IBM EMPLOYEE ATTRTION

Filippo Grandoni, Ludovico Amedeo Panariello, Alice Finotti, Leonardo Tonelli, Federico Giorgi

TABLE OF CONTENT

BIG DATA AND DATABASES PROJECT

DATA EXPLORATION

01 OVERVIEW & OBJECTIVE

OVERVIEW & OBJECTIVE

The dataset comprises a compilation of **1,470 entries**, each corresponding to a **unique employee** and encompassing a total of **35 variables**.

These measure **demographic factors** (e.g., age, gender, marital status, etc.) and **professional attributes** (e.g., department, job role, years at company, etc.), as well as **financial details** (e.g., monthly income, stock options) and **job satisfaction metrics** (e.g., job involvement, work-life balance, performance ratings).

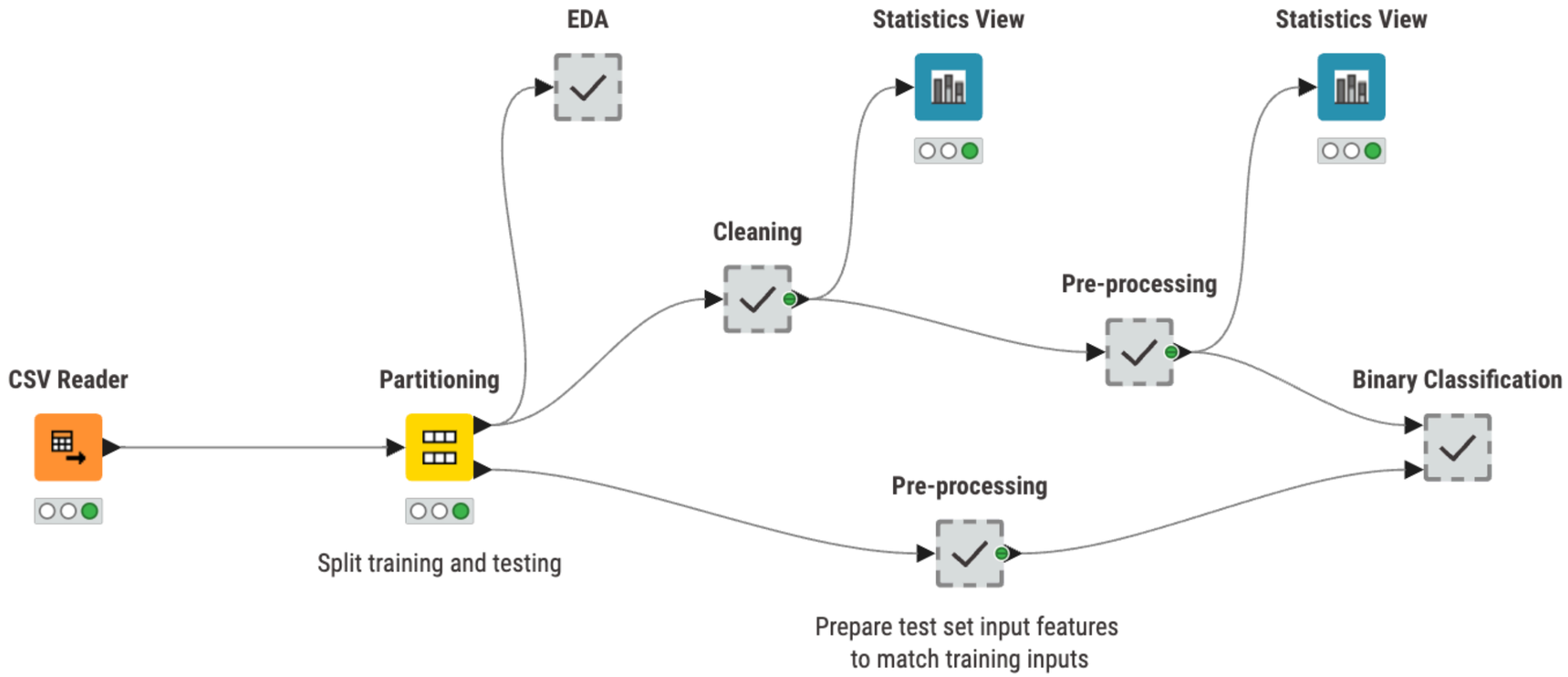
The primary objective of our analysis is to **predict an employee's likelihood of attrition** based on various attributes.

Attrition refers to the **tendency of an employee to leave the organization**, and understanding this behavior is critical for workforce management and strategic planning.

We begin by conducting a comprehensive **data analysis and visualization**. The insights gained will serve as a foundation for constructing **advanced machine learning models**. The ultimate goal of these models is to provide organizations with a reliable and efficient methodology to **anticipate employee turnover and improve retention**.

WORKFLOW

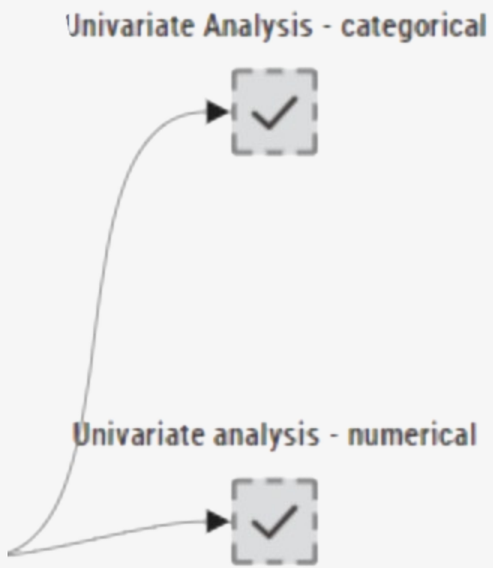
Data Preparation Page



02 UNIVARIATE ANALYSIS

UNIVARIATE ANALYSIS

Introduction



The univariate analysis focuses on examining each variable in the dataset individually to understand its **distribution**, **nature**, and **general patterns**. For each variable:

- **Categorical Variables**: The analysis summarizes proportions using pie charts, helping to visualize the relative frequency of categories.
- **Numerical Variables**: The analysis uses histograms to observe the distribution of data points, identifying trends such as skewness, concentration, or uniformity.

The univariate analysis helps identify:

- **Variable types:** categorical vs. numerical.
- **Patterns:** e.g., skewed distributions or balanced proportions.
- **Potential data imbalances**.

This step provides a **foundational understanding of the dataset**, setting the stage for more complex analyses.

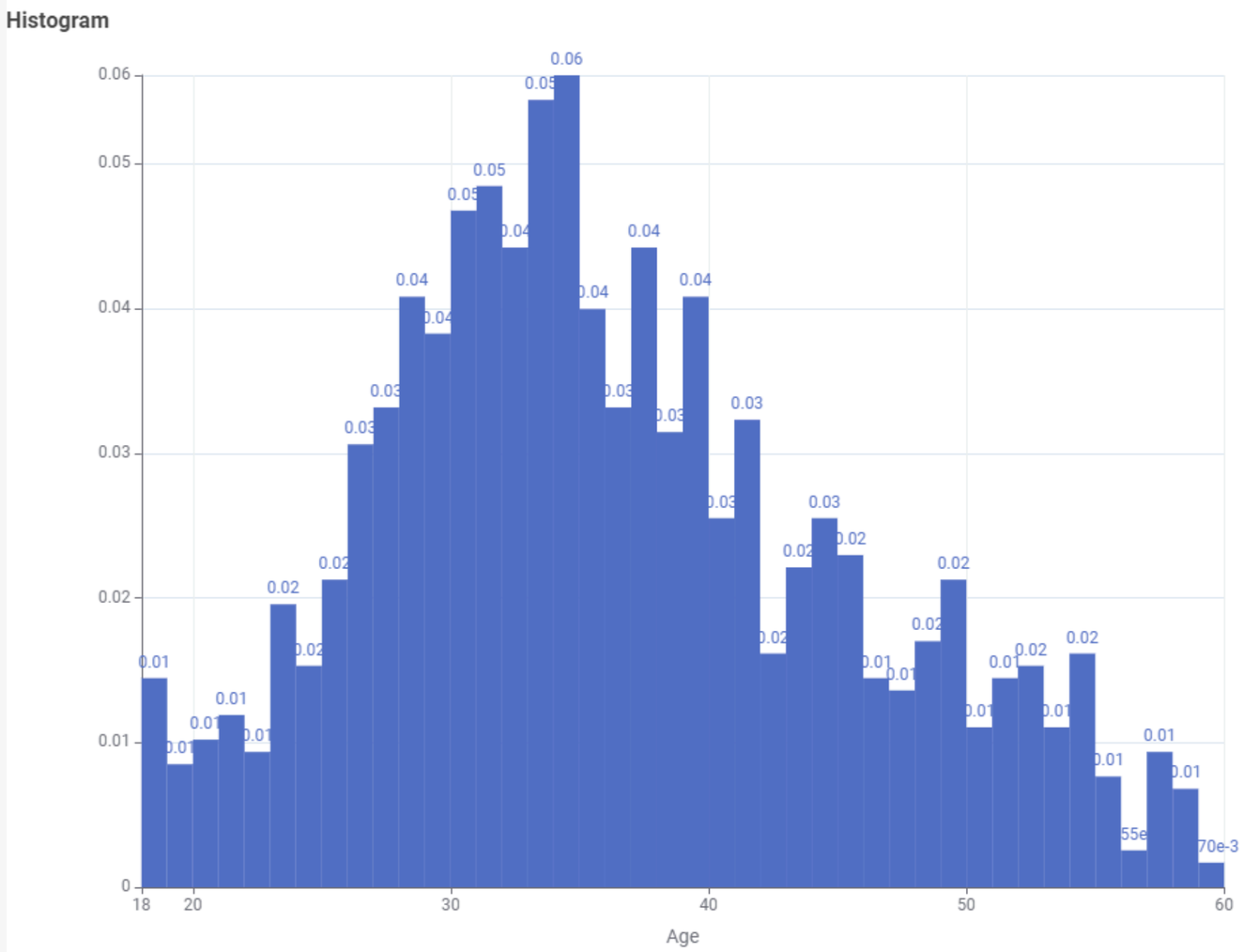
UNIVARIATE ANALYSIS

Age

Nature: Numerical, discrete

Range: 18 - 60

Insights: it is skewed slightly to the right, with the highest concentration of employees in their 30s, gradually reducing as age increases toward 60.

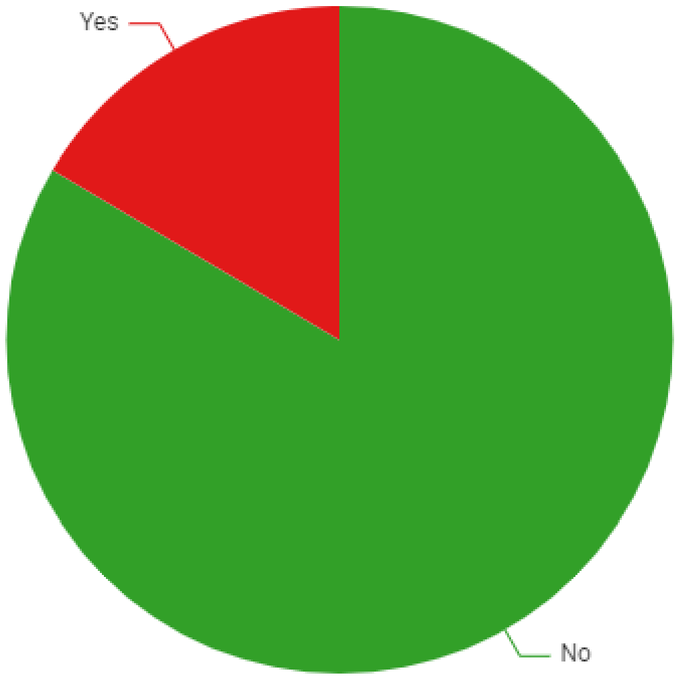


UNIVARIATE ANALYSIS

Attrition

Nature: Categorical, nominal
Description: Employee leaving the company
Categories: 0 = “no”, 1 = “yes”
Insights: it is imbalanced, with a significantly larger proportion of employees who stayed (**No**) compared to those who left (**Yes**).

Attrition Pie Chart



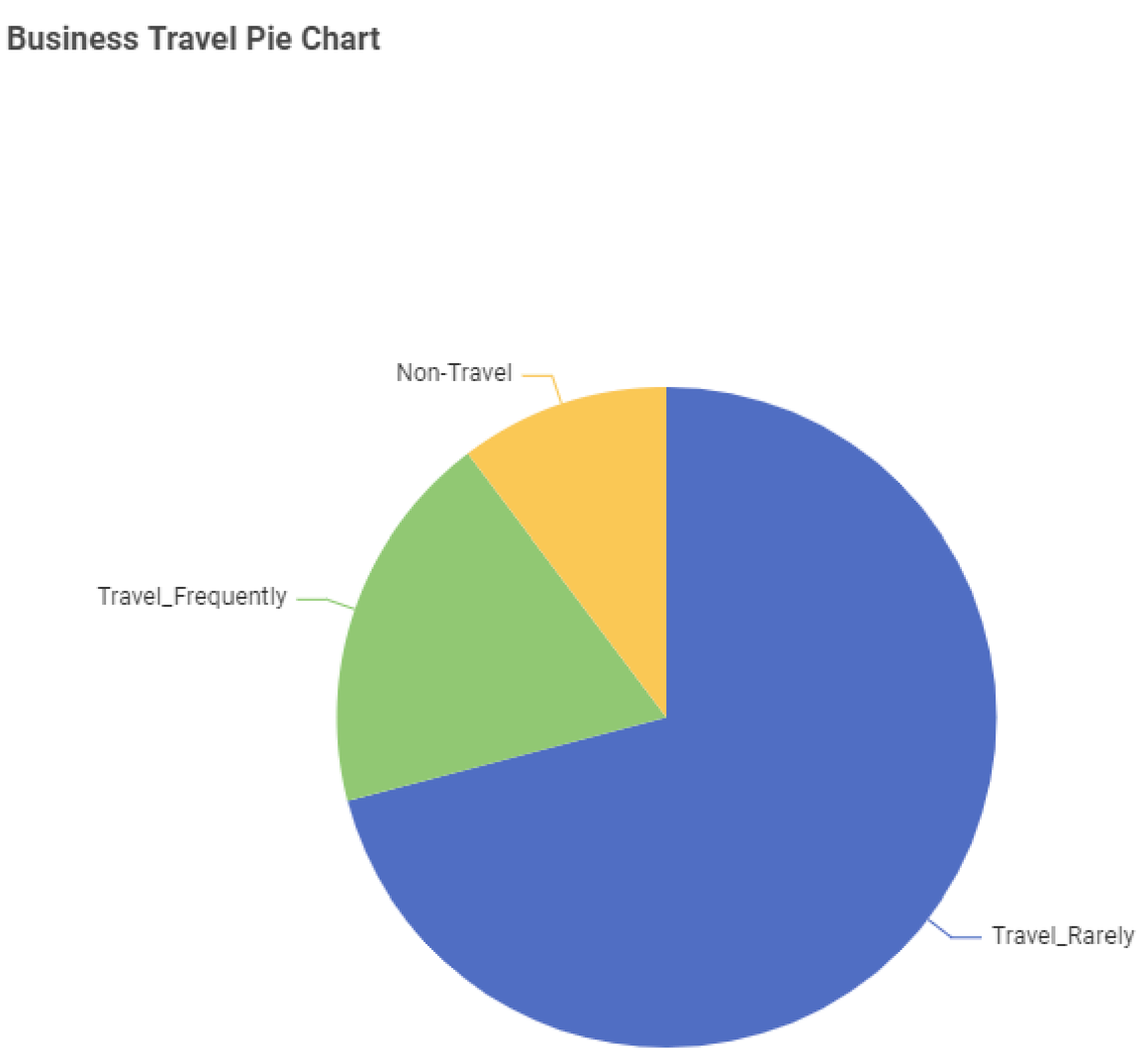
UNIVARIATE ANALYSIS

BusinessTravel

Nature: Categorical, nominal

Categories: 1=No Travel, 2=Travel Frequently, 3=Tavel Rarely

Insights: the majority of employees travel rarely, while fewer employees travel frequently or do not travel at all.



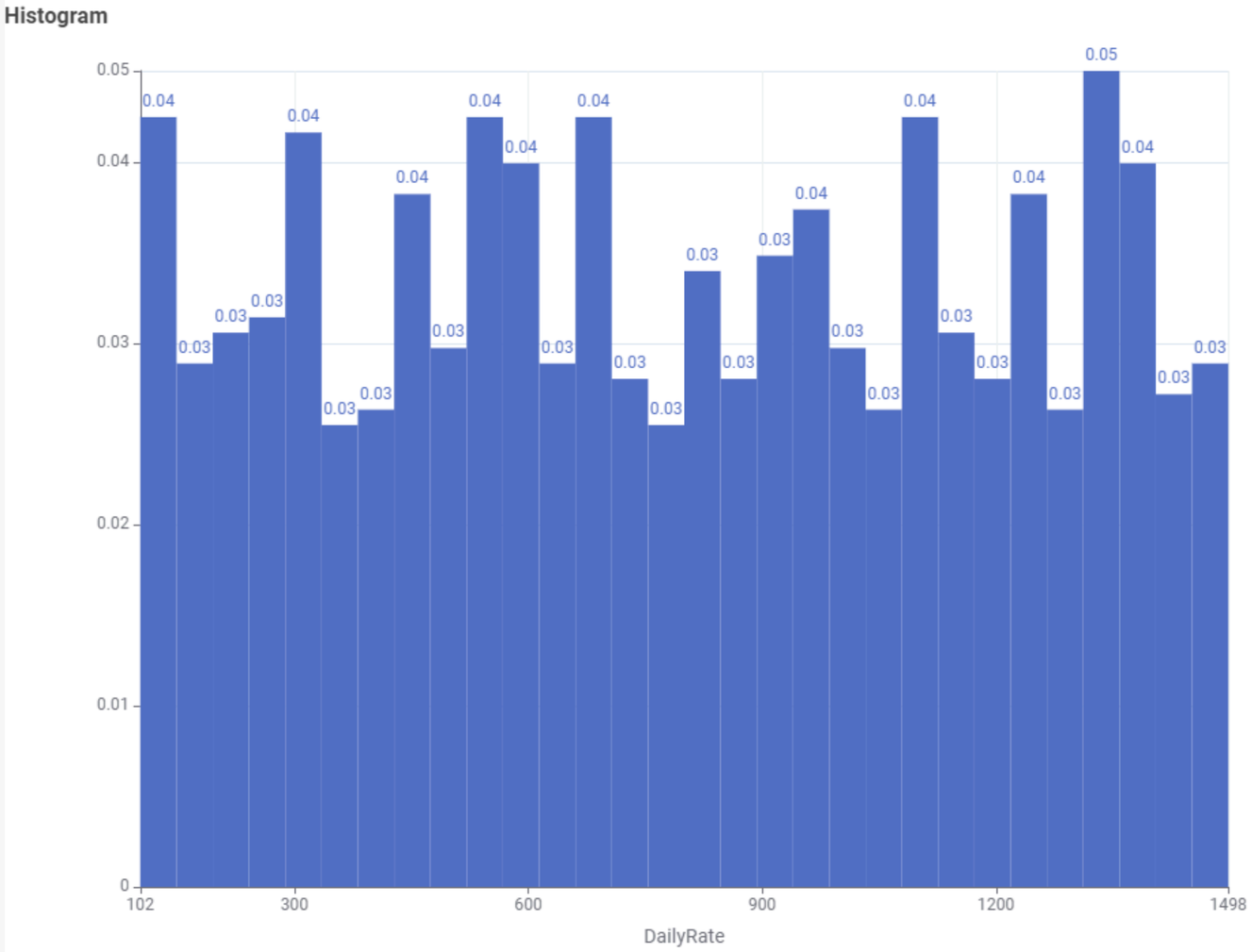
UNIVARIATE ANALYSIS

DailyRate

Nature: Numerical, discrete

Range: 102 - 1498

Insights: it is relatively uniform, indicating that employees are fairly evenly distributed across different levels of daily rates, without any significant concentration in specific ranges.



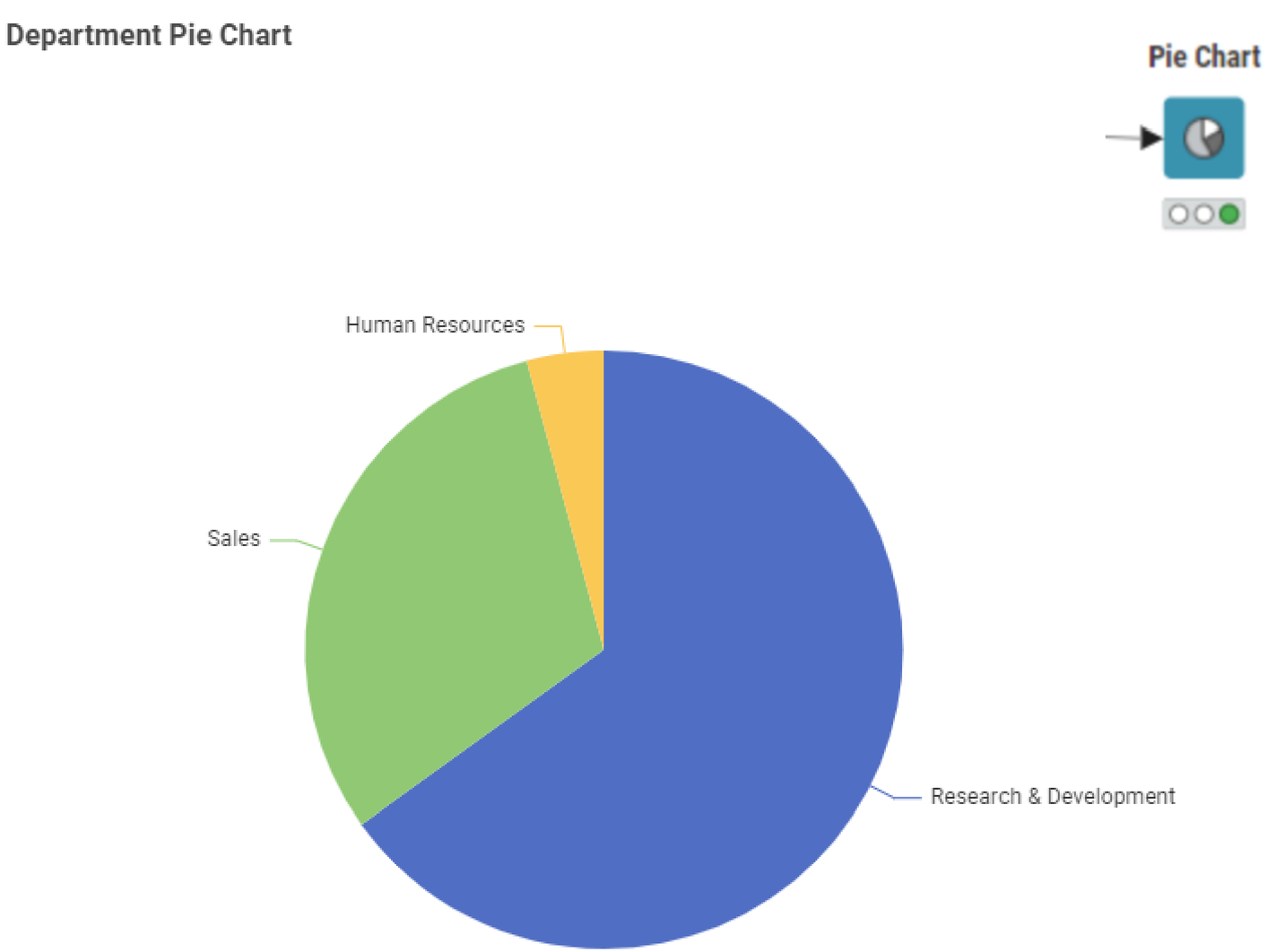
UNIVARIATE ANALYSIS

Department

Nature: Categorical, nominal

Categories: 1 = “HR”, 2 = “R&D”, 3 = “Sales”

Insights: the majority of employees work in Research & Development, followed by Sales, with the smallest proportion in Human Resources.



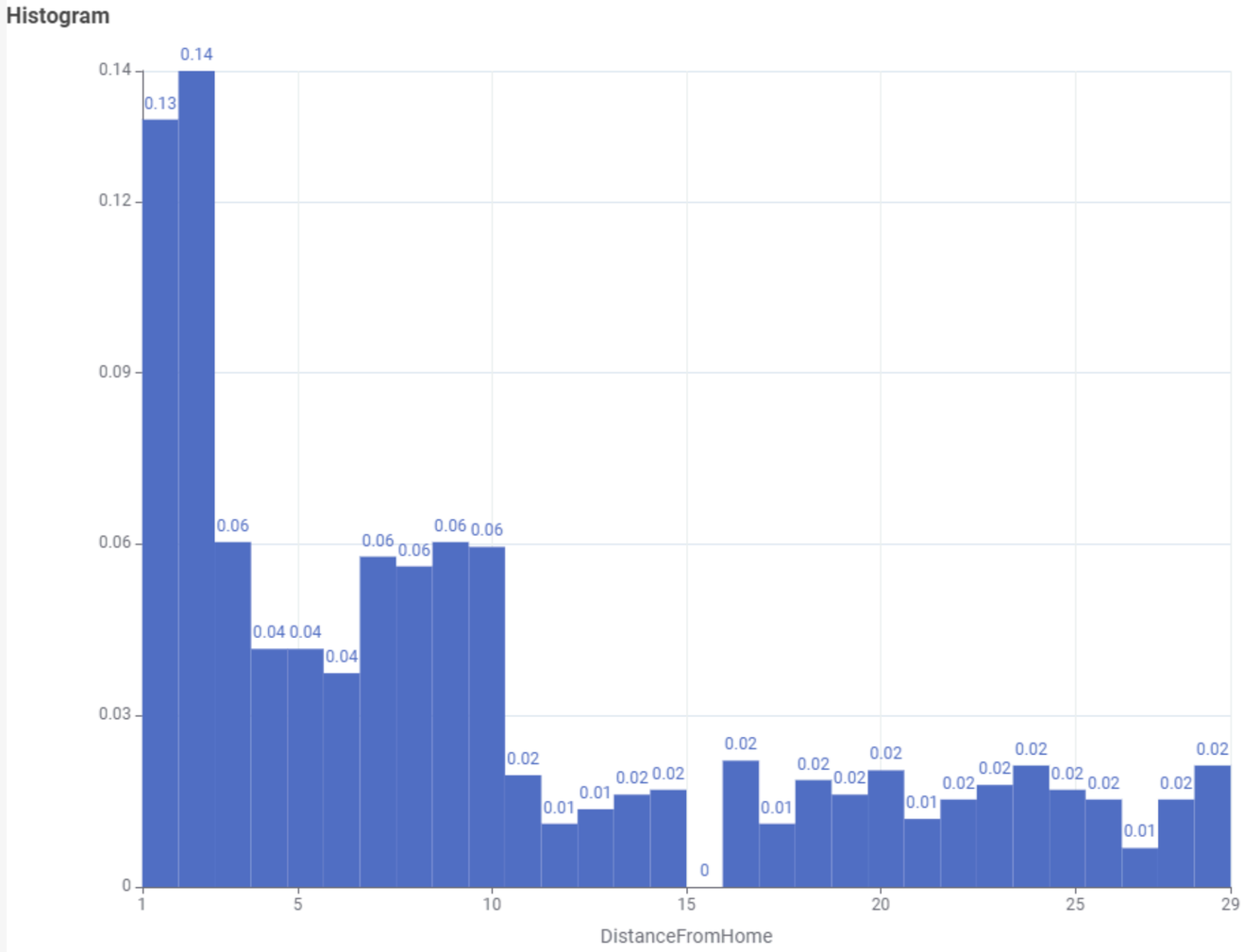
UNIVARIATE ANALYSIS

DistanceFromHome

Nature: Numerical, discrete

Range: 1 - 29

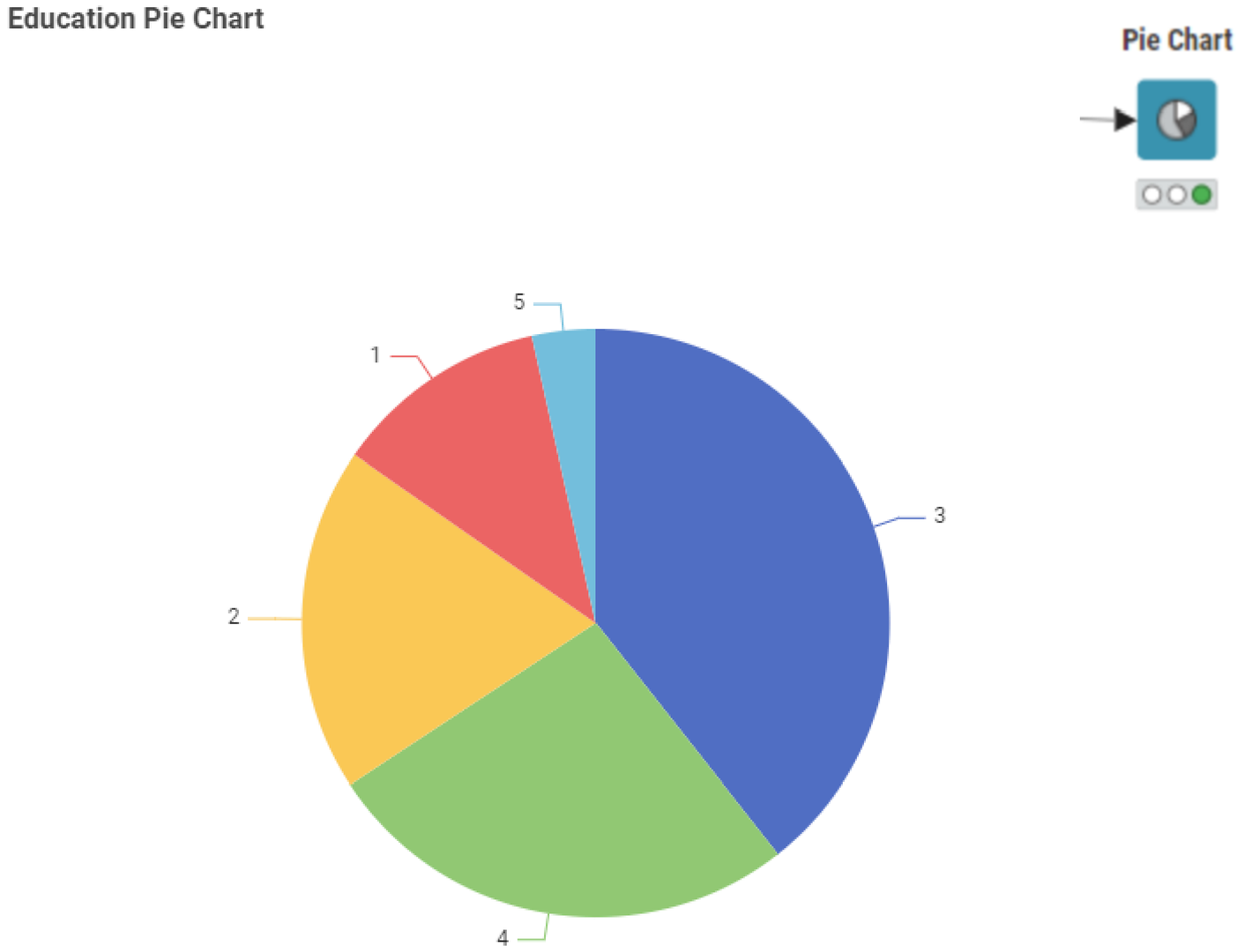
Insights: it is right-skewed, with the majority of employees living close to their workplace (within 1-5 units of distance) and fewer employees residing farther away.



UNIVARIATE ANALYSIS

Education

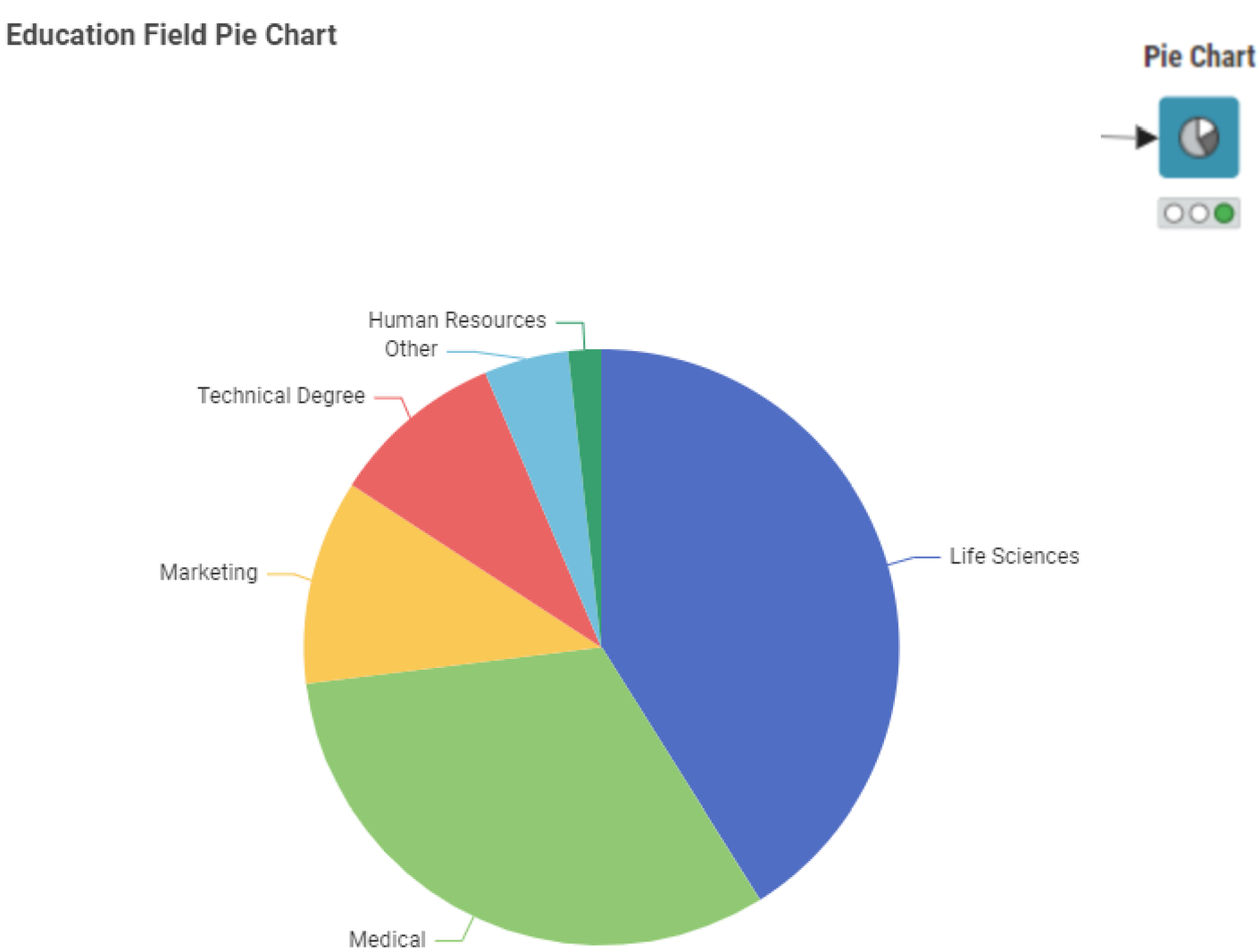
Nature: Categorical, ordinal
Categories: 1 - 5
Insights: the majority of employees fall into middle-level education categories (3 and 4), with fewer employees at the extremes of the education levels (1 and 5).



UNIVARIATE ANALYSIS

EducationField

Nature: Categorical, nominal
Categories: 1=HR, 2=Life Sciences, 3=Marketing, 4=Medical Sciences, 5=Others, 6= Technical
Insights: the majority of employees have a background in Life Sciences or Medical fields, with smaller proportions in Marketing, Technical Degrees, Human Resources, and Other fields.

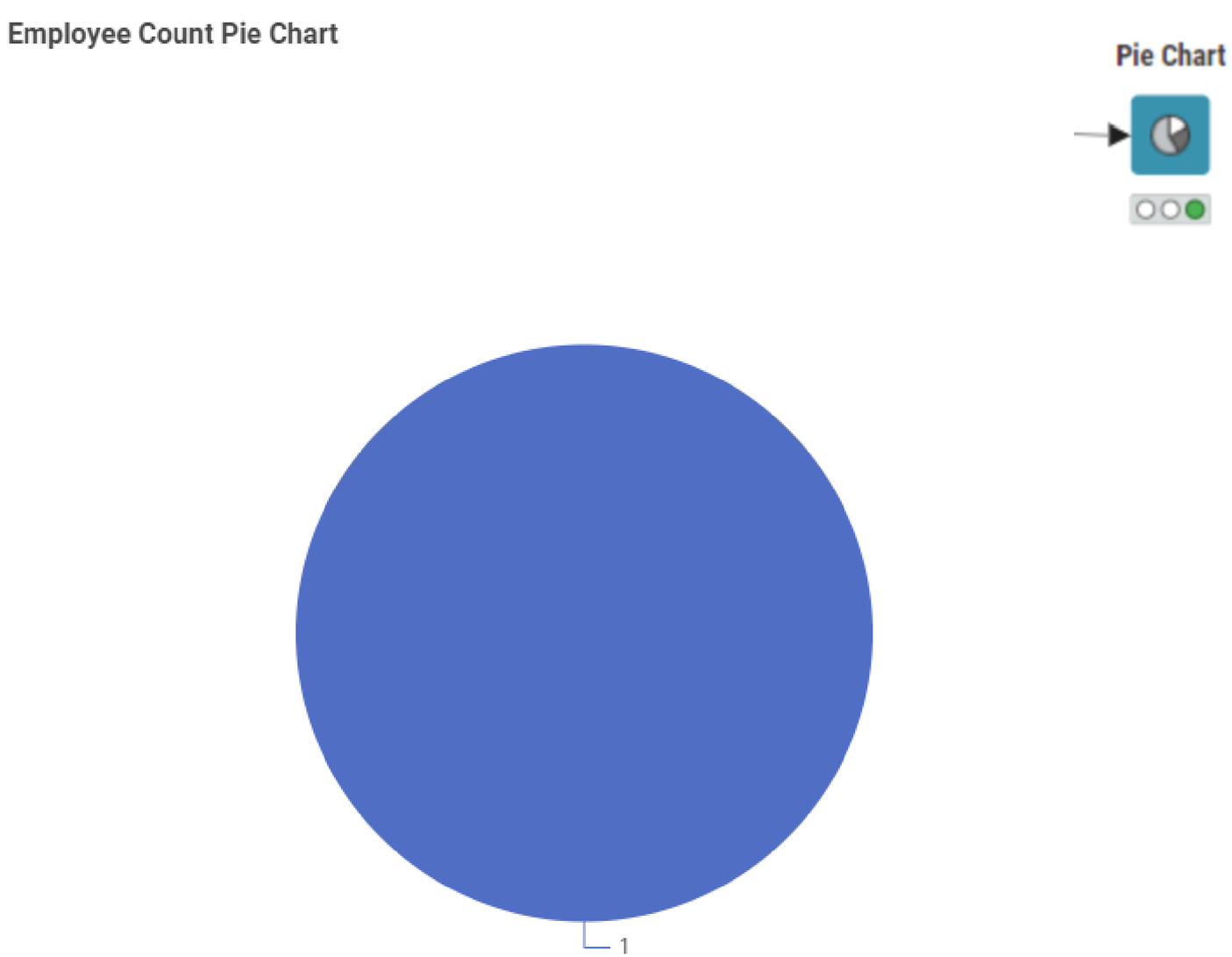


UNIVARIATE ANALYSIS

EmployeeCount

Nature: Numerical, discrete

Insights: All the entries are equal to 1 since each one is corresponding to a unique employee



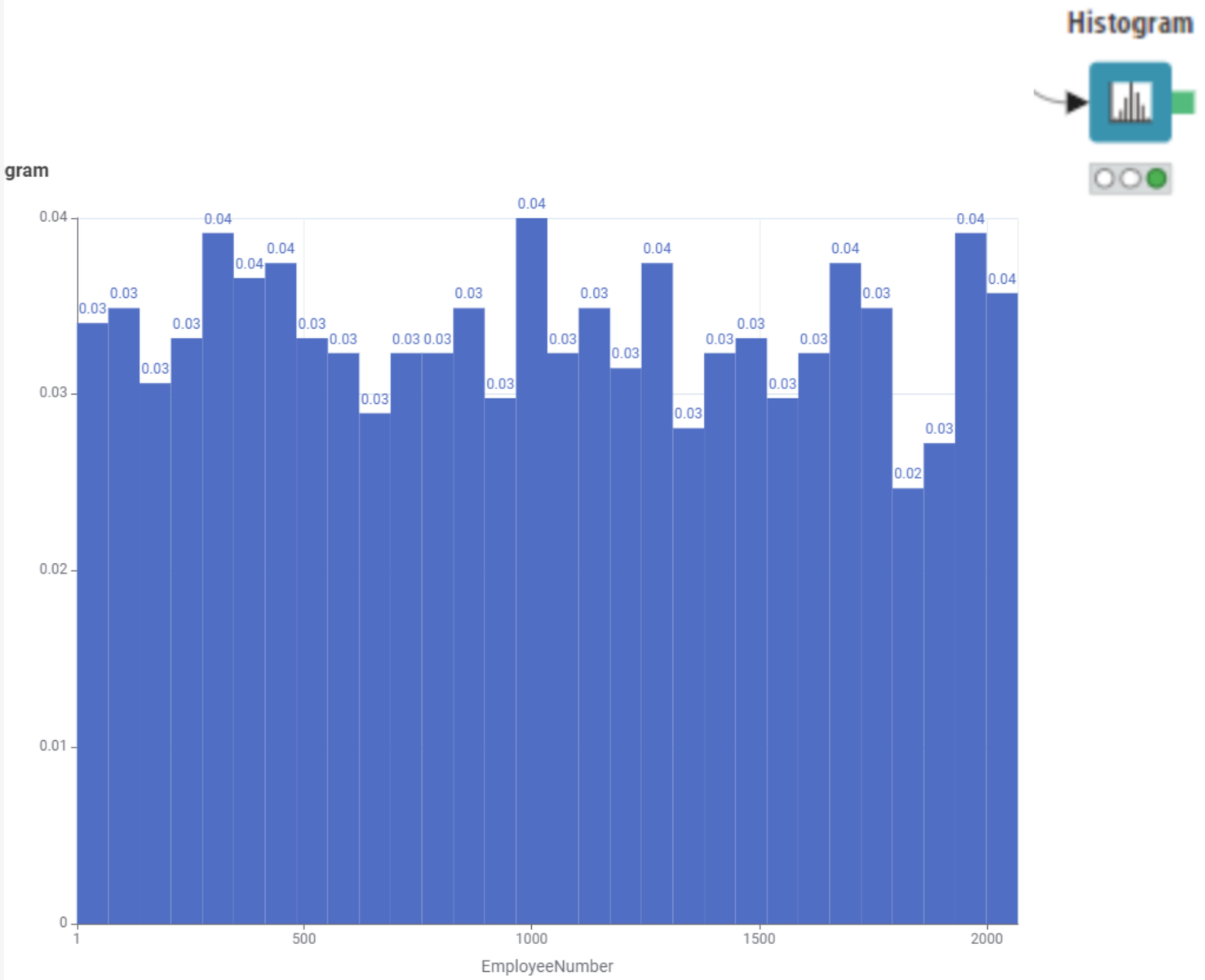
UNIVARIATE ANALYSIS

EmployeeNumber

Nature: Numerical, discrete

Range: 1 - 2068

Insights: it is uniformly distributed, representing unique identifiers for employees, with no concentration in any specific range.

