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

SID: 520077942

May 2024

Abstract The property rental market heavily relies on tenant ratings and perceptions, making it crucial to accurately predict the perceived quality of rental properties for informed decision-making. This study employs Ordinal Logistic Regression (OLR) to build prediction models based on extensive tenant reviews from Shitrentals.org, focusing on data from 2023. The analysis reveals that higher weekly rental prices and specific suburbs, particularly Redfern, are significantly associated with higher tenant satisfaction. These findings provide valuable insights for landlords or future property owners on key predictors to enhance rental quality scores and tenant satisfaction.

1 Introduction

The property rental market is significantly influenced by tenant ratings and perceptions, which makes predicting rental property quality crucial for informed decision-making (Australian Bureau of Statistics, 2023). This study aims to predict rental quality scores using Ordinal Logistic Regression (OLR) based on various attributes such as weekly rental prices, number of bedrooms, suburb, and tenant reviews from Shitrentals.org data collected in 2023. By understanding these factors, landlords and property owners can enhance their properties to achieve higher rental quality scores.

To explore the relationship between these independent variables and rental quality scores, a comprehensive Exploratory Data Analysis (EDA) was conducted. Based on the EDA, hypotheses were formulated and tested using multiple OLR models. Through iterative modeling, the analysis was refined to identify the most accurate and reliable model for predicting rental quality scores.

2 Methods

To identify factors influencing perceived rental property quality, exploratory data analysis (EDA) was conducted on the tenant reviews between independent variables and the dependent variable. This allowed for the formulation of hypotheses about what factors might influence the prediction of `score`. There was 203 rows missing for agency names as these were lessors who were private so the missing rows were replaced with 'Private'.

2.1 Explanatory Data Analysis

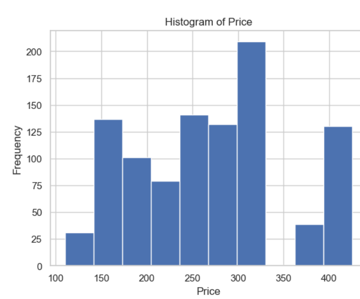


Figure 1: Overview of rental prices

An analysis of the rental price distribution for all properties in the study reveals that the majority of rental prices fall within the range of \$300 to \$325 per week. The highest rental price recorded was \$425 per week.

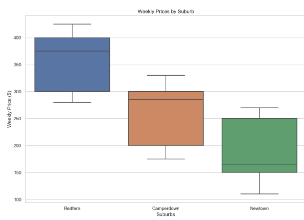


Figure 2: Weekly Price by Suburbs

Within this context, Redfern had the most expensive rentals, with the highest median weekly price, while Newtown had the cheapest rentals, with the lowest median weekly price. This variation in the median and range of prices between suburbs highlights that it could potentially influence rental scores.

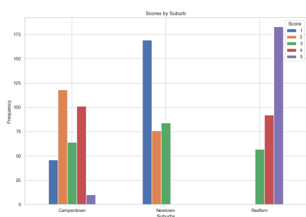


Figure 3: Scores by Suburb

This figure shows the distribution of rental quality scores across the three suburbs. In Camperdown, rentals receive lower to mid-range scores, with a peak at score 2, indicating possible issues in quality or amenities. We would need to analyze the review text to understand this better. Newtown rentals are perceived to be lower quality as their scores range between 1-3 and none of the properties have scored 3 indicating low tenant satisfaction. Redfern has high-quality perceptions, with all the scores above 3, indicating strong tenant satisfaction. This confirms that there is a direct relationship between suburb and score.

Building on the previous analysis of weekly rental prices, there is an observed relationship between suburb, price, and rental quality score. Redfern, with the highest quality scores, also has the highest weekly rent. This suggests a potential correlation between higher rental prices and better perceived quality. It brings to the next question of if the number of bedrooms also has an influence.

