORDER BY co_rf DESC) AS aaa
LIMIT 1) WHERE  wall = ' ' AND unlocateddata = 0 AND incompletecases = 0 AND notresidential =
0 AND otl_iqr_saleprice = 0 AND otl_iqr_buildarea = 0 AND OTL_IQR_landarea = 0 AND
invalidyearbuit = 0 AND duplicateddata = 0;
```

APPENDIX 3: AGGREGATION AND CLUSTERING TEMPORAL PRICE TREND

The aggregation and clustering of the data has been done in 5 steps.

Step 1: Download ABS data for age, income, and home ownership attributes. Recode the age and income attributes to a simpler classification as shown in Table 2.

Step 2: Establish a one-to-many relationship between sales data to their corresponding ABS boundaries using a spatial join.

Step 3: Create aggregate for median price and volume for each ABS boundary for each year from 2010 to 2020 using the relevant sales data points.

Step 4: Create cluster numbers for each ABS boundary by running the DTW algorithm on the 10-point median price trends.

Step 5: Manually inspect the clusters for each level of boundary and label them using an appropriate human understandable name.

The R programming code used for doing the aggregation and clustering is given below.

AGGREGATION AND CLUSTERING (R)

```r
library(tidyverse)
library(sf)
library(ggplot2)

sales ← read_csv("sales.csv") ▷ st_as_sf(coords = c("lon", "lat"), na.fail = FALSE, crs = 4326)

# Creating a combined ABS Census profile for both SA and NSW from the geopackages.

state ← rbind(
  st_read("census_data/G01_NSW_GDA2020.gpkg","G01_STE_2021_NSW") ▷
    mutate(children = Age_0_4_yr_P + Age_5_14_yr_P + Age_15_19_yr_P,
           young = Age_20_24_yr_P + Age_25_34_yr_P + Age_35_44_yr_P,
           middle = Age_45_54_yr_P + Age_55_64_yr_P,
           senior = Age_65_74_yr_P + Age_75_84_yr_P + Age_85ov_P) ▷
    select(code = STE_CODE_2021,
           name = STE_NAME_2021,
           population = Tot_P_P,
           male = Tot_P_M,
           female = Tot_P_F,
           children, young, middle, senior ) ▷
    left_join(st_read("census_data/G37_NSW_GDA2020.gpkg","G37_STE_2021_NSW") ▷
      st_set_geometry(NULL) ▷
      select(code = STE_CODE_2021,
             owned_outright = "O_OR_Total",
             mortgaged = "O_MTG_Total",
             rented = "R_Tot_Total",
             other = "Oth_ten_type_Total",
             not_stated ="Ten_type_NS_Total",
             total = "Total_Total"),
    by="code") ▷
    left_join(st_read("census_data/G57A_NSW_GDA2020.gpkg","G57A_STE_2021_NSW") ▷
      st_set_geometry(NULL) ▷
      mutate(low_income = NegNil_Inc_Tot + I_1_149_Tot + I_150_299_Tot + I_300_399_Tot +
                          I_400_499_Tot + I_500_649_Tot + I_650_799_Tot,
```

