**RECOMMENDATIONS ON IMPROVING THE HOTEL CAPACITY**

Contents

[**1.** **Executive summary** 2](#_Toc130132855)

[**2.** **Introduction** 3](#_Toc130132856)

[**3.** **Data Structure and Understanding** 4](#_Toc130132857)

[3.1. Structure of the data 4](#_Toc130132858)

[3.2. Data cleaning and Transformation 4](#_Toc130132859)

[*Figure 1: Data fields, types and quality* 5](#_Toc130132860)

[3.3. Descriptive analysis 5](#_Toc130132861)

[*Figure 2: Number of hotel reviews by Score and Traveler type.* 6](#_Toc130132862)

[*Figure 3: Average Score by User continent and Period of stay.* 7](#_Toc130132863)

[*Table 1: Descriptive statistics of Score factor.* 8](#_Toc130132864)

[*Figure 4: Hotel score distribution by user reviews.* 9](#_Toc130132865)

[**4.** **Predictive Analytics** 9](#_Toc130132866)

[4.1. Multiple linear regression model 9](#_Toc130132867)

[*Figure 5: Summary of multiple linear regression model 1.* 10](#_Toc130132868)

[*Figure 6: Summary of multiple linear regression model 2.* 11](#_Toc130132869)

[*Figure 7: Summary of multiple linear regression model 3.* 12](#_Toc130132870)

[*Figure 8: Analysis of Variance table.* 13](#_Toc130132871)

[**5.** **Prescriptive Analytics** 13](#_Toc130132872)

[**6.** **Conclusion and Recommendation** 15](#_Toc130132873)

[**References** 16](#_Toc130132874)

[**Appendix** 16](#_Toc130132875)

[Appendix a: Measures of Dispersion and Central Tendency for Score 16](#_Toc130132876)

# **Executive summary**

This report offers a thorough examination of Chrysalis, a four-star hotel in downtown Las Vegas, with an emphasis on descriptive, predictive, and prescriptive analysis. The study contains a descriptive analysis of Chrysalis' current condition, including occupancy rates, room rates, profitability, and infrastructure condition.

The paper also includes a predictive analysis that investigates the possible impact of renovating the hotel and casino, including the addition of extra facilities such as an infinity pool, gym renovation, spa, and tennis court.

Lastly, the report includes a prescriptive analysis, outlining specific recommendations for Chrysalis to improve its customer experience, increase occupancy rates, and generate more revenue. These include renovating the hotel, adding new facilities, and replacing slot machines in the casino. Additionally, the report recommends using TripAdvisor reviews to understand whether the current facilities meet customer demands and whether new facilities are necessary.

# **Introduction**

The hospitality industry is one of the most dynamic and competitive industries worldwide. The emergence of online travel agencies and review websites has made it easier for travelers to find and book their preferred accommodations. TripAdvisor is one of the most popular travel review websites, providing millions of reviews from travelers worldwide. This report provides a detailed analysis of the data obtained from TripAdvisor on the Chrysalis hotel, a 4-star hotel located in downtown Las Vegas. The data includes information such as user country, number of reviews, helpful votes, traveler style, and hotel facilities, among others. The aim of this report is to provide descriptive, predictive, and prescriptive analytics to support informed decision-making regarding the hotel's potential renovation. The report begins by presenting a structured overview of the data, including its sources and variables. The descriptive analysis of the data includes mean, median, mode, standard deviation, variance, and range of the Score variable, which provides insights into the satisfaction level of guests staying at the hotel. The predictive analysis of the data utilizes regression models to predict the occupancy rate of the hotel under different renovation scenarios. The prescriptive analysis involves examining the costs involved in renovating both the hotel and casino buildings, along with analyzing TripAdvisor reviewers' opinions to identify the need for new facilities. The report concludes by providing recommendations to Kate and Joe, the owners of Chrysalis hotel, on whether to renovate both the hotel and casino buildings or just the hotel, and whether to add new facilities or refurbish the old ones to meet guests' demands. The results of the analysis provide valuable insights that can help the owners make informed decisions that will increase their profits and improve their guests' experiences.

