Data Science for Innovation

Assessment Stage 1 | Report

Agustin Ferrari | Nathan Collins | Yasaman Mohammadi | Luca Sardo

Assessment 1:	Selecting dataset, Defining problem & Project requirements Literature Review Exploring, Summarising and Preparing data
Type:	Group Assessment
Length:	2198 words
Weight:	20%
Due:	Sunday, 26 March, 23:59

Section 1: Problem

[1.1] Definition of Rental Stress

"Rental stress" describes the financial strain experienced by an individual or household when allocating income towards rental settlement (Smith, 2019). The Australian Institute of Health and Welfare classifies unaffordable rent and, subsequently, the environment that induces rental stress as exceeding 30% or more of the total income designated for rent (AIHW., 2018). Higher instances of rent stress can limit the financial capital available to meet further necessary expenses, such as food, education, and healthcare.

[1.2] Primary Causes of Rent Stress

Rent stress typically arises from isolated events or a combination of factors. These may range from, but are not limited to:

Rising rental costs
A downstream effect precipitated by increasing demand for rental properties. This factor is largely determined by location, accessibility, and tourism. A coalescence of these features typically sees higher degrees of stress within metropolitan regions (Jones, 2021).

Low income
As the costs of living change at a different rate to the average income, lower wages result in less financial capital, shifting totals allotted to rent and basic necessities (Gonzalez, 2019).

Underemployment & Competition
Fewer employment opportunities that compensate adequately to handle the costs of living limit the financial capital available to allocate towards rent. Inversely more opportunities present the desirability to relocate proximally to a workplace, ultimately increasing the rental costs within the region (Johnson, 2020).

Existing financial obligations
Rent stress may arise from or be exacerbated when a household experiences financial turmoil through existing debt (Smith, 2019).

[1.3] Consequences of Rent Stress

Rental stress creates immediate and long-term penalties. By limiting financial security, a restriction is placed on basic provisions, creating downstream impacts on a household's lifestyle. Factors such as debt accumulation and bankruptcy have the terminal outcome of increasing the incidence of homelessness (Johnson et al., 2021).

These extended periods of financial hardship likewise impact overall physical health. Through limited access to healthcare and nutritional choice, a person's well-being lessens itself as a priority. Such neglect is often accompanied by anxiety, depression, and mental health deterioration. The sense of struggle has the capacity to fracture a household's support networks, creating feelings of isolation. These health concerns all bear the downstream effect of impacting overall lower life expectancies (Brown et al., 2020).

Rental stress also impedes productivity at work or school institutions. Parents unable to provide their children with stable environments foster further negative influences on educational and social outcomes. The distraction can severely cripple careers and financial stability (Brown et al., 2020).

[1.4] Rental Stress in New South Wales, Australia

As of 2021, the Australian Bureau of Statistics (ABS) summarises that over 40% of renters in New South Wales (NSW) experience rental stress, making it the highest in Australia. This number has increased by 21,000 households since 2016. Rent stress is concentrated in Sydney, where costs of living and housing are the state's highest. The total number of rental households in NSW has increased by over 17.5% since 2016 to now over 2 million, representing a third of total households (ABS, 2021).

In 2023, the reserve bank raised interest rates to 3.10%, a change that has altered existing landlord-tenant agreements in NSW. As landlords preserved their own financial stability, so too did they see increases in rental prices, further elevating existing rent stress pressures and an existing housing affordability problem (RBA, 2023).

[1.5] Government intervention in New South Wales

The NSW government has implemented interventions to reduce rent stress and provide access to affordable housing through policies and initiatives. Two of the more significant programs are:

Government Housing
Access to housing for low-income earners who cannot afford to rent. It comprises public housing and Aboriginal housing, both owned and managed by the government.

Community Housing
Access to and support for community housing providers, an affordable housing alternative for low to moderate-income earners. These providers work closely with government agencies.

Rental assistance programs are also accessible to eligible low-income earners. This includes private rental assistance through the NSW Government and rent assistance through the Commonwealth. The NSW Government also applies zoning regulations to combat rent stress by ensuring that there is an appropriate mix of housing options available in different areas (NSW Gov, 2021).

