# Problem Set 3
## Text Mining: Models and Algorithms

Gannavarapu ∘ Schindler ∘ Susanu ∘ Virgüez

Data Science for Decision Making, Class of '22
Barcelona School of Economics

Submission Date: Jan 29, 2022

---

# Part 1: Booking.com Scraping

## Event and Brief Data Description

The event that we chose for this analysis is the upcoming women's football match: *FC Barcelona Femení vs. Real Madrid Femenino*. The match is scheduled to take place on the 30$^{\text{th}}$ of March, 2022. Tickets for the event are set to be sold out (for the first time for a women's football match at the Camp Nou) with many visitors coming in to watch the match.

To analyze how prices would change in response to this event, we would need to analyze prices before and after the event to see if there is a significant change in them during the window around the event. We scrape hotel prices during the event, with a 7-day window prior to the match and after. Our scraping results indicate that there are 872 properties listed for availability on Booking.com during the period of the analysis.

## Identification Strategy

The match is scheduled on the 30$^{\text{th}}$ of March. We code the 30$^{\text{th}}$ as the *treatment*, to see if the prices are increased for that day. This strategy relies on the assumption that people would only come in to see the match and leave the next day. This is a strong assumption and we discuss it further in the section for potential problems.

Since the properties are heterogeneous, we would have to control for their individual characteristics. From our scraping, we are able to obtain features such as the type of room, distance to the metro, number of reviews, rating, etc. As the hotel features remain constant throughout, we need not consider any hotel fixed effects. However, hotel prices vary depending on the day of the week, hence we include day-of-the-week fixed effects, to incorporate a temporal element.

## Potential Problems

We consider stay lengths of a single day for one person. It might be the case that people coming in to watch the match, do so with their friends and family and stay longer. This might cause the prices for non-single rooms to increase and have no effect on single rooms. We also only consider the day of the event as treatment. People who come to see the match might stay for more than one day. Our identification strategy does not capture this aspect.

Another problem we identified is the use of cookies by such websites, which might change the

prices shown as the scraper connects to it multiple times. As such, the prices might not be indicative of the features of the hotel and the event, but rather the number of times we searched the website.

Another issues is that the timing at which scraping is done also influences results as room availabilities are dynamic. We noticed while testing the script, that a difference of an hour (everything else equal) resulted in less availabilities (for a particular search day). Thus searches made far into the future might not capture the effects of the event of interest.

# Model

$$Price_{gt} = \beta \boldsymbol{X_{gt}} + \zeta_t + \epsilon_{gt} \tag{1}$$

where $\boldsymbol{X_{gt}}$ is a vector of hotel characteristics, $\zeta_t$ are the temporal fixed effects and $\epsilon_{gt}$ is the error term.

# Results

The results of regressing equation (1) are presented in Table 1. Our results indicate that there is no significant effect of the football match on hotel prices. Of course, this result is subject to the assumptions and shortcomings of our model as mentioned in the previous sections.

As seen in column (1) of Table 1, running a naïve regression without taking into effect of the day-effect yields results that go in a completely different direction (change of sign of effect-of-interest).

Table 1: Hotel Prices

| | (1) price1 | (2) price1 |
|---|---|---|
| Event | -12.86*** | 0.522 |
| | (-18.82) | (0.79) |
| Apartment | 127.5** | 128.0** |
| | (2.58) | (2.60) |
| Rating | 40.19*** | 39.87*** |
| | (9.10) | (9.01) |
| Prepayment | 6.610* | 5.983* |
| | (2.35) | (2.26) |
| Distance | -0.00441** | -0.00429** |
| | (-2.87) | (-2.89) |
| # Revies | -0.0243 | -0.0249 |
| | (-0.99) | (-1.01) |
| $N$ | 10318 | 10318 |
| Day FE | No | Yes |

$t$ statistics in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

# Discussion

We were expecting to find positive results on hotel prices due to the first sold-out women's football match at the Camp Nou. Our preliminary naïve results do not support this hypothesis.

Some explanations could include that it may mainly be local people attending the match and hence they would not need hotel accomodation. In this way demand for hotels would not increase and hence neither would prices.

A further extension of the analysis could be to restrict hotel searches for only the local area of *Les Corts* which is the neighborhood where the stadium is located. Hotel prices in this area might be more strongly correlated with football matches. Unfortunately, time constraints did not allow for this.