# Statistical summaries

Abhijit Dasgupta

BIOF 339

# Where we've been

1. Understand what tidy data is
2. Manipulate data to make it tidy (tidyr, dplyr)
3. Transform particular variables
4. Write basic functions
5. High-throughput analyses
   - Lists of data sets
   - `map` to apply similar processes to each data set
   - for-loops to repeat same recipe on multiple data sets or objects

# Where we're going

1. Creating data summaries
2. Basic statistical comparisons between groups
3. Creating tables
    - Table 1
    - Tables for analytic results

The basic assumption we'll make is that we will start with a tidy data set.

# Statistical summaries

# Univariate summaries

## Single summaries

- Mean (`mean`)
- Variance(`var`)
- Standard deviation (`sd`)
- Count (`nrow` or `dplyr::n` or `dplyr::n_distinct`)

- Median ('median')
- Inter-quartile range (`IQR`)
- Mean absolute deviation (`mad`)
- Minimum (`min`) and Maximum (`max`)

## Multiple summaries

- Quantiles (`quantile`)
- Range (`range`)

# Summarizing the breast cancer expression dataset

# Mean

```
brca <- rio::import('../data/BreastCancer_Expression.csv')
brca %>%
  summarize(across(starts_with('NP'),
                   mean, na.rm=T))
```

```
  NP_958782 NP_958785 NP_958786 NP_000436 NP_958781 NP_958780 NP_958783
1 0.3202321 0.3269153 0.3264254 0.3236833 0.3270832 0.3263382 0.3259212
  NP_958784  NP_112598 NP_001611
1 0.3259995 -0.3074577 0.4578748
```

# Median

```
brca %>%
  summarize(across(starts_with('NP'),
                   median, na.rm=T))
```

```
  NP_958782 NP_958785 NP_958786 NP_000436 NP_958781 NP_958780 NP_958783
1 0.3236627 0.3269726 0.3269726 0.3302826 0.3269726 0.3269726 0.3269726
  NP_958784  NP_112598 NP_001611
1 0.3269726 -0.6021319 0.6948104
```

# Standard deviation

```
brca %>%
  summarize(across(starts_with('NP'),
                   sd, na.rm=T))
```

```
  NP_958782 NP_958785 NP_958786 NP_000436 NP_958781 NP_958780 NP_958783
1 0.9767777 0.9800721 0.9799358 0.9784656 0.9806001 0.9796277 0.9806739
  NP_958784 NP_112598 NP_001611
1 0.9807512  2.024663  1.496951
```

# Multiple summaries together

```
brca %>%
  summarize(across(starts_with('NP'),
                   c(mean,
                     median,
                     sd), na.rm=T))
```

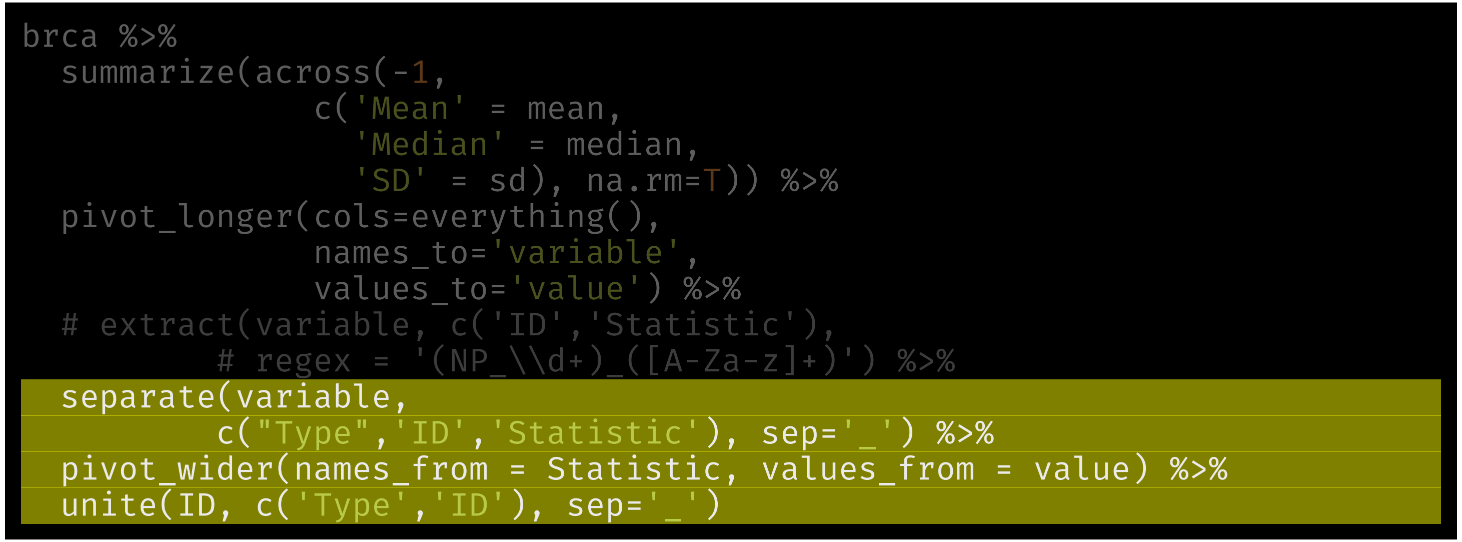
```
  NP_958782_1 NP_958782_2 NP_958782_3 NP_958785_1 NP_958785_2 NP_958785_3
1   0.3202321   0.3236627   0.9767777   0.3269153   0.3269726   0.9800721
  NP_958786_1 NP_958786_2 NP_958786_3 NP_000436_1 NP_000436_2 NP_000436_3
1   0.3264254   0.3269726   0.9799358   0.3236833   0.3302826   0.9784656
  NP_958781_1 NP_958781_2 NP_958781_3 NP_958780_1 NP_958780_2 NP_958780_3
1   0.3270832   0.3269726   0.9806001   0.3263382   0.3269726   0.9796277
  NP_958783_1 NP_958783_2 NP_958783_3 NP_958784_1 NP_958784_2 NP_958784_3
1   0.3259212   0.3269726   0.9806739   0.3259995   0.3269726   0.9807512
  NP_112598_1 NP_112598_2 NP_112598_3 NP_001611_1 NP_001611_2 NP_001611_3
1  -0.3074577  -0.6021319    2.024663   0.4578748   0.6948104    1.496951
```

