The Influence of Review Volume and Sentiment on Hotel Ratings: Evidence from Booking.com

Abstract

This study investigates the relationship between cumulative customer reviews or the number of reviews, sentiment, and perceived hotel quality using a large-scale dataset from Booking.com. We just picked hotels in the UK for our research. We try to examine the impact of cumulative review volume, lagged positive and negative sentiment, and reviewer behavior on the average score of hotels across time by combining pooled OLS and fixed effects panel regressions. In the last parts, we investigate the heterogeneous effect for the main types of travellers and also for their goals of the trip (Business vs Leisure). To validate the robustness of our findings and address concerns over potential fake reviews, we conduct graphical time-series inspections and exclude hotels with extreme review counts and sentiment metrics. Across all specifications, we find consistent evidence that cumulative reviews and positive word usage are positively associated with higher hotel scores, while negative sentiment significantly reduces scores.

Introduction:

Online reviews have become a critical element in shaping consumer perceptions and business reputations in the hospitality industry. As reviews accumulate over time, they serve not only as sources of information but also as signals of consistency, quality, and trustworthiness. However, while much attention has been paid to average ratings at a static point in time, less is known about how these scores evolve as new reviews are added. To capture this dynamic process, we shift focus from the aggregate average score to the individual reviewer scores that generate it, enabling a week-by-week analysis of changes within each hotel. This approach allows us to trace the trajectory of reputation development rather than merely comparing across hotels. By constructing a panel dataset at the hotel-week level and incorporating cumulative review metrics, sentiment indicators, and reviewer characteristics, we explore the mechanisms that underlie score variation over time. Weekly aggregation provides the granularity needed to detect meaningful fluctuations while maintaining data stability. Additionally, by incorporating linguistic sentiment and the timing of reviews, we go beyond numerical ratings to assess how both tone and temporal context contribute to reputation dynamics.

Literature Review

Online consumer reviews have emerged as a powerful influence on customer behavior and firm performance across various industries, particularly in the hospitality and service sectors. A large body of empirical literature has examined the quantitative and qualitative dimensions of review such as volume, sentiment, timing, and reviewer characteristics demonstrating their complex role in shaping consumer perceptions and purchase decisions. The current study contributes to this literature by adopting a time-variant, within-hotel framework to investigate how the cumulative number and content of reviews dynamically influence hotel ratings.

Several studies have demonstrated the economic value of online ratings and review volume. In his seminal paper, Luca (2011) finds that a one-star increase in Yelp ratings leads to a 5–9% increase in revenue for independent restaurants, highlighting the causal influence of average ratings on firm performance. Similarly, Anderson and Magruder (2012), in Learning from the Crowd, apply a regression discontinuity design to show that an additional half-star on Yelp significantly increases restaurant reservations, particularly in environments with limited alternative information. These studies underscore the reputational and revenue implications of online reviews, yet both rely on static or cross-sectional snapshots. In contrast, our research extends this stream by introducing a temporal dimension, tracking how a hotel's average rating evolves as more reviews accumulate, enabling the identification of longer-term reputation dynamics rather than one-off effects.

The role of review presentation and format has also attracted scholarly attention. Wang et al. (2019), in Scores vs. Stars, use a regression discontinuity framework to assess how visual representations of scores affect consumer decisions. They find that while higher numerical scores generally increase purchase probability, the conversion of scores into star ratings can introduce misleading discontinuities. This supports the notion that how reviews are presented can distort perception, complementing our investigation into whether cumulative sentiments and scores align over time with consumer experiences. Our study further adds to this domain by considering the semantic tone of review content as an element that often interacts with numerical ratings to shape consumer impressions.

Other studies have focused on volume effects in online marketplaces. Chevalier and Mayzlin (2006) investigate online book reviews and reveal that the number and valence of reviews significantly impact book sales, with negative reviews exerting a stronger influence than positive ones. Similarly, Duan, Gu, and Whinston (2008) find that while the persuasive effect of review scores is minimal, the sheer volume of reviews exerts a robust awareness effect on movie box office revenues. These findings align with our motivation to examine whether hotels with accumulating reviews experience changing average scores over time and whether early reviews exert a disproportionate influence on later consumer perceptions. Unlike these studies, however, our dataset and panel framework enable us to explore within-hotel score evolution, controlling for time-varying reviewer traits and review content.

Furthermore, the study Can Time Soften Your Opinion? by Chen and Lurie (2013) examines the timing of review submissions, showing that consumers who delay writing reviews tend to conform more closely to existing opinions, especially after negative experiences. This insight complements our temporal framework, where we analyze early vs. late reviews and test whether initial feedback anchors later ratings or whether reputation is recalibrated dynamically as more feedback accumulates.

Another closely related article is The Impact of Online Reviews and Volume Reviews on Consumer Purchase Decisions in Shopee, which finds that review volume significantly increases trust and purchase probability. Although focused on e-commerce, the findings reinforce the importance of cumulative feedback volume as an aspect central to our research question. While the Shopee study captures cross-sectional variation in purchasing behavior, our study leverages panel data to examine temporal patterns, thereby offering stronger causal insights into how reputation builds or decays over time.

Importantly, the psychological dimensions of review interpretation have been explored in the context of confirmation bias. Yin, Mitra, and Zhang (2016) find that consumers perceive reviews confirming their pre-existing beliefs (whether positive or negative) as more helpful. This behavioral lens provides context to our empirical findings on the weight of early reviews: if initial feedback is positive, subsequent reviewers may unconsciously reinforce this sentiment, which could explain reputation stickiness. Our temporal disaggregation helps unpack whether early ratings serve as anchors or if later reviews can shift the overall evaluation trajectory.

Finally, the panel-based study Do Online Reviews Matter? – An Empirical Investigation of Panel Data (Duan et al., 2008) directly informs our methodological approach. The use of panel data and time-variant controls allows for the identification of causal effects in the presence of unobserved heterogeneity. Our study mirrors this strategy by constructing a hotel-week panel dataset with cumulative and average measures of sentiment, review volume, and reviewer characteristics, enabling precise estimation of reputation evolution at the hotel level.

In summary, our study builds upon and complements existing research across multiple dimensions, review volume, sentiment, reviewer heterogeneity, temporal effects, and behavioral interpretation. By integrating these elements into a single, time-variant framework, we provide a nuanced understanding of how hotel reputations evolve in online platforms. This work contributes not only to hospitality and platform economics but also to broader discussions on digital reputation systems, online consumer behavior, and the dynamic role of user-generated content in shaping market outcomes.

Research Questions:

The first question of the research is related to quantity aspect of reviews, which is:

“How does the number of reviews affect the average score?”

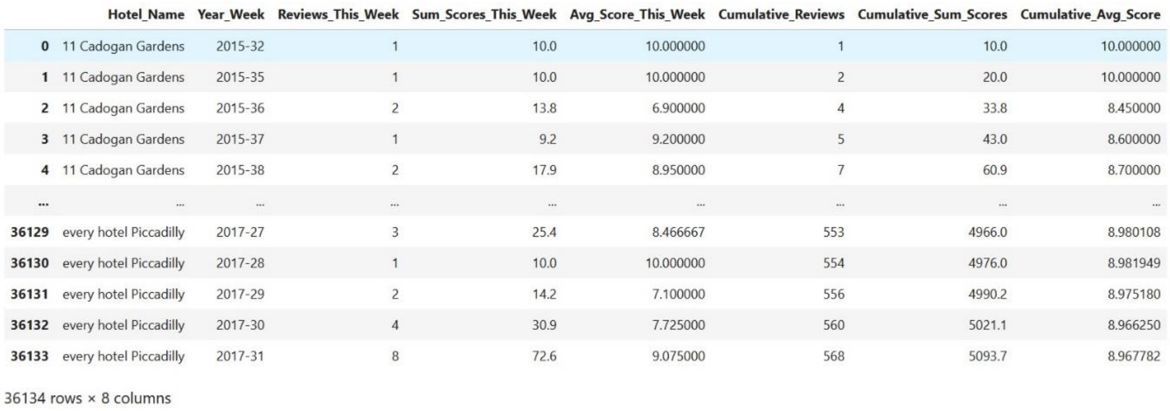
This question aims at how the number of the reviews can impact the average score. Does the high number of reviews mean more information for customers and an increase in their average score?

The challenge of this question is that the average score is the same for the same hotels. I mean, hotel X can have 500 reviews, but for all the reviews it has, the average score of Y and its average score don't change over time. To tackle this issue, we used Reviewer score (the score that every person gives to hotels) instead of Average score to create time variation in the data. So, as time goes on more reviews accumulate and the average score might change. So instead of just looking at for example, "Hotel A has 2000 reviews and an average score of 8.0," we can track “When Hotel A had 100 reviews, what was the average score?" "When it had 500 reviews, what was the average score?" "When it had 1000 reviews, what was the average score?" This enables us see the evolution of score for every hotels and answer better to the equation like do hotels tend to get higher/lower average scores as they accumulate more reviews? do first impressions (early reviews) stay stable, or does the reputation build/deteriorate over time?

As a short note, the booking.com has an algorithm in which the average score is just the average of the past reviewer score or the reviews that every person is given, and the approach we have here is consistent with what booking.com did before 2019. After 2019, they changed their algorithm and put more weight on the recent reviews.

This helps us to know how scores evolve as reviews accumulate, and we can see the dynamic effect of whether hotels build their reputation over time or start strong and fade.

To examine how hotel ratings or scores evolve over time as more customer feedback or reviews accumulates, we adopted a time variation framework in our analysis. Rather than comparing different hotels against each other, which would introduce confounding factors like quality, location, brand, and service level, we focused on how the average rating of each individual hotel changes week by week. This allowed us to isolate and understand the within-hotel dynamics of reputation formation and score shifts, making our estimates more internally valid.



	Hotel_Name	Year_Week	Reviews_This_Week	Sum_Scores_This_Week	Avg_Score_This_Week	Cumulative_Reviews	Cumulative_Sum_Scores	Cumulative_Avg_Score
0	11 Cadogan Gardens	2015-32	1	10.0	10.000000	1	10.0	10.000000
1	11 Cadogan Gardens	2015-35	1	10.0	10.000000	2	20.0	10.000000
2	11 Cadogan Gardens	2015-36	2	13.8	6.900000	4	33.8	8.450000
3	11 Cadogan Gardens	2015-37	1	9.2	9.200000	5	43.0	8.600000
4	11 Cadogan Gardens	2015-38	2	17.9	8.950000	7	60.9	8.700000
...	...	...	...	...	...	...	...	...
36129	every hotel Piccadilly	2017-27	3	25.4	8.466667	553	4966.0	8.980108
36130	every hotel Piccadilly	2017-28	1	10.0	10.000000	554	4976.0	8.981949
36131	every hotel Piccadilly	2017-29	2	14.2	7.100000	556	4990.2	8.975180
36132	every hotel Piccadilly	2017-30	4	30.9	7.725000	560	5021.1	8.966250
36133	every hotel Piccadilly	2017-31	8	72.6	9.075000	568	5093.7	8.967782

36134 rows × 8 columns

Table 1

To implement this, we constructed a panel dataset(Tab.1), where each row represents a specific hotel in a specific week. For each hotel-week combination, we calculated cumulative metrics like the total number of reviews received up to that point, the average score, and the cumulative sentiment characteristics of reviews (such as average number of positive or negative words per review). We also included reviewer-level control variables such as the average number of reviews previously written by reviewers (to proxy for reviewer experience), and the average days since each review was written.

A key design decision was to aggregate the data on a weekly basis. This choice is for a balance between granularity and stability. Daily aggregation, although very detailed, is often too sparse. Many hotels do not receive reviews every single day, especially smaller or mid-tier ones. Monthly aggregation, on the other hand, can smooth over short-term fluctuations and delay visibility into rapid changes in perception. Weekly aggregation allowed us to capture meaningful variations while avoiding excessive noise and data sparsity.

