Detect confounder

11.04.2022

Phuc Loi Luu, PhD

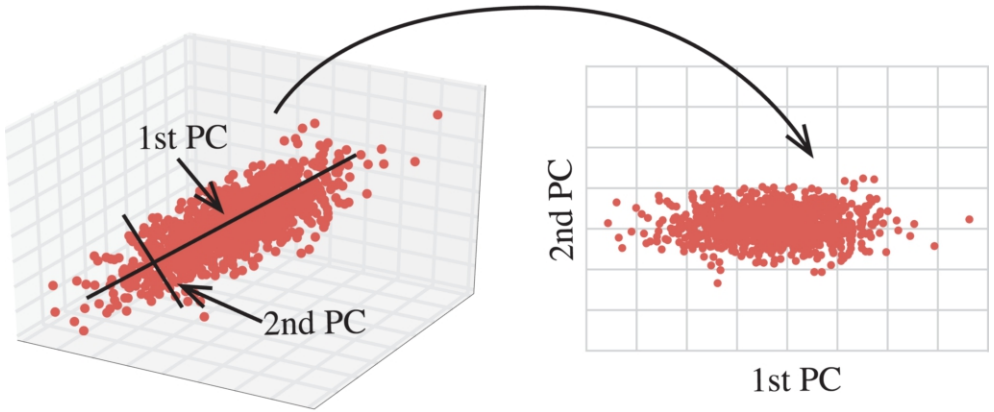
luu.p.loi@googlemail.com

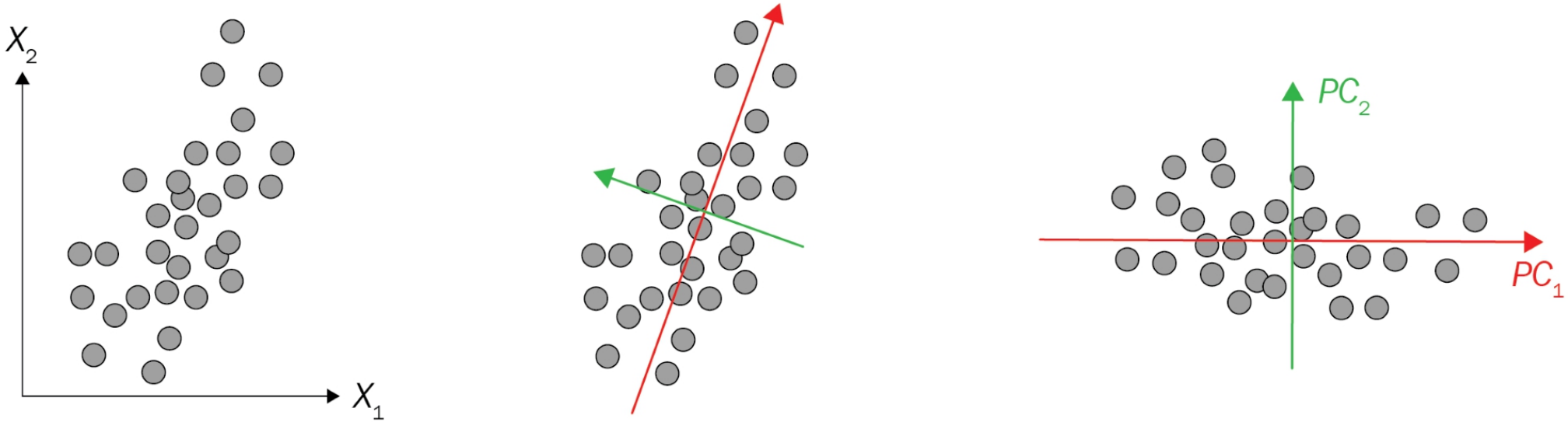
p.luu@garvan.org.au

Content

- PCA
- Variance explain
- How to calculate correlation?
- How to detect confounder?

PCA: dimension reduction method





Data in feature space → Find principal components → Data in **p**rincipal **c**omponents space

Data

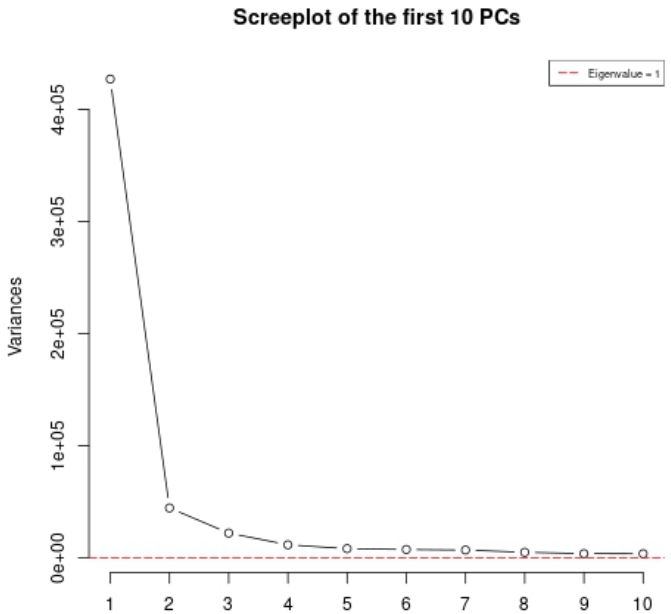
- DNA methylation beta value of 279 samples and 10000 probes (dim=10000x279)

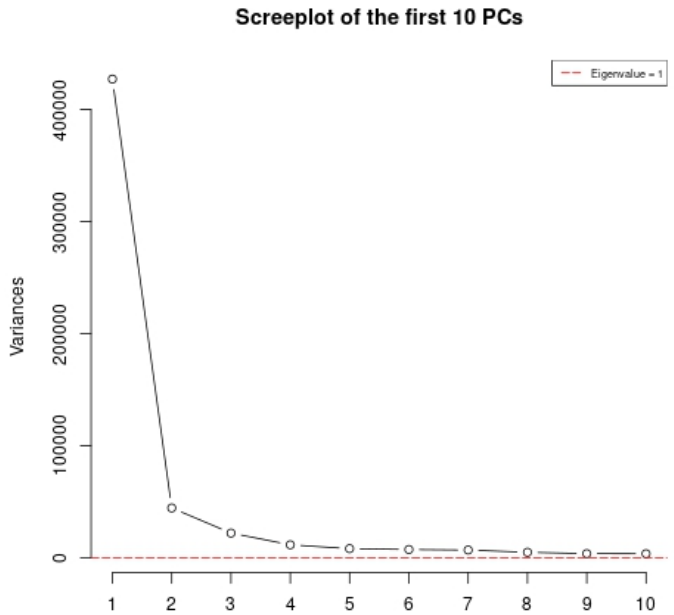
```
> head(m4[1:5,1:5])
  PC_1_EPIC_TA PC_1_EPIC_TA2 PC_1_EPIC_TB PC_1_EPIC_TB2 PC_1_EPIC_TC
1         0.36          0.31         0.32          0.36         0.30
3         0.87          0.87         0.90          0.92         0.91
4         0.92          0.91         0.90          0.90         0.90
5         0.95          0.94         0.93          0.95         0.94
6         0.10          0.12         0.12          0.12         0.13
```

