# 周恩来政府管理学院本科生 2018——2019 学年第一学期
# 《R 语言统计入门》课程期中作业

专业：_____学号：_____ 姓名：_____ 成绩：_____

## 一、数据输入（5 分）

请将下表在 R 中输入成 data frame 的格式，命名为 prac01。

| ID | Major | mid | final | total | Male |
|---|---|---|---|---|---|
| 1 | 英语 | 83 | 85 | 84 | TRUE |
| 2 | 法语 | 88 | 84 | 86 | TRUE |
| 3 | 社会工作 | 82 | 86 | 84 | FALSE |
| 4 | 应用心理学 | 89 | 88 | 88 | FALSE |
| 5 | 应用心理学 | 86 | 86 | 86 | TRUE |

**答案：**

## 二、命令练习（5 分）

（一）在 # 号后面解释前述命令的功能。

```
seq(1, 20, by = 2)  #
LETTERS[1:5] #
letters[5:1] #
rep(LETTERS[1:3], each = 3) #
rep(LETTERS[1:3], times = 3) #
dnorm(-2, 10, 2)  #

pnorm(-2, 10, 2)  #

pnorm(2.5)-pnorm(2)  #

rnorm(10, 20, 5)  #

qnorm(0.25, 10, 2)  #

cbind(dataA, dataB)  #

rbind(data, dataB)  #

quantile(x, probs = c(0.1, 0.6, 0.9)) #

scale(x, center = T, scale = T)  #

min(x)  #                    max(x)  #
```

```
mean(x)  #                    sd(x)

median(x)  #                  cor(x, y)

IQR(x)  #                     choose(n, k)  #

summary(x)                    factorial(n)  #

fivenum(x)                    options(digits = 5)  #

var(x)                        round(x)  #


choose(10, 6) * factorial(6)  #
```

（二）总结 Hadley Wickham 开发的 readxl 和 haven 包中的常用文件导入函数，可参考其个人网页整理。

http://hadley.nz/

## 三、数据操纵（45 分，前两题各 5 分，后两题各 15 分）

（一）安装 readxl 包，读入 rs2015.xlsx 文件，取名为 rs2015。该文件为 2015 年选修《R 语言统计应用入门》的所有同学的期中成绩（mid）和期末成绩（final）。写出读入命令的语句，并结合 dplyr 包完成以下任务：

（1）生成新变量“总评成绩”（total），其计算公式为期中成绩的 40%，加上期末成绩的 60%。注意总评成绩需要四舍五入为整数。

（2）生成新变量“等级成绩”（rank），取值方式如下：

| 总评成绩（total） | 等级成绩（rank） |
|---|---|
| 90~100 | A |
| 80~89 | B |
| 70~79 | C |
| 60~69 | D |
| 0~59 | E |

（3）生成新的虚拟变量 psy，凡是原文件中 major 为 PSY 的取值为 1，其余为 0。

（4）所有结果，先按 psy 变量升序排序，再按 total 变量降序排序。

（5）将上述结果写出为新的 .csv 文件，命名为 rs2015_arranged.csv。

（6）无放回地从心理学和非心理学的同学中，各自随机抽取 5 名同学，结果命名为 samples_by_psy。

（7）选出所有 major 为 PSY 的同学，另存为一个新数据框，命名为 psystudents。

**答案：**

（二）中国哪个省份的“胖子”最多？据中国疾病控制预防中心某年的一份报告，中国各省份中，肥胖率排名中天津排名第一，达到 21.2%。其中是否肥胖以 BMI 指数判定。按中国卫生与计划生育委员会 2013 年颁布的标准：
http://www.moh.gov.cn/ewebeditor/uploadfile/2013/08/20130808135715967.pdf
中国人的 BMI 指数取值与对应判定区间如下：

| **BMI** | **分类** |
|---|---|
| BMI < 18.5 | 体重过低 |
| 18.5 ≤ BMI < 24.0 | 体重正常 |
| 24.0 ≤ BMI < 28.0 | 超重 |
| BMI ≥ 28.0 | 肥胖 |

