# Statistics | HW1

```r
# data preprocessing
ceodata <- read.csv('ceo.csv')
ceodata$X <- NULL
salary <- ceodata$salary
head(ceodata)
```

```
##   salary totcomp tenure age  sales profits assets
## 1   3030    8138      7  61 161315    2956 257389
## 2   6050   14530      0  51 144416   22071 237545
## 3   3571    7433     11  63 139208    4430  49271
## 4   3300   13464      6  60 100697    6370  92630
## 5  10000   68285     18  63 100469    9296 355935
## 6   9375   42381      6  57  81667    6328  86100
```

## Problem 1

### Task 1

#### 1a

```r
mean(salary)
```

```
## [1] 2027.517
```

Mean(2027.517) - average salary.

```r
mean.default(salary, trim=0.1)
```

```
## [1] 1710.092
```

Trimmed mean(1710.092) - mean of salaries without lowest 10% and highest 10% of values.
Used to eliminate the impact of very large or very small salaries(called outliers) on the mean.

```r
median(salary)
```

```
## [1] 1600
```

Median(1600) - central salary. Half of salaries are smaller than median and half are larger.

```r
quantile(salary, c(0.25, 0.75))
```

```
##    25%    75%
## 1084.0 2347.5
```

Lower quartile(1084.0) - 25% of salaries are smaller than lower quartile (consequently 75% are larger)
Upper quartile(2347.5) - 75% of salaries are smaller than upper quartile (consequently 25% are larger)

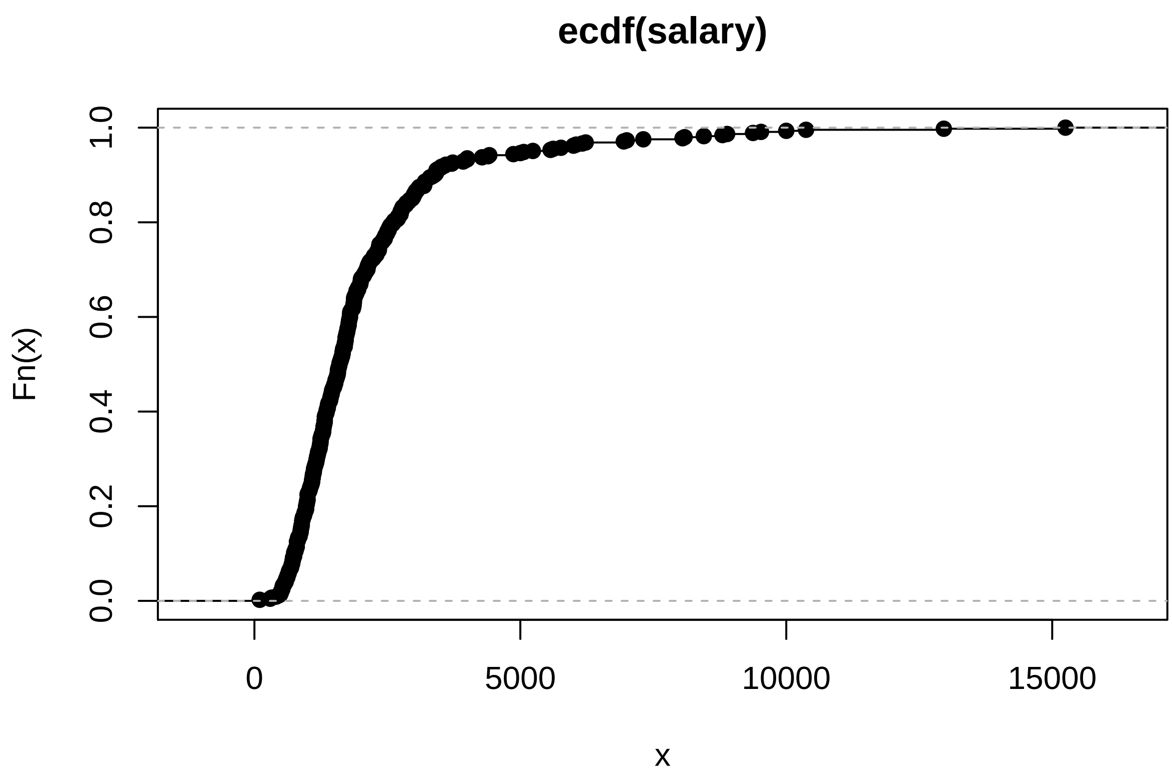
```r
quantile(salary, c(0.1,0.9))
```

```
##    10%    90%
##  750.0 3384.4
```

Lower 10%-quantile(750.0) - 10% of salaries are smaller than lower quartile (consequently 90% are larger)
Upper 10%-quantile(3384.4) - 90% of salaries are smaller than upper quartile (consequently 10% are larger)

### 1b

```r
Fn <- ecdf(salary)
plot(Fn)
```




Empirical cumulative distribution function of salaries.

```r
quantile(salary, c(0.2))
```

```
##   20%
## 976.2
```

```r
quantile(salary, c(0.8))
```

```
##  80%
## 2613
```

```r
Fn(1000)
```

```
## [1] 0.2237136
```

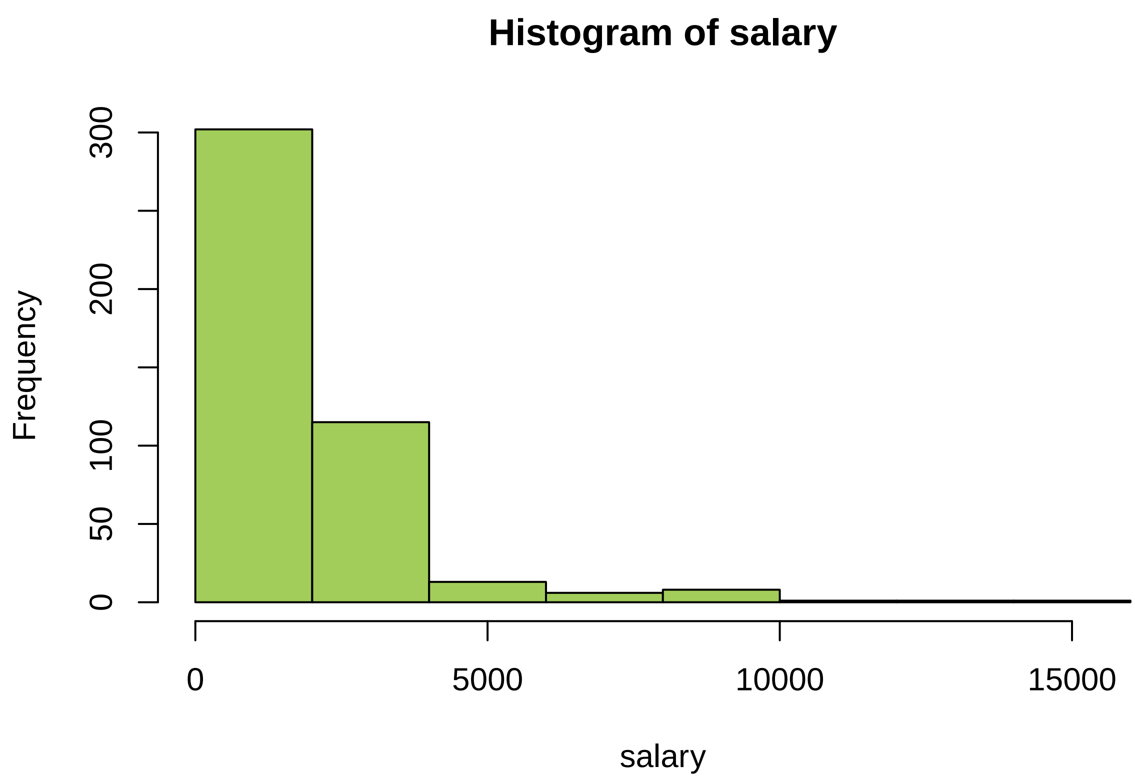
```r
1 - Fn(5000)
```

```
## [1] 0.05369128
```

