Southern University of Science and Technology

CS212 Intelligent Data Analysis

Director: Prof. Peter Tiňo

Investigating Properties of Houses from a New York Airbnb Dataset

Author:
Edward Yidong FANG

Student Number:
11510493

Email: 11510493@mail.sustc.edu.cn

Shenzhen, China July 7, 2017

Abstract

In these experiments, three basic methods of data analysis are implemented, which is Principle Components Anaylysis (PCA), Clustering and Self-organizing Map (SOM). By applying there methods, the relations among different attributes are discussed. After the dimension reduction, it was found that the attributes can be predicted by others to some extent, while some can not. The relations will be presented mostly as different firgures.

Contents

1 Introduction

Airbnb is an online marketplace and hospitality service, enabling people to lease or rent short-term lodging including vacation rentals, apartment rentals, homestays, hostel beds, or hotel rooms. The dataset analysed in there experiments is derived from Tom Slee's blog[2] and it is crawled from the website of Airbnb. And only a small part of data for New York, crawled on 05/06/2017, are analysed.

2 Data Preprocessing

2.1 Data Description

Here is the meaning for each column in the collected CSV file:

- room_id: A unique number identifying an Airbnb listing. The listing has a URL on the Airbnb web site of `http://airbnb.com/rooms/room_id`
- survey_id: A unique number identifying the behaviour of survey.
- host_id: A unique number identifying an Airbnb host. The host's page has a URL on the Airbnb web site of `http://airbnb.com/users/show/host_id`
- room_type: One of "Entire home/apt", "Private room", or "Shared room"
- country: the nation the room located in; acutually no data
- city: the city the room located in
- borough: A subregion of the city or search area for which the survey is carried out. The borough is taken from a shapefile of the city that is obtained independently of the Airbnb web site. For some cities, there is no borough information; for others the borough may be a number. If you have better shapefiles for a city of interest, please send them to me.
- neighborhood: As with borough: a subregion of the city or search area for which the survey is carried out. For cities that have both, a neighbourhood is smaller than a borough. For some cities there is no neighbourhood information.
- reviews: The number of reviews that a listing has received. Airbnb has said that 70% of visits end up with a review, so the number of reviews can be used to estimate the number of visits. Note that such an estimate will not be reliable for an individual listing (especially as reviews occasionally vanish from the site), but over a city as a whole it should be a useful metric of traffic.

- overall_satisfaction: The average rating (out of five) that the listing has received from those visitors who left a review.
- accommodates: The number of guests a listing can accommodate.
- bedrooms: The number of bedrooms a listing offers.
- bathrooms: The number of bathrooms a listing offers, actually not available.
- price: The price (in $US) for a night stay.
- minstay: The minimum stay for a visit, as posted by the host.
- name: The name of the room.
- property_type:"Apartment", "Loft", "Villa", "House", etc.
- latitude and longitude: The latitude and longitude of the listing as posted on the Airbnb site: this may be off by a few hundred metres.
- last_modified: the date and time that the values were read from the Airbnb web site.
- location: Unkown, certain number related to the location of the room.

The first line of the CSV file holds the column headings.

2.2 Data Preprocessing

Note that in this report only the data preprocessing process of the first problem is shown, but the other data preprocessing processes are all very similar.

2.2.1 Removing of the useless data and the Selection of the records

As the amount of data is really large (up to 40,730 rows), we just remove the data records without avaiable *reviews* or *overrall_satisfaction*. Since the prices higher than $200 per night are assumed as not reasonable, the records with those extremely high prices are removed. Also, as the columns *room_id, survey_id, host_id, country, city, last_modified* and *location* are meaningless, we removed them from the data. In addition, it is found that the values of column *bathrooms* and *minstay* are unavailable.

Finally, for the simplification of the problem, we just retrieve the records whose boroughs are "Manhattan" or "Brooklyn". And the experimental data are sampled from the original data under the sample fraction of 0.1.

	room_type	borough	accommodates	reviews	overall_satisfaction	bedrooms	price	longitude	latitude	property_type
0	Entire home/apt	Brooklyn	2	95	5.0	0.0	125.0	-73.943276	40.721256	Apartment
1	Entire home/apt	Brooklyn	3	3	4.5	1.0	165.0	-73.952168	40.723975	House
2	Entire home/apt	Manhattan	5	25	4.0	2.0	220.0	-73.962862	40.758275	Apartment
3	Entire home/apt	Manhattan	2	2	0.0	1.0	180.0	-74.003905	40.733196	Apartment
4	Entire home/apt	Brooklyn	4	3	5.0	1.0	132.0	-73.957279	40.733538	Apartment

Table 1: Data After preprocessing

The data at the end looks like the Table 1.

2.2.2 One Hot Encoding

To make use of the data field *room_type*, *borough* and *property_type*, both of the infomation are encoded using the method called One-Hot encoding to transform the data. By doing so, it is ensured that the infomation of all possible aspects of the room are included into the data matrix.
For example, if there two roon, whose types are "Private room" and "Entire home/apt", respectively, then the encoded result looks like Table 2. Three different columns are added to enumerate all possible room types. And the values are all zeros except for the room type the room belongs to.

	accommodates	...	latitude	room_type_Entire home/apt	room_type_Private room	room_type_Shared room	...
0	2	...	40.691398	0	1	0	...
1	4	...	40.811016	1	0	0	...

