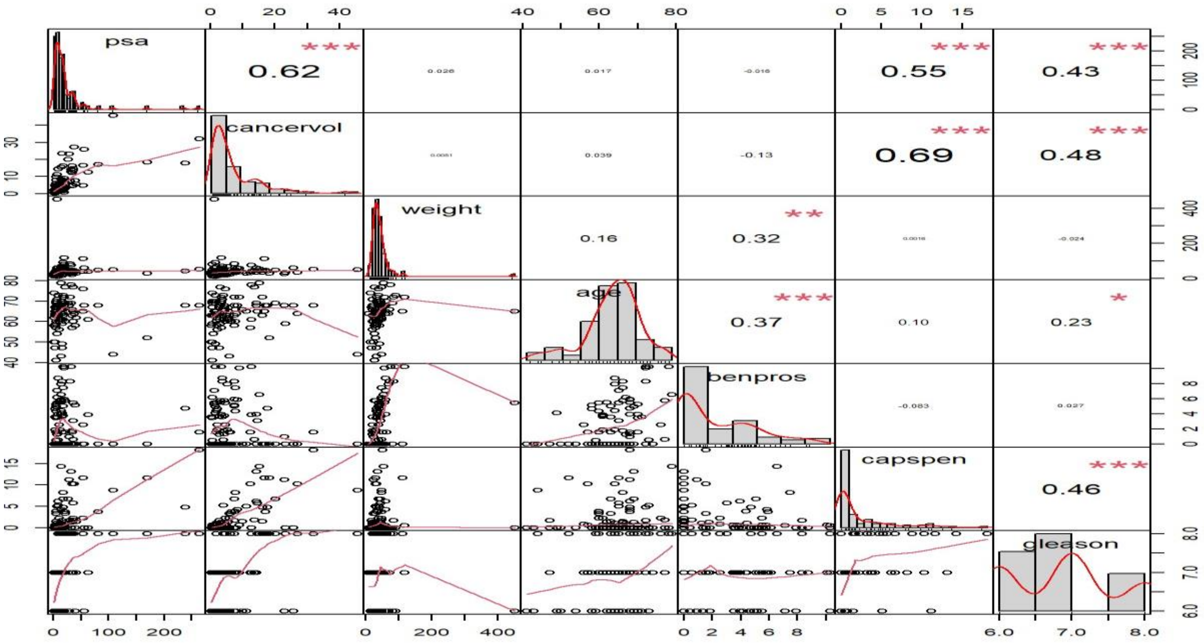
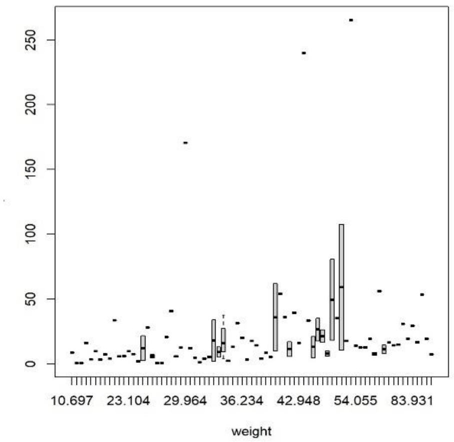
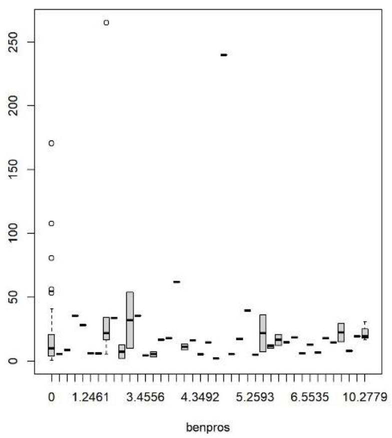
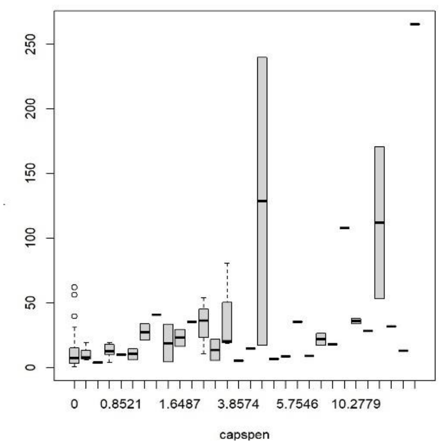
Q1A.

