4. (3 points)

An actuary wants to cluster six workers compensation classes based on excess ratios at two limits: 750,000 and 1,500,000. The actuary decides to use a weighted k-means algorithm with three clusters. Given the following:

Class	On-Leveled Earned Premium	Normalized Excess Ratio at 750,000 Limit	Normalized Excess Ratio at 1,500,000 Limit	Initial Cluster
1	7,000,000	0.250	0.095	A
2	8,000,000	0.200	0.080	B
3	7,500,000	0.170	0.062	C
4	6,000,000	0.130	0.048	A
5	5,000,000	0.450	0.217	A
6	8,500,000	0.240	0.098	B

- At the start of the algorithm, the actuary randomly assigned each class to a cluster.
- The actuary has observed that the 750,000 limit is selected by the insured 75% of the time, consistent across all six classes.

Briefly justify which measure to use to calculate the distance between vectors and using that measure, determine the cluster for each class after the first iteration of the weighted k-means algorithm.

QUESTION 4	
TOTAL POINT VALUE: 3	**LEARNING OBJECTIVE(S): A1**
SAMPLE ANSWERS	

Sample 1:

Recommend L^1:

- o Class 5 looks to be an outlier, and L^2 will penalize this outlier more | selection of limit being consistent across classes means that L^1 can be used
- o Minimize absolute error or relative error

Sample 2:

Recommend L^2:

- o Class 5 is not enough to be considered an outlier, okay to use L^2. Or could say that the limits are not equally likely to be selected by insureds so L^1 error is not minimized.
- o Minimize mean square error or variance between clusters
- o It is the more traditional approach so use since it doesn't make a significant difference as per the paper
- o Better reflects a skewed distribution

Determine the centroid of each initial cluster:

$R_A(750K) = (7000 * 0.25 + 6000 * 0.13 + 5000 * 0.45) / (7000 + 6000 + 5000) = 0.266$

$R_A(1.5M) = (7000 * 0.095 + 6000 * 0.048 + 5000 * 0.217) / (7000 + 6000 + 5000) = 0.113$

$R_B(750K) = (8000 * 0.2 + 8500 * 0.24) / (8000 + 8500) = 0.221$

$R_B(1.5M) = (8000 * 0.08 + 8500 * 0.098) / (8000 + 8500) = 0.089$

$R_C(750K) = 0.17$

$R_C(1.5M) = 0.062$

Sample 1 (Using L^1 Measure, even weights by limit):

Class	Distance to R_A	Distance to R_B	Distance to R_C	New Cluster

1	0.034 = abs((0.25-0.266)) + abs((0.095-0.113))	0.035	0.113	A
2	0.099	0.030	0.048	B
3	0.147	0.078	0.000	C
4	0.201	0.132	0.054	C
5	0.288	0.357	0.435	A
6	0.041	0.028	0.106	B

Sample 2 (Using L^2 Measure, even weights by limit):

Class	Distance to R_A	Distance to R_B	Distance to R_C	New Cluster
1	0.024 = [(0.25 - 0.266)^2 + (0.095 - 0.113)^2]^(1/2)	0.030	0.087	A
2	0.074	0.023	0.035	B
3	0.109	0.058	0.000	C
4	0.151	0.100	0.042	C
5	0.211	0.262	0.320	A
6	0.030	0.021	0.079	B

Sample 3 (Using L^1 Measure, 75/25 weights):

Class	Distance to R_A	Distance to R_B	Distance to R_C	New Cluster
1	0.016 =abs(.75*(0.25-0.266)) +abs(.25*(0.095-0.113))	0.023	0.068	A
2	0.057	0.018	0.027	B
3	0.084	0.045	0.000	C
4	0.118	0.078	0.034	C
5	0.164	0.204	0.249	A
6	0.023	0.017	0.062	B

Sample 4 (Using L^2 Measure, 75/25 weights):

Class	Distance to R_A	Distance to R_B	Distance to R_C	New Cluster

SAMPLE ANSWERS AND EXAMINER'S REPORT

1	0.016 = [.75*(0.25 - 0.266)^2 + .25*(0.095 - 0.113)^2]^(1/2)	0.026	0.071	A
2	0.059	0.018	0.027	B
3	0.087	0.046	0.000	C
4	0.122	0.081	0.035	C
5	0.168	0.209	0.255	A
6	0.023	0.017	0.063	B

EXAMINER'S REPORT

The candidates were expected to explain which distance measure should be used and how to calculate the means square error and clustering process.

Common mistakes included:
- Not adequately explaining which measure to use
- Not taking the square root for L^2
- Explaining the choice to weight with 75/25 rather than explaining the choice of L^1 or L^2
- Weighting the centroids together and then weighting the distance together: mathematically this gives the same answer when using L^1 measure but does not make sense when using the L^2 measure as it is simply the difference in this case
- Using only one limit instead of both limits to measure distance
- Not calculating all classes and not fully explaining how to compare the distances that weren't calculated.