**###1. (15 marks).###1.(a) Compute the variance, co-variance and correlation matrix of your random subset of 10 genes. Add an appropriate table to your report.**

```
## 'data.frame':    78 obs. of  10 variables:
##  $ NM_015957     : num  -0.067 -0.04 -0.06 0.147 -0.094 -0.052 -0.168 -
0.014 -0.161 -0.066 ...
##  $ Contig23913_RC: num  -0.214 -0.231 -0.093 -0.378 0.352 0.222 0.303 -
0.067 0.083 0.01 ...
##  $ Contig49076_RC: num  0.163 0.158 0.153 0.134 0.03 -0.024 -0.004 0.025
0.145 0.096 ...
##  $ NM_003034     : num  0.096 -0.09 -0.047 -0.011 -0.406 -0.244 -0.218 -
0.081 -0.016 -0.593 ...
##  $ NM_012396     : num  -0.269 -0.097 -0.166 -0.366 0.054 -0.09 0.212 -
0.056 0.178 -0.012 ...
##  $ U81599        : num  0.478 -0.151 -0.131 0.01 -0.104 -0.162 -0.116
0.061 0.014 -0.165 ...
##  $ NM_004417     : num  0.346 0.088 -0.152 -0.1 -0.216 0.327 0.338 -0.468
-0.113 -0.391 ...
##  $ AL157424      : num  0.016 -0.113 -0.173 -0.02 -0.065 -0.156 0.056 -
0.112 -0.248 -0.03 ...
##  $ NM_005512     : num  0.009 -0.212 -0.074 -0.475 0.189 -0.027 0.26 -
0.343 0.14 0.059 ...
##  $ Contig2339_RC : num  -0.048 -0.25 -0.08 -0.259 0.167 0.1 -0.054 0.026
0.122 -0.175 ...
```

*Summary of variance of 10 random genes*

| | NM_015957 | Contig23913_RC | Contig49076_RC | NM_003034 | NM_012396 | U81599 | NM_004417 | AL157424 | NM_005512 | Contig2339_RC |
|---|---|---|---|---|---|---|---|---|---|---|
| NM_015957 | 0.0217779 | -0.0098271 | 0.0003438 | 0.0049571 | -0.0084132 | 0.00787 13 | 0.0030423 | 0.00437 62 | -0.0075341 | -0.0006098 |
| Contig23913_RC | -0.0098271 | 0.0436823 | -0.0008189 | 0.0033713 | 0.0093477 | 0.00164 21 | 0.0019970 | -0.00357 58 | 0.0202705 | 0.0056441 |
| Contig49076_RC | 0.0003438 | -0.0008189 | 0.0170669 | -0.0100928 | -0.0000574 | -0.00367 64 | 0.0063650 | -0.00077 71 | -0.0003844 | -0.0008061 |

| | NM_015957 | Contig23913_RC | Contig49076_RC | NM_003034 | NM_012396 | U81599 | NM_004417 | AL157424 | NM_005512 | Contig2339_RC |
|---|---|---|---|---|---|---|---|---|---|---|
| NM_003034 | 0.0049571 | 0.0033713 | -0.0100928 | 0.1159729 | -0.0065707 | 0.0323088 | 0.0256664 | -0.0025205 | 0.0024442 | 0.0109860 |
| NM_012396 | -0.0084132 | 0.0093477 | -0.0000574 | -0.0065707 | 0.0349794 | -0.0073287 | -0.0027807 | 0.0008163 | 0.0017045 | 0.0045276 |
| U81599 | 0.0078713 | 0.0016421 | -0.0036764 | 0.0323088 | -0.0073287 | 0.0548965 | 0.0048851 | 0.0049212 | 0.0057488 | 0.0001311 |
| NM_004417 | 0.0030423 | 0.0019970 | 0.0063650 | 0.0256664 | -0.0027807 | 0.0048851 | 0.0855729 | 0.0034991 | 0.0159548 | 0.0231816 |
| AL157424 | 0.0043762 | -0.0035758 | -0.0007771 | -0.0025205 | 0.0008163 | 0.0049212 | 0.0034991 | 0.0324661 | 0.0096259 | 0.0062037 |
| NM_005512 | -0.0075341 | 0.0202705 | -0.0003844 | 0.0024442 | 0.0170445 | 0.0057488 | 0.0159548 | 0.0096259 | 0.0463120 | 0.0141650 |
| Contig2339_RC | -0.0006098 | 0.0056441 | -0.0008061 | 0.0109860 | 0.0045276 | 0.0001311 | 0.0231816 | 0.0062037 | 0.0141650 | 0.0652259 |

*Summary of covariance of 10 random genes*

| | NM_015957 | Contig23913_RC | Contig49076_RC | NM_003034 | NM_012396 | U81599 | NM_004417 | AL157424 | NM_005512 | Contig2339_RC |
|---|---|---|---|---|---|---|---|---|---|---|
| NM_015957 | 0.0217779 | -0.0098271 | 0.0003438 | 0.0049571 | -0.0084132 | 0.0007813 | 0.0030423 | 0.0043762 | -0.0075341 | -0.0006098 |
| Contig23913_RC | -0.0098271 | 0.0436823 | -0.0008189 | 0.0033713 | 0.0093477 | 0.0016421 | 0.0019970 | -0.0035758 | 0.0202705 | 0.0056441 |
| Contig49076_RC | 0.0003438 | -0.0008189 | 0.0170669 | -0.0100928 | -0.0000574 | -0.0036764 | 0.0063650 | -0.0007771 | -0.0003844 | -0.0008061 |
| NM_00 | 0.004 | 0.0033 | - | 0.115 | - | 0.03 | 0.025 | - | 0.002 | 0.0109 |

| | NM_015957 | Contig23913_RC | Contig49076_RC | NM_003034 | NM_012396 | U81599 | NM_004417 | AL157424 | NM_005512 | Contig2339_RC |
|---|---|---|---|---|---|---|---|---|---|---|
| 3034 | 9571 | 713 | 0.0100928 | 9729 | 0.0065707 | 23088 | 6664 | 0.00252 05 | 4442 | 860 |
| NM_012396 | -0.0084132 | 0.0093477 | -0.0000574 | -0.0065707 | 0.0349794 | -0.00732 87 | -0.0027807 | 0.00081 63 | 0.0170445 | 0.0045276 |
| U81599 | 0.0078713 | 0.0016421 | -0.0036764 | 0.0323088 | -0.0073287 | 0.05489 65 | 0.0048851 | 0.00492 12 | 0.0057488 | 0.0001311 |
| NM_004417 | 0.0030423 | 0.0019970 | 0.0063650 | 0.0256664 | -0.0027807 | 0.00488 51 | 0.0855729 | 0.00349 91 | 0.0159548 | 0.0231816 |
| AL157424 | 0.0043762 | -0.0035758 | -0.0007771 | -0.0025205 | 0.0008163 | 0.00492 12 | 0.0034991 | 0.03246 61 | 0.0096259 | 0.0062037 |
| NM_005512 | -0.0075341 | 0.0202705 | -0.0003844 | 0.0024442 | 0.0170445 | 0.00574 88 | 0.0159548 | 0.00962 59 | 0.0463120 | 0.0141650 |
| Contig2339_RC | -0.0006098 | 0.0056441 | -0.0008061 | 0.0109860 | 0.0045276 | 0.00013 11 | 0.0231816 | 0.00620 37 | 0.0141650 | 0.0652259 |

*Summary of correlation of 10 random genes*

| | NM_015957 | Contig23913_RC | Contig49076_RC | NM_003034 | NM_012396 | U81599 | NM_004417 | AL157424 | NM_005512 | Contig2339_RC |
|---|---|---|---|---|---|---|---|---|---|---|
| NM_015957 | 1.0000000 | -0.3186132 | 0.0178347 | 0.0986384 | -0.3048232 | 0.22765 06 | 0.0704738 | 0.16457 98 | -0.2372339 | -0.0161809 |
| Contig23913_RC | -0.3186132 | 1.0000000 | -0.0299919 | 0.0473655 | 0.2391357 | 0.03353 24 | 0.0326632 | -0.09495 22 | 0.4506770 | 0.1057381 |
| Contig49076_RC | 0.0178347 | -0.0299919 | 1.0000000 | -0.2268583 | -0.0023499 | -0.12010 70 | 0.1665527 | -0.03301 27 | -0.0136739 | -0.0241597 |
| NM_00 | 0.098 | 0.0473 | - | 1.000 | - | 0.40 | 0.257 | - | 0.033 | 0.1263 |

| | NM_015957 | Contig23913_RC | Contig49076_RC | NM_003034 | NM_012396 | U81599 | NM_004417 | AL157424 | NM_005512 | Contig2339_RC |
|---|---|---|---|---|---|---|---|---|---|---|
| 3034 | 6384 | 655 | 0.2268583 | 0000 | 0.1031645 | 49213 | 6428 | 0.0410763 | 3509 | 139 |
| NM_012396 | -0.3048232 | 0.2391357 | -0.0023499 | -0.1031645 | 1.0000000 | -0.1672422 | -0.0508260 | 0.0242222 | 0.4234789 | 0.0947878 |
| U81599 | 0.2276506 | 0.0335324 | -0.1201070 | 0.4049213 | -0.1672422 | 1.0000000 | 0.0712745 | 0.1165700 | 0.1140136 | 0.0021904 |
| NM_004417 | 0.0704738 | 0.0326632 | 0.1665527 | 0.2576428 | -0.0508260 | 0.0712745 | 1.0000000 | 0.0663854 | 0.2534414 | 0.3102885 |
| AL157424 | 0.1645798 | -0.0949522 | -0.0330127 | -0.0410763 | 0.0242222 | 0.1165700 | 0.0663854 | 1.0000000 | 0.2482449 | 0.1348106 |
| NM_005512 | -0.2372339 | 0.4506770 | -0.0136739 | 0.0333509 | 0.4234789 | 0.1140136 | 0.2534414 | 0.2482449 | 1.0000000 | 0.2577270 |
| Contig2339_RC | -0.0161809 | 0.1057381 | -0.0241597 | 0.1263139 | 0.0947878 | 0.0021904 | 0.3102885 | 0.1348106 | 0.2577270 | 1.0000000 |

The Variance matrix shows the variance between each gene helping to further research the biological combination between the different gene expressions. The variance value nearing zero shows significantly less variability.

