This assignment analyzes hotel prices in Madrid, considering ratings, stars, and distance from the city center using descriptive statistics, regressions, and model comparisons.

The analysis of hotel prices in Madrid was conducted using 708 observations from the hotel-europe_features csv file. The hotels-europe_price csv was merged with the previous file on *hotel_id* and later filtered to include only hotels located in Madrid based on the *city* and *city_actual* columns. This approach ensured the reliability of the data analysis by focusing exclusively on hotels within Madrid. Filtering for hotels allowed for a more meaningful comparison between different star rankings and whether a hotel was highly rated, without considering other types of establishments.

An exploratory analysis of the merged dataframe revealed some extreme values in price, which caused the distribution to be skewed to the right and had a high standard deviation. To address this issue, the interquartile range (IQR) method was applied to identify and eliminate extreme values that might distort the data. This allowed to retain only the prices that were relevant to the analysis and ensured the robustness of the findings.

As part of the data cleaning process, the dataset was filtered to include only hotels (based on the *accommodation type*) and instances from the year 2018 for which the *weekend*=1 flag was set. This focused the analysis on hotels in Madrid during 2018 weekends, providing a more relevant and focused context for evaluating the factors influencing hotel prices.

To assess the rating distribution, the binary variable, *highly_rated*, was created, and was assigned a value of 1 if the hotel's rating was 4 or above and 0 otherwise. Analyzing the dataset through the lens of this newly created variable revealed that approximately 58% of hotels were considered highly rated. Additionally, the average distance from the city center for highly rated hotels was 1.21 miles, while non-highly rated hotels had an average distance of 1.55 miles. This suggests a tendency for highly rated hotels to be located closer to the city center compared to their lower-rated counterparts.

The *linear probability model (LPM)* was employed to investigate the relationship between hotel distance from the city center, hotel star rating, and the probability of a hotel being considered highly rated. The results indicated that for each additional mile that a hotel is located further away from the city center, the probability of it being highly rated decreases by approximately 2%. Furthermore, for each additional star that a hotel has, the probability of it being highly rated increases by approximately 17.3%. Additionally, the constant term was 0.170, which indicated a positive and statistically significant ($p<.01$) relationship that suggests that even a hotel located in the city center with a single star still has a 17% chance of being considered highly rated.

The *Logit* and *probit* models were employed to estimate the probability of a hotel being considered highly rated. The *Logit* model results indicate that for each additional mile that a hotel is located from the city center, the probability of it being highly rated decreases by approximately 13.3%. Moreover, with a coefficient of 1.0061, we can say that for each additional star that a hotel has, the log odds of it being highly rated increase by 1.0061. This means that the odds of a 4-star hotel being highly rated are approximately 2.714 times higher than the odds of a 3-star hotel being highly rated*. The marginal effect for stars for the logit model shows that the probability of being highly rated increases by approximately 15.75% for each additional star the hotel has, while additional distance from the city center decreases the probability of being highly rated by approximately 2.09%

The *Probit* model further revealed that for each additional mile that a hotel is located from the city center, the probability of it being considered highly rated decreases by approximately 7.85%. Conversely, for each additional star that a hotel possesses, the probability of it being considered highly rated increases by approximately 61.74%. This positive and statistically significant ($p<.01$) relationship suggests that as the number of stars a hotel has increases, so does the likelihood of being considered highly rated. The marginal effect for stars for the *Probit* model indicates that the probability of being highly rated increases by approximately 16.01% for each additional star, and that for each additional mile away from the city center a hotel is, the chance of it being highly rated decreases by approximately 2.12%.

A comparative analysis of the *LPM*, *Logit*, and *Probit* models revealed that the *Probit* model outperformed the others, demonstrating a higher pseudo R-squared value and a lower log-loss value. While the log-loss values were relatively similar for all three models, suggesting comparable predictive accuracy, the *Probit* model emerged as the most robust in capturing the relationship between hotel distance from the city center, star rating, and the likelihood of being considered highly rated in Madrid. Overall, the findings from the three predictive models consistently indicate that hotels located closer to the city center have a higher probability of being highly rated, while hotels with a higher star rating also tend to receive higher ratings.

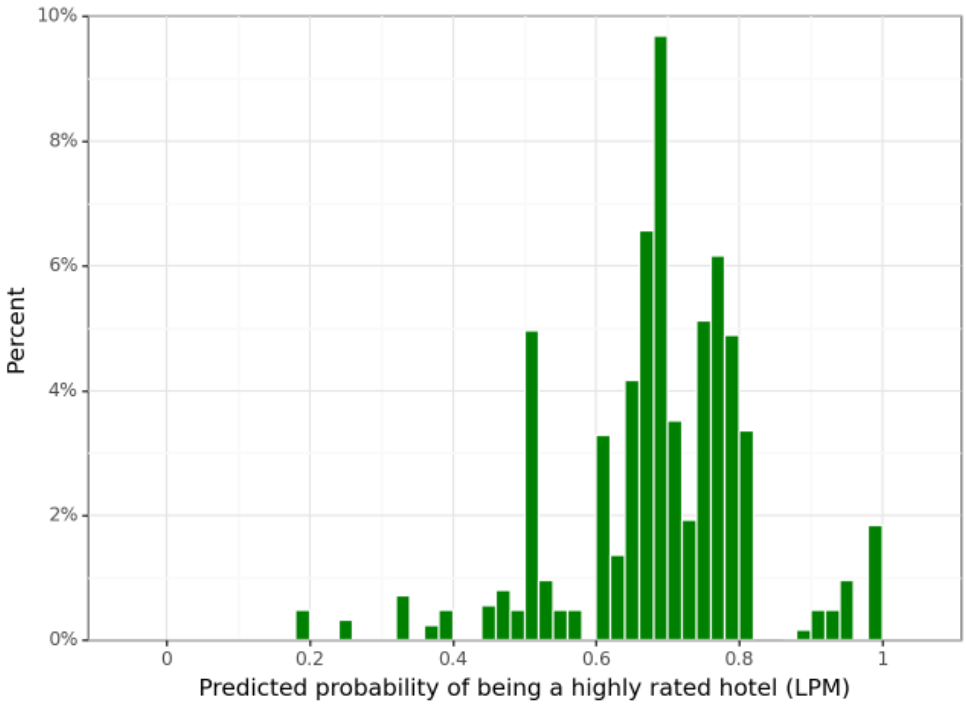
** Using Odds Ratio (Stars)=e^ (1.0061) ≈2.714*

