Data Collection and Analysis of Rental Real Estate

Chen Yu
Glasgow College
University of Electronic Science and Technology of China
Chengdu, China
2288975Y@student.gla.ac.uk

***Abstract*—In this paper, we collect the rental real estate data from the Internet and try to explore and analyse the data. We find some relation between the price of an apartment and its post date, distance from the nearest subway station, area, floor and equipment, in both visual way and quantitative way. We also find some relation between different equipment in the apartment. Finally, a price heatmap is created and some further work is proposed.**

***Index Terms*—Data Mining, Rental Real Estate, Data Visulation, Econometrics**

I. Introduction

With the development of e-commerce, an increasing number of people rent apartments online nowadays. This leads to an opportunity to get and analysis the huge dataset to generate political, scientific and other real-world values [1]. In this paper, we study the patterns of rental real estates based on the data collected from a popular online housing-trading platform Lianjia. Python and EViews are used for both data visualization and analysis. This paper is structured as follows. In Section II, I demonstrated the process of data collection and cleaning, as well as an overview of the dataset. And then an intuitive analysis based on data visualization is presented in Section III, followed by quantitative analysis in the next section. Finally, I would demonstrate an interactive heat map, conclude the results, and propose some further work.

II. Data Collection and Cleaning

A. Data Collection with a Web Crawler

We focus on Shanghais rental real estate section on the Lianjia, where 100 lists of house information are available. In the process of data collection, two libraries requests and BeautifulSoup are mainly used, while the program is disguised as a user with a header to visit the website. The data collection process consists of two steps: getting brief information from the list pages and the detail from the content pages. The data of a real estates district, district at a higher level, community, user-friendly representation of posting date and an associated link are firstly collected from the list pages. And then the program visits the content page for each real estate and collect some more detail information including the price, the way of renting, the indoor type, the area, the direction, the accurate posting date, the introduction, the way of checking in, the rent period, the time for visiting, the floor, the distance from the nearest subway station, the latest trading log in the same community and whether a lift, a garage, a pump electricity or fuel is available. In addition, the information of whether a television, refrigerator, washing machine, air conditioner, water heater, bed, heating, WIFI, wardrobe or natural gas is equipped would also be collected. As a consequence, 34 features of each real estate are described in a CSV table.

B. Coordinates Collection

For later application, we need the accurate coordinates of each real estate. The API from AutoNavi is utilized, hence with the help of which the longitude and latitude information could be obtained from a text description of the address. We first limit the searching are to Shanghai city. Since there are three levels of location description in our dataset, three types of address description would be requested in our algorism, including “Level I + Level II + Level III”, “Level I + Level III”, and “Level III”. In most cases, a correct response could be obtained with at least one type of description. Some examples are shown in the followed table (originally in Chinese). Therefore, the number of features of each real estate is expanded to 36. Note that to keep the table readable, most of the data are still not numerical values until now..

Request	Response
Xu Hui Hua Jing Dong Wan	121.458900,31.122795
Xu Hui Chang Qiao San Cun	121.441774,31.134567
Le Shan Yi Cun	121.428807,31.197827

C. Data Cleaning and Overview

While there are 2996 observations in our dataset, we then clean them into program-friendly forms. We transform the data of price, floor, area and distance from the nearest subway station to numerical values, and the availabilities of equipment are represented by a binary number, 1 and 0 for yes and no. It is notable that only 1243 out of 2996 observations contain the data of the distance from the nearest subway, which means the others do not exist a subway station nearby. As this would bias our data analysis, we construct a feature to represent whether a subway station is available nearby also in the form of a binary number. Also, some noisy data are also dropped in this step, such as some incorrect date. Some data spliced by attributes are shown as followed.

Price

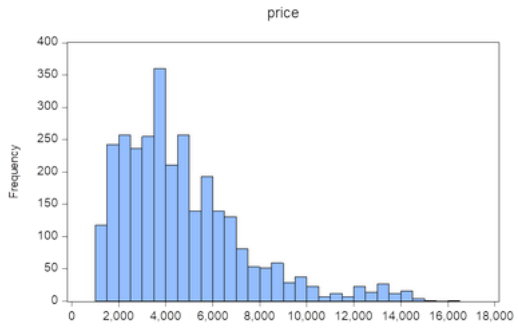

Fig. 1. Historgram of Price.

