DATA MANAGEMENT - 1 PROJECT

DATA PROFILING Using Talend Open Studio for Data Quality & Data Cleaning Using Open Refine

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 HARLEN	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 Apartm	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

Data Loading

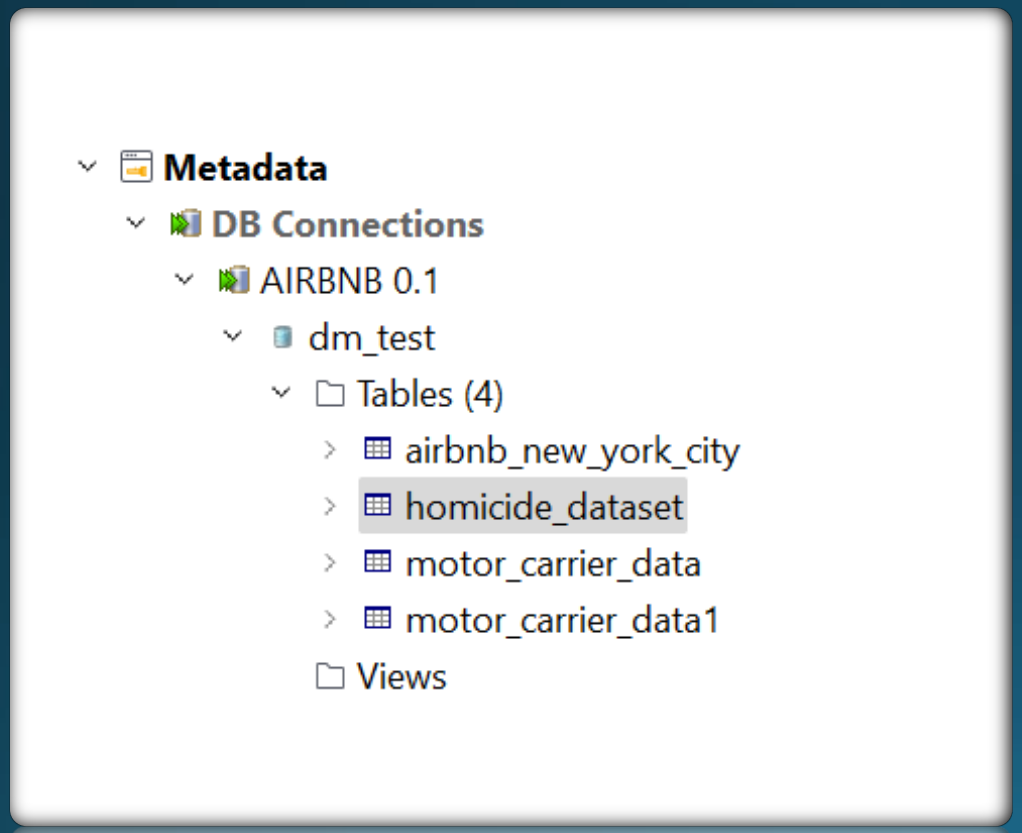
- Loaded the data into SQL workbench using code :

```sql
use dm_test;
show tables;
-- ------------Loading File data-----------
SHOW VARIABLES LIKE "secure_file_priv";
Load DATA INFILE 'C:/ProgramData/MySQL/MySQL Server 8.0/Uploads/AIRBNB_New_York_City.csv'
INTO TABLE airbnb_new_york_city
FIELDS TERMINATED BY ','
ENCLOSED BY '"'
LINES TERMINATED BY '\n'
IGNORE 1 ROWS;
SET SQL_SAFE_UPDATES = 0;
-- -------------------table ----------------------
select * from airbnb_new_york_city
```

Result Grid

count(`id`)
48895

Data Profiling using Talend Data prep

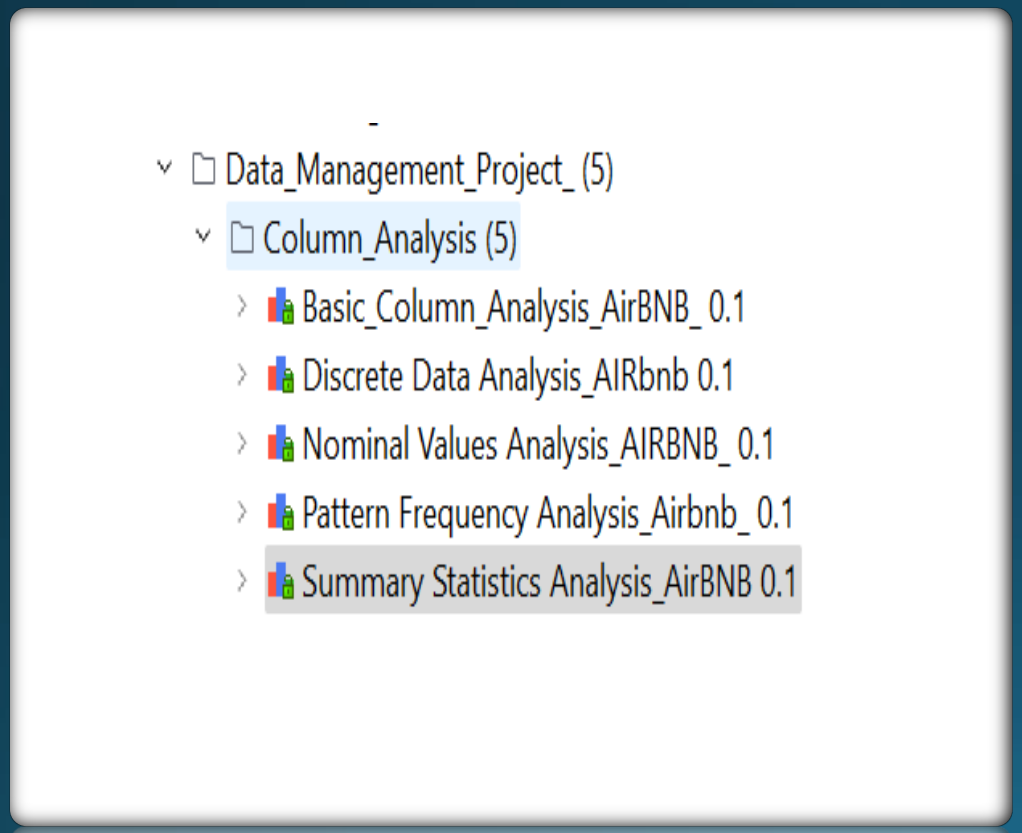


- The Dataset was loaded into Talend using DB connection (MySQL) using localhost server.