This time-based structure enabled us to estimate how cumulative reviews, sentiment, and reviewer characteristics influence a hotel’s average rating as it builds its reputation. In some versions of the analysis, we included hotel and time fixed effects to account for unobserved differences between hotels and seasonal or event-related shocks. In others, we ran regressions without fixed effects to test the robustness of the patterns we observed.

To better understand the underlying mechanisms behind customer ratings, we extended our analysis to explore the role of review content, specifically focusing on the number of positive and negative words used in customer feedback. Rather than relying solely on numerical scores, we aimed to examine how the linguistic tone of reviews, reflected in the frequency of emotionally charged words, contributes to shaping a hotel’s overall rating. So this enables us not only to focus on the quantitative aspect of reviews, but also the qualitative part and their impact on the average score.

Each review in the dataset includes a textual summary of both positive and negative experiences, written by the reviewer. We computed the word count for each of these segments to create two key indicators: Number of positive words per review and Number of negative words per review

To align these text-based variables with our time variation framework, we aggregated the data on a weekly basis for each hotel. This allowed us to construct:

The cumulative average number of positive words per review up to each week

The cumulative average number of negative words per review up to each week

By doing this, we were able to track how the tone of reviews evolves over time as the volume of feedback increases. Crucially, we did not treat these word counts as isolated events. Instead, we examined how the accumulation of linguistic sentiment, whether more enthusiastic or more critical, correlates with shifts in the cumulative average score.

Our regression model was then structured to assess whether the increasing presence of positive or negative language in reviews is associated with higher or lower average scores over time. This allowed us to answer a more nuanced question: not just whether more reviews matter, but whether what people say in those reviews matters just as much or perhaps more.

The results revealed clear patterns. A higher cumulative average of positive words was consistently associated with higher average scores, while a greater frequency of negative words predicted lower scores. This suggests that the emotional tone embedded in review text provides a meaningful signal to future customers and significantly influences the reputation trajectory of a hotel.

In short, by analysing the semantic content of reviews over time, we gained deeper insight into how customer experiences are communicated and how this communication, in turn, shapes overall perceptions. The accumulation of emotionally charged language whether positive or negative serves as a powerful channel through which a hotel's reputation is continually negotiated and recalibrated.

In this research, we are not only interested in the content and numbers of reviews, but also in the timing of those reviews, when they were posted in the life cycle of the hotel's review history. This might be important because early reviews are often disproportionately influential; they form the first impressions for future customers and can anchor the perception of quality long before the hotel accumulates a large number of reviews.

To capture this potential asymmetry in influence, we divided the cumulative review timeline into early and late periods, typically based on the total number of reviews a hotel eventually received. For example, we defined the early period as the time when the hotel had accumulated less than 25% of its final number of reviews, and the late period as the time after it had surpassed 75%. We then created interaction terms between these time period indicators and the number of cumulative reviews. This allowed us to directly compare whether the impact of reviews on average score was stronger in the early phase than in the later phase. This approach gave us a dynamic perspective on reputation building and helped us understand whether the power of reviews fades or intensifies over time. In addition to review timing, we aimed to control for characteristics of the reviewers themselves. since not all feedback carries equal weight. Two important reviewer-related variables were included:

Days Since Review: This variable captures how recent a review is by calculating the number of days between the review date and the date the data was collected. More recent reviews may be more reflective of a hotel's current quality or changes in service, whereas older reviews may no longer be relevant. Aggregating this variable weekly per hotel allowed us to account for any regency bias in scoring patterns.

Total Number of Reviews Reviewer Has Given: This serves as a proxy for reviewer experience. Experienced reviewers may rate more consistently or critically, while novice reviewers might give extreme scores. By averaging this measure weekly, we accounted for the evolving mix of reviewer types and their potential influence on average score.

Together, these controls helped us refine our model, making sure the observed patterns weren't driven simply by the age of reviews or the type of reviewers participating in different periods.

Finally, to test the robustness of our findings and explore possible heterogeneity in review behaviour, we analysed the Tags that accompany each review. These tags describe the purpose of the trip and the group type of the reviewer. Examples include:

Trip Purpose: "Business trip", "Leisure trip".

The main group type: "Couple".

We categorized these tags and computed the proportion of reviews with each tag on a weekly basis for each hotel. These proportions were then included in the regression to see whether the effects of cumulative reviews and sentiment varied by traveller type. For instance, business travellers may be more critical or focused on different aspects of service than leisure travellers, and solo travellers may have different experiences compared to those traveling with children or partners.

By including these tag-based variables, we could examine whether reputation dynamics and review sentiment function differently for different segments of the customer base. This contextual layer added further depth to our understanding of hotel ratings.

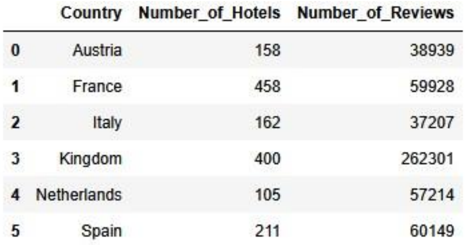
Data:

The data was scraped from Booking.com. All data in the file is publicly available to everyone already and is originally owned by Booking.com. This dataset contains 515,000 customer reviews and scores of 1493 hotels across Europe from 2015 to 2017. Meanwhile, the geographical location of hotels is also provided for further analysis.

Visualization of main dataset:

In this part, we will look at the statistics and details of the important columns or variables that we want to use in the research.

The first table (Table 2) summarizes the number of hotels and total reviews for all hotels that dataset has form different countries in Europe's which contain data from six countries. The United Kingdom has the highest number of reviews (262,301) despite having fewer hotels than France. France leads in hotel count (458), but its total reviews are significantly lower than the UK. Spain, Netherlands, and Austria follow with similar review counts around 57,000–60,000. Overall, this shows that review volume varies not just by hotel count but also by reviewer engagement in each country. Since there are many hotels from different European countries, we will pick the country with the most numbers of hotels which is United Kingdom as it seen in the below tables for our research.



	Country	Number_of_Hotels	Number_of_Reviews
0	Austria	158	38939
1	France	458	59928
2	Italy	162	37207
3	Kingdom	400	262301
4	Netherlands	105	57214
5	Spain	211	60149

Tab.2

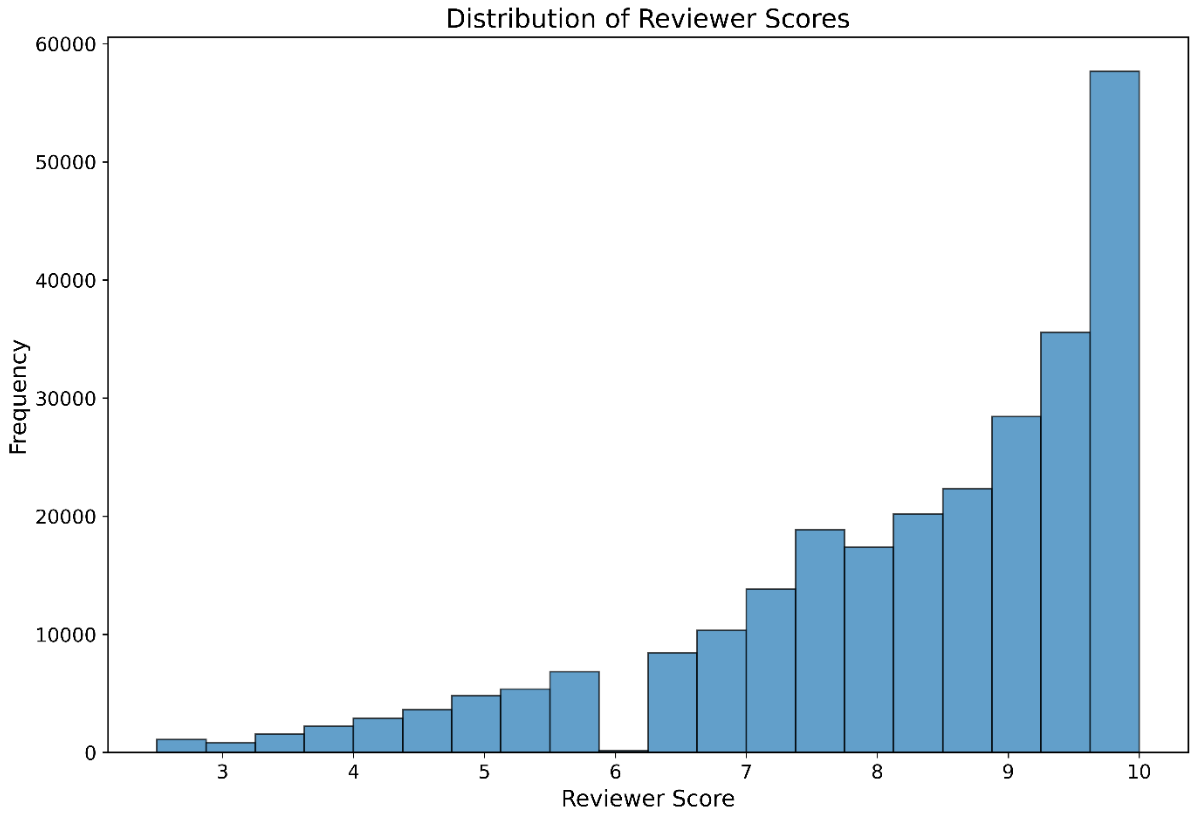


Fig .1

This histogram(Fig.1) shows the distribution of reviewer scores for UK left by hotel guests. The most frequent score is 10, indicating many guests were fully satisfied with their stay. There's a clear upward trend, which is almost steady increase from 2 to 10. Scores below 6 are relatively rare, suggesting that guests are less likely to leave very low ratings. Overall, the distribution is right-skewed, reflecting generally positive customer experiences.

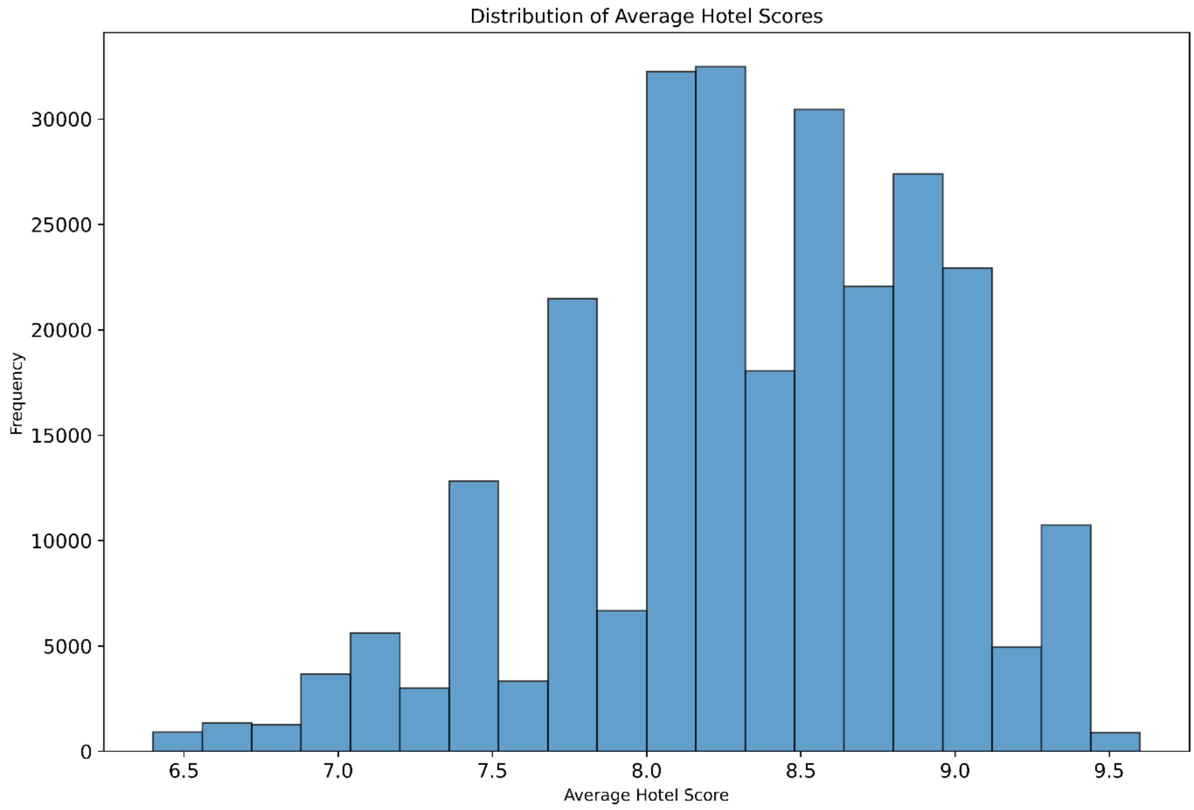


Fig2

This histogram (Fig 2) shows the distribution of average hotel scores across all hotels in UK. Most hotels have an average score between 8.0 and 9.0, with peaks around 8.2 and 8.6. The distribution is slightly left-skewed, meaning few hotels have very low average scores. Scores below 7.0 are relatively rare, suggesting most hotels maintain a good standard. Overall, the chart indicates that hotels tend to receive generally positive average ratings from guests.

