

# ESTIMATING ANNUAL EARNINGS IN SYDNEY'S HOUSING RENTAL MARKET

A stylized silhouette of a city skyline, featuring various building shapes and a Ferris wheel, rendered in shades of teal and orange against a teal background.

---

## PROJECT DOCUMENT

CAPSTONE PROJECT | INSTITUTE OF DATA

WENDY MARIA DSA | DATA SCIENTIST

# INTRODUCTION

## INDUSTRY OVERVIEW

Owning a residential property in Australia has historically been a safe and proven way to create personal wealth and has been shown to deliver a better return than the stock market over the long term. Results over a twenty-year period to December 2015 reveal that residential properties returned 10.5% per annum, beating both Australian shares at 8.7% and global shares at 7.6%<sup>1</sup>. Additionally, despite the global Covid19 pandemic and resulting widespread job losses, economic recession, and trade disruption 2021 has been an especially lucrative year for Australian property owners. The residential real estate market value surged to a record \$9 trillion plus, gaining the last \$1 trillion in just five months<sup>2</sup> with the Sydney market alone rising by 23.6%<sup>3</sup>. The residential market now sits 28.2% higher than the estimated value of superannuation, the ASX, and commercial real estate combined.

Alongside this increase in market value, 2021 also recorded an 8.9% increase in rental rates, the largest such increase since 2008<sup>4</sup>. While the combined increase in the Australian capital cities was 7.5%, the regional areas saw a mammoth 12.5% increase on the back of record migration to the regions during the Covid pandemic<sup>5</sup>. Weekly rents in more than 20 regional NSW markets jumped by 10% or more, with asking rents in five regions outstripping those in Greater Sydney which rose by 7.2%<sup>6</sup>.

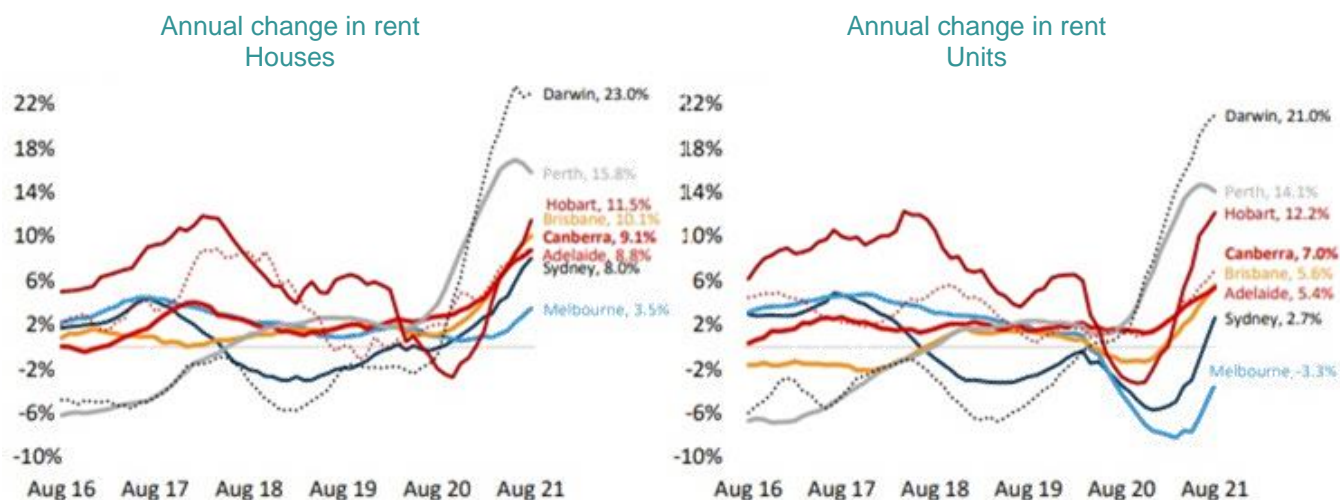


Figure 1: Australian Rent Prices August 2021. Screenshot courtesy of Corelogic.com.au

## PROBLEM STATEMENT

- **The Rental Market Conundrum**

Despite the rapid growth in rent values, gross rent yields have compressed nationally over the past year, due to the higher rapid rise in property values<sup>7</sup>. Rental yield measures the profit generated each year from investments as a percentage of its value. If a property rapidly increases in price but the asking rental increase does not follow the same trajectory, then the annual return or yield will fall. The level of rental return you can expect from an investment property is determined by several factors, including location,

type of property, and overall economic conditions. Thus, it is imperative that the rental income strategy is carefully planned. When you invest time and money into a property, getting the best possible return on investment is the main goal. This might mean opting for a short-term rental (STR) instead of a long-term rental (LTR) or vice versa. Regardless of the option chosen, there are some advantages and disadvantages to both options.

A STR (commonly called vacation rental) is a furnished self-contained living space or property that can be rented for upwards of a day to as long as 6 months. A LTR relates mostly to an unfurnished rental which is leased for at least 6 months or on annual fixed term contracts. With the advent of companies such as AirBnB and Stayz in Australia, STRs have become very popular as an alternative to the more traditional but pricier hotels in tourist destinations. They allow the owner greater flexibility in extending tenancy contracts weekly or monthly and is easier to keep up with repairs but are heavily affected by seasonality and shocks such as the Covid19 pandemic which saw yields plummet due to lockdowns or environmental disasters such as bushfires that slow down tourism. LTRs, on the other hand, provide more stability and consistency with long-term contracts and is not affected by seasonal variance that would make it difficult to rent the property, but will have more wear and tear from long-term occupancy and less flexibility in terms of owner's access to the property.

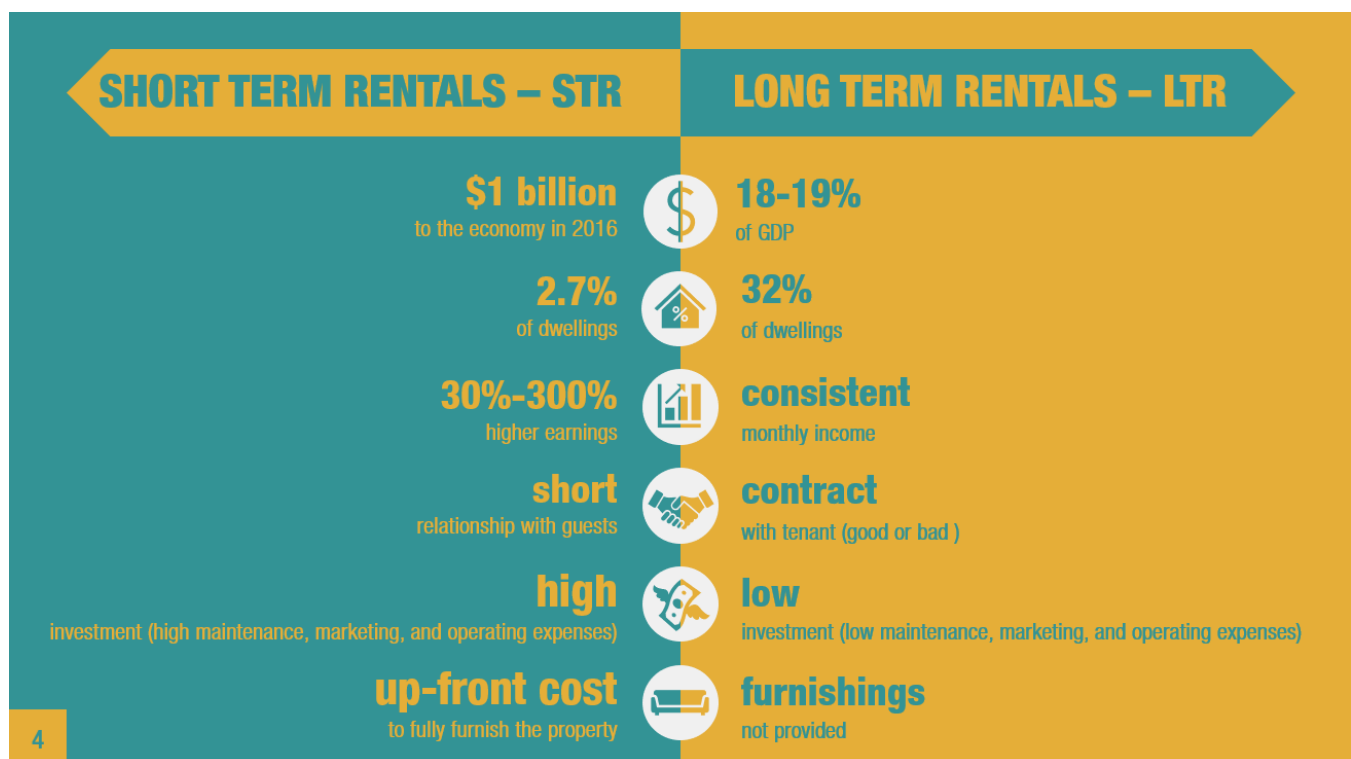
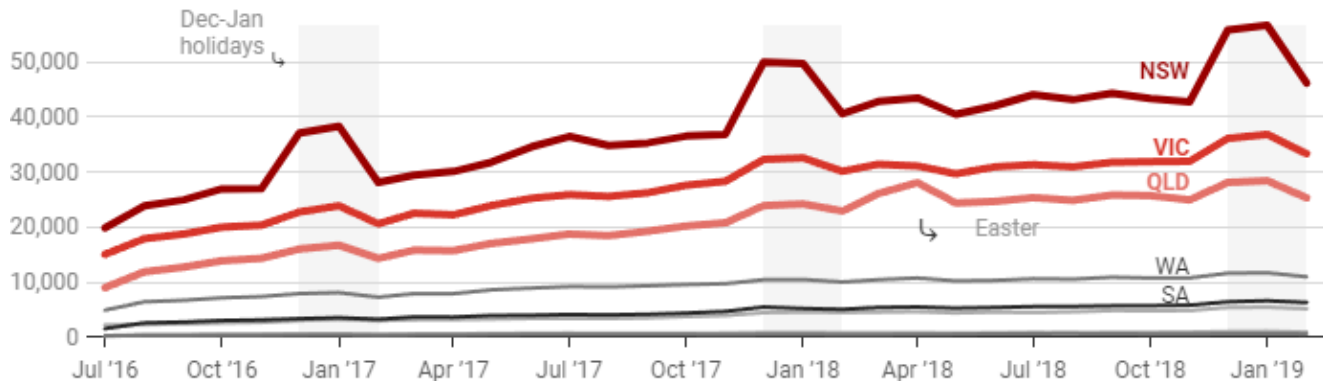


