Text Mining: Homework 1

Team Members:

Soledad Monge, Long Cheng, Iñigo Exposito

February 5, 2025

Part I: Scraping

1. The Event

We selected the Formula 1 Grand Prix in Barcelona (May 30 to June 1, 2025), as the key event for our analysis because it represents a major international sporting event with a significant impact on local tourism and accommodation demand. The Formula 1 Spanish Grand Prix, held at the Circuit de Barcelona-Catalunya, influx of both domestic and international visitors, including fans, teams, media personnel, and event staff, all of whom require accommodation in the city. Given the increased demand, it is reasonable to expect a notable rise in hotel and short-term rental prices in Barcelona during the race weekend compared to a regular weekend.

2. Time Periods and Control Group

By comparing accommodation prices in Barcelona during the Grand Prix weekend (May 30 – June 1, 2025) with a normal weekend (May 16 – May 18, 2025), and using Seville as a control city, we aim to quantify the price surge attributed to the event. We selected Seville as the control group because it shares several key characteristics with Barcelona, making it a suitable comparison for a Difference-in-Differences (DiD) estimation. Both cities:

- Are Major Tourist Destinations in Spain: They attract a high volume of domestic and international visitors year-round.
- Have a similar hospitality industry: Both cities have a diverse range of accommodations, including hotels, hostels, and vacation rentals, which are affected by tourism demand.

Unlike Barcelona, Seville does not host the Formula 1 Grand Prix, which means that its accommodation prices should remain stable in the absence of external shocks. In a DiD framework, Seville serves as a counterfactual, helping us isolate the effect of the Formula 1 even in Barcelona. If accommodation prices increase disproportionately in Barcelona compared to Seville during the event weekend, we can attribute this difference to the impact of the Formula 1 race rather than broader economic or seasonal factors affecting hotel prices in both cities.

3. Scraping Methodology

Starting with data organization. The scraped data (in English) are stored in a structured pandas *DataFrame* with columns for *hotel name*, *id*, *rating*, *price*, *link*, *description*, *check_in*, and *check_out*. This ensures consistency and facilitates downstream analysis. Data are saved to an Excel file (for each city-time pair such as *barcelona_16_18.xlsx*) for easy access and reproducibility. After controlling the same hotels for two time periods, we got 545 hotels for Seville and 701 hotels for Barcelona.

Next, automation. Key tasks (e.g., browser setup, date navigation, scrolling, and data extraction) are encapsulated in reusable functions like *start_up*(), *navigate_to_month*(), *scroll_and_load_all*() and *get_hotel_description*(). These functions handle browser setup, interactions, and data extraction automatically. Specifically, the *scroll_and_load_all*() function scrolls down to the bottom of the webpage and automatically clicks 'Load more results' to load the information for all hotels. The loop for iterating over all links in hotel containers extracts descriptions for all hotels.

Only packages covered in the course are used: *selenium* for browser automation; *pandas* for data structuring; *openpyxl* for Excel export. The browser used to scrape the data was Firefox as recommended in class.

Part II: Test Analysis

(a) Preprocess

We began by cleaning and standardizing the text data for each hotel description in both Barcelona and Seville. Specifically, we converted the text to lowercase, removed punctuation and numeric characters, and tokenized the text into individual words. We then filtered out common English stopwords as well as domain-specific terms (like ”hotel” and ”offers”) that do not add meaningful context, before applying a Porter Stemmer to consolidate words into their root forms. These steps help ensure the data is free from noise and that the most relevant terms receive more emphasis.

After preprocessing, we used the cleaned text to generate WordCloud visualizations for each city. Comparing the “before” and “after” WordClouds reveals that frequently repeated and less informative words (like “hotel” or “located”) are largely eliminated. As a result, the terms most indicative of the hotels' features—such as location-specific attributes or amenities—become more pronounced, enabling clearer insights into what each city's hotel offerings emphasize.

