《数据科学与数据分析》实验课作业

Week 11_Exercise 5

191820019 陈文杰

数据来源：请下载课件中的“作业数据”。

要求：请对“作业数据”中的“专利清洗作业”数据进行清洗，实现以下功能，并回答下列问题。

功能：

（1）清洗数据，利用**申请号将申请表中的信息和授权表中的信息匹配起来**，形成一张新的表，这张表要对应申请和授权两方面的信息。至于申请未授权，以及授权无申请信息的，酌情处理用于分析，无限制条件。

（2）**最少要实现**如课件第 4 章“南京专利数据清洗”案例中出现的**所有统计工作，并绘制图表**，可用于回答下列问题。

问题：

（1）怎样评估江北地区的核心技术优势？

（2）结合江北的部分产业规划，对江北地区的专利引进工作给出一定建议。**【简要作答，500 字以内】**

（3）数据清洗过程中遇到了哪些问题，你是如何解决的？谈一谈你对数据清洗工作的心得。**【最少 200 字】**

格式与内容要求：

（1）回答问题时，请先使用 Word 作答，另存为 PDF 文件，提交 **PDF** 版本，上传到“教学立方”；（只提交作业，不要提交数据！）

（2）**回答问题要有逻辑，条理清晰，注意格式，各级标题要有区分。不标序号且写成一篇作文或日记的作业分数会较低。**

一、数据清洗与图表绘制

1、数据清洗

```
library(dplyr)
library(xlsx)
library(readxl)  #不需要 java 环境 read_xlsx
apply.data = read_xlsx("E:\\Desktop\\Week 11_数据整理作业数据 (1)\\专利清洗作业.xlsx",2)
authorize.data = read_xlsx("E:\\Desktop\\Week 11_数据整理作业数据 (1)\\专利清洗作业.xlsx",4)

#1 数据清洗
#1.1 缺失值处理
which(is.na(apply.data),arr.ind=T)  #which 函数查看 df 总体是否有缺失值，arr.ind 返回坐标
sum(is.na(apply.data))  #求缺失值元素个数总和
sum(complete.cases(apply.data)) #求非空行数总和
table(is.na(apply.data$申请号)) #根据某列统计缺失的行数
table(complete.cases(apply.data$专利类型))  #True 代表非缺失
mean(is.na(apply.data$申请号))  #查看缺失值占比
#循环遍历，查询每列缺失值占比
for (i in colnames(apply.data)){
  print(table(is.na(apply.data[,i])))
  print(mean(is.na(apply.data[,i])))
}
apply.data <- select(apply.data,申请号:申请日)
authorize.data[!complete.cases(authorize.data),]  #列出有缺失值的行
nrow(authorize.data[!complete.cases(authorize.data),])  #计算有缺失值的样本量
#1.2 数据表合并
# merge(x, y, by = intersect(names(x), names(y)),
#       by.x = by, by.y = by, all = FALSE, all.x = all, all.y = all,
#       sort = TRUE, suffixes = c(".x",".y"),
#       incomparables = NULL, ...)
data = merge(apply.data,authorize.data,by = "申请号")
str(data)
#1.2 数据去重
index <- duplicated(data$申请号)
data[!index,]
write.csv(data,"E:\\Desktop\\data.csv")
```

2、统计分析与图表绘制

因图表过多，详见本 word 文档附录部分。

二、简答题

（1）怎样评估江北地区的核心技术优势？

行业名称	计数项:主分类号前三位 区属.x2					
	主分类号前三位	高淳区	鼓楼区	建邺区	江北新区	总计
测量仪器	G01	14	125	66	359	564
一般的物理或化学的方法或装置	B01	5	38	18	138	199
发电、变电或配电 H02 131 22 6	H02	6	42	24	119	191
建筑物	E04	16	40	21	101	178
基本电器元件	H01	3	25	13	123	164
输送；包装；贮存；搬运薄的或细丝	B65	11	19	23	107	160
机床；其他类目中不包括的金属加工	B23	3	4	4	111	122
水、废水、污水或污泥的处理	C02	6	32	19	52	109
电通信技术	H04	3	14	13	46	76
家具；家庭用的物品或设备；咖啡磨；香料磨；一般吸尘器	A47	1	2	1	43	47
医学或兽医学；卫生学	A61		20	1	20	41
	总计	68	361	203	1219	1851