Figure 2: Difference between STRs and LTRs. Source – author

While it may seem like a tiebreaker that can go either way, there is one notable factor that make STRs more popular and that is its earning potential. STRs can be a lucrative market as they can command from 30% - 300% higher rates than LTRs depending on the location<sup>8</sup>. In Australia, the largest proportion of accommodation was provided by NSW, which supplied over 2.5 million nights of accommodation in 2016 alone with Sydney dominating with the largest supply and demand for STRs<sup>9</sup>. Also, due to the high proportion of listings of entire homes, Sydney's average daily rate (ADR) is closest to that of hotels, as compared to other major cities, sometimes commanding between \$200 -\$450. In 2016-2017 the weekly

STR average for apartments in Sydney was \$733, while houses bought in an average of \$1211 per week. In contrast, the weekly LTR average for both apartments and houses were \$550.



Grey highlight ranges indicate the holiday months of December and January

Figure 3: Airbnb listings by state, July 2016 to February 2019. Source – The Conversation<sup>10</sup>.

- **Change in STR Laws**

STRs would seem the most lucrative rental income strategy to embrace, which has been true till now. The number of listings has grown by an average of 2.43% a month over the period between July 2016 and February 2019<sup>11</sup>. This proliferation of STRs has come with its own set of costs. The single biggest potential costs come in the form of higher housing costs if enough properties are converted from LTRs to STR accommodations. As property owners move from LTRs to STRs, the decrease in the supply of LTRs increases surrounding housing costs for the residents. Research shows that the already tight rental vacancy rates in high demand suburbs of Sydney have been exacerbated with STR homes amounting from 1.5 times the rental vacancy rates in central Sydney to nearly 4 times in the eastern beach suburbs<sup>12</sup>.

The NSW Government has thus implemented a new statewide regulatory framework for STRs accommodation, effective 1 November 2021<sup>13</sup>. The new legislation limits the number of days in which non-hosted STRs (host living off-site) can be rented out. The hosts of both hosted (hosts living on-site) and non-hosted residences will also have to register with the NSW Government. The register will capture the number of days a property is used as an STR and will be integrated with key STR booking platforms, allowing for improved monitoring of the policy's day limits. Furthermore, upon registration, hosts are required to confirm compliance with fire-safety standards. Hosted STRs can be rented 365 days per year, however non-hosted STRs can now be rented for only 180 days per year and for no more than three consecutive months in Greater Sydney and nominated regional NSW local government areas. These areas include the Greater Sydney Region, Ballina, Bega Valley, Newcastle, Dubbo, Clarence Valley, and Muswell brook. The government plans to extend these regulations eventually to the rest of NSW.

Furthermore, in April 2020, strata laws changed in relation to short-term rental accommodation. Owners' corporations are now permitted to implement by-laws that limit STRs in their strata scheme, by banning it in lots that are not the host's principal place of residence. However, if someone lives in a strata property as their principal place of residence, they will still be able to rent out their home or rooms while they live there, or while they are temporarily away<sup>14</sup>.



Figure 4: Overview of new laws for STRs in NSW. Source – author

### • The Stakeholders

The new NSW STR laws have now made it difficult for non-hosted property owners to get the high returns that they would normally earn from a year long STR listing. With rising home valuation and lower rental yields it has become even more important that owners are able to decide and choose the best option that would earn them the maximum potential rental income annually. Some such stakeholders are:



**A multiple investment property owner** who would now like to decide whether to stick with a STR strategy or move towards a LTR strategy across the investment portfolio or mix and match according to individual properties.



**A future investment property owner** who would like to understand which suburbs or areas would get the best rental yield.




**STR Management companies** who would like to decide on the best pricing to get the maximum return for their clients given the new STR laws.



**Real estate management companies** who would like to decide the best rental investment strategy for their clients

## BUSINESS QUESTION

The above problem statement gives rise to the following business question:



Short Term vs Long Term

**WHICH REAL ESTATE  
INVESTMENT IS RIGHT FOR YOU ?**

There are various online platforms such as AirDNA, Transparent, PriceLab, etc. that provide web-based applications that help design STR pricing strategy for AirBnB listings based on accurate historic data. Furthermore, you are generally required to input in your own expected average price ('base price'), and the algorithm will vary the daily price around that base price on each day depending on day of the week, seasonality, how far away the date is, and other factors. But there is not yet a platform that will assist stakeholders in deciding which rental strategy is best for them. This project aims to bridge this gap by aiming to inform the design of a minimal viable base application or decision-making framework that would assist stakeholders estimate their annual rental returns from both a STR and a LTR rental strategy and help decide which strategy to embrace after factoring in other miscellaneous costs.

While there are pros and cons to both STRs and LTRs, the fact remains that achieving the right balance not only benefits the property owner but also the economy at large. The growing STR sector contributed over \$1 billion revenue to the Australian economy in 2016 and supported more than 40,823 FTE jobs during the same year<sup>15</sup>. Consumption spending on housing services from LTRs, on the other hand, has contributed to about 18%-19% of GDP since 1990, including gross rents and utilities paid by renters, as well as utility payments and owners' imputed rents (an estimate of how much it would cost to rent owner-occupied dwellings). They also bring about diversity and community cohesion and boost the local economy<sup>16</sup>. Hence, achieving an accurate representation of LTRs and STRs within a particular suburb or LGA would provide information to state governments and local councils for future economic planning and development.

## DATA QUESTION

In order to provide a decision-making framework or application for stakeholders the first step is to estimate the earning potential of a STR rental strategy in line with the new NSW laws, i.e., predict the ADR of a STR based on various factors such as location, size, features, etc. and then multiply it with occupancy rates based on historical data. It is important to get the pricing right, particularly in big competitive cities like Sydney where even small differences in prices can make the difference between optimum occupancy and high earnings, which is rather difficult to achieve, or being priced out of the market.

