CLEAN DATA AND USE PYTHON TO DRAW CHARTS

Khuyen V.Vu

Vaasan Ammattikorkeakoulu

University Of Applied Sciences

Email: vankhuyenvu1904@gmail.com

*Abstract*— Data science is the most important thing in an organization. The bigger company, the greater data they have to solve. So we can consider that data analysis is the improving compass of a company.

# I. INTRODUCTION

In this project, we used python with most of the knowledge about string manipulation, how to curl data from a non-protect website, and how to basic analysis it by drawing charts then.

We will take the data from the website diemthi.hcm.edu.vn

about almost 75000 students’ data, clean it then put it in a csv file

# II. PROCESS

## Take data from the website

Using curl function and subprocess in Python to do it. You have to know the begin and the end value of student number



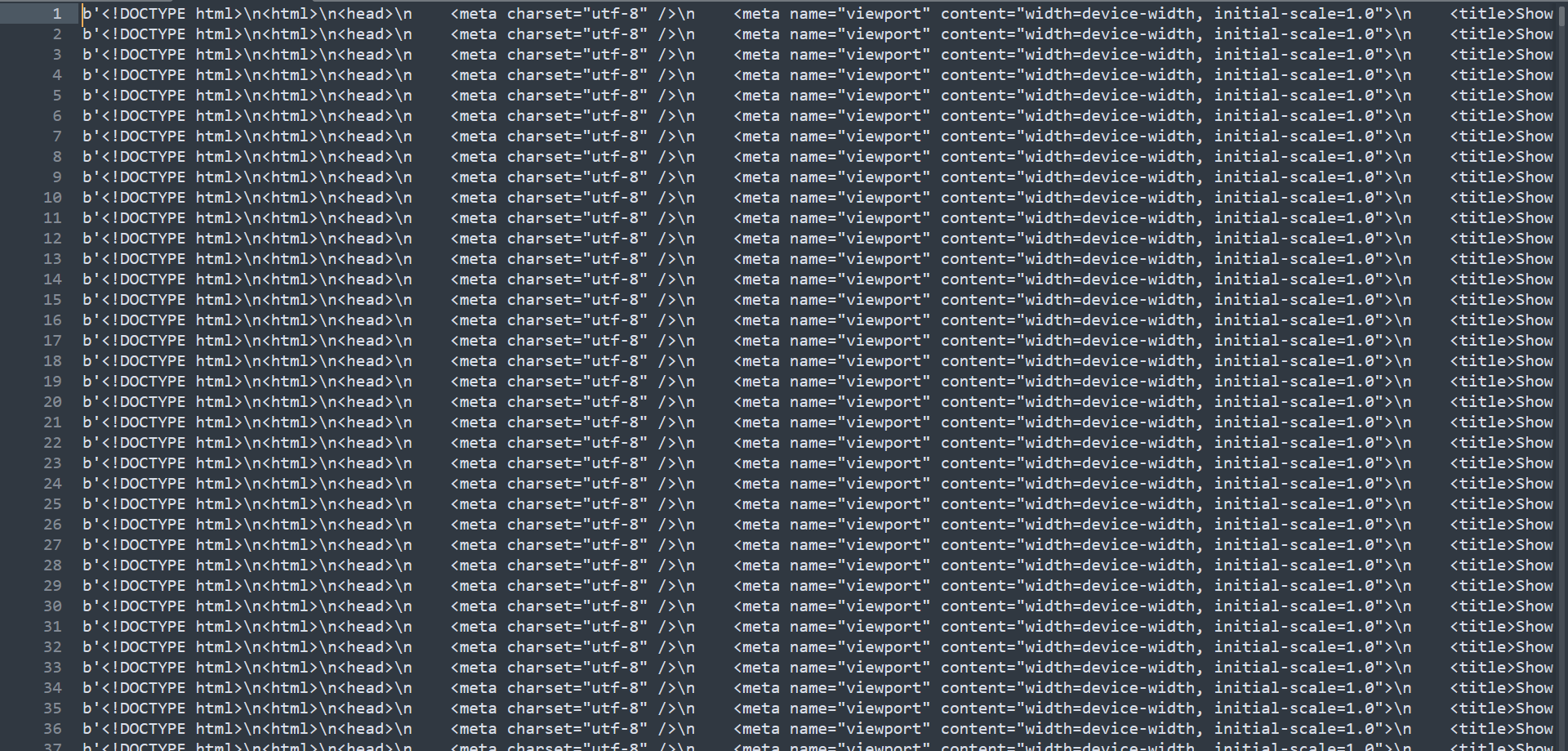
You have to have basic about website development to know how the command run, but I just do briefly, the main thing I want to focus is how can we work with the raw data