Table 2: Example Data After One-hot Encoding

2.2.3 Data Standardization

To see the significance of data standardization, two box plot was made, one for the raw data and another for the standardized data. Note that the one-hot encoded columns are standardized but are not included in the figures because the scal of the box plot will be influenced by those attributes and make the differece hard to distinguish.

As is obvious from the Figure 1, the values of attribute *reviews* are not in the same scale with other attributes. Neither does the longtitude and latitude. If we use these values in the matrix we will calculate later, the importance of attribute *reviews* ,*latitude* and *longtitude* will be extremly larger than ohter attributes. Thus, we need to apply the standardization to the raw data.

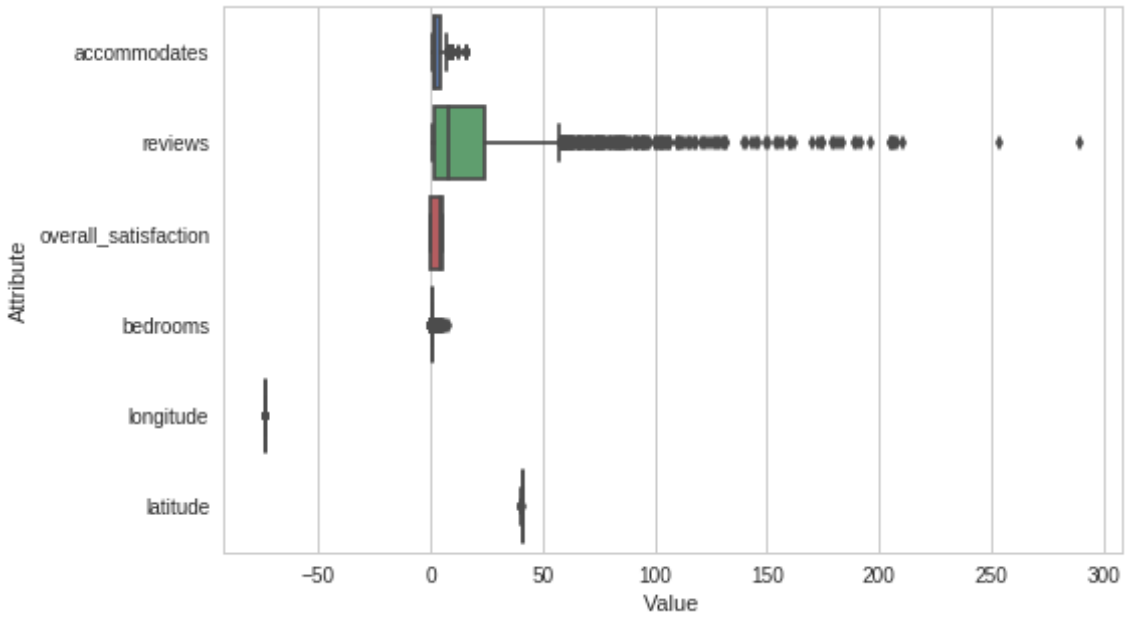


Figure 1: Data before standardization

Standardization can be done in many ways, the formula

$$NewValue = (OldValue - Mean)/StandardDeviation$$

is enough for this dataset. The result is shown in the Figure 2.

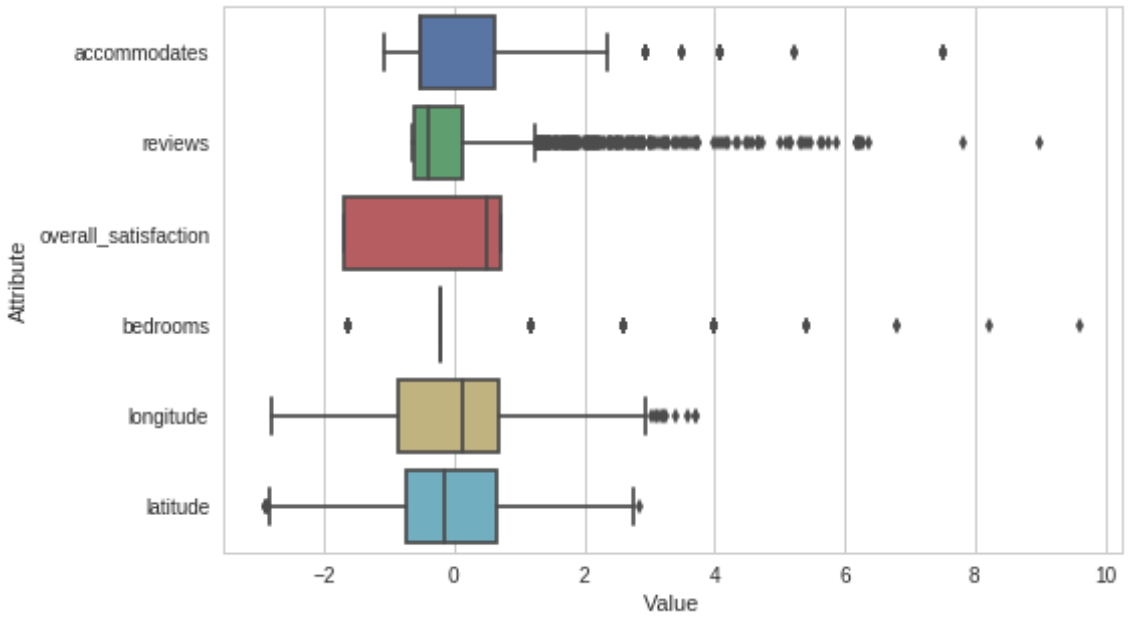


Figure 2: Data after standardization

3 Questions to Ask and Labeling

3.1 Questions to Ask

When people go out and want to find a room to stay in a city, it can be said that the most concerning aspects of a room is its condition, position and price. In the dataset analysed, the price is already there and the overall satisfaction can represent a combination of all the aspects. Thus, two questions are going to be explored in these expriment.

