Data Analysis 2: Assignment 2 (Moscow)

Ian Brandenburg; ID 2304791

Dávid Szabados; ID 2302806

Introduction:

The assignment investigates the relationship between the 'highly_rated' variable, derived from hotel ratings, and the variables 'stars', 'distance', and the logarithm of 'price'. 'Highly_rated' is set to 1 for ratings of 4 or higher and 0 otherwise. We will employ linear probability, logit, and probit models to analyze relative price impacts and correlations.

Filtering, Cleaning and basic attributes of the Data:

Data was refined to include only hotels in Moscow, from weekdays in November 2017, with prices under $1000 due to a long right tail distribution. Entries missing ratings or review counts were removed to avoid bias in the 'highly_rated' assessment. Details are in Appendix 1.

Models and Interpretation:

In Appendix 2, we applied three models to estimate the likelihood of a hotel being highly rated. The Linear Probability Model indicates that hotel stars significantly influence this probability, increasing it by roughly 15% per star, while a unit rise in lnPrice ups the chance by 7.5%. Distance plays a lesser role in this estimation.

Logit Marginal Effects Summary:

In a logit model, a 1% rise in price increases the likelihood of a hotel being highly rated by 7.37%, and an additional star increases it by 15.41%, both statistically significant. However, the impact of distance is minor (1.22%) and not statistically significant.

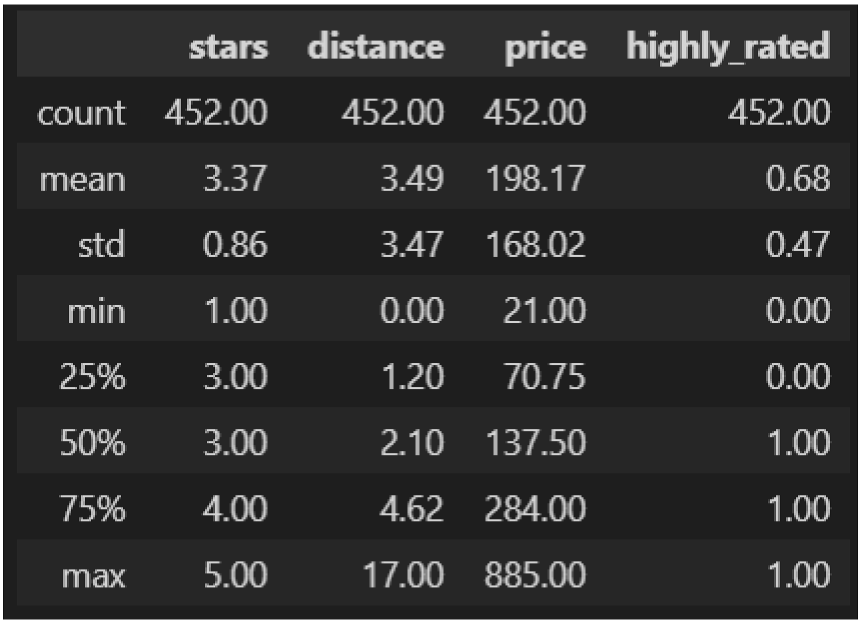
Probit Marginal Effects Summary:

In a probit model, a 1% price increase raises the probability of a hotel being highly rated by 7.19%, while an additional star increases it by 15.27%, both with strong statistical significance. Distance shows a negligible effect (1.08%) and lacks statistical significance.

Conclusion:

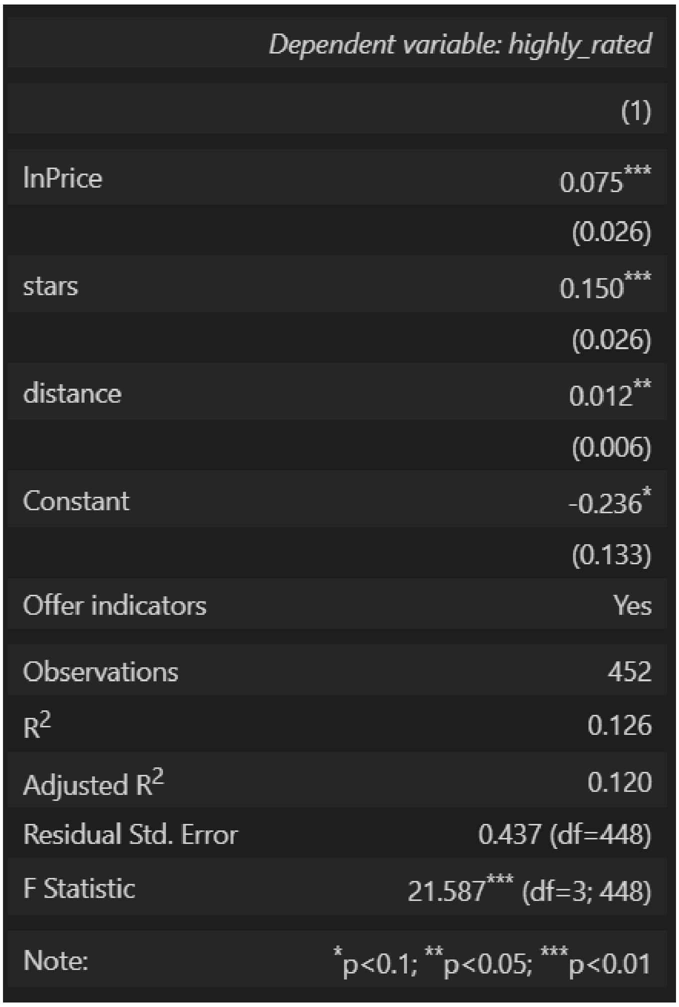
As shown in Appendix 4, both the probit and logit analyses predict very similarly. Overall, this analysis of Moscow hotels shows that hotel stars and price significantly impact guest ratings. Each additional star increases the likelihood of a high rating by about 15%, and a 1% price increase correlates with a 7% higher chance of being highly rated. In contrast, the distance from the city center has a negligible influence on ratings. These results highlight the primary importance of quality and value in determining hotel ratings.

Appendix 1



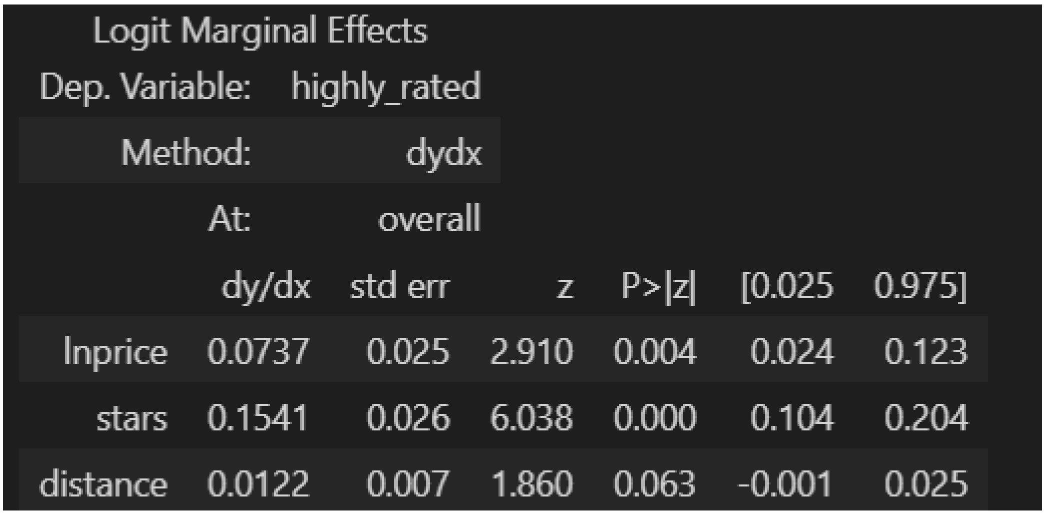
	stars	distance	price	highly_rated
count	452.00	452.00	452.00	452.00
mean	3.37	3.49	198.17	0.68
std	0.86	3.47	168.02	0.47
min	1.00	0.00	21.00	0.00
25%	3.00	1.20	70.75	0.00
50%	3.00	2.10	137.50	1.00
75%	4.00	4.62	284.00	1.00
max	5.00	17.00	885.00	1.00

