

# 时空分析与可视系统——长江学者信息数据分析与可视化

## 第九组

姓名:

学号:

队长: 马健  
队员: 雷一明  
队员: 申豪良  
队员: 姚洪良  
队员: 焦超

ZF1721248  
ZF1721213  
ZF1721255  
ZF1721403  
ZF1721209

# TOPICS

- ✓ 时空数据分析与可视化介绍
- ✓ 数据搜集与处理过程展示
- ✓ 数据可视化及分析

# 1. 时空数据分析与可视化介绍

## ➤ 时空数据分析与可视化介绍

- 空间自相关是指一些变量在同一个分布区内的观测数据之间潜在的相互依赖性。地理学第一定律指出，任何东西与别的东西之间都是相关的，但近处的东西比远处的东西相关性更强。
- 比如：分析我们的可视化效果图，从长江学者的籍贯分布图来看，南方地区的江苏省的样本点分布密集，临近的北方地区的山东省也很密集；而西北地区的甘肃省，距离江苏省很远，它的样本点分布很稀疏，同样处于西北地区的青海，内蒙古自治区的样本点分布也很稀疏。
- 由于这种相关性的传递，北方和南方地区的样本分布很稠密，距离较远的东北地区 and 西北地图样本点比较稀疏。



图1.1长江学者籍贯分布的空间关系

## ➤ 时空数据分析与可视化介绍

### ➤ 空间信息可视化

✓ 空间信息可视化是指运用计算机图形图像处理技术，将复杂的科学现象和自然景观及一些抽象概念图形化的过程。

### ➤ 空间信息可视化的形式

✓ 地图

✓ 图像

✓ 统计表

✓ 三维多媒体虚拟现实

我们采用的是地图的可视化形式，如图1.2所示

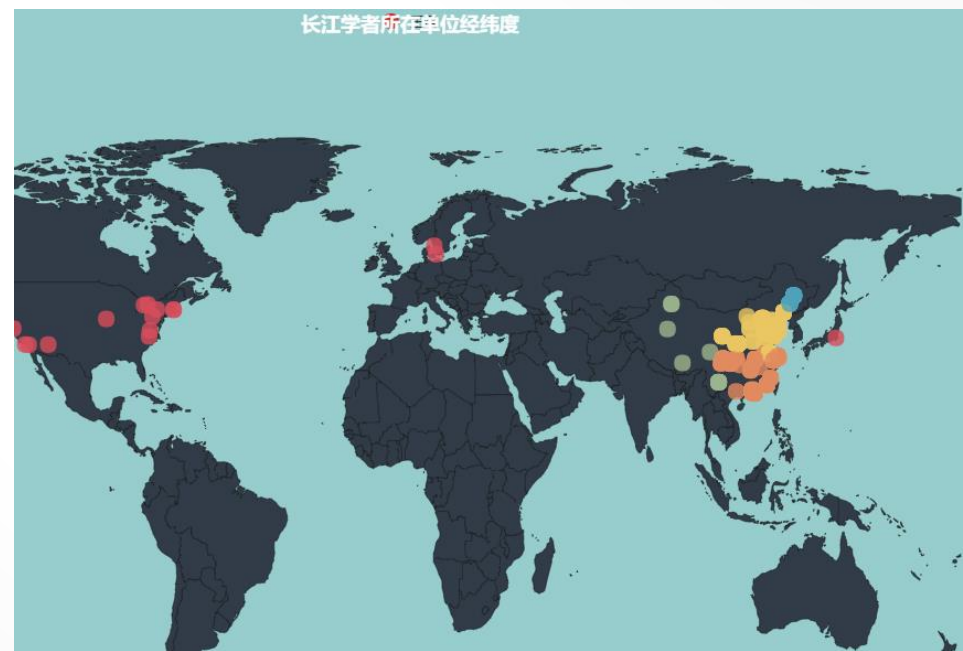


图1.2 长江学者分布图

## 2. 数据搜集与处理过程展示

## ➤ 数据搜集与处理过程展示

- 我们组的任务是进行长江学者信息数据分析与可视化。
- 首先我们组的组员在分工之后，在互联网上搜集长江学者的信息，在搜集数据，尤其是搜集经纬度位置坐标的过程中，我们用到的Geocoding API services有GPSpg。此外还用了高德地图搜索数据。
- 小组每个人分配的数据量是200人左右。最终我们汇总综合第八组和第十组搜集的信息，得到最终的数据。包括长江学者的年龄，性别，籍贯，学校所在地，工作单位所在地等。
- 右图为初始数据