# **Data Structure and Understanding**

* 1. Structure of the data

The dataset used in this analysis contains 505 rows of hotel review information collected from TripAdvisor. It includes diverse variables such as user country, hotel amenities, review score, and hotel details. The data is primarily categorical and comprises both nominal and ordinal variables. Its richness and diversity make it ideal for analyzing global hotel trends and identifying factors impacting customer experience. The dataset provides a comprehensive view of hotel reviews on TripAdvisor and is a valuable resource for conducting research in the hospitality industry.

* 1. Data cleaning and Transformation

The quality of data has a significant impact on the accuracy and validity of research outcomes. Hence, it is essential to evaluate the quality of data using multiple dimensions, including completeness. In this study, completeness was identified as one of the four dimensions to assess the quality of the data. Upon analysis of the dataset, it was found that the columns related to Nr. rooms, User continent, Member years, Review month, and Review weekday had insufficient data, with 96 blank values accounting for 19% of the completeness compared to other data columns. Additionally, to ensure the accuracy of the data, the words "New Zeland" and "Phillippines" were corrected to "New Zealand" and "Philippines," respectively, in the User country column. Further, it was identified that a data error existed in row 77, where the value in the Member years column was presented as -1806, which was not a reasonable number. In light of these findings, Nr. rooms, Member years, Review month, and Review weekday columns were eliminated from the study as these data fields did not contribute to the final findings desired by the hotel owners. Finally, to enhance the completeness of the User continent column, the VLOOKUP function was utilized on Excel to fill in the remaining blank values for further analysis. Through these measures, the data quality and completeness were improved, ensuring the reliability and validity of the subsequent descriptive, predictive, and prescriptive analyses conducted in this study.

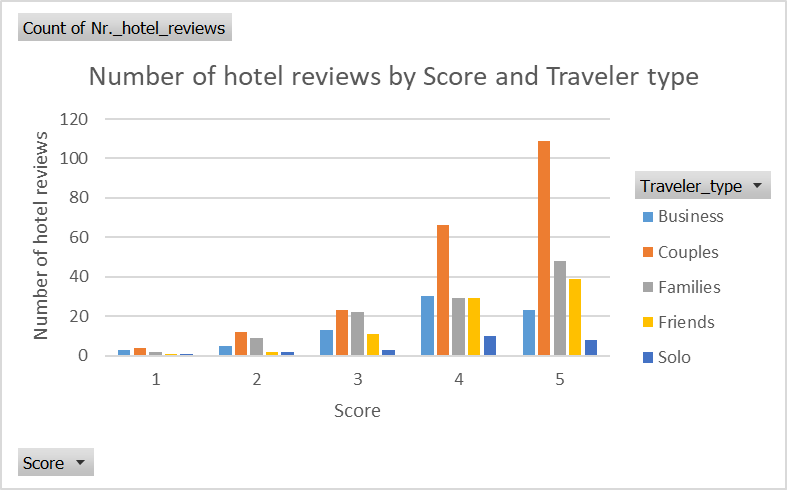
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Data** | **Data type** | **Number of values** | **Number of blank values** | **Completeness** |
| User country | Nominal | 504 | 0 | 100% |
| Nr. reviews | Ordinal | 504 | 0 | 100% |
| Nr. hotel reviews | Ordinal | 504 | 0 | 100% |
| Helpful votes | Ratio | 504 | 0 | 100% |
| Score | Ratio | 504 | 0 | 100% |
| Period of stay | Ratio | 504 | 0 | 100% |
| Traveler type | Nominal | 504 | 0 | 100% |
| Pool | Nominal | 504 | 0 | 100% |
| Gym | Nominal | 504 | 0 | 100% |
| Tennis court | Nominal | 504 | 0 | 100% |
| Spa | Nominal | 504 | 0 | 100% |
| Casino | Nominal | 504 | 0 | 100% |
| Free internet | Nominal | 504 | 0 | 100% |
| Hotel name | Nominal | 504 | 0 | 100% |
| Hotel stars | Ratio | 504 | 0 | 100% |
| Nr. rooms\* | Ratio | 408 | 96 | 81% |
| User continent | Nominal | 408 | 96 | 81% |
| Member years\* | Ratio | 408 | 96 | 81% |
| Review month\* | Nominal | 408 | 96 | 81% |
| Review weekday\* | Nominal | 408 | 96 | 81% |

