MassiveData

NYC Airbnb Project

Group 7:
Luming Peng,
Yifan Qian,
Ci Song,
Yuxuan Zhang,
Yifan Zhou

CONTENT

Business Proposition

01

Database Design

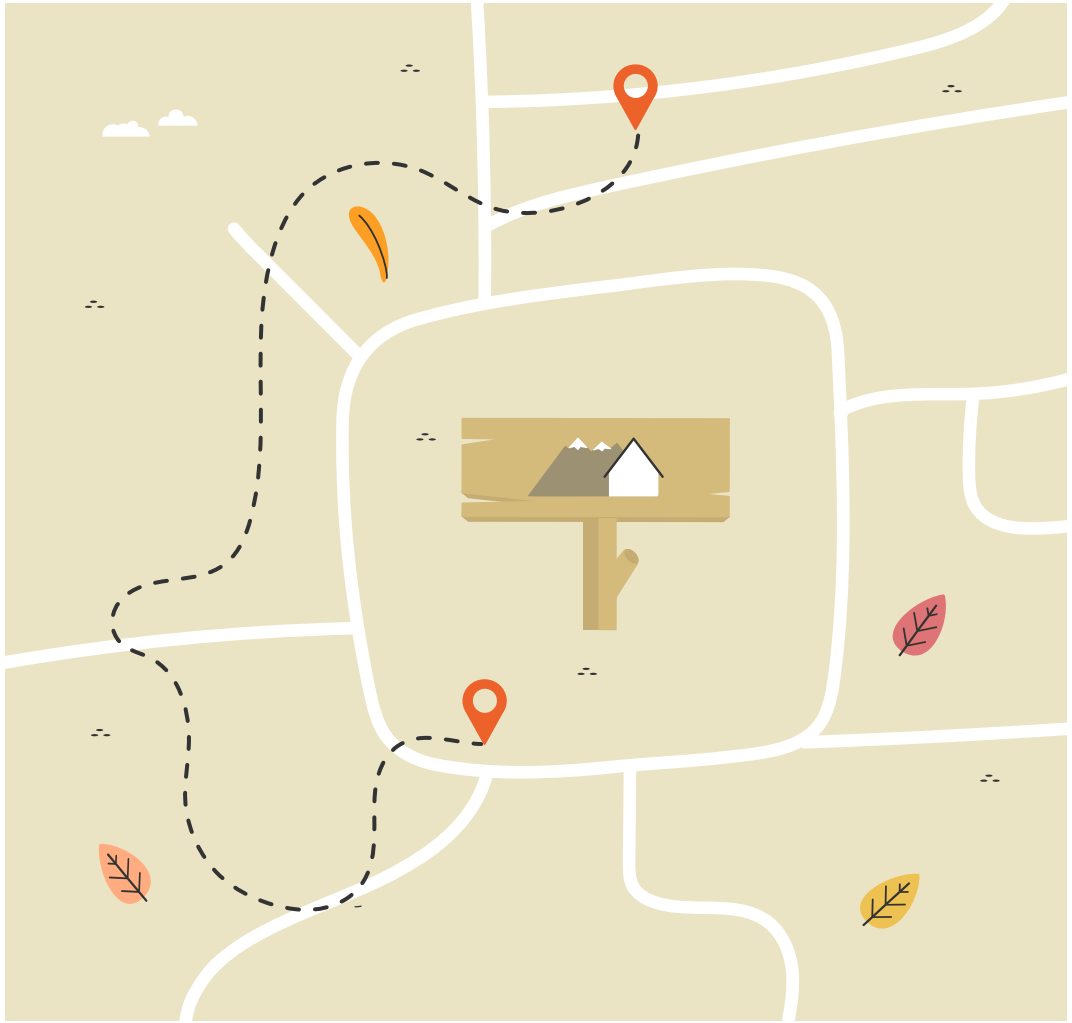
02

Data Analysis

03

Reflection & Future Steps

04



01

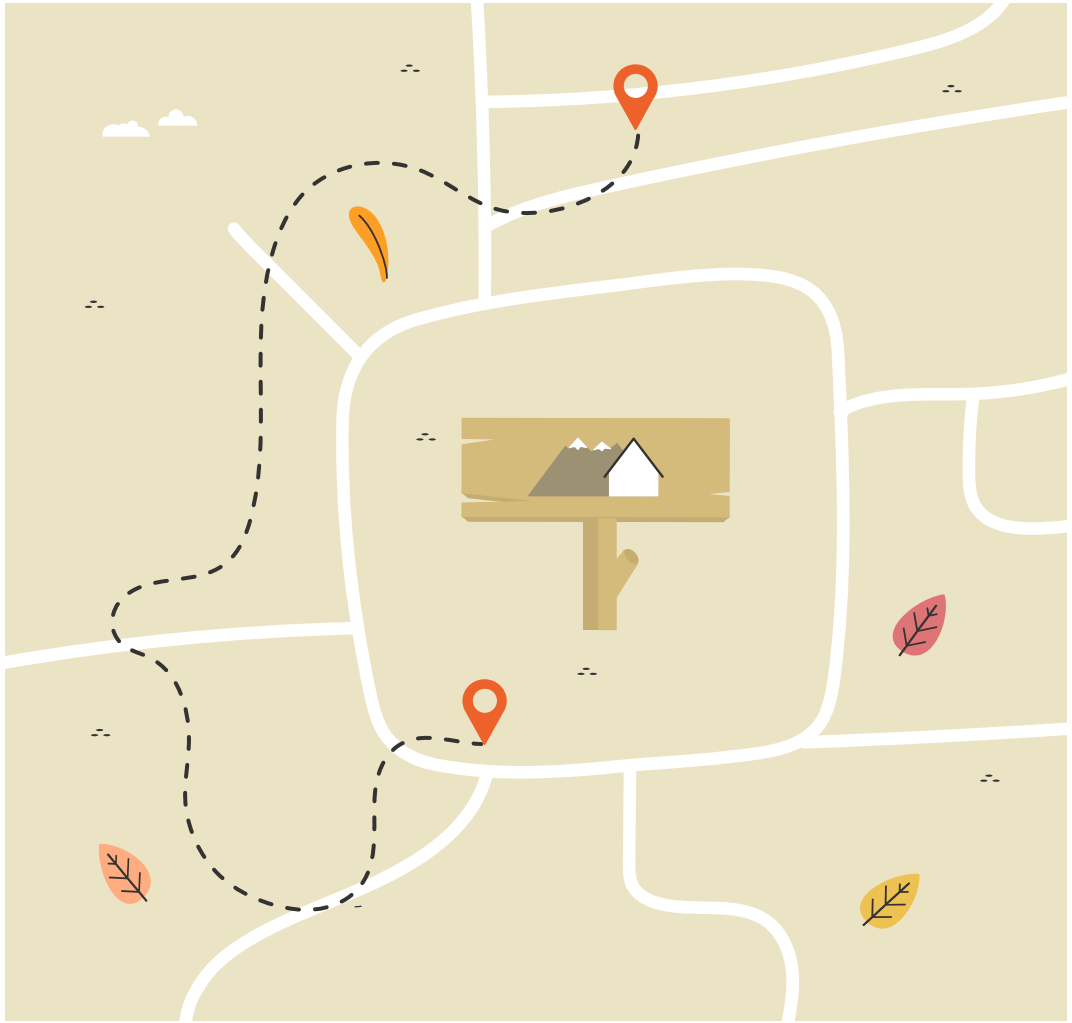
Business Proposition

MassiveData

Business Proposition and Value

- **Mission**: We want to provide a clear and focused analysis about what features of the homestay influences Airbnb guests to rate it and therefore influence the Airbnb hosts' profit at the end of the project.

