Data Analytics Using R

Project 2 - Property market in Prospect, SA, Australia

Students
Huyen Thi Thu Pham

Contents

1 Introduction

1.1 Objective:

The primary goal of this project is to analyze the property market in Prospect, South Australia (SA, 5082), focusing on trends in property prices and sales volumes over time. Additionally, the project seeks to determine whether the textual descriptions of properties contribute to making them more attractive to potential buyers.

2 Data Understanding and Visualisation

Data Description

The dataset was collected by scraping property sales information from the www.homely.com.au website. The dataset includes property that were listed and sold in the Prospect, SA (5082) from July 2014 to October 2024.

The dataset comprises 1,250 property records, covering various property types such as houses, townhouses, units, apartments, villas, land, and other, which distributes as Figure 1. The recorded property prices range from $211,500 to $2,400,000, although approximately 58.64% of the dataset (733 properties) has undisclosed prices. Most properties feature 2-3 bedrooms, 1-2 bathrooms, and 1-3 car spaces. Land sizes vary widely, with a median size of 618 sqm, though there are 513 entries with missing land size values.

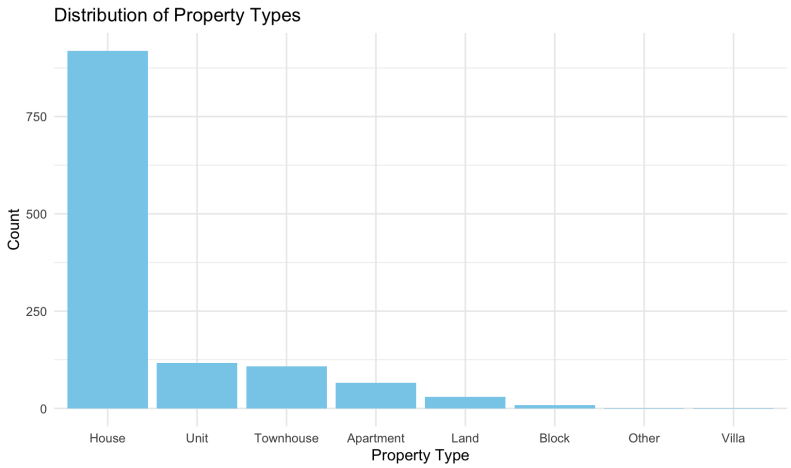

Figure 1: Distribution of Property Type

Top Ten Real Estate Companies in Prospect

Table 1 shows the top ten real estate companies by number of sales in Prospect, highlighting Fox Real Estate and Toop & Toop as the leading agencies.

Company Name	Count
Fox Real Estate	119
Toop & Toop - Norwood	80
Klemich Real Estate - Kent Town	76
Ouwens Casserly Real Estate	72
D B Philpott Real Estate	67
Harris Real Estate - Kent Town	65
Ray White - Norwood	53
LJ Hooker - Walkerville	47
First National Real Estate - Riggall	37
Refined Real Estate	20

Table 1: Top 10 Real Estate Companies by Count

Top Ten street have highest price in Prospect Figure 2 show that **Prospect Rd** stands out as the most expensive street, suggesting it might have larger or more luxurious properties. Closely examining the data, the highest price belongs to a combination of two properties, priced at $2,400,000. **Milner St and Alpha Rd** also have high average prices, indicating they are desirable location with potentially high property values. Following them are **Wilcox Ave, Gloucester St, Koonga Ave, Labrina Ave, Avenue Rd, Moore St and Clifford St**.

