Report

Barcelona School of Economics

Effects of an Event on Barcelona's Rental Prices

Homework 1: Mini Research Project

Research of the effects of a big annual event in Barcelona on rental prices on "booking.com"

Student: Mathieu Breier, Angelo Di Gianvito, Daniela Vélez

Course: 22DM014 Introduction to Text Mining and Natural Language Processing

February 4, 2024

1. Identify a (future) event that makes a lot of people come to Barcelona. Think about music festivals, local festivities etc. (2 points)

We have selected the Sónar festival. It is the 31st edition of the Barcelona International Festival of Advanced Music and Multimedia Art in 2024. This vibrant event takes place in Montjuic and attracts enthusiasts from all over the world to Barcelona to participate in its rich offer.

2. Think of the time periods to scrape and what second city to scrape for these same timer periods. Explain your choices in written. (2 points)

The festival unfolds on June 13, 14, and 15, and we have opted to analyze the period preceding it, that is the corresponding days on June 6, 7, and 8, 2024. Ensuring an equivalent number of days and proximity to the event dates is crucial for a meaningful comparison of similar scenarios. We also include Valencia as the second city to control for due to its proximity and similarities to Barcelona. Both cities are located on Spain's eastern coastline along the Mediterranean Sea and share similar geographical and cultural situation.

3. Design a careful scraping pipeline that follows the advises seen in class and TAs. The basic points to bear in mind are:

- ***Organize the data you need, format and structure to store it beforehand. Try to foresee how you will need to read in the data to answer your questions. If you want, you can include some few lines explaining your pipeline strategy at the beginning.***
- ***Codes should be as automated as possible. That is, you don't want to rely on human intervention to get your data.***
- ***Use only the packages we have seen in the course. Although Firefox is recommended, you can also use Chrome as your scraping browser.***
- ***Document your codes and make them robust and efficient.***

See **Pipeline (Step by Step)** section in **scraping_nb.ipynb** file.

4. Scrape date, room price, hotel name and hotel description. (5 points)

See **scraping_nb.ipynb** file.

5. Write down three equations. One with just a "treatment period" dummy, one with just a "treatment city" dummy and then one that adds both of them with their interaction. This last regression gives you a difference-in-difference estimate of the effect of the event on prices. Explain which coefficient captures this treatment effect and why you need a second city for this. (3 points)

First Regression

Treatment_city is a dummy that takes 1 if the city belongs to Barcelona and 0 if the city is Valencia.

$$\text{price} = \beta_0 + \beta_1 \cdot \text{Treatment\_city} + \epsilon \tag{1}$$

Second Regression

Treatment_period is a dummy that takes 1 if the days are between 13-15 June 2024 and 0 if the days are between 06-08 June 2024.

$$\text{price} = \beta_0 + \beta_1 \cdot \text{Treatment\_period} + \epsilon \tag{2}$$

Third Regression

$$\begin{aligned}\text{price} = \beta_0 + \beta_1 \cdot \text{Treatment\_period} + \beta_2 \cdot \text{Treatment\_city} \\ + \beta_3 \cdot (\text{Treatment\_period} \times \text{Treatment\_city}) + \epsilon\end{aligned} \tag{3}$$

The coefficient β_3 represents the difference-in-differences (DiD) estimate. It captures the net effect of the event on prices by considering the interaction between "Treatment_period" and "Treatment_city.", therefore this term allow us to isolate and quantify the unique effect of the treatment when both factors are present.

The need for a second city (control city) is crucial in DiD analysis to create a counterfactual. The control city allows us to account for trends or factors affecting both the treatment and control groups similarly that remain constant over time.

The interaction term (Treatment_period × Treatment_city) captures the differential effect by considering how the treatment effect varies across the treatment and control groups. Therefore, the DiD identify the treatment effect by contrasting the changes in the treatment city with the changes in the control city over the same period to avoid confounding variables and be more robust the estimation of the event's impact on prices.

