**《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**。

1. **数据清洗与图表绘制**
2. **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)

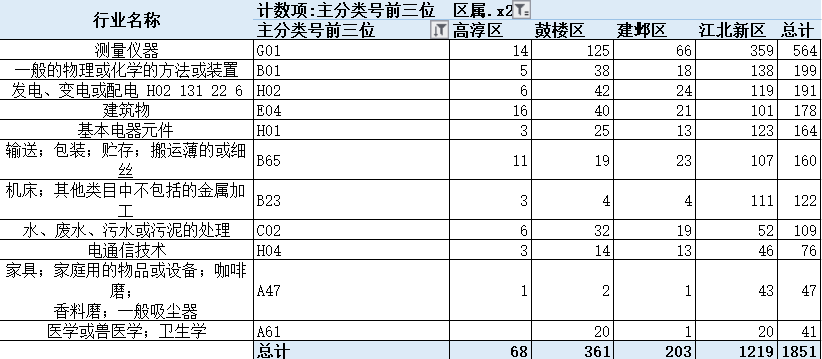
data[!index,]

write.csv(data,"E:\\Desktop\\data.csv")

1. **统计分析与图表绘制**

因图表过多，详见本word文档附录部分。

1. **简答题**
2. **怎样评估江北地区的核心技术优势？**



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。

