QBS177_ALGORITHMS_IN_BIOMED_DATASCIENCE

Breast Cancer Gene Expression Profiles Analysis

...

Presenters:

Avani Kuthe
Hanna Whitehouse
Hyunjoe Yoo
Parinitha Kompala
Xing Cheng

INTRODUCTION

- Breast cancer is the most frequent cancer among women, impacting 2.1 million women each year.
- Breast cancer causes the greatest number of cancer-related deaths among women.
- In 2018 alone, it is estimated that 627,000 women died from breast cancer.
- Cancers are associated with genetic abnormalities.
- Gene expression measures the level of gene activity in a tissue and gives information about its complex activities.
- Comparing the genes expressed in normal and diseased tissue can bring better insights into the cancer prognosis and outcomes.

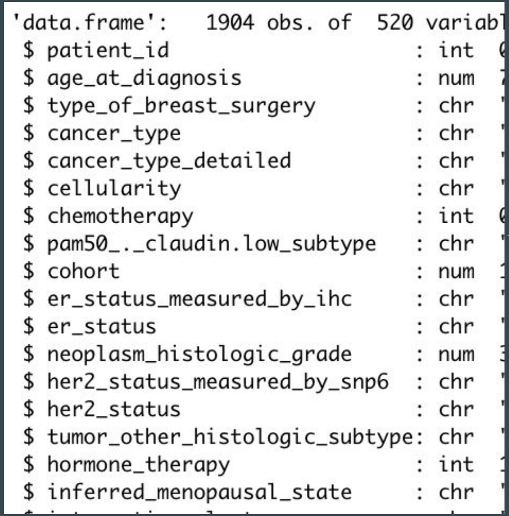
IN THIS PROJECT WE WILL BE DOING :

LINEAR REGRESSION, LOGISTIC REGRESSION,PCA, K-MEANS AND HIERARCHICAL CLUSTERING on the breast cancer gene expression dataset.

DATASET

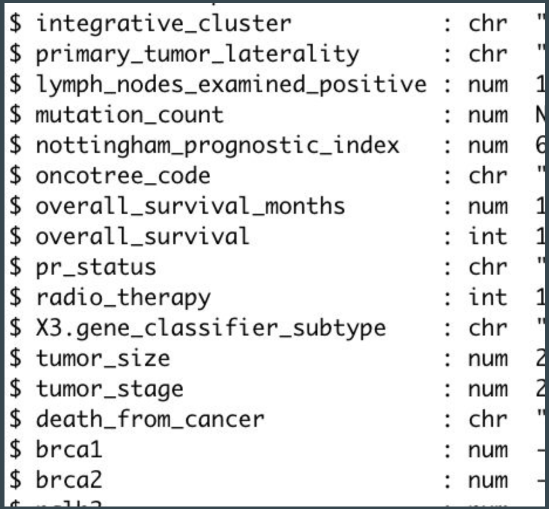
The Molecular Taxonomy of Breast Cancer International Consortium (METABRIC) database is a Canada-UK Project which contains targeted sequencing data of 1,980 primary breast cancer samples.

The dataset was collected by Professor Carlos Caldas from Cambridge Research Institute and Professor Sam Aparicio from the British Columbia Cancer Centre in Canada and published on Nature Communications



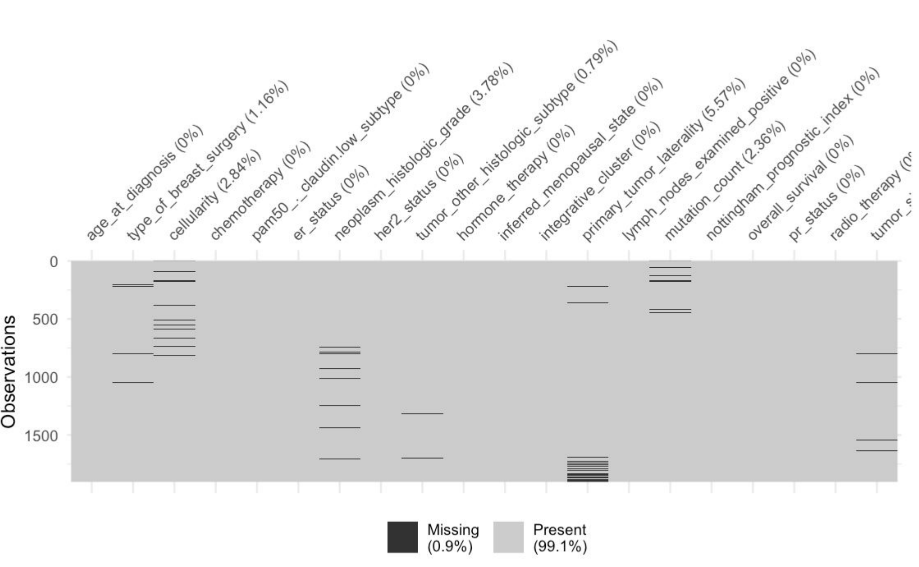
```
'data.frame':   1904 obs. of  520 variabl
 $ patient_id                   : int  0
 $ age_at_diagnosis             : num  7
 $ type_of_breast_surgery       : chr  "
 $ cancer_type                  : chr  "
 $ cancer_type_detailed         : chr  "
 $ cellularity                  : chr  "
 $ chemotherapy                 : int  0
 $ pam50_._claudin.low_subtype  : chr  "
 $ cohort                       : num  1
 $ er_status_measured_by_ihc    : chr  "
 $ er_status                    : chr  "
 $ neoplasm_histologic_grade    : num  3
 $ her2_status_measured_by_snp6 : chr  "
 $ her2_status                  : chr  "
 $ tumor_other_histologic_subtype: chr  "
 $ hormone_therapy              : int  1
 $ inferred_menopausal_state    : chr  "
```

