피마 인디언 당뇨병 데이터베이스

Doyoung Kim

2022-08-08

# 당뇨병

#### 혈액 속의 포도당이 세포 속으로 들어가 에너지원으로 이용되지 못하여 혈당이 비정상적으로 올라가는 질환

### 1) 탄수화물 섭취하면 위장에서 소화효소에 의해 탄수화물의 기본 구성성분인 포도당으로 변한 다음 혈액으로 흡수

### 2) 포도당은 기본적인 에너지원. 흡수된 포도당이 세포들에서 이용되기 위해 인슐린 호르몬이 필요

### 3) 인슐린은 췌장의 베타세포에서 분비되어 식사 후 올라간 혈당을 낮추는 기능

### 자료 출처

##### <https://www.kaggle.com/datasets/uciml/pima-indians-diabetes-database>

### 변수 소개

##### Pregnancies: 임신 횟수

##### Glucose: 포도당 부하 검사 수치

##### BloodPressure: 혈압(mm Hg)

##### SkinThickness: 팔 삼두근 뒤쪽의 피하지방 측정값(mm)

##### Insulin: 혈청 인슐린(mu U/ml)

##### BMI: 체질량지수(체중(kg)/키(m))^2

##### DiabetesPedigreeFunction: 당뇨 내력 가중치 값

##### Age: 나이

##### Outcome: 클래스 결정 값(0 또는 1)

## 파일 로딩 및 기본 정보 확인

df <- read.csv('./diabetes/diabetes.csv')  
str(df)

## 'data.frame': 768 obs. of 9 variables:  
## $ Pregnancies : int 6 1 8 1 0 5 3 10 2 8 ...  
## $ Glucose : int 148 85 183 89 137 116 78 115 197 125 ...  
## $ BloodPressure : int 72 66 64 66 40 74 50 0 70 96 ...  
## $ SkinThickness : int 35 29 0 23 35 0 32 0 45 0 ...  
## $ Insulin : int 0 0 0 94 168 0 88 0 543 0 ...  
## $ BMI : num 33.6 26.6 23.3 28.1 43.1 25.6 31 35.3 30.5 0 ...  
## $ DiabetesPedigreeFunction: num 0.627 0.351 0.672 0.167 2.288 ...  
## $ Age : int 50 31 32 21 33 30 26 29 53 54 ...  
## $ Outcome : int 1 0 1 0 1 0 1 0 1 1 ...

## 결측치 확인

sum(is.na(df))

## [1] 0

summary(df)

## Pregnancies Glucose BloodPressure SkinThickness   
## Min. : 0.000 Min. : 0.0 Min. : 0.00 Min. : 0.00   
## 1st Qu.: 1.000 1st Qu.: 99.0 1st Qu.: 62.00 1st Qu.: 0.00   
## Median : 3.000 Median :117.0 Median : 72.00 Median :23.00   
## Mean : 3.845 Mean :120.9 Mean : 69.11 Mean :20.54   
## 3rd Qu.: 6.000 3rd Qu.:140.2 3rd Qu.: 80.00 3rd Qu.:32.00   
## Max. :17.000 Max. :199.0 Max. :122.00 Max. :99.00   
## Insulin BMI DiabetesPedigreeFunction Age   
## Min. : 0.0 Min. : 0.00 Min. :0.0780 Min. :21.00   
## 1st Qu.: 0.0 1st Qu.:27.30 1st Qu.:0.2437 1st Qu.:24.00   
## Median : 30.5 Median :32.00 Median :0.3725 Median :29.00   
## Mean : 79.8 Mean :31.99 Mean :0.4719 Mean :33.24   
## 3rd Qu.:127.2 3rd Qu.:36.60 3rd Qu.:0.6262 3rd Qu.:41.00   
## Max. :846.0 Max. :67.10 Max. :2.4200 Max. :81.00   
## Outcome   
## Min. :0.000   
## 1st Qu.:0.000   
## Median :0.000   
## Mean :0.349   
## 3rd Qu.:1.000   
## Max. :1.000

## 0 <- NA로 바꾸기

df$Glucose[df$Glucose==0] <- NA  
df$BloodPressure[df$BloodPressure==0] <- NA  
df$SkinThickness[df$SkinThickness==0] <- NA  
df$BMI[df$BMI==0] <- NA

## 결측치 그래프

library(VIM)

