exam2

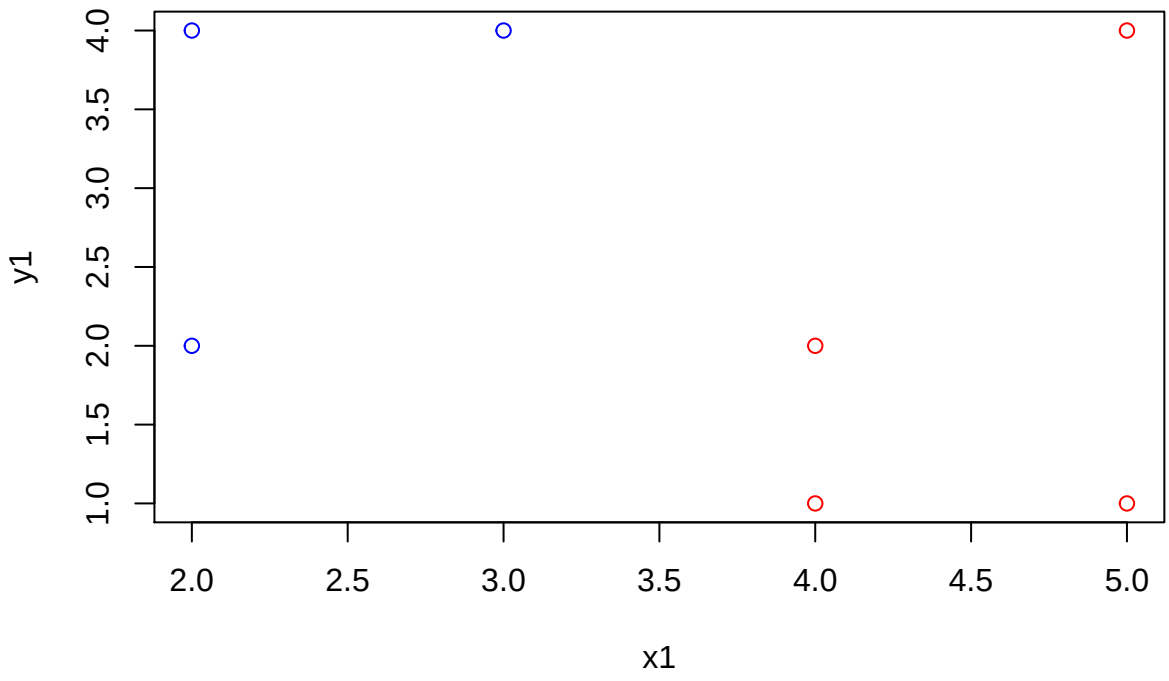
Sunny Lee

4/9/2021

1)

