5291A1

September 13, 2024

#

STAT 5291 Advanced Data Analysis Problem Set 1

Francis Zhang xz3279

9/13/2024

i)Determine whether there are outliers in the combined data, using boxplots or other suitable methods.

[28]:
```
install.packages("bootstrap")
```

```
Installing package into ‘/usr/local/lib/R/site-library’
(as ‘lib’ is unspecified)
```

[2]:
```
install.packages("Sleuth3", repos="http://R-Forge.R-project.org")
```

```
Installing package into ‘/usr/local/lib/R/site-library’
(as ‘lib’ is unspecified)
```

[3]:
```
library(Sleuth3)
summary(case0102)
```

```
     Salary          Sex
 Min.   :3900   Female:61
 1st Qu.:4980   Male  :32
 Median :5400
 Mean   :5420
 3rd Qu.:6000
 Max.   :8100
```

[4]:
```
stem(case0102$Salary)
```

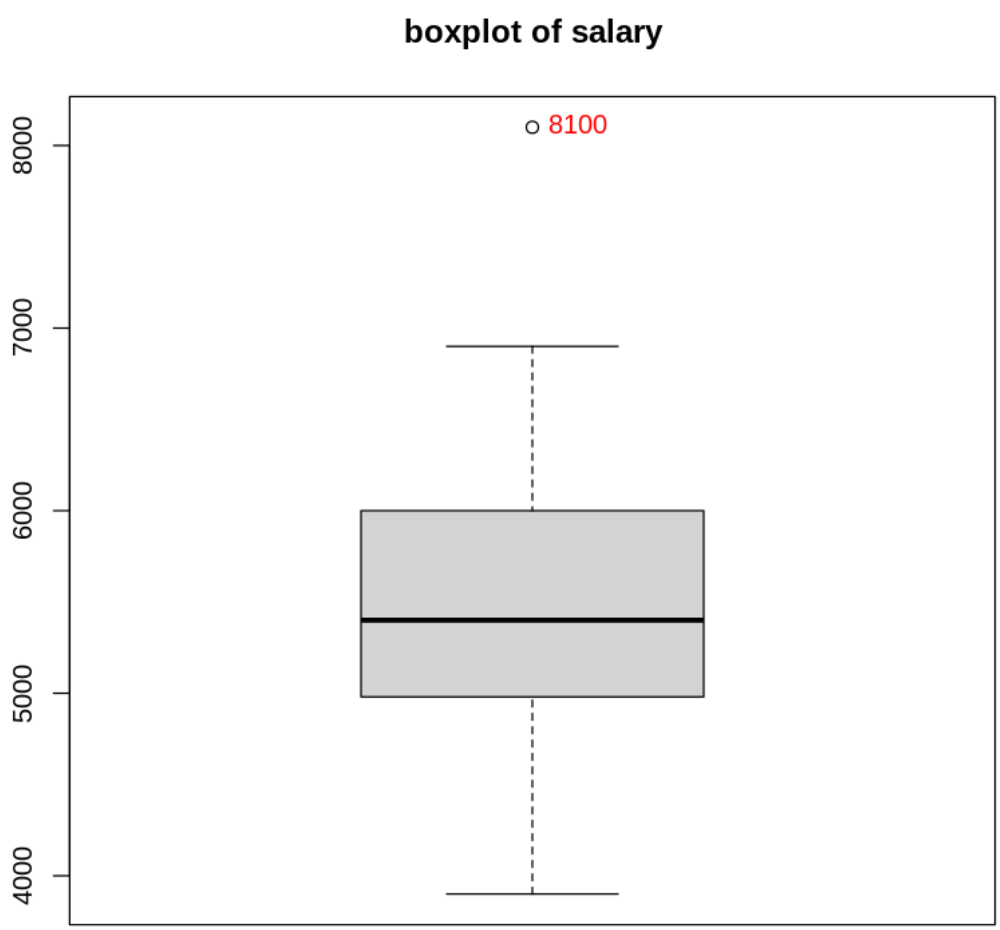
```
  The decimal point is 3 digit(s) to the right of the |

  3 | 9
  4 | 03444444
  4 | 55668888888888
```