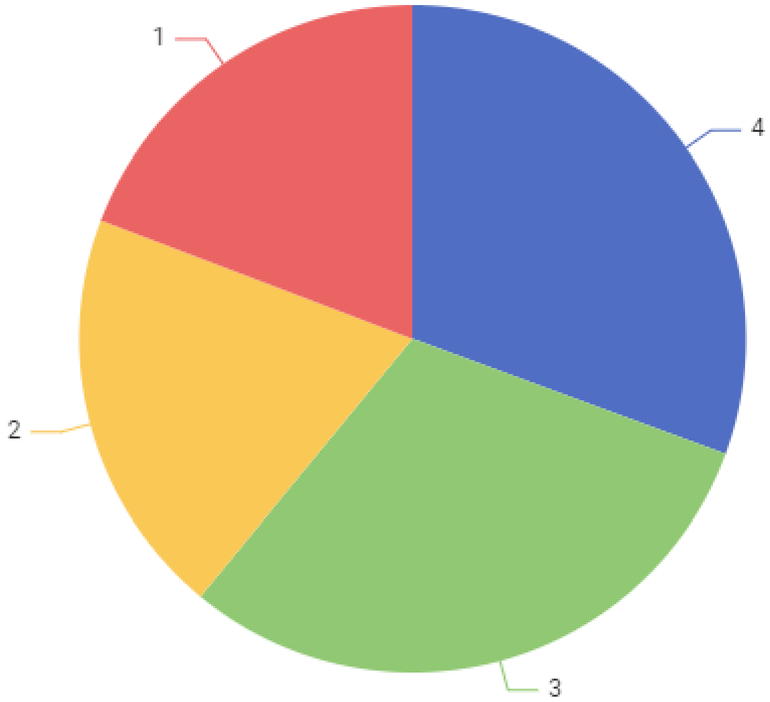


UNIVARIATE ANALYSIS

EnvironmentSatisfaction

Nature: Categorical, nominal
Description: Satisfaction With The Enviroment
Categories: 1 - 4
Insights: is fairly evenly distributed across the four satisfaction levels, with a slightly higher proportion of employees reporting the highest satisfaction level (4) and fewer reporting the lowest level (1).

EnvironmentSatisfaction Pie Chart



Pie Chart

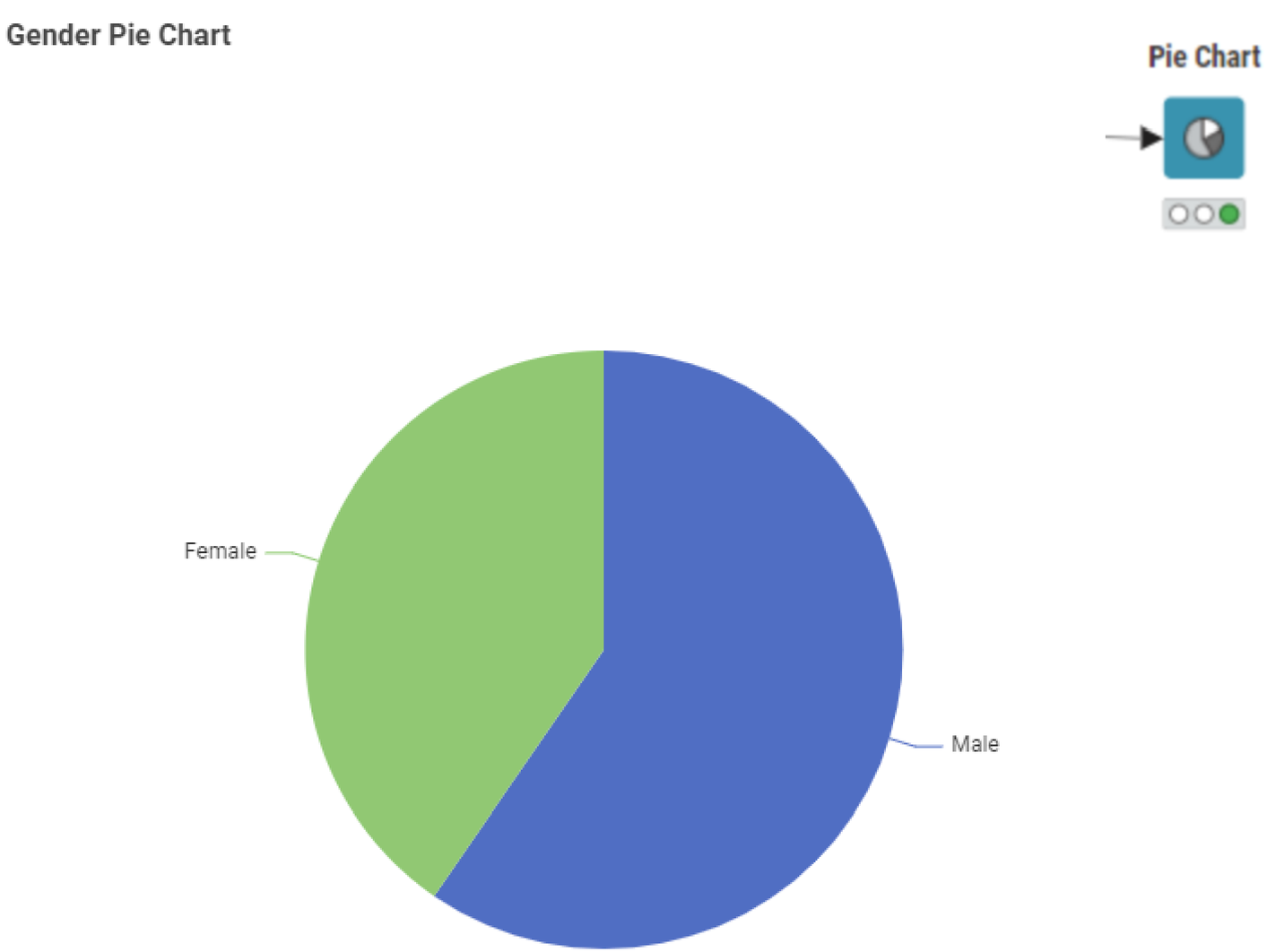
UNIVARIATE ANALYSIS

Gender

Nature: Categorical, nominal

Categories: 1=Female, 2=Male

Insights: it shows a higher proportion of males compared to females. This indicates a gender imbalance within the workforce.



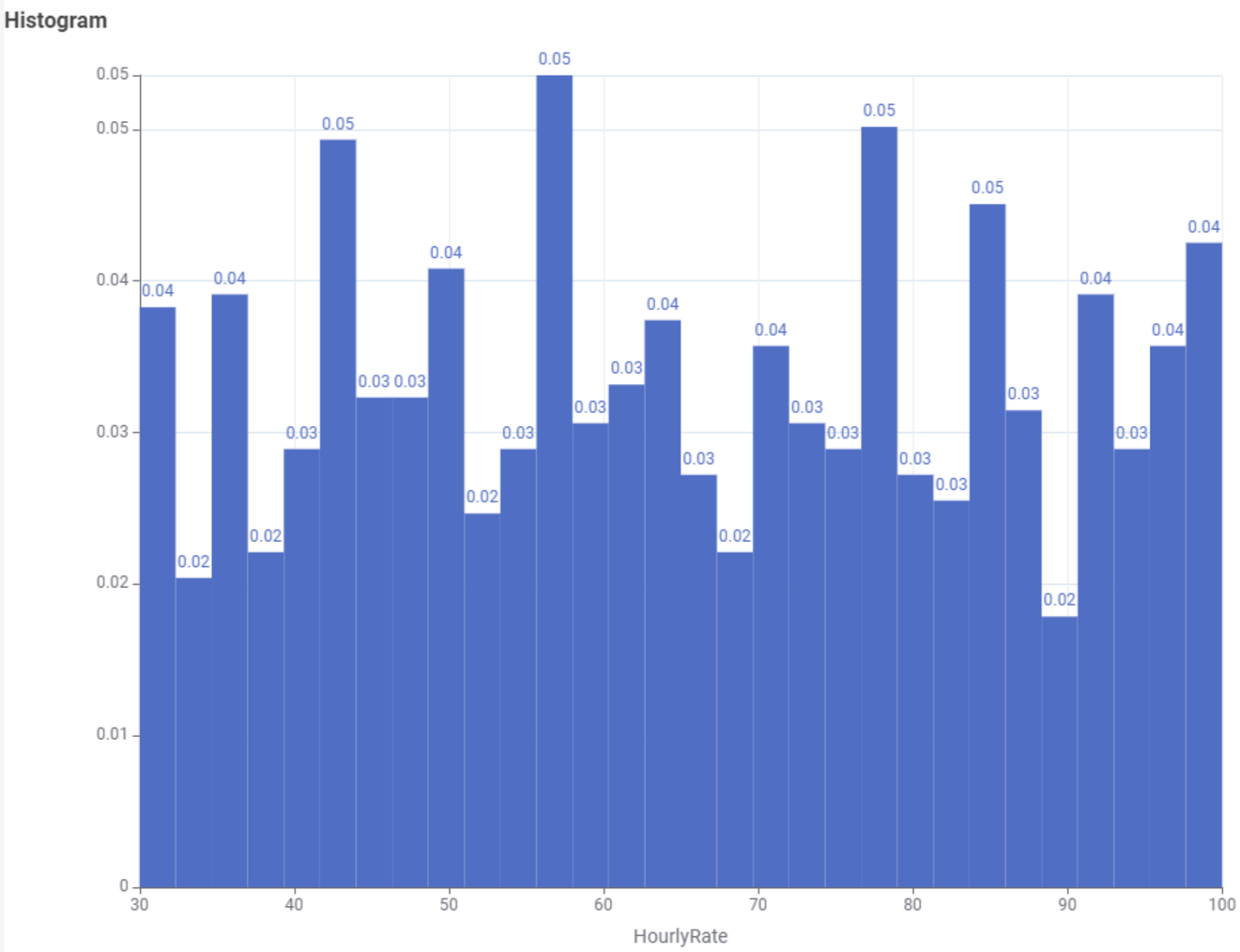
UNIVARIATE ANALYSIS

HourlyRate

Nature: Numerical, discrete

Range: 30 - 100

Insights: it is approximately uniformly distributed, indicating that employees' hourly rates are spread evenly across the range of values without any dominant concentration.



UNIVARIATE ANALYSIS

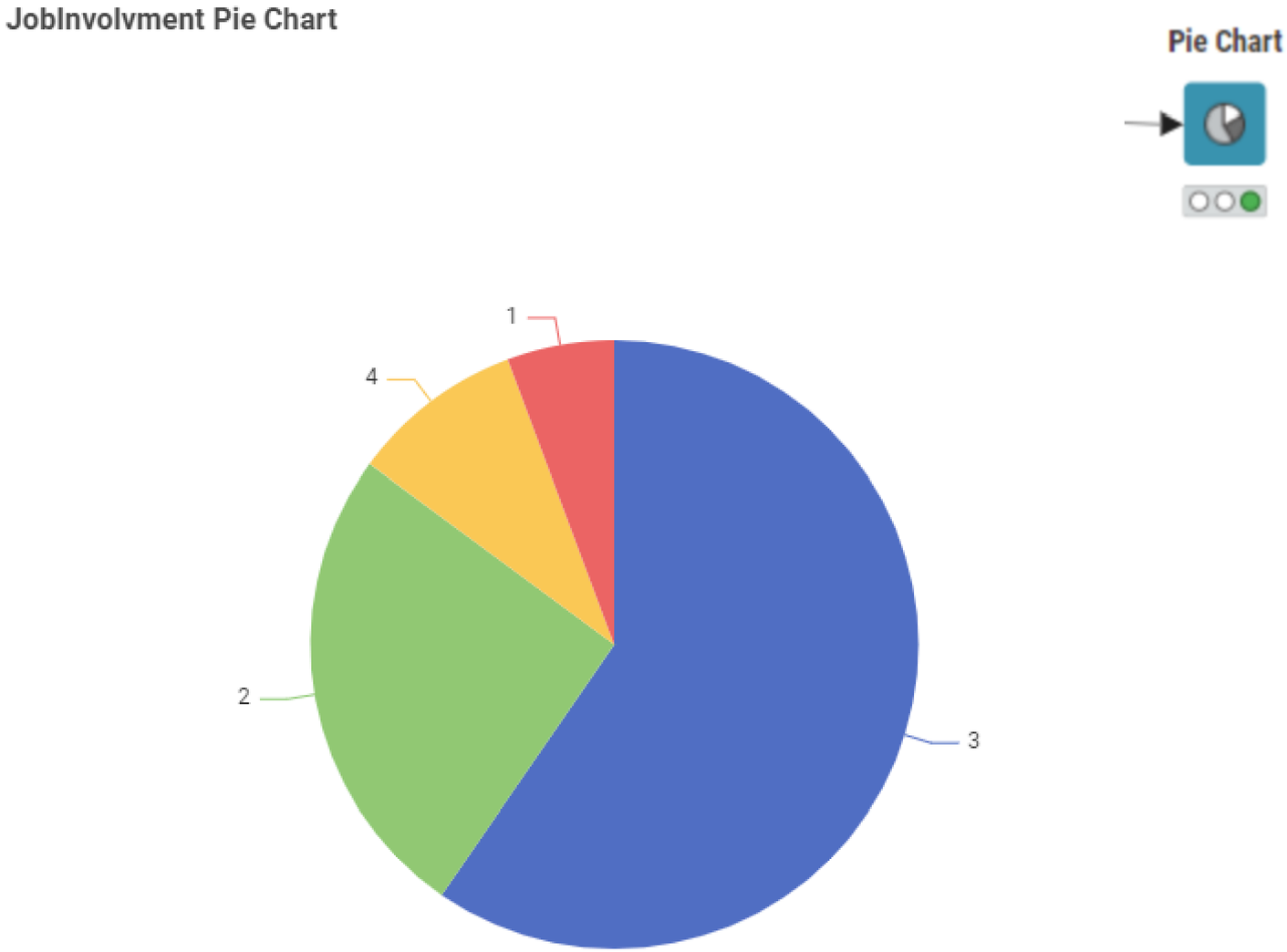
JobInvolvement

Nature: Categorical, ordinal

Description: Job Involvement

Categories: 1-4

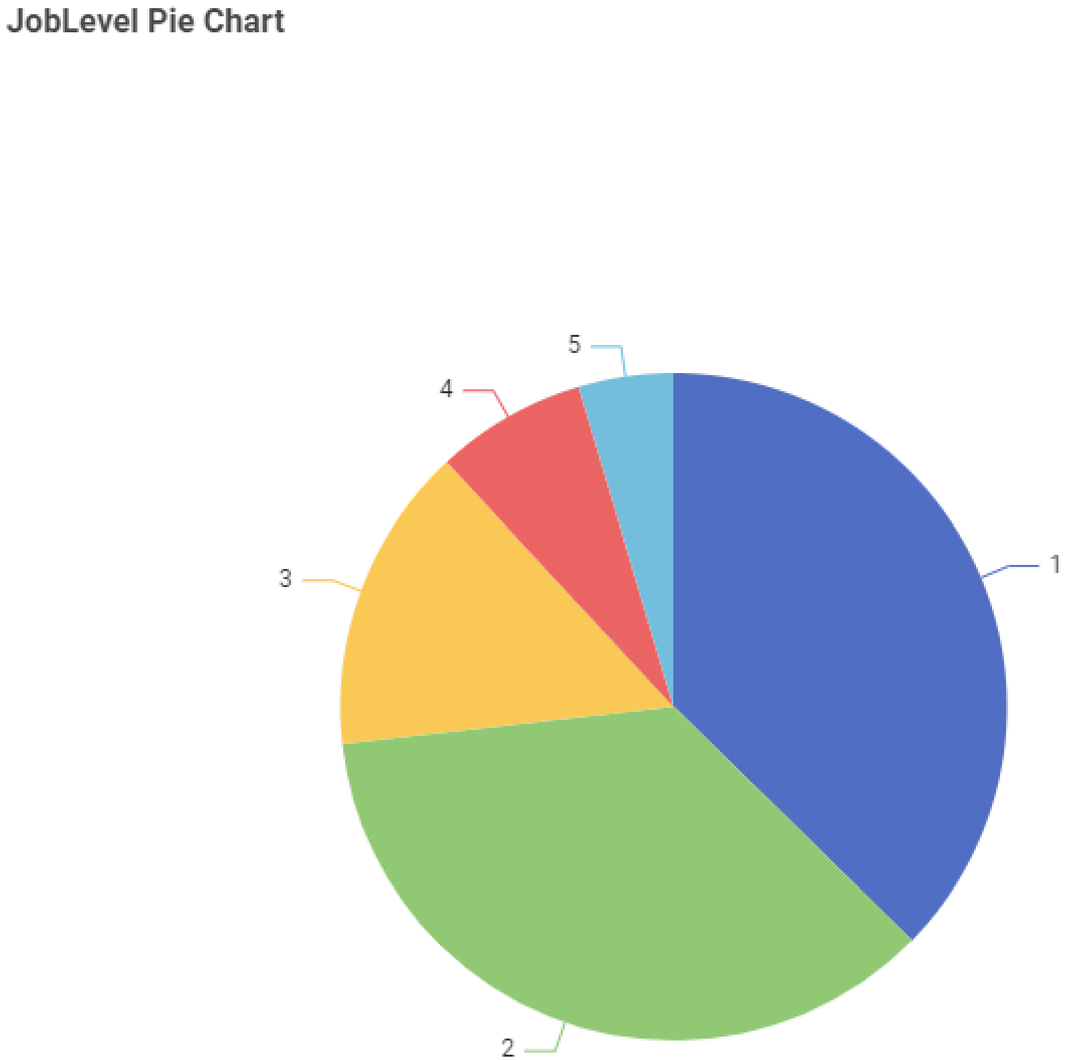
Insights: the majority of employees report moderate to high job involvement (levels 3 and 4), with fewer employees at the lower levels (1 and 2).



UNIVARIATE ANALYSIS

JobLevel

Nature: Categorical, nominal
Description: Level Of Job
Categories: 1-5
Insights: most employees are at the entry-level (1) or slightly higher levels (2 and 3), with relatively fewer employees in senior-level positions (4 and 5).



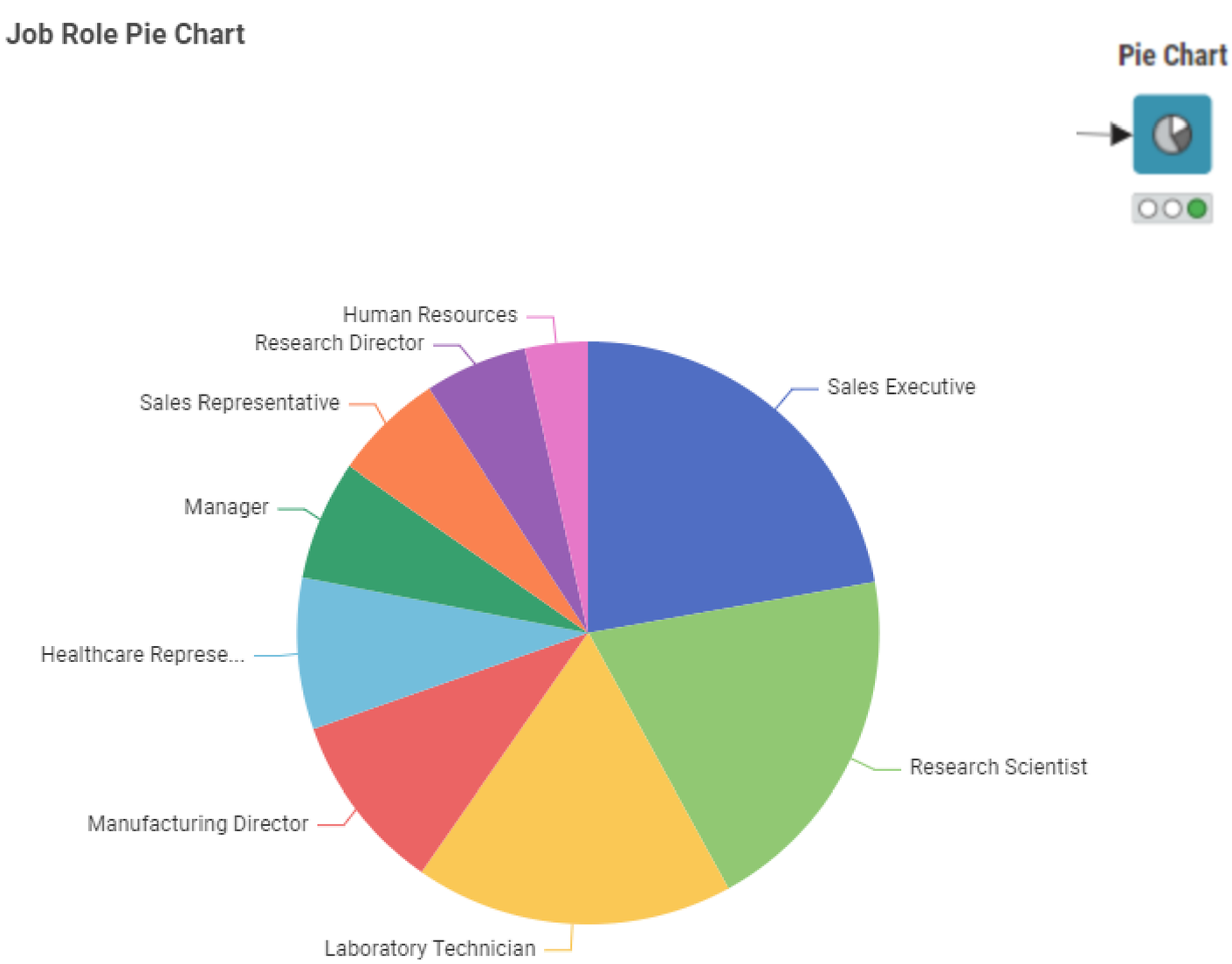
UNIVARIATE ANALYSIS

JobRole

Nature: Categorical, nominal

Categories: 1=Hc Rep, 2=Hr, 3=Lab Technician, 4=Manager, 5= Managing Director, 6= Reasearch Director, 7= Research Scientist, 8=Sales Executieve, 9= Sales Representative

Insights: the most common roles among employees are Sales Executive and Research Scientist, while roles like Research Director and Human Resources have smaller proportions.

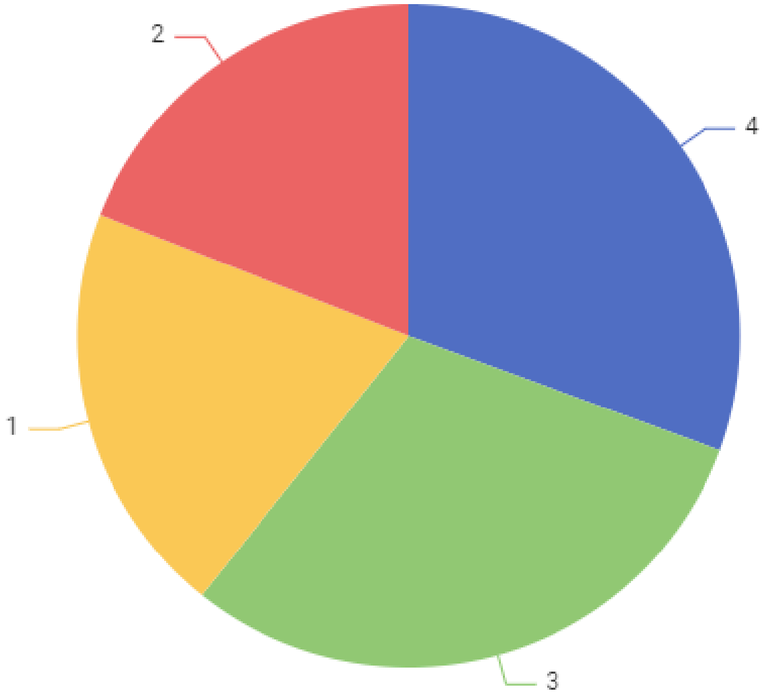


UNIVARIATE ANALYSIS

JobSatisfaction

Nature: Categorical, ordinal
Description: Satisfaction With The Job
Categories: 1-4
Insights: it is fairly evenly distributed, with a slight concentration of employees reporting high satisfaction levels (3 and 4) compared to lower levels (1 and 2).

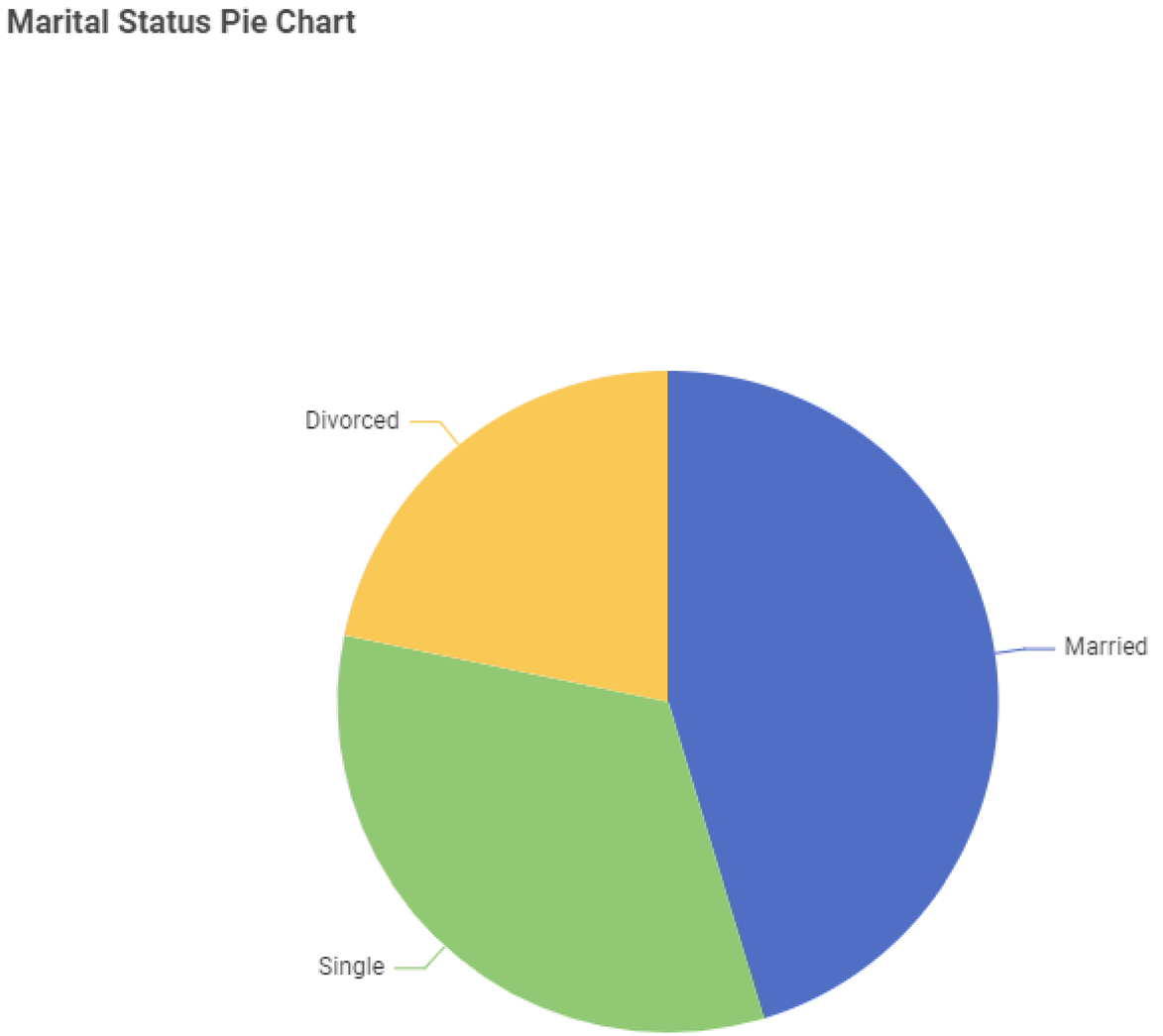
JobSatisfaction Pie Chart



UNIVARIATE ANALYSIS

MaritalStatus

Nature: Categorical, nominal
Categories:1 = Divorced, 2 =Married, 3 = Single
Insights: the majority of employees are married, followed by those who are single, with divorced employees forming the smallest group.



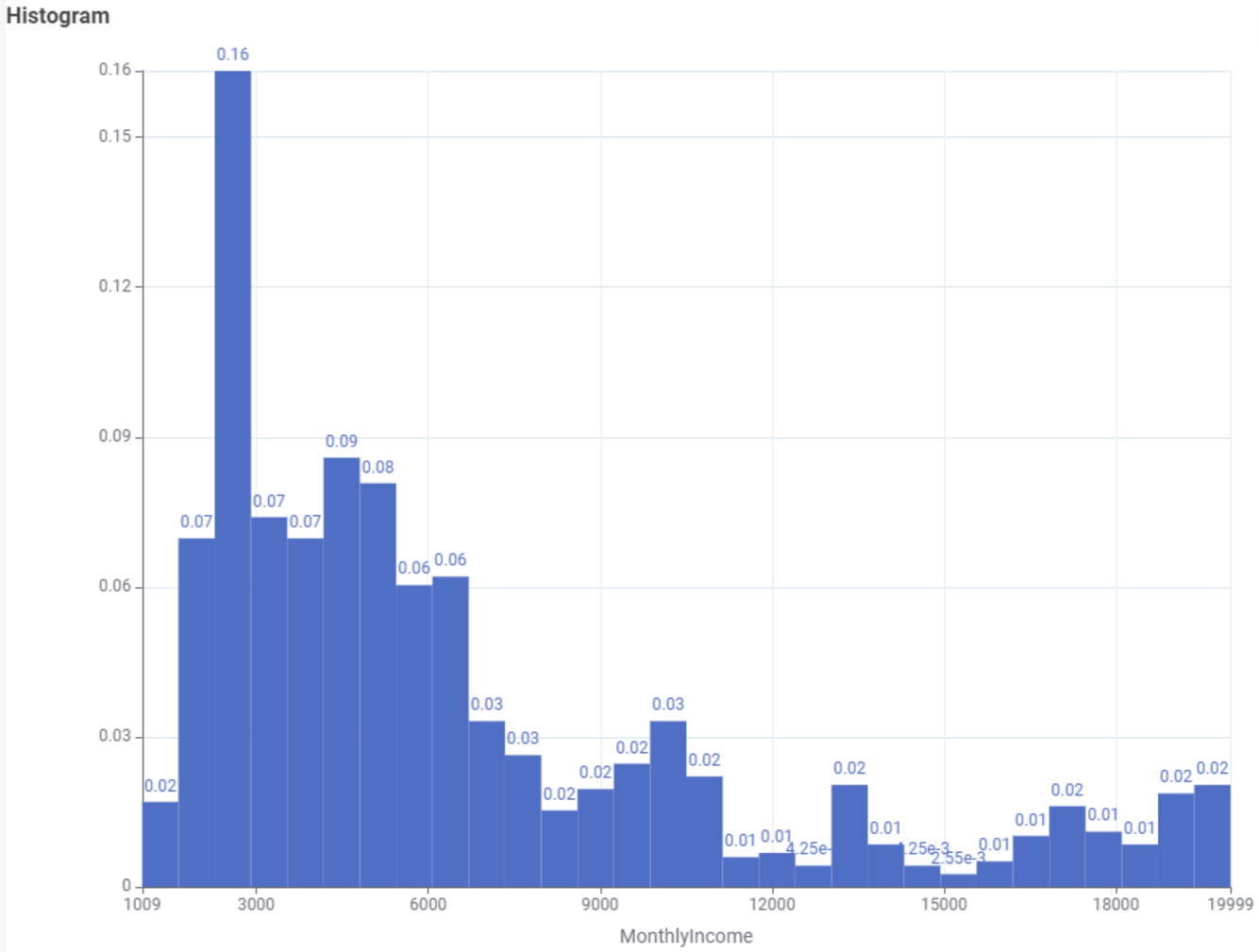
UNIVARIATE ANALYSIS

MonthlyIncome

Nature: Numerical, continous

Range: 1009 - 19999

Insights: it is highly right-skewed, with a significant proportion of employees earning around $3,000, while fewer employees earn higher salaries as income increases.



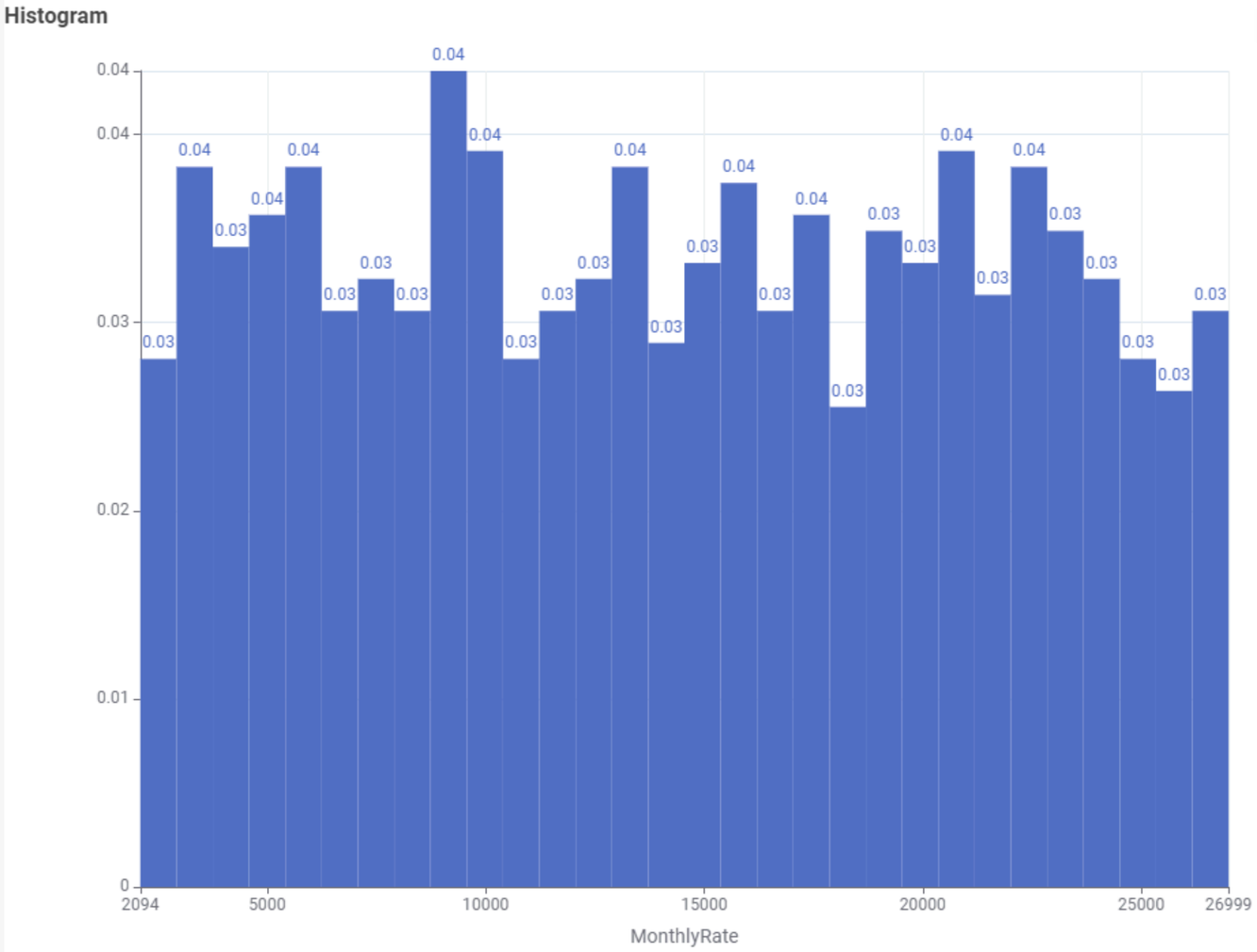
UNIVARIATE ANALYSIS

MonthlyRate

Nature: Numerical, discrete

Range: 2094 - 26999

Insights: it is approximately uniformly distributed, indicating that employees' monthly rates are evenly spread across the range without significant peaks or dips.



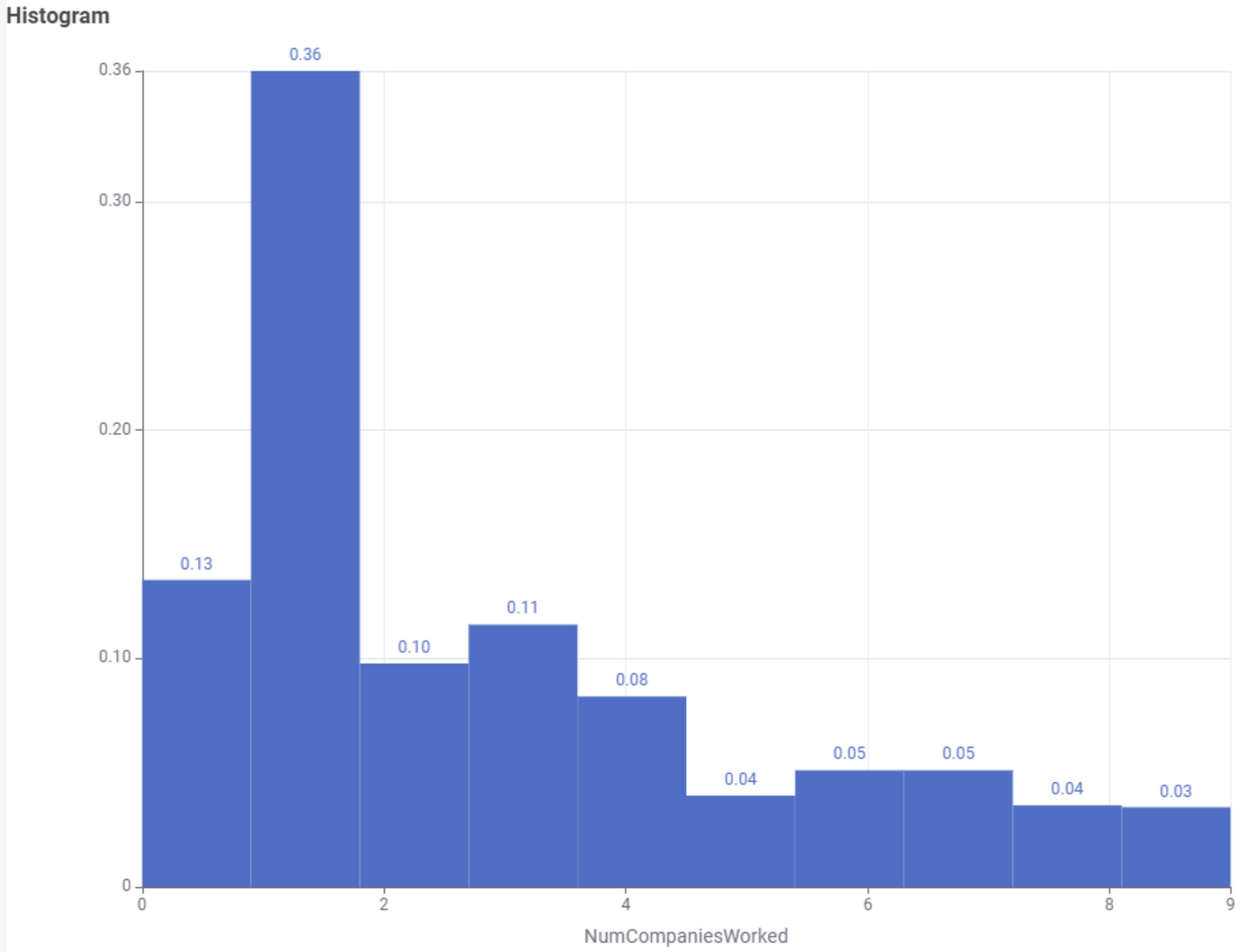
UNIVARIATE ANALYSIS

NumCompaniesWorked

Nature: Numerical, discrete

Range: 0 - 9

Insights: it is right-skewed, with the highest proportion of employees having worked for one company, followed by fewer employees as the number of companies worked increases.

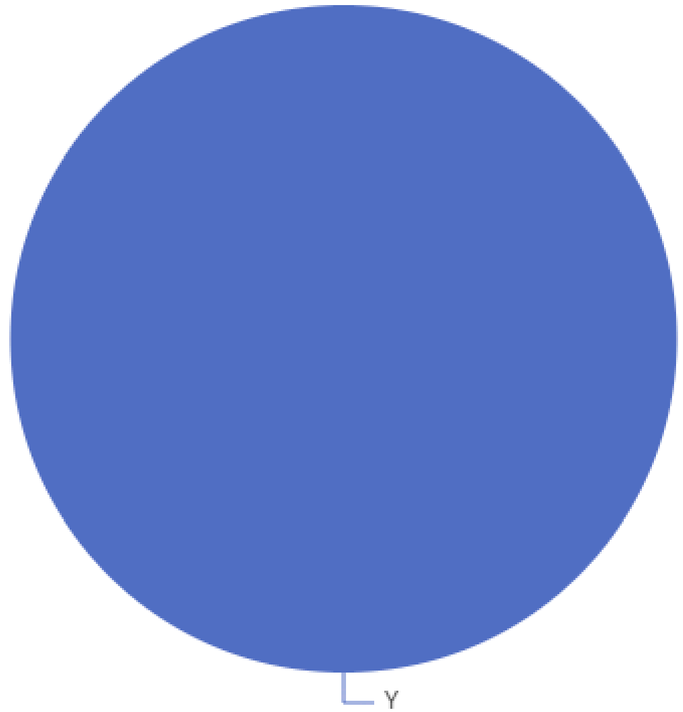


UNIVARIATE ANALYSIS

Over18

Nature: Categorical, nominal

Insights: all employees in the dataset are over 18 years old, as this is the only observed category.



