STATS 551 - HW1

Zhen Qin

5.

Compute the total number of quizzes for 10000 times. The sample mean and sample variance of these numbers are very close to the theoretical values: mean = 13.5 and variance = 6.75. Thus the results in question 4 are verified.

```
# set the number of trials
n = 10000

# create a function to generate results
quiz = function(){
  theta = runif(1,0,1)
  rbinom(1,1,theta)
}
total = rep(0,n)
for(i in 1:n){
  for(j in 1:27){
    total[i] = total[i] + quiz()
  }
}

# compute the sample mean and sample variance
mean(total)
```

```
## [1] 13.4783
```

```
var(total)
```

```
## [1] 6.564786
```

Exploratory Data Analysis

The dataset is available here. There are six variables **crab**(the ID of the observed crab), **y**(stands for the number of crabs near the observation), **weight**, **width**, **color** and **spine**.

Description

First, I loaded the data using R and extracted 6 rows. It is clear that **y**, **weight** and **width** are numerical variables, while others are categorical variables.

```
library(knitr)
library(ggplot2)
library(GGally)
```

```
## Warning: package 'GGally' was built under R version 3.4.3
```

```
# load the data and show 6 rows
crab = read.table("crabs.dat.txt", header = T)
kable(head(crab))
```

crab	y	weight	width	color	spine
1	8	3.05	28.3	2	3
2	0	1.55	22.5	3	3
3	9	2.30	26.0	1	1
4	0	2.10	24.8	3	3
5	4	2.60	26.0	3	3
6	0	2.10	23.8	2	3

Quantative Analysis

Second, I used quantative methods to show some properties. The median of **y** is less than the mean of **y**, which suggests **y** may have skewness. **weight** and **width** seem to have normality.

```
# data summary
crab = crab[,2:6]
crab$color = as.factor(crab$color)
crab$spine = as.factor(crab$spine)
kable(summary(crab))
```

y	weight	width	color	spine
Min. : 0.000	Min. :1.200	Min. :21.0	1:12	1: 37
1st Qu.: 0.000	1st Qu.:2.000	1st Qu.:24.9	2:95	2: 15
Median : 2.000	Median :2.350	Median :26.1	3:44	3:121
Mean : 2.919	Mean :2.437	Mean :26.3	4:22	NA
3rd Qu.: 5.000	3rd Qu.:2.850	3rd Qu.:27.7	NA	NA
Max. :15.000	Max. :5.200	Max. :33.5	NA	NA

Then I calculated the correlation matrix of numerical variables. It is worthwhile to note that the correlation of **weight** and **width** is big. Maybe they have colinearity.

```
# correlation matrix of numerical variables
kable(cor(crab[1:3]))
```

	y	weight	width
y	1.0000000	0.3692474	0.3398903
weight	0.3692474	1.0000000	0.8868715
width	0.3398903	0.8868715	1.0000000