# Multiple summaries together

```
brca %>%
  summarize(across(-1, # got tired of typing
               c('Mean'=mean,
                 'Median' = median,
                 'SD'=sd), na.rm=T))
```

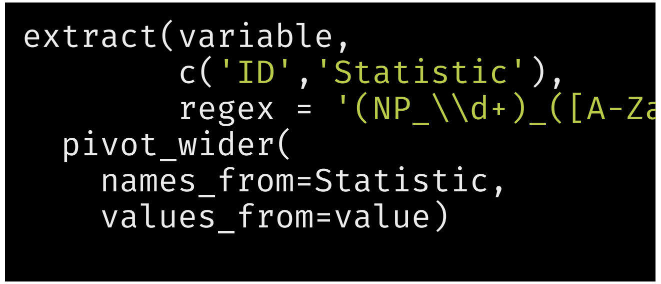
```
  NP_958782_Mean NP_958782_Median NP_958782_SD NP_958785_Mean NP_958785_Median
1      0.3202321        0.3236627    0.9767777      0.3269153        0.3269726
  NP_958785_SD NP_958786_Mean NP_958786_Median NP_958786_SD NP_000436_Mean
1    0.9800721      0.3264254        0.3269726    0.9799358      0.3236833
  NP_000436_Median NP_000436_SD NP_958781_Mean NP_958781_Median NP_958781_SD
1        0.3302826    0.9784656      0.3270832        0.3269726    0.9806001
  NP_958780_Mean NP_958780_Median NP_958780_SD NP_958783_Mean NP_958783_Median
1      0.3263382        0.3269726    0.9796277      0.3259212        0.3269726
  NP_958783_SD NP_958784_Mean NP_958784_Median NP_958784_SD NP_112598_Mean
1    0.9806739      0.3259995        0.3269726    0.9807512     -0.3074577
  NP_112598_Median NP_112598_SD NP_001611_Mean NP_001611_Median NP_001611_SD
1       -0.6021319     2.024663      0.4578748        0.6948104     1.496951
```

# Multiple summaries together

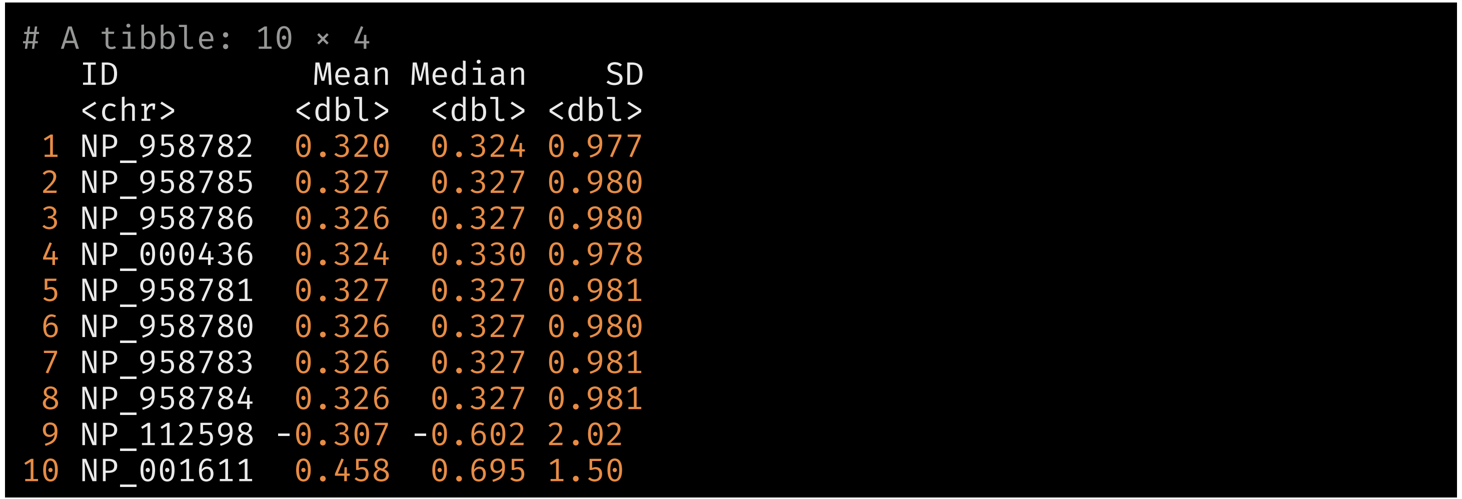


```
brca %>%
  summarize(across(-1,
                   c('Mean' = mean,
                     'Median' = median,
                     'SD' = sd), na.rm=T)) %>%
  pivot_longer(cols=everything(),
               names_to='variable',
               values_to='value') %>%
  # extract(variable, c('ID','Statistic'),
          # regex = '(NP_\\d+)_([A-Za-z]+)') %>%
  separate(variable,
          c("Type",'ID','Statistic'), sep='_') %>%
  pivot_wider(names_from = Statistic, values_from = value) %>%
  unite(ID, c('Type','ID'), sep='_')
```