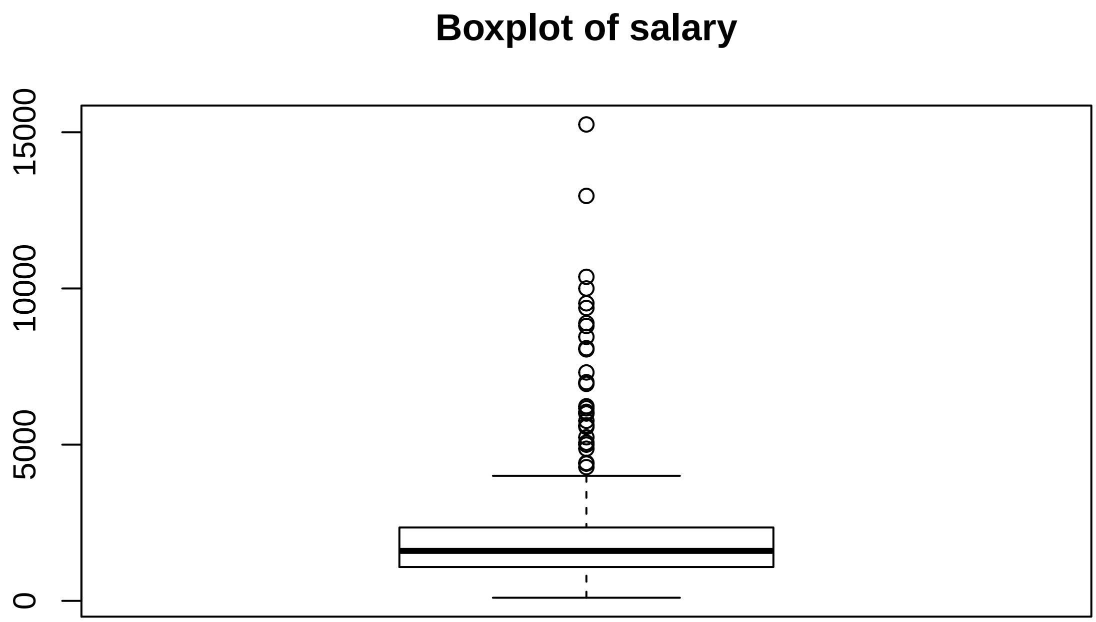
- $\hat{F}^{-1}(0.2) = 976.2$ - 20% of CEOs have at most \$976.2 salary.
  $\hat{F}^{-1}(0.8) = 2613$ - 80% of CEOs have at most \$2613 salary.
- $\hat{F}(1000) = 0.223$ - 22.3% of CEOs have at most \$1000 salary.
  $1 - \hat{F}(5000) = 0.053$ - 5.3% of CEOs have at least \$5000 salary.

**1c**

```r
hist(salary, col="darkolivegreen3")
```




```r
boxplot(salary, main="Boxplot of salary")
```




As we can see from histogram and boxplot, salary distribution is not symmetric.
Location measures:
**mean** is very sensitive to outliers and therefore meaningful only for symmetric data - not appropriate here.
**trimmed mean** is much more robust to outliers compared to the simple mean - appropriate here.
**median** is not as strongly influenced by outliers as mean - appropriate here.
**the interquartile range** is also robust to outliers. There are at least [n/2] of all observations in the interval - appropriate here.

```
library(moments)
skewness(salary)
```

```
## [1] 3.391005
```

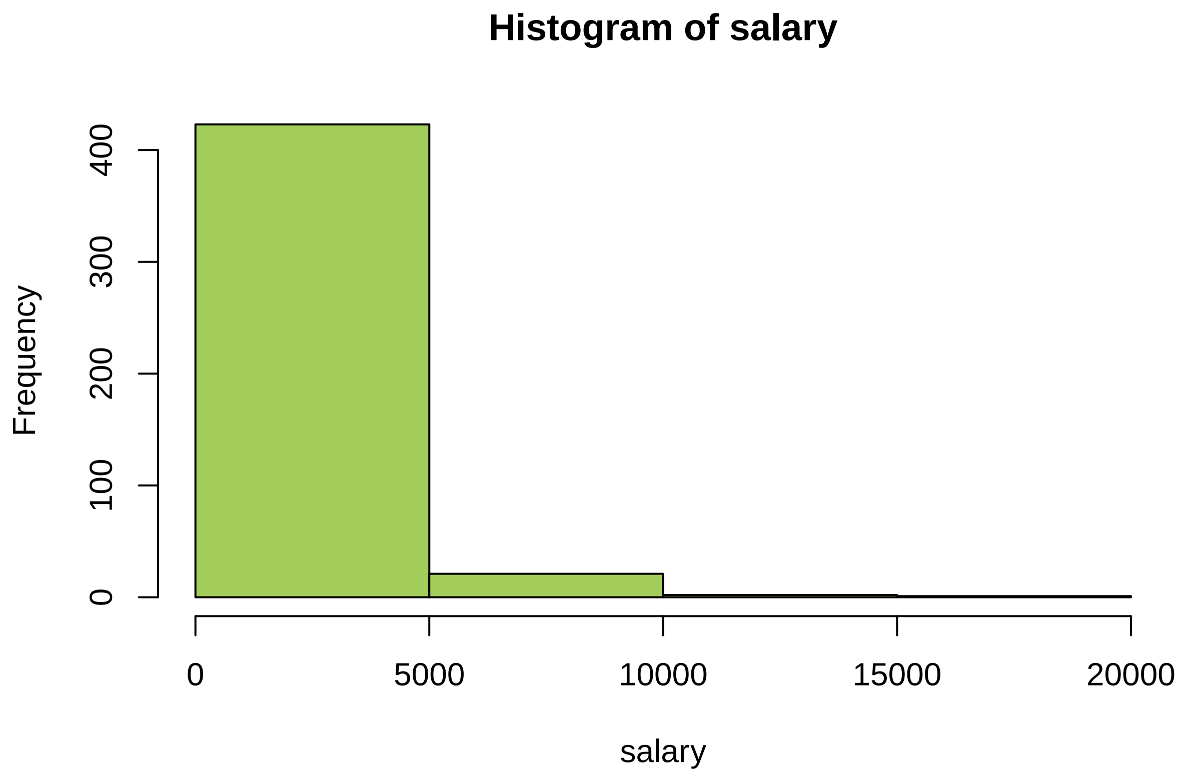
As a measure of symmetry we can use skewness. If skewness is larger than zero, then the distribution is right-skewed, therefor salary distribution is right-skewed.

