Data Analysis of IBM Attrition

IBM 사조

공준택 김순영 김혜린 오희준 장인아

CONTENTS

01 INTRO

	A	B	C	D	E	F	G	H
1	Age	Attrition	BusinessTr	DailyRate	Departme	DistanceFr	Education	Education
2	41	Yes	Travel_Rar	1102	Sales	1	2	Life Scien
3	49	No	Travel_Fre	279	Research &	8	1	Life Scien
4	37	Yes	Travel_Rar	1373	Research &	2	2	Other
5	33	No	Travel_Fre	1392	Research &	3	4	Life Scien
6	27	No	Travel_Rar	591	Research &	2	1	Medical
7	32	No	Travel_Fre	1005	Research &	2	2	Life Scien
8	59	No	Travel_Rar	1324	Research &	3	3	Medical
9	30	No	Travel_Rar	1358	Research &	24	1	Life Scien
10	38	No	Travel_Fre	216	Research &	23	3	Life Scien
11	36	No	Travel_Rar	1299	Research &	27	3	Medical
12	35	No	Travel_Rar	809	Research &	16	3	Medical
13	29	No	Travel_Rar	153	Research &	15	2	Life Scien
14	31	No	Travel_Rar	670	Research &	26	1	Life Scien
15	34	No	Travel_Rar	1346	Research &	19	2	Medical
16	28	Yes	Travel_Rar	103	Research &	24	3	Life Scien
17	29	No	Travel_Rar	1389	Research &	21	4	Life Scien
18	32	No	Travel_Rar	334	Research &	5	2	Life Scien
19	22	No	Non-Trave	1123	Research &	16	2	Medical

Data set:

IBM 이직 데이터

왜 이직 데이터를 분석하는가?

출처 : http://www.edaily.co.kr/news/NewsRead.edy?SCD=JC61&newsid=01872886615924656&DCD=A00306&OutLnkChk=Y

중소기업 조기 퇴사율 32.5%, 해결방안은?

입력시간 | 2017.05.01 11:08 | 정태선 기자 windy@edaily.co.kr

독자의견

회초년생 1년 내 퇴사율 27.7%, 2012년 대비 4.1% 증가

미스매치, 개인·기업 모두에 부정적인 결과 보여줘

다양한 온매치 정책 활용해 거시적인 성장 내다봐야

뉴스

Tweet G+1 1

How To / 리더십|조직관리 / 모바일 / 보안 / 분쟁|갈등 / 비즈니스|경제 / 소비자IT / 애플리케이션 / 이직/채용 / 인문학|교양

2017.05.02

'퇴사·이직과 함께 유출되는' 기업 기밀, 어떻게 지킬까

Andy Patrizio | CIO

'회사에서 가장 가치 있는 자산은 회사를 떠난다'는 격언이 있다. 그러나 최근 발표된 보안 조사 보고서에 따르면, 이 가장 가치 있는 자산을 따라 회사를 떠나 다시 돌아오지 않는 또다른 가치 있는 자산도 있다.

출처 : http://www.ciokorea.com/news/34064

왜 이직 데이터를 분석하는가?

지나친 이직은 기업의
비효율을 초래

주제 :

IBM 직원들의 이직 결정 요인들을 분석하여 지나친 이직을 막기 위한 방법을 제시

02 VARIABLES

<반응 변수>

Attrition	이직 여부

- Yes (237)
- No (1233)

<설명 변수>

- 모든 관측치에 대해 동일한 값을 가지는 변수 무시

Over18	StandardHours	EmployeeCount

- 여러 카테고리로 나눔

  - 현재 직장과 무관하게 정해지는 변수
  - 입사 후 정해지는 변수 (만족도, 업무 등)
  - 임금과 관련된 변수

1. 현재 직장과 무관하게 정해지는 변수

Age	나이	Gender	성별
TotalWorkingYears	경력	NumCompaniesWorked	일했던 직장 수
Education 1 : Below College 2 : College 3 : Bachelor 4 : Master 5 : Doctor	교육수준	DistanceFromHome	집과의 거리
EducationField Human Resources Life Sciences / Marketing Medical / Other Technical Degree	전공	MaritalStatus Single/Married/Divorced	결혼 수준

