1. Introduction

This assignment is to do with exploring Perth's residential property market and predicting house prices from the Perth House Prices dataset. In this project, we aim to build predictive models for Perth house prices using a dataset containing property attributes and their corresponding sale prices. Our approach includes Exploratory Data Analysis (EDA), data cleaning and preprocessing (such as handling missing values), feature importance estimation via a Random Forest model, and the development of neural network models. We compare models built using a reduced set of features (identified as most important by Random Forest) against models employing all available features.

2. Data Description

Data Source & Structure:
The dataset consists of 33,656 properties sold in Perth, each with details such as address, suburb, price, bedrooms, bathrooms, garage spaces, land and floor areas, build year, and distances to various amenities (Central Business District, nearest train station, nearest school), along with the school's ranking, and geospatial coordinates (latitude and longitude).

Key Columns:

- **price (target)**: The sale price of the property.
- **bedrooms, bathrooms, garage**: Quantitative room and parking attributes.
- **land_area, floor_area**: Property size indicators.
- **build_year**: Year the property was built.
- **cbd_dist**: Distance to the Perth CBD in metres.
- **nearest_stn_dist**: Distance to the nearest train station.
- **nearest_sch_dist, nearest_sch_rank**: Distance and academic rank of the nearest school.
- **latitude, longitude**: Geographic coordinates of the property.
- **postcode, suburb**: Location identifiers.
- **address**: Unique identifier of the property.

Data Types & Missing Values:

- The dataset comprises 19 columns, with a mix of numerical (int, float) and categorical (object) fields.
- Notable missing data:
    - **garage**: 2,478 missing values.
    - **build_year**: 3,155 missing values.
    - **nearest_sch_rank**: 10,952 missing values.

Statistical Summary:

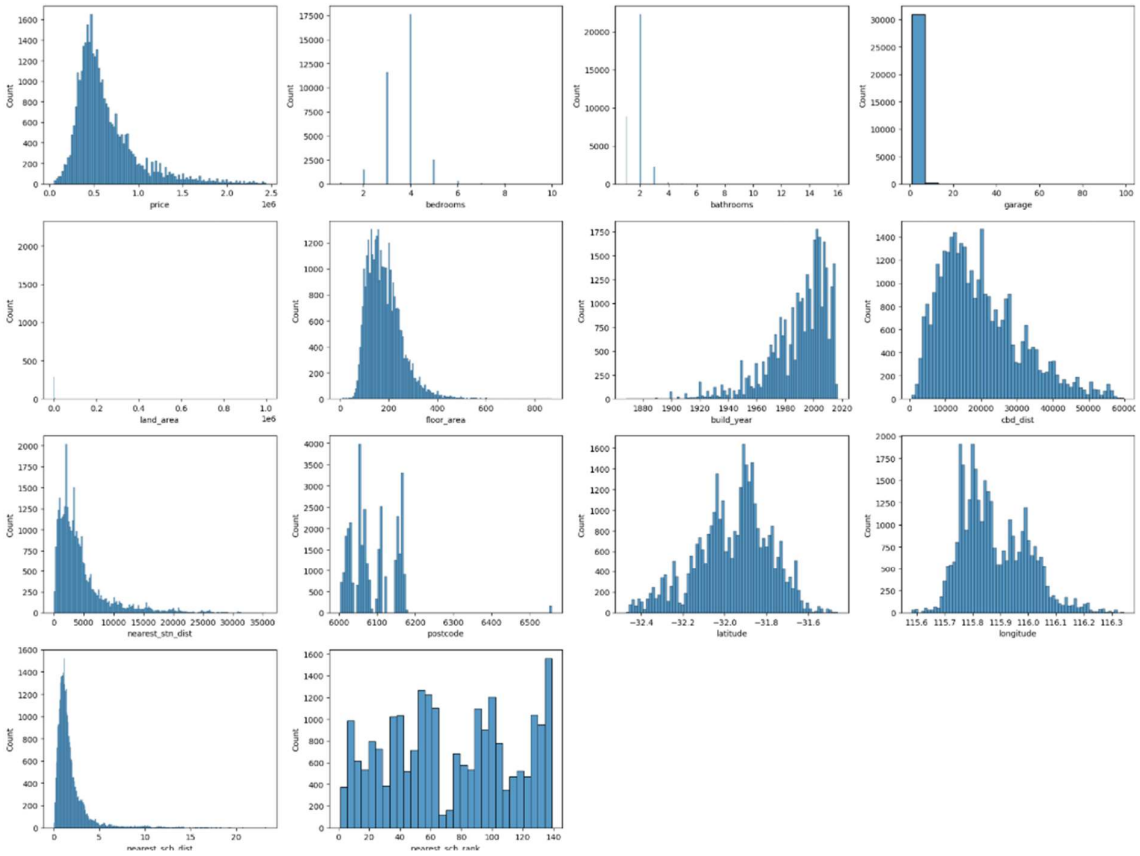
- **price**: Mean 637,072 AUD, std 355,826, range 51,000 to 2,440,000 AUD.
- **bedrooms**: Typically between 3 to 4 bedrooms.
- **land_area**: Highly skewed, ranging widely up to 999,999 m² (999,999 likely erroneous data), but with a median of 682 m².
- **floor_area**: Median 172 m².

