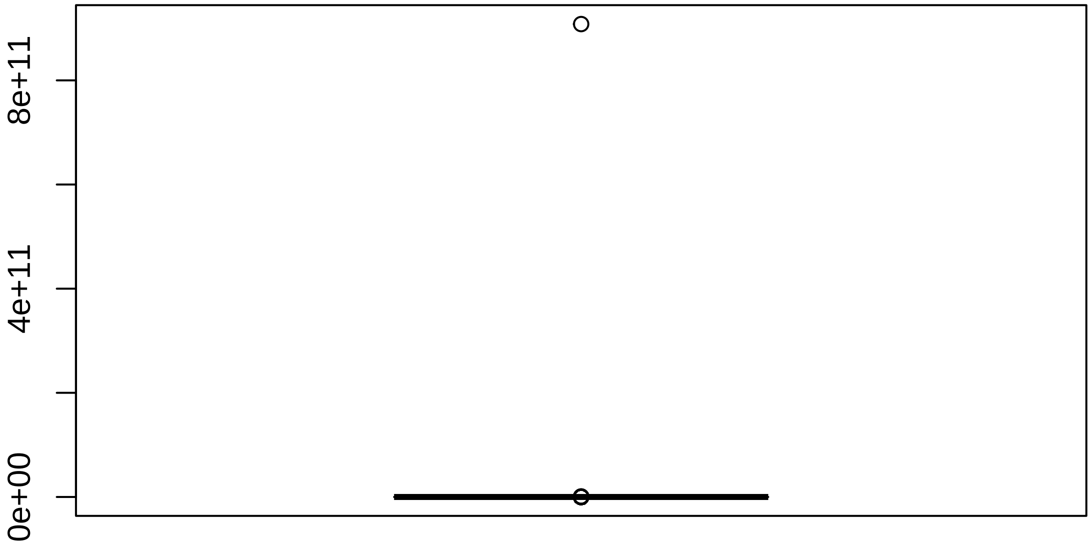
1. Calculate a 95 percent confidence interval for the "Gross output – Year 3 (Rs)"

Let's visualize the box plot for "Gross output – Year 3 (Rs)"

```
boxplot(Group_24$GOP_Year3)
```

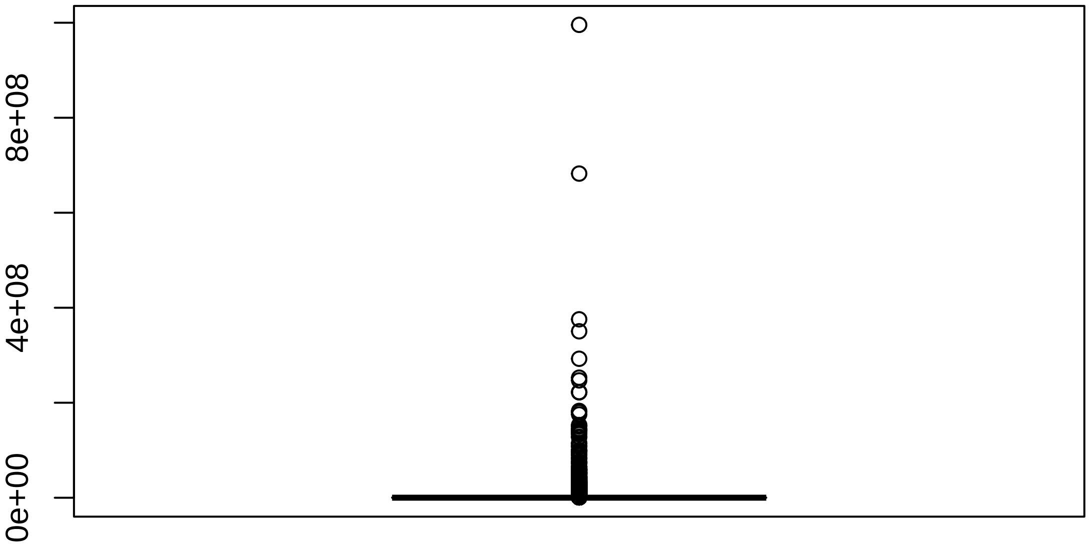


Clearly there is 1 outlier and we'll deal with it by removing it from our calculations:

```
GOP_Year3_OutliersRemoved = Group_24$GOP_Year3[!Group_24$GOP_Year3 %in% max(Group_24$GOP_Year3)]
Group_24 = Group_24[!Group_24$GOP_Year3 == max(Group_24$GOP_Year3),]
```

