R code clean dairy income data

Rae Zhang

9/25/2021

First get column names

```
(ColNames<-names(Dairy_DF))
```

```
##  [1] "X"                    "year"                 "state"
##  [4] "report"               "farmtype"             "category"
##  [7] "category_value"       "category2"            "category2_value"
## [10] "variable_id"          "variable_name"        "variable_sequence"
## [13] "variable_level"       "variable_group"       "variable_group_id"
## [16] "variable_unit"        "variable_description" "variable_is_invalid"
## [19] "estimate"             "median"               "statistic"
## [22] "rse"                  "unreliable_estimate"  "decimal_display"
```

```
for(name in 1:length(ColNames)){
  cat(ColNames[name], "\n")
}
```

```
## X
## year
## state
## report
## farmtype
## category
## category_value
## category2
## category2_value
## variable_id
## variable_name
## variable_sequence
## variable_level
## variable_group
## variable_group_id
## variable_unit
## variable_description
## variable_is_invalid
## estimate
## median
## statistic
## rse
## unreliable_estimate
## decimal_display
```

```
(NumColumns <-ncol(Dairy_DF))
```

```
## [1] 24
```

```
(NumRows <-nrow(Dairy_DF))
```

```
## [1] 60
```

Make tables of all the columns

MISSING VALUES

check the entire DF for missing values in total

Using an inline function and sapply (for simplify apply)

```
sapply(Dairy_DF, function(x) sum(is.na(x)))
```

```
##          year         state variable_name      estimate        median
##             0             0             0            17            44
```

Clean up missing values for each variable...

It has 17 NA value

```
table(Dairy_DF$estimate)
```

```
##
##    3508    6499    9767   20836   41528   42829   47983   50150   57428   63106
##       1       1       1       1       1       1       1       1       1       1
##   79578  125210  138566  148521  151002  176314  208412  212331  218553  221616
##       1       1       1       1       1       1       1       1       1       1
##  239996  258691  261149  270124  271052  279460  303906  360285  372207  683365
##       1       1       1       1       1       1       1       1       1       1
##  709389  800898  853612  867037  869851  918842  961495 1077314 1150240 1202771
##       1       1       1       1       1       1       1       1       1       1
## 1450632 3979949 6215313
##       1       1       1
```

```
summary(Dairy_DF$estimate)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##    3508  102394  258691  613519  827255 6215313      17
```

```
nrow(Dairy_DF)
```

```
## [1] 60
```

The data is missing 17 values out of 60. That is 28% which is quite a bit. Now make two tables by the two variable name "Gross cash farm income" and "Net farm income" to see under which category has the least NA value. And the three dataframes will be combined into one

```
Gross_Cash_Income <- Dairy_DF[Dairy_DF$variable_name=='Gross cash farm income',]
Gross_Cash_Income <- Gross_Cash_Income[,-3]
colnames(Gross_Cash_Income)[3] <- "Gross cash income"
colnames(Gross_Cash_Income)[4] <- "Gross cash income median"
head(Gross_Cash_Income, n=5)
```

```
##   year      state Gross cash income Gross cash income median
## 1 2019       Iowa           1450632                    504407
## 2 2019     Kansas                NA                        NA
## 3 2019 California           6215313                   4503308
## 4 2019  Wisconsin            800898                    292434
## 5 2019   Missouri            279460                        NA
```

```
Net_farm_income <- Dairy_DF[Dairy_DF$variable_name=='Net farm income',]
Net_farm_income <- Net_farm_income[,-3]
colnames(Net_farm_income)[3] <- "Net farm income"
colnames(Net_farm_income)[4] <- "Net farm income median"
head(Net_farm_income, n=5)
```

```
##    year     state Net farm income Net farm income median
## 46 2019      Iowa          270124                  125599
## 47 2019 Minnesota          208412                  109610
## 48 2019   Florida              NA                      NA
## 49 2019 Wisconsin          176314                   92903
## 50 2019  Missouri           50150                      NA
```

```
Other_related_income <- Dairy_DF[Dairy_DF$variable_name=='Other farm-related income',]
Other_related_income<- subset(Other_related_income, select=-c(variable_name))
colnames(Other_related_income)[3] <- "Other related income"
colnames(Other_related_income)[4] <- "Other related income median"
head(Other_related_income, n=5)
```

```
##    year     state Other related income Other related income median
## 16 2019    Kansas                   NA                         NA
## 17 2019 Minnesota                63106                      32000
## 18 2019   Florida                   NA                         NA
## 19 2019  Missouri                 3508                         NA
## 20 2019   Georgia                 9767                         NA
```

Merge three dataframes together

```
Dairy_DF <- merge(Gross_Cash_Income, Net_farm_income,
                  by.Gross_Cash_Income = "state", by.Net_farm_income = "state")

Dairy_DF <- merge(Dairy_DF, Other_related_income,
                  by.Dairy_DF = "state", by.Other_related_income = "state")


summary(Dairy_DF)
```

```
##       year         state           Gross cash income Gross cash income median
##  Min.   :2019   Length:15          Min.   : 279460   Min.   : 292434
##  1st Qu.:2019   Class :character   1st Qu.: 861732   1st Qu.: 451414
##  Median :2019   Mode  :character   Median : 961495   Median : 517179
##  Mean   :2019                      Mean   :1691831   Mean   :1457525
##  3rd Qu.:2019                      3rd Qu.:1326702   3rd Qu.:1523290
##  Max.   :2019                      Max.   :6215313   Max.   :4503308
##                                    NA's   :4         NA's   :11
##  Net farm income  Net farm income median Other related income
##  Min.   : 50150   Min.   : 92903        Min.   :  3508
##  1st Qu.:148003   1st Qu.:105433        1st Qu.: 15302
##  Median :234780   Median :117604        Median : 42829
##  Mean   :305138   Mean   :178745        Mean   : 75625
##  3rd Qu.:270820   3rd Qu.:190916        3rd Qu.: 71342
##  Max.   :867037   Max.   :386868        Max.   :303906
##  NA's   :5        NA's   :11            NA's   :4
##  Other related income median
##  Min.   :14897
##  1st Qu.:24784
##  Median :30040
##  Mean   :28494
##  3rd Qu.:33750
##  Max.   :39000
##  NA's   :11
```

From the summary, it shows gross cash income column and other related income column has fewer NA values than the other columns