- **Potential Clients**: Airbnb hosts and potential hosts in New York City who have a house type as an entire room and want to maximize their profit.

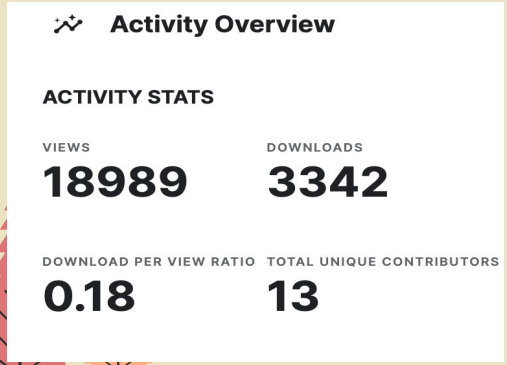


02

Database Design

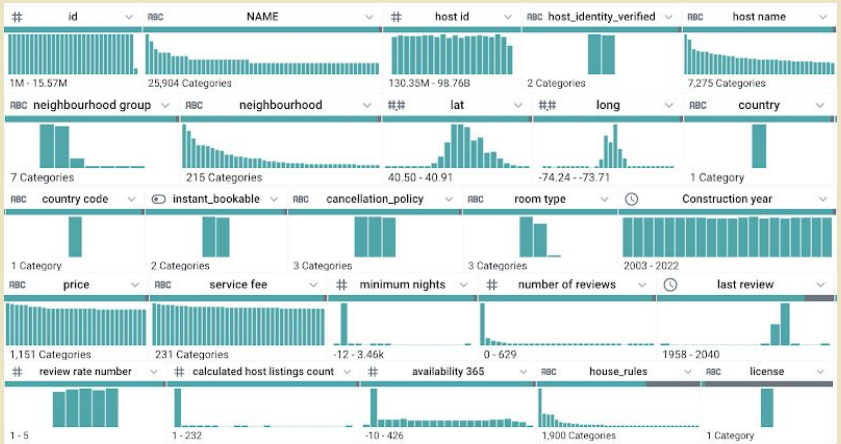
Data Overview

- **Data Source**: It has 18989 views and 3342 downloads on Kaggle. The original source of this dataset is Inside Airbnb, Inside Airbnb is a mission driven project that provides data and advocacy about Airbnb's impact on residential communities.



Column	Data Type	Description
HouseId	VARCHAR	/
HostId	VARCHAR	/
LocationID	VARCHAR	/
PolicyID	VARCHAR	Zip code belongs
HostIdentityVerified	VARCHAR	/
NeighbourhoodGroup	VARCHAR	Borough in Newyork
Neighbourhood	VARCHAR	Specfic area in each District
InstantBookable	TINYINT	/
CancellationPolicy	VARCHAR	/
ConstructionYear	DATE	/
Price	DECIMAL	/
ServiceFee	DECIMAL	/
MinNights	INT	Minimum number of rooms needs to order
NumOfReviews	INT	Number of reviews on Airbnb APP
LastReview	DATE	/
ReviewsPerMonth	DECIMAL	/
ReviewRateNumber	DECIMAL	/
CalHostListingsCount	INT	Number of advertisements
Availability365	INT	/

Data Preparation



Initial Data Quality

Final Data Quality

①Drop meaningless variables → **②Narrow down the range** → **③Imputation and deletion**

Final Data Quality

Accuracy: The data should be representative of actual, real-world scenarios. The incorrect data, 6/16/2040, and other incorrect dates in column “Last review” is deleted in the data preparation process to make the final data accurate.

Completeness: Completeness measures how well the data can give all of the required values that are currently available. After the data preparation process, by deleting and by setting the missing value to NULL or 0. There is no incompleteness data detected.

Consistency: When identical data values are kept in separate locations, they shouldn't clash with one another. After rechecking all 26 columns, there is no inconsistent data.

Validity: Information ought to be collected concurring to the proper format and dropped inside the correct range. During the data preparation process, the column “House Rules” is deleted, because it is text data and thus hard to analyze.

Uniqueness: Uniqueness guarantees that no values are duplicated or overlapped across all data sets. The column “Id” is kept instead of “Name”, which is easier to put in the SQL for analysis. And “Country”, and “Country Code” are removed.

Timeliness: Data needs to be updated in time to ensure that it is always available and accessible. In “Last Review”, 2018 to 2021 are kept, and other stale data and incorrect data of data is removed from the dataset.

Database Design

①Create six tables ⟹ ②Recoginize primary and foreign keys ⟹ ③Schemas ⟹ ④EER diagram

HostList

HostID	HostIdVerified
85098326012	unconfirmed
92037596077	verified
45498551794	verified
12801430940	verified
18824631834	verified
46551725984	verified
88653822946	verified
50357575975	verified
38981444696	unconfirmed

LocationList

LocationID	Neighborhood
01022323	Clinton Hill
00022399	East Harlem
02222323	Murray Hill
01122323	Chinatown
00023423	Upper West Side

NeighborhoodList

Neighborhood	NeighborhoodGroup
Clinton Hill	Brooklyn
East Harlem	Manhattan
Murray Hill	Manhattan
Chinatown	Manhattan
Upper West Side	Manhattan

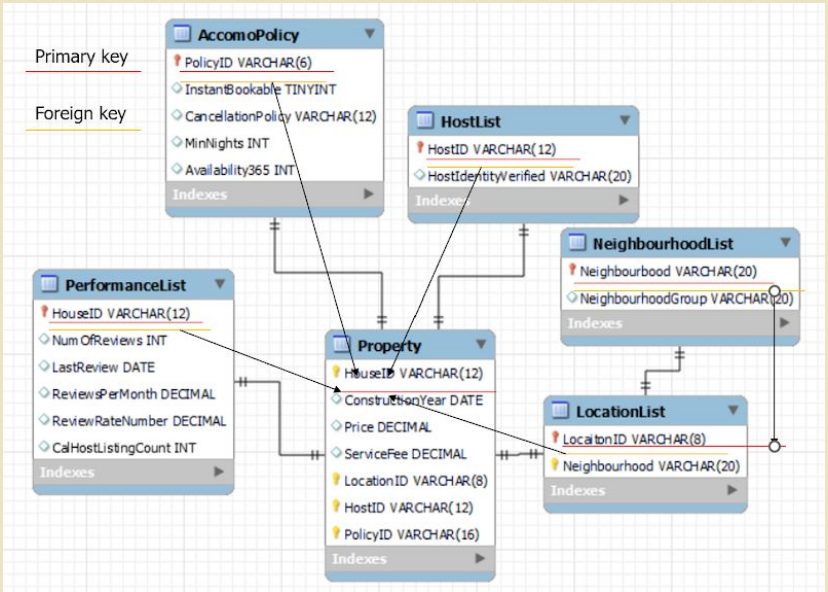
AccomoPolicy

PolicyID	InstantBookable	CancellationPolicy	MinNights	Availability365	ReviewRate	CalHostListingsCount
01022323	Clinton Hill	270	7/5/19	4.64	4	1
00022399	East Harlem	9	11/19/18	0.1	3	1
02222323	Murray Hill	74	6/22/19	0.59	3	1
01122323	Chinatown	160	6/9/19	1.33	3	4
00023423	Upper West Side	53	6/22/19	0.43	4	1

