Assessment 4_Part 1

Ameeta

2025-09-17

Download file for geneexpression.tsv The link for raw file were copied from the github and then downloaded in R-markdown using te following download.file function and then the file was saved in geneexpression.tsv using destfile argument.

```
# download the gene expression file
download.file("https://raw.githubusercontent.com/ghazkha/Assessment4/refs/heads/main/gene_expression.ts
```

Read in the file, making the gene identifiers the row names. Show a table of values for the first six genes. The downloaded file was read using the read.table function. and then it was saved to gene_data. Following that, the first 6 rows containing 6 gene-identifier were displayed using the head function.

```
# Read the downloaded file
gene_data <- read.table("geneexpression.tsv",     # path to the file
                        header=TRUE,       # indicates first row of the file contains column names
                        sep = "\t",        # \t is the standard for tsv files
                        row.names = 1,    #indicates that first colum contains row names
                        stringsAsFactors = FALSE )  # keep as plain character strings

# display the first 6 genes of the file
head(gene_data, 6) # 6 inidcates number of rows to be displayed
```

```
##                                GTEX.1117F.0226.SM.5GZZ7 GTEX.1117F.0426.SM.5EGHI
## ENSG00000223972.5_DDX11L1                             0                        0
## ENSG00000227232.5_WASH7P                            187                      109
## ENSG00000278267.1_MIR6859-1                           0                        0
## ENSG00000243485.5_MIR1302-2HG                         1                        0
## ENSG00000237613.2_FAM138A                             0                        0
## ENSG00000268020.3_OR4G4P                              0                        1
##                                GTEX.1117F.0526.SM.5EGHJ
## ENSG00000223972.5_DDX11L1                             0
## ENSG00000227232.5_WASH7P                            143
## ENSG00000278267.1_MIR6859-1                           1
## ENSG00000243485.5_MIR1302-2HG                         0
## ENSG00000237613.2_FAM138A                             0
## ENSG00000268020.3_OR4G4P                              0
```

Step 2: A new column was created by the name meanexpression for the mean of the other columns. Then the means were calculated using rowMeans functions and output was saved in $meanofcolumns in gene_data file.

```
gene_data$meanexpression <- rowMeans(gene_data) # rowMeans calculate means across all columns for each
head(gene_data, 6)
```

```
##                                GTEX.1117F.0226.SM.5GZZ7 GTEX.1117F.0426.SM.5EGHI
## ENSG00000223972.5_DDX11L1                             0                        0
## ENSG00000227232.5_WASH7P                            187                      109
```

```
## ENSG00000278267.1_MIR6859-1                          0                       0
## ENSG00000243485.5_MIR1302-2HG                        1                       0
## ENSG00000237613.2_FAM138A                            0                       0
## ENSG00000268020.3_OR4G4P                             0                       1
##                              GTEX.1117F.0526.SM.5EGHJ meanexpression
## ENSG00000223972.5_DDX11L1                            0      0.0000000
## ENSG00000227232.5_WASH7P                           143    146.3333333
## ENSG00000278267.1_MIR6859-1                          1      0.3333333
## ENSG00000243485.5_MIR1302-2HG                        0      0.3333333
## ENSG00000237613.2_FAM138A                            0      0.0000000
## ENSG00000268020.3_OR4G4P                             0      0.3333333
```

Step 3: List the 10 genes with the highest mean expression Firstly, the meanexpression in gene_data were ordered in the descending order using order function and then it was saved in gene_data_sorted file. After that, first 10 rows containing 10 genes with highest meanexpression were displayed in gene_data_sorted using a drop argument to ensure the datas are obtained in table format and not in vector format.

```
#Step 3: Listing 10 genes with the higest mean expression
# order the "meanexpression" of the gene-data in descending order and save in gene_data-sorted file
gene_data_sorted <- gene_data[order(-gene_data$meanexpression), ] # order(-gene_data$meanofcolumns) sor
# show 10 genes with the highest mea expression values
gene_data_sorted[1:10, "meanexpression", drop = FALSE] # drop =FALSE ensures datas are expressed in dat
```

```
##                           meanexpression
## ENSG00000198804.2_MT-CO1        529317.3
## ENSG00000198886.2_MT-ND4        514235.7
## ENSG00000198938.2_MT-CO3        504943.7
## ENSG00000198888.2_MT-ND1        403617.0
## ENSG00000198899.2_MT-ATP6       329751.7
## ENSG00000198727.2_MT-CYB        302254.0
## ENSG00000198763.3_MT-ND2        284217.7
## ENSG00000211445.11_GPX3         270141.7
## ENSG00000198712.1_MT-CO2        265678.0
## ENSG00000156508.17_EEF1A1       232187.3
```