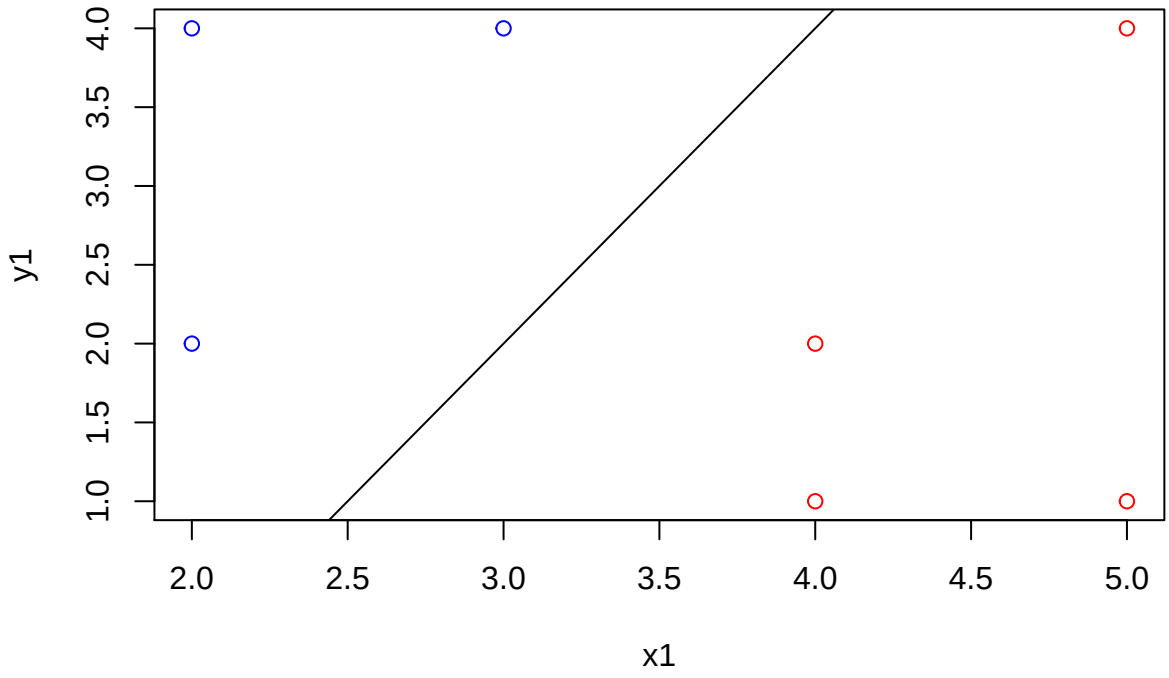
a)

```
x1 <- c(4, 5, 2, 4, 2, 3, 5)
y1 <- c(1, 1, 2, 2, 4, 4, 4)
plot(x1, y1, col = c("red", "red", "blue" ,"red","blue","blue", "red"))
abline(2, -4)
```



b)

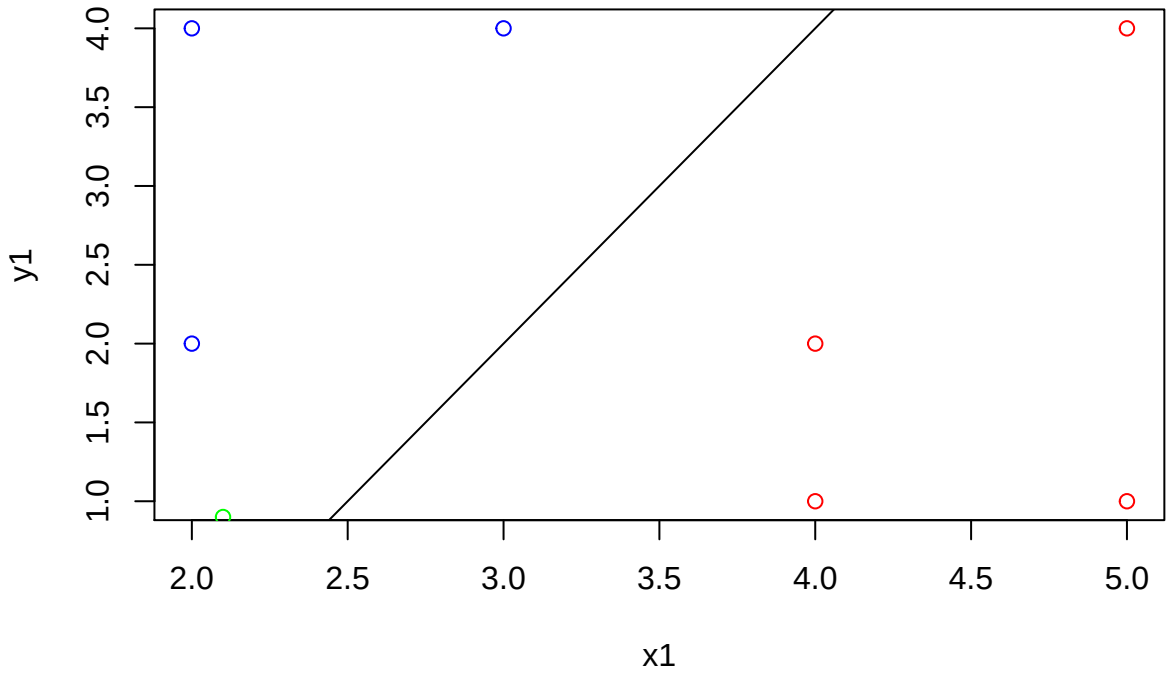
```
plot(x1, y1, col = c("red", "red", "blue" ,"red","blue","blue", "red"))
abline(-4, 2)
```



c) The equation of the hyperplane in the form $\beta_0 + \beta_1 X_1 + \beta_2 X_2 = 0$ is $4 - 2X_1 + X_2 = 0$.