## 필요한 패키지를 로딩중입니다: colorspace

## 필요한 패키지를 로딩중입니다: grid

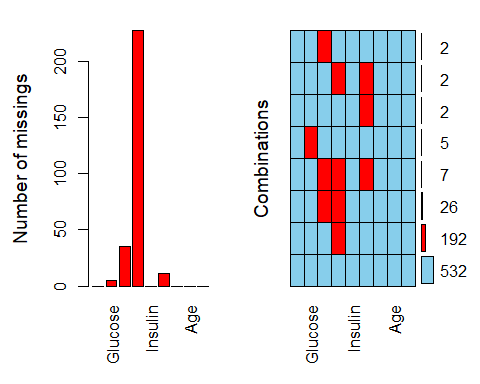
## VIM is ready to use.

## Suggestions and bug-reports can be submitted at: https://github.com/statistikat/VIM/issues

##   
## 다음의 패키지를 부착합니다: 'VIM'

## The following object is masked from 'package:datasets':  
##   
## sleep

aggr(df, numbers=T, prop=F)

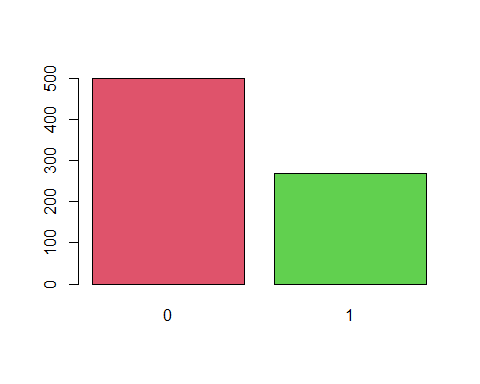


## 정상과 환자 구분하여 NA <- 각 변수 평균으로 채우기

df$Glucose[is.na(df$Glucose) & df$Outcome==0] <- mean(df$Glucose[df$Outcome==0], na.rm=T)  
df$BloodPressure[is.na(df$BloodPressure) & df$Outcome==0] <- mean(df$BloodPressure[df$Outcome==0], na.rm=T)  
df$SkinThickness[is.na(df$SkinThickness) & df$Outcome==0] <- mean(df$SkinThickness[df$Outcome==0], na.rm=T)  
df$BMI[is.na(df$BMI) & df$Outcome==0] <- mean(df$BMI[df$Outcome==0], na.rm=T)  
   
df$Glucose[is.na(df$Glucose) & df$Outcome==1] <- mean(df$Glucose[df$Outcome==1], na.rm=T)  
df$BloodPressure[is.na(df$BloodPressure) & df$Outcome==1] <- mean(df$BloodPressure[df$Outcome==1], na.rm=T)  
df$SkinThickness[is.na(df$SkinThickness) & df$Outcome==1] <- mean(df$SkinThickness[df$Outcome==1], na.rm=T)  
df$BMI[is.na(df$BMI) & df$Outcome==1] <- mean(df$BMI[df$Outcome==1], na.rm=T)

## 정상과 환자 수 그래프

barplot(table(df$Outcome), col = 2:3)



## 정상과 환자 비교

summary(df[df$Outcome==0, ])

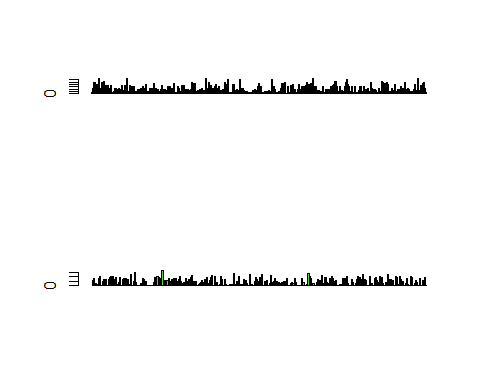
## Pregnancies Glucose BloodPressure SkinThickness   
## Min. : 0.000 Min. : 44.0 Min. : 24.00 Min. : 7.00   
## 1st Qu.: 1.000 1st Qu.: 93.0 1st Qu.: 63.50 1st Qu.:22.00   
## Median : 2.000 Median :107.5 Median : 70.88 Median :27.24   
## Mean : 3.298 Mean :110.6 Mean : 70.88 Mean :27.24   
## 3rd Qu.: 5.000 3rd Qu.:125.0 3rd Qu.: 78.00 3rd Qu.:31.00   
## Max. :13.000 Max. :197.0 Max. :122.00 Max. :60.00   
## Insulin BMI DiabetesPedigreeFunction Age   
## Min. : 0.00 Min. :18.20 Min. :0.0780 Min. :21.00   
## 1st Qu.: 0.00 1st Qu.:25.75 1st Qu.:0.2298 1st Qu.:23.00   
## Median : 39.00 Median :30.40 Median :0.3360 Median :27.00   
## Mean : 68.79 Mean :30.86 Mean :0.4297 Mean :31.19   
## 3rd Qu.:105.00 3rd Qu.:35.30 3rd Qu.:0.5617 3rd Qu.:37.00   
## Max. :744.00 Max. :57.30 Max. :2.3290 Max. :81.00   
## Outcome   
## Min. :0   
## 1st Qu.:0   
## Median :0   
## Mean :0   
## 3rd Qu.:0   
## Max. :0

