Final Project

Airbnb Recommendations

Student: Trinh Dinh Phuc

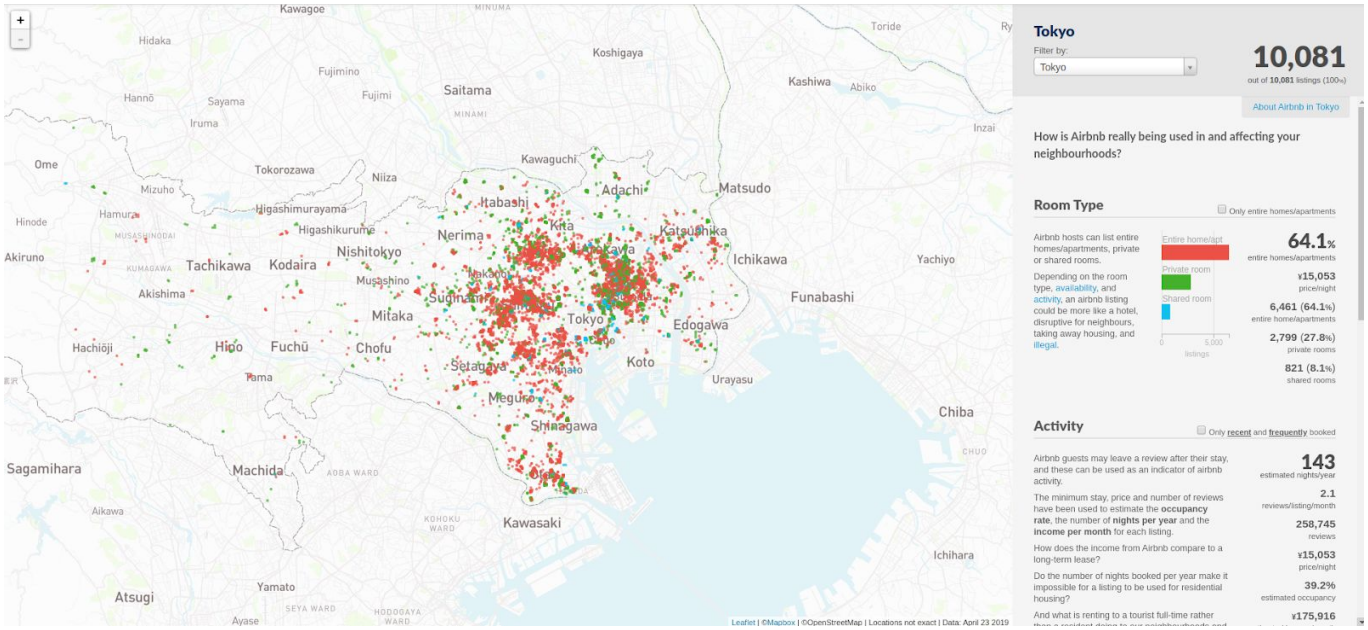
ID: 1101-949-014

OUTLINE:

1. Introduction

For the uninitiated, Airbnb is an internet marketplace for short-term home and apartment rentals. It allows you to, for example, rent out your home for a week while you're away, or rent out your spare bedroom to travelers. The company itself has grown rapidly from its founding in 2008 to a valuation near US$40 billion and is currently worth more than any hotel chain in the world.

In this work, I used the ALS machine-learning model to recommend suitable neighborhoods for every customer based on their reviews. The final input is described as follows: RDD[Rating], Rating(int customer_id, int neighbourhood_id, double rating).



The density of Airbnb providers in Tokyo.

2. Dataset description

Dataset Source: http://insideairbnb.com/get-the-data.html

Data	Description
listings.csv	Detailed Listings data for Tokyo
listings_summary.csv	Summary information and metrics for listings in Tokyo (good for visualizations).
neighbourhoods.csv	Neighborhood list for geo-filter. Sourced from city or open-source GIS files.
reviews.csv	Detailed Review Data for listings in Tokyo
reviews_summary.csv	Summary Review data and Listing ID (to facilitate time-based analytics and visualizations linked to a listing).