- **cbd_dist**: Median 17,500 m, indicating suburban spread.
- **nearest_sch_dist**: Median 1 km, with majority of properties relatively close to at least one school.

3. Exploratory Data Analysis (EDA)

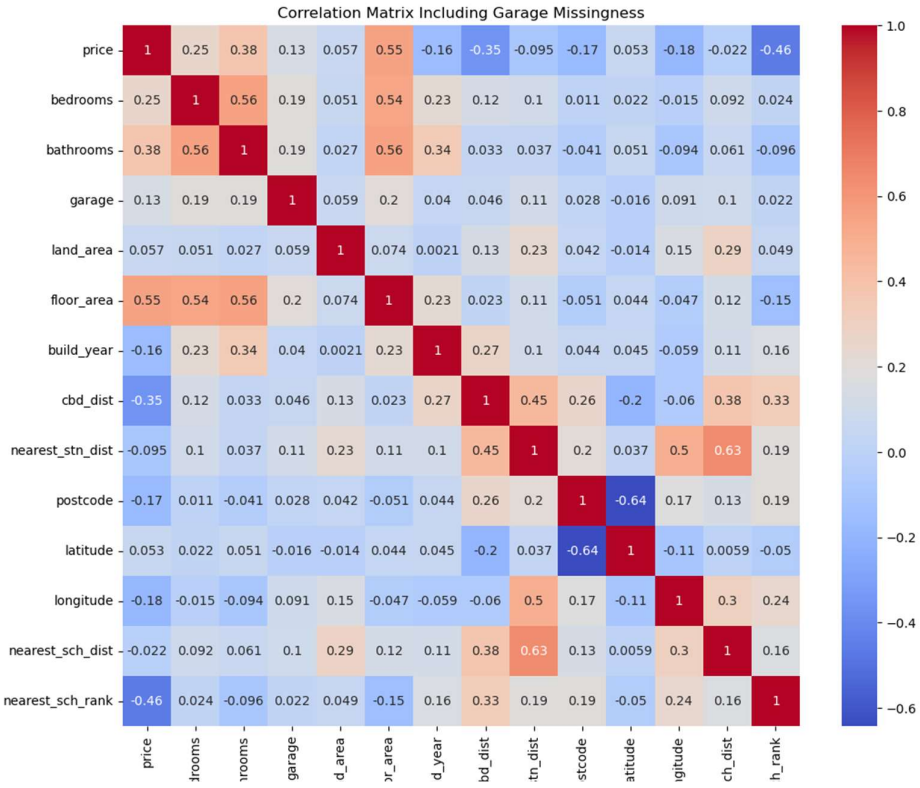
Univariate Analysis:

- **Price Distribution:** Slightly right-skewed with most properties under 1 million AUD.
- **Rooms and Facilities:** Bedrooms and bathrooms generally cluster around typical family-sized homes (3-4 bedrooms, 1-2 bathrooms). Garage spaces are mostly 2, though outliers exist.
- **Land and Floor Areas:** Land area is highly skewed due to some very large lots. Floor area tends to follow a more normal distribution centred around ~180 m².
- **Build Year:** Properties span a wide historical range with a median year around 1995-2000.
- **Geospatial Features:** Latitude and longitude distributions show concentration within Perth's urban centre.



Bivariate Analysis:

- **Price vs. Bedrooms/Bathrooms:** Increasing bedrooms and bathrooms correlates modestly with higher prices, though with diminishing returns.
- **Price vs. Floor_Area:** A stronger positive correlation (around 0.55) indicates that larger floor areas correspond to higher property values.
- **Price vs. cbd_dist:** Negative correlation 0.35), suggesting properties closer to the CBD are more expensive.
- **Price vs. nearest_sch_rank:** Negative correlation 0.46) indicates properties near highly ranked schools (low rank number) tend to command higher prices.
- **Price vs. garage & build_year:** Weaker but positive correlation with floor_area and negative or mild negative correlations with build_year (indicating newer homes do not necessarily always mean higher prices, possibly due to suburban locations).



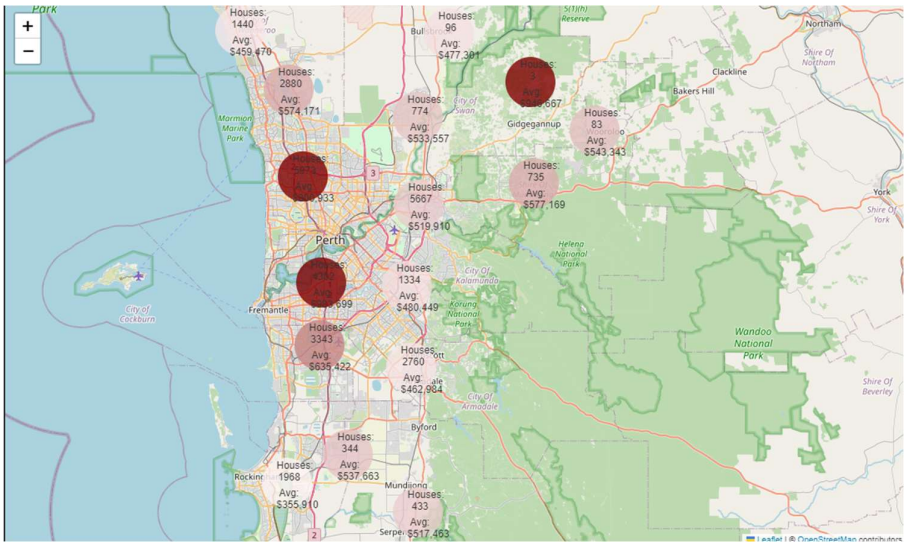
Strongest correlations with price:

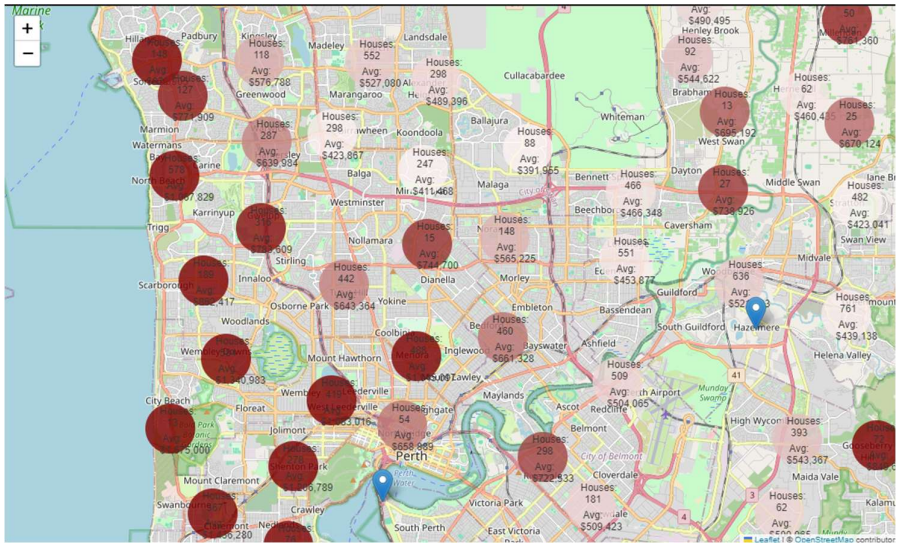
- **floor_area** (0.55)
- **bathrooms** (0.38) and **bedrooms** (0.25)
- **nearest_sch_rank** (-0.46) and **cbd_dist** (-0.35)

These factors suggest that both internal property features (floor area, number of bathrooms) and external location factors (distance to CBD, quality of nearest school) influence pricing.

We can also visualise where the properties are on a map with prices and land area on the interactive Folium map in the Jupyter Notebook.

Here are some screenshots at different zoom degrees:





4. Data Cleaning and Preprocessing

In this section, we will be examining ways to augment, clean and preprocess our dataset to prepare it for model training.

In terms of location-based variables, we have address, suburb and postcode, with latitude and longitude being geographical coordinates.

By checking the unique count of values per categorical variable, we can have a rough idea of what additional dimensions will be added during encoding.

Unique Values:

- address: 33,566
- suburb: 321

It is obvious that directly operating on address will lead to excessive dimensions, whilst using suburb will also increase our dimensions by 321 with one-hot encoding.

To try and reduce the number of dimensions added but still retain some form of location context to the dataset, I would like to introduce the concept of the SA2 Region.

Introducing the SA2 Region Concept

In Australia, the Australian Bureau of Statistics (ABS) has developed a standardised geographical framework known as the Australian Statistical Geography Standard (ASGS). Within this framework, the country is divided into hierarchical units for the purpose of collecting and analysing statistical information. One of the key levels in this hierarchy is the Statistical Area Level 2 (SA2).

What is an SA2 Region?

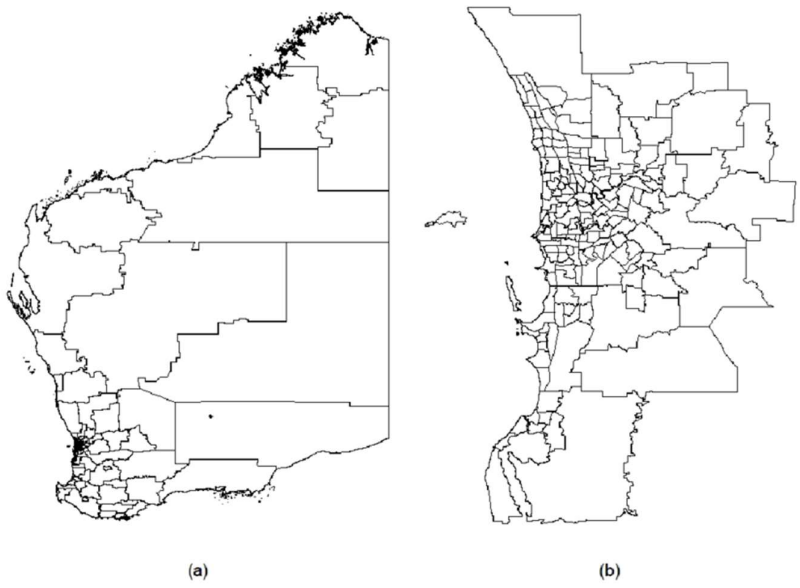
- **Definition:** A Statistical Area Level 2 is a medium-sized geographical region that generally represents a community with common social and economic characteristics. SA2 boundaries are designed to be stable over time and are often used as the building blocks for other ABS-defined regions, such as SA3s and SA4s (larger regions comprising of large urban centres).
- **Size and Coverage:** Each SA2 typically encompasses a population of around 3,000 to 25,000 individuals, depending on urban or regional contexts. In metropolitan areas, SA2s often reflect local neighbourhoods or clusters of suburbs, whereas in rural or remote areas, an SA2 might cover a much larger geographic area with lower population density.
- **Stability and Consistency:** Unlike many other geographical delineations (e.g., suburbs and postcodes), SA2 regions are reviewed and updated less frequently, ensuring that they remain as stable reference units over time

