EFFECTS OF FORMULA 1
ON BARCELONA'S RENTAL PRICES
(ACCORDING TO BOOKING DATA)

INTRODUCTION TO TEXT MINING AND NATURAL LANGUAGE PROCESSING

AUTHORS: VIKTORIA GAGUA, ALEJANDRO DELGADO TELLO, ALEX MALO

FEBRUARY 2025

BARCELONA SCHOOL OF ECONOMICS

Barcelona School of Economics

Event Selection

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

We chose the Formula One weekend in Barcelona, which starts on May 29th and ends on June 1st. This event presents an excellent opportunity to analyze price changes on Booking.com. The global appeal of Formula One attracts a diverse audience, from dedicated racing fans to high-spending international visitors, leading to a significant surge in accommodation demand in Barcelona. Due to limited availability during this period, guests are often willing to pay premium rates, making it an ideal moment for dynamic pricing strategies. Additionally, the event typically encourages longer stays due to pre- and post-race festivities, providing further opportunities to optimize revenue through extended bookings.
Despite the existence of bigger events such as the Mobile World Congress (MWC), which might yield better results, we decided against using it because the dates are too close to the current date, and most hotels are likely already fully booked. Therefore, our choice is the F1 event.

Time Periods and Control City Selection

Think of the time periods to scrape and what second city to scrape. The second city will be your control group. Explain your choices in writing.

We chose May 29 to June 2 to scrape Booking.com information in Barcelona, as this period best encapsulates the entire Formula One weekend experience. The potential check-out time is considered as June 2nd. On May 29, although the official race hasn't started, fans eagerly participate in the Promoter Pit Lane Walk—an exclusive behind-the-scenes experience. The race weekend officially starts on May 30 with Free Practices and Test Drives, drawing in a global audience and significantly boosting local accommodation demand. Since June 1 is dedicated to race day and post-race festivities, selecting June 2 as the check-out day ensures we capture data on extended stays, as many visitors linger to enjoy the full event experience.
For our control group, we chose Lisbon, a city where no extraordinary events are happening during the same period. Lisbon serves as an ideal control as it provides a baseline for typical booking trends without the influence of a major event, allowing us to isolate and compare the specific impact of the Formula One weekend in Barcelona.
In our Difference-in-Differences analysis, we examine the impact of the event during the treatment period from May 29 to June 2. The control group in our study is the city of Lisbon. Additionally, we scraped data for the previous weekend, from May 22 to May 26.
The purpose of collecting data from the previous weekend is as follows:

- **Pre-Event Trend Check:** Establishing a baseline ensures that the price trends in both Barcelona and Lisbon were similar before the event. This is crucial for the Diff-in-Diff approach, which assumes that, absent the event, both groups would have followed parallel trends.

- **Baseline Comparison:** Having a clear picture of pre-event price behavior allows us to confidently attribute significant changes during the treatment period to the event itself rather than random variations.

The weekend following the event falls in mid-June and is influenced by European summer vacations, leading to a general increase in accommodation prices. Choosing the previous weekend ensures our results are not biased by seasonal variations.
Overall, while Lisbon is our designated control group, scraping the previous weekend's data enhances our analysis by ensuring comparable pre-event conditions between the two groups.

Scraping Pipeline Design

Design a careful scraping pipeline that follows the advises seen in class and TAs.

To ensure a structured and efficient scraping process, we follow a carefully designed pipeline:

Set Up the Environment

We begin by configuring and launching a Firefox browser with specific settings (such as a custom download folder) to mimic real user browsing behavior on Booking.com.

Navigate and Prepare the Page

Once the browser is running, we load Booking.com, switch the language to English, and dismiss any cookie consent pop-ups to ensure a clean, user-friendly interface.

Input Search Criteria

We enter the cities (Barcelona or Lisbon) and specify our date ranges—from May 22 to May 26 and May 29 to June 2. These periods allow us to capture the full impact of the Formula One weekend and establish a control period.

Load All Hotel Listings

We simulate scrolling through the search results, automatically clicking the "Load More" button as needed to ensure all available hotel listings are fully loaded on the page.

