CS 432/532 Project 2 Report

The Percentage of Listings on AirBNB in New York, New York Containing a Review with the Host(s)' Name(s), and the Most Reviewed Neighborhoods on AirBNB in New York, New York

Jacob Helhoski

I. NoSQL Queries

Query 1:

For my first NoSQL query, I found the percentage of listings with at least one review that mentioned the host's name. To do this, I used an aggregation within the MongoDB Compass GUI.

My query first removed listings which had no reviews in order to make the following join more optimized. I also had to ensure that the "host_name" field was a string to prevent errors later on. The join was executed using the "$lookup" stage of the aggregation, which matched listings with reviews based on "listing_id". The resulting documents were then "unwound" to produce one document for each review.

In MongoDB, regular expressions cannot be dynamically generated, so I had to use "$indexOfCP" as a workaround, which finds the index of a substring within the review, returning -1 if it is not present. Using this method I could remove listings which did not have reviews mentioning the host from the aggregation. Finally, I grouped the results by listing and counted the number of listings returned, producing the values shown in fig.1.

Query 2:

For my second NoSQL query, I found the ten neighborhoods with the highest number of reviews. This was also done using an aggregation within MongoDB Compass.

I began similarly to my first query, removing listings with no reviews, making any future joins more optimized. I then performed a join between the listings and reviews collections based on matching listing id's, once again using the "$lookup" aggregation stage to facilitate the join, unwinding the resulting documents as well. A "$match" was then used to filter the items by date. The date field of the current documents would be checked to be greater than January 1st of the selected year, and less than December 31st of the selected year. Documents that did not pass this check were removed from the aggregation.

When grouping the documents, I had the choice between grouping by "neighborhood_group", which contained the boroughs of the listing, or "neighborhood", which contained a more specific neighborhood within a borough. I decided to group by the more specific neighborhood, since I felt this would give a more detailed geographic picture of the listings. During my grouping operation, I also produced a count of how many listings were in each neighborhood. The resulting neighborhoods were sorted in decreasing order by this listing count, producing the top ten neighborhood data seen in fig. 2.

II. NoSQL DATABASE AND DATASET

The dataset I used was titled “Airbnb New York - Jan 2024”. I found the dataset on Kaggle.com, and it was posted by Manjit Baishya. The dataset included four files:

1. a data dictionary describing data types and other metadata about the dataset
2. a set of listing data, including information about the neighborhood, geographic coordinates, host, listing name, and number of reviews, along with other data for over 30,000 listings
3. a set of reviews, the largest file, which contained information about reviewer name, the date of the review, and the review itself for over 1,000,000 reviews, all of which were corresponding to the listings in the previous file
4. a geojson file which associated the specific neighborhoods mentioned in the listing information with polygons that outlined the neighborhood and could be used in conjunction with a map program to visualize the neighborhood (this file was not used in my project)

I found the dataset to be relatively reliable, with few null values. The only issues I encountered with it were some mismatched types in the “host_name” field, which caused errors in my queries, and required those documents to be dropped from both aggregations.

For my NoSQL database, I used MongoDB, specifically the Compass GUI framework. A big difficulty in this project was understanding how MongoDB’s aggregation framework operated, as well as reading through MongoDB’s sometimes inconsistent online documentation. I was very satisfied with the performance of the database, although it required some optimization on my part. My first query originally implemented a join on the larger reviews file, which would not finish in a reasonable amount of time. This was fixed by joining on the smaller listings file, and creating indexes for the primary key values in each collection.

III. PROJECT OUTCOME

For query 1, my results, pictured below (fig. 1), show how the proportion of hosts that meet the requirements laid out in section I. Being mentioned in a review is assumed to be a good factor for a listing, as it shows a connection between the guest and the host. Around 55% of listings were found to have this quality.

For query 2, my results, pictured below (fig. 2), show that 3-4 main neighborhoods remained the most popular for the past 10 years. Those included Harlem, Bedford-Stuyvesant, and Williamsburg. Also note the drop in review in 2020, due to the COVID-19 pandemic. Sorting AirBNB ratings by the most reviewed neighborhoods can give someone a general sense of the popularity of staying in that neighborhood, and could correlate to how welcoming to tourists this area may be, making this query useful.