上表是针对特定的几个行业，进行区属之间的专利授权统计所得的结果，据此对江北地区的核心技术优势进行评估，得出以下几点结论：

①江北地区技术优势明显。可以看到，江北新区在精密仪器、生化、电力、水利、通信等行业的专利授权量明显高于其他的几个区属，具有较为明显的技术优势。

②产业分布不合理。这些技术仍然是偏向服务于传统产业，在这些领域拥有核心技术优势不代表其在新材料、新能源以及互联网等领域有较强竞争力，而新兴产业又是当今经济的主要增长点和驱动力，所以江北新区在保持现有技术优势的同时，也应考虑加大对新兴产业的投入，促进专业的提质升级。

③样本分布的影响。值得注意的是，江北新区包含了浦口区、六合区和栖霞区八卦洲街道，但是与之进行对比的都是单独的分区，所以即使江北新区总体核心技术优势明显，下属分区技术优势仍有待考量，基于此应关注新区内部的统一、协调发展，掌握核心技术、增强区属竞争力。

（2）结合江北的部分产业规划，对江北地区的专利引进工作给出一定建议。

①江北新区的产业规划

以往的江北地区主要承接江南的产业专业为主，收纳了许多夕阳产业，对江北地区的生态环境以及经济可持续发展造成了许多负面影响，因此在政府关于本轮打造江北新区的政策文件中，特意提出要着眼长远，对江北新区的产业进行高标准规划，实行清洁生产。

②专利引进工作建议

于此，江北地区的专利引进应着眼于包括互联网、新能源、新材料、通讯等

高新技术，为江北地区的产业提质升级注入生机，争取早日形成以新能源、信息技术、生物医药、汽车轨道交通以及节能环保在内的五大先进制造业，建立并完善现代产业体系。

（3）数据清洗过程中遇到了哪些问题，你是如何解决的？谈一谈你对数据清洗工作的心得。

#数据清洗过程中的问题与解决方法

①数据不完整。

首先进行数据概览，并分析缺失值在行列上的集中分布，进而决定采用适合的方式进行处理（删除、填充...）。在本次实验中，缺失值主要存在于部分记录的部分属性中，缺失值较少，而且这些属性值大多无关后续的统计分析，故未对缺失值进行删除或填充。

②数值不匹配。

以专利主题分类号为例，它大致分为两类，一类带字母，一类不带字母而是纯数字，难以进行合并分析。故通过人工查询，将纯数字的分类号转化为带字母的分类号，实现数值类型的统一，利于后续的专利类型统计分析。

③数据重复。

数据集难免出现重复数据，因而需要时刻注意数据的去重。

④数据无意义

在数据合并时如果未能选择合理的合并方式，可能得到笛卡尔积的数据统计表，里面的很多记录都无实际意义，因而选定特定的连接列、选用合适的合并方式至关重要。

#数据清洗心得

①数据清洗往往伴随数据采集进行，是完整数据分析流程的靠前期的工作，也是较为关键的部分，重要性不言而喻。

②数据清洗较为繁琐，一方面需要考虑数据的外况，看数据结构是否合理、数据缺失值是否有碍数据分析、数据是否有重复；另一方面需要结合现实，考察数据记录、数据值是否有意义，考察数据的内涵。