The Covariance summary table shows the co-variability between two genes. Positive values refer that the pair genes are directly proportional. Negative values refer that the pair genes are inversely proportional. Zero values refer to relationships that tend to be zero or no relationship.

The Correlation summary table shows the strength of the linear relationship between two genes. The positive values nearing 1 indicate the positive correlation between the two genes. The negative values near -1 indicate a negative correlation. The values nearing to zero have minimal correlation or no relationship between the genes. The table shows the mixed expression level of correlation between various pair of gene.

**###1.(b) Using R to calculate the distance matrix of your random subset of 10 genes. Add an appropriate table to your report.**

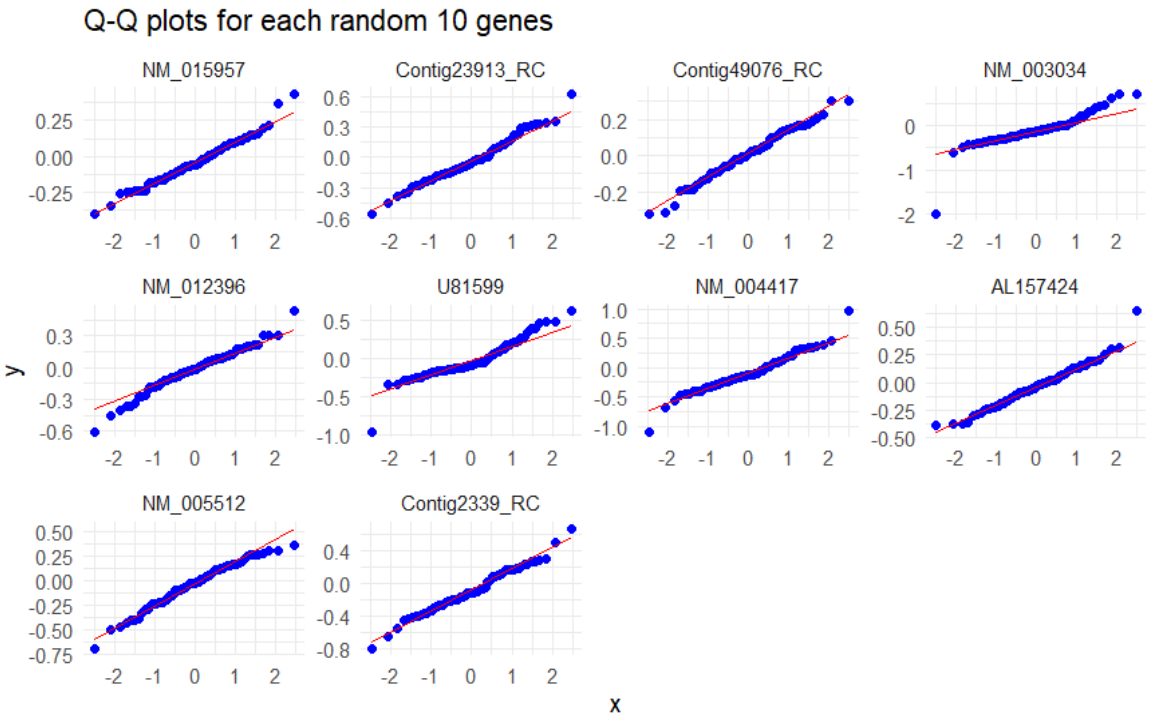
```
##           1        2        3        4        5        6        7
## 2  3.370944
## 3  3.426626 1.501139
## 4  3.902634 2.627362 3.175058
## 5  5.003077 4.209149 3.288939 5.886191
## 6  4.130907 3.192942 2.803918 4.870037 2.509188
## 7  4.850160 4.326752 4.092801 6.420849 2.508071 2.640296
## 8  4.181069 2.775190 2.261921 3.225263 3.645984 3.601745 4.844428
## 9  4.196262 3.387470 2.600202 5.448356 2.425005 2.976556 3.018795
## 10 4.651734 2.922856 2.349936 4.464692 2.452840 3.307339 3.640707
## 11 2.953858 2.226232 2.528990 3.860785 3.644734 2.338351 3.224704
## 12 4.406467 3.869780 3.334768 4.744192 3.591803 3.312076 3.871972
## 13 3.813948 3.327352 2.387039 4.507019 2.258460 2.907160 3.460316
## 14 4.093492 2.875545 2.473297 4.647108 2.275410 2.890386 2.649771
## 15 5.061616 2.411279 3.008690 3.789764 4.270927 3.885208 4.451399
## 16 4.806548 4.410721 3.715049 5.936159 2.962766 3.917193 3.697995
## 17 3.756859 3.771221 3.308198 5.793284 2.679401 3.243977 2.422202
## 18 3.949778 1.937191 2.252836 4.090679 3.329226 2.738793 3.037551
## 19 2.886815 3.533816 3.443580 5.262235 3.661063 3.585899 2.829810
## 20 5.565282 5.262685 5.265670 6.220054 5.530833 5.211583 5.540499
## 21 4.563513 3.199072 2.680618 4.863760 2.946509 3.918987 3.728556
## 22 4.181190 2.843058 2.609722 4.022379 2.782250 2.825798 3.344989
## 23 5.765010 4.023342 3.115429 6.054062 2.865402 3.397342 3.909940
## 24 4.180389 4.643158 4.615675 4.544745 5.709131 5.215692 5.687149
## 25 4.118420 2.590934 2.263187 3.819961 3.590024 2.362749 4.296975
## 26 4.179331 2.809395 2.293339 4.454267 2.396663 2.344139 2.651417
## 27 4.320908 3.258343 2.476380 5.300897 1.562495 1.993664 2.336141
## 28 4.462841 2.500601 2.936435 3.903077 4.490235 3.097616 4.537843
## 29 5.762669 4.995579 4.633465 7.064349 3.372618 3.744031 2.667324
## 30 6.009460 4.980556 4.072287 6.407733 2.212535 3.739219 3.523330
## 31 3.391930 1.498996 1.574126 3.551359 3.381544 2.638136 3.336878
## 32 3.861053 2.216835 2.661360 2.782449 4.920037 3.799532 5.101484
## 33 4.212009 1.972711 2.043845 3.879898 4.284102 3.868130 4.636400
## 34 4.934870 2.602078 3.373138 2.896076 5.142270 4.408699 5.319146
## 35 4.192958 4.445567 4.462382 4.592450 4.900735 5.494850 5.358581
## 36 4.344946 3.568373 2.659517 5.018822 2.182244 3.122434 3.319493
## 37 8.254187 6.788767 6.443914 7.766743 6.619177 7.201206 7.172454
## 38 4.796340 3.115025 3.455893 4.474537 4.461364 3.768543 4.291225
## 39 4.335133 4.945169 4.627133 5.747399 3.741979 4.696561 4.061767
## 40 3.046057 3.210047 3.233047 3.959753 4.133698 2.885298 4.087972
## 41 3.967017 2.746626 2.615017 2.945413 3.996625 3.563732 4.836018
## 42 4.639261 3.502910 3.159376 5.285955 2.268149 2.361142 2.213149
## 43 3.715517 2.789673 3.018775 4.170971 3.732406 3.586446 3.495992
## 44 6.584060 5.634228 5.689913 5.494626 5.963149 5.743194 6.153054
## 45 4.556705 3.466454 2.783075 4.922718 2.422528 2.900122 3.243318
## 46 5.175452 4.517219 3.797625 6.600894 1.777264 2.907288 1.849073
## 47 5.176337 4.562981 4.116670 5.884431 3.268467 3.747410 3.687237
```

```
## 48  4.802483  4.366479  3.794686  4.521191  4.225785  3.712252  5.255866
## 49  4.777783  3.608912  3.327013  3.141533  4.782216  4.382608  5.688579
## 50  4.341675  4.125990  3.651222  4.041829  4.606575  3.944897  5.352538
## 51  3.767987  3.037459  3.014739  4.970256  3.099799  1.908789  2.109103
## 52  5.289347  3.201662  2.911117  4.133576  4.083044  4.087806  4.973377
## 53  4.222814  4.347734  4.260596  5.330002  4.752525  4.102076  4.261583
## 54 10.204154  7.684294  7.984605  8.010937  8.699363  8.685312  9.579764
## 55  4.922834  2.796092  3.114054  3.709842  5.404444  4.862802  5.909426
## 56  3.317897  4.093218  3.910574  3.424944  5.165167  4.670960  5.850708
## 57  4.609092  3.703960  3.320547  3.445723  5.447679  4.330579  6.342101
## 58  3.153347  3.490673  3.517125  4.498625  4.629703  4.593762  4.661392
## 59  4.615829  3.189507  3.311491  4.271172  3.508587  3.656136  3.618331
## 60  4.164436  4.035372  3.390601  5.284753  2.325493  2.674249  3.288850
## 61  4.853858  2.738597  3.107838  4.443953  4.399977  4.595045  4.382258
## 62  2.966948  2.811422  2.890436  3.134132  4.246764  3.880409  4.628756
## 63  4.281773  5.242880  5.110542  4.529606  5.734275  5.137163  6.320914
## 64  3.605379  2.930980  2.770477  3.225622  5.076262  4.410445  5.466850
## 65  3.481614  3.733474  3.804177  4.203107  5.135740  3.812353  4.801782
## 66  4.544117  3.747441  3.621458  2.750351  5.796171  4.825924  6.645445
## 67  4.711811  3.901528  3.653147  4.982576  3.725418  3.596441  3.715045
## 68  4.413075  5.059231  4.436571  6.104049  3.325508  3.322473  3.662482
## 69  5.201499  4.386699  4.268136  6.084916  3.638678  4.005909  3.029650
## 70  2.656887  2.262097  2.648193  3.866435  3.918400  3.099224  3.493448
## 71  4.653395  3.458818  3.117816  4.392379  4.306089  3.300388  4.835350
## 72  3.901109  3.995196  3.328451  5.188785  2.540472  2.993046  3.276818
## 73  5.093311  4.770116  4.851930  6.576069  4.673829  2.813008  3.397224
## 74  4.009203  4.134561  4.128780  4.816895  5.092593  4.340272  4.515420
## 75  4.440872  2.803079  2.190770  4.709500  2.831812  3.361211  3.523199
## 76  5.093585  4.216437  3.718286  6.302756  2.411052  3.170383  1.911778
## 77  5.658084  5.028936  5.156068  5.562831  4.870196  5.430068  4.845067
## 78  5.443919  3.612064  3.211788  4.912953  3.027470  3.284813  3.927177
```

Note: The output of the distance matrix runs several lines. A minimal lines are shown above.

