Human Sources analysis

Mingwei (Show) Wu

Background

"Turnover" : turnover is that churn refers to the gradual loss of employees over a period of time. In a company, the employee turnover is the biggest issue facing HR and high costs Therefore, analysis the employee turnover is the way to prevent the damange and save mony to company. Usually, the comon reasons employee turnover is better opportunity, healh, rolocation, eaducation, and personal reasons etc.In addition, some hidden reasons of employees turnover includes percent salary hike, overtime, travel distance, career satisfaction, tenure, and supervisor's personality etc.

Data

emp_id: employees id

status: working status, Active and Inactive

location: location of working city

level: Job level in Company

gender: Male and Female

emp_age: employees age

rating: Internal work evaluation level

mar_rating: employees' manager internal work evaluation level

mgr_reportees: employees' manager report

mgr_age: employees' manager age

mgr_tenure: employees' manager tenure

compensation: salary

percent_hike: precentage of increase salary

hiring_score: hire interview score

hiring_sourece: platform for job

no_previous_companies_worked: number of previous work rompaneies

distance_from_homne: distance between home and work place

total_dependets: number of dependets

marital_status: status of marry

education: education level

promotion_last_2_years: the promotion of employee within last 2 years

no_leaves_taken: number of leaves have been taken

total_experience: total of work experience

monthly_overtime_hrs: total number of monthly overtime hours

date_of_joining: date of join the companey

the question, use the MySQL to select the dataset or use R to import the data for manipulation. Counting the status then use R to filter the highest proportion level to track. Also, check the significant factors for reason of turnover.

Data Importing

Importing the original data of organization, deal with 1954 observations and 34 column.

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union

## -- Attaching packages --------------------------------------- tidyverse 1.3.0 --

## v ggplot2 3.3.2     v purrr   0.3.4
## v tibble  3.0.4     v stringr 1.4.0
## v tidyr   1.1.2     v forcats 0.5.0

## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()

## [1] 1954   34

## [1] 1954   34
```

Fomular for turnover rate

Turnover rate = Number of employees who left / Total number of employees

counting the status from data frame. we know the active employee is 1557, and 397 employees left the company.

```
##     status    n
## 1   Active 1557
## 2 Inactive  397
```

calculate a mean of turnover_rate. the rate is approximation 18% for employees left

```
##   turnover_rate
## 1      0.203173
```

Approximation 22% of Analyst job level leaving and 15% of Specialist level leaving end of 12/31/2014 in the company.

```
## `summarise()` ungrouping output (override with `.groups` argument)
```

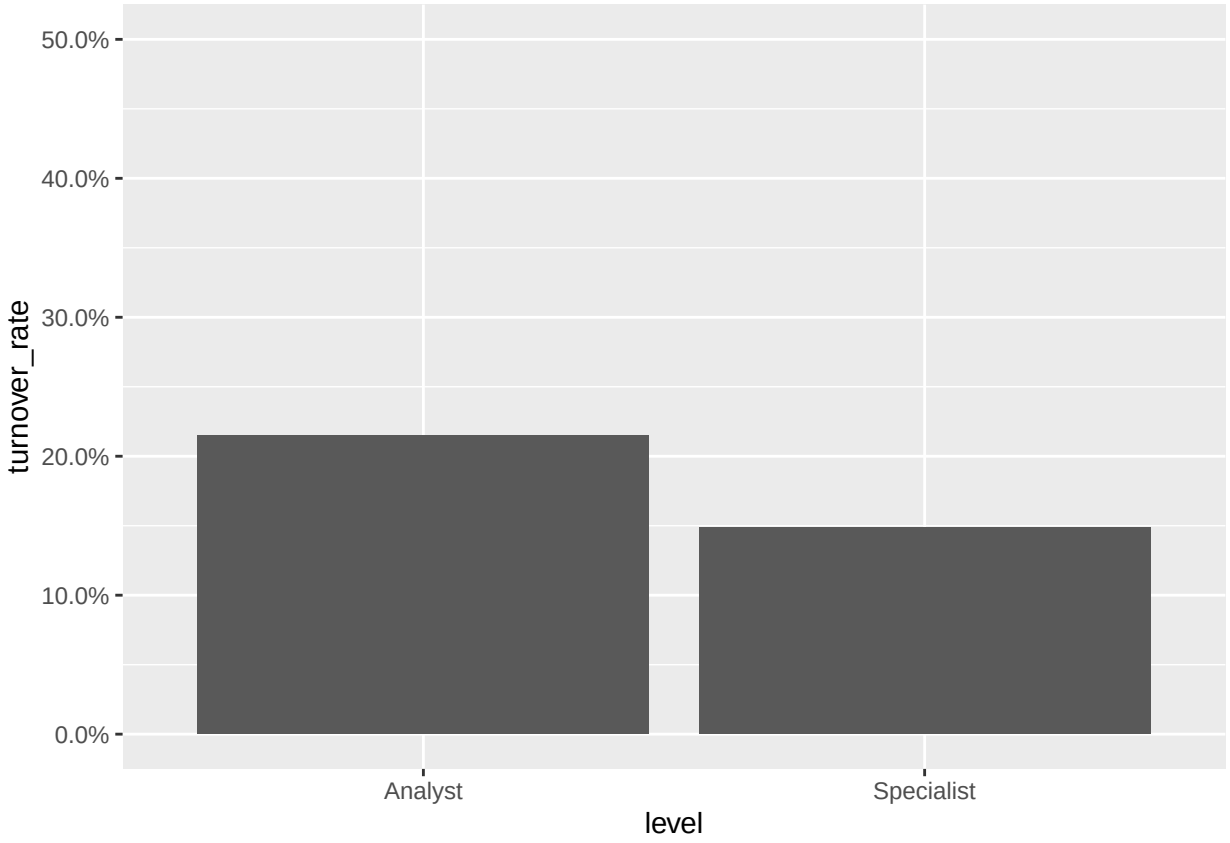
```
## # A tibble: 2 x 2
##   level      turnover_rate
##   <chr>              <dbl>
## 1 Analyst            0.215
## 2 Specialist         0.149
```

use graph for data visulization. the graph is showing the rate value between the analyst level and spcialist level.

```
##
## Attaching package: 'scales'
```

```
## The following object is masked from 'package:purrr':
##
##     discard
```

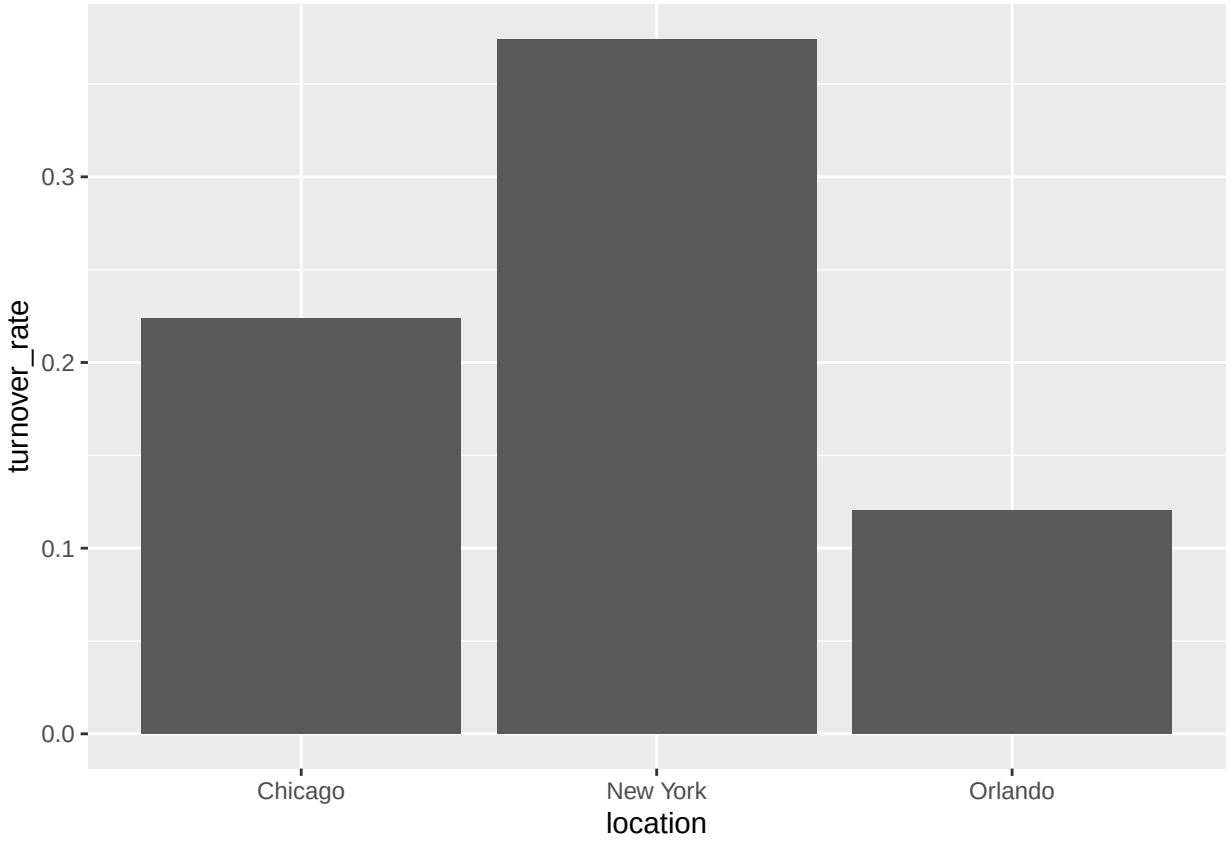
```
## The following object is masked from 'package:readr':
##
##     col_factor
```



