Prudential data

#### ***Lee-sohyun, Kwon-Yunhwan, Kim-jaehoon, Lee-jaehoo,***

#### ***2017-04-22***

prudential insurance

Prudential회사의 보험 데이터이다. 고객들의 **Response**를 예측하는 것이 이 데이터 분류의 목적이며, Response가 큰게 좋은 것인지, 나쁜 것인지는 알수 없다. 이 데이터의 특징은 몇몇 데이터가 정규화로 되어 있음

데이터 변수 설명

|  |  |
| --- | --- |
| 변수 | 설명 |
| **Id** | 회원들의 고유한 id |
| **Product\_Info\_1-7** | 보험 상품과 관련된 변수 |
| **Ins\_Age** | 회원들의 나이를 정규화해 놓은 변수 |
| **Ht** | 회원들의 키를 정규화한 변수 |
| **Wt** | 회원들의 몸무게를 정규화한 변수 |
| **BMI** | 회원들의 BMI를 정규화한 변수 |
| **Employment\_Info\_1-6** | 회원들의 과거 일자리와 관련된 변수 |
| **InsuredInfo\_1-6** | 회원들에 대하 제공된 정보 변수를 정규화 시켜놓은 것 |
| **Insurance\_History\_1-9** | 회원들이 보험 이력에 대한 변수 |
| **Family\_Hist\_1-5** | 회원들 가족들의 이력과 관련된 정규화 변수 |
| **Medical\_History\_1-41** | 회원들의 의료 이력을 정규화한 변수 |
| **Medical\_Keyword\_1-48** | 회원들과 관련된 의료 키워드의 존재/부재와 관련된 더미 변수 |
| **Response** | 회원들과 관련된 마지막 결정 변수, ordinal 변수 |

데이터 변수 요약

train <- fread("train.csv")  
test <- fread("test.csv")  
head(train)

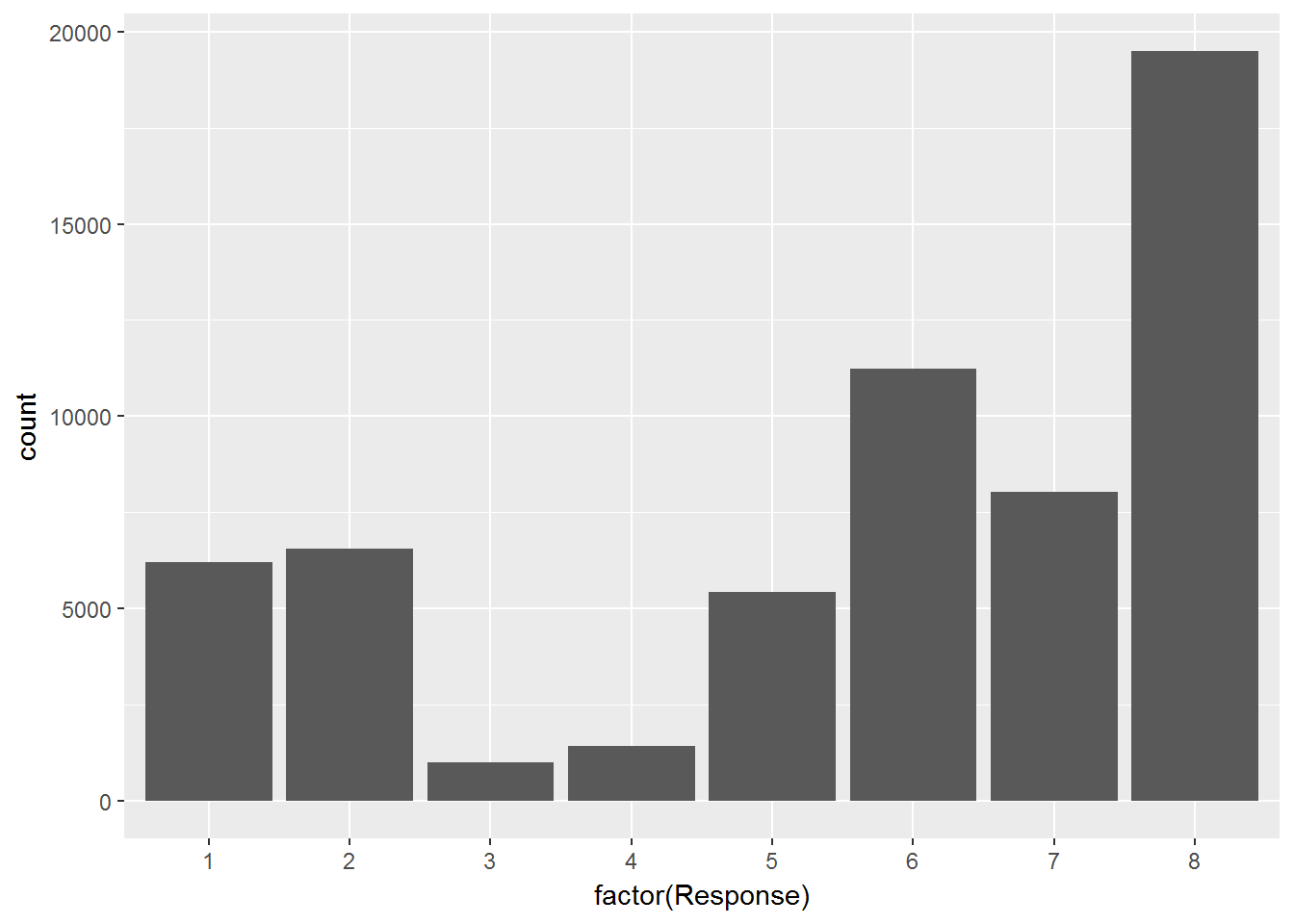
## Id Product\_Info\_1 Product\_Info\_2 Product\_Info\_3 Product\_Info\_4  
## 1: 2 1 D3 10 0.07692308  
## 2: 5 1 A1 26 0.07692308  
## 3: 6 1 E1 26 0.07692308  
## 4: 7 1 D4 10 0.48717949  
## 5: 8 1 D2 26 0.23076923  
## 6: 10 1 D2 26 0.23076923  
## Product\_Info\_5 Product\_Info\_6 Product\_Info\_7 Ins\_Age Ht  
## 1: 2 1 1 0.64179104 0.5818182  
## 2: 2 3 1 0.05970149 0.6000000  
## 3: 2 3 1 0.02985075 0.7454545  
## 4: 2 3 1 0.16417910 0.6727273  
## 5: 2 3 1 0.41791045 0.6545455  
## 6: 3 1 1 0.50746269 0.8363636  
## Wt BMI Employment\_Info\_1 Employment\_Info\_2  
## 1: 0.1485356 0.3230080 0.028 12  
## 2: 0.1317992 0.2722877 0.000 1  
## 3: 0.2887029 0.4287804 0.030 9  
## 4: 0.2050209 0.3524377 0.042 9  
## 5: 0.2343096 0.4240456 0.027 9  
## 6: 0.2991632 0.3648867 0.325 15  
## Employment\_Info\_3 Employment\_Info\_4 Employment\_Info\_5 Employment\_Info\_6  
## 1: 1 0 3 NA  
## 2: 3 0 2 0.0018  
## 3: 1 0 2 0.0300  
## 4: 1 0 3 0.2000  
## 5: 1 0 2 0.0500  
## 6: 1 0 2 1.0000  
## InsuredInfo\_1 InsuredInfo\_2 InsuredInfo\_3 InsuredInfo\_4 InsuredInfo\_5  
## 1: 1 2 6 3 1  
## 2: 1 2 6 3 1  
## 3: 1 2 8 3 1  
## 4: 2 2 8 3 1  
## 5: 1 2 6 3 1  
## 6: 1 2 8 3 1  
## InsuredInfo\_6 InsuredInfo\_7 Insurance\_History\_1 Insurance\_History\_2  
## 1: 2 1 1 1  
## 2: 2 1 2 1  
## 3: 1 1 2 1  
## 4: 2 1 2 1  
## 5: 2 1 2 1  
## 6: 1 1 2 1  
## Insurance\_History\_3 Insurance\_History\_4 Insurance\_History\_5  
## 1: 3 1 0.000666667  
## 2: 3 1 0.000133333  
## 3: 1 3 NA  
## 4: 1 3 NA  
## 5: 1 3 NA  
## 6: 3 2 0.005000000  
## Insurance\_History\_7 Insurance\_History\_8 Insurance\_History\_9  
## 1: 1 1 2  
## 2: 1 3 2  
## 3: 3 2 3  
## 4: 3 2 3  
## 5: 3 2 3  
## 6: 1 3 2  
## Family\_Hist\_1 Family\_Hist\_2 Family\_Hist\_3 Family\_Hist\_4 Family\_Hist\_5  
## 1: 2 NA 0.5980392 NA 0.5267857  
## 2: 2 0.1884058 NA 0.08450704 NA  
## 3: 3 0.3043478 NA 0.22535211 NA  
## 4: 3 0.4202899 NA 0.35211268 NA  
## 5: 2 0.4637681 NA 0.40845070 NA  
## 6: 2 NA 0.2941176 0.50704225 NA  
## Medical\_History\_1 Medical\_History\_2 Medical\_History\_3 Medical\_History\_4  
## 1: 4 112 2 1  
## 2: 5 412 2 1  
## 3: 10 3 2 2  
## 4: 0 350 2 2  
## 5: NA 162 2 2  
## 6: 6 491 2 2  
## Medical\_History\_5 Medical\_History\_6 Medical\_History\_7 Medical\_History\_8  
## 1: 1 3 2 2  
## 2: 1 3 2 2  
## 3: 1 3 2 2  
## 4: 1 3 2 2  
## 5: 1 3 2 2  
## 6: 1 3 2 2  
## Medical\_History\_9 Medical\_History\_10 Medical\_History\_11  
## 1: 1 NA 3  
## 2: 1 NA 3  
## 3: 2 NA 3  
## 4: 2 NA 3  
## 5: 2 NA 3  
## 6: 2 NA 3  
## Medical\_History\_12 Medical\_History\_13 Medical\_History\_14  
## 1: 2 3 3  
## 2: 2 3 3  
## 3: 2 3 3  
## 4: 2 3 3  
## 5: 2 3 3  
## 6: 2 3 3  
## Medical\_History\_15 Medical\_History\_16 Medical\_History\_17  
## 1: 240 3 3  
## 2: 0 1 3  
## 3: NA 1 3  
## 4: NA 1 3  
## 5: NA 1 3  
## 6: NA 1 3  
## Medical\_History\_18 Medical\_History\_19 Medical\_History\_20  
## 1: 1 1 2  
## 2: 1 1 2  
## 3: 1 1 2  
## 4: 1 1 2  
## 5: 1 1 2  
## 6: 2 1 2  
## Medical\_History\_21 Medical\_History\_22 Medical\_History\_23  
## 1: 1 2 3  
## 2: 1 2 3  
## 3: 1 2 3  
## 4: 2 2 3  
## 5: 1 2 3  
## 6: 2 2 3  
## Medical\_History\_24 Medical\_History\_25 Medical\_History\_26  
## 1: NA 1 3  
## 2: NA 1 3  
## 3: NA 2 2  
## 4: NA 1 3  
## 5: NA 2 2  
## 6: NA 1 3  
## Medical\_History\_27 Medical\_History\_28 Medical\_History\_29  
## 1: 3 1 3  
## 2: 3 1 3  
## 3: 3 1 3  
## 4: 3 1 3  
## 5: 3 1 3  
## 6: 3 1 3  
## Medical\_History\_30 Medical\_History\_31 Medical\_History\_32  
## 1: 2 3 NA  
## 2: 2 3 NA  
## 3: 2 3 NA  
## 4: 2 3 NA  
## 5: 2 3 NA  
## 6: 2 3 NA  
## Medical\_History\_33 Medical\_History\_34 Medical\_History\_35  
## 1: 1 3 1  
## 2: 3 1 1  
## 3: 3 3 1  
## 4: 3 3 1  
## 5: 3 3 1  
## 6: 3 1 1  
## Medical\_History\_36 Medical\_History\_37 Medical\_History\_38  
## 1: 2 2 1  
## 2: 2 2 1  
## 3: 3 2 1  
## 4: 2 2 1  
## 5: 3 2 1  
## 6: 2 2 1  
## Medical\_History\_39 Medical\_History\_40 Medical\_History\_41  
## 1: 3 3 3  
## 2: 3 3 1  
## 3: 3 3 1  
## 4: 3 3 1  
## 5: 3 3 1  
## 6: 3 3 3  
## Medical\_Keyword\_1 Medical\_Keyword\_2 Medical\_Keyword\_3 Medical\_Keyword\_4  
## 1: 0 0 0 0  
## 2: 0 0 0 0  
## 3: 0 0 0 0  
## 4: 0 0 0 0  
## 5: 0 0 0 0  
## 6: 0 0 0 0  
## Medical\_Keyword\_5 Medical\_Keyword\_6 Medical\_Keyword\_7 Medical\_Keyword\_8  
## 1: 0 0 0 0  
## 2: 0 0 0 0  
## 3: 0 0 0 0  
## 4: 0 0 0 0  
## 5: 0 0 0 0  
## 6: 0 0 0 0  
## Medical\_Keyword\_9 Medical\_Keyword\_10 Medical\_Keyword\_11  
## 1: 0 0 0  
## 2: 0 0 0  
## 3: 0 0 0  
## 4: 0 0 0  
## 5: 0 0 0  
## 6: 0 0 0  
## Medical\_Keyword\_12 Medical\_Keyword\_13 Medical\_Keyword\_14  
## 1: 0 0 0  
## 2: 0 0 0  
## 3: 0 0 0  
## 4: 0 0 0  
## 5: 0 0 0  
## 6: 0 0 0  
## Medical\_Keyword\_15 Medical\_Keyword\_16 Medical\_Keyword\_17  
## 1: 0 0 0  
## 2: 0 0 0  
## 3: 0 0 0  
## 4: 0 0 0  
## 5: 0 0 0  
## 6: 0 0 0  
## Medical\_Keyword\_18 Medical\_Keyword\_19 Medical\_Keyword\_20  
## 1: 0 0 0  
## 2: 0 0 0  
## 3: 0 0 0  
## 4: 0 0 0  
## 5: 0 0 0  
## 6: 0 0 0  
## Medical\_Keyword\_21 Medical\_Keyword\_22 Medical\_Keyword\_23  
## 1: 0 0 0  
## 2: 0 0 0  
## 3: 0 0 0  
## 4: 0 0 0  
## 5: 0 0 0  
## 6: 0 1 0  
## Medical\_Keyword\_24 Medical\_Keyword\_25 Medical\_Keyword\_26  
## 1: 0 0 0  
## 2: 0 0 0  
## 3: 0 0 0  
## 4: 0 0 0  
## 5: 0 0 0  
## 6: 0 0 0  
## Medical\_Keyword\_27 Medical\_Keyword\_28 Medical\_Keyword\_29  
## 1: 0 0 0  
## 2: 0 0 0  
## 3: 0 0 0  
## 4: 0 0 0  
## 5: 0 0 0  
## 6: 0 0 0  
## Medical\_Keyword\_30 Medical\_Keyword\_31 Medical\_Keyword\_32  
## 1: 0 0 0  
## 2: 0 0 0  
## 3: 0 0 0  
## 4: 0 0 1  
## 5: 0 0 0  
## 6: 0 0 0  
## Medical\_Keyword\_33 Medical\_Keyword\_34 Medical\_Keyword\_35  
## 1: 0 0 0  
## 2: 0 0 0  
## 3: 0 0 0  
## 4: 0 0 0  
## 5: 0 0 0  
## 6: 0 1 0  
## Medical\_Keyword\_36 Medical\_Keyword\_37 Medical\_Keyword\_38  
## 1: 0 0 0  
## 2: 0 0 0  
## 3: 0 0 0  
## 4: 0 0 0  
## 5: 0 0 0  
## 6: 0 0 0  
## Medical\_Keyword\_39 Medical\_Keyword\_40 Medical\_Keyword\_41  
## 1: 0 0 0  
## 2: 0 0 0  
## 3: 0 0 0  
## 4: 0 0 0  
## 5: 0 0 0  
## 6: 0 0 0  
## Medical\_Keyword\_42 Medical\_Keyword\_43 Medical\_Keyword\_44  
## 1: 0 0 0  
## 2: 0 0 0  
## 3: 0 0 0  
## 4: 0 0 0  
## 5: 0 0 0  
## 6: 0 0 0  
## Medical\_Keyword\_45 Medical\_Keyword\_46 Medical\_Keyword\_47  
## 1: 0 0 0  
## 2: 0 0 0  
## 3: 0 0 0  
## 4: 0 0 0  
## 5: 0 0 0  
## 6: 0 0 0  
## Medical\_Keyword\_48 Response  
## 1: 0 8  
## 2: 0 4  
## 3: 0 8  
## 4: 0 8  
## 5: 0 8  
## 6: 0 8

summary(train)

