Final Project EDA

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.0 --
## v ggplot2 3.3.3     v purrr   0.3.4
## v tibble  3.0.6     v dplyr   1.0.4
## v tidyr   1.1.2     v stringr 1.4.0
## v readr   1.4.0     v forcats 0.5.1
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(tidyverse)
library(dplyr)
library(ggplot2)
library(mice)
```

```
##
## Attaching package: 'mice'
## The following object is masked from 'package:stats':
##
##     filter
## The following objects are masked from 'package:base':
##
##     cbind, rbind
```

```
member = read_csv('data/member-data-2020-stat149.csv')
```

```
##
## -- Column specification --------------------------------------------------------
## cols(
##   .default = col_character(),
##   JSMtot = col_double(),
##   Age = col_double(),
##   AgeJoinedASA = col_double()
## )
## i Use `spec()` for the full column specifications.
```

```
head(member)
```

```
## # A tibble: 6 x 37
##   AnySection JSMtot USA.CAN DontPublish MEMTYPE   Age AgeJoinedASA Gender
##   <chr>       <dbl> <chr>   <chr>       <chr>   <dbl>        <dbl> <chr>
## 1 Yes             0 Yes     No          ILIFF      76           24 M
## 2 No              0 Yes     No          ILIFF      83           53 M
## 3 No              0 No      No          ILIFA      69           24 M
## 4 Yes             0 Yes     No          ISEN       71           24 M
## 5 Yes             5 No      No          ILIFF      74           27 M
## 6 No              3 No      No          IREG       74           29 M
## # ... with 29 more variables: EmploymentCategory <chr>, InChapter <chr>,
## #   P.SEC.BE <chr>, P.SEC.BIOM <chr>, P.SEC.BIOP <chr>, P.SEC.CNSL <chr>,
## #   P.SEC.COMP <chr>, P.SEC.EDUC <chr>, P.SEC.ENVR <chr>, P.SEC.EPI <chr>,
## #   P.SEC.GOVT <chr>, P.SEC.GRPH <chr>, P.SEC.HPSS <chr>, P.SEC.MDD <chr>,
## #   P.SEC.MHS <chr>, P.SEC.MKTG <chr>, P.SEC.NPAR <chr>, P.SEC.QP <chr>,
## #   P.SEC.SBSS <chr>, P.SEC.SDNS <chr>, P.SEC.SGG <chr>, P.SEC.SI <chr>,
## #   P.SEC.SIS <chr>, P.SEC.SLDM <chr>, P.SEC.SOC <chr>, P.SEC.SPES <chr>,
## #   P.SEC.SRMS <chr>, P.SEC.SSPA <chr>, P.SEC.TSHS <chr>
```

```
dim(member)
```

```
## [1] 17594    37
```

```
summary(member)
```

```
##   AnySection            JSMtot         USA.CAN          DontPublish
##  Length:17594       Min.   :0.0000   Length:17594       Length:17594
##  Class :character   1st Qu.:0.0000   Class :character   Class :character
##  Mode  :character   Median :0.0000   Mode  :character   Mode  :character
##                     Mean   :0.9689
##                     3rd Qu.:1.0000
##                     Max.   :5.0000
##
##    MEMTYPE               Age          AgeJoinedASA       Gender
##  Length:17594       Min.   : 11.00   Min.   :  6.00   Length:17594
##  Class :character   1st Qu.: 34.00   1st Qu.: 26.00   Class :character
##  Mode  :character   Median : 47.00   Median : 30.00   Mode  :character
##                     Mean   : 48.47   Mean   : 32.67
##                     3rd Qu.: 61.00   3rd Qu.: 37.00
##                     Max.   :105.00   Max.   :115.00
##                     NA's   :3399     NA's   :3400
##  EmploymentCategory  InChapter           P.SEC.BE          P.SEC.BIOM
##  Length:17594       Length:17594       Length:17594       Length:17594
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##
##   P.SEC.BIOP         P.SEC.CNSL         P.SEC.COMP         P.SEC.EDUC
##  Length:17594       Length:17594       Length:17594       Length:17594
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##
##   P.SEC.ENVR         P.SEC.EPI          P.SEC.GOVT         P.SEC.GRPH
##  Length:17594       Length:17594       Length:17594       Length:17594
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
```

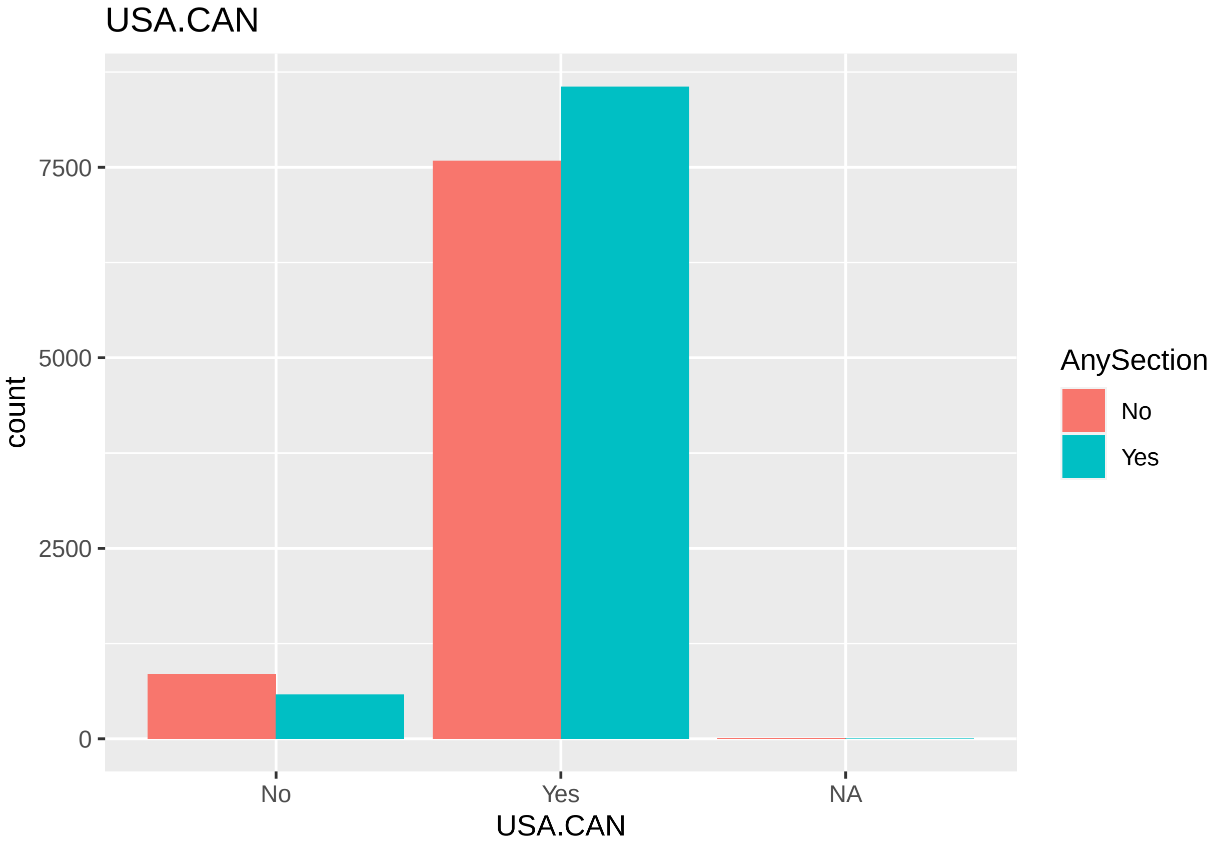
```
##
##   P.SEC.HPSS         P.SEC.MDD          P.SEC.MHS          P.SEC.MKTG
##  Length:17594       Length:17594       Length:17594       Length:17594
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##
##   P.SEC.NPAR          P.SEC.QP          P.SEC.SBSS         P.SEC.SDNS
##  Length:17594       Length:17594       Length:17594       Length:17594
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##
##   P.SEC.SGG           P.SEC.SI          P.SEC.SIS          P.SEC.SLDM
##  Length:17594       Length:17594       Length:17594       Length:17594
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##
##   P.SEC.SOC          P.SEC.SPES         P.SEC.SRMS         P.SEC.SSPA
##  Length:17594       Length:17594       Length:17594       Length:17594
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##
##   P.SEC.TSHS
##  Length:17594
##  Class :character
##  Mode  :character
##
##
##
##
```

