

# Finding Hotel Deals in Vienna using Machine Learning

Bruno Helmecky

12/02/2021

## Executive Summary

This report looks to find the 5 best “Deals” in Vienna, during holidays in December 2017. I look to answer “What are the 5 best offers in town during 2017 year-end holiday period ?” by applying a Linear Regression, a Classification & Regression Tree, & a Random Forest model, on data obtained from Trip Advisor. I define ‘best deals’ as those observations with largest negative residuals vs their expected price according to my selected model. I found the Random Forest model to vastly outperform both the OLS & CART model specification, reaching a 68.3% Adjusted R-squared value, 16.2 euros Mean Absolute Error & 24.5 euros RMSE. Finally, the model found 5 deals that represent bargains between 44-55 euros, in addition to being diverse in terms of pricing, stay length, star rating, distance from the centre & accommodation type.

## Introduction

Vienna is a popular destination during the (pre)-christmas period, famous for its christmas market, & beautiful scenery, glamorous reminders of a glorious age called the Austria-Hungarian Monarchy. As such a conveniently distanced, & suitable destination for couples’ & families’ from neighbouring countries, to relax. I look to answer “Which 5 hotels offer the best ‘Deal’ during 2018 year-end holiday period ?” by applying a Linear Regression, Classification & Regression Tree, & a Random Forest model, on data obtained from Trip Advisor. I define ‘Best Deal’ as the hotel offering the most-value for money. In this context, after modelling hotels’ prices based on variables related to hotel quality, distance from the city centre & guest reviews, I define the ‘best deals’ as those observations with largest negative residuals vs their expected price according to my selected model.

## Data Cleaning, Association Patterns, & Variable Transformations

**Cleaning:** My raw data comprises the hotels europe dataset available by clicking [here](#). This dataset includes hotel prices & various other features from 46 european cities on 10 different dates, & is scraped from a price comparison website. I filtered the dataset for a 2017, December, holiday noted period, for deals under 400 euros per night, netting 676 observations. Note that I included both 1-night & 4-night deals, as they represent different offerings, possibly with different value for money, even if offered by the same hotel. Furthermore, it is a realistic scenario for travellers to decide on the holidays’ length, based on the deals that they would find most to their liking. To avoid comparing apples & oranges however, I calculated the price per night for every deal, which represents my target variable. I cleaned the data 1st by Converting distance strings, & review ratings to numeric vectors; 2nd by Imputing missing Review Scores for a hotel with the datasets’ median, & adding flag variable columns to note hotels without reviews; & 3rd by converting accommodation type, offer type & neighbourhood variables to a series of dummy variables, resulting in 44 predictor variables used.

**Distributions & Association Patterns:** Please see the most important quantitative variables’ summary statistics below, as well as all variables’ histograms in the appendices. One may observe skewness with long right tails for prices per night, both distance variables, & the Nr. of reviews posted on Trip Advisor, while for the alternative rating measure skewness with long left tail is visible. Based on the summary statistics & histograms I changed the target variable to its’ log-form. Also due to observed summary statistics, histograms & association patterns, I log-transformed both distance measures & review counts, while taking the exponent of guest reviews. Finally, I pivoted neighbourhoods, offer categories, & accommodation type variables to dummy tables.

| Vars                   | min  | median | mean  | max  | sd    | skewness |
|------------------------|------|--------|-------|------|-------|----------|
| price                  | 30.0 | 262.0  | 359.1 | 1546 | 300.3 | 1.4      |
| pricepernight          | 30.0 | 128.0  | 147.4 | 386  | 67.8  | 1.2      |
| Nrnights               | 1.0  | 1.0    | 2.4   | 4    | 1.5   | 0.2      |
| starrating             | 1.0  | 3.5    | 3.4   | 5    | 0.7   | -0.3     |
| center1distance        | 0.0  | 1.5    | 1.7   | 13   | 1.6   | 3.1      |
| center2distance        | 0.5  | 3.5    | 3.7   | 13   | 1.6   | 1.5      |
| rating2_ta             | 2.5  | 4.0    | 4.0   | 5    | 0.4   | -0.7     |
| rating2_ta_reviewcount | 0.0  | 233.0  | 471.6 | 3262 | 613.5 | 1.9      |
| guestreviewsrating     | 1.0  | 4.0    | 4.0   | 5    | 0.5   | -1.3     |

## Modelling

**Model Comparisons:** With cleaning & transformations out of the way, 3 models were specified, each using the same group of predictor variables & predicting the natural logarithm of prices per night. Given the goal of this analysis is to find best deals in sample, no holdout data was sampled, however I performed 10-fold cross-validation. Please see the models' resulting summary statistics below, having already transformed log-price predictions to level-forms. One can observe the Regression Tree to beat the Multiple-Linear Regression model in terms of Mean absolute error (24.5 vs 30.9) & Root Mean Square error (35.3 vs 43.7), despite boasting slightly lower Adjusted R-squared (50.5% vs 52%). The Random Forest model however, using 7 variables at a time & a minimum node size of 5, vastly outperforms both, boasting ca. 33% higher adjusted R-squared (66.8% vs 50.5%), more than 8 euros less MAE (16.3 vs 24.5) & more than 10 euros less RMSE (24.8 vs 35.3). As such, the Random Forest is the clear winner & therefore is my chosen final model.

