다변량 자료분석 및 실험

당신의 정신은 건강한가요?

3조

통계학과 2014097022 김준섭

통계학과 2014024028 이준호

통계학과 2016111701 이승미

CONTENTS

1. 데이터설명

2. 탐색적 자료분석

3. 자료분석

4. 모델링

5. 결론

1. 데이터설명_dataset

1259개의 응답. 27개의 변수

Timestamp	Age	Gender	Country	state	self_employed	family_history	treatment	work_interfere	no_employees	remote_work	tech_company	benefits
2014-08-27 11:29	37	Female	United States	IL	NA	No	Yes	Often	06-25	No	Yes	Yes
2014-08-27 11:29	44	M	United States	IN	NA	No	No	Rarely	More than 1000	No	No	Don't know
2014-08-27 11:29	32	Male	Canada	NA	NA	No	No	Rarely	06-25	No	Yes	No
2014-08-27 11:29	31	Male	United Kingdom	NA	NA	Yes	Yes	Often	26-100	No	Yes	No
2014-08-27 11:30	31	Male	United States	TX	NA	No	No	Never	100-500	Yes	Yes	Yes
2014-08-27 11:31	33	Male	United States	TN	NA	Yes	No	Sometimes	06-25	No	Yes	Yes
2014-08-27 11:31	35	Female	United States	MI	NA	Yes	Yes	Sometimes	01-05	Yes	Yes	No
2014-08-27 11:32	39	M	Canada	NA	NA	No	No	Never	01-05	Yes	Yes	No
2014-08-27 11:32	42	Female	United States	IL	NA	Yes	Yes	Sometimes	100-500	No	Yes	Yes
2014-08-27 11:32	23	Male	Canada	NA	NA	No	No	Never	26-100	No	Yes	Don't know
2014-08-27 11:32	31	Male	United States	OH	NA	No	Yes	Sometimes	06-25	Yes	Yes	Don't know
2014-08-27 11:32	29	male	Bulgaria	NA	NA	No	No	Never	100-500	Yes	Yes	Don't know
2014-08-27 11:33	42	female	United States	CA	NA	Yes	Yes	Sometimes	26-100	No	No	Yes
2014-08-27 11:33	36	Male	United States	CT	NA	Yes	No	Never	500-1000	No	Yes	Don't know
2014-08-27 11:33	27	Male	Canada	NA	NA	No	No	Never	06-25	No	Yes	Don't know
2014-08-27 11:34	29	female	United States	IL	NA	Yes	Yes	Rarely	26-100	No	Yes	Yes
2014-08-27 11:34	23	Male	United Kingdom	NA	NA	No	Yes	Sometimes	26-100	Yes	Yes	Don't know

https://www.kaggle.com/osmi/mental-health-in-tech-survey

Survey on Mental Health in the Tech Workplace in 2014

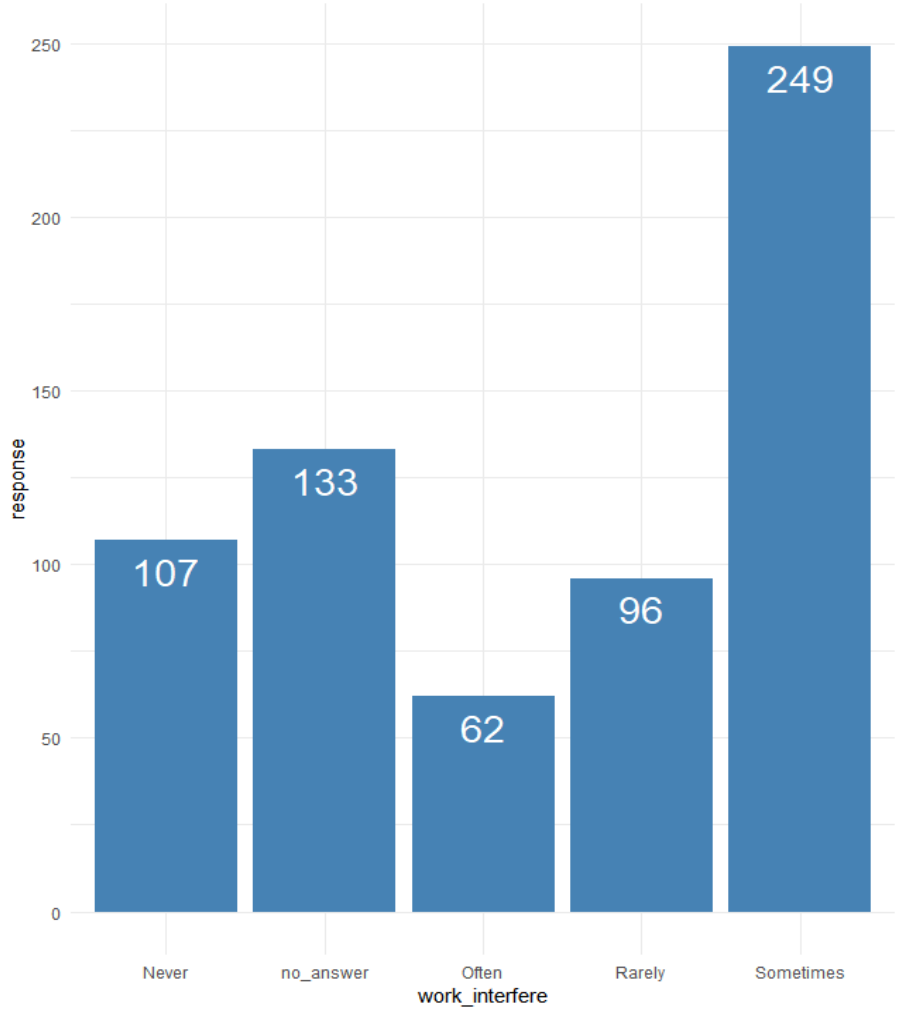
Attitudes towards mental health and frequency of mental health disorders

1. 데이터설명_변수설명

총 27개의 변수

변수명	설명
Timestamp	설문조사가 제출된 시간
Age	응답자의 나이
Gender	응답자의 성별
Country	응답자의 국적
state	미국에 거주하는 경우, 어떤 주나 지역에 살고 있습니까?
self_employed	자영업입니까?
family_history	정신질환에 대한 가족력이 있습니까?
treatment	정신 건강 상태에 대한 치료를 받으셨습니까?
work_interfere	정신 건강 상태 인 경우, 귀하의 업무에 방해가된다고 생각합니까?
no_employees	귀사에는 몇 명의 직원이 있습니까?
remote_work	적어도 50 % 이상 원격으로 (사무실 외부에서) 일하십니까?
tech_company	귀하의 고용주는 주로 기술회사입니까?
benefits	고용주가 정신 건강 혜택을 제공합니까?
care_options	고용주가 제공하는 정신 건강 치료 옵션을 알고 있습니까?
wellness_program	고용주가 직원 복지 프로그램의 일환으로 정신 건강에 관해 토론 한 적이 있습니까?
seek_help	고용주가 정신 건강 문제와 도움을 요청하는 방법에 대해 자세히 알아볼 수있는 자원을 제공합니까?
anonymity	정신 건강 또는 약물 남용 치료 자원을 활용하기로 선택하는 경우 익명성이 보호됩니까?
leave	정신 건강 상태에 대해 휴가를 갖는 것이 얼마나 쉽습니까?
mental_health_consequence	고용주와 정신 건강 문제를 논의하는 것이 부정적인 결과를 초래할 것이라고 생각합니까?
phys_health_consequence	고용주와의 신체 건강 문제를 논의하는 것이 부정적인 결과를 초래할 것이라고 생각합니까?
coworkers	직장 동료와 정신 건강 문제를 기꺼이상의 할 의향이 있습니까?
supervisor	직속 상사와 정신 건강 문제를 기꺼이상의 할 의향이 있습니까?
mental_health_interview	인터뷰에서 잠재적인 고용주와 함께 정신 건강 문제를 제기하겠습니까?
phys_health_interview	인터뷰에서 잠재적 인 고용주와 함께 신체 건강 문제를 제기하겠습니까?
mental_vs_physical	고용주가 정신 건강을 신체 건강만큼 심각하게 생각한다고 생각하십니까?
obs_consequence	직장에서 정신 건강 상태에있는 동료들에 대한 부정적인 결과를 들어 본 적이 있습니까?
comments	추가 메모 또는 의견

1. 데이터설명_변수설명


250
200
150
100
50
0
response
107
133
62
96
249
Never
no_answer
Often
Rarely
Sometimes
work_interfere

2. 탐색적자료분석_데이터정제

변수	수정
Age	경제활동인구(15세-65세)
Country	United States
State	NA 제거
Self_employed	No
Gender	M/ F

2. 탐색적자료분석_데이터정제

변수	수정
State1	$30-119 / $119-310 / 〉$310 (2014년 GDP)
State2	WEST / MIDWEST / NORTHWEST / SOUTH
Age1	10대 / 20대 / 30대 / 40대 / 50대 이상
Age2	사분위수