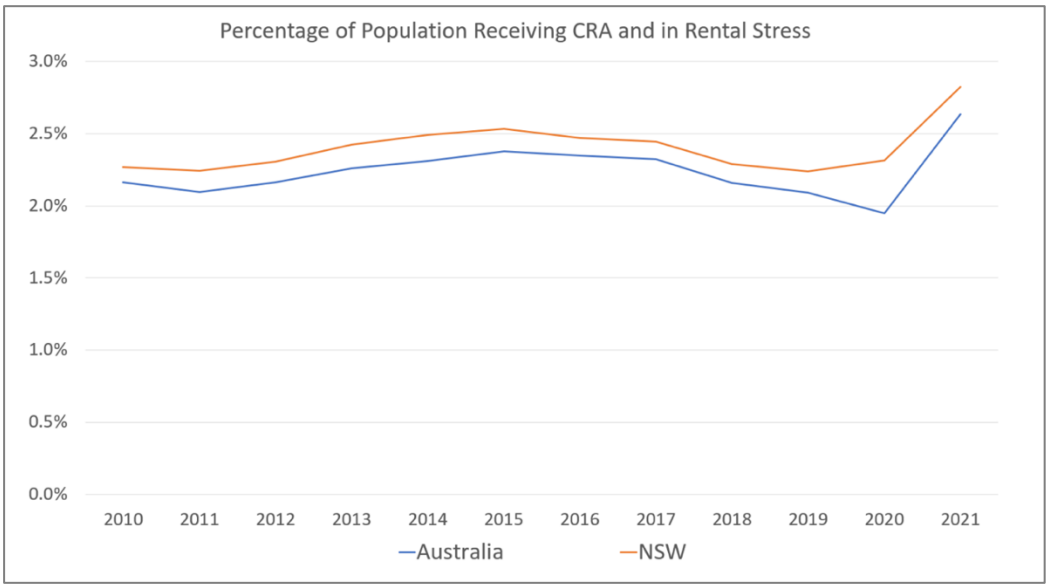

Figure 1 Percentage of the NSW Population vs. the Total Australian Population receiving Commonwealth Rental Assistance (CRA). This graph illustrates the recent increases (TU. NSW, 2021)

[1.6] Project Justification

The consequences of rent stress in NSW penetrating major proponents of lifestyle and health outcomes justify the grounds for statistical investigation towards the strengths of existing government interventions.

Research Questions formulated:

> *"What are the variables or characteristics that contribute to rent stress in New South Wales?"*
>
> *"Are existing government interventions helping alleviate rent stress in New South Wales?"*

The project will require a review of existing literature covering the nuances of rent stress within Australia and Internationally. Once completed, a scrape of ABS data packs will be conducted, then cleaned, explored, and visualised to gauge the effectiveness of government strategies for mitigating rent stress in NSW.

Section 2: Literature review

A literature review was conducted on seven papers, each concentrating on dynamics surrounding housing affordability.

The opening study examined general census data provided by the ABS in 2002, intending to identify determinants of **housing affordability stress** (HAS) in Australians over 55 (Tempe, 2008). The study highlighted particular characteristics related to housing affordability stress that subsequent studies later confirmed.

These were:

1. Stress induced by housing affordability is more pronounced in younger ages compared to older ones. Australians under 55 are four times more likely to afford housing and utilities.

2. Living alone places someone more at risk of housing affordability stress than being married to a couple, with no variance seen between genders.

3. A determining factor for this analysis is income; individuals receiving a “step income” with debt across multiple assets experience higher risks of stress.

Further investigation saw a compound of studies examining the influence of **public policies on housing expenditure** (Groenhart and Burke, 2014). Each touched on disruption in existing housing policies from the 1980s onwards, termed the 'neoliberal turn.' With the success of free-market theories, governments steadily reduced their investment in public housing sectors, resulting in a surge in privatised ownership within housing sectors. This led to the negative consequence of increased housing prices within lower socio-economic regions, creating fundamental determining factors that influence the prevalence of rent stress.

