New York State Hotel & Airbnb Analysis

Naibo Hu(Ray), Fan Fei(Fiona), Zihao Liu, Pengfei Gu(David)

Table of Contents

01 Executive Summary

The hospitality industry just experienced recession due to COVID-19 and is currently in a recovery stage. As people start to travel more, they face the problem of finding comfortable residences, convenient dine-in options, as well as safe environment.

Our group aims to analyze and visualize factors that characterize high quality hotels and Airbnbs in New York State and provide travellers with optimal decision that best suit their needs.

To provide insights that are data-driven and unbiased for decision making, our project will use Python, OpenRefine, R, Google Cloud Platform, SQL, and Tableau for data processing, data analysis, and data visualization.

02 Business Use Case

Travellers

Travellers can quickly access the information about hotel and Airbnb distribution across NY. Travellers can also filter out preferred residential options based on price, star-rating, room type, location, etc.

Our data analysis can further guide consumers' decision with tailored recommendation through accessing other information such as personal dining choices, other cost-efficient substitutes purposes of their trips.

Hotel Developers & Airbnb Owners

Our statistical model studies factors impacting popularity and price of hotels and Airbnb. Hotel developers and Airbnb hosts can utilize the models to determine pricing range and ideal locations to open their business.

03 Data Profile

	Dimensions	Size & Structure	Source	Matching Rules
hotels_ny.csv	11 attributes, 1,632 observations	218 KB Semi-structured	Google Dataset Search	Join with hotel_properties_ny on address
hotel_properties_ny.csv	20 attributes, 5,520 observations	855 KB Semi-structured	NYC Open Data	Join restaurant_ny on borough
airbnb_ny.csv	18 attributes, 36,924 observations	5391 KB Semi-structured	Inside Airbnb	Join hotel_properties_ny on borough
restaurant_ny.csv	9 attributes, 383,030 observations	83,829 KB Semi-structured	NYC Open Data	Join airbnb_ny on borough

03 Data Quality

Completeness

Any missing values?

Hotel_properties_ny.csv and airbnb_ny.csv have issues with names missing or no data

Validity

Does data match the rules?

All fields are checked and formatted to be the appropriate data type in our database.

Uniqueness

Are there duplicate values?

All hotels in hotels_ny.csv are unique. The rest datasets have repetitive rows but were removed.

Consistency

Consistent across various data stores?

All data is stored in and sourced from GCP, so it is consistent for all users in MySQL & Tableau.

Timeliness

Does data represent reality from required point in time?

Hotel_ny.csv is from 2017, but hotel information will not change significantly over time. The last update time for rest datasets is between 2020 and 2021.

Accuracy

Degree to which data represents reality

Hotels, Airbnbs, and restaurants ratings can be biased based on customers' experience.

04 Data Tools used

Data Extraction → **Data Cleaning** → **Database Design** → **Analysis** → **Visualization**

Google Dataset Search

NYC OpenData

Inside Airbnb
Adding data to the debate

NEW YORK STATE

data.world

OpenRefine, Excel, Python

OpenRefine, Excel, Python Numpy and pandas packages are used to process, organize and reformat the data

MySQL

Google Cloud Platform

MySQL Workbench and **Google Cloud Platform** are used to create a relational database and to store our data.

RStudio

mongoDB

MySQL

R studio, MongoDB, and MySql Workbench are used to generate insights from data

tableau

neo4j

Visualizations of the data are shown through **Tableau** and **neo4j**

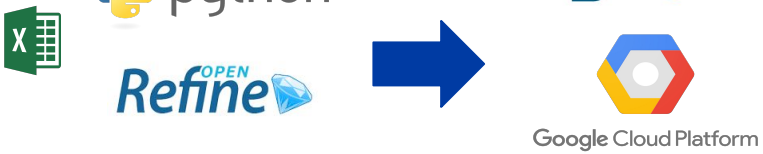
04 Data Processing (ETL)

Transform

- Use Python pandas to drop unnecessary columns, handle missing values, create organized dataframes, and select relevant attributes & rows
- Use Openrefine to clean and format dataset
- Use Excel to filter and remove duplicates

Load

- Use MySQL to connect tables and build Enhanced Entity Relationship model
- Upload datasets to GCP

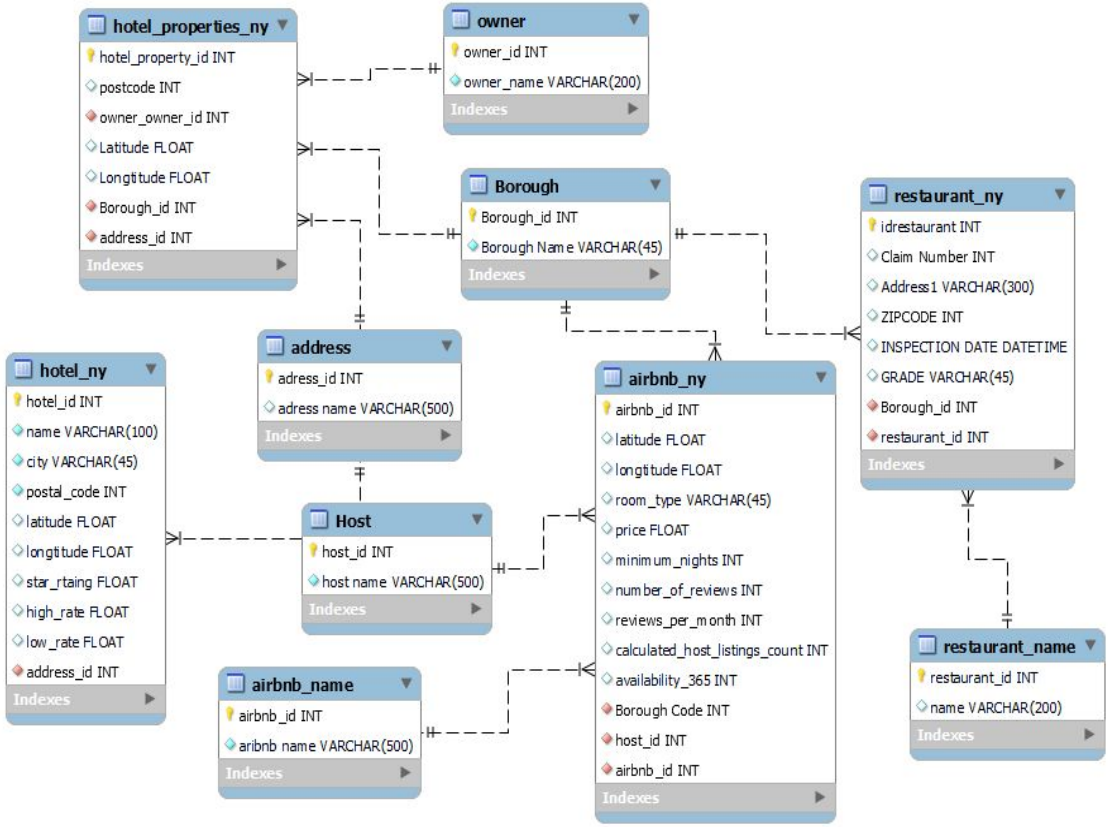


Dataset	Preprocessing Steps	Tables After Normalization
hotels_ny.csv	Delete “State”	“hotel_ny” “address”
hotel_properties_ny.csv	Delete ”PARID, BLOCK, LOT, TAXYEAR, BLDG_CLASS, TAXCLASS, Community Board, Council District, Census TRACK, BIN, BBL, NTA“. Drop rows with missing values. Merge street number & street name and rename the column to “address 1” Check and drop duplicate values based on address1 and OWNER_NAME Change Borough Name to the same format.	”owner“ “hotel_properties_ny” “borough”,
airbnb_ny.csv	Delete “license”, “last_review”, “neighborhood”. Remove duplicate rows, and missing values	“host”, “airbnb_name”, “room”, “airbnb_ny”
restaurant_ny.csv	Drop “Inspection Date”, “Violation Code”, “Violation Description” Remove duplicate claims. Cluster rows in OpenRefine	“restaurant_ny” “restaurant_name”

