Class 8

AUTHOR

Charlie Rezanka (A15837296)

Breast Cancer Project

Today we are going to explore some breast cancer data from the University of Wisconsin

```
wisc.data <- read.csv("WisconsinCancer.csv", row.names=1)
head(wisc.data)</pre>
```

	diagnosis	radius_mean	texture_mean	perimeter_mean	area_mean	
842302	М	17.99	10.38	122.80	1001.0	
842517	М	20.57	17.77	132.90	1326.0	
84300903	М	19.69	21.25	130.00	1203.0	
84348301	М	11.42	20.38	77.58	386.1	
84358402	М	20.29	14.34	135.10	1297.0	
843786	М	12.45	15.70	82.57	477.1	
	smoothness	_mean compa	ctness_mean c	oncavity_mean co	oncave.poir	nts_mean
842302	0.	11840	0.27760	0.3001		0.14710
842517	0.	08474	0.07864	0.0869		0.07017
84300903	0.	10960	0.15990	0.1974		0.12790
84348301	0.	14250	0.28390	0.2414		0.10520
84358402	0.	10030	0.13280	0.1980		0.10430
843786	0.	12780	0.17000	0.1578		0.08089
	symmetry_m	ean fractal	_dimension_mea	an radius_se te	kture_se pe	erimeter_se
842302	0.2	419	0.078	71 1.0950	0.9053	8.589
842517	0.1	812	0.056	67 0.5435	0.7339	3.398
84300903	0.2	.069	0.0599	99 0.7456	0.7869	4.585
84348301	0.2	597	0.097	44 0.4956	1.1560	3.445
84358402	0.1	809	0.058	83 0.7572	0.7813	5.438
843786	0.2	.087	0.076	13 0.3345	0.8902	2.217
	area_se sm	oothness_se	compactness_	se concavity_se	concave.po	oints_se
842302	153.40	0.006399	0.049	0.05373		0.01587
842517	74.08	0.005225	0.013	0.01860		0.01340
84300903	94.03	0.006150	0.040	0.03832		0.02058
84348301	27.23	0.009110	0.074	0.05661		0.01867
84358402	94.44	0.011490	0.024	0.05688		0.01885
843786	27.19	0.007510	0.033	45 0.03672		0.01137
	symmetry_s	e fractal_di	imension_se ra	adius_worst text	ture_worst	
842302	0.0300	3	0.006193	25.38	17.33	
842517	0.0138	9	0.003532	24.99	23.41	
84300903	0.0225	0	0.004571	23.57	25.53	
84348301	0.0596	3	0.009208	14.91	26.50	
84358402	0.0175	6	0.005115	22.54	16.67	
843786	0.0216	5	0.005082	15.47	23.75	
	perimeter_	worst area_v	worst smoothn	ess_worst compa	tness_wors	st

localhost:4368

5/2/23, 12:20 PM				Class 8	
842302	184.60	2019.0		0.1622	0.6656
842517	158.80	1956.0		0.1238	0.1866
84300903	152.50	1709.0		0.1444	0.4245
84348301	98.87	567.7		0.2098	0.8663
84358402	152.20	1575.0		0.1374	0.2050
843786	103.40	741.6		0.1791	0.5249
	concavity_worst	concave.poi	ints_worst	symmetry_worst	
842302	0.7119		0.2654	0.4601	
842517	0.2416		0.1860	0.2750	
84300903	0.4504		0.2430	0.3613	
84348301	0.6869		0.2575	0.6638	
84358402	0.4000		0.1625	0.2364	
843786	0.5355		0.1741	0.3985	
	fractal_dimension	on_worst			
842302		0.11890			
842517		0.08902			
84300903		0.08758			
84348301		0.17300			
84358402		0.07678			
843786		0.12440			

Q. How many patient samples are in this dataset?

```
nrow(wisc.data)
```

[1] 569

There are 569 patients in this dataset.

Q. How many patients have malignant tumors?

```
table(wisc.data$diagnosis)
```

в м 357 212

There are 357, 212 with malignant tumors.

