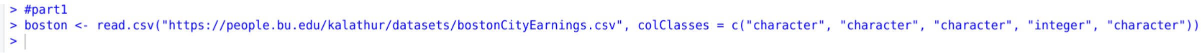
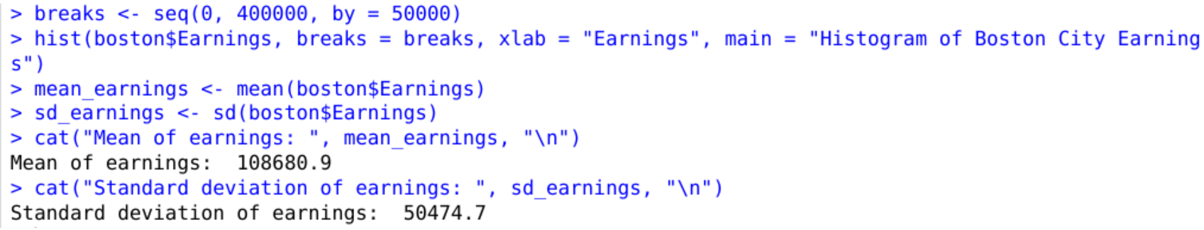
Part1) Central Limit Theorem (30 points) Initialize the city of Boston earnings dataset as shown below: boston <- read.csv("https://people.bu.edu/kalathur/datasets/bostonCityEarnings.csv", colClasses = c("character", "character", "character", "integer", "character")) The data in the file contains the total earnings of the employees of city of Boston.

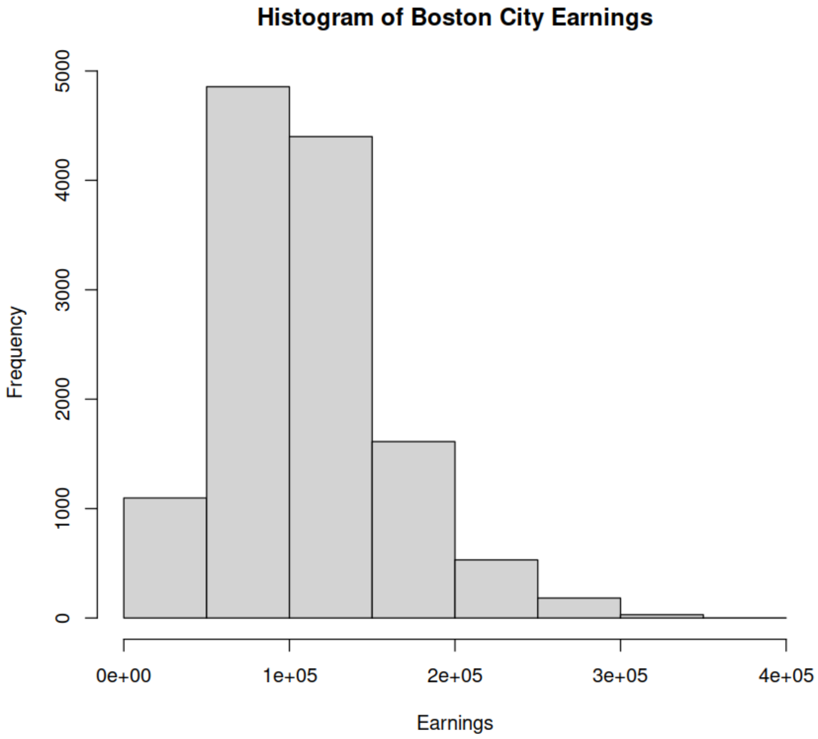


```
> #part1
> boston <- read.csv("https://people.bu.edu/kalathur/datasets/bostonCityEarnings.csv", colClasses = c("character", "character", "character", "integer", "character"))
> 
```

a) Show the histogram of the employee earnings. Use breaks from 0 to 400000 in steps of 50000 and show the corresponding tick labels on the x-axis. Compute the mean and standard deviation of this data. What do you infer from the shape of the histogram?