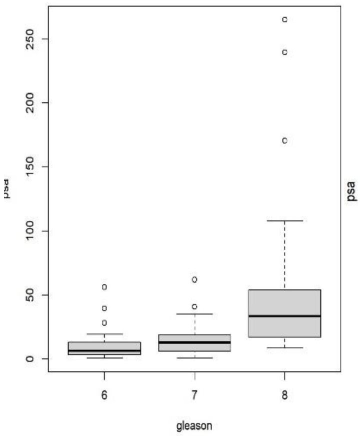
The dataset dimension is 97*9. It has no missing value. Three variable stands out. Firstly ,SubjectId or Column 1 of patient is not important and we would be leaving them out. Gleason is qualitative however it is treated as quantitative. Vesinv is purely qualitative. While observing boxplot we can see that variables has wide range and all of them have quite a few outliers.

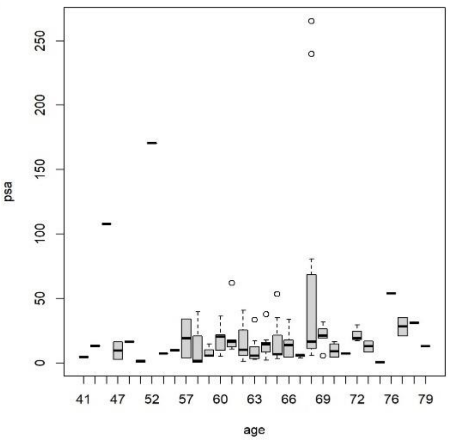


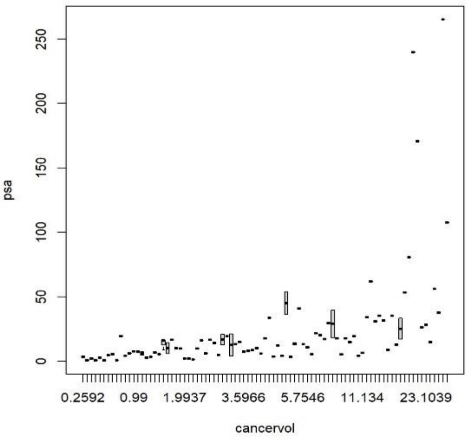








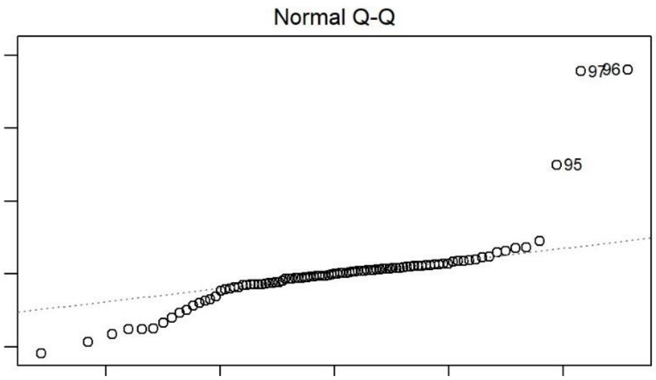




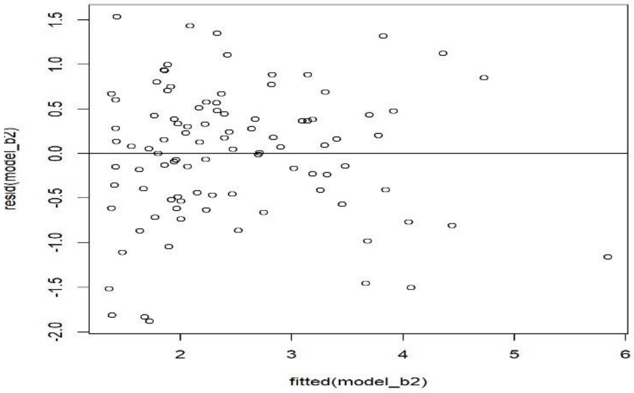
Q1B

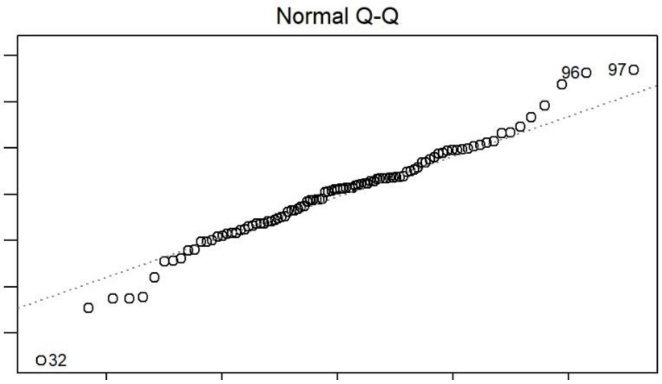
We are fitting multiple linear regression on psa including with vesinv which is a qualitative factor. By examining QQ plot we can see some deviations. Through the means of BOXCOX transformation on MASS package we have