We used two data sets. One for London restaurants in CSV format, and one for London hotels in XLSX (Excel) format. The data sets came from Kaggle and Data World respectively.

For both datasets we used Jupyter notebook to select the columns we needed and filter out the non-London establishments and non-five-star ratings. The only columns we used were Property Name, Review Rating, and Number of Reviews. We furthered polished the datasets by removing Location, renaming the columns, and exported to csv. One step that was specific to the hotel dataset was the creation of a new column showing the number of reviews per site. We did this by grouping the number of reviews by property name and then did a value count of the review ratings. We had to rename a few columns a second time then export to csv. Then we were able, while still in Jupyter Notebook, to connect to SQL server and create a database there by loading the csv files into MySQL.

Our original proposal stated that we would join the tables by city (London) and rating (5). We did not do this because we did not keep Location as a column and it did not make sense to us to merge them by rating.

Camilla Bahr

Sarah Sprole

Subhra Nandi