MiCM_data_wrangling_workshop

Lang Liu

10/10/2022

this a workshop

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.2 --
## v ggplot2 3.3.5      v purrr   0.3.4
## v tibble  3.1.8      v dplyr   1.0.10
## v tidyr   1.1.4      v stringr 1.4.0
## v readr   2.1.3      v forcats 0.5.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
data(iris)
head(iris)
```

```
##   Sepal.Length Sepal.Width Petal.Length Petal.Width Species
## 1          5.1         3.5          1.4         0.2  setosa
## 2          4.9         3.0          1.4         0.2  setosa
## 3          4.7         3.2          1.3         0.2  setosa
## 4          4.6         3.1          1.5         0.2  setosa
## 5          5.0         3.6          1.4         0.2  setosa
## 6          5.4         3.9          1.7         0.4  setosa
```

```
library(tibble)
iris_tibble = as_tibble(iris)
head(iris_tibble)
```

```
## # A tibble: 6 x 5
##   Sepal.Length Sepal.Width Petal.Length Petal.Width Species
##          <dbl>       <dbl>        <dbl>       <dbl> <fct>
## 1          5.1         3.5          1.4         0.2 setosa
## 2          4.9         3            1.4         0.2 setosa
## 3          4.7         3.2          1.3         0.2 setosa
## 4          4.6         3.1          1.5         0.2 setosa
## 5          5           3.6          1.4         0.2 setosa
## 6          5.4         3.9          1.7         0.4 setosa
```

```
class(iris_tibble)
```

```
## [1] "tbl_df"     "tbl"        "data.frame"
```

```
class(iris)
```

```
## [1] "data.frame"
```

```
iris$workshop
```

```
## NULL
```

```
iris_tibble$workshop
```

```
## Warning: Unknown or uninitialised column: ‘workshop‘.
```

```
## NULL
```

```
iris_tibble %>% summarise_all(~(sum(is.na(.))))
```

```
## # A tibble: 1 x 5
##   Sepal.Length Sepal.Width Petal.Length Petal.Width Species
##          <int>       <int>        <int>       <int>   <int>
## 1            0           0            0           0       0
```

```
#readr
df <- read_csv("../breast_cancer1.csv")
```

```
## Rows: 151 Columns: 32
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr  (1): type
## dbl (31): samples, 222859_s_at, 243182_at, 221157_s_at, 211521_s_at, 223297_...
##
## i Use ‘spec()‘ to retrieve the full column specification for this data.
## i Specify the column types or set ‘show_col_types = FALSE‘ to quiet this message.
```

```
path = "../breast_cancer_new.csv"
write_csv(df,path)
```

```
#dplyr
#filter
#these three expression are equivalent
filter(iris_tibble,Sepal.Length > 4)
```

```
## # A tibble: 150 x 5
##    Sepal.Length Sepal.Width Petal.Length Petal.Width Species
##           <dbl>       <dbl>        <dbl>       <dbl> <fct>
##  1          5.1         3.5          1.4         0.2 setosa
##  2          4.9         3            1.4         0.2 setosa
##  3          4.7         3.2          1.3         0.2 setosa
##  4          4.6         3.1          1.5         0.2 setosa
##  5          5           3.6          1.4         0.2 setosa
##  6          5.4         3.9          1.7         0.4 setosa
##  7          4.6         3.4          1.4         0.3 setosa
##  8          5           3.4          1.5         0.2 setosa
##  9          4.4         2.9          1.4         0.2 setosa
## 10          4.9         3.1          1.5         0.1 setosa
## # ... with 140 more rows
```

```
iris_tibble %>% filter(Sepal.Length > 4)
```

```
## # A tibble: 150 x 5
##    Sepal.Length Sepal.Width Petal.Length Petal.Width Species
##           <dbl>       <dbl>        <dbl>       <dbl> <fct>
##  1          5.1         3.5          1.4         0.2 setosa
##  2          4.9         3            1.4         0.2 setosa
##  3          4.7         3.2          1.3         0.2 setosa
##  4          4.6         3.1          1.5         0.2 setosa
##  5          5           3.6          1.4         0.2 setosa
##  6          5.4         3.9          1.7         0.4 setosa
##  7          4.6         3.4          1.4         0.3 setosa
##  8          5           3.4          1.5         0.2 setosa
##  9          4.4         2.9          1.4         0.2 setosa
## 10          4.9         3.1          1.5         0.1 setosa
## # ... with 140 more rows
```