You can take the data from this drive link:

<https://drive.google.com/file/d/1ZxFUyu2W3j6eMt7rQvJDsKC4RrCCfGZE/view?fbclid=IwAR1_kVtZjKbHYkrin9e_3dVt6iFDaRpjxV15qzskyQmQ8GVA1RdAGaT3_IU>

## Clean Data

Raw data:

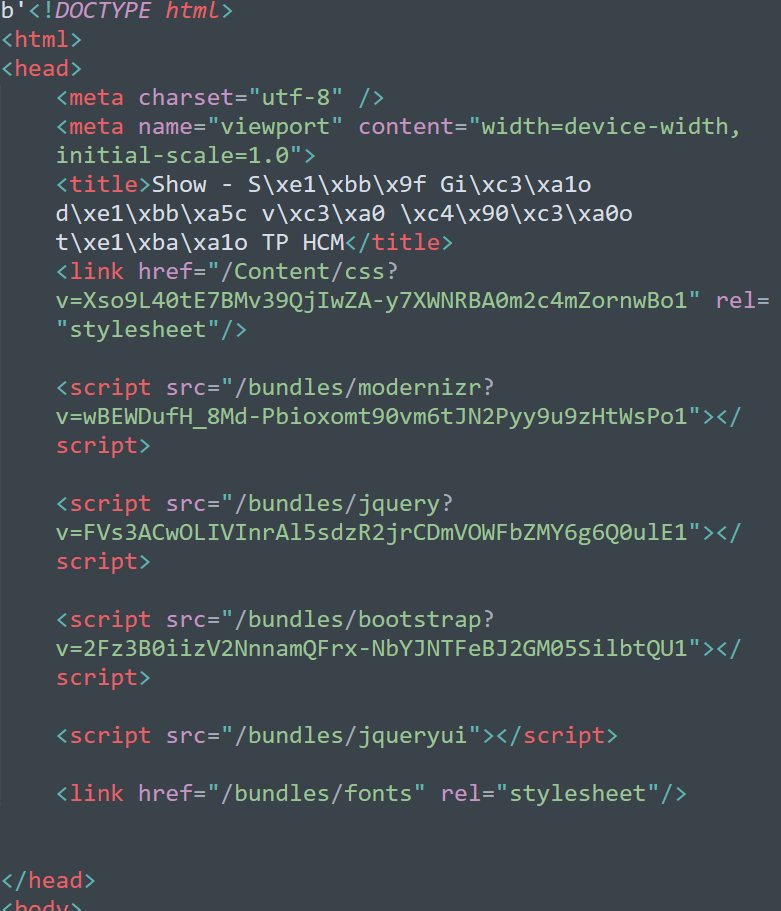


Because the raw data is about 75000 lines and it has the same structure, so I take 1 line and write it to a text file to work easily by using readline() function.

Then I used split() function to divide the whole string to a list which is divided by “\\n”.