"Listings.csv" Detailed listings data for Tokyo

```
val PATH_LISTINGS_DETAIL = "/home/harry/Documents/Airbnb/Airbnb_Japan/listings.csv"
//  +------------------+--------------------+--------------------+-----------+--------------------+--------------------+--------------------+--------------------+
//  |                id|         listing_url|          scrape_id|last_scraped|                name|             summary|               space|         desc
//  +------------------+--------------------+--------------------+-----------+--------------------+--------------------+--------------------+--------------------+
//  |             35303|https://www.airbn...|      20200428053647|  2020-04-28|La Casa Gaienmae ...|This shared flat ...|This apartment is...|This shared f
//  |I have been using...|                food| wine and cheeese!  |       null|                null|                null|                null|
//  +------------------+--------------------+--------------------+-----------+--------------------+--------------------+--------------------+--------------------+
//  only showing top 2 rows
```

“Listings_summary.csv” Summary information and metrics for listing in Tokyo (good for visualizations)

```
val PATH_LISTINGS = "/home/harry/Documents/Airbnb/Airbnb_Japan/listings_summary.csv"
//   +------+--------------------+-------+--------------+--------------------+-------------+--------+---------+---------------+-----+--------------+------
//   |    id|                name|host_id|     host_name|neighbourhood_group|neighbourhood|latitude|longitude|      room_type|price|minimum_nights|number
//   +------+--------------------+-------+--------------+--------------------+-------------+--------+---------+---------------+-----+--------------+------
//   | 35303|La Casa Gaienmae ...| 151977|        Miyuki|                null|   Shibuya Ku|35.67152|139.71203|   Private room| 4183|            28|
//   |197677|Oshiage Holiday A...| 964081|Yoshimi & Marek|                null|    Sumida Ku|35.71721|139.82596|Entire home/apt|11048|             3|
//   +------+--------------------+-------+--------------+--------------------+-------------+--------+---------+---------------+-----+--------------+------
// only showing top 2 rows
```

“Neighbourhood.csv” Neighbourhood list for geo-filter. Source from city or open-source GIS files.

```
val PATH_NEIGHBOURHOOD = "/home/harry/Documents/Airbnb/Airbnb_Japan/neighbourhoods.csv"
//    +--------------------+-------------+
//    |neighbourhood_group|neighbourhood|
//    +--------------------+-------------+
//    |                null|    Adachi Ku|
//    |                null|  Akiruno Shi|
//    +--------------------+-------------+
// only showing top 2 rows
```

“Reviews.csv” Detailed review data for listing in Tokyo

```
val PATH_REVIEWS_DETAIL = "/home/harry/Documents/Airbnb/Airbnb_Japan/reviews.csv"
//  +--------------------+--------------------+----------+----------+-------------+--------------------+
//  |          listing_id|                  id|      date|reviewer_id|reviewer_name|            comments|
//  +--------------------+--------------------+----------+----------+-------------+--------------------+
//  |               35303|              810980|2011-12-28|    1502908|        Firuz|Miyuki's has been...|
//  |Her place is very...|Harajuku stn and ...|      null|      null|         null|                null|
//  +--------------------+--------------------+----------+----------+-------------+--------------------+
// only showing top 2 rows
```

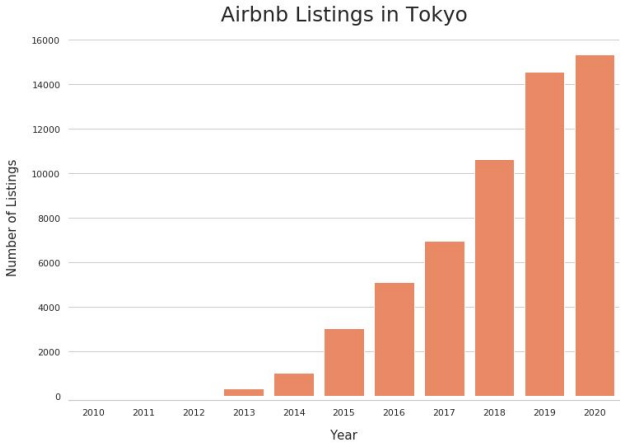
“Reviews_summary.csv” summary review data and listing ID (to facilitate time-based analytics and visualizations linked to a listing)

