Property Tycoon

Team members：

马志宇、刘润东、潘乔、吴铮、汪啸麟

CONTENTS


1
PART
Introduction

INTRODUCTION

NO.1 WHY PYTHON

NO.2 REAL ESTATE

NO.3 REFERENCE

1 REQUESTS

2 PANDAS

3 NUMPY

4 BEAUTIFULSOUP4

2 PART

Data collection

1.preparation

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

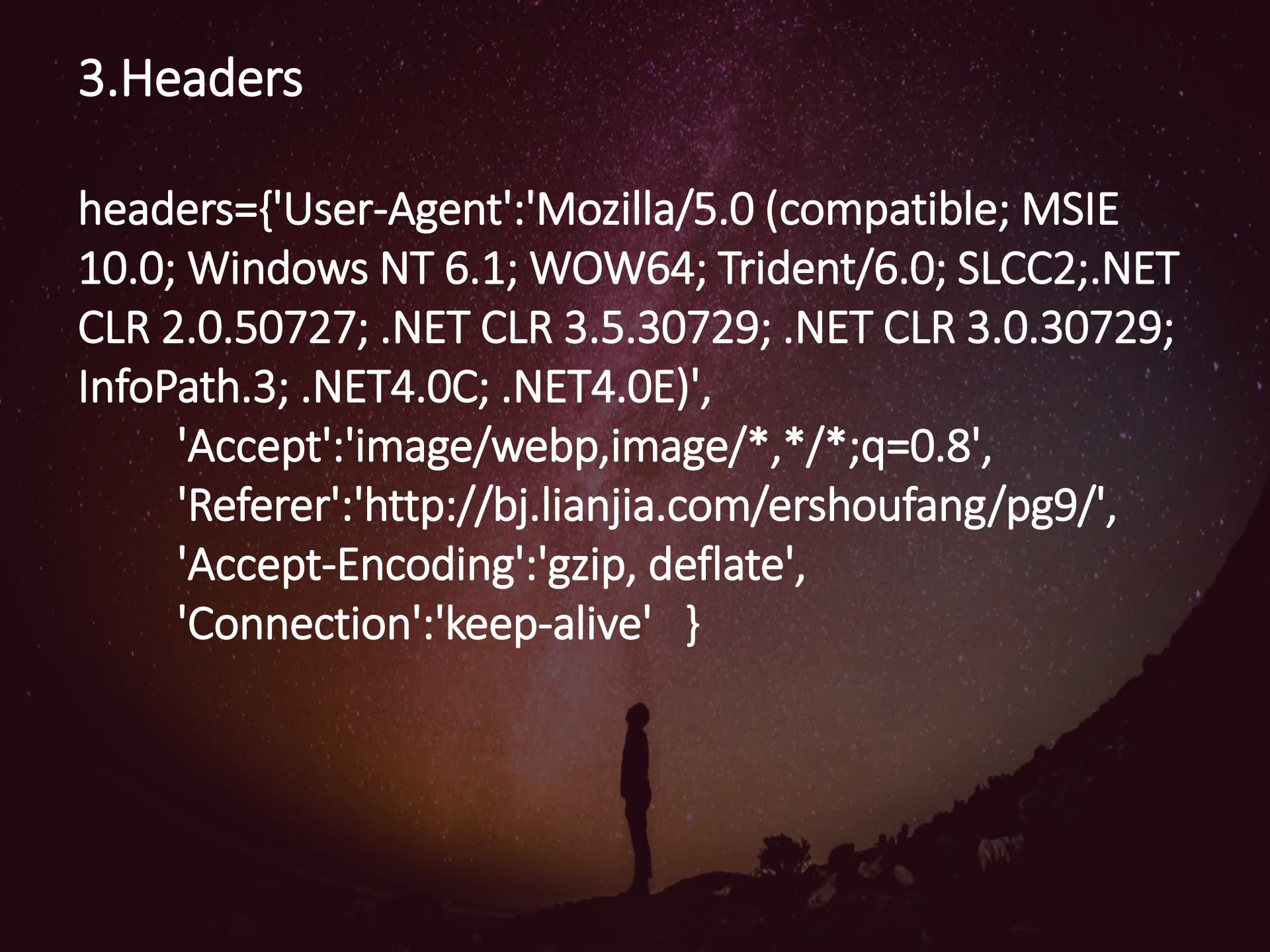
2.url

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

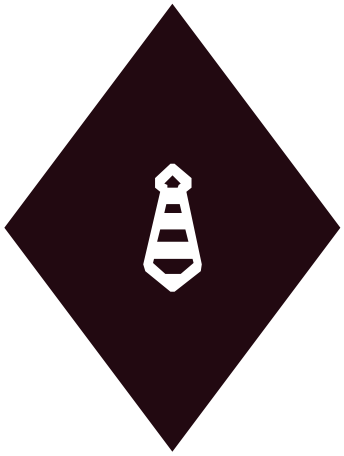
url='http://bj.lianjia.com/ershoufang/pg'Here we use ‘url’ as a name of our variable.