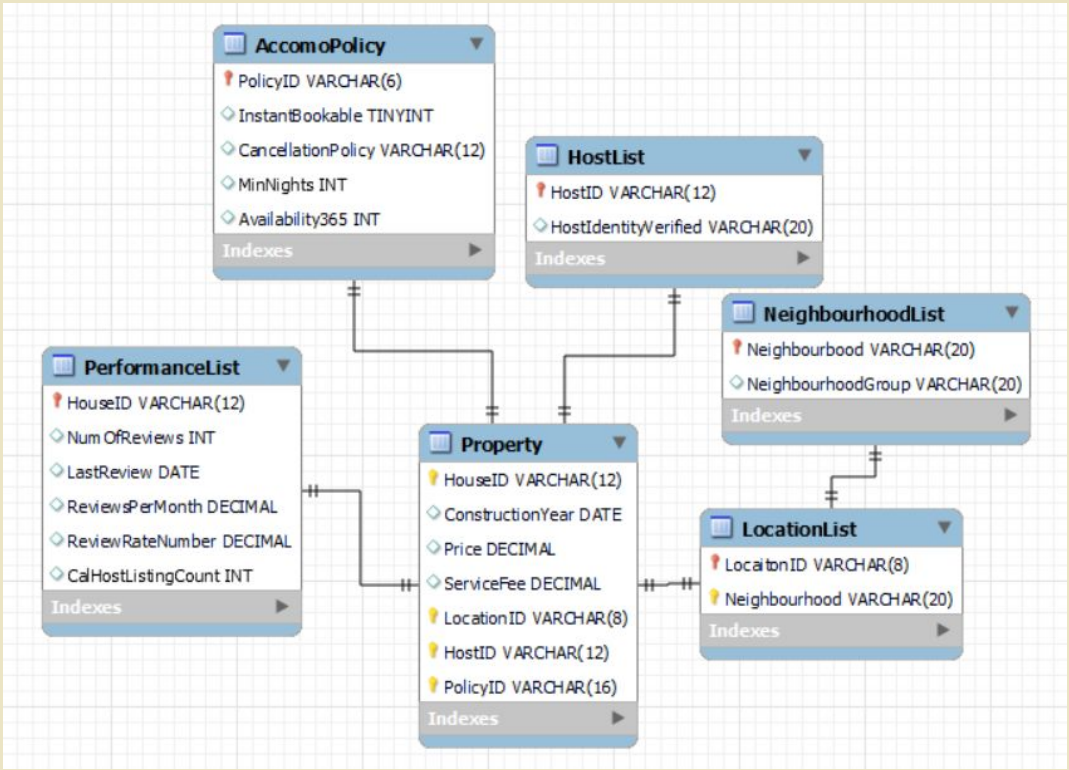
HouseID	ConstructionYear	Price	ServiceFee	LocationID	HostID	PolicyID
85098326012	2005	368.00	74.00	100003300012	300002234542	003663
92037596077	2009	204.00	41.00	100003300032	320002234542	003263
45498551794	2013	577.00	115.00	100003303012	303002234542	013663
12801430940	2004	319.00	64.00	102003300012	300002234549	003683
18824631834	2008	606.00	121.00	100003308012	300012234542	103663
46551725984	2018	578.00	124.00	100603300012	300002236542	003669
88653822946	2005	1,097.00	219.00	120003300012	300002274542	013689
50357575975	2020	370.00	74.00	100003300019	302002234542	003113

PerformanceList

LocationID	Neighborhood	NoOfReviews	LastReview	ReviewsPerMonth	ReviewRate	CalHostListingsCount
01022323	Clinton Hill	270	7/5/19	4.64	4	1
00022399	East Harlem	9	11/19/18	0.1	3	1
02222323	Murray Hill	74	6/22/19	0.59	3	1
01122323	Chinatown	160	6/9/19	1.33	3	4
00023423	Upper West Side	53	6/22/19	0.43	4	1



EER


AccomoPolicy
PolicyID VARCHAR(6)
InstantBookable TINYINT
CancellationPolicy VARCHAR(12)
MinNights INT
Availability365 INT
Indexes
HostList
HostID VARCHAR(12)
HostIdentityVerified VARCHAR(20)
Indexes
NeighbourhoodList
Neighbourbood VARCHAR(20)
NeighbourhoodGroup VARCHAR(20)
Indexes
PerformanceList
HouseID VARCHAR(12)
NumOfReviews INT
LastReview DATE
ReviewsPerMonth DECIMAL
ReviewRateNumber DECIMAL
CalHostListingCount INT
Indexes
Property
HouseID VARCHAR(12)
ConstructionYear DATE
Price DECIMAL
ServiceFee DECIMAL
LocationID VARCHAR(8)
HostID VARCHAR(12)
PolicyID VARCHAR(16)
Indexes
LocationList
LocaitonID VARCHAR(8)
Neighbourhood VARCHAR(20)
Indexes

Normalized 1, 2, 3NF

Property

HouseID	ConstructionYear	Price	ServiceFee	LocationID	HostID	PolicyID
85098326012	2005	368.00	74.00	100003300012	300002234542	003663
92037596077	2009	204.00	41.00	100003300032	320002234542	003263
45498551794	2013	577.00	115.00	100003303012	303002234542	013663
1280143094	2004	319.00	64.00	102003300012	300002234549	003683
18824631834	2008	606.00	121.00	100003308012	300012234542	103663
46551725984	2018	578.00	124.00	100603300012	300002236542	003669
88653822946	2005	1,097.00	219.00	120003300012	300002274542	013689
50357575975	2020	370.00	74.00	100003300019	302002234542	003113
38981444696	2017	589.00	118.00	100103300012	300002234592	013680

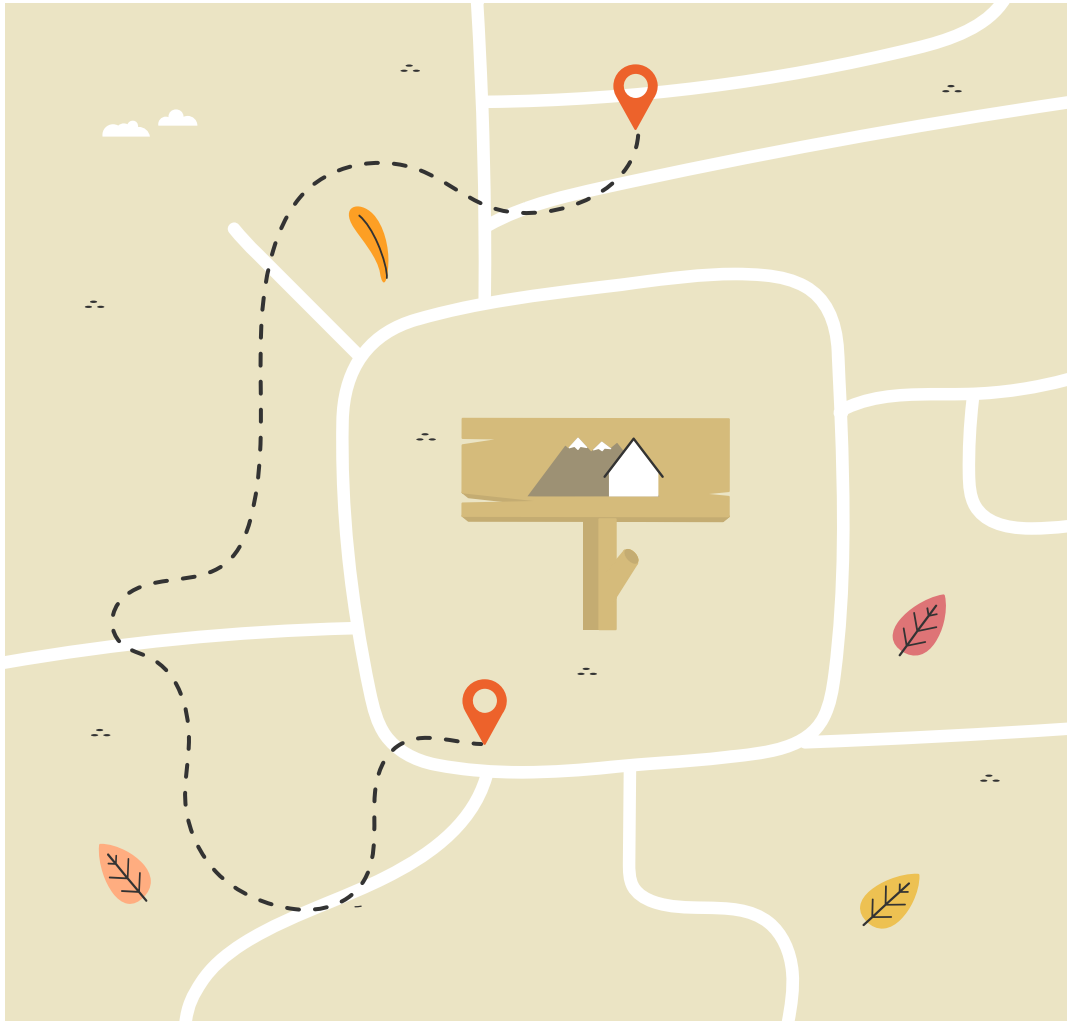
Property						
HouseID pk	HostID	LocationID	Construction_ Year	Price	Service_ Fee	PolicyID
1002755	85098326012	1	2005	$368.00	$74.00	A
1003689	92037596077	2	2009	$204.00	$41.00	B
1004098	45498551794	3	2013	$577.00	$115.00	C
1006859	1280143094	4	2004	$319.00	$64.00	D
1007411	18824631834	5	2008	$606.00	$121.00	E

