

DATA ANALYSIS

ASSIGNMENT 2

INTRODUCTION

We selected the city of Prague to conduct this analysis. From an initial number of 406 hotels, we worked on 320 units. We set a limit of USD600 per night, excluding 4 hotels from our analysis. Only hotels with proper ratings were kept.

75% of the Prague hotels selected are 3 stars and above. The average hotel rating is 4.0, seemingly high. Hotels are close to the city center (1.2 mile on average).

WORK

The first analysis of the data showed the strong correlation between price variation and the 3 variables: distance, rating, stars. See Scatter Plot (exhibit 1).

As far as ratings are concerned, we applied the three prediction tools LPM, Logit and Probit to establish what hotel features have an impact on rating and to what extent they determine the probability of getting a high rating (>4.0).

We looked at "Stars" and "Distance" both as absolute and binary values. We also explored promotional price cuts (not decisive) and number of reviews (marginal).

Whilst "Stars" and "Distance" showed high significance ($<1\%$) on linear prediction analyzed separately, we obtained unacceptable results with the combination of "Stars" and "Distance" (negative probability possibly due to the narrow distribution of hotels on both items).

Having observed the compact nature of the data, we selected 2 binary variables: "4 stars and above" on one hand and "under 1 mile" on the other hand (exhibit 2). Those choices of cuts were the result of all the preliminary analysis on the data (see Python code). We got satisfactory results with the combination of both binary variables.

CONCLUSION

To better understand the probability of getting a high rating (on average 60%), we come to the conclusion that:

- the probability to get a high rating is impacted by the hotel quality (4 star and above - +35pts)
- the distance from the city center (in our analysis "under 1 mile" is adding 15 pts to the probability).

The Probit (exhibit 3) and Logit give strong similar estimates. Results satisfy the 1% significance level. However some indicators are less conclusive like the probability distribution particularly with the LPM (data too compact?).

APPENDICES

Exhibit 1 - Hotel Sample

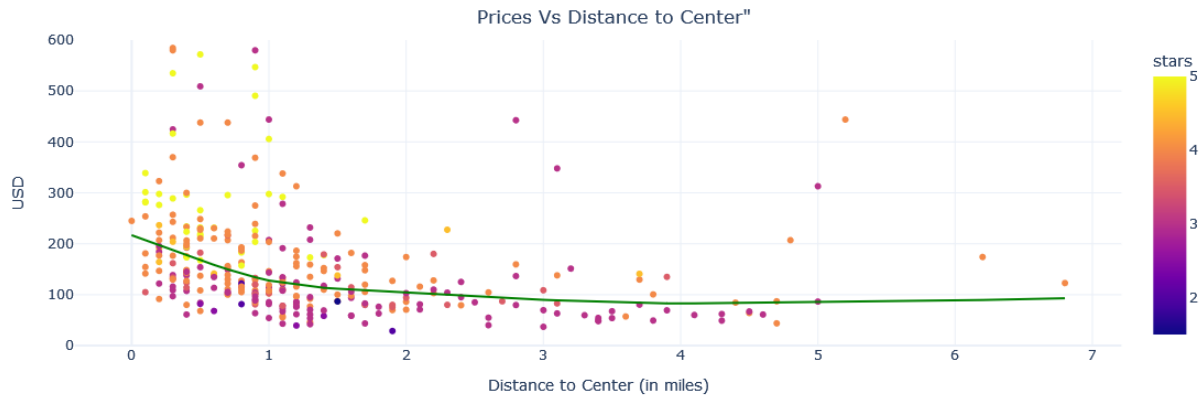


Exhibit 2 - LPM

(1) = Vs stars (number) and distance (miles)

(2) = Vs binary variables "4 stars +" and "under 1 mile"

Dependent variable: high_rating		
	(1)	(2)
stars	0.338*** (0.032)	
distance	-0.079*** (0.021)	
4 star +		0.418*** (0.053)
1 mile and under		0.158*** (0.052)
Constant	-0.524*** (0.141)	0.305*** (0.043)
Observations	320	320
R ²	0.305	0.247
Adjusted R ²	0.300	0.243
Residual Std. Error	0.407 (df=317)	0.423 (df=317)
F Statistic	97.440*** (df=2; 317)	55.365*** (df=2; 317)

Exhibit 3 - Probit Marginal Effect

Probit Marginal Effects						
Dep. Variable:	high_rating					
Method:	dydx					
At:	overall					
	dy/dx	std err	z	P> z	[0.025	0.975]
four_star_plus	0.3561	0.034	10.525	0.000	0.290	0.422
under_mile	0.1504	0.046	3.289	0.001	0.061	0.240

The marginal results are very close to the logit numbers (within 1 pt range) on both variables.

In both instances, the P value is close to zero, the CI interval does not include. We can reject the null.