****

**CONTENTS**

[**INTRODUCTION** 3](#_Toc200235454)

[**RESEARCH QUESTION & HYPOTHESIS** 3](#_Toc200235455)

[**DATA PROCESSING AND VARIABLE SELECTION** 3](#_Toc200235456)

[**THE VARIABLES** 3](#_Toc200235457)

[**STATISTICAL SUMMARY AND CORRELATION MATRIX** 4](#_Toc200235458)

[**HISTOGRAM:** 5](#_Toc200235459)

[**BAR CHART:** 6](#_Toc200235460)

[**CORRELATION MATRIX:** 6](#_Toc200235461)

[**MULTIVARIATE LINEAR REGRESSION** 7](#_Toc200235462)

[**SUB-SAMPLE ANALYSIS:** 8](#_Toc200235463)

[**EXPLANATION OF RESULT** 8](#_Toc200235464)

[**INTERPRETATION OF MODEL:** 9](#_Toc200235465)

[**SIGNIFICANCE OF PREDICTORS FOR EXPLAINING RESEARCH QUESTION:** 9](#_Toc200235466)

[**CONCLUSION** 9](#_Toc200235467)

[**MANAGERIAL IMPLICATIONS:** 9](#_Toc200235468)

**Data Analysis Assessment: Predicting Hotel Prices in European Cities**

**Submit to:** Mr. Khaled Eid

**Group Students:** Bhavik Goyal (8532229), Rohit Kude (8267479), Sharon Leon (8401640), Oliver Nosek (8805751), Amodini Kasturi (8219849)

# **INTRODUCTION**

The hospitality industry in Europe is considered to be a highly competitive, because of this it is significantly influenced by a wide range of factors, these include guest satisfaction, promotional offers, calendar events and many more. This analysis will be indicating how various booking conditions can influence hotel prices across different European cities. The dataset used in this study contains information on over 149,967 hotel stays across Europe, focusing on factors such as price, guest satisfaction reviews, promotional offers, and calendar events.

The main objective of this report is to identify any key drivers of price variation and offer insights to hotel managers on strategic pricing decisions.

# **RESEARCH QUESTION & HYPOTHESIS**

**How do guest satisfaction reviews, promotional offers, and calendar factors (weekend/holiday) affect hotel prices in European cities?**

**HYPOTHESIS:**

**Null Hypothesis (H0):** Guest satisfaction reviews, promotional offers, and calendar factors do not significantly affect the hotel prices.

**Alternative Hypothesis (H1):** Guest satisfaction reviews, promotional offers, and calendar factors have a significant effect on hotel prices.

# **DATA PROCESSING AND VARIABLE SELECTION**

The dataset used contains detailed information on over 149,967 single and multi-night hotel stays (4-night stay) in various cities across Europe. To ensure the consistency of the dataset, only the price for 1 night was selected for analysis. To eliminate the compounding effects of multi-night stays, the data with the prices of multi-night stays was dropped, since they were not in a significant amount, leaving only price for 1 night, which had also been removed due to it being a constant. Additionally, all missing values and all the insignificant columns in the dataset were also dropped. After cleaning, the dataset included detailed information on over 118,000 European hotel listings, focusing on factors like guest reviews, pricing, booking conditions, and offer levels.

## **THE VARIABLES:**

The variables used are:

**Dependent Variable (Y)**:

|  |  |  |  |
| --- | --- | --- | --- |
| Variable Name | Type | Definition | Measure |
| Price | Numeric | Listed price for a one-night stay | Continuous (in €) |

**Independent Variable (X):**

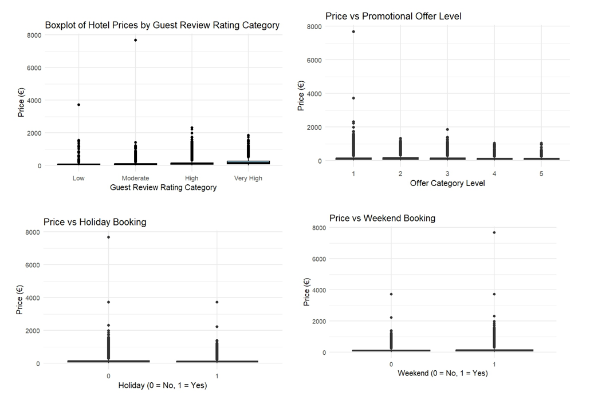
|  |  |  |  |
| --- | --- | --- | --- |
| Variable Name | Type | Definition | Measure |
| guestreviewratings | Numeric | Average guest review score | Continuous (Range: 1-5) |
| offer\_cat\_level | Numeric | Level of promotional offer applied | Integer scale: 1 (0%)to 5 (more than 75%) |
| weekend | Binary | Shows if the booking falls on a weekend | 0 = NO, 1 = YES |
| holiday | Binary | Shows if the booking falls on a holiday | 0 = NO, 1 = YES |