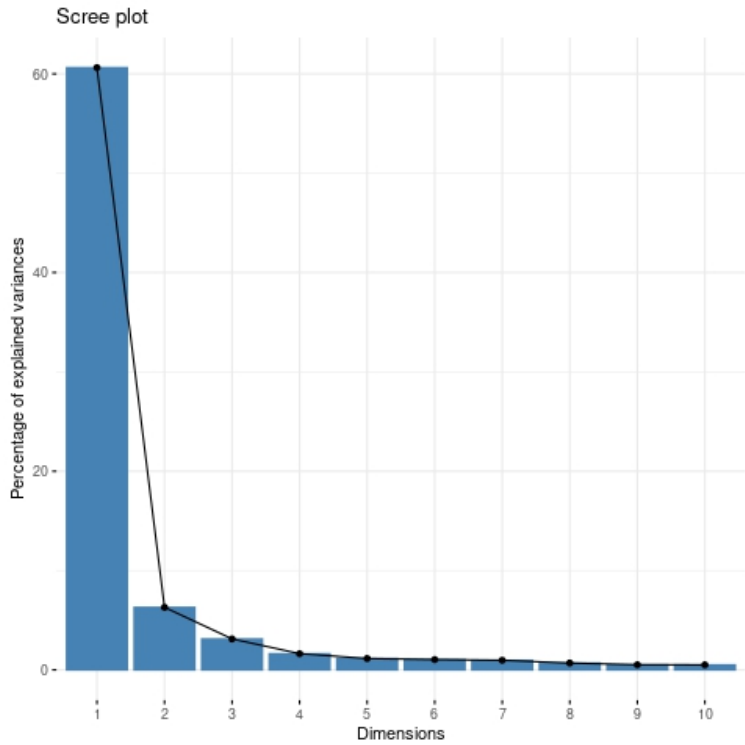
- Phenotype data of 279 patients and 6 phenotype columns (dim=279x6)

sample_ID	sex	age	race	phenotype	icud	study
PC_1_EPIC_TA	F	80	W	1	ICUd	Bernades
PC_1_EPIC_TA2	F	80	W	1	ICUd	Bernades
PC_1_EPIC_TB	F	80	W	1	non-ICUd	Bernades
PC_1_EPIC_TB2	F	80	W	1	non-ICUd	Bernades

