MAST30034 Applied Data Science- Real Estate Consulting Project Group 40

Overview

This investigation aims to identify significant driving factors behind rental property prices in Victoria, as well as notable suburbs which are predicted to experience the most growth within the next few years, and those which are the most liveable. We present our findings to both prospective renters and investors, in the hope that they may make informed decisions before entering the market.

This summary notebook outlines the strategy taken - different avenues of information we explored- and our main findings, as well as the assumptions and limitations of this project.

Contents

Warning: The anchor link might not work properly on Github

1. Data

This project was centered around data sourced from domain.com.au, with observations of each rental property listing including advertised prices and property details such as number of bedrooms, address, property type and coordinate location. We first scraped 12,000 property listings from the site, however found this an insufficient amount to properly address the task of predicting future suburb growth. An API was thus used to obtain a larger dataset of approximately 1 million rows.

The first few weeks of the project were spent exploring potential sources of publicly available data:

- ABS Census Data:
    - Population Estimates by postcode and SA2 and above
    - Income Summary statistics by geography

- Liveability indicator data:
  - Crime rates by LGA
  - Hospital locations
  - School locations
  - Public transport data(train, tram and bus stop locations)
  - Property Boundaries- used to calculate property area.
  - Open space dataset
- Datasets regarding statistical area boundaries(eg. SA2s, LGAs, suburbs, postcodes) and their conversion:
  - Postcode to suburb conversions- scraped from the Australia Post website
  - postcode_mapping.csv - postcode to sa2?
  - sa2_mapping.csv - sa2 to sa3s
  - Digital boundary

2. Assumptions & Limitations

Due to time constraints and project feasibility, the following assumptions were made:

- Domain property data was an accurate representation of the general Victorian rental market:
  - Advertised prices were the prices that properties ended up being leased for
  - Statistical independence (observations are not dependent of each other)
- The effects of COVID-19 on the Victorian rental market were abided.

Other considerations that were not explored due to time constraints or lack of available data:

- Age/condition/amenities of property
- Types of occupants (eg. families, subletters) - if used as a sharehouse, dining/living/sun rooms may have been used as bedrooms of occupants- therefore the total number of rooms may have been a relevant factor
- Other resident demographic factors(eg. age, ethnicity)
- Projected future number of dwellings
- Proximity to natural attractions (eg. beaches, lakes)
- Number of jobs in suburb/neighbouring suburbs
- Using Manhattan distance instead of openrouteservice distance for the distance to CBD due to lack of computing power.

3. Data Pipeline

Our data pipeline consists of these following steps:

- **Collection**:
  - Scraped and collected the data from domain.com.au. Collected ~**1,000,000 records** in total, which is **50 times bigger** than the average of other team's dataset
  - Generated a list of urls to visit and iterated through all the rental pages of all postcodes in domain,

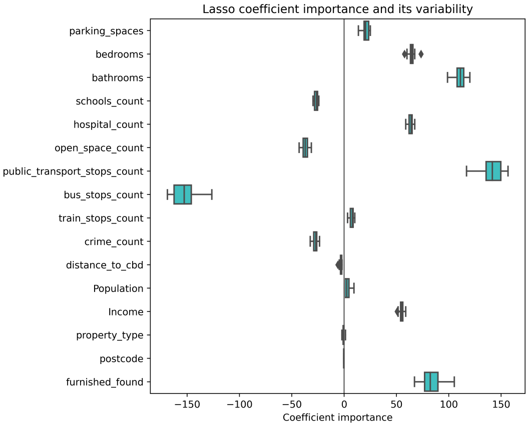
    - *Extensibility* : Accounting for minor technicalities such as the "read more" button appearing a few seconds later, and if the properties were not for sale then were skipped.
- **Outlier detection**:
    - Outliers were handled by first conducting a prior analysis on the distribution of rental prices, looking at the IQR, taking the top and bottom 2.5% (accumulatively 5%)
    - *Web verification*: double check the price on the web to make sure the data was retrieved without error and data is not incorrectly lost.
- **Partition**:
    - A stratified 80:20 split based on postcode was conducted, which ensures a good representation of the population.
    - Partitioning also helped with avoiding *information leak* by limiting the amount of sensitive information exposed at one time, which also keeps the training process unbiased and achieves the *privacy requirement*.
- **Imputation**:
    - Impute missing values based on the nearest neighbor. An apartment with missing values of bedroom will be imputed with the mode of bedrooms of all apartments within that suburb and similar records for other features such as the number of bedrooms and parking space.
    - *Extensibility* : Adopt for future extension for any random testing and validation set using the pipeline trained on training set. Simply call `fill_missing_values(validation_set, reference = train_set)`
- **Modelling**:
    - 2 machine learning methods are implemented and benchmarked: LASSO and LightGBM Regression
    - Similar to XGBoost and Random Forest, LightGBM is a gradient boosting algorithm which, similar to XGBoost and Random Forests, uses decision trees as its base learners but it grows trees leaf-wise, meaning it grows the tree by expanding with the maximum delta loss. It is widely used in many winning ML competitions.
    - LightGBM **outperforms** LASSO in all 3 metrics we used: MAPE (Mean Absolute Error), MSE (Mean Squared Error), MAE (Mean Absolute Error), which is partially because LASSO regression is not capable of learning the complex underlying relationship between features and response.
    - LightGBM will be used as the only model to answer the 3 questions given its strong performance