USER STORY

- As a Blogger, I want to know which host has more listing on the Airbnb platform, so that I can compare the reviews based on number of listings.
- As a Guest, I want to browse the area so that I can discover type of room which are most reviewed in a particular area.
- As a Guest, I want to see past reviews of the listing, and other listings by the same hosts, so that I can assess the quality of the host and the quality of the space.
- As a Guest, I want to browse listings so that I can discover unique places to stay that aren't hotel rooms.
- As an Analyst, I want to see the changes in the number of reviews for all listings over the months for a particular year.
- As an Analyst, I want to see the area where maximum listings exists.
- As a Guest I want to average prices of each Airbnb location broken down by the individual neighborhoods.

DATA PROFILING: ANALYSIS



The following Column Analysis were done on the dataset:

- Basic Column Analysis
- Discrete Data Analysis
- Nominal Value Analysis
- Pattern Analysis
- Summary Analysis

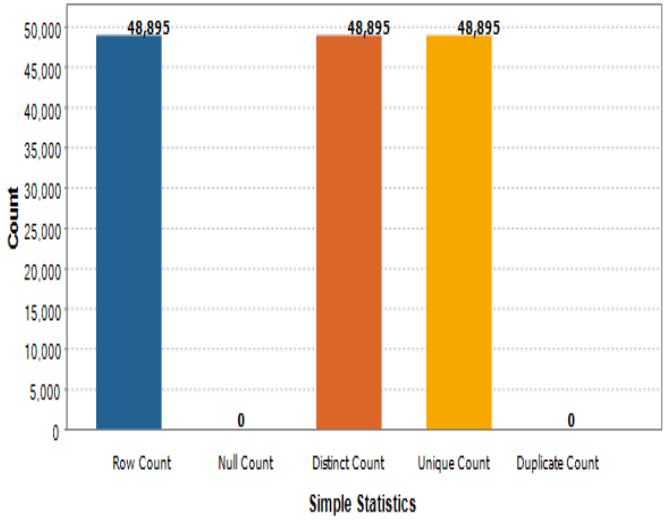
Basic Column Analysis

- Performed Basic Column Analysis on all the columns using simple statistics to identify the row count, null count, duplicate count etc.
- Data quality that can be identified using this analysis Completeness and Uniqueness.

Column: airbnb_new_york_city.id

Simple Statistics

Label	Count	%
Row Count	48895	100.00%
Null Count	0	0.00%
Distinct Count	48895	100.00%
Unique Count	48895	100.00%
Duplicate Count	0	0.00%



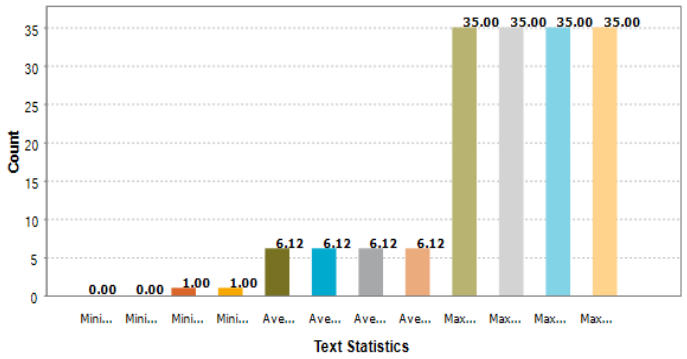
Nominal Value Analysis

- This is done only on the columns with Nominal data type. It gives the text statistics and Value frequency of the data in a particular column.
- Data quality that can be identified using this analysis Validity and Uniqueness.

Column: airbnb_new_york_city.host_name

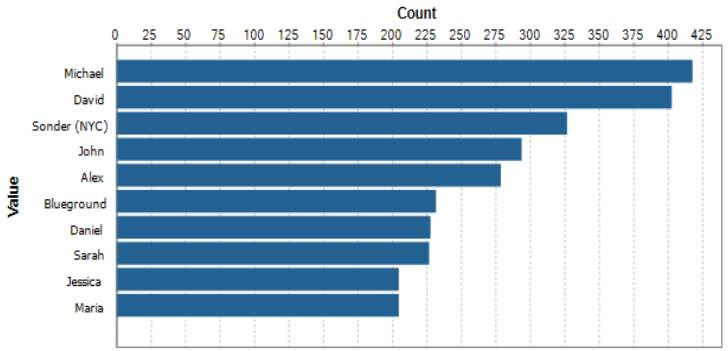
Text Statistics

Label	Value
Minimal Length With ...	0.00
Minimal Length With ...	0.00
Minimal Length With ...	1.00
Minimal Length	1.00
Average Length With ...	6.12
Average Length With ...	6.12
Average Length With ...	6.12
Average Length	6.12
Maximal Length With	35.00



Value Frequency

Value	Count	%
Michael	418	N/A
David	403	N/A
Sonder (NYC)	327	N/A
John	294	N/A
Alex	279	N/A
Blueground	232	N/A
Daniel	228	N/A
Sarah	227	N/A
Jessica	205	N/A



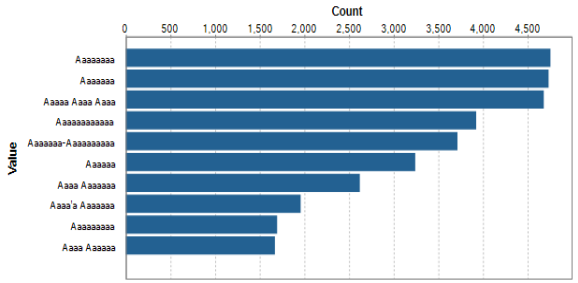
Pattern Frequency Analysis

- In pattern analysis, the pattern frequency of the data in columns are identified, and then we can also set regex/business rules according to the pattern.
- Data quality that can be identified using this analysis Validity and Consistency.