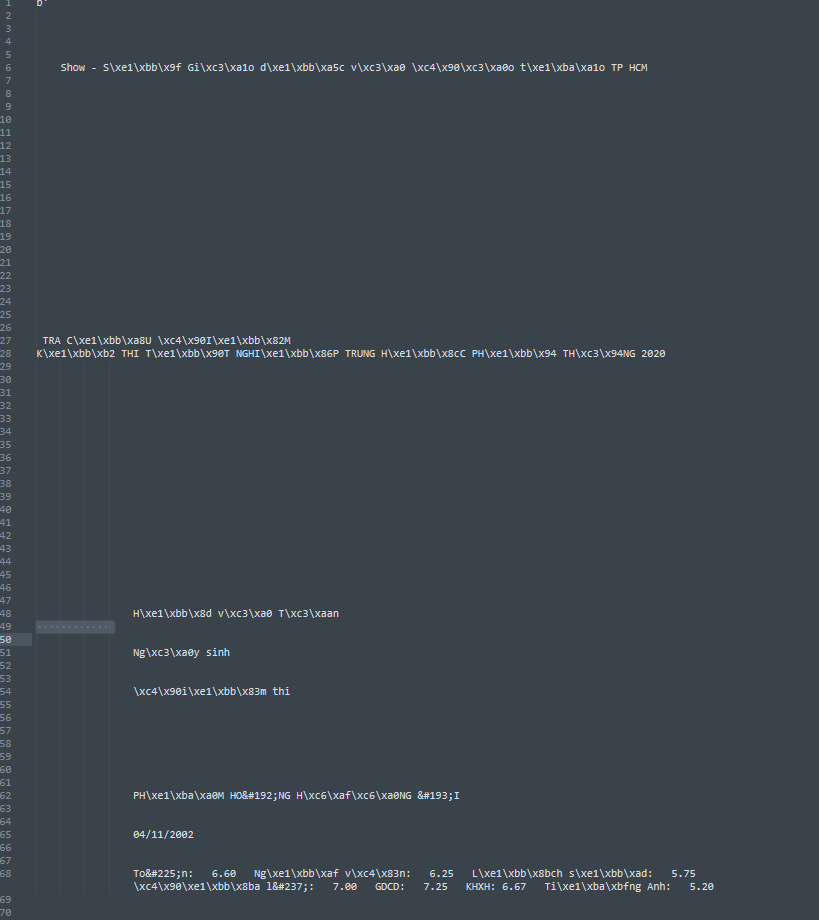
Next, I delete all the “\r”, “\t” characters.

The data after:



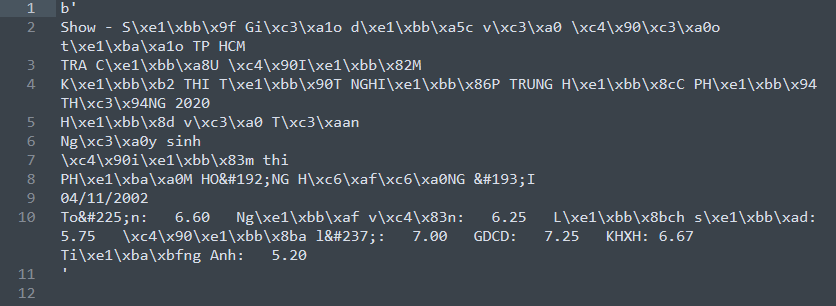
Then I deleted all the tags in data because we just need the name, date of birth and the scores of student to analyze.

Data after removing tags:



Next, I removed all the white space empty lines.

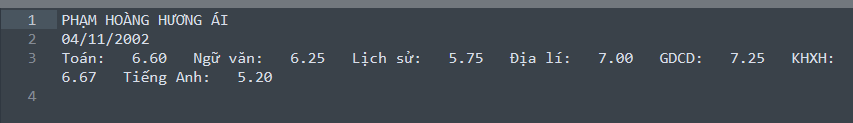
Data after :



Next, I decipher all the special characters and keep only the useful data.

I collected the characters from the internet then keep it in a text file, then use replace() function to replace all of it. And all of the characters started by &#... that we use chr() function to decipher it.