The bar chart (Fig 3) displays the top 10 reviewer nationalities by number of hotel reviews for UK hotels. The United Kingdom leads with more than 160,000 reviews, far surpassing all other countries. Countries like the USA, Australia, and Ireland follow, but with significantly lower counts. The rest of the nationalities are UAE, Saudi Arabia, Switzerland, Netherlands, Canada, and Kuwait which contribute modestly and fairly evenly. Overall, UK travellers are the most active reviewers in this dataset.

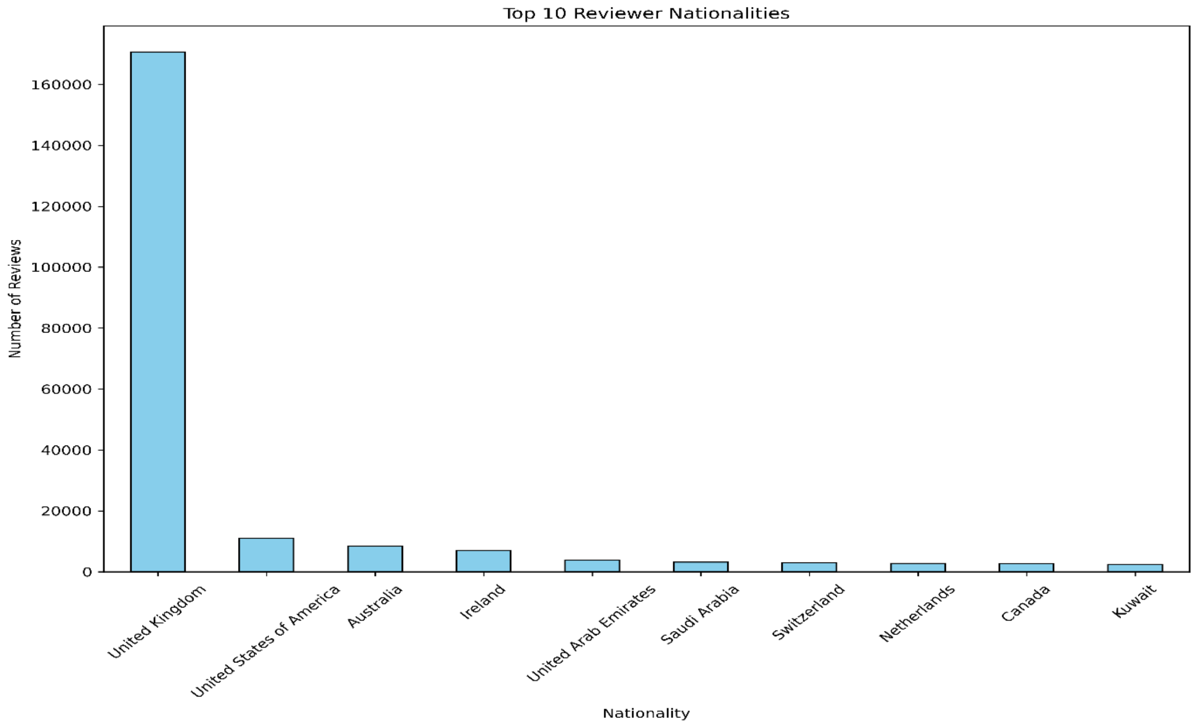


Fig 3

The histogram below (Fig. 4) shows the distribution of total negative word counts in hotel reviews for UK. Most reviews contain very few negative words, with a sharp peak close to zero. As the number of negative words increases, the frequency drops significantly. Very few reviews contain more than 100 negative words. The distribution suggests that most guests keep their negative comments brief.

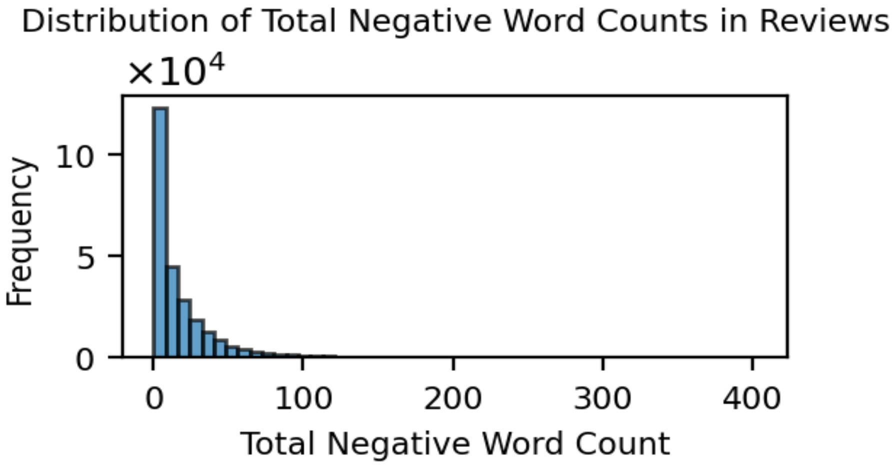


Fig 4

This histogram (Fig.5) presents the distribution of total positive word counts in UK hotel reviews. Most reviews include a small number of positive words, with a strong concentration near the lower end. The frequency drops sharply as the word count increases, showing that long positive reviews are rare. Very few reviews contain more than 100 positive words. This indicates that guests typically express their satisfaction using brief comments.

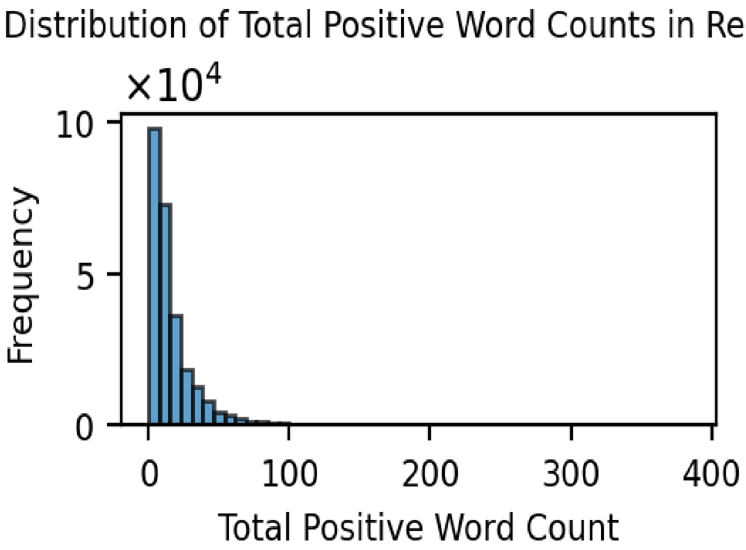


Fig 5

This histogram (Fig.6) illustrates how many additional scorings (ratings without written reviews) each hotel received. Most hotels tend to have between 200 and 500 such scorings, with a noticeable concentration around 400, indicating that it's common for hotels to get many silent ratings. The frequency gradually declines as the number of scorings increases beyond 500, showing that fewer hotels receive very large numbers of these ratings. Some outliers exist above 1500 and even over 2500, suggesting a few highly visible or popular hotels attract significantly more silent engagement. The overall shape is right-skewed, reflecting that while many hotels receive a moderate level of rating activity, only a small number receive exceptionally high counts.

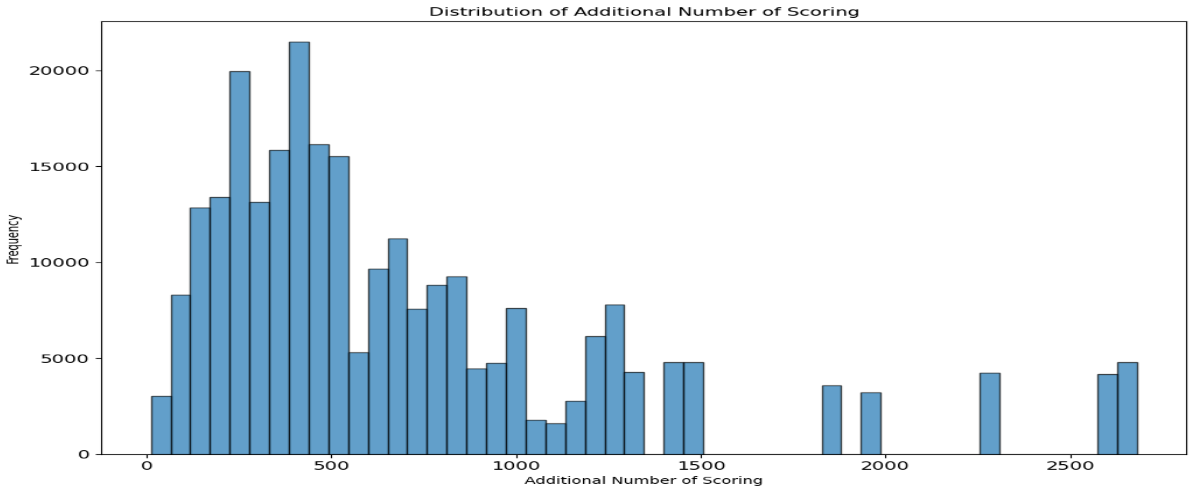


Fig.6

The scatter plot below explores the relationship between the total number of reviews and the average score for UK hotels. Each point represents one hotel, where the x-axis shows the number of total reviews and the y-axis shows the hotel's average score. Most hotels have fewer than 2,000 reviews and cluster in the average score range of 8.0 to 9.0, suggesting generally favourable impressions. There is no strong upward or downward trend, meaning somehow hotels with more reviews do not consistently receive higher or lower average scores. However, a few hotels with very high review counts (above 6,000) still maintain strong average scores, indicating consistent customer satisfaction despite greater visibility. The overall spread reflects that hotels of varying popularity can maintain similar quality ratings, and that review volume alone is not a clear indicator of rating level.