summary(df[df$Outcome==1, ])

## Pregnancies Glucose BloodPressure SkinThickness  
## Min. : 0.000 Min. : 78.0 Min. : 30.00 Min. : 7   
## 1st Qu.: 1.750 1st Qu.:119.0 1st Qu.: 68.00 1st Qu.:30   
## Median : 4.000 Median :140.5 Median : 75.32 Median :33   
## Mean : 4.866 Mean :142.3 Mean : 75.32 Mean :33   
## 3rd Qu.: 8.000 3rd Qu.:167.0 3rd Qu.: 82.00 3rd Qu.:36   
## Max. :17.000 Max. :199.0 Max. :114.00 Max. :99   
## Insulin BMI DiabetesPedigreeFunction Age   
## Min. : 0.0 Min. :22.90 Min. :0.0880 Min. :21.00   
## 1st Qu.: 0.0 1st Qu.:30.90 1st Qu.:0.2625 1st Qu.:28.00   
## Median : 0.0 Median :34.30 Median :0.4490 Median :36.00   
## Mean :100.3 Mean :35.41 Mean :0.5505 Mean :37.07   
## 3rd Qu.:167.2 3rd Qu.:38.77 3rd Qu.:0.7280 3rd Qu.:44.00   
## Max. :846.0 Max. :67.10 Max. :2.4200 Max. :70.00   
## Outcome   
## Min. :1   
## 1st Qu.:1   
## Median :1   
## Mean :1   
## 3rd Qu.:1   
## Max. :1

## 그래프 그리기

par(mfrow=c(2,1))  
barplot(df$Pregnancies[df$Outcome==0], col=2)  
barplot(df$Pregnancies[df$Outcome==1], col=3)

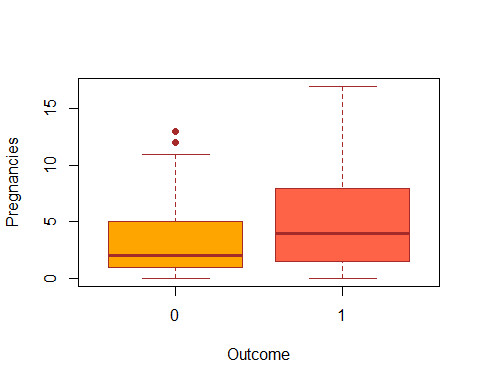


par(mfrow=c(1,1))

## 임신 횟수

##### 임신 중에 분비되는 태반 호르몬이 인슐린의 작용을 방해해 발생되는 인슐린 저항성 때문. 보통은 임신성 당뇨병의 경우 출산을 하게 되면 혈당이 정상화 되지만, 이후에 당뇨병으로 진행할 확률이 정상인의 4배

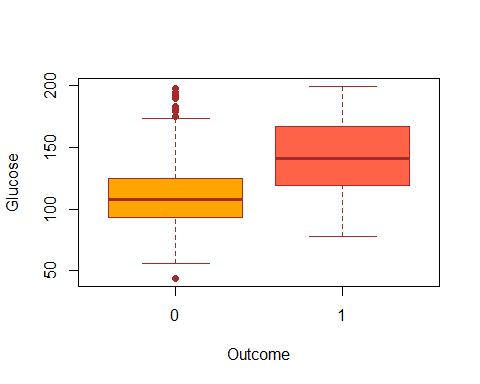
boxplot(Pregnancies ~ Outcome, data = df,  
 pch = 19, col = c('orange', 'tomato'), border = 'brown')