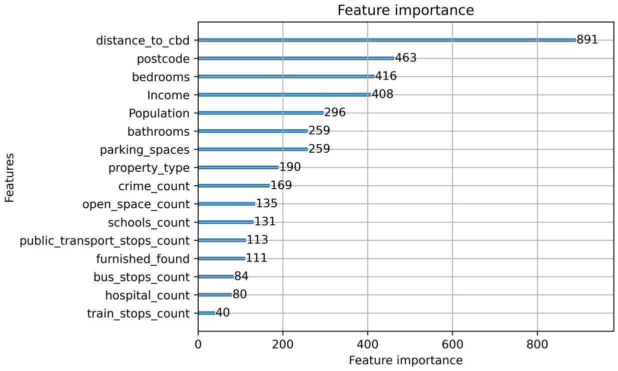
Benchmarking LightGBM and LASSO

Model	LASSO	**LightGBM**
MAE	138.75	**91.43**
MSE	91259.14	**52674.26**
MAPE	0.2306	**0.1435**

Q1: Finding the most significant features in predicting rental prices

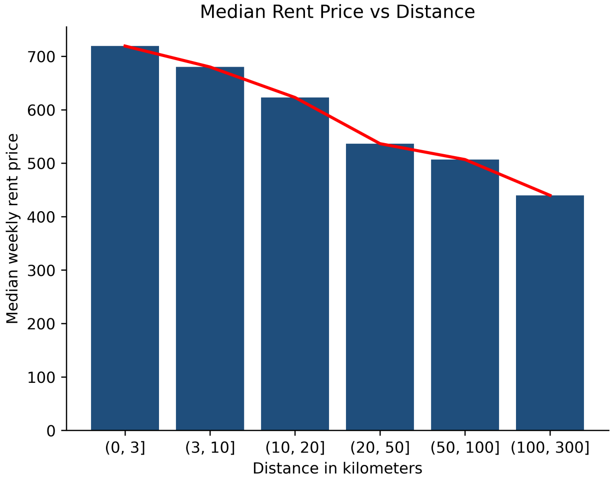
We also took into account the other top features that each selection model found, and LightGBM had more intuitively sensible returns. This combined with outperforming performance in evaluation metrics makes LightGBM the primary model used for this task and later tasks

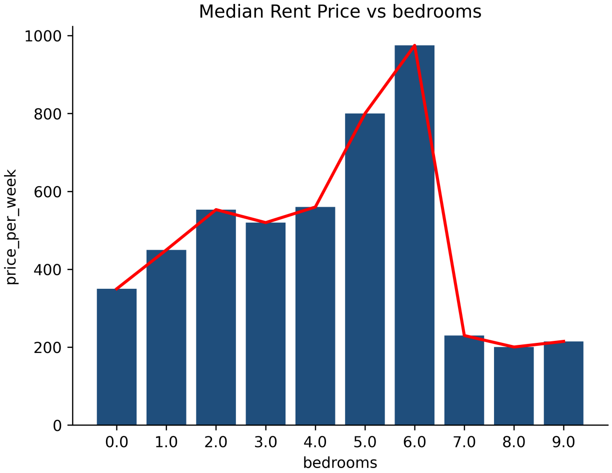




The top features for LightGBM are:

- **Distance to Melbourne Central (CBD)**: closer to distane, higher price
- **Postcode**: some postcodes are much more expensive than others: Docklands is much more expensive than Footscray
- **Number of bedrooms**: This is positively correlated with the property area, which is a highly useful for predicting rental price
- **Income**: Higher postcode income is postively correlated with higher rental price





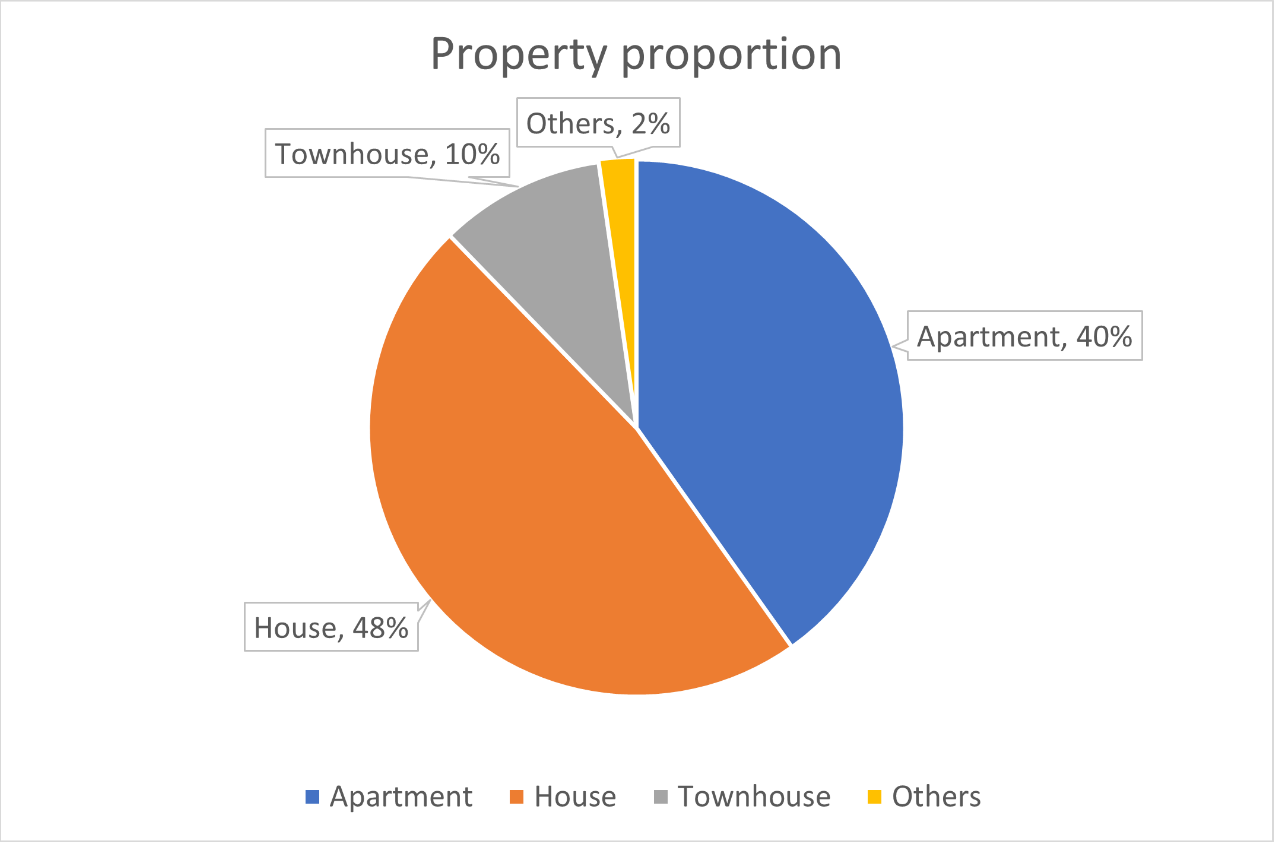
Q2: Predicting the fastest growing suburbs

2.1 Market Overview

- Real estate can be classified into many categories, but our team focuses on three major markets: townhouses, apartments, and houses. These three markets account for 98% of the total real estate market.

- We believe that breaking down the market in this way will attract more clients. It makes it easier for investors to find properties that fit their budget and investment goals. For example, some investors may only have the capacity to invest in a townhouse, while others may be able to invest in a large house or even a villa.

