Analysis of the USJudgeRatings data set

Adrien Toulouse & Paul-Antoine GIRARD

Problem 2: Analysis of the USJudgeRatings dataset

This exercise is open. You are asked to use the tools we have seen together to analyze the USJudgeRatings data set. This data set is provided in the package datasets. Your analysis should be reported here and include:

- an introduction
- a general description of the data
- the use of descriptive statistics
- the use of all techniques we have seen together that might be relevant
- a conclusion

Overall, your analysis, including the graphs and the codes should not exceed 15 pages in pdf.

Introduction

The USJudgeRatings dataset contains lawyers' ratings of state judges in the US Superior Court in 1977. The data is stored in a dataframe.

```
data(USJudgeRatings)
head(USJudgeRatings)
```

```
##                CONT INTG DMNR DILG CFMG DECI PREP FAMI ORAL WRIT PHYS RTEN
## AARONSON,L.H.   5.7  7.9  7.7  7.3  7.1  7.4  7.1  7.1  7.1  7.0  8.3  7.8
## ALEXANDER,J.M.  6.8  8.9  8.8  8.5  7.8  8.1  8.0  8.0  7.8  7.9  8.5  8.7
## ARMENTANO,A.J.  7.2  8.1  7.8  7.8  7.5  7.6  7.5  7.5  7.3  7.4  7.9  7.8
## BERDON,R.I.     6.8  8.8  8.5  8.8  8.3  8.5  8.7  8.7  8.4  8.5  8.8  8.7
## BRACKEN,J.J.    7.3  6.4  4.3  6.5  6.0  6.2  5.7  5.7  5.1  5.3  5.5  4.8
## BURNS,E.B.      6.2  8.8  8.7  8.5  7.9  8.0  8.1  8.0  8.0  8.0  8.6  8.6
```

```
str(USJudgeRatings)
```

```
## 'data.frame':    43 obs. of  12 variables:
##  $ CONT: num  5.7 6.8 7.2 6.8 7.3 6.2 10.6 7 7.3 8.2 ...
##  $ INTG: num  7.9 8.9 8.1 8.8 6.4 8.8 9 5.9 8.9 7.9 ...
##  $ DMNR: num  7.7 8.8 7.8 8.5 4.3 8.7 8.9 4.9 8.9 6.7 ...
##  $ DILG: num  7.3 8.5 7.8 8.8 6.5 8.5 8.7 5.1 8.7 8.1 ...
##  $ CFMG: num  7.1 7.8 7.5 8.3 6 7.9 8.5 5.4 8.6 7.9 ...
##  $ DECI: num  7.4 8.1 7.6 8.5 6.2 8 8.5 5.9 8.5 8 ...
##  $ PREP: num  7.1 8 7.5 8.7 5.7 8.1 8.5 4.8 8.4 7.9 ...
##  $ FAMI: num  7.1 8 7.5 8.7 5.7 8 8.5 5.1 8.4 8.1 ...
##  $ ORAL: num  7.1 7.8 7.3 8.4 5.1 8 8.6 4.7 8.4 7.7 ...
##  $ WRIT: num  7 7.9 7.4 8.5 5.3 8 8.4 4.9 8.5 7.8 ...
##  $ PHYS: num  8.3 8.5 7.9 8.8 5.5 8.6 9.1 6.8 8.8 8.5 ...
##  $ RTEN: num  7.8 8.7 7.8 8.7 4.8 8.6 9 5 8.8 7.9 ...
```

We are provided with n = 43 observations and p = 12 quantitative variables.
An observation is the different ratings received by a judge (given by his name) in the US Superior Court in 1977.

```
colnames(USJudgeRatings)
```

```
##  [1] "CONT" "INTG" "DMNR" "DILG" "CFMG" "DECI" "PREP" "FAMI" "ORAL" "WRIT"
## [11] "PHYS" "RTEN"
```

The variables are:

- CONT : Number of contacts of lawyer with judge
- INTG : Judicial integrity
- DMNR : Demeanor
- DILG : Diligence
- CFMG : Case flow managing
- DECI : Prompt decisions
- PREP : Preparation for trial
- FAMI : Familiarity with law
- ORAL : Sound oral rulings
- WRIT : Sound written rulings
- PHYS : Physical ability
- RTEN : Worthy of retention

General description of the data

```
sum(is.na(USJudgeRatings))
```

```
## [1] 0
```

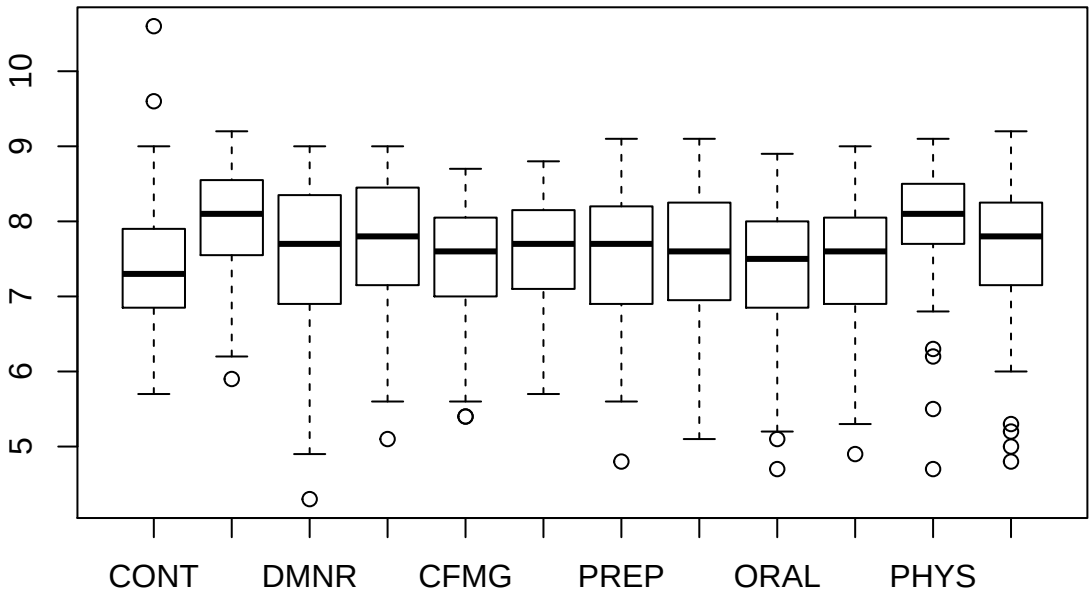
There are no missing values in the data frame.

```
summary(USJudgeRatings)
```

```
##       CONT             INTG            DMNR            DILG
##  Min.   : 5.700   Min.   :5.900   Min.   :4.300   Min.   :5.100
##  1st Qu.: 6.850   1st Qu.:7.550   1st Qu.:6.900   1st Qu.:7.150
##  Median : 7.300   Median :8.100   Median :7.700   Median :7.800
##  Mean   : 7.437   Mean   :8.021   Mean   :7.516   Mean   :7.693
##  3rd Qu.: 7.900   3rd Qu.:8.550   3rd Qu.:8.350   3rd Qu.:8.450
##  Max.   :10.600   Max.   :9.200   Max.   :9.000   Max.   :9.000
##       CFMG            DECI            PREP            FAMI
##  Min.   :5.400   Min.   :5.700   Min.   :4.800   Min.   :5.100
##  1st Qu.:7.000   1st Qu.:7.100   1st Qu.:6.900   1st Qu.:6.950
##  Median :7.600   Median :7.700   Median :7.700   Median :7.600
##  Mean   :7.479   Mean   :7.565   Mean   :7.467   Mean   :7.488
##  3rd Qu.:8.050   3rd Qu.:8.150   3rd Qu.:8.200   3rd Qu.:8.250
##  Max.   :8.700   Max.   :8.800   Max.   :9.100   Max.   :9.100
##       ORAL            WRIT            PHYS            RTEN
##  Min.   :4.700   Min.   :4.900   Min.   :4.700   Min.   :4.800
##  1st Qu.:6.850   1st Qu.:6.900   1st Qu.:7.700   1st Qu.:7.150
##  Median :7.500   Median :7.600   Median :8.100   Median :7.800
##  Mean   :7.293   Mean   :7.384   Mean   :7.935   Mean   :7.602
##  3rd Qu.:8.000   3rd Qu.:8.050   3rd Qu.:8.500   3rd Qu.:8.250
##  Max.   :8.900   Max.   :9.000   Max.   :9.100   Max.   :9.200
```

