brca_draft

Bulun Te

2024-04-20

```r
setwd(here::here())
load("CPD.rda")
library(readxl)
mamo_data = read.csv("mamo_data.csv")
library(purrr)

if (!requireNamespace("logistf", quietly = TRUE)) {
  install.packages("logistf")
}
library(logistf)
```

```
## Warning:   'logistf' R 4.3.3
```

```r
library(dplyr)
```

```
##
##     'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```r
library(corrplot)
```

```
## corrplot 0.92 loaded
```

```r
Data$BRCA -> brca
names(brca)
```

```
##  [1] "Cancer Code"            "Patient ID"
##  [3] "Age at Diagnosis (Years)"  "Pathologic Stage"
##  [5] "Gender"                 "Apoptosis"
##  [7] "Breast Reactive"        "Cell Cycle"
##  [9] "Core Reactive"          "DNA Damage Response"
```

```
## [11] "EMT"                      "Hormone Receptor"
## [13] "Hormone Signaling (Breast)" "PI3K/AKT"
## [15] "RAS/MAPK"                  "RTK"
## [17] "TSC/mTOR"                  "Survival Indicator"
```

```r
names(brca) <- make.names(names(brca))
brca_covar = brca %>% select(-Hormone.Receptor,
                             -Hormone.Signaling..Breast.,
                             -Cancer.Code,
                             -Core.Reactive)
map(brca_covar, ~sum(is.na(.)) )%>% unlist()
```

```
##            Patient.ID Age.at.Diagnosis..Years.      Pathologic.Stage
##                     0                       14                     7
##                Gender                 Apoptosis       Breast.Reactive
##                     0                        0                     0
##            Cell.Cycle      DNA.Damage.Response                   EMT
##                     0                        0                     0
##              PI3K.AKT                 RAS.MAPK                   RTK
##                     0                        0                     0
##              TSC.mTOR        Survival.Indicator
##                     0                        1
```

```r
brca_covar_na <- brca_covar %>% select('Age.at.Diagnosis..Years.',
                                       'Pathologic.Stage',
                                       'Survival.Indicator')
```

```r
brca_covar_na %>% filter(is.na(Age.at.Diagnosis..Years.)) %>% knitr::kable()
```

Age.at.Diagnosis..Years.	Pathologic.Stage	Survival.Indicator
NA	iib	0
NA	iia	0
NA	iib	0
NA	iia	0
NA	iia	0
NA	iia	0
NA	iiia	0
NA	iib	0
NA	iiia	0
NA	iia	0
NA	iib	0
NA	iib	0
NA	iia	0
NA	iib	0

```r
brca_covar_na %>% filter(is.na(Pathologic.Stage)) %>% knitr::kable()
```

Age.at.Diagnosis..Years.	Pathologic.Stage	Survival.Indicator
76	NA	0

Age.at.Diagnosis..Years.	Pathologic.Stage	Survival.Indicator
73	NA	0
51	NA	0
51	NA	0
45	NA	0
60	NA	0
50	NA	0

```
brca_covar_na %>% filter(is.na(Survival.Indicator)) %>% knitr::kable()
```

Age.at.Diagnosis..Years.	Pathologic.Stage	Survival.Indicator
47	iib	NA

```
skimr::skim(brca_covar)
```

Table 4: Data summary

Name	brca_covar
Number of rows	873
Number of columns	14
Column type frequency:	
character	1
factor	2
numeric	11
Group variables	None

Variable type: character

skim_variable	n_missing	complete_rate	min	max	empty	n_unique	whitespace
Patient.ID	0	1	27	27	0	873	0

Variable type: factor

skim_variable	n_missing	complete_rate	ordered	n_unique	top_counts
Pathologic.Stage	7	0.99	FALSE	12	iia: 291, iib: 214, iii: 132, ia: 62
Gender	0	1.00	FALSE	2	fem: 865, mal: 8

Variable type: numeric

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100	hist
Age.at.Diagnosis..Years.	14	0.98	58.26	13.32	26.00	48.50	58.00	67.00	90.00	
Apoptosis	0	1.00	30.80	1.78	28.28	29.65	30.43	31.39	43.69	

