HW2_Xu_Peng

Peng

2017/9/11

Problem 4

Version control could help me record and modify any temperate ideas inspired from the lectures and materials. If some problem happens, such as the system breakdown, it is easy to recover the data with version control. And with its help, I can also share my files with other people to make cooperation.

Problem 5

Part a

Table 1: Sensory data summary

Item	Person	value
Length:150	Length:150	Min. :0.700
Class :character	Class :character	1st Qu.:3.025
Mode :character	Mode :character	Median :4.700
NA	NA	Mean :4.657
NA	NA	3rd Qu.:6.000
NA	NA	Max. :9.400

Part b

Table 2: Long Jump data summary

YearCode	Year	dist
Min. :-4.00	Min. :1896	Min. :249.8
1st Qu.:21.00	1st Qu.:1921	1st Qu.:295.4
Median :50.00	Median :1950	Median :308.1
Mean :45.45	Mean :1945	Mean :310.3
3rd Qu.:71.00	3rd Qu.:1971	3rd Qu.:327.5
Max. :92.00	Max. :1992	Max. :350.5

Part c

Table 3: Brain/Body weight data summary

Brain	Body
Min. : 0.005	Min. : 0.10
1st Qu.: 0.600	1st Qu.: 4.25
Median : 3.342	Median : 17.25

Brain	Body
Mean : 198.790	Mean : 283.13
3rd Qu.: 48.203	3rd Qu.: 166.00
Max. :6654.000	Max. :5712.00

Part d

Table 4: Tomato data summary

Clone	Replicate	value	Variety
Length:18	Length:18	Length:18	Length:18
Class :character	Class :character	Class :character	Class :character
Mode :character	Mode :character	Mode :character	Mode :character

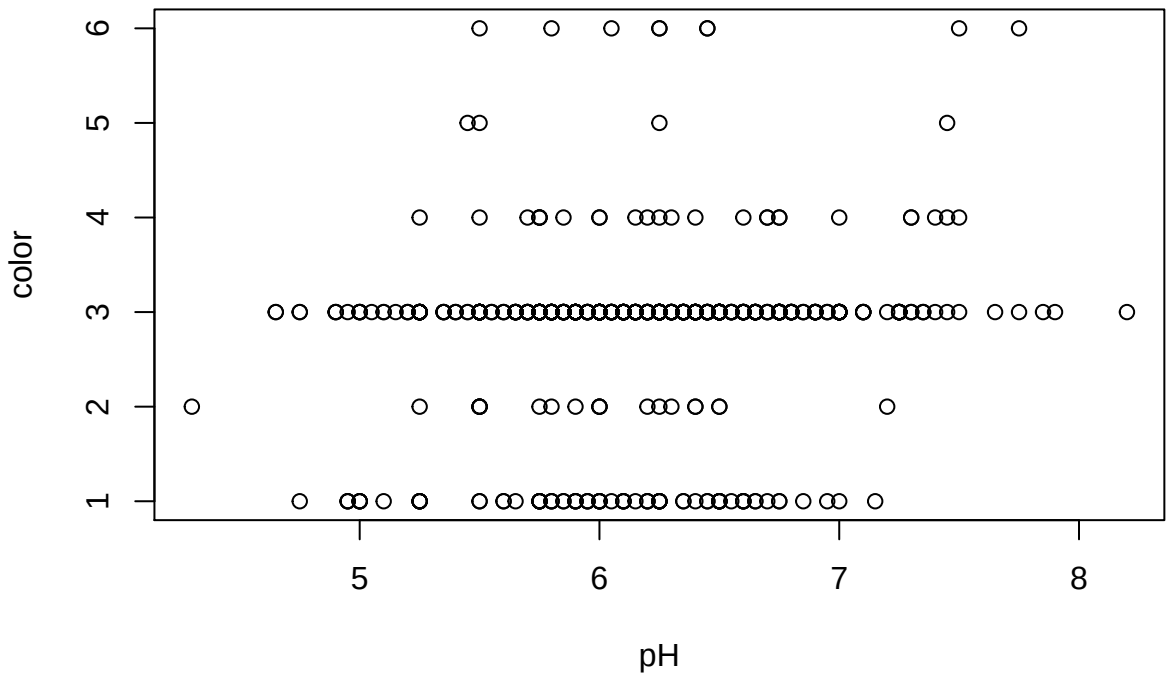
Problem 6

After importing the data, there are three columns related to our analysis. The mean pH value is adopted to combine pH_min and pH_max. So the problem is transformed to test the relationship between pH mean values and foliage colors.

The colors are symbolized with numbers from 1 to 6 according to their mean values, as shown below. Then the scattor plot could be drawn to show the distribution.

Table 5: Color Representation

color	No
Dark Green	1
Yellow-Green	2
Green	3
Gray-Green	4
Red	5
White-Gray	6



The lm function is used to build the linear model between pH and color. The coefficients are listed below.

```
##
## Call:
## lm(formula = color ~ pH, data = data.frame(plants_6))
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.0345  0.0404  0.1340  0.2088  3.2743
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.69628    0.28868   5.876 6.08e-09 ***
## pH           0.18716    0.04661   4.015 6.47e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7305 on 830 degrees of freedom
## Multiple R-squared:  0.01906,    Adjusted R-squared:  0.01787
## F-statistic: 16.12 on 1 and 830 DF,  p-value: 6.472e-05
```

Meanwhile, the ANOVA method is tried to explore the relationship, with details below.

```
##              Df Sum Sq Mean Sq F value   Pr(>F)
## pH            1    8.6   8.604   16.12 6.47e-05 ***
## Residuals   830  442.9   0.534
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Problem 7

Part d

Through the summary of data in 2017, there are 354 makes and 22275 models.

Part e

Use sort function to rank the defects. The largest five defects could be identified as followed.
K04 AC1 G05 RA2 K05
409467 323375 182410 177706 154179
The top make/model for these defects are:

Table 6: TOP Make/Models of Five Common Defects

Defect	Make_Model
K04	VOLKSWAGEN/POLO
AC1	VOLKSWAGEN/POLO
G05	VOLKSWAGEN/POLO
RA2	PEUGEOT/206; 1.4 3DRS
K05	VOLKSWAGEN/POLO

Part h

This workflow considers all the possible records at first, which lead to high volume of computing capacity. So if the contraint could be applied at first, such as the year 2017, the sample could be reduced a lot.