Drivers of Perceived Rental Property Quality: Analysing shitrentals.org Reviews

SID:500680807

This research report used the LLM ChatGPT to assist with proofreading, editing, drawing insights, and writing code.

***Abstract*—Australia is currently facing a persistent housing crisis, characterized by property prices escalating at a rate faster than household income [1]. In response, Australian TikToker Jordie established a publicly accessible database aimed at exposing rental properties that fail to meet minimum quality standards. This study seeks to elucidate the determinants of perceived rental property quality in Australia, utilizing data from the shitrentals.org website. Employing Ordinal Logistic Regression, we statistically analysed factors influencing the quality scores assigned to rental properties. The research process encompassed Exploratory Data Analysis, followed by the regression, selection, and optimization of potential models. The findings reveal that while weekly rental price significantly impacts quality scores, other variables in the model do not exhibit statistical significance. Further research is required to investigate additional factors that may influence rental property quality.**

1. Introduction

Australia is facing an ongoing housing crisis. Across the country, property prices escalated three times the pace of household income since June 2020. In 2023, Jordie an Australian TikToker created a Twitter database with free public access to expose rental properties that do not meet minimum standards of living quality and whose prices have been continuously increasing. This database allows lessees to review their own rental property or real estate agency, acting as 'rental cops' to inform people about the perceived quality of Australian rental properties [1]. This research report's motivation is to understand what drives the perceived quality of rental properties in Sydney, in 3 different suburbs. Indeed, low-quality rental housing is linked to a high mental health toll. An analysis by the Committee for Economic Development of Australia (CEDA) indicates that mold and dampness in homes can significantly adversely affect the health of occupants [2]. These conditions are reported 10% more frequently among renters compared to owner-occupiers. Thus, there is a critical need for improvements in rental property conditions. This study aims to identify the factors that influence perceived rental quality. By classifying quality scores, ranging from low to high, this research investigates explanatory factors that could assist lessees in predicting a property's quality. Such insights are crucial for enabling more efficient rental searches and ensuring lessors are held accountable for the quality of their properties.

In this research, we use Python to construct and test models based on shitrentals.org as our dataset. Ordinal Logistic Regression is used to analyze the score-affecting perceived drivers of rental property quality. The research process mainly includes Exploratory Data Analysis (EDA) which helps to understand the datasets and identify features in the dataset. Then, methods with the selection of multiple potential models that we meticulously compared and optimized thanks to backward elimination. We obtain mixed insights. Although we find that weekly rental price has a strong effect on scores, the other variables of our selected model are not significant. Further research is required to understand the variable of interest.

2. Methods

2.1. Data Collection

Shitrentals.org has 9797 observed values. With a total of 10 variables. Each variables have a total of 1000 values, except for agency_name which only has 797 values and 203 missing values. Variables included are such as weekly rent prices, lessor, agency name, suburb, review id, date, score, property type, number of bedrooms and rental reviews. We divided the dataset respectively: 60% of the data is set as training group, and 40% is set as test group.

2.2. EDA

EDA shows that properties have either 1 or 2 bedrooms and are classified into 3 distinct suburbs: Camperdown, Newtown, or Redfern. As shown in the Figure 1 and Appendix 1 (boxplot 2), scores are strongly correlated with weekly prices, reaching 0.96. Weekly price is also moderately strongly correlated with the number of bedrooms and there is a positive relationship between scores and price. Indeed, the higher is the price, the higher the scores tend to be, on average. Also, the -0.59 correlation between Newtown and Redfern indicates a negative relationship. The properties in Newtown are associated with lower weekly rental prices compared to the base category which is Camperdown. On the other hand, a positive correlation of 0.73 between Redfern and weekly prices shows a strong positive relationship. Thus, properties located in Redfern are associated with higher weekly rental prices compared to the base category, Camperdown. As shown on Appendix 1 (boxplot 4), the average price for properties in Redfern is higher ($375), while it is around $300 for Camperdown and around $160 for Newtown. Furthermore, the property lessors (Agency or Private) have similar average weekly rental prices, which are between $250 to $300.