### 1d

```
hist(salary, col="darkolivegreen3")
```

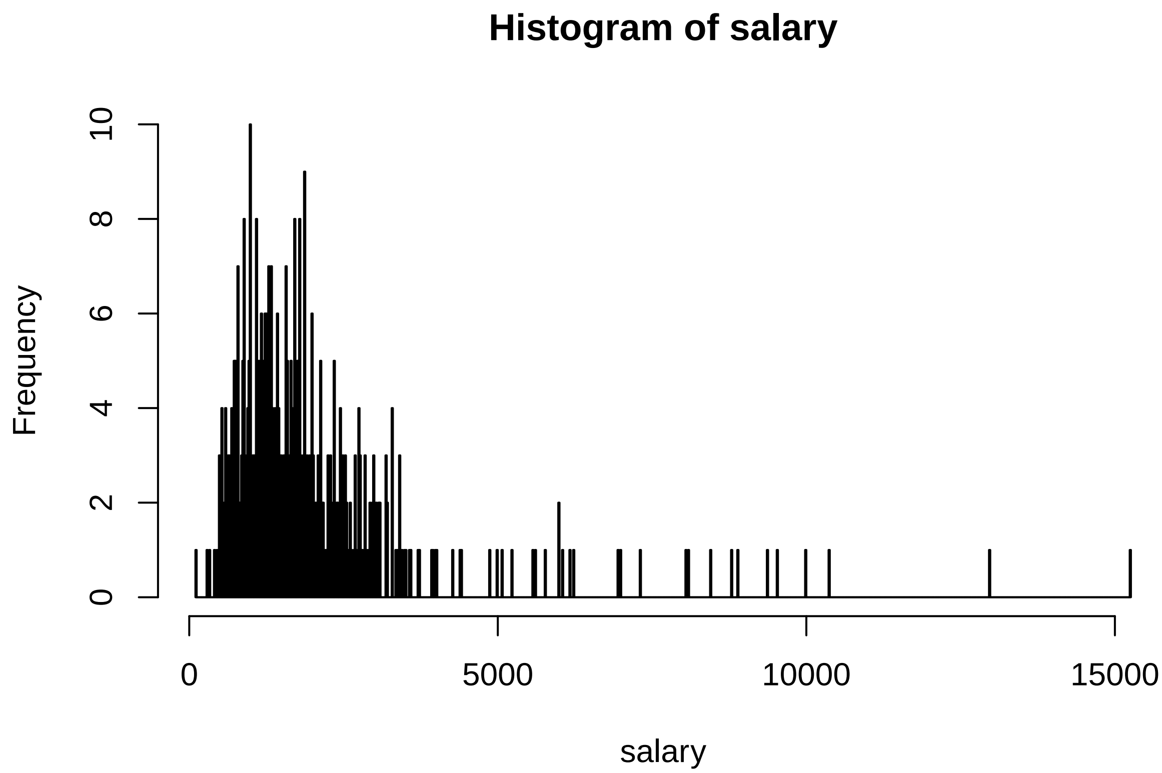

Histogram of salary. Default formula to compute number of bars is Sturges' formula ($k = \lceil \log_2 n \rceil + 1$) For salary data $n = 447$, $k = \lceil \log_2 447 \rceil + 1 = 10$.

```
hist(salary, breaks=4, col="darkolivegreen3")
```




Too rough histogram with only 4 bars.

```
hist(salary, breaks=1000, col="darkolivegreen3")
```



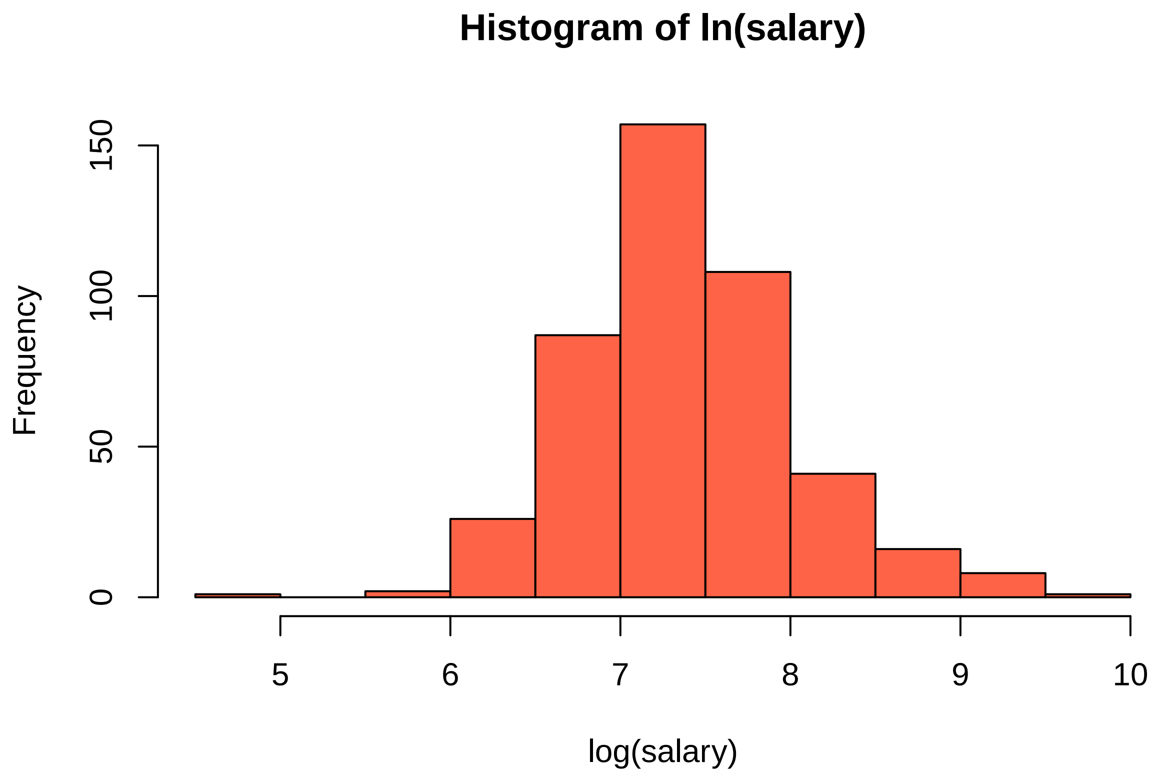

Too detailed histogram with 1000 bars.

We can see that too detailed histogram shows too much individual data and we can't clearly see the underlying pattern. On the other hand, too rough histogram has only 4 bars and again we are unable to find underlying pattern in the data.