```
iris_tibble[iris_tibble$Sepal.Length > 4,]
```

```
## # A tibble: 150 x 5
##    Sepal.Length Sepal.Width Petal.Length Petal.Width Species
##           <dbl>       <dbl>        <dbl>       <dbl> <fct>
##  1          5.1         3.5          1.4         0.2 setosa
##  2          4.9         3            1.4         0.2 setosa
##  3          4.7         3.2          1.3         0.2 setosa
##  4          4.6         3.1          1.5         0.2 setosa
##  5          5           3.6          1.4         0.2 setosa
##  6          5.4         3.9          1.7         0.4 setosa
##  7          4.6         3.4          1.4         0.3 setosa
##  8          5           3.4          1.5         0.2 setosa
##  9          4.4         2.9          1.4         0.2 setosa
## 10          4.9         3.1          1.5         0.1 setosa
## # ... with 140 more rows
```

```
#select
#select is for columns only
select(iris_tibble,Species,Sepal.Length)
```

```
## # A tibble: 150 x 2
##    Species Sepal.Length
##    <fct>          <dbl>
##  1 setosa           5.1
##  2 setosa           4.9
##  3 setosa           4.7
##  4 setosa           4.6
##  5 setosa           5
##  6 setosa           5.4
##  7 setosa           4.6
##  8 setosa           5
##  9 setosa           4.4
## 10 setosa           4.9
## # ... with 140 more rows
```

```
iris_tibble %>% select(Species,Sepal.Length)
```

```
## # A tibble: 150 x 2
##    Species Sepal.Length
##    <fct>          <dbl>
##  1 setosa           5.1
##  2 setosa           4.9
##  3 setosa           4.7
##  4 setosa           4.6
##  5 setosa           5
##  6 setosa           5.4
##  7 setosa           4.6
##  8 setosa           5
##  9 setosa           4.4
## 10 setosa           4.9
## # ... with 140 more rows
```

```
iris_tibble[,c("Species","Sepal.Length")]
```

```
## # A tibble: 150 x 2
##    Species Sepal.Length
##    <fct>          <dbl>
##  1 setosa           5.1
##  2 setosa           4.9
##  3 setosa           4.7
##  4 setosa           4.6
##  5 setosa           5
##  6 setosa           5.4
##  7 setosa           4.6
##  8 setosa           5
##  9 setosa           4.4
## 10 setosa           4.9
## # ... with 140 more rows
```

```
#slice
slice(iris_tibble,c(1:3,5:6))#skip row 4
```

```
## # A tibble: 5 x 5
##   Sepal.Length Sepal.Width Petal.Length Petal.Width Species
##          <dbl>       <dbl>        <dbl>       <dbl> <fct>
## 1          5.1         3.5          1.4         0.2 setosa
## 2          4.9         3            1.4         0.2 setosa
## 3          4.7         3.2          1.3         0.2 setosa
## 4          5           3.6          1.4         0.2 setosa
## 5          5.4         3.9          1.7         0.4 setosa
```

```
iris_tibble %>% slice(1:3)
```

```
## # A tibble: 3 x 5
##   Sepal.Length Sepal.Width Petal.Length Petal.Width Species
##          <dbl>       <dbl>        <dbl>       <dbl> <fct>
## 1          5.1         3.5          1.4         0.2 setosa
## 2          4.9         3            1.4         0.2 setosa
## 3          4.7         3.2          1.3         0.2 setosa
```

```
iris_tibble[c(1:3),]
```

```
## # A tibble: 3 x 5
##   Sepal.Length Sepal.Width Petal.Length Petal.Width Species
##          <dbl>       <dbl>        <dbl>       <dbl> <fct>
## 1          5.1         3.5          1.4         0.2 setosa
## 2          4.9         3            1.4         0.2 setosa
## 3          4.7         3.2          1.3         0.2 setosa
```