2. 입사 후 정해지는 변수 - 1

Relationship Satisfaction 1 : Low 2 : Medium 3 : High 4 : Very High	관계 만족도	YearsAtCompany	직장 내 연차
TotalWorkingYears	총 근무 기간(년)	YearsInCurrentRole	직무 연차
TrainingTimesLastYear	작년 직업 훈련시간	YearsSince LastPromotion	승진 후 경과 년 수
WorkLifeBalance 1 : Bad 2 : Good 3 : Better 4 : Best	일과 생활의 균형	YearsWithCurrManager	현재 상사와의 함께한 년 수

2. 입사 후 정해지는 변수 - 2

BusinessTravel Non-Travel / Travel_Frequently Travel_Rarely	출장	**JobLevel**	직급
Department Human Resources Research & Development / Sales	부서	**JobRole**	직업 역할
EnvironmentSatisfaction 1 : Low 2 : Medium 3 : High 4 : Very High	환경만족도	**JobSatisfaction** 1 : Low 2 : Medium 3 : High 4 : Very High	직업 만족도
JobInvolvement 1 : Low 2 : Medium 3 : High 4 : Very High	직장 소속감	**OverTime**	초과근무 여부
PerformanceRating 1 : Low 2 : Good 3 : Excellent 4 : Outstanding	업무평가		

3. 임금과 관련된 변수

HourlyRate	시급	DailyIncome	일급
MonthlyIncome	월 소득	MonthlyRate	월급
PercentSalaryHike	임금 상승률	StockOption	스톡옵션

03 VISUALIZATION

Factor 처리 되지 않은 변수들 → Factor 처리

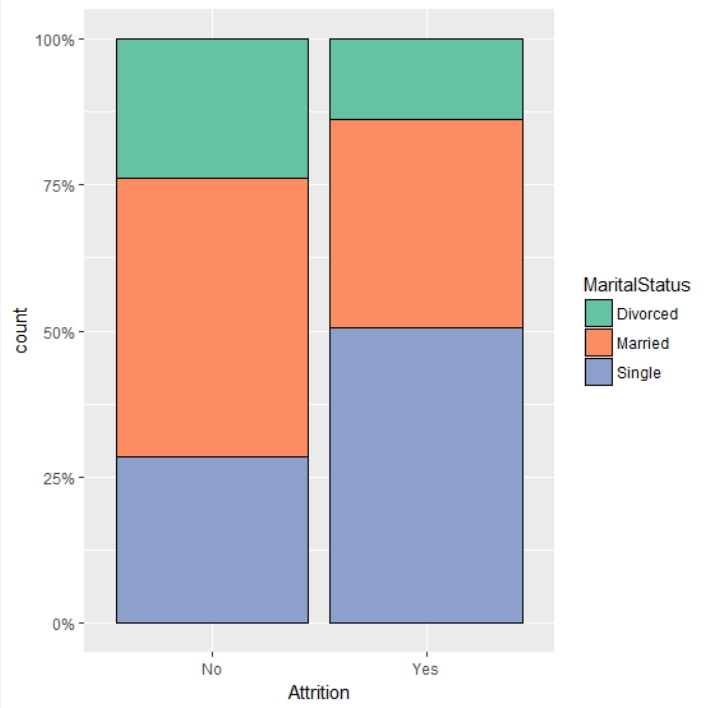
(시각화의 편의를 위해)

```
dat$Education<-factor(dat$Education,ordered=T)
dat$EnvironmentSatisfaction<-factor(dat$EnvironmentSatisfaction,ordered=T)
dat$JobInvolvement<-factor(dat$JobInvolvement,ordered=T)
dat$JobSatisfaction<-factor(dat$JobSatisfaction,ordered=T)
dat$RelationshipSatisfaction<-factor(dat$RelationshipSatisfaction,ordered=T)
dat$WorkLifeBalance<-factor(dat$WorkLifeBalance,ordered=T)
dat$PerformanceRating<-factor(dat$PerformanceRating,ordered=T)
dat$JobLevel<-factor(dat$JobLevel,ordered=T)
dat$StockOptionLevel<-factor(dat$StockOptionLevel,ordered=T)
```