All the variables (except the variable CONT) seem to be ranged between 0 and 10.
The last variable, RTEN, seems to conclude the analysis. In fact, it says if the lawyers think that the judge is worthy staying in the US Superior Court or not.
First, we can observe that each variable seems to follow a symetric distribution, since median and mean are always close. Are u sure? because sometimes the difference is big for values between 5 and 10.

```
Outvals = boxplot(USJudgeRatings)
```



We observe the presence of outliers for 10 of the 12 variables (with larger values for CONT and with lower values for the other variables).

We can take a look on some outliers.

```
max(USJudgeRatings$CONT)
```

```
## [1] 10.6
```

```
rownames(USJudgeRatings)[which.max(USJudgeRatings$CONT)]
```

```
## [1] "CALLAHAN,R.J."
```

The judge with the biggest number of contacts of lawyer is judge Callahan with a a number of 10.6 contacts.

```
min(USJudgeRatings$RTEN)
```

```
## [1] 4.8
```

```
rownames(USJudgeRatings)[which.min(USJudgeRatings$RTEN)]
```

```
## [1] "BRACKEN,J.J."
```

The judge with the lowest rating for worthiness of retention is judge Bracken with a rating of 4.8.

```
max(USJudgeRatings$RTEN)
```

```
## [1] 9.2
```

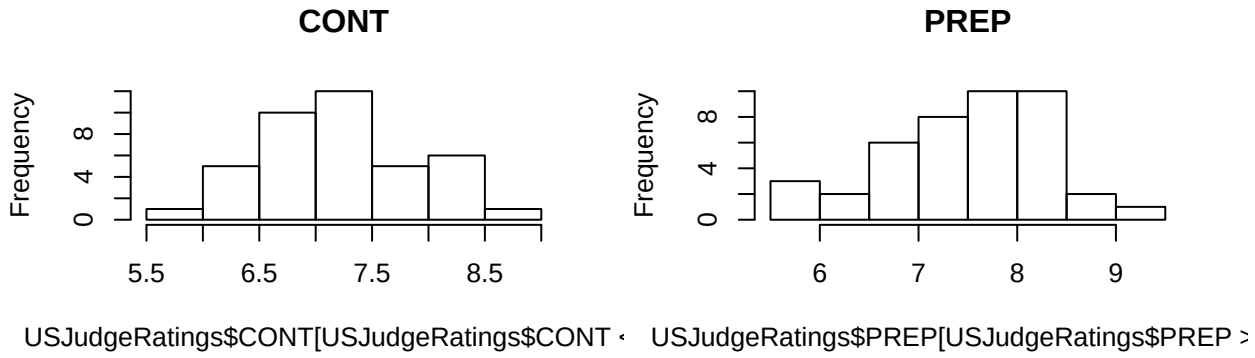
```
rownames(USJudgeRatings)[which.max(USJudgeRatings$RTEN)]
```

```
## [1] "RUBINOW,J.E."
```

The judge with the highest rating for worthiness of retention is judge Rubinow with a rating of 9.2.
We are not provided with extra information and we cannot check wether the outliers correspond to mistakes. Thus, we will assume that they aren't mistakes.