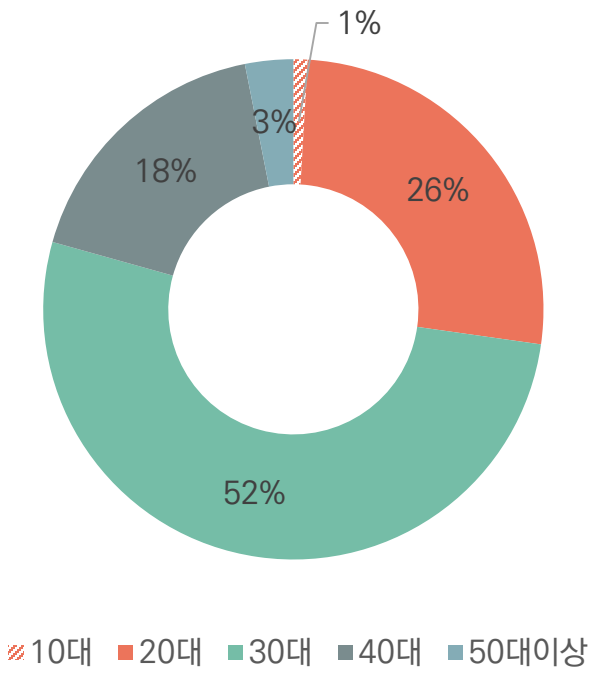
2. 탐색적자료분석_데이터정제

647개의 응답. 22개의 변수

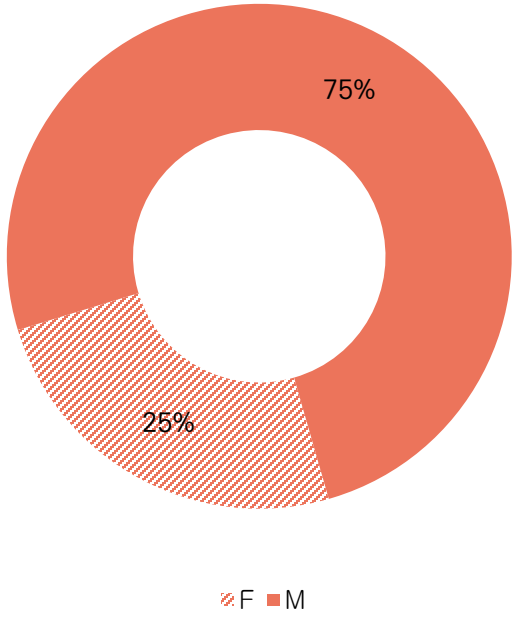
state_level	work_interfere	Age1	Gender	family_history	treatment	no_employees	remote_work	tech_company	benefits	care_options	wellness_program	seek_help
1	Often	40대	M	No	Yes	26to100	Yes	Yes	Yes	Yes	No	No
1	Sometimes	20대	M	Yes	Yes	100to500	No	Yes	Yes	Yes	No	No
2	Never	40대	M	No	Yes	over1000	No	No	Don't know	No	No	Don't know
1	Rarely	30대	M	Yes	Yes	26to100	No	Yes	Yes	Not sure	Don't know	Yes
1	Rarely	30대	F	Yes	Yes	6to25	Yes	Yes	Yes	Yes	Don't know	Don't know
2	Sometimes	30대	M	Yes	Yes	over1000	No	No	Yes	Yes	No	Don't know
2	no_answer	30대	M	No	No	1to5	No	Yes	Don't know	Not sure	No	Don't know
1	Sometimes	30대	M	No	Yes	26to100	Yes	Yes	Don't know	Not sure	No	Don't know
1	Sometimes	40대	M	Yes	Yes	26to100	Yes	Yes	Yes	Yes	Yes	Yes
2	Sometimes	40대	F	No	Yes	1to5	No	Yes	Yes	Yes	No	No
1	Rarely	20대	M	No	Yes	6to25	No	Yes	No	Yes	No	No
1	Rarely	30대	M	No	Yes	over1000	Yes	Yes	Yes	Yes	No	Yes
1	no_answer	50대이상	M	No	No	100to500	No	Yes	Yes	Yes	No	Don't know
1	no_answer	30대	M	No	No	over1000	Yes	Yes	Yes	Not sure	Don't know	Yes
2	no_answer	30대	M	No	No	6to25	No	Yes	No	No	No	No
1	Sometimes	30대	M	No	Yes	26to100	No	Yes	Don't know	No	No	No
1	Sometimes	30대	F	Yes	Yes	26to100	No	Yes	Yes	Yes	No	Yes

