MBA 547 Case Report, Homework 1

Topic: Multiple Linear Regression on Hotel Prices in Madrid – 2018 Data

Due Date: October 3, 2024

Submitted by: Ofuka Abung, Sadaf Vora, and Samyak Shah

Multiple Linear Regression on Hotel Prices in **Madrid** – 2018 Data

Executive Summary

We analyzed hotel pricing in Madrid for 2018, focusing on the key factors influencing prices, including star ratings, distance from the city center, and user ratings. Our dataset included 2,535 hotel observations. Distance from the city center and the number of stars a hotel has were the most significant factors in determining price. Specifically, we found that hotel prices drop by roughly €8.35 per kilometer further from the city center, while each additional star adds around €24.75 to the cost. While we also examined the effect of user ratings, they proved to be less consistent in impacting pricing.

Three (3) regression models were developed to understand price variation better. The first model used basic features and explained about 19.83% of the price variation. In our second model, we applied logarithmic transformations, which improved the model's performance, increasing the adjusted R-squared to 23.06%. The third model included polynomial terms to account for non-linear patterns, slightly improving accuracy and offering more insights into pricing behaviors. The second model proved to be the most effective of the three. It showed that a 1% increase in hotel rating could increase prices by €83.67, while a 1% increase in distance led to a price reduction of €29.31.

Our findings highlight that Madrid hotel managers should consider location and star rating when determining pricing strategies. Given the strong influence of these two factors, focusing on proximity to the city center and maintaining higher star ratings could provide competitive advantages. We also recommend further research into how specific amenities and seasonal trends affect hotel prices, as these could offer additional insights for refining pricing strategies.

Introduction

For this homework assignment, we were asked to analyze hotel pricing data in Madrid for 2018 using multiple linear regression. The goal was to determine how factors such as star rating, distance from the city center, and user ratings influence hotel prices. To achieve this, we merged two datasets: one

containing hotel features and another containing pricing details. We performed a preliminary descriptive analysis to identify key trends in the data and then built three regression models to explain price variation. This report will discuss the steps we took in the analysis, the models we tested, and our conclusions.

Data

The datasets we used for this analysis include hotel features (star rating, distance from the city center, and user rating) and hotel prices for various European cities. Our study focused on Madrid in 2018. After merging the two datasets based on a common column (hotel_id), we filtered the data for Madrid. The total sample size was 2,535 observations. We performed a correlation analysis and generated summary statistics to understand the relationships among the variables. We also created histograms and scatter plots to visualize the distribution of key variables. The data revealed a wide range of prices, with some hotels charging significantly more based on their star rating and proximity to the city center.

Fig 1.1

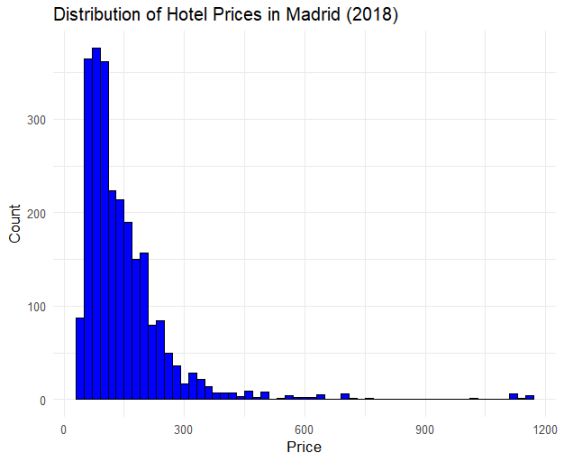

Fig 1.1 shows that most hotels were priced below €200. Many are budget-friendly, around the €50-100 range, and the distribution is right-skewed, indicating fewer high-priced hotels. Outliers appear above €600, representing luxury hotels, though they are rare.

Our analysis of the report confirms that hotel star ratings and proximity to the city center significantly impact prices. Higher-rated hotels closer to the center charge more, aligning with the observed price distribution.

