**三．操作实践部分：（共70分）**

根据题目要求使用R完成相关的任务。包括读入相关数据，进行数据处理，然后选择恰当的分析方法进行相应的分析和检验。除非电脑问题，结果以rmarkdown文件方式提交，除代码外，有简单的说明和注释。

文件名为姓名+学号，如“R2017李丁20110102.Rmd”，考试结束前作为附件发送到 [liding@ruc.edu.cn](mailto:liding@ruc.edu.cn) ，邮件主题为：“R2017考试姓名20110100”（姓名+学号）

**题目1（28分）：**

（1）请把下面的向量组合成一个数据框pdata（3分）。

（2）将其中的diebetes变量定义为无序因子变量（3分）

（3）先将status定义为有序因子变量（5分）。

（4）将年龄等于52的案例的diabetes取值改为 Type2（3分） 。

（5）将变量名patientID改为pid（3分）。

（6）增加一个变量age2等于age-18（3分）。

（7）对status变量进行recode，将Poor和Improved合并为 Fair（5分）。

（8）设定一个工作目录，并将数据框pdata写出成为pdata.csv文件（3分）。

patientID <- c(1, 2, 3, 4)

age <- c(25, 34, 28, 52)

diabetes <- c("Type1", "Type2", "Type1", "Type1")

status <- c("Poor", "Improved", "Excellent", "Poor")

**题目2（42分，每题6分）：**

考试数据来源于某商家219位用户2007年-2008年24个月5类刷卡消费（Grocery、Retail、Entertainment、Travel、Other）支付次数和支出总金额数据，每个人有（2\*12\*5条记录）。相关变量包括出生日期，性别，主卡类型，主卡申请日期等（这些信息对于同一个用户都是相同的）。我们想知道消费者前后两年的消费之间是否存在相关，不同特征的用户消费否存在差异。

（1）用R读入这个SPSS数据（提示：使用haven包和foreign包都需要注意有些变量的格式）

（2）在上述数据的基础上，筛选出Entertainment类的支出记录。尝试用ggplot作图，绘制出2007年和2008年各月份Entertainment支出的平均值的对比图（鼓励采用管道操作）。

（3）在第（1）问数据基础上汇总得到每个用户2007年和2008年的总支出，支出总次数，保留个体不变的那些基本信息（提示：得到的数据每个用户只有两条记录），存为一个新数据cred2。

（4）在cred2中新增一个变量表示每个客户每笔支出的单价（pspent=spent/items）。

（5）这些持卡客户2007年和2008年的支出总额均值存在差异吗？每笔支出单价的均值呢？请用合适的检验方式进行检验。

（6）在（4）更新的数据基础上，将长数据变成宽数据wcred。形成spent2007 spent2008 items2007 items2008 pspent2007 pspent2008之类的变量，注意保留其他基本信息。

（7）基于（6）得到的数据，你可以尝试探索哪些因素会影响pspent2008，你可以尝试分类汇总、作图、统计检验、回归分析等多种方法。

**题目3（附加题，20分，做对可得分，但卷面所得总分不超过100分）：**

尝试利用R爬虫提取 <http://bj.xiaozhu.com/fangzi/24730694703.html> 页面中的如下信息：标题、房源小区、单价、评价得分、房主名。

