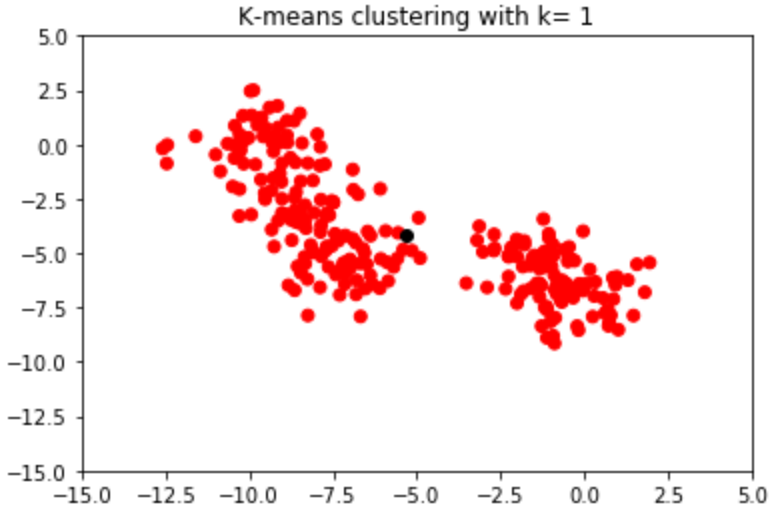
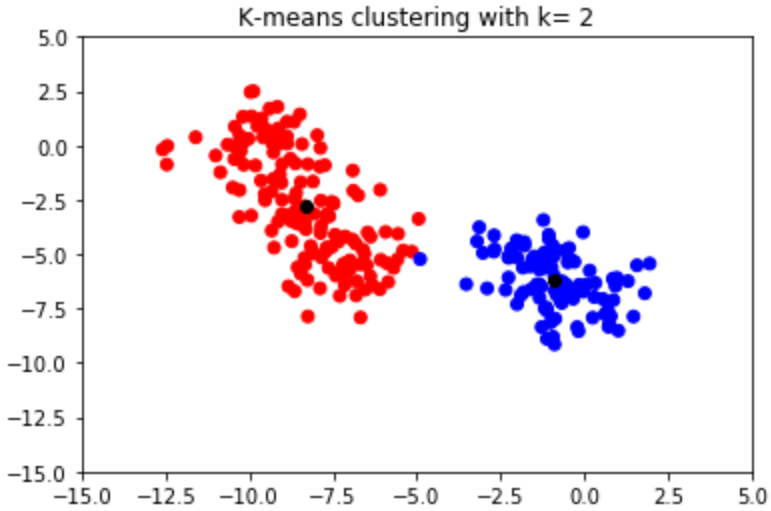
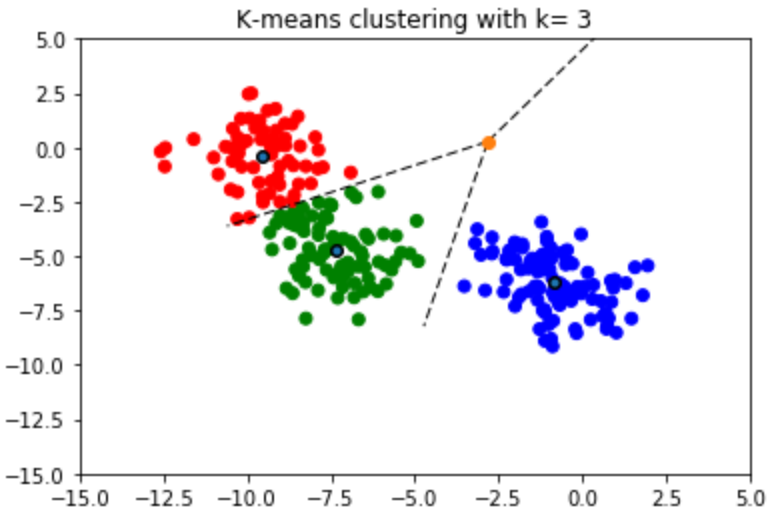
Oluwafemi Oladipupo
6.036 Project 3 Report

