IDS 575: MACHINE LEARNING STATISTICS

Airbnb Pricing Prediction

UIC UNIVERSITY OF ILLINOIS CHICAGO

GROUP 17

NAME	UIN
Giovanni Alvin Prasetya	678913250
Naga Bhaskar Muddana	665102758
Sagarika Ugendhar	672943645
Tushar Yadav	652509698

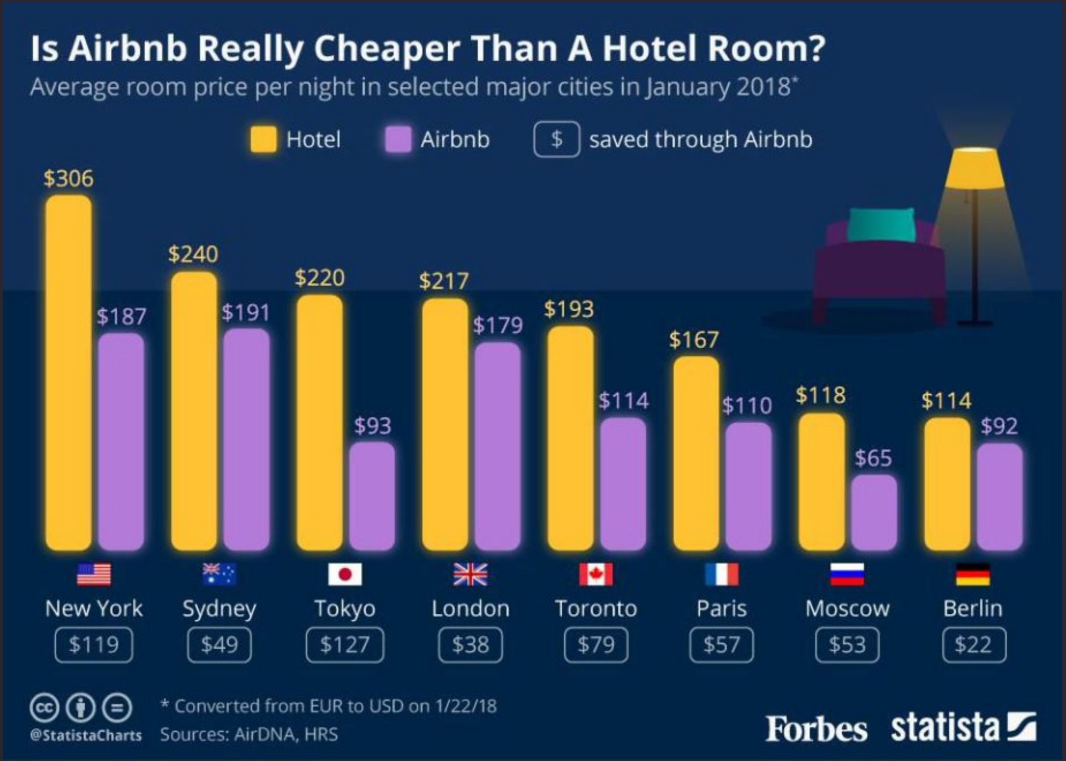
INTRODUCTION

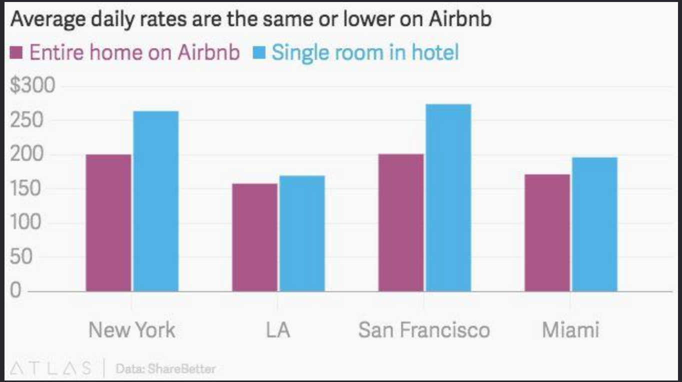
- Airbnb is a short-term rental platform that allows you to rent out a portion or all of your living space to others.
- Although Airbnb has been developing pricing tools for hosts since 2012, these tools have been relatively basic and have solely focused on simple parameters such as the number of rooms, surrounding properties, and amenities such as parking.
- Airbnb listings face competition from other Airbnbs rather than hotels.
- To strengthen our work background, we use another source of research related to this topic, which is predicting list prices on airbnb with Scikit-Learn. The purpose of this research is solely to generate competitive prices for a list of airbnb’s.
- Referring to this research, we would like to do some predictive analysis using regression algorithms.

