OPSM324 - Business Applications of Analytics

Project Report - Satisfaction & Employee Attrition

Geethika Jammula, Saivee Phatak, Siddhartha Reddy Kayitha, Sunishka Sil & Thwishaa Bansal

1. INTRODUCTION

For any company, regardless of its operations and size, its employee turnover is a significant challenge for many. In our project, we have considered the definition of employee turnover to be the ‘rate at which employees leave a company and need to be replaced by new hires’. It is a costly affair, both in terms of lost productivity and expenditure relating to recruiting and onboarding new employees. It involves not only the direct costs associated with recruitment, such as advertising, interviewing, and hiring, but also indirect costs like lost productivity during the transition period, training costs for new employees, and the impact on team morale and performance. Moreover, losing valuable employees can significantly impact organisational stability and performance. High turnover rates can lead to a loss of institutional knowledge, disruption in project timelines, and decreased overall productivity. Retaining talented employees who contribute positively to the company's success is crucial for maintaining a competitive edge in the market. Furthermore, by prioritising initiatives aimed at enhancing job satisfaction, such as providing opportunities for professional growth and development, fostering a positive work-life balance, recognizing and rewarding employees' contributions, and ensuring fair and transparent communication channels, organizations can significantly improve employee retention.

Hence, understanding employee behaviour and predicting which employee is at risk of attrition can help organisations to take proactive measures to retain valuable employees who add value to the firm and maintain organisational stability. In our project we propose to use a dataset containing information about employees and whether or not they left the organisation. We aim to build a model that predicts the probability that an employee will leave the company based on the characteristics exhibited by them. Using this predictive model, companies can identify the characteristics and factors associated with attrition, and organisations can tailor their recruitment, retention, and talent management strategies to address specific areas of concern. This model also incorporates data related to employee satisfaction levels, enabling organisations to identify at-risk employees and proactively intervene to address their concerns. This allows organisations to leverage data-driven decision-making processes in managing human resources. This data-driven approach enhances the effectiveness and efficiency of HR strategies, leading to better outcomes for both employees and the organisation as a whole, giving them a competitive advantage in the market.

In essence, the project is a data-driven paradigm in human resource management, advocating for the judicious integration of analytics in decision-making processes. By leveraging the insights gleaned from data analytics, organisations can cultivate a resilient and high-performing workforce, thus gaining a competitive advantage in the dynamic landscape of the modern marketplace.

2. METHODOLOGY

2.1 About Dataset: Relevance to the problem statement

The dataset has 10 employee-specific variables and 14,999 records. Each variable can be initially assumed to be a potential factor that can explain the causes or correlated reasons of high turnover risk of employees. The variables in this dataset are:

Variable	Description
Satisfaction Level (float64 - numerical)	The employee's satisfaction rating of their work experience at the company, i.e the dependent variable
Last Evaluation (float64 - numerical)	The employee's last evaluation score given by the company
Number of Projects (int64 - numerical)	Number of projects the employee has worked on throughout their tenure
Average Monthly Hours (int64 - numerical)	The employee's average hours of work in a month
Time Spent at the Company (int64 - numerical)	The number of years the employee has worked at the company,
Work Accidents (int64 - categorical)	Whether or not the employee has had any accidents at their workplace (0=no, 1=yes)
Promotion in the Last 5 Years (int64 - categorical)	Whether or not the employee has been promoted in the last five years (0=no, 1=yes)
Sales (obj - categorical)	Which of the 10 different departments the employee works in, the most recurrent department in the dataset is 'Sales'
Salary (obj - categorical)	The salary level of the employee - split between 'low', 'medium' and 'high'

Variable	Description
Left	Whether or not the employee has left the company (0=no, 1=yes), i.e the dependent variable

2.2 Research design and approach

In our research design and approach, we aim to explore the dynamics of employee turnover and satisfaction within the organisation through a comprehensive analysis of relevant data. Our focus revolves around two dependent variables: the categorical indicator of Employee Leaving Status, which signifies turnover, and the numerical measure of Satisfaction Level, offering insights into employee morale and engagement.

To analyse our dataset, we've selected two distinct models:

1. **Analysis of Variance (ANOVA)**: ANOVA enables the comparison of variances across the means of different groups. It determines if there is any significant difference between the means of different groups., ANOVA provides valuable insights into the factors influencing turnover and satisfaction levels.

2. **Multivariate Multiple Linear Regression (MMLR):** MMLR extends the standard Multiple Linear Regression (MLR) model to accommodate multiple dependent variables on a single set of predictor variables. In our case, we'll utilize Satisfaction and Employee Leaving Status as the dependent variables, examining the statistical significance of other variables in relation to these two key metrics. This approach allows us to discern the nuanced relationships between various factors and employee outcomes.

While MMLR and ANOVA offer robust methodologies for our analysis, it's essential to acknowledge their limitations. One notable constraint is the scarcity of alternative models suitable for analyzing multiple dependent variables, restricting our experimentation and potentially limiting the accuracy of our findings. Nevertheless, by leveraging these established techniques, we can extract meaningful insights into the complex interplay of factors affecting employee turnover and satisfaction.

For our data analysis, we'll employ R Studio for coding and Tableau for visualisations, utilizing a range of data preprocessing techniques and model fitting procedures. Additionally, we'll evaluate the goodness-of-fit, significance of predictors, and overall performance of the models. To further enrich our analysis, we'll also conduct separate Linear Regression analyses on each dependent variable, facilitating comparisons and enhancing the robustness of our conclusions.

2.3 Other methods are available:

In addition to Analysis of Variance (ANOVA) and Multivariate Multiple Linear Regression (MMLR), there are several other statistical models and machine learning algorithms available for analysing employee turnover and satisfaction data. Some alternative methods include:

1. **Logistic Regression:** Particularly useful for binary outcomes such as employee leaving status, logistic regression models the probability of the occurrence of a given event based on one or more predictor variables. It can provide insights into the likelihood of employees leaving the organisation based on various factors.
2. **Decision Trees:** Decision tree algorithms partition the data into subsets based on attributes or features, recursively creating a tree-like structure to make predictions. They are interpretable and can uncover complex relationships between predictors and outcomes, making them valuable for understanding employee behaviour.
3. **Random Forest:** Random Forest is an ensemble learning method that constructs multiple decision trees and combines their predictions to improve accuracy and robustness. It is effective for handling large datasets with numerous predictors and can capture nonlinear relationships effectively.
4. **Gradient Boosting Machines (GBM):** GBM is another ensemble learning technique that builds a sequence of weak prediction models, each correcting the errors of its predecessor. It is renowned for its high predictive accuracy and can handle complex interactions between predictors.
5. **Neural Networks:** Neural networks, especially deep learning architectures, can uncover intricate patterns and relationships in large-scale datasets. They are adept at capturing nonlinearities and interactions but may require substantial computational resources and data preprocessing.
6. **Survival Analysis:** Survival analysis models time-to-event data, making it suitable for studying employee tenure and the likelihood of turnover over time. It considers censoring, where some observations may not experience the event of interest during the study period.
7. **Clustering Algorithms:** Clustering algorithms such as K-means or hierarchical clustering can group employees based on similar characteristics, allowing for the identification of distinct segments with different turnover risks or satisfaction levels.

2.4 The advantage of Using ANOVA and MMLR:

1. **Control of Type I Error**: ANOVA and MMLR control Type I error rates effectively, ensuring that any observed effects are not due to random chance. This is particularly important in HR research, where decisions based on inaccurate findings could have significant consequences for organisations.
2. **Model Complexity**: ANOVA and MMLR are relatively simple models compared to some of the alternative machine learning algorithms. This simplicity makes them easier to implement, understand, and communicate to stakeholders who may not have a background in statistics or machine learning.
3. **Sample Size Considerations**: ANOVA and MMLR can be robust even with smaller sample sizes, making them suitable for studies with limited data availability. They do not require as much data as some machine learning algorithms, which may be advantageous in HR research where data collection can be challenging.
4. **Model Complexity**: ANOVA and MMLR are relatively simple models compared to some of the alternative machine learning algorithms. This simplicity makes them easier to implement, understand, and communicate to stakeholders who may not have a background in statistics or machine learning.
5. **Sample Size Considerations**: ANOVA and MMLR can be robust even with smaller sample sizes, making them suitable for studies with limited data availability. They do not require as much data as some machine learning algorithms, which may be advantageous in HR research where data collection can be challenging.