## 포도당 부하 검사 수치

##### 포도당을 부하했을 때, 당뇨병 환자이면 부하된 포도당을 정상인과 같이 신속하게 혈중에서 제거할 수 없다. 140mg/dL 미만 - 정상

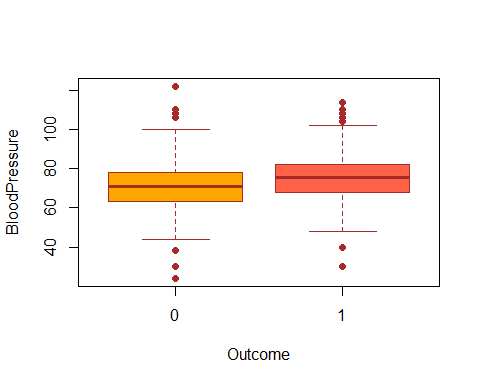
boxplot(Glucose ~ Outcome, data = df,  
 pch = 19, col = c('orange', 'tomato'), border = 'brown')



## 혈압

##### 인슐린에 대한 세포의 저항성이 고혈압이나 당뇨를 일으키는 원인으로 동일하게 작용하기 때문에 고혈압이 있는 사람은 향후 당뇨가 발생할 가능성이 일반인보다 약 2배 높다.

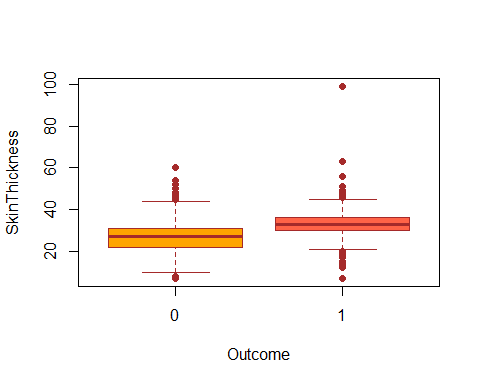
boxplot(BloodPressure ~ Outcome, data = df,  
 pch = 19, col = c('orange', 'tomato'), border = 'brown')



## 팔 삼두근 뒤쪽의 피하지방 측정값(mm)

##### 비만으로 인해 혈중 유리지방의 수치가 증가하게 되면서, 근육의 포도당 섭취가 저하하게 되고, 이로 인해 인슐린 감도가 약해진다. 또한 유리지방산의 증가로 인해 간에서 더 많은 포도당을 생산하게 되며 고혈당증을 더욱 부추기게 된다.

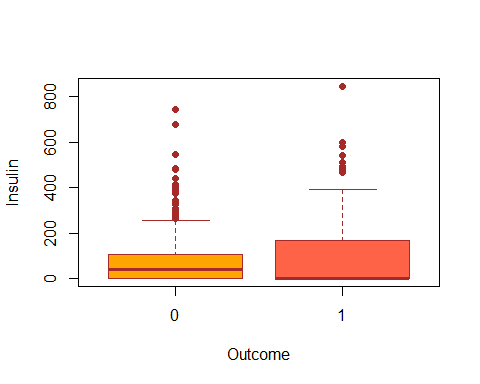
boxplot(SkinThickness ~ Outcome, data = df,  
 pch = 19, col = c('orange', 'tomato'), border = 'brown')



## 혈청 인슐린(mu U/ml)

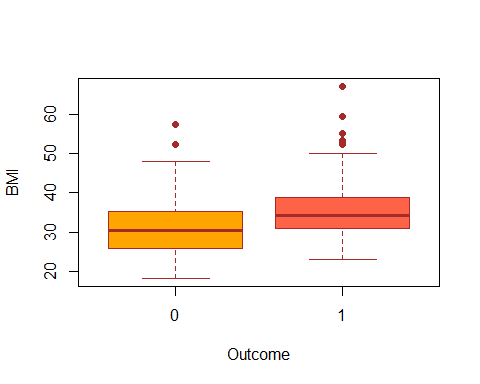
##### <https://terms.naver.com/entry.naver?docId=500275&cid=60408&categoryId=55558>

boxplot(Insulin ~ Outcome, data = df,  
 pch = 19, col = c('orange', 'tomato'), border = 'brown')



