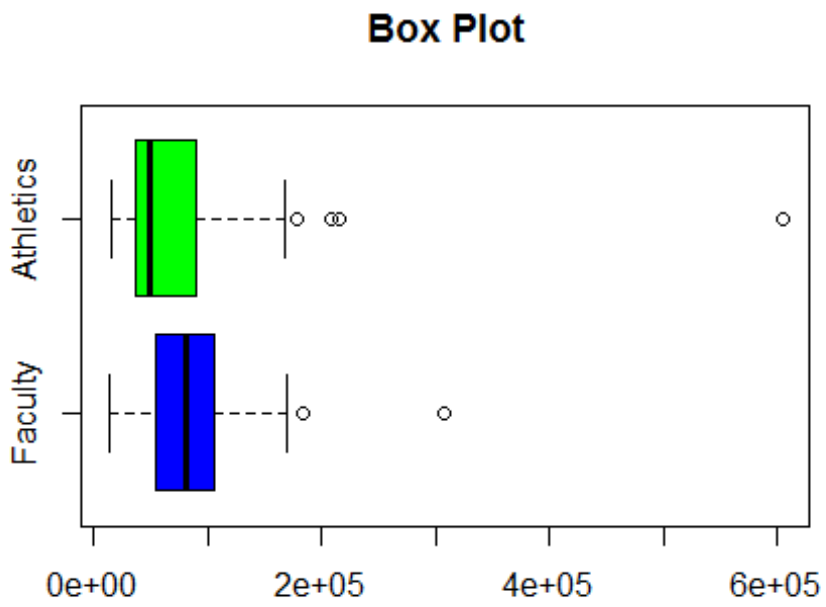
Take Home2 (S-520)

FNU Anirudh

November 9, 2015

Solution 1

```
salaries = read.table(file.choose(), header=TRUE)
Faculty = salaries$Salary[salaries$Job == "Faculty"]
Athletics = salaries$Salary[salaries$Job == "Athletics"]
boxplot(Faculty,Athletics,main="Box Plot",horizontal = TRUE,
        names=c("Faculty","Athletics"),col=c("blue","green"))
```



Median salary of Faculty is more than median salary of Athletics and there are more outliers in Athletics than Faculty, There is one huge outlier in athletics compared to faculty.

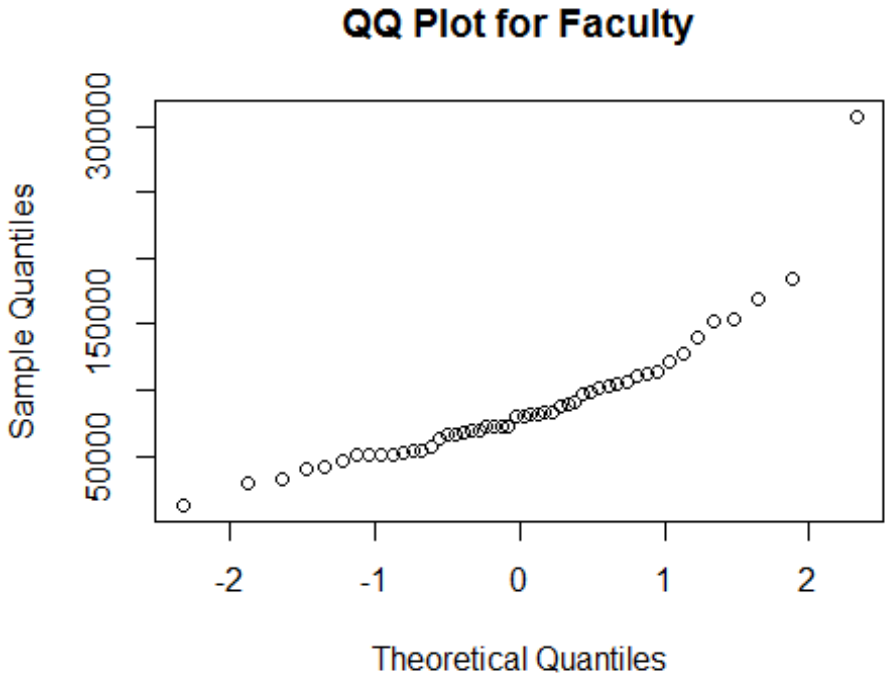
Solution 2

```
t.test(Faculty,Athletics)
```