*Figure 1: Data fields, types and quality*

\*: Data fields eliminated from the analysis after the data processing stage.

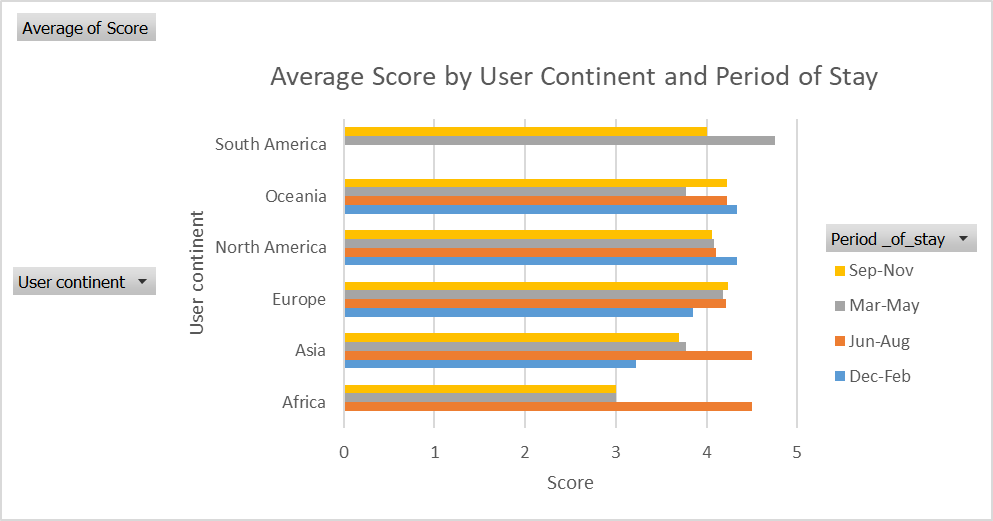
* 1. Descriptive analysis

Figure 2 provides an in-depth explanation of the number of reviews received for each possible combination of traveler type (Business, Couples, Families, Friends, Solo) and score (1-5), thereby facilitating a comparison of review counts across different traveler types and score levels. By analyzing this graph, valuable insights can be gleaned about the experiences of various traveler types and their satisfaction levels with the hotel. For example, this analysis can help to identify whether certain traveler types are more likely to give high or low scores, or whether certain scores are more prevalent among all traveler types. Overall, this figure provides a comprehensive overview of the distribution of reviews and can aid in understanding the dynamics of guest experiences at the hotel.



*Figure 2: Number of hotel reviews by Score and Traveler type.*

Figure 3 depicts the mean score for each possible combination of user continent and period of stay. This visual representation is a useful tool for gaining insights into the satisfaction levels of customers from diverse geographical regions during different times of the year. In particular, it can help to identify if guests from particular continents are more inclined to assign higher or lower scores overall, or if there are specific periods of the year that are associated with elevated or reduced satisfaction levels. Hence, this figure provides a nuanced understanding of guest satisfaction and can be utilised to develop tailored strategies for enhancing the guest experience at the hotel.