6. Estimate all three regressions. Make a standard regression table with 4 columns (3 for your answer here and one more below). Make sure you check how these regressions look like usually. Always report all coefficients. Then carefully interpret them for each regression and the changes you see. (4 points)

Table 1: Regression Results

	Dependent variable: Price			
	Model 1	Model 2	Model 3	Model 4
const	256.179***	332.899***	253.243***	158.997***
	(5.182)	(4.925)	(7.787)	(9.095)
Treatment_city	147.552***		120.979***	141.145***
	(6.628)		(9.597)	(9.435)
Treatment_period		26.705***	5.197	-15.831
		(6.930)	(10.360)	(10.008)
Interaction_term			58.004***	40.994***
			(13.217)	(12.656)
amenities				45.305***
				(6.805)
luxury				102.713***
				(7.391)
Observations	3188	3188	3188	3188
R^2	0.135	0.005	0.151	0.233
Adjusted R^2	0.134	0.004	0.150	0.232
Residual Std. Error	182.411 (df=3186)	195.631 (df=3186)	180.785 (df=3184)	171.832 (df=3182)
F Statistic	495.627*** (df=1; 3186)	14.850*** (df=1; 3186)	188.039*** (df=3; 3184)	193.377*** (df=5; 3182)
Note:				*p<0.1; **p<0.05; ***p<0.01

Interpretation of regression 1

The first regression provides with the average difference in hotel price between Barcelona and Valencia during the period starting from the 13th to the 15th of June 2024. The average difference in hotel price between the 2 cities is 147.6 euros and is significant. This indicates that on average, during that period, average prices in hotels will be 147.6 euros higher in Barcelona than in Valencia. However, it is important to mention that this first regression does not control for other potential confounding factors. Hence, the difference may not necessary originate from the event happening in that period.

Interpretation of regression 2

The second regression indicates the average difference in hotel price between the period starting from the 13th to the 15th of June 2024 and the previous period starting from the 6th to the 8th of June 2024. The average difference in hotel price between the 2 periods is 26.7 euros and is significant. In other words, hotel prices during the period going from the 13th to the 15th appear to be higher in both cities in comparison to the previous period.

Interpretation of regression 3

The results of the third regression shows a high, positive and significant coefficient for the interaction term. This indicates that there is a significant positive effect of treatment on prices. In other words, we can conclude that the Sonár festival happening in Barcelona during that period positively influences hotel prices with an associated increase of 58 euros specific to this event.

Interpretation of regression 4

In our fourth and final regression model, we incorporated a set of terms associated to luxury and amenities from the document term matrix that are correlated with hotel pricing. This analysis is enriched by the inclusion of both city and time period dummy variables, which serve to adjust for consistent effects over time, as well as an interaction term to understand the combined influences. This regression points out how certain descriptive terms in hotel listings are associated with pricing - whether they contribute to a price increase or decrease. These terms also allow us to control for difference specific to one city.

We observe that the interaction term remains significant. This suggests that the event happening in Barcelona during that time period does have a positive influence on the hotel pricing. We also observe that the interaction term coefficient slightly decreases from 58 to 40.9 after controlling for the 2 set of terms, luxury and amenities. This suggests that terms related to luxury and amenities were likely confounders with a positive bias on the price.

Furthermore, the analysis reveals some other findings:

- Terms related to luxury are associated to an increase of 102.7 euros in hotel prices. In other words, hotel prices are on average 102.7 more expensive when one of the terms included in the luxury vocabulary defined above is included in the hotel description. This is intuitive as more luxurious hotels are often more expensive.
- Similarly, the presence of terms related to amenities is positively correlated to the price, associated to increase of 45.3 in hotel prices. This is also intuitive, as hotels with more amenities would generally be more expensive.

7. For each of the hotel descriptions do the following:

(a) Extract at least two text features that can be useful controls in the regression. Think about the methods covered in class to transform text into numeric features and explain your decision. Show summary statistics for your feature for the different cities and time periods. (10 points)

In the latest regression analysis, we recognized the importance of introducing control variables to enhance the precision and reliability of our results. Specifically, we opted to incorporate two key hotel-specific characteristics as controls: 'luxury' and 'amenities.' To operationalize these variables, we generated dummy variables. The 'luxury' dummy variable assumes a value of 1 if any of the following words appear in the data—'elegant,' 'moderna,' 'moderno,' 'privado ducha,' 'piscina'—and 0 otherwise. Similarly, the 'amenities' dummy variable takes a value of 1 if any of the words 'conexión wifi,' 'wi fi,' 'park privado,' 'aparcamiento,' 'piscina' appear, and 0 otherwise.