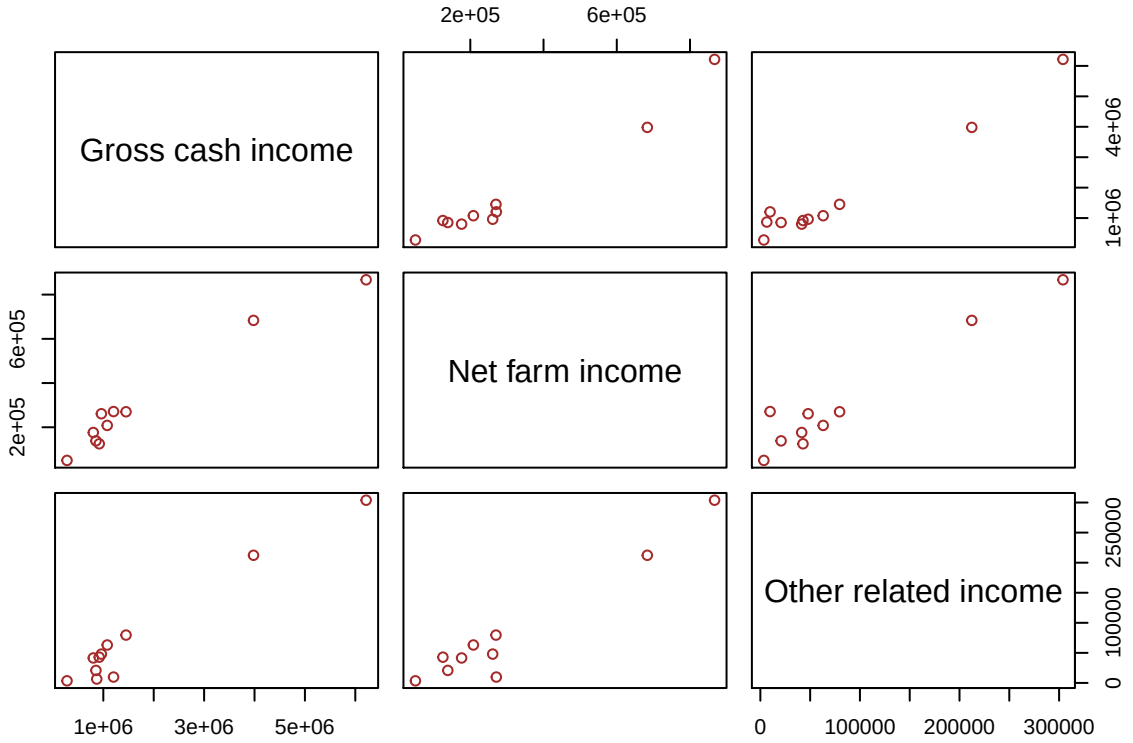
To estimate the missing values by correlations

```
str(Dairy_DF)
```

```
## 'data.frame':    15 obs. of  8 variables:
##  $ year                       : int  2019 2019 2019 2019 2019 2019 2019 2019 2019 2019 ...
##  $ state                      : chr  "Arkansas" "California" "Florida" "Georgia" ...
##  $ Gross cash income          : num  NA 6215313 NA 1202771 961495 ...
##  $ Gross cash income median   : num  NA 4503308 NA NA NA ...
##  $ Net farm income            : num  NA 867037 NA 271052 261149 ...
##  $ Net farm income median     : num  NA 386868 NA NA NA ...
```

```
##  $ Other related income       : num  NA 303906 NA 9767 47983 ...
##  $ Other related income median: num  NA 39000 NA NA NA ...
```

```
pairs(Dairy_DF[,c(3, 5, 7)],na.rm = T, col = "brown")
```



Use the "spearman" and "psych" library to calculate the correlations

```
(Temp<-Dairy_DF[,c(3, 5, 7)])
```

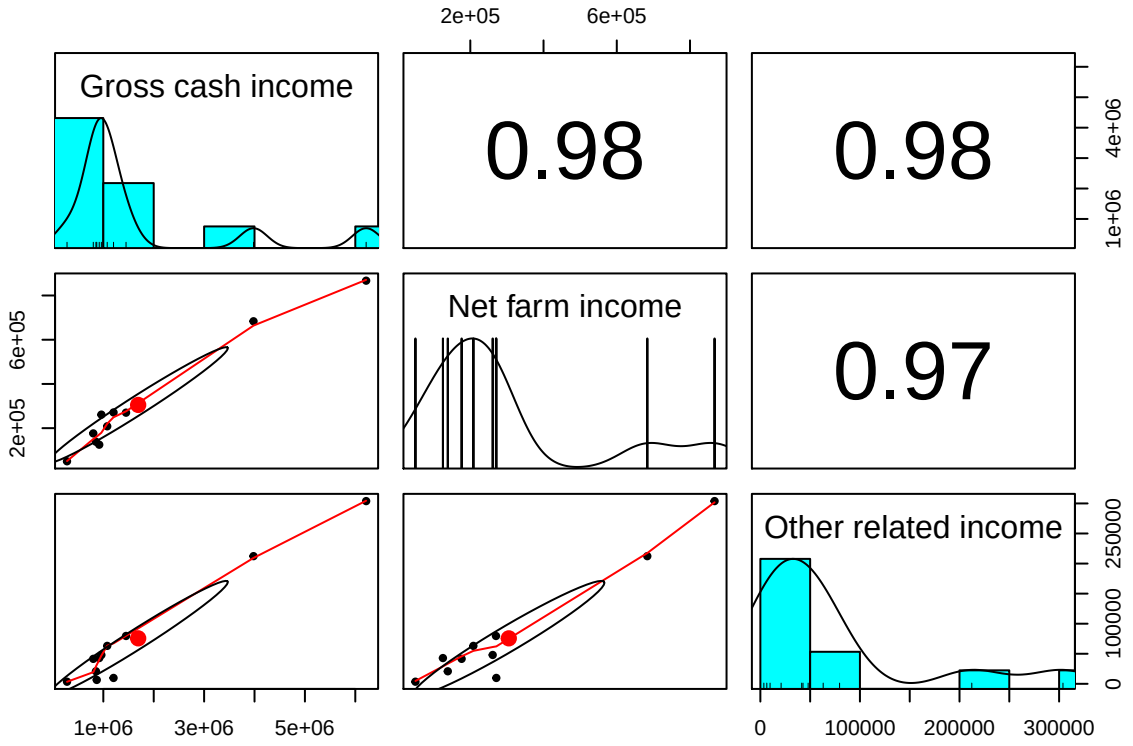
```
##    Gross cash income Net farm income Other related income
## 1                 NA              NA                    NA
## 2            6215313          867037                303906
## 3                 NA              NA                    NA
## 4            1202771          271052                  9767
## 5             961495          261149                 47983
## 6             853612          138566                 20836
## 7            1450632          270124                 79578
## 8                 NA              NA                    NA
## 9            1077314          208412                 63106
## 10            279460           50150                  3508
## 11                NA              NA                    NA
## 12            869851              NA                  6499
## 13           3979949          683365                212331
## 14            918842          125210                 42829
## 15            800898          176314                 41528
```

```
library(psych)
```

```
##
## Attaching package: 'psych'
```

```
## The following objects are masked from 'package:ggplot2':
##
##     %+%, alpha
```

```
pairs.panels(Temp)
```




```
corr.test(Temp, method = "spearman")
```

```
## Call:corr.test(x = Temp, method = "spearman")
## Correlation matrix
##                       Gross cash income Net farm income Other related income
## Gross cash income                  1.00            0.93                 0.81
## Net farm income                    0.93            1.00                 0.70
## Other related income               0.81            0.70                 1.00
## Sample Size
##                       Gross cash income Net farm income Other related income
## Gross cash income                    11              10                   11
## Net farm income                      10              10                   10
```

```
## Other related income                11             10                   11
## Probability values (Entries above the diagonal are adjusted for multiple tests.)
##                      Gross cash income Net farm income Other related income
## Gross cash income                    0            0.00                 0.01
## Net farm income                      0            0.00                 0.03
## Other related income                 0            0.03                 0.00
##
##  To see confidence intervals of the correlations, print with the short=FALSE option
```

The plot shows gross cash income, net farm income, and other related income have strong correlations.