```
#mutate
#creating new columns
mutate(iris_tibble,Sepal = Sepal.Length + Sepal.Width,
       Petal = Petal.Length + Petal.Width)
```

```
## # A tibble: 150 x 7
##    Sepal.Length Sepal.Width Petal.Length Petal.Width Species Sepal Petal
##           <dbl>       <dbl>        <dbl>       <dbl> <fct>   <dbl> <dbl>
##  1          5.1         3.5          1.4         0.2 setosa    8.6   1.6
##  2          4.9         3            1.4         0.2 setosa    7.9   1.6
##  3          4.7         3.2          1.3         0.2 setosa    7.9   1.5
##  4          4.6         3.1          1.5         0.2 setosa    7.7   1.7
##  5          5           3.6          1.4         0.2 setosa    8.6   1.6
##  6          5.4         3.9          1.7         0.4 setosa    9.3   2.1
##  7          4.6         3.4          1.4         0.3 setosa    8     1.7
##  8          5           3.4          1.5         0.2 setosa    8.4   1.7
##  9          4.4         2.9          1.4         0.2 setosa    7.3   1.6
## 10          4.9         3.1          1.5         0.1 setosa    8     1.6
## # ... with 140 more rows
```

```
iris_tibble %>% mutate(Sepal = Sepal.Length + Sepal.Width)
```

```
## # A tibble: 150 x 6
##    Sepal.Length Sepal.Width Petal.Length Petal.Width Species Sepal
##           <dbl>       <dbl>        <dbl>       <dbl> <fct>   <dbl>
##  1          5.1         3.5          1.4         0.2 setosa    8.6
##  2          4.9         3            1.4         0.2 setosa    7.9
##  3          4.7         3.2          1.3         0.2 setosa    7.9
##  4          4.6         3.1          1.5         0.2 setosa    7.7
##  5          5           3.6          1.4         0.2 setosa    8.6
##  6          5.4         3.9          1.7         0.4 setosa    9.3
##  7          4.6         3.4          1.4         0.3 setosa    8
##  8          5           3.4          1.5         0.2 setosa    8.4
##  9          4.4         2.9          1.4         0.2 setosa    7.3
## 10          4.9         3.1          1.5         0.1 setosa    8
## # ... with 140 more rows
```

```
iris_tibble["Sepal"] =  iris_tibble$Sepal.Length + iris_tibble$Sepal.Width
```

```
#all together
mutate(slice(select(filter(iris_tibble,Sepal.Length > 4),Species,Sepal.Length,Sepal.Width),1:3),Sepal =
```

```
## # A tibble: 3 x 4
##   Species Sepal.Length Sepal.Width Sepal
##   <fct>          <dbl>       <dbl> <dbl>
## 1 setosa           5.1         3.5   8.6
## 2 setosa           4.9         3     7.9
## 3 setosa           4.7         3.2   7.9
```

```
iris_tibble %>%
  select(Species,Sepal.Length,Sepal.Width) %>%
  slice(1:3) %>%
  mutate(Sepal = Sepal.Length + Sepal.Width) %>%
  filter(Sepal.Length > 4)
```

```
## # A tibble: 3 x 4
##   Species Sepal.Length Sepal.Width Sepal
##   <fct>          <dbl>       <dbl> <dbl>
## 1 setosa           5.1         3.5   8.6
## 2 setosa           4.9         3     7.9
## 3 setosa           4.7         3.2   7.9
```

```
iris_tibble_subset <- iris_tibble %>%
  select(Species,Sepal.Length,Sepal.Width) %>%
  slice(1:3) %>%
  mutate(Sepal = Sepal.Length + Sepal.Width) %>%
  filter(Sepal.Length > 4)
```

```
#Outliers
#create a sample tibble
element <- sample(letters[1:5], 1e4, replace=T)
value <- rnorm(1e4)
df <- tibble(element, value)
#calculate means in general
df %>%
  select(value) %>%
  summarise_all(list(avg=mean))
```

```
## # A tibble: 1 x 1
##      avg
##    <dbl>
## 1 0.0213
```

```
#after group_by(element)
df %>%
  group_by(element) %>%
  summarise_all(list(avg=mean,med = median))
```