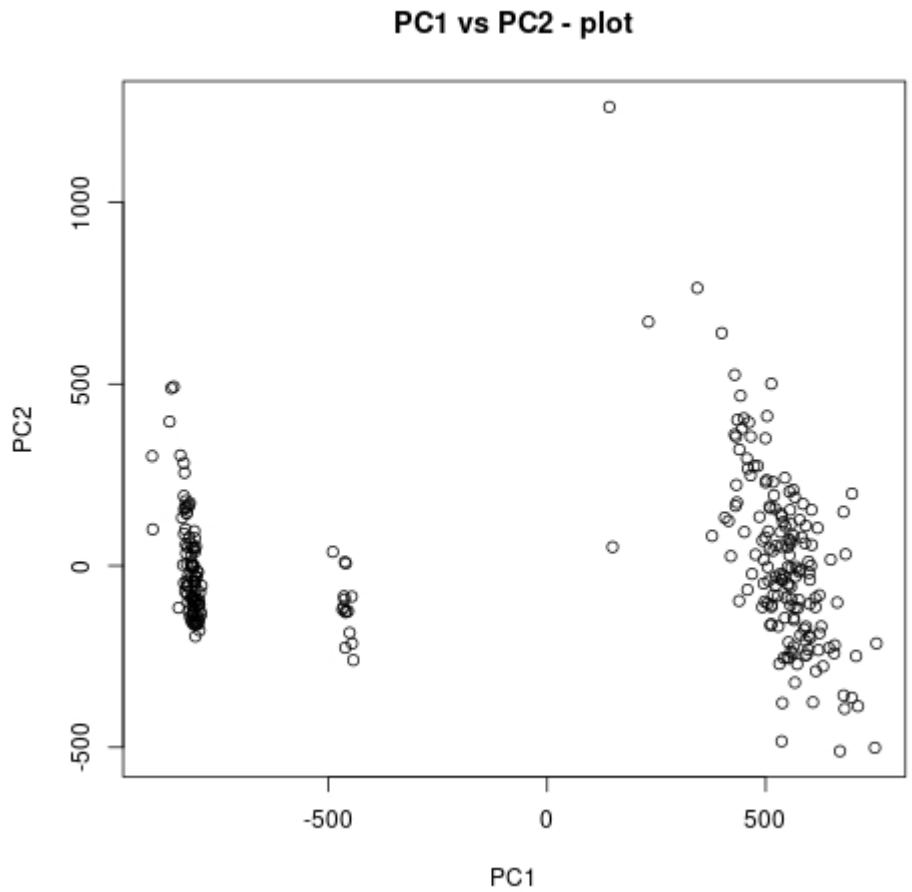
Variance explain: screeplot

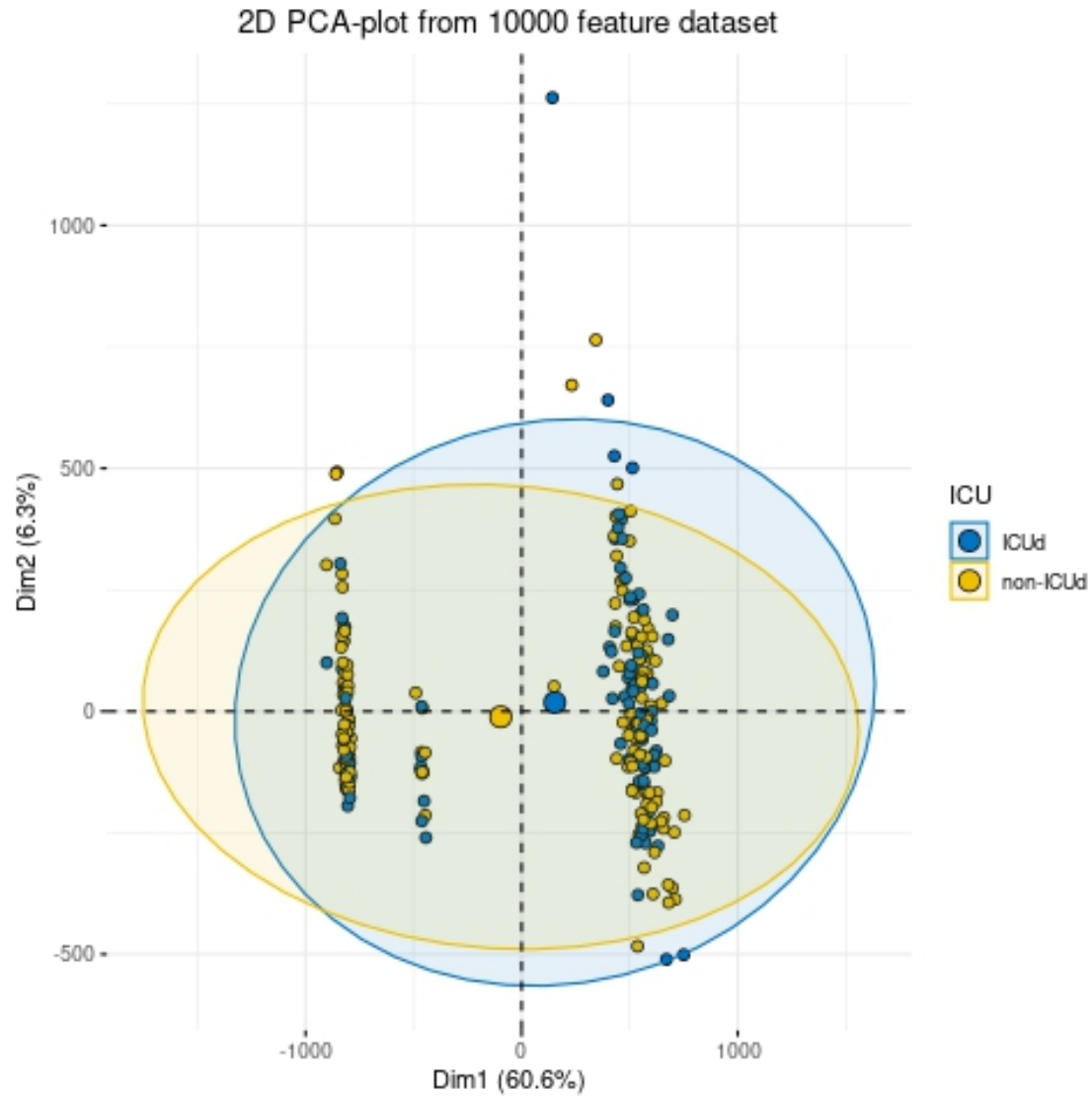




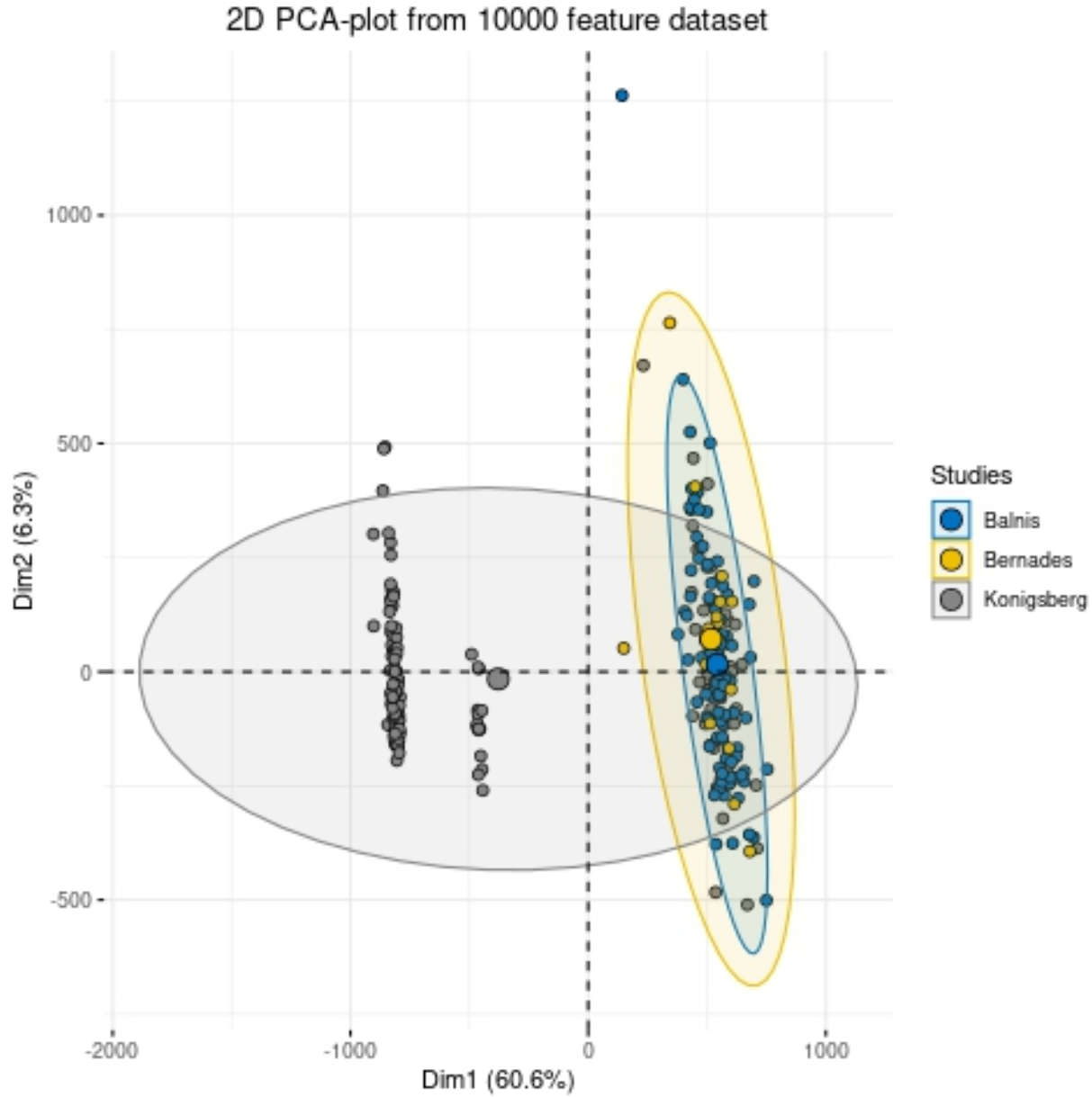


PCA plot

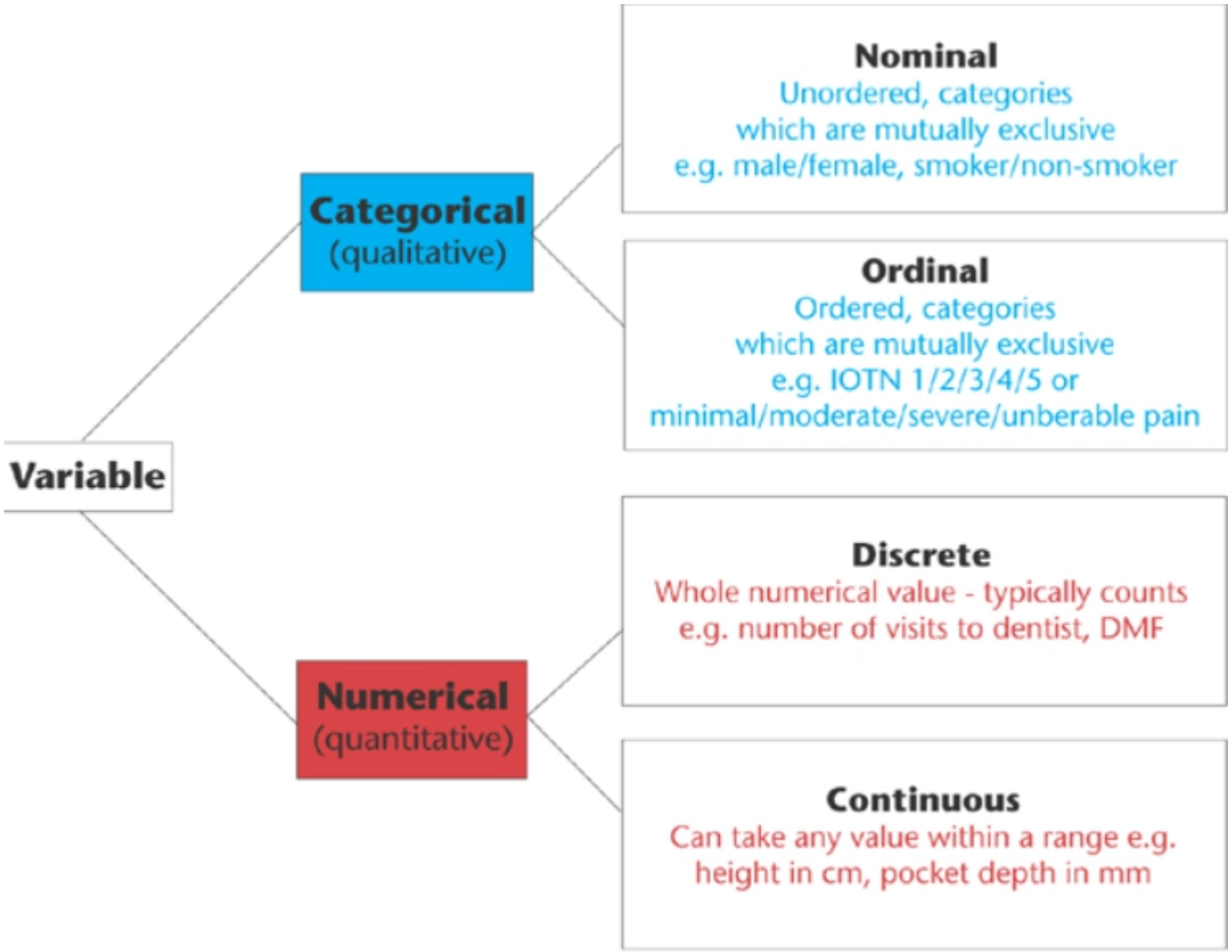

PC1 vs PC2 - plot
PC2
PC1
1000
500
0
-500
-500
0
500


2D PCA-plot from 10000 feature dataset
Dim2 (6.3%)
Dim1 (60.6%)
1000
500
0
-500
-1000
0
1000
ICU
ICUd
non-ICUd

Detect confounder by PCA plot?



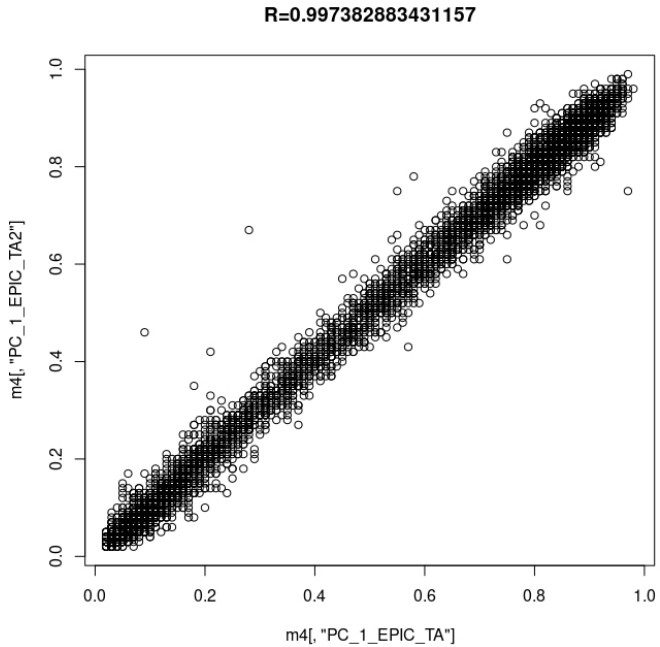
Variable Types



How to calculate correlation?

1. Continuous vs Continuous: weight vs height, methylation of PC_1_EPIC_TA vs PC_2_EPIC_TA

```
> head(m4[1:5,1:5])
  PC_1_EPIC_TA PC_1_EPIC_TA2 PC_1_EPIC_TB PC_1_EPIC_TB2 PC_1_EPIC_TC
1         0.36          0.31         0.32          0.36         0.30
3         0.87          0.87         0.90          0.92         0.91
4         0.92          0.91         0.90          0.90         0.90
5         0.95          0.94         0.93          0.95         0.94
6         0.10          0.12         0.12          0.12         0.13
```




```
> cor(m4[,"PC_1_EPIC_TA"], m4[,"PC_1_EPIC_TA2"])
[1] 0.9973829
```

```
> r <- cor(m4)
> r[1:4,1:4]