2.5 Limitations of ANOVA and MMLR:

1. **Assumptions**: Both ANOVA and MMLR rely on certain assumptions about the data. ANOVA assumes that the variances of the groups being compared are equal, and that the data are normally distributed. Violations of these assumptions can lead to inaccurate results. Similarly, MMLR assumes linearity, independence, homoscedasticity, and normally distributed errors. Departures from these assumptions can affect the validity of the results.
2. **Limited Flexibility**: ANOVA and MMLR are parametric methods with predefined functional forms. They may not capture complex relationships between predictors and outcomes as effectively as some non-parametric or machine learning approaches. This can result in reduced predictive accuracy if the true relationship in the data is nonlinear or complex.
3. **Sensitivity to Outliers**: Both ANOVA and MMLR can be sensitive to outliers in the data. Outliers can disproportionately influence the estimated coefficients or group means,

leading to biased results. Robustness techniques or alternative models may be needed to mitigate the impact of outliers.

4. **Limited Handling of Categorical Predictors**: While ANOVA can handle categorical predictors, it may not capture nonlinear or interactive effects between categorical variables and other predictors. MMLR can accommodate multiple dependent variables but may struggle with categorical predictors with many levels or interactions between categorical and continuous predictors.
5. **Multicollinearity**: In MMLR, multicollinearity (high correlation between predictor variables) can pose a problem by inflating standard errors and making coefficients difficult to interpret. Careful consideration of variable selection and regularization techniques may be necessary to address multicollinearity issues.
6. **Limited Predictive Power**: While ANOVA and MMLR can provide insights into relationships between variables, their predictive power may be limited compared to more complex machine learning algorithms. They may not capture all relevant interactions or nonlinearities in the data, potentially leading to suboptimal predictive performance.
7. **Difficulty Handling Missing Data**: Both ANOVA and MMLR require complete data for analysis. Missing data can pose challenges, and imputation methods mayintroduce bias or uncertainty into the results.

3. DATA ANALYSIS

3.1 Descriptive statistics:

No of Records	14999
No of Variables	10
Satisfaction	
Min	0.09
Max	1
Mean	0.612834
Variance	0.061817
Std Dev	0.248631
Left/Attrition	
Left	3571
Didn't leave	11428

3.2 Visualisation:

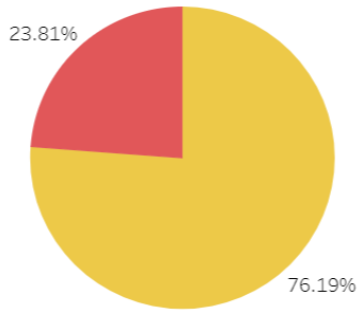


Figure 1: Percentage of People Left (red) vs. Remaining (yellow) in the company

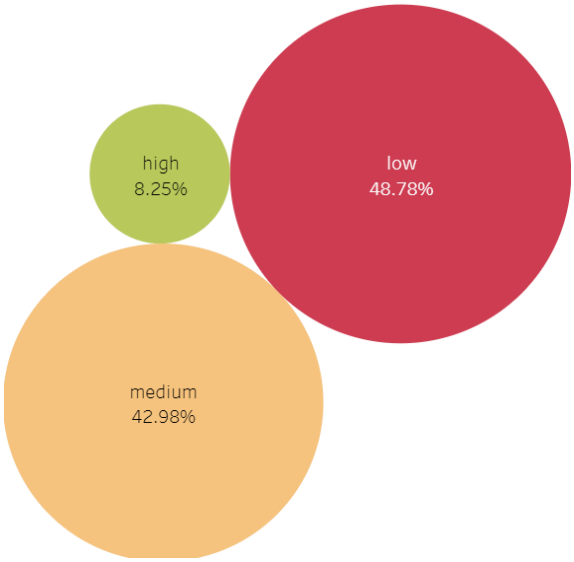


Figure 2: Percentage of Employees in High, Medium and Low Salary ranges

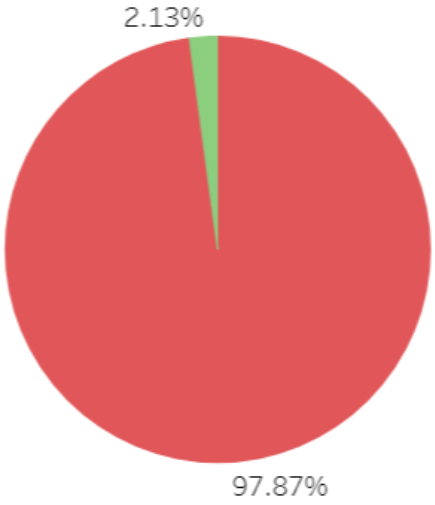


Figure 3: Percentage of Employees that got a Promotion in the Last 5 years (green) vs. no promotion (red)