Step 4. Determine the number of genes with a mean <10 The genes with meanexpression less than 10 was determined using a logical vector and then it was saved in gene_data_mean_10. Then the total number of genes with a meanexpression below 10 was calculated by summing the vectors.

```
# create logical vectors for genes with meanexpression <10
gene_data_mean_10 <- gene_data_sorted$meanexpression <10
# count the total number of genes with mean <10
sum(gene_data_mean_10)
```

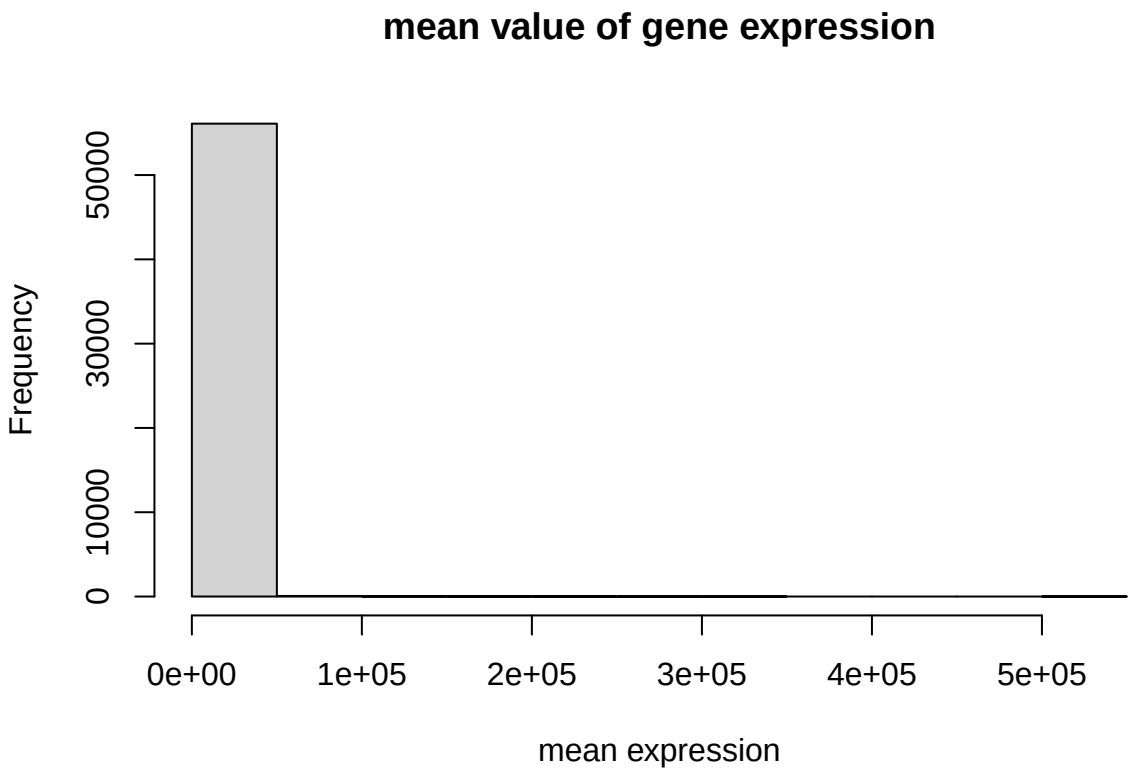
```
## [1] 35988
```

step 5: Make a histogram plot of the mean values and include it into your report. A histogram was generated using hist function to represent the mean of expression level of each genes where x-axis represent the mean expression values and y-axis represent the number of genes within each mean expression range (bin). As shown in the graph, the majority of the genes have a very low mean expression as indicated by the tallest bin near zero. However, there are a few genes which has extremely higher mean expression (5e+05) but appeared to be empty due to the dominance by the tall bar for low gene expression.

```
# Step 6: Histogram of mean values of gene expression
data(gene_data)
```

```
## Warning in data(gene_data): data set 'gene_data' not found
```

```r
hist(gene_data$meanexpression, # data to be plotted
     xlab="mean expression", # label for the x-axis
     main="mean value of gene expression") # Main title of the histogram
```