3.Headers

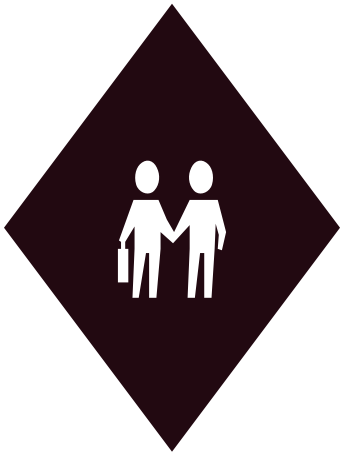
```
headers={'User-Agent':'Mozilla/5.0 (compatible; MSIE 10.0; Windows NT 6.1; WOW64; Trident/6.0; SLCC2;.NET CLR 2.0.50727; .NET CLR 3.5.30729; .NET CLR 3.0.30729; InfoPath.3; .NET4.0C; .NET4.0E)',
        'Accept':'image/webp,image/*,*/*;q=0.8',
        'Referer':'http://bj.lianjia.com/ershoufang/pg9/',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



3.1 Some web sites often judged by the UserAgent to the operating system



3.2Accept



3.3_Referer

3.4 Accept-Encoding

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 get. Here we set 0.5 pages per two seconds apart. Then the 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


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

```
#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})