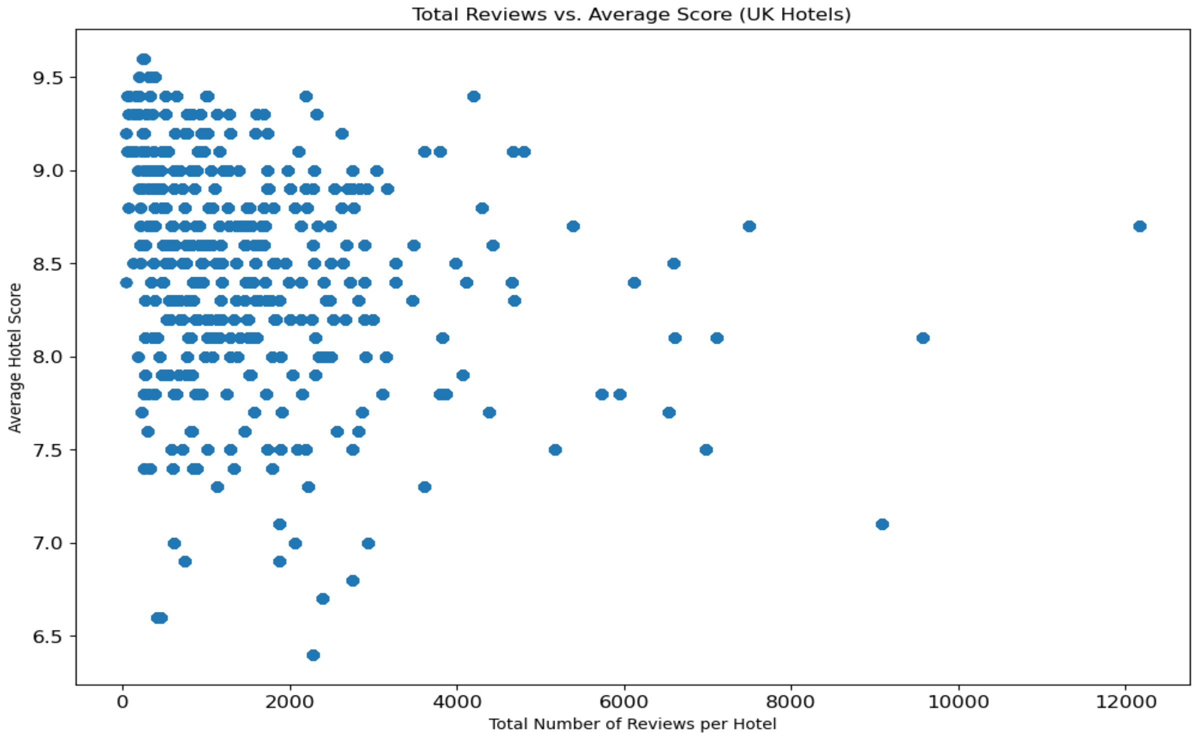


Fig 7

The two line charts (Figs 8 and 9) provide a comprehensive view of how the top 10 hotels in the UK have performed based on customer reviews. Together, they illustrate how cumulative average scores evolve over time and how they respond to the number of reviews a hotel receives.

In the first chart (Fig.8), titled "Cumulative Average Score Over Time", we see how each hotel's average rating changes from mid-2015 to mid-2017. Each line represents one hotel, with the vertical axis showing the cumulative average review score and the horizontal axis representing time by month. Most hotels start with some variability in their scores, particularly early in the timeline when fewer reviews have accumulated. However, as time passes and more reviews are received, the scores tend to stabilize, showing less fluctuation. For example, the Intercontinental London The O2 initially shows an extremely high rating, close to 9.8, but quickly drops and stabilizes around 9.4, maintaining one of the highest cumulative scores throughout the period. Similarly, hotels like the DoubleTree by Hilton and Park Plaza Westminster Bridge maintain strong and steady performance with cumulative scores above 8.5. On the other hand, some hotels like the Britannia International Hotel Canary Wharf and

Millennium Gloucester Hotel London consistently show lower ratings, clustering below the 7.5 mark. These may reflects the quality of service over time or possible changes in management, customer expectations, or service quality, as small fluctuation can be observed in several hotels around specific periods.

The second chart (Fig 9), titled “Cumulative Average Score vs. Cumulative Reviews”, presents a similar idea but from a different angle. In this chart we show how scores evolve over time, it shows how they change as more reviews accumulate. The x-axis shows the total number of reviews received, while the y-axis again displays the average score. This plot is especially useful in understanding how robust or sensitive each hotel’s rating is to increasing customer feedback. Early in the review accumulation process, there is considerable volatility, some hotels start with very high or low scores, which fluctuate sharply with the first few hundred reviews. For example, the Intercontinental again starts extremely high but drops quickly and then remains steady, reinforcing the pattern seen in the time-based plot. As more reviews accumulate, the lines flatten, illustrating how large numbers of reviews reduce the impact of each additional rating and stabilize the overall score.

By comparing the two charts, we can see that while time adds a temporal dimension to the reputation trajectory of a hotel, the number of reviews shows how sensitive that reputation is to new feedback. Hotels that maintain strong scores even after thousands of reviews, such as the Intercontinental and DoubleTree, appear to have consistent guest satisfaction, while those that decline or show instability may struggle with service consistency.

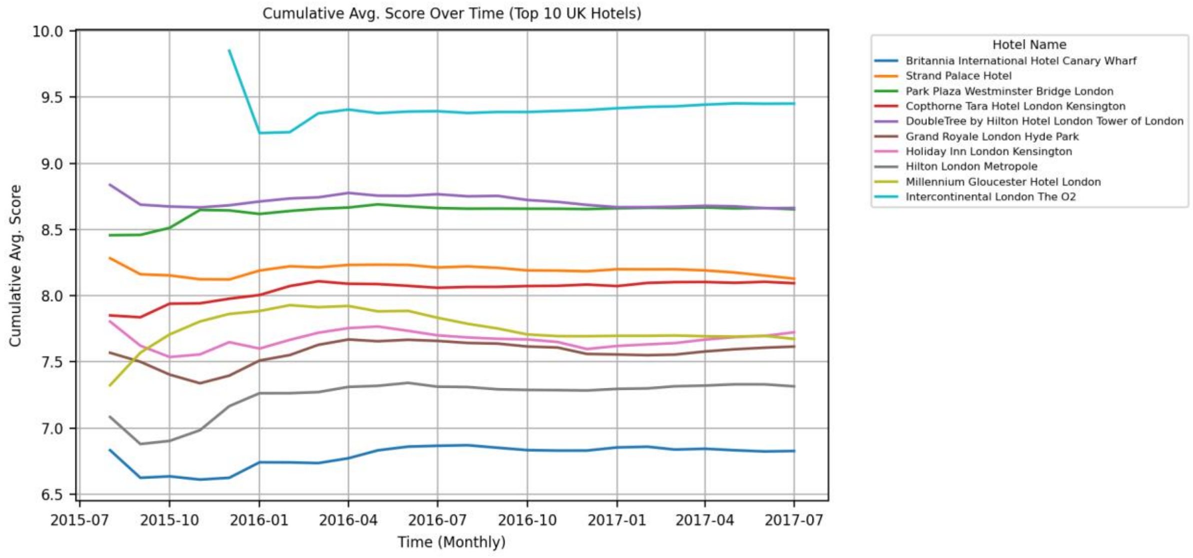


Fig 8

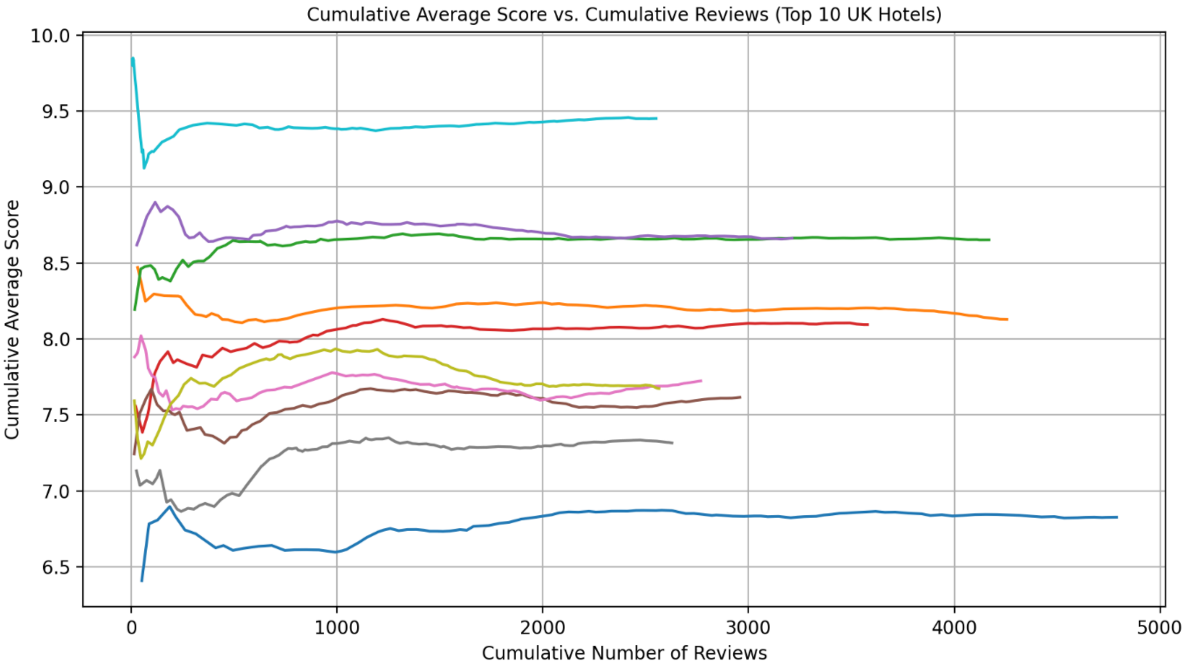


Fig 9

Methods:

In our analysis of hotel review data, we aimed to understand how various factors such as the number of reviews, the sentiment expressed in those reviews, and other dynamic features influence customer ratings. Because our dataset is structured as panel data, meaning we observe multiple reviews over time for each hotel, we needed a modelling approach that accounts for this structure. For this reason, we employed two econometric models: Pooled OLS regression and Fixed Effects regression.

We began with Pooled OLS as a baseline model. This approach treats all reviews as if they are independent observations, ignoring the fact that reviews are nested within individual hotels. In essence, it assumes that all hotels are similar in unobservable ways and that we can estimate a single relationship between the variables of interest (like total number of reviews, number of positive and negative words) and the outcome (reviewer score). This model is simple and helps establish general patterns across the dataset. However, it comes with a key limitation which is it does not account for hotel-specific characteristics that may influence review scores. For example, a luxury hotel in central London might consistently get high scores due to its inherent quality, while a budget hotel in a suburban area might receive lower scores, regardless of the number of reviews it accumulates. If we ignore such differences, the pooled regression may yield biased or misleading results.

To address this issue, we turned to the Fixed Effects model, which is particularly suited for panel data like ours. This model controls for all time-invariant characteristics of each hotel—those unchanging qualities like location, star rating, size, or overall service level. By doing so, the Fixed Effects model focuses only on within-hotel variation over time. That is, it examines how changes in variables like the number of reviews or review sentiment within the same hotel influence its reviewer scores, holding constant the hotel's baseline attributes. This approach allows us to ask a more meaningful question: When something changes for a given hotel (like when it receives a sudden influx of positive reviews), does that change affect how future guests rate it? In contrast to the pooled model, Fixed Effects gives us more causally credible insights because it removes the bias from unobserved, stable differences across hotels.

Together, these two models serve complementary purposes. The Pooled OLS model offers a broad overview of patterns across the full dataset, useful for descriptive insights. The Fixed Effects model, meanwhile, provides a more refined, within-hotel estimate that isolates the effect of time-varying variables, making it better suited for drawing conclusions about causal relationships. By using both, we can compare results and assess the robustness of our findings. If both models point to the same conclusions, we gain confidence in our results. If they differ, it's a signal that unobserved hotel-level traits are influencing the pooled estimates something the Fixed Effects model corrects for.

As I mentioned above in this analysis, we estimated two different regression models to understand what drives the average hotel review scores over time: the Pooled OLS model and the Fixed Effects panel model**.** Both models use panel data, where we observe multiple hotels (`i`) over multiple weeks (`t`), allowing us to track how review dynamics evolve for each hotel.

1: Pooled OLS

$\mathit{Cumulative\_Avg\_Score}_{it}$= β $_1$+β$_2$.log($\mathit{Cumulative\_Reviews}_{it}$)+β$_2$.log($\mathit{Lagged\_Cumulative\_Avg\_Positive}_{it-1}$)+β$_3$.($\mathit{Lagged\_Cumulative\_Avg\_Negative}_{it-1}$)+ε$_{it}$

This model assumes that all observations (hotels across weeks) are independent and identically distributed. It ignores hotel-specific or time-specific characteristics, meaning it pools all data together as if hotels are homogeneous. It provides a simple overview but may suffer from omitted variable bias due to unobserved factors like location or quality.

