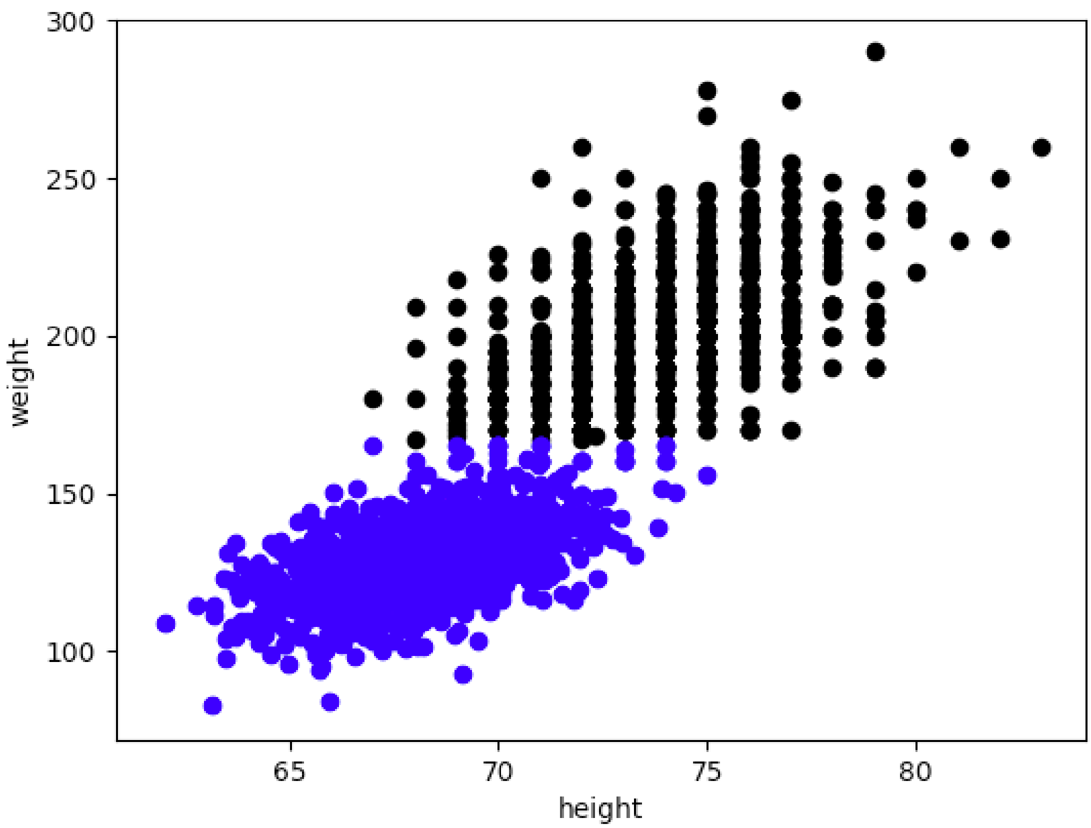
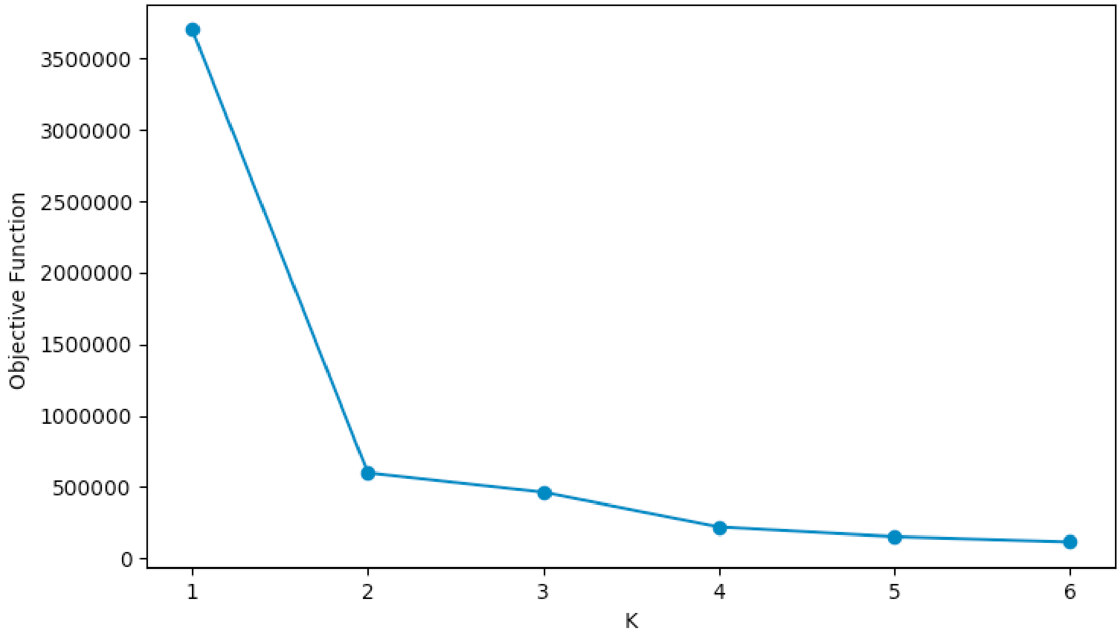
Brandon Peck
Machine Learning
Assignment 6