## Id Product\_Info\_1 Product\_Info\_2 Product\_Info\_3   
## Min. : 2 Min. :1.000 Length:59381 Min. : 1.00   
## 1st Qu.:19780 1st Qu.:1.000 Class :character 1st Qu.:26.00   
## Median :39487 Median :1.000 Mode :character Median :26.00   
## Mean :39507 Mean :1.026 Mean :24.42   
## 3rd Qu.:59211 3rd Qu.:1.000 3rd Qu.:26.00   
## Max. :79146 Max. :2.000 Max. :38.00   
##   
## Product\_Info\_4 Product\_Info\_5 Product\_Info\_6 Product\_Info\_7   
## Min. :0.00000 Min. :2.000 Min. :1.000 Min. :1.000   
## 1st Qu.:0.07692 1st Qu.:2.000 1st Qu.:3.000 1st Qu.:1.000   
## Median :0.23077 Median :2.000 Median :3.000 Median :1.000   
## Mean :0.32895 Mean :2.007 Mean :2.674 Mean :1.044   
## 3rd Qu.:0.48718 3rd Qu.:2.000 3rd Qu.:3.000 3rd Qu.:1.000   
## Max. :1.00000 Max. :3.000 Max. :3.000 Max. :3.000   
##   
## Ins\_Age Ht Wt BMI   
## Min. :0.0000 Min. :0.0000 Min. :0.0000 Min. :0.0000   
## 1st Qu.:0.2388 1st Qu.:0.6545 1st Qu.:0.2259 1st Qu.:0.3855   
## Median :0.4030 Median :0.7091 Median :0.2887 Median :0.4513   
## Mean :0.4056 Mean :0.7073 Mean :0.2926 Mean :0.4695   
## 3rd Qu.:0.5672 3rd Qu.:0.7636 3rd Qu.:0.3452 3rd Qu.:0.5329   
## Max. :1.0000 Max. :1.0000 Max. :1.0000 Max. :1.0000   
##   
## Employment\_Info\_1 Employment\_Info\_2 Employment\_Info\_3 Employment\_Info\_4  
## Min. :0.00000 Min. : 1.000 Min. :1.000 Min. :0.000   
## 1st Qu.:0.03500 1st Qu.: 9.000 1st Qu.:1.000 1st Qu.:0.000   
## Median :0.06000 Median : 9.000 Median :1.000 Median :0.000   
## Mean :0.07758 Mean : 8.642 Mean :1.301 Mean :0.006   
## 3rd Qu.:0.10000 3rd Qu.: 9.000 3rd Qu.:1.000 3rd Qu.:0.000   
## Max. :1.00000 Max. :38.000 Max. :3.000 Max. :1.000   
## NA's :19 NA's :6779   
## Employment\_Info\_5 Employment\_Info\_6 InsuredInfo\_1 InsuredInfo\_2   
## Min. :2.000 Min. :0.000 Min. :1.000 Min. :2.000   
## 1st Qu.:2.000 1st Qu.:0.060 1st Qu.:1.000 1st Qu.:2.000   
## Median :2.000 Median :0.250 Median :1.000 Median :2.000   
## Mean :2.143 Mean :0.361 Mean :1.209 Mean :2.007   
## 3rd Qu.:2.000 3rd Qu.:0.550 3rd Qu.:1.000 3rd Qu.:2.000   
## Max. :3.000 Max. :1.000 Max. :3.000 Max. :3.000   
## NA's :10854   
## InsuredInfo\_3 InsuredInfo\_4 InsuredInfo\_5 InsuredInfo\_6   
## Min. : 1.000 Min. :2.000 Min. :1.000 Min. :1.000   
## 1st Qu.: 3.000 1st Qu.:3.000 1st Qu.:1.000 1st Qu.:1.000   
## Median : 6.000 Median :3.000 Median :1.000 Median :1.000   
## Mean : 5.836 Mean :2.884 Mean :1.027 Mean :1.409   
## 3rd Qu.: 8.000 3rd Qu.:3.000 3rd Qu.:1.000 3rd Qu.:2.000   
## Max. :11.000 Max. :3.000 Max. :3.000 Max. :2.000   
##   
## InsuredInfo\_7 Insurance\_History\_1 Insurance\_History\_2  
## Min. :1.000 Min. :1.000 Min. :1.000   
## 1st Qu.:1.000 1st Qu.:1.000 1st Qu.:1.000   
## Median :1.000 Median :2.000 Median :1.000   
## Mean :1.039 Mean :1.728 Mean :1.056   
## 3rd Qu.:1.000 3rd Qu.:2.000 3rd Qu.:1.000   
## Max. :3.000 Max. :2.000 Max. :3.000   
##   
## Insurance\_History\_3 Insurance\_History\_4 Insurance\_History\_5  
## Min. :1.000 Min. :1.000 Min. :0.000   
## 1st Qu.:1.000 1st Qu.:1.000 1st Qu.:0.000   
## Median :3.000 Median :2.000 Median :0.001   
## Mean :2.147 Mean :1.959 Mean :0.002   
## 3rd Qu.:3.000 3rd Qu.:3.000 3rd Qu.:0.002   
## Max. :3.000 Max. :3.000 Max. :1.000   
## NA's :25396   
## Insurance\_History\_7 Insurance\_History\_8 Insurance\_History\_9  
## Min. :1.000 Min. :1.000 Min. :1.000   
## 1st Qu.:1.000 1st Qu.:1.000 1st Qu.:2.000   
## Median :1.000 Median :2.000 Median :2.000   
## Mean :1.902 Mean :2.048 Mean :2.419   
## 3rd Qu.:3.000 3rd Qu.:3.000 3rd Qu.:3.000   
## Max. :3.000 Max. :3.000 Max. :3.000   
##   
## Family\_Hist\_1 Family\_Hist\_2 Family\_Hist\_3 Family\_Hist\_4   
## Min. :1.000 Min. :0.000 Min. :0.00 Min. :0.000   
## 1st Qu.:2.000 1st Qu.:0.362 1st Qu.:0.40 1st Qu.:0.324   
## Median :3.000 Median :0.464 Median :0.52 Median :0.423   
## Mean :2.686 Mean :0.475 Mean :0.50 Mean :0.445   
## 3rd Qu.:3.000 3rd Qu.:0.580 3rd Qu.:0.60 3rd Qu.:0.563   
## Max. :3.000 Max. :1.000 Max. :1.00 Max. :0.944   
## NA's :28656 NA's :34241 NA's :19184   
## Family\_Hist\_5 Medical\_History\_1 Medical\_History\_2 Medical\_History\_3  
## Min. :0.00 Min. : 0.000 Min. : 1 Min. :1.000   
## 1st Qu.:0.40 1st Qu.: 2.000 1st Qu.:112 1st Qu.:2.000   
## Median :0.51 Median : 4.000 Median :162 Median :2.000   
## Mean :0.48 Mean : 7.962 Mean :254 Mean :2.102   
## 3rd Qu.:0.58 3rd Qu.: 9.000 3rd Qu.:418 3rd Qu.:2.000   
## Max. :1.00 Max. :240.000 Max. :648 Max. :3.000   
## NA's :41811 NA's :8889   
## Medical\_History\_4 Medical\_History\_5 Medical\_History\_6 Medical\_History\_7  
## Min. :1.000 Min. :1.000 Min. :1.00 Min. :1.000   
## 1st Qu.:1.000 1st Qu.:1.000 1st Qu.:3.00 1st Qu.:2.000   
## Median :2.000 Median :1.000 Median :3.00 Median :2.000   
## Mean :1.655 Mean :1.007 Mean :2.89 Mean :2.012   
## 3rd Qu.:2.000 3rd Qu.:1.000 3rd Qu.:3.00 3rd Qu.:2.000   
## Max. :2.000 Max. :3.000 Max. :3.00 Max. :3.000   
##   
## Medical\_History\_8 Medical\_History\_9 Medical\_History\_10 Medical\_History\_11  
## Min. :1.000 Min. :1.00 Min. : 0.0 Min. :1.000   
## 1st Qu.:2.000 1st Qu.:2.00 1st Qu.: 8.0 1st Qu.:3.000   
## Median :2.000 Median :2.00 Median :229.0 Median :3.000   
## Mean :2.044 Mean :1.77 Mean :141.1 Mean :2.994   
## 3rd Qu.:2.000 3rd Qu.:2.00 3rd Qu.:240.0 3rd Qu.:3.000   
## Max. :3.000 Max. :3.00 Max. :240.0 Max. :3.000   
## NA's :58824   
## Medical\_History\_12 Medical\_History\_13 Medical\_History\_14  
## Min. :1.000 Min. :1.000 Min. :1.000   
## 1st Qu.:2.000 1st Qu.:3.000 1st Qu.:3.000   
## Median :2.000 Median :3.000 Median :3.000   
## Mean :2.057 Mean :2.768 Mean :2.969   
## 3rd Qu.:2.000 3rd Qu.:3.000 3rd Qu.:3.000   
## Max. :3.000 Max. :3.000 Max. :3.000   
##   
## Medical\_History\_15 Medical\_History\_16 Medical\_History\_17  
## Min. : 0.0 Min. :1.000 Min. :1.000   
## 1st Qu.: 17.0 1st Qu.:1.000 1st Qu.:3.000   
## Median :117.0 Median :1.000 Median :3.000   
## Mean :123.8 Mean :1.328 Mean :2.978   
## 3rd Qu.:240.0 3rd Qu.:1.000 3rd Qu.:3.000   
## Max. :240.0 Max. :3.000 Max. :3.000   
## NA's :44596   
## Medical\_History\_18 Medical\_History\_19 Medical\_History\_20  
## Min. :1.000 Min. :1.000 Min. :1.000   
## 1st Qu.:1.000 1st Qu.:1.000 1st Qu.:2.000   
## Median :1.000 Median :1.000 Median :2.000   
## Mean :1.054 Mean :1.034 Mean :1.985   
## 3rd Qu.:1.000 3rd Qu.:1.000 3rd Qu.:2.000   
## Max. :3.000 Max. :3.000 Max. :3.000   
##   
## Medical\_History\_21 Medical\_History\_22 Medical\_History\_23  
## Min. :1.000 Min. :1.000 Min. :1.000   
## 1st Qu.:1.000 1st Qu.:2.000 1st Qu.:3.000   
## Median :1.000 Median :2.000 Median :3.000   
## Mean :1.109 Mean :1.982 Mean :2.528   
## 3rd Qu.:1.000 3rd Qu.:2.000 3rd Qu.:3.000   
## Max. :3.000 Max. :2.000 Max. :3.000   
##   
## Medical\_History\_24 Medical\_History\_25 Medical\_History\_26  
## Min. : 0.00 Min. :1.000 Min. :1.000   
## 1st Qu.: 1.00 1st Qu.:1.000 1st Qu.:3.000   
## Median : 8.00 Median :1.000 Median :3.000   
## Mean : 50.64 Mean :1.195 Mean :2.809   
## 3rd Qu.: 64.00 3rd Qu.:1.000 3rd Qu.:3.000   
## Max. :240.00 Max. :3.000 Max. :3.000   
## NA's :55580   
## Medical\_History\_27 Medical\_History\_28 Medical\_History\_29  
## Min. :1.00 Min. :1.000 Min. :1.000   
## 1st Qu.:3.00 1st Qu.:1.000 1st Qu.:3.000   
## Median :3.00 Median :1.000 Median :3.000   
## Mean :2.98 Mean :1.067 Mean :2.543   
## 3rd Qu.:3.00 3rd Qu.:1.000 3rd Qu.:3.000   
## Max. :3.00 Max. :3.000 Max. :3.000   
##   
## Medical\_History\_30 Medical\_History\_31 Medical\_History\_32  
## Min. :1.000 Min. :1.000 Min. : 0.00   
## 1st Qu.:2.000 1st Qu.:3.000 1st Qu.: 0.00   
## Median :2.000 Median :3.000 Median : 0.00   
## Mean :2.041 Mean :2.985 Mean : 11.97   
## 3rd Qu.:2.000 3rd Qu.:3.000 3rd Qu.: 2.00   
## Max. :3.000 Max. :3.000 Max. :240.00   
## NA's :58274   
## Medical\_History\_33 Medical\_History\_34 Medical\_History\_35  
## Min. :1.000 Min. :1.000 Min. :1.000   
## 1st Qu.:3.000 1st Qu.:3.000 1st Qu.:1.000   
## Median :3.000 Median :3.000 Median :1.000   
## Mean :2.805 Mean :2.689 Mean :1.002   
## 3rd Qu.:3.000 3rd Qu.:3.000 3rd Qu.:1.000   
## Max. :3.000 Max. :3.000 Max. :3.000   
##   
## Medical\_History\_36 Medical\_History\_37 Medical\_History\_38  
## Min. :1.000 Min. :1.000 Min. :1.000   
## 1st Qu.:2.000 1st Qu.:2.000 1st Qu.:1.000   
## Median :2.000 Median :2.000 Median :1.000   
## Mean :2.179 Mean :1.938 Mean :1.005   
## 3rd Qu.:2.000 3rd Qu.:2.000 3rd Qu.:1.000   
## Max. :3.000 Max. :3.000 Max. :2.000   
##   
## Medical\_History\_39 Medical\_History\_40 Medical\_History\_41  
## Min. :1.000 Min. :1.000 Min. :1.000   
## 1st Qu.:3.000 1st Qu.:3.000 1st Qu.:1.000   
## Median :3.000 Median :3.000 Median :1.000   
## Mean :2.831 Mean :2.968 Mean :1.641   
## 3rd Qu.:3.000 3rd Qu.:3.000 3rd Qu.:3.000   
## Max. :3.000 Max. :3.000 Max. :3.000   
##   
## Medical\_Keyword\_1 Medical\_Keyword\_2 Medical\_Keyword\_3 Medical\_Keyword\_4  
## Min. :0.000 Min. :0.000000 Min. :0.00000 Min. :0.00000   
## 1st Qu.:0.000 1st Qu.:0.000000 1st Qu.:0.00000 1st Qu.:0.00000   
## Median :0.000 Median :0.000000 Median :0.00000 Median :0.00000   
## Mean :0.042 Mean :0.008942 Mean :0.04927 Mean :0.01455   
## 3rd Qu.:0.000 3rd Qu.:0.000000 3rd Qu.:0.00000 3rd Qu.:0.00000   
## Max. :1.000 Max. :1.000000 Max. :1.00000 Max. :1.00000   
##   
## Medical\_Keyword\_5 Medical\_Keyword\_6 Medical\_Keyword\_7 Medical\_Keyword\_8  
## Min. :0.000000 Min. :0.0000 Min. :0.00000 Min. :0.00000   
## 1st Qu.:0.000000 1st Qu.:0.0000 1st Qu.:0.00000 1st Qu.:0.00000   
## Median :0.000000 Median :0.0000 Median :0.00000 Median :0.00000   
## Mean :0.008622 Mean :0.0126 Mean :0.01391 Mean :0.01041   
## 3rd Qu.:0.000000 3rd Qu.:0.0000 3rd Qu.:0.00000 3rd Qu.:0.00000   
## Max. :1.000000 Max. :1.0000 Max. :1.00000 Max. :1.00000   
##   
## Medical\_Keyword\_9 Medical\_Keyword\_10 Medical\_Keyword\_11  
## Min. :0.000000 Min. :0.00000 Min. :0.00000   
## 1st Qu.:0.000000 1st Qu.:0.00000 1st Qu.:0.00000   
## Median :0.000000 Median :0.00000 Median :0.00000   
## Mean :0.006652 Mean :0.03646 Mean :0.05802   
## 3rd Qu.:0.000000 3rd Qu.:0.00000 3rd Qu.:0.00000   
## Max. :1.000000 Max. :1.00000 Max. :1.00000   
##   
## Medical\_Keyword\_12 Medical\_Keyword\_13 Medical\_Keyword\_14  
## Min. :0.00 Min. :0.000000 Min. :0.000000   
## 1st Qu.:0.00 1st Qu.:0.000000 1st Qu.:0.000000   
## Median :0.00 Median :0.000000 Median :0.000000   
## Mean :0.01 Mean :0.005961 Mean :0.007848   
## 3rd Qu.:0.00 3rd Qu.:0.000000 3rd Qu.:0.000000   
## Max. :1.00 Max. :1.000000 Max. :1.000000   
##   
## Medical\_Keyword\_15 Medical\_Keyword\_16 Medical\_Keyword\_17  
## Min. :0.0000 Min. :0.00000 Min. :0.000000   
## 1st Qu.:0.0000 1st Qu.:0.00000 1st Qu.:0.000000   
## Median :0.0000 Median :0.00000 Median :0.000000   
## Mean :0.1905 Mean :0.01271 Mean :0.009161   
## 3rd Qu.:0.0000 3rd Qu.:0.00000 3rd Qu.:0.000000   
## Max. :1.0000 Max. :1.00000 Max. :1.000000   
##   
## Medical\_Keyword\_18 Medical\_Keyword\_19 Medical\_Keyword\_20  
## Min. :0.000000 Min. :0.000000 Min. :0.000000   
## 1st Qu.:0.000000 1st Qu.:0.000000 1st Qu.:0.000000   
## Median :0.000000 Median :0.000000 Median :0.000000   
## Mean :0.007494 Mean :0.009296 Mean :0.008134   
## 3rd Qu.:0.000000 3rd Qu.:0.000000 3rd Qu.:0.000000   
## Max. :1.000000 Max. :1.000000 Max. :1.000000   
##   
## Medical\_Keyword\_21 Medical\_Keyword\_22 Medical\_Keyword\_23  
## Min. :0.0000 Min. :0.00000 Min. :0.00000   
## 1st Qu.:0.0000 1st Qu.:0.00000 1st Qu.:0.00000   
## Median :0.0000 Median :0.00000 Median :0.00000   
## Mean :0.0146 Mean :0.03717 Mean :0.09778   
## 3rd Qu.:0.0000 3rd Qu.:0.00000 3rd Qu.:0.00000   
## Max. :1.0000 Max. :1.00000 Max. :1.00000   
##   
## Medical\_Keyword\_24 Medical\_Keyword\_25 Medical\_Keyword\_26  
## Min. :0.00000 Min. :0.00000 Min. :0.00000   
## 1st Qu.:0.00000 1st Qu.:0.00000 1st Qu.:0.00000   
## Median :0.00000 Median :0.00000 Median :0.00000   
## Mean :0.01889 Mean :0.08946 Mean :0.01344   
## 3rd Qu.:0.00000 3rd Qu.:0.00000 3rd Qu.:0.00000   
## Max. :1.00000 Max. :1.00000 Max. :1.00000   
##   
## Medical\_Keyword\_27 Medical\_Keyword\_28 Medical\_Keyword\_29  
## Min. :0.00000 Min. :0.00000 Min. :0.00000   
## 1st Qu.:0.00000 1st Qu.:0.00000 1st Qu.:0.00000   
## Median :0.00000 Median :0.00000 Median :0.00000   
## Mean :0.01186 Mean :0.01494 Mean :0.01175   
## 3rd Qu.:0.00000 3rd Qu.:0.00000 3rd Qu.:0.00000   
## Max. :1.00000 Max. :1.00000 Max. :1.00000   
##   
## Medical\_Keyword\_30 Medical\_Keyword\_31 Medical\_Keyword\_32  
## Min. :0.00000 Min. :0.0000 Min. :0.00000   
## 1st Qu.:0.00000 1st Qu.:0.0000 1st Qu.:0.00000   
## Median :0.00000 Median :0.0000 Median :0.00000   
## Mean :0.02504 Mean :0.0109 Mean :0.02117   
## 3rd Qu.:0.00000 3rd Qu.:0.0000 3rd Qu.:0.00000   
## Max. :1.00000 Max. :1.0000 Max. :1.00000   
##   
## Medical\_Keyword\_33 Medical\_Keyword\_34 Medical\_Keyword\_35  
## Min. :0.00000 Min. :0.00000 Min. :0.000000   
## 1st Qu.:0.00000 1st Qu.:0.00000 1st Qu.:0.000000   
## Median :0.00000 Median :0.00000 Median :0.000000   
## Mean :0.02284 Mean :0.02065 Mean :0.006938   
## 3rd Qu.:0.00000 3rd Qu.:0.00000 3rd Qu.:0.000000   
## Max. :1.00000 Max. :1.00000 Max. :1.000000   
##   
## Medical\_Keyword\_36 Medical\_Keyword\_37 Medical\_Keyword\_38  
## Min. :0.00000 Min. :0.00000 Min. :0.000000   
## 1st Qu.:0.00000 1st Qu.:0.00000 1st Qu.:0.000000   
## Median :0.00000 Median :0.00000 Median :0.000000   
## Mean :0.01041 Mean :0.06659 Mean :0.006837   
## 3rd Qu.:0.00000 3rd Qu.:0.00000 3rd Qu.:0.000000   
## Max. :1.00000 Max. :1.00000 Max. :1.000000   
##   
## Medical\_Keyword\_39 Medical\_Keyword\_40 Medical\_Keyword\_41  
## Min. :0.00000 Min. :0.00000 Min. :0.00000   
## 1st Qu.:0.00000 1st Qu.:0.00000 1st Qu.:0.00000   
## Median :0.00000 Median :0.00000 Median :0.00000   
## Mean :0.01366 Mean :0.05695 Mean :0.01005   
## 3rd Qu.:0.00000 3rd Qu.:0.00000 3rd Qu.:0.00000   
## Max. :1.00000 Max. :1.00000 Max. :1.00000   
##   
## Medical\_Keyword\_42 Medical\_Keyword\_43 Medical\_Keyword\_44  
## Min. :0.00000 Min. :0.00000 Min. :0.000000   
## 1st Qu.:0.00000 1st Qu.:0.00000 1st Qu.:0.000000   
## Median :0.00000 Median :0.00000 Median :0.000000   
## Mean :0.04554 Mean :0.01071 Mean :0.007528   
## 3rd Qu.:0.00000 3rd Qu.:0.00000 3rd Qu.:0.000000   
## Max. :1.00000 Max. :1.00000 Max. :1.000000   
##   
## Medical\_Keyword\_45 Medical\_Keyword\_46 Medical\_Keyword\_47  
## Min. :0.00000 Min. :0.000000 Min. :0.00000   
## 1st Qu.:0.00000 1st Qu.:0.000000 1st Qu.:0.00000   
## Median :0.00000 Median :0.000000 Median :0.00000   
## Mean :0.01369 Mean :0.008488 Mean :0.01991   
## 3rd Qu.:0.00000 3rd Qu.:0.000000 3rd Qu.:0.00000   
## Max. :1.00000 Max. :1.000000 Max. :1.00000   
##   
## Medical\_Keyword\_48 Response   
## Min. :0.0000 Min. :1.000   
## 1st Qu.:0.0000 1st Qu.:4.000   
## Median :0.0000 Median :6.000   
## Mean :0.0545 Mean :5.637   
## 3rd Qu.:0.0000 3rd Qu.:8.000   
## Max. :1.0000 Max. :8.000   
##

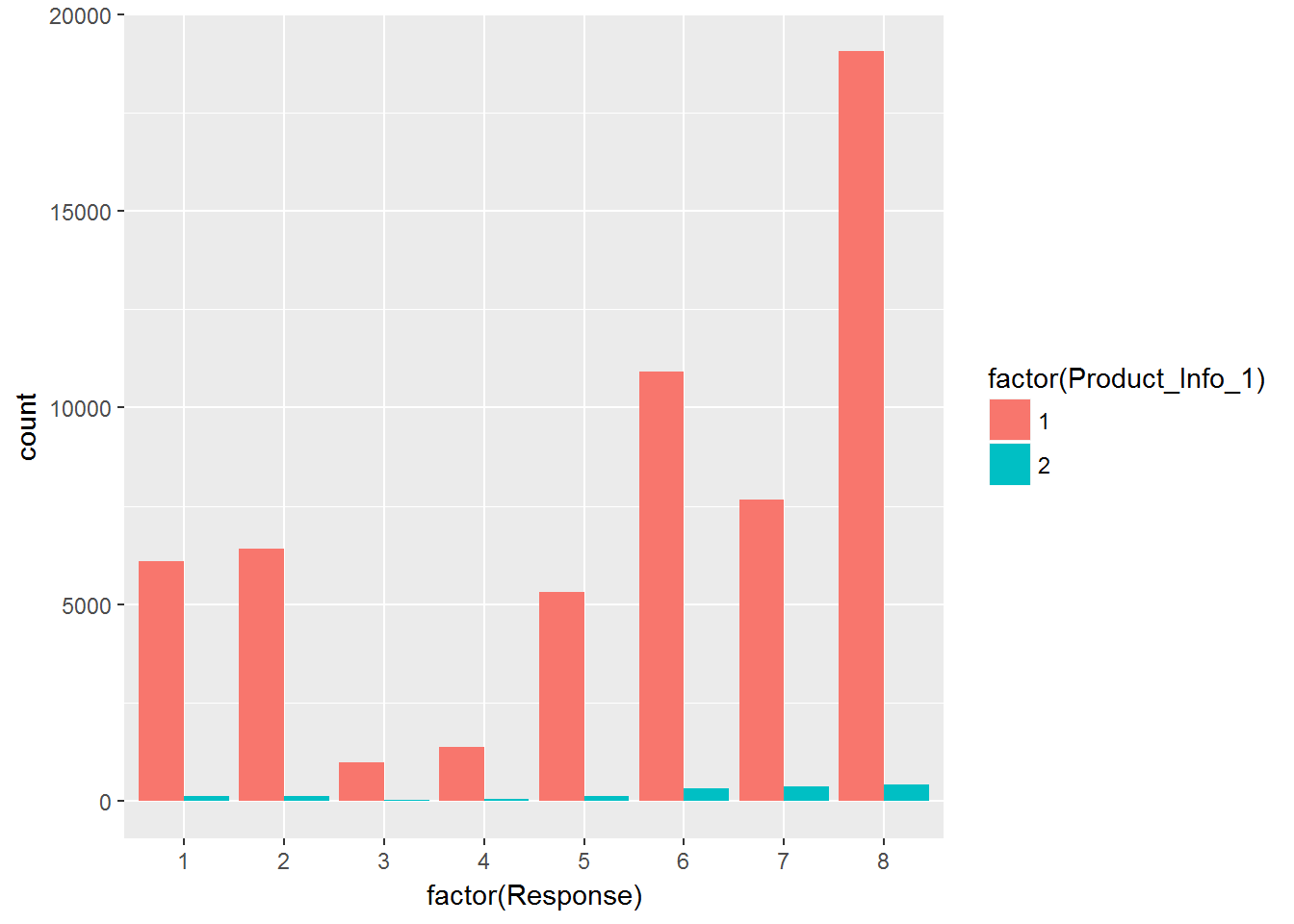
Response : 8이 가장 많으며 3, 4, 5 중간 번호가 적게 나타남

train %>% ggplot(aes(factor(Response))) + geom\_bar()



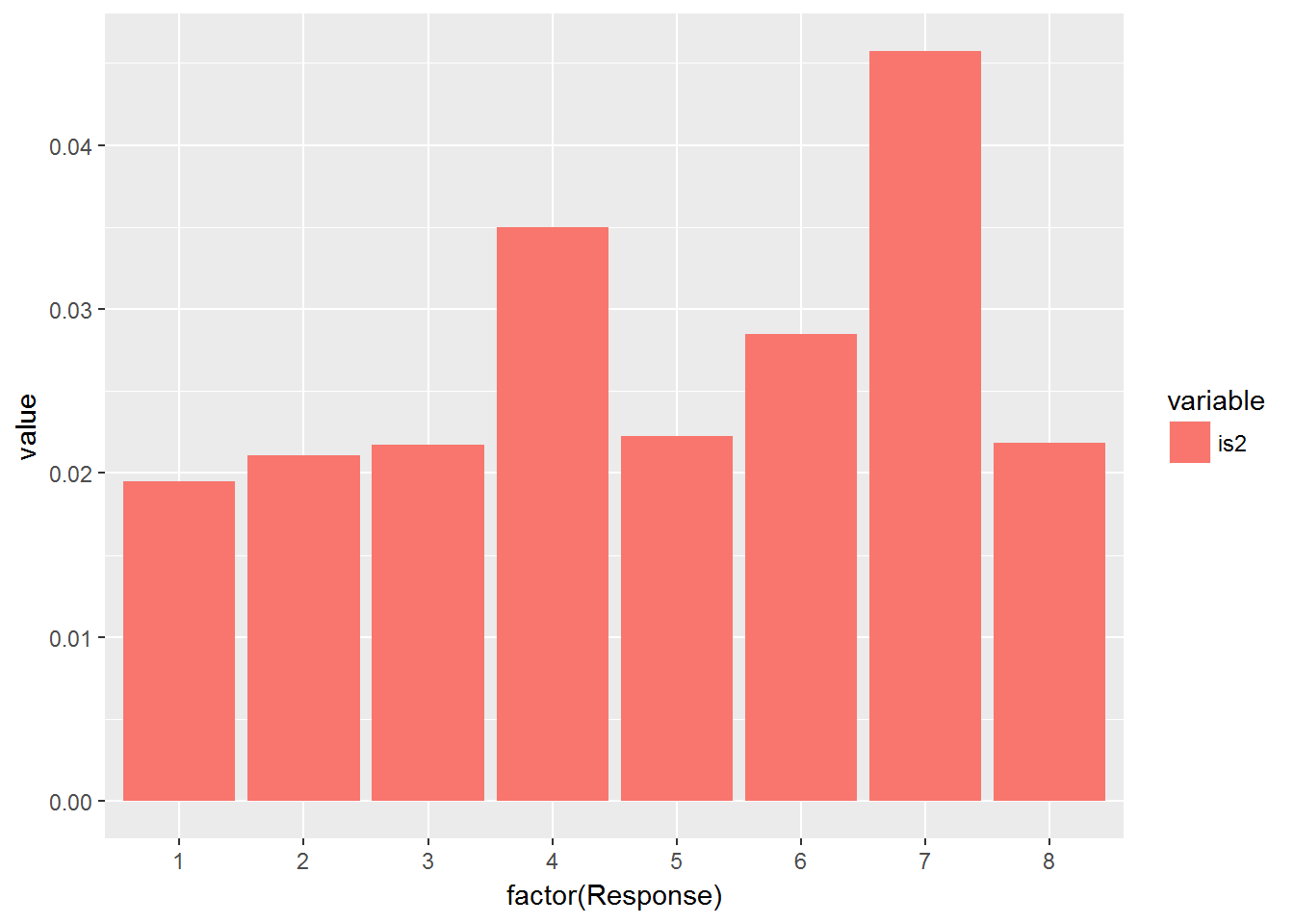
Product\_Info\_1 : Response가 8인 값이 1 빈도가 가장 많으며 2는 대체로 다 빈도수가 낮음

ggplot(train, aes(x=factor(Response),fill =factor(Product\_Info\_1))) +   
geom\_bar(position = "dodge")



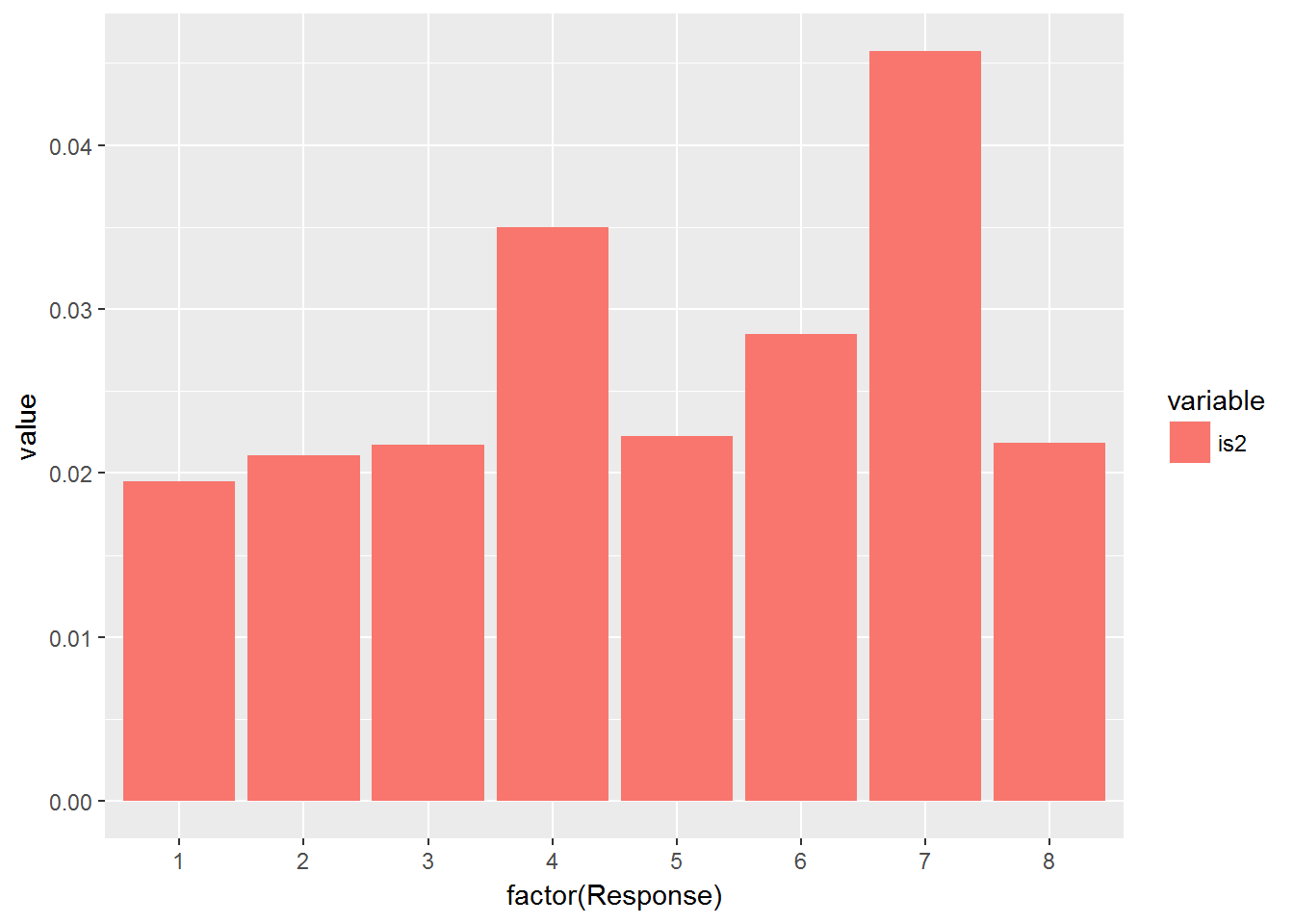
Product\_Info\_1 : 2만 보았을 때 Response 4, 7이 많게 나타남

train %>%  
 select(Response, Product\_Info\_1) %>%  
 group\_by(Response) %>%  
 summarise(is2 = sum(Product\_Info\_1 == 2)/n()) %>%  
 reshape2::melt(id.vars="Response",variable.names="is1+is2") %>%   
 ggplot(aes(x = factor(Response), y = value, fill = variable)) +   
 geom\_col(position = "dodge")



