MILESTONE 4 - GROUP 3

Ronald Adomako
Idris Dobbs
Narendra Chigili
Nikhil Srirama Sai

AGENDA

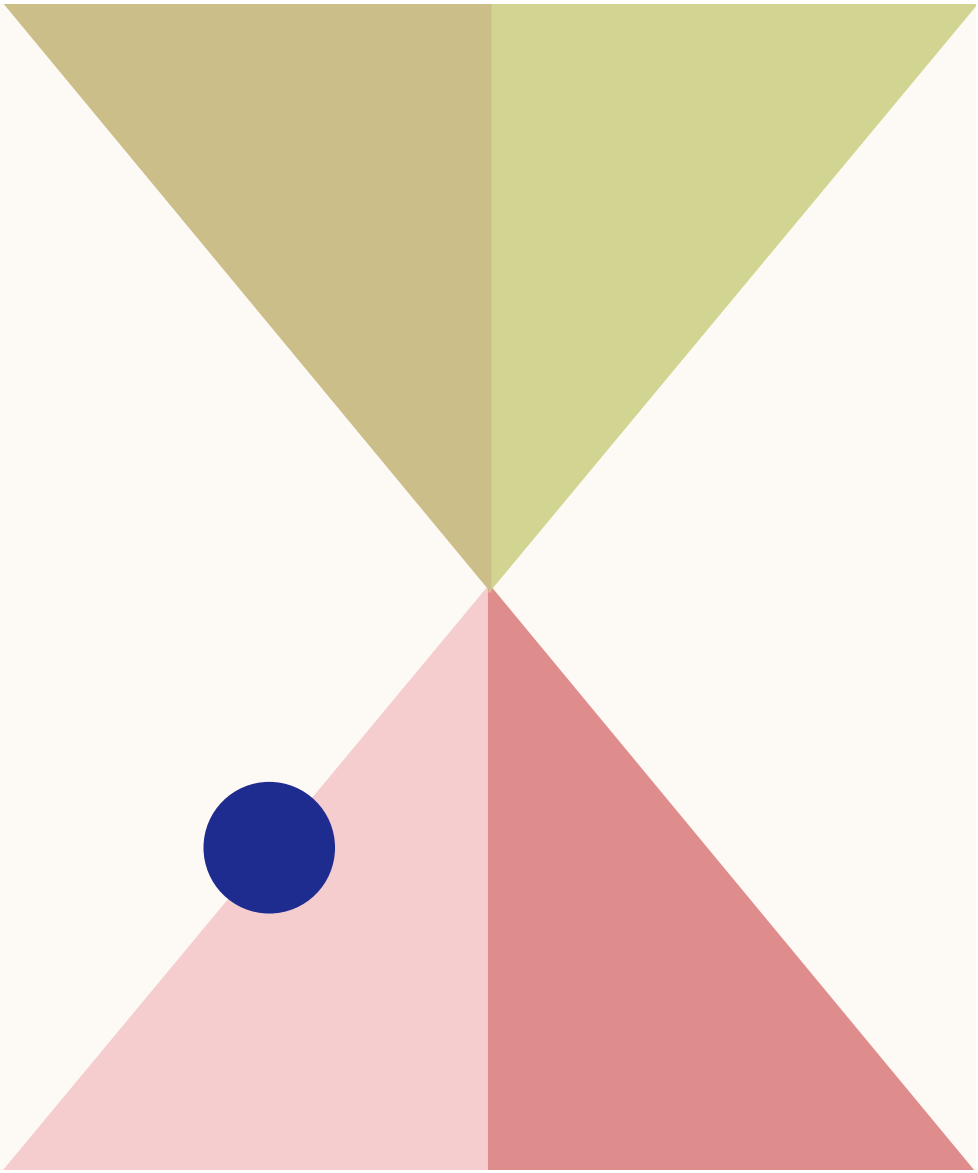
Teamwork

Background

Data Collection and Cleaning

Data Analyses Results

Conclusion



MEET OUR TEAM

Histogram analysis, feature extraction for statistics, leading team meetings, interpreting results for logistic regression and PCA, data processing and cleaning.

RONALD ADOMAKO

Software Engineer

Data cleaning & analysis by room type & neighborhood group/Choropleth mapping / outlier analysis/ box & whisker plots/ PCA

IDRIS DOBBS

Economist

Data visualization, Data cleaning, K-NN regression analysis & K-NN Classification, Test Accuracy Visualization

NARENDRA CHIGILI

Student

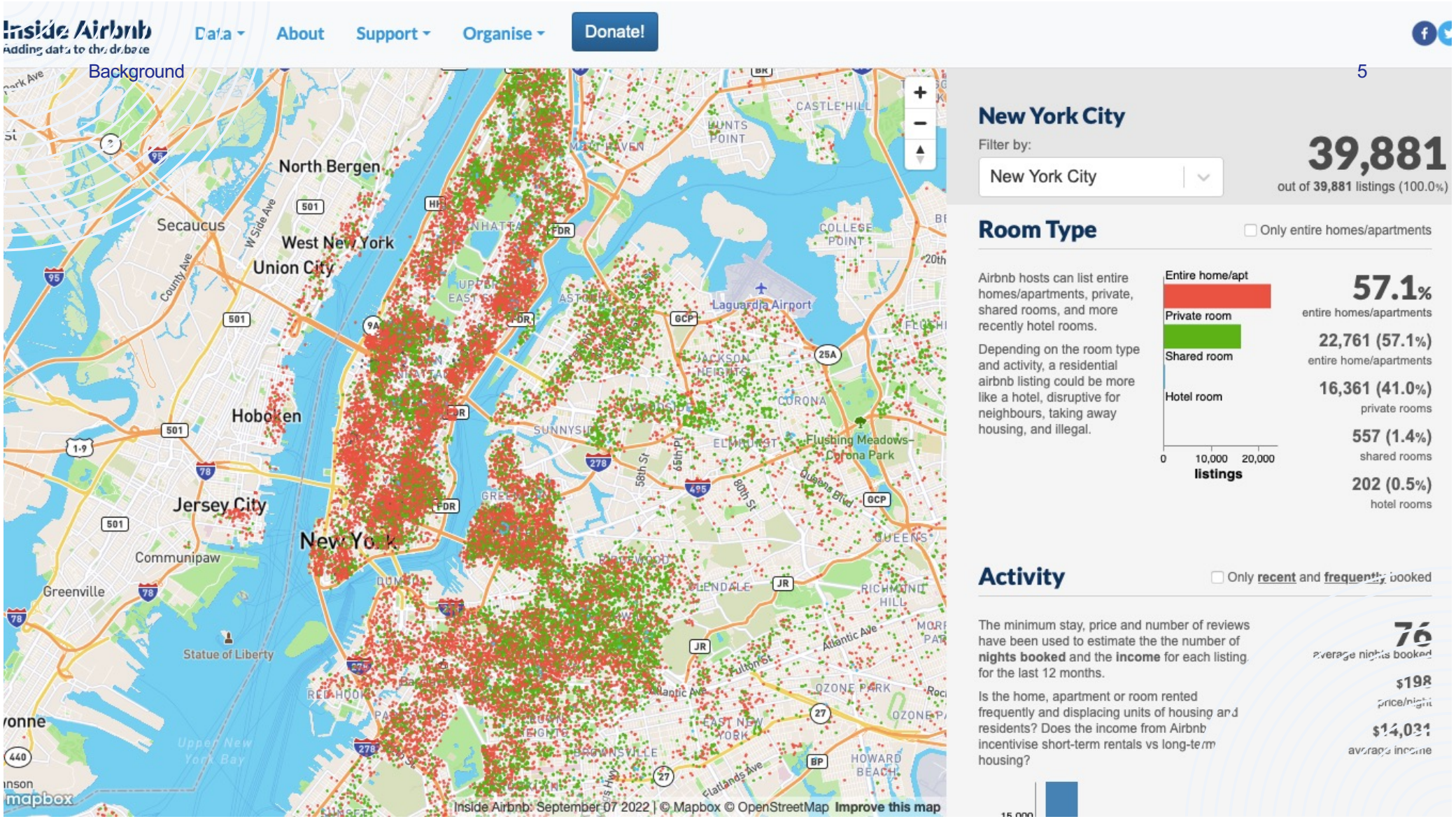
Extracting Zip codes using latitudes and longitudes/naïve bayes classification/ visualizatio n and analyzation of data using histogram.