Top Suburbs for House Price Growth in 2026



2.2 Approach

- Our group uses 2 separate and independent methods to find top suburbs rising in price: statistical and machine learning methods.

2.2.1 Statistics

- Statistical methods are based on the assumption that historical data can be used to predict future trends. In other words, if we assume that there will be no major disruptions to the market, such as the COVID-19 pandemic, then the suburbs that have experienced the highest rental price growth in the past are likely to continue to thrive and show significant price increases in the future.

2.2.2 Machine Learning

- Our strategy to predict future suburb growth was to use the aforementioned model to estimate rental prices for the next several years, given estimates of future income and population. These estimates were then compared to current prices in order to measure growth.

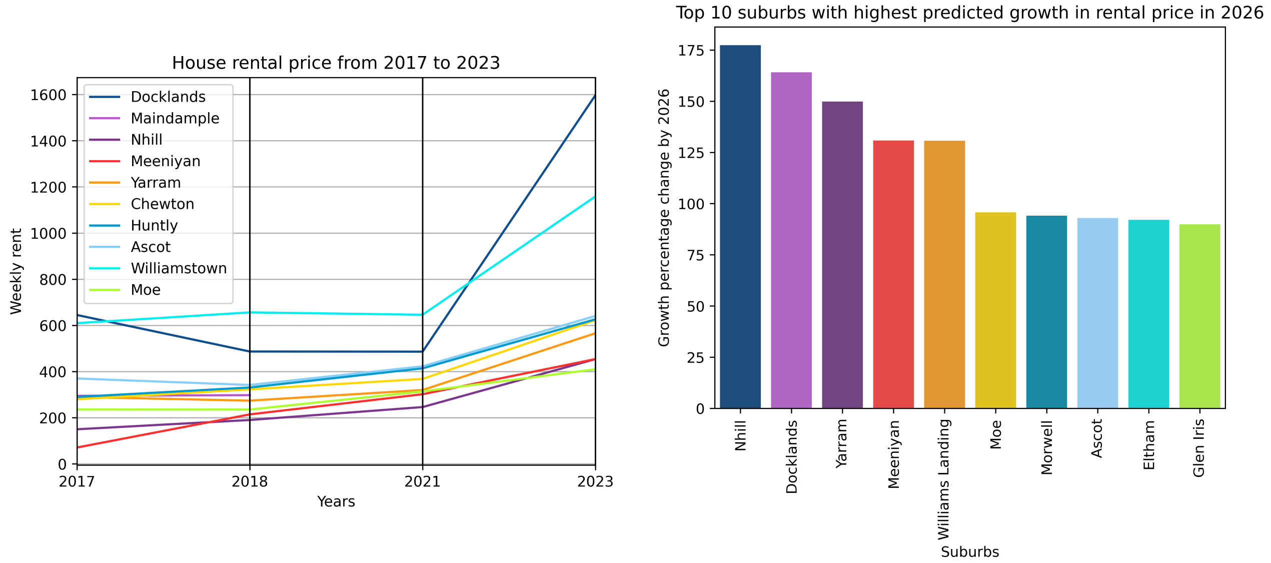
- Regarding **modelling**, a simple linear model was trained on historical population and median income collected from 2006 to 2021 ABS Census data for each postcode. The model was then used to predict values for the next several years. These new values were then used as input to the model in Q1, in order to obtain future rent price estimates. Other features (eg. infrastructure) were regarded as static, as we were unable to obtain data regarding future planned additions.

2.3 Results

- Two independent sets of suburbs with the highest rental price growth were obtained using the two methods. Although the order of the suburbs is not the same, the results of the two methods for house growth are very consistent, with 7 out of 10 suburbs overlapping.