```
member %>%
  count(AnySection)
```

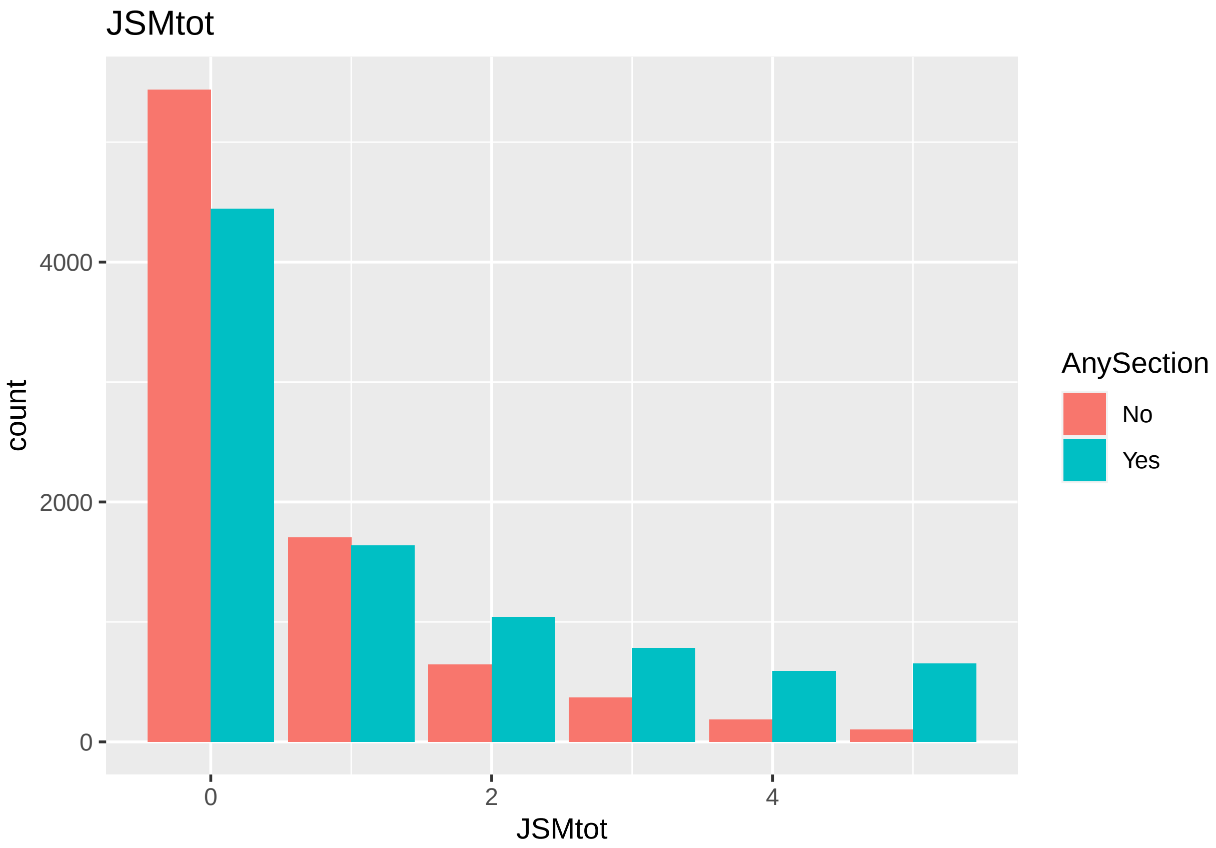
```
## # A tibble: 2 x 2
##   AnySection     n
## * <chr>      <int>
## 1 No          8447
## 2 Yes         9147
```

Visualization

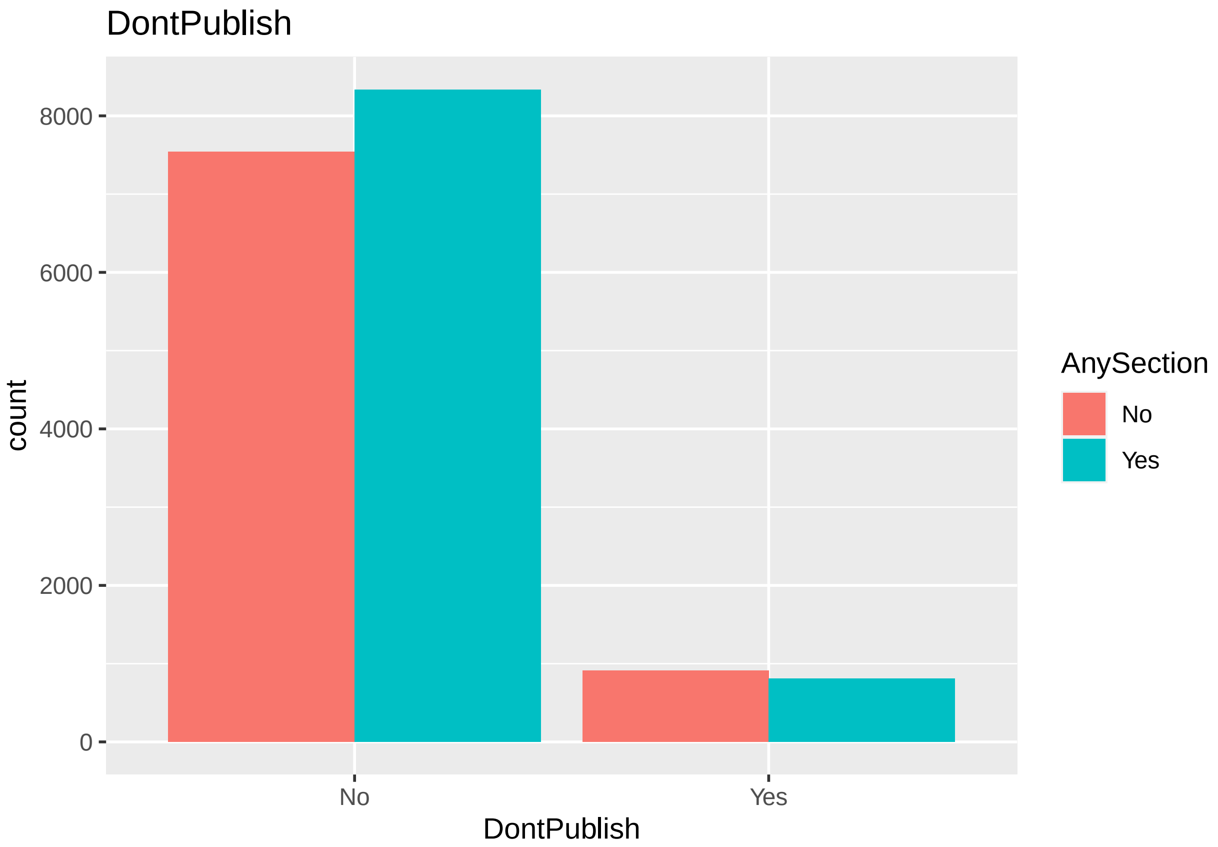
```r
ggplot(member, aes(USA.CAN, fill = AnySection)) +
   geom_bar(position = 'dodge')+ggtitle("USA.CAN")
```



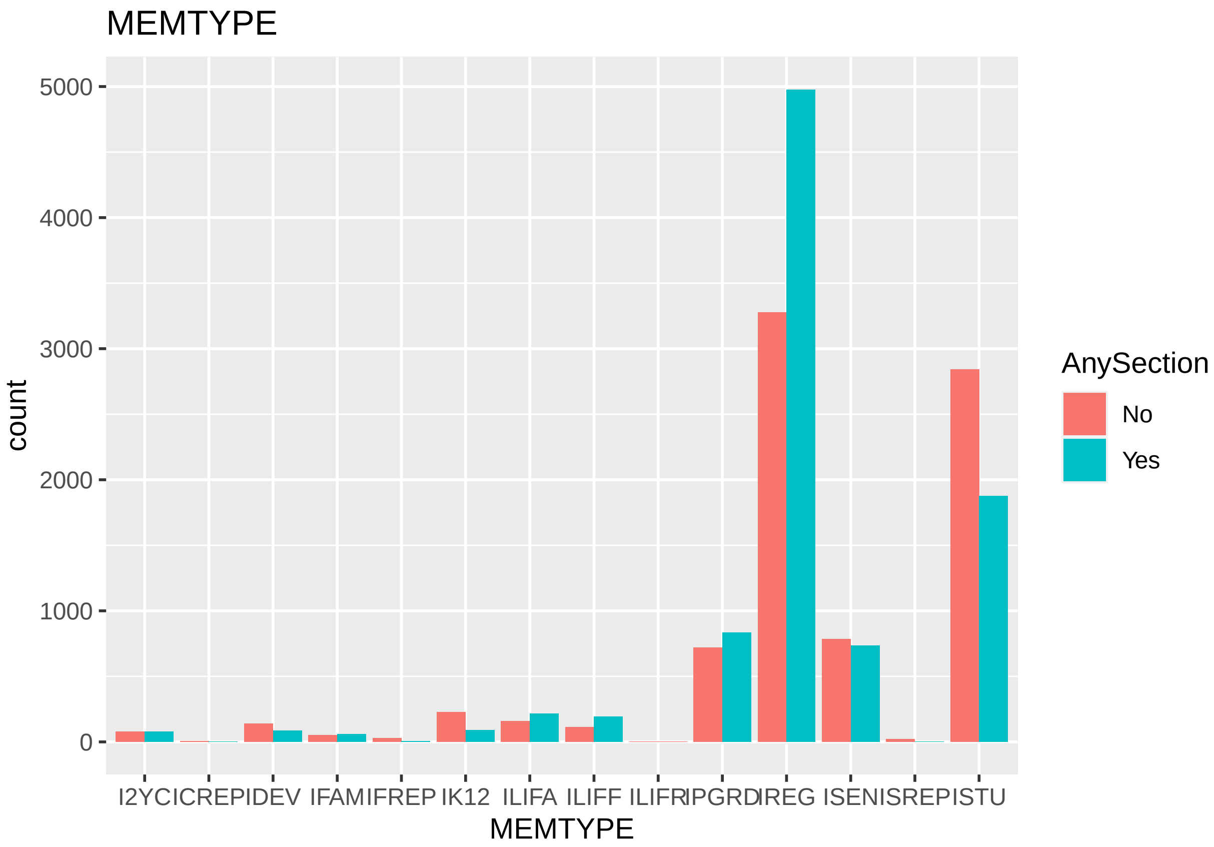
```r
ggplot(member, aes(JSMtot, fill = AnySection)) +
   geom_bar(position = 'dodge')+ggtitle("JSMtot")
```



