Analyzing Customer Sentiment to Enhance Satisfaction and Retention in the Hospitality Industry

Nicolas Fabre
Southern Utah University 2024

Abstract

This study investigates the feasibility of applying sentiment analysis to extract actionable insights from customer reviews and ratings through regression analysis, in order to enhance satisfaction and retention rates in the hospitality industry. By concentrating on how understanding consumer sentiment may lead to focused interventions, this research studies how sentiment analysis can be used to improve overall satisfaction and retention. The findings offer useful insights for firms looking to improve their client experience and overall success.

Introduction

“Your most unhappy customer are your greatest source of learning” – Bill Gates.

Understanding and managing customer dissatisfaction is critical in the hotel sector for enhancing service quality and retaining guests. This capstone project, titled "Analyzing Customer Sentiment to Enhance Satisfaction and Retention in the Hospitality Industry," seeks to investigate the correlations between several areas of hotel service and overall customer contentment. This study uses consumer feedback analytics to identify actionable insights that can assist hospitality companies in not only meeting, but exceeding, guest expectations.

The aim of this research is to explore the effectiveness of utilizing sentiment analysis, for extracting insights from customer reviews and feedback. This investigation, centered on the hospitality and marketing sectors delves into how grasping customer sentiments could result in tailored strategies that enhance satisfaction levels and customer retention. The primary objective of the study is to pinpoint factors that influence customer experiences and loyalty through a thorough evaluation of different hotel amenities and their influence, on customer feedback.

In addition to using sentiment analysis, it is critical to evaluate existing studies on the value of consumer feedback in the hospitality business. According to a study published in Tourism and Hospitality Research (2008), client satisfaction is heavily influenced by a variety of service aspects such as cleanliness, staff behavior, and overall value for money. This study demonstrates that knowing and responding to customer feedback can result in considerable improvements in service quality and guest retention. By combining these insights with sentiment analysis, this initiative hopes to give a comprehensive strategy to improving customer satisfaction and loyalty in the hotel business.

Additionally, employee or staff behavior has consistently been shown in research to have a significant impact on customer perceptions of service quality and overall satisfaction. According to Kattara, Weheba, and El-Said (2008), both positive and negative employee behaviors are strongly associated with customer satisfaction, regardless of gender, nationality, or reason of visit. This emphasizes the critical role employees play in influencing the client experience. In the hospitality sector, where service quality is crucial, hotel worker behaviors can either increase or detract from visitors' perceived value and satisfaction. According to the report, by prioritizing employee development and encouraging good behaviors, hospitality managers may strategically improve service quality and client retention.

Sentiment analysis has advanced dramatically and intensively, allowing tourists to make decisions based on rapid glimpses of visitor input. A recent study on sentiment analysis in hotel reviews revealed the need of employing natural language processing (NLP) and machine learning to capture the emotional nuances of visitor remarks and classify them as positive, negative, or neutral. This technology, also known as opinion mining, enables more efficient comparison of hotel options, hence improving the client experience by offering quick, complete insights into many service characteristics. Implementing sentiment analysis techniques can assist hoteliers, online travel agencies, and review platforms better analyze and respond to client perceptions (Altexsoft, 2021). By incorporating these innovative methods, this capstone project aims to analyze the relationships between different aspects of hotel service and guest satisfaction, offering practical recommendations to enhance customer experiences in the hospitality industry.

In the hospitality sector, recruiting new customers is substantially more expensive than retaining existing ones, but many experts have difficulty convincing executives to use customer feedback analytics to improve guest experiences. A thorough guide on customer feedback analytics states that the goal is to gather, clean, and analyze feedback data systematically in order to obtain useful insights. This guide shows how examining 50,000 TripAdvisor reviews can help identify areas for development and make practical recommendations. By exploiting such data-driven insights, hospitality companies can improve visitor satisfaction and long-term loyalty, resulting in a more profitable business ("Customer Feedback Analytics in Hospitality: Full Guide").

This capstone project intends to study the correlations between many areas of hotel service and client happiness using regression analysis, and to provide practical recommendations to improve customer experiences in the hospitality industry.

Data & Methodology

The information is structured in a CSV file containing details, about hotels, customer feedback and ratings. The main objective of utilizing this data is to assess sentiments and gather insights that could enhance customer loyalty and satisfaction within the hospitality industry.

Each hotel entry encompasses elements like the hotels name and ZIP code. Customers have rated aspects of their stay such as cleanliness, comfort, amenities, service quality and value for money on a scale from zero to ten. Additionally the dataset indicates the availability of amenities like WiFi, parking options or other facilities using Boolean and Integer variables.