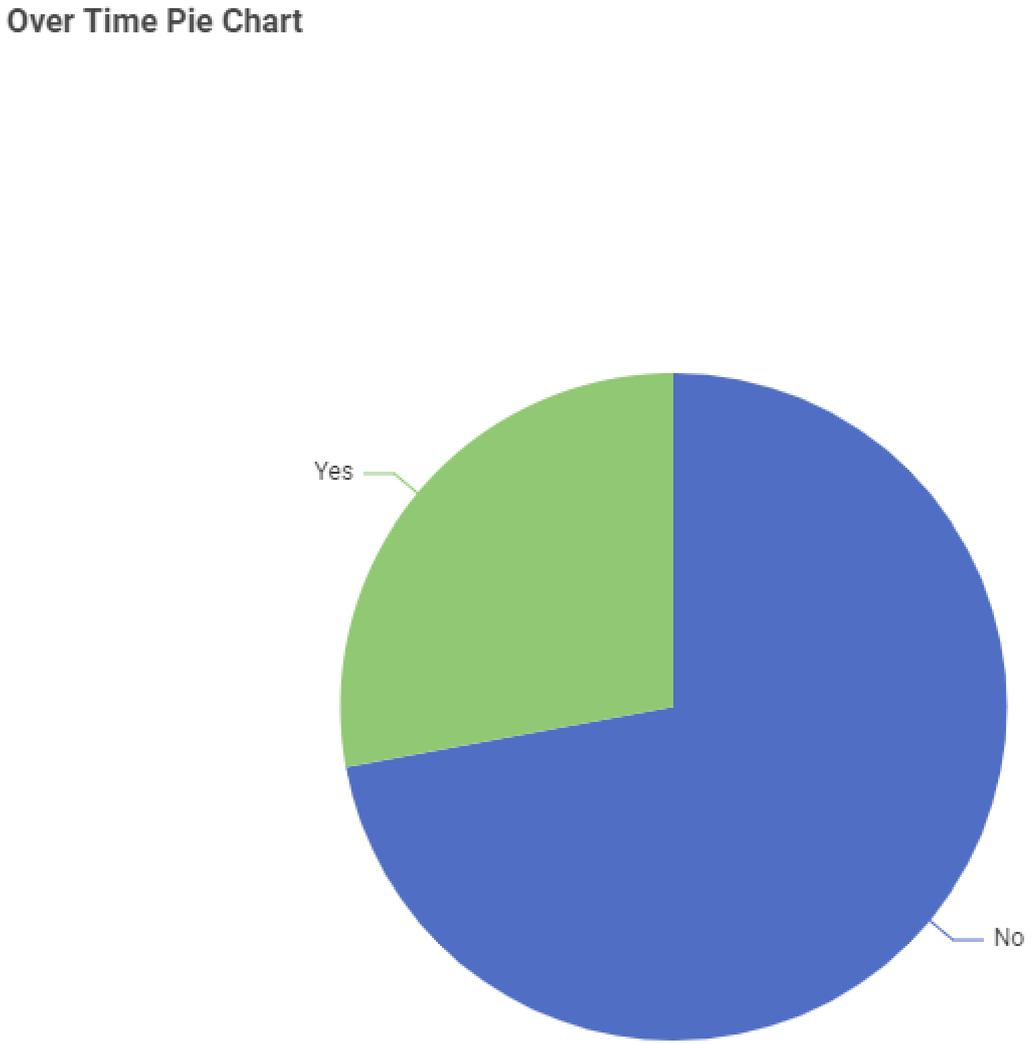
UNIVARIATE ANALYSIS

OverTime

Nature: Categorical, nominal

Categories: 1=No, 2=Yes

Insights: a majority of employees do not work overtime, while a smaller but significant proportion of employees report working overtime.



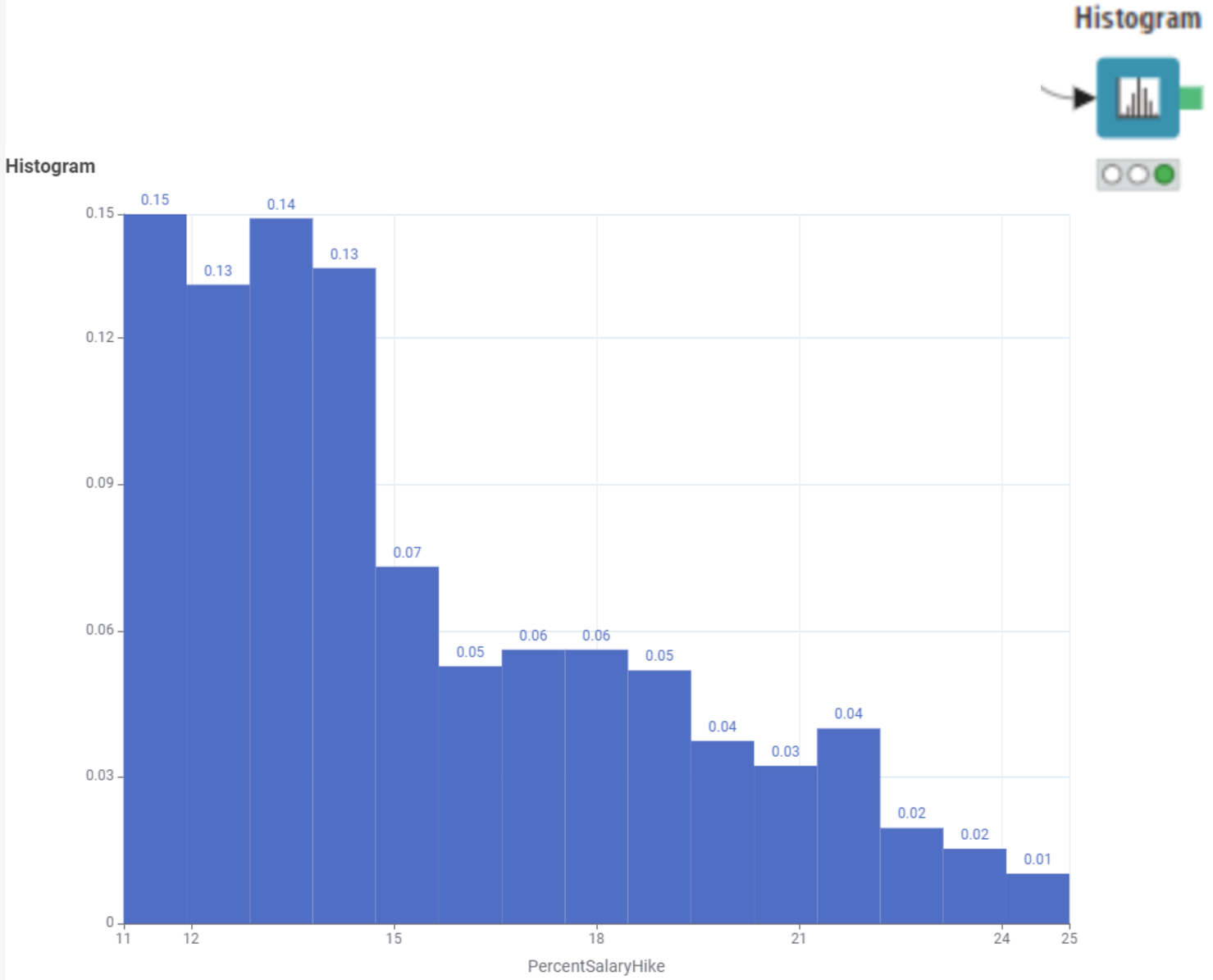
UNIVARIATE ANALYSIS

PercentSalaryHike

Nature: Numerical, discrete

Range: 11- 25

Insights: it shows a left-skewed distribution, with most employees receiving salary hikes between 11% and 15%, and fewer employees receiving higher percentage increases.

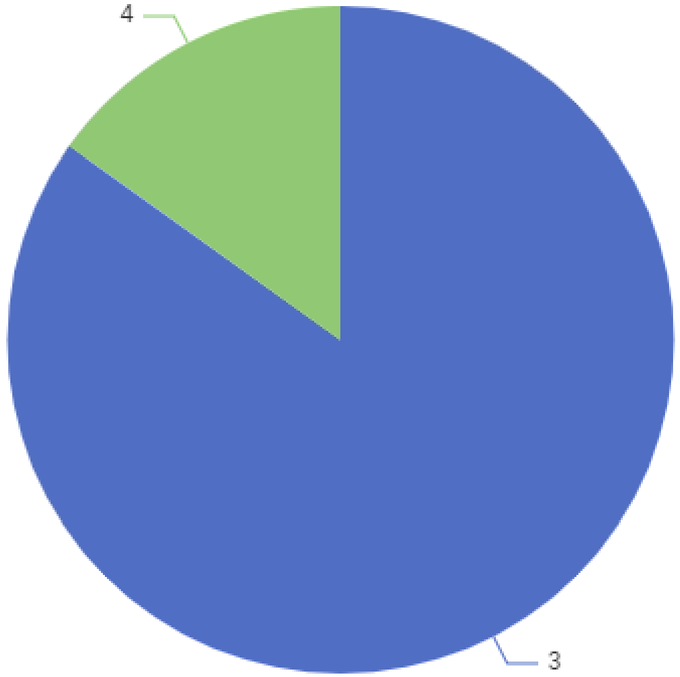


UNIVARIATE ANALYSIS

PerformanceRating

Nature: Categorical, ordinal
Description: Stock Options
Categories: 1-4
Insights: the majority of employees have a performance rating of 3, with a smaller proportion achieving the highest rating of 4.

PerformanceRatings Pie Chart

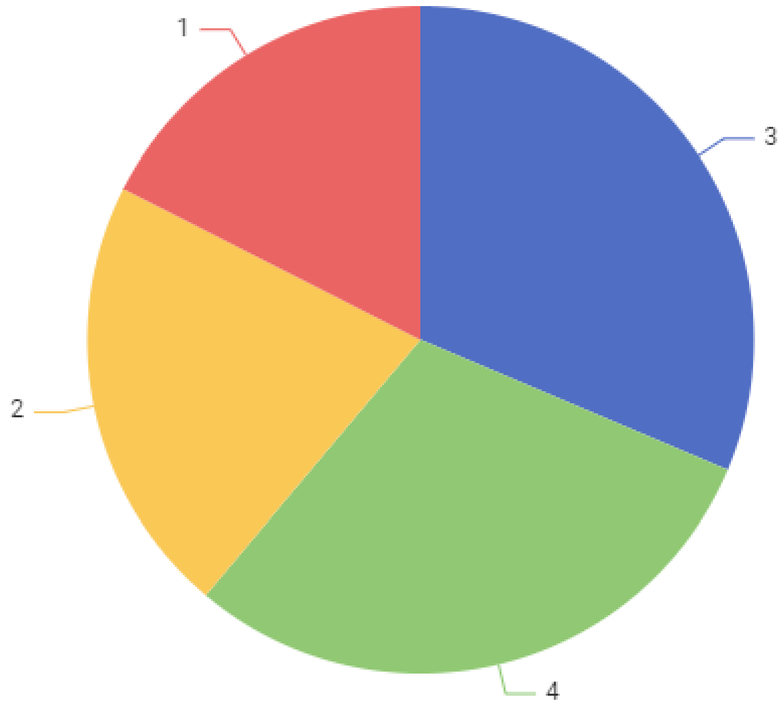


UNIVARIATE ANALYSIS

RelationshipSatisfaction

Nature: Categorical, ordinal
Description: Relations satisfaction
Categories: 1-4
Insights: it shows a relatively balanced distribution across all levels, with slightly more employees reporting higher satisfaction levels (3 and 4) compared to lower levels (1 and 2).

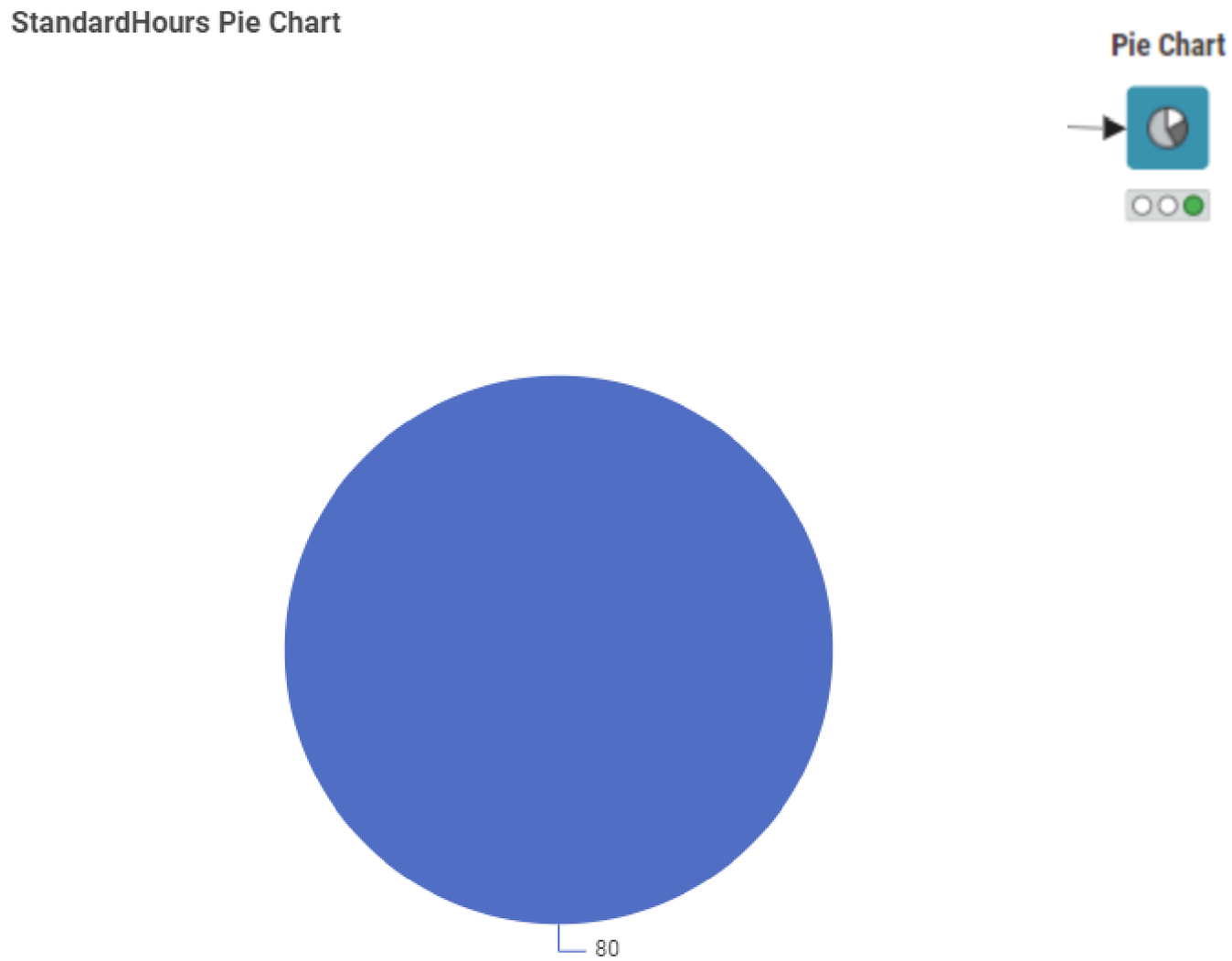
RelationshipSatisfaction Pie Chart



UNIVARIATE ANALYSIS

StandardHours

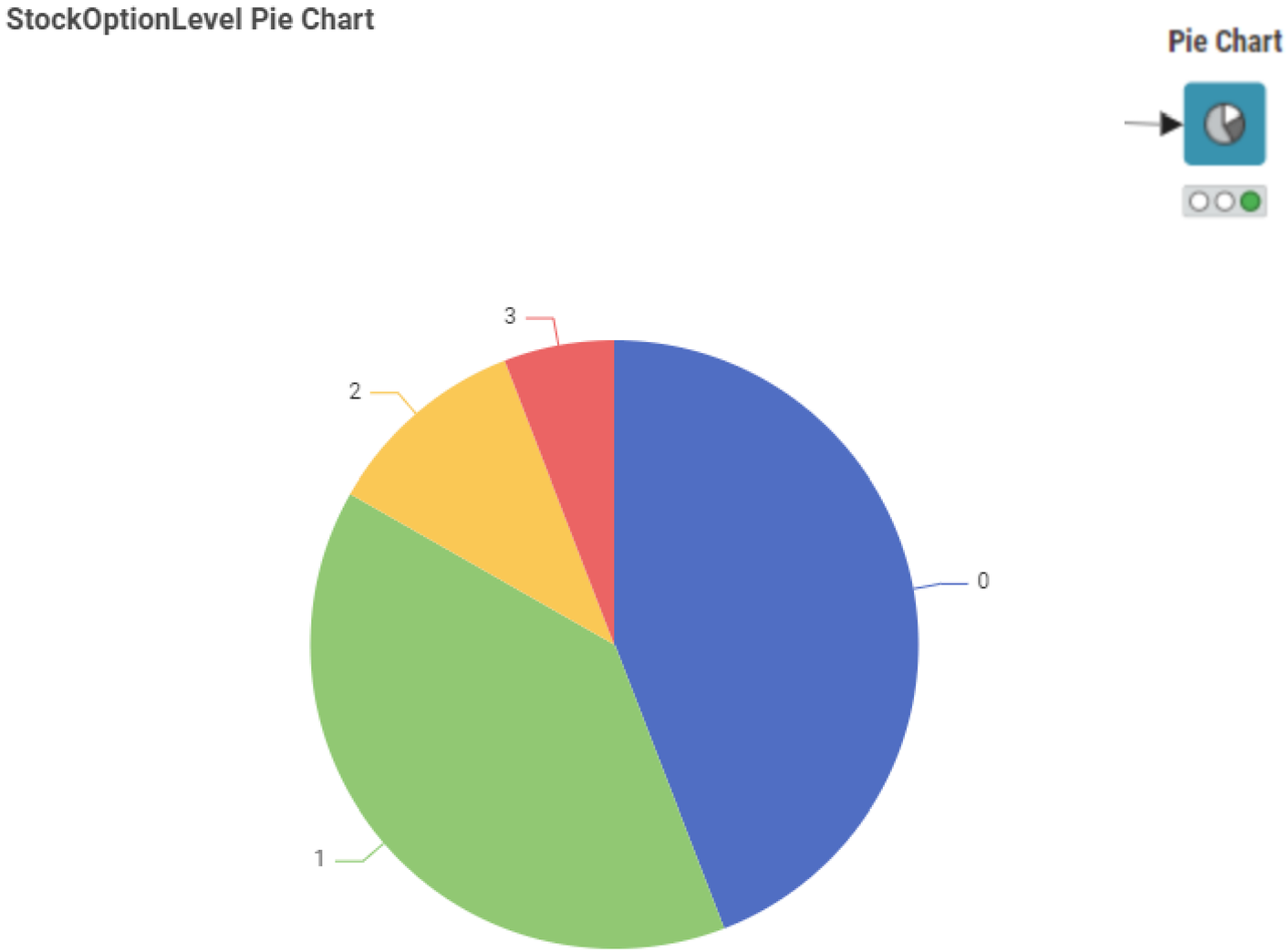
Nature: Numerical, discrete
Insights: it shows no variability, with all employees having a value of 80, indicating that standard hours are consistent across the dataset.



UNIVARIATE ANALYSIS

StockOptionLevel

Nature: Categorical, ordinal
Description: Stock Options
Categories: 0-3
Insights: most employees have stock option levels of 0 or 1, with fewer employees having higher levels (2 and 3).



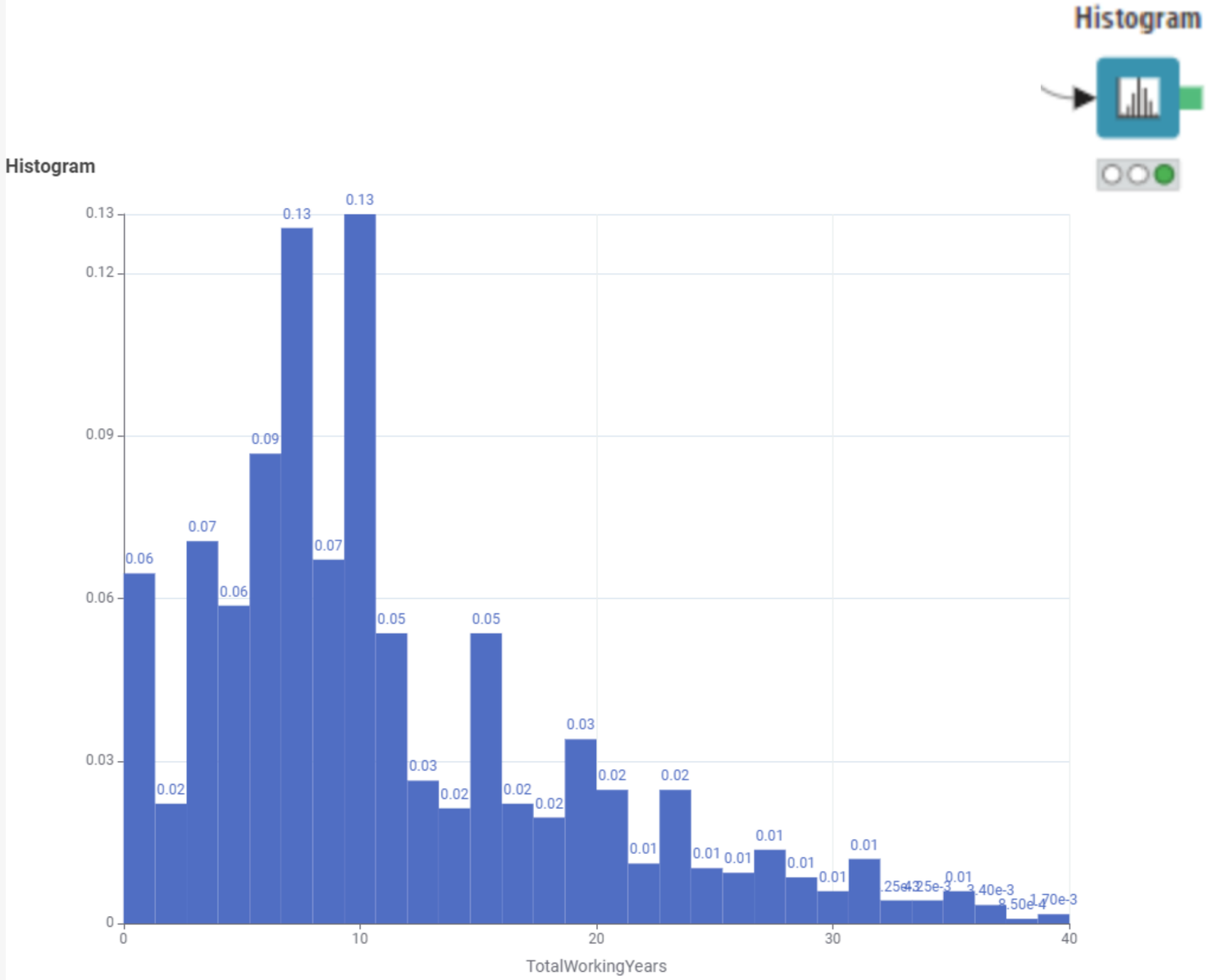
UNIVARIATE ANALYSIS

TotalWorkingYears

Nature: Numerical, discrete

Range: 0 - 40

Insights: it is right-skewed, with most employees having less than 15 years of total working experience, and fewer employees having significantly longer careers.



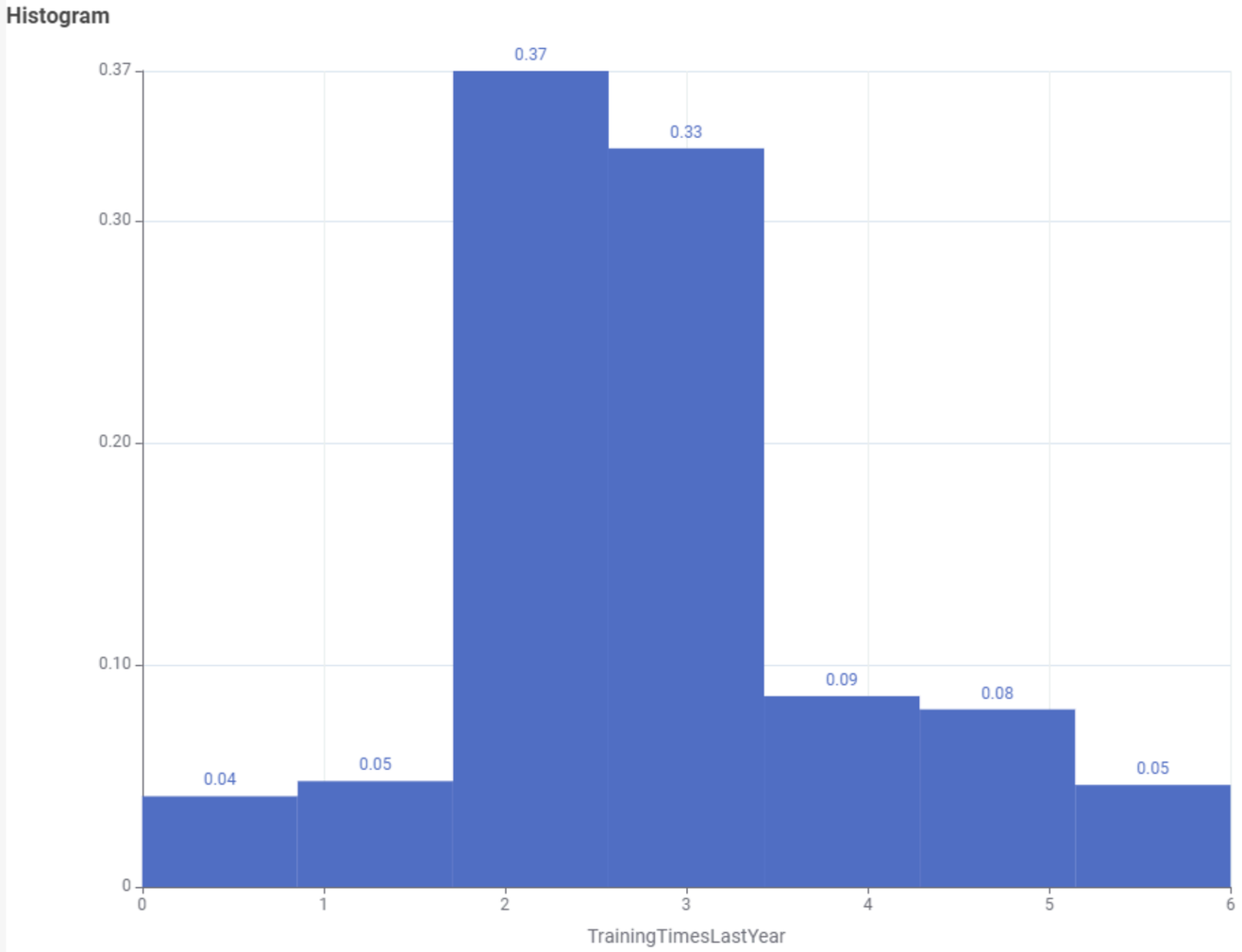
UNIVARIATE ANALYSIS

TrainingTimesLastYear

Nature: Numerical, discrete

Range: 0 - 6

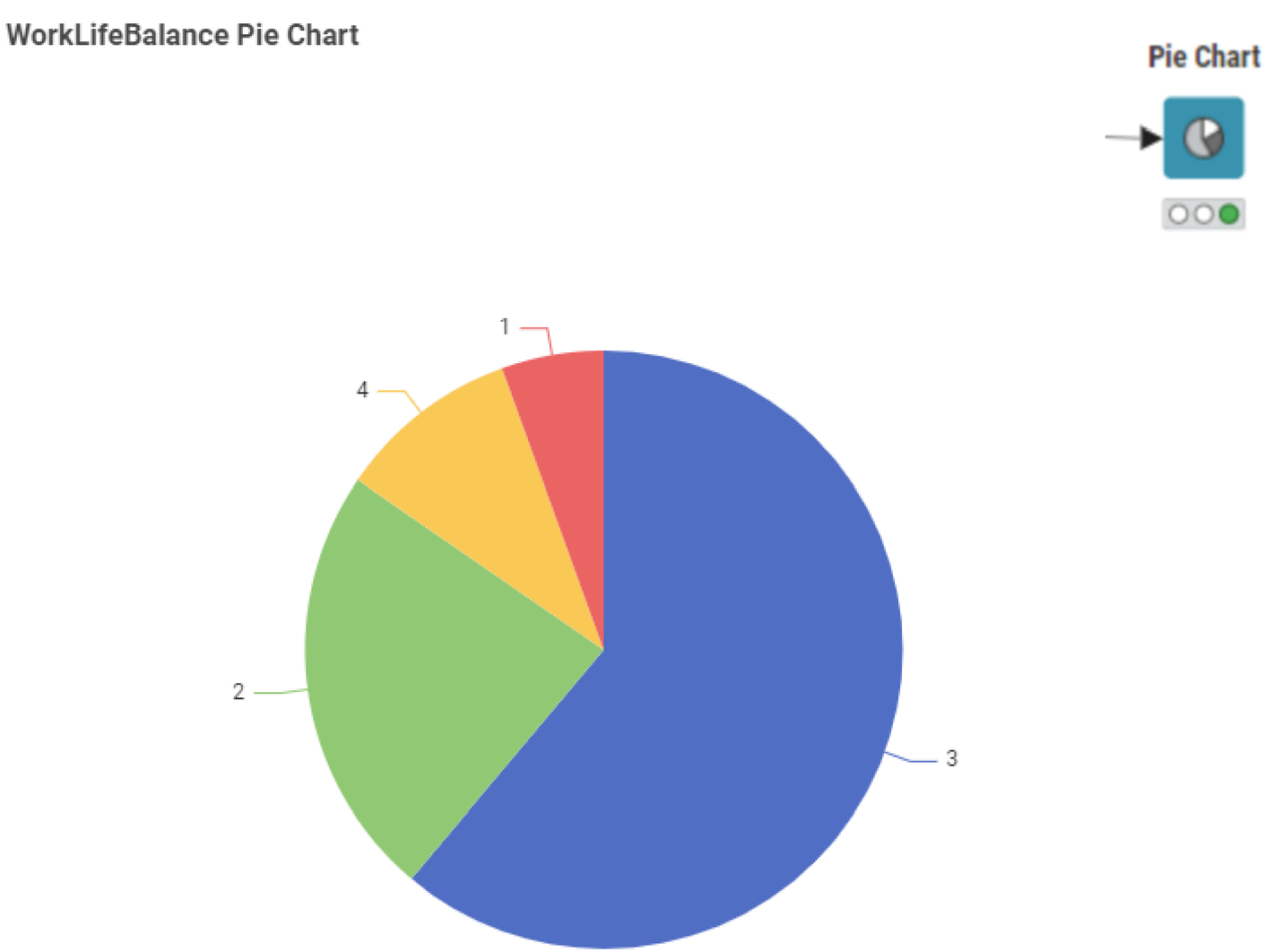
Insights: is centered around 2 and 3 sessions, indicating that most employees received moderate training, while fewer employees had either very low or very high training sessions.



UNIVARIATE ANALYSIS

WorkLifeBalance

Nature: Categorical, ordinal
Description: Time Spent Bewtween Work And Outside
Categories: 1-4
Insights: the majority of employees report a balance level of 3, indicating a relatively satisfactory work-life balance, while fewer employees report the lowest level (1) or the highest level (4).



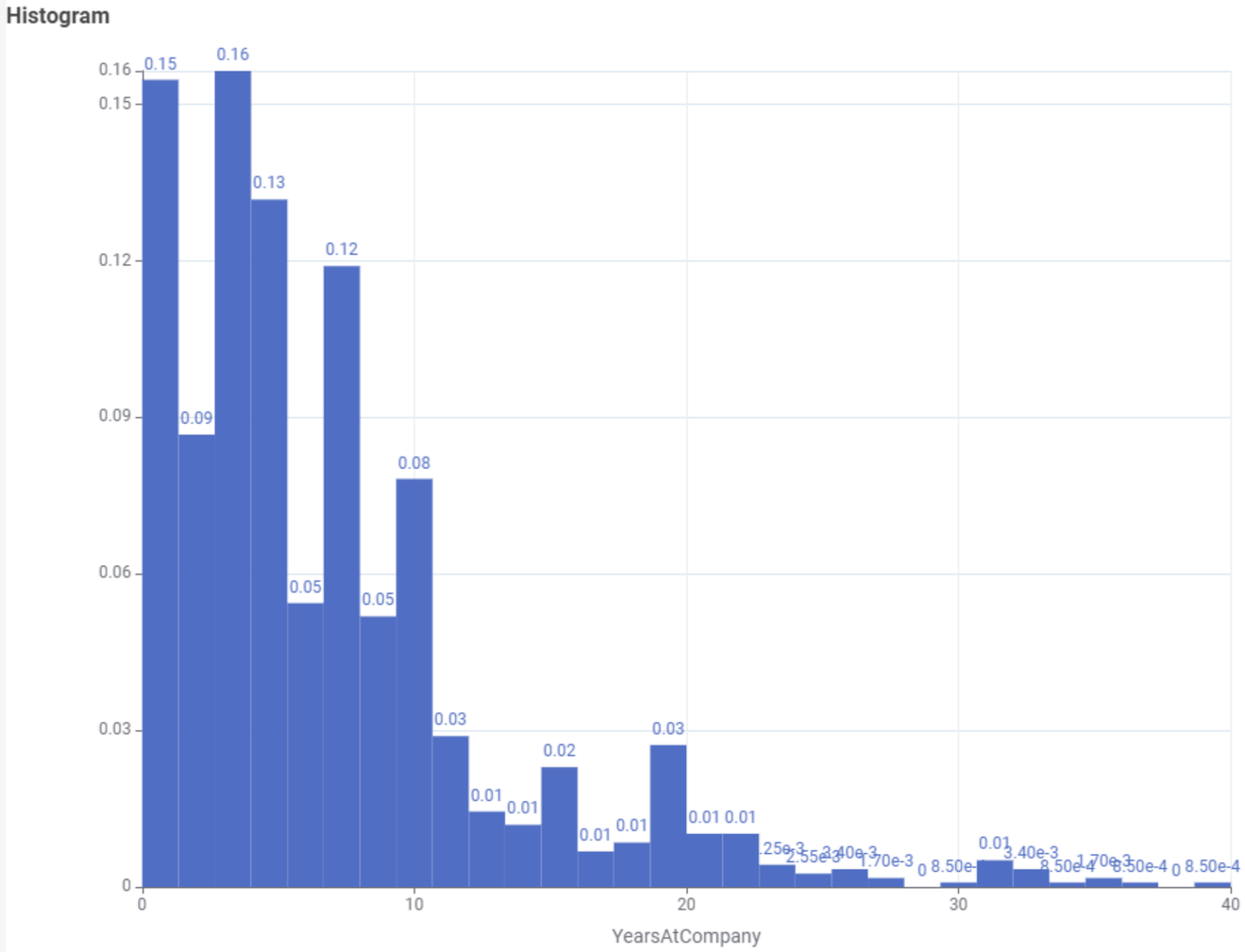
UNIVARIATE ANALYSIS

YearsAtCompany

Nature: Numerical, discrete

Range: 0 - 40

Insights: it is heavily right-skewed, with most employees having a short tenure of less than 5 years, and very few employees staying for 20 years or more.



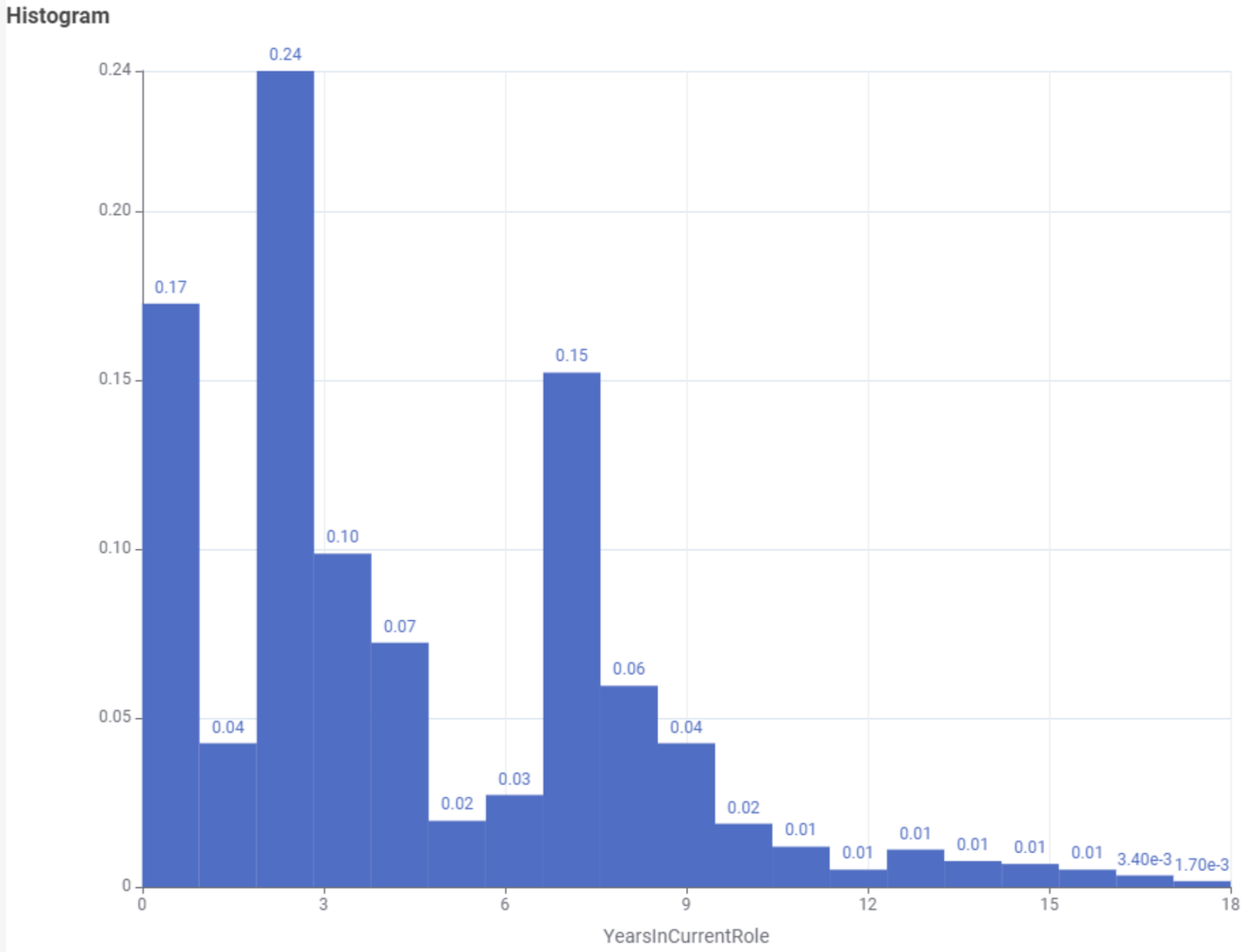
UNIVARIATE ANALYSIS

YearsInCurrentRole

Nature: Numerical, discrete

Range: 0 - 18

Insights: it is right-skewed, with the majority of employees having been in their current role for less than 5 years, and a sharp decline in frequency for employees with longer durations.



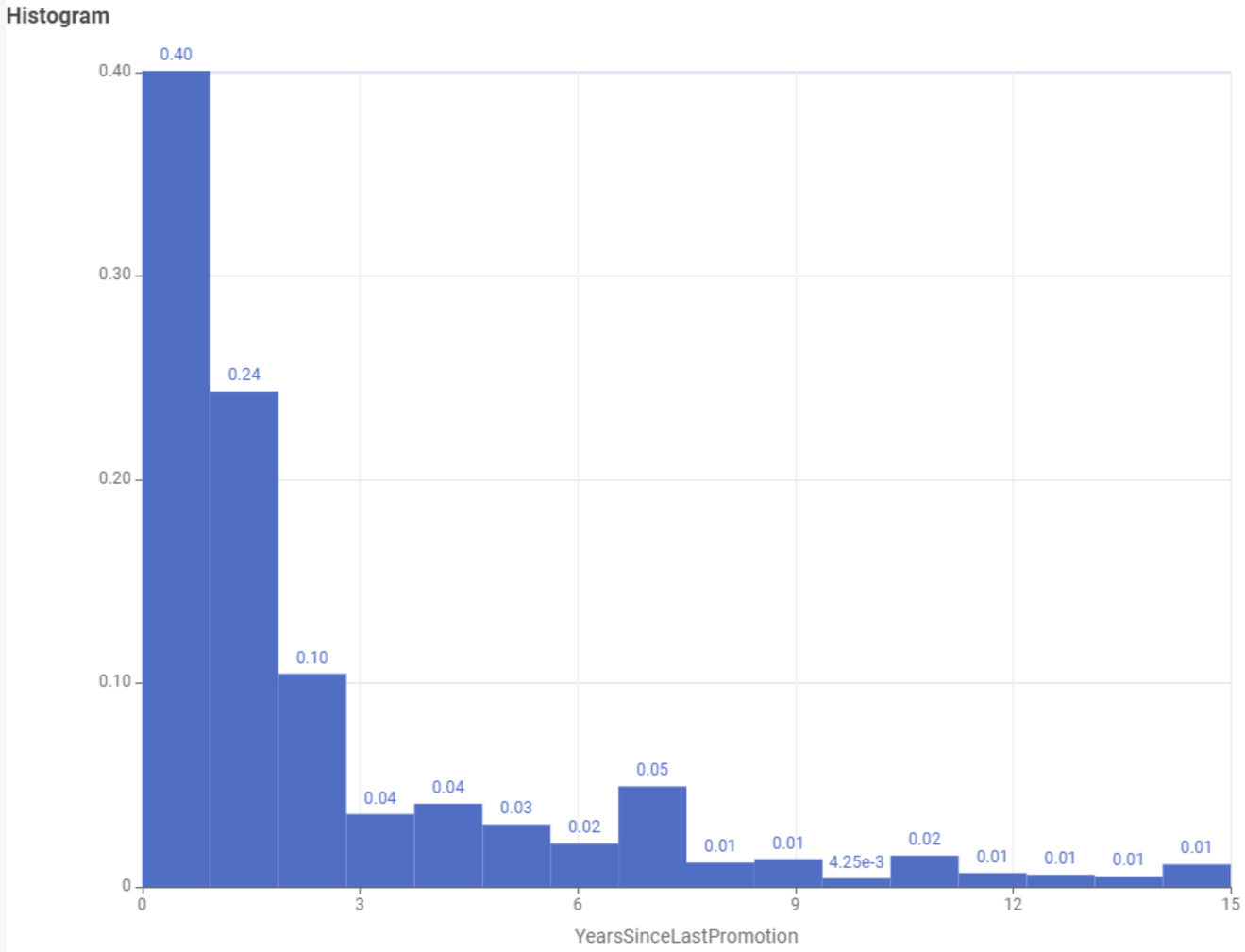
UNIVARIATE ANALYSIS

YearsSinceLastPromotion

Nature: Numerical, discrete

Range: 0 - 15

Insights: it is heavily right-skewed, with a large proportion of employees having been promoted within the last year, and fewer employees experiencing longer gaps since their last promotion.