For this project, we use data from AirBnB listings in December 2018 for Greater Sydney. We use AirBnB data since it is the dominant online STR marketplace with 50% of stock exclusively listed on the platform<sup>17</sup>. Just 25% of non-hosted entire home STR properties have a diversified distribution strategy across multiple platforms. This project aims to answer the following:

1. Based on AirBnB data and neighbourhood data how accurately can we predict the price of a STR in Sydney?
2. What features can help predict the price?
3. What can we do with the predictions?

---

# THE DATA SCIENCE PROCESS

## THE DATA

- **Data Acquisition**

The dataset used for this project comes from Insideairbnb.com, an anti-Airbnb lobby group that scrapes AirBnB listings, reviews, and calendar data for multiple cities around the world. The dataset used in this project was scraped on 7th December 2018 and contains information on Sydney AirBnB listings that were live on the site on that date.

The data is rather messy and has various limitations. The biggest among them is that it only includes the advertised price called the 'sticker' price – price being the target variable we want to predict. The sticker price is the nightly price that is advertised to potential guests, rather than the actual average amount actually paid per night by previous guests. The advertised prices can be set to any arbitrary amount, and hosts with less experience on AirBnB often set these to very low (\$0) or very high amounts (\$14,999). Nevertheless, this dataset can be used as a base proof of concept. The price from more accurate versions can be queried at a later stage using more precise data from the APIs of companies such as AirDNA, Transparent, PriceLab, etc. that sell higher quality AirBnB data. Since the dataset contains unique listing ids this can be used as the main feature during future querying.

1. **Original Scraped Dataset link:** <http://insideairbnb.com/get-the-data.html> This dataset has since been updated for 2021 and the 2018 dataset has been archived and is no longer available. For our predictions we would need to use pre-covid dataset as anything from January 2020 would not follow regular patterns and pricing in Sydney due to international and national border closure.
2. **Used dataset link:** <https://www.kaggle.com/tylerx/sydney-airbnb-open-data> This dataset was extracted from the above insideairbnb link and posted on kaggle. The dataset is from listings in Dec 2018.

- **Data Cleaning**

The original scrapped dataset contained 36,662 unique listings as observations with 96 features. The original features are listed in the appendix at the end. There are 23 integers, 10 boolean, 15 categorical, 6 currency, 6 datetime, 23 integers, 6 numeric, and 30 string data types.

NLP will not be used in the creation of the initial model since we are only interested in estimating earnings. Hence, the text columns were dropped, as were other columns not useful for predicting price (e.g. url, host name, and other features that are unrelated to the property). The features dropped in excel were:

1. **scraped data info:** scrape\_id
2. **listing information urls:** listing\_url, thumbnail\_url, medium\_url, picture\_url, xl\_picture\_url
3. **listing string descriptions:** name, summary, space, description, neighborhood\_overview, notes, transit, access, interaction, house\_rules



4. **other listing information:** experiences\_offered(blank), square\_feet (over 90% blank)
5. **host information urls:** host\_url, host\_thumbnail\_url, host\_picture\_url
6. **host string descriptions:** host\_name, host\_about, host\_verifications
7. **other host information not relevant:** host\_id, host\_acceptance\_rate (blank), host\_neighbourhood (similar to host\_location), host\_listings\_count (similar to host\_total\_listings\_count), calculated\_host\_listings\_count, host\_has\_profile\_pic, host\_identity\_verified
8. **redundant geographical information:** street (string information - similar to neighbourhood/neighbourhood\_cleansed/city), neighbourhood\_cleansed (similar to neighbourhood/city), neighbourhood\_group\_cleansed (blank), city (similar to neighbourhood), state, market, smart\_location (similar to neighbourhood), country\_code, country, is\_location\_exact
9. **calendar information:** calendar\_updated, calendar\_last\_scraped, has\_availability
10. **review information:** first\_review, last\_review, reviews\_per\_month (similar to number\_of\_reviews)
11. **business information:** requires\_license, license, jurisdiction\_names, is\_business\_travel\_ready

Additionally, since the project goal is to estimate earnings for non-hosted property owners, observations that had 'private' and 'shared room' types were removed. Also, any commercial property types such as hotels, hostels, bed and breakfasts, and boutique hotels were removed since the new rulings do not apply to commercial ventures. The remaining property types were further categorized as either 'House' or 'Apartment' and any which did not adhere to the above type were removed (eg. Yurt, Log-cabin, tent, etc.).

Some amenities are more important than others (e.g. a balcony is more likely to increase price than a fax machine), and some are likely to be fairly uncommon (e.g. 'Electric profiling bed'). A selection of the more important amenities was extracted and further grouped as below. One way to reduce the number of features to avoid the curse of dimensionality is to remove the amenities which add relatively little information or are relatively unhelpful in differentiating between different listings. The amenity features were one-hot encoded and those that contained fewer than 10% of listings were removed.

1. 24-hour check-in
2. Air conditioning/central air conditioning
3. Amazon Echo/Apple TV/DVD player/game console/Netflix/projector and screen/smart TV (i.e. non-basic electronics)
4. BBQ grill/fire pit/propane barbeque
5. Balcony/patio or balcony
6. Beach view/beachfront/lake access/mountain view/ski-in ski-out/waterfront (i.e. great location/views)
7. Bed linens
8. Breakfast
9. Cable TV/TV
10. Coffee maker/espresso machine
11. Cooking basics

12. Dishwasher/Dryer/Washer/Washer and dryer
13. Elevator
14. Exercise equipment/gym/private gym/shared gym
15. Family/kid friendly, or anything containing 'children'
16. Free parking on premises/free street parking/outdoor parking/paid parking off premises/paid parking on premises
17. Garden or backyard/outdoor seating/sun loungers/terrace
18. Host greets you
19. Hot tub/jetted tub/private hot tub/sauna/shared hot tub/pool/private pool/shared pool
20. Internet/pocket wifi/wifi
21. Long term stays allowed
22. Pets allowed/cat(s)/dog(s)/pets live on this property/other pet(s)
23. Private entrance
24. Safe/security system
25. Self check-in
26. Smoking allowed
27. Step-free access/wheelchair accessible, or anything containing 'accessible'
28. Suitable for events

The remaining cleaning process is meticulously described in the Python code notebook, which can be found at <https://github.com/wendydsa/Capstone-Estimating-Annual-Earnings-in-Sydneys-Housing-Rental-Market.git>

## • Feature Engineering

In addition to removing the above features 7 additional features were added based on geographical information. These were mainly geocoded using the web application <https://2kmfromhome.com/20km> with the centre of the suburb as the geographical centre to compute the radius. The features are as follows:

1. **'neighbourhood'** feature was cleaned and backward geo-verified wrt to city, zipcode, latitude, and longitude. 'neighbourhood' was changed to 'suburb'
2. **'lga'** feature was added to cluster the neighbourhood values according to respective LGA to bring down the number of neighbourhoods
3. **'dist\_to\_cbd'** was added as distance of suburb from cbd (townhall as centre) and further numerically encoded:  $\leq 5\text{km} = 1$ ,  $\leq 10\text{km} = 2$ ,  $\leq 15\text{km} = 3$ ,  $\leq 20\text{km} = 4$ ,  $\leq 25\text{km} = 5$ , and  $25\text{km}+ = 6$
4. **'cafe\_density'** was added from domain.com information (5 = best to 1 = worst). Source: <https://www.domain.com.au/liveable-sydney/sydneys-most-liveable-suburbs-2019/sydneys-top-suburbs-for-density-of-cafes-and-restaurants-908569/#>

5. **'transport\_rating'** was added from domain.com information (5 = best to 1 = worst) by averaging train access and bus access ratings. Source: <https://www.domain.com.au/liveable-sydney/sydneys-most-liveable-suburbs-2019/tackling-the-citys-transport-woes-sydneys-congested-suburbs-and-the-best-for-public-transport-revealed-in-new-study-907999/>
6. **'beach\_access'** was added by geocoding centre of suburb to beach access ( $\leq 2\text{km} = 5$ ,  $\leq 4\text{km} = 4$ ,  $\leq 6\text{km} = 3$ ,  $\leq 8\text{km} = 2$ ,  $8\text{km}+ = 1$ )
7. **'cultural\_access'** was added by geocoding centre of suburb to tourist centres (museums, arts, galleries, shopping, theatres...etc) ( $\leq 2\text{km} = 5$ ,  $\leq 4\text{km} = 4$ ,  $\leq 6\text{km} = 3$ ,  $\leq 8\text{km} = 2$ ,  $8\text{km}+ = 1$ )
8. **'nat\_park\_access'** was added by geocoding centre of suburb to large national parks and nature trails access ( $\leq 2\text{km} = 5$ ,  $\leq 4\text{km} = 4$ ,  $\leq 6\text{km} = 3$ ,  $\leq 8\text{km} = 2$ ,  $8\text{km}+ = 1$ )

The pipeline is quite reusable to query future data since the suburbs ultimately chosen adhered to the domain.com suburb listing. The numerical features were log transformed to try and normalise them, although features such as cleaning fee or security deposit would not be completely normalised due to the existence of \$0. Finally the categorical columns were one-hot encoded. The final dataset had 22,396 observations and 95 features.