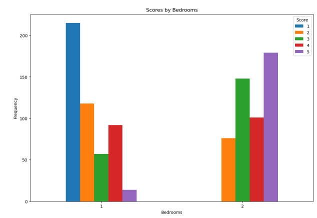


Figure 4: Scores by Number of Bedrooms

This figure shows the distribution of scores based on the number of bedrooms. It highlights that lower scores are more prevalent in 1-bedroom flats, while higher scores are more common in 2-bedroom flats. This suggests a significant difference in perceived quality for the lowest and highest scores based on the number of bedrooms, although the scores between 2 and 4 do not show as much variation.

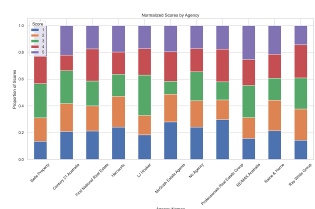


Figure 5: Scores by Agency

While the score distribution across agencies appears balanced, with no single agency dominating, some agencies may be more likely to achieve specific scores. For instance, REMAX Australia is likely to have the score 3 or 5 as it is the most prevalent frequency. However, to fully assess rental property quality, interaction with lessor must be considered and see if it influences tenant experience significantly. Thus, it has been incorporated into the initial model.

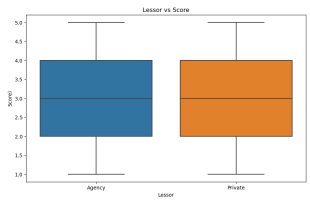


Figure 6: Lessors

The box plot shows that the distribution of scores is similar for properties managed by agencies and private lessors, with both having a median score around 3, indicating that the type of lessor does not significantly affect perceived rental quality and thus, it would likely be used as a control variable.

The sentiment analysis reveals that sentiment values 0 (Negative) and 1 (Positive) have a similar distribution across the scores, with minor variations indicating specific trends in the association of sentiment with certain scores. This indicates that sentiment analysis alone does not significantly influence the score rating.

Based on the analysis from the EDA, the following hypotheses have been formulated:

1. Higher weekly rental prices are linked to higher rental quality scores.

2. Rentals in Redfern exhibit higher quality scores compared to those in Camperdown and Newtown.

3. The sentiment expressed in review texts has a minor impact on rental quality scores, with a slightly higher score trend observed for more positive sentiments, but it is not a significant influencing factor.

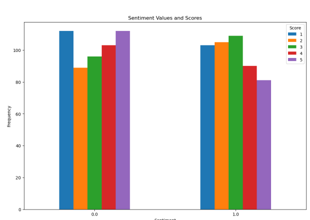


Figure 7: Sentiment Analysis

4. 2-bedroom rentals are more likely to achieve higher quality scores compared to 1-bedroom rentals.

2.2 Modelling Rental Quality Score with Ordinal Logistic Regression

As rental quality score ranges from 1 to 5 which is ordinal categories, OLR has been chosen for modelling. OLR can handle both continuous and categorical predictors, making it ideal for this dataset.

2.3 Assumptions of OLR

1. Linearity: The relationship between independent and dependent variables is linear.

2. Independence: Observations are independent of each other.

3. Homoscedasticity: Residuals have constant variance across all levels of the independent variables.

4. No Perfect Multicollinearity: Independent variables are not perfectly correlated.

5. Normality of Residuals: Residuals are normally distributed.

6. Exogeneity: Independent variables are uncorrelated with the error term.

2.4 Variables

In this study, the lessor variable was used as a control due to the minimal differences observed in rental quality scores across these categories. Therefore, to simplify the model and focus on variables with more substantial influence, we excluded the lessor and agency name from our analysis in all subsequent models, except for the first model, to assess their significance.

