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
```

```
member %>% count(MEMTYPE)
```

```
## # A tibble: 14 x 2
##    MEMTYPE     n
##  * <chr>   <int>
##  1 I2YC      154
##  2 ICREP       8
##  3 IDEV      226
##  4 IFAM      111
##  5 IFREP      31
##  6 IK12      318
##  7 ILIFA     373
##  8 ILIFF     306
##  9 ILIFR       2
## 10 IPGRD    1551
## 11 IREG     8255
## 12 ISEN     1518
## 13 ISREP      21
## 14 ISTU     4720
```

```
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)
```

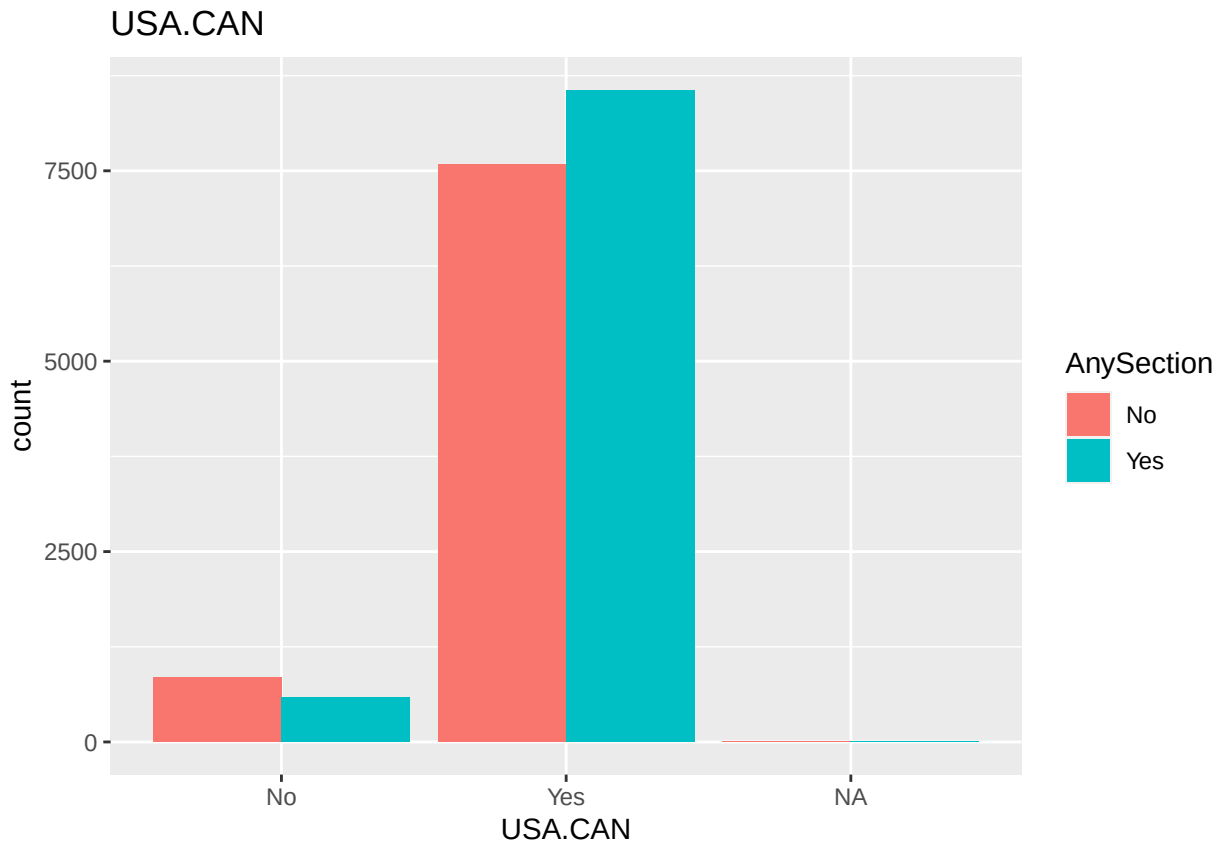
```
## # A tibble: 2 x 2
##   AnySection     n
## * <chr>      <int>
## 1 No          8447
## 2 Yes         9147
```

