2023 ADS2 Week 7 Data Cleaning Problem Set

ADS2

2023-11-02

Introduction

This R Markdown file contains a tutorial of how to do data cleaning with R. The input data is orignally from R package `ggplot2` (`diamonds` dataset). It contains the `X` (ID), `carat, cut, color, clarity, depth, table, price`, and measurements in `x, y, z` axes for more than 50,000 diamonds. In order to do this demo, 100 diamonds from the orignal dataset (with `set.seed=12`) were sampled and modifed some of the entries.

Setting up working directory

Before we start, we need to set up our working directory.**This is a absolute or relative directory?**

```
setwd("path.to.your.working.directory") #make sure change to your own corrected paths
```

Import data

The next thing we need to do is import our data. We have learnt in the lecture, there are several function to import data (including: `read.delim, read.delim2, read.csv, read.csv2`).

Since my data is in xxx.csv format, so I use `read.csv()` function

```
diamond_sample = read.csv("./Rdata_diamonds_samples100_mdf.csv")
head(diamond_sample)
```

```
##   X carat     cut color clarity depth table price    x    y    z
## 1 1  2.01    Fair     G     SI1  70.6  64.0 18574 7.43 6.64 4.69
## 2 2  2.08   Ideal     J      IF  61.0  55.0 17986 8.32 8.25 5.05
## 3 3  1.22   Ideal     G     VS2  61.4  56.0  8362 6.90 6.88 4.23
## 4 4  0.82 Premium     F     SI1  62.4  56.0  2962 6.01 5.98 3.74
## 5 5  0.32   Ideal     D     SI1  62.7  54.0   589 4.35 4.39 2.74
## 6 6  0.34   Ideal     E     VS2  62.1  54.1   758 4.47 4.50 2.78
```

R data types

Data types

Before we start data cleaning, let's get more ideas about R data types.

```r
#check the data type of X column
head(diamond_sample$X)
typeof(diamond_sample$X)
class(diamond_sample$X)
#check the data type of carat column
head(diamond_sample$carat)
class(diamond_sample$carat)
#check the data type of cut column
head(diamond_sample$cut)
class(diamond_sample$cut)
#What's the difference between -
#character and factor?
cut.chr=as.character(diamond_sample$cut)
head(cut.chr)
class(cut.chr)
```

Numeric vs integer?

Also, be aware of the difference between numeric vs integer

```r
var1 = c(2,3,5,6)
class(var1)
var2 = c(2.0,3.9,5.1,6.9)
class(var2)
var3 = c(2L,3L,5L,6L)
class(var3)
```

R data structure

Basic data structures

Numeric vectors

Before we start data cleaning, let's get more ideas about R data structure. First let's create some *numeric vectors*.

```r
############R data structure vector vs list
#create numeric vector
var.num = c(2.0,3.9,5.1,6.9)
print(var.num)
class(var.num)
```

Integer vectors

Then let's create some *integer vectors*.

```r
#create integer vector
var.int = c(2L,3L,5L,6L)
print(var.int)
class(var.int)
```

Character vectors

Then let's try to create some *character vectors*.

```
#create character vector
var.char = c("Hello",",","world","!")
class(var.char)
print(var.char)
```

Factor vectors

Then let's try to create some *factor vectors*.

```
#create factor vector
var.fac = factor(c("mX","mX","high","low"),
                 levels=c("low","mX","high"))
print(var.fac)
class(var.fac)
```

Lists

Finally, let combine all the vectors we just created, to generate a *list*.

```
#combine all vector above to create list
data.list=list(var.num,var.int,var.char,var.fac)
head(data.list)
data.list[[1]]
data.list[[1]][1]
data.list[[3]]
data.list[[3]][3]
length(data.list)
dim(data.list)
```

Matrix vs dataframes?

What's the difference between **matrix vs data frame**?

```
#convert dataframe to matrix
data.matrix = as.matrix(diamond_sample)
#look at the header of the matrix
head(data.matrix)
#check the matrix class
class(data.matrix[,1])
class(data.matrix[,2])
```

Now, let's remove the `data.matrix` and values that we created, and reload the data for Data cleaning exercise

```
rm(list = ls())
diamond_sample = read.csv("Rdata_diamonds_samples100_mdf.csv")
```

ADS2 Week 7 Problem set

Your task for this problem set will be as follows:

1. Please try to run the all the above R scipt yourself
2. Now, try to clean the data yourself. Answer the questions listed below.

Are there missing values?

Are there duplicated values?

Are there strange patterns in your data?