04 Enhanced Entity Relationship Diagram

DESIGN CONSIDERATIONS

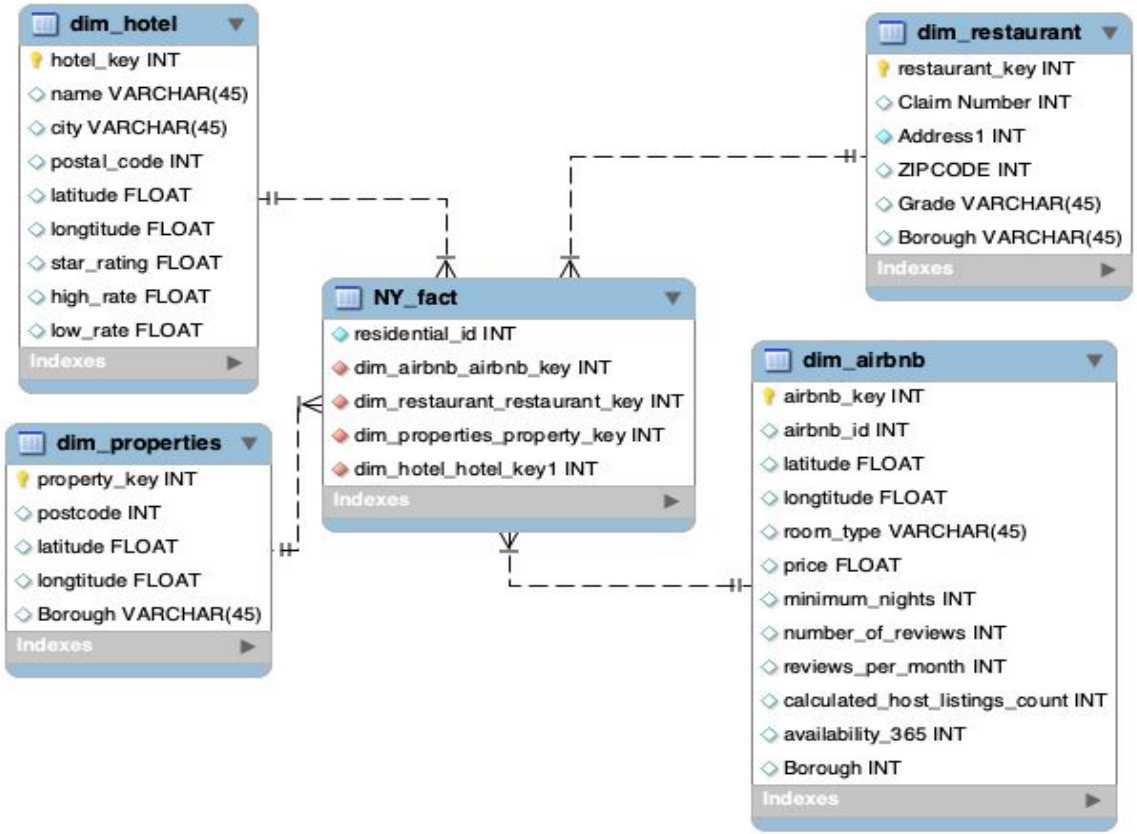
- Datasets are **normalized** for easier insertion, update and deletion
- **Address, host, airbnb_name, Borough, owner, restaurant_name** are likely to remain **static** for long periods of time, so they sit in their own tables.
- **Hotel_properties_ny, hotel_ny, airbnb_ny, and restaurant_ny** data are **dynamic**, so we stored them in separate tables to make frequent updates easier.
- While our database only includes data for New York, we built borough, host, address tables to accommodate more states with the **assumption this database would extend to more states when rolled out nationwide**



04 Dimensional Model

DESIGN CONSIDERATIONS

- **Dimensional Model** is developed based on the EER model that we constructed.
- Dimension **tables** include dim_hotel, dim_properties, dim_airbnb, and dim_restaurant.
- NY_fact table contains all the **table keys** from the dimension tables listed above.
- **Multi-dimensional views** help users with problem-solving, planning, and decision support



04 Database Design Considerations

Our group designs both OLTP and OLAP databases. We will use **OLTP** for database design as we expect a large number of data entries and modifications.

Subject-Oriented

This OLTP database will be used for the purpose of **responding immediately to the user requests** such as helping hotel & airbnb owners make sound decisions in terms of locating the premises to open their business

Quick Performance

Highly **effective in resolving insertion, updating and deletion** anomalies as well as support the application's data integrity.

Since Airbnb, hotels, and restaurants information change frequently over time, the OLTP database enables large number of data modifications.

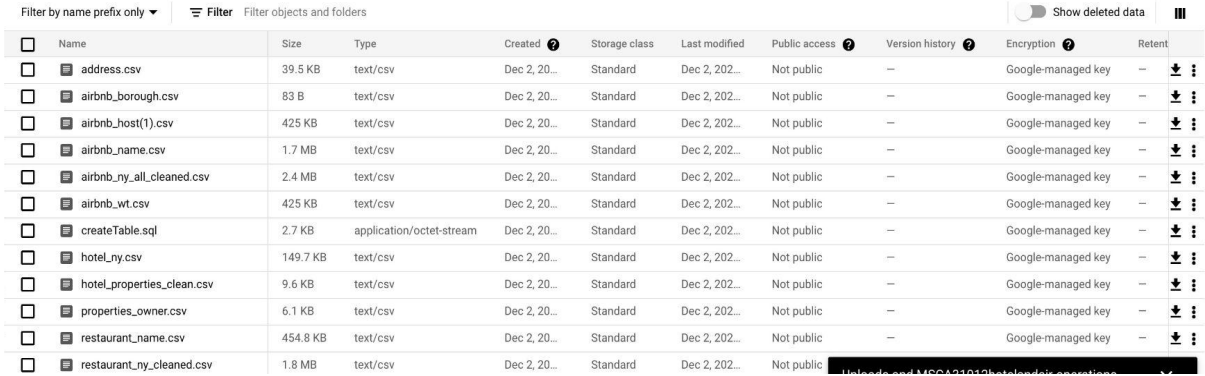
Concurrency & Atomicity

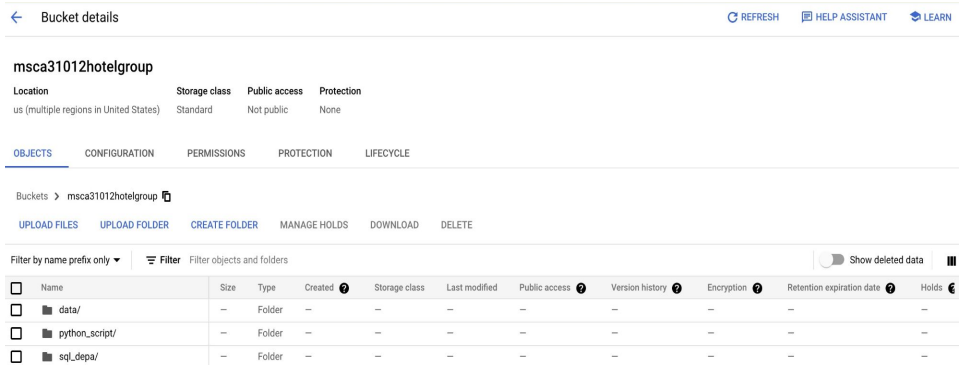
Ability to **handle many transaction requests** simultaneously and the ability to **reliably backup** and continue if part of the system fails.

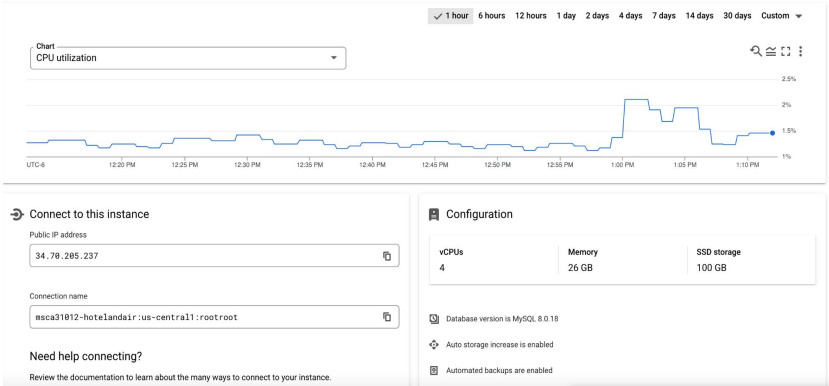
Since we expect large number of users and data entries, OLTP database ensures stable system performance.

04 Google Cloud Platform (GCP)

We uploaded **normalized datasets and coding scripts** to GCP to ensure that our works were saved properly. In addition, these data could be used for future collaborations.