Extract Hotel Information

With the complete set of listings displayed, we parse the page to extract key details for each hotel, including:

- Hotel name
- Price
- Rating
- Link to more information

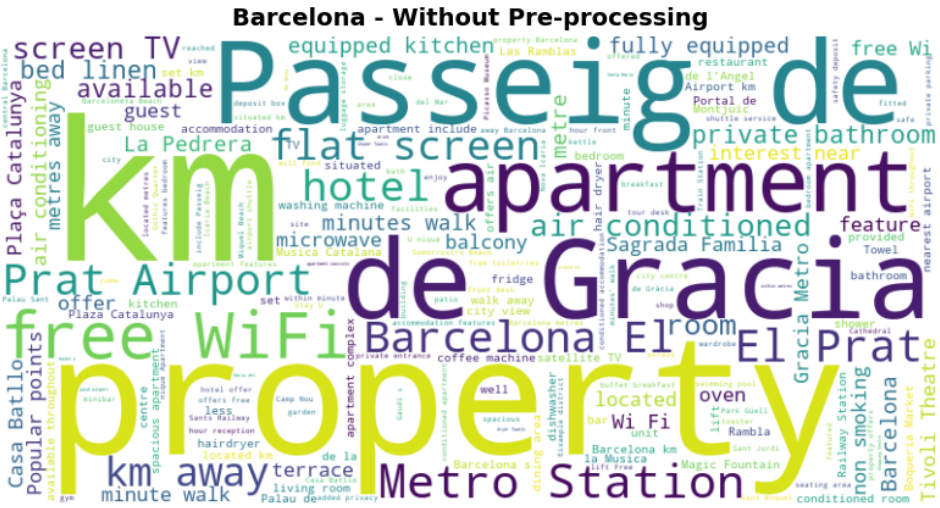
Scraping Hotel Descriptions

Finally, for each hotel, we visit its detailed page to scrape the hotel description. This step enriches our dataset by providing a fuller picture of each property and its amenities.

WordCloud Analysis

Create two wordclouds before and after pre-processing for each city (a total of four). Comment on the changes in the wordclouds.

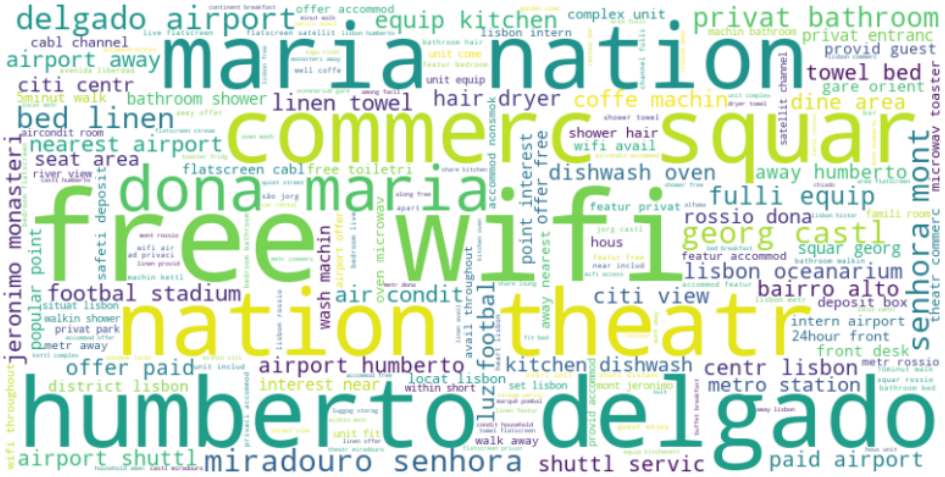
Before Pre-processing



After Pre-processing

Barcelona - Preprocessed

Lisbon - Preprocessed



Analysis of Changes

The WordCloud library in Python automatically removes stopwords by default. As a result, even in the non-processed word clouds, common words such as "and," "a," "is," "the," etc., do not appear. Before preprocessing, the word clouds contain frequently used terms that do not add significant value to our analysis, such as "stay," "property," "booking," "check," "guest," and "night"—a pattern observed in both Barcelona and Lisbon.

