Exercise Homework 1

Introduction to Text Mining and Natural Language Processing

Luis Francisco Alvarez, Vanessa Kromm, Clarice Mottet

February 4, 2024

1

In this project, we will conduct an analysis of accommodation prices in Barcelona during the annual **Primavera Sound Festival**, taking place from May 30th to June 1st. This three-day event features a lineup of renowned international artists, including *'Vampire Weekend'*, *'Rels B'*, *'The National'*, *'Lana del Rey'*, and more. Notably, the festival stands as a unique attraction in Europe, as it is the only place where it occurs outside of Latin America. This has the potential to draw a substantial influx of both national and international tourists. Our focus will be on examining the dynamics of accommodation pricing during this festival period, shedding light on trends in Barcelona prices, influential factors, and potential implications for visitors.

2

For this, we will use Booking as our primary data source. Our approach involves scraping the website to extract valuable information from various accommodation offers. Key features for our analysis will include prices, ratings, distance from the city center, and the textual descriptions associated with each listing. We plan to implement a Natural Language Processing (NLP) method. This approach will enable us to extract relevant information from the textual data in the descriptions, allowing us to integrate more refined details into our model and will be further discussed in section 7. The scraping section of the project involves extracting information for accommodations in two consecutive weeks (27/05/2024-02/06/2024) and (20/05/2024 - 26/05/2024), where the first one is when we expect an increase in prices due to more people coming in, and the week previous to use as a control. We decided to use the week before the event because we expect generally higher prices in June because it is part of the main tourist season. The underlying assumption is that tourism will have a demand-side effect on accommodations in the week of the event and the control week considered the situation in Barcelona will normalize. We will also obtain data from another city, specifically Rome, for those same two weeks and construct a counterfactual trend based on it. The decision to choose Rome as the control city was made based on several similarities between the cities in terms of total population, culinary offerings, cultural diversity, and proximity to each other. We believe that Rome can serve as a helpful benchmark for comparing Barcelona when analyzing the effects of the festival on hotel rates. Since the objective is to analyze the effect of the festival on accommodation prices, we will construct a control to act as a counterfactual in a differences-in-differences regression. Considering two relevant treatment variables, time and city. The time variable will have a value of one for observations in the week of the event (27/05-02/06) and zero otherwise. On the other hand, the city variable will be worth one for all accommodations in Barcelona and zero otherwise. Finally, by multiplying both variables, we can identify the lodgings from the festival's week located in Barcelona; we aim to account for any significant difference from the rest.

3

After opening booking.com in Firefox, the pop-up to accept the cookies is closed automatically. The pop-up to log in with the Google account has to be closed manually. It was not possible to close it the same way as the cookie pop-up, probably because it is embedded in an iframe. Since the place and the dates will have to be changed during the scraping process, we decided to define functions for that to avoid redundancy in the code. Notice that the function to choose the dates has a parameter that decides whether the calendar is scrolled to other months. This is only necessary for the first search. With the function get_information all hotels from all pages are scraped. The information that are extracted and saved in lists are hotel_name, rating, room_description, price, location_description and the hotel_link. If one of the elements is not available, a na-entry is added to the list. Like that, it is avoided that the code crashes in case one element is not extractable. After extracting all the information a dataframe is created. With the urls and the help of BeautifulSoup, the description of every hotel is extracted and also added to the dataframe. To then extract the prices for the control week, another function is used: this function extracts only the hotel_name and the price for the week. Like that, the code is kept efficient, because the other information are not needed again since they don't change (e.g. description). All the created dataframes are stored as csv files, so that other team members don't have to do the scraping again and can just load the data as csv to work with it. Further comments can be found in the Jupyter notebook.

4

The data we will gather from the accommodations is the following:

- Hotel name
- Rating
- Room Description
- Price
- Location Description
- Link
- Long Description

For this exercise, we decided to scrape data for the week of interest (the treatment week) and save all relevant information. For our analysis we have chosen to only include hotels that appear in both the control time period population (the week before) and the treatment time period population (the week during the event). By working with the same hotel population in both weeks, we're controlling for outliers that may effect our ability to determine price increases from one week to the next. Otherwise, we might have difficulty in being able to discern a variation in price tied to the event in question. For the control week, we limited

our data extraction to the name and price of the hotel, since we are going to use the same attributes as those for the hotels during the treatment week.

5

Price Prediction based only on Time Period:

$$Y = \beta_t X_t + \beta_{0t}$$

Where Y represents price, X_t is our input variable based on time period t. It acts as an indicator variable that takes the value of 1 during the treatment time period (during the week of the event) and 0 otherwise (the week previous). The regressor β_t captures the average treatment effect from one week to the next. Where β_0 is our typical y-intercept.