By incorporating these controls into our regression model, we aim to account for variations in hotel pricing that can be attributed to differences in luxury features and available amenities. This strategic inclusion ensures that the effects observed for our treatment variables are not confounded by differences in luxury or amenity offerings. Omitting 'luxury' and 'amenities' as control variables could introduce bias to our estimates, potentially distorting the true impact of our treatment variables on hotel prices. Including these controls serves as a robust method to mitigate any such biases, providing a more accurate understanding of the factors influencing our observed outcomes.

Summary Statistics for Luxury and Amenities

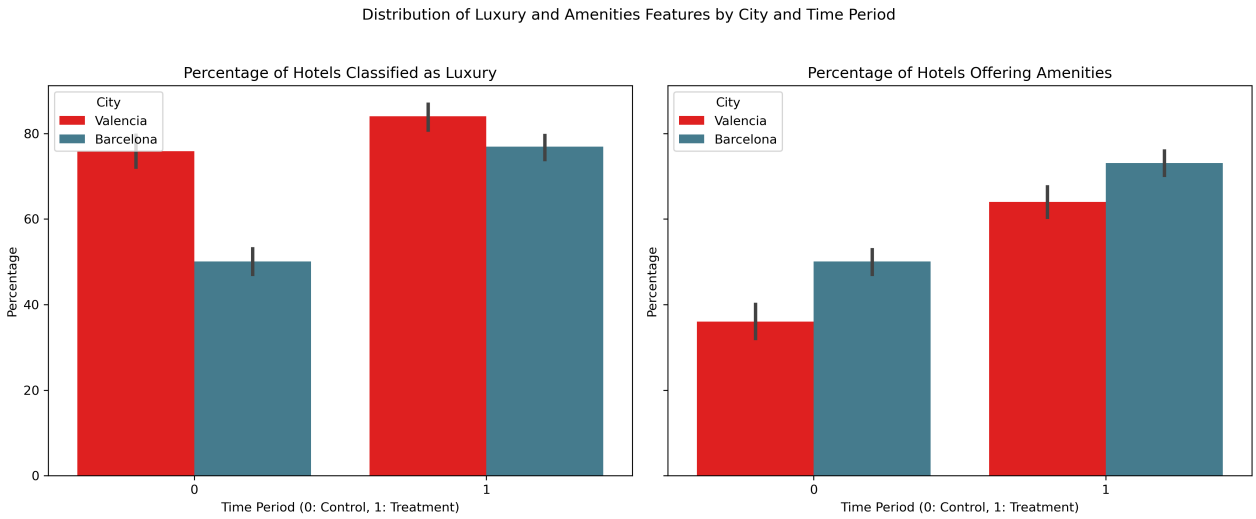

Figure 1: Average Hotel Prices in Madrid & Valencia

Table 2: Summary Statistics for 'luxury' and 'amenities' Features

		luxury			**amenities**		
		count	mean	sum	count	mean	sum
Treatment_city	Treatment_period						
0	0	539	0.758813	409	539	0.359926	194
	1	700	0.840000	588	700	0.640000	448
1	0	1039	0.500481	520	1039	0.500481	520
	1	910	0.769231	700	910	0.730769	665

(b) Run a 4th regression and add it to your table. This should add your controls to your 3rd regression. Interpret the coefficient on the controls and the change you see on your treatment variable. (2 points)

Interpretation of regression 4

See answer question 6) Interpretation of regression 4

8. Imagine that you instead run the regression with hotel fixed effects (no need to run). Explain why the treatment effect will change and why your regression with controls should be closer to this regression. (2 points)

Running a regression with hotel fixed effects means that we are controlling for the unobserved heterogeneity that are unique to each hotel and are constant across time, this reduces omitted variable bias.