## DATA ANALYSIS

A correlation matrix was plotted to ascertain the correlation between the various features. Large number of features had no correlation to each other. Some of the notable high correlations to the target price were beds, bedrooms, bathrooms, accommodates, security deposit and cleaning fee as well as with property\_type house. Some of the negative correlations to the target price were all the review scores, while property\_type apartment was the most negatively correlated, which made sense since higher the price less likely was it to be an apartment listing.

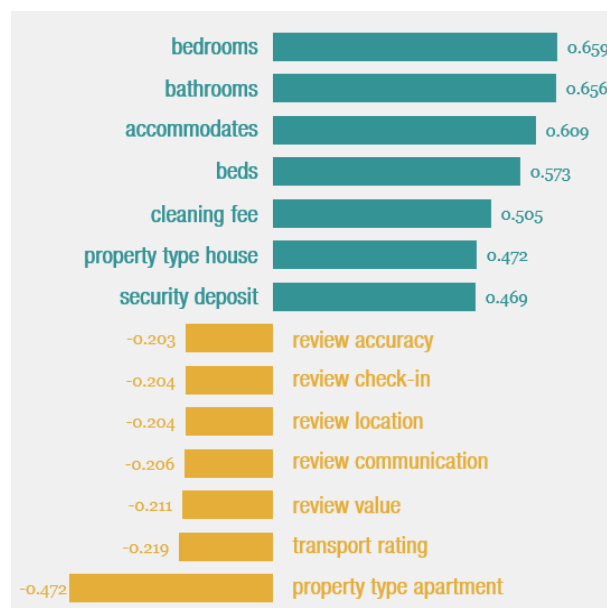


Figure 5: The Top 7 and the Bottom 7 correlated features to price. Source – Author

The top 20 suburbs were topped by the beach suburbs of Bondi and Manly and the centrally located suburb of Surry Hills. Analysis showed that most of the listings were located along the eastern and northern coasts of Sydney.

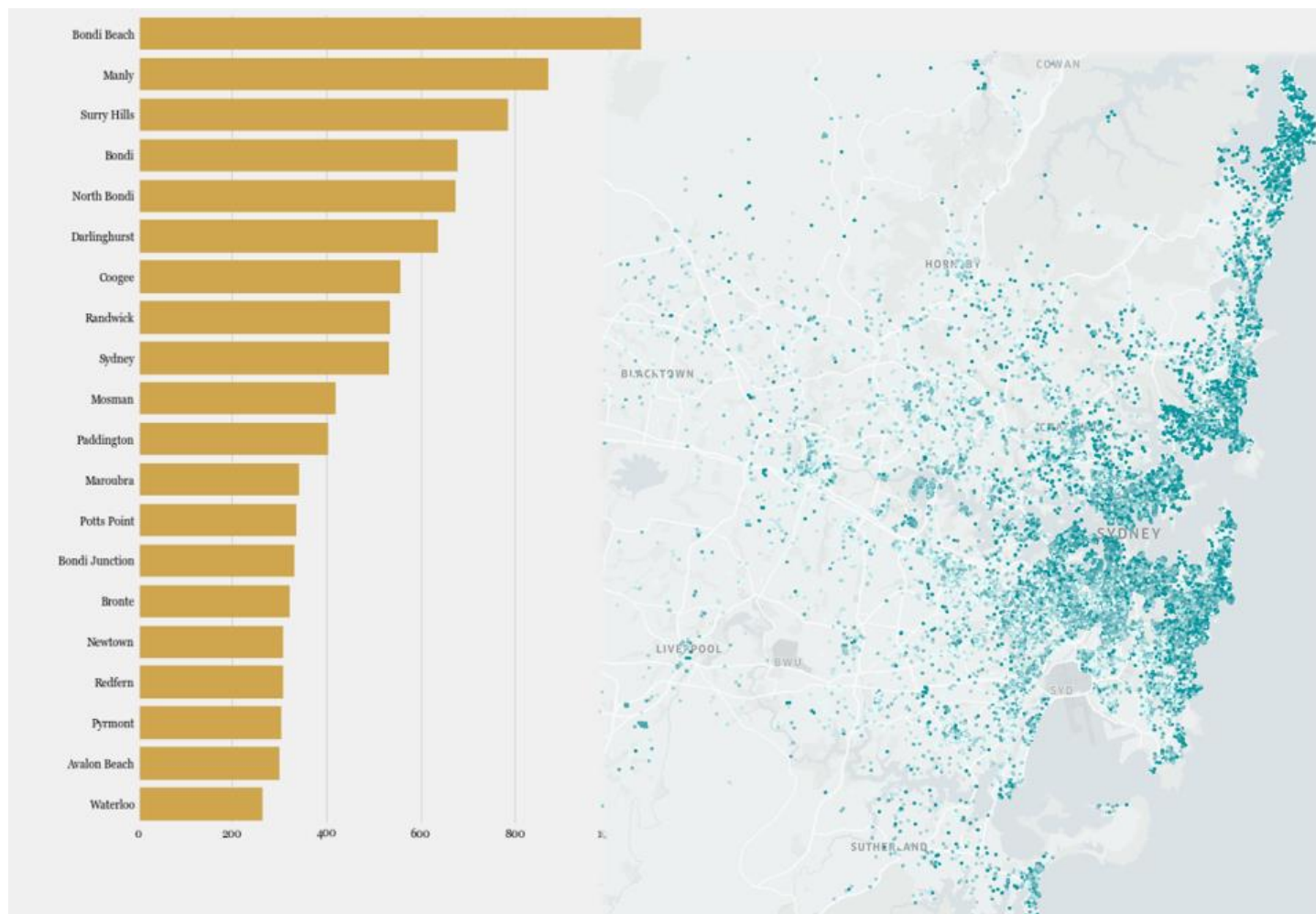


Figure 6: The Top suburbs with the highest listings. Source – Author

Prices in the dataset ranged from \$0 to \$14,999. The extreme ends of the range are due to hosts probably not using Airbnb advertised prices called 'sticker' prices correctly. The advertised prices can be set to any arbitrary amount, and these are the prices that show when dates are not entered on the site. Most notably the number of listings after the \$200 mark dropped considerably. There were only 0.5% observations above the \$1000 mark. There was also a notable drop in observations before the \$60 mark and before the \$75 mark. A little more than 0.5% of observations were below the \$60 mark and these were in some of the elite suburbs of Sydney with median high prices. The least and the maximum price marks were set at \$60 and \$1000 respectively and all observations beyond these ranges were imputed with these ranges.