Save the diagnosis for later use as a reference to comapare how well we do with PCA etc.

```
diagnosis <- as.factor(wisc.data$diagnosis)
#diagnosis</pre>
```

Now exclude the diagnosis column from the data.

localhost:4368 2/12

t to head(wisc) when rednering because it drops a ridiculous amount of data into the PDF otherwise

radius_mean text 17.99 20.57 19.69 11.42 20.29 12.45	ure_mean perim 10.38 17.77 21.25 20.38	122.80 132.90 130.00	rea_mean s 1001.0 1326.0	moothne	ess_mean 0.11840		
17.99 20.57 19.69 11.42 20.29	10.38 17.77 21.25	122.80 132.90	1001.0	moothne			
20.57 19.69 11.42 20.29	17.77 21.25	132.90			0.11840		
19.69 11.42 20.29	21.25		1226 A				
11.42 20.29		130 00			0.08474		
20.29	20 38		1203.0		0.10960		
	20.30	77.58	386.1		0.14250		
12.45	14.34	135.10	1297.0		0.10030		
12.19	15.70	82.57	477.1		0.12780		
ompactness_mean	concavity_mea	n concave.po	ints_mean	symmet	try_mean		
0.27760	0.300	1	0.14710)	0.2419		
0.07864	0.086	9	0.07017	,	0.1812		
0.15990	0.197	4	0.12790)	0.2069		
0.28390	0.241	4	0.10520)	0.2597		
0.13280	0.198	0	0.10430)	0.1809		
0.17000	0.157	8	0.08089)	0.2087		
ractal_dimensio	n_mean radius_	se texture_s	se perimet	er_se a	area_se		
0	.07871 1.09	50 0.905	i3	8.589	153.40		
0	.05667 0.54	35 0.733	19	3.398	74.08		
0	.05999 0.74	56 0.786	59	4.585	94.03		
0	.09744 0.49	56 1.156	50	3.445	27.23		
0	.05883 0.75	72 0.781	.3	5.438	94.44		
0	.07613 0.33	45 0.890)2	2.217	27.19		
moothness_se co	mpactness_se c	oncavity_se	concave.p	oints_s	se		
0.006399	0.04904	0.05373		0.0158	37		
0.005225	0.01308	0.01860		0.0134	10		
0.006150	0.04006	0.03832		0.020	58		
0.009110	0.07458	0.05661		0.0186	57		
0.011490	0.02461	0.05688		0.0188	35		
0.007510	0.03345	0.03672		0.0113	37		
ymmetry_se frac	tal_dimension_	se radius_wo	rst textu	re_wors	st		
0.03003	0.0061	93 25	5.38	17.3	33		
0.01389	0.0035	32 24	1.99	23.4	1 1		
0.02250				25.5	53		
0.05963							
0.01756							
0.02165							
184.60	2019.0	_	•	_			
	1956.0						
	•						
	0.13280 0.17000 fractal_dimension 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0	0.13280 0.198 0.17000 0.157 ractal_dimension_mean radius_ 0.07871 1.09 0.05667 0.54 0.05999 0.74 0.09744 0.49 0.05883 0.75 0.07613 0.33 moothness_se compactness_se compa	0.13280 0.1980 0.17000 0.1578 ractal_dimension_mean radius_se texture_s 0.07871 1.0950 0.905 0.05667 0.5435 0.733 0.05999 0.7456 0.786 0.09744 0.4956 1.156 0.05883 0.7572 0.781 0.07613 0.3345 0.896 moothness_se compactness_se concavity_se 0.006399 0.04904 0.05373 0.005225 0.01308 0.01860 0.006150 0.04006 0.03832 0.009110 0.07458 0.05661 0.011490 0.02461 0.05688 0.007510 0.03345 0.03672 ymmetry_se fractal_dimension_se radius_wc 0.03003 0.006193 25 0.01389 0.003532 24 0.02250 0.004571 23 0.05963 0.009208 14 0.05963 0.009208 14 0.01756 0.005115 22 