```
ggplot(member, aes(DontPublish, fill = AnySection)) +
  geom_bar(position = 'dodge')+ggtitle("DontPublish")
```

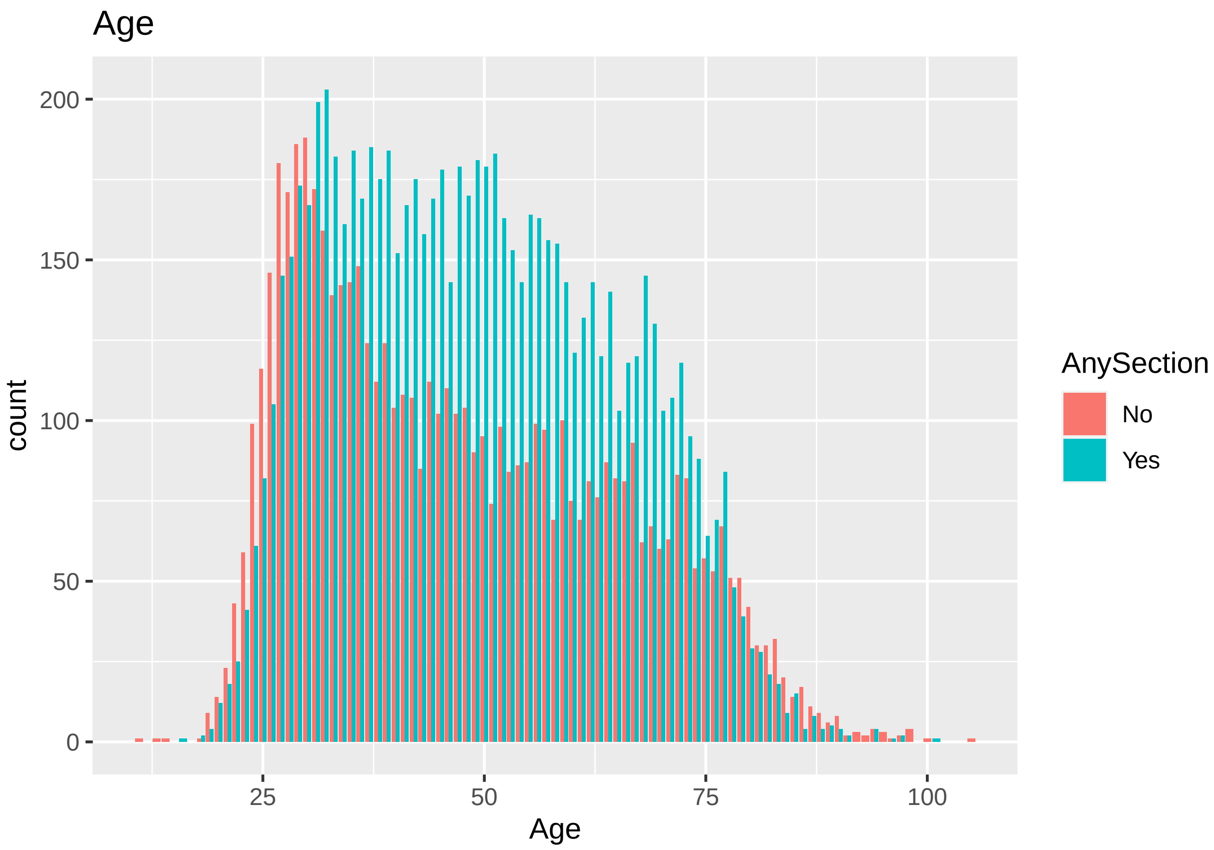