由于未能获得中国疾病控制中心的原始数据，这里拟采用 CGSS 2013 数据进行验证。该数据（`cgss2013.dta`）的 `s41` 变量为省份信息，省份编码与对应省份名称见文件 `label_cgss2013s41.csv`；`a13` 为身高（单位：cm）；`a14` 变量为体重（单位：斤，注意不是 kg，1 斤 = 0.5 kg）。试利用这两个文件，并结合 dplyr 包，计算各省份的肥胖率并进行排名。

（三）数据 PD201801.xls 是南开大学汪新建教授主持的《医患信任建设的社会心理机制研究》这一项目中《医患社会心态问卷：患方卷》的部分试调查数据，该数据为选修本课的相关同学自发调查所得的原始数据。该数据使用问卷星平台获得，下载 Excel 格式后我们将部分基础信息提取保存。部分数据显示如下：

| 序号 | 提交答卷时间 | 所用时间 | 来源 | 来源详情 | 来自 IP | …… |
|---|---|---|---|---|---|---|
| 1 | 2018/11/27 9:04:46 | 21 秒 | 链接 | 直接访问 | 117.131.219.42(天津-天津) | …… |
| 2 | 2018/11/27 9:05:44 | 16 秒 | 链接 | 直接访问 | 117.131.219.42(天津-天津) | …… |
| 3 | 2018/11/27 9:12:11 | 475 秒 | 链接 | 直接访问 | 106.47.113.139(天津-天津) | …… |
| 4 | 2018/11/27 9:27:05 | 14 秒 | 微信 | N/A | 106.47.228.5(天津-天津) | …… |
| 5 | 2018/11/27 9:27:26 | 17 秒 | 微信 | N/A | 106.47.228.5(天津-天津) | …… |

请完成以下问题：

（1）为了使问卷编码更为标准化，我们拟将所有变量名统一命名为英文，具体对应关系如下：

| 原变量名 | 新变量名 |
|---|---|
| 序号 | id |
| 提交答卷时间 | time1 |
| 所用时间 | time2 |
| 来源 | source1 |
| 来源详情 | source2 |
| 来自 IP | ip |
| 医德好的医生会对所有患者一视同仁 | dpt01 |
| 医生的很多做法是为了少担责任 | dpt02 |
| 即使声誉好的医院，医生也不一定敬业 | dpt03 |
| 医生对熟人会更尽心尽力 | dpt04 |
| 如果治疗出问题医院肯定会偏向医生 | dpt05 |
| 很多医生都是向钱看 | dpt06 |
| 普通人无法判断医生在治疗中是否尽力 | dpt07 |
| 医生会全力救治病人 | dpt08 |
| 跟医生熟识可以获得更好的医疗 | dpt09 |

请使用一条语句替换原始变量名。

注：在 PD201801.xls 的第二个 sheet 中有如上对应表。

（2）编制者在试调查之前已找少部分被试进行了预调查，估计较为认真作答的被试，一般应至少有 10 分钟左右方可填完问卷。被试的回答时间集中于 20~30 分钟以内，除了极少数老年被试，一般回答时长不会超过 60 分钟。鉴于此，规定此次问卷回答时长在 10~60 分钟之间（包含端点值）为有效问卷。请据此计算有效问卷数与有效问卷率。

（3）新生成两个变量，一为 province（省份），一为 city（省会城市或地级市），此部分信息均在 ip 一列中可以找到。统计各省份、各城市有多少被试参与调查。如果是直辖市（如天津-天津），不再区分下辖区，即 province 和 city 取值相同。

（4）在 PD201801.xls 中 dpt01-dpt09 这 9 个变量，实为我们编制的医患信任量表（患方版）的部分测量题目。该量表为 5 点计分式 Likert 量表，其中 dpt01 和 dpt08 的得分越高，则信任度越高；其他题项，得分越高，则信任度越低。这是一个总加量表，最后要统计被调查者的回答总分，并要求总分越高，信任度越高。这就涉及反向计分的过程，即除了 dpt01 和 dpt08 外，都需要用“6-原始分”的方式进行计分，最后再将此 9 道题目的总分累加得出个体的医患信任得分。新生成变量 dpt_total，用于储存每个个体的医患信任得分。注：如有负值（如-3），表示问卷星系统中的缺失值，请统一处理为 NA。