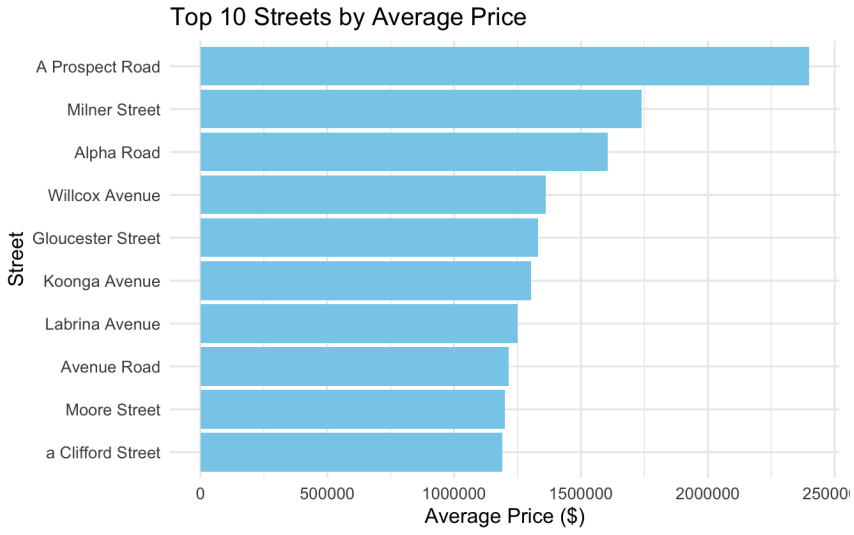

Figure 2: Top ten street have highest price in Prospect

3 Data Cleaning and Preparation

Data cleaning included handling missing values, formatting dates, and removing duplicates. Text in the property descriptions was pre-processed to standardize it for later analysis by converting to lowercase, removing special characters and punctuation, and eliminating stop words. Tokenization was also applied to support word frequency analysis.

After cleaning, the resulting dataset comprised 496 entries, distributed across various property types, as shown in Table 2.

3.1 Key Market Data Overview

Key Market Data	House	Townhouse	Unit	Apartment
Median price ($)	760,000	549,500	340,000	410,000
Number of Sales	373	42	52	29

Table 2: Key Market Data for Different Property Types

Table 2 illustrates the distribution of property types and their key market metrics.

- Houses have the highest median price of $760,000, indicating they are the most expensive property type in the dataset. This is further supported by the highest number of sales (373), suggesting they are the most sought-after properties.
- Townhouses, with a median price of $549,500, are moderately priced and account for 42 sales.
- Units emerge as the most affordable option, with a median price of $340,000 and 52 sales, while apartments have a median price of $410,000 and the fewest sales at 29, indicating lower demand.

Figure 3 shows the price distribution for different property types:

- Houses: The price distribution for houses is wide, with prices ranging from around $500,000 to over $2,000,000, reflecting significant variability likely attributable to factors such as location, size, and amenities.
- Townhouses: The price range for townhouses is narrower, mostly between $500,000 and $1,000,000. This suggests more consistency in townhouse prices.
- Units: Units have the lowest price range, mostly between $300,000 and $500,000, indicating they are generally more affordable.
- Apartments: The price distribution for apartments is similar to units but slightly higher, ranging from $400,000 to $600,000.

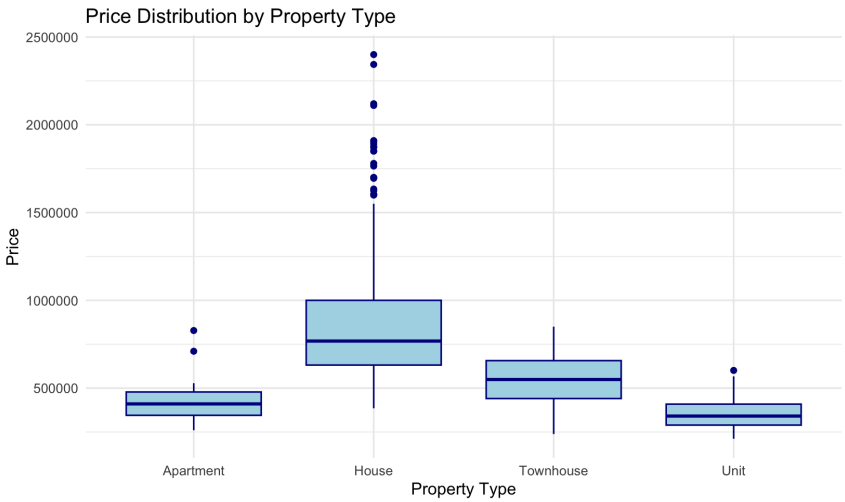

Figure 3: Distribution property type

3.2 Key Findings from Data Exploration

Price Trend

Figure 4 displays the distribution of property prices, revealing that the most common price range is between $ 400,000 to $800,000 and the overall spread of prices, helping us understand the pricing trends in the property market.

