The improvements of “the situation of China’s higher education”

Abstract

在对给定数据集进行可视化的过程中，需要考虑许多元素。既要兼顾数据的完整性和真实性，同样也要考虑数据的冗余和数据清洗。但是，最重要的是visualization的可读性。对于user来说，在阅读一个visualization时，能够清晰而准确地表达出数据与数据间的关系，让user在看到这个visualization的第一眼时不感到迷惑和烦躁，同时不破坏数据本身的结构，这个visualization可以说是比较成功的。

In the process of making a visualization of a given dataset, there would be a lot of elements that need to take in to consideration. For the user, when reading a visualization, the image can express the relationship between data clearly and accurately, so that the user will not feel confused and upset at the first sight of the visualization, not damaging the structure of the data itself at the same time. Then this visualization can be said to be relatively successful.

在这次project中，我们组选择了the story about the situation of Chinese higher education.我们使用了一张由新华社发表的visualization并且将它改造为了一张更加易读的visualization。但需要注意的一点是，原始的图像也并不能说是“失败的” visualization，因为在其源文章中，原始的图像起到的作用更多的是对于文章的修饰，也就是说其更注重的是图像的美观性。而这样做的结果便是其可读性大大地降低了，对于reader来说，对于这张图表的第一印象将是这张图表本身而不是它所反映的数据。

In this project, our team selected the story about the situation of Chinese higher education. We used a visualization published by Xinhua News Agency and transformed it into a more readable version. However, it should be noted that the original image is not actually a "failed" visualization, because in its original version, the original image focusing more on decorating the article, that is, it pays more attention to the beauty of the image. The result is that the readability is reduced. For the reader, the first impression of this chart will be the chart itself rather than the data it reflects.

当然，在经过我们的简单改造后，这张visualization就变得简单易读了。对于以学术为目的的information visualization的过程中，所重视的必然是图像的可读性。通过这样的对比，我们想说明的是信息可视化技术在学术论文中起到的作用。

Of course, after our simple transformation, this visualization becomes easier to read. In the process of information visualization for academic purposes, what we focusing on has to be the readability of the visualization. Through such comparison, what we want to illustrate is the importance of information visualization skills in academic papers.

我们首先将讲述graph中表达的story，然后是我们的改进，最后我们将总结整个report。

We will first describe the story expressed in the graph, then our improvement, and finally we will summarize the entire report.

The story of the graph

九月份是中国高校的开学季，许许多多的中国青年走进了高等院校的校园。中国的高等院校培养出了一代代年轻有为的中国青年，投身于中国发展的各行各业。

September is the back-to-school season for Chinese colleges and universities. Many Chinese young people go to the campus of colleges and universities. China's institutions of higher education have produced generations of promising young Chinese people who have devoted themselves to all walks of life in China's development.

中国现代高等教育波澜壮阔发展的序幕正式开启可以追溯到1895年10月２日，天津大学的前身北洋大学在天津诞生。

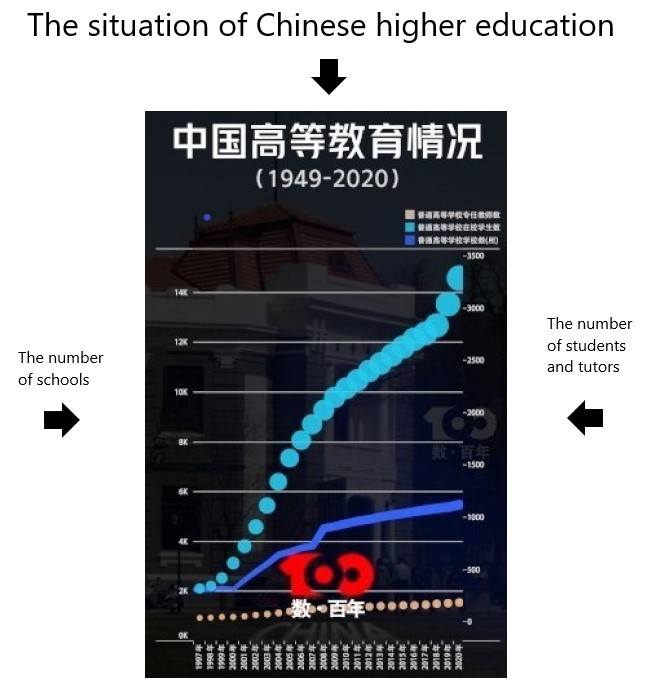
The spectacular development of China's modern higher education can be traced back to October 2, 1895, when the predecessor of Tianjin University, Imperial Tientsin University, was born in Tianjin.