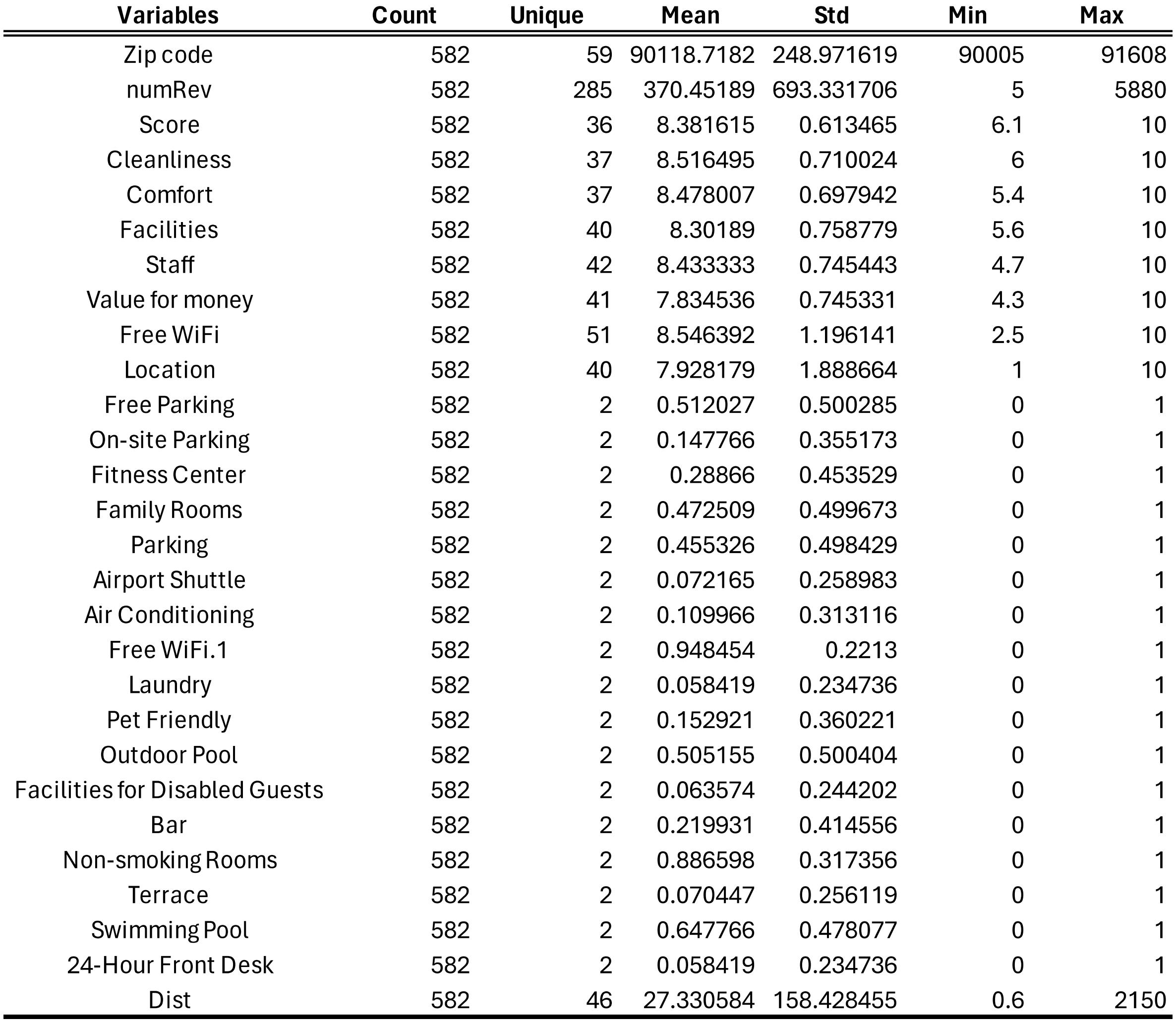
A significant component of the data is the score assigned to each hotel (Score) serving as our dependent variable for analysis. This extensive dataset allows us to explore how various hotel attributes influence customer contentment.

This refined dataset lays a groundwork, for our study aimed at identifying which hotel characteristics significantly impact review scores.

With the data, at hand our goal is to provide insights that can help hotel managers and marketers enhance service quality and guest satisfaction ultimately leading to retention rates. In my research work I utilized Python in conjunction with Excel and Google Colab (Notebook). This combination allowed me to generate summaries representations, like tables and charts and conduct regression analysis to evaluate different models.

Summary Table

Table 1. Summary of Kaggle booking.com dataset. (Excel)



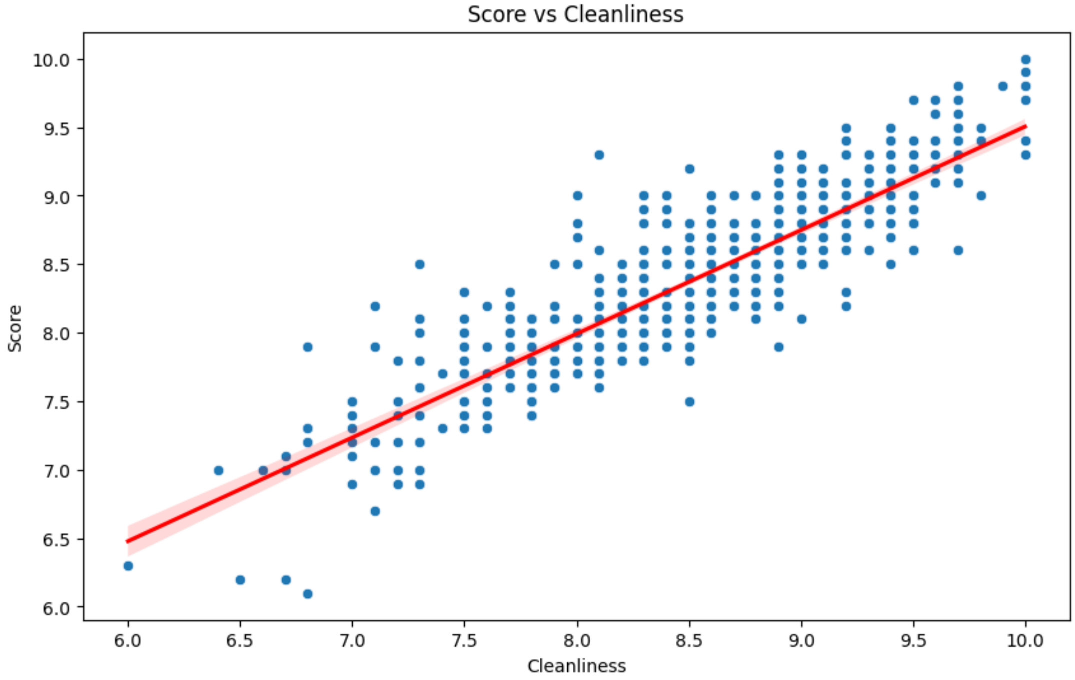
Variables	Count	Unique	Mean	Std	Min	Max
Zip code	582	59	90118.7182	248.971619	90005	91608
numRev	582	285	370.45189	693.331706	5	5880
Score	582	36	8.381615	0.613465	6.1	10
Cleanliness	582	37	8.516495	0.710024	6	10
Comfort	582	37	8.478007	0.697942	5.4	10
Facilities	582	40	8.30189	0.758779	5.6	10
Staff	582	42	8.433333	0.745443	4.7	10
Value for money	582	41	7.834536	0.745331	4.3	10
Free WiFi	582	51	8.546392	1.196141	2.5	10
Location	582	40	7.928179	1.888664	1	10
Free Parking	582	2	0.512027	0.500285	0	1
On-site Parking	582	2	0.147766	0.355173	0	1
Fitness Center	582	2	0.28866	0.453529	0	1
Family Rooms	582	2	0.472509	0.499673	0	1
Parking	582	2	0.455326	0.498429	0	1
Airport Shuttle	582	2	0.072165	0.258983	0	1
Air Conditioning	582	2	0.109966	0.313116	0	1
Free WiFi.1	582	2	0.948454	0.2213	0	1
Laundry	582	2	0.058419	0.234736	0	1
Pet Friendly	582	2	0.152921	0.360221	0	1
Outdoor Pool	582	2	0.505155	0.500404	0	1
Facilities for Disabled Guests	582	2	0.063574	0.244202	0	1
Bar	582	2	0.219931	0.414556	0	1
Non-smoking Rooms	582	2	0.886598	0.317356	0	1
Terrace	582	2	0.070447	0.256119	0	1
Swimming Pool	582	2	0.647766	0.478077	0	1
24-Hour Front Desk	582	2	0.058419	0.234736	0	1
Dist	582	46	27.330584	158.428455	0.6	2150