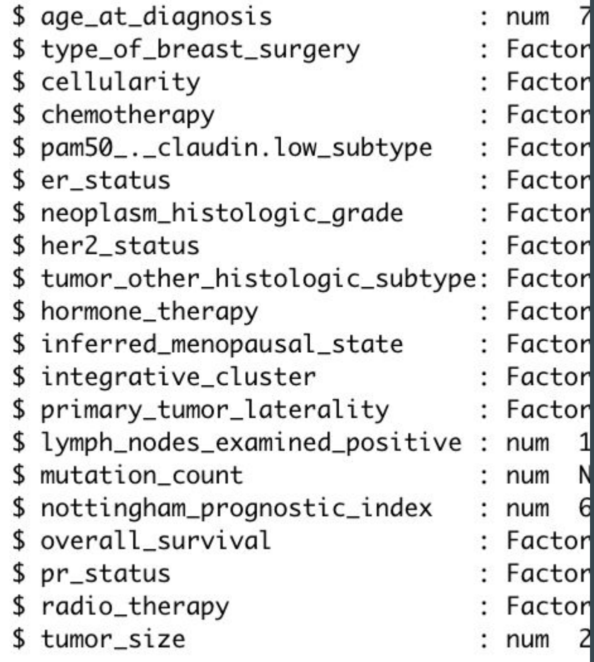
Structure of our dataset



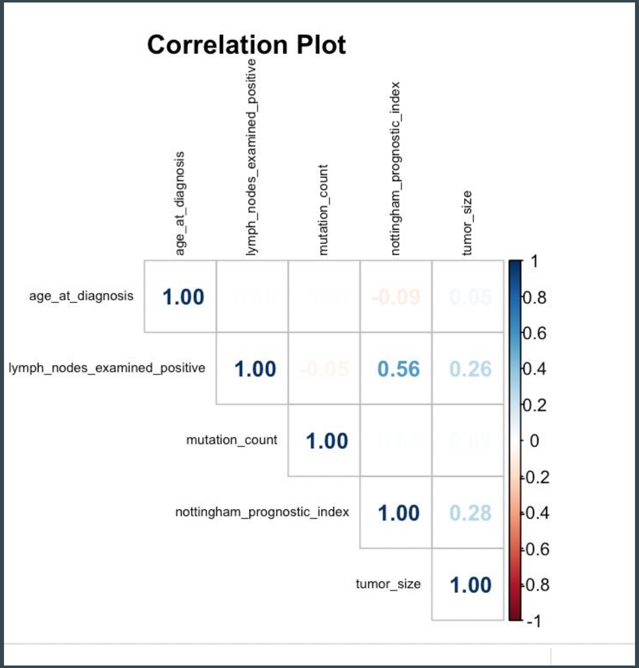
```
$ integrative_cluster          : chr  "
$ primary_tumor_laterality     : chr  "
$ lymph_nodes_examined_positive : num  1
$ mutation_count               : num  N
$ nottingham_prognostic_index  : num  6
$ oncotree_code                : chr  "
$ overall_survival_months      : num  1
$ overall_survival             : int  1
$ pr_status                    : chr  "
$ radio_therapy                : int  1
$ X3.gene_classifier_subtype   : chr  "
$ tumor_size                   : num  2
$ tumor_stage                  : num  2
$ death_from_cancer            : chr  "
$ brca1                        : num  -
$ brca2                        : num  -
```

DATA EXPLORATION AND VISUALIZATION



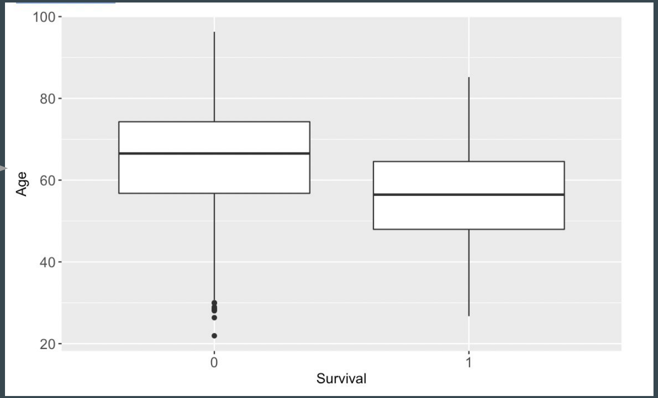


```
$ age_at_diagnosis              : num  7
$ type_of_breast_surgery        : Factor
$ cellularity                   : Factor
$ chemotherapy                  : Factor
$ pam50_._claudin.low_subtype   : Factor
$ er_status                     : Factor
$ neoplasm_histologic_grade     : Factor
$ her2_status                   : Factor
$ tumor_other_histologic_subtype: Factor
$ hormone_therapy               : Factor
$ inferred_menopausal_state     : Factor
$ integrative_cluster           : Factor
$ primary_tumor_laterality      : Factor
$ lymph_nodes_examined_positive : num  1
$ mutation_count                : num  N
$ nottingham_prognostic_index   : num  6
$ overall_survival              : Factor
$ pr_status                     : Factor
$ radio_therapy                 : Factor
$ tumor_size                    : num  2
```

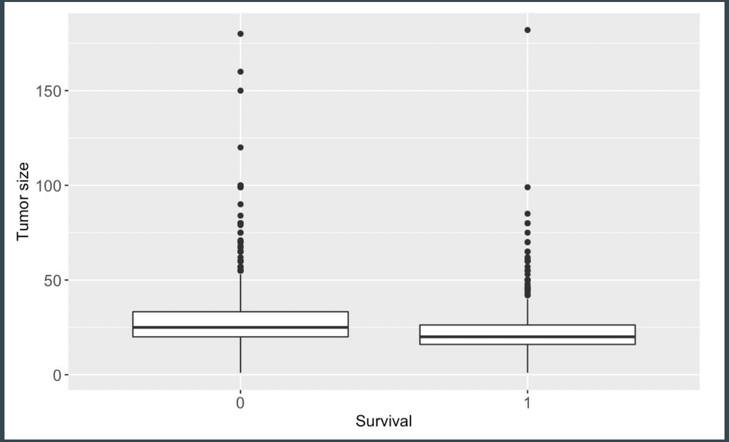


Correlation between few variables

A box plot of age vs survival



A box plot of Tumor size vs survival



PRINCIPAL COMPONENT ANALYSIS (PCA)

Why: genetics part of the dataset contains m-RNA levels z-score for 331 genes

Advantage:

- Speed up the computation
- Improve classification accuracy when multicollinearity exists in the dataset
- Visualize high-dimensional data

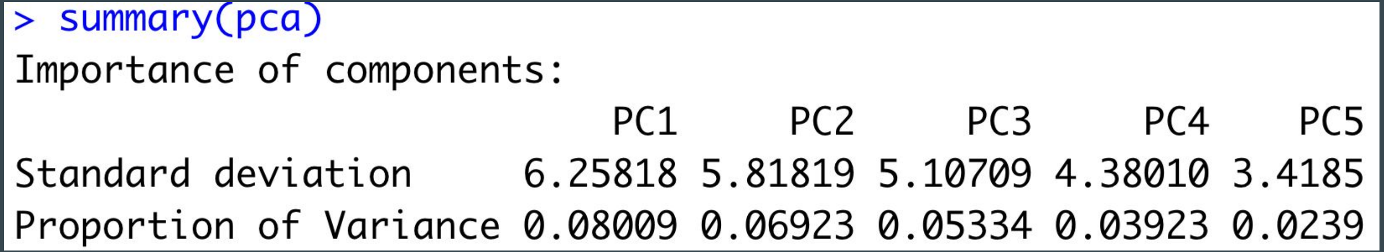
Limitations

- Low interpretability of principal components
- Trade-off between information loss and dimensionality reduction

PRINCIPAL COMPONENT ANALYSIS (PCA)

How:

pca=prcomp(genes,center=TRUE,scale=TRUE)