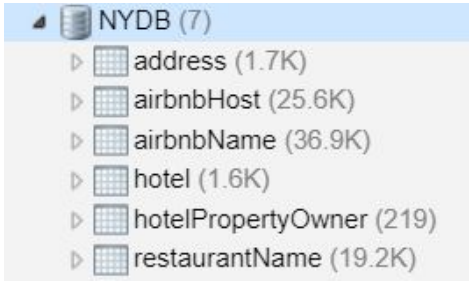






05 MongoDB: Comparison of NY Hotel & Airbnb

Load data to MongoDB



We loaded datasets to MongoDB through command prompt, and we named the database **"NYDB."** We queried the total number of hotels and Airbnbs in NY state. We found counts of hotels in NYC. we also queried the number of hotel owners and Airbnb hosts in NY.

As a result, the numbers of NY Airbnb and hosts are much greater than NY hotels and hotel owners. Apparently, **Airbnb offers more residential options for travellers and has a competitive advantage**.

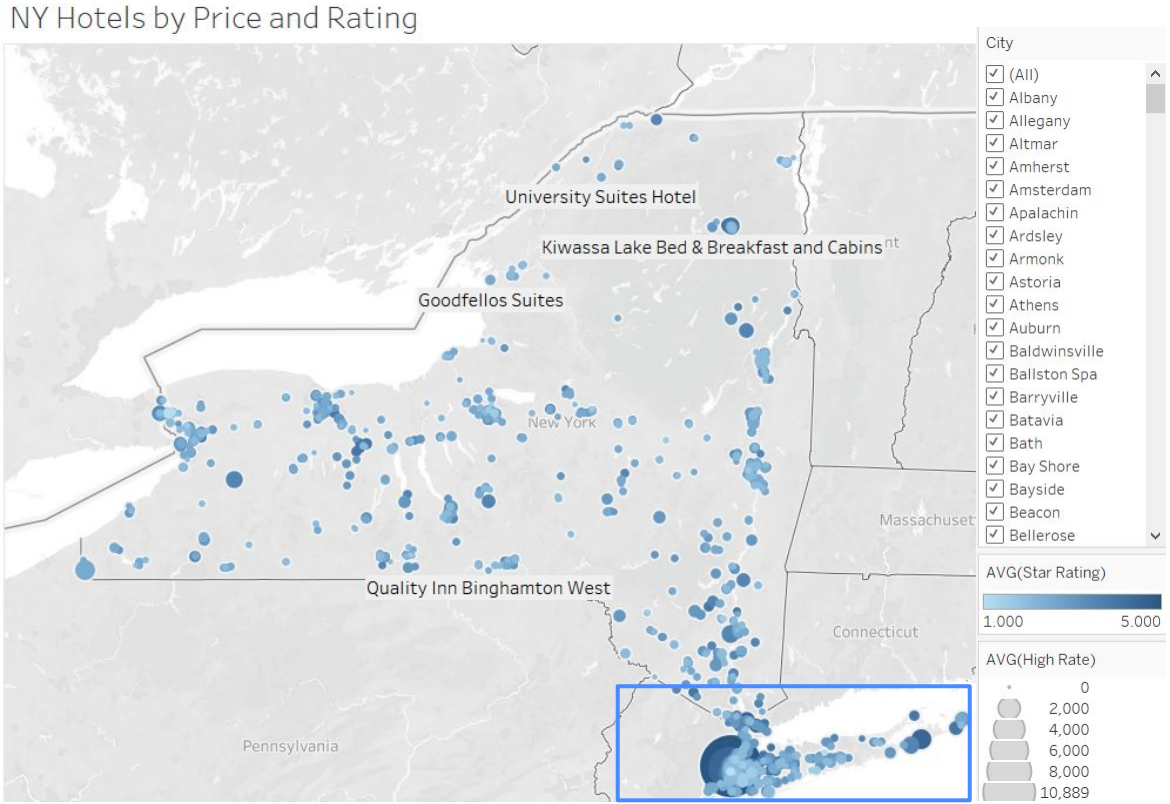
```
/* 1. count number of distinct airbnb_id */
db.airbnbName.distinct('_id').length

/* 2. count number of distinct hotel names */
db.hotel.distinct('name').length

/* 3. List total number of hotels that in the city 'New York" ?*/
db.hotel.find({city : "New York"}).count()

/* 4. Count the number of hotel owner */
db.hotelPropertyOwner.distinct('OWNER_NAME').length

/* 5. Count the number of Airbnb Host */
db.airbnbHost.distinct('host_id').length
```

Output:

36,882 Airbnbs in NY
1,602 hotels in NY,
401 hotels in New York City

219 NY hotel owners
25,650 NY Airbnb hosts

05 Tableau: Analysis of NY Hotels



This graph shows the **distribution of hotels in NY state.** The graph also displays hotel names, rate, and star rating.

Majority of hotels are in new york city, and **hotel prices in NYC are higher** than those in other regions.

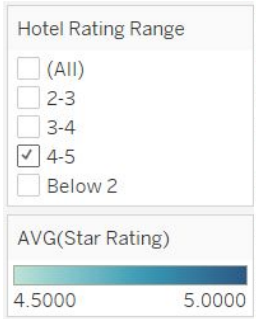
City was added as a filter, and travellers can use this filter to find desired hotels by region.

05 Tableau - Analysis of NY Hotel

Since hotel ratings are important factors impacting people's decision for hotel selection, we **divided rating into four categories** and added it as a filter. People can view the list of hotels by rating. In this case, we select NY hotels with 4-5 rating.

Hotel rating range calculated field

```
IF [Star Rating] <= 5 AND [Star Rating] > 4 THEN '4-5'
ELSEIF [Star Rating]  <=4  AND [Star Rating] > 3 THEN '3-4'
ELSEIF [Star Rating]  <=3  AND [Star Rating] > 2 THEN '2-3'
ELSEIF [Star Rating]  > 0 THEN 'Below 2'
END
```



NY Hotel by Rating

Hotel Name	
WestHouse New York	5.0000
Viceroy Central Park N..	5.0000
Trump Soho New York	5.0000
Trump International Ho..	5.0000
The William Vale Hotel	5.0000
The Towers of the Wald..	5.0000
The Towers at Lotte Ne..	5.0000
The Surrey	5.0000
The St. Regis New York	5.0000
The Ritz-Carlton New Y..	5.0000
The Ritz-Carlton New Y..	5.0000
The Ritz-Carlton New Y..	5.0000
the Quin	5.0000
The Plaza Hotel	5.0000
The Pierre, A Taj Hotel, ..	5.0000
The Peninsula New York	5.0000
The New York EDITION	5.0000
The Marmara Park Ave..	5.0000
The Mark	5.0000
The Lowell	5.0000
The Knickerbocker Hotel	5.0000
The Chatwal, a Luxury ..	5.0000
The Carlyle, A Rosewoo..	5.0000
The Beekman, A Thomp..	5.0000
Smyth, A Thompson Ho..	5.0000
Safehouse Suites Manh..	5.0000