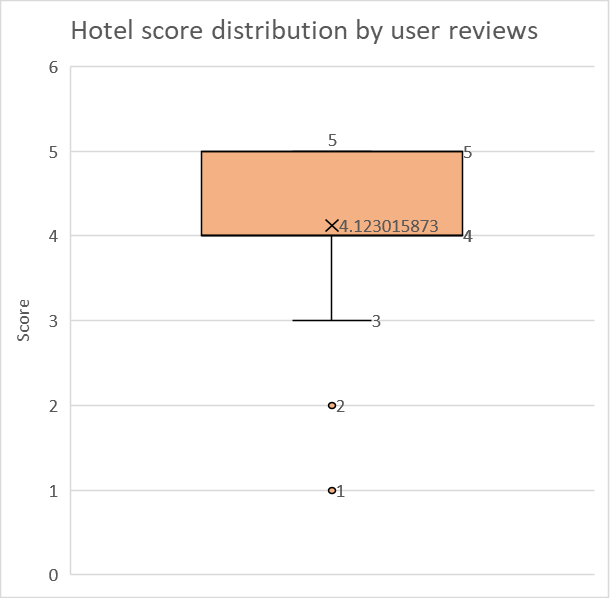
*Figure 3: Average Score by User continent and Period of stay.*

Table 1 presents a mean score of 4.12, suggesting that, on average, reviewers provide a positive score for hotels. Besides, the median score of 4.00, which is similar to the mean score, indicates that the distribution of scores is relatively symmetric. Additionally, the mode score of 5.00 suggests that a substantial number of reviewers assign a perfect rating to hotels. The standard deviation of 1.01 indicates that the scores are moderately dispersed around the mean, with a range of scores of 4.00, implying that the minimum and maximum scores given by reviewers were 1.00 and 5.00, respectively. Moreover, the standard error of 0.04 and the confidence level of 95.0% furnish information about the precision of the sample mean. The standard error denotes the expected error in the sample mean due to random variation in the data, while the confidence level represents the likelihood that the true population mean lies within a specific range of the sample mean. In this instance, the 95.0% confidence level implies that there is a 95% probability that the true population mean falls within plus or minus 0.09 of the sample mean of 4.12.

|  |  |
| --- | --- |
|  | **Score** |
| **Mean** | 4.12 |
| **Standard Error** | 0.04 |
| **Median** | 4.00 |
| **Mode** | 5.00 |
| **Standard Deviation** | 1.01 |
| **Sample Variance** | 1.01 |
| **Range** | 4.00 |
| **Minimum** | 1.00 |
| **Maximum** | 5.00 |
| **Sum** | 2078.00 |
| **Count** | 504.00 |
| **Largest(1)** | 5.00 |
| **Smallest(1)** | 1.00 |
| **Confidence Level(95.0%)** | 0.09 |

*Table 1: Descriptive statistics of Score factor.*

Figure 4 below further describes the distribution of Score rated on TripAdvisor by customers. In the box plot, two outliers were identified: 1 and 2 (<IQR\*1.5 below Q1). The IQR of the Score data is 1, which means that the middle 50% of the scores falls within a range of 1 between 4 and 5.