|               | Rsq   | MAE    | RMSE   | RMSE_norm |
|---------------|-------|--------|--------|-----------|
| OLS           | 0.520 | 30.855 | 43.738 | 0.297     |
| CART          | 0.505 | 24.472 | 35.336 | 0.240     |
| Random_forest | 0.668 | 16.295 | 24.777 | 0.168     |

**Final Model:** Please see the final models' summary statistics below, re-estimated on the complete dataset. Fitted to the complete dataset, Adjusted R-squared further improved, to 68.3%, with marginal improvements in both MAE & RMSE. Importantly the Root Mean Square Error only represents 16.6% of the average daily room price. With that, the model is ready to find the 5 best deals in town for what presumably is the christmas period.

| modelname      | Rsq   | MAE    | RMSE  | RMSE_norm |
|----------------|-------|--------|-------|-----------|
| RF Final Model | 0.683 | 16.179 | 24.45 | 0.166     |

**The 5 Best Deals:** Please see the 5 best deals according to the final model, as well as 5 hotels provided as best deals' according to the reference model for this course. Interestingly, the 5 best deals seem quite diverse in nature & price, incorporating 2 hotels in Innere Stadt of different star rating for 1 night, 1 apartment just over a mile from the centre that is quite budget-friendly, & 2 apartments further from the centre, however 1 of them for 4 nights. Thus, if one sees this analysis' purpose to be finding a diverse variety of alternatives that represent good value-for-money, then the final Random Forest model achieves that purpose, each deal offering a bargain between 44-55 euros according to our model. Meanwhile, the deals offered by the 5 hotels provided as best deals' according to a reference model only represented bargains of 11-18 euros, 1 of the even being over-priced by 28 euros according to the final model.

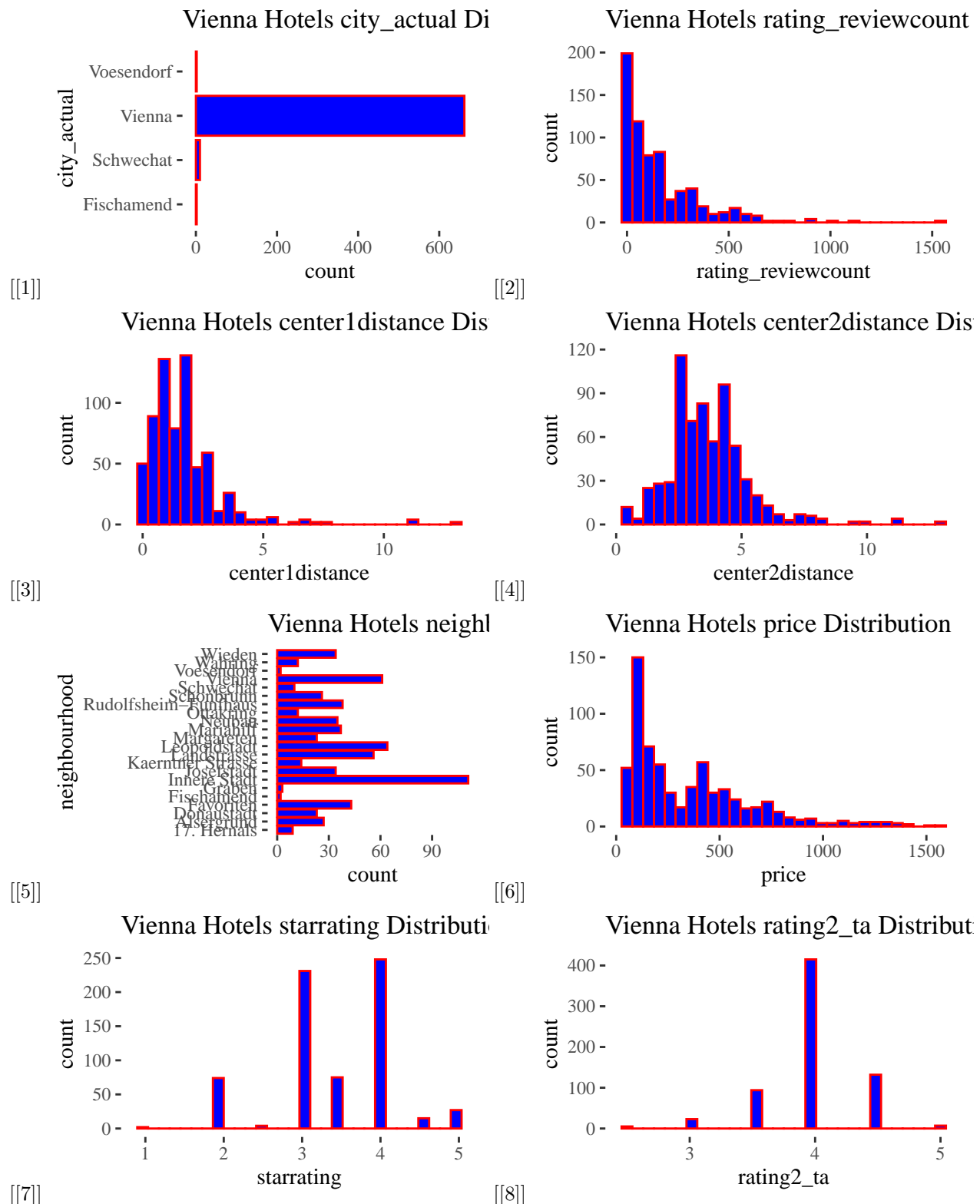
|     | Hotel_id | Stars | Avg.Price | Resid. | Nights | Type      | Where?          | Miles fr<br>Center | TA_Rating | Nr.Ratings |
|-----|----------|-------|-----------|--------|--------|-----------|-----------------|--------------------|-----------|------------|
| 1   | 21905    | 3.5   | 70        | -54.4  | 1      | Apartment | Alsergrund      | 1.2                | 4.0       | 0          |
| 2   | 21980    | 4.0   | 146       | -47.3  | 1      | Hotel     | Innere<br>Stadt | 0.1                | 4.0       | 714        |
| 3   | 21982    | 3.0   | 123       | -46.4  | 1      | Hotel     | Innere<br>Stadt | 0.5                | 4.5       | 985        |
| 4   | 22360    | 3.0   | 104       | -45.2  | 4      | Apartment | Vienna          | 2.8                | 4.0       | 0          |
| 5   | 21939    | 3.0   | 95        | -44.6  | 1      | Apartment | Favoriten       | 2.5                | 4.0       | 0          |
| 65  | 22184    | 3.0   | 85        | -18.6  | 1      | Hotel     | Leopoldstadt    | 0.7                | 4.0       | 827        |
| 79  | 21975    | 4.0   | 197       | -17.2  | 1      | Hotel     | Innere<br>Stadt | 0.1                | 4.5       | 211        |
| 84  | 22344    | 3.0   | 65        | -16.7  | 4      | Hotel     | Vienna          | 3.9                | 4.0       | 12         |
| 106 | 22344    | 3.0   | 57        | -14.8  | 1      | Hotel     | Vienna          | 3.9                | 4.0       | 12         |
| 150 | 21912    | 4.0   | 95        | -10.7  | 1      | Hotel     | Alsergrund      | 1.1                | 4.0       | 359        |
| 151 | 22080    | 3.0   | 65        | -10.7  | 1      | Hotel     | Josefstadt      | 1.1                | 3.0       | 85         |
| 588 | 21975    | 4.0   | 264       | 28.0   | 4      | Hotel     | Innere<br>Stadt | 0.1                | 4.5       | 211        |

