Fundamentals of Business Analytics

HR ANALYTICS

PREDICTING EMPLOYEE TURNOVER

Presented by: Group 3
Sarthak Singh (23PGBAN022)
Alankrit sharma (23PGBAN005)
Tanmay Malhotra (23PGBAN028)
Vansh Gandhi (23PGBAN026)
Vyom Maheswari (23PGBAN028)
Tanisha Goyal (23PGBAN004)

CONCEPTS IN HUMAN RESOURCE

EMPLOYEE TURNOVER

Employee turnover refers to the total number of employees who leave the organization over a period of time.

Turnover Cost

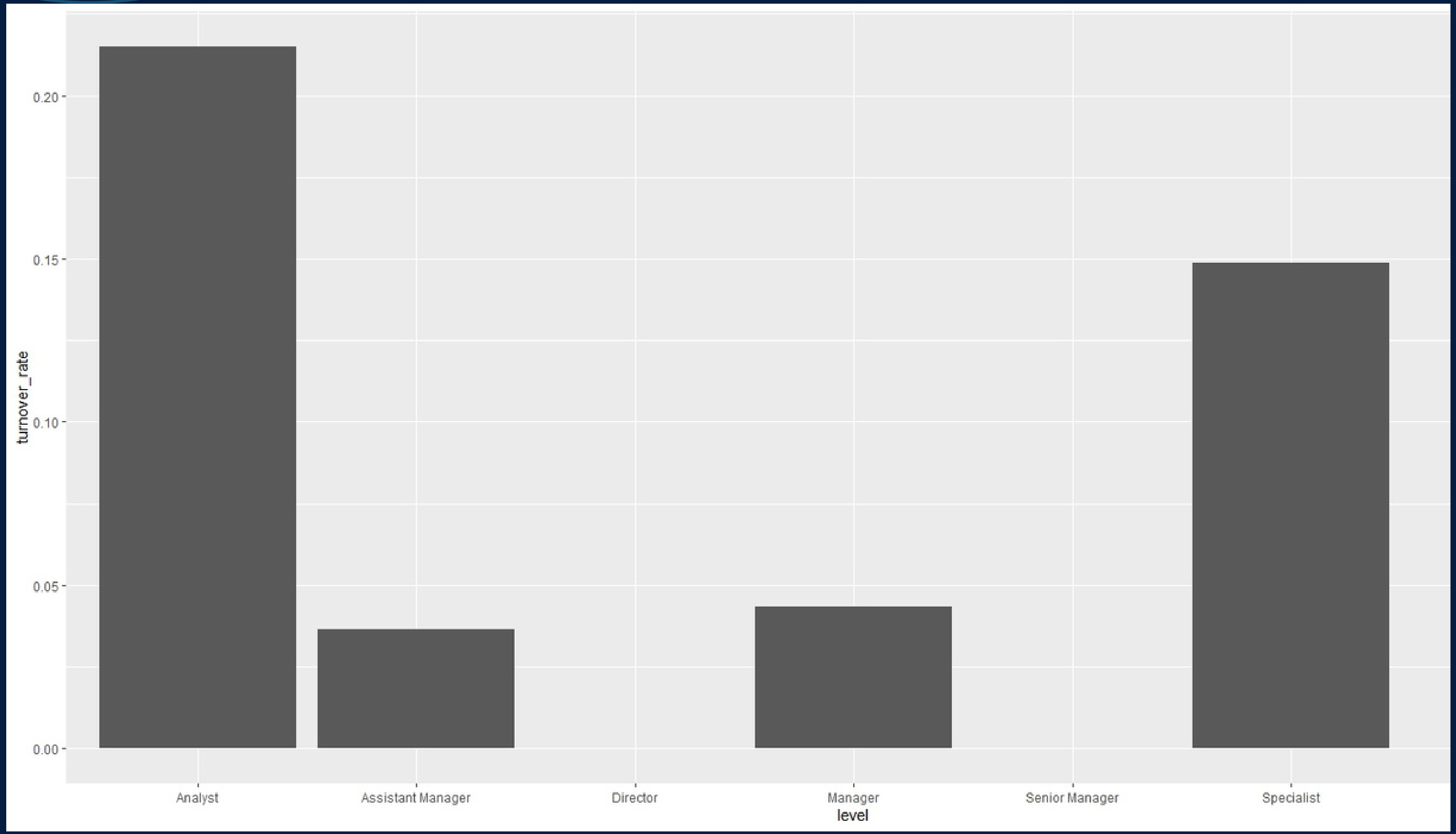
Employees are one of the most important assets to a company. Therefore turnovers are costly. In recent times there has been a spike in employee turnover. As per society of human resource management a turnover costs 100-200% of an employee monthly salary to an organization.

Affects of turnover

Turnover adversely affects the efficiency, productivity, profitability, and morale of the company.