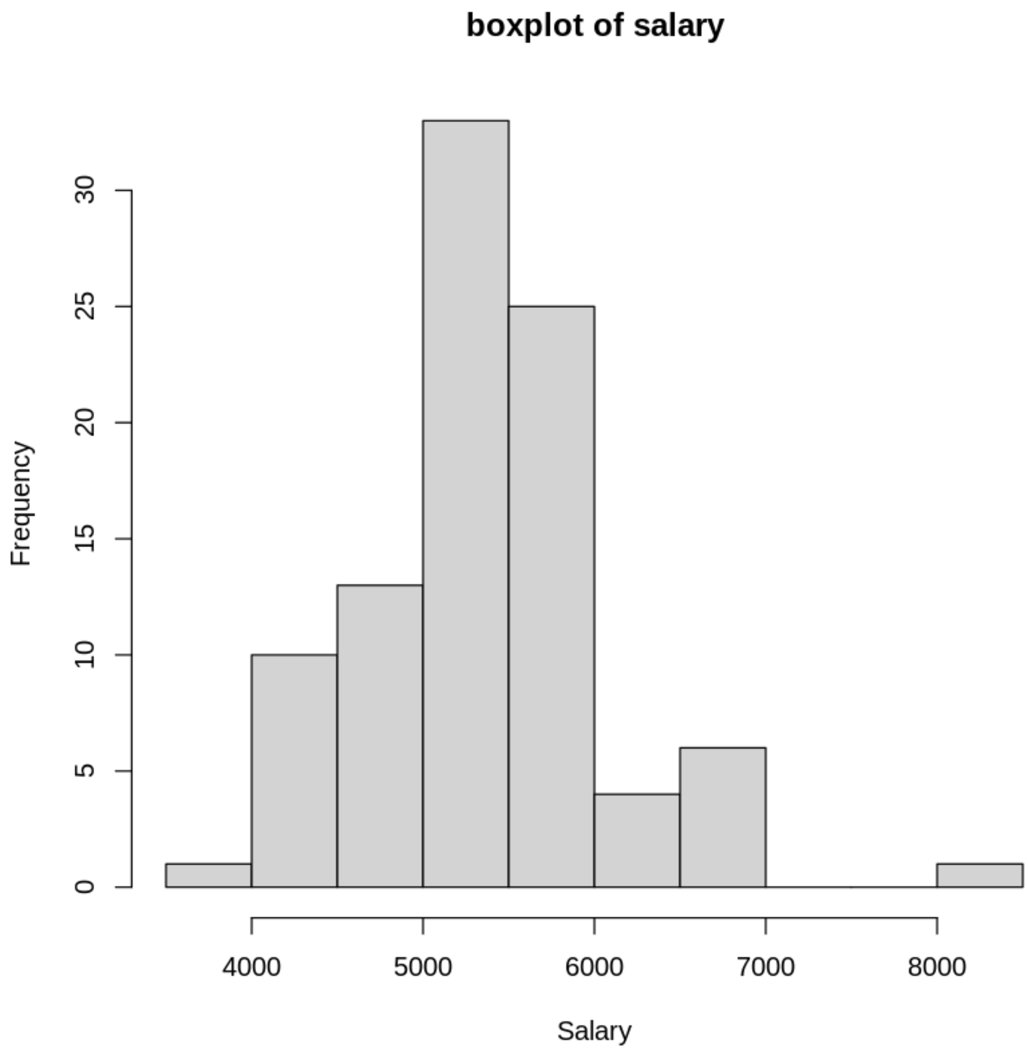
```
5 | 001111111122223334444444444444444
5 | 5566777777
6 | 000000000000001333
6 | 666899
7 |
7 |
8 | 1
```

[6]:

```r
box <- boxplot(case0102$Salary, main = "boxplot of salary")
outliers <- box$out
text(rep(1, length(outliers)), outliers, labels = outliers, pos = 4, col =
↪"red")
```