Also, the state Arkansas, Florida, Kansas, and Nebraska have no data, so the columns are delted.

```
Dairy_DF <- subset(Dairy_DF, state!="Arkansas" & state!="Florida" & state!="Kansas"
                   & state!="Nebraska")

summary(Dairy_DF)
```

```
##       year         state           Gross cash income Gross cash income median
##  Min.   :2019   Length:11          Min.   : 279460   Min.   : 292434
##  1st Qu.:2019   Class :character   1st Qu.: 861732   1st Qu.: 451414
##  Median :2019   Mode  :character   Median : 961495   Median : 517179
##  Mean   :2019                      Mean   :1691831   Mean   :1457525
##  3rd Qu.:2019                      3rd Qu.:1326702   3rd Qu.:1523290
##  Max.   :2019                      Max.   :6215313   Max.   :4503308
##                                                      NA's   :7
##  Net farm income  Net farm income median Other related income
##  Min.   : 50150   Min.   : 92903         Min.   :  3508
##  1st Qu.:148003   1st Qu.:105433         1st Qu.: 15302
##  Median :234780   Median :117604         Median : 42829
##  Mean   :305138   Mean   :178745         Mean   : 75625
##  3rd Qu.:270820   3rd Qu.:190916         3rd Qu.: 71342
##  Max.   :867037   Max.   :386868         Max.   :303906
##  NA's   :1        NA's   :7
##  Other related income median
##  Min.   :14897
##  1st Qu.:24784
##  Median :30040
##  Mean   :28494
##  3rd Qu.:33750
##  Max.   :39000
##  NA's   :7
```

The estimation of the relationship between gross cash income and net farm income is net farm income is about 0.128 times than gross cash income, so the missing values in net farm income will be filled by the correlation

```
Dairy_DF$`Net farm income` <-
  ifelse(is.na(Dairy_DF$`Net farm income`),
         0.128*Dairy_DF$`Gross cash income`,
         Dairy_DF$`Net farm income`)
head(Dairy_DF, n=10)
```

```
##    year          state Gross cash income Gross cash income median
## 2  2019     California           6215313                   4503308
## 4  2019        Georgia           1202771                        NA
## 5  2019       Illinois            961495                        NA
## 6  2019        Indiana            853612                        NA
## 7  2019           Iowa           1450632                    504407
## 9  2019      Minnesota           1077314                    529951
## 10 2019       Missouri            279460                        NA
## 12 2019 North Carolina            869851                        NA
## 13 2019          Texas           3979949                        NA
## 14 2019     Washington            918842                        NA
##    Net farm income Net farm income median Other related income
## 2         867037.0                 386868                303906
## 4         271052.0                     NA                 9767
## 5         261149.0                     NA                47983
## 6         138566.0                     NA                20836
## 7         270124.0                 125599                79578
## 9         208412.0                 109610                63106
## 10         50150.0                     NA                 3508
## 12        111340.9                     NA                 6499
## 13        683365.0                     NA               212331
## 14        125210.0                     NA                42829
##    Other related income median
## 2                        39000
## 4                           NA
## 5                           NA
## 6                           NA
## 7                        28080
## 9                        32000
## 10                          NA
## 12                          NA
## 13                          NA
## 14                          NA
```

The median columns doesn't provide too much details, so they are deleted

```
head(Dairy_DF, n=10)
```

```
##    year          state Gross cash income Gross cash income median
## 2  2019     California           6215313                   4503308
```

```
## 4  2019        Georgia           1202771                      NA
## 5  2019       Illinois            961495                      NA
## 6  2019        Indiana            853612                      NA
## 7  2019           Iowa           1450632                  504407
## 9  2019      Minnesota           1077314                  529951
## 10 2019       Missouri            279460                      NA
## 12 2019 North Carolina            869851                      NA
## 13 2019          Texas           3979949                      NA
## 14 2019     Washington            918842                      NA
##    Net farm income Net farm income median Other related income
## 2         867037.0                 386868                303906
## 4         271052.0                     NA                  9767
## 5         261149.0                     NA                 47983
## 6         138566.0                     NA                 20836
## 7         270124.0                 125599                 79578
## 9         208412.0                 109610                 63106
## 10         50150.0                     NA                  3508
## 12        111340.9                     NA                  6499
## 13        683365.0                     NA                212331
## 14        125210.0                     NA                 42829
##    Other related income median
## 2                        39000
## 4                           NA
## 5                           NA
## 6                           NA
## 7                        28080
## 9                        32000
## 10                          NA
## 12                          NA
## 13                          NA
## 14                          NA
```

VISUALIZATION

```
Dairy_DF <- subset(Dairy_DF, select=-c(`Gross cash income median`,
                                       `Net farm income median`,
                                       `Other related income median`))
head(Dairy_DF, n=5)
```

```
##   year      state Gross cash income Net farm income Other related income
## 2 2019 California           6215313          867037                303906
## 4 2019    Georgia           1202771          271052                  9767
## 5 2019   Illinois            961495          261149                 47983
## 6 2019    Indiana            853612          138566                 20836
## 7 2019       Iowa           1450632          270124                 79578
```

Here the reshape library and melt function is being used to make grouped bar graph

```
Dairy_DF_melt <- melt(Dairy_DF[,c('state','Gross cash income','Net farm income',
                                  'Other related income')],id.vars = 1)
```