Data after:

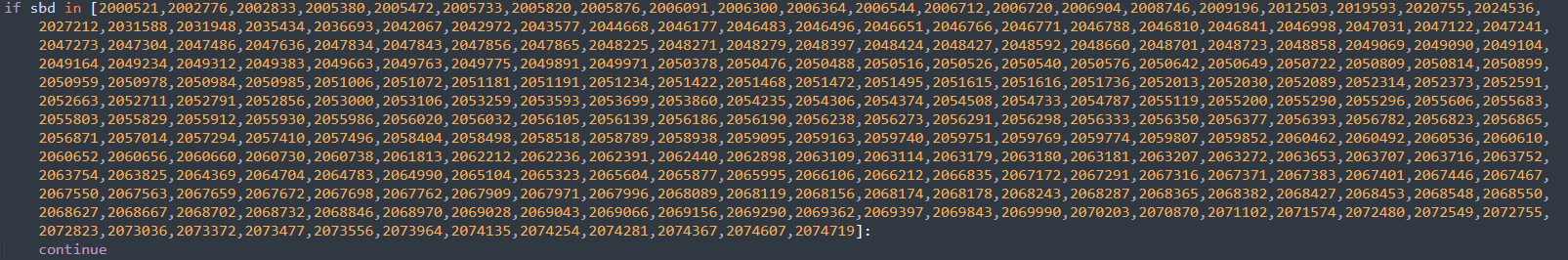


Then we arrange the data to clean data and put it in 1 row

Data after:



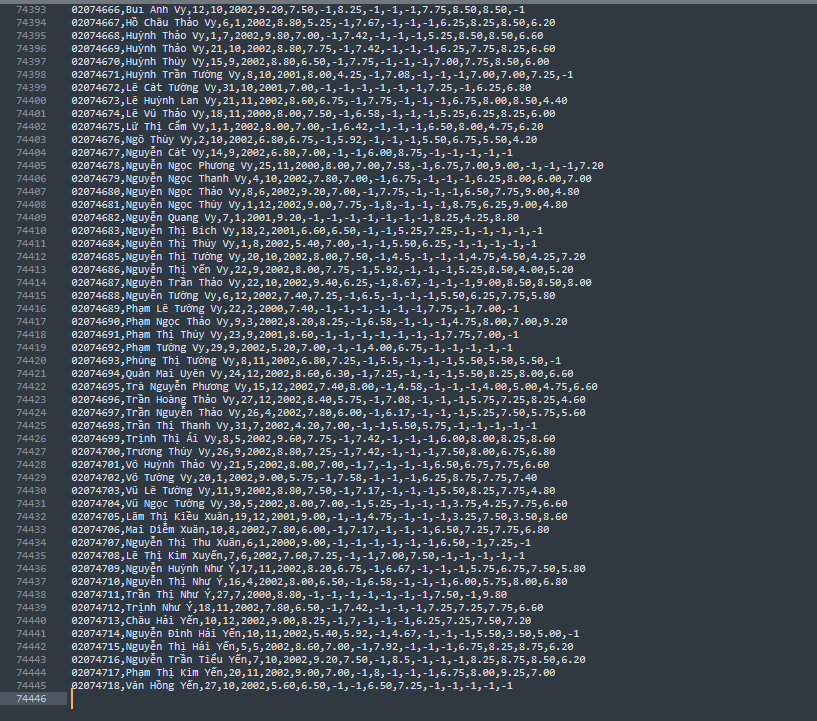
The last thing we have to do is using a loop to go through all the students. But I saw the problem is that we have some students who do not do the exam, so we use the try: except: function to filter the trash data, and I received:



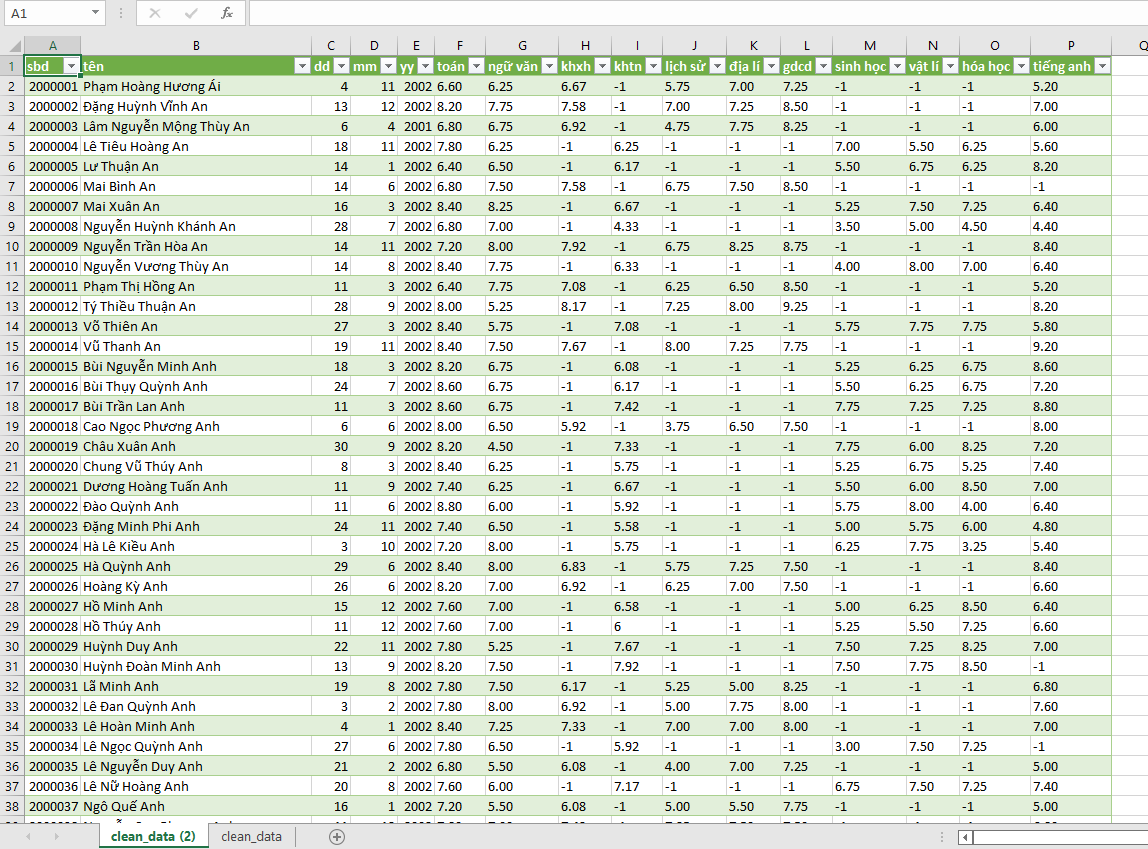
Above are all the students who did not do the exam, I ignore it by using the continue function. And change the text file to csv file.

You should take the code from my github to check easily.

Data after:



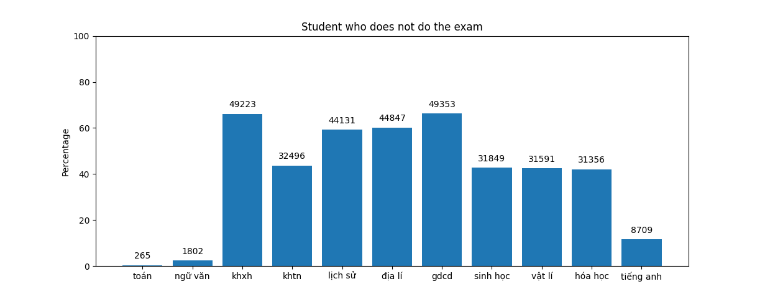
CSV file after clean:



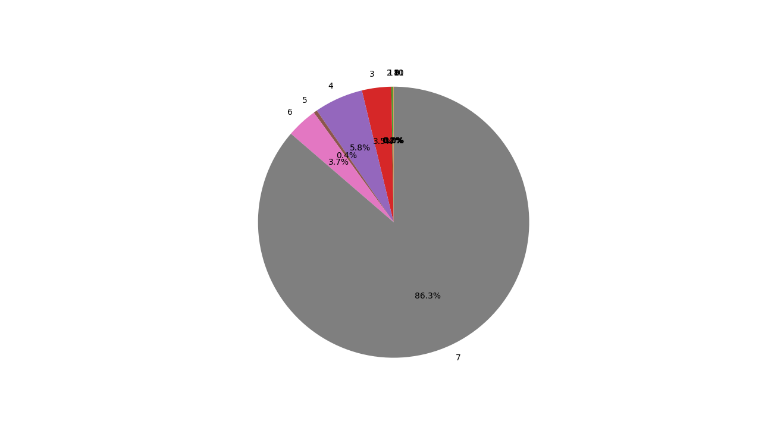
## Draw charts

I used matplotlib to draw all the chart.

