MSiA 400 Lab 7

Exploratory Data Analysis

Huiyu Wu

11/14/2020

NORTHWESTERN
UNIVERSITY

Dataset

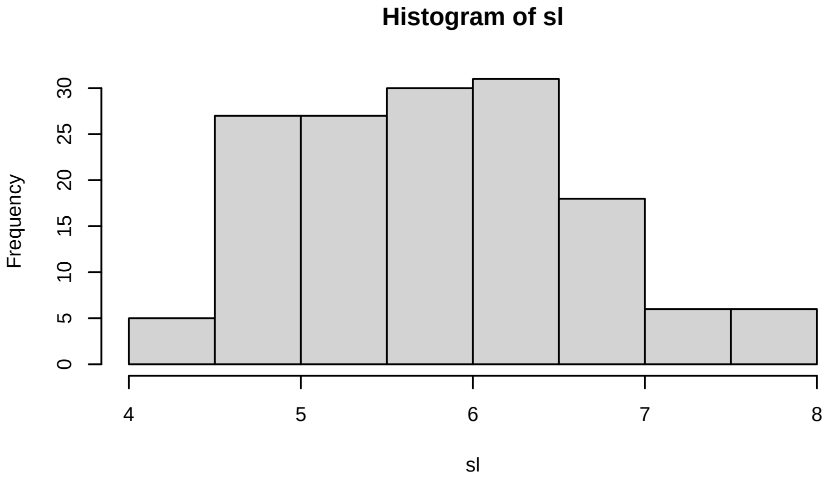
- Iris flowers from 3 species: Iris setosa, versicolor, and virginica
  - 50 flowers from each species
- Sepal length & width; petal length & width
  - In cm

```
data(iris); iris$Species = factor(iris$Species)
head(iris)
```

```
##   Sepal.Length Sepal.Width Petal.Length Petal.Width Species
## 1          5.1         3.5          1.4         0.2  setosa
## 2          4.9         3.0          1.4         0.2  setosa
## 3          4.7         3.2          1.3         0.2  setosa
## 4          4.6         3.1          1.5         0.2  setosa
## 5          5.0         3.6          1.4         0.2  setosa
## 6          5.4         3.9          1.7         0.4  setosa
```

Histogram

```
par(cex=0.7); sl=iris$Sepal.Length
hist(sl)
```



Stem-and-Leaf Plot

```r
stem(sl, width=75)
```

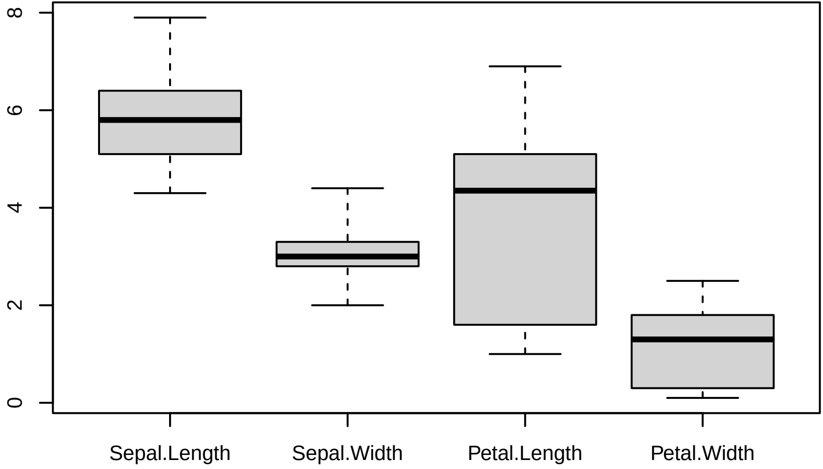
```
##
##   The decimal point is 1 digit(s) to the left of the |
##
##   42 | 0
##   44 | 0000
##   46 | 000000
##   48 | 00000000000
##   50 | 0000000000000000000
##   52 | 00000
##   54 | 0000000000000
##   56 | 00000000000000
##   58 | 0000000000
##   60 | 000000000000
##   62 | 0000000000000
##   64 | 000000000000
##   66 | 0000000000
##   68 | 0000000
##   70 | 00
##   72 | 0000
##   74 | 0
##   76 | 00000
```

Interquartile range

- $IQR = Q3 - Q1$
- Outliers: $(-\infty, Q1 - 1.5IQR,] \cup [Q3 + 1.5IQR, \infty)$

