**Sentiment Analysis for the Hospitality Industry**

*By Guilherme Osorio, July 2023*

*Fundamentals of Data Science, University of the Cumberlands*

**Business Understanding**

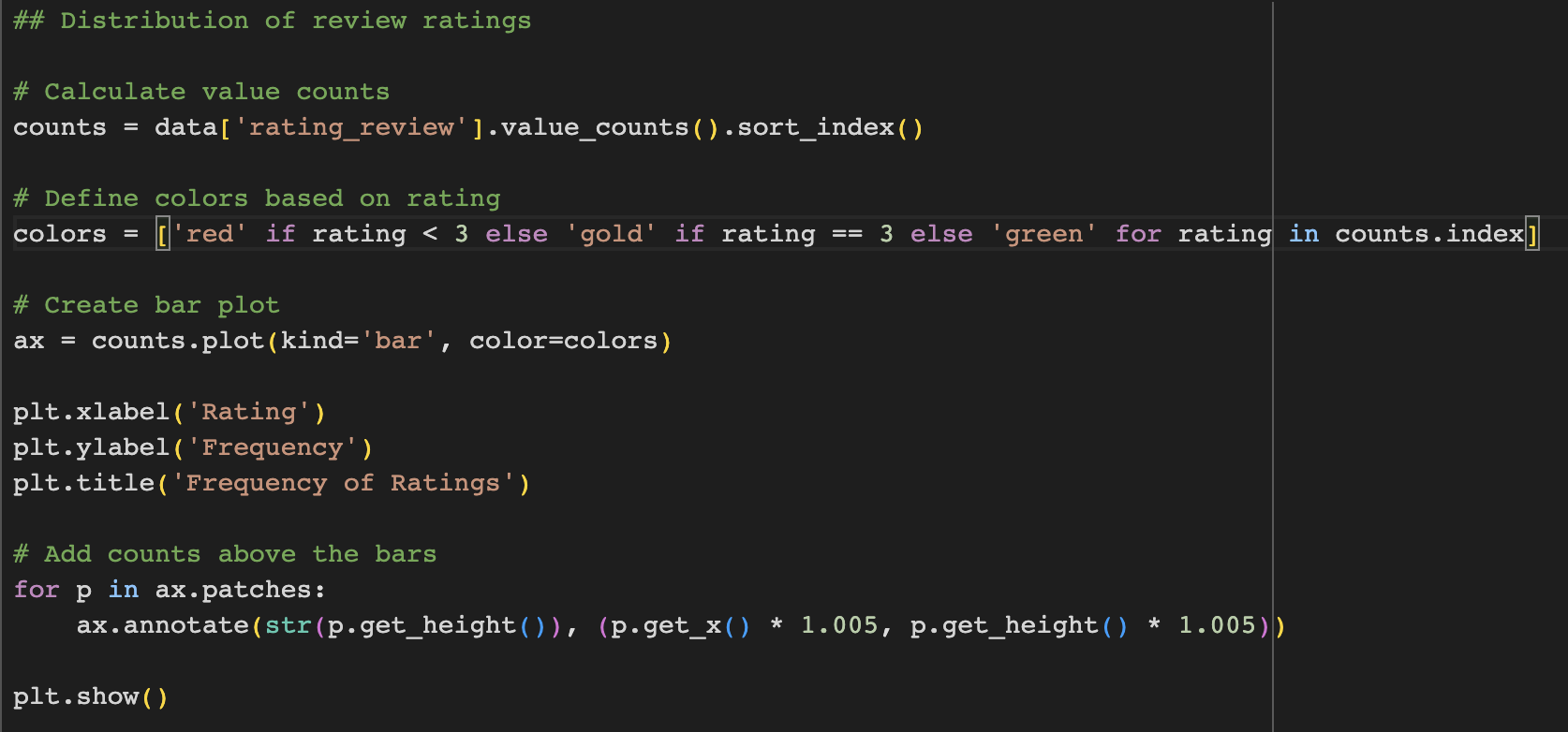
Airbnb is a direct competitor to hotels in the hospitality industry. One way it creates value is by differentiating from hotels. When doing so, it is important to understand what customers like about hotels, and what customers dislike. This dataset of Trip Advisor reviews contains data which could be useful for this analysis – some data exploration could give hints about what is important to guests, why makes their stay enjoyable, and what does not. Airbnb can use this information to develop a differentiation strategy which prioritizes the key aspects customers have in mind. Currently, Airbnb is a strong brand and benefits from a large network economy, all of this without owning significant real estate like hotels do. With insights from this analysis, Airbnb can combine the current competitive advantages it already holds with new characteristics with the goal of expanding its customer base. This will enhance the current advantages outlined above even further, allowing Airbnb to charge premium prices.

