Business Analytics Homework 1

Instructor: Zach Zhizhong ZHOU

1. 假设你手头有 3 个数据库，分别是
数据库 1：上海市所有楼盘 2014 年每天每个小时时段的用电量。
数据库 2：上海市所有交通监控设备搜集到的 2014 年每天每分钟时段的车流量。
数据库 3：上海市所有地铁站出入口搜集到的 2014 年每天每分钟时段的人流量。
使用这 3 个数据库，你认为可以从中提取出何种有价值的商业信息？（不一定 3 个全用，可以使用其中 1 个或者 2 个）

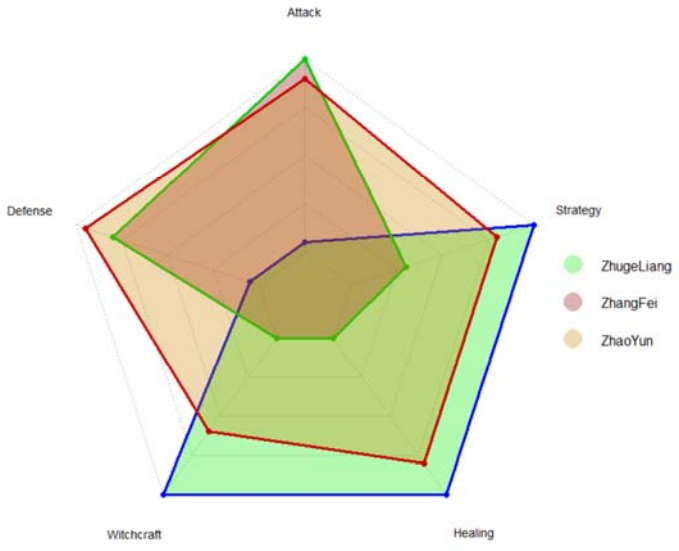
2. 在 D 盘建立目录：D:/BA/Homework/HW01，将课程文件夹/BA/Homework/HW01 中的所有文件拷贝到 D 盘新建目录 D:/BA/Homework/HW01 中（也可以把本邮件的所有附件拷贝到新建目录中）。PengPai_HTML.txt 是澎湃新闻网站的首页源代码。打开 PengPai.R 文件，这是一个不完整的 R Script 代码文件。该代码文件中含有中文注释，使用 GB2312 编码。你应该在 R Studio 当中选择 Tools --> Global Options... --> Default text encoding: GB2312，这样中文注释才不是乱码。如果你在 R Studio 中选择默认编码是其他编码，那么在打开该 R Script 代码文件之后再选择 File-->Reopen with Encoding...-->GB2312。
完整的代码文件是从澎湃新闻首页上提取每天热门新闻、每三天热门新闻以及每周热门新闻的标题。该代码文件在以下几行缺了一些代码，请你补充完整：
第 7 行：Hot00 <- xpathSApply(doc, "//ul[ ]/li/a" ,xmlValue) #提取每日热门新闻
第 10 行：Hot01 <- xpathSApply(doc, ,xmlValue) #提取三天内热门新闻
第 13 行：Hot02 <- xpathSApply(doc, ,xmlValue)
请你将这些缺失的代码填上。运行完整的代码文件之后，提取出的热门新闻标题存入 HotTopics.csv 中。运行代码之后得到的 csv 文件应该和课程文件夹/BA/Homework/HW01 中的 HotTopics.csv 是一样的。
最后用 R 语言代码计算 30 条热门新闻标题当中有多少条是不重复的（答案是 22 条）。将计算不重复新闻标题的代码写在 R 语言代码文件当中：
Hot <- c(Hot00,Hot01,Hot02)
#这一条语句计算结果应该是 22 条。

3. 下面是诸葛亮、张飞和赵云各项指标数值：
诸葛亮：攻击 5，防守 5，妖术 100，医术 100，谋略 100
张飞：攻击 100，防守 80，妖术 0，医术 0，谋略 30
赵云：攻击 90，防守 95，妖术 60，医术 60，谋略 80。
SpiderPlot_SanGuo.R 是不完整的代码文件，当中删掉了几段代码，请你直接在代码文件上把被删掉的代码补齐。

```
library(fmsb)
MyData <- data.frame(Name = c('ZhugeLiang','ZhangFei','ZhaoYun'),
                     Attack=c(5,100,90),
                     Defense=c(5,80,95),
                     Witchcraft = c(100,0,60),
                     Healing= c(100,0,80),
```