2. 탐색적자료분석_응답자정보

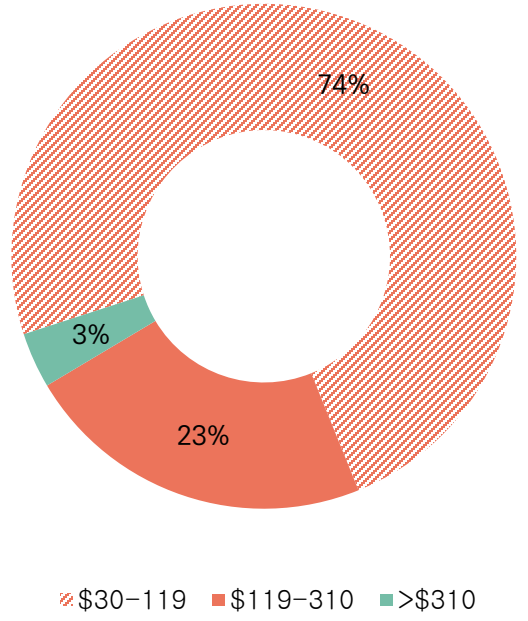
Unit : billion



Age

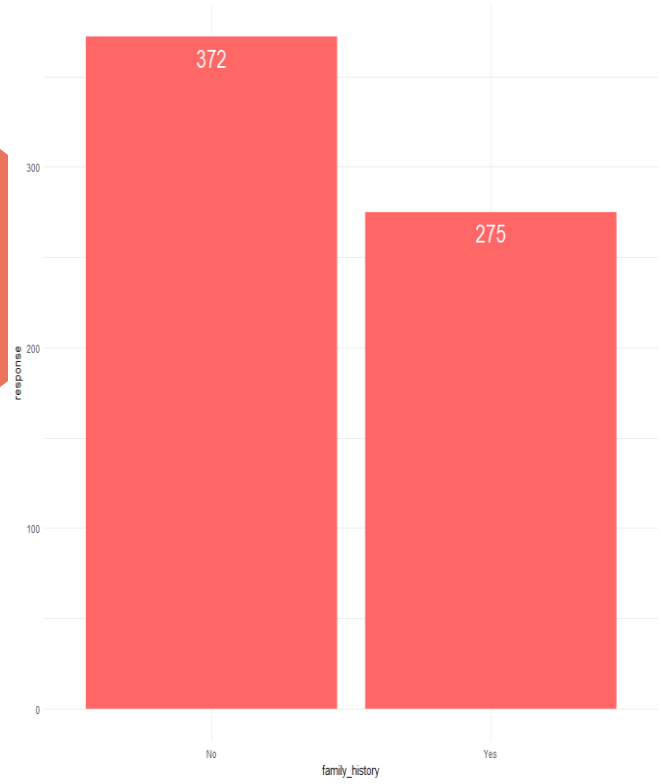


Gender

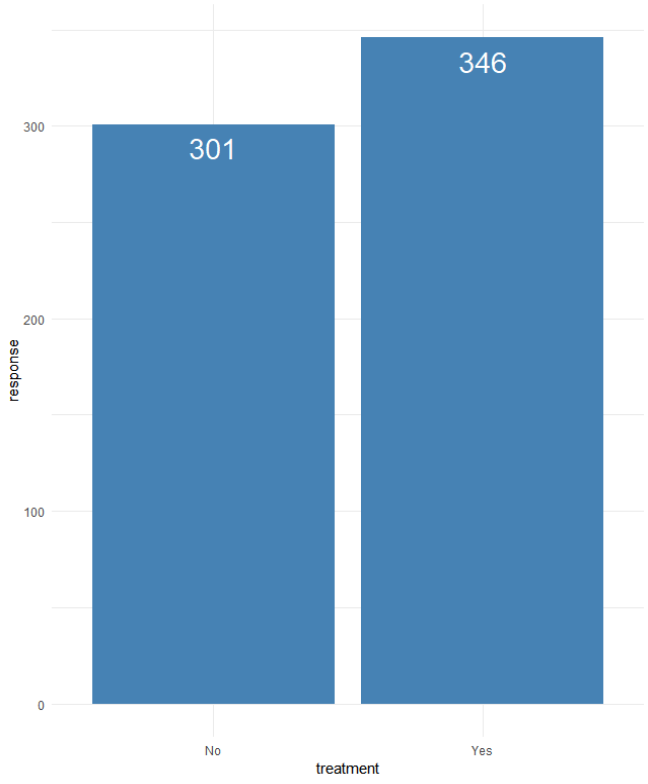


State

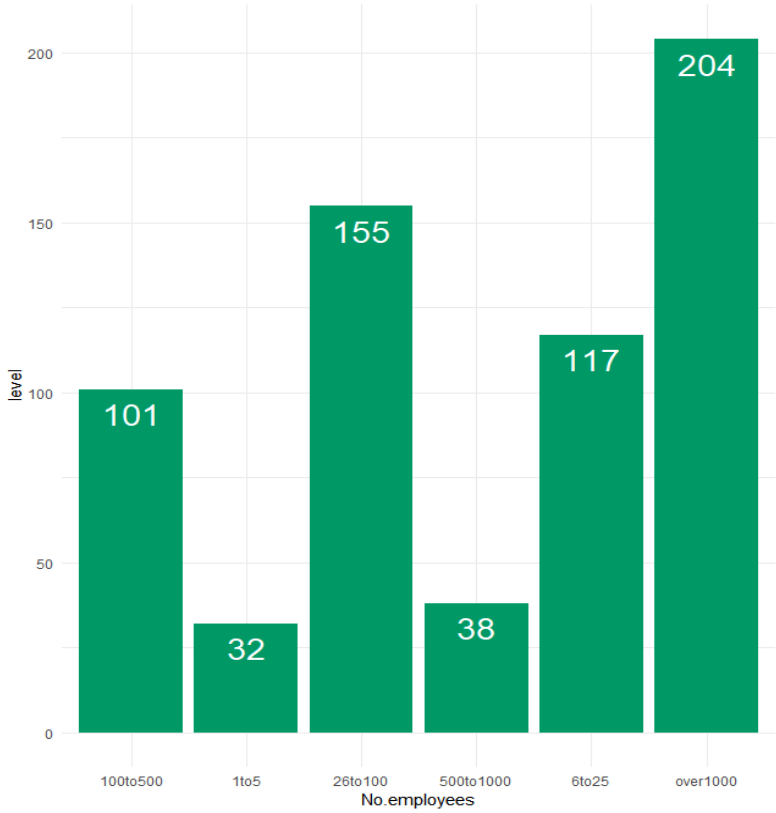
2. 탐색적자료분석_응답현황



family_history

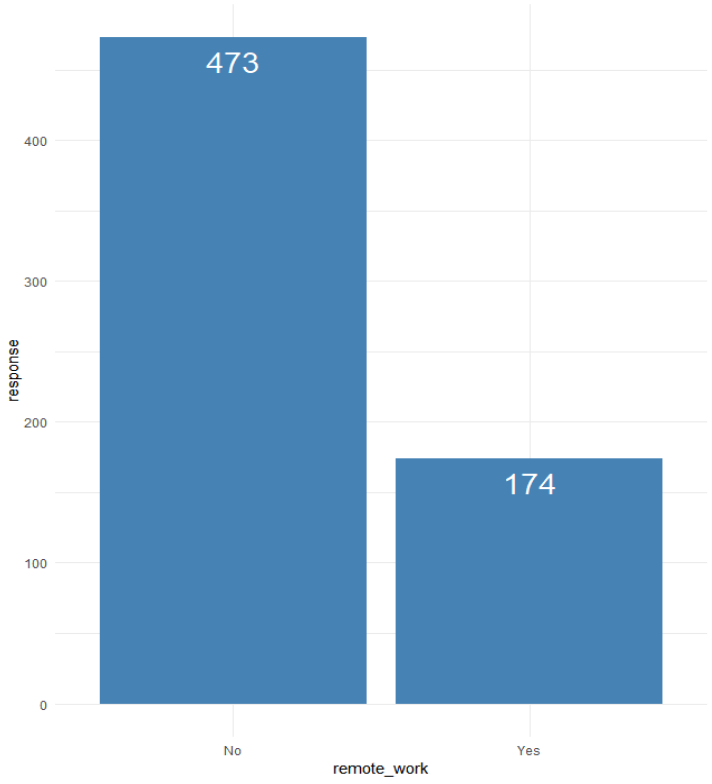


treatment

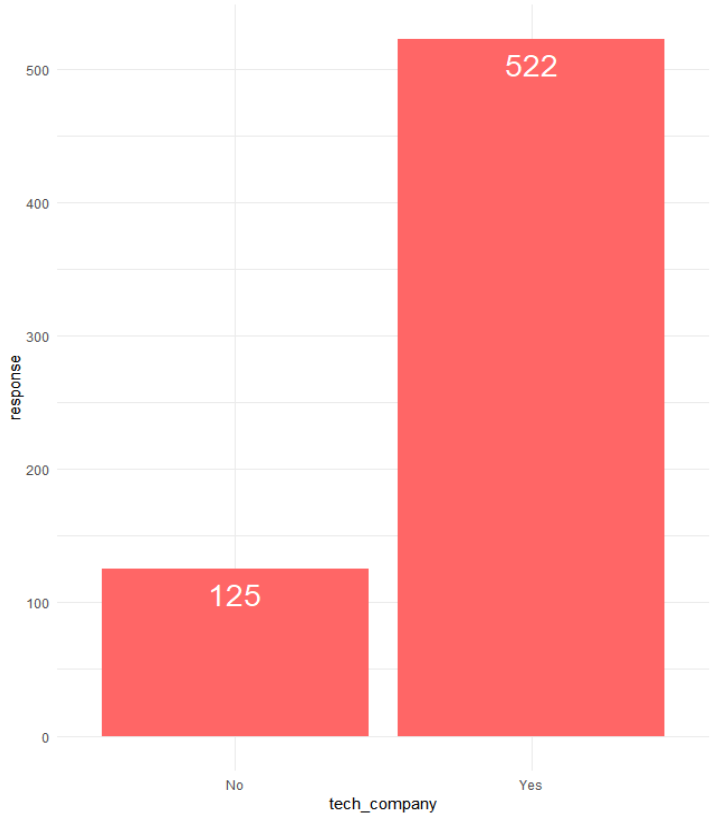


no_employees

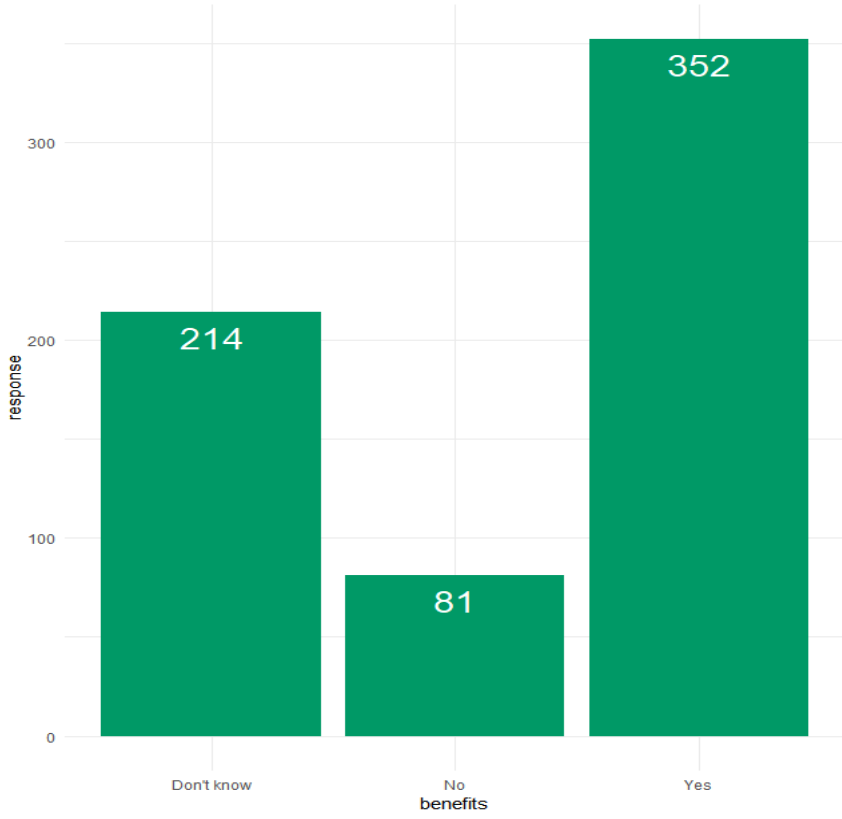
2. 탐색적자료분석_응답현황



remote_work

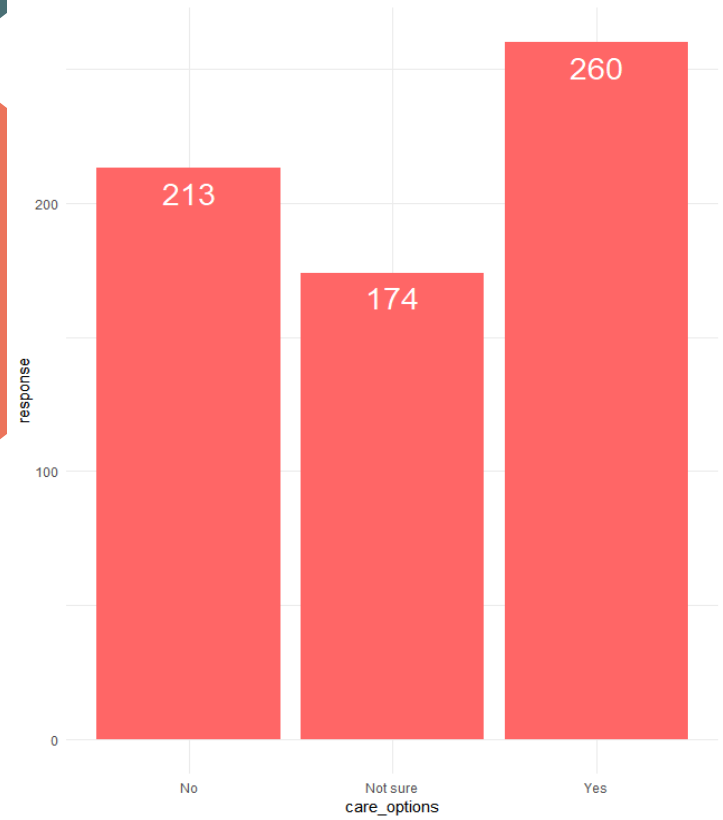


tech_company

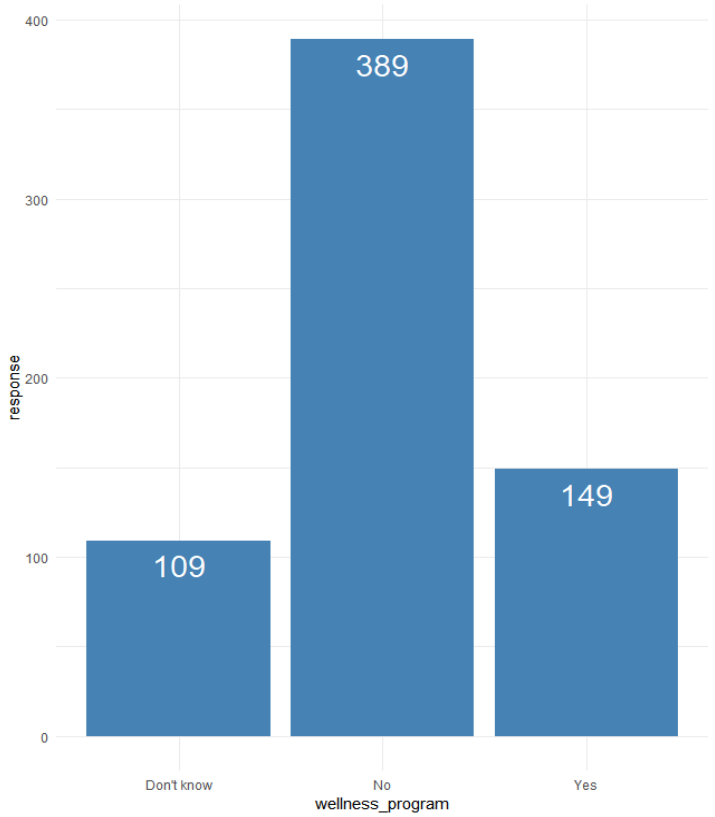


benefits

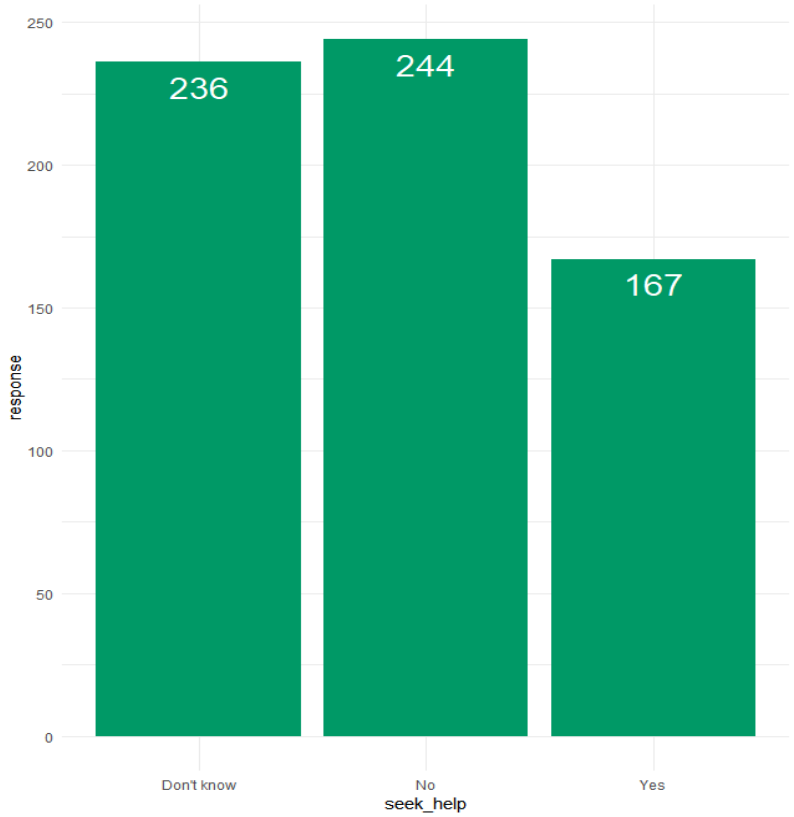
2. 탐색적자료분석_응답현황



care_options

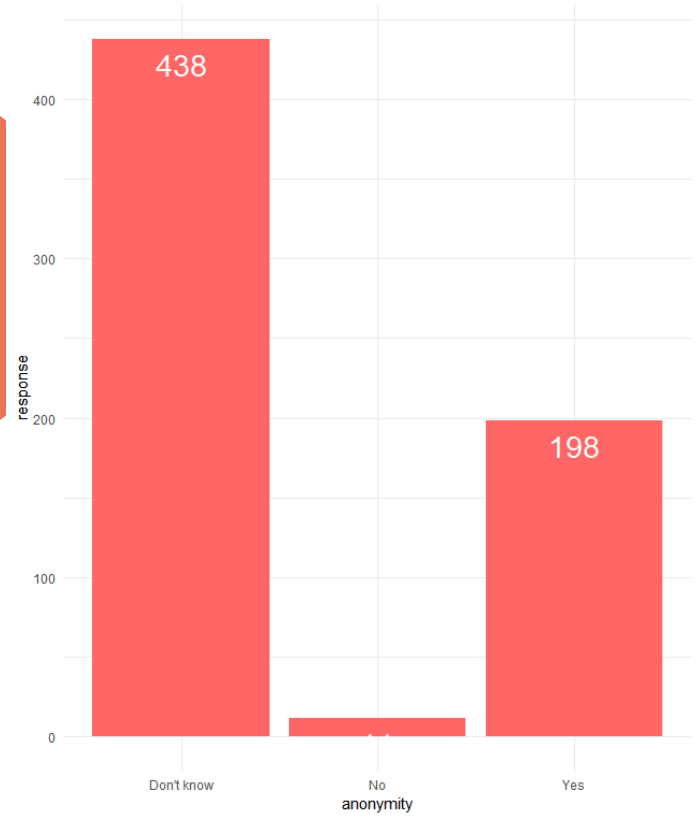


wellness_program

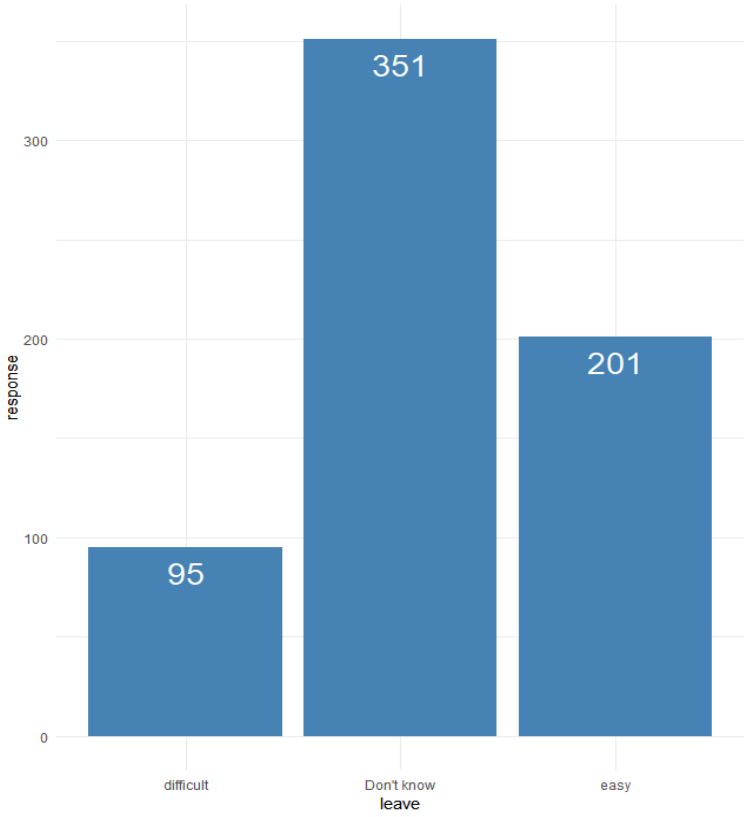


seek_help

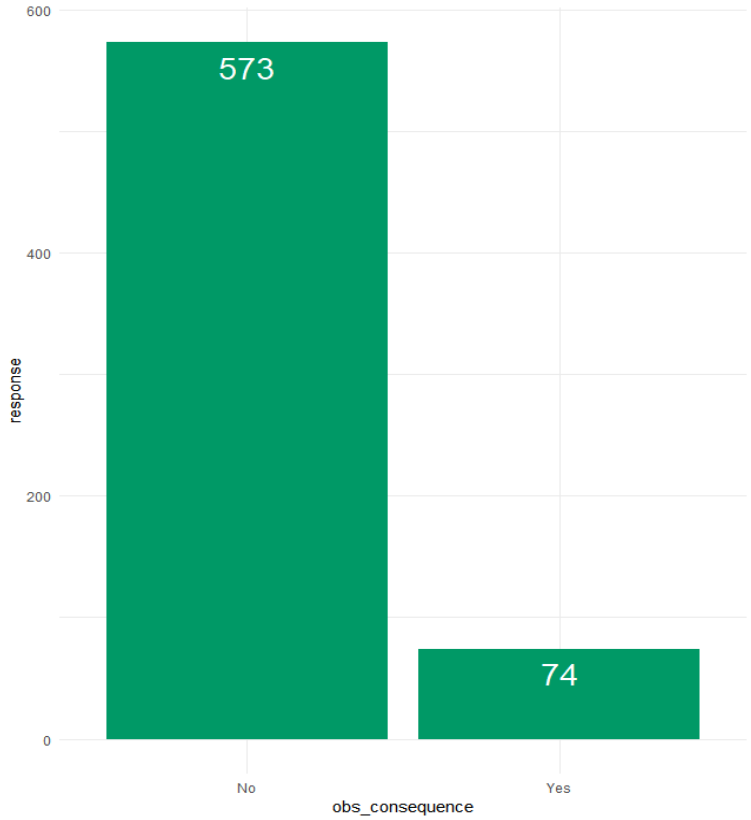
2. 탐색적자료분석_응답현황



anonymity

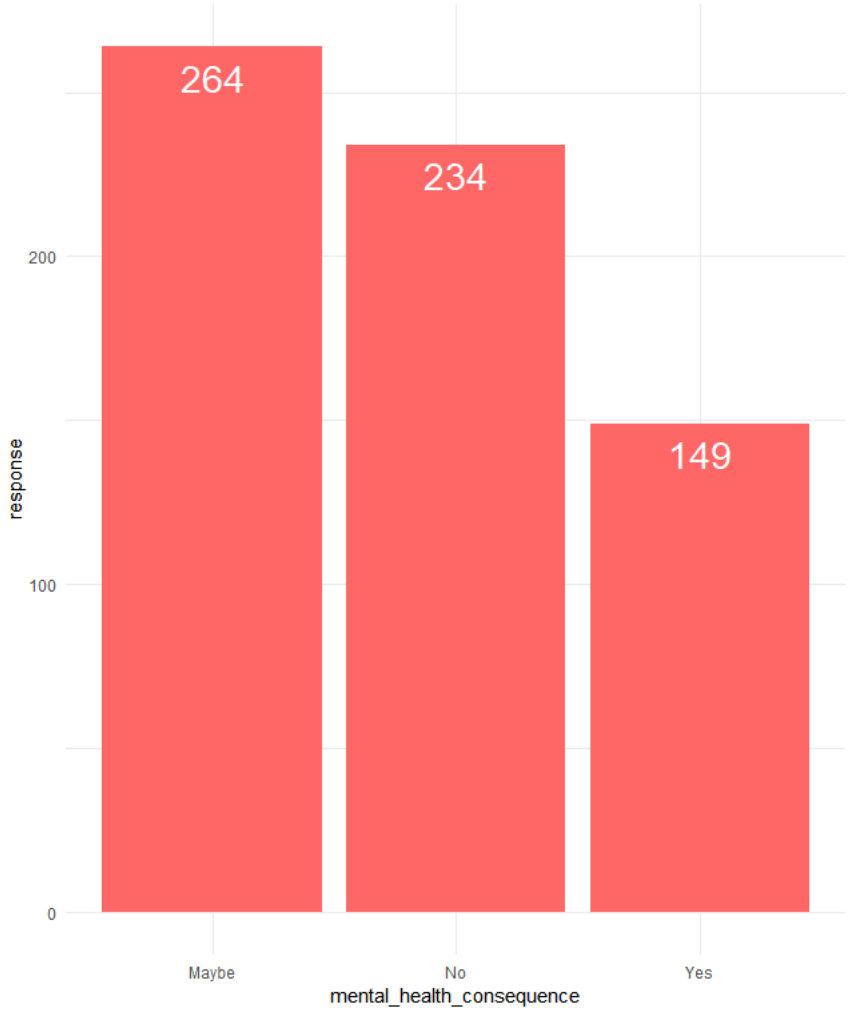


leave

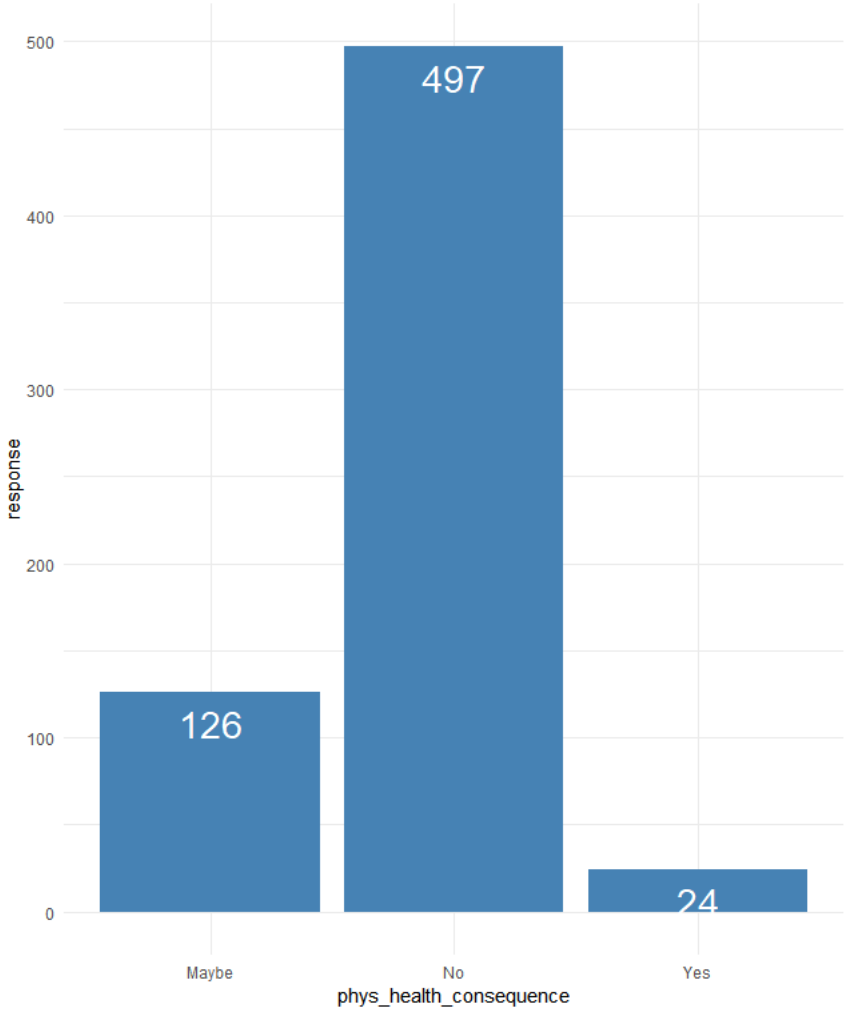


obs_consequence

2. 탐색적자료분석_응답현황

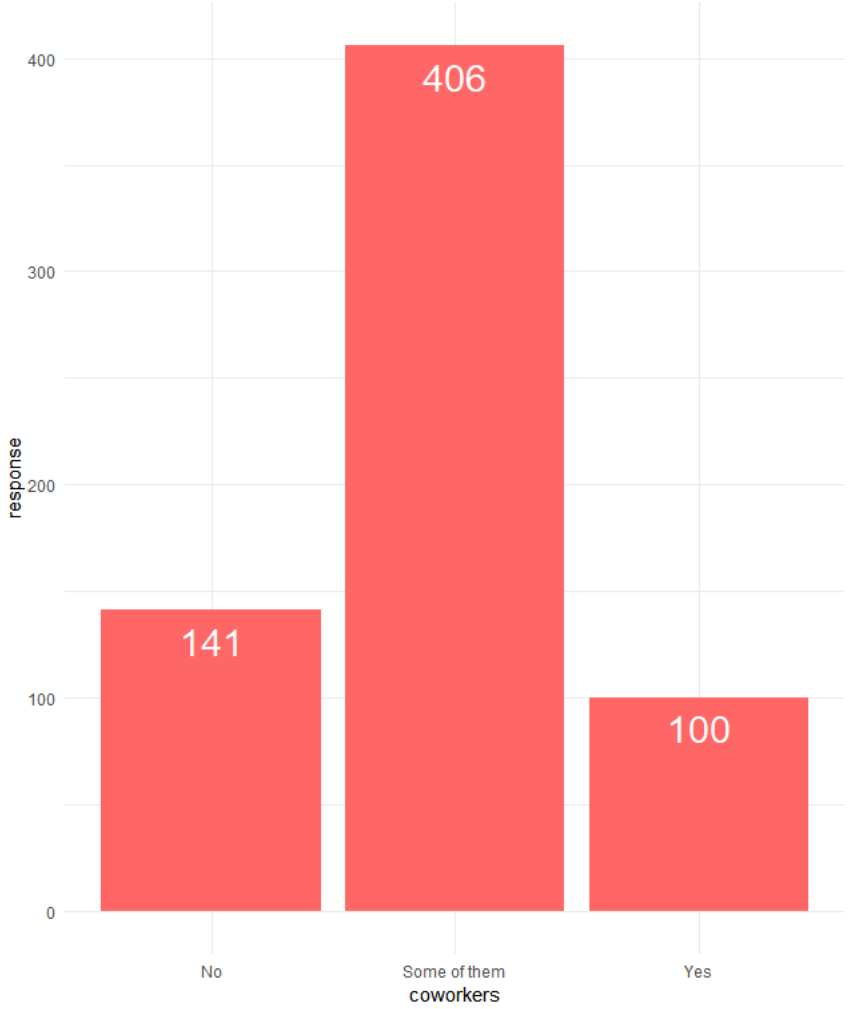


