Data Analysis 2 Assignment 2
Submitted by: Zariza Chowdhury (ID: 2500086)

Determinants of High Hotel Ratings in Paris

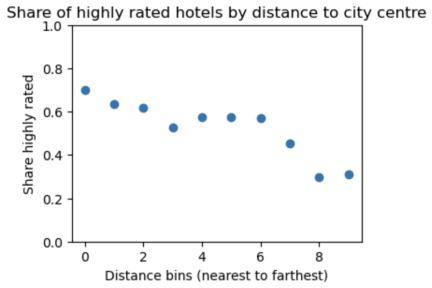
Data and Sample Construction

I combined hotel features and price data from the hotels-europe project to study the determinants of high user ratings. To ensure a consistent hotel-level unit of observation, I filtered prices to a baseline booking condition (one night, weekday, non-holiday). Paris was selected as it has more than 250 hotels after filtering. After removing observations with missing values in rating, stars, and distance to the city centre, the final sample consists of 1,440 hotels. A hotel is classified as highly rated if its average user rating is at least 4.

Descriptive Evidence

About 53% of hotels are highly rated. Highly rated hotels are closer to the city centre, have higher star ratings, and charge higher prices on average. Bin-scatter plots show that the share of highly rated hotels decreases with distance and increases sharply with star classification.

Fig 1: Bin-scatter plot of share of highly rated hotels by distance to city centre



Regression Models

I estimated a linear probability model, a logit model, and a probit model using distance and hotel stars as explanatory variables. Across all models, distance is negatively associated with being highly rated, while star ratings are positively associated. The linear probability model implies a 2.1 percentage point decrease in probability per kilometre and a 28.8 percentage point increase per additional star, with logit and probit marginal effects showing similar magnitudes.

Table 1: Regression results: Logit model with marginal effects at the mean

Logit Marginal Effects

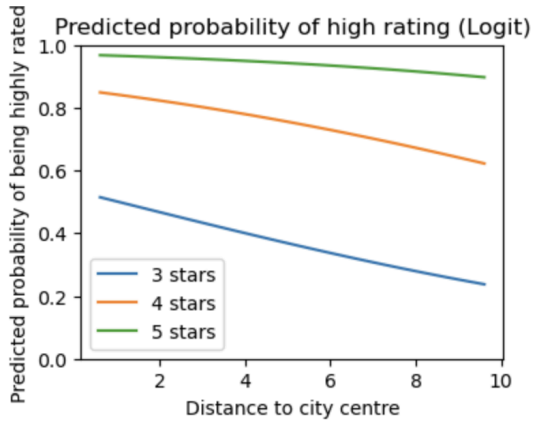
Dep. Variable:	highly_rated					
Method:	dydx					
At:	mean					
	dy/dx	std err	z	P>|z|	[0.025	0.975]
distance	−0.0338	0.006	−5.870	0.000	−0.045	−0.023
stars	0.4144	0.025	16.310	0.000	0.365	0.464

Predicted Probabilities and Conclusion

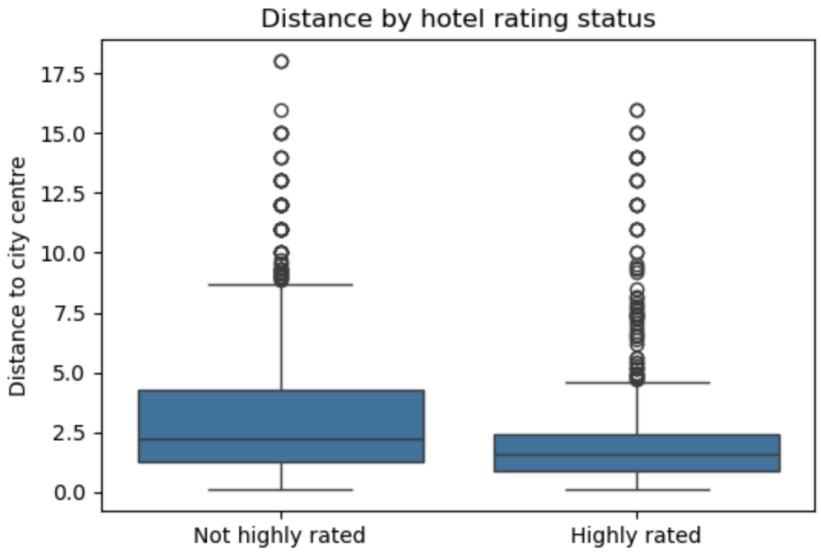
Predicted probabilities from the logit model show that the likelihood of being highly rated declines with distance for all star categories, while higher-star hotels consistently have higher predicted probabilities.