Graphic Analysis

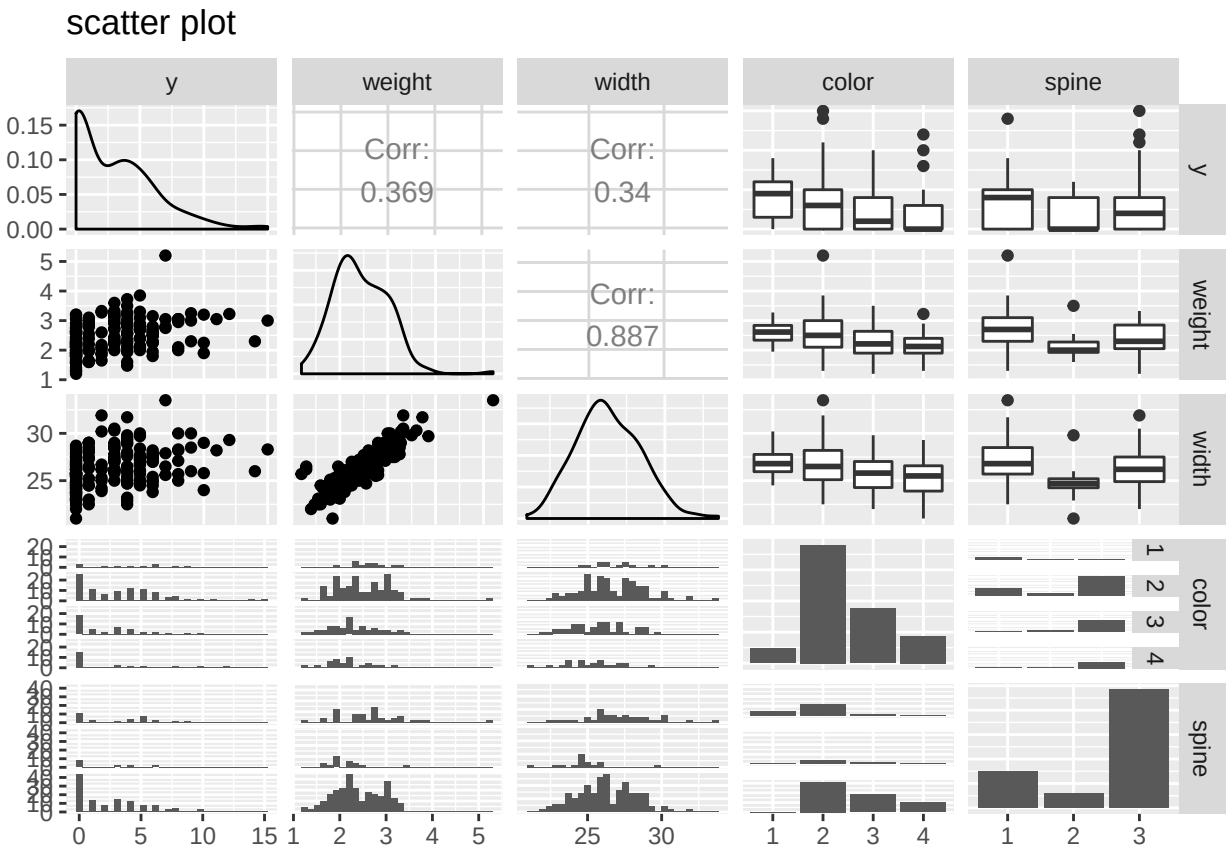
Third, I used graphic methods to explore.

The scatter plot matrix shows pairwise relationship of variables. This plot verifies the guess of quantative analysis.

```
# scatter plot matrix
ggpairs(crab) + ggtitle("scatter plot")
```

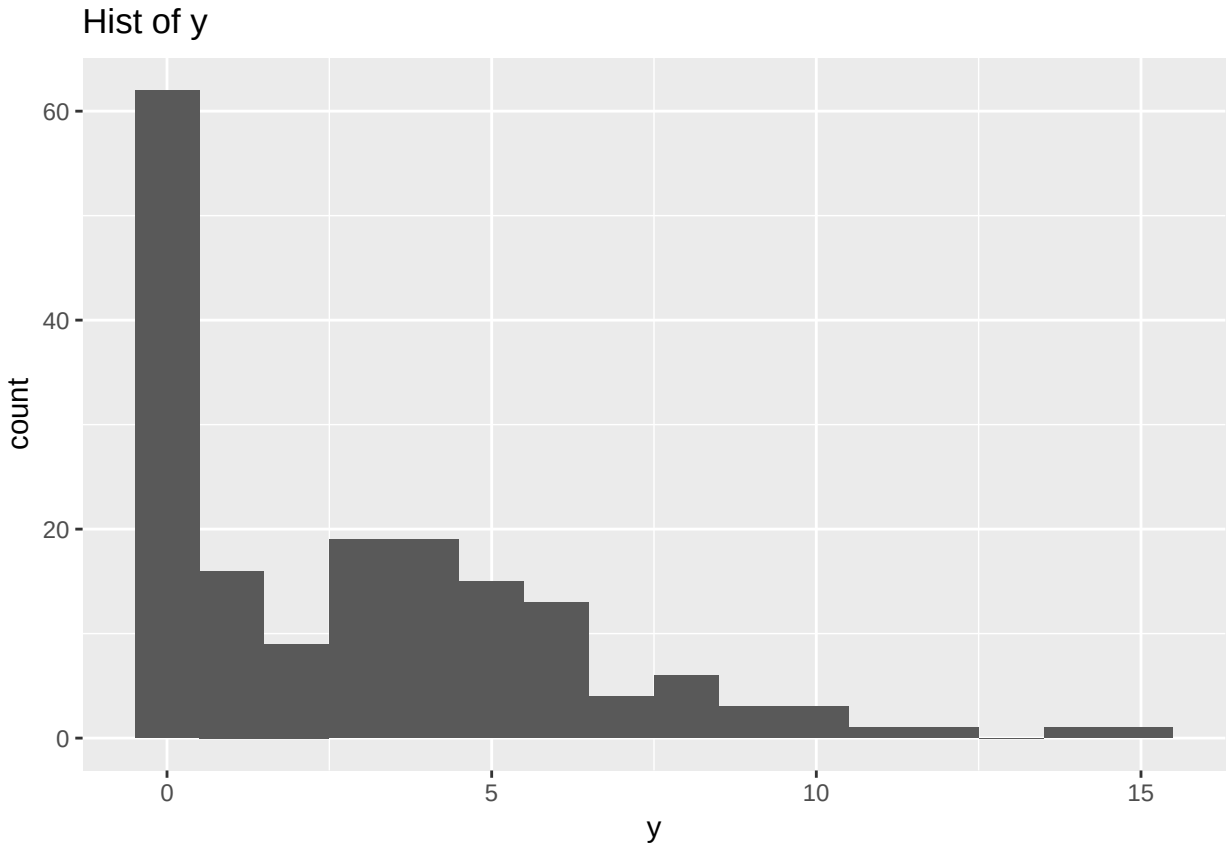
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