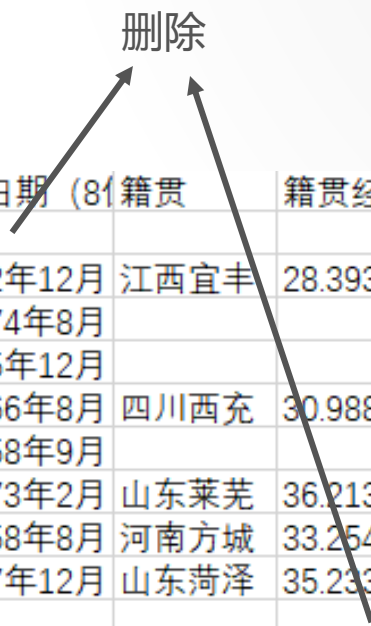
姓名	性别	出生日期 (81	籍贯	籍贯经纬度	单位
李葆春	男				多伦多大学
李葆明	男	1962年12月	江西宜丰	28.393800	复旦大学
李斌	男	1974年8月			电子科技大学
李波	男	1965年12月			香港科技大学
李波	男	1966年8月	四川西充	30.988009	北京航空航天大学
李伯耿	男	1958年9月			浙江大学
李传峰	男	1973年2月	山东莱芜	36.213590	中国科学院
李春英	女	1958年8月	河南方城	33.254390	第四军医大
李春忠	男	1967年12月	山东菏泽	35.233630	华东理工大学
李丹	女				大连理工大学
李道亮	男	1971年2月	山东垦利	37.586787	清华大学
李稻葵	男	1963.12.22	安徽凤阳	32.868048	清华大学
李德才	男	1966年2月	河北沧州	38.303831	16300,116.83
李德润	男				清华大学
李登峰	男	1965年10月	广西博白	22.275984	福州大学

图2.1 长江学者初始数据

## ➤ 数据搜集与处理过程展示

➤ 搜集好的初始数据存在较多缺失，主要包含籍贯，出生年月，我们利用Excel等工具进行数据清理，去掉了信息缺失过多的样本数据。

➤ 为保证数据的有效性，清洗原则是每个样本来说，缺失三条以上的数据就删除，如果一条数据不存在就设置为Null。



The diagram shows two arrows pointing from the '出生日期' (Date of Birth) and '籍贯' (Place of Origin) columns to the word '删除' (Delete) at the top. This indicates that rows with missing data in these columns are being removed from the dataset.

姓名	性别	出生日期 (81	籍贯	籍贯经纬度	单位
李葆春	男				多伦多大学
李葆明	男	1962年12月	江西宜丰	28.393800	复旦大学
李斌	男	1974年8月			电子科技大学
李波	男	1965年12月			香港科技大学
李波	男	1966年8月	四川西充	30.988009	北京航空航天大学
李伯耿	男	1958年9月			浙江大学
李传峰	男	1973年2月	山东莱芜	36.213590	中国科学院
李春英	女	1958年8月	河南方城	33.254390	第四军医大学
李春忠	男	1967年12月	山东菏泽	35.233630	华东理工大学
李丹	女				大连理工大学
李道亮	男	1971年2月	山东垦利	37.586787	清华大学
李稻葵	男	1963.12.22	安徽凤阳	32.868048	清华大学
李德才	男	1966年2月	河北沧州	38.303831	6300,116.83
李德润	男				清华大学
李登峰	男	1965年10月	广西博白	22.275984	福州大学

图2.2 对长江学者相关数据进行清洗



## ➤ 数据搜集与处理过程展示

- 考虑到数据的可视化的要求，我们对数据再次进行清洗，将包含籍贯与学校经纬度信息较多的数据进行了提取，最后得到数据是不存在缺失值的数据。（也考虑了缺失值的众数性赋值）