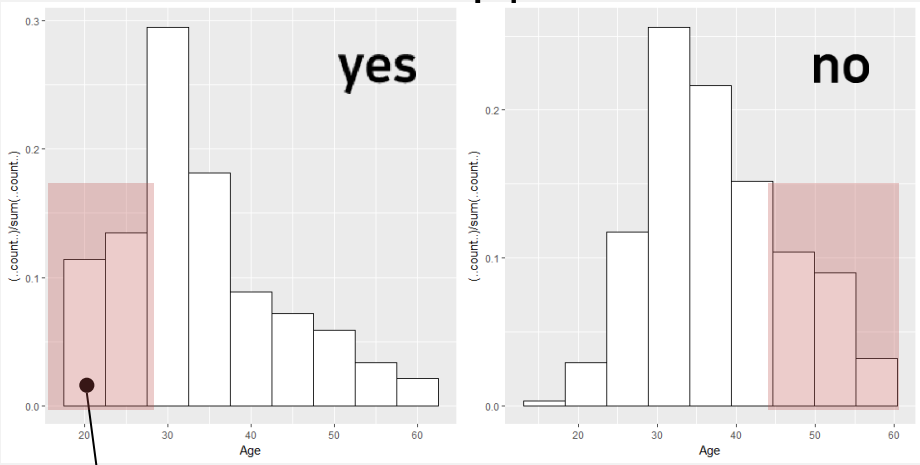
순서형 변수 처리

1. 현재 직장과 무관하게 정해지는 변수

<결혼 수준>



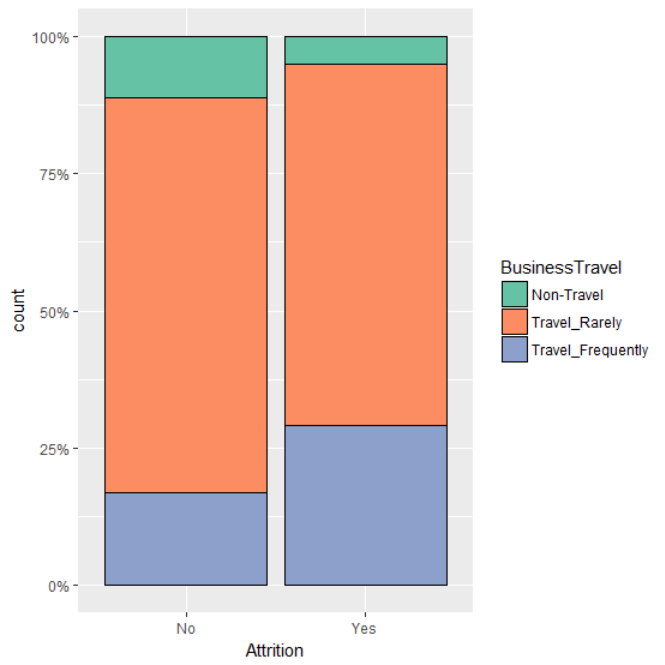
<나이>



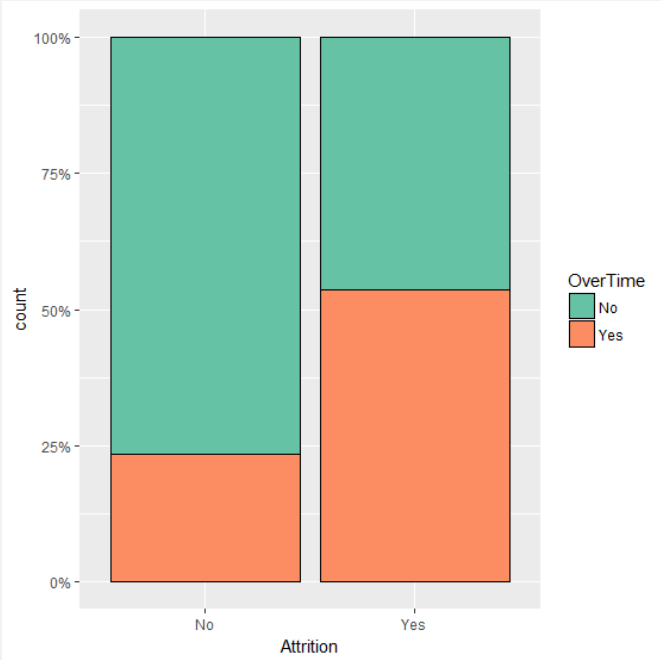
이직을 하지 않은 사람들 보다
이직을 한 사람들의 연령대가 더 낮음

2. 입사 후에 정해지는 변수