The analysis focuses on the variation in the treatment effect within hotels over time. This allows us to have a more precise estimate. The magnitude of the treatment effect may change

Table 3: Regression Results

	Dependent variable: Price
	Model 4
const	158.997***
	(9.095)
Interaction_term	40.994***
	(12.656)
Treatment_city	141.145***
	(9.435)
Treatment_period	-15.831
	(10.008)
amenities	45.305***
	(6.805)
luxury	102.713***
	(7.391)
Observations	3188
R^2	0.233
Adjusted R^2	0.232
Residual Std. Error	171.832 (df=3182)
F Statistic	193.377*** (df=5; 3182)
Note:	*p<0.1; **p<0.05; ***p<0.01

because the fixed effects model is capturing and accounting for additional sources of variation, we could split the treatment effect from other cofounding factors.

Our regression with controls should be closer to this regression because if this happens this suggest that the control variables capture some of the unobserved heterogeneity that the fixed effects model is addressing. The fixed effects model is more robust in dealing with unobserved heterogeneity.

EXTRA

Summary Statistics

Table 5: Top 5 Terms with Highest Correlation to Price

Terms	**Correlation**
estrella	0.571812
vario	0.477926
piscina	0.404673
servicio habitacion	0.380868
hotel estrella	0.355597

Table 4: Top 20 most frequent terms present in the descriptions

Barcelona Period 1	Barcelona Period 2	Valencia Period 1	Valencia Period 2	General
hostal	rambla	minuto pie	establecimiento	playa
cataluña	moderno	ciencia	solo	desayuno
plaza cataluña	información	cuenta	moderna	10
además	barrio	solo	aqua	15
sagrada	sagrada familia	desayuno	km alojamiento	tren
sagrada familia	sagrada	moderna	piscina	dispon
diagon	familia	10	incluyen	rambla
familia	tren	restaurant	centro comerci	conexión
zona	dispon	aqua	minuto coch	decoración
sirv	información turística	coch	arena	barrio
fuert	servicio habitacion	dispon	comerci	gratuito
aeropuerto barcelona	terraza	habitacion air	toda habitacion	vista
caja fuert	gracia	art ciencia	autobú	apartamento
caja	instalacion	interé	equipada	bar
bañera	prat	ciudad art	alberga	alberga
terraza	toda habitacion	ropa cama	jardín	coch
art	barcelona prat	grati alojamiento	zona	artículo
equipada	cerca	minuto coch	ilunion	vía
tren	turística	km aeropuerto	instalacion	vía satélit
parada	centro barcelona	cafetera	conexión	satélit