## 체질량지수(체중(kg)/키(m))^2

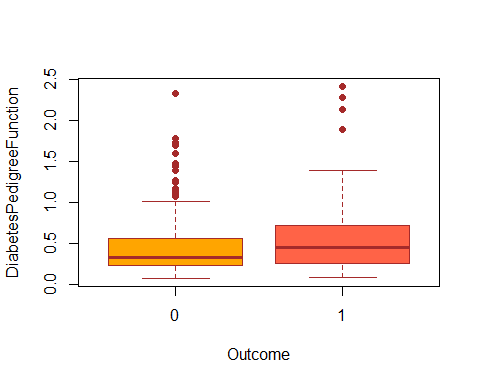
boxplot(BMI ~ Outcome, data = df,  
 pch = 19, col = c('orange', 'tomato'), border = 'brown')



## 당뇨 내력 가중치 값

##### 부모가 모두 제2형 당뇨병인 경우 자녀에게서 제2형 당뇨병이 발병할 가능성은 30% 정도, 부모 중 한 사람만 제2형 당뇨병인 경우 자녀에게 제2형 당뇨병이 발병할 가능성은 15% 정도다.

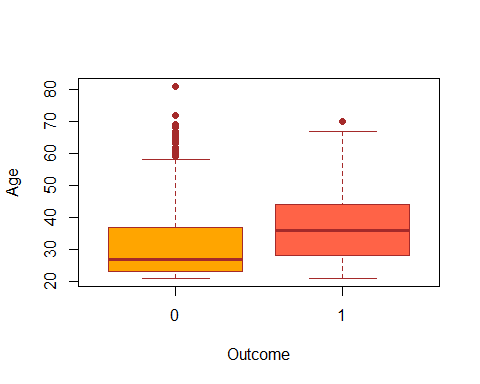
boxplot(DiabetesPedigreeFunction ~ Outcome, data = df,  
 pch = 19, col = c('orange', 'tomato'), border = 'brown')



## Age: 나이

##### 제1형 당뇨병은 15세 미만의 소아청소년에게 주로 발생하는 당뇨병, 제2형 당뇨병은 30대 이상 주로 발생하는 당뇨병

boxplot(Age ~ Outcome, data = df,  
 pch = 19, col = c('orange', 'tomato'), border = 'brown')



## df를 각각 10개의 데이터로 샘플링

set.seed(2022)  
df.origin <- data.frame(id=1:nrow(df),   
 Outcome=df$Outcome,  
 Pregnancies=df$Pregnancies,  
 Glucose=df$Glucose,  
 BloodPressure=df$BloodPressure,  
 SkinThickness=df$SkinThickness,  
 Insulin=df$Insulin,  
 BMI=df$BMI,  
 DiabetesPedigreeFunction=df$DiabetesPedigreeFunction,  
 Age=df$Age)  
idx <- sample(1:nrow(df.origin), size=20)  
df.sample <- df.origin[idx,]  
df.sample

## id Outcome Pregnancies Glucose BloodPressure SkinThickness Insulin BMI  
## 228 228 1 3 162 52.00000 38.00000 0 37.2  
## 435 435 0 1 90 68.00000 8.00000 0 24.5  
## 718 718 0 10 94 72.00000 18.00000 0 23.1  
## 708 708 0 2 127 46.00000 21.00000 335 34.4  
## 703 703 1 1 168 88.00000 29.00000 0 35.0  
## 766 766 0 5 121 72.00000 23.00000 112 26.2  
## 476 476 0 0 137 84.00000 27.00000 0 27.3  
## 123 123 0 2 107 74.00000 30.00000 100 33.6  
## 233 233 0 1 79 80.00000 25.00000 37 25.4  
## 270 270 1 2 146 75.32143 33.00000 0 27.5  
## 248 248 0 0 165 90.00000 33.00000 680 52.3  
## 7 7 1 3 78 50.00000 32.00000 88 31.0  
## 498 498 0 2 81 72.00000 15.00000 76 30.1  
## 112 112 1 8 155 62.00000 26.00000 495 34.0  
## 470 470 0 6 154 78.00000 41.00000 140 46.1  
## 513 513 0 9 91 68.00000 27.23546 0 24.2  
## 307 307 1 10 161 68.00000 23.00000 132 25.5  
## 449 449 1 0 104 64.00000 37.00000 64 33.6  
## 378 378 0 1 87 60.00000 37.00000 75 37.2  
## 669 669 0 6 98 58.00000 33.00000 190 34.0  
## DiabetesPedigreeFunction Age  
## 228 0.652 24  
## 435 1.138 36  
## 718 0.595 56  
## 708 0.176 22  
## 703 0.905 52  
## 766 0.245 30  
## 476 0.231 59  
## 123 0.404 23  
## 233 0.583 22  
## 270 0.240 28  
## 248 0.427 23  
## 7 0.248 26  
## 498 0.547 25  
## 112 0.543 46  
## 470 0.571 27  
## 513 0.200 58  
## 307 0.326 47  
## 449 0.510 22  
## 378 0.509 22  
## 669 0.430 43