# **STATISTICAL SUMMARY AND CORRELATION MATRIX**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Variables | Mean | Median | SD | Min | Max | NOTE |
| price | 179.15 | 113 | 255.63 | 5 | 39996 | Continuous variable representing price in euros (€) |
| month | 6.87 | 6 | 4.18 | 1 | 12 | Categorical (1 = Jan, ..., 12 =Dec) |
| weekend | 0.66 | 1 | 0.47 | 0 | 1 | Binary (0 = No, 1 = Yes) |
| holiday | 0.22 | 0 | 0.41 | 0 | 1 | Binary (0 = No, 1 = Yes) |
| guestreviewsrating | 3.91 | 4 | 0.62 | 1 | 5 | Guest satisfaction score (scale: 1 to 5) |
| offer\_cat\_level | 2.12 | 2 | 1.1 | 1 | 5 | Promotional offer intensity (1 = none to 5 = 75%+ discount) |

Hotel prices average €179.15 (median €113, SD = €255.63), indicating strong variability and right-skew due to high-end listings (max = €3999). 66% of bookings occur on weekends, and 22% of these are on holidays, highlighting peak demand periods. Guest satisfaction is generally high, with an average rating of 3.91 (SD = 0.62) on a 5-point scale. The mean offer level is 2.12 (on a 1–5 scale), indicating that moderate discounts are commonly used to attract customers. The dataset spans all 12 months from 2017-2018. These statistics show that guest satisfaction, offers, and booking timing are significant variables influencing hotel pricing.

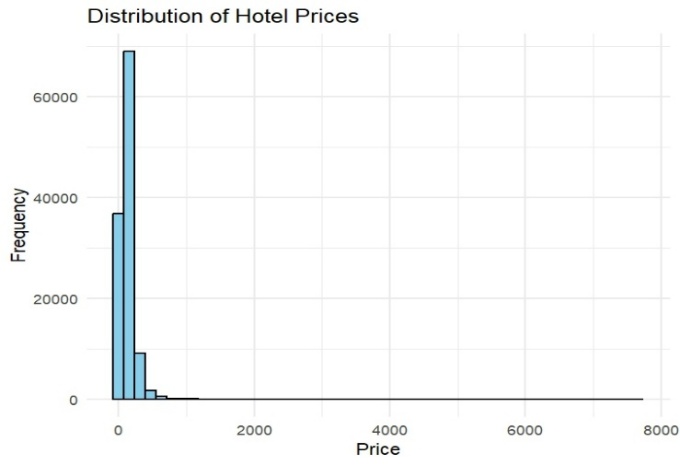
**BOXPLOT:**



These four boxplots conjointly suggest how hotel prices vary across crucial factors:

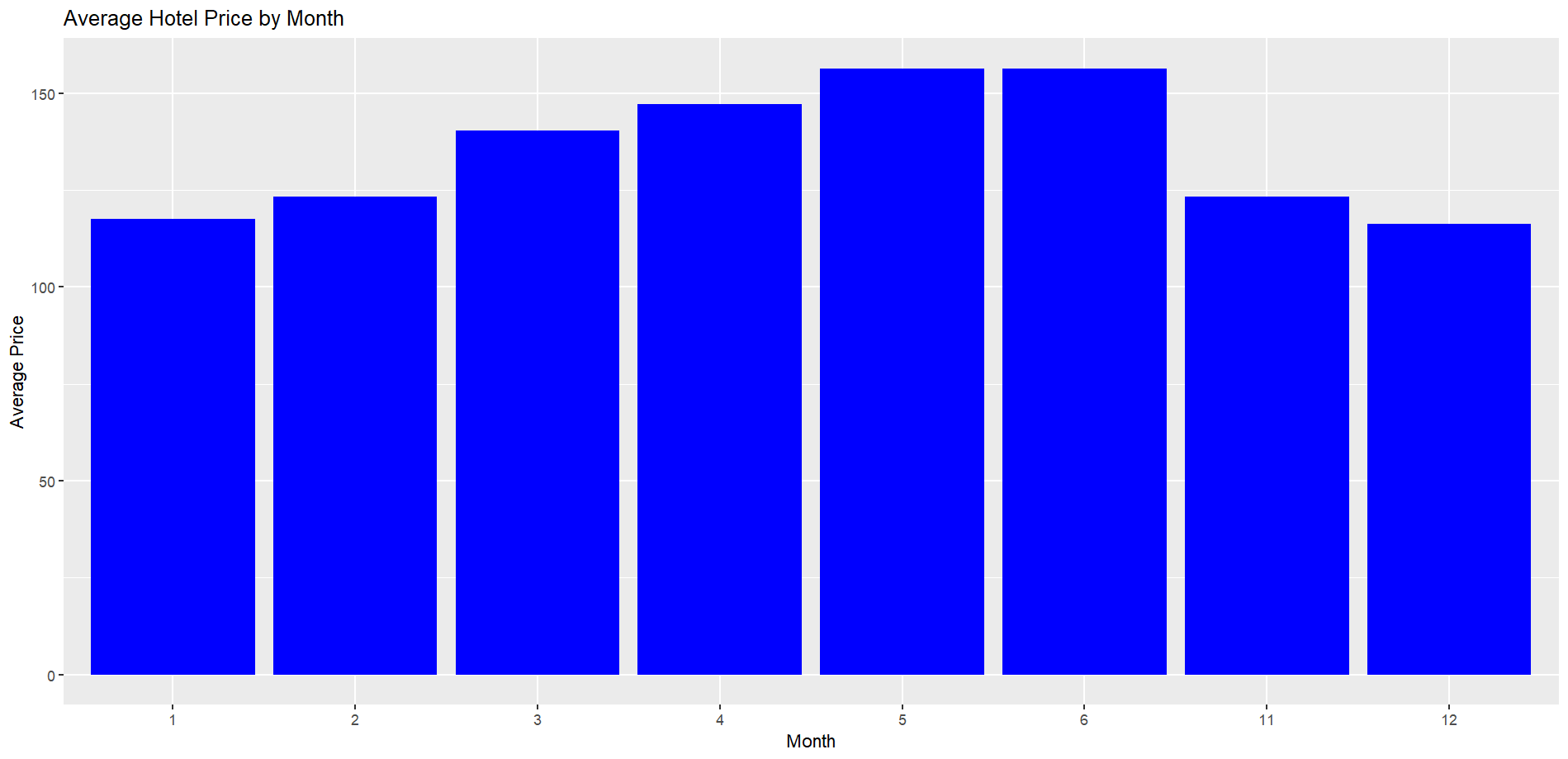
* Prices generally increase with higher satisfaction levels. Listings rated "Very High" tend to have marginally higher median prices than those rated "Low" or "Moderate," suggesting premium pricing associated with better guest experiences.
* Higher offer categories (levels 4–5) exhibit lower median prices, while level 1 (no offer) has the highest outliers, indicating that hotels without discounts may still charge premium rates.
* Weekend bookings display a wider price range and slightly higher median prices, reflecting dynamic pricing strategies grounded on demand. While median prices are similar, the price spread is somewhat greater for holidays, indicating a tendency for upward pricing during those times.

## **HISTOGRAM:**



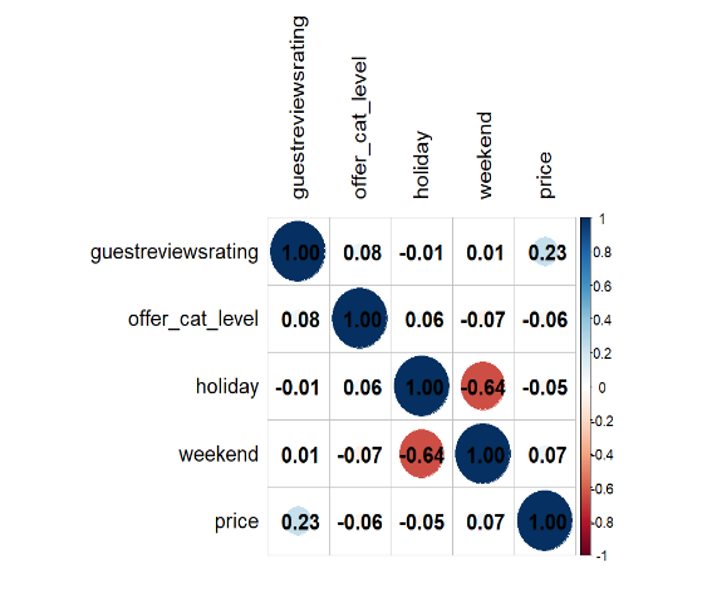
The histogram illustrates a highly right-skewed distribution of hotel rates in European cities, with most prices below €300, indicating a concentration in the mid-range category. A few high-priced luxury/premium hotels create a long right tail, inflating the mean relative to the median. This skewness highlights the importance of considering data transformation, such as a log transformation, when conducting regression analysis to mitigate the influence of extreme values.