You could replace the highlighted code with



```
extract(variable,
        c('ID','Statistic'),
        regex = '(NP_\\d+)_([A-Za
  pivot_wider(
    names_from=Statistic,
    values_from=value)
```



```
# A tibble: 10 × 4
   ID          Mean Median    SD
   <chr>      <dbl>  <dbl> <dbl>
 1 NP_958782  0.320  0.324 0.977
 2 NP_958785  0.327  0.327 0.980
 3 NP_958786  0.326  0.327 0.980
 4 NP_000436  0.324  0.330 0.978
 5 NP_958781  0.327  0.327 0.981
 6 NP_958780  0.326  0.327 0.980
 7 NP_958783  0.326  0.327 0.981
 8 NP_958784  0.326  0.327 0.981
 9 NP_112598 -0.307 -0.602 2.02
10 NP_001611  0.458  0.695 1.50
```

# Summarizing a data set

# Data set summary

There is a function `summary` that will give you summaries of all the variables. It's nice for looking at the data, but the output format isn't very good for further manipulation

```
summary(brca[,-1]) # Omit first column
```

```
   NP_958782          NP_958785          NP_958786          NP_000436      
 Min.   :-1.9478   Min.   :-1.9527   Min.   :-1.9552   Min.   :-1.9478  
 1st Qu.:-0.4549   1st Qu.:-0.4421   1st Qu.:-0.4440   1st Qu.:-0.4385  
 Median : 0.3237   Median : 0.3270   Median : 0.3270   Median : 0.3303  
 Mean   : 0.3202   Mean   : 0.3269   Mean   : 0.3264   Mean   : 0.3237  
 3rd Qu.: 0.9181   3rd Qu.: 0.9238   3rd Qu.: 0.9238   3rd Qu.: 0.9180  
 Max.   : 2.7651   Max.   : 2.7797   Max.   : 2.7797   Max.   : 2.7980  
   NP_958781          NP_958780          NP_958783          NP_958784      
 Min.   :-1.9576   Min.   :-1.9552   Min.   :-1.9552   Min.   :-1.9552  
 1st Qu.:-0.4440   1st Qu.:-0.4458   1st Qu.:-0.4440   1st Qu.:-0.4440  
 Median : 0.3270   Median : 0.3270   Median : 0.3270   Median : 0.3270  
 Mean   : 0.3271   Mean   : 0.3263   Mean   : 0.3259   Mean   : 0.3260  
 3rd Qu.: 0.9277   3rd Qu.: 0.9238   3rd Qu.: 0.9238   3rd Qu.: 0.9238  
 Max.   : 2.7870   Max.   : 2.7797   Max.   : 2.7834   Max.   : 2.7834  
   NP_112598          NP_001611      
 Min.   :-4.9527   Min.   :-2.5751  
 1st Qu.:-1.6741   1st Qu.:-0.5216  
 Median :-0.6021   Median : 0.6948  
 Mean   :-0.3075   Mean   : 0.4579  
 3rd Qu.: 0.8696   3rd Qu.: 1.4394  
 Max.   : 4.9557   Max.   : 3.4365  
```

# Maybe an easier way?

# The `tableone` package

The `tableone` package is meant to create, you guessed it, Table 1.

It is quite a convenient package for most purposes and saves gobs of time

# The tableone package

```r
library(tableone)
tab1 <- CreateTableOne(data=brca[,-1])
tab1
```

```
                      Overall
n                        83
NP_958782 (mean (SD))  0.32 (0.98)
NP_958785 (mean (SD))  0.33 (0.98)
NP_958786 (mean (SD))  0.33 (0.98)
NP_000436 (mean (SD))  0.32 (0.98)
NP_958781 (mean (SD))  0.33 (0.98)
NP_958780 (mean (SD))  0.33 (0.98)
NP_958783 (mean (SD))  0.33 (0.98)
NP_958784 (mean (SD))  0.33 (0.98)
NP_112598 (mean (SD)) -0.31 (2.02)
NP_001611 (mean (SD))  0.46 (1.50)
```

# The tableone package

```
library(tableone)
tab1 <- CreateTableOne(data = brca[-1])
print(tab1, nonnormal = names(brca)[-1])
```