Why Use SA2 Instead of Suburbs or Postcodes Alone?

1. **Standardisation and Stability:**
   Suburbs and postcodes are often defined for administrative and postal purposes and can be subject to changes over time. Postcode boundaries may be altered due to changes in postal services or the introduction of new delivery routes. Suburbs, while familiar to residents, may not be defined in a standardised manner across different local government areas. SA2s, on the other hand, are specifically designed for statistical analysis, ensuring stable, well-documented, and methodologically consistent boundaries.

Better Population and Socio-Economic Representation:
SA2 regions are constructed with the aim of bringing together communities with relatively homogeneous social and economic characteristics. By grouping properties into SA2s, we leverage a geographical unit that is likely to capture meaningful patterns, such as local school catchment areas, public amenities, and overall neighbourhood quality rather than relying on arbitrary lines drawn for non-statistical purposes.

2. **Dimensionality Reduction and Neighbourhood Context:**
   Using individual suburbs or postcodes as categorical variables in a predictive model can lead to an explosion in the number of features, especially in large metropolitan areas with hundreds of suburbs. This introduces sparsity and complexity that can hamper model performance. By aggregating data to the SA2 level, we can effectively reduce the number of unique categorical units, simplifying the dataset. Additionally, since SA2s are often formed based on socio-demographic similarities, they naturally serve as a higher-quality grouping mechanism that captures the essence of a neighbourhood and how it can influence property prices more effectively than a mere administrative or postal boundary.



SA2 map of (a) Western Australia (b) metropolitan Perth

Due to the locking of statistical region data behind paywall, I will be using the open-source database that contains postcode and corresponding SA data via this website (https://www.matthewproctor.com/australian_postcodes).

After merging the SA2 data with our dataset, the number of unique SA2 regions stands at 83, which is a significant reduction from suburbs in our dataset (321) whilst still retaining geographical context.

An additional benefit of adding the SA2 contextual information is we can utilise it to impute missing values in our dataset for the garage and build_year features, instead of using broad imputes like the median count across the whole dataset.

Data Transformations:

We should do log transformation on features that are heavily skewed, such as:

- land_area
- nearest_stn_dist
- nearest_sch_dist
- cbd_dist

Also, we can split the date_sold feature into 2 features, year sold and month sold.

Missing Data Imputation:

- **garage:** About 7% (2478 values). Imputed with the median garage count at the SA2 level. For example, in "Alkimos - Eglinton," the average garage count was 2, so missing values within that SA2 were set to 2.
- **build_year:** About 9% missing (3155 values). Imputed similarly with the median build year by SA2. Areas like "Baldivis - South" showed an average build year of 2011, so missing values there were filled accordingly.
- **nearest_sch_rank:** About 32.5% missing (10952 values). Schools that are eligible to be ranked are ATAR-applicable (Australian Tertiary Admission Rank) schools where students can study WACE (Western Australia Certificate of Education) courses. These WACE courses are designed to prepare students for university entrance and contribute to a student's ATAR score. As such, schools that are **unranked** are typically schools that are all or a combination of:
    - do not offer sufficient WACE courses that contribute to their ATAR score
    - lack qualified teaching staff
    - do not meet the curriculum and assessment requirements set by SCSA (School Curriculum and Standards Authority)

  Missing values for nearest_sch_rank can be imputed with a "bad rank" value of 999 (1 is best rank) as the nearest school is not eligible to be ATAR-applicable.

Outlier Detection & Removal:

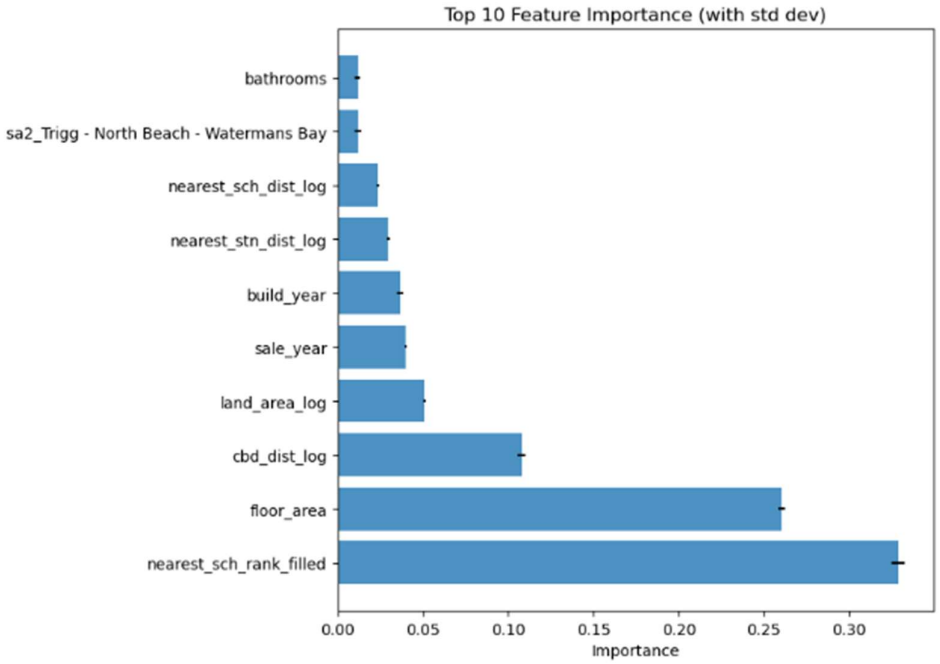
Another benefit of having SA2 in our dataset is we can use this information to find price and land_area outliers more granularly.

We can use a multivariate Z-score threshold within each SA2 region to detect and remove price and land_area outliers. For instance, certain properties with disproportionately large/erroneous land areas or anomalous prices given their SA2 neighbourhood were removed. This step can help to stabilise model training by reducing the impact of extreme values unlikely to generalise. A more conservative threshold of 3.5 was used to only flag out extreme outliers as some real estate might have legitimate high-value properties that set it aside from others. Nonetheless, this threshold can be adjusted as needed for further experimentation.

5. Feature Set

One way to experiment with the feature set is to try and identify the most influential predictors by using a Random Forest model after data cleaning. Preliminary results indicate:

- **nearest_sch_rank_filled , floor_area, cbd_dist_log** are among the top predictors of price.
- **land_area_log**, **sale_year, build_year** show weaker relationships



We can use these insights for the creation of two different modelling scenarios:

1. **Reduced Feature Set Model:** Using only top 10 predictive features from Random Forest
2. **Full Feature Set Model:** Using all available features.

We will use the same train validation test split across all model training to prevent any data leakage.

6. Network Architectures

Base Architecture

The base architecture (BaseNet) implements a straightforward feedforward neural network with three layers. Its progressive narrowing structure (input→64→32→1) serves as an effective baseline for price prediction. The architecture incorporates ReLU activation functions and modest dropout regularisation (0.2). This architecture's simplicity allows for quick training and serves as a benchmark for more complex models.

Deep Architecture

The deep architecture (DeepNet) extends the basic model with additional capacity and regularisation techniques. Its structure (input→256→128→64→1) includes batch normalization after each hidden layer, coupled with dropout implementations. This design allows the network to capture intricate relationships between housing features while maintaining training stability through batch normalisation.

Residual Architecture

The residual architecture (ResidualNet) introduces skip connections, allowing the network to learn both direct and transformed feature relationships. This design is relevant for housing price prediction, where some features (such as location or land size) might directly influence price while others require more complex transformation. The architecture maintains a consistent width (128 neurons) through its main processing layers before final prediction, allowing it to preserve important feature information throughout the network.

Wide Architecture

The wide architecture (WideNet) emphasises breadth over depth, utilising two wide layers of 512 neurons each. This design allows for extensive feature interaction within each layer, making it suitable for capturing complex relationships between multiple housing characteristics. The increased dropout rate (0.3) provides necessary regularisation for the larger parameter space, helping prevent overfitting to historical price patterns.

Pyramid Architecture

The pyramid architecture (PyramidNet) implements a systematically narrowing structure (256→128→64→32→1) with LeakyReLU activations. This design creates a controlled feature distillation process, potentially beneficial for gradually extracting relevant pricing factors from raw housing features. The architecture's use of LeakyReLU activation functions helps maintain stable gradients throughout training.

Dense Architecture

The dense architecture (DenseNet) implements connections from each layer to all subsequent layers. This design allows the network to simultaneously consider features at multiple levels of abstraction, potentially capturing complex interactions between housing characteristics that influence price.

7. Optimisation Strategy

The training process employs two optimiser configurations (Adam and AdamW) along with two learning rate scheduling approaches (Plateau and Cosine). This combination provides flexibility in handling different aspects of the training process:

- Adam optimiser provides adaptive learning rates for each parameter, valuable for handling features of different scales in housing data
- AdamW adds weight decay regularisation, helping prevent overfitting to historical price patterns
- Plateau scheduler adapts learning rates based on validation performance, useful for finding optimal parameters without overfitting
- Cosine scheduler provides structured learning rate variation, potentially helping models escape local optima

8. Test Results

Reduced Feature Set Training:

Model Performance (Sorted by Test R²):

model_name	test_rmse	test_r2	test_mae
deep_adam_plateau	$157,044.25	0.799680	$100,390.27
pyramid_adam_cosine	$159,678.22	0.792904	$101,649.70
pyramid_adamw_plateau	$159,743.84	0.792733	$100,781.60
pyramid_adam_plateau	$160,970.83	0.789537	$101,925.57
wide_adamw_cosine	$161,504.89	0.788138	$102,395.38
deep_adamw_cosine	$161,529.27	0.788074	$103,095.77
pyramid_adamw_cosine	$161,693.88	0.787642	$103,536.81
residual_adamw_cosine	$161,769.53	0.787443	$104,096.36
wide_adam_plateau	$161,829.41	0.787286	$103,354.99
residual_adamw_plateau	$161,950.95	0.786966	$104,551.22
wide_adamw_plateau	$162,008.78	0.786814	$102,310.98
wide_adam_cosine	$162,852.34	0.784588	$103,443.38
residual_adam_cosine	$163,773.56	0.782144	$105,580.84
dense_adamw_plateau	$164,859.72	0.779245	$105,902.84
deep_adam_cosine	$164,956.03	0.778987	$107,500.45
deep_adamw_plateau	$164,962.52	0.778970	$105,185.57
residual_adam_plateau	$165,090.06	0.778628	$106,111.38
dense_adam_cosine	$167,455.64	0.772238	$107,554.51
base_adam_cosine	$168,084.56	0.770524	$110,369.84
base_adam_plateau	$169,990.72	0.765290	$111,009.49
base_adamw_plateau	$170,157.42	0.764829	$111,770.51
dense_adam_plateau	$170,335.56	0.764337	$109,102.88
dense_adamw_cosine	$171,069.27	0.762302	$110,260.29
base_adamw_cosine	$176,455.08	0.747100	$117,825.58

Full Feature Set Training:

Model Performance (Sorted by Test R²):

model_name	test_rmse	test_r2	test_mae
dense_adamw_plateau	$136,158.59	0.849419	$85,274.26
dense_adam_plateau	$136,455.75	0.848761	$85,064.29
residual_adam_plateau	$136,674.86	0.848275	$86,018.03
residual_adamw_plateau	$137,820.67	0.845720	$86,711.34
residual_adam_cosine	$138,365.36	0.844498	$88,478.70
pyramid_adamw_plateau	$138,691.45	0.843764	$86,608.05
pyramid_adam_plateau	$138,913.88	0.843263	$86,665.94
residual_adamw_cosine	$139,370.88	0.842230	$88,842.00
deep_adamw_plateau	$139,836.58	0.841174	$88,709.95
deep_adamw_cosine	$139,844.55	0.841156	$86,717.10
wide_adam_plateau	$139,902.66	0.841023	$86,065.15
wide_adamw_plateau	$140,395.52	0.839901	$85,676.13
deep_adam_plateau	$140,937.98	0.838662	$87,434.70
wide_adam_cosine	$141,661.53	0.837001	$88,212.55
deep_adam_cosine	$142,206.50	0.835745	$92,723.38
base_adamw_plateau	$142,585.94	0.834867	$91,550.20
dense_adam_cosine	$142,658.22	0.834699	$90,130.09
pyramid_adamw_cosine	$143,689.16	0.832301	$89,348.66
base_adam_plateau	$143,930.58	0.831738	$92,414.07
base_adamw_cosine	$144,201.34	0.831104	$91,676.19
dense_adamw_cosine	$144,359.56	0.830733	$91,418.02
pyramid_adam_cosine	$145,101.02	0.828990	$89,974.13
wide_adamw_cosine	$146,446.81	0.825803	$93,023.77
base_adam_cosine	$147,260.09	0.823863	$91,876.83

Key findings from the model performance tables:

The models trained on top 10 selection of features selected by our Random Forest model underperformed the models trained with all features, suggesting that there are intrinsic relations and interactions amongst the wider set of features.

We will examine the models trained on the full feature set.

1. Best performing model: dense_adamw_plateau
    - RMSE: $136,158.59 (root mean squared error)
    - R^2: 0.849419 (84.9% of variance explained)
    - MAE: $85,274.26 (average absolute prediction error)
2. Clear patterns in model architecture performance:
    - Dense and residual architectures perform best
    - Plateau learning rate schedules generally outperform cosine

The R^2 value of 0.849 indicates the model is performing quite well, explaining about 85% of the variance in house prices. This suggests the model has captured most of the important patterns in the data.

However, looking at the error metrics:

- The RMSE of $136,159 shows there are some significant deviations in individual predictions
- The MAE of $85,274 indicates that, on average, the predictions are off by about $85K

The fact that RMSE is notably higher than MAE suggests there are some outlier predictions where the model makes larger errors, since RMSE penalises large errors more heavily.