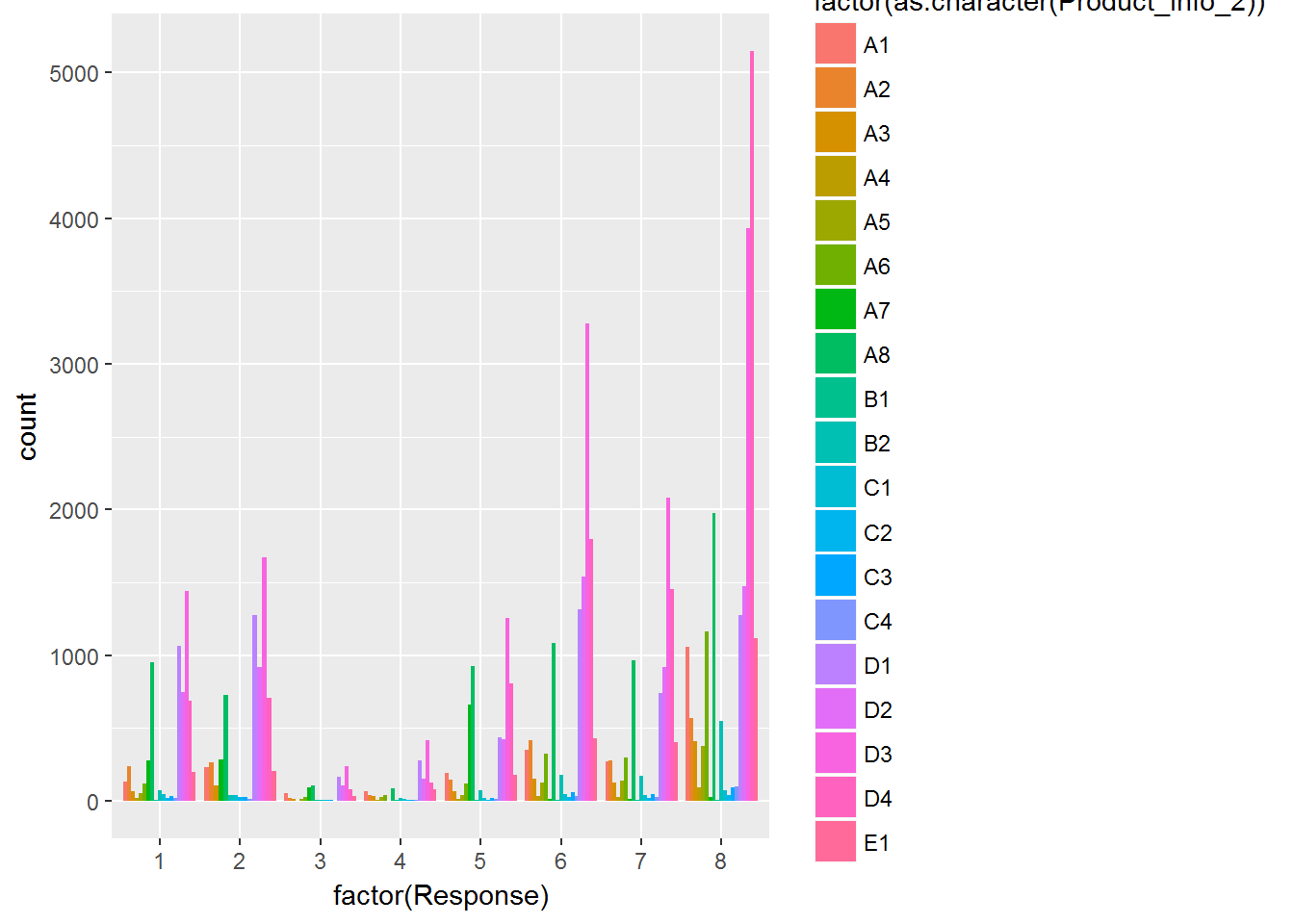
Product\_Info\_1 : 2만 보았을 때 Response 4, 7이 많게 나타남

train %>%  
 select(Response, Product\_Info\_1) %>%  
 group\_by(Response) %>%  
 summarise(is2 = sum(Product\_Info\_1 == 2)/n()) %>%  
 reshape2::melt(id.vars="Response",variable.names="is1+is2") %>%   
 ggplot(aes(x = factor(Response), y = value, fill = variable)) +   
 geom\_col(position = "dodge")



Product\_Info\_2 : Response 의 카운트에 따라 상품 갯수가 많이 나오며, 특히나 D3, D4 등의 상품이 많음

ggplot(train, aes(x=factor(Response), fill=factor(as.character(Product\_Info\_2))))+  
 geom\_bar(position = "dodge")



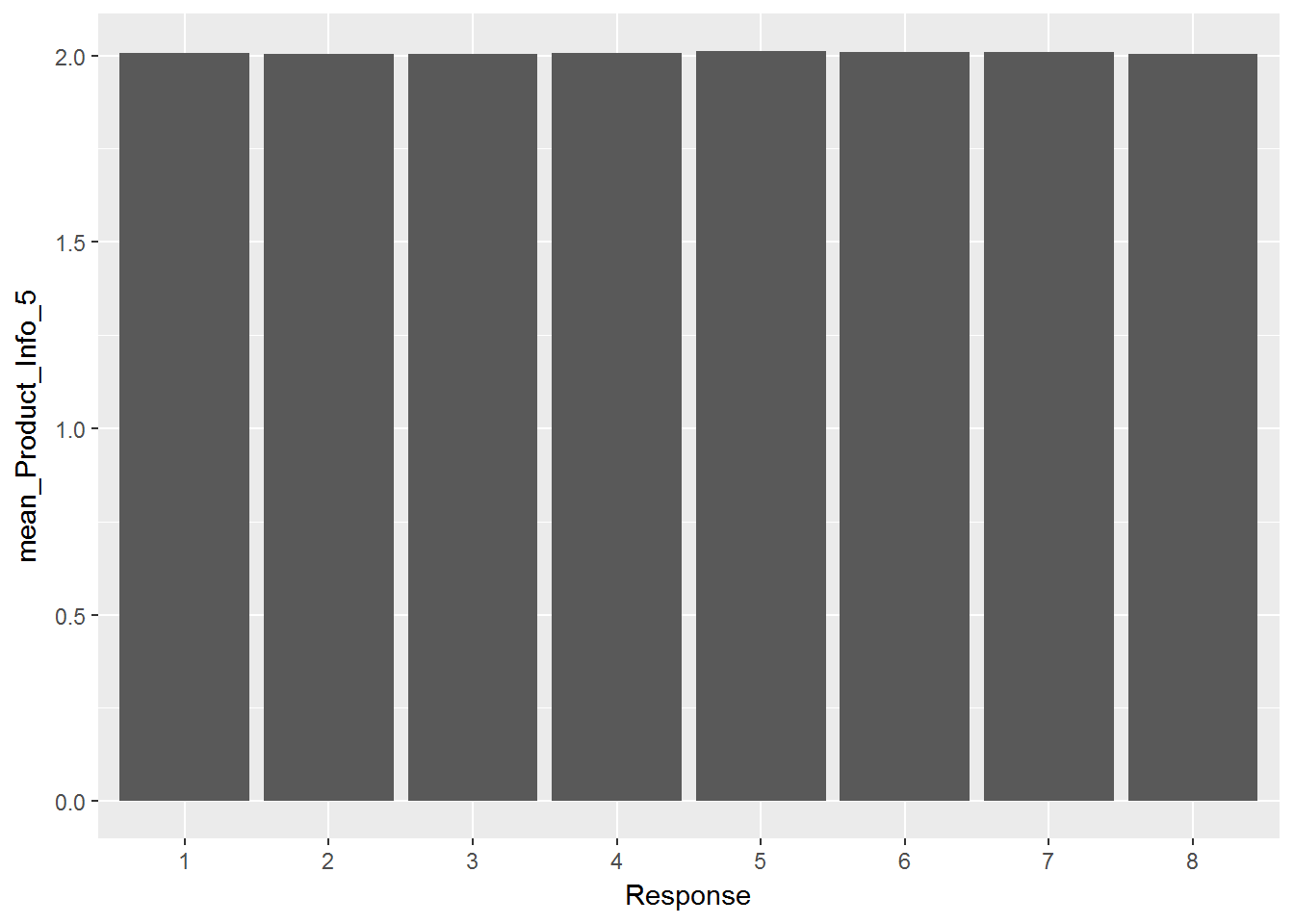
Product\_Info\_4 : 8의 평균값이 가장 높고 1에 가까울 수록 낮은 성향이 있음 / 변수로 활용하기 용이하다고 판단됨

ggplot(train[,list(mean\_Product\_Info\_5=mean(Product\_Info\_4)),by="Response"],  
 aes(x=Response, y=mean\_Product\_Info\_5)) + geom\_bar(stat="identity")



Product\_Info\_5 : 반면, Product\_Info\_5의 평균값이 대부분 2를 나타냄

ggplot(train[,list(mean\_Product\_Info\_5=mean(Product\_Info\_5)),by="Response"], aes(x=Response, y=mean\_Product\_Info\_5), fill=Response) + geom\_bar(stat="identity")



summary(aov(Product\_Info\_4 ~ Response, data= train)) # ANOVA 테스트

## Df Sum Sq Mean Sq F value Pr(>F)   
## Response 7 236 33.69 443.9 <2e-16 \*\*\*  
## Residuals 59373 4505 0.08   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

bartlett.test(Product\_Info\_4 ~ Response, data= train) # 오차의 등분산성 가정 만족

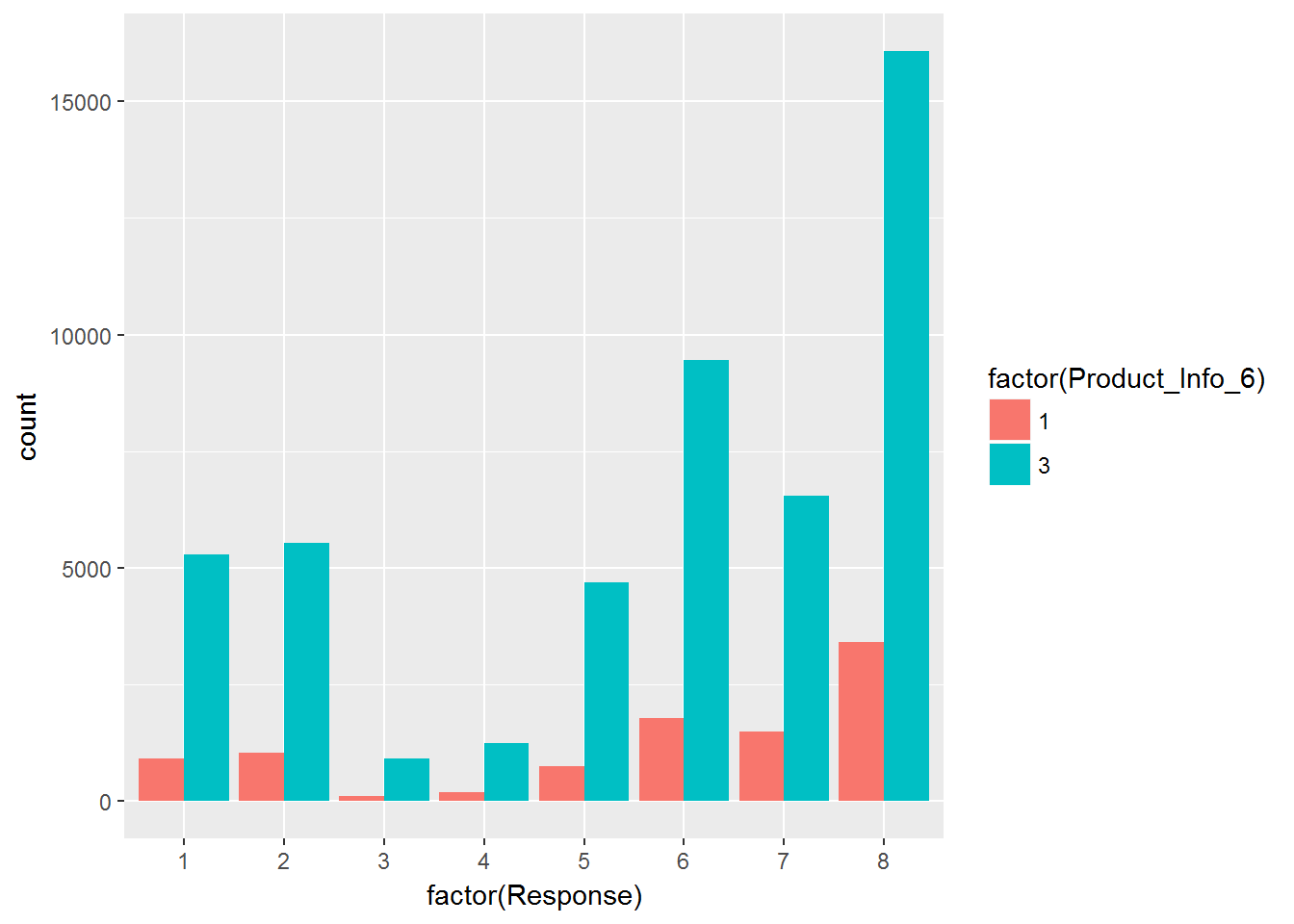
##   
## Bartlett test of homogeneity of variances  
##   
## data: Product\_Info\_4 by Response  
## Bartlett's K-squared = 669.62, df = 7, p-value < 2.2e-16

LDuncan(aov(Product\_Info\_4 ~ Response, data= train), "group") # 사후검정

##   
## DUNCAN TEST TO COMPARE MEANS   
##   
## Confidence Level: 0.95   
## Dependent Variable: Product\_Info\_4  
## Variation Coefficient: 83.73924 %   
##   
##   
## Independent Variable: Response   
## Factors Means   
## 8 0.399571535685515 a   
## 7 0.349039343015572 b   
## 6 0.339379173576338 c   
## 4 0.256563411668067 d   
## 1 0.253301987630739 de   
## 2 0.249476076569444 de   
## 5 0.247065582743373 e   
## 3 0.214226992757157 f

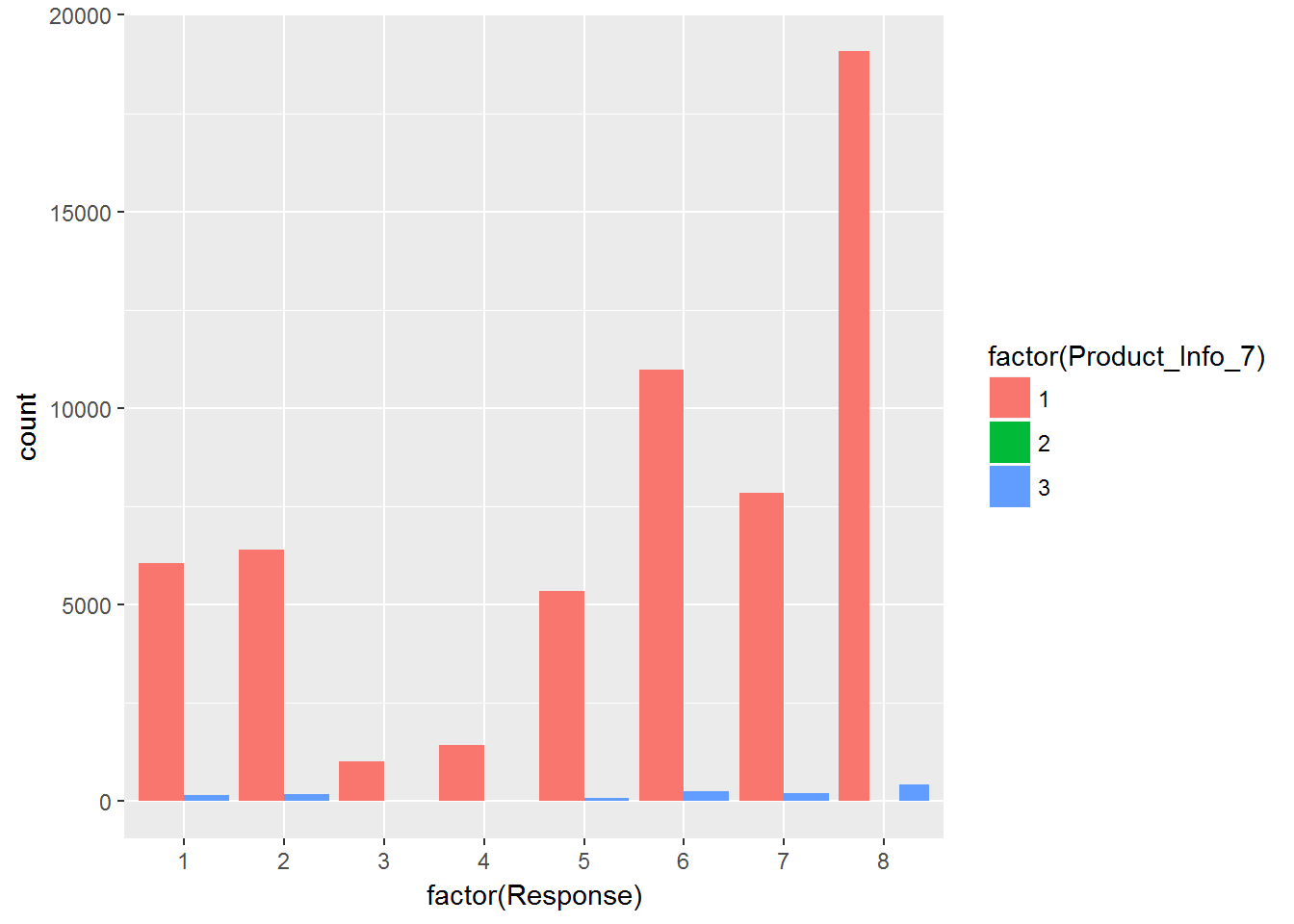
Product\_Info\_6 : 1과 3이 가장 많게 나타나며, Response 3,4는 차이가 크지 않고, 8번이 크게 나타남

ggplot(train, aes(x=factor(Response), fill=factor(Product\_Info\_6)))+  
 geom\_bar(position = "dodge")



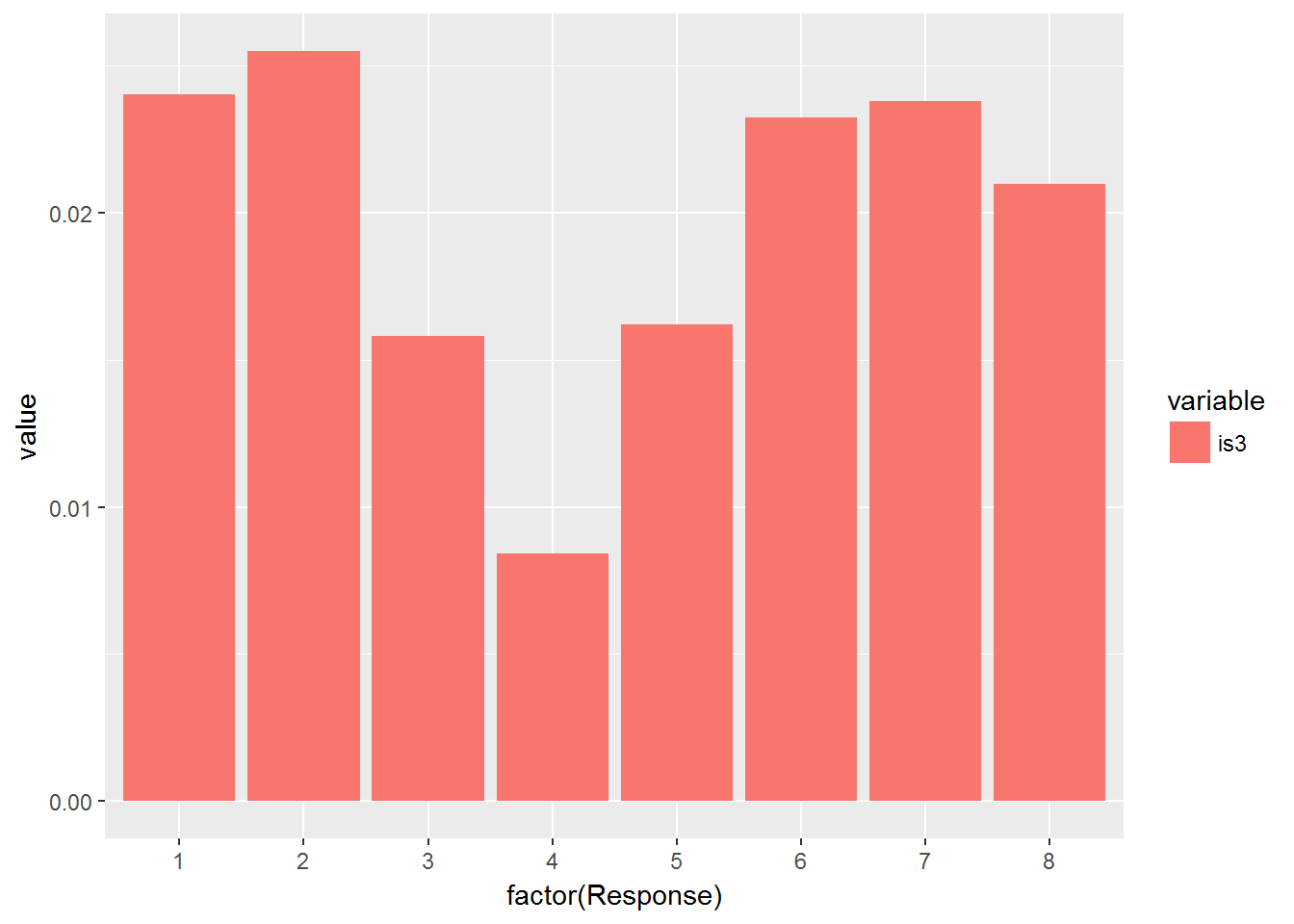
Product\_Info\_7 : 2번은 거의 없으며, Response 8을 제외하고는 존재하지 않음

ggplot(train, aes(x=factor(Response), fill=factor(Product\_Info\_7)))+  
 geom\_bar(position = "dodge")



Product\_Info\_7 : 3번은 Response가 3~6일때 카운트가 적음

train %>%  
 select(Response, Product\_Info\_7) %>%  
 group\_by(Response) %>%  
 summarise(is3 = sum(Product\_Info\_7 == 3)/n()) %>%  
 reshape2::melt(id.vars="Response",variable.names="is3") %>%   
 ggplot(aes(x = factor(Response), y = value, fill = variable)) +   
 geom\_col(position = "dodge")



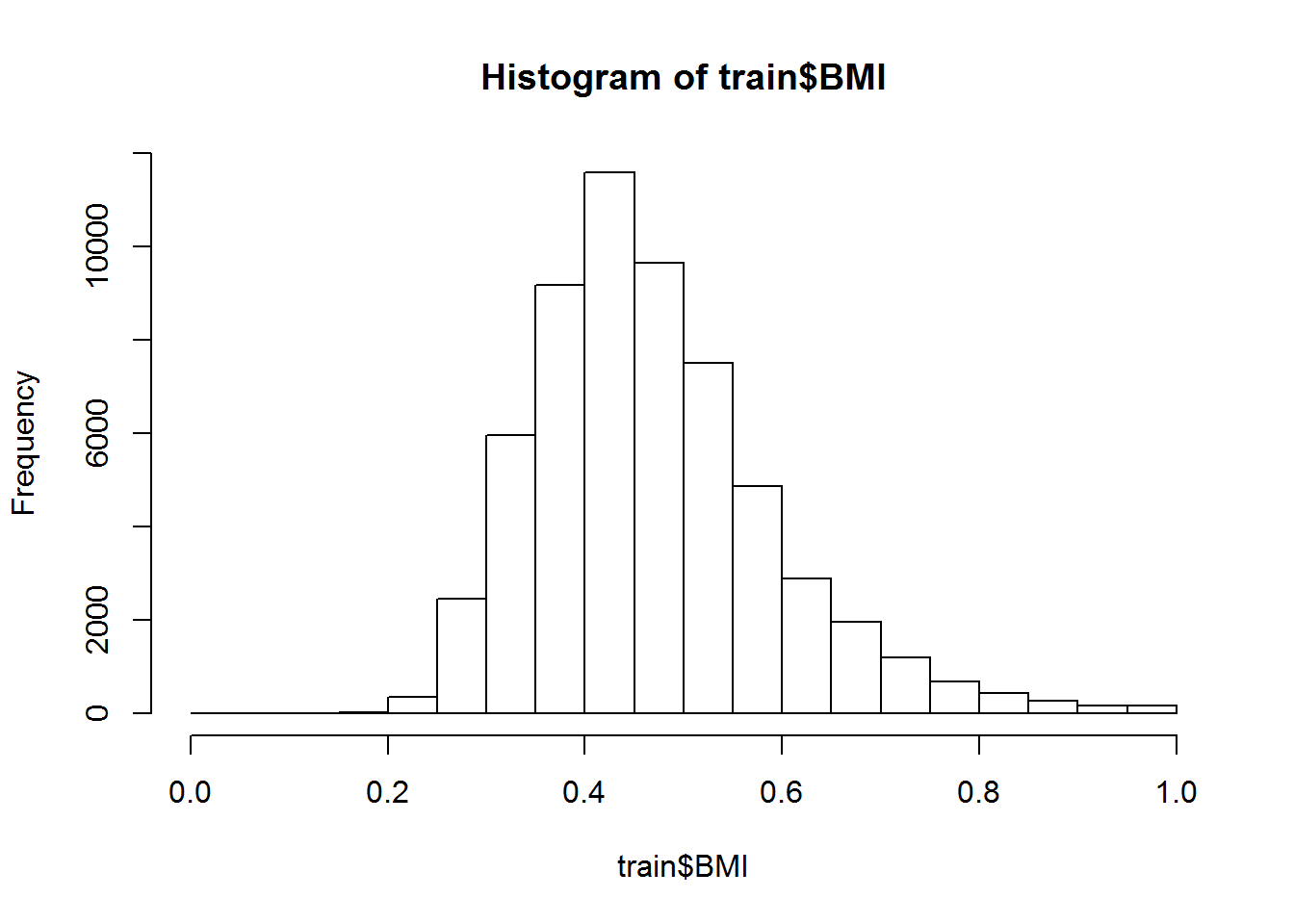
Ins\_age : 나이는 평균적으로 Response가 1에 가까울 수록 높게 나타났으며 8일 수록 적게 나타남

ggplot(train[,list(mean\_Ins\_Age=mean(Ins\_Age)),by="Response"],  
 aes(x=Response, y=mean\_Ins\_Age)) + geom\_bar(stat="identity")