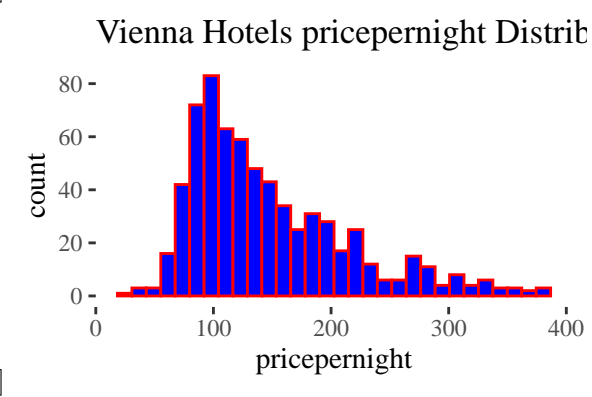
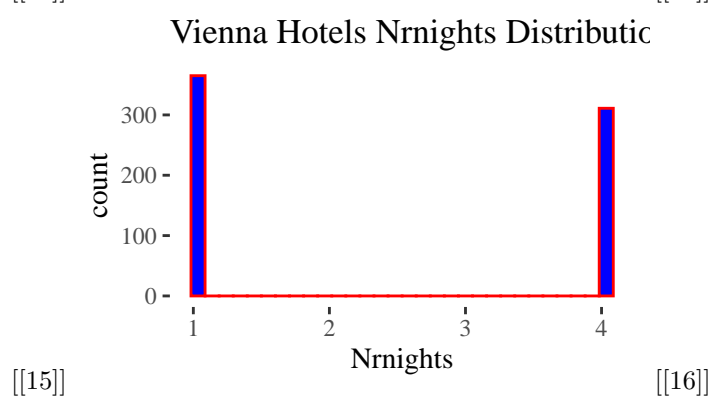
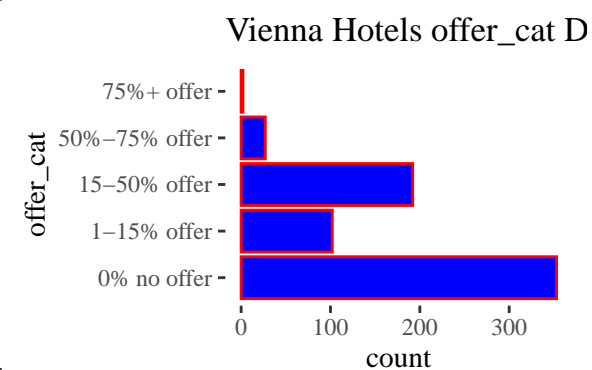
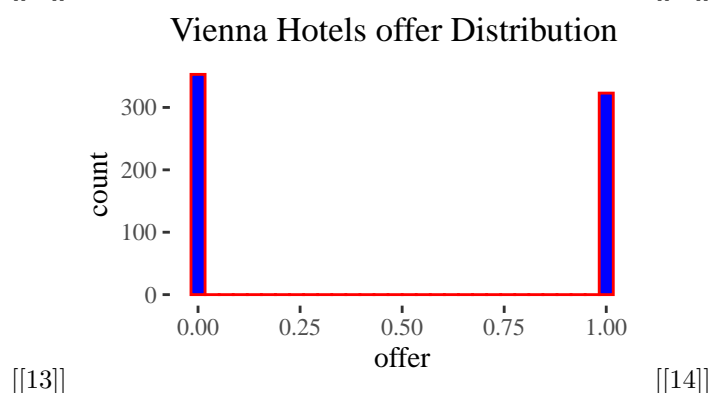
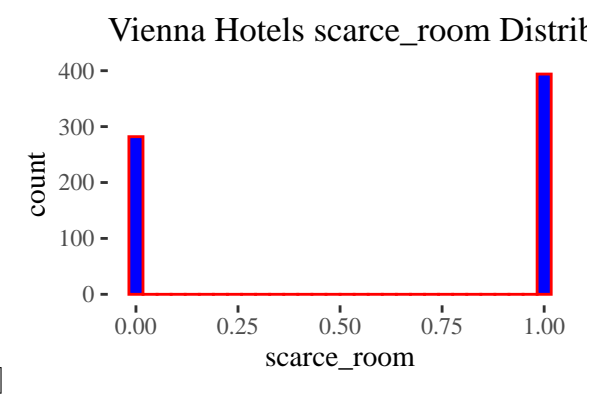
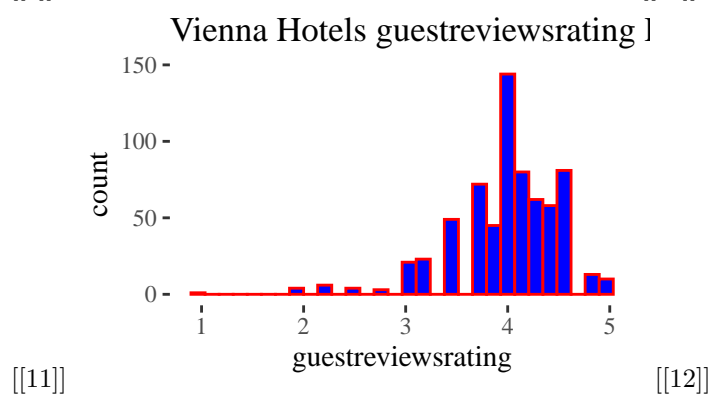
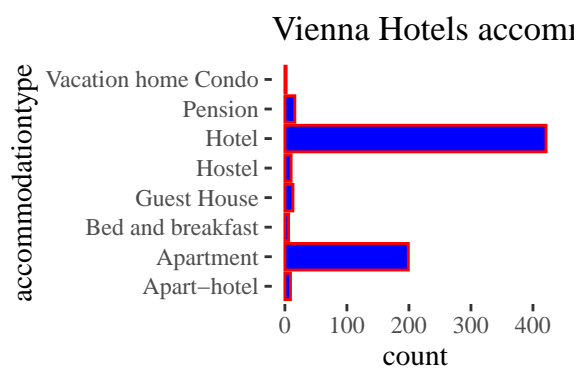
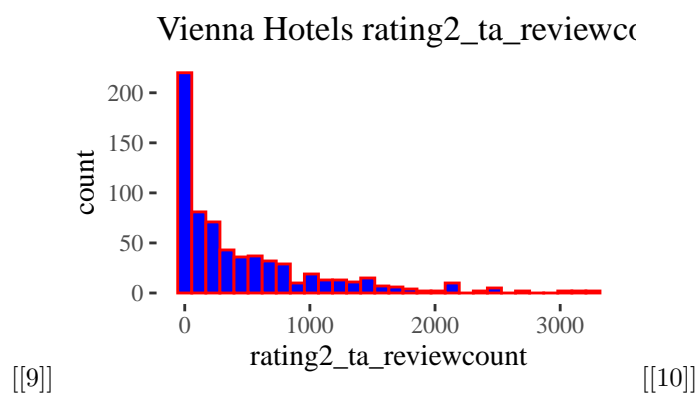
## Conclusion / Summary