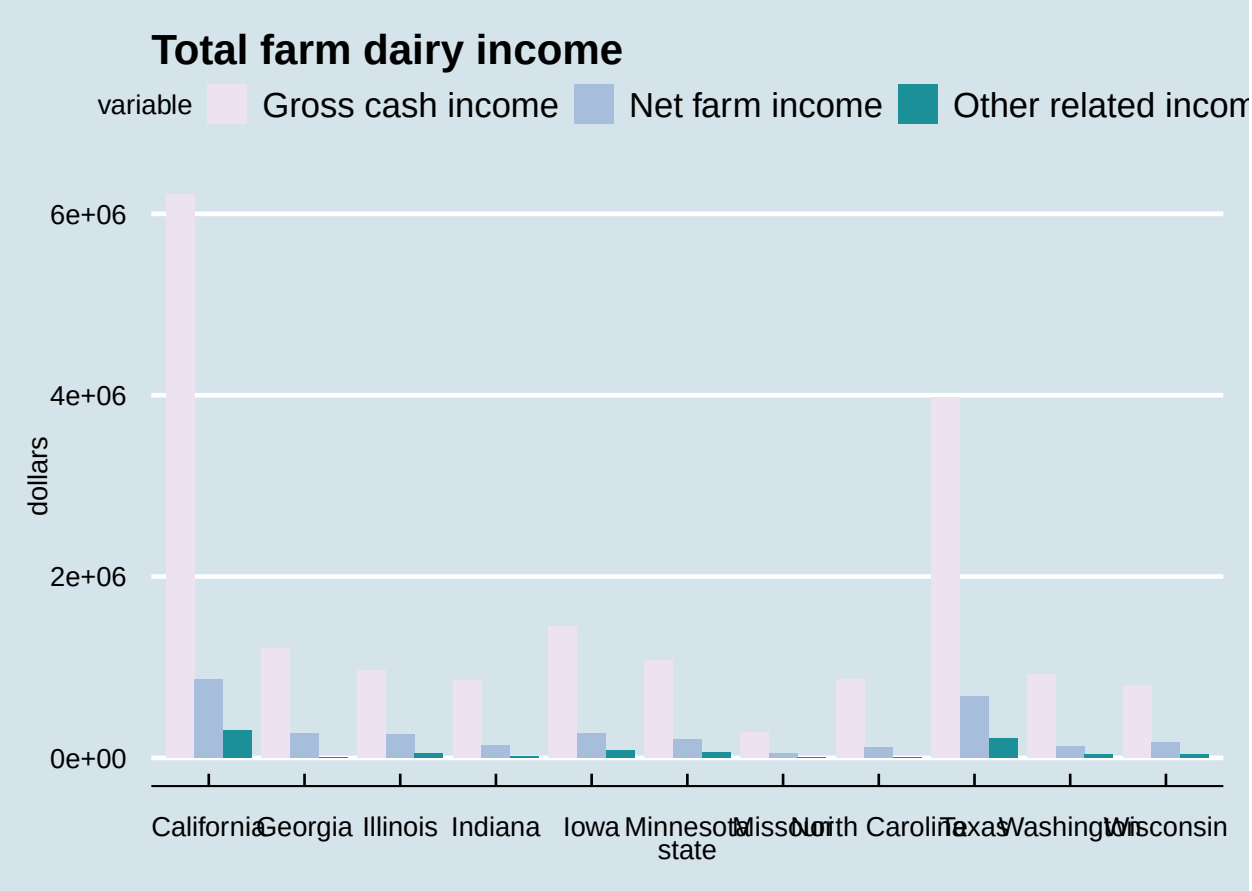
Round columns

```
Dairy_DF <- Dairy_DF %>% mutate(across(where(is.numeric), round, 0))
Dairy_DF_melt <- Dairy_DF_melt %>% mutate(across(where(is.numeric), round, 0))
head(Dairy_DF_melt, n=5)
```

```
##        state          variable   value
## 1 California Gross cash income 6215313
## 2    Georgia Gross cash income 1202771
## 3   Illinois Gross cash income  961495
## 4    Indiana Gross cash income  853612
## 5       Iowa Gross cash income 1450632
```

Visualization

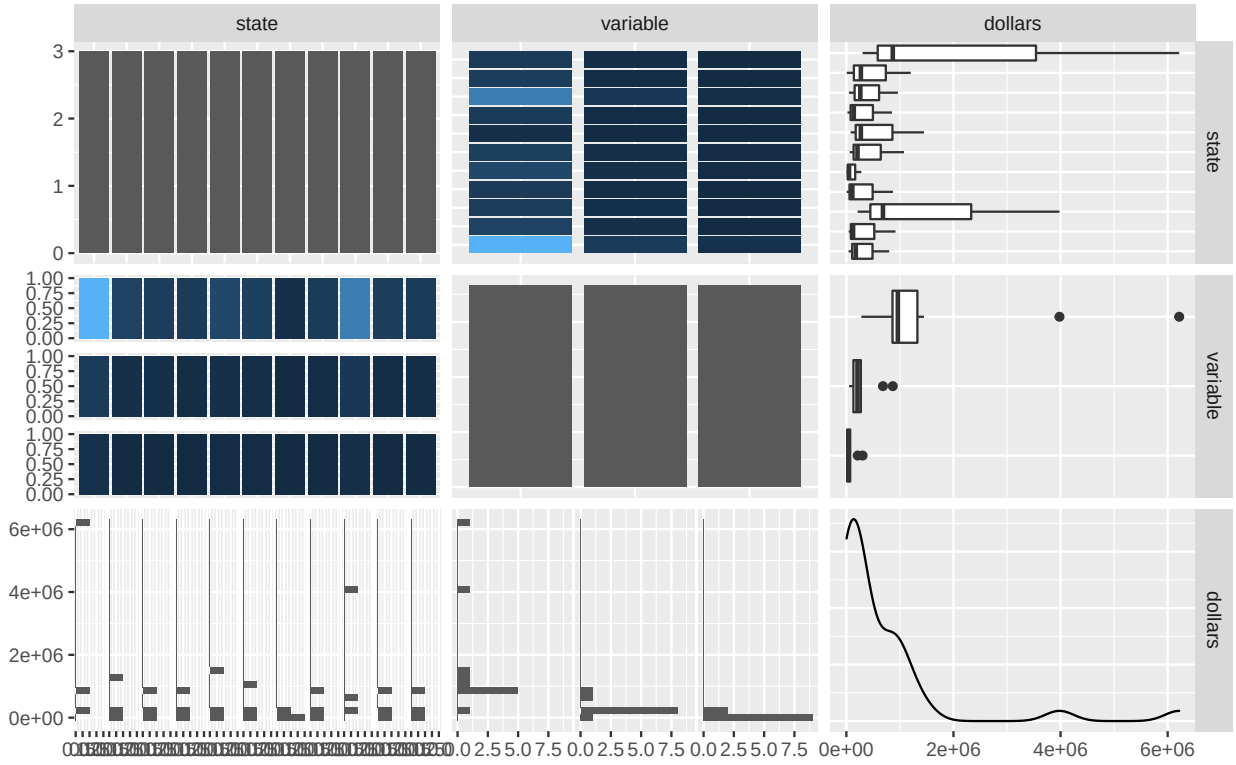
```
colnames(Dairy_DF_melt)[3] <- "dollars"

ggplot(Dairy_DF_melt, aes(x = state,y = `dollars`)) +
  geom_bar(aes(fill = variable),stat = "identity",position = "dodge")+
  ggtitle("Total farm dairy income") +
  theme_economist() +
  scale_color_economist() +
  scale_fill_brewer(palette = "PuBuGn")
```



More correlations

```
(ggpairs(Dairy_DF_melt, mapping=ggplot2::aes(color = `dollars`)))
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



ADD MORE DATASET

With gross cash income and net farm income, it's hard to tell the farm average net income by states. Therefore, another .CSV file containing farms number is needed to be read and exploited. After divided by farm number, the net income would be the net income per farm in each state, which gives better understanding on dairy farms' well-beings within different states.

```
Farm_number <- read.csv("numberofdairyfarm_2019.csv", na.string=c("", " "))
Farm_number <- subset(Farm_number, select=c(state, estimate))
head(Farm_number, n=10)
```

```
##             state estimate
## 1            Iowa      821
## 2      California     1239
## 3         Florida       NA
## 4  North Carolina      129
## 5         Georgia      300
## 6      Washington      958
## 7        Nebraska       NA
## 8       Minnesota     1934
## 9       Wisconsin     8078
## 10        Indiana      906
```

It also needs to delete the states values to match with the Dairy_DF dataset.

```
Farm_number <- subset(Farm_number, state!="Arkansas" & state!="Florida" &
                        state!="Kansas" & state!="Nebraska")
colnames(Farm_number)[2] <- "farm count"
head(Farm_number, n=10)
```

```
##             state farm count
## 1            Iowa        821
## 2      California       1239
## 4  North Carolina        129
## 5         Georgia        300
## 6      Washington        958
## 8       Minnesota       1934
## 9       Wisconsin       8078
## 10        Indiana        906
## 12       Missouri        812
## 13          Texas        640
```

MERGE DATASET

Merge the farm number data with Dairy_DF.