applied log transformation and through the means of QQ plot and residual plot we can see that assumptions are held reasonably. Residual plot also shows that approximately E(∈)=0 and var(∈)=σ^2.

Residual Plot and Normal QQ Before log Transformation



Residual Plot and Normal QQ After log Transformation

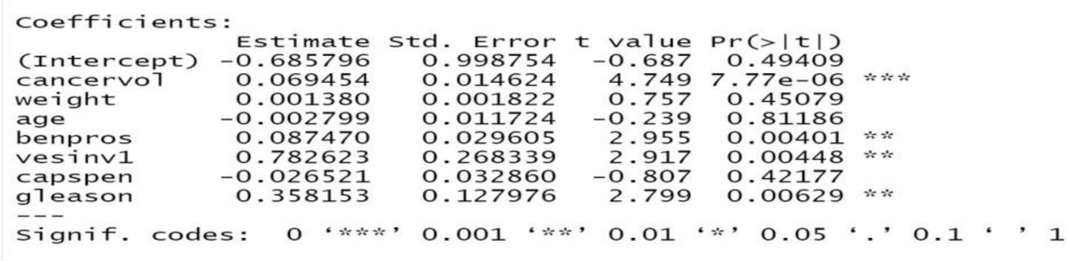




Q1C.We used log(psa) as response value and found out that age, weight and benpros have no statical significance.

Predictor	Estimate	Standard Error	R^2 Adj	T-value	p-value>|t|	Significant
Cancervol	0.096	0.011	0.426	8.47	2.69E-13	Yes
weight	0.003	0.002	0.00446	1.195	0.235	No
Gleason	0.8408	0.1348	0.2831	6.237	1.23E-08	Yes
Vesinv(1)	1.5783	0.2356	0.3136	6.698	1.48E-09	Yes
benpros	0.059	0.03856	0.01451	1.554	0.124	No
age	0.02633	0.01567	0.01865	1.68	0.0961	No
Caspen	0.15796	0.02676	0.2606	5.903	5.5e-.08	Yes