1. **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。

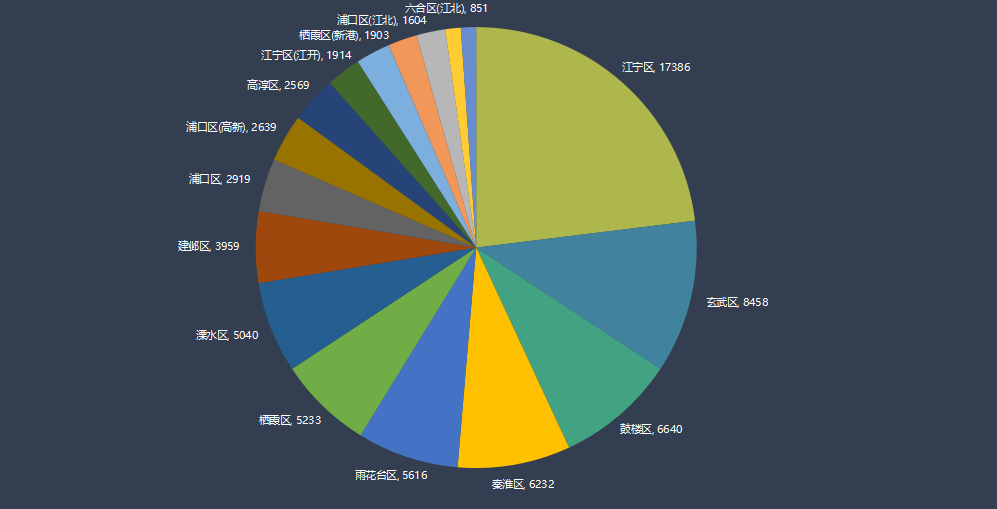
**【附录：实验结果】**

**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）**



**表1：专利类型统计表**



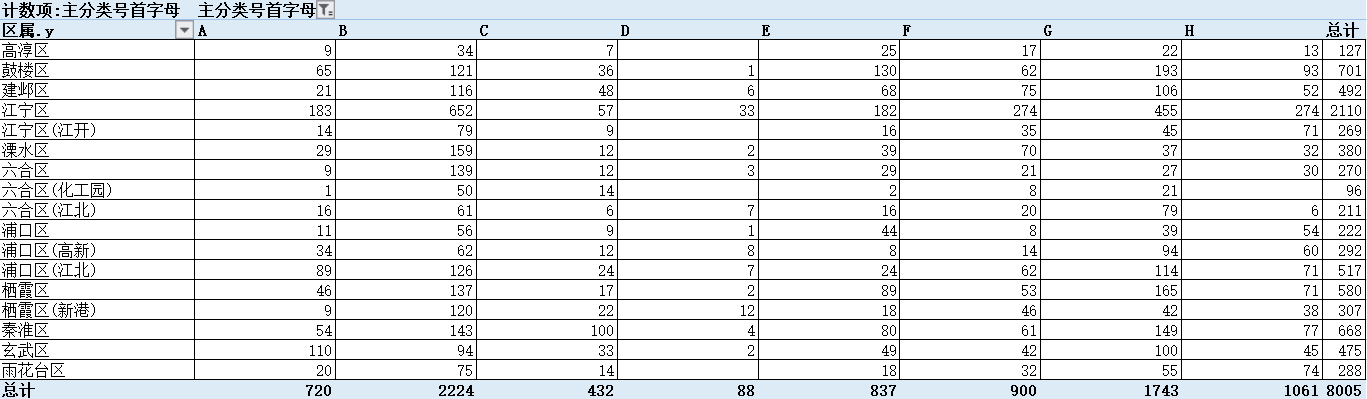
**表2：申请人类型统计表（数值）**



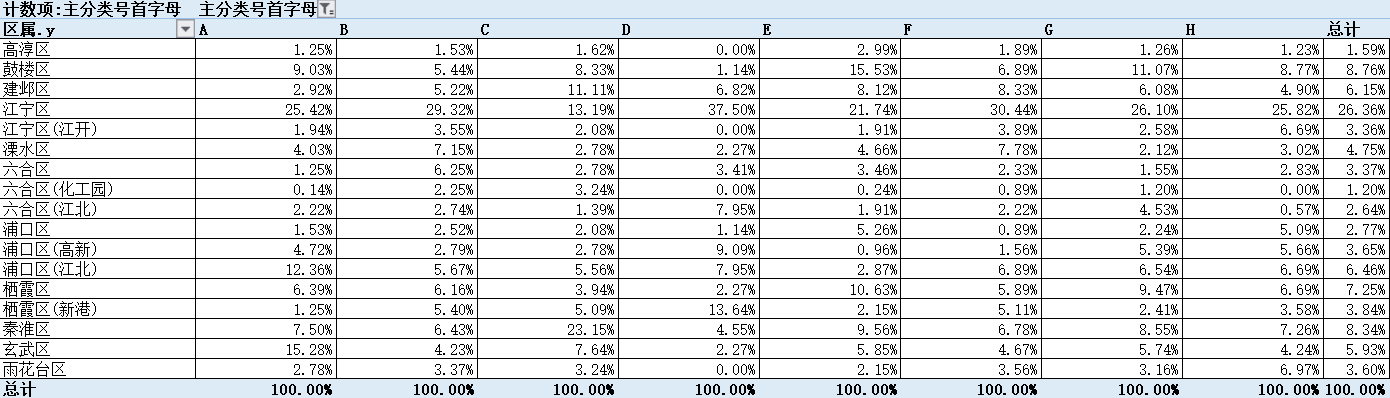