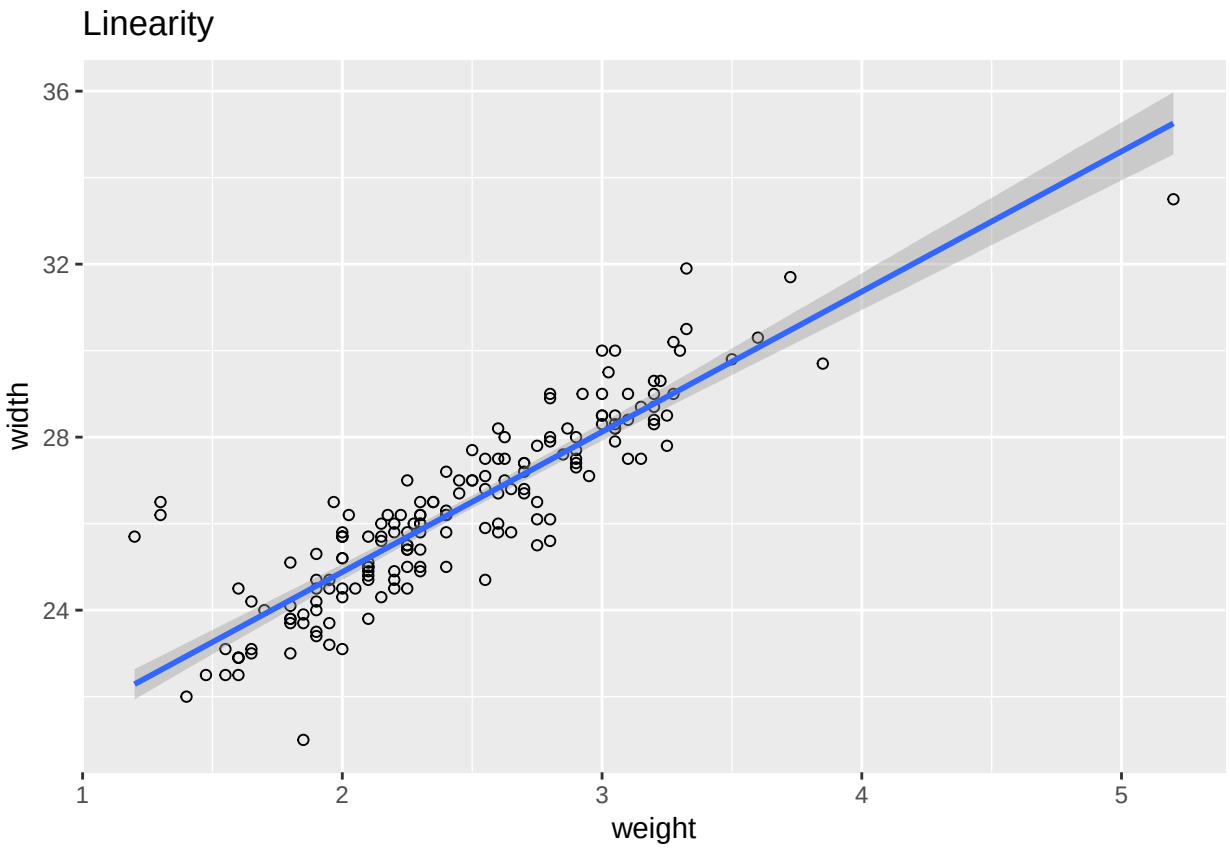
The histogram of **y** shows there are many 0s in **y**, so I cannot use any model that assumes **y** has normality. This may be a potential difficult feature of the data.

```
# histogram of y
ggplot(data = crab) + geom_histogram(aes(x = y), binwidth = 1) + ggtitle('Hist of y')
```