With an overall reduction in government support within the housing sector, the 2008 **global financial crisis** further altered the housing affordability stress dynamic and, subsequently, the incidence of rent stress (Wood et al., 2015). The groups most affected throughout the period were:

1. Lower-income families with children aged between 0-5 and 15-24 were most exposed to housing affordability stress.

2. Migrants from non-English speaking countries and groups experiencing difficulty accessing work.

When exploring a **regional cross-section of rent stress**, populations residing in New South Wales were consistently indicated as enduring peak HAS of all Australian states, with three-quarters residing in metropolitan regions (Troy et al., 2019). A number of factors are proposed to contribute (some of which are explored in this project), though all are predominantly derived from and compounded by population density. Studies of these regions furthermore illustrate the political orientation of these districts, with majorities tending towards a Labour government (Thackway and Randolph, 2021).

Recent explorations of 2021 HAS factors contain **data influenced by COVID-19**. A highlight revealed that 1 in 15 households endured either homelessness or severe rent stress; a trend, if left unchanged, would grow by another 300,000 individuals within 20 years (Van den Nouwelant et al., 2022). While the incidence of these will be highest within urban regions of Sydney (11% increase in HAS over 5 years), it will also penetrate other state cities, such as Brisbane and Melbourne.

A concluding consideration is the **definition and variables factored into rent stress**. As boundaries pertaining to rent stress are not categorised, they become subjective to the researcher and, therefore, can vary depending on their intent or the study's country of origin. In Australia, rent stress constitutes a >30% proportion of income allotted to cover rental expenses, though this threshold sometimes fluctuates to be lower in some studies at around 20-25% or higher at 40%. Comparing similar investigations in other countries, other expenses are occasionally factored in. In the United States, "rent burden" is instead calculated by the inclusion of utility expenses and rental insurance, which inflates total remuneration (Eggers and Moumen, 2010). If this share exceeds 50% of income, then the rent burden is considered excessive.

Section 3: Approach

[3.1] CRISP-DM

The project approach is based on the “Cross-Industry Standard Process for Data Mining”, a guideline for data mining across a variety of disciplines. This process facilitates recurrent review over project milestones as new information surfaces, insights improve, and objectives are completed.

For instance, the data cleaning, transformation or formatting phase is an iterative process that may require revisiting and improvement according to the needs that arise during the analysis.

[3.2] Feature Engineering

In order to fulfil Part 2’s requirements, feature engineering will be conducted, where some features will be converted into more comprehensive variables. Throughout the EDA process, current categorical features available reflect the highest level of educational attainment, but these features were scraped as smaller composites, like “Year 11” and “Year 12”. Feature engineering will be performed on education and health variables in order to provide the machine learning model with improved and comprehensive variables.

[3.3] Pre-processing

In the machine learning preparation phase, cross-validation remains paramount in practice. Data will be split into batches to help train, test, and validate a model’s output. First, it’ll be divided into a training set (80%) and a testing set (20%). This training dataset will then be split into multiple folds by using cross-validation, with the model trained and tested on different subsets of the data to mitigate overfitting. The testing set will be allotted for model assessment. Before predicting the model, however, all features will first be standardised, which will be achieved by extracting the mean and dividing it by the standard deviation.

[3.4] Multiple Linear Regression & Accounting for Non-Linearity

To meet the demands of our research questions, multiple linear regression models will be constructed. The objective is to fit a parametric, non-flexible model in effort to understand and interpret the relationships between explanatory variables and the overall variability of rent stress. Multiple linear regression models will serve as a foundation to identify features that may require further engineering.

Furthermore, non-linear relationships will be combined with multiple linear models to investigate possible non-linearities between the variables in rent stress (e.g., “age” squared, as alluded to in existing literature).

[3.5] Regularisation, Non-parametric Learning & Hyperparameter Tuning

The research will emphasise feature selection and machine-learning models with interpretation simplicity, including Lasso, Ridge, or Elastic Net regression models. In addition, more flexible machine-learning models will be constructed to gauge if loss values can be minimised and improve the overall understanding of rent stress predictors. Finally, hyperparameter tuning will be performed in order to increase general performance.