2.5 Iterative Modelling

In OLR Model 1 (Figure 12), high standard errors relative to coefficient values for the number of bedrooms and suburbs suggest multicollinearity with another predictor (Grace-Martin, 2019). This indicates redundant information, making it difficult to isolate the unique effect of each variable on the dependent variable. Consequently, the model can create a false sense of confidence in the predictions, leading to high precision in the coefficients while simultaneously introducing bias, thereby making the model inaccurate overall (Farrar & Glauber, 1967). To address this issue, the suburb variable was excluded in the subsequent model (Figure 14). This decision aligns with the challenges identified during Exploratory Data Analysis (EDA), where the strong association between weekly price and suburbs was found to be problematic for OLR analysis.

	weekly_price	n_bedrooms	suburb_Newtown	suburb_Redfern
weekly_price	1.000000	0.630756	-0.589853	0.729361
n_bedrooms	0.630756	1.000000	-0.024758	0.007101
suburb_Newtown	-0.589853	-0.024758	1.000000	-0.493648
suburb_Redfern	0.729361	0.007101	-0.493648	1.000000

Figure 8: Correlation Matrix between Price and Suburb

Through the correlation matrix, it can be confirmed there is a relationship between suburb and weekly price. This means that properties located in Redfern are associated with higher weekly rental prices while Newtown is associated with lower weekly rental prices, both relative to Camperdown. Thus, Model 1 (Figure 12) is not considered due to its high AIC (309.2) and BIC (392.8), which indicate the poorest fit to the data among the four models, in addition to the aforementioned issues with multicollinearity.

Model 2 (see Figure 14), while slightly better than Model 1 with an AIC of 305.2 and BIC of 380.0, still does not compare favorably with Models 3 (see Figure 9) and 4 (see Figure 16). It has the same accuracy (0.85) (see Figure 15) and similar precision, recall, and f1-score as Model 1, making it less competitive.

Model 3 was selected for the following reasons:

- **Balance of Fit and Performance:** Despite Model 3 having slightly higher AIC and BIC values compared to Model 4, it demonstrates the highest accuracy (0.87) and highest precision across all 5 outcomes and strong performance metrics.

OrderedModel Results

Dep. Variable:	score	**Log-Likelihood:**	-139.33
Model:	OrderedModel	**AIC:**	292.7
Method:	Maximum Likelihood	**BIC:**	323.4
Date:	Wed, 22 May 2024		
Time:	12:40:54		
No. Observations:	600		
Df Residuals:	593		
Df Model:	3		

	coef	std err	z	P>|z|	[0.025	0.975]
weekly_price	0.3460	0.034	10.163	0.000	0.279	0.413
n_bedrooms	-0.3259	0.399	-0.817	0.414	-1.108	0.456
sentiment_1.0	0.1586	0.311	0.510	0.610	-0.451	0.769
1/2	65.7368	6.607	9.949	0.000	52.787	78.687
2/3	3.0051	0.103	29.127	0.000	2.803	3.207
3/4	2.7836	0.102	27.170	0.000	2.583	2.984
4/5	1.9055	0.097	19.662	0.000	1.716	2.095

Figure 9: Selected Model 3

- **Model Fit vs. Accuracy:** In ordinal logistic regression, it is essential to strike a balance between model fit and accuracy. A model with excellent fit statistics but poor performance is not practical, and vice versa. Model 3 successfully achieves this balance, making it a reliable choice for practical application.

	precision	recall	f1-score	support
0	0.98	0.93	0.95	43
1	0.85	0.91	0.88	32
2	0.78	0.84	0.81	51
3	0.82	0.74	0.78	43
4	0.97	0.97	0.97	31
accuracy			0.87	200
macro avg	0.88	0.88	0.88	200
weighted avg	0.87	0.87	0.87	200

Figure 10: Accuracy of Model 3

The precision of Model 3 is the highest for all 5 ordinal outcomes among all models, making it the best suitable model for this prediction. In this, the predictor variables are weekly price, number of bedrooms, and sentiment. Model 4, while having the best fit statistics with the lowest AIC (291.3) and BIC (317.7), has slightly lower accuracy (0.86) (see Figure 17) compared to Model 3. Although its performance metrics are similar, the slightly better accuracy and f1-score of Model 3 make it the preferred choice.