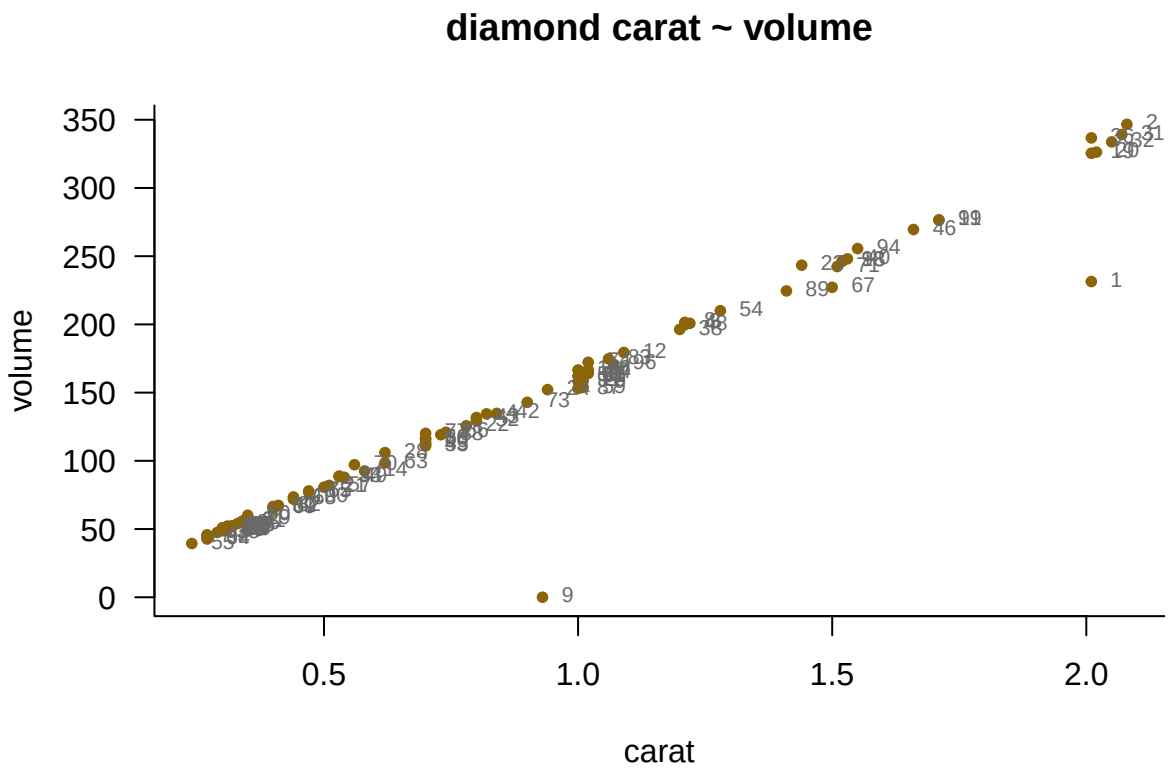
Hint: We know that diamond and it's volume have a linear relationship. Thus, we would like to investigate the relationship between `carat` vs. `volume`. In order to test this idea, we need to generate a new vector called `volume = x * y * z`.

```
diamond_sample$volume = diamond_sample$x * diamond_sample$y * diamond_sample$z %>%
  round(2)
head(diamond_sample)
```

```
##   X carat     cut color clarity depth table price    x    y    z    volume
## 1 1  2.01    Fair     G     SI1  70.6  64.0 18574 7.43 6.64 4.69 231.38209
## 2 2  2.08   Ideal     J      IF  61.0  55.0 17986 8.32 8.25 5.05 346.63200
## 3 3  1.22   Ideal     G     VS2  61.4  56.0  8362 6.90 6.88 4.23 200.80656
## 4 4  0.82 Premium     F     SI1  62.4  56.0  2962 6.01 5.98 3.74 134.41485
## 5 5  0.32   Ideal     D     SI1  62.7  54.0   589 4.35 4.39 2.74  52.32441
## 6 6  0.34   Ideal     E     VS2  62.1  54.1   758 4.47 4.50 2.78  55.91970
```

We then plot scatterplot of `carat` vs `volume` to see whether they have a linear relationship.

```
plot(x = diamond_sample$carat, y = diamond_sample$volume,
     pch = 20, col = "darkgoldenrod4",
     las = 1, xlab = "carat", ylab = "volume",
     main = "diamond carat ~ volume", bty = "l")
text(diamond_sample$carat, diamond_sample$volume,
     labels = diamond_sample$X, col = "dimgray",
     cex = 0.7, pos =4)
```



What do you think about it?

Correct typos in the dataset.

After running all the procedure above (clean missing, duplicated, strange data), first we want to see whether there is any typo. **Please check those character/factor vectors in the diamonds data, see whether you can find any typos and then correct those typo in R? Remember to document any edit you do properly. Use screen-diagnosis-treat-document strategy.**

Find outliers in the dataset.

After removing the missing data, duplicated data, strange data and typos, we now want to see whether there is any outliers. For example, is there any outlier if we investigate the relationship between carat vs price. Since we know, the diamond price is positively correlate with its carat! (the bigger the diamond is the more expensive).

What to do:

1. **screen for out outliers**
2. **diagnosis for out outliers**
3. **treat out outliers**
4. **documentation**

Hint: If the data looks suspicious, and you don't know whether you should remove it or not, you can generate a new indicator vector to the dataframe to indicate whether this observation is suspicious (but you don't have evXence to delete it).

Bonus Question (Optional)

For the outliers you identified above, those have strange pattern of `carat ~ price`, try to make more plot to see whether this strange pattern is actually correlate with other features.

Task: try to plot the relationship between `carat ~ price`, but seperate data points by their clarity. (**Hint**: use `ggplot2`, `facet_grX()` function).

Originally created by Wanlu Liu in 2019.

Last update by Dmytro Shytikov in 2023