Predicting Suburbs That Are Expected To Experience Significant Growth Or Demand In Near Furture

Student Name: Shreeyans Karki

Student ID: s8073324

Unit Code: NIT3171

Unit Name: ICT Business Analysis and Data Visualization

# 1. Introduction

The Melbourne housing market is ever changing and evolving with the population, infrastructural and housing demand changes. As a real estate developer and agent, it's really hard to predict where the growth for the market might occur next. And it is crucial to spot such growth as soon as possible as it means rising demands and skyrocketing property values. This can let the agents or developers make a smart investment and develop strategies asap.

This business analysis aims to identify suburbs that are expected to have a growth spurt or demand in the future. Using data of historical sales and which suburb it was done in, our goal is to locate trends and suggest the regions that will become the next central property investment area.

Sounds good, but how do we actually accomplish this? We will be applying a combination of supervised learning and time series forecasting methods. These can help magnify the underlying relation between price changes and property features over time. This will allow us to detect any abnormal trend that signals good momentum for a suburb.

This report covers:

1. Business Analysis Problem/Task
2. Methodology
3. Results and Discussions

By the end, readers will have a data driven result to read and analyze for themselves and understand which suburb is most likely to become the next hotspot of property management, in turn allowing better investment planning and data driven decision making.

# 2. Business Analysis Problem/Task

In the real estate sector, identifying future growth areas is a must for effective investment, development planning and infrastructure allocation. As Melbourne's suburban areas are rapidly developing, the estate agents and investors are in need of data driven insights to guide their decisions and their resource allocations.

The main business analysis problem that is addressed in this report is the identification of emerging suburbs- the suburbs most likely to experience increased housing demand and price appreciation soon. To understand where these shifts are occurring enables better strategic planning and reduces the risk of missed opportunities in a competitive property market.

To resolve this dilemma, the report is going to:

* Analyze suburb-level property trends using historical house sales data.
* Forecast future house prices for individual suburbs using time series modeling.
* Model and evaluate suburb characteristics (e.g., distance to CBD, average land size, property count) to understand which factors contribute to price growth.
* Rank suburbs based on their forecasted price growth to spotlight high-potential areas.

By completing the above analysis, we can expect to get results such as:

* A ranked list of suburbs showing strong upward price trends.
* A set of predictive factors that explain suburb-level growth patterns.
* Actionable insights that help stakeholders prioritize areas for future development or investment.

The main benefit of this business analysis job is that it gives you data-driven proof for finding growth suburbs early on. This encourages more intelligent real estate choices, makes it possible to distribute development funds more effectively, and helps create better urban planning throughout Melbourne's metropolitan areas.

# 3. Methodology

This study uses supervised learning algorithms and time series forecasting models to solve the business analysis problem of identifying Melbourne's rising suburbs. This hybrid approach makes it possible to identify important suburb traits that are correlated with growth as well as project future property prices.

