Financial Engineering and Risk Management (2021) 4: 59-64
Clausius Scientific Press, Canada

DOI: 10.23977/ferm.2021.040408
ISSN 2523-2576

Design and implementation of second-hand housing data statistical analysis system

Jun Zhang[1, *], Taizhi Lv[2]

[1]*College of Information Technology, Jiangsu Maritime Institute, Jiangsu Nanjing, 211170, China*

[2]*Nanjing Longyuan Microelectronic Company Limited, Jiangsu Nanjing 211106, China*

Corresponding author: 1052871890@qq.com

Keywords: web crawlers, housing price, Python, data visualization, KNN.

Abstract: The content of this paper is the statistical analysis of the housing price data in Wuxi. Obtain data on the net, visualize the data, to see the prices clearly, judge the influence prices of each element, use linear regression to find out the price per square meter and the relationship between the building area, through the KNN algorithm to divided into high-grade village, compare the Euclidean distance and the Manhattan distance of the differences in house prices problem. This system is based on Python language, MongoDB stores data, uses MySQL to process relevant data, uses PyCharm as the development tool, Python 3.9 as the running environment, uses Scrapy framework to crawl the second-hand house data of LianJia network, and stores the data into MongoDB. After dirty data processing, we use lightweight Web application framework Flask and Echarts to conduct visual analysis on the Web page. Finally, the linear regression algorithm is used to find out the elements related to the price, and the KNN classification algorithm is used to divide the residential area into three grades by the level of the housing price.

1. Instruction

The network provides us with a lot of data information. Using crawlers to crawl the online housing information, integrate, clean and statistically analyze the data. Useful information can help people see the house price trend more clearly and intuitively, and consider all aspects of elements more comprehensively to find the most suitable housing [1].

As the third most important city in important economy province of China, the fluctuation trend of house prices in Wuxi has always affected people's hearts, but it is not as exaggerated as that in first tier cities, so there is less research on house prices in Wuxi. Therefore, by analyzing the data of second-hand houses in Wuxi, we can clearly and intuitively see the past and present situation of house prices in Wuxi, analyze the current situation of house prices from various factors, find out the law and predict the trend of house prices in the future. Secondly, for people who rent and buy a house, deciding what kind of house to choose needs to comprehensively consider all factors affecting house prices. The classification of residential areas in this paper has reference value for those who choose housing. People with small budget can choose to choose houses with high-cost performance in medium and low-grade residential areas, while people with sufficient budget can choose houses in

high-grade residential areas.

2. Data crawling and processing

2.1 Data crawling

The object of this project is lianjia.com, which provides comprehensive house price information all over the country. This time, data analysis is conducted for second-hand houses in Wuxi, and information such as location, property right, orientation, community name, house type, floor, area, price, house type, listing time, etc. are obtained. Based on the scrapy framework [2], the system crawls the second-hand house price data in Wuxi in five steps.

Step 1: Firstly, the core of the framework - engine processes the data flow of the system and triggers transactions.

Step 2: Subsequently, the scheduler receives the request and puts the URL of the received house information list into the queue, and then he performs the crawling strategy to determine the crawling order.

Step 3: Send the obtained URL encapsulation request to the downloader through the engine.

Step 4: Download the house price list through the downloader, encapsulate and return the response, which is parsed by the crawler, and send the data to the project pipeline.

Step 5: The data processed by the project pipeline is stored in MongoDB.

Step 6: Repeat the first step. After crawling the URL of the house information list, crawl the specific house information in each URL.

2.2 Data storage

The system uses MongoDB to store the crawled data without creating a table. After obtaining the connection with mongoclient() in the pipeline of Scrapy. The system performs the MongoDB insertion operation self.client.fang.wx.insert (item), where "fang" is the library name and "wx" is the table name. After execution, the table will be created automatically to insert the data.