d) Algebraically, when we plug in our new observation, we get: $4 - 2(2.1) + .9 = .7$, and since our value is positive, we will assign the new observation into the Blue Group. Plotting this new observation, we come to the same conclusion as we can clearly see that the new observation is above the hyperplane where the blue observations are.

```
plot(x1, y1, col = c("red", "red", "blue" ,"red","blue","blue", "red"))
points(2.1, .9, col = "green")
abline(-4, 2)
```



2)

a) Following the tree diagram, we have the years greater than 4.5, the hits is greater than 117.5, walks is greater than 52.5, RBI is less than 80.5 and years is less than 6.5, thus we would predict the salary of this new player would be 6.459 million dollars.

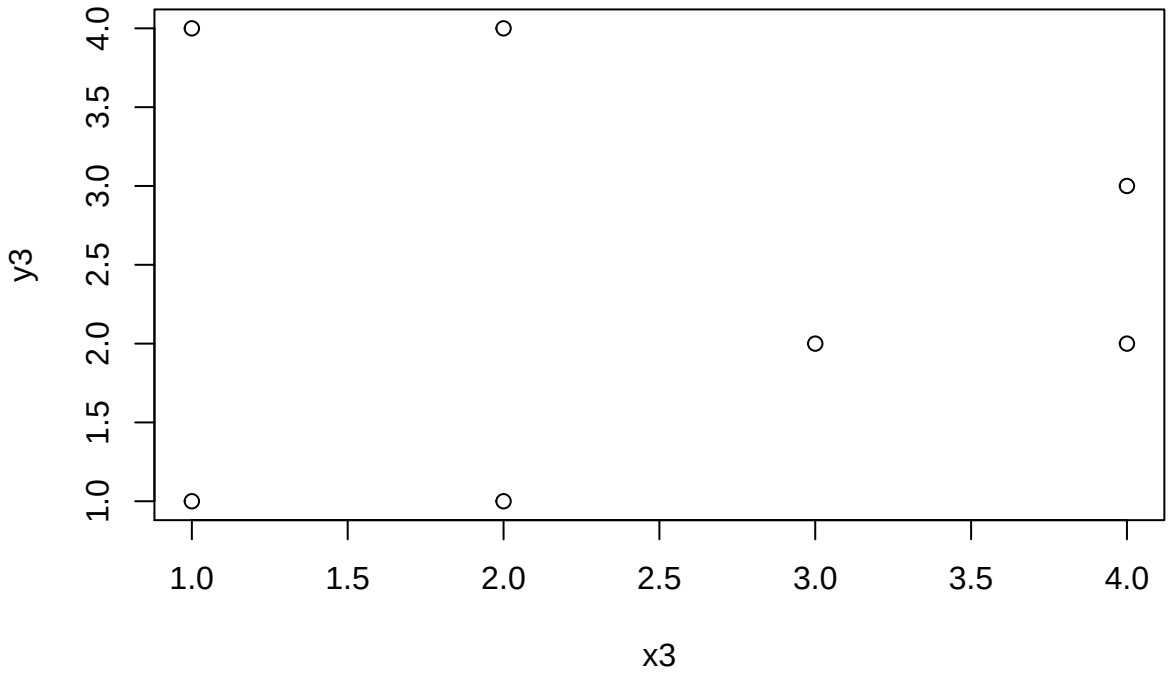
b) The recursive binary split which gives the greatest reduction in RSS is the first split, where we test whether Years is less than 4.5.

c) Yes, as instead of continuous variables in the terminal nodes, we would use categorical variables which would lead the tree to output classify a new observation rather than predict a new observation.