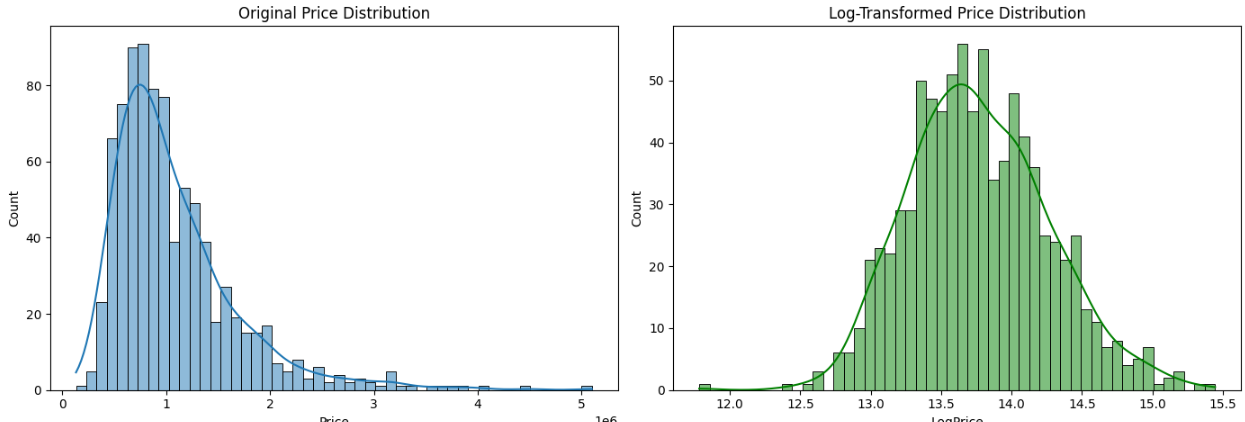
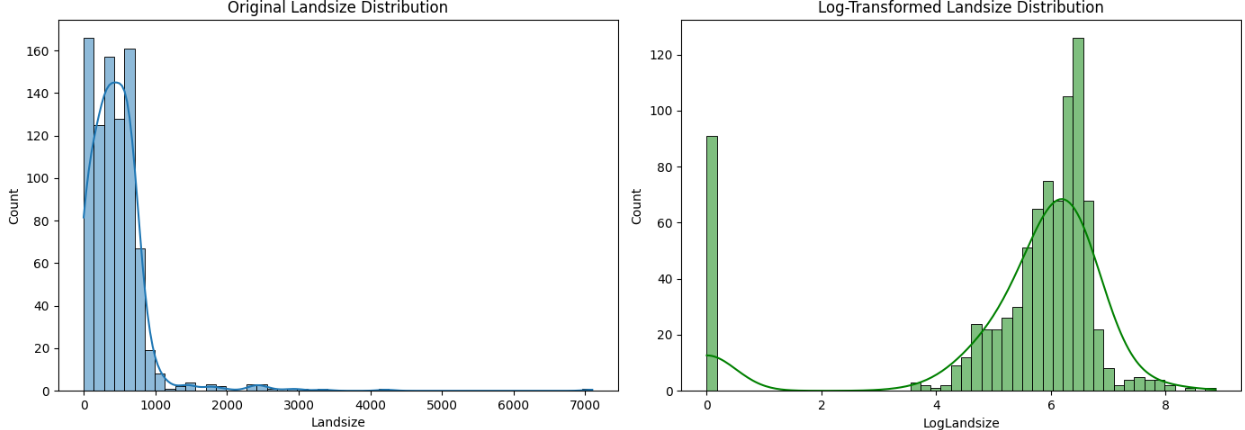
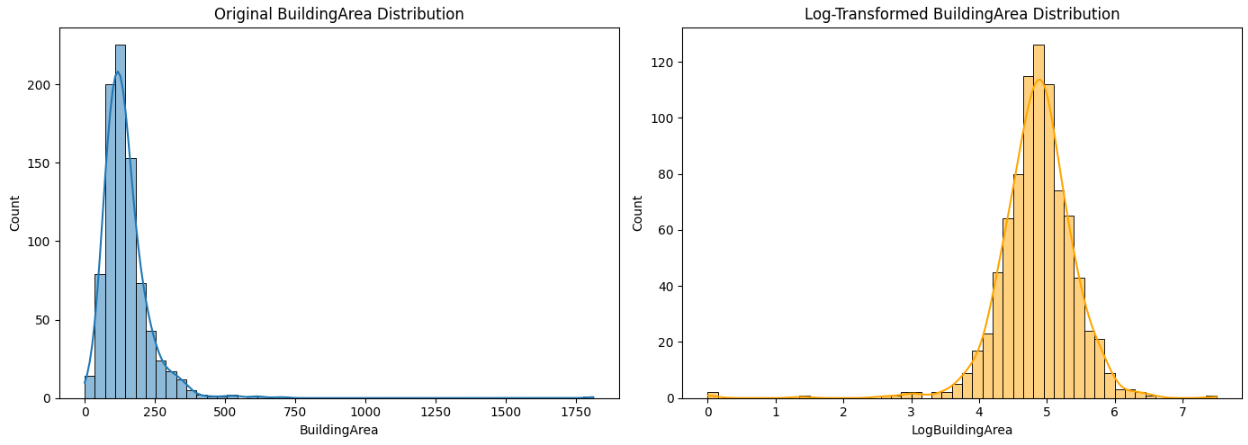
We also applied log transformation to highly skewed attributes such as Price, Landsize and BuildingArea. The results are as follows:  
  


Figure: Histograms showing original price distribution and the log-transformed price distribution

Figure: Histograms showing original Landsize distribution and the log-transformed Landsize distribution

Figure: Histograms showing original BuildingArea distribution and the log-transformed BuildingArea distribution

What data mining method/algorithm/model is used?

