Case study: Gene Expression Data Analysis

Experimental setup

There are two varieties of plant xyz, one is **resistant (R)**to a fungal infection while other is susceptible to the fungal **disease (D).** The experiment was conducted to study the differential expression of 490 genes (G1, G2, G3,.., G490), which are involved in 3 different pathways (P1, P2 and P3).

The gene expression was measured in the two mentioned varieties of plants, **R** and **D,** on **Day 2, Day 4, Day 6 and Day 8**.

The gene expression values for variety R on these days are given as **R2, R4, R6 and R8** and that for variety D are given as **D2, D4, D6 and D8.**

The gene expression data is stored in a file "***path.txt***". The first 6 rows of the file are shown below.

Gene	R2	R4	R6	R8	D2	D4	D6	D8	Path
G1	0.01921	0.315849	-0.0778	-0.92117	-0.17061	-0.12	-0.46325	-0.2612	P1
G2	0.032333	0.130992	1.087859	-1.59053	-0.4574	-0.21103	-0.17103	0.016065	P1
G3	-0.83635	0.923642	0.057925	-1.53209	-0.21717	-0.24977	-0.75996	-1.22076	P1
G4	-0.07571	0.182417	0.592473	-1.16334	-0.46948	0.279369	0.185046	0.227038	P1
G5	-0.04873	0.397049	-0.38733	-0.18519	0.06131	-0.10354	-0.47377	-0.37664	P1
G6	-0.5445	0.388928	-1.54556	0.251004	-0.08671	0.176119	-0.18591	-0.03225	P1

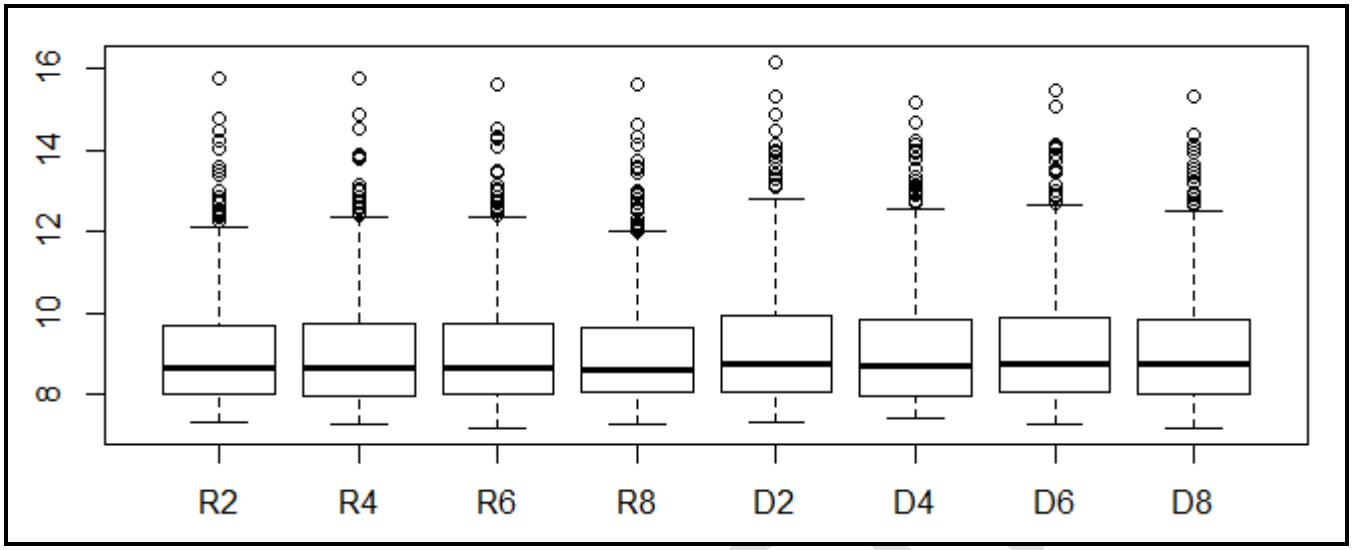
Objective

Data exploration, visualization and analysis of gene expression data, in a given file, using R to find out those genes, which are differentially expressed on 3 or more days between two varieties, **R** and **D.**