NIKHIL SRIRAMA SAI

Student

BACKGROUND

Group-3



New York City

Filter by:

New York City

39,881

out of **39,881** listings (100.0%)

Room Type

☐ Only entire homes/apartments

Airbnb hosts can list entire homes/apartments, private, shared rooms, and more recently hotel rooms.

Depending on the room type and activity, a residential airbnb listing could be more like a hotel, disruptive for neighbours, taking away housing, and illegal.

57.1%
entire homes/apartments

22,761 (57.1%)
entire home/apartments

16,361 (41.0%)
private rooms

557 (1.4%)
shared rooms

202 (0.5%)
hotel rooms

Activity

☐ Only **recent** and **frequently** booked

The minimum stay, price and number of reviews have been used to estimate the the number of **nights booked** and the **income** for each listing for the last 12 months.

Is the home, apartment or room rented frequently and displacing units of housing and residents? Does the income from Airbnb incentivise short-term rentals vs long-term housing?

76
average nights booked

$198
price/night

$14,031
average income

INTRODUCTION

AirBnB is a PaaS for the short term rental market. Users use the platform to list residences for short term rentals. We noticed that for big cities, such as New York City, there are many host and some may want to use AirBnB for lucrative means. Given location and characteristics of a property, a new host would want to know whether he or she is charging the optimal amount to rent the space to lodgers.

- What are the largest determinants / predictors of AirBnB rental prices?
- How can we optimize rental revenue based on rental location and other characteristics?
- What price should be charged based on rental location/ characteristics?

DATA COLLECTION & CLEANING

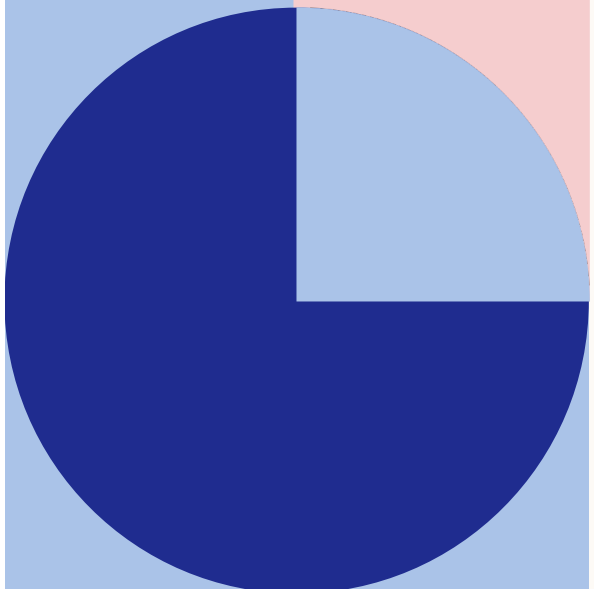
We want to build a model that indicates whether host are charging an optimal price for their AirBnB rental.

- http://insideairbnb.com/new-york-city/
- Inside Airbnb Data Dictionary
- http://insideairbnb.com/get-the-data
- http://data.insideairbnb.com/united-states/ny/new-york-city/2022-09-07/visualisations/listings.csv

DATA DICTIONARY

Other relevant data variables used:

- **'accommodates'** - # people the listing can accommodate
- **'bedrooms'** - # bedrooms in the property
- **'beds'** - # beds on the property



	Features	Description
0	id	Airbnb's unique identifier for the listing
1	name	Name of the listing
2	host_id	Airbnb's unique identifier for the host/user
3	host_name	Name of the host. Usually just the first name(s).
4	neighbourhood_group	Borough
5	neighbourhood	Neighborhood equivalent for zip code group
6	latitude	Uses the World Geodetic System (WGS84) projection for latitude and longitude.
7	longitude	Uses the World Geodetic System (WGS84) projection for latitude and longitude.
8	room_type	[Entire home/apt|Private room|Shared room|Hotel]
9	price	daily price in local currency
10	minimum_nights	minimum number of night stay for the listing (calendar rules may be different)
11	number_of_reviews	The number of reviews the listing has
12	last_review	The date of the last/newest review
13	reviews_per_month	The number of reviews the listing has over the lifetime of the listing
14	calculated_host_listings_count	The number of listings the host has in the current scrape, in the city/region geography.
15	availability_365	avaliability_x. The availability of the listing x days in the future as determined by the calend...
16	number_of_reviews_ltm	The number of reviews the listing has (in the last 12 months)
17	license	The licence/permit/registration number

DATA BY ROOM TYPE & BOROUGH

```
Int64Index: 39881 entries, 0 to 39880
Data columns (total 21 columns):
 #   Column                          Non-Null Count  Dtype
---  ------                          --------------  -----
 0   id                              39881 non-null  int64
 1   name                            39868 non-null  object
 2   host_id                         39881 non-null  int64
 3   host_name                       39831 non-null  object
 4   neighbourhood_group             39881 non-null  object
 5   neighbourhood                   39881 non-null  object
 6   latitude                        39881 non-null  float64
 7   longitude                       39881 non-null  float64
 8   room_type                       39881 non-null  object
 9   price                           39881 non-null  int64
 10  minimum_nights                  39881 non-null  int64
 11  number_of_reviews               39881 non-null  int64
 12  last_review                     31519 non-null  object
 13  reviews_per_month               31519 non-null  float64
 14  calculated_host_listings_count  39881 non-null  int64
 15  availability_365                39881 non-null  int64
 16  number_of_reviews_ltm           39881 non-null  int64
 17  license                         5 non-null      object
 18  accommodates                    39881 non-null  int64
 19  bedrooms                        36098 non-null  float64
 20  beds                            38997 non-null  float64
dtypes: float64(5), int64(9), object(7)
memory usage: 6.7+ MB
```

```
Int64Index: 39851 entries, 0 to 39880
Data columns (total 14 columns):
 #   Column                          Non-Null Count  Dtype
---  ------                          --------------  -----
 0   id                              39851 non-null  int64
 1   neighbourhood_group             39851 non-null  object
 2   latitude                        39851 non-null  float64
 3   longitude                       39851 non-null  float64
 4   room_type                       39851 non-null  object
 5   price                           39851 non-null  int64
 6   minimum_nights                  39851 non-null  int64
 7   number_of_reviews               39851 non-null  int64
 8   calculated_host_listings_count  39851 non-null  int64
 9   availability_365                39851 non-null  int64
 10  number_of_reviews_ltm           39851 non-null  int64
 11  accommodates                    39851 non-null  int64
 12  bedrooms                        36098 non-null  float64
 13  beds                            38997 non-null  float64
dtypes: float64(4), int64(8), object(2)
memory usage: 4.6+ MB
None
```

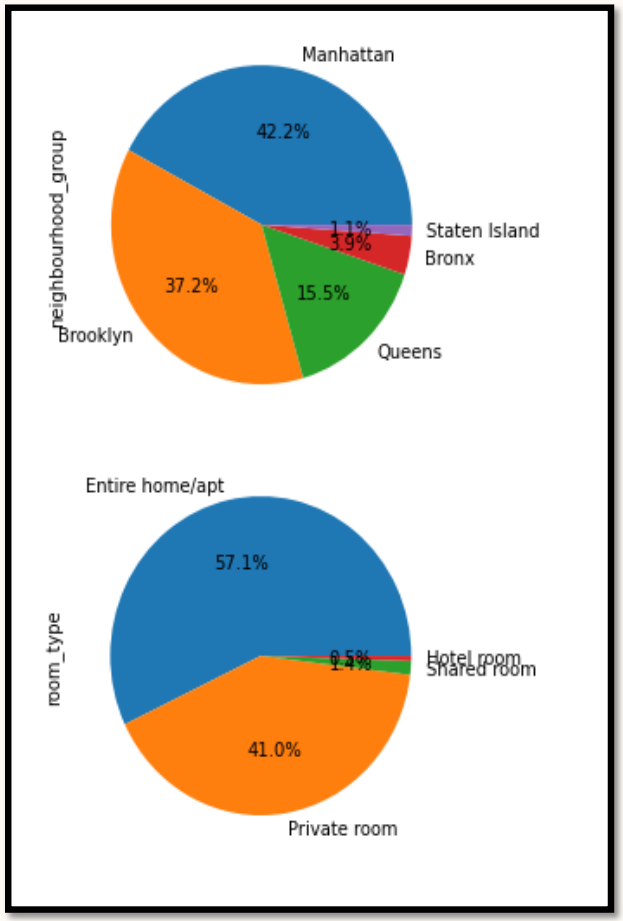
- Data Sources
  - http://data.insideairbnb.com/united-states/ny/new-york-city/2022-03-05/visualisations/listings.csv
  - http://data.insideairbnb.com/united-states/ny/new-york-city/2022-03-05/data/listings.csv.gz
- 21 variables including the index
- 39881 observations
- Variables added: accommodates, bedrooms, & beds
- Variables eliminated:
  - 'last_review' and 'reviews_per_month' have 31519 observations
  - License has 5 observations
  - Categorical variables that would cause dimensionality issues
    - Host name, hostid, neighbourhood(for now)
- Rationale: low information content or other variables that measure the same thing
- Reduced from 21 variables to 14
- Removed observatons with prices of zero

