Employee Satisfaction: How Can It Be Improved?

Hasan Kayra Mike
INF356 - Introduction to Data Analysis
Galatasaray University
Computer Engineering
kayramike@gmail.com

I. Introduction

A. Overview

This document is a data analysis report conducted on an employee satisfaction survey and it provides valuable insights into the factors influencing employee satisfaction.

B. Motivation

Employee satisfaction is an important topic for companies. The satisfaction which an employee gets from work is correlated with their productivity; therefore, low employee satisfaction will hinder the growth of the company. On another note, an unsatisfied employee, in the best case scenario, experiences some form of burn-out syndrome or silent quits; in the worst case scenario, commits suicide, which is not uncommon.

The motivation of this report is to help companies understand the employees and their needs in order to prevent any satisfaction based consequences that are detrimental to the employee or the company; moreover, to help companies create better overall workplaces.

C. Research Questions

1) Are there specific combinations of attributes that strongly correlate with high or low satisfaction?
   - Does salary level impact employee satisfaction?
   - Do higher performance evaluation scores relate to higher job satisfaction levels?
   - Is there a relationship between the number of years spent with the company and employee satisfaction?
   - How does the number of projects and average monthly hours correlate with employee satisfaction levels?
   - Are there differences in satisfaction levels between employees who have experienced work accidents and those who haven't?
   - Do employees who received promotions report higher satisfaction levels than those who didn't?
   - Is there a difference in satisfaction levels across different departments?
2) Can the satisfaction level of an employee be predicted?

II. Method

A. Dataset

1) Story/Overview: The Employee Satisfaction Survey dataset was obtained from an annual employee satisfaction survey conducted by the company's HR department. Employees were asked to voluntarily participate to provide their job satisfaction levels. The dataset also contains a comprehensive collection of information regarding employees within the company; information such as last year's performance evaluations, project involvement, average monthly work hours, tenure with the company, work accidents, promotions received in the last five years, departmental affiliations, and salary levels.

2) Attributes:

- ”id”: Identification number of an employee. Nominal
- ”satisfaction_level”: Employee's self-reported job satisfaction level. Continuous.
- ”last_evaluation”: Employee's most recent performance evaluation score. Continuous.
- ”number_project”: Number of projects the employee is currently working on. Discrete.
- ”average_monthly_hours”: Average number of hours worked per month by the employee. Continuous.
- ”time_spend_company”: Number of years the employee has spent with the company. Discrete.
- ”work_accident”: Indicates whether the employee has experienced a work accident. Binary categorical.
- ”promotion_last_5years”: Indicates whether the employee has received a promotion in the last 5 years. Binary categorical.
- ”dept”: The department or division in which the employee works. Nominal.
- ”salary”: Employee's salary level. Ordinal.

B. Method for Answering Research Questions

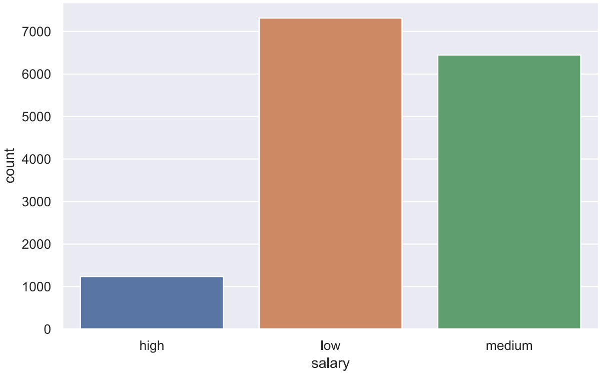
1) Handle null values of the dataset.
2) Provide a descriptive analysis on the dataset.
3) Check the distribution of ”satisfaction_level” to see if it is normal.
4) Calculate the correlation coefficients between ”satisfaction_level” and other attributes.
5) Conduct ANOVA test and then pairwise t tests for the unique values of the ”dept” attribute.
6) Do t tests to figure out the roles of ”work_accident” and ”promotions_last_5years”.
7) Perform linear regression and test the reliability of the linear regression using MSE.