PROJECT REFERENCE

- This research is based on previous research related to Optimization of Airbnb Dynamic Pricing
- Airbnb Dynamic Pricing solely based on two objectives, which was Business and Analytics Purposes
- This research analytical purposes was to create a model that was as flexible as possible by determining price at the scale of the smallest possible rental period at daily basis
- After creating the suitable model, the project focused on maximizing yearly profit for a listing on Airbnb
- The analytical model implemented for this project are linear regression, SVR and some more regression algorithms

AIRBNB STATISTICS

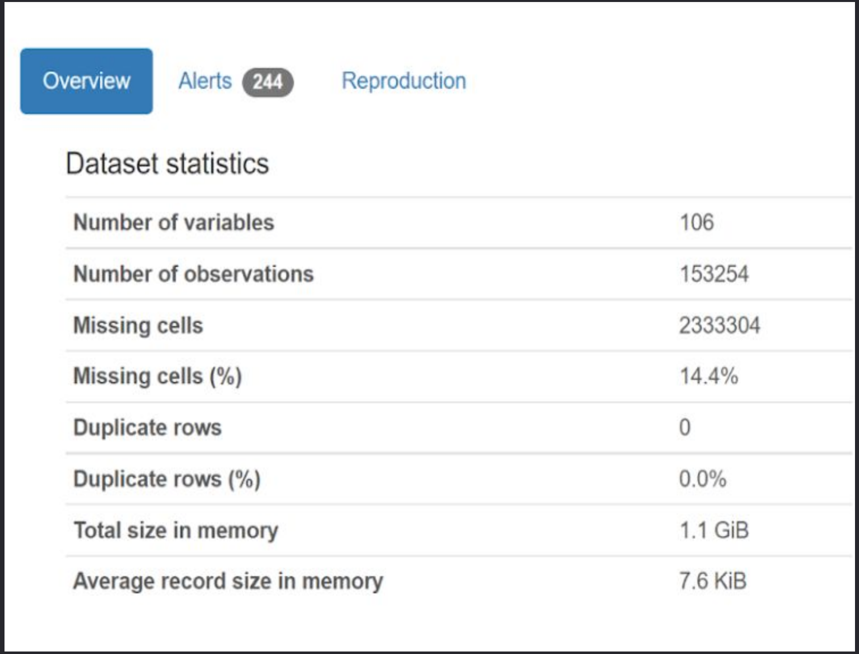




- Based on these real world chart we decided to analyze the price component for New York to see how prices are distributed across various neighborhoods and room types
- The interesting part of this is the gap average between hotel and airbnb in New York

DATA EXPLORATION & CLEANUP

Raw Data



Dataset statistics

Number of variables	106
Number of observations	153254
Missing cells	2333304
Missing cells (%)	14.4%
Duplicate rows	0
Duplicate rows (%)	0.0%
Total size in memory	1.1 GiB
Average record size in memory	7.6 KiB

