Property Tycoon

Team members：马志宇、刘润东、潘乔、吴铮、汪啸麟

Introduction

1. Project Motivation

In business studies, getting data is central to some problems. With the advent of the era of big data, the efficient data acquisition and analysis has become an indispensable ability. Python is a productive and convenient operating programming language, and we expect to use Python crawlers and Python data visualization to obtain data and perform preliminary data analysis and presentation. We take the real estate industry as the subject of our research. From the leading internal website, such as lianjia.com, fang.com, we are able to obtain a period of time and a area of place of housing data. After the data export and visualization, we can provide reference for investors, industry researchers, or those who have the intention to purchase.

2. Project Tools

Libraries for web crawlers：requests，beautifulsoup　　Libraries for data visualization：ggplot（r-ggplot2），matplotlib

Data collection

1. Preparation

Start the preparation before you crawl, and import the library files you need to use. Here are two main requests and BeautifulSoup. The Time library is responsible for setting the rest time per scraping .We'll introduce pandas and numpy when we use them .

```
import request
import time
from bs4 import BeautifulSoup
import pandas as pd
import numpy as np
```

2.url

A Uniform Resource Locator (URL)， is a reference to a web resource that specifies its location on a computer network and a mechanism for retrieving it. URLs occur most commonly to reference web pages (http), but are also used for file transfer (ftp),

email (mailto), database access (JDBC), and many other applications

url='http://bj.lianjia.com/ershoufang/pg'

Here we use ‘url’ as a name of our variable.

3.headers

http protocol is the abbreviation for “Hypertext Transfer Protocol”， and it is used throughout the world wide web, and almost all of what you see in your browser is transferred through http Protocol .

```
headers={'User-Agent':'Mozilla/5.0 (compatible; MSIE 10.0; Windows NT 6.1; WOW64; Trident/6.0; SLCC2;.NET CLR 2.0.50727; .NET CLR 3.5.30729; .NET CLR 3.0.30729; InfoPath.3; .NET4.0C; .NET4.0E)',
        'Accept':'image/webp,image/*,*/*;q=0.8',
        'Referer':'http://bj.lianjia.com/ershoufang/pg9/',
        'Accept-Encoding':'gzip, deflate',
        'Connection':'keep-alive'      }
```

3.1 Some web sites often judged by the UA to the operating system, different transmission different browsers and different pages, it may cause some normal page cannot be displayed in a browser, but UA can bypass the detection by disguise.

3.2Accept

Specifies the type of content that the client can receive, and the order in the content type indicates the order in which the client receives the order.

3.3 Referer

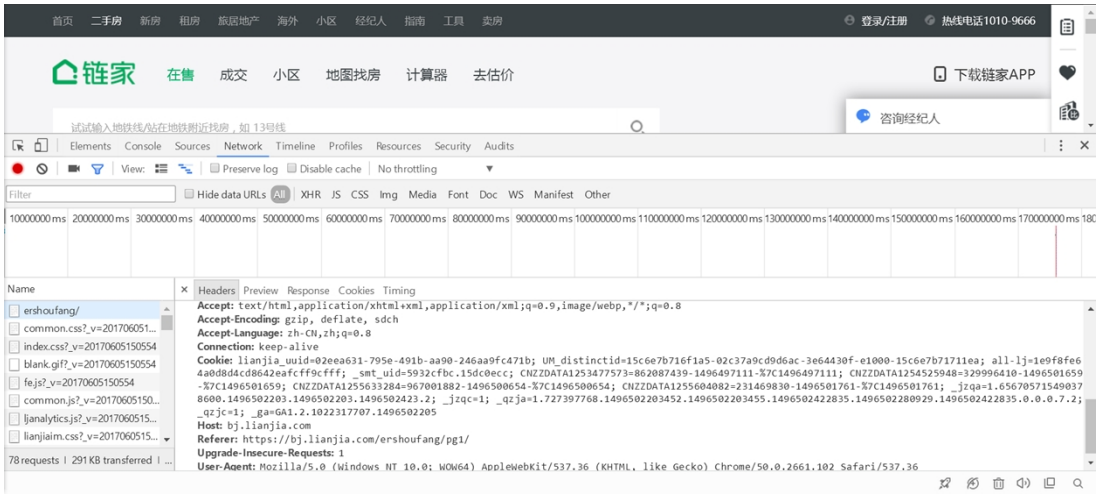
Contains a URL where the user starts from the page represented by the URL and accesses the page of the current request

3.4 Accept-Encoding

Specifies that the web server that the client browser can support returns the content compression encoding type. Indicates that the server is allowed to compress before sending the output to the client to conserve bandwidth. Here is the returned compression format that the client browser can support.