Column: airbnb_new_york_city.neighbourhood

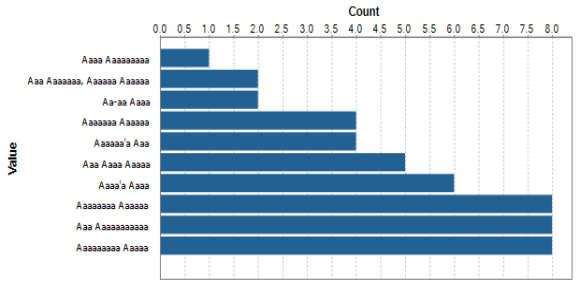
Pattern Frequency

Value	Count	%
Aaaaaaaa	4755	N/A
Aaaaaaa	4734	N/A
Aaaaa Aaaa Aaaa	4680	N/A
Aaaaaaaaaaaa	3924	N/A
Aaaaaaa-Aaaaaaaaaa	3714	N/A
Aaaaaa	3241	N/A
Aaaa Aaaaaaa	2621	N/A
Aaaa'a Aaaaaaa	1958	N/A
Aaaaaaaaa	1694	N/A



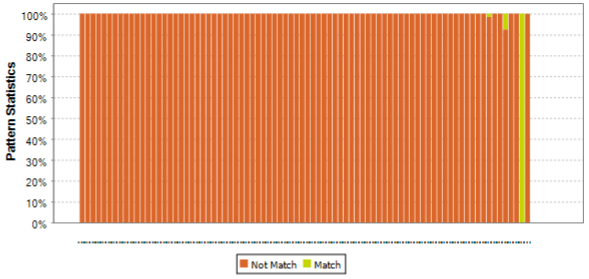
Pattern Low Frequency

Value	Count	%
Aaaa Aaaaaaaaa	1	N/A
Aaa Aaaaaaa, Aaaaaa ...	2	N/A
Aa-aa Aaaa	2	N/A
Aaaaaaa Aaaaaa	4	N/A
Aaaaaa'a Aaa	4	N/A
Aaa Aaaa Aaaaa	5	N/A
Aaaa'a Aaaa	6	N/A
Aaaaaaaa Aaaaaa	8	N/A
Aaa Aaaaaaaaaaa	8	N/A



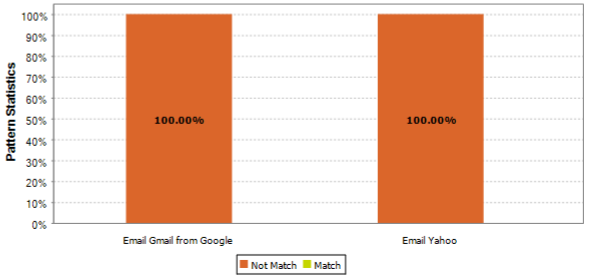
Pattern Matching

Label	Match...	Not M...	Match	Not M...
BE Code postal	0.00%	100.0...	0	48895
Companies House	0.00%	100.0...	0	48895
DE Postleitzahl (postal...	0.00%	100.0...	0	48895
FR Code postal	0.00%	100.0...	0	48895
FR Insee Code	0.00%	100.0...	0	48895
Postal code or Pin cod...	0.00%	100.0...	0	48895
Swiss Zip Code validat...	0.00%	100.0...	0	48895
US State Codes	0.00%	100.0...	0	48895
US Zipcode Validation	0.00%	100.0...	0	48895



SQL Pattern Matching

Label	Match...	Not M...	Match	Not M...
Email Gmail from Goo...	0.00%	100.0...	0	48895
Email Yahoo	0.00%	100.0...	0	48895



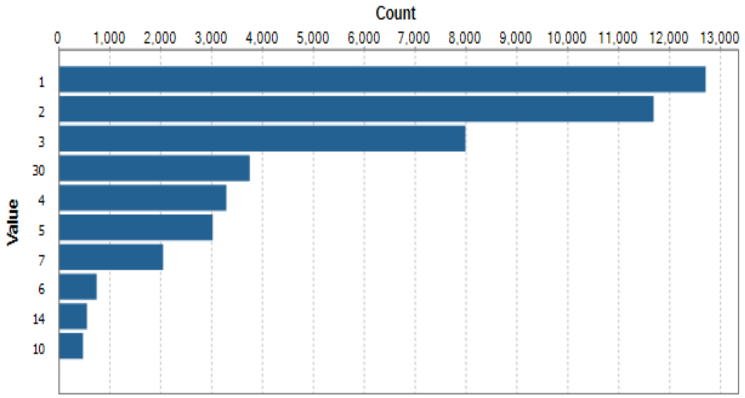
DISCRETE DATA ANALYSIS

- Discrete data analysis provides analysis of NUMERICAL data
- It gives the Bin frequency, helping us to find out the group of data present in the column
- It is useful in identifying the user stories easily.