The dataset includes 582 entries detailing aspects of the hotel. The "Score" feature stands out with a rating of 8.38 and a standard deviation of 0.61 suggesting a level of overall customer satisfaction. "Cleanliness" also receives feedback averaging 8.52 with a deviation of 0.71 emphasizing the importance guests place on cleanliness. Similarly "Comfort" garners ratings, averaging 8.48 with a score slightly lower at 5.4 indicating that while most hotels offer comfortable accommodations some may need improvement. The average rating, for "facilities" is 8.30; however the larger standard deviation of 0.76 suggests variations, in amenities provided.

The average rating, for the "Staff" category is 8.43 indicating interactions overall despite some instances of subpar service with a score of 4.7. Ratings for "Value for money" are slightly lower at 7.83 suggesting that while pricing is generally considered fair by guests there are cases where expectations may not be met in terms of cost. "Free WiFi" receives an score of 8.55; however the standard deviation of 1.20 implies varying internet quality across different locations. The "Location" category has a score of 7.93 indicating that while most hotel locations are satisfactory some may not be as ideally positioned as others. These key factors suggest that while hotels generally excel in areas like satisfaction, cleanliness, comfort, and staff service improvements could be made in aspects such as value, for money and WiFi reliability.

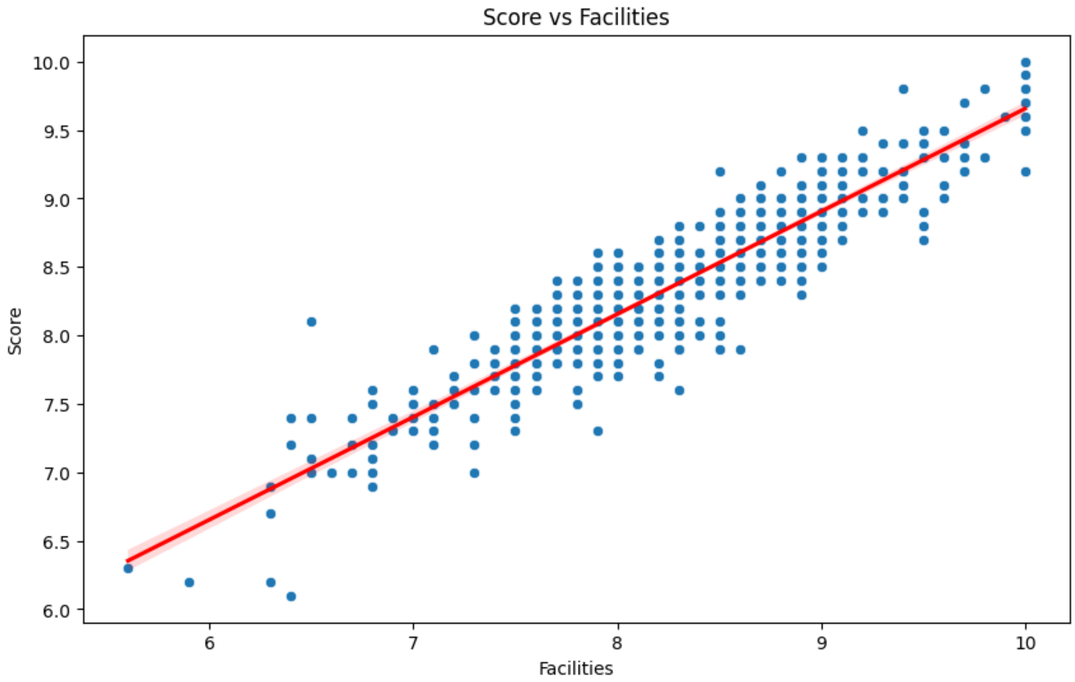
Chart Regression & Analysis

Scatter Plot 1 - Score vs Cleanliness.