```
ggplot(member, aes(MEMTYPE, fill = AnySection)) +
   geom_bar(position = 'dodge')+ggtitle("MEMTYPE")
```



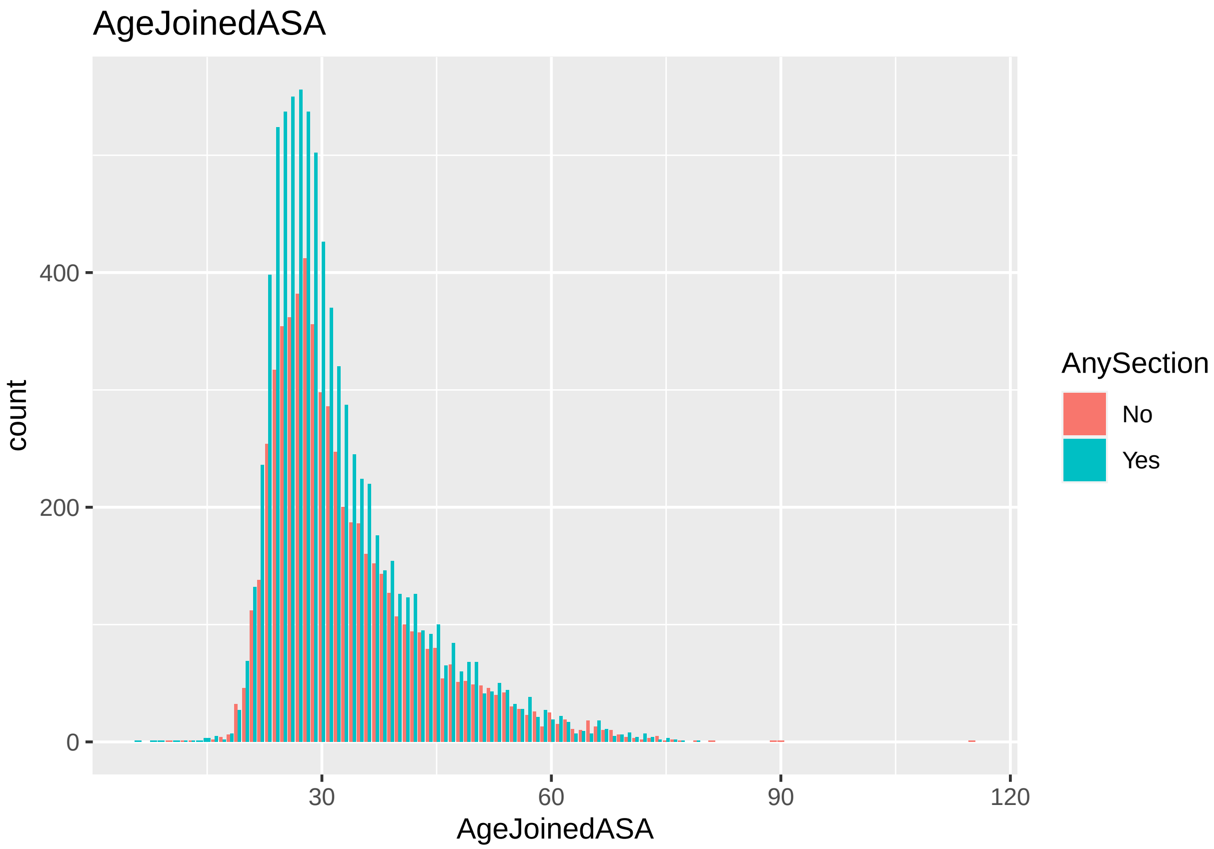
```
ggplot(member, aes(Age, fill = AnySection)) +
   geom_bar(position = 'dodge')+ggtitle("Age")
```

