클린업 3주차 패키지 오답노트

서희나

2023-01-18

Chapter 1 : Data Preprocessing & EDA

문제1. data1을 fread 함수를 통해 불러오고, 데이터 구조를 파악해보세요.

- 특정 데이터셋은 NA로 처리된 값이 blank 형태로 되어있는 경우가 있습니다.
- Fread 함수 내에서 불러올 때, na.strings 옵션을 통해서 NA값을 갖도록 데이터셋을 불러오세요

```
pacman::p_load(tidyverse,
               caret,
               data.table,
               magrittr,
               corrplot,
               cluster,
               Rtsne)
```

```
data<-fread('data.csv',na.strings=c('',NA))
```

```
data %>% str
```

```
## Classes 'data.table' and 'data.frame':   26457 obs. of  20 variables:
##  $ index        : int  0 1 2 3 4 5 6 7 8 9 ...
##  $ gender       : chr  "F" "F" "M" "F" ...
##  $ car          : chr  "N" "N" "Y" "N" ...
##  $ reality      : chr  "N" "Y" "Y" "Y" ...
##  $ child_num    : int  0 1 0 0 0 2 0 0 1 0 ...
##  $ income_total : num  202500 247500 450000 202500 157500 ...
##  $ income_type  : chr  "Commercial associate" "Commercial associate" "Working" "Commercial a
ssociate" ...
##  $ edu_type     : chr  "Higher education" "Secondary / secondary special" "Higher education"
"Secondary / secondary special" ...
##  $ family_type  : chr  "Married" "Civil marriage" "Married" "Married" ...
##  $ house_type   : chr  "Municipal apartment" "House / apartment" "House / apartment" "House
/ apartment" ...
##  $ DAYS_BIRTH   : int  -13899 -11380 -19087 -15088 -15037 -13413 -17570 -14896 -15131 -15785
...
##  $ DAYS_EMPLOYED: int  -4709 -1540 -4434 -2092 -2105 -4996 -1978 -5420 -1466 -1308 ...
##  $ FLAG_MOBIL   : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ work_phone   : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ phone        : int  0 0 1 1 0 0 0 0 0 0 ...
##  $ email        : int  0 1 0 0 0 1 1 1 1 0 ...
##  $ occyp_type   : chr  NA "Laborers" "Managers" "Sales staff" ...
##  $ family_size  : num  2 3 2 2 2 4 1 2 3 2 ...
##  $ begin_month  : num  -6 -5 -22 -37 -26 -18 -41 -53 -38 -5 ...
##  $ credit       : num  1 1 2 0 2 1 2 0 2 2 ...
##  - attr(*, ".internal.selfref")=<externalptr>
```

문제2. 변수별 결측치의 개수를 파악하고, index 변수를 제거해주세요. 또한 결측치가 많은 열을 제거해주세요.

```
data %>% is.na %>% colSums
```

```
##         index        gender           car       reality     child_num
##             0             0             0             0             0
##  income_total   income_type      edu_type   family_type    house_type
##             0             0             0             0             0
##    DAYS_BIRTH DAYS_EMPLOYED    FLAG_MOBIL    work_phone         phone
##             0             0             0             0             0
##         email    occyp_type   family_size   begin_month        credit
##             0          8171             0             0             0
```

```
data %<>% select(-c("index", "occyp_type"))
```

문제3. 각 변수별로 unique한 값이 몇 개씩 존재하는지 파악해주세요.

```
data %>% lapply(n_distinct)
```

```
## $gender
## [1] 2
##
## $car
## [1] 2
##
## $reality
## [1] 2
##
## $child_num
## [1] 9
##
## $income_total
## [1] 249
##
## $income_type
## [1] 5
##
## $edu_type
## [1] 5
##
## $family_type
## [1] 5
##
## $house_type
## [1] 6
##
## $DAYS_BIRTH
## [1] 6621
##
## $DAYS_EMPLOYED
## [1] 3470
##
## $FLAG_MOBIL
## [1] 1
##
## $work_phone
## [1] 2
##
## $phone
## [1] 2
##
## $email
## [1] 2
##
## $family_size
## [1] 10
##
## $begin_month
## [1] 61
##
## $credit
## [1] 3
```

문제4. Child_num은 자녀의 수입니다. 자녀의 수가 6명 이상인 사람은 이상치로 간주하여 제외해주세요.

```
data %<>% filter(!data$child_num>=6)
```

문제5. 범주형 변수 중 factor의 개수가 하나인 변수는 분석에 의미가 없으므로 제외해주세요.

```
data %<>%
  select_if(summarise_all(.,n_distinct)!=1)
```

문제6. DAYS_BIRTH 변수는 현재 시점 기준으로 몇일 전 태어났는지를 나타내는 변수입니다. 만 나이로 파생변수를 만들고 기존 변수는 삭제해주세요. (버림 사용)

```
nrow(data)==sum(data$DAYS_BIRTH<0)
```

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

```
data %<>%
  mutate(age=((DAYS_BIRTH %>% abs)%/%365)) %>%
  select(-DAYS_BIRTH)
```

문제7. DAYS_EMPLOYED 변수는 현재 시점 기준으로 며칠 전 취업을 했는지 나타내는 변수입니다. 만 근속연수에 대한 파생변수를 만들고 기존 변수는 삭제해주세요. (버림 사용)

- DAYS_EMPLOYED 변수 내에서 양수 값이 존재하는지 확인해주세요.
- 양수 값이 존재한다면 현재 고용 상태가 아닌 인원이므로, 값을 0으로 대체해주세요.

```
sum(data$DAYS_EMPLOYED>0)
```

```
## [1] 4438
```

```
data$DAYS_EMPLOYED %<>%ifelse((.)>0,0,.)
```

```
data %<>%
  mutate(YEARS_EMPLOYED=((DAYS_EMPLOYED %>% abs)%/%365)) %>%
  select(-DAYS_EMPLOYED)
```

문제8. Begin_month 변수는 현재 시점 기준으로 신용카드를 몇 달 전에 발급받았는지 나타냅니다.양수로 변환해주세요.

```
data %<>% mutate_at(vars(begin_month),abs)
```

