1. hypothesis

Based on the number of Linux Kernel active programmers per month in each time zone, the number of Linux Kernel active programmers for the next month per region is projected

2. Data source

- For purpose：

The data source for this article is the log data of Linux Kernel. Linux is an open source computer operating system kernel. It is a POSIX-compliant UNIX-like operating system written in C. Linux was first developed by Linus Torvalds of Finland as an attempt to provide a free UNIX-like operating system on Intel's x86 architecture. The project began in 1991 with the assistance of some Minix hackers in its early days, and today countless programmers around the world are contributing to the project for free. Therefore, this article can calculate the number of programmers active in the Linux community in each time zone (country) based on the monthly contribution of programmers in each time zone to Linux. This calculates the monthly and active regression curves for each time zone.

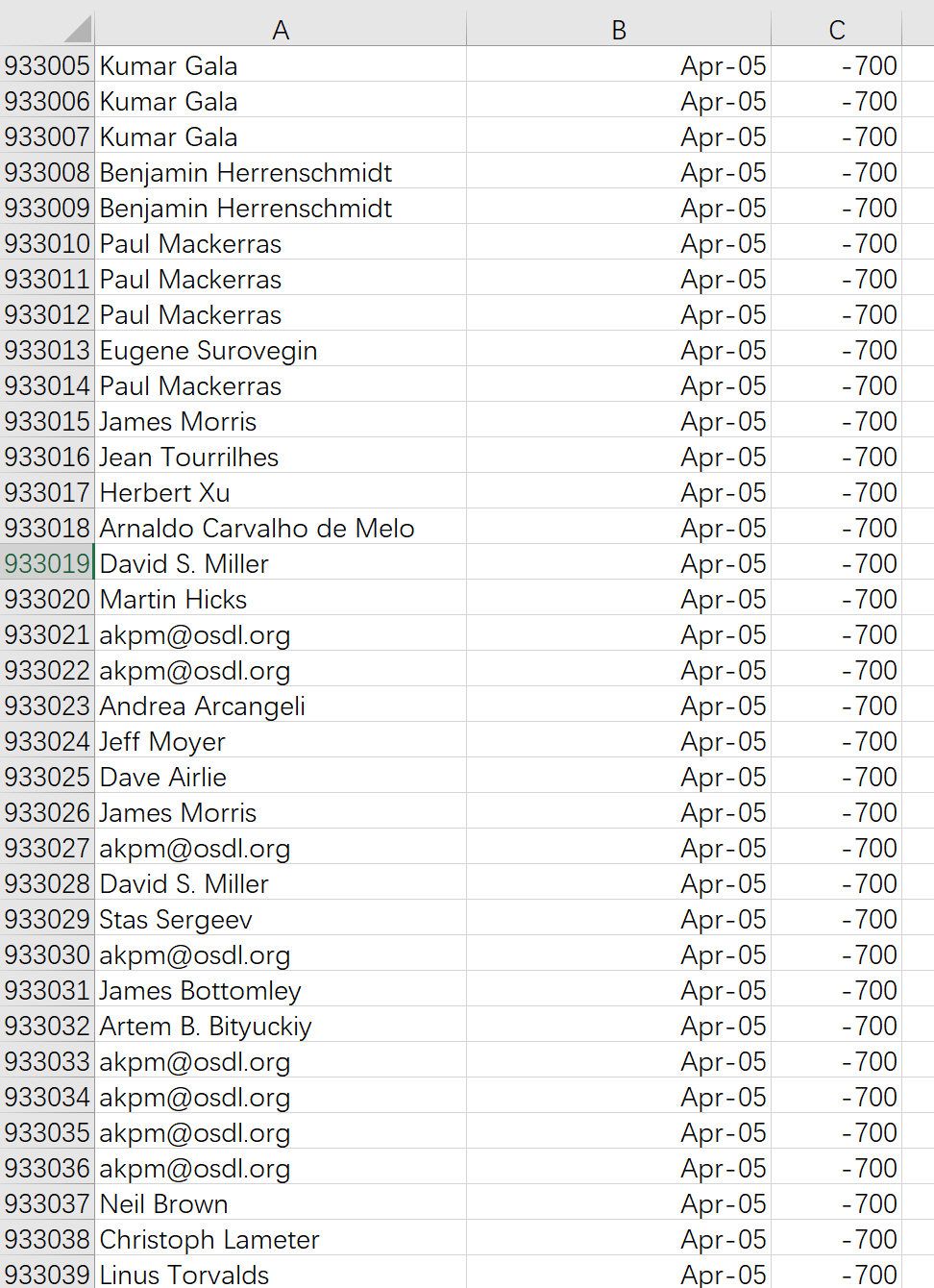
- Step：

1. Pull the Linux Stable warehouse