- Can the ranges of price be roughly predicted using other attributes of the data?
- Can the overall stasfications be roughly predicted using other attributes of data?
- What is the most crucial factors that influence the price and overall satisfaction among the attributes we have?
- How much the difference will be for the data in two different borough?

3.2 Labeling Schemes

3.2.1 Label of Prices

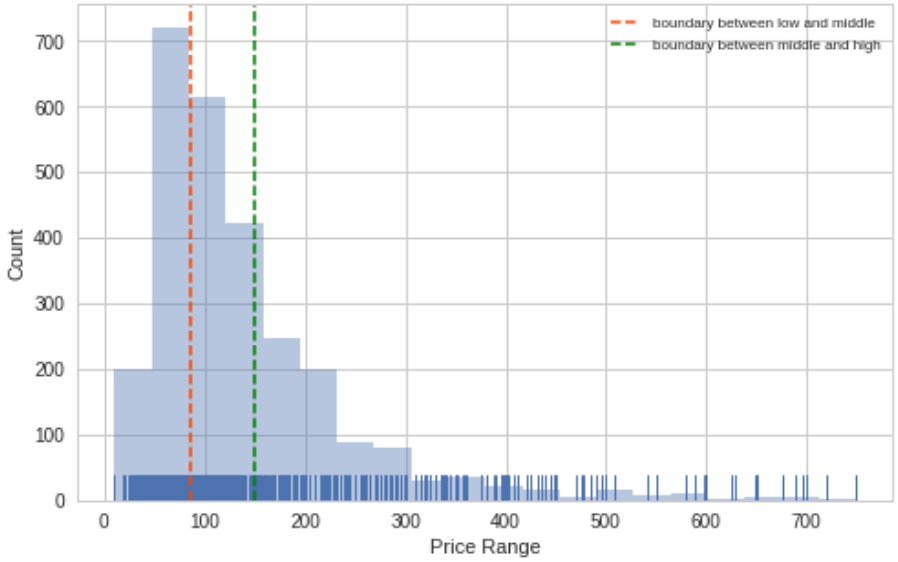


Figure 3: Distribution of prices that is lower than $750

The price level can be classify into three level according to the distribution of the prices. Fig.3 shows the distribution and according to it, the data are labelled into three approximately-equal-size classes: "low price", "middle price" and "high price".

Note that for the convenience of visualization, we just plot the distribution of prices that is lower than $750, and only a few records' prices are higher than $750.

After some trying, it is finnally found that these three intervals are best for the data to be evenly devided and labelled: $[0, 85)$, $[85, 149)$ and $[149, \infty)$.

3.2.2 Label of Overall Satisfaction

The label of overall satisfication is set to be "low" if *overall_satisfaction* is equal to or less than 4, "median" if is equal to 4.5, "high" if equal to 5.

4 Data Analysis

After the preprocessing of the data, the Principle Component Anaylysis (PCA) algorithm was used to generate the eigenvalues, eigenvectors, and coefficients to the eigenvalues, which actually are the variances that contianed in each directions that eigenvectors expend.

4.1 Coordinate Projections

First some simple projections of just two attribues of the data are used.
As most of our attributes are discrete, some of the important can be distinguished as a form of points' labels.
From most of these figures such as Fig.4a and Fig.4e it can be found that the average room price in Manhattan is generally higher than that in Brookly. This is perhaps due to the fact that Manhattan is the center of the city. Also, Fig.4a indicates that the larger the accomodates number is, the higher the price is. The trend of price over number of bedrooms in Fig.4e is similar, which is intutively true.

However, Fig. **??**, **??**, 4f, 4g and 4h also show some relation between price and location, overall satisfication, room type and reviews number. So which of there attributes can influence the price mostly, i.e. the importance of these attributes. Then it can be figured out by the technique of PCA.

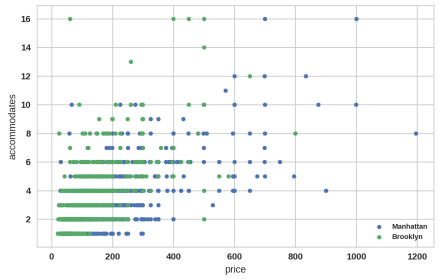
4.2 Principal Component Analysis

4.2.1 Prices of the rooms

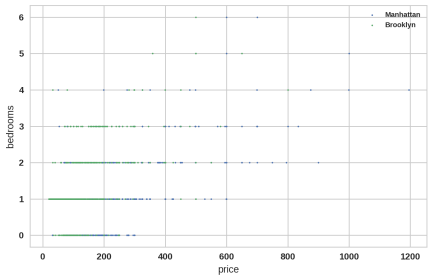
It can be noticed from Fig.7 that the distribution of variance of the principle components is not concentrated to the first serval principal components very well. Thus, if the data is projected on to the first two or three principle components, it can be predicted that the structure of results will not very similar to that of the original data.