mental_health_consequence

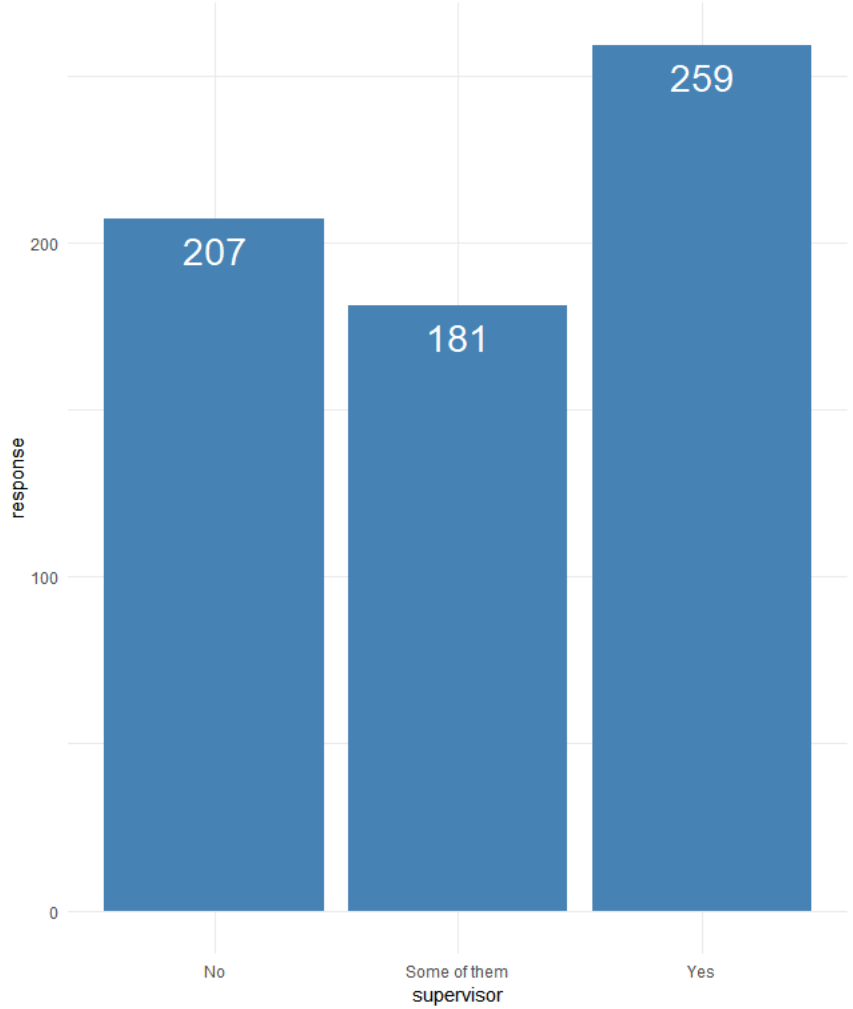


phys_health_consequence

2. 탐색적자료분석_응답현황

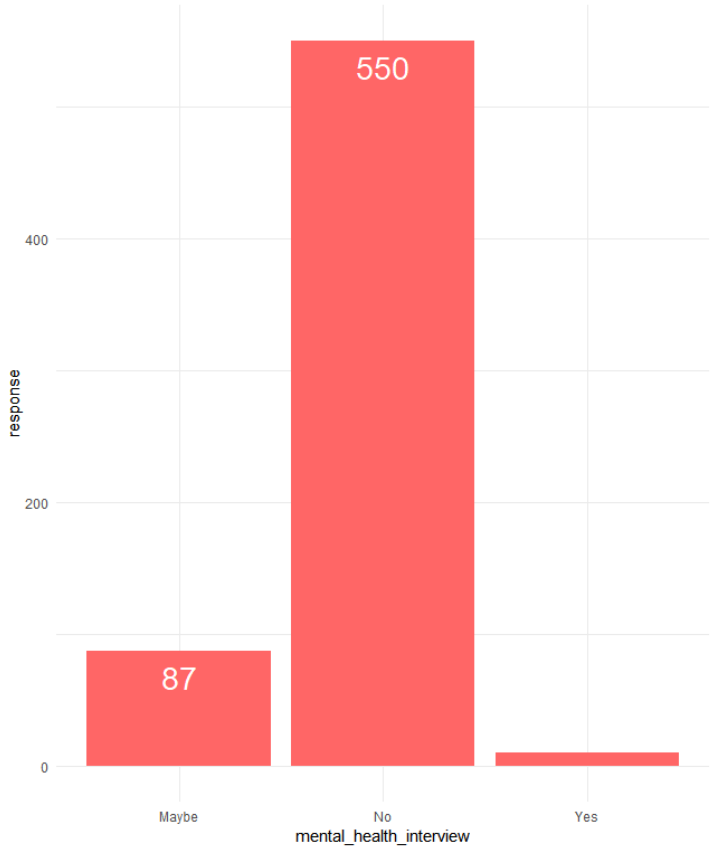


coworkers

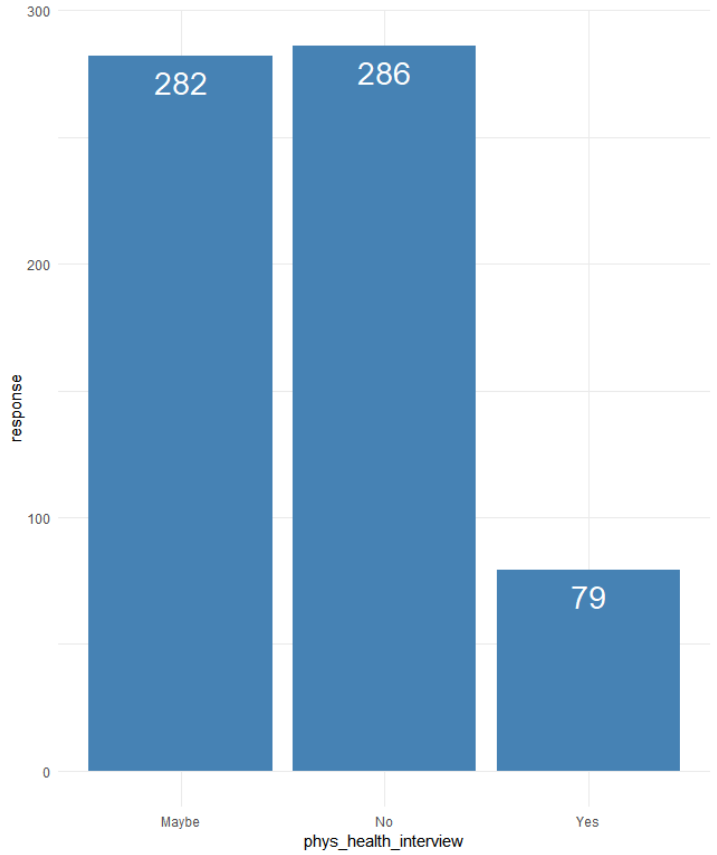


supervisor

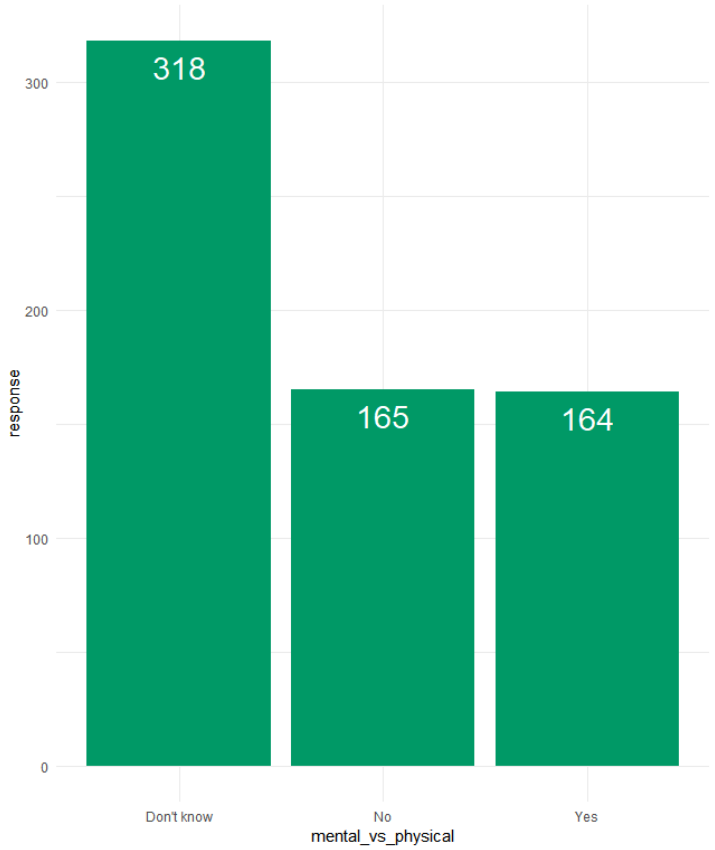
2. 탐색적자료분석_응답현황



mental_health_interview

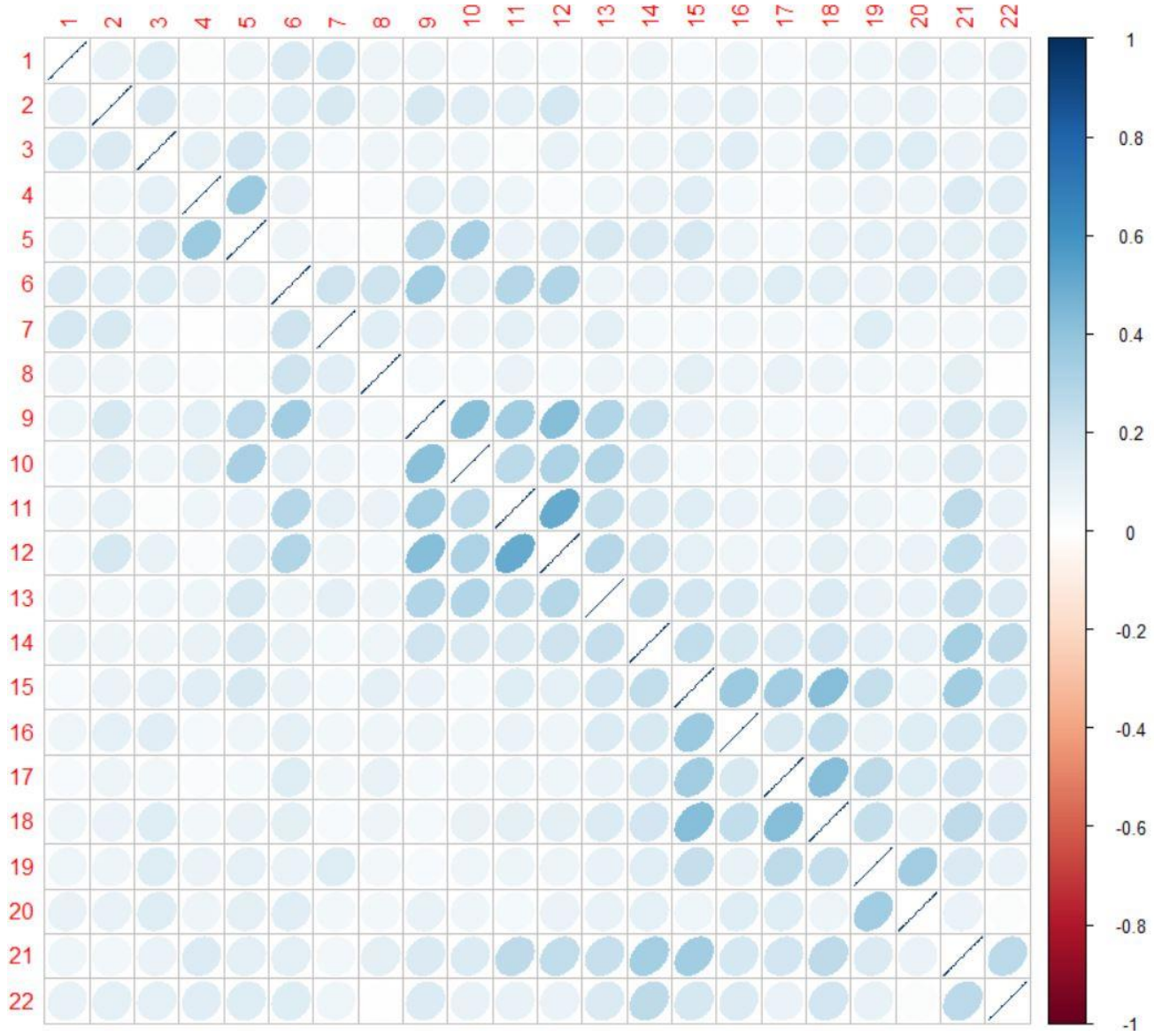


phys_health_interview



mental_vs_physical

3. 자료분석_상관행렬


1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22
1 0.8 0.6 0.4 0.2 0 -0.2 -0.4 -0.6 -0.8 -1

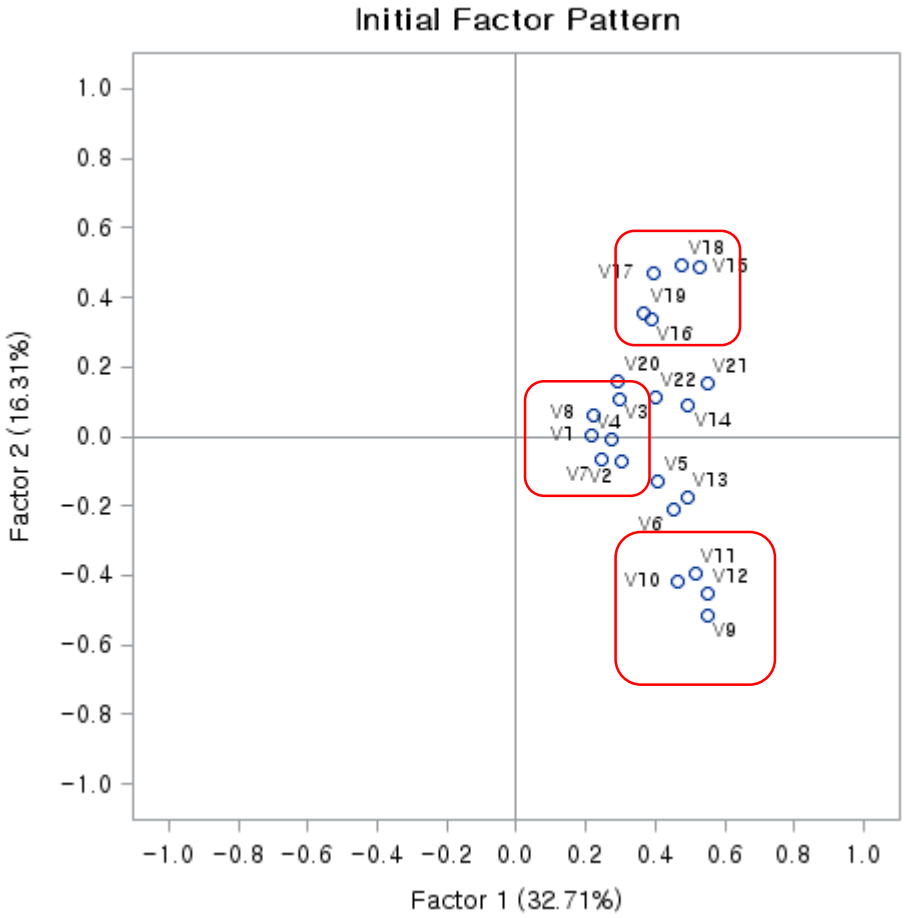
3. 자료분석_주성분분석

	Comp.1	Comp.2	Comp.3	Comp.4	Comp.5
state_level	-0.01378	0.274906	-0.06104	0.257257	-0.28818
Age1	0.063502	0.17541	-0.08364	0.096637	-0.04814
Gender	-0.08173	0.1619	0.200153	0.054072	0.022616
family_history	0.01822	-0.00519	0.493867	0.062157	0.321267
treatment	0.124853	-0.04964	0.533216	-0.0627	0.19017
no_employees	0.193297	0.108417	-0.30741	-0.00159	0.246256