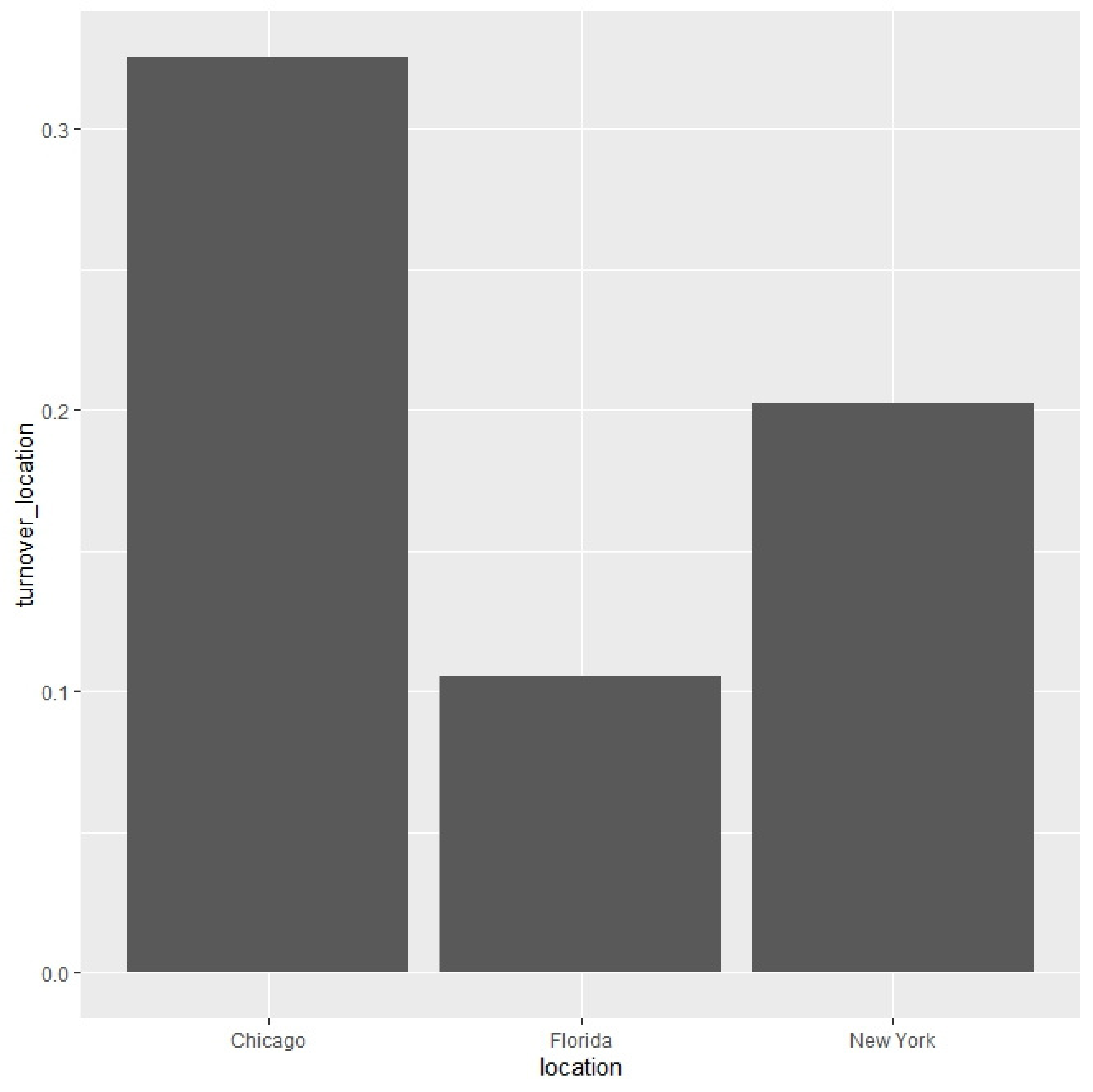
Turnover rate at different levels

- From Graph we can see that turnover rate at analyst and specialist is highest.
- Thus we will be further focusing our analysis on employees belonging to analyst and specialist level.



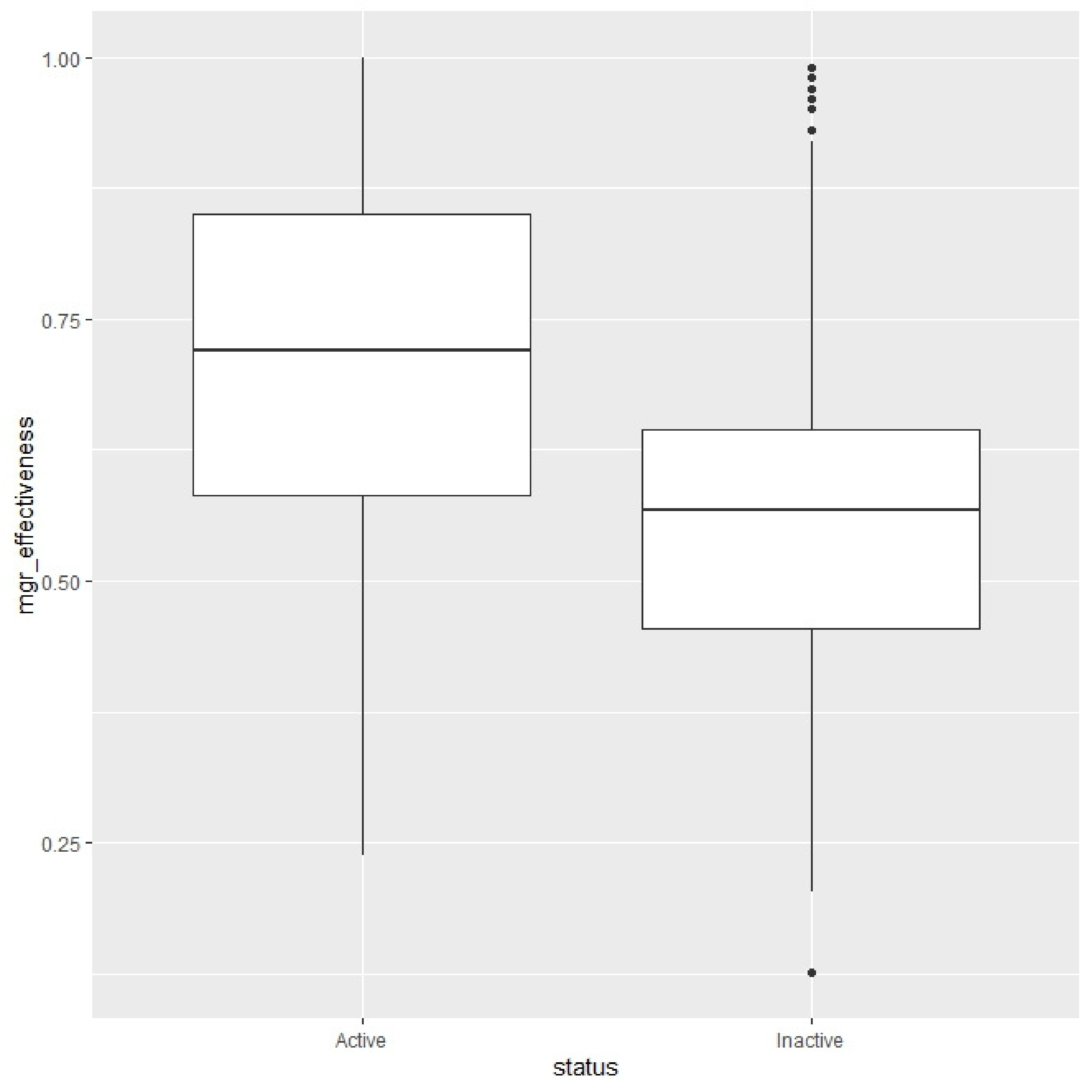
Turnover rate in different cities

- Here we can see the employees in chicago and new york regions has higher turnover rate as compared to employees in florida.
- The probable reason behind this observation can be that there are more oppurtunities available in Chicago and New york regions as compared to florida.