Let's visualize the box plot for "Gross output – Year 3 (Rs)"

```
boxplot(GOP_Year3_OutliersRemoved)
```



T distribution would be used as we are using almost the sample set of 10000 samples and Standard deviation σ of population is not given.

To find the confidence interval with t-distribution we use the following formula:

$$CI = \bar{x} \pm t_{\alpha/2} \frac{S}{\sqrt{n}}$$

Total number of data in sample (n) = 9999

degree of freedom = n-1 = 9998

Sample mean of GOP_Year3$(\bar{x})$ =

```
mean(GOP_Year3_OutliersRemoved)
```

```
## [1] 1965767
```

Standard deviation (S) of GOP_Year3 =

```
sd(GOP_Year3_OutliersRemoved)
```

```
## [1] 16712686
```

Given that 95 % confidence hence $\alpha = 0.05$ and n = 9999, $t_{\alpha/2}$ value from t table is 1.9602

The confidence interval for the “Gross output – Year 3(Rs)” can be *manually* calculated by the formula mentioned above.

A function called *t.test* provided in R can be used to compute the same, where we pass in the samples and confidence level.

```r
t.test(GOP_Year3_OutliersRemoved, conf.level = 0.95)
```

```
##
##  One Sample t-test
##
## data:  GOP_Year3_OutliersRemoved
## t = 11.762, df = 9998, p-value < 2.2e-16
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  1638148 2293386
## sample estimates:
## mean of x
##   1965767
```

With 95% confidence the population mean is between 1638148 and 2293386, based on 9999 samples

---

2. Define two different measures that you consider most appropriate for measuring the performance of the units. This definition is up to you. These can be the variables that are already in the data or new variables defined based on the existing variables. For example, you can define a metric “Gross output per employee” by dividing the gross output of the year by the total number of employees. Please explain in one paragraph why you have selected these two measures and why you think they are most appropriate. Remaining analysis is to be carried out based on these definitions.

The two measures that we’ll use for computations are as follows:

1. Gross output of Year3 per employee(GOP_Year3_PerEmp) = GOP_Year3 / EMP_TOTAL

To measure the performance of small-scale units, in general, could consider some metrics associated with financial performance, profitability, and customer retention. In this case, as we do not have metrics associated with customer retention, we ignore that the aspect of customer and would like to consider the overall output in terms of a productivity gain achieved by the total output generated by an employee as this can directly affect the financial performance and profitability.

```r
Group_24[['GOP_Year3_PerEmp']] = Group_24[['GOP_Year3']]/Group_24[['EMP_TOTAL']]
```

2. Fixed Asset Gain Amount(Fixed_Asset_Gain) = MKT_VAL_FA - ORI_PURC_VAL_PM

In similar terms, fixed asset cost growth can help us to determine the value of the asset over a period of time. some assets value can increase while some can depreciate, both have a direct impact on the performance metrics. Accordingly, we would like to consider the below two measures to determine the performance of unitsGross output per employeeFixed Asset cost growth

```
Group_24[['Fixed_Asset_Gain']] = Group_24[['MKT_VAL_FA']] - Group_24[['ORI_PURC_VAL_PM']]
```

---

3. Calculate 99% confidence interval for the population mean for each of the two metrics defined by you. Interpret these confidence intervals in terms of their relevance to the management.

