NLP and Text Mining

Scraping Booking.com: Insights on Hotel Prices

Andrew Bennett, Luke Atazona, Miguel Handt Fueyo

February 4, 2024

1 The event: Primavera Sound Festival

The chosen event for our study is the Primavera Sound Festival, renowned as one of Barcelona's largest music events, showcasing a wide array of musical genres. It attracts an extensive, varied crowd, including both local residents and international visitors, across different ages and genders. Scheduled to run from May 30th to June 2nd, 2024, the festival's estimated 500,000 attendants (Primavera Sound, 2024) are expected to significantly boost the demand for accommodations such as hotels and hostels, leading to increased bookings and therefore influencing the prices. Given the festival's nature as a leisure event, it would not be surprising to see spillover effects beyond the earmarked days of the event.

2 Time period and control city

2.1 Choice of Event and Non-event Time Periods

Our time periods - the events and non-event weeks - are chosen to allow for direct comparison between regular weeks and a major event week. Also, while the event takes place from the 30th of May to the 2nd of June 2024, we operationally define our "eventweek" as going from the 25th of May to the 2nd of June to account for prebooking and extended stays price effects. This allows us to uncover the adequate price dynamics arising from the event. Our choice of non-event week takes into account the need to have a week in the same season but also that is appropriately distant from other major events that could contaminate our treatment. We define our "non-eventweek" to be from the 11th to the 19th of May (a Sat to a Sun), ensuring that there is no spillovers from the event (See Figure 2) . Our choice of the weeks is also to ensure that effects arising from major overlapping events with Primavera Sound are isolated. This way, observed differences in rental prices can be attributed to Primavera Sound Festival in Barcelona.

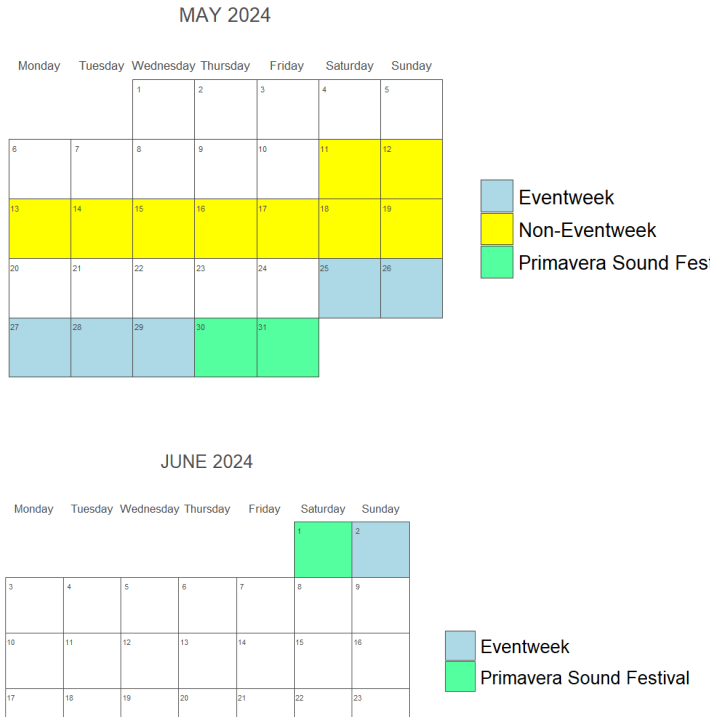

Figure 1: The Timeframe

2.2 Choice of Control City

Our control city is Valencia, located on the eastern coast of Spain, approximately 350km away from Barcelona. Valencia provides an appropriate counterfactual for Barcelona for a number of reasons. Valencia is geographically and culturally close and similar to Barcelona. The city has comparable tourist appeal; like Barcelona, it is a major tourist destination in Spain with a rich cultural scene, and a major football team. Indeed, Valencia, like Barcelona, experiences peak tourist seasons during the summer months, typically from June to August.

Valencia, also does not have major events occurring at the time of the event week in Barcelona, and similarly is without major events in the control week. The largest events of impact during the event week are local concerts, and a basketball game. In the non-event week, there is also no noteable events of impactVisit Valencia (2024).