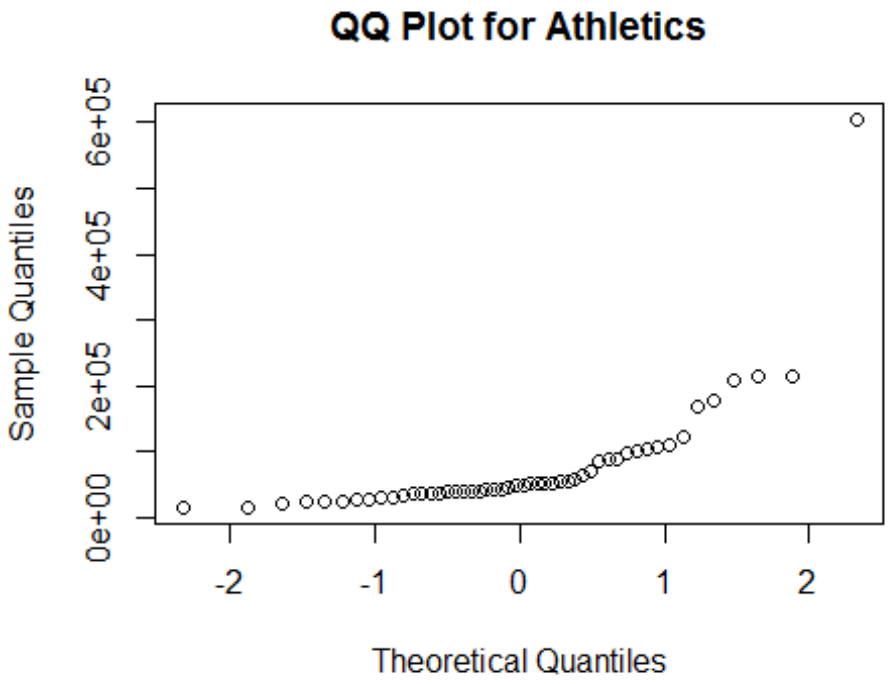
```
##
##  Welch Two Sample t-test
##
## data:  Faculty and Athletics
```

```
## t = 0.63883, df = 74.075, p-value = 0.5249
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -19826.38  38539.46
## sample estimates:
## mean of x mean of y
##  87608.22  78251.68
```

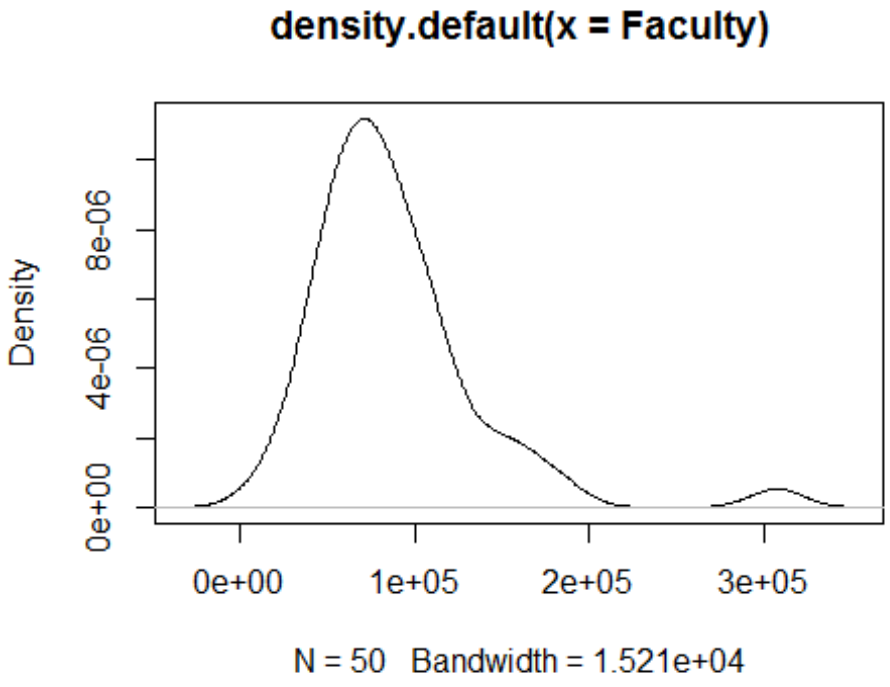
```
qqnorm(Faculty,main = "QQ Plot for Faculty")
```



