# lab_08

Alexis Saldivar (PID: A69038083)

## Table of contents

### Background

something

### Libraries

```
library(ggplot2)
```

### Data import

Remove patient id and diagnosis from the dataset.

```
cancer_raw_df <- read.csv("WisconsinCancer.csv", row.names=1)
new_samples <- read.csv("new_samples.csv")
```

### Data Cleaning

Remove diagnosis column from the dataset

```
diagnosis <- as.factor(cancer_raw_df$diagnosis)
cancer_df <- cancer_raw_df[,-1]
```

### Data summary

```
obsvs <- nrow(cancer_df)
vars <- ncol(cancer_df)
vars_w_mean <- length(grep("_mean$", names(cancer_df), value=TRUE))
n_malignant <- sum(diagnosis=="M")
sprintf("Q1. There are %s obsvervations in the dataset", obsvs)
```

```
[1] "Q1. There are 569 obsvervations in the dataset"
```

```
sprintf("Q2. There are %s malignant diagnosis", n_malignant)
```

```
[1] "Q2. There are 212 malignant diagnosis"
```

```
sprintf("Q3. There are %s vars with suffix _mean", vars_w_mean)
```

```
[1] "Q3. There are 10 vars with suffix _mean"
```

## PCA

```
round(colMeans(cancer_df), 4)
```

```
            radius_mean            texture_mean          perimeter_mean
                14.1273                 19.2896                 91.9690
              area_mean         smoothness_mean        compactness_mean
               654.8891                  0.0964                  0.1043
         concavity_mean     concave.points_mean           symmetry_mean
                 0.0888                  0.0489                  0.1812
 fractal_dimension_mean               radius_se              texture_se
```

```
                 0.0628                  0.4052                  1.2169 
           perimeter_se                 area_se           smoothness_se 
                 2.8661                 40.3371                  0.0070 
         compactness_se            concavity_se       concave.points_se 
                 0.0255                  0.0319                  0.0118 
            symmetry_se    fractal_dimension_se            radius_worst 
                 0.0205                  0.0038                 16.2692 
          texture_worst         perimeter_worst              area_worst 
                25.6772                107.2612                880.5831 
       smoothness_worst       compactness_worst         concavity_worst 
                 0.1324                  0.2543                  0.2722 
   concave.points_worst          symmetry_worst fractal_dimension_worst 
                 0.1146                  0.2901                  0.0839 
```

```r
apply(cancer_df, 2, sd)
```

```
            radius_mean            texture_mean          perimeter_mean 
           3.524049e+00            4.301036e+00            2.429898e+01 
              area_mean         smoothness_mean        compactness_mean 
           3.519141e+02            1.406413e-02            5.281276e-02 
         concavity_mean     concave.points_mean           symmetry_mean 
           7.971981e-02            3.880284e-02            2.741428e-02 
 fractal_dimension_mean               radius_se              texture_se 
           7.060363e-03            2.773127e-01            5.516484e-01 
           perimeter_se                 area_se           smoothness_se 
           2.021855e+00            4.549101e+01            3.002518e-03 
         compactness_se            concavity_se       concave.points_se 
           1.790818e-02            3.018606e-02            6.170285e-03 
            symmetry_se    fractal_dimension_se            radius_worst 
           8.266372e-03            2.646071e-03            4.833242e+00 
          texture_worst         perimeter_worst              area_worst 
           6.146258e+00            3.360254e+01            5.693570e+02 
       smoothness_worst       compactness_worst         concavity_worst 
           2.283243e-02            1.573365e-01            2.086243e-01 
   concave.points_worst          symmetry_worst fractal_dimension_worst 
           6.573234e-02            6.186747e-02            1.806127e-02 
```

```r
cancer_pca <- prcomp(cancer_df, scale=TRUE)
cancer_pca_summary <- summary(cancer_pca)
pc1_var <- cancer_pca_summary$importance[2, 1]
sprintf("Q4. PC1 explains %s of the variance", pc1_var)
```

```
[1] "Q4. PC1 explains 0.44272 of the variance"
```

```
# Get the PCs that explain 70% of the variance
sum_var <- 0
pc_n <- 0
while(sum_var<=0.7) {
  pc_n <- pc_n + 1
  sum_var <- sum_var + cancer_pca_summary$importance[2, pc_n]
}
sprintf("Q5. The first %s PCs explain %s of the variance", pc_n, sum_var)
```