```
member[duplicated(member[,]),]
```

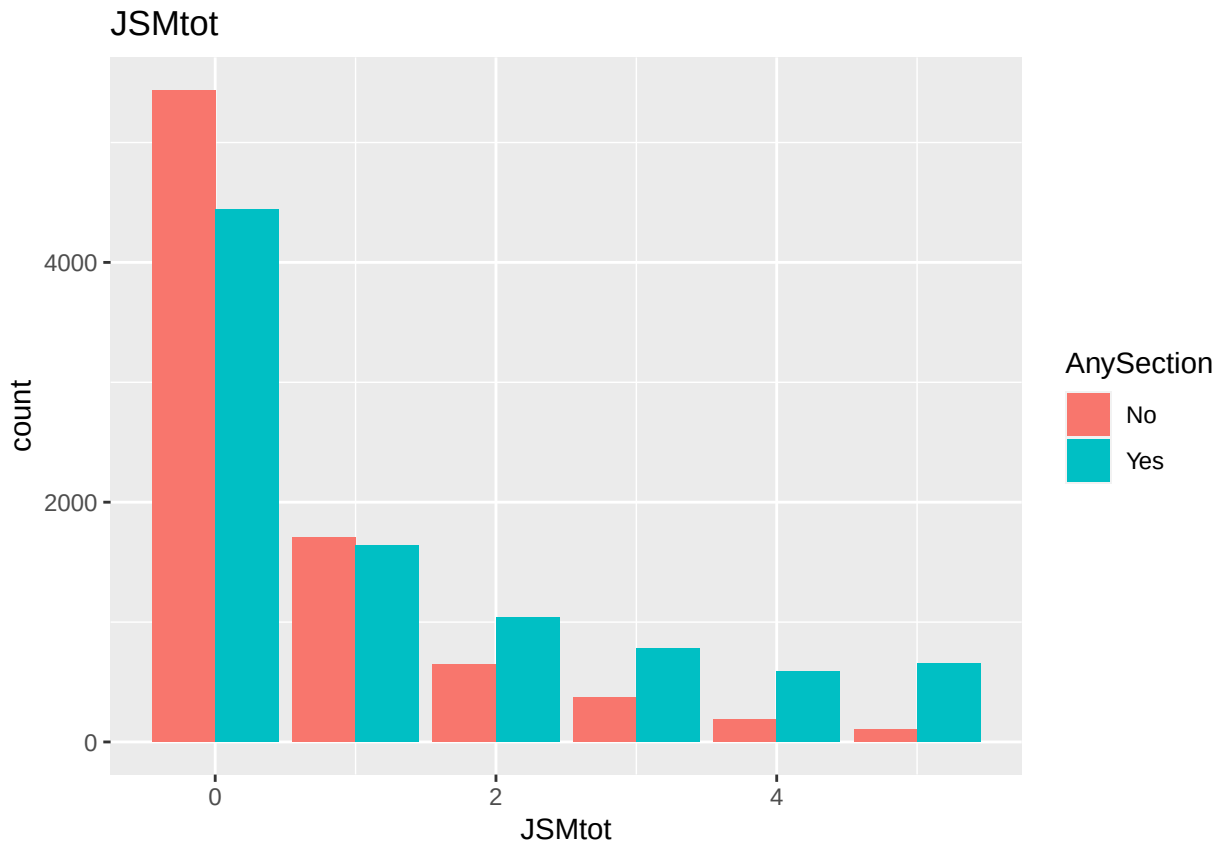
```
## # A tibble: 3,656 x 37
##    AnySection JSMtot USA.CAN DontPublish MEMTYPE   Age AgeJoinedASA Gender
##    <chr>       <dbl> <chr>   <chr>       <chr>   <dbl>        <dbl> <chr>
##  1 No              0 Yes     No          ISEN       NA           NA <NA>
##  2 No              0 Yes     No          ISEN       76           29 M
##  3 No              0 Yes     No          ISEN       79           29 M
##  4 No              0 Yes     No          ILIFA      NA           NA <NA>
##  5 Yes             1 Yes     No          ISEN       73           30 M
##  6 No              0 Yes     No          ISEN       71           23 M
##  7 No              0 Yes     No          ILIFA      NA           NA <NA>
##  8 No              0 Yes     No          ISEN       79           29 M
##  9 No              0 Yes     No          ISEN       72           25 M
## 10 No              0 Yes     No          ISEN       78           29 M
## # ... with 3,646 more rows, and 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>
```

Visualization

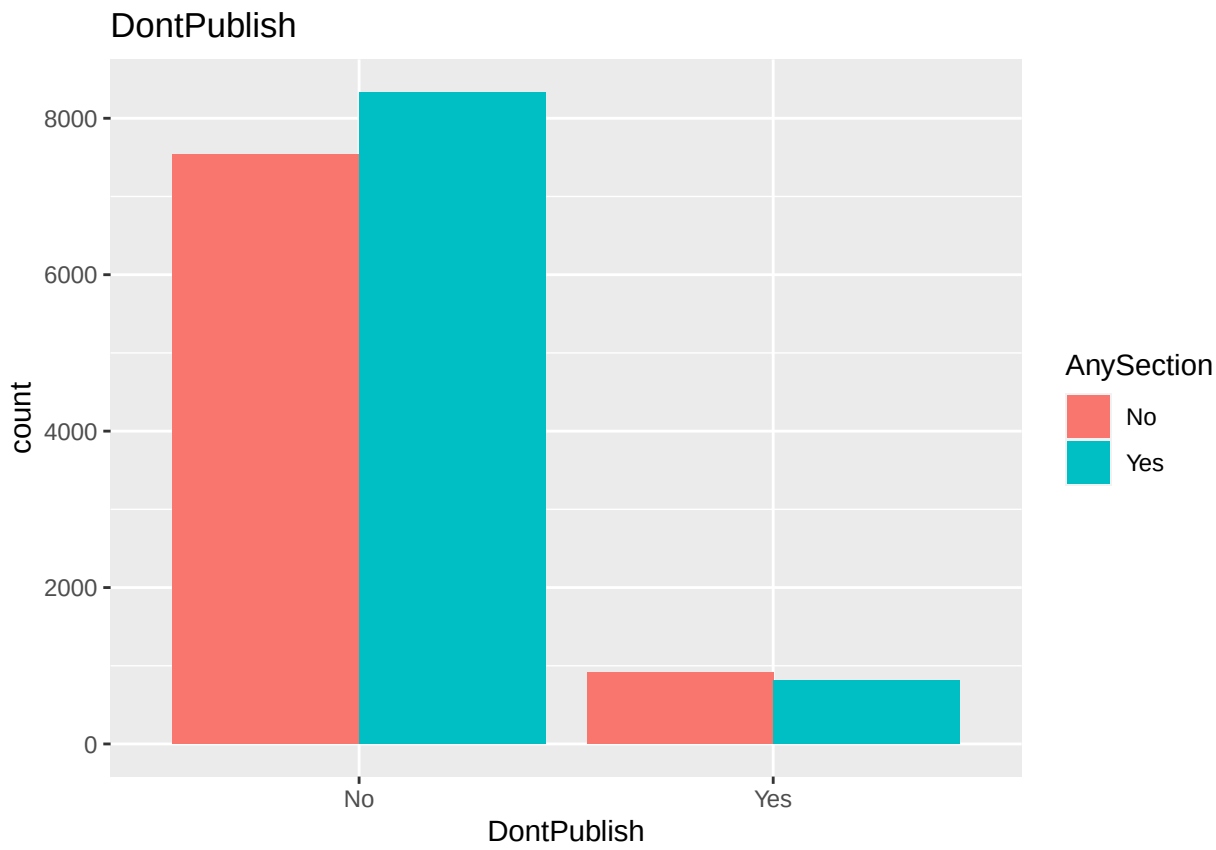
```
ggplot(member, aes(USA.CAN, fill = AnySection)) +
   geom_bar(position = 'dodge')+ggtitle("USA.CAN")
```



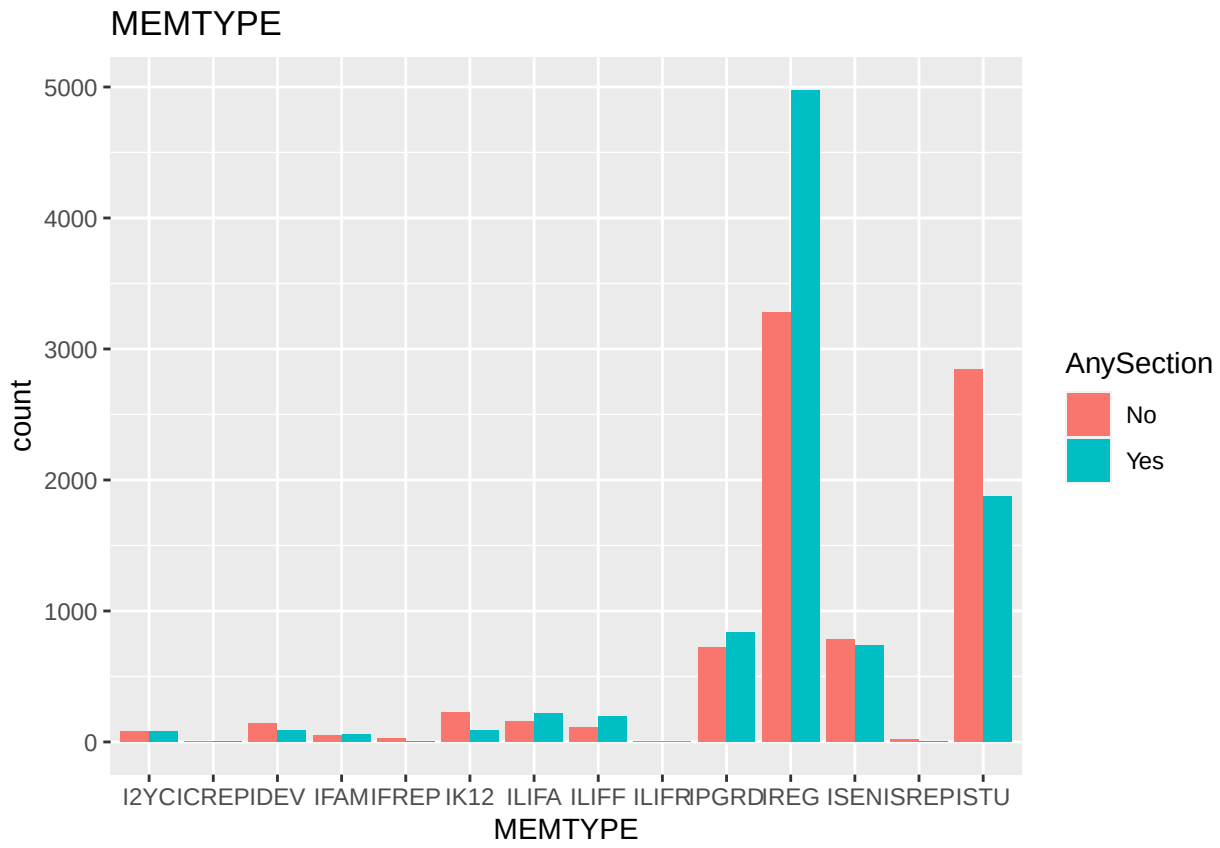