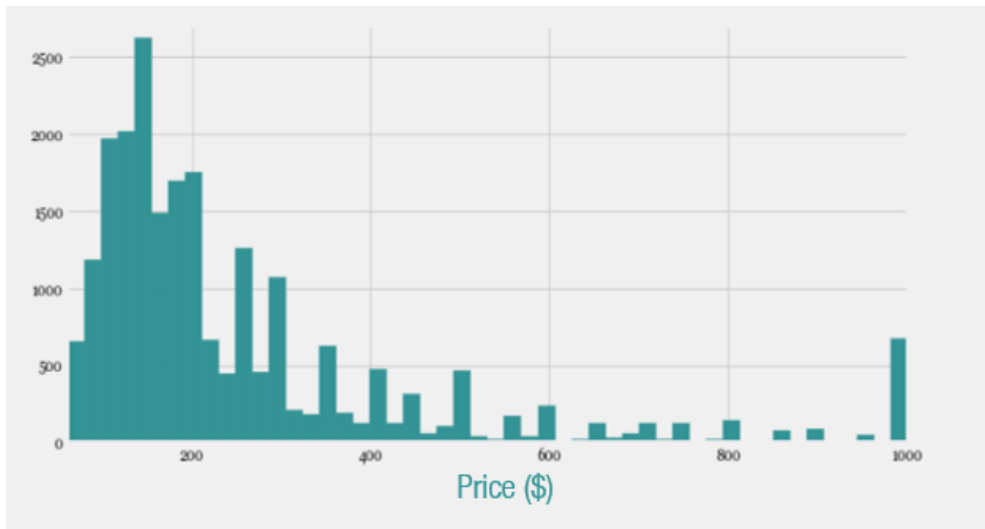


Figure 7: Price distribution range. Source – Author

Further analysis showed that apartments dominated the listings but ranged mostly from \$60 to \$ 300. Houses covered the entire spectrum of price range but featured most prominently in the higher price categories. The other notable features were security deposit and cleaning fees. Most of the distribution for security deposit was in the lower price range between \$100 - \$ 500 which showed that people were willing to pay even with the prospect of a security deposit. The cleaning fee also followed a similar trajectory with most of the listing between the range of \$50 to \$150.

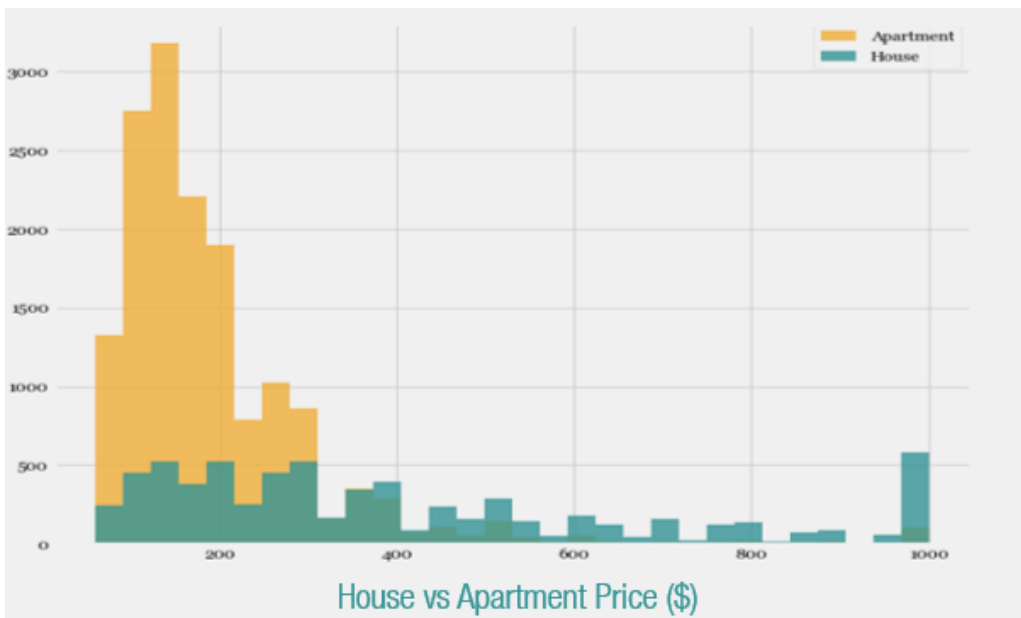


Figure 8: Price distribution range of House versus Apartments. Source – Author

The LGA feature aggregated the suburbs and price. The priciest LGA is Mosman followed by the Northern beaches, Hunters Hill, Lane Cove, and the eastern beach suburb of Woollahra. The cheapest STRs can be found in the western and southwestern LGAs of Sydney, most notably in Fairfield, Cumberland, and Blacktown. The innerwest LGA of Burwood and the northern LGA of Ku-ring-gai also had much cheaper STRs to rent. The top 5 priciest LGA's are: Mosman, Northern Beaches, Hunters Hill, Lane Cove, and Woollahra. Mosman has much lower listings than the eastern suburbs but commands higher price - usually rent out houses at the top end of the price range. Suburbs in the Northern Beaches barely touch the Top 20 listings numbers but are much pricier than the eastern beach suburbs.

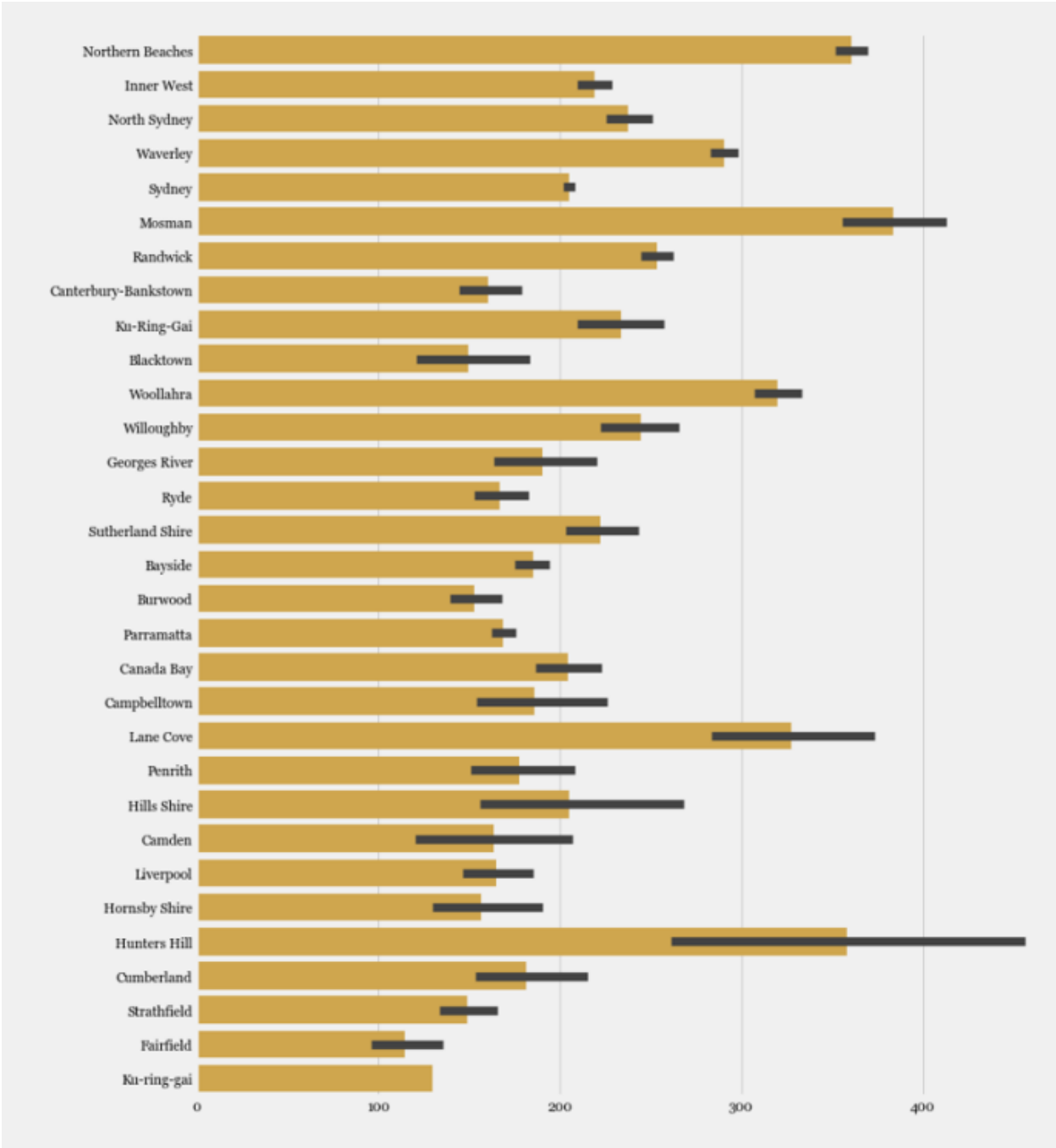


Figure 9: Price distribution within LGAs. Source – Author

In terms of size and spatial features of a listing, the greater number of bedrooms and bathrooms the higher the price. This also is true for the number of guests accommodated although the returns diminish after 12 guests. The median common set up is 2 beds in a 2-bedroom property with 1 bathroom and which accommodates 4 guests.