3.5 Connection

Indicates whether a persistent connection is required



4 .Use the for loop to generate 1-100 of the numbers, convert the format to the previous URL fixed part, and spell the URL you want to scrap. Here we set 0.5 pages per two seconds apart. The scrapped pages are saved in html.

```
 for i in range(1,100):
     if i == 1:
         i=str(i)
         html=requests.get(url=url+i+'/',headers=headers).content
     else:
         i=str(i)
         html2=requests.get(url=url+i+'/',headers=headers).content
         html=html+html2
     time.sleep(0.5)
```

Time.sleep

Defer the specified time to run the thread. Unit is' per second'

5.**Parse pages and extract information**

When the page is finished, it can not read and extract data directly, but also need to parse the page. We use BeautifulSoup to parse pages. Become what we see in the browser's source code view

Beautiful Soup is a Python library for pulling data out of HTML and XML files. It works with your favorite parser to provide idiomatic ways of navigating, searching,

and modifying the parse tree.

lj=BeautifulSoup(html,**'html.parser'**)

After you have finished parsing the page, you can extract the key information from the page. Here we are on the housing price, housing information and attention of the three parts extracted. That is the three part of the red bid.

The class=priceInfo part of the page div tag is extracted and the for loop is used to store the total data in each of the listings in tp

Here we use the parameter 'attrs 'in the function' find_all()'to define a dictionary parameter to search for tags that contain special attributes; and use 'a.span.string' , ' b.get_text ' to get data and text information.

```
price=lj.find_all('div',attrs={'class':'priceInfo'})
tp=[]
for a in price:
    totalPrice=a.span.string
    tp.append(totalPrice)
```

Extraction of housing information and attention methods, and extraction of housing prices similar to the method, the following is the specific code, housing information stored in Hi, attention is stored in the fi.

```
#提取房源信息
houseInfo=lj.find_all('div',attrs={'class':'houseInfo'})
hi=[]
for b in houseInfo:
    house=b.get_text()
    hi.append(house)

#提取房源关注度
followInfo=lj.find_all('div',attrs={'class':'followInfo'})
fi=[]
for c in followInfo:
    follow=c.get_text()
    fi.append(follow)
```

6.Create data tables and clean data

Import the pandas library, collect the listings, the total price and the attention before, and then generate the data table. Easy to analyze later.

```
house=pd.DataFrame({'totalprice':tp,'houseinfo':hi,'followinfo':fi})
```

Before the analysis, the data should be extracted and cleaned. Such as housing information, in the table, each housing name of the District, Huxing, area, direction and other information in a field, can not be used directly. The operation needs to be done first. The rule here is obvious, each information is based on the vertical segmentation, so we only need to be disaggregated to the vertical line, the types of housing information has become a separate field.。

```
houseinfo_split=pd.DataFrame((x.split('|') for x in house.houseinfo),\
                             index=house.index,\
                             columns=['小区','户型','mianji','朝向','装修','电梯'])

house=pd.merge(house,houseinfo_split,right_index=True,left_index=True)

followinfo_split=pd.DataFrame((x.split('/') for x in house.followinfo),\
                             index=house.index,\
                             columns=['guanzhu','热度','日期'])
house=pd.merge(house,followinfo_split,right_index=True,left_index=True)

print(house)
```