③在数据清洗阶段充分利用可视化图表，一方面提高清洗的效率和准度，另一方面可以加深对数据集的认知，为后续的模型搭建与研究分析奠定基础。

【附录：实验结果】

图 1：专利申请数区属分布条形图（1）

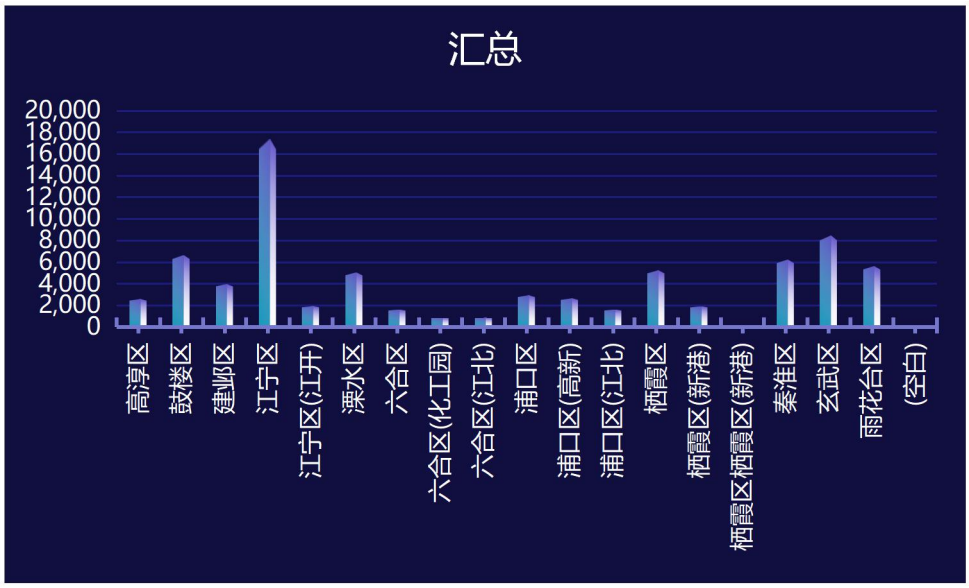


图 2：专利申请数区属分布条形图（降序）

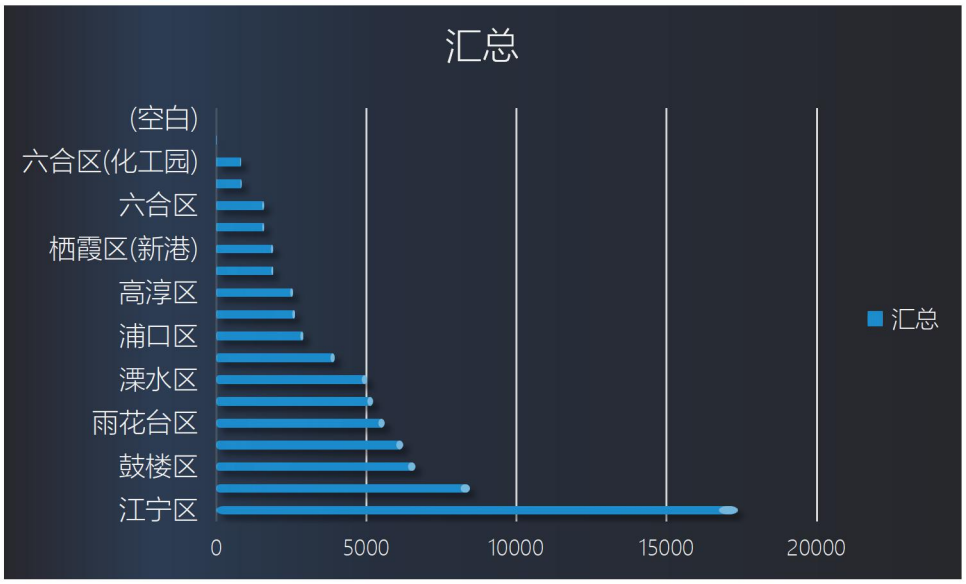


图 3：专利申请数区属分布饼状图（显示百分比，部分区属未组合）

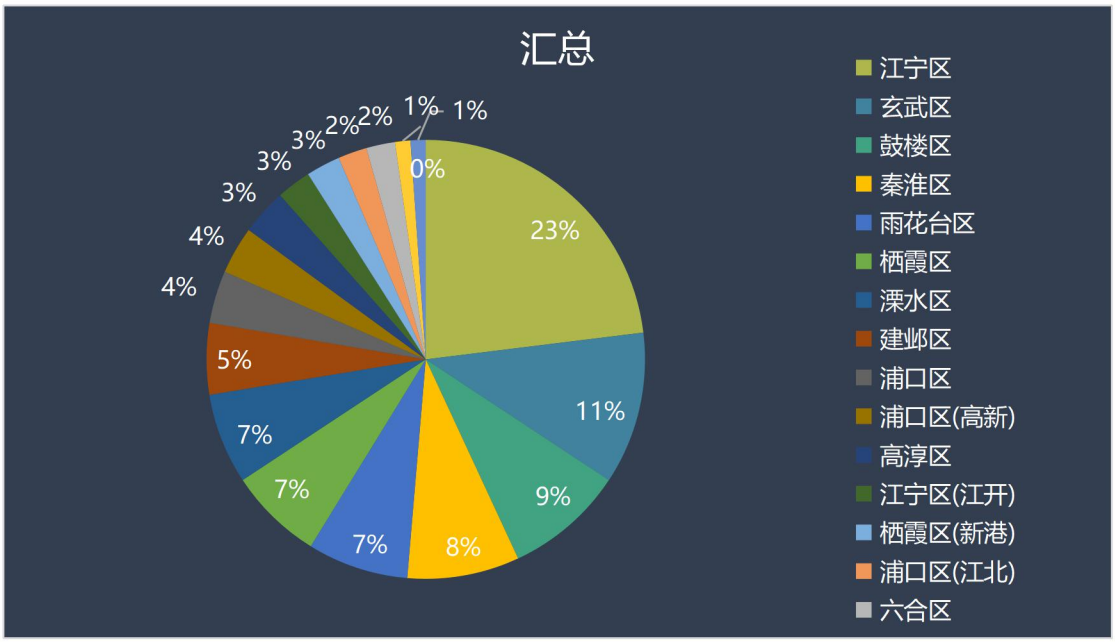


图 4：专利申请数区属分布饼状图（显示数量）

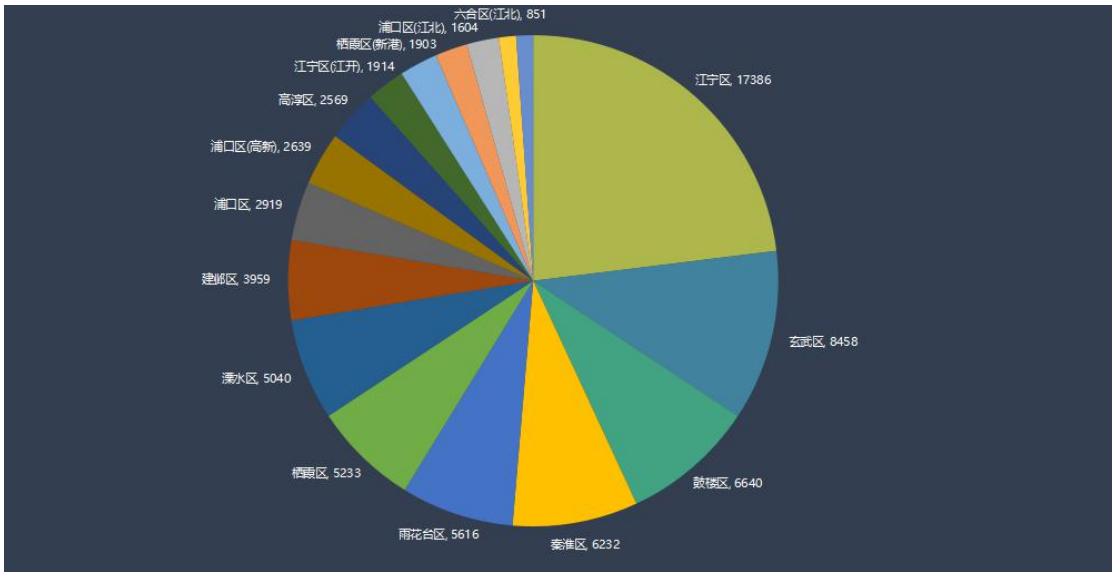