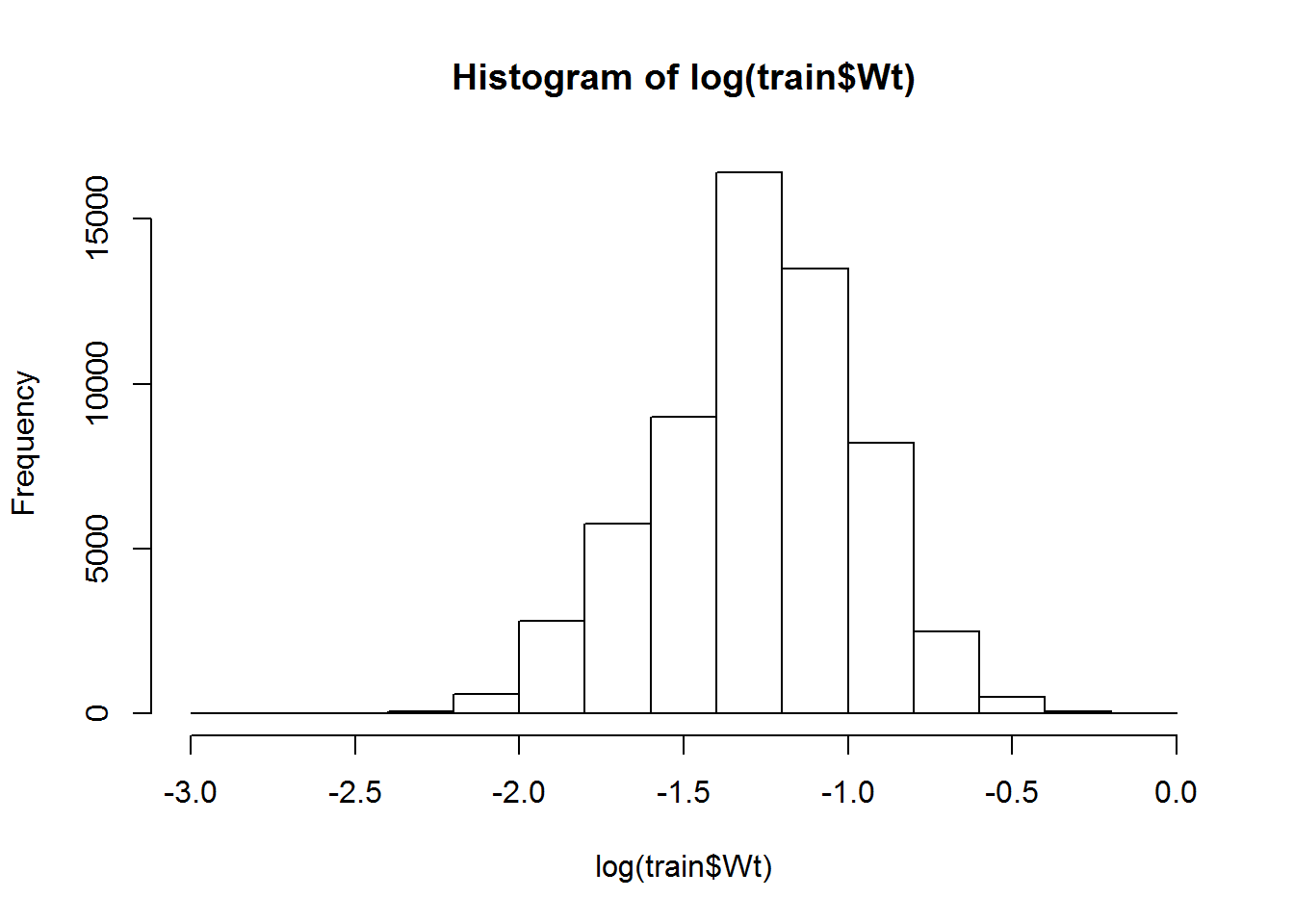


BMI, Tt, Wt, Ins\_Age : 정규 분포

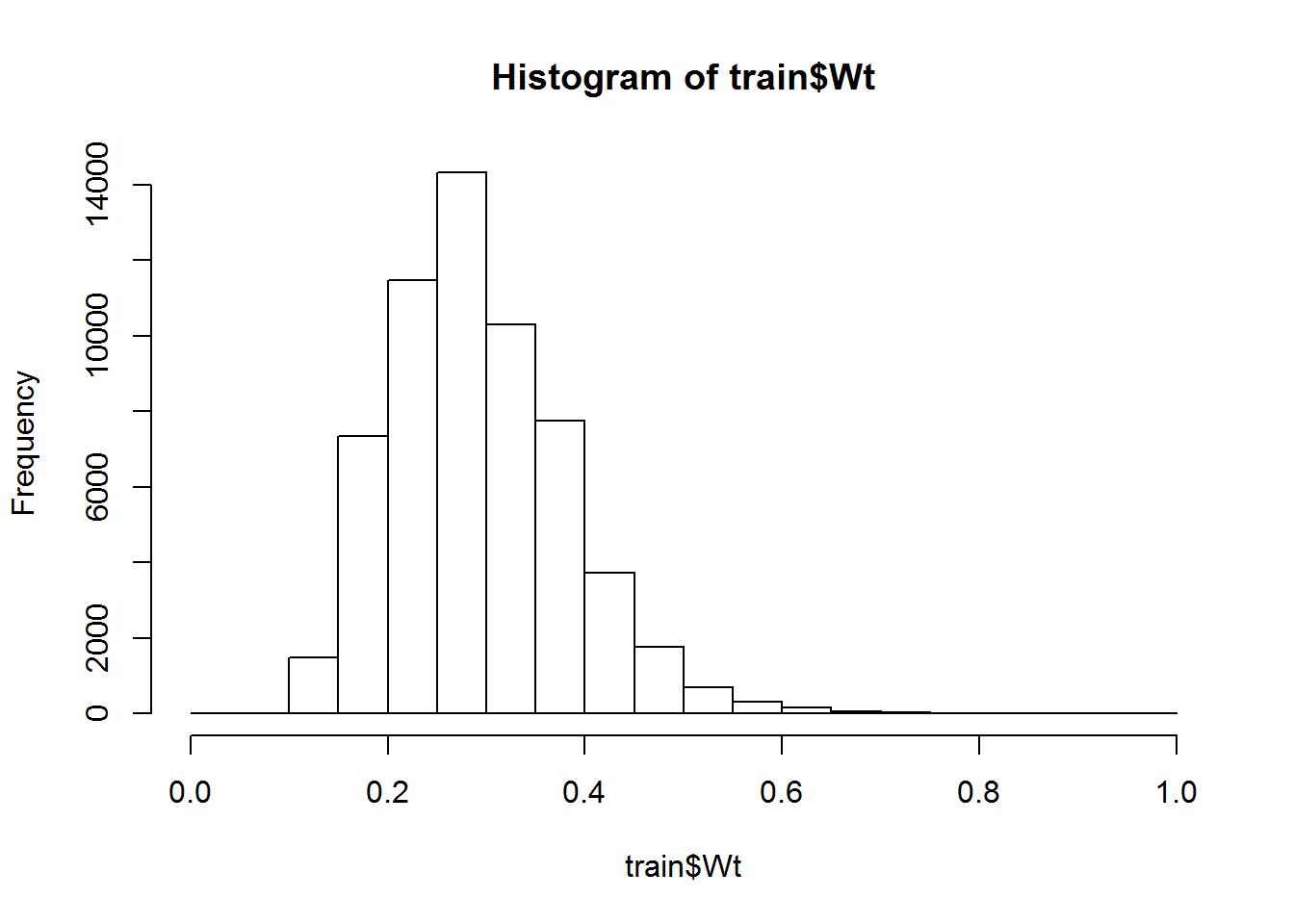
hist(train$BMI)



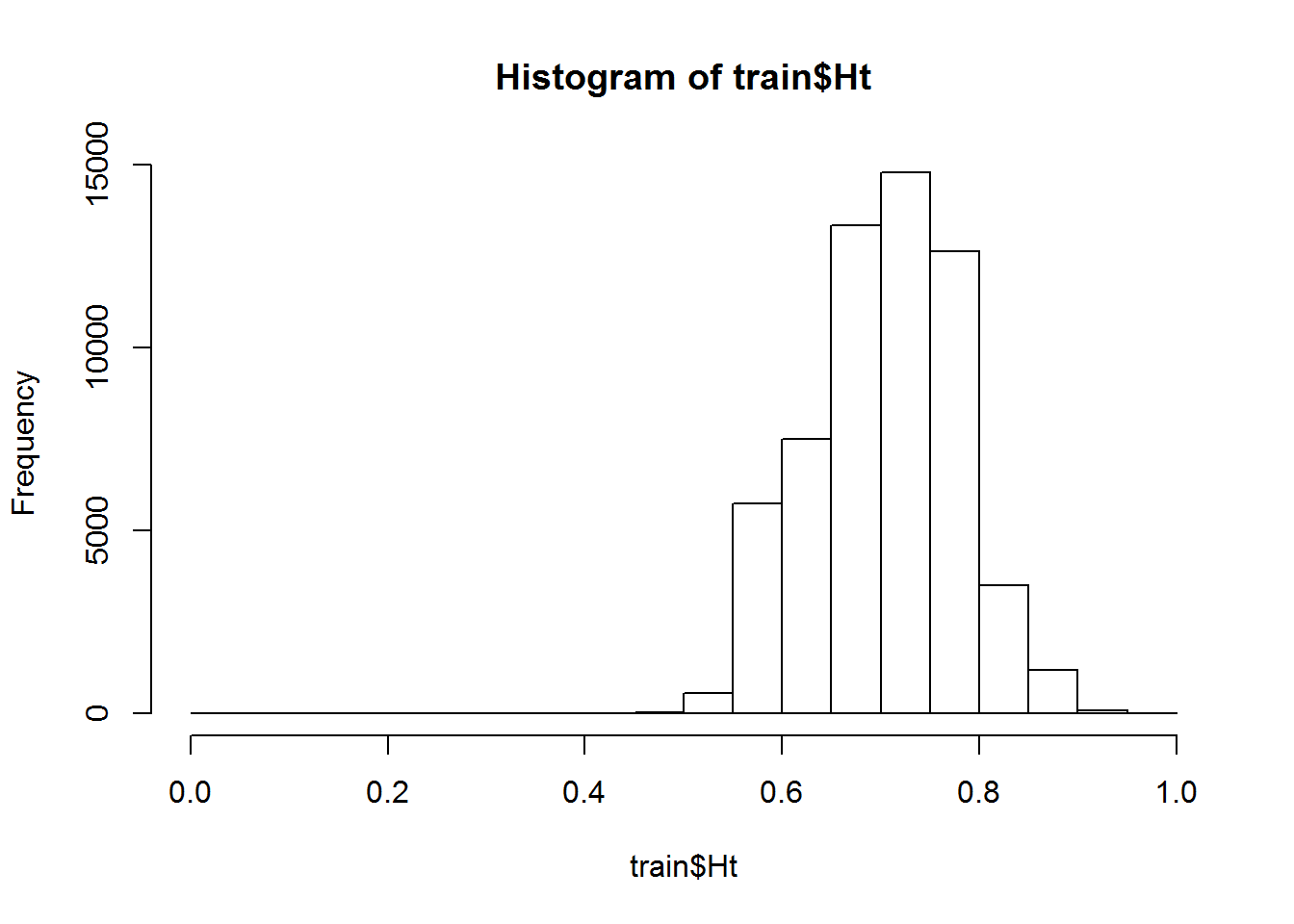
hist(log(train$Wt))



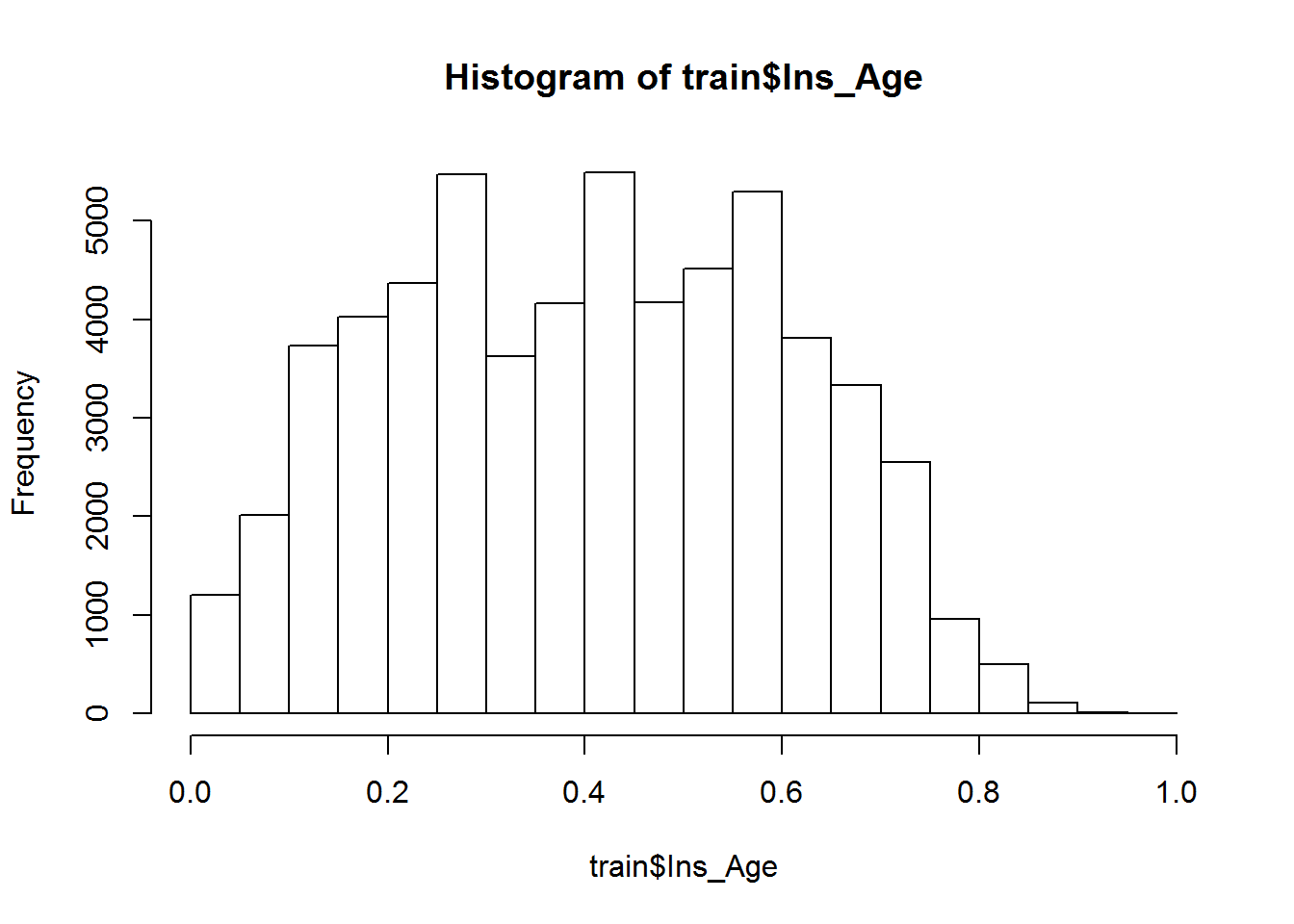
hist(train$Wt)



hist(train$Ht)



hist(train$Ins\_Age)



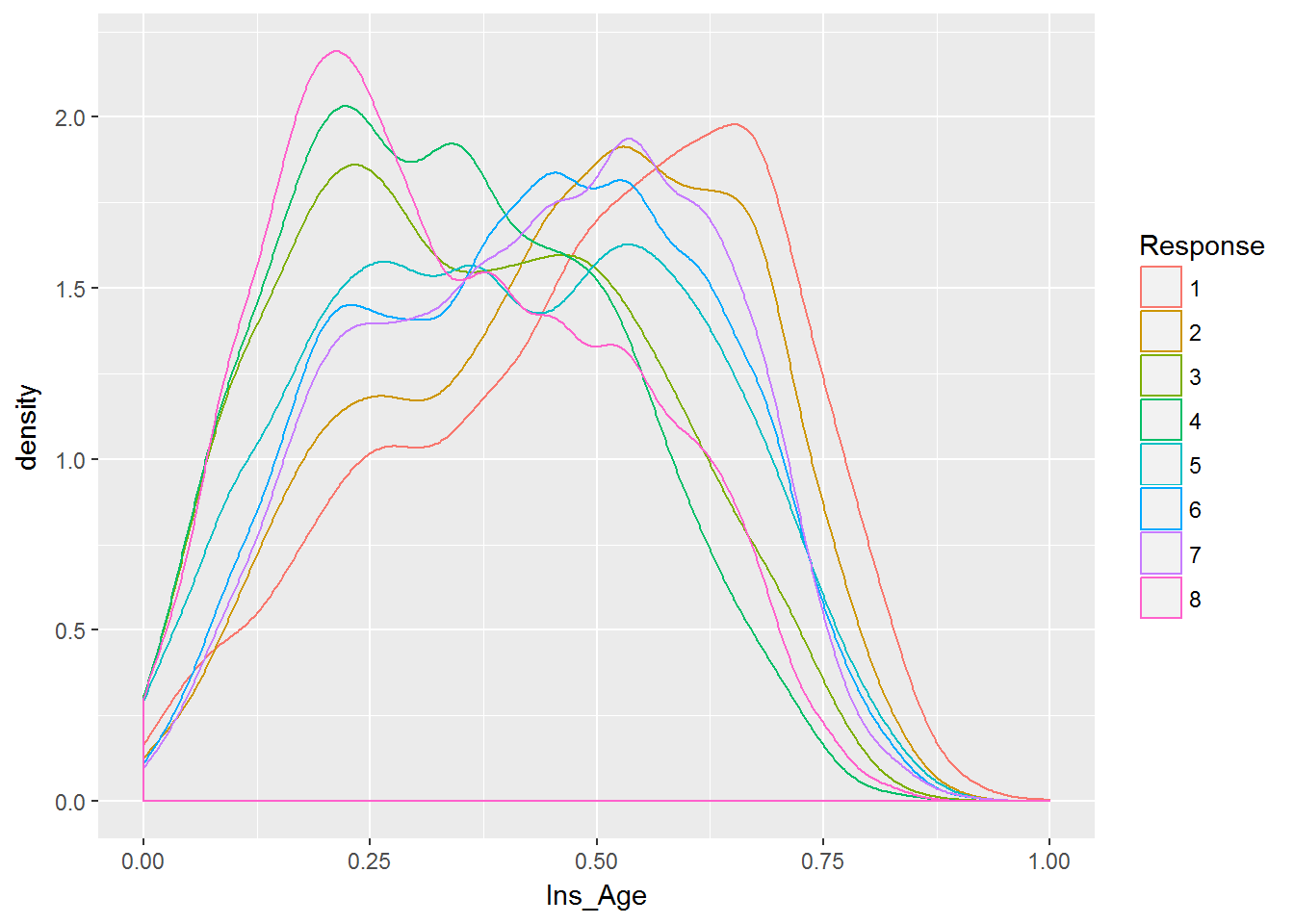
BMI : Response별 평균 BMI

ggplot(train[,list(mean\_BMI=mean(BMI)),by="Response"],  
 aes(x=Response, y=mean\_BMI)) + geom\_bar(stat="identity")

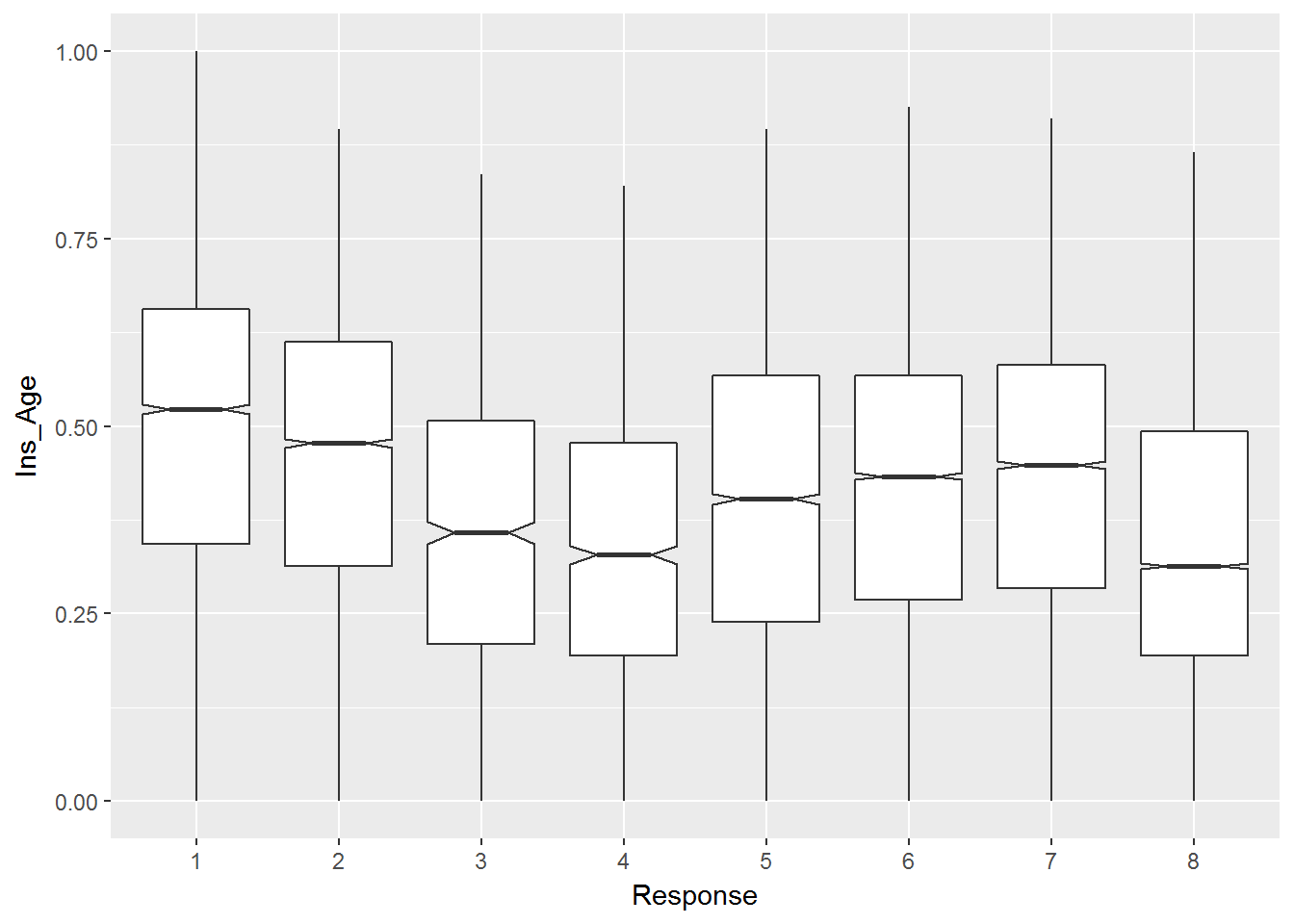


Ins\_age, Density Plot : age가 0.6이상인 경우에는 response 1등급이 가장 많으며, age가 0.3미만인 경우에는 response 8등급이 가정 많다.

ggplot(train, aes(Ins\_Age , colour = Response)) + geom\_density()

