HR analysis

# 배경

OOO회사 2021년 9월 이직률은 약 23%(총 종업원 14,999명 중 3,571명)이다.

table(HR$left)

0 1   
 11428 3571

이직을 낮추기 위하여 각 변수들 간 상관관계를 분석하여 해결방법을 제시하였다.

**도입**

## import library

library(ggplot2)

library(dplyr) 필요한 라이브러리를 가져온다.

## import Dataset

HR = read.csv(‘HR\_comma\_sep.csv’) 데이터 셋을 입력한다.

## check Dataset

head(HR, 5)

satisfaction\_level last\_evaluation number\_project average\_montly\_hours  
 1 0.38 0.53 2 157  
 2 0.80 0.86 5 262  
 3 0.11 0.88 7 272  
 4 0.72 0.87 5 223  
 5 0.37 0.52 2 159  
 time\_spend\_company Work\_accident left promotion\_last\_5years sales salary  
 1 3 0 1 0 sales low  
 2 6 0 1 0 sales medium  
 3 4 0 1 0 sales medium  
 4 5 0 1 0 sales low  
 5 3 0 1 0 sales low

summary(HR)

satisfaction\_level last\_evaluation number\_project average\_montly\_hours  
 Min. :0.0900 Min. :0.3600 Min. :2.000 Min. : 96.0   
 1st Qu.:0.4400 1st Qu.:0.5600 1st Qu.:3.000 1st Qu.:156.0   
 Median :0.6400 Median :0.7200 Median :4.000 Median :200.0   
 Mean :0.6128 Mean :0.7161 Mean :3.803 Mean :201.1   
 3rd Qu.:0.8200 3rd Qu.:0.8700 3rd Qu.:5.000 3rd Qu.:245.0   
 Max. :1.0000 Max. :1.0000 Max. :7.000 Max. :310.0   
 time\_spend\_company Work\_accident left promotion\_last\_5years  
 Min. : 2.000 Min. :0.0000 Min. :0.0000 Min. :0.00000   
 1st Qu.: 3.000 1st Qu.:0.0000 1st Qu.:0.0000 1st Qu.:0.00000   
 Median : 3.000 Median :0.0000 Median :0.0000 Median :0.00000   
 Mean : 3.498 Mean :0.1446 Mean :0.2381 Mean :0.02127   
 3rd Qu.: 4.000 3rd Qu.:0.0000 3rd Qu.:0.0000 3rd Qu.:0.00000   
 Max. :10.000 Max. :1.0000 Max. :1.0000 Max. :1.00000   
 sales salary   
 Length:14999 Length:14999   
 Class :character Class :character   
 Mode :character Mode :character

str(HR)

'data.frame': 14999 obs. of 10 variables:  
 $ satisfaction\_level : num 0.38 0.8 0.11 0.72 0.37 0.41 0.1 0.92 0.89 0.42 ...  
 $ last\_evaluation : num 0.53 0.86 0.88 0.87 0.52 0.5 0.77 0.85 1 0.53 ...  
 $ number\_project : int 2 5 7 5 2 2 6 5 5 2 ...  
 $ average\_montly\_hours : int 157 262 272 223 159 153 247 259 224 142 ...  
 $ time\_spend\_company : int 3 6 4 5 3 3 4 5 5 3 ...  
 $ Work\_accident : int 0 0 0 0 0 0 0 0 0 0 ...  
 $ left : int 1 1 1 1 1 1 1 1 1 1 ...  
 $ promotion\_last\_5years: int 0 0 0 0 0 0 0 0 0 0 ...  
 $ sales : chr "sales" "sales" "sales" "sales" ...  
 $ salary : chr "low" "medium" "medium" "low" ...

dim(HR)

[1] 14999 10

sapply(HR,class)

satisfaction\_level last\_evaluation number\_project   
 "numeric" "numeric" "integer"   
 average\_montly\_hours time\_spend\_company Work\_accident   
 "integer" "integer" "integer"   
 left promotion\_last\_5years sales   
 "integer" "integer" "character"   
 salary   
 "character"

## change variable’s Class for analysis

분석을 위해서 class를 변경해준다.

- satisfaction\_level : 0~1사이의 값으로 표현됨. (만족 0% ~ 100%로 이해하면 편하기 떄문에 num에서 타입을 유지함.)

- last\_evaluation : 0~1사이의 값으로 표현됨. (평가 0% ~ 100%로 이해하면 편하기 떄문에 num에서 타입을 유지함.)

