

# Data Analysis : Assignment 2

Ghazal Ayobi and Shah Ali Gardezi

## Introduction

The Question of this case study is how hotels stars is related to highly rated hotels. For this assignment we use Hotels-Europe data. This data set contains two tables **Features** and **Price**. We joined the two tables using left join.

## Data Transformation

As a process of filtering and data transformation, we use hotel user rating as the dependent variables and transformed it to a binary variable called *highly\_rated* which equals to one if *rating* is more than 4, 0 otherwise. We selected **Paris** City and considered **Hotels** as accommodation type. Moreover, we excluded hotels with less than *USD 600* per night, and we removed null and duplicated values from the data set. In order to understand what functional form to include in the regression we examined *Lowess* regression with highly rated hotels and distance. *Figure 1* in the appendix shows that distance and highly rated hotels have negative relationship from city center to 1.2 miles away from center, however *highly\_rated* hotels does not indicated any relationship between 1.2 and 3 miles. After 3 miles, as distance changes by one mile highly rated declines. We created a binary variable for *Stars* called *top\_stars* which equals to one if *stars* is more than 3, 0 otherwise. Other variables are *price* which was transformed to log of Price and *weekend* which is binary variable

## Analysis

Table 2 shows five regression mpdels: lpm, logit, marginal logit, probit and marginal probit regression results.

- Filtered the data for **Paris**
- Selected Hotels and Apartments for accommodation type
- Price is less than 600
- Removed null values from *stars*
- Removed duplicates
- Removed Null values from *rating*, *stars*, and *distance*
- Created log of price, *lnprice*
- Created *highly\_rated* if rating  $\geq 4$

Table 1: Summary Statistics

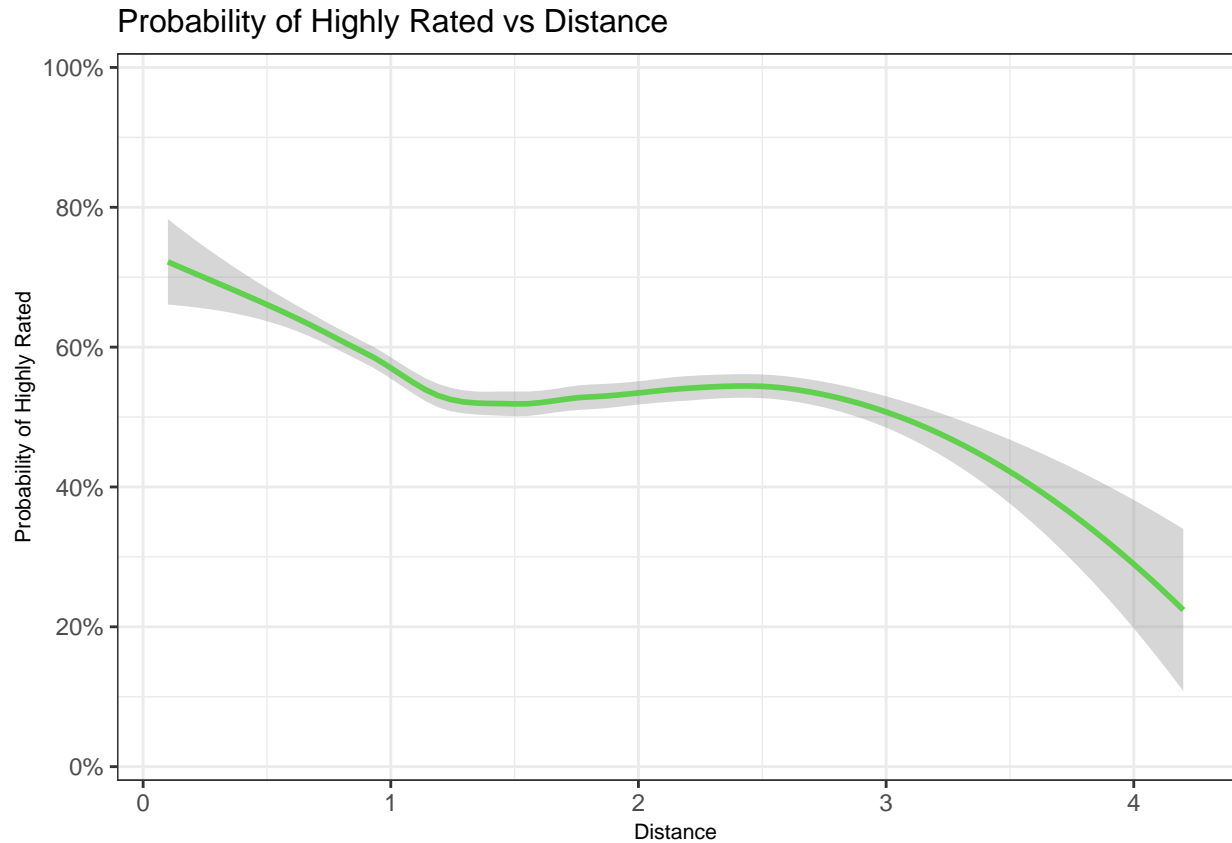
	Mean	SD	Min	Max	Median	P95	N
highly_rated	0.56	0.50	0.00	1.00	1.00	1.00	11397
distance	1.62	0.78	0.10	4.20	1.60	2.90	11397
stars	3.22	0.78	1.00	5.00	3.00	4.00	11397