A distance matrix is the distance matrix calculating the distance between each gene pair based on similarity and dissimilarity between the genes. This matrix refers to a quantitative measure of dissimilarity between pairs of genes. Thus, a higher value indicates greater dissimilarity. This matrix depicts the fundamental step to analyze the relationship between genes and gene expression data.

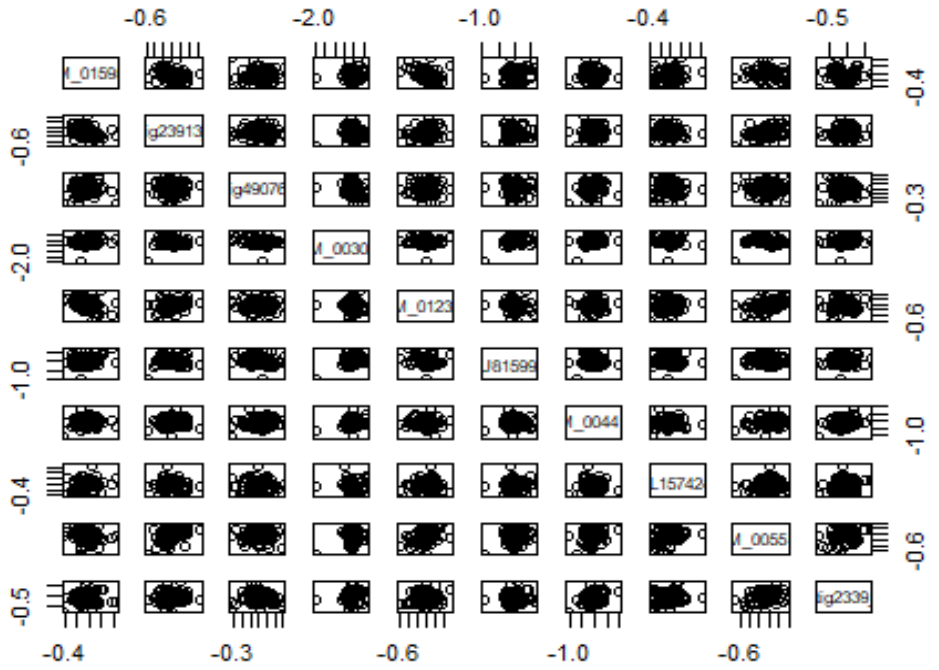
**###(c) Using R to calculate univariate Q-Q-plots and a Q-Q-plot based on the generalised distance for the observations of your random subset of 10 genes. Add appropriate figures to your report.**

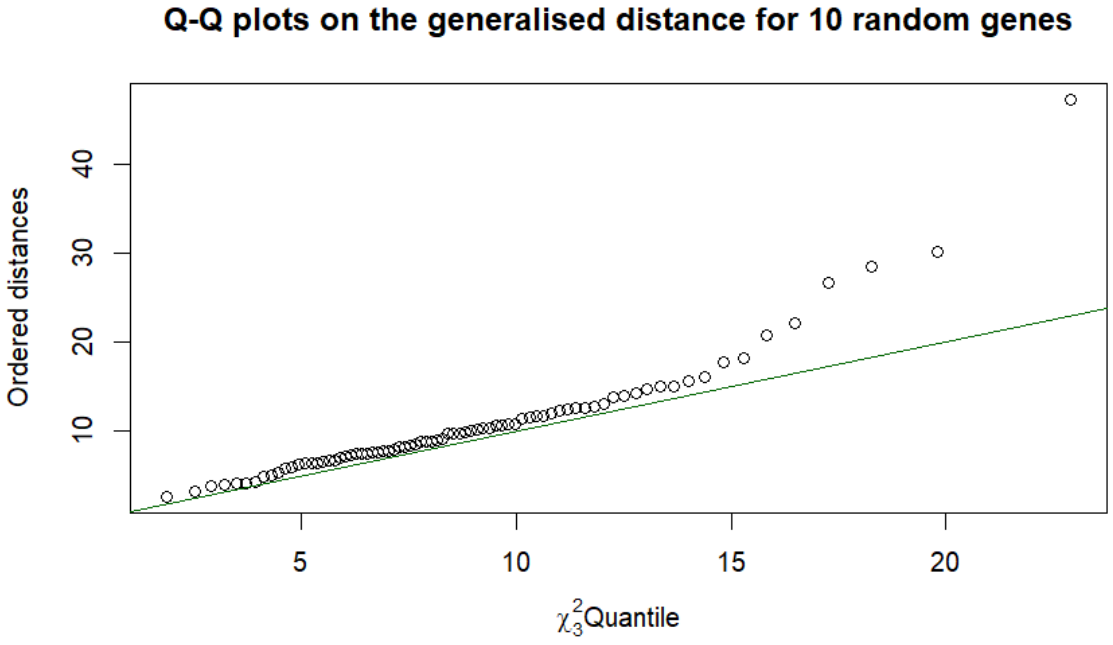



The univariate Q-Q plots are generated for a random subset of 10 genes. These plots depict the quantities of the observed data to the theoretical normal distribution. Each plot represent each gene of a random subset of 10 genes under the variable name my.gene.subset.

The blue points represent the observed quantities and the red line represents quantities of the normal distribution. When the blue points fall closely to the red line, inferring that the data nearly follows the normal distribution.While for some genes the blue points are deviated from the red line representing a high-tailed distribution.

This Q-Q plot graph helps to infer the normality assumptions and further statistical methods to apply to further research.


-0.6
-2.0
-1.0
-0.4
-0.5
f_0159
g23913
g49076
f_0030
f_0123
J81599
f_0044
L15742
f_0055
g2339
-0.4
-0.3
-0.6
-1.0
-0.6


Q-Q plots on the generalised distance for 10 random genes
Ordered distances
10
20
30
40
5
10
15
20
$\chi^2_3$Quantile

The pair plot graph shows the significant pairwise relationships between the randomly selected 10 genes. This gives us a clear picture to identify the patterns and outliers in the existing data structure.

The Mahalanobis distance is used to calculate the generalized distance here. The corresponding observations, mean values and covariance of the dataset are used. The Q-Q plot helps to assess the distribution using this Mahalanobis distance.

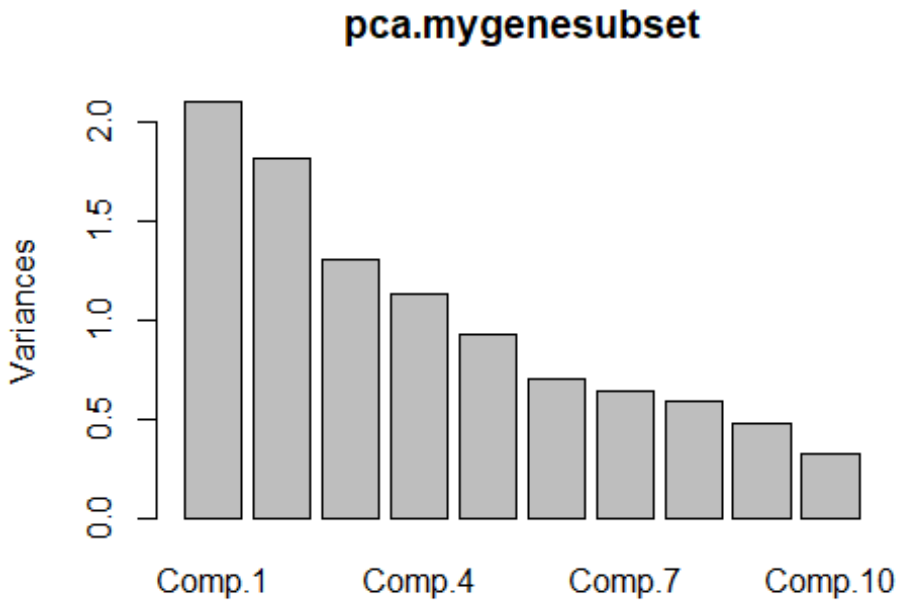
This plot compares the observed Mahalanobis distance to the chi-square distribution, which represents the degree of freedom as 9(10 random gene variable -1). The green line represents the reference line showing the exact match point of observed and expected quantiles. The observed Mahalanobis distance, mostly closely on the green line (reference line), depicted that they are chi-square distributed, further indicating a multivariate normality of the data set containing 10 random genes.

The few observations falling apart from the green line indicate that there are few outliers in the dataset. However, this doesn't affect the data much. This helps to evaluate the multivariate assumptions of observations in the dataset and any potential data anomalies before proceeding observations with further analysis.