*Figure 4: Hotel score distribution by user reviews.*

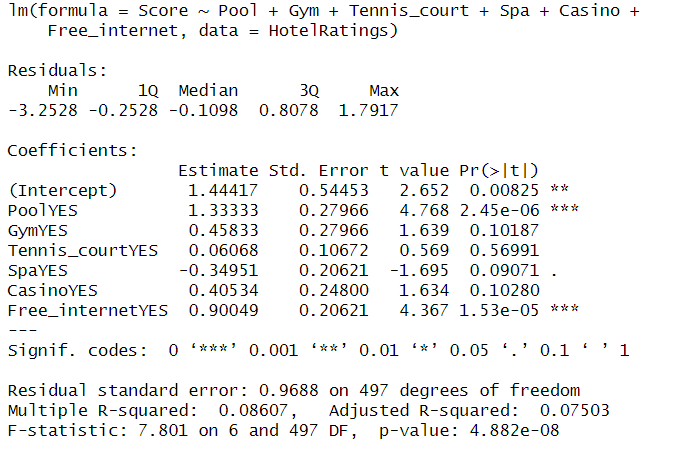
# **Predictive Analytics**

* 1. Multiple linear regression model

In this sector, score is elected as a dependent variable of two multiple regression models below because it is the ultimate opinion from the reviews about the hotels that they stayed. Characteristics such as Pool, Gym, Tennis court, Spa, Casino, Free internet are considered being included in this model since those are facilities affecting the reason for customers scoring the hotel according to the data given. Indeed, there would be more elements influencing customer’s opinions about the hotel that the data could not show.

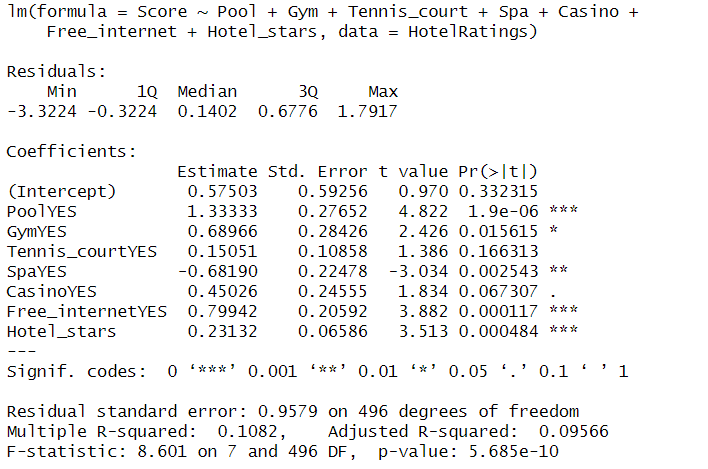
The lower the p-value, the stronger the evidence against the null hypothesis and the greater the significance of the independent variable in explaining variance in the dependent variable (Score). The adjusted R-squared is a variation of the R-squared that takes the number of independent variables included in the model into account (Tranmer and Elliot, 2008). The proportion of variance in the dependent variable explained by the independent variables is represented by R-squared. According to the model (Figure 5), The adjusted R-squared value of 0.07503 indicates that the model's independent variables only explain 7.5% of the total variation in the dependent variable.

The F-statistic is another test for the model's overall significance. A significant F-statistic indicates that at least one of the independent variables is related to the dependent variable in a substantial way (Tranmer and Elliot, 2008). From the below model, with a p-value of 4.882e-08, the F-statistic indicates that the model is statistically significant overall.



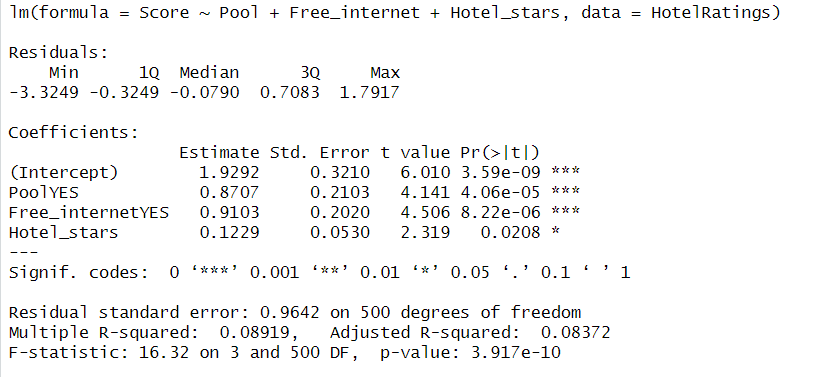
*Figure 5: Summary of multiple linear regression model 1.*

In the second model, Hotel stars was added in considering that the more the hotel stars, the better quality of the facilities which might alter guests’ viewpoints. According to Figure 6, except for one extra predictor variable (Hotel stars), the interpretation of the p-values and adjusted R-squared is similar to the first model. Apart for Tennis court and Casino, all predictor variables in this model are statistically significant. This model's adjusted R-squared value is 0.09566, which is slightly higher than the first model but still indicates that the model is not a good fit for the data. Overall, these findings indicate that the set of predictor factors in the models is not very good at predicting the outcome variable (Score), and that more variables may need to be added or the models updated in some other way.