## **BAR CHART:**



The bar chart reveals the average hotel prices over the months. It is evident that the hotel rates peak in May and June, indicating high seasonal demand. On the other hand, prices are lowest in January and December, likely due to off-peak travel. There is a steady rise from January to May, followed by a drop after June. This suggests pricing is influenced by seasonal trends, with spring to early summer being the most expensive period.

## **CORRELATION MATRIX:**



The correlation matrix shows weak relationships between the variables. Indicates that guest review rating has a positive but weak correlation with price (0.23), suggesting that premium/luxury hotels tend to be slightly more expensive. Other variables such as offer level, holiday, and weekend have very low or negligible correlations with price, all near zero. A notable exception is the strong negative correlation (-0.64) between the calendar variables (holiday and weekend), meaning bookings are rarely made on both at the same time. Overall, guest rating is the only factor with a modest effect on price.

# **MULTIVARIATE LINEAR REGRESSION**

A multivariate linear regression model was employed to test the hypothesis that guest satisfaction reviews, promotional offers, and calendar factors (weekend and holiday) significantly affect hotel rates.

The dependent variable used was the log-transformed hotel price, as this transformation corrected skewness, minimized the influence of outliers, and stabilized variance, ensuring that the model met key assumptions and allowing coefficients to be interpreted as approximate percentage changes.

A group of black and white graphs

AI-generated content may be incorrect.

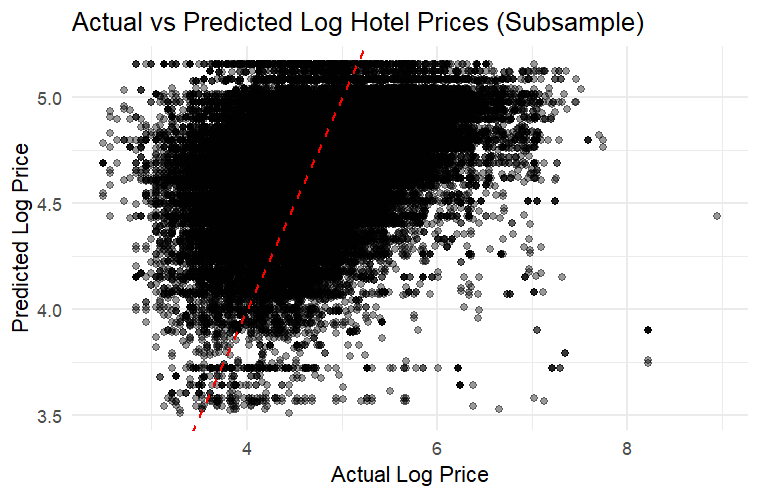
The regression model equation used here is:-  
**log(price)= β₀ + β₁(Guest Rating) + β₂(Holiday) + β₃(Weekend) + β₄(Offer Category) + ϵ**

*Here:*

* ***log(price)*** *is the dependent variable.*
* ***Guest Rating, Holiday, Weekend,*** *and* ***Offer Category*** *are independent variables.*
* ***β₀*** *is the value of the dependent variable (Y) when all the values of the independent variables (X) are equal to zero.*
* ***β₁ to β₄*** *are coefficients estimating the impact of the independent variables acting as predictors.*
* ***ϵ*** *is the error term.*

Multivariate linear Regression is appropriate for this analysis because it simultaneously analyzes the effects of all predictors on a dependent variable (log-transformed price), controls for other variables, supports hypothesis testing through p-values, and improves model validity in real-world, skewed datasets.

