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 6, 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[1] 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.

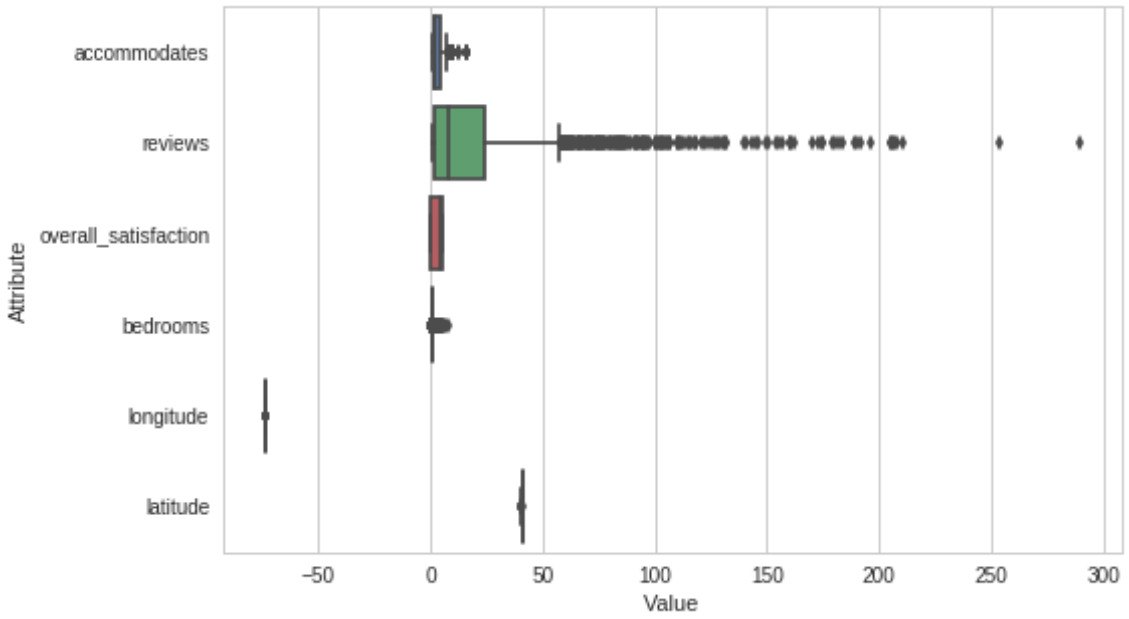

Figure 1: Data before standardization

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.

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?

