Research on Second-hand Housing Market Based on Big Data Technology--A Case Study in Shenyang

Yitong Liu*

Corresponding author: qiya841030290@sina.com*

Dept. of Management, Shenyang Jianzhu University, 25 Hunnan Middle Road, Shenyang, China 110168

Abstract: With the application of computer technology in various aspects of regional governance and economic operation, data has become a key factor affecting regional development. It is increasingly important to obtain and analyze data. With the help of Selenium automated testing technology, Python data processing and visualization methods, this paper collects the public housing information of second-hand houses in Shenyang, and analyzes the market situation. The results show that Python+Selenium technology has powerful functions in data acquisition, data analysis and data visualization, and the code is simple. Computer technology provides an effective tool for scientific research and local government supervision of the real estate market.

Keywords: Big data, second-hand houses, Python, Selenium, data analysis

1. Introduction

In the real estate industry, the information of the sellers and buyers is asymmetric in many cases. With the rapid development of the Internet, many real estate portals are building communication channels between sellers and buyers, which is particularly evident in the second-hand housing market. Many second-hand housing owners hang the housing information in the portal websites, which makes the scattered and difficult collected second-hand real estate information can be quickly extracted and analyzed through big data technology.

The information on the Internet is presented in the form of Web pages and cannot be directly used for scientific research analysis. It is inefficient and error-prone to obtain research data by manual extraction. There have been many studies on obtaining raw information from web pages and processing them into data that can be directly used for analysis. The popular extraction tools include MDR[1], improved method Delta[2], etc. Mei X. et al.[3] based on the design criteria of web template have proposed the method of fully automatic generation of web page information extraction wrapper - PSNT (extraction based on template Structure and tag tree). Wang J.Z. and Lai W.J.[4], Nie L.J. et al.[5], Feng C.J.[6], Zhang H.Z. et al.[7], Wang F.C.[8] use Python built-in or third-party libraries such as Requests, Scrapy, Numpy, Matplotlib, Beautiful Soup, ArcPy to automatically extract data and analyze real estate information.

With the widespread use of dynamic web pages, Selenium, as an open source automated testing tool, provides a basic framework for data acquisition of Web sites. Combined with the

ICBBEM 2023, May 19-21, Hangzhou, People's Republic of China

DOI 10.4108/eai.19-5-2023.2334271

application of Python library related with data analysis and visualization, they can efficiently screen, clean and analyze the huge and complex information. Therefore, this paper takes the second-hand houses in Shenyang as an example. Through Python+Selenium, we can obtain and comprehensively analyze the public housing information of the second-hand houses in Shenyang, such as the price, floor area, house type, location and the distance from the subway, so as to provide basic information of the second-hand housing market for social science research and provide reference for local supervision.

2. Methodology

2.1 Selenium automated test framework

With the wide application of JavaScript technology, more and more web pages rely on JavaScript scripts to load data dynamically. A method of data acquisition based on Selenium automated testing has been proposed. Selenium is a set of tools mainly used for automated testing of web applications. WebDriver is a part of Selenium and also the main part of browser behavior simulation[9]. WebDriver supports Firefox, Chrome, IE and other mainstream browsers, Windows, Linux, MacOS and Android systems, as well as Python, Java, R and other programming languages. It can automatically manipulate the browser to simulate the interaction between people and browser, and then realize the behavior of visiting websites.

2.2 Python libraries related with data analysis and visualization

Numpy is the abbreviation of Numerical Python and is a common scientific calculation module. However, in terms of statistical analysis, Numpy only provides basic functions, and it needs to be used in conjunction with Pandas. Pandas is a data statistical analysis library based on Numpy. Pandas incorporates a large number of standardized models and methods for efficient data processing. It is an effective tool for processing multidimensional data.

Matplotlib is a data visualization module based on Numpy's array operation function that can generate high-quality graphics[10]. Seaborn is a Python data visualization library based on Matplotlib. It provides a highly interactive interface for drawing attractive and informative statistical charts. Seaborn carries out more advanced API encapsulation on the basis of Matplotlib, which makes drawing easier[11]. Pyechart is an open source data visualization tool. With the help of 400+map files and native Baidu maps, it can provide strong support for geographic data visualization.