(b) WordCloud Analysis

Before Preprocessing

At first glance, the wordclouds for both Seville and Barcelona show many references to each city's most prominent landmarks and amenities. In the before preprocessing images, these raw texts are teeming with generic hospitality terms—such as "property," "hotel," and "room"—alongside city-specific phrases like "Seville Airport" or "El Prat Airport." They also reflect capitalization inconsistencies and contain multi-word phrases with punctuation intact. For Seville, especially, famous sites like the Giralda, the Alcázar, or the Triana Bridge stand out; whereas in Barcelona, "Passeig de Gracia," "La Pedrera," and "Sagrada Familia" dominate the landscape. Taken as a whole, these unprocessed wordclouds combine both the local attractions (showcasing each city's tourist draw) and the everyday descriptors typical of hotel listings.

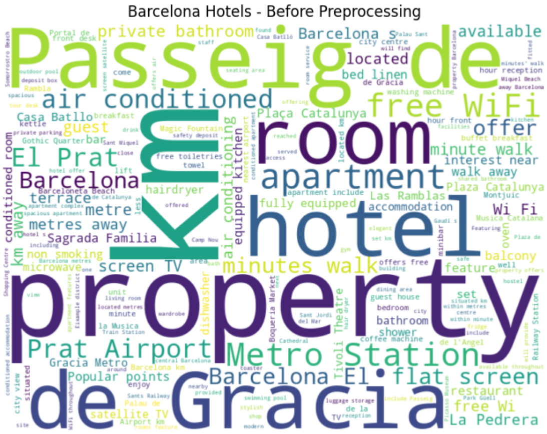

Barcelona Hotels - Before Preprocessing

Figure 1: Barcelona Hotels-Before Preprocessing

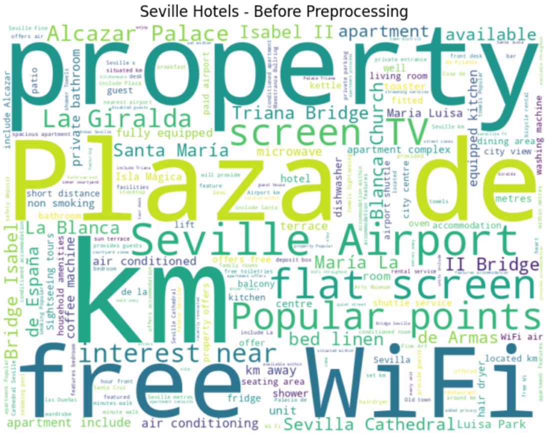

Seville Hotels - Before Preprocessing

Figure 2: Seville Hotels-Before Preprocessing

After Preprocessing

Once the text is processed—through lowercasing, removing stopwords and filler terms, and possibly applying stemming or lemmatization—the resulting wordclouds become more streamlined. Words such as "property," "km," "room," and "hotel" become less prominent or transform into normalized tokens. In Seville's processed cloud, we can still see the distinctive references to its heritage ("girald," "alcazar," "plaza," and "triana bridg," for example), but they appear in simpler forms. Barcelona's processed cloud retains mentions of "passeig de gracia," "el prat," "sagrada familia," and "la pedrera," clearly foregrounding the top local attractions and amenities like "free wifi" and "metro station." By trimming away repetitive filler and standard hospitality vocabulary, these processed clouds more directly spotlight the features that truly characterize each city's lodging options.