0.02165 0.005082 15 erimeter_worst area_worst smoothness_wors 184.60 2019.0 0.162 158.80 1956.0 0.123 152.50 1709.0 0.144 98.87 567.7 0.205 152.20 1575.0 0.137 103.40 741.6 0.179 oncavity_worst concave.points_worst symmetry_worst concave.points_worst symmetry 0.7119 0.2654 0.2416 0.1860 0.4504 0.2430	0.13280 0.1980 0.10438 0.17000 0.1578 0.08089 ractal_dimension_mean radius_se texture_se perimet	0.13280 0.1980 0.10430 0.17000 0.1578 0.08089 ractal_dimension_mean radius_se texture_se perimeter_se a	0.13280 0.1980 0.10430 0.1809 0.17000 0.1578 0.08089 0.2087 ractal_dimension_mean radius_se texture_se perimeter_se area_se 0.07871 1.0950 0.9053 8.589 153.40 0.05667 0.5435 0.7339 3.398 74.08 0.05999 0.7456 0.7869 4.585 94.03 0.09744 0.4956 1.1560 3.445 27.23 0.05883 0.7572 0.7813 5.438 94.44 0.07613 0.3345 0.8902 2.217 27.19 moothness_se compactness_se concavity_se concave.points_se 0.006399 0.04904 0.05373 0.01587 0.005225 0.01308 0.01860 0.01340 0.006150 0.04006 0.03832 0.02058 0.009110 0.07458 0.05661 0.01867 0.011490 0.02461 0.05688 0.01885 0.007510 0.03345 0.03672 0.01137 ymmetry_se fractal_dimension_se radius_worst texture_worst 0.03003 0.006193 25.38 17.33 0.01389 0.003532 24.99 23.41 0.02250 0.004571 23.57 25.53 0.05963 0.009208 14.91 26.50 0.01756 0.00515 22.54 16.67 0.02165 0.005082 15.47 23.75 erimeter_worst area_worst smoothness_worst compactness_worst 184.60 2019.0 0.1622 0.6656 158.80 1956.0 0.1238 0.1866 152.50 1709.0 0.1444 0.4245 98.87 567.7 0.2098 0.8663 152.20 1575.0 0.1374 0.2050 0.01340 0.2460 0.1791 0.5249 oncavity_worst concave.points_worst symmetry_worst 0.7119 0.2654 0.4601 0.2416 0.1860 0.2750 0.4504 0.2430 0.3613	0.13280 0.1980 0.10430 0.1809 0.17000 0.1578 0.08089 0.2087 ractal_dimension_mean radius_se texture_se perimeter_se area_se 0.07871 1.0950 0.9053 8.589 153.40 0.05667 0.5435 0.7339 3.398 74.08 0.05999 0.7456 0.7869 4.585 94.03 0.09744 0.4956 1.1560 3.445 27.23 0.05883 0.7572 0.7813 5.438 94.44 0.07613 0.3345 0.8902 2.217 27.19 moothness_se compactness_se concavity_se concave.points_se 0.006399 0.04904 0.05373 0.01587 0.005225 0.01308 0.01860 0.01340 0.006150 0.04006 0.03832 0.02058 0.009110 0.07458 0.05661 0.01867 0.011490 0.02461 0.05688 0.01885 0.007510 0.03345 0.03672 0.01137 ymmetry_se fractal_dimension_se radius_worst texture_worst 0.03003 0.006193 25.38 17.33 0.01389 0.003532 24.99 23.41 0.02250 0.004571 23.57 25.53 0.05963 0.009208 14.91 26.50 0.01756 0.005815 22.54 16.67 0.02165 0.005082 15.47 23.75 erimeter_worst area_worst smoothness_worst compactness_worst 184.60 2019.0 0.1622 0.6656 158.80 1956.0 0.1238 0.1866 152.50 1709.0 0.1444 0.4245 98.87 567.7 0.2098 0.8663 152.20 1575.0 0.1374 0.2050 0.0719 0.2554 0.1374 0.2050 0.07119 0.2654 0.4601 0.2416 0.1860 0.2750 0.4504 0.2430 0.3613