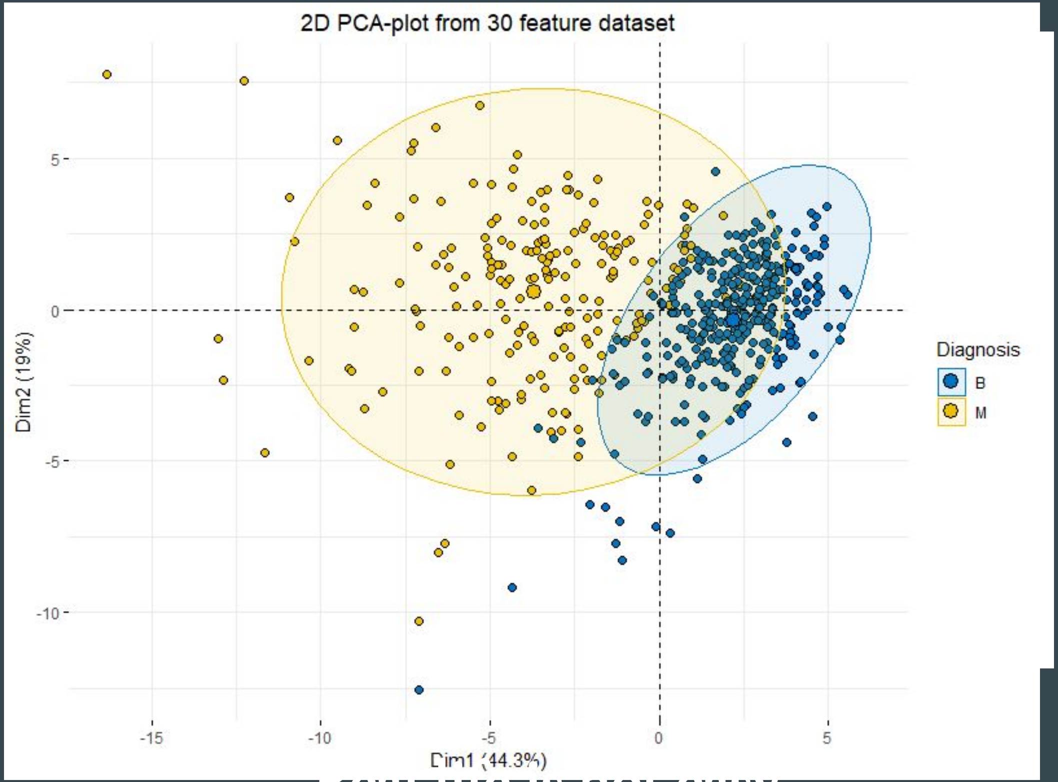


```
> summary(pca)
Importance of components:
                           PC1     PC2     PC3     PC4    PC5
Standard deviation     6.25818 5.81819 5.10709 4.38010 3.4185
Proportion of Variance 0.08009 0.06923 0.05334 0.03923 0.0239
```

Standard deviation: sqrt(eigenvalue), eigenvalue = SS(distances for PC)

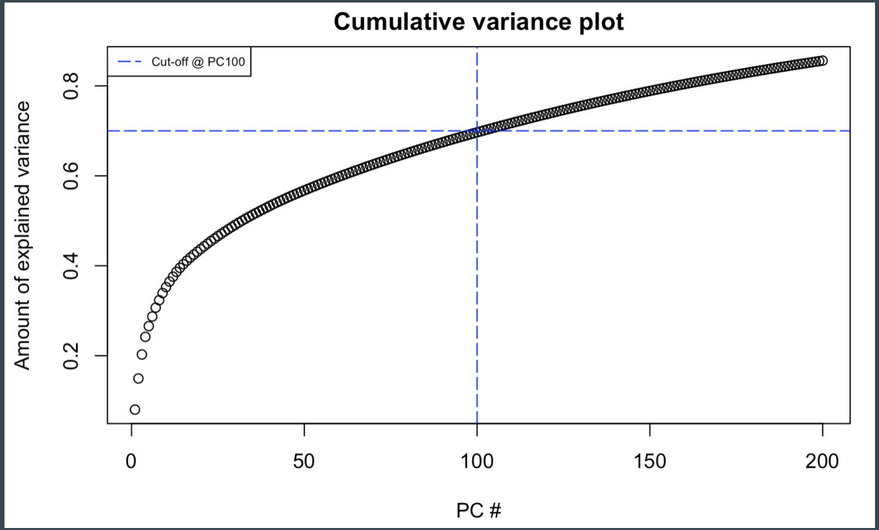
Proportion of Variance: the amount of variance the component accounts for

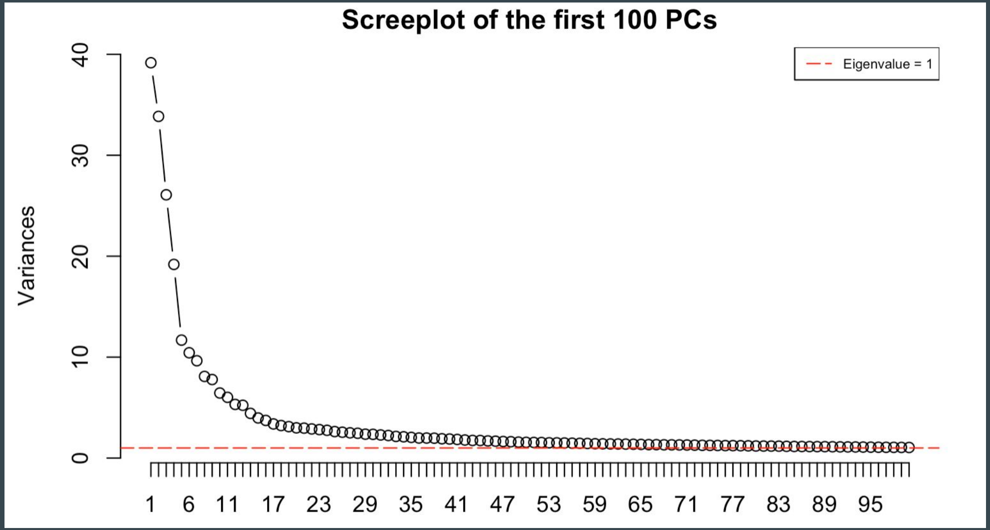
PRINCIPAL COMPONENT ANALYSIS (PCA)



Real life is not easy

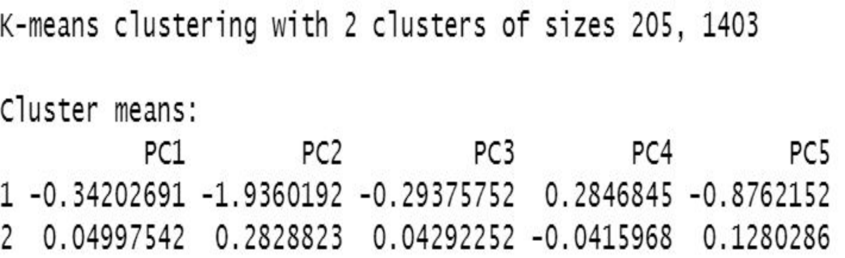
PRINCIPAL COMPONENT ANALYSIS (PCA)