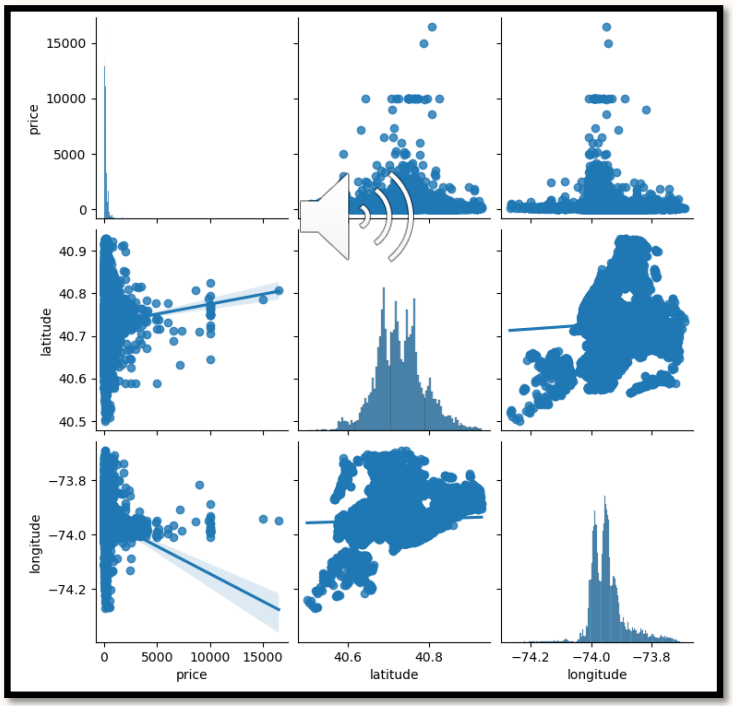
DATA ANALYSIS & RESULTS

Group-3

PCA

DATA ANALYSIS BY NYC BOROUGH & ROOM TYPE





Distribution by Borough & Room Type

- Nearly 80% of listings in Manhattan & Brooklyn
- Room Type Distribution
  - 57% - 'Entire home/apt'
  - 41% - 'Private room'

Latitude, Longitude & Price

- Latitude or longitude did not seem to show a significant relationship with Price

DETERMINE & REMOVE THE OUTLIERS

Data Distributed by Room Type

room_type	count	mean	std	min	25%	50%	75%	max	IQR	upper_outlier	lower_outlier
Entire home/apt	22761.0	251.546022	338.044654	10.0	128.00	180.0	275.0	15000.0	147.00	495.500	-92.500
Hotel room	172.0	436.470930	282.491141	100.0	258.75	308.0	501.5	1998.0	242.75	865.625	-105.375
Private room	16361.0	122.936495	356.373737	10.0	55.00	75.0	110.0	16500.0	55.00	192.500	-27.500
Shared room	557.0	119.398564	454.106078	10.0	42.00	66.0	92.0	10000.0	50.00	167.000	-33.000

- Outliers determined by 'room_type'
- Interquartile rule applied by room type
- All observations with a price of zero were removed
- Major Room Types ('Entire home/apt', 'Private room')
- Major NYC Burroughs ('Manhattan', 'Brooklyn)

Data Distributed by Room Type (outliers removed)

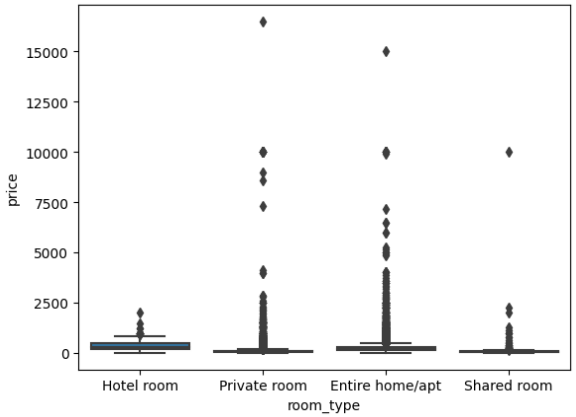
room_type	count	mean	std	min	25%	50%	75%	max
Entire home/apt	21117.0	196.250793	95.382230	10.0	125.0	175.0	250.0	496.0
Hotel room	155.0	363.490323	163.471288	100.0	248.5	307.0	470.0	826.0
Private room	14751.0	78.448648	34.294310	10.0	52.0	72.0	97.0	192.0
Shared room	501.0	66.179641	32.916921	10.0	40.0	60.0	85.0	165.0

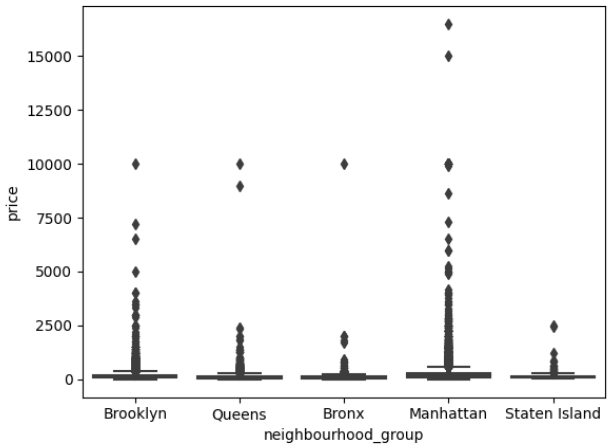
Data Distributed by NYC Borrough (outliers removed)

neighbourhood_group	count	mean	std	min	25%	50%	75%	max
Bronx	1530.0	108.036601	68.561010	10.0	60.0	89.0	135.00	450.0
Brooklyn	14192.0	133.914459	88.425233	10.0	69.0	110.0	175.00	529.0
Manhattan	14426.0	180.040690	105.605595	10.0	100.0	150.0	235.00	826.0
Queens	5954.0	114.189620	79.028227	10.0	60.0	90.0	145.00	495.0
Staten Island	422.0	113.646919	61.827482	33.0	69.0	99.5	140.75	412.0

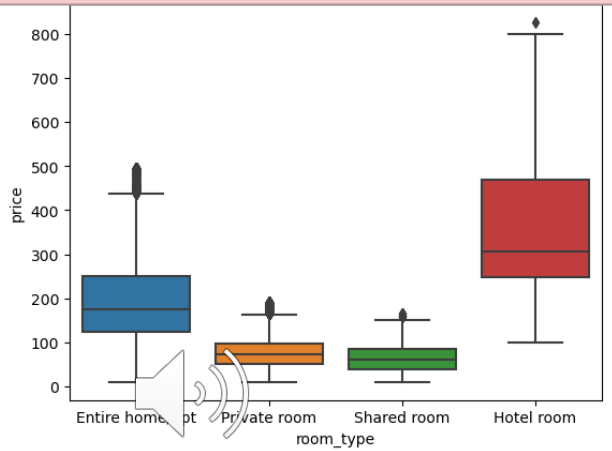
DATA ANALYSIS BY NYC BOROUGH & ROOM TYPE: OUTLIER ANALYSIS