localhost:4368 3/12

5/2/23, 12:20 PM			Class 8
84358402	0.4000	0.1625	0.2364
843786	0.5355	0.1741	0.3985
	fractal_dimension_wors	t	
842302	0.1189	9	
842517	0.0890	2	
84300903	0.08758	3	
84348301	0.17300	9	
84358402	0.07678	3	
843786	0.1244	9	

Q. How many "dimensions", "variables", "columns" are present in the dataset?

```
ncol(wisc)
```

[1] 30

There are 30 dimensions in the dataset.

Principal Component Analysis

To perform PCA in R we can use the prcomp() function, it takes as input a numeric dataset and an optional scale=FALSE/TRUE argument.

We generally always want to set scale=TRUE but lets make sure by checking if the mean and standard deviation values are different accross these 30 columns.

round(colMeans(wisc))

perimeter_mean	texture_mean	radius_mean
92	19	14
compactness_mean	smoothness_mean	area_mean
0	0	655
symmetry_mean	concave.points_mean	concavity_mean
0	0	0
texture_se	radius_se	<pre>fractal_dimension_mean</pre>
1	0	0
smoothness_se	area_se	perimeter_se
0	40	3
concave.points_se	<pre>concavity_se</pre>	compactness_se
0	0	0
radius_worst	<pre>fractal_dimension_se</pre>	symmetry_se
16	0	0
area_worst	perimeter_worst	texture_worst
881	107	26
concavity_worst	compactness_worst	smoothness_worst
0	0	0
<pre>fractal_dimension_worst</pre>	symmetry_worst	concave.points_worst
0	0	0

localhost:4368 4/12

```
pca <- prcomp(wisc, scale=TRUE)
summary(pca)</pre>
```

```
Importance of components:
```

```
PC1
                                 PC2
                                         PC3
                                                  PC4
                                                          PC5
                                                                  PC6
                                                                          PC7
Standard deviation
                       3.6444 2.3857 1.67867 1.40735 1.28403 1.09880 0.82172
Proportion of Variance 0.4427 0.1897 0.09393 0.06602 0.05496 0.04025 0.02251
Cumulative Proportion 0.4427 0.6324 0.72636 0.79239 0.84734 0.88759 0.91010
                           PC8
                                  PC9
                                         PC10
                                                 PC11
                                                         PC12
                                                                 PC13
                                                                         PC14
Standard deviation
                       0.69037 0.6457 0.59219 0.5421 0.51104 0.49128 0.39624
Proportion of Variance 0.01589 0.0139 0.01169 0.0098 0.00871 0.00805 0.00523
Cumulative Proportion 0.92598 0.9399 0.95157 0.9614 0.97007 0.97812 0.98335
                                  PC16
                                          PC17
                                                   PC18
                          PC15
                                                           PC19
                                                                   PC20
                                                                          PC21
Standard deviation
                       0.30681 0.28260 0.24372 0.22939 0.22244 0.17652 0.1731
Proportion of Variance 0.00314 0.00266 0.00198 0.00175 0.00165 0.00104 0.0010
Cumulative Proportion
                       0.98649 0.98915 0.99113 0.99288 0.99453 0.99557 0.9966
                          PC22
                                  PC23
                                         PC24
                                                  PC25
                                                          PC26
                                                                  PC27
                                                                          PC28
Standard deviation
                       0.16565 0.15602 0.1344 0.12442 0.09043 0.08307 0.03987
Proportion of Variance 0.00091 0.00081 0.0006 0.00052 0.00027 0.00023 0.00005
Cumulative Proportion 0.99749 0.99830 0.9989 0.99942 0.99969 0.99992 0.99997
                          PC29
                                  PC30
Standard deviation
                       0.02736 0.01153
Proportion of Variance 0.00002 0.00000
Cumulative Proportion 1.00000 1.00000
```

```
attributes(pca)
```

```
$names
```

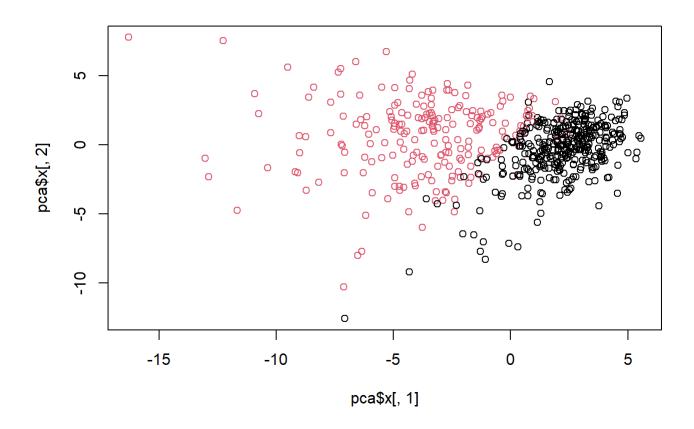
[1] "sdev" "rotation" "center" "scale" "x"