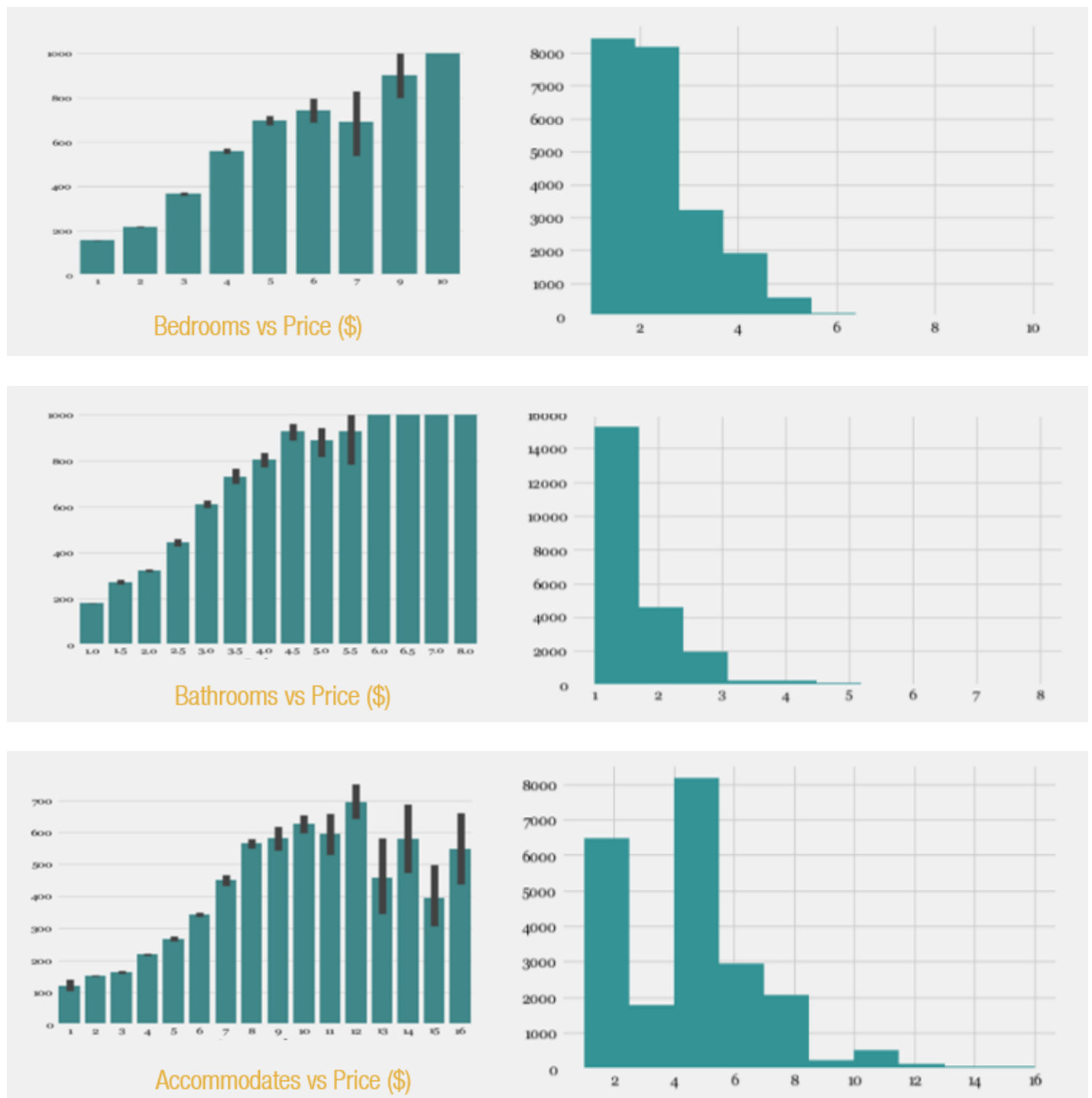


Figure 10: Distribution of Bedrooms, bathrooms and number of guests accommodated. Source – Author

## DATA MODELING

The test cases were set up to follow 3 iterations, mainly:

1. Model with all features.
2. Model with notable features selected
3. Model with the selected features were hyper tuned using Randomized CV.

All 3 iterations were modeled using four algorithms, namely Random Forest Regressor (RF), XGBoost Regressor (XGB), Super Vector Machine Regressor (SVM), and K Nearest Neighbour Regressor (KNN).

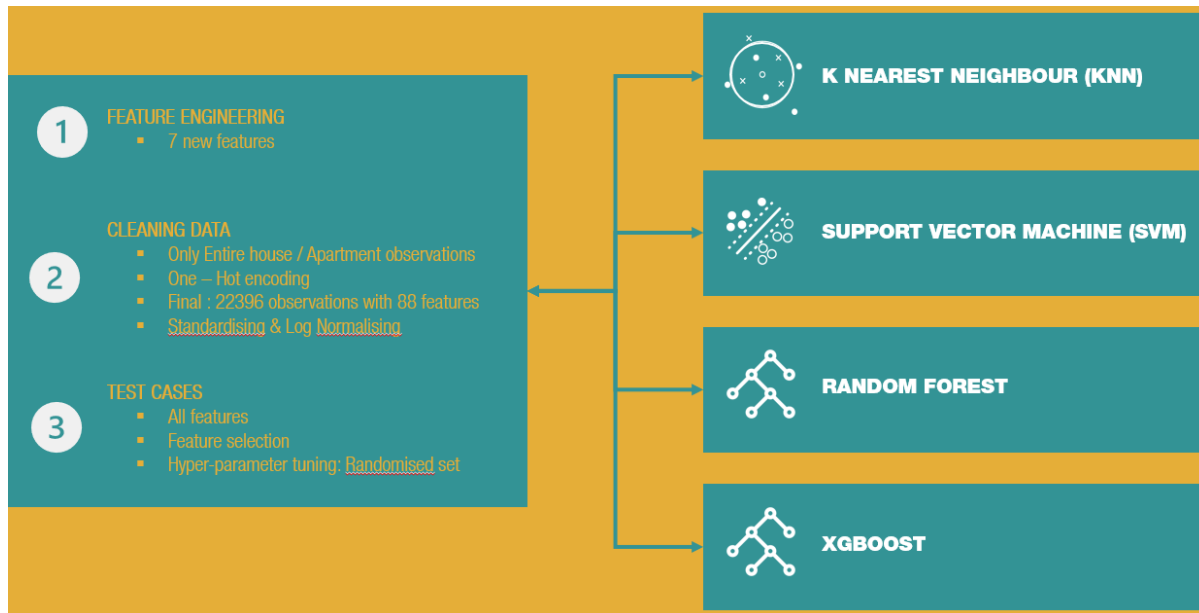


Figure 11: The Modeling Process. Source – Author

The top 9 features for the 2<sup>nd</sup> and 3<sup>rd</sup> iteration were chosen using Random Forest's feature selection process. XGBoost's feature selection was also used to compare the two. While most of the top features were similar, there were a few additional features in XGBoost (11 features). Since we plan to use the features in an application it was best to limit the features to 9 or 10 features.

