Exercise Homework 1 - Introduction to Text Mining

Team Members: Oriol Gelabert, Enzo Infantes & Tarang Kadyan

February 05, 2025

1 Part 1: Scraping

1.1 Experimental design

Our starting point is identifying a popular event at the city of Barcelona. As a global city, plenty of events take place during the whole year but one of the most worldwide known, which attracts a lot of companies and people is the Mobile World Congress. It is one of the biggest and most significant events in the telecommunications and technology industries where global leaders from tech, mobile, and innovation industries gather to showcase the latest in mobile technology, 5G, AI, etc.

Thus, we decided to choose this event that attracts biggest tech companies to Barcelona to study how the presence of a mass event can affect the prices on hotel stays at Barcelona. As we decided Mobile World Congress, we have to select the dates used for our research. To avoid possible variation between weekdays we will select a whole week as our study period during MWC. As MWC is hosted between 3-6 March, our study period will be from 1st to 8th of this same month.

We have to think about a control city to obtain comparable results for our experiment. To avoid the effect of MWC we will not use Spanish cities as some people could use the opportunity of visiting Barcelona to visit adjacent cities like Madrid or Valencia. Also, an important factor to consider is that the control city has not to host important events during the studied period. After some research, and comparing cities with similar number of visitors, population and hotels, we decided to use Milan as our control group. This city, also famous worldwide as Barcelona, has a number of hotels and visitors similar to Barcelona and thus is an appropriate comparison for this subject. Additionally it does not host any important event during MWC, which assures us that the flow of visitors is not incremented by the presence of a festival, congress, ...

After fixing control city we have to think about a comparable time span. As explained before to be as rigorous as possible we will use a similar time span, in our case one week to compare the prices. Then, we have to take into consideration that neither Barcelona or Milan have to host events during those days. As a result, the first days of August, the week between 1st and 8th (where no mass events are hosted in any of the two cities), will be used as comparison period to track the changes on both cities with respect to the study period where the studied event is hosted.

1.2 Scrapping

After defining from which periods and cities we are required to gather information, we start constructing a web scrapping pipeline to extract information from the *Booking.com* website. In this pipeline, we will use *selenium* package in python to perform *"human actions"* in order to browse for hotels in the selected cities and periods. Once the browsing has been done, then we retrieve some information from the properties shown in the hotel list displayed after searching. The extracted information is:

- Hotel name : Get a identifier for each hotel, which will be the name
- Stars classification: The number of stars assigned to each hotel or apartment.
- Rating : The rating that guests have given to an hotel or apartment.
- Location: The neighborhood where an hotel is located.
- Distance to the center: Distance to the city center of each hotel
- Price: The price of one week stay for 2 guests for each hotel.
- Link: Extract the link (https://) for each hotel.

Also we are required to extract the description of each hotel considered. In order to do that, using *sellenium* would be very time consuming as we should enter each hotel webpage, go back, enter the following, etc. To avoid all these steps we can use python packages *requests* and *bs4* to access each website using the links (that we extracted previously using *sellenium*) and then retrieving the html content of the website from which we can retrieve the text in the description.

To allow simpler manipulation, each period and city information is stored into its singular data frame along with the description provided for the corresponding hotel. Thus, we result in 4 data frames that are finally stored in **.csv** files in order to be exported or used independently in the second and third part of this project.

1.3 Pipelines

We created a project to perform these steps efficiently. For this reason, we have the following **.py** files, which contain important functions used to search, extract, and save all the data about hotels. The next schema represent all the folders and files we used on this project:

```
textmining_booking/
|-- booking/
|   |-- packages/
|   |   |-- __pycache__/
|   |   |-- __init__.py          # Package initialization file
|   |   |-- dataloading.py       # Data loading and cleaning
|   |   |-- processing.py        # Data processing
|   |   |-- scraper.py           # Web scraper from Booking
|   |   |-- selenium_setup.py    # Selenium setup for scraping
|   |-- Barcelona_MWC.csv        # Data extracted from Barcelona
|   |-- Milan_MWC.csv            # Data extracted from Milan
|   |-- geckodriver.exe          # Selenium driver for Firefox
|-- ITM_HW1.ipynb                # Principal Notebook
|-- hw1.pdf                      # Homework 1 questions
|-- README.md                    # Project structure file
|-- requirements.txt             # Dependencies required to run
```

```
|-- setup.py                    # Installation and setup script
```

In the `packages` folder, there are the files `processing.py`, `selenium_setup.py`, `scraper.py`, and `dataloading.py`. These Python files contain all the functions needed to search for, extract, and process hotel data.