### 1e

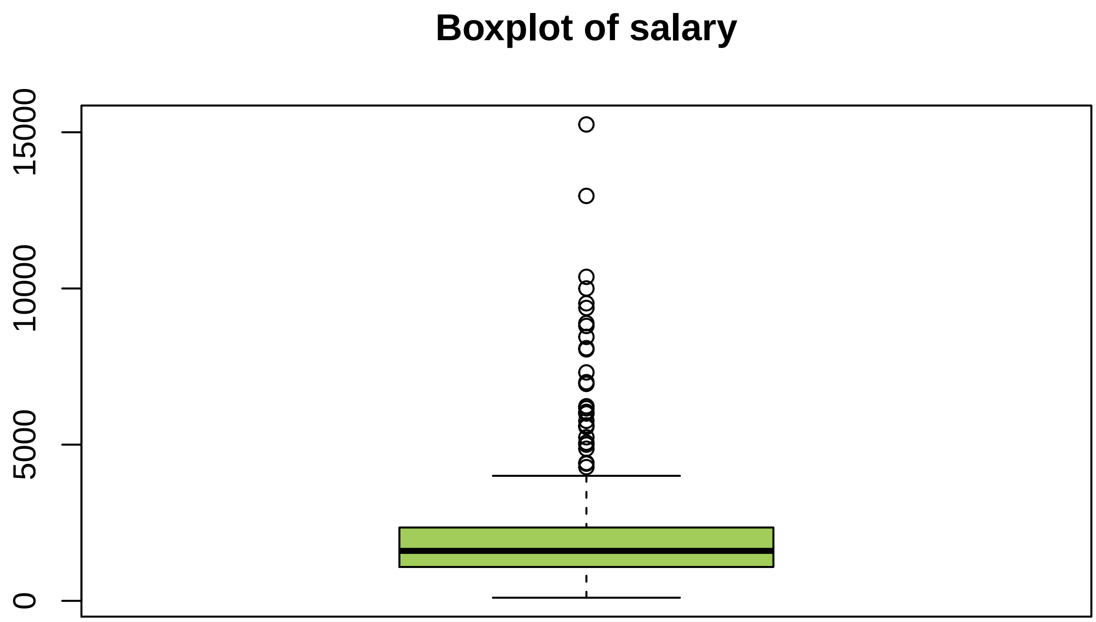
```
hist(salary, col="darkolivegreen3")
```


```
hist(log(salary), col="tomato", main="Histogram of ln(salary)")
```

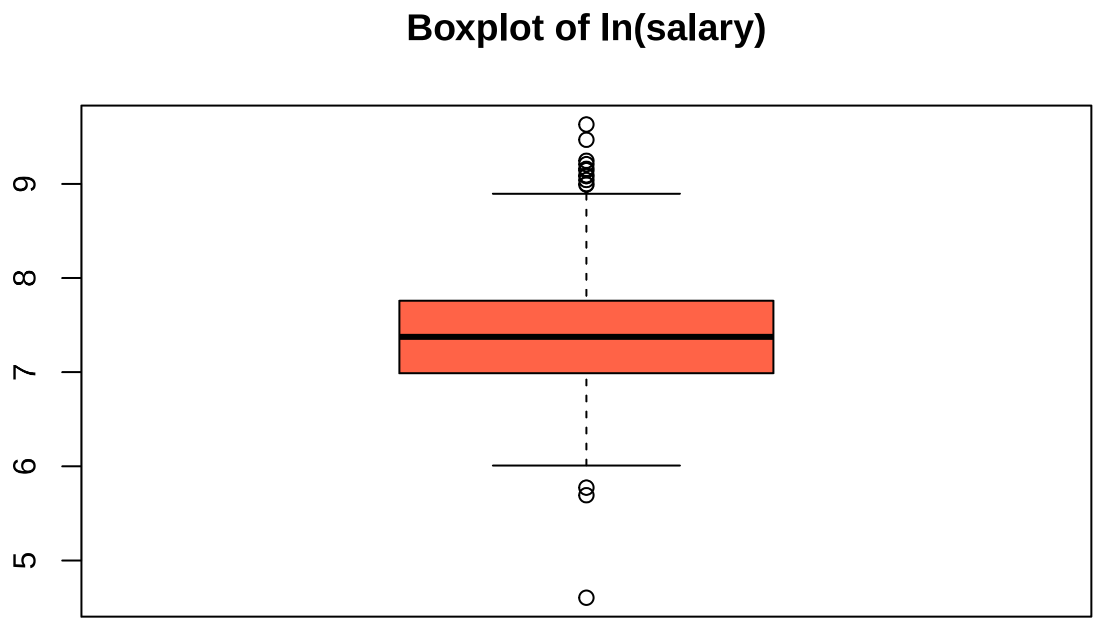



Histogram of ln(salary)

```
boxplot(salary, main="Boxplot of salary",col="darkolivegreen3")
```




```r
boxplot(log(salary),main="Boxplot of ln(salary)", col="tomato")
```




Boxplot of ln(salary). We can see that this is almost symmetric distribution.

```
mean(log(salary))
```

```
## [1] 7.391898
```

```
median(log(salary))
```