**###2. (15 marks). Use R for a principal component analysis of your random subset of 10 genes. Add appropriate tables and figures to your report.**

```
## Importance of components:
##                           Comp.1    Comp.2    Comp.3    Comp.4     Comp.5
## Standard deviation     1.4480969 1.3470669 1.1428088 1.0608945 0.96007165
## Proportion of Variance 0.2096985 0.1814589 0.1306012 0.1125497 0.09217376
## Cumulative Proportion  0.2096985 0.3911574 0.5217586 0.6343083 0.72648206
##                            Comp.6     Comp.7     Comp.8     Comp.9
Comp.10
## Standard deviation     0.83796135 0.80171080 0.76479687 0.69242320
0.57087291
## Proportion of Variance 0.07021792 0.06427402 0.05849142 0.04794499
0.03258959
## Cumulative Proportion  0.79669998 0.86097400 0.91946542 0.96741041
1.00000000
##
## Loadings:
##                Comp.1 Comp.2 Comp.3 Comp.4 Comp.5 Comp.6 Comp.7 Comp.8
Comp.9
## NM_015957       0.366  0.337  0.224  0.186  0.103  0.028  0.553  0.572
0.128
## Contig23913_RC -0.463 -0.044 -0.270 -0.144  0.291 -0.531  0.129  0.293
0.345
## Contig49076_RC  0.008 -0.150  0.562 -0.335  0.565  0.076  0.144 -0.347
0.269
```

```
## NM_003034      -0.025  0.515 -0.351 -0.255 -0.077  0.328 -0.121 -0.119
0.588
## NM_012396      -0.444 -0.223 -0.027  0.197 -0.040  0.690  0.346  0.033
0.069
## U81599          0.038  0.503 -0.288  0.112  0.461  0.009  0.192 -0.377 -
0.435
## NM_004417      -0.191  0.382  0.388 -0.424 -0.053  0.166 -0.378  0.357 -
0.259
## AL157424       -0.086  0.236  0.340  0.701  0.046 -0.102 -0.375 -0.133
0.341
## NM_005512      -0.569  0.133  0.061  0.186  0.201  0.011 -0.044  0.167 -
0.262
## Contig2339_RC  -0.290  0.274  0.290 -0.101 -0.566 -0.299  0.448 -0.369 -
0.011
##                Comp.10
## NM_015957       0.082
## Contig23913_RC -0.318
## Contig49076_RC  0.105
## NM_003034       0.240
## NM_012396      -0.332
## U81599         -0.265
## NM_004417      -0.345
## AL157424       -0.207
## NM_005512       0.694
## Contig2339_RC  -0.007
```



Principal Component Analysis is generally used to reduce the complex dimensions and identify the high dimensions of the dataset. The PCA is performed on the randomly selected 10 genes consisting of two main tasks in the analysis; importance of components and loading of each gene on the principal components.

The importance of the Components table shows the quantitative measure of standard deviation, Proportion of variance and the cumulative proportion of each component.

The standard deviation is calculated by the square root of eigenvalues of the covariance matrix.

The proportion of variance is the square of the standard deviation divided by the sum of squares of all standard deviations.

The cumulative proportion of variance is the cumulative sum of proportions of variance explained by each principal component.

Loadings refers to the contribution of each gene to principal component. High value depicts the high association between the gene and the respective principal component.

In principal component 1, there exists the highest value in standard deviation, Proportion of Variance and Cumulative Proportion showing the high variability in the data. And the trend is decreasing in further component proceedings, which clearly indicates less variability in the data. Each component gives valuable information for analyzing the overall structure of this data set.

While analyzing the loadings in component 1, genes Contig11075_RC, NM_004358, and L36069 contribute significant variations in Component 1. Similarly, in Component 2, Contig54232_RC, NM_004056, Contig11075_RC and NM_001218 genes have higher values. The loading table illustrates the relationship between the genes and the principal components. The high loading values of genes represents high correlation and negative values are negatively correlated.

This PCA model gives us the underlying structure of this gene expression data. The repeated pattern of loading of genes to the components and analyzing the importance of components give us the key pattern and relationship between the genes. The graph clearly shows the variance in the data that has decreased with increasing components.

**###3. (15 marks).**

**###3.(a) Fit a multivariate analysis of variance model (MANOVA) to your random subset of 10 genes. Investigate if there is a difference between invasive (label 1) and noninvasive (label 2) cancer. Note: You need to add column 4949 containing the information invasive and noninvasive cancer to your random subset of 10 genes. Add appropriate tables and figures to your report.**

```
##  [1] 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2
## [39] 2 2 2 2 2 2 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
```

```
1 1 1
## [77] 1 1

##                       Df  Pillai approx F num Df den Df  Pr(>F)
## (Intercept)            1 0.26760  2.44806     10     67 0.01479 *
## my.gene.subset$class   1 0.07814  0.56792     10     67 0.83421
## Residuals             76
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The MANOVA fit model helps to investigate whether there is a significant difference in the expression levels of 10 random genes between invasive and non-invasive cancer.

In MANOVA, intercepts represent the estimated intercepts of the regression equation for each dependent variable. Similar to the intercepts in linear regression models, the position where the expected variable of each dependent variable when all other independent variables tend to zero, intercepts are formed.

F-Statistics are determined to test the overall significance of the model. In the overall MANOVA test, F-statistics are calculated by comparing the variability between groups to the variability within groups.

The degree of freedom is associated with the number of groups, the number of dependent variables and parameters estimated in the MANOVA model. A multivariate test statistic termed Pillai trace is used in MANOVA to evaluate the overall significance of the model or individual effects. Ranges from 0 to 1 with larger values depicting greater discrimination between the groups. The p-value indicates the probability of observing the data with test statistics. Here, the smaller the p-value depicts that there are significant differences between the groups or effect in the respective model. Pr's value, otherwise known as the Probability Ratio, is another measure of significance. The ratio of the observed value of a test statistic to its expected value in the null hypothesis. If a Pr value close to 1 indicates a prediction of a null hypothesis is approved when the Pr values differ far from 1, the most likely null hypothesis is rejected.

The intercept of Pillai traces 0.54587, indicates a moderate-to-large effect size. The approximate F-statistic is 8.0534 and the degree of freedom is 10 and 67 for the numerator and denominator, respectively. The associated p-value (Pr(>F)) is highly significant ($p < 0.001$), showing that there is a significant overall effect when considering all variables.

The intercept of Pillai traces 0.17394 for variable class, which indicates smaller effect size when compared to the intercept.

The approximate F-statistic is 1.4108 and the degree of freedom is 10 and 67 for the numerator and denominator, respectively. The associated p-value (Pr(>F)) is 0.1948, which is not significant ($p > 0.05$). This suggests that there is no significant difference in gene expression levels of 10 random genes between invasive and noninvasive cancer after controlling for other variables.

The residual 76 shows the unexplained variances after counting on the intercepts.

To conclude, the MANOVA results show there is a significant impact while considering all variables, while specific variable-class (invasive vs. noninvasive cancer) does not significantly impact the differences in gene expression levels.

Therefore, the H0 null Hypothesis is not rejected, concluding that there is no significant difference in the expression levels of the random subset of 10 genes between invasive (label 1) and noninvasive (label 2) cancer.


The above box plot shows the interquartile range of gene expression for each cancer type (invasive and noninvasive).

The Q1 and Q3 are the top and bottom edges of the box and the middle line(line inside the box) represents the median value of the gene expression. The data points scattered outside the box are termed as outliers. Label 1 represents noninvasive cancer and Label 2 represents invasive cancer.

From the MANOVA table:

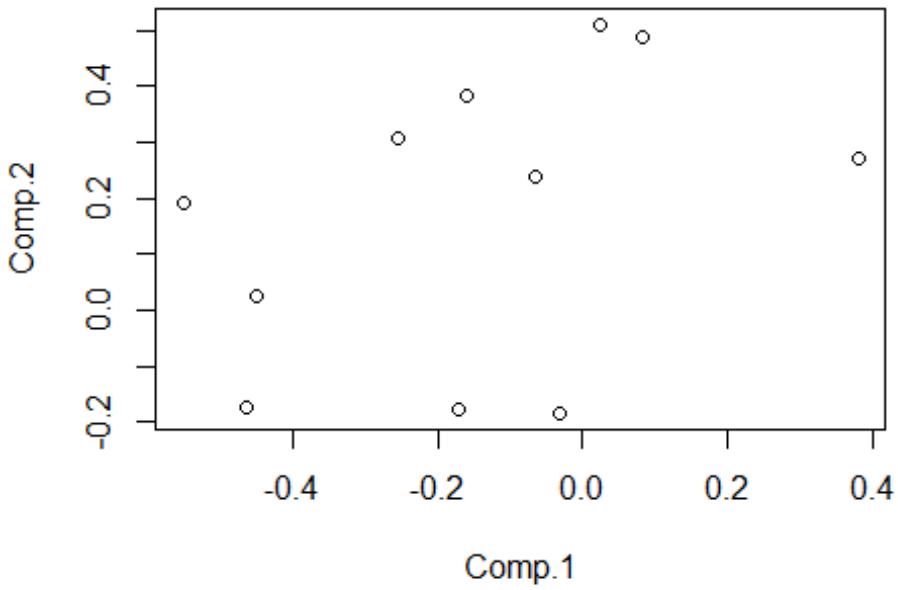
For the intercept (overall gene expression levels), the Pillai's trace statistic is 0.54587, indicating a significant difference between the two groups ($p < 0.001$).

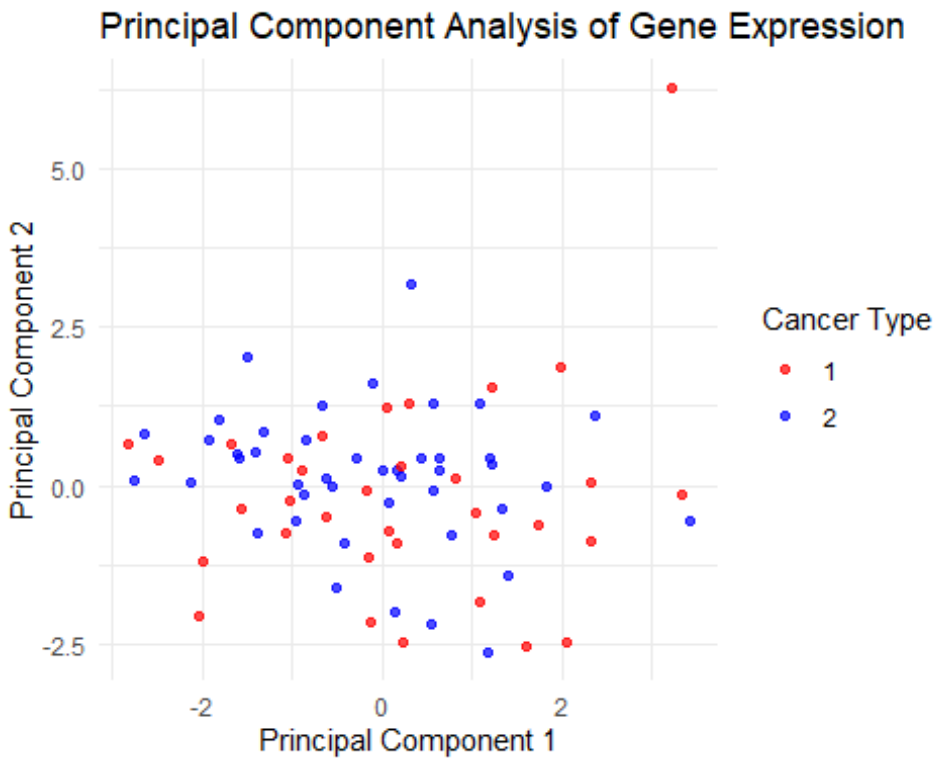
For the factor my.gene.subset$class (cancer type), the Pillai's trace statistic is 0.17394, indicating a non-significant difference between the two groups ($p = 0.1948$).

The box plot shows substantial overlap in label 1 and label 2 which clearly shows that there is o significant difference in the expression levels of the random subset of 10 genes between invasive (label 1) and noninvasive (label 2) cancer. Hence, H0 Null hypothesis is accepted.

**###3.(b) Use the first and second principal component to illustrate, if there is a difference between invasive and noninvasive cancer. Add appropriate tables and figures to your report.**