Feature Engineering

Imputation -Handling missing values

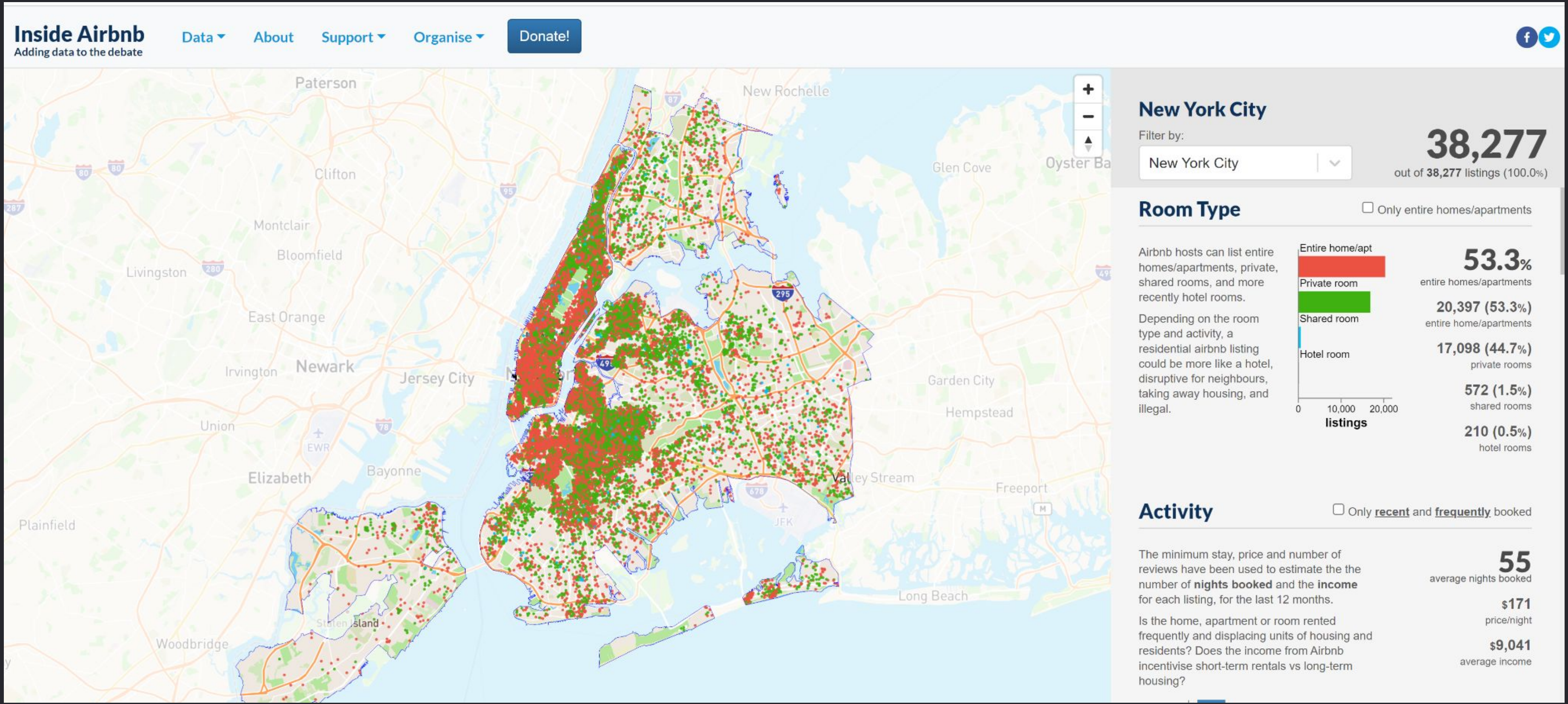
➔ Replacing ' ' or (Blanks) with NaN
➔ Drop columns that has only URL
➔ Drop columns that has 60% or more missing values
➔ Map columns having True/False to 1/0
➔ Removing special characters from dataset
➔ Converting string features having numerical values to Float

Handling Outliers - Removing extreme values

Categorical Encoding - Encode categorical features into numerical values using One hot encoding

Scaling - Standardizing the data using StandardScalar() from sklearn.preprocessing()

GEOGRAPHIC DISTRIBUTION OF NYC LISTINGS



New York City

Filter by: New York City

38,277
out of **38,277** 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.

53.3%
entire homes/apartments

20,397 (53.3%)
entire home/apartments

17,098 (44.7%)
private rooms

572 (1.5%)
shared rooms

210 (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.

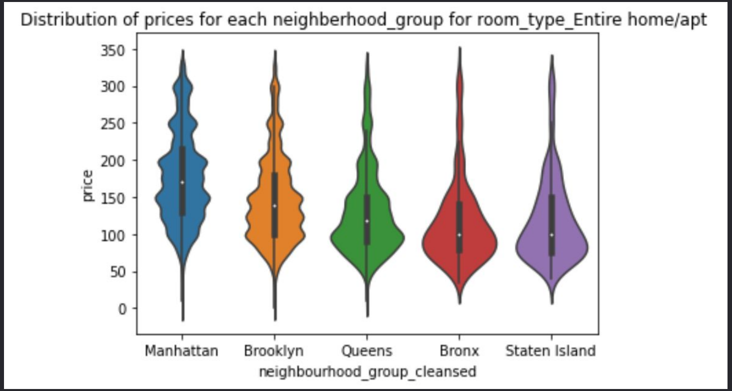
Is the home, apartment or room rented frequently and displacing units of housing and residents? Does the income from Airbnb incentivise short-term rentals vs long-term housing?