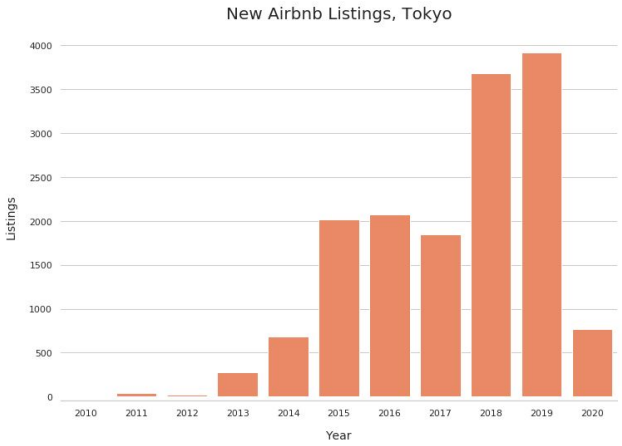
```
val PATH_REVIEWS = "/home/harry/Documents/Airbnb/Airbnb_Japan/reviews_summary.csv"
//    +----------+-------------------+
//    |listing_id|               date|
//    +----------+-------------------+
//    |     35303|2011-12-28 00:00:00|
//    |     35303|2012-10-01 00:00:00|
//    +----------+-------------------+
// only showing top 2 rows
```

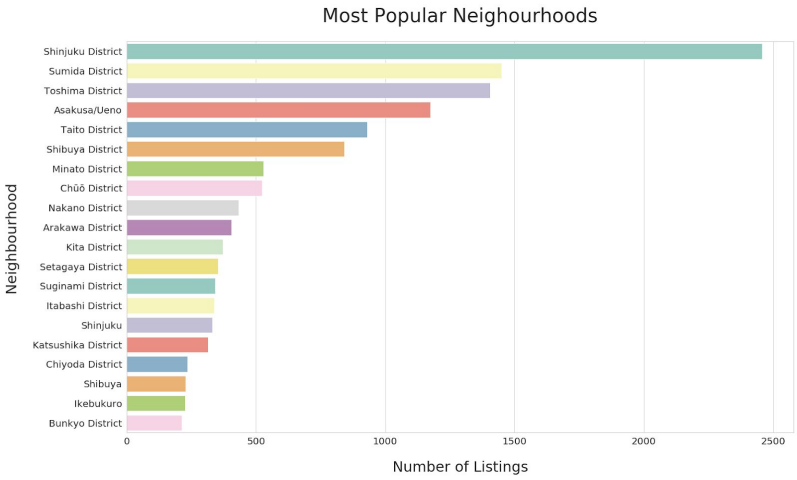
The data shape before being train is 300.000 rows.

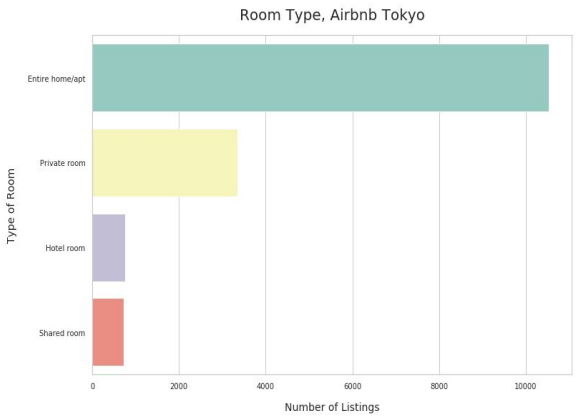
3. EDA - Exploratory Data Analysis

Airbnb listings over the years (predicted - left, actual -right)