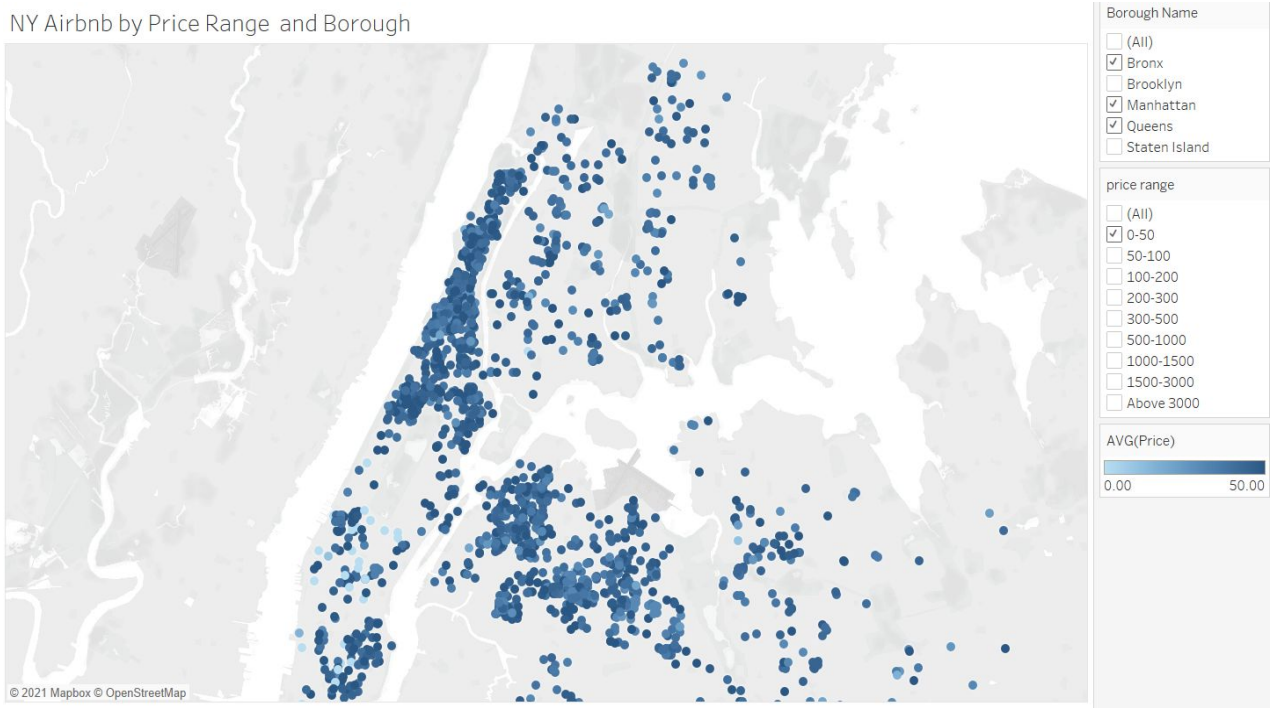
NY Hotel by Hotel Owner

Owner Name	Hotel Name	
11 WEST 51 REALTY, LLC	The Jewel facing Roc..	Manhattan
54M 33W37 LLC	Marriott Vacation Cl..	Manhattan
AXIT CAPITAL LLC	Jet Luxury at the Tru..	Manhattan
	Trump Soho New Yo..	Manhattan
DUNDEE INVEST LLC	Jet Luxury at the Tru..	Manhattan
	Trump Soho New Yo..	Manhattan
HA TRUONG LLC	Jet Luxury at the Tru..	Manhattan
	Trump Soho New Yo..	Manhattan
SOHO 1111 LLC	Jet Luxury at the Tru..	Manhattan
	Trump Soho New Yo..	Manhattan

We also listed hotels in manhattan area and found that Trump Soho New York was owned by three companies, including **"AXIT CAPITAL LLC," "DUNDEE INVEST LLC," and "SOHO 1111 LLC."**

05 Tableau: Analysis of NY Airbnb

We divided Airbnb prices into nine ranges and added **“borough” and “price range” as filters**. In this case, we filtered Airbnbs in **Bronx, Manhattan, and Queens** with price **below $50** per night. Travellers can use these filters to search for desired Airbnbs based on price and location.

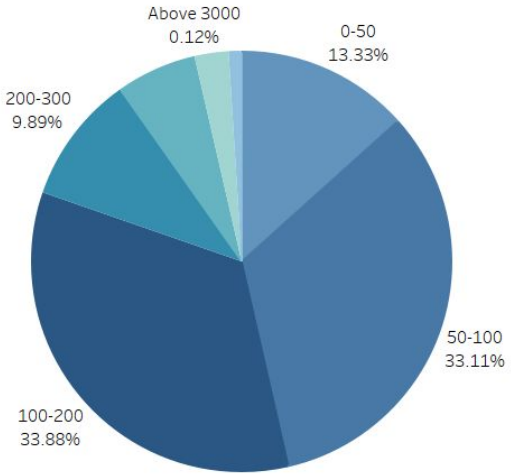


Price range calculated field:

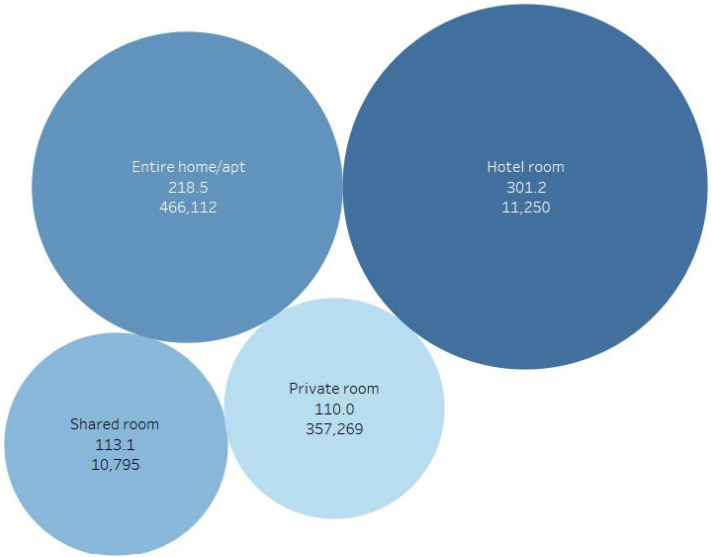
```
IF [Price] <= 50 AND [Price] >= 0 THEN '0-50'
ELSEIF [Price]  <= 100 AND [Price] > 50 THEN
'50-100'
ELSEIF [Price]  <= 200 AND [Price] > 100 THEN
'100-200'
ELSEIF [Price]  <= 300 AND [Price] > 200 THEN
'200-300'
ELSEIF [Price]  <= 500 AND [Price] > 300 THEN
'300-500'
ELSEIF [Price]  <= 1000 AND [Price] > 500
THEN '500-1000'
ELSEIF [Price]  <= 1500 AND [Price] > 1000
THEN '1000-1500'
ELSEIF [Price]  <= 3000 AND [Price] > 1500
THEN '1500-3000'
ELSEIF [Price]  > 3000 THEN 'Above 3000'

END
```

05 Tableau: Analysis of NY Airbnb

NY Airbnb Count Proportion by Price Range



NY Airbnb Average Price and Sum of Reviews by Room Type



The left graph shows that **majority (67%) Airbnbs are priced between $50 - $200** per night.

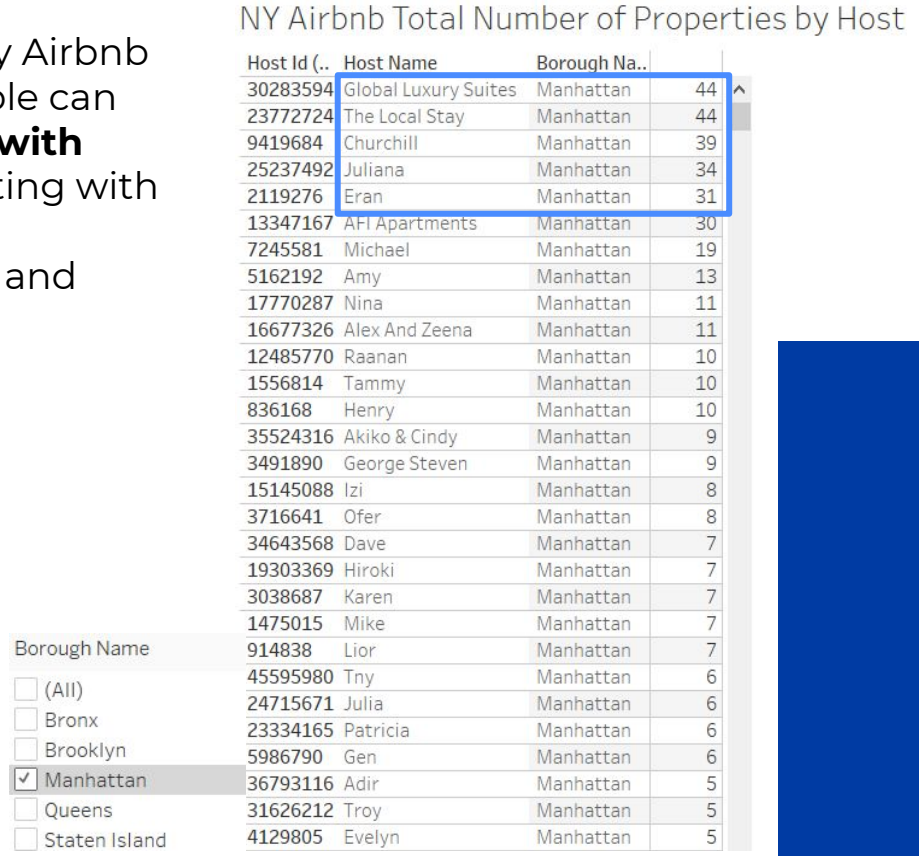
The right graph shows that **Hotel room is the most expensive room type** ($301.2 per night).

Entire home/apt is the most popular room type, as it has the highest number of reviews (466,112 reviews).