*Figure 6: Summary of multiple linear regression model 2.*

From the two above model, it can be seen that Gym, Tennis court, Spa and Casino, which possess p-values greater than 0.05, are not statistically significant at the 95% confidence level in predicting the reviewers’ score and can be removed from the model to improve accuracy. Thus, Figure 7 represents three predictors: Pool, Free\_internet, and Hotel\_stars. The coefficients for Pool and Free\_internet are both positive and statistically significant, indicating that hotels with these amenities tend to have higher ratings than those without. The coefficient for Hotel\_stars is also positive and statistically significant, suggesting that hotels with higher star ratings tend to have higher ratings from customers. The adjusted R-squared is slightly lower, indicating that the inclusion of these predictors did not substantially improve the fit of the model compared to a model with just the intercept. The F-statistic for the model is statistically significant, indicating that at least one of the predictors has a significant effect on the outcome. As a result, while the model suggests that having a pool, free internet, and higher star ratings are associated with higher ratings from customers, the model does not explain much of the variance in ratings and there may be other factors that are important in predicting hotel ratings.

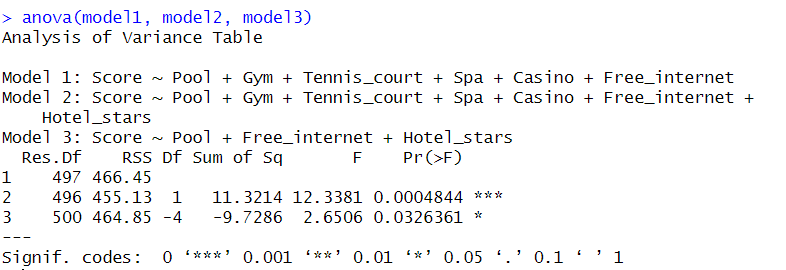


*Figure 7: Summary of multiple linear regression model 3.*

* 1. Linear regression model evaluation

In order to identify the most suitable regression model for the business, an ANOVA test was conducted in RStudio in order to identify which model is the most significant among the three. The analysis of variance table portrays the results of comparing the three models using a significance level of 0.05.

Drawing on the analysis of variance table from Figure 8, it is apparent that Model 2 is the optimal fit, given its possession of the minimum residual sum of squares (RSS) and the maximum F-statistic in contrast to the other models. Moreover, the F-statistic's p-value attains statistical significance at a significance level of 0.05, signifying that, at the very least, one of the coefficients in Model 2 holds statistical significance in expounding the variance in the response variable (Rouder el at., 2016; Ferreira el at., 2014). Consequently, one may deduce that Model 2 yields a superior fit compared to both Model 1 and Model 3.



*Figure 8: Analysis of Variance table.*

In order to ascertain the critical amenities offered by a hotel, it is necessary to analyse the best fitting regression model, which is Model 2. In Model 2, "Pool" and "Free internet" have emerged as the features with the lowest p-values, underscoring their robust and statistically significant association with the dependent variable, "Score." A smaller p-value implies a stronger refutation of the null hypothesis, which posits the absence of a substantial connection between the attribute and the target variable, thus indicating greater significance. As a result, these two characteristics are considered to be the most pivotal predictors of "Score" within this model.