2. Use

**git log --pretty=format:'"%an","%ad"" --date=format:'"%m,%y","%Z"' >> ~Desktop/date\_data.csv**

The Git log data on The Linux kernel was extracted according to the author's name, change the submission time and the author's time zone as columns, with a total of 933,039 pieces of data. And keep all the data in a CSV file.



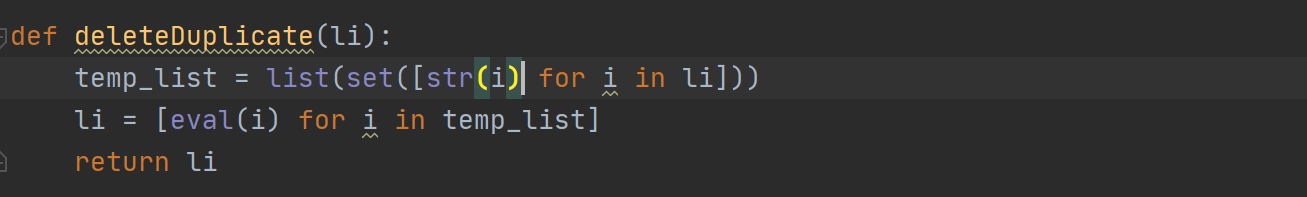
- Data cleaning

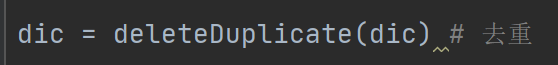
- For purpose;

Data cleansing is the process of re-examining and validating data to remove duplicates, correct errors, and provide data consistency. Since this article calculates the number of active users per month in each time zone, if a programmer commits code multiple times within a month, I add it to the list as an active user. Secondly, the data date source of Linux kernel is Git, which logs in directly from the programmer's computer date, so invalid values may appear. For example, the time value appears 2085-7.Therefore, the corresponding data cleaning work should be carried out.

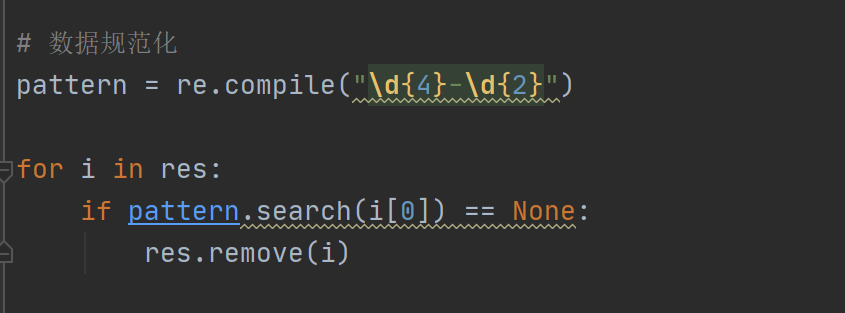
- Step:

1. Data deweighting

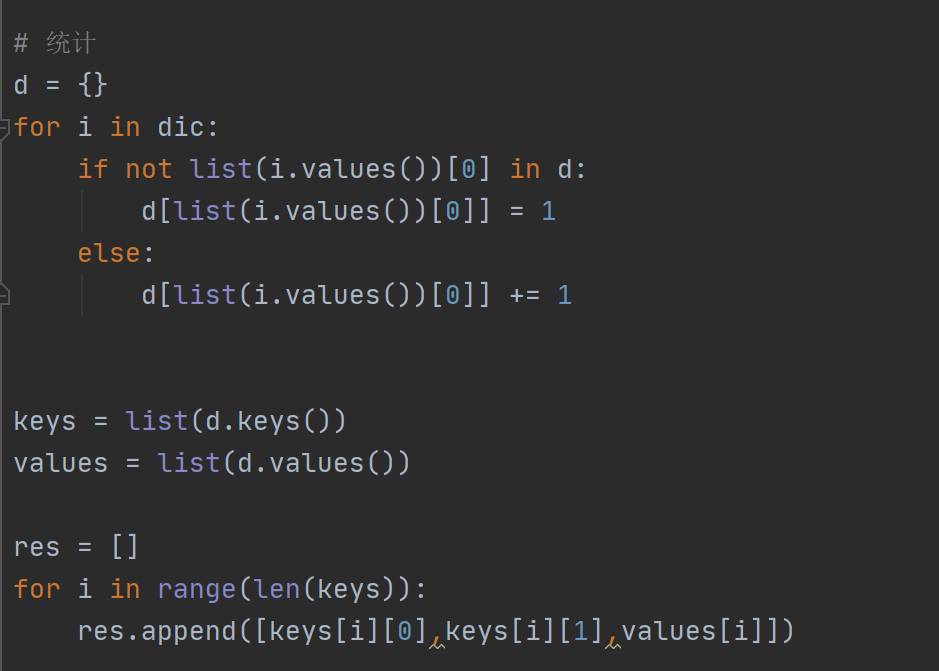




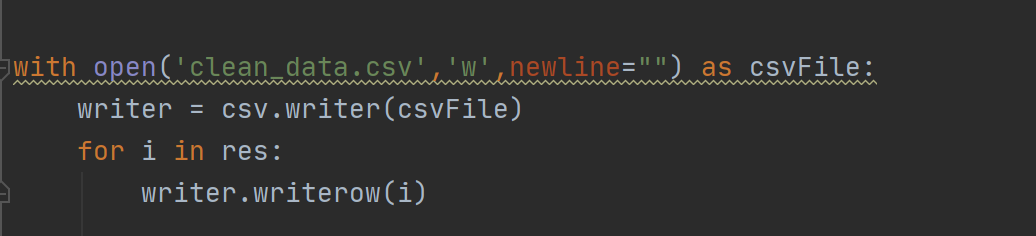
2. Standardize the time data

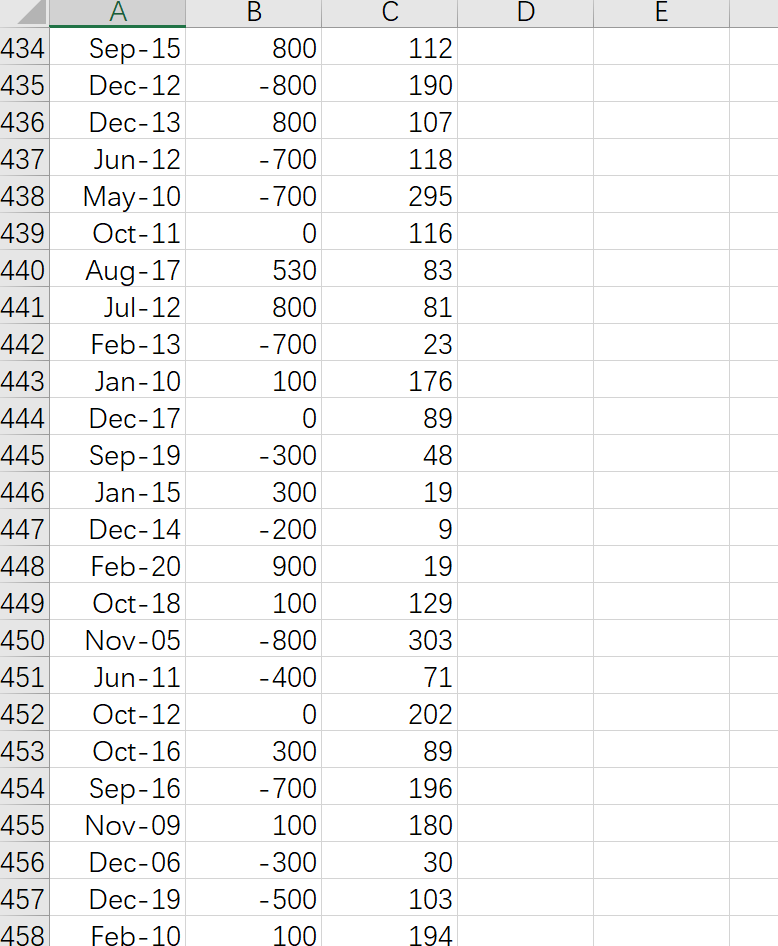


3. Do preliminary statistics -- count the number of active people in each time zone



4. Generate CSV files





After preliminary data cleaning, 933,039 data were sorted into 3118 data

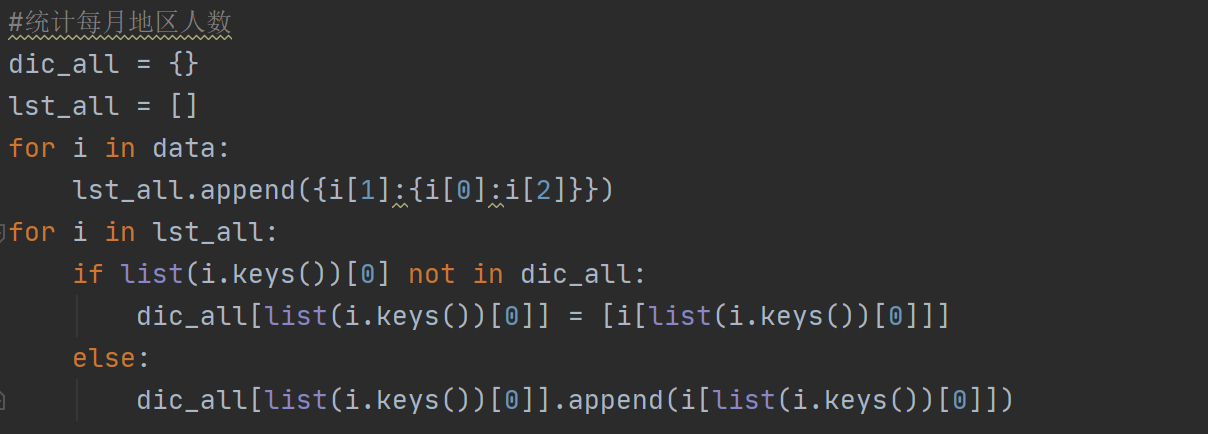