Descriptive statistics analysis of the dataset

```
par(mfrow=c(2,2))
hist(USJudgeRatings$CONT[USJudgeRatings$CONT<9], main="CONT")
hist(USJudgeRatings$PREP[USJudgeRatings$PREP>5], main="PREP" )
hist(USJudgeRatings$WRIT, main="WRIT")
hist(USJudgeRatings$RTEN, main="RTEN")
```




```
round(sqrt(diag(var(USJudgeRatings))),2)
```

```
## CONT INTG DMNR DILG CFMG DECI PREP FAMI ORAL WRIT PHYS RTEN
## 0.94 0.77 1.14 0.90 0.86 0.80 0.95 0.95 1.01 0.96 0.94 1.10
```

```
print('The smallest standard deviation is: ')
```

```
## [1] "The smallest standard deviation is: "
```

```
min(round(sqrt(diag(var(USJudgeRatings))),2))
```

```
## [1] 0.77
```

```
print('The largest standard deviation is: ')
```

```
## [1] "The largest standard deviation is: "
```

```
max(round(sqrt(diag(var(USJudgeRatings))),2))
```

```
## [1] 1.14
```

Regarding the dispersion, we look at the interquartile range (given by the boxplots) and the empirical standard deviation. Overall, the dispersions are not very high (around 1). We find that the variables DMNR and RTEN have the largest standard deviation, while the DECI variable has the smallest.

Let's measure the correlations between the 11 first variables and the variable RTEN.

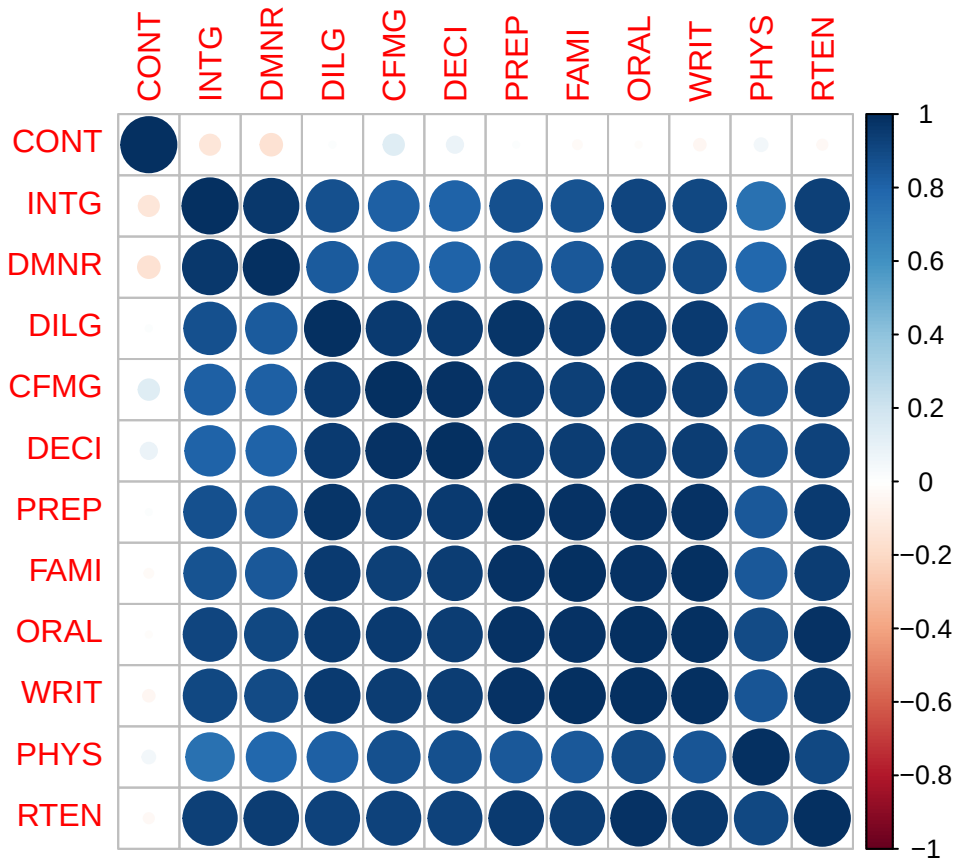
```
round(cor(USJudgeRatings),2)
```

```
##       CONT  INTG  DMNR DILG CFMG DECI PREP  FAMI  ORAL  WRIT PHYS  RTEN
## CONT  1.00 -0.13 -0.15 0.01 0.14 0.09 0.01 -0.03 -0.01 -0.04 0.05 -0.03
## INTG -0.13  1.00  0.96 0.87 0.81 0.80 0.88  0.87  0.91  0.91 0.74  0.94
## DMNR -0.15  0.96  1.00 0.84 0.81 0.80 0.86  0.84  0.91  0.89 0.79  0.94
## DILG  0.01  0.87  0.84 1.00 0.96 0.96 0.98  0.96  0.95  0.96 0.81  0.93
## CFMG  0.14  0.81  0.81 0.96 1.00 0.98 0.96  0.94  0.95  0.94 0.88  0.93
## DECI  0.09  0.80  0.80 0.96 0.98 1.00 0.96  0.94  0.95  0.95 0.87  0.92
## PREP  0.01  0.88  0.86 0.98 0.96 0.96 1.00  0.99  0.98  0.99 0.85  0.95
## FAMI -0.03  0.87  0.84 0.96 0.94 0.94 0.99  1.00  0.98  0.99 0.84  0.94
## ORAL -0.01  0.91  0.91 0.95 0.95 0.95 0.98  0.98  1.00  0.99 0.89  0.98
## WRIT -0.04  0.91  0.89 0.96 0.94 0.95 0.99  0.99  0.99  1.00 0.86  0.97
## PHYS  0.05  0.74  0.79 0.81 0.88 0.87 0.85  0.84  0.89  0.86 1.00  0.91
## RTEN -0.03  0.94  0.94 0.93 0.93 0.92 0.95  0.94  0.98  0.97 0.91  1.00
```

```
library(corrplot)
```