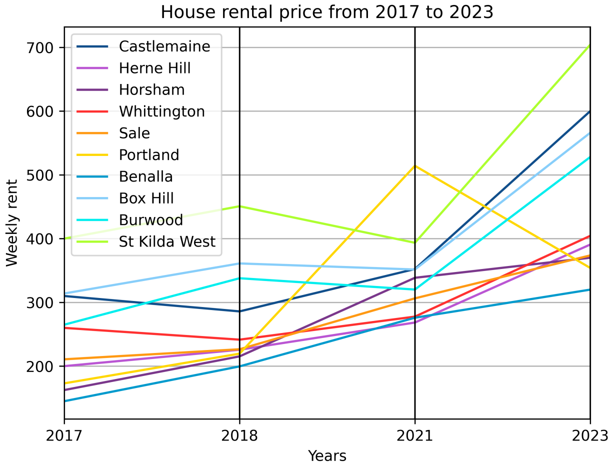
- Therefore, our team recommends that investors consider buying houses in suburbs such as Docklands, Nhill, and Meeniyan and rent out these properties to tenants, as these suburbs are predicted to have the highest rental price growth in 2026.

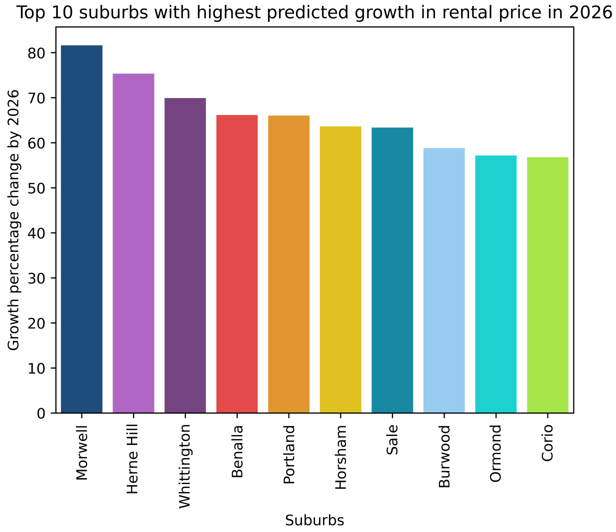
Top Suburbs for House Price Growth in 2026



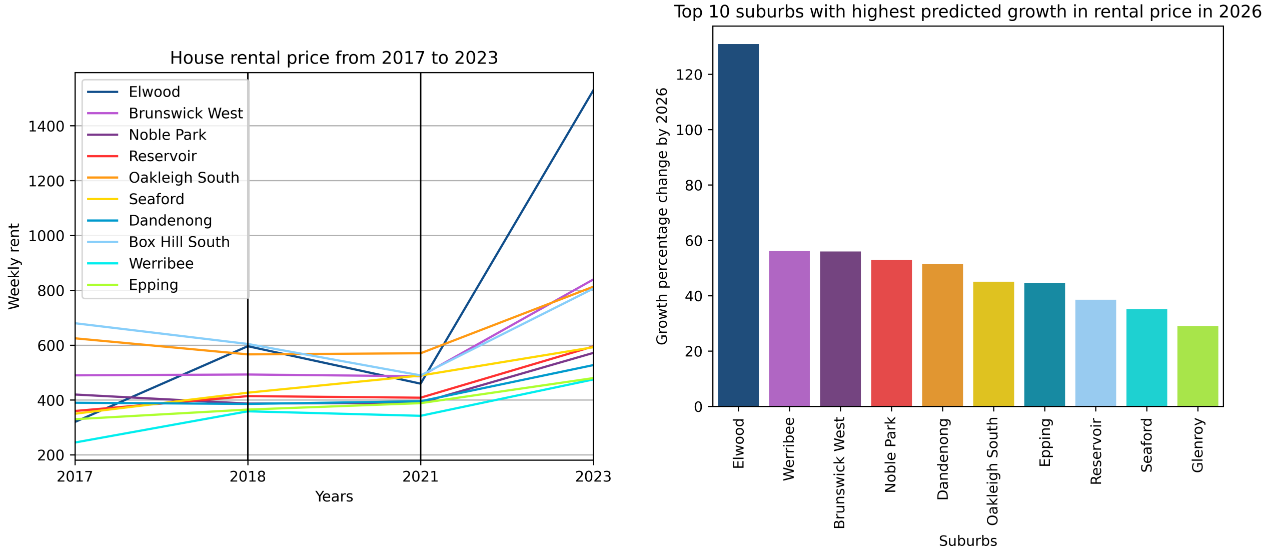
- Similarly, the same analysis was also applied to the other two property categories. The results of the two independent methods are highly consistent, which demonstrates the reliability and confidence of the results. For apartments, investors should consider investing in Portland, Heme Hill, and Sale. For townhouses, Elwood, Dandenong, and Oakleigh South are highly recommended.