문제9. 범주형 변수는 factor형으로, 수치형 변수는 numeric으로 바꿔주세요.

```
data %>%
  select_if(summarise_all(.,n_distinct)<100) %>%
  lapply(unique)
```

```
## $gender
## [1] "F" "M"
##
## $car
## [1] "N" "Y"
##
## $reality
## [1] "N" "Y"
##
## $child_num
## [1] 0 1 2 3 4 5
##
## $income_type
## [1] "Commercial associate" "Working"              "State servant"
## [4] "Pensioner"            "Student"
##
## $edu_type
## [1] "Higher education"              "Secondary / secondary special"
## [3] "Incomplete higher"             "Lower secondary"
## [5] "Academic degree"
##
## $family_type
## [1] "Married"               "Civil marriage"       "Separated"
## [4] "Single / not married" "Widow"
##
## $house_type
## [1] "Municipal apartment" "House / apartment"   "With parents"
## [4] "Co-op apartment"     "Rented apartment"    "Office apartment"
##
## $work_phone
## [1] 0 1
##
## $phone
## [1] 0 1
##
## $email
## [1] 0 1
##
## $family_size
## [1] 2 3 4 1 5 6 7
##
## $begin_month
##  [1]  6  5 22 37 26 18 41 53 38 40 51 60  2 14  7 35  4 13 57 47 33 30 20  8 39
## [26] 21 19 24 48 12 10 42 29  3 23 25  1 15 32 59 54 34  0 27 45 56 46  9 44 36
## [51] 43 49 11 55 58 28 52 17 50 16 31
##
## $credit
## [1] 1 2 0
##
## $age
##  [1] 38 31 52 41 36 48 40 43 32 27 62 35 53 24 63 37 54 58 39 61 47 60 55 45 59
## [26] 34 51 57 50 64 65 42 44 46 33 28 26 49 25 29 66 56 30 22 23 67 68 21
##
## $YEARS_EMPLOYED
```

```
##  [1] 12  4  5 13 14  3  6  0 11  2 16  1 10  9  8 25  7 15 33 23 27 17 32 20 30
## [26] 22 21 18 19 29 31 24 26 28 36 41 34 35 40 39 37 42 43 38
```

```
data %<>%
  mutate_if(summarise_all(.,n_distinct)<=6,as.factor) %>%
  mutate_if(!(summarise_all(.,n_distinct)<=6),as.numeric) %>%
  mutate_at(vars(child_num),as.numeric)
```

```
data %>% glimpse
```

```
## Rows: 26,451
## Columns: 17
## $ gender         <fct> F, F, M, F, F, F, F, M, M, F, F, M, F, F, F, M, M, M, F…
## $ car            <fct> N, N, Y, N, Y, N, N, N, Y, N, N, Y, Y, N, N, Y, N, N, N…
## $ reality        <fct> N, Y, Y, Y, Y, Y, N, Y, Y, Y, Y, Y, Y, Y, N, Y, Y, Y, N, Y…
## $ child_num      <dbl> 1, 2, 1, 1, 1, 3, 1, 1, 2, 1, 1, 2, 2, 1, 1, 3, 1, 1, 1…
## $ income_total   <dbl> 202500, 247500, 450000, 202500, 157500, 270000, 315000,…
## $ income_type    <fct> Commercial associate, Commercial associate, Working, Co…
## $ edu_type       <fct> Higher education, Secondary / secondary special, Higher…
## $ family_type    <fct> Married, Civil marriage, Married, Married, Married, Mar…
## $ house_type     <fct> Municipal apartment, House / apartment, House / apartme…
## $ work_phone     <fct> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0…
## $ phone          <fct> 0, 0, 1, 1, 0, 0, 0, 0, 0, 0, 1, 0, 0, 1, 1, 1, 1, 0, 0…
## $ email          <fct> 0, 1, 0, 0, 0, 1, 1, 1, 1, 0, 1, 0, 0, 0, 0, 0, 0, 1, 0…
## $ family_size    <dbl> 2, 3, 2, 2, 2, 4, 1, 2, 3, 2, 1, 3, 3, 2, 1, 4, 2, 2, 2…
## $ begin_month    <dbl> 6, 5, 22, 37, 26, 18, 41, 53, 38, 5, 40, 51, 60, 41, 41…
## $ credit         <fct> 1, 1, 2, 0, 2, 1, 2, 0, 2, 2, 2, 2, 0, 2, 2, 0, 1, 2, 2…
## $ age            <dbl> 38, 31, 52, 41, 41, 36, 48, 40, 41, 43, 52, 32, 32, 27,…
## $ YEARS_EMPLOYED <dbl> 12, 4, 12, 5, 5, 13, 5, 14, 4, 3, 6, 0, 5, 6, 0, 11, 12…
```

문제10. 수치형 변수와 범주형 변수들에 대해서 각각 상관계수를 구하고, 시각화를 진행해주세요.

- 절댓값이 0.5 이상의 높은 상관관계를 보이는 변수들에 대해 설명해주세요.
- Child_num 변수를 제거해주세요.

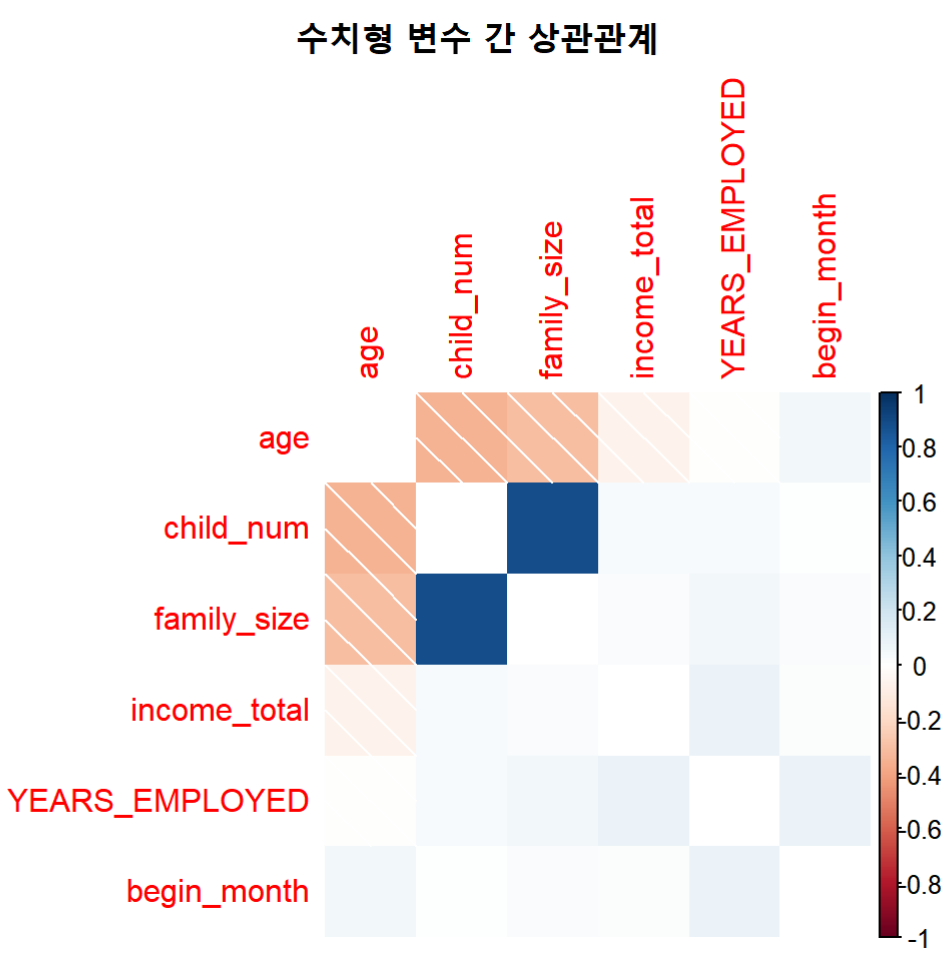
```
library(corrplot)
```

```
num_corr<-cor(data %>% select_if(is.numeric),method='pearson')
num_corr
```

```
##                  child_num income_total family_size begin_month          age
## child_num       1.000000000   0.03344159  0.88613545 0.007353323 -0.343914366
## income_total    0.033441586   1.00000000  0.02445778 0.017992531 -0.064074414
## family_size     0.886135448   0.02445778  1.00000000 0.023426227 -0.303831435
## begin_month     0.007353323   0.01799253  0.02342623 1.000000000  0.057298656
## age            -0.343914366  -0.06407441 -0.30383143 0.057298656  1.000000000
## YEARS_EMPLOYED  0.039223241   0.08257037  0.05142162 0.083362310 -0.003145562
##                YEARS_EMPLOYED
## child_num          0.039223241
## income_total       0.082570372
## family_size        0.051421615
## begin_month        0.083362310
## age               -0.003145562
## YEARS_EMPLOYED     1.000000000
```