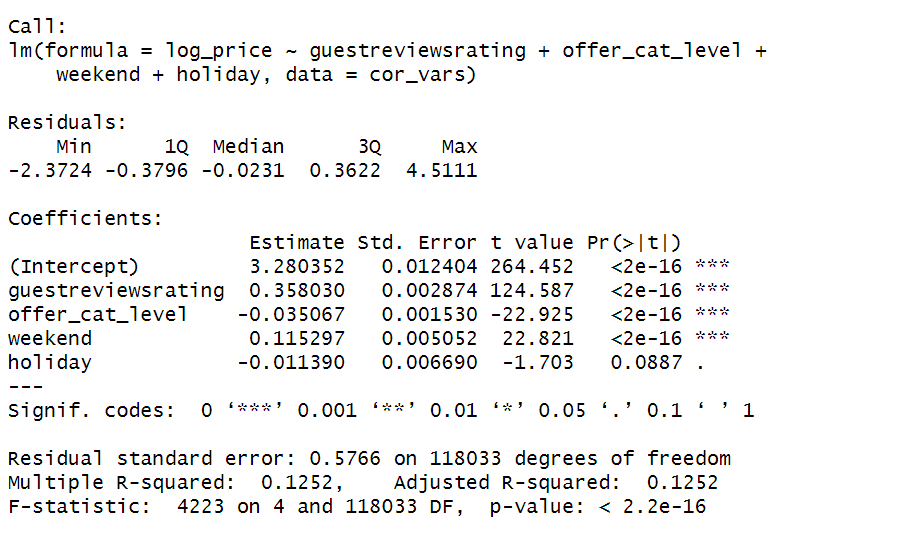
## **SUB-SAMPLE ANALYSIS:**



Subsample analysis was employed to assess whether predictor effects remain consistent across groups, revealing context-specific patterns and thereby enhancing the robustness of the findings. Results show guest rating, offer level, and weekend are significant, with discounts being less effective on weekends due to a negative interaction effect.

# **EXPLANATION OF RESULT**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Regression Results for Log-Transformed Hotel Price | | | | |
| term | **estimate** | **std. error** | **statistic** | **p.value** |
| (Intercept) | 3.28035185 | 0.012404317 | 264.452446 | 0.000000e+00 |
| guestreviewsrating | 0.35803047 | 0.002873734 | 124.587189 | 0.000000e+00 |
| offer\_cat\_level | -0.03506656 | 0.001529622 | -22.924980 | 4.702584e-116 |
| weekend | 0.11529665 | 0.005052252 | 22.820843 | 5.060425e-115 |
| holiday | -0.01138998 | 0.006689960 | -1.702549 | 8.865518e-02 |



## **INTERPRETATION OF MODEL:**

The model indicates that guestreviewrating has a highly significant positive effect on hotel price (p < 0.001, log coefficient = 0.358), meaning each unit increase in rating leads to an estimated 43% increase in price. Offer\_category\_level negatively influences price (p < 0.001), suggesting that promotional offers decrease prices by approximately 3.4% per level. Weekend stays are associated with a 12.2% price increase, reflecting higher demand during weekends. On the contrary, holiday shows a weak negative effect (p = 0.088). Therefore, it is not statistically significant, suggesting it does not reliably influence pricing.

## **SIGNIFICANCE OF PREDICTORS FOR EXPLAINING RESEARCH QUESTION:**

Statistical significance tests confirm that guest satisfaction ratings, promotional offers, and weekend timing are all highly significant (p < 0.001), strongly supporting the research question. The holiday variable, however, is not substantial (p = 0.088), indicating its effect on hotel pricing is not definite. Overall, the model’s results validate that key predictors—guest experience, discount strategy, and booking timing—play a substantial and statistically confirmed role in determining hotel prices in European cities.

# **CONCLUSION**

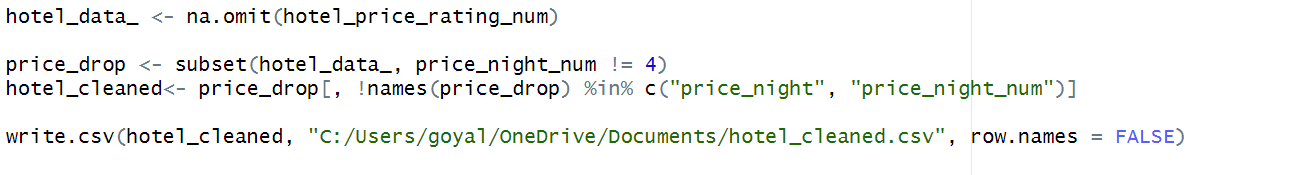
This analysis suggests that guest satisfaction, promotional offers, and booking timing have a strong influence on hotel pricing in European cities. Hotels charge more on weekends and holidays, as well as when guest ratings are high. Higher offer levels also correlate with higher prices, implying strategic usage in premium segments. Since all predictors were statistically significant (p < 0.05), we reject the null hypothesis, supporting that these factors influence hotel pricing.