Column: airbnb_new_york_city.minimum_nights

Bin Frequency

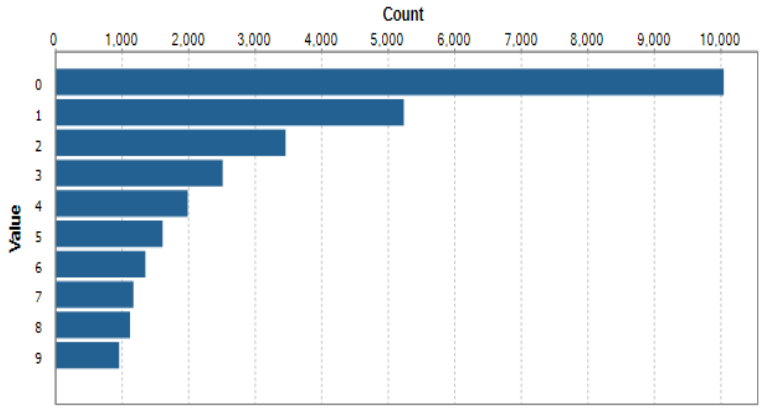
Value	Count	%
1	12720	N/A
2	11696	N/A
3	7999	N/A
30	3760	N/A
4	3303	N/A
5	3034	N/A
7	2058	N/A
6	752	N/A
14	562	N/A



Column: airbnb_new_york_city.number_of_reviews

Bin Frequency

Value	Count	%
0	10052	N/A
1	5244	N/A
2	3465	N/A
3	2520	N/A
4	1994	N/A
5	1618	N/A
6	1357	N/A
7	1179	N/A
8	1127	N/A



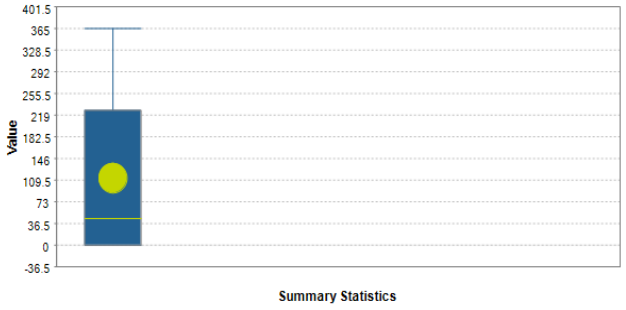
SUMMARY ANALYSIS

- Discrete data analysis provides analysis of numerical data
- It gives the range, the inter quartile range and the mean and median values
- It is useful in identifying the outliers.

Column: airbnb_new_york_city.availability_365

Summary Statistics

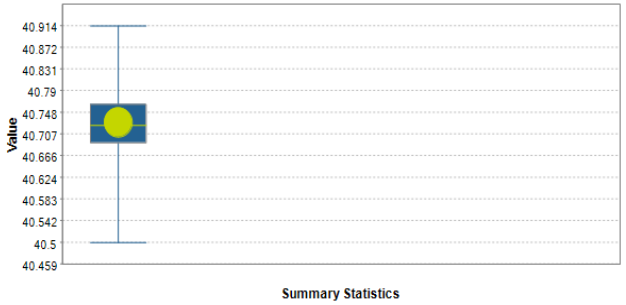
Label	Value
Mean	112.78132733408324
Median	45.0
Inter Quartile Range	227.0
Lower Quartile	0.0
Upper Quartile	227.0
Range	365.0
Minimum	0
Maximum	365



Column: airbnb_new_york_city.latitude

Summary Statistics

Label	Value
Mean	40.7289488806627
Median	40.72307
Inter Quartile Range	0.07301999999999964
Lower Quartile	40.6901
Upper Quartile	40.76312
Range	0.41327000000000425
Minimum	40.49979
Maximum	40.91306



CROSS-TABLE ANALYSIS: REDUNDANCY ANALYSIS

- In this analysis the two tables are to be compared to find out the primary key and foreign key based on functional dependency.