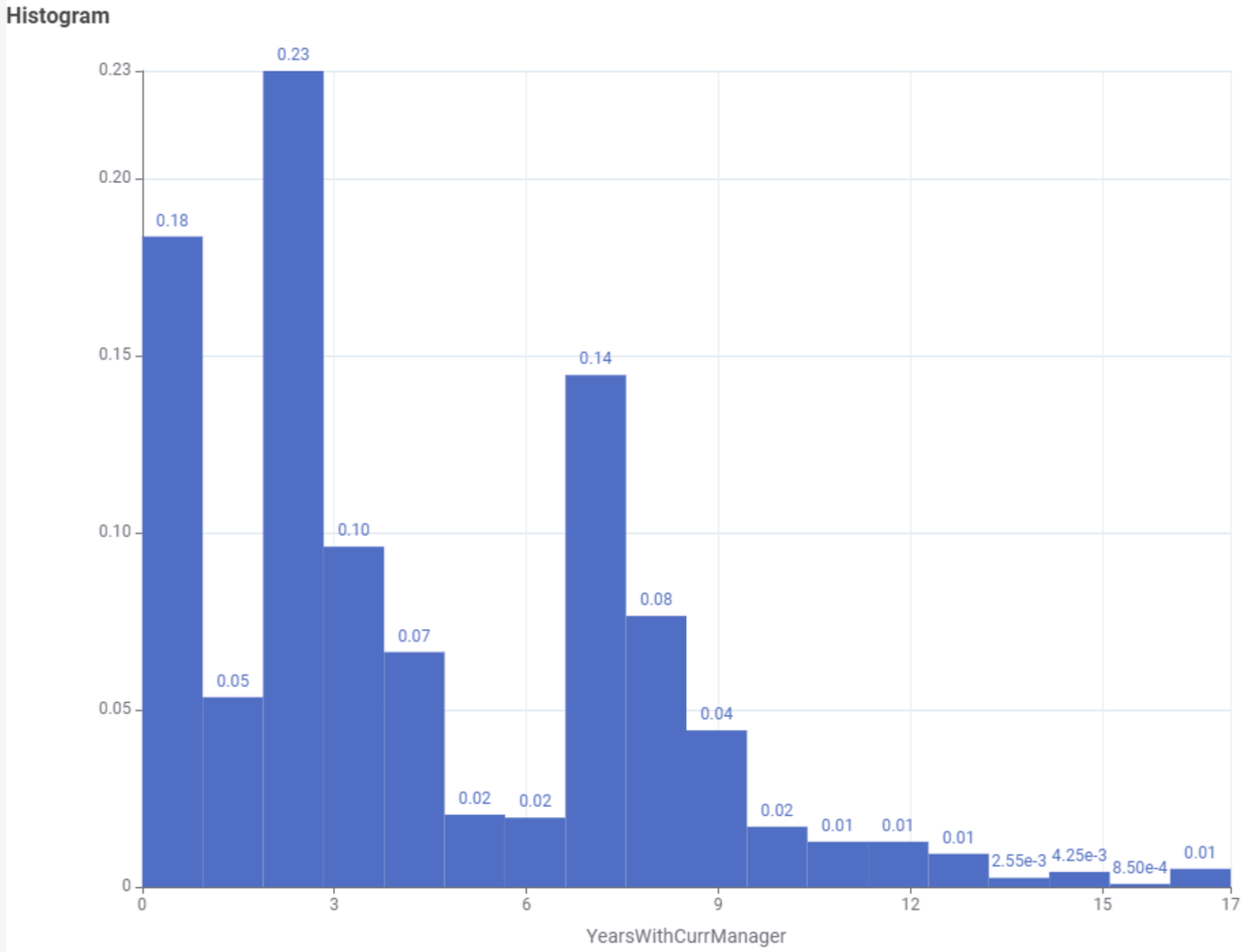
UNIVARIATE ANALYSIS

YearsWithCurrManager

Nature: Numerical, discrete

Range: 0- 17

Insights: it is right-skewed, with most employees having worked with their current manager for fewer than 5 years, and very few employees exceeding 10 years.



03 BIVARIATE ANALYSIS

BIVARIATE ANALYSIS

CONTINUOUS VS. CONTINUOUS VARIABLES

For continuous variables, the relationships were examined through **Pearson's correlation analysis**. This allowed us to quantify the strength and direction of linear associations between variable pairs.

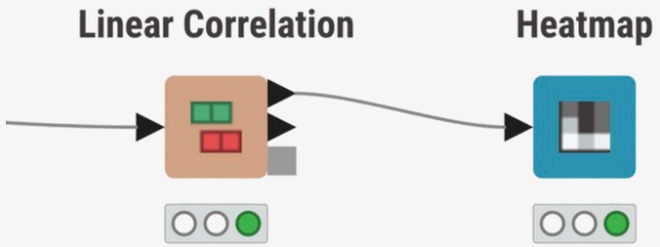
To **assess statistical significance**, **p-values were calculated** for each correlation coefficient. A low p-value indicated a significant correlation, suggesting a meaningful relationship between the variables.

Additionally, **scatter plots** were used to visualize these relationships and to **identify any potential patterns or deviations from linearity**.

This operation was done to detect **multicollinearity** which occurs when two or more variables are highly correlated, leading to **redundancy in the information they provide**. This can significantly affect regression analysis by inflating the variance of coefficient estimates, which **reduces** their **precision**. Furthermore, multicollinearity complicates the interpretation of results, as it becomes difficult to isolate the individual effect of each variable on the dependent variable. Addressing multicollinearity is essential for ensuring that models are reliable.

BIVARIATE ANALYSIS

Correlation Heatmap



Heatmap

	Age	DailyRate	EmployeeNumber	MonthlyRate	PercentSalaryHike	YearsAtCompany
DailyRate	0.01					
DistanceFromHome	-0.02	-0.01				
EmployeeNumber	-0.02	-0.06	0.05			
HourlyRate	0.02	3.67e-5	0.03	0.07		
MonthlyIncome	0.49	0.01	-0.02	-0.04	-0.01	
MonthlyRate	0.05	-0.03	0.03	2.58e-3	-3.26e-3	0.04
NumCompaniesWorked	0.29	0.05	-0.03	-7.94e-4	-3.95e-3	0.13
PercentSalaryHike	0.01	0.01	0.02	-0.03	4.68e-3	-0.03
TotalWorkingYears	0.68	0.01	-3.56e-3	-0.03	-0.01	0.77
TrainingTimesLastYear	-0.02	0.02	-0.04	0.02	-0.01	-0.03
YearsAtCompany	0.32	-0.05	0.01	-0.03	-0.02	0.51
YearsInCurrentRole	0.21	1.80e-3	0.01	-0.03	-0.02	0.36
YearsSinceLastPromotion	0.20	-0.04	-3.54e-3	-0.04	-0.02	0.33
YearsWithCurrManager	0.20	-0.04	0.01	-0.03	-0.02	0.34

	MonthlyRate	PercentSalaryHike	YearsAtCompany
NumCompaniesWorked	0.02		
PercentSalaryHike	-0.01	0.02	
TotalWorkingYears	0.05	0.22	-0.02
TrainingTimesLastYear	-0.03	-0.05	2.42e-4
YearsAtCompany	-0.02	-0.13	-0.06
YearsInCurrentRole	-0.02	-0.11	-0.03
YearsSinceLastPromotion	0.01	-0.06	-0.03
YearsWithCurrManager	-0.04	-0.12	-0.02

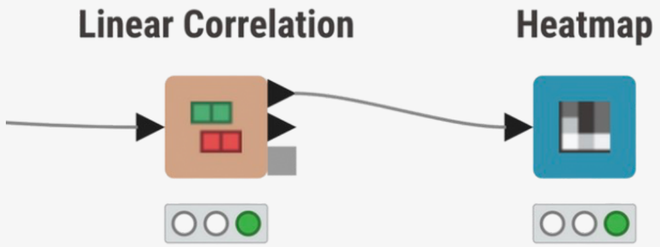
	YearsAtCompany
YearsAtCompany	0.63
YearsInCurrentRole	0.46
YearsSinceLastPromotion	0.38
YearsWithCurrManager	0.46

	YearsInCurrentRole	YearsSinceLastPromotion	YearsWithCurrManager
YearsInCurrentRole	1.95e-3		
YearsSinceLastPromotion	-0.01	0.62	0.55
YearsWithCurrManager	-4.32e-3	0.76	0.71	0.49

1
-0

BIVARIATE ANALYSIS

P-value heatmap



Heatmap

	Age	DailyRate	EmployeeNumber	MonthlyRate	PercentSalaryHike	YearsAtCompany
DailyRate	0.71					
DistanceFromHome	0.49	0.65				
EmployeeNumber	0.46	0.06	0.07			
HourlyRate	0.44	1.00	0.33	0.02		
MonthlyIncome	0	0.69	0.48	0.15	0.63	
MonthlyRate	0.08	0.28	0.27	0.93	0.91	0.17
NumCompaniesWorked	0	0.08	0.33	0.98	0.89	3.57e-6
PercentSalaryHike	0.81	0.68	0.48	0.29	0.87	0.30
TotalWorkingYears	0	0.72	0.90	0.28	0.81	0
TrainingTimesLastYear	0.42	0.47	0.16	0.40	0.65	0.37
YearsAtCompany	0	0.09	0.84	0.30	0.52	0
YearsInCurrentRole	2.19e-13	0.95	0.75	0.25	0.45	0
YearsSinceLastPromotion	5.33e-12	0.13	0.90	0.23	0.55	0
YearsWithCurrManager	1.82e-12	0.14	0.78	0.34	0.46	0

DailyRate													
DistanceFromHome													
EmployeeNumber													
HourlyRate													
MonthlyIncome													
MonthlyRate													
NumCompaniesWorked	0.52												
PercentSalaryHike	0.81	0.49											
TotalWorkingYears	0.11	7.39e-14	0.45										
TrainingTimesLastYear	0.37	0.08	0.99	0.36									
YearsAtCompany	0.50	5.94e-6	0.06	0	0.77								
YearsInCurrentRole	0.60	1.03e-4	0.31	0	0.95	0							
YearsSinceLastPromotion	0.77	0.04	0.27	0	0.69	0	0						
YearsWithCurrManager	0.13	1.91e-5	0.44	0	0.88	0	0	0					

1

0

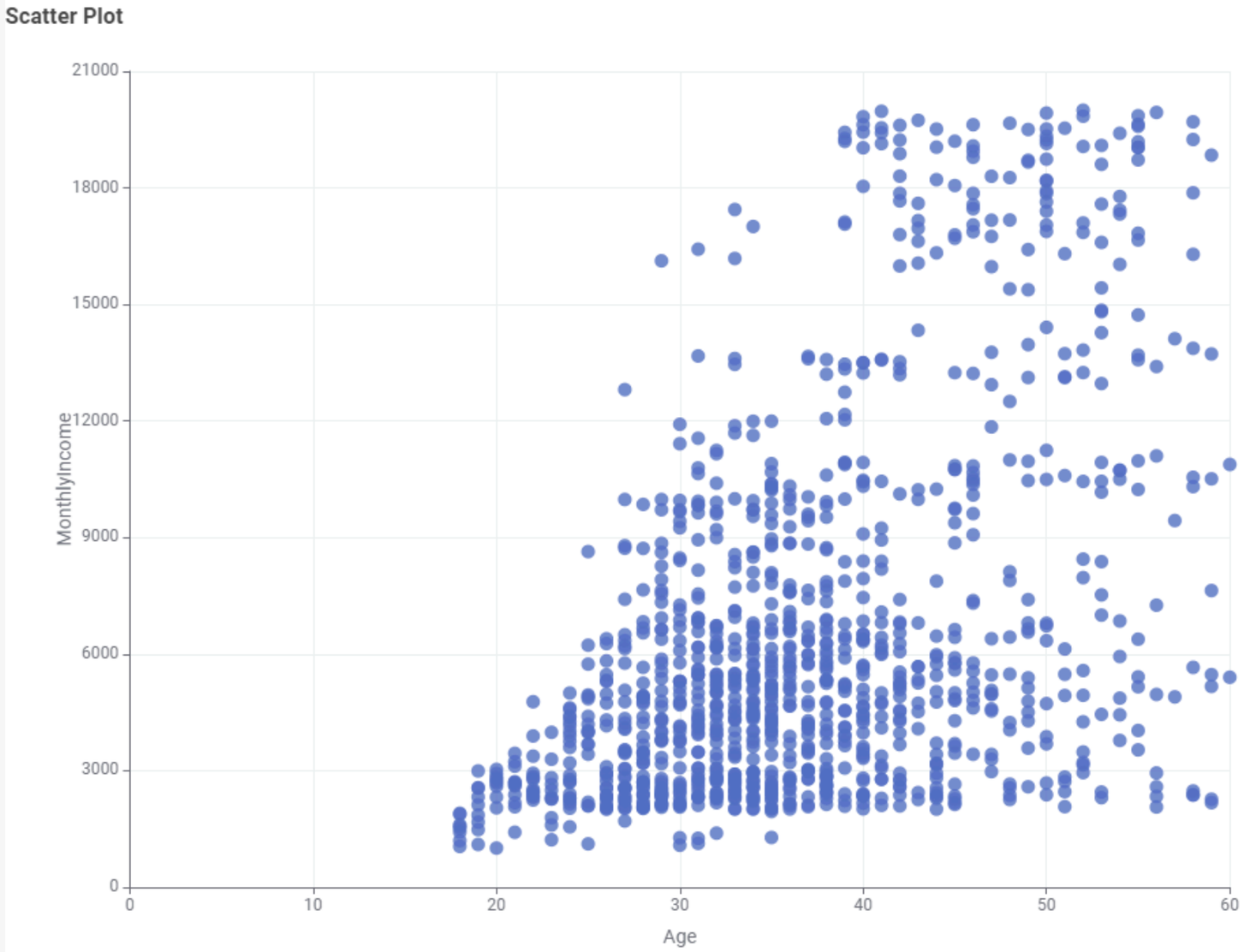
BIVARIATE ANALYSIS

Age vs MonthlyIncome

Correlation: 0.49

P-value: 0

Insights: older employees tend to have higher monthly incomes. However, there is significant variability, especially at higher ages, indicating other factors influencing income levels.



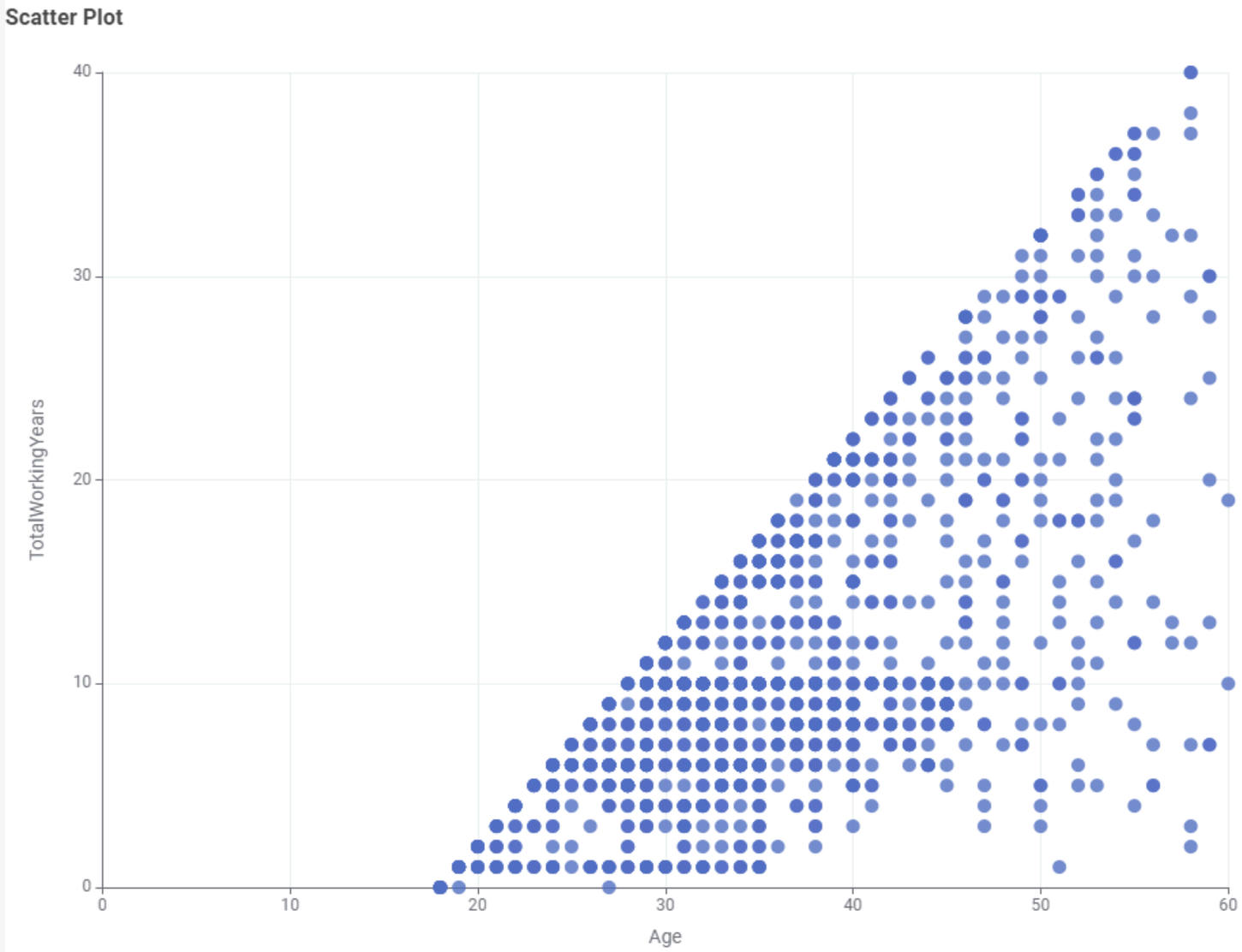
BIVARIATE ANALYSIS

Age vs TotalWorkingYears

Correlation: 0.68

P-value: 0

Insights: there is a clear linear boundary along the 45-degree line, representing the constraint that total working years cannot exceed the employee's age. This relationship reflects the logical limit imposed by an individual's age on their career duration.



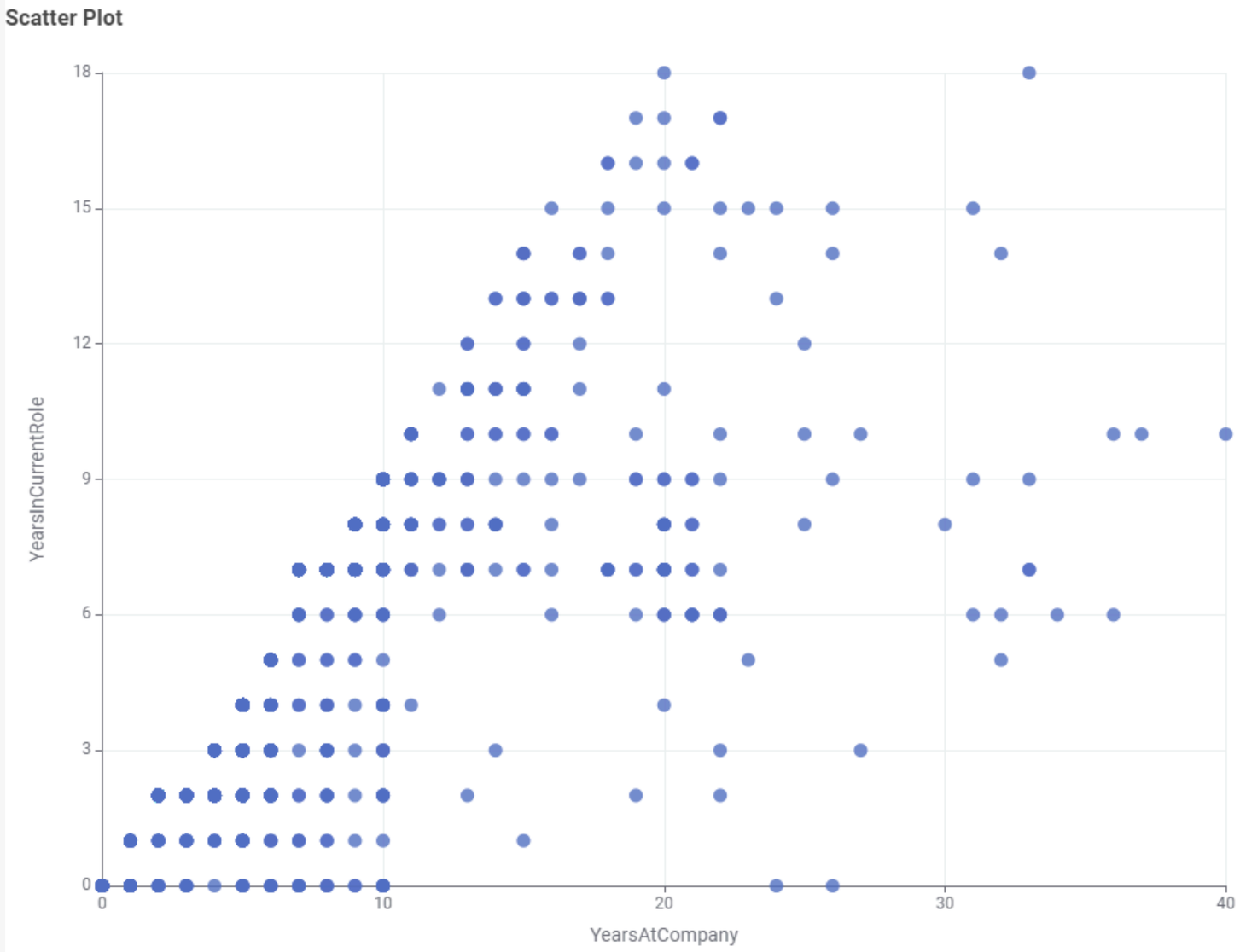
BIVARIATE ANALYSIS

YearsAtCompany vs YearsInCurrentRole

Correlation: 0.76

P-value: 0

Insights: all data points lie under or along the 45-degree line. This is due to an inherent constraint: an employee's years in their current role cannot exceed their total years at the company. This creates a clear boundary in the plot.



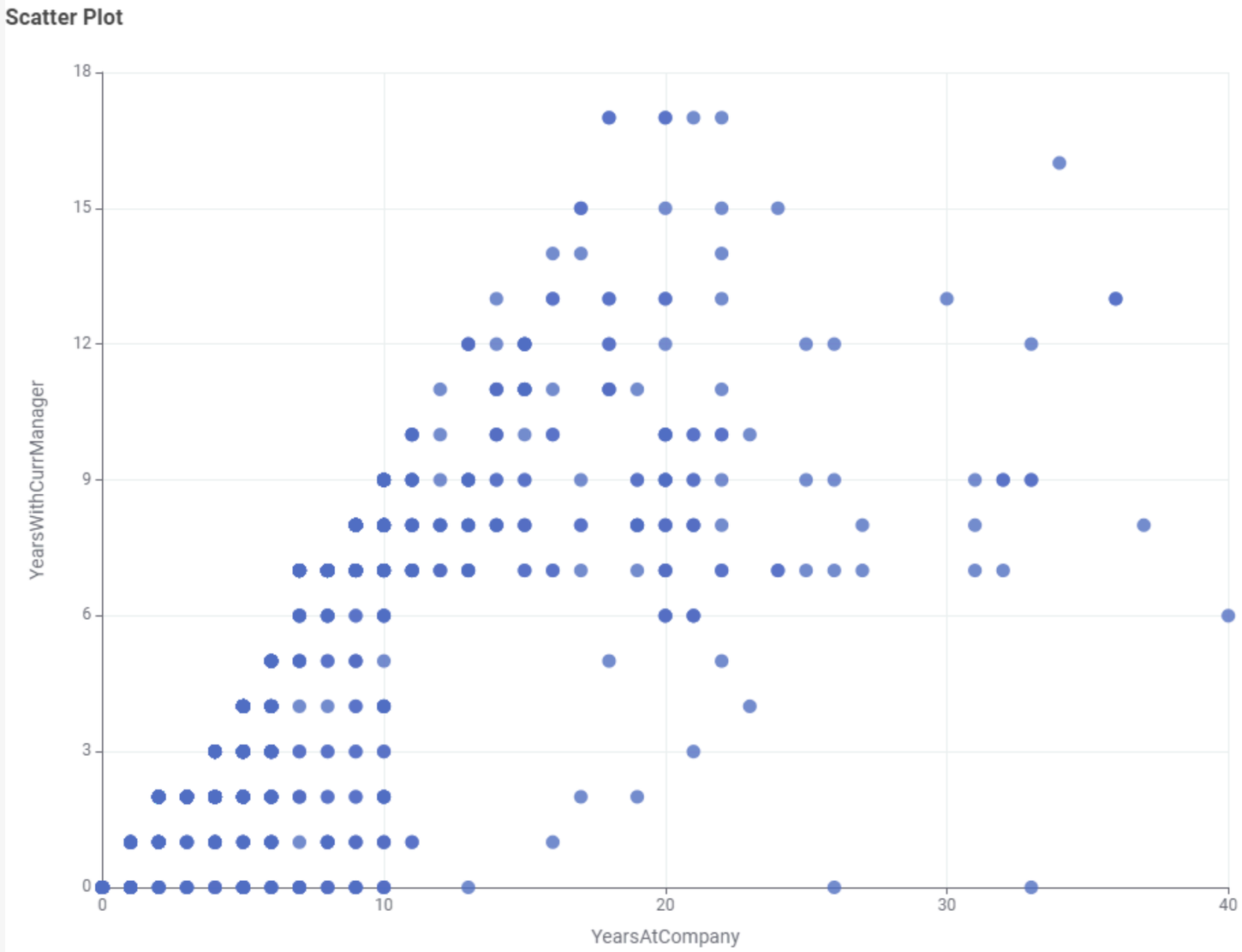
BIVARIATE ANALYSIS

YearsWithCurrManager vs YearsInCurrentRole

Correlation: 0.76

P-value: 0

Insights: all data points lying under or along the 45-degree line, reflecting the natural constraint that the number of years with the current manager cannot exceed the total years at the company. This boundary highlights the hierarchical relationship between the two variables.



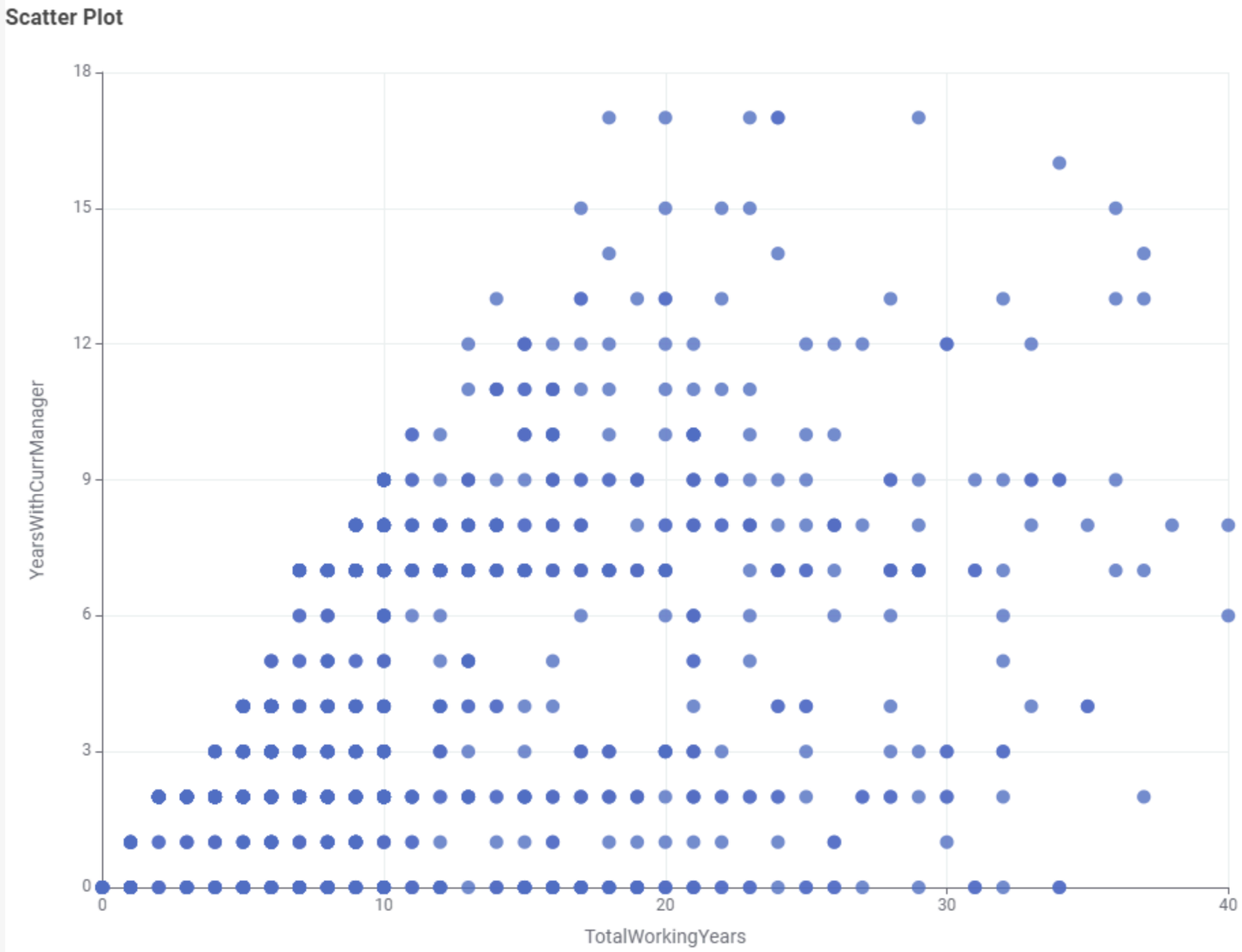
BIVARIATE ANALYSIS

YearsWithCurrManager vs TotalWorkingYears

Correlation: 0.46

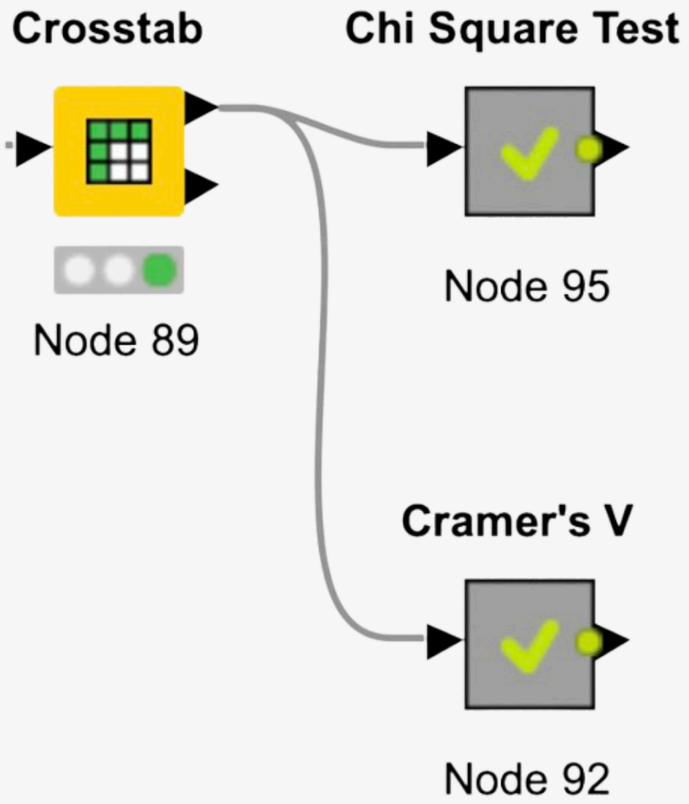
P-value: 0

Insights: all data points fall below the diagonal line, reflecting the constraint that the number of years with the current manager cannot exceed the total working years of an employee. The spread indicates variability in the proportion of career time employees spend with their current manager.



BIVARIATE ANALYSIS

CATEGORICAL VS. CATEGORICAL VARIABLES



To perform the bivariate analysis between **categorical** variables, we implemented two key statistical tests:

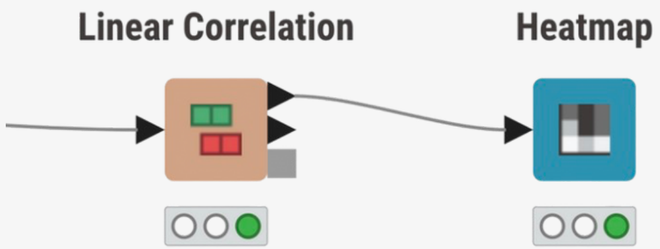
- **Cramer's V**, a measure ranging from 0 to 1, evaluates the strength of association between two categorical variables by analyzing how records are distributed across the combinations of their categories. This provides a clear indication of the relationship's intensity.

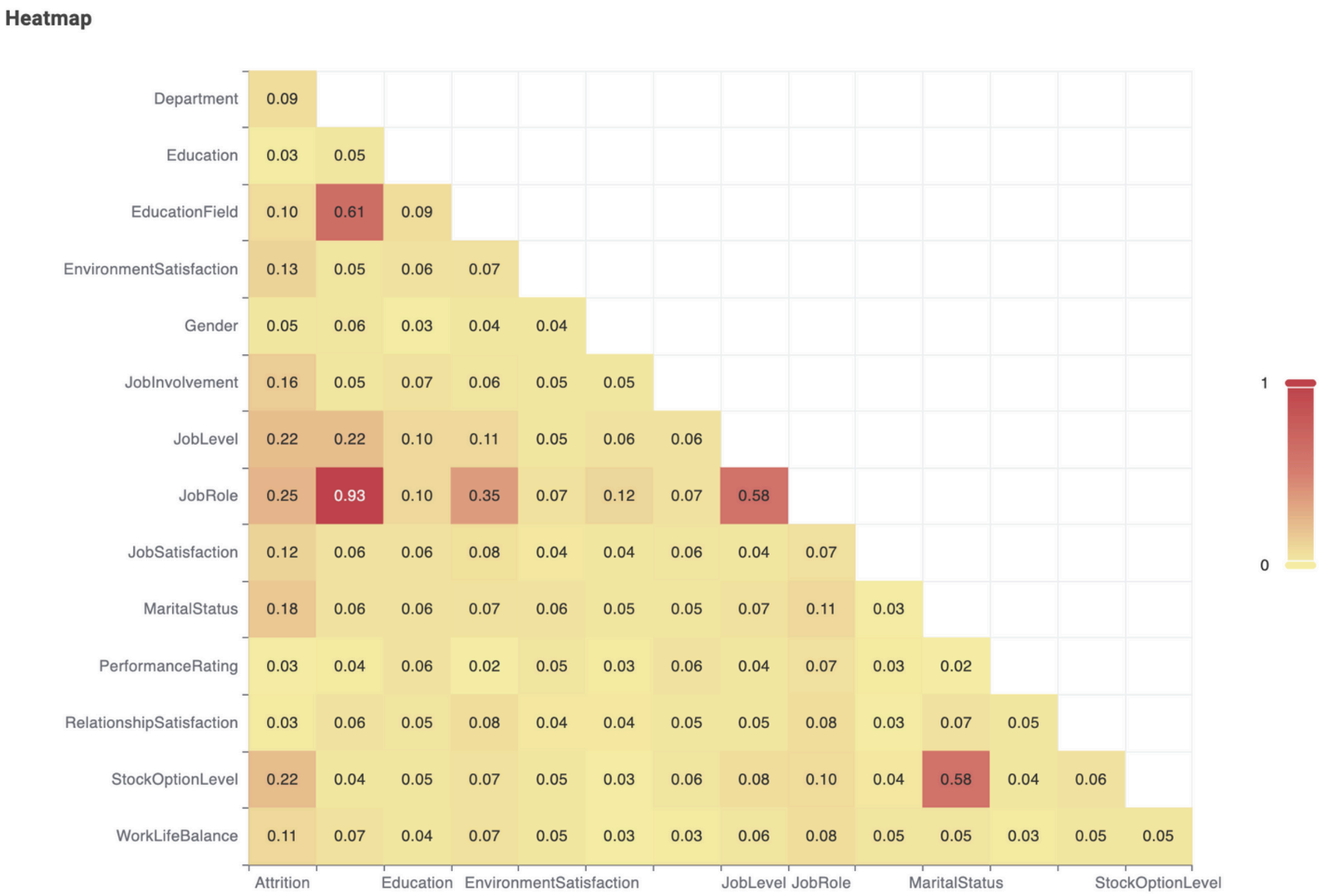
- The **Chi-Square test** complements Cramer's V by assessing the statistical significance of the observed associations. It ensures that the detected relationships are unlikely to have occurred by chance, validating the **reliability** of Cramer's V.

By combining these two tests, we gain a **comprehensive understanding of how categorical variables interact with the target feature**, allowing us to quantify the **strength of associations while ensuring their statistical validity.** This approach offers both precision and confidence in identifying meaningful patterns within the data.

BIVARIATE ANALYSIS

Cramer’s V matrix





BIVARIATE ANALYSIS

Attrition - JobLevel

The analysis reveals a weak but statistically significant association between job level and attrition (Cramer's V = 0.22, Chi-Square = 69.09, p < 0.0001). Entry-level employees exhibit the highest attrition rate (10.37%), significantly exceeding expectations, while attrition decreases substantially at higher job levels, such as "Top." These findings suggest targeted retention strategies are necessary for entry-level roles, addressing potential issues like career growth and job satisfaction.

CRAMER V

Statistics

Frequency Expected Percent	entry-level	high	low	mid	top	Total
No	309	80	393	142	54	978
	358.4337	69.0255	364.2551	138.051	48.2347	
	26.2755%	6.8027%	33.4184%	12.0748%	4.5918%	83.1633%
Yes	122	3	45	24	4	198
	72.5663	13.9745	73.7449	27.949	9.7653	
	10.3741%	0.2551%	3.8265%	2.0408%	0.3401%	16.8367%
Total	431	83	438	166	58	1,176
	36.6497%	7.0578%	37.2449%	14.1156%	4.932%	100%

Chi-Squared Test

Statistic	DF	Value	Prob
Chi-Square	4	69.0929	3.53E-14

Total sample size: 1176.0

BIVARIATE ANALYSIS

Attrition - JobRole

The analysis shows a moderate relationship between job role and attrition (Cramer's V = 0.25, Chi-Square = 68.00, p < 0.0001). Sales-related roles, including Sales Executives (22.45%) and Representatives (5.78%), exhibit higher attrition rates compared to other roles, such as Managers and Research Scientists. This suggests that attrition is more prevalent in sales-oriented positions, potentially due to performance pressure or job demands, warranting targeted interventions to improve retention in these areas.

Statistics

Frequency Expected Percent	Healthcare Representative	Human Resources	Laboratory Technician	Manager	Manufacturing Director	Research Director	Research Scientist	Sales Executive	Sales Representative	Total
No	95	35	153	74	112	64	188	214	43	978
	84.8265	36.5918	171.3163	64.8673	98.1327	54.8878	191.2755	219.551	56.551	
	8.0782%	2.9762%	13.0102%	6.2925%	9.5238%	5.4422%	15.9864%	18.1973%	3.6565%	83.1633%
Yes	7	9	53	4	6	2	42	50	25	198
	17.1735	7.4082	34.6837	13.1327	19.8673	11.1122	38.7245	44.449	11.449	
	0.5952%	0.7653%	4.5068%	0.3401%	0.5102%	0.1701%	3.5714%	4.2517%	2.1259%	16.8367%
Total	102	44	206	78	118	66	230	264	68	1,176
	8.6735%	3.7415%	17.517%	6.6327%	10.034%	5.6122%	19.5578%	22.449%	5.7823%	100%

CRAMER V

Chi-Squared Test

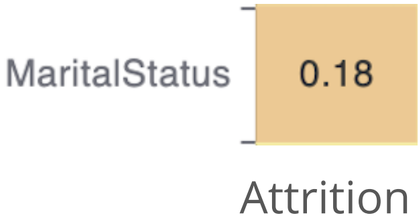
Statistic	DF	Value	Prob
Chi-Square	8	68.0029	1.23E-11

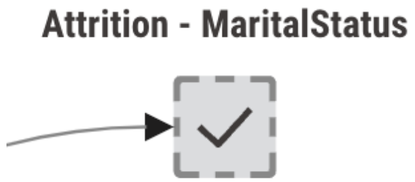
BIVARIATE ANALYSIS

Attrition - MaritalStatus

The bivariate analysis shows a weak relationship between marital status and attrition (Cramer's V = 0.18, Chi-Square = 40.44, p < 0.0001). Singles have the highest attrition rate (8.59%), followed by those who are divorced (2.47%), while married employees exhibit the lowest attrition rate (5.78%). This trend suggests that marital status may influence an employee's likelihood of leaving, potentially due to varying commitments and priorities.

CRAMER V:





Statistics:

Frequency Expected Percent	Divorced	Married	Single	Total
No	235	469	274	978
	219.551	446.5867	311.8622	
	19.983%	39.881%	23.2993%	83.1633%
Yes	29	68	101	198
	44.449	90.4133	63.1378	
	2.466%	5.7823%	8.5884%	16.8367%
Total	264	537	375	1,176
	22.449%	45.6633%	31.8878%	100%

Chi-Squared test:

Statistic	DF	Value	Prob
Chi-Square	2	40.4396	1.65E-9

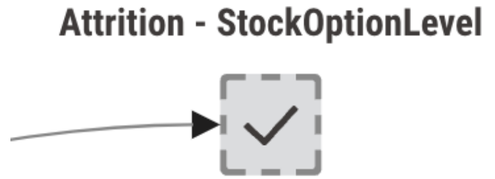
Total sample size: 1176.0

BIVARIATE ANALYSIS

Attrition - StockOptionLevel

The analysis reveals a moderate relationship between stock option level and attrition (Cramer's V = 0.22, Chi-Square = 55.41, $p < 0.0001$). Employees with low stock option levels experience the highest attrition (3.83%), followed by those with high stock options (1.11%), while attrition is minimal among employees with mid-level stock options (0.85%). The findings suggest that stock options may play a role in employee retention, with insufficient stock options potentially contributing to higher attrition.

CRAMER V:



Statistics:

Frequency Expected Percent	high	low	mid	unknown	Total
No	53	439	112	374	978
	54.8878	402.5102	101.4592	419.1429	
	4.5068%	37.3299%	9.5238%	31.8027%	83.1633%
Yes	13	45	10	130	198
	11.1122	81.4898	20.5408	84.8571	
	1.1054%	3.8265%	0.8503%	11.0544%	16.8367%
Total	66	484	122	504	1,176
	5.6122%	41.1565%	10.3741%	42.8571%	100%

Chi-Squared test:

Statistic	DF	Value	Prob
Chi-Square	3	55.4148	5.60E-12

Total sample size: 1176.0

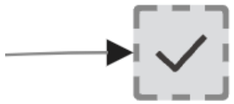
BIVARIATE ANALYSIS

Department - JobLevel

The analysis shows a moderate association between department and job level (Cramer's V = 0.22, Chi-Square = 109.44, p < 0.0001). The Research & Development department dominates across job levels, particularly at entry and low levels, accounting for 65.05% of the workforce. Sales represent 30.78%, concentrated at the low and mid-levels. Human Resources make up the smallest share (4.17%), primarily at entry-level. This highlights distinct hierarchical distributions among departments, with Research & Development showing a broader representation across higher levels.