Two categories of models were used:

1. Time Series Forecasting
   1. Model: Exponential Smoothing (ETS) and ARIMA (Auto Regressive Integrated Moving Average)
   2. Purpose: Forecast the average housing price in each suburb over a future horizon (6–12 months).
2. Supervised Learning
   1. Model: Random Forest Regression and Linear Regression
   2. Purpose: Model the relationship between suburb features (e.g., distance to CBD, rooms, land size) and the forecasted price growth percentage.

Brief introduction of the method/algorithm/model

* ARIMA is a classical time series forecasting model that captures trends and seasonality by analyzing dependencies between past values. It is effective for stable, linear trends in time series data.
* Exponential Smoothing (ETS) uses weighted averages of past observations, giving more weight to recent data, and is well-suited for short-term forecasting of smooth economic indicators like house prices.
* Random Forest Regressor is an ensemble machine learning model that builds multiple decision trees and averages their outputs. It handles non-linear relationships well and is robust to outliers.
* Linear Regression is a simple, interpretable model that explains how input variables linearly relate to the output. It is useful for understanding which features most influence housing growth.

Why were these methods selected?

* Time series models (ARIMA and ETS) are ideal for forecasting numeric values over time and were used to predict future price levels per suburb based on historical trends.
* Random Forest Regression was selected for its high accuracy and ability to model complex feature interactions in non-linear housing data.
* Linear Regression complements this by offering interpretability — helping identify which factors contribute most to price growth.
* This combination provides both predictive power and explanatory insight, aligning with the goal of identifying and justifying emerging suburbs.

What dataset is used?

The dataset used is the Melbourne Housing Market dataset, containing detailed property transaction records across various suburbs. Key attributes include:

* Transaction details: Price, Date of Sale, Method of Sale
* Property features: Rooms, Bathroom, Car spaces, Landsize, BuildingArea
* Location info: Suburb, Distance to CBD, Council area, Property count

The dataset spans multiple years and includes thousands of records across hundreds of suburbs.

Pre-processing of the dataset

To ensure data quality and model readiness, the following preprocessing steps were performed:

1. Missing Value Handling:

Records with null values in Price, Suburb or Date were removed. Columns with high missingness (e.g., BuildingArea) were conditionally included.

1. Date Formatting:

The Date field was converted to datetime format to support monthly aggregation.

1. Monthly Aggregation:

Data was grouped by Suburb and Month, calculating mean values for Price, Rooms, Bathroom, Car, Landsize, etc.

1. Log Transformation:

Applied to skewed columns (Price, Landsize, BuildingArea) using log1p to normalize distributions.

1. Feature Extraction:

A ready-to-use dataset was created with:

* 1. Inputs: Aggregated suburb-level attributes
  2. Target: Forecasted % growth in average price over the next 6–12 months (from time series model)

This preprocessed dataset was then used for training and evaluation in the supervised learning stage.

Justification for Not Using the Sliding Window Method

A sliding window approach is commonly used in time series forecasting and supervised learning, especially for sequence modeling (predicting stock prices) and models like LSTM, RNNs or regression with lagged variables.

However, in our context, forecasting aggregated monthly suburb-level prices, the sliding window method is not optimal because our task isn’t to predict the next price point based on lag of previous data. We also are forecasting using models fitted per suburb and aren't building global sequence model. We are more interested in trend extrapolation and correlation of features and not short-term prediction.

So, forecasting the time series of suburbs using ARIMA/ETS aligns better with our business goals.

# Results & Discussion

The results of the analysis and the interpreted meaning and implications for identifying the emerging Melbourne suburbs are highlighted in this section. The project combines time series forecasting and supervised learning to evaluate suburb-level growth potential based on historical housing data.