```
## `summarise()` ungrouping output (override with `.groups` argument)
```

```
## # A tibble: 3 x 2
##   location turnover_rate
##   <chr>            <dbl>
## 1 Chicago          0.224
## 2 New York         0.374
## 3 Orlando          0.121
```

```
location%>%  #histogram to data visulization
  ggplot(aes(location,turnover_rate))+ geom_col()
```



chekcing rating relationship for turnover_rate

from the data calculating, the internal work evulation rate are showing that the unacceptable is 63% highest proportion of turnover_rate. The number 2 higher proportion is below average rating. On the contraction, the acceptable is 22%, above average 13% and excellent only 3%

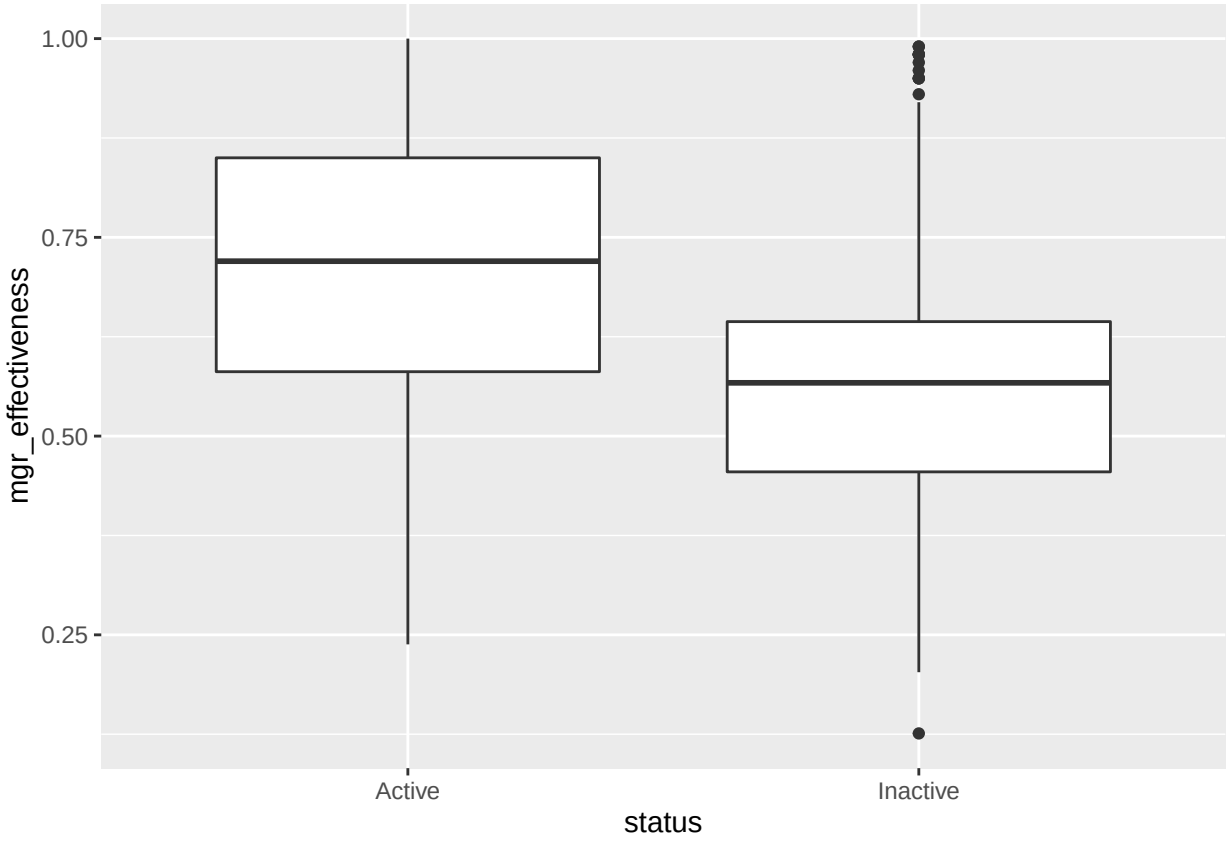
```
## `summarise()` ungrouping output (override with `.groups` argument)
```