Fig 1.2

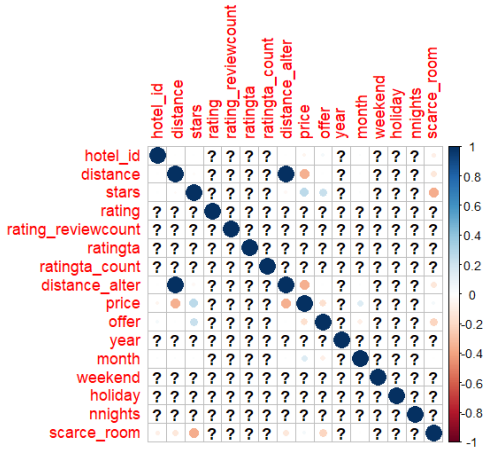

Fig 1.2 shows a correlation matrix of the relationships between different variables in the dataset, with each cell representing the correlation coefficient between two variables. Larger, darker circles indicate strong correlations. Higher hotel star ratings are positively correlated with higher prices, while distance from the city center is negatively correlated with price, meaning hotels farther from the center are generally cheaper. The matrix was used to identify relationships between variables before further analysis, ensuring that no significant multicollinearity affected the regression models built afterward.

Fig 1.3

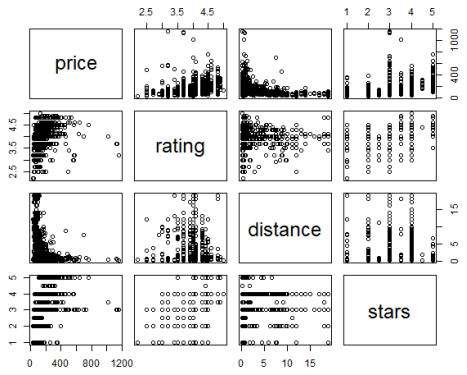

Fig 1.3 displays pairwise relationships between price, rating, distance, and stars for the hotels. Some key insights include:

Price vs. Distance: There is a clear negative relationship. Hotels farther from the city center tend to have lower prices.

Price vs. Stars: Higher-star hotels generally have higher prices, showing a positive relationship.

Rating vs. Price: A weak positive correlation indicates that higher-rated hotels tend to have slightly higher prices but with significant variation.

Distance vs. Stars/Rating: No strong trends are visible between these variables, indicating limited correlation.

Analysis

We built three linear regression models to explain hotel price variation:

Model 1: Used the **test indicator variables** and had an adjusted R-squared of **0.1983**.

For every 1-point increase in hotel rating, the price increases by €28.05. An additional kilometer away from the city center decreases the price by €8.35, and an extra star increases the price by €24.75.

All variables are statistically significant, meaning they strongly affect hotel prices. However, the model explains only 19.93% of the price variation, suggesting other factors also influence hotel prices.

```
Call:
lm(formula = price ~ rating + distance + stars, data = madrid_2018_data)

Residuals:
    Min      1Q  Median      3Q     Max
-110.34  -47.49  -20.46   17.74 1015.35

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -18.6788    18.7274  -0.997    0.319
rating       28.0524     5.0897   5.512 3.93e-08 ***
distance     -8.3541     0.4522 -18.473  < 2e-16 ***
stars        24.7454     2.3475  10.541  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 101.5 on 2435 degrees of freedom
  (96 observations deleted due to missingness)
Multiple R-squared:  0.1993,   Adjusted R-squared:  0.1983
F-statistic:   202 on 3 and 2435 DF,  p-value: < 2.2e-16
```

Model 2: Applied **log transformations** to rating and distance to capture potential non-linear relationships, and we had an adjusted R-squared of **0.2306**

A 1% increase in rating leads to a rise in price by €83.67, while a 1% increase in distance from the city center decreases the price by €29.31. Finally, an additional star increases the cost by €32.56.