Using the main notebook for this project (`ITM_HW1.ipynb`), we called all these functions to retrieve and process the data. For instance, in `scraper.py`, there is the `BookingScraper` class, which is crucial to our process because it calls other functions within it to select the destination, choose the check-in and check-out dates, reject all cookies, and search for all available results.

1.3.1 Installation and Usage

To install the required dependencies, run:

```
pip install -r requirements.txt
```

1.3.2 Usage

Selenium Setup

- Download `geckodriver.exe` for Firefox or use the appropriate driver for Chrome.
- Place it in the project folder.

Run the Files

```
python packages/scraper.py
python packages/dataloading.py
python packages/processing.py
```

These files generate searches on Booking webpages, extract data according to our delimitations, and preprocess the description of each hotel.

Data Analysis

- Run the `ITM_HW1.ipynb` notebook in Jupyter Notebook to visualize exploratory analysis, data cleaning, and the DiD regression.
- In this notebook, we use pipelines to call all the functions from the `.py` files.

Finally, we can see that the code follows the instructions on automation, efficiency, and the rest of the best practices for this kind of projec

2 Part 2: Text Analysis

2.1 Tokenization and Stemming

After the extraction of hotel descriptions as described in the previous section, now we can use this data in order to obtain some ideas about the more common words and concepts present in the hotel descriptions of each of the cities.

In order to be able to analyze the text, some pre-processing must be done in advance to achieve reliable results on this task. First of all, we must tokenize the words of each description by the following steps:

- Transform the descriptions to lower case to avoid capitalization mismatches.
- Eliminate punctuation signs from the words that do not provide information.

Once the words are tokenized, we can filter some stopwords of spanish language as well as some inferred words like *"hotel", "habitacion", "huespedes", "ubicacion", 'barcelona','milán'* which we identify that can be in an important fraction of the descriptions and thus are not providing relevant information for our study.

Finally we lemmatize our tokenized words in order to extract the *lemma* of our words, this is considering words with same semanthic origin as equal so words like *"playa"* or *"playas"* are considered as the same word in our pre-processed descriptions.

2.2 Wordclouds

Once the pre-processing has been done, we can generate worldclouds for each city, before and after pre-processing that might help us to identify some structural or important words that we can use as controls on the DiD as well as what words are used more by the hotels in their descriptions.

The WordClouds for Barcelona are ploted as following:

Figure 1: WordCloud for Barcelona hotels before Pre-processing.

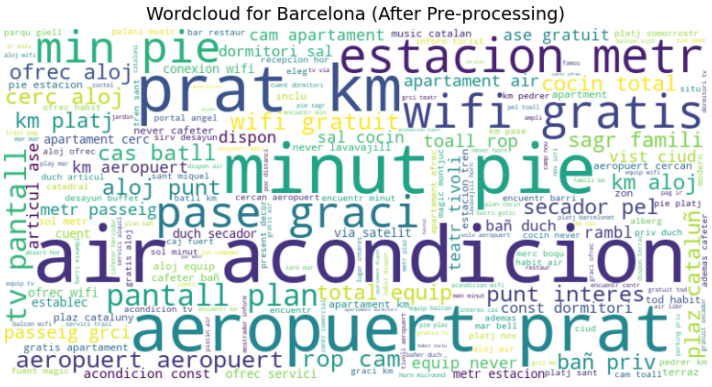

Figure 2: WordCloud for Barcelona hotels after Pre-processing.