It could be obtained from the dataset that the minimum price is 1000 yuan, the maximum price is 16000 yuan, with a mean of 4585.086 and a standard deviation of 2634.77. From the histogram as Fig. 1, it could be observed that a most significant number of real estates are with a price under 6000 yuan, and most of them are around 4000 yuan a month.

Floor

The maximum floor of the houses is 50, the minimum of

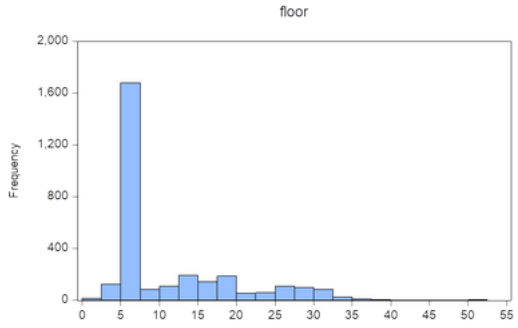

Fig. 2. Historgram of Floor.

which is 1. It is clear from the figure Fig. 2 that more than 1600 apartments are on the floor between 5 and 8.

Area

The average value of the area comes out to be 62.85 meters

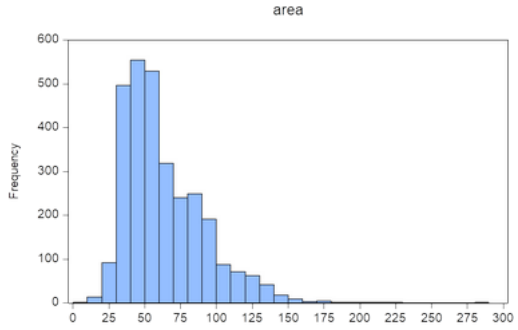

Fig. 3. Historgram of Area.

square, the median is 56, the maximum is 288 and the minimum is 7. The histogram of the data is also shown as Fig. 3.

Equipment and Station Nearby

The dataset demonstrates that 30.1% of the apartments are equipped with a television, 34.6 with a refrigerator, 34.7% with a washing machine, 35.6% with an air conditioner, 34.8% with a water heater, 35.9% with a bed, 13.7% with a heating, 17.6% with WIFI, 35.6% with wardrobe, and 29.2% with natural gas. In addition, a share of 58.5% real states is near subway stations, with a distance of 1 meter to 1194 meters.

III. Data Visualization and Exploration

Since the prices of real estates are usually varied with time, we plot the scatter diagram versus posted date, as shown as Fig. 4. It is hard to find a trend of the price accounting to the diagram. However, the number of apartments increases significantly with the date moving forward.

Fig. 4. Price versus date.

To find some other pattern, we colour the point where an apartment equipped with television is represented with red, and the other with blue as Fig. 5.

We could find that most of the red points are distributed in the upper area of the diagram, which means that apartments equipped with television are usually charged a higher price. This would be described quantitively in the next section.

Similarly, we could use a red colour to represent apartments with refrigerator and blue to represent the other as Fig. 6. It also illiterates the fact that apartments with a higher price

Fig. 5. Price versus date classified by equipment of television.

Fig. 6. Price versus date classified by equipment of refrigerator.

are more likely to be equipped with refrigerators, which is intuitive. We could also find that when the price is low, apartments posted earlier are also more likely to be well-equipped.

We then apply the same technique to the washing machine, air conditioner, water heater, bed, heating, WIFI, wardrobe, and natural gas. They all illustrate similar patterns, as shown in Fig. 7, Fig. 8, Fig. 9, Fig. 10, Fig. 11, Fig. 12, Fig. 13 and Fig. 14.

Fig. 7. Price versus date classified by equipment of washing machine.

When it comes to the classification by the existence of a subway station nearby, the trend becomes more obvious that apartments near a station are usually more worthy, as shown in the figure Fig. 15 below. For those which is near the station, we plot a scatter diagram of prices versus distances from the stations, as Fig. 16. It could be observed that the price range is relatively wide when the distance is short. And while the distance rises up, the prices tend to decay overall.

We also plot a scatter diagram of prices (yuan) versus area (meter square) of the apartments as Fig. 17. A strong positive relation between area and price is shown. It is also interesting that a blank area occurs where the area is around 120 and the price is around 500. It means that if the real estate is with an area of around 120 meters square, it could be utterly expensive or cheap.