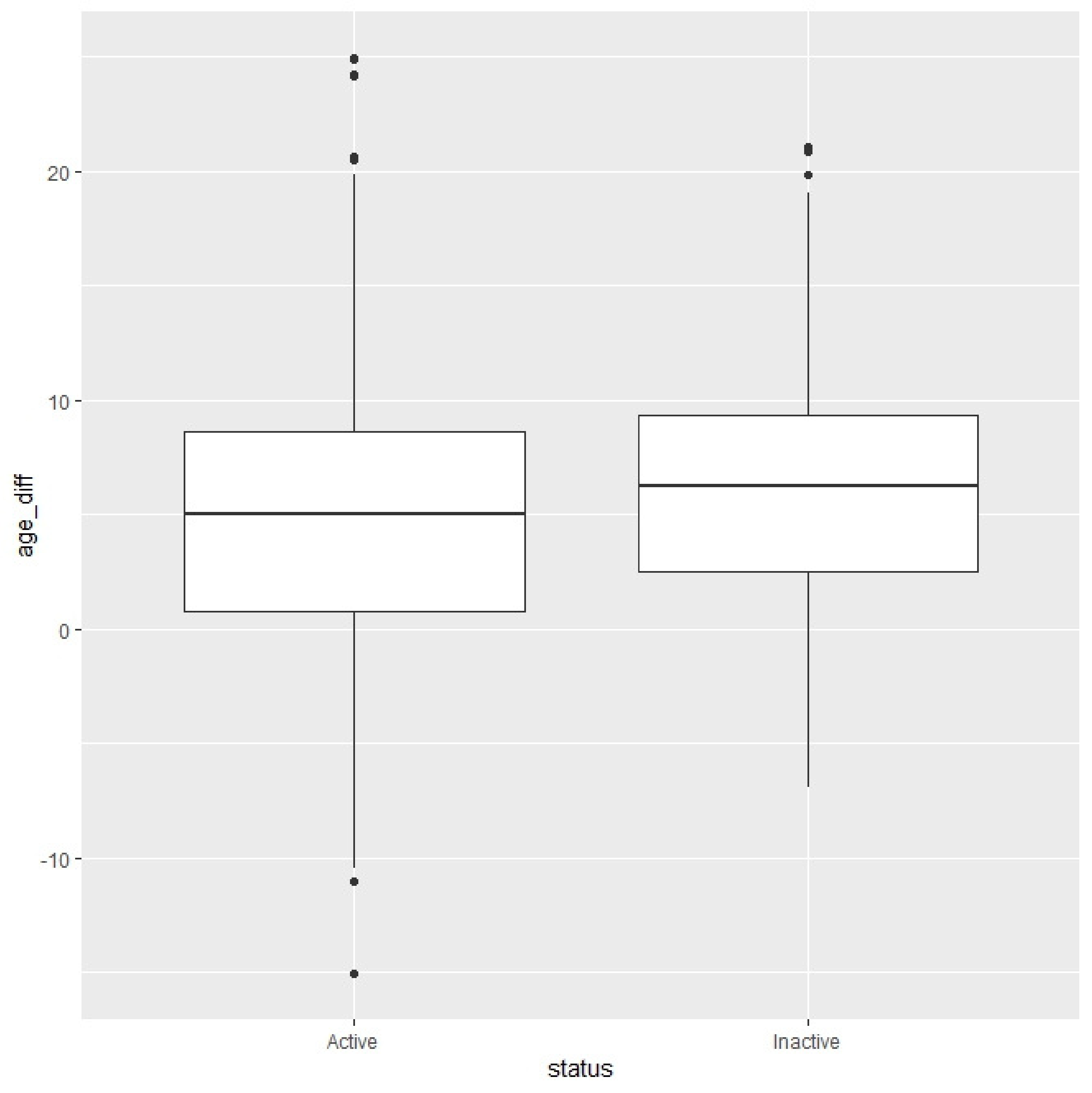
Manager's Effectiveness VS Status

- In Gallup's comprehensive 2015 study, they found a harsh truth: "75% of people quit their job to get away from their manager at some point in their career."
- Job satisfaction to a large extent depends upon the manager.
- From the boxplot we can see managers of active employees have higher score as compared to that of inactive ones.