All variables are highly significant. The model explains 23.16% of the price variation, slightly better than the previous model. However, it still leaves room for other factors to influence hotel prices.

```
Call:
```

```
lm(formula = price ~ log(rating) + log(distance) + stars, data =
madrid_2018_data_clean)

Residuals:
    Min      1Q  Median      3Q     Max
-128.60  -43.26  -17.36   13.47  984.93

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   -68.113     23.077  -2.952  0.00319 **
log(rating)    83.666     18.225   4.591 4.64e-06 ***
log(distance) -29.314      1.355 -21.637  < 2e-16 ***
stars          32.562      2.311  14.092  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 99.46 on 2435 degrees of freedom
Multiple R-squared:  0.2316,   Adjusted R-squared:  0.2306
F-statistic: 244.6 on 3 and 2435 DF,  p-value: < 2.2e-16
```

Model 3: Included **polynomial terms** for rating and distance to test for non-linear patterns and had an adjusted R-squared of **0.2292**

The first-degree polynomial for rating increases the price by €486.57, showing a strong positive effect. Meanwhile, the second-degree polynomial for rating further increases the price by €292.96, indicating a non-linear, upward trend. Finally, the first-degree polynomial for distance decreases the price by €1844.38, indicating a significant negative relationship.

Regarding distance, the second-degree polynomial for distance increases the price by €931.49, suggesting a complex non-linear effect. Each additional star increases the price by €27.81.

All variables are highly significant. Similar to Model 2, the model explains 23.08% of the price variation, capturing both linear and non-linear relationships between the variables and hotel prices.

```
Call:
lm(formula = price ~ poly(rating, 2) + poly(distance, 2) + stars,
    data = madrid_2018_data_clean)

Residuals:
    Min      1Q  Median      3Q     Max
-122.18  -41.67  -16.21   13.96  998.80

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)            52.811      8.049   6.561 6.51e-11 ***
poly(rating, 2)1      486.574    111.981   4.345 1.45e-05 ***
poly(rating, 2)2      292.963    102.385   2.861  0.00425 **
poly(distance, 2)1 -1844.376    101.249 -18.216  < 2e-16 ***
poly(distance, 2)2   931.493    102.416   9.095  < 2e-16 ***
stars                 27.813      2.363  11.771  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 99.56 on 2433 degrees of freedom
```

```
Multiple R-squared:  0.2308,   Adjusted R-squared:  0.2292
F-statistic:   146 on 5 and 2433 DF,  p-value: < 2.2e-16
```

All three models indicated that hotel distance from the city center and star rating were significant price predictors, while user ratings had a weaker effect. We also calculated each model's variance inflation factors (VIFs) to assess multicollinearity. The VIFs were all within acceptable limits, suggesting that the variables are not highly correlated.

Multicollinearity Test Results: We conducted multicollinearity tests for the 3 models using the Variance Inflation Factor (VIF).

Model 1

Rating: 1.23, Distance: 1.01, Stars: 1.22

All VIF values are below 5, indicating no multicollinearity issues.

Model 2

log(Rating): 1.21, log(Distance): 1.02, Stars: 1.23

Again, all values are below 5, suggesting no significant multicollinearity.

Model 3

poly(Rating, 2): 1.07, poly(Distance, 2): 1.02, Stars: 1.13

The GVIF values are also low, indicating no multicollinearity concerns.

Conclusions/Recommendations

While we found that distance and star ratings were the most significant factors influencing prices, with hotels closer to the city center and those with higher star ratings charging more, user ratings had less of an impact on pricing.

After testing three regression models, we selected **Model 2** as the best fit. This model, which applied logarithmic transformations, explained 23.06% of the price variation and effectively captured non-linear relationships. It showed that a 1% increase in rating raised prices by €83.67, while a 1% increase in distance lowered prices by €29.31. This model also showed a VIF score of below 5, indicating an absence of multicollinearity (even though model 3 is the best).