Top Suburbs for Apartment Price Growth in 2026





Top Suburbs for Townhouse Price Growth in 2026



- Overall, our team provides investors with a quantitative solution segmented to each property market. This allows investors to choose which channel (house, townhouse, and apartment) to invest in depending on their financial budget and portfolio preferences.

Q3: The most liveable and affordable suburbs in Victoria

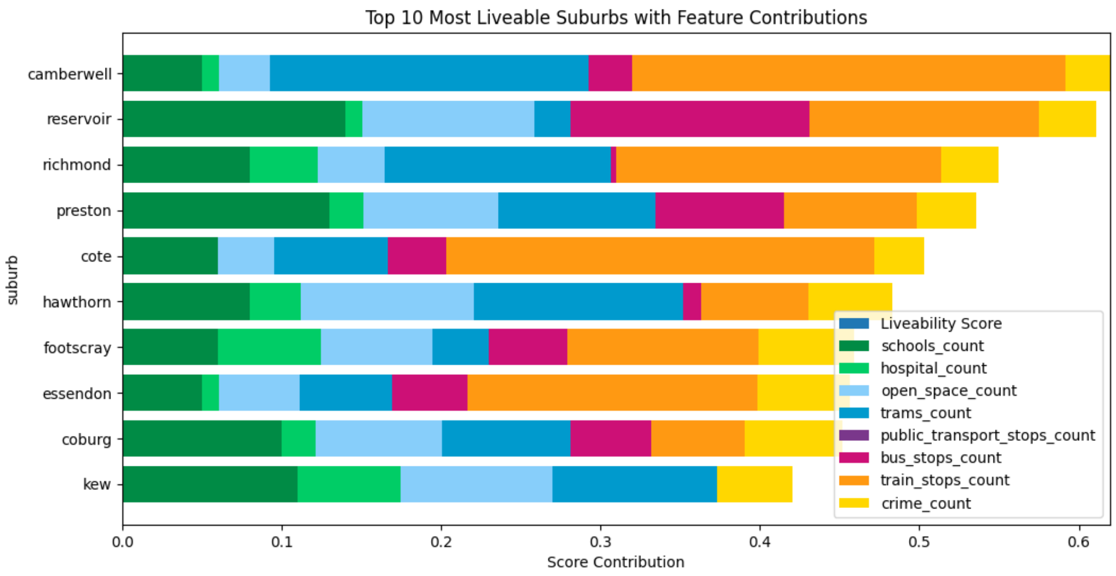
Measuring Liveability

To define liveability, we drew upon the Australian Urban Observatory's method of calculating it, which are linked to the United Nations' Sustainable Development Goals. These factors include:

- Access to healthcare and hospitals
- Access to schools, childcares, and universities
- Access to open and green space
- Connectivity through public transport
- Absence of crime

Link: https://auo.org.au/portal/metadata/urban-liveability-index/

All of these features were then standardised and given weightings based upon liveability impact. For example, the number of train stops has a higher weighting than the number of bus stops, given trains provide faster travel connections into the city and surrounding suburbs. Crime was also given a negative weighting, since it is an undesirable feature. Each suburb was then assigned a liveability score out of 100, with the highest ranking suburb being 100 and the lowest ranking 0. This allows us to easily compare and rank suburbs.