[7]:
```
hist(case0102$Salary, main = "boxplot of salary", breaks = 10, xlab = "Salary")
```



[8]:
```
case0102[which.max(case0102$Salary),]
```

A data.frame: 1 × 2

	Salary <int>	Sex <fct>
93	8100	Male

[9]:
```
IQR(case0102$Salary)
```

1020

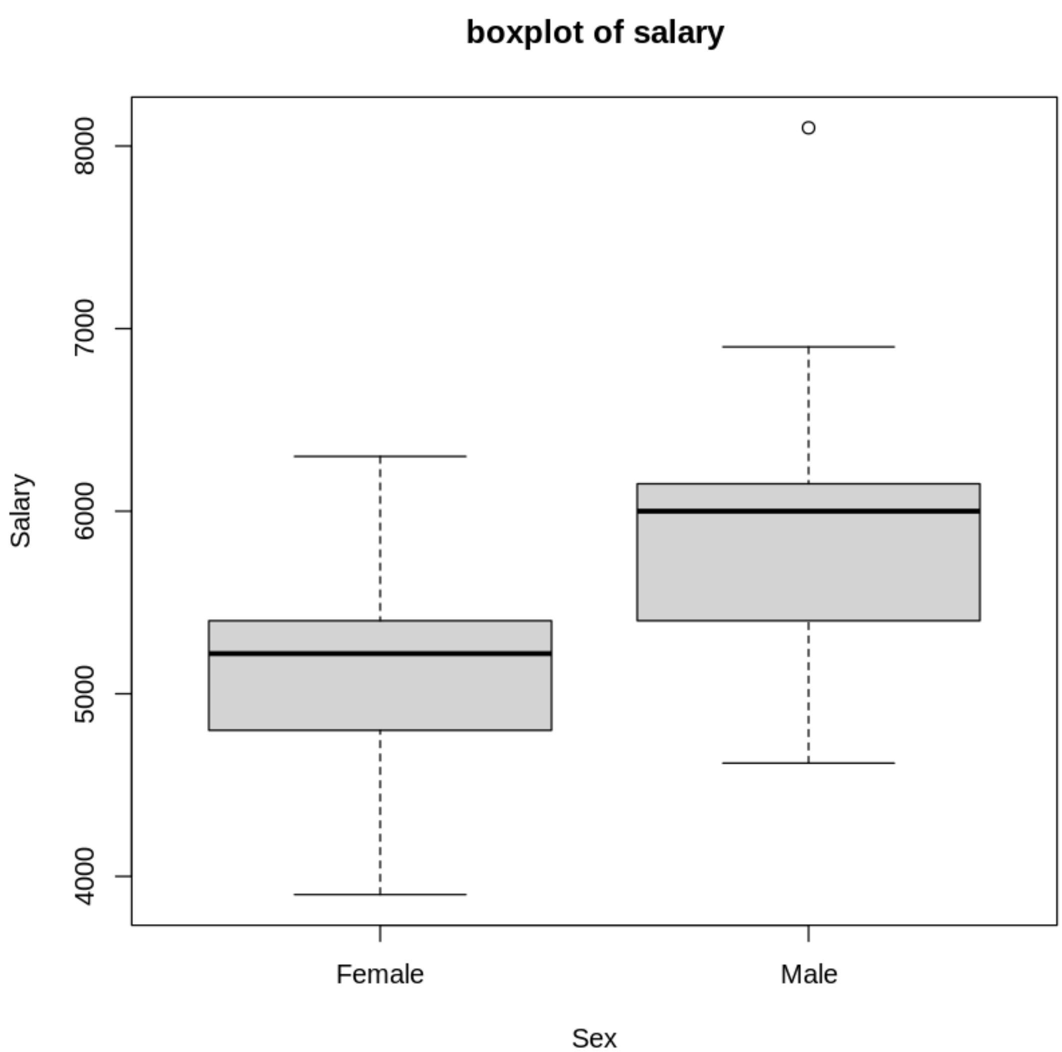
From above plots, we find that there is one outlier with salary of 8100 and sex as male. It can be proved since the IQR is 1020 and third quantile is 6000, 1.5*1020 + 6000 = 7530 < 8100 which means the point is not in the range.

ii)Perform separate EDA, and compute the sample coefficient of variation and median for Salary in each group (i.e., Males and Females).