- 下图即为完整的数据

	name	籍贯	籍贯纬度	籍贯经度	单位城市	单位城市	在单位纬	在单位经	科学学校纬	科学学校经	研究生学校纬	研究生学校经	博士学校纬	博士学校经度
0	郁建兴	桐乡	30.63	120.565	30.307	120.082	30.271	120.195	30.872801	120.123	30.313559	120.39	31.336472	121.499
1	郁文贤	松江	31.032	121.228	31.23	121.402	31.206	121.425	28.229031	113.001	28.229031	113.001	28.229031	113.001
2	郁振华	上海	31.23	121.474	31.23	121.402	31.23	121.391	31.230452	121.391	31.230452	121.391	31.230452	121.391
3	喻国明	上海	31.23	121.474	23.101	113.293	23.128	113.348	23.120075	113.316	23.120075	113.316	23.120075	113.316
4	喻景权	义乌	29.307	120.075	30.307	120.082	30.271	120.195	30.272337	120.196	32.261948	99.949	32.261948	99.949
5	袁宏永	麻城	31.173	115.008	39.905	116.407	40.003	116.327	28.174528	112.926	31.843676	117.296	30.537860	114.365
10	袁小平	仪征	32.272	119.185	39.905	116.407	34.221	117.2	34.221153	117.2	34.221153	117.2	34.221153	117.2
14	岳晓奎	河南	34.766	113.754	30.677	108.94	34.242	108.905	34.241802	108.905	34.241802	108.905	34.241802	108.905
15	岳珠峰	丹阳	32.01	119.607	30.677	108.94	34.242	108.905	34.241802	108.905	34.241802	108.905	34.241802	108.905
16	翟婉明	靖江	31.983	120.277	30.776	103.956	29.576	103.446	29.575670	103.446	29.575670	103.446	29.575670	103.446
17	翟学伟	南京	32.06	118.797	32.059	118.755	32.055	118.779	39.103693	117.167	39.103693	117.167	39.103693	117.167

图2.3 长江学者最终数据

## ➤ 数据搜集与处理过程展示

➤ 地名匹配不确定性的来源有：

✓ 简称

少数民族自治区 少数民族 遗漏

✓ 地区/市 名称混乱

✓ 地区/市 名称的历史沿革

考虑到数据中籍贯信息的参差不齐，考虑的信息的对等性，我们另外将所有籍贯统一到了省份，然后用于可视化展示。

籍贯省份	name
上海市	10
云南省	5
内蒙古自治区	8
北京市	6
吉林省	11
四川省	27
天津市	2
安徽省	35

图 2.4 籍贯的统计结果

### 3.数据可视化及分析

## ➤ 数据可视化及分析

- 我们把搜集到的信息清洗后，采用Kmeans聚类算法将数据分为4类，又综合考虑中国四大地理分区，为了方便计算，我们将中国分为四个地区，分别是 东北、西部、北方、南方和国外五个地区，其范围为  
东北：北纬42-53度 东经103-135度 记为1  
西部：北纬18-53度 东经73-103度 记为2  
北方：北纬23-42度 东经103-135度 记为3  
南方：南方18-32度 东经103-135度 记为4  
国外：其他 记为5。