3 Results and Discussion

Model 3 was chosen due to its highest precision level for all ordinal outcomes, providing the most accurate and reliable results. The ordinal logistic regression model for predicting rental quality scores can be expressed as:

$$\text{logit}[P(Y \leq j)] = \alpha_j - \beta X$$

Where:

- Y is the ordinal rental quality score, ranging from 1 to 5.
- $j \in \{1, 2, 3, 4\}$ represents the levels of the ordinal outcome variable, excluding the highest level J (which is 5 in this case).
- α_j are the intercepts (thresholds) corresponding to each level j of the ordinal outcome variable.
- β are the coefficients for the predictors X, which include variables such as weekly rental price, suburb, number of bedrooms, and sentiment of review text.

Coefficients of Predictors

- **Weekly Price**: The coefficient for *weekly_price* is 0.3460 and is statistically significant (p ¡ 0.05). This suggests that as the weekly price increases, the log-odds of being in a higher category of the outcome variable (*score*) also increase. Specifically, for each one-unit increase in *weekly_price*, the log-odds of a higher *score* increase by 0.3460.
- **Number of Bedrooms (*n_bedrooms*)**: The coefficient for *n_bedrooms* is -0.3259, which is not statistically significant (p ¿ 0.05).
- **Sentiment (*sentiment_1.0*)**: The coefficient for *sentiment_1.0* is 0.1586, which is not statistically significant (p ¿ 0.05). This suggests that sentiment and number of bedrooms does not have a significant effect on the log-odds of the outcome variable (*score*).

Although the coefficients for *n_bedrooms* and *sentiment_1.0* are not statistically significant, it is beneficial to include them in the model for the following reasons:

- **EDA Analysis**: Through this, it has been identified these variables to have potential influence on the quality scores and thus, it is better to include them (Beyer, 1981).
- **Model Completeness**: Excluding variables with potential effects, even if not statistically significant in this model, can lead to omitted variable bias and distort other coefficient estimates. Additionally, relying solely on statistically significant variables like *weekly_price* may result in an underfitted model that fails to capture the data's complexity (Harrell, 2015).
- **Future Research**: These variables may become significant with larger sample sizes or different data, and their inclusion provides a basis for future research comparisons (Box & Draper, 1987).

Thresholds in Ordinal Logistic Regression (OLR)

Thresholds (or cut points) in OLR are essential for understanding how the probability of falling into or below a certain category of the dependent variable changes with predictors (X).

The thresholds (α_j) are:

- **Threshold 1/2**: 65.7368 (Standard Error: 6.607, z-value: 9.949, p-value: 0.000)
- **Threshold 2/3**: 3.0051 (Standard Error: 0.103, z-value: 29.127, p-value: 0.000)
- **Threshold 3/4**: 2.7836 (Standard Error: 0.102, z-value: 27.170, p-value: 0.000)
- **Threshold 4/5**: 1.9055 (Standard Error: 0.097, z-value: 19.662, p-value: 0.000)

These thresholds show the differentiation between score categories. A high threshold value like 65.7368 suggests a significant increase in log-odds is needed to move from a score of 1 to 2. In contrast, lower thresholds indicate less pronounced differentiation.

Calculation with Predictor: Weekly Price

Given a significant predictor, *weekly_price*, with a coefficient of 0.3460:

For a property with a current weekly rental price of \$300 and a rental quality score of 1:

$$\Delta \text{log-odds} = 65.7368$$

$$\Delta \text{weekly\_price} = \frac{65.7368}{0.3460} \approx 190.03$$

Thus, to move from score 1 to 2, the weekly price needs to increase by approximately \$190, making the new weekly rental price \$490.03. This indicates the difficulty of improving the score from 1 to 2, reflecting the model's threshold values.

To achieve a higher score, property owners can make enhancements through:

- **Property Upgrades**: Renovations, modern appliances.
- **Enhanced Amenities**: Adding in-unit or building amenities.
- **Furnishings and Fixtures**: Upgrading furniture and fixtures.
- **Maintenance and Safety**: Regular maintenance and safety improvements as this was the most prevalent theme in the reviews.