We can identify important differences between pre-processed wordcloud and the raw text one. As expected, raw descriptions contained a huge amount of stopwords like *"de","la","en"* or even directly the name of the city. This caused the wordcloud to give importance on meaningless words and not focus in words that can provide us more information.

If we observe the wordcloud generated after pre-processing, we can identify that the words with more weight in the wordcloud are now not prepositions or articles but commodities of the hotel or some distinctive treats. We can distinguish not words but lemmas like:
"air acondicion","aeropuert","estacion","metr","gratis","wifi" among others.

If we look at Milan generated wordclouds now we can observe the following results:

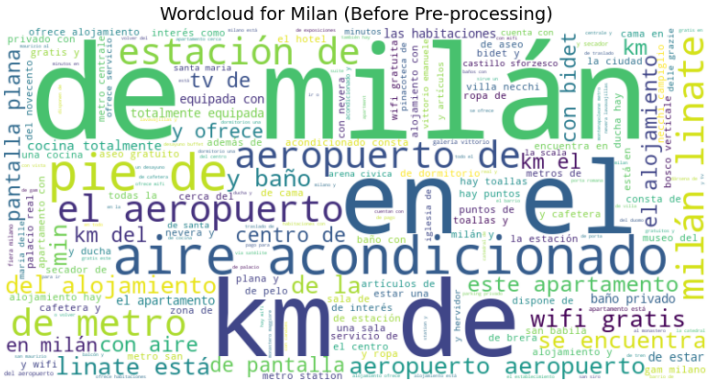

Figure 3: WordCloud for Milan hotels before Pre-processing.

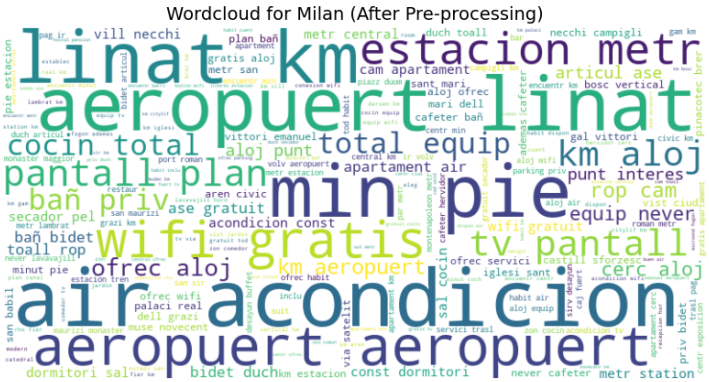

Figure 4: WordCloud for Milan hotels after Pre-processing.

As deducted in Barcelona wordclouds, Milan ones exhibit a similar behaviour. In the raw text we can identify a huge presence of stopwords and non-informative words, while in the pre-processed text characteristics of the hotels and amenities now arouse and can be easily identified.

Now, using the pre-processed wordclouds we can gather some extra information from the description of the hotels. A sentiment analysis could be performed for each description, but this might be too complex for this project. Instead, we decided to identify the presence of some clue words, that can help us give an idea of what the hotel is offering and how that variables can impact the price change. The selected variables, consisting in one or more words are the following:

- Desayuno (Breakfast) : Hotels offering breakfast tend to charge a premium.
- Cocina (Kitchen) : The presence of a kitchen may influence pricing, especially for long-term stays.
- Air (Air Conditioning) : In warmer climates, like Barcelona, air conditioning is often a necessity, affecting demand.
- Piscina (Swimming Pool) : Pools, particularly in luxury hotels, are often associated with higher prices.

Thus we filter the descriptions for each hotel and assign to each one of the variables a 1 if at least one of the words conforming the list is present in the description and a 0 otherwise.

Each of these features was converted into binary variables (1 if the word was present in the description, 0 if absent). These were then added as explanatory variables in the regression to assess their influence on hotel prices.

3 Part 3: DiD

Last part of this project consists of using the information extracted via web scrapping and the pre-processed text in order to understand how the presence of an event might alter the price of hotels in Barcelona. To study this phenomena, we apply a difference in difference regression using the treated group (Barcelona) and the control one (Milan) on two different time periods (MWC and first week of August).