The dataset being analyzed contains hotel reviews from customers from the new Delhi region in India. The goal of this analysis is to provide a company like Airbnb with some understanding of what makes customers happy and what makes customers unsatisfied. After the data exploration stage, some pre-trained BERT based models were fine-tuned and tested on the task of sentiment analysis. By automating the sentiment analysis task to a high accuracy, Airbnb could use this model in many ways, such as flagging hosts with constant negative reviews, and rewarding hosts with great reviews (this is just an example, deciding how to use this model would be the next step in the analysis).

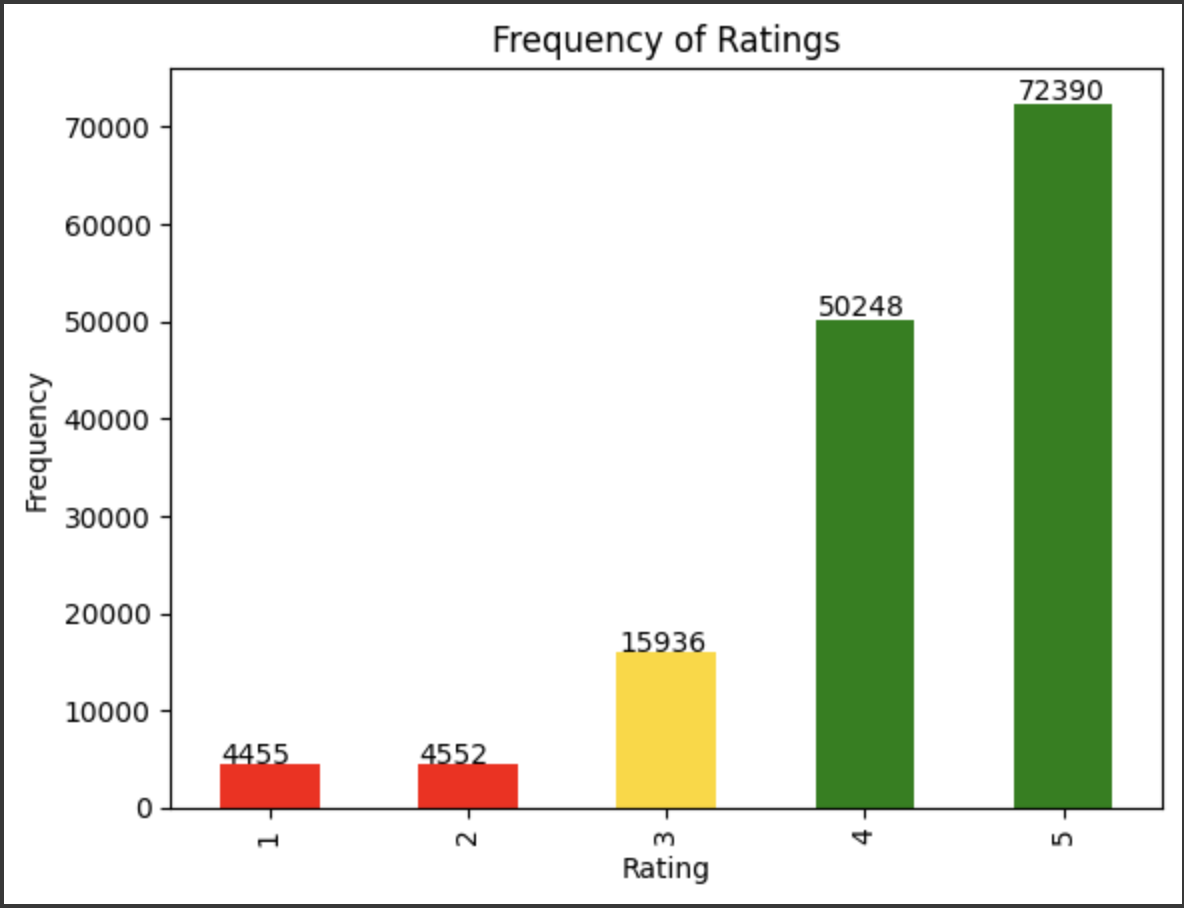
**Data Discovery & Understanding**

To help Airbnb create new competitive advantages and make customers happy, allowing them to charge premium prices, I believe using the dataset found on Kaggle ([here](https://www.kaggle.com/datasets/arnabchaki/tripadvisor-reviews-2023?resource=download)) can be useful.

After loading the data into a Google Colab notebook (this will be useful due to the access to a GPU, which will allow for faster training of the models later), the data exploration stage starts. The original data only has two fields: “rating\_review” (numeric, containing the rating of the review, from 1 to 5) and “review\_full” (containing the full review in text). The dataset has a total of 147581 data points, which should be plenty for our analysis. The describe() method from pandas helps us understand the review column. With a mean of 4.23 and a standard deviation of 0.9, most of the reviews in the dataset seem to be positive, creating some imbalance. To confirm this, the ratings can be visualized through a bar plot:

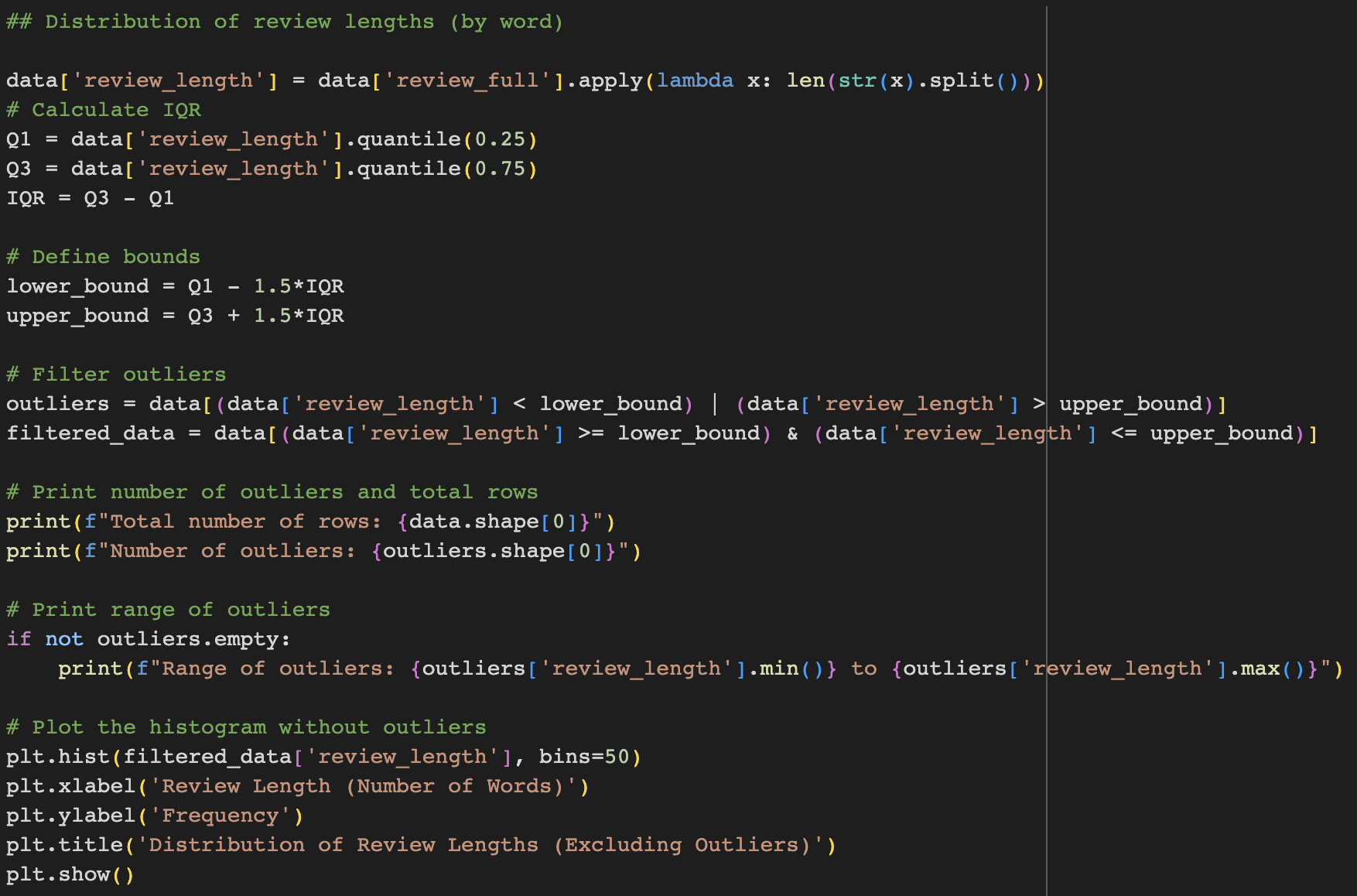


*Figure 1: Python code to produce the “Frequency of Ratings” bar plot*

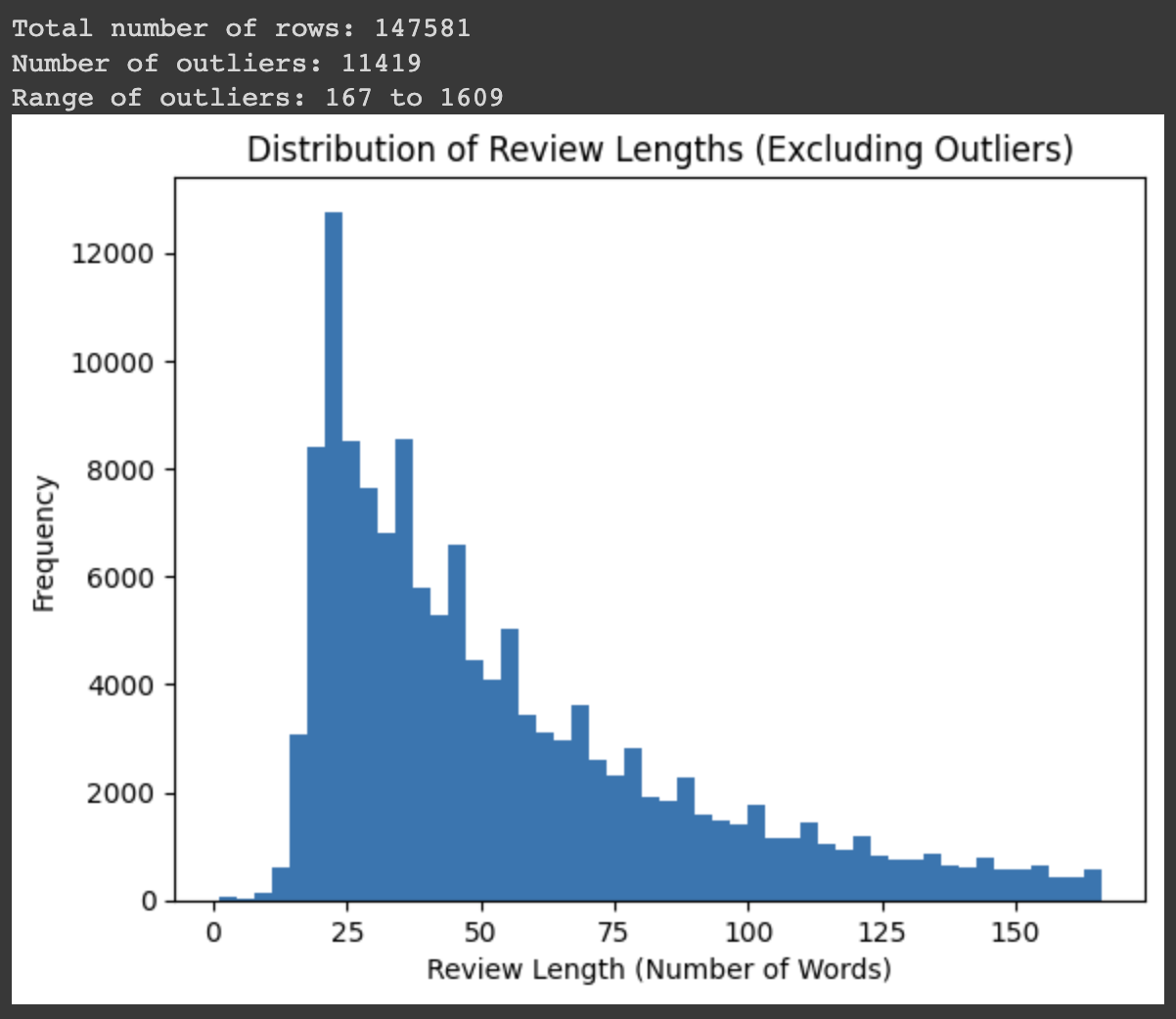


*Figure 2: Frequency of Ratings bar plot*

