Link to the github.
Data Preparation and Cleaning

Our analysis will be based on the hotels europe dataset. We are only interested in hotels in Rome, so we filtered the data to include only those hotels. We also dropped any extreme values, such as those with a distance of more than 10 miles. However, we did not drop any values related to price, as there did not appear to be any obvious extreme values. Additionally, we only included accommodations with a type of hotel.
We added a dummy variable called highly_rated, which indicates whether the user rating is greater than or equal to 4. We also added another dummy variable for stars (premium_hotel -> if the hotel has 4 or 5 stars), because stars is an ordinal variable, which means that there is a ranking but the difference between 1 star and 2 stars can be different than between 3 and 4 stars.

Modeling and Analysis
We created 2 LPM models.
Interpretation: From the results, we can see that distance and premium_hotel are significant at 1%. The R-squared is better for the 2nd model when we add the premium_hotel variable.
(2nd model) Our constant (0.44) means that there is an approx. 44 percent probability of the hotel being highly-rated. The distance coefficient shows that when the distance increases by 1 mile, the probability of being a highly-rated hotel decreases by approximately 4 percent. SE defines the confidence interval. For a 95% CI, it is [-0.052, -0.036], which means that a 1-mile distance affects the probability of being highly rated by around 4-5% negatively. It doesn't include 0, so we can also reject Ho based on that (of course, we can do that by looking at p as well).

The coefficient of premium_hotel indicates that if the hotel is premium (stars >=4), there is a 24% higher probability of the hotel being highly rated.

We also built logit and probit models. The results of these models were similar to the results of the LPMs.

Logit Interpretation: We obtained similar results for the logit model as well. The p-values indicate that the coefficients are significant, and the confidence interval values are consistent. However, for the 'premium_hotel' coefficient, the confidence interval is slightly narrower. The coefficient for 'distance' is -0.046, which is slightly larger in absolute value compared to the coefficient in LPM, while the 'premium_hotel' coefficient is a bit smaller.

Probit Interpretation: We obtain similar results as with logit and LPM. The confidence interval of distance is slightly narrower for probit compared to logit. The coefficients are significant at the 1% level.

Visualizations
After building all three models, we predicted the probability of a hotel being highly_rated for each observation in the dataset. We then created a graph that shows the predicted probabilities from each model.
Interpretation of graph: The predicted probabilities from the logit and the probit are very similar. Their range is between 0.11 and 0.68, which is somewhat narrower than the predicted probabilities from the LPM (their range is 0.04 to 0.68). The narrower ranges are the result of the S-shaped curves of the logit and the probit that approach zero and one slowly, in contrast with the straight line produced by the LPM.

Summary
In the end we built a table that shows 4 measures of fit of our 3 models: lpm, logit and probit.

Overall, the Logit and Probit models seem to perform similarly and better than the LPM model in terms of explaining variance, predictive accuracy, goodness of fit. The differences between Logit and Probit are minimal in these statistics. R-squared is higher for Logit and Probit compared to the LPM, Brier-score is around the same, Log-loss is slightly smaller in absolute value for Logit and Probit compared to the LPM