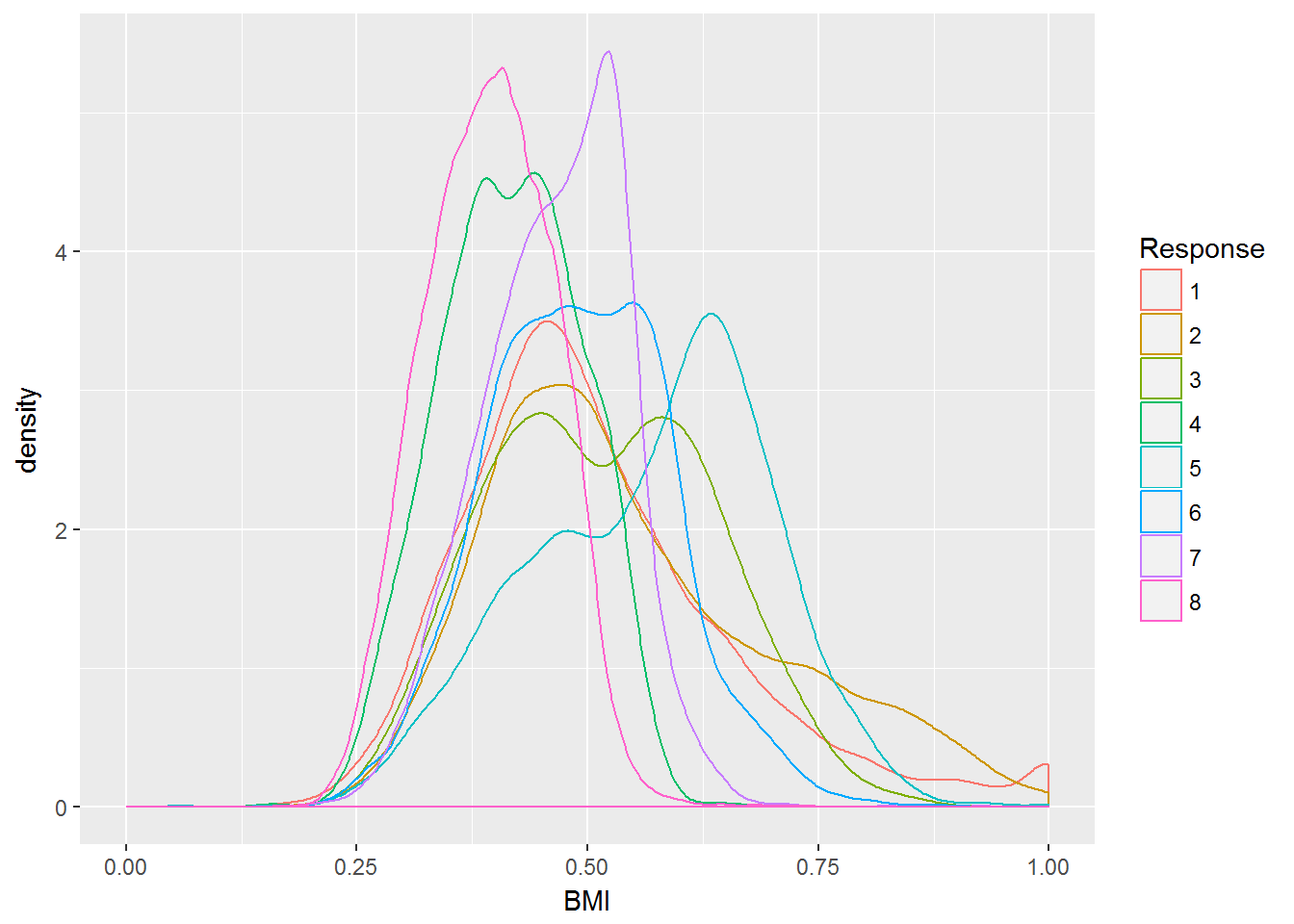


ggplot(train, aes(x=Response, y=Ins\_Age)) + geom\_boxplot(notch=TRUE)

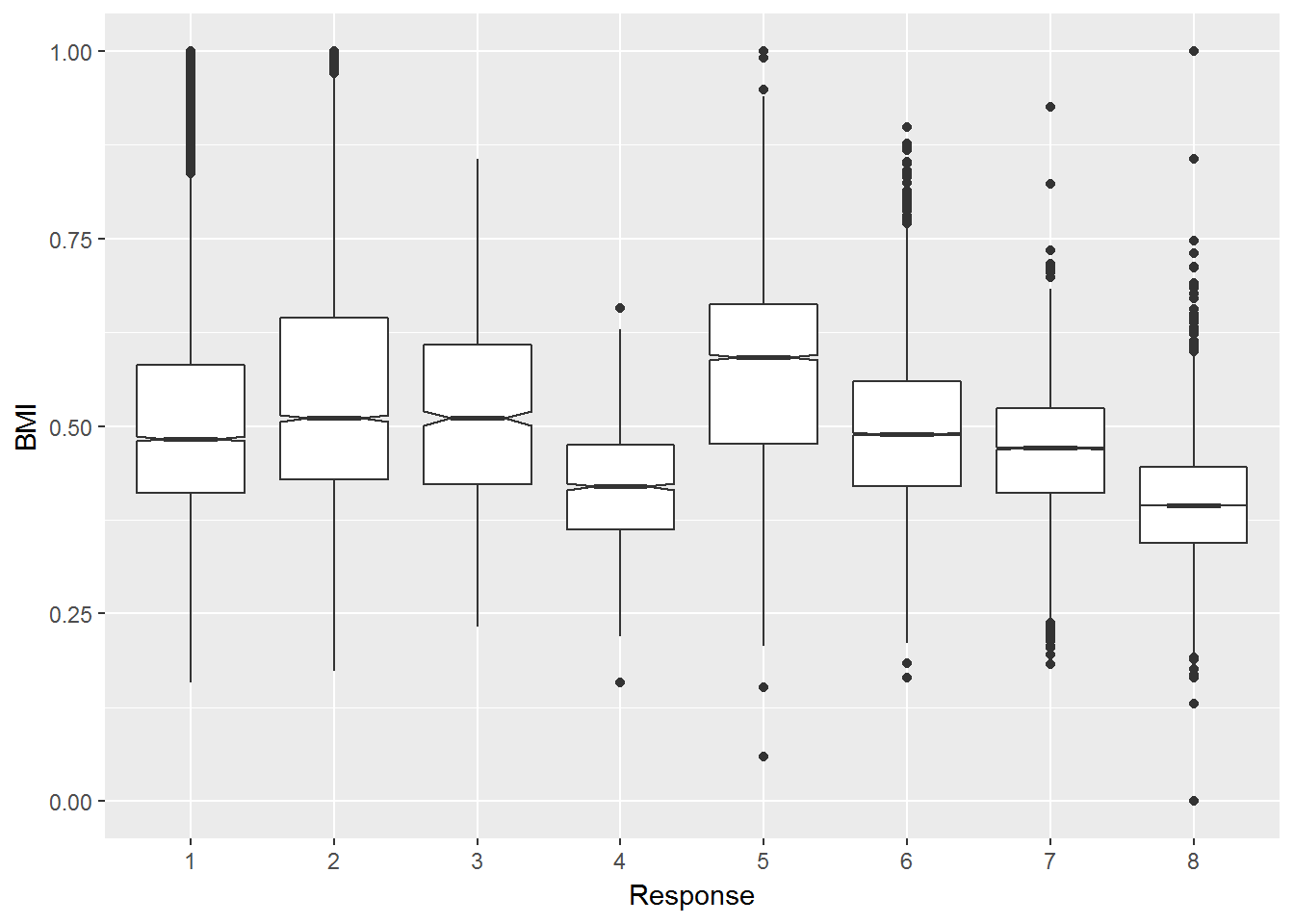


BMI, Density Plot : BMI 0.4미만이면 8등급의 인원 비율이 가장 높으며 BMI가 0.8이상인 경우는 대부분 1,2,3등급에 속한다

ggplot(train, aes(BMI, colour = Response)) + geom\_density()

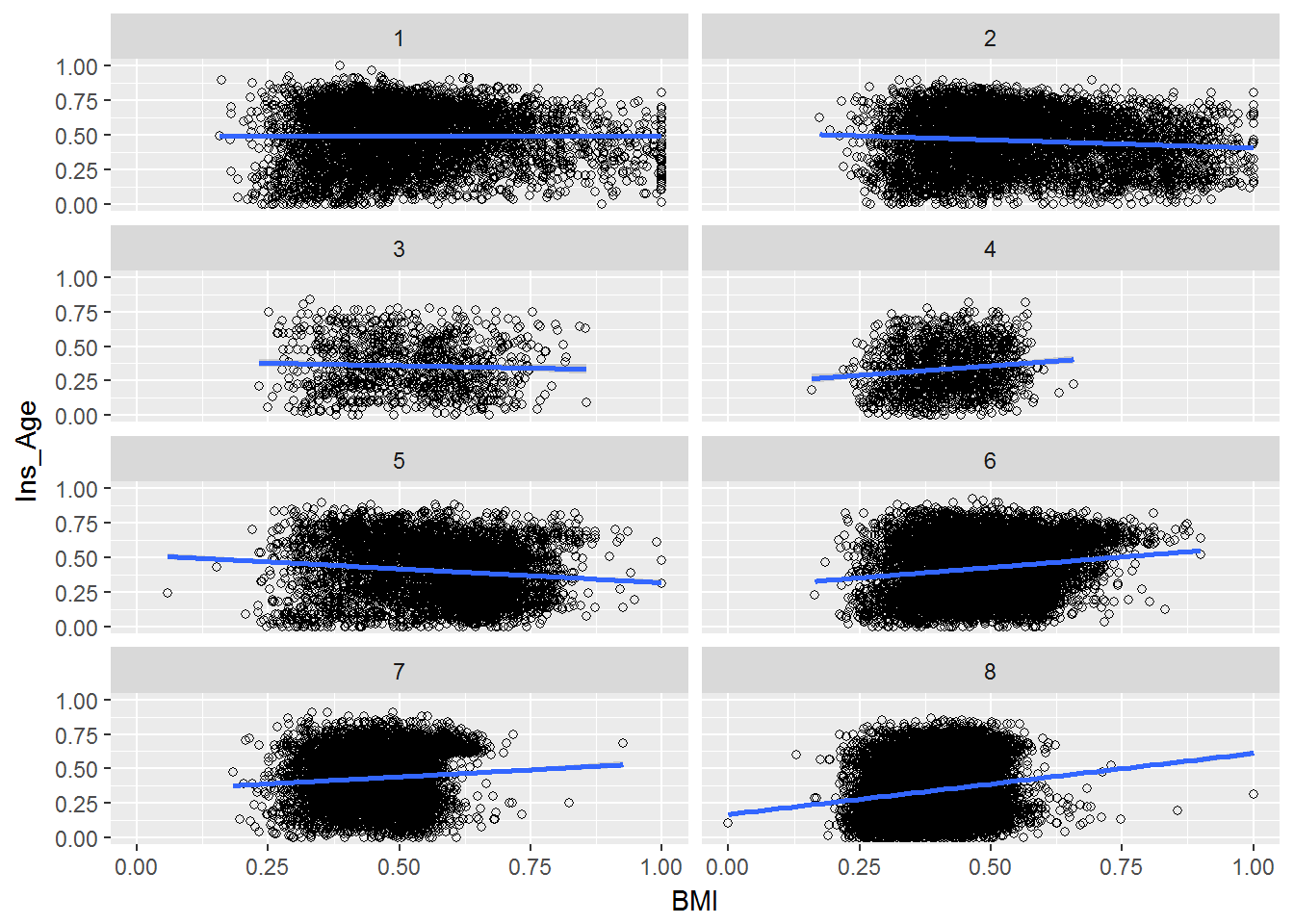


ggplot(train, aes(x=Response, y=BMI)) + geom\_boxplot(notch=TRUE)

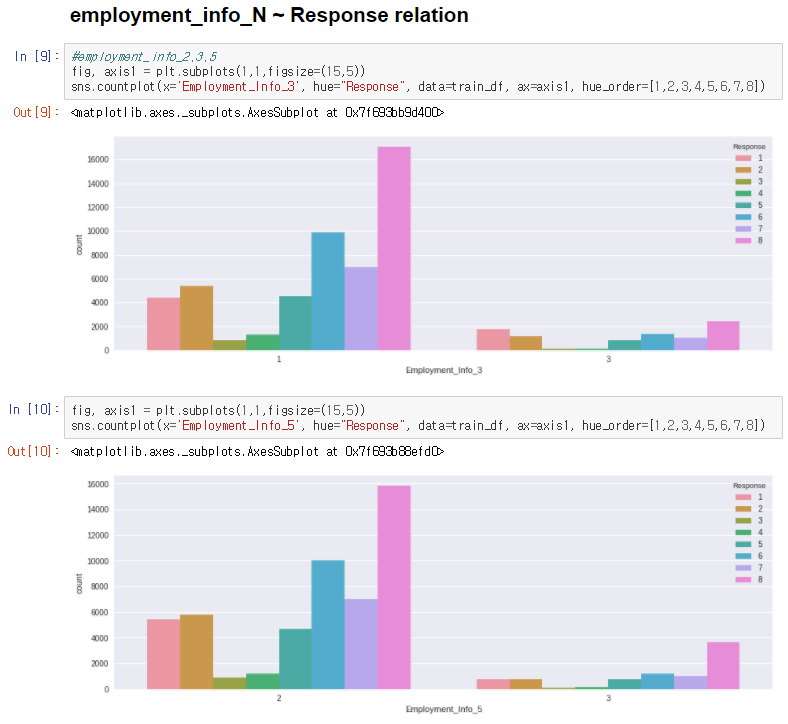


BMI & Ins\_Age : 전체적으로 크게 상관 없는듯

ggplot(train, aes(x=BMI, y= Ins\_Age)) +  
 geom\_point(shape=1) +   
 geom\_smooth(method=lm) + facet\_wrap(~ Response , ncol=2)



Employment : Response 와 비슷하게 나타남



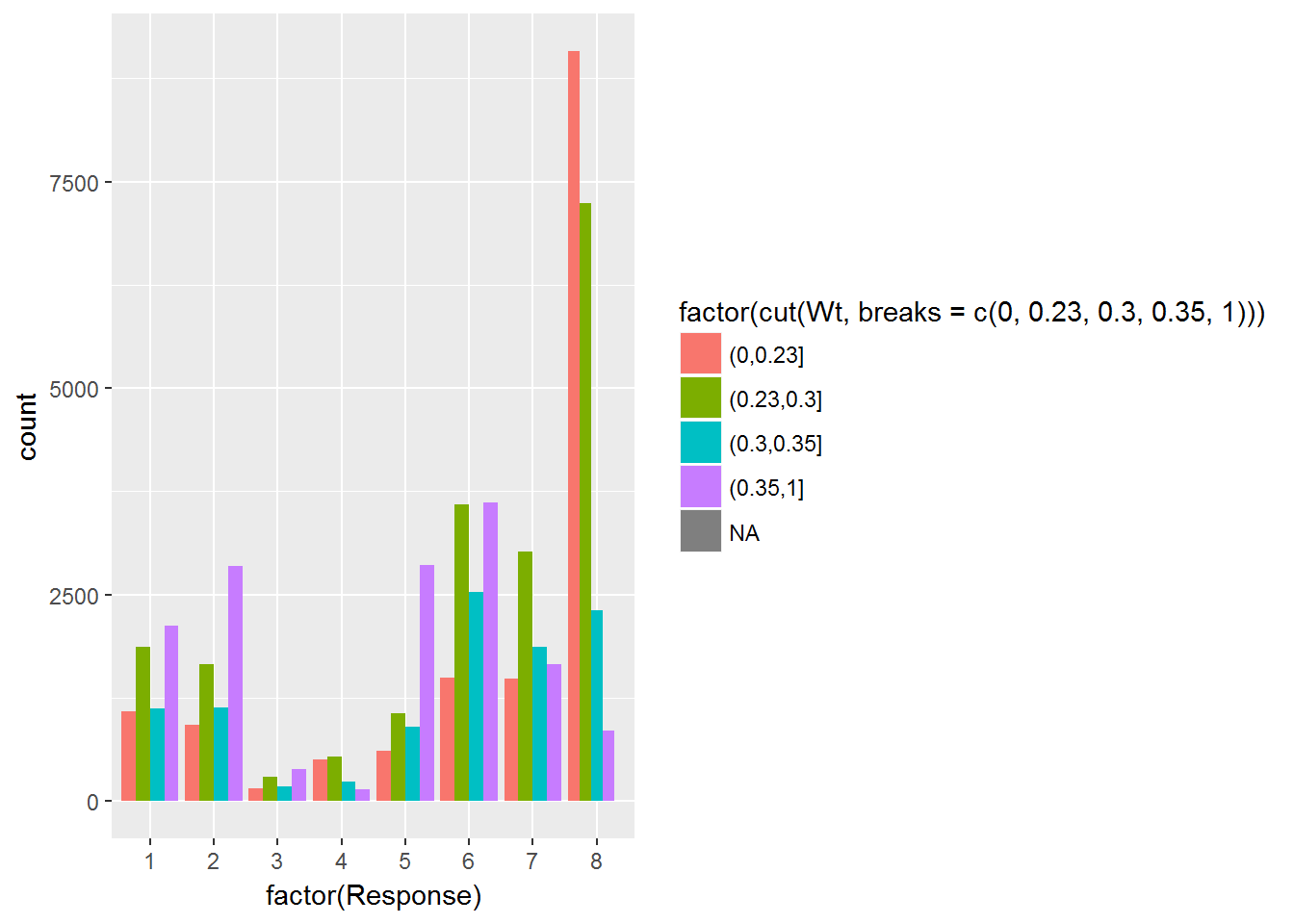
BMI\_Ins\_Age

Ht : Response 8의 경우 0~0.65 구간의 값이 가장 많음

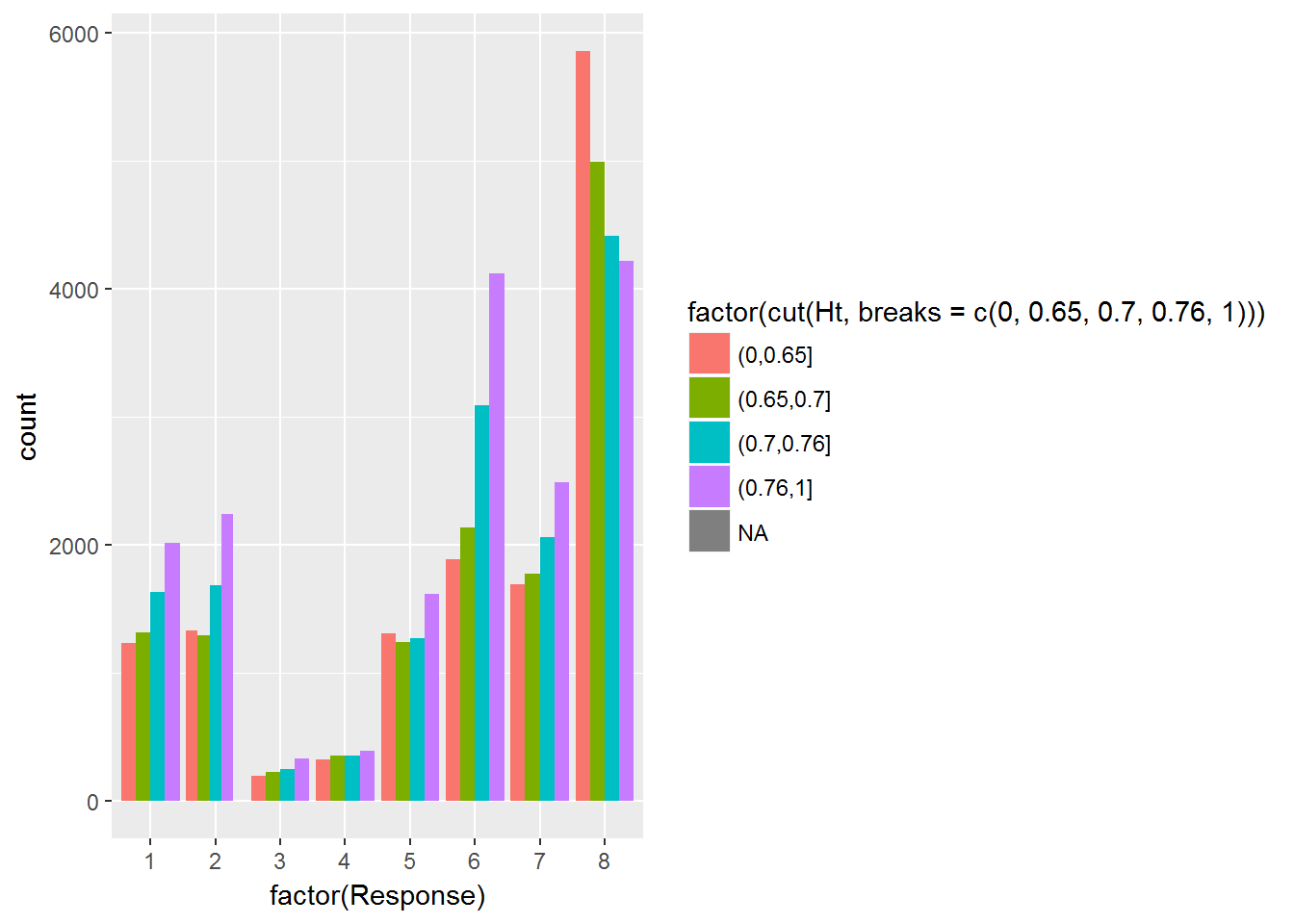
Ins\_Age : Response 8의 경우 0.2 ~04 구간이 많게 나타남

BMI : Response 8의 경우 0 ~ 0.385 구간이 역시 많음 / 0.53~1은 매우 적음

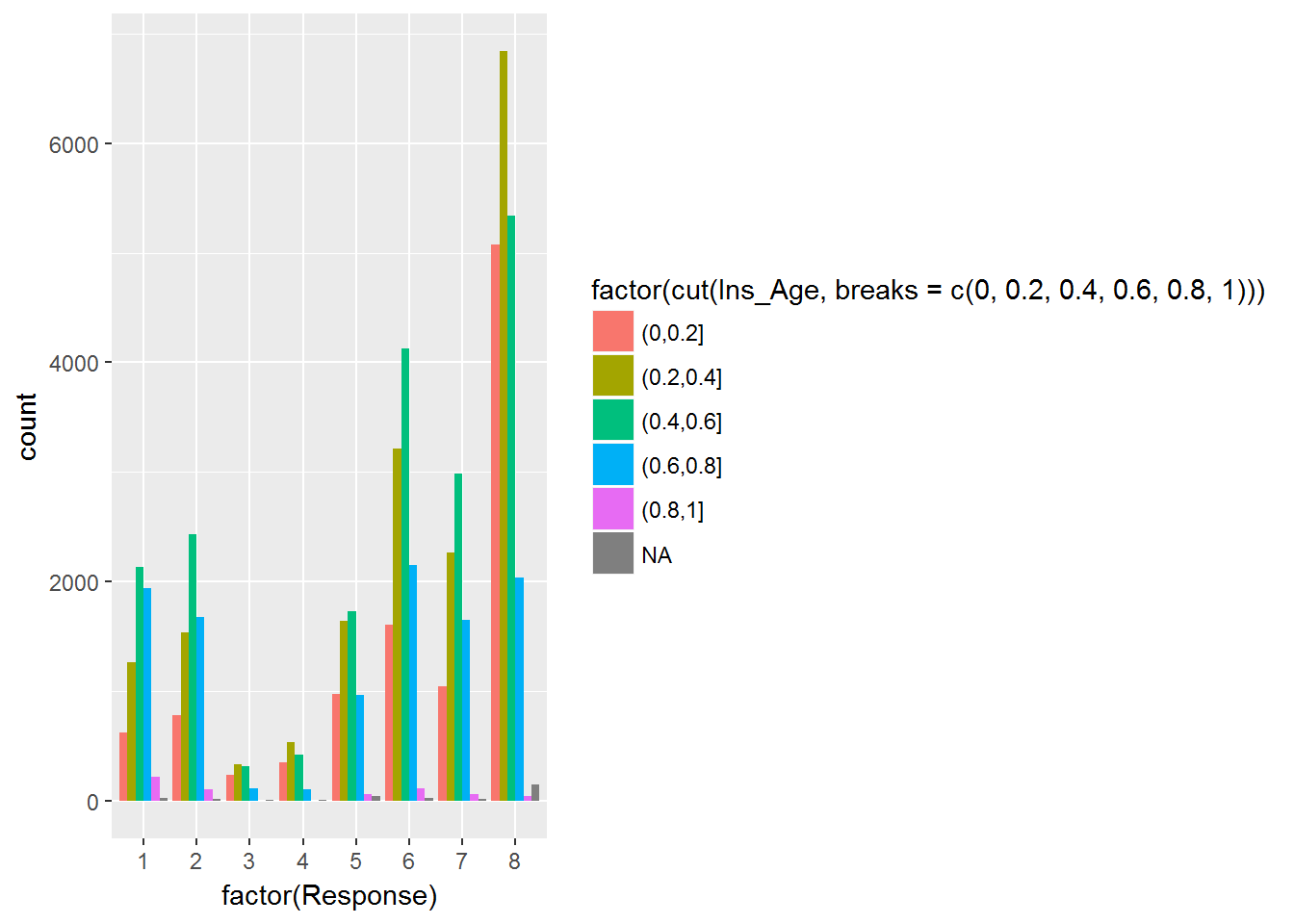
ggplot(train, aes(x=factor(Response), fill=factor(cut(Wt,breaks = c(0,.23,.3,.35,1.0)))))+  
 geom\_bar(position = "dodge")



