Un-cleaning / Creating Raw data

TASK B

Ahmad Saud Azmi -
1108047

Data Source

- The data set is downloaded from: https://www.kaggle.com/datasets/dgomonov/new-york-city-airbnb-open-data

- The data set describes the listing activity and metrics in NYC for 2019. It includes information about hosts, geographical availability, reviews, and rating. It has 16 columns that provide data about host id, host name, latitude and longitude, reviews, room type, availability of 365 days, etc.

id	name	host_id	host_name	neighbourhood_group	neighbourhood	latitude	longitude	room_type	price	minimum_nights	number_of_reviews	last_review	reviews_per_month	calculated_host_listings_count	availability_365
2539	Clean & quiet apt home	2787	John	Brooklyn	Kensington	40.64749	-73.97237	Private room	149	1	9	19-10-2018	0.21	6	365
2595	Skylit Midtown Castle	2845	Jennifer	Manhattan	Midtown	40.75362	-73.98377	Entire home/a	225	1	45	21-05-2019	0.38	2	355
3647	THE VILLAGE OF HARLEM	4632	Elisabeth	Manhattan	Harlem	40.80902	-73.9419	Private room	150	3	0			1	365
3831	Cozy Entire Floor of Brov	4869	LisaRoxan	Brooklyn	Clinton Hill	40.68514	-73.95976	Entire home/a	89	1	270	05-07-2019	4.64	1	194
5022	Entire Apt: Spacious Stu	7192	Laura	Manhattan	East Harlem	40.79851	-73.94399	Entire home/a	80	10	9	19-11-2018	0.1	1	0
5099	Large Cozy 1 BR Apartme	7322	Chris	Manhattan	Murray Hill	40.74767	-73.975	Entire home/a	200	3	74	22-06-2019	0.59	1	129
5121	BlissArtsSpace!	7356	Garon	Brooklyn	Bedford-Stuy	40.68688	-73.95596	Private room	60	45	49	05-10-2017	0.4	1	0
5178	Large Furnished Room N	8967	Shunichi	Manhattan	Hell's Kitchen	40.76489	-73.98493	Private room	79	2	430	24-06-2019	3.47	1	220
5203	Cozy Clean Guest Room	7490	MaryEllen	Manhattan	Upper West S	40.80178	-73.96723	Private room	79	2	118	21-07-2017	0.99	1	0
5238	Cute & Cozy Lower East	7549	Ben	Manhattan	Chinatown	40.71344	-73.99037	Entire home/a	150	1	160	09-06-2019	1.33	4	188

Quality Dimensions

The following quality dimensions were kept in mind:

- Validity
- Consistency
- Conformity
- Completeness
- Uniqueness
- Accuracy

Quality Dimensions

❑ **Completeness**:

- To make the data incomplete the Staten Island neighbourhood group was filtered.
- Some of the columns, such as latitude, number of reviews, and number of reviews per month, have been left blank.

❑ **Consistency:**

- To make data inconsistent selected rows by filtering the neighbourhood group Queens.
- In some rows, the unique column id is replaced with duplicate values.
- Instead of DD-MM-YYYY, the date format was changed to MM-DD-YYYY.

Quality Dimensions

❑ **Accuracy**:

- To make data inaccurate selected rows by filtering the neighbourhood group Bronx.
- Replaced Lat, long with wrong co-ordinates.
- Replaced neighbourhood “Queen” spelling with “Qeens”

❑ **Validity:**

- To make data invalid selected rows by filtering the neighbourhood group Manhattan for neighbourhood Lower East Side
- Replaced price with negative value
- To make data invalid selected rows by filtering the neighbourhood group Bronx.
- Replaced Minimum Nights with negative values
- Replaced availability_365 columns by add few values (total number of days in a year more than 365)

Quality Dimensions

❑ **Uniqueness**:

- To make data not unique selected rows by filtering the neighbourhood group Queens.
- Unique column id is replaced with duplicate values in all the rows.

❑ **Conformity:**

- To remove conformity selected rows by filtering the neighbourhood group Queens.
- Replaced the date with MM-DD-YYYY instead of DD-MM-YYYY.

Recipe/ Steps

Filter

0. Create project
1. Text transform on 373 cells in column latitude: grel:value.replace(value,'')
2. Text transform on 373 cells in column number_of_reviews: grel:value.replace(value,'')
3. Text transform on 314 cells in column reviews_per_month: grel:value.replace(value,'')
4. Text transform on 939 cells in column last_review: grel:value.toDate('yyyy-mm-dd','dd-MM-yyyy').toString('dd-mm-yyyy')
5. Text transform on 362 cells in column id: grel:value.replace(value,'39593')
6. Text transform on 1091 cells in column latitude: grel:value+(toNumber(2*(row.index)))
7. Text transform on 1091 cells in column longitude: grel:value+(toNumber(2*(row.index)))
8. Text transform on 1091 cells in column availability_365: grel:toNumber(value)+toNumber(80)
9. Text transform on 1091 cells in column minimum_nights: grel:toNumber(value)*-1)
10. Text transform on 911 cells in column price: grel:toNumber(value)*-1
11. Text transform on 4574 cells in column last_review: grel:value.replace(value,"dates")
12. Text transform on 537 cells in column neighbourhood_group: grel:value.replace(value[1],'')

File:

Unclean_step.json

Calculations

- 48895 rows excluding header were present in the document
- Around 26 % of entire dataset i.e., 12713 records /data were either removed or applied variations into it. 1-1.5 % extra to involve margin of error.
- As per the steps performed ,we have used neighbourhood group column as a text facet for various locations and performed the majority steps.
- The total number of records that has been cleanup is close to 12777 hence the dataset is made 25% Raw/Unclean based on the mentioned dimensions