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100	hist
Breast.Reactive	0	1.00	18.47	3.41	14.61	15.63	17.46	20.39	29.64	
Cell.Cycle	0	1.00	21.34	2.59	18.66	19.51	20.49	22.14	32.64	
DNA.Damage.Response	0	1.00	47.74	5.59	40.83	43.77	46.06	49.94	72.41	
EMT	0	1.00	21.47	2.47	17.96	19.48	20.82	23.23	32.82	
PI3K.AKT	0	1.00	21.56	1.98	19.05	19.99	21.02	22.57	30.36	
RAS.MAPK	0	1.00	35.77	4.63	30.42	32.38	34.37	37.85	56.52	
RTK	0	1.00	7.21	0.91	6.59	6.69	6.84	7.25	11.55	
TSC.mTOR	0	1.00	8.00	0.52	7.49	7.67	7.83	8.14	11.32	
Survival.Indicator	1	1.00	0.03	0.16	0.00	0.00	0.00	0.00	1.00	

```r
brca_covar$Pathologic.Stage <- as.factor(brca_covar$Pathologic.Stage)
brca_covar$Gender <- as.factor(brca_covar$Gender) %>% as.numeric()
brca_covar <- brca_covar %>% select(-Patient.ID)

brca_covar_numeric <- brca_covar %>%
  select_if(is.numeric)

# corrplot(cor(brca_covar_numeric), method = "number")

library(GGally)