2.3 Data clean

Data cleaning is the first step of data analysis. After crawling data, a large number of data often have many problems, and there will be a lot of meaningless data or non-standard data. Data cleaning, deletion and modification of duplicate data, incomplete data and unreasonable data are required. Generally, dirty data is divided into three categories: incomplete data, wrong data, duplicate data [4]. Through data cleaning, the problems of missing value, out of bounds value, inconsistent code, duplicate data and so on are solved.

3. Data analysis

3.1 Data statistics

The system uses the flash framework to make web pages. First, create a flask instance: app = flask (_name_), with the current module (_name_) as the parameter. Start the flash program through run(), and record the URL through the origin to access the page [3]. In the system, @ app.route ('/ index. HTML'), where "/ index. HTML" is the route of the system home page, so enter the URL of http: / / localhost: 5000 / index.html, and the route function will be called to obtain the return value and return to the browser. After accessing the URL, you need to return to an HTML page. In the static page, the

template needs to be used to modify the information to meet the different information displayed on the page. The writing of variables and logic code on the page needs to be handled by jinja2, with {}} representing variables and {%%} representing code logic.

The visualization of data is realized by the combination of flask framework and echarts. Firstly, the database is connected through cursor operation, the written SQL statements are stored in the cursor for processing, and the fetchall () method is used to accept all results and return items to the defined MySQL () method. Call render_ The template () template transmits the query values in the database to the front end. Take the page named "charts" in the system as an example: render_ Template ('charts. HTML ', items2 = items2), where the data is sent to the front end in the form of a dictionary. Figure 1 shows the proportion of each orientation of housing information, and Figure 2 shows the relationship between housing orientation and price.

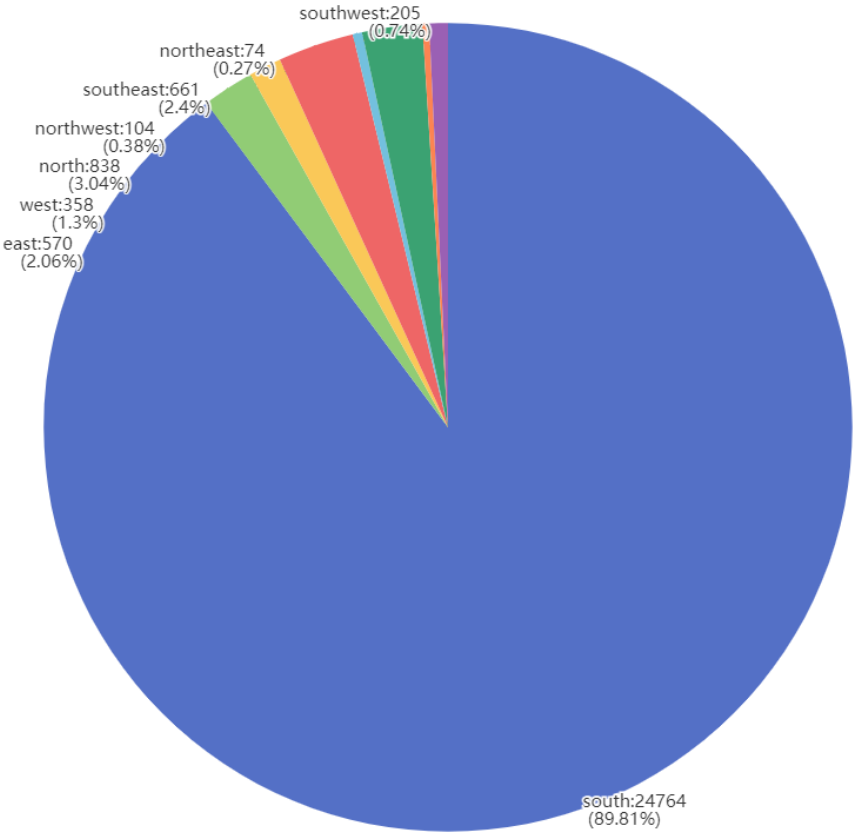

Figure 1: Proportion of each orientation of the house

It can be seen from figures 1that houses facing south account for the largest proportion, followed by those facing north and Southeast. From Figure 2, it is found that houses facing south and southeast generally have high prices, while houses facing north have the lowest average price. It can be concluded that houses facing south have good daylighting and are popular. Most of the houses facing north are apartments, which are characterized by economic benefits.