图 5：专利申请数区属分布饼状图（显示百分比）

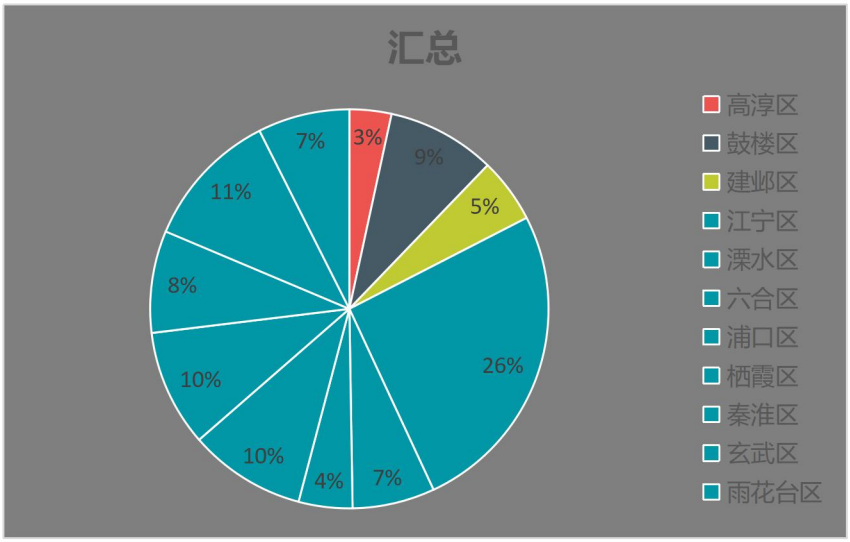


图 6：专利申请数区属分布条形图（部分区属汇总）

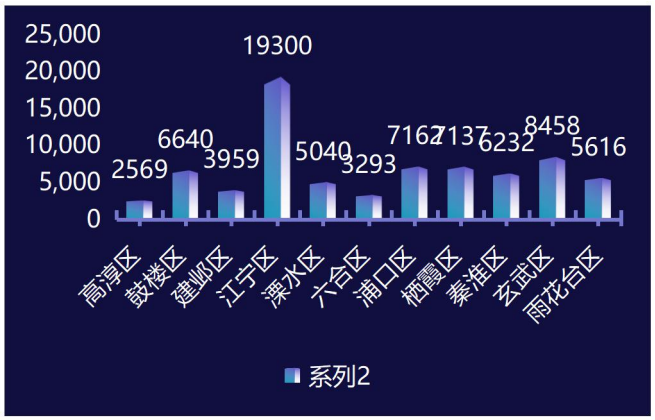
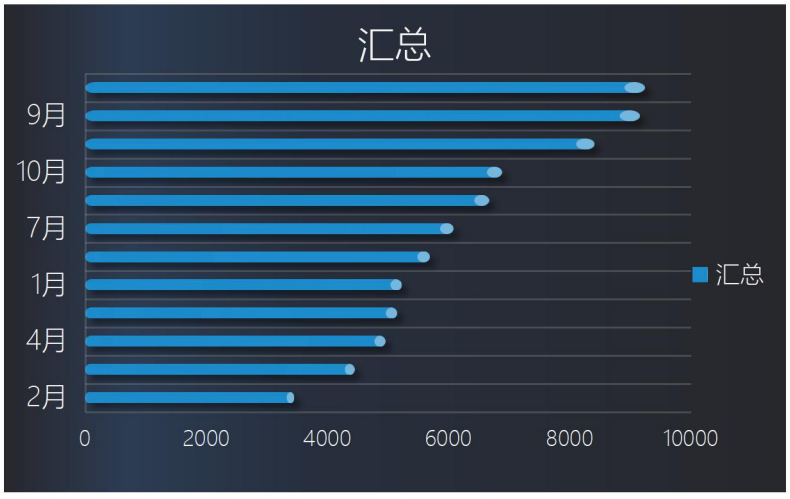


图 7：专利申请月份统计条形图（部分区属汇总）



计数项:区属	专利类型				
区属	发明	实用新型	外观设计	(空白)	总计
高淳区	1526	838	205		2569
鼓楼区	3404	2733	503		6640
建邺区	1388	1526	1045		3959
江宁区	7449	7140	2797		17386
江宁区(江开)	508	621	785		1914
溧水区	2131	2051	858		5040
六合区	514	938	149		1601
六合区(化工园)	391	401	49		841
六合区(江北)	460	326	65		851
浦口区	1348	1472	99		2919
浦口区(高新)	1152	1349	138		2639
浦口区(江北)	695	854	55		1604
栖霞区	2741	2171	321		5233
栖霞区(新港)	778	1036	90		1904
秦淮区	3806	2068	358		6232
玄武区	6080	1949	429		8458
雨花台区	2915	2168	533		5616
(空白)					
总计	37286	29641	8479		75406