```
ggplot(member, aes(JSMtot, fill = AnySection)) +
   geom_bar(position = 'dodge')+ggtitle("JSMtot")
```




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

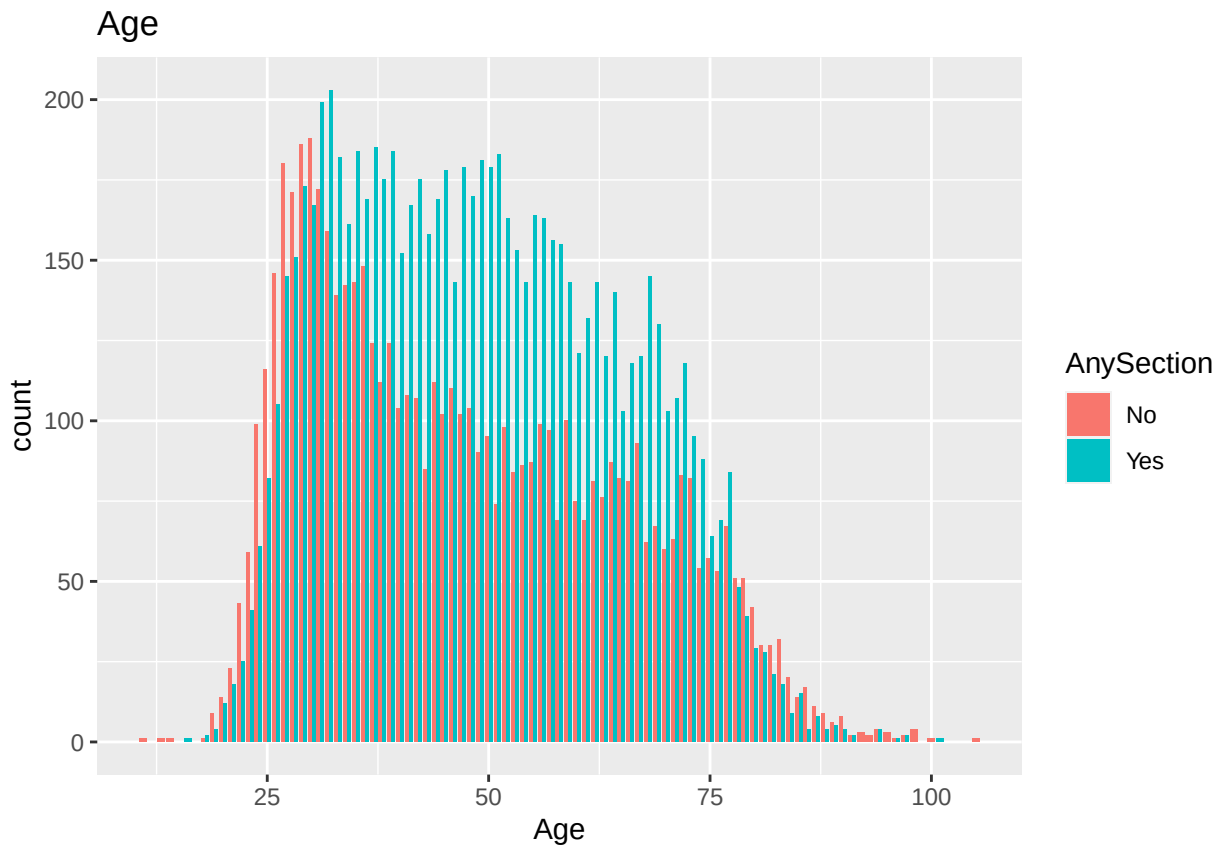