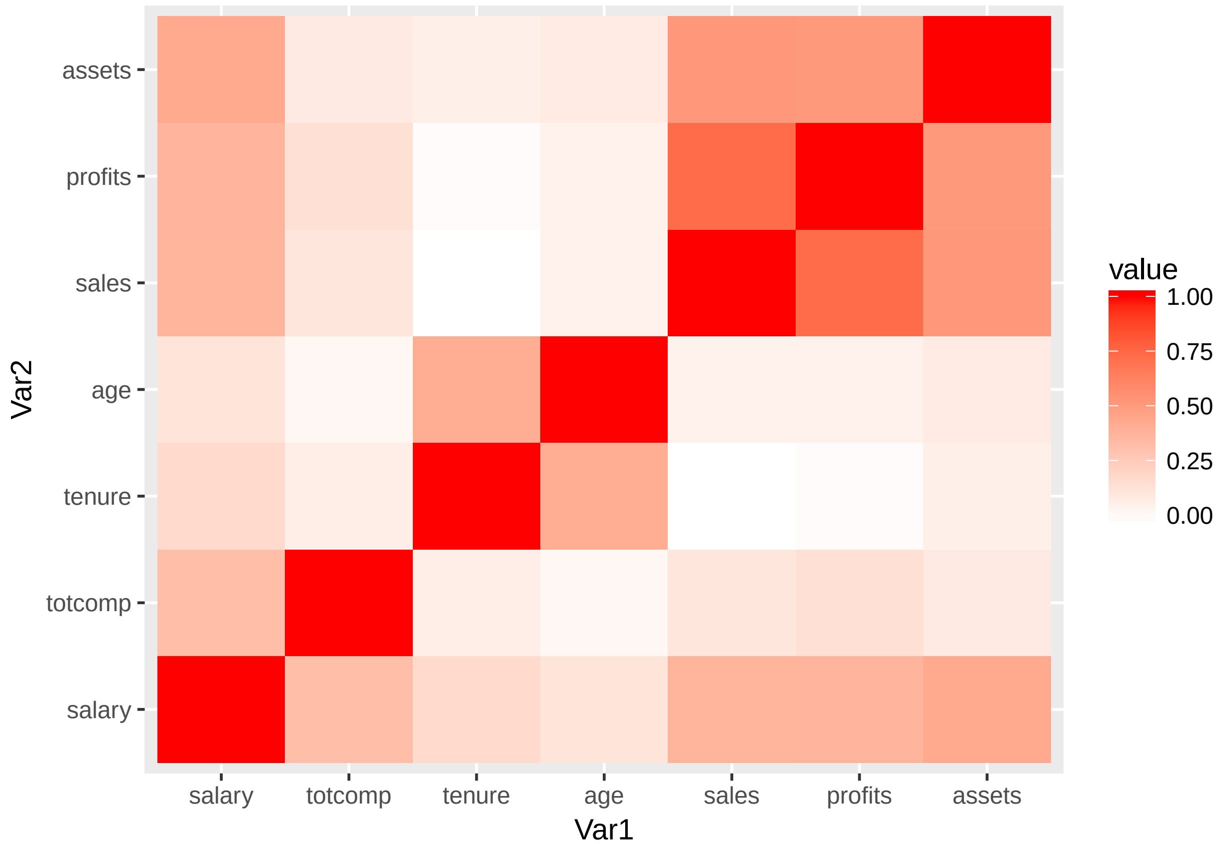
```
## [1] 7.377759
```

Mean and median of ln(salary).
In a symmetric distribution, the mean and median fall at the same point. As we can see mean is pretty much the same as median.

## Task 2

### 2a

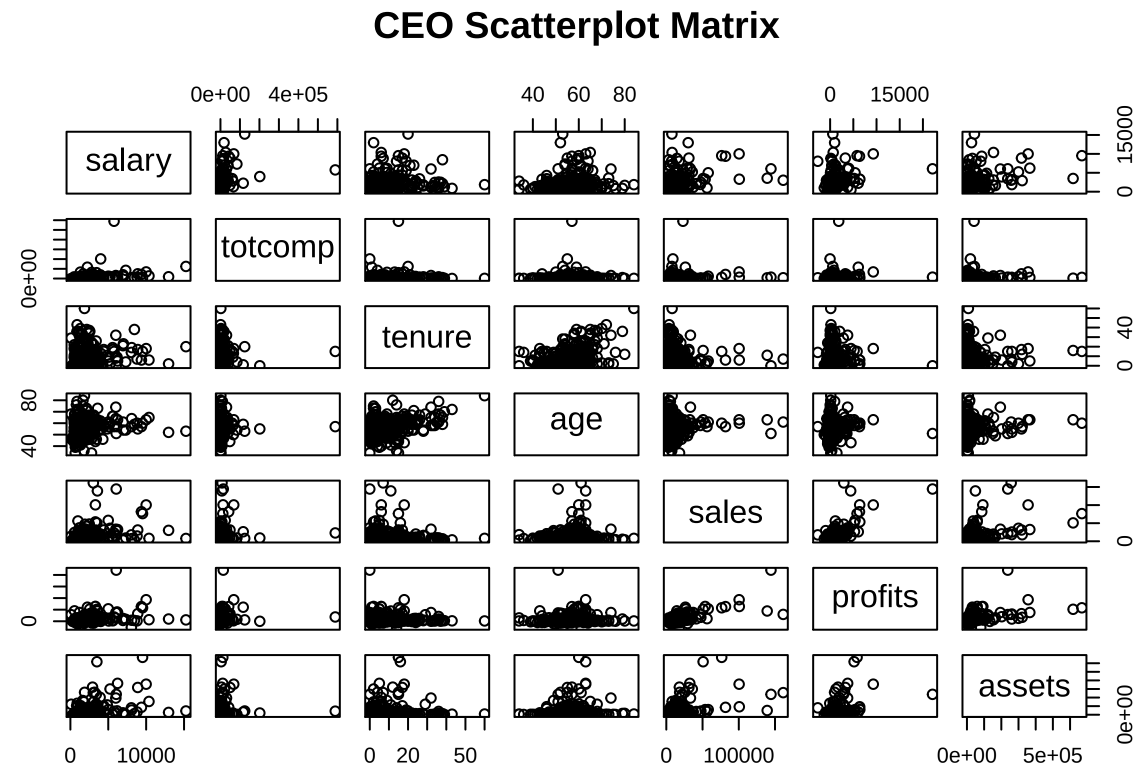
```
library(ggplot2)
library(reshape2)
pearson_correlations = cor(ceodata, method="pearson")
qplot(x=Var1, y=Var2, data=melt(pearson_correlations), fill=value, geom="tile") + scale_fill_gradient(l
```




Heatmap of Pearson correlations. The more bright red square, the bigger correlation.
For example, we can see that there are pretty strong correlation between sales and profits, which is very logical.
Note: Pearson correlation evaluates the linear relationship between two continuous variables.

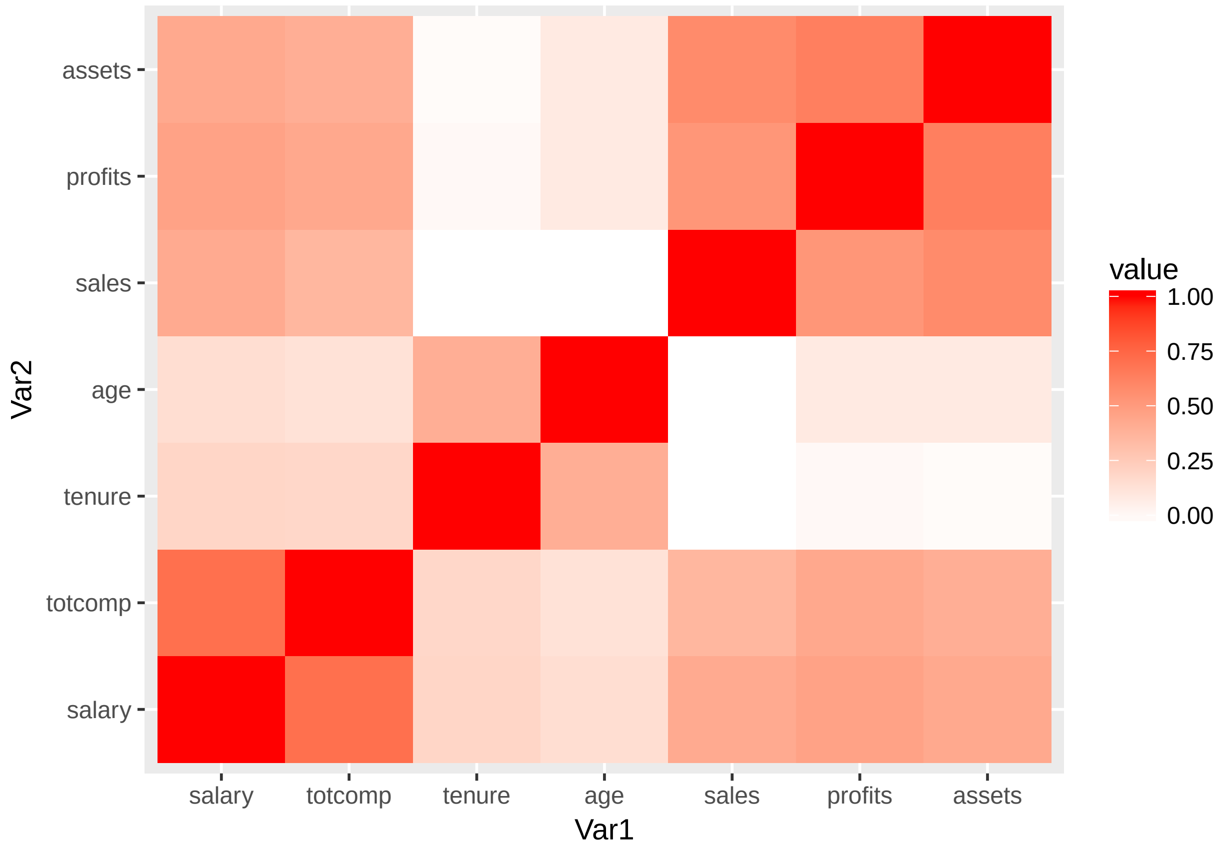
### 2b

```
pairs(~salary + totcomp + tenure + age + sales + profits + assets, data=ceodata, main="CEO Scatterplot M
```