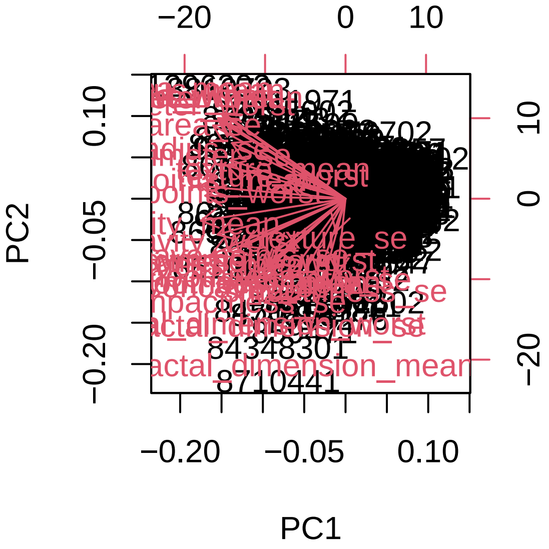
```
[1] "Q5. The first 3 PCs explain 0.72636 of the variance"
```

```
# Get the PCs that explain 90% of the variance
sum_var <- 0
pc_n <- 0
while(sum_var<=0.9) {
  pc_n <- pc_n + 1
  sum_var <- sum_var + cancer_pca_summary$importance[2, pc_n]
}
sprintf("Q6. The first %s PCs explain %s of the variance", pc_n, sum_var)
```

```
[1] "Q6. The first 7 PCs explain 0.9101 of the variance"
```