LocationList

LocationID	Neighborhood
01022323	Clinton Hill
00022399	East Harlem
02222323	Murray Hill
01122323	Chinatown
00023423	Upper West Side

NeighborhoodList

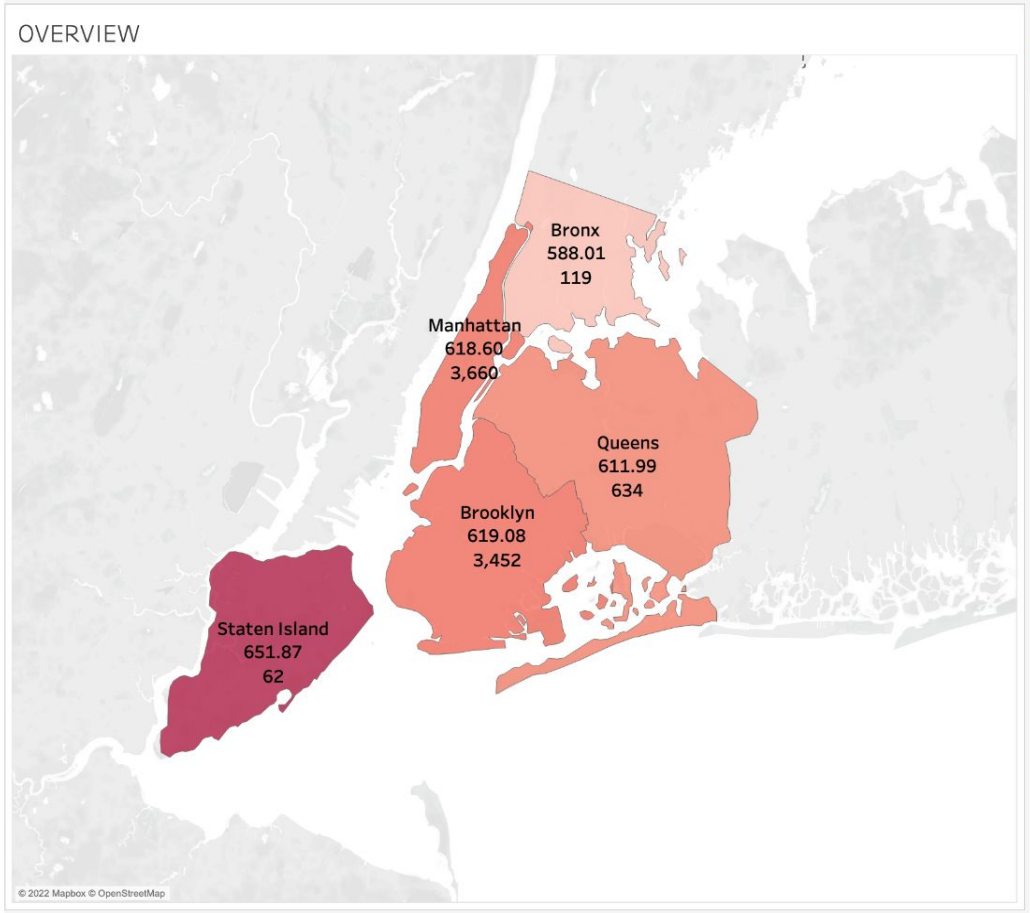
Neighborhood	NeighborhoodGroup
Clinton Hill	Brooklyn
East Harlem	Manhattan
Murray Hill	Manhattan
Chinatown	Manhattan
Upper West Side	Manhattan