You have to give the variable names of those you think are non-normally distributed and need to be summarized by the median

```
                         Overall
n                           83
NP_958782 (median [IQR])  0.32 [-0.45, 0.92]
NP_958785 (median [IQR])  0.33 [-0.44, 0.92]
NP_958786 (median [IQR])  0.33 [-0.44, 0.92]
NP_000436 (median [IQR])  0.33 [-0.44, 0.92]
NP_958781 (median [IQR])  0.33 [-0.44, 0.93]
NP_958780 (median [IQR])  0.33 [-0.45, 0.92]
NP_958783 (median [IQR])  0.33 [-0.44, 0.92]
NP_958784 (median [IQR])  0.33 [-0.44, 0.92]
NP_112598 (median [IQR]) -0.60 [-1.67, 0.87]
NP_001611 (median [IQR])  0.69 [-0.52, 1.44]
```

# The tableone package

```r
library(tableone)
tab1 <- CreateTableOne(data = brca[-1])
kableone(print(tab1, nonnormal = names(brca)[-1]),
         format='html')
```

| | Overall |
|---|---|
| n | 83 |
| NP_958782 (median [IQR]) | 0.32 [-0.45, 0.92] |
| NP_958785 (median [IQR]) | 0.33 [-0.44, 0.92] |
| NP_958786 (median [IQR]) | 0.33 [-0.44, 0.92] |
| NP_000436 (median [IQR]) | 0.33 [-0.44, 0.92] |
| NP_958781 (median [IQR]) | 0.33 [-0.44, 0.93] |
| NP_958780 (median [IQR]) | 0.33 [-0.45, 0.92] |
| NP_958783 (median [IQR]) | 0.33 [-0.44, 0.92] |
| NP_958784 (median [IQR]) | 0.33 [-0.44, 0.92] |
| NP_112598 (median [IQR]) | -0.60 [-1.67, 0.87] |
| NP_001611 (median [IQR]) | 0.69 [-0.52, 1.44] |

# Mixed data

Let's first put the expression and clinical data together

```
library(rio)
brca1 <- import('../data/clinical_data_breast_cancer_hw.csv')
brca2 <- import('../data/BreastCancer_Expression.csv')
brca <- left_join(brca1, brca2, by=c('Complete.TCGA.ID' = 'TCGA_ID')) %>%
  mutate(Age.at.Initial.Pathologic.Diagnosis =
           as.numeric(Age.at.Initial.Pathologic.Diagnosis)) %>%
  mutate(ER.Status = ifelse(ER.Status %in% c('Positive','Negative'),
                            ER.Status, NA))

summary(brca)
```

```
 Complete.TCGA.ID      Gender          Age.at.Initial.Pathologic.Diagnosis
 Length:108         Length:108         Min.   :30.00
 Class :character   Class :character   1st Qu.:49.00
 Mode  :character   Mode  :character   Median :58.00
                                       Mean   :58.72
                                       3rd Qu.:66.50
                                       Max.   :88.00
                                       NA's   :1
  ER.Status          PR.Status         HER2.Final.Status     Tumor
 Length:108         Length:108         Length:108         Length:108
 Class :character   Class :character   Class :character   Class :character
 Mode  :character   Mode  :character   Mode  :character   Mode  :character




     Node            Metastasis          AJCC.Stage        Vital.Status
 Length:108         Length:108         Length:108         Length:108
 Class :character   Class :character   Class :character   Class :character
 Mode  :character   Mode  :character   Mode  :character   Mode  :character
```

Let's first put the expression and clinical data together

```
library(rio)
brca1 <- import('../data/clinical_data_breast_cancer_hw.csv')
brca2 <- import('../data/BreastCancer_Expression.csv')
brca <- left_join(brca1, brca2, by=c('Complete.TCGA.ID' = 'TCGA_ID')) %>%
  mutate(Age.at.Initial.Pathologic.Diagnosis =
           as.numeric(Age.at.Initial.Pathologic.Diagnosis)) %>%
  mutate(ER.Status = ifelse(ER.Status %in% c('Positive','Negative'),
                            ER.Status, NA),
         HER2.Final.Status = ifelse(HER2.Final.Status=='Equivocal',
                                    NA, HER2.Final.Status)) %>%
  mutate(across(is.character, as.factor)) %>%
  mutate(Complete.TCGA.ID = as.character(Complete.TCGA.ID))

str(brca)
```

```
'data.frame':	108 obs. of  23 variables:
 $ Complete.TCGA.ID                  : chr  "TCGA-A2-A0T2" "TCGA-A2-A0CM" "TCGA-BH-A18V" "TCGA-BH-A18Q" ...
 $ Gender                            : Factor w/ 2 levels "FEMALE","MALE": 1 1 1 1 1 1 1 1 1 1 ...
 $ Age.at.Initial.Pathologic.Diagnosis: num  66 40 48 56 38 57 74 60 61 NA ...
 $ ER.Status                         : Factor w/ 2 levels "Negative","Positive": 1 1 1 1 1 1 1 1 1 1 ...
 $ PR.Status                         : Factor w/ 2 levels "Negative","Positive": 1 1 1 1 1 1 1 1 1 1 ...
 $ HER2.Final.Status                 : Factor w/ 2 levels "Negative","Positive": 1 1 1 1 1 1 1 1 1 1 ...
 $ Tumor                             : Factor w/ 4 levels "T1","T2","T3",..: 3 2 2 2 3 2 3 2 2 2 ...
 $ Node                              : Factor w/ 4 levels "N0","N1","N2",..: 4 1 2 2 4 1 1 1 1 1 ...
 $ Metastasis                        : Factor w/ 2 levels "M0","M1": 2 1 1 1 1 1 1 1 1 1 ...
 $ AJCC.Stage                        : Factor w/ 11 levels "Stage I","Stage IA",..: 11 5 6 6 10 5 6 5 5 5 ...
 $ Vital.Status                      : Factor w/ 2 levels "DECEASED","LIVING": 1 1 1 1 2 2 2 2 2 2 ...
 $ Days.to.Date.of.Last.Contact      : int  240 754 1555 1692 133 309 425 643 775 964 ...
 $ Days.to.date.of.Death             : int  240 754 1555 1692 NA NA NA NA NA NA ...
 $ NP_958782                         : num  NA 0.683 NA 0.195 NA ...
 $ NP_958785                         : num  NA 0.694 NA 0.215 NA ...
 $ NP_958786                         : num  NA 0.698 NA 0.215 NA ...
```