```
Dairy_DF <- merge(Dairy_DF, Farm_number,
                  by.Dairy_DF = "state", by.Farm_number = "state")
```

Create three more columns to show the average net income per farm.

```
Dairy_DF$`Gross cash income per farm` <- Dairy_DF$`Gross cash income`/Dairy_DF$`farm count`
Dairy_DF$`Net farm income per farm` <- Dairy_DF$`Net farm income`/Dairy_DF$`farm count`
Dairy_DF$`Other related income per farm` <- Dairy_DF$`Other related income`/Dairy_DF$`farm count`
head(Dairy_DF, n=10)
```

```
##             state year Gross cash income Net farm income Other related income
## 1      California 2019           6215313          867037                303906
## 2         Georgia 2019           1202771          271052                  9767
## 3        Illinois 2019            961495          261149                 47983
## 4         Indiana 2019            853612          138566                 20836
## 5            Iowa 2019           1450632          270124                 79578
## 6       Minnesota 2019           1077314          208412                 63106
## 7        Missouri 2019            279460           50150                  3508
## 8  North Carolina 2019            869851          111341                  6499
## 9           Texas 2019           3979949          683365                212331
## 10     Washington 2019            918842          125210                 42829
##    farm count Gross cash income per farm Net farm income per farm
```

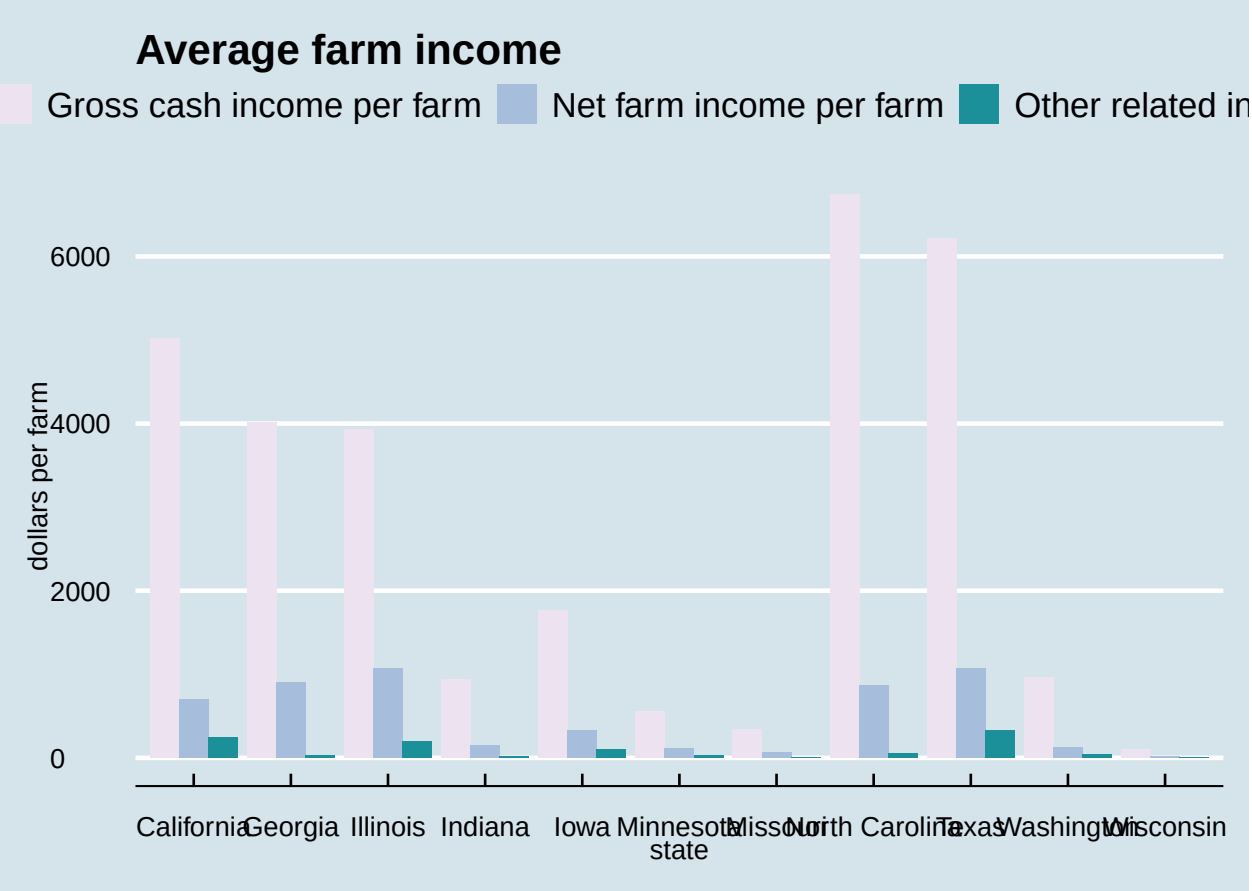
```
## 1        1239                  5016.3947                699.78773
## 2         300                  4009.2367                903.50667
## 3         245                  3924.4694               1065.91429
## 4         906                   942.1766                152.94260
## 5         821                  1766.9086                329.01827
## 6        1934                   557.0393                107.76215
## 7         812                   344.1626                 61.76108
## 8         129                  6743.0310                863.10853
## 9         640                  6218.6703               1067.75781
## 10        958                   959.1253                130.69937
##    Other related income per farm
## 1                      245.283293
## 2                       32.556667
## 3                      195.848980
## 4                       22.997792
## 5                       96.928136
## 6                       32.629783
## 7                        4.320197
## 8                       50.379845
## 9                      331.767187
## 10                      44.706681
```

VISUALIZATION

Create new grouped bar plot to get the information and comparison of the average

net come per farm in selected states.

```
Dairy_DF_melt2 <- melt(Dairy_DF[,c('state','Gross cash income per farm','Net farm income per farm',
                                   'Other related income per farm')],id.vars = 1)

colnames(Dairy_DF_melt2)[3] <- "dollars per farm"

ggplot(Dairy_DF_melt2, aes(x = state,y = `dollars per farm`)) +
  geom_bar(aes(fill = variable),stat = "identity",position = "dodge")+
  ggtitle("Average farm income") +
  theme_economist() +
  scale_color_economist() +
  scale_fill_brewer(palette = "PuBuGn")
```



Transformation and normalization

TO find out which state's farm's average gross cash income is above $1000 and below $100,

here to add two labels onto the dataset - "more than one thousand", "below one thousand"