Step 6: Import this csv file into an R object. What are the column names? The csv raw file for growth data was downloaded using download.file function and was saved as growth_data. The column names were checked using the colnames functions. The column names for this file are site, TreeID, Circumf_2005_cm, Circumf_2010_cm, Circumf_2015_cm, and Circumf_2020_cm.

```r
# download the growth data file
download.file("https://raw.githubusercontent.com/ghazkha/Assessment4/refs/heads/main/growth_data.csv", d
growth_data <- read.csv("growthdata.csv")
colnames(growth_data)
```

```
## [1] "Site"            "TreeID"          "Circumf_2005_cm" "Circumf_2010_cm"
## [5] "Circumf_2015_cm" "Circumf_2020_cm"
```

Step 7: Calculate the mean and standard deviation of tree circumference at the start and end of the study at both sites.

```r
# calculate mean and standard deviation for Northeast site at the start (Circumf_2005_cm)
meannortheast_start<- mean(growth_data$Circumf_2005_cm[growth_data$Site== "northeast"])
meannortheast_start
```

```
## [1] 5.292
```

```r
sdnortheast_start <- sd(growth_data$Circumf_2005_cm[growth_data$Site== "northeast"])
sdnortheast_start
```

```
## [1] 0.9140267
```

```
#calculate mean and standard deviation for norteast at the end ()
meannortheast_end <- mean(growth_data$Circumf_2020_cm[growth_data$Site== "northeast"])
meannortheast_end
```

```
## [1] 54.228
```

```
sdnortheast_end <- sd(growth_data$Circumf_2020_cm[growth_data$Site== "northeast"])
sdnortheast_end
```

```
## [1] 25.22795
```

```
# calculate mean and standard deviation for southwest at the start
meansouthwest_start<- mean(growth_data$Circumf_2005_cm[growth_data$Site== "southwest"])
meansouthwest_start
```

```
## [1] 4.862
```

```
sdsouthwest_start <- sd(growth_data$Circumf_2005_cm[growth_data$Site== "southwest"])
sdsouthwest_start
```

```
## [1] 1.147471
```

```
# calculate # calculate mean and standard deviation for southwest at the end
meansouthwest_end<- mean(growth_data$Circumf_2020_cm[growth_data$Site== "southwest"])
meansouthwest_end
```

```
## [1] 45.596
```

```
sdsouthwest_end <- sd(growth_data$Circumf_2020_cm[growth_data$Site== "southwest"])
sdsouthwest_end
```

```
## [1] 17.87345
```

step 8: Make a box plot of tree circumference at the start and end of the study at both sites.

```
``` r
northeast <- growth_data[growth_data$Site == "northeast", ]
northeast
```