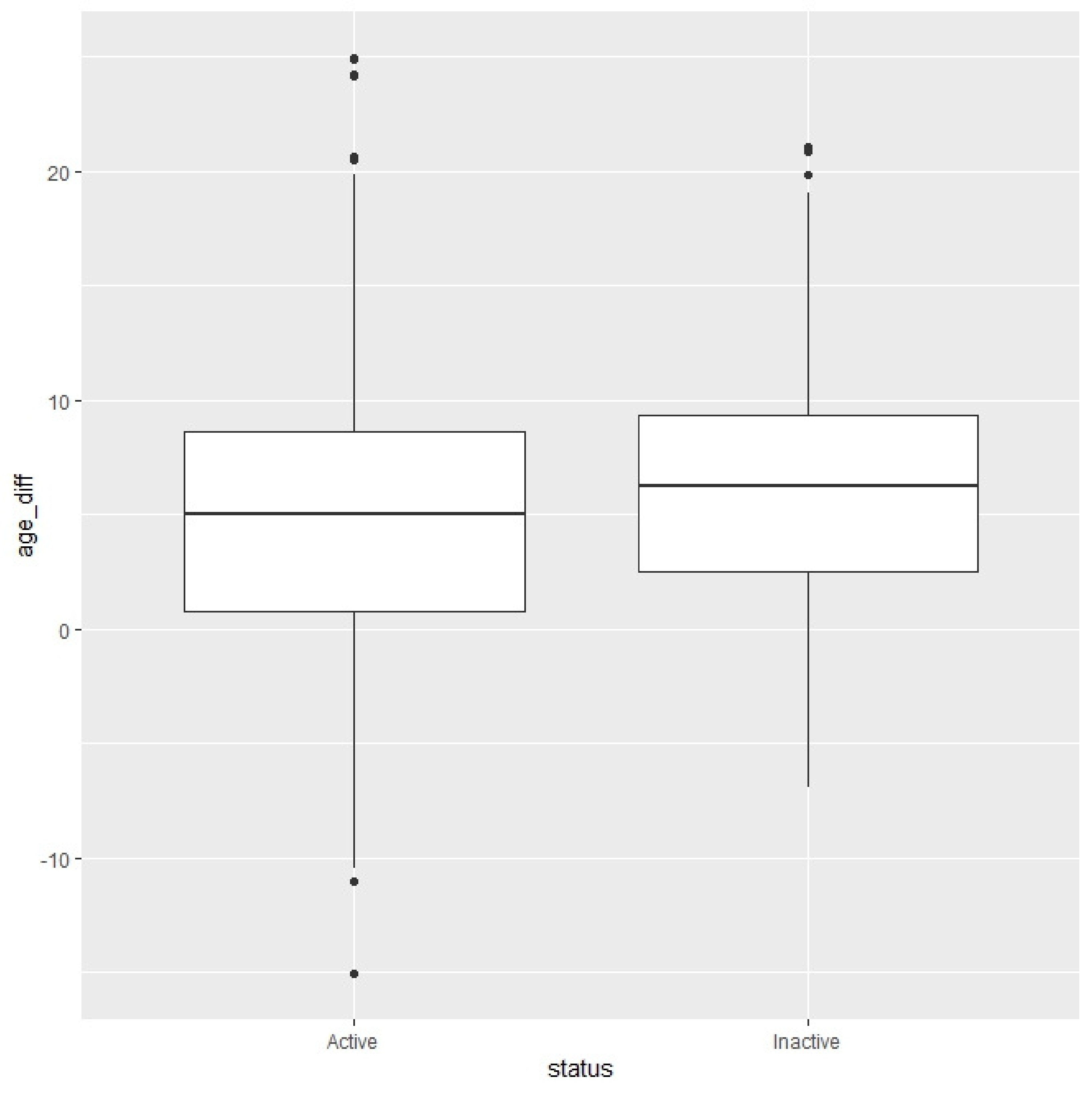
Age difference VS Status

- Different generations of the workforce have different views and opinions regarding managing and delivering work and handling pressure.
- New-generation managers need to learn to motivate and manage the pool of older workers. Employees reporting to younger managers generally report overconfidence, lack of respect, and inexperience of manager as a reason for turnover while as per younger managers, older employees wallow in their past achievements and do not accept the pace of change in the work environment.
- From Graph we can see there is not much difference between the ages of manager and employees who are active and inactive.



Age difference VS Status

- Different generations of the workforce have different views and opinions regarding managing and delivering work and handling pressure.
- New-generation managers need to learn to motivate and manage the pool of older workers. Employees reporting to younger managers generally report overconfidence, lack of respect, and inexperience of manager as a reason for turnover while as per younger managers, older employees wallow in their past achievements and do not accept the pace of change in the work environment.
- From Graph we can see there is not much difference between the ages of manager and employees who are active and inactive.



