## R数据处理基础 实验四

一、实验目的

1. 从网上下载文件，联网获取数据

2. 数据预处理（dplyr和tidyr基本应用）

二、实验时间

2019年11月26日

三、实验要求

1. 使用R Markdown 生成word版实验报告，命名“学号+姓名+实验四.docx”。

2. 实验报告及相应的Rmd文件于2019年11月27日前上传至ftp：//202.204.121.57相应文件夹下。

四、实验内容

1. 在tidyr包中有一个who数据集，是世界卫生组织调查的结核病数据，这个数据集我们在第八讲的4-6页曾经提到过，这是一个不符合整洁数据的数据集，我们要把它变整洁。首先用who查看一下数据是什么样的(提示：>后直接输入who并回车)；然后查看who的说明（?who）；说明中有关于变量的描述，该数据集的60列中，第5-60列包含了3个变量，“new\_”后面的代码标识不同的诊断方法“method of diagnosis”（rel= relapse, sn= negative pulmonary smear, sp= positive pulmonary smear, ep= extra pulmonary），数字前面紧跟的是性别代码（m，f），数字代表年龄分组（0-14，15-24，25-34…55-64，65+）。
2. 假如我们对采用relapse方法检测到的记录感兴趣，选取字段iso2，year，newrel\_m014:newrel\_f65；
3. 试用gather和separate将它转成整洁的数据集。可以分步做，最后的结果数据集包含6个字段: iso2，year，method=”rel”，gender=”m”/”f”，agegroup=(”014”…”65”) ，count。

提示：你需要用到的函数有select()，gather()，separate()，mutate()，substr()， length()，如果不清楚用法，可以用类似?substr查看。

1. 从网上下载数据，教务处网上的“关于公布大学生创新创业训练项目结题验收结论的通知”里最下面有一个“附件2:北京林业大学生创新创业训练项目结题验收汇总表.xlsx”（ http://jwc.bjfu.edu.cn/docs/2019-07/20190705110207787523.xls），
2. 试着下载该文件并保存到当前目录下，命名为”IETP.xlsx”;

library(xlsx）

fileUrl <-<http://jwc.bjfu.edu.cn/docs/2019-07/20190705110207787523.xls>

download.file(fileUrl,destfile= " IETP.xlsx ", mode= " wb " )

1. 将下载后的文件读进R，注意该表是中文的，先不指定encoding=参数，试着读取该表并赋给一个名为ietp的数据框，用head（）函数查看读进来的记录是否乱码？然后利用encoding=”UTF-8”参数重新读入数据赋给一个名为ietp1的数据框，并查看读入的数据是否正确，如果换成“GB2312”编码行吗？

ietp<-read.xlsx("IETP.xlsx",sheetIndex = 1, header=TRUE)

head(ietp)

ietp1<-read.xlsx("IETP.xlsx",sheetIndex = 1, header=TRUE, encoding = "UTF-8")

View(ietp1)

ietp2<-read.xlsx("IETP.xlsx",sheetIndex = 1, header=TRUE, encoding = " GB2312")

View(ietp2)

1. 用一个表格列出各学院“国家级”、“北京市级”、“校级”项目数(group\_by函数或其他办法)；将汇总结果存入相应的数据框中并存成相应的以逗号分隔的文本文件。

write.table(mydata, "c:/mydata.txt", sep="\t")