CRAMER V:

JobLevel - Department



Statistics:

Frequency Expected Percent	entry-level	high	low	mid	top	Total
Human Resources	28	1	12	4	4	49
	17.9583	3.4583	18.25	6.9167	2.4167	
	2.381%	0.085%	1.0204%	0.3401%	0.3401%	4.1667%
Research & Development	340	57	230	95	43	765
	280.3699	53.9923	284.9235	107.9847	37.7296	
	28.9116%	4.8469%	19.5578%	8.0782%	3.6565%	65.051%
Sales	63	25	196	67	11	362
	132.6718	25.5493	134.8265	51.0986	17.8537	
	5.3571%	2.1259%	16.6667%	5.6973%	0.9354%	30.7823%
Total	431	83	438	166	58	1,176
	36.6497%	7.0578%	37.2449%	14.1156%	4.932%	100%

Chi-Squared test:

Statistic	DF	Value	Prob
Chi-Square	8	109.4394	4.96E-20

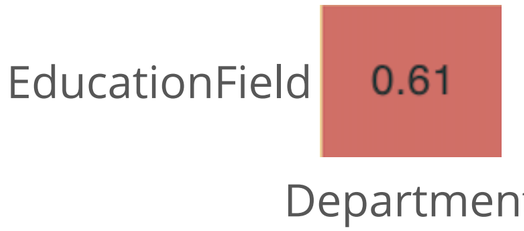
Total sample size: 1176.0

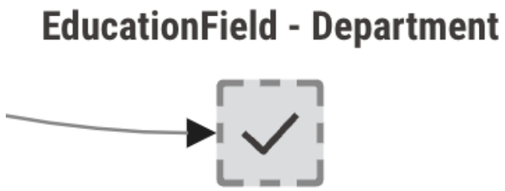
BIVARIATE ANALYSIS

Department - EducationField

The analysis demonstrates a strong association between department and education field (Cramer's V = 0.61, Chi-Square = 732.35, p < 0.0001). Employees in Research & Development predominantly come from Life Sciences (41.07%), reflecting the specialized skills needed in this department. Sales are more evenly distributed but have notable representation from Marketing (10.97%) and Medical (5.95%). Human Resources has minimal representation and is distributed across various education fields. These findings highlight the alignment between educational background and departmental needs, particularly the critical link between technical expertise and roles in R&D.

CRAMER V:





Statistics:

Frequency Expected Percent	Human Resources	Life Sciences	Marketing	Medical	Other	Technical Degree	Total
Human Resources	22	12		10	2	3	49
	0.9167	20.125		15.2917	2.8333	4.4583	
	1.8707%	1.0204%		0.8503%	0.1701%	0.2551%	4.1667%
Research & Development		349		287	53	76	765
		314.1964		238.7372	44.2347	69.6046	
		29.6769%		24.4048%	4.5068%	6.4626%	65.051%
Sales		122	129	70	13	28	362
		148.6786	39.7092	112.9711	20.932	32.9371	
		10.3741%	10.9694%	5.9524%	1.1054%	2.381%	30.7823%
Total	22	483	129	367	68	107	1,176
	1.8707%	41.0714%	10.9694%	31.2075%	5.7823%	9.0986%	100%

Chi-Squared test:

Statistic	DF	Value	Prob
Chi-Square	10	732.3456	7.12E-151

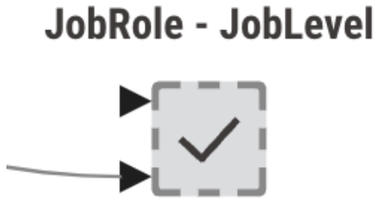
Total sample size: 1176.0

BIVARIATE ANALYSIS

JobRole - JobLevel

The analysis highlights a significant relationship between job roles and levels, with Cramer's V = 0.58 and a highly significant Chi-Square statistic of 1167.23 ($p < 0.0001$). Certain job roles, like Research Scientist and Sales Executive, show a wide distribution across job levels, emphasizing career progression within these roles. Conversely, roles like Laboratory Technician and Human Resources exhibit higher concentration in entry-level and low job levels. This indicates structured growth opportunities and potential disparities in career advancement depending on job roles.

CRAMER V:



Statistics:

Frequency Deviation Percent	entry-level	high	low	mid	top	Total
Healthcare Representative		7	63	32		102
		-0.199	25.0102	17.602		
		0.5952%	5.3571%	2.7211%		8.6735%
Human Resources	28		12	4		44
	11.8741		-4.3878	-2.2109		
	2.381%		1.0204%	0.3401%		3.7415%
Laboratory Technician	159		46	1		206
	83.5017		-30.7245	-28.0782		
	13.5204%		3.9116%	0.085%		17.517%
Manager		33		11	34	78
		27.4949		-0.0102	30.1531	
		2.8061%		0.9354%	2.8912%	6.6327%
Manufacturing Director		9	73	36		118
		0.6718	29.051	19.3435		
		0.7653%	6.2075%	3.0612%		10.034%
Research Director		25		17	24	66
		20.3418		7.6837	20.7449	
		2.1259%		1.4456%	2.0408%	5.6122%
Research Scientist	181		48	1		230
	96.7058		-37.6633	-31.466		
	15.3912%		4.0816%	0.085%		19.5578%
Sales Executive		9	191	64		264
		-9.6327	92.6735	26.7347		
		0.7653%	16.2415%	5.4422%		22.449%
Sales Representative	63		5			68
	38.0782		-20.3265			
	5.3571%		0.4252%			5.7823%
Total	431	83	438	166	58	1,176
	36.6497%	7.0578%	37.2449%	14.1156%	4.932%	100%

Chi-Squared test:

Statistic	DF	Value	Prob
Chi-Square	32	1,167.2299	8.43E-225

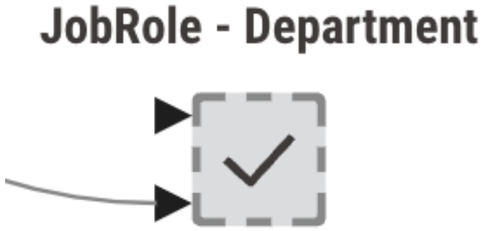
Total sample size: 1176.0

BIVARIATE ANALYSIS

JobRole - Department

The bivariate analysis demonstrates a strong association between Job Roles and Departments, with a very high Cramer's V value of 0.93, indicating near-complete dependence. The Chi-Square statistic (1,625.76, p < 0.001) confirms this significant relationship. Specific job roles, such as Laboratory Technicians and Research Scientists, are heavily concentrated within the Research & Development department, while Sales Executives dominate the Sales department. This reflects clear alignment of job roles with departmental functions, underscoring specialization within organizational structures.

CRAMER V:



Statistics:

Frequency Deviation Percent	Human Resources	Research & Development	Sales	Total
Healthcare Representative		102		102
		35.648		
		8.6735%		8.6735%
Human Resources	44			44
	42.1667			
	3.7415%			3.7415%
Laboratory Technician		206		206
		71.9949		
		17.517%		17.517%
Manager	5	43	30	78
	1.75	-7.7398	5.9898	
	0.4252%	3.6565%	2.551%	6.6327%
Manufacturing Director		118		118
		41.2398		
		10.034%		10.034%
Research Director		66		66
		23.0663		
		5.6122%		5.6122%
Research Scientist		230		230
		80.3827		
		19.5578%		19.5578%
Sales Executive			264	264
			182.7347	
			22.449%	22.449%
Sales Representative			68	68
			47.068	
			5.7823%	5.7823%
Total	49	765	362	1,176
	4.1667%	65.051%	30.7823%	100%

Chi-Squared test:

Statistic	DF	Value	Prob
Chi-Square	16	1,625.7557	0.0

Total sample size: 1176.0

BIVARIATE ANALYSIS

JobRole - EducationField

The bivariate analysis reveals a moderate association between Job Roles and Education Fields, with a Cramer's V value of 0.35, indicating a notable relationship. The Chi-Square statistic (584.33, $p < 0.001$) highlights a significant dependency. Certain education fields correspond strongly to specific job roles, such as individuals with technical degrees or life sciences backgrounds clustering in specialized roles. This emphasizes the alignment between educational qualifications and career paths, showcasing the importance of field-specific education in determining job roles within an organization.

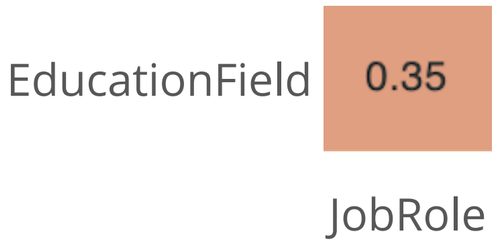
Statistics

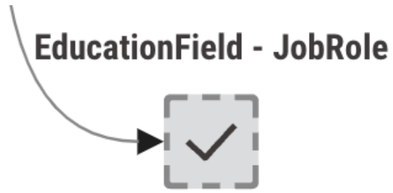
Frequency Deviation Percent	Healthcare Representative	Human Resources	Laboratory Technician	Manager	Manufacturing Director	Research Director	Research Scientist	Sales Executive	Sales Representative	Total
Human Resources		18		4						22
		17.1769		2.5408						
		1.5306%		0.3401%						1.8707%
Life Sciences	46	12	95	26	58	29	106	90	21	483
	4.1071	-6.0714	10.3929	-6.0357	9.5357	1.8929	11.5357	-18.4286	-6.9286	
	3.9116%	1.0204%	8.0782%	2.2109%	4.932%	2.466%	9.0136%	7.6531%	1.7857%	41.0714%
Marketing				13				98	18	129
				4.4439				69.0408	10.5408	
				1.1054%				8.3333%	1.5306%	10.9694%
Medical	36	9	78	26	42	30	81	48	17	367
	4.1684	-4.7313	13.7126	1.6582	5.1752	9.4031	9.2228	-34.3878	-4.2211	
	3.0612%	0.7653%	6.6327%	2.2109%	3.5714%	2.551%	6.8878%	4.0816%	1.4456%	31.2075%
Other	8	2	18	5	7	3	12	12	1	68
	2.102	-0.5442	6.0884	0.4898	0.1769	-0.8163	-1.2993	-3.2653	-2.932	
	0.6803%	0.1701%	1.5306%	0.4252%	0.5952%	0.2551%	1.0204%	1.0204%	0.085%	5.7823%
Technical Degree	12	3	15	4	11	4	31	16	11	107
	2.7194	-1.0034	-3.7432	-3.0969	0.2636	-2.0051	10.0731	-8.0204	4.8129	
	1.0204%	0.2551%	1.2755%	0.3401%	0.9354%	0.3401%	2.6361%	1.3605%	0.9354%	9.0986%
Total	102	44	206	78	118	66	230	264	68	1,176

Chi-Squared Test

Statistic	DF	Value	Prob
Chi-Square	40	606.7209	2.25E-102

CRAMER V



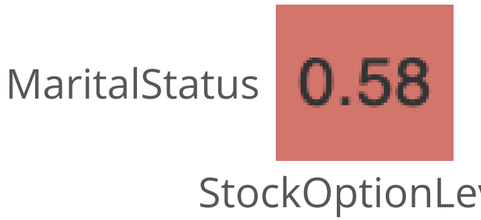


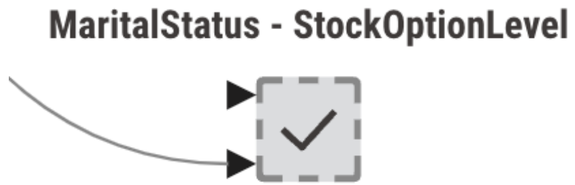
BIVARIATE ANALYSIS

MaritalStatus - StockOptionLevel

The analysis shows a strong association between marital status and stock option levels (Cramer's V = 0.58, significant Chi-Square test). Divorced individuals are mostly in the low level (51%), married employees are distributed across low (27%) and unknown (10%), while singles dominate the unknown category (32%). These patterns suggest that stock option allocation may reflect organizational policies tailored to marital status or differing employee needs.

CRAMER V:





Statistics:

Frequency Deviation Percent	high	low	mid	unknown	Total
Divorced	37	160	61	6	264
	22.1837	51.3469	33.6122	-107.1429	
	3.1463%	13.6054%	5.1871%	0.5102%	22.449%
Married	29	324	61	123	537
	-1.1378	102.9898	5.2908	-107.1429	
	2.466%	27.551%	5.1871%	10.4592%	45.6633%
Single				375	375
				214.2857	
				31.8878%	31.8878%
Total	66	484	122	504	1,176
	5.6122%	41.1565%	10.3741%	42.8571%	100%

Chi-Squared test:

Statistic	DF	Value	Prob
Chi-Square	6	584.325	5.61E-123

Total sample size: 1176.0

BIVARIATE ANALYSIS

CONTINUOUS VS. CATEGORICAL VARIABLES

The analysis of relationships **between continuous and categorical variables** was conducted by evaluating how the distribution of the continuous target variable varied across the categories of the categorical variable. This approach helps uncover whether specific groups **exhibit distinct patterns or behaviors in the target variable.**

Boxplots were employed as a visual tool to **compare distributions across categories**. These plots provide insights into the **central tendency** (e.g., median), variability (e.g., interquartile range), and the presence of outliers within each group. This visual representation allows for an intuitive understanding of differences between categories.

To statistically validate these observations, we performed a **One-Way ANOVA** (Analysis of Variance). This test determines whether the means of the continuous target variable differ significantly across the categories of the categorical variable. A significant result indicates that at least one group mean is different, suggesting a meaningful relationship between the variables.

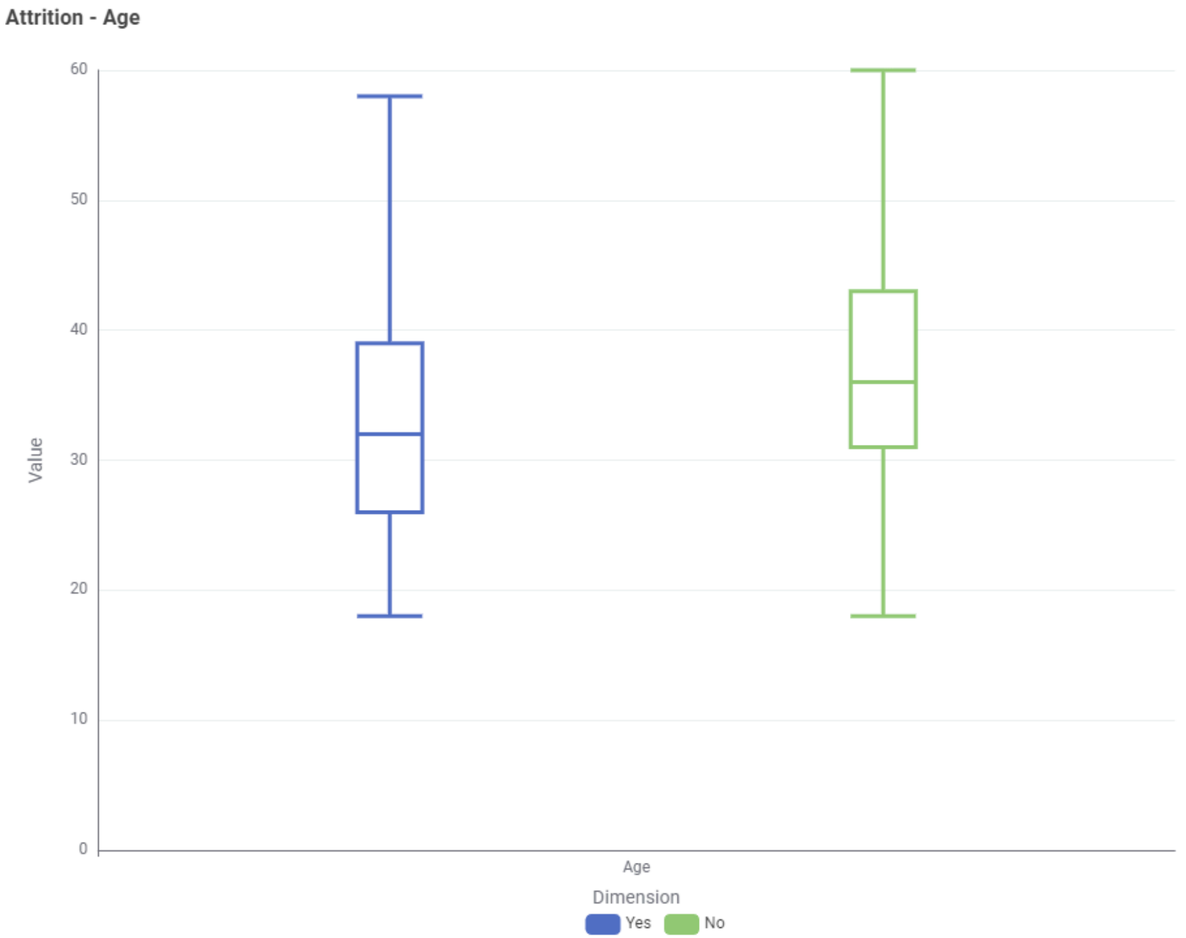
By combining **visual** and **statistical** approaches, this method ensures both **clarity** and **rigor** in identifying and **interpreting** the **interactions** between continuous and categorical variables.

BIVARIATE ANALYSIS

Attrition vs Age

Insights: employees who left the company (Attrition = Yes) tend to be younger, with a median age lower than those who stayed. The age range of employees who stayed is broader, indicating that attrition is more common among younger employees.

Test Column *String*	Group *String*	N *Number (integer)*	Mean *Number (double)*	Standard Deviation *Number (double)*
Age	Yes	198	33.348	9.809
Age	No	978	37.543	8.811
Age	Total	1176	36.837	9.118

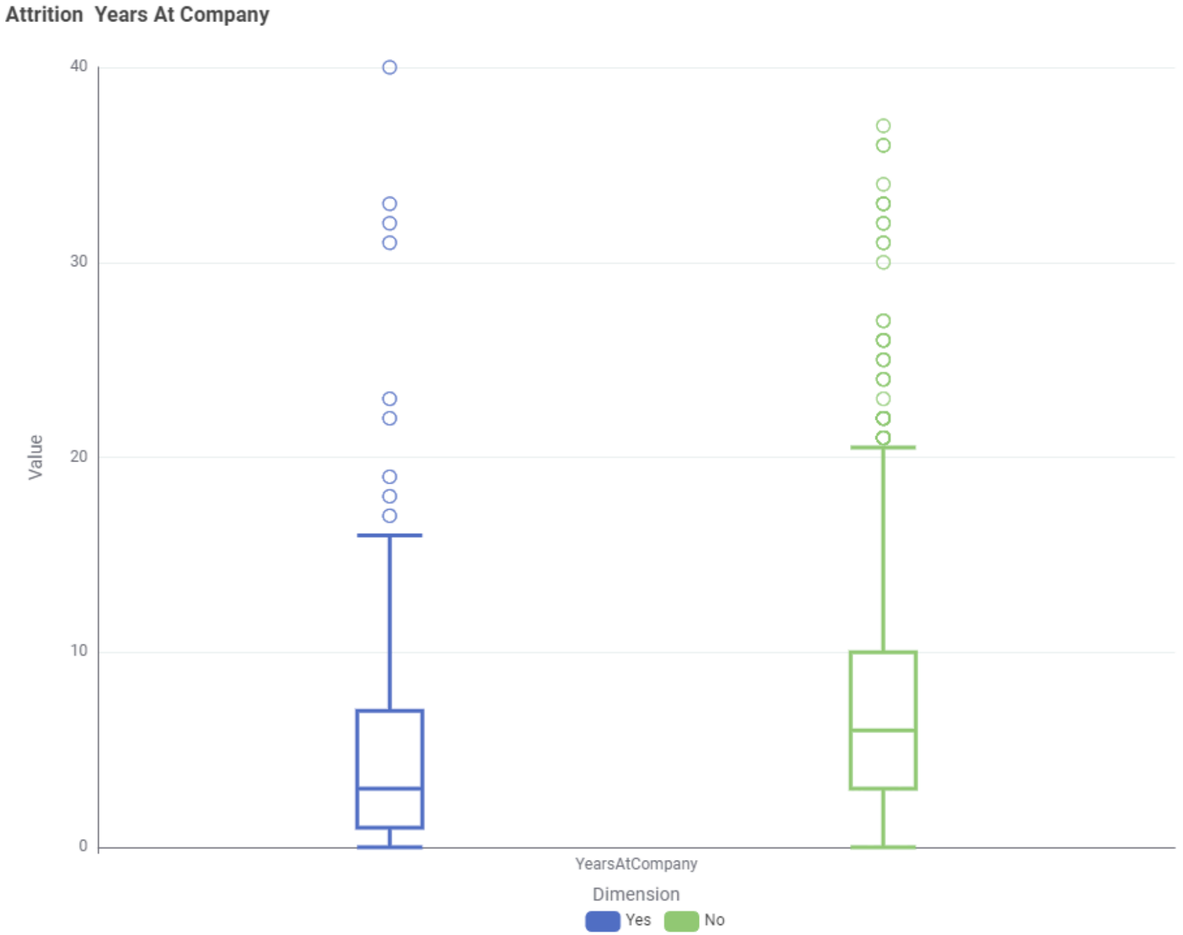


BIVARIATE ANALYSIS

Attrition vs YearsAtCompany

Insights: employees who left (Attrition = Yes) generally had shorter tenures, with a median significantly lower than those who stayed. The group of employees who stayed (Attrition = No) includes a broader range of tenures, including employees with very long durations at the company.

Test Column *String*	Group *String*	N *Number (integer)*	Mean *Number (double)*	Standard Deviation *Number (double)*
YearsAtCompany	Yes	198	5.025	5.971
YearsAtCompany	No	978	7.4	6.208
YearsAtCompany	Total	1176	7	6.23

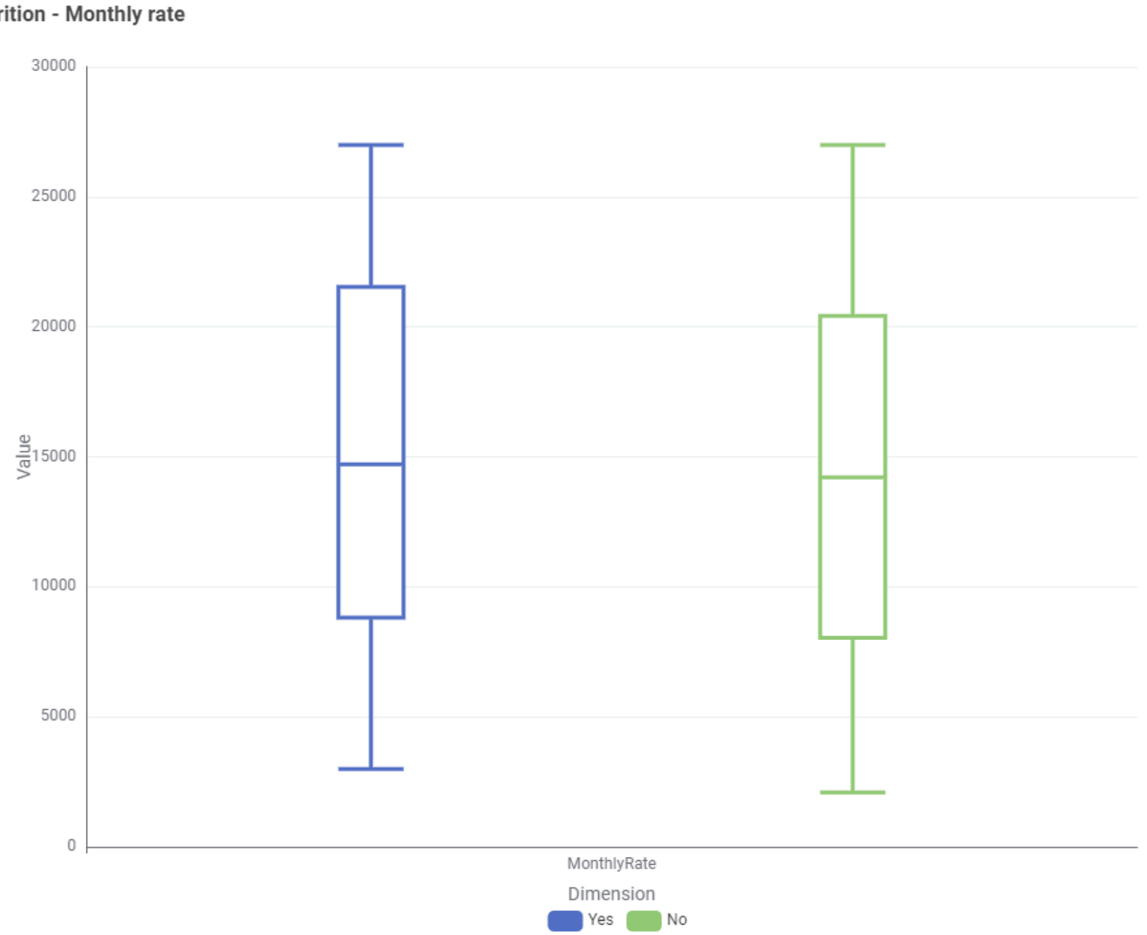


BIVARIATE ANALYSIS

Attrition vs MonthlyRate

Insights: there is no significant difference in the distribution of monthly rates between employees who left (Attrition = Yes) and those who stayed (Attrition = No). Both groups have similar medians and interquartile ranges, suggesting that monthly rate might not be a strong predictor of attrition.

Test Column *String*	Group *String*	N *Number (integer)*	Mean *Number (double)*	Standard Deviation *Number (double)*
MonthlyRate	Yes	198	14,759.101	7,286.817
MonthlyRate	No	978	14,266.292	7,088.825
MonthlyRate	Total	1176	14,349.265	7,121.791

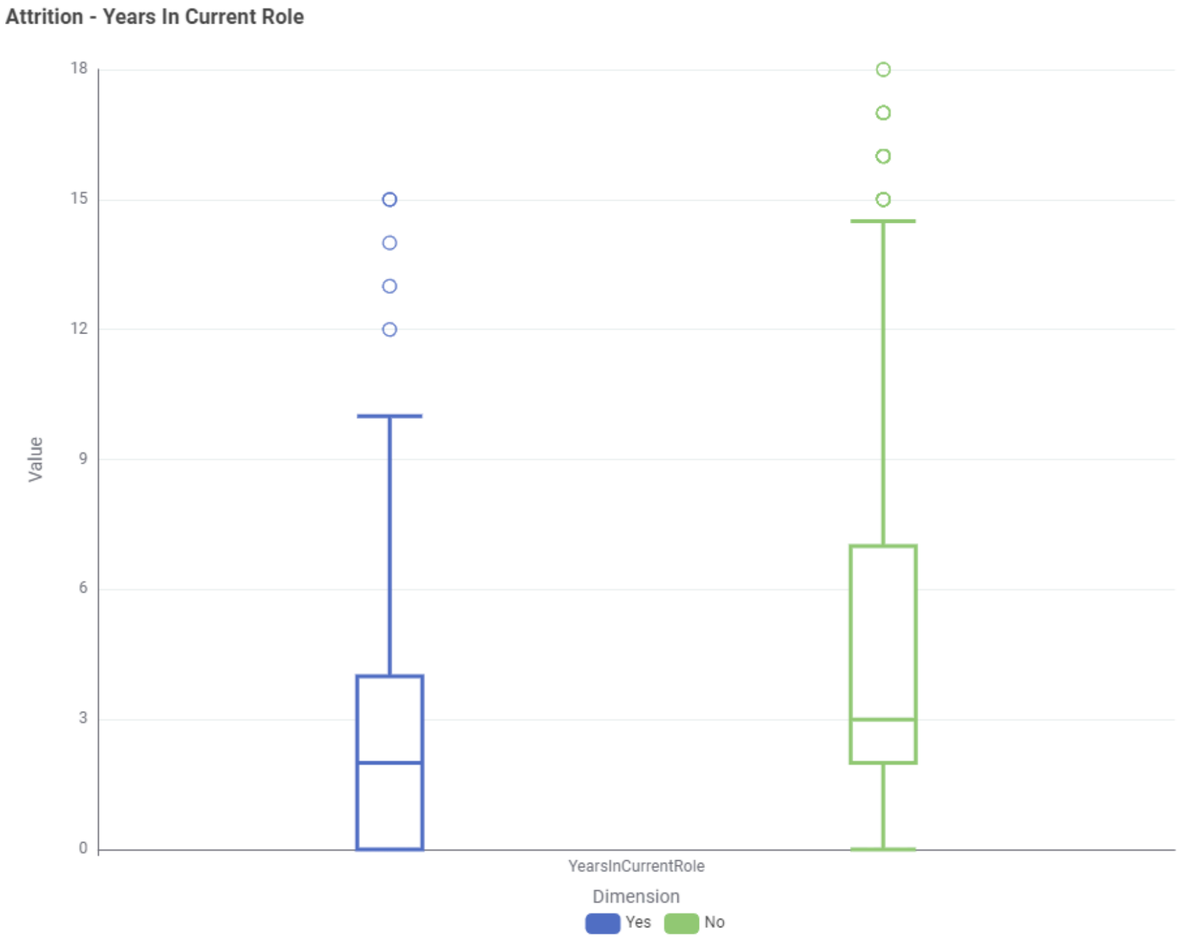


BIVARIATE ANALYSIS

Attrition vs YearsInCurrentRole

Insights: employees who have left the organization tend to have spent fewer years in their current role compared to those who stayed, as indicated by the lower median and interquartile range for the "Yes" group. Longer tenure in a role appears to correlate with lower attrition rates.

Test Column *String*	Group *String*	N *Number (integer)*	Mean *Number (double)*	Standard Deviation *Number (double)*
YearsInCurrentRole	Yes	198	2.879	3.189
YearsInCurrentRole	No	978	4.465	3.694
YearsInCurrentRole	Total	1176	4.198	3.661

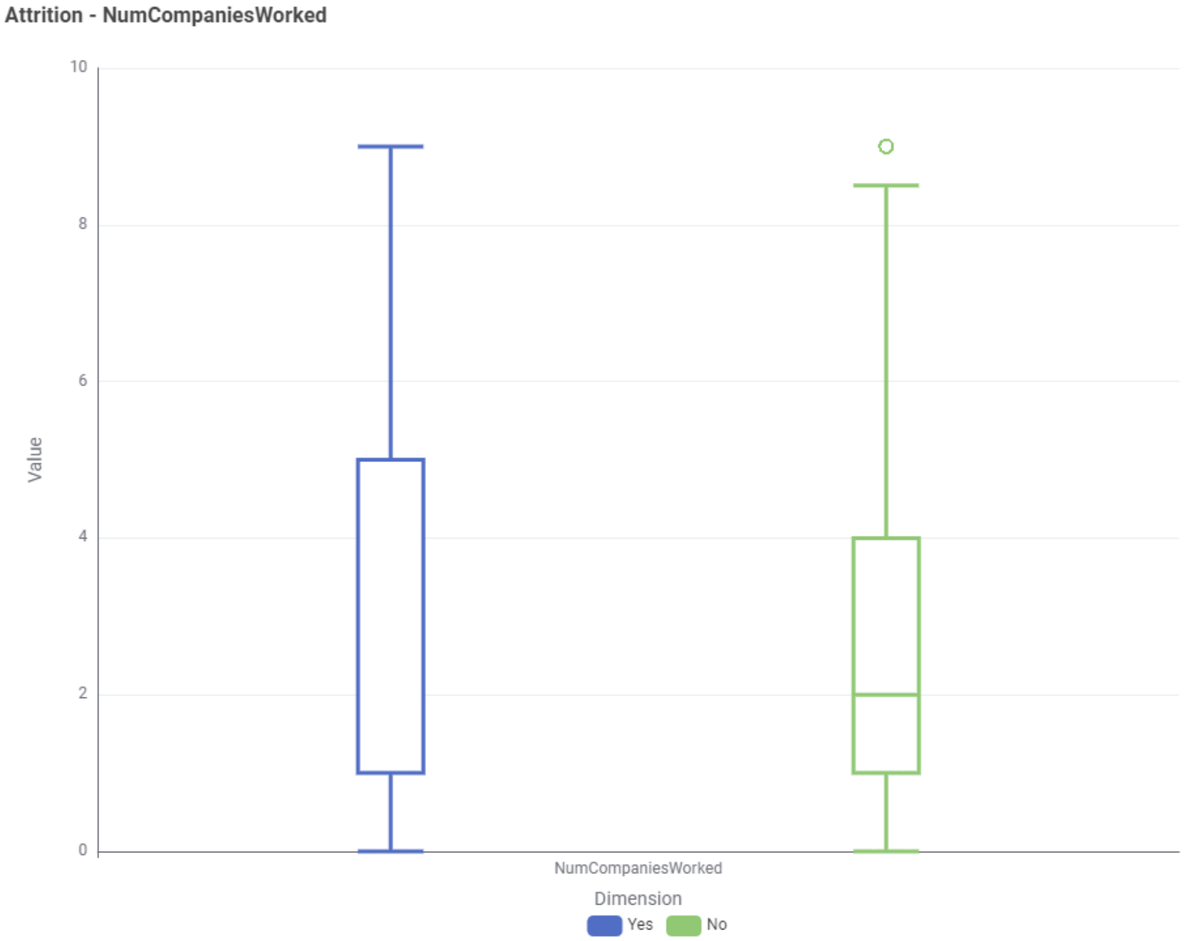


BIVARIATE ANALYSIS

Attrition vs NumCompaniesWorked

Insights: employees who have left the organization ("Yes" group) generally tend to have worked for a higher number of companies compared to those who stayed ("No" group), as evidenced by the higher median and broader distribution in the "Yes" group. This suggests that higher mobility between companies may be associated with higher attrition.

Test Column *String*	Group *String*	N *Number (integer)*	Mean *Number (double)*	Standard Deviation *Number (double)*
NumCompaniesWorked	Yes	198	2.929	2.701
NumCompaniesWorked	No	978	2.645	2.472
NumCompaniesWorked	Total	1176	2.693	2.513

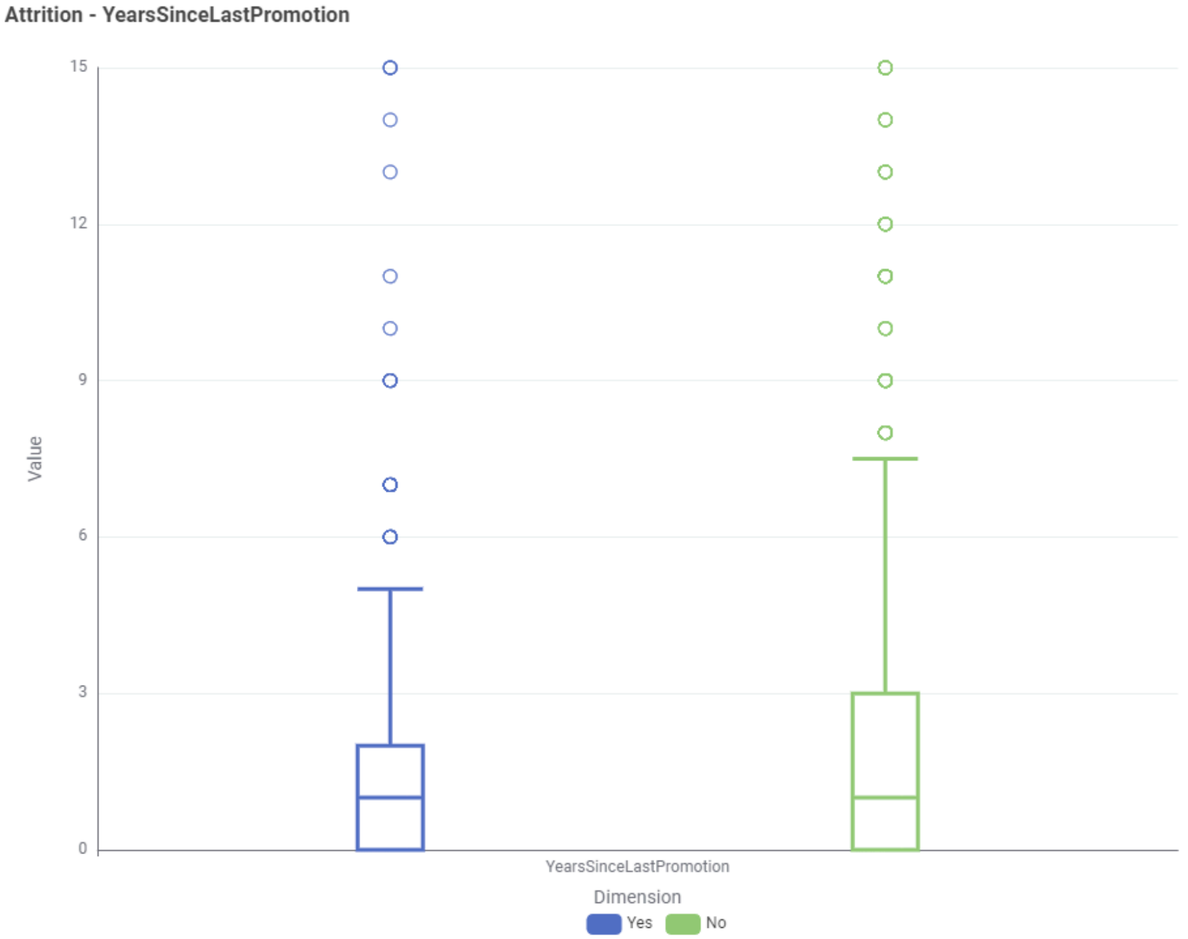


BIVARIATE ANALYSIS

Attrition vs YearsSinceLastPromotion

Insights: employees who left the organization ("Yes" group) generally have a shorter time since their last promotion compared to those who stayed ("No" group), as indicated by the lower median and smaller range in the "Yes" group. This could imply dissatisfaction with recent promotions or shorter tenure periods before leaving.

Test Column *String*	Group *String*	N *Number (integer)*	Mean *Number (double)*	Standard Deviation *Number (double)*
YearsSinceLastPromotion	Yes	198	1.955	3.166
YearsSinceLastPromotion	No	978	2.218	3.252
YearsSinceLastPromotion	Total	1176	2.173	3.238

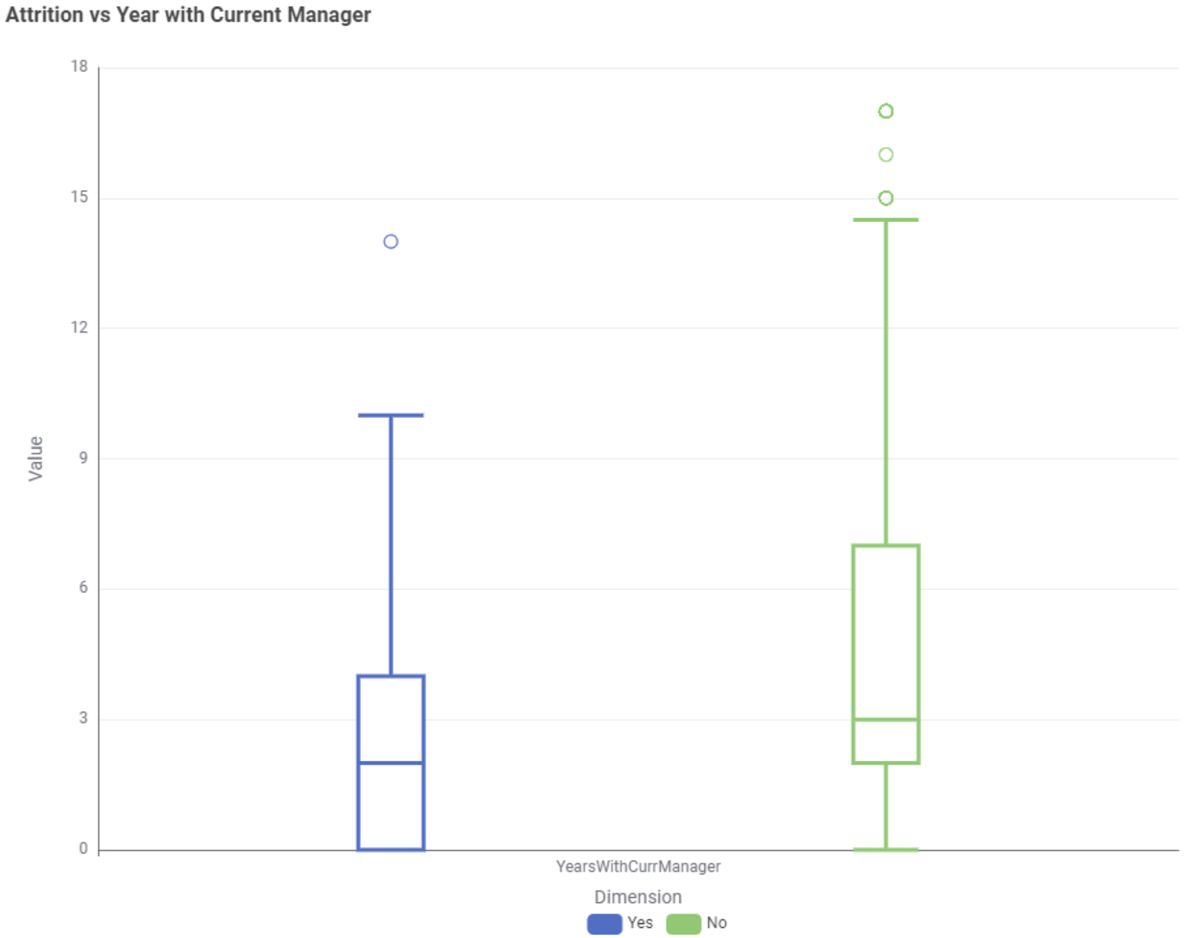


BIVARIATE ANALYSIS

Attrition vs YearsWithCurrManager

Insights: employees who left the organization (attrition = "Yes") have significantly fewer years with their current manager, as compared to those who stayed (attrition = "No"). This suggests a possible correlation between tenure with the current manager and employee retention.