```
## # A tibble: 5 x 2
##   rating        turnover_rate
##   <chr>                 <dbl>
## 1 Above Average         0.131
```

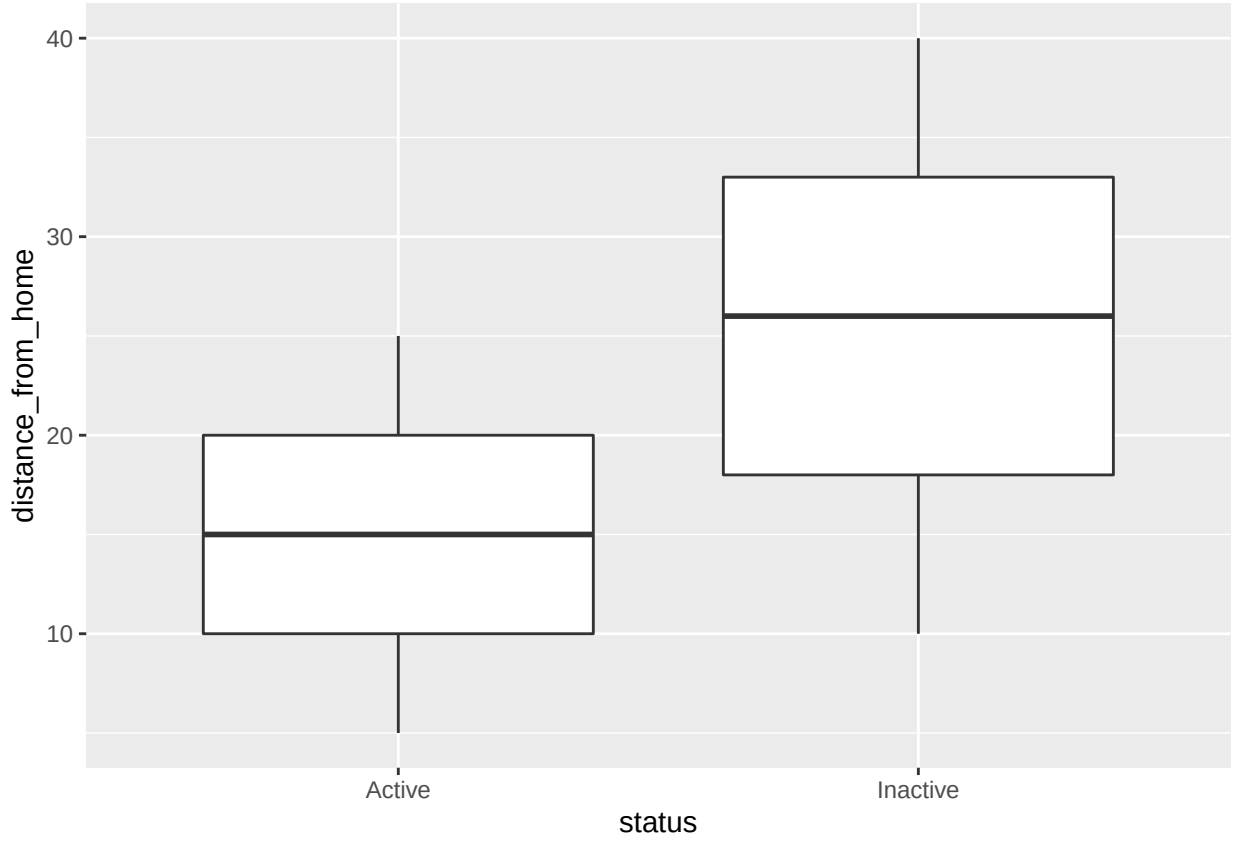
```
## 2 Acceptable           0.221
## 3 Below Average        0.385
## 4 Excellent            0.0305
## 5 Unacceptable         0.633
```

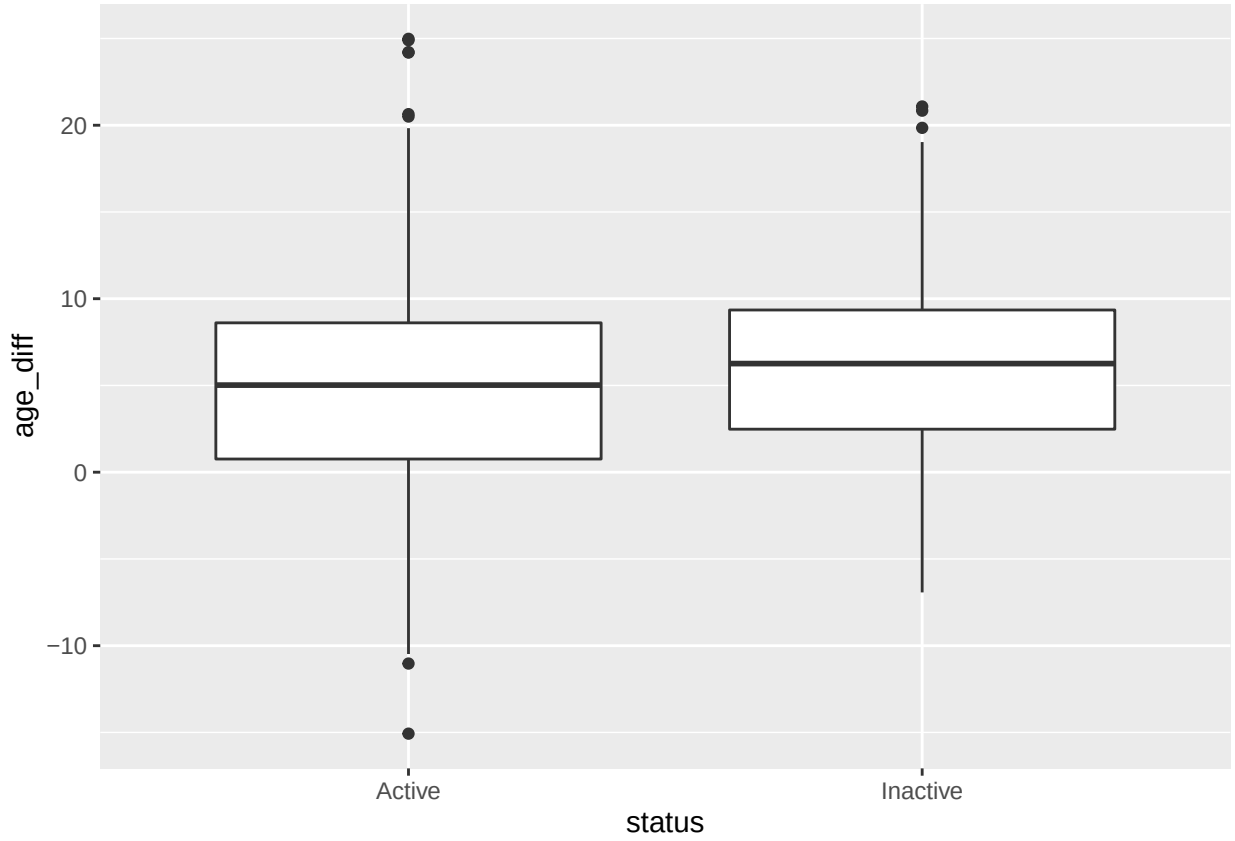
Question: is the work evulation as main factor for employee turnover? the employee was fire by company?

the grph showing manager effectiveness also affct the employee turnover. the box-plot shows the outliers of inactive. however, overall shows that the mean of manger effectiveness of active is higher than inactive. according to box-plot, we know the manger effectiveness also affect turnover of employees.

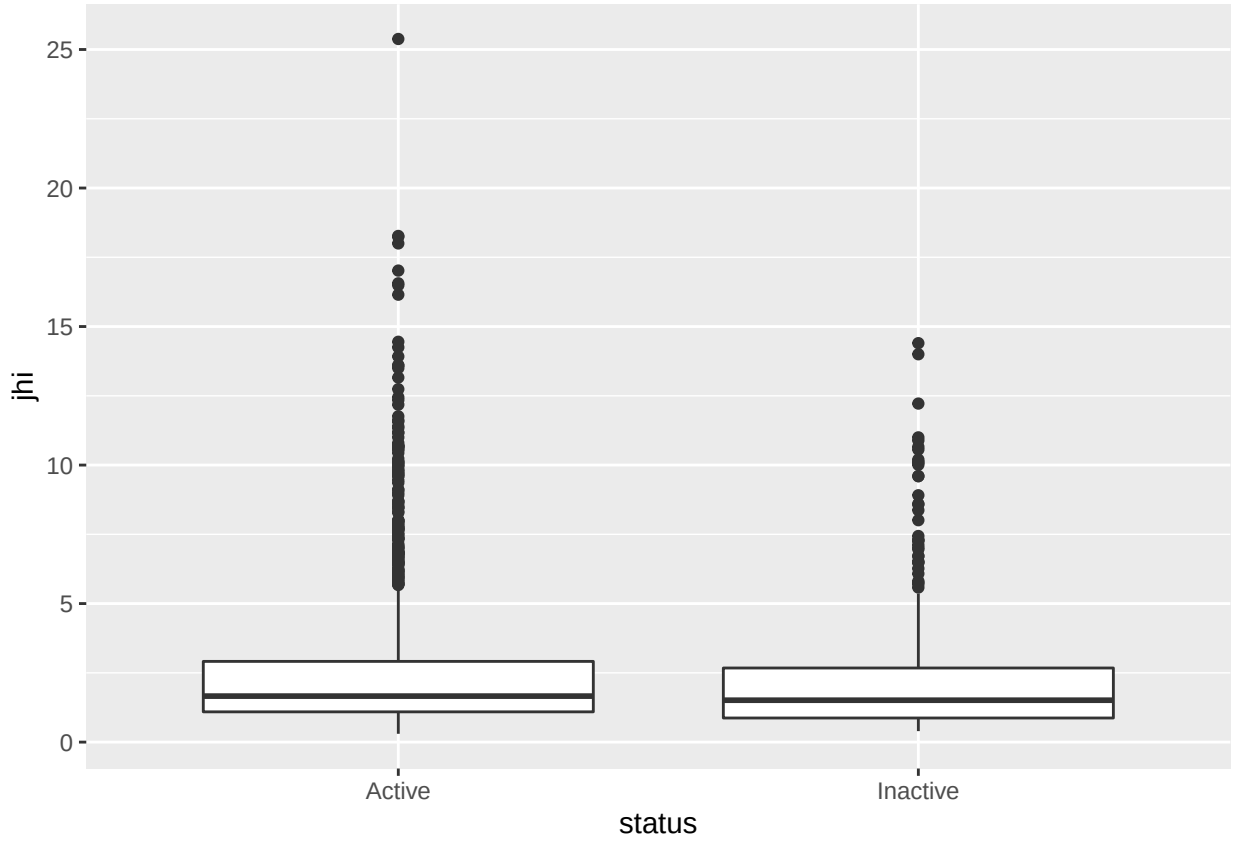


checking the distance between work place and home either effect to turnover or not.