Interpreting Log-Odds

For a specific weekly price, the model equation can be used to find log-odds. For instance, with a weekly price of \$10:

$$\text{logit}[P(Y \leq 1)] = 65.7368 - (0.3460 \times 10) = 62.2768$$

Converting log-odds to probability:

$$P(Y \leq 1) = \frac{e^{62.2768}}{1 + e^{62.2768}}$$

Given the high value of $e^{62.2768}$, $P(Y \leq 1) \approx 1$, indicating a nearly 100% probability of a score of 1 with a weekly price of \$10. This suggests low rental prices are strongly associated with the lowest rental quality scores, highlighting the need for significant price and quality improvements to achieve higher scores.

3.1 Conclusion and Limitations

This study employed ordinal logistic regression (OLR) to predict perceived rental property quality using data from Shitrentals.org. Key factors influencing tenant satisfaction were identified, including weekly rental price, number of bedrooms, and sentiment in reviews. Model 3 emerged as the most suitable, balancing model fit and performance with an AIC of 292.7, BIC of 323.4, and the highest accuracy (0.87). Higher weekly rental prices were positively associated with higher perceived quality scores, suggesting that better amenities and investments lead to greater tenant satisfaction. Although the number of bedrooms and sentiment were not statistically significant, their inclusion provided a comprehensive understanding of factors affecting rental quality. The study offers actionable recommendations for landlords: investing in property upgrades, enhanced amenities, and regular maintenance can boost rental quality scores, potentially increasing rental income and tenant satisfaction.

Despite these valuable findings, this study has several limitations:

- **Data Source Bias:** Data from Shitrentals.org may not represent the entire rental market, and tenant reviews could be biased. The dataset used is only focused on 3 suburbs.
- **Sample Size:** While sufficient for this analysis, the sample size might not capture all variations in rental property quality and tenant satisfaction. Larger datasets could yield more robust results. Redfern had the largest sample size which can make the results more skewed towards Redfern.
- **Variable Significance:** Some variables, like the number of bedrooms and sentiment, were not statistically significant, possibly due to non-representative sample size.
- **Data Source Date:** The study uses data from 2023, and rental market conditions can change

over time. Longitudinal studies could provide deeper insights into how these factors influence rental quality perceptions.

Addressing these limitations in future research could enhance the understanding of factors affecting rental property quality and provide even more actionable insights for landlords and property managers.

4 Bibliography

References

[1] Australian Bureau of Statistics. (2023, April 24). *New insights into the rental market — Australian Bureau of Statistics*. Www.abs.gov.au. Retrieved from https://www.abs.gov.au/statistics/detailed-methodology-information/information-papers/new-insights-rental-market

[2] Beyer, H. (1981). Tukey, John W.: Exploratory Data Analysis. Addison-Wesley Publishing Company Reading, Mass. — Menlo Park, Cal., London, Amsterdam, Don Mills, Ontario, Sydney 1977, XVI, 688 S. *Biometrical Journal*, *23*(4), 413–414. https://doi.org/10.1002/bimj.4710230408

[3] Box, G. E. P., & Draper, N. R. (1987). *Empirical Model-Building and Response Surfaces*. Wiley. Retrieved from https://books.google.com.au/books/about/Empirical_Model_Building_and_Response_Su.html?id=QO2dDRufJEAC\&redir_esc=y

[4] Farrar, D. E., & Glauber, R. R. (1967). Multicollinearity in Regression Analysis: The Problem Revisited. *The Review of Economics and Statistics*, *49*(1), 92. https://doi.org/10.2307/1937887

[5] Grace-Martin, K. (2019, February 25). *Eight Ways to Detect Multicollinearity - The Analysis Factor*. The Analysis Factor. Retrieved from https://www.theanalysisfactor.com/eight-ways-to-detect-multicollinearity/