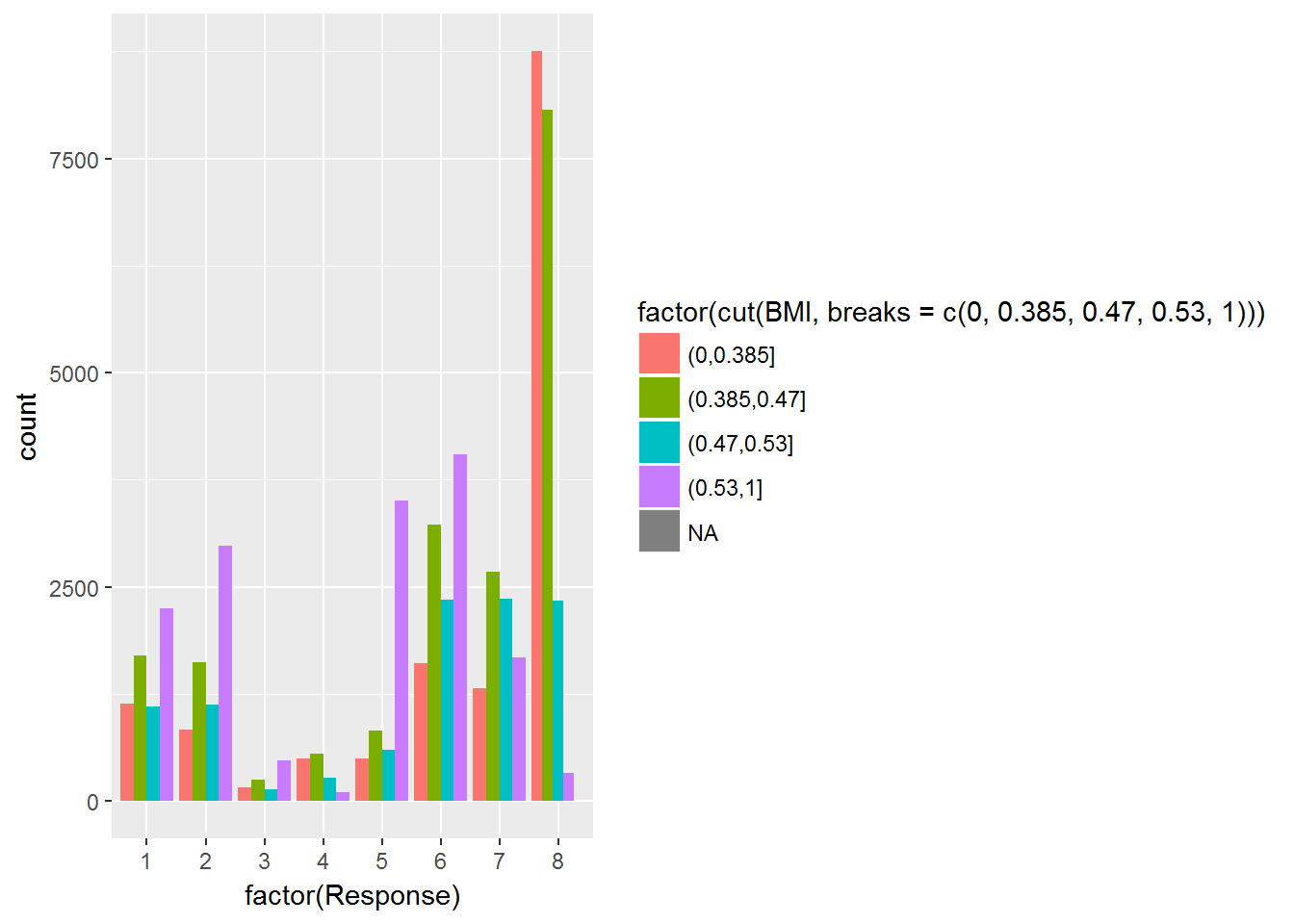
ggplot(train, aes(x=factor(Response), fill=factor(cut(Ht,breaks = c(0,.65,.7,.76,1.0)))))+  
 geom\_bar(position = "dodge")



ggplot(train, aes(x=factor(Response), fill=factor(cut(Ins\_Age,breaks = c(0,.2,.4,.6,.8,1.0)))))+  
 geom\_bar(position = "dodge")

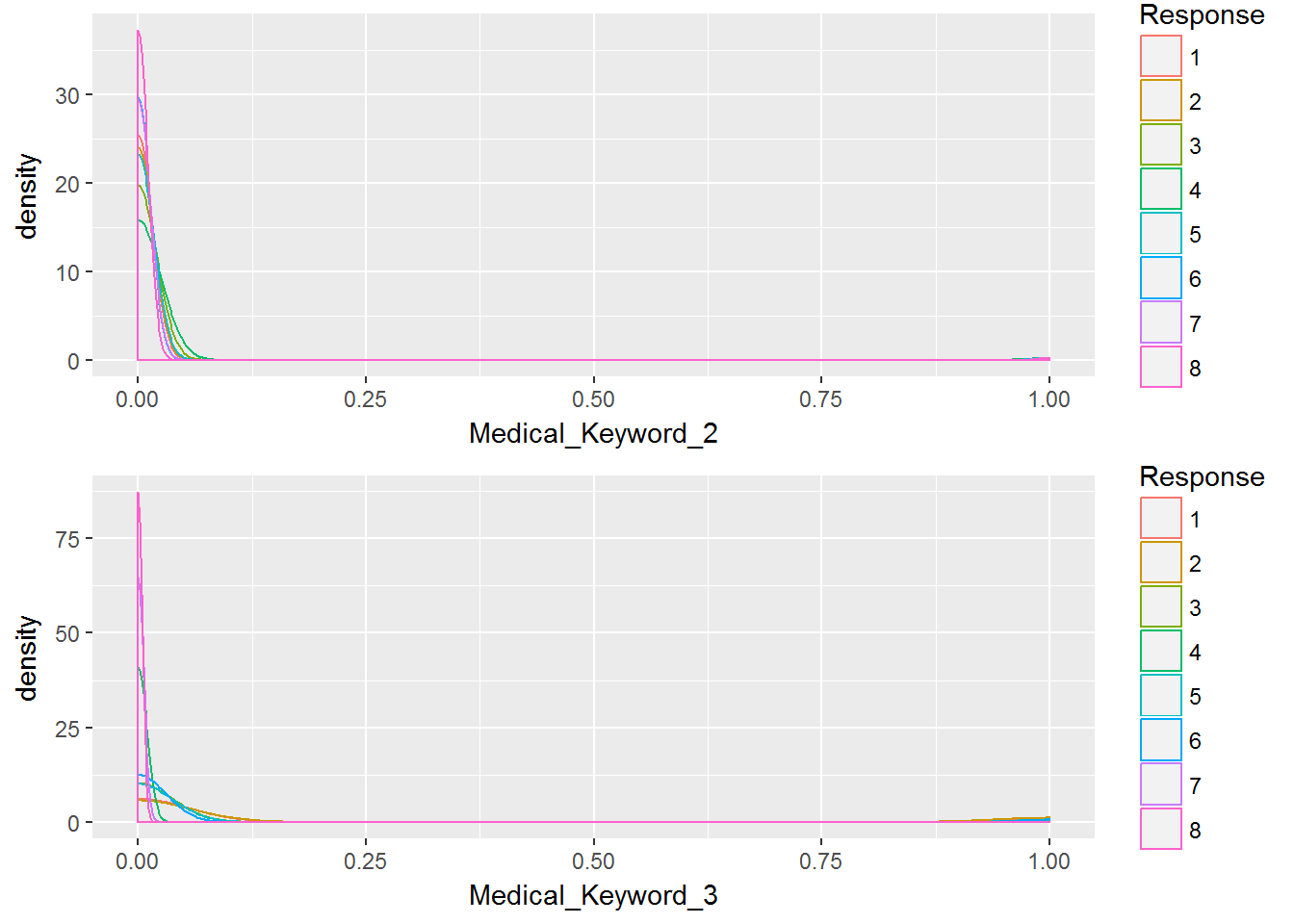


ggplot(train, aes(x=factor(Response), fill=factor(cut(BMI,breaks = c(0,.385,.47,.53,1.0)))))+  
 geom\_bar(position = "dodge")

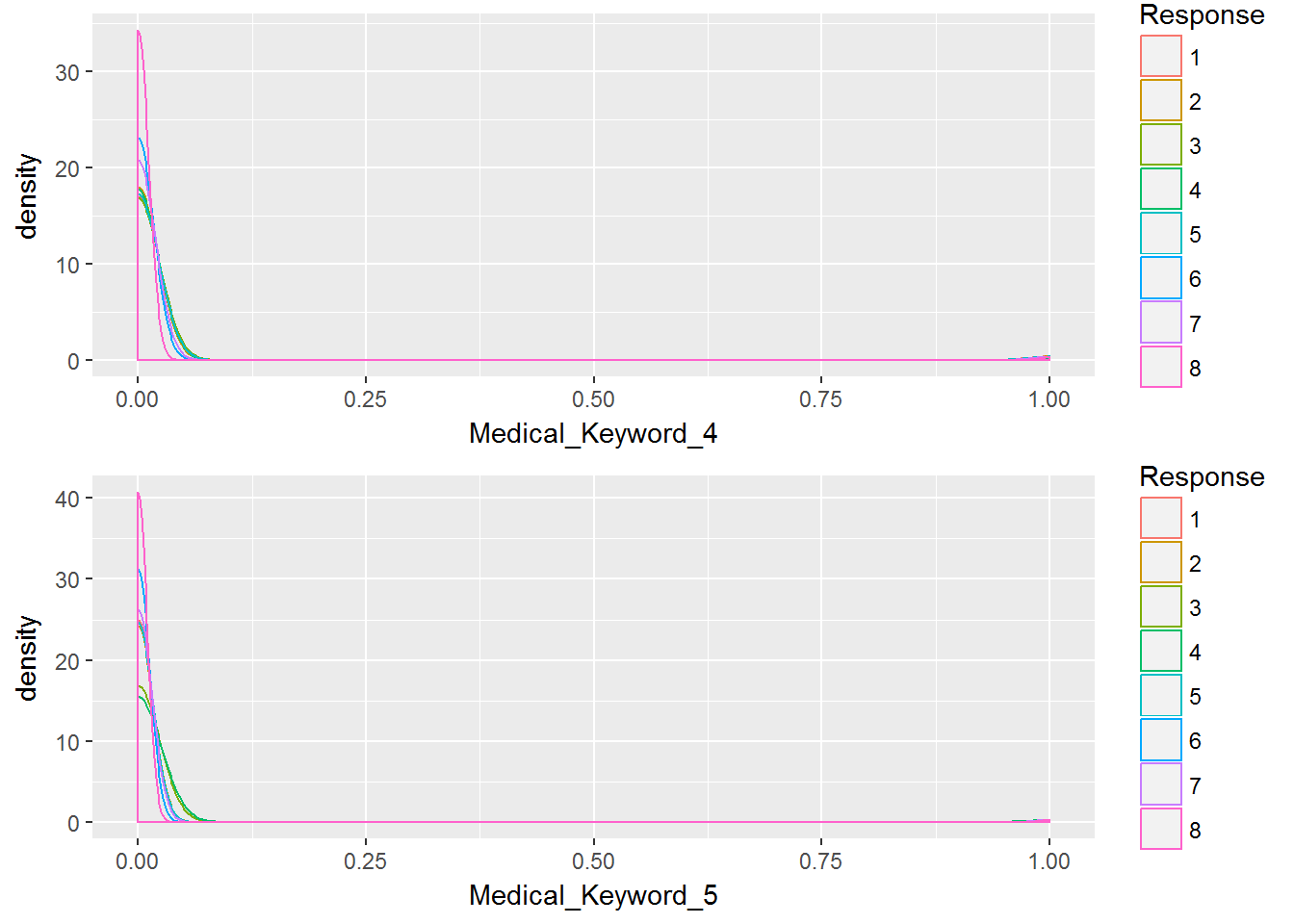


Medical\_Keyword

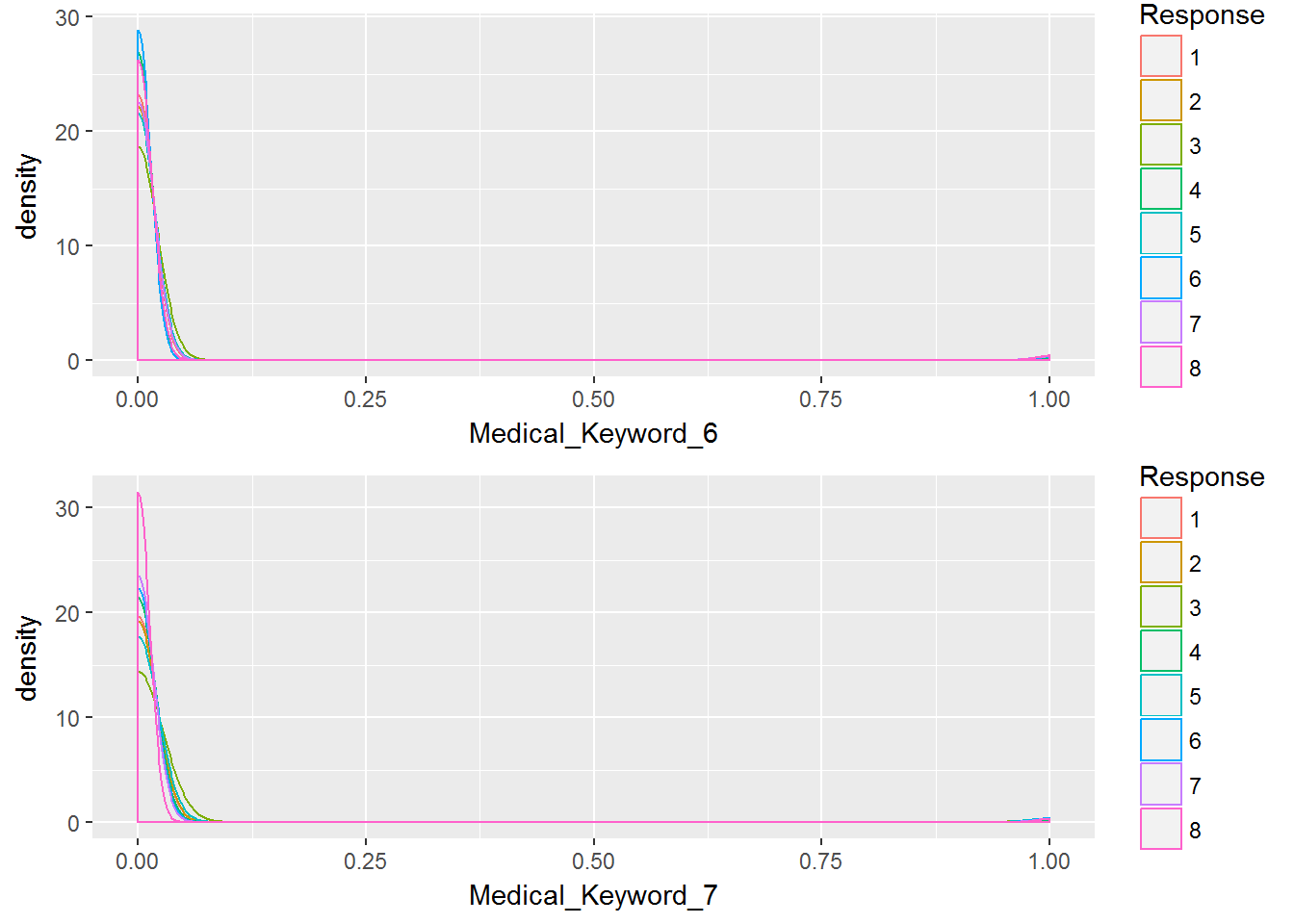
grid.arrange(ggplot(train, aes(Medical\_Keyword\_2 , colour = Response)) + geom\_density(),   
 ggplot(train, aes(Medical\_Keyword\_3 , colour = Response)) + geom\_density())



grid.arrange(ggplot(train, aes(Medical\_Keyword\_4 , colour = Response)) + geom\_density(),   
 ggplot(train, aes(Medical\_Keyword\_5 , colour = Response)) + geom\_density())

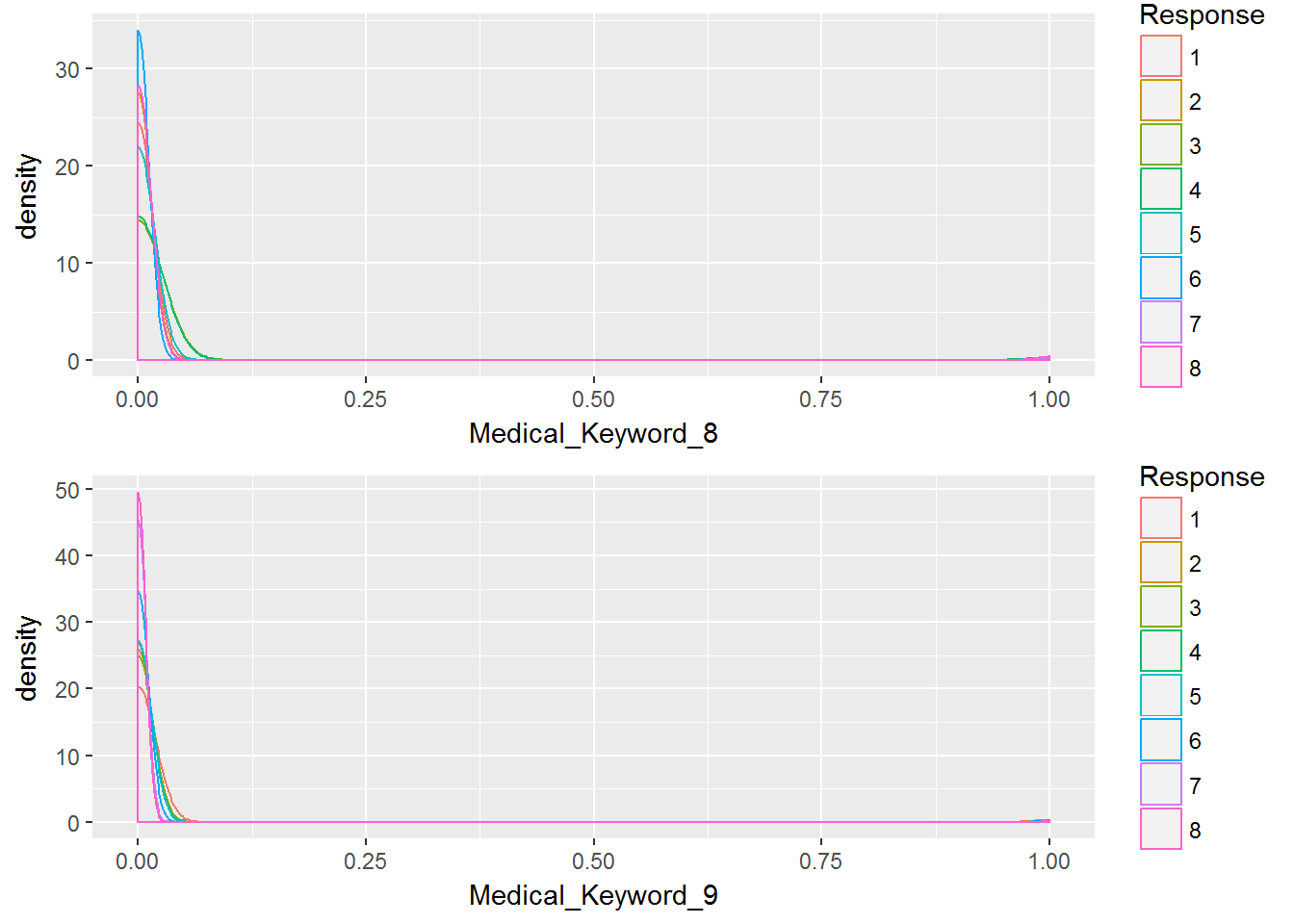


grid.arrange(ggplot(train, aes(Medical\_Keyword\_6 , colour = Response)) + geom\_density(),   
 ggplot(train, aes(Medical\_Keyword\_7 , colour = Response)) + geom\_density())



Medical\_keyword\_8 : Response5에서 높게 나타남

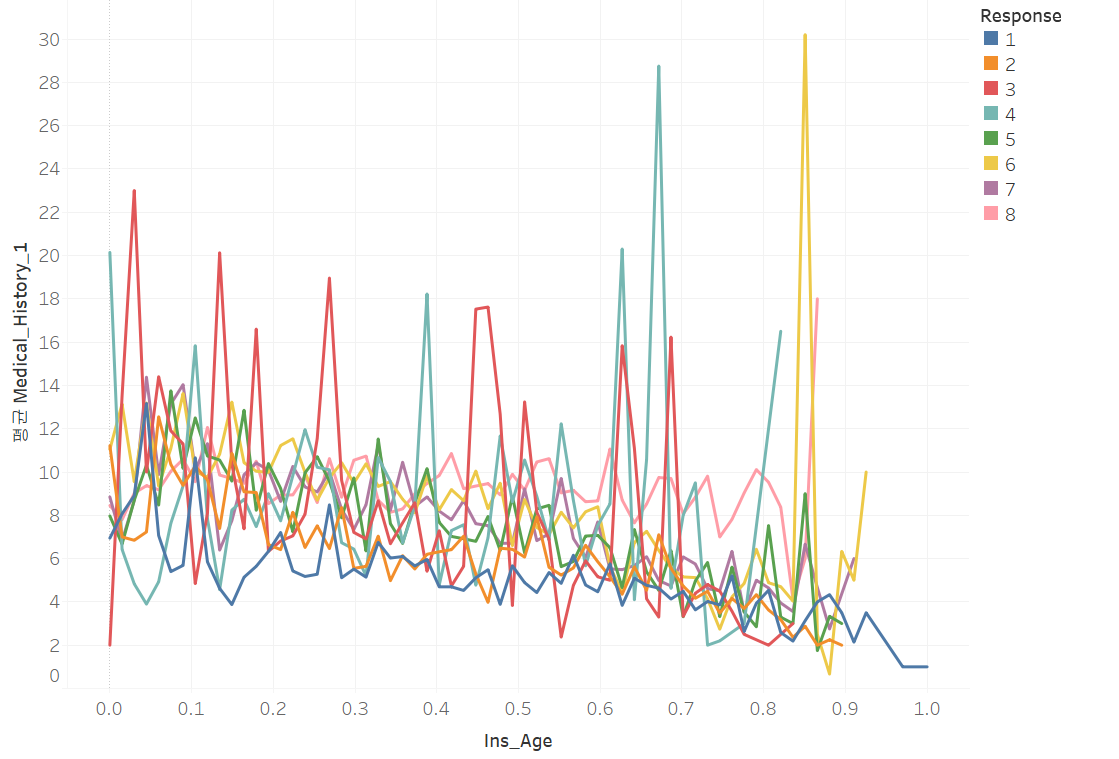
grid.arrange(ggplot(train, aes(Medical\_Keyword\_8 , colour = Response)) + geom\_density(),   
 ggplot(train, aes(Medical\_Keyword\_9 , colour = Response)) + geom\_density())



Int\_Age & Medical\_History1 : Ins\_Age가 높고 Medical History1이 높을 수록 Response는 6이 많다