[3.6] Hypothesis Testing

As the project’s second research question targets the influence of government intervention, the null hypothesis would be established as:

> “*The government does not lower rent stress through social and community housing.*”

Hypothesis testing investigates whether to reject or accept it.

Section 4: Data

[4.1] Data Source & Description

Data was attained from “Quick Stats”, a publication by the Australian Bureau of Statistics (ABS) documenting census data from 2021. The data was categorised into selected geographical areas, where scrapings were obtained from each zone within the New South Wales state and concatenated into a singular statistical area known as Statistical Area 1 (SA1).

This resulted in a data frame with 19123 rows that represented every statistical area in the state of New South Wales, with 84 columns. 83 are explanatory variables (such as income, distinctive traits, health, etc.), and 1 is the outcome variable (Household Rental Affordability Index, RAID).

[4.2] Data Acquisition Process

All ABS data describing each geographical zone provides an embedded “Statistical Area 1” ID in its URL. A data scraping script was created to utilise these existing ABS URL formats.

Template URL:

> https://www.abs.gov.au/census/find-census-data/quickstats/2021/ + SA1**CODE**.

The list of corresponding (SA1) codes were sourced from ABS datapacks.

Data pack source:

> https://www.abs.gov.au/census/find-census-data/datapacks

Once extracted with Python, corresponding SA1 codes from these data packs were concatenated with their SA1 to acquire 19,000 HTML pages.

A Python script was then developed to scrape each URL and load the HTML content into Pandas. The script would subset the table number, with its rows and columns, based on the set features for the study.

Table 1 Data used and corresponding source from ABS repositories.

Data	Source
SA Codes	Datapacks
Sq Kilometres	Shapefiles
Rent stress interventions	Table Builder
Remain	Quick Stats

[4.3] Exploratory Analysis

All missing values within each category were first scouted for, and if there were any, whether they were randomised. All features comprising too many missing values were dropped, a disappointingly necessary process as some features appeared only in around half of all overall observations. The graphical tool MSNO facilitated a visual representation of all null values.

Table 2 Python packages imported for data exploration and visualisation.

Package	Function
Pandas	analysis
NumPy	analysis
Matplotlib	visualisation
Seaborn	visualisation
Missingno	visualisation of null values

Following the data cleansing process, "Car, as a passenger" remained the only variable with null values, where an impute method was exercised to resolve any missing features. The variable's distribution was examined prior to resolving through an aggregation metric. Skewness was also encountered and resolved by using the median instead of the mean for imputation. This process illustrated the importance of normalisation and standardisation of the dataset prior to training a machine learning algorithm.

Synopsis of the data preparation process:

Acquisition

Data was prepared by URL conversions and consolidated into a data frame, further using loops to concatenate.

Dropping & Symbol removal

Following visualisation, missing-value columns were dropped (with the conversion of one, "Car as a passenger"), and symbols within float columns (i.e., "$", "%", ",") were removed.

Transformation

The dependent variable was transformed from "absolute households" to "percentage of total households" to facilitate comparability between each feature.

Quality Checking

A "quality check" loop was created to set feature ranges between 0 to 100, with exceptions for "Density" or "Area sqkm."

Now with a complete data set, the outcome variable was described using the Python “Pandas” package, providing summary statistics and a general structure of each feature. By examining the difference between the mean, the median and value ranges, a right-skewed spread was identified. While the range of values is 79%, 50% of the observations fall between 7% and 20%.

```
Household RAID in percentage summary statistics
count    18721.000000
mean         8.810213
std          6.441095
min          0.000000
25%          4.455675
50%          7.548938
75%         11.562030
max         79.012346
Name: Household RAID, dtype: float64
```

Figure 2 Summary statistics of Household Rent Affordability Indicator (RAID), curated through Pandas .describe() function.

[4.4] Preliminary Data Insights

Following the initial exploratory phase, the first key insight was a distribution difference between Greater Sydney and the rest of NSW, emphasising a higher statistical prevalence for rent stress in Greater Sydney.