              PC_1_EPIC_TA PC_1_EPIC_TA2 PC_1_EPIC_TB PC_1_EPIC_TB2
PC_1_EPIC_TA     1.0000000     0.9973829    0.9950750     0.9977626
PC_1_EPIC_TA2    0.9973829     1.0000000    0.9960812     0.9976331
PC_1_EPIC_TB     0.9950750     0.9960812    1.0000000     0.9952973
PC_1_EPIC_TB2    0.9977626     0.9976331    0.9952973     1.0000000
```

How to calculate correlation?

2. Continuous vs Nominal: height vs gender, age vs race

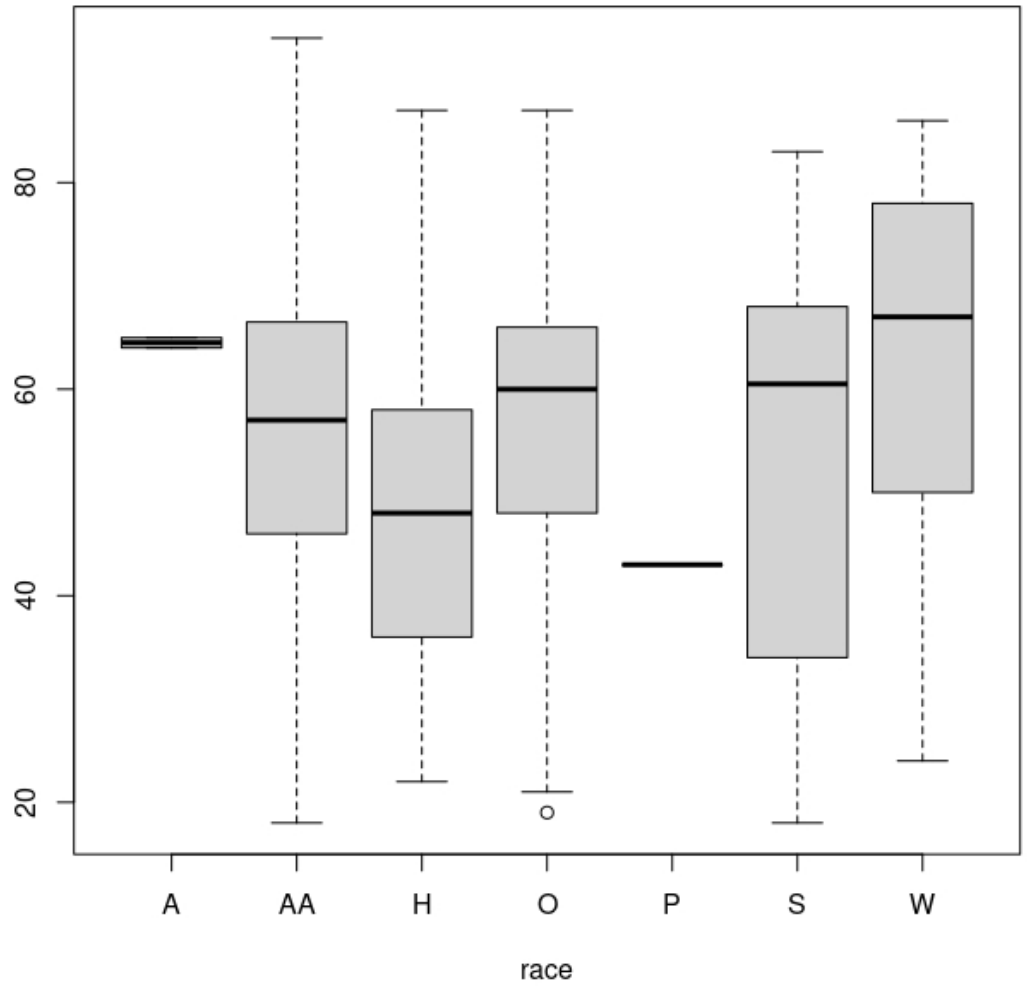
```
> TargetsTable <- read.table(file=TargetsTablep, header=TRUE, sep="\t")
> head(TargetsTable)
      sample_ID sex age race phenotype     icud    study
1  PC_1_EPIC_TA   F  80    W         1     ICUd Bernades
2 PC_1_EPIC_TA2   F  80    W         1     ICUd Bernades    > table(TargetsTable$race)
3  PC_1_EPIC_TB   F  80    W         1 non-ICUd Bernades
4 PC_1_EPIC_TB2   F  80    W         1 non-ICUd Bernades      A  AA   H   O   P   S   W
5  PC_1_EPIC_TC   F  80    W         1 non-ICUd Bernades      2  36 109  26   1  12  93
6 PC_2_EPIC_TA2   M  57    W         1     ICUd Bernades
> cor(TargetsTable[,"age"], TargetsTable[,"race"])
Error in cor(TargetsTable[, "age"], TargetsTable[, "race"]) :
  'y' must be numeric
```