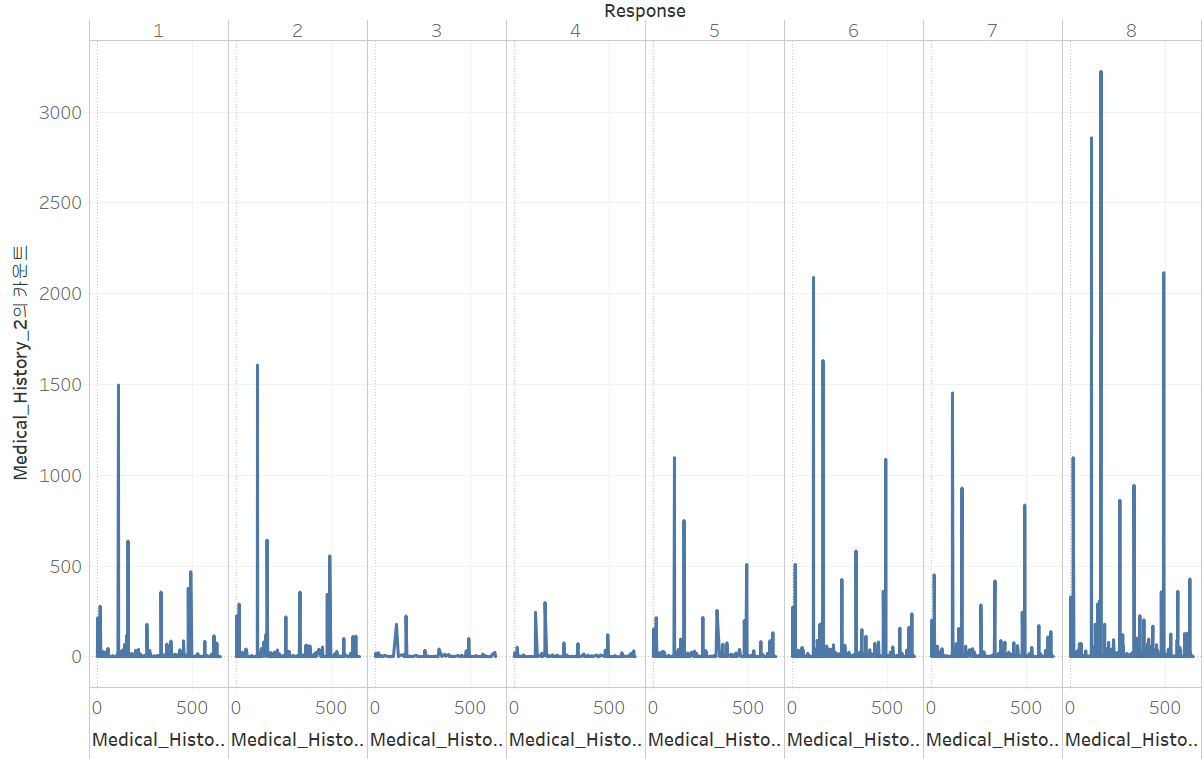
6예측 용이, 0.6 ~0.7은 4가 많음

특히나 Int\_Age\_MH1에 따른 편차가 존재 / Response 5는 Ins\_Age가 커질수록 빈도가 줄어드는 경향이 있음



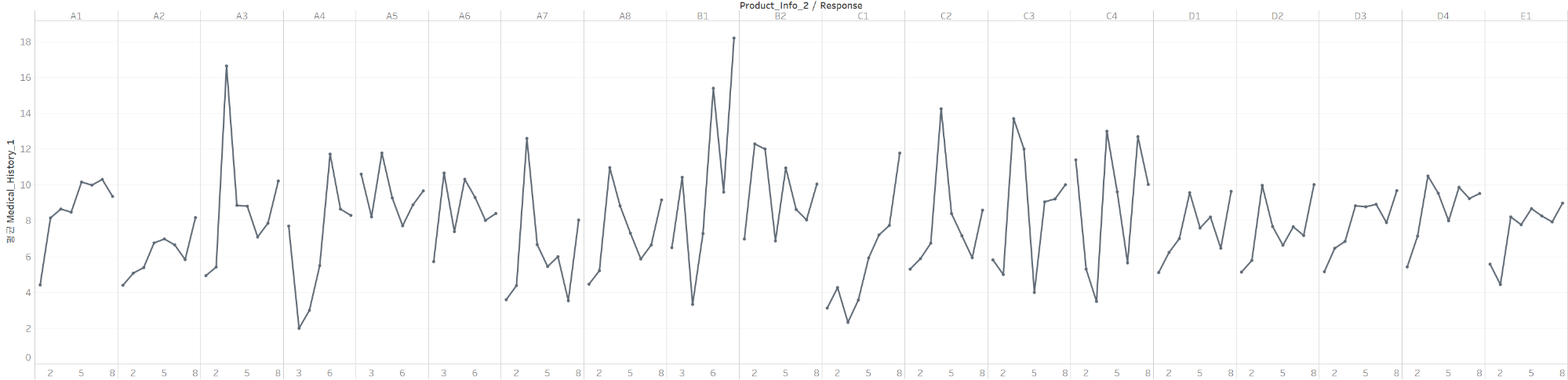
Int\_Age\_MH1

Medical\_History2 : Response에 따라 편차 존재



MH2\_Count

Product Information2 & Medical History1 : 상품에 따라 등급이 다양하게 나타남



PI2\_MH1

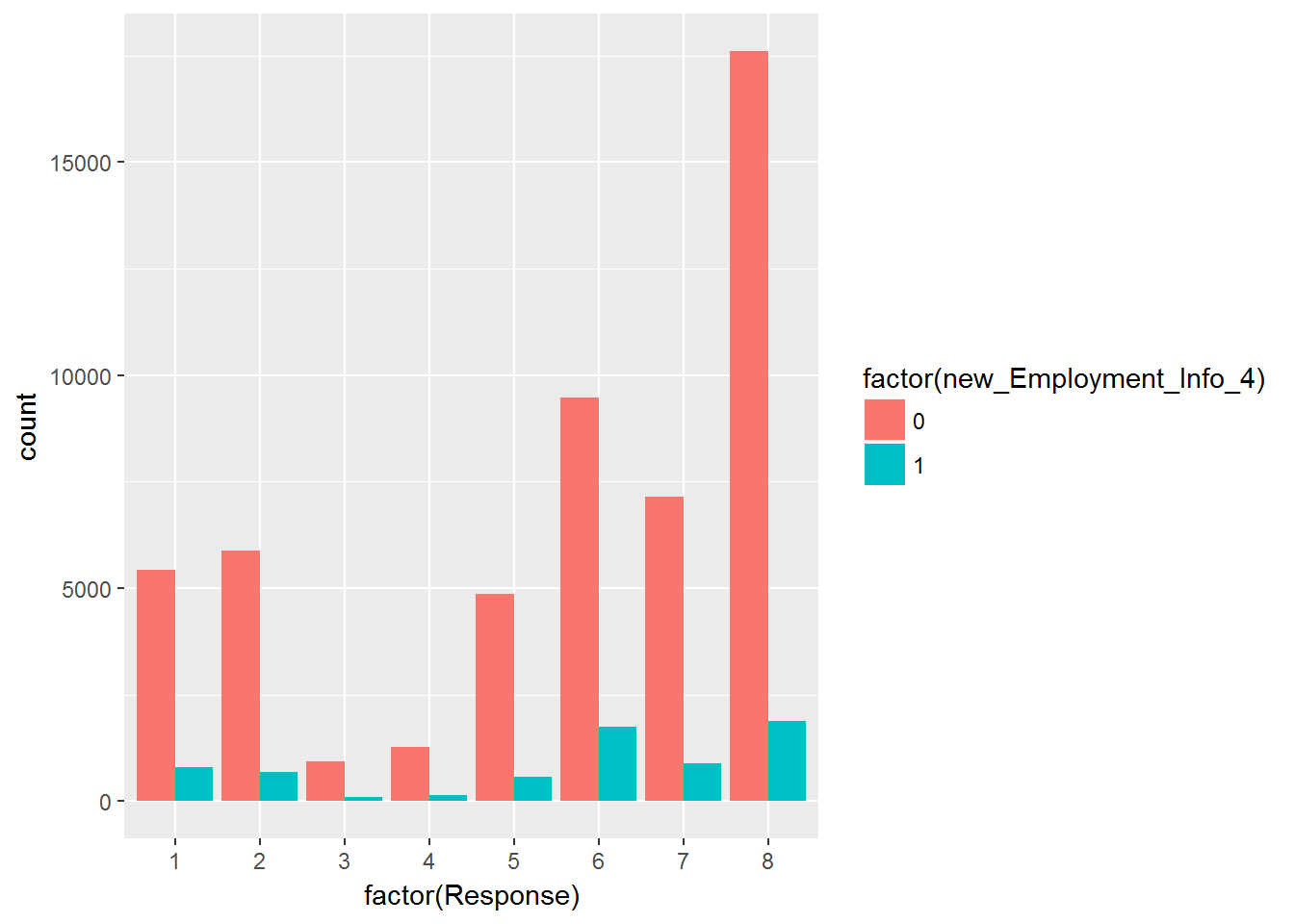
많은 na를 가지고 있는 변수들이 있다는 것

어떤 변수는 na값들이 더 많은 경우도 있다. 그래서 na의 값과 na가 아닌 값들을 비교했다.

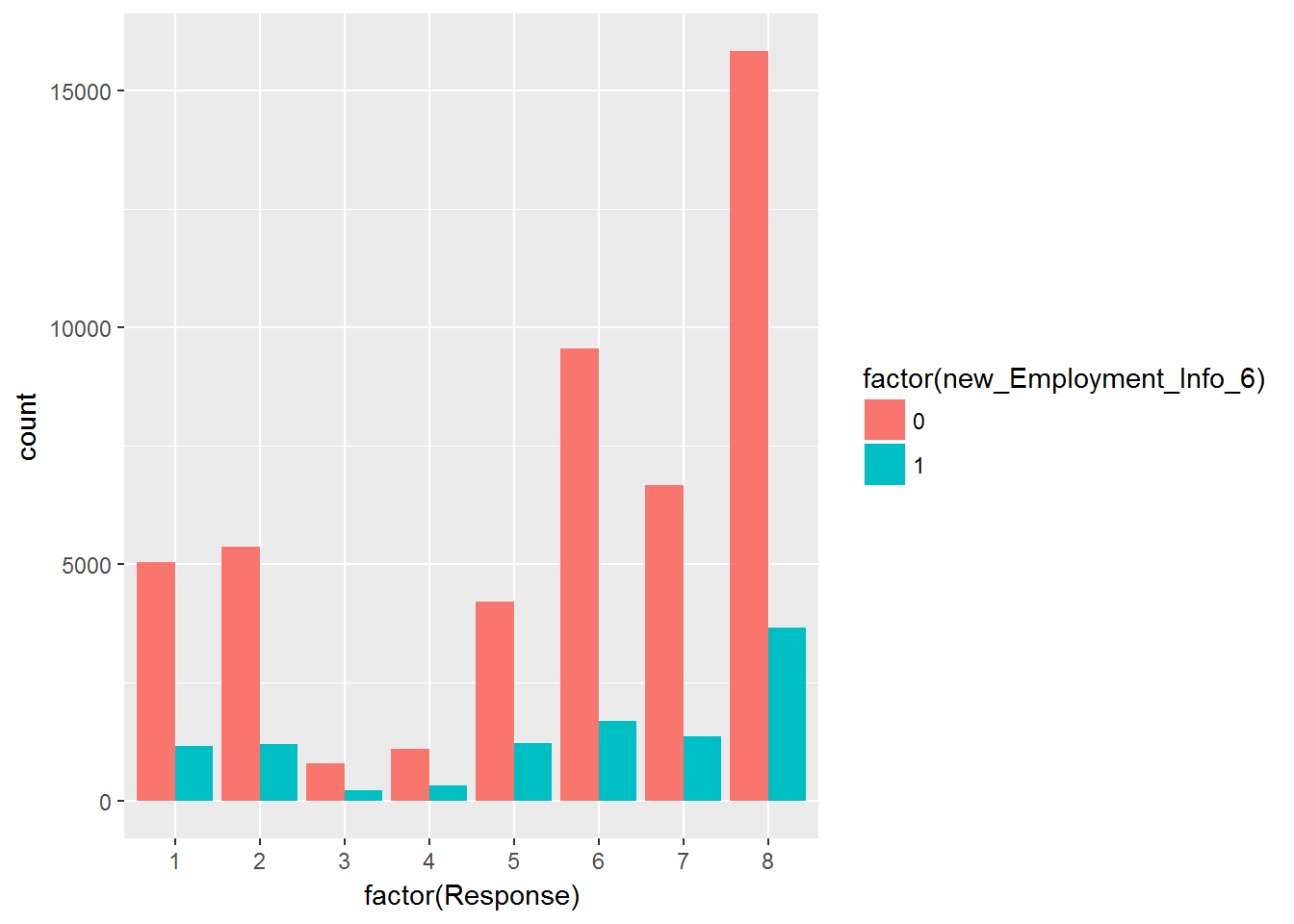
train %>%  
 mutate(new\_Employment\_Info\_4 = as.numeric(is.na(Employment\_Info\_4))) %>%  
 mutate(new\_Employment\_Info\_6 = as.numeric(is.na(Employment\_Info\_6))) %>%  
 mutate(new\_Insurance\_History\_5 = as.numeric(is.na(Insurance\_History\_5))) %>%  
 mutate(new\_Family\_Hist\_2 = as.numeric(is.na(Family\_Hist\_2))) %>%  
 mutate(new\_Family\_Hist\_3 = as.numeric(is.na(Family\_Hist\_3))) %>%  
 mutate(new\_Family\_Hist\_4 = as.numeric(is.na(Family\_Hist\_4))) %>%  
 mutate(new\_Family\_Hist\_5 = as.numeric(is.na(Family\_Hist\_5))) %>%  
 mutate(new\_Medical\_History\_1 = as.numeric(is.na(Medical\_History\_1))) %>%  
 mutate(new\_Medical\_History\_10 = as.numeric(is.na(Medical\_History\_10))) %>%  
 mutate(new\_Medical\_History\_15 = as.numeric(is.na(Medical\_History\_15))) %>%  
 mutate(new\_Medical\_History\_24 = as.numeric(is.na(Medical\_History\_24))) %>%  
 mutate(new\_Medical\_History\_32 = as.numeric(is.na(Medical\_History\_32))) -> train

na(1)와 not-na(0)의 비교 시각화

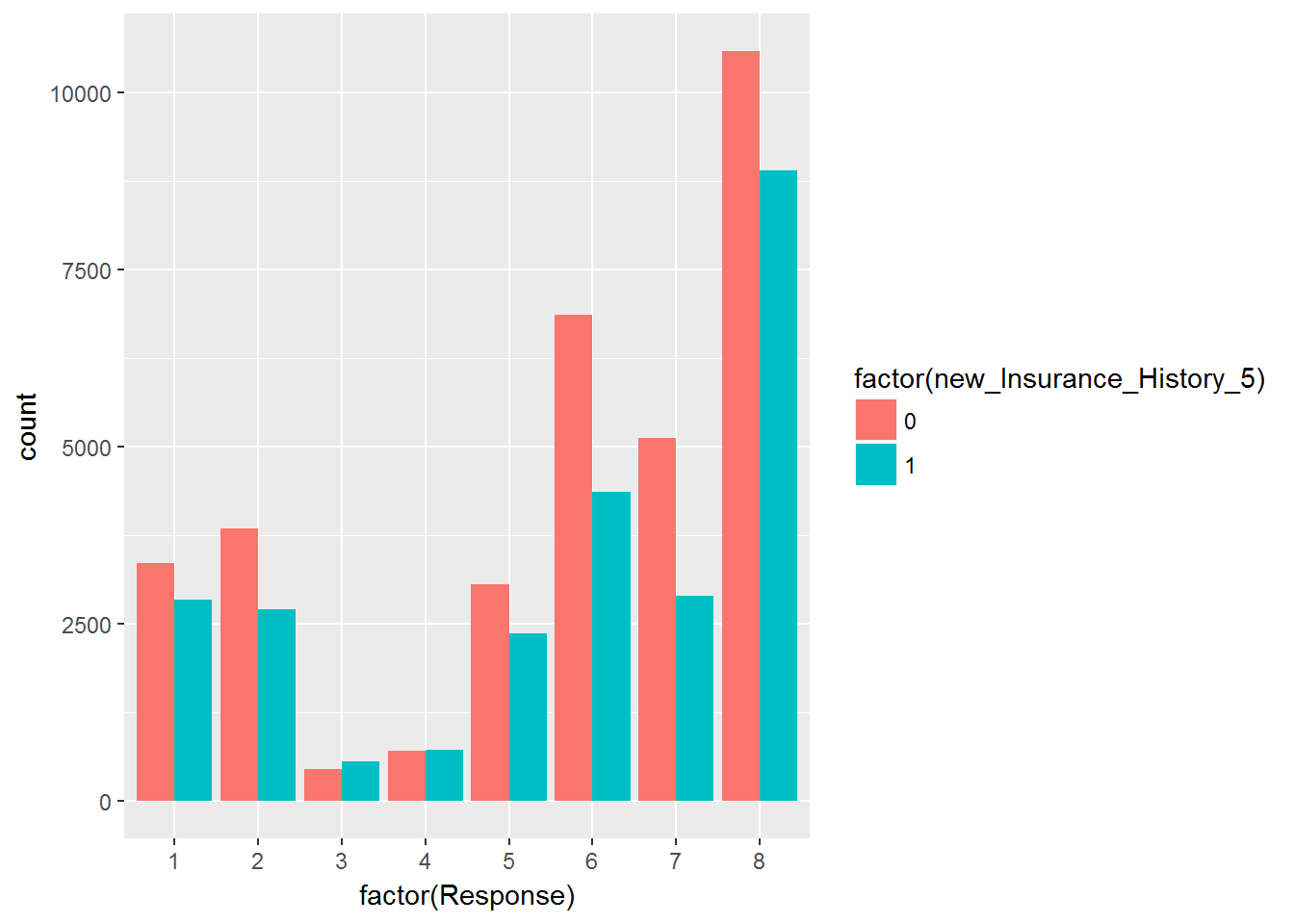
ggplot(data = train, aes(factor(Response), fill = factor(new\_Employment\_Info\_4))) +  
 geom\_bar(position = "dodge")



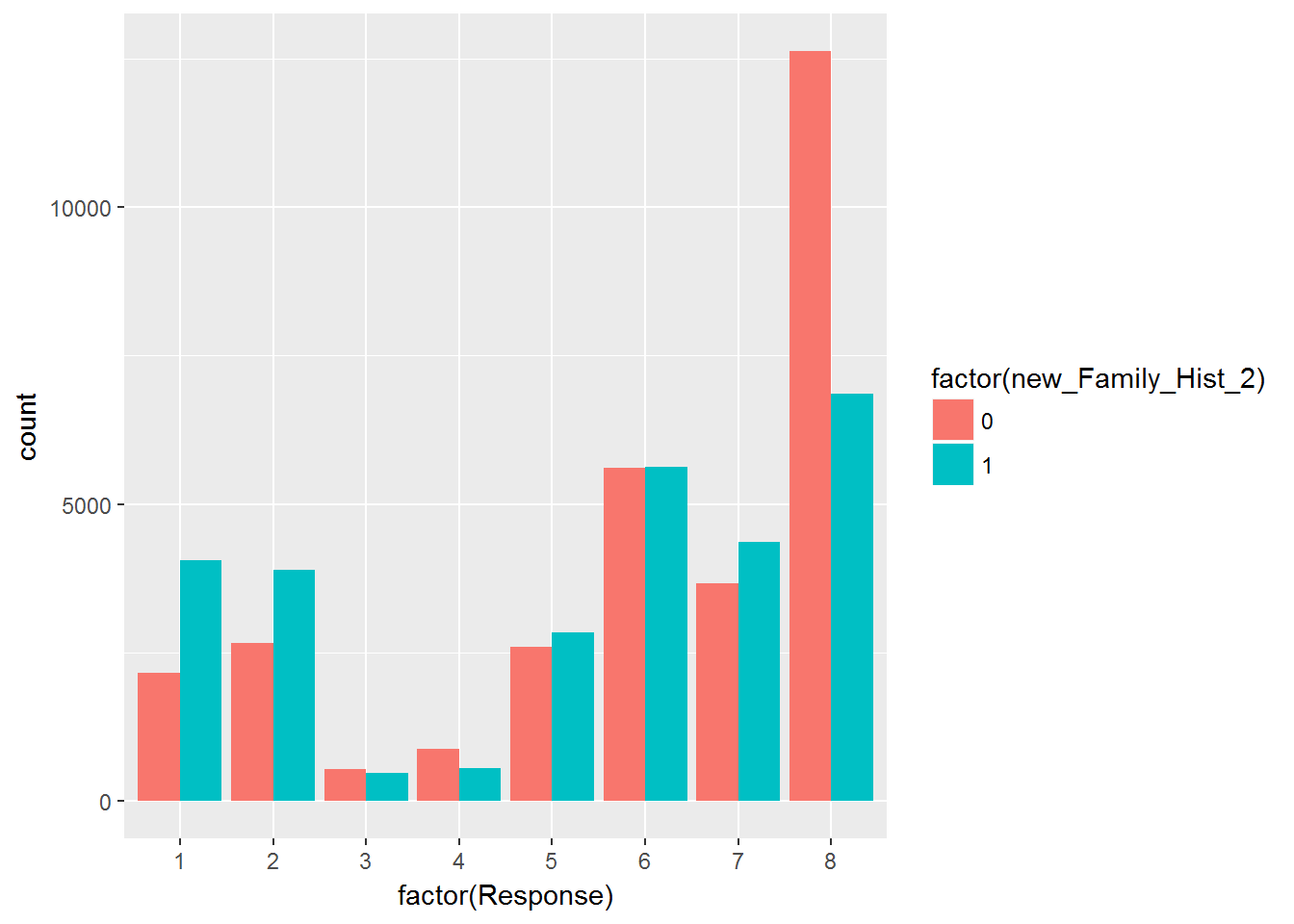
ggplot(data = train, aes(factor(Response), fill = factor(new\_Employment\_Info\_6))) +  
 geom\_bar(position = "dodge")



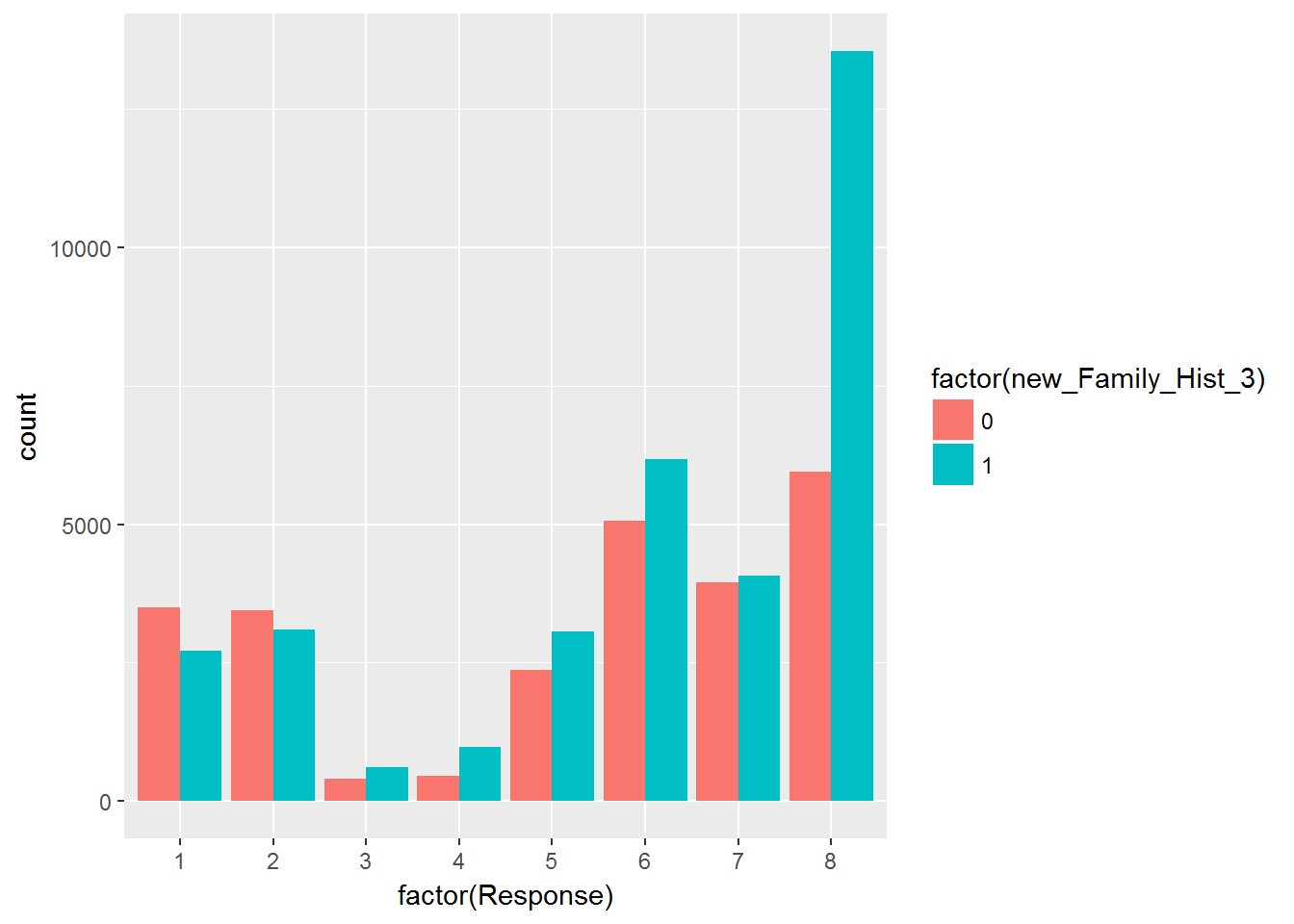
ggplot(data = train, aes(factor(Response), fill = factor(new\_Insurance\_History\_5))) +  
 geom\_bar(position = "dodge")



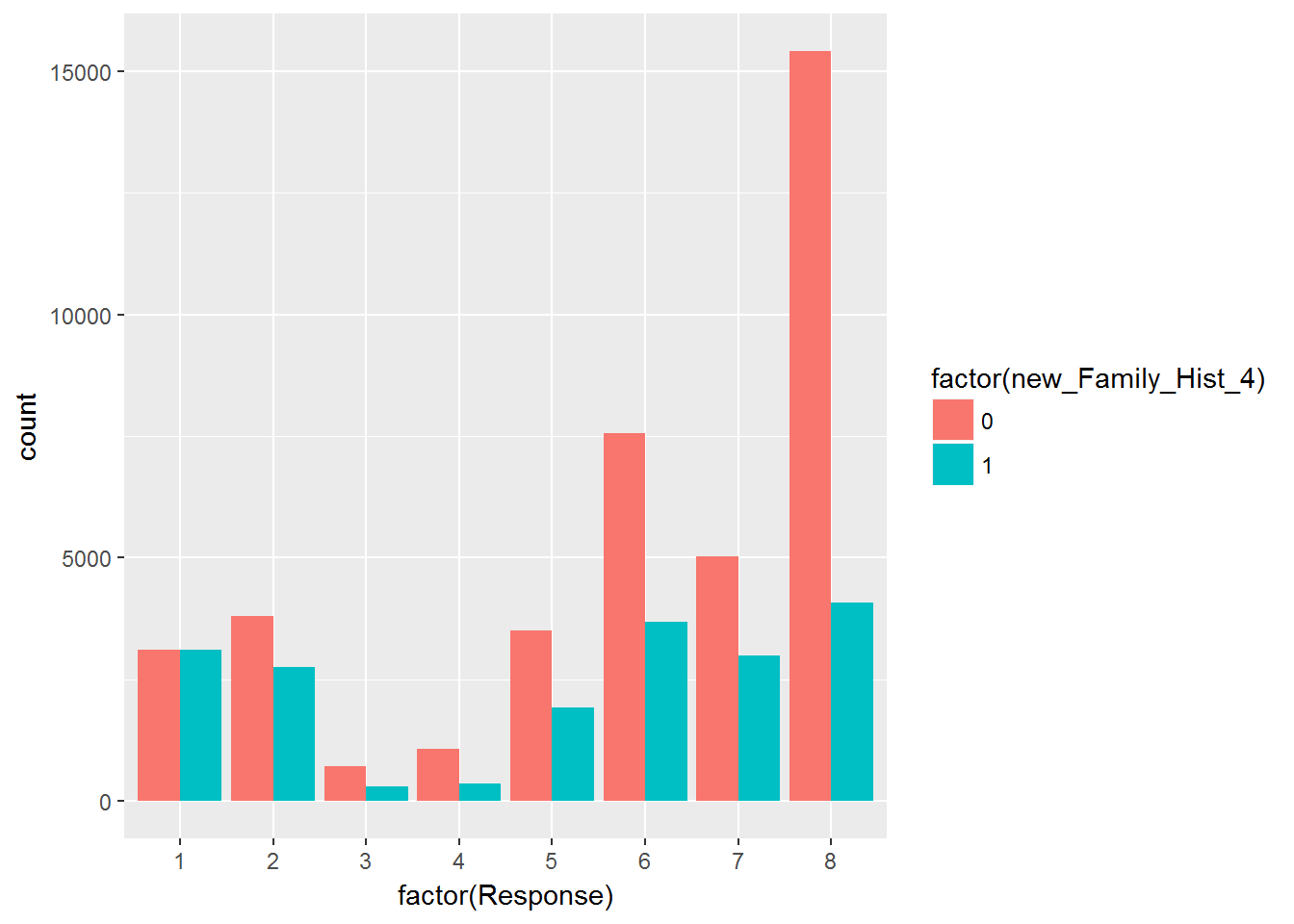
ggplot(data = train, aes(factor(Response), fill = factor(new\_Family\_Hist\_2))) +  
 geom\_bar(position = "dodge")