수치형 변수 경우에는 절댓값이 0.5이상의 높은 상관관계를 보이는 변수가 없다.

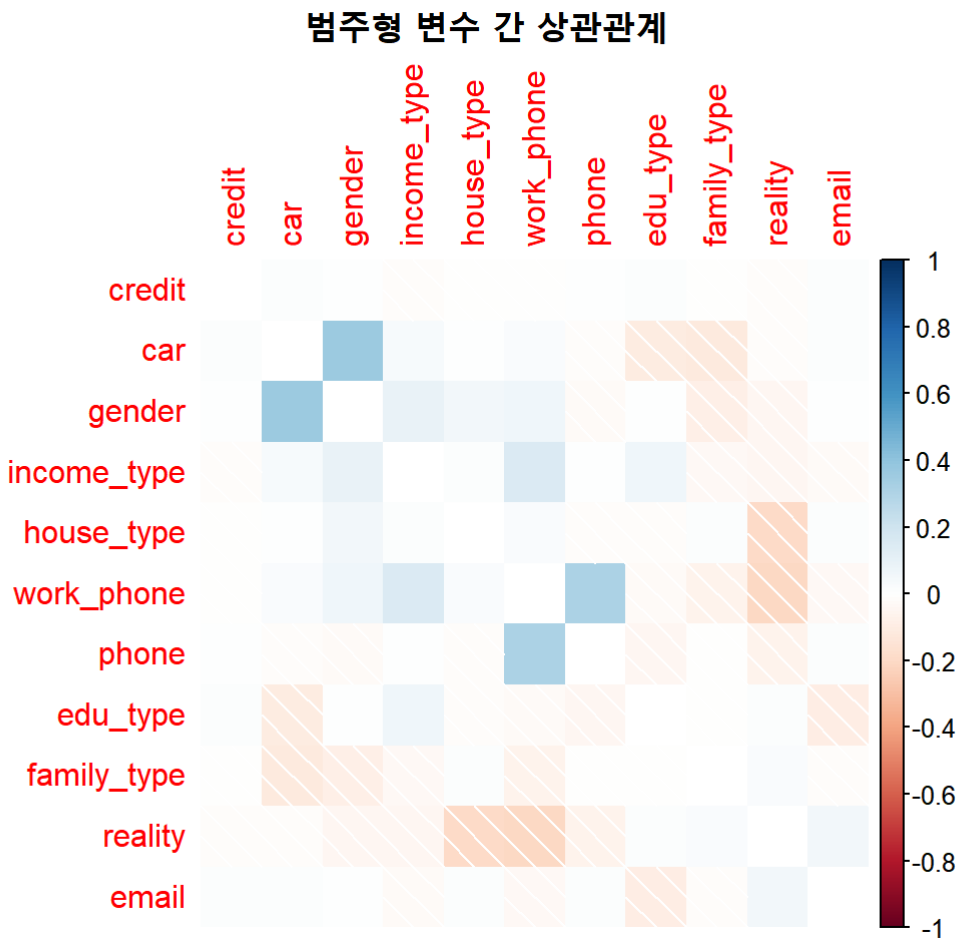
```
corrplot(num_corr,
         order='AOE',
         method='shade',
         diag = FALSE,
         title='수치형 변수 간 상관관계',
         mar=c(0,0,2,0))
```



```
cate_corr<-cor(data %>%
                 select_if(is.factor) %>%
                 mutate_if(is.factor,as.numeric),
               method='spearman')
cate_corr
```

```
##                    gender          car      reality  income_type    edu_type
## gender        1.0000000000  0.362149394 -0.04910970  0.093680445  0.00554969
## car           0.3621493942  1.000000000 -0.01676820  0.039046939 -0.10374334
## reality      -0.0491096977 -0.016768202  1.00000000 -0.043679250  0.01148677
## income_type   0.0936804452  0.039046939 -0.04367925  1.000000000  0.06398159
## edu_type      0.0055496899 -0.103743345  0.01148677  0.063981593  1.00000000
## family_type  -0.0855718067 -0.110956705  0.02548739 -0.030900305 -0.00493877
## house_type    0.0559731846  0.007564998 -0.19415487  0.019052872 -0.01205442
## work_phone    0.0645735300  0.026375915 -0.20852026  0.150176914 -0.02478204
## phone        -0.0268584704 -0.013503097 -0.06539264  0.003978208 -0.04470437
## email         0.0001333333  0.018924306  0.05109024 -0.026321773 -0.09481430
## credit        0.0025166601  0.011080633 -0.01295038 -0.013052742  0.01426985
##                family_type   house_type   work_phone        phone         email
## gender       -0.085571807  0.055973185  0.064573530 -0.026858470  0.0001333333
## car          -0.110956705  0.007564998  0.026375915 -0.013503097  0.0189243056
## reality       0.025487385 -0.194154868 -0.208520258 -0.065392644  0.0510902399
## income_type  -0.030900305  0.019052872  0.150176914  0.003978208 -0.0263217727
## edu_type     -0.004938770 -0.012054416 -0.024782042 -0.044704374 -0.0948142973
## family_type   1.000000000  0.012122818 -0.060711844 -0.008038982 -0.0178243279
## house_type    0.012122818  1.000000000  0.021784683 -0.017659621  0.0127344008
## work_phone   -0.060711844  0.021784683  1.000000000  0.310005140 -0.0310077476
## phone        -0.008038982 -0.017659621  0.310005140  1.000000000  0.0125285230
## email        -0.017824328  0.012734401 -0.031007748  0.012528523  1.0000000000
## credit       -0.005418205 -0.007988294 -0.003584115  0.006810132  0.0151953952
##                     credit
## gender        0.002516660
## car           0.011080633
## reality      -0.012950381
## income_type  -0.013052742
## edu_type      0.014269852
## family_type  -0.005418205
## house_type   -0.007988294
## work_phone   -0.003584115
## phone         0.006810132
## email         0.015195395
## credit        1.000000000
```

```
corrplot(cate_corr, method='shade',
         diag = FALSE,
         order = 'AOE',
         mar =c(0,0,2,0),
         title='범주형 변수 간 상관관계')
```




```
data %<>% select(-child_num)
```