IV. Quantitive Analysis of Data

After exploring the dataset, we utilize EViews to analysis the data quantitively, with the help of which could be more ef-

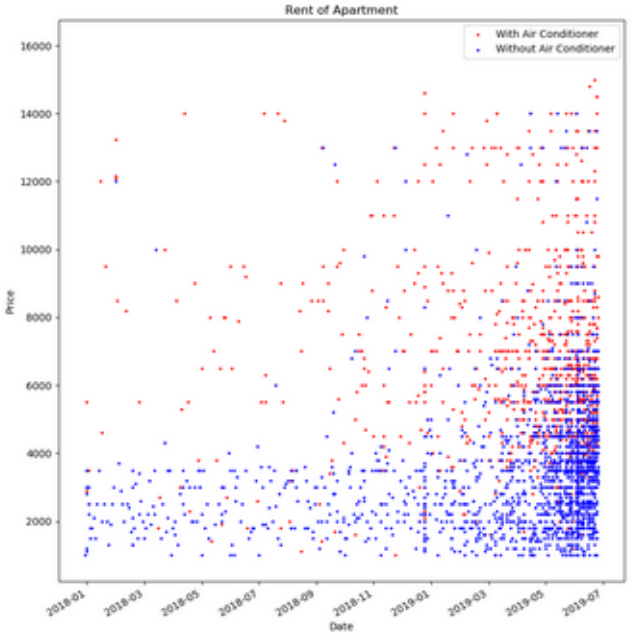

Fig. 8. Price versus date classified by equipment of air conditioner.

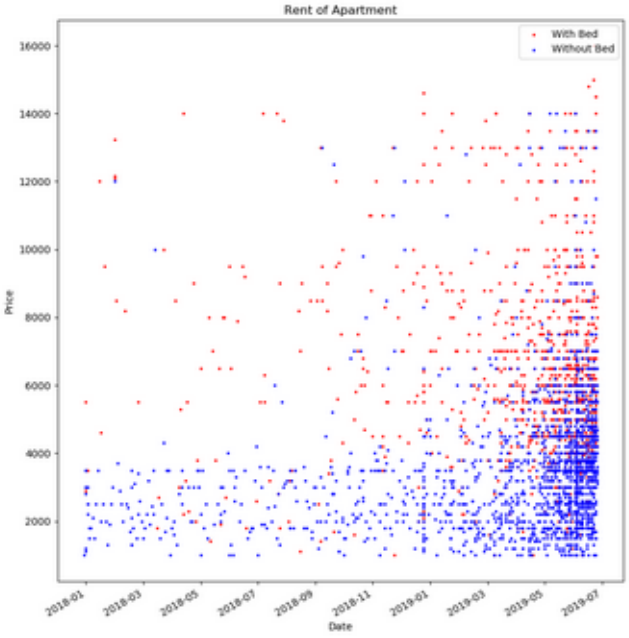

Fig. 10. Price versus date classified by equipment of bed.

Fig. 9. Price versus date classified by equipment of water heater.

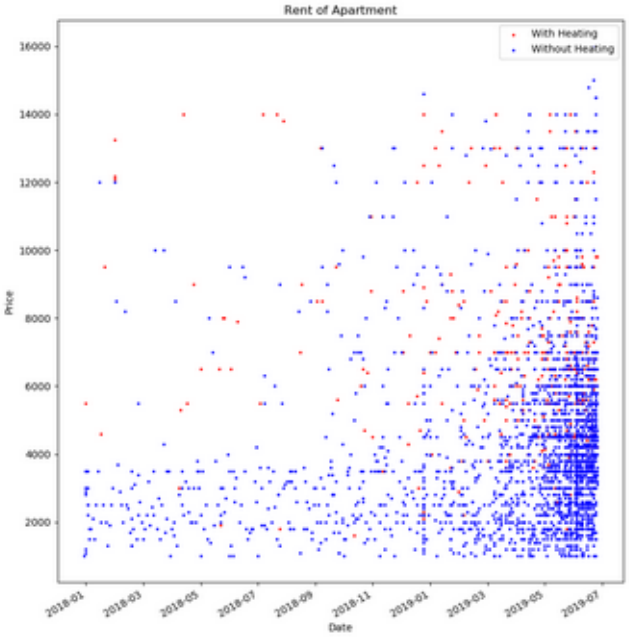

Fig. 11. Price versus date classified by equipment of heating.