05 Tableau: Analysis of NY Airbnb

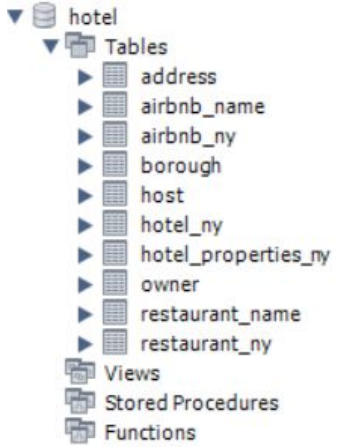
We counted the number of properties by Airbnb host and added borough as a filter. People can use this filter to **identify regional hosts with large number of properties.** By connecting with these hosts, people can have a better understanding to regional Airbnb prices and availability.

In this case, the table shows that **“Global Luxury Suites”, “The Local Stay”, “Churchill”, “Juliana”, and “Eran”** are the top 5 airbnb hosts who possess highest number of properties in Manhattan, which means that they together own significant amount of Airbnb market share in that region.



05 MySQL Workbench: Analysis of NY Restaurant

Load Data to hotel schema



We created **hotel schema**, loaded csv datasets to sql workbench through Import Wizard, and queried **number of restaurants by borough and grade.**

We find that **Manhattan has the largest number of restaurants**, while Staten Island has the least number of restaurants. In addition, 23,267 of restaurants have A grade.

# of Restaurants by borough

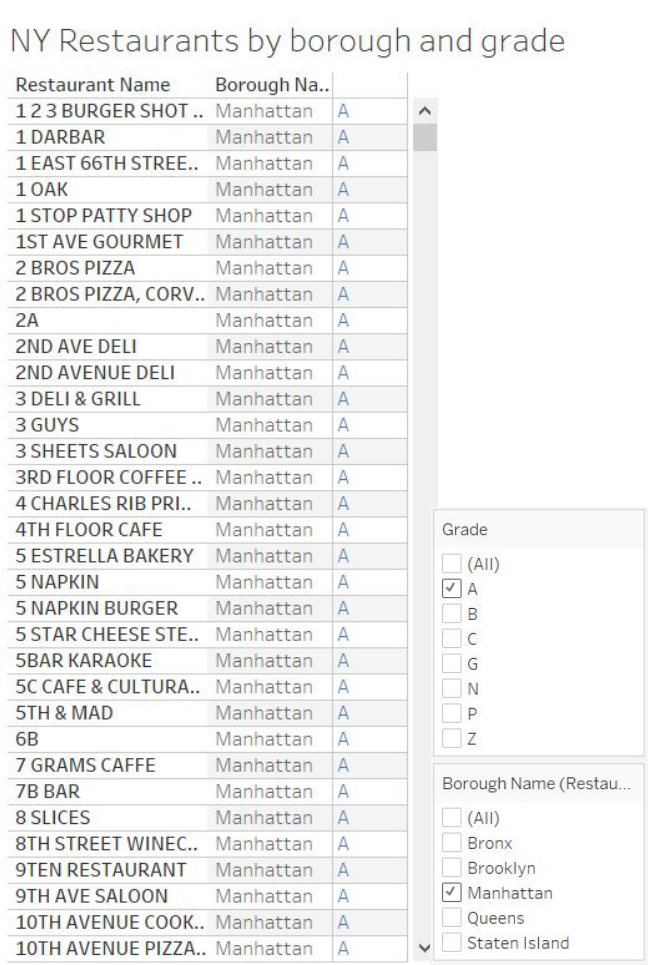
borough_name	total_count
Manhattan	12315
Brooklyn	8292
Queens	7432
Bronx	3070
Staten Island	1105

# of Restaurants by grade

grade	total_count
A	23267
B	4443
C	1746
P	1356
N	1000
Z	401
G	1

```sql
# count number of restaurants group by borough
SELECT
    r.borough_name, COUNT(r.restaurant_name) AS total_count
FROM
    restaurant_ny AS r

GROUP BY r.borough_name
ORDER BY total_count DESC;
```

```sql
# count number of restaurant group by grade
SELECT
    r.grade, COUNT(r.restaurant_name) AS total_count
FROM
    restaurant_ny AS r

GROUP BY r.grade
ORDER BY total_count DESC;
```



05 Tableau: Analysis of NY Restaurant

We created a table displaying each restaurant and corresponding grade. We added grade, borough as filters so that people can use them to **search for qualitative restaurants in different regions**. In this graph we show a list of restaurants with A grade in Manhattan.

The filters would also be great for people to look up restaurants that they should avoid due to sanitize issues.

05 Statistical Analysis: Simple Linear Regression

With R studio, we analyzed **NY Airbnbs** using **simple linear regression in** and studied factors that associate with Airbnb prices. We splitted data into 80% training data and 20% testing data. To choose related variables, we used **stepwise forward method** for model selection

Model: price~borough+room_type+ minimum nights+number of reviews

The significant p-values indicate that borough code, room_type, minimum nights, and number of reviews are important factors impacting Airbnb price.

The R^2 is low while p-value is statistically significant, indicating that the relationship between y and x could be non-linear. (However, from **random forest**(non linear model), the % of variation explained is still low, **13.69%**) It means further research needs to be done to find more related variables.

```
Call:
 randomForest(formula = price ~ ., data = train_airbnb, mtry = 2)
               Type of random forest: regression
                     Number of trees: 500
No. of variables tried at each split: 2

          Mean of squared residuals: 76677.68
                    % Var explained: 13.69
```

```
Coefficients:
                              Estimate Std. Error t value Pr(>|t|)
(Intercept)                  263.35450    3.34157  78.812  < 2e-16 ***
Borough.Code2                -89.48395   10.53995  -8.490  < 2e-16 ***
Borough.Code3                -58.78587    3.76538 -15.612  < 2e-16 ***
Borough.Code4                -81.02425    5.37196 -15.083  < 2e-16 ***
Borough.Code5                -78.21210   19.16785  -4.080 4.51e-05 ***
`room_typeHotel room`         53.88415   19.83130   2.717  0.00659 **
`room_typePrivate room`      -97.17720    3.49315 -27.819  < 2e-16 ***
`room_typeShared room`       -90.98001   13.62407  -6.678 2.47e-11 ***
minimum_nights                -0.32162    0.05761  -5.582 2.39e-08 ***
number_of_reviews             -0.25273    0.03468  -7.288 3.22e-13 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 290.8 on 29495 degrees of freedom
Multiple R-squared:  0.04816,   Adjusted R-squared:  0.04787
F-statistic: 165.8 on 9 and 29495 DF,  p-value: < 2.2e-16
```

05 Statistical Analysis: Logistic Regression

With R studio, we analyzed **NY hotels** using **logistic regression** and studied factors that associate with hotel rating. We splitted data into 80% training data and 20% testing data.

Model: hotel_level~high_rate+low_rate+I(high_rate*low_rate)

After testing variables, we find **high price , low price, and the correlation term are important factors.** The model accuracy for test dataset is **90%**

```
Call:
glm(formula = hotel_level ~ high_rate + low_rate + I(high_rate *
    low_rate), family = binomial, data = train_hotel)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.2817  -0.3894  -0.2705  -0.1903   2.5376

Coefficients:
                          Estimate Std. Error z value Pr(>|z|)
(Intercept)             -5.080e+00  2.772e-01 -18.326  < 2e-16 ***
high_rate                7.165e-03  8.920e-04   8.033 9.49e-16 ***
low_rate                 7.266e-03  1.566e-03   4.639 3.50e-06 ***
I(high_rate * low_rate) -1.721e-06  1.845e-07  -9.325  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1147.85  on 1303  degrees of freedom
Residual deviance:  686.29  on 1300  degrees of freedom
AIC: 694.29

Number of Fisher Scoring iterations: 7
```

We split hotel rating into two categories and named the variable **“hotel_level”**

hotel_ny$hotel_level=ifelse(hotel_ny$star_rating>=4,1,0)

```
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0 265  22
         1  12  27

               Accuracy : 0.8957
                 95% CI : (0.8573, 0.9267)
```

05 Statistical Analysis: Text Mining

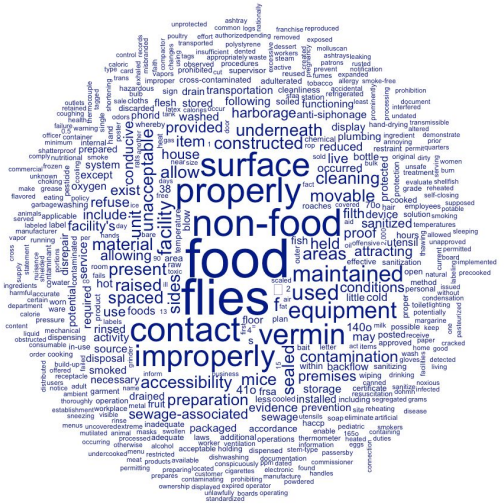
We use **restaurant dataset** for text analysis. We first inspect violation comments and count word frequencies to identify common problems. Then, we build a word cloud to visualize.