Figure 1. Correlation Heatmap

Properties with 2 bedrooms have an average price of $300 while 1-bedroom properties cost around $200 (Appendix 1 boxplot 3). Finally, Appendix 2 shows that the average rental prices of properties and their different agency name are similar.

We checked for potential high degree of correlation between variables. According to the correlation matrix Figure 1, no multicollinearity is present between numerical variables. However, we must check the correlation between numerical and dummy variables. Indeed, the Appendix 1 (boxplot 2), shows that there is a possible high correlation between rental price and suburbs. Hence, we tested for multicollinearity thanks to a Point-Biserial Correlation test. The results indicate a correlation of -0.5 between price and Newtown and a correlation of 0.73 between price and Redfern. A correlation of 0.73 is high and could indicate a risk of multicollinearity.

2.3. Data cleaning & feature engineering

2.3.1. Data Cleaning. The dataset contained null values in agency_name, due to the private and anonymous nature of some agencies. Hence, we replaced those names by the term 'Missing_name'. We checked for outliers by creating boxplots and we dropped review_id as it is only a unique identifier and date as it only gives us data on the year 2023.

2.3.2. Dummy variables. Categorical variables were converted to numerical variables using dummy variables. The suburb feature was transformed in 2 dummies (Newtown and Redfern), with Camperdown as a base variable. Lessor was transformed as Private (base variable = Public). Agency names were transformed with Belle Property as a base variable.

2.3.3. text_review. The variable text_review showing rental reviews on properties was transformed in a sentiment analysis using TextBlob Python package widely used for NLP development. It processes textual data and uses 2 dimensions polarity and subjectivity to access the positive or negative nature of comment. This new variable called 'opinion' indicates 1 if the review is overall positive and 0 if it is negative [3].

2.4. Model Selection

Suppose that lessees of properties in Newtown, Redfern and Camperdown are asked to rate how satisfied they are of rental properties. Their scores are scaled from 1 to 5, with 5 indicating the highest level of a property's perceived quality. Score is not a continuous variable but an ordinal categorical variable. Hence, we must use a classification regression to predict and classify values into classes based on different parameters [4]. Models such as a Binary logistic regression is not adequate as it cannot predict the probability of more than 2 events. Additionally, a multinomial regression model is not appropriate as the ordering of the categories is ignored. The most suitable model is the Ordinal Logistic Regression model (OLR). The Ordinal Probit Regression could also be used but is harder to interpret [5]. OLR applies when the dependent variable is an ordinal variable that specifies an order with two or more categories. Also, one or more of the independent variables are either continuous, categorical, or ordinal. There should be no multicollinearity in the model. A key assumption of OLR is proportional odds, meaning the effect of an independent variable remains constant across each level of the response. According to our EDA, the risk of multicollinearity between price and Redfern is high. Hence, we remove the dummy variables for suburbs to avoid the risk of model assumption failure.

Modelling the Original Model (OM) including all features (except review_id and date): (see Figure 2)

LL = -264.07, AIC = 566.1, Accuracy score = 88%

The regression modelling output of the OLR with all the features is shown in Figure 2.