However, the PCA technique indeed reduce some dimensions, as it can be concluded from the Fig.7 that the first 15 eigenvectors can preserve the 90% of the variance of the dataset.

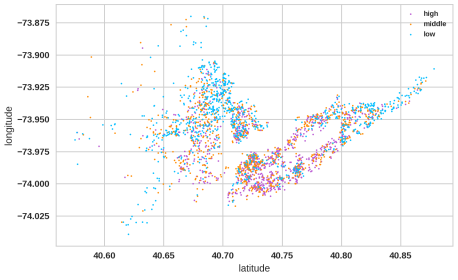
Figure 6a projects the 23 dimensional data onto a sub-space spanned by the first and second eigenvector of the covariance matrix of X



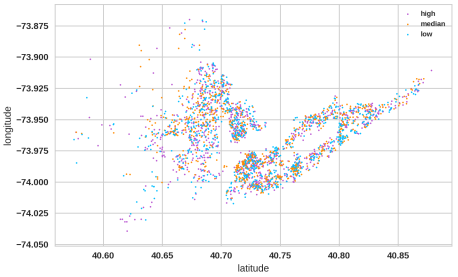
(a) prices and accomdates with borough



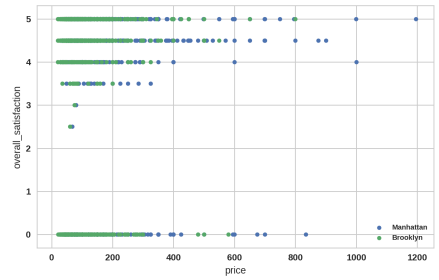
(b) prices and bedrooms with borough



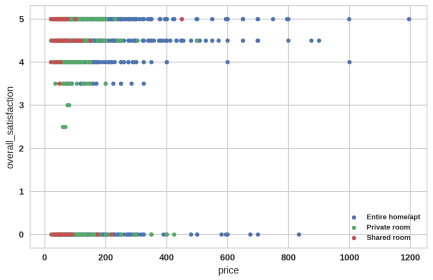
(c) latitude and longitude with price



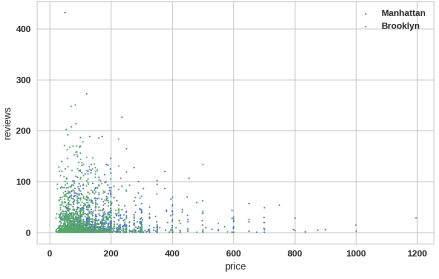
(d) latitude and longitude with satisfiction



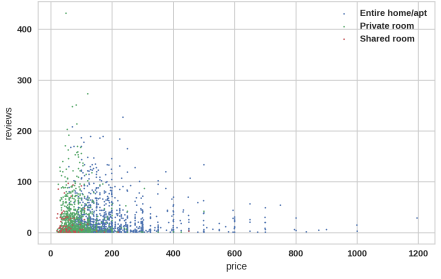
(e) latitude and longitude with price



(f) satisfiction and price with room type



(g) number of reviews and price with borough



(h) number of reviews and price with room type

Figure 4: Co-ordinate projections

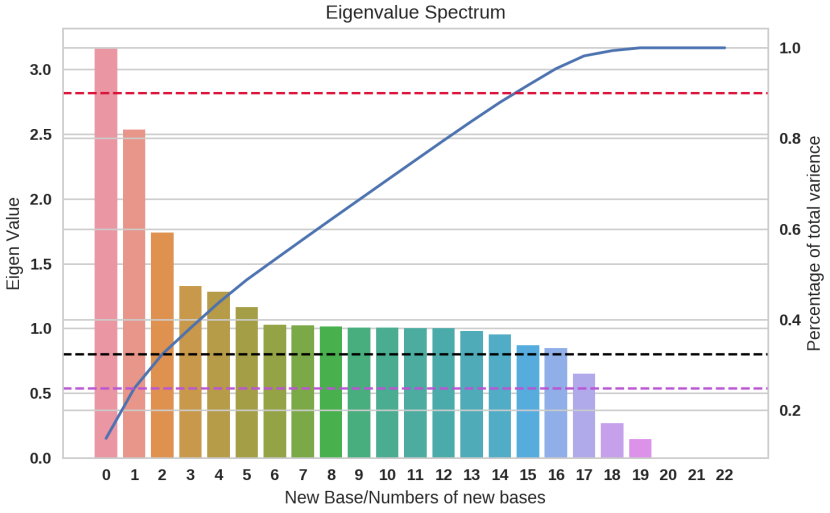
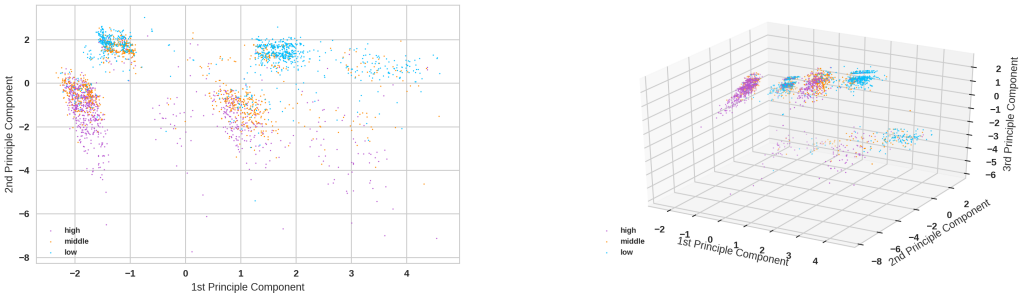