We find key words like **flies and vermin(mouse)** are frequently mentioned. By putting the key words in context, we find that vermin only refers to no protection against mouse pollution. but flies indicate that besides no protection against files, clients actually see flies in restaurants..

We group comments by different restaurants. We build a dictionary only include “flies” and “vermin” to see the top restaurant to avoid in nyc: **Locanda Vini E Olii**

```
dfm_sort(pos_dfm,decreasing = TRUE, margin = "features"
)
```

```
## Document-feature matrix of: 20,515 documents, 1 feature (44.39% sparse) and 1 docvar.
##                                                                 features
## docs                                                             bad
##                                                                   33
##   (lewis Drug Store) Locanda Vini E Olii                           3
##   (public Fare) 81st Street And Central Park West (delacorte Theatre)   3
```



06 Lessons Learned

01

The “hotel_ny” dataset contains information about city, while the rest of datasets include borough as geographic information. For better visualization and analysis purpose, we will make sure that **all datasets are consistent** in the future.

02

Our linear regression model has low R-square value and does not fit Airbnb data well. We should **add more independent variables** to improve the fit.

03

The Import Wizard feature in MySQL workbench is **time-consuming** and involves many manual works, which make large data import **infeasible**.

04

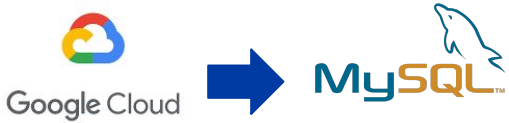
We have a better understanding to **data pipeline** after incorporating different data tools in this project.

06 Recommendations

Currently, our database are suitable for users to perform operations such as insert, update, and delete hotel and airbnb data.

01

As volume of data increases, use **GCP to load the data into SQL** instead of using import wizard.



02

Use python to **web scrape** and collect Airbnb and hotel real time data to **update our database** with most recent information. Some ideal websites include Booking, Tripadvisor, Airbnb, Expedia, etc.

03

With **real-time data**, we can conduct **time series analysis** to identify trends and systematic patterns as well as making predictions

07 Our Team

Naibo Hu(Ray)

Fan Fei (Fiona)

Zihao Liu

Pengfei Gu (David)

08 Appendix

Data ETL Python: NY Airbnb

NY airbnb

In [2]:
```
df = pd.read_csv ('airbnb_ny.csv')
df.head(10)
```

In [3]:
```
#drop column last_review, license
df = df.drop(columns=['license','last_review','neighbourhood'])
```

In [4]:
```
#change name from neighbourhood_group to Borough
df = df.rename({'neighbourhood_group' : 'Borough'}, axis = 1)
df.head()
```

Out[4]:

	id	name	host_id	host_name	Borough	latitude	longitude	room_type	price	minimum_nights	number_of_reviews	reviews_per_month
0	2595	Skylit Midtown Castle	2845	Jennifer	Manhattan	40.75356	-73.98559	Entire home/apt	150	30	48	0.33
1	3831	Whole flr w/private bdrm, bath & kitchen(pls r...	4869	LisaRoxanne	Brooklyn	40.68494	-73.95765	Entire home/apt	76	1	408	5.03
2	5121	BlissArtsSpace!	7356	Garon	Brooklyn	40.68535	-73.95512	Private room	60	30	50	0.54

Data ETL Python: NY Airbnb

```
#create borough table
borough = df.Borough.drop_duplicates()
borough
```

```
0          Manhattan
1           Brooklyn
23            Queens
89     Staten Island
91             Bronx
Name: Borough, dtype: object
```

```
#assign id to each Borough
borough_name = ['Manhattan','Bronx','Brooklyn','Queens','State Island']
id = ['1','2','3','4','5']
Borough = pd.DataFrame()
Borough['id'] = id
Borough['Name'] = borough_name
Borough
```

```
#create Room type table
room = df.room_type.drop_duplicates()
room
```

```
0       Entire home/apt
2          Private room
187          Hotel room
315         Shared room
Name: room_type, dtype: object
```

```
#assign id to each Room Type
type_name = ['Entire home/apt','Private room','Hotel room','Shared room']
type_id = ['1','2','3','4']
Room = pd.DataFrame()
Room['id'] = type_id
Room['Room Type'] = type_name
Room
```

Data ETL Python: NY Airbnb

```
#Create host table
host = df[['host_id','host_name']]
host.head(10)
```

	host_id	host_name
0	2845	Jennifer
1	4869	LisaRoxanne
2	7356	Garon
3	7378	Rebecca
4	8967	Shunichi
5	7490	MaryEllen
6	9744	Laurie
7	15991	Allen & Irina
8	16104	Kae
9	16800	Cyn

```
#Create Airbnb Table
airbnb = df[['id','name']]
airbnb.columns =["id","Name of Airbnb"]
airbnb.head(10)
```

	id	Name of Airbnb
0	2595	Skylit Midtown Castle
1	3831	Whole flr w/private bdrm, bath & kitchen(pls r...
2	5121	BlissArtsSpace!
3	5136	Spacious Brooklyn Duplex, Patio + Garden
4	5178	Large Furnished Room Near B'way
5	5203	Cozy Clean Guest Room - Family Apt
6	5803	Lovely Room 1, Garden, Best Area, Legal rental
7	6848	Only 2 stops to Manhattan studio
8	6872	Uptown Sanctuary w/ Private Bath (Month to Month)
9	6990	UES Beautiful Blue Room

Data ETL Python: NY Hotel

NY hotel

```
hotel = pd.read_csv ('hotels_ny.csv')
hotel.head()
```

	ean_hotel_id	name	address1	city	state_province	postal_code	latitude	longitude	star_rating	high_rate	low_rate
0	269955	Hilton Garden Inn Albany/SUNY Area	1389 Washington Ave	Albany	NY	12206	42.68751	-73.81643	3.0	154.0272	124.0216
1	113431	Courtyard by Marriott Albany Thruway	1455 Washington Avenue	Albany	NY	12206	42.68971	-73.82021	3.0	179.0100	134.0000
2	108151	Radisson Hotel Albany	205 Wolf Rd	Albany	NY	12205	42.72410	-73.79822	3.0	134.1700	84.1600
3	254756	Hilton Garden Inn Albany Medical Center	62 New Scotland Ave	Albany	NY	12208	42.65157	-73.77638	3.0	308.2807	228.4597
4	198232	CrestHill Suites SUNY University Albany	1415 Washington Avenue	Albany	NY	12206	42.68873	-73.81854	3.0	169.3900	89.3900

```
hotel=hotel.drop(columns=['state_province'])
```

```
# convert to CSV
df.to_csv("airbnb_ny_all.csv", index=False)
borough.to_csv("airbnb_borough.csv",index=False)
room.to_csv("airbnb_room.csv", index=False)
host.to_csv("airbnb_host.csv", index=False)
airbnb.to_csv("airbnb_name.csv", index=False)
hotel.to_csv("hotel_ny.csv", index=False)
```

Data ETL Python: NY Hotel

Hotel Properties - hotel_properties_ny.csv

```
import pandas as pd
```

```
df = pd.read_csv ('hotel_properties_ny.csv')
df.head(10)
```

```
df.to_csv("hotel_properties_clean_v1.csv", index = False )
```

```
df1 = pd.read_csv ('hotel-properties-clean-v3.csv')
df1
```

```
#drop column last_review, license
df = df.drop(columns=['PARID','BLOCK','LOT','TAXYEAR','BLDG_CLASS','TAXCLASS','Community Board',
                      'Council District','Census Tract','BIN','BBL','NTA'])
df
```

```
df = df.dropna()
```

```
df["address1"] = df["STREET NUMBER"] + " " + df["STREET NAME"]
df
```

```
df = df.drop(columns=['STREET NUMBER','STREET NAME'])
df
```

```
df = df.drop_duplicates(subset=['OWNER_NAME','address1'], keep=False)
df
```

```
#create owner table
owner = df1[['OWNER_NAME']]

owner['owner_id'] = owner.groupby(['OWNER_NAME']).ngroup()
owner = owner.drop_duplicates()
owner = owner.sort_values(["owner_id"], ascending=True)
owner
```