```
ggplot(member, aes(MEMTYPE, fill = AnySection)) +
   geom_bar(position = 'dodge')+ggtitle("MEMTYPE")
```



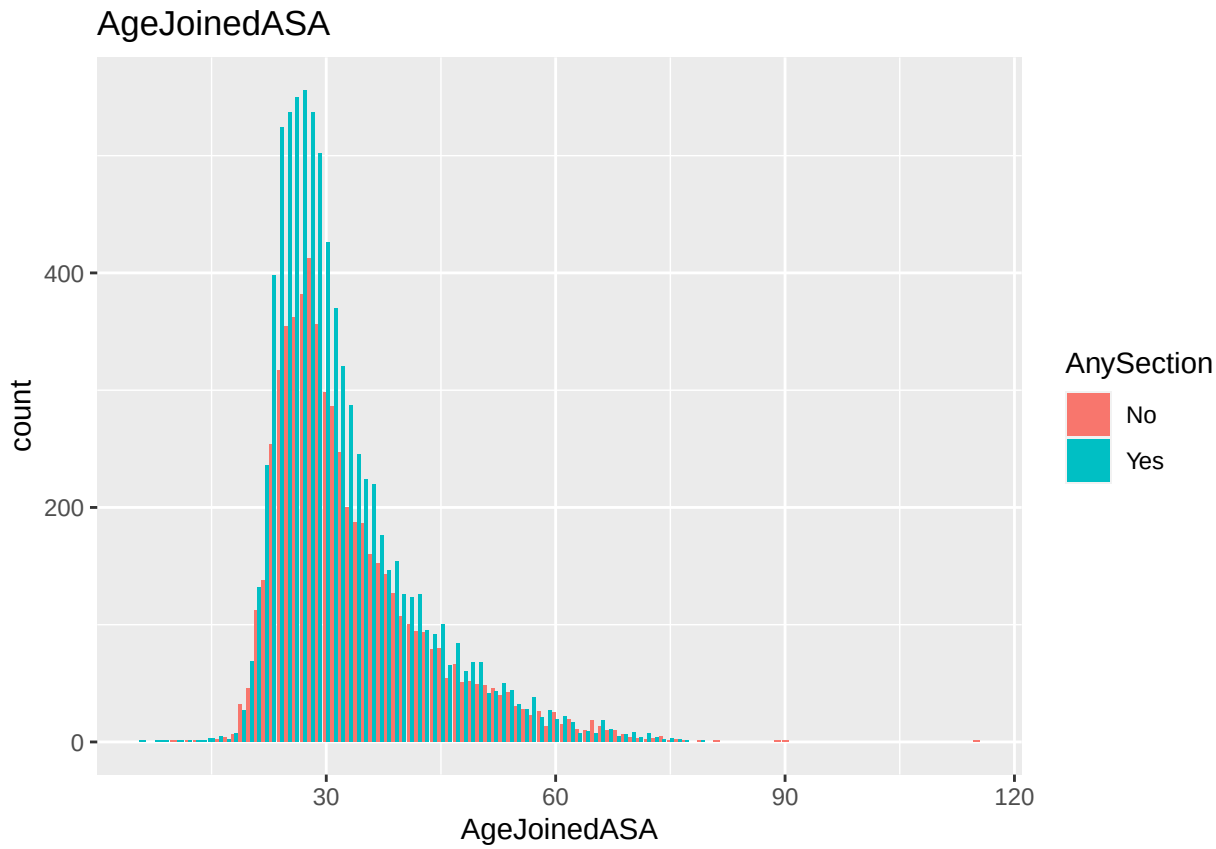