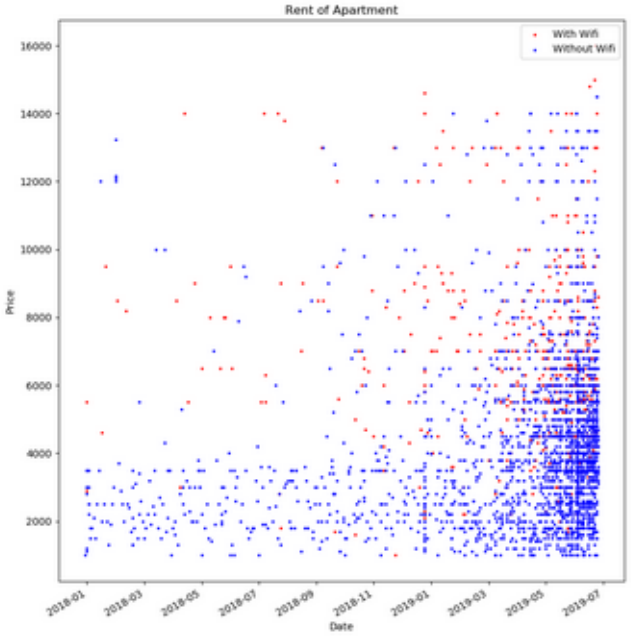

Fig. 12. Price versus date classified by equipment of WIFI.

Fig. 13. Price versus date classified by equipment of wardrobe.

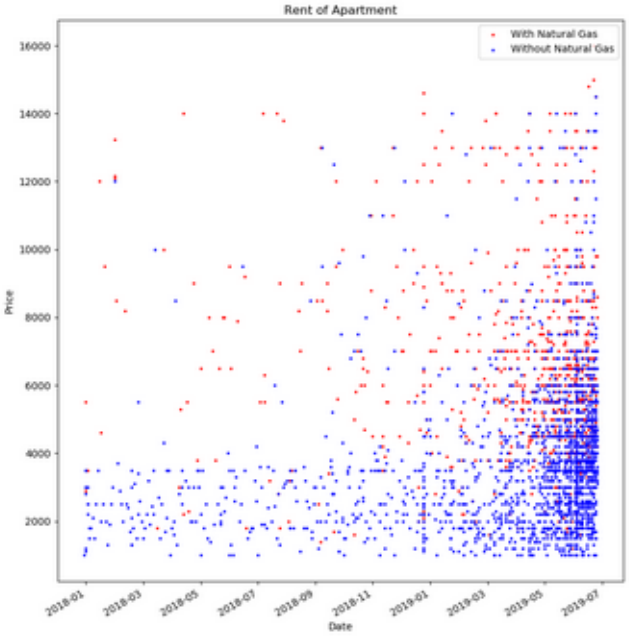

Fig. 14. Price versus date classified by equipment of natural gas.

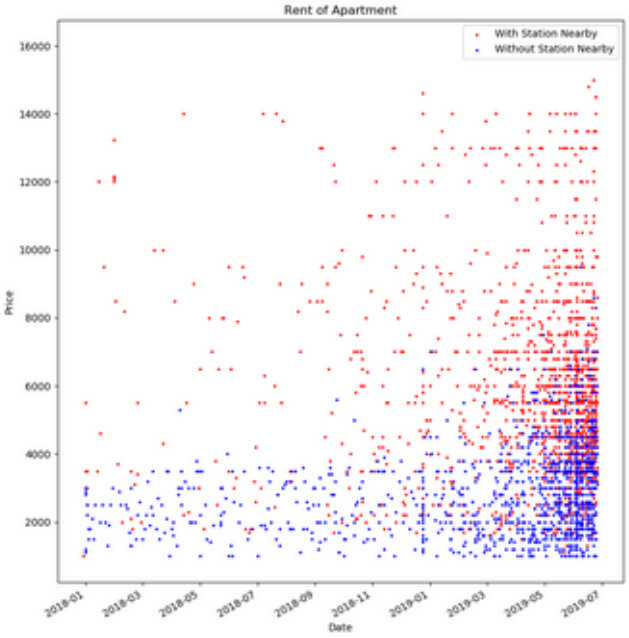

Fig. 15. Price versus date classified by existence of a subway station nearby.