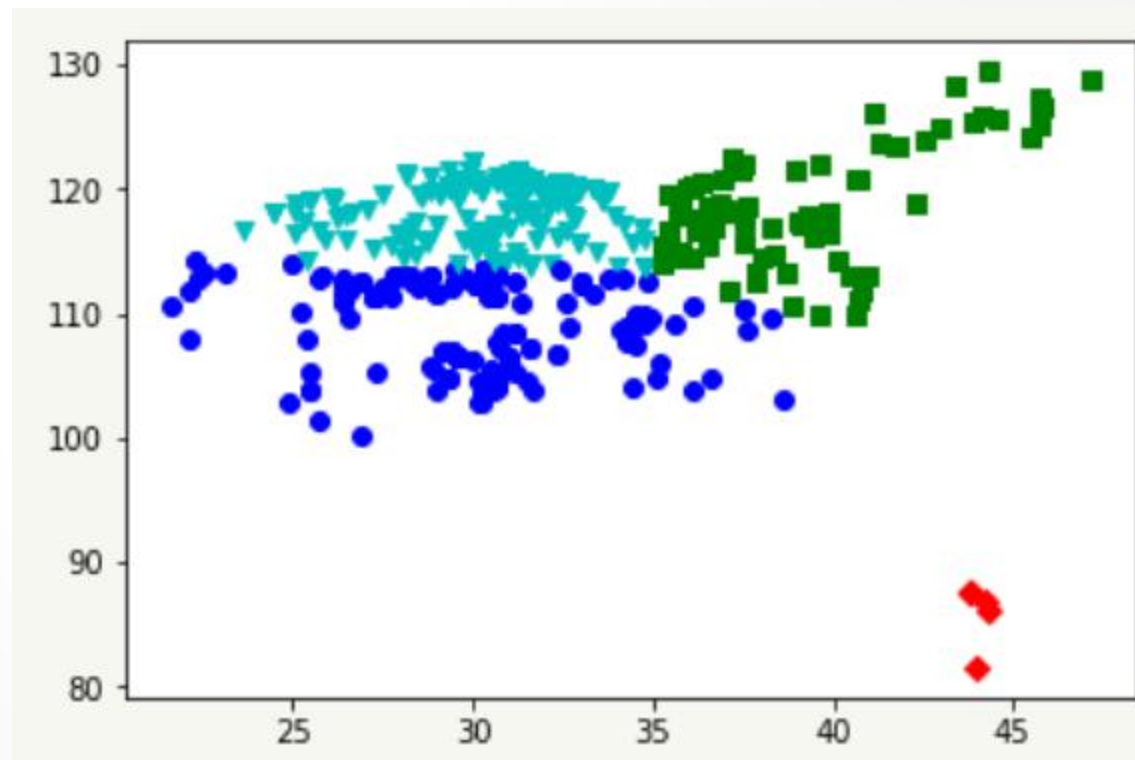


图3.1 利用聚类分析对数据分类

## ➤ 数据可视化及分析

➤ 利用Pyecharts工具生成可视化图形。

从右图可以看出，30-35岁和51-65岁左右的长江学者占少部分，总和仅占16%，36-40岁的长江学者占20%左右，说明很少有人可以在40岁之前就可以评上长江学者，而超过55岁被评上长江学者的可能性也会很小。41-45岁的长江学者占比超过40%，说明这是最有可能成为长江学者的年龄，46-50岁的长江学者占22%左右，这个年龄段的长江学者比41-45岁的人少，年龄的中位数在这个时间段，说明在这个年龄附近的人会更有可能会成为长江学者。综上，36-50岁被评为长江学者地人数最多，45岁应该是长江学者地“黄金年龄”，在这个年龄被评为长江学者的可能性很高。

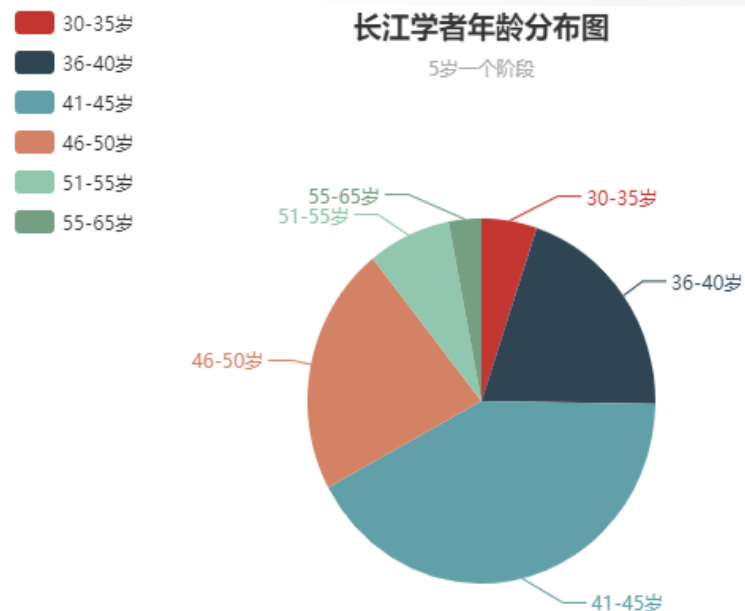


图3.2 长江学者年龄分布图

## ➤ 数据可视化及分析

➤ 可以通过右图得出结论，女性在长江学者中仅占总人数的5%左右。对女性来说，成为长江学者比男性更为困难。女性成为长江学者的可能性很低（在我们有效数据的分析之后，可能性不到5%）。