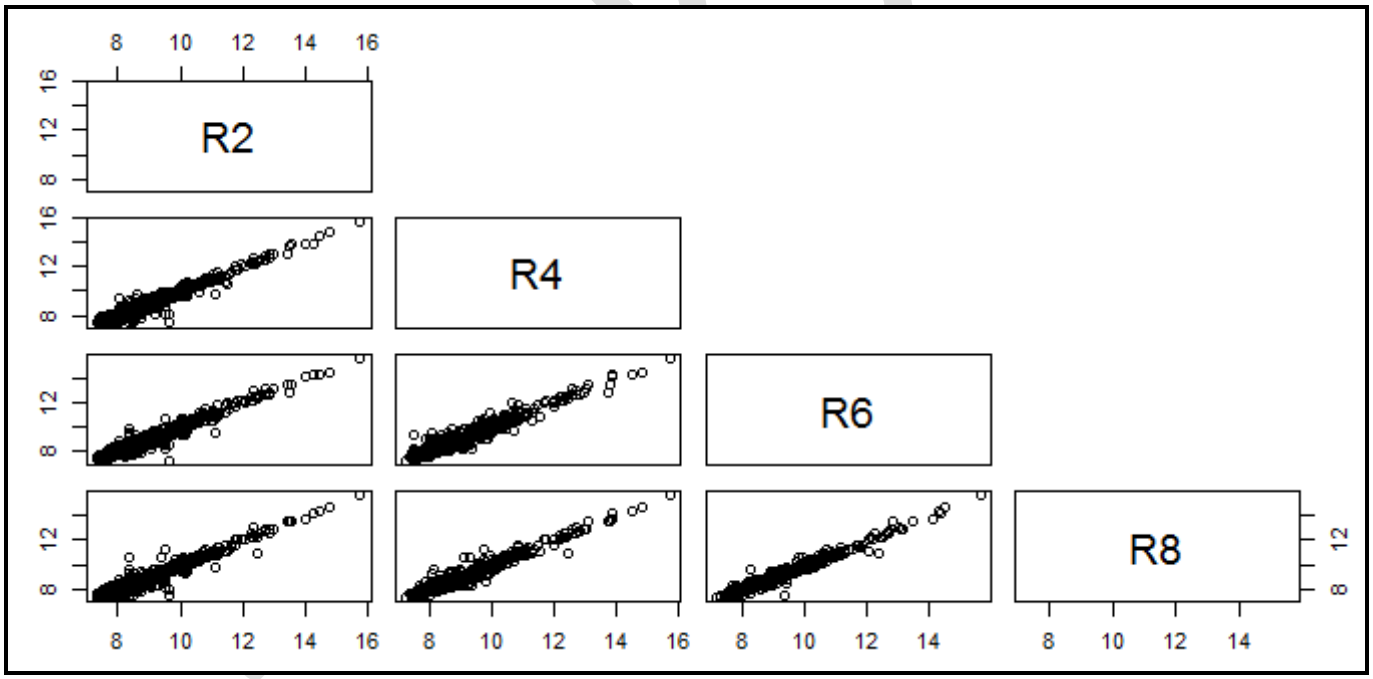
Steps

1. Import the file "path.txt" into R.
2. What is the data structure of imported file?
3. How many rows and columns are there?
4. What are column names?
5. Find out minimum, first quantile, median, third quantile, mean and maximum of expression values on each day. Store the result in a file.