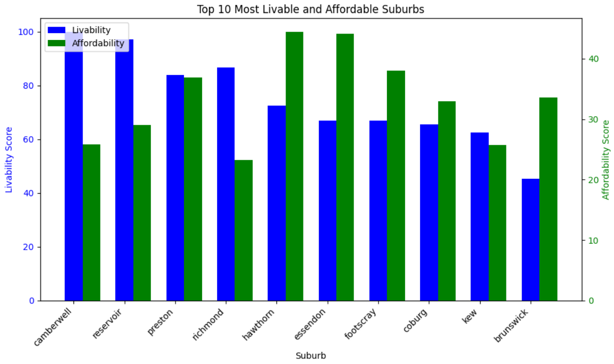


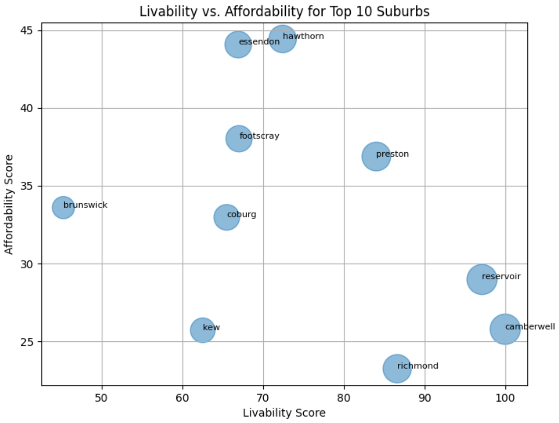
A significant contributor to the high scores of Camberwell, Richmond and Cote was the number of train stops. Further, Camberwell, Richmond and Hawthorn contain a high number of tram stops. These areas offer superior public transport options, and therefore may be more suitable for students and professionals commuting into the CBD for work or university. On the other hand, suburbs containing more schools, hospitals and open space like Resevoir, Preston and Kew may be more ideal for families or retirees.

Measuring Affordability

Affordability was calculated in line with the Cambridge Dictionary definition of affordability: 'the state of being cheap enough for people to be able to buy'. As we know, affordability is relative; what may be considered cheap in one area could be expensive in another.

Similr to the liveability calculations, we used the median income and median rent of a suburb as features, which were then standardised and weighted. Each suburb was again given a score between 0 and 100.





Camberwell, Resevoir and Hawthorn indicated significantly greater liveability as compared to affordability. This indicates that although the suburbs may offer greater convenience and superior amenities, they are less affordable for their residents. Renters in these areas may be less tolerant of further rent increases. In contrast, Hawthorn, Essendon and Footscray display greater affordability than liveability - while they are more affordable, their liveability score is still relatively high, and this may suggest undervaluation of rental properties.

7. Challenges

A major obstacle was that the data by area needed to be converted to suburb in order to address questions 2 and 3 . External datasets also came in different **granularities** so conversion between the 3 types (postcode, suburb, sa2) were needed, and transformations were not always 1-to-1.

Additionally, domain listings tended to have **inconsistent/missing data**, which required preprocessing. Addresses did not have consistent formatting- missing street numbers, misspelled street names, street type abbreviations and incorrect postcodes were a prevalent issue. These were dealt with using regex and a list of possible suburb names. Rent price was mixed up when sometimes price_per_week, price_per_month, price_per_year

While some property listings did provide **internal and land area data**, this was missing from most. We attempted to impute these values with estimations for each property, which had to me matched to listings by address. Property boundary data was accurate for determining land area of houses, and apartment area was estimated by dividing the total land area by number of apartments on the floor of that apartment. The latter proved to be not accurate enough- likely because it did could not account for the space taken up by hallways, elevators and gardens. As such, these values were insufficient to be used in the final model.

8. Future exploration

- Other Features: Consider other factors such as property area (in progress under `notebooks/supporting_notebooks/merging_area_data.ipynb`), expected future dwellings, expected net immigration, interests rates, gov. incentives

- Granularity: Instead of using affordability based on average income, divide population into different categories by income, then determine affordability for each of these groups
- Model Choice: Deep Learning Time Series forecasting: Transformer-based models (eg. Autoformer, N-BEATS). This will require much more massive data as the number of years coverd up to date is 6 years only.

9. Conclusion

This project allowed us to gain a more nuanced understanding of the Victorian rental property market. We found that indicators such as distance to the **CBD, postcode, income and number of bedrooms** were the most significant driving factors behind rental prices.

Additionally, we identified the fastest growing suburbs for each property type and investors should invest into these suburbs (refer to Q2 for more) and rent out to tenants because these suburbs will rise most in the future. On the other hand, if you are tenants/renters, avoid these suburbs because the expected increase in rental price is higher compared to other suburbs.

- House: Docklands, Nhill, Meeniyan, ...
- Apartment: Portland, Sale, Heme Hill, ...
- Townhouse: Elwood, Dandenong, Oakleigh South, ...

House is the market with most growth while apartment is the one that is the most stable among the 3 markets in 2026.

Following our analysis of **liveability and affordability**, our recommendations for investors are as follows:

- Rental prices in Hawthorn, Essendon and Footscray should be revised and brought in line with current market conditions.
- Camberwell and Reservoir rent should not be lifted, at risk of renters looking elsewhere.