- Data extraction (take +08:00 time zone as an example)

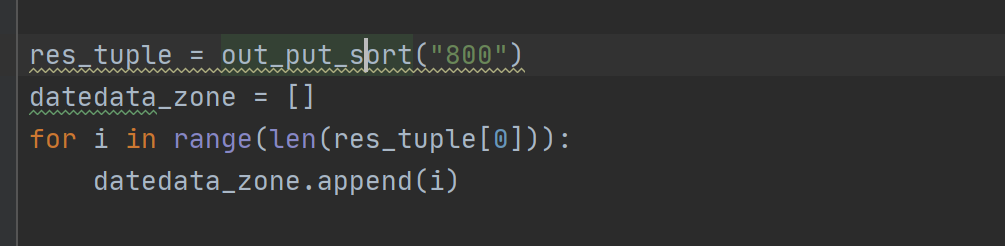
- Purpose:

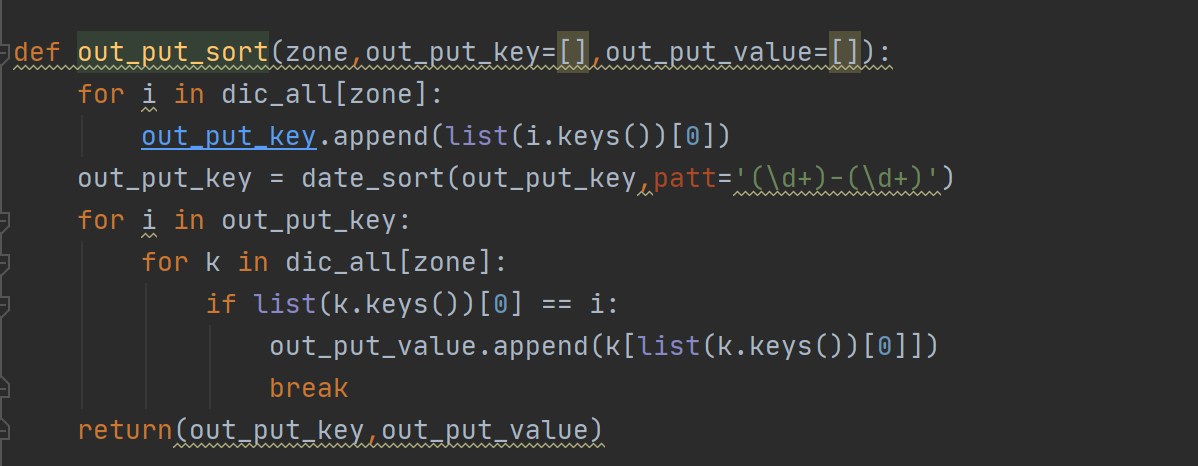
The available 3118 pieces of data were extracted by blocks, and the time period was used as the key word for sorting. In addition, the date is digitized according to the earliest data time as 0 point and every month as the unit.