```
## # A tibble: 5 x 3
##   element     avg       med
##   <chr>     <dbl>     <dbl>
## 1 a       0.0120   0.000272
## 2 b       0.00287  0.00716
## 3 c       0.0296   0.0337
## 4 d       0.0106  -0.0115
## 5 e       0.0515   0.0595
```

```
#before removing outliers
df %>%
  group_by(element) %>%
  summarise_all(list(mean=mean,sd=sd))
```

```
## # A tibble: 5 x 3
##   element    mean    sd
##   <chr>     <dbl> <dbl>
## 1 a       0.0120  0.998
## 2 b       0.00287 0.980
## 3 c       0.0296  1.03
## 4 d       0.0106  1.02
## 5 e       0.0515  1.01
```

```
#after removing outliers
df %>%
  group_by(element) %>%
  filter(!(abs(value - median(value)) > 2*sd(value))) %>%
  summarise_all(list(mean=mean,sd=sd))
```

```
## # A tibble: 5 x 3
##   element      mean    sd
##   <chr>      <dbl> <dbl>
## 1 a       -0.0103  0.859
## 2 b        0.0191  0.853
## 3 c        0.0421  0.899
## 4 d        0.00399 0.892
## 5 e        0.0519  0.875
```

```
#distinct
iris_tibble %>%
  distinct(Species,.keep_all = TRUE)
```

```
## # A tibble: 3 x 6
##   Sepal.Length Sepal.Width Petal.Length Petal.Width Species    Sepal
##          <dbl>       <dbl>        <dbl>       <dbl> <fct>      <dbl>
## 1          5.1         3.5          1.4         0.2 setosa       8.6
## 2          7           3.2          4.7         1.4 versicolor  10.2
## 3          6.3         3.3          6           2.5 virginica    9.6
```

```
#summarise
iris_tibble %>%
  filter(Sepal.Length > 4) %>%
  select(Species,Sepal.Length,Sepal.Width) %>%
  slice(1:3) %>%
  mutate(Sepal = Sepal.Length + Sepal.Width) %>%
  summarise(sum_length = sum(Sepal.Length),
            sum_width = sum(Sepal.Width),
            sum_sepal = sum(Sepal))
```

```
## # A tibble: 1 x 3
##   sum_length sum_width sum_sepal
##        <dbl>     <dbl>     <dbl>
## 1       14.7       9.7      24.4
```

```
iris_tibble %>%
  filter(Sepal.Length > 4) %>%
  select(Species,Sepal.Length,Sepal.Width) %>%
  slice(1:3) %>%
  mutate(Sepal = Sepal.Length + Sepal.Width) %>%
  select(-Species) %>%
  summarise_all(list(sum=sum))
```

```
## # A tibble: 1 x 3
##   Sepal.Length_sum Sepal.Width_sum Sepal_sum
##              <dbl>           <dbl>     <dbl>
## 1             14.7             9.7      24.4
```

```
#summarise_all
iris_tibble %>%
  filter(Sepal.Length > 4) %>%
  select(Species,Sepal.Length,Sepal.Width) %>%
  slice(1:3) %>%
  mutate(Sepal = Sepal.Length + Sepal.Width) %>%
  select(Sepal.Length,Sepal.Width,Sepal) %>%
  summarise_all(list(total=sum))
```

```
## # A tibble: 1 x 3
##   Sepal.Length_total Sepal.Width_total Sepal_total
##                <dbl>             <dbl>       <dbl>
## 1               14.7               9.7        24.4
```

```
#group_by
iris_tibble %>%
  group_by(Species) %>%
  summarise_all(list(avg = mean,total = sum))
```

```
## # A tibble: 3 x 11
##   Species    Sepal.Len~1 Sepal~2 Petal~3 Petal~4 Sepal~5 Sepal~6 Sepal~7 Petal~8
##   <fct>            <dbl>   <dbl>   <dbl>   <dbl>   <dbl>   <dbl>   <dbl>   <dbl>
## 1 setosa            5.01    3.43    1.46   0.246    8.43    250.    171.    73.1
## 2 versicolor        5.94    2.77    4.26   1.33     8.71    297.    138.   213
## 3 virginica         6.59    2.97    5.55   2.03     9.56    329.    149.   278.
## # ... with 2 more variables: Petal.Width_total <dbl>, Sepal_total <dbl>, and
## #   abbreviated variable names 1: Sepal.Length_avg, 2: Sepal.Width_avg,
## #   3: Petal.Length_avg, 4: Petal.Width_avg, 5: Sepal_avg,
## #   6: Sepal.Length_total, 7: Sepal.Width_total, 8: Petal.Length_total
```