```
##                      Comp.1      Comp.2
## NM_015957       0.38109481  0.27144942
## Contig23913_RC -0.45068214  0.02491045
## Contig49076_RC -0.03193337 -0.18453938
## NM_003034       0.02387999  0.50967008
## NM_012396      -0.46376064 -0.17307506
## U81599          0.08265174  0.48641437
## NM_004417      -0.15975281  0.38380452
## AL157424       -0.06462211  0.23647308
## NM_005512      -0.55125016  0.19267420
## Contig2339_RC  -0.25570444  0.30590978
## class          -0.17087613 -0.17718019
```






In order to compare the first and second principal components, we are extracting the coefficients of the first principal component (PC1) and second principal component (PC2). These coefficients are the contributions of each variable to their respective principal component. The positive coefficient represents positive correlation and the negative coefficient represents negative correlation.

The plot graph of coefficients of PC1 and PC2 helps us to visualize the difference between the two groups of invasive and noninvasive cancer types. PCA is performed on the extracted PC1 and PC2, which is a subset of genes excluding the class variables, and of ggplots are used to give a clear visualization of differences between the two groups, invasive and noninvasive cancer types.

The inferences from the above analysis is that the PCA coefficient showing the correlation between each gene and PC1 and PC2 and insights in the gene contribute more variation in PC1 and PC2. The graph of the coefficients of the first two principal components gives the relative contribution of each gene to overall variation in the dataset, which is more of a scattered plot without any overlaps.

However, through the PC1 and PC2 of gene expression in a reduced dimensional-space, we see there are significant clustered plot patterns providing insights into the potential discriminatory power of the selected subset (differentiated by red and blue points in the graph) of genes for distinguishing between invasive and noninvasive cancer types.

**###4. (30 marks).**

**###4.(a) Apply LDA to your random subset of 10 genes and the class variable (invasive (label 1) and noninvasive (label 2) cancer). Calculate a confusion matrix, sensitivity, specificity and misclassification error. Add appropriate tables and figures to your report.**

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction 1 2
##          1 2 7
##          2 8 6
##
##                Accuracy : 0.3478
##                  95% CI : (0.1638, 0.5727)
##     No Information Rate : 0.5652
##     P-Value [Acc > NIR] : 0.9896
##
##                   Kappa : -0.3424
##
##  Mcnemar's Test P-Value : 1.0000
##
##             Sensitivity : 0.20000
##             Specificity : 0.46154
##          Pos Pred Value : 0.22222
##          Neg Pred Value : 0.42857
##              Prevalence : 0.43478
##          Detection Rate : 0.08696
##    Detection Prevalence : 0.39130
##       Balanced Accuracy : 0.33077
##
##        'Positive' Class : 1
##
```

```
## <NA>
##   NA
```

```
## <NA>
##   NA
```

```
##  Accuracy
## 0.6521739
```

The above output of the Linear Discriminant Analysis (LDA) helps to analyze the performance to classify the cancer types as invasive and noninvasive. LDA is applied to a subset of 10 random genes with class variables indicating invasive and noninvasive cancer types.

The data has been split into training represented as Data.1 (70 percentage of the data) and testing as Data.2(remaining 30 percentage of the data). LDA is applied to Data.1 training dataset with class variables as a response and predictors as gene expression features. The trained LDA model is used to predict the class labels of the testing set (data.2) of the data. In order to analyze the performance of the LDA model, a confusion matrix is computed.

The confusion matrix shows the numeric data of the performance of the LDA model as follows:

Accuracy is 34.78 percent, which is considerably low, referring to a low performance model. The 95 percent confidence interval for accuracy is between 16.38 percent to 57.27 percent, representing a poor model's predictive capabilities. The kappa statistic is -0.3424, indicating poor agreement beyond the chance of current and actual class. This show's model's performance is not significant. The sensitivity (true positive rate) is 20 percent, which is a measure of the proportion of actual invasive cases identified by the model. The specificity is 46.15 percent, which is a measure of the proportion of actual noninvasive identified by the model. The misclassification error is miscalculated instances which computes to 65.21 percent.

To conclude, the above LDA model trained has poor performance with the testing data. Further classification accuracy has to be improved by exploring more alternative classification algorithms.

**###4.(b) Apply Quadratic discriminant analysis (QDA) to your random subset of 10 genes and the class variable (invasive (label 1) and noninvasive (label 2) cancer). Calculate a confusion matrix, sensitivity, specificity and misclassification error. Add appropriate tables and figures to your report.**

```
## Call:
## qda(class ~ NM_015957 + Contig23913_RC + Contig49076_RC + NM_003034 +
##     NM_012396 + U81599 + NM_004417 + AL157424 + NM_005512 + Contig2339_RC,
##     data = my.gene.subset)
##
## Prior probabilities of groups:
##         1         2
## 0.4358974 0.5641026
##
## Group means:
##     NM_015957 Contig23913_RC Contig49076_RC   NM_003034   NM_012396
U81599
## 1 -0.03179412    -0.03700000    -0.01829412 -0.07973529 -0.05770588 -
```

```
0.00550000
## 2 -0.05097727    -0.02204545     0.02965909 -0.12490909  0.01156818 -
0.04068182
##     NM_004417    AL157424   NM_005512 Contig2339_RC
## 1 -0.08888235 -0.02911765 -0.06652941   -0.07514706
## 2 -0.09161364 -0.03984091 -0.01652273   -0.09272727

##
## Predict.qdaclass  1  2
##                1 24  4
##                2 10 40

## Confusion Matrix and Statistics
##
##           Reference
## Prediction  1  2
##          1 24  4
##          2 10 40
##
##                Accuracy : 0.8205
##                  95% CI : (0.7172, 0.8983)
##     No Information Rate : 0.5641
##     P-Value [Acc > NIR] : 1.532e-06
##
##                   Kappa : 0.6276
##
##  Mcnemar's Test P-Value : 0.1814
##
##             Sensitivity : 0.7059
##             Specificity : 0.9091
##          Pos Pred Value : 0.8571
##          Neg Pred Value : 0.8000
##              Prevalence : 0.4359
##          Detection Rate : 0.3077
##    Detection Prevalence : 0.3590
##       Balanced Accuracy : 0.8075
##
##        'Positive' Class : 1
##

## Sensitivity
##   0.7058824

## Specificity
##   0.9090909

##  Accuracy
## 0.1794872
```

The above Quadratic Discriminant Analysis (QDA) helps to develop a predictive model to classify the invasive and noninvasive cancer type based on the gene expression data. A subset of gene expression data is created and converted to the matrix 'x'. The prior probabilities of the group and group means are calculated. The QDA fitting is performed with the predictor variables specified as gene expressions, and the response variable as 'class'. The classes are predicted from the QDA fitted model.

A contingency table is created to compare the predicted classes with actual classes. A confusion matrix is generated to evaluate the performance of the QDA predictions. The classification is as follows:

24 invasive cancer cases as True Positives, 40 non-invasive cancer cases as True Negatives, 4 non-invasive cancer cases as False Positives, 10 invasive cancer cases as False Negatives.

The QDA model accuracy computes to 82.05 percent, which is significantly a high performance of the model. The 95 percent confidence interval for accuracy is between 71.72 percent to 89.73 percent, representing a decent model's predictive capabilities. The kappa statistic is 0.6276, which shows agreement between predicted and actual classes beyond chance turns to be moderate. Sensitivity for this QDA model is 70.59 percent, which determines the true positives of invasive cancer cases. Specificity identified by this QDA model is 90.91 percent of noninvasive cancer cases, which determines the ability to avoid false alarm in noninvasive cases. The calculated misclassification error of the overall QDA model is 17.95 percent, which is the proportion of incorrectly classified cases.

In order to infer from this QDA model, there is a promising accurate classification of cancer types based on gene expression data. High performance, sensitivity and specificity values are remarkably high, depicting good performance in detecting invasive and noninvasive cancer cases.

**###4.(c) Discuss the difference between LDA and QDA using the results on your random subset of 10 genes and the class variable (invasive (label 1) and noninvasive (label 2) cancer).**

The LDA model is based on the assumption that the predictors have a multivariate normal distribution within each class and the covariance matrix across classes is equal. The QDA model proceeds with no assumption, allowing each class to have its own covariance matrix.

With respect to the accuracy of the overall model, the LDA model shows a low accuracy level of 34.78 percent comparatively on the other hand, the QDA model has a high accuracy of 82.05 percent. The kappa statistics for the LDA model is -0.3424 and the QDA model is 0.6276. The kappa statistic generally determines the agreement between the predicted and actual classes. In this case, the QDA model exhibits high agreement beyond the chance of the positive value 0.6276 to that of LDA model in negative value -0.3424.

The 95 percent confidence interval for accuracy range in the LDA model is low, between 16.38 percent to 57.27 percent, whereas the QDA model comes with a better accuracy

range from 71.72 percent to 89.83 percent. It is evident that the LDA model has poor predictive capabilities when compared to the QDA model.