- number\_project : 2~7사이의 정수로 구성되어 있음. (이는 연속변수가 아님. 따라서 factor로 하는 것이 좋다고 판단함)

- average\_monthly\_hours : 96~310시간으로 구성되어 있음 (연속변수는 아니지만, 214개의 값이 들어감. int로 유지)

- time\_spend\_company : 재직 연수로 2년~10년임. (이는 연속변수가 아님. 따라서 factor로 하는 것이 좋다고 판단함)

- Work\_accident : 사고 유무로 판단하는데 0,1값만 존재 (logical로 설 판단함. True, False이므로)

- left : 이직유무도 0.1만 존재함. (logical로 설정하는 것이 좋음. True, False이므로)

- promotion\_last\_5years : 0,1만 존재함. (logical로 설정하는 것이 좋음. True, False이므로)

- sales : 직군으로 구성되어 있음. (일반적인 문자열이지만, factor로 보는 것이 좋음. 8개 직군이므로)

- salary : high, low, medium으로 3가지로 나뉨. (문자열이지만 factor로 설정하는 것이 더 분석에 어울림.)

따라서 satisfaction\_level, last\_evaluation, average\_monthly\_hours의 타입은 유지하고 나머지 7개를 변경해준다.

## Check

잘 바뀌었는지 확인한다.

sapply(HR,class)

**satisfaction\_level last\_evaluation number\_project**   
 "numeric" "numeric" "factor"   
 **average\_montly\_hours time\_spend\_company Work\_accident**   
 "integer" "factor" "logical"   
 **left promotion\_last\_5years sales**   
 "logical" "logical" "factor"   
 **salary**   
 "factor"

## Corr Matrix (만족도, 인사평가, 월 평균근속시간 )

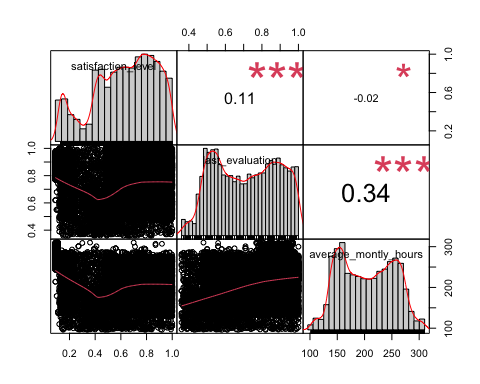
1. 분석의 편의를 위해서 새로운 DF를 추출하고 cor matrix를 뽑아본다.
2. correlaion matrix를 출력한다. (변수는 만족도, 인사평가, 월평균 근로시간)

cor(HR\_cor)

satisfaction\_level last\_evaluation average\_montly\_hours  
 satisfaction\_level 1.00000000 0.1050212 -0.02004811  
 last\_evaluation 0.10502121 1.0000000 0.33974180  
 average\_montly\_hours -0.02004811 0.3397418 1.00000000

1. corr matrix + performance\_Analysis

chart.Correlation(HR\_cor, histogram=TRUE, pch=1)



여기서 우리는 만족도와 인사평가의 상관성이 0.11로미약하게 존재하고 인사평가와 평균 근무 시간의 상관성이 0.34로 신뢰할만한 수준임을 파악할 수 있었다.

이제 이직율을 낮추기 위한 솔루션을 도출하고자 한다.

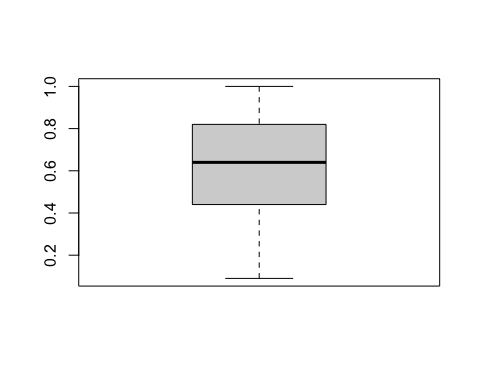
# 분석파트

## 1. 적절한 업무량과 업무시간을 도출.

### 1-1. 프로젝트 수와 만족도 상관관계 분석

* 전직원들의 평균 업무량과 월 평균 근무시간 도출

boxplot(HR$satisfaction\_level)  
boxplot(HR$satisfaction\_level)$stat



[,1]  
 [1,] 0.09  
 [2,] 0.44  
 [3,] 0.64  
 [4,] 0.82  
 [5,] 1.00

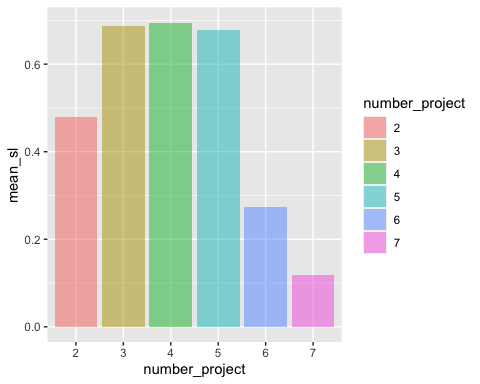
전체 직원의 만족도 중앙값은 0.64임을 확인할 수 있다.

