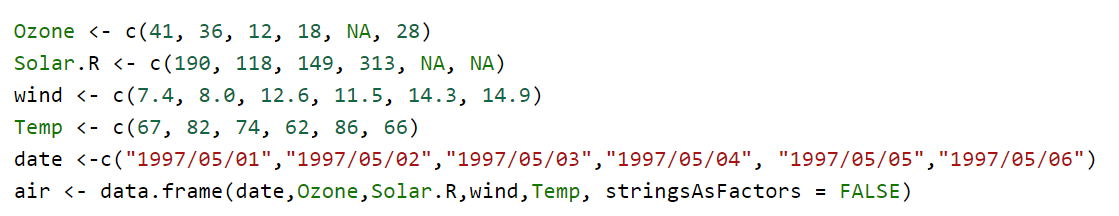
实验二 R语言与描述性统计

【实验目的】

1. 掌握数据预处理的相关操作，包括查看数据集的缺失值，对缺失值进行处理，对异常变量重新赋值。
2. 掌握使用可视化图表方法进行探索性统计分析。
3. 掌握常用的概率分布，并学会用R语言编程求解概率分布问题。
4. 掌握常用的抽样方法，并学会用R语言编程实现抽样。

【实验内容与实现】

1. 根据下述代码，生成数据，并按要求做统计分析。



1. 使用summary()函数查看哪一列含有缺失值的个数，并查看‘wind’列的最值。
2. 使用class()函数查看date的类型，并将其转化为date型，格式为年/月/日。
3. 使用subset()根据‘Temp’列获取air中‘Temp’大于65小于80的数据。
4. 去掉air中含有缺失值的列。
5. 找出 ‘Temp’ 中大于 85 的值将其设置为缺失值。
6. 创建一个数据框 air1 包含两个列 q1，q2。用 cbind 函数添加到 air 中。
7. 使用 names（） 函数查看 air 的列名，再自己创建一个新的数据框，新数据框的列名要与air的列名相同。使用 rbind（） 函数添加到 air 中。
8. 科学家正在开展一项模拟车窗着色对侧窗视觉影响的实验，测得的数据在data.txt，数据中含有以下几列：

case：观测序号；

id：研究车辆的识别号（1-26）；

age：受测者年龄；

sex：表示受测者性别的因子变量，具有f（女性）和m（男性）两个水平；

tint：表示着色程度的有序因子变量，no<lo<hi；

target：表示对比度的因子变量，水平为：locon（低对比度），hicon（高对比度）；

it：检查时间，执行简单识别任务所需的时间（单位为毫秒）；

csoa：临界刺激开始异步，识别字母数字目标的时间（单位为毫秒）；

agegp：年龄阶段因子变量，水平为younger（21-27岁）、older（70-78岁）；

完成下述任务：

1. 读取数据，并逐列对数据进行描述性统计分析；
2. 分别探索it和csoa的数据分布情况；
3. 利用图表探索变量tint对it和csoa的影响，并给出相应的文字解析；
4. 作出age和it的散点图, 并进一步完成下面的操作:

i. 用loess()函数作出拟合线;

ii. 对图的坐标刻度进行调整;

1. 探索描述变量age、it和csoa两两之间是否存在线性关系。
2. agpop.csv文件是美国政府每五年做一次有关农业的普查，收集50个州的所有农场的有关数据，共3078个美国县级或与县级规模相当的农场数据。共有4个区域（region/rnum），50个州（state/snum）,以及3041个县（county/cnum）。我们用到的变量有:县（county/cnum）, 州（state/snum）, 区域（region/rnum）,1992年每个县的耕地面积（acres92）, 1987年每个县的耕地面积（acres87）,1992年每个县拥有的农场个数（farms92）。完成下述任务：
   * 1. 导入原始文件并对进行异常值做剔除处理；
     2. 利用以下抽样方法，获取抽样样本：

i. 利用不放回简单随机抽样，抽取三个样本容量为300的样本，并分别输出每个样本的前五行数据；

ii. 使用分层抽样方法，以‘region’为分层变量，从每层简单随机抽取75个样本单元，构成一个样本，并查看样本中‘region’每层的样本数量；

iii. 采用等距抽样方法抽取三个样本容量为300的样本，并分别输出每个样本的前五行数据；（注：等距抽样方法，也称系统抽样法，先将总体的全部单元按照一定顺序排列，采用简单随机抽样抽取第一个样本单元(或称为随机起点)，再顺序抽取其余的样本单元）。

1. 蒙特卡洛模拟是一大类计算算法的集合，依靠重复的随机抽样来获得数值结果。在统计学中，蒙特卡洛模拟也被广泛应用，可以用来验证各种统计定理和方法的正确性。请自学下面内容https://zhuanlan.zhihu.com/p/619959058，实现用蒙特卡洛模拟验证大数定理和中心极限定理的基本操作。

【作业】

1. 我们以奶牛为例，执行下列操作：
   1. 生成如下的数据：80个正态分布的随机数（即长度为80的向量），该正态分布满足均值为400.0以及标准差为18.2的条件。将该向量命名为‘weight’并输出他的前20个值。
   2. 计算并输出反映‘weight’向量数据的集中趋势、离散趋势和描述分布的各统计量。
   3. 通过‘weight’向量生成一个直方图。要求x轴标签为‘Cows weight’且直方图的柱状体为蓝色。
   4. 生成‘weight’向量的箱线图。要求y轴对应标签为‘Cows weight’，设置颜色为蓝色。
   5. 分析c,d绘制的两种图。通过你的分析，那种图生成了更多的信息？
2. 设Z〜N（0,1）。请使用R模拟(通过产生的随机数求得) 来计算以下概率。

a. P(Z ≤ 0.82)

b. P(-0.74 ≤ Z ≤ 0.82)

c. P(-0.6 ≤ Z ≤0.9)

d. P(-0.614 ≤ Z)

1. 用命令rnorm( )命令产生1000个均值为10, 方差为4的正态分布随机数,

用直方图呈现数据的分布并添加核密度曲线，并用QQPlot方法判断该组数据是否符合正态分布。

1. 对ggplot2 package里面自带的 mpg 数据进行描述性统计分析:

(1) 理解并描述数据集：请用R代码查看该数据集、并用简练的语言描述出该数据集的基本信息（包括但不限于观察值和变量数目、类型，各变量代表什么含义）；

（2）计算每个变量的基本统计量，数据分布，用图、表形式展示（注：描述离散变量和连续变量的统计量和数据分布是不一样的）；

（3）利用普通散点图和条件散点图方式探索不同类型汽车（class）的排量（displ）和每加仑高速里程（hwy）的关系，并用简短的文字描述一下（注：普通散点图用不同颜色或者形状的点表示不同类型的汽车，不要忘记加图例和x、y轴的label）。