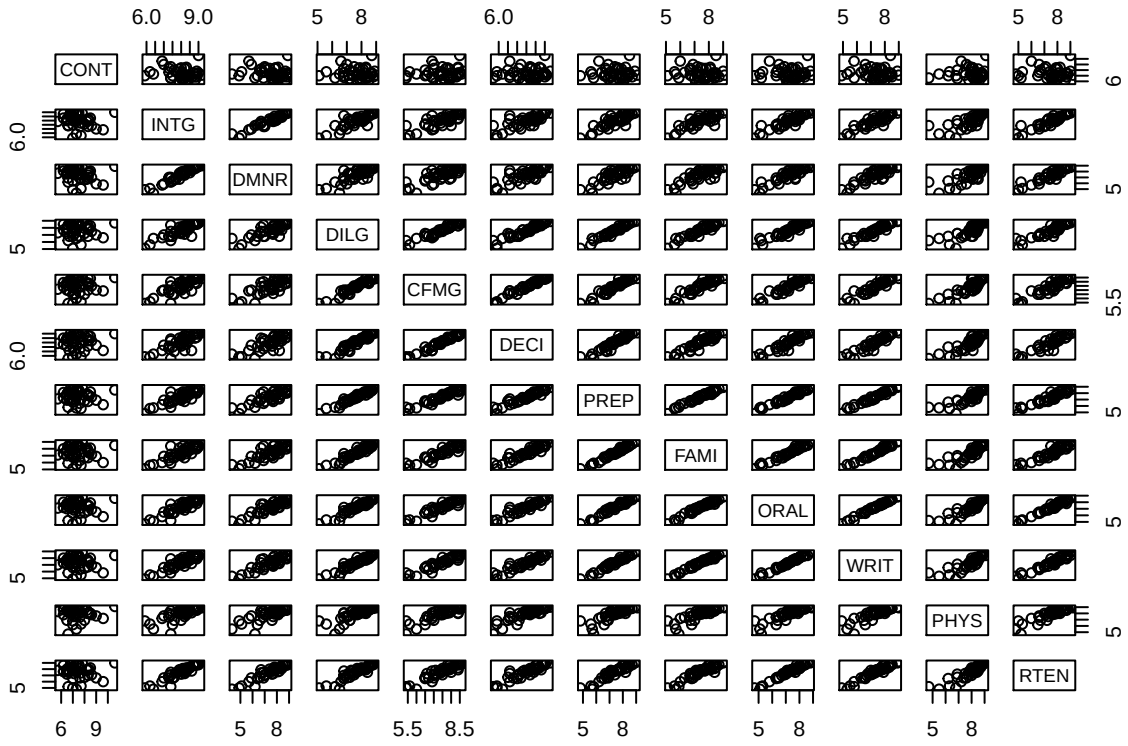
```
## corrplot 0.84 loaded
```

```
corrplot(cor(USJudgeRatings))
```