remote_work	0.043855	0.264153	-0.25241	0.114447	-0.15124
tech_company	-0.04556	0.155258	-0.25293	0.134838	0.417127
benefits	0.406824	-0.12299	-0.01368	-0.10668	0.029245
care_options	0.33941	-0.13924	0.146703	-0.17212	0.018294
wellness_program	0.325011	-0.19194	-0.24637	-0.10661	0.08013
seek_help	0.361151	-0.19406	-0.1946	-0.15416	-0.01185
anonymity	0.195583	-0.26774	-0.00476	-0.06666	-0.21421
leave	-0.00455	-0.33166	0.040491	0.173374	-0.33098
mental_health_consequence	-0.28505	-0.33803	-0.06027	0.023171	0.195405
phys_health_consequence	-0.21632	-0.16879	-0.08083	0.112127	0.030785
coworkers	-0.29002	-0.17707	-0.18078	-0.24241	0.18313
supervisor	-0.2951	-0.27258	-0.12173	-0.07718	0.159865
mental_health_interview	-0.24476	0.019793	0.007931	-0.48256	-0.23973
phys_health_interview	-0.11726	0.158962	0.066379	-0.51588	-0.31118
mental_vs_physical	-0.04434	-0.39304	-0.0616	0.154904	-0.09181
obs_consequence	-0.0436	-0.20019	0.104565	0.413544	-0.30511

```
                          Comp.1    Comp.2    Comp.3    Comp.4
Standard deviation     1.9979181 1.7524086 1.4870655 1.25806525
Proportion of Variance 0.1814399 0.1395880 0.1005165 0.07194219
Cumulative Proportion  0.1814399 0.3210279 0.4215444 0.49348658
                          Comp.5    Comp.6    Comp.7    Comp.8
Standard deviation     1.11382483 1.10271846 1.0450390 0.99671051
Proportion of Variance 0.05639117 0.05527218 0.0496412 0.04515599
Cumulative Proportion  0.54987775 0.60514993 0.6547911 0.69994713
                          Comp.9   Comp.10   Comp.11   Comp.12
Standard deviation     0.9773187 0.92270650 0.86040200 0.85321026
Proportion of Variance 0.0434160 0.03869942 0.03364962 0.03308944
Cumulative Proportion  0.7433631 0.78206254 0.81571216 0.84880161
```

Comp.1 : Benefit
Comp.2 : mental_vs_physical
Comp.3 : family_history, treatment

3. 자료분석_인자분석



일부 변수들의 연관성이 보인다.
1. state_level, Age, Gender, family_history, no_employees, remote_work –직원 정보, 회사 성질