Chapter 2 : Unsupervised Learning (Clustering)

문제0. cluster, Rtsne 패키지를 설치해주고 불러와주세요.

문제1. 원활한 클러스터링을 위해 범주형 변수들 중 gender와 income_type을 제외한 변수를 제거해주세요.

```
remove<-setdiff(data %>% select_if(is.factor) %>% colnames,
                c('gender','income_type'))
data %<>% select(-all_of(remove))
```

문제2. 변수의 영향력을 조정하기 위해 수치형 변수들에 대해 Min-Max Scaling을 진행해주세요.

(HINT2) caret 패키지의 preProcess 함수를 이용하면 편합니다. (range 방법 이용)

```
minmax<-preProcess(data,method='range')
data_sc<-predict(minmax,data)
```

문제3. 빠른 클러스터링을 위해 임의로 5,000개의 sample을 추출해주세요.

- set.seed(2930) 설정
- sample_n 함수 사용, 5,000개 sample 추출

```
set.seed(2930)
data_sample<-data_sc %>% sample_n(.,5000)
```

문제4. 데이터셋에 대해서 gower distance를 계산하고, 결과값을 matrix로 변환해 저장해주세요.

- 수치형 자료들에 대한 클러스터링은 일반적으로 Euclidian distance를 사용합니다. 하지만 범주형 자료가 섞인 혼합형 자료의 경우 Euclidian distance에 대해서 명확히 정의를 내리기가 어렵습니다. 따라서 Gower distance를 통해 거리를 구하여 자료간 유사도와 비유사도를 계산합니다.
- Gower distance에 대해 자세한 설명은 구글링을 통해 찾아보시면 좋습니다!
- cluster 패키지 내의 daisy 함수 사용, metric = “gower”

```
gower_distance<-daisy(data_sample,metric='gower') %>% as.matrix
```

```
# 가장 유사한 조합
data_sample[
  which(gower_distance == min(gower_distance[gower_distance != min(gower_distance)]),
        arr.ind = TRUE)[1, ], ]
```

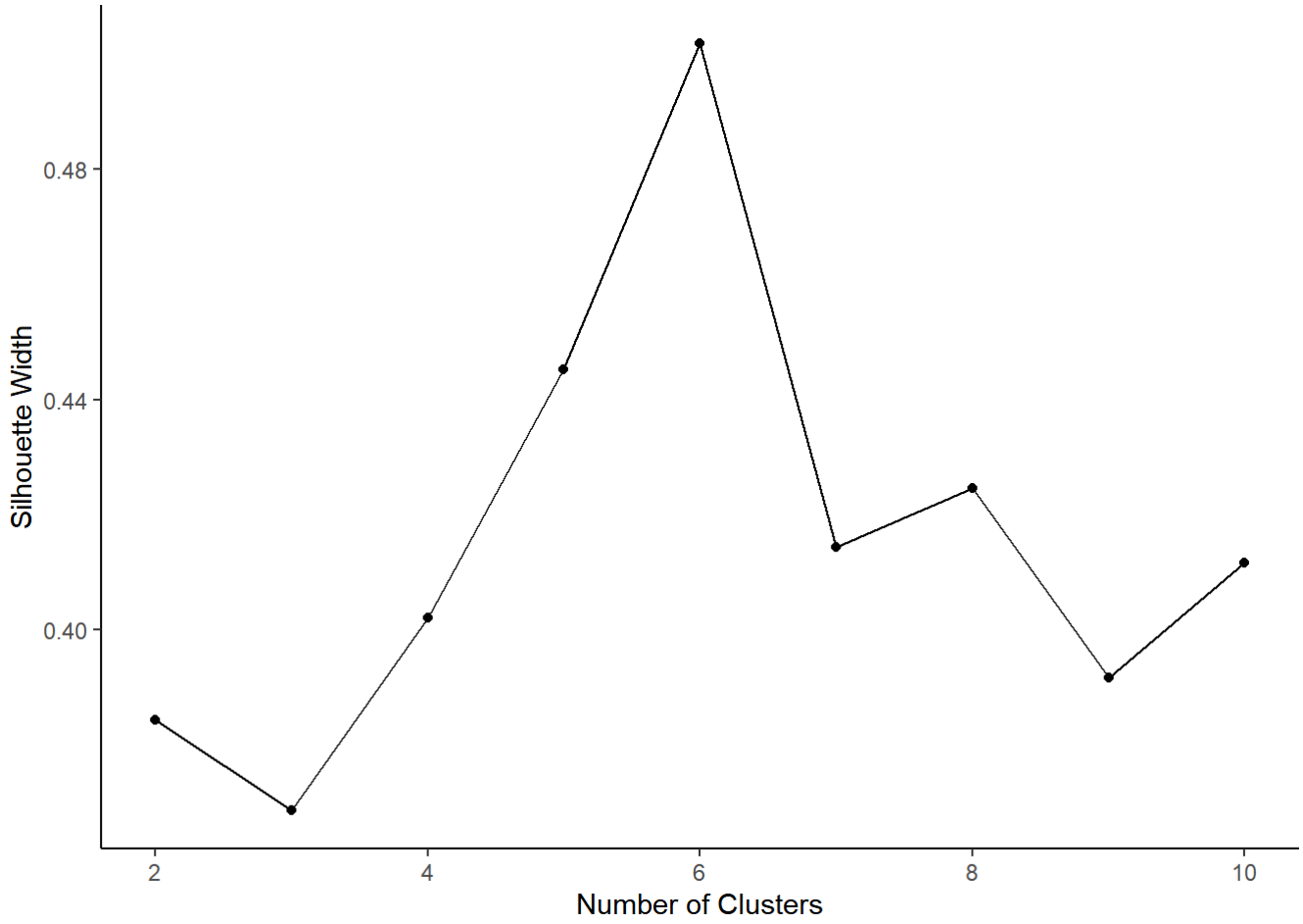
```
##    gender income_total income_type family_size begin_month       age
## 1:      F   0.03488372   Pensioner           0   0.3833333 0.9148936
## 2:      F   0.03197674   Pensioner           0   0.3833333 0.9148936
##    YEARS_EMPLOYED
## 1:              0
## 2:              0
```

```
# 가장 유사하지 않은 조합
data_sample[
  which(gower_distance == max(gower_distance[gower_distance != max(gower_distance)]),
        arr.ind = TRUE)[1, ], ]
```

```
##    gender income_total          income_type family_size begin_month       age
## 1:      M    0.5639535 Commercial associate         0.5   0.1666667 0.1276596
## 2:      F    0.1279070              Working         0.0   1.0000000 0.7872340
##    YEARS_EMPLOYED
## 1:     0.04651163
## 2:     0.83720930
```

```
silhouette<-c()

for(i in 2:10){
  fit<-pam(gower_distance,diss=T,k=i)
  silhouette<-c(silhouette,fit$silinfo$avg.width)
}
```

```
ggplot()+
  geom_point(aes(x=c(2:10),y=silhouette))+
  geom_line(mapping=aes(x=c(2:10),y=silhouette))+
  labs(x='Number of Clusters',y='Silhouette Width')+
  theme_classic()
```




```
set.seed(2930)
cluster<-pam(gower_distance,diss=T,k=6)
```