```
## Warning: Removed 186 rows containing non-finite values (stat_boxplot).
```

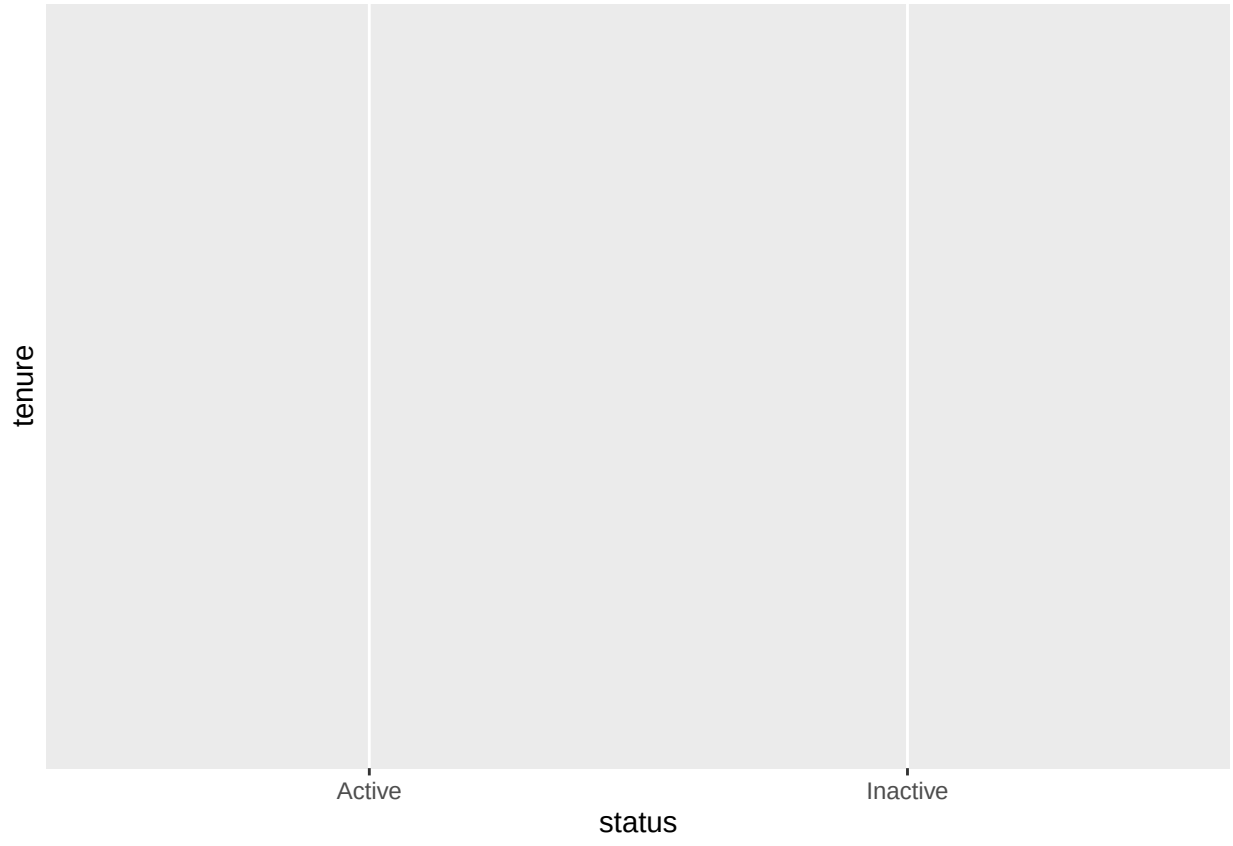


Question: Any suggestion to improve a employee turnover?

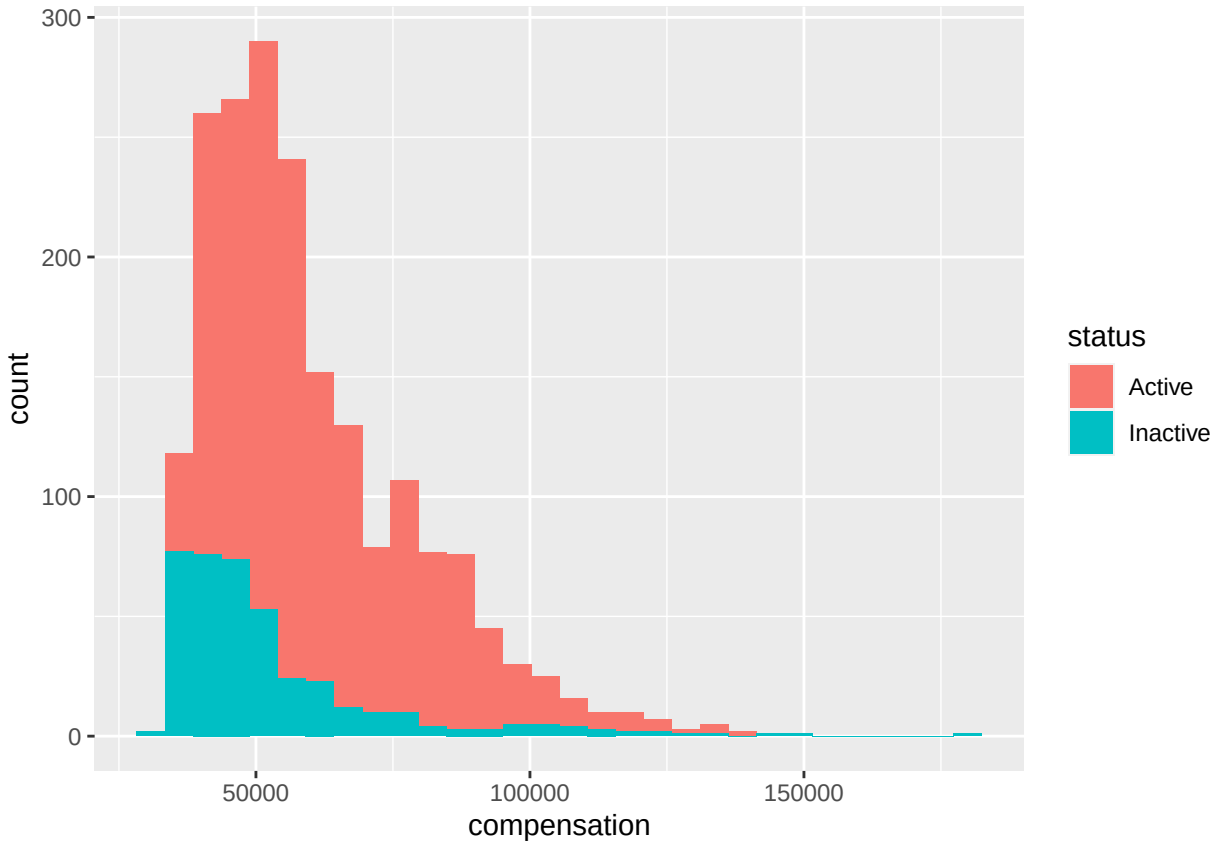
Calulate the empolyee tenure. According to below box-plot, we can see the Q1 percentile of active is almost equal the Q3 percentile of Inactive employee. In inactive employee, only 50% work like a year in the company. However, active empolyees are working for a long time than inactive employee. As a result. we can assume that the inactive employee domain percentage is new employee. According that, we have to improve the Junior employees in their first year.

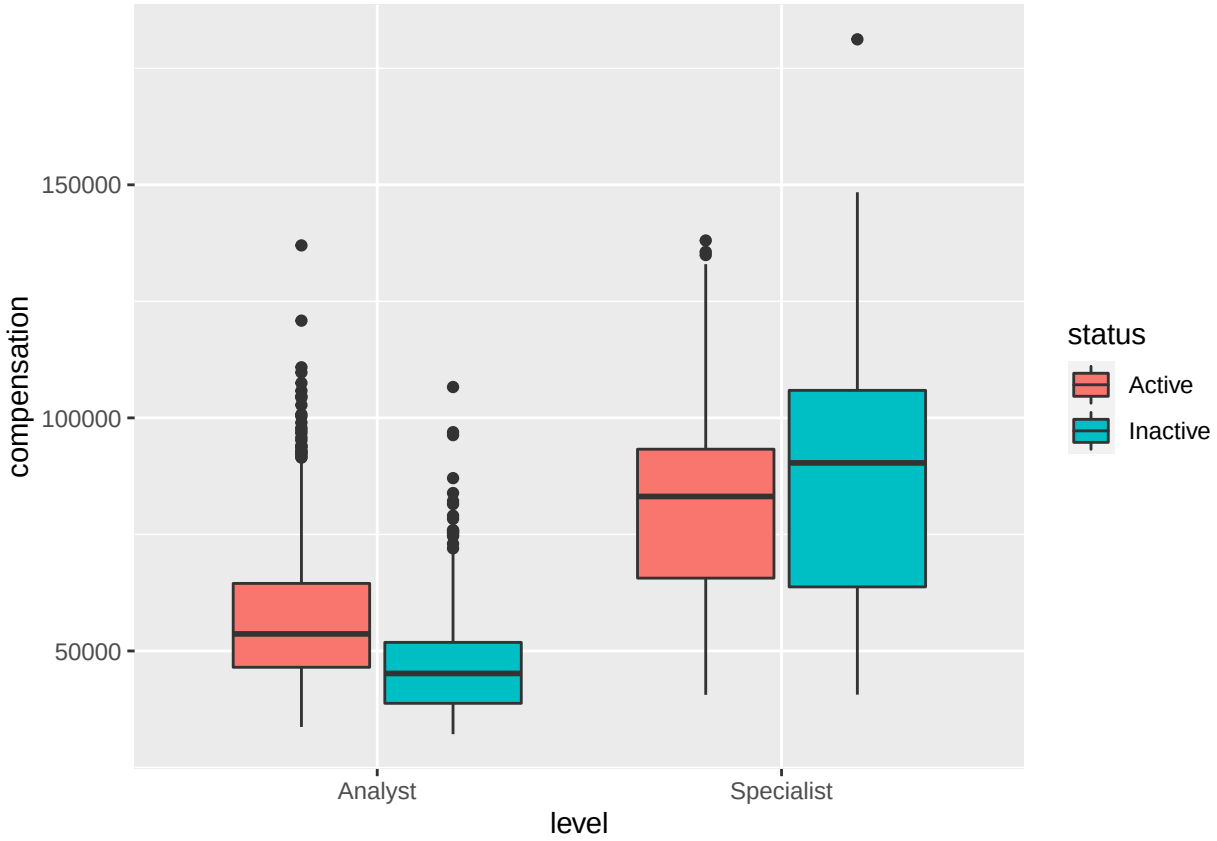
```
##
## Attaching package: 'lubridate'

## The following objects are masked from 'package:base':
##
##     date, intersect, setdiff, union

## Warning: Problem with `mutate()` input `tenure`.
## i All formats failed to parse. No formats found.
## i Input `tenure` is `ifelse(...)`.