To find the confidence interval with t-distribution we use the following formula:

$$CI = \bar{x} \pm t_{\alpha/2}\frac{S}{\sqrt{n}}$$

1. **Gross output of Year3 per employee(GOP_Year3_PerEmp)**

Total number of data in sample $(n) = 9999$

degree of freedom = n-1 = 9998

Sample mean of GOP_Year3_PerEmp $(\bar{x}) =$

```
mean(Group_24$GOP_Year3_PerEmp)
```

```
## [1] 158610.4
```

Standard deviation (S) of GOP_Year3_PerEmp =

```
sd(Group_24$GOP_Year3_PerEmp)
```

```
## [1] 1094651
```

Given that 99 % confidence, hence $\alpha = 0.01$ and n = 9999, $t_{\alpha/2}$ value from t table is 2.576

The confidence interval for the "Gross output of Year3 per employee" can be *manually* calculated by the formula mentioned above.

A function called *t.test* provided in R can be used to compute the same, where we pass in the samples and confidence level.

```
t.test(Group_24$GOP_Year3_PerEmp, conf.level = 0.99)
```

```
##
##  One Sample t-test
##
## data:  Group_24$GOP_Year3_PerEmp
## t = 14.489, df = 9998, p-value < 2.2e-16
## alternative hypothesis: true mean is not equal to 0
## 99 percent confidence interval:
##  130407.3 186813.6
## sample estimates:
## mean of x
##  158610.4
```

With 99% confidence the population mean is between 130407.3 and 186813.6, based on 9999 samples

1. **Fixed Asset Gain Amount(Fixed_Asset_Gain)**

Total number of data in sample (n) = 9999

degree of freedom = n-1 = 9998

Sample mean of Fixed_Asset_Gain $(\bar{x})$ =

```r
mean(Group_24$Fixed_Asset_Gain)
```

```
## [1] 876748.5
```

Standard deviation (S) of Fixed_Asset_Gain =

```r
sd(Group_24$Fixed_Asset_Gain)
```

```
## [1] 32584436
```

Given that 99 % confidence, hence $\alpha = 0.01$ and n = 9999, $t_{\alpha/2}$ value from t table is 2.576

The confidence interval for the "Fixed Asset Gain Amount" can be *manually* calculated by the formula mentioned above.

A function called *t.test* provided in R can be used to compute the same, where we pass in the samples and confidence level.

```r
t.test(Group_24$Fixed_Asset_Gain, conf.level = 0.99)
```

```
##
##  One Sample t-test
##
## data:  Group_24$Fixed_Asset_Gain
## t = 2.6906, df = 9998, p-value = 0.007145
## alternative hypothesis: true mean is not equal to 0
## 99 percent confidence interval:
##    37226.84 1716270.24
## sample estimates:
## mean of x
##  876748.5
```

With 99% confidence the population mean is between 37226.84 and 1716270.24, based on 9999 samples

---

4(a). What is the probability that a firm selected at random is a SSSBE unit?

The column to consider is UNIT_TYPE which contains values 1 and 2, which signifies SSI and SSSBE unit types respectively.

No. of SSSBE =

```r
nrow(Group_24[Group_24$UNIT_TYPE == 2,])
```

```
## [1] 2125
```

Probability firm selected is SSSBE unit =

```r
prob_sssbe = nrow(Group_24[Group_24$UNIT_TYPE == 2,])/nrow(Group_24)
prob_sssbe
```

```
## [1] 0.2125213
```

4(b). What is the probability that a firm selected at random is GOOD in performance? (Calculate the average of the first performance measure that you had defined in question 2 above. If the firm's performance is above this average, it considered good. If it is below average, it is considered Bad)

The column considered is GOP_Year3_PerEMP. Mean of GOP_Year3_PerEMP =

```r
mean_GOP_Year3_PerEmp = mean(Group_24$GOP_Year3_PerEmp)
mean_GOP_Year3_PerEmp
```

```
## [1] 158610.4
```

Good Performance firms are considered when GOP_Year3_PerEMP is greater than or equal to mean_GOP_Year3_PerEmp

Probability of Good Performance Firms =

```r
prob_good = nrow(Group_24[Group_24$GOP_Year3_PerEmp >= mean_GOP_Year3_PerEmp,])/nrow(Group_24)
prob_good
```

```
## [1] 0.1406141
```

4(c). What is the probability that a firm selected is a SSSBE Unit and ALSO GOOD in performance?