```
ggplot(member, aes(Age, fill = AnySection)) +
   geom_bar(position = 'dodge')+ggtitle("Age")
```

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

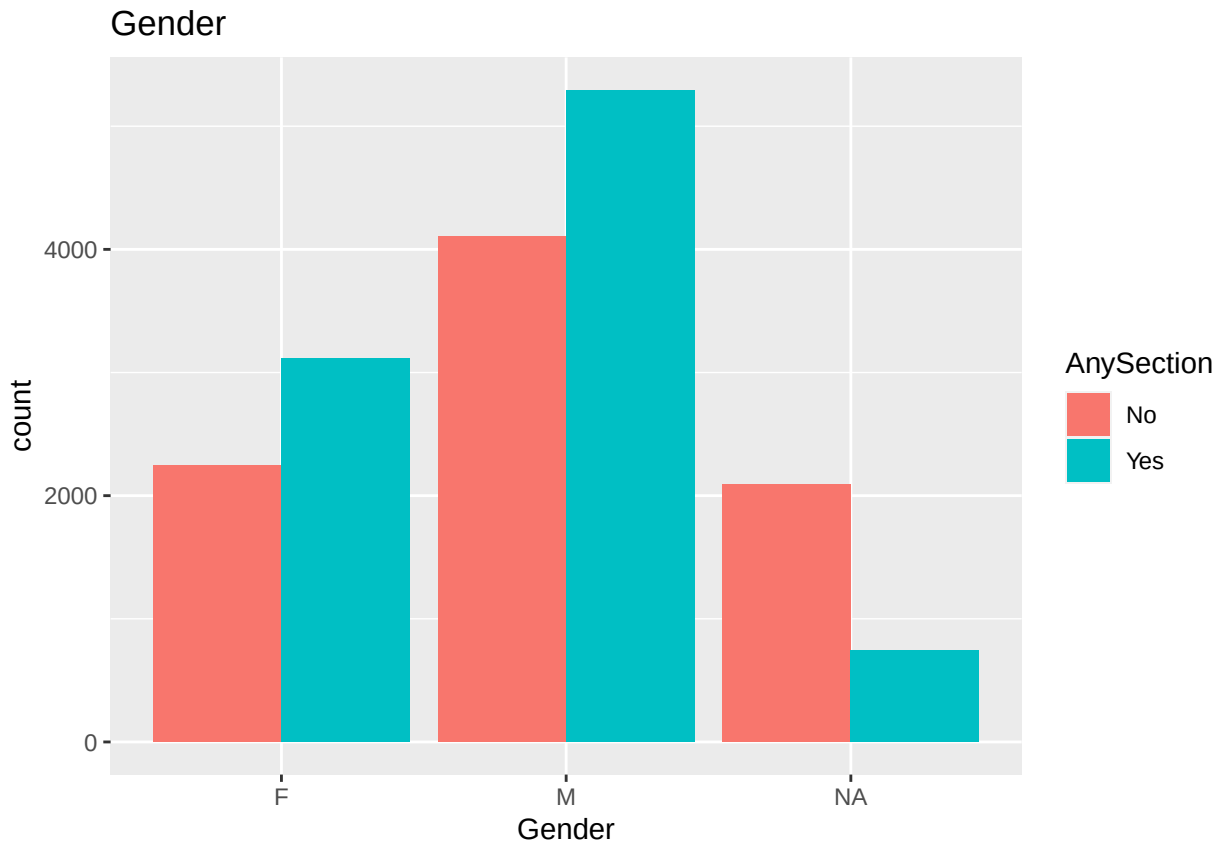