2: Fixed effect

$\mathit{Cumulative\_Avg\_Score}_{it}$= β $_1$+β$_2$.log($\mathit{Cumulative\_Reviews}_{it}$)+β$_2$.log($\mathit{Lagged\_Cumulative\_Avg\_Positive}_{it-1}$)+β$_3$.($\mathit{Lagged\_Cumulative\_Avg\_Negative}_{it-1}$) +η $_i$ + λ$_t$+ε$_{it}$

This model extends the first by including:

η_i : Hotel fixed effects, capturing all unobserved, time-invariant characteristics of each hotel (like brand, location, design).

λ_t : Time fixed effects, controlling for week-level shocks that affect all hotels (Like holiday seasons, pandemic effects). This model allows us to estimate how within-hotel changes over time (like review count or sentiment) affect review scores, net of stable differences across hotels

$\mathit{Cumulative\_Avg\_Score}_{it}$: The average review score for hotel i in week t (dependent variable).

log($\mathit{Cumulative\_Reviews}_{it}$):The log of the total number of reviews a hotel has received up to time t. Measures popularity or exposure.

log($\mathit{Lagged\_Cumulative\_Avg\_Positive}_{it-1}$):The log of the average number of positive words in reviews from the previous week. Captures the sentiment momentum.

($\mathit{Lagged\_Cumulative\_Avg\_Negative}_{it-1}$): The log of the average number of negative words in the previous week. Reflects trailing dissatisfaction.

ε$_{it}$: The error term.

In both models, we used log transformations for the main continuous predictors, such as the number of cumulative reviews and the average number of positive and negative words. This decision was driven by three key reasons. First, log transformations make interpretation easier and coefficients become elasticities, so we can understand the percentage change in review score resulting from a 1% change in a predictor. Second, logs help stabilize variance by compressing large values and reducing the skew found in variables like review count or word frequencies, which often range from single digits to thousands. Third, using logs reflects diminishing returns; for instance, the first few reviews have a larger influence than later ones, and the log scale captures this declining marginal impact naturally.

We also incorporated lagged variables for review sentiment, specifically, the average number of positive and negative words from the previous week. This was essential for maintaining correct temporal ordering in the model. We wanted to ensure that we're predicting current review scores based on past sentiment, not the other way around. Lagging the sentiment variables helps us avoid simultaneity bias, which can occur when both the predictor and outcome are determined at the same time. Furthermore, using lagged sentiment captures momentum or trends in how guests are feeling over time, reflecting whether a hotel is gaining or losing favor among its customers.

Result and Discussion:

The first two regression tables (tab.3 and 4) present the results of modelling average hotel review scores using Ordinary Least Squares (OLS) and Fixed Effects (PanelOLS) approaches. They tell two different stories about how review volume and sentiment affect hotel review scores; one broad and descriptive, the other focused and internally valid.

In the OLS model (Table 3), we see strong and statistically significant effects for all predictors. The coefficient for log-transformed positive words is very high (1.226), while the effect of negative words is strongly negative (−1.065). Similarly, the number of cumulative reviews (LogReviews) shows a small but significant positive impact (0.005). This model explains 61.3% of the variation in scores, suggesting high explanatory power. But this explanatory strength comes at a cost. OLS assumes that each observation is independent and treats differences between hotels as random variation rather than structured differences. This means that any systematic variation across hotels, like consistent quality differences, branding, location, or service standards is not accounted for. As a result, the OLS coefficients are likely inflated, capturing both the true within-hotel effects and any unobserved, time-invariant differences between hotels. The Fixed Effects model (Table 4) corrects for this by controlling for those time-invariant hotel-specific characteristics. When we shift to this framework, the coefficients for the same variables shrink dramatically: the effect of positive words drops from 1.226 to 0.403, and negative words from −1.065 to −0.300. Likewise, the coefficient for LogReviews increases notably, from 0.005 to 0.049. Why does this happen? This change suggests that OLS is overestimating the impact of sentiment because it attributes between-hotel variation to within-hotel dynamics. For example, a luxury hotel might consistently attract more glowing reviews than a budget hotel, but that doesn't mean sentiment is increasing for that hotel over time. The Fixed Effects model isolates only the within-hotel changes, such as how sentiment or review count shifts affect scores for a specific hotel across weeks.

The increase in the LogReviews coefficient in the fixed effects model is especially interesting. It implies that as an individual hotel accumulates more reviews over time, its average score tends to increase more than what OLS suggests. OLS likely underestimates this dynamic because it averages across hotels at different stages in their review accumulation.

In summary, the discrepancies between these tables arise because OLS captures both between-hotel and within-hotel variation, while Fixed Effects isolate the latter. Although the OLS model looks stronger statistically, it is more prone to omitted variable bias. The PanelOLS, despite a lower R^2, gives a cleaner estimate of how review sentiment and volume influence reputation trajectories within each hotel—making it the more trustworthy model for causal interpretation.

Table 3: OLS regression predicting Review Score

Variable	*B*	*SE*	*LL*	*UL*	*t*	*p*
Intercept	7.996	0.035	7.928	8.065	229.582	< .001
LogReviews	0.005	0.002	0.001	0.008	2.662	< .01
LogPositive (t-1)	1.226	0.009	1.208	1.245	131.016	< .001
LogNegative (t-1)	-1.065	0.007	-1.078	-1.051	-154.329	< .001

Note. B = unstandardized regression coefficient; SE = standard error; LL & UL = Lower and Upper bounds of 95% Confidence Interval. p-values: *p* < .05, **p** < .01, ***p*** < .001.

Table 4: Fixed Effects Regression Predicting Review Score

Variable	*B*	*SE*	*LL*	*UL*	*t*	*p*
Intercept	7.88	0.026	7.829	7.93	303.64	< .001
LogReviews	0.049	0.004	0.042	0.056	14.141	< .001
LogPositive (t-1)	0.403	0.006	0.391	0.416	64.704	< .001
LogNegative (t-1)	-0.3	0.004	-0.308	-0.291	-68.943	< .001

Note. B = unstandardized regression coefficient; SE = standard error; LL & UL = Lower and Upper bounds of 95% Confidence Interval. p-values: *p* < .05, **p** < .01, ***p*** < .001.

Tables 5 and 6 incorporate a crucial dynamic into the analysis of online hotel reviews to see how the effect of cumulative reviews changes depending on whether they are submitted in the early phase of a hotel's review lifecycle.

In the OLS model (Table 5), early reviews (the first 25% reviews) are associated with significantly lower scores (−0.229), and the interaction term between LogReviews and the Early Period is positive (0.027). This implies that during the early period, accumulating more reviews is linked with rising scores, whereas later reviews show a small negative trend (main LogReviews = −0.026). However, this OLS pattern may be distorted by the between-hotel differences it cannot control for. Newer or smaller hotels with lower visibility might naturally receive harsher scrutiny or face higher uncertainty early on, which could depress their initial average scores. Thus, OLS may misleadingly suggest that early reviews are an uphill battle when, in reality, unobserved heterogeneity is at play.

The Fixed Effects model (Table 6), however, provides a more credible story by comparing each hotel to itself over time. Here, the early period is associated with higher scores (0.072), and the interaction term is negative (−0.012), meaning that while early reviews tend to be more positive, the marginal effect of each new review on the overall score is weaker during that phase. This means in the early stage, hotels often go above and beyond to please customers, and early customers who are often loyal, enthusiastic, or first-movers may reciprocate with high scores. These reviews reflect a honeymoon period where both sides are more invested. But as time progresses and reviews become more representative of the broader customer base, the average scores become more stable, more critical voices enter the conversation, and the strong impact of each additional review in boosting the score declines.

In this light, the Fixed Effects model reveals the deeper mechanism: early reviews inflate initial reputation, but it is the later reviews that truly build or recalibrate it, reflecting the collective judgment of a more diverse customer base. This pattern is consistent with how reputations evolve in real-world settings in a way that the early impressions matter, but they eventually give way to broader consensus. The OLS model captures part of this dynamic but is confounded by cross-hotel variation. The Fixed Effects model, by contrast, isolates the temporal transition in sentiment and credibility, offering a more reliable account of how online ratings genuinely shift over a hotel's lifecycle.

Table 5: OLS Regression with Interaction Term for early reviews

Variable	*B*	*SE*	*LL*	*UL*	*t*	*p*
Intercept	8.245	0.038	8.171	8.32	216.756	< .001
LogReviews	-0.026	0.003	-0.031	-0.021	-9.71	< .001
LogPositive (t-1)	1.204	0.009	1.185	1.222	127.926	< .001
LogNegative (t-1)	-1.063	0.007	-1.076	-1.049	-154.368	< .001
Early Period	-0.229	0.022	-0.272	-0.186	-10.421	< .001
Interaction: Log Reviews × Early Period	0.027	0.005	0.018	0.036	5.894	< .001

Note. B = unstandardized regression coefficient; SE = standard error; LL & UL = Lower and Upper bounds of 95% Confidence Interval. p-values: *p* < .05, **p** < .01, ***p*** < .001.

Table 6: Fixed Effects Regression with Interaction Terms for early reviews

Variable	*B*	*SE*	*LL*	*UL*	*t*	*p*
Intercept	7.785	0.029	7.728	7.843	265.18	< .001
LogReviews	0.064	0.004	0.056	0.072	15.63	< .001
LogPositive (t-1)	0.404	0.006	0.392	0.417	64.823	< .001
LogNegative (t-1)	-0.297	0.004	-0.305	-0.288	-67.512	< .001
Early Period	0.072	0.011	0.051	0.093	6.746	< .001
Interaction: Log Reviews × Early Period	-0.012	0.002	-0.016	-0.008	-5.668	< .001

Note. B = unstandardized regression coefficient; SE = standard error; LL & UL = Lower and Upper bounds of 95% Confidence Interval. p-values: *p* < .05, **p** < .01, ***p*** < .001.

The two tables below (Tab 7 and 8) present the results of modelling average hotel review scores using both OLS and Fixed Effects (PanelOLS) regressions, with an expanded specification that introduces two new explanatory variables: reviewer experience and review recency. In the OLS model (Table 7), the age of the review has a significant negative effect (−0.024), suggesting that older reviews are associated with lower average scores. This could be interpreted as guests placing more trust in newer reviews and discounting older ones when forming perceptions about a hotel. The coefficient on reviewer experience is slightly negative, but not significant.

However, these results shift in the Fixed Effects model (Table 8), where both variables lose their significance. This difference reflects the model’s focus on within-hotel variation over time, which provides a cleaner lens to assess whether these factors actually cause score changes, as opposed to simply correlating with them due to differences across hotels.

The lack of significance for reviewer experience in the fixed effects model makes sense when we consider that each hotel tends to attract a consistent type of guest. A business hotel will consistently receive reviews from seasoned travelers, while a leisure resort may regularly be reviewed by more casual or less experienced guests. Since the composition of reviewers doesn’t change much from week to week within the same hotel, there’s very little variation in reviewer experience to explain shifts in scores. That’s why its effect disappears when controlling for fixed hotel traits.

A similar logic applies to days since review. While the OLS model may suggest that older reviews drag down average scores, this could be because hotels with older reviews may simply be lower-quality or receive less frequent attention, not because time itself affects scores. Within a hotel, however, the age of the reviews likely doesn’t vary enough week to week to have a meaningful impact. A few reviews being slightly older or newer isn’t enough to shift the average score, which explains the non-significance in the Fixed Effects model.

Ultimately, these results emphasize that between-hotel comparisons can be misleading. What appears to be a strong effect in pooled OLS may actually reflect stable differences between types of hotels, rather than true causal effects within hotels. The Fixed Effects model helps clarify that, within each hotel, reviewer experience and recency simply don't fluctuate enough to impact scores in a significant or consistent way.