Figure 2: Original Model Regression Output

	coef	std err	z	P> |z|	[0.025	0.975]
weekly_price	0.3047	0.022	13.700	0.000	0.261	0.348
n_bedrooms	18.4703	340.329	0.054	0.957	-648.563	685.504
Newtown	-11.5122	274.389	-0.042	0.967	-549.306	526.281
Redfern	18.7268	340.330	0.055	0.956	-648.308	685.761
Private	-0.4247	0.480	-0.885	0.376	-1.365	0.515
Century 21 Australia	-0.2599	0.548	-0.474	0.636	-1.335	0.815
First National Real Estate	0.0080	0.562	0.014	0.989	-1.094	1.110
Harcourts	-0.3816	0.558	-0.684	0.494	-1.476	0.712
LJ Hooker	-0.4844	0.568	-0.853	0.394	-1.598	0.629
McGrath Estate Agents	-0.2200	0.543	-0.405	0.685	-1.284	0.844
Professionals Real Estate Group	-0.1750	0.583	-0.300	0.764	-1.318	0.968
RE/MAX Australia	-0.3351	0.568	-0.590	0.555	-1.448	0.778
Raine & Horne	-0.1940	0.562	-0.345	0.730	-1.296	0.908
Ray White Group	0.5674	0.568	0.998	0.318	-0.546	1.681
opinion	0.0086	0.227	0.038	0.970	-0.435	0.453
1/2	76.7029	340.338	0.225	0.822	-590.347	743.752
2/3	3.2122	8.108	0.396	0.692	-12.678	19.103
3/4	3.2432	10.712	0.303	0.762	-17.751	24.237
4/5	1.8228	0.074	24.763	0.000	1.679	1.967

This summary regression output gives the estimated log-odd coefficients of each of the independent variables shown in the 'coef 'output section. Weekly_price is highly significant due to a very low p-value 0.000.

TABLE 1. ALL FEATURES SUMMARY STATISTICS FOR BACKWARD ELIMINATION

Log-Likelihood	AIC	Accuracy score in %	Elimination	Feature
-500.33	1037.0	not calculated	remove	weekly_price
-264.08	564.2	88%	remove	n_bedrooms
-264.09	562.2	88%	remove	suburb (Newtown + Redfern)
-264.46	564.9	88%	remove	Private
-266.63	553.3	88.35%	remove	agency_name
-264.07	564.1	86.97%	remove	opinion

We assume that it has a major impact in predicting rental scores. Table 1, represents the backward elimination process of all predictor variables. It gives the log-likelihoods (LL), AIC scores and accuracy of OLR models where we dropped features one by one from the Original Model. The distribution of scores classes are relatively balanced. Because there is no class imbalance, we can use the accuracy score to check the model's effectiveness.

2.4.1. weekly_price. To illustrate Table 1: dropping weekly_price from OM, gives us a LL of -500.33 which is very low compared to OM. The AIC is also high. However, lower LL values and high AIC values indicate a worst fitting model. Score is highly correlated with weekly rental price which is significant. Hence, we expect to worsen the model fit if we drop price. Our model is not properly defined if we drop price.

2.4.2. review_text. Dropping the feature opinion (dummy variable for review_text) gives the same LL as OM, -264.0, but a slightly better AIC = 564.1. However, 86.97% accuracy score is lower than OM. Hence, we are better off keeping opinion in our model.

2.4.3. agency_name. Dropping agency_name which includes all agencies such as LJ Hooker and Harcourts, gives a better AIC, but the LL is slightly worst. However, it improves the accuracy score by 0.35%. Thus, we drop agency_name.

2.4.4. suburb. In our EDA, we found a risk of multicollinearity between suburbs and score. Hence, we drop Newtown and Redfern (dummy variables for suburb). Also, dropping suburb gives a better AIC, and the LL is unchanged. We drop suburb as it doesn't improve the model.

2.4.5. Private. Dropping lessor, which is represented by the Private dummy variable, gives a better AIC and a worst LL. The accuracy is the same. From our EDA, we know that the difference between the average weekly price and private or agency lessors is not significant, meaning that lessor won't significantly affect prices and indirectly scores. The risk of multicollinearity is low hence, we keep Private as the AIC is improving.

2.4.6. n_bedrooms. We drop n_bedrooms as dropping it gives a slightly better AIC. Also, according to our EDA there are only units with either 1 or 2 bedrooms which doesn't make a major difference in predicting scores (correlation = 0.57).

