Stat 407 Quiz 5 Name _____

1. *(1 pt)* The parallel coordinate plot can possibly be scaled in what way(s)? (Choose just one answer) ____

   (a) By subtracting mean and dividing by the standard deviation of each variable

   (b) By subtracting the minimum value and dividing by the range, for each variable, producing numbers between 0 and 1.

   (c) By subtracting the minimum value and dividing by the range, of all variables, producing numbers between 0 and 1.

   (d) All of the above.

   (e) None of the above.

2. *(1 pt)* For the following data matrix,

$$\mathbf{X} = \begin{bmatrix} 4 & -3 & 0 & -1 & 2 \\ 0 & 2 & 2 & -1 & 1 \\ 2 & -1 & -1 & 1 & 0 \end{bmatrix}$$

   Compute the Euclidean distance ($d(\mathbf{X}, \mathbf{Y}) = \sqrt{(\mathbf{X}-\mathbf{Y})'(\mathbf{X}-\mathbf{Y})}$) between observations 2 and 3. ____

3. *(2 pts)* Answer these questions for the following variance-covariance matrix.

$$\mathbf{S} = \begin{bmatrix} 4 & -2 & 4 \\ -2 & 3 & 3 \\ 4 & 3 & 20 \end{bmatrix}$$

   (a) *(1pt)* Which of the following two results is most likely those from PCA on the variance-covariance matrix, **S**? Explain your answer.

   _____

   _____

A

	PC1	PC2	PC3
V1	-0.74	0.01	0.67
V2	0.49	0.69	0.53
V3	-0.46	0.72	-0.52
Var	1.76	1.18	0.05

B

	PC1	PC2	PC3
V1	-0.32	0.69	-0.65
V2	-0.03	-0.69	-0.72
V3	-0.95	-0.21	0.24
Var	20.21	4.85	0.23

   (b) *(1pt)* If an observation has the values [3 2 0] calculate its score on the first principal component of analysis A. (Assume that the means for each variables are 0.) _____

4. *(2 pts)* True or false

   (a) Quadratic discriminant analysis assumes that the samples are take from a multivariate normal population with possibly different means, and different variance-covariance matrices. **T** or **F**

   (b) PCA summarizes only linear dependence. **T** or **F**

5. *(2 pts)* The following matrix, **D**, represents shows the interpoint distances between 5 points, A-E.

$$\mathbf{D} = \left[\begin{array}{c|ccccc} & A & B & C & D & E \\ \hline A & 0 & 1 & 2 & 3 & 3 \\ B & 1 & 0 & 3 & 4 & 5 \\ C & 2 & 3 & 0 & 2 & 3 \\ D & 3 & 4 & 2 & 0 & 2 \\ E & 3 & 5 & 3 & 2 & 0 \end{array}\right]$$

   (a) Points A and B would be joined at the first step of hierarchical clustering. Compute the intercluster distance between cluster AB and point C, using complete linkage. ___

   (b) What would be the height of the dendrogram where A and B are joined? ___

6. *(2 pts)* The following problem was conducted on data collected from the 2012 Australian Open women's tennis tournament. It contains standardized data, of statistics measuring the performance of players in the first three rounds of play. We want to predict if a player will make it to the quarterfinals.

```
> tennis.rp<-rpart(quarters~Aces+Double.Faults+Unforced.Errors+Winners+
     Receiving.Points.Won, data=tennis.sub, method="class")
> tennis.rp
n= 128

node), split, n, loss, yval, (yprob)
      * denotes terminal node

1) root 128 8 FALSE (0.9375000 0.0625000)
  2) Receiving.Points.Won< 49.165 108 0 FALSE (1.0000000 0.0000000) *
  3) Receiving.Points.Won>=49.165 20 8 FALSE (0.6000000 0.4000000)
    6) Aces< 1.535 8 1 FALSE (0.8750000 0.1250000) *
    7) Aces>=1.535 12 5 TRUE (0.4166667 0.5833333) *
```

   (a) *(1)* Calculate the error rate for the classification tree. ___

   (b) *(1)* Draw the classification tree.