Examining worst test set predictions:

Worst Test Set Predictions:

	Original Index	Actual Price	Predicted Price	Absolute Error	Percentage Error	bedrooms	bathrooms	floor_area
0	15274	2,400,000.00	825,059.50	1,574,940.50	65.62	1	1	120
1	14300	2,400,000.00	1,060,913.25	1,339,086.75	55.80	3	2	188
2	330	1,980,000.00	857,840.25	1,122,159.75	56.67	2	2	114
3	3735	2,360,000.00	1,394,902.00	965,098.00	40.89	4	3	285
4	20914	1,790,000.00	844,406.88	945,593.12	52.83	4	2	130

Index 15274:

address	274 Pinjar Road
suburb	Mariginiup
price	2400000
bedrooms	1
bathrooms	1
garage	1
land_area	60341
floor_area	120
build_year	1935
cbd_dist	26400
nearest_stn	Currambine Station
nearest_stn_dist	5500
date_sold	2014-11-01 00:00:00
postcode	6078
latitude	-32
longitude	116
nearest_sch	JOSEPH BANKS SECONDARY COLLEGE
nearest_sch_dist	3
nearest_sch_rank	92
sa2_name_2021	Carabooda - Pinjar
nearest_sch_rank_filled	92
land_area_log	11
nearest_stn_dist_log	9
nearest_sch_dist_log	1
cbd_dist_log	10
sale_year	2014
sale_month	11

This property is a floral farmland (https://www.realestate.com.au/property/274-pinjar-rd-mariginiup-wa-6078/), suggesting that the built-up areas (floor_area, bedrooms, bathrooms, garage) are not actually for residential, but perhaps just amenities for a tender during work hours. The large land area can now be understood as the area required for the farmland. Our models cannot predict well for such outlier cases currently.

Index 14300:

address	25/1 Corkhill Street
suburb	North Fremantle
price	2400000
bedrooms	3
bathrooms	2
garage	2
land_area	295
floor_area	188
build_year	2002
cbd_dist	12900
nearest_stn	North Fremantle Station
nearest_stn_dist	778
date_sold	2017-02-01 00:00:00
postcode	6159
latitude	-32
longitude	116
nearest_sch	JOHN CURTIN COLLEGE OF THE ARTS
nearest_sch_dist	2
nearest_sch_rank	25
sa2_name_2021	Fremantle
nearest_sch_rank_filled	25
land_area_log	6
nearest_stn_dist_log	7
nearest_sch_dist_log	1
cbd_dist_log	9
sale_year	2017
sale_month	2

This property is a luxury penthouse (https://www.realestate.com.au/sold/property-house-wa-north+fremantle-124734854) and is situated in SA2 Fremantle. Its price is accurate based on the real estate website. The statistical description of SA2 Fremantle is as such:

```
merged_df[merged_df['sa2_name_2021'] == 'Fremantle'].describe()
```

	price	bedrooms	bathrooms	garage	land_area	floor_area	build_year	cbd_dist	nearest_stn_dist
count	140	140	140	140	140	140	140	140	140
mean	1041554	3	2	2	455	179	1974	13948	1127
min	270000	1	1	1	125	63	1886	12400	164
25%	778750	3	1	2	295	128	1938	13100	623
50%	922500	3	2	2	362	170	1993	13650	966
75%	1201250	3	2	2	499	223	2000	14900	1600
max	2400000	5	3	7	2198	400	2015	16000	3200
std	408320	1	1	1	348	67	38	996	689

The property's price is an outlier in the region and our dataset might not have enough data points for these kind of luxury properties in the region.

Index 330:

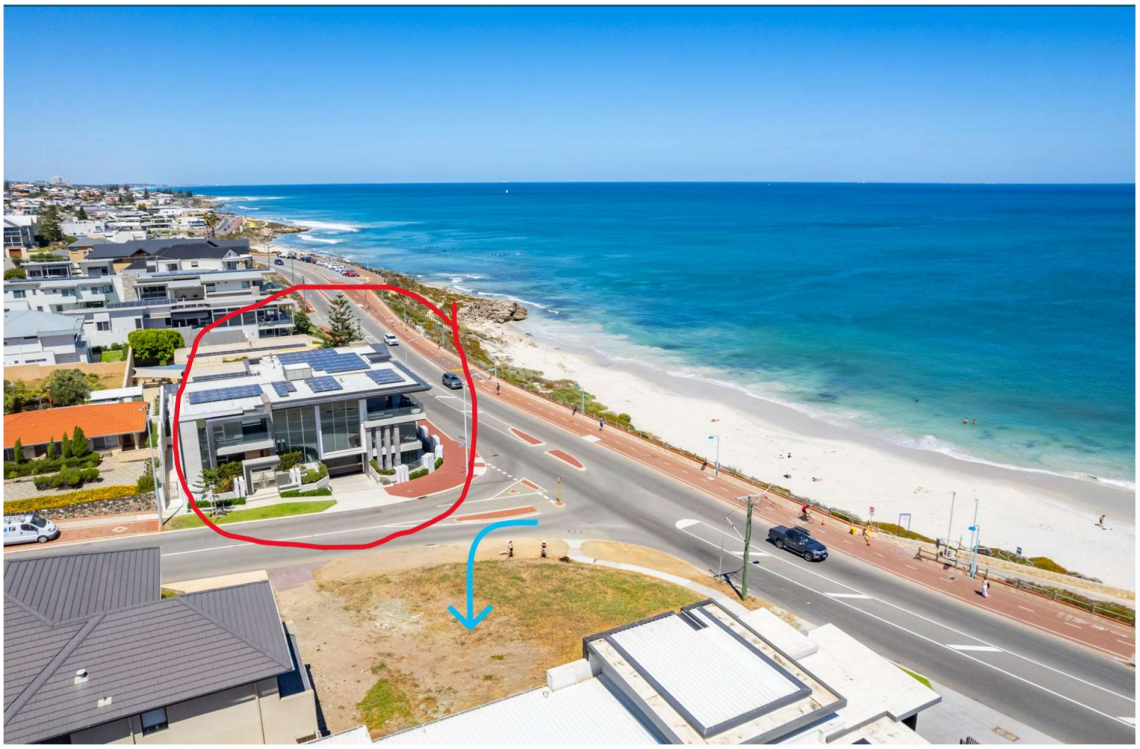
address	1 Malcolm Street
suburb	North Beach
price	1980000
bedrooms	2
bathrooms	2
garage	1
land_area	637
floor_area	114
build_year	1983
cbd_dist	14600
nearest_stn	Warwick Station
nearest_stn_dist	4300
date_sold	2018-08-01 00:00:00
postcode	6020
latitude	-32
longitude	116
nearest_sch	CARINE SENIOR HIGH SCHOOL
nearest_sch_dist	2
nearest_sch_rank	47
sa2_name_2021	Trigg - North Beach - Watermans Bay
nearest_sch_rank_filled	47
land_area_log	6
nearest_stn_dist_log	8
nearest_sch_dist_log	1
cbd_dist_log	10
sale_year	2018
sale_month	8