```r
            middle_income = I_800_999_Tot + I_1000_1249_Tot + I_1250_1499_Tot +
                            I_1500_1749_Tot + I_1750_1999_Tot + I_2000_2499_Tot +
                            I_2500_2999_Tot,
            high_income = I_3000_3499_Tot + I_3500_3999_Tot) ▷
      select(code = STE_CODE_2021, high_income, middle_income, low_income),
    by="code") ▷
    left_join(st_read("census_data/G57B_NSW_GDA2020.gpkg","G57B_STE_2021_NSW") ▷
      st_set_geometry(NULL) ▷
      mutate(high_income2 = I_4000more_Tot) ▷
      select(code = STE_CODE_2021, high_income2),
    by="code") ▷
    mutate(high_income = high_income  +  high_income2) ▷
    select(-high_income2),
  st_read("census_data/G01_SA_GDA2020.gpkg","G01_STE_2021_SA") ▷
    mutate(children = Age_0_4_yr_P + Age_5_14_yr_P + Age_15_19_yr_P,
           young = Age_20_24_yr_P + Age_25_34_yr_P + Age_35_44_yr_P,
           middle = Age_45_54_yr_P + Age_55_64_yr_P,
           senior = Age_65_74_yr_P + Age_75_84_yr_P + Age_85ov_P) ▷
    select(code = STE_CODE_2021,
           name = STE_NAME_2021,
           population = Tot_P_P,
           male = Tot_P_M,
           female = Tot_P_F,
           children, young, middle, senior ) ▷
    left_join(st_read("census_data/G37_SA_GDA2020.gpkg","G37_STE_2021_SA") ▷
      st_set_geometry(NULL) ▷
      select(code = STE_CODE_2021,
             owned_outright = "O_OR_Total",
             mortgaged = "O_MTG_Total",
             rented = "R_Tot_Total",
             other = "Oth_ten_type_Total",
             not_stated ="Ten_type_NS_Total",
             total = "Total_Total"),
    by="code") ▷
    left_join(st_read("census_data/G57A_SA_GDA2020.gpkg","G57A_STE_2021_SA") ▷
      st_set_geometry(NULL) ▷
      mutate(low_income = NegNil_Inc_Tot + I_1_149_Tot + I_150_299_Tot + I_300_399_Tot +
I_400_499_Tot + I_500_649_Tot + I_650_799_Tot,
             middle_income = I_800_999_Tot + I_1000_1249_Tot + I_1250_1499_Tot +
I_1500_1749_Tot + I_1750_1999_Tot + I_2000_2499_Tot + I_2500_2999_Tot,
             high_income = I_3000_3499_Tot + I_3500_3999_Tot) ▷
      select(code = STE_CODE_2021, high_income, middle_income, low_income),
    by="code") ▷
    left_join(st_read("census_data/G57B_SA_GDA2020.gpkg","G57B_STE_2021_SA") ▷
      st_set_geometry(NULL) ▷
      mutate(high_income2 = I_4000more_Tot) ▷
      select(code = STE_CODE_2021, high_income2),
    by="code") ▷
    mutate(high_income = high_income  +  high_income2) ▷
    select(-high_income2)
) ▷ filter(!st_is_empty(geom))

################################################################################
# The above operations are repeated for GCCSA, SA4, SA3, SA2 and LGA
################################################################################


# Aggregating and sales volume and median price for all 5 levels

sales_joined_uf ← sales ▷ st_transform(crs="epsg:7844") ▷
  st_join(state ▷ select(state = code)) ▷
  st_join(lga ▷ select(lga = code)) ▷
  st_join(gccsa ▷ select(gccsa = code)) ▷
```

```r
  st_join(sa4 ▷ select(sa4 = code)) ▷
  st_join(sa3 ▷ select(sa3 = code)) ▷
  st_join(sa2 ▷ select(sa2 = code))

sales_joined ← sales_joined_uf ▷
  filter(!(price══0 | is.na(gccsa) | gccsa %in% c("1GSYD","4GADE") | st_is_empty(geometry)))
▷ filter(date < '2021-01-01')

state_summary ← sales_joined ▷ st_set_geometry(NULL) ▷
  mutate(year = format(date,"%Y")) ▷
  group_by(state,year) ▷ summarise(price = median(price), sales = n()) ▷
  pivot_wider(names_from = "year", values_from = c("price","sales"))

lga_summary ← sales_joined ▷ st_set_geometry(NULL) ▷
  mutate(year = format(date,"%Y")) ▷
  group_by(lga,year) ▷ summarise(price = median(price), sales = n()) ▷
  pivot_wider(names_from = "year", values_from = c("price","sales"))

gccsa_summary ← sales_joined ▷ st_set_geometry(NULL) ▷
  mutate(year = format(date,"%Y")) ▷
  group_by(gccsa,year) ▷ summarise(price = median(price), sales = n()) ▷
  pivot_wider(names_from = "year", values_from = c("price","sales"))

sa4_summary ← sales_joined ▷ st_set_geometry(NULL) ▷
  mutate(year = format(date,"%Y")) ▷
  group_by(sa4,year) ▷ summarise(price = median(price), sales = n()) ▷
  pivot_wider(names_from = "year", values_from = c("price","sales"))