```
#arrange can reorder sample
#by default, the order is ascending
#to get descending order, call desc() on subject
iris_tibble %>%
  group_by(Species) %>%
  summarise_all(list(avg = mean,total = sum)) %>%
  arrange(desc(Sepal.Width_avg))
```

```
## # A tibble: 3 x 11
##   Species    Sepal.Len~1 Sepal~2 Petal~3 Petal~4 Sepal~5 Sepal~6 Sepal~7 Petal~8
##   <fct>            <dbl>   <dbl>   <dbl>   <dbl>   <dbl>   <dbl>   <dbl>   <dbl>
## 1 setosa            5.01    3.43    1.46   0.246    8.43    250.    171.    73.1
## 2 virginica         6.59    2.97    5.55   2.03     9.56    329.    149.   278.
## 3 versicolor        5.94    2.77    4.26   1.33     8.71    297.    138.   213
## # ... with 2 more variables: Petal.Width_total <dbl>, Sepal_total <dbl>, and
## #   abbreviated variable names 1: Sepal.Length_avg, 2: Sepal.Width_avg,
## #   3: Petal.Length_avg, 4: Petal.Width_avg, 5: Sepal_avg,
## #   6: Sepal.Length_total, 7: Sepal.Width_total, 8: Petal.Length_total
```

```
#pivot_longer
#cols selects columns that will go into the rows
#names_to names the columns of the new column
#values_to defines the column name of values associated with selected columns
iris_tibble %>%
  group_by(Species) %>%
  summarise_all(list(avg = mean,total = sum)) %>%
  pivot_longer(cols = !Species,names_to = "measure", values_to = "value")
```

```
## # A tibble: 30 x 3
##    Species measure              value
##    <fct>   <chr>                <dbl>
##  1 setosa  Sepal.Length_avg     5.01
##  2 setosa  Sepal.Width_avg      3.43
##  3 setosa  Petal.Length_avg     1.46
##  4 setosa  Petal.Width_avg      0.246
##  5 setosa  Sepal_avg            8.43
##  6 setosa  Sepal.Length_total 250.
##  7 setosa  Sepal.Width_total  171.
##  8 setosa  Petal.Length_total  73.1
##  9 setosa  Petal.Width_total   12.3
## 10 setosa  Sepal_total        422.
## # ... with 20 more rows
```

```
#another way to select columns
iris_tibble %>%
  group_by(Species) %>%
  summarise_all(list(avg = mean,total = sum)) %>%
  pivot_longer(cols = contains("_"),names_to = "measure", values_to = "value")
```

```
## # A tibble: 30 x 3
##    Species measure              value
##    <fct>   <chr>                <dbl>
##  1 setosa  Sepal.Length_avg     5.01
##  2 setosa  Sepal.Width_avg      3.43
##  3 setosa  Petal.Length_avg     1.46
##  4 setosa  Petal.Width_avg      0.246
##  5 setosa  Sepal_avg            8.43
##  6 setosa  Sepal.Length_total 250.
##  7 setosa  Sepal.Width_total  171.
##  8 setosa  Petal.Length_total  73.1
##  9 setosa  Petal.Width_total   12.3
## 10 setosa  Sepal_total        422.
## # ... with 20 more rows
```

```
#pivot_wider()
#id_col selects the column that is repetitive
#names_from selects column associated with id_col
#values_from select values
iris_tibble %>%
  group_by(Species) %>%
  summarise_all(list(avg = mean,total = sum)) %>%
  pivot_longer(cols = contains("_"),names_to = "measure", values_to = "value") %>%
  pivot_wider(id_col = measure, names_from = Species, values_from = value)
```