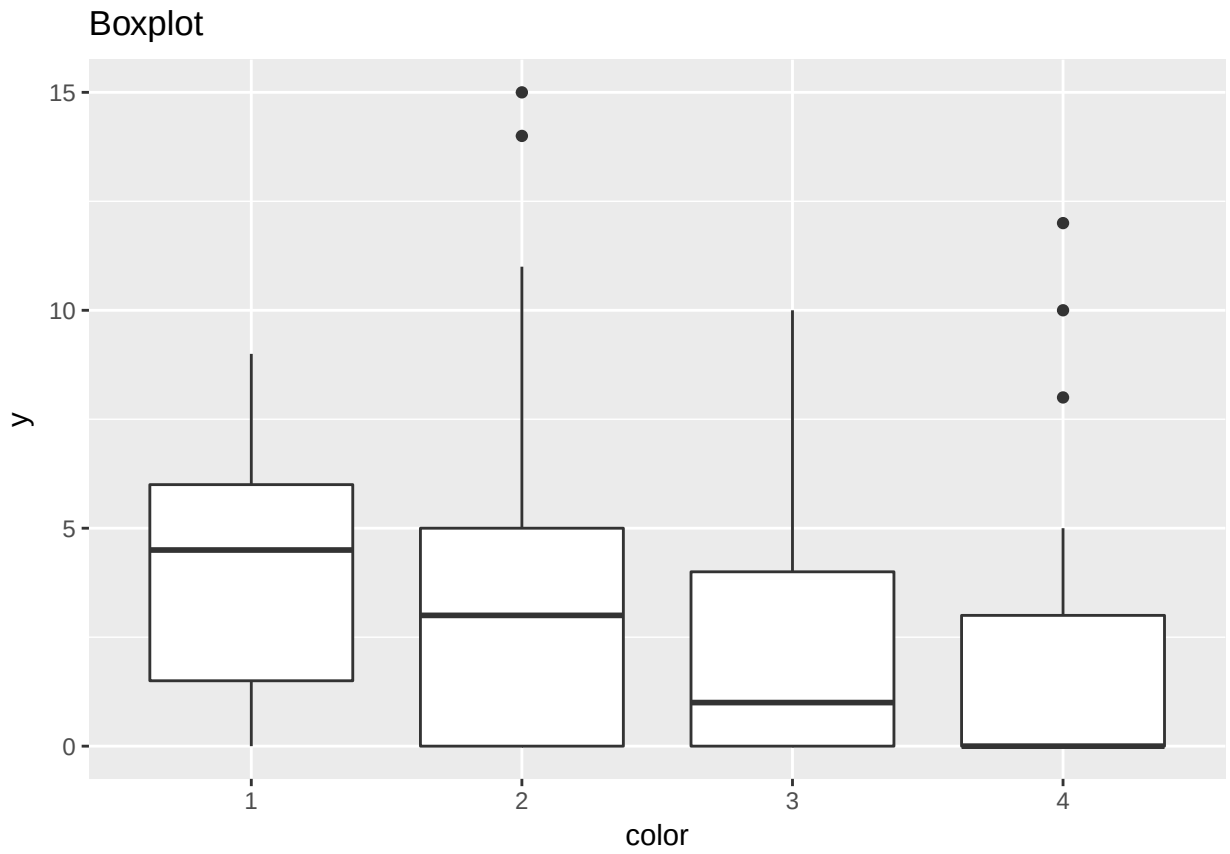
The regression line in the following plot has good fit, so these varibles have colinearity.

```
# linearity between variables
ggplot(data = crab, aes(x = weight, y = width)) + geom_point(shape = 1) + geom_smooth(method = lm) + gg
```