Test Column *String*	Group *String*	N *Number (integer)*	Mean *Number (double)*	Standard Deviation *Number (double)*
YearsWithCurrManager	Yes	198	2.758	3.05
YearsWithCurrManager	No	978	4.365	3.606
YearsWithCurrManager	Total	1176	4.094	3.569

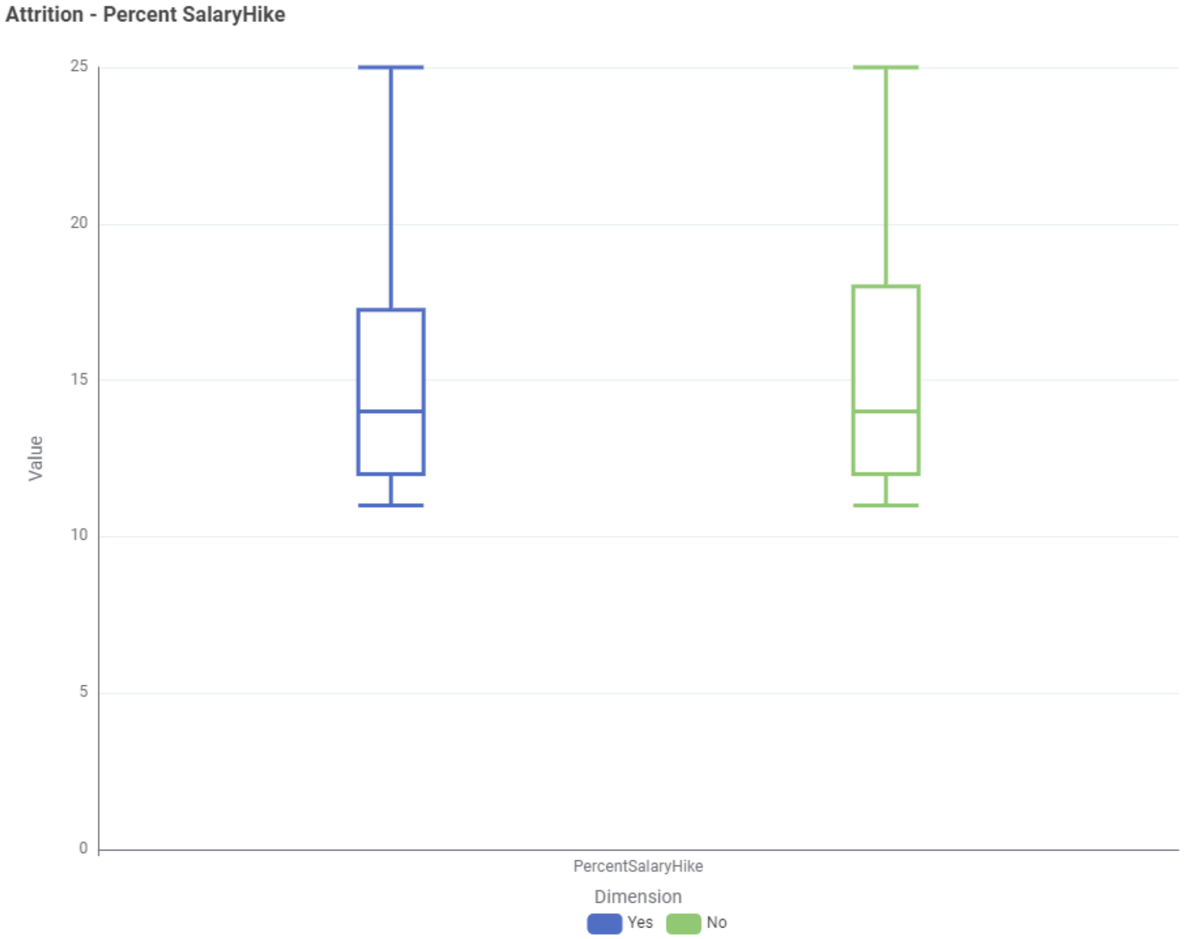


BIVARIATE ANALYSIS

Attrition vs PercentSalaryHike

Insights: the distribution of percentage salary hikes appears similar between employees who stayed and those who left. There is no significant difference in the median or range, suggesting salary hikes may not directly influence attrition.

Test Column *String*	Group *String*	N *Number (integer)*	Mean *Number (double)*	Standard Deviation *Number (double)*
PercentSalaryHike	Yes	198	15.101	3.842
PercentSalaryHike	No	978	15.22	3.62
PercentSalaryHike	Total	1176	15.2	3.657

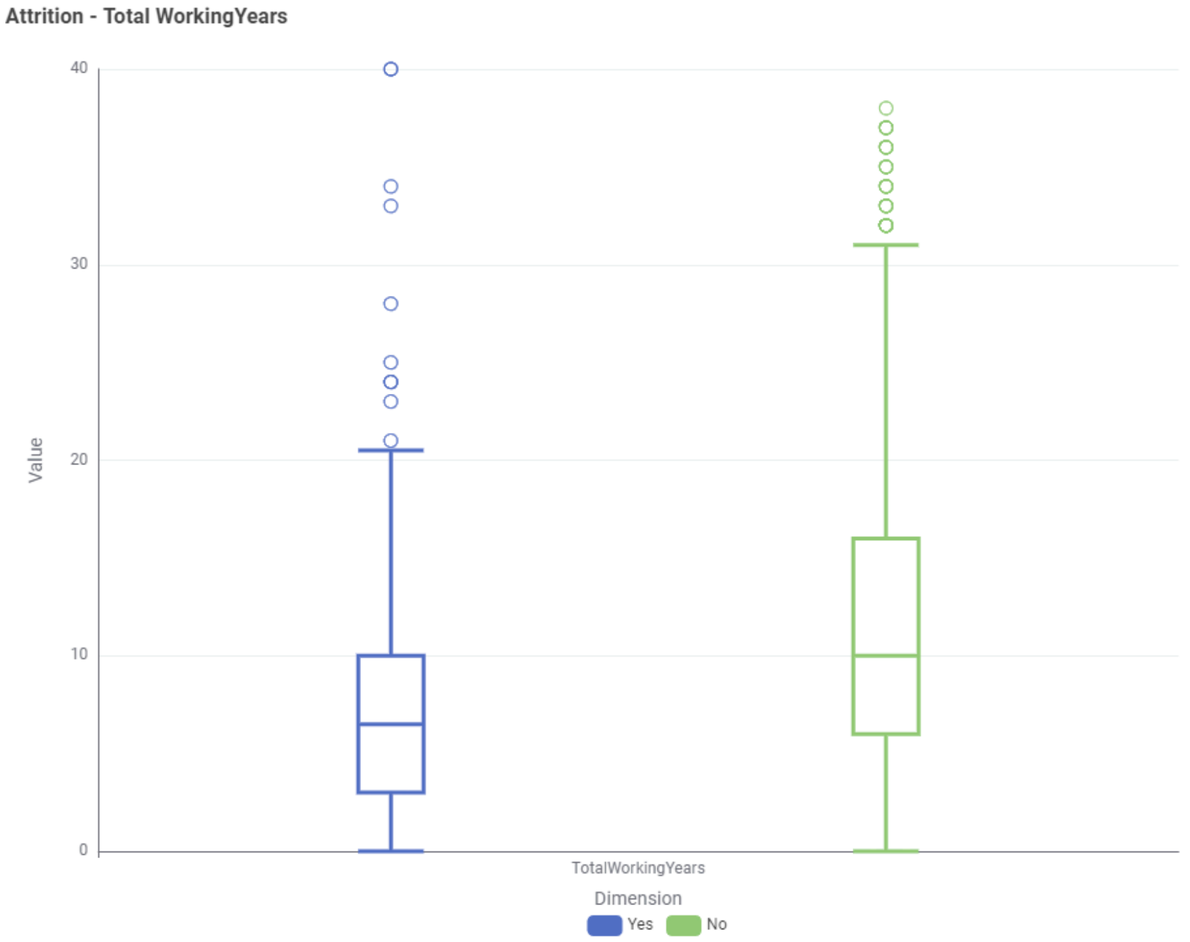


BIVARIATE ANALYSIS

Attrition vs TotalWorkingYears

Insights: employees who left the organization tend to have fewer total working years than those who stayed. The median total working years for attrition cases is significantly lower, indicating that less experienced employees might be more prone to leaving.

Test Column *String*	Group *String*	N *Number (integer)*	Mean *Number (double)*	Standard Deviation *Number (double)*
TotalWorkingYears	Yes	198	7.955	6.977
TotalWorkingYears	No	978	11.952	7.876
TotalWorkingYears	Total	1176	11.279	7.873

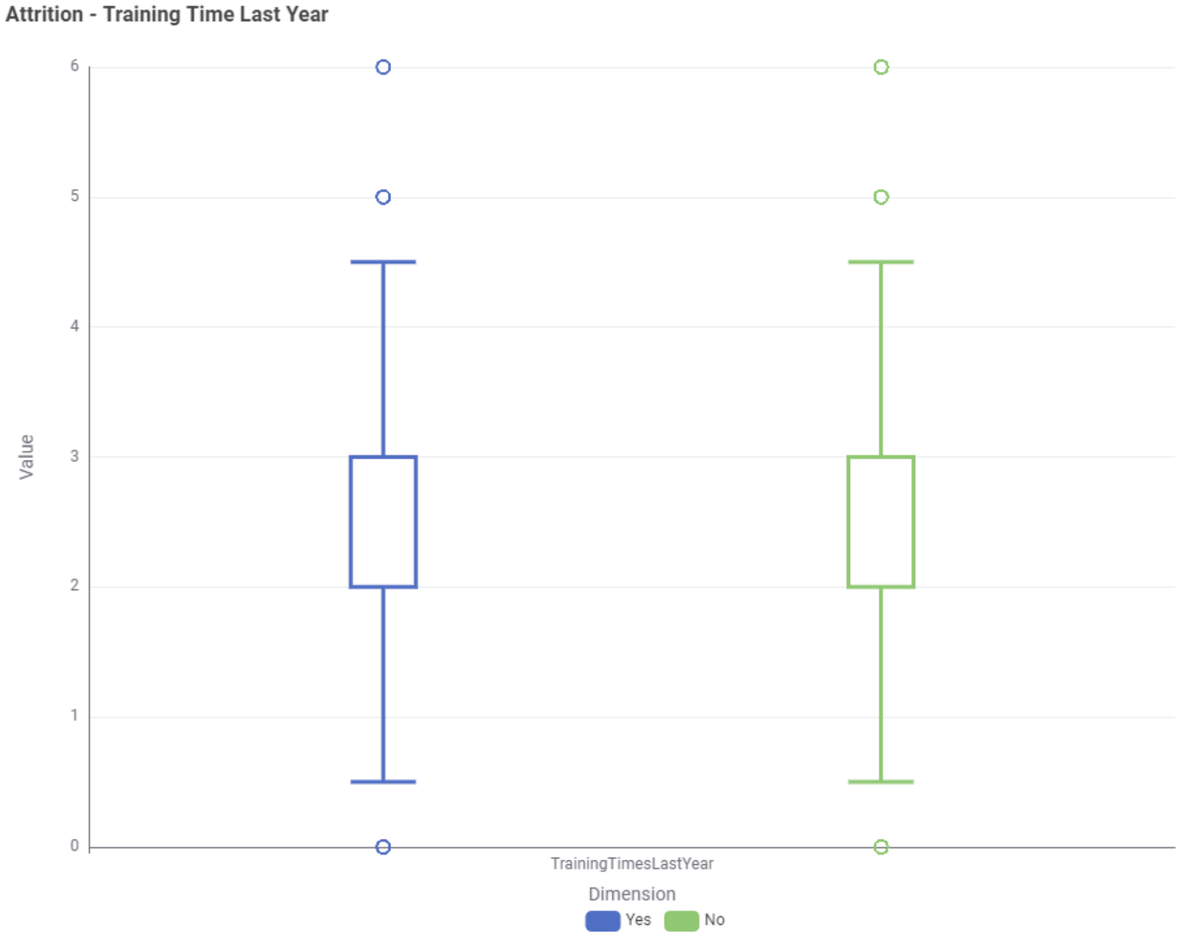


BIVARIATE ANALYSIS

Attrition vs TrainigTimeLastYear

Insights: the median training times for both employees who left and those who stayed are approximately equal, indicating that the amount of training provided last year does not significantly differ between these groups. The variability and outliers, however, suggest some employees received either very low or very high training.

Test Column *String*	Group *String*	N *Number (integer)*	Mean *Number (double)*	Standard Deviation *Number (double)*
TrainingTimesLastYear	Yes	198	2.631	1.302
TrainingTimesLastYear	No	978	2.832	1.304
TrainingTimesLastYear	Total	1176	2.798	1.306

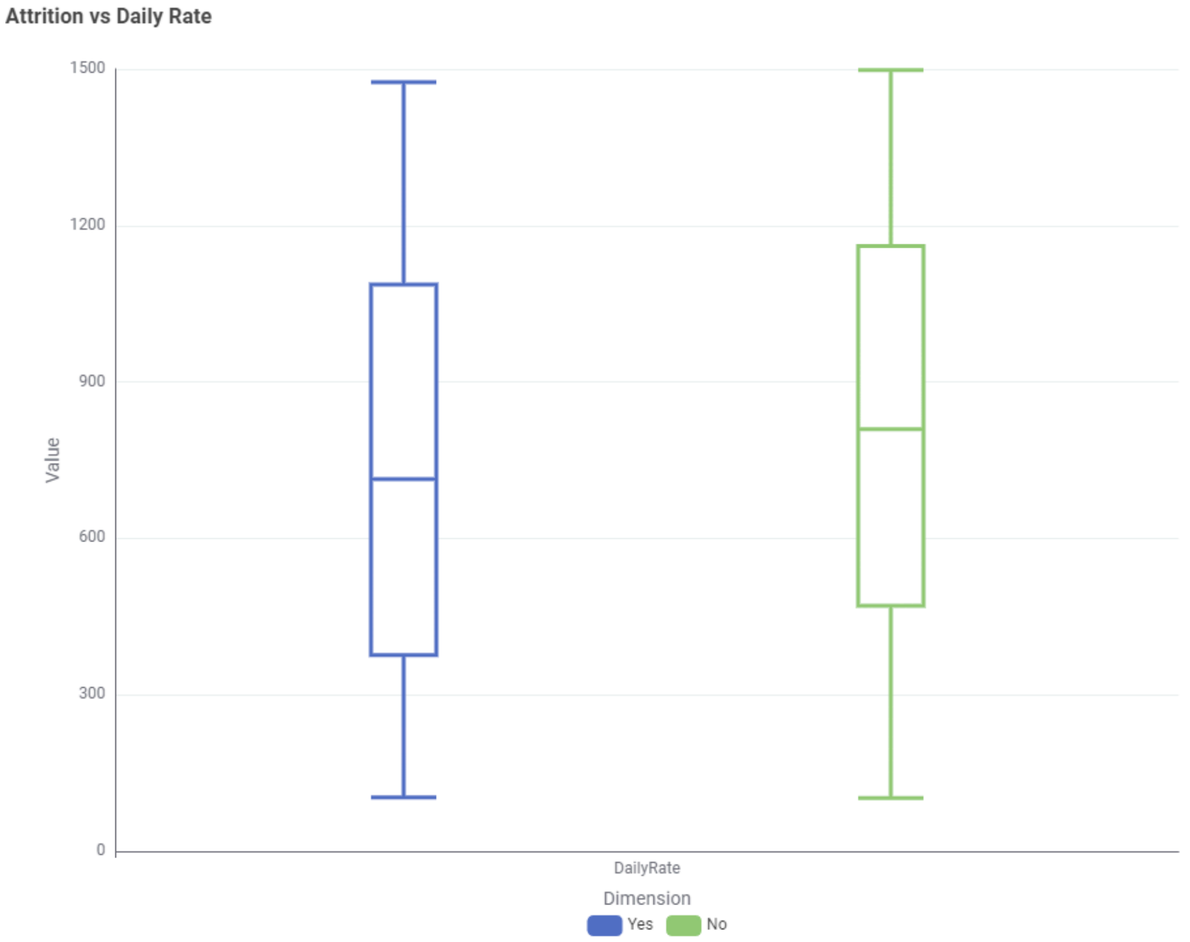


BIVARIATE ANALYSIS

Attrition vs DailyRate

Insights: there is no significant difference in the daily rate between employees who left (attrition) and those who stayed. The distribution and median values for both groups appear similar, suggesting that daily rate may not be a strong predictor of attrition.

Test Column *String*	Group *String*	N *Number (integer)*	Mean *Number (double)*	Standard Deviation *Number (double)*
DailyRate	Yes	198	741.394	397.749
DailyRate	No	978	808.176	405.068
DailyRate	Total	1176	796.932	404.451

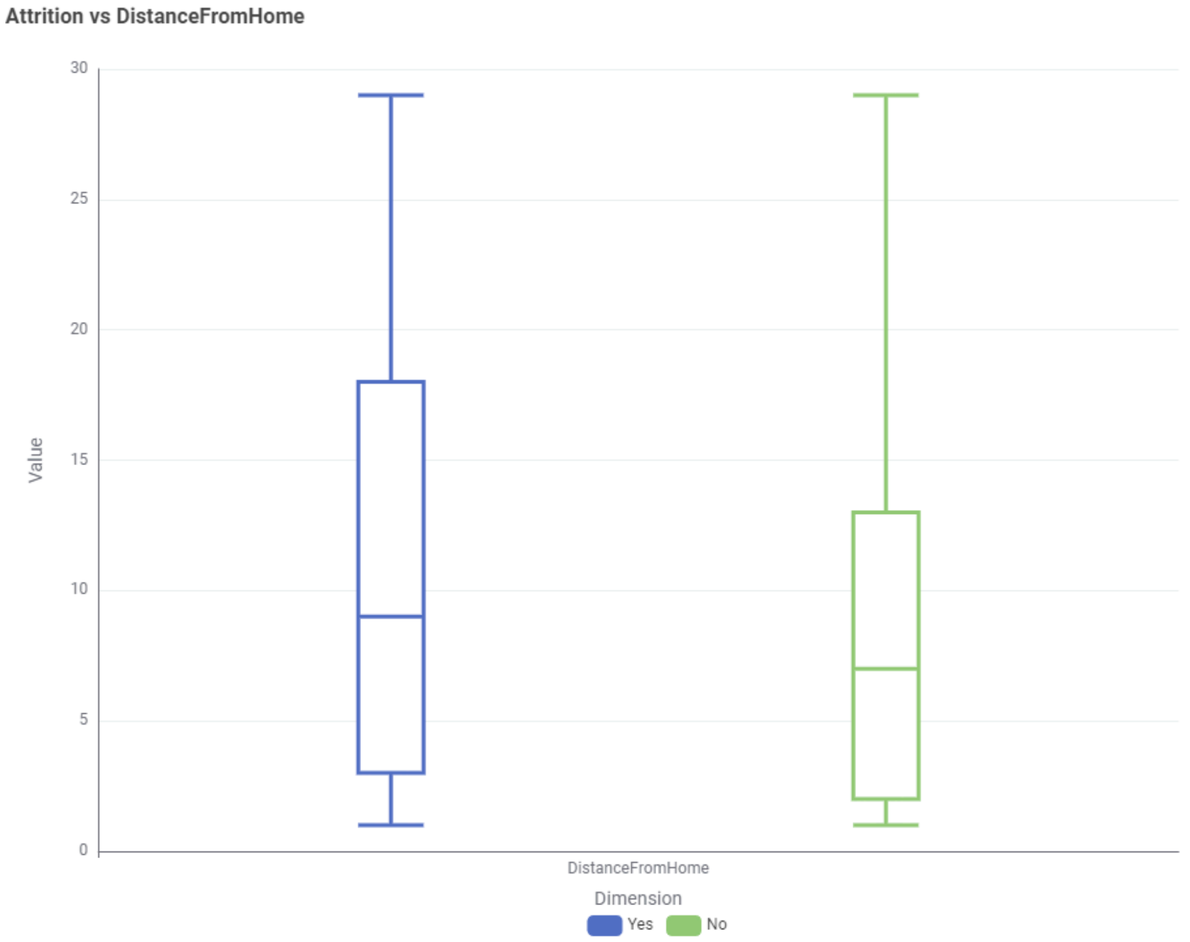


BIVARIATE ANALYSIS

Attrition vs DistanceFromHome

Insights: the distance from home shows a slightly wider distribution for employees who left (attrition), but the medians of both groups are quite close. This indicates that distance from home may have a minor influence on attrition but is not a strong determinant.

Test Column *String*	Group *String*	N *Number (integer)*	Mean *Number (double)*	Standard Deviation *Number (double)*
DistanceFromHome	Yes	198	10.949	8.61
DistanceFromHome	No	978	9.029	8.025
DistanceFromHome	Total	1176	9.352	8.154

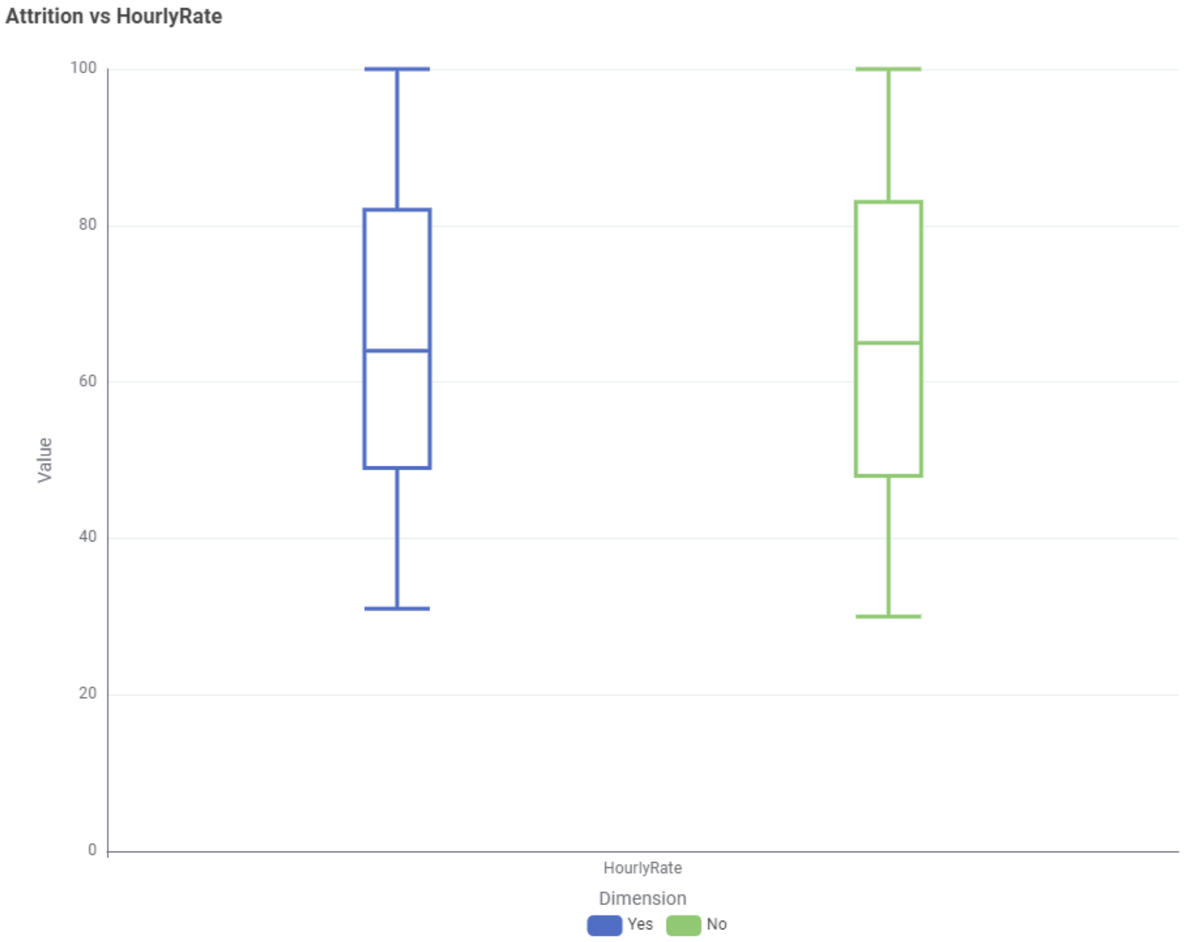


BIVARIATE ANALYSIS

Attrition vs HourlyRate

Insights: the boxplot for hourly rate shows no significant difference between employees who left (attrition) and those who stayed. Both groups have similar distributions, suggesting that hourly rate is unlikely to have a strong direct influence on attrition.

Test Column *String*	Group *String*	N *Number (integer)*	Mean *Number (double)*	Standard Deviation *Number (double)*
HourlyRate	Yes	198	64.914	19.99
HourlyRate	No	978	65.377	20.319
HourlyRate	Total	1176	65.299	20.256

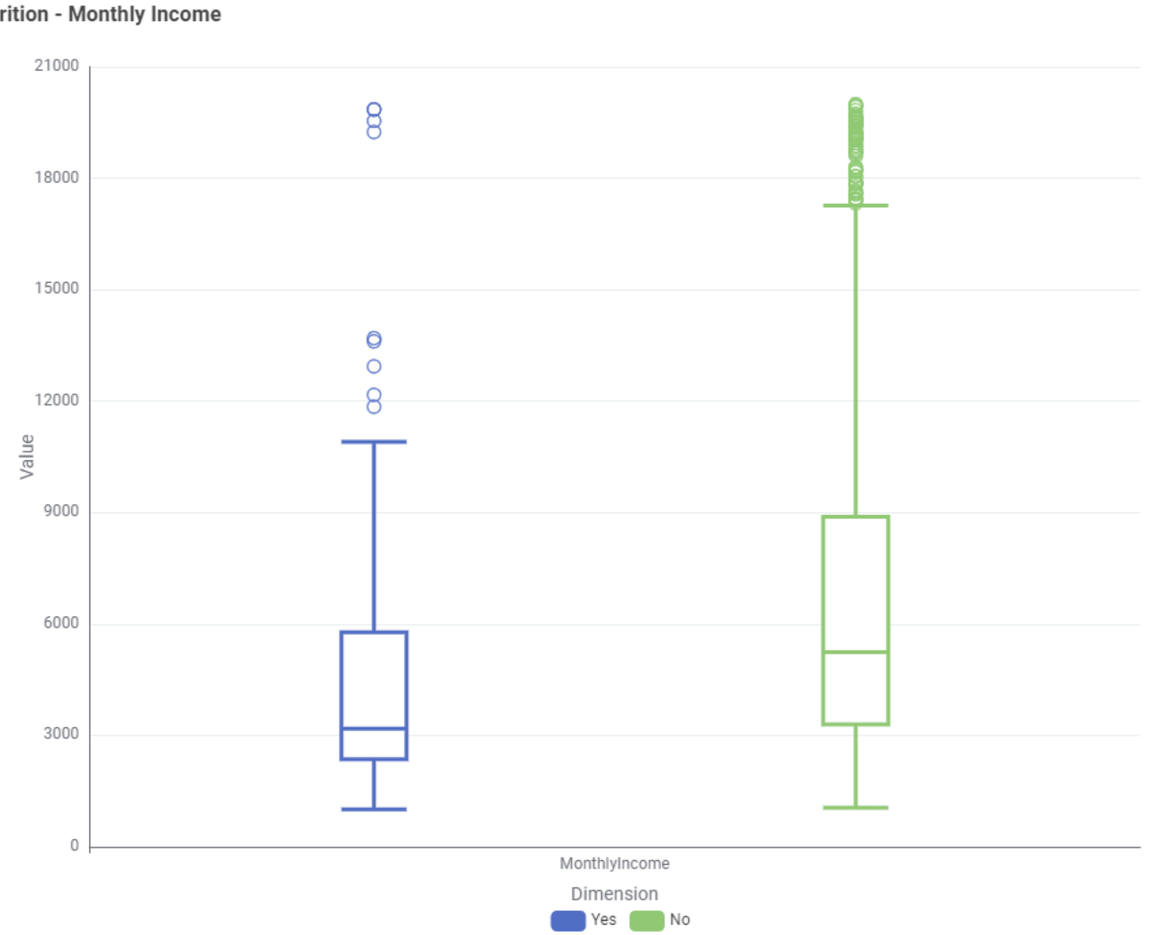


BIVARIATE ANALYSIS

Attrition vs MonthlyIncome

Insights: employees who left (Attrition = Yes) tend to have lower monthly incomes compared to those who stayed (Attrition = No). The median income is lower for the attrition group, and the spread of incomes is narrower, indicating that higher earners are less likely to leave.

Test Column *String*	Group *String*	N *Number (integer)*	Mean *Number (double)*	Standard Deviation *Number (double)*
MonthlyIncome	Yes	198	4,642.253	3,517.996
MonthlyIncome	No	978	6,897.284	4,853.687
MonthlyIncome	Total	1176	6,517.611	4,730.338



DATA CLEANING

01 REDUNDANT COLUMNS

02 MISSING VALUES

03 OUTLIERS

01 REDUNDANT COLUMNS

REDUNDANT COLUMNS

In displaying the “Statistics View” of our dataset, and focusing on unique values of each feature, we suddenly noticed that 3 variables presented no variation in their values. These are:

- “EmployeeCount”: possibly the number of employees referred to each entry information, and this case are all ones, signaling only one employee information for each observation
- “Over18”: whether the individual is over 18 years old or not, in this case everyone is
- “StandardHours”: could be the number of hours worked for each employee by contract, and for this instance, everyone has the same amount of contracted hours (80).

Name	Type	# Unique values ↑
EmployeeCount	Number (integer)	1
Over18	String	1
StandardHours	Number (integer)	1

02 MISSING VALUES

MISSING VALUES

Name	# Missing values
Age	0
Attrition	0
BusinessTravel	0
DailyRate	0
Department	0
DistanceFromHome	0
Education	0
EducationField	0
EmployeeCount	0
EmployeeNumber	0
EnvironmentSatisfaction	0
Gender	0

Name	# Missing values
HourlyRate	0
JobInvolvement	0
JobLevel	0
JobRole	0
JobSatisfaction	0
MaritalStatus	0
MonthlyIncome	0
MonthlyRate	0
NumCompaniesWorked	0
Over18	0
OverTime	0
PercentSalaryHike	0

Name	# Missing values
PerformanceRating	0
RelationshipSatisfaction	0
StandardHours	0
StockOptionLevel	0
TotalWorkingYears	0
TrainingTimesLastYear	0
WorkLifeBalance	0
YearsAtCompany	0
YearsInCurrentRole	0
YearsSinceLastPromotion	0
YearsWithCurrManager	0

Treating missing values during data preparation is crucial to prevent biases and improve the reliability and performance of models. While performing univariate analysis, with the “Statistics” node we counted the number of missing values for each variable and observed that there are none.

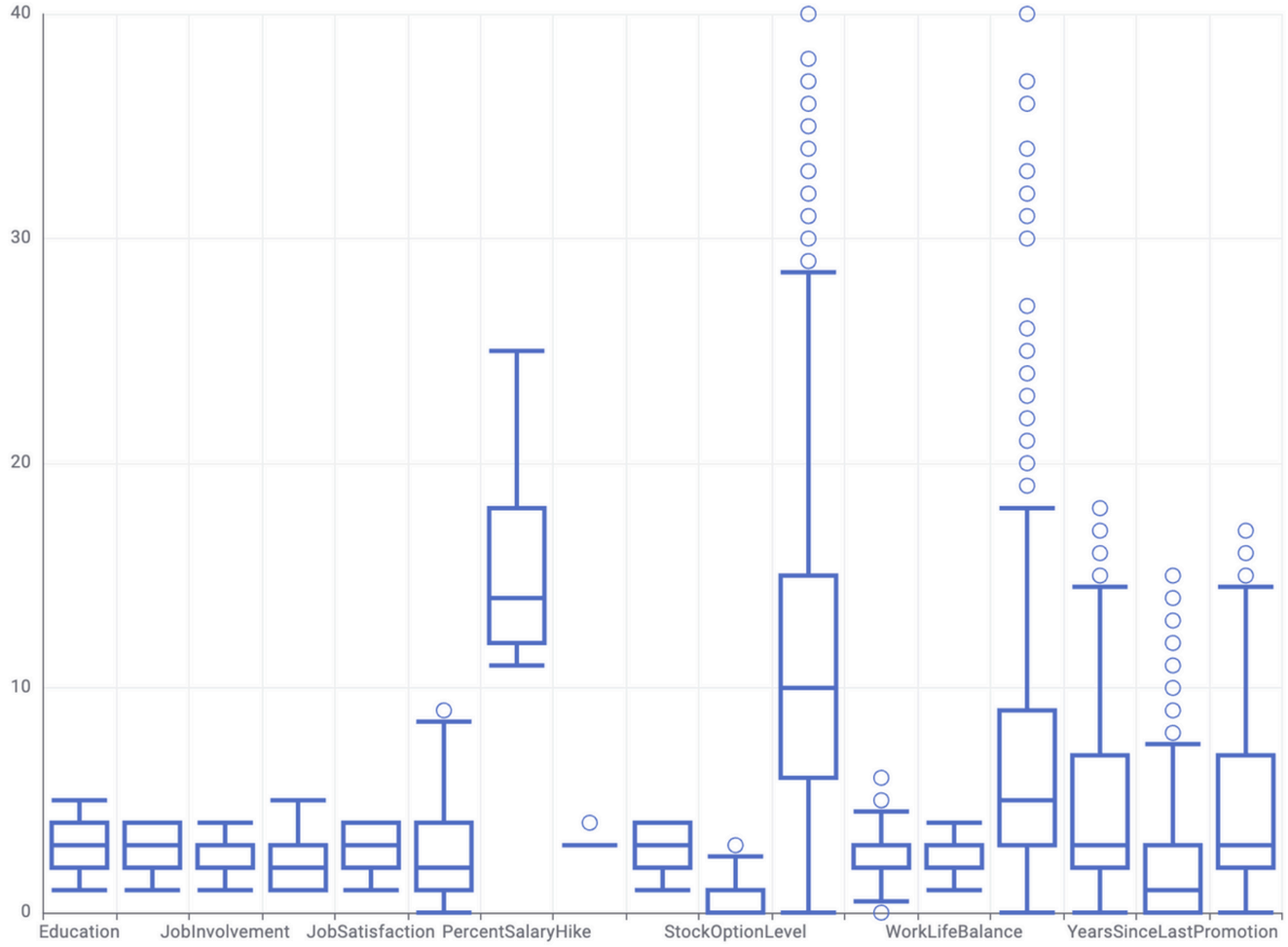
Statistics View

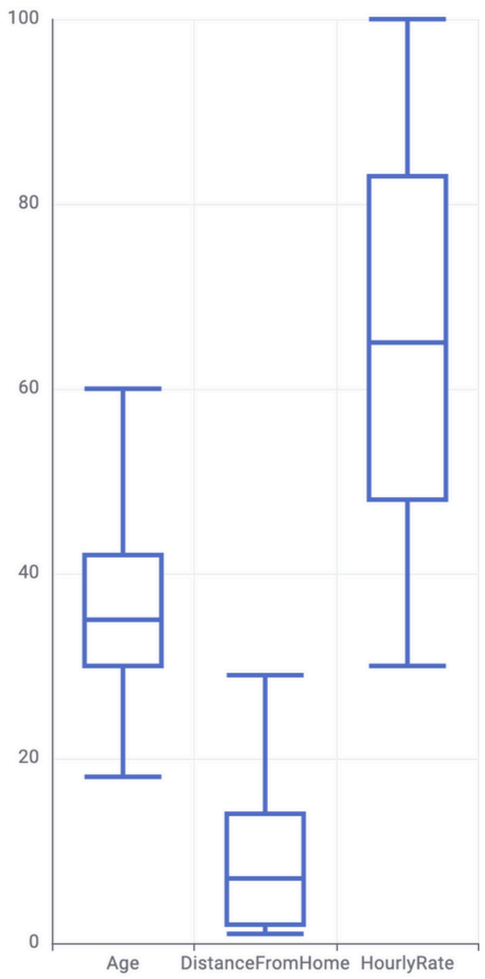
03 OUTLIERS

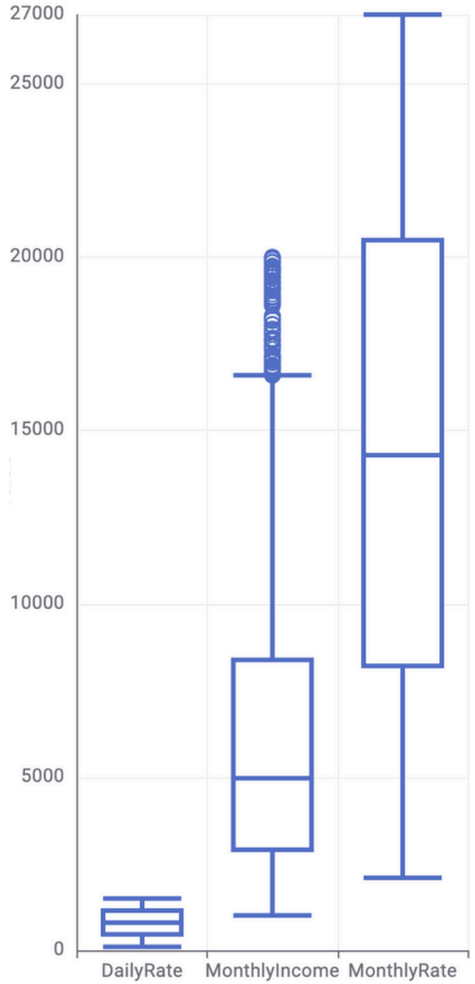
OUTLIERS

We consider as outliers values that are beyond 1.5 times the Inter Quartile Range.

To visually identify them, we analyse the Box Plots:







OUTLIERS

Once we identified which variables presented outliers, we used the Numeric Outliers node to count the number of outliers for each of these variables

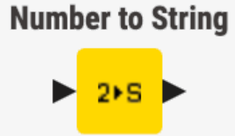
Outlier column *String*	Member count *Number (integer)*	Outlier count *Number (integer)*	Lower bound *Number (double)*	Upper bound *Number (double)*
MonthlyIncome	1176	99	-4,785	15,583
NumCompaniesWorked	1176	38	-3.5	8.5
PerformanceRating	1176	178	3	3
StockOptionLevel	1176	67	-1.5	2.5
TotalWorkingYears	1176	51	-7.5	28.5
TrainingTimesLastYear	1176	200	0.5	4.5
YearsAtCompany	1176	82	-6	18
YearsInCurrentRole	1176	17	-5.5	14.5
YearsSinceLastPromotion	1176	170	-3	5
YearsWithCurrManager	1176	11	-5.5	14.5

OUTLIERS

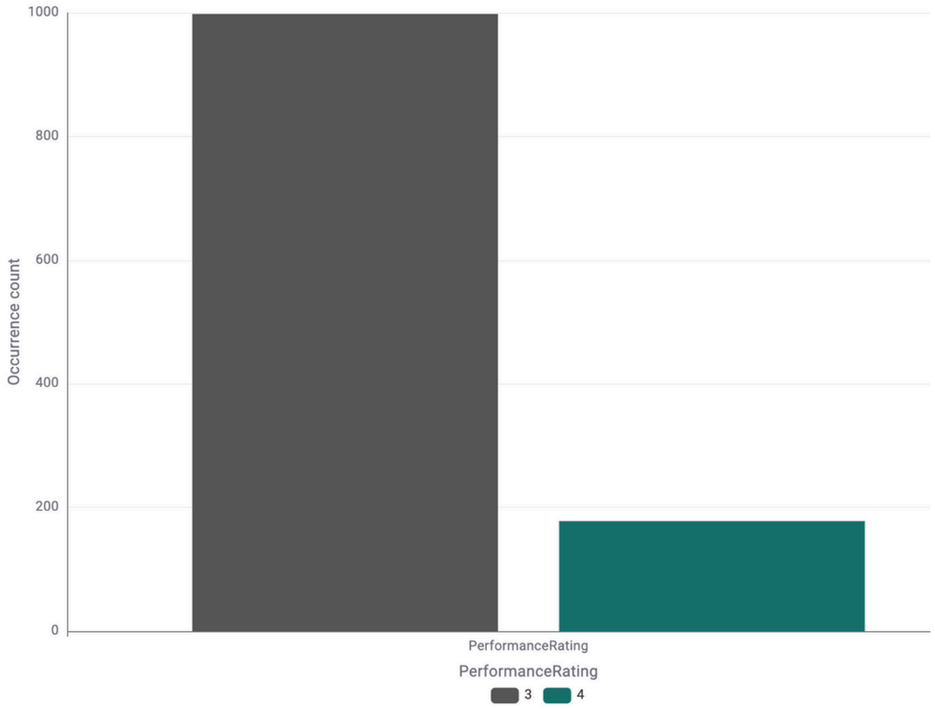
Performance Ratings and Stock Option Level

We then proceeded to treat outliers

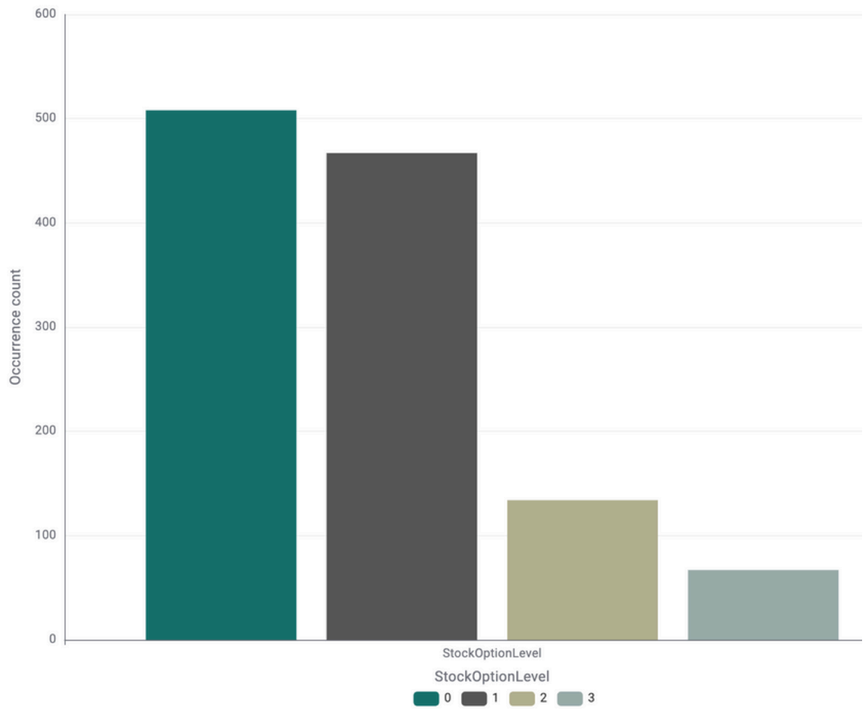
Performance Ratings and Stock Option Level already were categorical variables with a small number of discrete levels. Observations were concentrated on the low end of the scale, as it is reasonable when considering how companies typically function. We therefore decided not to treat them, and decided to transform them as categorical variables, using the”Number to String” node.