<출장>

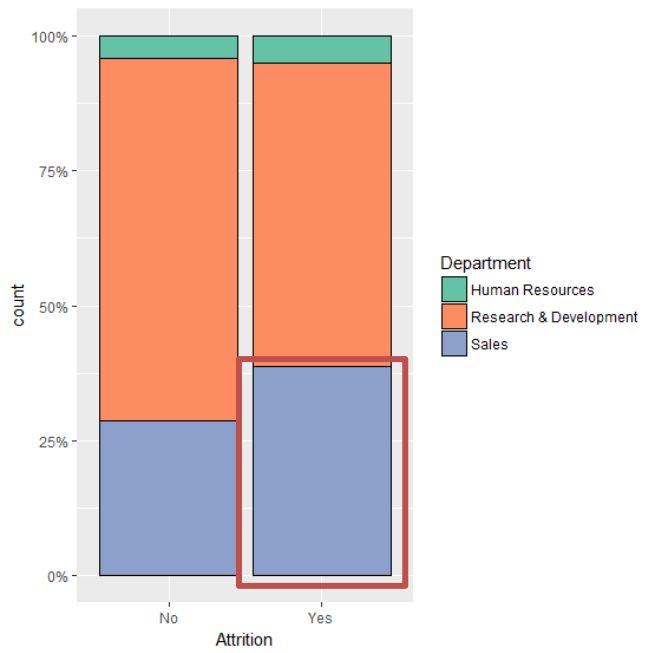


<초과 근무>



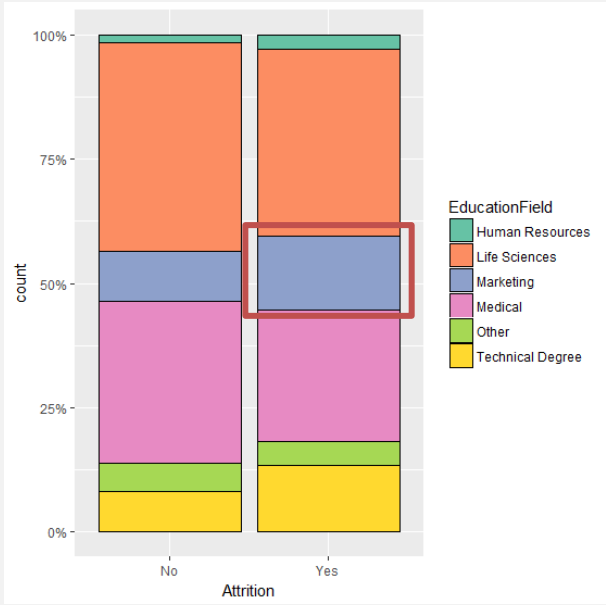
2. 입사 후에 정해지는 변수 – 업무

<부서>



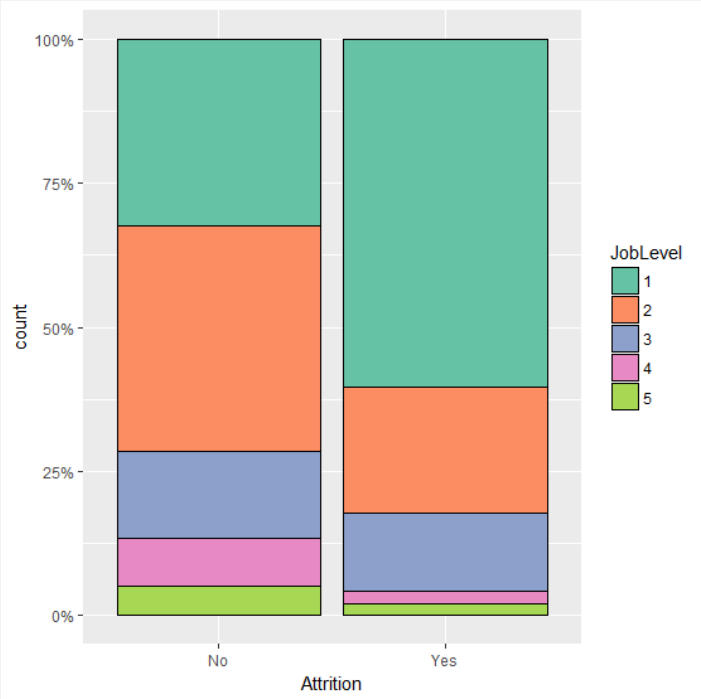
Marketing을 전공한
영업직의 직원들이
이직을 많이 하는 것은 아닐까?