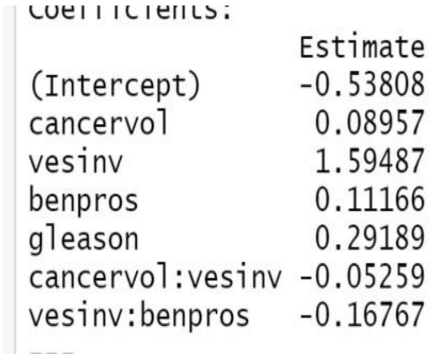
Q1d We have compared the full model with vesinvl having value of 1



```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) -0.685796   0.998754  -0.687  0.49409    
cancervol    0.069454   0.014624   4.749 7.77e-06 ***
weight       0.001380   0.001822   0.757  0.45079    
age         -0.002799   0.011724  -0.239  0.81186    
benpros      0.087470   0.029605   2.955  0.00401 ** 
vesinv1      0.782623   0.268339   2.917  0.00448 ** 
capspen     -0.026521   0.032860  -0.807  0.42177    
gleason      0.358153   0.127976   2.799  0.00629 ** 
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Q1e The final model is chosen after dropping variables based on their p-value and less interactions. Namely weight ,capsen,gleason and vesinv were dropped. We chose remaining on the basis of Adjusted R^2 . After that to get the good model we ran the model diagnostic tools.

Q1f

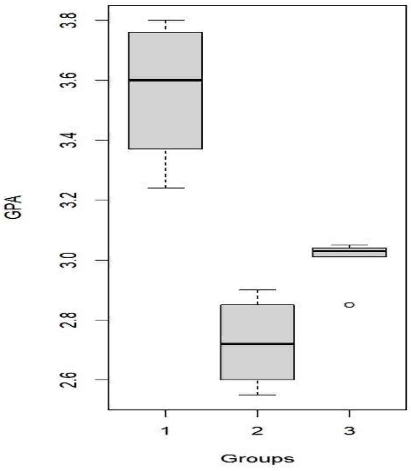


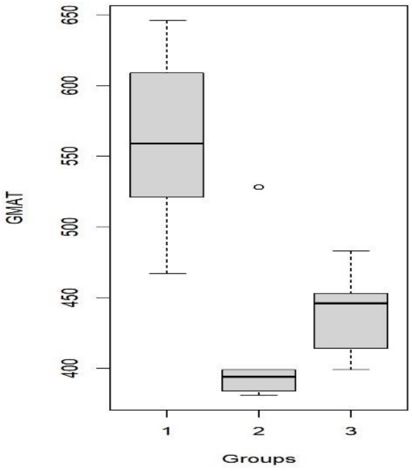
```
Coefficients:
                 Estimate
(Intercept)      -0.53808
cancervol         0.08957
vesinv            1.59487
benpros           0.11166
gleason           0.29189
cancervol:vesinv -0.05259
vesinv:benpros   -0.16767
---
```

*Based on this results we can write the final model as log(psa)= -.53808+.08957*cancervol+1.59487*I(vesinv=1)+.11166*benpros+.29819*gleason-.05259*cancervol*I(vesinv=1)-.16767*benpros*(vesinv=1)*

Q1g We use mean values of the most significant quantitative values namely cancervol, benpros and gleason and we get predicted value of 10.79634 for psa.

Q2

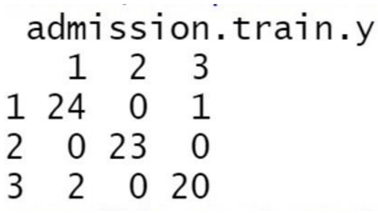




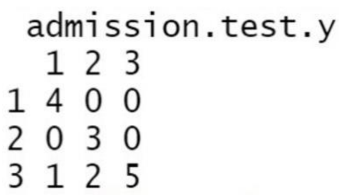
Q2a Group variable is the response variable with three levels: 1 (admit), 2 (do not admit), and 3 (borderline). While GPA and GMAT are predictors. Training set has 70 observation out of 85. We can see that and also can intuitively deduce that Admit has higher GPA and GMAT scores followed by borderline and do not admit.

Q2b.

Misclassification rate for training and test data for LDA is .042=3/70 and .25=3/12, respectively. Training error is 2 while test error is 1 and 2 for admit and do not admit for borderline respectively.

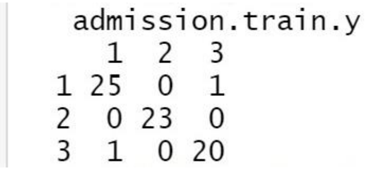


```
   admission.train.y
     1  2  3
  1 24  0  1
  2  0 23  0
  3  2  0 20
```

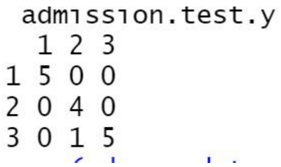


```
 admission.test.y
   1 2 3
 1 4 0 0
 2 0 3 0
 3 1 2 5
```

Q2c Misclassification rate for training and test data for QDA is .028=2/70 and .066=1/15, respectively. Training error is 1 while test error is 1 regarding do not admit for borderline .

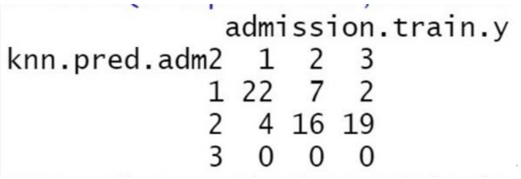


```
   admission.train.y
     1  2  3
  1 25  0  1
  2  0 23  0
  3  1  0 20
```

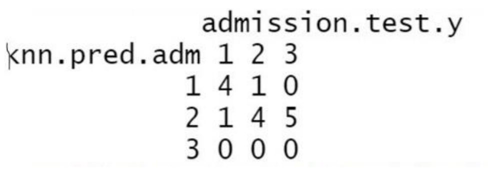