Data ETL OpenRefine: Hotel Properties & Restaurant

0. Create project
1. Text transform on 86 cells in column Borough: value.replace("1","MANHATTAN")
2. Text transform on 4 cells in column Borough: value.replace("2","BRONX")
3. Text transform on 24 cells in column Borough: value.replace("3","BROOKLYN")
4. Text transform on 23 cells in column Borough: value.replace("4","QUEENS")
5. Text transform on 1 cells in column Borough: value.replace("5","STATEN ISLAND")
6. Text transform on 3 cells in column Borough: value.replace("STATEN IS","STATEN ISLAND")

0. Create project
1. Rename column CAMIS to Claims
2. Rename column Claims to Claim Number
3. Rename column DBA to Restaurant Name
4. Rename column BORO to Borough_Code
5. Rename column STREET to Address1
6. Rename column Borough_Code to Borough Code
7. Remove column VIOLATION CODE
8. Remove column VIOLATION DESCRIPTION
9. Mass edit 26330 cells in column Restaurant Name
10. Mass edit 5586 cells in column Restaurant Name
11. Mass edit 9823 cells in column Restaurant Name

		Last modified	Name
✕	About	2021-11-14 18:28 PM	restaurant ny csv
✕	About	2021-11-14 17:07 PM	hotel properties clean v2 csv
✕	About	2021-11-14 16:35 PM	hotel properties clean v1 csv

SQL Data Import

address

address_id	address_name
1	26 ANN STREET
2	130 DUANE STREET
3	119 ORCHARD STREET
4	246 SPRING STREET
5	273 WEST 11 STREET
6	440 WEST 33 STREET
7	325 WEST 33 STREET
8	338 WEST 39 STREET
9	140 WEST 28 STREET
10	1227 BROADWAY
11	60 WEST 37 STREET
12	33 WEST 37 STREET
13	27 WEST 38 STREET
14	893 BROADWAY
15	86 MADISON AVENUE

airbnb_name

airbnb_id	airbnb_name
2595	Skylit Midtown Castle
3831	Whole flr w/private bdrm, bath & kitchen(pls read)
5121	BlissArtsSpace!
5136	Spacious Brooklyn Duplex, Patio + Garden
5203	Cozy Clean Guest Room - Family Apt
5803	Lovely Room 1, Garden, Best Area, Legal rental
6848	Only 2 stops to Manhattan studio
6872	Uptown Sanctuary w/ Private Bath (Month to M...
6990	UES Beautiful Blue Room
7064	Amazing location! Wburg. Large, bright & tranquil
7097	Perfect for Your Parents: Privacy + Garden
7750	2 Furnished bedrooms next to Central Park
7801	Sweet and Spacious Brooklyn Loft
8490	Maison des Sirenes1,bohemian, luminous apart...
9357	Midtown Pied-a-terre

borough

borough_code	borough_name
1	Manhattan
3	Brooklyn
4	Queens
5	Staten Island
2	Bronx

host

host_id	host_name
2845	Jennifer
4869	LisaRoxanne
7356	Garon
7378	Rebecca
7490	MaryEllen
9744	Laurie
15991	Allen & Irina
16104	Kae
16800	Cyn
17297	Joelle
17571	Jane
17985	Sing

SQL Data Import

airbnb_ny

airbnb_id	host_id	borough_code	latitude	longitude	room_type	price	minimum_nights	number_of_reviews	calculated_host_listings_count	availability_365
2595	2845	1	40.75356	-73.98559	Entire home/apt	150	30	48	3	340
3831	4869	3	40.68494	-73.95765	Entire home/apt	76	1	408	1	208
5121	7356	3	40.68535	-73.95512	Private room	60	30	50	1	365
5136	7378	3	40.66265	-73.99454	Entire home/apt	275	5	2	1	204
5203	7490	1	40.8038	-73.96751	Private room	75	2	118	1	0
5803	9744	3	40.66801	-73.98784	Private room	98	4	194	3	333
6848	15991	3	40.70935	-73.95342	Entire home/apt	89	30	182	1	274
6872	16104	1	40.80107	-73.94255	Private room	65	30	0	2	365
6990	16800	1	40.78778	-73.94759	Private room	62	30	234	1	287

hotel_ny

hotel_id	hotel_name	address_id	city	postal_code	latitude	longtitude	star_rating	high_rate	low_rate
269955	Hilton Garden Inn Albany/SUNY Area	124	Albany	12206	42.68751	-73.81643	3	154.0272	124.0216
113431	Courtyard by Marriott Albany Thruway	125	Albany	12206	42.68971	-73.82021	3	179.01	134
108151	Radisson Hotel Albany	126	Albany	12205	42.7241	-73.79822	3	134.17	84.16
254756	Hilton Garden Inn Albany Medical Center	127	Albany	12208	42.65157	-73.77638	3	308.2807	228.4597
198232	CrestHill Suites SUNY University Albany	128	Albany	12206	42.68873	-73.81854	3	169.39	89.39
125200	The Desmond Hotel Albany	129	Albany	12211	42.72874	-73.79807	3.5	189.0266	153.0644
109728	Ramada Plaza Albany	130	Albany	12206	42.68031	-73.78444	3	158.6321	89.036
235037	Hampton Inn & Suites Albany-Downtown	131	Albany	12210	42.65334	-73.75142	2.5	225.47	224.47
106464	Albany Marriott	132	Albany	12205	42.72111	-73.80036	3.5	158.8856	128.9077
106922	Best Western Sovereign Hotel - Albany	133	Albany	12203	42.67807	-73.82819	3	139.0244	78.3255

SQL Data Import

hotel_properties_ny

hotel_property_id	postcode	owner_id	borough_code	latitude	longtitude	address_id
1	10038	11	1	40.710764	-74.007708	1
2	10013	35	1	40.716114	-74.00737	2
3	10002	4	1	40.719454	-73.98956	3
4	10013	82	1	40.725616	-74.005253	4
5	10013	105	1	40.725616	-74.005253	4
6	10014	182	1	40.736068	-74.004399	5

restaurant_ny

claim_number	restaurant_name	borough_name	address	zipcode	inspection_date	grade
50086079	SUNSHINE PARADISE	Brooklyn	UTICA AVENUE	11203	NULL	N
40925478	BELLA VIA	Queens	VERNON BOULEVARD	11101	NULL	A
40608746	KANA TAPAS BAR & RESTAURANT	Manhattan	SPRING STREET	10013	NULL	A
50094131	ESSEX RESTAURANT	Manhattan	RIVINGTON STREET	10002	NULL	C
41447856	WILD	Manhattan	HUDSON STREET	10014	NULL	A

Owner

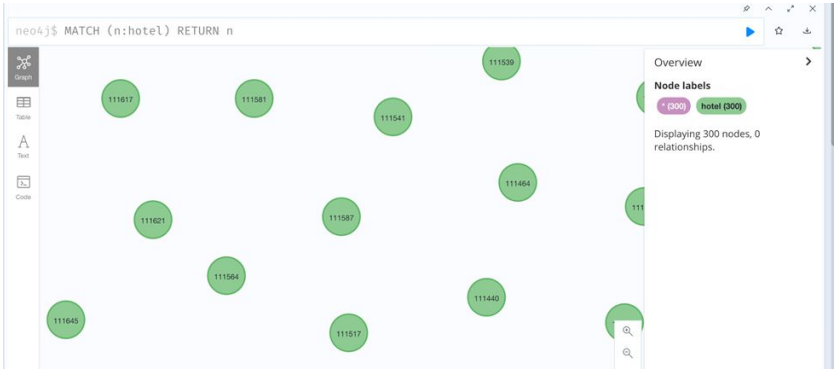
owner_id	owner_name
0	11 WEST 51 REALTY, LLC
1	1107 D LLC
2	111 EAST 48TH STREET HOLDINGS, LLC
3	117 WEST 79TH OWNER LLC
4	119 ORCHARD PROPERTY, INC.