<전공>

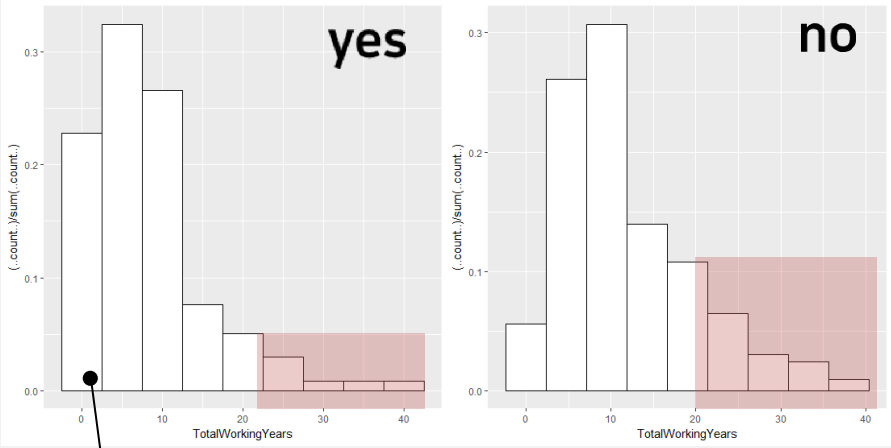


2. 입사 후에 정해지는 변수 – 업무

<직급>



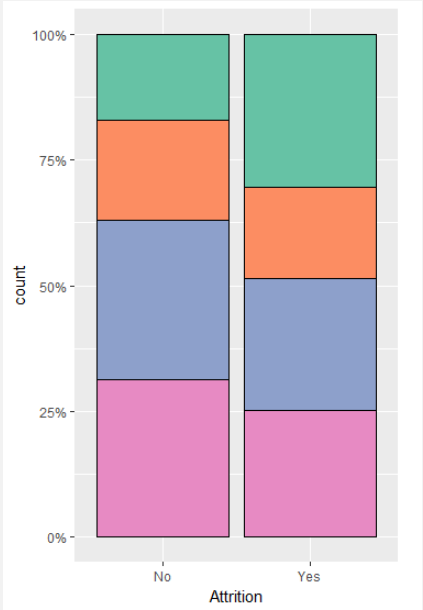
<총 근무 기간>



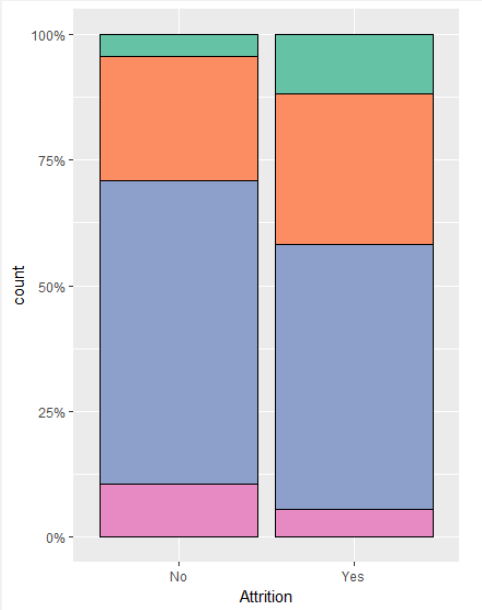
업무 상 어느 정도 안정적으로
되었을 때 이직을 하지 않는다

2. 입사 후에 정해지는 변수 – 만족도

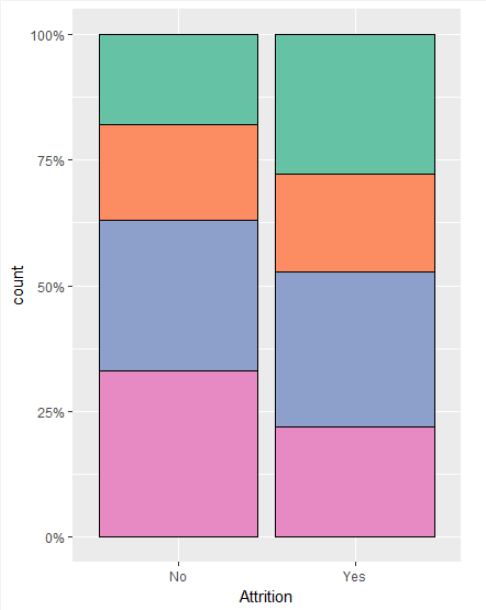
<환경만족도>



<업무 소속감>



<직업만족도>



매우 불만족

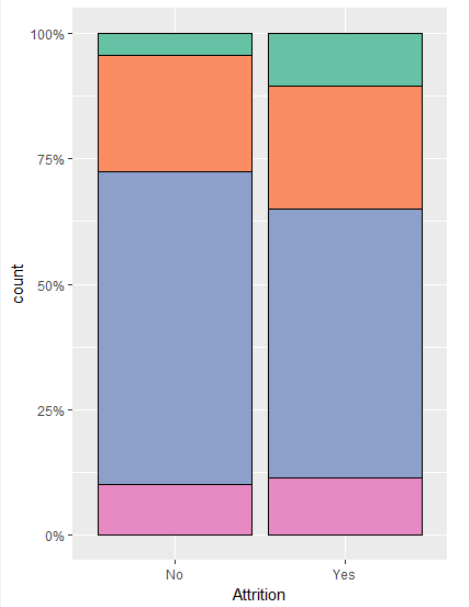
매우 만족

2. 입사 후에 정해지는 변수 – 만족도

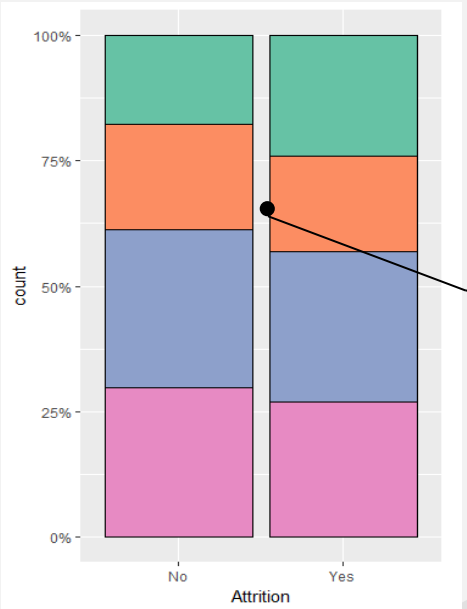
매우 불만족

매우 만족

<일과 삶의 균형>

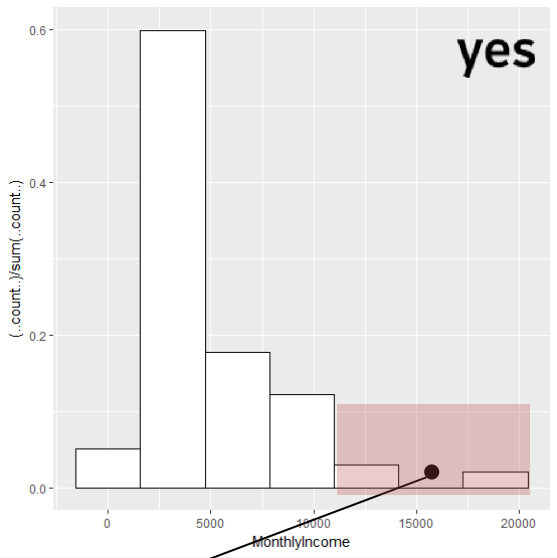


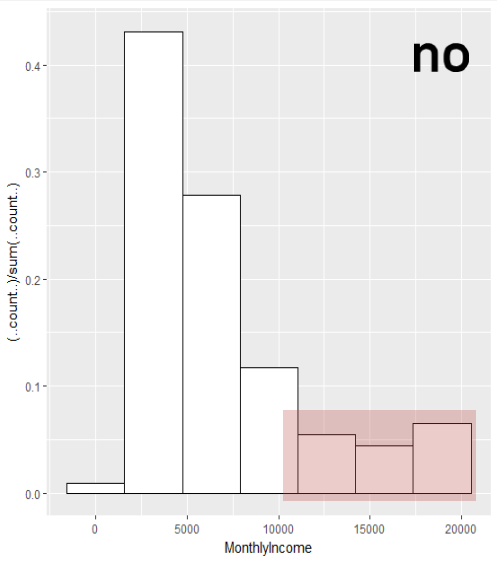
<관계만족도>



이 두 변수는 이직한 사람과 이직하지 않은 사람 사이의 차이가 적음

3. 임금과 관련된 변수





대체적으로 월수입이 높은 사람보다 낮은 사람이 이직한 비율이 높음