Figure 5: A scree plot showing the percentage variance of each principle component, and a running cumulative total

The more detialed infomation of the first three eigenvectors are shown in the Table 3. In the first eigenvector the importance of borough is very important, this is why it can be discovered that in the following PCA projection figues there are 2 similar part of cluster (such as in Fig. 6a and 6b). Also, the latitude and longtitude are actully related to the borough it located in, thus they increase the distinction or distance in the projection.

The price differences are mainly interpreted by the second principal eigenvector. The most siginificant components in the 2nd vector is the room type. Since "Entire" type and "Private" type are negtive and positive respectively, we can conclude from the PCA projection result that the prices of Entire home/apt rooms are higher much higher than those of Private room. Also, the more accommodates and bedrooms, the higher price the room will have. The importance of accommodate numbers is greater than that of room numbers. And the overall stasifcation does influence a little on the price, which reveal the "effect of market".

For the 3rd most siginificant eigenVector, the point is that the room with larger value has lower price. So this vector illustrate the property type influence on the price. The price of property type from lower to higher is apartment, house, loft and townhouse.



(a) Two dimensions (b) Three dimensions

Figure 6: A projection of the PCA results

1st Principle Component Eignenvalue 3.15901473		2nd Principle Component Eigenvalue 2.53613192		3rd Principle Component Eigenvalue 1.73809388	
borough_Manhattan	-0.51484551	room_type_Entire home/apt	-0.52678505	property_type_Apartment	0.60738014
borough_Brooklyn	0.51484551	room_type_Private room	0.51272818	property_type_House	-0.36991721
latitude	-0.41706656	accommodates	-0.50934697	property_type_Loft	-0.28433712
property_type_Apartment	-0.29132238	bedrooms	-0.34241624	borough_Manhattan	-0.26485932
longitude	0.26209607	overall_satisfaction	-0.14247044	borough_Brooklyn	0.26485932
property_type_House	0.25374459	latitude	0.11644046	latitude	-0.22556836
room_type_Private room	0.15672518	reviews	-0.0980652	property_type_Townhouse	-0.20936499
room_type_Entire home/ap	-0.15512859	property_type_Apartment	0.09209157	room_type_Entirehome/apt	0.1495919
property_type_Loft	0.10969607	longitude	0.08412272	property_type_Other	-0.14047292

Table 3: Details of top three eigenvectors for price label

1st Principle Component Eignenvalue 3.35927066		2nd Principle Component Eigenvalue 2.87945626		3rd Principle Component Eigenvalue 1.74548804	
borough_Manhattan	-0.42064644	accommodates	-0.42181393	property_type_Apartment	-0.60670587
borough_Brooklyn	0.42064644	bedrooms	-0.34494455	property_type_House	0.36586291
price	-0.35066038	room_type_Entire home/apt	-0.32707405	property_type_Loft	0.28124627
room_type_Entire home/apt	-0.33406302	room_type_Private room	0.31176258	borough_Manhattan	0.25276884
room_type_Private room	0.32657947	latitude	0.3067403	borough_Brooklyn	-0.25276884
latitude	-0.30536557	price	-0.3003249	latitude	0.22054766
longitude	0.27052436	borough_Manhattan	0.29921848	property_type_Townhouse	0.21576297
accommodates	-0.23111635	borough_Brooklyn	-0.29921848	room_type_Entirehome/apt	-0.20696862
property_type_Apartment	-0.18760868	property_type_Apartment	0.24644235	room_type_Private room	0.17363995

Table 4: Details of top three eigenvectors for satisfication label

4.2.2 Overall satisfaction

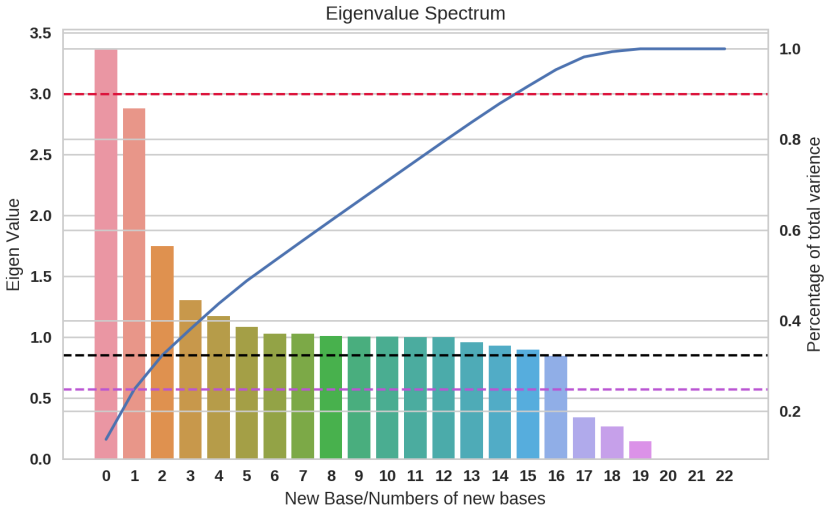
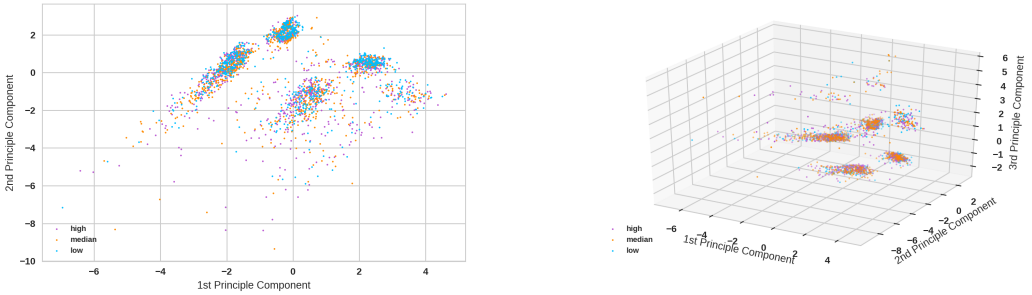