表 1：专利类型统计表

计数项:区属	申请人类型						
区属	大专院校	个人	机关团体	科研机构	企业	(空白)	总计
高淳区	12	212	5	7	2333		2569
鼓楼区	2026	1029	593	640	2352		6640
建邺区	343	345	23	51	3197		3959
江宁区	5502	3241	142	185	8316		17386
溧水区	197	400		74	4369		5040
六合区	22	315	31		1233		1601
六合区(化工园)	57	25	4	15	740		841
六合区(江北)	172	36			643		851
浦口区	701	924	6	4	1284		2919
浦口区(高新)	808	84			1747		2639
浦口区(江北)	637	274		27	666		1604
栖霞区	2179	629	31	141	2253		5233
栖霞区(新港)	1	38	1	29	1835		1904
秦淮区	1988	1599	227	290	2128		6232
玄武区	4055	1950	316	1108	1029		8458
雨花台区	17	762	315	268	4254		5616
总计	18717	11866	1694	2903	40226		75406

表 2：申请人类型统计表（数值）

计数项:申请人类型	申请人类型				
区属	大专院校	机关团体	科研机构	企业	总计
高淳区	0.47%	0.19%	0.27%	90.81%	100.00%
鼓楼区	30.51%	8.93%	9.64%	35.42%	100.00%
建邺区	8.66%	0.58%	1.29%	80.75%	100.00%
江宁区	31.65%	0.82%	1.06%	47.83%	100.00%
江宁区(江开)	0.00%	0.00%	3.34%	96.50%	100.00%
溧水区	3.91%	0.00%	1.47%	86.69%	100.00%
六合区	1.37%	1.94%	0.00%	77.01%	100.00%
六合区(化工园)	6.78%	0.48%	1.78%	87.99%	100.00%
六合区(江北)	20.21%	0.00%	0.00%	75.56%	100.00%
浦口区	24.02%	0.21%	0.14%	43.99%	100.00%
浦口区(高新)	30.62%	0.00%	0.00%	66.20%	100.00%
浦口区(江北)	39.71%	0.00%	1.68%	41.52%	100.00%
栖霞区	41.64%	0.59%	2.69%	43.05%	100.00%
栖霞区(新港)	0.05%	0.05%	1.52%	96.38%	100.00%
秦淮区	31.90%	3.64%	4.65%	34.15%	100.00%
玄武区	47.94%	3.74%	13.10%	12.17%	100.00%
雨花台区	0.30%	5.61%	4.77%	75.75%	100.00%
总计	**24.82%**	**2.25%**	**3.85%**	**53.35%**	**100.00%**

表 3：申请人类型统计表（横向百分比）

计数项:申请人类型	申请人类型					
区属	大专院校	个人	机关团体	科研机构	企业	总计
高淳区	0.06%	1.79%	0.30%	0.24%	5.80%	3.41%
鼓楼区	10.82%	8.67%	35.01%	22.05%	5.85%	8.81%
建邺区	1.83%	2.91%	1.36%	1.76%	7.95%	5.25%
江宁区	29.40%	27.31%	8.38%	6.37%	20.67%	23.06%
江宁区(江开)	0.00%	0.03%	0.00%	2.20%	4.59%	2.54%
溧水区	1.05%	3.37%	0.00%	2.55%	10.86%	6.68%
六合区	0.12%	2.65%	1.83%	0.00%	3.07%	2.12%
六合区(化工园)	0.30%	0.21%	0.24%	0.52%	1.84%	1.12%
六合区(江北)	0.92%	0.30%	0.00%	0.00%	1.60%	1.13%
浦口区	3.75%	7.70%	0.35%	0.14%	3.10%	3.87%
浦口区(高新)	4.32%	0.71%	0.00%	0.00%	4.34%	3.50%
浦口区(江北)	3.40%	2.31%	0.00%	0.93%	1.66%	2.13%
栖霞区	11.64%	5.30%	1.83%	4.86%	5.60%	6.94%
栖霞区(新港)	0.01%	0.32%	0.06%	1.00%	4.56%	2.52%
秦淮区	10.62%	13.48%	13.40%	9.99%	5.29%	8.26%
玄武区	21.66%	16.43%	18.65%	38.17%	2.56%	11.22%
雨花台区	0.09%	6.42%	18.60%	9.23%	10.58%	7.45%
总计	**100.00%**	**100.00%**	**100.00%**	**100.00%**	**100.00%**	**100.00%**