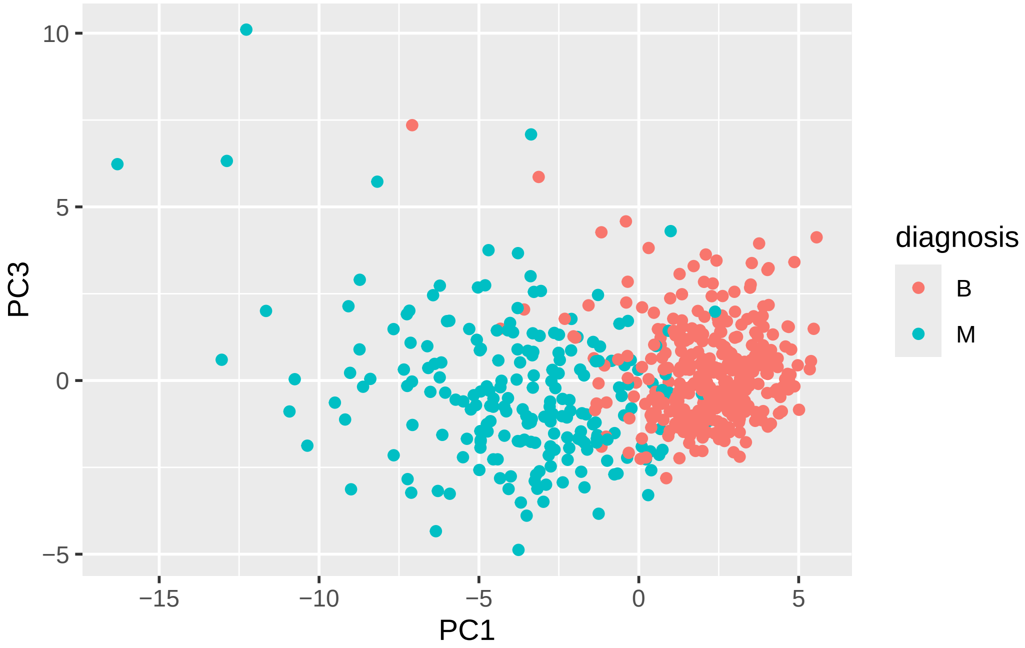
Q7.

```
biplot(cancer_pca)
```

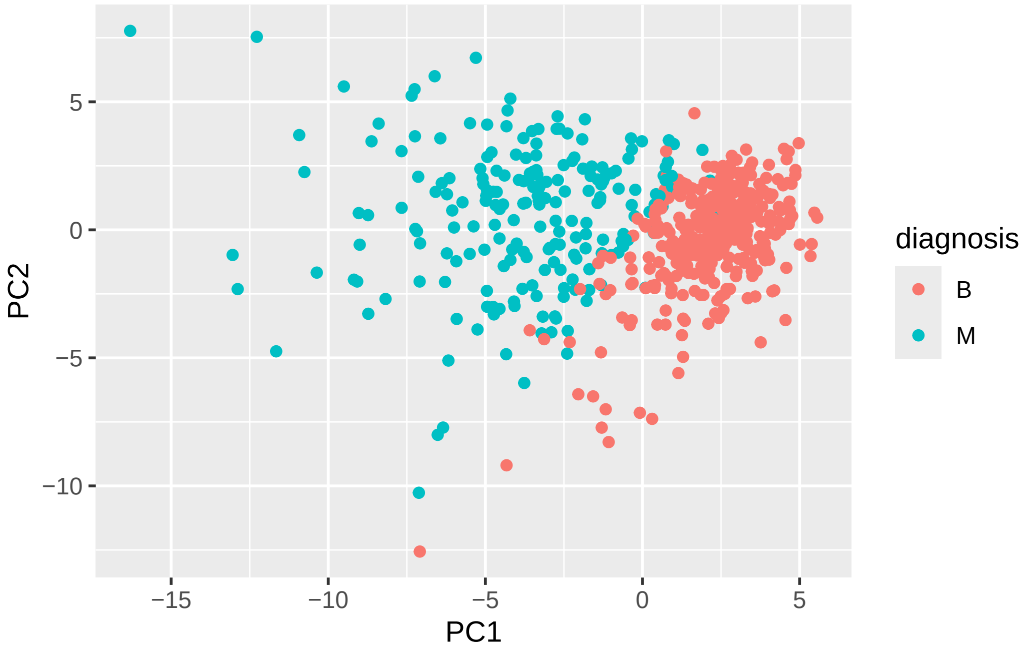


Q8.

```
ggplot(cancer_pca$x) +
  aes(PC1, PC3, col=diagnosis) +
  geom_point()
```



```
ggplot(cancer_pca$x) +
  aes(PC1, PC2, col=diagnosis) +
  geom_point()
```



```
lo <- cancer_pca$rotation["concave.points_mean", 1]
sprintf("Q9. The contribution of concave.points_mean to PC1 is %s", round(lo,2))
```