```
##          Site TreeID Circumf_2005_cm Circumf_2010_cm Circumf_2015_cm
## 1   northeast   A012             5.2            10.1            19.9
## 4   northeast   A087             3.8             6.5            10.9
## 6   northeast   A008             5.9            10.0            16.8
## 7   northeast   A035             4.4             9.9            22.2
## 8   northeast   A053             5.3             9.0            15.2
## 12  northeast   A046             3.5             6.8            13.3
## 17  northeast   A070             5.0             8.5            14.3
## 18  northeast   A092             7.2            16.3            36.7
## 21  northeast   A044             5.4            12.2            27.5
## 24  northeast   A052             6.3            10.7            18.0
## 26  northeast   A016             4.8            10.8            24.2
## 28  northeast   A011             5.1            11.5            25.8
## 30  northeast   A099             5.4            12.2            27.4
## 31  northeast   A066             5.8             9.8            16.6
## 34  northeast   A100             6.2            12.2            23.8
## 35  northeast   A071             5.0             8.4            14.3
## 40  northeast   A072             5.7            11.2            22.0
```
```

```
## 44  northeast   A045            5.6             9.5            16.1
## 45  northeast   A030            5.4            10.6            20.8
## 46  northeast   A049            5.7             9.6            16.2
## 48  northeast   A090            6.9            15.4            34.8
## 49  northeast   A055            3.9             6.7            11.3
## 52  northeast   A091            5.2            10.1            19.8
## 53  northeast   A082            6.2            15.9            40.6
## 55  northeast   A042            3.6             9.1            23.4
## 56  northeast   A098            6.6            14.9            33.6
## 57  northeast   A007            6.6            16.9            43.2
## 58  northeast   A025            5.1            12.9            33.2
## 59  northeast   A077            4.1            10.6            27.1
## 60  northeast   A062            4.4             8.7            17.0
## 61  northeast   A086            3.9             8.9            19.9
## 62  northeast   A076            5.4            10.6            20.7
## 68  northeast   A060            6.1            13.7            30.8
## 69  northeast   A034            3.8             7.4            14.5
## 71  northeast   A040            5.1            11.4            25.7
## 73  northeast   A003            6.4            14.3            32.3
## 75  northeast   A005            6.0            15.3            39.3
## 76  northeast   A019            4.1             9.3            20.9
## 77  northeast   A051            6.4            14.5            32.6
## 79  northeast   A083            5.5            14.1            36.1
## 81  northeast   A079            4.2             8.3            16.2
## 82  northeast   A059            4.7             9.3            18.2
## 88  northeast   A068            5.0            12.8            32.7
## 90  northeast   A026            5.6            14.4            37.0
## 91  northeast   A022            5.7            12.8            28.7
## 93  northeast   A057            4.5             8.8            17.2
## 94  northeast   A054            6.4            12.5            24.5
## 95  northeast   A094            6.1            15.6            39.9
## 97  northeast   A093            5.0            11.3            25.4
## 100 northeast   A063            5.4            12.1            27.2
##     Circumf_2020_cm
## 1              38.9
## 4              18.5
## 6              28.4
## 7              50.0
## 8              25.8
## 12             26.0
## 17             24.2
## 18             82.5
## 21             61.8
## 24             30.5
## 26             54.6
## 28             58.0
## 30             61.7
## 31             28.0
## 34             46.7
## 35             24.1
## 40             43.2
## 44             27.2
## 45             40.9
## 46             27.4
```

```
## 48             78.2
## 49             19.0
## 52             38.8
## 53            103.9
## 55             59.8
## 56             75.5
## 57            110.6
## 58             84.9
## 59             69.3
## 60             33.3
## 61             44.8
## 62             40.6
## 68             69.4
## 69             28.5
## 71             57.7
## 73             72.6
## 75            100.6
## 76             47.0
## 77             73.4
## 79             92.4
## 81             31.8
## 82             35.7
## 88             83.8
## 90             94.6
## 91             64.6
## 93             33.7
## 94             48.0
## 95            102.0
## 97             57.2
## 100            61.3
```