Selected Model (SM) equation: Score = B1 weekly_price + B2 opinion + B3 Private + error

3. Results and Discussion

LL = -267.18 AIC = 548.4 Accuracy score = 88%

TABLE 2. SELECTED MODEL REGRESSION OUTPUT

	coef	**std err**	**z**	**P> |z|**	**[0.025**	**0.975]**
weekly_price	0.3046	0.022	13.783	0.000	0.261	0.348
Private	-0.2750	0.284	-0.967	0.334	-0.833	0.283
opinion	0.0430	0.221	0.195	0.846	-0.390	0.476
1/2	58.4058	4.278	13.652	0.000	50.021	66.791
2/3	2.8804	0.074	38.777	0.000	2.735	3.026
3/4	2.6383	0.076	34.591	0.000	2.489	2.788
4/5	1.8275	0.073	24.940	0.000	1.684	1.971

After model comparison and context-based justification, 3 regressors were chosen for the stepwise regression from 7 available features. The Selected Model has a smaller LL by 0.03% compared to the OM. This is not very significant compared to the 0.1% improvement that the SM AIC has. This AIC is the lowest from all our backward selection model selection process, which indicates that SM is the best fitting model.

Coefficients interpretation:

For every one unit increase in weekly price, the odds of having a higher rental score increase by 35.6%, holding constant all other variables. This result is statistically significant at a 5% level of significance (p-value = 0.000). We conclude that weekly price has a major impact on scores.

For properties with a Private lessor, the odds of having a higher rental score (1-5) is 24.03% (10.7597=0.24031*100%) lower that of lessor that are Agencies, holding constant all other variables. However, the effect is not significant given its p-value. We do not have sufficient evidence to conclude that listing a property as Private significantly impacts scores. The decrease in odds is not statistically significant.

For properties with a positive opinion (coded = 1), the odds of having a higher rental score is associated with 1.044 times the odds of having a higher score compared to negative opinion (coded = 0). But this result is not statistically significant (p-value = 0.846).

In OLR, thresholds are used to make the difference between adjacent levels of scores. The interpretation of the first cut point coefficient is:

Properties that had a value of 58.41 or less on the underlying unobserved variables that gave rise to the score would be classified as lower scoring given that properties had an agency lessor with a negative opinion (baseline variables), setting all other variables to zero [6]. 58.41

is a very high for the 1/2 threshold. It suggests that a substantial shift is needed to move from score 2 to score 1, while moderate values for other thresholds indicate easier transitions between categories.

The findings from this study emphasize the importance of weekly price in determining scores classification. Higher weekly prices significantly increase the likelihood of having a higher score. On the other hand, the non-significant impact of lessor and opinion dummy variables illustrate that receiving a positive opinion or having an agency lessor does not with confidence increase scores.

4. Conclusion and Limitations

In this study, the OLR analysis was implemented to explore and examine the connection among the rental scores and its drivers. The significant role of weekly price in predicting score was highlighted, while the effect of the opinion and lessor dummy variables remain inconclusive. These findings provide a clear direction for property managers to prioritize pricing strategies in their efforts to achieve higher ratings. Further research is needed to uncover other influential factors that may contribute to the outcome variable's classification. Limitations:

Our dataset only included 2023 data and only one type of property Unit/flat. Also, human error in filling up the rental survey can give us false insights. For example, a person wanted to give a bad review but put down a score 5 or clicked on the wrong score. Also, reviews are subjective and some part of the review could have a negative tone but the person would be overall satisfied with the property. Furthermore, people are more likely to give bad reviews rather than positive ones which could include some bias in our analysis [7]. One major limitation of our analysis is whether the proportional odds assumption (POA) is tenable or not. The POA mandates that the relationship between each pair of outcome groups is the same. Hence, we only have one set of coefficients for each pair of outcome groups. The violation of this assumption can give biased estimates and wrong conclusions. We would need different sets of coefficients to describe relationships [8]. A second analysis risk is linked to the sentiment analysis with Textblob. Indeed, it only describes polarity and subjectivity which could diminish accuracy. Lastly, potential missing relationships and features might be missing. According to [9]. A, the age of a building as an impact on the building quality and depreciation rate, thus, potentially on the perceived quality of a property and its score. Finally, researching on different and larger datasets would reinforce the findings and provide more detailed insights.