```

```
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')
```


3
PART
Difficulties

1. learning of different libraries.

requests, beautifulsoup4, pandas and numpy

convert the form of data into the attribute of function

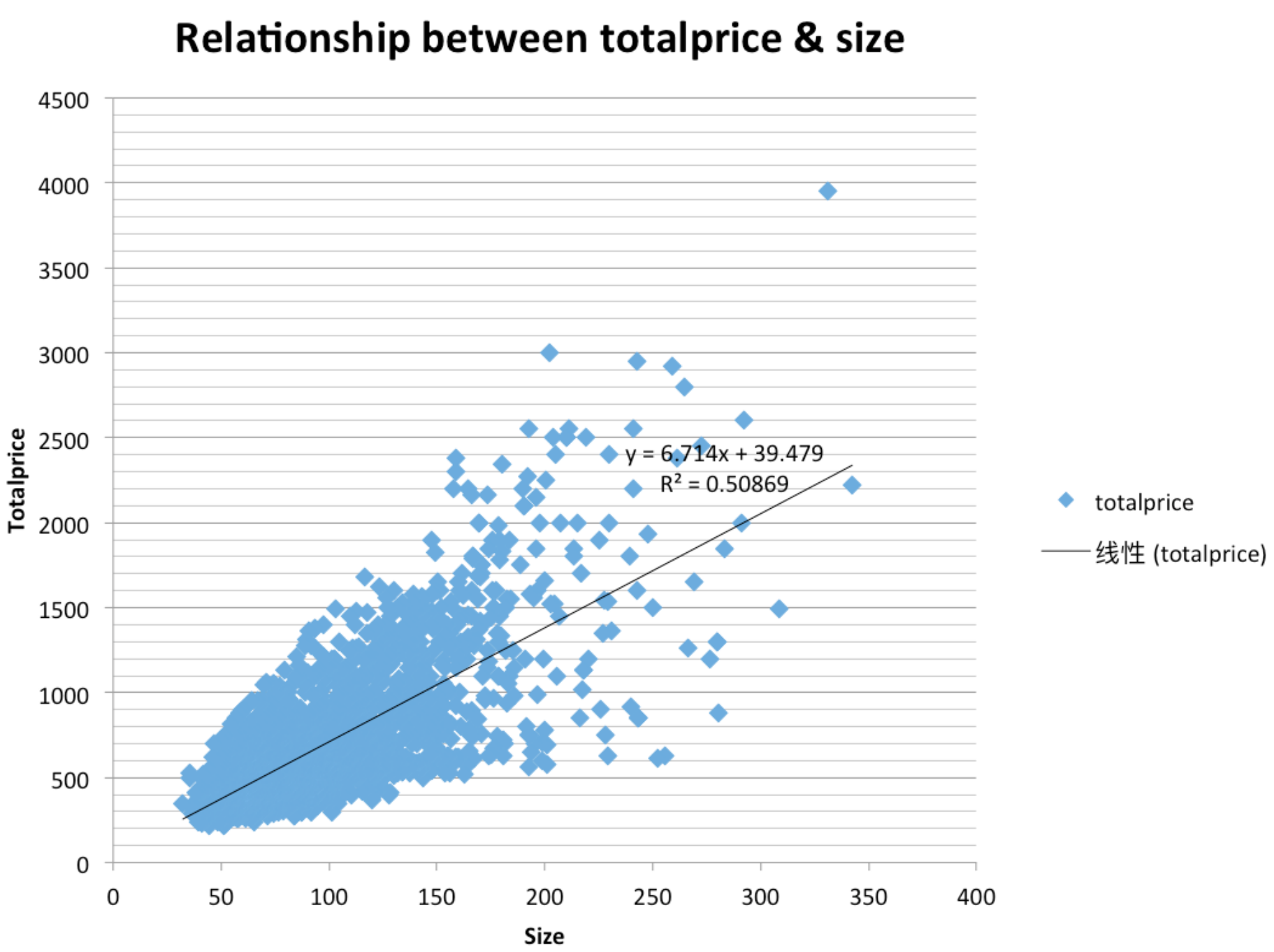
2. The locked IP addresses of our computers

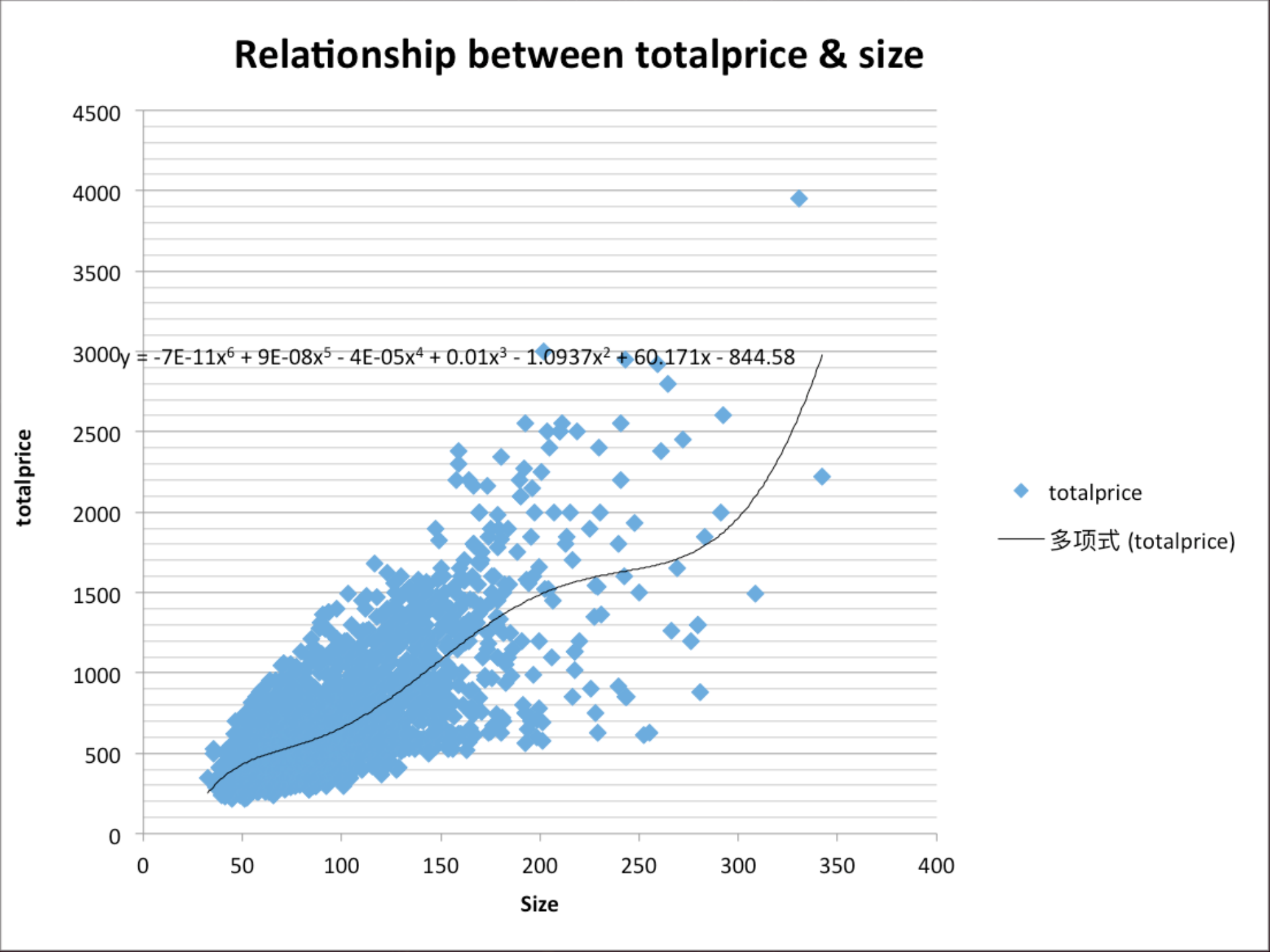
lianjia.com have already locked our IP address