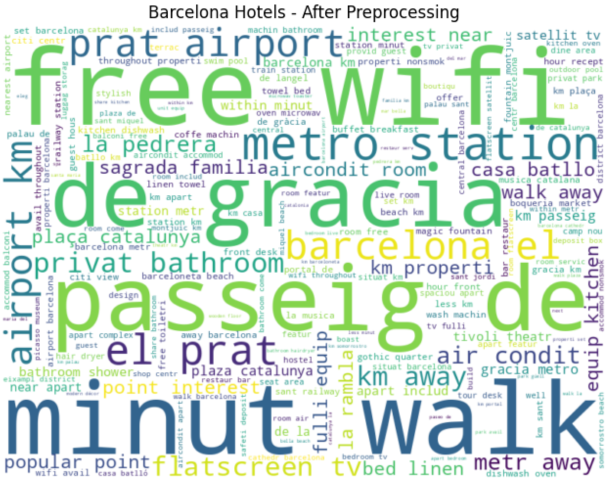

Barcelona Hotels - After Preprocessing

Figure 3: Barcelona Hotels-After Preprocessino

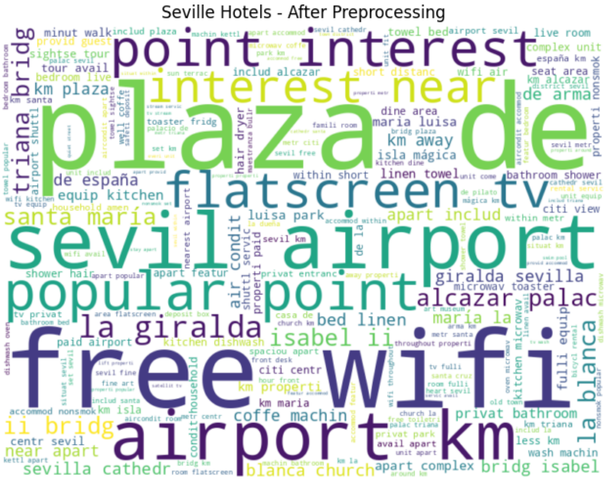

Seville Hotels - After Preprocessing

Figure 4: Seville Hotels-After Preprocessino

Comparison

Comparing the "before" and "after" wordclouds side by side reveals how preprocessing can help unearth deeper insights about a dataset. The raw wordclouds capture the full, unfiltered snapshot of hotel descriptions—complete with capitalized terms, repeated stock phrases, and a mixture of essential and extraneous words. After preprocessing, however, the text is more uniform and meaningful, allowing city-specific highlights—landmarks, neighborhoods, and services—to emerge more clearly. Overall, this transformation demonstrates why text normalization is so valuable in uncovering which topics or keywords matter most, especially when comparing descriptions from hotels in different destinations.

Part III: DiD

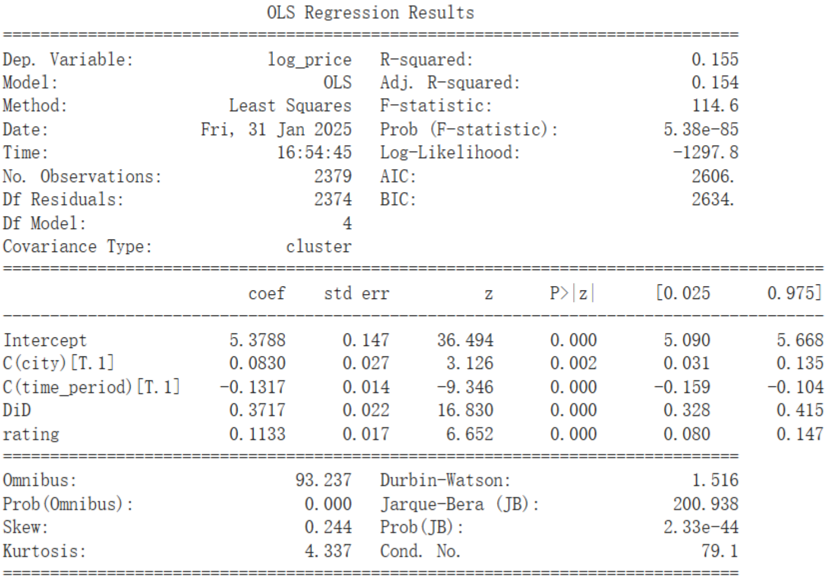
(a) A Fixed Effects Regression Equation

In order to identify the effect of our chosen event (Formula 1 Grand Prix in Barcelona) on hotel prices, we specify a difference-in-differences (DiD) model that includes fixed effects for cities and time periods. We use a generic version of the equation with city and time fixed effects as:

$$\text{Ln}(\text{Price})_{i,t} \;=\; \alpha \;+\; \gamma\, C(\text{city}_i) \;+\; \delta\, C(\text{time}_t) \;+\; \beta\,\big(C(\text{city}_i) \times C(\text{time}_t)\big) \;+\; \mathbf{X}_{i,t}\boldsymbol{\theta} \;+\; \varepsilon_{i,t},$$

where $\text{Ln}(\text{Price})_{i,t}$ is the natural log of price of hotel i at time t, $C(\text{city}_i)$ is a dummy variable indicating which city hotel i is located in (treated vs. control), $C(\text{time}_t)$ is a dummy variable for the period (pre-event vs. post-event), and the interaction term $C(\text{city}_i) \times C(\text{time}_t)$ captures the DiD effect, denoted by β. Additionally, $\mathbf{X}_{i,t}$ are other controls, such as hotel rating or text-based amenities(in part (b)).

We need a second city because DiD relies on comparing changes in price over time between two groups: one that experiences the event (treated city) and one that does not (control city). This “control” group is essential for constructing a counterfactual trend, representing what would have happened in the absence of the event. By subtracting out the general time trend (in Seville) and the general city difference in the pre-period, we can isolate the *causal effect* of the event on prices, contingent on the assumption of parallel trends.

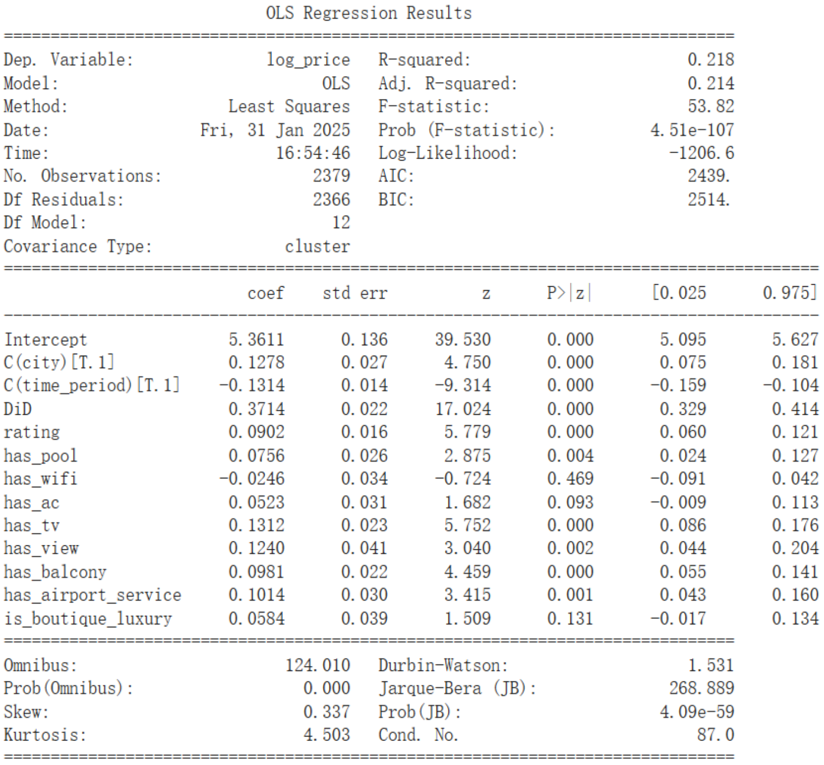


```
                            OLS Regression Results
==============================================================================
Dep. Variable:              log_price   R-squared:                       0.155
Model:                            OLS   Adj. R-squared:                  0.154
Method:                 Least Squares   F-statistic:                     114.6
Date:                Fri, 31 Jan 2025   Prob (F-statistic):           5.38e-85
Time:                        16:54:45   Log-Likelihood:                -1297.8
No. Observations:                2379   AIC:                             2606.
Df Residuals:                    2374   BIC:                             2634.
Df Model:                           4
Covariance Type:              cluster
=======================================================================================
                          coef    std err          z      P>|z|      [0.025      0.975]
---------------------------------------------------------------------------------------
Intercept               5.3788      0.147     36.494      0.000       5.090       5.668
C(city)[T.1]            0.0830      0.027      3.126      0.002       0.031       0.135
C(time_period)[T.1]    -0.1317      0.014     -9.346      0.000      -0.159      -0.104
DiD                     0.3717      0.022     16.830      0.000       0.328       0.415
rating                  0.1133      0.017      6.652      0.000       0.080       0.147
==============================================================================
Omnibus:                       93.237   Durbin-Watson:                   1.516
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              200.938
Skew:                           0.244   Prob(JB):                     2.33e-44
Kurtosis:                       4.337   Cond. No.                         79.1
==============================================================================
```