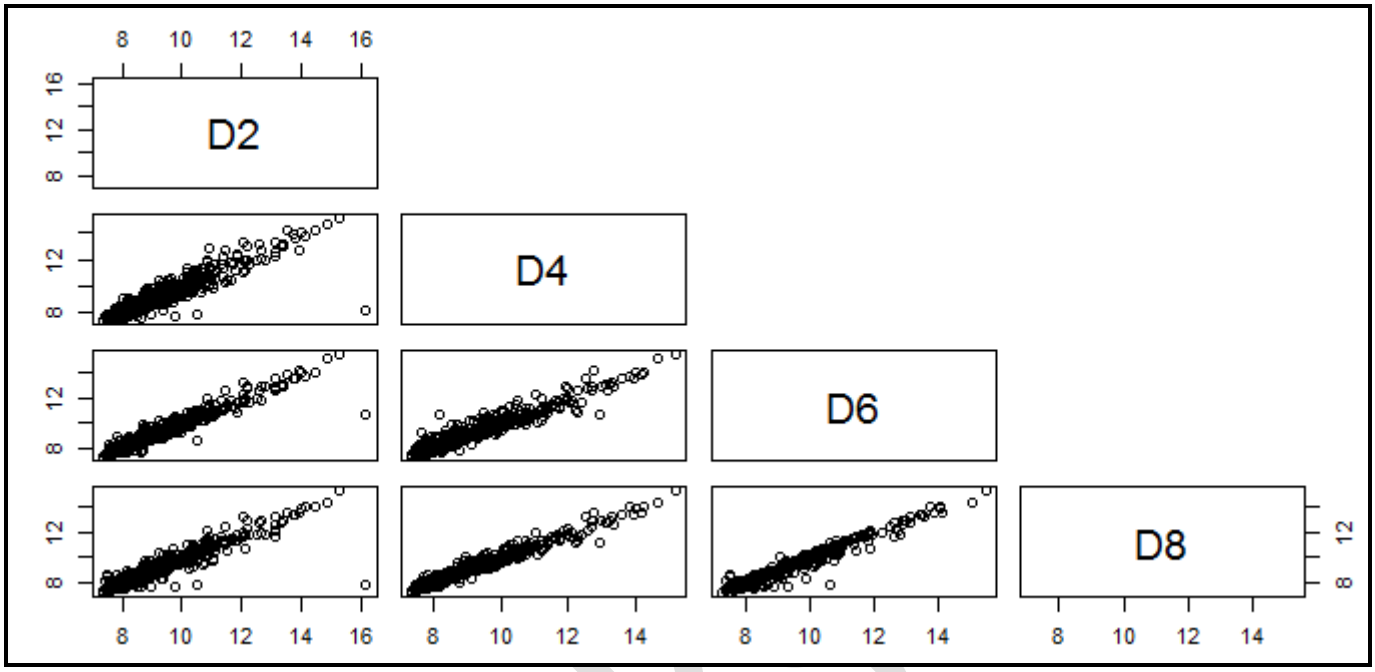
6. Visualization of gene expression on 2,4,6,8 days of D and R plants using boxplot.



7. Visualization of pairwise correlation of gene expressions among R2, R4, R6 and R8.



8. Visualization of pairwise correlation of gene expressions among D2, D4, D6 and D8.



9. Calculate the pairwise correlation coefficient values among R2, R4, R6 and R8.

   Ans:

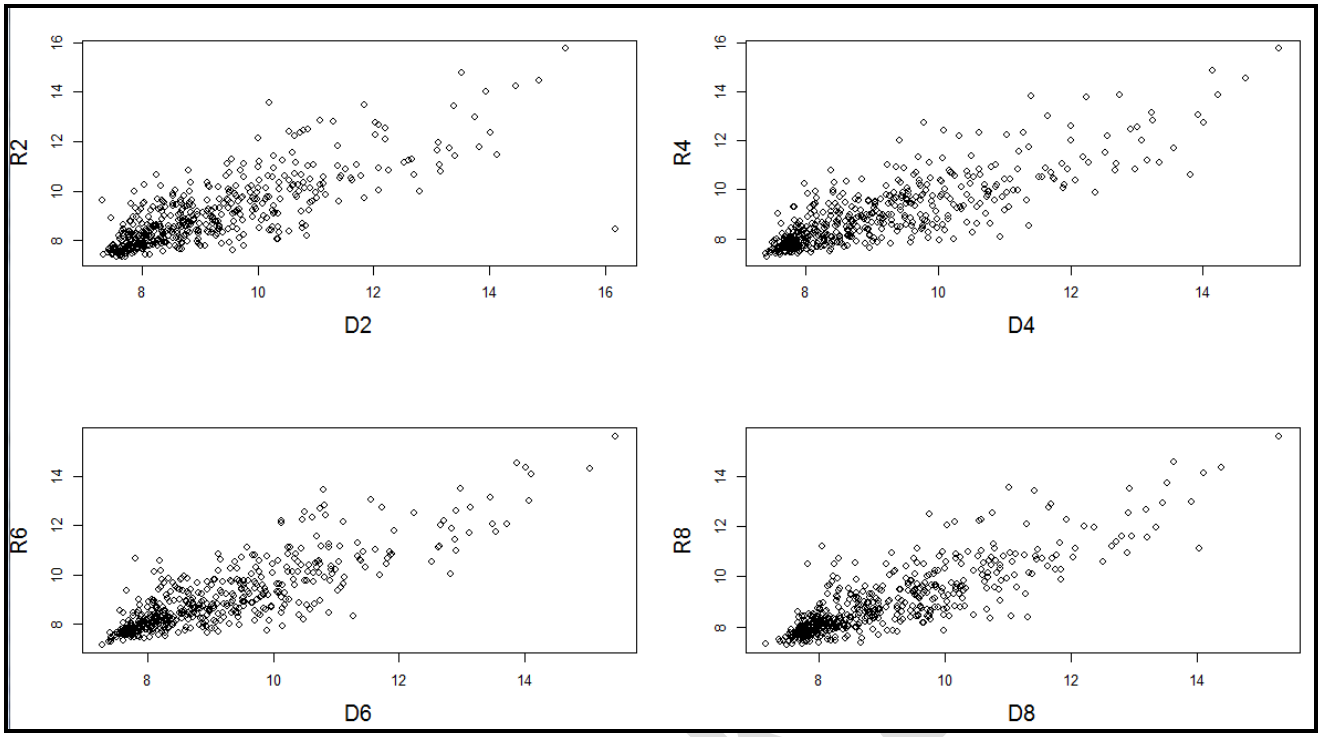
```
          R2        R4        R6        R8
R2 1.0000000 0.9711281 0.9704722 0.9665528
R4 0.9711281 1.0000000 0.9679812 0.9748109
R6 0.9704722 0.9679812 1.0000000 0.9838464
R8 0.9665528 0.9748109 0.9838464 1.0000000
```

10. Calculate the pairwise correlation coefficient values among D2, D4, D6 and D8.

    Ans:

```
          D2        D4        D6        D8
D2 1.0000000 0.9225403 0.9617992 0.9298054
D4 0.9225403 1.0000000 0.9574108 0.9795738
D6 0.9617992 0.9574108 1.0000000 0.9768943
D8 0.9298054 0.9795738 0.9768943 1.0000000
```