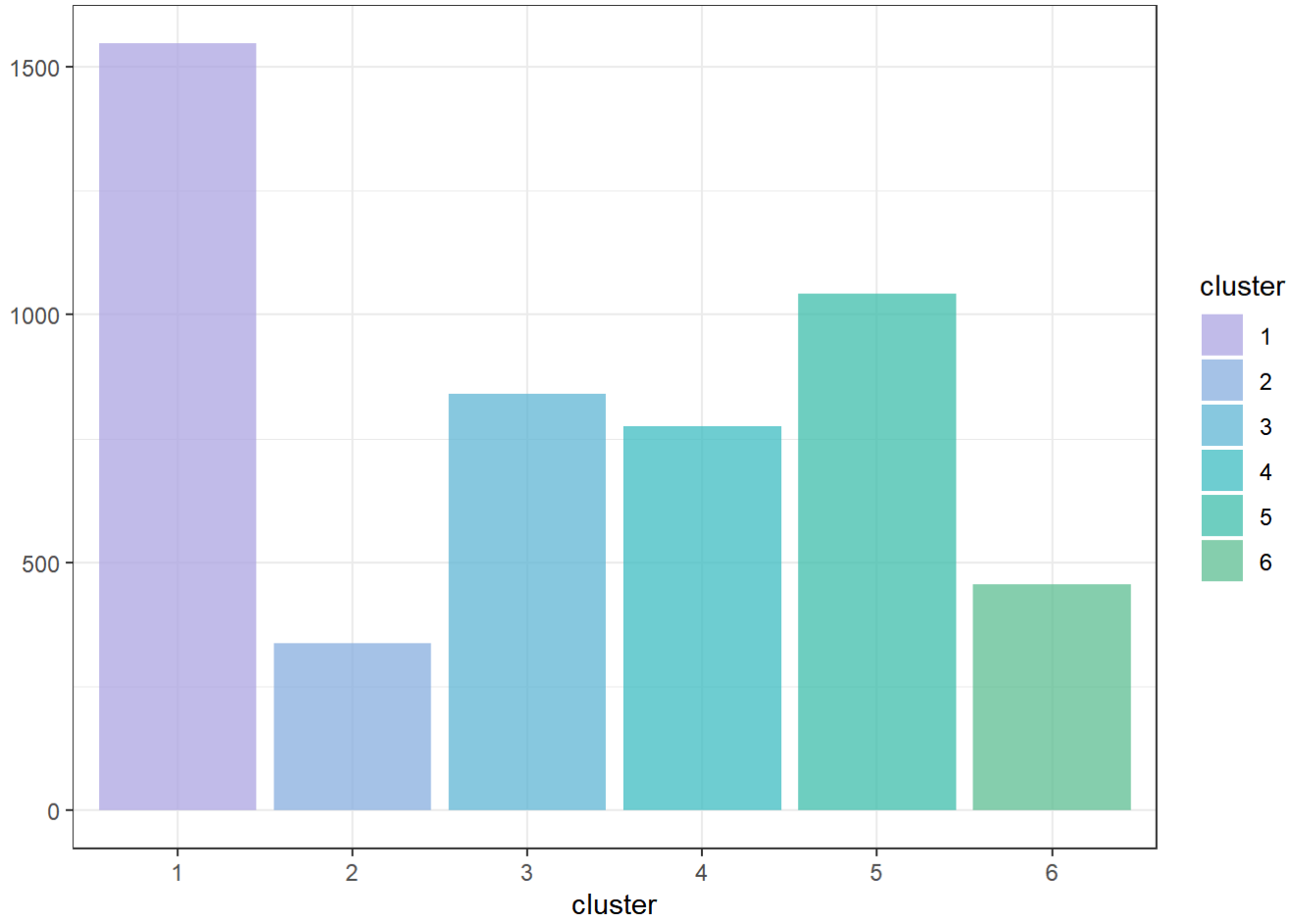
```
result<-data_sample %>%
  mutate(cluster=cluster$clustering) %>%
  group_by(cluster) %>%
  do(summary=summary(.))
```

```
result$summary
```

```
## [[1]]
##  gender    income_total                   income_type    family_size
##  F:1547   Min.   :0.00000   Commercial associate:   0   Min.   :0.0000
##  M:   0   1st Qu.:0.05523   Pensioner           :   0   1st Qu.:0.1667
##           Median :0.07849   State servant       :   0   Median :0.1667
##           Mean   :0.09206   Student             :   0   Mean   :0.2102
##           3rd Qu.:0.11337   Working             :1547   3rd Qu.:0.3333
##           Max.   :0.41860                               Max.   :0.8333
##   begin_month          age          YEARS_EMPLOYED       cluster
##  Min.   :0.0000   Min.   :0.02128   Min.   :0.00000   Min.   :1
##  1st Qu.:0.1833   1st Qu.:0.25532   1st Qu.:0.04651   1st Qu.:1
##  Median :0.4167   Median :0.40426   Median :0.11628   Median :1
##  Mean   :0.4432   Mean   :0.42695   Mean   :0.17349   Mean   :1
##  3rd Qu.:0.6667   3rd Qu.:0.57447   3rd Qu.:0.23256   3rd Qu.:1
##  Max.   :1.0000   Max.   :0.95745   Max.   :0.95349   Max.   :1
##
## [[2]]
##  gender   income_total                   income_type   family_size
##  F:309   Min.   :0.008721   Commercial associate:  0   Min.   :0.0000
##  M: 29   1st Qu.:0.055233   Pensioner           :  0   1st Qu.:0.1667
##          Median :0.098837   State servant       :338   Median :0.1667
##          Mean   :0.108919   Student             :  0   Mean   :0.2244
##          3rd Qu.:0.140262   Working             :  0   3rd Qu.:0.3333
##          Max.   :0.491279                              Max.   :0.8333
##   begin_month          age          YEARS_EMPLOYED      cluster
##  Min.   :0.0000   Min.   :0.04255   Min.   :0.0000   Min.   :2
##  1st Qu.:0.2500   1st Qu.:0.29787   1st Qu.:0.1163   1st Qu.:2
##  Median :0.4833   Median :0.42553   Median :0.2326   Median :2
##  Mean   :0.4867   Mean   :0.45040   Mean   :0.2593   Mean   :2
##  3rd Qu.:0.7292   3rd Qu.:0.59574   3rd Qu.:0.3488   3rd Qu.:2
##  Max.   :1.0000   Max.   :0.89362   Max.   :0.9535   Max.   :2
##
## [[3]]
##  gender   income_total                   income_type   family_size
##  F:732   Min.   :0.001453   Commercial associate:  0   Min.   :0.0000
##  M:109   1st Qu.:0.040698   Pensioner           :841   1st Qu.:0.0000
##          Median :0.069767   State servant       :  0   Median :0.1667
##          Mean   :0.079126   Student             :  0   Mean   :0.1266
##          3rd Qu.:0.098837   Working             :  0   3rd Qu.:0.1667
##          Max.   :0.418605                              Max.   :0.6667
##   begin_month          age          YEARS_EMPLOYED         cluster
##  Min.   :0.0000   Min.   :0.1064   Min.   :0.0000000   Min.   :3
##  1st Qu.:0.2000   1st Qu.:0.7660   1st Qu.:0.0000000   1st Qu.:3
##  Median :0.3833   Median :0.8085   Median :0.0000000   Median :3
##  Mean   :0.4190   Mean   :0.8108   Mean   :0.0005254   Mean   :3
##  3rd Qu.:0.6333   3rd Qu.:0.8723   3rd Qu.:0.0000000   3rd Qu.:3
##  Max.   :1.0000   Max.   :1.0000   Max.   :0.3488372   Max.   :3
##
## [[4]]
##  gender   income_total                   income_type   family_size
##  F:775   Min.   :0.002907   Commercial associate:775   Min.   :0.0000
##  M:  0   1st Qu.:0.069767   Pensioner           :  0   1st Qu.:0.1667
##          Median :0.098837   State servant       :  0   Median :0.1667
##          Mean   :0.110355   Student             :  0   Mean   :0.2009
##          3rd Qu.:0.127907   Working             :  0   3rd Qu.:0.3333
```

```
##          Max.   :1.000000                                  Max.   :0.8333
##   begin_month          age          YEARS_EMPLOYED       cluster
##  Min.   :0.0000   Min.   :0.02128   Min.   :0.00000   Min.   :4
##  1st Qu.:0.2167   1st Qu.:0.23404   1st Qu.:0.04651   1st Qu.:4
##  Median :0.4167   Median :0.40426   Median :0.09302   Median :4
##  Mean   :0.4474   Mean   :0.40382   Mean   :0.13806   Mean   :4
##  3rd Qu.:0.6667   3rd Qu.:0.53191   3rd Qu.:0.18605   3rd Qu.:4
##  Max.   :1.0000   Max.   :0.93617   Max.   :0.97674   Max.   :4
##
## [[5]]
##  gender    income_total                    income_type   family_size
##  F:   0   Min.   :0.01163   Commercial associate:  0   Min.   :0.0000
##  M:1042   1st Qu.:0.06977   Pensioner           :  7   1st Qu.:0.1667
##           Median :0.09884   State servant       : 49   Median :0.1667
##           Mean   :0.11480   Student             :  0   Mean   :0.2271
##           3rd Qu.:0.14244   Working             :986   3rd Qu.:0.3333
##           Max.   :0.56395                              Max.   :0.8333
##   begin_month          age          YEARS_EMPLOYED       cluster
##  Min.   :0.0000   Min.   :0.02128   Min.   :0.00000   Min.   :5
##  1st Qu.:0.1833   1st Qu.:0.19149   1st Qu.:0.04651   1st Qu.:5
##  Median :0.3833   Median :0.34043   Median :0.09302   Median :5
##  Mean   :0.4135   Mean   :0.36834   Mean   :0.13547   Mean   :5
##  3rd Qu.:0.6167   3rd Qu.:0.51064   3rd Qu.:0.18605   3rd Qu.:5
##  Max.   :1.0000   Max.   :0.95745   Max.   :0.86047   Max.   :5
##
## [[6]]
##  gender   income_total                    income_type   family_size
##  F:  0   Min.   :0.04070   Commercial associate:418   Min.   :0.0000
##  M:457   1st Qu.:0.09012   Pensioner           :  7   1st Qu.:0.1667
##          Median :0.12209   State servant       : 32   Median :0.1667
##          Mean   :0.14410   Student             :  0   Mean   :0.2243
##          3rd Qu.:0.15698   Working             :  0   3rd Qu.:0.3333
##          Max.   :0.85465                              Max.   :0.8333
##   begin_month          age         YEARS_EMPLOYED       cluster
##  Min.   :0.0000   Min.   :0.0000   Min.   :0.00000   Min.   :6
##  1st Qu.:0.2667   1st Qu.:0.1702   1st Qu.:0.04651   1st Qu.:6
##  Median :0.4667   Median :0.3191   Median :0.09302   Median :6
##  Mean   :0.4903   Mean   :0.3535   Mean   :0.11419   Mean   :6
##  3rd Qu.:0.7167   3rd Qu.:0.4894   3rd Qu.:0.16279   3rd Qu.:6
##  Max.   :1.0000   Max.   :0.9574   Max.   :0.58140   Max.   :6
```