Identify which variables are categorical (factors) and which are continuous (numeric)

```
catvars <- brca %>% select(where(is.factor)) %>% names()
ctsvars <- brca %>% select(where(is.numeric)) %>% names()
```

```
CreateCatTable(vars = catvars, data = brca)
```

```

                                 Overall
  n                              108
  Gender = MALE (%)                 2 ( 1.9)
  ER.Status = Positive (%)         69 (64.5)
  PR.Status = Positive (%)         55 (50.9)
  HER2.Final.Status = Positive (%) 28 (26.2)
  Tumor (%)
     T1                            16 (14.8)
     T2                            67 (62.0)
     T3                            19 (17.6)
     T4                             6 ( 5.6)
  Node (%)
     N0                            54 (50.0)
     N1                            30 (27.8)
     N2                            15 (13.9)
     N3                             9 ( 8.3)
  Metastasis = M1 (%)               2 ( 1.9)
  AJCC.Stage (%)
     Stage I                        3 ( 2.8)
     Stage IA                       7 ( 6.5)
     Stage IB                       2 ( 1.9)
     Stage II                      11 (10.2)
     Stage IIA                     32 (29.6)
     Stage IIB                     23 (21.3)
     Stage III                      4 ( 3.7)
     Stage IIIA                    12 (11.1)
     Stage IIIB                     6 ( 5.6)
     Stage IIIC                     6 ( 5.6)
     Stage IV                       2 ( 1.9)
  Vital.Status = LIVING (%)        97 (89.8)
```

```
CreateContTable(vars = ctsvars, data = brca)
```

```
                                                  Overall          
  n                                               108              
  Age.at.Initial.Pathologic.Diagnosis (mean (SD))   58.72 (13.21)  
  Days.to.Date.of.Last.Contact (mean (SD))         806.37 (667.70) 
  Days.to.date.of.Death (mean (SD))               1254.45 (678.05) 
  NP_958782 (mean (SD))                              0.32 (0.99)   
  NP_958785 (mean (SD))                              0.33 (1.00)   
  NP_958786 (mean (SD))                              0.33 (1.00)   
  NP_000436 (mean (SD))                              0.32 (0.99)   
  NP_958781 (mean (SD))                              0.33 (1.00)   
  NP_958780 (mean (SD))                              0.33 (1.00)   
  NP_958783 (mean (SD))                              0.33 (1.00)   
  NP_958784 (mean (SD))                              0.33 (1.00)   
  NP_112598 (mean (SD))                             -0.30 (2.06)   
  NP_001611 (mean (SD))                              0.38 (1.46)   
```

```
brca <- brca %>%
  rename(
    'Age'='Age.at.Initial.Pathologic.Diagnosis',
    'Last.Contact' = 'Days.to.Date.of.Last.Contact',
    'Death' = 'Days.to.date.of.Death'
  )
ctsvars <- brca %>%
  select(where(is.numeric))%>% names()
CreateContTable(vars = ctsvars, data = brca)
```

```
                         Overall
n                        108
Age (mean (SD))            58.72 (13.21)
Last.Contact (mean (SD))  806.37 (667.70)
Death (mean (SD))        1254.45 (678.05)
NP_958782 (mean (SD))       0.32 (0.99)
NP_958785 (mean (SD))       0.33 (1.00)
NP_958786 (mean (SD))       0.33 (1.00)
NP_000436 (mean (SD))       0.32 (0.99)
NP_958781 (mean (SD))       0.33 (1.00)
NP_958780 (mean (SD))       0.33 (1.00)
NP_958783 (mean (SD))       0.33 (1.00)
NP_958784 (mean (SD))       0.33 (1.00)
NP_112598 (mean (SD))      -0.30 (2.06)
NP_001611 (mean (SD))       0.38 (1.46)
```