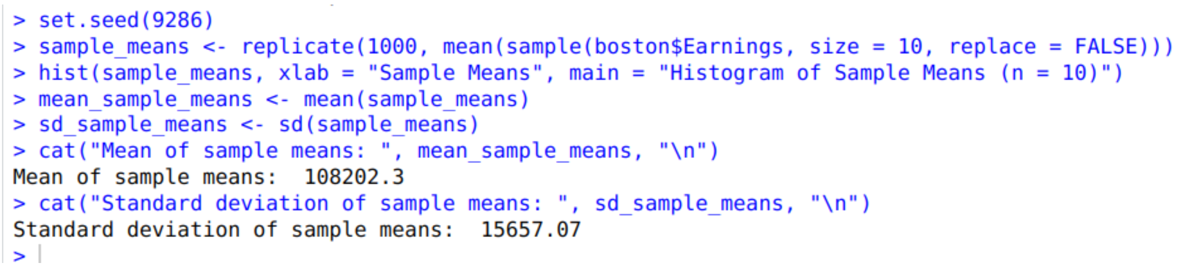
```
> breaks <- seq(0, 400000, by = 50000)
> hist(boston$Earnings, breaks = breaks, xlab = "Earnings", main = "Histogram of Boston City Earning
s")
> mean_earnings <- mean(boston$Earnings)
> sd_earnings <- sd(boston$Earnings)
> cat("Mean of earnings: ", mean_earnings, "\n")
Mean of earnings:  108680.9
> cat("Standard deviation of earnings: ", sd_earnings, "\n")
Standard deviation of earnings:  50474.7
```



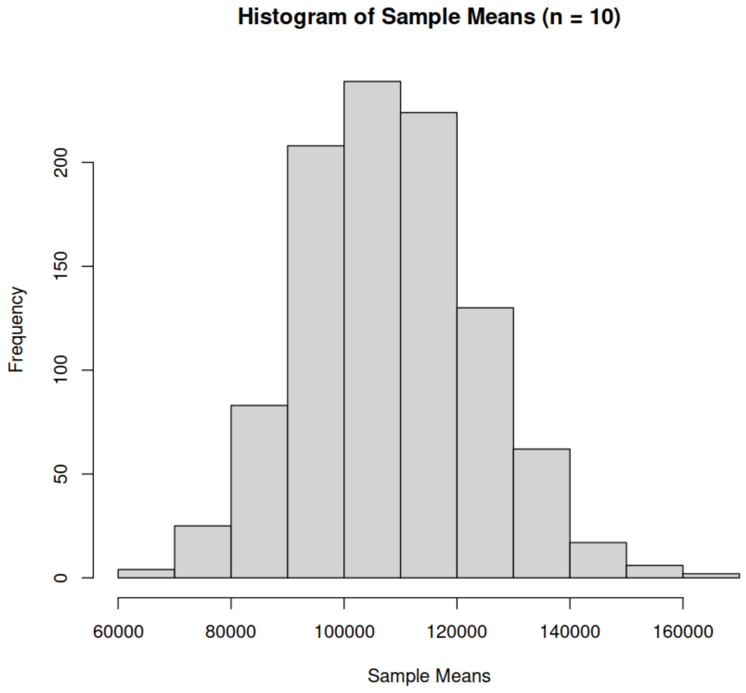
The shape of the histogram indicates that there may be some high-earning outliers that are pulling the mean towards the right.

b) Draw 1000 samples of this data of size 10, show the histogram of the sample means. Compute the mean of the sample means and the

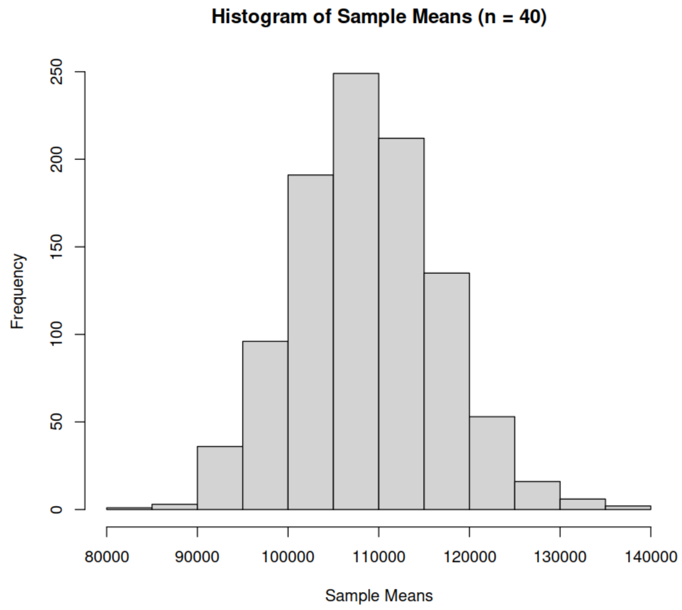
standard deviation of the sample means. Use sample() function with replace as FALSE for drawing the samples. Set the start seed for random numbers as the last 4 digits of your BU id.