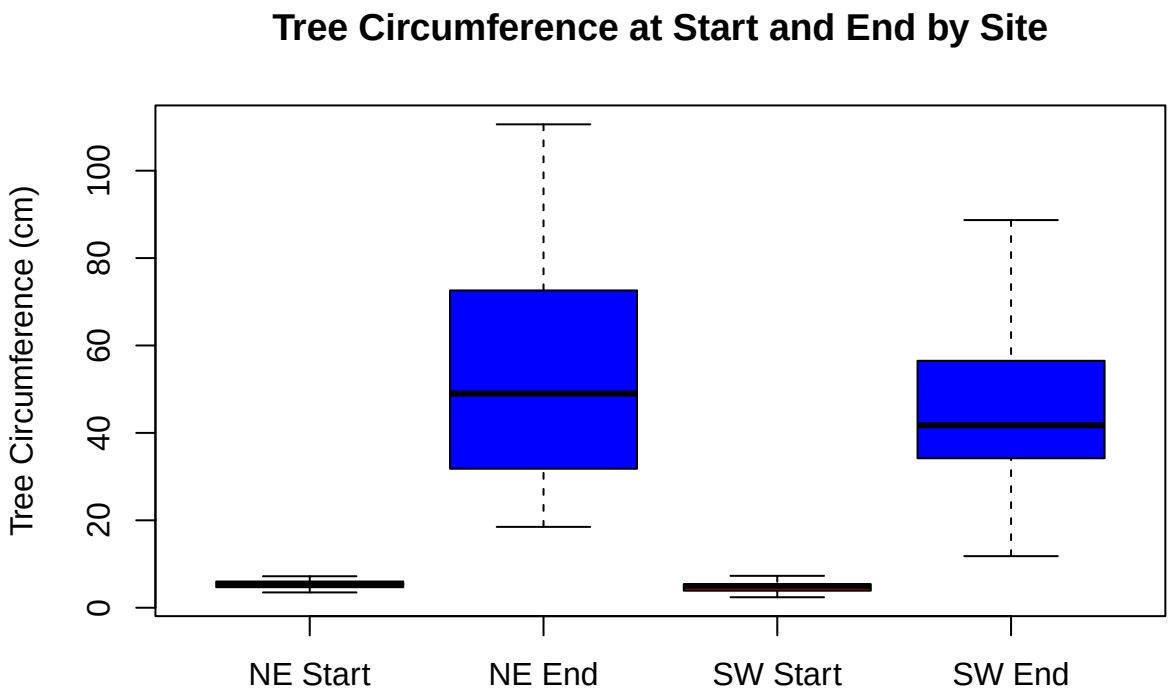
```
southwest <- growth_data[growth_data$Site == "southwest", ]
southwest
```

```
##         Site TreeID Circumf_2005_cm Circumf_2010_cm Circumf_2015_cm
## 2  southwest   A039             4.9             9.6            18.9
## 3  southwest   A010             3.7             7.3            14.3
## 5  southwest   A074             3.8             6.4            10.9
## 9  southwest   A023             7.1            12.0            20.2
## 10 southwest   A024             3.8             7.4            14.5
## 11 southwest   A021             5.4            10.5            20.6
## 13 southwest   A085             2.4             4.1             7.0
## 14 southwest   A064             5.9            13.3            30.0
## 15 southwest   A097             6.5            11.0            18.6
## 16 southwest   A048             2.9             5.7            11.1
## 19 southwest   A095             5.0            11.2            25.1
## 20 southwest   A081             6.5            12.8            25.0
## 22 southwest   A058             6.6            15.0            33.7
## 23 southwest   A096             4.7            10.6            23.8
## 25 southwest   A065             5.1            10.0            19.6
## 27 southwest   A050             3.7             8.2            18.5
## 29 southwest   A015             5.3            11.8            26.6
## 32 southwest   A009             4.3             9.7            21.8
## 33 southwest   A036             4.5            10.2            23.0
## 36 southwest   A002             5.5            10.8            21.2
```

```
## 37 southwest   A047            4.8            8.1           13.7
## 38 southwest   A029            3.2            6.3           12.4
## 39 southwest   A038            4.8            9.4           18.4
## 41 southwest   A032            4.7            9.2           18.1
## 42 southwest   A017            5.3            8.9           15.1
## 43 southwest   A006            5.5            9.3           15.7
## 47 southwest   A061            5.4           12.1           27.2
## 50 southwest   A001            5.1           10.0           19.7
## 51 southwest   A089            5.3           13.5           34.6
## 54 southwest   A033            5.1           11.5           25.9
## 63 southwest   A067            4.7           12.1           31.0
## 64 southwest   A069            4.2            9.4           21.1
## 65 southwest   A028            3.9            7.6           14.9
## 66 southwest   A041            5.0           12.8           32.9
## 67 southwest   A075            3.9            9.9           25.3
## 70 southwest   A073            7.3           14.3           28.0
## 72 southwest   A004            5.3           12.0           27.0
## 74 southwest   A018            2.8            5.5           10.9
## 78 southwest   A056            3.7            8.3           18.6
## 80 southwest   A031            4.7           10.6           23.8
## 83 southwest   A088            5.4           10.6           20.8
## 84 southwest   A080            3.9           10.1           25.9
## 85 southwest   A027            3.3            7.4           16.6
## 86 southwest   A037            6.1           12.0           23.6
## 87 southwest   A013            4.7           11.9           30.5
## 89 southwest   A020            6.0           11.8           23.1
## 92 southwest   A014            6.2           12.2           23.9
## 96 southwest   A084            5.0            9.9           19.4
## 98 southwest   A043            7.0           13.7           26.9
## 99 southwest   A078            3.2            7.3           16.4
##    Circumf_2020_cm
## 2             37.0
## 3             28.1
## 5             18.4
## 9             34.2
## 10            28.4
## 11            40.5
## 13            11.8
## 14            67.6
## 15            31.4
## 16            21.8
## 19            56.5
## 20            49.1
## 22            75.7
## 23            53.6
## 25            38.4
## 27            41.7
## 29            59.9
## 32            49.0
## 33            51.8
## 36            41.5
## 37            23.1
## 38            24.2
## 39            36.0
```

```
## 41            35.5
## 42            25.5
## 43            26.5
## 47            61.2
## 50            38.6
## 51            88.7
## 54            58.3
## 63            79.3
## 64            47.6
## 65            29.2
## 66            84.2
## 67            64.9
## 70            54.9
## 72            60.8
## 74            21.3
## 78            41.8
## 80            53.5
## 83            40.8
## 84            66.2
## 85            37.4
## 86            46.2
## 87            78.0
## 89            45.3
## 92            46.8
## 96            38.0
## 98            52.8
## 99            36.8
```

```r
boxplot(northeast$Circumf_2005_cm, northeast$Circumf_2020_cm,
        southwest$Circumf_2005_cm, southwest$Circumf_2020_cm,
        names = c("NE Start","NE End", "SW Start", "SW End"),
        col = c("red", "blue", "red", "blue"),
        ylab = "Tree Circumference (cm)",
        main = "Tree Circumference at Start and End by Site")
```