sa3_summary ← sales_joined ▷ st_set_geometry(NULL) ▷
  mutate(year = format(date,"%Y")) ▷
  group_by(sa3,year) ▷ summarise(price = median(price), sales = n()) ▷
  pivot_wider(names_from = "year", values_from = c("price","sales"))

sa2_summary ← sales_joined ▷ st_set_geometry(NULL) ▷
  mutate(year = format(date,"%Y")) ▷
  group_by(sa2,year) ▷ summarise(price = median(price), sales = n()) ▷
  pivot_wider(names_from = "year", values_from = c("price","sales"))


state_final ← left_join(state_summary,state,by = c("state" = "code"))
lga_final ← left_join(lga_summary,lga,by = c("lga" = "code"))
gccsa_final ← left_join(gccsa_summary,gccsa,by = c("gccsa" = "code" ))
sa4_final ← left_join(sa4_summary,sa4,by = c("sa4" = "code"))
sa3_final ← left_join(sa3_summary,sa3,by = c("sa3" = "code"))
sa2_final ← left_join(sa2_summary,sa2,by = c("sa2" = "code"))

# Clustering temporal trends using DTW library

library(dtw)

cluster_label ← function(table, code_column, n_cluster) {
  table_cleaned ← table ▷
    select(all_of(code_column), starts_with("price_")) ▷
    pivot_longer(starts_with("price_"), values_drop_na = FALSE) ▷
    mutate(value = ifelse(value > 1500000, NA, value)) ▷
    group_by_at(code_column) ▷
    fill(value,.direction = "downup") ▷
    pivot_wider(all_of(code_column))
  price_matrix ← table_cleaned ▷ ungroup() ▷ select(starts_with("price")) ▷
as_data_frame()
  distance_matrix ← dist(price_matrix, method = "DTW")
  clusters ← hclust(distance_matrix, method = "complete")
  labels ← cutree(clusters, k = n_cluster)
```

```
  return(table ▷ ungroup() ▷ mutate(cluster = labels))
}

lga_clusters ← cluster_label(lga_final, "lga", 5)
sa4_clusters ← cluster_label(sa4_final, "sa4", 5)
sa3_clusters ← cluster_label(sa3_final, "sa3", 5)
sa2_clusters ← cluster_label(sa2_final, "sa2", 5)

# Plotting the results and manually labelling the clusters.

sa4_clusters ▷ select(sa4,cluster,starts_with("price_")) ▷
pivot_longer(starts_with("price_")) ▷ mutate(name = str_replace(name,"price_","")) ▷
  filter(value < 2000000) ▷ ggplot() + geom_line(aes(name,value,group=sa4)) + facet_grid(. ~
cluster)

sa3_clusters ▷ select(sa3,cluster,starts_with("price_")) ▷
pivot_longer(starts_with("price_")) ▷ mutate(name = str_replace(name,"price_","")) ▷
filter(value < 2000000) ▷ ggplot() + geom_line(aes(name,value,group=sa3)) + facet_grid(. ~
cluster)

sa2_clusters ▷ select(sa2,cluster,starts_with("price_")) ▷
pivot_longer(starts_with("price_")) ▷ mutate(name = str_replace(name,"price_","")) ▷
filter(value < 2000000) ▷ ggplot() + geom_line(aes(name,value,group=sa2)) + facet_grid(. ~
cluster)

lga_clusters ▷ select(lga,cluster,starts_with("price_")) ▷
pivot_longer(starts_with("price_")) ▷ mutate(name = str_replace(name,"price_","")) ▷
filter(value < 2000000) ▷ ggplot() + geom_line(aes(name,value,group=lga)) + facet_grid(. ~
cluster)