Last Successful Execution: 1

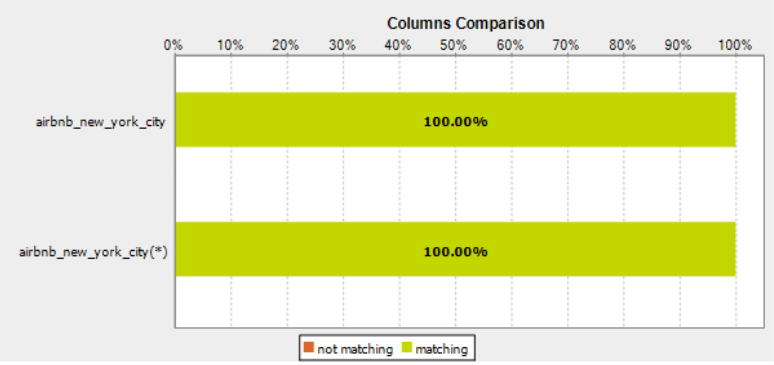
Analyzed Column Sets

Element(s) from airbnb_new_york_...	Element(s) from airbnb_new_york_...
id	id

Analysis Results

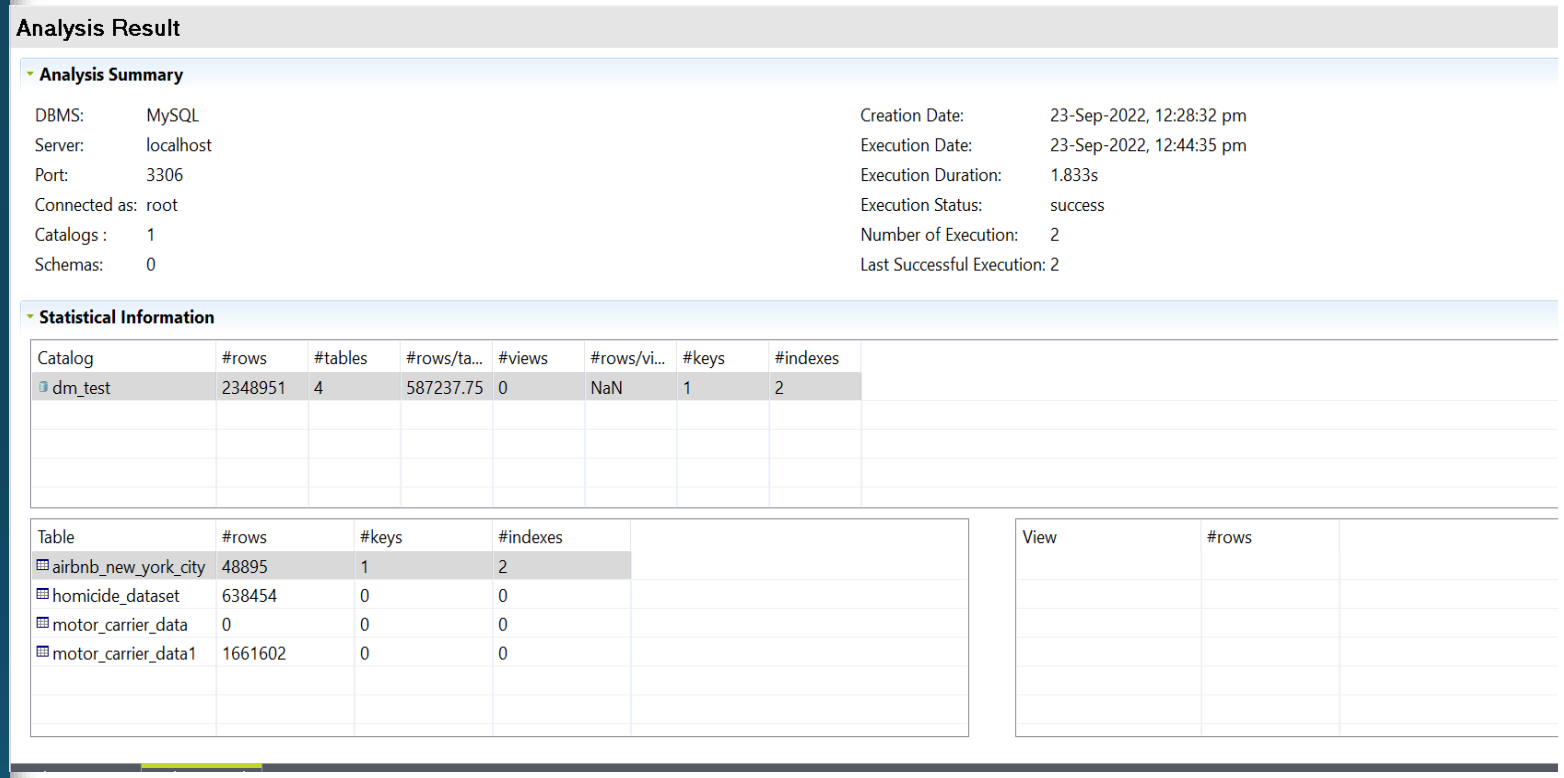
100.00% of the data from the A set (airbnb_new_york_city) are found in data from the B set (airbnb_new_york_city(*))
100.00% of the data from the B set (airbnb_new_york_city(*)) are found in data from the A set (airbnb_new_york_city)

	airbnb_new_yor...	airbnb_new_yor...
%Match	100.00%	100.00%
%NotMat...	0.00%	0.00%
#Match	48895	48895
#NotMatch	0	0
#Rows	48895	48895



STRUCTURAL ANALYSIS: CONNECTION ANALYSIS:

- Analyses the MySQL Database structure with number of rows, tables, keys, indexes etc.



Analysis Result

Analysis Summary

DBMS:	MySQL	Creation Date:	23-Sep-2022, 12:28:32 pm
Server:	localhost	Execution Date:	23-Sep-2022, 12:44:35 pm
Port:	3306	Execution Duration:	1.833s
Connected as:	root	Execution Status:	success
Catalogs :	1	Number of Execution:	2
Schemas:	0	Last Successful Execution:	2

Statistical Information

Catalog	#rows	#tables	#rows/ta...	#views	#rows/vi...	#keys	#indexes
dm_test	2348951	4	587237.75	0	NaN	1	2

Table	#rows	#keys	#indexes
airbnb_new_york_city	48895	1	2
homicide_dataset	638454	0	0
motor_carrier_data	0	0	0
motor_carrier_data1	1661602	0	0

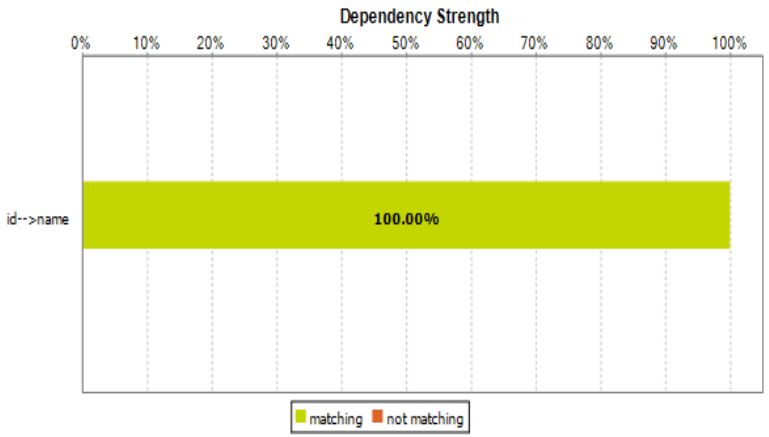
View	#rows

FUNCTIONAL DEPENDENCY ANALYSIS

- In this Analysis we find that column is the Primary key by having functional dependency with maximum columns having greater number of match %. In this, column “Id” has 100% match with other columns(here we have checked for name column)

Analysis Results

Dependency	#Match	%Match	#row
id-->name	48895	100.00%	48895



CORRELATIONAL ANALYSIS: NOMINAL ANALYSIS

- This analysis works with only nominal values and shows the relations between pairs of data.
- No conclusion as count is 1

Analysis Result

Analysis Summary

Connection: AIRBNB
Catalog: dm_test
Table(s): airbnb_new_york_city
View(s):

Analysis Result

Graphics

Simple Statistics

Data

reviews_per_month	last_review	room_type	neighbourhood	neighbourhood_group	host_name	name	COUNT(*)
		Private room	Bedford-Stuyvesant	Brooklyn	Sal	Cozy room...	1
4.27	2019-06-16	Private room	East Harlem	Manhattan	Erika	Room 1/2 ...	1
0.16	2016-11-04	Private room	Harlem	Manhattan	Leomaris	Beautiful ro...	1
0.09	2018-06-08	Entire home/apt	Sunnyside	Queens	Sen	Vintage NY...	1
		Private room	Williamsburg	Brooklyn	Rebecca	Available 2...	1
0.10	2019-06-09	Entire home/apt	East Village	Manhattan	Alexandra	GREAT APT...	1
		Entire home/apt	Kips Bay	Manhattan	Ramy	New 1BR: ...	1
0.13	2017-04-30	Entire home/apt	Kensington	Brooklyn	Sarah	Sunny Spac...	1

NUMERICAL CORRELATIONAL ANALYSIS

- This analysis works with both Provide NOMINAL and NUMERICAL.
- It is useful for user stories identification.