```
ggplot(member, aes(AgeJoinedASA, fill = AnySection)) +
   geom_bar(position = 'dodge')+ggtitle("AgeJoinedASA")
```

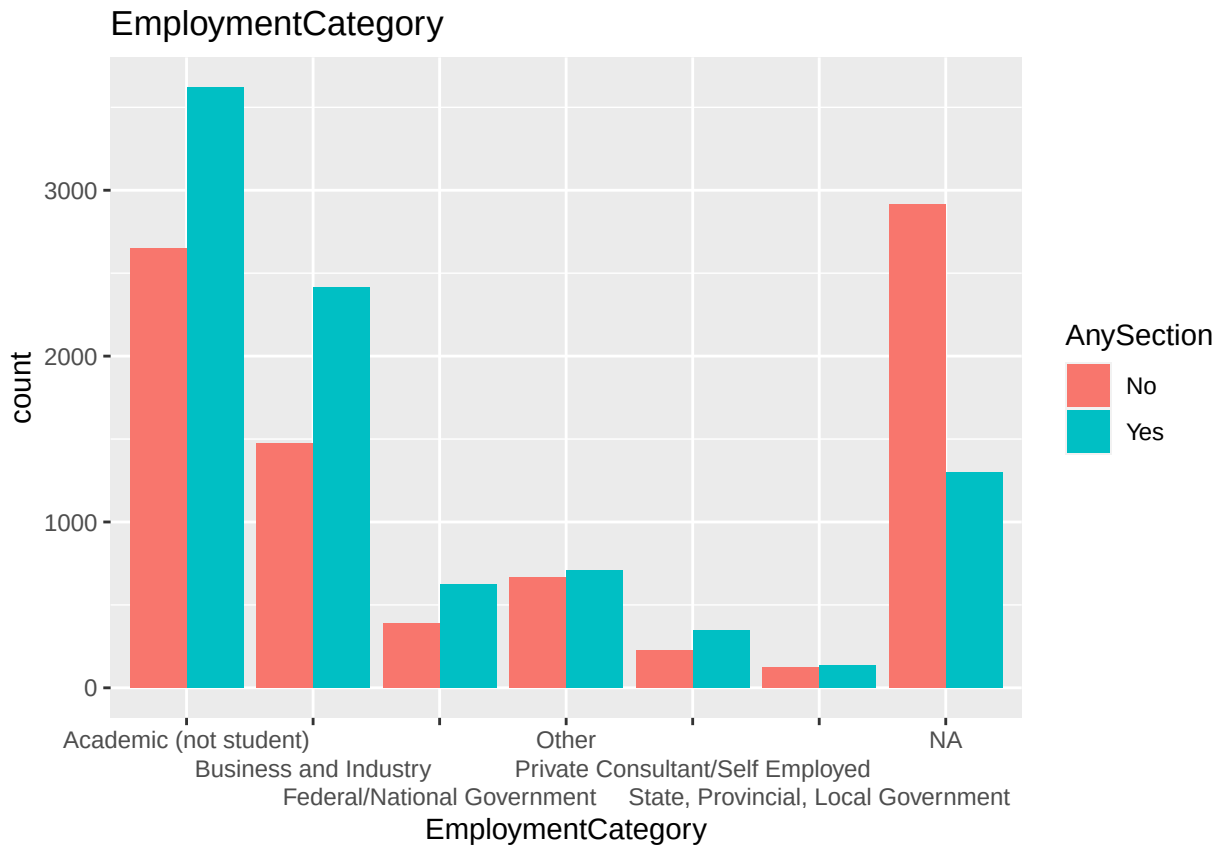
```
## Warning: Removed 3400 rows containing non-finite values (stat_count).
```



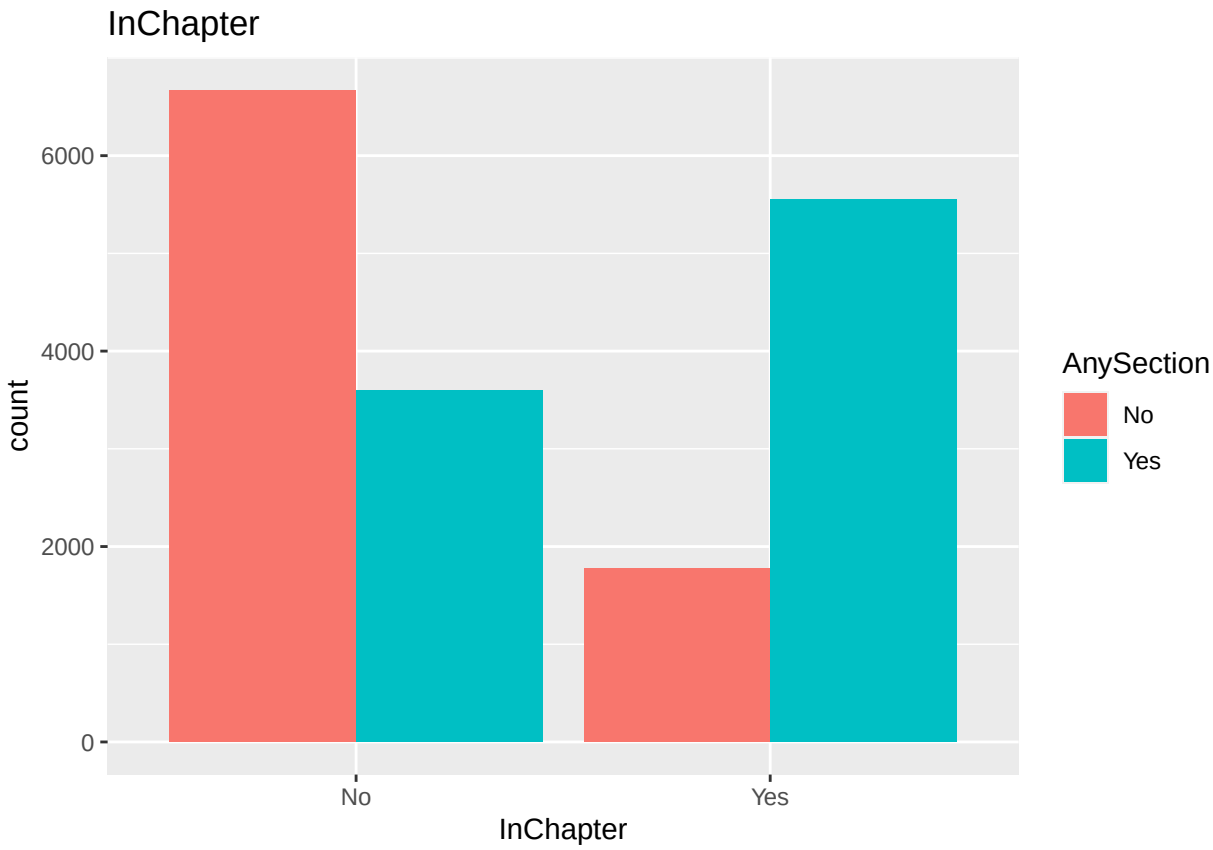

```
ggplot(member, aes(Gender, fill = AnySection)) +
   geom_bar(position = 'dodge')+ggtitle("Gender")
```




```
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

Turn char columns to numeric factor values

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

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

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

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

Missing value

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))
```

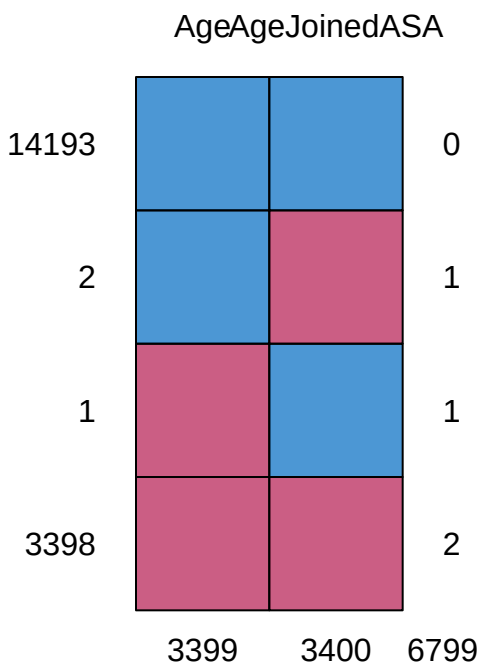
```
## [1] 2834
```

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

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

Impute the missing value of Age and AgeJoinedASA using MICE method. From the pattern matrix, we could see that the missing values in these two columns are systematic.