➤ 为了分析原因，我们横向的对比了第一组和第二组做的中国科学院和中国工程院的相关数据，女院士仅占院士总人数的5%左右，查阅国家杰出青年基金的数据，截止2016年来资助3004人，其中女性256人，仅占8%左右。我们推断造成这种现象的原因并不是先天性的，首要原因应该是在第一批长江学者接受高等教育的年代，女子受到教育的比例比较低，根据教育部1998年的数据，当时的女大学生毕业人数比男大学少了近79万人<sup>[1]</sup>。我们推断另一个重要的原因是因为长期以来形成的以男性为主的科研环境，这种状况短时间内不易改变。

➤ [1]数据来源：教育部<http://old.moe.gov.cn/publicfiles/business/htmlfiles/moe/s8493/index.html>

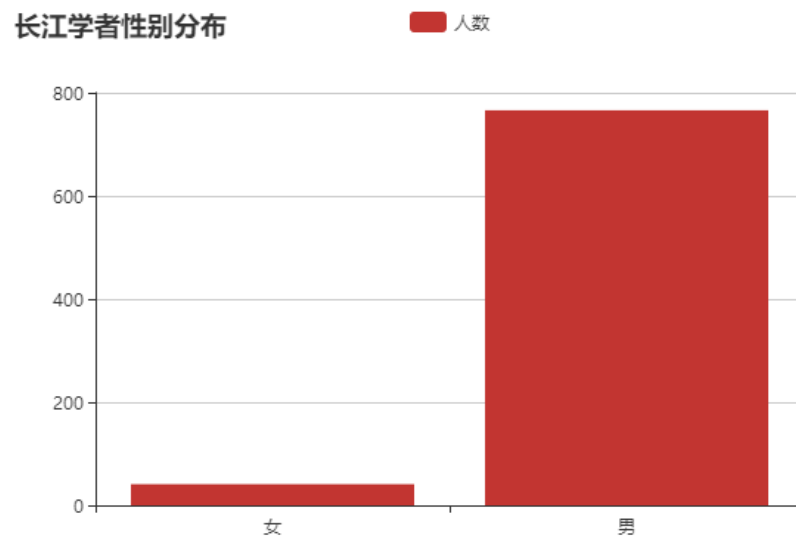


图3.3 长江学者性别分布图

## ➤ 数据可视化及分析

从右图可以得到以下信息，除去没有查询到地数据，长江学者的本科学校在北方和南方的分布最为集中。分别为44%和43%，通过经纬度分布图可以看到在北方地区主要集中在北京，天津，西安，济南等地，这些地方有很多985，211高校，十分有利于人才的培养。南方地区则主要集中在上海，杭州，长度，广州等城市，同样地，这些地区地教育水平也很发达，而东北地区本科学校占比仅为6%左右，这些地区仅有哈尔滨，沈阳等城市地教育水平较高，211院校也不多，占比少也在情理之中。国外地占比很少，仅仅不到1%，说明很少有长江学者一开始就在国外接受本科教育，99%以上的长江学者都是在国内接受本科教育的。

