part1

Question 1

Download file `gene_experession.tsv` from github, read file csv, make row name is column one and print first 6 genes.

```
download.file("https://github.com/markziemann/SLE712_files/raw/master/bioinfo_asst3_part1_files/gene_ex
df <- read.csv('gene_expression.tsv', sep='\t', stringsAsFactors = FALSE, row.names = 1)
head(df, 6)
```

```
##                 SRR5150592 SRR5150593
## ENSG00000223972          1          0
## ENSG00000227232          0          1
## ENSG00000278267          0          0
## ENSG00000243485          0          0
## ENSG00000284332          0          0
## ENSG00000237613          0          0
```

Try to access a gene by gene name.

```
df['ENSG00000223972', ]
```

```
##                 SRR5150592 SRR5150593
## ENSG00000223972          1          0
```

Make a new column is mean of other columns

```
df$mean <- rowMeans(df[, 1:2])
head(df)
```

```
##                 SRR5150592 SRR5150593 mean
## ENSG00000223972          1          0  0.5
## ENSG00000227232          0          1  0.5
## ENSG00000278267          0          0  0.0
## ENSG00000243485          0          0  0.0
## ENSG00000284332          0          0  0.0
## ENSG00000237613          0          0  0.0
```

Question 3

Create sorted dataframe by mean column order

```
sorted_df <- df[order(df$mean), ]
```

Show top 10-highest mean genes

```
top10genes <- row.names(tail(sorted_df, 10))
top10genes
```

```
##  [1] "ENSG00000108821" "ENSG00000198712" "ENSG00000196924" "ENSG00000198786"
##  [5] "ENSG00000198804" "ENSG00000137801" "ENSG00000198886" "ENSG00000075624"
##  [9] "ENSG00000210082" "ENSG00000115414"
```

Question 4

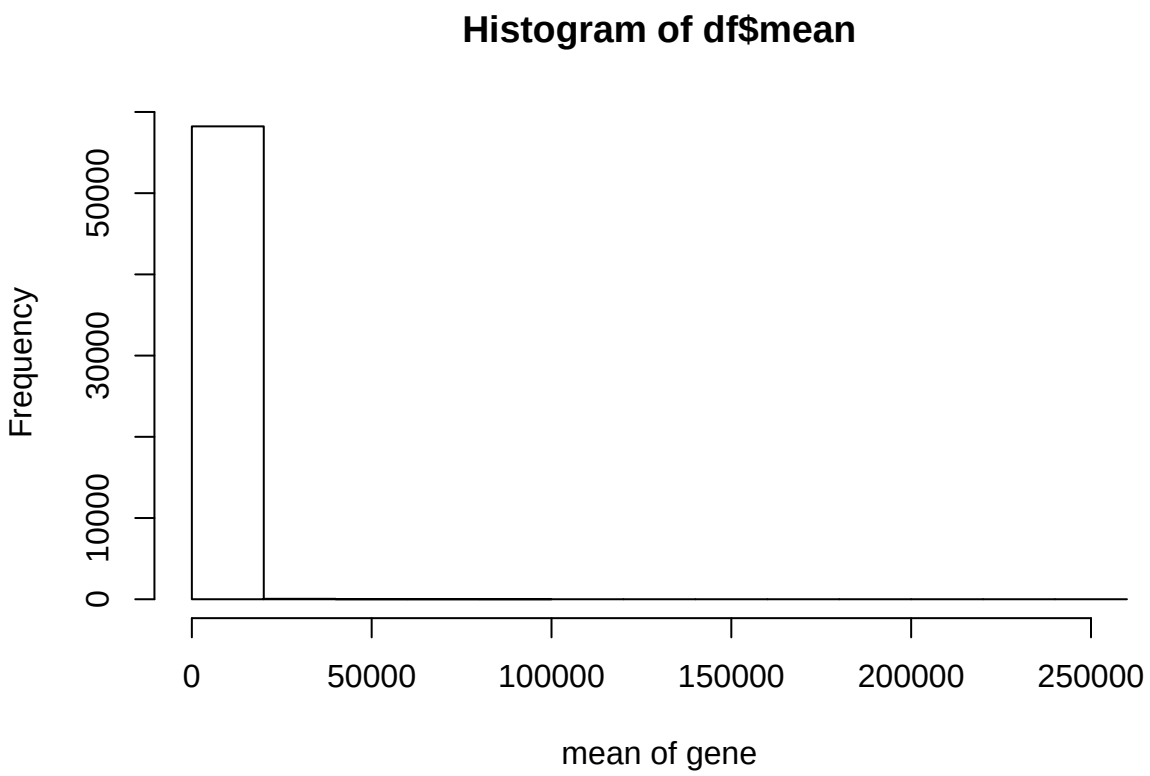
Take mean column after compare this column with 10 (<10). The resutl will be a list of boolean array. Number of genes with mean lower than 10 is sum of this list.

```
number_genes <- sum(df$mean < 10)
number_genes
```

```
## [1] 43124
```