Export data to CSV files and perform further analysis in excel

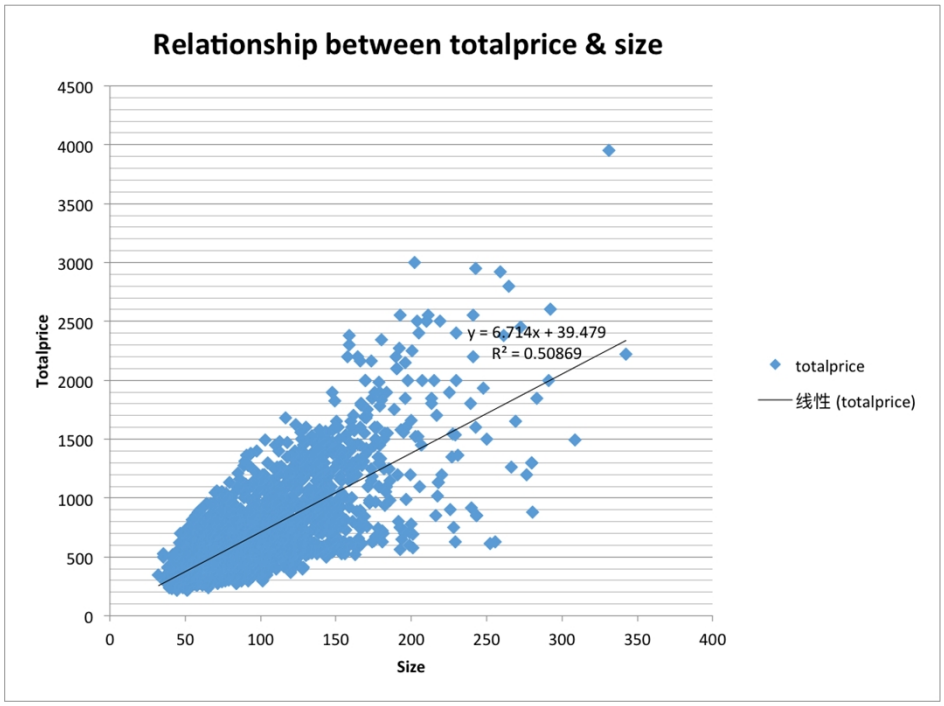
```
house.to_csv('final.csv')
```

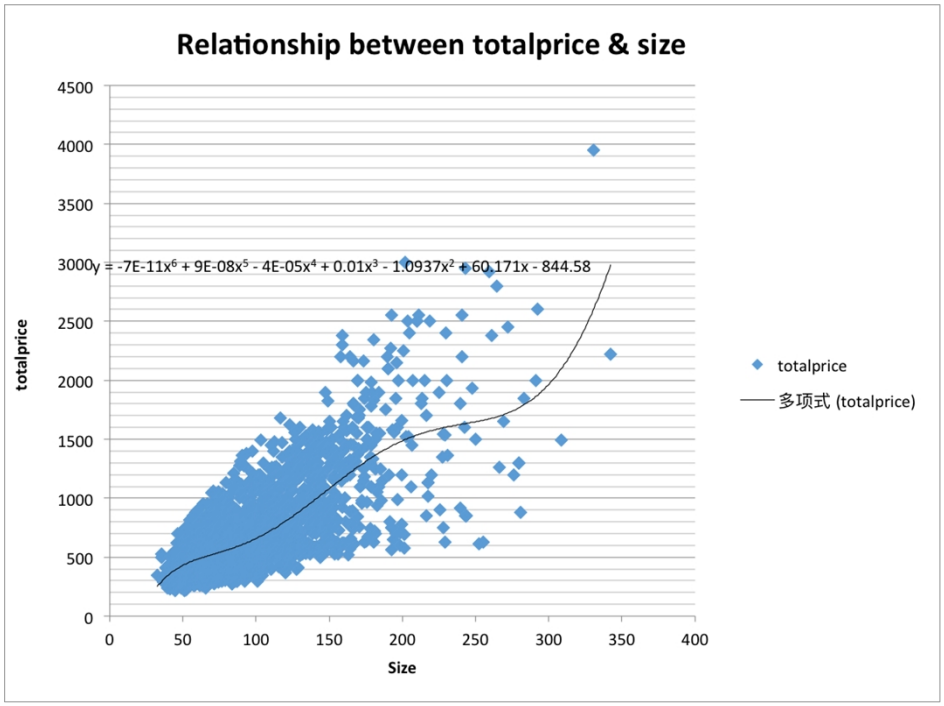
Data analysis

		followinf	house totalprice	小区	户型	Size	朝向	Decoratior	Elevator	Attention c	Heat degree	日期
1												
2	$	51人关注	万泉新 3950	万泉新新家[	4室2厅	330.86	南 北	精装	有电梯	51	35	1个月以前发布
3	641.00	52人关注	国奥村 3000	国奥村	3室2厅	202.09	南 北	简装	有电梯	52	57	1个月以前发布
2365	427.00	180人关注	行宫园 240	行宫园三里	2室1厅	65.53	南 北	其他	无电梯	180	63	1个月以前发布
2366	506.00	97人关注	古城南 235	古城南路	2室1厅	41.08	南	其他	无电梯	97	12	5天以前发布
2367	1162.00	43人关注	艺苑西 235	艺苑西街9号	1室1厅	39.21	南 北	简装	无电梯	43	14	14天以前发布
2368	198.00	193人关注	北环里 230	北环里小区	1室1厅	41.28	东	精装	无电梯	193	92	1个月以前发布
2369	78.00	85人关注	建设巷 226	建设巷小区	2室1厅	52.04	南 北	简装	无电梯	85	25	15天以前发布
2370	988.00	255人关注	富强东 220	富强东里	2室1厅	44.58	南	其他		255	53	1个月以前发布
2371	2266.00	54人关注	昊文温 219	昊文温泉家[	2室1厅	51	东北	毛坯	无电梯	54	29	2个月以前发布