Next, the data was explored to hypothesize a simple theory: “negative reviews tend to be long, while positive reviews tend to be short”. To figure this out, the first step was to create a new column “review\_length”, with the count of words in “review\_full” and visualize its distribution:



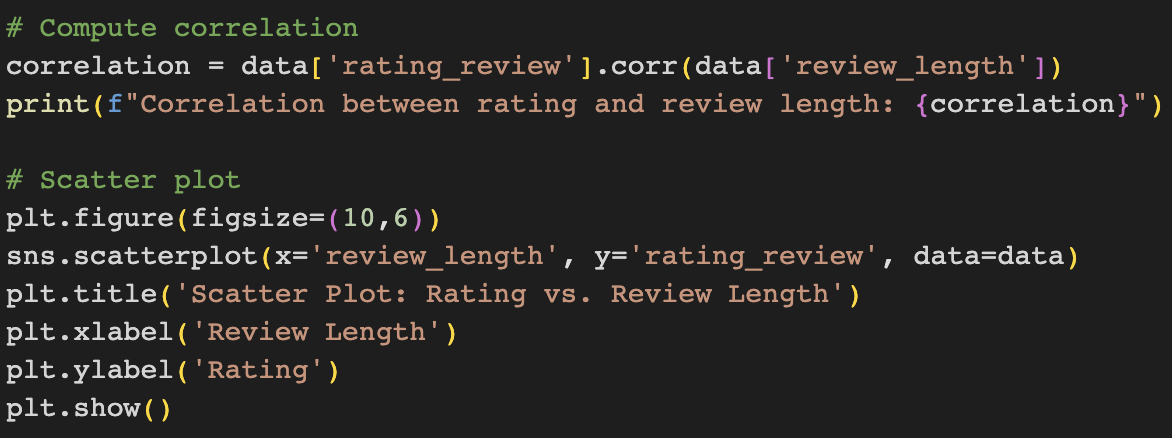
*Figure 3: Code to visualize the distribution of the review lengths*