[10]:
```
#make sure if there is no missing data
sum(is.na(case0102))
```

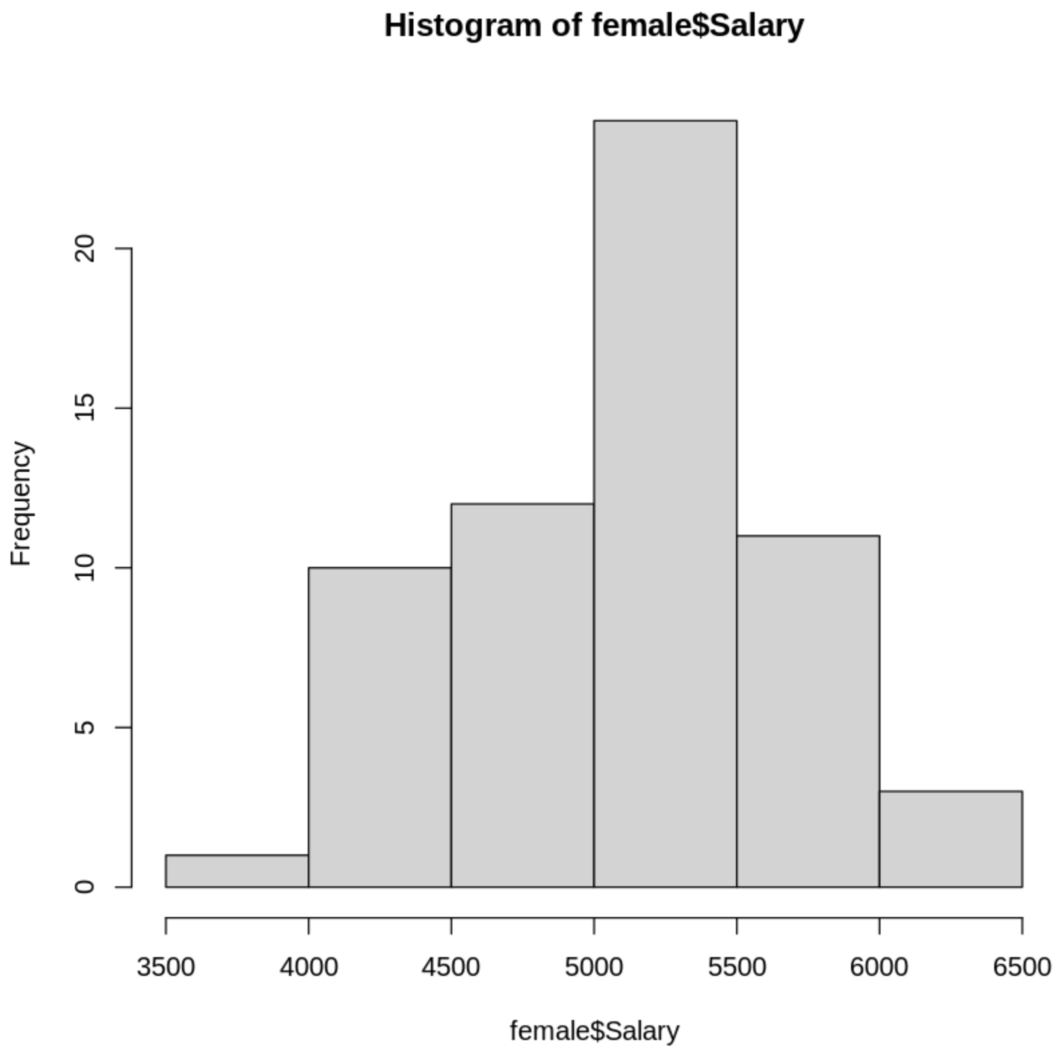
0

[11]:
```
female <- case0102[case0102$Sex == "Female",]
male <- case0102[case0102$Sex == "Male",]
boxplot(Salary~Sex, data = case0102, main = "boxplot of salary")
```