中华人民共和国成立之初，中国仅有普通高等学校205所，经过数十年发展，尤其是1999年中国高等院校开始实行扩大招生之后，高等院校的数量、招生人数、在校生人数、毕业生人数以及学校的规模都有了爆炸式的增长。

At the beginning of the People's Republic of China, there were only 205 general colleges and universities in China. After decades of development, especially after the Chinese general colleges and universities began enrollment expansion in 1999, the number of general colleges and universities, the number of students, the number of graduates and the size of schools have all exploded.

为了了解相关情况，我们找到了一张图表（<http://www.news.cn/datanews/20211015/C999058529900001C39214B0FE601D68/20211015C999058529900001C39214B0FE601D68_1211351537_1630390571638_title.jpg>）。

然而，再研究图表的相关数据时，我们团队发现了很多该图再可视化方面很多不合理的地方。



In order to understand the specific information, We found a graph (<http://www.news.cn/datanews/20211015/C999058529900001C39214B0FE601D68/20211015C999058529900001C39214B0FE601D68_1211351537_1630390571638_title.jpg>)

However, when reviewing the data of the graph, our team found a lot of inadequacies in the visualization of the graph.

该图的X轴是年份信息，从1997年一直到2020年。Y轴则分为左右两边，左侧表示学校的数量；右侧表示教师和学生的人数，单位为“万人”。左侧Y轴对应的是图表中一条深蓝色的折线，代表普通高等学校数；右侧的Y轴对应的是图表中的两条散点，蓝色的散点表示普通高等学校在校学生数，棕色的散点表示普通高等学校专任教师数。

On the X-axis is the year information, from 1997 to 2020. The Y-axis is divided into the left and right sides, the left side represents the number of schools. On the right, indicates the number of teachers and students. The unit is ten thousand. The Y-axis on the left corresponds to the dark blue broken line in the graph, which represents the number of general colleges and universities; The Y-axis on the right corresponds to the two scatter points in the graph. The blue scatter points indicates the number of students in general colleges and universities, and the brown scatter points indicates the number of full-time teachers in general colleges and universities.

这些数据看上去井井有条，但实际上在可视化方面却有着一些不足。首先，我们团队发现，它的X轴的相对长度短，导致表示年份的X轴坐标垂直了，不便于我们查看。其次，散点表示人数时，变化趋势并没有特别直观，并且散点太大，不便于我们直接观察不同年份的具体人数数据。最后，两侧的Y轴坐标的原点高低不同，左侧的Y轴坐标区间太大（6K~14K的坐标甚至没有任何作用），右侧的坐标没有直线引导不便于观察数据。

The data looks neat, but it actually has some shortcomings in visualization. First, our team found that the relative length of its X-axis was short, which made the X-axis coordinates representing the year vertical and not convenient for us to observe. Secondly, when the scatter points represent the number of people, the change trend is not particularly intuitive, and the scatter points are too large for us to directly observe the specific data of the number of people in different years. Finally, the origin’s height of the Y-axis coordinates on both sides is different, the coordinate interval of the left Y-axis is too large (the coordinates of 6K~14K even have no effect), and the coordinate on the right is not guided by a straight line, which is not convenient for data observation.

针对上面的这些问题，我们的团队决定改进这张可视化图表。

In response to these problems, our team decided to improve the visualization graph.