```
Dairy_DF$`gross income status`<-
  cut(Dairy_DF$`Gross cash income per farm`, breaks = c(0, 100, 1000, Inf),
      labels = c("below one handard", "below one thousand", "more than one thousand"))
head(Dairy_DF, n=5)
```

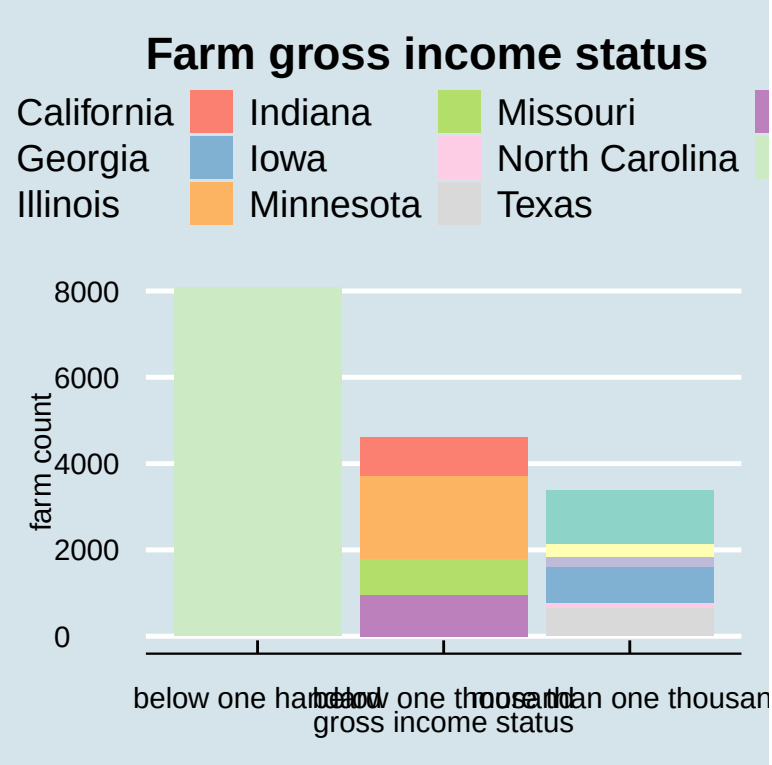
```
##        state year Gross cash income Net farm income Other related income
## 1 California 2019           6215313          867037                303906
## 2    Georgia 2019           1202771          271052                  9767
## 3   Illinois 2019            961495          261149                 47983
## 4    Indiana 2019            853612          138566                 20836
## 5       Iowa 2019           1450632          270124                 79578
##   farm count Gross cash income per farm Net farm income per farm
```

```
## 1       1239                  5016.3947                 699.7877
## 2        300                  4009.2367                 903.5067
## 3        245                  3924.4694                1065.9143
## 4        906                   942.1766                 152.9426
## 5        821                  1766.9086                 329.0183
##   Other related income per farm    gross income status
## 1                     245.28329 more than one thousand
## 2                      32.55667 more than one thousand
## 3                     195.84898 more than one thousand
## 4                      22.99779     below one thousand
## 5                      96.92814 more than one thousand
```

VISUALIZATION

Visualize farm's gross income status using stack bar plot

```
ggplot(Dairy_DF, aes(fill=state, y=`farm count`, x=`gross income status`)) +
  geom_bar(position="stack", stat="identity") +
  ggtitle("Farm gross income status") +
  theme_economist() +
  scale_color_economist() +
  scale_fill_brewer(palette = "Set3")
```



TRANSFORMATION

Create a new column named "expenses"

```
Dairy_DF$Expenses <-
  Dairy_DF$`Gross cash income` - Dairy_DF$`Net farm income`

Dairy_DF$`Expenses per farm` <-
  Dairy_DF$`Gross cash income per farm` - Dairy_DF$`Net farm income per farm`
head(Dairy_DF, n=5)
```

```
##        state year Gross cash income Net farm income Other related income
## 1 California 2019           6215313          867037                303906
## 2    Georgia 2019           1202771          271052                  9767
## 3   Illinois 2019            961495          261149                 47983
## 4    Indiana 2019            853612          138566                 20836
## 5       Iowa 2019           1450632          270124                 79578
##   farm count Gross cash income per farm Net farm income per farm
## 1       1239                  5016.3947                699.7877
## 2        300                  4009.2367                903.5067
## 3        245                  3924.4694               1065.9143
## 4        906                   942.1766                152.9426
## 5        821                  1766.9086                329.0183
##   Other related income per farm    gross income status Expenses
## 1                     245.28329 more than one thousand  5348276
## 2                      32.55667 more than one thousand   931719
## 3                     195.84898 more than one thousand   700346
## 4                      22.99779     below one thousand   715046
## 5                      96.92814 more than one thousand  1180508
##   Expenses per farm
## 1          4316.607
## 2          3105.730
## 3          2858.555
## 4           789.234
## 5          1437.890
```

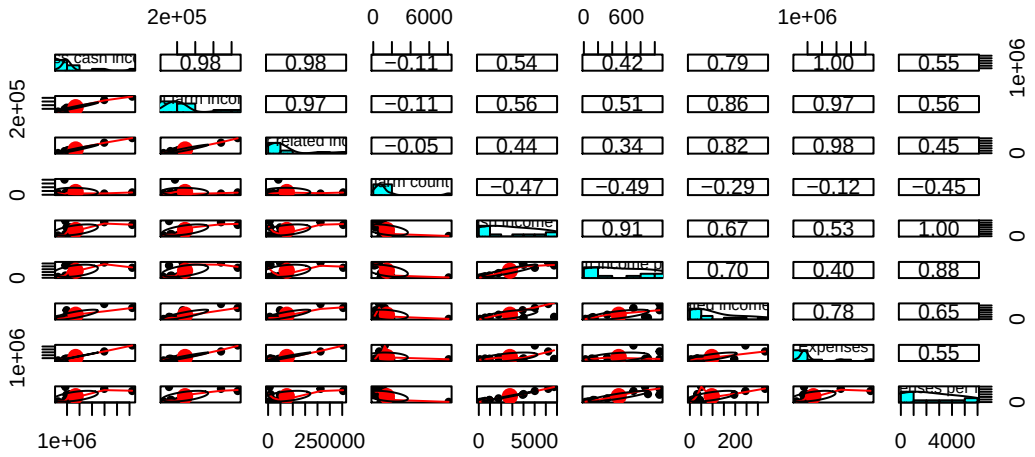
Use the "spearman" and "psych" library to calculate the correlations

```
(Temp<-Dairy_DF[,c(3, 4, 5, 6, 7, 8, 9, 11, 12)])
```

```
##    Gross cash income Net farm income Other related income farm count
## 1            6215313          867037                303906       1239
## 2            1202771          271052                  9767        300
## 3             961495          261149                 47983        245
## 4             853612          138566                 20836        906
```

```
## 5            1450632          270124                79578        821
## 6            1077314          208412                63106       1934
## 7             279460           50150                 3508        812
## 8             869851          111341                 6499        129
## 9            3979949          683365               212331        640
## 10            918842          125210                42829        958
## 11            800898          176314                41528       8078
##    Gross cash income per farm Net farm income per farm
## 1                   5016.39467                699.78773
## 2                   4009.23667                903.50667
## 3                   3924.46939               1065.91429
## 4                    942.17660                152.94260
## 5                   1766.90865                329.01827
## 6                    557.03930                107.76215
## 7                    344.16256                 61.76108
## 8                   6743.03101                863.10853
## 9                   6218.67031               1067.75781
## 10                   959.12526                130.69937
## 11                    99.14558                 21.82644
##    Other related income per farm Expenses Expenses per farm
## 1                     245.283293  5348276        4316.60694
## 2                      32.556667   931719        3105.73000
## 3                     195.848980   700346        2858.55510
## 4                      22.997792   715046         789.23400
## 5                      96.928136  1180508        1437.89038
## 6                      32.629783   868902         449.27715
## 7                       4.320197   229310         282.40148
## 8                      50.379845   758510        5879.92248
## 9                     331.767187  3296584        5150.91250
## 10                     44.706681   793632         828.42589
## 11                      5.140876   624584          77.31914
```