55
average nights booked

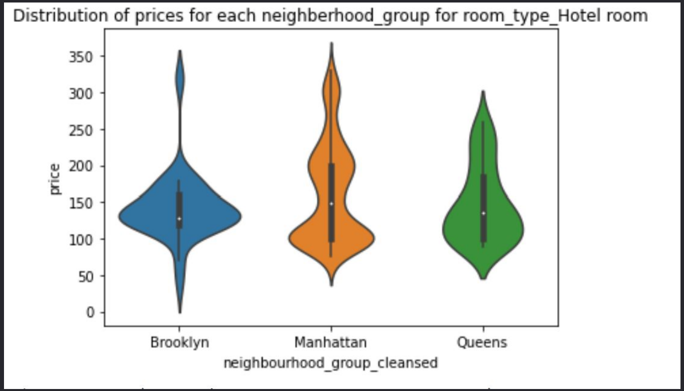
$171
price/night

$9,041
average income

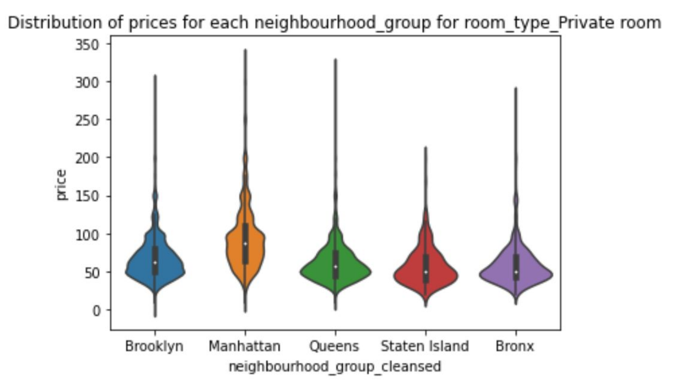
EXPLORATORY DATA ANALYSIS



Violin plot



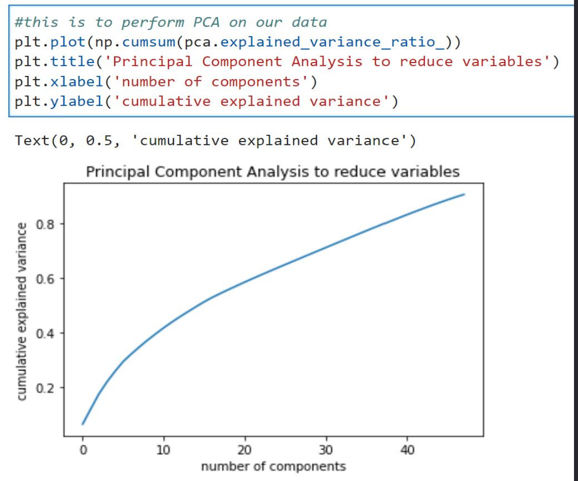
Graphs shows Price Distribution of each neighbourhood_group filtered by Room type



Room Types

❑ Entire Home/Apt
❑ Hotel_room
❑ Private_room

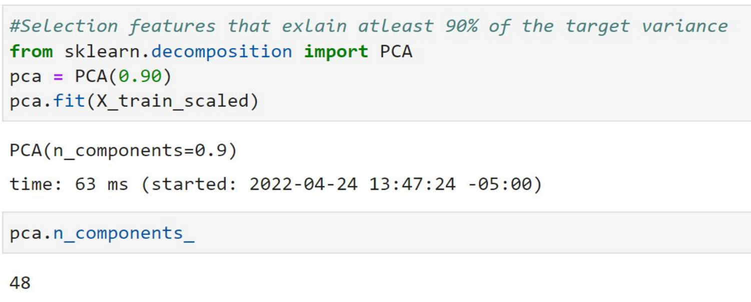
PRINCIPAL COMPONENT ANALYSIS (PCA)



```
#this is to perform PCA on our data
plt.plot(np.cumsum(pca.explained_variance_ratio_))
plt.title('Principal Component Analysis to reduce variables')
plt.xlabel('number of components')
plt.ylabel('cumulative explained variance')
```

```
Text(0, 0.5, 'cumulative explained variance')
```