## 샘플링한 두 데이트프레임 병합

set.seed(2022)  
df.x <- df.sample[sample(1:nrow(df.sample), size=10), c(1,2,3:6)]  
df.x[order(df.x$id), ]

## id Outcome Pregnancies Glucose BloodPressure SkinThickness  
## 7 7 1 3 78 50 32  
## 112 112 1 8 155 62 26  
## 233 233 0 1 79 80 25  
## 248 248 0 0 165 90 33  
## 307 307 1 10 161 68 23  
## 378 378 0 1 87 60 37  
## 449 449 1 0 104 64 37  
## 669 669 0 6 98 58 33  
## 708 708 0 2 127 46 21  
## 766 766 0 5 121 72 23

df.y <- df.sample[sample(1:nrow(df.sample), size=10), c(1,2,7:10)]  
df.y[order(df.y$id), ]

## id Outcome Insulin BMI DiabetesPedigreeFunction Age  
## 112 112 1 495 34.0 0.543 46  
## 228 228 1 0 37.2 0.652 24  
## 435 435 0 0 24.5 1.138 36  
## 449 449 1 64 33.6 0.510 22  
## 476 476 0 0 27.3 0.231 59  
## 513 513 0 0 24.2 0.200 58  
## 669 669 0 190 34.0 0.430 43  
## 703 703 1 0 35.0 0.905 52  
## 718 718 0 0 23.1 0.595 56  
## 766 766 0 112 26.2 0.245 30

df.merge <- merge(x=df.x, y=df.y, by=c('id', 'Outcome'), all = T)  
df.merge

## id Outcome Pregnancies Glucose BloodPressure SkinThickness Insulin BMI  
## 1 7 1 3 78 50 32 NA NA  
## 2 112 1 8 155 62 26 495 34.0  
## 3 228 1 NA NA NA NA 0 37.2  
## 4 233 0 1 79 80 25 NA NA  
## 5 248 0 0 165 90 33 NA NA  
## 6 307 1 10 161 68 23 NA NA  
## 7 378 0 1 87 60 37 NA NA  
## 8 435 0 NA NA NA NA 0 24.5  
## 9 449 1 0 104 64 37 64 33.6  
## 10 476 0 NA NA NA NA 0 27.3  
## 11 513 0 NA NA NA NA 0 24.2  
## 12 669 0 6 98 58 33 190 34.0  
## 13 703 1 NA NA NA NA 0 35.0  
## 14 708 0 2 127 46 21 NA NA  
## 15 718 0 NA NA NA NA 0 23.1  
## 16 766 0 5 121 72 23 112 26.2  
## DiabetesPedigreeFunction Age  
## 1 NA NA  
## 2 0.543 46  
## 3 0.652 24  
## 4 NA NA  
## 5 NA NA  
## 6 NA NA  
## 7 NA NA  
## 8 1.138 36  
## 9 0.510 22  
## 10 0.231 59  
## 11 0.200 58  
## 12 0.430 43  
## 13 0.905 52  
## 14 NA NA  
## 15 0.595 56  
## 16 0.245 30

## 기존 데이터와 샘플링 데이터 평균 상관계수 비교

library(psych)  
describe(df.merge)[c(3:10),3]

## [1] 3.6000 117.5000 65.0000 29.0000 86.1000 29.9100 0.5449 42.6000

describe(df)[-9,3]

## [1] 3.8450521 121.6973577 72.4281410 29.2470424 79.7994792 32.4464201  
## [7] 0.4718763 33.2408854

cor(describe(df.merge)[c(3:10),3], describe(df)[-9,3])

## [1] 0.9914949