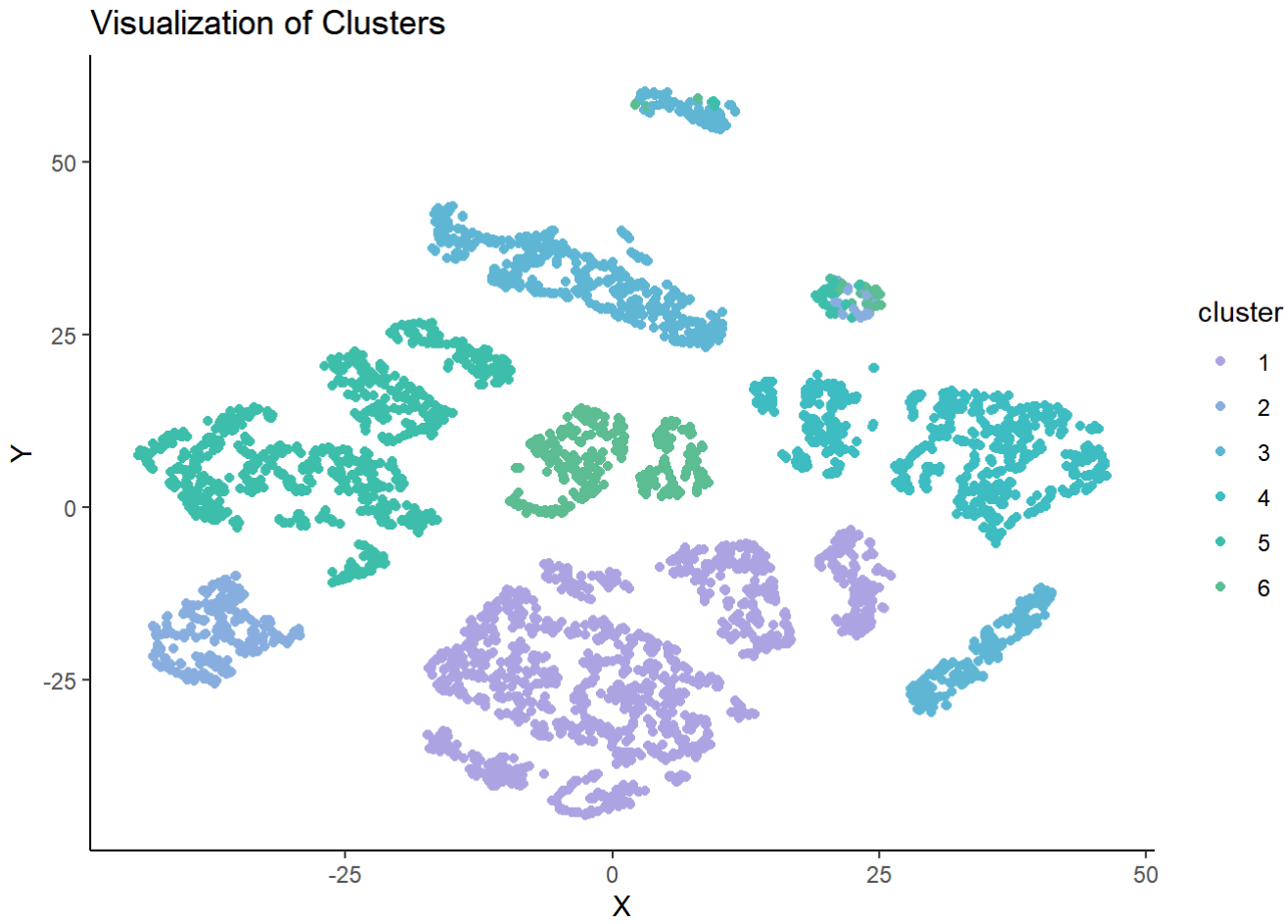
```
h6 <- hcl.colors(6, palette = "Cold")

cluster$clustering %>%
  table %>%
  as.data.frame %>%
  rename('cluster'='.') %>%
  ggplot(aes(x=cluster,y=Freq,fill=cluster))+
  geom_bar(stat='identity',alpha=0.75)+
  scale_fill_manual(values=h6)+
  labs(y=NULL)+
  theme_bw()
```




```
set.seed(2930)
tsne<-Rtsne(gower_distance,is_distance=T)
```

```
visualize<-data.frame(
  X=tsne$Y[,1],
  Y=tsne$Y[,2],
  cluster=cluster$clustering %>% as.factor
)
```

```
visualize %>% ggplot(aes(x=X,y=Y))+
  geom_point(aes(color=cluster))+
  theme_classic()+
  ggtitle("Visualization of Clusters")+
  scale_color_manual(values=h6)
```




```
  theme(plot.title=element_text(hjust=0.5,size=20,face='bold'))
```

```
## List of 1
##  $ plot.title:List of 11
##   ..$ family       : NULL
##   ..$ face         : chr "bold"
##   ..$ colour       : NULL
##   ..$ size         : num 20
##   ..$ hjust        : num 0.5
##   ..$ vjust        : NULL
##   ..$ angle        : NULL
##   ..$ lineheight   : NULL
##   ..$ margin       : NULL
##   ..$ debug        : NULL
##   ..$ inherit.blank: logi FALSE
##   ..- attr(*, "class")= chr [1:2] "element_text" "element"
##  - attr(*, "class")= chr [1:2] "theme" "gg"
##  - attr(*, "complete")= logi FALSE
##  - attr(*, "validate")= logi TRUE
```