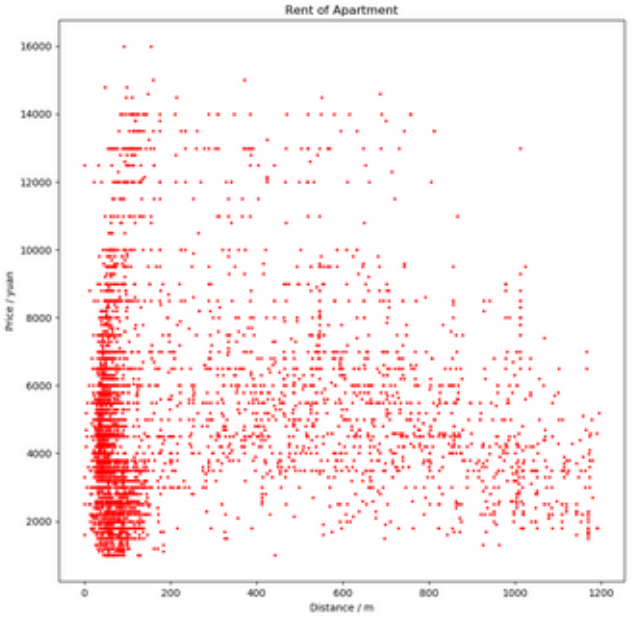

Fig. 16. Price versus distance from a nearest subway station.

Fig. 17. Price versus area.

fective to demonstrate some economic interpretation compared to Python.

A. Prince Estimation

Since the price evaluation of real estate is always attracts huge attentions, we first try to implement the linear regression between prices and other variables, including area, floor, whether a subway station nearby, whether an air conditioner equipped, equipment of beds, equipment of heating, equipment of natural gas, equipment of refrigerator, equipment of television, equipment of wardrobe, equipment of washing machine, equipment of water heater and equipment of WIFI. These variables are represented as PRICE, AREA, FLOOR, SUBWAY, AIR_CONDITIONER, BED, HEATING, NATURAL_GAS, REFRIGERATOR, TELEVISION, WARDROBE, WASHING_MACHINE, WATER_HEATER and WIFI. And hence the equation could be formulated as:

$$PRICE = \beta_0 PRICE + \beta_1 AREA + \beta_2 FLOOR + \beta_3 SUBWAY + \beta_4 AIR\_CONDITIONER + \beta_5 BED + \beta_6 HEATING + \beta_7 NATURAL\_GAS + \beta_8 REFRIGERATOR + \beta_9 TELEVISION + \beta_{10} WARDROBE + \beta_{11} WASHING\_MACHINE + \beta_{12} WIFI + C$$

The method of Least Squares is chosen and the result is shown as Fig. 18.

Dependent Variable: PRICE
Method: Least Squares
Date: 08/01/19 Time: 16:47
Sample: 1 2996
Included observations: 2996

Variable	Coefficient	Std. Error	t-Statistic	Prob.
AREA	15.48407	1.281239	12.08523	0.0000
FLOOR	86.62942	4.331202	20.00124	0.0000
SUBWAY	2251.936	77.30753	29.12958	0.0000
AIR_CONDITIONER	903.3610	351.6985	2.568567	0.0103
BED	-1417.481	399.4906	-3.548221	0.0004
HEATING	253.9783	145.1080	1.750271	0.0802
NATURAL_GAS	117.5028	172.6894	0.680429	0.4963
REFRIGERATOR	683.4454	324.8995	2.103559	0.0355
TELEVISION	-121.3964	190.1767	-0.638335	0.5233
WARDROBE	884.2669	427.2336	2.069751	0.0386
WASHING_MACHINE	654.8161	333.0494	1.966123	0.0494
WATER_HEATER	221.7074	295.2493	0.750916	0.4528
WIFI	-27.77081	147.2388	-0.188611	0.8504
C	635.9064	107.6418	5.907615	0.0000

R-squared	0.519396	Mean dependent var	4585.086
Adjusted R-squared	0.517301	S.D. dependent var	2634.770
S.E. of regression	1830.547	Akaike info criterion	17.86728
Sum squared resid	9.99E+09	Schwarz criterion	17.89534
Log likelihood	-26751.18	Hannan-Quinn criter.	17.87737
F-statistic	247.8998	Durbin-Watson stat	1.834612
Prob(F-statistic)	0.000000		

Fig. 18. Regresion result from Eviews.