state_final$cluster ← c("Mid-High","Low-Mid")
gccsa_final$cluster ← c("Mid-High","Low-Mid")
sa4_final$cluster ← recode(sa4_clusters$cluster, `1` = "Mid-High", `2` = "Low-Mid", `3` =
"Low-Low", `4` = "zHigh-VHigh", `5` = "Mid-VHigh")
sa3_final$cluster ← recode(sa3_clusters$cluster, `4` = "Mid-High", `2` = "Low-Mid", `5` =
"Low-Low", `1` = "zHigh-VHigh", `3` = "Mid-VHigh")
sa2_final$cluster ← recode(sa2_clusters$cluster, `1` = "Mid-High", `5` = "Low-Mid", `4` =
"Low-Low", `3` = "zHigh-VHigh", `2` = "Mid-VHigh")
lga_final$cluster ← recode(lga_clusters$cluster, `1` = "Mid-High", `3` = "Low-Mid", `4` =
"Low-Low", `5` = "zHigh-VHigh", `2` = "Mid-VHigh")

# Exporting the results to Geo-Packages.

state_final ▷ st_write("state.gpkg")
lga_final ▷ st_write("lga.gpkg")
gccsa_final ▷ st_write("gccsa.gpkg")
sa4_final ▷ st_write("sa4.gpkg")
sa3_final ▷ st_write("sa3.gpkg")
sa2_final ▷ st_write("sa2.gpkg")
```

APPENDIX 4 - USER MANUAL FOR THE REGIONAL PROPERTY MARKET DASHBOARD

The Regional Property Market Dashboard is a bespoke, interactive, map-based dashboard which visualises the data and analysis on the regional property markets across New South Wales and South Australia. The dashboard was developed as part of the Australian Property Research and Education Fund Grant (APREF) from Australian Property Institute (API). The basic structure