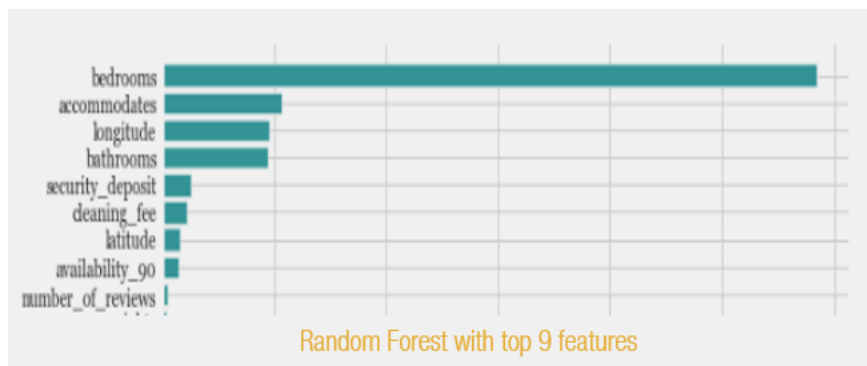



Figure 11: The selected features from Random Forest. Source – Author



It is not surprising that the most important feature is how many people the property accommodates, as that's one of the main things you would use to search for properties with in the first place. It is also not surprising that features related to location and reviews are in the top ten. It is also notable that two other fee types - cleaning, and security also make the top 9 feature list. It is likely that these are positive relationships, and that when a host sets a higher price for the nightly stay, they are also likely to set other prices high. The top 9 features for deployment will be:

1. Bedrooms
2. Accommodates
3. Longitude
4. Bathrooms
5. Security deposit
6. Cleaning fee
7. Latitude
8. Availability in 90 days
9. Number of reviews

Below are the metrics from the modeling process. The top performing model was XGBoost with all features, while the top performing model with limited features was Random Forest model that was hyper tuned with randomized CV. Since, the models trained quickly different features were added and removed to try improving the models but there were no significant differences in the metrics.

	size	train rmse	test rmse	train r2	test r2
 <b>RANDOM FOREST</b>	All features	0.38	0.39	0.65	0.63
	Imp features	0.38	0.39	0.64	0.63
	Hyper-param	0.34	0.38	0.71	0.66
 <b>XGBOOST</b>	All features	0.34	0.36	0.71	0.68
	Imp features	0.37	0.38	0.67	0.65
	Hyper-param	0.34	0.37	0.71	0.66
 <b>K NEAREST NEIGHBOUR (KNN)</b>	All features	0.39	0.42	0.63	0.58
	Imp features	0.37	0.39	0.66	0.62
	Hyper-param	0.37	0.40	0.65	0.61
 <b>SUPPORT VECTOR MACHINE (SVM)</b>	All features	0.28	0.37	0.8	0.66
	Imp features	0.37	0.38	0.67	0.66
	Hyper-param	0.37	0.38	0.66	0.64

## MODEL OUTCOME

The R2 test scores are average. This is one of those situations where most of the features in the dataset do not affect the price. Even in the best performing model, the model was only able to explain 68% of the variation in price albeit with all the features. The remaining 32% is probably made up of features that are not present in the data. It is likely that a significant proportion of this unexplained variance is due to variations in the listing photos. The photos of properties on AirBnB are very important in encouraging guests to book, and so can also be expected to have a significant impact on price-better photos (primarily better-quality properties and furnishings, but also better-quality photography) equal higher prices.



To try and improve the model the following steps should be implemented:

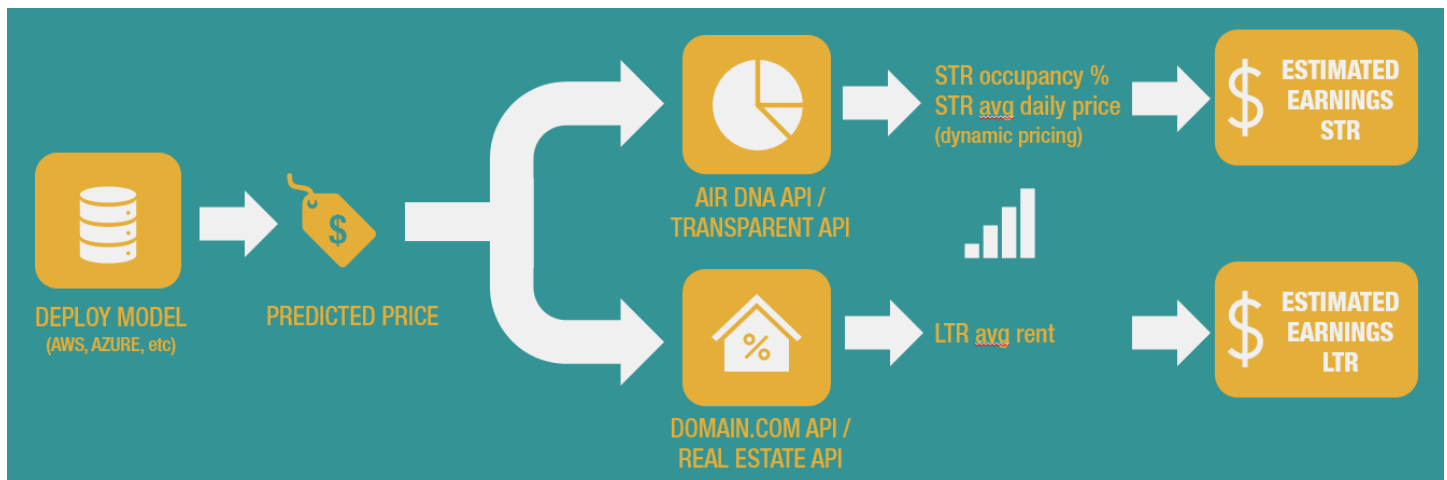
1. Find a way to incorporate image quality into the model, e.g. by using the output of a convolutional neural network (CNN) model to assess image quality as an input into the pricing model.
2. Use better quality/more accurate data which includes the actual average prices paid per night, perhaps using the APIs of AirDNA or Transparent.
3. Augment the model with natural language processing (NLP) of listing descriptions and/or reviews, e.g. for sentiment analysis or looking for keywords that could explain the value of the listing and its effect on price.
4. Tailor the model more specifically to new listings in order to help hosts set prices for new properties, by removing features that would not be known at the time - e.g. other fees, availability and reviews.

## ANSWERING THE QUESTIONS

1. Based on the dataset, the model could explain only 66% - 68% of the variation in price. The most important features were number of bedrooms, bathrooms, beds, and guests accommodated.
2. The Exploratory Data Analysis process shows that the top suburbs to invest in STRs are the northern and eastern beach suburbs. Research shows that these beaches tend to have high occupancy rates throughout the year. Price being inelastic, i.e. people will still pay with increase in price, can be set dynamically. The calendars for STRs can be scattered throughout the year based on occupancy rates and seasonality.
3. Listings in the western and southwestern suburbs would be better off using a LTR rental strategy. The same goes for the Inner West listings although suburbs near Universities can use switch between LTRs and STRs based on returns after 1-2 years.

## FUTURE END TO END SOLUTION

In addition to predicting base prices, a sequence model could be created to calculate daily rates using data on seasonality and occupancy, which would allow the creation of actual pricing software. The predictions from the output could then be hooked up to the AirDNA API for the occupancy rates and ADRs across different suburbs to calculate the estimated earnings for a STR. In parallel, using the Domain.com API of average LTR rents in different suburbs can be used to calculate the estimated earnings from a LTR strategy. The stakeholder can then take the decision as to which rental strategy would best suit his purpose.



## REFERENCES

- **Code access**

The code, notebooks, documents, data files, and presentation to the project can be found at the following site.