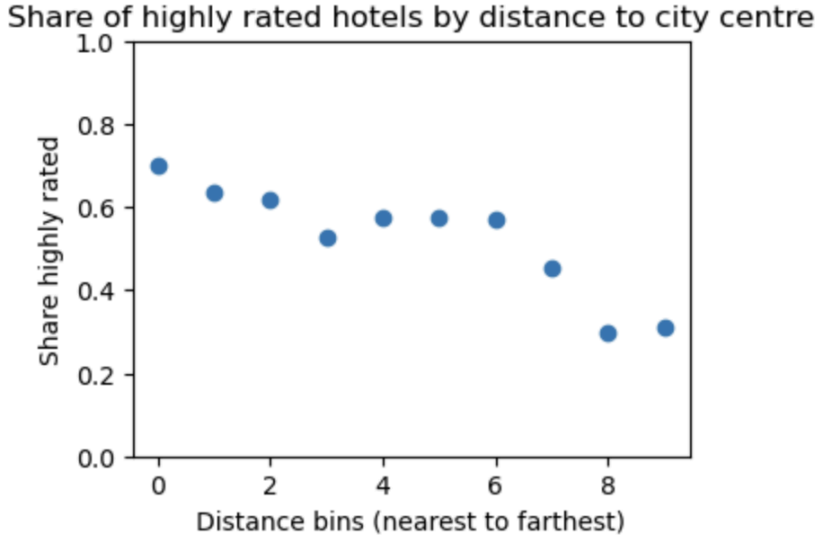
Overall, the analysis shows strong descriptive associations between hotel location, star ratings, and user evaluations. While the results are not causal, they suggest that proximity to the city centre and formal quality ratings align closely with user ratings.

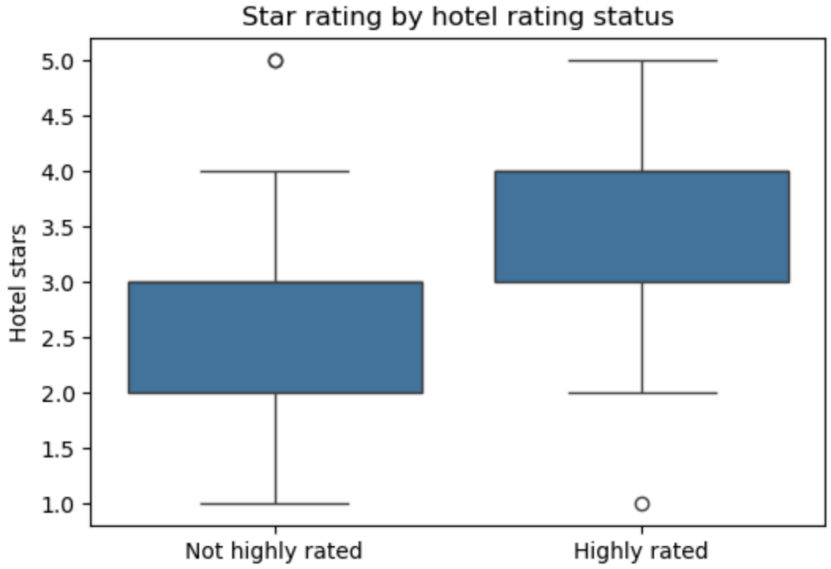
Fig 2: Predicted probability of being highly rated by distance and star rating (Logit model)