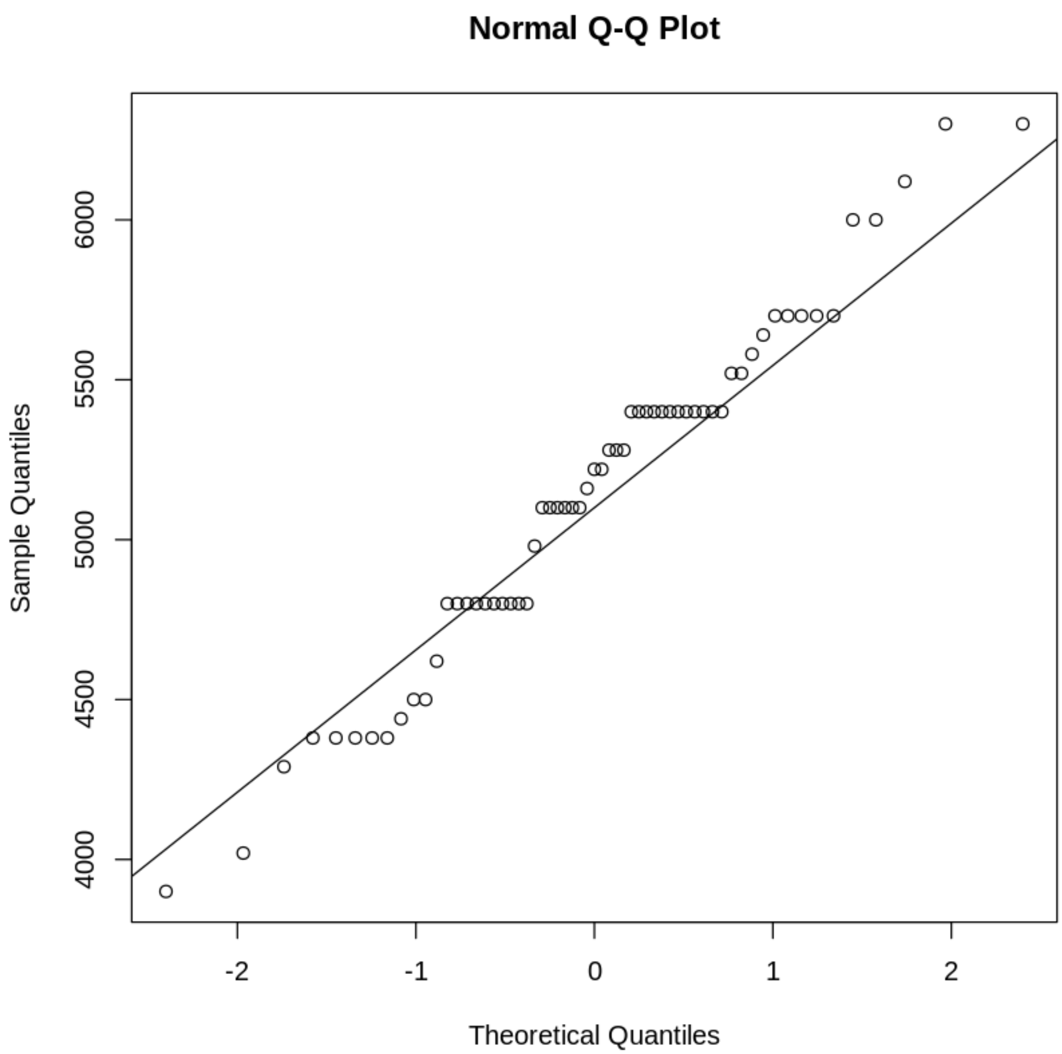
[12]:

```
hist(female$Salary)
```



[15]:

```
qqnorm(female$Salary)
qqline(female$Salary)
```



[16]:
```
stem(female$Salary)
```

```
The decimal point is 2 digit(s) to the right of the |

38 | 0
40 | 2
42 | 988888
44 | 400
46 | 2
48 | 00000000008
50 | 0000006
52 | 22888
```

```
54 | 000000000000228
56 | 400000
58 |
60 | 002
62 | 00
```