## Warning: All formats failed to parse. No formats found.
```

```
## Warning: Problem with `mutate()` input `tenure`.
## i All formats failed to parse. No formats found.
## i Input `tenure` is `ifelse(...)`.

## Warning: All formats failed to parse. No formats found.

## Warning: Problem with `mutate()` input `tenure`.
## i All formats failed to parse. No formats found.
## i Input `tenure` is `ifelse(...)`.

## Warning: All formats failed to parse. No formats found.

## Warning: Problem with `mutate()` input `tenure`.
## i All formats failed to parse. No formats found.
## i Input `tenure` is `ifelse(...)`.

## Warning: All formats failed to parse. No formats found.

## Warning: Removed 1954 rows containing non-finite values (stat_boxplot).
```



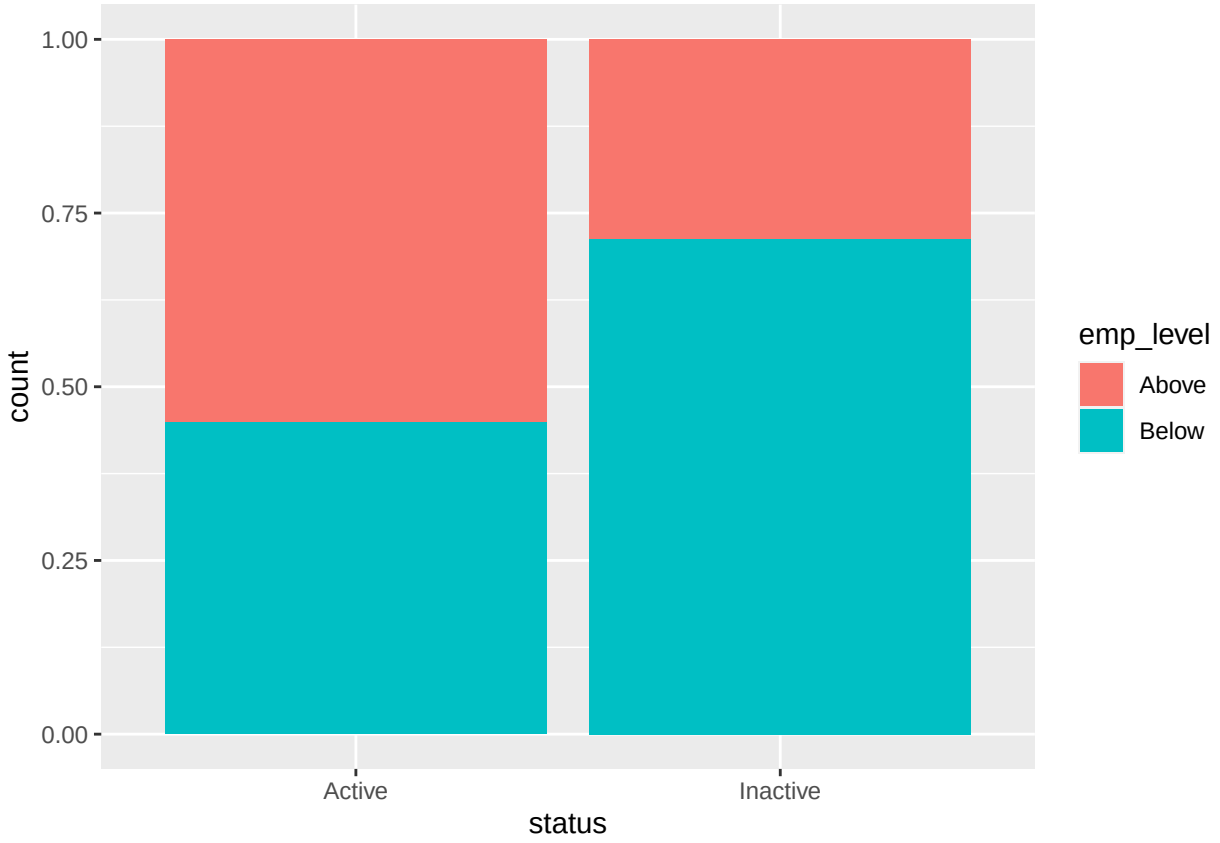

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```


300
200
100
0
count
50000
100000
150000
compensation
status
Active
Inactive




```
## # A tibble: 2 x 2
## # Groups:   level [2]
##   level      median_compensation
##   <chr>                    <dbl>
## 1 Analyst                  51840
## 2 Specialist               83496
```

```
emp_final<- emp_ratio%>%
  mutate(emp_level = ifelse( compa_ratio > 1, "Above", "Below")) # add compa level , if compa_ration ge
emp_final%>%
  ggplot(aes(status, fill = emp_level))+geom_bar(position = "fill") #compare compa level between active
```



Unstanding information value : measure of predictive power of independent variable to accurately predict the dependent variable

Information value = sima(% of non-events-% of events))* log(% of non-events/% of events)

information value : less than 0.15 meaning predictive power is poor, if 0.15 < IV < 0.4 id moderate, else greater than 0.4 meaning strong.

```
## [1] "Variable emp_id was removed because it is a non-numeric variable with >1000 categories"
## [1] "Variable department was removed because it has only 1 unique value"
## [1] "Variable cutoff_date was removed because it has only 1 unique value"
## [1] "Variable tenure was removed because it has only 1 unique level"
```

```
##                      Variable           IV
## 12               percent_hike 1.144784e+00
## 17            total_dependents 1.088645e+00
## 21             no_leaves_taken 9.404533e-01
## 27          mgr_effectiveness 6.830020e-01
## 11               compensation 6.074885e-01
## 34                compa_ratio 4.768892e-01
## 24             date_of_joining 4.330804e-01
## 6                       rating 3.869373e-01
```

```
## 23         monthly_overtime_hrs 3.786644e-01
## 8                mgr_reportees 3.620543e-01
## 2                     location 2.963023e-01
## 35                   emp_level 2.940446e-01
## 26                      mgr_id 2.820235e-01
## 5                      emp_age 2.275477e-01
## 16           distance_from_home 1.470549e-01
## 30            work_satisfaction 1.378953e-01
## 22             total_experience 1.345781e-01
## 19                    education 1.253865e-01
## 20       promotion_last_2_years 9.979915e-02
## 9                      mgr_age 9.816205e-02
## 29           perf_satisfaction 7.099511e-02
## 13                hiring_score 6.684727e-02
## 31                    age_diff 6.634065e-02
## 32                         jhi 6.586588e-02
## 10                  mgr_tenure 5.918048e-02
## 28          career_satisfaction 3.539857e-02
## 3                         level 2.726491e-02
## 33          median_compensation 2.726491e-02
## 18               marital_status 2.588063e-02
## 7                    mgr_rating 2.172222e-02
## 15 no_previous_companies_worked 1.729893e-02
## 14                hiring_source 8.773529e-03
## 4                        gender 3.959968e-05
## 1                        status 0.000000e+00
## 25            last_working_date 0.000000e+00
```

logistic regression

As a summary showing the overall for the multiple linear regression, we can see the p-value to design which is significant factors. Compensation, career_satisfaction, rating, work_satification and promotion_last_2_years are not significant to this regression.