[6] Harrell, F. (2015). *Regression Modeling Strategies: With Applications to Linear Models, Logistic and Ordinal Regression, and Survival Analysis* (2nd ed.). Springer. Retrieved from https://warin.ca/ressources/books/2015_Book_RegressionModelingStrategies.pdf

5 Appendices

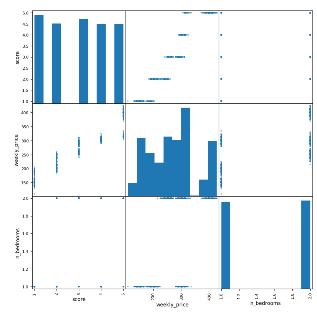


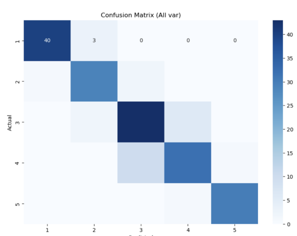
Figure 11: Numerical Variables as a Whole

OrderedModel Results

Dep. Variable:	score	**Log-Likelihood:**	-135.61
Model:	OrderedModel	**AIC:**	309.2
Method:	Maximum Likelihood	**BIC:**	392.8
Date:	Wed, 22 May 2024		
Time:	12:40:53		
No. Observations:	600		
Df Residuals:	581		
Df Model:	15		

	coef	std err	z	P>|z|	[0.025	0.975]
weekly_price	0.3533	0.035	10.033	0.000	0.284	0.422
n_bedrooms	21.9368	406.032	0.054	0.957	-773.870	817.744
suburb_Newtown	-14.3298	324.355	-0.044	0.965	-650.054	621.395
suburb_Redfern	22.3061	406.033	0.055	0.956	-773.505	818.117
agency_name_Century 21 Australia	-0.8463	0.821	-1.031	0.303	-2.455	0.762
agency_name_First National Real Estate	-0.0484	0.808	-0.060	0.952	-1.632	1.535
agency_name_Harcourts	-1.1027	0.896	-1.230	0.219	-2.859	0.654
agency_name_LJ Hooker	-0.3495	0.796	-0.439	0.661	-1.909	1.210
agency_name_McGrath Estate Agents	-0.4708	0.764	-0.616	0.538	-1.969	1.028
agency_name_No Agency	-0.2899	0.702	-0.413	0.680	-1.665	1.086
agency_name_Professionals Real Estate Group	0.6024	0.854	0.705	0.481	-1.072	2.276
agency_name_RE/MAX Australia	-0.8685	0.928	-0.936	0.349	-2.688	0.950
agency_name_Raine & Horne	-0.7878	0.806	-0.978	0.328	-2.367	0.792
agency_name_Ray White Group	0.1140	0.815	0.140	0.889	-1.483	1.711
sentiment_1.0	0.1369	0.317	0.431	0.666	-0.485	0.759
1/2	88.8925	406.045	0.219	0.827	-706.942	884.727
2/3	3.3574	8.506	0.395	0.693	-13.315	20.030
3/4	3.4324	10.480	0.328	0.743	-17.108	23.973
4/5	1.8994	0.097	19.610	0.000	1.710	2.089

Figure 12: OLR Model 1



	precision	recall	f1-score	support
0	0.98	0.93	0.95	43
1	0.85	0.91	0.88	32
2	0.78	0.84	0.81	51
3	0.82	0.74	0.78	43
4	0.97	0.97	0.97	31
accuracy			0.87	200
macro avg	0.88	0.88	0.88	200
weighted avg	0.87	0.87	0.87	200