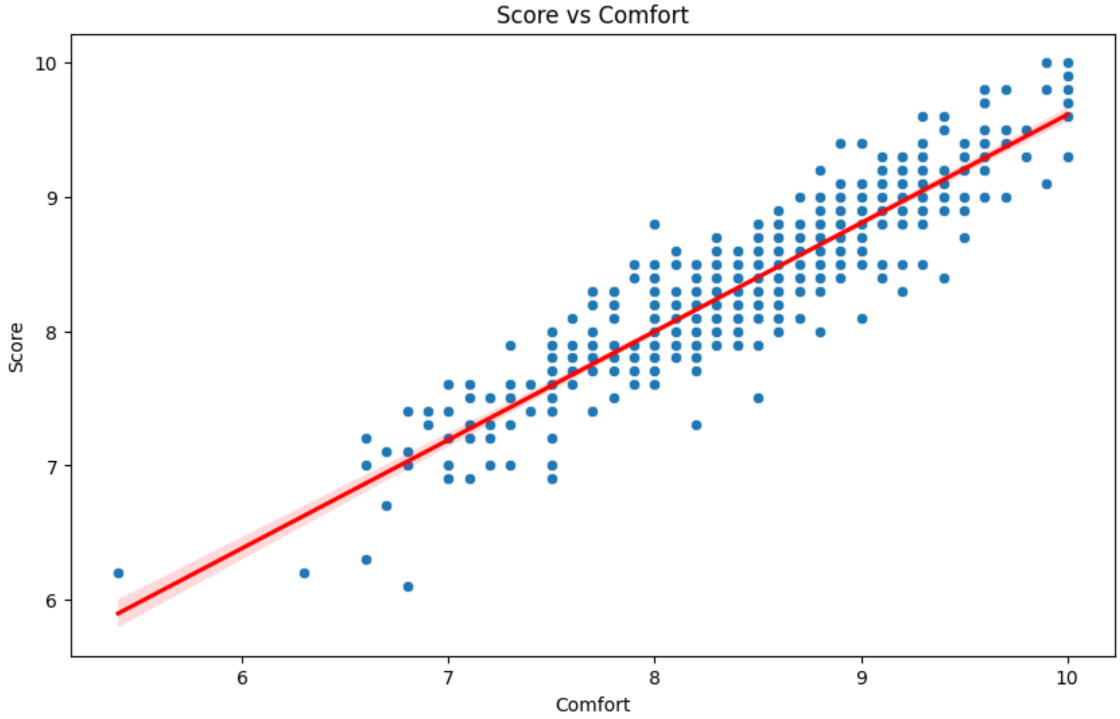
Observation: This scatter plot depicts the association between Cleanliness and the overall score of hotels in the dataset. Each blue dot represents a single hotel review score, with the cleanliness rating on the x-axis and the total Score on the y-axis. The trend line in red illustrates the regression line and how the variable cleanliness influences the overall score of hotels. The scatter plot shows that there is a strong positive link between the variable cleanliness and the overall score, which influences customer happiness. Based on this scatter plot, I conclude that when cleanliness grows, so does the overall Score, reinforcing the favorable association between these factors. This insight suggests that cleanliness plays a significant contribution in increasing customer positive experience within the hospitality industry.

Scatter Plot 2 - Score vs Facilities.



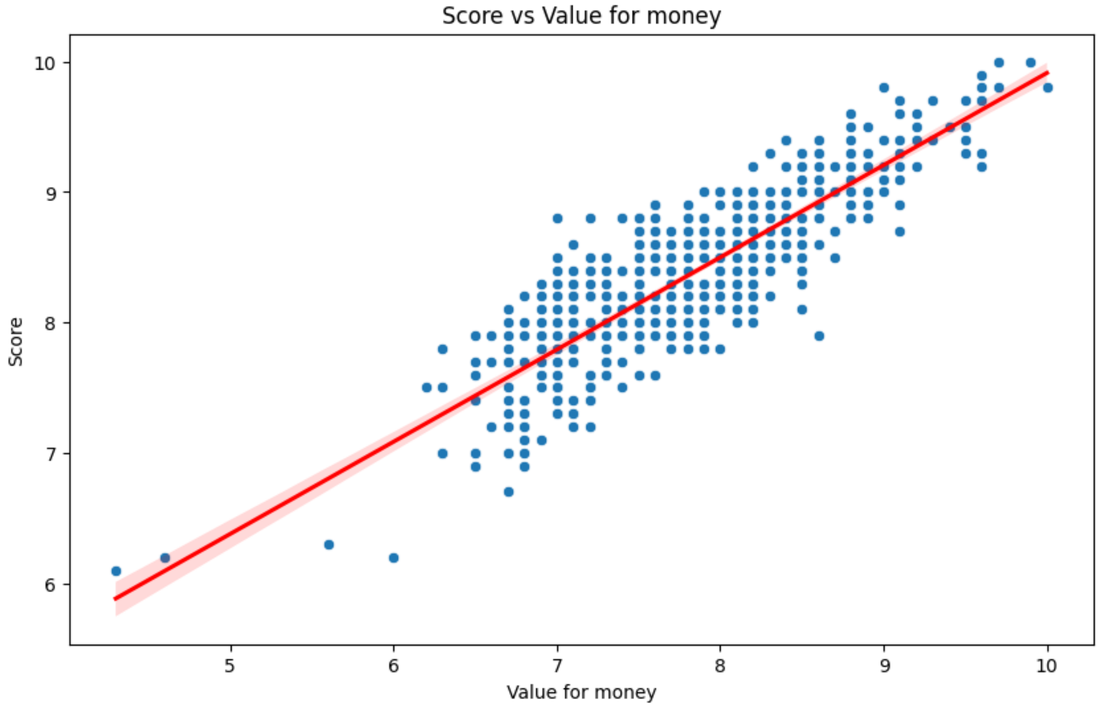
Observation: The following scatter plot depicts the relationship between the variable "Facilities" and the overall score assigned by customers to hotels in the dataset. This scatter plot revealed a strong positive association between facility quality and overall score. The red regression line demonstrates that as the quality of the facilities improves, so does the score, increasing the positivity of these factors. I discovered that the majority of blue dots in ratings range from 7 to 9, indicating that many Los Angeles hotels offer good facilities. Furthermore, this scatter plot implies that having improved facilities will always be perceived as a positive impact on a customer journey at a hotel, as evidenced by the trending line. It is important that the hospitality industry invest in good quality facility for customer to enjoy their stay and leaves with a positive impression as it predicts better overall Score in reviews. These findings provide valuable insights for hotel management and policymakers focused on improving customer satisfaction.