Analysis Result

Analysis Summary

Connection: AIRBNB
Catalog: dm_test
Table(s): airbnb_new_york_city
View(s):

Creation Date: 23-Sep-2022, 1:08:18 pm
Execution Date: 23-Sep-2022, 7:52:37 pm
Execution Duration: 0.171 s
Execution Status: success
Number of Execution: 3
Last Successful Execution: 3

Analysis Result

Graphics

Simple Statistics

Data

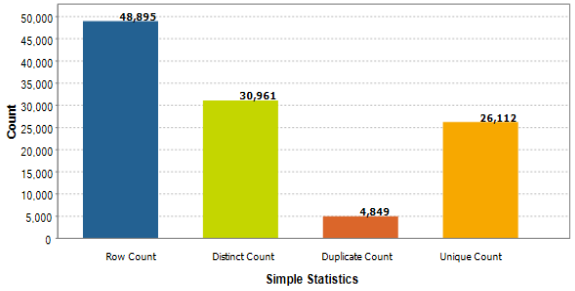
neighbourhood_group	AVG(availability_365)	COUNT(availability_365)	SUM(CASE WHEN availability_365 IS NULL THEN 1 EL...	AVG(calculated_host_listings_count)	COUNT(calculated_host_listings_count)	SUM(CASE WHEN calcu
Brooklyn	100.2323	20104	0	2.2844	20104	0
Manhattan	111.9794	21661	0	12.7913	21661	0
Queens	144.4518	5666	0	4.0602	5666	0
Staten Island	199.6783	373	0	2.3190	373	0
Bronx	165.7589	1091	0	2.2337	1091	0

TIME CORRELATIONAL ANALYSIS

- It also includes Nominal column, hence can analyze user story. “As a Guest, I want to browse the area so that I can discover type of room which are most reviewed in a particular area “ user story can be analyzed from this.

Simple Statistics

Label	Count	%
Row Count	48895	100.00%
Distinct Count	30961	63.32%
Duplicate Count	4849	9.92%
Unique Count	26112	53.40%



Data

reviews_per_month	neighbourhood_group	last_review	COUNT(*)
0.21	Brooklyn	2018-10-19	1
0.38	Manhattan	2019-05-21	1
	Manhattan		5029
4.64	Brooklyn	2019-07-05	1
0.10	Manhattan	2018-11-19	1
0.59	Manhattan	2019-06-22	2
0.40	Brooklyn	2017-10-05	1
3.47	Manhattan	2019-06-24	1
0.99	Manhattan	2017-07-21	1
1.33	Manhattan	2019-06-09	2

Analysis Settings | Analysis Results

Quality Dimensions

Column_name	Primary Key	Completeness	Consistency	Validity	Conformity	Accuracy	Uniqueness
id	Primary Key (Functional Depedency)	X	Yes	X	X	X	Yes
name	NA	Yes	X	X	Yes	Yes	NA
host_id	NA	X	Yes	Yes	X	X	NA
host_name	NA	Yes	Yes	X	X	Yes	NA
neighbourhood_group	NA	X	Yes	Yes	Yes	Yes	NA
neighbourhood	NA	X	Yes	Yes	Yes	X	NA
latitude	NA	X	X	X	X	X	NA
longitude	NA	X	X	X	X	X	NA
room_type	NA	X	X	X	X	X	NA
price	NA	X	Yes	X	X	Yes	NA
minimum_nights	NA	X	Yes	X	X	X	NA
number_of_reviews	NA	X	Yes	X	X	X	NA
last_review	NA	Yes	X	X	X	X	NA
reviews_per_month	NA	Yes	X	X	X	Yes	NA
calculated_host_listings_count	NA	X	Yes	X	X	X	NA
availability_365	NA	X	Yes	Yes	X	Yes	NA

Data Cleaning
Using OpenRefine

Recipe/ Steps

- Text Transformation - to Titlecase, replace characters on “name” column (By Nominal Value Analysis - Consistency)
- Text Transformation – fix names with by removing special characters on “name” column (By Nominal Value Analysis –, Conformity)
- Text Transformation - remove decimal on “price” column (by Discrete data Analysis - Consistency, Accuracy)
- Text Transformation - fix spelling of Queens in “neighbourhood_group” column(by Nominal Value Analysis - Accuracy)
- Text Transformation - updated the rows from ‘New York’ to ‘Manhattan’ in “neighbourhood_group” column(by Nominal Value Analysis - Accuracy, conformity, Validity)
- Text Transformation - updated blank cells with “NA” in column “last_review”(by Basic Column Analysis - Completeness)
- Text Transformation - to Titlecase, replace characters on “host_name” column (By Nominal Value Analysis - Consistency)
- Text Transformation - replaced null value with 0 and rounded the value unto 2 decimal point in “reviews_per_month” column (by pattern frequency analysis - completeness, consistency, accuracy)
- Edit column - added new column based on this column, added column by fetching URL on column “longitude” and “latitude”(Basic column Analysis - completeness, conformity)