\$class

[1] "prcomp"

```
plot(pca$x[,1], pca$x[,2], col=diagnosis)
```

localhost:4368 5/12



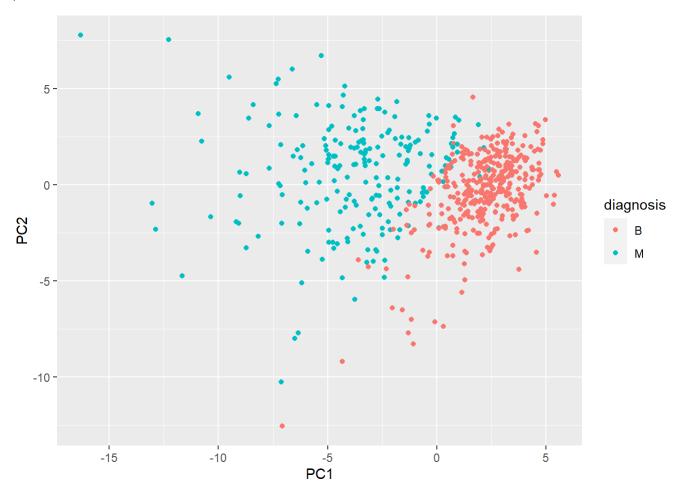
#Red points are malignant, black are red.

```
library(ggplot2)

x <- as.data.frame(pca$x)

ggplot(x) +
  aes(PC1, PC2, col=diagnosis) +
  geom_point()</pre>
```

localhost:4368 6/12



#PC1 is the component describing the most variation, PC2 describes the component with the 2nd mos

Q. How much variance is captured in the top 3 PCs?

The top 3 PCs caputure 73% of the total variance.

Q9. For the first principal component, what is the component of the loading vector (i.e. wisc.pr\$rotation[,1]) for the feature concave.points_mean? This tells us how much this original feature contributes to the first PC.

```
pca$rotation["concave.points_mean",1]
```

[1] -0.2608538

```
attributes(pca)
```

\$names

[1] "sdev" "rotation" "center" "scale" "x"

localhost:4368 7/12

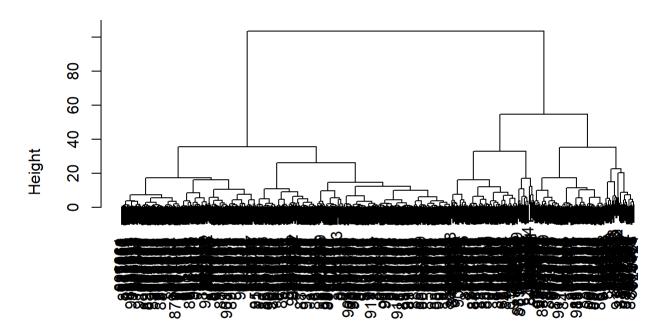
```
$class
[1] "prcomp"
```

Combine PCA results with clustering

We can use our new PCA variables (i.e. the scores along with the PCs contained in t pca\$x) as input for other methods such as clustering.

```
#Hclust needs a distance matrix as input and a method
d <- dist(pca$x[,1:3])
hc <- hclust(d, method="ward.D2")
plot(hc)</pre>
```

Cluster Dendrogram



d hclust (*, "ward.D2")

To get our cluster membership vector we can use the <code>cutree()</code> function and specify a height (h) or number of groups (k).

```
grps <- cutree(hc, h=80, k=2)
table(grps)</pre>
```

grps

1 2

localhost:4368 8/12

203 366

I want to find out how many diagnosis "M" amd "B" are in ech grp?

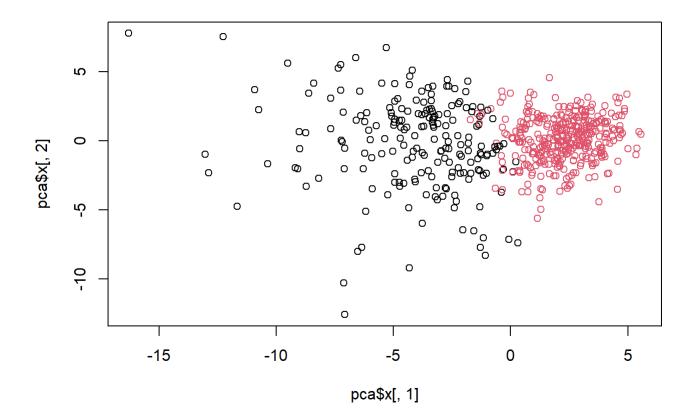
```
table(diagnosis, grps)
```

```
grps
diagnosis 1 2
B 24 333
M 179 33
```

Group 1 has 179 patients with malignant tumors, group 2 has 33.