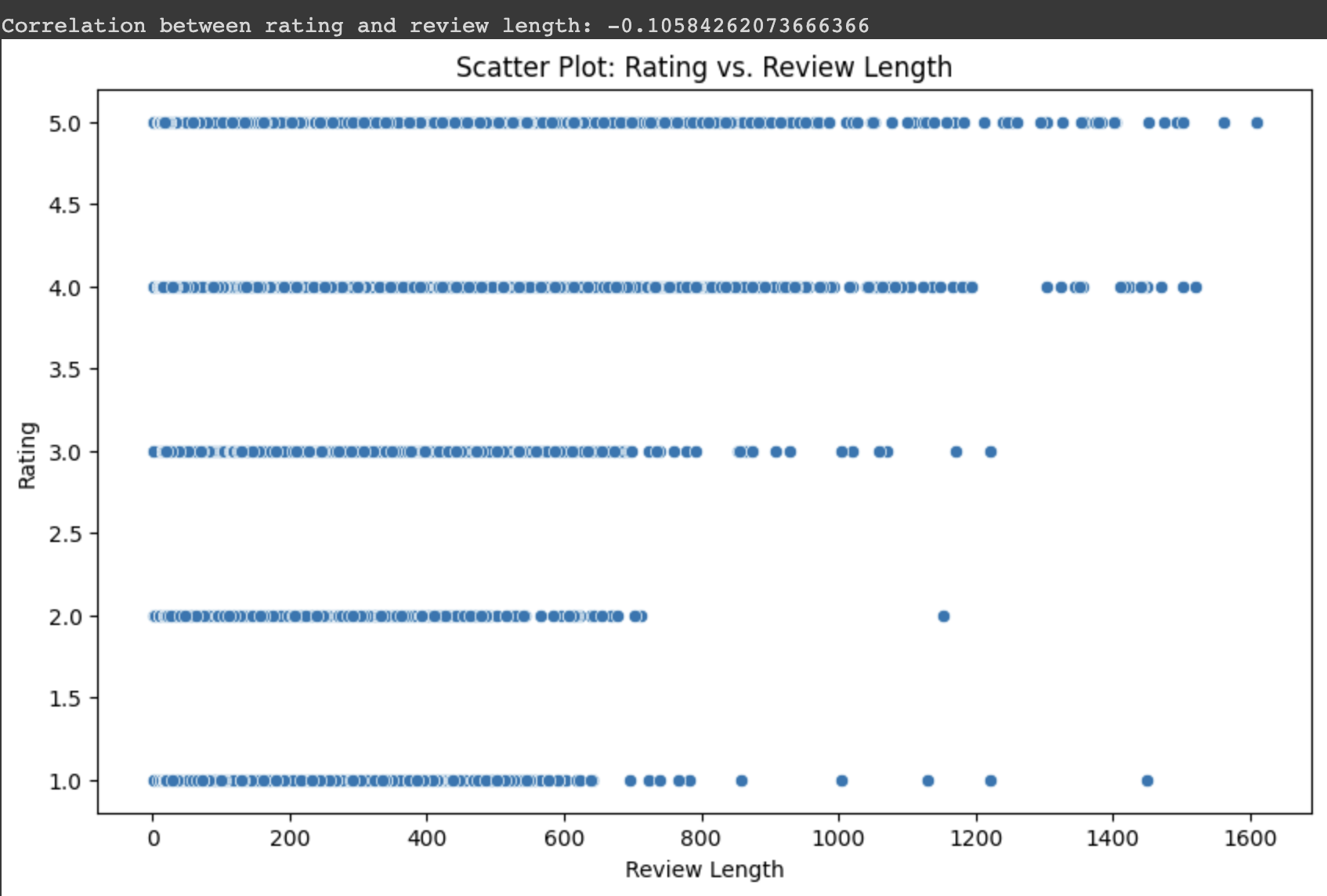
*Figure 4: Distribution of review lengths*

The data shows that this variable is positively skewed, with most reviews being short and some reviews being exceptionally long. A vast majority of the outliers are in the group of long reviews, on the right/positive side of the distribution.

To decide if our hypothesis would be worth testing for, another visualization was created – a scatter plot of the review’s rating on the y-axis and the review’s length on the x-axis -, as well as computing the correlation of the review’s rating with the review’s length.