Performance Ratings

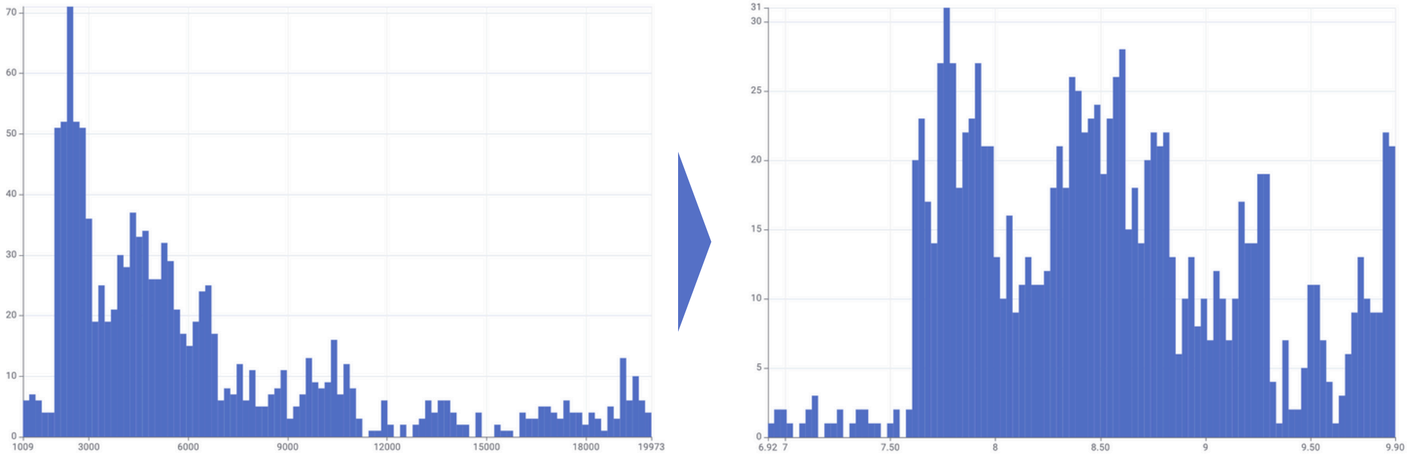


Stock Option Level

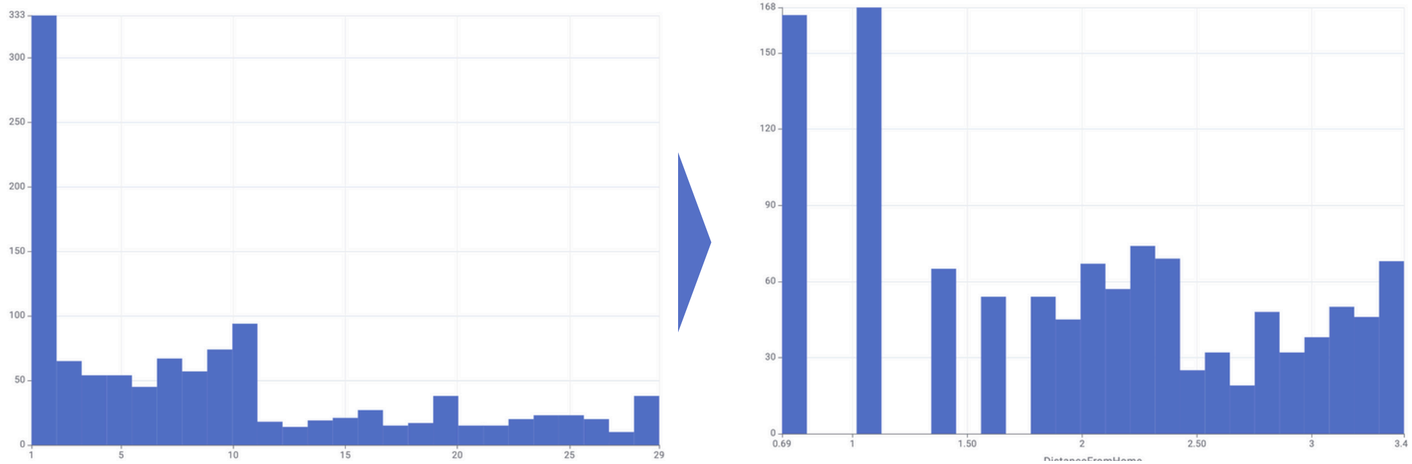


OUTLIERS

Monthly Income



Distance from Home



Monthly Income and Distance from Home

During data analysis, we observed that Monthly Income exhibited a significantly right-skewed distribution. To address both outliers and skewness, we applied a logarithmic transformation, which balances the distribution and aligns it with the assumptions of many machine learning models, thereby improving their performance and interpretability. For the same reason, we also applied this transformation to Distance from Home, which, although not affected by outliers, also displayed a right-skewed distribution.

To do so, we used the “Math Formula (Multi Column)” node

Since all observations for both variables were positive, the transformations could be applied without the risk of generating missing values or errors.

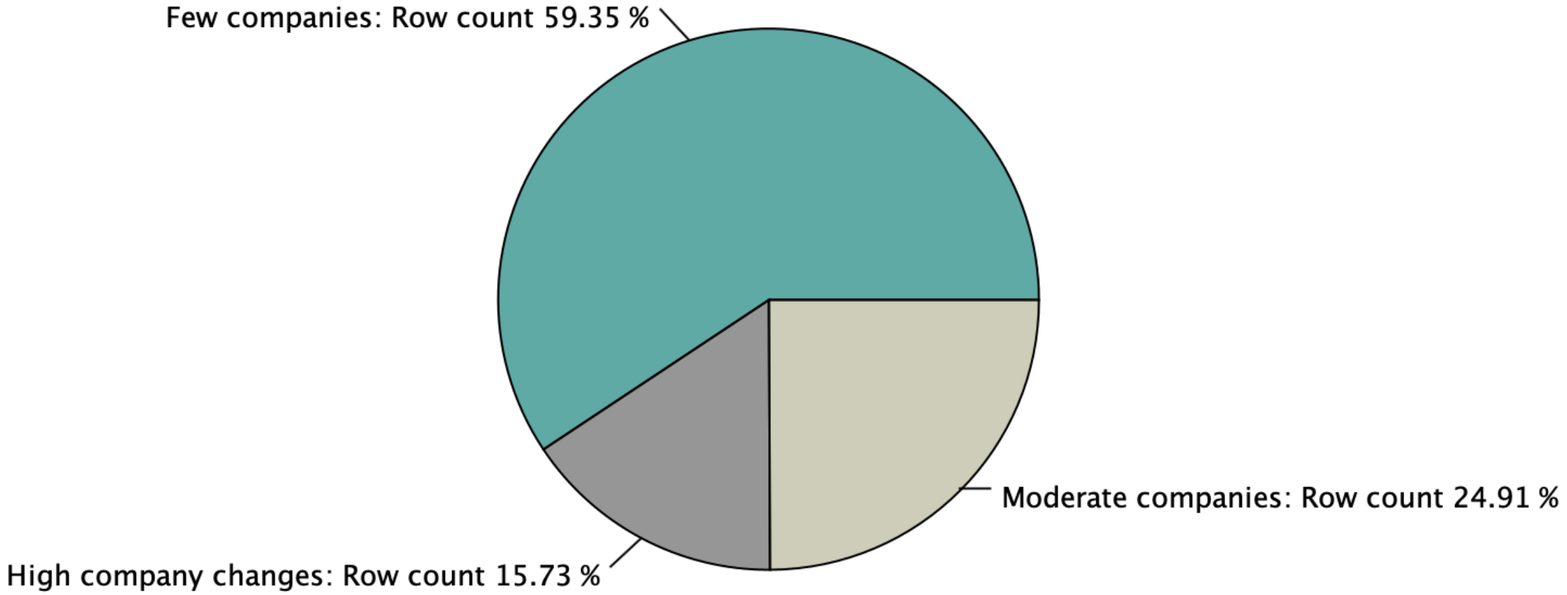
OUTLIERS

Since the remaining variables all presented an ordinal structure, we opted to transform them in categorical variables, to both reduce impact of outliers and improve interpretability. The transformations were performed using the “Numeric Binner” node.

Number Companies Worked

As we observed during the data analysis, the variable presented a right-skewed distributions and assumed vaues between 1 and 9. We decided to bin observations in three categories, to capture meaningful career transitions:

```
Few companies : ] −∞ ... 3.0 [
Moderate companies : [ 3.0 ... 6.0 [
High company changes : [ 6.0 ... ∞ [
```



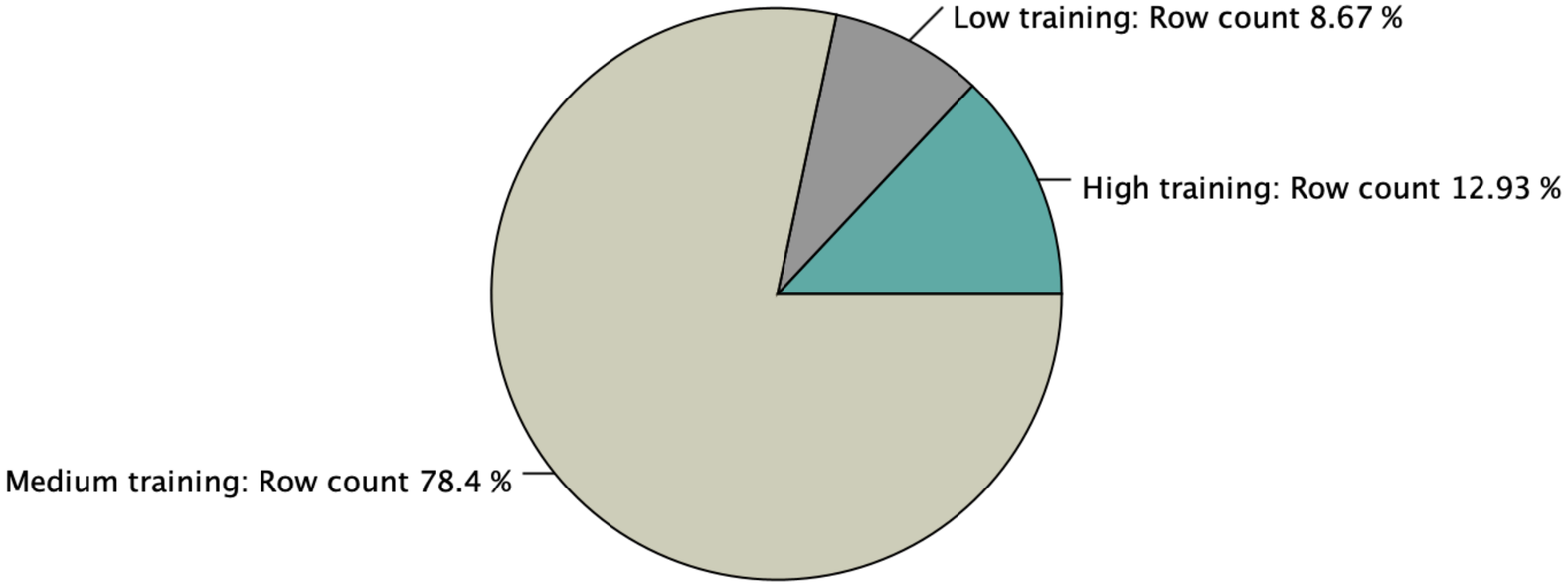
OUTLIERS

Since the remaining variables all presented an ordinal structure, we opted to transform them in categorical variables, to both reduce impact of outliers and improve interpretability. The transformations were performed using the “Numeric Binner” node.

Training Times Last Year

For the variable Training Times Last Year observations range between 0 and 6 and are concentrated in the center, around 2 and 3. We grouped them in three categories:

```
Low training : ] −∞ ... 2.0 [
Medium training : [ 2.0 ... 5.0 [
High training : [ 5.0 ... ∞ [
```



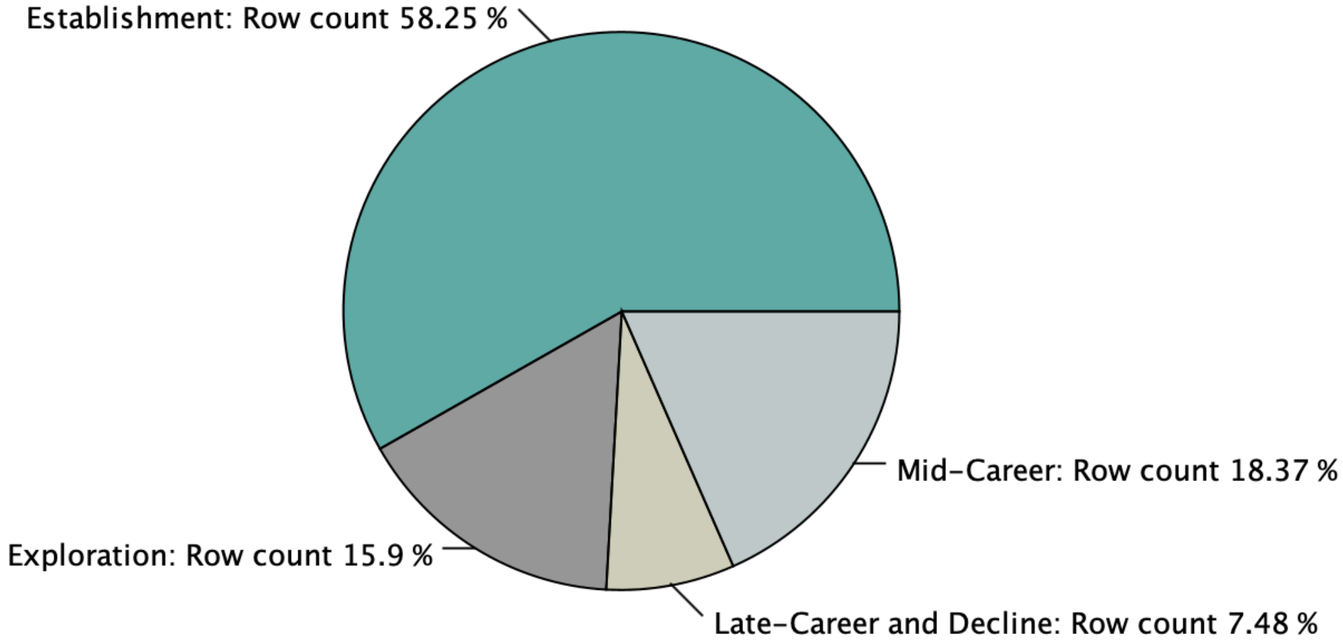
OUTLIERS

Since the remaining variables all presented an ordinal structure, we opted to transform them in categorical variables, to both reduce impact of outliers and improve interpretability. The transformations were performed using the “Numeric Binner” node.

Total Working Years

The variable has a right-skewed distributions with values ranging from 0 to 40. The categorisation we applied was based on the five stages frequently used to descrive careers. However, to account for the skewness of the variable, we grouped together the last two stages, Late Career and Decline:

```
Exploration : ] -∞ ... 5.0 [
Establishment : [ 5.0 ... 15.0 [
Mid-Career : [ 15.0 ... 25.0 [
Late-Career and Decline : [ 25.0 ... ∞ [
```



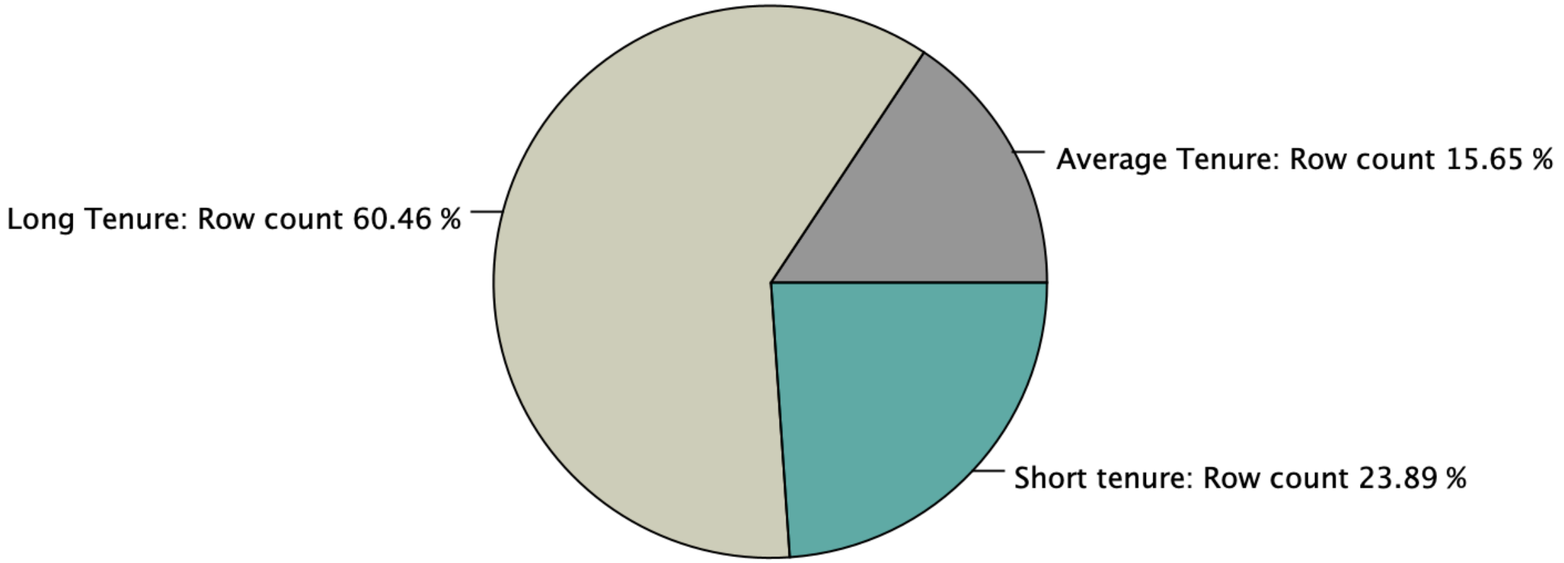
OUTLIERS

Since the remaining variables all presented an ordinal structure, we opted to transform them in categorical variables, to both reduce impact of outliers and improve interpretability. The transformations were performed using the “Numeric Binner” node.

Years at Company

The variable has a right-skewed distributions with values ranging from 0 to 40. To group observations, we considered that the average tenure for an employee in the IT sector is 3-4 years, therefore we defined the following categories:

```
Short tenure : ] −∞ ... 3.0 [
Average Tenure : [ 3.0 ... 5.0 [
Long Tenure : [ 5.0 ... ∞ [
```



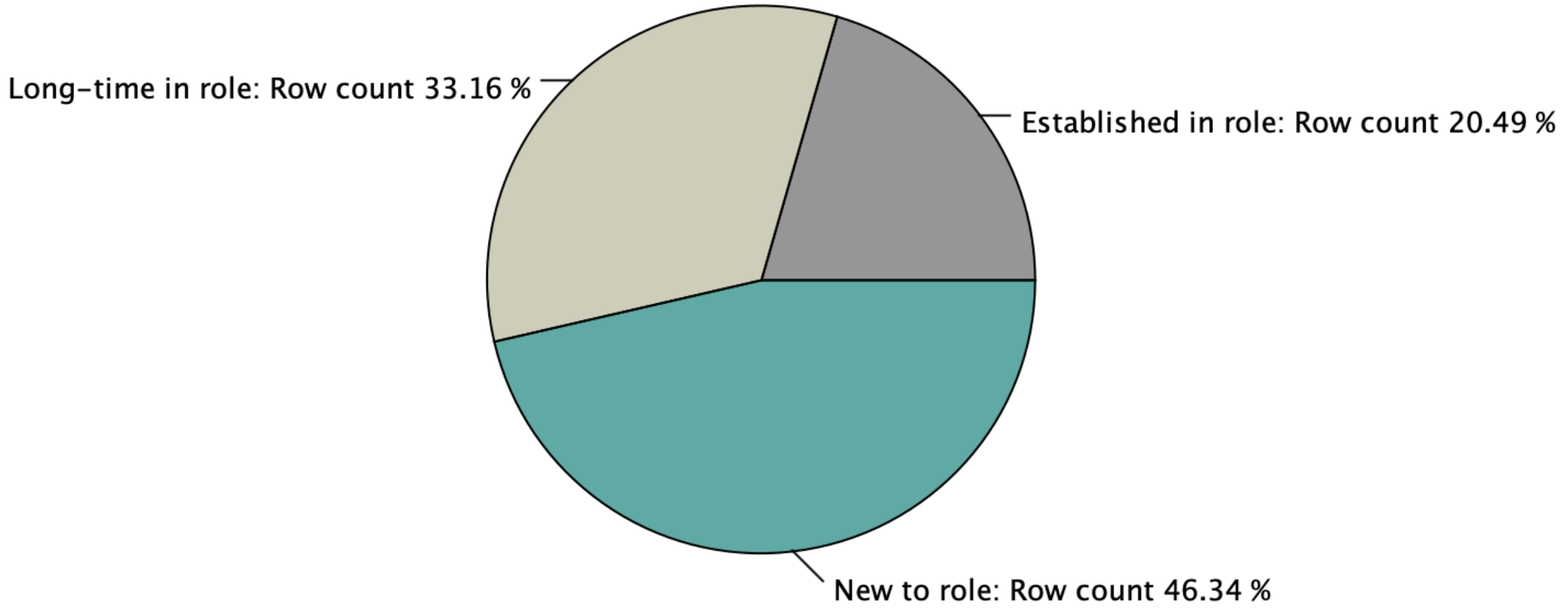
OUTLIERS

Since the remaining variables all presented an ordinal structure, we opted to transform them in categorical variables, to both reduce impact of outliers and improve interpretability. The transformations were performed using the “Numeric Binner” node.

Years in Current Role

In the sample Years in Current Role takes value between 0 and 18, with a distribution concentrated on the left-hand side. To capture role stability and progression, the following bins were adopted for categorisation:

```
New to role : ] −∞ ... 3.0 [
Established in role : [ 3.0 ... 7.0 [
Long-time in role : [ 7.0 ... ∞ [
```



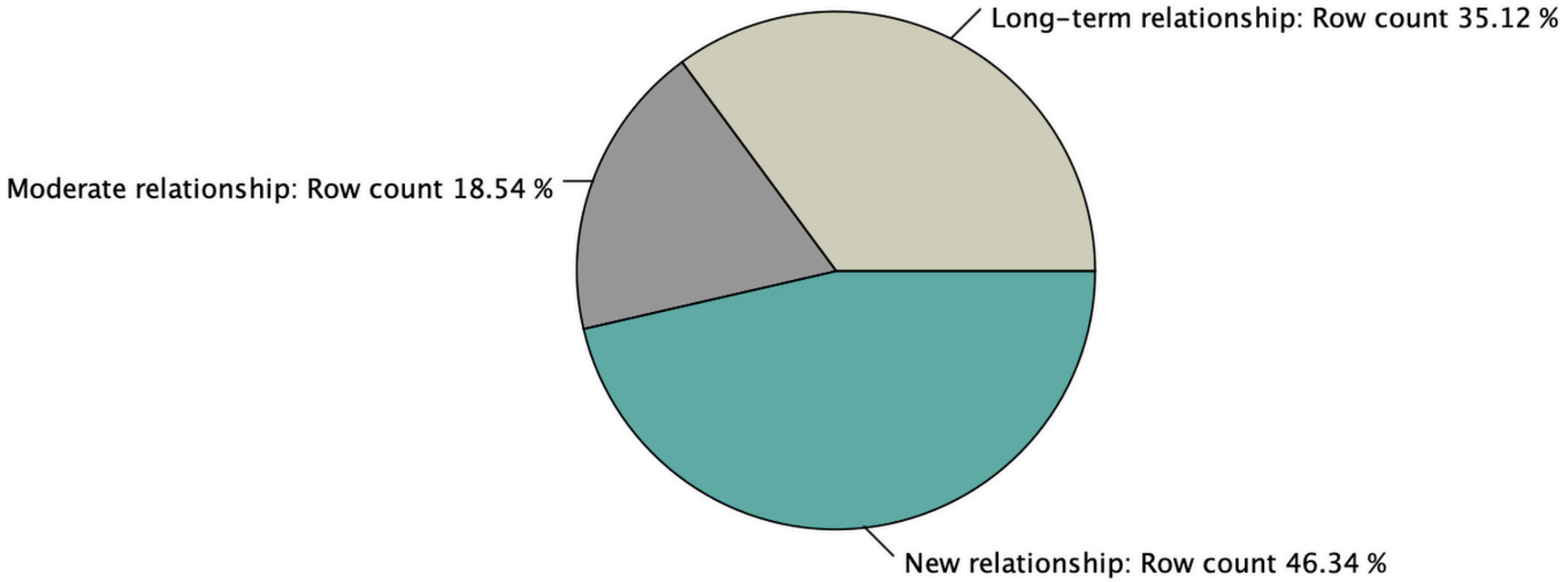
OUTLIERS

Since the remaining variables all presented an ordinal structure, we opted to transform them in categorical variables, to both reduce impact of outliers and improve interpretability. The transformations were performed using the “Numeric Binner” node.

Years with Current Manager

Years with Current Manager shows a heavily right-skewed distribution and a range extending from 0 to 15. Observations were grouped in the following categories:

```
New relationship : ] −∞ ... 3.0 [
Moderate relationship : [ 3.0 ... 6.0 [
Long-term relationship : [ 6.0 ... ∞ [
```



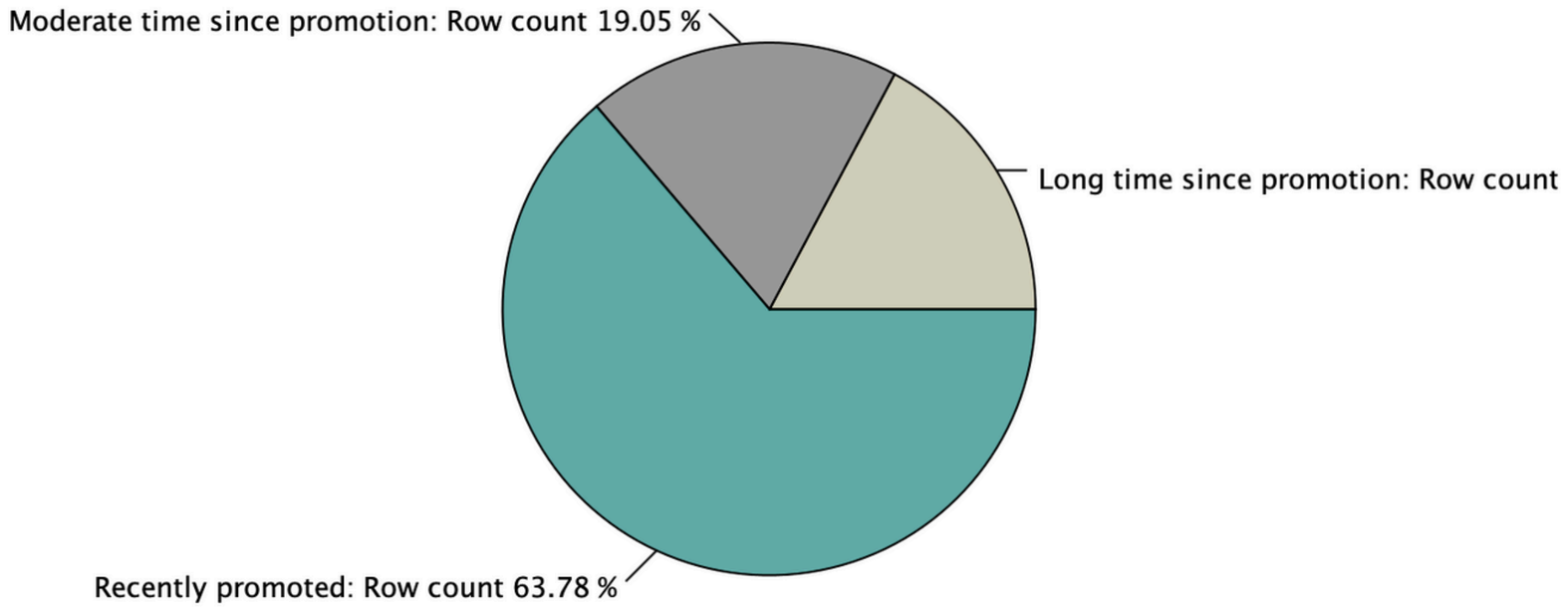
OUTLIERS

Since the remaining variables all presented an ordinal structure, we opted to transform them in categorical variables, to both reduce impact of outliers and improve interpretability. The transformations were performed using the “Numeric Binner” node.

Years since Last Promotion

For this variable, the distribution is bimodal with peaks around 2–3 and 6–7, a right-skewed tail, and values ranging from 0 to approximately 17. Taking into consideration that the average time between promotions in IT companies it 2 to 4 years, we grouped data in 3 categories:

```
Recently promoted : ] −∞ ... 2.0 [
Moderate time since promotion : [ 2.0 ... 5.0 [
Long time since promotion : [ 5.0 ... ∞ [
```



OUTLIERS

Alternative handling methods

There are a few other ways in which we could have treated outliers

Elimination of observations

Replacing them with other values

such as the mean, the median or the mode

Treating them as a separate group

However, we determined that variable transformation was the most appropriate method for our dataset due to its specific characteristics. The dataset, with a moderate number of observations, was already clean and free from inconsistencies. The identified outliers did not appear to be errors but rather **natural deviations** that could be logically explained by typical company dynamics. Given this context, transforming the variables allowed us to address the influence of outliers without compromising their inherent meaning or the integrity of the data.

DATA PRE-PROCESSING

01 FEATURE ENGINEERING

02 ENCODING

03 RESAMPLING

04 PARTITIONING

01 FEATURE ENGINEERING

FEATURE ENGENEERING

ENCODING

Two variables, “OverTime” and “Attrition”, the target variable, presented a Yes/No structure. Using a “Rule Engine” node, we transformed them in binary variables, with 1 representing yes and 0 representing no.

```
$OverTime$ = "Yes" => 1
$OverTime$ = "No" => 0
```

```
$Attrition$ = "Yes" => 1
$Attrition$ = "No" => 0
```

Similarly, using a “Rule Engine” node, we introduced a new binary variable “Female”, assuming value 1 for all observations where variable “Gender” was equal to Female, and 0 for Male. We then dropped the “Gender” column, that had beconme superfluous.

```
$Gender$ = "Female" => 1
$Gender$ = "Male" => 0
```

FEATURE ENGENEERING

INTERACTION TERMS

We enhanced the model's predictive power by adding interaction terms to capture complex relationships between features. Using the “Math Formula” Node, we added two terms:

- **OverTime * JobSatisfaction:** captures how job satisfaction moderates the impact of overtime on attrition, as dissatisfaction may amplify the negative impact of overtime on retention
- **Female * EnvironmentSatisfaction**: highlights gender-specific effects of environment satisfaction on attrition.

NEW VARIABLES

Using the same node, we also introduced two new features, in order to capture additional insights and patterns not directly evident in the original variables:

- **OverallSatisfaction:** computed as the average of JobSatisfaction, EnvironmentSatisfaction and RelationshipSatisfaction rounded to the nearest integer, aims to provide a holistic view of an employee’s overall satisfaction level, which could be strongly related to attrition.
- **CommuteStress**: computed as DistanceFromHome*(5 - JobSatisfaction), reflects the stress of commuting relative to how satisfied an employee is with their job. Employees with long commutes and low job satisfaction may be more likely to leave.

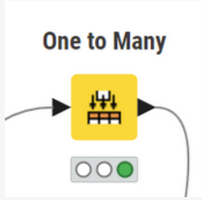
02 ENCODING

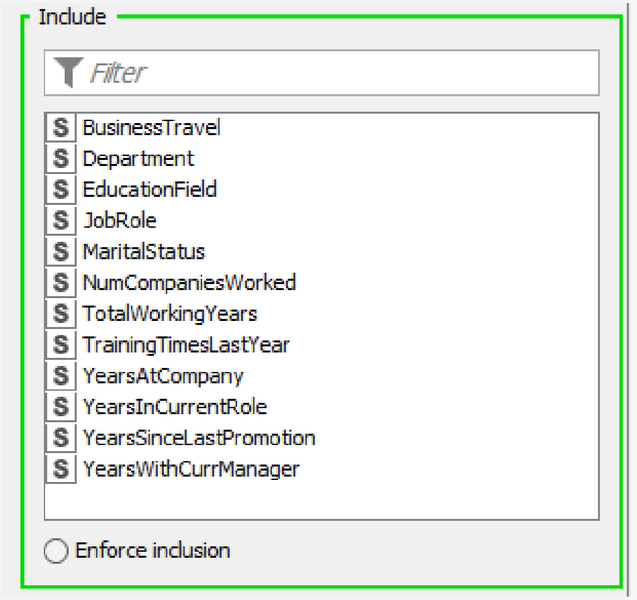
VARIABLE ENCODING

PRE-PROCESSING

Most of Machine Learning models take as inputs numerical variables, posing the problem of categorical significant features for our prediction task. The usual approach is to encode those variables with **One Hot Encoding technique**, which consists in creating as many columns as the unique values of the categorical column and for each observation assign 1 to the category the individual is from, while leaving zeros to the others.
Some of the variables are **already encoded**, like education, job involvement and job level, while others need to be taken into account with One Hot Encoding, specifically with the **“One to Many”** node.

Many models do it ineherntly but we decided to **make preprocessing steps as explicit as possible**: crucial when evaluating and debugging projects like this.





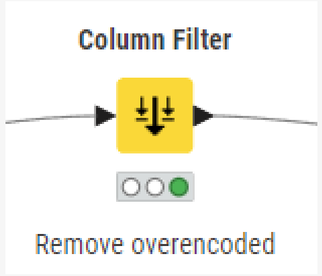
Categorical variables encoded

VARIABLE ENCODING

PRE-PROCESSING

One thing to be careful about is that knime one hot encoding, **no column is dropped from the newly created**. This poses problems of **multicollinearity** in our design matrix, generating invertibility problems for models that require invertibility of the matrix (e.g. logistic regression with closed form estimation). Even in models that do not require the full rank of the matrix, the redundancy in the one hot encoding could lead to problems in *convergence or efficiency* of the training. To solve this issue we can just drop one column (corresponding to a specific category) for each categorical one hot encoded feature.

Dropping one category avoids this problem while preserving all essential information for modeling. We decided to drop the **most standard category** for each categorical feature. We can do it simply by filtering the columns of our dataset with the **“Filter columns”** node.



Values dropped for each categorical variable

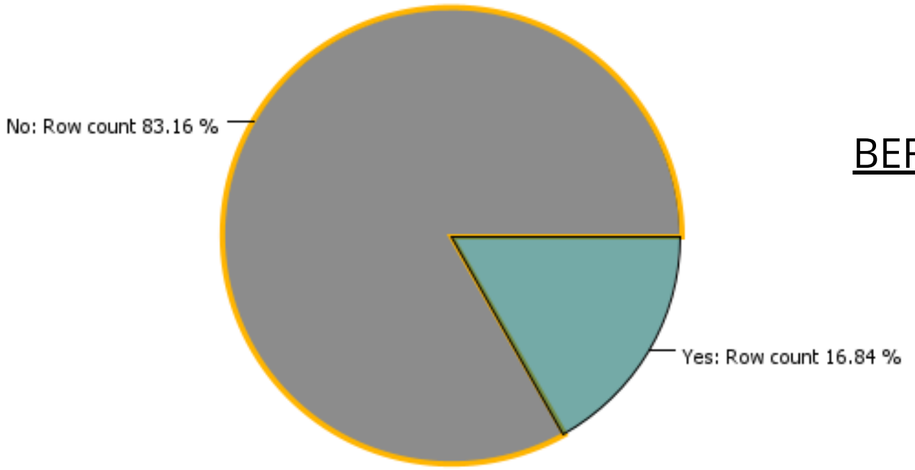
- BusinessTravel → NonTravel
- Department → R&D
- EducationField → Other
- JobRole → Sales Executive
- MaritalStatus → Single
- NumCompaniesWorked → FewCompanies
- TotalWorkingYears → Exploration
- TrainingTimesLastYear → Medium
- YearsAtCompany → Average Tenure
- YearsInCurrentRole → Established
- YearsSinceLastPromotion → Moderate time
- YearsWithCurrManager → Moderate relationship

03 RESAMPLING

RESAMPLING

PRE-PROCESSING

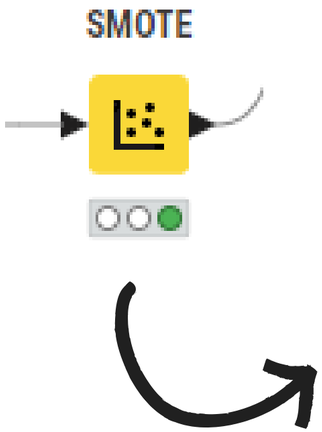
In binary classification a common problem is **target class imbalance**. This imbalance can lead to biased models that perform well on the majority class but **poorly on the minority class**, as standard machine learning algorithms tend to optimize for overall accuracy rather than balanced performance across classes. In an employee attrition prediction project, this issue is critical because attrition cases (employees leaving) are typically much fewer than retention cases. An imbalanced dataset might cause the model to overlook the minority class (attrition), predicting that most employees will stay, which **defeats the purpose of the project**. Effective handling of the imbalance is essential to ensure that the model accurately identifies employees at risk of leaving, enabling timely interventions and informed decision-making.

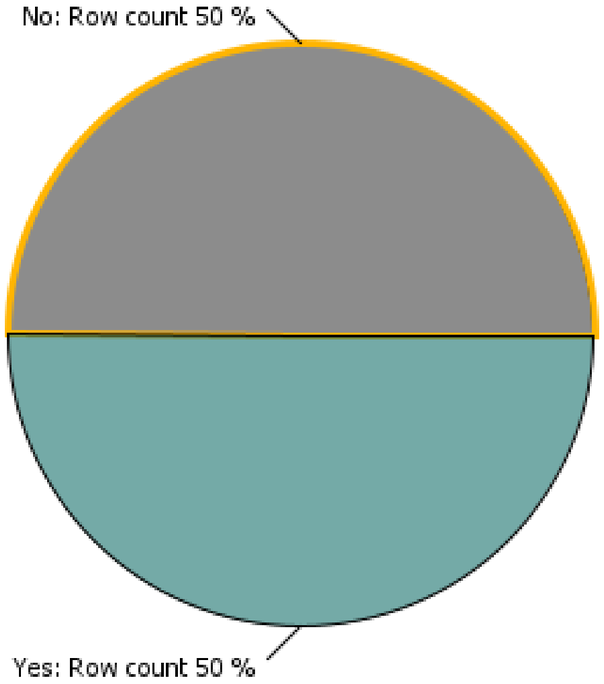


BEFORE RESAMPLING

RESAMPLING

PRE-PROCESSING





AFTER RESAMPLING

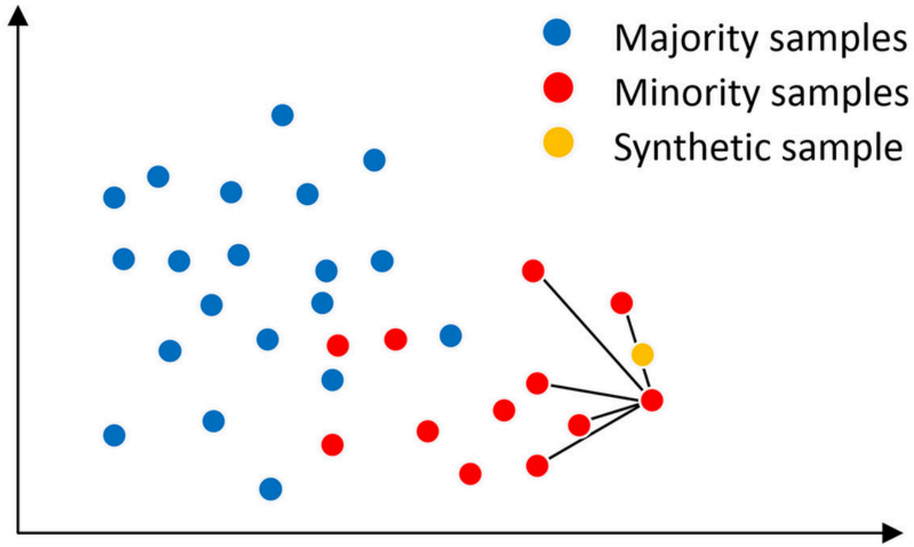
To address the issue of dataset imbalance, there are two common approaches:

- **Resampling** the dataset, such as using SMOTE.
- Using **models specifically tuned** for imbalanced datasets, such as Weighted Random Forest (WRF).

We chose to proceed with **SMOTE** because KNIME does not currently offer a node for implementing WRF or other weighted models. This allows us to **effectively balance** the dataset and improve model performance.

RESAMPLING

PRE-PROCESSING



What's SMOTE?

SMOTE is a method used to address class imbalance in datasets. It works by creating synthetic samples for the **minority class.**

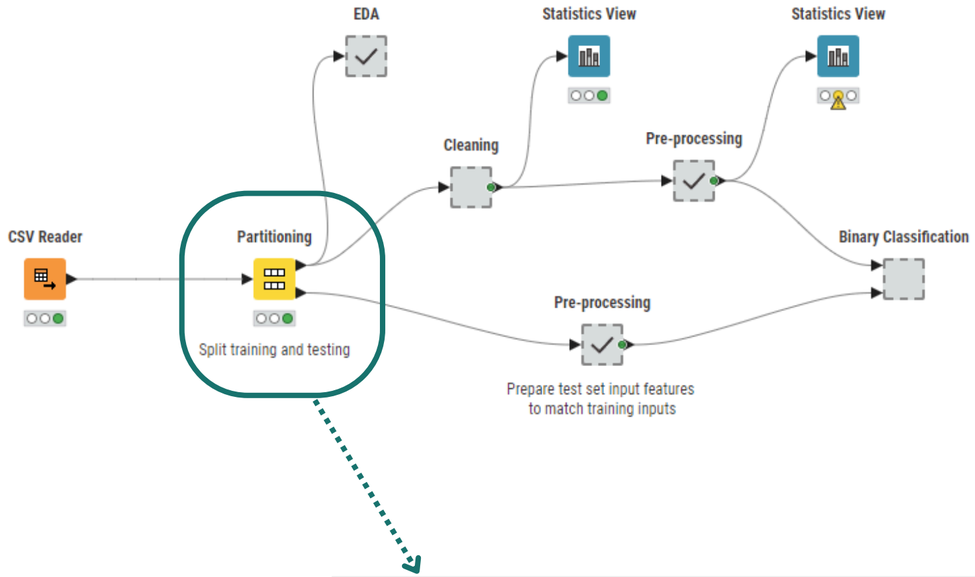
It selects a random data point from the minority class and finds its k-nearest neighbors. Then, it interpolates new samples by randomly selecting one of the neighbors and creating a **new point** along the line between the original point and the neighbor.

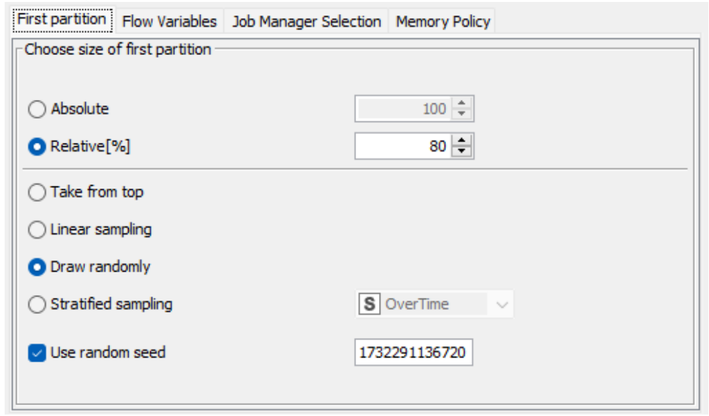
This **boosts the representation of the minority class** (as you can see in the pie chart) without simply duplicating existing samples, helping **models learn better** from imbalanced data.