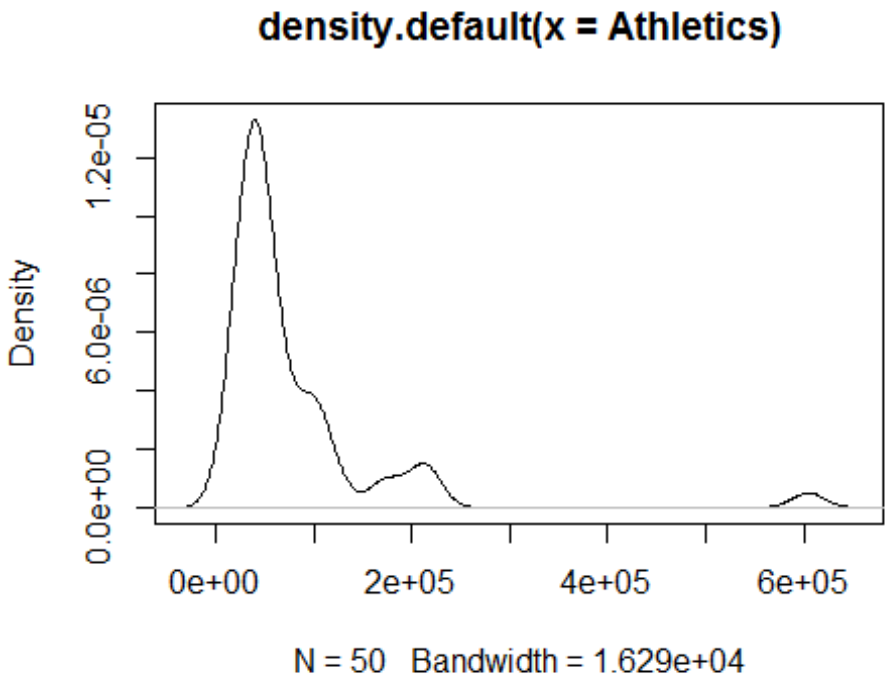

```
qqnorm(Athletics,main = "QQ Plot for Athletics")
```




```
plot(density(Faculty))
```




```
plot(density(Athletics))
```