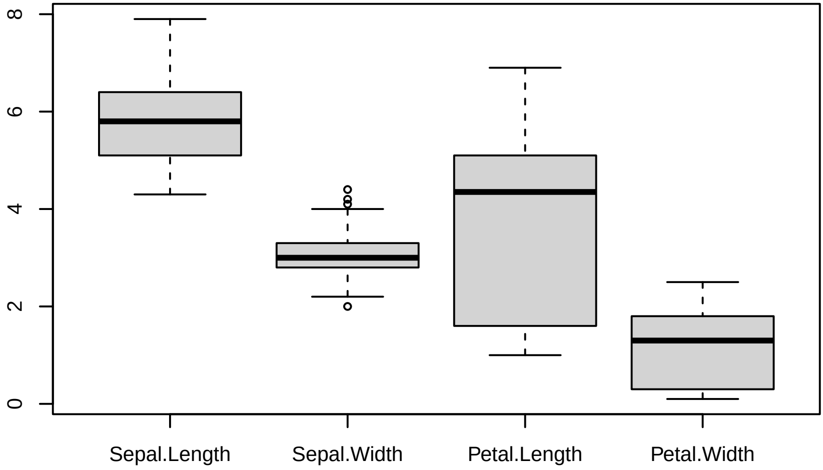
Boxplot

```r
par(cex=0.7)
boxplot(iris[,1:4], range=0)
```



Box-and-Whisker plot

```
par(cex=0.7)
boxplot(iris[,1:4])
```



5-Number Summary

```r
summary(iris[,1:4])
```

```
##   Sepal.Length    Sepal.Width     Petal.Length    Petal.Width   
##  Min.   :4.300   Min.   :2.000   Min.   :1.000   Min.   :0.100  
##  1st Qu.:5.100   1st Qu.:2.800   1st Qu.:1.600   1st Qu.:0.300  
##  Median :5.800   Median :3.000   Median :4.350   Median :1.300  
##  Mean   :5.843   Mean   :3.057   Mean   :3.758   Mean   :1.199  
##  3rd Qu.:6.400   3rd Qu.:3.300   3rd Qu.:5.100   3rd Qu.:1.800  
##  Max.   :7.900   Max.   :4.400   Max.   :6.900   Max.   :2.500  
```

Skewness & Kurtosis

- Skewness: $\frac{E\left[(X-\mu)^3\right]}{\sigma^3}$
- Kurtosis: $\frac{E\left[(X-\mu)^4\right]}{\sigma^4}$

```
library(e1071)
```

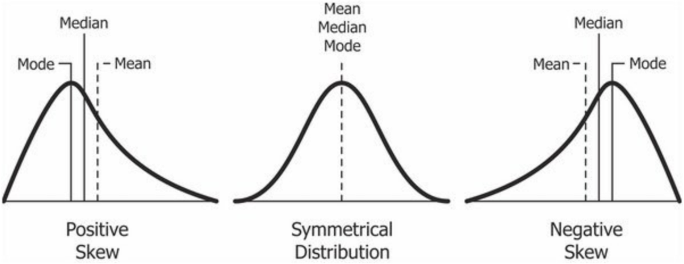
```
skewness(sl)
```

```
## [1] 0.3086407
```

```
kurtosis(sl)
```

```
## [1] -0.6058125
```

Skewness

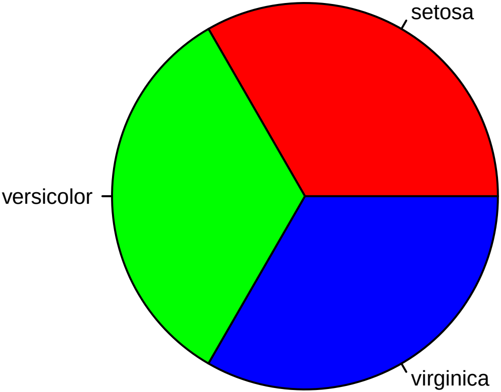

Median
Mode
Mean
Positive
Skew
Mean
Median
Mode
Symmetrical
Distribution
Median
Mean
Mode
Negative
Skew

Kurtosis

- Standard normal has kurtosis 3
- R gives you excess Kurtosis
- $\frac{E\left[(X-\mu)^4\right]}{\sigma^4} - 3$

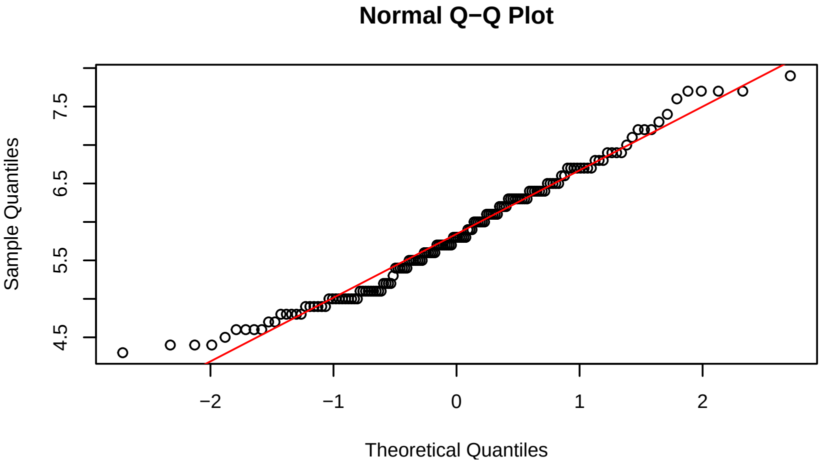
Pie Chart

```
par(cex=0.7)
t=table(iris$Species)
pie(t,labels=names(t),col=rainbow(length(t)))
```