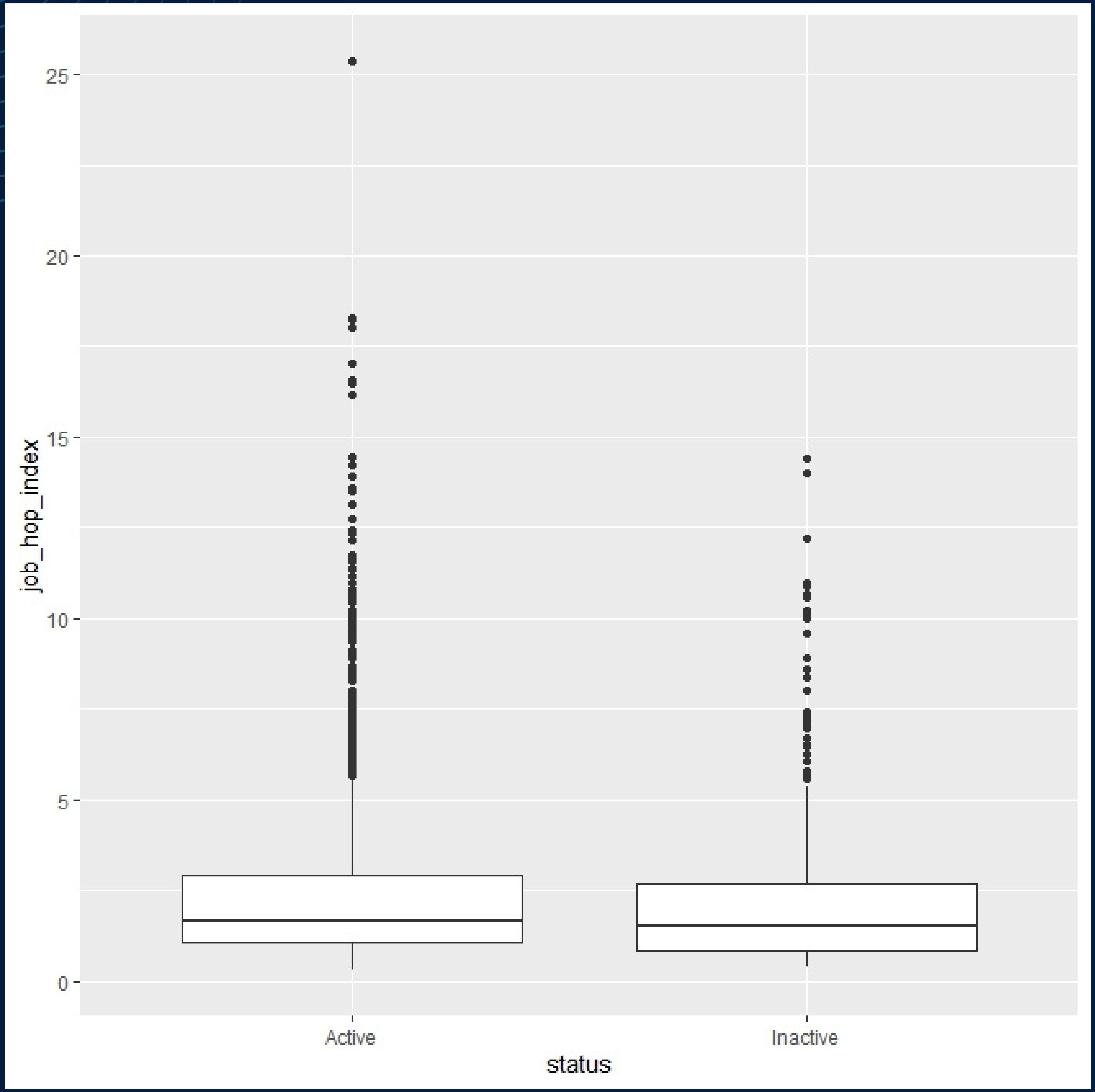
Feature Engineering

Feature engineering is a technique through which we derive a new feature through existing features.

- Job hop index refers to how frequently an employee changes companies, and there's a common understanding that those who frequently switch jobs are likely to continue doing so in the future.
- Compensation plays a significant role in determining employee turnover. If an employee at a certain level within an organization receives lower compensation compared to their peers, they are more likely to leave the company. Notably, for specialists, inactive employees tend to have higher average salaries than active ones, while for analysts, active employees generally have higher average salaries than inactive ones.
- Different generations within the workforce hold distinct perspectives on managing work, handling pressure, and delivering results. There's a need for newer-generation managers to understand how to motivate and manage older workers effectively. Employees reporting to younger managers often mention overconfidence, lack of respect, and managerial inexperience as reasons for turnover. Conversely, according to younger managers, older employees tend to dwell on past achievements and struggle to adapt to the changing work environment's pace.

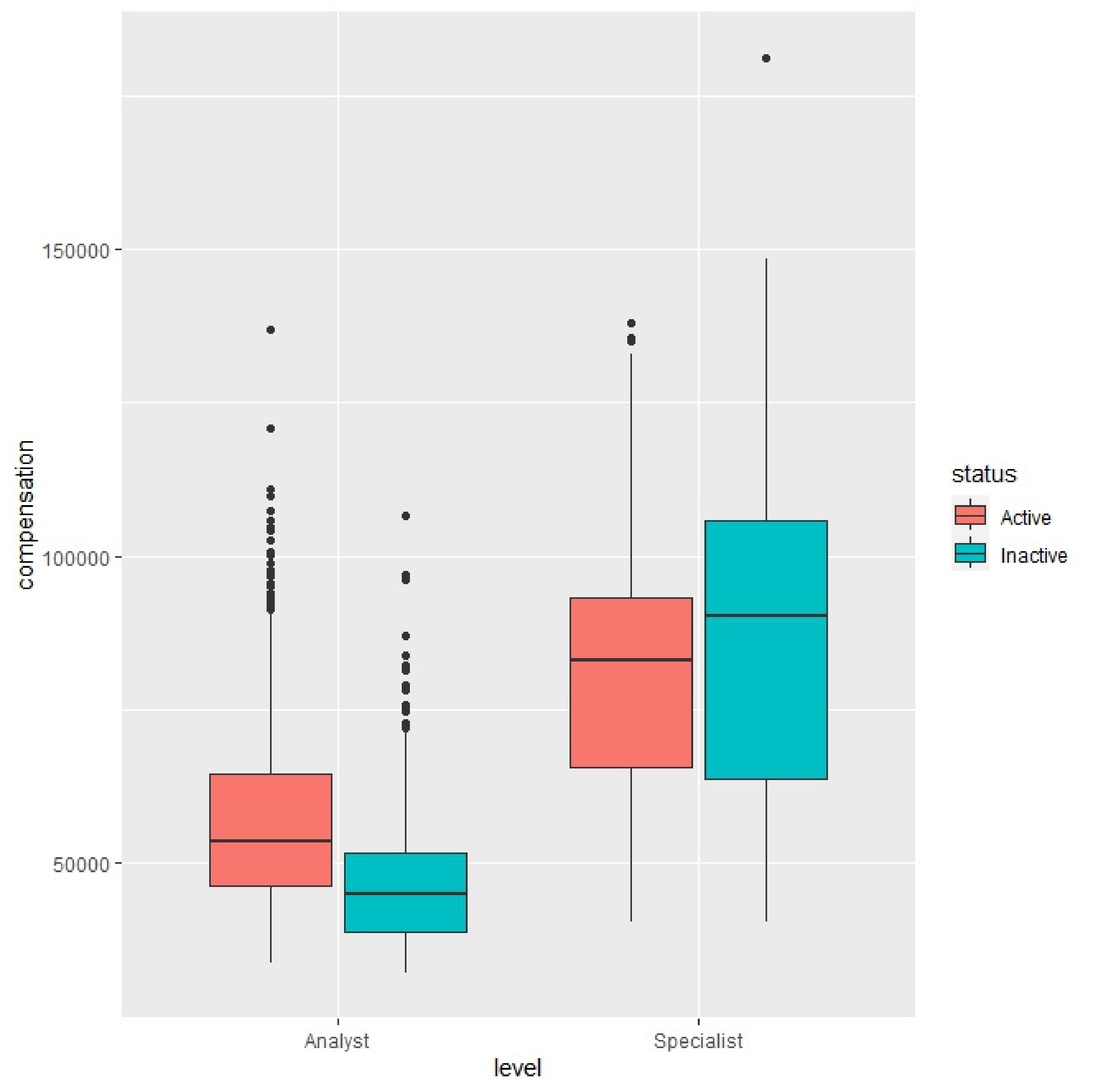
Job Hop Index VS Status

- **Job_hop_index** :- How frequently an employee leaves the company.
- From box plot we can conlude that active and unactive employees have similar job hop index.