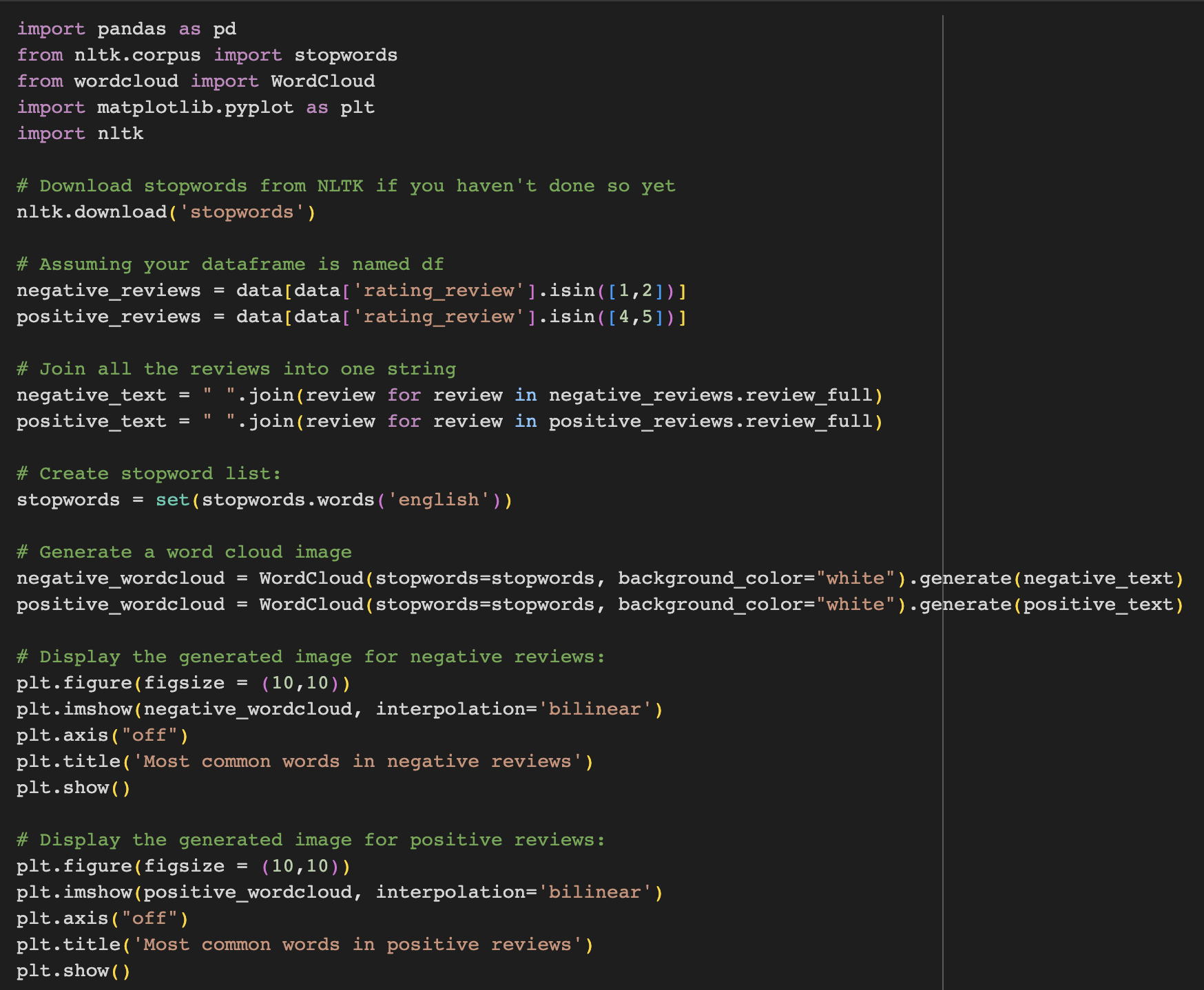
*Figure 5: Code for the scatter plot of rating vs length*



*Figure 6: Scatter plot of rating vs review length*

Given that the correlation between the two variables seems to be negligible, and that the scatter plot does not seem to give much evidence to prove our hypothesis, the relationship between the two variables was not explored further.