Here we can see almost linear relationship between some variables, for instance sales and profits.
At the same time there are nonlinear correlations and no correlations.
So I think that for some variables linear correlation coefficients are appropriate here.

```
spearman_correlation = cor(ceodata, method="spearman")
qplot(x=Var1, y=Var2, data=melt(spearman_correlation), fill=value, geom="tile") + scale_fill_gradient(l
```




Heatmap of Spearman correlations. Here we can see more strong correlations between some variables than on heatmap using Pearson correlations.
This could happen because of nonlinear but monotonic relationship between variables.
For example, totcomp and salary or totcomp and profits.
Note: The Spearman correlation evaluates the monotonic relationship between two continuous variables.

```
sorted_salaries = sort(ceodata$salary)
min(which(sorted_salaries==6000))
```

```
## [1] 429
```

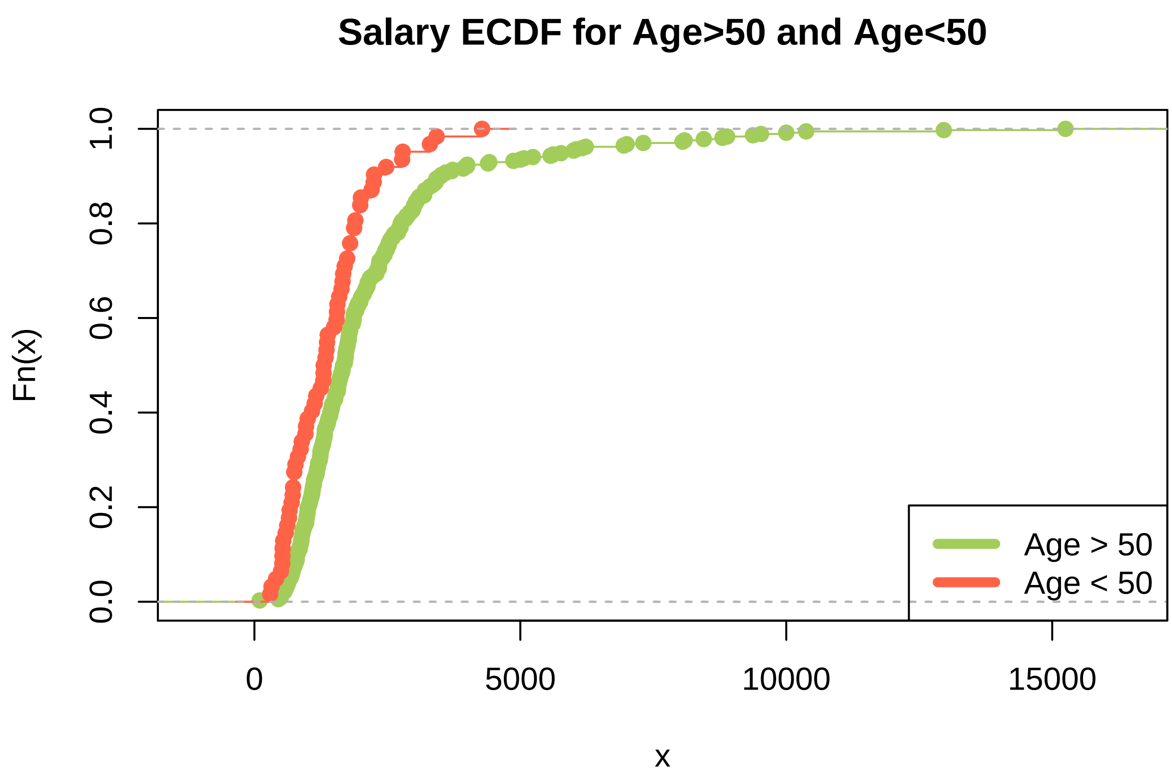
Min rank of salary=6000.

### 2c

```
hist(salary[which(ceodata$age > 50)],  col="darkolivegreen3", main="Age>50 and Age<50 Histograms Of Sala
hist(salary[which(ceodata$age < 50)],  col="tomato", add=T)
legend("topright", legend=c("Age > 50", "Age < 50"),col=c("darkolivegreen3", "tomato"), lwd=3)
```


Histograms of two groups: Age > 50 and Age < 50.
From this plot we can see that there are far less CEOs with age < 50 comparing to CEOs with age > 50.

```
plot(ecdf(salary[which(ceodata$age > 50)]), col="darkolivegreen3", main="Salary ECDF for Age>50 and Age<
plot(ecdf(salary[which(ceodata$age < 50)]), add=T, col="tomato")
legend("bottomright", legend=c("Age > 50", "Age < 50"),col=c("darkolivegreen3", "tomato"), lwd=5)
```




ECDF plots of two groups: Age > 50 and Age < 50.
The distribution is pretty much the same for small salaries.
However, more CEO's with age > 50 get larger salaries. This can clearly be seen from ecdf plot. For example, let's took 0.95, we can see that 95% of CEOs > 50 years old get at most 5905.6
while 95% of CEOs < 50 years old get at most 2787.5

## Task 3

### 3a

```
grouped_data <- data.frame(salary=ceodata$salary, age=ceodata$age)
# group data by categories
grouped_data$age <- ifelse(grouped_data$age < 50, "a1", "a2")
grouped_data$salary[suppressWarnings(as.integer(grouped_data$salary)) < 2000] <- "s1"
grouped_data$salary[suppressWarnings(as.integer(grouped_data$salary) >= 2000) & suppressWarnings(as.inte
grouped_data$salary[suppressWarnings(as.integer(grouped_data$salary)) >= 4000] <- "s3"
```

```
con_table <- xtabs(~age+salary, data=grouped_data)
addmargins(con_table)
```

```
##      salary
## age    s1  s2  s3 Sum
##   a1   52   9   1  62
##   a2  248 107  30 385
##   Sum 300 116  31 447
```

Contigency table with absolute frequencies.

```r
con_table <- xtabs(~age+salary, data=grouped_data)
addmargins(con_table / nrow(grouped_data))
```

```
##      salary
## age            s1          s2          s3         Sum
##   a1  0.116331096 0.020134228 0.002237136 0.138702461
##   a2  0.554809843 0.239373602 0.067114094 0.861297539
##   Sum 0.671140940 0.259507830 0.069351230 1.000000000
```

Contigency table with relative frequencies.