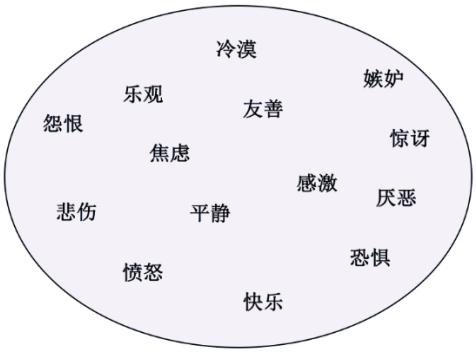
答案（请附代码）：

（四）在 PD201802.xls 这一文件中，储存了关于医患社会情绪感知的回答信息，其基本形式如下：

| 序号 | 您是否有以下感受？…… |
|---|---|
| 1 | (跳过) |
| …… | …… |
| 9 | 冷漠〖5〗┇焦虑〖7〗┇愤怒〖6〗 |
| …… | …… |
| 22 | 感激〖对医生心存感激〗┇友善〖医生态度友善〗┇平静〖平静的接受〗 |
| …… | …… |
| 25 | 感激〖1-1-1〗┇乐观〖10-10-10〗┇友善〖10-10-10〗 |

读入此文件并保存为同名对象，并将变量名“序号”改成“id”,“……感受”改成“emotions”。在此基础上，完成如下工作。此题的原题形式如下：

请回顾您最近 6 个月来在“医患关系”上的亲身经历或所见所闻，您是否有以下感受？请在大圆中圈出 3 个您最先想到的情绪名词。




现在请在您选好的情绪名词后面的线段数字上画“√”标明您的感受程度。数字越大，感受程度越强烈。

怨恨　1　2　3　4　5　6　7　8　9　10

感激　1　2　3　4　5　6　7　8　9　10

……

由于问卷星系统的问题，其输出的答案即是 PD201802.xls 所示形式。

（1）要求新生成 14 个变量，变量名即为上述 14 个情绪词的名称，然后要求将 PD201802.xls 转化为如下形式：

| ID | 怨恨 | 感激 | 悲伤 | 乐观 | 冷漠 | 友善 | …… |
|---|---|---|---|---|---|---|---|
| 1 | NA | NA | NA | NA | NA | NA | …… |
| 2 | NA | 6 | NA | 7 | NA | 9 | …… |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| 3 | 6 | NA | 7 | NA | NA | NA | …… |
| 4 | NA | NA | NA | 6 | NA | 7 | …… |
| 5 | NA | NA | NA | 9 | NA | NA | …… |
| 6 | NA | NA | NA | NA | 7 | NA | …… |
| 7 | NA | NA | 1 | NA | NA | NA | …… |
| 8 | 4 | 2 | NA | NA | 7 | NA | …… |

也就是说，如果被调查者没有选中该情绪词，则取值为 NA；如选中，则取值为被调查回答的情绪强度值。例如，前面第 9 个被调查者答案为：

9　冷漠〖5〗┇焦虑〖7〗┇愤怒〖6〗

这说明他/她选中了冷漠（强度为 5）、焦虑（强度为 7）和愤怒（强度为 6）这三个情绪词。如果出现第 22 个被调查的回答：

22　感激〖对医生心存感激〗┇友善〖医生态度友善〗┇平静〖平静的接受〗

即本来应当填入分值，但却填入文字，则统一处理为 NA。如出现第 25 个被调查者的回答：

25　感激〖1-1-1〗┇乐观〖10-10-10〗┇友善〖10-10-10〗

也统一处理为 NA。即只保留能够准确识别其分值的个案。

（2）统计各情绪词的选中频次，并做成列表，按选中频次从高至低排序。

（3）阅读《中国人的医患社会情绪体验及其影响因素》一文（PDF 另行提供），思考其中关于医患社会情绪的分析方式。请对此文的分析方法做出优缺点评价并提出进一步的分析建议。

吕小康, 刘颖, 汪新建, 张慧娟, 张子睿.（2018）. 中国人的医患社会情绪体验及其影响因素. 载王俊秀主编，《中国社会心态研究报告（2018）》，pp. 146－175. 北京：社会科学文献出版社。