```
 admission.test.y
   1 2 3
 1 5 0 0
 2 0 4 0
 3 0 1 5
```

Q2eMisclassification rate for training and test data for KNN is .45=32/70 and .04=6/15, respectively. Training and test error both seem to be 0.

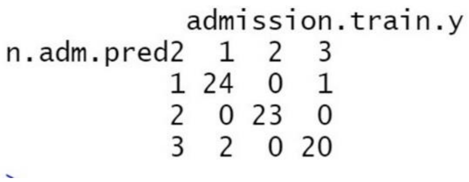


```
               admission.train.y
knn.pred.adm2   1  2  3
            1  22  7  2
            2   4 16 19
            3   0  0  0
```

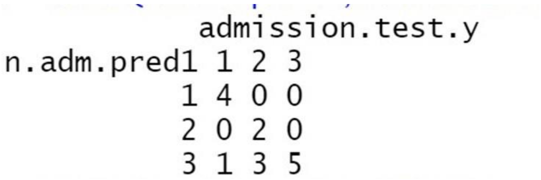


```
               admission.test.y
knn.pred.adm 1 2 3
           1 4 1 0
           2 1 4 5
           3 0 0 0
```

Q2d. Misclassification rate for training and test data for Naïve Bayes is .042=3/70 and .26=4/15, respectively. Training and test error both seem to be 0.

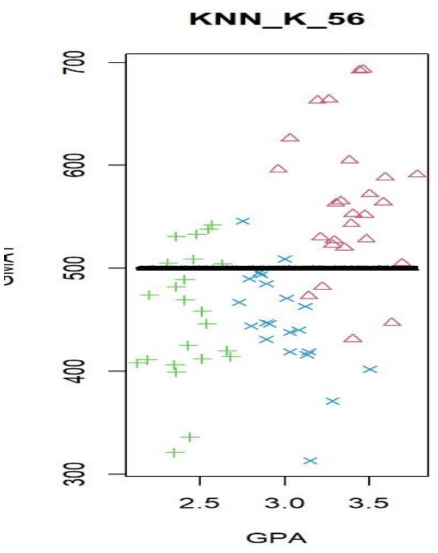


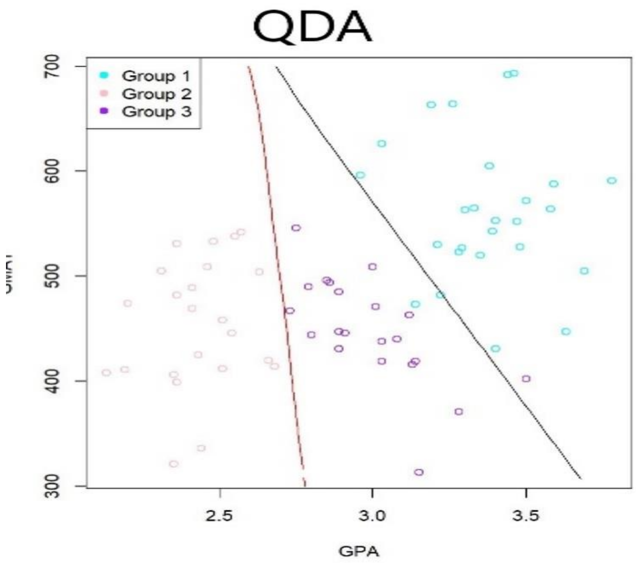
```
             admission.train.y
n.adm.pred2   1  2  3
          1  24  0  1
          2   0 23  0
          3   2  0 20
```

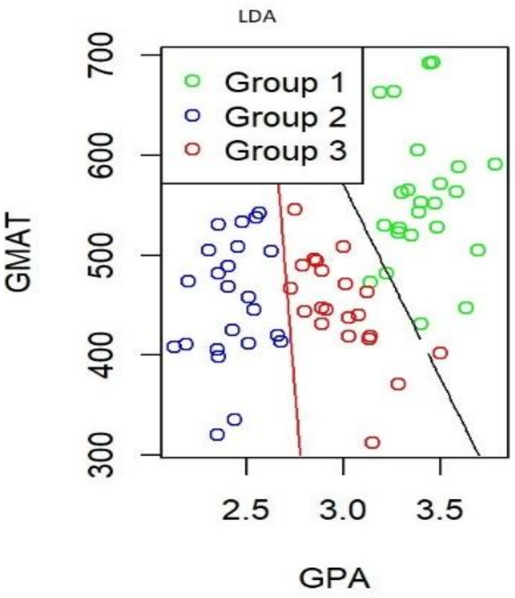


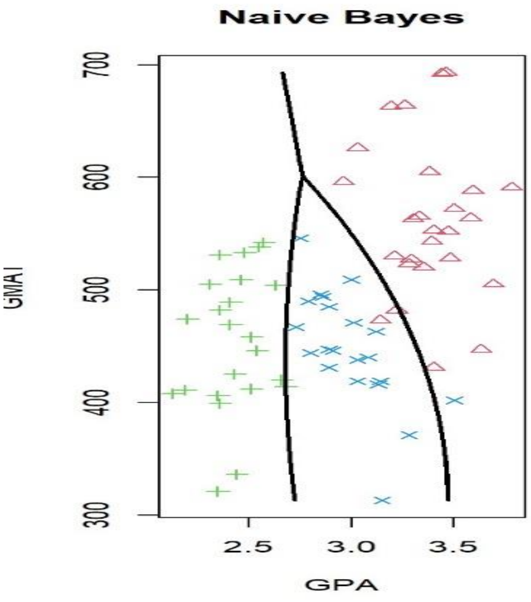
```
             admission.test.y
n.adm.pred1 1 2 3
          1 4 0 0
          2 0 2 0
          3 1 3 5
```

Graphs with decision boundary for above questions for aforementioned classifiers..









We can see that the decision boundary of LDA , Naïve Bayes and QDA was sensible as they clearly demarcated the respective group. However KNN's decision boundary was not as successful in doing so.

Q2f. Based on graphs that showed decision boundary, misclassification rate and the analysis of confusion matrix I came to conclusion that QDA would be the best classifier