[17]:
```r
summary(female)
```

```
    Salary          Sex    
 Min.   :3900   Female:61  
 1st Qu.:4800   Male  : 0  
 Median :5220              
 Mean   :5139              
 3rd Qu.:5400              
 Max.   :6300              
```

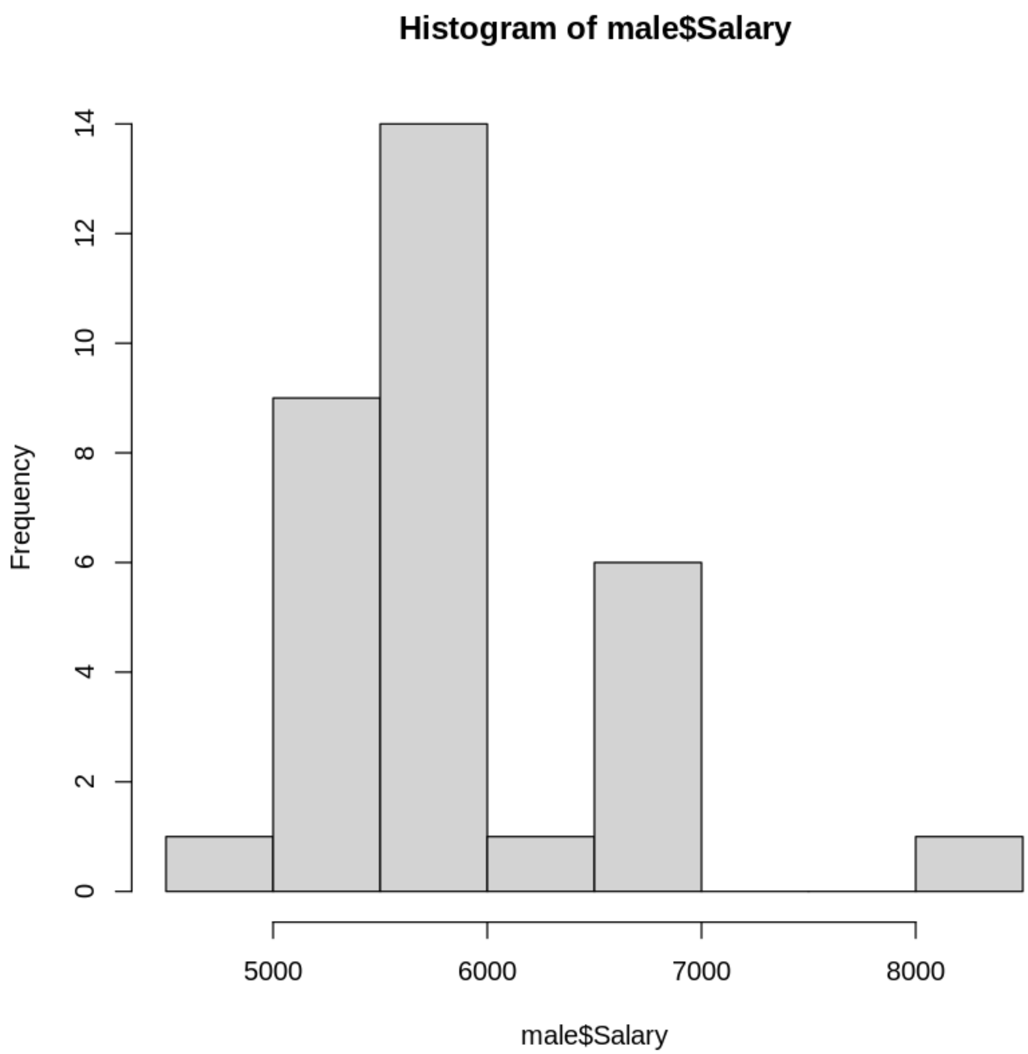
[19]:
```r
cv_female <- sd(female$Salary) / mean(female$Salary)
```