hr1 : 프로젝트 개수별로 출력한 만족도의 평균치

hr1  
 number\_project mean\_sl  
 1 2 0.479  
 2 3 0.688  
 3 4 0.695  
 4 5 0.679  
 5 6 0.273  
 6 7 0.119

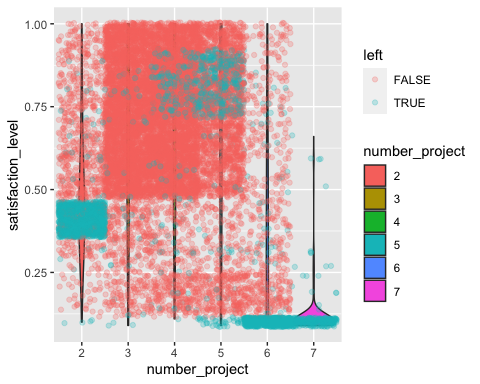
hr1으로 출력한 col\_plot (프로젝트 개수 별 만족도의 평균치)

ggplot(data = hr1, aes(x=number\_project, y=mean\_sl)) +   
 geom\_col(aes(fill=number\_project), alpha = 0.5)



우리는 이직을 결정하는 요소를 보는 것이기 때문에 이직여부를 추가하여 분석에 활용할 필요가 있음.

ggplot(data = HR, aes(x=number\_project, y=satisfaction\_level)) +   
 geom\_violin(aes(fill=number\_project), trim=TRUE,   
 adjust=10) +  
 geom\_jitter(position = position\_jitter(.5),  
 aes(col = left),  
 alpha=0.25)



**전체직원 만족도 중앙값 0.6**

2개 : 평균만족도는 0.479. (어느정도 자신의 일이 주어져야 만족도가 높은 듯)

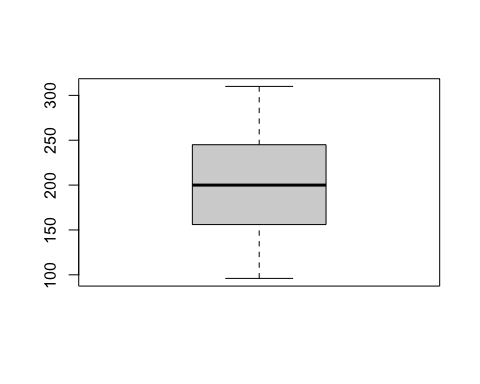
3~5개 : 평균만족도 0.68 ~ 0.695

6개 : 평균만족도 : 0.119 ~ 0.273

#### 1-1. conclusion

* 직원들의 평균 업무량은 3~5개가 적당
* 하지만, 이직을 보면, 4~5개 프로젝트를 수행한 인원 중 만족도가 높은 부분에 이직이 발생한 것도 주목해야 함.

### 1-2. 평균근무시간과 만족도 상관관계 분석

boxplot(HR$average\_montly\_hours)  
boxplot(HR$average\_montly\_hours)$stat

[,1]  
 [1,] 96  
 [2,] 156  
 [3,] 200  
 [4,] 245  
 [5,] 310

전체직원의 근무시간 중앙값 200

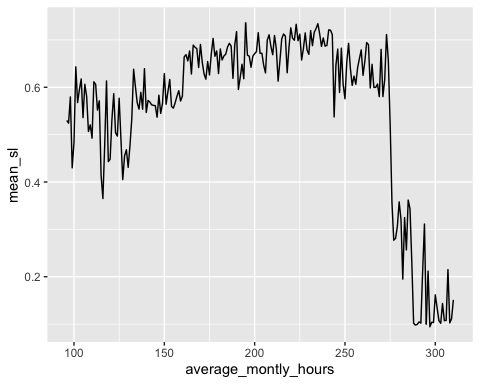
hr2 : 근무 시간 별 만족도의 평균 수치를 출력

hr2

# A tibble: 215 × 2  
 average\_montly\_hours mean\_sl  
 <int> <dbl>  
 1 195 0.737  
 2 235 0.734  
 3 223 0.733  
 4 220 0.725  
 5 234 0.724  
 6 241 0.721  
 7 242 0.720  
 8 231 0.720  
 9 190 0.718  
 10 233 0.716  
 # … with 205 more rows

hr2로 출력힌 line\_plot (근무 시간 별 만족도의 평균치)

ggplot(data = hr2, aes(x=average\_montly\_hours, y=mean\_sl)) + geom\_line()