From the housing data collected in the excel file, we are able to do some data analyses about the house price in Beijing.

1.Relationship between totalprice & size

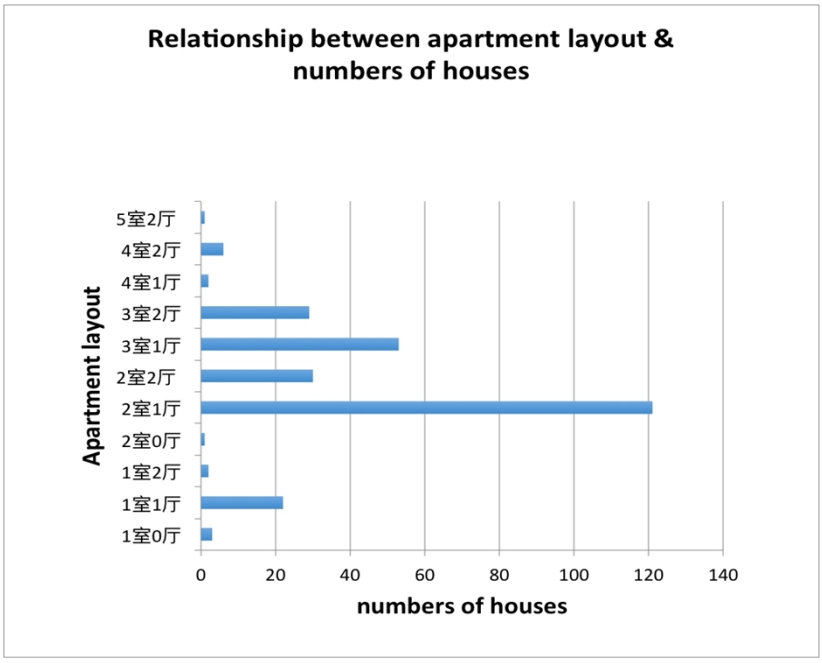




The picture above shows that the variable cost of the house is ￥67,140 per square meter and the fix cost is ￥394,790. Also, we employ the Polynomial Fitting Model, as this model has the minimum error. We can see the price of unit area is changing at different size. It is quite hard for young people to buy an apartment in Beijing, as they have to spend twenty years of their salaries for the smallest apartment, which is only 50 square meters large. This statement is based on the assumption that these young

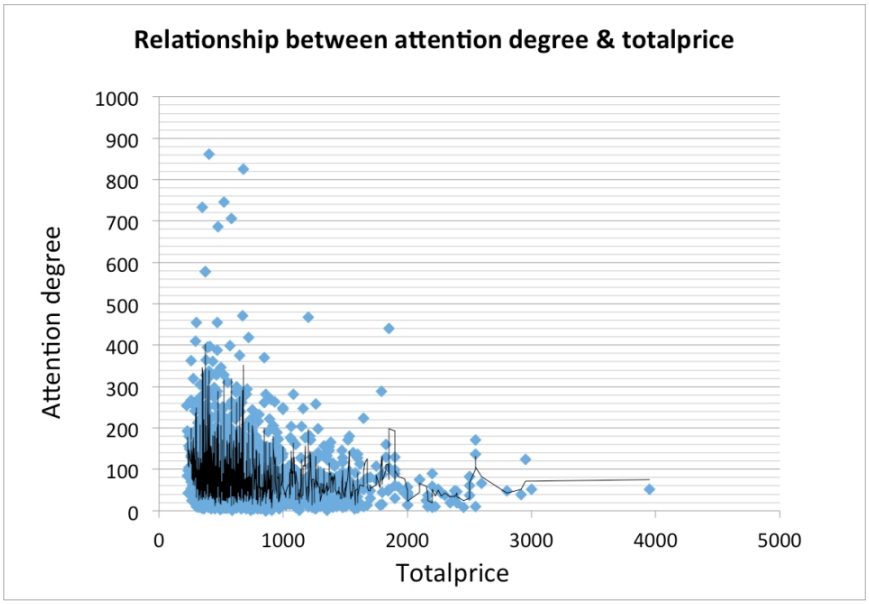
people can earn ￥200,000 per year and they don't have to spend money on their eating or drinking, which is quite unrealistic. It reveals the great difficulty in owning a house in Beijing.