This property is an interesting outlier case, whereby the premise of the price is leaning towards the sale of land (with existing property as a secondary proposition) with the view of rebuilding a new property that will have a higher valuation.

The sale reflected in our dataset (https://www.realestate.com.au/sold/property-house-wa-north+beach-127095818) has the following excerpt from its description, which is indicative of redevelopment of the land along with a photo of ongoing redevelopment nearby (circled in red):

“With your driveway positioned off Malcolm Street, your block maximises its coastal location with ease of parking and access, enveloped in the beauty and activity of West Coast Drive. Enjoy the comfortable two-bedroom home currently at the property while you design the residence of your dreams or take advantage of the architecturally designed DA approved plans available for viewing from the current vendors.”

The property was eventually torn down and the land sold for development at $2,950,000 on May 7, 2024, with nearby redevelopment completed (circled in red) (https://www.realestate.com.au/property/1-malcolm-st-north-beach-wa-6020/). This might mean future property prices in this area will see an uptrend.



This property shows that there can be an in-between situation where the property sold has existing housing in place but the valuation is based on the potential appreciation of land and new property to be built, rather than a property being priced based on existing residential infrastructure.

Index 3735:

address	12 Lynn Street
suburb	Trigg
price	2360000
bedrooms	4
bathrooms	3
garage	4
land_area	847
floor_area	285
build_year	1998
cbd_dist	13600
nearest_stn	Warwick Station
nearest_stn_dist	4800
date_sold	2020-10-01 00:00:00
postcode	6029
latitude	-32
longitude	116
nearest_sch	ST MARY'S ANGLICAN GIRLS' SCHOOL
nearest_sch_dist	1
nearest_sch_rank	7
sa2_name_2021	Trigg - North Beach - Watermans Bay
nearest_sch_rank_filled	7
land_area_log	7
nearest_stn_dist_log	8
nearest_sch_dist_log	1
cbd_dist_log	10
sale_year	2020
sale_month	10

This property's price is on the high side of the SA2 region and suggests that the price in dataset should be log transformed to address the skewed distribution.

Index 20914:

address	4 Hepworth Road
suburb	Trigg
price	1790000
bedrooms	4
bathrooms	2
garage	2
land_area	462
floor_area	130
build_year	2009
cbd_dist	13000
nearest_stn	Stirling Station
nearest_stn_dist	5200
date_sold	2016-12-01 00:00:00
postcode	6029
latitude	-32
longitude	116
nearest_sch	ST MARY'S ANGLICAN GIRLS' SCHOOL
nearest_sch_dist	1
nearest_sch_rank	7
sa2_name_2021	Trigg - North Beach - Watermans Bay
nearest_sch_rank_filled	7
land_area_log	6
nearest_stn_dist_log	9
nearest_sch_dist_log	1
cbd_dist_log	9
sale_year	2016
sale_month	12

This property's SA2 region is once again from Trigg – North Beach – Watermans Bay, which makes it the 3rd entry from the same SA2 out of the top 5 worst test predictions. The price is on the high side and aside from addressing the skewed distribution with log transformation for price, it might also be helpful to take a deeper look into this SA2 region to see if there is any trend in redevelopment by property developers and look to break

down this SA2 region into more granular SA1 regions to capture the context of specific areas being redeveloped or an underlying price context that requires more granularity than the current SA2 region.

9. Recommendations:

1. Examine property trends in SA2 regions that are experiencing redevelopment or SA2 regions that are still too broad to capture specific local context and experiment to see if breaking down these specific SA2 regions into more granular SA1 regions might be beneficial.
2. Consider changing the Z-score to exclude more outliers.
3. Remove outliers where properties are farmland.
4. Log transformation of price to better handle skewed distribution.
5. Further iterate and experiment with other neural network architectures, optimisers and schedulers.
6. Explore the usage of non-neural network models.