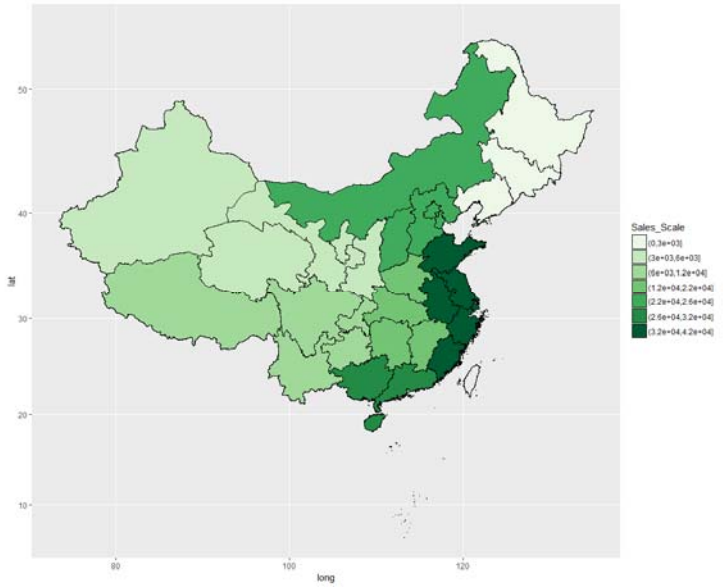
```
                                Strategy=c(100,30,80),stringsAsFactors=FALSE)
MinScore =0; MaxScore=100;
[redacted]
#define the max and min value for each column.

MyColor <- function (Myclr, ClrTransparency) {
   TT <- col2rgb(Myclr)/255;
   return(rgb(TT[1],TT[2],TT[3],ClrTransparency))
}
#define a function to caculate RGB Hex constants for a specific color with a specific transparency
MyColor("lightyellow",0.2);

colors_border=c( "blue", "green3", "red3")
colors_in=c( MyColor("green2",0.3) , MyColor("red4",0.3) , MyColor("orange3",0.3))
[redacted]
```

画出来的蛛网图效果应该是这样的：



4. 某企业在中国内地各个地区的产品销售情况如下：
华东地区（包括山东、江苏、安徽、浙江、福建、上海）：40,000
华南地区（包括广东、广西、海南）：30,000
华中地区（包括湖北、湖南、河南、江西）：20,000
华北地区（包括北京、天津、河北、山西、内蒙古）：25,000

西北地区（包括宁夏、新疆、青海、陕西、甘肃）：5,000
西南地区（包括四川、云南、贵州、西藏、重庆）：10,000
东北地区（包括辽宁、吉林、黑龙江）：2,000
将销量数据画成中国分区统计图。画分区统计图前应将销量取值划分成 7 个区间，分别是(0,3000], (3000,6000], (6000,12000], (12000,22000], (22000,26000], (26000,32000], (32000,42000]。分区统计图应使用深浅不同的绿色画出这 7 个取值区间，取值越高的区间绿色越深（见下图）。注意中国地图不要漏掉台湾。
你会用到课程文件夹/BA/CHNMap/当中的 ProvIndex.csv 和 ProvMapData.csv 数据文件，但是中国各大区销量的数据文件需要你自己填写。



提交作业的截止时间：11 月 2 日周三下午 17:00 分。
应该通过电子邮件提交到 zachzhoucourse@hotmail.com。提交作业的邮件标题是：“BA HW01 你的名字”。
提交文件列表：
1、 第 1 题的解答，将解答放在一个 Word 或者 PDF 文档里面，将文档命名为 HW01_Q1
2、 PengPai.R 完整代码。
3、 SpiderPlot_SanGuo.R 完整代码。
4、 画图的完整代码，将代码文件命名为 CHN_Sales.R。
以上代码确保放在以下目录当中可以成功执行：D:/BA/ Homework/HW01。如果你修改工作目录，导致代码无法在上面的目录中执行，TA 将不会检查你的代码而是直接给你 0 分！

你可以独立完成作业，也可以组队完成，每个队伍人数不得超过 4 人。人数越少的队伍，在期末总评时会有适当加分奖励（奖励 0.5 – 3.5 分，最多可以将你的成绩提升一个绩点比如从 B+提高到 A-，但不会提升 2 个或者以上的绩点）。