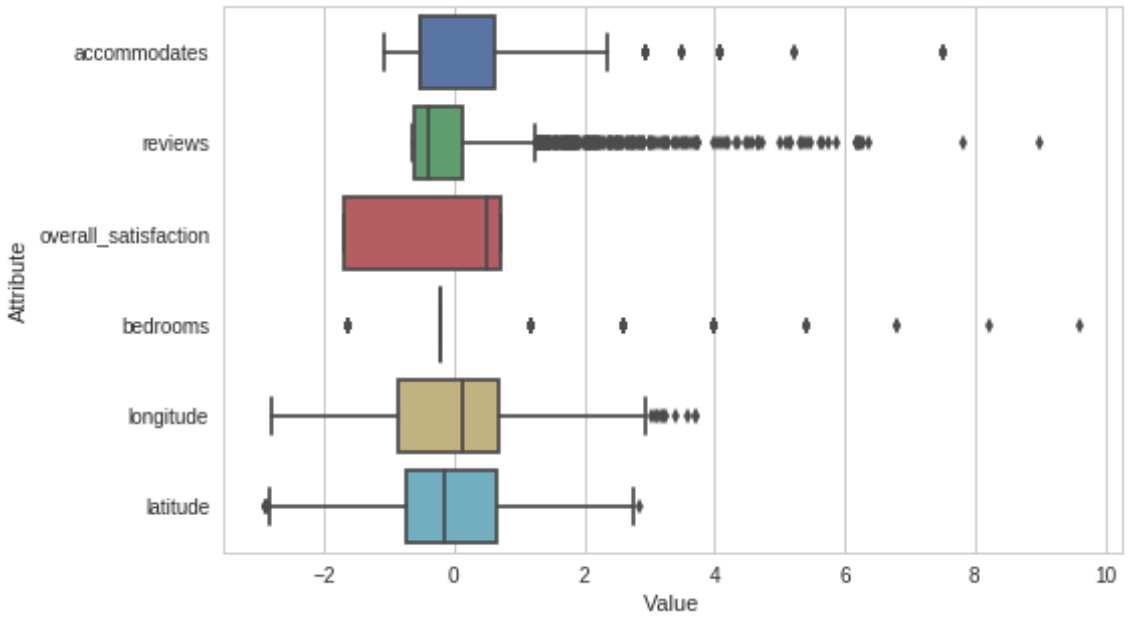

Figure 2: Data after standardization

- 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

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 $[149, \infty)$.

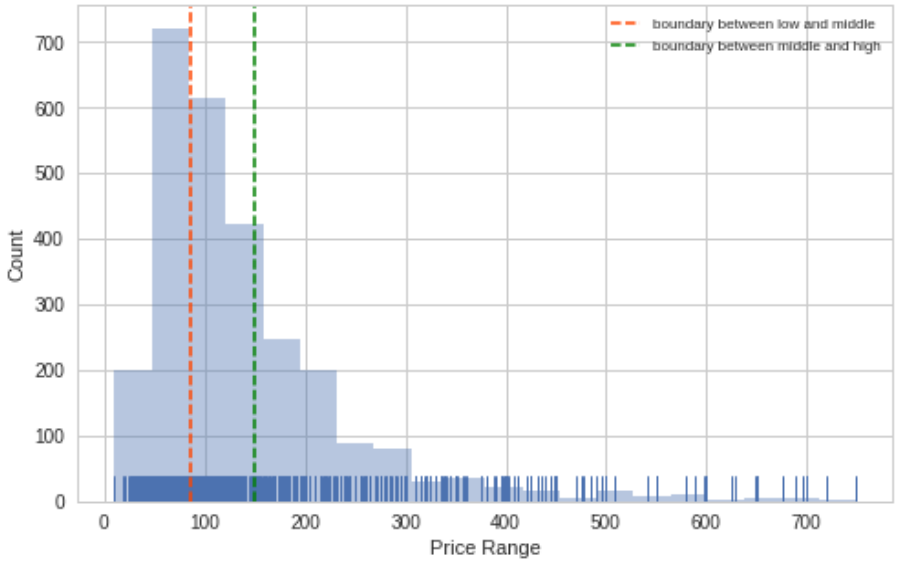

Figure 3: Distribution of prices that is lower than $750

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.

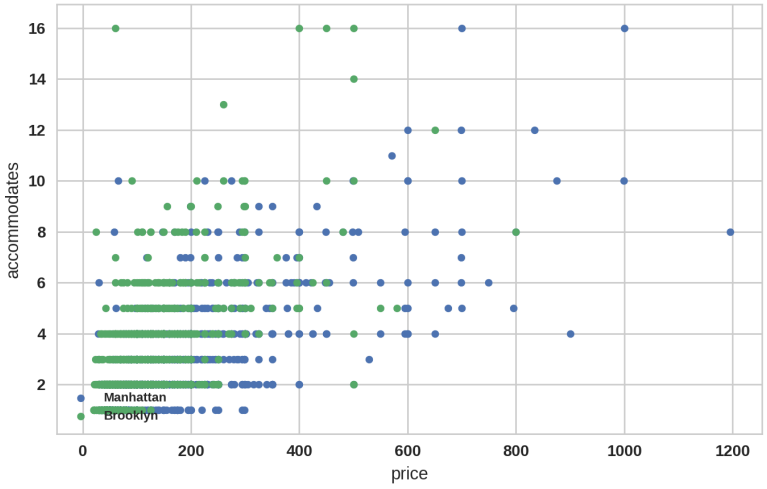

Figure 4: A scree plot showing the percentage variance of each principle component, and a running cumulative total

4.2 Prices of the rooms

4.2.1 Principal Component Analysis

It can be noticed from Fig.5 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.5 that the first 15 eigenvectors can preserve the 90% of the variance of the dataset.