```
md.pattern(member[,6:7])
```




```
##        Age AgeJoinedASA
## 14193    1            1    0
## 2        1            0    1
## 1        0            1    1
## 3398     0            0    2
##       3399         3400 6799
```

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

```
## Warning: Number of logged events: 2
```

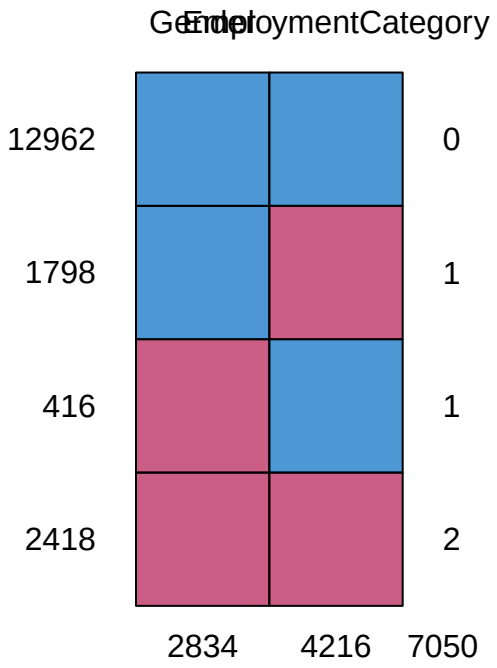
```
member.imp <- complete(imp)
summary(member.imp)
```

```
##    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 : 45.00   Median : 30.00   Mode  :character
##  Mean   :10.33   Mean   : 47.37   Mean   : 32.45
##  3rd Qu.:13.00   3rd Qu.: 60.00   3rd Qu.: 36.75
##  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.   :0.00000   Min.   :0.000000
##  1st Qu.:0.000000   1st Qu.:0.000000   1st Qu.:0.00000   1st Qu.:0.000000
##  Median :0.000000   Median :0.000000   Median :0.00000   Median :0.000000
##  Mean   :0.007787   Mean   :0.009208   Mean   :0.01267   Mean   :0.006252
##  3rd Qu.:0.000000   3rd Qu.:0.000000   3rd Qu.:0.00000   3rd Qu.:0.000000
##  Max.   :1.000000   Max.   :1.000000   Max.   :1.00000   Max.   :1.000000
##    P.SEC.ENVR        P.SEC.EPI         P.SEC.GOVT        P.SEC.GRPH
##  Min.   :0.000000   Min.   :0.000000   Min.   :0.000000   Min.   :0.00000
##  1st Qu.:0.000000   1st Qu.:0.000000   1st Qu.:0.000000   1st Qu.:0.00000
##  Median :0.000000   Median :0.000000   Median :0.000000   Median :0.00000
##  Mean   :0.004092   Mean   :0.009151   Mean   :0.005058   Mean   :0.01051
##  3rd Qu.:0.000000   3rd Qu.:0.000000   3rd Qu.:0.000000   3rd Qu.:0.00000
##  Max.   :1.000000   Max.   :1.000000   Max.   :1.000000   Max.   :1.00000
##    P.SEC.HPSS        P.SEC.MDD         P.SEC.MHS         P.SEC.MKTG
##  Min.   :0.000000   Min.   :0.000000   Min.   :0.000000   Min.   :0.000000
##  1st Qu.:0.000000   1st Qu.:0.000000   1st Qu.:0.000000   1st Qu.:0.000000
##  Median :0.000000   Median :0.000000   Median :0.000000   Median :0.000000
##  Mean   :0.004149   Mean   :0.001592   Mean   :0.001989   Mean   :0.003751
##  3rd Qu.:0.000000   3rd Qu.:0.000000   3rd Qu.:0.000000   3rd Qu.:0.000000
##  Max.   :1.000000   Max.   :1.000000   Max.   :1.000000   Max.   :1.000000
##    P.SEC.NPAR         P.SEC.QP         P.SEC.SBSS        P.SEC.SDNS
##  Min.   :0.000000   Min.   :0.000000   Min.   :0.000000   Min.   :0.000000
##  1st Qu.:0.000000   1st Qu.:0.000000   1st Qu.:0.000000   1st Qu.:0.000000
##  Median :0.000000   Median :0.000000   Median :0.000000   Median :0.000000
##  Mean   :0.003694   Mean   :0.004206   Mean   :0.007844   Mean   :0.002217
##  3rd Qu.:0.000000   3rd Qu.:0.000000   3rd Qu.:0.000000   3rd Qu.:0.000000
##  Max.   :1.000000   Max.   :1.000000   Max.   :1.000000   Max.   :1.000000
##    P.SEC.SGG          P.SEC.SI         P.SEC.SIS          P.SEC.SLDM
##  Min.   :0.000000   Min.   :0.00000   Min.   :0.000000   Min.   :0.000000
##  1st Qu.:0.000000   1st Qu.:0.00000   1st Qu.:0.000000   1st Qu.:0.000000
##  Median :0.000000   Median :0.00000   Median :0.000000   Median :0.000000
##  Mean   :0.001648   Mean   :0.00125   Mean   :0.004036   Mean   :0.007957
##  3rd Qu.:0.000000   3rd Qu.:0.00000   3rd Qu.:0.000000   3rd Qu.:0.000000
##  Max.   :1.000000   Max.   :1.00000   Max.   :1.000000   Max.   :1.000000
##    P.SEC.SOC         P.SEC.SPES         P.SEC.SRMS         P.SEC.SSPA
##  Min.   :0.000000   Min.   :0.000000   Min.   :0.000000   Min.   :0.000000
##  1st Qu.:0.000000   1st Qu.:0.000000   1st Qu.:0.000000   1st Qu.:0.000000
##  Median :0.000000   Median :0.000000   Median :0.000000   Median :0.000000
##  Mean   :0.004547   Mean   :0.004036   Mean   :0.007332   Mean   :0.004377
##  3rd Qu.:0.000000   3rd Qu.:0.000000   3rd Qu.:0.000000   3rd Qu.:0.000000
##  Max.   :1.000000   Max.   :1.000000   Max.   :1.000000   Max.   :1.000000
##    P.SEC.TSHS
##  Min.   :0.000000
##  1st Qu.:0.000000
##  Median :0.000000
##  Mean   :0.005172
##  3rd Qu.:0.000000
##  Max.   :1.000000
```

Description of member.imp So far member.imp has JSMtot imputed by mean imputation, Age and AgeJoinedASA imputed by MICE method. Gender and EmploymentCategory still have missing values. From the pattern matrix, we see an large overlap of missing values in the two columns.