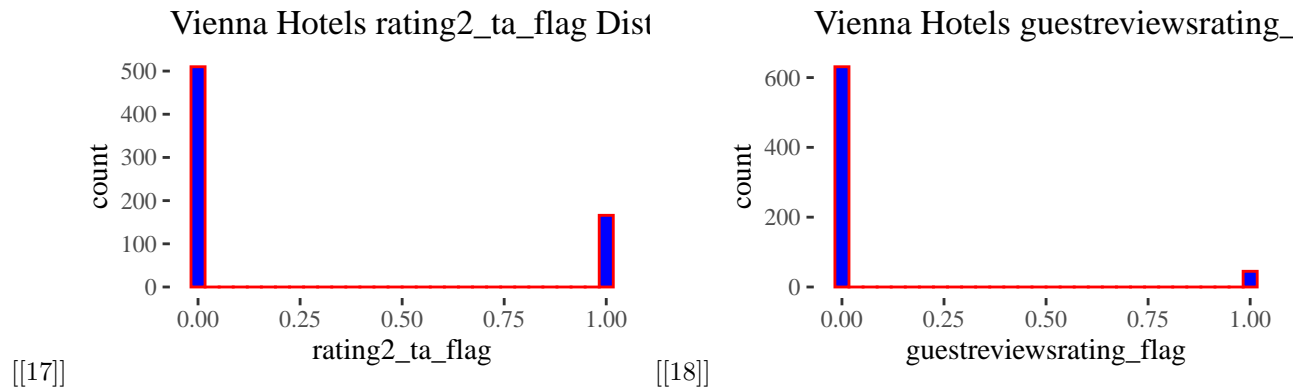
This report investigated the best hotel deals available in Vienna, during the 2017 December holiday period, using data originally scraped from a price comparison website, & also considering apartments, not just hotels, while sticking to a daily budget of max 400 euros, while being flexible about staying for 1, or 4 nights. After comparing performance of an Ordinary Least Squares, a Classification & Regression Tree, & a Random Forest model, I found Random Forest to vastly outperform OLS & CART. Meanwhile, re-estimated on the whole dataset, the final model improved slightly further, reaching adjusted R-Squared of 68.3%, mean absolute error of 16.2 euros, & root mean squared error of 24.45 euros. Using this model, I found the 5 ‘best deals’ for the period by looking at largest negative residuals, each deviating between 45-55 euros from their respective predicted prices. Interestingly, the deals found were quite diverse in both nature & pricing, while the 5 best deals according to the reference model provided for this course represented bargains of about 11-18 euros.