Price Prediction based only on City:

$$Y = \beta_c X_c + \beta_{0c}$$

Where Y represents price, X_c is our input variable for city c and acts as an indicator variable that takes the value of 1 if the observation is from the Barcelona and 0 if the observation is from Rome. The regresor β_c captures the average treatment effect of a hotel stay in Rome to a hotel stay in Barcelona. Again β_0 is our typical y-intercept.

Difference in Difference regression based on Time Period and City:

$$Y = \beta_t X_t + \beta_c X_c + \beta_{t,c} X_{t,c} + \beta_{0t,c}$$

As before, Y is the price, X_t, is a city indicator input, X_c is a time period indicator input. In this equation we include an interaction term $X_{t,c}$ to capture the case when X_t and X_c are the treatment for both the time period and city. Here, $X_{t,c}$ indicates when X_t is 1 (week during the event) and when X_c is 1 (in Barcelona) and is 0 otherwise. The β_0 value represents the average price of hotel stay during our control time period (X_t is 0, week before Primavera Sound). The β_t represents the average price change in our control city from the control week to the treatment week. The β_c represents the average price change from our control city to our treatment city (Barcelona) during the control week. Lastly, $\beta_{t,c}$ represents the average price change in our treatment city (Barcelona) from the control time period to the treatment time period. This coefficient captures the average treatment effect on prices, isolating the impact that cannot be explained by time trends or pre-existing differences between the treatment and control city. We will be able to use these β values to determine price changes from the control time period to the treatment time period and with respect to both cities to determine if the Primavera Sound event correlates to an increase in hotel stay prices in Barcelona.

It is crucial having a second city so we are able to construct the counterfactual. For example, if we saw a 20% increase in prices for Barcelona observations, we would not be able to confer the effect to the Primavera Sound if we observed a similar trend in Rome although there is no festival there. Otherwise, without Rome, there is no scenario where we have data **without** the festival taking place.

6

The output from the three regressions is presented in Table 1.
In both Model 1 and Model 2, when we consider just one treatment variable, the difference in price is significant. We also see there is a significant difference in average prices for hotel stay in Rome and Barcelona, as well as in time period independent of each other.
In Model 1 we're looking at both the Barcelona and Rome to understand how the overall average price of hotels changes from the week before the Primavera Sound event and the week during the event. We're seeing the overall average price of hotels the week before the Primavera Sound event is approximately 1,713€ as given by the intercept of the equation. The following week has an average hotel price increase of approximately 150€ given by the coefficient of our time period treatment indicator variable. This tells us that on average, we see an increase in price from week to week.
In Model 2 we are not distinguishing between weeks to understand how the price of hotels differs between cities. With our city indicator $X_c = 1$ if Barcelona and 0 otherwise, we see that the average price of a hotel stay for a week in Rome is approximately 1,449€ as given by the intercept term. We're seeing that the average price of a hotel stay for a week in Barcelona is on average approximately 676€ more than what a stay in Rome would be, as given the regression coefficient value β_c.

In Model 3, we've created an interaction term specifically for the case where we're looking to see the hotel price rate change for the case where a person is in Barcelona and during the event. First we look at the intercept which has the value 1,461€ representing the average stay in a hotel in Rome the week before the Primavera Event. For the same week, we see that a stay in Barcelona would be on average 1,963€ before the event. For the week during the event we see that the average price in Barcelona increases by 355€ as indicated by the $\beta_{t,c}$ coefficient. This marks an 18% increase between the two weeks in Barcelona.

The average price in Rome goes from 1,461€ before the event to 1435.9€ during the week of the event with a 25.60€ reduction in price resulting in an approximate 2% decrease.
While we cannot attribute the increase in price in Barcelona to the event, we can say that hotel prices do increase in Barcelona between the two weeks whereas we do not see the same magnitude of price change in another European travel destination.
Equation 3 details the difference in difference analysis that we're looking for because it enables us to look at the percent increase in price between the time period in question and have an analogous city to compare it to. In the first two equations we were unable to determine how to attribute a change in rate depending on city or time period, but in

equation two, we're able to look at the averages of the 4 cases: (before event, Barcelona), (during event, Barcelona), (before event, Rome) and (during event, Rome) corresponding to zeros and ones with the dummy variables to conduct a hotel price behavior analysis.
Finally, Equation 4 consists of the DiD regressions with controls. Results are presented, although further explanation is in the following exercises.