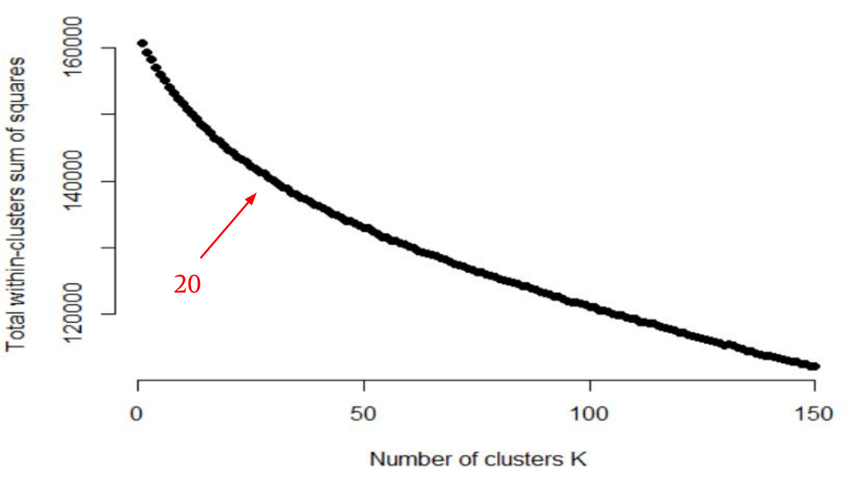


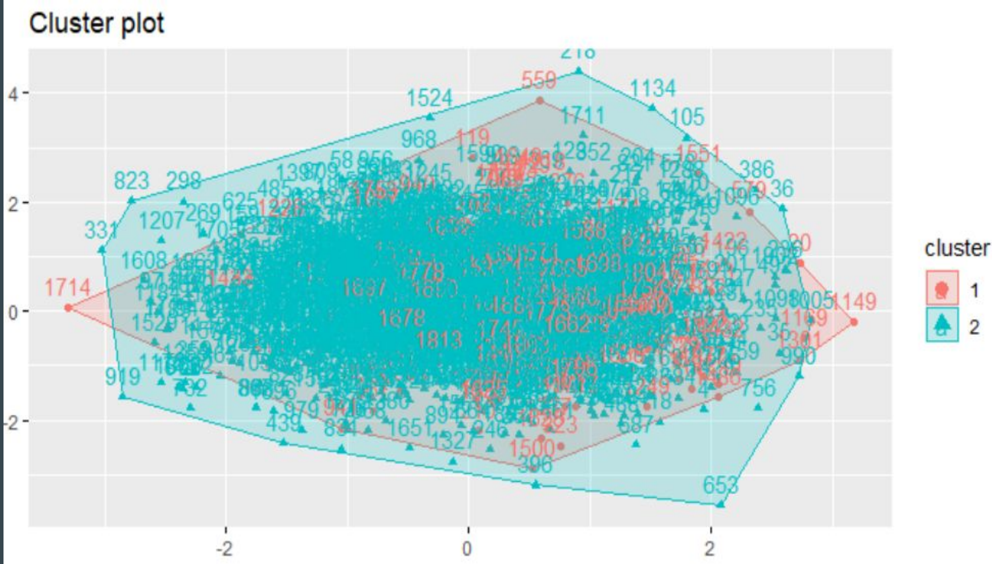
Select the first 100 PCs, which are able to explain 70% variance of genes data

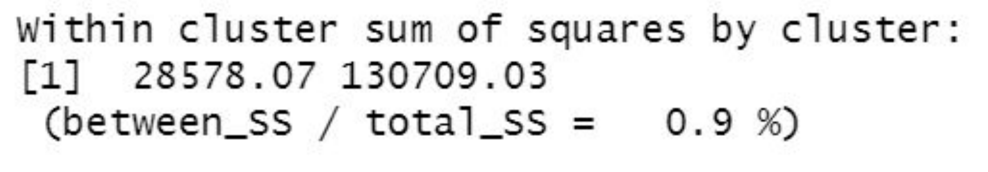
K-MEANS



```
K-means clustering with 2 clusters of sizes 205, 1403

Cluster means:
          PC1        PC2         PC3        PC4        PC5
1 -0.34202691 -1.9360192 -0.29375752  0.2846845 -0.8762152
2  0.04997542  0.2828823  0.04292252 -0.0415968  0.1280286
```

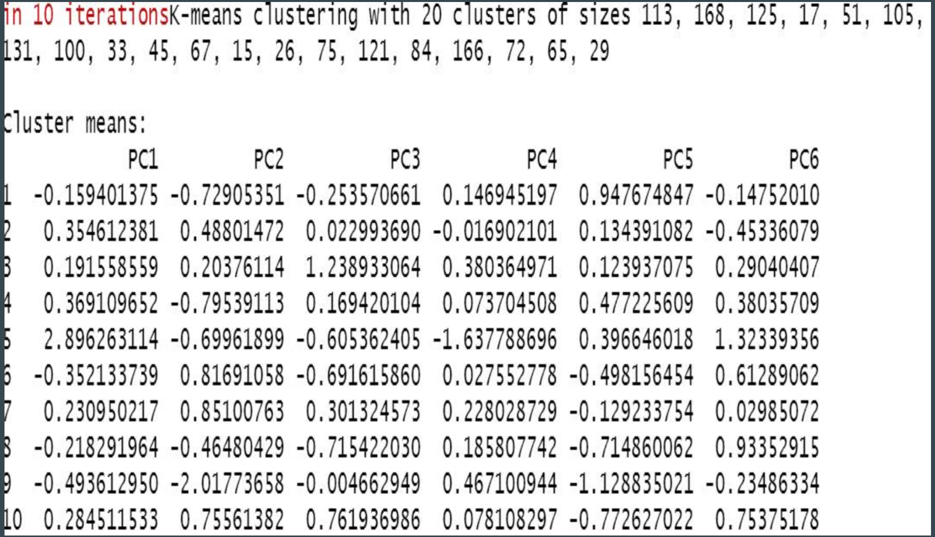




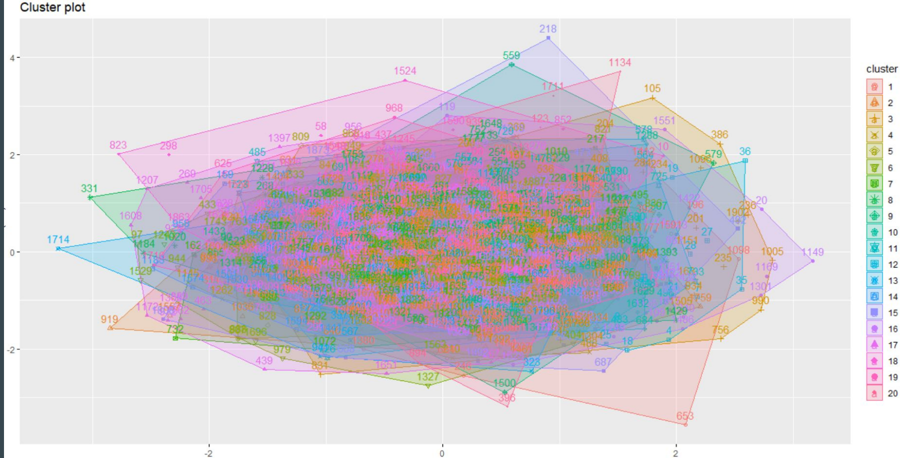


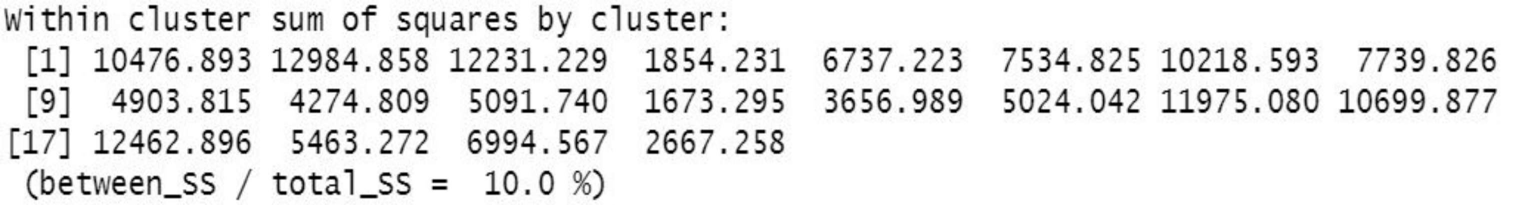
```
Within cluster sum of squares by cluster:
[1]  28578.07 130709.03
 (between_SS / total_SS =   0.9 %)
```