Q-Q Plot

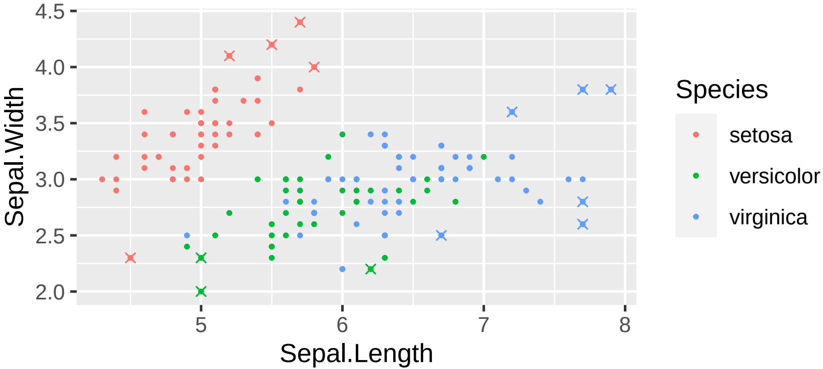
```
par(cex=0.7); qqnorm(sl)
abline(mean(sl),b=sd(sl),col="red")
```



Outliers

```r
library(ggplot2); library(ggpmisc)
```

```r
ggplot(data=iris, aes(Sepal.Length, Sepal.Width, color=Species)) +
  geom_point(size=.5) +
  stat_dens2d_filter(geom="point", shape=4, keep.fraction=.1)
```



K-Means Clustering

```r
set.seed(400)
km = kmeans(iris[,1:4], 3) # 3 clusters
km$size # size of each cluster
```

```
## [1] 62 38 50
```

```r
km$centers # center of each cluster
```

```
##   Sepal.Length Sepal.Width Petal.Length Petal.Width
## 1     5.901613    2.748387     4.393548    1.433871
## 2     6.850000    3.073684     5.742105    2.071053
## 3     5.006000    3.428000     1.462000    0.246000
```

```r
km$withinss # within cluster sum-of-squars
```

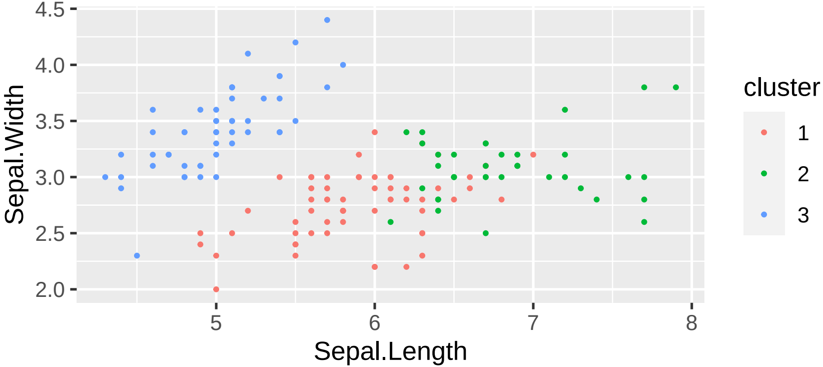
```
## [1] 39.82097 23.87947 15.15100
```

```r
km$betweenss/km$totss
```

```
## [1] 0.8842753
```

K-Means Clustering

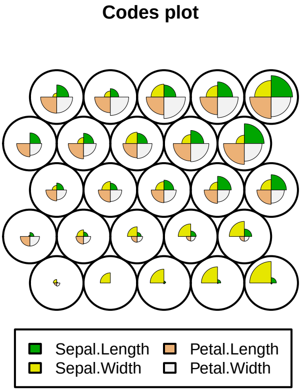
```
iris$cluster=factor(km$cluster)
ggplot(data=iris, aes(Sepal.Length, Sepal.Width, color=cluster)) +
  geom_point(size=.5)
```



Self-Orgnizing Maps (SOM)

```r
library(kohonen)
```

```r
som_grid = somgrid(xdim=5, ydim=5, topo="hexagonal")
som_model = som(scale(iris[,1:4]), grid=som_grid)
par(cex=0.5); plot(som_model)
```


Codes plot
Sepal.Length
Sepal.Width
Petal.Length
Petal.Width