In the first specification, we regress the natural log of price (ln(price)) on city fixed effects, time fixed effects, and an interaction term (DiD) that captures the difference-in-differences effect of the event, plus a rating control. The city dummy itself (about 0.08) suggests an 8% higher baseline price in the treated city before the event, and the time dummy (about -0.13) indicates that, on average, prices of hotels in Seville drop by around 13% between pre- and post-periods. The positive and significant DiD coefficient (roughly 0.37) implies that hotels in the treated city exhibit an estimated 37% higher price (in log terms) post-event, relative to what would have occurred without the event (as inferred from the control city). Finally, the prices of hotels in Barcelona increase by 24% (37%-13%) in the post-periods. Moreover, rating has a positive coefficient of about 0.11, implying that each one-point increase in the rating is associated with roughly an 11% higher price.

(b) Using Text Features as Controls

In hotel listings, text descriptions often contain important information about amenities and characteristics that might otherwise go unobserved. By extracting features—such as whether a hotel has a pool, free WiFi, or whether the description mentions “luxury” or “boutique”—we can control for these unobserved qualities in our DiD regression. Including these variables helps ensure that any difference in prices after the event is not driven by compositional differences in the type of hotels across the treated and control cities or over time.



```
                            OLS Regression Results
==============================================================================
Dep. Variable:              log_price   R-squared:                       0.218
Model:                            OLS   Adj. R-squared:                  0.214
Method:                 Least Squares   F-statistic:                     53.82
Date:                Fri, 31 Jan 2025   Prob (F-statistic):          4.51e-107
Time:                        16:54:46   Log-Likelihood:                -1206.6
No. Observations:                2379   AIC:                             2439.
Df Residuals:                    2366   BIC:                             2514.
Df Model:                          12
Covariance Type:              cluster
=======================================================================================
                          coef    std err          z      P>|z|      [0.025      0.975]
---------------------------------------------------------------------------------------
Intercept               5.3611      0.136     39.530      0.000       5.095       5.627
C(city)[T.1]            0.1278      0.027      4.750      0.000       0.075       0.181
C(time_period)[T.1]    -0.1314      0.014     -9.314      0.000      -0.159      -0.104
DiD                     0.3714      0.022     17.024      0.000       0.329       0.414
rating                  0.0902      0.016      5.779      0.000       0.060       0.121
has_pool                0.0756      0.026      2.875      0.004       0.024       0.127
has_wifi               -0.0246      0.034     -0.724      0.469      -0.091       0.042
has_ac                  0.0523      0.031      1.682      0.093      -0.009       0.113
has_tv                  0.1312      0.023      5.752      0.000       0.086       0.176
has_view                0.1240      0.041      3.040      0.002       0.044       0.204
has_balcony             0.0981      0.022      4.459      0.000       0.055       0.141
has_airport_service     0.1014      0.030      3.415      0.001       0.043       0.160
is_boutique_luxury      0.0584      0.039      1.509      0.131      -0.017       0.134
==============================================================================
Omnibus:                      124.010   Durbin-Watson:                   1.531
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              268.889
Skew:                           0.337   Prob(JB):                     4.09e-59
Kurtosis:                       4.503   Cond. No.                         87.0
==============================================================================
```