```
## Warning: Removed 3399 rows containing non-finite values (stat_count).
```

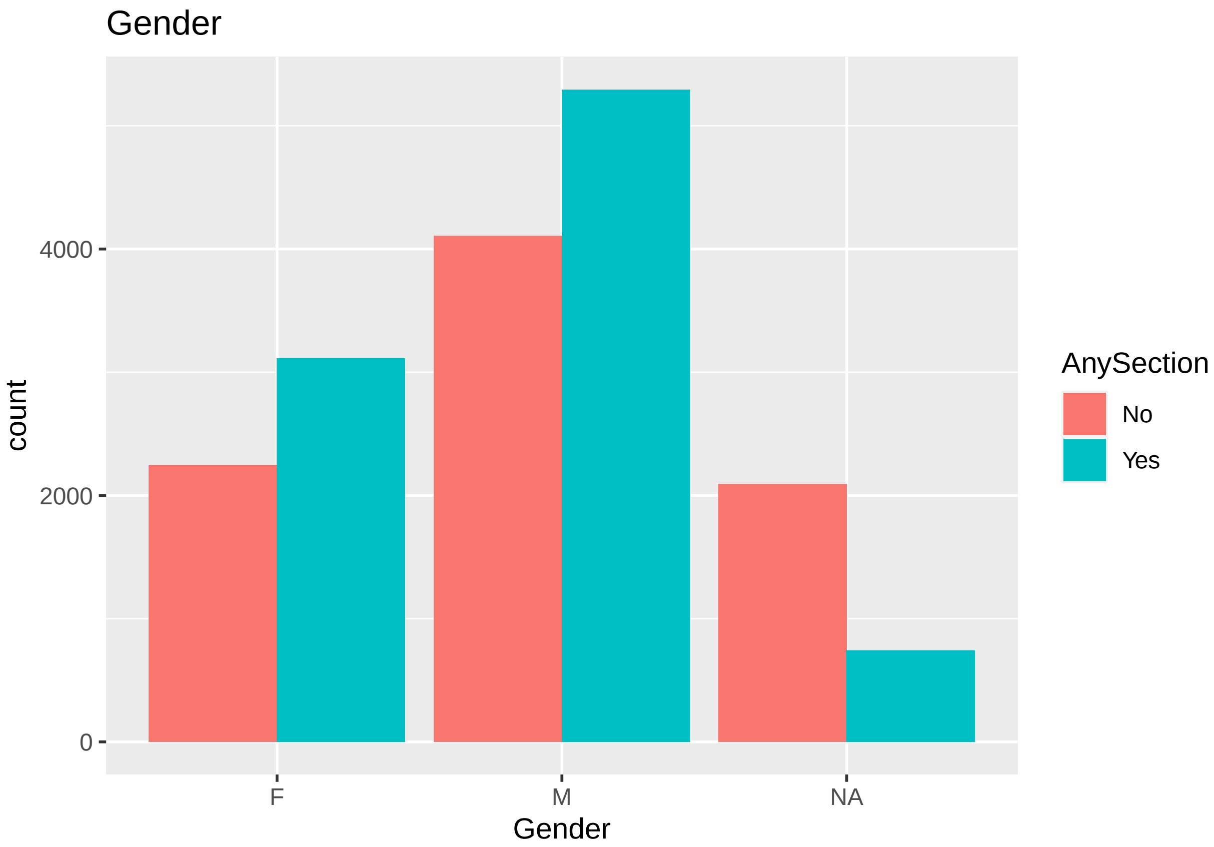


```
ggplot(member, aes(AgeJoinedASA, fill = AnySection)) +
   geom_bar(position = 'dodge')+ggtitle("AgeJoinedASA")
```

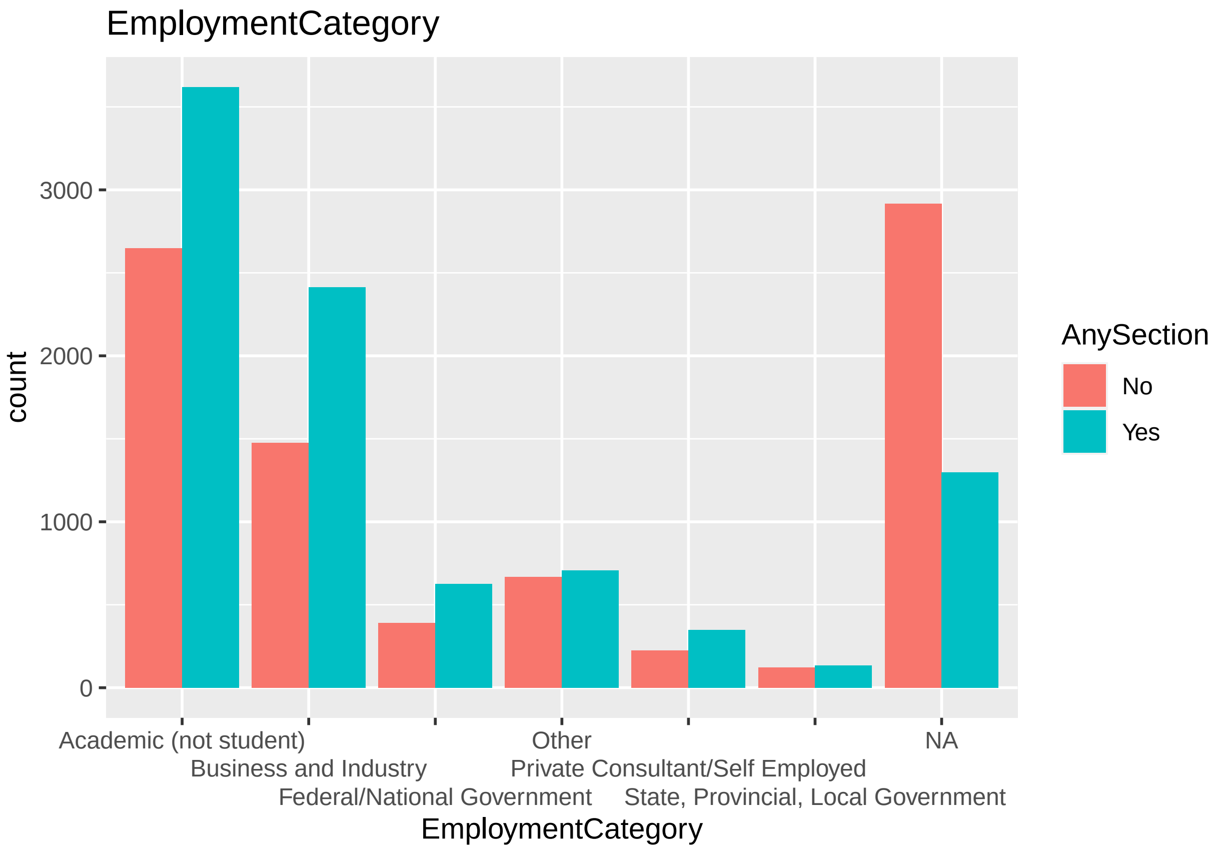
```
## Warning: Removed 3400 rows containing non-finite values (stat_count).
```



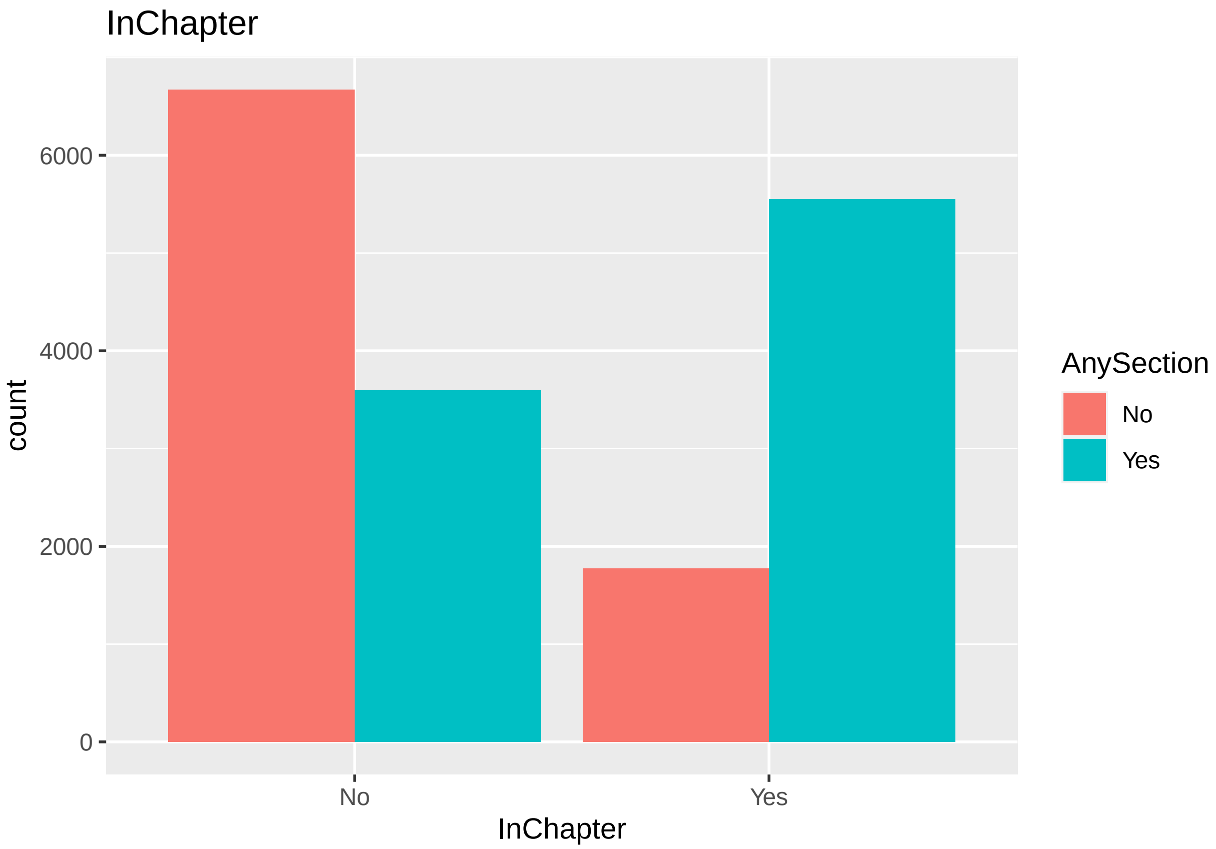
```
ggplot(member, aes(Gender, fill = AnySection)) +
   geom_bar(position = 'dodge')+ggtitle("Gender")
```