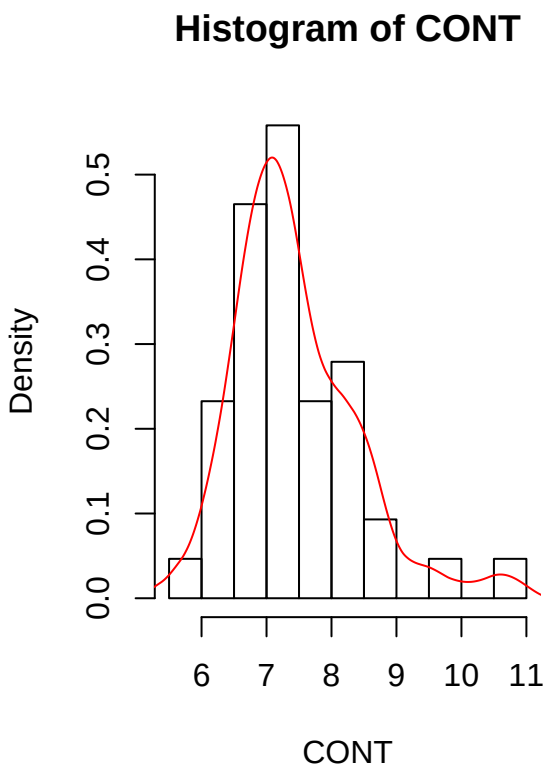


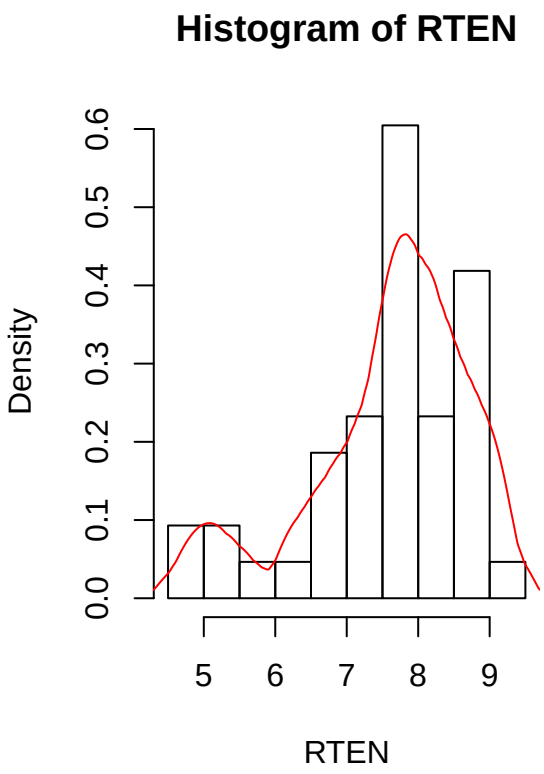
All the variables have a strong positive correlation two by two except the variable CONT which is not correlated to all the other variables. The number of contacts of a lawyer with the judge doesn't seem to explain the ratings received by the judge.

```
pairs(USJudgeRatings)
```