Let's create a column for GOOD and BAD performance firms.

```r
Group_24$goodPerformance <- with(Group_24, ifelse(Group_24$GOP_Year3_PerEmp >=
                                                    mean_GOP_Year3_PerEmp,
                                                  'GOOD', 'BAD'))
```

Let's draw a contingency table for the two variables, Unit type and GoodPerformance firms:

```r
table(Group_24$UNIT_TYPE, Group_24$goodPerformance)
```

```
##
##      BAD GOOD
##   1 6539 1335
##   2 2054   71
```

71 out of 10000 firms are both SSSBE and Good performing.

So probability that a firm selected is a SSSBE Unit and ALSO GOOD in performance:

```r
71/9999
```

```
## [1] 0.00710071
```

4(b)What can you say about the performance of the SSSBE units in terms of GOOD or BAD based on the probabilities calculated above?

Let's have a look at the contingency table once again:

```r
table(Group_24$UNIT_TYPE, Group_24$goodPerformance)
```

```
##
##       BAD GOOD
##   1 6539 1335
##   2 2054   71
```

The second Row with UNIT_TYPE = 2 are SSSBE units. *71* out of 2054 + 71 = *2125* units are SSSBE units, and out of those just 71 are categorized as performing good.

Percentage of those GOOD performing SSSBE units is:

```r
label_percent(accuracy = 0.001)(71/2125)
```

```
## [1] "3.341%"
```

Only 3.34% of the SSSBE units are GOOD in performance, remaining 99.66% of SSSBE units are BAD in performance

---

5. Test the null hypothesis that the population average of the variable "Value of Exports for Year 3" = 87,300. Carry out a one sided test. Clearly state your null and alternate hypotheses.

H_0:"Value of Exports for Year 3" <= 87,300

H_1: "Value of Exports for Year 3" > 87,300

```r
t.test(Group_24$VOE_Year3 ,mu = 87300, alternative = "greater")
```

```
##
##  One Sample t-test
##
## data:  Group_24$VOE_Year3
## t = -22.281, df = 9998, p-value = 1
## alternative hypothesis: true mean is greater than 87300
## 95 percent confidence interval:
##  4727.581      Inf
## sample estimates:
## mean of x
##  10404.65
```

If the p-value is less than (or equal to) α, then the null hypothesis is rejected in favor of the alternative hypothesis. And, if the p-value is greater than α, then the null hypothesis is not rejected.

In our case, p-value is 1 which is greater than α which is 0.05, hence we do not reject the null hypothesis.

---

6. There is a feeling within the Central Government Department for promotion of small scale units, be it SSSBE or SSI, that if the population proportion is less than 25%, there is a need for providing special incentives. Based on your sample, would you recommend these special incentives for SSSBE or SSI or both?

Null Hypothesis H_0: population proportion of SSSBE or SSI >= 25%

Alternative Hypothesis H_1: population proportion of SSSBE or SSI < 25%

Now we calculate the test statistic using the following formula:

$$z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$$

```
z_SSSBE = (((nrow(Group_24[Group_24$UNIT_TYPE == 2,]))/nrow(Group_24))-0.25) /
  sqrt(0.25*(1-0.25)/nrow(Group_24))
z_SSSBE
```

```
## [1] -8.654913
```

```
z_SSI = (((nrow(Group_24[Group_24$UNIT_TYPE == 1,]))/nrow(Group_24))-0.25) /
  sqrt(0.25*(1-0.25)/nrow(Group_24))
z_SSI
```

```
## [1] 124.1192
```

Let the level of significance be 5%, therefore Lower tail of Z(0.05) = -1.65

For SSI units: z_SSI > -Z(0.05)" therefore we fail to reject H_0. There is no significant reason to provide special incentives for the SSI units.