```
##
## Call:
## glm(formula = turnover ~ emp_age + percent_hike + hiring_score +
##     compensation + distance_from_home + total_dependents + total_experience +
##     monthly_overtime_hrs + career_satisfaction + perf_satisfaction +
##     work_satisfaction + location + rating + marital_status +
##     education + promotion_last_2_years, family = "binomial",
##     data = org)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.3728  -0.2997  -0.1162  -0.0249   3.3442
##
## Coefficients:
##                             Estimate Std. Error z value Pr(>|z|)
## (Intercept)               -2.184e+00  2.733e+00  -0.799 0.424291
## emp_age                   -2.738e-01  5.781e-02  -4.737 2.17e-06 ***
## percent_hike              -4.669e-01  4.891e-02  -9.545  < 2e-16 ***
## hiring_score               6.108e-02  3.086e-02   1.980 0.047760 *
```

```
## compensation              -5.404e-06  6.901e-06  -0.783 0.433585
## distance_from_home         2.013e-01  1.491e-02  13.496  < 2e-16 ***
## total_dependents           7.443e-01  7.223e-02  10.304  < 2e-16 ***
## total_experience           1.170e-01  5.737e-02   2.040 0.041316 *
## monthly_overtime_hrs       1.723e-01  2.533e-02   6.804 1.02e-11 ***
## career_satisfaction        4.193e-01  9.003e-01   0.466 0.641378
## perf_satisfaction         -2.455e+00  7.882e-01  -3.115 0.001841 **
## work_satisfaction         -2.979e-01  9.881e-01  -0.302 0.763024
## locationNew York           1.396e+00  2.764e-01   5.051 4.39e-07 ***
## locationOrlando           -8.959e-01  2.417e-01  -3.706 0.000210 ***
## ratingAcceptable          -1.681e-01  2.391e-01  -0.703 0.481963
## ratingBelow Average       -1.531e+00  4.318e-01  -3.546 0.000391 ***
## ratingExcellent           -3.753e-01  6.233e-01  -0.602 0.547124
## ratingUnacceptable        -2.634e+00  7.617e-01  -3.458 0.000544 ***
## marital_statusSingle       1.684e+00  3.366e-01   5.002 5.67e-07 ***
## educationMasters           1.751e+00  3.843e-01   4.558 5.18e-06 ***
## promotion_last_2_yearsYes -4.267e-02  2.748e-01  -0.155 0.876607
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1972.64  on 1953  degrees of freedom
## Residual deviance:  800.29  on 1933  degrees of freedom
## AIC: 842.29
##
## Number of Fisher Scoring iterations: 7


##
## Call:
## glm(formula = turnover ~ emp_age + percent_hike + hiring_score +
##     distance_from_home + total_dependents + total_experience +
##     monthly_overtime_hrs + perf_satisfaction + work_satisfaction +
##     location + marital_status + education, family = "binomial",
##     data = org)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.4686  -0.3170  -0.1320  -0.0318   3.0960
##
## Coefficients:
##                      Estimate Std. Error z value Pr(>|z|)
## (Intercept)          -2.20596    2.63514  -0.837  0.40252
## emp_age              -0.28830    0.05351  -5.388 7.14e-08 ***
## percent_hike         -0.32873    0.03007 -10.930  < 2e-16 ***
## hiring_score          0.04918    0.03016   1.630  0.10300
## distance_from_home    0.19869    0.01459  13.617  < 2e-16 ***
## total_dependents      0.73606    0.07098  10.370  < 2e-16 ***
## total_experience      0.10135    0.05626   1.801  0.07163 .
## monthly_overtime_hrs  0.16180    0.02452   6.600 4.12e-11 ***
## perf_satisfaction    -2.37872    0.58944  -4.036 5.45e-05 ***
## work_satisfaction    -0.34245    0.95838  -0.357  0.72085
## locationNew York      1.40716    0.26298   5.351 8.76e-08 ***
## locationOrlando      -0.85796    0.23221  -3.695  0.00022 ***
```

```
## marital_statusSingle  1.62941    0.33074   4.927 8.37e-07 ***
## educationMasters      1.78722    0.37628   4.750 2.04e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1972.64  on 1953  degrees of freedom
## Residual deviance:  822.69  on 1940  degrees of freedom
## AIC: 850.69
##
## Number of Fisher Scoring iterations: 7
```

In the catogrical variables, 1 is yes, and 0 is baseline. For example, Marital_status summary only show Single, the married as baseline. we could state that when employee is single, it is associate 1.63 increase turnover with employee is single.

turnover= -2.21x-0.288emp_age-0.329percent_hike+0.049hiring_sore+0.199distance_from_homw+0.736to 2.379per_satisfaction-0.342work_satisfaction+1.41New york or -0.858 Orlando+1.63Sigale+1.787 Masters.

```
##          (Intercept)              emp_age         percent_hike
##          -2.20595505          -0.28829617          -0.32872906
##         hiring_score   distance_from_home     total_dependents
##           0.04917649           0.19869275           0.73606333
##     total_experience monthly_overtime_hrs    perf_satisfaction
##           0.10135401           0.16180301          -2.37872259
##    work_satisfaction     locationNew York      locationOrlando
##          -0.34245419           1.40716345          -0.85796263
## marital_statusSingle     educationMasters
##           1.62941476           1.78722459
```

after we assume a realtionship to employees' self, we can add the manager information in the model.

```
##
## Call:
## glm(formula = turnover ~ emp_age + percent_hike + hiring_score +
##     distance_from_home + total_dependents + total_experience +
##     monthly_overtime_hrs + perf_satisfaction + work_satisfaction +
##     location + marital_status + education + mgr_rating + mgr_reportees +
##     mgr_age + mgr_tenure, family = "binomial", data = org)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -2.77652  -0.30516  -0.12154  -0.02717   3.04969
##
## Coefficients:
##                         Estimate Std. Error z value Pr(>|z|)
## (Intercept)             -4.35211    2.85720  -1.523  0.12771
## emp_age                 -0.29176    0.05636  -5.177 2.26e-07 ***
```

```
## percent_hike            -0.33470    0.03131 -10.690  < 2e-16 ***
## hiring_score             0.04341    0.03148   1.379  0.16790
## distance_from_home       0.19817    0.01516  13.074  < 2e-16 ***
## total_dependents         0.72643    0.07467   9.728  < 2e-16 ***
## total_experience         0.09618    0.05881   1.635  0.10199
## monthly_overtime_hrs     0.16441    0.02591   6.344 2.23e-10 ***
## perf_satisfaction       -2.54492    0.64147  -3.967 7.27e-05 ***
## work_satisfaction       -0.52446    1.07491  -0.488  0.62562
## locationNew York         1.51056    0.28523   5.296 1.18e-07 ***
## locationOrlando         -0.64191    0.24213  -2.651  0.00802 **
## marital_statusSingle     1.65180    0.35593   4.641 3.47e-06 ***
## educationMasters         1.75622    0.39708   4.423 9.74e-06 ***
## mgr_ratingAcceptable     0.21178    0.23285   0.910  0.36308
## mgr_ratingBelow Average -0.36585    0.40655  -0.900  0.36817
## mgr_ratingExcellent      0.39011    0.33076   1.179  0.23823
## mgr_ratingUnacceptable   0.35634    0.85852   0.415  0.67809
## mgr_reportees            0.12535    0.01918   6.536 6.32e-11 ***
## mgr_age                  0.03771    0.02362   1.597  0.11034
## mgr_tenure              -0.03958    0.02673  -1.481  0.13862
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1972.64  on 1953  degrees of freedom
## Residual deviance:  768.39  on 1933  degrees of freedom
## AIC: 810.39
##
## Number of Fisher Scoring iterations: 7


##
## Call:
## glm(formula = turnover ~ emp_age + percent_hike + distance_from_home +
##     total_dependents + monthly_overtime_hrs + perf_satisfaction +
##     location + marital_status + education + mgr_reportees, family = "binomial",
##     data = org)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -2.73695  -0.30182  -0.11947  -0.02871   3.12217
##
## Coefficients:
##                      Estimate Std. Error z value Pr(>|z|)
## (Intercept)          -2.10268    1.27825  -1.645  0.09998 .
## emp_age              -0.21065    0.03360  -6.270 3.61e-10 ***
## percent_hike         -0.33307    0.03097 -10.754  < 2e-16 ***
## distance_from_home    0.19458    0.01475  13.192  < 2e-16 ***
## total_dependents      0.72930    0.07310   9.976  < 2e-16 ***
## monthly_overtime_hrs  0.15724    0.02513   6.256 3.95e-10 ***
## perf_satisfaction    -2.41213    0.51329  -4.699 2.61e-06 ***
## locationNew York      1.64616    0.26924   6.114 9.71e-10 ***
## locationOrlando      -0.63414    0.23525  -2.696  0.00703 **
## marital_statusSingle  1.71471    0.34570   4.960 7.04e-07 ***
## educationMasters      1.72941    0.39196   4.412 1.02e-05 ***
```

```
## mgr_reportees         0.12310    0.01872   6.577 4.80e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1972.6  on 1953  degrees of freedom
## Residual deviance:  781.5  on 1942  degrees of freedom
## AIC: 805.5
##
## Number of Fisher Scoring iterations: 7
```

70% Training data and 30% Test data

split the data 70% into train and 30% into test. we got 1367 observations in training data

```
## [1] 1367
```

we do ramdonly select train data and test data into sample

After we calculating the propoertion in train date in status, we can see in proportion that we lost approximation 20.6% to our employee.