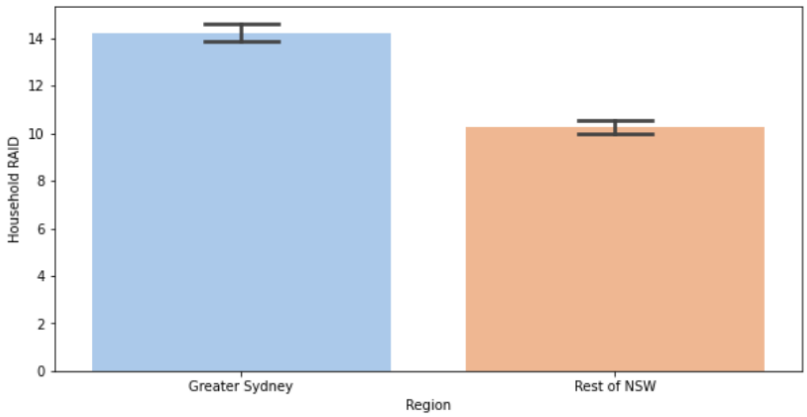

Figure 3 Rent Stress mean across NSW regions.

With Python’s Seaborn package, scatter plots with embedded linear regressions were generated to plot explanatory variables against the target. While features such as “tenure types” highlighted strong relationships to the target variable, the feature depicted ownership and was otherwise self-explanatory. Such features did not contribute to further insights into rent stress.

Other variables, such as the “percentage of married” or “never married”, were sidelined with scepticism, as they were observed as potential proxies for representing a group’s characteristics. This is due to individuals still sharing rental costs regardless of matrimonial commitment. Additionally, other columns with strong correlations, such as "average number of vehicles", were perceived as an indicator of wealth. Moderate correlations were identified between household composition features as well; however, these require further exploration, as "household groups" represented negative relationships, while "household families" were positive.

The ten most correlated features with the dependent variable were:

1. 'Tenure_type_rented',
2. 'Tenure_type_owned_outright',
3. 'Avg_num_vehicles',
4. 'Pct_never_married',
5. 'Pct_married',
6. 'households_composition_group',
7. 'households_composition_family',
8. 'Median_age',
9. 'Tenure_type_owned_mortgage',
10. 'Density'.

Relationships were also visualised with data concerning the percentage of households rented through community or public housing, yielding surprisingly positive correlations. Areas with higher proportions of public housing also returned higher rental stress. The linear model between these features alone opens exploratory pathways to uncover further confounding variables that may influence this relationship.

It's here that a learning algorithm that accounts for all features would prove valuable, as a simple linear model does not provide adequate proof that government interventions are necessarily lowering rent stress.

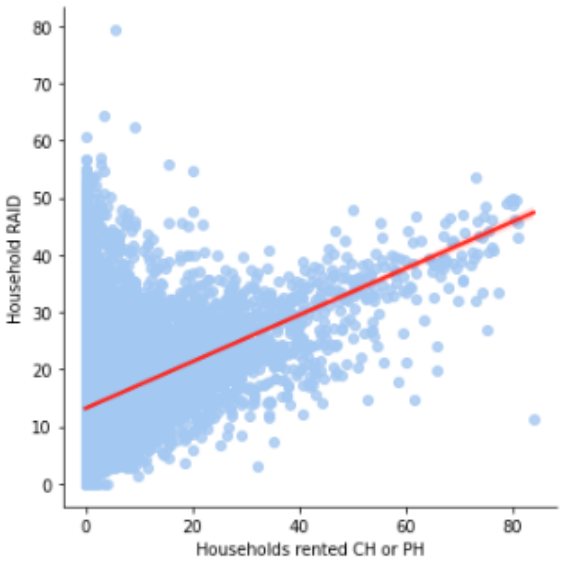

Figure 4 Relationship between Government Intervention through Community Housing (CH) and Public Housing (PH), against Rent Stress.