K-Means



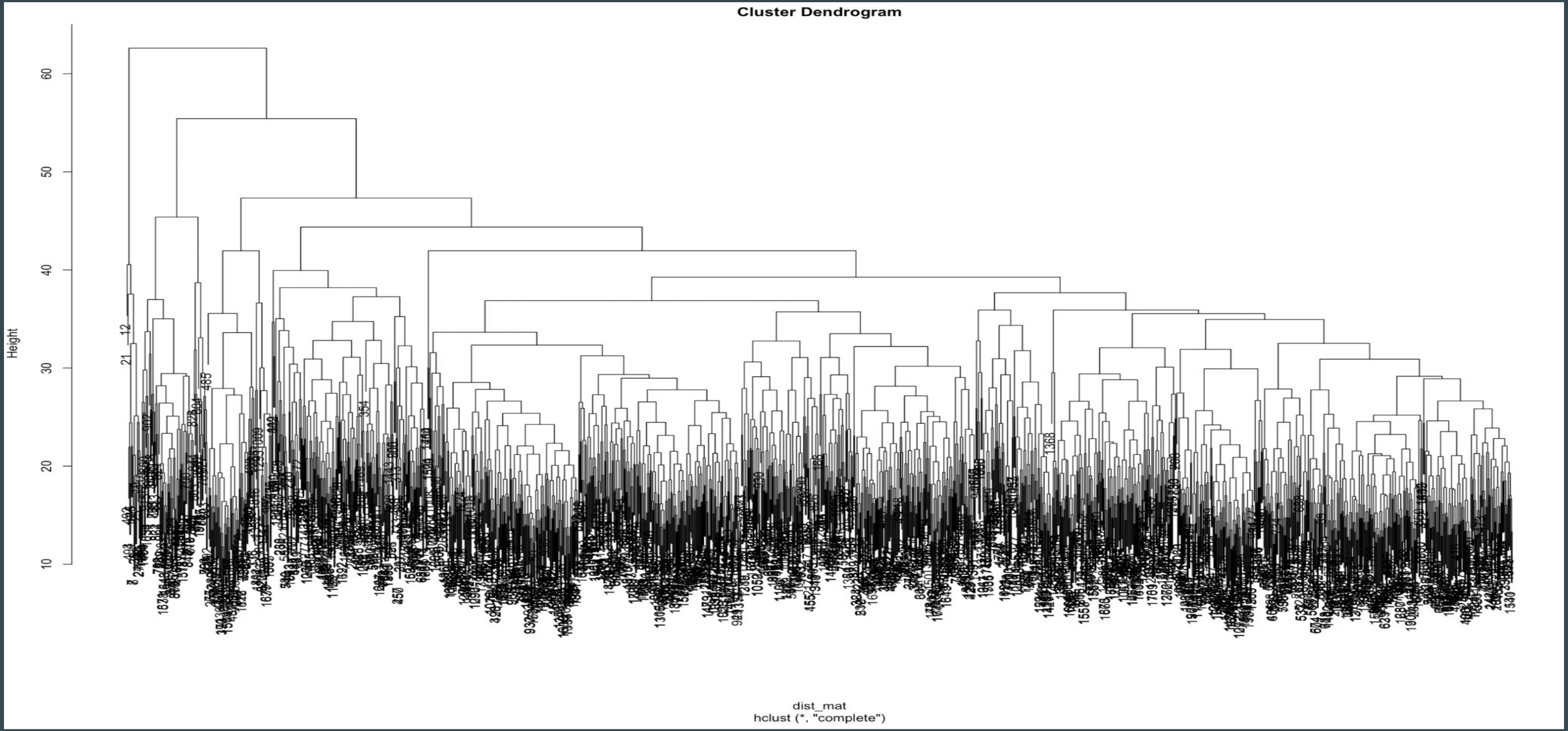
```
in 10 iterationsK-means clustering with 20 clusters of sizes 113, 168, 125, 17, 51, 105,
131, 100, 33, 45, 67, 15, 26, 75, 121, 84, 166, 72, 65, 29

Cluster means:
          PC1         PC2          PC3          PC4          PC5         PC6
1  -0.159401375 -0.72905351 -0.253570661  0.146945197  0.947674847 -0.14752010
2   0.354612381  0.48801472  0.022993690 -0.016902101  0.134391082 -0.45336079
3   0.191558559  0.20376114  1.238933064  0.380364971  0.123937075  0.29040407
4   0.369109652 -0.79539113  0.169420104  0.073704508  0.477225609  0.38035709
5   2.896263114 -0.69961899 -0.605362405 -1.637788696  0.396646018  1.32339356
6  -0.352133739  0.81691058 -0.691615860  0.027552778 -0.498156454  0.61289062
7   0.230950217  0.85100763  0.301324573  0.228028729 -0.129233754  0.02985072
8  -0.218291964 -0.46480429 -0.715422030  0.185807742 -0.714860062  0.93352915
9  -0.493612950 -2.01773658 -0.004662949  0.467100944 -1.128835021 -0.23486334
10  0.284511533  0.75561382  0.761936986  0.078108297 -0.772627022  0.75375178
```

