Determinants of High Hotel Ratings: Using data about Berlin hotels

by: Alina Imanakhunova

---

1. Introduction and Sample Description

The analysis sample consists of **559 hotels located in Berlin**, drawn from the *hotels-europe* dataset after applying cleaning and selection criteria. The dataset combines hotel features with pricing information, allowing for an examination of factors associated with high customer ratings in a major European city.

The dependent variable in the analysis is a binary indicator, `highly_rated`, which equals 1 if a hotel's average user rating is **at least 4 out of 5**, and 0 otherwise. In the final sample, approximately **56.4% of hotels are classified as highly rated**, indicating a relatively high overall level of customer satisfaction among Berlin hotels.

Key explanatory variables include **distance to the city center (in kilometers)** and **hotel star classification**, which serve as proxies for location attractiveness and service quality, respectively. The average hotel in the sample is located **2.80 km from the city center** and has **3.29 stars**, suggesting that most hotels are mid-range and centrally accessible.

Hotel prices vary substantially across the sample. The median-based hotel-level price measure has a mean of **€128.28 per night**, with a right-skewed distribution and a small number of high-price observations exceeding **€1,500**. To avoid overweighting hotels with many price observations, prices were aggregated to the hotel level using the **median price** across booking conditions. No further transformation of prices is used in the main analysis, as price serves only as a descriptive characteristic rather than a dependent variable.

Overall, the sample is sufficiently large and heterogeneous to support econometric analysis using both linear and nonlinear probability models. The binary nature of the outcome variable motivates the use of **Linear Probability, Logit, and Probit models**, allowing for a comparison of results across alternative estimation approaches.

---

2. Data and Cleaning Decisions

This dataset focuses on hotels located in Berlin. The original Berlin sample contains **579 hotels** in the features dataset. Hotel characteristics were merged with the price dataset using a unique hotel identifier, `hotel_id`.

Hotels with missing or implausible values were removed. Specifically, ratings were restricted to the interval $(0, 5]$, star classifications to $(0, 5]$, and distance to the city center to non-negative values. After cleaning, the final analysis sample consists of **559 hotels**.

A binary outcome variable was constructed as follows:

$$highly\_rated_i = \begin{cases} 1 & \text{if } rating_i \geq 4 \\ 0 & \text{otherwise} \end{cases}$$

---

3. Descriptive Statistics

Approximately **56.4%** of Berlin hotels are classified as highly rated, indicating a relatively high overall quality level. The average hotel is located **2.8 km** from the city center and has slightly above **3 stars**.

Variable	Mean	Std. Dev.	Min	Max
Highly rated	0.564	0.496	0	1
Rating	3.92	0.55	1.0	5.0
Distance (km)	2.80	2.11	0.3	14.0
Stars	3.29	0.80	1.0	5.0
Price (EUR)	128.3	125.1	39	1546

The distribution of prices is right-skewed, while ratings cluster around higher values, consistent with a competitive hotel market in a major European capital.

---

4. Econometric Models

The probability that a hotel is highly rated is modeled as a function of distance to the city center and hotel star classification:

$$P(highly\_rated_i = 1) = f(distance_i, stars_i)$$

Three econometric models are estimated:

1. **Linear Probability Model (LPM)**

$$P(\text{highly\_rated}_i = 1) = \alpha + \beta_1 \, \text{distance}_i + \beta_2 \, \text{stars}_i$$

2. **Logit model**

$$\text{logit}\big(P(\text{highly\_rated}_i = 1)\big) = \alpha + \beta_1 \, \text{distance}_i + \beta_2 \, \text{stars}_i$$

3. **Probit model**

$$\text{probit}\big(P(\text{highly\_rated}_i = 1)\big) = \alpha + \beta_1 \, \text{distance}_i + \beta_2 \, \text{stars}_i$$

The LPM provides a simple and easily interpretable benchmark, while logit and probit models account for the binary nature of the dependent variable by constraining predicted probabilities to the $[0, 1]$ interval.

---

5. Regression Results

Logit and Probit Estimates

Variable	Logit Coef.	Logit ME	Probit Coef.	Probit ME
Distance	-0.036	-0.009	-0.024	-0.009
Stars	1.356***	0.330***	0.813***	0.317***
Observations	559		559	

*** p < 0.01

Distance to the city center has a negative but statistically insignificant effect across all models. In contrast, hotel star classification is strongly and positively associated with the probability of being highly rated.