Table 7: OLS Regression with Reviewer Experience & Recency

Variable	*B*	*SE*	*LL*	*UL*	*t*	*p*
Intercept	8.221	0.041	8.141	8.301	201.87	< .001
LogReviews	-0.004	0.002	-0.008	-0.0	-2.115	< .05
LogPositive (t-1)	1.215	0.009	1.196	1.233	129.007	< .001
LogNegative (t-1)	-1.067	0.007	-1.081	-1.054	-154.706	< .001
LogReviewer Experience	-0.001	0.003	-0.007	0.004	-0.406	0.685
Log Days Since Review	-0.024	0.002	-0.029	-0.02	-10.58	< .001

Note. B = unstandardized regression coefficient; SE = standard error; LL & UL = Lower and Upper bounds of 95% Confidence Interval. p-values: *p* < .05, **p** < .01, ***p*** < .001.

Table 8: Fixed Effects Regression with Reviewer Experience & Recency

Variable	*B*	*SE*	*LL*	*UL*	*t*	*p*
Intercept	7.926	0.104	7.722	8.129	76.215	< .001
LogReviews	0.049	0.004	0.042	0.056	14.129	< .001
LogPositive (t-1)	0.403	0.006	0.391	0.416	64.683	< .001
LogNegative (t-1)	-0.3	0.004	-0.308	-0.291	-68.945	< .001
LogReviewer Experience	0.001	0.001	-0.001	0.003	0.68	0.497
Log Days Since Review	-0.008	0.018	-0.044	0.027	-0.466	0.641

Note. B = unstandardized regression coefficient; SE = standard error; LL & UL = Lower and Upper bounds of 95% Confidence Interval. p-values: *p* < .05, **p** < .01, ***p*** < .001.

As a conclusion, it might be said that variation in guest type or review timing isn't large enough within a single hotel to drive meaningful score changes week to week. The true drivers of score changes are more likely to be review volume and sentiment, rather than the characteristics of who wrote the review or when.

Heterogeneous effect:

To explore the heterogeneous effects of review dynamics across different types of hotel guests, we utilized the Tags column in our dataset to distinguish between business and leisure travelers. Understanding heterogeneity is critical in this context because different types of guests may have different expectations, behaviors, and reactions to hotel quality. For example, business travelers might be more sensitive to positive words, while leisure guests might prioritize number of reviews and negative words or vice versa . By identifying the purpose of travel, we can examine whether the impact of sentiment, review volume, or time trends differs meaningfully between these two segments. This allows us to run subgroup regressions or incorporate interaction terms to test whether effects such as the influence of review positivity or recency vary by guest type, thereby capturing richer, more policy-relevant insights than would be possible with a one-size-fits-all model.

To do this, we leveraged the Tags column, which includes unstructured text entries describing each guest's travel context, such as 'Leisure trip', 'Business trip'. These tags provide valuable qualitative context, but in their raw form are not directly usable for analysis. We began by cleaning the tags—removing brackets and quotation marks and splitting them into individual strings, which we then standardized by stripping extra spaces.

From this cleaned list, we created two binary variables at the review level: one to flag whether the review came from a business trip, and another for a leisure trip. If the tag list contained 'Business trip', we marked that review as a business review (is_business_trip = 1), and similarly for leisure trips.

To translate this into a time-based panel structure, we extracted the year-week from each review date and grouped reviews by Hotel_Name and Year_Week. We then calculated, for each hotel-week, the total number of reviews and the number of business and leisure reviews. This allowed us to compute two new variables: Share_Business_Trip and Share_Leisure_Trip, representing the proportion of reviews each week that came from guests traveling for business or leisure, respectively.

We merged these travel purpose shares into our existing dataset to enrich the time-series data with behavioral context. Finally, we used these shares to segment the data into business-dominant weeks (Share_Business_Trip > 0.5) and leisure-dominant weeks (Share_Leisure_Trip > 0.5). This segmentation allowed us to test whether the drivers of review scores, such as sentiment, review volume, and reviewer experience operate differently for different types of guests, offering a robust way to analyze heterogeneous treatment effects in the customer review process.

Through this approach, we successfully transformed unstructured textual data into analytically meaningful subgroup indicators, enabling a more nuanced investigation of how hotel reputation evolves across different customer segments.

To better understand how travel purpose influences review behavior, we segmented our data into business-dominant and leisure-dominant weeks and ran both OLS and fixed effects regressions. The goal was to capture whether patterns in review volume and sentiment differ by trip type and whether modeling approaches yield consistent conclusions. The OLS regressions show that for both groups, cumulative reviews and sentiment variables are statistically significant, but the magnitudes and implications differ.

The tables (9,10,11,12 in the appendix) compare OLS and Fixed Effects regression results, split between business-dominant and leisure-dominant weeks, showing how review volume, sentiment, recency, and reviewer experience influence hotel ratings. The OLS models (Tabs 9 & 11) capture broad patterns across hotels, reflecting between-hotel variation. In contrast, the Fixed Effects models isolate within-hotel changes over time, revealing behavioral responses to evolving review signals.

In OLS, more cumulative reviews are associated with lower scores, especially during business and leisure weeks. However, this likely reflects structural differences in which hotels with many reviews may be older, larger, or more exposed to criticism. When we apply Fixed Effects, the relationship flips: within a hotel, accumulating more reviews correlates with higher scores, indicating a reputation-building effect over time.

Sentiment shows a similar pattern. In OLS, positive and negative word counts have very strong effects, especially for leisure weeks. But in Fixed Effects, these coefficients shrink, revealing that OLS exaggerates sentiment's influence due to unobserved hotel traits. Yet, tone still matters more in leisure settings, where guests seek vivid, emotionally rich cues to reduce uncertainty.

Business and leisure travellers engage with reviews differently, and this divergence helps explain the shifting magnitude and direction of coefficients across models and segments. Business travellers, typically efficiency-driven and risk-averse, are more responsive to depth (volume of reviews) and freshness (regency) than to narrative tone. Their decision-making is often time-constrained; for instance, a consultant arriving late at night seeks immediate reliability, not a poetic recount of past experiences. In such contexts, volume serves as a proxy for credibility, while recency signals up-to-date service conditions, such as renovations or management changes. Moreover, business guests often operate under standardized expectations. They for example prioritize Wi-Fi reliability, check-in speed, and breakfast hours which making aggregated, recent feedback especially informative. Since their expenses are usually covered by employers, they are less sensitive to marginal sentiment shifts and more focused on minimizing risk.

In contrast, leisure travellers tend to respond more strongly to sentiment and tone, as reflected in the larger coefficients for linguistic variables during leisure-dominant weeks. These guests are typically planners and aspirational thinkers. They embark on choice-rich search journeys, where reviews serve as storytelling mechanisms. A family evaluating a beach resort may carefully examine adjectives, narrative structure, and emotional tone to imagine the experience. Because leisure travel often involves discretionary spending, guests are sensitive to regret and uncertainty; vivid, emotionally expressive language reduces perceived risk. Additionally, leisure travellers seek experiential fit. They rely on tone to infer ambiance; which star ratings alone cannot capture. As a result, they tend to reward warm, enthusiastic language and penalize harsh or negative descriptions, far more than they respond to abstract metrics such as "1,000 reviews." These distinct cognitive and emotional heuristics explain why the same review variables yield different patterns in OLS versus Fixed Effects models and across traveller segments.

Another broader strategy to explore heterogeneous effects in hotel reviews is by identifying not only why guests travel (business vs. leisure), but also who they travel with their group type. The main group type in the data set was ‘Couple’ and we conducted the regression on this group.

These regression results focus on weeks when couples made up the majority of hotel guests, allowing us to explore heterogeneous effects by traveler type. In both the OLS and Fixed Effects models(13,14) , positive sentiment strongly increases average scores, while negative sentiment significantly lowers them confirming that couples are responsive to emotional tone in reviews. However, the effect sizes shrink in the Fixed Effects model, indicating that the pooled OLS may overstate the impact due to unobserved hotel-level differences.

Couple travellers are typically more emotionally invested in their hotel experience than solo business travellers. They're often traveling for leisure, romance, or shared relaxation, and so they tend to pay close attention to how others describe the mood, tone, and overall vibe of the hotel. That's why the regression results show that positive and negative words in reviews have a big impact on their ratings especially in the OLS model, which captures broad patterns across all hotels.

However, the OLS model also includes between-hotel differences and it doesn’t separate out whether couples are reacting to actual changes in a hotel over time, or just that some hotels consistently get more glowing or harsh reviews than others. For example, a romantic resort will always get more positive words than a noisy city hotel, but that doesn’t mean sentiment is changing within those hotels. When we shift to the Fixed Effects model, which only looks at changes within the same hotel over time, the influence of sentiment (positive or negative words) becomes smaller but still statistically significant. This tells us that while tone still matters to couples, the large effects we saw in OLS were partly driven by the type of hotel they were staying at not just by changes in review tone over time.

Interestingly, review volume has a small but significant negative effect in both models, suggesting that for couples, higher review counts do not necessarily signal better experiences. The recency of reviews (log_Avg_Days_Since_Review) is significant in OLS but becomes insignificant in Fixed Effects, implying that any apparent influence of old reviews is likely explained by stable hotel traits rather than time-specific changes.

Gaming or fake reviews:

For conducting any causal or statistical analysis using online review data, it is crucial to verify the authenticity of the data itself, especially to ensure there are no signs of fake or manipulated reviews. Review platforms can sometimes be subject to artificial inflation, where hotels or third parties submit fraudulent reviews to boost scores or popularity.

To check the integrity of the dataset, we visualized the monthly number of reviews for the middle-range group of 10 hotels. The first graph(Fig.10) shows expected variability, with periodic peaks aligning with known high-travel seasons such as summer and the beginning of each year.

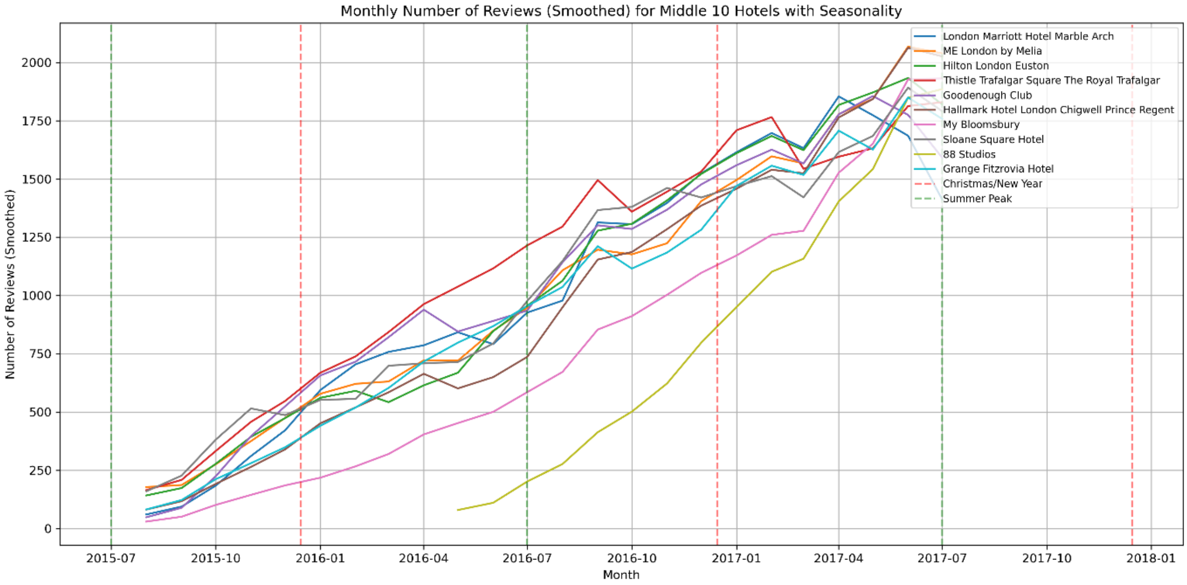


Fig 10

There are no sudden, unexplained jumps in review counts for individual hotels, which would have indicated suspicious activity. It incorporates seasonality markers (like Christmas/New Year, summer) and shows a gradual, consistent rise in reviews across all middle-tier hotels. This consistency, combined with natural seasonality, supports the credibility of the data. Overall, the visual evidence suggests that the review patterns are organic and trustworthy.