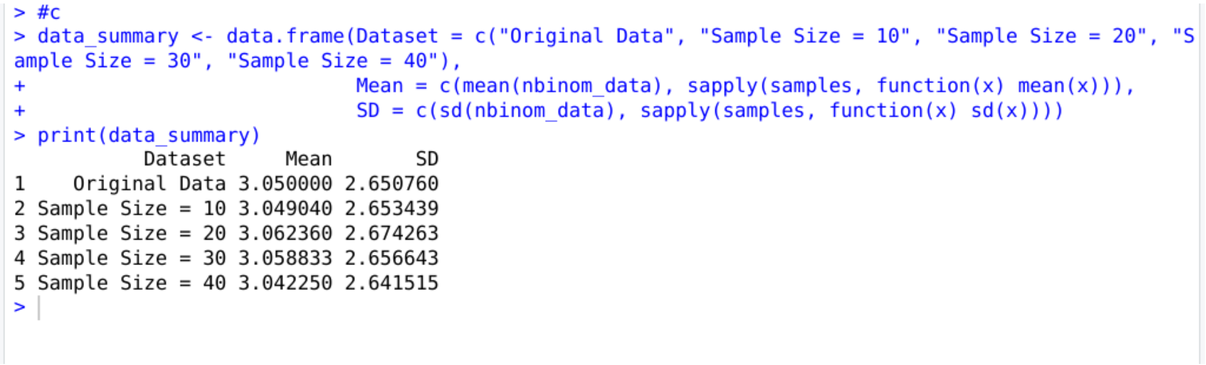
```
> set.seed(9286)
> sample_means <- replicate(1000, mean(sample(boston$Earnings, size = 10, replace = FALSE)))
> hist(sample_means, xlab = "Sample Means", main = "Histogram of Sample Means (n = 10)")
> mean_sample_means <- mean(sample_means)
> sd_sample_means <- sd(sample_means)
> cat("Mean of sample means: ", mean_sample_means, "\n")
Mean of sample means:  108202.3
> cat("Standard deviation of sample means: ", sd_sample_means, "\n")
Standard deviation of sample means:  15657.07
> 
```



c) Draw 1000 samples of this data of size 40, show the histogram of the sample means. Compute the mean of the sample means and the standard deviation of the sample means. Use sample() function with replace as FALSE for drawing the samples. Set the start seed for random numbers as the last 4 digits of your BU id.



d) Compare of means and standard deviations of the above three distributions.



```
> #c
> data_summary <- data.frame(Dataset = c("Original Data", "Sample Size = 10", "Sample Size = 20", "S
ample Size = 30", "Sample Size = 40"),
+                              Mean = c(mean(nbinom_data), sapply(samples, function(x) mean(x))),
+                              SD = c(sd(nbinom_data), sapply(samples, function(x) sd(x))))
> print(data_summary)
           Dataset     Mean       SD
1    Original Data 3.050000 2.650760
2 Sample Size = 10 3.049040 2.653439
3 Sample Size = 20 3.062360 2.674263
4 Sample Size = 30 3.058833 2.656643
5 Sample Size = 40 3.042250 2.641515
> 
```

Compared of means and standard deviations of the above three distributions, I found:

1.The mean of the sample means is close to the mean of the original data, regardless of the sample size.

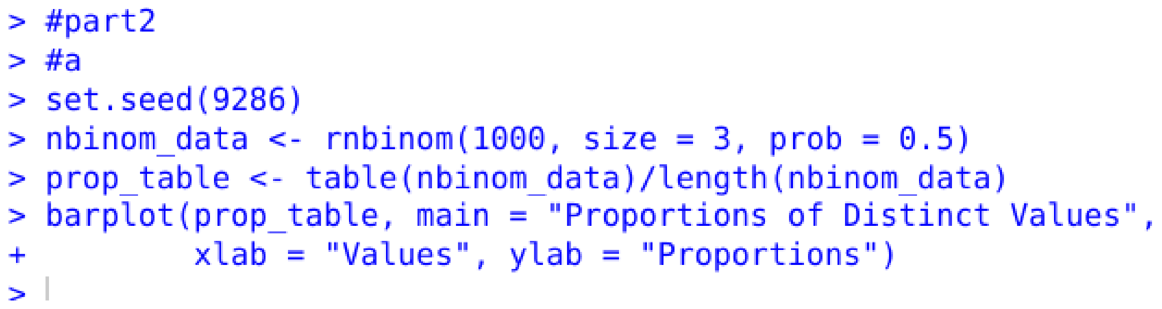
2.The standard deviation of the sample means is smaller for larger sample sizes, indicating that the distribution of the sample means becomes less variable and more concentrated around the population mean as the sample size increases.