Figure 4: Density Plot

Property Market Evolution

Figure 5 illustrates the evolution of property prices and sales volumes in Prospect from 2014 to 2024.

- A slight decline in average prices coupled with a significant increase total sales from **2015 to 2016**, suggesting a period of low demand.
- From **2016 to 2019**, total sales experienced a slight decline, while prices remained relatively stable, indicating a potential market slowdown during this time.
- A gradual recovery in **2020**, likely influenced by external economic factors, including reduced interest rates intended to stimulate the housing market.
- A significant surge in prices and sales in **2021**, possibly exacerbated by the COVID-19 pandemic's effects on housing demand.
- Continued growth in both prices and sales from **2021 to 2024**, suggesting a robust demand and recovery in the property market.

Between 2016 and 2019, both average prices and total sales experienced a slight decline, followed by a gradual increase in 2020. This trend sharply escalated in 2021, likely due to the COVID-19 pandemic's impact, including reduced interest rates to stimulate buying. From 2021 to 2023, both prices and sales continued to rise, suggesting robust demand in the property market.

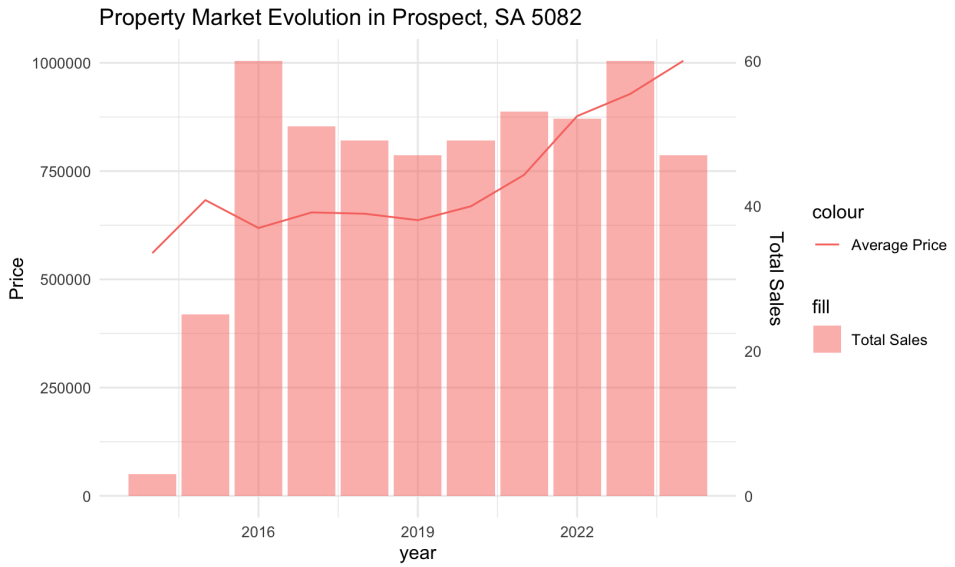

Figure 5: Property Market Evolution in Prospect, SA 5082

Season Sales Trends:

Figure 6 reveals that sales tend to peak in November and dip in June. This trend suggests that sales activity is highest in the spring season, potentially due to favorable weather and the end-of-year holiday period, which often encourages market activity.