The weather is also generally identical. Since Hotel prices are crucially driven by seasonality, it is ideal that both the treatment and control cities have comparable climates and access to beaches. This makes it ideal as a control city to better isolate the effect of Primavera Sound on Barcelona's rental price. While Madrid has comparable size, we believe it does not provide adequate similarities to serve as appropriate counterfactual for our analysis, such as the tourist seasons. A baseline balance check of covariates, in particularly temperature pattern confirms these observations (see Figure 2).

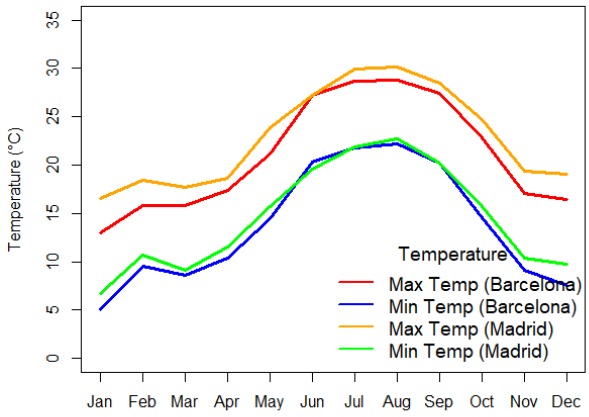

Figure 2: Monthly Average Temperatures for Barcelona and Valencia

3 Scraping pipeline

The scraping pipeline begins by opening a Firefox browser. Then the browser will attempt to exit all popup windows. It begins by accepting the cookies. Then it refreshes the page 3 times in an attempt to trigger the promotional popup. After, it will accept the promotional popup if it exists, and otherwise it will begin exiting the login prompt. The browser switches contexts to the Google sign-in iframe, declines the sign in then gets ready to search for a hotels.

The scraping pipeline then inputs a city, then selects the dates, then searches for the selected city and dates. It waits an adequate amount of time (depending on internet connection), then proceeds to collect data. The scraper will iterate through each page of the hotel site, and pull relevant information from each div container that displays hotel information. Most importantly, the scraper is collecting links for each of the hotel detail pages.

The final step is for the scraper to visit each of the hotel description pages and pull all relevant information. Items of particular value are: latitude and longitude, description, and ratings. Once the scraper has completed the collection of the data it will return a dictionary that will be turned into a dataframe. To speed the computational workload, the process has been threaded to execute individual hotel requests as separate processes.

See code for more details.

4 Scrape date, room price, hotel name and hotel description

We created a pipeline that is computationally efficiently, in order to gather the data, we had to implement packages and use web page analysis to gather specific data points.

We gathered the dates from the initial search as input parameters. We used xpaths with selenium to gather the name, prices and short description along with a few other data points from the hotel card container. In order to capture more detail for each hotel, we switched to the requests library, capturing the html page data, then porting that in to BeautifulSoup. Within BeautifulSoup, we relied heavily on CSS class names in order to capture data points in the hotel detail pages.

See code for more details.

5 Regressions

5.1 Regression 1

$$\text{Price} = \alpha_0 + \alpha_1 \cdot \text{TreatmentPeriod} + \varepsilon$$

α_1 captures the average difference in hotel rental prices between the event week and the non-event weeks, ignoring the city-specific effects.

5.2 Regression 2

$$\text{Price} = \gamma_0 + \gamma_2 \cdot \text{TreatmentCity} + u_i$$

γ_2 measures the average difference in hotel rental prices between Barcelona and Valencia, the control city, ignoring the effect of the event week.

5.3 Regression 3

$$\text{Price} = \beta_0 + \beta_1 \cdot \text{TreatmentCity} + \beta_2 \cdot \text{TreatmentPeriod} + \beta_3 \cdot (\text{TreatmentCity} \times \text{TreatmentPeriod}) + \nu_i$$

β_3 is the Difference-in-Differences (DiD) estimate of the event's effect on prices. This coefficient represents the difference in the before-and-after price changes in Barcelona relative to the before-and-after changes in Valencia. It captures the causal effect of the event on hotel booking prices in Barcelona, controlling for the overall time effect and the difference between the two cities.