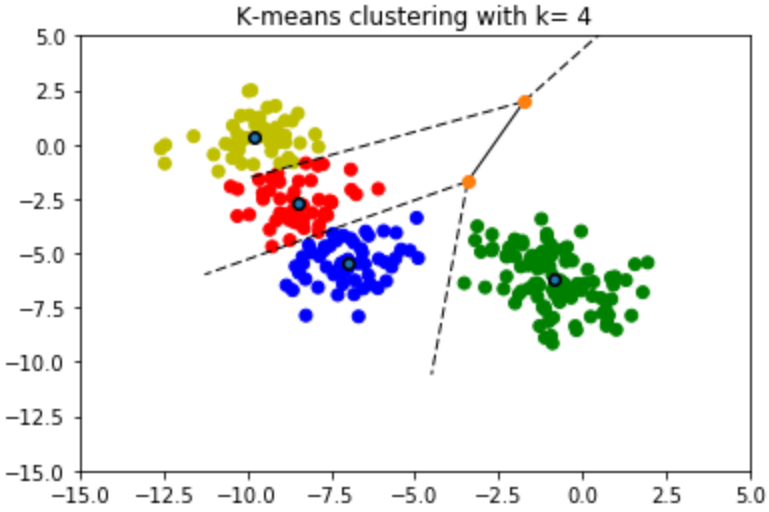
(1) K-Means Versus EM

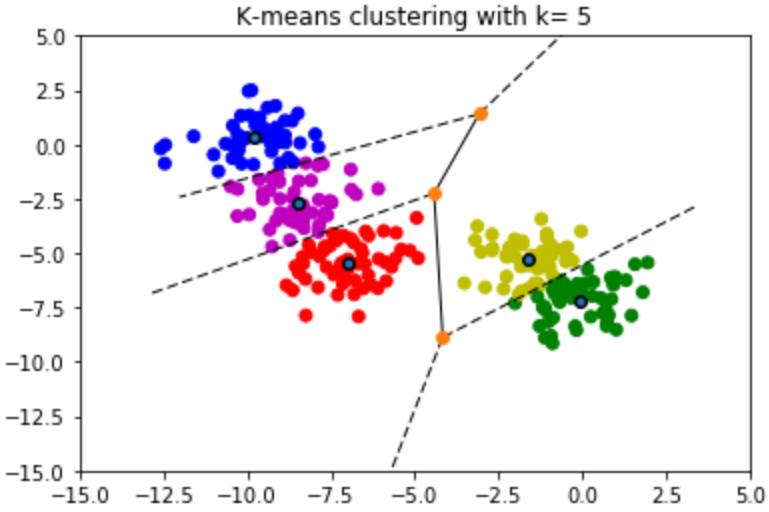
1. K-Means Graphs & Cost
    a. 1099.0
    b. 556.8
    c. 391.0
    d. 340.3
    e. 293.4