Scatter Plot 3 - Score vs Comfort.



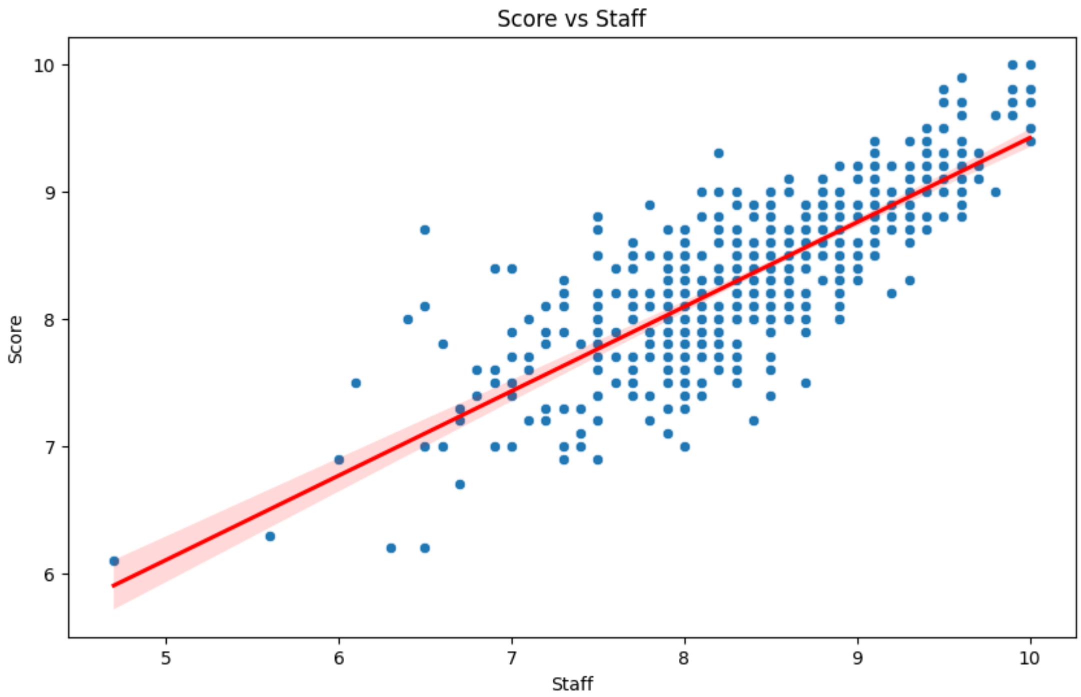
Observation: This scatter plot depicts the association between the variable comfort and the overall score provided by hotel clients. This scatter plot illustrates that when Comfort increases, so does the total score, indicating that a higher comfort experience correlates with higher overall client satisfaction. This graph shows that several blue dots cluster between 8 and 9 on the x-axis, indicating that hotels in the Los Angeles area provide high-quality comfort according to customer reviews. Although, there are certain variabilities for lower Comfort rating, this scatter plot gives insurance in the positive and strong association between Comfort and total Score. It emphasizes that comfort has a substantial impact on consumer satisfaction.

Scatter Plot 4 - Score vs Value for Money.