综上，北方和南方的有很多985，211重点院校，所以这些地区的长江学者本科学校的占比比较高。

本科学校在各个地区的数目

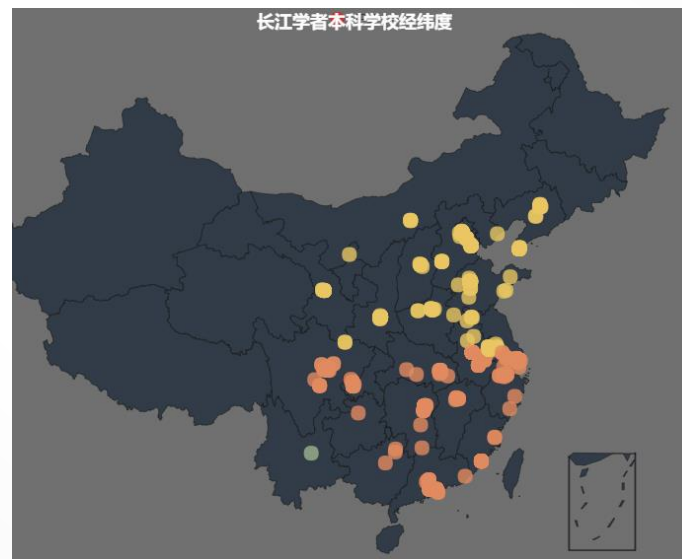
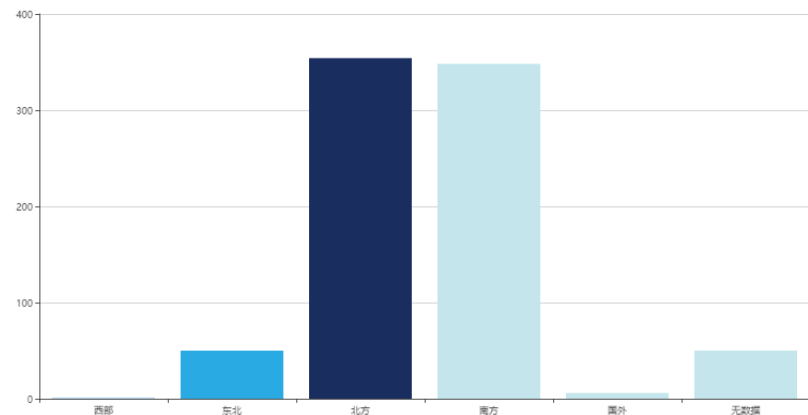


图3.4 长江学者本科学校数据分布图



## ➤ 数据可视化及分析

同样从右图可以得到以下信息，除去没有查询到地数据，长江学者的研究生学校在北方和南方的分布也最为集中。分别为44%和29%，通过经纬度分布图可以看到北方地区比如北京和西安有很多985，211高校，这些都是长江学者研究生阶段的母校。从这个占比来看，南方地区略低于北方地区。西北地区和东北地区占比仅为7%左右，原因和上一条的分析相似。国外的占比虽然不多，仅占6%但是明显比本科阶段的占比高，说明有很多人选择出国，留学深造，以费城，华盛顿，波士顿，西雅图，伦敦，剑桥这些地区为主。

研究生学校在各个地区的数目

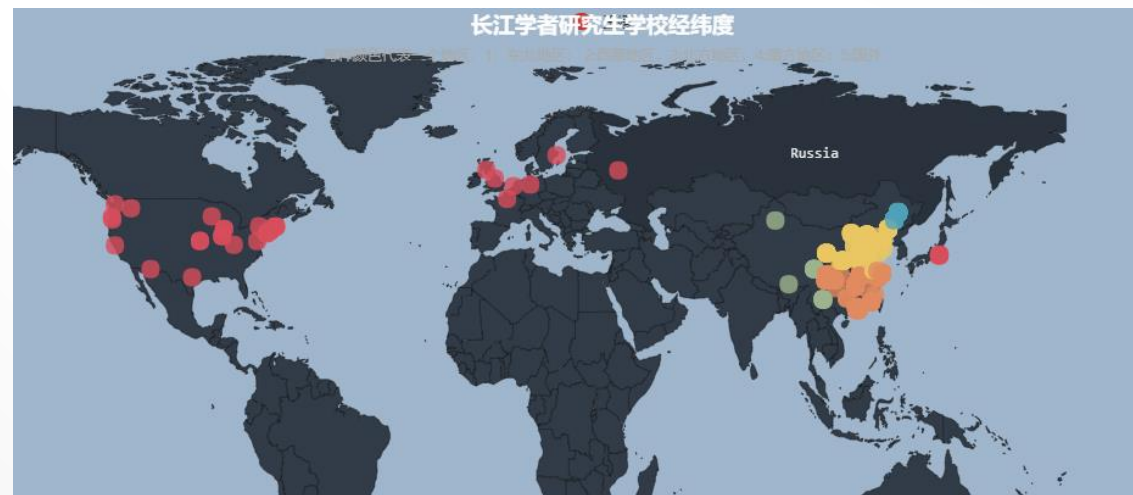
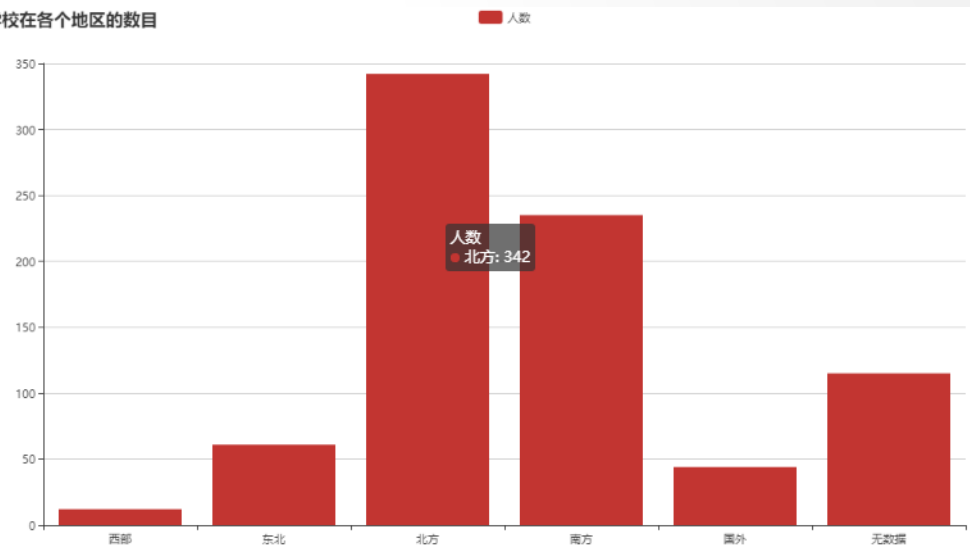