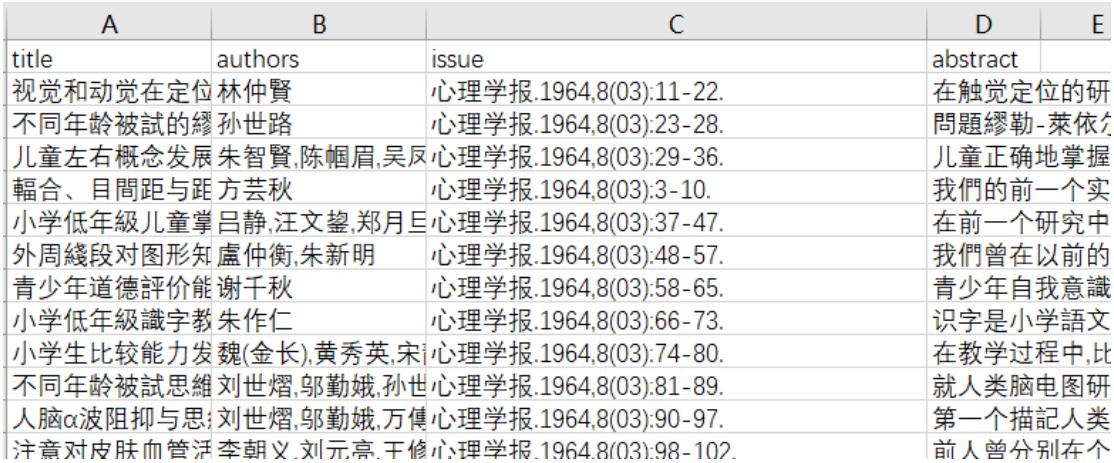
答案（请附代码）：

# 四、爬虫与文本分析（40 分，每题各 20 分）

（一）爬取如下四个期刊网站至创刊起（以网站实际提供的期数为准）至 2018 年底所有期的文章：

| 期刊名称 | 官方网址 |
|---|---|
| 《社会学研究》 | http://www.shxyj.org/ |
| 《社会》 | http://www.society.shu.edu.cn |
| 《心理学报》 | http://journal.psych.ac.cn/xlxb/CN/0439-755X/home.shtml |
| 《心理科学进展》 | http://journal.psych.ac.cn/xlkxjz/CN/1671-3710/home.shtml |

要求包含标题（title）、作者（authors）、卷期号与页码（issue）及摘要（abstract）这四项信息，并储存为如下形式（以《心理学报》为例）：



| A | B | C | D | E |
|---|---|---|---|---|
| title | authors | issue | abstract | |
| 视觉和动觉在定位 | 林仲賢 | 心理学报.1964,8(03):11-22. | 在触觉定位的研 | |
| 不同年龄被試的繆 | 孙世路 | 心理学报.1964,8(03):23-28. | 問題繆勒-萊依 | |
| 儿童左右概念发展 | 朱智賢,陈帼眉,吴 | 心理学报.1964,8(03):29-36. | 儿童正确地掌握 | |
| 輻合、目間距与距 | 方芸秋 | 心理学报.1964,8(03):3-10. | 我們的前一个实 | |
| 小学低年級儿童掌 | 吕静,汪文鋆,郑月 | 心理学报.1964,8(03):37-47. | 在前一个研究中 | |
| 外周綫段对图形知 | 盧仲衡,朱新明 | 心理学报.1964,8(03):48-57. | 我們曾在以前的 | |
| 青少年道德評价能 | 谢千秋 | 心理学报.1964,8(03):58-65. | 青少年自我意識 | |
| 小学低年級識字教 | 朱作仁 | 心理学报.1964,8(03):66-73. | 识字是小学語文 | |
| 小学生比较能力发 | 魏(金长),黄秀英,宋 | 心理学报.1964,8(03):74-80. | 在教学过程中,比 | |
| 不同年龄被試思維 | 刘世熠,邬勤娥,孙世 | 心理学报.1964,8(03):81-89. | 就人类脑电图研 | |
| 人脑α波阻抑与思 | 刘世熠,邬勤娥,万 | 心理学报.1964,8(03):90-97. | 第一个描記人类 | |
| 注意对皮肤血管活 | 李朝义,刘元亭,王修 | 心理学报.1964,8(03):98-102. | 前人曾分别在个 | |