04 LOGISTIC REGRESSION

step0 Over18 변수 factor 해제

Factor의 level이 하나인 경우는 Logistic Regression이 실행되지 않음

step1 단계적 알고리즘을 사용해 AIC를 기준으로 모델 선택

```
null<-glm(Attrition~1,data=dat,family="binomial")
full<-glm(Attrition~.,data=dat,family="binomial")
step(null,scope = list(lower=null,upper=full),direction = "both")
```

```
Call:  glm(formula = Attrition ~ OverTime + JobRole + MaritalStatus +
    EnvironmentSatisfaction + JobSatisfaction + JobInvolvement +
    BusinessTravel + YearsInCurrentRole + YearsSinceLastPromotion +
    DistanceFromHome + NumCompaniesWorked + Age + WorkLifeBalance +
    RelationshipSatisfaction + TrainingTimesLastYear + YearsWithCurrManager +
    Gender + EducationField + TotalWorkingYears + YearsAtCompany +
    StockOptionLevel, family = "binomial", data = dat)
```

⇒ 선택된 변수가 너무 많음

step2 BIC를 기준으로 한 번 더 단계적 변수 선택

```
step(null,scope = list(lower=null,upper=aic.fit),direction = "both",k=log(nrow(dat)))
```

Penalty가 더 강한 BIC로 stepwise

최종 선택된 모형 (총 12개의 변수 선택)

```
Call:  glm(formula = Attrition ~ OverTime + TotalWorkingYears + MaritalStatus +
    EnvironmentSatisfaction + JobSatisfaction + JobInvolvement +
    BusinessTravel + NumCompaniesWorked + DistanceFromHome +
    YearsSinceLastPromotion + YearsInCurrentRole + RelationshipSatisfaction,
    family = "binomial", data = dat)
```