这张图表展示了中国现代高等教育（1997~2020）从精英化走向大众化再走向普及化，成为世界上最大规模的高等教育大国的趋势，是了解、研究中国现代高等教育发展的重要图表材料。因此，对该图进行改进具有重大的意义。

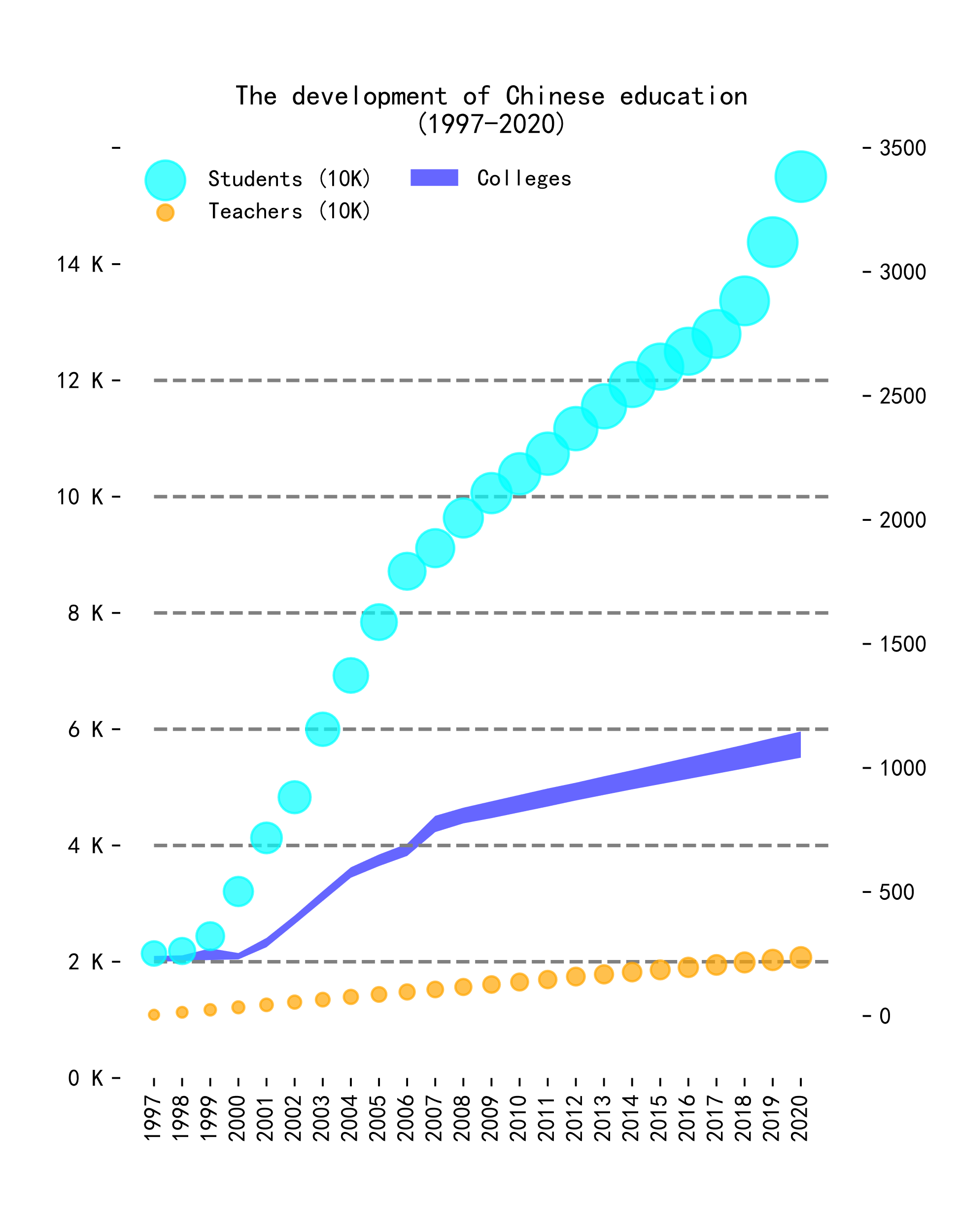
This graph shows the trend of China's modern higher education (1997-2020) from elitism to popularization and then to popularization. China becomes one of the largest countries of higher education in the world. It is an important graph material for understanding and studying the development of China's modern higher education. Therefore, it is of great significance to improve the graph.