#### 3b

We can see that there are small amount of CEOs < 50 years old.
Also, CEOs < 50 years get smaller salary comparing to CEOs > 50 years.
We can see that there are 62 CEOs under 50 years and 52 of them have < 2000 salary.

#### 3c

```r
con_table <- xtabs(~age+salary, data=grouped_data)
chisq.test(con_table)
```

```
## Warning in chisq.test(con_table): Chi-squared approximation may be
## incorrect
```

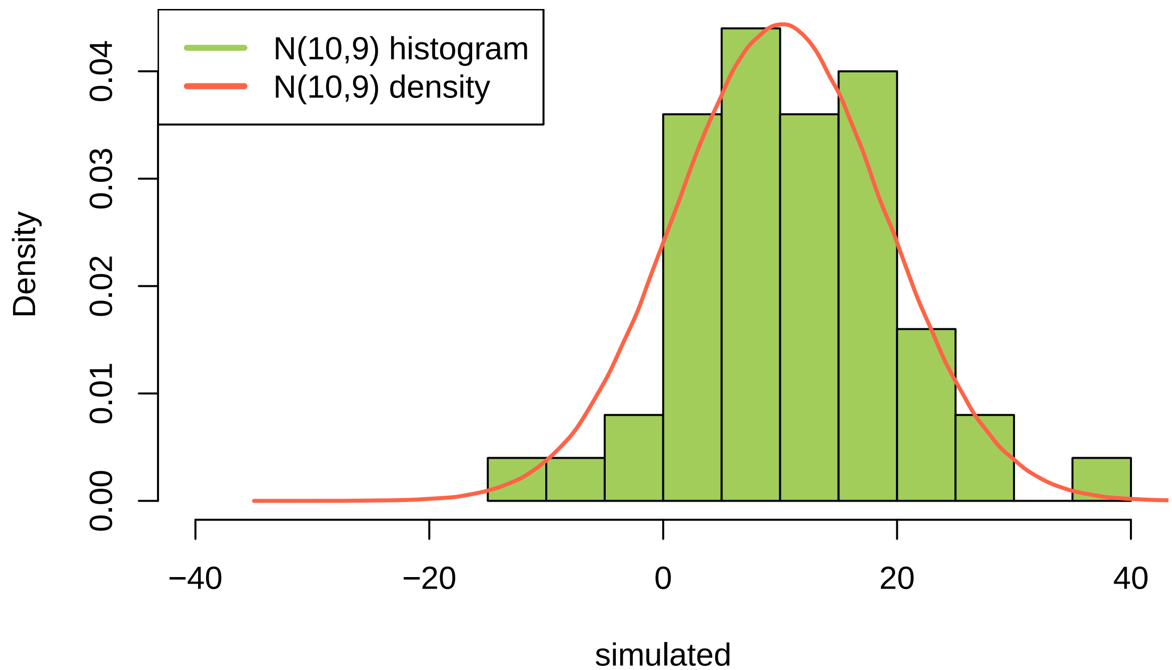
```
##
##  Pearson's Chi-squared test
##
## data:  con_table
## X-squared = 9.5787, df = 2, p-value = 0.008318
```

# 2 Problem

## Task 1

#### 1a

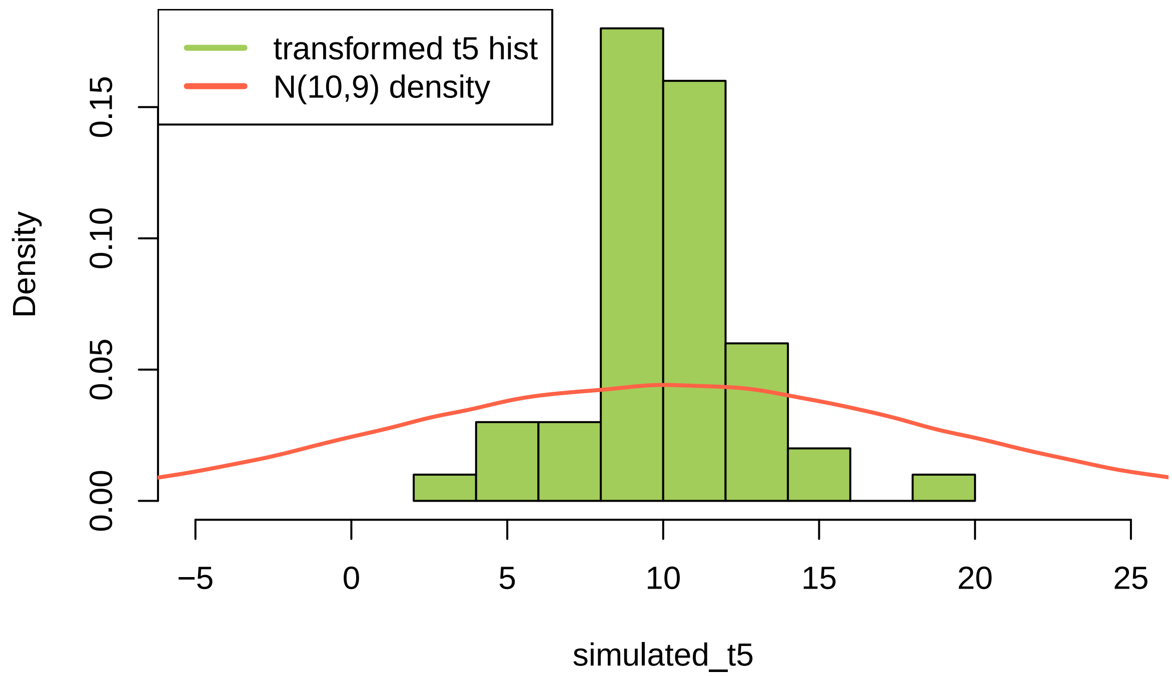
```r
set.seed(19)
simulated <- rnorm(50, 10, 9)
normal <- rnorm(1000000, 10, 9)
hist(simulated, col="darkolivegreen3",prob=TRUE, xlim=c(-40, 40), main="")
lines(density(normal), col="tomato", lwd=2)
legend("topleft", legend=c("N(10,9) histogram", "N(10,9) density"),col=c("darkolivegreen3", "tomato"), l
```




We can see that density plot of N(10, 9) is pretty the same as hist of simulated N(10,9).
However, we have only 50 samples in simulated N(10, 9), so It's not really clear that simulated is normal distribution.

**1b**

```
set.seed(19)
normal <- rnorm(100000, 10, 9)
simulated_t5 <- rt(50, df=5)
simulated_t5 <- 10 + 3 * sqrt(3 / 5) * simulated_t5
hist(simulated_t5, col="darkolivegreen3", prob=TRUE, xlim = c(-5, 25), main="")
lines(density(normal), col="tomato", lwd=2)
legend("topleft", legend=c("transformed t5 hist", "N(10,9) density"),col=c("darkolivegreen3", "tomato")
```




Here we can see hist of transformed t5 distribution and density plot of N(10, 9).
It is obvious that transformed t5 has higher density that N(10, 9).