First of all, we must consider the data that we gathered. As MWC is a close event, many of the hotels scrapped from August have not been scrapped for MWC as they were no longer available. If we use all the data without selecting it then, we would introduce a huge bias on the study, as probably the hotels left now are not as good as the ones that we have not been able to study and are present on August. With this objective, from now we only consider hotels that were present in both MWC and August dataframes and thus we will be working with less number of observations.

In order to be able to perform a complete study, we collapse our 4 dataframes into a singular one which contains all the information. For this we create three new variables: *Treatment*,*Period* and a *DiD* variable. *Treatment* is a binary variable that will refer with a 1 for hotels in Barcelona and a 0 for those in Milan. Similarly, *Period* will input a 0 for hotels in August dates and a 1 for those of MWC dates. The DiD variable is simply an interaction term between this two variables used for DiD Regressions.

Then the regression model will look like this:

$$\text{price}_{itp} = \beta_0 + \beta_1(\text{treated}_t) + \beta_2(\text{period}_p) + \beta_3(\text{treated}_t \times \text{period}_p) + C_i B_i + \epsilon_{itp}$$

In this equation:

- *price* is the dependent variable.
- β_0 is the intercept
- β_1 indicates the price difference between cities.
- β_2 captures the price difference between periods.
- β_3 is the DiD term that measures the effect of the event on Barcelona.
- C_i is the vector of control variables.
- B_i is the corresponding coefficients of this additional control variables.
- ϵ_{itp} is the term.

The indices in the equation correspond to: $\boldsymbol{i}$ for each observed hotel, $\boldsymbol{t}$indicates treated city with 1 (Barcelona) and 0 to the control (Milan) and $\boldsymbol{p}$ for period (1 during MWC dates, 0 otherwise).

We have to note that we used the logarithm of this regression equation, as it helps us to understand the effects in terms of price percentage rather than price absolute value, which give us a clearer idea of the effect.

3.1 First approach

Our first approach we will only use along the DiD variables, only information extracted directly from the hotel and not from the description as controls. Thus, the control variables used for this first regression will be:

- Hotel Ratings.
- Stars Classification.
- Distance to city center.

We are going to use this control variables as they can give us an idea of the quality of an hotel which surely has an implication on its price. Good ratings and star classification give us prior information of in a hotel might be good or bad, which implies that consumers might be willing to pay a higher price for a room. Centrality of the hotel might also be an important variable, but could also be affected by the fact that MWC is hosted far from the city center so its congress assistants might not be necessarily interested in centric hotels as typical tourist do.

Applying an OLS regression on the described model the results obtained are the following:

OLS Regression Results

Dep. Variable:	Price	R-squared:	0.548
Model:	OLS	Adj. R-squared:	0.547
Method:	Least Squares	F-statistic:	385.2
Date:	Wed, 05 Feb 2025	Prob (F-statistic):	8.02e-304
Time:	19:03:58	Log-Likelihood:	-578.52
No. Observations:	1571	AIC:	1171.
Df Residuals:	1564	BIC:	1209.
Df Model:	6		
Covariance Type:	HC1		

	coef	std err	z	P>|z|	[0.025	0.975]
Intercept	5.6324	0.145	38.819	0.000	5.348	5.917
Treated	0.2132	0.030	7.169	0.000	0.155	0.272
Period	-0.0124	0.022	-0.559	0.576	-0.056	0.031
DiD	0.5417	0.037	14.822	0.000	0.470	0.613
Rating	0.1420	0.018	8.040	0.000	0.107	0.177
Stars	0.1873	0.013	14.575	0.000	0.162	0.212
Center_Distance	-0.0988	0.007	-14.548	0.000	-0.112	-0.085

Omnibus:	452.098	Durbin-Watson:	1.631
Prob(Omnibus):	0.000	Jarque-Bera (JB):	1488.714
Skew:	1.414	Prob(JB):	0.00

...