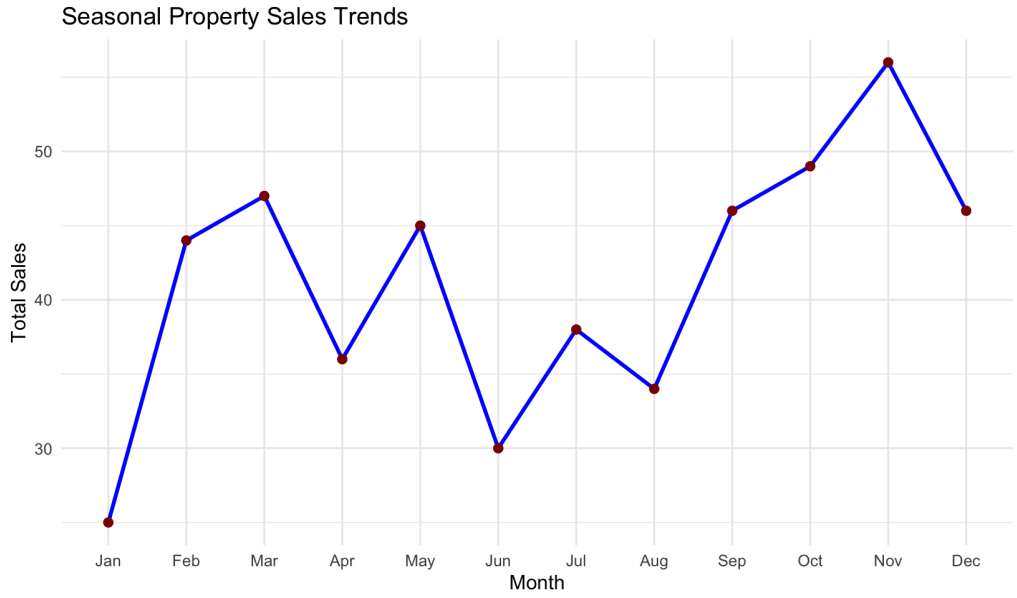

Figure 6: Seasonal Property Sales Trends

4 Clustering Analysis

Property Clustering Based on Categorical Attributes

To explore how property type and location may impact price, we used the K-Modes clustering algorithm. K-Modes is effective for handling nominal data, making it well-suited for grouping properties based on categorical attributes. Using this method, properties were grouped into two clusters based on type and other categorical features. Table 3 provides summary statistics for these clusters.

Cluster	Avg. Price	Avg. Bed	Avg. Bath	Avg. Car	Avg. Land
1	822,471	3.11	1.53	2.22	547
2	551,940	2.48	1.34	1.51	280

Table 3: Cluster Summary Statistics

This clustering approach reveals clear patterns in the property data:

- Cluster 1: This cluster has a higher average price and generally includes properties with larger land sizes, more bedrooms, bathrooms, and car spaces, suggesting it may represent higher-value or more spacious.
- Cluster 2: With a lower average price, Cluster 2 comprises properties with fewer bedrooms, bathrooms, car spaces, and a smaller land size, likely indicative of more compact or affordable properties.

5 Prediction model

Feature Correlation Analysis

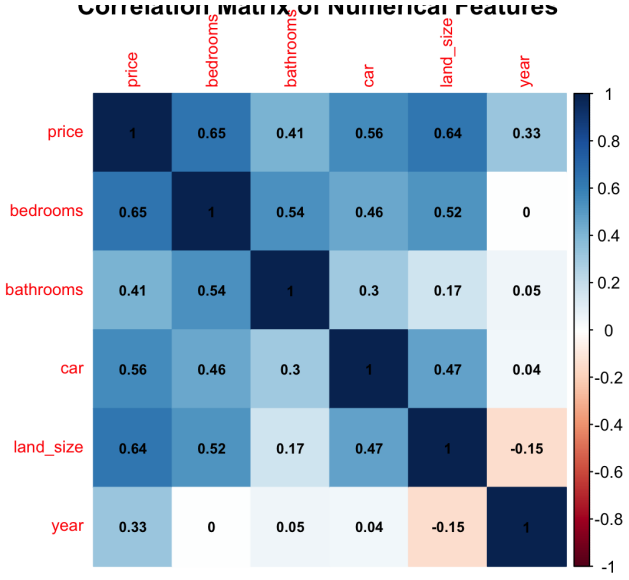

Figure 7: Correlation Matrix of Numerical Features

Figure 7 presents the correlation matrix for numerical features in the dataset, indicating several important relationships:

- Positive Correlation with Price: Price shows a positive correlation with the number of bedrooms, bathrooms, parking spaces, and land size, indicating that properties with more space and features generally have higher prices.
- Negative Correlation with Year: Year exhibits a negative correlation with most property features, suggesting that more recently sold properties may be smaller or less feature-rich than older ones when sold for the same price.