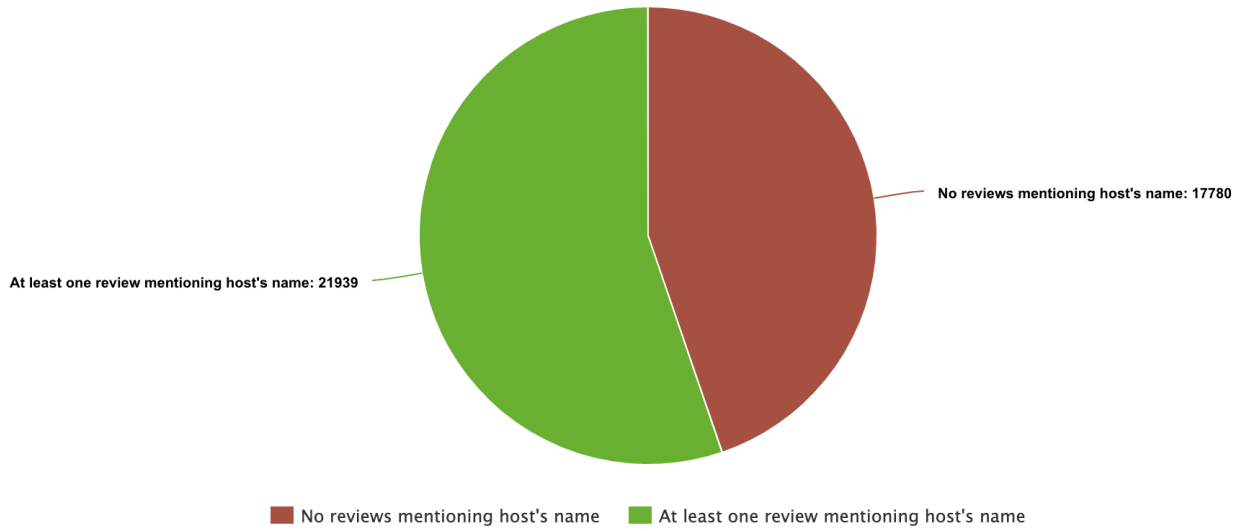

Figure 1: Proportion of listings with at least one review mentioning the host's name

2014		2015		2016		2017		2018	
Bedford	1545	Bedford	3287	Bedford	5619	Bedford	8042	Bedfor	10220
Harlem	1336	Harlem	2345	Harlem	3995	Harlem	5468	Harlem	7128
Williamsburg	1273	Williamsburg	2229	Williamsburg	3754	Williamsburg	4705	Williamsburg	6534
East Village	675	East Village	1282	Crown Heigh	2009	Crown Heigh	2568	Bushwick	3510
Crown Heigh	626	Upper West S	1250	East Village	1901	Bushwick	2369	Crown Heigh	3502
Upper West S	585	Crown Heigh	1186	Upper West S	1836	East Village	2271	Hell's Kitcher	3240
Clinton Hill	493	Bushwick	991	Hell's Kitcher	1732	Hell's Kitcher	2213	East Village	2930
Bushwick	458	Hell's Kitcher	977	Bushwick	1554	Upper West S	2189	Upper West S	2519
East Harlem	413	East Harlem	863	East Harlem	1261	East Harlem	1658	East Harlem	2437
West Village	375	Upper East S	733	Upper East S	1140	Astoria	1644	Astoria	2278
2019		2020		2021		2022		2023	
Bedford	13569	Bedford	5391	Bedford	10856	Bedford	19644	Bedford	20927
Harlem	9002	Midtown	2833	Harlem	5346	Harlem	10350	Harlem	11681
Williamsburg	8153	Williamsburg	2607	Midtown	5170	Williamsburg	9271	Midtown	9709
Bushwick	5205	Harlem	2531	Williamsburg	4744	Midtown	8365	Williamsburg	9537
Crown Heigh	4333	Bushwick	1732	Financial Dist	4463	Bushwick	7452	Crown Heigh	8198
Hell's Kitcher	3962	Hell's Kitcher	1571	Crown Heigh	3813	Crown Heigh	7249	Bushwick	8127
East Harlem	3824	Crown Heigh	1429	Bushwick	3674	Hell's Kitcher	6289	Hell's Kitcher	7748
East Village	3657	East Village	1266	Hell's Kitcher	3003	Financial Dist	4125	Chelsea	5370
Upper West S	2908	SoHo	1200	East Village	2733	East Village	4053	Lower East S	5239
Washington I	2839	East Harlem	1151	Chelsea	2172	Astoria	4031	Astoria	4850

Figure 2: Top ten most reviewed neighborhoods from 2014-2023

REFERENCES

1. https://www.kaggle.com/datasets/manjitbaishya001/airbnb-new-york-jan-2024?select=neighbourhoods.geojson