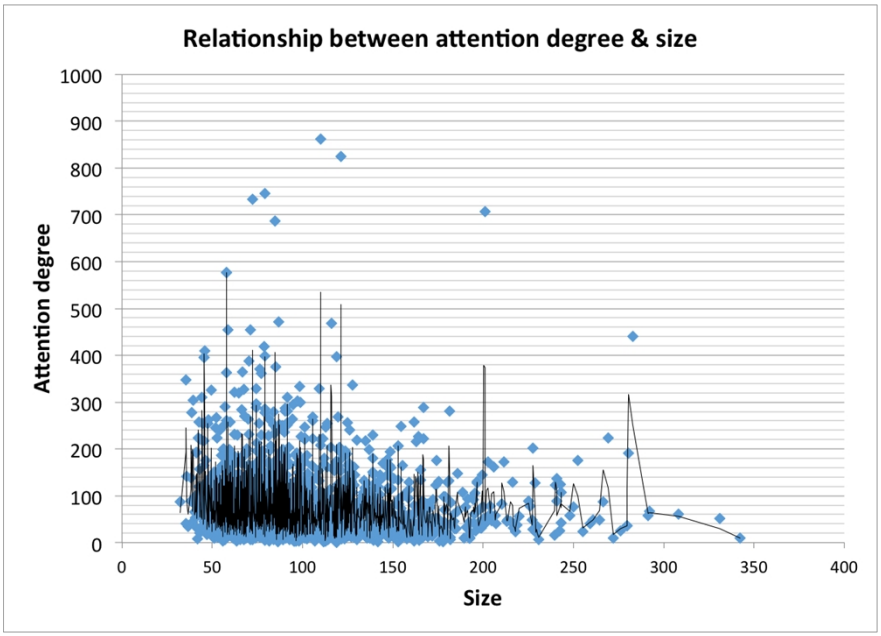
2. Relationship between apartment layout & numbers of houses



We can easily find out that an apartment with two bedrooms and a living room is most popular. The second popular apartment layout is the one with three bedrooms and a living room. It can be explained by the common three-member-family pattern in China, where the child lives in one bedroom and the parents live in the other. In the apartment with three bedrooms and a living room, if there comes a guest, another bedroom can be prepared for him.

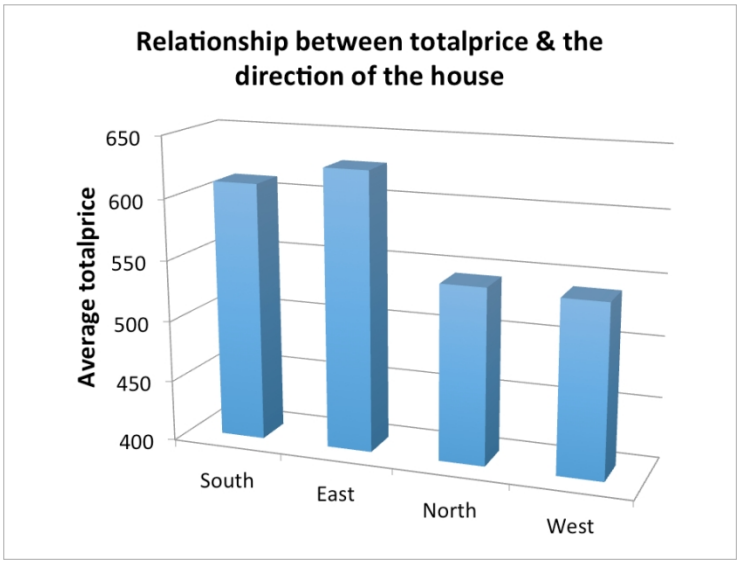
3.Relationship between attention degree & total price, attention degree & size