图3.5 长江学者研究生学校数据分布图



## ➤ 数据可视化及分析

从右图可以得到以下信息，长江学者的博士学校在北方的占比为39%，通过经纬度分布图可以看出北方地区的北京最为密集；长江学者的博士学校在北方的占比为24%，西北地区地区和东北地区占比仅为7%左右，原因和上一条的分析相似。国外的占比为11%，综合可以看到，对于长江学者来说，出国读博的人并没有想象中的占比那么高。

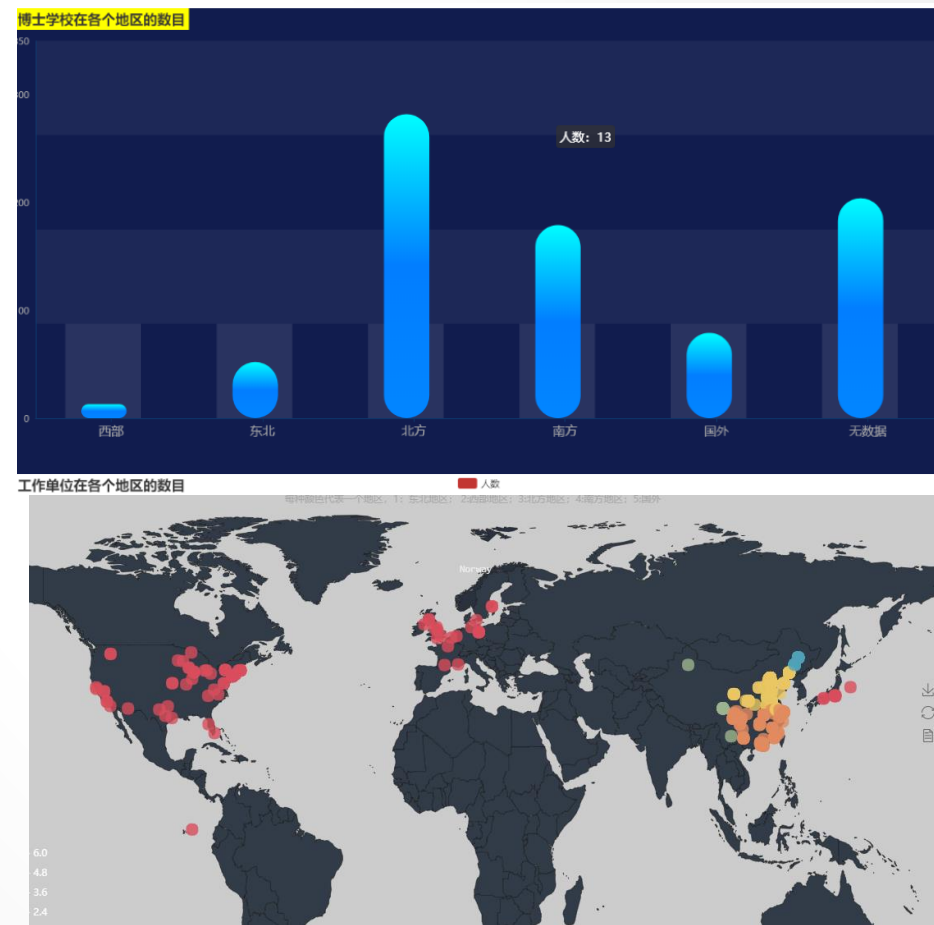


图3.6 长江学者博士学校分布图

## ➤ 数据可视化及分析

- 在地图中按照导入的长江学者工作单位的数据生成分布图，可以得到以下结论，超过80%的长江学者被国内重点高校聘用，也有少部分长江学者被美国和欧洲的一些院校聘用。