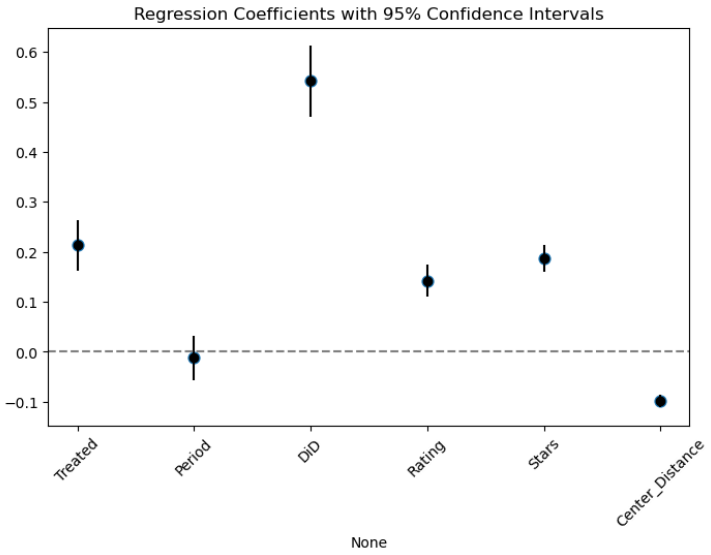

Figure 5: Coefficients and confidence intervals on this first approach.

Model Summary:

- **R-squared** = 0.548 → 54.8% of the variance in price is explained by the model.
- **Adjusted R-squared** = 0.547 → Adjusted for great number of predictors, indicating good model fit.
- **F-statistic** = 316.5 → Highly significant, meaning the model overall explains a significant amount of variation in price.

Coefficient Interpretations:

- **Treated:** Being in the treated city (Barcelona) increases log(Price) by 0.2132. This means that hotels in Barcelona tend to be 23% more expensive that Milan ones.
- **Period:** The effect of time is statistically insignificant (p = 0.587), meaning price changes over time alone are not significant.
- **DiD:** The DiD coefficient is positive and significant, meaning prices in Barcelona increased significantly during the MWC event, translating to a 72% increase in price.
- **Rating:** Each point increase in consumers rating means an increase on the price by 15.3%.
- **Stars:** Hotels with higher star classification charge higher prices. A one-star increase leads to a 20.6% price increase.
- **Center_Distance:** Negative and significant. It indicates that hotels with an increasing distance from the city center charge cheaper fees. An increase of 1 km distance reduces price by 9.4%.

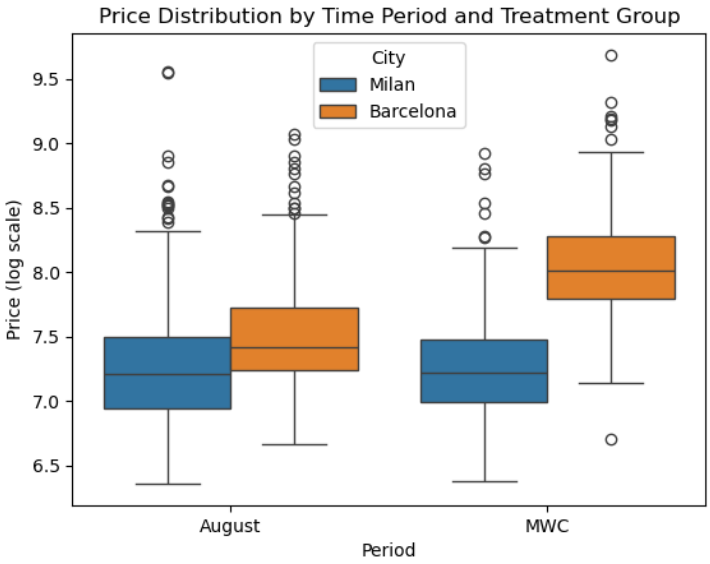

Figure 6: Prices comparison box plots.