## Task 2

### 2a

```
set.seed(19)
normal <- rnorm(50, 10, 9)
list <- c(normal)
p <- 49
for (i in 0:49){
  list = c(list, 16 + i * (24 - 16)/p)
}
mean(normal)
```

```
## [1] 11.09455
```

```
mean(list)
```

```
## [1] 15.54728
```

```
median(normal)
```

```
## [1] 10.80788
```

```r
median(list)
```

```
## [1] 17.71429
```

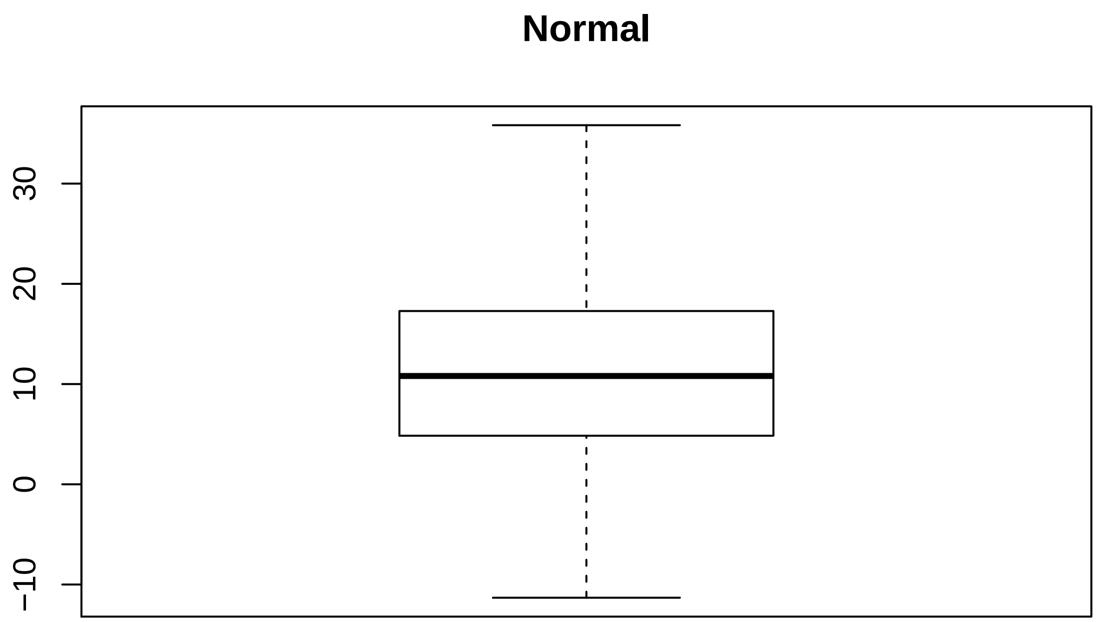
```r
var(normal)
```

```
## [1] 84.77684
```

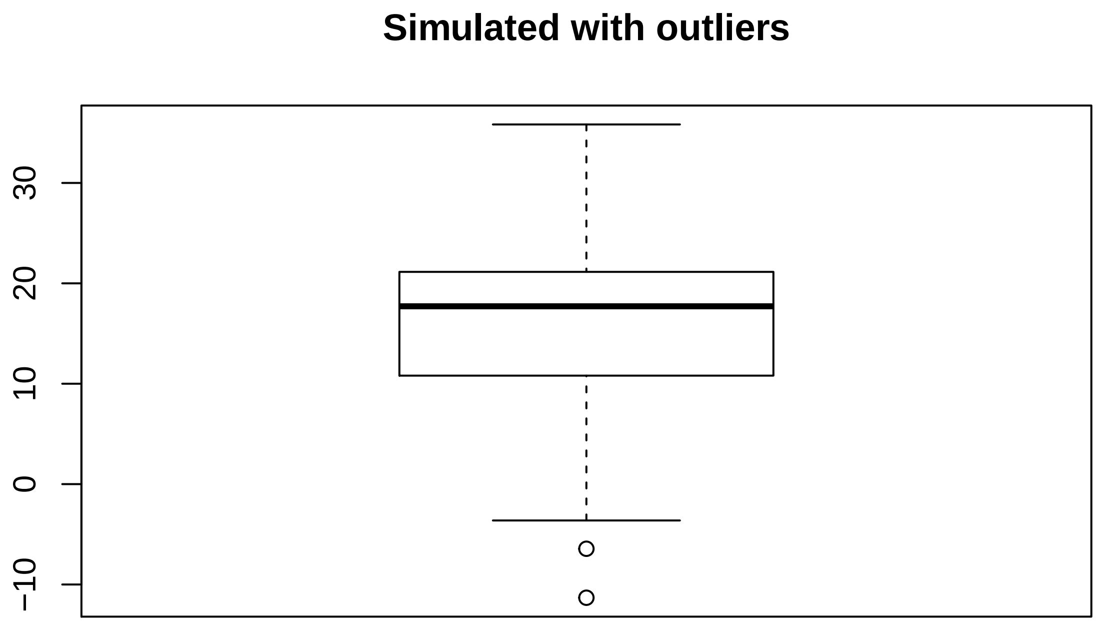
```r
var(list)
```

```
## [1] 64.79083
```

```r
boxplot(normal, main="Normal")
```




```r
boxplot(list, main="Simulated with outliers")
```




Here on boxplots we can clearly see the impact of adding outliers.
Median and mean become significantly larger.
Variance also changes significantly.

### 2b, 2c, 2d

Interactive graphics links:

- Boxplot animation
- Histogram animation

Here we can see how adding outliers changes location measures.

## Task 3

### 3b

```
u <- rnorm(50, 0, 1)
set.seed(10)
v <- rnorm(50, 0, 1)
p <- 0.7

v <- p*u + sqrt(1 - p * p) * v
```

**3c**

```r
set.seed(19)
u <- rnorm(10000, 0, 1)
set.seed(44)
v <- rnorm(10000, 0, 1)
p <- 0.75
v_tansformed <- p * u + sqrt(1 - p^2) * v
cor(data.frame(u, v_tansformed), method="pearson")
```

```
##                      u v_tansformed
## u            1.000000     0.744237
## v_tansformed 0.744237     1.000000
```

```r
cor(data.frame(u, v_tansformed), method="spearman")
```

```
##                       u v_tansformed
## u            1.0000000    0.7250769
## v_tansformed 0.7250769    1.0000000
```

Simulated U and V*. Pearson and Spearman correlation coefficients.

```r
# exp transformation
v_exp <- exp(v_tansformed)
cor(data.frame(u, v_exp), method="pearson")
```

```
##              u    v_exp
## u     1.000000 0.574222
## v_exp 0.574222 1.000000
```

```r
cor(data.frame(u, v_exp), method="spearman")
```

```
##               u     v_exp
## u     1.0000000 0.7250769
## v_exp 0.7250769 1.0000000
```

Here we can see that Spearman correlation coefficients remain unchanged and Pearson correlation coefficients changed.
This is because Spearman correlation is stable to transformations.