have changed several laptops to run the code

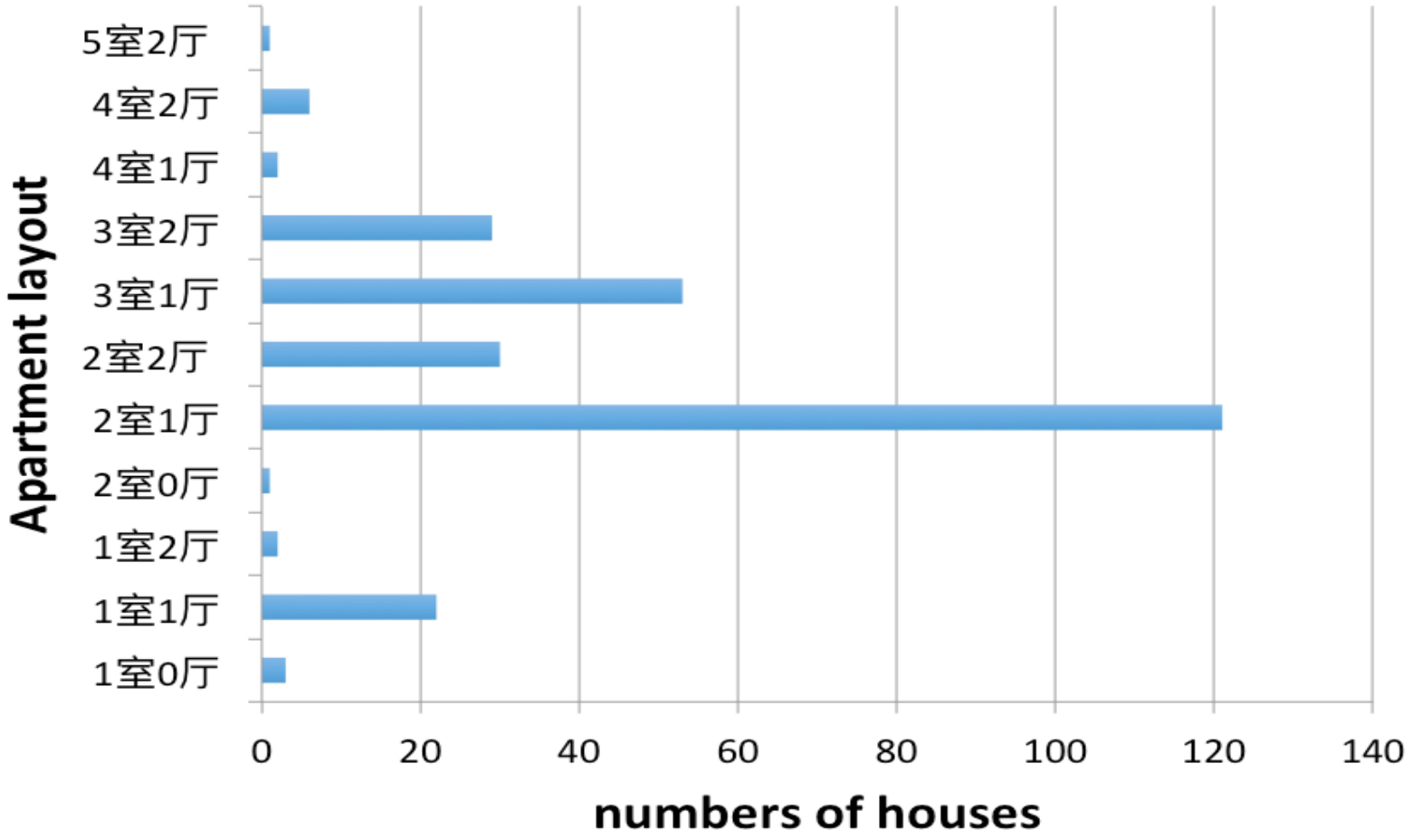
Finally, we got a csv file with over 2,000 pieces of data

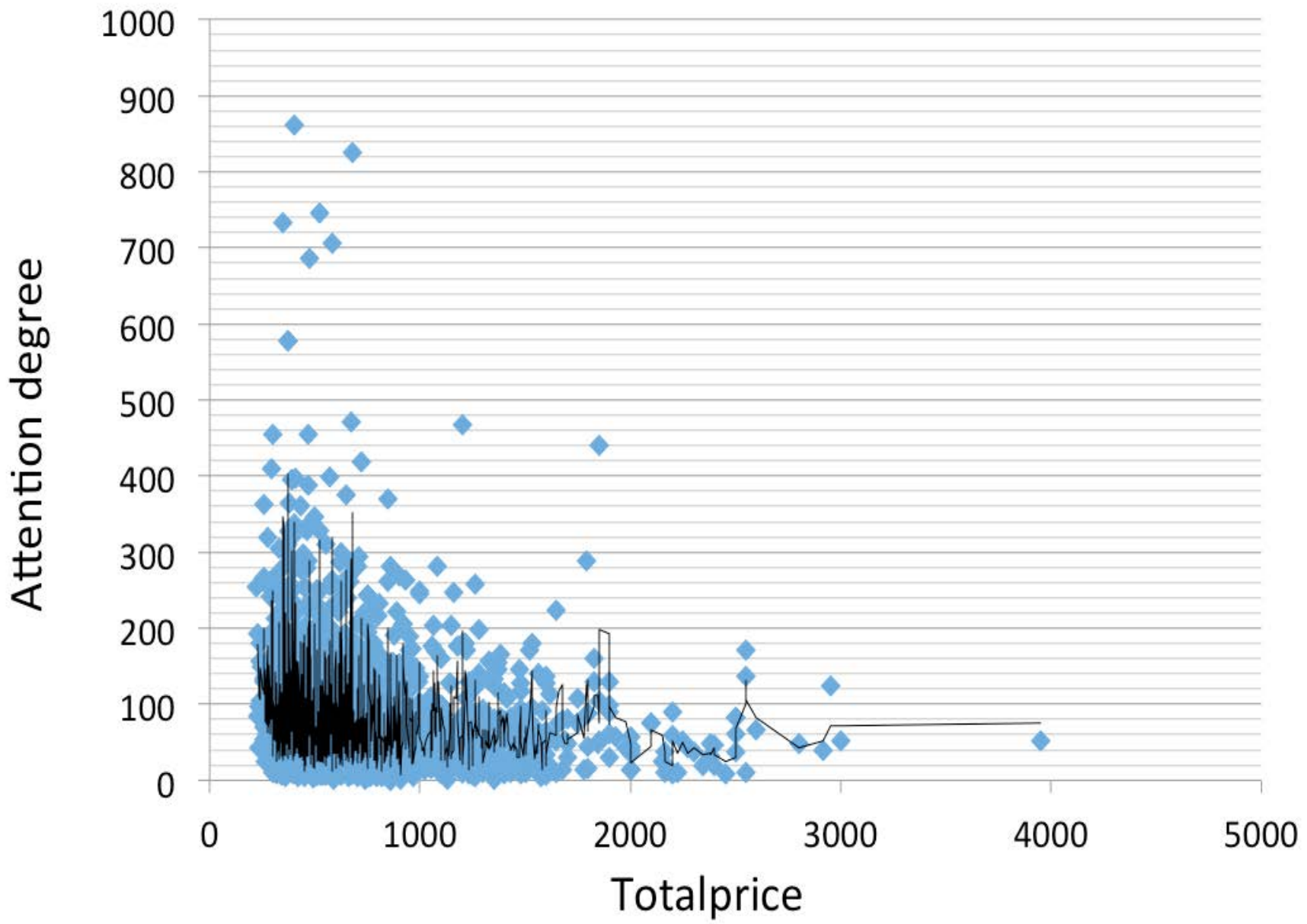

4
PART
Data analysis

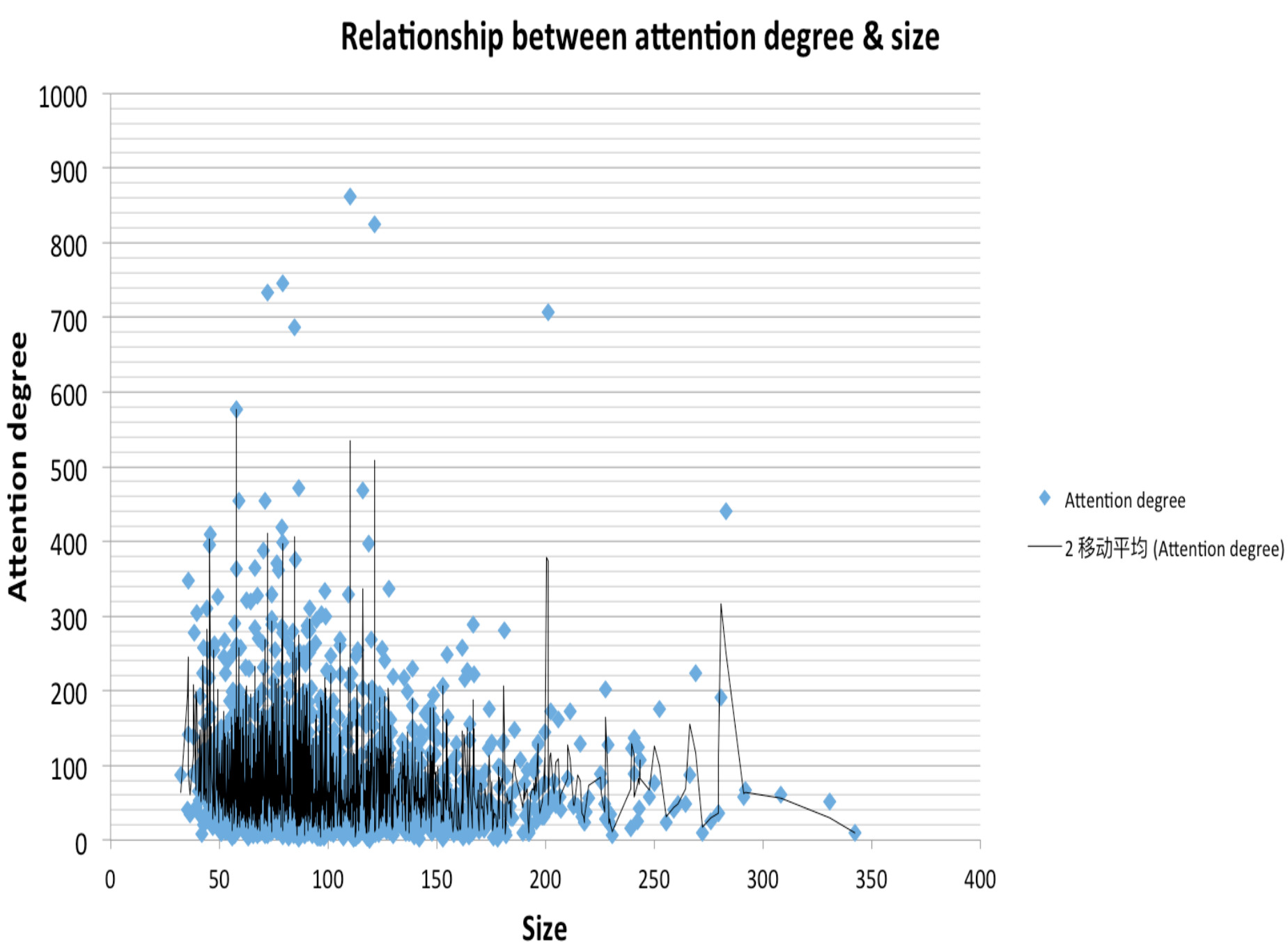

Relationship between totalprice & size
Totalprice
Size
4500
4000
3500
3000
2500
2000
1500
1000
500
0
0
50
100
150
200
250
300
350
400
y = 6.714x + 39.479