The Sensitivity(True Positive Rate), specificity(True Negative Rate) and misclassification error of the LDA model are 20 percent,46.15 percent and 65.21 percent. And the QDA model has 70.59 percent,90.91 percent and 17.94 percent respectively. Sensitivity and specificity are significantly higher in the QDA model compared to the LDA model. This is a true positive rate of invasive cancer cases and a true negative rate of noninvasive cancer cases in a random 10 gene subset. The misclassification error is comparatively lower in the QDA model than the LDA model, which determines that the QDA model possess more accurate classification of cancer types. The LDA model approaches with the assumption that classes are linear. However, the QDA model showcases more flexibility for complex decisions, which leads to better classification performance.

In order to conclude, the above LDA and QDA models. In this case, the difference in assumption of the predictors component makes the QDA model to be more flexible with respect to the covariance matrix. This further enhances the performance of the QDA model compared to the LDA model for this analysis. The LDA model seems to be more promising for simpler computation of a dataset where the assumptions of equal covariance matrices is satisfied. But the QDA model has become a more optimal model to use due to its flexibility in negotiating the assumption of covariance matrices.

This supports the QDA model to be more reliable for complex relationship modelling between predictors and classes.

**###5. (25 marks). Use the median of the first principal component of your random subset of 10 genes to predict the class variable (invasive (label 1) and noninvasive (label 2) cancer). Use Fisher's Exact test and sensitivity,specificity and Youden index. Add appropriate tables and figures to your report.**

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction  0  1
##          0  0 16
##          1  0 18
##
##                Accuracy : 0.5294
##                  95% CI : (0.3513, 0.7022)
##     No Information Rate : 1
##     P-Value [Acc > NIR] : 1.0000000
##
##                   Kappa : 0
##
##  Mcnemar's Test P-Value : 0.0001768
```

```
##
##             Sensitivity :     NA
##             Specificity : 0.5294
##          Pos Pred Value :     NA
##          Neg Pred Value :     NA
##              Prevalence : 0.0000
##          Detection Rate : 0.0000
##    Detection Prevalence : 0.4706
##       Balanced Accuracy :     NA
##
##        'Positive' Class : 0
##

## NULL

## NULL

## numeric(0)

##
##  Fisher's Exact Test for Count Data
##
## data:  Fisher.table
## p-value = 0.8196
## alternative hypothesis: true odds ratio is not equal to 1
## 95 percent confidence interval:
##  0.3010991 2.1825224
## sample estimates:
## odds ratio
##  0.8137785
```

The main aim of this analysis is to predict the relationship between predicted classifications based on the first principal component of gene expression data and the actual class using Fischer's exact test. Fischer's exact test is used to determine the significant association between two categorical values when the sample size is small and chi-square assumptions (no random sampling, low count in contingency table) are violated. A significant tool for analysis when other methods are not appropriate.

Null Hypothesis(H0): There is no association between the predicted class based on the first principal component and the actual class from the dataset.

Alternative Hypothesis(H1): There is an association between the predicted class based on the first principal component and the actual class from the dataset.

A new data frame is created as Data.5 by excluding the last column from my.gene.subset data frame, an essential step to perform Principal Component Analysis on the gene expression data only. Principal Component Analysis is performed on new data frame Data.5. Further, the First Principal Component is extracted from the result and the respective median is computed. This median value is a threshold for classifying the data

points into two groups based on their position relative to the median along the first principal component axis. The data points are further classified as 1 for points above the median, 0 for points below the median. The actual class labels are extracted from the original data frame and are converted to character type for analysis. A contingency table is created with predicted and actual classes to execute the Fisher's exact test to check the association between the predicted and actual class labels.

The Confusion matrix shows the True negatives is 0, False Positives is 16, False Negatives is 0, True positives is 18. The Accuracy of classification of sample is 52.94 percent for this model. Kappa value is 0 which represents that there is no agreement beyond the chance between the predicted and actual class. Specificity marks 52.94 percent, representing the negative samples of the model. McNemar's Test P-value is 0.0001768 which shows that the significant difference in error between the predicted and actual classes. Youden's Index is 0 as its cannot be calculated as sensitivity is not available.

The above Fisher's Exact test has a p-value 0.8196, which shows insufficiency to reject the null hypothesis. This suggests a likelihood of no association between the predicted class on the first principal component and the actual class. The 95 percent confidence interval for the odds ratio ranges from 0.3010991 to 2.1825224. And the estimated odds ratio calculated as 0.8137785 depicts that the odds of the correct class based on the first component are slightly lower than the odds of the incorrect class. The difference is not significant.

To conclude, on the above Fisher's Exact test, there is insufficient evidence to conclude the association. Alternative methods or larger datasets are required for further validation. Hence, it concludes that there is no association between the predicted class based on the first principal component and the actual class from the dataset.

**Appendix:**

**R code below:**

**Read the data into RStudio (or R) using the read.csv R command**

```
InitialData <- read.csv(file="gene-expression-invasive-vs-noninvasive-cancer.csv")
```

**Check using the str, dim and dimnames command it worked - parts of the generated output are added as comments; lines starting with "#" comments and not R code.**

```
str(InitialData)
```

```
## 'data.frame':    78 obs. of  4949 variables:
##  $ J00129        : num  -0.448 -0.48 -0.568 -0.819 -0.112 -0.391 -0.624 -
0.528 -0.811 -0.839 ...
##  $ Contig29982_RC: num  -0.296 -0.512 -0.411 -0.267 -0.67 -0.31 -0.12 -
0.447 -0.536 2 ...
##  $ Contig42854   : num  -0.1 -0.031 -0.398 0.023 0.421 -0.06 -0.236 -0.254
-0.211 0.147 ...
##  $ Contig42014_RC: num  -0.177 -0.075 0.116 -0.23 -0.19 -0.164 -0.175
0.017 -0.201 -0.325 ...
##  $ Contig27915_RC: num  -0.107 -0.104 -0.092 0.198 0.032 -0.173 0.253
0.654 0.287 -0.303 ...
##  $ Contig20156_RC: num  -0.11 -0.234 -0.166 -0.51 0.281 -0.034 -0.125
0.364 -0.08 -0.061 ...
##  $ Contig50634_RC: num  -0.095 -0.225 0.036 0.529 0.31 -0.091 -0.127 0.068
-0.15 0.097 ...
##  $ Contig42615_RC: num  -0.076 -0.094 0.397 0.354 0.056 0.036 -0.02 0.181
0.045 0.006 ...
##  $ Contig56678_RC: num  -0.134 0.115 -0.194 -0.261 0.116 0.346 0.047 -1.14
-0.11 0.176 ...
##  $ Contig48659_RC: num  -0.14 0.019 -0.128 0.012 0.074 0.007 -0.15 -0.111
-0.072 -0.084 ...
##  $ Contig49388_RC: num  0.006 0.15 0.139 -0.26 0.041 0.251 0.266 -0.153
0.471 0.114 ...
##  $ Contig1970_RC : num  0.111 0.038 -0.033 -0.069 0.067 0.229 0.246 -0.415
-0.096 -0.081 ...
##  $ Contig26343_RC: num  -0.236 0.092 0.039 -0.115 0.279 0.297 0.142 0.111
0.047 -0.071 ...
##  $ Contig53047_RC: num  -0.866 -1.035 -1.114 -1.021 -1.006 ...
##  $ Contig43945_RC: num  0.126 -0.062 0.011 -0.999 0.211 -0.1 -0.194 -0.053
0.096 -0.121 ...
##  $ Contig19551   : num  -0.692 -0.21 -0.462 0.273 0.242 -0.883 0.206 0.174
-0.355 0.23 ...
##  $ Contig10437_RC: num  0.132 -0.139 -0.185 0.159 0.276 -0.146 -0.301 -
0.075 0.253 0.022 ...
##  $ Contig47230_RC: num  0.095 0.068 -0.168 -0.398 -0.604 0.382 -0.549 -
0.635 0.856 0.515 ...
##  $ Contig20749_RC: num  0.252 0.268 -0.289 -0.734 0.08 0.403 -0.012 -0.586
0.105 0.138 ...
##  $ AL157502      : num  0.139 -0.179 -0.378 -0.427 0.372 -0.014 -0.022 -
0.821 -0.294 -0.165 ...
##  $ Contig36647_RC: num  -0.097 0.181 -0.494 0.848 -0.01 0.6 -0.984 0.077 -
0.15 0.58 ...
##  $ D31887        : num  0.113 0.06 -0.211 -0.338 0.076 -0.025 0.075 -0.03
-0.275 0.14 ...
##  $ AB033006      : num  -0.209 -0.198 -0.331 -0.239 -0.118 -0.317 -0.25 -
0.082 -0.017 -0.32 ...
##  $ AB033007      : num  0.107 -0.04 0.114 0.081 -0.072 0.134 0.131 0.069
0.177 0.21 ...
##  $ M83822        : num  0.098 0.147 -0.121 -0.09 0.075 0.295 0.024 -0.39 -
0.171 0.03 ...
```