To ensure the integrity of our review data, we examined the monthly number of positive words used in reviews for both the top 10 and middle 10 hotels. This sentiment-based check helps us identify potential fake or manipulated reviews, which might show up as unnatural surges in positivity.

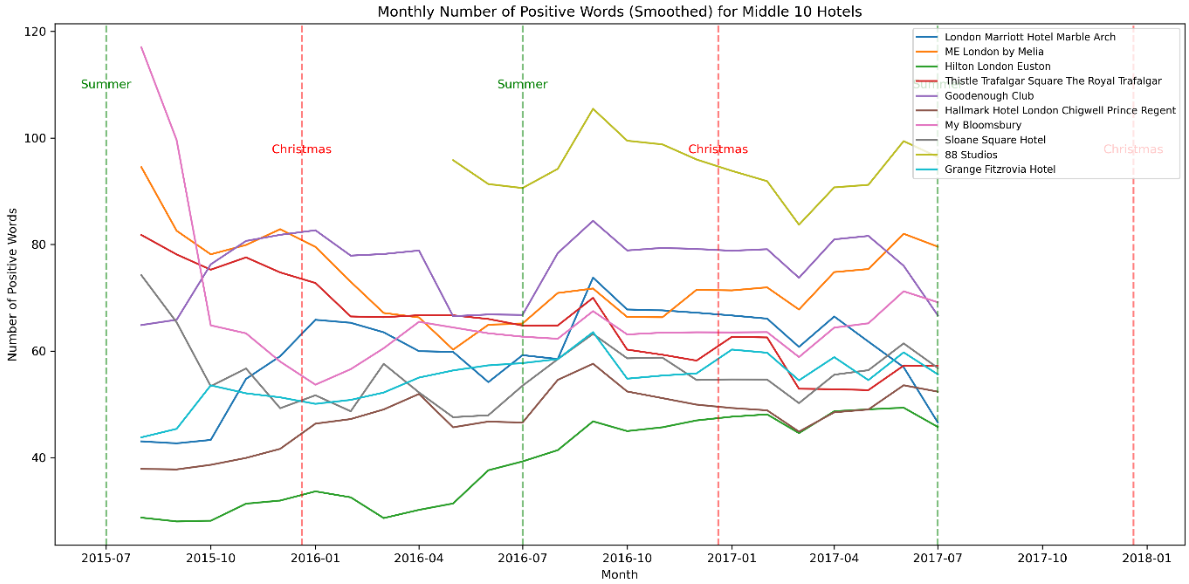


Fig.11

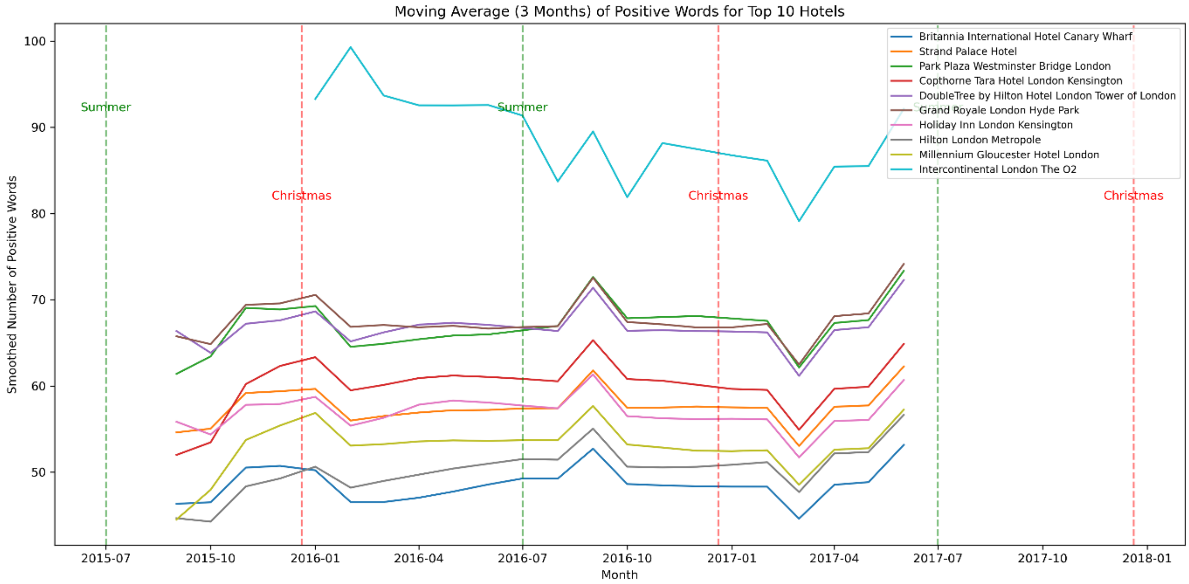


Fig.12

In the second graph(Fig 11), which covers the middle-tier hotels, we observe natural fluctuations over time with modest peaks during major travel seasons such as summer and Christmas marked clearly by dashed vertical lines. These changes align with expected tourism patterns and do not suggest artificial sentiment inflation. The first graph(Fig.12), representing the top 10 hotels, reveals a more structured seasonal rhythm, where spikes in positive words consistently occur around holidays. Importantly, these peaks are synchronized across different hotels, suggesting a shared external influence rather than manipulation. None of the hotels show sharp, isolated surges in positivity that might raise suspicion. Instead, the trends appear stable, cyclical, and realistic. Together, these visual patterns support the credibility of the reviews, providing further evidence that the sentiment expressed by guests follows expected and organic seasonal dynamics.

The below graph(fig.13) below visualize the monthly number of negative words found in reviews for the top 10 hotels, helping us detect suspicious or fake reviewing behavior.

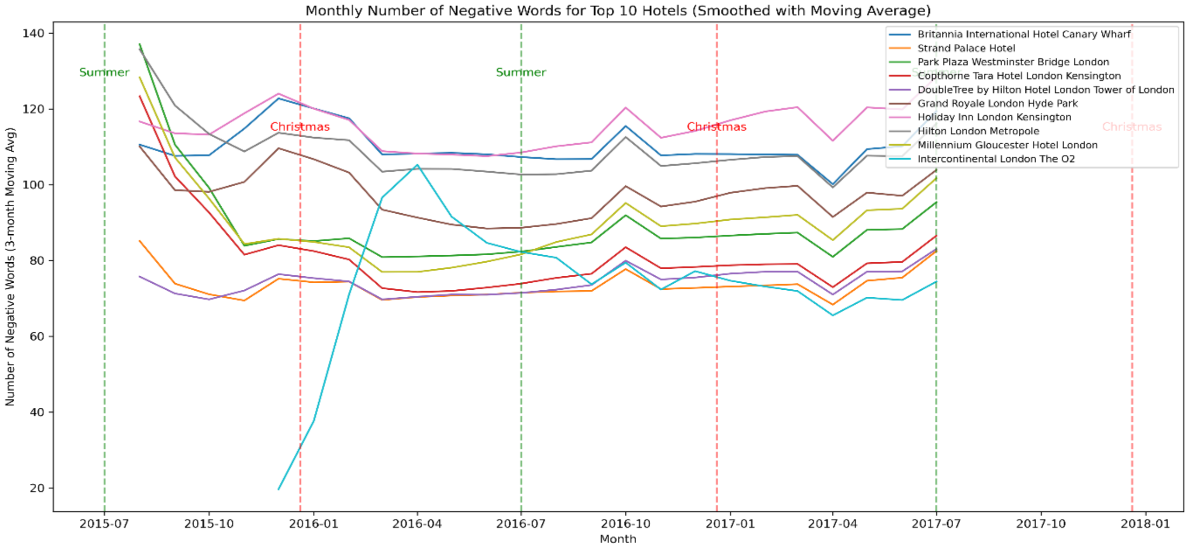


Fig.13

Figure 13 illustrates the monthly number of negative words in reviews for the top 10 hotels, smoothed with a moving average to highlight underlying trends. The overall pattern across these hotels is stable and consistent, with most properties showing a decline or plateau in negative sentiment over time. Notably, small increases occur around high-demand periods like Christmas and summer, which is expected given the influx of guests and operational strain during holidays. These seasonal fluctuations are moderate and relatively synchronized across hotels, indicating that they likely reflect genuine customer experiences rather than manipulation.

Crucially, there are no abrupt spikes or sudden drops that would suggest suspicious behavior such as fake review injection or large-scale review suppression. The lines follow smooth trajectories, with no single hotel displaying a dramatic deviation from the group norm. This consistency suggests that the reviews for these top-performing hotels are organically generated and reflect real-time service dynamics and seasonal crowding, rather than artificial attempts to boost or damage reputation. Overall, these visualizations support the notion that review behavior appears natural, with patterns closely tied to expected seasonal trends.

To ensure the reliability of our findings and rule out the influence of potentially fake or manipulated reviews, we implemented a robustness check by systematically removing hotels that could disproportionately affect the results.

Specifically, we excluded all hotels that belonged to any of the following three categories: those with the highest number of total reviews (10 hotels), those with the most positive words(10 hotels), and those with the most negative words(10 hotels). These hotels are most likely to be either highly commercial, overly managed in terms of online presence, or possibly subject to artificial inflation of reviews. This step is essential because extreme values, whether in volume or sentiment, can distort average behaviors and relationships, potentially leading to biased regression estimates.

After removing these high-impact hotels, we re-estimated both the OLS(Tab.15) and fixed effects models(Tab.16) and compared the outcomes to those derived from the full dataset. The comparison reveals a remarkable consistency in the results. In the OLS model (tab.15), the R-squared remains high

(0.662), and the signs and significance of the coefficients on cumulative reviews, positive words, and negative words remain unchanged. This indicates that the explanatory power of the model is not being artificially driven by extreme cases. Similarly, the fixed effects model(Tab.16), which controls for unobserved time-invariant characteristics of hotels and common shocks over time, also shows no meaningful deviation from the full-sample estimates. Coefficients remain statistically significant and directionally consistent, reaffirming the robustness of our results.

This approach thus not only helps us identify and account for potentially unreliable patterns in the data but also confirms that our key relationships, particularly the effects of cumulative reviews and sentiment on average score, are stable and not driven by a handful of extreme hotels. The overall narrative remains intact: customer sentiment and review volume are meaningful predictors of hotel reputation, and this holds true even when we exclude the most extreme contributors.

Implication:

The strong business implication of this research is that hotels can no longer rely solely on maintaining high average star ratings; instead, they must continuously monitor the tone, volume, and recency of guest reviews to understand how reputation evolves in real time. Our findings show that reputation is shaped not just by the number of reviews, but by the emotional content of those reviews and when they are written. This means hotel managers need to actively manage and guide the review process, prompting timely, sentiment-rich feedback from guests to sustain a positive trajectory.

Rather than treating reviews as passive reflections, businesses can use them strategically to shape perception, build trust, and differentiate in competitive markets. For example, encouraging recent reviews after service improvements can help reset negative impressions, while amplifying authentic positive sentiment can strengthen brand identity. In short, our work offers actionable insights into how review dynamics affect reputation over time, empowering hospitality firms to shift from reactive reputation management to proactive reputation strategy.