表 4：申请人类型统计表（纵向百分比）

计数项:主分类号首字母	主分类号首字母								
区属.y	A	B	C	D	E	F	G	H	总计
高淳区	9	34	7		25	17	22	13	127
鼓楼区	65	121	36	1	130	62	193	93	701
建邺区	21	116	48	6	68	75	106	52	492
江宁区	183	652	57	33	182	274	455	274	2110
江宁区(江开)	14	79	9		16	35	45	71	269
溧水区	29	159	12	2	39	70	37	32	380
六合区	9	139	12	3	29	21	27	30	270
六合区(化工园)	1	50	14		2	8	21		96
六合区(江北)	16	61	6	7	16	20	79	6	211
浦口区	11	56	9	1	44	8	39	54	222
浦口区(高新)	34	62	12	8	8	14	94	60	292
浦口区(江北)	89	126	24	7	24	62	114	71	517
栖霞区	46	137	17	2	89	53	165	71	580
栖霞区(新港)	9	120	22	12	18	46	42	38	307
秦淮区	54	143	100	4	80	61	149	77	668
玄武区	110	94	33	2	49	42	100	45	475
雨花台区	20	75	14		18	32	55	74	288
总计	720	2224	432	88	837	900	1743	1061	8005

表 5：专利分类统计表（数值）

计数项:主分类号首字母	主分类号首字母								
区属.y	A	B	C	D	E	F	G	H	总计
高淳区	1.25%	1.53%	1.62%	0.00%	2.99%	1.89%	1.26%	1.23%	1.59%
鼓楼区	9.03%	5.44%	8.33%	1.14%	15.53%	6.89%	11.07%	8.77%	8.76%
建邺区	2.92%	5.22%	11.11%	6.82%	8.12%	8.33%	6.08%	4.90%	6.15%
江宁区	25.42%	29.32%	13.19%	37.50%	21.74%	30.44%	26.10%	25.82%	26.36%
江宁区(江开)	1.94%	3.55%	2.08%	0.00%	1.91%	3.89%	2.58%	6.69%	3.36%
溧水区	4.03%	7.15%	2.78%	2.27%	4.66%	7.78%	2.12%	3.02%	4.75%
六合区	1.25%	6.25%	2.78%	3.41%	3.46%	2.33%	1.55%	2.83%	3.37%
六合区(化工园)	0.14%	2.25%	3.24%	0.00%	0.24%	0.89%	1.20%	0.00%	1.20%
六合区(江北)	2.22%	2.74%	1.39%	7.95%	1.91%	2.22%	4.53%	0.57%	2.64%
浦口区	1.53%	2.52%	2.08%	1.14%	5.26%	0.89%	2.24%	5.09%	2.77%
浦口区(高新)	4.72%	2.79%	2.78%	9.09%	0.96%	1.56%	5.39%	5.66%	3.65%
浦口区(江北)	12.36%	5.67%	5.56%	7.95%	2.87%	6.89%	6.54%	6.69%	6.46%
栖霞区	6.39%	6.16%	3.94%	2.27%	10.63%	5.89%	9.47%	6.69%	7.25%
栖霞区(新港)	1.25%	5.40%	5.09%	13.64%	2.15%	5.11%	2.41%	3.58%	3.84%
秦淮区	7.50%	6.43%	23.15%	4.55%	9.56%	6.78%	8.55%	7.26%	8.34%
玄武区	15.28%	4.23%	7.64%	2.27%	5.85%	4.67%	5.74%	4.24%	5.93%
雨花台区	2.78%	3.37%	3.24%	0.00%	2.15%	3.56%	3.16%	6.97%	3.60%
总计	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%