将爬虫结果分别储存为 papers_xb、papers_jz、papers_socrs、papers_soc 四个对象，并写出为 csv 文档。注：xb、jz、socre 和 soc 分别表示《心理学报》、《心理科学进展》、《社会学研究》、《社会》四个期刊。

**注：此题仅需要附《心理学报》和《社会学研究》两个网站的爬虫命令即可。**

（二）对 `papers_xb.csv` 等文件，利用 stringr 包，提取出期刊名（journal）、年份（year）、卷号（vol）、期号（no）、起始页（start_page）、终止页（end_page）、文章占据的页数（pages）的信息，均另存为一列，并按年份、卷号、期号的顺序升序排序，最后结果另存为 `papers_xb_arranged.csv` 的数据框，如下所示：

| title | authors | issue | abstract | journal | year | vol | no | start_page | end_page | pages |
|---|---|---|---|---|---|---|---|---|---|---|
| 发挥集体 | 潘菽 | 心理学报. | 中国心理 | 心理学报 | 1956 | 1 | 0 | 3 | 12 | 10 |
| 关于心理 | 朱智賢 | 心理学报. | 在苏联,关 | 心理学报 | 1956 | 1 | 0 | 13 | 22 | 10 |
| 关于心理 | 臧玉海 | 心理学报. | 現在心理 | 心理学报 | 1956 | 1 | 0 | 23 | 28 | 6 |
| 关于心理 | 潘菽 | 心理学报. | 关于心理 | 心理学报 | 1956 | 1 | 0 | 29 | 36 | 8 |
| 人底心理 | 叶麐 | 心理学报. | 我們所提 | 心理学报 | 1956 | 1 | 0 | 37 | 47 | 11 |
| 对唯心主 | 荆其誠,叶 | 心理学报. | 感觉是認 | 心理学报 | 1956 | 1 | 0 | 49 | 59 | 11 |
| 刺激过程 | 曹日昌 | 心理学报. | 感觉的过 | 心理学报 | 1956 | 1 | 0 | 61 | 71 | 11 |
| 在用电流 | 沈迺璋,邵 | 心理学报. | 在本实驗 | 心理学报 | 1956 | 1 | 0 | 73 | 85 | 13 |
| 远近差交 | 刘范,彭瑞 | 心理学报. | 不同的訓 | 心理学报 | 1956 | 1 | 0 | 87 | 96 | 10 |
| 运动动力 | 李家治,赫 | 心理学报. | 巴甫洛夫 | 心理学报 | 1956 | 1 | 0 | 97 | 108 | 12 |
| 德意志民 | 陈立,曹日 | 心理学报. | <正>一.一 | 心理学报 | 1957 | 1 | 0 | 4 | 13 | 10 |
| 兒童第一 | 吳江霖,刘 | 心理学报. | <正>巴甫 | 心理学报 | 1957 | 1 | 0 | 14 | 30 | 17 |
| 詞在兒童 | 张述祖,詹 | 心理学报. | <正>問題 | 心理学报 | 1957 | 1 | 0 | 31 | 39 | 9 |
| 预测运动 | 曹日昌,荆 | 心理学报. | <正>一. | 心理学报 | 1957 | 1 | 0 | 40 | 54 | 15 |

注：在形如如下信息中，

心理学报. 1956, 1(00): 97-108.

1 为卷号，（）中的 00 为期号。即年份后的数字为卷号，括号中的数字为期号， - 符号两侧的数字为起止页。

请制作四个清洁的_arranged.csv 文件，并据此统计四个期刊所有文章的每篇文章的（1）平均作者数及年代变化趋势（即分年度统计平均作者数），（2）平均的标题（包括正副标题和标点）字数及其年代变化趋势。

此题仅需附一个文件的分析代码即可，然后即可给出结果。

## 五、开放性问题（5 分，不少于 100 字）

此题请用电脑打字，最后请说明电脑统计的字数（不计空格）。

结合自身的专业背景，谈一谈你对大数据技术对本专业研究的应用前景，以及 R 在其中可扮演的角色。并由此谈一谈你对本专业数据处理类课程设置、教学内容、教学方式的建议。