```
md.pattern(member[,8:9])
```




```
##       Gender EmploymentCategory
## 12962      1                  1    0
## 1798       1                  0    1
## 416        0                  1    1
## 2418       0                  0    2
##         2834               4216 7050
```

Correlation Matrix

```
library("Hmisc")
```

```
## Loading required package: lattice
```

```
## Loading required package: survival
```

```
## Loading required package: Formula
```

```
##
## Attaching package: 'Hmisc'
```

```
## The following objects are masked from 'package:dplyr':
##
##     src, summarize
```

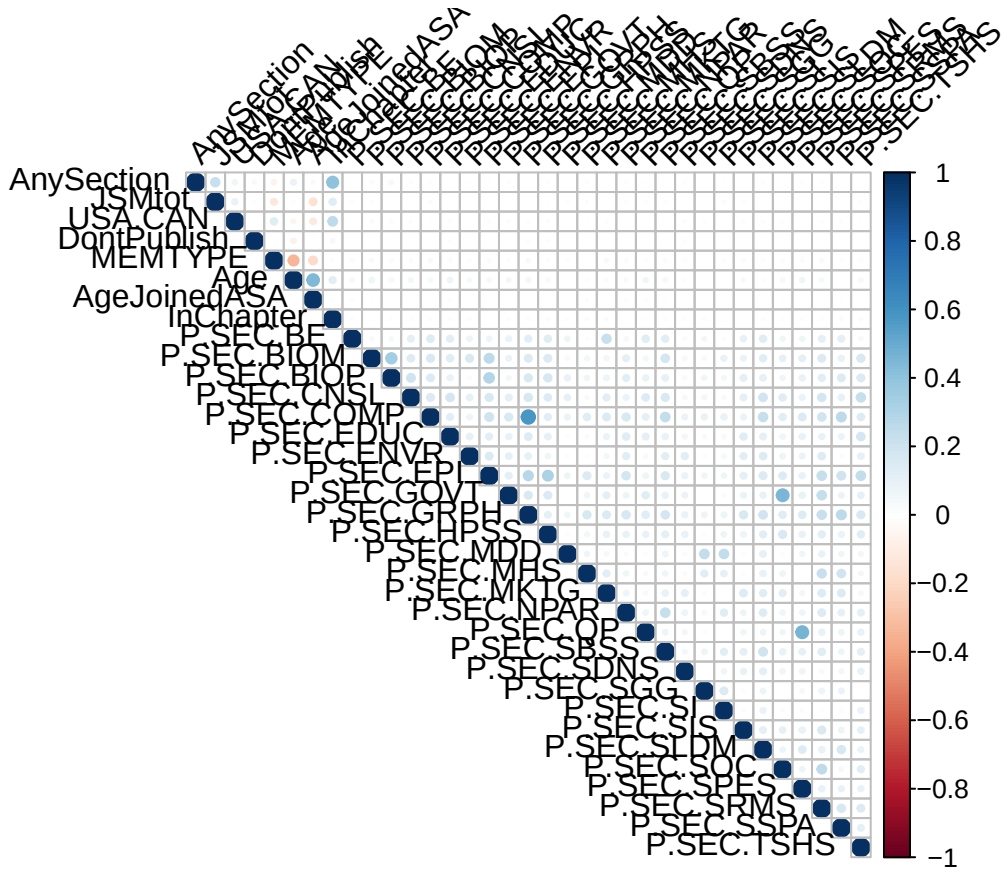
```
## The following objects are masked from 'package:base':
##
##     format.pval, units
```

```
library(corrplot)
```

```
## corrplot 0.84 loaded
```

```
sub.member.imp =cbind(member.imp[,0:7],member.imp[,10:37])
cor <- cor(as.matrix(sub.member.imp))
```

```
corrplot(cor, type = "upper",
         tl.col = "black", tl.srt = 45)
```




```
# ++++++++++++++++++++++++++++
# flattenCorrMatrix
# ++++++++++++++++++++++++++++
# cormat : matrix of the correlation coefficients
# pmat : matrix of the correlation p-values
flattenCorrMatrix <- function(cormat, pmat) {
  ut <- upper.tri(cormat)
  data.frame(
    row = rownames(cormat)[row(cormat)[ut]],
    column = rownames(cormat)[col(cormat)[ut]],
    cor  =(cormat)[ut],
    p = pmat[ut]
    )
}
cor2<-rcorr(as.matrix(sub.member.imp))
cor_table = flattenCorrMatrix(cor2$r, cor2$P)
head(cor_table[order(abs(cor_table$cor),decreasing = TRUE),],10)
```

```
##            row       column        cor p
## 149 P.SEC.COMP   P.SEC.GRPH  0.5810733 0
## 489   P.SEC.QP   P.SEC.SPES  0.4668699 0
## 452 P.SEC.GOVT    P.SEC.SOC  0.4596058 0
## 21         Age AgeJoinedASA  0.4438895 0
```

```
## 22  AnySection    InChapter  0.4020789 0
## 55  P.SEC.BIOM   P.SEC.BIOP  0.3484721 0
## 15     MEMTYPE          Age -0.3308769 0
## 169  P.SEC.EPI   P.SEC.HPSS  0.3188039 0
## 116 P.SEC.BIOP    P.SEC.EPI  0.2970723 0
## 152  P.SEC.EPI   P.SEC.GRPH  0.2709819 0
```