Appendix 2

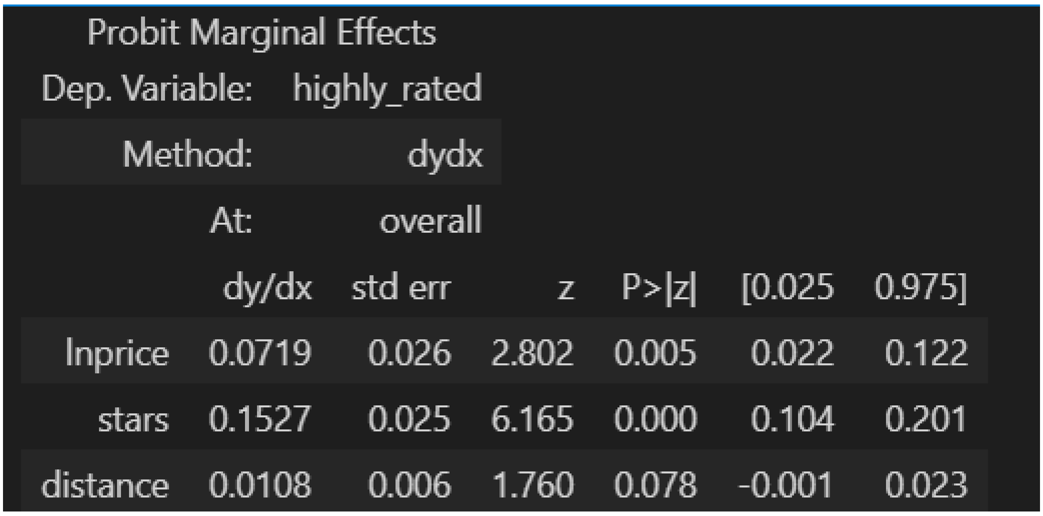


	Dependent variable: highly_rated
	(1)
lnPrice	0.075^{***}
	(0.026)
stars	0.150^{***}
	(0.026)
distance	0.012^{**}
	(0.006)
Constant	-0.236^{*}
	(0.133)
Offer indicators	Yes
Observations	452
R^2	0.126
Adjusted R^2	0.120
Residual Std. Error	0.437 (df=448)
F Statistic	21.587^{***} (df=3; 448)
Note:	$^{*}p<0.1$; $^{**}p<0.05$; $^{***}p<0.01$

Appendix 3.a/b

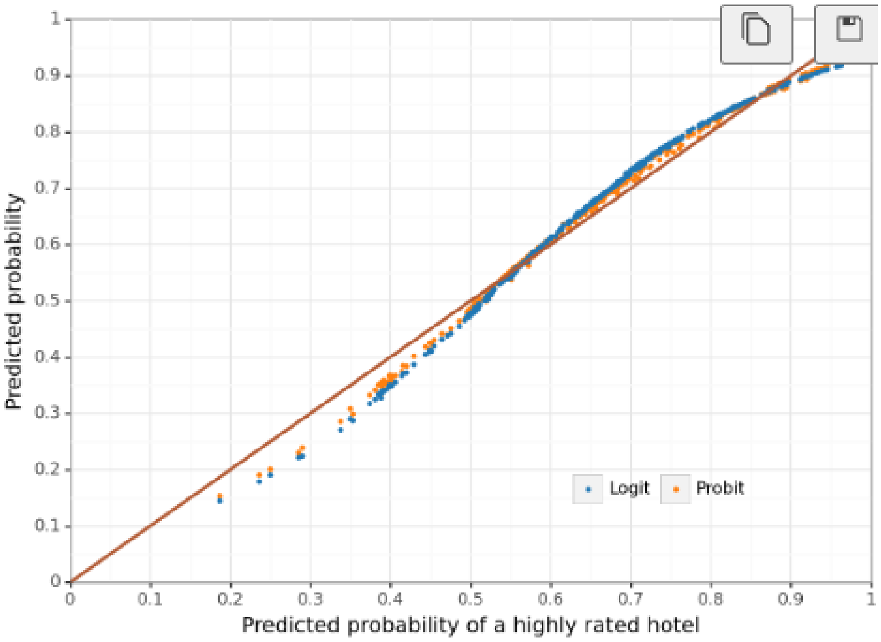


```
         Logit Marginal Effects
  Dep. Variable:    highly_rated
        Method:             dydx
            At:          overall
             dy/dx   std err      z   P>|z|   [0.025   0.975]
  lnprice   0.0737     0.025  2.910   0.004    0.024    0.123
    stars   0.1541     0.026  6.038   0.000    0.104    0.204
 distance   0.0122     0.007  1.860   0.063   -0.001    0.025
```



```
         Probit Marginal Effects
  Dep. Variable:    highly_rated
        Method:             dydx
            At:          overall
             dy/dx   std err      z   P>|z|   [0.025   0.975]
  lnprice   0.0719     0.026  2.802   0.005    0.022    0.122
    stars   0.1527     0.025  6.165   0.000    0.104    0.201
 distance   0.0108     0.006  1.760   0.078   -0.001    0.023
```

Appendix 4


Predicted probability
Predicted probability of a highly rated hotel
Logit
Probit