# Appendices

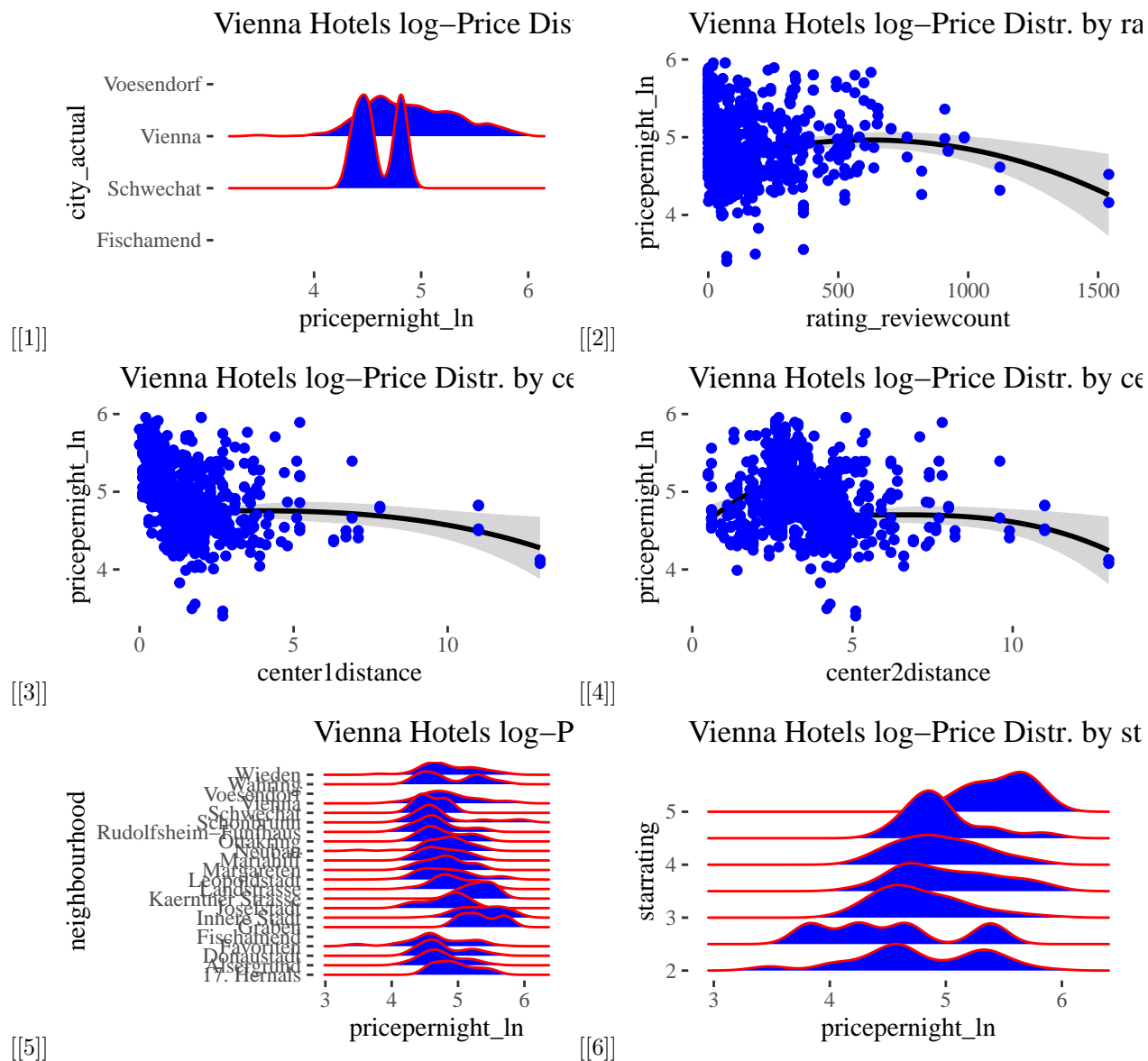
## 1) Variable Histograms

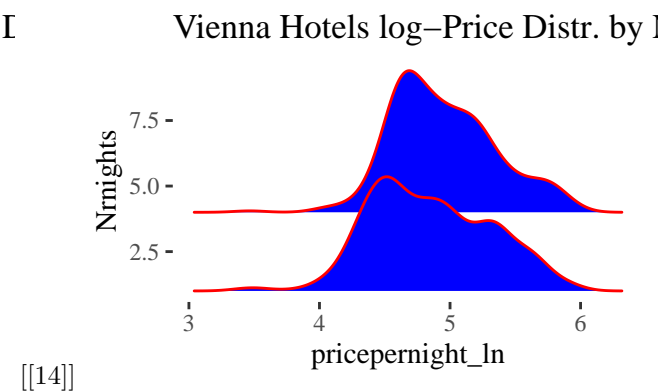
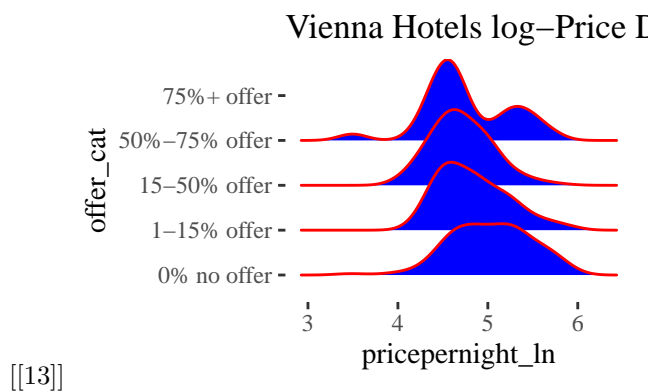
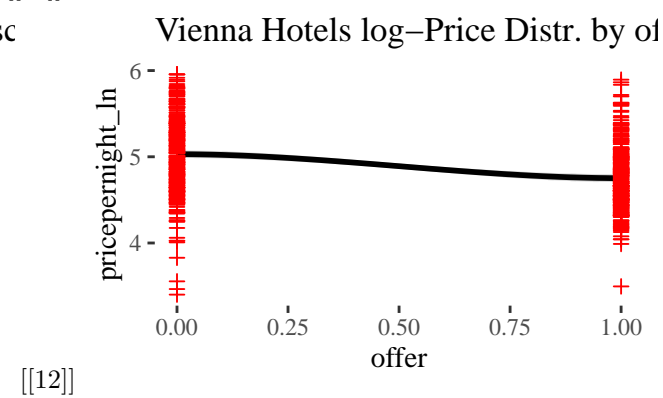