We recommend that hotel managers focus on location and star ratings when setting prices. Additionally, incorporating factors like amenities and seasonality could further refine pricing strategies, ensuring competitiveness and revenue maximization.

Appendix

```
Preliminary descriptive statistics
> summary(madrid_2018_data)
    hotel_id        city              distance          stars
rating         country          city_actual
 Min.   :8877   Length:2535        Min.   : 0.100   Min.   :1.000
Min.   :2.200   Length:2535        Length:2535
 1st Qu.:9030   Class :character   1st Qu.: 0.400   1st Qu.:3.000   1st
Qu.:3.700   Class :character   Class :character
 Median:9246   Mode  :character   Median : 1.200   Median :3.000
Median :4.000   Mode  :character   Mode  :character
 Mean   :9235                      Mean   : 3.551   Mean   :3.293
Mean   :3.989
 3rd Qu.:9418                      3rd Qu.: 5.600   3rd Qu.:4.000   3rd
Qu.:4.300
 Max.   :9584                      Max.   :19.000   Max.   :5.000
Max.   :5.000

NA's   :96
 rating_reviewcount center1label       center2label       neighbourhood
ratingta     ratingta_count
 Min.   :   1.0     Length:2535        Length:2535        Length:2535
Min.   :1.500   Min.   :   2.0
 1st Qu.:  48.0     Class :character   Class :character   Class :character
1st Qu.:3.500   1st Qu.: 152.0
 Median : 104.0     Mode  :character   Mode  :character   Mode  :character
Median :4.000   Median : 396.0
 Mean   : 167.8
Mean   :3.894   Mean   : 633.1
 3rd Qu.: 217.0
3rd Qu.:4.500   3rd Qu.: 910.0
 Max.   :1855.0
Max.   :5.000   Max.   :4211.0
 NA's   :96
NA's   :112     NA's   :112
 distance_alter   accommodation_type     price            offer
offer_cat             year          month
 Min.   : 0.000   Length:2535        Min.   :  31.0   Min.   :0.000
Length:2535         Min.   :2018   Min.   :1.000
 1st Qu.: 0.400   Class :character   1st Qu.:  79.5   1st Qu.:0.000
Class :character   1st Qu.:2018   1st Qu.:2.000
 Median : 1.100   Mode  :character   Median : 117.0   Median :1.000
Mode  :character   Median :2018   Median :3.000
 Mean   : 3.506                     Mean   : 147.0   Mean   :0.716
Mean   :2018   Mean   :3.397
 3rd Qu.: 5.400                     3rd Qu.: 181.0   3rd Qu.:1.000
3rd Qu.:2018   3rd Qu.:5.000
 Max.   :19.000                     Max.   :1159.0   Max.   :1.000
Max.   :2018   Max.   :6.000

    weekend     holiday     nnights   scarce_room
 Min.   :1   Min.   :0   Min.   :1   Min.   :0.0000
```

```
 1st Qu.:1   1st Qu.:0   1st Qu.:1   1st Qu.:0.0000
 Median :1   Median :0   Median :1   Median :0.0000
 Mean   :1   Mean   :0   Mean   :1   Mean   :0.3444
 3rd Qu.:1   3rd Qu.:0   3rd Qu.:1   3rd Qu.:1.0000
 Max.   :1   Max.   :0   Max.   :1   Max.   :1.0000
```

```
> # Calculate VIF for model 1
> vif(model1)
  rating distance    stars
1.228534 1.007208 1.220628
> # Calculate VIF for model 2
> vif(model2)
  log(rating) log(distance)         stars
     1.213444      1.023163      1.232209
> # Calculate VIF for model 3
> vif(model3)
                      GVIF Df GVIF^(1/(2*Df))
poly(rating, 2)   1.329207  2        1.073738
poly(distance, 2) 1.093560  2        1.022611
stars             1.286221  1        1.134117
```

```
> AIC(model1, model2, model3)
       df      AIC
model1  5 29465.62
model2  5 29365.37
model3  7 29371.88
```