Problem 8

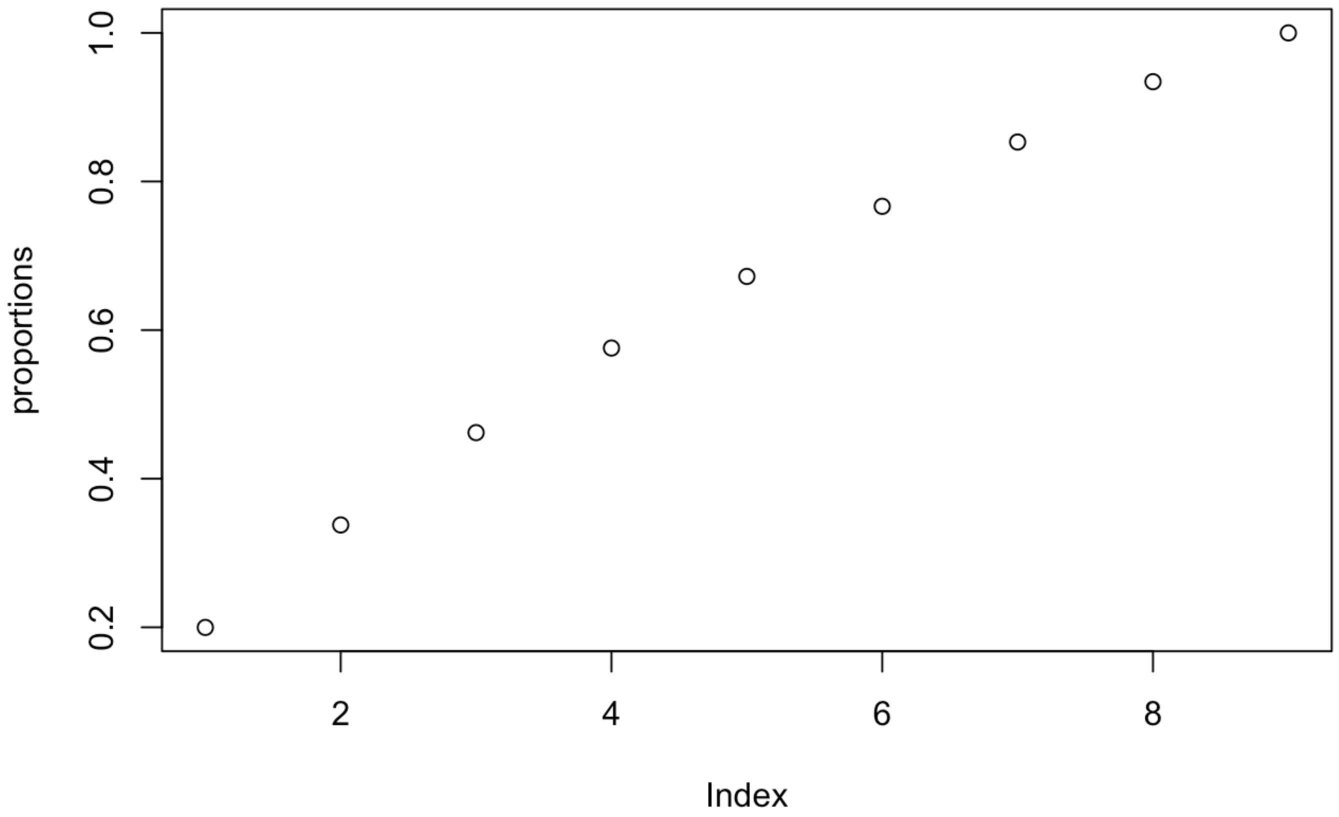
Part 1

```
training_features <- read.csv('training_features.csv')

# Replace data with a new data frame with all NAs imputed with column medians
training_features <- as.data.frame(lapply(training_features, function(x) {x[is.na(x)] <- median(x, na.rm=TRUE); x}))

# Exclude subject.id 525450 from the analysis
training_features <- training_features[training_features$subject.id != 525450,]

training_features <- training_features[,c("q1_speech.slope", "q2_salivation.slope", "q3_swallowing.slope", "q4_handwriting.slope", "q5a_cutting_without_gastrostomy.slope", "q6_dressing_and_hygiene.slope", "q7_turning_in_bed.slope", "q8_walking.slope", "q9_climbing_stairs.slope")]

pca = prcomp(training_features, scale=TRUE)
```