While our study provides an insights into how review volume, sentiment, and timing influence hotel scores across different contexts, it also opens several avenues for future research. One key limitation is that we relied solely on observable review data without access to actual booking outcomes, which constrains our ability to link perception to behavior. Additionally, while fixed effects allowed us to control for unobserved hotel characteristics, we could not fully disentangle platform design effects or detect manipulation mechanisms with certainty. Future research could explore how platform algorithms, such as those prioritizing recent reviews, shape visibility and consumer trust. Investigating the role of reviewer networks, elite user influence, or even AI-generated content could also shed light on emerging credibility challenges. Further, matching reviews to booking data would allow a more direct study of how sentiment affects purchasing. Expanding the linguistic analysis to capture emotional nuance or exploring cross-segment spillover effects would also enrich our understanding of review dynamics. Lastly, exploring regional and cultural differences in review interpretation could support more tailored hospitality strategies in global markets.

Conclusion:

This study examined how different review characteristics, such as sentiment, volume, recency, and reviewer experience, shape average hotel scores across various traveller types and travel purposes, using rich observational data from Booking.com. Our analysis combined pooled OLS and fixed effects panel regressions, offering both broad associations and more rigorous within-hotel causal insights.

We found that cumulative reviews and sentiment variables significantly impact perceived hotel quality, but their effects differ meaningfully between business and leisure travelers.

Positive sentiment has a stronger influence on average scores in leisure-dominant weeks, where emotionally motivated travelers are more responsive to joyful experiences and expressive language. In contrast, business travelers are more sensitive to numbers of reviews and tend to penalize poor service more heavily, suggesting higher expectations for consistency and professionalism. Reviewer experience and recency of reviews also matter, but their effects are more pronounced in OLS models and weaken under fixed effects, especially for business-dominant data.

When examining couples specifically, the findings reveal that they respond meaningfully to both positive and negative sentiment, with consistent effects across OLS and fixed effects models. Couples seem to value emotional tone but are less reactive to review volume or freshness, suggesting a more holistic, stable approach to evaluating hotel quality.

To test the robustness of our results and rule out fake review distortions, we visually explored review patterns over time. We assessed seasonal trends (like around Christmas and summer) and found no major anomalies that would suggest systematic review manipulation. Additionally, we removed hotels with the highest volumes of reviews, positive words, and negative words to reduce the potential influence of extreme or artificially inflated review behavior. Even after excluding these outliers, our core findings remained consistent and validating the credibility of our results.

Overall, this research demonstrates the importance of sentiment, volume, and temporal patterns in shaping hotel ratings and shows that both traveller intent and group composition moderate these effects. It also underscores the value of combining fixed effects models with visual diagnostics to ensure credible, interpretable insights in platforms vulnerable to review inflation or manipulation.

References:

Luca, Michael. "Reviews, reputation, and revenue: The case of Yelp. com." Com (March 15, 2016). Harvard Business School NOM Unit Working Paper 12-016 (2016).

Anderson, M., & Magruder, J. (2012). Learning from the crowd: Regression discontinuity estimates of the effects of an online review database. The Economic Journal, 122(563), 957-989.

Wang, W., Li, F., & Yi, Z. (2019). Scores vs. stars: A regression discontinuity study of online consumer reviews. Information & Management, 56(3), 418-428.

Chevalier, J. A., & Mayzlin, D. (2006). The effect of word of mouth on sales: Online book reviews. Journal of marketing research, 43(3), 345-354.

Duan, W., Gu, B., & Whinston, A. B. (2008). Do online reviews matter?—An empirical investigation of panel data. Decision support systems, 45(4), 1007-1016.

Li, H., Qi, R., Liu, H., Meng, F., & Zhang, Z. (2021). Can time soften your opinion? The influence of consumer experience valence and review device type on restaurant evaluation. International Journal of Hospitality Management, 92, 102729.

Simanjuntak, S. M., Luthfiyyah, S. P., Wulanda, A., & Situmorang, S. H. THE IMPACT OF ONLINE REVIEWS AND VOLUME REVIEWS ON CONSUMER PURCHASE DECISIONS IN SHOPEE: A QUANTITATIVE ANALYSIS.

Yin, D., Mitra, S., & Zhang, H. (2016). Research note—When do consumers value positive vs. negative reviews? An empirical investigation of confirmation bias in online word of mouth. Information Systems Research, 27(1), 131-144.

Duan, W., Gu, B., & Whinston, A. B. (2008). Do online reviews matter?—An empirical investigation of panel data. Decision support systems, 45(4), 1007-1016.

Appendix:

The csv file contains 17 fields. The description of each field is as below:

Hotel_Address: Address of hotel.

Review_Date: Date when reviewer posted the corresponding review.

Average_Score: Average Score of the hotel, calculated based on the latest comment in the last year.

Hotel_Name: Name of Hotel

Reviewer_Nationality: Nationality of Reviewer

Negative_Review: Negative Review the reviewer gave to the hotel. If the reviewer does not give the negative review, then it should be: 'No Negative'

Review_Total_Negative_Word_Counts: Total number of words in the negative review.

Positive_Review: Positive Review the reviewer gave to the hotel. If the reviewer does not give the negative review, then it should be: 'No Positive'

Review_Total_Positive_Word_Counts: Total number of words in the positive review.

Reviewer_Score: Score the reviewer has given to the hotel, based on his/her experience

Total_Number_of_Reviews_Reviewer_Has_Given: Number of Reviews the reviewers have given in the past.

Total_Number_of_Reviews: Total number of valid reviews the hotel has.

Tags: Tags reviewer gave the hotel.

days_since_review: Duration between the review date and scrape date.

Additional_Number_of_Scoring: There are also some guests who just made a scoring on the service rather than a review. This number indicates how many valid scores without review in there.

lat: Latitude of the hotel

lng: longtitude of the hotel

Table 9: OLS Regression: Business-Dominant Weeks

Variable	***B***	***SE***	***LL***	***UL***	***t***	***p***
Intercept	9.142	0.169	8.811	9.473	54.13	< .001
LogReviews	-0.072	0.01	-0.091	-0.053	-7.485	< .001
LogPositive (t-1)	1.115	0.036	1.044	1.186	30.857	< .001
LogNegative (t-1)	-1.081	0.027	-1.134	-1.029	-40.192	< .001
LogReviewer Experience	0.016	0.01	-0.003	0.035	1.628	0.104
Log Days Since Review	-0.099	0.012	-0.123	-0.074	-7.943	< .001

Note. B = unstandardized regression coefficient; SE = standard error; LL & UL = Lower and Upper bounds of 95% Confidence Interval. p-values: *p* < .05, **p** < .01, ***p*** < .001.

Table 10: OLS Regression: Leisure-Dominant Weeks

Variable	***B***	***SE***	***LL***	***UL***	***t***	***p***
Intercept	8.115	0.044	8.028	8.201	183.36	< .001
LogReviews	-0.005	0.002	-0.009	-0.001	-2.406	< .05
LogPositive (t-1)	1.24	0.01	1.22	1.26	120.574	< .001
LogNegative (t-1)	-1.052	0.008	-1.067	-1.038	-140.288	< .001
LogReviewer Experience	-0.002	0.003	-0.008	0.004	-0.593	0.553
Log Days Since Review	-0.021	0.002	-0.026	-0.016	-8.727	< .001

Note. B = unstandardized regression coefficient; SE = standard error; LL & UL = Lower and Upper bounds of 95% Confidence Interval. p-values: *p* < .05, **p** < .01, ***p*** < .001.

Table 11: Fixed Effects Regression: Business-Dominant Weeks

Variable	*B*	*SE*	*LL*	*UL*	*t*	*p*
Intercept	7.954	0.649	6.681	9.227	12.251	< .001
LogReviews	0.086	0.017	0.053	0.118	5.114	< .001
LogPositive (t-1)	0.362	0.028	0.308	0.416	13.106	< .001
LogNegative (t-1)	-0.243	0.02	-0.282	-0.205	-12.506	< .001
LogReviewer Experience	-0.0	0.004	-0.008	0.007	-0.137	0.891
Log Days Since Review	-0.068	0.115	-0.294	0.157	-0.596	0.551

Note. B = unstandardized regression coefficient; SE = standard error; LL & UL = Lower and Upper bounds of 95% Confidence Interval. p-values: *p* < .05, **p** < .01, ***p*** < .001.

Table 12: Fixed Effects Regression: Leisure-Dominant Weeks

Variable	*B*	*SE*	*LL*	*UL*	*t*	*p*
Intercept	7.936	0.109	7.722	8.149	72.88	< .001
LogReviews	0.045	0.004	0.037	0.052	11.774	< .001
LogPositive (t-1)	0.418	0.007	0.405	0.432	61.609	< .001
LogNegative (t-1)	-0.303	0.005	-0.312	-0.294	-64.807	< .001
LogReviewer Experience	0.002	0.001	-0.0	0.004	1.851	0.064
Log Days Since Review	-0.009	0.019	-0.046	0.028	-0.475	0.634

Note. B = unstandardized regression coefficient; SE = standard error; LL & UL = Lower and Upper bounds of 95% Confidence Interval. p-values: *p* < .05, **p** < .01, ***p*** < .001.

Table 13: OLS Regression: Couple-Only Sample

Variable	*B*	*SE*	*LL*	*UL*	*t*	*p*
Intercept	8.22	0.109	8.006	8.434	75.351	< .001
LogReviews	-0.015	0.007	-0.028	-0.001	-2.136	< .05
LogPositive (t-1)	1.006	0.027	0.952	1.059	36.974	< .001
LogNegative (t-1)	-0.803	0.016	-0.834	-0.771	-49.803	< .001
LogReviewer Experience	0.026	0.009	0.009	0.043	3.05	< .01
Log Days Since Review	-0.036	0.008	-0.052	-0.021	-4.591	< .001

Note. B = unstandardized regression coefficient; SE = standard error; LL & UL = Lower and Upper bounds of 95% Confidence Interval. p-values: *p* < .05, **p** < .01, ***p*** < .001.

Table 14: Fixed Effects Regression: Couple-Only Sample

Variable	*B*	*SE*	*LL*	*UL*	*t*	*p*
Intercept	9.262	0.593	8.099	10.424	15.62	< .001
LogReviews	-0.037	0.016	-0.069	-0.006	-2.322	< .05
LogPositive (t-1)	0.067	0.021	0.026	0.108	3.196	< .01
LogNegative (t-1)	-0.114	0.011	-0.137	-0.092	-10.122	< .001
LogReviewer Experience	-0.002	0.005	-0.012	0.007	-0.484	0.628
Log Days Since Review	-0.056	0.104	-0.26	0.148	-0.536	0.592

Note. B = unstandardized regression coefficient; SE = standard error; LL & UL = Lower and Upper bounds of 95% Confidence Interval. p-values: *p* < .05, **p** < .01, ***p*** < .001.

Table 15: OLS Regression Excluding Top 10 Hotels (Robustness Check)

Variable	*B*	*SE*	*LL*	*UL*	*t*	*p*
Intercept	8.038	0.036	7.968	8.109	223.559	< .001
LogReviews	0.021	0.002	0.018	0.025	11.429	< .001
LogPositive (t-1)	1.296	0.01	1.277	1.315	134.947	< .001
LogNegative (t-1)	-1.179	0.007	-1.193	-1.165	-164.796	< .001

Note. B = unstandardized regression coefficient; SE = standard error; LL & UL = Lower and Upper bounds of 95% Confidence Interval. p-values: *p* < .05, **p** < .01, ***p*** < .001.

Table 16: Fixed Effects Regression Excluding Top 10 Hotels (Robustness Check)

Variable	*B*	*SE*	*LL*	*UL*	*t*	*p*
Intercept	7.838	0.027	7.785	7.891	288.49	< .001
LogReviews	0.053	0.004	0.046	0.061	14.15	< .001
LogPositive (t-1)	0.425	0.006	0.412	0.437	64.889	< .001
LogNegative (t-1)	-0.308	0.005	-0.317	-0.299	-66.188	< .001

Note. B = unstandardized regression coefficient; SE = standard error; LL & UL = Lower and Upper bounds of 95% Confidence Interval. p-values: *p* < .05, **p** < .01, ***p*** < .001.