References

[1] A. Goh, "Rotting foundations, stained carpets, and black mold: Meet the millennial who's exposing terrible house rentals across Australia," 2023.

[2] T. Pappu, "The mental health toll of Australia's low-quality rental housing," 2024.

[3] J. Praveen Gujjar and H. R. Prasanna Kumar, "Sentiment Analysis:Textblob For Decision Making," *© 2021 IJSRET 1097 International Journal of Scientific Research & Engineering Trends*, vol. 7, no. 2, 2021.

[4] "Regression vs Classification in Machine Learning - Javatpoint."

[5] E. Pinzon, "The Stata Blog probit or logit: ladies and gentlemen, pick your weapon," 2016.

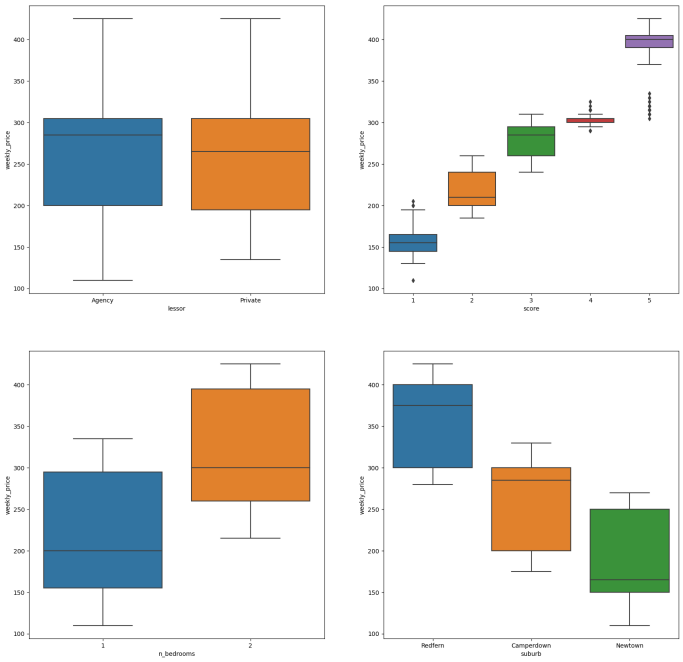
[6] U. of St Andrews, "ANALYSING LIKERT SCALE/TYPE DATA, ORDINAL LOGISTIC REGRESSION EXAMPLE IN R."

[7] nationalstrat, "The Psychology of Customer Reviews | National Strategic Group," June 2018.

[8] UCLA, "Ordinal Logistic Regression | R Data Analysis Examples."

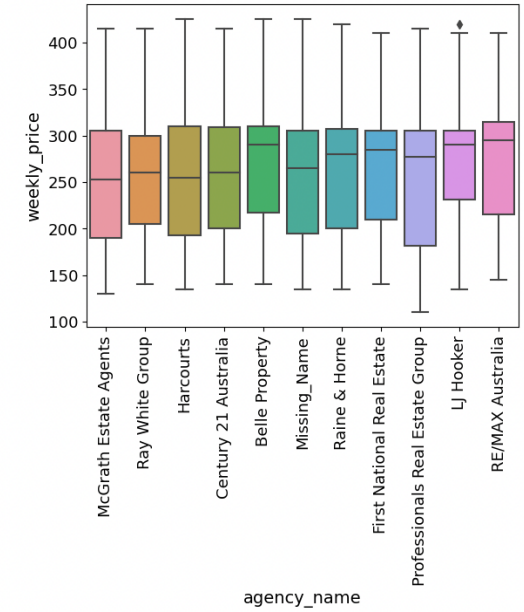
[9] A. Baum, "Quality, depreciation, and property performance," *Journal of Real Estate Research*, vol. 8, no. 4, pp. 541–565, 1993.

Appendix



Appendix 1: weekly price boxplots analysis

Appendix



Appendix 2: weekly_price vs agency_name boxplot