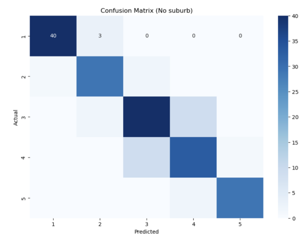
Figure 13: Model 1 Accuracy and Fit

OrderedModel Results

Dep. Variable:	score	**Log-Likelihood:**	-135.61
Model:	OrderedModel	**AIC:**	305.2
Method:	Maximum Likelihood	**BIC:**	380.0
Date:	Wed, 22 May 2024		
Time:	12:40:54		
No. Observations:	600		
Df Residuals:	583		
Df Model:	13		

	coef	std err	z	P>|z|	[0.025	0.975]
weekly_price	0.3535	0.035	10.053	0.000	0.285	0.422
n_bedrooms	-0.3691	0.417	-0.885	0.376	-1.186	0.448
agency_name_Century 21 Australia	-0.8462	0.821	-1.031	0.303	-2.455	0.763
agency_name_First National Real Estate	-0.0499	0.808	-0.062	0.951	-1.634	1.534
agency_name_Harcourts	-1.1024	0.896	-1.230	0.219	-2.859	0.655
agency_name_LJ Hooker	-0.3496	0.796	-0.439	0.660	-1.910	1.210
agency_name_McGrath Estate Agents	-0.4705	0.765	-0.615	0.538	-1.969	1.028
agency_name_No Agency	-0.2898	0.702	-0.413	0.680	-1.666	1.086
agency_name_Professionals Real Estate Group	0.6033	0.854	0.706	0.480	-1.071	2.278
agency_name_RE/MAX Australia	-0.8687	0.928	-0.936	0.349	-2.688	0.951
agency_name_Raine & Horne	-0.7876	0.806	-0.977	0.328	-2.367	0.792
agency_name_Ray White Group	0.1141	0.815	0.140	0.889	-1.483	1.712
sentiment_1.0	0.1368	0.317	0.431	0.666	-0.485	0.759
1/2	66.6262	6.824	9.764	0.000	53.252	80.001
2/3	3.0325	0.105	28.961	0.000	2.827	3.238
3/4	2.8112	0.104	27.075	0.000	2.608	3.015
4/5	1.8998	0.097	19.633	0.000	1.710	2.089

Figure 14: OLR Model 2



	precision	recall	f1-score	support
0	0.98	0.93	0.95	43
1	0.85	0.91	0.88	32
2	0.78	0.78	0.78	51
3	0.75	0.77	0.76	43
4	0.97	0.94	0.95	31
accuracy			0.85	200
macro avg	0.87	0.86	0.86	200
weighted avg	0.86	0.85	0.86	200

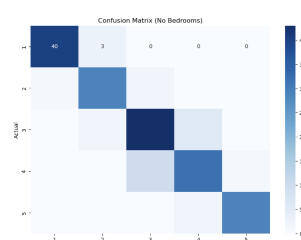
Figure 15: Model's 2 Accuracy and Fit

OrderedModel Results

Dep. Variable:	score	**Log-Likelihood:**	-139.66
Model:	OrderedModel	**AIC:**	291.3
Method:	Maximum Likelihood	**BIC:**	317.7
Date:	Wed, 22 May 2024		
Time:	12:40:55		
No. Observations:	600		
Df Residuals:	594		
Df Model:	2		

	coef	std err	z	P>|z|	[0.025	0.975]
weekly_price	0.3475	0.034	10.160	0.000	0.280	0.415
sentiment_1.0	0.1777	0.310	0.574	0.566	-0.429	0.785
1/2	66.3632	6.615	10.033	0.000	53.399	79.328
2/3	3.0252	0.099	30.480	0.000	2.831	3.220
3/4	2.7794	0.103	26.955	0.000	2.577	2.982
4/5	1.9047	0.097	19.624	0.000	1.714	2.095

Figure 16: OLR Model 4



	precision	recall	f1-score	support
0	0.98	0.93	0.95	43
1	0.85	0.91	0.88	32
2	0.78	0.84	0.81	51
3	0.80	0.74	0.77	43
4	0.97	0.94	0.95	31
accuracy			0.86	200
macro avg	0.88	0.87	0.87	200
weighted avg	0.87	0.86	0.87	200

Figure 17: Model 4's Accuracy and Fit