After preprocessing—where we removed stopwords (including custom terms like "booking"), applied stemming, and handled punctuation and special characters—the updated word clouds reveal more meaningful terms relevant to our analysis.

For Barcelona, we now see terms like "Passeig de Gràcia," one of the city's most expensive locations for hotels due to its prime location and limited supply. Additionally, keywords such as "metro station," "private bathroom," and "Prat airport" become prominent.

Similarly, for Lisbon, we observe "Comércio Square," an area comparable to Passeig de Gràcia in terms of exclusivity and high hotel prices. Other notable terms include "National Theater Dona Maria" and "Humberto Delgado Airport." Finally, in both cities, the most frequently mentioned feature across all hotels is "free WiFi."

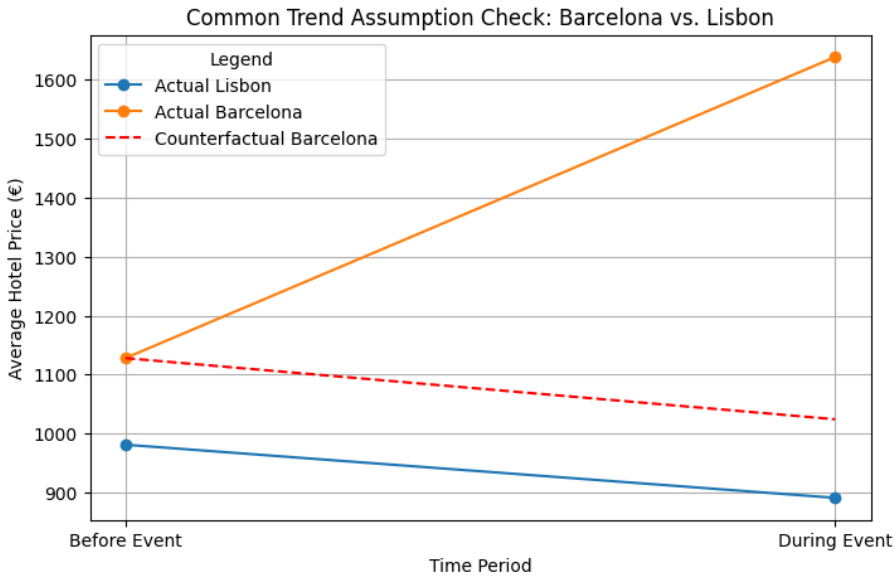
Fixed Effects Regression for Difference-in-Differences Estimate

Write down a fixed effects regression equation that allows you to derive a difference-in-difference estimate of the effect of the event on prices. Think of controls to add, why is this relevant? Explain why you need a second city for this.

Control variables are much needed in our analysis because they allow us to isolate the effect of the event on prices from other factors that could also be influencing those prices. For example, hotels differ in terms of the amenities they offer, such as free Wi-Fi, swimming pool, or gym, and these differences can affect their pricing independently of any external event. In addition, factors like the hotel's star rating, overall quality or even a specific season can influence prices. By including these variables as controls in our regression, we account for the observable differences across hotels that might otherwise confound the estimated impact of the event. So, using fixed effects and additional controls ensure that we account for both observed and unobserved confounding factors.

On the other hand, inclusion of Lisbon as a control group is crucial for the DiD approach. The second city provides a counterfactual scenario—representing what the treatment city's price trends might have looked like in the absence of the event. By comparing the changes in prices over time in Barcelona (which experienced the event) to Lisbon (which did not), we can better isolate the effect of the event. This helps to control for common trends and other external factors that might affect prices across both cities. Without a second city, it would be much more difficult to distinguish the event's impact from other time-varying influences.

Expanding on this, for Lisbon to serve this purpose as a valid control group, the key requirement is that, in the absence of treatment (Formula 1 in this case), the price difference between Barcelona and Lisbon would have remained constant over time. This ensures that the estimated effect of the event is causal rather than influenced by other factors. To validate this parallel trends assumption, we conducted Welch's test (included in the code attached to the zip folder) and found that Lisbon may not be an ideal control, as its prices declined instead of remaining stable. We can see this trend in the following graph:



Using Text Features from Descriptions as Controls

How would you use text features from the description as controls? Think about the text in the descriptions you scraped. How would this help? Why would terms like "Barcelona" not help?