**表3：申请人类型统计表（横向百分比）**



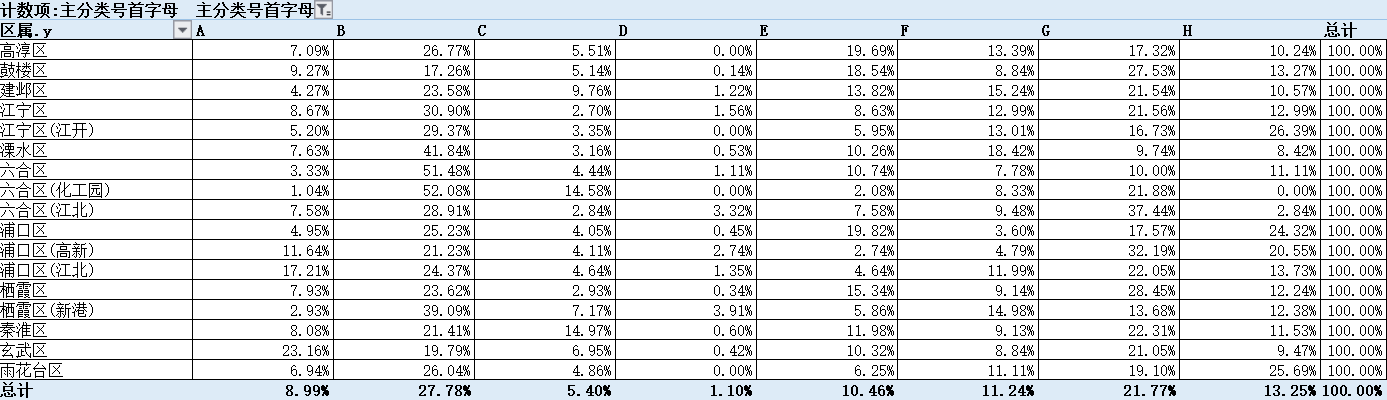
**表4：申请人类型统计表（纵向百分比）**



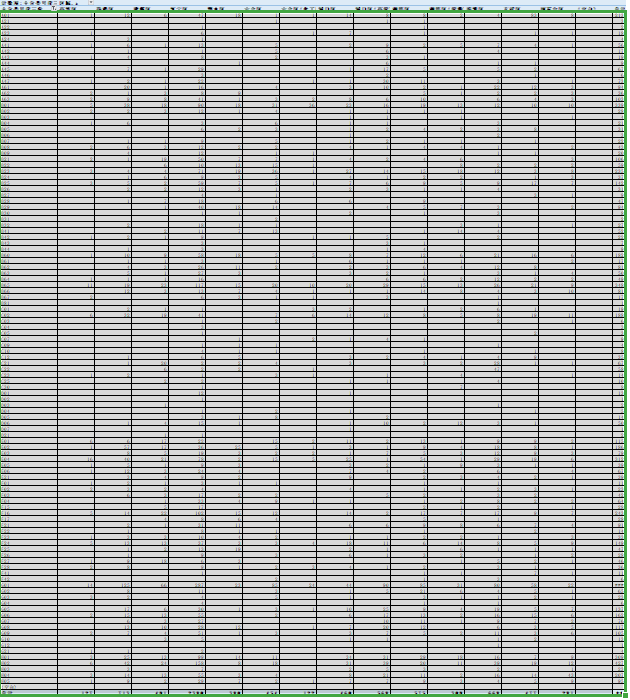
**表5：专利分类统计表（数值）**



**表6：专利分类统计表（纵向百分比）**



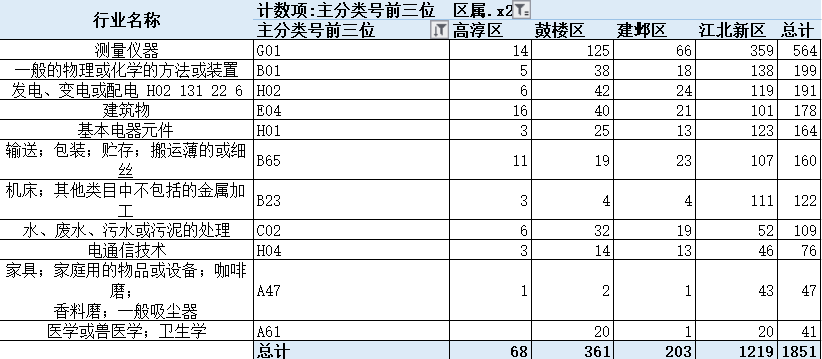
**表7：专利分类统计表（横向百分比）**



**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）**



**表10：专利行业归属统计表**



**表11：特定产业（电通信技术）统计表（数值+百分比）**