Compensation Vs Status

- **Compensation**:- This also determines the turnover of the employees.
- If an employee at a specific level in an organization has lower compensation as compared to peers then he/she is most likely to leave organization.
- For specialists, we can observe inactive employees have a higher average salary than active ones
- For analysts, we can observe active employees have a higher average salary than inactive ones



INFORMATION VALUE

Information Value (IV) is a metric widely used in predictive modeling and especially in the field of credit scoring and risk analytics. It measures the strength of the relationship between a predictor variable (an independent variable) and the target variable (the dependent variable or the outcome of interest).

Criteria to classify predictive power

1.) IV > 0.5: Strong predictive power
2.) 0.3 < IV ≤ 0.5: Moderate predictive power
3.) 0.1 < IV ≤ 0.3: Weak predictive power
4.) Very weak predictive power

Feature Importance through Information value

	Variable	IV	
	----------------------------	--------------------------	
	percent_hike	1.144784e+00	
	total_dependents	1.088645e+00	
	no_leaves_taken	9.404533e-01	
	tenure	9.332570e-01	
	mgr_effectiveness	6.830020e-01	
	compensation	6.074885e-01	
	campa_ratio	4.768892e-01	
	rating	3.869373e-01	
	monthly_overtime_hrs	3.786644e-01	
	mgr_reportees	3.620543e-01	
	location	2.963023e-01	
	compa_level	2.940446e-01	
	mgr_id	2.820235e-01	
	emp_age	2.275477e-01	
	distance_from_home	1.470549e-01	
	work_satisfaction	1.378953e-01	
	total_experience	1.345781e-01	
	education	1.253865e-01	
	promotion_last_2_years	9.979915e-02	
	mgr_age	9.816205e-02	
	perf_satisfaction	7.099511e-02	
	hiring_score	6.684727e-02	
	age_diff	6.634065e-02	
	job_hop_index	6.586588e-02	
	mgr_tenure	5.918048e-02	
	career_satisfaction	3.539857e-02	
	level	2.726491e-02	
	median_compensation	2.726491e-02	
	marital_status	2.588063e-02	
	mgr_rating	2.172222e-02	
	no_previous_companies_worked	1.729893e-02	
	hiring_source	8.773529e-03	
	gender	3.959968e-05	
	status	0.000000e+00	