- Step:

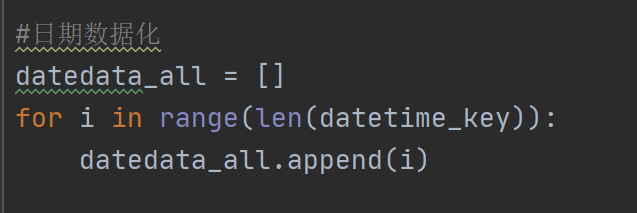
1. Sorting key words based on time period



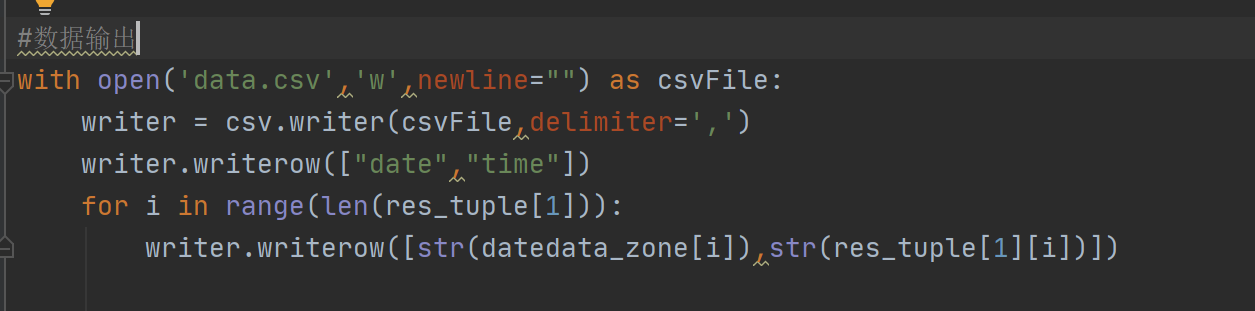


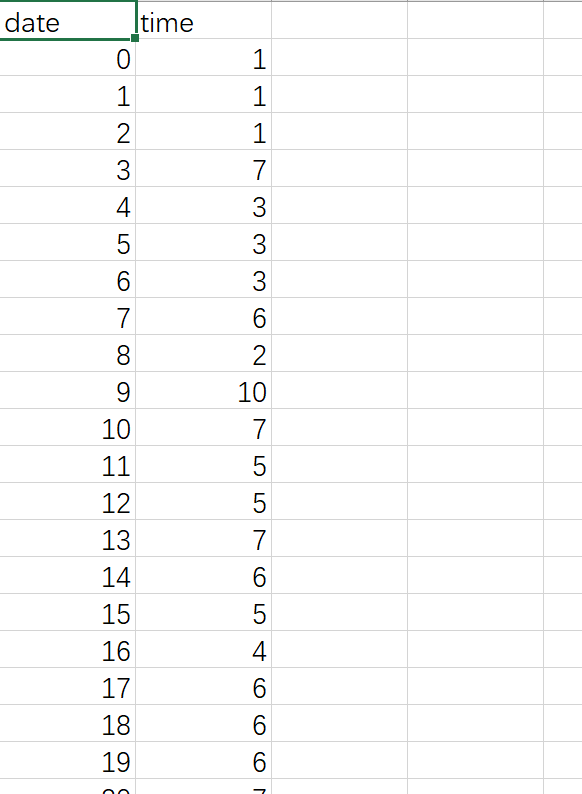


2. Digitize the date



3. Output as CSV file

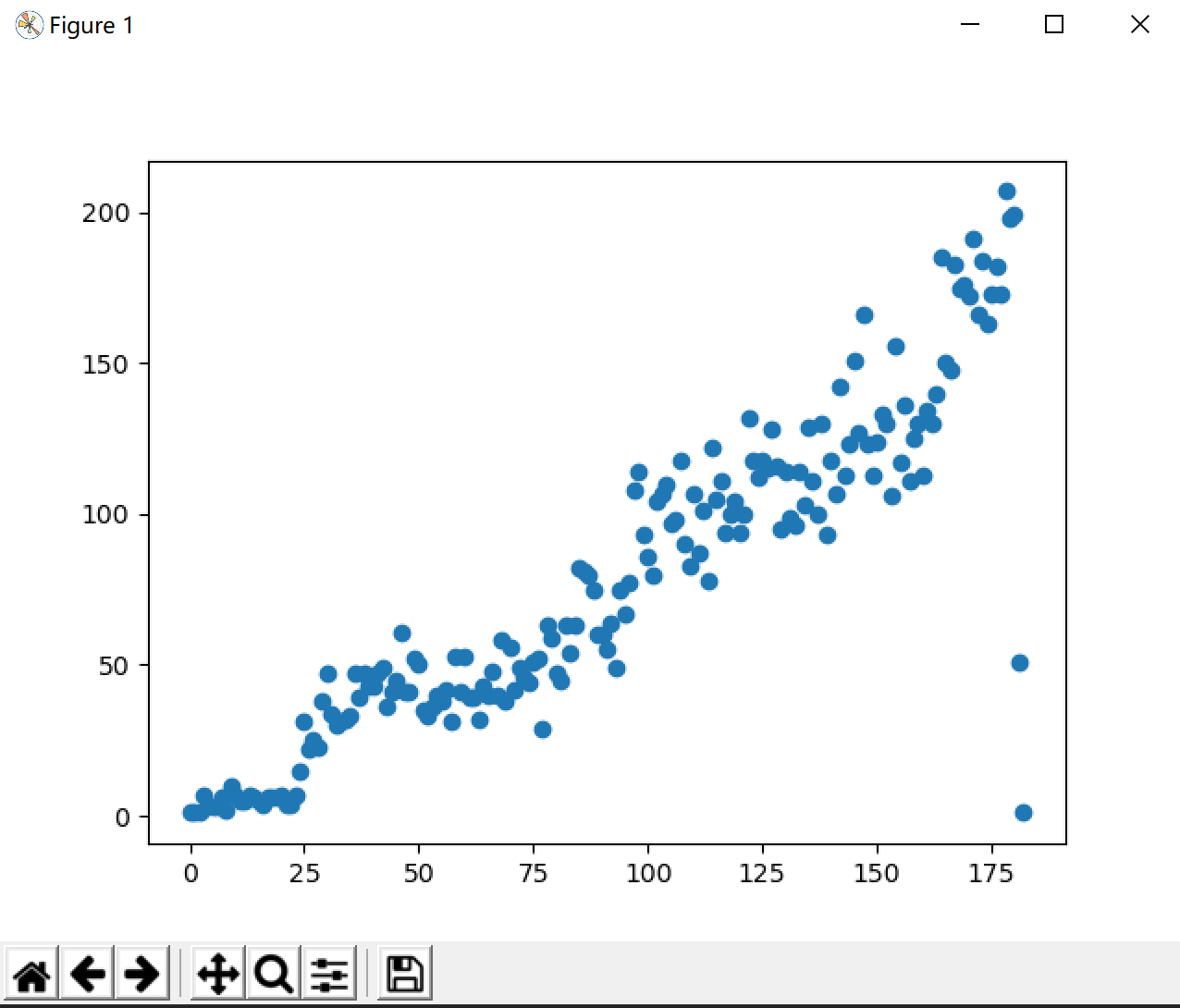




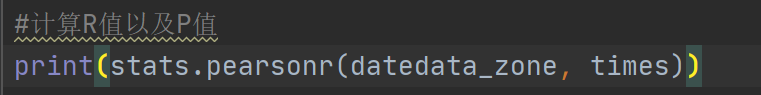
- Model making (take +08:00 time zone as an example)

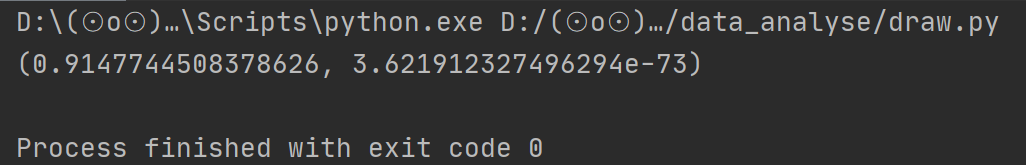
1. Determine the linear relationship

According to the above CSV file for scatter plot drawing, we can get the following scatter plot:



And by calculating R values and P values





R can be obtained as 0.9147744508378626

P was 3.621912327496294 e-73

It can be seen that in the region of +80:00, there is a linear relationship between the monthly factor and the number of active people in the region.