Important takeaways from our results:

- **The MWC Event Increases Prices:** The DiD coefficient is strongly positive and significant, confirming that hotel prices in Barcelona increase significantly (around 70 percent) during MWC. This suggests that high demand during the event leads to price surges, making Barcelona an expensive destination for that period.
- **Hotel Star Rating and Location:** Hotels with more stars charge higher prices, and hotels farther from the city center tend to have lower prices.
- **Significant Price Variation Explained:** The model explains a significant portion of price variation, confirming that event-driven demand, hotel quality, and location are key price determinants.

In conclusion, the MWC event creates a significant price surge, reinforcing the idea that major events drive up accommodation costs in Barcelona. As we can see in Figure 6, There is a high price increase in hotels from Barcelona during MWC while the price of hotels in Milan looks very similar.

Lastly we have to consider why it is important to consider why using a second city to compare the prices is crucial. In DiD analysis, the second city helps us to stablish a trend that would had been maintained in the absence of the event. By this we mean that the second city provides us of a non-treatment situation of Barcelona on the dates of MWC if the event was never hosted. This permits us then to compare the real situation and the 'created' one and thus estimate the impact of the event on the hotel prices of the studied city, in our case Barcelona.

3.2 Second Approach

In a second approach we are required to use the text features that have been scraped in the OLS regressor. Our approach then will be using the binary variables that we created in part 2, indicating the presence of certain words in the hotel description. In this way we might identify if a hotel offers breakfast or has amenities like swimming pool or air conditioner. We run an OLS regression maintaining the same variables as previous approach but this time we include the following controls:

After the inclusion of this additional controls we obtained the following regression:

OLS Regression Results

Dep. Variable:	Price	R-squared:	0.566
Model:	OLS	Adj. R-squared:	0.563
Method:	Least Squares	F-statistic:	222.0
Date:	Wed, 05 Feb 2025	Prob (F-statistic):	1.94e-309
Time:	19:08:47	Log-Likelihood:	-547.64
No. Observations:	1571	AIC:	1119.
Df Residuals:	1559	BIC:	1184.
Df Model:	11		
Covariance Type:	HC1		

	coef	std err	z	P>|z|	[0.025	0.975]
Intercept	5.7344	0.144	39.953	0.000	5.453	6.016
Treated	0.2143	0.031	7.014	0.000	0.154	0.274
Period	-0.0123	0.022	-0.551	0.582	-0.056	0.031
DiD	0.5416	0.036	15.246	0.000	0.472	0.611
Rating	0.1319	0.017	7.598	0.000	0.098	0.166
Stars	0.1829	0.014	13.356	0.000	0.156	0.210
Center_Distance	-0.0971	0.007	-14.518	0.000	-0.110	-0.084
Desayuno	-0.0466	0.022	-2.166	0.030	-0.089	-0.004
Air	-0.0171	0.025	-0.688	0.492	-0.066	0.032
Piscina	0.0358	0.030	1.207	0.227	-0.022	0.094
Cocina	0.1202	0.020	6.042	0.000	0.081	0.159
...						

After this second approach similar results than previous one are found. If we consider the added control variables, both *Air* and *Piscina* do not exhibit statistically significant p-values. *Desayuno* seems to have a negative effect in the price, but could be due cheap hotels announcing buffet breakfast as an important amenity and top-hotels not doing that. Lastly, *Cocina*, which has a statistically significant p-value, exhibits a coefficient of 0.12 on the logarithmic scale which indicates that the apartments or hotel rooms offering a kitchen increase the price of the stay.

Why Use Text Features?
On the gathered information we do not have a list or indicator of hotel services and facilities in a structured way.Thus, we need to analyze the descriptions, from where we can extract can extract key amenities that influence prices. To incorporate text-based information into our model, we identified the presence of specific keywords that indicate valuable hotel attributes and turned them into binary variables.
We selected only a few group of indicators but more could be used, even a sentiment analysis could be performed on the whole description text to cluster hotels in different groups. We selected the words by ourself because we have a previous knowledge of what good hotels can offer and we know that this words might not be present in every hotel description. That is why words like 'Barcelona','Milan', 'hotel' or 'room' would not have given any information as they are present in most of the hotels descriptions and thus do not help us to differentiate an hotel from another one. Also this common words give no information on the quality of the hotel and can induce us to some extra bias.