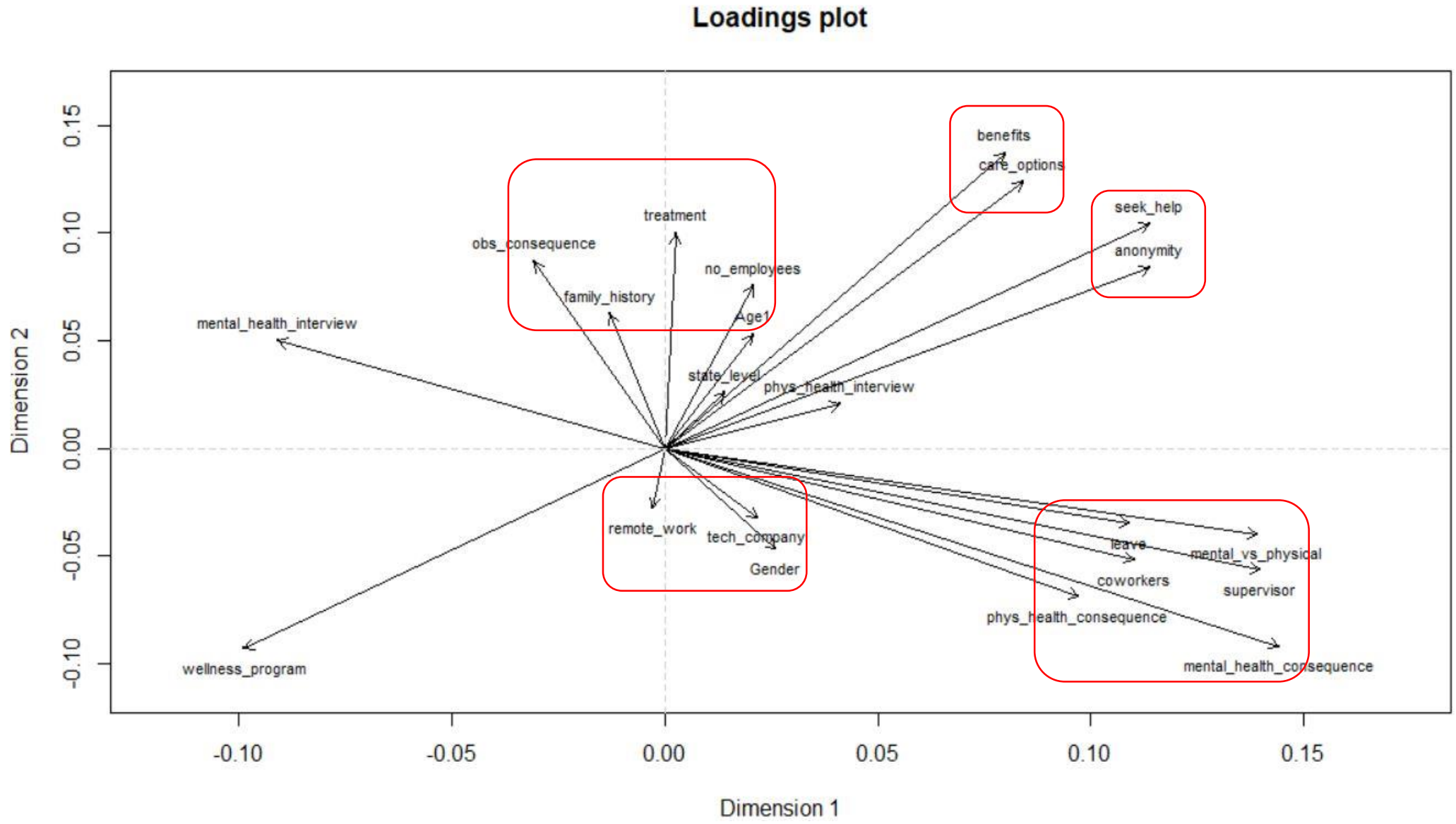
2. mental_health_consequence, leave, phys_health_consequence, coworkers, supervisor –직원의 생각과 태도

3. benefits, care_options, wellness_program, tech_company –회사의 복지

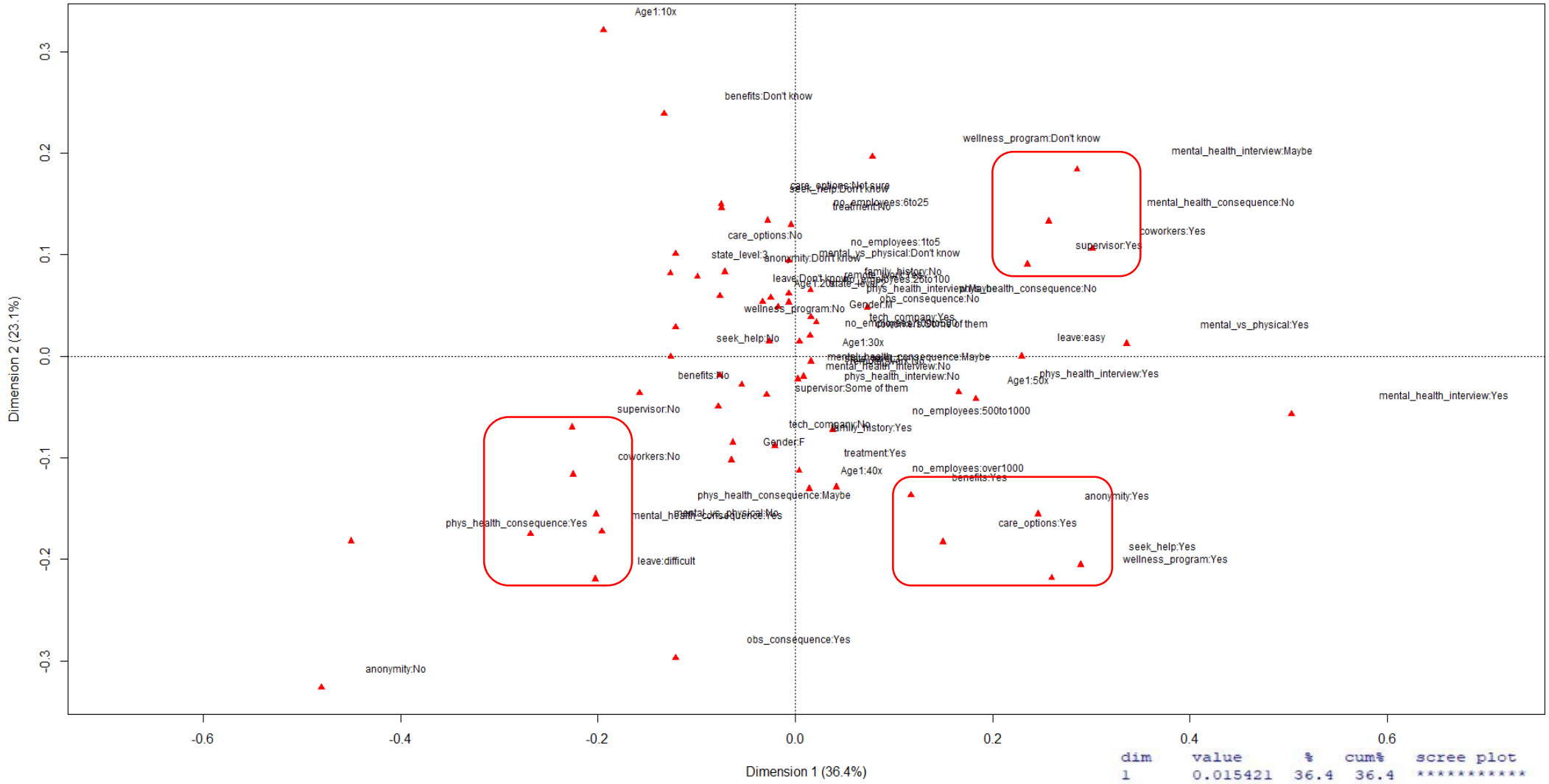
Eigenvalues of the Correlation Matrix: Total = 22 Average = 1

	Eigenvalue	Difference	Proportion	Cumulative
1	3.85241962	1.93115857	0.1751	0.1751
2	1.92126105	0.48816528	0.0873	0.2624
3	1.43309578	0.05195770	0.0651	0.3276
4	1.38113807	0.22535537	0.0628	0.3904
5	1.15578270	0.12574015	0.0525	0.4429
6	1.03004255	0.02767171	0.0468	0.4897
7	1.00237084	0.04475620	0.0456	0.5353
8	0.95761464	0.03969580	0.0435	0.5788
9	0.91791884	0.02196199	0.0417	0.6205
10	0.89595685	0.04408584	0.0407	0.6613

3. 자료분석_대응분석


Loadings plot
Dimension 2
Dimension 1
0.15
0.10
0.05
0.00
-0.05
-0.10
-0.10
-0.05
0.00
0.05
0.10
0.15
benefits
care_options
seek_help
anonymity
treatment
obs_consequence
no_employees
family_history
Age1
mental_health_interview
state_level
phys_health_interview
remote_work
tech_company
Gender
leave
mental_vs_physical
coworkers
supervisor
phys_health_consequence
mental_health_consequence
wellness_program

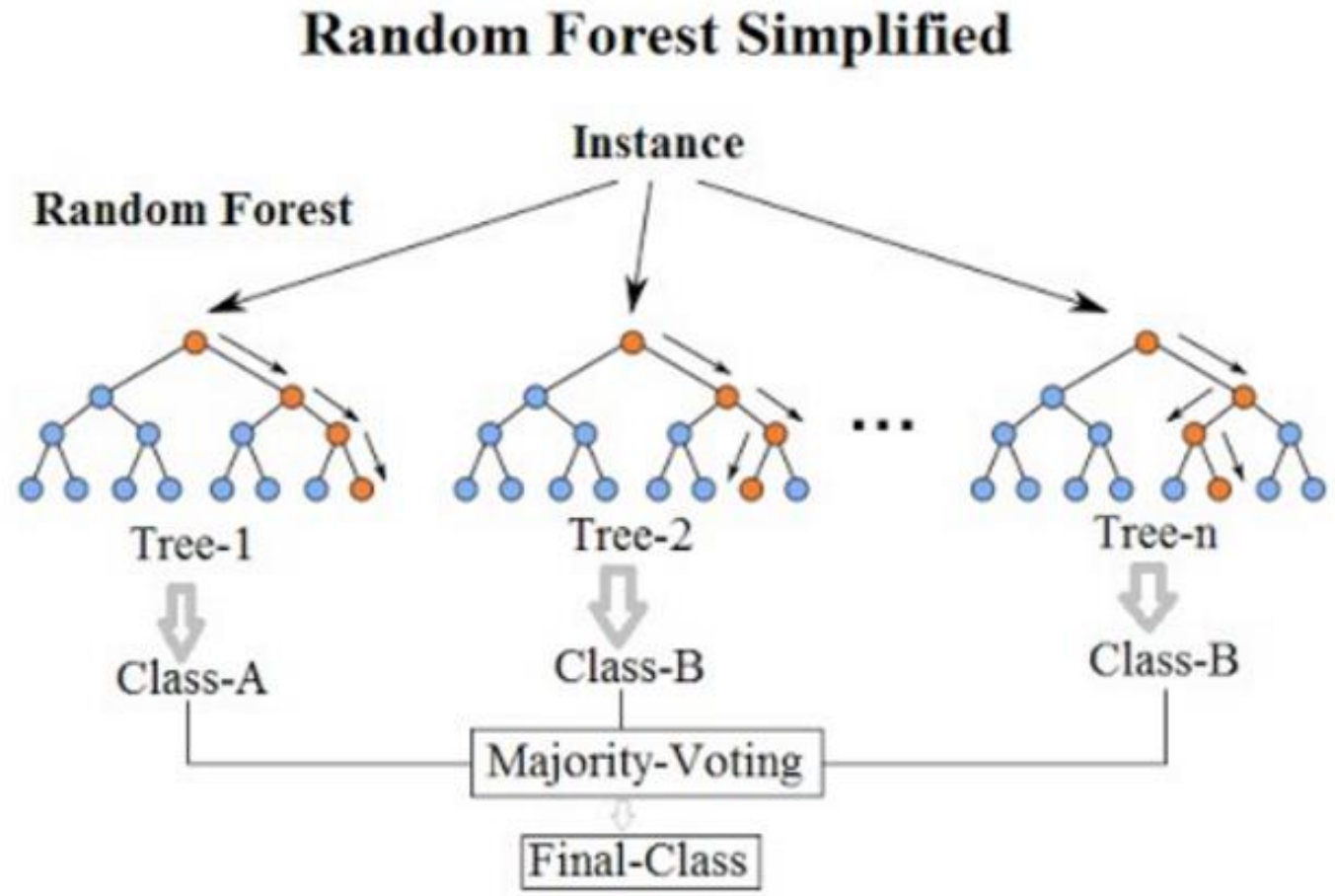
3. 자료분석_대응분석



dim	value	%	cum%	scree plot
1	0.015421	36.4	36.4	**********
2	0.009766	23.1	59.5	*******
3	0.005318	12.6	72.0	****
4	0.000995	2.3	74.4	*
5	0.000741	1.7	76.1	*
6	0.000552	1.3	77.4	