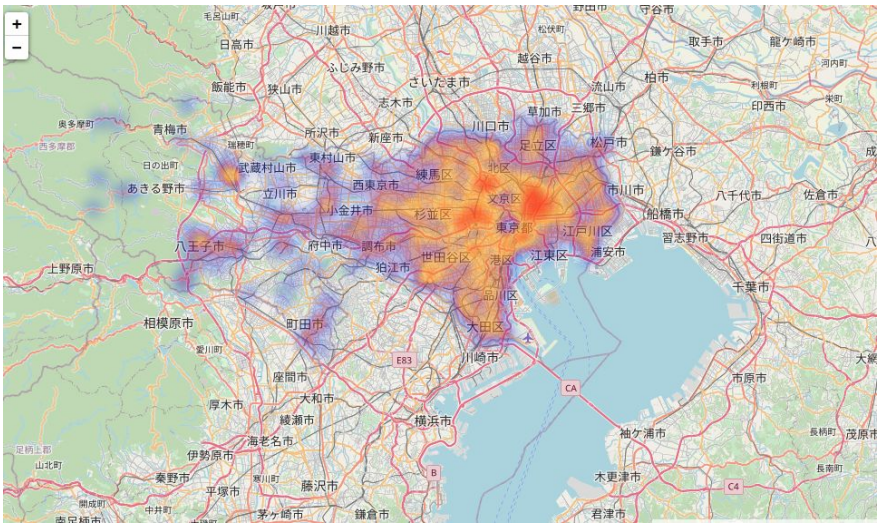




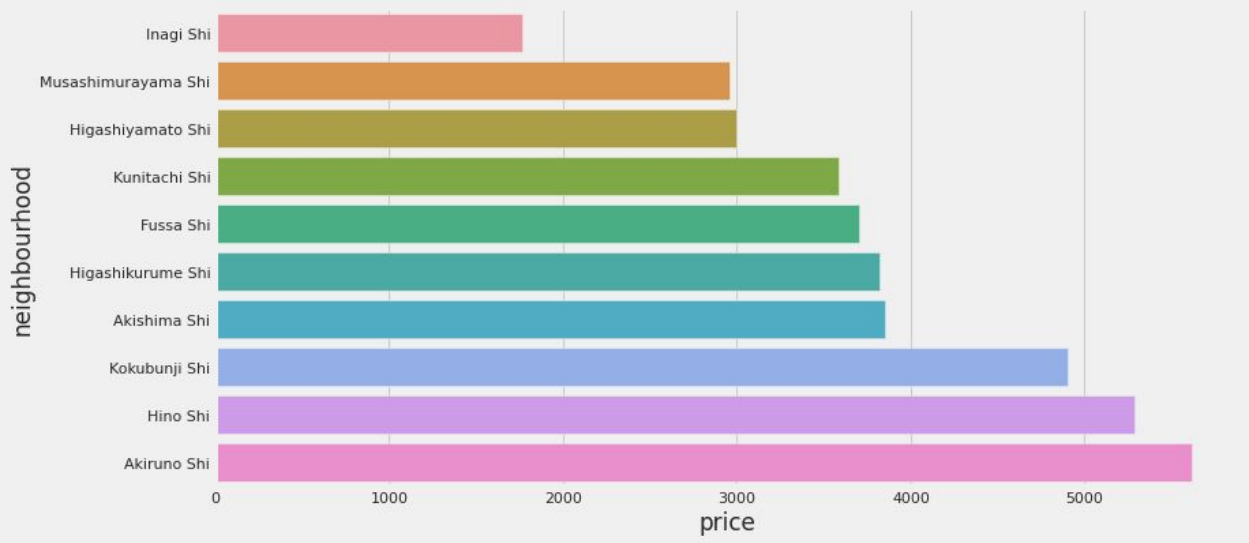




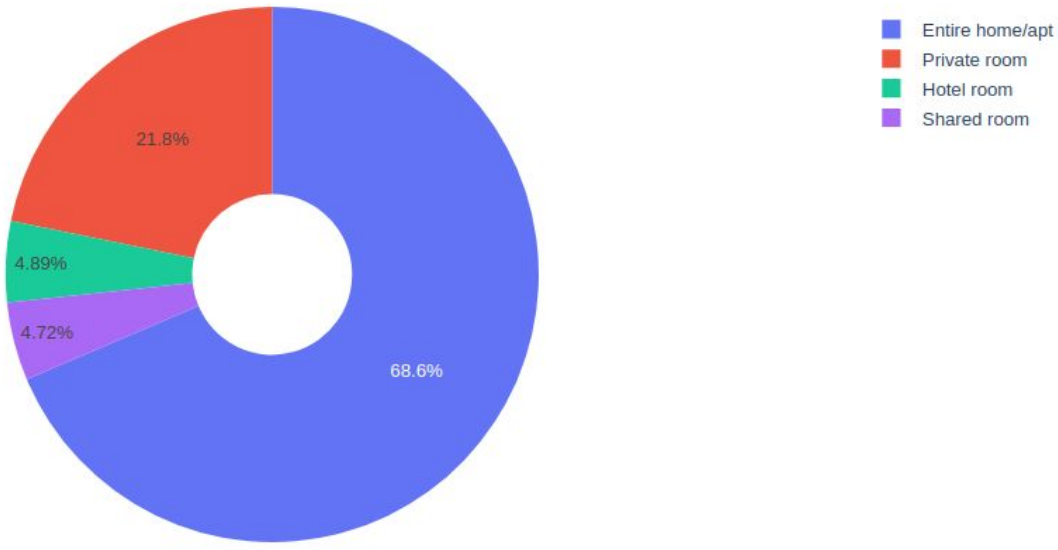
The highest Density areas are marked in red and lowest density areas are marked in blue color.