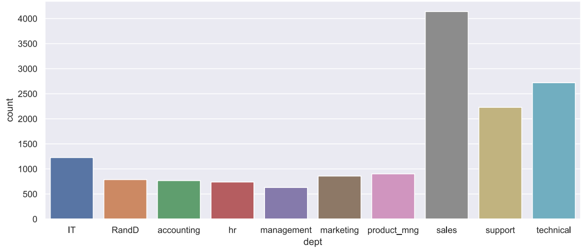
III. Results

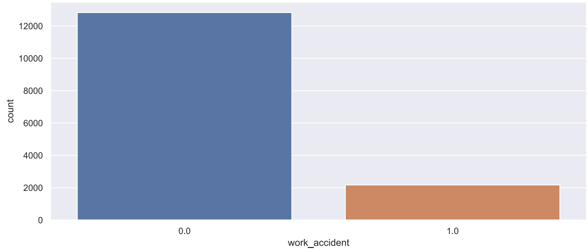
A. Descriptive Analysis

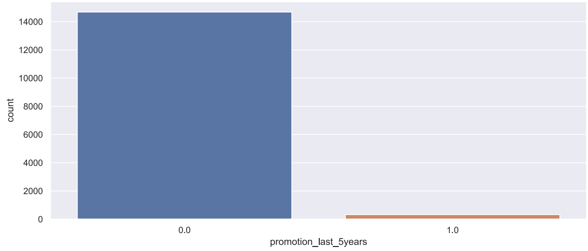
The dataset has 15787 instances and 10 attributes. Since the participation to the survey was voluntary, there are 788 null instances, meaning, 788 people didn't participate; therefore, the analysis will ignore them and end up with 14999 non-null instances.

1) Overview of the categorical attributes:



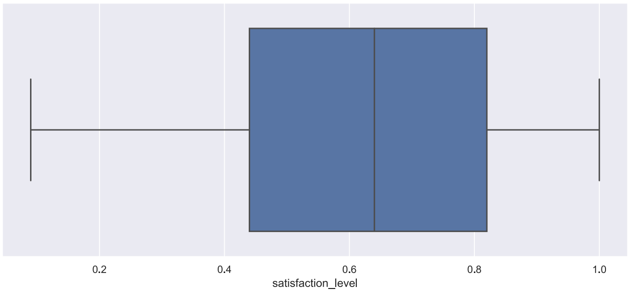


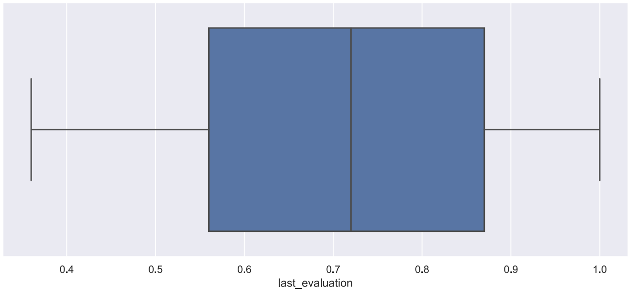


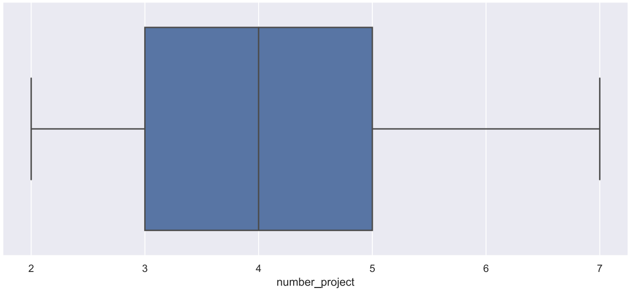


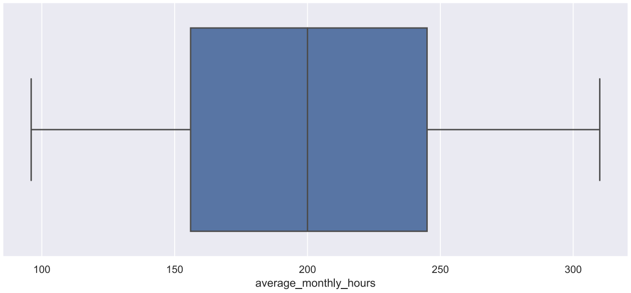
2) Descriptive overview of the quantitative attributes:

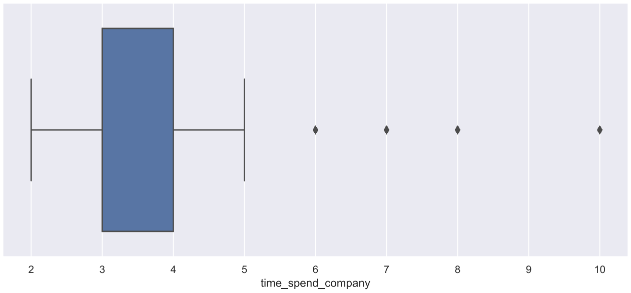
	count	mean	std	min	25%	50%	75%	max
id	14999.0	7500.000000	4329.982679	1.00	3750.50	7500.00	11249.50	14999.0
satisfaction_level	14999.0	0.612834	0.248631	0.09	0.44	0.64	0.82	1.0
last_evaluation	14999.0	0.716102	0.171169	0.36	0.56	0.72	0.87	1.0
number_project	14999.0	3.803054	1.232592	2.00	3.00	4.00	5.00	7.0
average_monthly_hours	14999.0	201.050337	49.943099	96.00	156.00	200.00	245.00	310.0
time_spend_company	14999.0	3.498233	1.460136	2.00	3.00	3.00	4.00	10.0
work_accident	14999.0	0.144610	0.351719	0.00	0.00	0.00	0.00	1.0
promotion_last_5years	14999.0	0.021268	0.144281	0.00	0.00	0.00	0.00	1.0











B. Analysis for Answering Research Questions

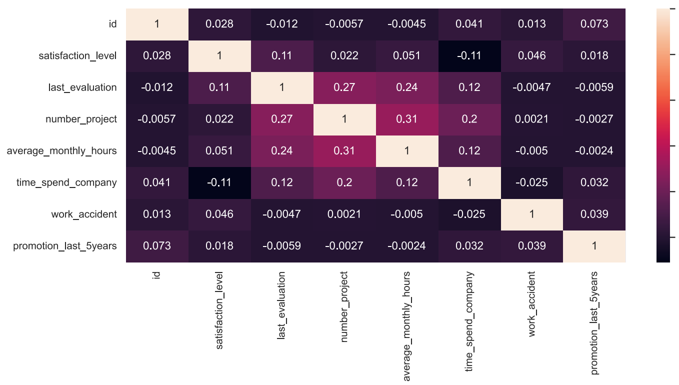
Start with checking the normality assumption, using Shapiro-Wilk test, to figure out which method of correlation coefficient calculation is needed.

H_0 : ”satisfaction_level” = normal distribution.

H_1 : ”satisfaction_level” $\neq$ normal distribution.

$p - value = 0.000... < \alpha = 0.05 \implies H_0$ is rejected.

Since the H_0 is rejected, non-parametric correlation coefficient calculation methods are more appropriate. This analysis will use the Spearman method.



As seen on the correlation matrix above, none of the attributes have significant correlations with the ”satisfaction_level” attribute.

Now, using the Mann-Whitney U non-parametric t test, let’s see if the promotions or work accidents have any effect on the satisfaction level.

- **”work_accident”:**
  H_0 : work accidents have no effect on satisfaction.
  H_1 : work accidents affect satisfaction.
  $p - value = 5.242x10^{-12} < \alpha = 0.05 \implies H_0$ is rejected.
  $\mu_{work\_accident=1} = 0.648326$
  $\mu_{work\_accident=0} = 0.606833$
  Per the Mann-Whitney U test, those who experienced work accidents report higher satisfaction levels.
- **”promotion_last_5years”:**
  H_0 : promotions have no effect on satisfaction.
  H_1 : promotions affect satisfaction.
  $p - value = 0.006301 < \alpha = 0.05 \implies H_0$ is rejected.
  $\mu_{work\_accident=1} = 0.611895$
  $\mu_{work\_accident=0} = 0.656019$
  Per the Mann-Whitney U test, those who got promoted in the last 5 years report higher satisfaction levels.