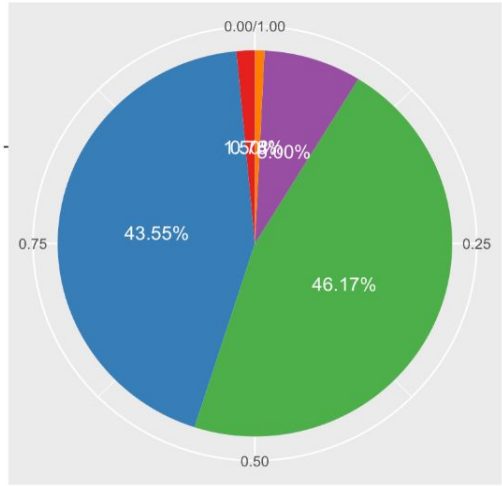


03

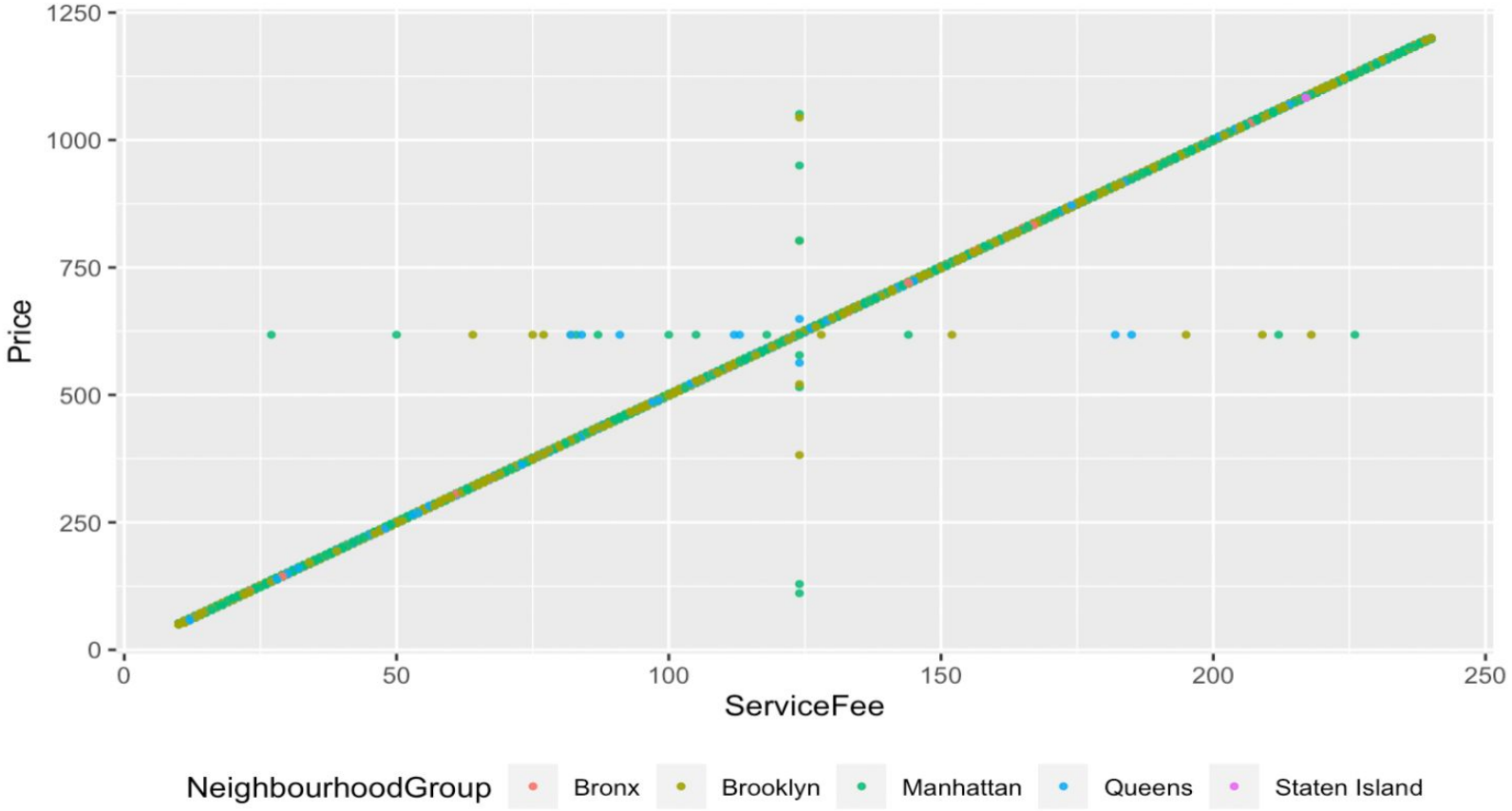
Data Analysis

NYC Airbnb Overview





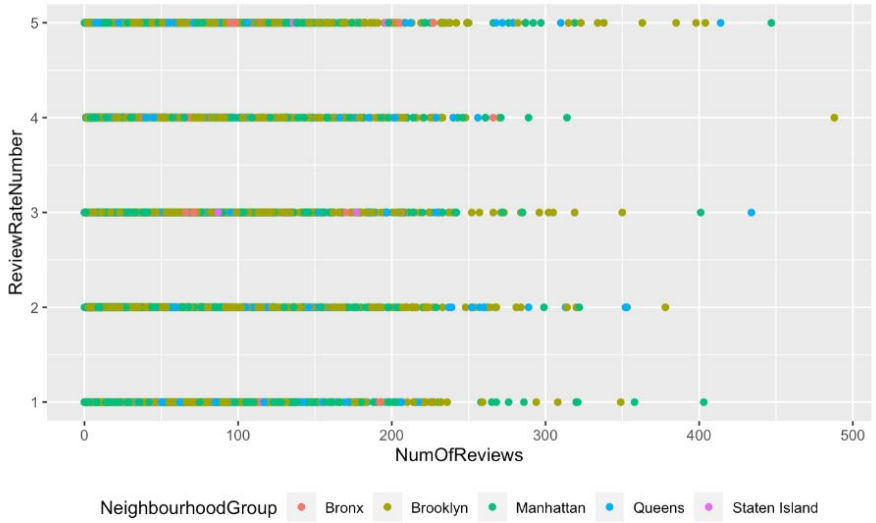
Price vs. Service Fee

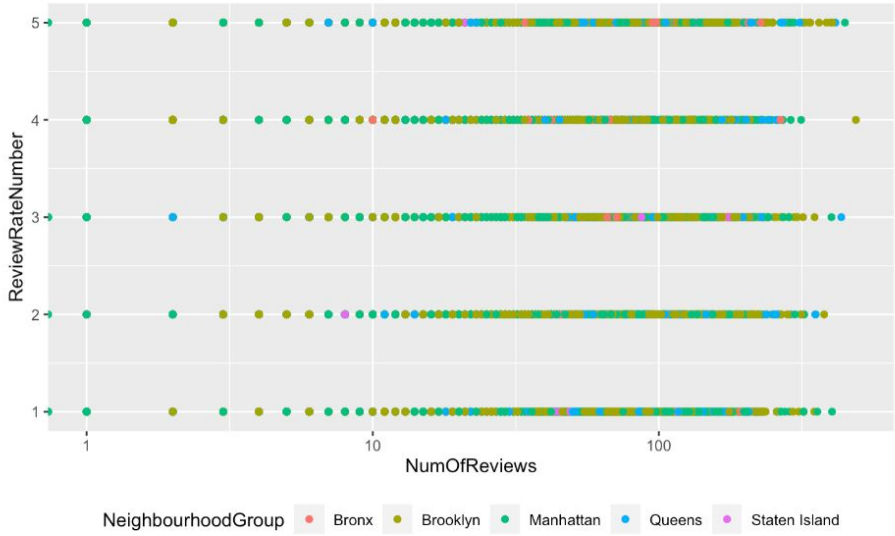


-

-

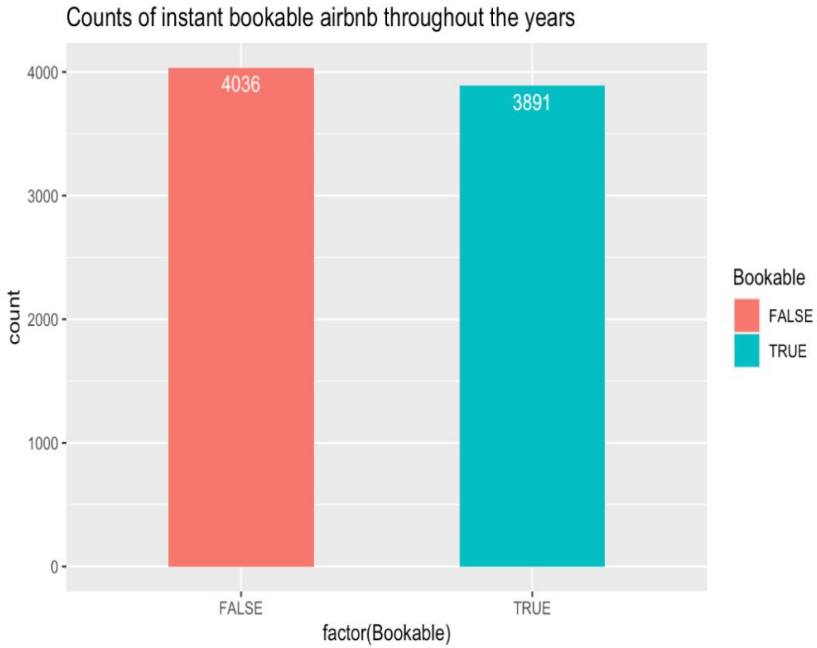
Number of Reviews vs. Review Rate Number