```
##     status    n      prop
## 1   Active 1079 0.7893197
## 2 Inactive  288 0.2106803
```

```
##     status   n      prop
## 1   Active 478 0.8143101
## 2 Inactive 109 0.1856899
```

Variance Inflation Factor

```
library(car)
```

```
## Loading required package: carData
```

```
##
## Attaching package: 'car'
```

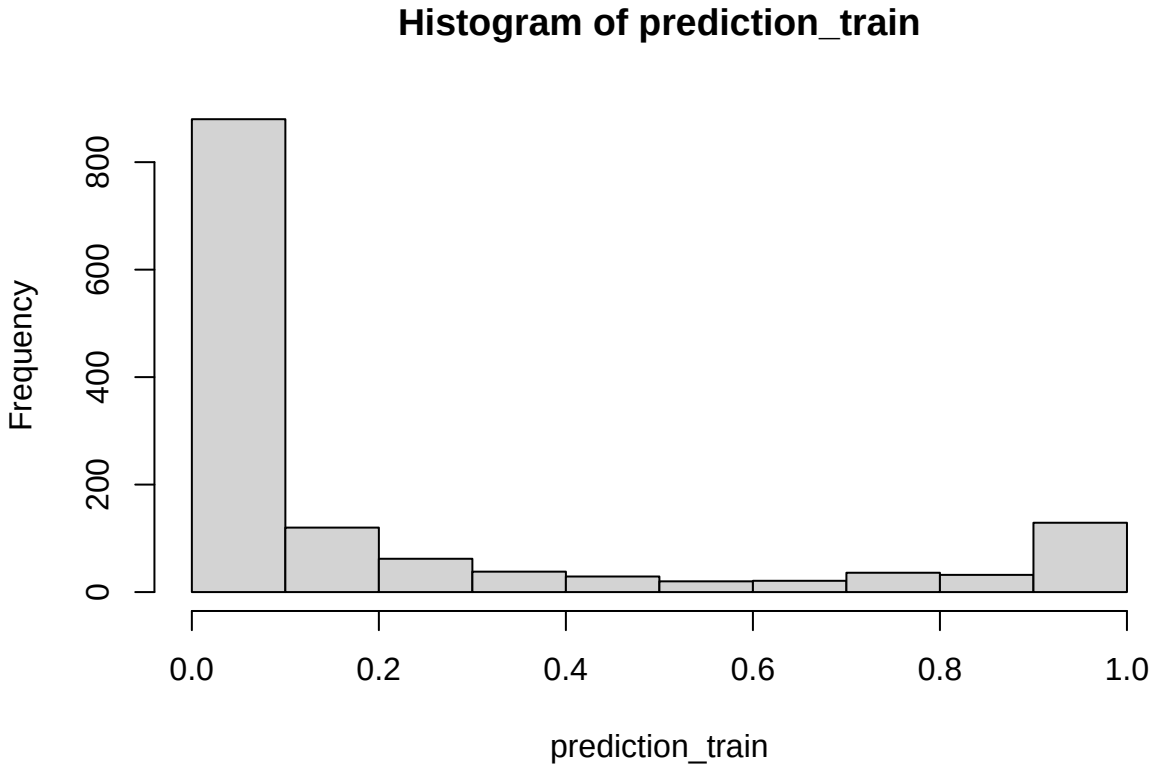
```
## The following object is masked from 'package:purrr':
##
##     some
```

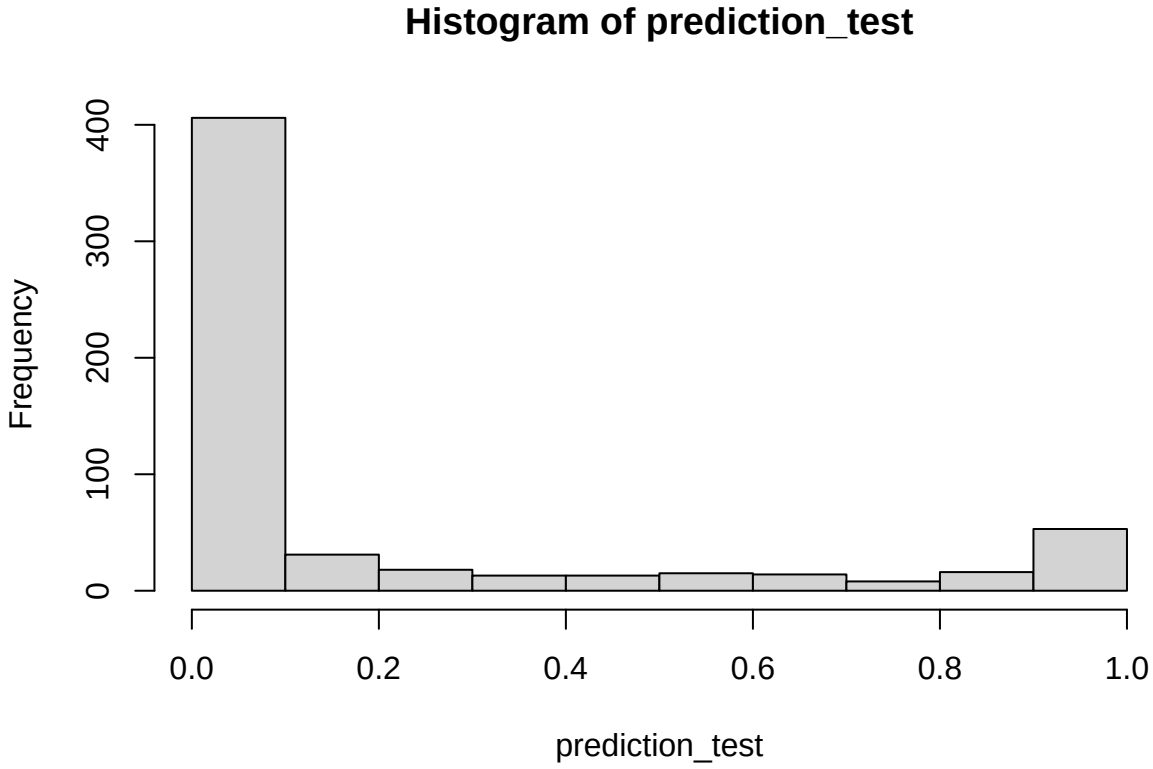
```
## The following object is masked from 'package:dplyr':
##
##     recode
```

```r
vif(logistic_mgr1) #check a multicollinearity, the vif for each variables is greater than 1 but less th
```

```
##                          GVIF Df GVIF^(1/(2*Df))
## emp_age              1.556914  1        1.247764
## percent_hike         1.091015  1        1.044517
## distance_from_home   1.080413  1        1.039429
## total_dependents     1.859150  1        1.363507
## monthly_overtime_hrs 1.026944  1        1.013382
## perf_satisfaction    1.048421  1        1.023924
## location             1.468268  2        1.100782
## marital_status       2.027701  1        1.423974
## education            1.116719  1        1.056749
## mgr_reportees        1.052921  1        1.026119
```

below attachment is prediction range of train data.





```
##
## pre_cut   0   1
##       0 461  20
##       1  17  89
```

```r
n<-sum(conf_matrix) #number of instances
nc<- nrow(conf_matrix) # number of classes
diag <- diag(conf_matrix) # number of correctly classified instances per class
rowsums <- apply(conf_matrix, 1, sum) # number of instances per class
colsums <- apply(conf_matrix, 2, sum) # number of predictions per class
p <- rowsums / n # distribution of instances over the actual classes
q <- colsums / n # distribution of instances over the predicted classes
```

```
## [1] 0.9369676
```

```
##         0         1
## 0.9644351 0.8165138
```

```
## [1] 0.9369676
```

```
##         0         1
## 0.9644351 0.8165138
```

Question: After we analysis the result at above, what is next step?

when we know the accuracy and precision to our analysis, we have to provide the feeback and improve our strategy.

create retension strategy

analysis the probability of employee turnover.