```
##  $ AB033025      : num  0.11 0.087 -0.141 -0.61 0.236 -0.094 -0.067 -0.116
-0.175 -0.774 ...
##  $ AF114264      : num  0.096 0.051 -0.164 -0.047 0.245 -0.165 -0.072 -
0.427 -0.249 -0.372 ...
##  $ Contig40673_RC: num  0.305 -0.056 -0.124 -0.02 -0.19 0.016 -0.246 0.181
1.48 -0.199 ...
##  $ Contig17345_RC: num  0.055 -0.031 -0.031 0.251 -0.06 -0.104 -0.254
0.408 -0.003 0.002 ...
##  $ AB033034      : num  -0.137 -0.05 -0.188 0.153 0.181 -0.231 -0.032 -
0.024 -0.305 -0.249 ...
##  $ AB033035      : num  -0.056 -0.162 0.06 -0.249 -0.046 -0.129 0.15 0.088
-0.286 -0.343 ...
##  $ AF227899      : num  -0.001 0.11 -0.395 0.175 0.411 -0.024 -0.246 -
0.385 -0.465 0.044 ...
##  $ AB033043      : num  0.108 0.105 0.079 -0.223 0.109 0.201 0.361 -0.224
0.02 0.188 ...
##  $ AB033049      : num  0.329 0.049 0.177 -0.307 0.3 0.046 -0.139 0.036 -
0.154 0.069 ...
##  $ Contig55834_RC: num  0.078 0.175 -0.375 0.017 0.255 0.035 -0.044 -0.379
-0.302 0.071 ...
##  $ Contig67229_RC: num  -0.098 -0.107 0.612 -0.107 0.143 0.065 0.061 0.358
0.016 0.025 ...
##  $ Contig3396_RC : num  -0.097 -0.068 0.071 -0.113 -0.021 0.4 0.031 -0.043
-0.004 -0.199 ...
##  $ AB033050      : num  -0.019 0.104 0.023 0.582 0.128 0.216 0.091 -0.186
-0.156 0.032 ...
##  $ AB033055      : num  0.208 0.003 -0.051 0.024 0.161 -0.073 -0.167 -0.28
-0.071 -0.099 ...
##  $ AF009314      : num  -0.021 0.025 -0.279 -0.226 0.09 0.026 0.121 -0.541
-0.106 0.041 ...
##  $ AB033062      : num  0.113 -0.166 -0.153 -0.695 0.101 -0.037 0.366
0.139 -0.05 -0.202 ...
##  $ AB033066      : num  0.178 0.065 -0.077 0.176 -0.089 0.018 -0.144 -
0.077 -0.157 0.008 ...
##  $ Contig46243_RC: num  -0.081 0.015 -0.262 -0.209 0.366 0.292 -0.058 -
0.017 -0.082 0.113 ...
##  $ Contig26077_RC: num  0.272 0.197 -0.218 -0.038 -0.254 0.107 0.14 -0.4
0.26 0.219 ...
##  $ U45975        : num  0.737 0.268 -0.064 -0.226 -0.511 0.646 0.36 -0.432
0.25 -0.178 ...
##  $ Contig43679_RC: num  0.122 0.099 -0.646 0.169 -0.128 -0.055 -0.21 -
0.353 -0.12 0.12 ...
##  $ AB033073      : num  -0.152 0.018 -0.028 -0.167 0.307 0.157 0.261 -
0.297 0.129 0.282 ...
##  $ AF018081      : num  -0.063 -0.212 -0.129 -0.347 0.133 -0.128 0.05 -
0.389 -0.116 0.145 ...
##  $ AB033079      : num  -0.009 0.07 -0.159 0.004 0.018 -0.26 0.069 -0.18 -
0.378 -0.092 ...
##  $ X56210        : num  -0.146 -0.06 -0.045 -0.13 0.185 -0.099 0.04 -0.168
-0.146 -0.095 ...
```

```
##  $ AB033086      : num  0.223 -0.098 0.271 -2 -0.156 0.102 -0.016 -0.358
0.153 -0.055 ...
##  $ AB033091      : num  -0.176 0.018 -0.194 0.212 0.097 -0.202 0.073 -
0.481 -0.391 -0.014 ...
##  $ AB033092      : num  0.015 -0.064 -0.312 -0.08 0.138 -0.071 0.012 0.025
-0.333 -0.013 ...
##  $ NM_003004     : num  -0.47 -0.576 -0.064 -0.104 0.134 -0.09 -0.072
0.854 -0.068 -0.049 ...
##  $ Contig57877_RC: num  0.212 -0.053 -0.088 -0.356 0.007 -0.12 -0.14 -
0.368 -0.277 -0.401 ...
##  $ NM_003010     : num  -0.211 -0.331 -0.355 -0.064 0.395 0.227 0.118 -
0.106 -0.089 0.261 ...
##  $ NM_003012     : num  0.238 -0.26 0.141 -0.306 -0.098 -0.114 -0.026
0.263 -0.095 -0.222 ...
##  $ NM_003014     : num  0.039 -0.039 0.112 -0.273 -0.051 -0.047 0.317 -
0.54 0.029 -0.16 ...
##  $ Contig43806_RC: num  -0.734 -0.661 -0.632 -0.944 -0.94 -0.58 -0.494 -
0.924 -0.616 -0.7 ...
##  $ Contig29226_RC: num  0.045 -0.135 -0.041 -0.54 0.185 -0.033 -0.064
0.084 0.048 -0.04 ...
##  $ NM_003020     : num  -0.103 -0.255 -0.034 -0.548 -0.067 -0.237 -0.002 -
0.351 -0.362 -0.164 ...
##  $ NM_003022     : num  0.292 0.092 -0.049 0.318 -0.051 0.259 -0.002 -
0.284 -0.158 0.125 ...
##  $ Contig54847_RC: num  0.181 -0.208 -0.178 -0.692 0.129 0.198 0.436 -
0.838 -0.029 0.166 ...
##  $ Contig33260_RC: num  0.056 0.297 -0.342 0.007 0.175 0.168 0.41 -0.089 -
0.035 -0.006 ...
##  $ NM_002300     : num  -0.434 -0.316 -0.525 0.033 -0.178 -0.023 -0.149
0.245 -0.131 -0.457 ...
##  $ Contig14658_RC: num  0.043 0.087 -0.036 -0.115 0.208 0.058 0.331 -0.161
0.599 0.23 ...
##  $ NM_003033     : num  0.236 0.031 0.34 0.023 -0.247 0.123 -0.309 -0.077
-0.119 -0.23 ...
##  $ NM_003034     : num  0.096 -0.09 -0.047 -0.011 -0.406 -0.244 -0.218 -
0.081 -0.016 -0.593 ...
##  $ NM_002306     : num  -0.271 0.066 0.092 -0.185 -0.01 0.025 0.094 0.152
0.156 0.185 ...
##  $ NM_003035     : num  -0.385 -0.08 0.053 -0.032 -0.071 -0.184 -0.534
0.581 -0.283 -0.253 ...
##  $ NM_002308     : num  -0.237 -0.269 0.203 0.312 0.088 -0.399 -0.076
0.425 -0.054 0.116 ...
##  $ NM_003038     : num  0.131 0.275 0.065 0.043 -0.171 0.192 -0.046 0.086
0.384 0.015 ...
##  $ NM_002313     : num  -0.047 -0.036 0.109 0.516 -0.197 0.063 0.026 -
0.108 -0.185 0.143 ...
##  $ Contig54839_RC: num  0.13 -0.101 0.224 -0.149 0.01 -0.05 -0.104 0.083
0.156 0.028 ...
##  $ NM_002318     : num  -0.386 0.189 -0.122 -0.75 0.039 -0.236 0.343 -
0.285 0.148 -0.205 ...
```

```
##  $ NM_003051     : num  0.299 -0.173 0.193 -0.02 -0.155 0.005 -0.375 -0.3
-0.363 0.064 ...
##  $ NM_003056     : num  0.116 -0.073 0.03 -0.041 -0.164 -0.049 0.113 0.167
0.035 -0.314 ...
##  $ Contig66143_RC: num  -0.294 0.55 0.642 -0.087 -0.381 -0.417 -0.137 -
0.081 -0.27 -0.859 ...
##  $ Contig51809_RC: num  0.169 -0.086 0.129 -0.3 -0.054 -0.104 0.076 0.176
0.184 -0.298 ...
##  $ NM_002332     : num  0.025 -0.141 -0.113 -0.389 0.257 0.047 0.341 -
0.346 0.221 0.215 ...
##  $ NM_001605     : num  -0.101 -0.138 -0.054 0.232 -0.147 -0.083 -0.082 -
0.103 -0.113 -0.213 ...
##  $ NM_003064     : num  -0.065 -0.107 -0.033 0.069 -0.019 0.006 -0.052
0.305 0.527 0 ...
##  $ NM_002336     : num  -0.005 -0.162 -0.015 -0.024 -0.051 0.122 0.11 -
0.039 -0.079 0.209 ...
##  $ NM_002337     : num  -0.083 -0.024 -0.17 0.023 -0.029 -0.019 0.119
0.034 0.192 -0.016 ...
##  $ NM_003066     : num  -0.131 -0.093 -0.026 0.028 -0.029 -0.016 -0.097
0.319 0.432 0.004 ...
##  $ NM_001609     : num  0.081 -0.026 -0.133 0.077 -0.191 0.102 -0.207 -
0.595 -0.107 -0.131 ...
##  $ Contig50846_RC: num  0.064 -0.051 -0.083 -0.009 -0.063 -0.019 -0.019
0.306 -0.005 -0.219 ...
##  $ NM_001611     : num  -0.712 -0.435 -0.532 -0.097 -0.278 0.323 -0.371
0.188 -0.033 -0.062 ...
##  $ NM_003070     : num  0.09 0.028 -0.042 -0.261 0.151 0.078 0.188 -0.177
-0.254 0.227 ...
##  $ NM_002341     : num  -0.269 -0.731 -0.177 0.369 -0.48 -0.455 -0.133
0.219 0.114 -0.513 ...
##  $ NM_001613     : num  -0.143 0.053 -0.05 -0.492 0.074 -0.121 0.277 -
0.331 -0.07 -0.145 ...
##  $ NM_003071     : num  -0.08 -0.125 0.097 -0.012 -0.003 0.381 -0.355
0.127 0.181 -0.159 ...
##  $ NM_001614     : num  -0.064 0.102 -0.031 -0.112 -0.196 -0.114 0.122
0.052 -0.141 0.13 ...
##  $ NM_002343     : num  -0.58 -1.26 -0.261 -0.356 -0.547 -0.371 -0.026
0.722 -0.657 0.314 ...
##  $ NM_001615     : num  -0.75 -0.23 -0.071 -0.999 -0.573 -0.933 -0.514 -
0.696 -0.841 -0.529 ...
##  $ NM_002345     : num  -0.177 0.053 -0.251 -0.124 0.261 -0.182 0.045 -
0.552 -0.2 -0.134 ...
##  $ NM_002346     : num  -0.339 -0.08 0.253 0.393 -0.099 -0.159 -0.129 0.07
-0.002 0.057 ...
##  $ NM_001618     : num  -0.292 -0.242 -0.125 0.085 0.181 -0.177 -0.141
0.09 -0.327 0.02 ...
##  $ Contig52320   : num  -0.01 0.311 -0.024 0.191 0.064 -0.096 0.12 -0.481
-0.306 -0.07 ...
##   [list output truncated]
```

**'data.frame': 78 obs. of 4949 variables:**

**$ J00129 : num -0.448 -0.48 -0.568 -0.819 -0.112 -0.391 -0.624 -0.528 -0.811 -0.839 …**

**$ Contig29982_RC: num -0.296 -0.512 -0.411 -0.267 -0.67 -0.31 -0.12 -0.447 -0.536 2 …**

