Sonja Lindberg & Erica Zhou
Kerbs: sonj, ezhou
Git Repo: https://github.com/ericazhou7/6.s080-labs/
Commit Hash: 21b8c7f652d17aca941482bd5296532d0ba7f068

Lab 5 Write Up

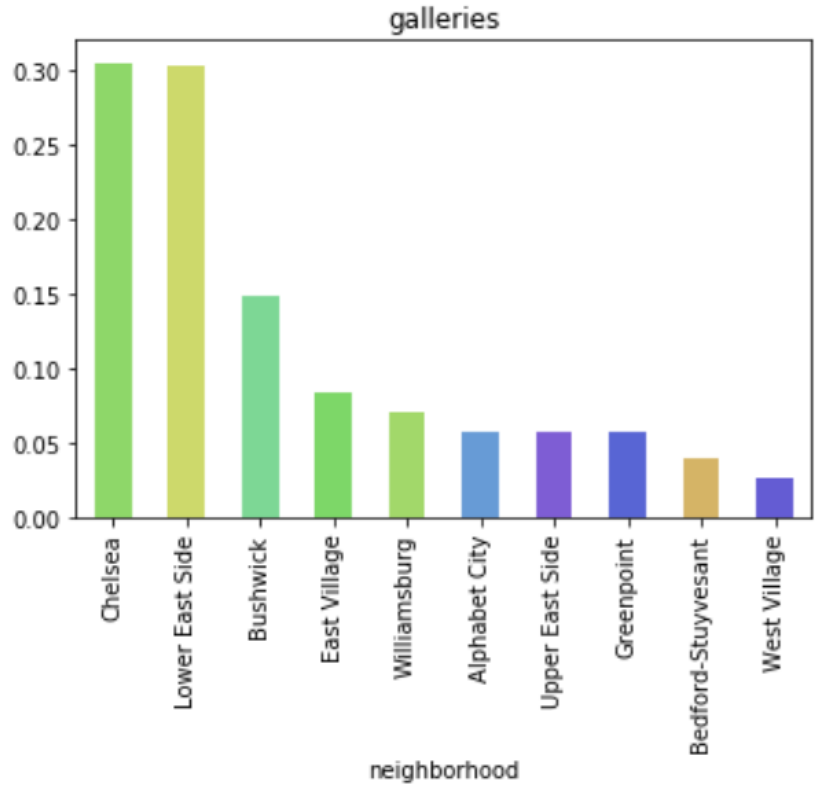
Please view our visualization in **visualizations.pdf** or by running the jupyter notebook **nyc_listings.ipynb**. To run the notebook, you may need to install the following packages: **pandas, numpy, nltk, matplotlib, seaborn, wordcloud.** The data for this visualization was originally found at http://insideairbnb.com/get-the-data.html.
We've also included screenshots of one type of visualizations we did (top 10 neighborhoods for a given search term) at the end of this write up.

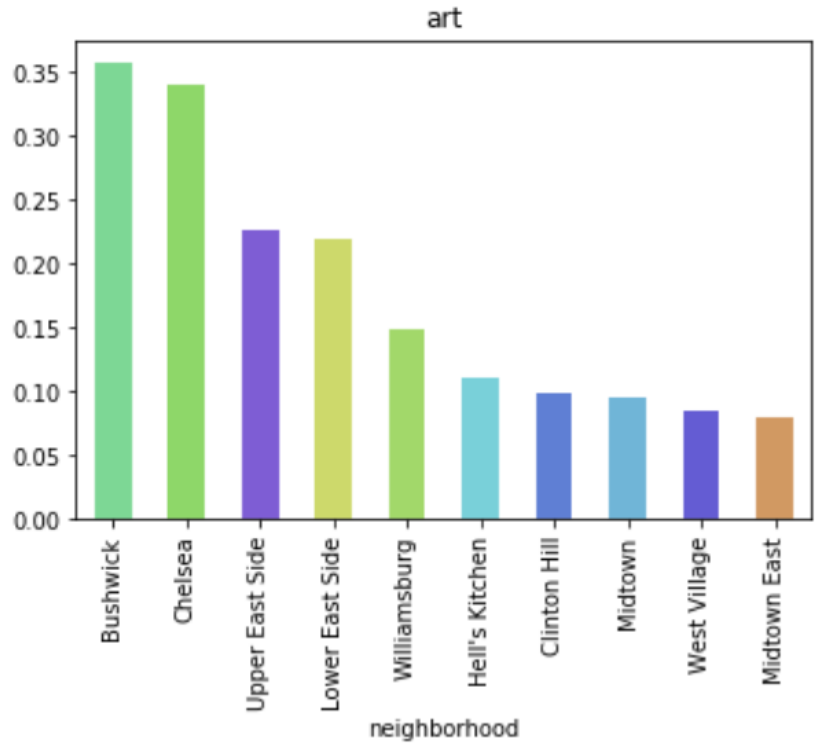
Our visualization shows the frequency with which several terms appear in Airbnb listing neighborhood descriptions. This representation gives an interesting insight into the "flavor" of different neighborhoods, revealing which neighborhoods are known for / described as having great coffee or museums (for example). Other terms like "greek" or "art" reveal the less easily quantifiable identification of neighborhood communities. Our visualization thus uses expert local data (the Airbnb host knowledge) to show neighborhood trends and characteristics for different qualities an NYC visitor might want to find or explore.

We built this visualization in two major steps: cleaning/wrangling and visualization. We began by using the Airbnb listing information and narrowing our focus to the most rich information: the host descriptions. We then transformed these descriptions into strings with in pandas data frames, pre-processing the text to filter out stopwords. When considering what stopwords we had to remove, we started with a basic filter of removing punctuation and then added NYC- and Airbnb-specific stopwords. For example, the NYC subways are often mentioned, and so we filtered out words like "subway" and "train", along with numbers and single letters like "l" or "r" that represented subways lines. We then calculated the frequency of each term per neighborhood, normalized by the ratio of number words per neighborhood to number total words in all the descriptions. These frequencies were used by our visualization method to show bar graphs for each selected term in NYC neighborhoods.

We considered using word clouds as an alternative presentation of term frequency in several selected neighborhoods, but decided that the use case of searching for neighborhoods matching a term was more interesting and realistically more useful as a tool. We also felt that the bar graphs gave better quantitative information.

Sonja Lindberg & Erica Zhou
Kerbs: sonj, ezhou
Git Repo: https://github.com/ericazhou7/6.s080-labs/
Commit Hash: 21b8c7f652d17aca941482bd5296532d0ba7f068





Sonja Lindberg & Erica Zhou
Kerbs: sonj, ezhou
Git Repo: https://github.com/ericazhou7/6.s080-labs/
Commit Hash: 21b8c7f652d17aca941482bd5296532d0ba7f068