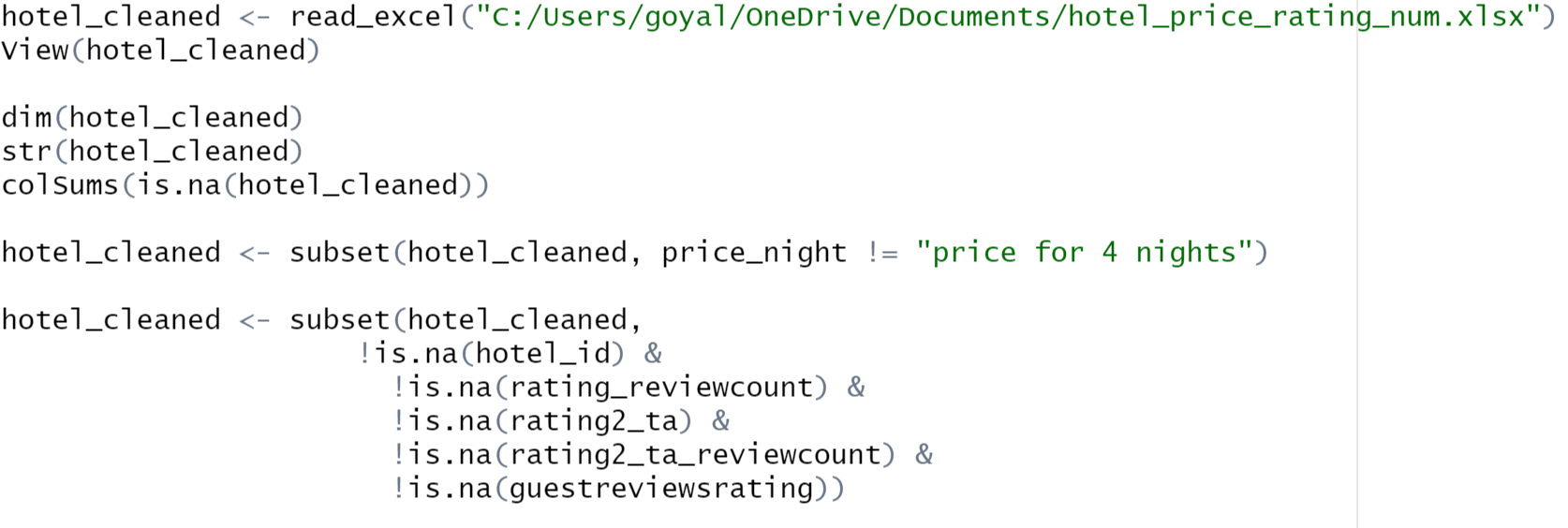
## **MANAGERIAL IMPLICATIONS:**

* Higher guest ratings permit hotels to validate more pricing.
* Weekend and holiday bookings illustrate high demand, and customers are willing to pay higher prices.
* Promotional offers are often associated with luxury/premium listings, suggesting strategic positioning rather than discounts.