In our project we extracted word clouds from the hotel descriptions and filtered the text using a specific list of amenity-related words. This allowed us to generate control variables that capture important features of each hotel's offering, such as mentions of "breakfast," "swimming pool," "gym," or "parking" and so on. These amenity keywords provide valuable information about the quality and services offered by the hotel, which can influence its pricing.

Incorporating these text-derived features as controls helps us account for differences in hotel quality and amenities that are not fully captured by standard numerical variables. This reduces omitted variable bias in our regression analysis, ensuring that our estimates of the event's impact on prices are more accurate.

On the other hand, terms like "Barcelona" would not be particularly useful as control variables because they simply indicate the hotel's location—a factor already controlled for in our fixed effects or other location-specific variables. By focusing on the amenity-related keywords instead, we capture unique, descriptive aspects of the hotels that more directly affect their pricing strategies.

Regression Models

Model 1 - Impact of Treatment Period

$$\text{price}_i = \beta_0 + \beta_1 \left(\text{Treatment\_period}_i\right) + \varepsilon_i$$

In this model, $\text{Treatment\_period}_i$ is a dummy variable equal to 1 if the observation falls on or after the specified treatment date (e.g., May 29, 2025), and 0 otherwise. The coefficient β_1 measures how much prices change, on average, after the treatment period begins, relative to the pre-treatment period. Results are as follows:

OLS Regression Results

```
==============================================================================
OLS Regression Results
==============================================================================
Dep. Variable:                  Price   R-squared:                       0.025
Model:                            OLS   Adj. R-squared:                  0.025
Method:                 Least Squares   F-statistic:                     101.9
Date:                Wed, 05 Feb 2025   Prob (F-statistic):           1.11e-23
Time:                        17:32:16   Log-Likelihood:                -31400.
No. Observations:                3971   AIC:                         6.280e+04
Df Residuals:                    3969   BIC:                         6.282e+04
Df Model:                           1
Covariance Type:                  HC1
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept      1055.4316     11.093     95.144      0.000    1033.690    1077.174
treatment_period   210.1788     20.819     10.096      0.000     169.375     250.982
==============================================================================
Omnibus:                     1936.491   Durbin-Watson:                   1.341
Prob(Omnibus):                  0.000   Jarque-Bera (JB):            17504.040
Skew:                           2.126   Prob(JB):                         0.00
Kurtosis:                      12.365   Cond. No.                         2.62
==============================================================================

Notes:
[1] Standard Errors are heteroscedasticity robust (HC1)
```

This simple regression shows that prices increased by about $210 during the treatment period compared to the pre-treatment baseline of $1055, a finding that's statistically significant but only explains 2.5% of price variation. The low R-squared and diagnostic issues like autocorrelation and non-normal residuals suggest this model is missing crucial variables. The huge difference between this simple model and the upcoming difference-in-differences specification (which will show city-specific effects and interactions) highlights **the importance of controlling for location-specific trends and heterogeneity**. Adding controls will improve the **model's explanatory power** and provide more reliable estimates of the treatment effect. The current specification's flaws are evident in the positive autocorrelation indicated by the Durbin-Watson statistic of 1.341 and the substantial non-normality shown by the high Jarque-Bera statistic.

Model 2 - Impact of Treatment City

The following equation represents the price model:

$$\text{price}_i = \beta_0 + \beta_1 \left(\text{Treatment\_city}_i\right) + \varepsilon_i$$

Where, Treatment_city_i is a dummy variable equal to 1 if the hotel is in the treatment city (e.g., Barcelona) and 0 otherwise (e.g., Lisbon). The coefficient β_1 indicates the average price difference for hotels in the treatment city compared to the control city, before any treatment period begins. results are as follows:

OLS Regression Results

Dep. Variable:	Price	R-squared:	0.113
Model:	OLS	Adj. R-squared:	0.113
Method:	Least Squares	F-statistic:	507.4
Date:	Wed, 05 Feb 2025	Prob (F-statistic):	7.80e-106
Time:	17:32:35	Log-Likelihood:	-31212.
No. Observations:	3971	AIC:	6.243e+04
Df Residuals:	3969	BIC:	6.244e+04
Df Model:	1		
Covariance Type:	HC1		

	coef	std err	z	P>|z|	[0.025	0.975]
Intercept	935.6778	12.367	75.660	0.000	911.439	959.916
treatment_city	447.6347	19.873	22.525	0.000	408.685	486.584