Appendixes

Table 1. LPM of Highly Rated Hotel on Distance and Stars.

	Dependent variable: highly_rated
	(1)
distance	-0.021***
	(0.005)
stars	0.173***
	(0.014)
Constant	0.170***
	(0.057)
Observations	1371
R^2	0.121
Adjusted R^2	0.119
Residual Std. Error	0.398 (df=1368)
F Statistic	96.548*** (df=2; 1368)
Note:	*p<0.1; **p<0.05; ***p<0.01

Graph 1. Histogram and Predicted Probabilities



** Using Odds Ratio (Stars)=e^ (1.0061) ≈2.714*

Graph 2. Predicted Probabilities from the probit and logit models.

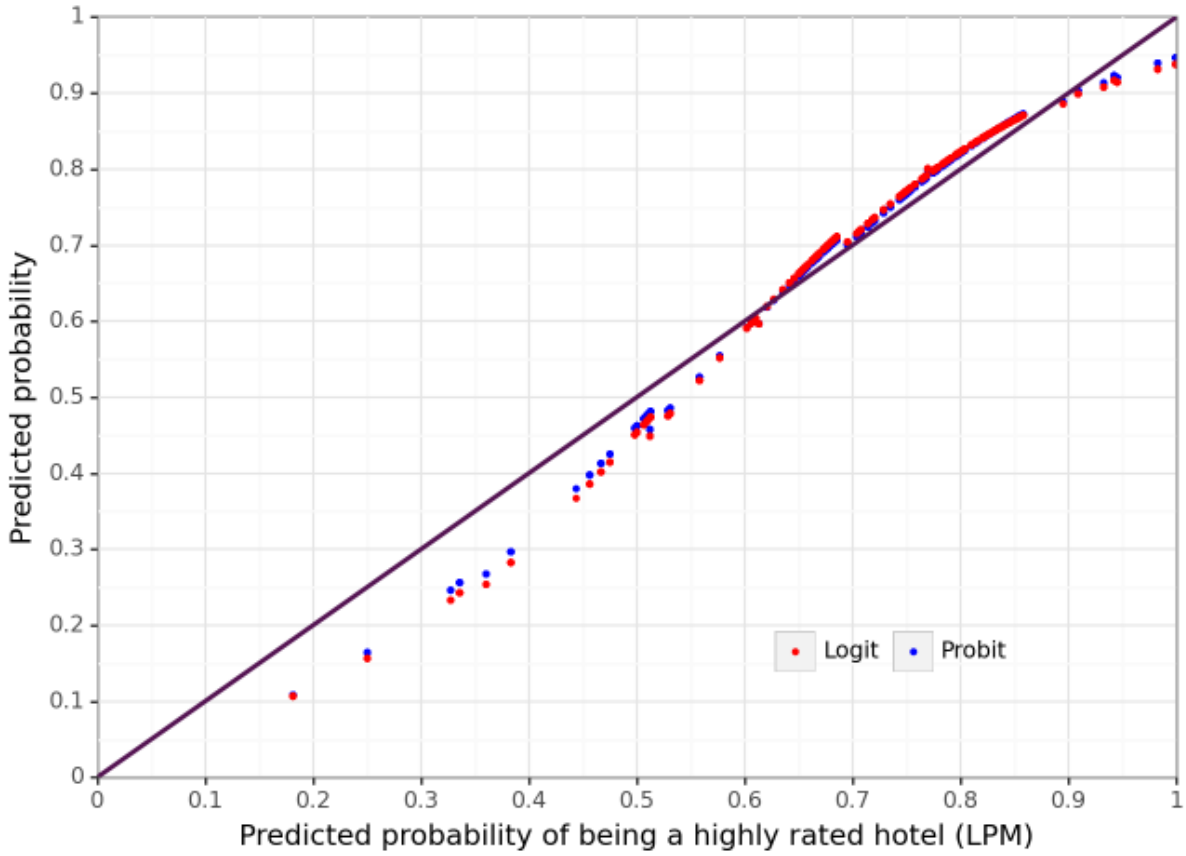


Table 2. LPM, Logit, and Probit comparison.

	LPM	Logit	Probit
R-squared	0.121	0.120	0.121
Brier-score	0.158	0.158	0.158
Pseudo R-squared	NaN	0.111	0.112
Log-loss	-0.498	-0.484	-0.483

Github repository: https://github.com/Alejandra-savagebriz/DA2/blob/main/Assignment2_DA2.ipynb

** Using Odds Ratio (Stars)=e^ (1.0061) ≈2.714*