--> How can we calculate correlation of continuous vs Nominal???

How to calculate correlation of continuous vs Nominal?

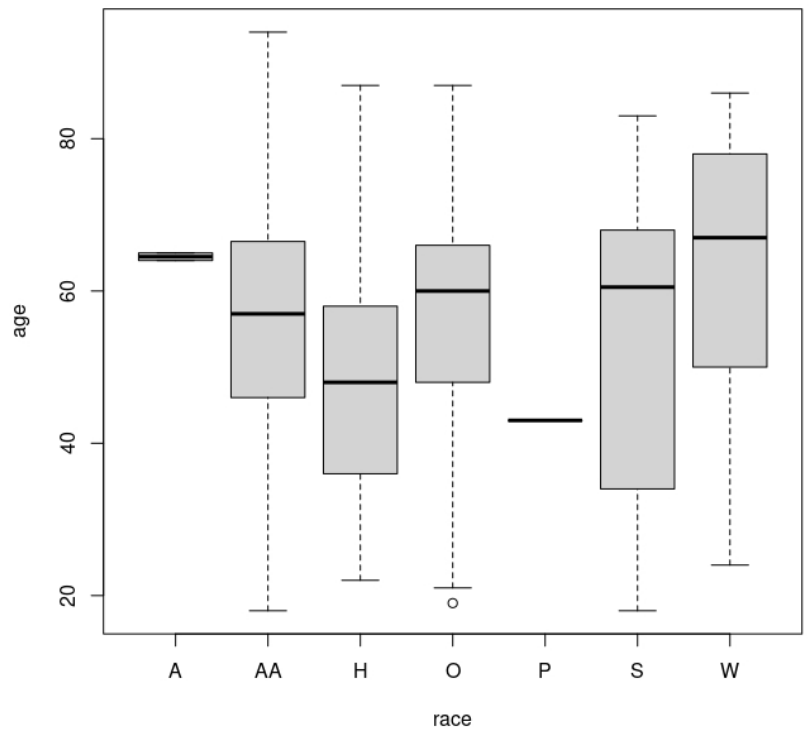
# plot relation between age and study
boxplot(age~race,data=TargetsTable)

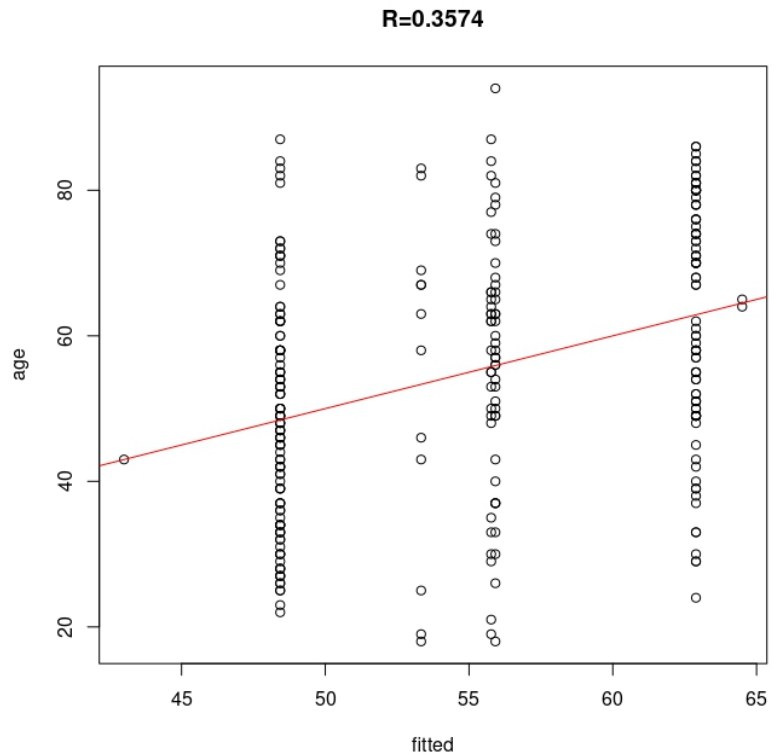
```
> model <- lm(age ~ race, data=TargetsTable)
summary(model)

Call:
lm(formula = age ~ race, data = TargetsTable)

Residuals:
    Min      1Q  Median      3Q     Max
-38.892 -12.431   0.569  11.826  38.569

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   64.500     11.690   5.517 8.01e-08 ***
raceAA        -8.583     12.011  -0.715    0.475
raceH        -16.069     11.797  -1.362    0.174
raceO         -8.731     12.132  -0.720    0.472
raceP        -21.500     20.248  -1.062    0.289
raceS        -11.167     12.627  -0.884    0.377
raceW         -1.608     11.815  -0.136    0.892
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 16.53 on 272 degrees of freedom
Multiple R-squared:  0.1277,    Adjusted R-squared:  0.1085
F-statistic: 6.639 on 6 and 272 DF,  p-value: 1.443e-06
```



How to calculate correlation of continuous vs Nominal?

```
> # correlation of age vs race
rsq <- summary(model)$r.squared

# plot the correlation of age vs race
plot(x=model$fitted, y=TargetsTable[, "age"], xlab="fitted", ylab="age",
      main=paste0("R=", round(sqrt(rsq), 4)))
abline(lm(TargetsTable[,"age"] ~ model$fitted), col="red")
```