[20]:
```r
cv_female
```

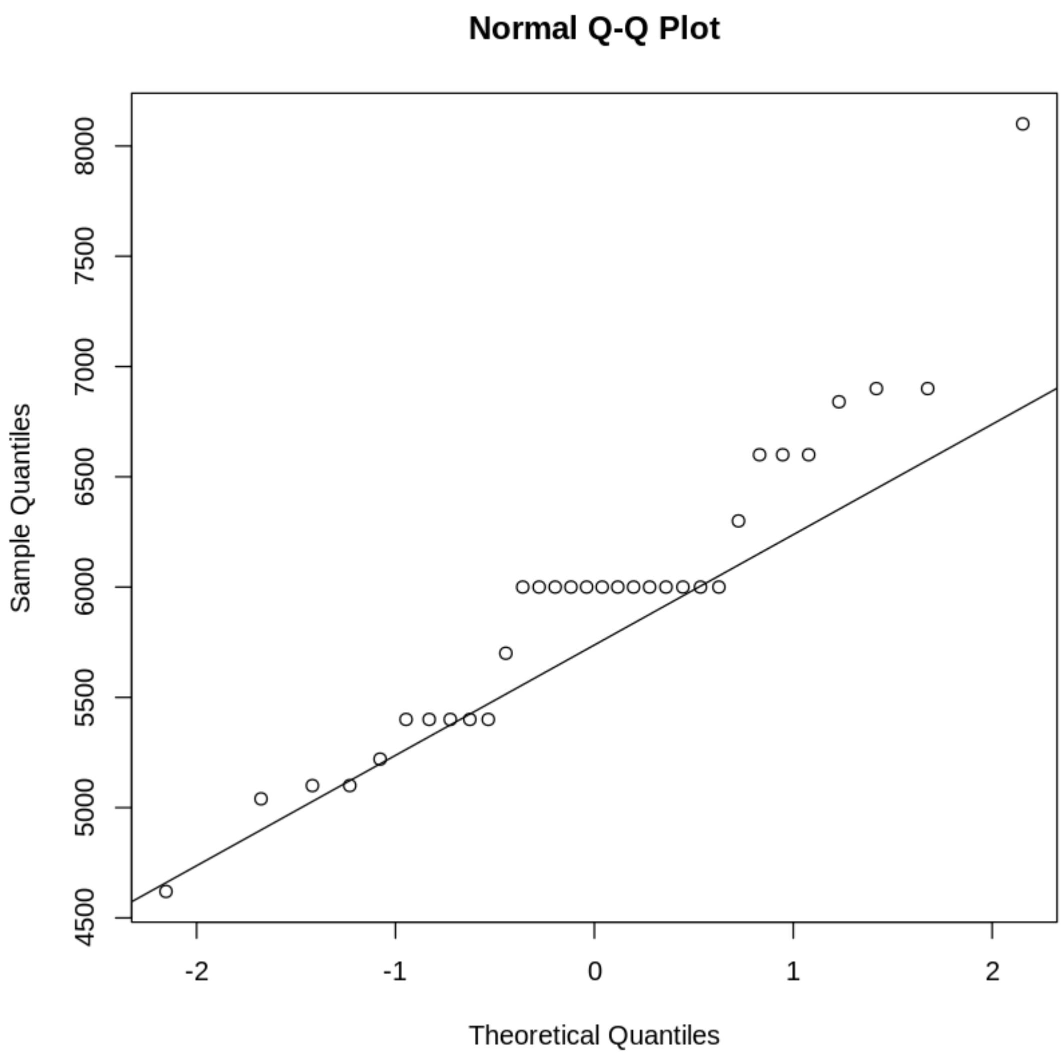
0.105056656669599

The sample coefficient of variation for Salary is 0.10506 and median for Salary is 5220 in female group.