Question 5

```
hist(df$mean, xlab='mean of gene')
```



Question 6

Download growth data from github and load it into a dataframe

```
download.file("https://github.com/markziemann/SLE712_files/raw/master/bioinfo_asst3_part1_files/growth_
df <- read.csv('growth_data.csv')
head(df)
```

```
##        Site TreeID Circumf_2004_cm Circumf_2009_cm Circumf_2014_cm
## 1 northeast   A003             5.2            10.1            19.9
## 2 northeast   A005             4.9             9.6            18.9
## 3 northeast   A007             3.7             7.3            14.3
## 4 northeast   A008             3.8             6.5            10.9
## 5 northeast   A011             3.8             6.4            10.9
## 6 northeast   A012             5.9            10.0            16.8
##   Circumf_2019_cm
## 1            38.9
## 2            37.0
## 3            28.1
## 4            18.5
## 5            18.4
## 6            28.4
```

Print column names of dataframe

```r
colnames(df)
```

```
## [1] "Site"            "TreeID"          "Circumf_2004_cm" "Circumf_2009_cm"
## [5] "Circumf_2014_cm" "Circumf_2019_cm"
```

Mean and standard deviation at 2004 (start)

```r
mean_2004 <- mean(df$Circumf_2004_cm)
cat('Mean at 2004: ', mean_2004)
```

```
## Mean at 2004:  5.077
```

```r
cat('\n')
```

```r
sd_2004 <- sd(df$Circumf_2004_cm)
cat('SD at 2004: ', sd_2004)
```

```
## SD at 2004:  1.054462
```

Mean and standard deviation at 2019 (end)

```r
mean_2019 <- mean(df$Circumf_2019_cm)
cat('Mean at 2019: ', mean_2019)
```

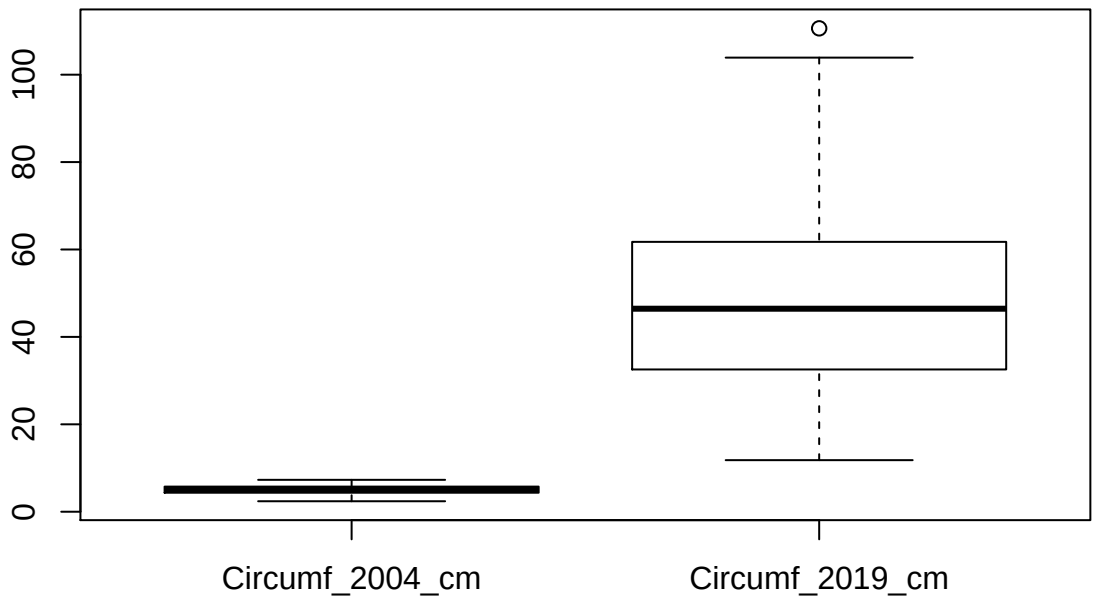
```
## Mean at 2019:  49.912
```

```r
cat('\n')
```

```r
sd_2019 <- sd(df$Circumf_2019_cm)
cat('SD at 2019: ', sd_2019)
```

```
## SD at 2019:  22.17979
```

```r
boxplot(df$Circumf_2004_cm, df$Circumf_2019_cm,
    names=c("Circumf_2004_cm", "Circumf_2019_cm"))
```




```
df$growth <- df$Circumf_2019_cm - df$Circumf_2009_cm
north_growth <- df[df$Site=="northeast", ]$growth
south_growth <- df[df$Site=="southwest", ]$growth
mean_northeast <- mean(north_growth)
mean_southwest <- mean(south_growth)

cat("Mean growth of Northeast over the past  10 years: ", mean_northeast, '\n')
```

```
## Mean growth of Northeast over the past  10 years:  30.076
```

```
cat("Mean growth of Southwest over the past  10 years: ", mean_southwest)
```

```
## Mean growth of Southwest over the past  10 years:  48.354
```

```
t_test_res <- t.test(north_growth, south_growth)
t_test_pvalue <- t_test_res$p.value
wilcox_test_res <- wilcox.test(north_growth, south_growth)
wilcox_pvalue <- wilcox_test_res$p.value

cat('p value of t.test: ', t_test_pvalue, '\n')
```

```
## p value of t.test:  1.712524e-06
```

```r
cat('p value of wilcox.test: ', wilcox_pvalue)
```

```
## p value of wilcox.test:  4.6264e-06
```