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

The more detialed infomation of the first three eigenvectors are shown in the Table.**??**

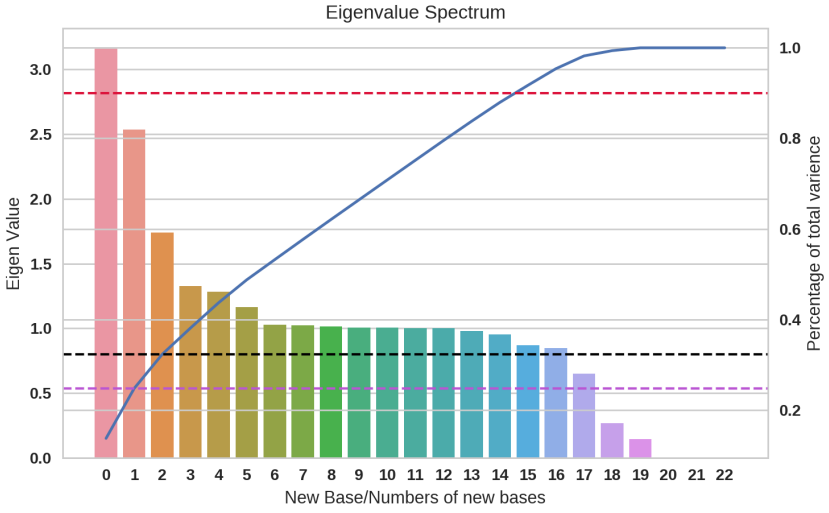

Figure 5: A scree plot showing the percentage variance of each principle component, and a running cumulative total

4.2.2 Clustering

4.2.3 Self-orgnization map

4.3 Overall satisfaction

4.3.1 Principal Component Analysis

4.3.2 Clustering

4.3.3 Self-orgnization map

Each cell is the normalised sum of the distances between a neuron and its neighbours.

4.4 Differences between two boroughs

5 Conclusion

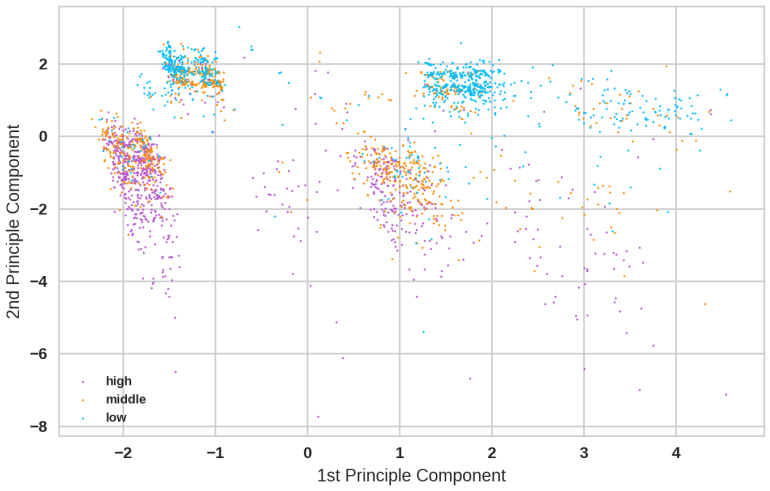

Figure 6: A 2D projection of the PCA results

References

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

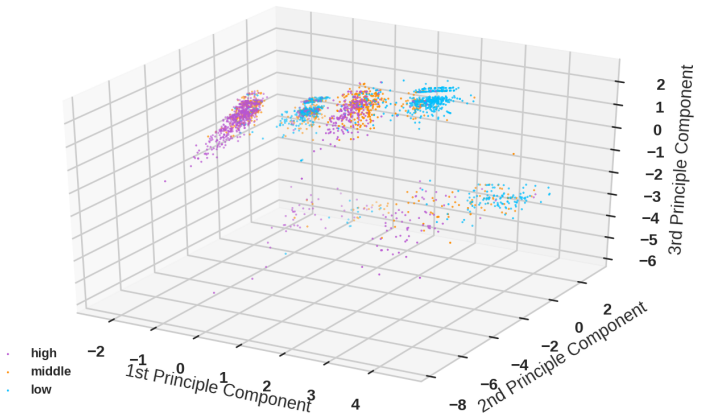

Figure 7: A 3D projection of the PCA results