

Due: Friday, September 29 11:59 PM

AIRBNB PRICING IN ASHEVILLE, NC

Airbnb wants to help new hosts set prices for their Airbnb listings in Asheville, NC. They have hired your data science consulting company to build a model to generate prices based on a variety of factors.

For this assignment, you will write your report in two parts: 1) a report (1-2 pages) describing your model to (non-technical) Airbnb executives, and 2) a report (3-4 pages) justifying your model to your (technical) data science team

THE DATASET

[Click here to download the data](#)

The data is from [Inside Airbnb](#).

[Click here for the data dictionary](#)

CLEANING AND MODEL REQUIREMENTS

You must include the following variables in your model:

- room_type (you may need to combine categories)
- number of bedrooms
- number of bathrooms (note that this variable needs to be cleaned, as the bathrooms variable is empty. I recommend using the `str_sub()` function from the [stringr package](#) and `as.numeric()` to extract the number of bathrooms
- create a new variable that gives the distance to downtown. There may be multiple ways to do this, but you can use the code below. This uses the `apply` function, which is a useful function in R to perform an operation on all rows (or columns) of a matrix. Then it uses the `distm()` function in the [geosphere package](#) to calculate the distance in meters from a latitude and longitude in downtown Asheville. Finally, it multiplies by the appropriate constant to convert the value to miles.

```
library(geosphere) #you will need to install this package
airbnb$dist_to_dt <- apply(airbnb[,c("longitude","latitude")],1,function(x)
  c(-82.55481168521978,35.59701329976918),
  x,fun=distHaversine)*0.00062137
```

Note that you will also need to clean the `price` variable. You can use the `str_sub()` function again here.

Choose at least one other variable to include in your model. Consider what is appropriate for a new host to set a price.

Extra credit

Earn extra points on this assignment by finding a way to incorporate amenity features in your model. For example, a host may want to change the price based on whether or not they allow pets.

As you fit and assess your model, consider the following elements that we have discussed in class:

- Is this a prediction or inference problem? Should model interpretability be prioritized in this situation?
- Look at the diagnostic plots for your model. Determine if you need to transform predictor(s) or the outcome variable to improve the model.
- Evaluate influential points and multicollinearity and make adjustments accordingly. If you remove any observations, be sure to include this information in your report.
- Which model metric(s) are appropriate to assess the model?

DELIVERABLES

REPORT FOR AIRBNB EXECUTIVES (1-2 PAGES)

The report for the Airbnb executives should explain your model to a non-technical audience. No code or “raw” R output should be in the report.

Specifically, your report should contain the following elements:

- **Introduction:** Provide an overview of the dataset and the goals of the analysis. Provide basic information about the data (e.g., sample size). You may choose to include summary statistics or basic plots.
- **Methods:** Explain the model you used to analyze the data without getting into technical details. Why did you decide to use that model for this dataset and how does it accomplish their goal? Which variables did you include in the model and why?
- **Results:** Justify your model with the appropriate model metric(s). Explain them in non-technical terms. Then, provide an example of how the model can be used by giving the projected price for a particular combination of variables in your model (e.g., “for a listing with 2 bedrooms, 2 bathrooms, ..., the price would be —”)
- **Conclusion:** Do you feel that the model is good enough to be deployed? If not, can you think of additional data that Airbnb could collect to improve the model?

REPORT FOR DATA SCIENCE TEAM

The second part of the assignment will be a 3-4 page report that is suitable for other data scientists. Here, you will present details of your model to justify the conclusions you presented to the client. This section should present technical details that someone with a data science background can understand. This report must include the following, though you may wish to provide additional details relevant to the analysis:

- **Introduction:** Provide details about the dataset, including any data cleaning that needed to be done. Was there any missing data? Did you make any assumptions during the data cleaning process?
- **Methods:** Describe your model assessment and building process. Did you include any interaction terms? Why or why not? Did you transform any variables? If so, why? Provide the model diagnostic plots. Did you exclude any observations? Did you make any adjustments because of multicollinearity? How did you assess your model?
- **Conclusion:** What do you conclude about the validity of this analysis?

Submission & Formatting Instructions/Tips

You will submit two files to gradescope: 1) a PDF that contains the two requested reports, and 2) the qmd file you used to produce the reports

- All code must be hidden in the PDF. You can hide all code by adding the following to your YAML Header:
execute:
 echo: false
- Your quarto document should be rendered directly to PDF, not to HTML and then saved as a PDF
- The PDF should be 5-6 pages: 1-2 pages for the non-technical report and 3-4 pages for the technical report. These ranges are intentional; reports that fall outside of this range will be penalized.
- Any plots (including diagnostic plots) must be appropriately formatted (axis labels, legends where appropriate)
- "Raw" R output and variable names should not be in the report. For example, in the writing, tables, plots, etc, you should say "room type" instead of "room_type"
- The quarto website includes lots of helpful information for generating your report. I recommend that you use the [visual editor](#) to make formatting easier. The visual editor has several features that look like a generic word processor
- [PDF Basics](#)
- [Gallery for advanced Quarto formatting \(not necessary, but could be helpful\)](#)

EXAMPLE OF GOOD FORMATTING

This is a different assignment structure, but notice the following elements of this report:

- code is hidden
- tables/plots are labeled with the variable descriptions instead of the variable names
- sections are labeled
- variables in the text are referred to by description instead of variable name
- All output is presented in text, table, or plot (no raw output)