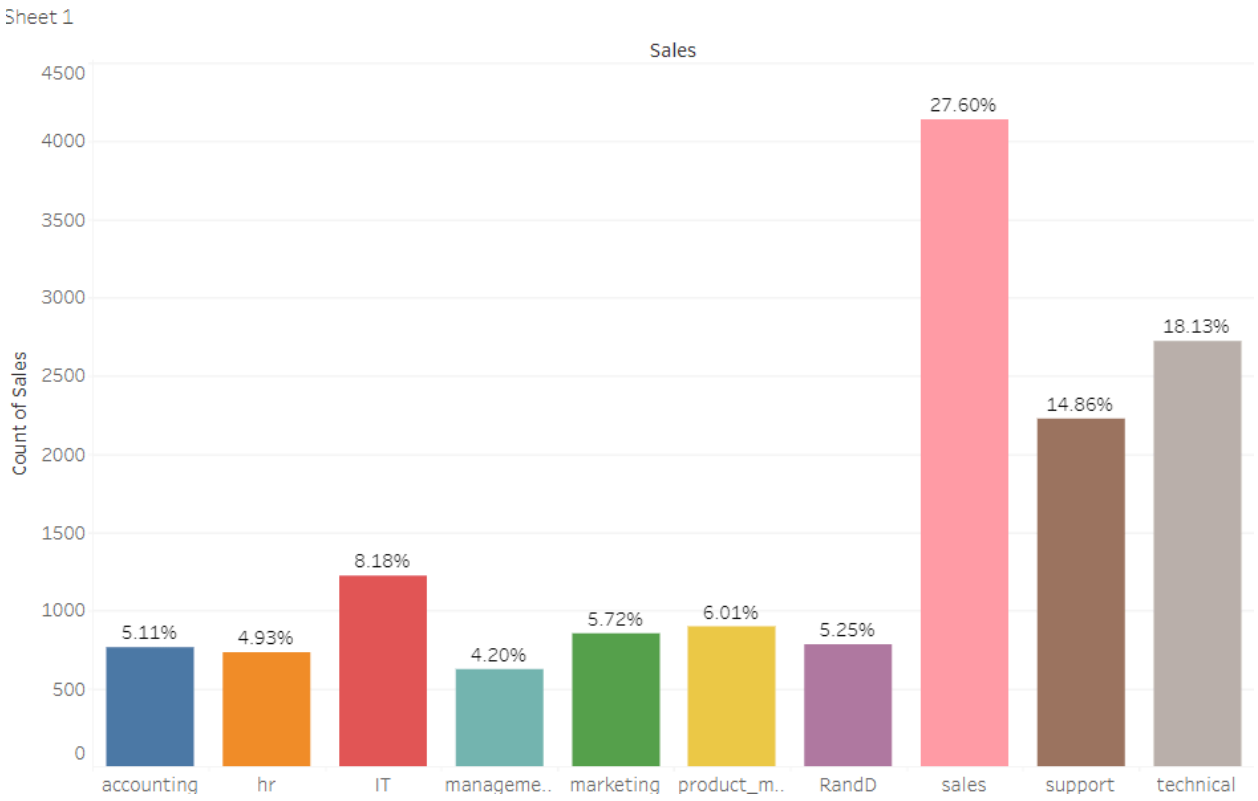


Figure 4: Percentage of attrited employees in every Department

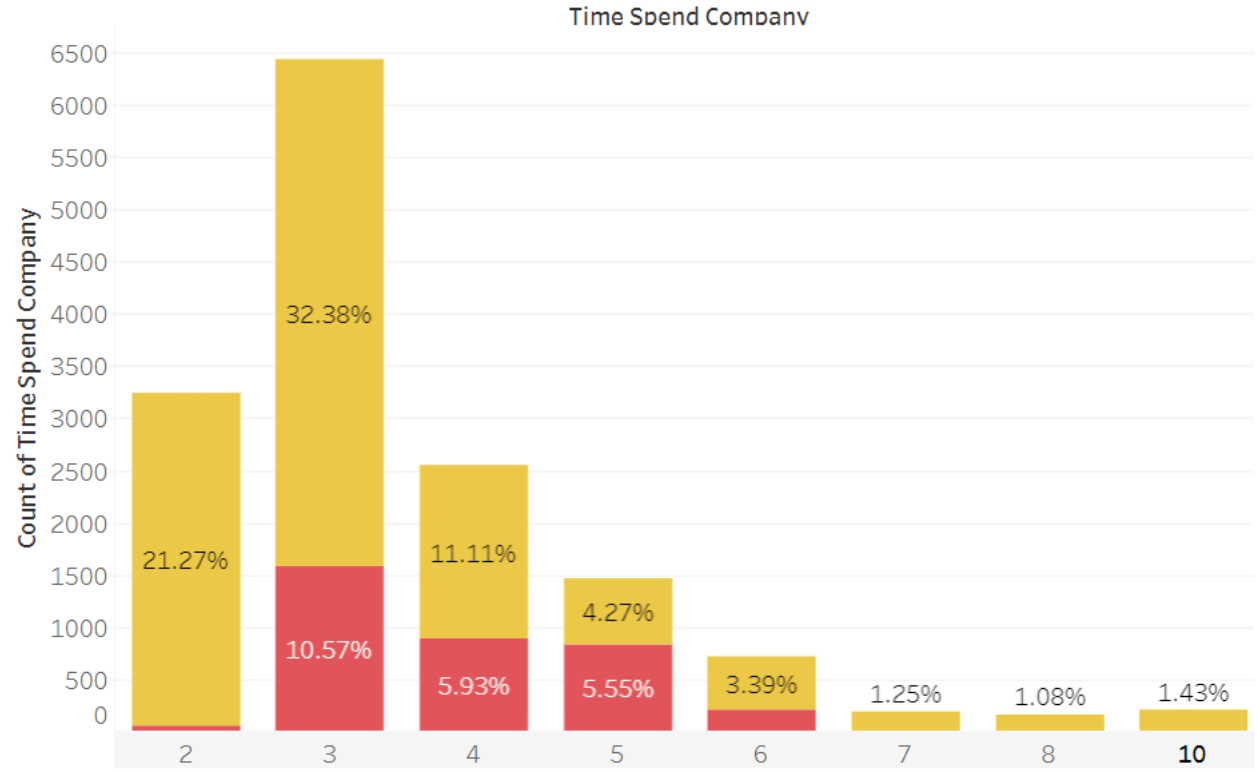


Figure 5: Percentage of Employees Left vs Remaining; against Years Spent at the company

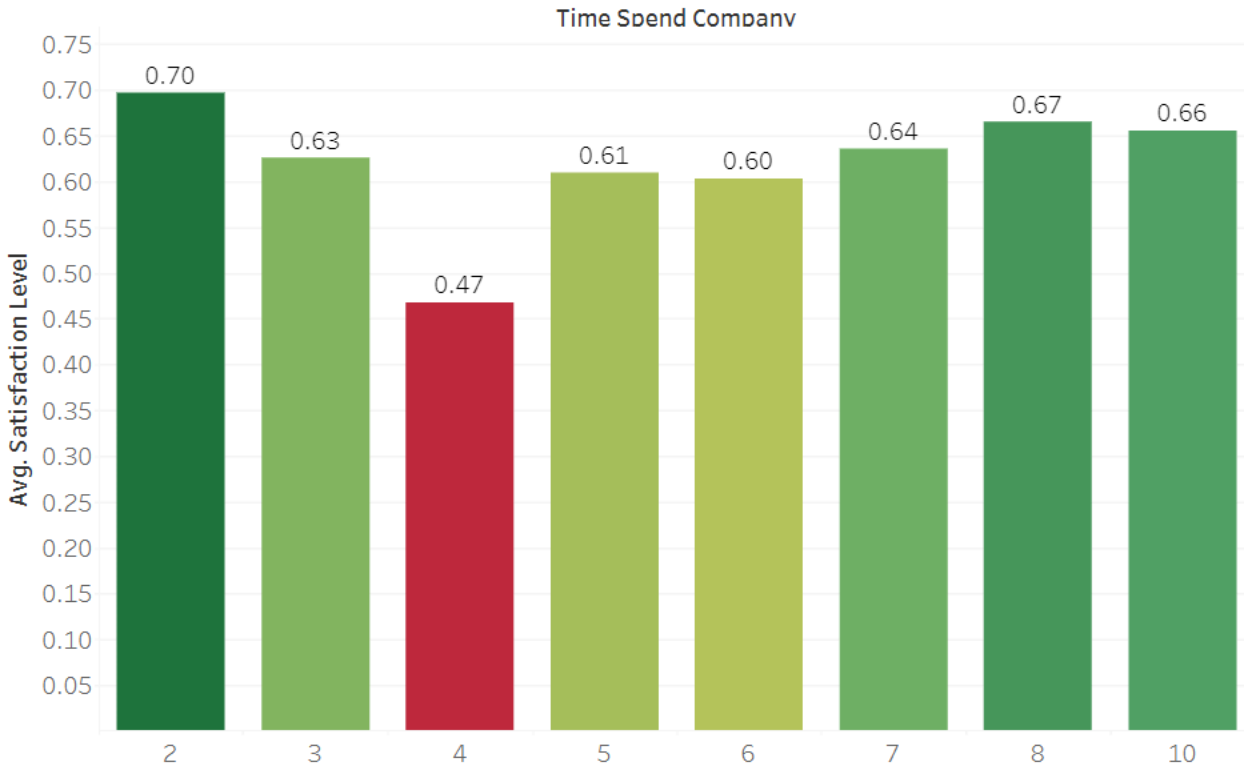


Figure 6: Average Satisfaction Level of Employees at different tenures

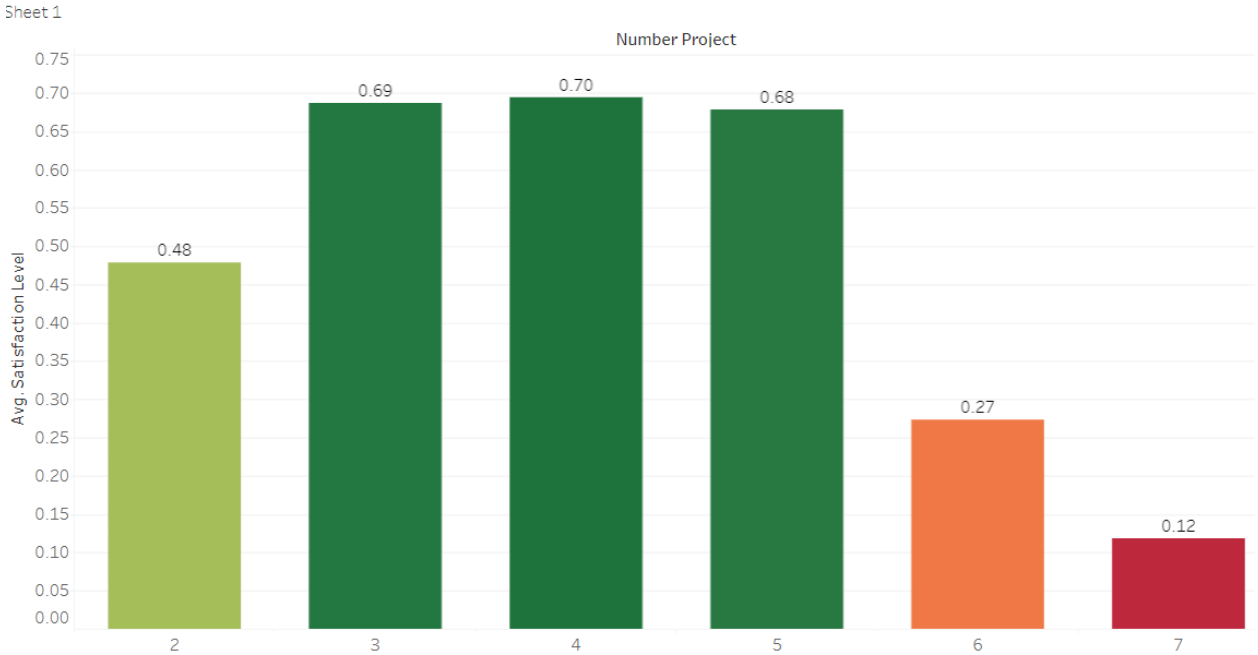


Figure 7: Average Satisfaction Levels of Employees with different numbers of projects assigned