Table 1: Regression Results Summary

	Time Treatment	City Treatment	DiD	DiD with Controls
β_0	1713.9^{***}	1448.8^{***}	1461.5^{***}	964.3^{***}
	(30.5)	(29.5)	(41.4)	(52.9)
β_t	149.5^{***}		-25.6	-24.8
	(43.4)		(58.8)	(48.3)
β_c		675.8^{***}	501.05^{***}	143.1^{**}
		(41.7)	(58.3)	(50.4)
$\beta_{t,c}$			355.1^{***}	380.1^{***}
			(83)	(68.3)
Controls	No	No	No	Yes

Standard errors are in parenthesis; [*] p <0.05; [**] p <0.01; [***] p <0.001

7

Our goal with adding in control features is to further infer whether the price change from one week to another by city is due to the event or due to these features. Potential control variables are the following: Ratings, Distance from City Centre, Number of Beds, Free Cancellation, Breakfast Included, and whether "few" were left at this price.

7.1 Feature Discussion

Ratings was a feature easily scraped. Out of the 1975 hotels we scrapped, 47 did not have rating information; we filled in these null values with the average across all hotels regardless of location. Ratings might have an effect on the price where hotels with higher ratings might have higher prices.
We were able to extract the Distance from City Centre (in meters) using regex from the "Long Description" field for all hotels that we have represented. Distance from city centre might have an effect on the price where hotels with a lower distance may be more expensive due to being in a more desired part of the city.
Within the "Room Description" field, we were able to extract: number of double beds, single beds, sofa beds, and bunk beds using a team-built function relying heavily on regex and split methods. There were only approximately 14 out of the 1975 instances where number of beds fields were manually determined. In cases where the room description stated "multiple beds" or did not have information, we filled in the specific bed count features (number of double beds, single beds, sofa beds, and bunk beds) with the mode and filled in the total number of beds with the mode across all information gathered as well. These features may have

an effect on price, where looking for a hotel stay with sofa beds or bunk beds may indicate people looking for more comprehensive yet affordable accommodations.
If the following phrases appeared in the room_description, we one-hot encoded these to be potentially used in the regression: "Free Cancellation", "X left at this price" (as there were times where it was one or five beds left at a certain price), and "Breakfast Included". All these phrases have the potential to have an effect on price.

7.2 Feature Summary

We provide summary statistics on all of these potential regression features by city. To determine which features we will include in the model, we create a LASSO regression to predict price and use the magnitude of the coefficients and also detect sparsity from our explanatory variables to gain insight into which features have the largest impact on price. By incorporating the two features with the largest LASSO regression coefficient into our third regression equation from problem 6, we hope to further analyze the magnitude of price change that could be attributable to the event in Barcelona and not these other hotel features.

Table 2: Barcelona Summary: General Features

Metric	Rating	Distance From City Centre	Total Number of Beds
mean	7.92	1612.54	2.34
std	1.12	1037.56	1.48
min	1.00	100.00	1.00
max	10.00	7700.00	13.00

Table 3: Rome Summary: General Features

Metric	Rating	Distance From City Centre	Total Number of Beds
mean	8.45	2237.82	1.56
std	0.68	1486.91	0.86
min	5.50	150.00	1.00
max	10.00	10000.00	8.00

An interesting observation is that Barcelona has an 2.34 average total number of beds whereas Rome has an average of 1.56 meaning that on average hotels in Barcelona usually have one more bed available than hotels in Rome; this can have an impact on prices and is evidence in favour of adding the control variable.

Table 4: Barcelona Summary: Number of Bed Types

Metric	Double Beds	Single Beds	Sofa Beds	Bunk Beds
mean	0.78	1.25	0.21	0.08
std	0.75	1.35	0.45	0.47
min	0.00	0.00	0.00	0.00
max	5.00	9.00	4.00	8.00

Table 5: Rome Summary: Number of Bed Types

Metric	Double Beds	Single Beds	Sofa Beds	Bunk Beds
mean	0.88	0.44	0.19	0.01
std	0.58	0.80	0.47	0.12
min	0.00	0.00	0.00	0.00
max	6.00	3.00	4.00	2.00

Looking at the mean of bed types between the two cities we see a generally similar distribution except for the number of single beds where Barcelona has a higher occurrence of hotels offering accommodations that include a single bed.

Table 6: Barcelona Summary: Additional Features

Metric	Few Left at this Price	Free Cancellation	Breakfast Included
mean	0.74	0.5	0.04
std	0.44	0.5	0.19
min	0.00	0.0	0.00
max	1.00	1.0	1.00

Table 7: Rome Summary: Additional Features

Metric	Few Left at this Price	Free Cancellation	Breakfast Included
mean	0.95	0.77	0.21
std	0.22	0.42	0.41
min	0.00	0.00	0.00
max	1.00	1.00	1.00

Another difference we can see between the two cities and not the time period is that 4% of hotels in our population have breakfast included in Barcelona whereas 20% of the hotels in Rome do. Free cancellation is also more common at 77% in Rome instead of 50% in Barcelona. Between both subsets, "Few Left at this Price" is relatively high with 95% in Rome and 74% in Barcelona.