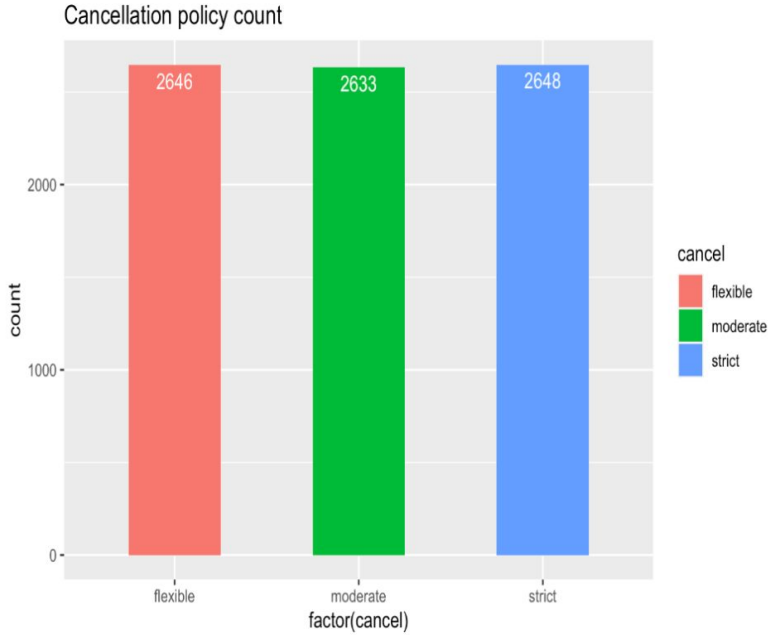




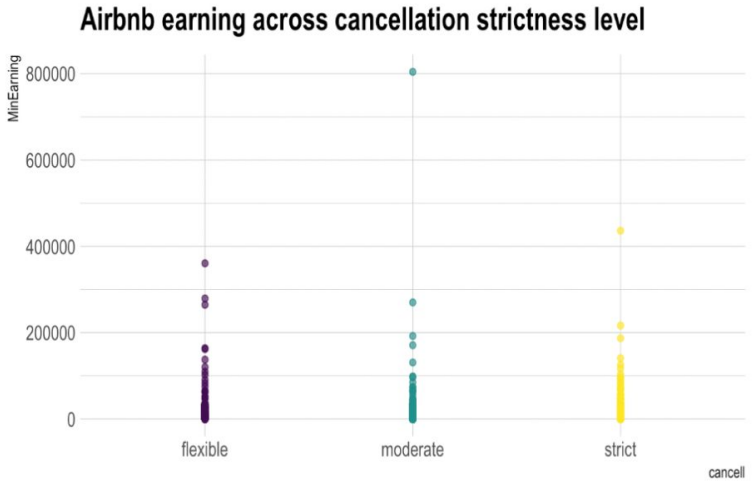
-

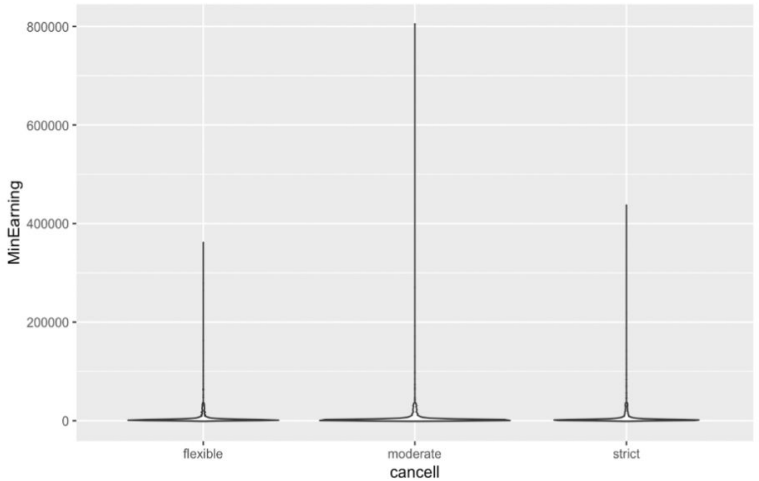
Accommodation policies don’t tend be a strong indicator of airbnb owners’ earning





Accommodation policies don’t tend be a strong indicator of airbnb owners’ earning

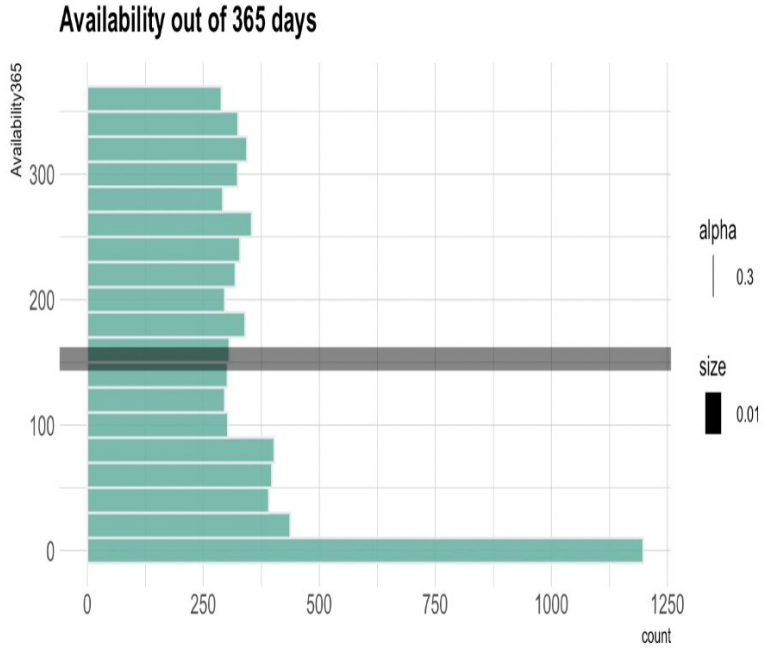




The cancellation policy seems to be evenly distributed, most of the apartments have a strict policy regarding cancellations. When it comes to revenue (#ofnights *price), an apartment that has a flexible policy tends to have more stable profits even though three categories are nicely distributed.

```
minProfit<- airbnb %>%
  mutate(profit = airbnb$MinNights* airbnb$Price)
MinEarning<-minProfit$profit
```

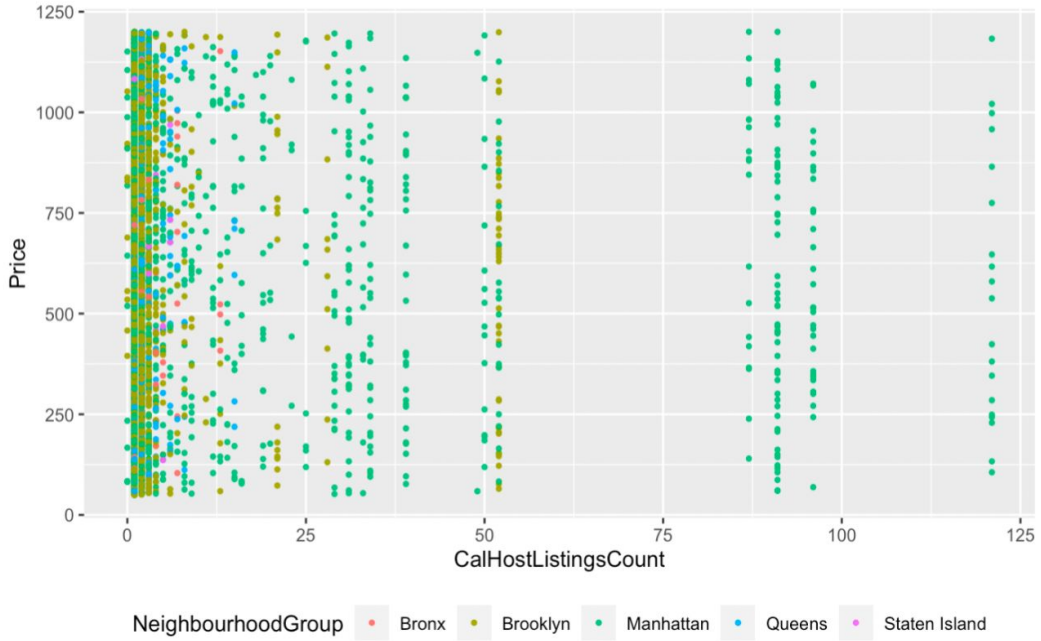
Accommodation policies don’t tend be a strong indicator of airbnb owners’ earning



A lot of Airbnb has availabilities of 0 days which can be strongly influenced by covid-19 and inactivity of business etc…

The grey line indicated the mean available days around 150 /365. Most Airbnb are below the mean available days.