Key Results

1. Top Emerging Suburbs by Predicted Growth:  
    The top 5 suburbs predicted to show the highest growth in housing prices are:
   1. Balwyn North (
   2. Bentleigh East
   3. Glenroy
   4. Pascoe Vale
   5. South Yarra

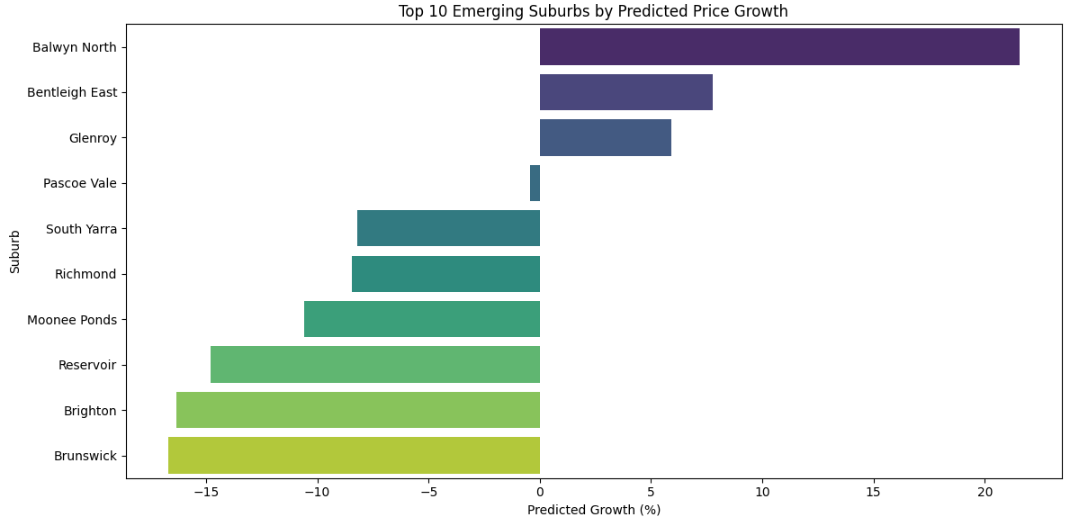


Figure: Top 10 Emerging Suburbs by predicted growth (%)

These were identified based on their high predicted percentage growth from a trained Random Forest model.

1. Feature Importance:  
    According to the Random Forest model:
   1. LogLandsize, Propertycount, and LogBuildingArea were the most influential features in predicting future price growth.
   2. Rooms, Bathroom, and Distance had moderate impact, while Car showed minimal influence.

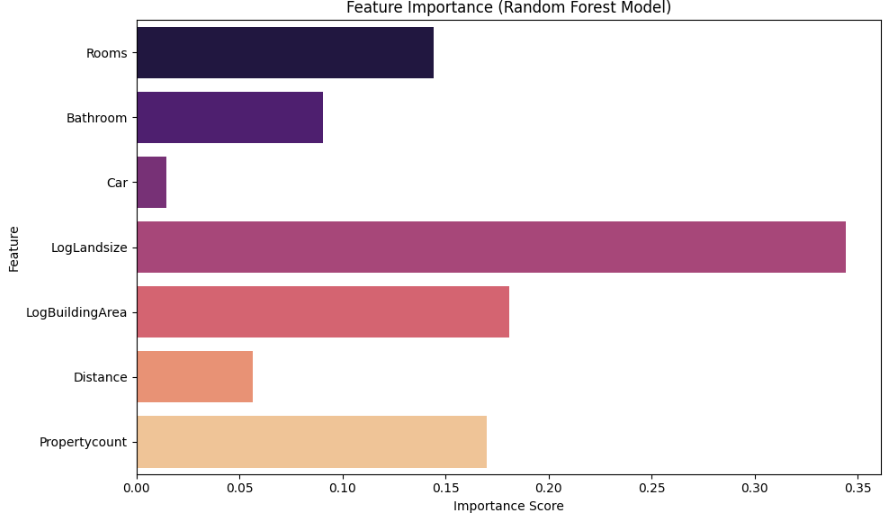


Figure: Feature Importance Plot

1. Model Evaluation:  
    The predictive model was evaluated using both training and test data:
   1. Train R^2: 0.828
   2. Train RMSE: 7.708
   3. Test R^2: 0.068
   4. Test RMSE: 16.910