7.3 Feature Selection Using LassoCV

Table 8: Lasso Feature Importance

Feature	Lasso Coefficient	Lasso Coefficient Importance Scaled
total_number_of_beds	473.627064	1.000000
number_of_double_beds	314.806519	0.664672
distance_city_centre_meter	-254.199722	0.536709
timeperiod_city_interaction	139.727296	0.295015
city	134.586470	0.284161
rating	126.904341	0.267941
free_cancellation	87.969794	0.185736
breakfast_included	79.063701	0.166932
number_of_sofa_beds	-71.298977	0.150538
number_of_bunk_beds	-36.856743	0.077818
few_left_at_this_price	-32.231826	0.068053
number_of_single_beds	0.000000	0.000000
timeperiod	0.000000	0.000000

Using `LassoCV` we use 5-fold cross validation to determine the best alpha parameter to use for a price prediction model. Taking the absolute values of the coefficients and dividing by the maximum, we calculate a feature importance metric. The most important feature by far in predicting price appears to be *Total Number of Beds* followed by *Number of Double Beds* and *Distance from City Centre*. For our fourth regression model, we will be including all three of these features to further analyze the price change in hotels from one week to the next with an event in Barcelona. The high importance of the features can also be seen graphically below:

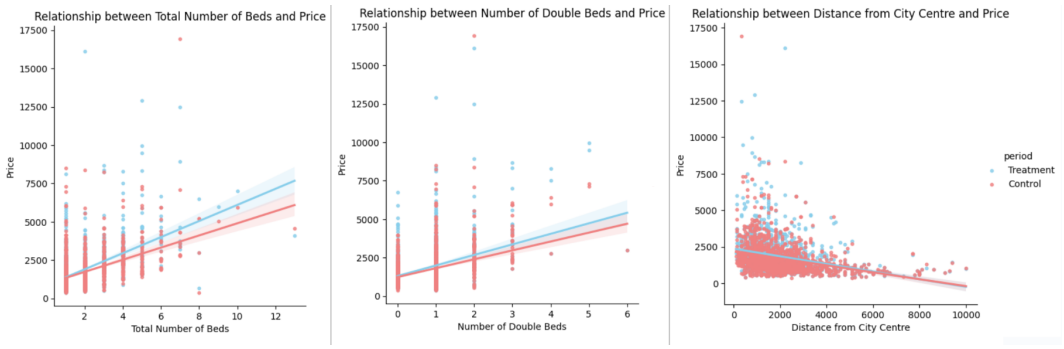

Figure 1: Relationship between the most important features and price

7.4 Model Analysis and Comparison

Table 9: Model 4 Parameter Values

Parameter	Coefficient
Intercept	964.278740
Timeperiod	-24.827301
City	143.124614
Interaction	380.095868
Total Number of Beds	342.084583
Number of Double Beds	470.727212
Distance From City Centre	-0.200821

Due to the "Total Number of Beds" coefficients and the "Number of Double Beds" coefficients being quite large, we can already see that the relationship of price between time period and city has decreased.

To compare this model to model 3, we will assume "Total Number of Beds", "Number of Double Beds" and "Distance from City Centre" is zero.

With this model we see that the average price of a hotel stay in Rome the week before the Primavera Sound event (in BCN) is approximately 964.28€ and has a -24.82€ decrease in price going into the next week for a 2.5% decrease in price from one week to the next. We see that the average price of a hotel stay in Barcelona the week before the Primavera Sound event is approximately 1,107.40€ (964.28€ + 143.12€ and increases by 380.10€ in price for the week of the event resulting in a 34% price increase.

Previously we saw a 2% (-25.6/1461.50) decrease in price between the weeks in Rome and an 18% (355.1/(1461.50+501.05)) increase in price between the weeks in Barcelona. In this model, we're seeing the same price change behavior between time period in Rome but for Barcelona we're seeing a more drastic price increase.

By adding in more information through control variables, we're able to grasp a better understanding of price changes from one week to the next that is not attributable to number of beds or distance from the city centre.

In this model, we're seeing a more compelling argument (34% price increase instead of 18%) that during the week of the Primavera Sound event in Barcelona, hotel prices increase as compared to a similar European travel destination, Rome.

8

In our case, we already ran the regressions with hotel fixed effects where the we only included hotels that appeared in both our time period control and treatment population. The treatment effect like this is different to the treatment effect when you wouldn't use the exact same hotels, because you may have different observations coming into play in the treatment and/or control group, that will certainly influence the means differently. When we compare a DiD regression to FE, the later one will capture all the individual effects from the hotels; rather than the average effect between weeks and cities. That is why, when we introduce controls, therefore keeping constant some individual characteristics from the accommoda-

tions in the regression , part of this distortion would be equalised and the estimation would be closer than if we just use the treatments as predictors.