Figure 7: A scree plot showing the percentage variance of each principle component, and a running cumulative total

Then the same technique is applied to the infomation without overall satisfection data. Although the satisfication data are dropped and the price data are added, the most parts of PCA results almost remains the same. This means that the overall satisfications are not strongly related to the prices of the rooms.
From the Fig. 7, 8a and 8b it can be found that the 2nd and 3rd eigenvector has the largest influence on the overall satisfication data.

It may be found that with low 2nd vector component but high 3rd vector component, the overall satisfication tends to be higher and with high 2nd vector component and



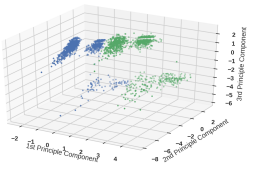
(a) Two dimensions (b) Three dimensions

Figure 8: A projection of the PCA results

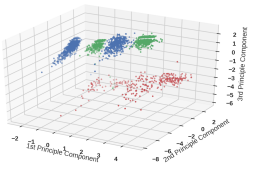
high 3rd vector component, the stisfication tends to be lower. It can be known with lower 2nd vector, the price can be higher. Thus it can be Speculated that the overrall satisfication may related to the combination effects of price and property type, this can be something not reflected in the data such as "Environmental health status", "Convenience of transportation", etc..

4.3 Clustering

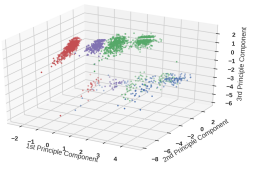
After doing the basic overview of the data by PCA, the technique of clustering is used. Here the *KMeans* tools in python package *sklearn* is applied.
The Kmeans clustering will first be applied to the dataset and visiualize it in a 3D PCA projcetoin figure. Then the results will be compared with that of KMeans directly applied to the PCA projected data.
The Figure 9 shows the results of KMeans on the whole dataset. And Figure 10 shows the resultes of KMeans on the PCA-projected data. Also, the quantization errors has been calculated and are shown in Figure 11.
From the results, we know that after the 3D-PCA projection, some infomation loss lead to the increase of number clusters where the elbow point appears. And the KMeans on the high dimensions doesn't successfully distingish the two clusters in the 3D projection which should be obvious, this may be caused by the adjacency of these points in the high dimension compared to other points.



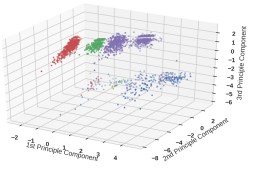
(a) Cluser number 2



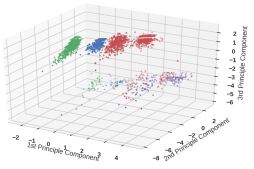
(b) Cluser number 3



(c) Cluser number 4

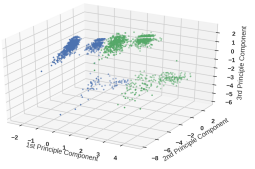


(d) Cluser number 5

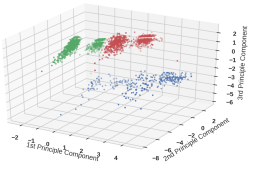


(e) Cluser number 6

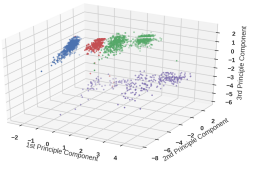
Figure 9: KMeans results on the whole dataset