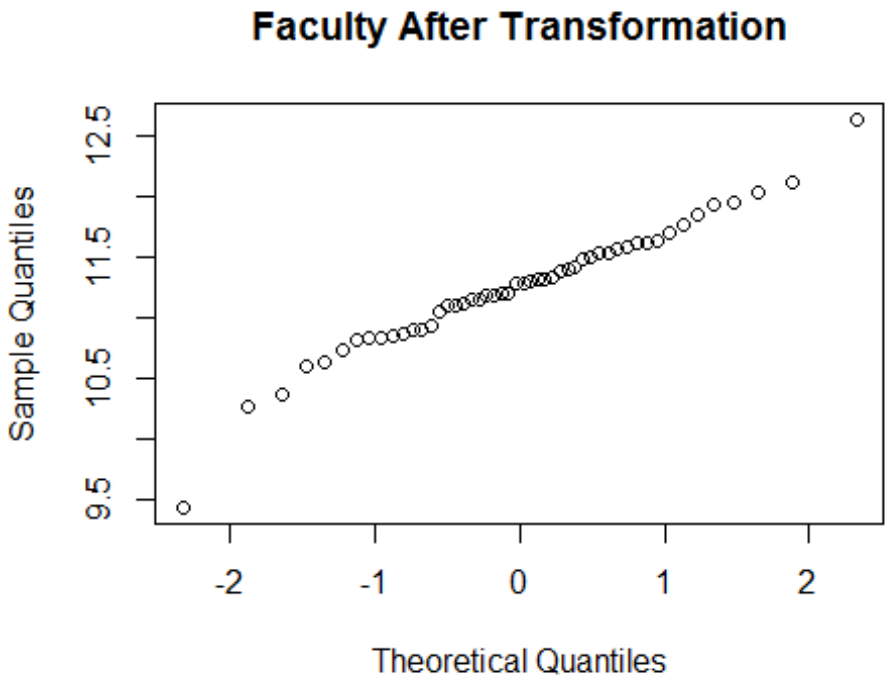


Welch's Two-Sample t-test is based on assumption that samples are normally distributed.We cannot trust p -value because we can see from QQ Plot and density plot that

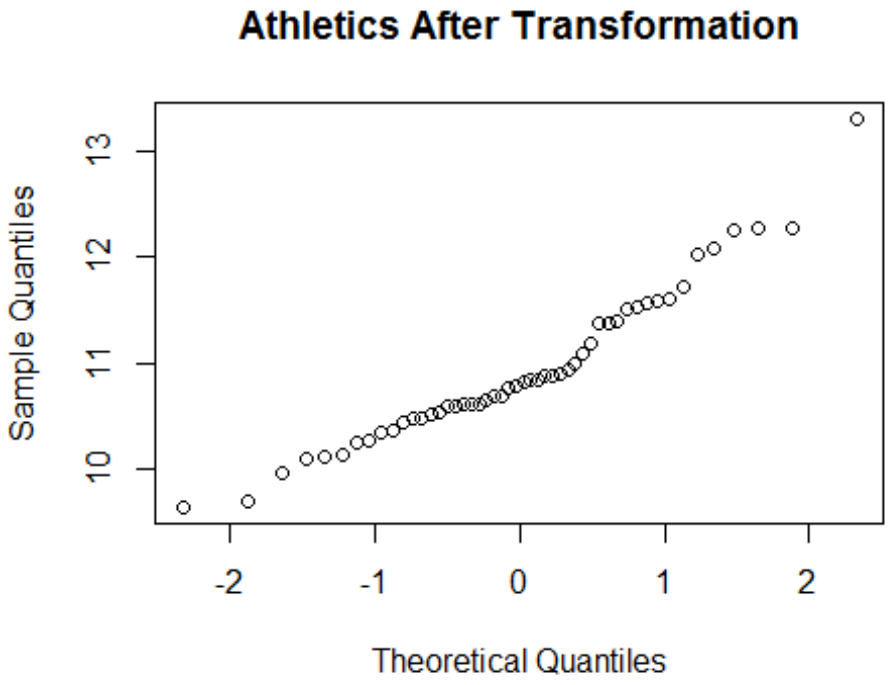
- both faculty and athletics are not normally distributed.
- sample size is relatively small
- There are huge outliers in both faculty and athletics (There is huge outlier in athletics which will effect mean to very large extent) and welch's t-test involves calculations with mean.