```
## Adding missing grouping variables: `level`
```

```
## # A tibble: 10 x 3
## # Groups:   level [2]
##    level      emp_id   fit
##    <chr>      <chr>  <dbl>
##  1 Analyst    E6406  0.976
##  2 Analyst    E2163  0.932
##  3 Analyst    E13342 0.906
##  4 Specialist E682   0.883
##  5 Analyst    E5524  0.881
##  6 Analyst    E3160  0.832
##  7 Specialist E3360  0.793
##  8 Specialist E8607  0.749
##  9 Specialist E6475  0.736
## 10 Specialist E3389  0.648
```

Improve data frame.

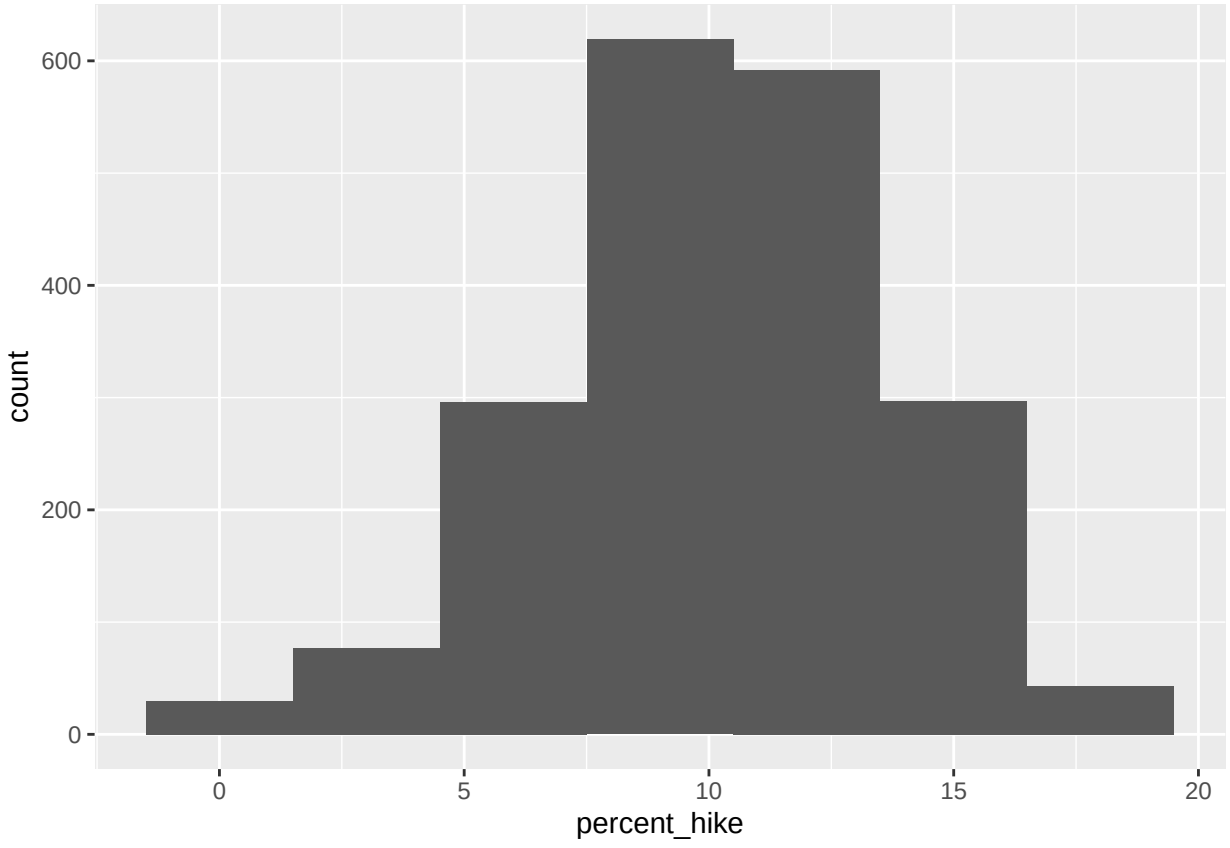
```
## # A tibble: 8 x 3
## # Groups:   risk_bucket [4]
##   level      risk_bucket     n
##   <chr>      <fct>       <int>
## 1 Analyst    no-risk      1162
## 2 Analyst    low-risk       60
## 3 Analyst    medium-risk    22
## 4 Analyst    high-risk      15
## 5 Specialist no-risk       283
## 6 Specialist low-risk        7
## 7 Specialist medium-risk     4
## 8 Specialist high-risk       4
```

Provide a plan to improve employee turnover.

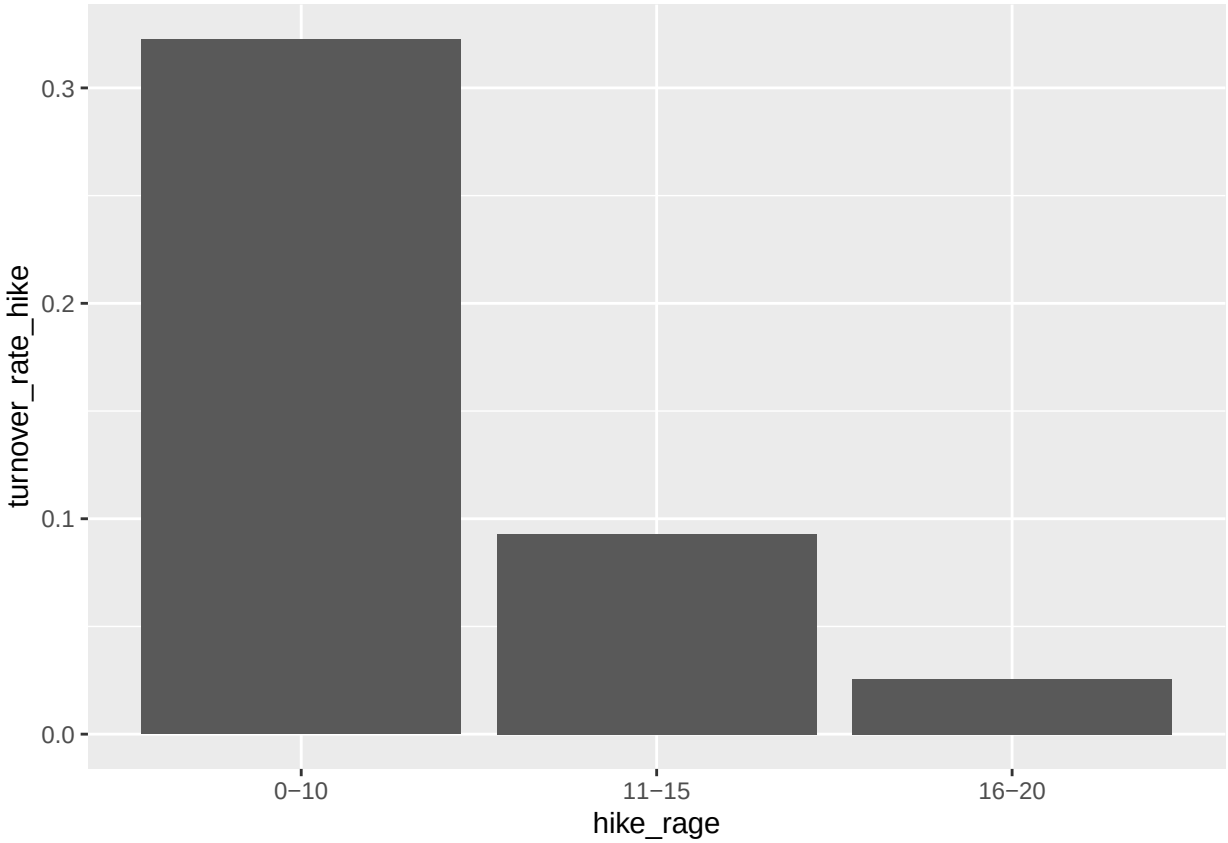
The histogram table shows normal distribution.

ROI: retun on investment

ROI = Program Benifits / Program Cost




```
## `summarise()` ungrouping output (override with `.groups` argument)
```




```
## # A tibble: 1 x 3
## # Groups:   level [1]
##   level   median_compensation     n
##   <chr>                 <dbl> <int>
## 1 Analyst               51840  1604

## Adding missing grouping variables: `level`

## # A tibble: 6 x 2
## # Groups:   level [1]
##   level   compensation
##   <chr>          <int>
## 1 Analyst        32148
## 2 Analyst        32304
## 3 Analyst        33696
## 4 Analyst        33768
## 5 Analyst        33768
## 6 Analyst        33900
```

At the above anaylsis, we identified that the turnover rate of employees in 0 to 10 percent salary hike range is higher compared 11 to 15 and 16 to 20 percent salary hike range. Assumming all employees who received a salary hike between 0 and 10% were instead offered a hike, there is a very good chance we would have been retain most of the employees. Provide a strategy to increase 5% to all employees' salay.

```
## [1] 54432
```

```
## [1] 40000
```

ROI plan

The strategy still make returns exceed 73% costs.

```
## The return on investment is 73%!
```