For SSSBE units; z_SSSBE < -Z(0.05)" therefore we reject H_0. There is a significant reason to provide special incentives for the SSSBE units.

---

7. Some male chauvinists like to think that a larger proportion of SSSBEs are managed by men as compared to women. Do you agree with this contention? Explain your answer with appropriate statistical evidence.

Given that the SSSBEs are managed by either men or women, we can state the following:

Null Hypothesis H_0: population proportion of SSSBEs managed by men <= 50%

Alternative Hypothesis H_1: population proportion of SSSBEs managed by men > 50%

Now we calculate the test statistic using the following formula:

$$z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$$

```
z_SSSBE_men = (((nrow(Group_24[Group_24$MAN_BY == 1,]))/nrow(Group_24))-0.50) /
  (sqrt(0.50*(1-0.50)/(nrow(Group_24[Group_24$UNIT_TYPE == 2,]))))

z_SSSBE_men
```

```
## [1] 43.41457
```

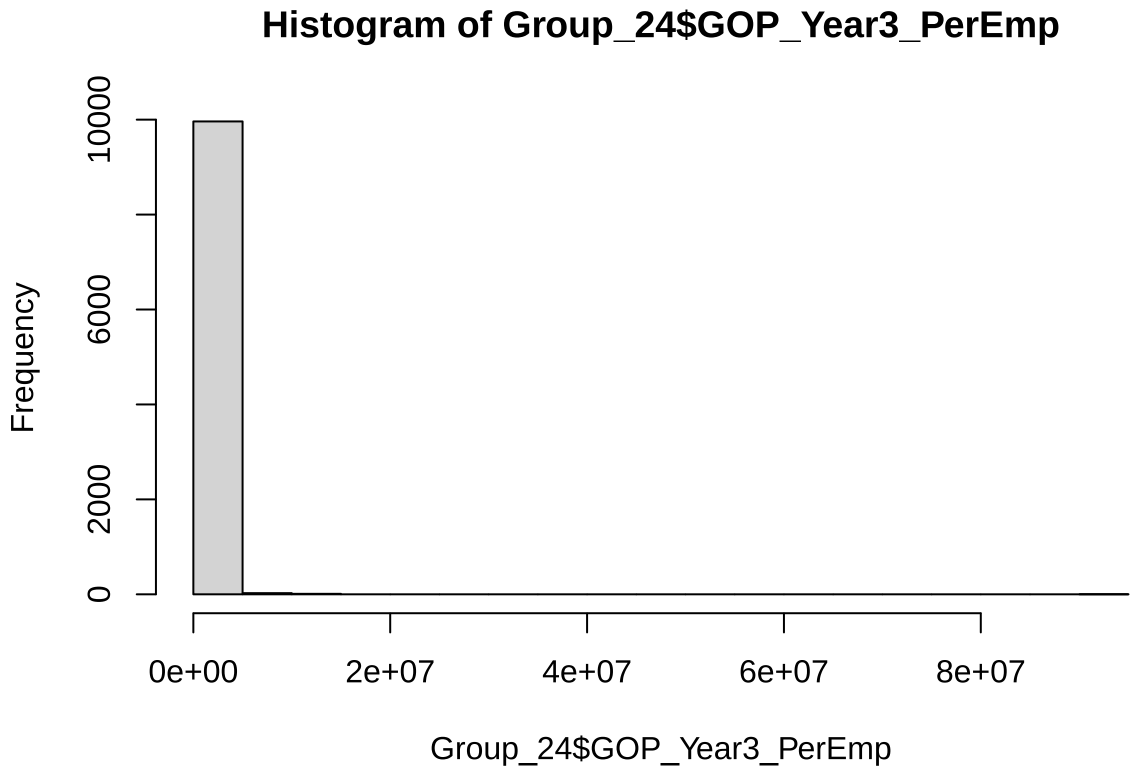
Let the level of significance be 5%, therefore upper tail of Z(0.05) = 1.65

For the SSSBE units; "$z\_SSSBE\_men >$ Z(0.05)" therefore we reject H_0. There is a significant reason to believe that "Larger proportion of SSSBEs are managed by men as compared to women".

