DS 300 Project 1

1. Dataset Analysis

Column names and data types

- id (integer)
- name (string)
- host_id (integer)
- host_name (string)
- neighbourhood_group (string)
- neighbourhood (string)
- latitude (float)
- longitude (float)
- room_type (string)
- price (integer)
- minimum_nights (integer)
- number_of_reviews (integer)
- last_review (string)
- reviews_per_month (float)
- calculated_host_listings_count (integer)
- availability_365 (integer)

Analysis of quantitative variables
While id, host_id, latitude, and longitude are technically quantitative, in most situations it does not make sense to find certain statistics (mean, max, etc). As such they are left out of this analysis.

Variable	**Min**	**Mean**	**Median**	**Max**
price	0.00000	152.720687	106.00000	10000.00000
minimum_nights	1.00000	7.029962	3.00000	1250.00000
number_of_reviews	0.00000	23.274466	5.00000	629.00000
reviews_per_month	0.01000	1.373221	0.72000	58.50000
calculated_host_listings_count	1.00000	7.143982	1.00000	327.00000
availability_365	0.00000	112.781327	45.00000	365.00000

From this table, it would seem that most values for these variables would land between the mean and median value of the entire dataset.

Analysis of qualitative variables

- As name is the name of the listing, it is generally a unique value for every listing.
- A given host_name may not be unique within itself but host_id distinguishes multiple hosts with the same name.

- neighbourhood_group is the boroughs of New York and so only has 5 possible values: Bronx, Brooklyn, Manhattan, Queens, and Staten Island.
- neighbourhood is the specific area within neighbourhood_group and may be unique within the dataset.
- room_type only has 3 possible values: Entire house/apt, Private room, or Shared room.
- last_review is the date of the latest review for that booking. Values may not be unique given the number of listings in the entire dataset but it is possible.

Preserved utility
After anonymization, we should be able to look for aggregate data such as mean, median, etc. on numerical variables as shown in table above. Queries about specific information about some qualitative variables, such as room_type or availability_365 should still be accessible.

2. Identification of Private Information

Explicit Identifiers:

- id - This is the listing id and is unique to every listing
- name - This is the actual name of the listing on airBnb

Quasi Identifiers:

- host_id - An anonymized label for each host. Can be used to identify a host
- host_name - Even though this is only the first name of the host, it can be used to identify a listing if a name is unique
- latitude and longitude - Together they can pinpoint the exact location of the listing
- neighbourhood_group - Can help de-anonymize the dataset with additional information
- neighbourhood - Similar to above only it’s a little more specific with the location
- number_of_reviews - This is publicly available information that could potentially be used to de-anonymize the dataset when used in conjunction with other quasi identifiers.
- last_review - Same as above: publically available so can be used for identification if unique for a listing
- price - It is necessary to have this available somehow without revealing the exact listing if unique
- minimum_nights - Same as above, publically available so can be used for identification

Sensitive Information

- calculated_host_listings_count - The number of rentals the host has. I would consider this private information that the host wouldn’t want other people to know.

Non-sensitive Information:

- room_type - The type of rental available to users
- reviews_per_month - Not much use for de-anonymization since the time frame of the reviews isn’t known.
- availability_365 - Assuming there is a lag between when this value is entered and when this value is reported then since this value changes every day, I wouldn’t consider it sensitive

3. Anonymization Approach

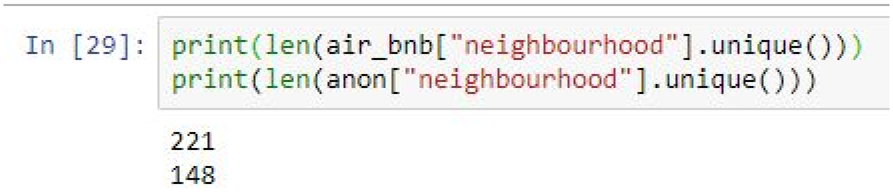
The first thing our anonymization approach does is to suppress explicit identifiers and redundant quasi-identifiers. The explicit identifiers id and name were suppressed to ***** for all values. Since latitude and longitude are redundant because neighbourhood can replace the two values with one value that is as accurate, they were also suppressed. Lastly, host_name was suppressed because it cannot be generalized in a simple manner because of many unique values.

Then quasi identifier host_id was hashed so there is a non obvious link between a host's id and name and his or her listing(s). last_review was also generalized to month to prevent a unique date from revealing a listing. Since there were many values for neighbourhood, it was anonymized to be 15-anonymous, with unique values becoming "other".

Lastly, price, minimum_nights, and number_of_reviews had some noise placed on it using a laplace distribution with each variable's standard deviation. This creates noise for each value so a query will return a value close to the real value without constantly getting a different answer each time.

4. Analysis of Utility Loss

- While the original dataset had 221 unique neighborhoods, some were grouped into the other category because of a lack of data, resulting in only 148 unique neighborhoods after anonymization



```
In [29]: print(len(air_bnb["neighbourhood"].unique()))
         print(len(anon["neighbourhood"].unique()))

         221
         148
```

- Seeing the change in distribution can be done by repeating the variable statistics for numerical variables.

Variable	Min	Mean	Median	Max
price	-9.377066	152.717607	107.516903	10010.0
minimum_nights	-26.755354	7.021309	3.229610	1243.0
number_of_reviews	-43.986851	23.290540	6.966905	629.0
reviews_per_month	0.010000	1.373221	0.720000	58.0
calculated_host_listings_count	1.000000	7.143982	1.000000	327.0
availability_365	0.000000	112.781327	45.000000	365.0

Overall, the original distribution is the same with a few shifts in values and some impossible negative values. This is especially true for reviews per month and days available which we left untouched.

- The listings' exact coordinates were removed so detail is now only available at the neighborhood level instead of the street level.

- The data can still be analyzed by owner through the hashed host ID, although the identity of the listings and the owners are masked