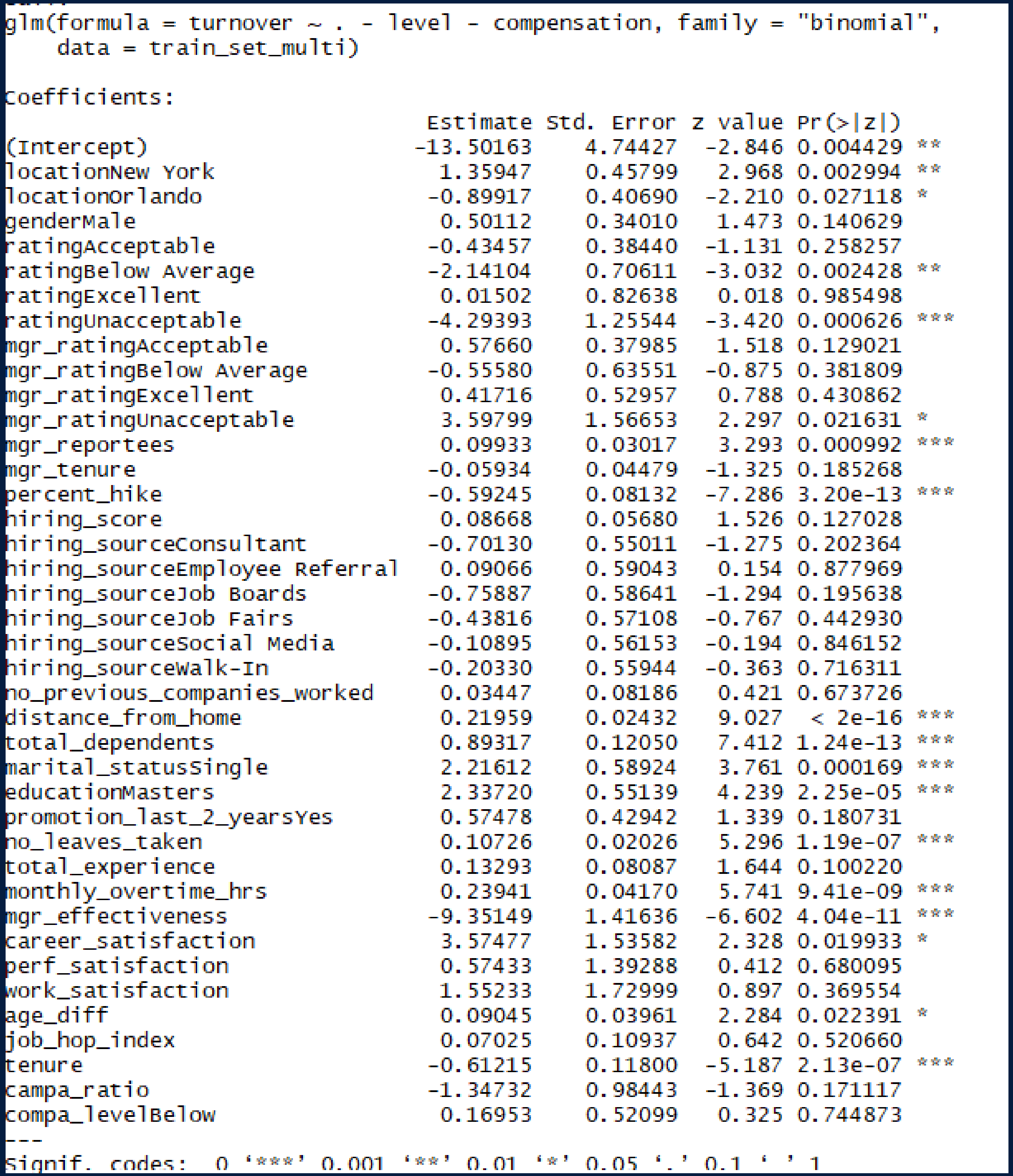
MODEL SELECTION

We've framed our analysis as a classification problem aimed at predicting employee turnover, which involves a categorical target variable with binary values (0 and 1). To address this, we've opted for the logistic regression model as it's well-suited for handling binary classification tasks by estimating the likelihood of an employee either staying (0) or leaving (1) based on various predictor variables.

Multicollinearity

Due to the presence of 34 variables in our dataset, there's a likelihood of collinearity among several of these variables. To address this concern, we applied the variance inflation factor (VIF) to detect and mitigate highly correlated features. Our analysis using VIF revealed a strong correlation between the variables "compensation" and "level." As a result, steps were taken to address this high collinearity between these specific features.

Logistic Regression Model Summary



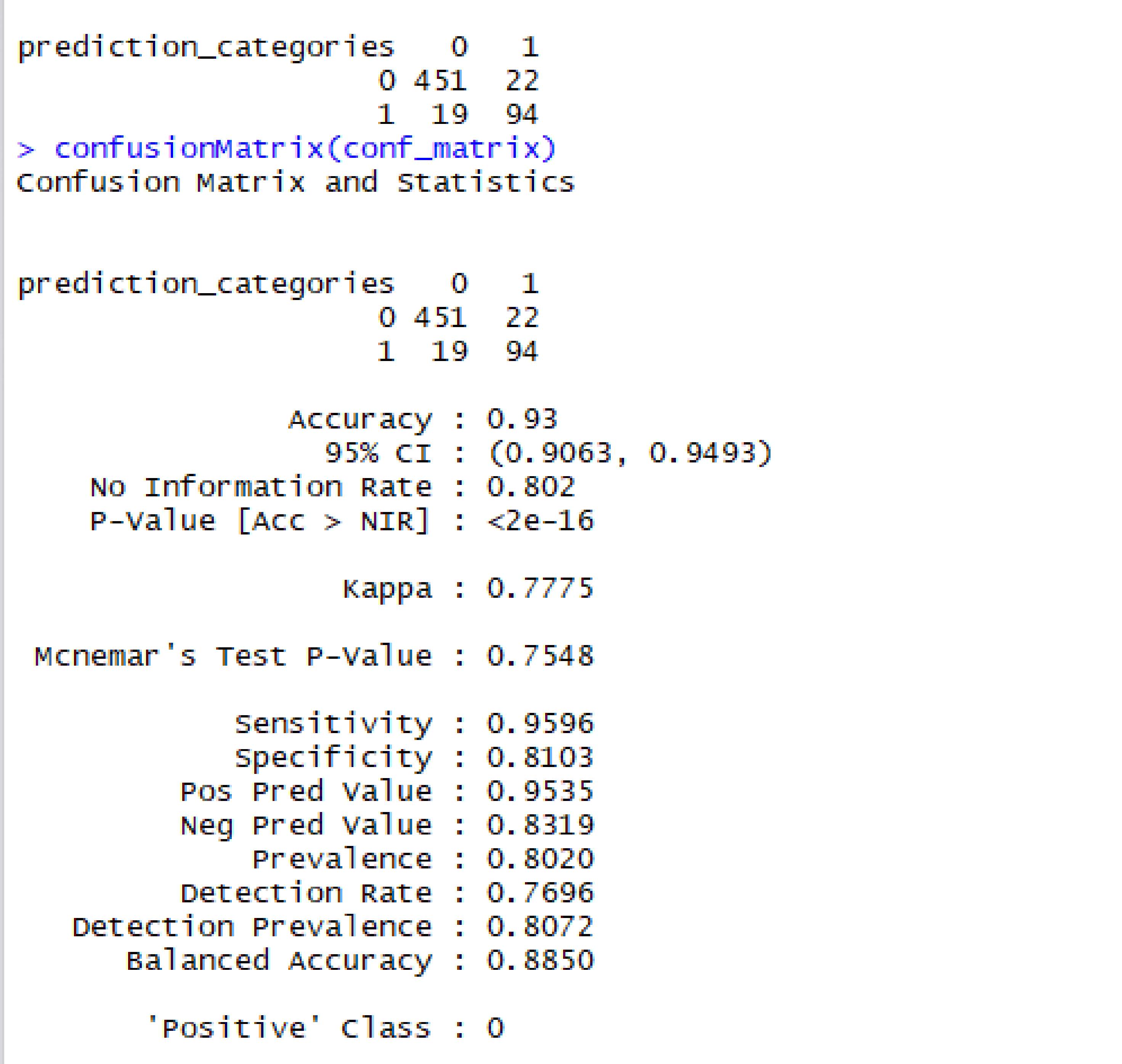
```
glm(formula = turnover ~ . - level - compensation, family = "binomial",
    data = train_set_multi)

Coefficients:
                              Estimate Std. Error z value Pr(>|z|)    
(Intercept)                  -13.50163    4.74427  -2.846 0.004429 ** 
locationNew York               1.35947    0.45799   2.968 0.002994 ** 
locationOrlando               -0.89917    0.40690  -2.210 0.027118 *  
genderMale                     0.50112    0.34010   1.473 0.140629    
ratingAcceptable              -0.43457    0.38440  -1.131 0.258257    
ratingBelow Average           -2.14104    0.70611  -3.032 0.002428 ** 
ratingExcellent                0.01502    0.82638   0.018 0.985498    
ratingUnacceptable            -4.29393    1.25544  -3.420 0.000626 ***
mgr_ratingAcceptable           0.57660    0.37985   1.518 0.129021    
mgr_ratingBelow Average       -0.55580    0.63551  -0.875 0.381809    
mgr_ratingExcellent            0.41716    0.52957   0.788 0.430862    
mgr_ratingUnacceptable         3.59799    1.56653   2.297 0.021631 *  
mgr_reportees                  0.09933    0.03017   3.293 0.000992 ***
mgr_tenure                    -0.05934    0.04479  -1.325 0.185268    
percent_hike                  -0.59245    0.08132  -7.286 3.20e-13 ***
hiring_score                   0.08668    0.05680   1.526 0.127028    
hiring_sourceConsultant       -0.70130    0.55011  -1.275 0.202364    
hiring_sourceEmployee Referral  0.09066    0.59043   0.154 0.877969    
hiring_sourceJob Boards       -0.75887    0.58641  -1.294 0.195638    
hiring_sourceJob Fairs        -0.43816    0.57108  -0.767 0.442930    
hiring_sourceSocial Media     -0.10895    0.56153  -0.194 0.846152    
hiring_sourceWalk-In          -0.20330    0.55944  -0.363 0.716311    
no_previous_companies_worked   0.03447    0.08186   0.421 0.673726    
distance_from_home             0.21959    0.02432   9.027  < 2e-16 ***
total_dependents               0.89317    0.12050   7.412 1.24e-13 ***
marital_statusSingle           2.21612    0.58924   3.761 0.000169 ***
educationMasters               2.33720    0.55139   4.239 2.25e-05 ***
promotion_last_2_yearsYes      0.57478    0.42942   1.339 0.180731    
no_leaves_taken                0.10726    0.02026   5.296 1.19e-07 ***
total_experience               0.13293    0.08087   1.644 0.100220    
monthly_overtime_hrs           0.23941    0.04170   5.741 9.41e-09 ***
mgr_effectiveness             -9.35149    1.41636  -6.602 4.04e-11 ***
career_satisfaction            3.57477    1.53582   2.328 0.019933 *  
perf_satisfaction              0.57433    1.39288   0.412 0.680095    
work_satisfaction              1.55233    1.72999   0.897 0.369554    
age_diff                       0.09045    0.03961   2.284 0.022391 *  
job_hop_index                  0.07025    0.10937   0.642 0.520660    
tenure                        -0.61215    0.11800  -5.187 2.13e-07 ***
campa_ratio                   -1.34732    0.98443  -1.369 0.171117    
compa_levelBelow               0.16953    0.52099   0.325 0.744873    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

For training our model we split our data into testing and training data. For training data we took 70 percent of data and for testing we took 30 percent of data. The trained logistic regression model gives probabilities as output. We took 0.5 probability as threshold value

The logistic regression analysis suggests that the variables location New York, locationOrlando, ratingBelow Average, ratingUnacceptable, mgr_ratingUnacceptable, mgr_reportees, percent_hike, distance_from_home, total_dependents, marital_statusSingle, educationMasters, no_leaves_taken, monthly_overtime_hrs, tenure, age_diff, career_satisfaction, mgr_effectiveness, and monthly_overtime_hrs exhibit a statistical significance level exceeding 99%.

Performance on test data