```
library(psych)
pairs.panels(Temp)
```



```r
corr.test(Temp, method = "spearman")
```

```
## Call:corr.test(x = Temp, method = "spearman")
## Correlation matrix
##                               Gross cash income Net farm income
## Gross cash income                          1.00            0.90
## Net farm income                            0.90            1.00
## Other related income                       0.81            0.77
## farm count                                -0.05            0.05
## Gross cash income per farm                 0.63            0.45
## Net farm income per farm                   0.63            0.57
## Other related income per farm              0.81            0.65
## Expenses                                   0.95            0.79
## Expenses per farm                          0.63            0.45
##                               Other related income farm count
## Gross cash income                                0.81      -0.05
## Net farm income                                  0.77       0.05
## Other related income                             1.00       0.35
## farm count                                       0.35       1.00
## Gross cash income per farm                       0.26      -0.65
## Net farm income per farm                         0.30      -0.73
## Other related income per farm                    0.76      -0.28
## Expenses                                         0.73       0.05
## Expenses per farm                                0.26      -0.65
##                               Gross cash income per farm
## Gross cash income                                   0.63
## Net farm income                                     0.45
## Other related income                                0.26
## farm count                                         -0.65
## Gross cash income per farm                          1.00
## Net farm income per farm                            0.88
## Other related income per farm                       0.77
## Expenses                                            0.62
## Expenses per farm                                   1.00
##                               Net farm income per farm
## Gross cash income                                 0.63
## Net farm income                                   0.57
## Other related income                              0.30
## farm count                                       -0.73
## Gross cash income per farm                        0.88
## Net farm income per farm                          1.00
## Other related income per farm                     0.75
## Expenses                                          0.51
## Expenses per farm                                 0.88
##                               Other related income per farm Expenses
## Gross cash income                                      0.81     0.95
## Net farm income                                        0.65     0.79
## Other related income                                   0.76     0.73
## farm count                                            -0.28     0.05
## Gross cash income per farm                             0.77     0.62
## Net farm income per farm                               0.75     0.51
## Other related income per farm                          1.00     0.71
## Expenses                                               0.71     1.00
```

```
## Expenses per farm                                    0.77     0.62
##                               Expenses per farm
## Gross cash income                          0.63
## Net farm income                            0.45
## Other related income                       0.26
## farm count                                -0.65
## Gross cash income per farm                 1.00
## Net farm income per farm                   0.88
## Other related income per farm              0.77
## Expenses                                   0.62
## Expenses per farm                          1.00
## Sample Size
## [1] 11
## Probability values (Entries above the diagonal are adjusted for multiple tests.)
##                               Gross cash income Net farm income
## Gross cash income                          0.00            0.01
## Net farm income                            0.00            0.00
## Other related income                       0.00            0.01
## farm count                                 0.89            0.87
## Gross cash income per farm                 0.04            0.16
## Net farm income per farm                   0.04            0.07
## Other related income per farm              0.00            0.03
## Expenses                                   0.00            0.00
## Expenses per farm                          0.04            0.16
##                               Other related income farm count
## Gross cash income                          0.08       1.00
## Net farm income                            0.15       1.00
## Other related income                       0.00       1.00
## farm count                                 0.30       0.00
## Gross cash income per farm                 0.43       0.03
## Net farm income per farm                   0.37       0.01
## Other related income per farm              0.01       0.40
## Expenses                                   0.01       0.89
## Expenses per farm                          0.43       0.03
##                               Gross cash income per farm
## Gross cash income                                   0.66
## Net farm income                                     1.00
## Other related income                                1.00
## farm count                                          0.61
## Gross cash income per farm                          0.00
## Net farm income per farm                            0.00
## Other related income per farm                       0.01
## Expenses                                            0.04
## Expenses per farm                                   0.00
##                               Net farm income per farm
## Gross cash income                                 0.66
## Net farm income                                   0.79
## Other related income                              1.00
## farm count                                        0.26
## Gross cash income per farm                        0.01
## Net farm income per farm                          0.00
## Other related income per farm                     0.01
## Expenses                                          0.11
## Expenses per farm                                 0.00
```

```
##                               Other related income per farm Expenses
## Gross cash income                                      0.08     0.00
## Net farm income                                        0.58     0.11
## Other related income                                   0.16     0.26
## farm count                                             1.00     1.00
## Gross cash income per farm                             0.15     0.66
## Net farm income per farm                               0.17     1.00
## Other related income per farm                          0.00     0.31
## Expenses                                               0.01     0.00
## Expenses per farm                                      0.01     0.04
##                               Expenses per farm
## Gross cash income                          0.66
## Net farm income                            1.00
## Other related income                       1.00
## farm count                                 0.61
## Gross cash income per farm                 0.00
## Net farm income per farm                   0.01
## Other related income per farm              0.15
## Expenses                                   0.66
## Expenses per farm                          0.00
##
##  To see confidence intervals of the correlations, print with the short=FALSE option
```

Doesn't show more information

Create a new column calculating net profit margin = net income / gross income

```
Dairy_DF$`profit margin%`<- 100*Dairy_DF$`Net farm income per farm`/
                            Dairy_DF$`Gross cash income per farm`
head(Dairy_DF, n=5)
```

```
##        state year Gross cash income Net farm income Other related income
## 1 California 2019           6215313          867037                303906
## 2    Georgia 2019           1202771          271052                  9767
## 3   Illinois 2019            961495          261149                 47983
## 4    Indiana 2019            853612          138566                 20836
## 5       Iowa 2019           1450632          270124                 79578
##   farm count Gross cash income per farm Net farm income per farm
## 1       1239                   5016.3947                 699.7877
## 2        300                   4009.2367                 903.5067
## 3        245                   3924.4694                1065.9143
## 4        906                    942.1766                 152.9426
## 5        821                   1766.9086                 329.0183
##   Other related income per farm    gross income status Expenses
## 1                     245.28329 more than one thousand  5348276
## 2                      32.55667 more than one thousand   931719
## 3                     195.84898 more than one thousand   700346
## 4                      22.99779     below one thousand   715046
## 5                      96.92814 more than one thousand  1180508
##   Expenses per farm profit margin%
## 1          4316.607       13.95001
```

