For this project, I downloaded a Kaggle dataset on New York City Airbnb listings. Along with importing the necessary packages and previewing the data, I did some research on the source. Understanding the source and its content is an important step that allows us to determine the quality and make informed decisions when cleaning data. As an Airbnb user, I am generally familiar with information included in listings, which helped me figure out the data quality. For example, records that show zero availability but have reviews don't necessarily mean bad data because there could be hosts who happened to set their property as unavailable temporarily when the data was collected.

When thinking about missing values, I also had to make decisions involving impact on analysis. Considering how many listings don't have reviews, I might want to remove these listings from an analysis specifically looking at reviews, such as finding correlation between reviews per month and number of reviews. While the 0 values are valid here, they would just be unnecessary noise in this analysis. On the other hand, if we were comparing reviews per month with price, the listings with 0 reviews might have lower prices, which would be relevant to our analysis. In that scenario, I would want to keep the 0 values.

One of the challenging steps for me was fuzzy matching. My dataset didn't have an obvious example to use to perform this, so I researched different ways to compare the listing names within the 'name' column. I ended up creating a list from the name column, and then a smaller subset from that to work with in order to extract matches based on a chosen string.

References:

Dgomonov. (2019). New York City Airbnb Open Data. Kaggle. Retrieved from https://www.kaggle.com/dgomonov/new-york-city-airbnb-open-data

Kazil, J. & Jarmul, K. (2016). Data wrangling with Python [Kindle ed]. Sebastopol, CA: O'Reilly Media, Inc.