3. Data acquisition and cleaning

Through Python, Selenium was called to obtain the unit price, total price, floor area, house type, location, adjacent to subway and number of followers of second-hand houses in Shenyang from the website. Import WebDriver and install Selenium library in Python. Xpath method was used to locate the target information, and the obtained information was stored in the form of a dictionary. For the houses not adjacent to the subway, used try and except statements to write 0 under the corresponding column, and finally used the dictionary to build a DataFrame object, and wrote this DataFrame into a CSV file.

Because the data obtained from the website may have duplicate values, missing values, non-standard data and other problems, it is necessary to standardize these data before data analysis. Data cleaning was mainly carried out from the following aspects:

First, duplicate values were deleted and missing values were filled in. Drop_duplicates method was used to delete the duplicate data. Isnull method was used to find the missing value. Missing values were filled with Fillna method, and filled in 0 or None according to the data type.

Second, data type was transformed. Among the collected data, the floor area, unit price, total price and number of followers were all text data with units, which couldn't be directly used for data analysis. Replace method was used to remove the unit information, and astype ("float") was used to convert the text type to floating point data.

Third, outliers were removed. The Drop method were used to delete outliers with a floor area of more than 300 m^2 and a total price of more than 4.5 million yuan.

After data cleaning, 2994 second-hand housing data were retained.

4. Data Mining

Describe function was used to calculate the number, mean value, variance, minimum value, maximum value and quartile value of floor area, unit price and total price. The statistical results are shown in Table 1. It can be seen from Table 1 that the average floor area of second-hand houses in Shenyang is 84.22 m^2, the average unit price is 9738.68 yuan/m^2, the average total price is 839200 yuan.

Table 1. Descriptive statistical analysis of floor area, unit price and total price of second-hand houses in Shenyang.

Item	Floor area/m^2	Unit price/Yuan/m^2	Total price/Ten thousand yuan
count	2994.00	2994.00	2994.00
mean	84.22	9738.68	83.92
std	25.69	3293.24	47.83
min	28.69	3139.00	16.90
25%	69.64	7559.25	55.00
50%	83.13	9113.50	75.00
75%	90.45	11223.75	97.72
max	210.51	27178.00	440.00

4.1 Unit price

Use the Skev and Kurt functions to calculate skewness and kurtosis of the unit price of second-hand houses in Shenyang. Then draw the histogram of the unit price and its kernel density estimation curve through the Displot function (as shown in Figure 1). The calculated skewness is 1.24, which is greater than zero, and its kernel density estimation curve is skewed

to the right. It can be seen that the mode of the unit price is less than the mean, which means the unit price of most second-hand houses in Shenyang is less than 9738.68 yuan/m^2. From the perspective of kurtosis, the calculation result is 2.46, less than 3. The data distribution of the unit price of second-hand houses is slightly flat than the normal distribution, and the extreme value is less.

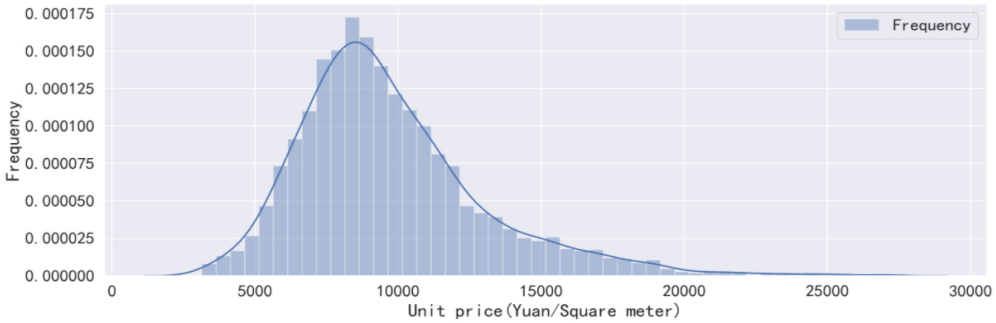

Figure 1. Histogram of house unit price and its kernel density estimation curve.

4.2 Total price and floor area

Concat method was used to splice the total price and floor area series, and make a scatter chart, as shown in Figure 2. The X axis represents the floor area, and the Y axis represents the total price. There is a positive correlation between the total price of houses and the floor area as a whole. Intensive housing supply is concentrated in the floor area of 25-100 m^2 with a total price of 0-2 million yuan. The total price with a floor area of 25-50 m^2 do not exceed 1 million yuan. The total price with a floor area of more than 125 m^2 is relatively discrete, especially the houses with a floor area of 150-175 m^2 have the highest degree of dispersion. On the one hand, the supply of second-hand houses with floor area more than 125 m^2 is reduced, and on the other hand, the demand is non-rigid. Therefore, the total price is more affected by other factors, showing an increased dispersion of the price between the houses with similar floor areas.