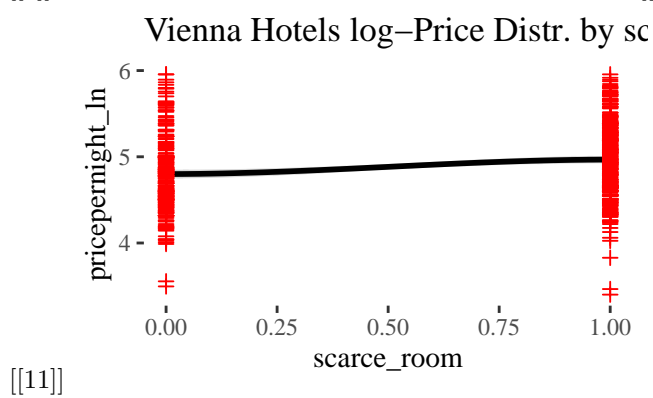
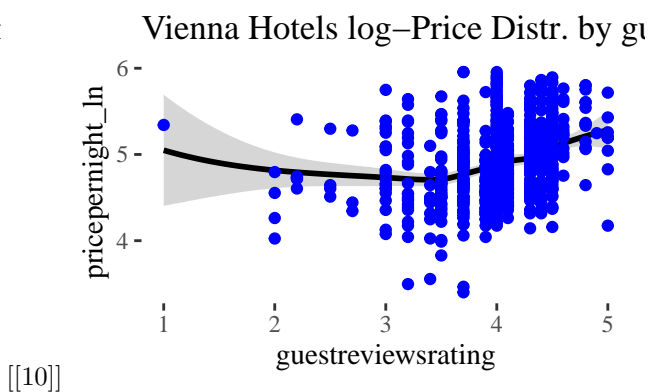
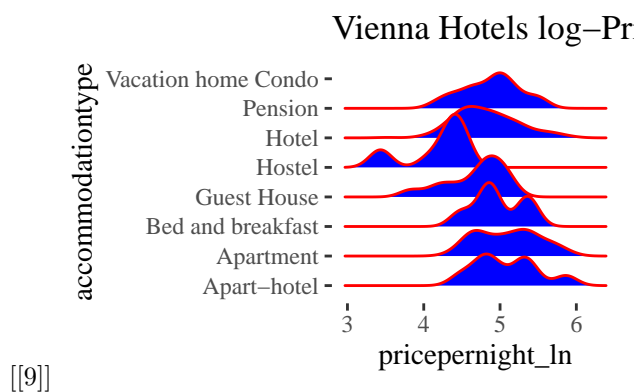
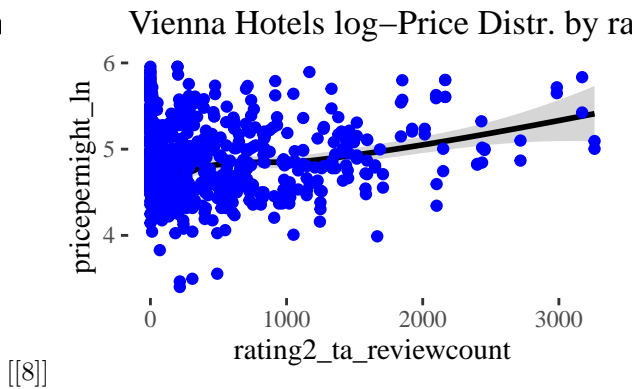
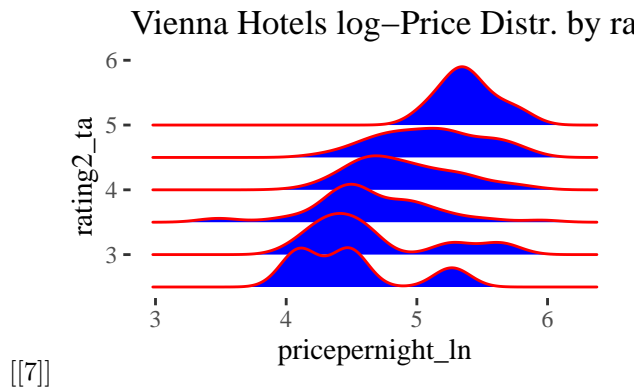