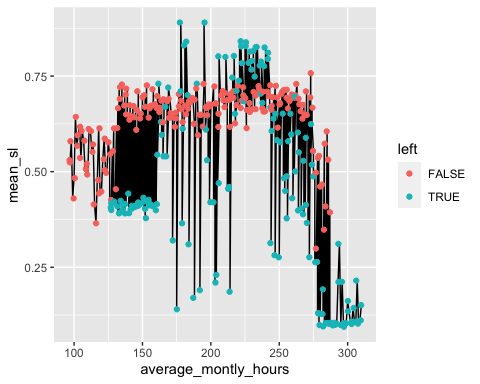


hr3 : 근무 시간을 순차적으로 나열하고 이직 여부를 반영한, 만족도 평균치를 추출.

hr3

# A tibble: 356 × 3  
 # Groups: average\_montly\_hours [215]  
 average\_montly\_hours left mean\_sl  
 <int> <lgl> <dbl>  
 1 96 FALSE 0.53   
 2 97 FALSE 0.524  
 3 98 FALSE 0.580  
 4 99 FALSE 0.43   
 5 100 FALSE 0.483  
 6 101 FALSE 0.643  
 7 102 FALSE 0.568  
 8 103 FALSE 0.595  
 9 104 FALSE 0.618  
 10 105 FALSE 0.536  
 # … with 346 more rows

hr3로 출력힌 col\_plot (근무 시간 별 만족도의 평균치)

 여기서는 만족도가 너무 높아도, 너무 낮아도 이직이 발생하는 것을 확인할 수 있다. 대체적으로 사람들은 0.625 ~ 0.75 정도의 만족도를 가진 경우 회사에 남아있는 것을 택한다. (월 근무시간은 130~275시간 사이)

* 전체직원의 근무시간 중앙값 200
* 100~150 hours : 0.4 ~ 0.6
* 150~275 hours : higher satisfaction (upper 0.6)
* upper 275 : dramatically lower satisfaction (0.2 ~ 0.4)

#### 1-2. conclusion

* 직원들의 월평균 근무시간은 150시간에서 275시간 유지 필요

### 1. conclusion

* **프로젝트 개수를 3~5개로 낮춰서** 업무 가중을 경감시킬 필요가 있음.
  + 하지만, 4~5개의 업무를 수행한 인원 중 이직이 높은 현상도 있었음.
* 월평균 **근무시간을 조정**하여 적절한 워라밸을 형성해야 함.

### 2.이직을 낮추기 위한 솔루션

분석에 앞서 우리는 이직을 결정하는 변수가 어떤 것이 있는지 상관관계 분석을 통해 도출하고자 한다.

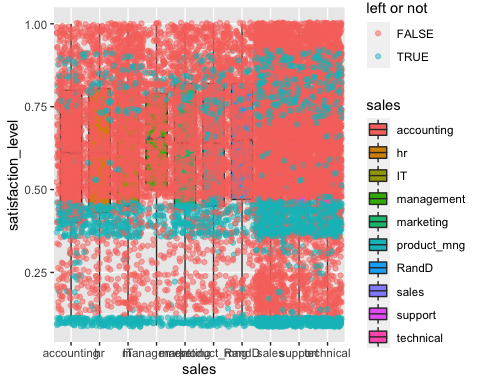
이직여부기준 상관관계 분석이 필요할 것이다.

먼저 이직이 부서별로 차이가 있는지, 또 그 부서의 만족도는 어떠한지 분포를 살펴보았다.

1. 부서가 이직에 미치는 영향
2. 급여가 이직에 미치는 영향
3. 프로젝트 수가 이직에 미치는 영향

#### 2-1. 이직여부를 기준으로 부서 및 만족도 상관관계조사

table(HR$left) FALSE : TRUE = 11428 : 3571



해석

\* 대체로 만족도는 0.38 ~ 1.00에 많이 포진되어 있는 편임. 또한 0.05~0.1 에 위치한 사람들도 많이 보임. 직군 중에서는 sales, support, technical에서 사람들이 많이 있는 듯 함.