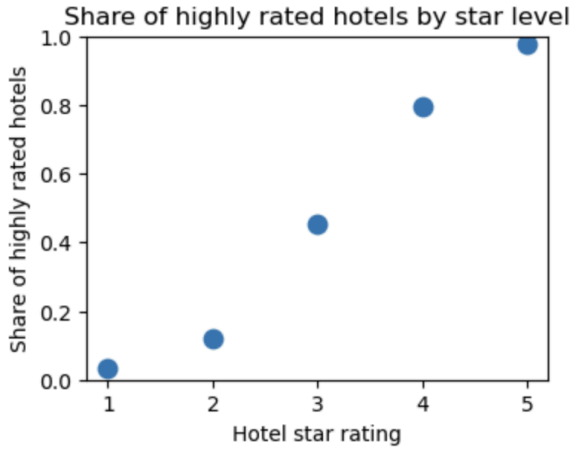


Appendix


Distance by hotel rating status
Distance to city centre
17.5
15.0
12.5
10.0
7.5
5.0
2.5
0.0
Not highly rated
Highly rated


Share of highly rated hotels by distance to city centre
Share highly rated
1.0
0.8
0.6
0.4
0.2
0.0
0
2
4
6
8
Distance bins (nearest to farthest)


Star rating by hotel rating status
Hotel stars
5.0
4.5
4.0
3.5
3.0
2.5
2.0
1.5
1.0
Not highly rated
Highly rated



Regression Result of Linear Probability Model:

OLS Regression Results

Dep. Variable:	highly_rated	R-squared:	0.272
Model:	OLS	Adj. R-squared:	0.271
Method:	Least Squares	F-statistic:	497.1
Date:	Sun, 21 Dec 2025	Prob (F-statistic):	8.23e-165
Time:	11:57:17	Log-Likelihood:	-814.11
No. Observations:	1440	AIC:	1634.
Df Residuals:	1437	BIC:	1650.
Df Model:	2		
Covariance Type:	HC1		

	coef	std err	z	P>|z|	[0.025	0.975]
Intercept	-0.3418	0.042	-8.105	0.000	-0.424	-0.259
distance	-0.0210	0.004	-5.320	0.000	-0.029	-0.013
stars	0.2879	0.011	26.918	0.000	0.267	0.309

Omnibus:	856.233	Durbin-Watson:	1.874
Prob(Omnibus):	0.000	Jarque-Bera (JB):	83.709
Skew:	-0.095	Prob(JB):	6.65e-19
Kurtosis:	1.834	Cond. No.	21.5

Notes:
[1] Standard Errors are heteroscedasticity robust (HC1)

Logit Regression Results

Dep. Variable:	highly_rated	No. Observations:	1440
Model:	Logit	Df Residuals:	1437
Method:	MLE	Df Model:	2
Date:	Sun, 21 Dec 2025	Pseudo R-squ.:	0.2347
Time:	11:57:17	Log-Likelihood:	-761.75
converged:	True	LL-Null:	-995.32
Covariance Type:	nonrobust	LLR p-value:	3.656e-102

	coef	std err	z	P>|z|	[0.025	0.975]
Intercept	-4.8620	0.333	-14.582	0.000	-5.515	-4.208
distance	-0.1362	0.023	-5.878	0.000	-0.182	-0.091
stars	1.6674	0.102	16.278	0.000	1.467	1.868

Probit Regression Results

Dep. Variable:	highly_rated	No. Observations:	1440
Model:	Probit	Df Residuals:	1437
Method:	MLE	Df Model:	2
Date:	Sun, 21 Dec 2025	Pseudo R-squ.:	0.2349
Time:	11:57:17	Log-Likelihood:	-761.52
converged:	True	LL-Null:	-995.32
Covariance Type:	nonrobust	LLR p-value:	2.911e-102

	coef	std err	z	P>|z|	[0.025	0.975]
Intercept	-2.8692	0.185	-15.504	0.000	-3.232	-2.506
distance	-0.0803	0.013	-5.979	0.000	-0.107	-0.054
stars	0.9843	0.056	17.646	0.000	0.875	1.094

Logit Marginal Effects

Dep. Variable:	highly_rated		
Method:	dydx		
At:	mean		

	dy/dx	std err	z	P>|z|	[0.025	0.975]
distance	-0.0338	0.006	-5.870	0.000	-0.045	-0.023
stars	0.4144	0.025	16.310	0.000	0.365	0.464

Probit Marginal Effects

Dep. Variable:	highly_rated		
Method:	dydx		
At:	mean		

	dy/dx	std err	z	P>|z|	[0.025	0.975]
distance	-0.0319	0.005	-5.975	0.000	-0.042	-0.021
stars	0.3910	0.022	17.652	0.000	0.348	0.434