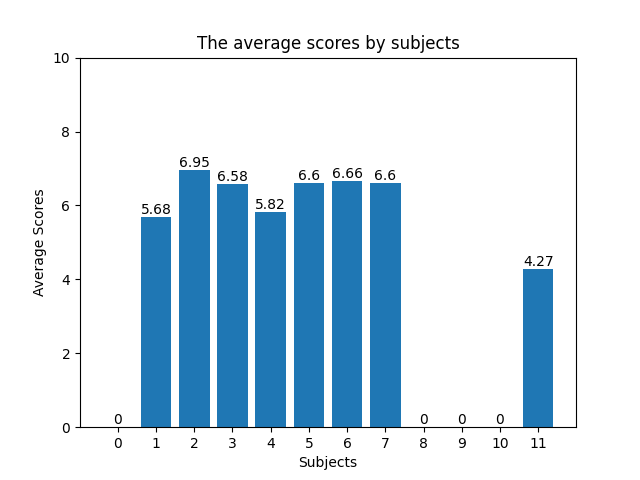
* Barchart: Student who does not do the exam.



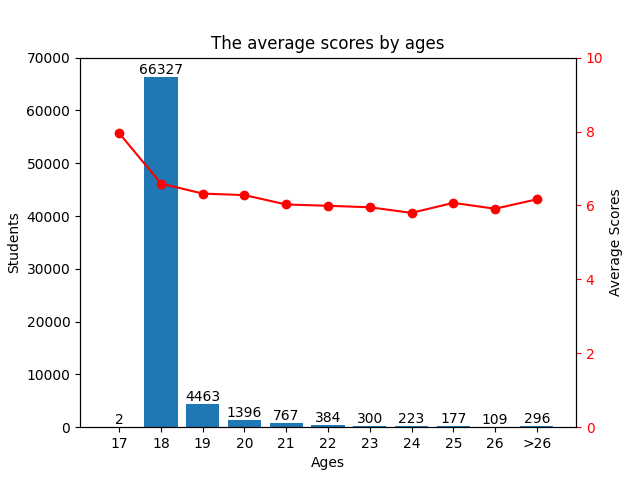
* Piechart: Percentage of the subject student often do.



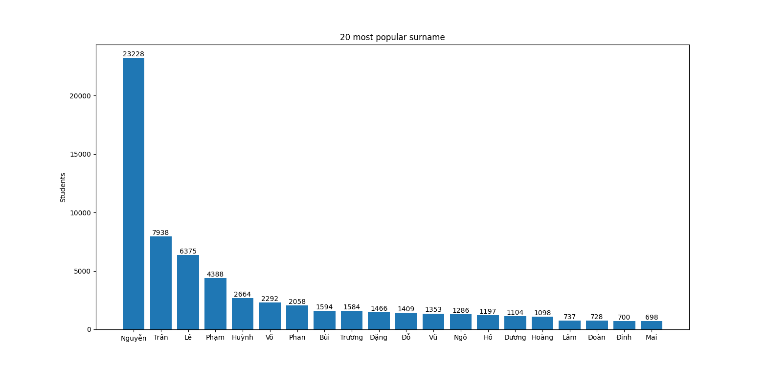
* Barchart: Average scores.



* Barchart – Linechart: Average scores by ages.



* Barchart: Most popular surname.



# III. DISCUSSION

* Process of a basic data project: take the data – clean – analyze – show the finish.
* About the web crawling, because this is quite of easy to take the data from that website, so that is the disadvantage I have to improve.
* String manipulation is the most important skill to clean data, we can do everything with python.
* Using matplotlib to draw the charts depend on what is your purpose.