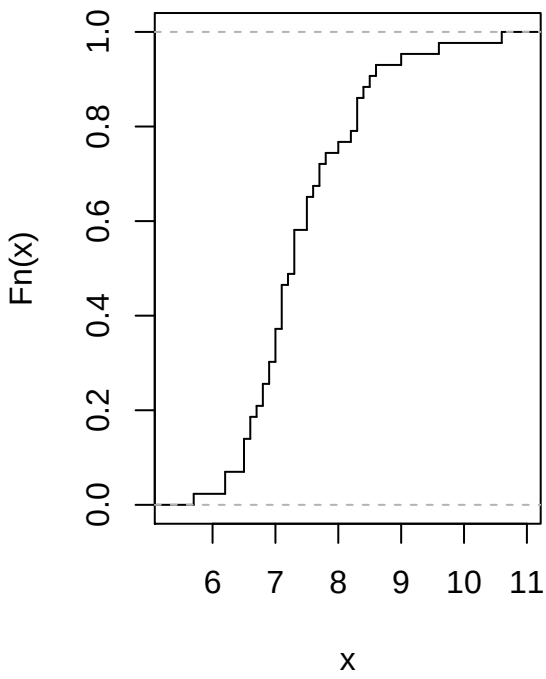


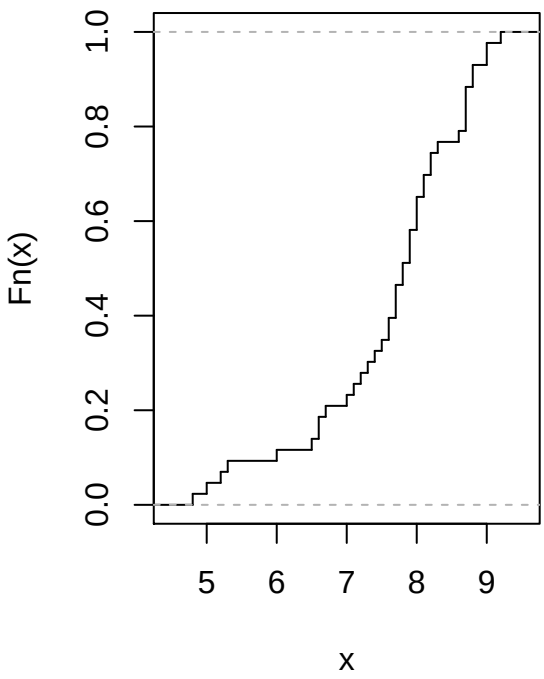

```
par(mfrow=c(1,2))
hist(USJudgeRatings$CONT, probability= TRUE, main="Histogram of CONT", xlab="CONT")
d = density(USJudgeRatings$CONT, kernel = 'c', bw = 0.3)
lines(d, col="red")

hist(USJudgeRatings$RTEN, probability= TRUE, main="Histogram of RTEN" , xlab="RTEN")
d = density(USJudgeRatings$RTEN, kernel = 'o', bw = 0.3)
lines(d, col="red")
```