Chapter 3 : Supervised Learning (Support Vector Machine, SVM)

```
library(e1071)
```

문제1. 앞의 신용점수 데이터를 그대로 사용하여, 해당 사용자의 성별을 예측하는 모델링을 진행합니다. 아래와 같은 전처리를 진행해주세요.

- Index, occyp_type, FLAG_MOBIL, car, reality, income_type, edu_type, family_type,house_type, work_phone, phone, email 변수 제거
- DAYS_BIRTH에서 출생일을 만나이로 변환 / 고용상태를 판단하는 DAYS_EMPLOYED 변수 변환
- Child_num이 6명 이상인 경우 이상치로 판단하여 제외
- begin_month에 대해 양수로 처리 범주형 변수들에 대해 factor형으로 변환

```
data<-fread('data.csv',na.strings=c('',NA))

data %<>%
  select(-c("index", "occyp_type",'FLAG_MOBIL', 'car', 'reality', 'income_type', 'edu_type', 'f
amily_type','house_type', 'work_phone', 'phone', 'email')) %>%
  filter(!data$child_num>=6) %>%
  select_if(summarise_all(.,n_distinct)!=1) %>%
  mutate(age=((DAYS_BIRTH %>% abs)%/%365)) %>%
  select(-DAYS_BIRTH) %>%
  mutate_at(vars(DAYS_EMPLOYED),
            function(x){ifelse((x)>0,0,x)}) %>%
  mutate(YEARS_EMPLOYED=((DAYS_EMPLOYED %>% abs)%/%365)) %>%
  select(-DAYS_EMPLOYED) %>%
  mutate_at(vars(begin_month),abs) %>%
  mutate_if(summarise_all(.,n_distinct)<=6,as.factor) %>%
  mutate_if(!(summarise_all(.,n_distinct)<=6),as.numeric) %>%
  mutate_at(vars(child_num),as.numeric)
```

문제2. 데이터를 Train set과 valid set으로 구분해주세요. (seed : 2930, 7:3 비율)

```
set.seed(2930)
index<-createDataPartition(data$gender,p=0.7,list=F)
train<-data[index,]
test<-data[-index,]
```

문제4. svm 함수를 이용해 성별을 예측하여 분류하는 모델링을 진행해보세요

```
svm_fit<-svm(gender~.,data=train,type='C-classification',
             kernel='radial',gamma=0.1,cost=10)
```

```
predict_test<-predict(svm_fit,newdata=test)
```

```
cfmatrix<-confusionMatrix(predict_test,test$gender)
cfmatrix
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction    F    M
##          F 4871 1902
##          M  437  725
##
##                Accuracy : 0.7052
##                  95% CI : (0.6951, 0.7152)
##     No Information Rate : 0.6689
##     P-Value [Acc > NIR] : 2.178e-12
##
##                   Kappa : 0.2254
##
##  Mcnemar's Test P-Value : < 2.2e-16
##
##             Sensitivity : 0.9177
##             Specificity : 0.2760
##          Pos Pred Value : 0.7192
##          Neg Pred Value : 0.6239
##              Prevalence : 0.6689
##          Detection Rate : 0.6139
##    Detection Prevalence : 0.8536
##       Balanced Accuracy : 0.5968
##
##        'Positive' Class : F
##
```

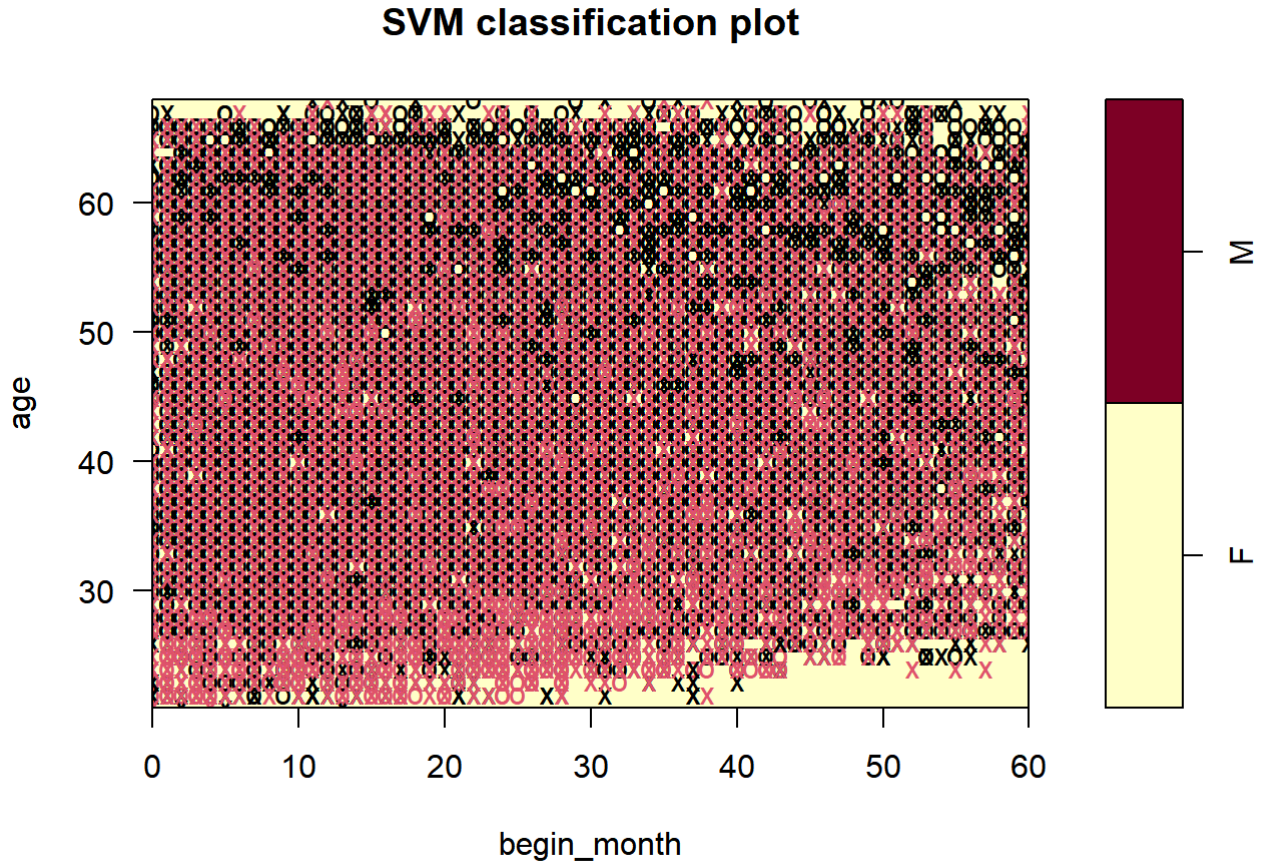
```
obj<-tune(svm,gender~.,data=train,
          ranges = list(cost = c(0.1,1), gamma = c(0.1,1),
                        kernel = c('radial', 'linear')),
          tunecontrol = tune.control(sampling = "fix"))
```

```
print(obj$best.parameters)
```

```
##   cost gamma kernel
## 4    1     1 radial
```

```
model<-svm(gender~.,data=data,type='C-classification',
            kernel='radial',gamma=1,cost=1)
```

```
plot(model,data,age~begin_month)
```


SVM classification plot
age
60
50
40
30
0
10
20
30
40
50
60
begin_month
M
F