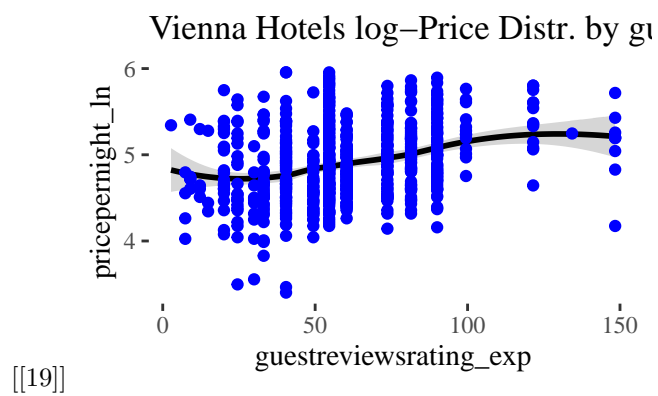
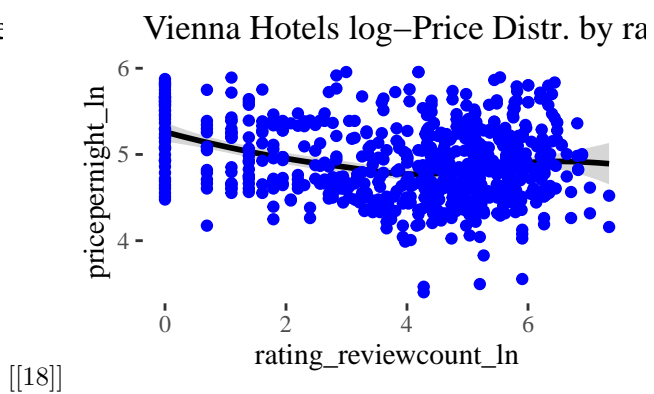
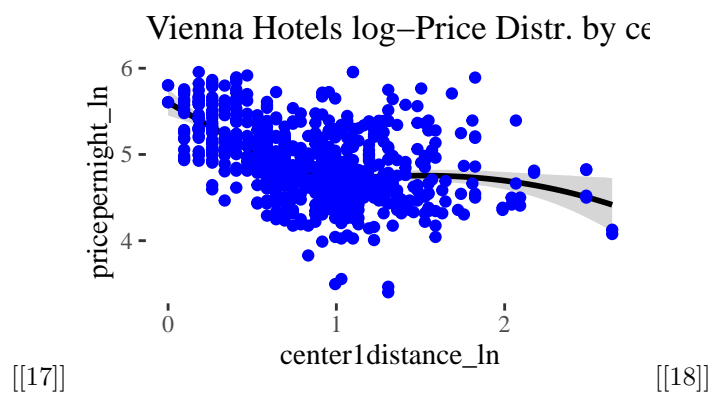
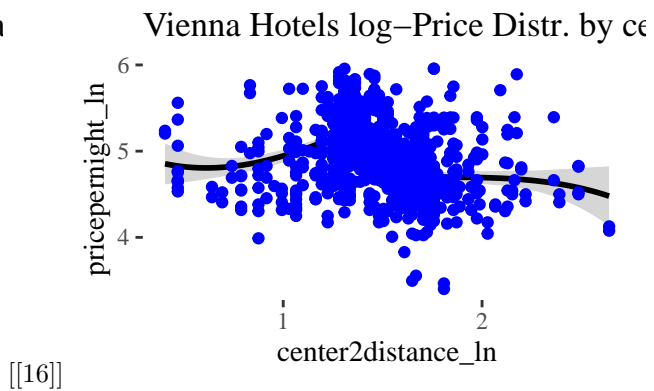
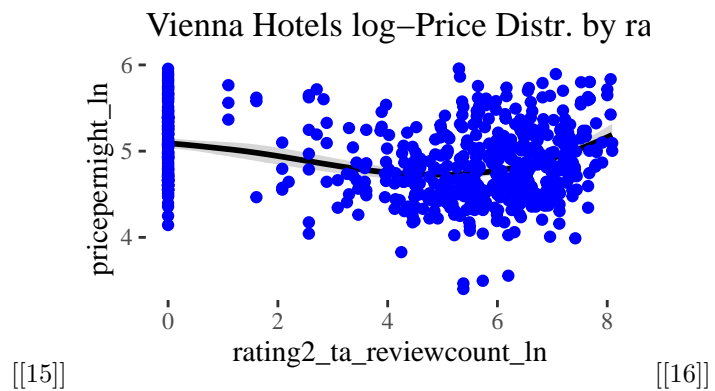