Price Vs. Host Listing Count



Most Airbnb get a listing number between 0-50

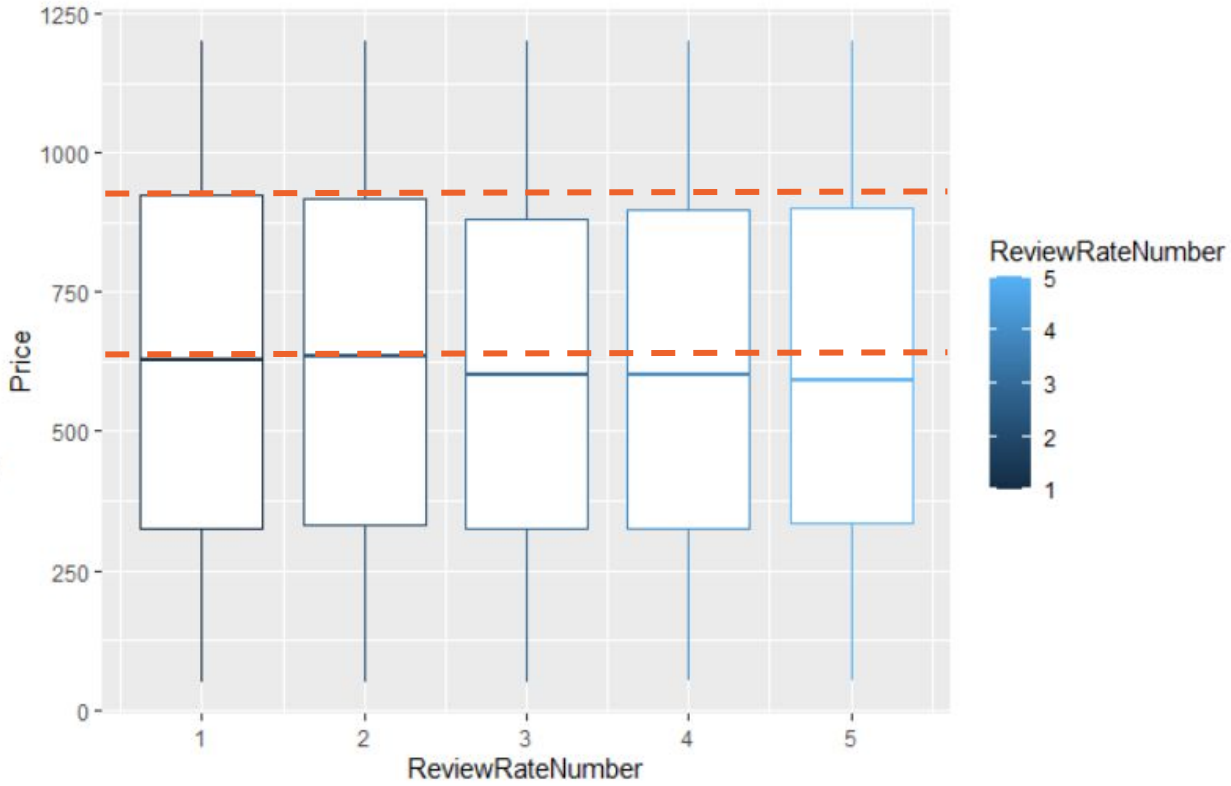
The majority of Brooklyn Airbnb get a listing of 0

Manhattan gets the most listing counts compared to the rest of the neighborhoods.

Price vs. rating number

Middle point of the price for 5 star is the lowest, the third quantile of the price for 1 star is highest.

A reasonable price may lead to a higher review rating, which cause more guests selected your homestays.

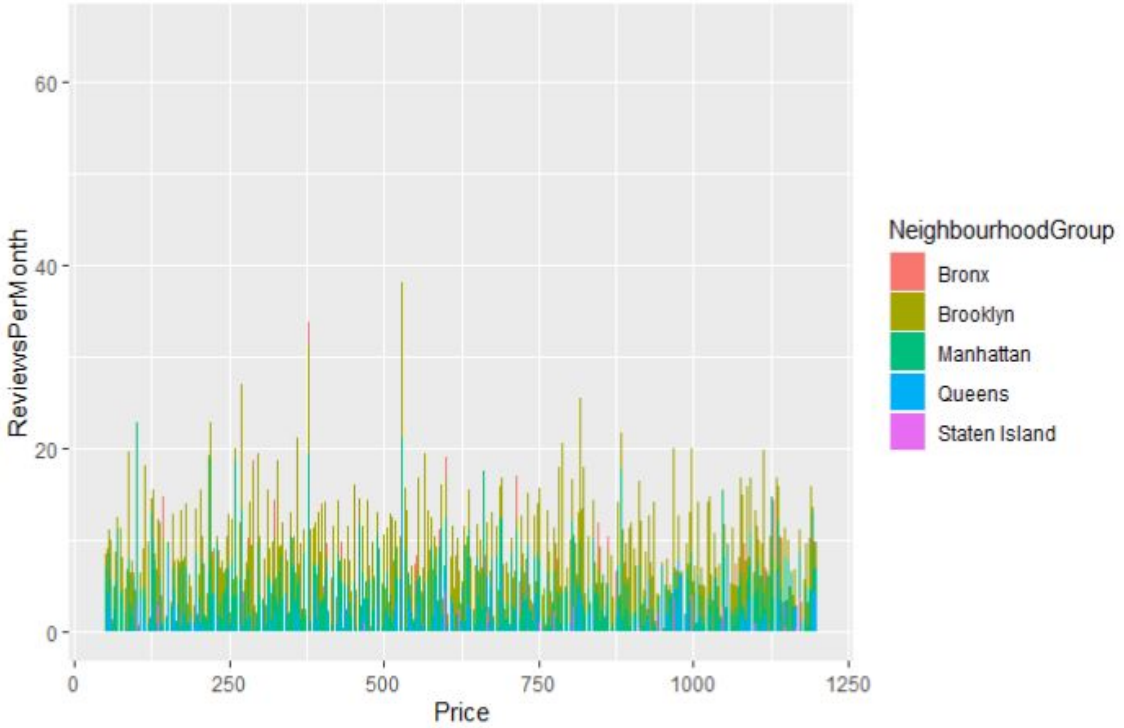


Price vs. reviews per month

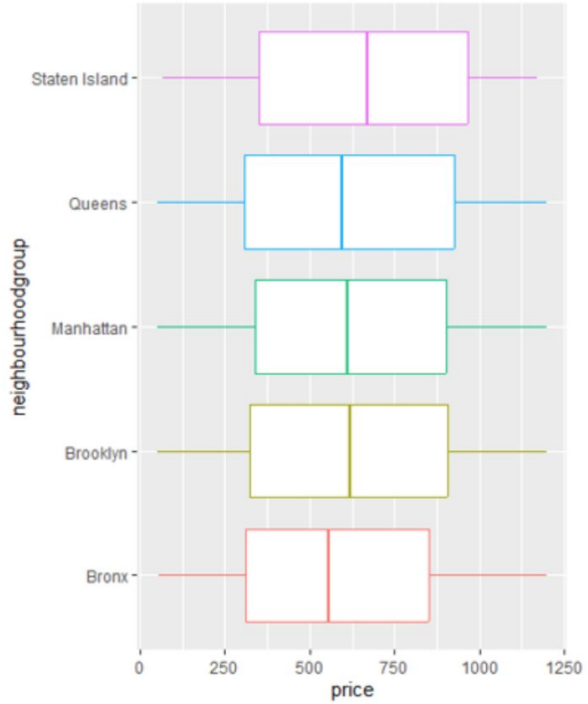
Indirectly test the airbnb guests activity;