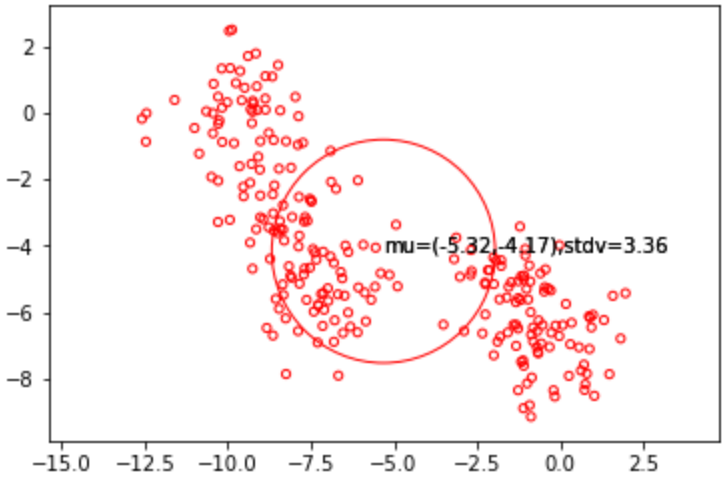


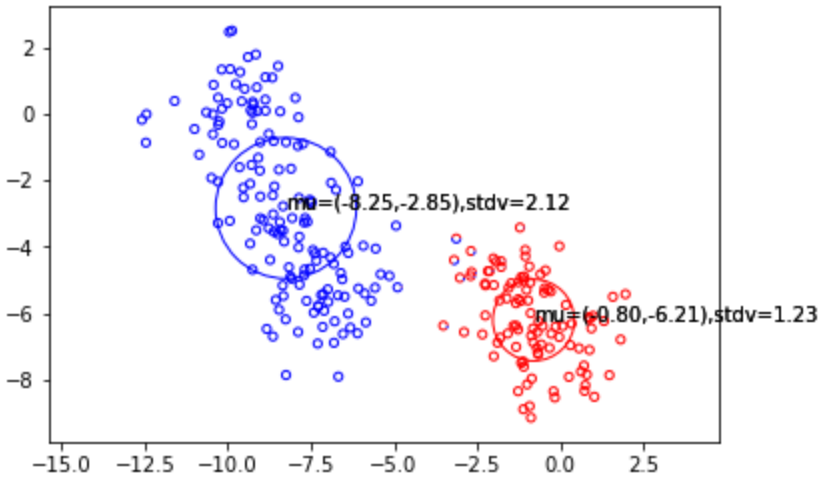


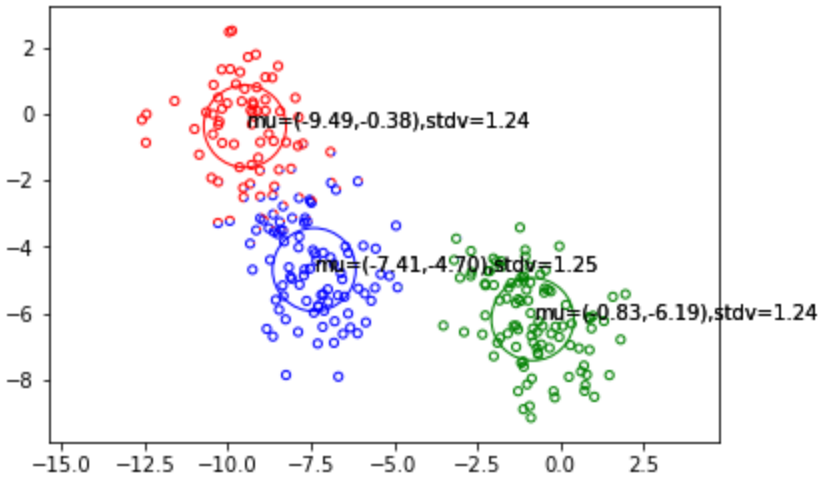


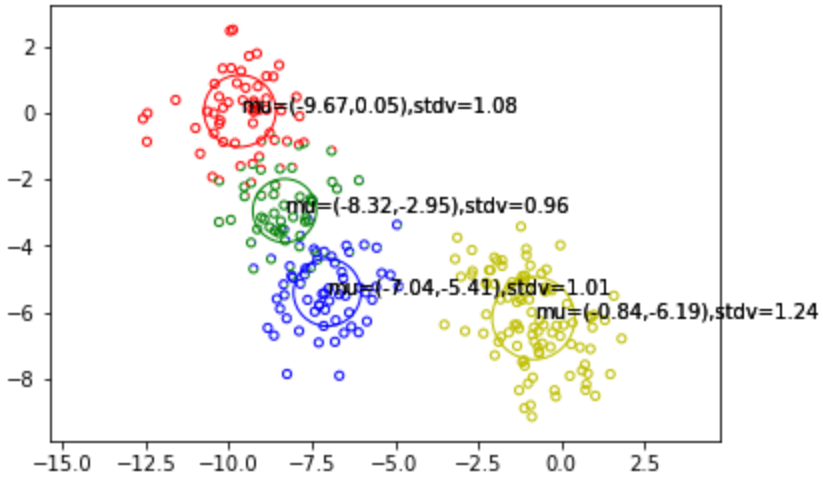
2. GMM implemented
3. Tests Passed
4. L values:

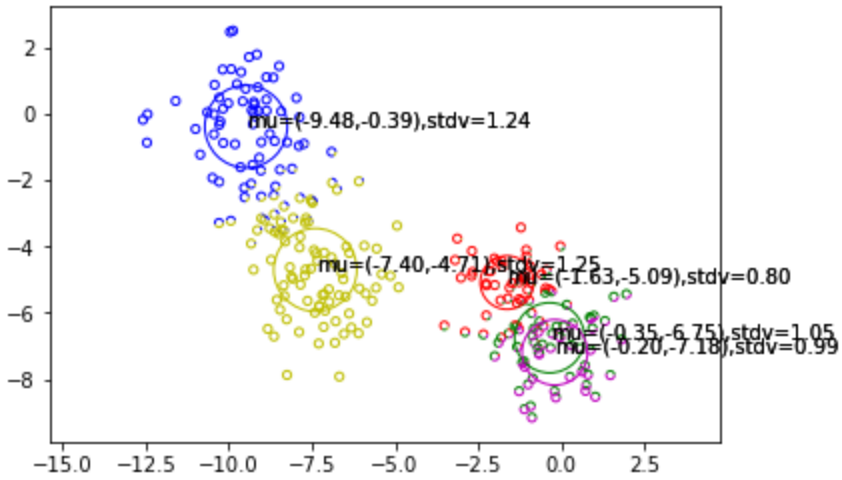
a. -1315.3
b. -1139.7
c. -1072.5
d. -1059.1
e. -1058.8











5. The clustering of K-means and EM are very similar but differ in how they divide at k=5. Where KM creates less overlap in the clustering, EM has heavy overlap in the bottom left grouping of points. The KM seems to be better fit for k=5, as there is a much more consistent decrease in error/cost from k=4 to k=5. For EM there is little change in cost from k=4 to k=5. The heavy overlap in the EM clustering is due to the EM soft-assigning/assigning with probability points to clusters, while the K-means clustering hard-assigns every point to a single cluster. This enables overlapping in the EM algorithm.

(2) Clustering Census Data

1. Redundant features artificially increase the weight of the features, increasing the distance between clusters based on these features. This increased distance creates larger margins of separation of data based on these redundant features, making clusters more confident than they would be with non-redundant feature sets.

$$p(X, z|\pi, a) = \prod_{i}^{N} p(z|\pi) \prod_{d}^{D} p(x_d|a_{kd})$$

$$\sum_{x_d\ missing} \prod_{i}^{N} p(z|\pi) \prod_{d}^{D} p(x_d|a_{kd}) = \prod_{i}^{N} p(z|\pi) \sum_{x_d\ missing} \prod_{d}^{D} p(x_d|a_{kd}) = \prod_{i}^{N} p(z|\pi)$$

So missing (NaN) should be handled as multiplying by 1

2. E-Step Posterior distribution

$$p(z|x,\pi,a) = p(x,\pi,a|z)p(z)/p(x,\pi,a)$$
$$p(z) = \pi[k]$$
$$p(x,\pi,a|z) = \prod_d^D p(x_d|a_{kd}) = \prod_d^D a_{kd}[x]$$
$$p(x,\pi,a) = \sum_i^N p(x,\pi,a|z)p(z) = \sum_i^N \pi[k] \prod_d^D a_{kd}[x]$$