# Putting it together

```
CreateTableOne(vars = c(catvars, ctsvars),
               data = brca)
```

```
                                   Overall
  n                                    108
  Gender = MALE (%)                      2 ( 1.9)
  ER.Status = Positive (%)              69 (64.5)
  PR.Status = Positive (%)              55 (50.9)
  HER2.Final.Status = Positive (%)      28 (26.2)
  Tumor (%)
     T1                                 16 (14.8)
     T2                                 67 (62.0)
     T3                                 19 (17.6)
     T4                                  6 ( 5.6)
  Node (%)
     N0                                 54 (50.0)
     N1                                 30 (27.8)
     N2                                 15 (13.9)
     N3                                  9 ( 8.3)
  Metastasis = M1 (%)                    2 ( 1.9)
  AJCC.Stage (%)
     Stage I                             3 ( 2.8)
     Stage IA                            7 ( 6.5)
     Stage IB                            2 ( 1.9)
     Stage II                           11 (10.2)
     Stage IIA                          32 (29.6)
     Stage IIB                          23 (21.3)
     Stage III                           4 ( 3.7)
```

# Putting it together

```
CreateTableOne(data = brca[,-1])
```

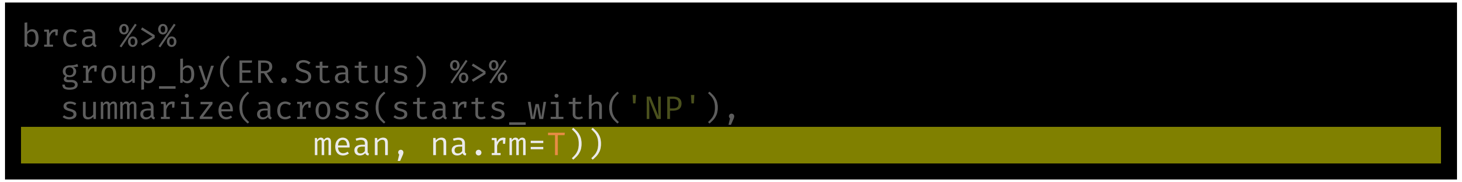
```
                                     Overall
  n                                    108
  Gender = MALE (%)                      2 ( 1.9)
  Age (mean (SD))                    58.72 (13.21)
  ER.Status = Positive (%)              69 (64.5)
  PR.Status = Positive (%)              55 (50.9)
  HER2.Final.Status = Positive (%)      28 (26.2)
  Tumor (%)
     T1                                 16 (14.8)
     T2                                 67 (62.0)
     T3                                 19 (17.6)
     T4                                  6 ( 5.6)
  Node (%)
     N0                                 54 (50.0)
     N1                                 30 (27.8)
     N2                                 15 (13.9)
     N3                                  9 ( 8.3)
  Metastasis = M1 (%)                    2 ( 1.9)
  AJCC.Stage (%)
     Stage I                             3 ( 2.8)
     Stage IA                            7 ( 6.5)
     Stage IB                            2 ( 1.9)
     Stage II                           11 (10.2)
     Stage IIA                          32 (29.6)
     Stage IIB                          23 (21.3)
     Stage III                           4 ( 3.7)
```

# Grouped summaries

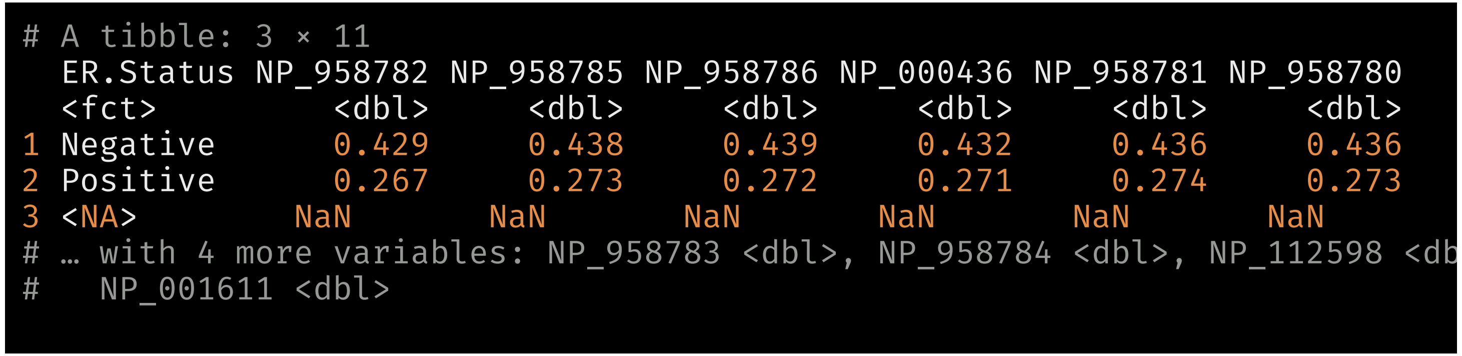
```
brca %>%
  group_by(ER.Status) %>%
  summarize(across(starts_with('NP'),
                   mean))
```

```
# A tibble: 3 × 11
  ER.Status NP_958782 NP_958785 NP_958786 NP_000436 NP_958781 NP_958780
  <fct>         <dbl>     <dbl>     <dbl>     <dbl>     <dbl>     <dbl>
1 Negative         NA        NA        NA        NA        NA        NA
2 Positive         NA        NA        NA        NA        NA        NA
3 <NA>             NA        NA        NA        NA        NA        NA
# … with 4 more variables: NP_958783 <dbl>, NP_958784 <dbl>, NP_112598 <db
#   NP_001611 <dbl>
```

There are missing values now, so we have to use `na.rm=T`.