and architecture of the dashboard was developed for the Australian Property Market Explorer under the Australian Housing Data Analytics Platform project. This document details all the necessary information required for the users to access and utilise the tool.

ACCESS

The dashboard is available at https://dashboard.regionalpropertymarkets.com and requires a username and password to access. Access is managed through Australian Property Institute and requested through https://www.api.org.au/about/get-in-touch. A typical login screen should like below,
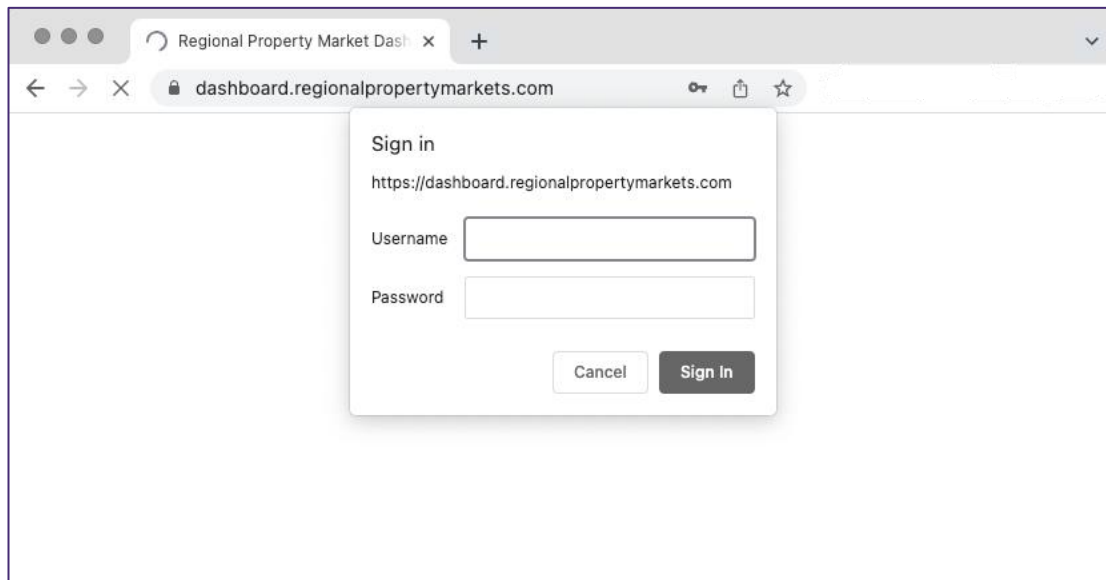
Typical Login Screen for the dashboard

FEATURES

The dashboard has four major components - the title and navigation bar at the top, the full screen map panel, the information panel on the right and the control panel at the bottom. Figure below highlights these components on a typical screen. The title and navigation bar has two actionable items – 'About' button and 'Logout' button. The about button brings up a modal screen to show information about the dashboard along with links to the data used and the outputs. The map panel has two layers currently 'cluster layer' and 'median price' layer as shown in the screenshots below,
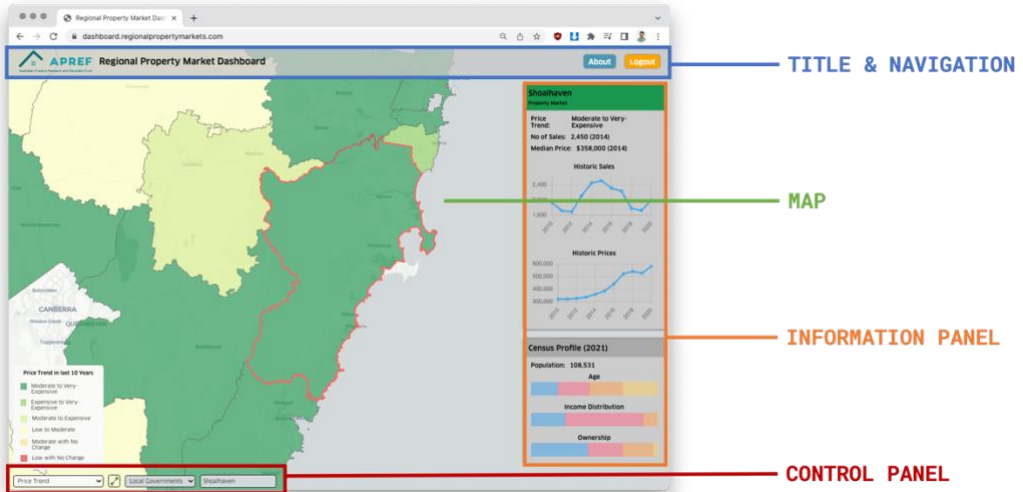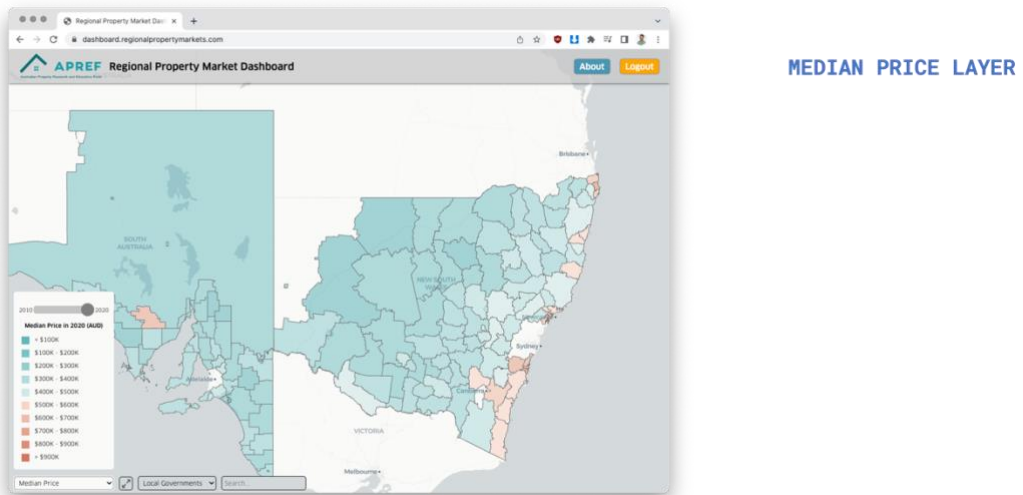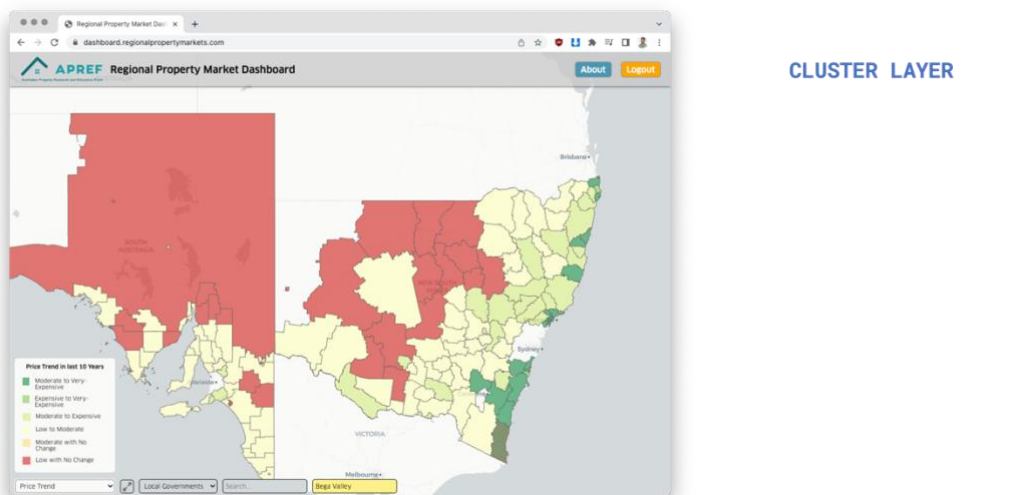