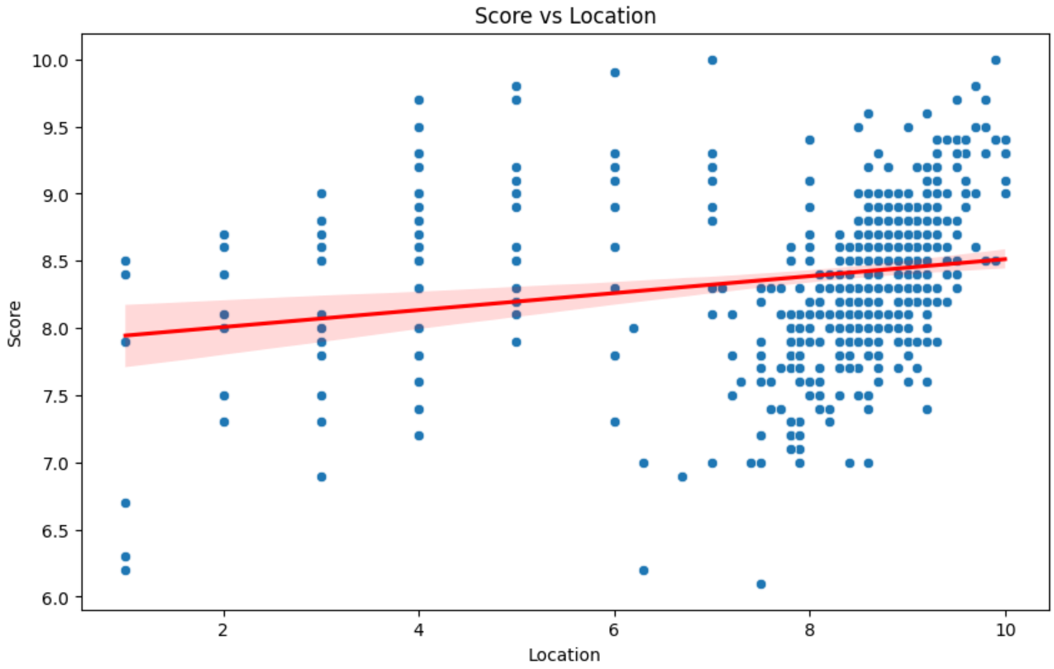
Observation: The following scatter plot depicts the link between the Score and Value for Money variables. It demonstrates that when value for money improves, so does customer satisfaction. It implies that consumers who perceive greater value for their money have a better hotel experience, and hence receive a higher review rating. This scatter plot shows that data points are primarily clustered toward the upper end of the x-axis, indicating that hotels in the Los Angeles area are judged to be good value. The graphs exhibit a linear and positive trendline, indicating a positive association between Value for Money and the overall score provided to hotels by clients. The role for the variable Value for Money plays a significant part when it comes to rate hotels by customers. It is important that experts in the hospitality industry acknowledge these findings to enhance customers' experience. The scatter plot shows that when a hotel gets a lower score when it comes to Value for money the overall score tends to be lower as well.

Scatter Plot 5 - Score vs Staff.



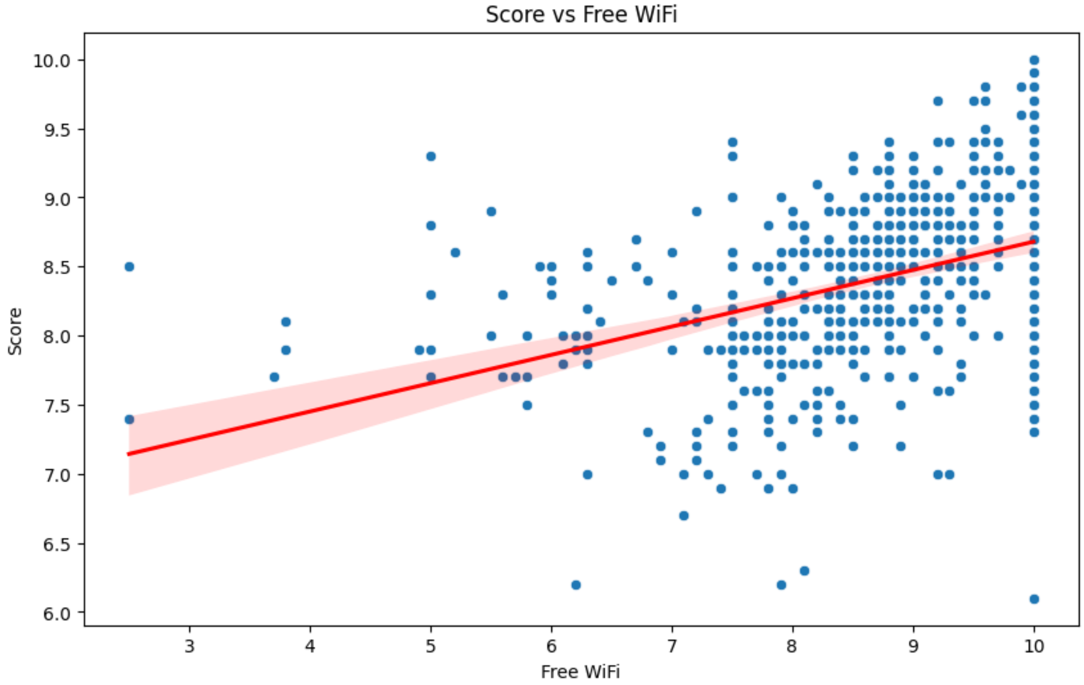
Observation: The following scatter plot depicts the association between Staff and total Score in hotels. The scatter plot reveals a linear, positive association between the two variables. Each blue dot represents a specific hotel, with staff ratings on the x-axis and overall scores on the y-axis. The regression line in this plot has a significant upward slope, indicating that higher staff ratings are generally associated with a higher overall Score. Based on this graph, I concluded that having pleasant and rigorous staff is critical, particularly in the hospitality industry. When a consumer perceives exceptional staff service, it inevitably leads to high customer satisfaction.

Scatter Plot 6 - Score vs Location.



Observation: The graph above illustrates how the location variable is linked to consumer ratings. It seems that there is a connection, between the two with higher location ratings often corresponding to higher overall scores. The data points are quite spread out for location ratings but they become more clustered for higher ratings (between 8 and 10 on the scale). This analysis hints that while location does play a role in scores its impact is relatively minor compared to factors, like comfort or facilities. The widespread distribution of points suggests that various other elements also influence guests satisfaction levels during their hotel stays.

Scatter Plot 7 - Score vs Free WiFI.