restaurant_name

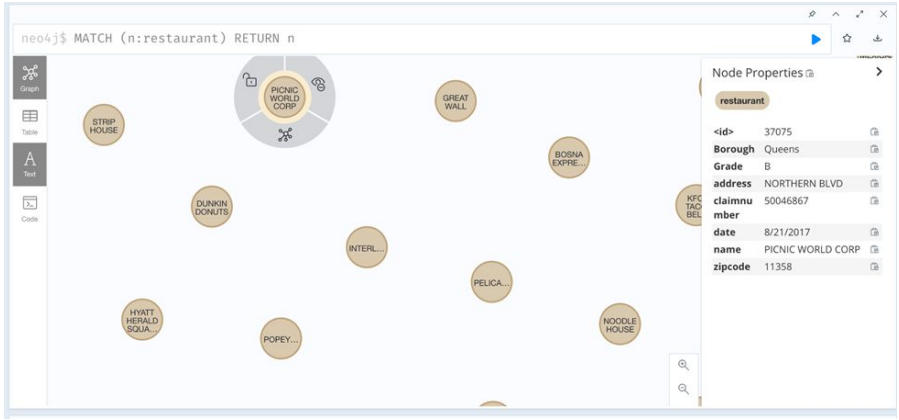
restaurant_id	restaurant_name
1	SUNSHINE PARADISE
2	BELLA VIA
3	KANA TAPAS BAR & RESTAURANT
4	ESSEX RESTAURANT

Neo4j Data Import

hotel_ny



restaurant_ny



airbnb_ny

Connect Sql to Tableau

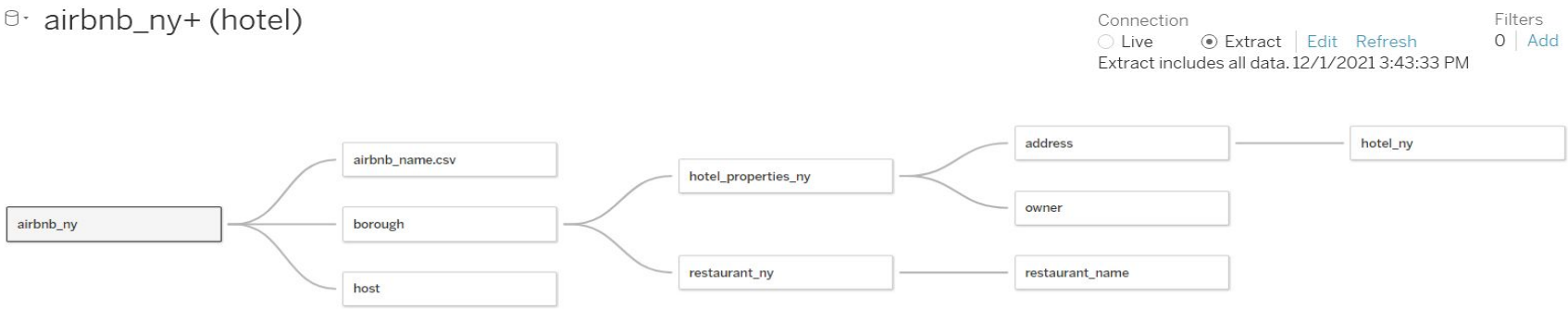
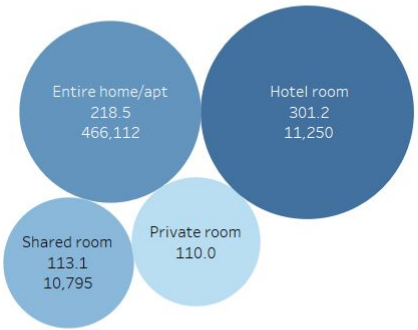


Tableau Dashboard

NY Total Number of Hotel and Airbnb

Distinct count of Airbnb Id	36,88
Distinct count of Hotel Name	1,58

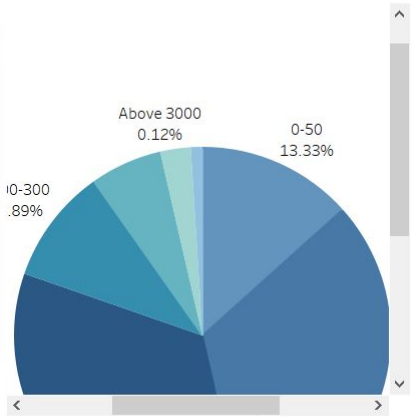
NY Airbnb Average Price and Sum of Reviews by Room Type



NY Restaurants by borough and grade

Restaurant Name	Borough Na..	
1 2 3 BURGER SHOT ..	Manhattan	A
1 DARBAR	Manhattan	A
1 EAST 66TH STREE..	Manhattan	A
1 OAK	Manhattan	A
1 STOP PATTY SHOP	Manhattan	A
1ST AVE GOURMET	Manhattan	A
2 BROS PIZZA	Manhattan	A
2 BROS PIZZA, CORV..	Manhattan	A
2A	Manhattan	A
2ND AVE DELI	Manhattan	A
2ND AVENUE DELI	Manhattan	A
3 DELI & GRILL	Manhattan	A
3 GUYS	Manhattan	A
3 SHEETS SALOON	Manhattan	A
3RD FLOOR COFFEE ..	Manhattan	A
4 CHARLES RIB PRI..	Manhattan	A

NY Airbnb Count Proportion by Price Range



NY Hotel by Hotel Owner

Owner Name	Hotel Na..	
11 WEST 51 REALTY, LLC	The Jewe..	Manhattan
54M 33W37 LLC	Marriott ..	Manhattan
AXIT CAPITAL LLC	Jet Luxur..	Manhattan
	Trump S..	Manhattan

NY Hotel by Rating

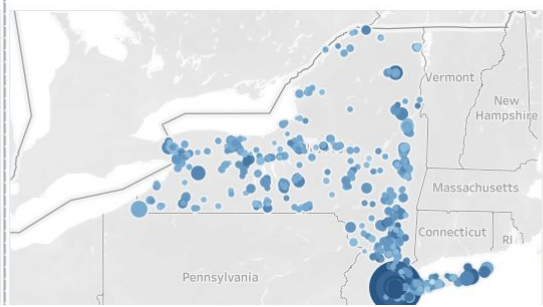
Hotel Name	
The Ritz-Carlton New Y..	5.0000
The Ritz-Carlton New Y..	5.0000
The Ritz-Carlton New Y..	5.0000
the Quin	5.0000
The Plaza Hotel	5.0000

NY Airbnb Total Number of Properties by Host

Host Id (..	Host Na..	Borough Na..	
24715671	Julia	Manhattan	6
23334165	Patricia	Manhattan	6
5986790	Gen	Manhattan	6
36793116	Adir	Manhattan	5
31626212	Troy	Manhattan	5
4129805	Evelyn	Manhattan	5
2559004	Sergio	Manhattan	5
292204	Blanca	Manhattan	5
81335	Evan	Manhattan	5
47022650	Jillian	Manhattan	4
44350279	J.P.	Manhattan	4

NY Airbnb by Price Range and Borough

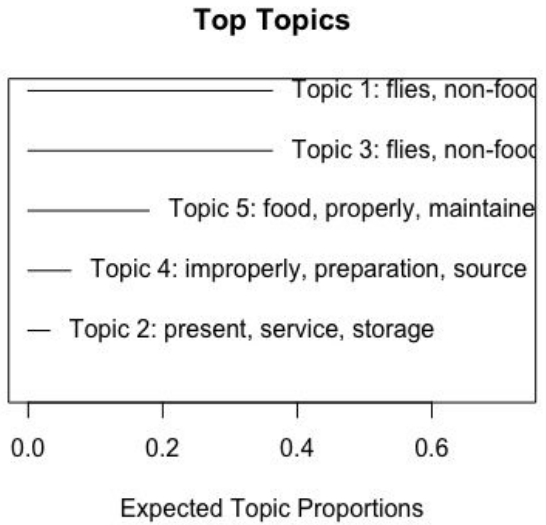
NY Hotel Price and Rating



Borough Name
(All)
Bronx
price range
(All)
Borough Name
(All)
City
Ardsley
Hotel Rating Range
4-5
Below 2

Topic MODEL For Text Analysis

```{r}

set.seed(100)
if (require("stm")) {
   my_lda_fit5 <-
stm(mydfm_nostops, K = 5,
verbose = FALSE)
   plot(my_lda_fit5)
}

```



We set **k=5** for topic model from deep learning to find potential topics of inspection comments. We believe there are 3 main themes from human review: food service, food storage,and food cleanliness. We choose not to put this in main part because further study and better data sets for comments are needed to get a more precise insights.

08 Reference

NY Hotel Dataset:
https://datasetsearch.research.google.com/search?query=new%20york%20hotels&docid=L2cvMTFqOWMwOHkzNQ%3D%3D

NY Hotel Properties:
https://data.cityofnewyork.us/City-Government/Hotels-Properties-Citywide/tjus-cn27

NY Airbnb:
http://insideairbnb.com/get-the-data.html

NY Restaurant:
https://data.cityofnewyork.us/Health/restaurant-data-set-2/f6tk-2b7a