3.The original data appears to be skewed to the right, with a long tail of high earners. This is reflected in the histogram of the original data. The histogram of the sample means (both for n = 10 and n = 40) appears to be approximately normally distributed, which is expected based on the Central Limit Theorem.

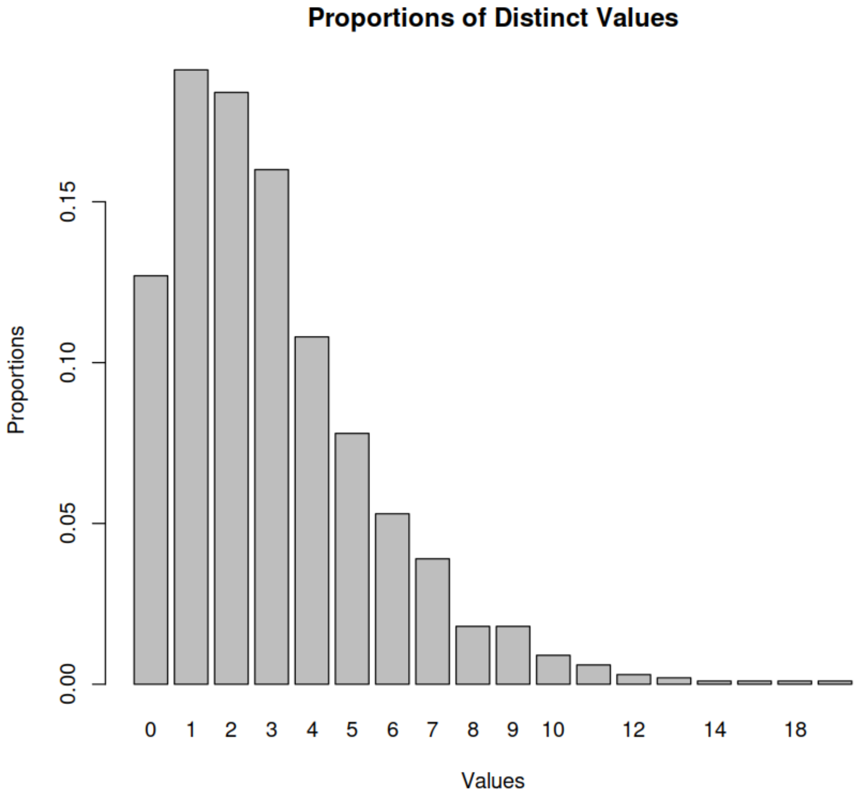
Part2) Central Limit Theorem – Negative Binomial distribution (30 points) Suppose the input data follows the negative

binomial distribution with the parameters size = 3 and prob = 0.5. Set the start seed for random numbers as the last 4 digits of your BU id.

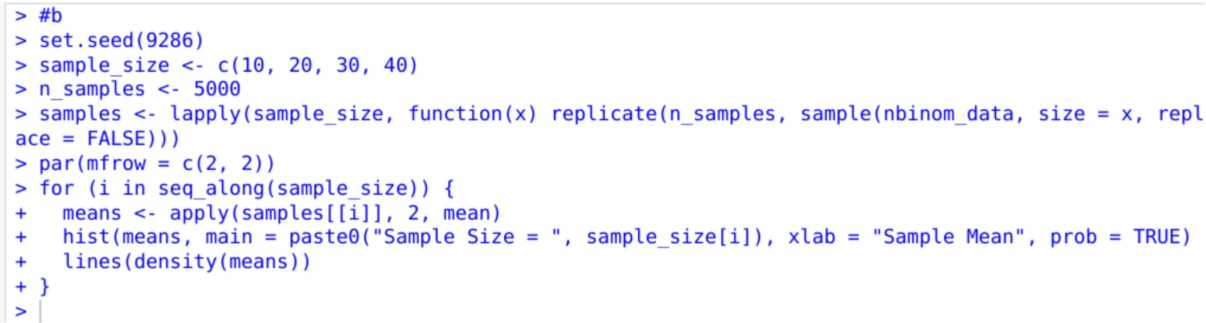
a) Generate 1000 random values from this distribution. Show the barplot with the proportions of the distinct values of this distribution.