# **Prescriptive Analytics**

* 1. Decision Tree

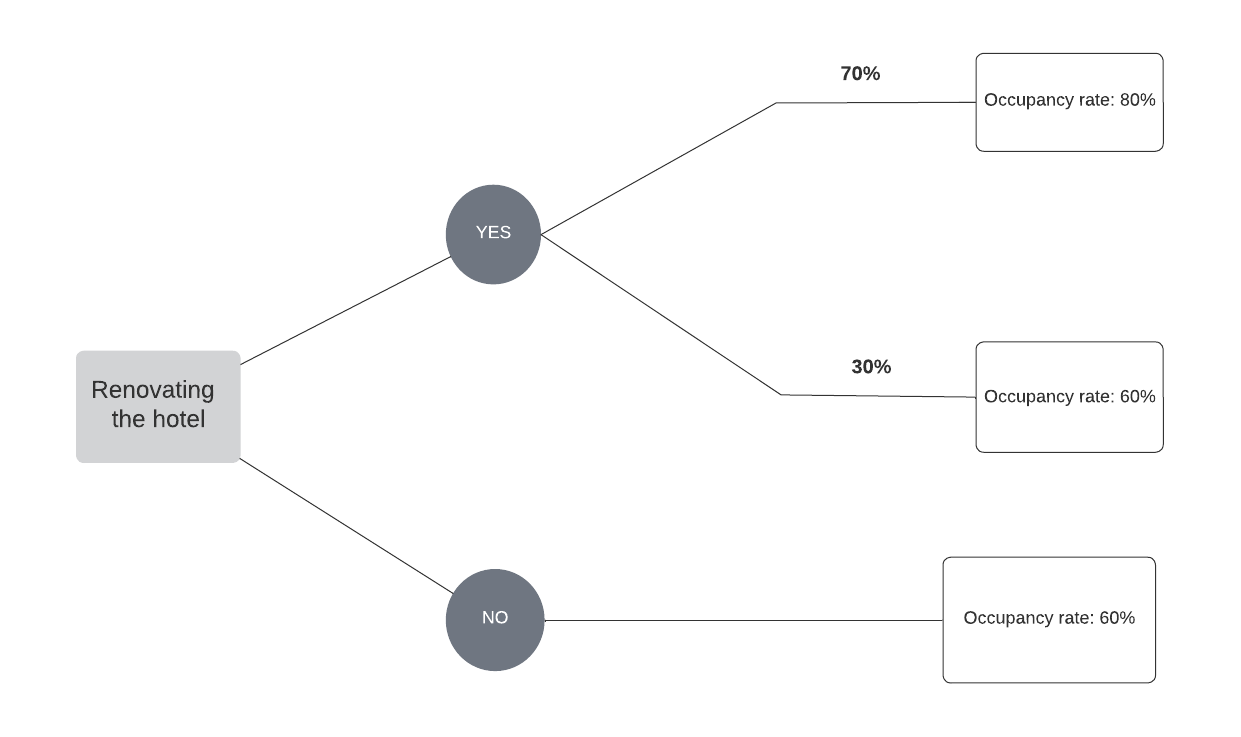
According to the outcomes of the predictive analytics analysis conducted earlier, pool and free internet were determined as significant variables that the hotel owners should take into consideration for renovation decision to meet customer expectations. Additionally, it is recommended that 20 slot machines in the casino be substituted. As a result, the criteria for making the decision are delineated as follows:

Cost:

* Renovation (hotel only): $10 million
* Infinity pool: $220,000
* Free internet: $20,000
* Slot machine (each): $25,000

Annual profit when:

* The occupancy rate is 80%: 1000 x 0.8 x $60 x 365 = $17,520,000
* The occupancy rate is 60%: 1000 x 0.6 x $60 x 365 = $13,140,000



From the above decision tree, the cost and revenue can be calculated as below:

Scenario 1: If the hotel is renovated:

* Cost: (20 x $25,000) + $220,000 + $20,000 + $10 million = $10,740,000
* Annual revenue: (0.7 x $17,520,000) + (0.3 x $13,140,000) = $16,206,000
* Annual profit: $16,206,000 - $10,740,000 + $600,000 (casino’s yearly profit) = $6,066,000

Scenario 2: If the hotel is not renovated:

* Cost: (20 x $25,000) + $220,000 + $20,000 = $740,000
* Revenue: 1000 x 0.6 x $60 x 365 = $13,140,000
* Profit: $13,140,000 - $740,000 + $600,000 (casino’s yearly profit) = $13,000,000

Thus, it can be seen that the profit resulting from not renovating the hotel is greater, exceeding twice the profit gained from renovating the hotel.