```
prediction_categories   0   1
                    0 451  22
                    1  19  94
> confusionMatrix(conf_matrix)
Confusion Matrix and Statistics


prediction_categories   0   1
                    0 451  22
                    1  19  94

               Accuracy : 0.93
                 95% CI : (0.9063, 0.9493)
    No Information Rate : 0.802
    P-Value [Acc > NIR] : <2e-16

                  Kappa : 0.7775

 Mcnemar's Test P-Value : 0.7548

            Sensitivity : 0.9596
            Specificity : 0.8103
         Pos Pred Value : 0.9535
         Neg Pred Value : 0.8319
             Prevalence : 0.8020
         Detection Rate : 0.7696
   Detection Prevalence : 0.8072
      Balanced Accuracy : 0.8850

       'Positive' Class : 0
```

For training our model we split our data into testing and training data. For training data we took 70 percent of data and for testing we took 30 percent of data. The trained logistic regression model gives probabilities as output. We took 0.5 probability as threshold value

The logistic regression analysis suggests that the variables location New York, locationOrlando, ratingBelow Average, ratingUnacceptable, mgr_ratingUnacceptable, mgr_reportees, percent_hike, distance_from_home, total_dependents, marital_statusSingle, educationMasters, no_leaves_taken, monthly_overtime_hrs, tenure, age_diff, career_satisfaction, mgr_effectiveness, and monthly_overtime_hrs exhibit a statistical significance level exceeding 99%.

Performance of model

- The logistic regression model exhibits an overall accuracy of 93%, indicating that 93% of the total predictions are accurately classified. The model's sensitivity, standing at 96%, signifies its capability to correctly predict potential turnover among employees who might leave the organization. This aspect is particularly crucial given the high costs associated with employee turnover.
- However, the model's specificity at 81% implies that it is 81 times more likely to correctly identify employees who will not leave the organization. While specificity is an essential metric, it holds less significance in the context of our model's primary objective, which focuses on accurately predicting employees who are likely to depart from the company.

Strategies to reduce employee turnover

- We have segmented the data based on the active status of employees to predict the likelihood of them leaving the organization.
- For employees with a predicted turnover probability of less than 0.5, we are categorizing them as "no-risk employees."
- Employees with a turnover probability ranging between 0.5 and 0.6 will be classified as "low-risk," while those with probabilities falling between 0.6 and 0.8 will be categorized as "medium-risk."
- Lastly, employees with a turnover probability exceeding 0.8 will be labeled as "high-risk."
- Our strategies for managing these risk categories differ:
- For high-risk employees, immediate action plans will be implemented, including one-on-one conversations and informing their reporting manager promptly.
- Medium-risk employees will entail medium-term action planning, monitoring for any behavioral changes, and conducting either one-on-one discussions or open-house sessions.
- Low-risk employees will have a standard monitoring approach, with interventions or actions taken if there are notable changes in their behavior or performance.