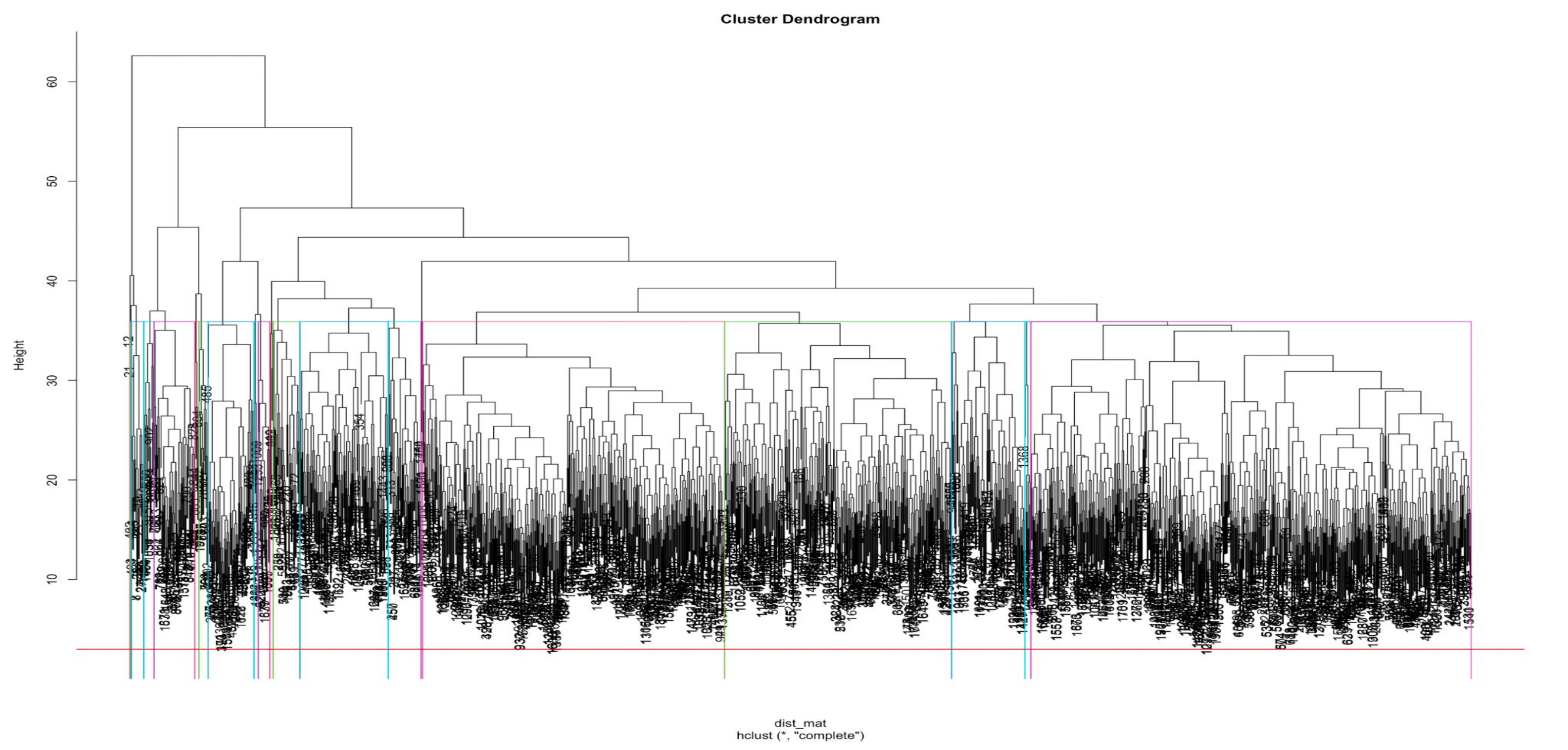


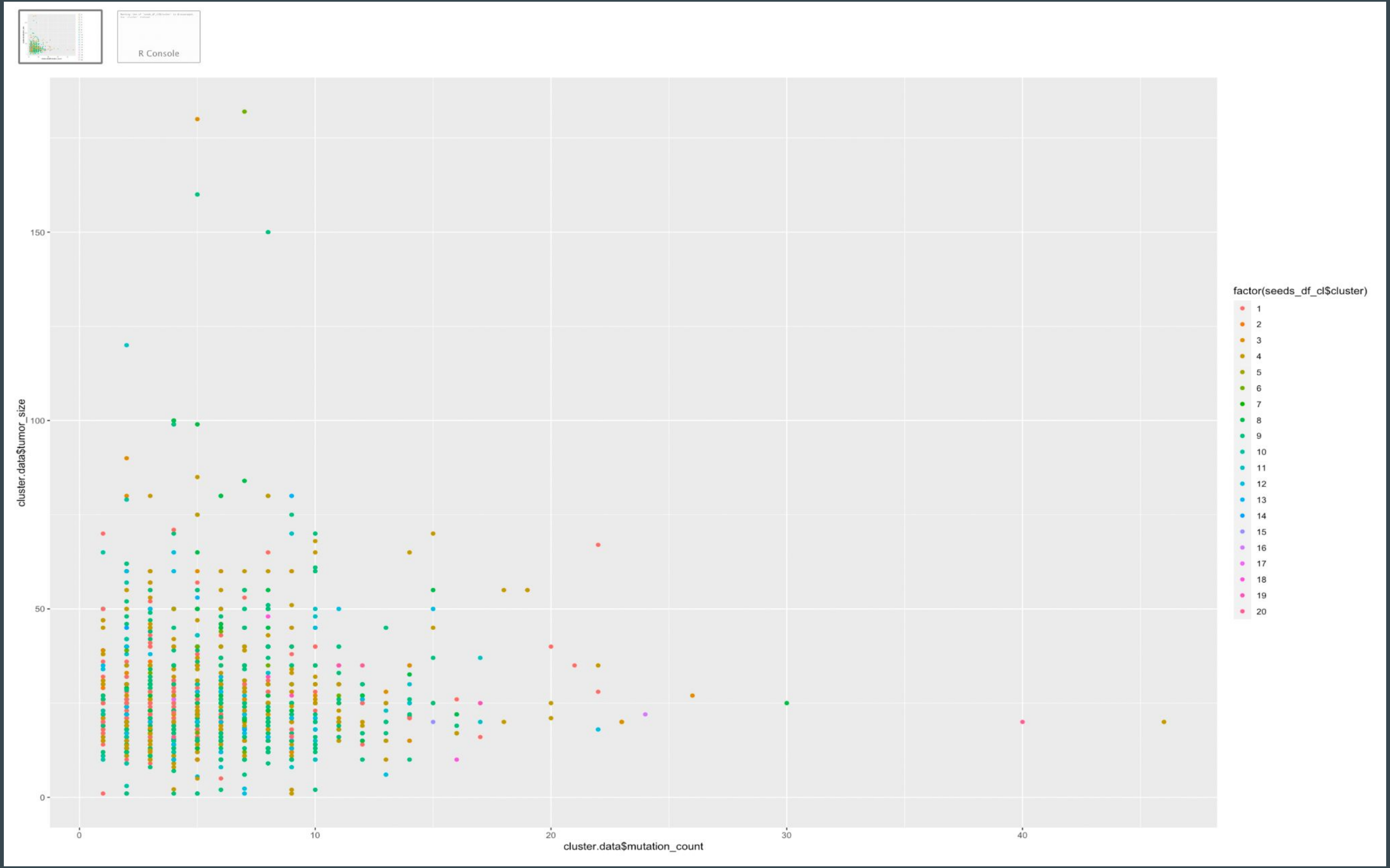


```
Within cluster sum of squares by cluster:
 [1] 10476.893 12984.858 12231.229  1854.231  6737.223  7534.825 10218.593  7739.826
 [9]  4903.815  4274.809  5091.740  1673.295  3656.989  5024.042 11975.080 10699.877
[17] 12462.896  5463.272  6994.567  2667.258
 (between_SS / total_SS =  10.0 %)
```

HIERARCHICAL CLUSTERING




Cluster Dendrogram
Height
60
50
40
30
20
10
12
21
485
354
250
1368
1530
dist_mat
hclust (*, "complete")