Categorizing based on Price

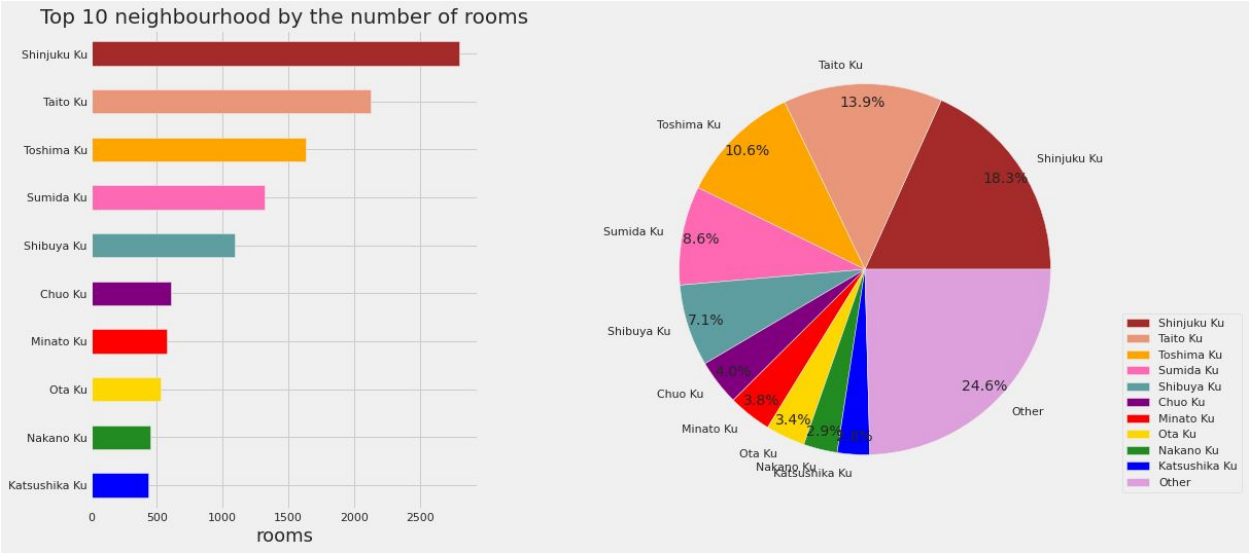


We can see that entire home apartment has highest share followed by private room and least prefered is shared room