Omnibus:	2206.597	Durbin-Watson:	1.474
Prob(Omnibus):	0.000	Jarque-Bera (JB):	24199.878
Skew:	2.444	Prob(JB):	0.00
Kurtosis:	14.062	Cond. No.	2.63

Notes:
[1] Standard Errors are heteroscedasticity robust (HC1)

This regression focuses solely on the city-specific price differences, showing that hotels in the treatment city (Barcelona) command a substantial premium of about $448 over the control city's baseline price of $936, with this difference being highly statistically significant. The model explains 11.3% of price variation - better than the time-only model but still leaving substantial unexplained variance. Similar to the previous model, diagnostic tests reveal concerning patterns: positive autocorrelation (Durbin-Watson of 1.474), significant non-normality in residuals (Jarque-Bera statistic of 24199.878), and notable skewness and kurtosis, suggesting that city differences alone don't capture the full complexity of hotel pricing dynamics. While this specification confirms significant baseline price differences between cities, its limitations again show the need for a more comprehensive model that accounts for both temporal trends and city-specific effects, as well as their interaction, which will be captured more effectively in the difference-in-differences specification ahead.

Model 3 - Difference-in-Differences (DiD) Model

The following equation represents the extended price model:

$$\text{price}_i = \beta_0 + \beta_1 \left(\text{Treatment\_period}_i\right) + \beta_2 \left(\text{Treatment\_city}_i\right) + \beta_3 \left(\text{Treatment\_period}_i \times \text{Treatment\_city}_i\right) + \varepsilon_i$$

In this model, $\text{Treatment\_period}_i$ captures any overall price shift after the event date, Treatment_city_i measures the baseline difference in the treatment city relative to the control city, and the interaction term $\text{Treatment\_period}_i \times \text{Treatment\_city}_i$ is the difference-in-differences component that estimates how hotel prices in the treatment city change **relative** to the control city, over and above any common time trends or baseline city differences. Results are as follows:

OLS Regression Results

Dep. Variable:	Price	R-squared:	0.189
Model:	OLS	Adj. R-squared:	0.189
Method:	Least Squares	F-statistic:	209.9
Date:	Wed, 05 Feb 2025	Prob (F-statistic):	2.39e-126
Time:	17:32:41	Log-Likelihood:	-31033.
No. Observations:	3971	AIC:	6.207e+04
Df Residuals:	3967	BIC:	6.210e+04
Df Model:	3		
Covariance Type:	HC1		

	coef	std err	z	P>|z|	[0.025	0.975]
Intercept	981.2113	17.959	54.637	0.000	946.013	1016.410
treatment_period	-90.1069	24.671	-3.652	0.000	-138.461	-41.753
treatment_city	146.6576	22.031	6.657	0.000	103.477	189.838
interaction	600.4837	38.022	15.793	0.000	525.962	675.006

Omnibus:	2135.328	Durbin-Watson:	1.613
Prob(Omnibus):	0.000	Jarque-Bera (JB):	23650.440
Skew:	2.334	Prob(JB):	0.00
Kurtosis:	14.007	Cond. No.	6.89

Notes:
[1] Standard Errors are heteroscedasticity robust (HC1)

In the model Baseline price (Intercept) is \$981.21 as depicted. The treatment_period coefficient (-90.11) means that during the treatment period, on average, prices in BOTH cities dropped by about \$90. This could be due to seasonal effects, economic conditions, or other factors affecting both cities.

However, the **large positive interaction term** (600.48) means that the treatment city responded very differently to the treatment than the control city: Control city during treatment only experienced the treatment_period effect (-90.11), while Treatment city during treatment experienced **both the treatment_period effect AND the interaction effect** (-90.11 + 600.48 = +510.37 net increase). So while there was an overall downward trend in prices during the treatment period, the treatment city actually saw a substantial price increase relative to this baseline trend. This could happen due to our proposed event: F1 grand prix. This pattern is exactly what makes difference-in-differences analysis **powerful** in our case - it detects treatment effects even when they run counter to the prevailing trends in the data.