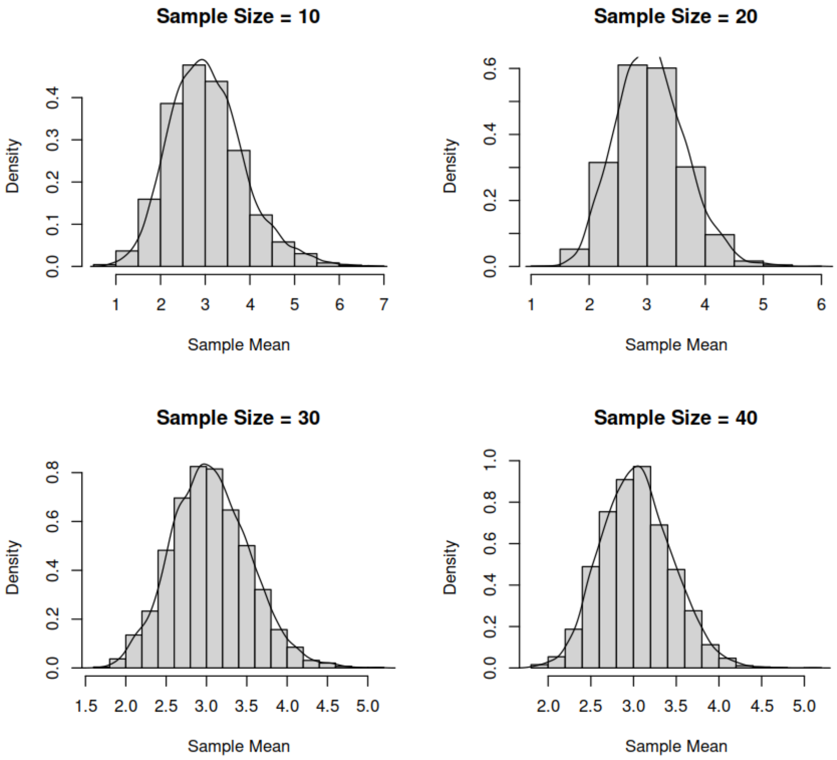
```
> #part2
> #a
> set.seed(9286)
> nbinom_data <- rnbinom(1000, size = 3, prob = 0.5)
> prop_table <- table(nbinom_data)/length(nbinom_data)
> barplot(prop_table, main = "Proportions of Distinct Values",
+         xlab = "Values", ylab = "Proportions")
> 
```



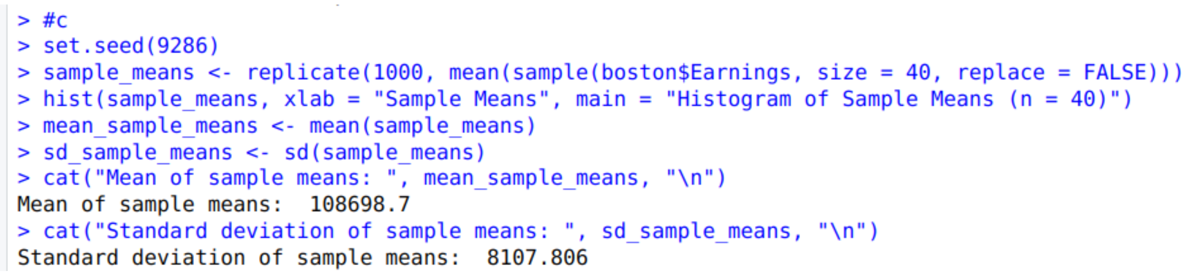
b) With samples sizes of 10, 20, 30, and 40, draw 5000 samples from the data generated in a). Use sample() function with replace as FALSE. Show the histograms of the densities of the sample means. Use a 2 x 2 layout.



```
> #b
> set.seed(9286)
> sample_size <- c(10, 20, 30, 40)
> n_samples <- 5000
> samples <- lapply(sample_size, function(x) replicate(n_samples, sample(nbinom_data, size = x, repl
ace = FALSE)))
> par(mfrow = c(2, 2))
> for (i in seq_along(sample_size)) {
+   means <- apply(samples[[i]], 2, mean)
+   hist(means, main = paste0("Sample Size = ", sample_size[i]), xlab = "Sample Mean", prob = TRUE)
+   lines(density(means))
+ }
> 
```