[21]:
```r
hist(male$Salary)
```



[22]:
```
qqnorm(male$Salary)
qqline(male$Salary)
```



[23]:
```
stem(male$Salary)
```

```
The decimal point is 3 digit(s) to the right of the |

4 | 6
5 | 011244444
5 | 7
6 | 00000000000003
6 | 666899
7 |
7 |
8 | 1
```

[24]:
```
summary(male)
```

```
     Salary          Sex     
 Min.   :4620   Female: 0  
 1st Qu.:5400   Male  :32  
 Median :6000              
 Mean   :5957              
 3rd Qu.:6075              
 Max.   :8100              
```

[25]:
```
cv_male <- sd(male$Salary) / mean(male$Salary)
```

[26]:
```
cv_male
```

0.115955648890936

The sample coefficient of variation for Salary is 0.1160 and median for Salary is 6000 in female group.

iii)For each of the estimates computed in (ii) above, determine the bias and variance using each of the following methods: Jackknife, Bootstrap.

Jackknife for male group

[29]:
```
library(bootstrap)
jackknife(male$Salary,sd)$jack.bias
```

-11.280106757317

[30]:
```
jackknife(male$Salary,sd)$jack.se^2
```

15578.9863225273

The bias of SD under the jackknife method is -11.28011. The variance of estimator is 15578.99.

[31]:
```
jackknife(male$Salary,IQR)$jack.bias
```

1162.5

[32]:
```
jackknife(male$Salary,IQR)$jack.se^2
```

130781.25

The bias of IQR under the jackknife method is 1162.5. The variance of estimator is 130781.2.

Jackknife for female group

[33]:
```
jackknife(female$Salary,sd)$jack.bias
```

-1.94673751133223

[34]:
```
jackknife(female$Salary,sd)$jack.se^2
```

2101.90976108041

The bias of SD under the jackknife method is -1.946738. The variance of estimator is 2101.91.

[35]:
```
jackknife(female$Salary,IQR)$jack.bias
```

0

[36]:
```
jackknife(female$Salary,IQR)$jack.se^2
```

0

The bias of IQR under the jackknife method and the variance of estimator is zero.

Bootstrap for male group

[40]:
```
set.seed(0)
B <- 1000
n <- length(male$Salary)
est <- vector(length = B)
for (i in 1:B) {
sample.new <- sample(male$Salary, n, replace = T)
est[i] <- sd(sample.new)
}
b.sd.bias <- mean(est) - sd(male$Salary)
b.sd.bias
```

-20.8345519726006

[41]:
```
b.sd.var <- var(est)
b.sd.var
```

11643.3151848787

The bias of SD under the bootstrap method is -20.8346. The variance of estimator is 11643.32.

[42]:
```
set.seed(0)
B <- 1000
n <- length(male$Salary)
est <- vector(length = B)
for (i in 1:B) {
  sample.new <- sample(male$Salary, n, replace = T)
  est[i] <- IQR(sample.new)
}
b.IQR.bias <- mean(est) - IQR(male$Salary)
b.IQR.bias
```

37.155

[43]:
```
b.IQR.var <- var(est)
b.IQR.var
```

81888.1691441442

The bias of IQR under the bootstrap method is 37.155. The variance of estimator is (SE)^2 = 81888.17.

Bootstrap for female group

[44]:
```
set.seed(0)
B <- 1000
n <- length(female$Salary)
est <- vector(length = B)
for (i in 1:B) {
sample.new <- sample(female$Salary, n, replace = T)
est[i] <- sd(sample.new)
}
b.sd.bias <- mean(est) - sd(female$Salary)
b.sd.bias
```

-5.36415003305865

[45]:
```
b.sd.var <- var(est)
b.sd.var
```

2018.2772811078

The bias of SD under the bootstrap method is -5.3642. The variance of estimator is 2018.28.

[46]:
```
B <- 1000
n <- length(female$Salary)
est <- vector(length = B)
for (i in 1:B) {
sample.new <- sample(female$Salary, n, replace = T)
est[i] <- IQR(sample.new)
}
b.IQR.bias <- mean(est) - IQR(female$Salary)
b.IQR.bias
```

69.42

[47]:
```
b.IQR.var <- var(est)
b.IQR.var
```

13457.1207207207

The bias of IQR under the bootstrap method is 69.42. The variance of estimator is 13457.12..