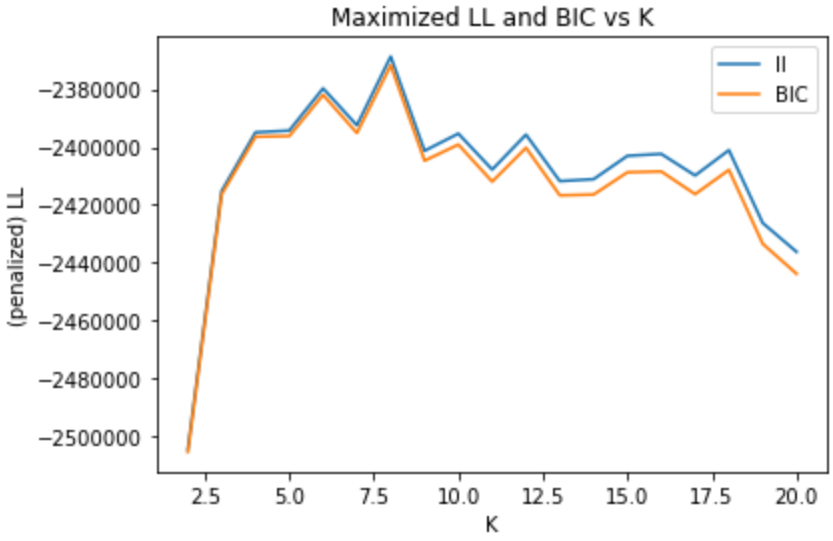
3. M-Step Updates
   a. Pi Update

$$n_j = \sum_i^N p(z=k|x)$$
$$\pi_j = n_j/N$$

   b. Alpha Update

$$a_{dk} = p(z=k|x)x / \sum_i^N p(z=k|x)x$$
$$= p_{zT} \bullet dum(data[:,d]) / \sum_i^N p_{zT} \bullet dum(data[:,d])$$

4. Implemented in Project3.py
5. Max LL is maximum at k = 8 with -2.37e6 and minimum at k=2 with -2.50e6. Otherwise it hovers around -2.4e6.
6. Choosing Model
   a. Implemented BIC. Counting free parameters with np-1 instead of np as the np parameters must sum to 1, and so the last parameter is not independent.
   b. K = 8 is the best for both LL and BIC. They do agree.



7. Results
   a. For K=8 clusterings. Clusters seem primarily separated by age, gender, and income.
      i. 0-12 male of european ancestry

ii. 0-12 female with hispanic ancestry. Born in America (not US)
iii. 13-19 male of european ancestry
iv. 20-29 male of hispanic ancestry. Born in America (not US). HS degree. Private employer. <$15k.
v. 20-29 female of european ancestry. HS degree. Private employer. <$15k.
vi. 30-39 male of european ancestry. HS degree. Private employer. $30-$60k.
vii. 30-39 female of european ancestry. Naturalized. HS degree. Private employer. <$15k.
viii. >65 female of european ancestry. HS degree. <$15k.

b. For K=6 clusterings. Demographics are relatively similar, and all age groups are still represented. There are only minor changes between groups that have the same age/gender.
   i. 0-12 male of european ancestry
   ii. 13-19 female of european ancestry
   iii. 20-29 female of european ancestry. HS degree. Private employer. <$15k.
   iv. 30-39 male of european ancestry. HS degree. Private employer. $30-$60k.
   v. 30-39 female of hispanic ancestry. Born in America (not US). HS degree. Private employer. <$15k.
   vi. >65 female of european ancestry. HS degree. <$15k.
c. The age groups are the most stable features. Income is also very stable in relation to age and gender. Every cluster above the age of 19 has a high school education. All employers are n/a or for profit. Ancestry is primarily western european or hispanic. Overall the clusters are very stable.
d. When used on states, the clusters could be compared to determine the largest group of people in a state. Ex: having different age'd clusters would indicate varying age between the state's populations.