5.4 Why The Need for a Second City

The choice of a second city, Valencia, is to serve as a control or as counterfactual for our treatment city, Barcelona. This helps to isolate the effect of Primavera Sound Festival from other confounders that might affect hotel booking or rental prices. By comparing Barcelona to Valencia (without Primavera Sound Festival), differences in hotel rental or booking prices can be attributed to the pure effects of Primavera Sound Festival, rather than to other city-specific factors or time variations.

6 Estimation

Before any estimation can be made, we preprocess our data. For better readability, we divide the prices by the number of nights (8). We tackle an issue with different units in the scraped "central distance" feature, where some distances under one kilometer are written in meters, while everything above is in kilometers (f.ex: 300m vs. 1.2km). We use KNN imputation to impute missing values and finally reduce our dataset to include only those accommodations that appear in both the eventweek and non-eventweek in order to ensure that they are comparable.

The results for the first three regressions are shown in the table below. The first shows higher prices on average during the defined *eventweek* (p-value below 0.01), which of course is not the coefficient we are looking for since the binary variable is not accounting for the different cities. On the other hand, the second regression only accounts for the difference in average prices between the two cities, where clearly Barcelona has a higher average price (p-value below 0.01), as we would expect. Finally, we take the two previously used variables and add their interaction term (*cityxevt*). This coefficient as we have mentioned is our DiD estimate of the event's effect on prices. The estimated coefficient is higher than the coefficient for *eventweek* in regression 1, which makes sense since there is no significant event happening in Valencia and thus average prices are very similar for both periods there. This is closely related to *eventweek* not being significant in regression 3, as then the entire effect on prices is being driven by the event happening in Barcelona, as reflected by *cityxevt* (p-value below 0.01).

	Dependent variable: price		
	(1)	(2)	(3)
city		104.364***	71.519***
		(5.351)	(6.985)
cityxevt			106.518***
			(10.539)
const	228.532***	180.868***	177.647***
	(3.448)	(4.268)	(5.892)
eventweek	47.823***		6.224
	(5.508)		(8.190)
Observations	3277	3277	3277
R^2	0.023	0.104	0.177
Adjusted R^2	0.022	0.104	0.176
Residual Std. Error	153.921 (df=3275)	147.359 (df=3275)	141.290 (df=3273)
F Statistic	75.384*** (df=1; 3275)	380.407*** (df=1; 3275)	234.398*** (df=3; 3273)
Note:			*p<0.1; **p<0.05; ***p<0.01

7 Final Estimation

We try the following methods to extract more features to control for differences between accommodations that can drive prices (See also the table below with the main statistics on the engineered features):

1. Extra controls: Apart from prices and hotel descriptions, we have also scraped: Ratings, distance to city center, review counts, latitude and longitude.

2. Distance to event: With the latitude and longitude data we compute the distance of the accommodations to the event location. Of course, this only makes sense for hotels in Barcelona, so the only way to make this feature work is to interact it with the *cityxevt* dummy. A statistically significant coefficient would then tell us that during the eventweek, being closer to the location is on average associated with higher prices. Unfortunately, this variable is not significant and we leave it out of the 4th regression.

3. Major brands: We have a csv of the most important hotel brands (Hilton, Novotel, Ibis, etc.) and create a dummy which is 1 if the accomodation belongs to one of the chains and 0 otherwise. We see that the mean price when this variable is 1 is aprox. 11€ (p. night) higher than when it is zero.

4. NLP based features: These features are chosen based on their tf-idf score, their p-value and coefficient in a regression on price (this is computed with a for loop that tries the best n-grams according to tf-idf as binary variables as another independent variable in regression (3)), and common sense.

   (a) Apartamento: Since we have scraped listings that are not only hotels but also apartments (typically more expensive), we think this is a very important control to include. In total, 2227 listings include this word, 1239 in Barcelona. The average price per night is 24€ higher for apartments.