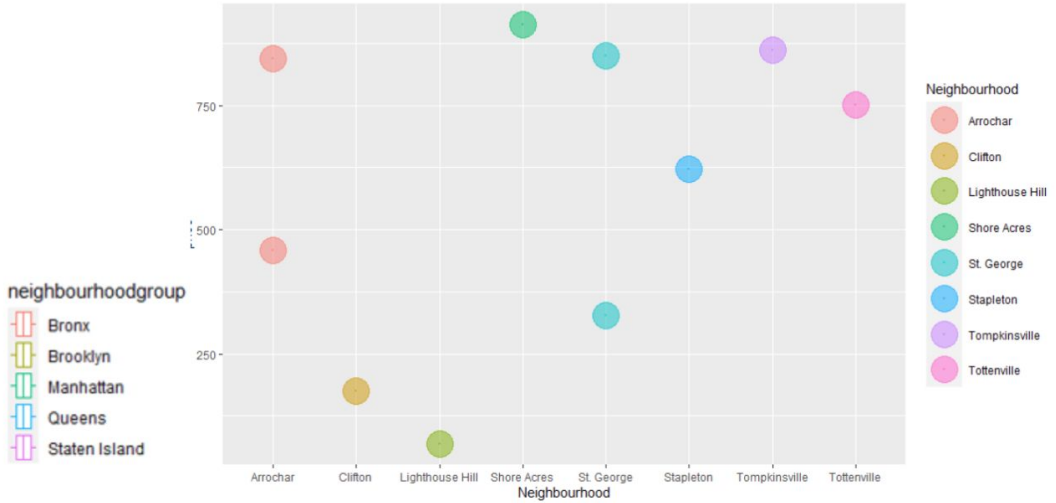
Tread that the reviews per month is higher when the price in the mid-low range.

More reviews per month could support the credit of homestays rating number.



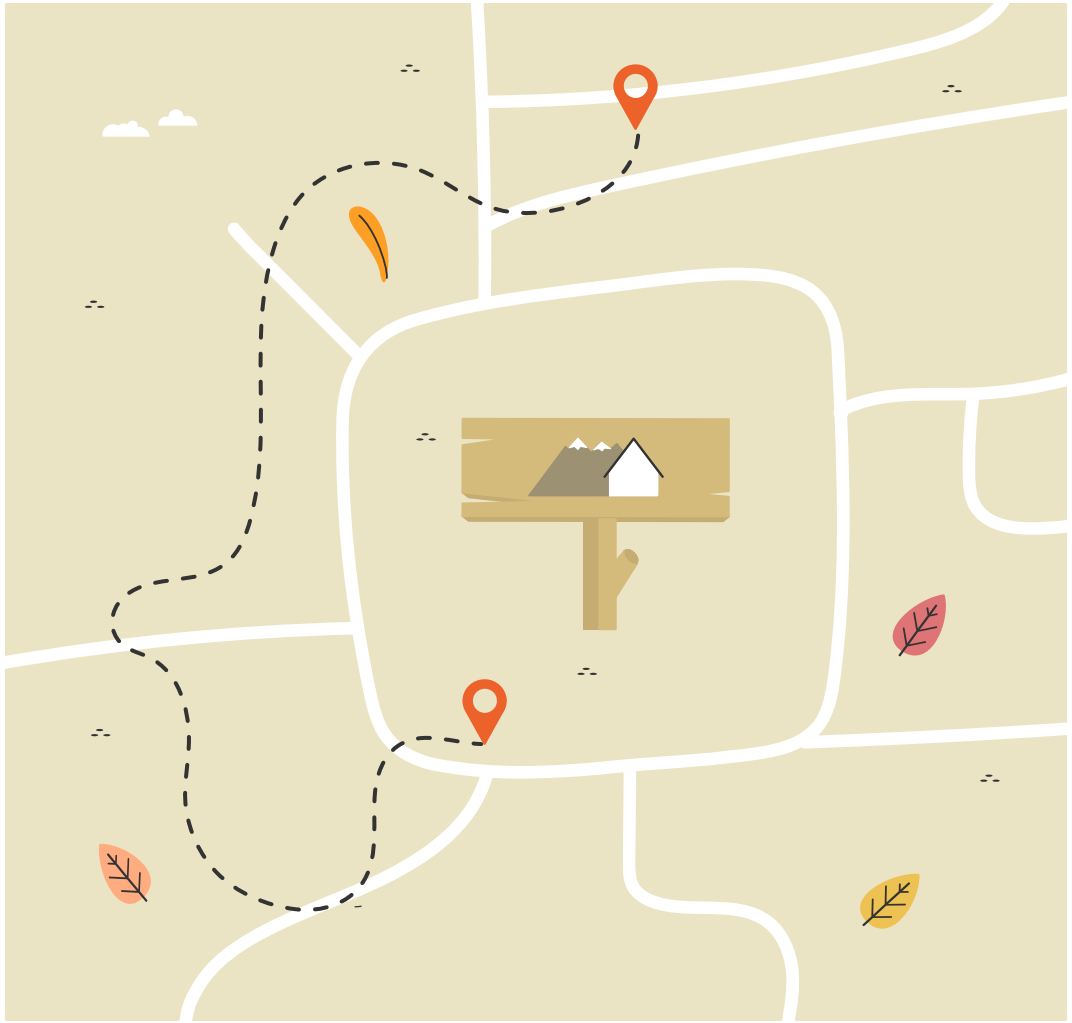
Price vs. Neighborhood





Staten Island not only has the highest average price, but also the highest median, and the largest spread.

Shore Acres has the highest price, while Lighthouse Hill has the lowest.



04

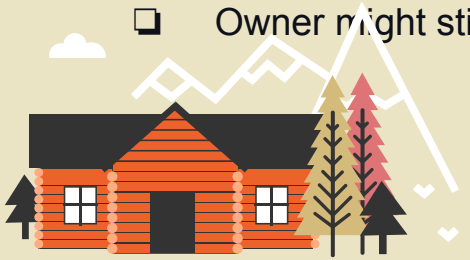
Reflection & Future Steps

Reflection & Future Steps

- More guests will be attracted by the reviews and high ratings and choose your homestay, especially price per night is around the average price
  - Develop public image, publicity
  - Increase their service and customer loyalty

Reflection & Future Steps

- Manhattan is still the most popular area to start an Airbnb business. Staten Island is still unexplored and pricing on Airbnb units are higher than another region
  - Even though the competition is high, Manhattan gets more attention by listing on the website.
    - Ads listing
  - Potential owner might not invest in Staten island
- Accommodation policy doesn't appear to be a determinator of businesses' financial performances
  - Owner might still want to implement strictness for liability purpose



Thanks!

Do you have any questions?