We cna also plo our results using our clustering vector grps.

```
plot(pca$x[,1], pca$x[,2], col=grps)
```



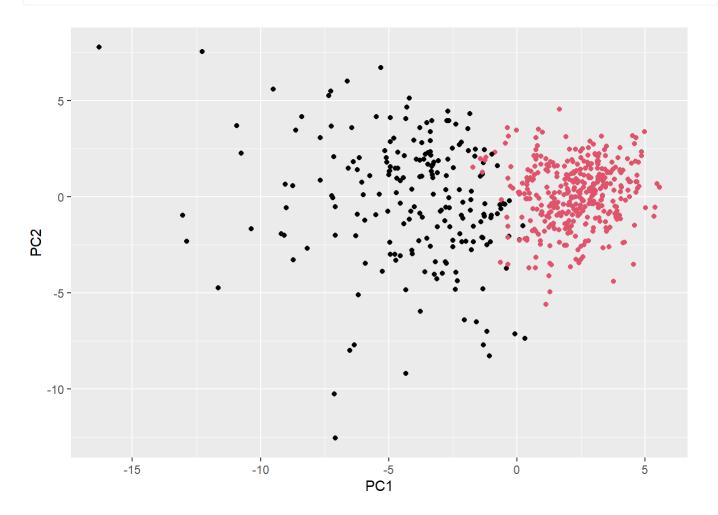
```
library(ggplot2)

x <- as.data.frame(pca$x)

ggplot(x) +</pre>
```

localhost:4368 9/12

```
aes(PC1, PC2) +
geom_point(col=grps)
```



#overlap between red and black points due to unplotted 3rd dimension

Q15. What is the specificity and sensitivity of our curent results?

The sensitivity of our current clustering is:

```
179/(179+24)
```

[1] 0.8817734

The specificity of our current clustering is:

```
333/(333+33)
```

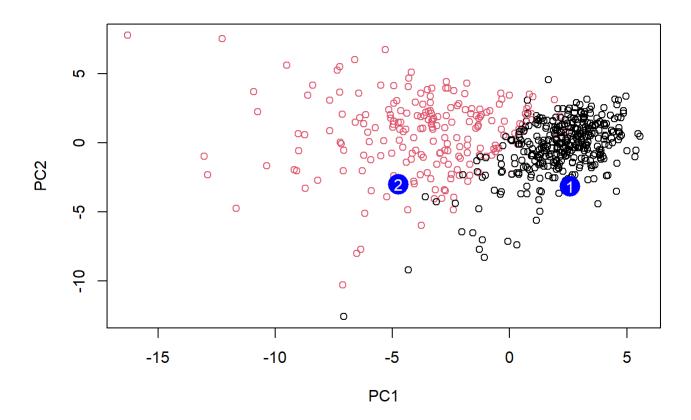
[1] 0.9098361

Prediction

localhost:4368

```
url <- "https://tinyurl.com/new-samples-CSV"</pre>
new <- read.csv(url)</pre>
npc <- predict(pca, newdata=new)</pre>
npc
                   PC2
                              PC3
                                         PC4
                                                                       PC7
          PC1
                                                  PC5
                                                             PC6
[1,] 2.576616 -3.135913 1.3990492 -0.7631950 2.781648 -0.8150185 -0.3959098
[2,] -4.754928 -3.009033 -0.1660946 -0.6052952 -1.140698 -1.2189945 0.8193031
           PC8
                    PC9
                              PC10
                                        PC11
                                                 PC12
                                                           PC13
                                                                   PC14
[1,] -0.2307350 0.1029569 -0.9272861 0.3411457 0.375921 0.1610764 1.187882
[2,] -0.3307423 0.5281896 -0.4855301 0.7173233 -1.185917 0.5893856 0.303029
         PC15
                   PC16
                               PC17
                                          PC18
                                                      PC19
                                                                 PC20
[1,] 0.3216974 -0.1743616 -0.07875393 -0.11207028 -0.08802955 -0.2495216
PC21
                    PC22
                               PC23
                                         PC24
                                                     PC25
                                                                  PC26
[1,] 0.1228233 0.09358453 0.08347651 0.1223396 0.02124121 0.078884581
[2,] -0.1224776 0.01732146 0.06316631 -0.2338618 -0.20755948 -0.009833238
            PC27
                       PC28
                                    PC29
                                                PC30
[1,] 0.220199544 -0.02946023 -0.015620933 0.005269029
[2,] -0.001134152  0.09638361  0.002795349 -0.019015820
plot(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")
```

localhost:4368 11/12



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

We should prioritize patient 2, as they are statistically more likely to be a true positive diagnosis based on their variation relative to other malignant cancer patients.

localhost:4368