Solution 3

```
logfac=log(Faculty)
logath=log(Athletics)
qqnorm(logfac,main="Faculty After Transformation")
```




```
qqnorm(logath,main = "Athletics After Transformation")
```




```
var1=var(logfac)
var2=var(logath)
```

We should use Welch's two sample test since data is close to normal distribution after taking log and variances of the samples are not equal. Student's t-test could have been used if variance values were equal or closer hence Welch's two sample test seems plausible.

Solution 4

```
m1=mean(logfac)
m2=mean(logath)
Delta = mean(logfac) - mean(logath)
se = sqrt(var(logfac)/50 + var(logath)/50)
Tw = Delta/se
nu =
(var(logfac)/50+var(logath)/50)^2/((var(logfac)/50)^2/49+(var(logath)/50)^2/4
9)
Pvalue = 2*(1-pt(abs(Tw),df=nu))
Pvalue
```

```
## [1] 0.01648609
```

```
# Welch 95% confidence interval
q = qt(0.975, df=nu)
lower = Delta - q*se
lower
```

```
## [1] 0.05839702
```

```
upper = Delta + q*se
upper
```

```
## [1] 0.5657975
```

Since P-value is less than 0.05, we can reject our null hypothesis.

Solution 5

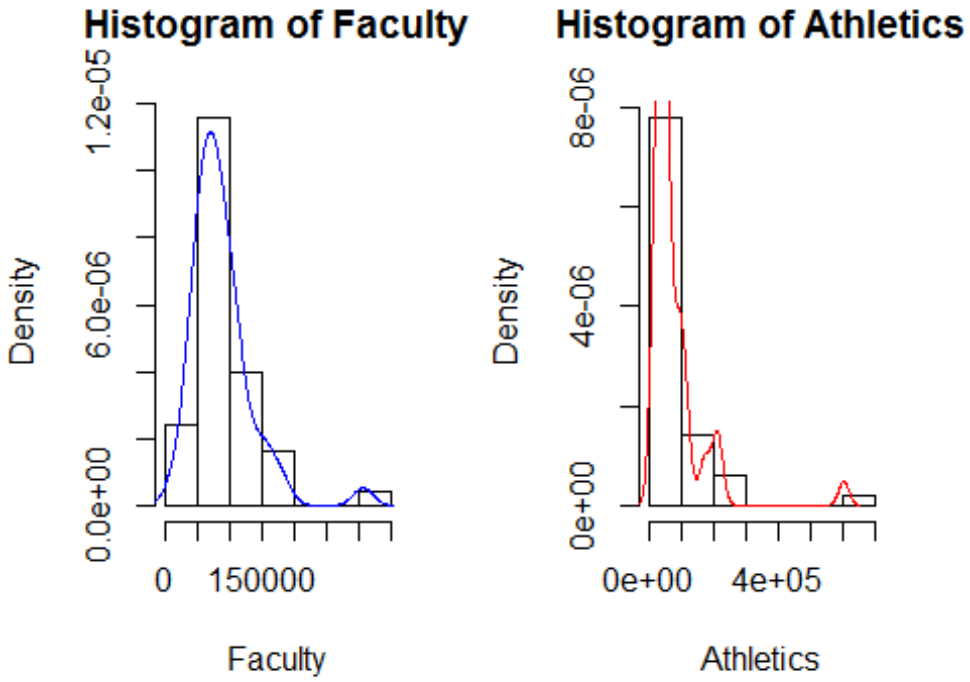
```
summary(Faculty)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   12560   55060   79910   87610  104600  307700
```

```
summary(Athletics)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   15270   35960   48900   78250   89000  604900
```

```
par(mfrow=c(1,2))
hist(Faculty,prob=TRUE)
lines(density(Faculty),col="blue")
hist(Athletics,prob=TRUE)
lines(density(Athletics),col="red")
```



We can see that Mean and median for Faculty is more than Athletics and there is huge outlier in Athletics (double of outlier in faculty) as seen in summary. From Histogram we can say that faculty and athletics don't have same distribution.