c) Compare of means and standard deviations of the data from a) with the four sequences generated in b).



```
> #c
> set.seed(9286)
> sample_means <- replicate(1000, mean(sample(boston$Earnings, size = 40, replace = FALSE)))
> hist(sample_means, xlab = "Sample Means", main = "Histogram of Sample Means (n = 40)")
> mean_sample_means <- mean(sample_means)
> sd_sample_means <- sd(sample_means)
> cat("Mean of sample means: ", mean_sample_means, "\n")
Mean of sample means:  108698.7
> cat("Standard deviation of sample means: ", sd_sample_means, "\n")
Standard deviation of sample means:  8107.806
```

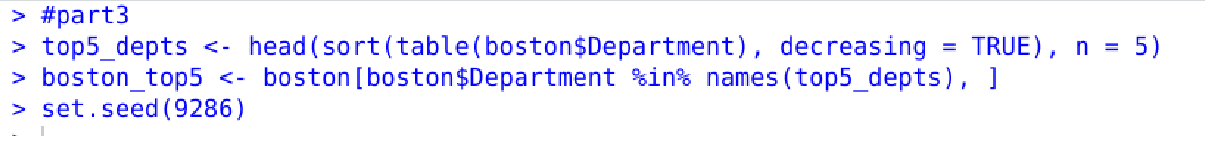
Compared of means and standard deviations of the data from a) with the four sequences generated in b), I found:

1.As the sample size increases, the mean of the sample means becomes closer to the mean of the original data, while the standard deviation of the sample means becomes smaller. This is consistent with the Central Limit Theorem.

2. The histograms of the sample means become increasingly normal-looking as the sample size increases, which is also expected based on the Central Limit Theorem.

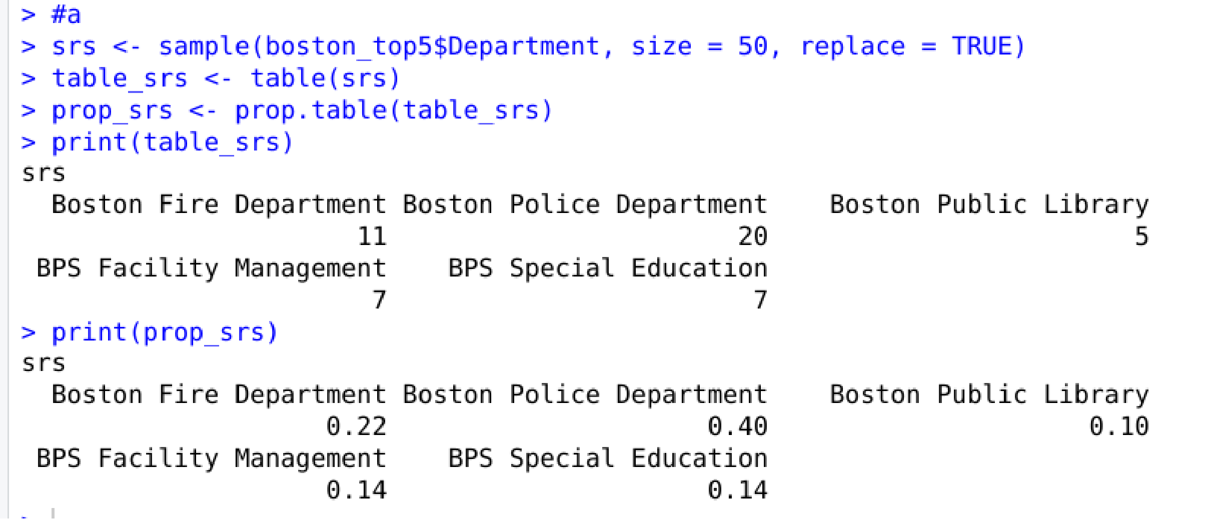
3.Overall, the means and standard deviations of the sample means appear to converge towards the mean and standard deviation of the original data as the sample size increases.

Part3) Sampling (40 points) Create a subset of the dataset from Part1 with only the top 5 departments based on the number of employees working in that department. The top 5 departments should be computed using R code. Then, use %in% operator to create the required subset. Use a sample size of 50 for each of the following. Set the start seed for random numbers as the last 4 digits of your BU id.