Improvements

要想改进这张visualization，首先要将它还原出来。我们使用了matplotlib来完成这项工作。

下面是我们还原出的图表，我们将它以 .svg的格式输出出来。我们将原图像放在了旁边以供对照。

To improve this visualization, first we need to restore it in to a vector diagram. We used matplotlib to do this. Here is the diagram we restored. We output it in a .svg format. We put the original image aside for comparison.





也许你会发现图中的数据大小略有不同，那是因为我们实际上没有原始的数据集，而是编造了一组近似的数据集来使用。这是因为原文中并没有给出具体的数据集，我们不得不使用了自己编造的数据。但这其实并不是重点，我们想展现的是对这张图的改进，而不是真实的数据。

You may find that the data size in the graph is slightly different, because we actually do not have the original dataset, but fabricate a set of approximate datasets for use. This is because there is no specific data set in the original text, so we have to use the data we made up. But this is not the key point. We want to show the improvement of this chart, not the real data.

在前面的上下文中，我们提到了原始可视化的三个不足之处。首先就是两条y轴的使用。一左一右的两条y轴干扰了对于单位的分辨。解决方式也很简单：将这一张plot分为两张subplot。一张用于描述左侧坐标轴的数据：大学的数量，一张用于描述右侧坐标轴的数据：导师与学生的数量。

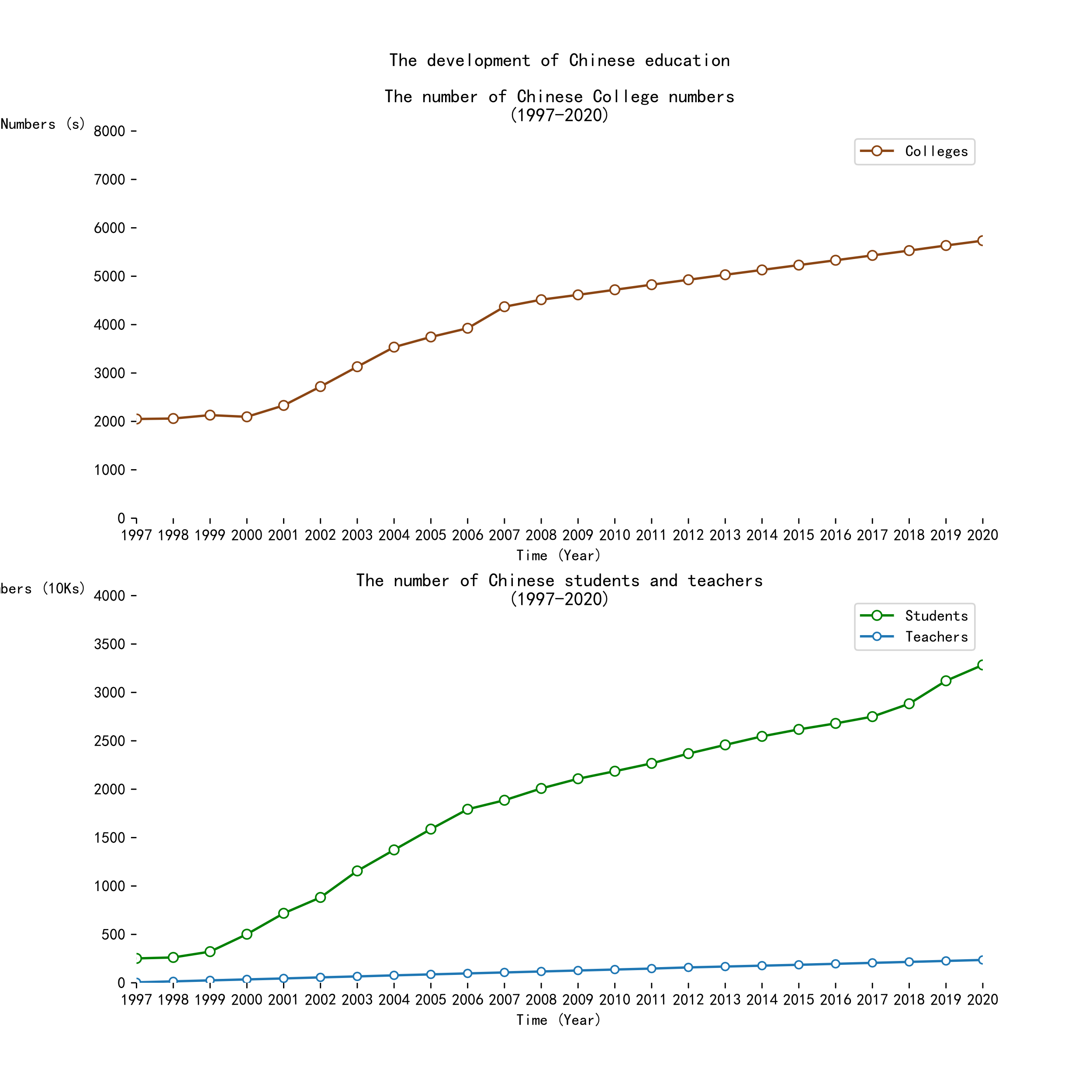
In the previous context, we mentioned three shortages of the original visualization. The first is the usage of the two y-axes. The two y-axes of one on the left and one on the right interfere with the resolution of units. The solution is also simple: divide this plot into two subplots. One plot is used to describe the data of the left y-axis: the number of universities, and one plot is used to describe the data of the right y-axis: the number of tutors and students.