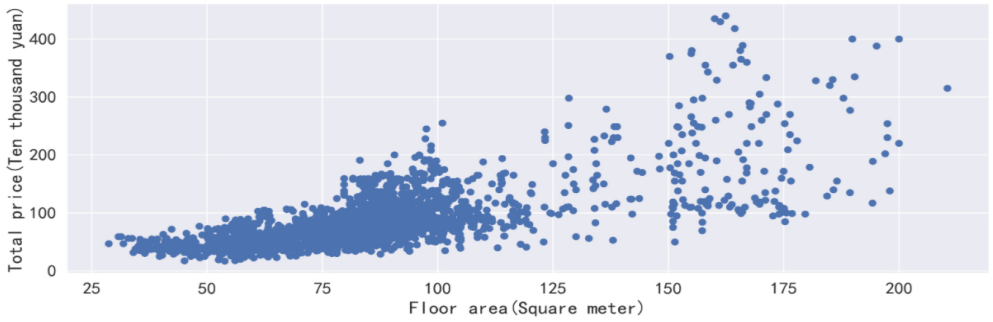

Figure 2. Scatter chart of total price and floor area of second-hand houses in Shenyang.

4.3 House type

Value_counts function was used to calculate the types of second-hand houses in Shenyang. The top five types of houses are two bedrooms and one living room, two bedrooms and two

living rooms, one bedroom and one living room, three bedrooms and two living rooms, four bedrooms and two living rooms, and their number is 1544, 1000, 178, 129 and 53 respectively. The floor area, unit price, total price and number of followers of each house type are shown in Table 2.

Table 2. Mean of floor area, unit price, total price and number of followers of each house type.

House type	House/ Number	Floor area/ m^2	Unit price/ Yuan/m^2	Total price/ Ten thousand yuan	Followers / Number
two bedrooms and one living room	1544	76.63	9279.15	71.96	10.97
two bedrooms and two living rooms	1000	88.00	10186.83	89.96	11.45
one bedroom and one living room	178	50.30	9425.98	46.12	8.76
three bedrooms and two living rooms	129	143.45	11654.60	171.82	22.59
four bedrooms and two living rooms	53	162.71	11957.43	195.73	21.06

From Table 2, we can see that the supply of two bedrooms and one living room, two bedrooms and two living rooms is large, but the demand concern is not high. Due to the larger floor area and higher attention, the average unit price and total price of three bedrooms and two living rooms and four bedrooms and two living rooms are higher among the five types of houses.

4.4 Location

In order to analyze the spatial distribution of second-hand houses, first use the Groupby method and Count function to aggregate the data according to the location, and use Map_visualmap function to map the number of second-hand houses in each districts of Shenyang, as shown in Figure 3. From Figure 3, we can see that the second-hand houses are mainly concentrated in Hunnan District, Yuhong District and Tiexi District, followed by Heping District, Huanggu District and Shenbei District, while the second-hand houses in Dadong District, Shenhe District and Sujiatun District are fewer.

Then use the for loop to iterate through all the information of the second-hand houses in each district, and use Sort_values function to sort according to the number of second-hand houses in each sub-district. It can be found that there is a large supply of second-hand houses in Aoti Sub-district, Daoyi Sub-district and Changbai Sub-district. The number of second-hand houses in each of them is more than 200. Changqing, Xinshifu, Yuhongxincheng, Jingjikaifaqu, and Xisantaizi these 5 sub-districts each have a supply of 101-200 houses. The total supply in these 8 sub-districts reaches 1244, accounting for 41.55% of the total second-hand housing market.

Groupby. agg (np. mean) method was used to calculate the average of the floor area, unit price and total price of second-hand houses in each sub-district. The data are sorted according to the average value of the total price, and a bar chart is made to show the data of the top 3, bottom 3

and main distributed sub-districts. In order to better display the data in the chart, the average value of the unit price is divided by 100 to reduce the order of magnitude. The bar chart is shown in Figure 4. Dongwulihe, Sanhaojie and Changbai are the top three sub-districts in total price, and Changbai is also the third sub-district in total supply. Both the unit price and floor area in Dongwulihe and Sanhaojie are high, so that the total price is significantly higher than that in other sub-districts. The houses in Changbai are generally high in unit price, small in floor area and high in total price. Xita, Puhexincheng and Maguanqiao are the last three sub-districts in the total price. The unit price of houses in Xita is higher and the floor area is smaller, while unit price in Puhexincheng and Maguanqiao is lower and the floor area is larger. As the sub-districts where the supply of second-hand houses are concentrated, the floor area and unit price of houses in Aoti are significantly higher than those in Daoyi.