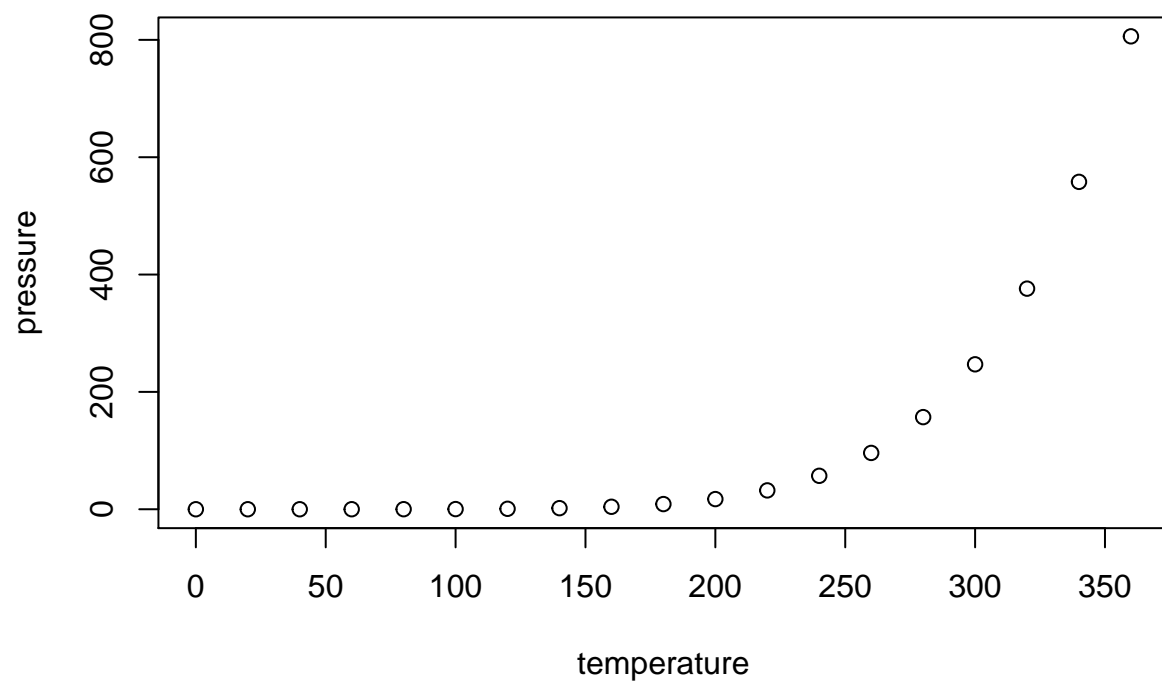
## 2) Scatterplot Associations & Ridge Distributions











Note that the `echo = FALSE` parameter was added to the code chunk to prevent printing of the R code that generated the plot.