Analysis of Hotels In München Using Different Regression Techniques

This study aimed to explore the correlation between the binary variable "highly_rated" and key factors such as stars, distance, rating review count and price in Munich hotels. The analysis found that a simpler Linear Probability Model performed similarly to more complex models in predicting hotel ratings based on attributes.

Part 1: Data Processing:

Utilizing two datasets from OSF hotels-europe, we merged hotel features and prices based on 'hotel_id' for a comprehensive analysis. We filtered the whole dataset for Munich only, this city was selected as it aligned with the task description of selecting a city in the top 10 most observations. We created the binary variable "highly_rated" (1 for ratings over 4, 0 otherwise), which will serve for us to check various key variables against. This refined dataset will then be utilized to examine the correlations between hotel attributes and their ratings.

Part 2: Analysis of our chosen dependants

Here distribution analyses were created for the dependant variables, namely: Price, Distance, Rating, Stars. These revealed certain patterns in the data that where noteworthy, one of these was acted upon, namely Price and Distance revealed the presence of extreme values. We created a summary of our statistical measures to further evaluate the chosen variables [Please See Figure 1 in the Appendix].

Part 3: Model Estimation

Three regression models were employed: Linear Probability Model (LPM), Logit and the Probit model, with the aim of comparing their predictions. To make them comparable at a later stage, the marginal effects of the logit and the probit models where obtained. For the LPM, a histogram was obtained to show the distribution of Predicted Probabilities [Please See Figure 2 in the Appendix].

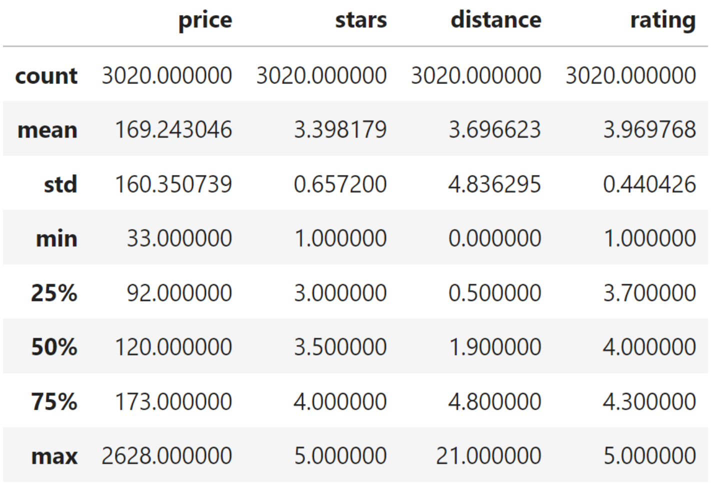
Part 4: Analysis and Interpretation

The predicted probabilities from the logistic and probit models generally concur with those from the Linear Probability Model (LPM), as most data points are distributed closely around the diagonal line representing agreement between the models. There is no clear evidence of systematic bias in either the logistic or probit model compared to the LPM, indicating that both non-linear models provide similar probability predictions for the outcome. The distribution of points across the plot suggests that both logistic and probit models align with the LPM across the full range of predicted probabilities, with no significant divergence at the extremes, which implies consistent model performance [Please See Figure 3 in the Appendix].

The differences in the marginal effects across the Linear Probability Model (LPM), Logistic, and Probit models are generally minimal for all the variables. This indicates that, for this dataset and these specific predictors, choosing between the LPM, Logistic, and Probit models does not lead to significantly different estimations. The direction of the effects for each predictor variable is consistent across all three models. The small magnitude of the differences in marginal effects between the models suggests that the linear approximation provided by the LPM is quite close to the non-linear estimations of the Logistic and Probit models. This similarity indicates that the LPM can serve as a reasonable approximation for the effects of these predictors, despite its simplicity compared to the Logistic and Probit models [Please See Figure 4 in the Appendix].

Appendix:

Figure 1:



	price	stars	distance	rating
count	3020.000000	3020.000000	3020.000000	3020.000000
mean	169.243046	3.398179	3.696623	3.969768
std	160.350739	0.657200	4.836295	0.440426
min	33.000000	1.000000	0.000000	1.000000
25%	92.000000	3.000000	0.500000	3.700000
50%	120.000000	3.500000	1.900000	4.000000
75%	173.000000	4.000000	4.800000	4.300000
max	2628.000000	5.000000	21.000000	5.000000

Figure 2:

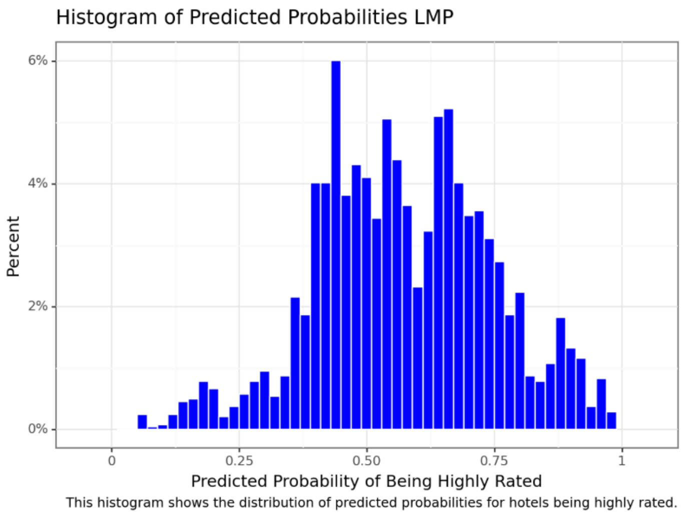


Figure 3:

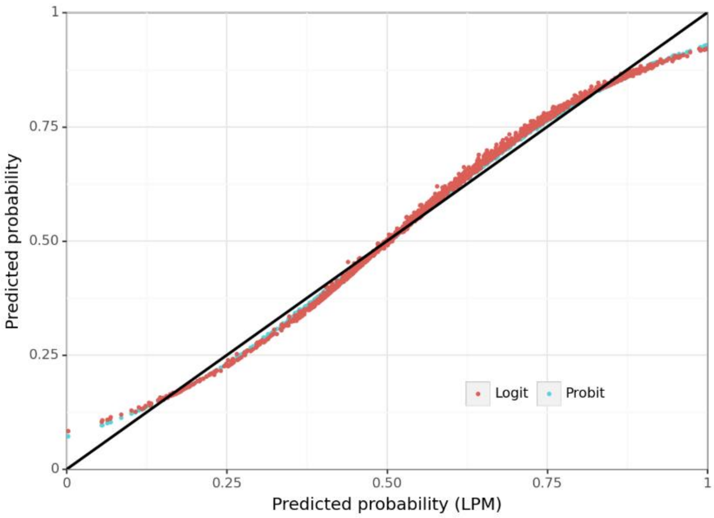
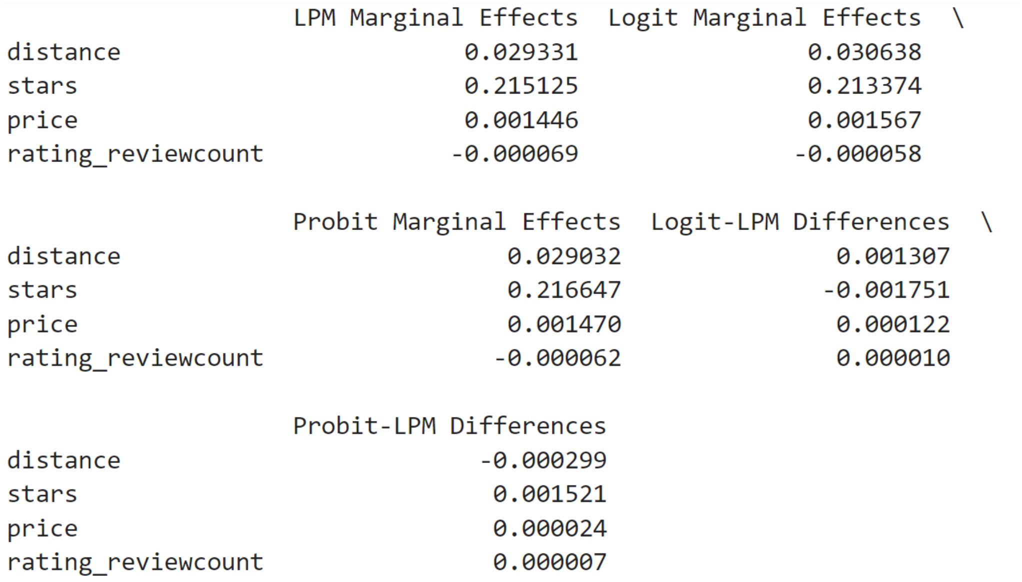


Figure 4:



```
                     LPM Marginal Effects  Logit Marginal Effects  \
distance                         0.029331                0.030638
stars                            0.215125                0.213374
price                            0.001446                0.001567
rating_reviewcount              -0.000069               -0.000058

                     Probit Marginal Effects  Logit-LPM Differences  \
distance                            0.029032               0.001307
stars                               0.216647              -0.001751
price                               0.001470               0.000122
rating_reviewcount                 -0.000062               0.000010

                     Probit-LPM Differences
distance                          -0.000299
stars                              0.001521
price                              0.000024
rating_reviewcount                 0.000007
```