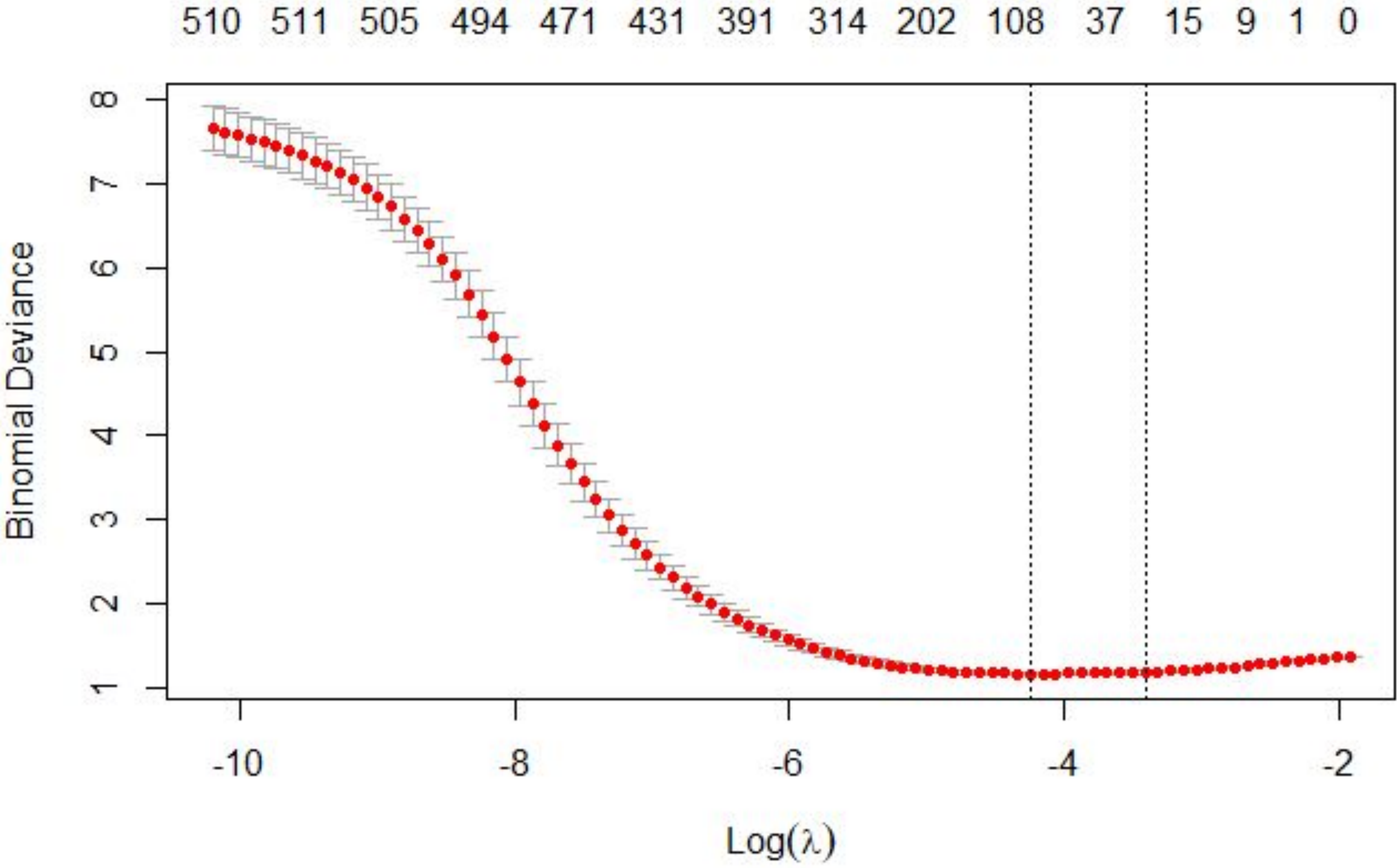

R Console
cluster.data$tumor_size
150
100
50
0
0
10
20
30
40
cluster.data$mutation_count
factor(seeds_df_cl$cluster)
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20

Logistic Regression

Overall Survival Status: Dead or Alive

Lasso

- lambda.min = 0.01431327
- 96 out of 515 variables selected
- Example variables: type of breast surgery, ER status, inferred menopausal state


510 511 505 494 471 431 391 314 202 108 37 15 9 1 0
Binomial Deviance
1
2
3
4
5
6
7
8
-10
-8
-6
-4
-2
Log(λ)

Results

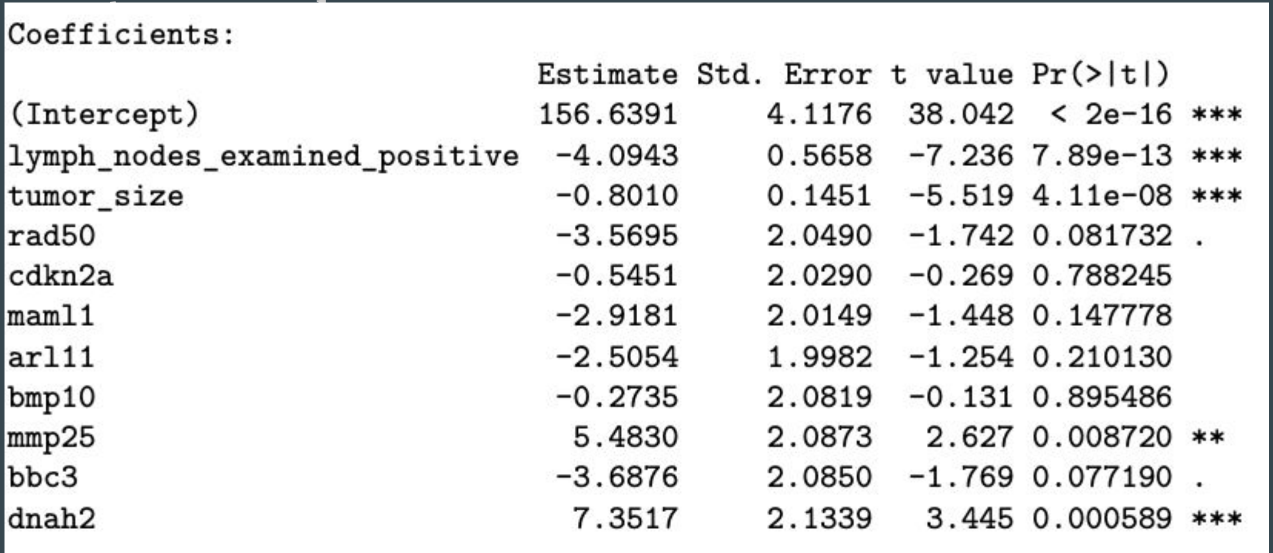
- AIC: 1440.3
- 22 coefficients with p-value < 0.05.
- 5 Largest ORs:
    - Radio therapy: 1.645868
    - cdkn2a: 1.533894
    - chek2: 1.508313
    - hsd3b7: 1.457797
    - rps6ka2: 1.405400

Results

- 5 smallest ORs:
  - ptprm: 0.7346325
  - nras: 0.7379200
  - bche: 0.7447645
  - dll3: 0.7831545
  - afdn: 0.8016692