TITLE & NAVIGATION

MAP

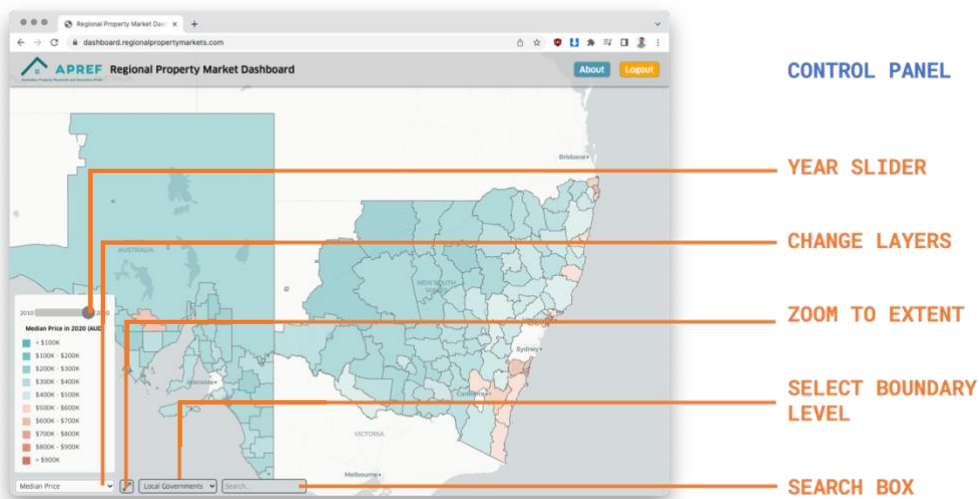INFORMATION PANEL

CONTROL PANEL

MEDIAN PRICE LAYER

CLUSTER LAYER

The map is interactive where users can zoom and pan the map using either the mouse or touch screen interface. The layers are interactive where the users can hover over a boundary to highlight it with the name shown in the control panel. The users can also click to select/un-select the boundary to toggle the display of the corresponding information panel.

The first drop down menu of the control panel allows the user to switch between the two layers. When on the median price layer one can use the year slider to change the year between 2010 – 2020. The second button resets the zoom and pan to the initial setting. The third dropdown allows the user to select the level of the boundary displayed on the map layers. The last text input box allows the user to list, search and select a particular area from the map.

The information panel is displayed when a boundary/area is selected on the map or using the search panel. It consists of 3 distinct sections – Market summary at the top, two interactive charts showing the historic prices in the middle, and three interactive bar charts showing the demographic profile of the market as shown in Figure below,
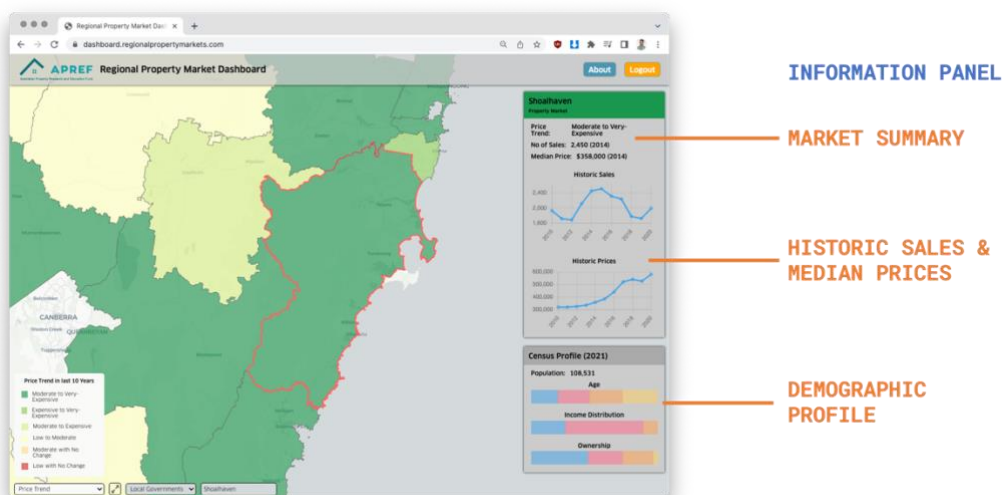