Q3

Q3a

$$= \log\left(\frac{\pi_k \exp\left(-\frac{1}{2}(x-\mu_k)^T\mathbf{\Sigma}^{-1}(x-\mu_k)\right)}{\pi_K \exp\left(-\frac{1}{2}(x-\mu_K)^T\mathbf{\Sigma}^{-1}(x-\mu_K)\right)}\right) = \log\left(\frac{\pi_k}{\pi_K}\right) - \frac{1}{2}(x-\mu_k)^T\mathbf{\Sigma}^{-1}(x-\mu_k) + \frac{1}{2}(x-\mu_K)^T\mathbf{\Sigma}^{-1}(x-\mu_K)$$

$$= \log\left(\frac{\pi_k}{\pi_K}\right) - \frac{1}{2}(\mu_k+\mu_K)^T\mathbf{\Sigma}^{-1}(\mu_k-\mu_K) + x^T\mathbf{\Sigma}^{-1}(\mu_k-\mu_K)$$

$$= a_k + \sum_{j=1}^{p} b_{kj}x_j, \qquad where\ a_k = \log\left(\frac{\pi_k}{\pi_K}\right) - \frac{1}{2}(\mu_k+\mu_K)^T\mathbf{\Sigma}^{-1}(\mu_k-\mu_K)$$

Hence proved

Similarly,

$$\log\left(\frac{\Pr\,(Y=k \mid X=x)}{\Pr\,(Y=K \mid X=x)}\right) = \log\left(\frac{\pi_k f_k(x)}{\pi_K f_K(x)}\right)$$

$$= \log\left(\frac{\pi_k \prod_{j=1}^{p} f_{kj}(x_j)}{\pi_K \prod_{j=1}^{p} f_{Kj}(x_j)}\right) = \log\left(\frac{\pi_k}{\pi_K}\right) + \sum_{j=1}^{p} \log\left(\frac{f_{kj}(x_j)}{f_{Kj}(x_j)}\right) == a_k + \sum_{j=1}^{p} g_{kj}(x_j)$$

Hence proved.

Q3b If LDA is assuming Gaussian distributions and it has same covariance matrix but different means then we can say it is a special case of naïve Bayes classifier.

Q3c If some Naïve Bayes has Gaussian distributions, all predictors are independent and same covariance for different class then we can say it is a special case of LDA where the covariance matrix has to be diagonal. These assumptions must be satisfied for naïve Bayes to be special case of LDA.