finalproject

Seungjun Lee

2021 5 29

```
# origin up은 피피티에 언급안되어 있으나 일단 뺐습니다.
# 이게 생각보다 어떤 변수를 빼야하는지 명확하게 언급이 안되어 있어서 (특히 forcast part) 일단 NA많은 forecast는 전부 뺐습니다.
# 혹시 빼면 안되는 변수가 있는데 뺐으면 알려주세요..!

#channel_sales는 일단 빼서 수치형 자료만 남겼습니다.
# selected_data의 모든 값에 절댓값을 취해주고 나중에 다시 데이터에 channel_sales열을 넣었습니다.
selected_data = data %>% select(-c(id,activity_new,origin_up,forecast_base_bill_ele,forecast_base_bill_year,
                                   forecast_bill_12m,forecast_meter_rent_12m,forecast_cons,forecast_price_energy_p1,
                                   forecast_price_energy_p2,forecast_price_pow_p1,imp_cons ,margin_gross_pow_ele,num_years_antig,
                                   channel_sales))
head(selected_data)
```

```
##   cons_12m cons_gas_12m cons_last_month no_consumption churn days_active
## 1    22034            0            3084              0     0        2224
## 2     4060            0               0              1     0        2511
## 3     7440            0            1062              0     0        1165
## 4  4199490       728810          456462              0     0        2192
## 5    11272            0               0              1     0        2192
## 6   104657            0            6760              0     0        1461
##   days_since_last_modification forecast_cons_12m forecast_cons_year
## 1                         1863            729.06                425
## 2                          732            597.77                  0
## 3                          -68           1311.16               1062
## 4                         1833          11776.27              17393
## 5                         1833           1671.41                  0
## 6                         1096          10378.44               6760
##   forecast_discount_energy has_gas margin_net_pow_ele nb_prod_act net_margin
## 1                        0       0              43.08           1      81.42
## 2                        0       0              24.42           1      61.58
## 3                       30       0              38.58           2      81.61
## 4                        0       1              -2.80           2     897.08
## 5                        0       0              29.76           1     157.99
## 6                        0       0              -4.41           1     700.71
##   pow_max
## 1  17.250
## 2  13.200
## 3  13.856
## 4  33.000
## 5  13.200
## 6  70.000
```

```
summary(selected_data) # 결측치가 별로 없다.
```

```
##     cons_12m         cons_gas_12m     cons_last_month   no_consumption
##  Min.   : -125276   Min.   :  -3037   Min.   : -91386   Min.   :0.0000
##  1st Qu.:    5906   1st Qu.:      0   1st Qu.:      0   1st Qu.:0.0000
##  Median :   15332   Median :      0   Median :    900   Median :0.0000
##  Mean   :  194846   Mean   :  31920   Mean   :  19466   Mean   :0.3246
##  3rd Qu.:   50222   3rd Qu.:      0   3rd Qu.:   4127   3rd Qu.:1.0000
##  Max.   :16097108   Max.   :4188440   Max.   :4538720   Max.   :1.0000
##
##      churn          days_active     days_since_last_modification
##  Min.   :0.00000   Min.   :-41444   Min.   :-42398
##  1st Qu.:0.00000   1st Qu.:  1461   1st Qu.:    55
##  Median :0.00000   Median :  1833   Median :   732
##  Mean   :0.09912   Mean   :  2013   Mean   :  1246
##  3rd Qu.:0.00000   3rd Qu.:  2401   3rd Qu.:  1827
##  Max.   :1.00000   Max.   :  5925   Max.   : 42373
##
##  forecast_cons_12m  forecast_cons_year forecast_discount_energy
##  Min.   :-16689.3   Min.   :-85627     Min.   : 0.000
##  1st Qu.:   513.2   1st Qu.:     0     1st Qu.: 0.000
##  Median :  1179.2   Median :   378     Median : 0.000
##  Mean   :  2370.5   Mean   :  1907     Mean   : 0.984
##  3rd Qu.:  2691.7   3rd Qu.:  1994     3rd Qu.: 0.000
##  Max.   :103801.9   Max.   :175375     Max.   :50.000
##
##     has_gas       margin_net_pow_ele  nb_prod_act       net_margin
##  Min.   :0.0000   Min.   :-615.66    Min.   : 1.000   Min.   :-4148.99
##  1st Qu.:0.0000   1st Qu.:  11.95    1st Qu.: 1.000   1st Qu.:   51.97
##  Median :0.0000   Median :  20.97    Median : 1.000   Median :  119.68
##  Mean   :0.1841   Mean   :  21.46    Mean   : 1.348   Mean   :  217.99
##  3rd Qu.:0.0000   3rd Qu.:  29.64    3rd Qu.: 1.000   3rd Qu.:  275.81
##  Max.   :1.0000   Max.   : 374.64    Max.   :32.000   Max.   :24570.65
##                   NA's   :9                           NA's   :11
##     pow_max
##  Min.   :  1.00
##  1st Qu.: 12.50
##  Median : 13.86
##  Mean   : 20.58
##  3rd Qu.: 19.80
##  Max.   :500.00
##
```

```
# 모든 값에 절댓값 취해주기
selected_data = abs(selected_data)
dim(selected_data)
```

```
## [1] 16092    15
```