Model 4 - Fixed Effects Regression Model

The following equation represents the difference-in-differences model:

$$y_{it} = \alpha_i + \lambda_t + \beta\,(\text{Treatment}_i \times \text{Post}_t) + \varepsilon_{it}$$

Where:

- y_{it} is the price for hotel i at time t.
- α_i is the hotel (entity) fixed effect capturing all time-invariant characteristics of hotel i.
- λ_t is the time fixed effect capturing time-specific factors that affect all hotels.
- Treatment_i is a dummy variable equal to 1 if hotel i is in Barcelona (the treatment group), and 0 otherwise.
- Post_t is a dummy variable equal to 1 for dates on/after May 29, 2025 (post-event period), and 0 for dates before.
- $(\text{Treatment}_i \times \text{Post}_t)$ is the interaction term (the difference-in-differences term) whose coefficient β measures the treatment effect.
- ε_{it} is the error term.

Results are as follows:

PanelOLS Estimation Summary

Dep. Variable:	price	R-squared:	0.3090
Estimator:	PanelOLS	R-squared (Between):	0.1401
No. Observations:	3971	R-squared (Within):	0.3826
Date:	Wed, Feb 05 2025	R-squared (Overall):	0.1377
Time:	17:32:58	Log-likelihood	-2.396e+04
Cov. Estimator:	Robust		
		F-statistic:	593.92
Entities:	2641	P-value	0.0000
Avg Obs:	1.5036	Distribution:	F(1,1328)
Min Obs:	1.0000		
Max Obs:	2.0000	F-statistic (robust):	375.01
		P-value	0.0000
Time periods:	2	Distribution:	F(1,1328)
Avg Obs:	1985.5		
Min Obs:	1974.0		
Max Obs:	1997.0		

Parameter Estimates

	Parameter	Std. Err.	T-stat	P-value	Lower CI	Upper CI
did	332.51	17.170	19.365	0.0000	298.83	366.19

F-test for Poolability: 17.611
P-value: 0.0000

```
Distribution: F(2641,1328)

Included effects: Entity, Time
```

The coefficient on the did term is about 332.51. This implies that, once the event starts, hotels in the treatment group (Barcelona) charge, on average, around €332 more relative to what we observe in the control group (Lisbon), above and beyond any baseline price differences or common time trends.

The 95% confidence interval ranges from about €299 to €366, suggesting we are very confident that the true treatment effect is within that range.

The p-value is 0.0000 (essentially less than 0.0001), indicating that the effect is statistically significant at any conventional level (e.g., 1%, 5%, or 10%). In other words, **there is strong evidence that the event is associated with a real shift in prices**, rather than being due to chance.

The R-squared of around 0.309 (or about 31%) means the model explains roughly one-third of the total variation in hotel prices across time and entities. In panel data settings, this is often considered quite reasonable, given that many unobservable factors can influence prices.

Taken together, these results strongly suggest that **Formula 1 has a large and significant impact on Barcelona hotel prices**. Compared to Lisbon (the control group), Barcelona hotels will see their post-event prices increase substantially—by about €332 more than any price changes that would have happened absent the event.

Model 5 - Enhanced DiD model with additional controls for amenities

The following equation represents the model:

$$\text{price}_i = \beta_0 + \beta_1\,(\text{treatment}_i) + \beta_2\,(\text{post}_i) + \beta_3\,(\text{did}_i) + \beta_4\,(\text{amenities\_count}_i) + \varepsilon_i$$

Where:

- price_i is the dependent variable (the price of hotel i).
- treatment_i is a dummy variable (1 if the hotel is in Barcelona, 0 otherwise).
- post_i is a dummy variable (1 if the date is on/after May 29, 2025, 0 otherwise).
- did_i is the interaction term ($\text{treatment}_i \times \text{post}_i$).
- amenities_count_i is the total number of listed amenities for hotel i.
- $\beta_0, \beta_1, \ldots, \beta_4$ are the coefficients to be estimated.
- ε_i is the error term.