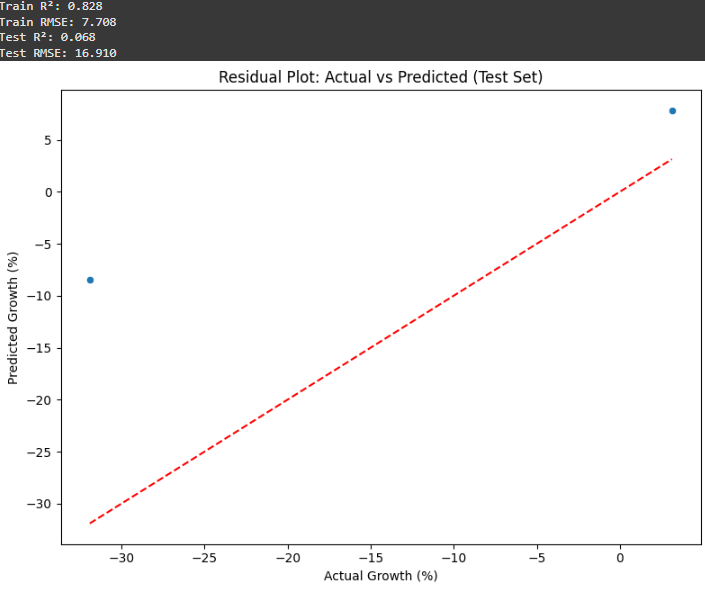


Figure: Predicted growth (%) vs Actual growth (%)

These scores indicate that the model fits the training data well but has limited generalization power on unseen data. The low-test R^2 suggests that the model may be overfitting or that additional variables are needed to improve accuracy.

Discussion

The research offers a useful but preliminary understanding of which suburbs could be classified as "emerging." Consistent market trends and land availability make certain suburbs, like Bentleigh East and Balwyn North, stand out.

The log transformation applied to Price, Landsize, and BuildingArea effectively normalized their distributions, improving model stability and interpretability. Histogram comparisons before and after the transformation visually confirm reduced skewness.

Nonetheless, particularly in extreme situations, the residual plot displays a wide range between the test sets actual and projected values. This demonstrates that, despite its intelligence, the current model is not very predictive when applied to data other than training.

Contributing factors to the low-test performance may include:

* Small sample size in some suburbs
* Uncaptured external influences like upcoming infrastructure, school zones, or policy changes
* Lack of up-to-date data post-2020 market changes

Suggestions for Business Management

* Give top-ranked suburbs with strong indications of future demand, such Bentleigh East and Balwyn North, priority for marketing and development.
* This model should be used as a guide, not as a final decision-maker, particularly when combined with market research, census updates, and local government plans.
* To increase prediction accuracy, update the model by adding additional datasets (such as school evaluations, transportation projects, and population trends).
* As fresh data becomes available, periodically retrain the model to account for recent changes in the market and consumer behavior.

# Conclusion

The business analytical task of identifying new suburbs in Melbourne's housing market that are anticipated to see future demand and price increases was the focus of this report. For developers, investors, and planners looking to make quick and wise real estate decisions, this is a major obstacle.

We used a hybrid data mining technique that combined supervised learning (Random Forest Regression) with time series forecasting (Exponential Smoothing) to address this issue. To increase the dependability of the model, skewed variables were log-transformed and historical housing data was aggregated at the suburb-month level. Short-term growth estimates were generated using suburb-wise forecasts and further modeled using attributes such as property size, location to the central business district, and number of rooms.

Suburban areas including Glenroy, Bentleigh East, and Balwyn North were identified as major growth prospects in the results. The main predictors of expansion, according to feature importance analysis, were land size, building area, and property density. However, the test performance (R^2 = 0.068) indicated possible overfitting and space for improvement, even though the model had a high R^2 score on training data (0.828).

In conclusion, the experiment showed the viability and promise of integrating machine learning and forecasting to identify real estate growth patterns. Although future research should concentrate on including external factors and improving the model for improved generalization, the results can enhance evidence-based decision-making.