

Airbnb Reviews Data Analysis — Study Tasks 1

Introduction

The goal of this study is to analyze how data scientists approach solving data analysis problems.

We will ask you to analyze a dataset consisting of Airbnb reviews. To perform the analysis, you will be given access to a Colab Notebook, a type of Jupyter Python notebook that runs on the Google cloud, and have access to AI code generation facilities.

Before you begin

We have summarized [here](#) for you the basic instructions for running Colab, and for using its generative AI features.

We recommend that you read these instructions before approaching the analysis tasks, and that you keep them handy, open on a browser tab for your reference during the user study.

You can access the notebook using this [link](#).

What You Have to Do

You are given a dataset of AirBnb reviews, and you are asked to perform four analysis tasks. Each task produces, as an answer, either a figure or a numeric answer.

Preliminary step – Load the dataset and explore the data

Before starting the analytics tasks, load the Airbnb dataset. Here is a [link](#) to the dataset:

<https://docs.google.com/spreadsheets/d/1a2c9vCMFFfDXmhjoEoX2EwS2lYtbqE4WfZY72TXW9co/edit>

You can tell Colab to load the dataset from that URL.

Explore the dataset to understand its structure, contents, and key characteristics.

An Airbnb listing is a property, room, or experience that is advertised on AirBnb and can be booked by guests; guests can leave reviews on listings they have booked. Each row in the dataset corresponds to a review written by a user for a listing.

Task 1 — High-Quality Listings

Context. You want to identify which listings perform best based on review score.

Questions.

Q1.1. First, determine which property type (e.g., Apartment, Bed & Breakfast, Loft) receives the highest average review score.

Q1.2. Then, focus only on listings with more than 100 reviews. Among those listings, find the 3 that have the highest average review score. Print the IDs of the three listings.

Task 2 — Most Active Reviewers

Context

Some users write many more reviews than others. You want to identify them.

Questions.

Q2.1 Who is the most active reviewer? Specify the reviewer's name.

Q2.2 What is the difference in the number of reviews between the first and the third most active reviewers?

Task 3 — Review Length and Ratings

Context

You want to test whether longer reviews are associated with higher ratings.

Consider only reviews where the property type (e.g., Apartment, Bed & Breakfast, Loft) and review comments are available. The review length is simply the length of the review comments.

Questions:

Q3.1. Compute the Pearson correlation between review length and review rating. Report the correlation coefficient and answer: is the correlation strongly positive? That is, are longer reviews associated with higher ratings?

Q3.2. Now, focus on property types.

- (a) What is the property type with the highest average review length?
- (b) What is the property type with the highest average review rating?
- (c) Display a bar graph giving, for each property type, in blue the average rating, and in red, the average review length.

Task 4 — Trends Over Time

Context

We want to explore how the number of reviews written changes over time, and whether this behavior differs across property types.

Questions

Q4.1. Which combination of month and property type has the highest number of distinct listings reviewed?

Q4.2. Plot the number of reviews each property type receives each month.

Produce a plot with one line per property type; each line indicates the number of reviews received each month by that property type.

From the plot, which property type shows the largest variation in activity over time?

Share your experience

Once you finish, please fill out this form with your experience: <https://forms.gle/hHyXke5pDxazSneP6>