In the second specification, we incorporate additional controls derived from text descriptions (pool, WiFi, air conditioning, etc.). These features improve the model’s explanatory power (the adjusted R^2 rises from about 0.15 to about 0.21), suggesting that amenities mentioned in the text help explain cross-hotel price differences. The DiD coefficient remains strongly positive (approximately 0.37), indicating that even after accounting for these hotel-specific features, the event leads to a roughly 37% increase in prices in the treated city, above and beyond the control city’s trend. Certain amenities, such as having a balcony or an airport shuttle, are associated with noticeable price premiums, while some features (e.g. free WiFi) are not statistically significant in this sample.

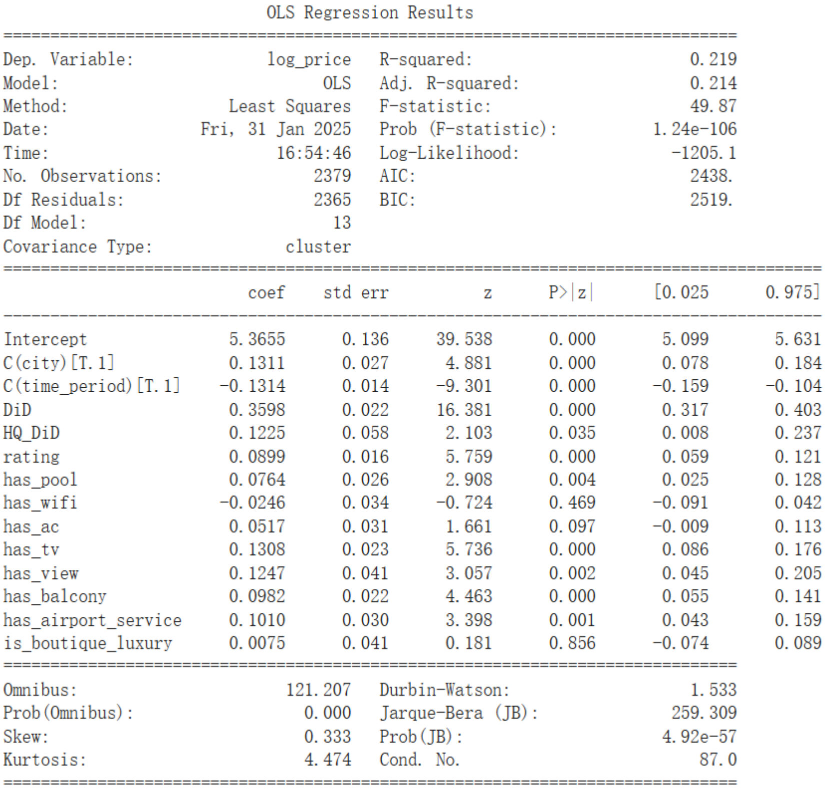
(c) Decomposing the Treatment Effect by Hotel Quality

To investigate heterogeneous treatment effects by hotel quality, one can extend the DiD model by interacting a "quality" measure with the DiD term. Suppose we define HighQuality_i as an indicator derived from text descriptions, capturing whether a hotel is "boutique," "luxury," or mentions premium amenities. We then estimate an augmented model:

$$\text{Ln}(\text{Price})_{i,t} = \alpha + \gamma\, C(\text{city}_i) + \delta\, C(\text{time}_t) + \beta\left(C(\text{city}_i) \times C(\text{time}_t)\right)$$

$$+\phi\left(C(\text{city}_i) \times C(\text{time}_t) \times \text{HighQuality}_i\right) + \mathbf{X}_{i,t}\boldsymbol{\theta} + \varepsilon_{i,t}.$$