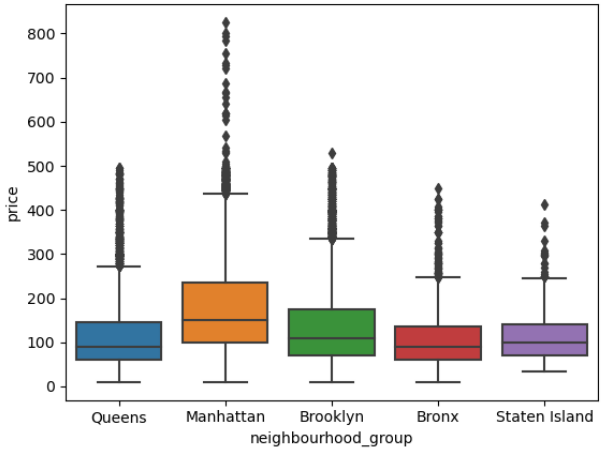
Box &Whiskers with outliers





Box &Whiskers with outliers removed

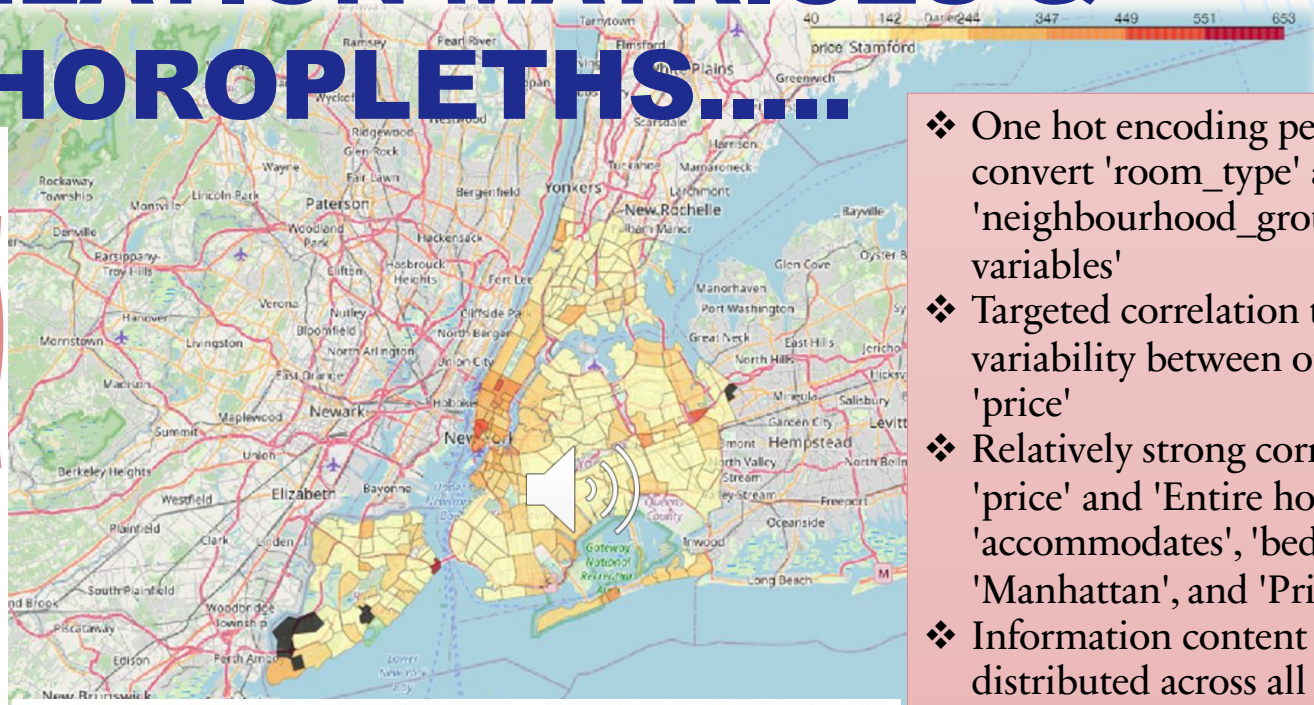


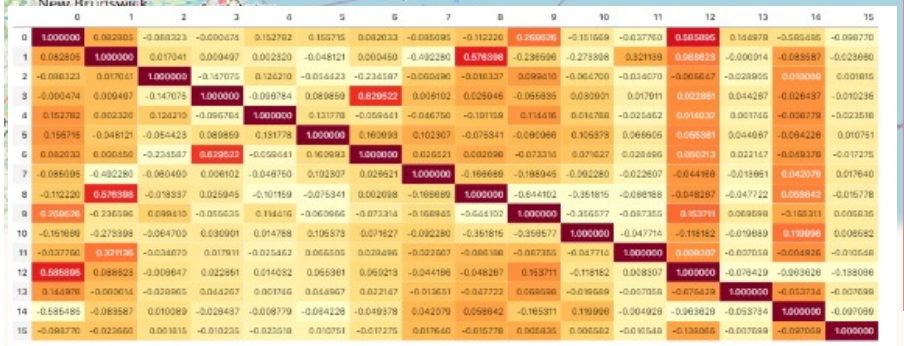


- Significant variability in the data prior to outlier removal.
- Significant variability still remains in the data even with outliers removed
- **Signals:**
  - Hotels & entire homes have higher prices
  - Manhattan has higher prices
  - Brooklyn has 2nd highest price, but not significantly higher than the other boroughs

CORRELATION MATRICES & CHOROPLETHS.....

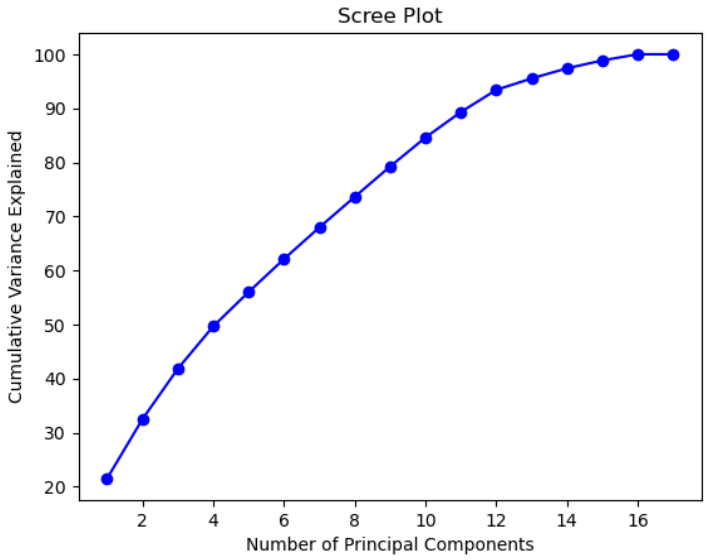
	price
price	1.000000
Entire home/apt	0.607875
accommodates	0.522097
bedrooms	0.422809
beds	0.397797
Manhattan	0.264734
availability_365	0.158387
Hotel room	0.153587
calculated_host_listings_count	0.138000
number_of_reviews_ltm	0.079055
number_of_reviews	-0.002248
Staten Island	-0.035411
Bronx	-0.082539
minimum_nights	-0.085153
Shared room	-0.100743
Brooklyn	-0.104841
Queens	-0.147348
Private room	-0.606871



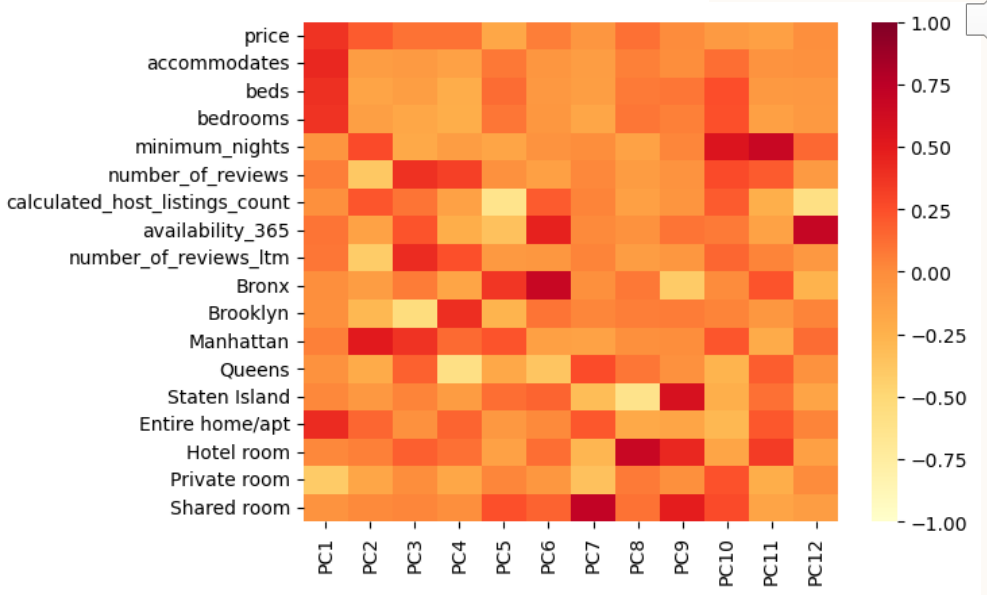


- One hot encoding performed to convert 'room_type' and 'neighbourhood_group' into 'dummy variables'
- Targeted correlation to measure variability between other variables and 'price'
- Relatively strong correlations between 'price' and 'Entire homes/apt', 'accommodates', 'bedrooms', 'beds', 'Manhattan', and 'Private room'
- Information content is widely distributed across all the principal loadings
- **Choropleth** map shows variability in price by neighborhood.