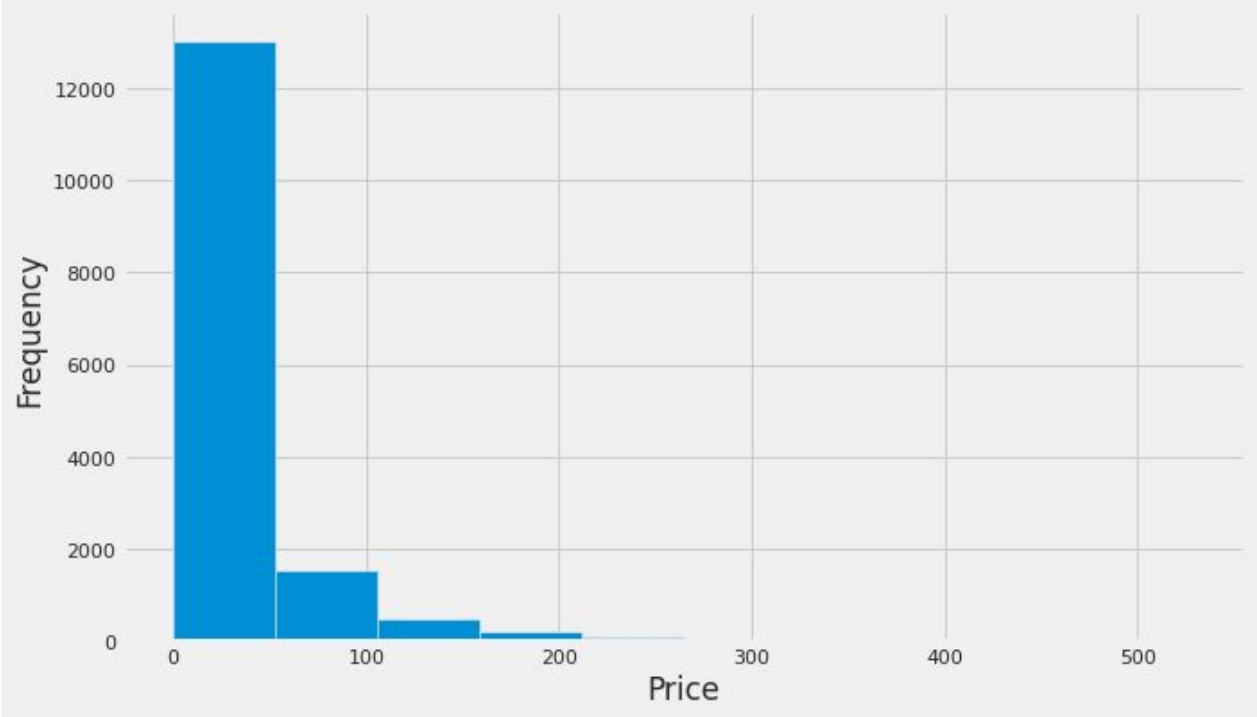


Our word cloud shows the words that are more often used in the Name of this listings

The 10 most expensive neighborhoods to book on Airbnb

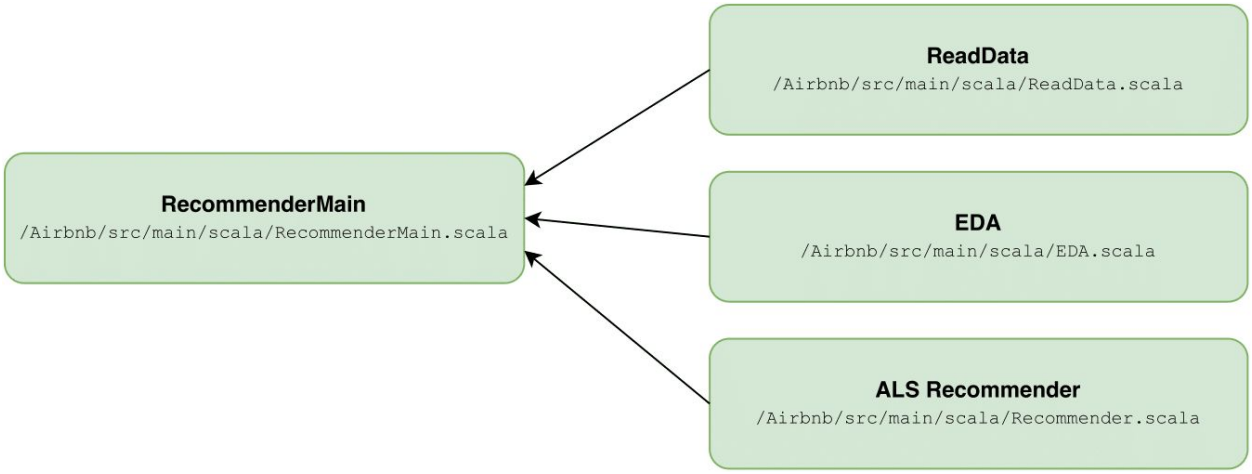


We can see that low-cost rooms or in range 0-50 $ have more reviews.