```
ggplot(member, aes(EmploymentCategory, fill = AnySection)) +
   scale_x_discrete(guide = guide_axis(n.dodge=3))+
   geom_bar(position = 'dodge')+ggtitle("EmploymentCategory")
```



```
ggplot(member, aes(InChapter, fill = AnySection)) +
   geom_bar(position = 'dodge')+ggtitle("InChapter")
```



Data processing

Missing value

These columns do not have missing value. We turn yes/no to 1/0.

```
member$AnySection = as.numeric(as.factor(member$AnySection))-1
member$DontPublish = as.numeric(as.factor(member$DontPublish))-1
member$MEMTYPE = as.numeric(as.factor(member$MEMTYPE))-1
member$InChapter = as.numeric(as.factor(member$InChapter))-1
member$P.SEC.BE = as.numeric(as.factor(member$P.SEC.BE))-1
member$P.SEC.BIOM = as.numeric(as.factor(member$P.SEC.BIOM))-1
member$P.SEC.BIOP = as.numeric(as.factor(member$P.SEC.BIOP))-1
member$P.SEC.CNSL = as.numeric(as.factor(member$P.SEC.CNSL))-1

member$P.SEC.COMP = as.numeric(as.factor(member$P.SEC.BIOM))-1
member$P.SEC.EDUC = as.numeric(as.factor(member$P.SEC.BIOP))-1
member$P.SEC.ENVR = as.numeric(as.factor(member$P.SEC.CNSL))-1
member$P.SEC.EPI = as.numeric(as.factor(member$P.SEC.BIOM))-1
member$P.SEC.GOVT = as.numeric(as.factor(member$P.SEC.BIOP))-1
member$P.SEC.GRPH = as.numeric(as.factor(member$P.SEC.CNSL))-1

member$P.SEC.HPSS = as.numeric(as.factor(member$P.SEC.BIOM))-1
member$P.SEC.MDD = as.numeric(as.factor(member$P.SEC.BIOP))-1
member$P.SEC.MHS = as.numeric(as.factor(member$P.SEC.CNSL))-1
member$P.SEC.MKTG = as.numeric(as.factor(member$P.SEC.BIOM))-1
member$P.SEC.NPAR = as.numeric(as.factor(member$P.SEC.BIOP))-1
member$P.SEC.QP = as.numeric(as.factor(member$P.SEC.CNSL))-1
```

```
member$P.SEC.SBSS = as.numeric(as.factor(member$P.SEC.BIOM))-1
member$P.SEC.SDNS= as.numeric(as.factor(member$P.SEC.BIOP))-1
member$P.SEC.SGG = as.numeric(as.factor(member$P.SEC.CNSL))-1
member$P.SEC.SI = as.numeric(as.factor(member$P.SEC.BIOM))-1
member$P.SEC.SIS = as.numeric(as.factor(member$P.SEC.BIOP))-1
member$P.SEC.SLDM = as.numeric(as.factor(member$P.SEC.CNSL))-1

member$P.SEC.SOC = as.numeric(as.factor(member$P.SEC.BIOM))-1
member$P.SEC.SPES= as.numeric(as.factor(member$P.SEC.BIOP))-1
member$P.SEC.SRMS = as.numeric(as.factor(member$P.SEC.CNSL))-1
member$P.SEC.SSPA = as.numeric(as.factor(member$P.SEC.BIOM))-1
member$P.SEC.TSHS = as.numeric(as.factor(member$P.SEC.BIOP))-1
```

category1: Business and Economics government BE, CNSL, category2:

```
# Business and Economics Statistics
```

```
member
```

```
## # A tibble: 17,594 x 37
##    AnySection JSMtot USA.CAN DontPublish MEMTYPE   Age AgeJoinedASA Gender
##         <dbl>  <dbl> <chr>         <dbl>   <dbl> <dbl>        <dbl> <chr>
##  1          1      0 Yes               0       7    76           24 M
##  2          0      0 Yes               0       7    83           53 M
##  3          0      0 No                0       6    69           24 M
##  4          1      0 Yes               0      11    71           24 M
##  5          1      5 No                0       7    74           27 M
##  6          0      3 No                0      10    74           29 M
##  7          0      1 Yes               0       7    77           33 M
##  8          1      1 Yes               0      10    70           25 M
##  9          0      0 Yes               0      11    73           27 M
## 10          0      0 No                0      11    78           26 M
## # ... with 17,584 more rows, and 29 more variables: EmploymentCategory <chr>,
## #   InChapter <dbl>, P.SEC.BE <dbl>, P.SEC.BIOM <dbl>, P.SEC.BIOP <dbl>,
## #   P.SEC.CNSL <dbl>, P.SEC.COMP <dbl>, P.SEC.EDUC <dbl>, P.SEC.ENVR <dbl>,
## #   P.SEC.EPI <dbl>, P.SEC.GOVT <dbl>, P.SEC.GRPH <dbl>, P.SEC.HPSS <dbl>,
## #   P.SEC.MDD <dbl>, P.SEC.MHS <dbl>, P.SEC.MKTG <dbl>, P.SEC.NPAR <dbl>,
## #   P.SEC.QP <dbl>, P.SEC.SBSS <dbl>, P.SEC.SDNS <dbl>, P.SEC.SGG <dbl>,
## #   P.SEC.SI <dbl>, P.SEC.SIS <dbl>, P.SEC.SLDM <dbl>, P.SEC.SOC <dbl>,
## #   P.SEC.SPES <dbl>, P.SEC.SRMS <dbl>, P.SEC.SSPA <dbl>, P.SEC.TSHS <dbl>
```

USA.CAN column has 12 missing values, which are less than 1% of the total data. We will impute the missing value with the mean of this column.

```
sum(is.na(member$USA.CAN))
```

```
## [1] 12
```

```
member$USA.CAN[is.na(member$USA.CAN)] = 'Yes'
# any better way to convert yes/no to 1/0 ??
member$USA.CAN = as.numeric(as.factor(member$USA.CAN))-1
member %>%
  count(USA.CAN)
```

```
## # A tibble: 2 x 2
##   USA.CAN     n
## *   <dbl> <int>
## 1       0  1434
## 2       1 16160
```

Gender column has 2834 missing values. EmploymentCategory column has 4216 missing values.

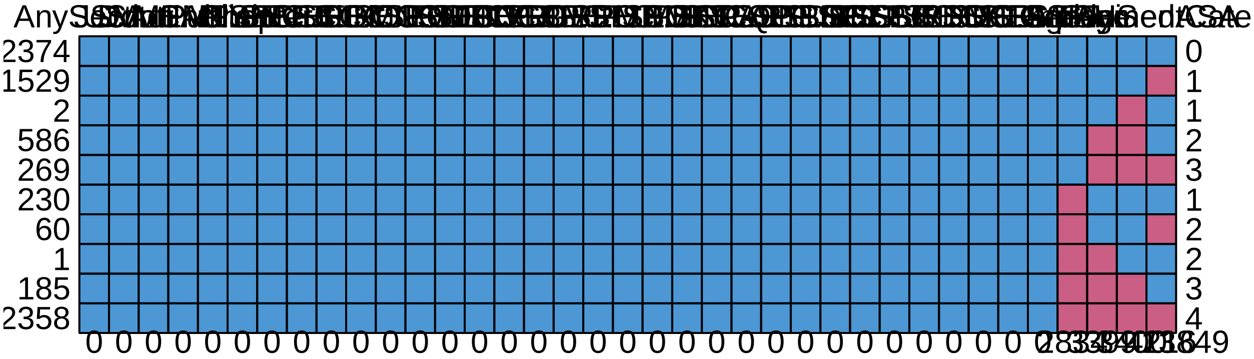
```
sum(is.na(member$Gender))
```

```
## [1] 2834
```

```
sum(is.na(member$EmploymentCategory))
```

```
## [1] 4216
```

```
md.pattern(member)
```



```
##       AnySection JSMtot USA.CAN DontPublish MEMTYPE InChapter P.SEC.BE
## 12374          1      1       1           1       1         1        1
## 1529           1      1       1           1       1         1        1
## 2              1      1       1           1       1         1        1
## 586            1      1       1           1       1         1        1
## 269            1      1       1           1       1         1        1
## 230            1      1       1           1       1         1        1
## 60             1      1       1           1       1         1        1
## 1              1      1       1           1       1         1        1
## 185            1      1       1           1       1         1        1
## 2358           1      1       1           1       1         1        1
##                0      0       0           0       0         0        0
##       P.SEC.BIOM P.SEC.BIOP P.SEC.CNSL P.SEC.COMP P.SEC.EDUC P.SEC.ENVR
## 12374          1          1          1          1          1          1
## 1529           1          1          1          1          1          1
## 2              1          1          1          1          1          1
## 586            1          1          1          1          1          1
## 269            1          1          1          1          1          1
## 230            1          1          1          1          1          1
## 60             1          1          1          1          1          1
## 1              1          1          1          1          1          1
## 185            1          1          1          1          1          1
## 2358           1          1          1          1          1          1
##                0          0          0          0          0          0
##       P.SEC.EPI P.SEC.GOVT P.SEC.GRPH P.SEC.HPSS P.SEC.MDD P.SEC.MHS P.SEC.MKTG
## 12374         1          1          1          1         1         1          1
## 1529          1          1          1          1         1         1          1
## 2             1          1          1          1         1         1          1
## 586           1          1          1          1         1         1          1
## 269           1          1          1          1         1         1          1
## 230           1          1          1          1         1         1          1
## 60            1          1          1          1         1         1          1
## 1             1          1          1          1         1         1          1
## 185           1          1          1          1         1         1          1
## 2358          1          1          1          1         1         1          1
##               0          0          0          0         0         0          0
##       P.SEC.NPAR P.SEC.QP P.SEC.SBSS P.SEC.SDNS P.SEC.SGG P.SEC.SI P.SEC.SIS
## 12374          1        1          1          1         1        1         1
## 1529           1        1          1          1         1        1         1
## 2              1        1          1          1         1        1         1
## 586            1        1          1          1         1        1         1
## 269            1        1          1          1         1        1         1
## 230            1        1          1          1         1        1         1
## 60             1        1          1          1         1        1         1
## 1              1        1          1          1         1        1         1
## 185            1        1          1          1         1        1         1
## 2358           1        1          1          1         1        1         1
##                0        0          0          0         0        0         0
##       P.SEC.SLDM P.SEC.SOC P.SEC.SPES P.SEC.SRMS P.SEC.SSPA P.SEC.TSHS Gender
## 12374          1         1          1          1          1          1      1
## 1529           1         1          1          1          1          1      1
## 2              1         1          1          1          1          1      1
## 586            1         1          1          1          1          1      1
## 269            1         1          1          1          1          1      1
## 230            1         1          1          1          1          1      0
## 60             1         1          1          1          1          1      0
## 1              1         1          1          1          1          1      0
## 185            1         1          1          1          1          1      0
## 2358           1         1          1          1          1          1      0
##                0         0          0          0          0          0   2834
##        Age AgeJoinedASA EmploymentCategory
## 12374    1            1                  1     0
## 1529     1            1                  0     1
## 2        1            0                  1     1
## 586      0            0                  1     2
## 269      0            0                  0     3
## 230      1            1                  1     1
## 60       1            1                  0     2
## 1        0            1                  1     2
## 185      0            0                  1     3
## 2358     0            0                  0     4
##       3399         3400               4216 13849
```