Observation: The scatter plot displays how the overall score relates to the rating, for WiFi in hotels. Each blue dot on the graph represents a hotel with its Free WiFi rating plotted on the x overall score on the y axis. The red line shows the trend of the data through regression.The plot indicates a link between WiFi ratings and total score although it is not as strong as correlations seen in categories like Comfort or Facilities. The slope of the regression line suggests that higher Free WiFi ratings often correspond to scores. However there is variation in data points around the regression line at higher Free WiFi ratings indicating that overall scores can differ significantly regardless of WiFi quality. This analysis suggests that while Free WiFi ratings impact scores to some extent other factors likely have a significant influence, on hotel satisfaction.

Regression Analysis

These regression models are created to forecast the rating based on elements, like cleanliness, comfort, facilities, staff service, value for money complimentary WiFi and location. By using a variety of models I aim to uncover the impact of these factors on the ratings and gain insights into the hospitality sector. I will gradually introduce these variables in models to assess their importance and influence on the outcome.

Model 1

From the first regression model, $\beta 0$ represents the intercept indicating the base level of score when $\beta 1$ and $\beta 2$ equal to zero. In this model the first two variables that I will evaluate are Cleanliness and Comfort. The error u term indicates the presence of unobserved factors that might impact the model.

$$Score = \beta 0 + \beta 1 + \beta 2 + \ u$$

Regression Output

Regression Statistics	
Multiple R	0.942
R Square	0.888
Adjusted R Square	0.888
Standard Error	0.206
Observations	582

	Coefficients	*Std. Error*	*t Stat*	*P-value*
Intercept	1.07	0.11	9.92	0.00
Cleanliness	0.33	0.02	15.54	0.00
Comfort	0.53	0.02	25.06	0.00

The first model has an r-square of 0.888, implying that the incorporated independent variables account for approximately 88.8% changes in the dependent variable and the model. The adjusted r-square value is still at 0.888 indicating that the model is good-fitted, and a rational explanation of the dependent variable can be drawn from all these independent variables together. The intercept term is equal to 1.075 with p-value lower than 0.05 implying that it has statistical significance. Basically, this means that when all other variables are zero, there will still be a base case value of 1.075. In other words, if everything else was held constant the initial score would have been 1. 075.

The coefficient for Cleanliness among explanatory variables is equal to .326 and its p-value is less than .05 which suggests statistical significance. In other words, cleanliness affects overall score by less than one unit and given that all other factors remain constant. Similarly, Comfort has a coefficient of 0.535, and its p-value is less than 0.05, indicating statistical significance. However, if we were to increase the comfort by one unit then the overall score would increase by about half a point unit. This clearly shows how both cleanliness and comfort play a positive role in influencing this variable with greater emphasis on the comfort variable having a slightly larger effect than the variable cleanliness.

Model 2

The second model will accentuate the evaluation of variables and how it affects the dependent variable Score. In this model I will add three more variables such as Facilities, Staff and Value for money, taking into considerations the error term u regarding inexplicable factors that might influence the model.

$$Score = \beta 0 + \beta 1 \cdot Cleanliness + \beta 2 \cdot Comfort + \beta 3 \cdot Facilities + \beta 4 \cdot Staff + \beta 5 \cdot Value\ Money + u$$

Regression Output:

Regression Statistics	
Multiple R	0.988
R Square	0.976
Adjusted R Square	0.976
Standard Error	0.096
Observations	582

	Coefficients	*Standard Error*	*t Stat*	*P-value*
Intercept	0.635	0.053	11.969	0.000
Cleanliness	0.166	0.011	15.166	0.000
Comfort	0.206	0.014	14.728	0.000
Facilities	0.194	0.013	15.095	0.000
Staff	0.184	0.008	23.233	0.000
Value for money	0.182	0.009	19.484	0.000

The regression model shows a r-squared value of 0.976 meaning that 97.6% of the Score variation can be explained by the independent variables used. The Adjusted r-squared value also stands high at 0.976 indicating that the model fits well and the included variables together offer an explanation, for the Score. The intercept is at 0.635. Its p value being than 0.05 signifies its statistical significance implying that when all other variables are zero the baseline Score is 0.635. Notably Cleanliness has a coefficient of 0.166 with a p value below 0.05 suggesting that increasing Cleanliness by one unit corresponds to a 0.166 unit rise in the Score while keeping factors constant. Similarly, Comfort has a coefficient of 0.206 with a p value below 0.05; hence elevating Comfort by one unit leads to a corresponding increase of 0.206 units, in the Scores value without impacting other variables significantly.

The Staff has a coefficient of 0.184 and its p value being, below 0.05 shows that it is statistically important. This suggests that an increase of one unit in Staff corresponds to a 0.184 unit rise in the Score. Similarly, Value for Money has a coefficient of 0.182 and its p value being than 0.05 indicates its significance. This means that a one unit increase in Value for Money leads to a 0.182 unit increase in the Score. The model emphasizes the influence of Cleanliness, Comfort, Facilities, Staff and Value for Money on the Score with each factor making a substantial and positive contribution, to the overall rating.

Model 3

In the third model β0 stands for the starting point while β1 to β7 represent how each factor affects the rating score. β1 to β7 I will evaluate the variables Cleanliness, Comfort, Facilities, Staff, Value for money, Free Wifi and Location respectively. The error term ε takes into account any variation not explained by these factors. This third model will provide a more thorough insights, into how aspects of hotel services and amenities contribute to customer satisfaction and overall ratings.

$$Score = \beta 0 + \beta 1 \cdot Cleanliness + \beta 2 \cdot Comfort + \beta 3 \cdot Facilities + \beta 4 \cdot Staff + \beta 5 \cdot Value\ for\ money + \beta 6 \cdot Free\ WiFi + \beta 7 \cdot Location + u$$

Regression Output:

Regression Statistics	
Multiple R	0.988
R Square	0.977
Adjusted R Square	0.977
Standard Error	0.094
Observations	582

	Coefficients	*Standard Error*	*t Stat*	*P-value*
Intercept	0.563	0.054	10.512	0.000
Cleanliness	0.165	0.011	15.360	0.000
Comfort	0.205	0.014	14.989	0.000
Facilities	0.191	0.013	15.189	0.000
Staff	0.180	0.008	22.668	0.000
Value money	0.177	0.009	18.887	0.000
Free WiFi	0.014	0.004	3.960	0.000
Location	0.008	0.002	3.900	0.000

The regression model shows a power with an r-squared value of 0.977 suggesting that almost 98% of the variation, in hotel scores can be accounted for by the independent variables included. The Adjusted r-squared value remains stable at 0.977 indicating a fitted model where the independent variables collectively offer an explanation for the dependent variable.