```
# channel_sales 값이 일단 너무 길어서 factor로만 바꿔봤습니다.
# factor 자체에 order가 있는건 아니고 그냥 일단 임의로 바꿨습니다.

table(data$channel_sales) # 8개의 종류 확인
```

```
##
##                                epumfxlbckeskwekxbiuasklxalciiuu
##                             4216                                4
## ewpakwlliwisiwduibdlfmalxowmwpci fixdbufsefwooaasfcxdxadsiekoceaa
##                              966                                2
## foosdfpfkusacimwkcsosbicdxkicaua lmkebamcaaclubfxadlmueccxoimlema
##                             7375                             2073
## sddiedcslfslkckwlfkdpoeeailfpeds usilxuppasemubllopkaafesmlibmsdf
##                               12                             1444
```

```
levels(data$channel_sales) = as.factor(1:8)
levels(data$channel_sales)
```

```
## [1] "1" "2" "3" "4" "5" "6" "7" "8"
```

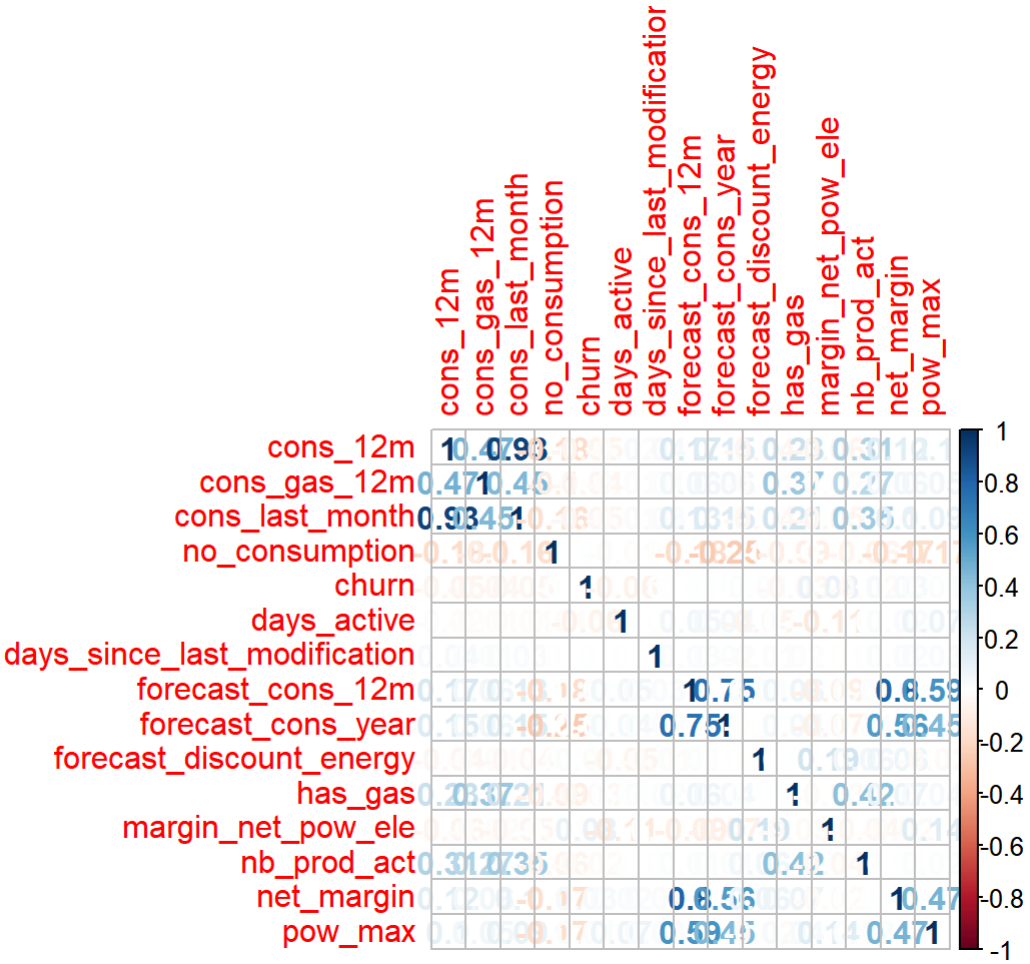
```
# channels_sales와 합친 후 결측치 전부 제거
selected_data = na.omit(cbind(selected_data,channel_sales = data$channel_sales))
dim(selected_data) # 11개의 데이터가 제거되었다.
```

```
## [1] 16081    16
```

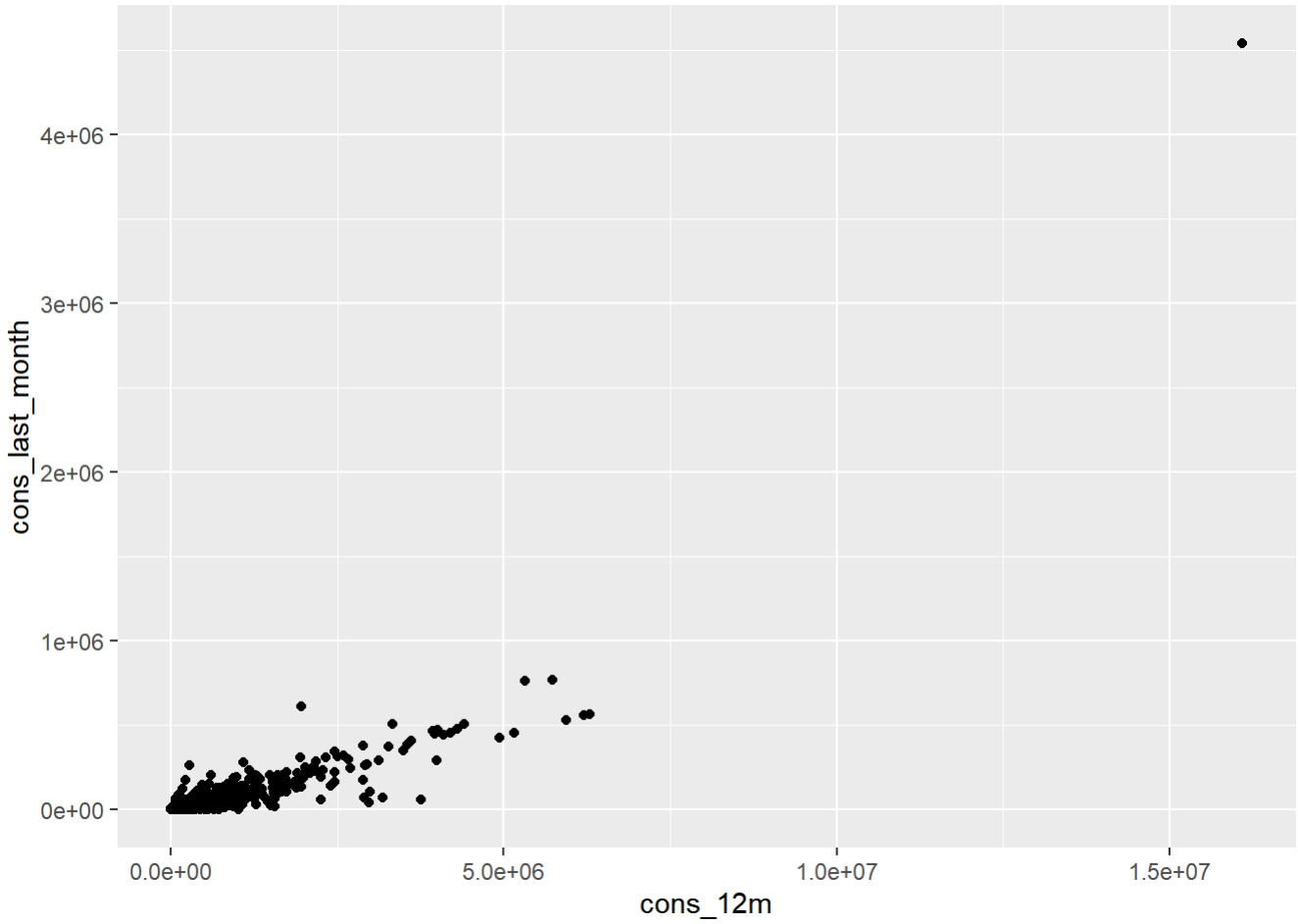
```
summary(selected_data)
```

```
##     cons_12m         cons_gas_12m     cons_last_month   no_consumption
##  Min.   :       0   Min.   :      0   Min.   :      0   Min.   :0.0000
##  1st Qu.:    5923   1st Qu.:      0   1st Qu.:      0   1st Qu.:0.0000
##  Median :   15359   Median :      0   Median :    916   Median :0.0000
##  Mean   :  193474   Mean   :  31919   Mean   :  19546   Mean   :0.3247
##  3rd Qu.:   50181   3rd Qu.:      0   3rd Qu.:   4159   3rd Qu.:1.0000
##  Max.   :16097108   Max.   :4188440   Max.   :4538720   Max.   :1.0000
##
##      churn          days_active    days_since_last_modification
##  Min.   :0.00000   Min.   :  365   Min.   :    0
##  1st Qu.:0.00000   1st Qu.: 1461   1st Qu.:  146
##  Median :0.00000   Median : 1831   Median :  732
##  Mean   :0.09919   Mean   : 2023   Mean   : 1489
##  3rd Qu.:0.00000   3rd Qu.: 2401   3rd Qu.: 1827
##  Max.   :1.00000   Max.   :41444   Max.   :42398
##
##  forecast_cons_12m  forecast_cons_year forecast_discount_energy
##  Min.   :     0.0   Min.   :     0     Min.   : 0.0000
##  1st Qu.:   514.6   1st Qu.:     0     1st Qu.: 0.0000
##  Median :  1182.0   Median :   385     Median : 0.0000
##  Mean   :  2379.1   Mean   :  1929     Mean   : 0.9833
##  3rd Qu.:  2693.2   3rd Qu.:  2000     3rd Qu.: 0.0000
##  Max.   :103801.9   Max.   :175375     Max.   :50.0000
##
##     has_gas       margin_net_pow_ele  nb_prod_act       net_margin
##  Min.   :0.0000   Min.   :  0.00     Min.   : 1.000   Min.   :    0.00
##  1st Qu.:0.0000   1st Qu.: 12.36     1st Qu.: 1.000   1st Qu.:   52.99
##  Median :0.0000   Median : 21.11     Median : 1.000   Median :  121.20
##  Mean   :0.1841   Mean   : 24.18     Mean   : 1.348   Mean   :  222.93
##  3rd Qu.:0.0000   3rd Qu.: 29.76     3rd Qu.: 1.000   3rd Qu.:  277.40
##  Max.   :1.0000   Max.   :615.66     Max.   :32.000   Max.   :24570.65
##
##     pow_max       channel_sales
##  Min.   :  1.00   5      :7375
##  1st Qu.: 12.50   1      :4211
##  Median : 13.86   6      :2068
##  Mean   : 20.57   8      :1443
##  3rd Qu.: 19.80   3      : 966
##  Max.   :500.00   7      :  12
##                   (Other):   6
```

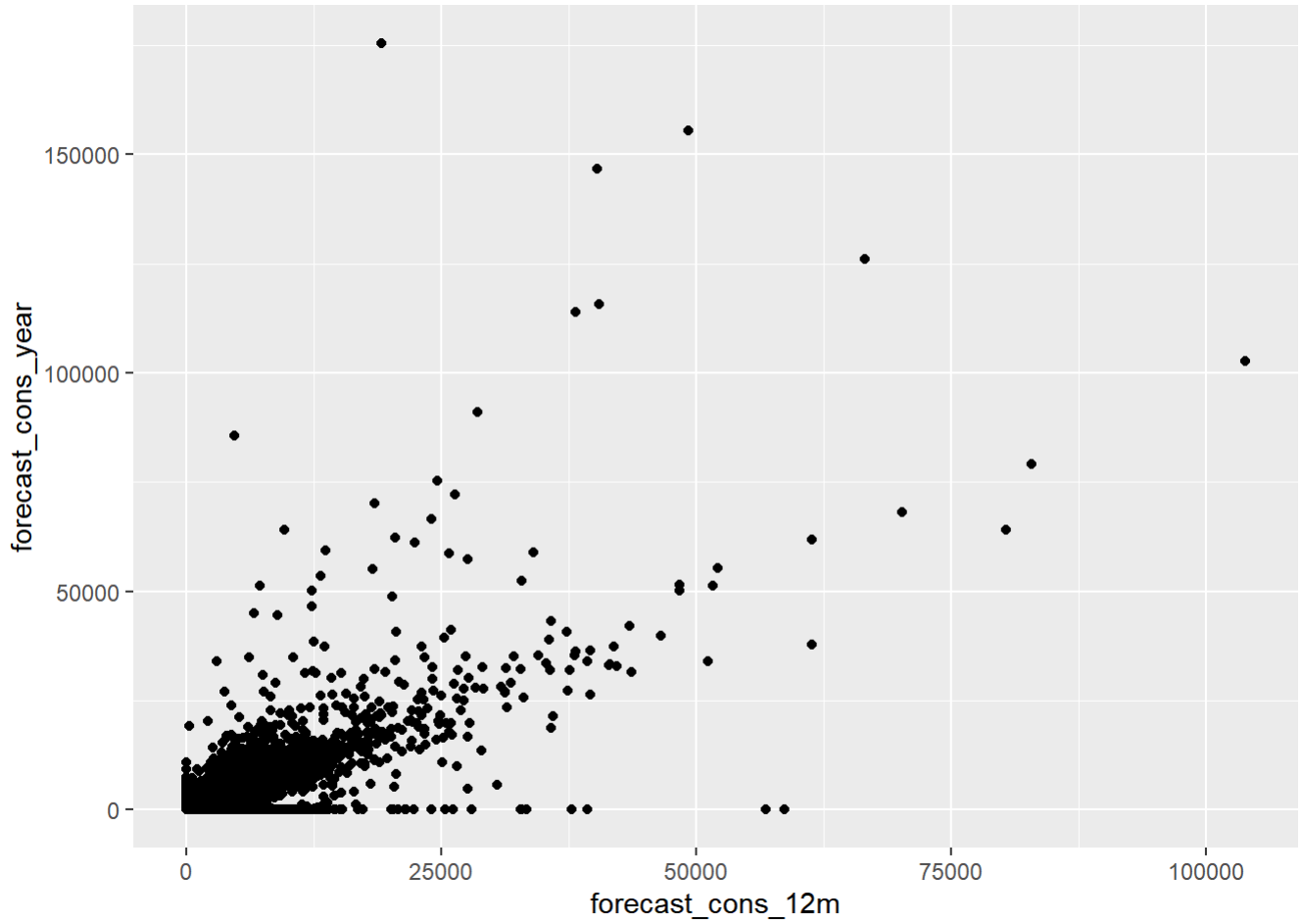
```
numeric = selected_data %>% select(-channel_sales)
cor = cor(numeric)
corrplot(cor,method='number')
```



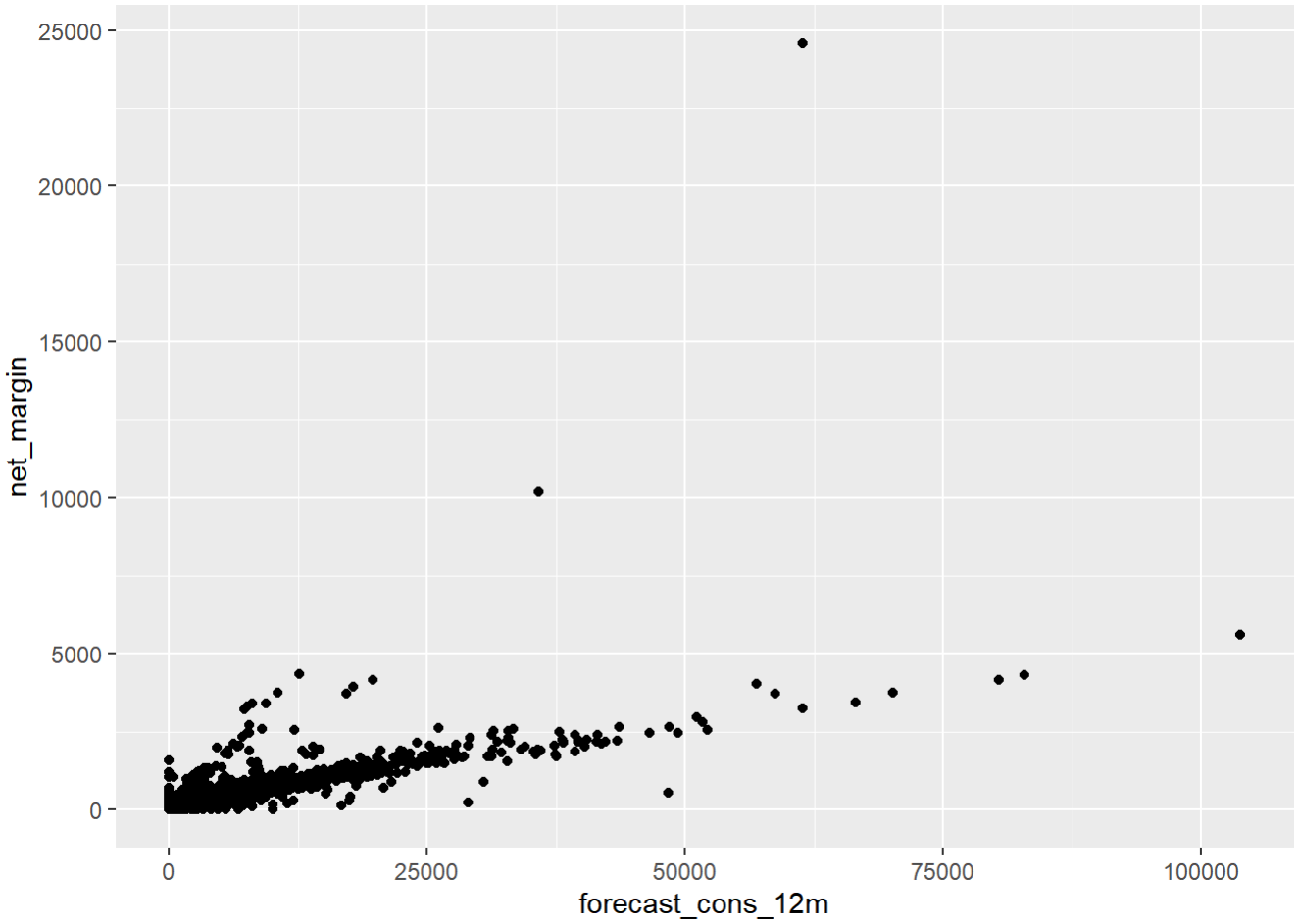

```
ggplot(selected_data, aes(x=cons_12m,y=cons_last_month))+
        geom_point() # 상관계수 : 0.93
```




```
ggplot(selected_data, aes(x=forecast_cons_12m,y=forecast_cons_year))+
  geom_point() # 상관계수 : 0.75
```




```
ggplot(selected_data, aes(x=forecast_cons_12m,y=net_margin))+
  geom_point() # 상관계수 : 0.8
```


25000
20000
15000
10000
5000
0
net_margin
0
25000
50000
75000
100000
forecast_cons_12m