Recipe/ Steps

Filter

0. Create project
1. Text transform on 40669 cells in column name: value.toTitlecase()
2. Text transform on 13 cells in column name: grel:value.replace("....","")
3. Text transform on 6218 cells in column name: grel:value.replace("!","")
4. Text transform on 102 cells in column name: grel:value.replace(";",",")
5. Text transform on 1534 cells in column price: jython:return(value).split(".")[0]
6. Text transform on 5666 cells in column neighbourhood_group: grel:value.replace("Qeens","Queens")
7. Text transform on 249 cells in column neighbourhood_group: grel:value.replace("New York","Manhattan")
8. Text transform on 10052 cells in column last_review: grel:if(value==null,'NA',value)
9. Text transform on 20 cells in column name: value.toTitlecase()
10. Text transform on 2852 cells in column name: grel:if(value.contains("And"),value.replace

Filter

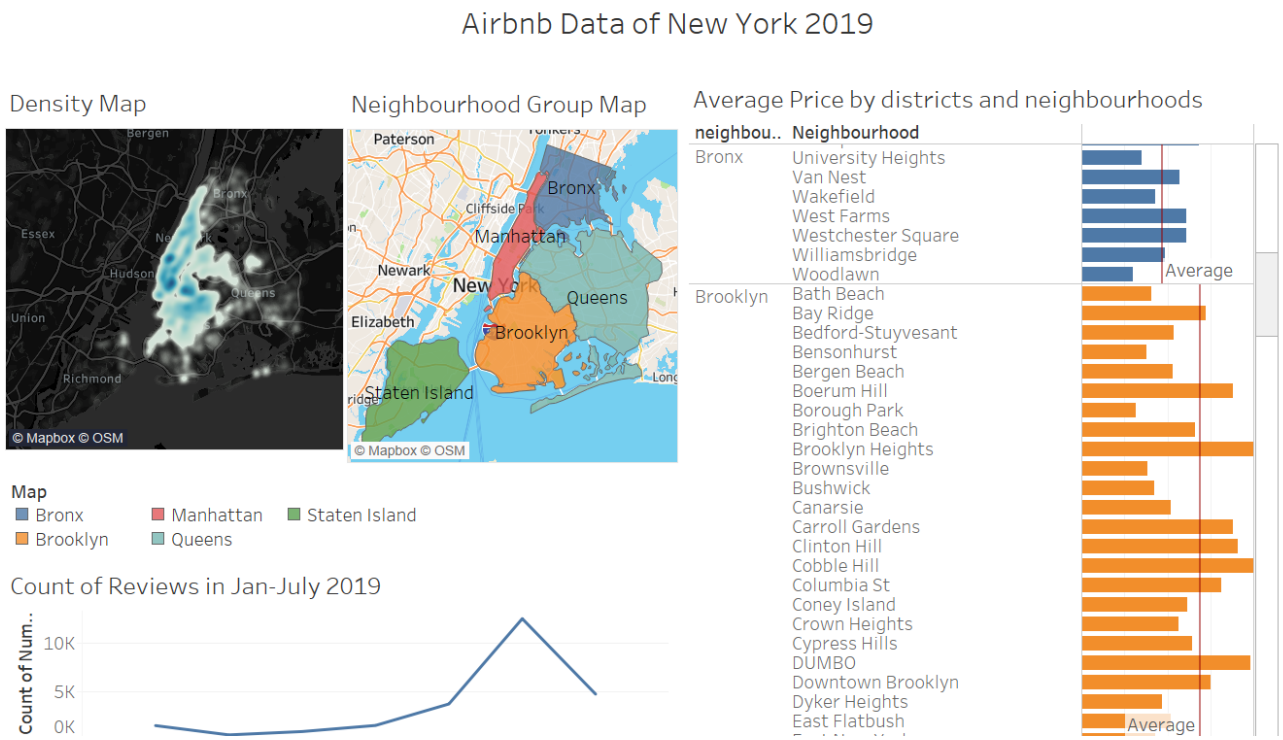
10. Text transform on 2852 cells in column name: grel:if(value.contains("And"),value.replace
11. Text transform on 6038 cells in column name: grel:if(value.contains("-"),value.replace("-","&"),value)
12. Text transform on 1287 cells in column name: grel:if(value.contains("+"),value.replace("+
13. Text transform on 974 cells in column host_name: value.toTitlecase()
14. Text transform on 1170 cells in column host_name: grel:if(value.contains("And"),value.replace
15. Text transform on 186 cells in column host_name: grel:if(value.contains("-"),value.replace("-","&"),value)
16. Text transform on 32 cells in column host_name: grel:if(value.contains("+"),value.replace("+
17. Text transform on 131 cells in column name: grel:if(value.contains("bdrm"),value.replac
18. Blank down 587 cells in column reviews_per_month
19. Text transform on 10639 cells in column reviews_per_month: grel:if(isNull(value),"0",value)
20. Text transform on 4758 cells in column reviews_per_month: jython:if(value=="0"): return value else: return str("%.2f"%round(float(value),2))

17. Text transform on 131 cells in column name: grel:if(value.contains("bdrm"),value.replac
18. Blank down 587 cells in column reviews_per_month
19. Text transform on 10639 cells in column reviews_per_month: grel:if(isNull(value),"0",value)
20. Text transform on 4758 cells in column reviews_per_month: jython:if(value=="0"): return value else: return str("%.2f"%round(float(value),2))
21. Create new column Coordinates based on column latitude by filling 18 rows with jython:return value
22. Text transform on 18 cells in column Coordinates: grel:cells ['latitude'].value + ' , '+cells['longitude'].value
23. Create new column API_URL based on column Coordinates by filling 18 rows with grel:"https://nominatim.openstreetmap.org q="+escape(value,"url")+"&format=jsonv2
24. Create column API_Data at index 9 by fetching URLs based on column API_URL using expression grel:value
25. Create new column Display_name based on column API_Data by filling 18 rows with jython:import json value1=json.loads(value) return value1[0]['display_name'].split(',')[3]