4. 모델링_random forest

- 알고리즘 소개



4. 모델링_random forest

- 알고리즘 선택이유
  - 범주형 자료에서 사용가능
  - 분류(classification)문제 사용가능
  - 이상치 자체를 하나의 경우로 분류하기 때문에 이상치에 안정적
  - **인자분석과 주성분분석을 통한 차원축소 불가**
  - **상관관계에 따른 변수제거 및 차원축소를 하지않고 변수의 개수를 선택해서 모델링**

4. 모델링_random forest

- K-fold
  - 과적합(overfitting)을 방지하기 위해 비복원 추출법인 K-fold CV 사용

K-Fold Cross Validation (K중 교차 타당성) 방법

EX
전체 데이터 1000개 -> 데이터 10등분 -> 훈련 데이터 900개, 시험 데이터 100개
-> 실제 데이터는 1000개지만 10,000개 데이터를 분석한 효과를 볼 수 있다.

장점
적은 수의 데이터의 모형 성능 평가 및 향상 기법, 안정적

4. 모델링_random forest

- Hyperparameter optimization
  - mtry : 랜덤하게 뽑을 변수의 개수(보통 classification에서는 $\sqrt{p}$, regression에서는 $\frac{1}{3}$ p) -〉 [1, 2, · · · , 15]
    * p is number of variable
  - ntree : tree의 개수 -〉 [300, 500, 1000, 1500]
  - p : train_set의 비율 -〉 [0.6, 0.7, 0.8]
  - k : k-fold 의 k -〉 [5, 10]

4. 모델링_random forest

- Hyperparameter optimization

mtry	ntree	p	k
1	300	0.6	5
.	500	0.7	10
.			
8	1000	0.8	
.	1500		
.			
15			

```
k-fold number= 5
p= 0.7
ntree= 1000

Random Forest

452 samples
 22 predictor
  5 classes: 'Never', 'no_answer', 'Often', 'Rarely', 'Sometimes'

No pre-processing
Resampling: Cross-Validated (5 fold, repeated 3 times)
Summary of sample sizes: 361, 361, 362, 361, 363, 364, ...
Resampling results across tuning parameters:

  mtry  Accuracy   Kappa
   1    0.4226252  0.0006155297
   2    0.4851745  0.1698101686
   3    0.5061252  0.2397529220
   4    0.5171223  0.2720288081
   5    0.5200207  0.2828100806
   6    0.5179049  0.2855790960
   7    0.5105112  0.2781250707
   8    0.5223264  0.2965535419
   9    0.5066761  0.2756222159
  10    0.5149005  0.2896087900
  11    0.5126866  0.2881350744
  12    0.5140779  0.2907706506
  13    0.5111141  0.2870994708
  14    0.5125956  0.2894855601
  15    0.5141096  0.2926608651
```

4. 모델링_random forest

- 예측 결과

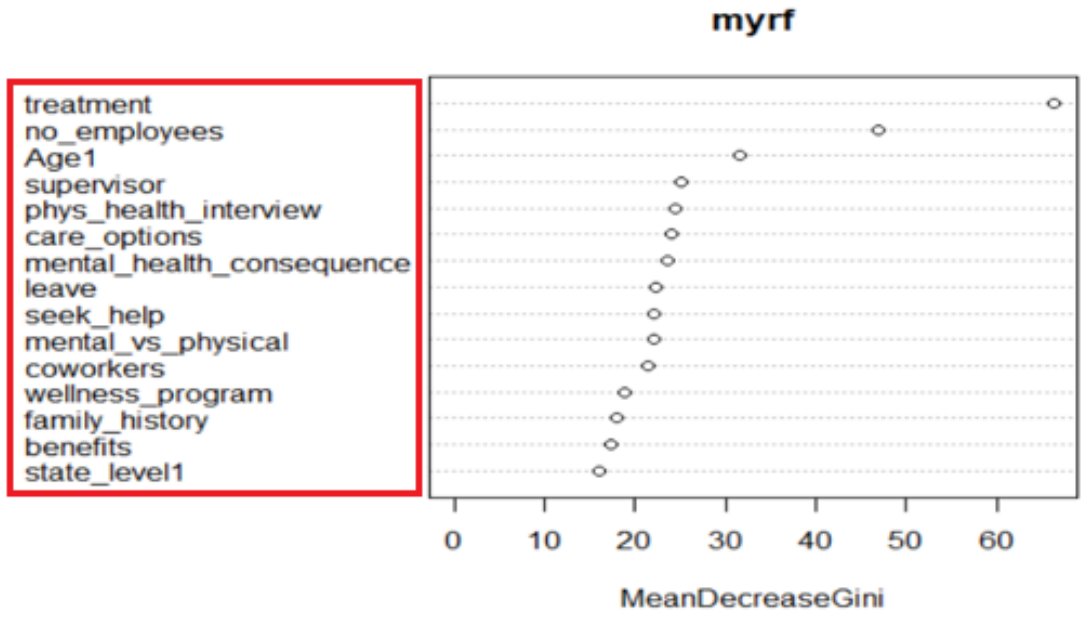
```
Confusion Matrix and Statistics

            Reference
Prediction   Never no_answer Often Rarely Sometimes
  Never          9         8     0      2         0
  no_answer     12        26     0      3         3
  Often          0         0     0      0         4
  Rarely         0         0     2      3         4
  Sometimes     10         5    15     24        65

Overall Statistics

                Accuracy : 0.5282
                  95% CI : (0.4556, 0.5999)
    No Information Rate : 0.3897
    P-Value [Acc > NIR] : 6.151e-05

                   Kappa : 0.3184
 Mcnemar's Test P-Value : NA
```



4. 모델링_random forest

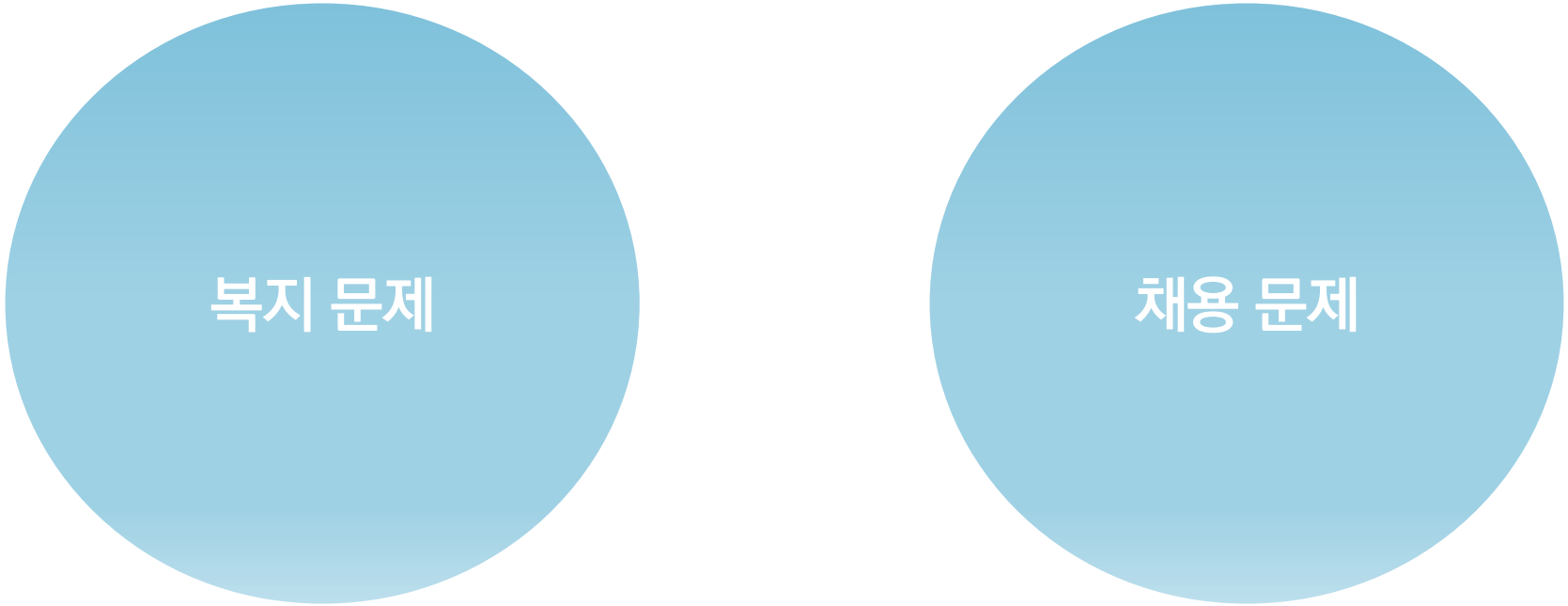
- 영향력 있는 변수들

Imp_var1	Imp_var2	Imp_var3	Imp_var4	Imp_var5	Imp_var6
Care_options	age1	Treatment	Coworkers	No_emplyees	Mental_health_consequence
Leave	State_level1	Family_history	Supervisor		Mental_vs_physical
Seek_help			Phys_health_interview		
Wellness_program					
benefits					

Imp_var1 – 회사의 정신건강 복지 실태
Imp_var2 – 응답자 정보
Imp_var3 – 응답자의 정신건강상태
Imp_var4 – 의견제시 경향
Imp_var5 – 회사 정보
Imp_var6 – 회사 정신건강 복지 실태에 대한 응답자의 생각

5. 결론

- 업무에 대한 기업적 측면



THANK YOU