<https://github.com/wendydsa/Capstone-Estimating-Annual-Earnings-in-Sydneys-Housing-Rental-Market.git>

- **Resources Used**

The following libraries and algorithms were used in the project:

1. **Libraries:** Pandas, Numpy, Matplotlib, Seaborn, Scipy, Statsmodel, Math, Datetime, and Re.
2. **SkLearn Preprocessing Libraries:** Standard Scalar, Train test split, Randomised CV, Mean squared error, and R2 Score
3. **Algorithms:** Random Forest Regressor, XGBoost Regressor, SVM Regressor (SVR), and KNeighbors Regressor

- **Appendix**

Features	Data Type	Description
id	integer	Airbnb's unique identifier for the listing
listing_url	string	URL of listing
scrape_id	integer	Inside Airbnb "Scrape" this was part of
last_scraped	datetime	UTC. The date and time this listing was "scraped".
name	string	Name of the listing
summary	string	Summary description of the listing
space	string	Description of the space for rent
description	string	Detailed description of the listing
experiences_offered	string	Description of the experiences offered
neighborhood_overview	string	Host's description of the neighbourhood
notes	string	Notes for the listing.
transit	string	Notes for the listing
access	string	Access notes
interaction	string	Notes on host interaction
house_rules	string	Notes on house rules
thumbnail_url	string	URL to the Airbnb hosted thumbnail image for the listing
medium_url	string	URL to the Airbnb hosted medium sized image for the listing
picture_url	string	URL to the Airbnb hosted regular sized image for the listing
xl_picture_url	string	URL to the Airbnb hosted XL sized image for the listing
host_id	integer	Airbnb's unique identifier for the host/user
host_url	string	The Airbnb page for the host
host_name	string	Name of the host. Usually just the first name(s).
host_since	date	The date the host/user was created.
host_location	string	The host's self-reported location
host_about	string	Description about the host
host_response_time	category	How fast does the host respond
host_response_rate	integer	Rate at which the host responds

Features	Data Type	Description
host_acceptance_rate	numeric	That rate at which a host accepts booking requests.
host_is_superhost	boolean	[t=true; f=false]
host_thumbnail_url	string	URL to the host profile
host_picture_url	string	URL to the host picture
host_neighbourhood	string	Location of the host
host_listings_count	string	The number of listings the host has
host_total_listings_count	string	The number of listings the host has
host_verifications	string	Channels to contact host
host_has_profile_pic	boolean	[t=true; f=false]
host_identity_verified	boolean	[t=true; f=false]
street	category	Address of the listing
neighbourhood	category	Neighbourhood of the listing
neighbourhood_cleansed	category	The neighbourhood as geocoded using the latitude and longitude against neighborhoods as defined by open or public digital shapefiles.
neighbourhood_group	category	The neighbourhood group as geocoded using the latitude and longitude against neighborhoods as defined by open or public digital shapefiles.
city	category	Listing city
state	category	Listing state
zipcode	category	Listing zipcode
market	category	Listing market
smart_location	category	Location of listing
country_code	category	Listing country code
country	category	Listing country
latitude	numeric	Uses the World Geodetic System (WGS84) projection for latitude and longitude.
longitude	numeric	Uses the World Geodetic System (WGS84) projection for latitude and longitude.
is_location_exact	boolean	Possibly related to the host's choice for displaying the listing's location. See references."
property_type	category	Self-selected property type.
room_type	category	[Entire home/apt Private room Shared room Hotel]

Features	Data Type	Description
accommodates	integer	The maximum capacity of the listing
bathrooms	numeric	The number of bathrooms in the listing
bedrooms	integer	The number of bedrooms
beds	integer	The number of bed(s)
bed_type	string	The type of bed(s)
amenities	string	Amenities provided by listing
square_feet	numeric	Area of the listing
price	currency	daily price in local currency
weekly_price	currency	weekly price in local currency
monthly_price	currency	monthly price in local currency
security_deposit	currency	security deposit in local currency
cleaning_fee	currency	cleaning fee in local currency
guests_included	integer	number of guests included in the price
extra_people	currency	amount in local currency for additional guests
minimum_nights	integer	minimum number of nights stay for the listing (calendar rules may be different)
maximum_nights	integer	maximum number of nights stay for the listing (calendar rules may be different)
calendar_updated	date	date the host calendar was last updated
has_availability	boolean	[t=true; f=false]. Whether listing is available
availability_30	integer	availability_x. The availability of the listing x days in the future as determined by the calendar. Note a listing may not be available because it has been booked by a guest or blocked by the host.
availability_60	integer	availability_x. The availability of the listing x days in the future as determined by the calendar. Note a listing may not be available because it has been booked by a guest or blocked by the host.
availability_90	integer	availability_x. The availability of the listing x days in the future as determined by the calendar. Note a listing may not be available because it has been booked by a guest or blocked by the host.
availability_365	integer	availability_x. The availability of the listing x days in the future as determined by the calendar. Note a listing may not be available because it has been booked by a guest or blocked by the host.
calendar_last_scraped	date	Scrape date of calendar
number_of_reviews	integer	The number of reviews the listing has

Features	Data Type	Description
first_review	date	The date of the first/oldest review
last_review	date	The date of the last/newest review
review_scores_rating	integer	Scores rating out of 100
review_scores_accuracy	integer	Accuracy score rating out of 10
review_scores_cleanliness	integer	Cleanliness score rating out of 10
review_scores_checkin	integer	Checkin score rating out of 10
review_scores_communication	integer	Communication score rating out of 10
review_scores_location	integer	Location score rating out of 10
review_scores_value	integer	Value score rating out of 10
requires_license	boolean	[t=true; f=false]. Whether the listing/jurisdiction requires a license
license	string	The licence/permit/registration number
jurisdiction_names	string	Legal jurisdiction
instant_bookable	boolean	[t=true; f=false]. Whether the guest can automatically book the listing without the host requiring to accept their booking request. An indicator of a commercial listing.
is_business_travel_ready	boolean	[t=true; f=false]
cancellation_policy	category	The cancellation policy of the host
require_guest_profile_picture	boolean	[t=true; f=false]
require_guest_phone_verification	boolean	[t=true; f=false]
calculated_host_listings_count	integer	The number of listings the host has in the current scrape, in the city/region geography.
reviews_per_month	numeric	The number of reviews the listing has over the lifetime of the listing



## SOURCES

---

- <sup>1</sup> [https://www.bluewealth.com.au/the\\_research\\_blog/australian-residential-property-unique-asset-class/](https://www.bluewealth.com.au/the_research_blog/australian-residential-property-unique-asset-class/)
- <sup>2</sup> <https://www.mortgagebusiness.com.au/breaking-news/16123-housing-market-surpasses-9-trillion-mark>
- <sup>3</sup> <https://propertyupdate.com.au/property-investment-sydney/#sydney-house-prices>
- <sup>4</sup> <https://www.news.com.au/finance/real-estate/renting/australian-rental-market-worst-in-13-years-new-report-finds/news-story/8d1913c5527d23fcc28263e2db542dbb>
- <sup>5</sup> <https://www.news.com.au/finance/real-estate/renting/australian-rental-market-worst-in-13-years-new-report-finds/news-story/8d1913c5527d23fcc28263e2db542dbb>
- <sup>6</sup> <https://www.domain.com.au/news/soaring-rental-prices-creating-housing-crisis-in-regional-nsw-1070834/>
- <sup>7</sup> <https://managecasa.com/articles/australia-housing-market/>
- <sup>8</sup> <https://www.telegraph.co.uk/property/landlord-guide/pros-and-cons-of-short-term-letting/amp/>
- <sup>9</sup> <https://propertyupdate.com.au/air-bnb-phenomenon-australias-short-term-rental-report/#core-findings>
- <sup>10</sup> <https://www.rent.com.au/blog/airbnbs-australia>
- <sup>11</sup> <https://www.rent.com.au/blog/airbnbs-australia>
- <sup>12</sup> <https://sheltersnsw.org.au/short-term-rentals-and-local-housing-markets/>
- <sup>13</sup> <https://www.savings.com.au/news/nsw-airbnb-restrictions-november>
- <sup>14</sup> <https://www.fairtrading.nsw.gov.au/about-fair-trading/legislation-and-publications/changes-to-legislation/laws-for-short-term-rental-accommodation>
- <sup>15</sup> <https://propertyupdate.com.au/air-bnb-phenomenon-australias-short-term-rental-report/#core-findings>
- <sup>16</sup> [https://www.internationalhousingassociation.org/fileUpload\\_details.aspx?contentTypeID=3&contentID=254952&subContentID=695501&channelID=38488](https://www.internationalhousingassociation.org/fileUpload_details.aspx?contentTypeID=3&contentID=254952&subContentID=695501&channelID=38488)
- <sup>17</sup> <https://seetransparent.com/blog/whats-up-down-under-australia-short-term-rental-report/>