Principal Component Analysis to reduce variables
cumulative explained variance
0.2
0.4
0.6
0.8
0
10
20
30
40
number of components



```
#Selection features that exlain atleast 90% of the target variance
from sklearn.decomposition import PCA
pca = PCA(0.90)
pca.fit(X_train_scaled)
```

```
PCA(n_components=0.9)
time: 63 ms (started: 2022-04-24 13:47:24 -05:00)
```

```
pca.n_components_
```

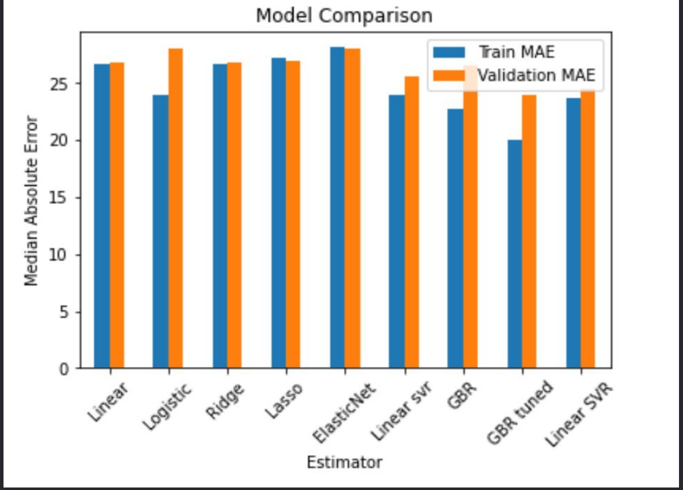
```
48
```

- ➢ Dimensionality reduction technique for reducing the number of features in a dataset
- ➢ Performed PCA with 0.9 for the number of components parameter
- ➢ Scikit-learn chooses the minimum number of principal components such that 90% of variance is retained
- ➢ Thereby reducing the features from 101 to 48
- ➢ Apply the mapping (transform) to both training and test dataset

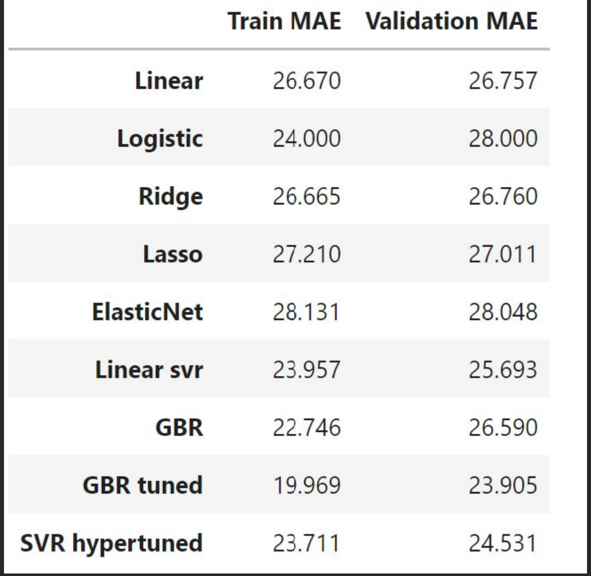
MAKING PREDICTIONS WITH SCIKIT-LEARN

Models used for comparison

- ❖ LinearRegression()
- ❖ LogisticRegression()
- ❖ Ridge()
- ❖ Lasso()
- ❖ ElasticNet()
- ❖ LinearSVR()
- ❖ GradientBoostingRegressor()
- ❖ Linear SVR



Train MAE and Validation MAE for model comparison



	Train MAE	Validation MAE
Linear	26.670	26.757
Logistic	24.000	28.000
Ridge	26.665	26.760
Lasso	27.210	27.011
ElasticNet	28.131	28.048
Linear svr	23.957	25.693
GBR	22.746	26.590
GBR tuned	19.969	23.905
SVR hypertuned	23.711	24.531

Evaluation Metric

- We chose Median Absolute Error(MAE) as Evaluation metric to evaluate model performance

- Median Absolute error is less sensitive to outliers than other metrics like Mean Squared Error(MSE)

- Looking at the graph we can say most of the models being able to predict the price with a median error around 20 to 30 dollars

Thank you

Motivation, explain the features,creativity, missing values to fill the knn