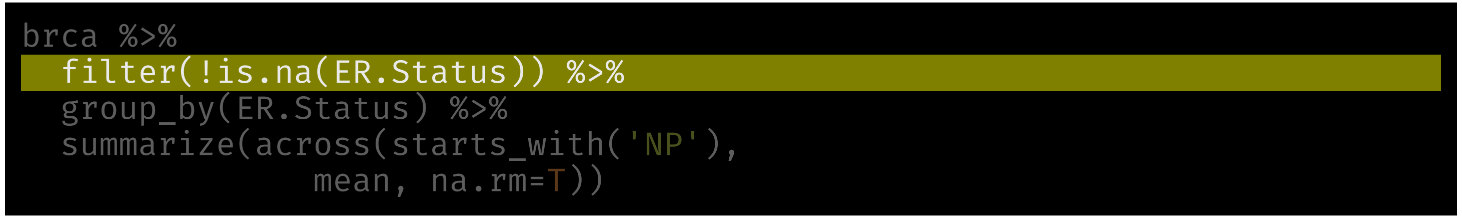


```
brca %>%
  group_by(ER.Status) %>%
  summarize(across(starts_with('NP'),
                mean, na.rm=T))
```

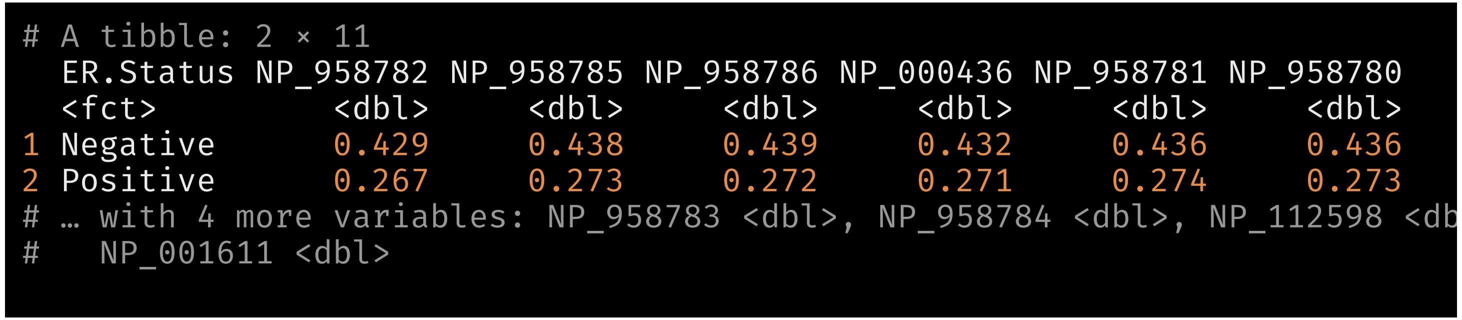


```
# A tibble: 3 × 11
  ER.Status NP_958782 NP_958785 NP_958786 NP_000436 NP_958781 NP_958780
  <fct>         <dbl>     <dbl>     <dbl>     <dbl>     <dbl>     <dbl>
1 Negative      0.429     0.438     0.439     0.432     0.436     0.436
2 Positive      0.267     0.273     0.272     0.271     0.274     0.273
3 <NA>        NaN       NaN       NaN       NaN       NaN       NaN
# … with 4 more variables: NP_958783 <dbl>, NP_958784 <dbl>, NP_112598 <db
#   NP_001611 <dbl>
```

We still have a row for the missing values of ER.Status

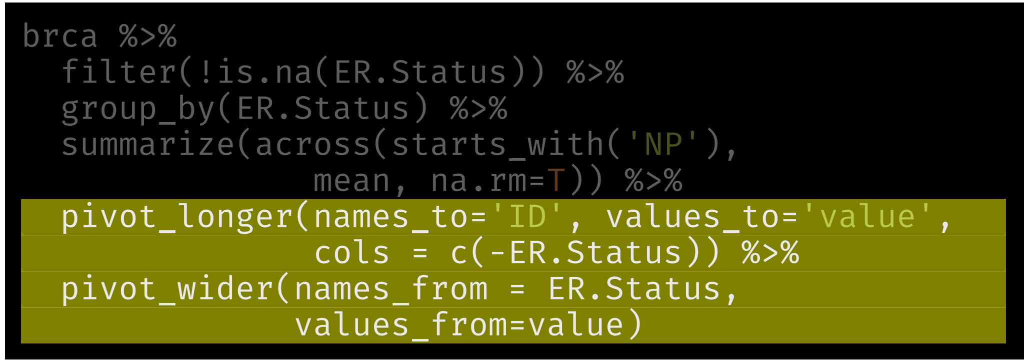


```
brca %>%
  filter(!is.na(ER.Status)) %>%
  group_by(ER.Status) %>%
  summarize(across(starts_with('NP'),
                   mean, na.rm=T))
```