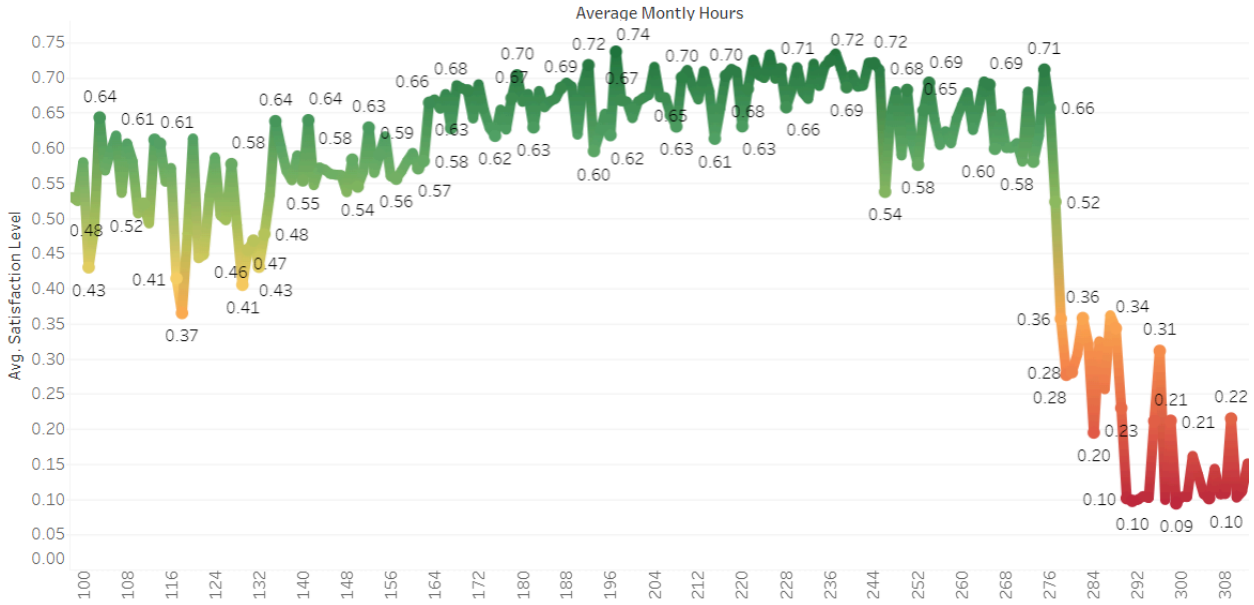


Figure 8: Average Satisfaction Levels across Employees with increasing Average Monthly Hours

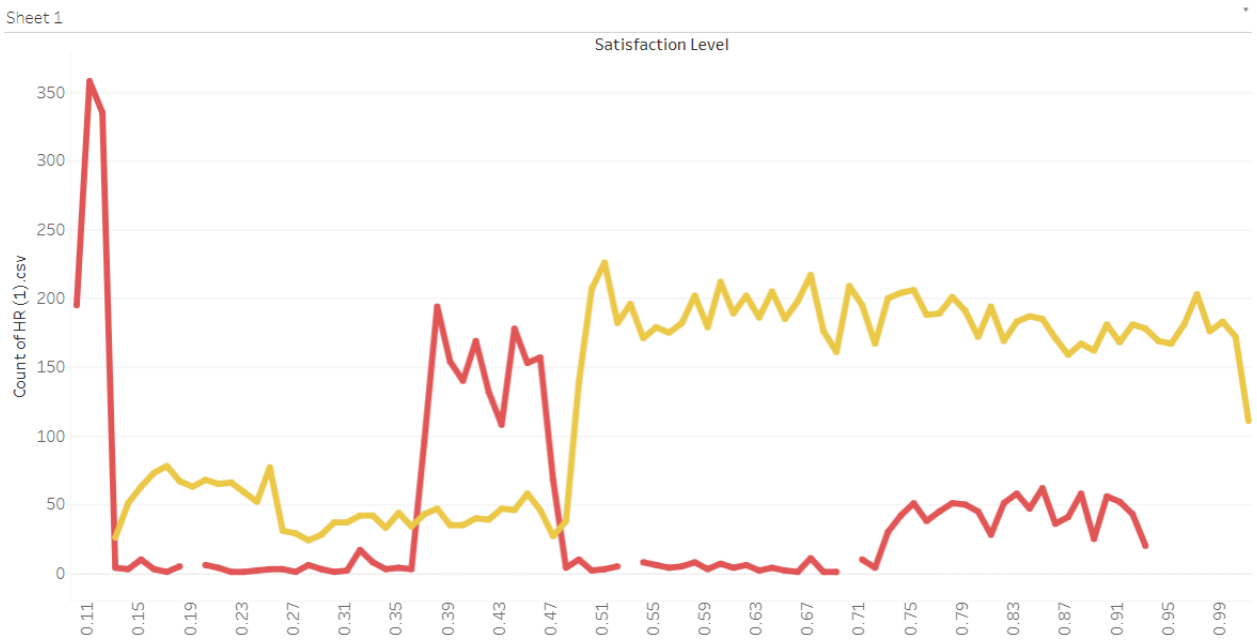


Figure 9: Satisfaction Levels of Employees left (red) vs. remaining (yellow)

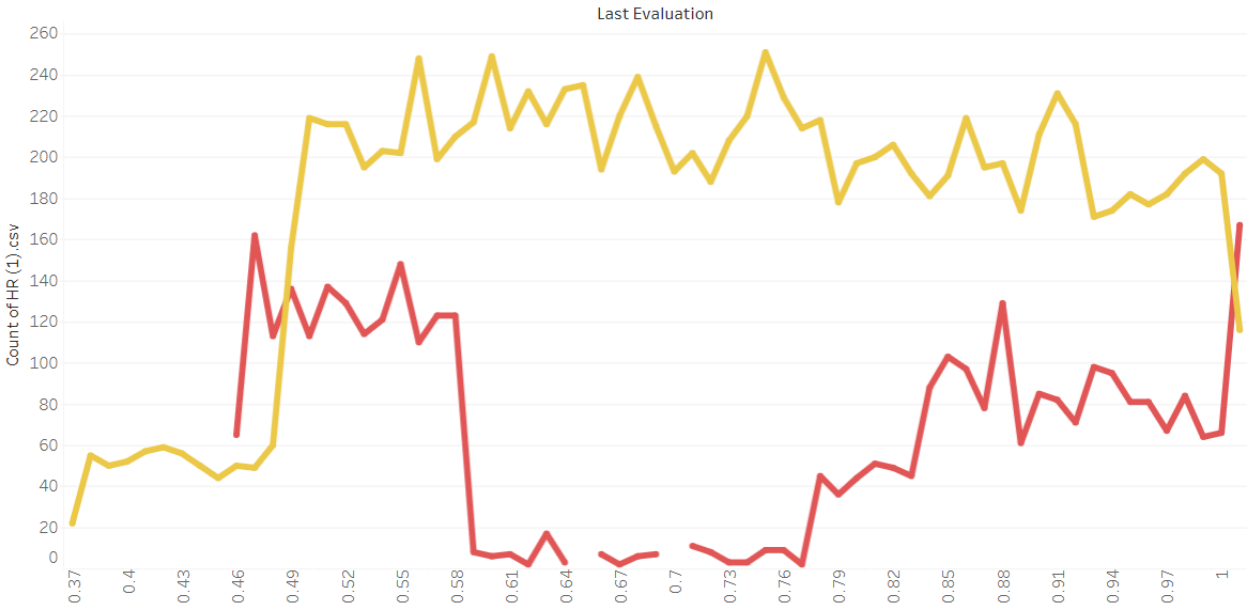


Figure 10: Last Evaluation scores of number of Employees Left (red) vs Remaining (yellow)