```
imp <- mice(member, method = "pmm", m = 5, maxit = 50, printFlag = FALSE)
```

```
## Warning: Number of logged events: 25
```

```
member.imp <- complete(imp)
summary(member.imp)
```

```
##    AnySection         JSMtot          USA.CAN        DontPublish
##  Min.   :0.0000   Min.   :0.0000   Min.   :0.0000   Min.   :0.00000
##  1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:1.0000   1st Qu.:0.00000
##  Median :1.0000   Median :0.0000   Median :1.0000   Median :0.00000
##  Mean   :0.5199   Mean   :0.9689   Mean   :0.9185   Mean   :0.09776
##  3rd Qu.:1.0000   3rd Qu.:1.0000   3rd Qu.:1.0000   3rd Qu.:0.00000
##  Max.   :1.0000   Max.   :5.0000   Max.   :1.0000   Max.   :1.00000
##     MEMTYPE           Age          AgeJoinedASA       Gender
##  Min.   : 0.00   Min.   : 11.00   Min.   :  6.00   Length:17594
##  1st Qu.:10.00   1st Qu.: 33.00   1st Qu.: 26.00   Class :character
##  Median :10.00   Median : 46.00   Median : 30.00   Mode  :character
##  Mean   :10.33   Mean   : 47.85   Mean   : 32.66
##  3rd Qu.:13.00   3rd Qu.: 61.00   3rd Qu.: 37.00
##  Max.   :13.00   Max.   :105.00   Max.   :115.00
##  EmploymentCategory   InChapter         P.SEC.BE          P.SEC.BIOM
##  Length:17594       Min.   :0.0000   Min.   :0.000000   Min.   :0.00
##  Class :character   1st Qu.:0.0000   1st Qu.:0.000000   1st Qu.:0.00
##  Mode  :character   Median :0.0000   Median :0.000000   Median :0.00
##                     Mean   :0.4164   Mean   :0.005911   Mean   :0.01
##                     3rd Qu.:1.0000   3rd Qu.:0.000000   3rd Qu.:0.00
##                     Max.   :1.0000   Max.   :1.000000   Max.   :1.00
##    P.SEC.BIOP          P.SEC.CNSL          P.SEC.COMP     P.SEC.EDUC
##  Min.   :0.000000   Min.   :0.000000   Min.   :0.00   Min.   :0.000000
##  1st Qu.:0.000000   1st Qu.:0.000000   1st Qu.:0.00   1st Qu.:0.000000
##  Median :0.000000   Median :0.000000   Median :0.00   Median :0.000000
##  Mean   :0.007787   Mean   :0.009208   Mean   :0.01   Mean   :0.007787
##  3rd Qu.:0.000000   3rd Qu.:0.000000   3rd Qu.:0.00   3rd Qu.:0.000000
##  Max.   :1.000000   Max.   :1.000000   Max.   :1.00   Max.   :1.000000
##    P.SEC.ENVR          P.SEC.EPI      P.SEC.GOVT          P.SEC.GRPH
##  Min.   :0.000000   Min.   :0.00   Min.   :0.000000   Min.   :0.000000
##  1st Qu.:0.000000   1st Qu.:0.00   1st Qu.:0.000000   1st Qu.:0.000000
##  Median :0.000000   Median :0.00   Median :0.000000   Median :0.000000
##  Mean   :0.009208   Mean   :0.01   Mean   :0.007787   Mean   :0.009208
##  3rd Qu.:0.000000   3rd Qu.:0.00   3rd Qu.:0.000000   3rd Qu.:0.000000
##  Max.   :1.000000   Max.   :1.00   Max.   :1.000000   Max.   :1.000000
##    P.SEC.HPSS     P.SEC.MDD          P.SEC.MHS          P.SEC.MKTG
##  Min.   :0.00   Min.   :0.000000   Min.   :0.000000   Min.   :0.00
##  1st Qu.:0.00   1st Qu.:0.000000   1st Qu.:0.000000   1st Qu.:0.00
##  Median :0.00   Median :0.000000   Median :0.000000   Median :0.00
##  Mean   :0.01   Mean   :0.007787   Mean   :0.009208   Mean   :0.01
##  3rd Qu.:0.00   3rd Qu.:0.000000   3rd Qu.:0.000000   3rd Qu.:0.00
##  Max.   :1.00   Max.   :1.000000   Max.   :1.000000   Max.   :1.00
##    P.SEC.NPAR           P.SEC.QP          P.SEC.SBSS     P.SEC.SDNS
##  Min.   :0.000000   Min.   :0.000000   Min.   :0.00   Min.   :0.000000
##  1st Qu.:0.000000   1st Qu.:0.000000   1st Qu.:0.00   1st Qu.:0.000000
##  Median :0.000000   Median :0.000000   Median :0.00   Median :0.000000
##  Mean   :0.007787   Mean   :0.009208   Mean   :0.01   Mean   :0.007787
##  3rd Qu.:0.000000   3rd Qu.:0.000000   3rd Qu.:0.00   3rd Qu.:0.000000
##  Max.   :1.000000   Max.   :1.000000   Max.   :1.00   Max.   :1.000000
##    P.SEC.SGG            P.SEC.SI      P.SEC.SIS           P.SEC.SLDM
##  Min.   :0.000000   Min.   :0.00   Min.   :0.000000   Min.   :0.000000
##  1st Qu.:0.000000   1st Qu.:0.00   1st Qu.:0.000000   1st Qu.:0.000000
##  Median :0.000000   Median :0.00   Median :0.000000   Median :0.000000
##  Mean   :0.009208   Mean   :0.01   Mean   :0.007787   Mean   :0.009208
##  3rd Qu.:0.000000   3rd Qu.:0.00   3rd Qu.:0.000000   3rd Qu.:0.000000
##  Max.   :1.000000   Max.   :1.00   Max.   :1.000000   Max.   :1.000000
##    P.SEC.SOC      P.SEC.SPES         P.SEC.SRMS          P.SEC.SSPA
##  Min.   :0.00   Min.   :0.000000   Min.   :0.000000   Min.   :0.00
##  1st Qu.:0.00   1st Qu.:0.000000   1st Qu.:0.000000   1st Qu.:0.00
##  Median :0.00   Median :0.000000   Median :0.000000   Median :0.00
##  Mean   :0.01   Mean   :0.007787   Mean   :0.009208   Mean   :0.01
##  3rd Qu.:0.00   3rd Qu.:0.000000   3rd Qu.:0.000000   3rd Qu.:0.00
##  Max.   :1.00   Max.   :1.000000   Max.   :1.000000   Max.   :1.00
##    P.SEC.TSHS
##  Min.   :0.000000
##  1st Qu.:0.000000
##  Median :0.000000
##  Mean   :0.007787
##  3rd Qu.:0.000000
##  Max.   :1.000000
```

```
#member.na <-na.convert.mean(member)
```