11. Draw a four panel plot depicting four scatterplots of **R2 Vs D2**, **R4 Vs D4**, **R6 Vs D6 and R8 Vs D8**.



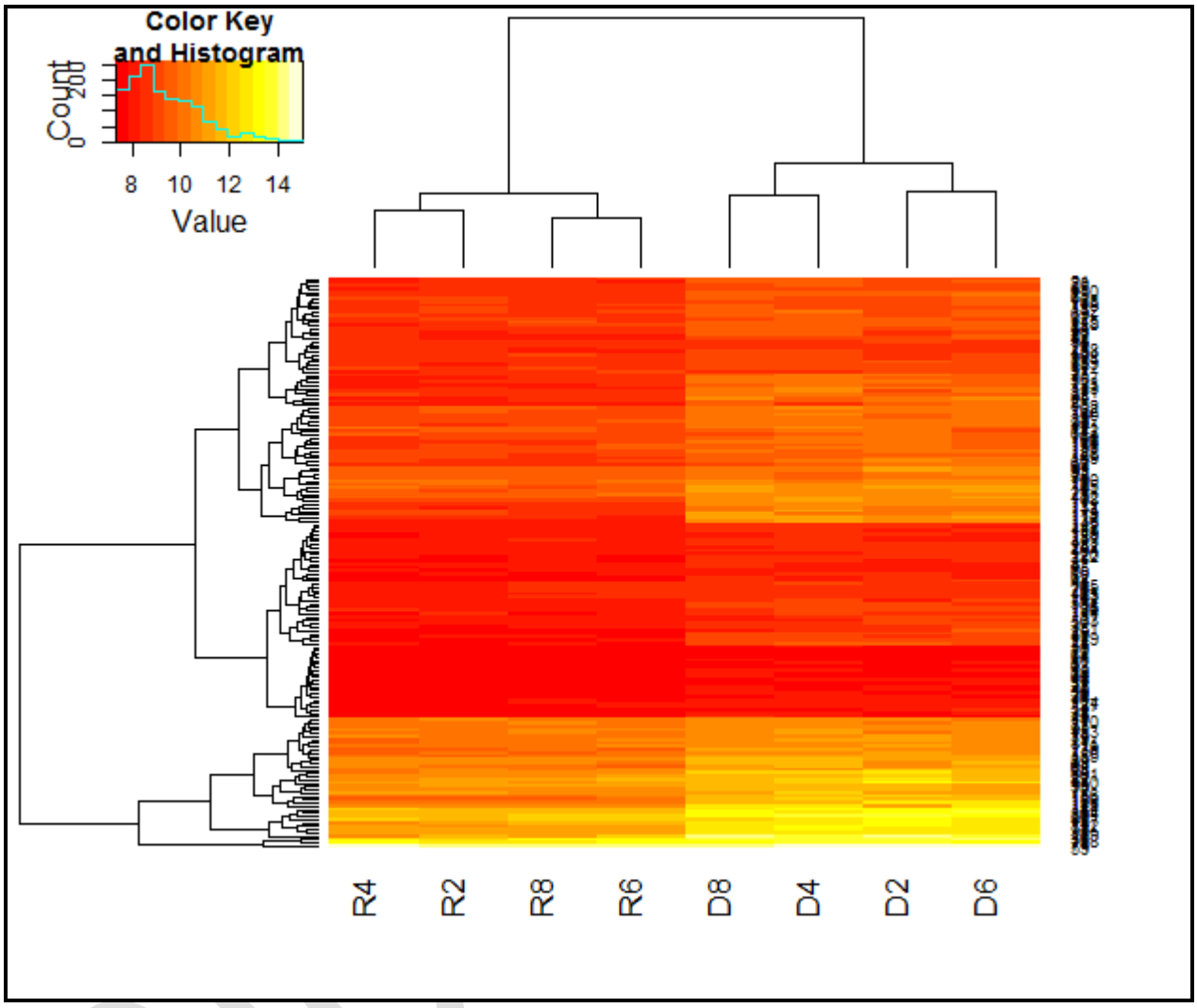
12. Filter those genes that are up-regulated in D variety on all days i.e. (D2-R2)>0; (D4-R4)>0; (D6-R6)>0 and (D8-R8)>0. Write the differential expression values of these filtered genes in a file, *up.txt*.

    Ans: 172 genes.

13. Count the pathway wise gene count for the genes, which are filtered in step 12.
    Ans:

```
P1  P2  P3
106  43  23
```

14. Plot the heatmap showing clustering of genes filtered in step 12. Save the heatmap image.



15. You are provided with an annotation file, anno.txt, of all the genes containing information of gene name, description and accession number. Retrieve the annotations for genes filtered in step 12 from *anno.txt* file**. Hint:** Search "`%in%`" in help and try to understand from the given example.

16. Group the genes as per their pathways. Arrange the values for each group according expression on D2. Write the arranged data in a file, *Genes_arranged.txt.*