PCA: FINDINGS



- Max # of principle components = 17
- 12 principal components needed to describe ~94% of the variability in the data
- Effective marginal contribution to data variability ends with ~ 16 principal components



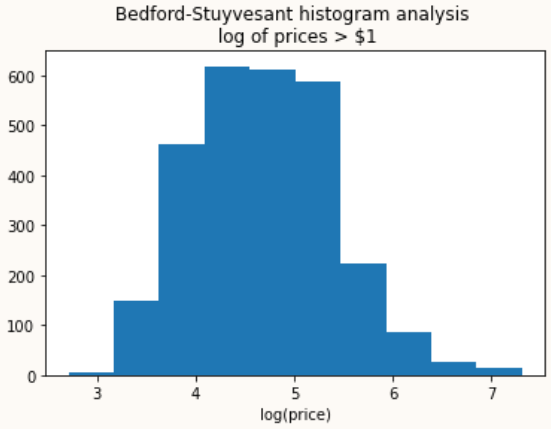
- Variables with greatest positive variability with 'price' are, 'Entire home/apt', accommodates, bedrooms, beds, & 'Manhattan'
- 'Private room' is negatively correlated with price
- Most variables have one or two dimensions which show some extreme variability
- Information content is widely distributed across all the principal loadings

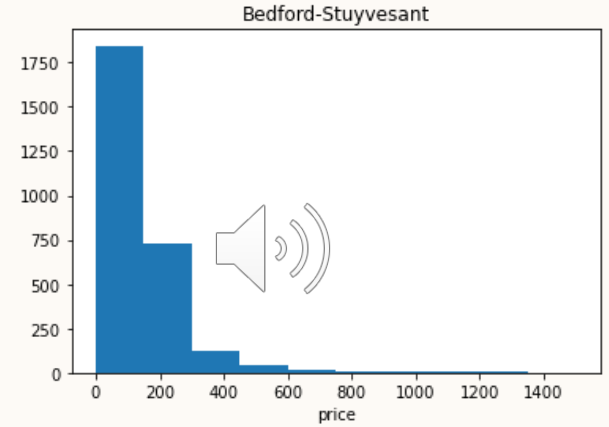
REGRESSION

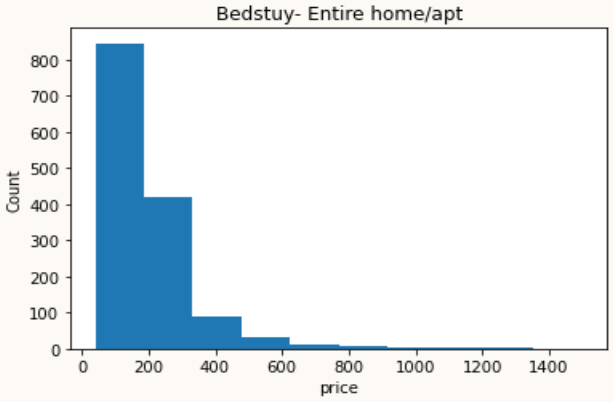
What price should be charged based on rental location/ characteristics?

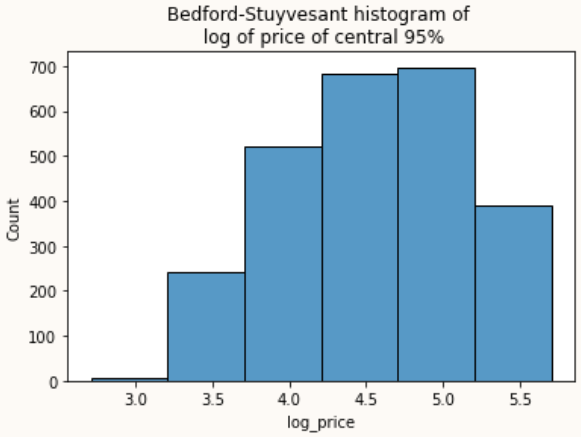
DATA SUMMARIZATION & VISUALIZATION

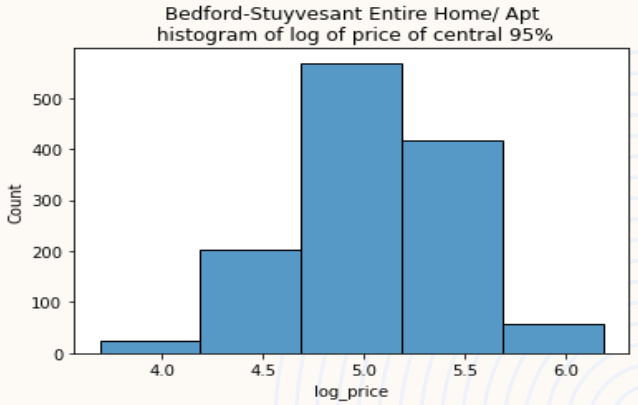
	pop_rank	neighbourhood	count	percent	price
12	0	Bedford-Stuyvesant	2779	6.968230	140.85
237	1	Williamsburg	2456	6.158321	183.21
108	2	Harlem	1878	4.709009	164.74
144	3	Midtown	1701	4.265189	381.70
30	4	Bushwick	1657	4.154861	114.35

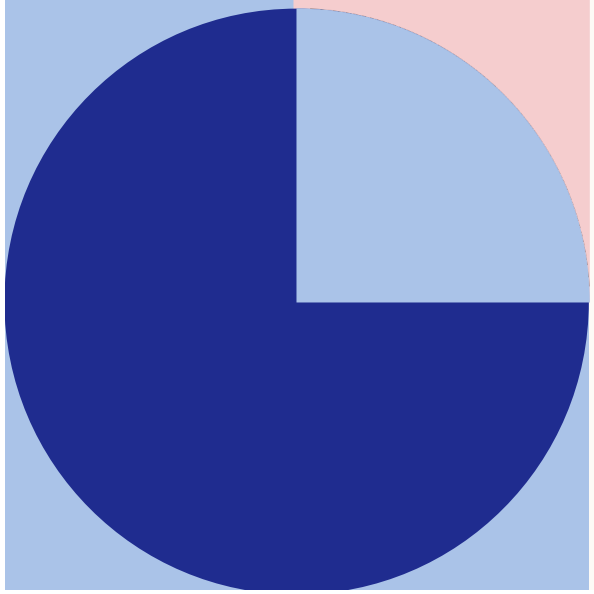












DATA MANIPULATION BY NEIGHBORHOOD

neighbourhood	price
Allerton	118.78
Arden Heights	13.36
Arrochar	132.03
Arverne	230.26
Astoria	109.01
...	...
Windsor Terrace	175.40
Woodhaven	94.53
Woodlawn	141.00
Woodrow	115.00
Woodside	84.72

We have reduced 39881 into 244 rows of manageable data!

```
1 df.stb.freq(['neighbourhood'])
```

	neighbourhood	count	percent	cumulative_count	cumulative_percent
0	Bedford-Stuyvesant	2779	6.968230	2779	6.968230
1	Williamsburg	2456	6.158321	5235	13.126551
2	Harlem	1878	4.709009	7113	17.835561
3	Midtown	1701	4.265189	8814	22.100750
4	Bushwick	1657	4.154861	10471	26.255610
...	...	...	...	...	...
239	Fort Wadsworth	1	0.002507	39877	99.989970
240	Ferry Point Park	1	0.002507	39878	99.992478
241	Country Club	1	0.002507	39879	99.994985
242	Bull's Head	1	0.002507	39880	99.997493
243	Bloomfield	1	0.002507	39881	100.000000