Part 2

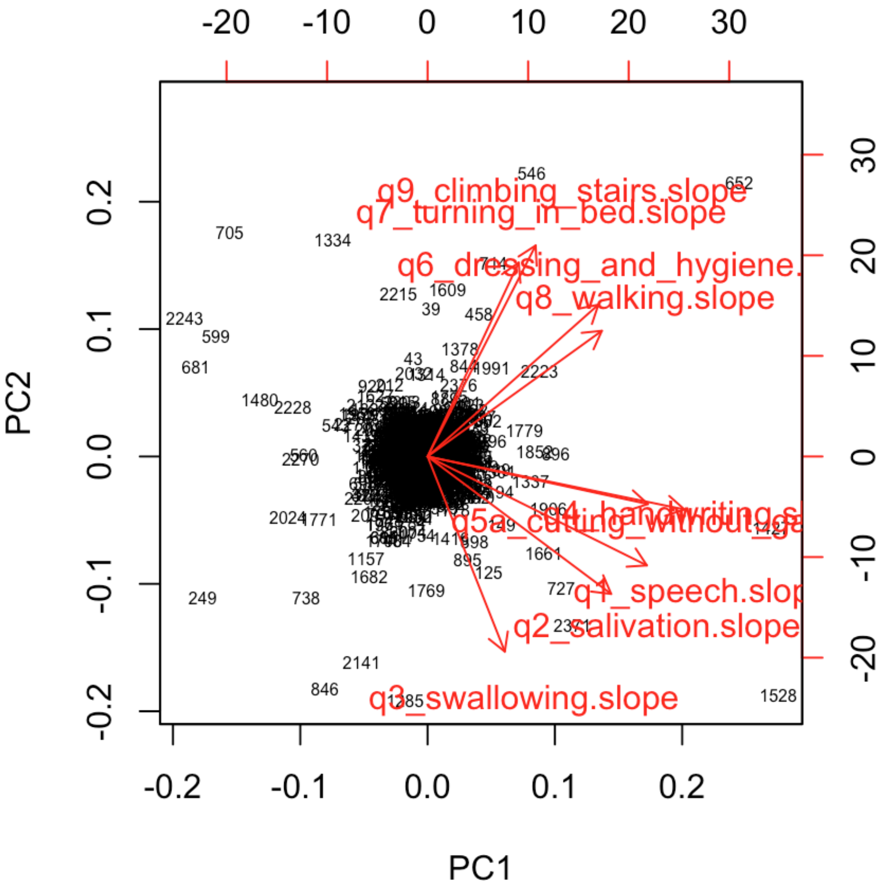
No, the top 2 principal components do not capture most of the variance.

```
vars <- apply(pca$x, 2, var)
props <- vars / sum(vars)
proportions = cumsum(props)
plot(proportions)
```



Part 3

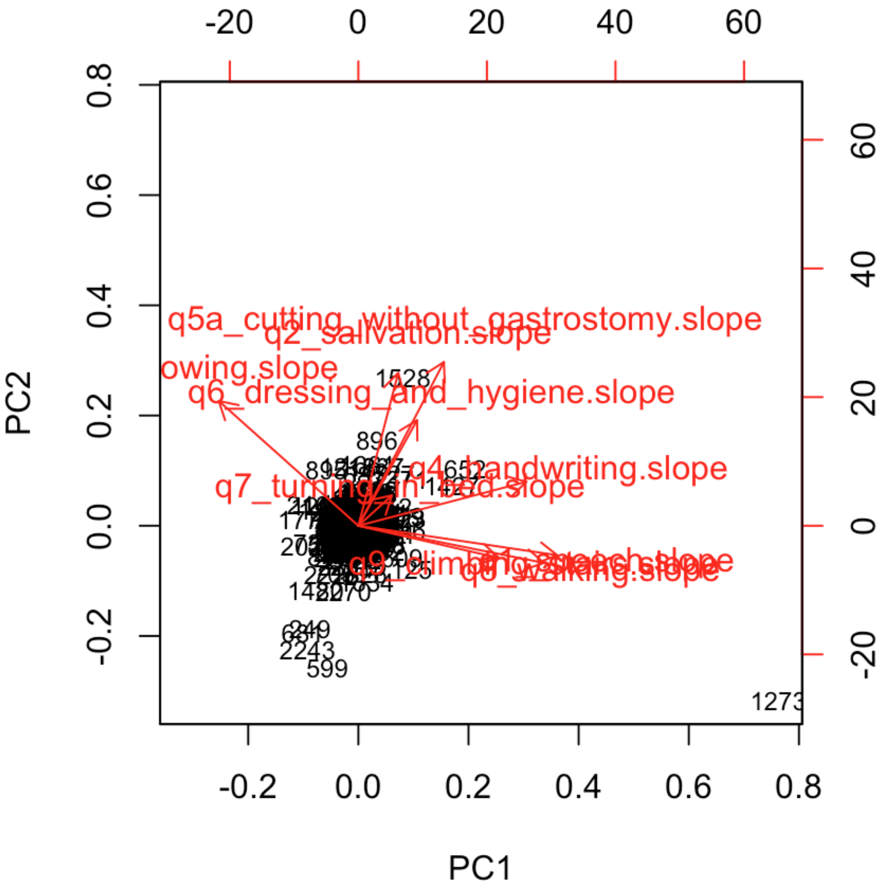
```
biplot(pca, scale=1, cex=c(1/2, 1))
```



Part 4

Patient #525450 is a major outlier from the rest of the data. It is so different that it completely changes the direction of the principal component!

```
training_features <- read.csv('training_features.csv')

# Replace data with a new data frame with all NAs imputed with column medians
training_features <- as.data.frame(lapply(training_features, function(x) {x[is.na(x)] <- median(x, na.rm=TRUE); x}))

training_features <- training_features[,c("q1_speech.slope", "q2_salivation.slope", "q3_swallowing.slope", "q4_handwriting.slope", "q5a_cutting_without_gastrostomy.slope", "q6_dressing_and_hygiene.slope", "q7_turning_in_bed.slope", "q8_walking.slope", "q9_climbing_stairs.slope")]

pca = prcomp(training_features, scale=TRUE)
biplot(pca, scale=1, cex=c(3/4, 1))
```


-20
0
20
40
60
0.8
0.6
0.4
0.2
0.0
-0.2
60
40
20
0
-20
PC2
q5a_cutting_without_gastrostomy.slope
q2_salivation.slope
owing.slope
q6_dressing_and_hygiene.slope
1528
896
q4_handwriting.slope
652
q7_turning_in_bed.slope
1427
q9_climbing_stairs.slope
q1_speech.slope
q8_walking.slope
125
1482
2270
249
631
2243
599
1273
-0.2
0.0
0.2
0.4
0.6
0.8
PC1