Model Training and Evaluation

To create a robust predictive model for property prices, we added the cluster labels generated by the K-Modes clustering as an additional categorical feature. This step helps capture complex patterns associated with property type and location. The

dataset was then split into training (80%) and testing (20%) sets to evaluate the model's performance on unseen data.

We trained a linear regression model using bedrooms, bathrooms, car, land size, year, and cluster as predictor variables.

Residuals: Min = -515,578, Max = 859,284, indicating a range in the prediction errors. Residual Standard Error (RSE): 195,400 on 389 degrees of freedom.

Coefficient	**Estimate**	**Std. Error**	**t value**	$Pr(>\|t\|)$	**Signif.**
(Intercept)	-9.991e+07	7.132e+06	-14.008	¡ 2e-16	***
bedrooms	1.109e+05	1.737e+04	6.388	4.80e-10	***
bathrooms	7.656e+04	1.984e+04	3.859	0.000133	***
car	4.132e+04	9.083e+03	4.549	7.22e-06	***
land_size	6.267e+02	4.824e+01	12.991	¡ 2e-16	***
year	4.944e+04	3.532e+03	13.998	¡ 2e-16	***
cluster	-2.274e+03	2.002e+04	-0.114	0.909652	

Table 4: Summary of Model Coefficients

Table 4 provides detailed coefficients for each variable:

- Each additional bedroom is associated with a price increase of about $110,900, holding other variables constant. This variable is statistically significant (p ¡ 0.001).
- Each additional bathroom is associated with an increase in price of approximately $76,560, also significant (p ¡ 0.001).
- Each additional car space adds roughly $41,320 to the price, significant as well.
- For each additional square meter of land, the price is expected to increase by about $626, which is highly significant.
- Each additional year adds around $49,440 to the price, and indicate an upward trend in property prices over time.
- The coefficient for cluster is -2,274, but it has a high p-value (0.909652), suggests that cluster membership doesn't contribute meaningful information to predict price.

Model Performance

The model's performance was evaluated using the Mean Squared Error (MSE) and R^2 metrics on the test dataset:

- Mean Squared Error (MSE): The model achieved an MSE of approximately 30.46 billion, indicating the average squared difference between predicted and actual property prices.

- R^2 Score: The R^2 score of 0.783 suggests that the model explains around 78.3% of the variance in property prices on the test dataset. This high R^2, close to the training set R^2, indicates the model has strong predictive power without significant overfitting.

Overall, the model provides reliable predictions of property prices based on key features, highlighting the influence of bedrooms, bathrooms, car spaces, land size, and sale year on property value. The clustering variable, however, did not show a significant effect on price prediction.

6 Text Analysis:

To deepen our understanding of property values, we analyzed property descriptions, focusing on commonly used words to identify key themes and examine their potential influence on property prices. By leveraging clustering and text analysis techniques, we can assess how descriptive language correlates with the desirability and perceived value of properties.

To prepare the data, we first created a vocabulary from the descriptions, removing stop words, numbers, and special characters. Words with fewer than three occurrences were excluded to ensure relevance. Each word was then stemmed and converted into a binary presence-absence format, resulting in a document-term matrix where each property's features are represented.

6.1 Common Themes in Property Descriptions :

Examining the most frequently used words in property descriptions provides insight into features that real estate agents prioritize to attract buyers. Figure 8, reveals common themes that real estate agents emphasize to appeal to buyers.

- Core Features: Words like "bedroom," "home," "kitchen," and "bathroom" emphasize essential household amenities.
- Location & Accessibility: Terms like "park," "zone," "transport," and "suburb" stress proximity to amenities and convenience.
- Lifestyle & Quality: Words such as "modern," "secure," "quality," and "entertain" appeal to buyers looking for comfort, lifestyle, and investment value.
- Utilities: Terms like "gas," "water," and "storage" underline essential utilities.

Figure 8: Word Cloud

6.2 Clustering Analysis:

To determine whether textual descriptions meaningfully differentiate properties, we converted word which has clean and tokenize from description into a numeric vector and calculated binary distances between properties based on the word matrix. Using hierarchical clustering with Ward's method. The dendrogram in figure 9 revealed two primary clusters, representing the highest level of dissimilarity between groups of properties.