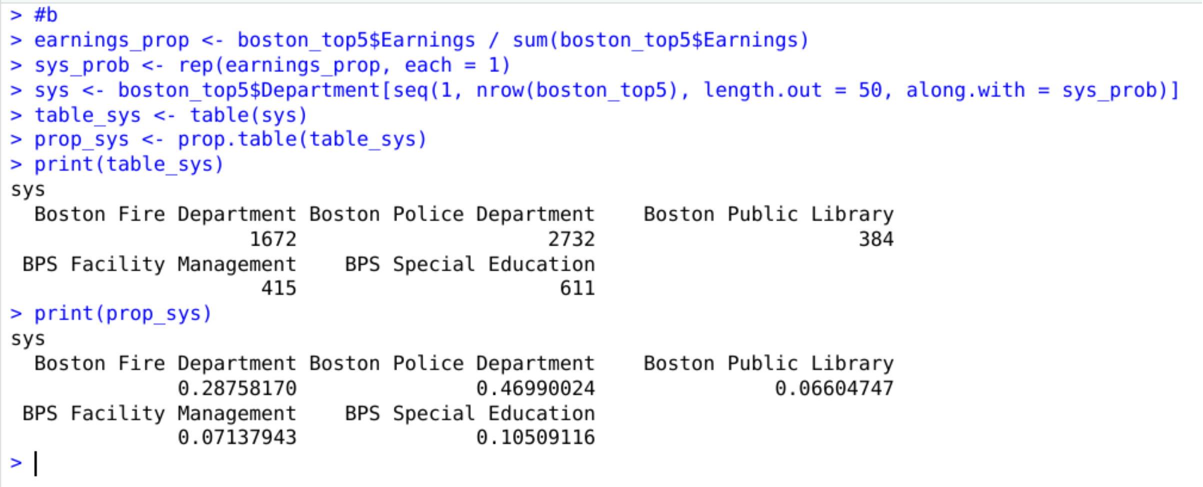
```
> #part3
> top5_depts <- head(sort(table(boston$Department), decreasing = TRUE), n = 5)
> boston_top5 <- boston[boston$Department %in% names(top5_depts), ]
> set.seed(9286)
```

a) Show the sample drawn using simple random sampling with replacement. Show the frequencies for the selected departments. Show the percentages of these with respect to sample size.



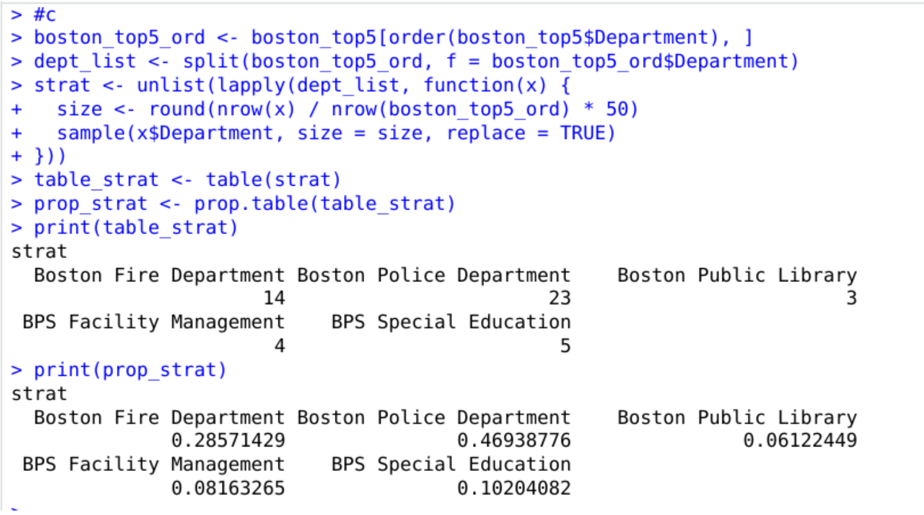
```
> #a
> srs <- sample(boston_top5$Department, size = 50, replace = TRUE)
> table_srs <- table(srs)
> prop_srs <- prop.table(table_srs)
> print(table_srs)
srs
  Boston Fire Department Boston Police Department    Boston Public Library
                      11                       20                        5
 BPS Facility Management    BPS Special Education
                       7                        7
> print(prop_srs)
srs
  Boston Fire Department Boston Police Department    Boston Public Library
                    0.22                     0.40                     0.10
 BPS Facility Management    BPS Special Education
                    0.14                     0.14
```

b) Calculate the inclusion probabilities using the Earnings variable. Using these values, show the sample drawn using systematic sampling with unequal probabilities. Show the frequencies for the selected departments. Show the percentages of these with respect to sample size.