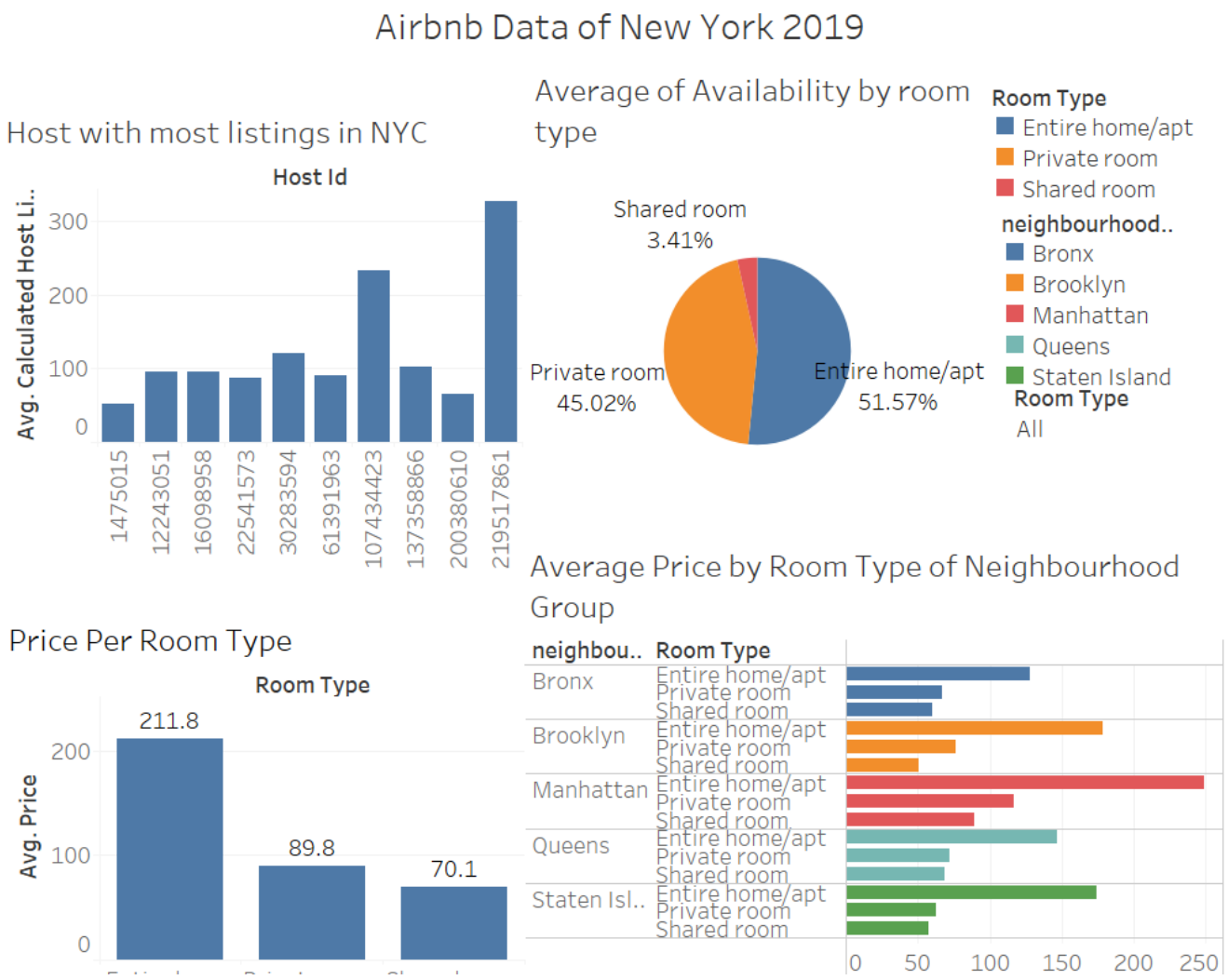
Text transform on 48877 cells in column Coordinates: grel:cells ['latitude'].value + ' , '+cells['longitude'].value

Create new column API_NEW based on column Coordinates by filling 48895 rows with grel:"https://nominatim.openstreetmap.or q="+escape(value,"url")+"&format=jsonv2

Dashboard


Airbnb Data of New York 2019
Density Map
Neighbourhood Group Map
Average Price by districts and neighbourhoods
© Mapbox © OSM
Paterson
Bronx
Manhattan
Newark
New York
Queens
Elizabeth
Brooklyn
Staten Island
Map
Bronx
Manhattan
Staten Island
Brooklyn
Queens
Count of Reviews in Jan-July 2019
Count of Num..
10K
5K
0K
neighbou..
Neighbourhood
Bronx
University Heights
Van Nest
Wakefield
West Farms
Westchester Square
Williamsbridge
Woodlawn
Average
Brooklyn
Bath Beach
Bay Ridge
Bedford-Stuyvesant
Bensonhurst
Bergen Beach
Boerum Hill
Borough Park
Brighton Beach
Brooklyn Heights
Brownsville
Bushwick
Canarsie
Carroll Gardens
Clinton Hill
Cobble Hill
Columbia St
Coney Island
Crown Heights
Cypress Hills
DUMBO
Downtown Brooklyn
Dyker Heights
East Flatbush
Average

Dashboard


Airbnb Data of New York 2019
Host with most listings in NYC
Host Id
Avg. Calculated Host Li..
300
200
100
0
1475015
12243051
16098958
22541573
30283594
61391963
107434423
137358866
200380610
219517861
Average of Availability by room type
Shared room
3.41%
Private room
45.02%
Entire home/apt
51.57%
Room Type
Entire home/apt
Private room
Shared room
neighbourhood..
Bronx
Brooklyn
Manhattan
Queens
Staten Island
Room Type
All
Price Per Room Type
Room Type
Avg. Price
211.8
89.8
70.1
200
100
0
Average Price by Room Type of Neighbourhood Group
neighbou..
Room Type
Bronx
Brooklyn
Manhattan
Queens
Staten Isl..
Entire home/apt
Private room
Shared room
0
50
100
150
200
250

Questions?

Thank You