1)
Q3) 2-Means Cluster Scatter



The 2-means scatter initializes two nodes randomly and iteratively moves these nodes until they have reached the “center of gravity” of closest points and no nods move across clusters in two successive iterations.

Q4) Objective Function



This is calculated by taking the squared difference from each point in the data to their respective cluster centroids, and summing these values for each cluster. This gives an idea of which cluster number, k, results in the least error.

Q5)
Cluster 1 purity – 0.999001
Cluster 2 purity – 0.972498
This is calculated by taking the maximum number of same output nodes in a cluster and dividing that number by the total nodes in the cluster. This gives an idea of how "pure" a cluster is.

2) Questions

1)
(a)
Here the x's represent the two cluster centroids. Boundaries are decided based on which nodes are closes to each of the centroids. So, the dashed line shows any nodes on one side of the line would belong to the cluster whose centroid is on that side of the line. The centroids are random initially, then divide based on which nodes are closest to them. They then move to the centers of the nodes that are closest to them. This continues iteratively until the centroids will no longer exchange nodes in their clusters.

2)

(a)

There are 3 centroids since there are 3 partitions. Out of 6 positions to choose from there are 3 positions that will be filled. This results in 6 choose 3.

(b)

No – Dividing in this way will cause the centroids to move on the next iteration	None - There is no such way we can place starting configurations and partition the nodes. Since {a,b,e} will divide the {c,d} partition	0
Yes - Because of the min distances to the centroids have been achieved so they will not move	{b,c,e} f is closest to c d is closest to e a is closest to b	4 {b,c,e} {a,d,f} {a,d,c} {b,e,f}
Yes - Because of the min distances to the centroids have been achieved so they will not move	{a,b,c} d is closest to a e is closest to b f is closest to c	8 {a,b,c} {a,e,f} {a,e,c} {a,b,f} {d,e,f} {d,e,c} {d,b,f} {d,b,c}
Yes - Because of the min distances to the centroids have been achieved so they will not move	{a,b,d} a is closest to a d is closest to d e,f,c are closest to b	2 {a,d,e} {a,d,b}
Yes - Because of the min distances to the centroids can be achieved in some way so they will not move	None – We cannot place centroids on the nodes and achieve an outcome	0
Yes - Because of the min distances to the centroids have been achieved so they will not move	{a,c,f} a is closest to d and b c is closest to c f is closest to e	1 {a,c,f}

3)

(a)

$H(Emotion|Wig = Y) = -(P(\text{S}|\text{Y})\log_2(\text{P}(\text{S}|\text{Y}) + P(\text{H}|\text{Y})\log_2(\text{P}(\text{H}|\text{Y})) = -(.5\ln(.5) + .5\ln(.5)) = 1$

(b)

$H(Emotion|Ears = 3) = -(P(\text{S}|3)\log_2(\text{P}(\text{S}|3) + P(\text{H}|3)\log_2(\text{P}(\text{H}|3)) = -(0 + 0) = 0$

(c)

Color most evenly divides over the output.

(d)

The drawing given is correct because if the color is green we can predict sad and if it is red we can predict happy. The only case where the output divides is over blue. We can then split the output half and half over blue since there is one sad and one happy instance for two blue inputs.

(e)

With two outputs the training error maximum is 50% because this is what we could by randomly choosing an output. If our model has greater error than this we can always use the trivial random model.

(f)

4)

(a)

$H(Edible|Odor = 1\ or\ Odor = 3) = -(P(\text{Yes}|\ 1\text{ or }3)\log_2(\text{P}(\text{Yes}|1\text{ or }3) + P(\text{No}|1\text{ or }3)\log_2(\text{P}(\text{No}|1\text{ or }3)) = -(.5\log_2(.5) + .5\log_2(.5)) = 1$

(b)

Odor because this splits the output on 1 and 3 and 2 is the only value for which the output is divided in a 2:3 = yes:no ratio.

(C)

The figure uses odor as the root and must divide the output over the value 2. It chooses color as the next level because color has more features which can narrow the output. Finally shape divide the output when color is blue.