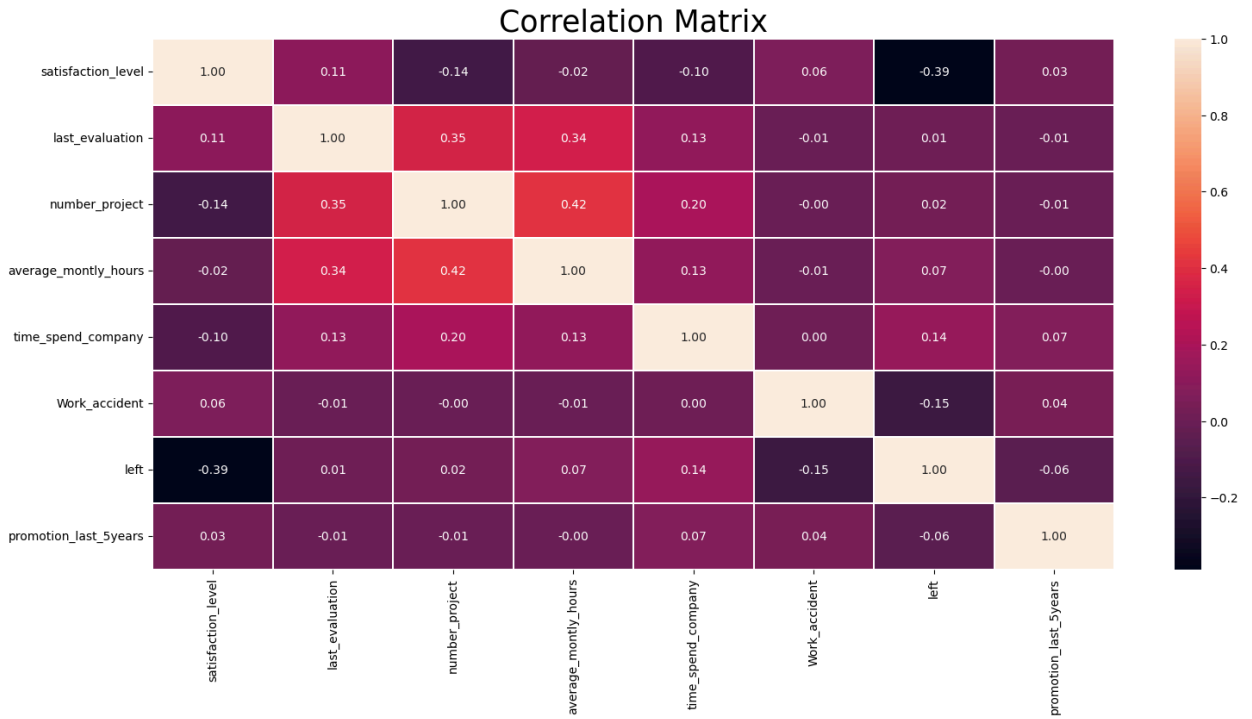


Figure 11: Dataset Correlation Matrix

3.3 Insights from EDA Visualisations:

- The number of current employees is disproportionately higher than the number of employees that left the organisation in this dataset
- The majority of the employees in this dataset come from medium or low-salaried categories, work in the sales department, have not suffered work accidents, and have not received a promotion in the last 5 years.
- Employees who have left the company make up for larger numbers and have lower satisfaction levels compared to employees continuing to work at the organisation.
- Attrition of an employee is negatively correlated to their reported satisfaction level. All other variables do not show any significant correlations with each other in the matrix.
- Employees with both high and low magnitudes in evaluation scores and monthly hours have previously left the company.
- Employees with the lowest average satisfaction level are either at the 4 year-mark of their tenure, have 6+ projects assigned to them, or are working more than 270+ hours on average per month
- Most of the previous employees have left at the 3-year and 4-year mark of their tenure.

3.4. Models:

As mentioned in the methodology section, 4 models were run for this project, namely - ANOVA for satisfaction along with Tukey's HSD, Multivariate Multiple Linear Regression, Multiple Linear Regression for Satisfaction, and Multiple Linear Regression for Attrition.

ANOVA:

ANOVA was used to see if there are significant differences in satisfaction levels across different levels of salary (low, medium and high), department (sales, HR, accounting, marketing, etc), work accident (yes and no), and promotion in last 5 years (yes and no).

Anova was run only on the dependent variable satisfaction since the attrition variable wasn't continuous numeric in nature. Additionally, only 4 independent variables were chosen for ANOVA since only these variables had sub-levels in them.

ANOVA on satisfaction and salary revealed that there is a significant difference, with higher salaries having higher satisfaction, according to Tukey's HSD test. Similarly, significant differences were highlighted in satisfaction and department, with departments like management and IT having higher satisfaction and sales and support having lower satisfaction.

```
               Df Sum Sq Mean Sq F value   Pr(>F)
salary          2    2.3  1.1693   18.96 5.97e-09 ***
Residuals   14996  924.8  0.0617
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
> TukeyHSD(mod_2)
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = satisfaction_level ~ salary, data = data)

$salary
                    diff         lwr          upr     p adj
low-high    -0.03671654 -0.05461098 -0.018822102 0.0000046
medium-high -0.01565305 -0.03372130  0.002415191 0.1049520
medium-low   0.02106349  0.01111999  0.031006988 0.0000021
```

ANOVA: Satisfaction and Salary

```
              Df Sum Sq Mean Sq F value Pr(>F)
sales          9    1.2 0.13334   2.158 0.0219 *
Residuals   14989  925.9 0.06177
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
> TukeyHSD(mod_3)
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = satisfaction_level ~ sales, data = data)

$sales
                               diff           lwr         upr     p adj
hr-accounting            0.0166579630 -0.0238793796 0.05719531 0.9538782
IT-accounting            0.0359905707 -0.0002091571 0.07219030 0.0528192
management-accounting    0.0391979678 -0.0030876751 0.08148361 0.0966120
marketing-accounting     0.0364501600 -0.0026292545 0.07552957 0.0919908
product_mng-accounting   0.0374829077 -0.0011439970 0.07610981 0.0659094
RandD-accounting         0.0376708707 -0.0022319090 0.07757365 0.0833786
sales-accounting         0.0322956213  0.0013803638 0.06321088 0.0322262
support-accounting       0.0361484474  0.0032268515 0.06907004 0.0184129
technical-accounting     0.0257458202 -0.0064060715 0.05789771 0.2503949
IT-hr                    0.0193326077 -0.0172866847 0.05595190 0.8122838
management-hr            0.0225400047 -0.0201053687 0.06518538 0.8112309
marketing-hr             0.0197921970 -0.0196761817 0.05926058 0.8549549
product_mng-hr           0.0208249447 -0.0181954352 0.05984532 0.8022261
RandD-hr                 0.0210129077 -0.0192708869 0.06129670 0.8230009
sales-hr                 0.0156376583 -0.0157678416 0.04704316 0.8602051
support-hr               0.0194904843 -0.0138919026 0.05287287 0.7046532
technical-hr             0.0090878572 -0.0235356991 0.04171141 0.9970235
management-IT            0.0032073971 -0.0353383163 0.04175311 0.9999999
marketing-IT             0.0004595893 -0.0345389401 0.03545812 1.0000000
product_mng-IT           0.0014923371 -0.0330001867 0.03598486 1.0000000
RandD-IT                 0.0016803000 -0.0342352711 0.03759587 1.0000000
sales-IT                -0.0036949494 -0.0292576090 0.02186771 0.9999867
support-IT               0.0001578767 -0.0277979438 0.02811370 1.0000000
technical-IT            -0.0102447505 -0.0372899084 0.01680041 0.9727590
```