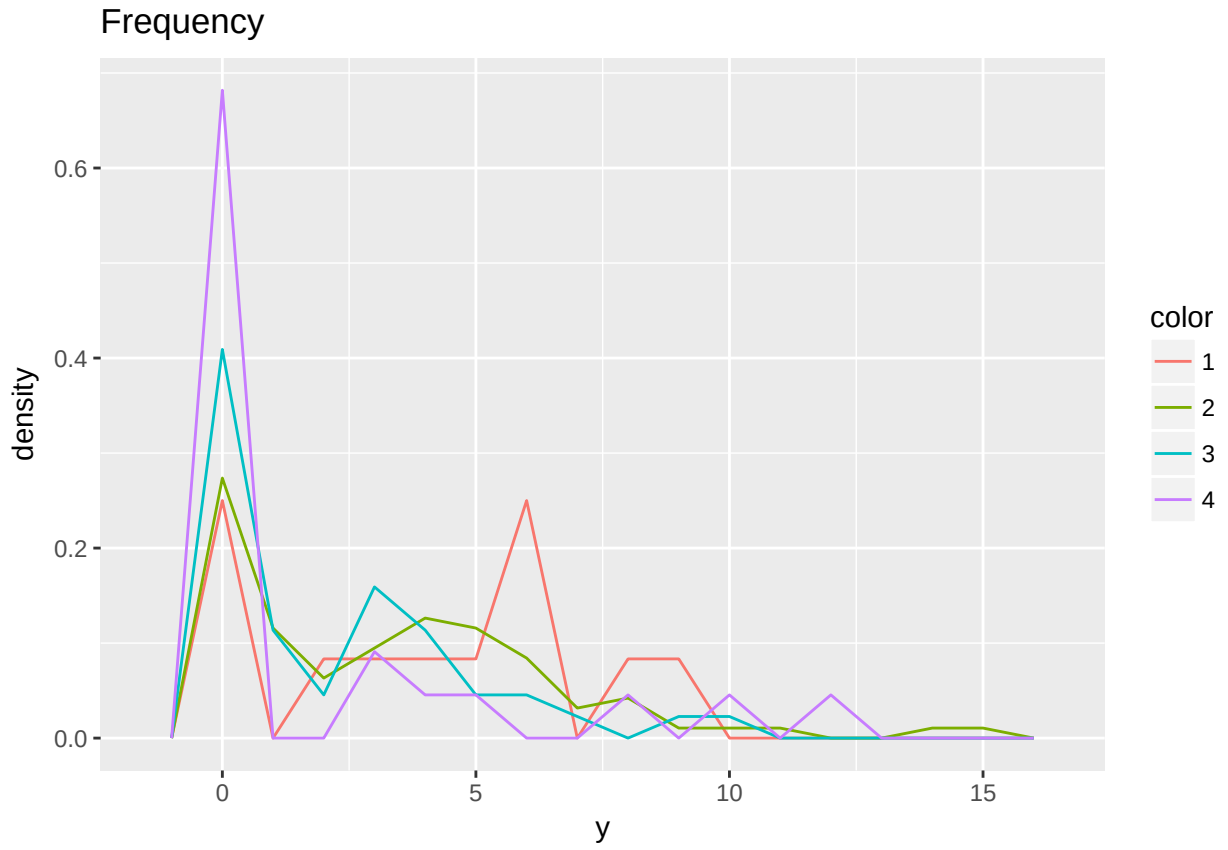
The boxplot shows **y** has considerable skewness when **color** is 2, 3 or 4.

```
# boxplot by color
ggplot(data = crab, aes(x = color, y = y))+geom_boxplot()+ggtitle('Boxplot')
```