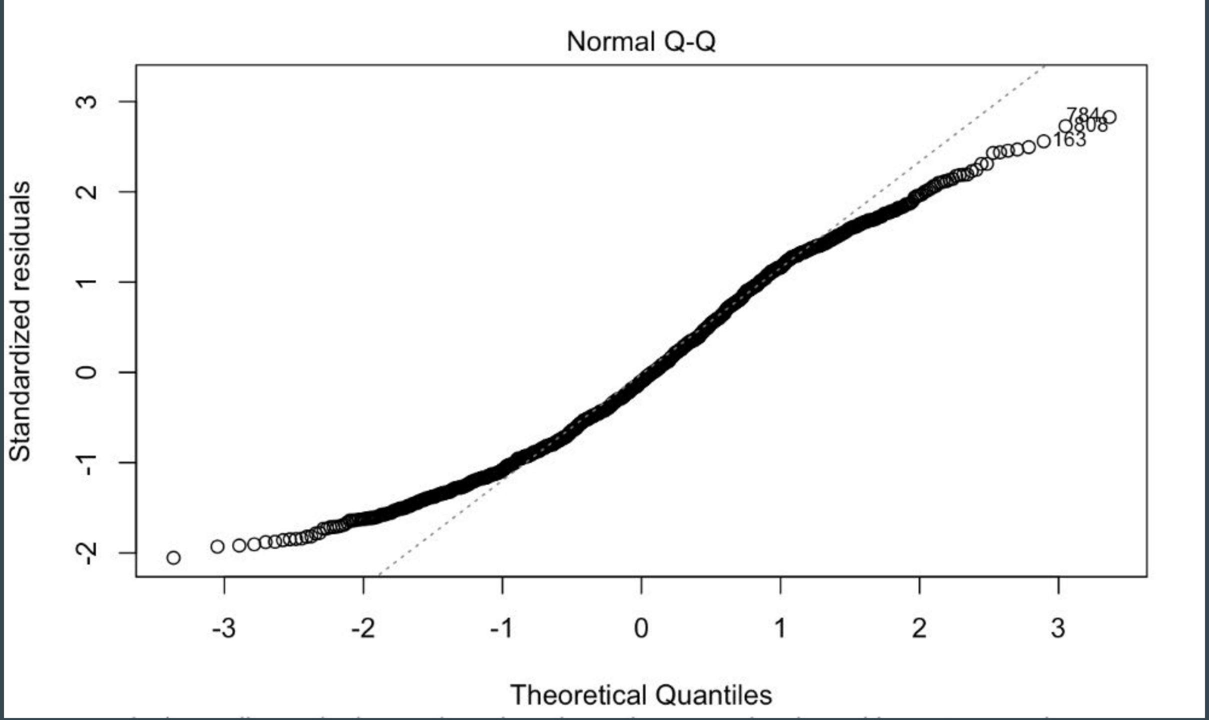
Linear regression

- Top 10 variables with p values < 0.05 from logistic regression.
- Total month of surviving
  - Top 5 variables:

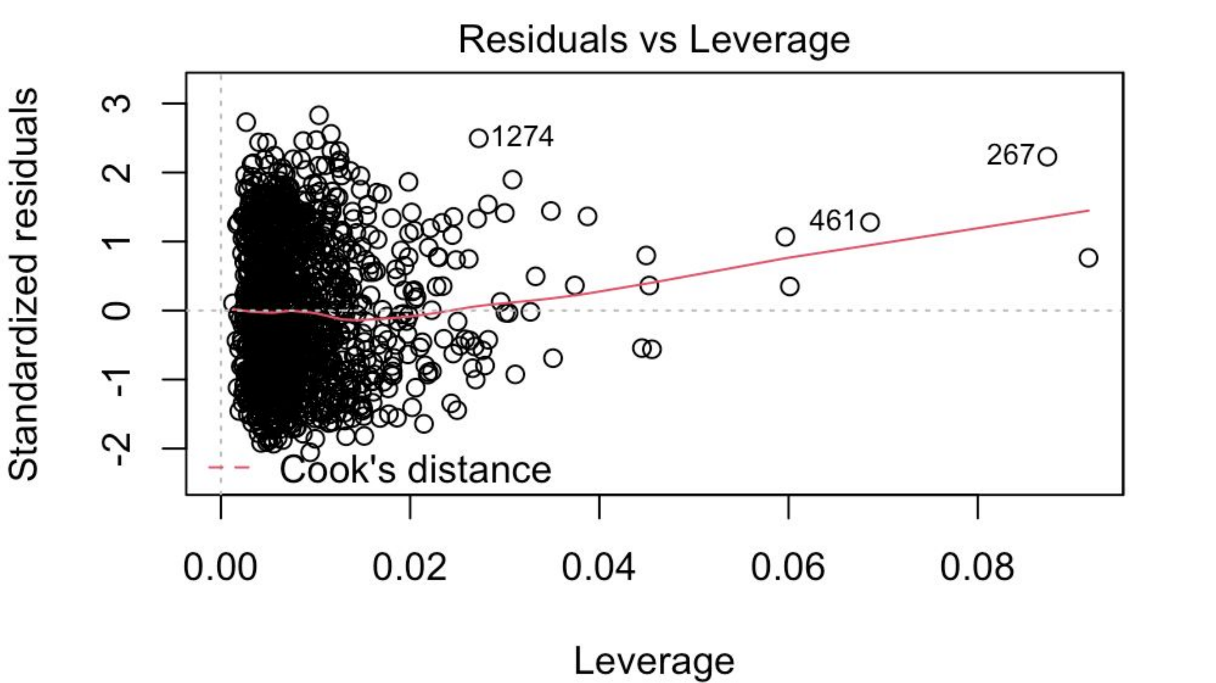


```
Coefficients:
                            Estimate Std. Error t value Pr(>|t|)    
(Intercept)                 156.6391     4.1176  38.042  < 2e-16 ***
lymph_nodes_examined_positive -4.0943     0.5658  -7.236 7.89e-13 ***
tumor_size                   -0.8010     0.1451  -5.519 4.11e-08 ***
rad50                        -3.5695     2.0490  -1.742 0.081732 .  
cdkn2a                       -0.5451     2.0290  -0.269 0.788245    
maml1                        -2.9181     2.0149  -1.448 0.147778    
arl11                        -2.5054     1.9982  -1.254 0.210130    
bmp10                        -0.2735     2.0819  -0.131 0.895486    
mmp25                         5.4830     2.0873   2.627 0.008720 ** 
bbc3                         -3.6876     2.0850  -1.769 0.077190 .  
dnah2                         7.3517     2.1339   3.445 0.000589 ***
```

QQ Plot


Normal Q-Q
Standardized residuals
Theoretical Quantiles
-2
-1
0
1
2
3
-3
-2
-1
0
1
2
3
784
808
163

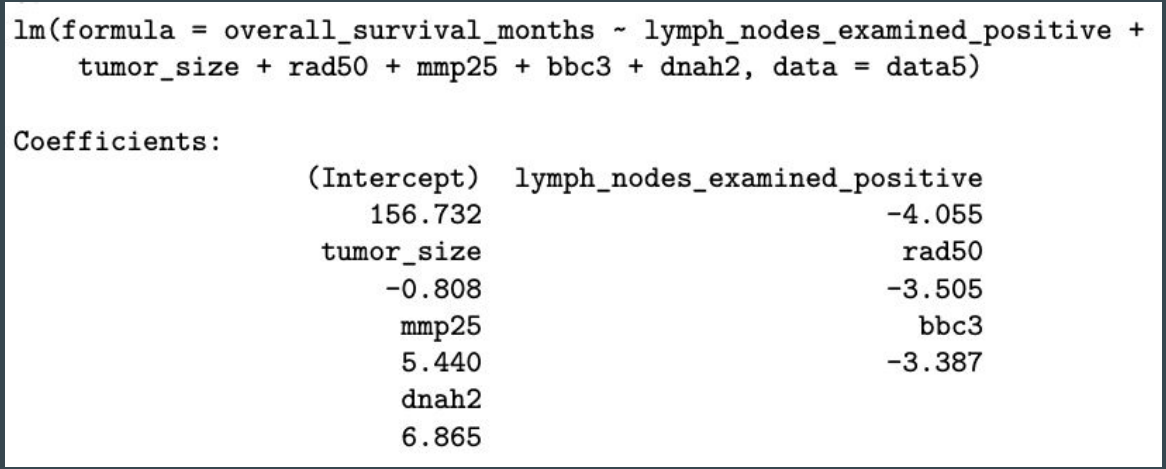
Leverage Plot



Stepwise AIC

Importance

- Dnah2
- Mmp25
- Tumor_size
- Bbc3
- Rad50
- lymph_nodes_examined_positive



```
lm(formula = overall_survival_months ~ lymph_nodes_examined_positive +
    tumor_size + rad50 + mmp25 + bbc3 + dnah2, data = data5)

Coefficients:
                  (Intercept)  lymph_nodes_examined_positive
                      156.732                         -4.055
                   tumor_size                          rad50
                       -0.808                         -3.505
                        mmp25                           bbc3
                        5.440                         -3.387
                        dnah2
                        6.865
```

Thank You