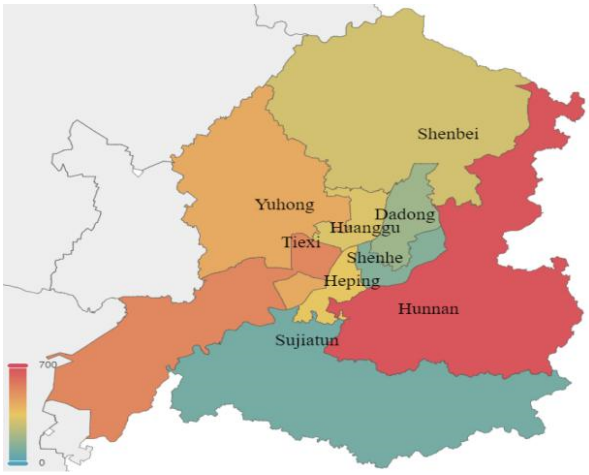

Figure 3. Distribution map of second-hand houses in Shenyang urban districts.

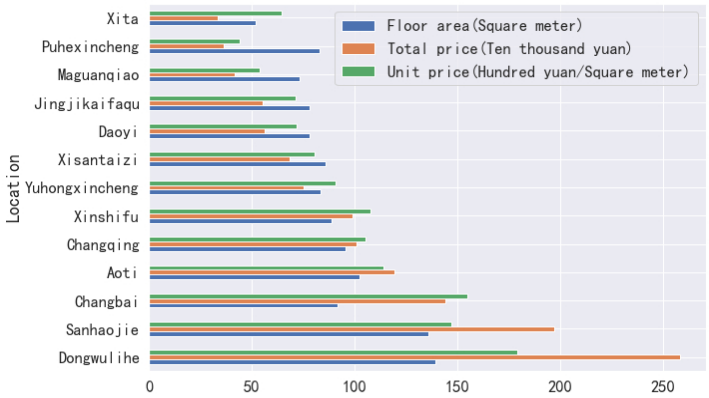

Figure 4. Bar chart of floor area, unit price and total price of second-hand houses in parts of Shenyang.

4.5 Distance from the subway

By Groupby method, Second-hand houses of adjacent subway and nonadjacent subway were divided. Further used Get_group method and Value_Counts function, found that the second-hand houses near the subway in Shenyang are mainly concentrated in the Aoti, Changbai and Changqing Sub-district, with 197, 96 and 85 houses respectively. The nonadjacent-subway second-hand houses are mainly concentrated in Daoyi, Changbai and Xinshifu Sub-district, with the number of houses 151, 129 and 115 respectively. The mean, skewness and kurtosis of the floor area, total price and number of followers of the houses near the subway and the houses faraway the subway are calculated by Mean, Skiv and Kurt functions. The results are shown in Table 3.

Table 3. Mean, skewness and kurtosis of the floor area, total price and number of followers of the houses adjacent to subway and nonadjacent to subway.

	House/ Number	Floor area			Total price			Followers		
		Mean/ m^2	Skewness	Kurtosis	Mean/ Ten thousand yuan	Skewness	Kurtosis	Mean / Number	Skewness	Kurtosis
Adjacent to subway	1207	85.34	1.37	3.28	92.24	2.44	10.29	14.33	3.49	16.33
Nonadjacent to subway	1787	83.46	1.73	4.68	78.29	3.28	15.49	9.97	16.10	438.56

From Table 3, it can be seen that the mean, skewness and kurtosis of the floor area of the houses adjacent to subway and nonadjacent to subway are not significantly different. On the whole, the floor area of most nonadjacent-subway second-hand houses is slightly smaller than that of the houses adjacent to subway, and the degree of dispersion is greater. In terms of total prices, the difference between the average price of houses adjacent to subway and nonadjacent to subway is 139500 yuan. Because the skewness and kurtosis of the nonadjacent-subway houses are greater than those of the adjacent-subway houses, the price difference between most of them is higher than 139500 yuan. In terms of the degree of concern, the mean, skewness and kurtosis of number of followers show that the degree of concern of the adjacent-subway houses is significantly higher than that of the nonadjacent-subway houses.