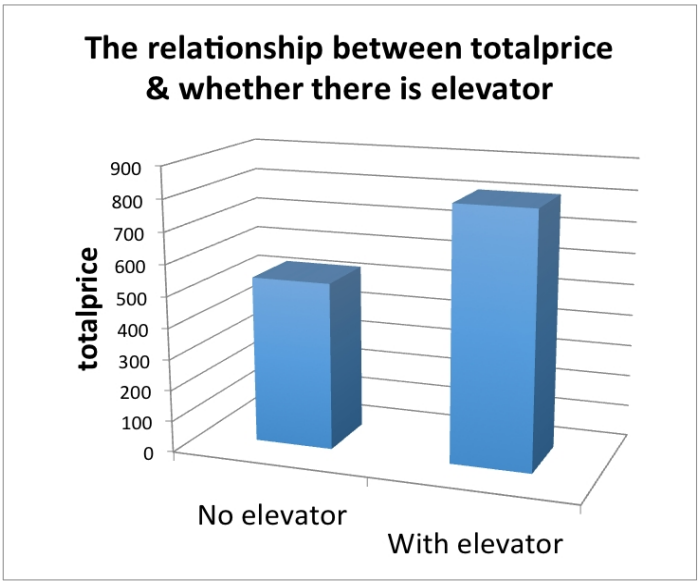
The First chart shows that the attention degree decreases generally when the total price increases. The latter one shows that people are most interested in apartment around 75 square meters. Small apartment is the most popular choice when people confront with the incredible high living price in Beijing.

4.Relationship between total price & direction of the house



The house's cost rank (from high to low) of the four directions of the apartment is east, south, north, west. People are willing to pay around 15% more for a apartment facing towards south or east, because these two orientation can provide the apartment with more sunlight.

5. Relationship between total price & whether there is elevator



It is surprising that people are willing to pay 51% more for an apartment with elevator. There are many tall buildings in Beijing and people living there are so willing to pay a large amount of money for the convenience of no need for climbing

stairs.

Difficulties

1. learning of different libraries.

During the process of writing our program, we learned several useful libraries of python which can be used in different fields of data analysis. Such as requests, beautifulsoup4, pandas and numpy. We face difficulties in learning how to use it and convert the form of data into the attribute of function in these libraries.

2. The locked IP addresses of our computers

After many debugs, we are very upset to find that the website of lianjia.com have already locked our IP address. During the process of this program, we have changed several laptops to run the code. Finally, we got a csv file with over 2,000 pieces of data about different variables of real estate in Beijing.

Advantages of our project

Since we can collect all the data on the Lianjia website, up to 2372 different apartments, with this code in a few minutes, it saves quite a lot of energy on collecting data. Otherwise, people have to collect thousands of data by hand, which is quite a heavy work.

Also, our web scrapping code is quite flexible. We can collect whatever type of data on the website as we like. In our project, we collect data such as the attention degree, heat degree, total price, size, direction, decoration style, whether equipped with elevator or not. We can even do a little change to collect housing data on other housing websites, such as sh.fang.com, anjuke.com, which can largely enhance the accuracy of our data analysis.

Last but not the least, the web scrapping program is quite a useful tool in our future study. We can collect all kinds of data with the program easily and conveniently.

Disadvantages of our project

The web scrapping program can only collect the data on the main list. It cannot get into every page of each apartment. As a result, we can only collect a few data of a specific apartment, which is not enough for further study.

The data on the website is not absolutely correct. People hoping to purchase an apartment will usually bargain with the seller, and as result the actual price of the apartment will be lower that the price shown on the website.

We are not capable of complex data analysis with python. Although we collect the data and save it in the Excel film, we don't know how to make some of the charts with python. Instead, we have to make them with MS Excel, which is not convenient enough for a project.

Summary

To begin with, we import all kinds of libraries, such as ch, requests, time, pandas, numpy, matplotlib.pyplot and beautifulsoup.

Then, we conduct the website-choosing and data collection. Firstly, we observe the target page structure, especially the format of URL. Through the for cycle, we can switch the URL variable part of the page code, and set the header information in a HTTP request, grasp list page, and save in the HTML. Secondly, through the pandas library, we create, clean, and generate the data tables. Finally, we export it to the CSV file with pandas for later processing.

The last process is the visualization of the collected data. We use MS Excel to draw the relevant chart in order to study the relationships between different type of values.