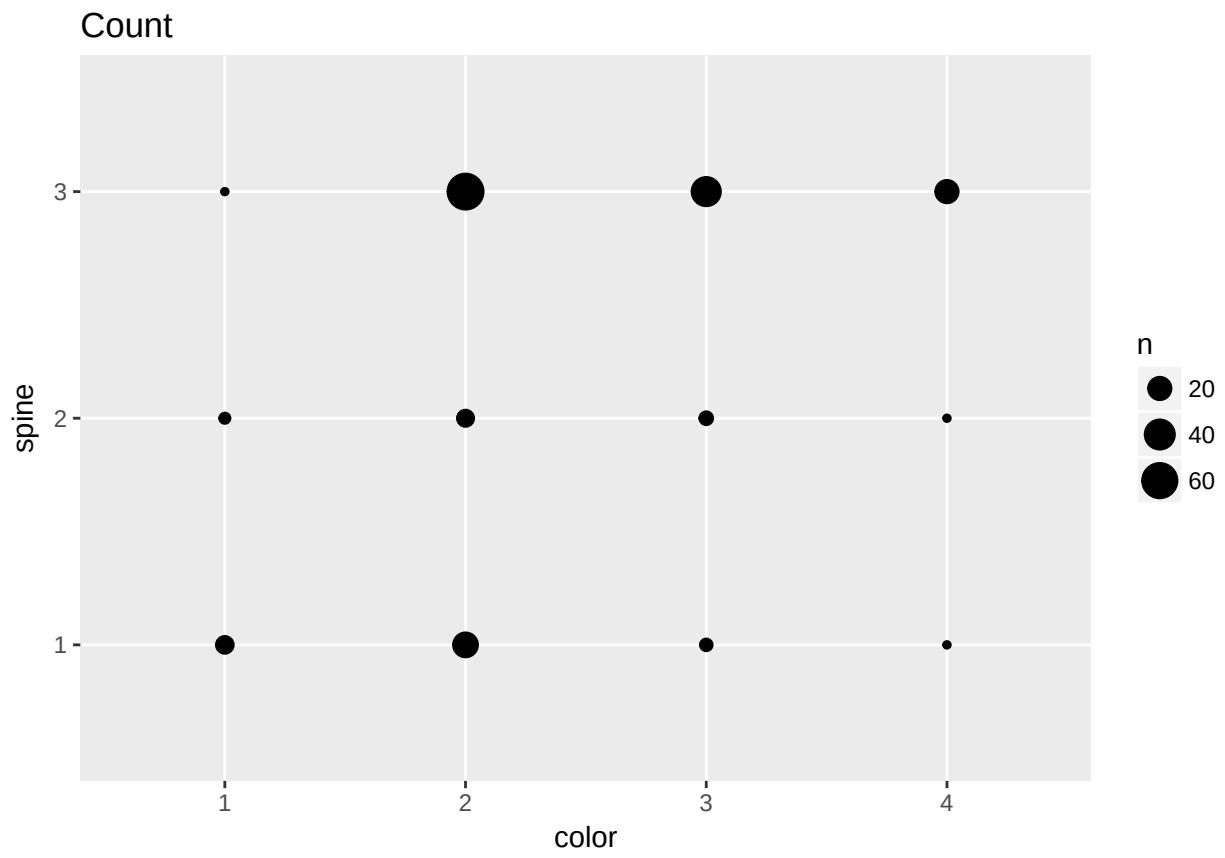
The freqency plot verifies this idea.

```
# frequency plot
ggplot(data = crab, aes(x = y, y = ..density..))+geom_freqpoly(aes(colour = color),binwidth = 1)+ggtitl
```