Here we include the count of amenities to control for potential price differences arising from varying numbers of amenities offered by each hotel. The regression tells us the estimated effect of the event (represented by did) on hotel prices, while holding constant the differences in the number of amenities. The amenities coefficient will show how each **additional** amenity correlates with hotel price, independently of the event effect.

OLS Regression Results

Dep. Variable:	price	R-squared:	0.221
Model:	OLS	Adj. R-squared:	0.221
Method:	Least Squares	F-statistic:	169.1
Date:	Wed, 05 Feb 2025	Prob (F-statistic):	1.72e-163
Time:	17:33:16	Log-Likelihood:	-30953.
No. Observations:	3971	AIC:	6.192e+04
Df Residuals:	3965	BIC:	6.196e+04
Df Model:	5		
Covariance Type:	HC1		

	coef	std err	z	P>|z|	[0.025	0.975]
Intercept	259.8539	131.315	1.979	0.048	2.482	517.226
treatment	69.1006	22.351	3.092	0.002	25.293	112.908
post	-79.8225	23.462	-3.402	0.001	-125.807	-33.838
did	618.3999	36.859	16.778	0.000	546.158	690.642
amenities_count	81.2443	7.509	10.820	0.000	66.528	95.961
rating	60.7542	16.358	3.714	0.000	28.693	92.815

Omnibus:	2254.224	Durbin-Watson:	1.531
Prob(Omnibus):	0.000	Jarque-Bera (JB):	27517.001
Skew:	2.472	Prob(JB):	0.00
Kurtosis:	14.910	Cond. No.	87.2

```
Notes:
[1] Standard Errors are heteroscedasticity robust (HC1)
```

The "treatment" coefficient of approximately €69.1 indicates that in the pre-event period, Barcelona hotels are on average €69 more expensive than Lisbon hotels (holding the other variables constant). The large "did" coefficient of about €618.4 means that, on top of Barcelona's baseline price difference, the post-event period in Barcelona is associated with an additional average increase of nearly €618, net of the broader drop signaled by the "post" coefficient. The "amenities_count" coefficient of about €81.24 indicates that each additional amenity is associated with an €81 increase in a hotel's price, controlling for everything else in the model. Overall, the R-squared of 0.221 suggests that the model explains about 22.1% of the variation in hotel prices, with the remaining variation due to other unobserved factors.

Model 6 - Heterogeneous Treatment Effects

This regression includes a hotel's rating and an interaction term between the DiD variable and `rating`. By including `did:rating`, we allow the event's effect on price to vary ("heterogeneous effects") depending on each hotel's rating. Higher-rated hotels, for instance, might see a different price impact from the event than lower-rated hotels.

$$\text{price}_i = \beta_0 + \beta_1\,(\text{treatment}_i) + \beta_2\,(\text{post}_i) + \beta_3\,(\text{did}_i) + \beta_4\,(\text{rating}_i) + \beta_5\,(\text{did}_i \times \text{rating}_i) + \beta_6\,(\text{amenities\_count}_i) + \varepsilon_i$$

Where:

- $\text{treatment}_i = 1$ if hotel i is in the treatment city (0 otherwise).
- $\text{post}_i = 1$ if date is on/after the event start (0 otherwise).
- $\text{did}_i = \text{treatment}_i \times \text{post}_i$.
- rating_i = the hotel's rating.
- $\text{did}_i \times \text{rating}_i$ = the heterogeneous effect of the event, depending on rating.
- amenities_count_i = total number of amenities offered by hotel i.

Results are as follows:

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                  price   R-squared:                       0.264
Model:                            OLS   Adj. R-squared:                  0.262
Method:                 Least Squares   F-statistic:                     207.7
Date:                Wed, 05 Feb 2025   Prob (F-statistic):          5.21e-231
Time:                        17:33:47   Log-Likelihood:                -30843.
No. Observations:                3971   AIC:                         6.170e+04
Df Residuals:                    3964   BIC:                         6.174e+04
Df Model:                           6
Covariance Type:                  HC1
================================================================================
                     coef    std err          z      P>|z|      [0.025      0.975]
--------------------------------------------------------------------------------
Intercept      -1110.1514    110.812    -10.018      0.000   -1327.338    -892.965
treatment        107.2886     21.281      5.042      0.000      65.580     148.998
post             -79.1382     22.821     -3.468      0.001    -123.866     -34.411
did             3142.4778    199.811     15.727      0.000    2750.855    3534.100
rating           229.9428     13.862     16.588      0.000     202.774     257.112
did:rating      -312.4008     24.605    -12.697      0.000    -360.626    -264.176
amenities_count   62.9459      7.057      8.920      0.000      49.115      76.776
==============================================================================
Omnibus:                     2341.394   Durbin-Watson:                   1.632
Prob(Omnibus):                  0.000   Jarque-Bera (JB):            34119.787
Skew:                           2.534   Prob(JB):                         0.00
Kurtosis:                      16.436   Cond. No.                         202.
==============================================================================

Notes:
[1] Standard Errors are heteroscedasticity robust (HC1)
```

This is a very interesting regression model, because these results suggest that once the event ("did") kicks in for the treatment city, **lower-rated hotels experience a large jump in price, while higher-rated hotels in the treatment city see a more modest price increase**. In particular, the large positive coefficient on did (about +€3142) indicates a strong event effect at a hypothetical rating of zero, but the negative interaction (did:rating approximately = –€312) means that each extra rating point reduces that large event-driven price boost. Hence, higher-rated hotels in Barcelona after the event still see a price increase but not as big as lower-rated hotels. This observed effect may be attributed to the composition of the F1 fanbase, which often includes a substantial segment of budget-conscious travelers primarily focused on attending the race rather than seeking premium accommodations. Consequently, lower-rated hotels experience a pronounced increase in demand—end thus a larger price uptick — during the event. In contrast, higher-rated (and typically more expensive) hotels may already benefit from steady, year-round demand, leaving less room for an event-driven price surge. As a result, while these hotels still see a positive impact, **the magnitude of the increase is more modest compared to lower-rated alternatives.**

Separately, the positive coefficient on rating (+€229) indicates that—outside of the event effect—each additional rating point (e.g., going from 7 to 8) is associated with a higher price. Likewise, the large positive effect of amenities_count (+€62) confirms that hotels offering more amenities tend to charge more. The treatment coefficient (+€107) shows that, even before the event, Barcelona hotels were already more expensive on average.

Conclusion

This study analyzed the impact of the Formula 1 event in Barcelona on hotel prices using a Difference-in-Differences (DiD) framework, leveraging Lisbon as a control city. The results indicate a substantial price increase in Barcelona's hotels during the event period, with lower-rated hotels experiencing a more pronounced price surge compared to higher-rated ones, as revealed by the interaction term in our regression models.

Additionally, amenities played a significant role in price determination, as demonstrated by the positive correlation between the number of amenities and hotel prices. Our text-based features extracted from hotel descriptions helped control for variations in hotel offerings, improving the accuracy of our estimations.

Limitations and Potential Enhancements

Despite the robustness of our findings, certain limitations must be acknowledged:

- **Data Constraints:** The analysis is based on data from a single comparison date, which may limit generalizability. Many other latent external factors, such as local economic conditions or coinciding tourism trends, could also influence pricing. Analyzing multiple events across different years would provide a more comprehensive understanding of event-driven price fluctuations.
- **Control City Assumption:** While Lisbon was chosen as a control city due to the absence of a major event, differences in general tourism trends, economic conditions, or seasonality effects between the two cities could still affect results. Using multiple control cities instead of just Lisbon could enhance the robustness of the findings.
- **Modeling Limitations:** The regression models assume parallel trends between Barcelona and Lisbon before the event, a necessary condition for the validity of the DiD approach. While we checked for pre-treatment trends, potential unobserved factors might still bias results.

Besides the possible changes mentioned above, some potential enhancements for future research could improve results by:

- **Exploring Alternative Control Groups:** Refining the **Text Analysis:** Incorporating advanced NLP techniques such as sentiment analysis or topic modeling could further improve control variable selection from hotel descriptions.
- **Testing Different Model Specifications:** More complex econometric models, including machine learning approaches, could be used to predict price changes with higher precision.