Table 2: The Probability of Highly rated hotels and top stars - LMP, Logit, and Probit models

	(1) LMP0	(2)LMP	(3) logit coeffs	(4) logit Marg	(5) Probit	(6) Probit Marg
Constant	0.137** (0.011)	0.027 (0.027)	-2.551** (0.146)		-1.451** (0.084)	
top_stars	0.497** (0.012)	0.421** (0.012)	2.145** (0.075)	0.436** (0.012)	1.266** (0.041)	0.432** (0.011)
lspline(distance, c(1.2, 3))1		-0.108** (0.021)	-0.518** (0.110)	-0.104** (0.022)	-0.340** (0.065)	-0.113** (0.021)
lspline(distance, c(1.2, 3))2		0.018* (0.009)	0.096* (0.043)	0.019* (0.009)	0.061* (0.026)	0.020* (0.009)
lspline(distance, c(1.2, 3))3		-0.251** (0.050)	-1.268** (0.259)	-0.254** (0.050)	-0.802** (0.155)	-0.268** (0.049)
price		0.001** (0.000)	0.006** (0.000)	0.001** (0.000)	0.003** (0.000)	0.001** (0.000)
weekend		0.101** (0.010)	0.481** (0.049)	0.098** (0.010)	0.313** (0.030)	0.106** (0.010)
Num.Obs.	11 397	11 397	11 397	11 397	11 397	11 397

\*  $p < 0.05$ , \*\*  $p < 0.01$

```
## fitting null model for pseudo-r2
## fitting null model for pseudo-r2
```



	Model 1	Model 2	Model 3
Constant	0.027 (0.027)	-2.551** (0.146)	-1.451** (0.084)
top_stars	0.421** (0.012)	2.145** (0.075)	1.266** (0.041)
lspline(distance, c(1.2, 3))1	-0.108** (0.021)	-0.518** (0.110)	-0.340** (0.065)
lspline(distance, c(1.2, 3))2	0.018* (0.009)	0.096* (0.043)	0.061* (0.026)
lspline(distance, c(1.2, 3))3	-0.251** (0.050)	-1.268** (0.259)	-0.802** (0.155)
price	0.001** (0.000)	0.006** (0.000)	0.003** (0.000)
weekend	0.101** (0.010)	0.481** (0.049)	0.313** (0.030)
Num.Obs.	11 397	11 397	11 397
R2	0.185		
PseudoR2		0.149	0.148

\*  $p < 0.05$ , \*\*  $p < 0.01$