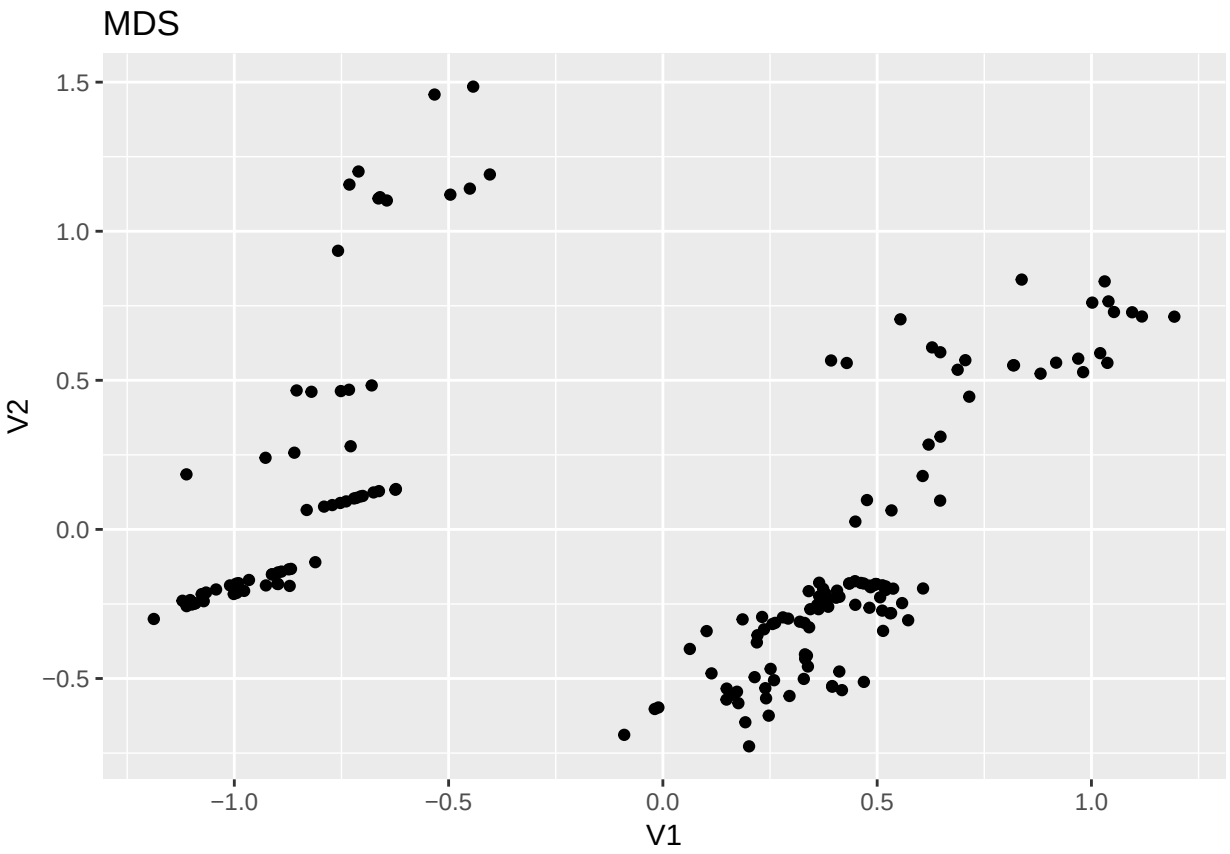
The dot plot shows some categories have few cases, e.g. **color** = 1, **spine** = 3. Maybe they have relationship with **y** because all of them are not distributed evenly.

```
# number of cases by color and spine
ggplot(data = crab, aes(x = color, y = spine)) + geom_count() + ggtitle('Count')
```



The scatter plot of MDS coordinates shows that points can be divided into 2 groups in a way.

```
# MDS plot
mds = cmdscale(dist(crab,method = 'canberra'))
mds = as.data.frame(mds)
ggplot(data = mds, aes(x = V1, y = V2)) + geom_point() + ggtitle('MDS')
```



Conclusion

Here are some interesting properties of the data:

- In **y** majority is equal to 0
- The two variables **weight** and **width** have colinearity. And they both have normality
- The categorical variables may be influential with respect to **y**

I will not use this dataset to do the final project. The main reason is that the data is relatively simple but I would like to do some challenging works. In addition, it is hard to get proper prior knowledge of crabs, so bayesian modeling may be not good for the data.