Analyzing the coefficients of the model provides insights into how different factors influence hotel scores. The intercept stands at 0.56 with a p value of 0.00 signaling its significance. This means that when all other variables are zero the baseline score is set at 0.56.

Looking at variables Cleanliness has a coefficient of 0.16 (p value = 0.00) indicating that increasing Cleanliness by one unit is linked to a 0.16 unit rise in Score while holding other factors constant. Similarly Comfort has a coefficient of 0.20 (p value = 0.00) showing that boosting Comfort by one unit leads to a Score increase of 0.20 units. Considering Facilities with a coefficient of 0.19 (p value = 0) it suggests that improved facilities add around 0.19 units, to the Score.

Staff services also play a role with a coefficient of 0.18 (p value = 0.00) indicating that enhanced staff services contribute 0.18 units to the score. Value, for Money shows a coefficient of 0.18 (p value = 0.00) suggesting that value for money results in a 0.18 unit increase in the score by unit added.

Although Free WiFi and Location are significant, they have impacts on the score. Free WiFi has a coefficient of 0.01 (p value = 0.00) demonstrating a noteworthy impact on the score. Similarly, Location also has a coefficient of 0.01 (p value = 0.00) indicating a significant relationship between location and score.

In essence this regression model underscores the influence of Cleanliness, Comfort, Facilities, Staff and Value for Money on hotel scores as each factor contributes positively and significantly to the score. While Free WiFi and Location are also factors their impact is less pronounced, than other variables outlined in the model analysis provided valuable insights into key factors that boost customer satisfaction and overall hotel ratings.

Models Table Output:

	Model (1)	**Model (2)**	**Model (3)**
	Score	Score	Score
Attributes			
Cleanliness	0.325***	0.165***	0.165***
Comfort	0.534***	0.206***	0.205***
Facilities		0.194***	0.192***
Staff		0.184***	0.181***
Value money		0.182***	0.177***
Free WiFi			0.014***
Location			0.008***
Constant	1.076***	0.635***	0.563***
Observations	582	582	582
R-squared	**0.888**	**0.976**	**0.977**

The development of the models shows an enhancement, in their ability to explain, as seen in the rising r-squared values. Initially Model 1 with an r-squared value of 0.888 explains a part of the Scores variance. However by adding factors like Facilities, Staff, Value for Money, Free WiFi and Location Models 2 and 3 exhibit progress with r-squared values of 0.976 and 0.977 respectively. This indicates that Model 3 which includes the range of attributes offers the most accurate prediction, for hotel guests overall satisfaction scores.

Conclusion

This research study lends an in-depth review of the relationship between hotel ratings and various important factors, including Cleanliness, Comfort, Facilities, Staff, Value for Money, Free WiFi, and Location. In this paper, I use regression analysis through different models to identify the roles of variables contributing to customer satisfaction and, as a whole, hotel rating. The results obtained indicate that Cleanliness, Comfort, Facilities, Staff, and Value for Money do have an influence on the hotel scores, each statistically significant. While factors such as Free

WiFi and Location are important, their effects are relatively limited compared with that of the factors. With a high r-squared value and the adjusted r-squared in our models, we can anticipate that the regression method will work fine. Such insights will enable the hotel manager and the stakeholders to prioritize their investment in enhancing areas of customer satisfaction in order to yield better ratings overall.

Further research can do more in-depth investigation of other determinants of hotel scores, such as personalized services or sustainability practices. Longitudinal studies could also capture the variations overtime due to changes in clients' tastes or preferences arising out of the changing patterns or trends within the hospitality industry. This would be helpful in gaining insight, considering geographical locations and types of accommodation. Through this way, factors that influence hotel satisfaction can be understood. By getting into these aspects, upcoming studies can increase our understanding of how the hospitality sector functions so that more effective strategies are developed for enhancing customers' experience and contentment.

References:

Dataset, Kaggle.com

https://www.kaggle.com/datasets/thedevastator/sentiment-analyses-of-city-hotels

Nina, N. (2019, March 25). This Is Why Bill Gates Says "Your Most Unhappy Customers Are Your Greatest Source Of Learning". *Medium*.

https://medium.com/@ninathena9/this-is-why-bill-gates-says-your-most-unhappy-customers-are-your-greatest-source-of-learning-fbf3ef93e6e9

Kattara, H. S., Weheba, D., & El-Said, O. A. (2008). The impact of employee behaviour on customers' service quality perceptions and overall satisfaction. Tourism and Hospitality Research, 8(4), 309–323.

http://www.jstor.org/stable/23745454

AltexSoft. (2021, March 11). Sentiment analysis in hotel reviews: Developing a decision-making assistant for travelers.

https://www.altexsoft.com/blog/sentiment-analysis-hotel-reviews/
Lexalytics. (n.d.). *Customer feedback analytics in hospitality: Full guide"*
https://www.lexalytics.com/blog/customer-feedback-analytics-guide-hospitality/