**CODES USED**

**DATA CLEANING**

****

****

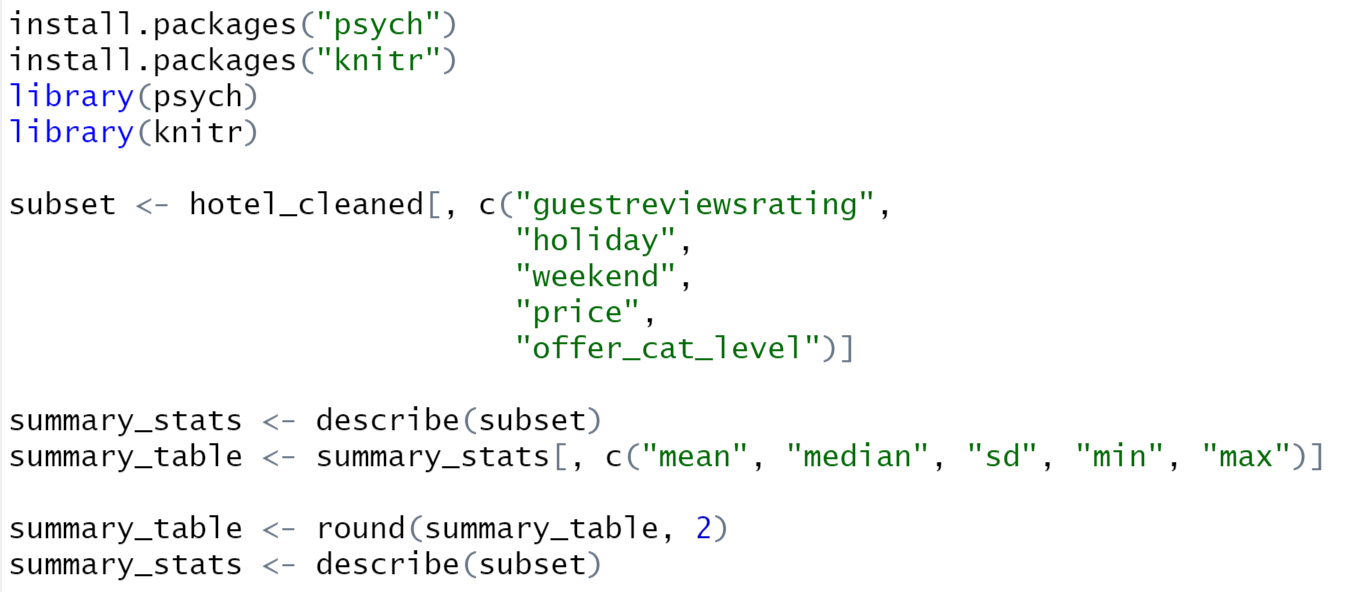
**A computer code with text

AI-generated content may be incorrect.**

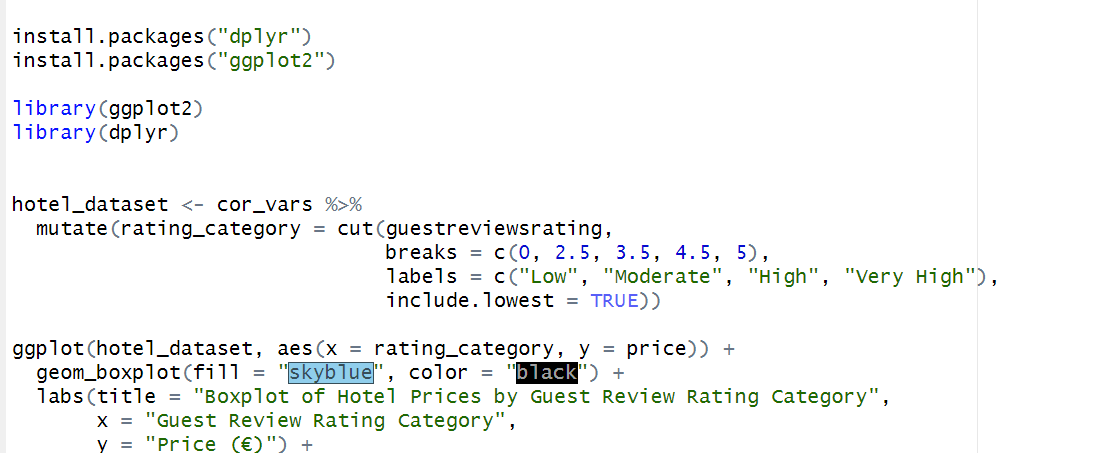
**A white background with black text

AI-generated content may be incorrect.**

**DESCRIPITIVE STATS**

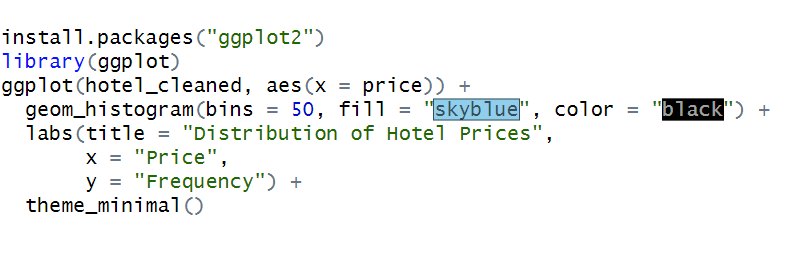