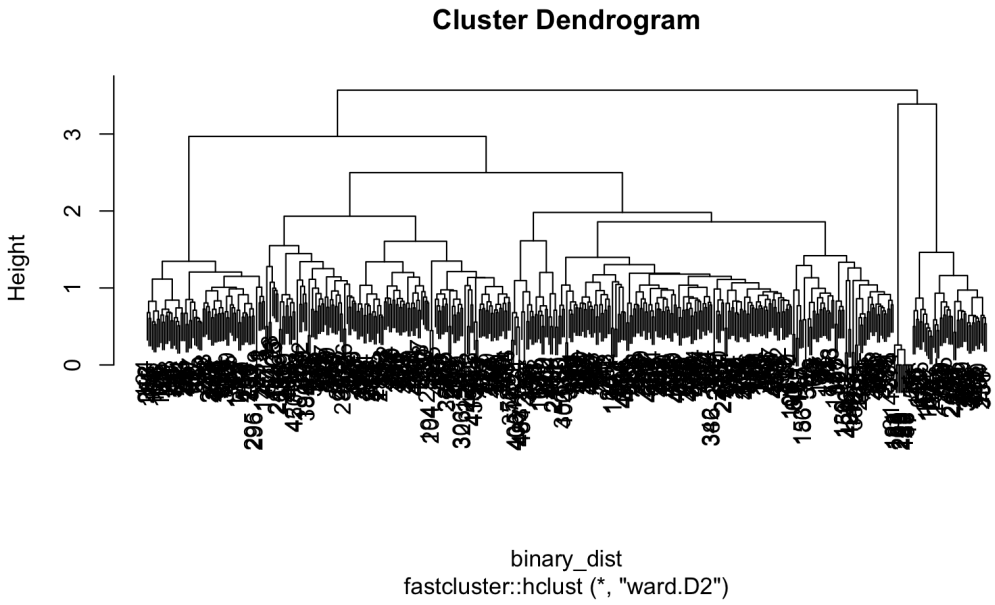

Figure 9: Cluster Dendrogram

From Table 5, we can see:

1. Cluster 1: Commonly associated with family-oriented buyers, properties in this cluster highlight aspects like *bedroom, kitchen, home*, and other household features. This cluster likely reflects buyers looking for family homes with relevant amenities.
2. Cluster 2: Contains terms like *assess, develop, legal, market* that suggest a focus on real estate professionals or investors interested in the business and legal dimensions of property transactions.

Table 5: Most Common Words by Cluster

Cluster	Word 1	Proportion 1	Word 2	Proportion 2
1	bedroom	0.386	kitchen	0.368
1	prospect	0.351	bathroom	0.312
1	built	0.265	dine	0.257
1	school	0.255	shop	0.253
1	offer	0.239	north	0.236
1	rear	0.236	plan	0.250
2	free	0.0515	local	0.0515
2	team	0.0459	assess	0.0412
2	conduct	0.0412	db	0.0412
2	direct	0.0412	due	0.0412
2	legal	0.0412	market	0.0412
2	philpott	0.0412	develop	0.0412

7 Property Market Predictions using Numerical Data with Cluster Analysis

To further analyze the significance of these clusters, we conducted a linear regression using cluster as a predictor. The model predict price based on predictors included the number of bedrooms, bathrooms, car space, land size, and year built, along with cluster.

Regression Results

Table 6: Linear Regression Coefficients

Coefficient	Estimate	Std. Error	t value	$Pr(>\|t\|)$
(Intercept)	−9.836e+07	6.344e+06	−15.505	< 2e-16 ***
bedrooms	1.205e+05	1.493e+04	8.068	5.57e-15 ***
bathrooms	7.308e+04	1.779e+04	4.107	4.69e-05 ***
car	4.485e+04	7.380e+03	6.077	2.47e-09 ***
land_size	5.996e+02	4.042e+01	14.835	< 2e-16 ***
year	4.866e+04	3.141e+03	15.492	< 2e-16 ***
cluster2	−4.616e+04	2.763e+04	−1.671	0.0954 .