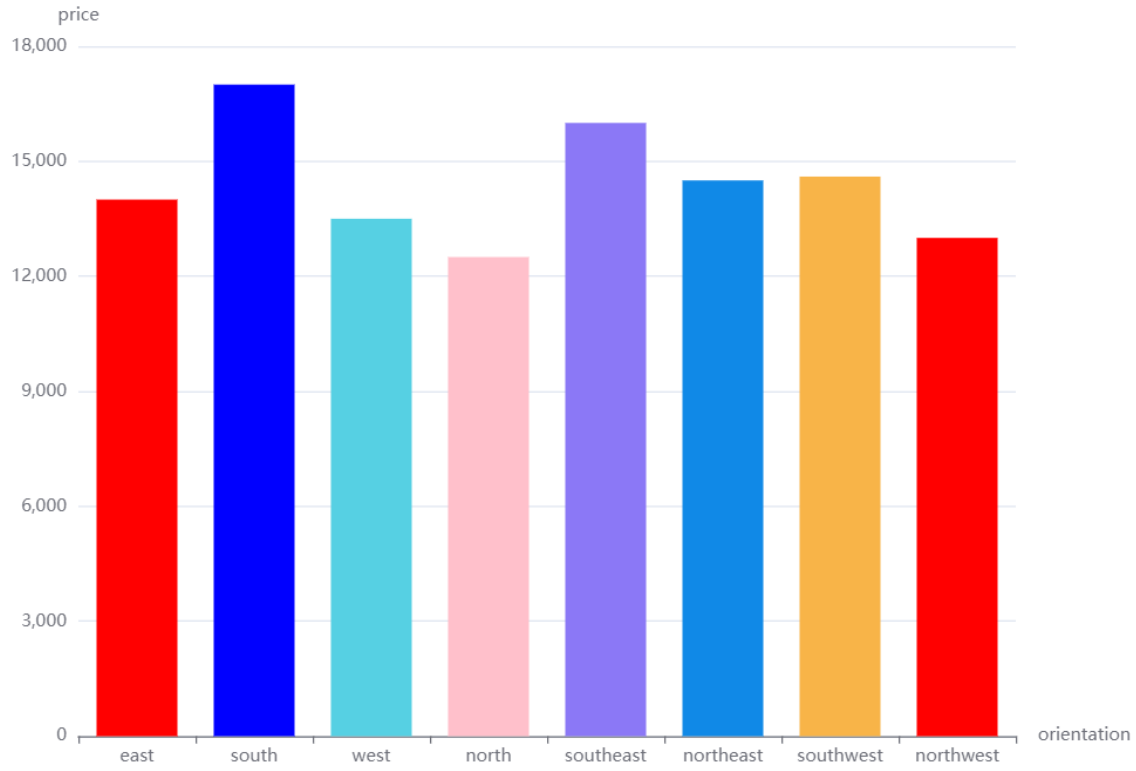

Figure 2: Relationship between orientation and house price

3.2 House classification analysis

The crawling house price data does not classify the community, and it is necessary to consider the community type when selecting houses. Therefore, KNN classification algorithm is used to classify the community. In this analysis, the grade of the community is classified according to the unit price per square meter of houses. The number of rooms, direction, decoration and floors are taken. These elements are all factors affecting the house price. In order to facilitate the calculation, the data are processed. The low, medium and high floors are converted into 1, 2 and 3 respectively according to the price, and the eight directions are converted into numbers 1-8 according to the price. The decoration sets the hardcover with the highest price as 1, the blank as 2 and the simple decoration as 3. The grade of the community is divided into three grades: high, middle and low. The community is classified in three aspects: comprehensive decoration, floor and orientation. Similarly, the grade is converted from high to low into numbers, which are replaced by 0, 1 and 2 respectively.