244 rows × 5 columns

CLEAN DATASET BY NEIGHBORHOOD

- Added columns for mean, median, mode and count
- Statistics was extracted from neighborhood (244 rows) groupby method for **feature extraction**.
- Extracted statistics were merged onto 39881 observations.
- Observations were reduced for respective neighborhoods that have more than 30 observations
  - ❖ Mode relevance ~ "boot-strapability"
  - ❖ https://machinelearningmastery.com/a-gentle-introduction-to-the-bootstrap-method/

```
1 dat_strap = dat[dat['count_hood']>30]
2 dat_strap
```

calculated_host_listings_count	availability_365	number_of_reviews_ltm	mean_price	mode_price	median_price	count_hood
30	217	4	198.15	150	150.5	708
9	356	0	151.76	35	85.0	117
6	365	0	106.47	50	78.0	315
1	83	0	136.97	65	99.5	92
1	219	4	134.1	80	100.0	188
...	...	...	...	...	...	...
7	365	0	198.28	260	178.0	79
1	0	5	127.24	90	100.0	147
1	152	18	231.74	150	157.0	1630
1	311	0	183.21	200	136.0	2456
1	0	4	198.28	260	178.0	79

DATA ANALYSIS & RESULTS

Logit Marginal Effects

```
=====================================
Dep. Variable:                      y
Method:                          dydx
At:                           overall
==============================================================================
                dy/dx    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
price         -0.0031   2.19e-05   -140.080      0.000      -0.003      -0.003
minimum_nights     0.0010   8.41e-05     11.377      0.000       0.001       0.001
number_of_reviews     0.0003   4.38e-05      6.338      0.000       0.000       0.000
availability_365     -0.0003   1.47e-05    -17.786      0.000      -0.000      -0.000
number_of_reviews_ltm    -0.0016      0.000     -8.665      0.000      -0.002      -0.001
==============================================================================
```

Logit Regression Results

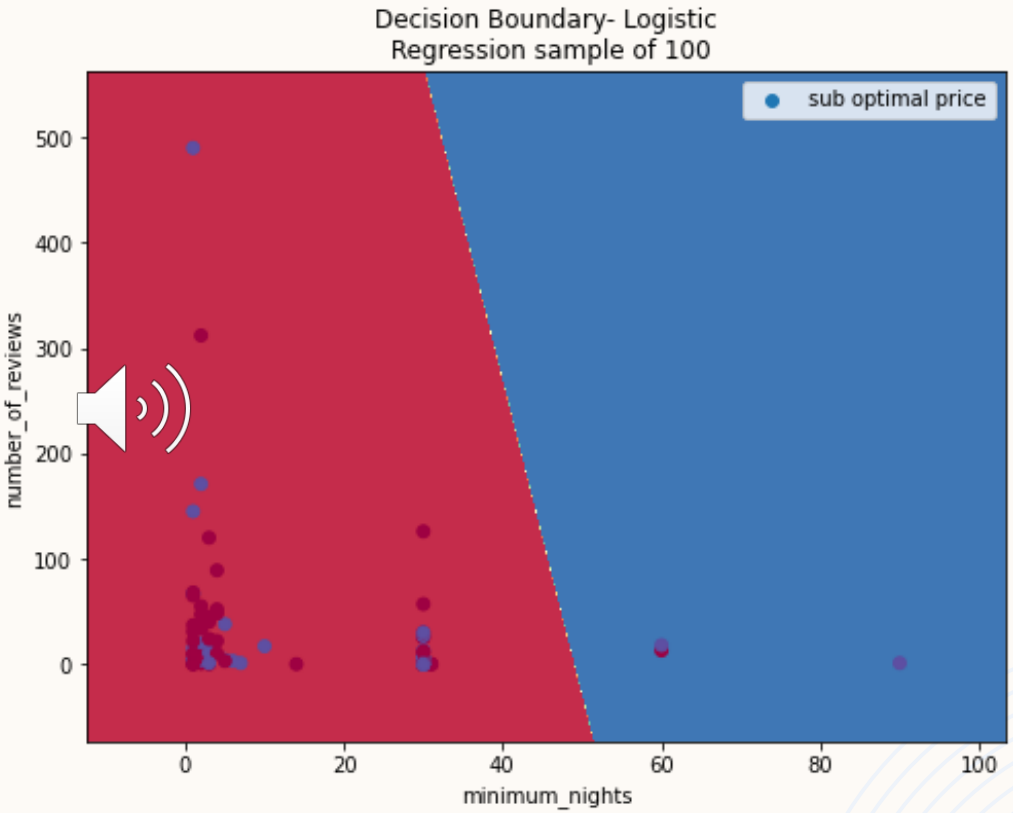
Dep. Variable:	y	**No. Observations:**	38545
Model:	Logit	**Df Residuals:**	38539
Method:	MLE	**Df Model:**	5
Date:	Sat, 03 Dec 2022	**Pseudo R-squ.:**	inf
Time:	15:06:09	**Log-Likelihood:**	-inf
converged:	True	**LL-Null:**	0.0000
Covariance Type:	nonrobust	**LLR p-value:**	1.000

	coef	std err	z	P>|z|	[0.025	0.975]
Const	2.3755	0.036	65.699	0.000	2.305	2.446
price	-0.0202	0.000	-81.518	0.000	-0.021	-0.020
minimum_nights	0.0063	0.001	11.306	0.000	0.005	0.007
number_of_reviews	0.0018	0.000	6.325	0.000	0.001	0.002
availability_365	-0.0017	9.82e-05	-17.491	0.000	-0.002	-0.002
number_of_reviews_ltm	-0.0103	0.001	-8.629	0.000	-0.013	-0.008

CONCLUSION

- Parsimonious-like model after dropping columns for id, name, and license no.
- Trained on
  - price
  - minimum_nights
  - number_of_reviews
  - availability_365
  - number_of_reviews_ltm

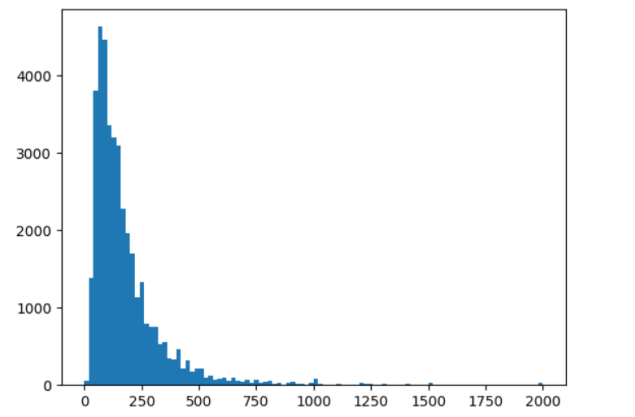
- Mostly flat **marginal effects**. Variables are of equal importance.
- p-values are of significance
- **Prices are sub-optimal for places with greater than 50 nights.**



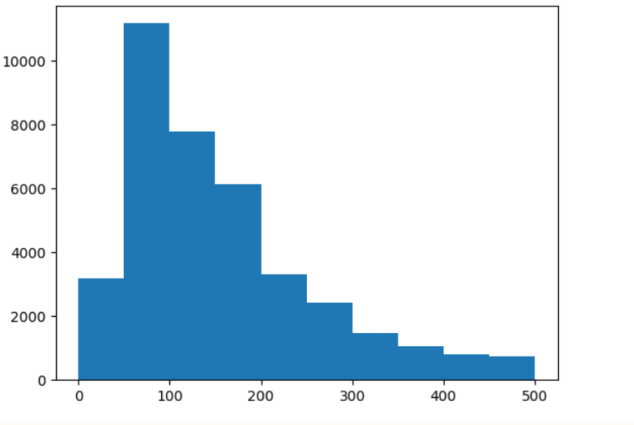
KNN

K-NN REGRESSION ANALYSIS

VALUE COUNT OF 'PRICE'>500 WAS VERY LESS



RESTRICTED THE DATASET WITH PRICE<=500

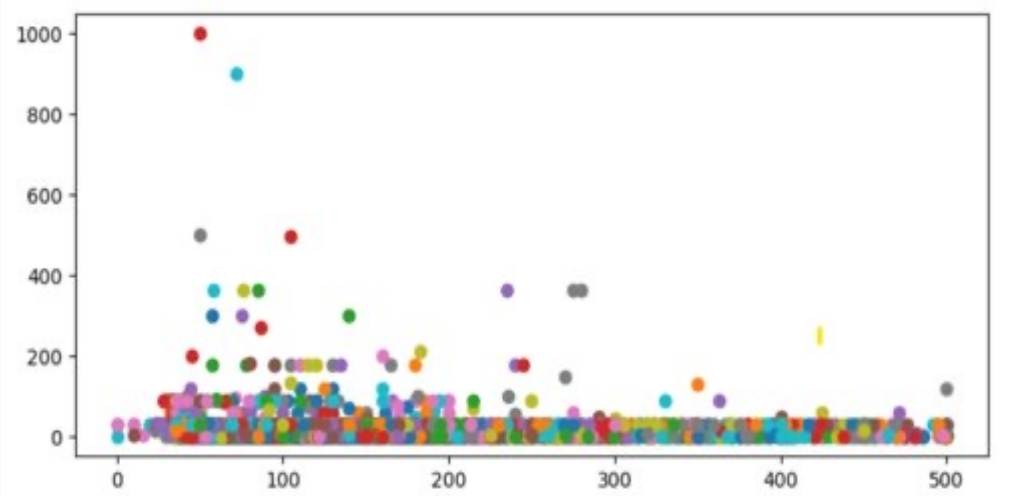


ANALYSIS

Tried to predict the price based on the room_type and neighbourhood group

- Root mean square error rate were very high.
- Tried to tune the performance using GridSearch, but there was no significant change.
- Scatter plot for price using test variables depict that price is linearly distributed.
- K-NN regression did not show any good results.

```
RMSE_train: 3.799804249904882
RMSE_test: 4.582044594624339
```