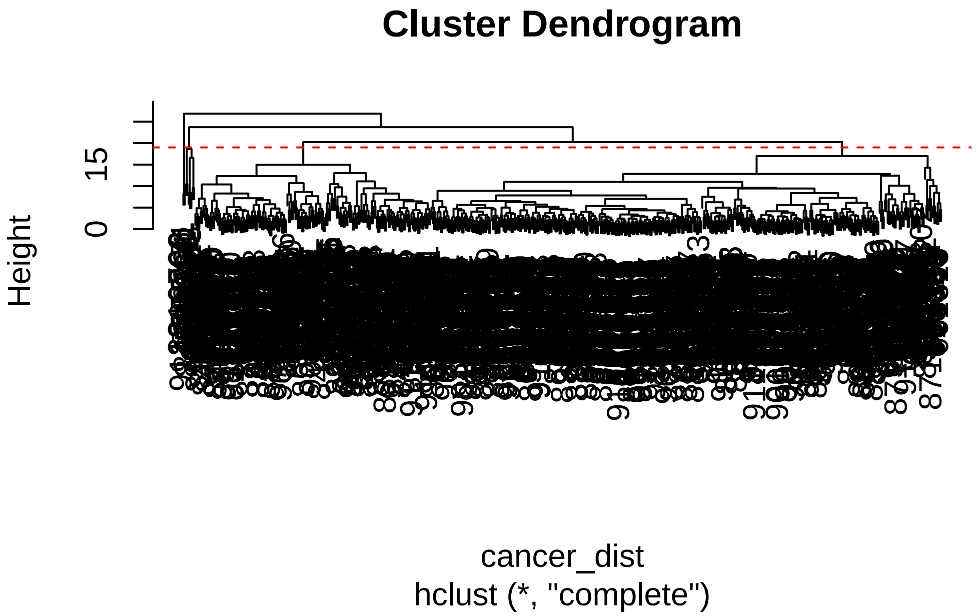
```
[1] "Q9. The contribution of concave.points_mean to PC1 is -0.26"
```

### Hierarchical Clustering

Q10. what is the height at which the clustering model has 4 clusters?

19

```
cancer_df_scaled <- scale(cancer_df)
cancer_dist <- dist(cancer_df_scaled, method="euclidian")
cancer_hclust <- hclust(cancer_dist, method="complete")
plot(cancer_hclust)
abline(19, 0, col="red", lty=2)
```



```
cancer_hclust_cut <- cutree(cancer_hclust, 2)
table(cancer_hclust_cut, diagnosis)
```

```
                 diagnosis
cancer_hclust_cut   B   M
                1 357 210
                2   0   2
```

```
cancer_dist <- dist(cancer_df_scaled, method="euclidian")
cancer_hclust <- hclust(cancer_dist, method="ward.D2")
cancer_hclust_cut <- cutree(cancer_hclust, 2)
table(cancer_hclust_cut, diagnosis)
```

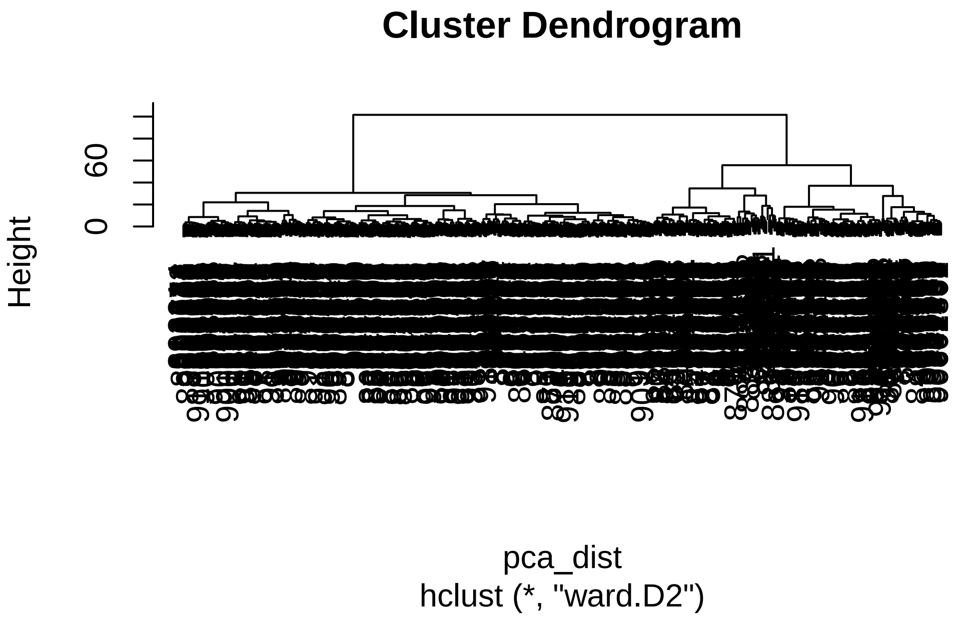
```
                 diagnosis
cancer_hclust_cut   B   M
                1  20 164
                2 337  48
```

> Q12. Which method gives your favorite results for the same data.dist dataset? Explain your reasoning.