```
## # A tibble: 10 x 4
##    measure             setosa versicolor virginica
##    <chr>                <dbl>      <dbl>     <dbl>
##  1 Sepal.Length_avg     5.01        5.94      6.59
##  2 Sepal.Width_avg      3.43        2.77      2.97
##  3 Petal.Length_avg     1.46        4.26      5.55
##  4 Petal.Width_avg      0.246       1.33      2.03
##  5 Sepal_avg            8.43        8.71      9.56
##  6 Sepal.Length_total 250.        297.      329.
##  7 Sepal.Width_total  171.        138.      149.
##  8 Petal.Length_total  73.1       213       278.
##  9 Petal.Width_total   12.3        66.3     101.
## 10 Sepal_total        422.        435.      478.
```

```
#another example of pivot_wider
df <- data.frame(player=rep(c('A', 'B'), each=2),
                 stat=rep(c('points', 'assists'), times=2),
                 amount=c(14, 6, 18, 7))
df %>% pivot_wider(id_cols = player, names_from = stat, values_from = amount)
```

```
## # A tibble: 2 x 3
##   player points assists
##   <chr>   <dbl>   <dbl>
## 1 A          14       6
## 2 B          18       7
```

```
df %>% pivot_wider(id_cols = stat, names_from = player, values_from = amount)
```

```
## # A tibble: 2 x 3
##   stat         A     B
##   <chr>   <dbl> <dbl>
## 1 points      14    18
## 2 assists      6     7
```

```
#missing values detection
x <- c(1,NA,2)
is.na(x)
```

```
## [1] FALSE  TRUE FALSE
```

```
sum(is.na(x))
```

```
## [1] 1
```

```
iris_tibble %>% summarise_all(~sum(is.na(.)))
```

```
## # A tibble: 1 x 6
##   Sepal.Length Sepal.Width Petal.Length Petal.Width Species Sepal
##          <int>       <int>        <int>       <int>   <int> <int>
## 1            0           0            0           0       0     0
```

```
df <- data.frame(player=rep(c('A', 'B'), each=2),
                 stat=rep(c('points', 'assists'), times=2),
                 amount=c(14,NA, 18, NA))
df %>% summarise_all(~sum(is.na(.)))
```

```
##   player stat amount
## 1      0    0      2
```

```
#drop_na
df %>% drop_na(amount)
```

```
##   player   stat amount
## 1      A points     14
## 2      B points     18
```

```
#fill
df %>% fill(amount)
```

```
##   player    stat amount
## 1      A  points     14
## 2      A assists     14
## 3      B  points     18
## 4      B assists     18
```

```
df %>% fill(amount,.direction="up")
```

```
##   player    stat amount
## 1      A  points     14
## 2      A assists     18
## 3      B  points     18
## 4      B assists     NA
```

```
#replace_na
df$amount <- df$amount %>% replace_na(999)
```

```
#union
a1 <- data.frame(a = 1:5, b=letters[1:5])
a2 <- data.frame(a = 1:3, b=letters[1:3])
#INNER JOIN
merge(a1,a2,by="a",all=FALSE)
```

```
##   a b.x b.y
## 1 1   a   a
## 2 2   b   b
## 3 3   c   c
```

```
#OUTER JOIN
merge(a1,a2,by="a",all=TRUE)
```

```
##   a b.x  b.y
## 1 1   a    a
## 2 2   b    b
## 3 3   c    c
## 4 4   d <NA>
## 5 5   e <NA>
```

```
#LEFT JOIN
merge(a1,a2,by="a",all.x=TRUE)
```

```
##   a b.x  b.y
## 1 1   a    a
## 2 2   b    b
## 3 3   c    c
## 4 4   d <NA>
## 5 5   e <NA>
```

```
#RIGHT JOIN
merge(a1,a2,by="a",all.y=TRUE)
```

```
##   a b.x b.y
## 1 1   a   a
## 2 2   b   b
## 3 3   c   c
```

```
#dplyr
#difference
a1 %>% anti_join(a2,by = "a")
```

```
##   a b
## 1 4 d
## 2 5 e
```

```
a1 %>% semi_join(a2,by = 'a')
```

```
##   a b
## 1 1 a
## 2 2 b
## 3 3 c
```

```
a2 %>% anti_join(a1,by = "a")
```

```
## [1] a b
## <0 rows> (or 0-length row.names)
```