Bibliography

Literature Review

1. Eggers, F. & Moumen F. (2010), Investigating very high rent burdens among renters in the American Housing Survey, U.S. Department of Housing & Urban Development. https://www.census.gov/content/dam/Census/programs-surveys/ahs/publications/High_rent_burdens_v2.pdf

2. Groenhart, L. and Burke, T. (2014), What has happened to Australia's public housing? Thirty years of policy and outcomes, 1981 to 2011. Australian Journal of Social Issues, 49: 127-149. https://onlinelibrary.wiley.com/doi/10.1002/j.1839-4655.2014.tb00305.x

3. Temple, J.B. (2008), Correlates of housing affordability stress among older Australians. Australasian Journal on Ageing, 27: 20-25. https://doi.org/10.1111/j.1741-6612.2007.00268.x

4. Thackway, W. & Randolph, B., (2021). Housing, Financial Stress and Electoral Geography: An analysis of the spatial distribution of housing-associated financial stress in Australia. https://cityfutures.ada.unsw.edu.au/documents/660/Housing_Financial_Stress_and_Electoral_Geography_Report__FINAL_V4.pdf

5. Troy, L., van den Nouwelant, R., Randolph B. (2019), Estimating need and costs of social and affordable housing delivery. City Futures Research Centre https://cityfutures.ada.unsw.edu.au/documents/522/Modelling_costs_of_housing_provision_FINAL.pdf

6. Van den Nouwelant, R., Troy, L., Soundararaj, B. (2022), Quantifying Australia’s unmet housing need, City Futures Research Centre https://cityfutures.ada.unsw.edu.au/social-and-affordable-housing-needs-costs-and-subsidy-gaps-by-region/

7. Wood, G., Ong, R., and Cigdem, M. (2015) Factors shaping the dynamics of housing affordability in Australia 2001–11, AHURI Final Report No. 244, Australian Housing and Urban Research Institute Limited, Melbourne https://www.ahuri.edu.au/research/final-reports/244

Section 1

1. Australian Bureau of Statistics, ABS. (2021). Housing occupancy and costs, 2019-2020. Retrieved from https://www.abs.gov.au/statistics/people/housing/housing-occupancy-and-costs/latest-release

2. Australian Institute of Health and Welfare. (2018). Rental housing: A supplementary data analysis of Australia's rental environment. Canberra: Author.

3. Brown, K. & Jones, M. (2020). The relationship between housing affordability and health outcomes: A systematic review. Journal of Health Economics and Policy, 35(2), 87-105. https://doi.org/10.1016/j.jhep.2019.12.002

4. Gonzalez, M. (2019). The effects of income inequality on housing affordability: A study of low-wage workers. Journal of Poverty and Public Policy, 27(3), 227-243. https://doi.org/10.1002/pop4.237

5. Johnson, L. (2020). The impact of underemployment on housing affordability: A case study of urban centers. Journal of Urban Economics, 45(2), 87-105. https://doi.org/10.1016/j.jue.2019.12.002

6. Johnson, T. & Davis, J. (2021). The impact of rental stress on homelessness: An empirical study. Journal of Housing and Homelessness, 45(3), 167-181. https://doi.org/10.1177/10964581211005020

7. Jones, S. (2021). Understanding the impact of rising rental costs in metropolitan regions. Journal of Housing Studies, 38(2), 87-95. https://doi.org/10.1080/02673037.2021.1234567

8. NSW Government. (2021). Affordable housing. Retrieved from https://www.nsw.gov.au/topics/housing-and-property/affordable-housing

9. Reserve Bank of Australia, RBA. (2023, February 2). Media release: Statement by Philip Lowe, Governor: Monetary policy decision. Retrieved from https://www.rba.gov.au/media-releases/2023/mr-23-02.html

10. Smith, J. (2019). Rental stress: Understanding financial strain when allocating income towards rental settlement. Journal of Housing Economics, 42, 23-35. https://doi.org/10.1016/j.jhe.2018.10.002

Section 1 Figure

Tenants' Union of New South Wales. (2021, June 22). Census 2021: Renters are fastest growing tenure in Australia. Retrieved from https://www.tenants.org.au/blog/census-2021-renters-are-fastest-growing-tenure-australia

Data Scraping

Australian Bureau of Statistics. (2021). QuickStats: Australia. Retrieved from https://www.abs.gov.au/census/find-census-data/quickstats/2021/