```
## 2          3105.730       22.53563
## 3          2858.555       27.16072
## 4           789.234       16.23290
## 5          1437.890       18.62113
```

Round the columns before visualization

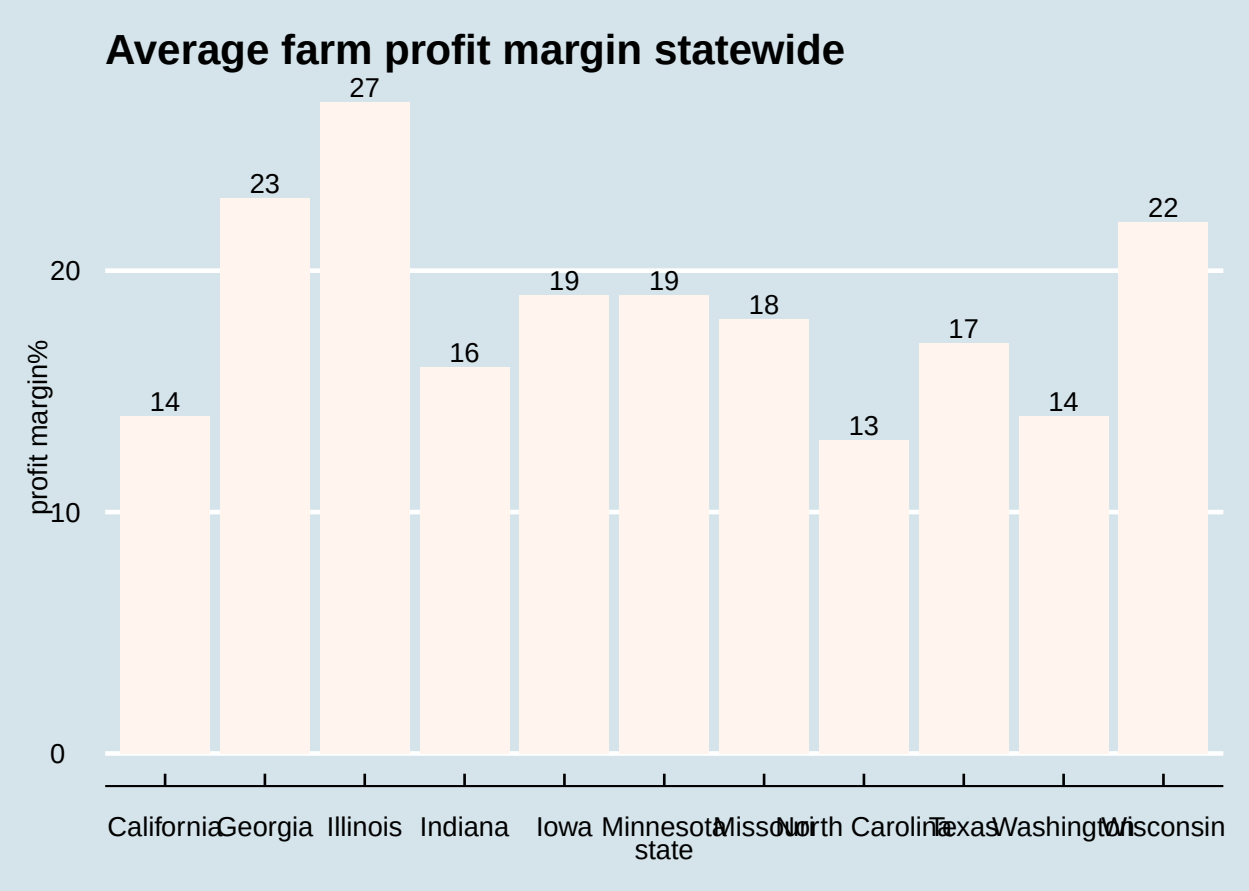
```
Dairy_DF <- Dairy_DF %>% mutate(across(where(is.numeric), round, 0))
head(Dairy_DF, n=5)
```

```
##        state year Gross cash income Net farm income Other related income
## 1 California 2019           6215313          867037                303906
## 2    Georgia 2019           1202771          271052                  9767
## 3   Illinois 2019            961495          261149                 47983
## 4    Indiana 2019            853612          138566                 20836
## 5       Iowa 2019           1450632          270124                 79578
##   farm count Gross cash income per farm Net farm income per farm
## 1       1239                       5016                      700
## 2        300                       4009                      904
## 3        245                       3924                     1066
## 4        906                        942                      153
## 5        821                       1767                      329
##   Other related income per farm    gross income status Expenses
## 1                           245 more than one thousand  5348276
## 2                            33 more than one thousand   931719
## 3                           196 more than one thousand   700346
## 4                            23     below one thousand   715046
## 5                            97 more than one thousand  1180508
##   Expenses per farm profit margin%
## 1             4317             14
## 2             3106             23
## 3             2859             27
## 4              789             16
## 5             1438             19
```

VISUALIZATION

Visualize the profit margin per farm to find out which state has the most profit margin in dairy production

```
ggplot(Dairy_DF, aes(x=state, y=`profit margin%`)) +
 geom_bar(fill="seashell", stat="identity") +
  geom_text(aes(label=`profit margin%`), position=position_dodge(width=0.9), vjust=-0.25) +
  theme_economist() +
  scale_color_economist() +
  ggtitle("Average farm profit margin statewide") +
  scale_fill_brewer(palette = "PuBuGn")
```



NORMALIZATION

```
Dairy_just_numeric <- Dairy_DF[,c(3:12)]
Dairy_just_numeric <- subset(Dairy_just_numeric, select=-c(`gross income status`))
head(Dairy_just_numeric, n=5)
```

```
##   Gross cash income Net farm income Other related income farm count
## 1           6215313          867037                303906       1239
## 2           1202771          271052                  9767        300
## 3            961495          261149                 47983        245
## 4            853612          138566                 20836        906
## 5           1450632          270124                 79578        821
##   Gross cash income per farm Net farm income per farm
## 1                       5016                      700
## 2                       4009                      904
## 3                       3924                     1066
## 4                        942                      153
## 5                       1767                      329
##   Other related income per farm Expenses Expenses per farm
## 1                           245  5348276              4317
```

```
## 2                            33   931719             3106
## 3                           196   700346             2859
## 4                            23   715046              789
## 5                            97  1180508             1438
```

Create function - for min-max

```
My_Min_Max_Function <- function(x) {
  MyMax=max(x)
  MyMin = min(x)
  Diff = MyMax - MyMin
  normVal = x/(Diff)
  return(normVal)
}

Dairy_just_numeric<-
    My_Min_Max_Function(Dairy_just_numeric)
head(Dairy_just_numeric, n=5)
```

```
##   Gross cash income Net farm income Other related income   farm count
## 1         1.0000006      0.13950022          0.048896362 1.993465e-04
## 2         0.1935175      0.04361038          0.001571442 4.826791e-05
## 3         0.1546979      0.04201706          0.007720131 3.941880e-05
## 4         0.1373402      0.02229431          0.003352368 1.457691e-04
## 5         0.2333966      0.04346107          0.012803547 1.320932e-04
##   Gross cash income per farm Net farm income per farm
## 1               0.0008070395             1.126251e-04
## 2               0.0006450202             1.454473e-04
## 3               0.0006313443             1.715120e-04
## 4               0.0001515612             2.461664e-05
## 5               0.0002842980             5.293381e-05
##   Other related income per farm  Expenses Expenses per farm
## 1                  3.941880e-05 0.8605004      0.0006945753
## 2                  5.309471e-06 0.1499071      0.0004997338
## 3                  3.153504e-05 0.1126808      0.0004599932
## 4                  3.700540e-06 0.1150459      0.0001269446
## 5                  1.560663e-05 0.1899355      0.0002313642
```

WRTIE THE THREE DATAFRAMES INTO CSV FILES

```
write.csv(Dairy_just_numeric, "Dairy_just_numeric.csv")
write.csv(Dairy_DF, "Dairy_DF.csv")
write.csv(Dairy_DF_melt, "Dairy_DF_melt.csv")
```