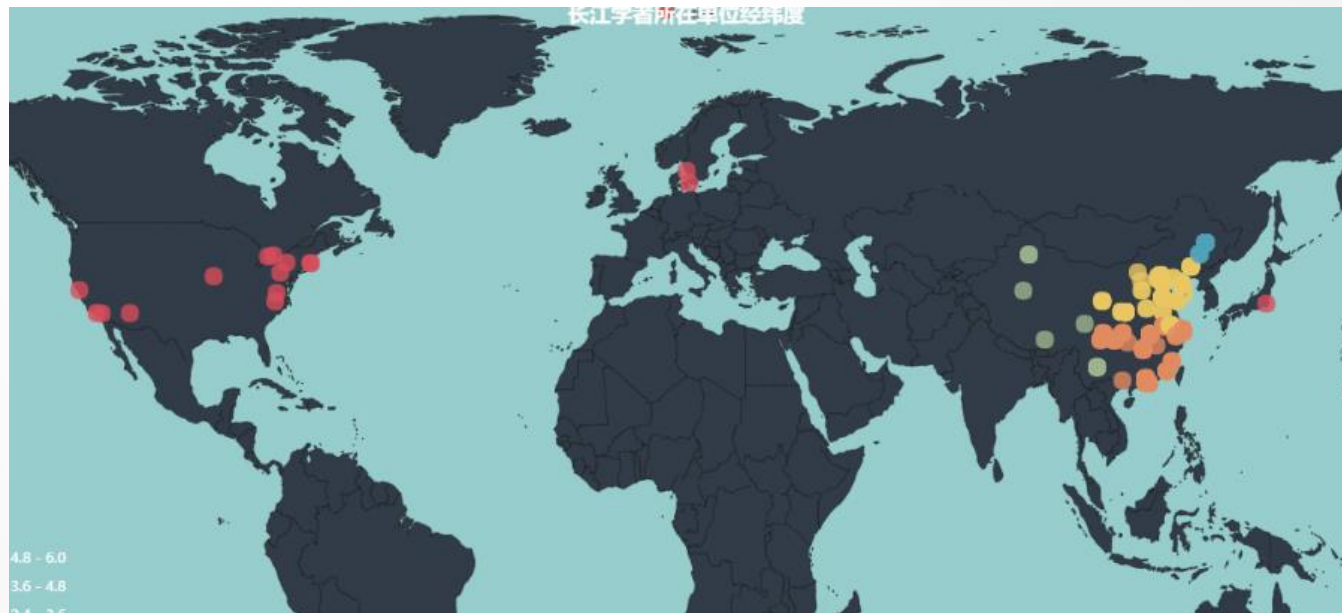


图3.7 长江学者的工作单位分布图

## ➤ 数据可视化及分析

综合分析以上数据，通过这些学校的人数密度来看，本科密度明显比研究生，博士密度低，并且研究生、博士的学校主要集中在985学校和国外高校中，这说明有很多的长江学者在本科时候的学校并不是很好，但是在获得更好的教育资源后依然能够通过自身努力成为长江学者，所以受到良好的高等教育可以算是成为长江学者的一个必要条件，而且从这个规律中也可以看出，虽然出身不好，但是通过自身的努力后，仍然可以达到很高的高度，这可以作为我们的激励使我们好好学习。

## ➤ 总结

通过这次作业，我们小组数据对数据可视化的重要性有了较好的认识与理解，并掌握了数据可视化的技巧与方法，这对我们以后的学习与研究具有很好的启发性意义。