The final visualizations done were a word cloud of the most common words in positive reviews and a word cloud with the most common words in negative reviews – for this analysis, positive reviews were considered to be the ones with a rating of 4 and 5, negative reviews were the ones with a rating of 1 and 2, and neutral reviews had a rating of 3.



*Figure 7: Code to produce the word cloud visualizations*



*Figure 8: Word cloud of the most frequently used words in negative reviews*



*Figure 9: Word cloud of the most frequently used words in positive reviews*

The word clouds help give us an overview of some of the topics customers mention the most in both negative and positive reviews. Words such as “place”, “food”, “restaurant”, “service”, “staff” and “ambience” tells us a story: customers value the food in the hotel’s restaurants, the friendliness and assertiveness of the staff and service provided, and the location of the hotel the most when writing a review of their stay. To have a better understanding of the division of topics in the data, a topic modeling analysis could be done with a tool like [BertTopic](https://maartengr.github.io/BERTopic/index.html) in future research.

Once satisfied with the data exploration stage, the data had to be prepared to train the BERT-based models for sentiment analysis. To clean the data, only one step had to be taken: converting non-string type reviews to strings, due to some reviews consisting of empty values.

The next and last step of the analysis was to evaluate the accuracy of a baseline pre-trained Roberta transformer model for sentiment analysis and fine-tuning and evaluating two other models on our data: a fine-tuned version of the Roberta model used as a baseline, and a fine-tuned version of a MiniLM model. The main library used for training was Py Torch, and the models were loaded from Hugging Face.

Fine-tuning of the pre-trained models was done using a training set with 30,000 randomly selected reviews from the original dataset, while 10,000 reviews were used for evaluation. Full code and details of this part of the analysis can be found on the Jupyter notebook provided. The fine-tuned models are also available on the Hugging Face platform: [Roberta fine-tuned](https://huggingface.co/gosorio/robertaSentimentFT_TripAdvisor), [MiniLM fine-tuned](https://huggingface.co/gosorio/minilmFT_TripAdvisor_Sentiment).

To summarize this stage, here are the evaluation metrics for the 3 models:

* Roberta baseline
  + Accuracy: .8724
  + F-1 score: .8484
* Fine-tuned Roberta
  + Accuracy: .9111
  + F-1 score: .9061
* Fine-tuned MiniLM:
  + Accuracy: .9018
  + F-1 score: .8956

**Conclusion & Next Steps**

Considering the explorations and analyses carried out in this study, two prominent features surfaced: the findings from the word cloud visualizations and the development of BERT-based models for sentiment analysis. The word clouds provided an overview of frequently appearing terms in both positive and negative reviews, highlighting key areas of interest for hotel guests such as "place", "food", "restaurant", "service", "staff", and "ambience". This sheds light on what guests value during their stay, thus underlining potential areas for Airbnb to support their differentiation strategy.

From a technical standpoint, the implementation of BERT-based models demonstrated noteworthy results, particularly with the fine-tuned Roberta and MiniLM models. These models classified sentiment with high accuracy, revealing the potential of such approaches in the continuous evaluation of host performance and guest satisfaction within the Airbnb platform.

Looking ahead, the scope of this research can be expanded in several ways. Firstly, topic modeling could be applied as a follow-up to the word cloud analysis. This approach may uncover deeper insights by identifying common and persistent topics present in positive and negative reviews. Secondly, the model developed in this study could be integrated into Airbnb's existing operational pipeline. By doing so, it can serve a multitude of purposes such as recognizing exemplary hosts and those requiring improvement, comprehending customer needs across different regions, pinpointing locations where guests are most and least satisfied, and much more.

In addition, future research could also investigate the relationship between review sentiment and other parameters such as pricing or location. This could contribute to a more granular understanding of guest preferences and experiences.

In sum, this study provides a stepping-stone to further research and application of sentiment analysis in the hospitality industry. Through continual learning and improvement, platforms like Airbnb can utilize such data-driven insights to cultivate superior guest experiences and strengthen their competitive position in the market.