2. The convolutional neural network is used to predict the value

BP(Back Propagation) neural network is a concept proposed by scientists led by Rumelhart and McClelland in 1986. It is a multi-layer feedforward neural network trained in error back propagation algorithm. It is the most widely used neural network.

- Advantage:

BP neural network has been mature in both network theory and performance. Its outstanding advantages are strong nonlinear mapping ability and flexible network structure. The number of middle layers and neurons in each layer of the network can be set arbitrarily according to the specific situation, and its performance varies with the difference of structure.

- Step:

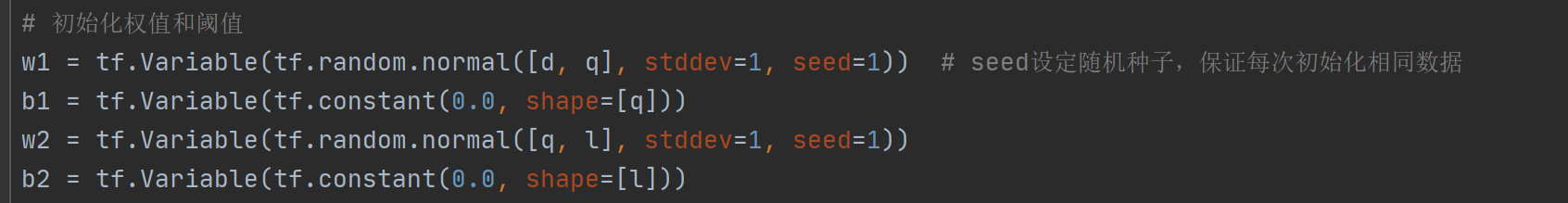
1. Define the structure and parameters of neural network

A typical three-layer neural network (input layer + hidden layer + output layer) was selected, and the number of nodes was set as 3-7-1:

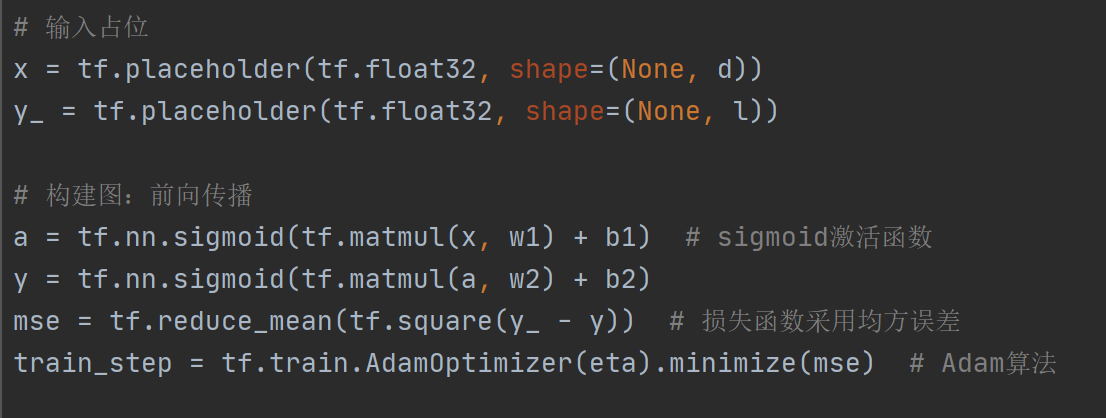


2. Initialize weights and thresholds

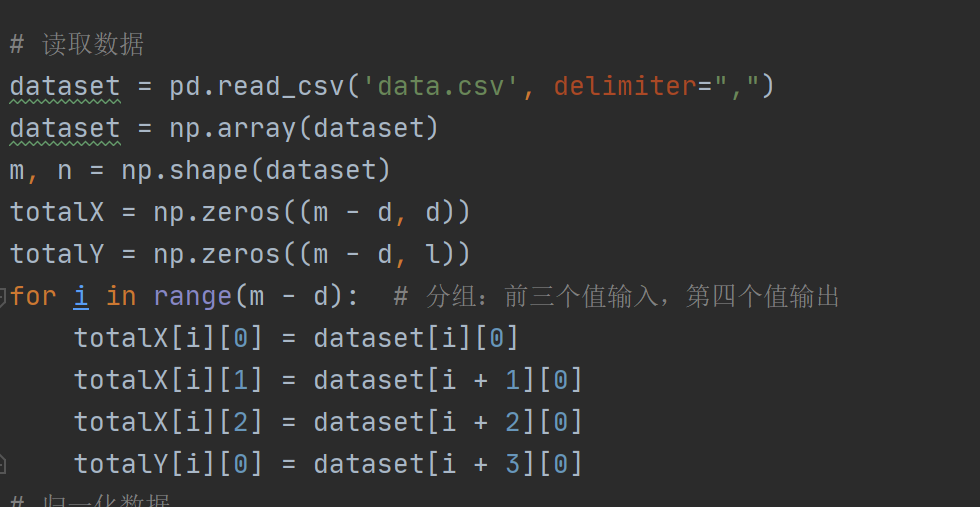
The weight from the input layer to the hidden layer is W1, and the matrix size is d\*q. The weight from the hidden layer to the output layer is W2, and the matrix size is Q \* L. The hidden layer threshold is B1, and the output layer threshold is B2.



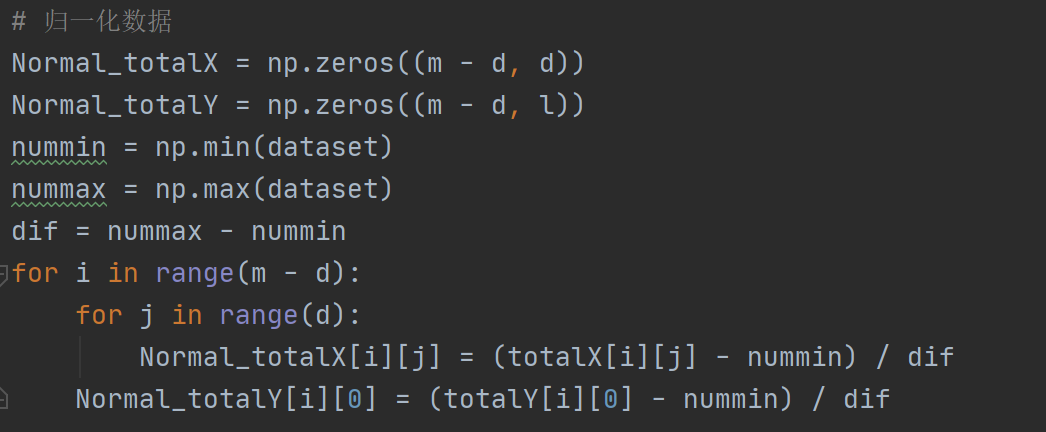
3. Build the figure



4. Input the data



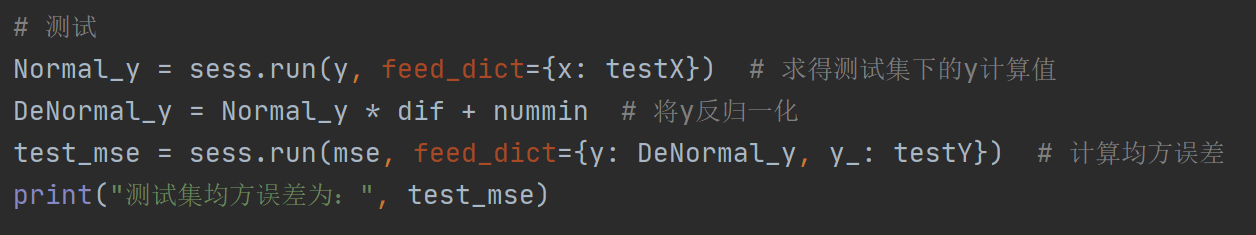
5. Data normalization



6. Create session execution diagrams and train the neural network



7. Test



8. Forecast