表 6：专利分类统计表（纵向百分比）

计数项:主分类号首字母	主分类号首字母								
区属.y	A	B	C	D	E	F	G	H	总计
高淳区	7.09%	26.77%	5.51%	0.00%	19.69%	13.39%	17.32%	10.24%	100.00%
鼓楼区	9.27%	17.26%	5.14%	0.14%	18.54%	8.84%	27.53%	13.27%	100.00%
建邺区	4.27%	23.58%	9.76%	1.22%	13.82%	15.24%	21.54%	10.57%	100.00%
江宁区	8.67%	30.90%	2.70%	1.56%	8.63%	12.99%	21.56%	12.99%	100.00%
江宁区(江开)	5.20%	29.37%	3.35%	0.00%	5.95%	13.01%	16.73%	26.39%	100.00%
溧水区	7.63%	41.84%	3.16%	0.53%	10.26%	18.42%	9.74%	8.42%	100.00%
六合区	3.33%	51.48%	4.44%	1.11%	10.74%	7.78%	10.00%	11.11%	100.00%
六合区(化工园)	1.04%	52.08%	14.58%	0.00%	2.08%	8.33%	21.88%	0.00%	100.00%
六合区(江北)	7.58%	28.91%	2.84%	3.32%	7.58%	9.48%	37.44%	2.84%	100.00%
浦口区	4.95%	25.23%	4.05%	0.45%	19.82%	3.60%	17.57%	24.32%	100.00%
浦口区(高新)	11.64%	21.23%	4.11%	2.74%	2.74%	4.79%	32.19%	20.55%	100.00%
浦口区(江北)	17.21%	24.37%	4.64%	1.35%	4.64%	11.99%	22.05%	13.73%	100.00%
栖霞区	7.93%	23.62%	2.93%	0.34%	15.34%	9.14%	28.45%	12.24%	100.00%
栖霞区(新港)	2.93%	39.09%	7.17%	3.91%	5.86%	14.98%	13.68%	12.38%	100.00%
秦淮区	8.08%	21.41%	14.97%	0.60%	11.98%	9.13%	22.31%	11.53%	100.00%
玄武区	23.16%	19.79%	6.95%	0.42%	10.32%	8.84%	21.05%	9.47%	100.00%
雨花台区	6.94%	26.04%	4.86%	0.00%	6.25%	11.11%	19.10%	25.69%	100.00%
总计	8.99%	27.78%	5.40%	1.10%	10.46%	11.24%	21.77%	13.25%	100.00%

表 7：专利分类统计表（横向百分比）

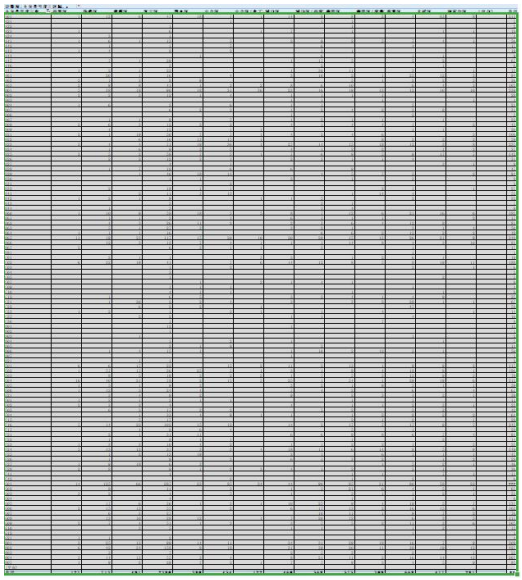


表 8：专利分类按区属统计表（1）

表 9：专利分类按区属统计表（2）

行业名称	计数项:主分类号前三位 区属.x2					
	主分类号前三位	高淳区	鼓楼区	建邺区	江北新区	总计
测量仪器	G01	14	125	66	359	564
一般的物理或化学的方法或装置	B01	5	38	18	138	199
发电、变电或配电 H02 131 22 6	H02	6	42	24	119	191
建筑物	E04	16	40	21	101	178
基本电器元件	H01	3	25	13	123	164
输送；包装；贮存；搬运薄的或细丝	B65	11	19	23	107	160
机床；其他类目中不包括的金属加工	B23	3	4	4	111	122
水、废水、污水或污泥的处理	C02	6	32	19	52	109
电通信技术	H04	3	14	13	46	76
家具；家庭用的物品或设备；咖啡磨； 香料磨；一般吸尘器	A47	1	2	1	43	47
医学或兽医学；卫生学	A61		20	1	20	41
	总计	68	361	203	1219	1851

表 10：专利行业归属统计表

区属.x2	计数项:主分类号前三位2	计数项:主分类号前三位
高淳区	272	2.59%
鼓楼区	894	8.51%
建邺区	577	5.49%
江宁区	3389	32.26%
江宁区(江开)	6	0.06%
溧水区	659	6.27%
江北新区	2847	27.10%
六合区(江北)	3	0.03%
浦口区(江北)	1	0.01%
秦淮区	811	7.72%
玄武区	645	6.14%
雨花台区	402	3.83%
总计	10506	100.00%

表 11：特定产业（电通信技术）统计表（数值+百分比）