+ **여기서 주목할 점은 만족도가 0.75~0.93, 0.35~0.44?, 0.1 미만인 곳에서 이직이 이뤄진다는 점이다.**

+ 이는 전 직군에 공통적으로 발생하고 있다. 또한, sales, suppport, technical에서는 상대적으로 주목할 부분 외에도 이직이 간간히 발생한다.

+ 이에 대한 해석으로는 아무래도 직군에 종사하는 인원의 차이 때문인 듯 하다. 따라서 인원이 얼마나 있는지 살펴보았다.

HR\_count

HR$sales n  
 1 accounting 767  
 2 hr 739  
 3 IT 1227  
 4 management 630  
 5 marketing 858  
 6 product\_mng 902  
 7 RandD 787  
 8 sales 4140  
 9 support 2229  
 10 technical 2720

##### 2-1. conclusion

추론한 결과로는 **sales(4140), technical(2720), support(2229), IT(1227)** 순으로 많았음. 하지만, 직군이 만족도에 영향을 미치는 편은 적은 듯 하다.

따라서, 잠깐 직군이 만족도에 미치는 영향을 살펴보았다.

ggplot(HR\_test1, aes(x=sales, y=mean\_sl))+  
 geom\_col(aes(fill=sales), alpha=.5, show.legend = TRUE)+  
 ylim(0,1)



직군 별 만족도의 평균은 어떤지 비교해보았지만 0.58~0.621로 오차 수준도 작은 편이었다.

**따라서, 직군과 만족도의 관계는 없다고 해석하는 것이 좋다고 판단하였다.**

#### 2-2. 이직여부를 기준으로 임금수준 도출, 임금 수준에 따른 만족도 구분

이번에는 임금수준이 이직에 어떤 영향을 주는지 살펴보았다. (+그래프에 만족도를 추가)

해석

이직을 3가지 케이스로 구분(만족도를 1분위 ~ 3분위수로 표현하였다.)

\* (1) low : 0.23 ~ 0.73

\* (2) medium : 0.125 ~ 0.75

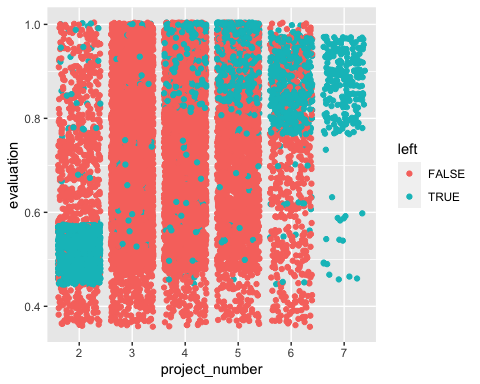
\* (3) high : 0.375 ~ 0.45

##### 2-2. conclusion

**“low”와 “medium”의 만족도를 높이기 위한 솔루션 필요**

#### 2-3. 이직여부 기준 프로젝트 수, 평가점수 그래프

ggplot(HR,aes(x=number\_project, y=last\_evaluation)) +  
 geom\_jitter(aes(col = left)) +   
 labs(col = "left") + xlab("project\_number") + ylab("evaluation")



해석

(1) 2개 프로젝트를 수행하면서 평가점수가 0.58 이하에서 이직이 발생했다.

(2) 4 ~ 5개의 프로젝트를 수행하면서 평가점수가 0.8 이상에서 이직이 발생했다.

(3) 7개의 프로젝트를 수행하는 인원은 전부 이직을 선택하였다.

##### 2-3. conclusion

* 2개 프로젝트를 수행하는 직원의 만족도를 높이기 위해 업무 배정의 필요성 존재.
* 7개 프로젝트를 수행하는 직원의 이직 방지를 솔루션 필요
* 7개 프로젝트를 수행하는 직원의 업무 강도를 낮춰주는 형태도 중요

#### 2. conclusion

1. 부서가 이직에 미치는 영향 : 미비하다.
2. 급여가 이직에 미치는 영향 : 급여가 높지 않은 중간, 하위 계층의 만족도를 높여야 이직을 감소시킬 수 있다.
3. 프로젝트 수가 이직에 미치는 영향 : 너무 적은 업무와 과도한 업무는 이직을 유발한다. 따라서 업무를 재배치할 필요가 있다.