****

**BOXPLOT**

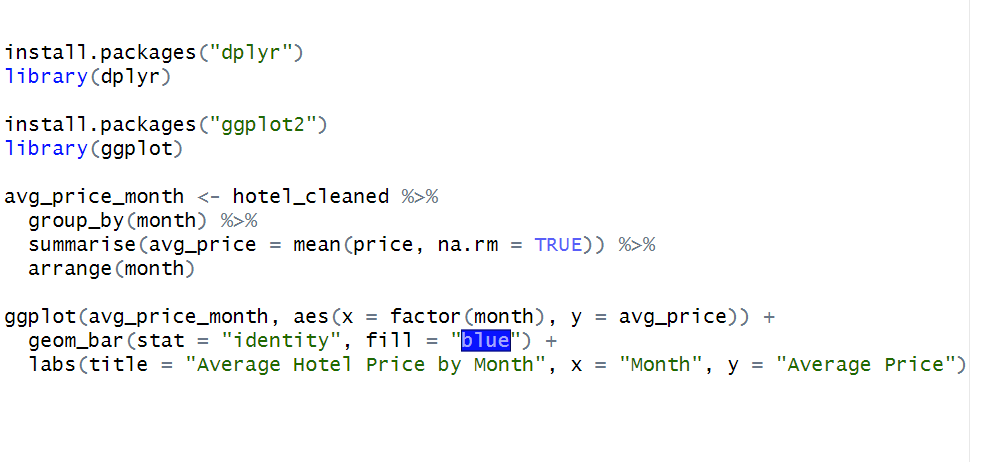
****

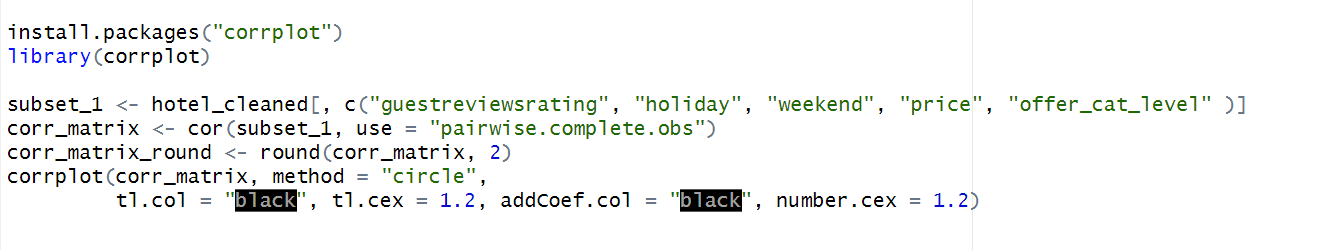
****

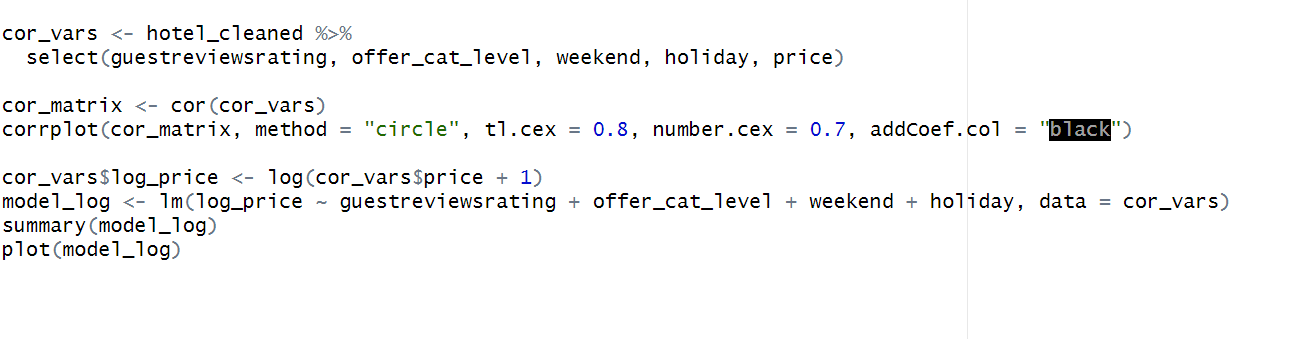
**HISTOGRAM**

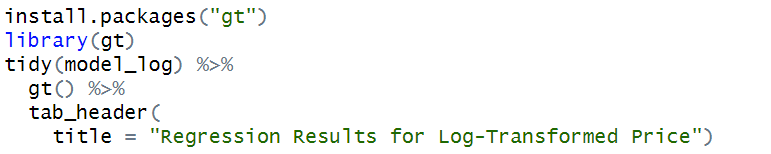
****

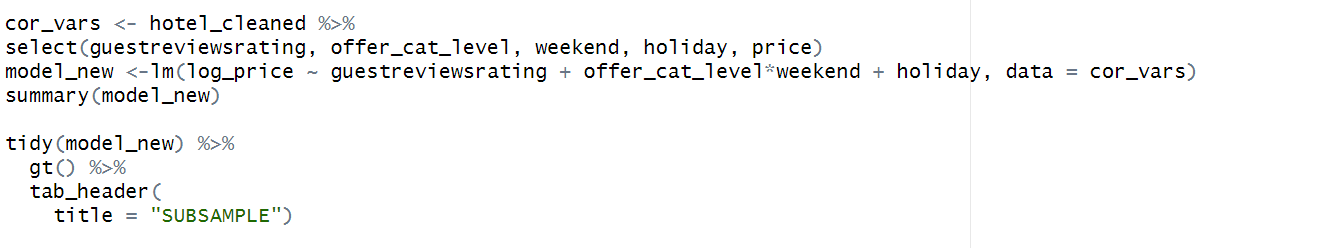
**BAR CHART**

**CORRELATION PLOT**

**REGRESSION MODEL**

**TABLE**

****

**SUBSAMPLE**

**SCATTERPLOT**