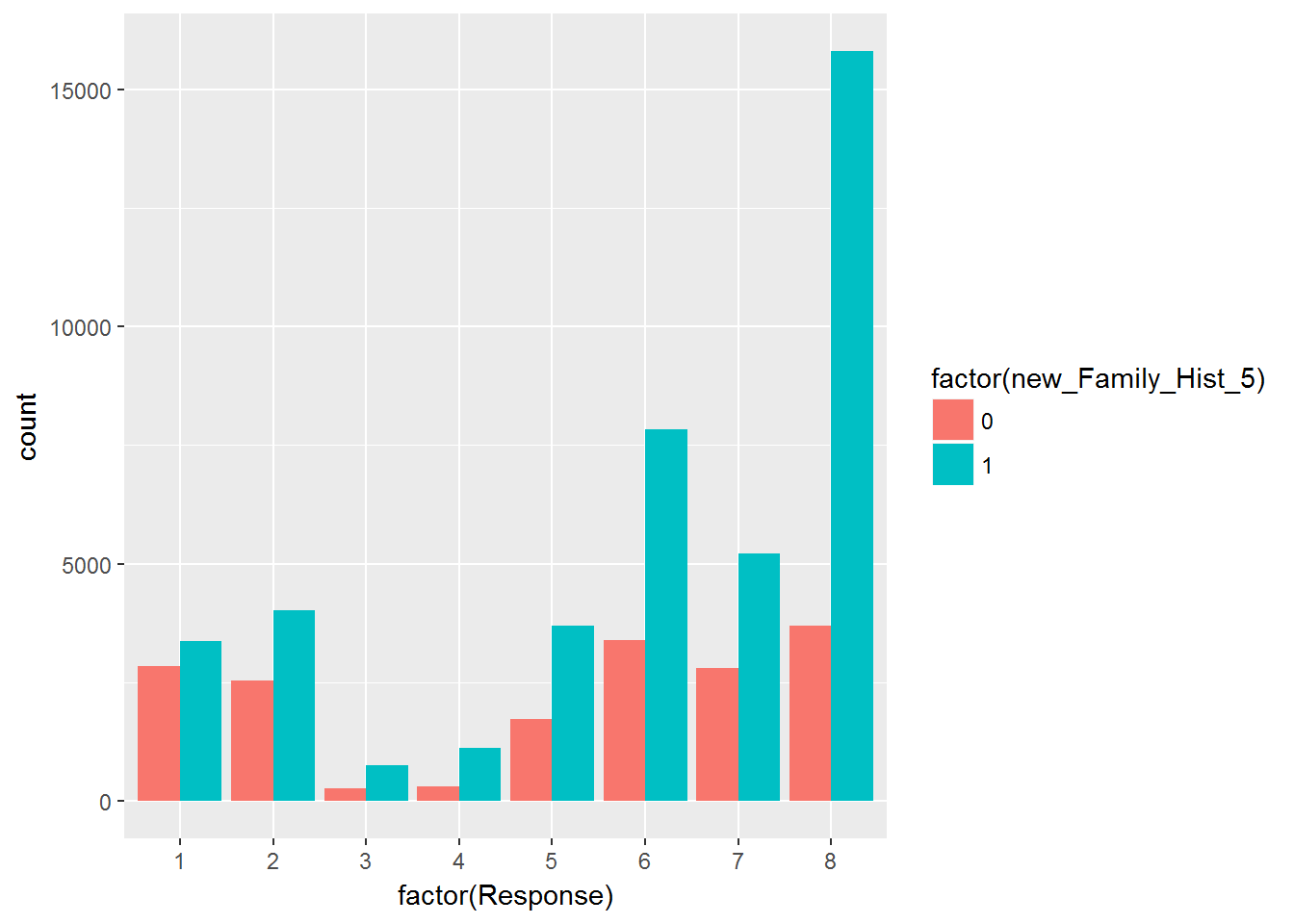
ggplot(data = train, aes(factor(Response), fill = factor(new\_Family\_Hist\_3))) +  
 geom\_bar(position = "dodge")



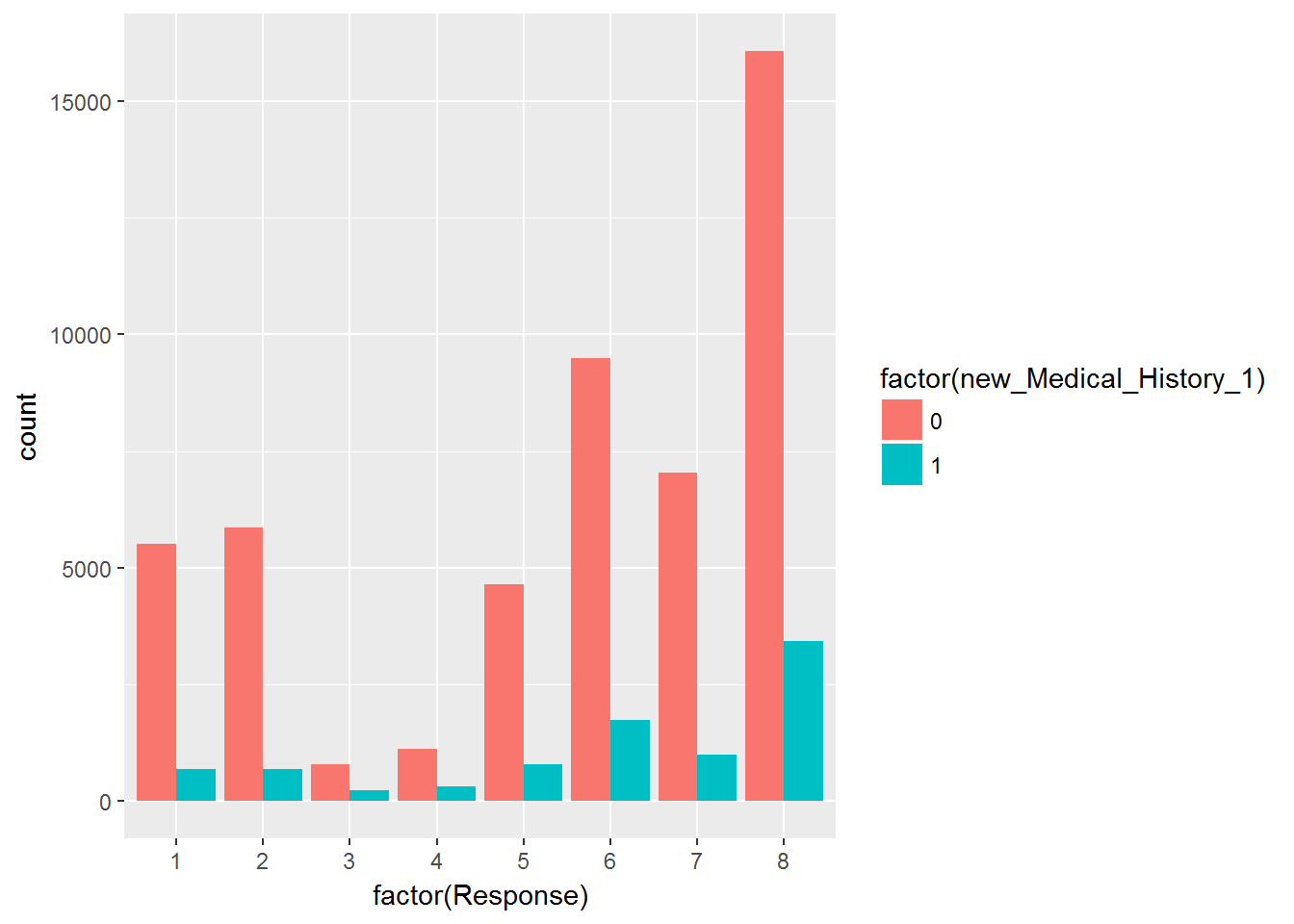
ggplot(data = train, aes(factor(Response), fill = factor(new\_Family\_Hist\_4))) +  
 geom\_bar(position = "dodge")



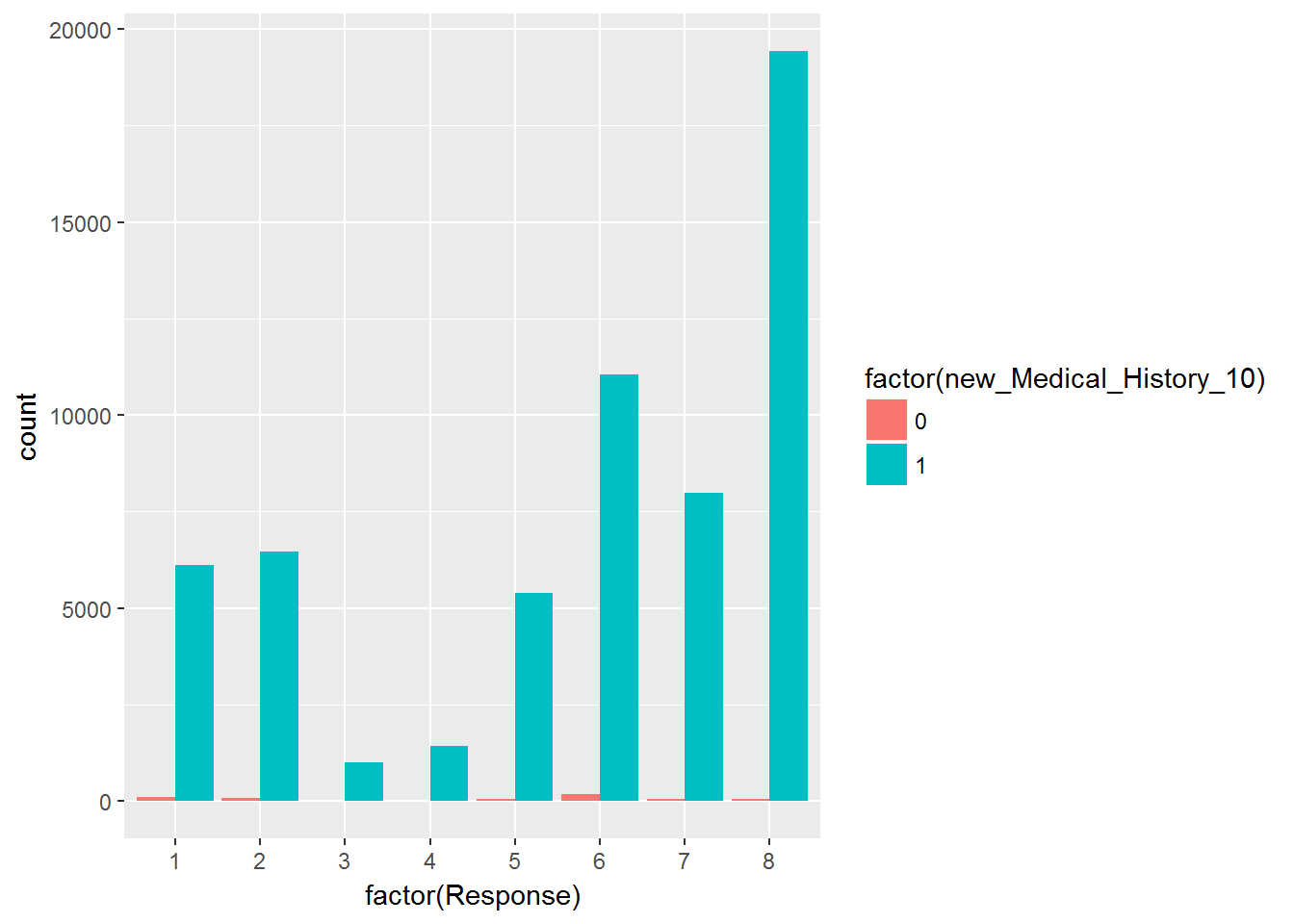
ggplot(data = train, aes(factor(Response), fill = factor(new\_Family\_Hist\_5))) +  
 geom\_bar(position = "dodge")



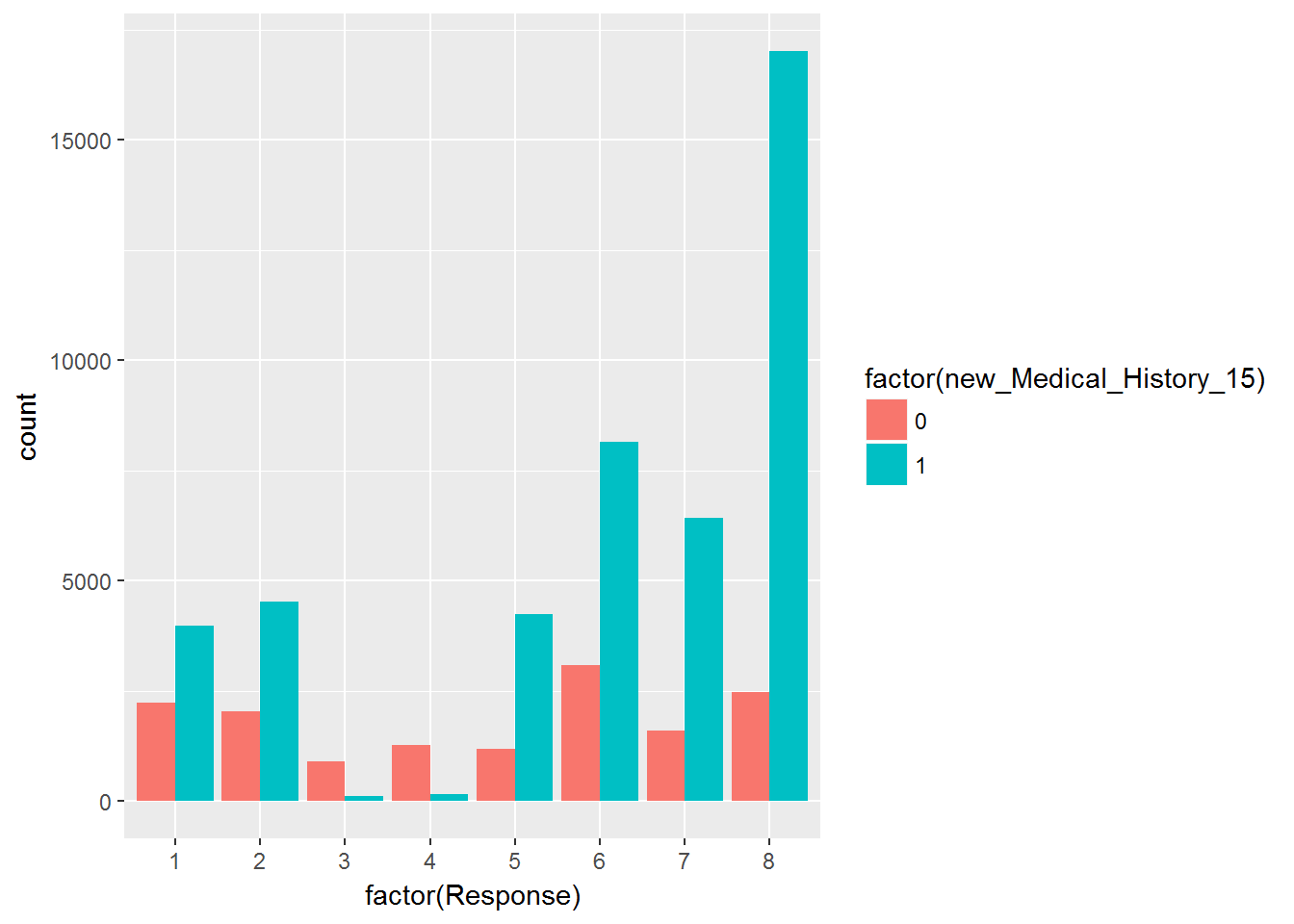
ggplot(data = train, aes(factor(Response), fill = factor(new\_Medical\_History\_1))) +  
 geom\_bar(position = "dodge")



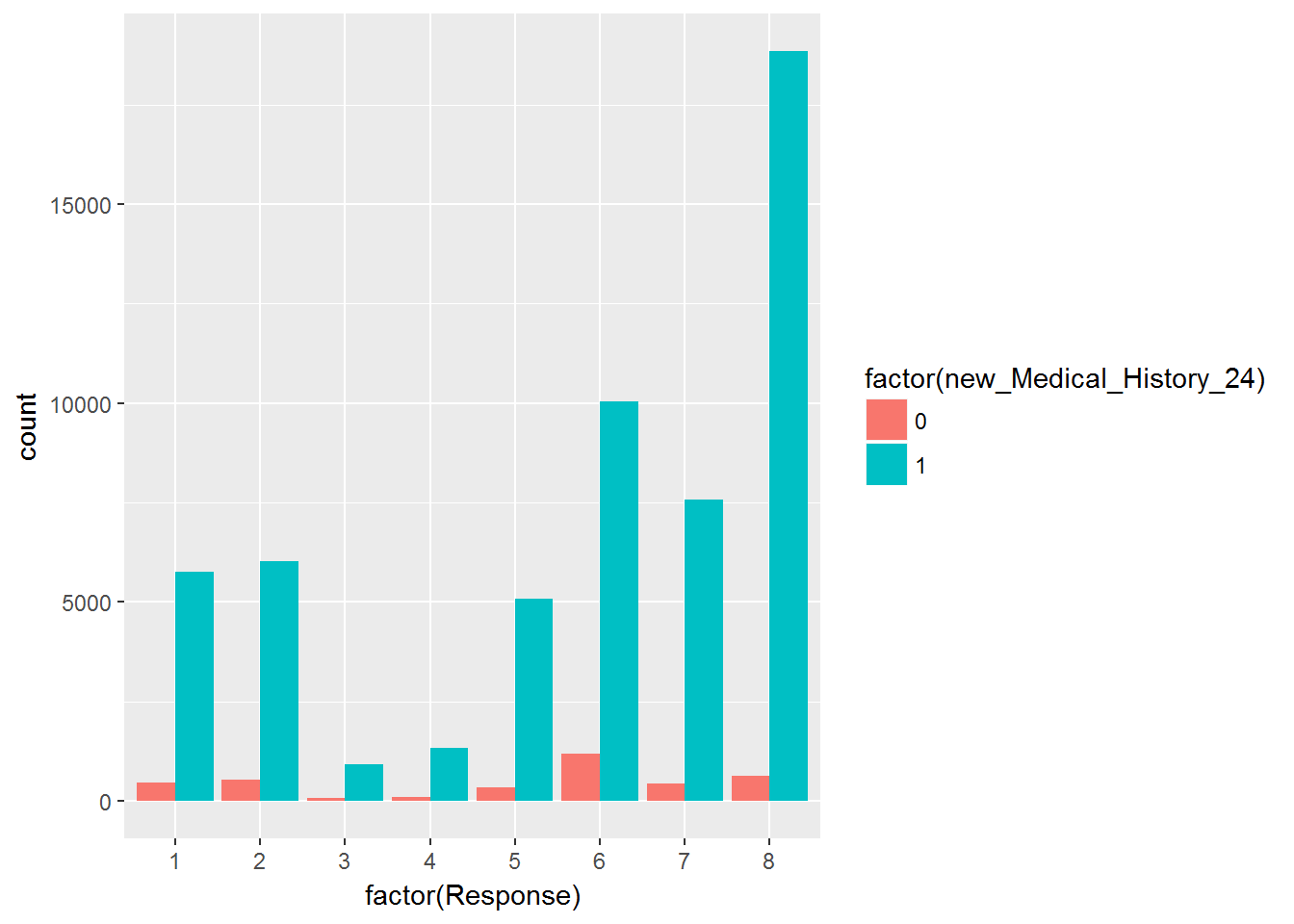
ggplot(data = train, aes(factor(Response), fill = factor(new\_Medical\_History\_10))) +  
 geom\_bar(position = "dodge")



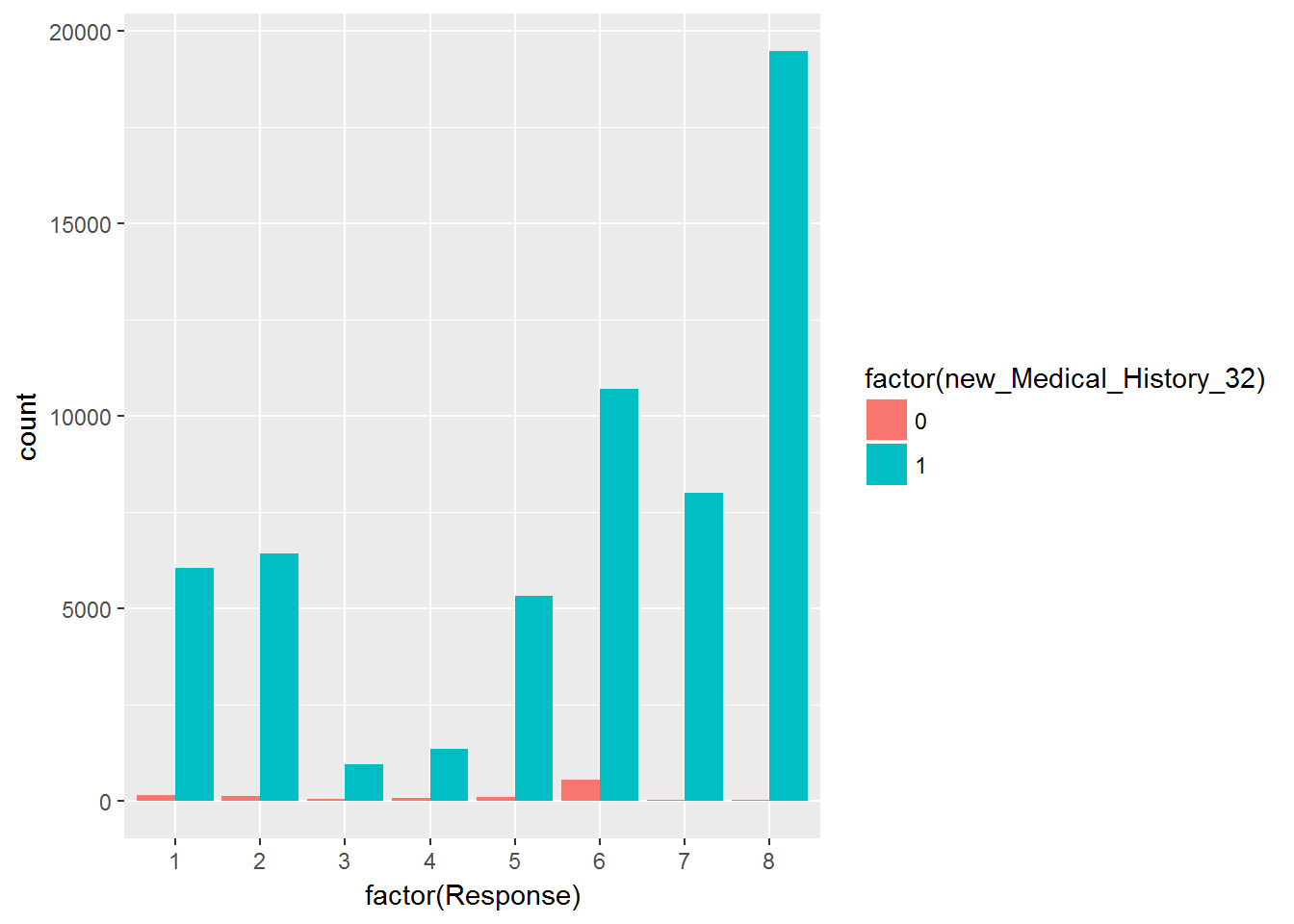
ggplot(data = train, aes(factor(Response), fill = factor(new\_Medical\_History\_15))) +  
 geom\_bar(position = "dodge")



ggplot(data = train, aes(factor(Response), fill = factor(new\_Medical\_History\_24))) +  
 geom\_bar(position = "dodge")



ggplot(data = train, aes(factor(Response), fill = factor(new\_Medical\_History\_32))) +  
 geom\_bar(position = "dodge")



is.na 변수 갯수 히스토그램

qplot(rowSums(is.na(train)), geom="histogram",binwidth = 1, col=I("red"))