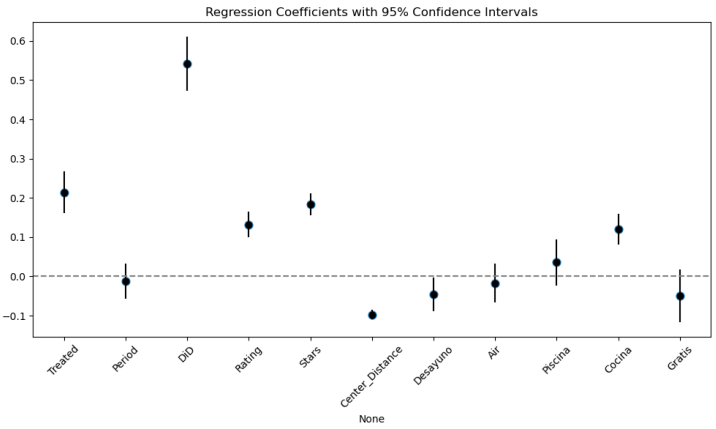

Figure 7: Coefficients and confidence intervals on this second approach.

3.3 Third approach

We are lastly asked to try to use controls to decompose the treatment effect by hotel quality. We need then to separate the dataframe we created into two groups, the hotels that are considered high quality and the ones that not. To account for this, we use text description and introduce a Luxury variable in our regression model. We use text analysis on hotel descriptions to detect words associated with luxury like: *spa*, *lujo* (luxury), *vistas* (views), *azotea* (rooftop)... and we created a dummy variable that identified the presence of that words in the description with a 1 and the absence with a 0.

Then we are able to separate the DiD dataframe in two single ones, one with the hotels considered high quality and one with ones that are not considered. We consider now the following regression model for each dataframe:

$$\text{price}_{itp} = \beta_0 + \beta_1(\text{treated}_t) + \beta_2(\text{period}_p) + \beta_3(\text{treated}_t \times \text{period}_p) + \beta_5(\text{stars}_i) + \epsilon_{itp}$$

In this way we are also considering the stars classification as a pre-known reliable indicator of a hotel quality.

We run a separate regression to each dataframe obtaining the following results:
For the luxury hotels:

	coef	std err	z	P>|z|	[0.025	0.975]
Intercept	6.4918	0.077	84.304	0.000	6.341	6.643
Treated	0.1253	0.045	2.759	0.006	0.036	0.214
Period	-0.0415	0.043	-0.965	0.335	-0.126	0.043
DiD	0.6301	0.058	10.837	0.000	0.516	0.744
Stars	0.2490	0.020	12.750	0.000	0.211	0.287

And for the non-luxury ones:

	coef	std err	z	P>|z|	[0.025	0.975]
Intercept	6.4743	0.064	100.819	0.000	6.348	6.600
Treated	0.3430	0.044	7.781	0.000	0.257	0.429
Period	0.0027	0.032	0.084	0.933	-0.060	0.065
DiD	0.4558	0.055	8.245	0.000	0.347	0.564
Stars	0.2109	0.018	11.695	0.000	0.176	0.246

From these results, we can extract some conclusions. First, the difference in percentage price between non-luxury hotels in Barcelona and Milan is bigger than the one existing between luxury hotels of both cities. In the other hand both luxury and non luxury hotels exhibit a similar dependence on star classification. But the most interesting statistically significant result is that DiD variable exhibits a bigger coefficient for luxury hotels than for non-luxury. This indicates that the price of luxury hotels receives a bigger impact by the presence of MWC than the non-luxury ones. We can attribute this to the fact that many congress assistants or visitors are wealthy people that demand luxury hotels instead of non-luxury ones, which increases the demand and in consequence the prices of the hotels that we perceive as luxurious.