It shows that with a significance level of 5%, the coefficients of HEATING, NATURAL_GAS, TELEVISION, WARDROBE, WATER_HEATER and WIFI are not statistic significance. The other variables and their corresponding coefficients with interpretations are shown in the table below.

variables	coefficients	interpretations
AREA	15.48	The price will rise by 15.48 yuan with one additional area.
FLOOR	88.63	One floor higher would lead to 88.63 yuan more expensive
SUBWAY	2251.94	If a subway station is nearby, the apartment is worth more 2251.94 yuan
AIR_ CONDITIONER	903.36	The price would increase by 903.36 yuan when an air conditioner is equipped
BED	-1417.48	The price would decrease by 1417.48 yuan when a bed is equipped
REFRIGERATOR	683.45	The price would increase by 683.45 yuan when a refrigerator is equipped
WARDROBE	884.27	The price would increase by 683.45 yuan when a wardrobe is equipped
WASHING_ MACHINE	654.82	The price would increase by 654.82 yuan when a washing machine is equipped

B. Correction Analysis of Equipment

We then also compute Pearson's correlation coefficient of each two equipment. The formula of that between variables A and B is shown as below:

$$\frac{cov(A,B)}{\sigma_A \sigma_B}$$

And the result is shown in Fig. 19:

Correlation

	AIR_CONDI...	BED	HEATING	NATURAL_...	REFRIGER...	TELEVISION	WARDROBE	WASHING_...	WATER_H...	WIFI
AIR_CONDI...	1.000000	0.968568	0.535622	0.848277	0.954893	0.871211	0.974942	0.960612	0.963478	0.615321
BED	0.968568	1.000000	0.530601	0.837371	0.963746	0.875093	0.985008	0.963672	0.956469	0.611797
HEATING	0.535622	0.530601	1.000000	0.591466	0.546525	0.598341	0.536005	0.546129	0.543361	0.735188
NATURAL_...	0.848277	0.837371	0.591466	1.000000	0.850878	0.879429	0.842891	0.851601	0.851845	0.688212
REFRIGER...	0.954893	0.963746	0.546525	0.850878	1.000000	0.884945	0.962795	0.966693	0.943547	0.628388
TELEVISION	0.871211	0.875093	0.598341	0.879429	0.884945	1.000000	0.879282	0.884028	0.869277	0.687319
WARDROBE	0.974942	0.985008	0.536005	0.842891	0.962795	0.879282	1.000000	0.962750	0.958422	0.613976
WASHING_...	0.960612	0.963672	0.546129	0.851601	0.966693	0.884028	0.962750	1.000000	0.955170	0.625647
WATER_H...	0.963478	0.956469	0.543361	0.851845	0.943547	0.869277	0.958422	0.955170	1.000000	0.624737
WIFI	0.615321	0.611797	0.735188	0.688212	0.628388	0.687319	0.613976	0.625647	0.624737	1.000000

Fig. 19. Correction of equipment from EViews.

It could be obtained that some pairs of equipment have high correlation coefficients (greater than 0.95): air conditioner and bed, air conditioner and refrigerator, air conditioner and wardrobe, bed and refrigerator, bed and wardrobe, bed and washing machine, bed and water heater, refrigerator and wardrobe, refrigerator and washing machine, refrigerator and water heater, wardrobe and washing machine, wardrobe and water heater, washing machine and water heater. It means that these devices usually occur in pair in the apartment.

V. Data Interactive Visualization

In this section, we try to visualize the data in a more interactive way. We locate each apartment on a map and mark the prices with different colours, while the highest is red, the medium is orange and the lowest id green. Our method is to first transform the coordinate data with GCJ-02 standard into BD-09 standard and then output the results in the appropriate format, which contributes as one part of the HTML code. API from Baidu Map is also utilized. The demonstration is Fig. 20 as followed, which show intuitively that the real estates in the centre of the city maintain higher prices than other overall. This map could be accessed through http://www.whoischen.com/heatmap/.

Fig. 20. Price heatmap.

VI. Conclusion

In this paper, we try to collect the data of rental real estate in Shanghai from Lianjia, clean the data, explore the data and do some analysis to find some pattern in the collected information. However, since the time is limited, the analysis does not go too much in the mathematical direction, and hence this has been remained to be the further work of this research. Since the coordinate of each real estate has been calculated, finding some useful pattern based on the location could also be tried in the future.

References

[1] H. Zhang, N. Parikh, G. Singh, and N. Sundaresan, "Chelsea won, and you bought a t-shirt: characterizing the interplay between twitter and e-commerce," in *IEEE/ACM International Conference on Advances in Social Networks Analysis & Mining*, 2013.