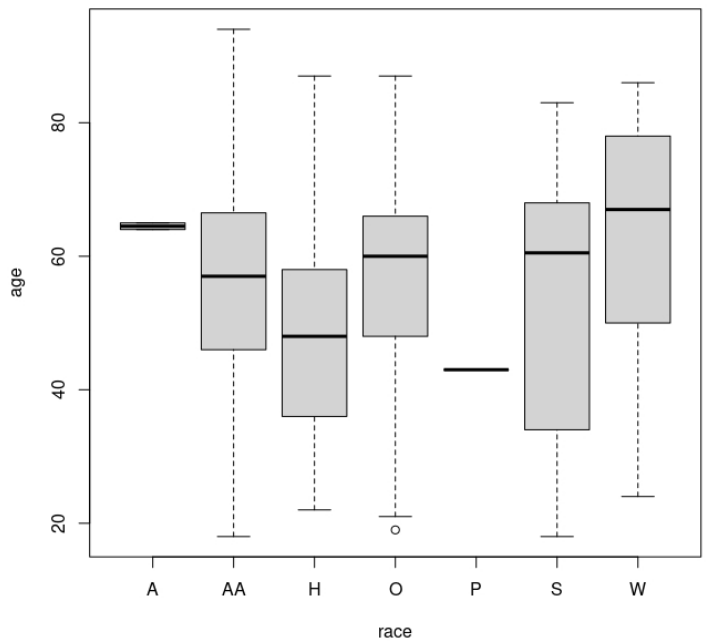


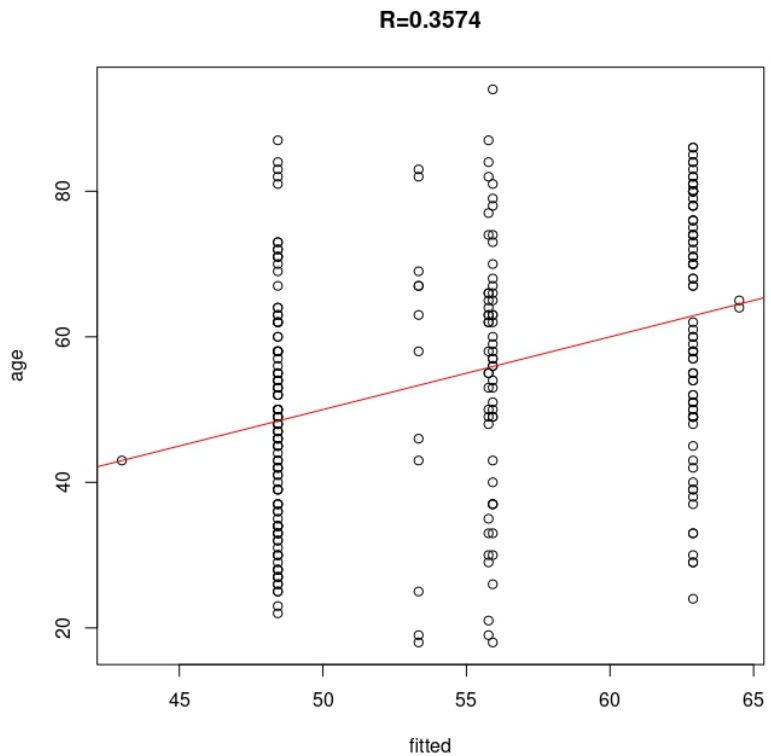


How to calculate correlation of continuous vs Nominal?

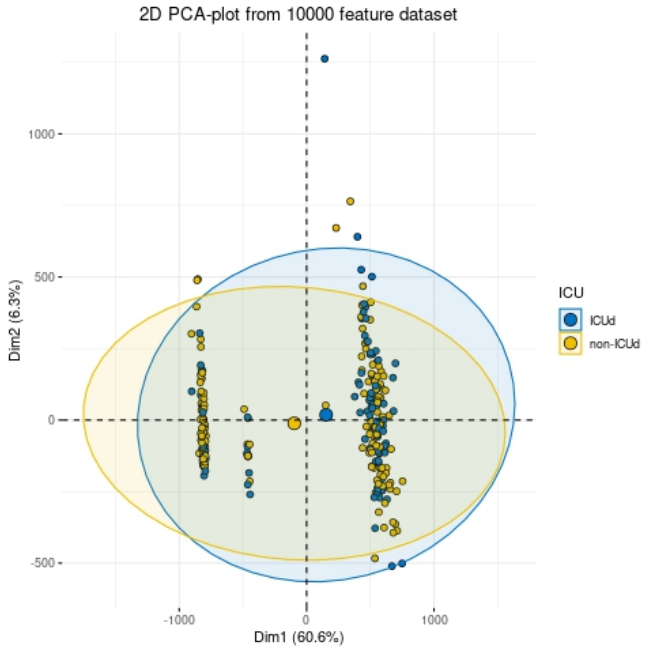
```
> # get overall pvalue
lmp <- function (modelobject) {
    if (class(modelobject) != "lm") stop("Not an object of class 'lm' ")
    f <- summary(modelobject)$fstatistic
    p <- pf(f[1],f[2],f[3],lower.tail=F)
    attributes(p) <- NULL
    return(p)
}

lmp(model)
[1] 1.442529e-06
```

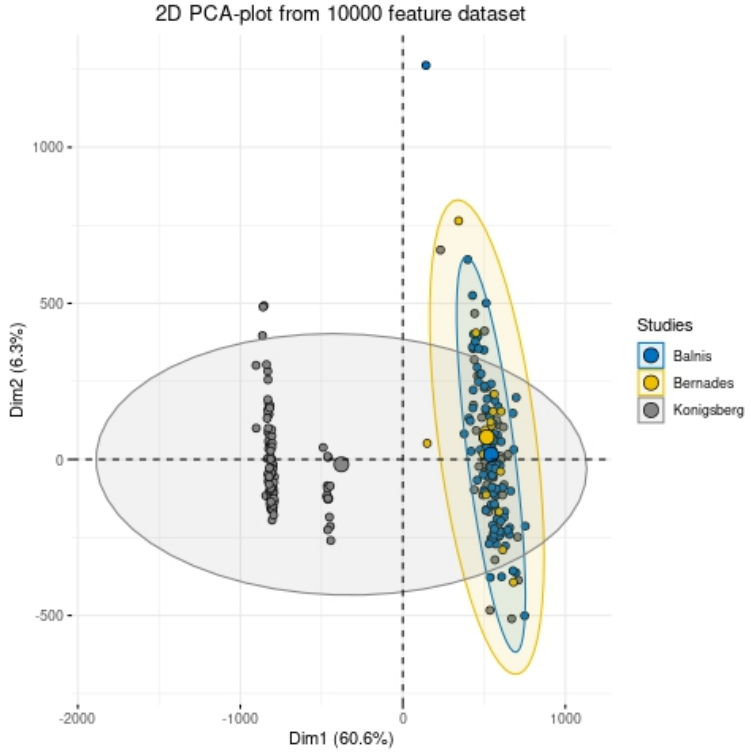
```
> # correlation of age vs race
rsq <- summary(model)$r.squared
> rsq
[1] 0.1277426
```





Automate confounder detection

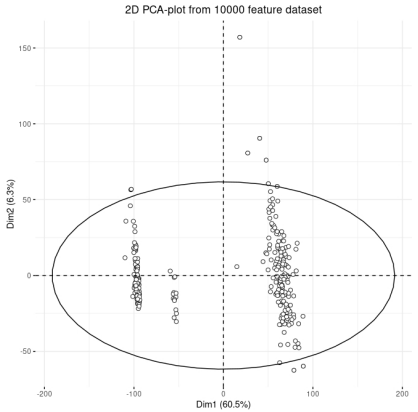




Is age (continuous) confounder ?

1. Compute PCA on data
2. Calculate correlation (R) of PC1/PC2/PC3/PC4/PC5 vs age ?
3. if pvalue < 0.05 AND |R| > r: Confounder otherwise NOT (r=0.3, 0.5, 0.8 ...)

Automate confounder detection: Compute PCA on data



Is age (continuous) confounder ?