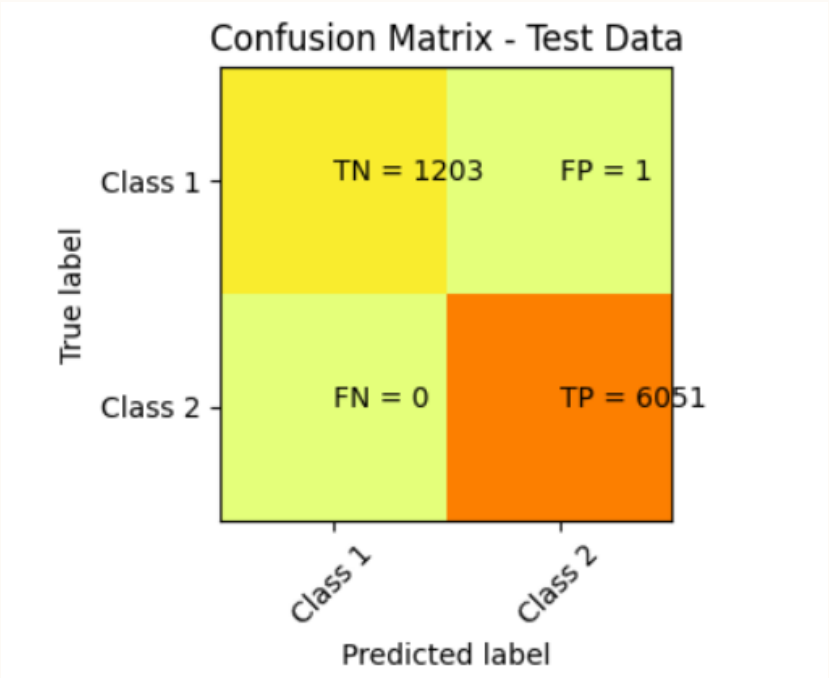


K-NN CLASSIFICATION

Tried to predict the price using classification based on neighborhood group.

- Price is classified in to two classes; price<250 & price>250 within the range of 500.

- Classification metrics report shows the results are 100% correct. That means there's no scope for prediction of price using current features.

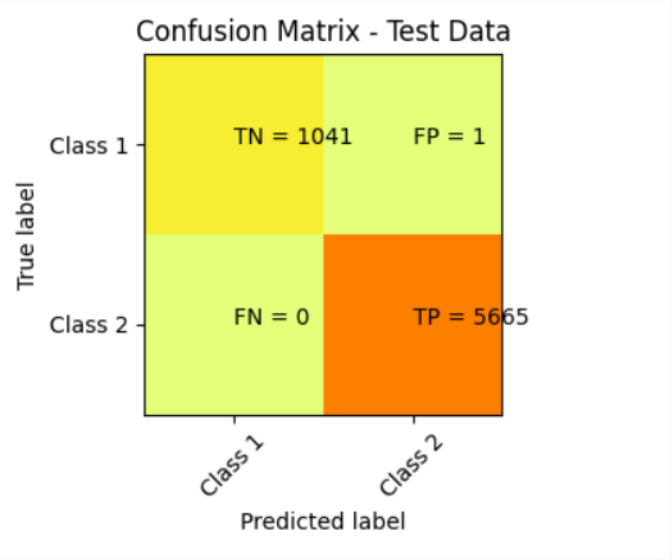


	precision	recall	f1-score	support
0.0	1.00	1.00	1.00	1204
1.0	1.00	1.00	1.00	6051
accuracy			1.00	7255
macro avg	1.00	1.00	1.00	7255
weighted avg	1.00	1.00	1.00	7255

COMPARING 2 MODELS

K=2 was the best performing parameter. Drawn Confusion matrix for this K value.

Compared the k=2 model with k=4. K=2 has higher accuracy & precision but same recall. That means recall at any k value greater than 1 is 100%.

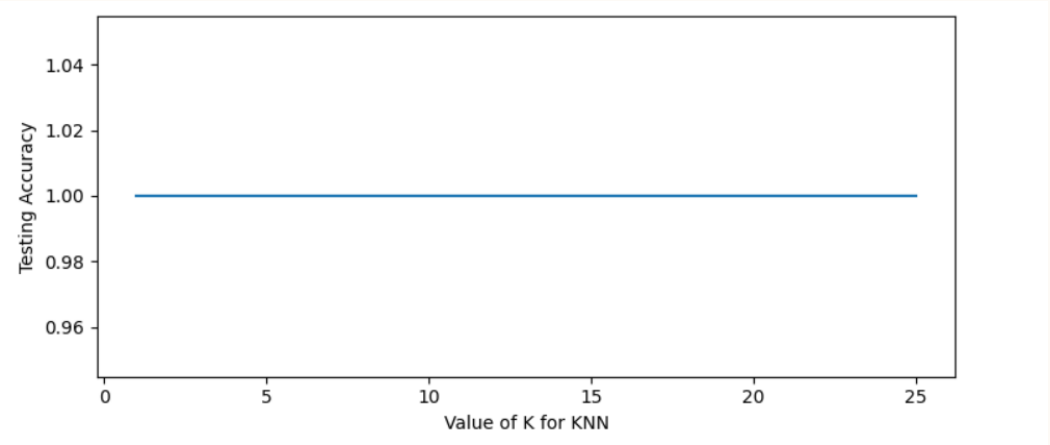


	Accuracy	Precision	Recall	F1 Score
k-NN (k=4)	0.998509	0.998238	1.0	0.999118
k-NN (k=2)	0.999851	0.999824	1.0	0.999912

TESTING ACCURACY

- For KNN models, complexity is determined by the value of K (lower value = more complex)
- Testing accuracy penalizes models that are too complex or not complex enough
- Training accuracy rises as model complexity increases.

Gathered Accuracy scores for k = 1 to 26. And Visualized the testing accuracy



CONCLUSION

We need more variables that can show more impact on the Price. The current dataset have very less features which can explain about the price. The only feature that can show significant effect in price is the location in this dataset. We cannot draw optimal price prediction using these variables. Considering this fact, we need to include variables like amenities, bedrooms, etc which are reasonable variables for the explaining the price difference between room types in real time also.

ZIPCODES

DATA PREPARATION FOR PRICE BASED ON ZIPCODES

- we've first started using ZipCodes to analyze the data apart using neighborhoods.

- For the analysis ZipCodes have been pulled up with the help of latitudes and longitudes and using a technique Geocoding. It is the process of transforming a description of a location (such as a physical address, or a name of a place) into a pair of latitude and longitude on the Earth's surface for that place.

DATA ANALYSIS USING ZIPCODES

- Included all the raw data points for analysis and found that there are extreme outliers due to divergent data that is available.
- Then we have considered mode prices instead of actual price.
- To maintain the skewdness in data log values of prices are considered and the graphs are plotted to get an overview of the analysis that we are about to do.