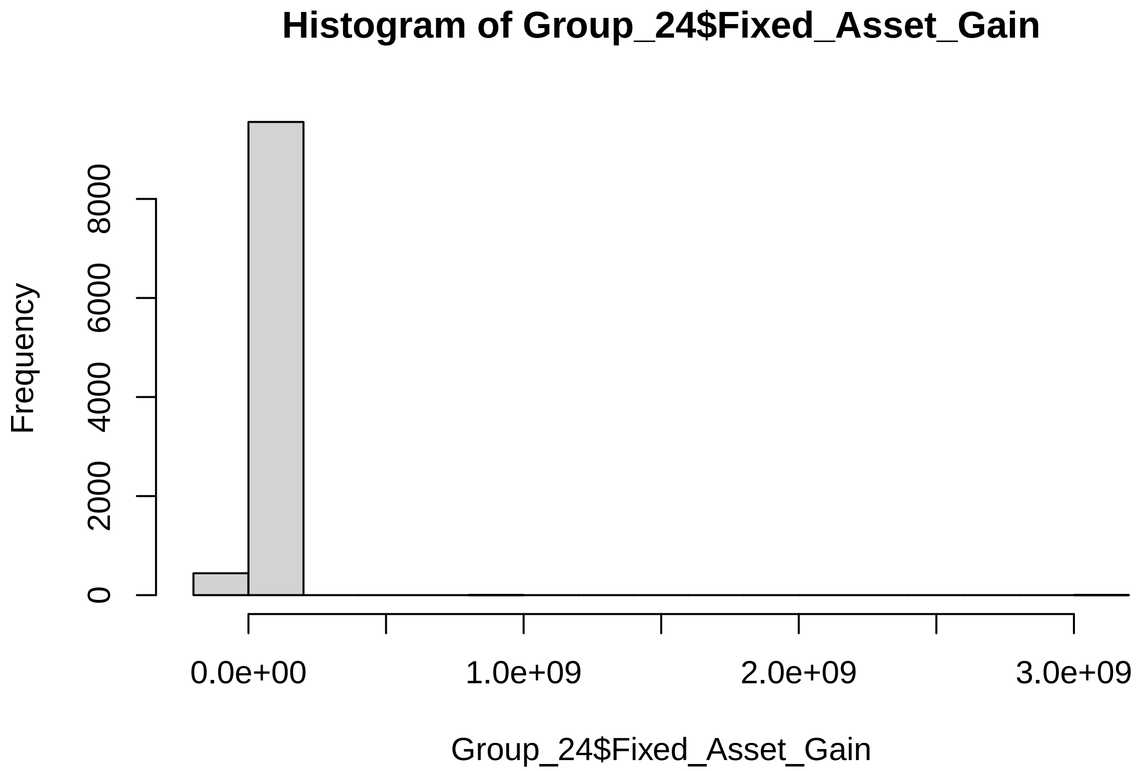
---

8. Comment on the distribution of the two metrics (variables) created by you. Can you conclude that they follow normal distribution? Explain the rationale for your conclusion.

```
hist(Group_24$GOP_Year3_PerEmp)
```



```
hist(Group_24$Fixed_Asset_Gain)
```



The histograms represent the shape of the distribution of the measures:

1. *Gross output of Year3 per employee (GOP_Year3_PerEmp)*
2. *Fixed Asset Gain Amount (Fixed_Asset_Gain)*

Clearly we see that the variables do not follow the Normal Distribution. The distributions are highly right or positively skewed.