ANOVA: Satisfaction and Department

ANOVA on satisfaction and work accidents, as well as, ANOVA on satisfaction and promotion in the last 5 years saw significant differences, rejecting the null hypothesis that there is no significant difference between them.

```
                          Df Sum Sq Mean Sq F value  Pr(>F)
promotion_last_5years      1    0.6  0.6079   9.839 0.00171 **
Residuals              14997  926.5  0.0618
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> TukeyHSD(mod_4)
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = satisfaction_level ~ promotion_last_5years, data = data)

$promotion_last_5years
          diff       lwr        upr     p adj
1-0 0.04412371 0.0165508 0.07169663 0.0017119
```

ANOVA: Satisfaction and promotion

```
                Df Sum Sq Mean Sq F value   Pr(>F)
Work_accident    1    3.2   3.194   51.85 6.28e-13 ***
Residuals    14997  923.9   0.062
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> TukeyHSD(mod_5)
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = satisfaction_level ~ Work_accident, data = data)

$Work_accident
          diff        lwr        upr p adj
1-0 0.04149321 0.03019809 0.05278834     0
```

ANOVA: Satisfaction and work accidents

Multivariate Multiple Linear Regression:

```
Coefficients:
                        Estimate Std. Error t value Pr(>|t|)
(Intercept)            0.5177089  0.0293702  17.627  < 2e-16 ***
last_evaluation        0.2013398  0.0243915   8.254  < 2e-16 ***
number_project        -0.0476825  0.0035248 -13.528  < 2e-16 ***
average_montly_hours   0.0006042  0.0000858   7.042 2.01e-12 ***
time_spend_company     0.0295047  0.0026603  11.091  < 2e-16 ***
Work_accident         -0.1416517  0.0107604 -13.164  < 2e-16 ***
promotion_last_5years -0.1176387  0.0266085  -4.421 9.92e-06 ***
saleshr                0.0395765  0.0239885   1.650   0.0990 .
salesIT               -0.0170734  0.0212935  -0.802   0.4227
salesmanagement       -0.0411773  0.0253335  -1.625   0.1041
salesmarketing         0.0153255  0.0231032   0.663   0.5071
salesproduct_mng       0.0006232  0.0226886   0.027   0.9781
salesRandD            -0.0542847  0.0235805  -2.302   0.0213 *
salessales             0.0125764  0.0181674   0.692   0.4888
salessupport           0.0203324  0.0193439   1.051   0.2932
salestechnical         0.0176432  0.0189056   0.933   0.3507
salarylow              0.2064583  0.0149134  13.844  < 2e-16 ***
salarymedium           0.1452645  0.0149490   9.717  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3864 on 10366 degrees of freedom
Multiple R-squared:  0.07222,   Adjusted R-squared:  0.07069
F-statistic: 47.46 on 17 and 10366 DF,  p-value: < 2.2e-16
```

On running satisfaction and left as dependent variables and the remaining 8 variables as predictors, the following was obtained from the model:

- Multiple R-square: 0.07222
- Adjusted R-square: 0.07069

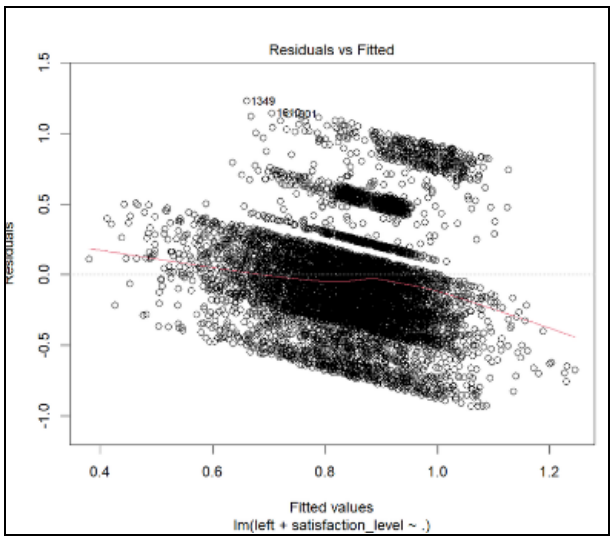
The multiple R-square value of 0.07222 indicates that approximately 7.22% of total variance in the dependent variables is explained by the independent variables collectively. The adjusted R-square value of 0.07069, which is slightly lower than the multiple R-square value indicates that the addition of independent variables to the model doesn't substantially improve its explanatory power.

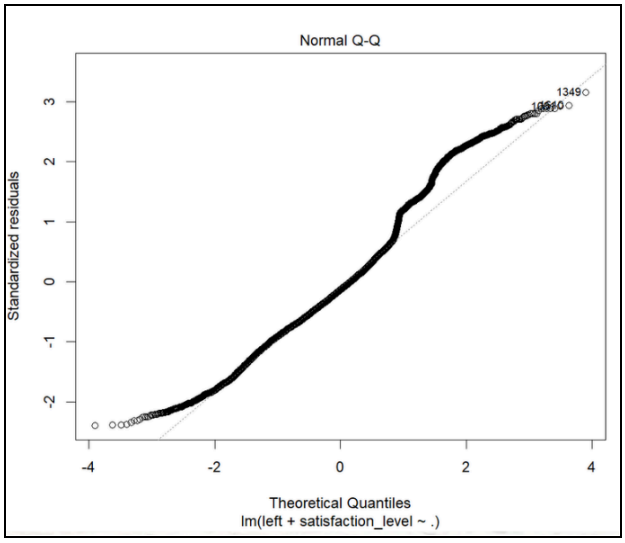
In fact, these were the highest R-squared values obtained from the model. Backward elimination method of this model saw relatively lower values. The significant variables according to this model are - Significant Variables:

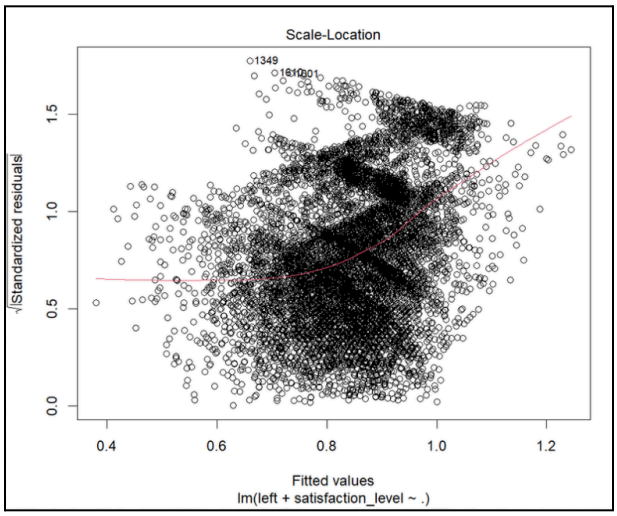
- Last evaluation
- Number of Projects
- Average Monthly Hours worked
- Time Spent in the Company
- Work Accident
- Promotions in the last 5 years
- R&D department (dummy variable from department)
- Salary

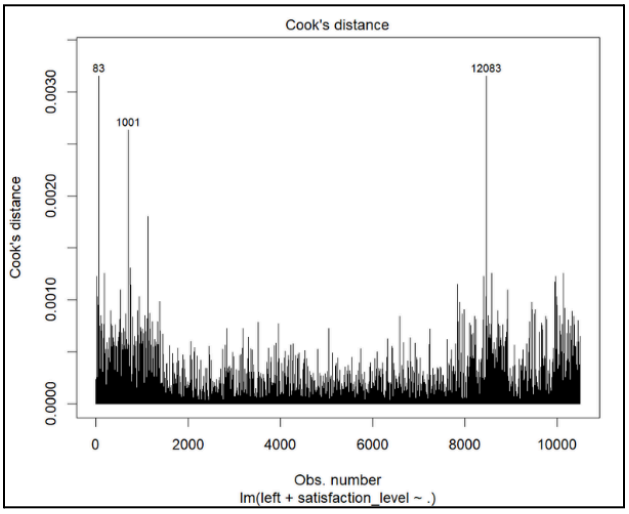
Analysing the assumptions of linear regression on the MMLR model:

- The Q-Q Plot is mostly on the line, indicating normal distribution
- The Residuals vs Fitted graph and Scale-Location graph indicate heteroscedasticity. Additionally, it is seen that the residuals are evenly spread horizontally but not vertically in the Scale-Location plot.
- Breusch-Pagan test further confirms heteroscedasticity with p-value being less than 0.05.
- Cook's Distance plot indicates three main outliers only with no impact.









-

```
> bptest(mmlr_model)

        studentized Breusch-Pagan test

data:  mmlr_model
BP = 1669.7, df = 17, p-value < 2.2e-16
```

Interpretation of the significant variables of the MMLR model:

- **Last Evaluation**: Higher performance evaluations generally correlate with higher satisfaction levels and lower attrition rates. Poor evaluations may indicate dissatisfaction and risk of leaving.
- **Number of Projects**: More projects are associated with lower satisfaction levels and higher attrition rates, likely due to increased workload and pressure.
- **Average Monthly Hours**: Longer work hours may lead to higher satisfaction levels, but employees with fewer hours are at risk of leaving, possibly due to dissatisfaction.
- **Time Spent at Company**: Longer tenure is associated with higher satisfaction and lower attrition, indicating that satisfied employees tend to stay longer.
- **Work Accidents**: Work accidents correlate with lower satisfaction and higher attrition, suggesting negative experiences and an increased likelihood of leaving.
- **Promotions in Last Five Years**: Surprisingly, promotions are associated with lower satisfaction and higher attrition, possibly due to dissatisfaction with salary adjustments or workplace experiences post-promotion.
- **Salary (Low and Medium)**: Lower and medium salary levels are significant predictors of satisfaction and attrition, likely contributing to employees' decisions to leave for better-paying roles.
- **Department (R&D)**: Employees in the R&D department tend to have lower satisfaction levels and higher attrition rates compared to other departments, suggesting potential dissatisfaction with the company.

Multiple Linear Regression of Satisfaction:

The MLR model with satisfaction as the dependent variable and the same 8 independent variables produced:

```
Call:
lm(formula = satisfaction_level ~ last_evaluation + number_project +
    average_montly_hours + time_spend_company + Work_accident +
    promotion_last_5years + sales + salary, data = train_data)

Residuals:
     Min       1Q   Median       3Q      Max
-0.63782 -0.18387  0.01992  0.19744  0.59856

Coefficients:
                        Estimate Std. Error t value Pr(>|t|)
(Intercept)            5.990e-01  1.815e-02  33.006  < 2e-16 ***
last_evaluation        2.613e-01  1.508e-02  17.327  < 2e-16 ***
number_project        -3.639e-02  2.183e-03 -16.672  < 2e-16 ***
average_montly_hours   5.452e-05  5.328e-05   1.023  0.30619
time_spend_company    -1.732e-02  1.666e-03 -10.395  < 2e-16 ***
Work_accident          4.514e-02  6.675e-03   6.762 1.43e-11 ***
promotion_last_5years  3.655e-02  1.628e-02   2.246  0.02475 *
saleshr                1.327e-02  1.472e-02   0.902  0.36729
salesIT                3.874e-02  1.314e-02   2.948  0.00320 **
salesmanagement        4.574e-02  1.575e-02   2.904  0.00369 **
salesmarketing         2.924e-02  1.416e-02   2.064  0.03904 *
salesproduct_mng       4.458e-02  1.406e-02   3.171  0.00152 **
salesRandD             3.050e-02  1.460e-02   2.088  0.03679 *
salessales             3.501e-02  1.120e-02   3.126  0.00178 **
salessupport           3.755e-02  1.195e-02   3.142  0.00168 **
salestechnical         3.034e-02  1.166e-02   2.601  0.00931 **
salarylow             -3.623e-02  9.206e-03  -3.935 8.36e-05 ***
salarymedium          -1.462e-02  9.254e-03  -1.579  0.11427
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2412 on 10481 degrees of freedom
Multiple R-squared:  0.06233,   Adjusted R-squared:  0.06081
F-statistic: 40.98 on 17 and 10481 DF,  p-value: < 2.2e-16
```

Multiple R-square: 0.06233

Adjusted R-square: 0.06081

This indicates that 6.2% of the variance in satisfaction is explained by the independent variables. Since the value is very low, it indicates that the model is not really effective in explaining the variance. According to it, all independent variables except *average monthly hours*, *salarymedium,* and *saleshr* have statistical significance.

Multiple Linear Regression of Attrition:

```
Call:
lm(formula = left ~ last_evaluation + number_project + average_montly_hours +
    time_spend_company + Work_accident + promotion_last_5years +
    sales + salary, data = train_data)

Residuals:
     Min       1Q   Median       3Q      Max
-0.69444 -0.27625 -0.18071  0.09029  1.03187

Coefficients:
                       Estimate Std. Error t value Pr(>|t|)
(Intercept)           -6.667e-02  3.070e-02  -2.172 0.029896 *
last_evaluation       -8.077e-02  2.551e-02  -3.167 0.001547 **
number_project        -9.762e-03  3.693e-03  -2.644 0.008211 **
average_montly_hours   6.367e-04  9.013e-05   7.064 1.71e-12 ***
time_spend_company     4.849e-02  2.818e-03  17.205  < 2e-16 ***
Work_accident         -1.818e-01  1.129e-02 -16.098  < 2e-16 ***
promotion_last_5years -1.451e-01  2.754e-02  -5.270 1.39e-07 ***
saleshr                2.215e-02  2.490e-02   0.890 0.373732
salesIT               -5.328e-02  2.223e-02  -2.397 0.016568 *
salesmanagement       -9.686e-02  2.664e-02  -3.635 0.000279 ***
salesmarketing        -2.565e-02  2.396e-02  -1.071 0.284368
salesproduct_mng      -6.022e-02  2.379e-02  -2.532 0.011362 *
salesRandD            -1.074e-01  2.471e-02  -4.347 1.39e-05 ***
salessales            -3.505e-02  1.895e-02  -1.850 0.064367 .
salessupport          -2.366e-02  2.022e-02  -1.170 0.241845
salestechnical        -1.218e-02  1.973e-02  -0.617 0.536958
salarylow              2.206e-01  1.557e-02  14.163  < 2e-16 ***
salarymedium           1.320e-01  1.565e-02   8.435  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.408 on 10481 degrees of freedom
Multiple R-squared:  0.08378,   Adjusted R-squared:  0.0823
F-statistic: 56.38 on 17 and 10481 DF,  p-value: < 2.2e-16
```

The MLR model with attrition as the dependent variable and the same 8 independent variables produced:

Multiple R-square: 0.08378
Adjusted R-square: 0.0823

This indicates that 8.3% of the variance in attrition is explained by the independent variables. Since the value is very low, it indicates that the model is not effective in explaining the variance. According to it, all independent variables except *salessupport*, *salestechnical*, *salesmarketing* and *saleshr* have statistical significance.

It is also worth noting that this model has a higher multiple R-squared value compared to the MLR model of satisfaction and the MMLR model.

Multicollinearity:

```
> vif_values <- car::vif(mmlr_model)
> # Display VIF values
> print(vif_values)
                         GVIF Df GVIF^(1/(2*Df))
last_evaluation      1.206029  1        1.098194
number_project       1.305364  1        1.142525
average_montly_hours 1.273118  1        1.128325
time_spend_company   1.069800  1        1.034311
Work_accident        1.003248  1        1.001623
promotion_last_5years 1.033363  1        1.016545
sales                1.083310  9        1.004456
salary               1.052630  2        1.012906
```

Multicollinearity was checked in the dataset, with all variables having VIF value < 5, indicating low multicollinearity.

4. DISCUSSION

The findings from our various models give us a deeper insight into the different factors that affect and influence employee satisfaction and attrition. The insights into these various factors can be leveraged by a company to implement decisions that will help with employee retention and satisfaction. From allocating the number of projects per month to monitoring monthly work hours, the company can make adjustments according to employee preferences to better increase satisfaction.

The MLR Models can be used for predictive analysis, to see which employees might be at risk of leaving the company based on the current factors. The analysis can also be used for prescriptive modelling to encourage employees to stay. The company can also foster tailored programs for employee improvement and growth. Employee feedback can be taken through discussion/focus groups, surveys, and quarterly reviews.

The major limitation or constraint that we dealt with was the heteroscedasticity in our data, which did have an effect on our R^2. Some of the reasons for the significance of certain variables remain unexplained, which can maybe be further identified only through qualitative research.