KNN refers to the K neighbors most similar to the test data. Among the K neighbors, find the category with the highest frequency, and the test data can be classified into this category to realize the classification of data [5]. Finding the most similar K neighbors is to calculate the distance between the measured data and other data. There are many formulas to calculate the distance between the measured data and other data. This paper mainly analyzes Manhattan distance and Euclidean distance, and will compare the two distances to find the result with the highest accuracy. The Manhattan distance formula is d1=$\sum_{k=1}^{m}|a-b|$, and the Euclidean distance is $d2=\sqrt{\sum_{k=1}^{m}(a-b)^2}$.

Finally, the test calculates their accuracy for different K values and different distances. Here, set the K value range to 1-4 and traverse the cyclic K value. The final results are shown in Figure 3.

	k	Distance Function	Accuracy
0	1	l1_distance	0.845024
1	2	l1_distance	0.840131
2	3	l1_distance	0.862969
3	4	l1_distance	0.854812
4	1	l2_distance	0.848287
5	2	l2_distance	0.848287
6	3	l2_distance	0.869494
7	4	l2_distance	0.859706

Figure 3: Distance function and K value result diagram

The l1_distance in the Figure 3 is Manhattan distance, l2_ distance is Euclidean distance. It is concluded that when k value is 3 and Euclidean distance is selected, the highest accuracy is 0.869494. Comparing two different distance functions, it can be seen that in the case of different K values, although the difference between the two results is small, the accuracy of European distance is slightly higher than that of Manhattan distance.

4. Conclusion

This paper studies from crawler to data analysis, learns the scrapy framework, deeply studies the principle of crawler, crawls the second-hand house price data in Wuxi, crawls 30000 pieces of data in total, processes these data, visualizes the data and analyzes the current situation. KNN algorithm is used to learn cell classification, and the accuracy of learning results is about 86%. The difference and error between European distance and Manhattan distance are studied. It is found that the error difference between them is not very large, but there are differences and cannot replace each other. On the problem of cell classification, the accuracy of European distance is slightly higher than that of Manhattan distance, and European distance is more suitable for this kind of problem.

Acknowledgements

This work was financially supported by the funding of Qianfan science and technology team of Jiangsu Maritime Institute (Big data analysis and application research team), young academic leaders of Jiangsu colleges and universities QingLan project, excellent teaching team of Jiangsu colleges and universities QingLan project (Innovative teaching team of software technology specialty) , and the big data collaborative innovation center of Jiangsu Maritime Institute.

References

[1] Xu, Lulin, and Zhongwu Li. "A new appraisal model of Second-Hand housing prices in China's First-Tier cities based on machine learning algorithms." Computational Economics 57.2 (2021): 617-637.

[2] Gan, Lu, et al. "Coupling coordination analysis with data-driven technology for disaster–economy–ecology system: an empirical study in China." Natural Hazards (2021): 1-31.

[3] Choudhury, Aakash, et al. "HealthSaver: a neural network based hospital recommendation system framework on flask webapplication with realtime database and RFID based attendance system." Journal of Ambient Intelligence and Humanized Computing (2021): 1-14..

[4] Q. Jiang, S. Yan, H. Cheng and X. Yan, "Local–Global Modeling and Distributed Computing Framework for Nonlinear Plant-Wide Process Monitoring With Industrial Big Data," IEEE Transactions on Neural Networks and Learning Systems, vol. 32, no. 8, pp. 3355-3365.
[5] Xing, Wenchao, and Yilin Bei. "Medical health big data classification based on KNN classification algorithm." IEEE Access 8 (2019): 28808-28819.