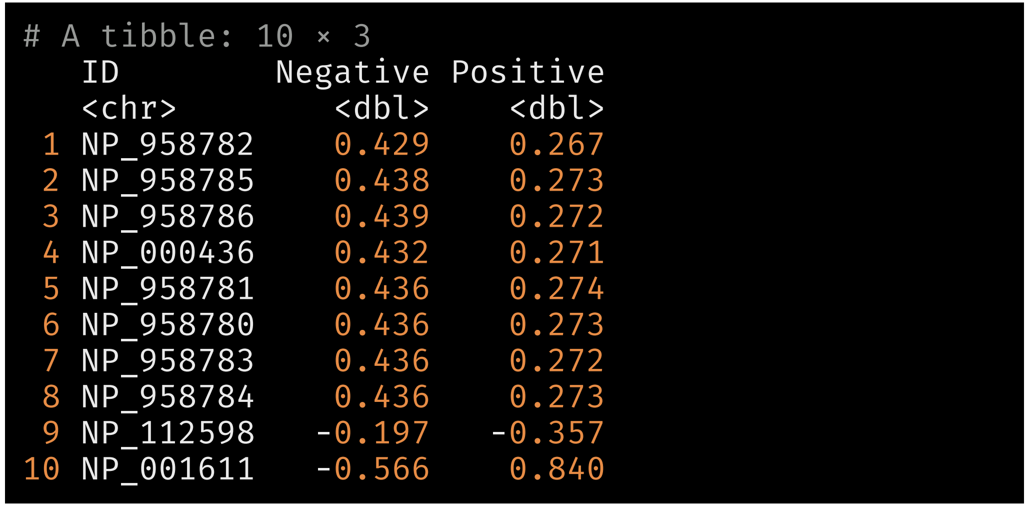


```
# A tibble: 2 × 11
  ER.Status NP_958782 NP_958785 NP_958786 NP_000436 NP_958781 NP_958780
  <fct>         <dbl>     <dbl>     <dbl>     <dbl>     <dbl>     <dbl>
1 Negative      0.429     0.438     0.439     0.432     0.436     0.436
2 Positive      0.267     0.273     0.272     0.271     0.274     0.273
# … with 4 more variables: NP_958783 <dbl>, NP_958784 <dbl>, NP_112598 <db
#   NP_001611 <dbl>
```

How about reversing the rows and columns for readability



```
brca %>%
  filter(!is.na(ER.Status)) %>%
  group_by(ER.Status) %>%
  summarize(across(starts_with('NP'),
                   mean, na.rm=T)) %>%
  pivot_longer(names_to='ID', values_to='value',
               cols = c(-ER.Status)) %>%
  pivot_wider(names_from = ER.Status,
              values_from=value)
```



```
# A tibble: 10 × 3
   ID        Negative Positive
   <chr>        <dbl>    <dbl>
 1 NP_958782    0.429    0.267
 2 NP_958785    0.438    0.273
 3 NP_958786    0.439    0.272
 4 NP_000436    0.432    0.271
 5 NP_958781    0.436    0.274
 6 NP_958780    0.436    0.273
 7 NP_958783    0.436    0.272
 8 NP_958784    0.436    0.273
 9 NP_112598   -0.197   -0.357
10 NP_001611   -0.566    0.840
```

# Using tableone

```
CreateTableOne(
  data = brca %>% filter(!is.na(ER.Status)),
  vars = brca %>%
    select(starts_with('NP')) %>%
    names(),
  strata = 'ER.Status', # single quotes, not backticks
  test = F)
```

```
                       Stratified by ER.Status
                        Negative     Positive
  n                        38           69
  NP_958782 (mean (SD))  0.43 (1.13)  0.27 (0.93)
  NP_958785 (mean (SD))  0.44 (1.14)  0.27 (0.93)
  NP_958786 (mean (SD))  0.44 (1.14)  0.27 (0.93)
  NP_000436 (mean (SD))  0.43 (1.14)  0.27 (0.93)
  NP_958781 (mean (SD))  0.44 (1.14)  0.27 (0.93)
  NP_958780 (mean (SD))  0.44 (1.14)  0.27 (0.93)
  NP_958783 (mean (SD))  0.44 (1.14)  0.27 (0.93)
  NP_958784 (mean (SD))  0.44 (1.14)  0.27 (0.93)
  NP_112598 (mean (SD)) -0.20 (2.28) -0.36 (1.97)
  NP_001611 (mean (SD)) -0.57 (1.54)  0.84 (1.19)
```

# Alternatives to tableone

- table1
- gtsummary
- flextable
- arsenal

# arsenal

```
library(arsenal)
summary(tableby(ER.Status ~ ., data = brca[,-1])) # Here . implies all other variables.
```

| | Negative (N=38) | Positive (N=69) | Total (N=107) | p value |
|---|---|---|---|---|
| Gender | | | | 0.289 |
| FEMALE | 38 (100.0%) | 67 (97.1%) | 105 (98.1%) | |
| MALE | 0 (0.0%) | 2 (2.9%) | 2 (1.9%) | |
| Age | | | | 0.101 |
| N-Miss | 1 | 0 | 1 | |
| Mean (SD) | 55.919 (12.269) | 60.348 (13.573) | 58.802 (13.245) | |
| Range | 36.000 - 82.000 | 30.000 - 88.000 | 30.000 - 88.000 | |
| PR.Status | | | | < 0.001 |
| Negative | 38 (100.0%) | 14 (20.3%) | 52 (48.6%) | |
| Positive | 0 (0.0%) | 55 (79.7%) | 55 (51.4%) | |
| HER2.Final.Status | | | | 0.281 |
| N-Miss | 0 | 1 | 1 | |
| Negative | 26 (68.4%) | 53 (77.9%) | 79 (74.5%) | |