---

6. Interpretation and Model Comparison

Marginal effects from the nonlinear models indicate that an additional star increases the probability of a hotel being highly rated by approximately **32–33 percentage points**, holding distance constant. This effect is economically large and highly statistically significant.

The similarity of marginal effects between the logit and probit models reflects the well-known scaling difference between these specifications rather than substantive disagreement. Average predicted probabilities are nearly identical across models:

Model	Mean Predicted Probability
LPM	0.5635
Logit	0.5635
Probit	0.5678

These values closely match the observed share of highly rated hotels in the data.

---

7. Conclusion

The results show that **hotel quality**, as measured by star classification, is the primary determinant of high user ratings in Berlin (see Chart.1 in Appendix). Distance from the city center plays a limited role once service quality is taken into account.(see Chart.2 in Appendix)

The consistency of results across linear and nonlinear probability models strengthens confidence in the findings. Overall, the analysis highlights the importance of **service quality over location** in explaining high customer satisfaction among hotels in Berlin.

Appendix

The appendix includes the full notebook code that i will upload together with this report, complete regression outputs, and supplementary figures. All results are fully reproducible using the provided Jupyter Notebook.

Chart.1

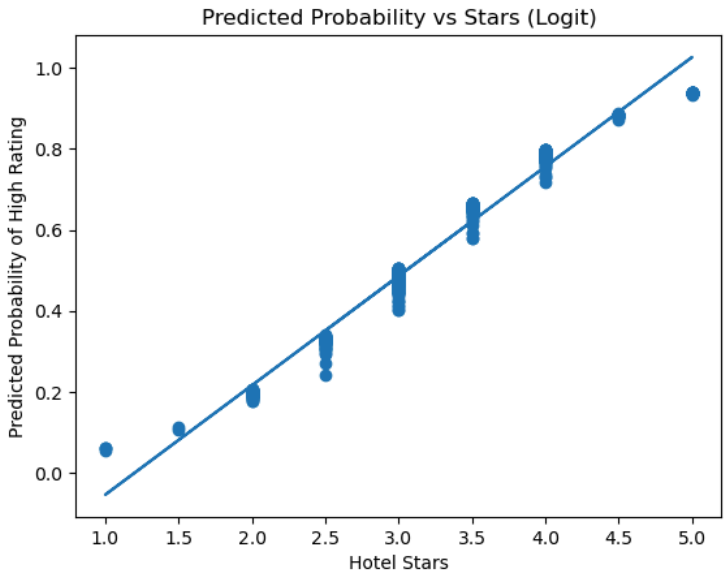
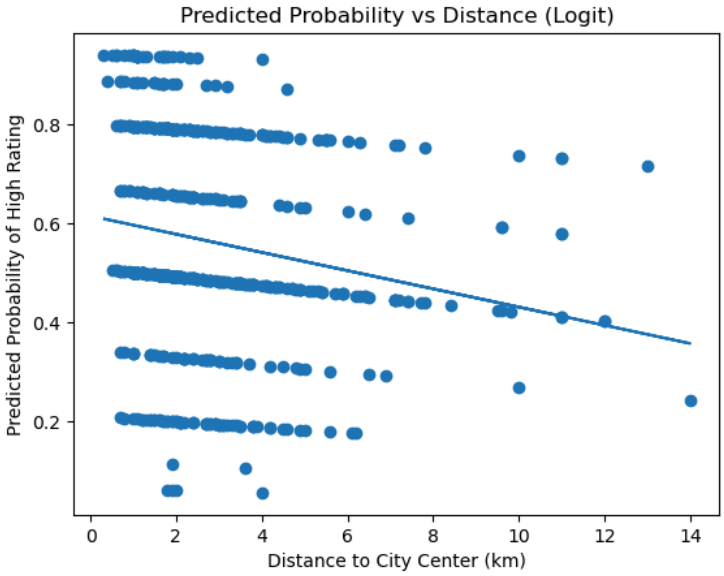


Chart.2


Predicted Probability vs Distance (Logit)
Predicted Probability of High Rating
0.2
0.4
0.6
0.8
0
2
4
6
8
10
12
14
Distance to City Center (km)