Rooms with the most number of reviews

4. Workflow

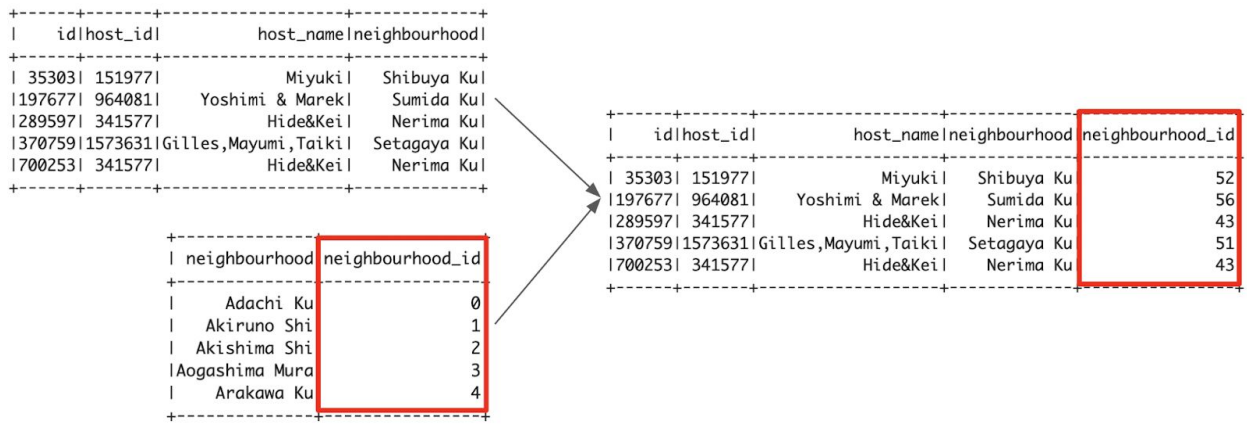


Look at the workflow figure above, we have **"RecommenderMain"** is the main function of the scala project which includes 3 different objects **"ReadData", "EDA", "ALS Recommender".** We summarize the functions of each sub-functions as follows:

- **ReadData:** This class is used to read data and transforms Dataframe into Map[Long, String], and includes 4 functions: *loadReviewsDetail, loadListings, getNeighbourhoodMap, getReviewerMap.*
- **EDA:** This class is used to clean missing values in the Dataframe.
- **ALS Recommender:** Includes *getRating* function and ALS machine-learning model.

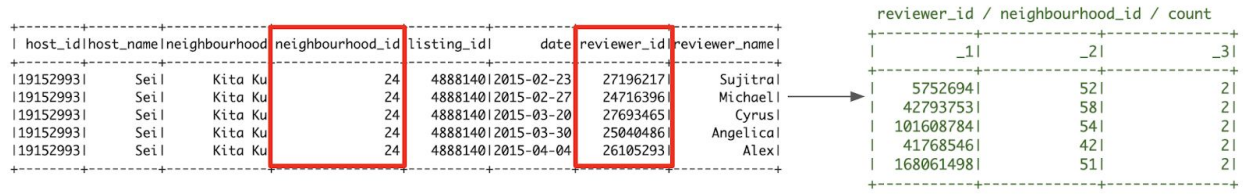
5. Summary of Results

Joins the accommodation data frame and the local data frame to generate a data frame with the local ID.