5. Conclusion

This paper uses Python+Selenium technology to obtain effective data and carries out data mining on the second-hand housing market of Shenyang from different aspects. The results show that Selenium automated testing technology has greatly improved the efficiency of data acquisition. In terms of data processing, Python combines a large number of third-party libraries, making it have powerful functions and simple code. Its ability to provide data

visualization is no less than C language. Python+Selenium technology provides an effective tool for scientific research and local government supervision based on quantitative analysis of big data.

References

[1]Liu, B., Crossman, R., Zhai Y.H. (2003) Mining Data Records in Web Pages. In: Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data. Washington, D.C. pp. 601-606. https://doi.org/10.1145/956750.956826.

[2]Zhai Y.H., Liu, B. (2005) Web Data Extraction Based on Partial Tree Alignment. In: Proceedings of the 14th international conference on World Wide Web. Chiba Japan. pp. 76-85.https://doi.org/10.1145/1060745.1060761.

[3]Mei, X., Cheng, X.Q., Guo Y., Zhang, G., Ding, G.D. (2008) Fully Automatic Wrapper Generation for Web Information Extraction. Journal of Chinese Information Processing. Commun., 01:22-29. https://kns.cnki.net/kcms2/article/abstract?v=3uoqIhG8C44YLTlOAiTRKgchrJ08w1e7VSL-HJEdEx3uI1gV1mfPRB5MledPdJFvvY5IwEhBvnppt0uNVoyy7K7oCcU4cLYQ&uniplatform=NZKPT.

[4]Wang, J.Z., Lai, W.J. (2017) Data Acquisition and Development of Online Analysis Tools Based on Real Estate Transaction Websites. Computer Technology and Development. Commun., 27:154-159. https://doi.org/10.3969/j.issn.1673-629X.2017.05.032.

[5]Nie, L.J., Fang, Z.W., Li, R.X. (2022) Implementation of Web Crawler Capture Based on Scrapy Framework. Software. Commun., 43:18-20. https://doi.org/10.3969/j.issn.1003-6970.2022.11.006.

[6]Feng, X.J. (2021) Current Situation Analysis and Hot Research on Second-hand Housing Market Based on Python Crawler: A Case Study in Nanjing. Computer & Telecommunication. Commun., 11:65-68. https://doi.org/10.15966/j.cnki.dnydx.2021.11.010.

[7]Zhang, H.Z., Zhang, Q., Dong, Q.H., Zhou, X.P., Wang, B. (2017) Analysis of the price trend of second-hand houses based on the data obtained from housing trading websites under big data: Take Shanghai as an example. Scientific and Technological Innovation. Commun., 21:142-144. https://kns.cnki.net/kcms2/article/abstract?v=3uoqIhG8C44YLTlOAiTRKibYlV5Vjs7iAEhECQAQ9aTiC5BjCgn0RpjkbkauycMRuJ5zuKr_anpnb-CVeje1BPpmTEbKKm1G&uniplatform=NZKPT.

[8]Wang, F.C., Qi, P. (2020) Information crawling and data analysis of second-hand houses in Hefei based on Python. Journal of Jiujiang University(Natural Science Edition). Commun., 35:49-51+80. https://doi.org/10.19717/j.cnki.jjun.2020.03.013.

[9]Zheng, H.P. (2021) Application of a Selenium data automatic acquisition technology. Journal of Suihua University. Commun.,41:157-160.https://kns.cnki.net/kcms2/article/abstract?v=3uoqIhG8C44YLTlOAiTRKibYlV5Vjs7iy_Rpms2pqwbFRRUtoUImHU97aJbRbuAEvppjuxOn2N0DKRD5rHGwyjeY1s1CdhDy&uniplatform=NZKPT.

[10]Luo, B.W. (2019) Data visualization based on Python. Information Recording Materials. Commun., 20:72-74. https://doi.org/10.16009/j.cnki.cn13-1295/tq.2019.12.042.

[11]Dai, Y., Zheng, C.X. (2021) Crawling and analyzing Nanjing second hand house data with Python. Computer Era. Commun., 01:37-40+45. https://doi.org/10.16644/j.cnki.cn33-1094/tp.2021.01.009.