其次是所使用的plot类型为散点图。虽然它可以粗略表现数据变化的趋势，但它不能更细致地表现数据的大小。我们这里选用了折线图配合节点进行数据可视化。一方面它更为清晰地展现出数据的涨幅，另一方面它也能更为准确的表现数据的大小。同时我们将图例统一，改用不同颜色来分辨不同数据。

Second, the plot type used is a scatter plot. Although it can roughly show the trend of data changes, it cannot show the size of data accurate enough. Here, we choose the line plot cooperating with nodes for data visualization. On the one hand, it can show the growth of data more clearly, on the other hand, it can show the size of data more accurately. At the same time, we will unify the legend and use different colors to distinguish different data.

最后，就是坐标轴宽度太窄。我们将横轴拉长，使plot可以恰好将所有年份横向输出而不发生重叠。下面是我们最终的完成图。

Finally, the width of the x-axis is too narrow. We lengthen the horizontal axis so that the plot can output all the years horizontally without overlapping. Here is our final finished picture.



这就是the final version of our improvement of the original plot. 将其与original plot进行对比，你就会发现它对于数据的展现更加直观了。你可以清晰地看到数据的涨幅，同时可以准确地获取数据的大小。相比于original plot要简洁了许多。

This is the final version of our improvement of the original plot. Comparing it with the original plot, and you will find that it is more intuitive for data presentation. You can see the growth of the data clearly and obtain the size of the data accurately. Compared with the original plot, it is much cleaner.

Summary

Overall, 我们说明了选择这张visualization的原因，这张visualization的故事以及对visualization的改进。相信看到这里，你已经对于information visualization技术有了一个初步的印象。在对于数据的处理和展示中，以实用目的为主的information visualization技术可以使数据展现得更加cleaner，高效。使readers能够轻易地从visualization中获取所需要的信息，story,从而对所展示的dataset有一个更深层次的理解。

Overall, we stated the reason for choosing this visualization, the story of visualization and the improvement of visualization. I believe you have got a preliminary impression of information visualization technics. In the process of data processing and presentation, information visualization skills for practical purposes can make data presentation cleaner and more efficient. Readers can easily obtain the required information and stories from the visualization, so that they can have a deeper understanding of the dataset displayed.