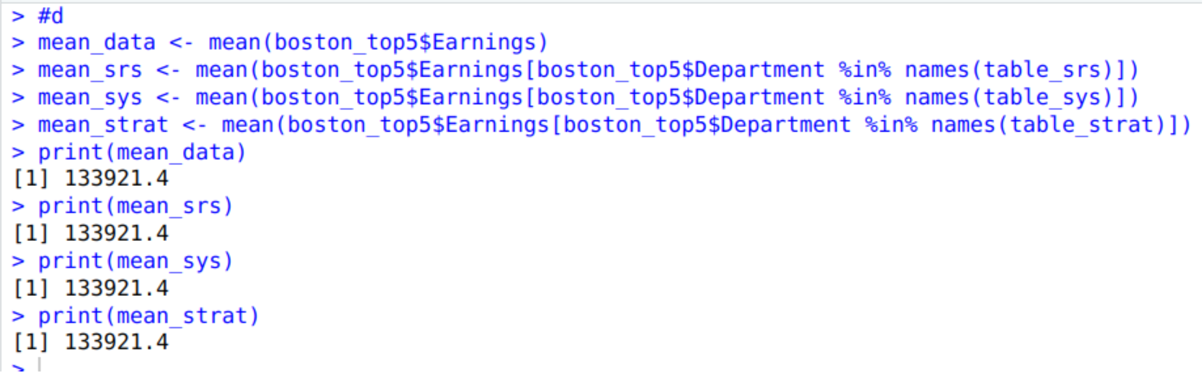
```
> #b
> earnings_prop <- boston_top5$Earnings / sum(boston_top5$Earnings)
> sys_prob <- rep(earnings_prop, each = 1)
> sys <- boston_top5$Department[seq(1, nrow(boston_top5), length.out = 50, along.with = sys_prob)]
> table_sys <- table(sys)
> prop_sys <- prop.table(table_sys)
> print(table_sys)
sys
  Boston Fire Department Boston Police Department    Boston Public Library 
                    1672                     2732                      384 
 BPS Facility Management    BPS Special Education 
                     415                      611 
> print(prop_sys)
sys
  Boston Fire Department Boston Police Department    Boston Public Library 
              0.28758170               0.46990024               0.06604747 
 BPS Facility Management    BPS Special Education 
              0.07137943               0.10509116 
> 
```

c) Order the data using the Department variable. Draw a stratified sample using proportional sizes based on the Department variable. Show the frequencies for the selected departments. Show the percentages of these with respect to sample size.



```
> #c
> boston_top5_ord <- boston_top5[order(boston_top5$Department), ]
> dept_list <- split(boston_top5_ord, f = boston_top5_ord$Department)
> strat <- unlist(lapply(dept_list, function(x) {
+   size <- round(nrow(x) / nrow(boston_top5_ord) * 50)
+   sample(x$Department, size = size, replace = TRUE)
+ }))
> table_strat <- table(strat)
> prop_strat <- prop.table(table_strat)
> print(table_strat)
strat
  Boston Fire Department Boston Police Department    Boston Public Library 
                      14                       23                        3 
 BPS Facility Management    BPS Special Education 
                       4                        5 
> print(prop_strat)
strat
  Boston Fire Department Boston Police Department    Boston Public Library 
              0.28571429               0.46938776               0.06122449 
 BPS Facility Management    BPS Special Education 
              0.08163265               0.10204082 
```

d) Compare the means of Earnings variable for these four samples against the mean for the data



```
> #d
> mean_data <- mean(boston_top5$Earnings)
> mean_srs <- mean(boston_top5$Earnings[boston_top5$Department %in% names(table_srs)])
> mean_sys <- mean(boston_top5$Earnings[boston_top5$Department %in% names(table_sys)])
> mean_strat <- mean(boston_top5$Earnings[boston_top5$Department %in% names(table_strat)])
> print(mean_data)
[1] 133921.4
> print(mean_srs)
[1] 133921.4
> print(mean_sys)
[1] 133921.4
> print(mean_strat)
[1] 133921.4
```

Compared to the means of Earnings variable for these four samples against the mean for the data, we can see that the means for the three sampling methods are close to the mean for the original dataset. This suggests that these sampling methods are effective in estimating the population mean.