R² = 0.50869
totalprice
线性 (totalprice)


Relationship between totalprice & size
y = -7E-11x⁶ + 9E-08x⁵ - 4E-05x⁴ + 0.01x³ - 1.0937x² + 60.171x - 844.58
totalprice
Size
4500
4000
3500
3000
2500
2000
1500
1000
500
0
0
50
100
150
200
250
300
350
400
totalprice
多项式 (totalprice)

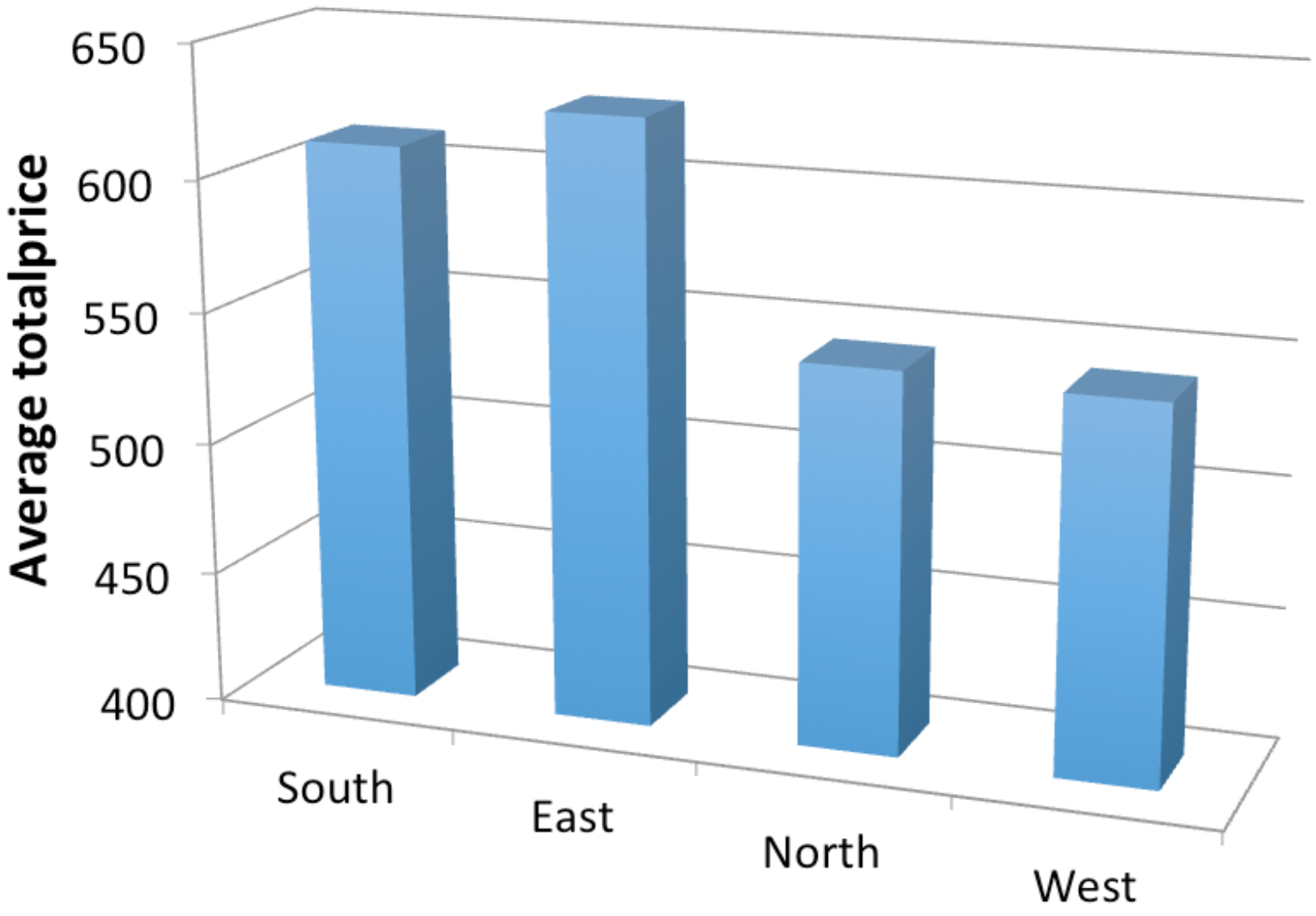
Relationship between apartment layout & numbers of houses