Visualisations

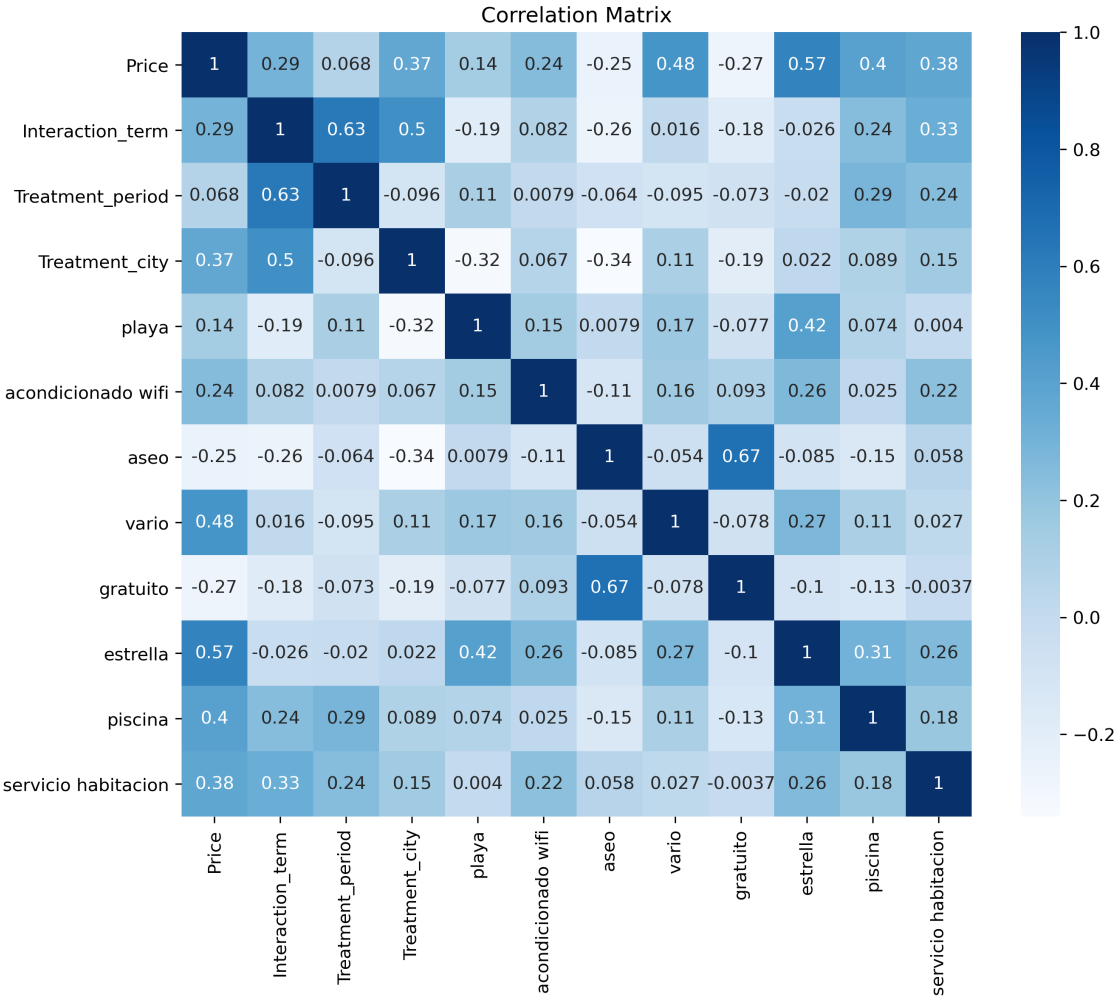

Figure 2: Correlation Matrix

Figure 3: Word Cloud

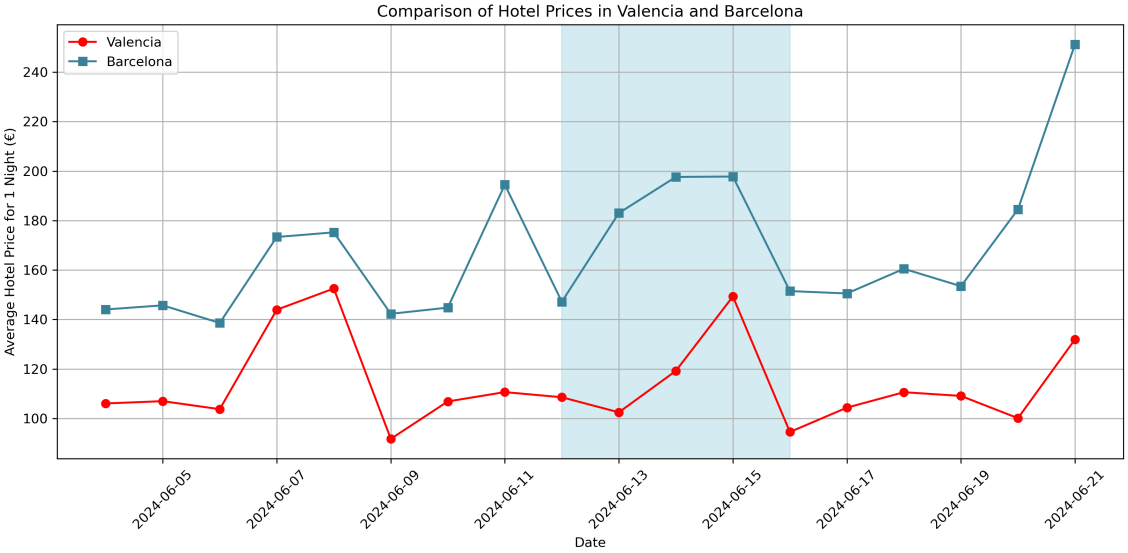

Figure 4: Average Hotel Prices in Madrid & Valencia