TROUBLESHOOTING & SUPPORT

The dashboard is a 'front-end native' tool where most of the application logic is carried out in the browser of the user. All the customisations and state are ephemeral and the dashboard resets to original condition when the browser window is refreshed. If there are any issues while accessing the tool, please follow the below steps to troubleshoot.

1. Refresh the browser window.
2. Restart the browser.
3. Clear cache and hard reload (Open the developer tools from chrome menu and long press refresh button to bring up reload options as shown in Figure below)
4. Use a different browser (This has been developed and tested for Google Chrome)
5. Use a different computer.

If all the above steps fail, please contact the developer or maintainer listed at the 'about' page with screenshots of errors in each step of troubleshooting. For login related issues contact API through https://www.api.org.au/about/get-in-touch/
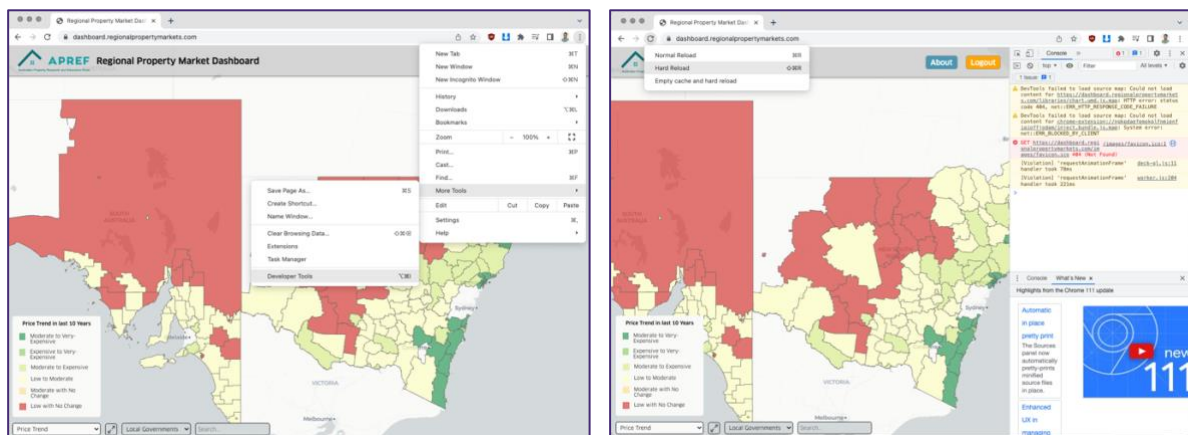
Clearing cache and doing a hard reload of the browser window.

DEVELOPMENT AND CONTRIBUTION

Although the data used in the visualisation is distributed under a non-commercial license, the source code is distributed under gnu-gpl3. Please email the maintainer if you'd like to get involved with the development of the tool and contribute to the code.