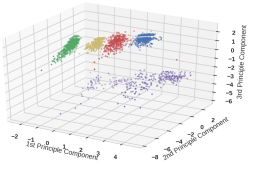
(a) Cluser number 2



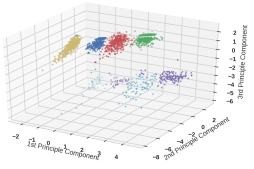
(b) Cluser number 3



(c) Cluser number 4

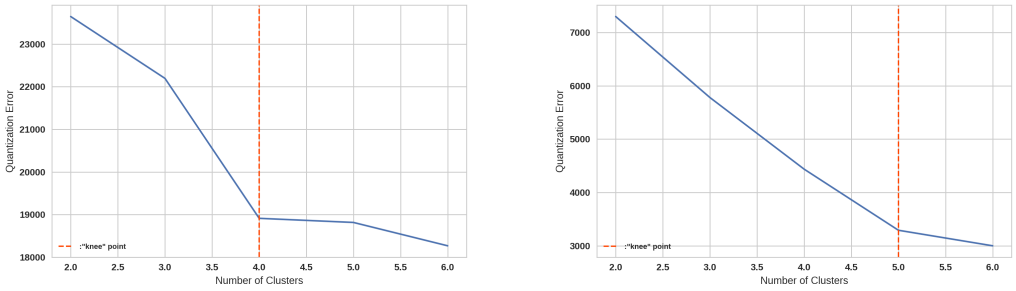


(d) Cluser number 5



(e) Cluser number 6

Figure 10: KMeans results on the PCA-projected data



(a) the whole dataset (b) the PCA-projected 3D data

Figure 11: Quantization error of KMeans results

4.4 Self-orgnaization Map

The SOM (Self-organizationing Map anyalysis is done with the help of python package *minisom*[1], this packages first normolize the data and can then use the self-organization algorithem to iterate and get the weight of each component of the codebook vectors. The package supports the 2-D SOM algorithm, so it can be used to plot the high dimensional data in a two dimensinal plannar. Each cell in the figures is the normalised sum of the distances between a codebook vector and its neighbours.

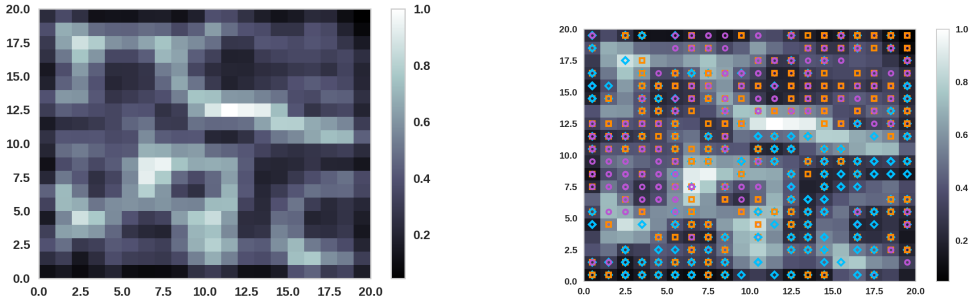
4.4.1 SOM on the whole dataset

It can be seen from the Fig.12 that there are approximately four cluster, which confirms the results in the last clustering on the whole dataset. However, althoush some clusters have been shown in the Fig.13, there is no obvious rules of the distribution of the overall stasification lablel.

4.4.2 SOM on the 3D-PCA projection

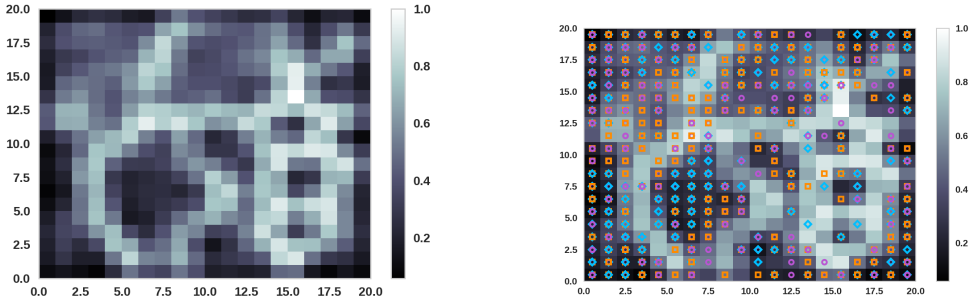
Compared the SOM result in Fig.14 with the result of SOM on the whole dataset, it can be concluded that the structure of the data become more simple after the projection on to a low dimension, but it can be seen that the number of the cluster is almost the same, so most of the infomation is not lost.

Then, the weights of the codebook vectors are also projected to the 2D plannar using



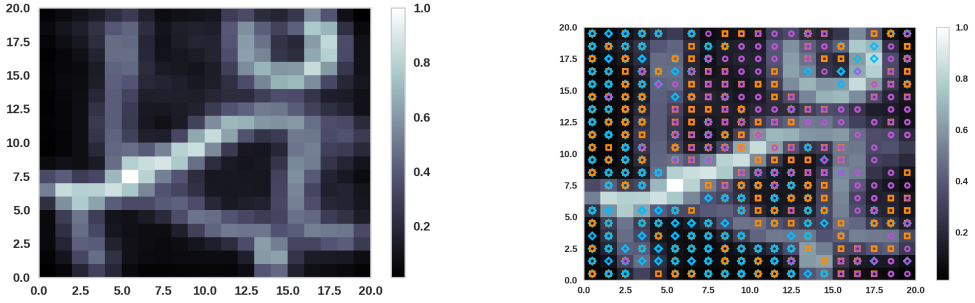
(a) Distance Map of the 2D SOM (b) Labelled Distance Map of the 2D SOM

