**Week2**

data<-read.csv("data\_structuralerror.csv",header=TRUE,skip=4,strip.white=TRUE,

na.strings = c("N/A","NA","-",""))

#读取数据，跳过前四行，删除前导和尾随的空格，并将"n/a"，“na”，"-"视为缺失值

library(tidyr)

library(dplyr)

names(data)<-tolower(names(data))#将所有列名转换为小写

data$dishwasher<-gsub("[[:punct:]]","",data$dishwasher)#将所有的标点符号替换为空格

data$dishwasher<-tolower(data$dishwasher)#将这一列所有的大写都换成小写

data$no.of.mugs.plates<-gsub("[^0-9,-]","",data$no.of.mugs.plates)#将所有的非数字和非逗号，非破折号的字符替换成空字符

data$dishwasher<-tolower(data$dishwasher)

data<-read.csv("data\_structuralerror.csv",header=TRUE,skip=4,strip.white=TRUE,

na.strings = c("N/A","NA","-",""))

require(tidyverse)

data<-data%>%rename\_with(str\_to\_lower)#将所有列名转化为小写

data<-data%>%mutate(dishwasher=str\_replace\_all(data$dishwasher,"[[:punct:]]",""))#将所有标点符号替换成空字符

data<-data%>%mutate(dishwasher=str\_to\_lower(data$dishwasher))#将这一列中的所有字符转化为小写

data<-data%>%mutate(no.of.mugs.plates=str\_replace\_all(data$no.of.mugs.plates,"[^0-9,-]",""))#将所有的非数字和非逗号，非破折号的字符替换成空字符

data<-data%>%distinct()#删除重复行

###处理时间

#方法1：

cleandate<-pmax(

strptime(data$date,format = "%d/%m/%y"),

strptime(data$date,format = "%d-%m-%y"),

strptime(data$date,format = "%d-%m-%Y"),

strptime(data$date,format = "%d/%m/%Y"),

strptime(data$date,format = "%d/%b/%y"),

strptime(data$date,format = "%d %b %Y"),

strptime(data$date,format = "%d %b %y"),na.rm=TRUE)

#方法2：

data<-data%>%mutate(clean.date=as.Date(cleandate))

#快捷键（处理时间）

require(lubridate)

require(tidyverse)

data<-data%>%mutate(clean.date=parse\_date\_time(date,"dmy"))

#根据某些条件对数据进行分类：如果sink.table 列含sink或both，它将会返回yes，如果有na将返回NA，其他情况no

data<-data%>%mutate(sink=case\_when(str\_detect(sink.table,"sink|both")==TRUE~"yes",

is.na(sink.table)==TRUE~NA,.default = "no"))

#用separate（）将no.of.mugs.plates这一列的数据分成两个新列

data<-data%>%separate(no.of.mugs.plates,c("lower","upper"),"-")

#选择特定的列用select（）函数，并将他们存在新的数据框中

data<-data%>%select(clean.date,name,sink,table,no.of.mugs.plates,note)

#加载数据文件，文件名字为：data\_missing.Rda

load("data\_missing.Rda")

#统计缺失值的总数sum() 和is.na()

sum(is.na(originaldata))

# 来删除包含缺失值的行（3种方法）

newdata<-na.omit(missingdata)#用na.omit

newdata2<-missingdata%>%drop\_na()#drop\_na()

newdata3<-missingdata[complete.cases(missingdata),]#complete.cases()

#计算每列缺失值的数量colSums()和is.na()

colSums(is.na(missingdata))

#选择数据框种的特定列（2种方法）

newdata4<-missingdata[,c(2,4,5)]#使用列序号

newdata5<-missingdata%>%select(V2,V4,V5)#使用列名

##处理缺失值的方法（5种）

#使用均值来覆盖一列种的缺失值

imputedata$V1[is.na(missingdata$V1)]<-mean(missingdata$V1,na.rm = TRUE)#忽略na值

#找到一列中的所有na值，并用数字3来替换，常数插补

table(missingdata$V3)#查看这一列种各分类的总数

imputedata$V3[is.na(missingdata$V3)]<-3

#KNN插补，找到具有相似特征的观察值，并使用这些观察值的平均值来插补缺失值

install.packages("VIM")

imputeknn<-VIM::kNN(missingdata)

#多重插补：使用模型估计和重复模拟来生成一组完整的数据集

install.packages("mice")

micemodel<-mice::mice(missingdata)

imputemice<-mice::complete(micemodel)

#用predict()函数来填充缺失值，用model预测值来填充

model<-lm(V1~V2,data = missingdata)##线性回归分析创建一个线性模型，V1是因变量，V2是自变量

imputedata$V1[is.na(missingdata$V1)]<-predict(model,new.df)

#处理文本文件

require(readr)

data<-read\_tsv("data\_outlier.txt")#读取文本文件

datadf<-as.data.frame(data)#将数据转化为数据框

#离群值检测的方法

out<-boxplot(datadf$hwy)$out#创建箱线图，离群值被储存在out变量中

which(datadf$hwy %in% out)#返回数据中找到离群值的位置

install.packages("outliers")

#grubbs检测是一种常用的检测单个离群值的统计方法

outliers::grubbs.test(datadf$hwy)#使用grubbs.test函数来进行检测，检测离群值

outliers::grubbs.test(datadf$hwy,opposite = TRUE)#检测数据中的最小值是否为离群值

#rosnerTest 函数进行Rosner检测 检测多个离群值的统计方法

install.packages("EnvStats")

EnvStats::rosnerTest(datadf$displ,k=5)#k=5表示检测的离群值的数量

##grubbs和rosner检测的区别

model<-lm(V1~V2,data = missingdata)

influence.measures(model)#计算线性模型的影响度量，这些度量可以检查回归拟合的质量

