   (b) Piscina: In order to control for major quality differences in the hotels we add this binary which is one if the accommodation has the word "piscina" (swimming pool) in its description. This is the case for 367 listings, out of which 239 are in Barcelona. The mean price (p. night) for the listings that include this word is aprox. 19€ higher.

   (c) Punto interés: A very important factor driving hotel prices is the proximity to cultural landmarks or more generally "points of interest". With this variable we hope to identify accommodations that are particularly close to those locations. An indication that this works well is that the average price is aprox. 33€ higher for listings containing "punto interés".

   (d) Wifi gratis: Again, we attempt to measure a quality difference in the availability of free wifi. Wifi is very common for hotels of average and higher quality so this is more of a control on the lower spectrum of listings. 1838 hotels have free wifi according to their descriptions, 625 of those are in Barcelona.

5. Sentiment Scores: We have implemented sentiment scoring, however the highest ranking description only achieves a score of 0.01. This could be down to the fact that while most descriptions are logically positive, they are written in a formal, professional style. We do not consider the sentiment scores in our final regression.

	count	mean	value count ones	value count Bcn	mean price 0	mean price 1
is major brand	3277.0	0.16	539	293	245.44	256.55
piscina	3277.0	0.11	367	239	245.48	261.45
wifi gratis	3277.0	0.56	1838	928	247.81	246.84
apartamento	3277.0	0.68	2227	1239	230.93	254.97
punto interes	3277.0	0.31	1002	625	237.21	270.11

The final regression then is as shown below. We see that the included independent variables are mostly statistically significant, with the exception of eventweek, which makes sense as explained before, and "is major event" which while it is not statistically significant we still want to consider as a control. Our DiD coefficient "cityxevt" has increased a bit more, according to it the event Primavera Sound Festival happening is associated with an increase in hotel prices of 113.314€ (p. night) on average, a sizable effect. By controlling for some characteristics that we have extracted through NLP, we are confident to have increased the precision of our estimate.

	Dependent variable: price	
	(1)	(2)
apartamento		56.319***
		(5.515)
city	71.519***	91.306***
	(6.985)	(7.086)
cityxevt	106.518***	113.314***
	(10.539)	(10.236)
const	177.647***	-19.745
	(5.892)	(20.739)
eventweek	6.224	5.511
	(8.190)	(7.938)
is major brand		10.708
		(6.517)
piscina		24.807***
		(7.776)
punto interes		29.332***
		(6.330)
rating		15.690***
		(2.285)
wifi gratis		11.127*
		(6.269)
Observations	3277	3277
R^2	0.177	0.229
Adjusted R^2	0.176	0.227
Residual Std. Error	141.290 (df=3273)	136.823 (df=3267)
F Statistic	234.398*** (df=3; 3273)	108.117*** (df=9; 3267)
Note:		*p<0.1; **p<0.05; ***p<0.01

8 Hotel fixed effects instead

If you run a regression with hotel fixed effects in the DiD setting, this controls for unobserved, time-invariant characteristics specific to each hotel that could influence the price dynamics. This further isolates the pure causal effect of the event, Primavera Sound Festival. This may change the treatment effect since it now controls for additional hotel-specific attributes or features that affect prices accounts.

The regression with controls extracted from hotel descriptions (such as amenities, rating, proximity to event location, etc) explicitly captures some of the hotel-specific characteristics unlike the fixed effect specification that captures same implicitly. Assuming these hotel descriptive controls are good ones, adequately capturing variation in prices due to hotel specific characteristics, the estimated treatment effect would be closer to the regression with hotel fixed effects as both are accounting for, or isolating effects on prices specific to each hotel from the event's effects.

References

Primavera Sound. (2024). *20th anniversary of primavera sound - an event of special interest.* Retrieved September 25, 2023, from https://www.primaverasound.com/en/news/20-aniversario-de-primavera-sound--acontecimiento-de-especial-interes

Visit Valencia. (2024). *Events in valencia in may 2024.* Retrieved September 25, 2023, from https://www.visitvalencia.com/en/events-valencia/month-events?date=2024-05