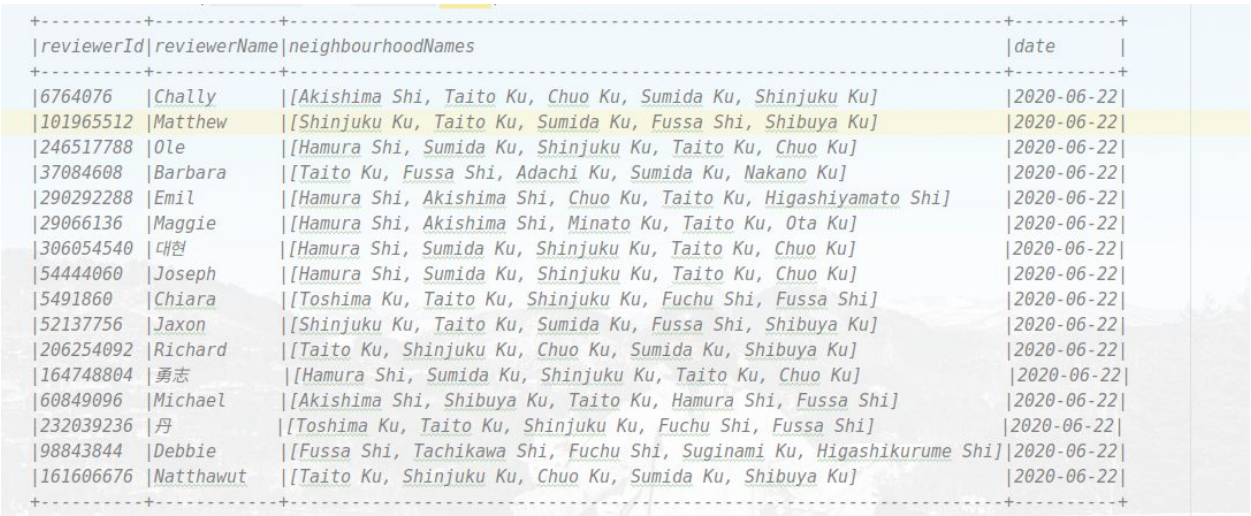


Add customer information to the DataFrame generated in this way by joining the history of accommodation use by each customer.

For each customer, groupBy, and count which area they visited.



Create travel destination recommendation locations by customers using Spark Mllib's recommendation algorithm model (ALS). Finally, the following customer-specific recommended data can be obtained.



reviewerId	reviewerName	neighbourhoodNames	date
6764076	Chally	[Akishima Shi, Taito Ku, Chuo Ku, Sumida Ku, Shinjuku Ku]	2020-06-22
101965512	Matthew	[Shinjuku Ku, Taito Ku, Sumida Ku, Fussa Shi, Shibuya Ku]	2020-06-22
246517788	Ole	[Hamura Shi, Sumida Ku, Shinjuku Ku, Taito Ku, Chuo Ku]	2020-06-22
37084608	Barbara	[Taito Ku, Fussa Shi, Adachi Ku, Sumida Ku, Nakano Ku]	2020-06-22
290292288	Emil	[Hamura Shi, Akishima Shi, Chuo Ku, Taito Ku, Higashiyamato Shi]	2020-06-22
29066136	Maggie	[Hamura Shi, Akishima Shi, Minato Ku, Taito Ku, Ota Ku]	2020-06-22
306054540	대현	[Hamura Shi, Sumida Ku, Shinjuku Ku, Taito Ku, Chuo Ku]	2020-06-22
54444060	Joseph	[Hamura Shi, Sumida Ku, Shinjuku Ku, Taito Ku, Chuo Ku]	2020-06-22
5491860	Chiara	[Toshima Ku, Taito Ku, Shinjuku Ku, Fuchu Shi, Fussa Shi]	2020-06-22
52137756	Jaxon	[Shinjuku Ku, Taito Ku, Sumida Ku, Fussa Shi, Shibuya Ku]	2020-06-22
206254092	Richard	[Taito Ku, Shinjuku Ku, Chuo Ku, Sumida Ku, Shibuya Ku]	2020-06-22
164748804	勇志	[Hamura Shi, Sumida Ku, Shinjuku Ku, Taito Ku, Chuo Ku]	2020-06-22
60849096	Michael	[Akishima Shi, Shibuya Ku, Taito Ku, Hamura Shi, Fussa Shi]	2020-06-22
232039236	丹	[Toshima Ku, Taito Ku, Shinjuku Ku, Fuchu Shi, Fussa Shi]	2020-06-22
98843844	Debbie	[Fussa Shi, Tachikawa Shi, Fuchu Shi, Suginami Ku, Higashikurume Shi]	2020-06-22
161606676	Natthawut	[Taito Ku, Shinjuku Ku, Chuo Ku, Sumida Ku, Shibuya Ku]	2020-06-22

6. **References**:

Dataset Source: http://insideairbnb.com/get-the-data.html

Readers can find source code for more details.

https://github.com/TrinhDinhPhuc/AirbnbPredictionWithSpark