Table 6 indicate that,

- Adding a bedroom increases the price by around $120,500. This variable is statistically significant ($p < 0.05$).
- Adding a bathroom increases the price by about $73,080, also significant ($p < 0.05$).

- Each car space adds roughly $44,850 to the price, significant as well.
- Each additional square meter increases price by $599, which is highly significant.
- Each additional year adds around $48,660 to the price, and indicate an upward trend in property prices over time.
- The coefficient for cluster 2 is -46,160, suggesting that properties in this cluster might be priced slightly lower, but this was not statistically significant (p = 0.0954).

Model Performance

The model achieved an R^2 of 0.728, indicating that it explains about 72.8% of the variance in property prices. This suggests that while the numerical variables are strong predictors of price, cluster membership based on description did not significantly improve predictive power.

Our analysis of property descriptions provided valuable insights into marketing themes and buyer preferences. While clustering helped differentiate properties based on descriptive language, these clusters did not meaningfully enhance price prediction when combined with quantitative features. This suggests that while descriptions offer qualitative insights, numerical variables remain the primary drivers of property valuation.

This finding implies that while property descriptions can support targeted marketing, they may lack the quantitative depth to significantly impact price prediction models.

8 Interactive DashBoard

8.1 Dashboard Overview:

The Prospect Real Estate Dashboard provides a comprehensive analysis of property trends from 2014 to 2024. Built using Shiny, ggplot2, and other R libraries, this interactive tool offers valuable insights into the real estate market.

8.2 Key Fearures:

1. **Overview Tab**
    - **User Inputs:** Allows users to input the number of bedrooms, bathrooms, car spaces, land size, and year of sale to predict property prices.
    - **Price Distribution:** Displays a histogram and density plot of property prices, highlighting the distribution over the years.

- **Value Boxes:** Summarizes key metrics such as average price, total properties sold, and the most common property type.

2. **Analysis Tab**
    - **Price Trends Over Time:** A line plot showing the average price trends over the years, accompanied by a bar plot of total sales.
    - **Monthly Sales Trends:** A line plot illustrating seasonal property sales trends.
    - **Feature Importance:** A bar plot of linear regression coefficients, indicating the importance of various features (bedrooms, bathrooms, car spaces, land size, year) in predicting property prices.
    - **Top Streets by Price:** A box plot showing the price distribution for the top 10 streets, helping identify the most and least expensive areas.

8.3 Data and Model:

- The dataset includes key features such as price, bedrooms, bathrooms, car spaces, land size, and year.
- A linear regression model is used to predict property prices based on these features.
- The data is split into training and testing sets to ensure robust model performance.

9 Conclusions and Insights:

This study provides a comprehensive analysis of the property market in Prospect, South Australia, from 2014 to 2024. By leveraging data analytics techniques, we have identified key trends and factors influencing property prices and sales volumes.

Key Findings:

- **Price Trends:** Property prices in Prospect have shown a general upward trend over the past decade, with significant increases observed post-2020, likely influenced by the COVID-19 pandemic and associated economic factors.
- **Seasonal Trends:** Sales volumes peak in November and dip in June, suggesting seasonal variations in market activity.
- **Clustering Analysis:** Properties were grouped into two clusters based on categorical attributes, revealing distinct patterns in property features and prices. Cluster 1, with higher average prices, includes properties with more bedrooms, bathrooms, and larger land sizes, indicating higher-value properties.

- **Prediction Model:** The linear regression model identified bedrooms, bathrooms, car spaces, land size, and year of sale as significant predictors of property prices. The model achieved a high R^2 score, indicating strong predictive power.
- **Text Analysis:** Common themes in property descriptions, such as location, features, and lifestyle, were identified. However, the inclusion of text-based cluster labels did not significantly improve the predictive accuracy of the model.

Interactive Dashboard: The Prospect Real Estate Dashboard offers an interactive tool for exploring property trends and making informed decisions. It provides valuable insights into price distributions, seasonal trends, and the importance of various property features.

Limitations: The study faced several limitations, including a high proportion of properties with undisclosed prices and missing data for certain features. Future research could address these gaps and explore additional factors influencing property prices.