Ward.D2, because it split the largest number of B and M diagnosis into different clusters.

## Clustering PCA

```
pca_dist <- dist(cancer_pca$x[,1:7], method="euclidian")
pca_hclust <- hclust(pca_dist, method="ward.D2")
plot(pca_hclust)
```



> Q13. How well does the newly created model with four clusters separate out the two diagnoses?

This model has a higher True positive value

```
pca_hclust_cut <- cutree(pca_hclust, 2)
table(pca_hclust_cut, diagnosis)
```

```
              diagnosis
pca_hclust_cut   B   M
             1  28 188
             2 329  24
```

```
sensitivity <- round(188 / (188 + 28), 2)
specificity <- round(324 / (324 + 24), 2)
sprintf("sensitivity: %s, specificity: %s", sensitivity, specificity)
```

```
[1] "sensitivity: 0.87, specificity: 0.93"
```

> Q14. How well do the hierarchical clustering models you created in previous sections (i.e. before PCA) do in terms of separating the diagnoses?

This model has less True positive but also less false positives

```r
cancer_dist <- dist(cancer_df_scaled, method="euclidian")
cancer_hclust <- hclust(cancer_dist, method="ward.D2")
cancer_hclust_cut <- cutree(cancer_hclust, 2)
table(cancer_hclust_cut, diagnosis)
```

```
                 diagnosis
cancer_hclust_cut   B   M
                1  20 164
                2 337  48
```

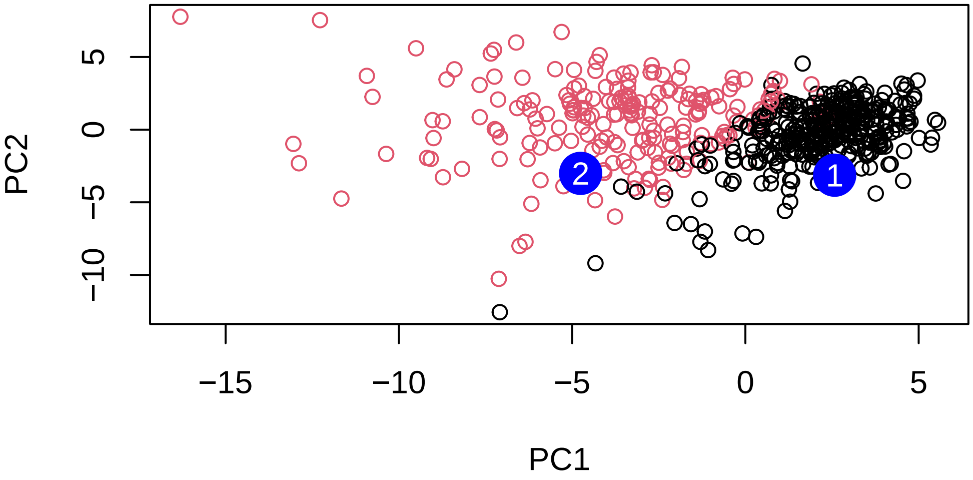
```r
sensitivity <- round(164 / (164 + 20), 2)
specificity <- round(337 / (337 + 48), 2)
sprintf("sensitivity: %s, specificity: %s", sensitivity, specificity)
```

```
[1] "sensitivity: 0.89, specificity: 0.88"
```

### Prediction

```r
url <- "https://tinyurl.com/new-samples-CSV"
new <- read.csv(url)
npc <- predict(cancer_pca, newdata=new)
```

```r
plot(cancer_pca$x[,1:2], col=diagnosis)
points(npc[,1], npc[,2], col="blue", pch=16, cex=3)
text(npc[,1], npc[,2], c(1,2), col="white")
```



Q16. Which of these new patients should we prioritize for follow up based on your results?

Patient 2 should be prioritized.