Step 9: Calculate the mean growth over the last 10 years at each site.

```
# calculate growth data from 2010 to 2020
growth_data$ growth_10_years <- (growth_data$Circumf_2020_cm - growth_data$Circumf_2010_cm)
growth_data$ growth_10_years
```

```
##   [1] 28.8 27.4 20.8 12.0 12.0 18.4 40.1 16.8 22.2 21.0 30.0 19.2  7.7 54.3 20.4
##  [16] 16.1 15.7 66.2 45.3 36.3 49.6 60.7 43.0 19.8 28.4 43.8 33.5 46.5 48.1 49.5
##  [31] 18.2 39.3 41.6 34.5 15.7 30.7 15.0 17.9 26.6 32.0 26.3 16.6 17.2 17.7 30.3
##  [46] 17.8 49.1 62.8 12.3 28.6 75.2 28.7 88.0 46.8 50.7 60.6 93.7 72.0 58.7 24.6
##  [61] 35.9 30.0 67.2 38.2 21.6 71.4 55.0 55.7 21.1 40.6 46.3 48.8 58.3 15.8 85.3
##  [76] 37.7 58.9 33.5 78.3 42.9 23.5 26.4 30.2 56.1 30.0 34.2 66.1 71.0 33.5 80.2
##  [91] 51.8 34.6 24.9 35.5 86.4 28.1 45.9 39.1 29.5 49.2
```

```
#Calculate mean growth data for last 10 years for each site
North_east_growth_data <- (growth_data$growth_10_years[growth_data$Site == "northeast"])
North_east_growth_data
```

```
##  [1] 28.8 12.0 18.4 40.1 16.8 19.2 15.7 66.2 49.6 19.8 43.8 46.5 49.5 18.2 34.5
## [16] 15.7 32.0 17.7 30.3 17.8 62.8 12.3 28.7 88.0 50.7 60.6 93.7 72.0 58.7 24.6
## [31] 35.9 30.0 55.7 21.1 46.3 58.3 85.3 37.7 58.9 78.3 23.5 26.4 71.0 80.2 51.8
## [46] 24.9 35.5 86.4 45.9 49.2
```

```
North_east_mean_growth_data <- mean(North_east_growth_data)
North_east_mean_growth_data
```

```
## [1] 42.94
```

```
South_west_growth_data <- (growth_data$growth_10_years[growth_data$Site == "southwest"])
South_west_growth_data
```

```
##  [1] 27.4 20.8 12.0 22.2 21.0 30.0  7.7 54.3 20.4 16.1 45.3 36.3 60.7 43.0 28.4
## [16] 33.5 48.1 39.3 41.6 30.7 15.0 17.9 26.6 26.3 16.6 17.2 49.1 28.6 75.2 46.8
## [31] 67.2 38.2 21.6 71.4 55.0 40.6 48.8 15.8 33.5 42.9 30.2 56.1 30.0 34.2 66.1
## [46] 33.5 34.6 28.1 39.1 29.5
```

```
Southwest_mean_growth_data <- mean (South_west_growth_data)
Southwest_mean_growth_data
```

```
## [1] 35.49
```

Step 10 Use the t.test to estimate the p-value that the 10 year growth is different at the two sites. The p-value for the two sites were determined using t-tests. The north_east_growth_data which contains the last 10 years growth data for northeast site and South_west_growth_data which contains last 10 years growt data for southwest site were used to compare the growth between the two sites.The study observed higher mean 10-year growth of 42.94 at the northeast site compared to soutwest site (35.49 cm) but the growth difference was not statistically significant at 5% significance level (t-value = 1.8882, df = 87.978, and p-value = 0.06229).

```
t.test(North_east_growth_data,South_west_growth_data)
```

```
##
##  Welch Two Sample t-test
##
## data:  North_east_growth_data and South_west_growth_data
## t = 1.8882, df = 87.978, p-value = 0.06229
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.3909251 15.2909251
## sample estimates:
## mean of x mean of y
##     42.94     35.49
```