```
dim(InitialData)

## [1]   78 4949
```

**[1] 78 4949**

**Command ——————————————————————**

```
dimnames(InitialData)[[2]][4947:4949]

## [1] "NM_000898" "AF067420"  "Class"
```

**[1] "NM_000898" "AF067420" "Class"**

**The data set has 78 rows (patients) 4949 columns with 4948 gene expression measurement of cancer tissue, each column representing a 'gene'with column 4949 having the information of a class variable with two values: 1 and 2.**

```
table(InitialData[4949])

## Class
##  1  2
## 34 44
```

**Class**

**1 2**

**34 44**

**To gurantee that your data narrative is individual we ask you to set the seed of the random number generator with R function set.seed using your Registration Number;for example "2244222" - replace by your registration number**

```
set.seed(2310158)
```

**Note: R function set.seed generates no R output. Make sure that you run your final data analysis with the submitted R code and that all tables and figures are generated by the final code. You may use R Mark Down or similar R tools to support this.**

**Selects your random individual subset of 50 genes**

```
my.gene.subset <- InitialData[,rank(runif(1:4948))[1:10]]
str(my.gene.subset)
```

```
# To calculate variance of random subset of 10 genes.
variance<-var(my.gene.subset)
# To calculate co-variance of random subset of 10 genes.
```

```
covar.cov<-cov(as.matrix(my.gene.subset))
# To calculate correlation of random subset of 10 genes
correlation<-cor(as.matrix(my.gene.subset))

library(knitr)
```

```r
# Creating a table to show the variance, covariance,correlation of the random
10 genes
Variance_table<-as.data.frame(variance)
kable(Variance_table,caption = "Summary of variance of 10 random genes")

covariance_table<-as.data.frame(covar.cov)
kable(covariance_table,caption = "Summary of covariance of 10 random genes")

correlation_table<-as.data.frame(correlation)
kable(correlation_table,caption = "Summary of correlation of 10 random
genes")



# To obtain the distance matrix for randomly selected 10 genes.
distance_matrix<-dist(scale(my.gene.subset))
print(distance_matrix)
```

```r
library(ggplot2)

## Warning: package 'ggplot2' was built under R version 4.3.2

library(reshape2)

## Warning: package 'reshape2' was built under R version 4.3.2

#Reshape the data set
data.reshape<-melt(my.gene.subset[1:10])

## No id variables; using all as measure variables

#Univariate Q-Q plot
QQplot.Univariate<-ggplot(data.reshape,aes(sample=value))+
  geom_qq(col='blue')+
  geom_qq_line(col='red')+
  facet_wrap(~ variable, scales = "free")+
  theme_minimal()+
  labs(title = 'Q-Q plots for each random 10 genes')
plot(QQplot.Univariate)


#To create Q-Q plots based on generalised distance
pairs(my.gene.subset)

# cm - mean value for column variable of the dataset.
cm <-colMeans(my.gene.subset)

# covariance matrix of the dataset
S<-cov(my.gene.subset)
```

```r
# Mahalanobis distances is used to calculate the distance between

#the point and the distribution
d<-apply(my.gene.subset,MARGIN = 1, function(my.gene.subset)
+t(my.gene.subset-cm)%*% solve(S)%*%(my.gene.subset-cm))

# plot to assess Mahalanobis distance
plot(qchisq(((1:nrow(my.gene.subset))-1/2)/nrow(my.gene.subset), df=9),
     sort(d),
     xlab = expression(paste(chi[3]^2, "Quantile")),
     ylab = "Ordered distances");abline(a = 0, b = 1,col='dark
Green',title("Q-Q plots on the generalised distance for 10 random genes"))




# To obtain Principal Component Analysis for the subset my.gene.subset
pca.mygenesubset<-princomp(my.gene.subset,cor = TRUE)

# Summarize the Principal Component Analysis with loadings and cutoff
summary(pca.mygenesubset,loadings= TRUE,cutoff=.0)

# Graphical representation of the Principal Component Analysis
plot(pca.mygenesubset)



#3.(a) To find the MANOVA of random subset of 10 genes.
#Null Hypothesis H0:There is no significant difference in the expression
levels of the random subset of 10 genes between invasive (label 1) and
noninvasive (label 2) cancer.
#Alternative Hypothesis (H1): There is a significant difference in the
expression levels of the random subset of 10 genes between invasive (label 1)
and noninvasive (label 2) cancer.
my.gene.subset$class<-InitialData$Class
my.gene.subset$class

x<-as.matrix(my.gene.subset[,c(1:10)])

# To fit the MANOVA
Manova.fit<-manova(x~my.gene.subset$class)

# Summarize the MANOVA fit
summary(Manova.fit,intercept=TRUE)

#Box plot representation of MANOVA

boxplot(x~my.gene.subset$class)
```

```r
#3.b Taking the first two Principal components
pca.mygenesubset<-princomp(my.gene.subset,cor=TRUE)

# Extracting the coefficients for first and second principal components
coeff.PC1.PC2<-pca.mygenesubset$loadings[,1:2]

#Print the coefficients of first(PC1) and second(PC2) principal components
print(coeff.PC1.PC2)

# Graphical representation of the difference between invasive vs noninvasive
plot(coeff.PC1.PC2)



#
gene_class_var<-InitialData$Class

gene_exclu_class<-my.gene.subset[,1:10]
gene_pca<-prcomp(gene_exclu_class,scale=TRUE)

#Set a dataframe of extracted PC1 and PC2 from PCA results
gene_set<-as.data.frame(gene_pca$x[,1:2])
gene_set$class<-gene_class_var
# To convert class to factor variable for plotting purpose
gene_set$class<-as.factor(gene_set$class)
# ggplot to show the PCA plot
library(ggplot2)
ggplot(gene_set,aes(x=PC1,y=PC2,color=class))+geom_point(alpha=0.7)+
  theme_minimal()+
  labs(title="Principal Component Analysis of Gene Expression",
       x="Principal Component 1",
       y="Principal Component 2",
       color="Cancer Type")+
  scale_color_manual(values=c("1"="red","2"="blue"))



#4.(a)
library(MASS)

library(caret)

# Converting the class to factor type

my.gene.subset$class<-as.factor(my.gene.subset$class)

# Split the data into training and testing sets (optional, but recommended
for performance evaluation)
set.seed(2310158) # Use a seed for reproducibility
indexes <- createDataPartition(my.gene.subset$class, p = 0.7, list = FALSE)
```

```r
Data.1 <- my.gene.subset[indexes, ]
Data.2 <- my.gene.subset[-indexes, ]

# Training set = Data.1, Testing set= Data.2 ,data split in 70:30 ratio
# Apply LDA on the training data
lda.model<-lda(class ~.,data = Data.1)

# Predictions
Predict.lda<-predict(lda.model,Data.2)
Predict.class<-Predict.lda$class

#confusion matrix

Matrix.confusion<-confusionMatrix(Predict.class,Data.2$class)
print(Matrix.confusion)



# Sensitivity, Specificity, and Misclassification Error from the confusion
matrix
print(sensitivity.lda<-Matrix.confusion$byClass['sensitivity'])

print(specificity.lda<-Matrix.confusion$byClass['specificity'])

print(misclassify.lda.error<-1- Matrix.confusion$overall['Accuracy'])



#4.(b)
#Set a subset from the gene expression data
x<-as.matrix(my.gene.subset[,c(1:10)])

#Fitting a QDA(Quadratic Discriminant Analysis) model
#Predictor variable as specified gene expression, response variable as class
Qda.model<-
qda(class~NM_015957+Contig23913_RC+Contig49076_RC+NM_003034+NM_012396+U81599+
NM_004417+AL157424+NM_005512+Contig2339_RC , data = my.gene.subset)

#Print the QDA model details
print(Qda.model)



# Predictions on QDA model
Predict.qdaclass<-predict(Qda.model)$class

# Creating a contingency table comparing predicted class to actual class
table(Predict.qdaclass,my.gene.subset$class)
```

```r
#QDA confusion matrix
Confusion.matrix.QDA<-confusionMatrix(Predict.qdaclass,my.gene.subset$class)
print(Confusion.matrix.QDA)



#Sensitivity,specificity and misclassification error
Sensitivity.QDA<-Confusion.matrix.QDA$byClass['Sensitivity']
Specificity.QDA<-Confusion.matrix.QDA$byClass['Specificity']
misclassify.QDA.error<- 1- Confusion.matrix.QDA$overall['Accuracy']
print(Sensitivity.QDA)

print(Specificity.QDA)

print(misclassify.QDA.error)

5
# Performing PCA excluding the Class variable
Data.5<-my.gene.subset[,-ncol(my.gene.subset)]
Result<-prcomp(Data.5,scale=TRUE)
First.Component<-Result$x[,1]
my.gene.subset$class<-as.factor(my.gene.subset$class)



# Determining the median of PC1
Median.First.Component<-median(First.Component)

# Classifying median based on first component
Predictclassify.Median<-
ifelse(First.Component>Median.First.Component,'1','0')#1 above median, 0
below median

Actualclassify.Median<-as.character(my.gene.subset$class)

#Null Hypothesis(H0):There is no association between the predicted class
based on the first principal component and the actual class from the dataset.



#Alternative Hypothesis(H1):There is an association between the predicted
class based on the first principal component and the actual class from the
dataset.

# Fisher's Exact Test
Fisher.table<-
table(Predicted=Predictclassify.Median,Actual=Actualclassify.Median)
Fishertest.result<-fisher.test(Fisher.table)



# Convert to factors with consistent levels
Predictclassify.Median <- factor(Predictclassify.Median, levels = c("0",
```

```r
"1"))
Actualclassify.Median <- factor(Actualclassify.Median, levels = c("0", "1"))

# To determine Confusion matrix,sensitivity,specificity and youden_index
con_matrix<-confusionMatrix(Predictclassify.Median,Actualclassify.Median)
sensitivity <- con_matrix$byclass['Sensitivity']
specificity <- con_matrix$byclass['Specificity']
youden_index <- sensitivity + specificity - 1

print(con_matrix)

print(sensitivity)

print(specificity)

print(youden_index)

print(Fishertest.result)
```