**武汉大学经济与管理学院**

**2020-2021学年第二学期《R语言数据分析方法与实验》  
期末考试试题**

**注意事项**

1. 交卷格式为：代码+输出结果（可使用Word、PDF、R markdown，要求一个代码对应一个结果），code务必可以复制运行
2. 文件命名为“学号姓名”，例如“2018301051188张三”

3.  提交方式为腾讯微云，具体链接由主考老师安排

4.  没有按时提交者考试为0分，请大家控制好时间

**第一题 (共35分)**

探索nycflights13数据集

1. 从flights数据中找出到达时间延误2小时或者更多的所有航班，并将生成的新数据保存为flight\_arr2hr。(5分)
2. 将生成的flight\_arr2hr数据集根据目的地（dest）进行分组，统计出抵达每个目的地的航班数量，筛选出抵达航班数量前十名的目的地，将结果命名为top10\_dest。(5分)
3. 从weather表中挑选出以下变量：year, month, day, hour, origin, humid, wind\_speed，并将其与flight\_arr2hr表根据共同变量进行左连接, 生成的新数据保存为flight\_weather (5分)
4. 基于flight\_weather数据集，根据不同出发地(origin)在平行的三个图中画出风速 wind\_speed（x轴）和出发延误时间dep\_delay（y轴）的散点图，以及平滑曲线。 (5分)
5. flights中每家航空公司在2013年有多少班次的航班被取消了？提示：依据dep\_time来判断某班次航班是否被取消(5分)
6. 找出flights中每一家航空公司的航班最常去的目的地机场，以及flights中每家航空公司飞往最常去的目的地机场的航班数量 (10分)

**第二题（共20分）**

探索diamonds数据集

1. 对diamonds数据集，生成一个新变量id，用于存储每条观测值所在的行数。挑选出id, x, y, z四个变量，将宽数据转换为长数据：将x, y, z的变量名存为新变量dimension，将x, y, z的值存为新变量length。转换后的长数据存为xyz\_long。（5分）
2. 将xyz\_long数据集转换回宽数据xyz\_wide。宽数据xyz\_wide包含id, x, y, z四个变量。（5分）
3. 写代码找出diamonds中最常见和最不常见的color，即出现次数最多和最少的color。然后，按照下列规则对color列重新定义——将color中其它颜色的因子合并为“Other Color”，最常见和最不常见的color分别设置为“Common”和“Rare”，将新生成的数据框保存为diamonds\_color\_redefine。最后，再将“Common”设置为diamonds\_color\_redefine中color因子的基准水平(reference level)。（10分）

**第三题 （共25分）**

请加载 stringr 包

1. 根据stringr::words数据，统计每个单词的长度，按照其长度的中位数分为 “短单词” 和“长单词”两类，然后统计出每个单词的元音个数，以及元音比例（=元音个数/单词长度），将上述生成的数据保存为word\_type。该数据的变量名依次为word, word\_length, word\_type, num\_vowel, proportion\_vowel. (5 分)
2. 将单词类型分为“短单词” 和“长单词”。现在你想要检验两个“单词类型”对应的单词

“元音比例” 是否显著不同, 请根据不同的单词类型对元音比例画直方图 (注意：不要画成两个图) (5)

1. 请从words 中每次取x个单词，统计辅音结尾的比率，并将其重复n次，将其写成函数。 注意，函数变量有两个，为：“抽取单词个数” x 和 “重复次数” n。请用x=10, n=5000测试你的函数并显示它是可运行的。要求运行结果产生一个新表,里面变量nonvowel\_ratio, 同时产生一个直方图，并伴有密度曲线。请问它是什么分布？为什么？ (15分)

**第四题（20分+ 30分附加题）**

使用 dbCRAN <- tools::CRAN\_package\_db() 获取cran上R packages 的相关数据,。dbCRAN的每一行是对一个包的信息的描述。其中，Package，Maintainer, Author, Depends 四列分别表示每一行所代表的包的名称，维护者，作者和该包依赖于哪些包以及R的版本，将这四列保存为 dbName，然后执行以下操作：

1. 写一个函数cleanNames 用来：去除一列中的各种括号以及括号里面的内容，并去除换行和行尾空格(提示：可以sample\_n 几行来测试你的函数结果)；(10分)

2 对 dbNames的后三列，不用显式循环，执行上面的函数，并保存为 一个tibble，名为 dbClean；(5分)

3. 对 dbClean, 找出维护最多包的 Maintainer 前10，使用条形图(barplot) 图形展示；(5分)

**以下为附加题，如总分超过100分，则计入平时成绩**

4. 对 dbClean 的 Depends, 利用 “,” 来分离(split)，建立tibble，并找出被依赖最多的10个包。(10分)

5. **如果问题1-2没有完成，dbClean没有得到，本题也可以基于dbNames来完成。如果已经得到dbClean，请基于dbClean来完成。**

首先，设置自己学号的末尾四位为随机数种子，从 dbClean中随机抽取Depends列不为NA的100行形成一个子集，保存为dbSample。其次，从dbSample的Depends列中提取出所有出现在该列的包（假设为 N 种），并将“R”及其版本号（如果有的话）去掉。最后，在数据集dbSample的Depends列之后增加 N 列，分别命名为“dep\_包的名字”，这些字段的类型为逻辑型，TRUE和FALSE分别表示某一行所表示的包是否依赖于该包（即该包是否出现在Depends列中）。（20分）