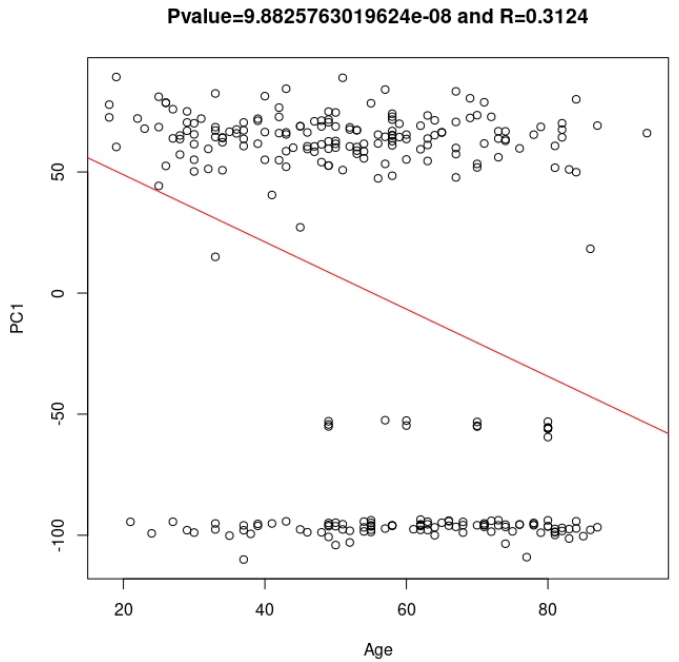
1. Compute PCA on data

```
> head(loadings[1:4,1:4])
                    PC1        PC2       PC3      PC4
PC_1_EPIC_TA  -55.57493 -15.276187 -22.94758 42.55954
PC_1_EPIC_TA2 -55.50855 -11.276034 -26.84154 41.47697
PC_1_EPIC_TB  -59.46157   2.940322 -24.55429 32.45723
PC_1_EPIC_TB2 -55.99016 -15.123830 -24.25384 44.04253
```

sample_ID	sex	age	race	phenotype	icud	study
PC_1_EPIC_TA	F	80	W	1	ICUd	Bernades
PC_1_EPIC_TA2	F	80	W	1	ICUd	Bernades
PC_1_EPIC_TB	F	80	W	1	non-ICUd	Bernades
PC_1_EPIC_TB2	F	80	W	1	non-ICUd	Bernades

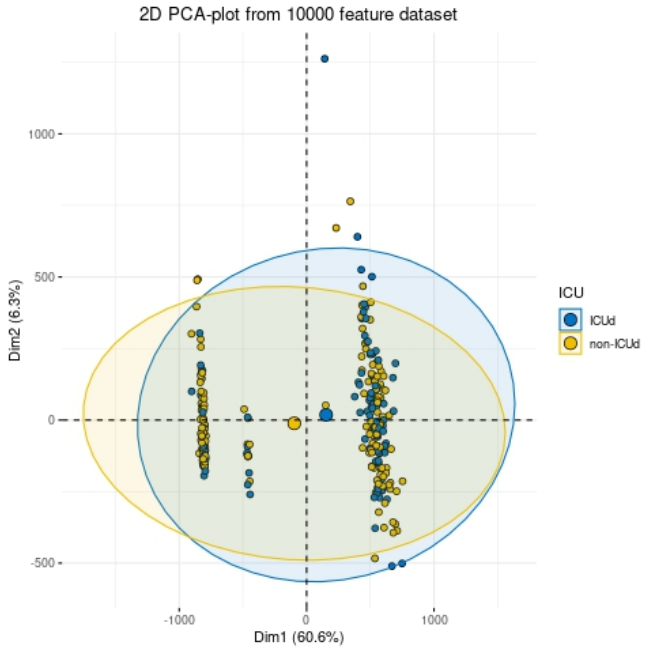
2. Calculate correlation (R) of PC1 vs age ?

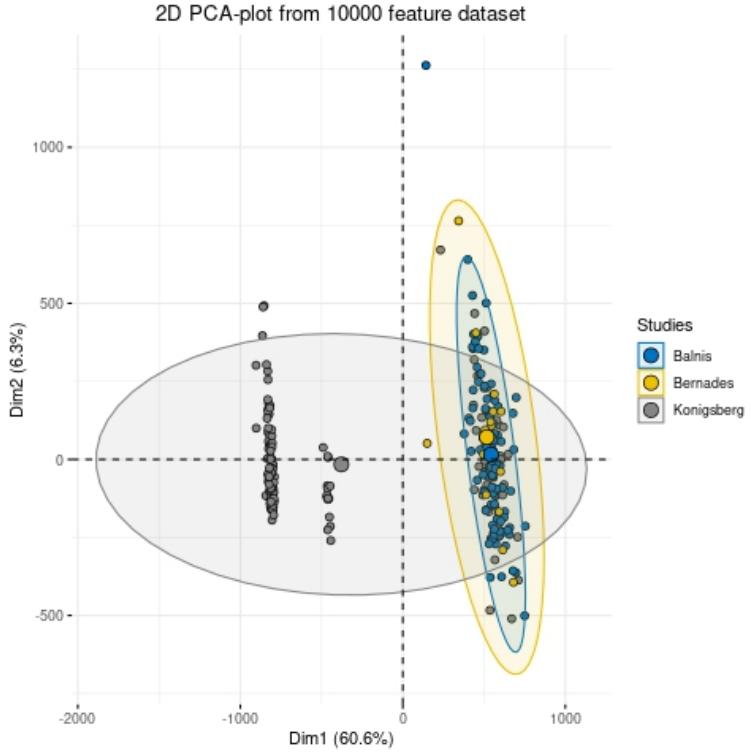
```
> # calculate correlation PC1 vs age
dt <- merge(loadings, TargetsTable, by.x="row.names", by.y="sample_ID")
model <- lm(PC1 ~ age, data=dt)
> # correlation of age vs race
(rsq <- round(sqrt(summary(model)$r.squared), 4))
# p value
(pvalue <- lmp(model))
[1] 0.3124
[1] 9.882576e-08
```



3. if Pvalue < 0.05 and |R|=0.3124 < 0.5:
age is NOT Confounder

Automate confounder detection

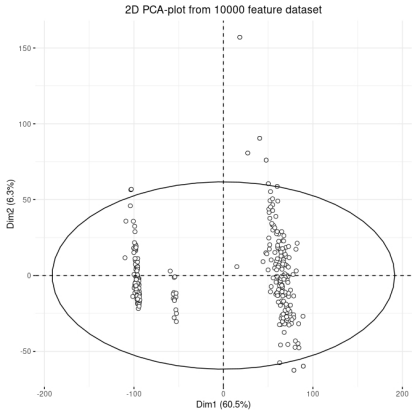




Is race (Nominal) confounder ?

1. Compute PCA on data
2. Calculate correlation (R) of PC1/PC2/PC3/PC4/PC5 vs race ?
3. if |R| > 0.3: Confounder otherwise NOT

Automate confounder detection: Compute PCA on data



Is race (Nominal) confounder ?