# **Conclusion and Recommendation**

According to the previous analysis, while it is advisable for Kate and Joe to consider renovating only the infinity pool, free internet, and replacing 20 slot machines, it is uncertain whether these changes will yield a higher return on investment.

Within the domain of statistical inference, more precisely in the area of predictive inference, a prediction interval can be defined as an approximation of a specific range wherein an upcoming observation is anticipated to lie, based on the available data, with a certain level of probability (Phillips, 2017). From the best regression model (Model 2), a forecast has been made (find: RStudio script) to determine the probable average rating (Score) that a hotel possessing particular attributes (such as a pool, gym, tennis court, spa, casino, free internet, and a 4-star rating) could garner. The resulting projected Score is 4.24, and there exists a 95% probability that the actual Score of such a hotel would reside within the interval of 2.35 and 6.13.

To achieve a comprehensive understanding of the factors that influence customer satisfaction in hotels, it is essential to consider factors beyond just the available amenities. Customer service, cleanliness, food quality, room comfort are some other factors that play also crucial roles in shaping customers' experiences (Li et al., 2013). Therefore, the p-values acquired from three models does not have significant and noticeable figures. To provide accurate and informed recommendations, it is important to conduct a thorough analysis using various sources of information beyond just TripAdvisor. Hotel owners can seek out additional sources such as Google and social media reviews to obtain a more precise and comprehensive picture of customer satisfaction.

In general, to optimize their business outcomes, hotel owners need to adopt a holistic approach towards improving their services. While the renovation of hotel facilities is undoubtedly important, it is not the sole determinant of customer satisfaction; hence, hotel owners should also pay attention to other aspects as mentioned above. By offering exceptional services and amenities that cater to the diverse needs and preferences of their customers, hotel owners can create a unique and memorable experience that not only attracts new customers but also cultivates customer loyalty. Additionally, a focus on customer satisfaction can lead to positive word-of-mouth marketing, which can have a significant impact on the success of the business in the long run. Ultimately, by focusing on these key elements, hotel owners can create a competitive advantage for their business, ensuring a sustainable and profitable future (Kandampully and Suhartanto, 2000).

# **References**

Ferreira, E.B., Cavalcanti, P.P. and Nogueira, D.A., (2014). ExpDes: an R package for ANOVA and experimental designs. Applied Mathematics, 5(19), p.2952.

Kandampully, J. and Suhartanto, D., (2000). Customer loyalty in the hotel industry: the role of customer satisfaction and image. International journal of contemporary hospitality management.

Li, H., Ye, Q. and Law, R., (2013). Determinants of customer satisfaction in the hotel industry: An application of online review analysis. Asia Pacific Journal of Tourism Research, 18(7), pp.784-802.

Rouder, J.N., Engelhardt, C.R., McCabe, S. and Morey, R.D., (2016). Model comparison in ANOVA. Psychonomic bulletin & review, 23, pp.1779-1786.

Tranmer, M. and Elliot, M., (2008). Multiple linear regression. The Cathie Marsh Centre for Census and Survey Research (CCSR), 5(5), pp.1-5.

# **Appendix**

## Appendix a: Measures of Dispersion and Central Tendency for Score

|  |  |
| --- | --- |
| **Score Dispersion** | **Score** |
| Min | 1 |
| Q1 | 4 |
| Median | 4 |
| Mean | 4.12 |
| Q3 | 5 |
| Max | 5 |
| IQR | 1 |
| IQR\*1.5 | 1.5 |
| Upper limit | 6.5 |
| Lower limit | 2.5 |