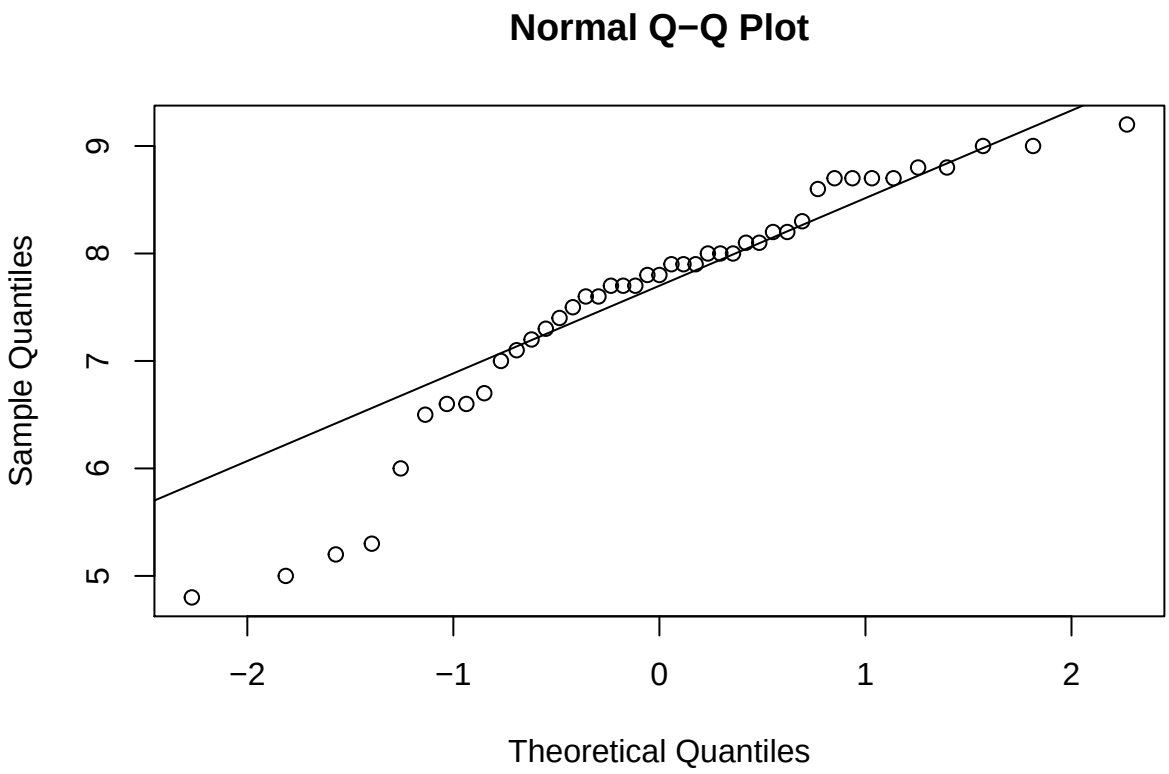



```
par(mfrow=c(1,2))
plot(ecdf(USJudgeRatings$CONT), verticals = TRUE, do.points = FALSE, main = "ECDF CONT")
plot(ecdf(USJudgeRatings$RTEN), verticals = TRUE, do.points = FALSE, main = "ECDF RTEN")
```






```
qqnorm(USJudgeRatings$RTEN)
qqline(USJudgeRatings$RTEN)
```



The QQ plots suggests that the RTEN variable is Gaussian.

Explaining the RTEN variable with a regression model

We will use RTEN as our dependent variable and try to explain it by fitting a regression model. We will try to find which of the other 11 variables explain the best our dependant variable and therefore which criterion are the most important for lawyers when evaluating if a judge is fit to stay at the Supreme Court.

```
library(e1071)
kurtosis
```

```
## function (x, na.rm = FALSE, type = 3)
## {
##     if (any(ina <- is.na(x))) {
##         if (na.rm)
##             x <- x[!ina]
##         else return(NA)
##     }
##     if (!(type %in% (1:3)))
##         stop("Invalid 'type' argument.")
##     n <- length(x)
##     x <- x - mean(x)
##     r <- n * sum(x^4)/(sum(x^2)^2)
##     y <- if (type == 1)
##         r - 3
##     else if (type == 2) {
##         if (n < 4)
##             stop("Need at least 4 complete observations.")
##         ((n + 1) * (r - 3) + 6) * (n - 1)/((n - 2) * (n - 3))
```

```
##     }
##     else r * (1 - 1/n)^2 - 3
##     y
## }
## <bytecode: 0x7fcbd8af2798>
## <environment: namespace:e1071>
```

```
skewness
```

```
## function (x, na.rm = FALSE, type = 3)
## {
##     if (any(ina <- is.na(x))) {
##         if (na.rm)
##             x <- x[!ina]
##         else return(NA)
##     }
##     if (!(type %in% (1:3)))
##         stop("Invalid 'type' argument.")
##     n <- length(x)
##     x <- x - mean(x)
##     y <- sqrt(n) * sum(x^3)/(sum(x^2)^(3/2))
##     if (type == 2) {
##         if (n < 3)
##             stop("Need at least 3 complete observations.")
##         y <- y * sqrt(n * (n - 1))/(n - 2)
##     }
##     else if (type == 3)
##         y <- y * ((1 - 1/n))^(3/2)
##     y
## }
## <bytecode: 0x7fcbd8b1b560>
## <environment: namespace:e1071>
```