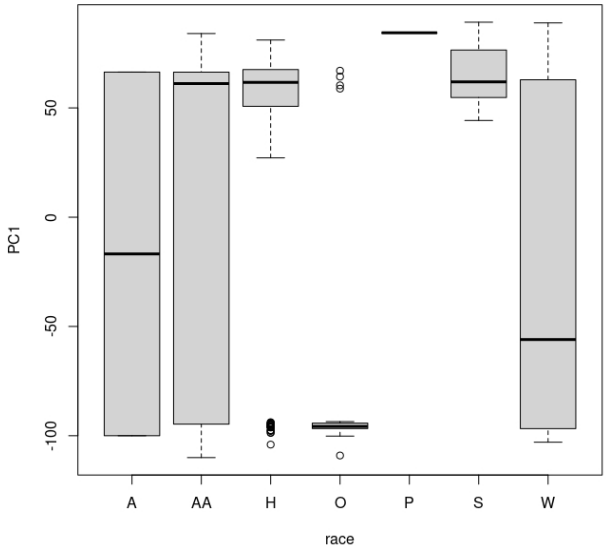
1. Compute PCA on data

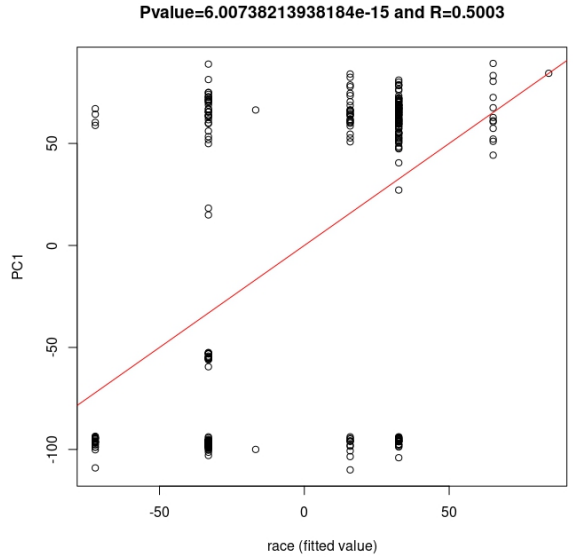
```
> head(loadings[1:4,1:4])
                    PC1        PC2       PC3      PC4
PC_1_EPIC_TA  -55.57493 -15.276187 -22.94758 42.55954
PC_1_EPIC_TA2 -55.50855 -11.276034 -26.84154 41.47697
PC_1_EPIC_TB  -59.46157   2.940322 -24.55429 32.45723
PC_1_EPIC_TB2 -55.99016 -15.123830 -24.25384 44.04253
```

sample_ID	sex	age	race	phenotype	icud	study
PC_1_EPIC_TA	F	80	W	1	ICUd	Bernades
PC_1_EPIC_TA2	F	80	W	1	ICUd	Bernades
PC_1_EPIC_TB	F	80	W	1	non-ICUd	Bernades
PC_1_EPIC_TB2	F	80	W	1	non-ICUd	Bernades

2. Calculate correlation (R) of PC1 vs race ?

```
> model <- lm(PC1 ~ race, data=dt)
# summary(model)

# correlation of PC1 vs race
(rsq <- round(sqrt(summary(model)$r.squared), 4))
# p value
(pvalue <- lmp(model))
[1] 0.5003
[1] 6.007382e-15
```

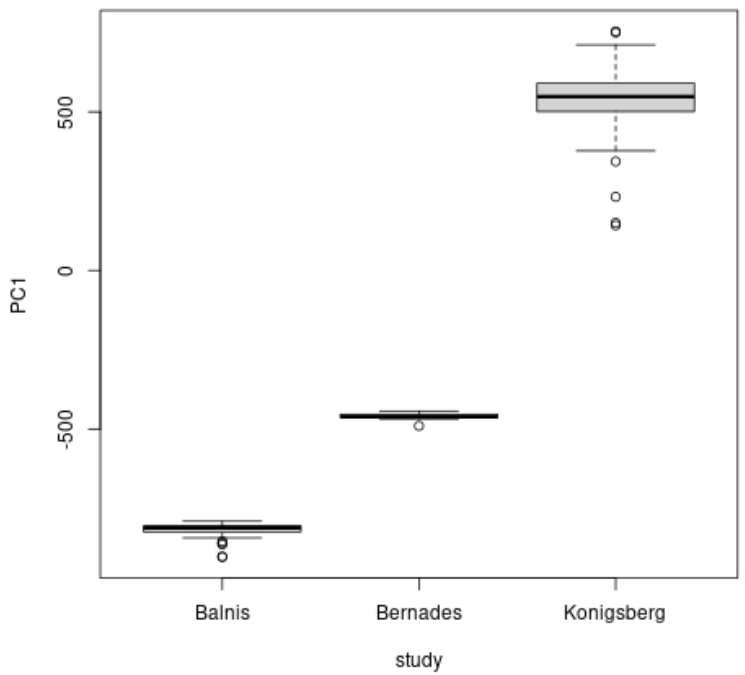


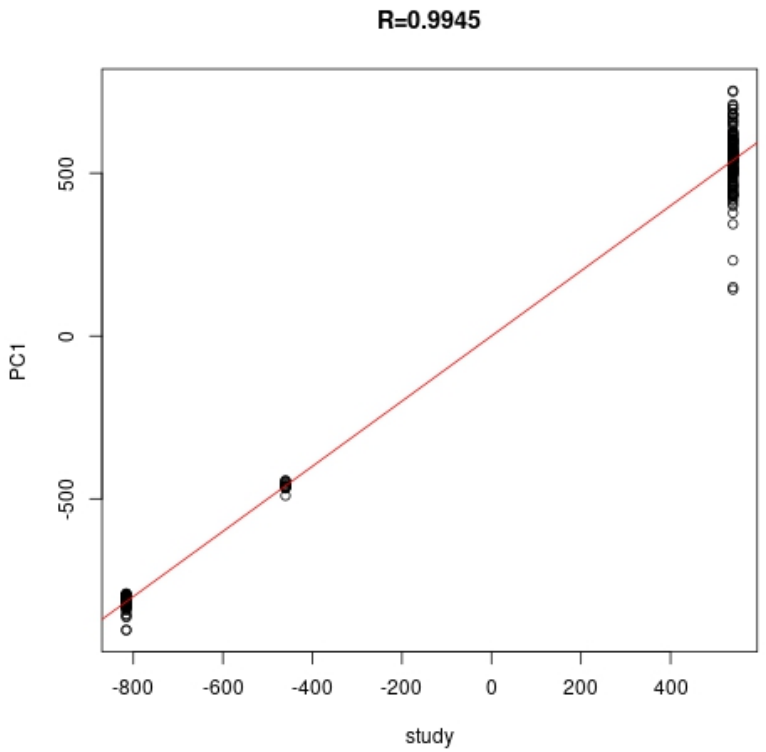


3. if Pvalue < 0.05 and |R|=0.5003 > 0.5: race is Confounder

Exercise: Is study (Nominal) confounder ?

sample_ID	sex	age	race	phenotype	icud	study
PC_1_EPIC_TA	F	80	W	1	ICUd	Bernades
PC_1_EPIC_TA2	F	80	W	1	ICUd	Bernades
PC_1_EPIC_TB	F	80	W	1	non-ICUd	Bernades
PC_1_EPIC_TB2	F	80	W	1	non-ICUd	Bernades





Homework: using ComplexHeatmap package
1. Heatmap Pvalue (row=PC1-10, col=sex, age, race, icud, study)
2. Heatmap Correlation (row=PC1-10, col=sex, age, race, icud, study)

sample_ID	sex	age	race	phenotype	icud	study
PC_1_EPIC_TA	F	80	W	1	ICUd	Bernades
PC_1_EPIC_TA2	F	80	W	1	ICUd	Bernades
PC_1_EPIC_TB	F	80	W	1	non-ICUd	Bernades
PC_1_EPIC_TB2	F	80	W	1	non-ICUd	Bernades