3)

a)

```
x3 <- c(1, 2, 3, 4, 1, 2, 4)
y3 <- c(1, 1, 2, 2, 4, 4, 3)
plot(x3, y3)
```



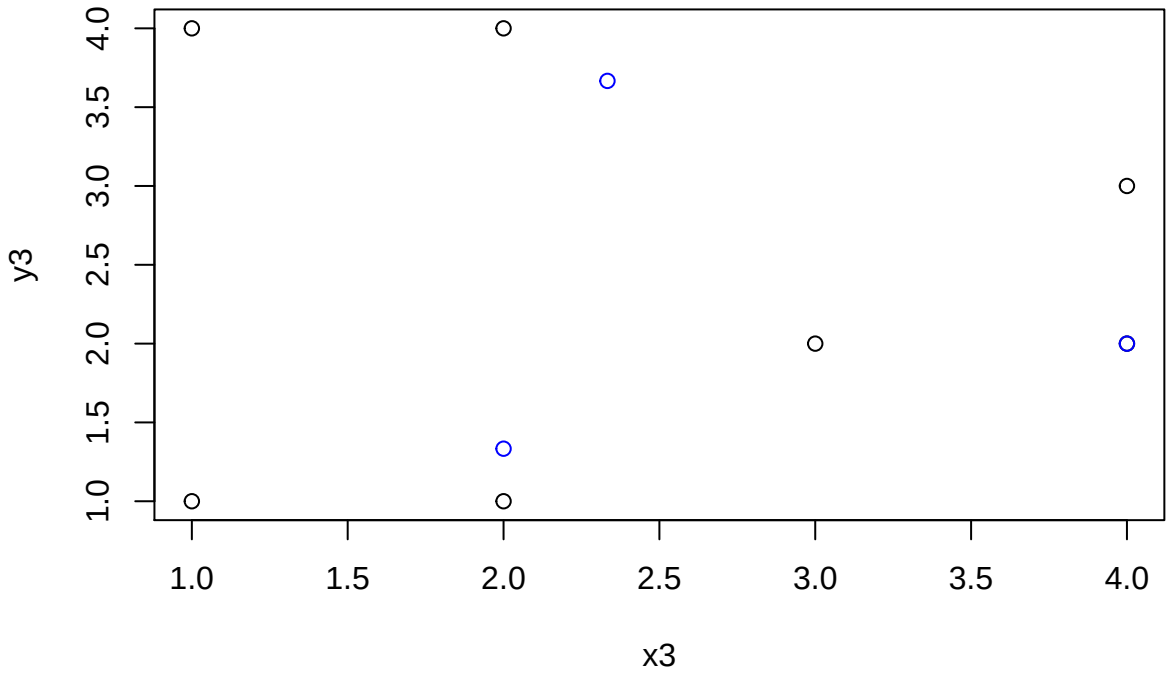

```
centroid_x <- c(mean(x3[1:3]), mean(x3[5:7]), mean(x3[4]))
centroid_y <- c(mean(y3[1:3]), mean(y3[5:7]), mean(y3[4]))
centroid_x
```

```
## [1] 2.000000 2.333333 4.000000
```

```
centroid_y
```

```
## [1] 1.333333 3.666667 2.000000
```

```
plot(x3, y3)
points(centroid_x, centroid_y, col = "blue")
```



From the initialization, we find that C_1 is the point $(2, 1.33)$, C_2 is the point $(2.33, 3.67)$ and C_3 is the point $(4, 2)$. Performing the first clustering algorithm with these initialized points:

```
new_centroid <- rep(0, 7)
for (i in 1:7)
{
  dist <- rep(0, 3)
  for (j in 1:3)
  {
    dist[j] <- sqrt((x3[i] - centroid_x[j])^2 + (y3[i] - centroid_y[j])^2)
  }
  new_centroid[i] <- which.min(dist)
}

new_centroid
```

```
## [1] 1 1 3 3 2 2 3
```

and we see that observations 1 and 2 are grouped to centroid 1, observations 3, 4, 7 are grouped to centroid 3 and obsevations 5, 6 are grouped to centroid 2. Calculating where the new centroids will be:

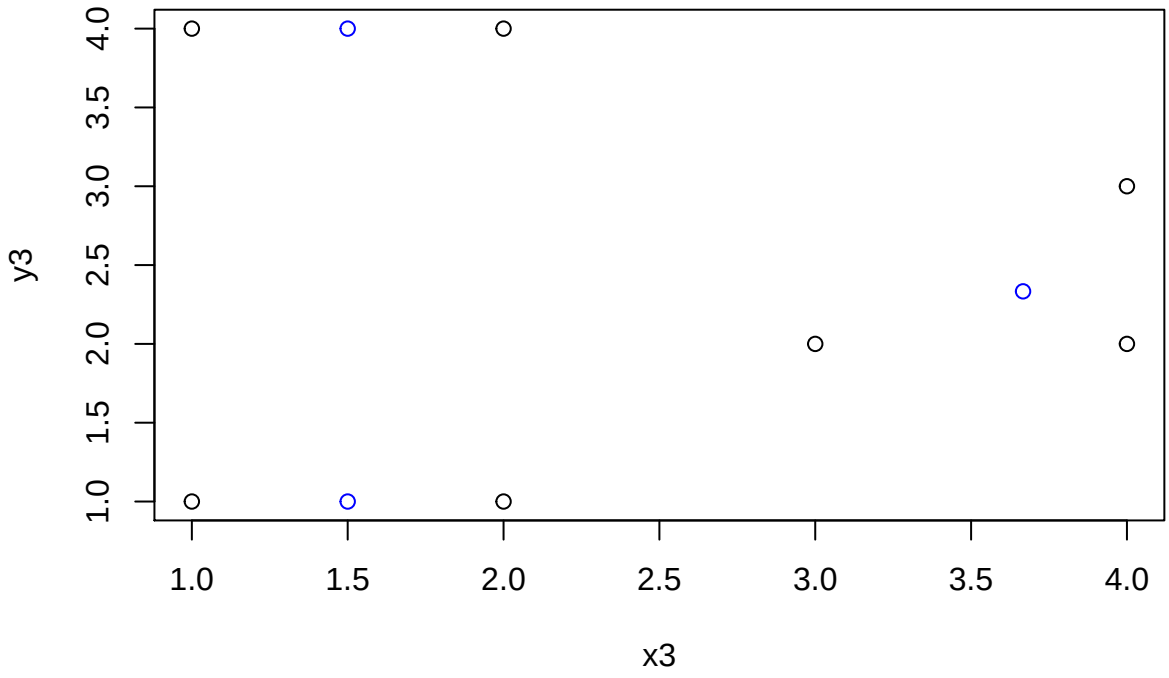
```
centroid_x <- c(mean(x3[1:2]), mean(x3[5:6]), mean(x3[c(3, 4, 7)]))
centroid_y <- c(mean(y3[1:2]), mean(y3[5:6]), mean(y3[c(3, 4, 7)]))
centroid_x
```

```
## [1] 1.500000 1.500000 3.666667
```

```
centroid_y
```

```
## [1] 1.000000 4.000000 2.333333
```

```
plot(x3, y3)
points(centroid_x, centroid_y, col = "blue")
```



We see that the new centroids are: $C_1 = (1.5, 1)$, $C_2 = (1.5, 4)$, $C_3 = (3.67, 2.33)$, and when we run the next iteration, we see no change in the groups, and thus our algorithm ends.

```
new_centroid <- rep(0, 7)
for (i in 1:7)
{
  dist <- rep(0, 3)
  for (j in 1:3)
  {
    dist[j] <- sqrt((x3[i] - centroid_x[j])^2 + (y3[i] - centroid_y[j])^2)
  }
  new_centroid[i] <- which.min(dist)
}

new_centroid
```

```
## [1] 1 1 3 3 2 2 3
```

b) This is unsupervised learning, as we do not have a specific response variable.