Let’s perform non-parametric ANOVA test (Kruskal-Wallis H test) to figure out if there are differences in satisfaction levels across different departments and salary levels.

- **”salary”:**
  H_0 : different salary levels have the same satisfaction level.
  H_1 : satisfaction levels change over salary levels.
  $p - value = 1.947x10^{-7} < \alpha = 0.05 \implies H_0$ is rejected.
  Per the Kruskal-Wallis H test, different salary levels don’t have similar satisfaction levels. Now performing pairwise Mann-Whitney U tests:
  - **”low-medium”:**
    H_0 : low salary = medium salary in terms of satisfaction.
    H_1 : low salary $\neq$ medium salary in terms of satisfaction.
    $p - value = 3.707x10^{-6} < \alpha = 0.05 \implies H_0$ is rejected.
    $\mu_{low} = 0.600753$
    $\mu_{medium} = 0.621817$
    Per the Mann-Whitney U test, those who have low salaries report lower satisfaction levels than those who have medium salaries.
  - **”low-high”:**
    H_0 : low salary = high salary in terms of satisfaction.
    H_1 : low salary $\neq$ high salary in terms of satisfaction.
    $p - value = 2.623x10^{-5} < \alpha = 0.05 \implies H_0$ is rejected.
    $\mu_{low} = 0.600753$
    $\mu_{high} = 0.637470$
    Per the Mann-Whitney U test, those who have low salaries report lower satisfaction levels than those who have high salaries.
  - **”medium-high”:**
    H_0 : medium salary = high salary in terms of satisfaction.
    H_1 : medium salary $\neq$ high salary in terms of satisfaction.
    $p - value = 0.1286 > \alpha = 0.05 \implies H_0$ is failed to reject.
    $\mu_{medium} = 0.600753$
    $\mu_{high} = 0.637470$
    Per the Mann-Whitney U test, those who have medium salaries report similar satisfaction levels to those who have high salaries.
- **”dept”:**
  H_0 : departments have no effect on satisfaction.
  H_1 : departments affect satisfaction.
  $p - value = 0.03189 < \alpha = 0.05 \implies H_0$ is rejected.
  $\mu_{work\_accident=1} = 0.611895$
  $\mu_{work\_accident=0} = 0.656019$
  Per the Kruskal-Wallis H test, different departments don’t have similar satisfaction levels. With the Mann-Whitney U test applied pairwise to each two groups of departments, the conclusion is that only the accounting department differs from the rest in terms of satisfaction levels while other departments show similar satisfaction levels. By evaluating the mean satisfaction level of each department, it is evident that accounting department shows lower satisfaction levels than the other departments.
- **Linear Regression**

  1) Let’s continue with the prediction of satisfaction levels using linear regression, starting with encoding

the categorical attributes (”dept” and ”salary”) using LabelEncoder.

2) The satisfaction level will be dropped from the dataset and assigned to a new variable.
3) The satisfaction level and the rest of the data set will be each divided into two groups: train and test with 20% of the data being used for the testing.
4) While testing the algorithm, MSE and R^2 will be calculated.

Here are the results:
MSE = 0.05773700044478266
R^2 = 0.06276783301384192

IV. Conclusion

A. Summary and Future Works

While there aren’t any significant correlations between satisfaction levels and other attributes, it is concluded that employees of the accounting department, low salary employees, employees that have not experienced work accidents and employees that didn’t get a promotion in the last 5 years report lower satisfaction levels than their counterparts.

Also, using the linear regression model to predict the satisfaction levels of employees didn’t perform really well.

R-squared value of the linear regression indicates that the model explains approximately 6.28% of the variance in the dependent variable (satisfaction_level). This suggests that the model’s ability to predict satisfaction levels based on the provided features is limited.

While having a small MSE is generally desirable, an R-squared score of 0.0628 indicates that a large portion of the variance in the satisfaction levels remains unexplained by the model.

In order to perform a much more useful analytics, other metrics and features must be taken into consideration, such as amenities at the workplace.