Since there are many ZipCodes we have considered the most populus ZipCodes for analysis

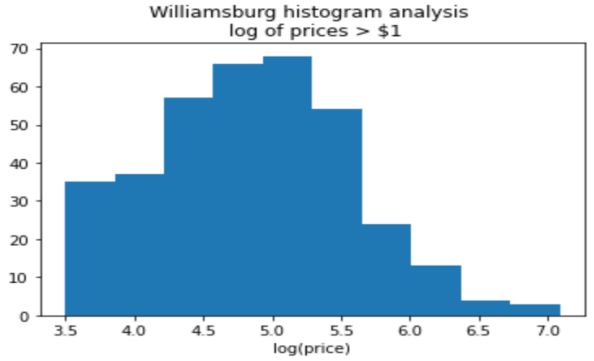
```
df2_s1 = df2.groupby('ZipCodes').get_group(11238)
df2_s2 = df2.groupby('ZipCodes').get_group(11206)
df2_s3 = df2.groupby('ZipCodes').get_group(10026)
```

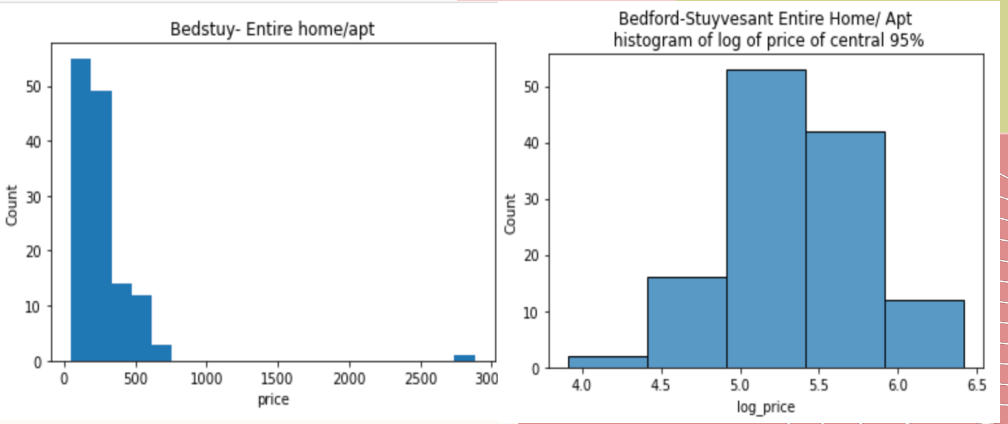
We have eliminated outliers and tried plotting graph.

After eliminating outliers, plotted graphs for different room types like home/apt.

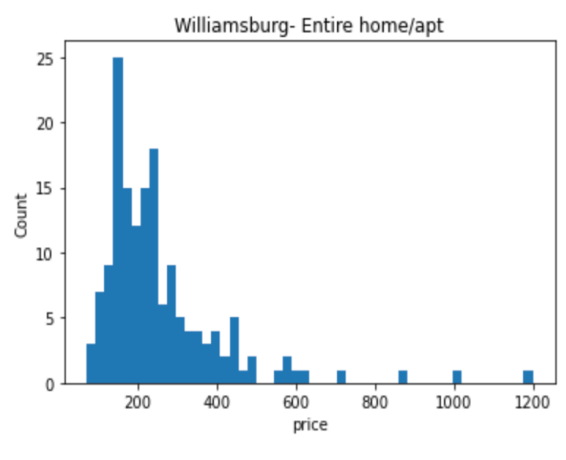
groupby('room_type').get_group('Entire home/apt')

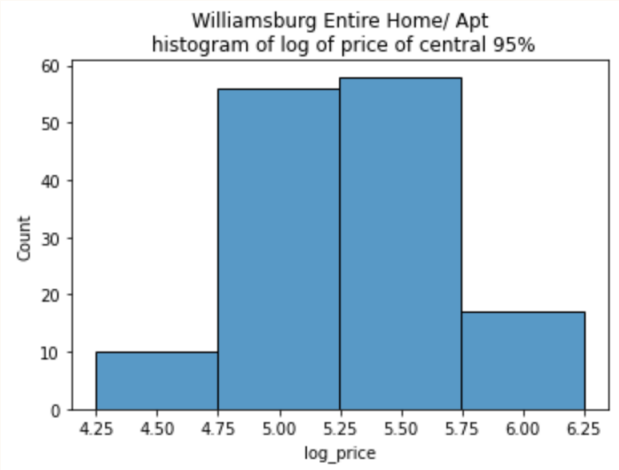
```
Log of outlier for uncorrected prices: (3.4965075614664802, 7.090076835776092)
Log of outlier for corrected prices: (3.5, 7.09)
```

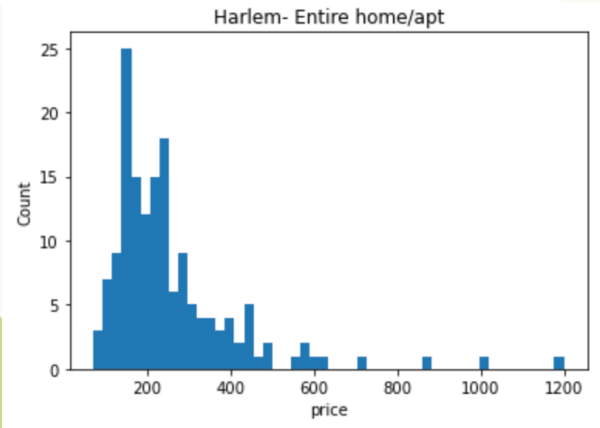


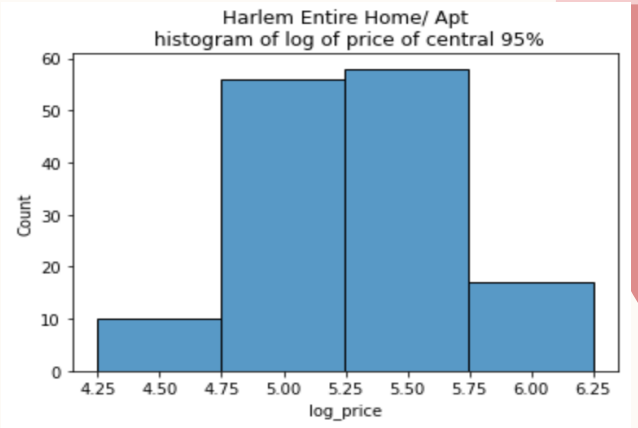


In the similar manner analyzed room types for locations Williamsburg and Harlem.



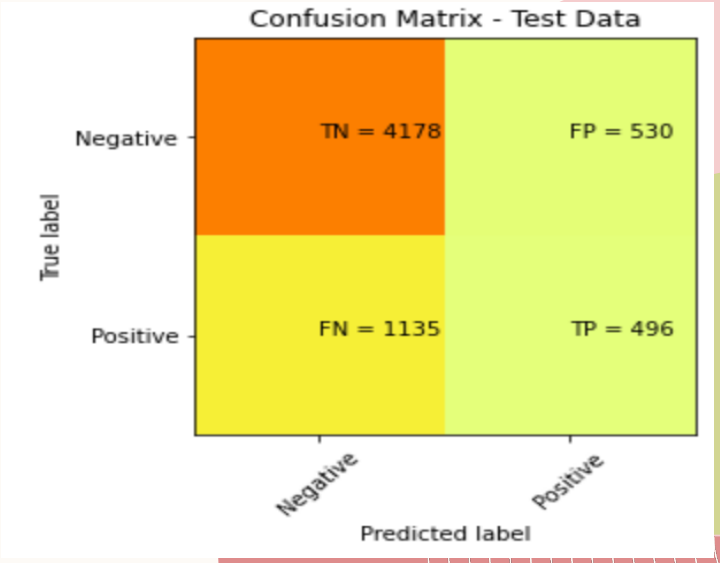






NAÏVE BAYES CLASSIFICATION

- Started using two files and creating a new dataframe adding the zipcodes and review column for analysis.
- In creating the column we have takan average review ratings that are >1 as "1" and the ratings that are <1 are considered "0".
- Then created confusion matrix accordingly and did calculated accuracy, precision, recall for analysis.



	precision	recall	f1-score	support
0	0.97	1.00	0.99	4708
1	1.00	0.92	0.96	1631
accuracy			0.98	6339
macro avg	0.99	0.96	0.97	6339
weighted avg	0.98	0.98	0.98	6339

CONCLUSION

The above classification demonstrates that customer reviews are more significant in determining price.

It is assumed that properties closer to Manhattan or other well-known attractions will, on average, have higher list prices because the majority of reviews mention the neighborhood, how accessible and close they are to popular destinations, as well as amenities and how big (spacious) they are.

CONCLUSIONS

Group-3

SUMMARY

- Prices are suboptimal with rentals greater than 50 nights
- Price prediction based on room type and location has 100% accuracy, which is not good. Considering other variables which have significant impact on price can improvise optimal price prediction.

THANK YOU

Follow us on Github

Adomako, @adomakor412

Dobbs, @idobbs-2012

Chigili, @naren1610

Sai, @SaiNikhilPalaparthi