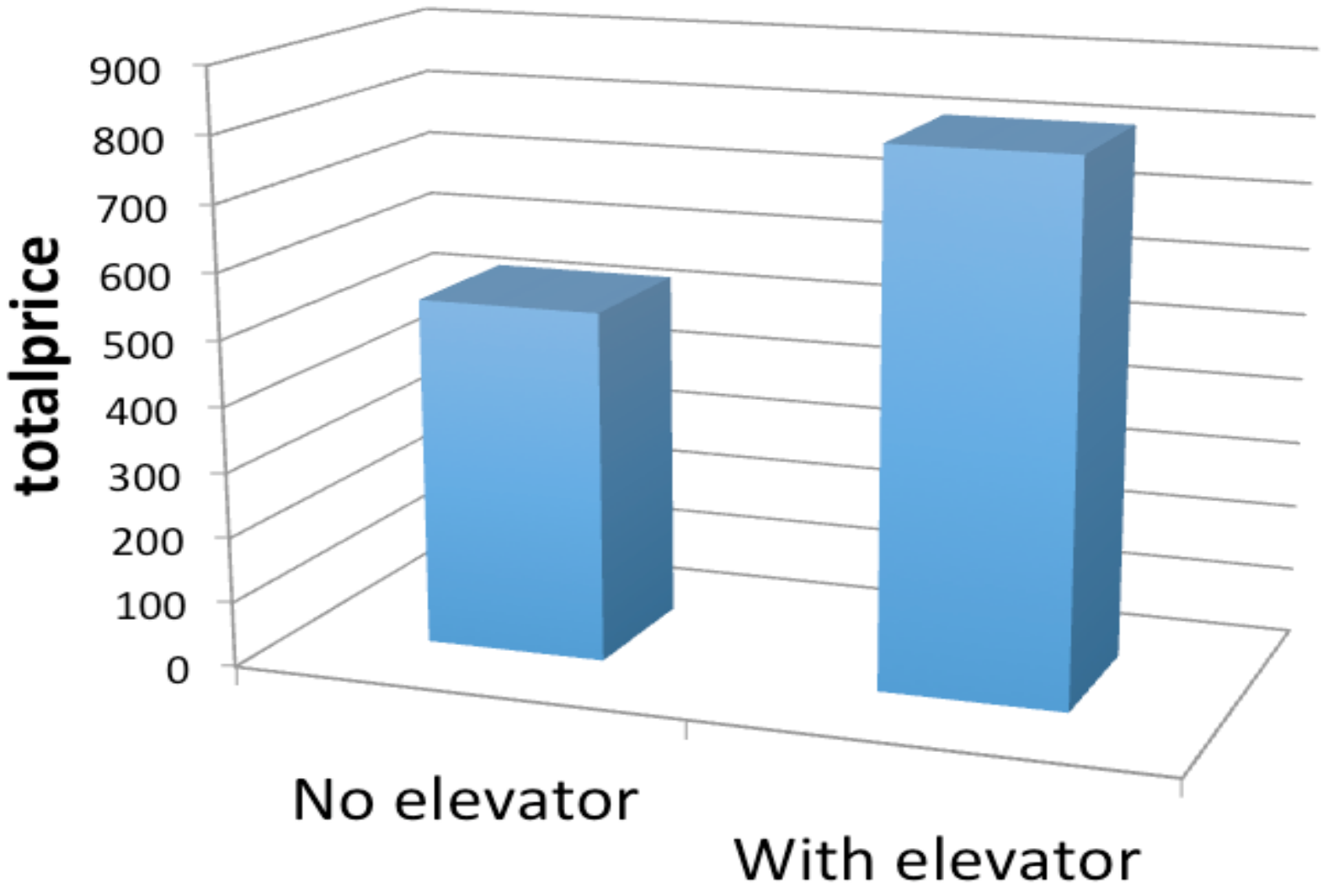

Relationship between attention degree & totalprice
Attention degree
1000
900
800
700
600
500
400
300
200
100
0
0
1000
2000
3000
4000
5000
Totalprice


Relationship between attention degree & size
Attention degree
Size
0
100
200
300
400
500
600
700
800
900
1000
0
50
100
150
200
250
300
350
400
Attention degree
2 移动平均 (Attention degree)

Relationship between totalprice & the direction of the house



The relationship between totalprice & whether there is elevator




5
PART
Advantages
and
disadvantages

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.

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.

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 Matplotlib to draw the relevant chart in order to study the relationships between different type of values.


THANK
YOU