Finding Hotel Deals in Vienna using Machine Learning

Bruno Helmeczy

12/02/2021

Executive Summary

This report looks to find the best "Hotel Deal" in Vienna, during holidays in December 2018. 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 the 'best deals' as those observations with largest negative residuals vs their expected price according to my selected model.

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

My raw data comprises the hotel europe dataset available by clicking **here**

This is an R Markdown document. Markdown is a simple formatting syntax for authoring HTML, PDF, and MS Word documents. For more details on using R Markdown see http://rmarkdown.rstudio.com.

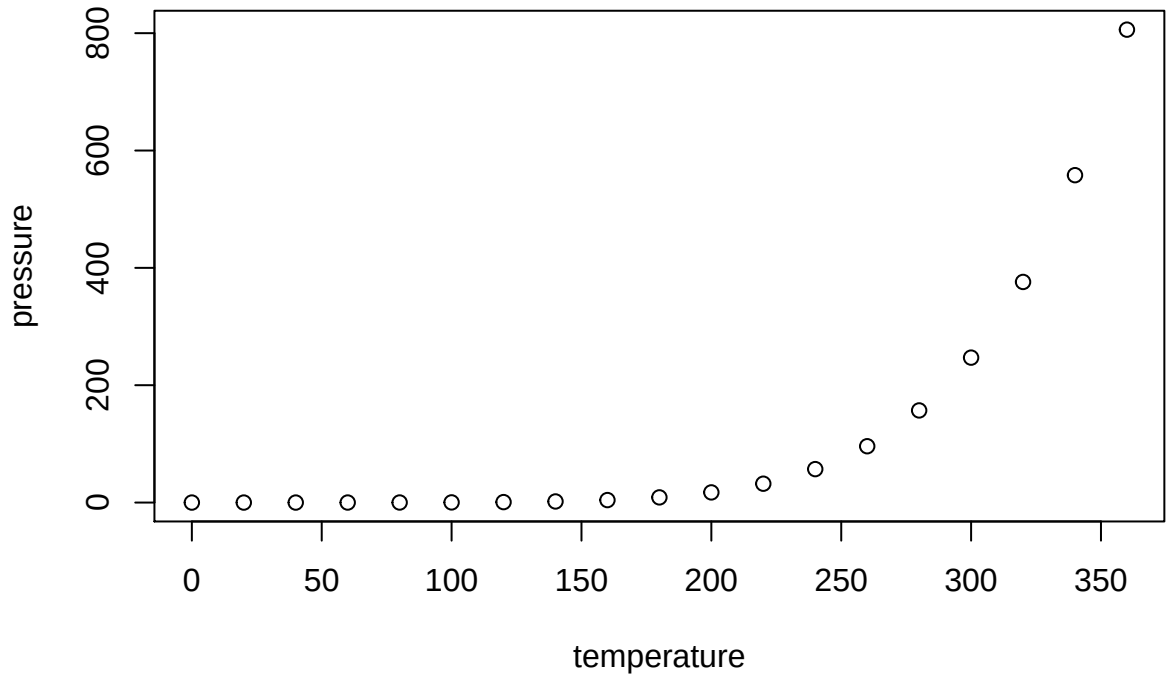
When you click the **Knit** button a document will be generated that includes both content as well as the output of any embedded R code chunks within the document. You can embed an R code chunk like this:

```
summary(cars)
```

```
##      speed           dist       
##  Min.   : 4.0   Min.   :  2.00  
##  1st Qu.:12.0   1st Qu.: 26.00  
##  Median :15.0   Median : 36.00  
##  Mean   :15.4   Mean   : 42.98  
##  3rd Qu.:19.0   3rd Qu.: 56.00  
##  Max.   :25.0   Max.   :120.00
```

Including Plots

You can also embed plots, for example:



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