Another limitation we have faced is the diversity in the different work environments in different regions, professions and departments. The workload, number of projects and in general job descriptions differ from department to department. Moving forward, we need to keep in mind these requirements and tailor the independent variables accordingly.

Qualitative research will prove to be extremely useful for future research, especially to capture general sentiment of employees in an organization. The continuous gathering and updating of data will also be key in maintaining a valid matrix for employee satisfaction and retention.

5. CONCLUSION

In conclusion, our research has revealed important information on the dynamics of employee happiness and attrition, highlighting crucial areas in which organizations need to make improvements. Important findings are as follows

R&D Department Review -

Although the R&D department's low attrition rate is positive indicating that our employees are not leaving, our data shows us a pattern of poor satisfaction levels in regards to this department. To mitigate this discrepancy, a thorough evaluation procedure must be followed, employee input must be gathered via surveys, and incentive-based initiatives must be put in place to boost retention and satisfaction.

Streamlining the Promotion Process -

The importance of timely promotions in enhancing employee satisfaction is underscored by our ANNOVA and MMLR models. Extended promotion schedules raise the likelihood of attrition, even when promotions within five years are associated with higher satisfaction. Therefore, it can be beneficial to satisfaction scores to optimize the promotion process to happen ideally within two to three years.

Workload Balance -

The unfavourable impact of workload intensity on satisfaction is highlighted by the negative coefficient linked to the number of projects. On the other hand, our data indicates a marginal rise in satisfaction with higher monthly hours worked, highlighting the necessity of a balanced

workload. Organisations can do that through efficient work practices, and promoting well-being initiatives.

Fair Compensation -

Important results from our MMLR model (SalaryLow and SalaryMedium) highlight how crucial fair compensation is to raising retention and satisfaction levels. Employers need to assess and modify their pay plans and salary structure in order to meet the needs of their workforce and create a positive work atmosphere that will help retain employees over the long run.

Additionally, we found that the overall explanatory power of our MMLR model, as indicated by the Multiple R-square and Adjusted R-square values, was relatively low, suggesting that the included predictors accounted for only a small proportion of the variance in employee satisfaction and attrition. This highlights the complexity of these phenomena and suggests that additional factors not captured in our model may also play significant roles.

To conclude, we created a model which enables organizations to successfully forecast and mitigate employee turnover risks through the application of ANOVA, multiple linear regression, and multivariate multiple linear regression methodologies. This facilitates data-driven decision-making in HR management. Moreover, the incorporation of qualitative insights strengthens the practical relevance of our results, establishing a connection between theoretical understanding and actual business difficulties.

Analytics has the ability to significantly change organizational strategies and promote a continuous improvement culture. Reflecting on this project we aim to use our knowledge and skills to promote innovation and constructive change in the field of business analytics. It deepened our appreciation for the complexities of human behavior and organizational dynamics. While our analysis has provided some valuable insights, it has also underscored the need for humility, curiosity, and ongoing inquiry in the field of business analytics. As we move forward, we are committed to leveraging these lessons to drive continuous improvement and innovation in our analytical practices, ultimately contributing to the advancement of both theory and practice in the field.

6. REFERENCES

https://www.kaggle.com/datasets/aminaidhm23386/employee-worker

7. APPENDIX

1. Data and Code

1.1 Raw Data

satisfaction_level	last_evaluation	number_projec	average_montly_h	time_spend_co	Work_accident	left	promotion_last	sales	salary
0.38	0.53	2	157	3	0	1	0	sales	low
0.8	0.86	5	262	6	0	1	0	sales	medium
0.11	0.88	7	272	4	0	1	0	sales	medium
0.72	0.87	5	223	5	0	1	0	sales	low
0.37	0.52	2	159	3	0	1	0	sales	low
0.41	0.5	2	153	3	0	1	0	sales	low
0.1	0.77	6	247	4	0	1	0	sales	low
0.92	0.85	5	259	5	0	1	0	sales	low

HR Dataset

1.2 R-Studio Code

```
library(Ecdat)
library(ggplot2)
library(ggcorrplot)
library(GGally)
library(dplyr)
library(caret)
library(lmtest)
library(rpart)
library(rpart.plot)
library(caret)
library(car)
library(pscl)
library(MLmetrics)
library(corrplot)

# Set the working directory and load the CSV file
setwd("/Users/saiveephatak/Downloads/")
getwd()
data1 <- read.csv("HR.csv", stringsAsFactors = TRUE)

# Check the structure of the dataset
str(data1)
```

```
# Convert 'sales' into dummy variables
dummy_variables <- model.matrix(~ sales - 1, data = data1)
# Convert 'salary' into dummy variables
dummy_variables <- model.matrix(~ salary - 1, data = data1)

# Remove the intercept column (if needed)
# dummy_variables <- dummy_variables[, -1]

# Bind the dummy variables to the original dataset
data <- cbind(data1, dummy_variables)

# Fit MANOVA model
manova_model <- manova(cbind(left, satisfaction_level) ~
last_evaluation+number_project+average_montly_hours+time_spend_company+Work_a
ccident+promotion_last_5years+sales+salary, data = data)

# Print summary of MANOVA model
summary(manova_model)

library(caTools)

# Splitting the data into training and testing sets
set.seed(123) # For reproducibility
split <- sample.split(data, SplitRatio = 0.7) # Splitting 70% for training, 30% for testing
train_data <- subset(data, split == TRUE)
test_data <- subset(data, split == FALSE)

# Extracting predictor variables (X) and response variables (Y) for training set
X_train <- train_data[, c("last_evaluation", "number_project", "average_montly_hours",
"time_spend_company", "Work_accident", "promotion_last_5years", "sales", "salary")]
Y_train <- train_data[, c("left", "satisfaction_level")]

# Extracting predictor variables (X) and response variables (Y) for testing set
X_test <- test_data[, c("last_evaluation", "number_project", "average_montly_hours",
"time_spend_company", "Work_accident", "promotion_last_5years", "sales", "salary")]
Y_test <- test_data[, c("left", "satisfaction_level")]

train_data <- cbind(X_train, Y_train)
```

```
# Fit MMLR model on training set
mmlr_model <- lm(left + satisfaction_level ~ ., data = train_data)

# Print summary of MMLR model
summary(mmlr_model)

mmlr_model_1 <- lm(left + satisfaction_level ~
last_evaluation+number_project+average_montly_hours+time_spend_company+Work_a
ccident+promotion_last_5years+salary, data = train_data)
summary(mmlr_model_1)

vif(mmlr_model)

data <- subset(data, select = -salary)
data <- subset(data, select = -sales)

# Calculate correlation matrix
correlation_matrix <- cor(data)
correlation_matrix

# Create a correlation plot
corrplot(correlation_matrix, method = "color", type = "upper", tl.col = "black", tl.srt =
45)

# Predict on the testing set
Y_pred <- predict(mmlr_model, newdata = X_test)
Y_pred

plot(mmlr_model, 1)
plot(mmlr_model, 2)
plot(mmlr_model, 3)
plot(mmlr_model, 4)

# Predictions on testing data
predictions <- predict(mmlr_model, newdata = test_data)

# Extract predicted satisfaction level and attrition status
predicted_satisfaction <- predictions[, "satisfaction_level"]
predicted_attrition <- predictions[, "left"]
```

```
# Predictions on testing data
predictions <- predict(mmlr_model, newdata = test_data)


# Calculate evaluation metrics (e.g., RMSE, MAE) for satisfaction level prediction
satisfaction_rmse <- sqrt(mean((test_data$satisfaction_level - predicted_satisfaction)^2))
satisfaction_mae <- mean(abs(test_data$satisfaction_level - predicted_satisfaction))

# Calculate evaluation metrics (e.g., accuracy, confusion matrix) for attrition status
prediction
attrition_accuracy <- mean(ifelse((test_data$left - predicted_attrition) == 0, 1, 0))
attrition_confusion_matrix <- table(test_data$left, predicted_attrition)

# Print evaluation metrics
print(paste("Satisfaction Level RMSE:", satisfaction_rmse))
print(paste("Satisfaction Level MAE:", satisfaction_mae))
```

2. Presentation Slides

Data-Driven Employee Turnover Prediction.pdf