: 초과 근무, 총 근무 년 수, 결혼 상태, 환경만족도, 직업만족도, 직업소속감, 출장, 일한 회사의 개수, 집과의 거리, 지난 승진부터 지난 년 수, 직무 연차, 관계 만족도

```
Coefficients:
                                Estimate Std. Error z value Pr(>|z|)
(Intercept)                     0.509338   0.609768   0.835 0.403550
OverTimeYes                     1.755423   0.175447  10.005  < 2e-16 ***
TotalWorkingYears              -0.111046   0.017428  -6.372 1.87e-10 ***
MaritalStatusMarried            0.323972   0.243162   1.332 0.182753
MaritalStatusSingle             1.299079   0.243304   5.339 9.33e-08 ***
EnvironmentSatisfaction        -0.393657   0.076620  -5.138 2.78e-07 ***
JobSatisfaction                -0.382299   0.075345  -5.074 3.90e-07 ***
JobInvolvement                 -0.588540   0.115341  -5.103 3.35e-07 ***
BusinessTravelTravel_Frequently 1.862590   0.394330   4.723 2.32e-06 ***
BusinessTravelTravel_Rarely     1.070638   0.367880   2.910 0.003611 **
NumCompaniesWorked              0.152567   0.034432   4.431 9.38e-06 ***
DistanceFromHome                0.039182   0.009934   3.944 8.00e-05 ***
YearsSinceLastPromotion         0.177097   0.035831   4.943 7.71e-07 ***
YearsInCurrentRole             -0.134431   0.035535  -3.783 0.000155 ***
RelationshipSatisfaction       -0.246547   0.076561  -3.220 0.001281 **
```

대부분 유의한 변수 선택

step3 Likelihood Test (모형 적합성 검정)

```
> lrtest(bic.fit)
Likelihood ratio test

Model 1: Attrition ~ OverTime + TotalWorkingYears + MaritalStatus + EnvironmentSatisfaction
+
    JobSatisfaction + JobInvolvement + BusinessTravel + NumCompaniesWorked +
    DistanceFromHome + YearsSinceLastPromotion + YearsInCurrentRole +
    RelationshipSatisfaction
Model 2: Attrition ~ 1
  #Df  LogLik  Df  Chisq Pr(>Chisq)
1  15 -478.48
2   1 -649.29 -14 341.63  < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

null model 과 fit model 비교해서 모델의 적합도 검정

⇒ p-value가 매우 작다.

⇒ fit model이 유의하다.

step4 Durbin-Watson 검정 (독립성 검정)

```
> dwtest(bic.fit)

        Durbin-Watson test

data:  bic.fit
DW = 1.9212, p-value = 0.06553
alternative hypothesis: true autocorrelation is greater than 0
```

$$Y_t = \beta_0 + \beta_1 X_t + u_t$$

$$u_t = \alpha u_{t-1} + \omega_t$$

$$u_t = \omega_t + \alpha\omega_{t-1} + \alpha u_{t-2}$$

$$u_t = \omega_t + \alpha\omega_{t-1} + \alpha^2\omega_{t-2} + \alpha^2 u_{t-3}$$

$$u_t = \omega_t + \alpha\omega_{t-1} + \alpha^2\omega_{t-2} + \alpha^3\omega_{t-3} + \dots$$

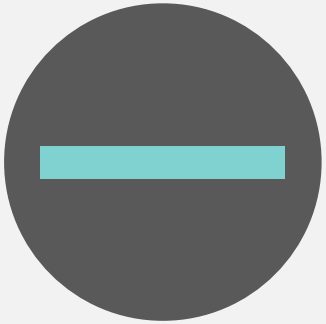
OLS method에서
Autocorrelation (자기상관성) 없다는 가정 필요

H_0 : α = 0 ⇔ 가정 만족

p-value 0.064 ⇔ 가정 약하게 만족

OverTime (초과 근무)
MaritalStatus (결혼 상태)
BusinessTravel (출장)
NumCompaniesWorked (일했던 직장 수)
DistanceFromHome (집과의 거리)
YearsSinceLastPromotion (승진 후 경과 시간)

당장 IBM을 떠날 거야!



TotalWorkingYears (총 근무 기간)
EnvironmentSatisfaction (환경 만족도)
JobInvolvement (직업 소속감)
JobSatisfaction (직업 만족도)
YearsInCurrentRole (직무 연차)
RelationshipSatisfaction (관계 만족도)

IBM에 계속 있어야지!

1. 어떤 변수가 이직에 중요한 영향을 주는지 알아볼 수 없을까?
   (Tree 모형)

2. IBM 직원들의 지나친 이직을 막을 수 있는 방법 제안

[THANK YOU]