Here, ϕ measures how the DiD effect differs for higher-quality hotels. A positive and significant ϕ would imply that high-quality hotels respond more strongly to the event in terms of raising prices. This is often of interest when events (e.g., major conferences, festivals) cause large spikes in demand that upscale hotels can exploit more readily.



```
                            OLS Regression Results
==============================================================================
Dep. Variable:              log_price   R-squared:                       0.219
Model:                            OLS   Adj. R-squared:                  0.214
Method:                 Least Squares   F-statistic:                     49.87
Date:                Fri, 31 Jan 2025   Prob (F-statistic):          1.24e-106
Time:                        16:54:46   Log-Likelihood:                -1205.1
No. Observations:                2379   AIC:                             2438.
Df Residuals:                    2365   BIC:                             2519.
Df Model:                          13
Covariance Type:              cluster
=======================================================================================
                          coef    std err          z      P>|z|      [0.025      0.975]
---------------------------------------------------------------------------------------
Intercept               5.3655      0.136     39.538      0.000       5.099       5.631
C(city)[T.1]            0.1311      0.027      4.881      0.000       0.078       0.184
C(time_period)[T.1]    -0.1314      0.014     -9.301      0.000      -0.159      -0.104
DiD                     0.3598      0.022     16.381      0.000       0.317       0.403
HQ_DiD                  0.1225      0.058      2.103      0.035       0.008       0.237
rating                  0.0899      0.016      5.759      0.000       0.059       0.121
has_pool                0.0764      0.026      2.908      0.004       0.025       0.128
has_wifi               -0.0246      0.034     -0.724      0.469      -0.091       0.042
has_ac                  0.0517      0.031      1.661      0.097      -0.009       0.113
has_tv                  0.1308      0.023      5.736      0.000       0.086       0.176
has_view                0.1247      0.041      3.057      0.002       0.045       0.205
has_balcony             0.0982      0.022      4.463      0.000       0.055       0.141
has_airport_service     0.1010      0.030      3.398      0.001       0.043       0.159
is_boutique_luxury      0.0075      0.041      0.181      0.856      -0.074       0.089
==============================================================================
Omnibus:                      121.207   Durbin-Watson:                   1.533
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              259.309
Skew:                           0.333   Prob(JB):                     4.92e-57
Kurtosis:                       4.474   Cond. No.                         87.0
==============================================================================
```

We test for heterogeneous treatment effects by interacting a "high-quality" indicator (captured by `is_boutique_luxury`) with the DiD term (labeled HQ_DiD). In this specification, the main DiD coefficient of about 0.36 still indicates a sizable 36% price increase for *non-luxury* hotels in the treated city. However, the positive HQ_DiD coefficient of around 0.12 suggests that boutique/luxury hotels see an *additional* 12% price boost, making their total event-driven increase closer to 48%. Thus, higher-end properties appear to capitalize more on the event's heightened demand, underscoring how quality segments can respond differently to the same external shock.

A potential drawback of deriving such a "quality" indicator from text is that the listing descriptions might be incomplete or inconsistently written. Some hotels may not mention their star rating or particular amenities, even though they exist, while others might overemphasize certain qualities as a marketing strategy. Consequently, the quality measure could be subject to measurement error if the text does not accurately reflect the true attributes of each hotel.

Moreover, textual descriptions can differ in length and level of detail for reasons unrelated to actual hotel quality. Smaller or budget-oriented hotels may supply very brief text, leaving out important information, whereas luxury hotels often highlight numerous features, which might inflate their text-derived "quality" scores. These discrepancies can introduce bias or noise in the quality indicator, making it essential to cross-validate text-based measures with more standardized data sources, such as official star ratings or user-generated reviews.