Figure 12: SOM results on the whole dataset with price label



(a) Distance Map of the 2D SOM (b) Labelled Distance Map of the 2D SOM

Figure 13: SOM results on the whole dataset with overall satisfication label



(a) Distance Map of the 2D SOM (b) Labelled Distance Map of the 2D SOM

Figure 14: SOM results on the PCA-projected data with price label

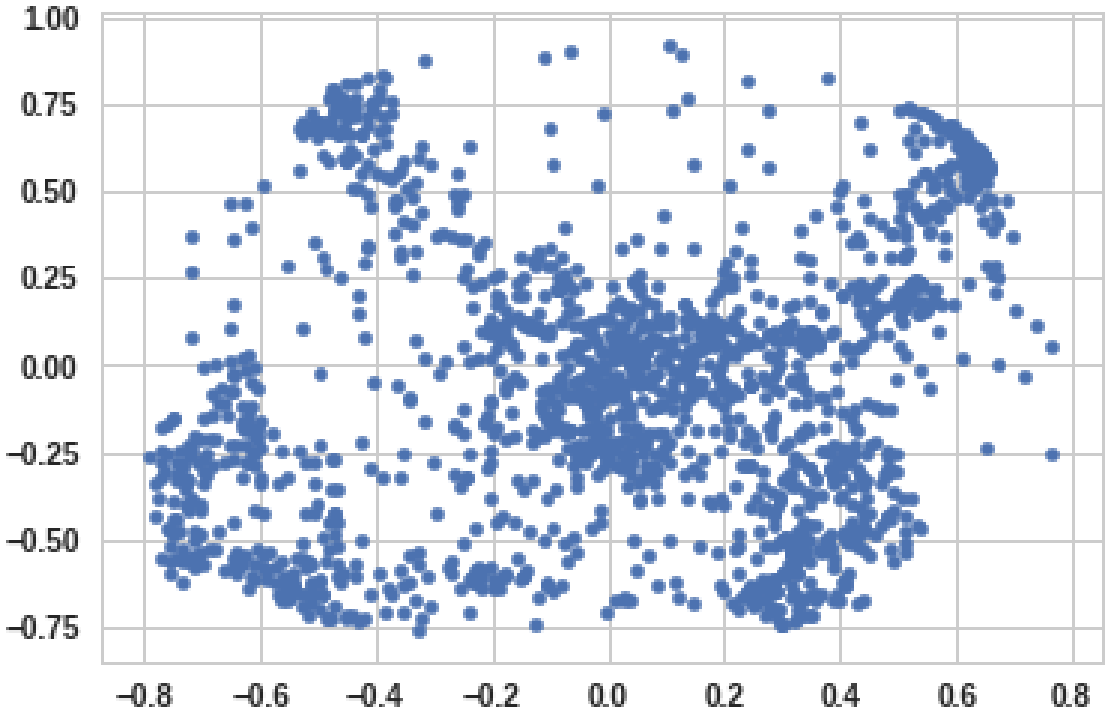
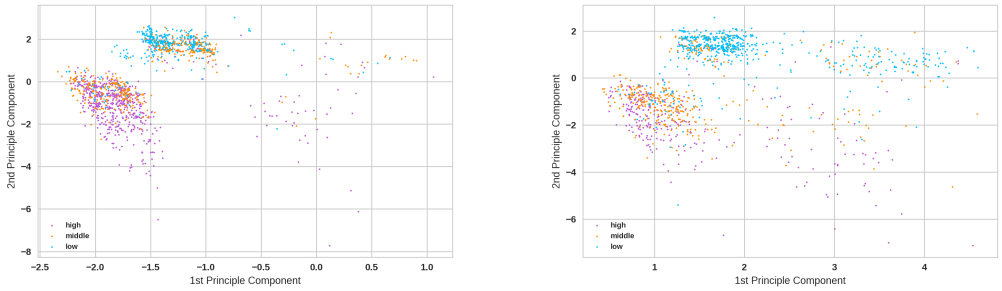


Figure 15: Nomalized codebook vector of the SOM on the whole dataset projected on to the first two eigenvectors

the eigenvectors of the previous principle component analysis. The result is shown in the Fig. 15. The structure of this figure is very similar to that of Fig 6a, which is derived by PCA. Both of these two techniques can visulize the high-dimension data in the low dimension space and make the following anaylysis more easy.

4.5 Differences between two boroughs

Although the distinction of two borough in the preivious analysis is very obvious, it can be depicted more precisely if they are shown in the two differe figures.
This two Figure (16a and 16b) have almost the same structure but different scale over 1st principle component, this probabably is casued by different locations.



(a) Manhattan (b) Brooklyn

Figure 16: PCA projection on first two pinciple components

5 Conclusion

With this airbnb dataset, what can be achieved is limited, beacuse the lack of some important attributes such as the health condition and the convenience of public tarnsportation. The preprocessing of the dataset can be really important for the data crawled directly from the Internet, the presedures include dealing with missing values, standardization and some special transformation of some attributes such as one-hot encoding.

The co-ordinate projections can really give some infomation, but they are limited, unable to describe the overview of the whole dataset. With the technique of PCA and SOM, the data can be visualized in a low dimension spcae as a whole. Besides, the SOM results can illustrate the structure of the data better than PCA, and PCA has better ability in determining the importances of different attributes.

The different results of clustering over the whole dataset and the projected data is very likely cased by the large infomation loss during the projection because only about 30% of the toltal variance remains.

References

[1] justglowing, 2017. URL https://github.com/JustGlowing/minisom.

[2] T. Slee. Airbnb data collection: Get the data, 2017. URL http://tomslee.net/airbnb-data-collection-get-the-data.