ggpairs(brca_covar_numeric) %>% ggsave("brca_covar_numeric.png",.,width=20)
```
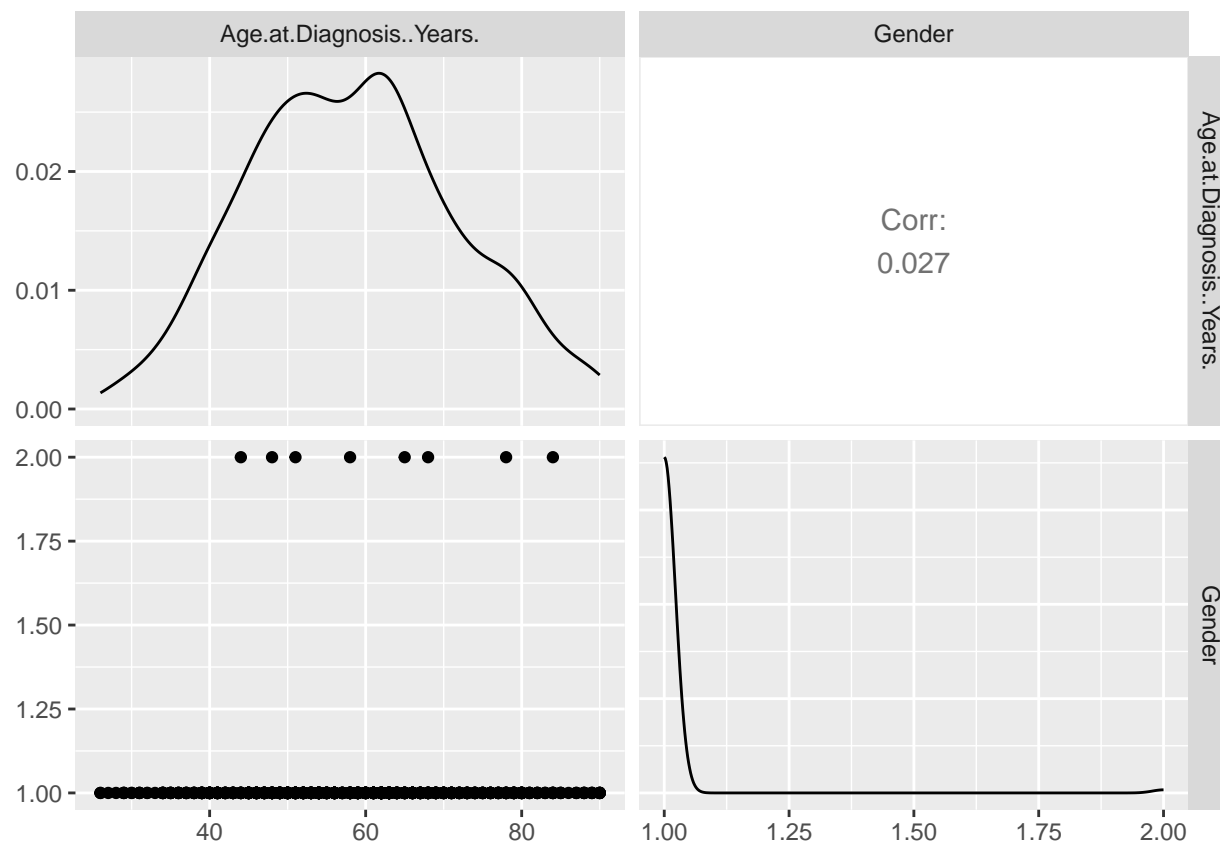
Checking the distribution of age and gender

```r
brca_covar %>% select(Age.at.Diagnosis..Years.,Gender) %>% ggpairs()
```

```
## Warning: Removed 14 rows containing non-finite values (`stat_density()`).


## Warning in ggally_statistic(data = data, mapping = mapping, na.rm = na.rm, :
## Removed 14 rows containing missing values


## Warning: Removed 14 rows containing missing values (`geom_point()`).
```

```
brca_covar$Gender %>% table()
```
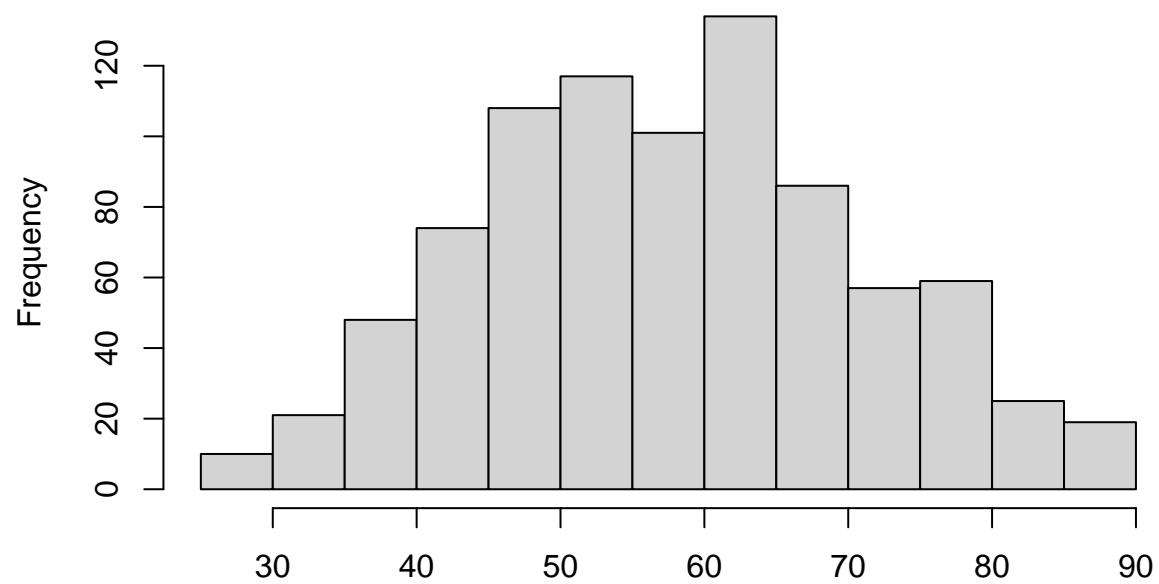
```
## .
##   1   2
## 865   8
```

```
brca_covar$Age.at.Diagnosis..Years. %>% hist()
```

Histogram of .

.

```r
brca_covar %>% select(Gender,Survival.Indicator) %>% table()
```

```
##       Survival.Indicator
## Gender   0   1
##      1 841  23
##      2   8   0
```