04 PARTITIONING

PARTITIONING

PRE-PROCESSING





Finally, we talk about an essential part in any prediction task, model selection and testing. In order to test trained algorithms to fresh data, we **must partition** our dataset randomly in two parts: training set, where we will be training our model, and testing set, which will **represent unseen data** and therefore will be our fresh observations where we can test our models onto, without bothering collecting new data just for testing a procedure. A crucial step is to understand **at which stage we should split our dataset**, because drawing the line too deep in the workflow could lead to **data leakage**, and bias our models performances.

To ensure accuracy, we split the dataset at the beginning, performed data cleaning (including outlier removal) on the training set, and then applied the **same preprocessing steps** to the test set. This ensured that the **test data was properly prepared** with the correct data types and formats for the model.

MODELING

01 MODELS

02 FINE TUNING

03 TARGET METRICS

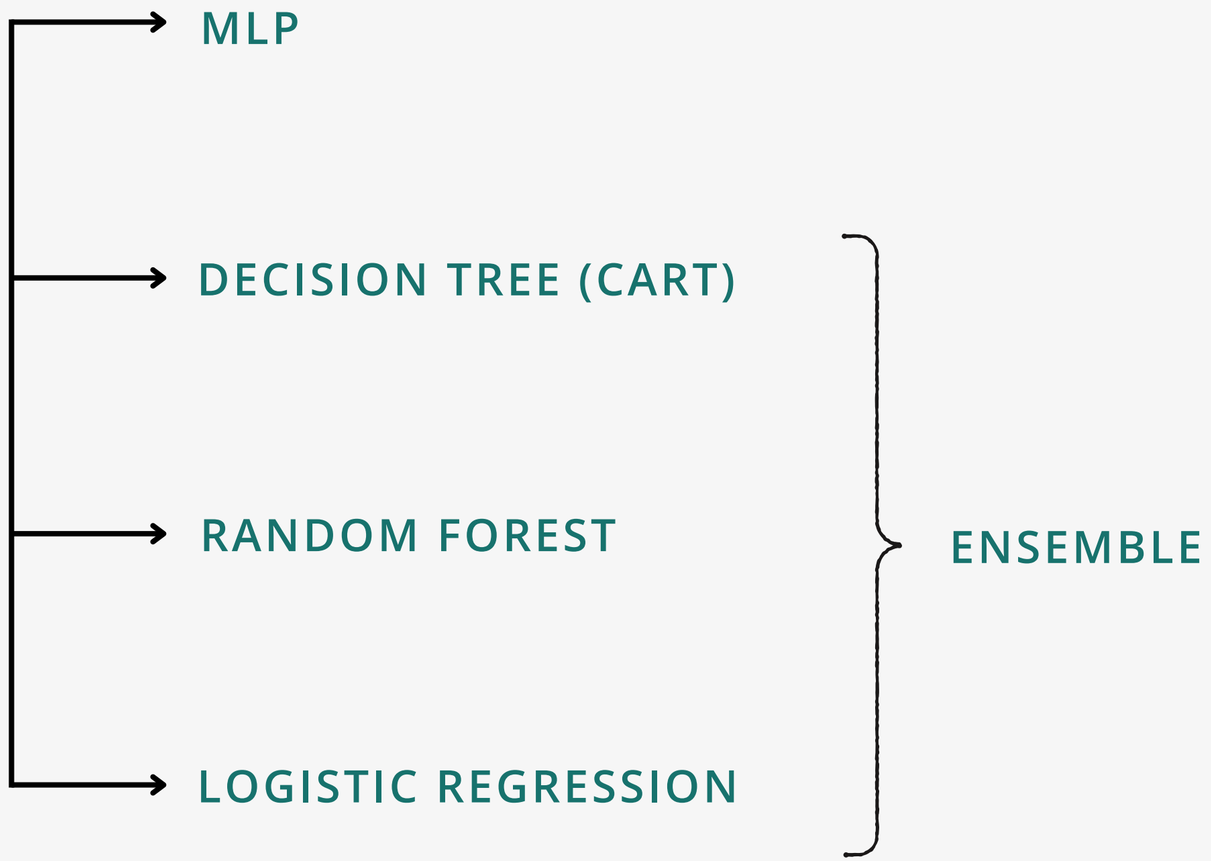
04 FEATURE OPTIMIZATION

05 MODELS COMPARISON

06 EXTRA LEARNER

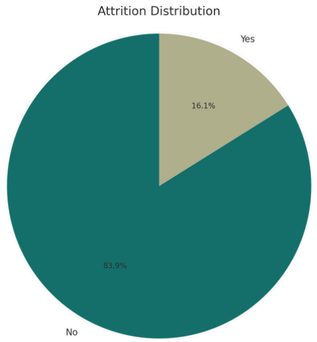
01 MODELS

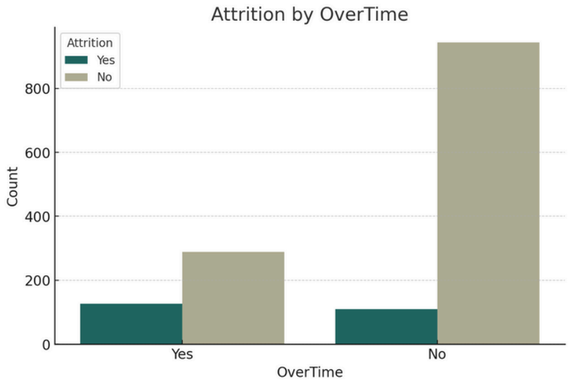
BINARY CLASSIFICATION MODELS

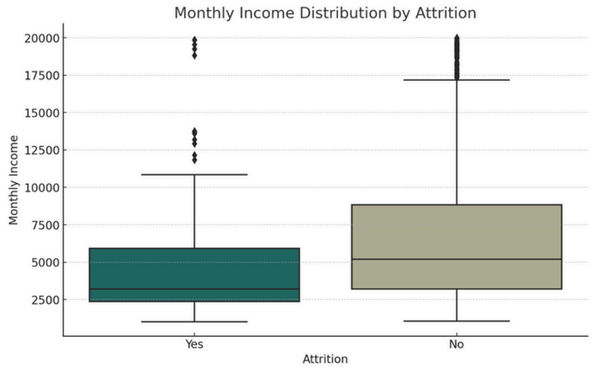


OVERVIEW

TARGET VARIABLE







To predict employee attrition, we employed multiple machine learning models, each leveraging unique methodologies to analyze the dataset.

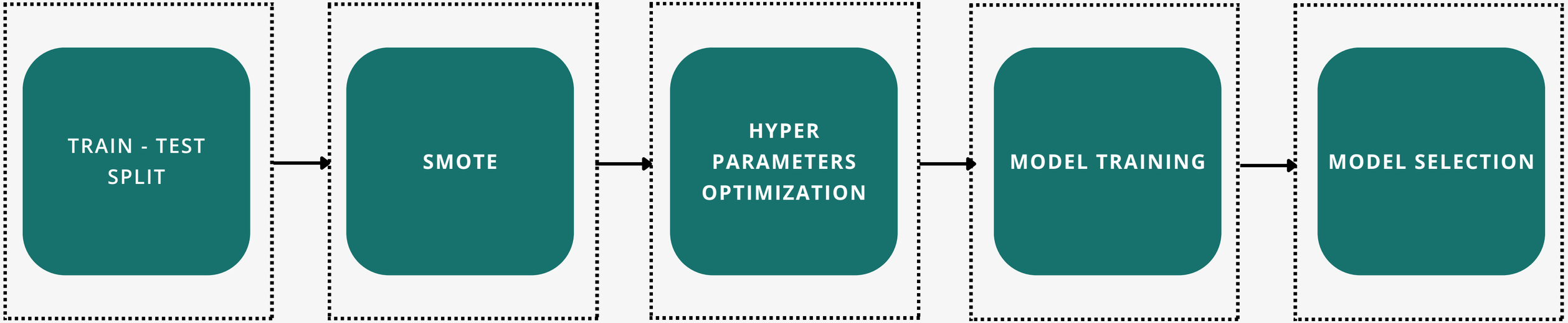
The chosen models include:

- **Logistic Regression**: A baseline probabilistic model that identifies relationships between employee characteristics and the likelihood of attrition, providing interpretable coefficients for actionable insights.
- **Decision Tree**: A rule-based model that splits the data into hierarchical decisions, offering easy visualization and identification of key factors influencing attrition.
- **Random Forest**: An ensemble model that combines multiple decision trees to improve accuracy and reduce overfitting, balancing sensitivity and specificity.
- **Multi-Layer Perceptron** (MLP): A neural network that captures complex, non-linear patterns in the data, suitable for uncovering deeper insights.

Each model was trained and evaluated using metrics such as sensitivity, specificity, precision, and AUC-ROC to assess their predictive performance and suitability for guiding retention strategies.

PIPELINE

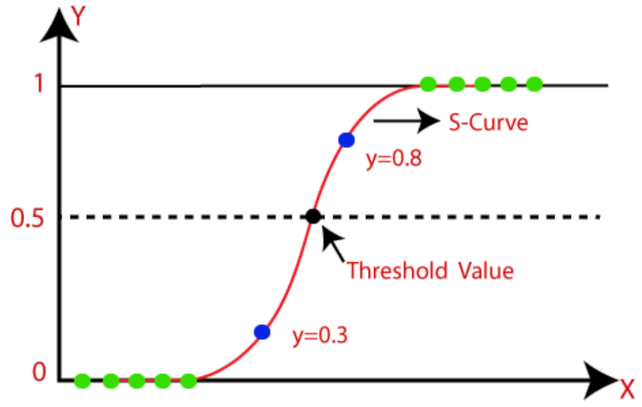
MODELLING



01 MODELS: LOGISTIC REGRESSION

TRAINING

LOGISTIC REGRESSION



The Logistic Regression training pipeline predicts employee attrition using a probabilistic framework. It begins with Normalization, ensuring consistent feature scaling to improve training stability. The Normalizer computes scaling parameters from training data, applied to test data through the Normalizer (Apply) node.

The Logistic Regression Learner trains the model by estimating feature weights (coefficients) that influence the probability of attrition based on inputs like tenure or satisfaction. The Logistic Regression Predictor uses this model to generate probabilities and classifications on the test data.

Model performance is evaluated using the Binary Classification Inspector, which provides metrics such as accuracy, sensitivity, and AUC-ROC. The Math Formula and Column Filter refine outputs for concise and actionable reporting. This pipeline ensures interpretable attrition predictions.

COEFFICIENTS

LOGISTIC REGRESSION

Variable *String*	Coeff. *Number (doub...*	Std. Err. *Number (double)*	z-score *Number (double)*	P>|z| ↑ *Number (double)*
Travel_Frequently_BusinessTravel	1.029	0.13	7.908	0
High company changes_NumCompaniesWorked	0.544	0.08	6.811	0
JobInvolvement	-0.49	0.072	-6.795	0
Married_MaritalStatus	-0.608	0.092	-6.582	0
Divorced_MaritalStatus	-0.62	0.102	-6.08	0
WorkLifeBalance	-0.411	0.072	-5.721	0
Travel_Rarely_BusinessTravel	0.699	0.13	5.365	0
OverTime	0.953	0.184	5.166	0
Long time since promotion_YearsSinceLastPromotion	0.428	0.093	4.585	0
MonthlyIncome	-0.877	0.193	-4.535	0
Technical Degree_EducationField	0.424	0.109	3.879	0
Short tenure_YearsAtCompany	0.387	0.101	3.837	0
PercentSalaryHike	-0.423	0.112	-3.792	0
Research Director_JobRole	-0.64	0.176	-3.635	0
EnvironmentSatisfaction	-0.474	0.131	-3.631	0
Low training_TrainingTimesLastYear	0.258	0.072	3.584	0
RelationshipSatisfaction	-0.346	0.114	-3.026	0.002
Manufacturing Director_JobRole	-0.576	0.192	-3.004	0.003
Healthcare Representative_JobRole	-0.519	0.182	-2.861	0.004

Coefficients in logistic regression indicate the impact of each variable on the **likelihood of employee attrition**. A positive coefficient means the variable increases the probability of leaving, while a negative coefficient reduces it. The p-values assess statistical significance, with smaller values (e.g., $p < 0.05$) suggesting strong evidence for the variable's importance.

Key insights include:

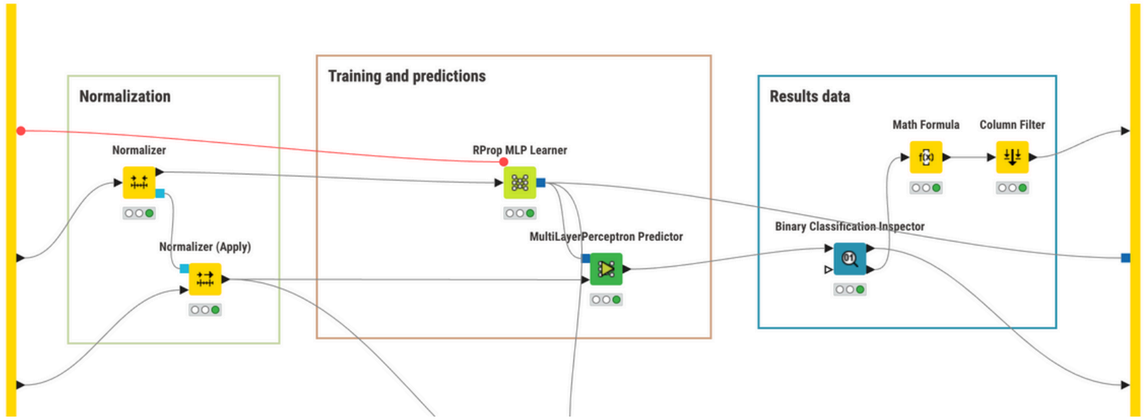
- **Travel_Frequently_BusinessTravel** (Coeff: 1.029): Employees traveling frequently are more likely to leave, likely due to stress or work-life imbalance.
- **MonthlyIncome** (Coeff: -0.877): Higher income reduces attrition, likely reflecting job satisfaction and stability.
- **OverTime** (Coeff: 0.953): Overtime strongly increases attrition, possibly due to burnout.

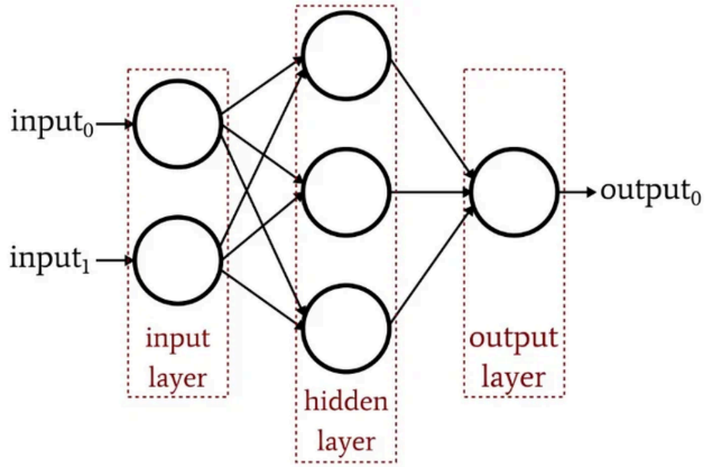
These significant coefficients provide actionable insights for reducing attrition by addressing workload, travel policies, and compensation.

01 MODELS: MLP

TRAINING

MLP





The MLP training pipeline involves three key stages: **preprocessing**, **model training**, and **evaluation**. First, normalization ensures all input features are scaled to comparable ranges using parameters derived from the training data, improving the stability of neural network training.

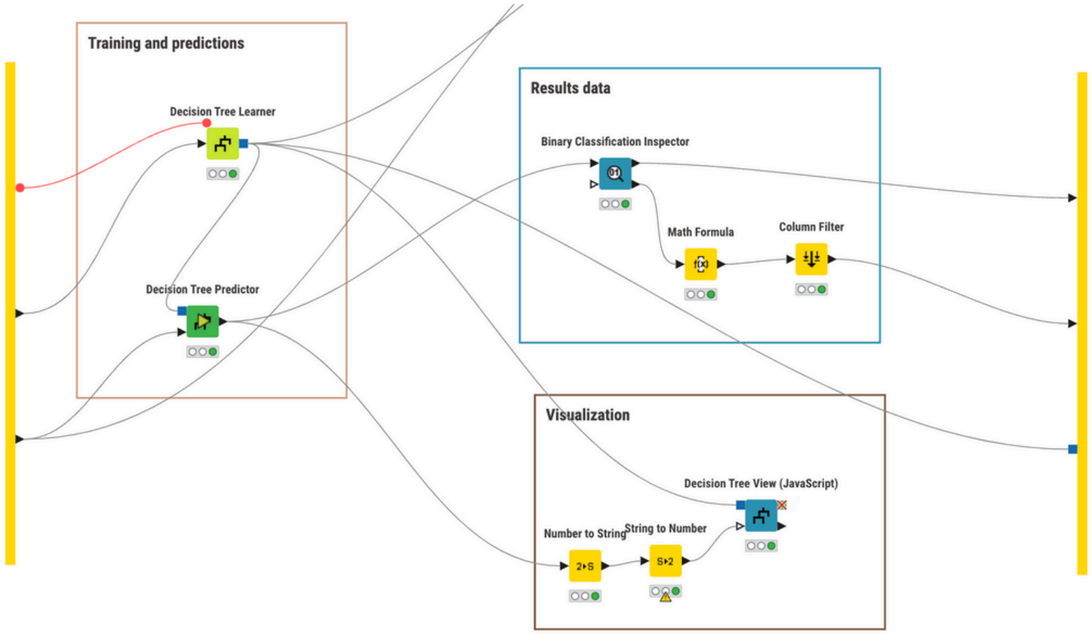
The **RProp MLP Learner** trains the Multi-Layer Perceptron using the Resilient **Backpropagation** algorithm, optimizing the network's weights efficiently without being affected by gradient magnitudes.

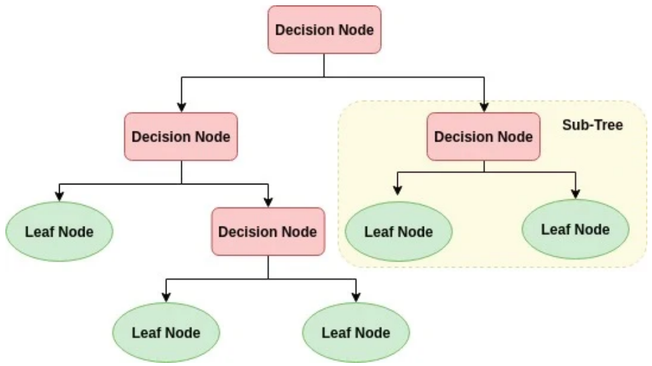
The trained model generates predictions through the MLP Predictor, including probabilities for binary classification. The Binary Classification Inspector evaluates model performance using metrics providing a complete view of its ability to classify employees at risk of leaving. Finally, formulas and column filtering refine outputs for performance statistics adding ad-hoc measure (cost).

01 MODELS: DECISION TREE (CART)

TRAINING

DECISION TREE





The Decision Tree training pipeline models employee attrition by splitting data into a tree structure of decision nodes and leaf nodes. Using the Decision Tree Learner, the model recursively partitions the data to minimize impurity or **maximize information gain** at each split. Features such as tenure or job satisfaction determine the splits, while the leaf nodes provide the final predictions of whether an employee will stay or leave. The Decision Tree Predictor applies the trained model to the test data, generating predictions and probabilities.

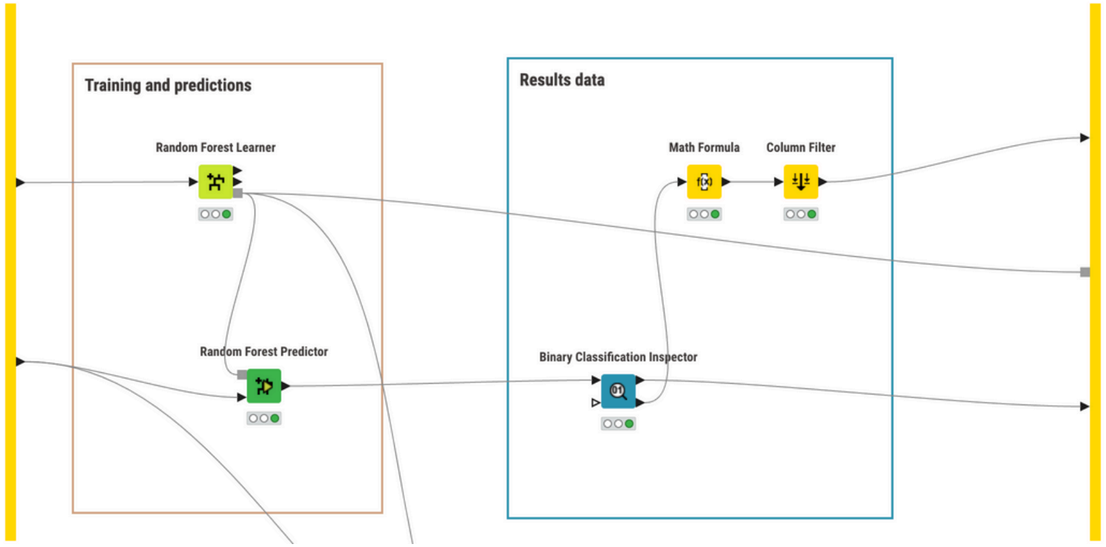
Performance evaluation is conducted through the Binary Classification Inspector, calculating metrics like accuracy, sensitivity, and AUC-ROC to assess the model's predictive capabilities. Additional steps refine outputs for **interpretability**.

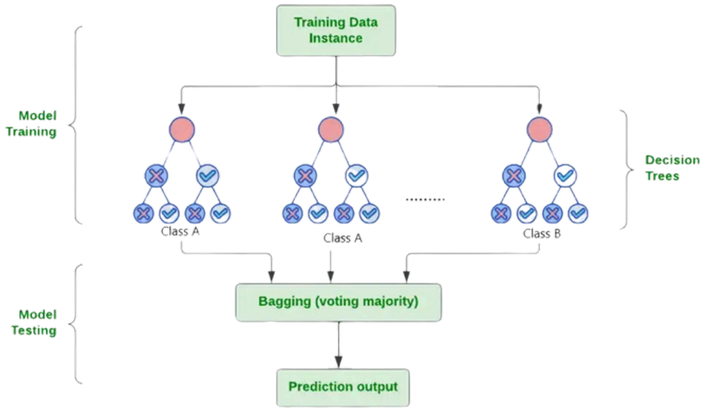
Finally, the Decision Tree View visualizes the model, **highlighting key features and decision paths**. This interactive visualization aids understanding of the model's behavior, making it an effective tool for attrition analysis

01 MODELS: RANDOM FOREST

TRAINING

RANDOM FOREST





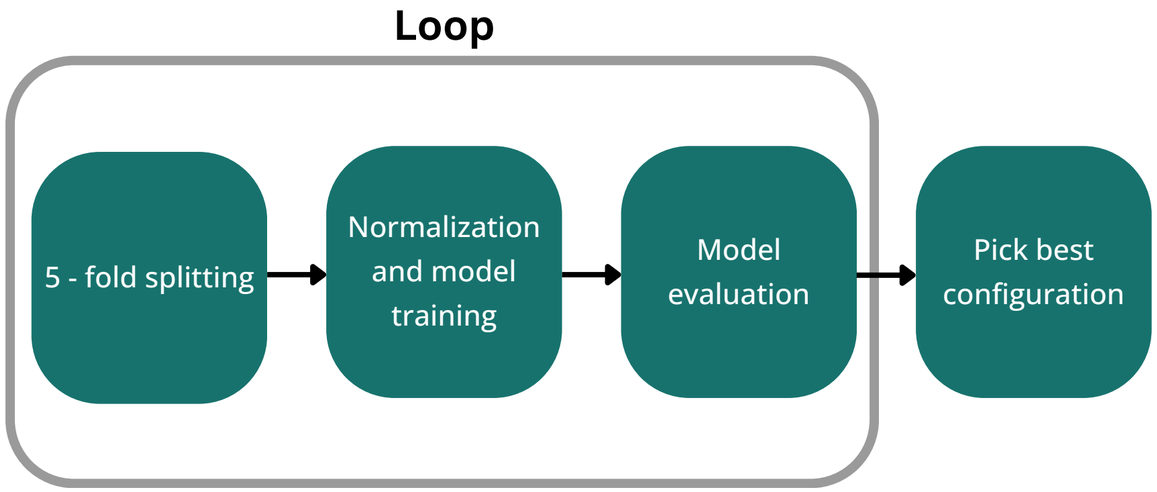
The Random Forest training pipeline creates a robust ensemble model by combining multiple decision trees, leveraging their collective power to enhance prediction accuracy and reduce overfitting. The Random Forest Learner trains the model by randomly selecting subsets of data and features to **grow diverse decision trees**. Each tree independently predicts whether an employee will leave or stay, and the final prediction is determined by majority voting across all trees, **improving both stability and performance.**

The trained model is applied to the test data using the Random Forest Predictor, producing probabilities and classifications. Performance is evaluated through the Binary Classification Inspector. Additional nodes like Math Formula and Column Filter refine the results for clear analysis and reporting. This ensemble approach combines **high accuracy with interpretability**, making it a powerful tool for employee attrition prediction.

02 FINE TUNING

MLP

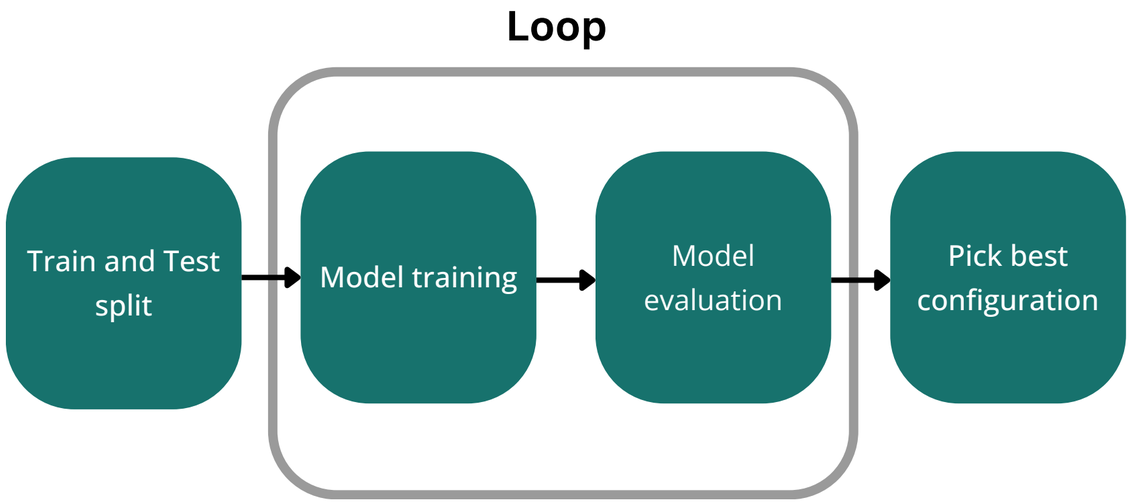
PARAMETER OPTIMIZATION



With this approach we aimed to capture the best parameter for our MLP model using cross validation. At each iteration of the loop we trained our model with different parameters and then evaluated it using AUC. At the end of the loop, we collected the results and we picked the **best configuration**.
The considered parameters are the number of units per layer, in a range from 10 to 50, and the number of layers of the network, in a range from 2 to 10.
To give a more detailed look of the pipeline: we first start the loop choosing the parameter to be optimized, then proceed to do the 5-fold split and then at each iteration the data are normalized and the model is trained with the current configuration of parameters. After this the AUC of each configuration is collected and the best parameters configuration is saved and used in our MLP Model.
Optimal architecture: 2 layers, 50 units

DECISION TREE

PARAMETER OPTIMIZATION



Again, we did the same procedure for the Decision Tree. This time we looked for the best value for **MinNodeSize** that is the parameter that determine the minimum possible number of observation in a node. Usually, a low value for this parameter means a deeper and more complex tree, while a high value indicates a simpler and shallower tree.
We analysed a range of values going from 1 to 500.

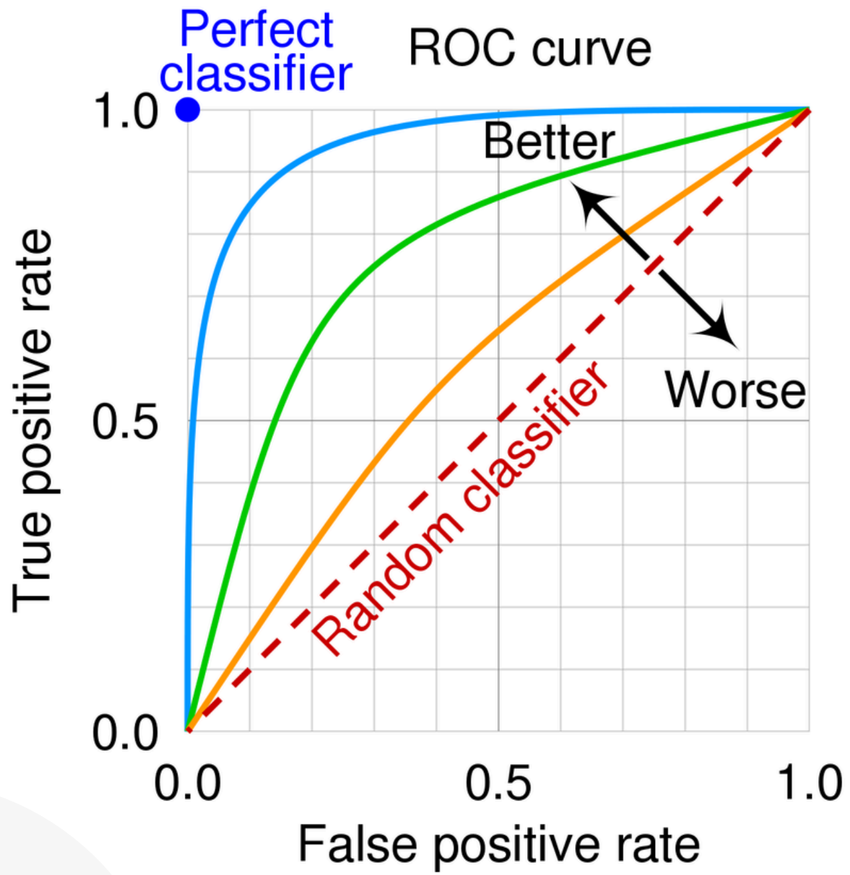
This time to reduce computational cost we avoided the cross validation, going instead for a simple partition of the data before the entering the loop.

Optimal parameter: 6 MinNodeSize

03 TARGET METRICS

AUC

PERFORMANCE METRIC



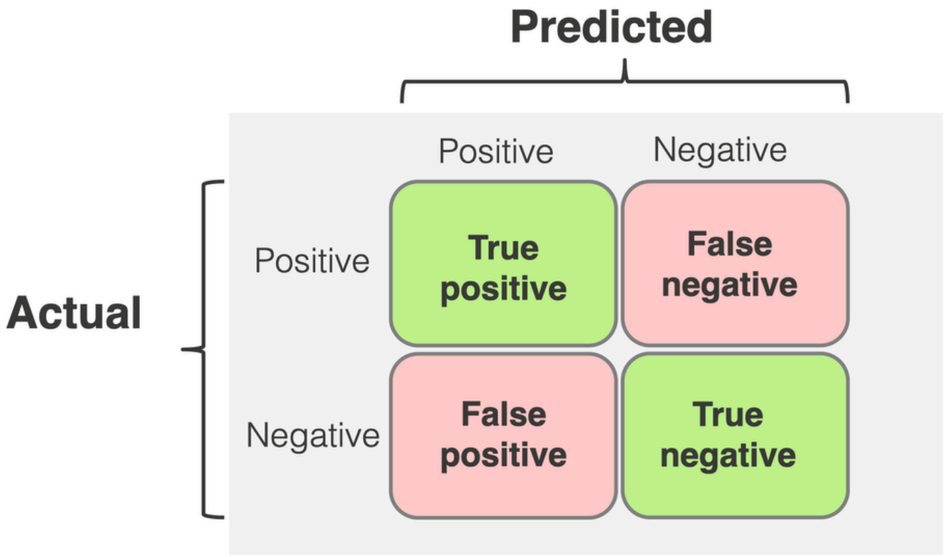
Another important metric is the AUC-ROC, which assesses our model's **ability to discriminate** between employees who will leave and those who will stay. The ROC curve plots the True Positive Rate against the False Positive Rate across various classification thresholds, illustrating the trade-off between correctly identifying leavers and incorrectly flagging stayers.

Mathematically, the AUC represents the probability that a **randomly selected employee** who **left** is **assigned** a **higher risk score than** a **randomly selected** employee who **stayed**. By integrating over all possible thresholds, the AUC provides a single scalar value summarizing the model's overall ranking performance.

Maximizing the AUC enhances the model's **discriminative power**, enabling us to prioritize employees based on their calculated likelihood of leaving.

CONFUSION MATRIX

PERFORMANCE METRIC



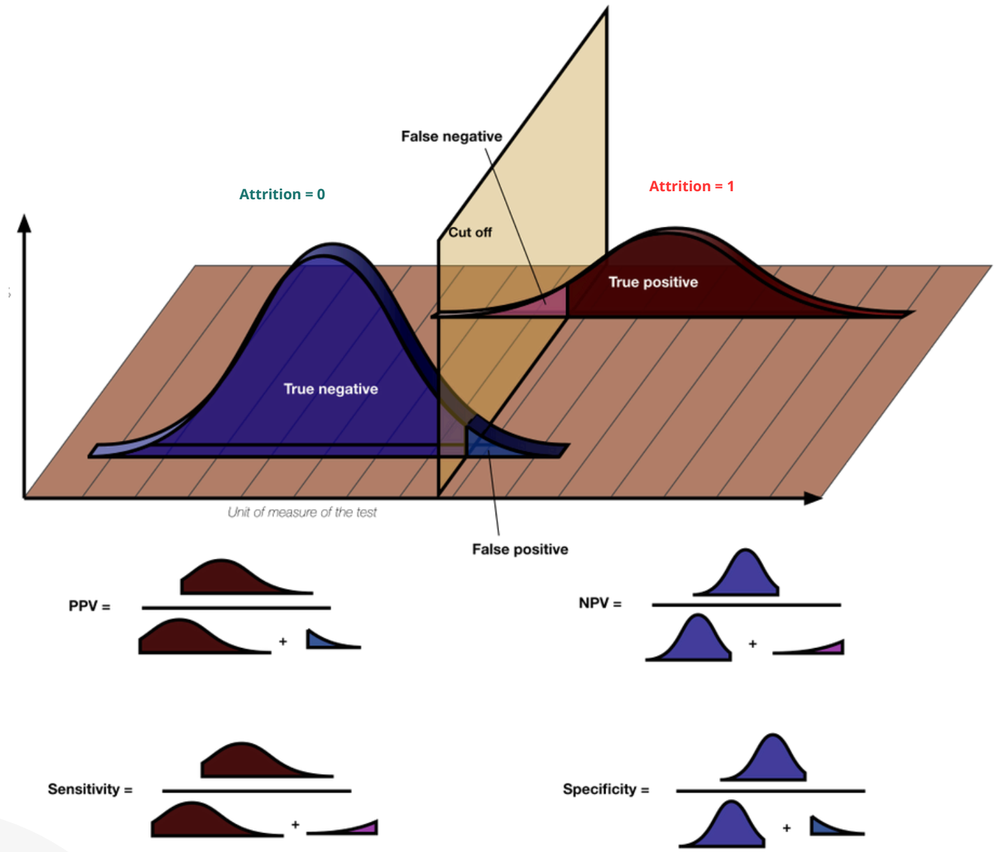
Another important tool we should keep into consideration is the Confusion Matrix, which represents the effectiveness of our model in classifying both employees who are likely to leave and those who are likely to stay.

Accurately classifying employees who are likely to stay has its own advantages. A correct evaluation allows the company to focus retention efforts where they're most needed and to better plan for future workforce stability.

The Confusion Matrix provides a clear visualization of these classifications, helping us understand the number of **true positives**, **false positives**, **true negatives**, and **false negatives** in our predictions. This insight is crucial for improving our model and making informed decisions to enhance employee retention.

SENSITIVITY

PERFORMANCE METRIC



In our problem, a **positive result** (denoted as 1) represents an employee who is **likely** to **leave** the company. As with many anomaly detection tasks, our goal is to identify as many positives as possible among all actual positives. This means we aim to minimize the number of employees predicted to stay but who actually leave—these are our false negatives.

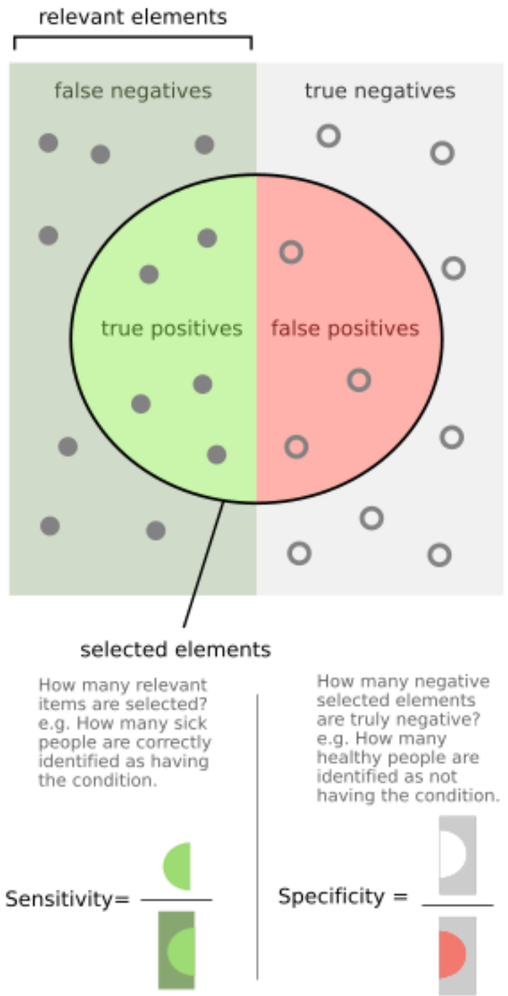
To achieve this, we focus on **maximizing sensitivity**, calculated as:

$$\text{Sensitivity} = \frac{\text{True Positives (TP)}}{\text{True Positives (TP)} + \text{False Negatives (FN)}}$$

By maximizing sensitivity, we **increase** the number of **departing employees** accurately **identified** and reduce the number of potential leavers that the model fails to detect. This approach helps us proactively address attrition by implementing targeted retention strategies, thereby minimizing the loss of valuable talent from the organization.

SPECIFICITY

PERFORMANCE METRIC



Specificity, or the **true negative rate**, is a crucial metric that provides a summary of a model's ability to correctly identify employees who are **not likely to leave** the organization. It is particularly helpful when you need to quickly understand how effectively your attrition prediction model distinguishes between those who will stay and those who might depart. Specificity ranges from 0 to 1, with the maximum value reached when the model accurately classifies all non-attriting employees in a test.

Calculated as:

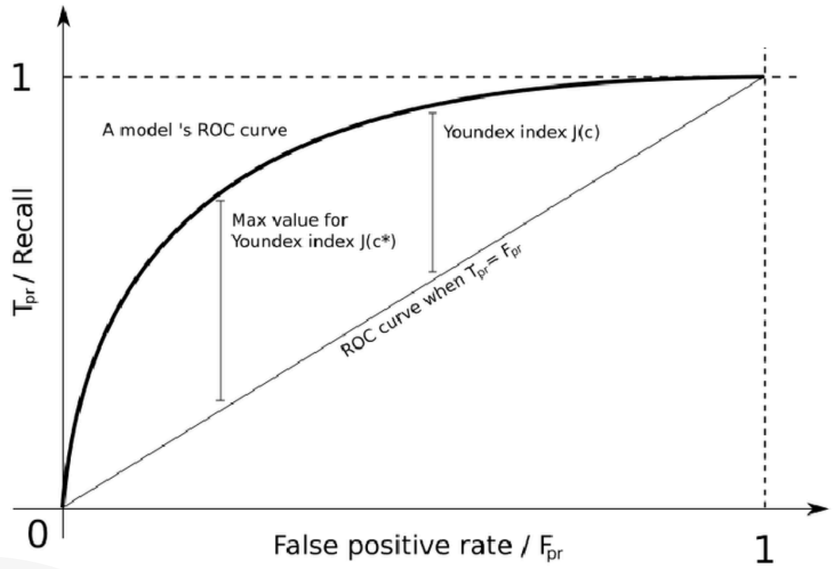
$$\text{Specificity} = \frac{\text{True Negatives (TN)}}{\text{True Negatives (TN)} + \text{False Positives (FP)}}$$

- True Negatives (**TN**): Employees correctly predicted to stay who actually stayed.
- False Positives (**FP**): Employees incorrectly predicted to leave but actually stayed.

YOUDEN INDEX

PERFORMANCE METRIC

$$
\begin{aligned}
J &= \text{Sensitivity} + \text{Specificity} - 1 \\
&= \frac{\text{TP}}{\text{TP} + \text{FN}} + \frac{\text{TN}}{\text{TN} + \text{FP}} - 1
\end{aligned}
$$



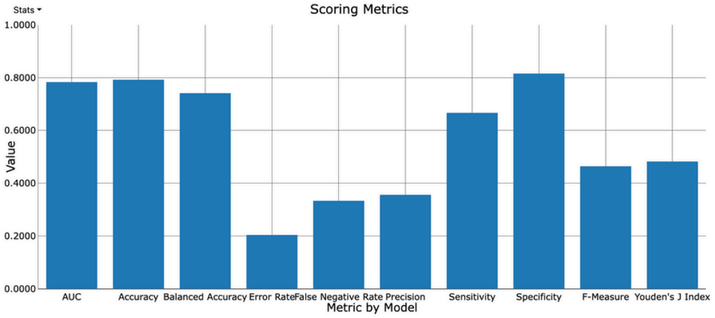
Also known as Youden's J statistic, this metric provides a concise **summary of a model's overall performance**, making it helpful for quickly assessing the **predictive power** of a process. It incorporates both **specificity** (the ability to correctly identify negative cases) and **sensitivity** (the ability to correctly identify positive cases) into a single value that ranges from **0 to 1**. A maximum value of 1 is achieved when a model perfectly classifies all records in a test dataset.

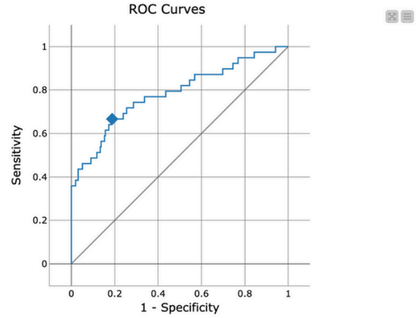
While attaining a Youden's J value of 1 is an ideal scenario and often unattainable in practice, keeping this metric in use is beneficial for explanatory purposes. Its straightforward **interpretation** makes it easily understandable for non-technical stakeholders, aiding in clear communication about model effectiveness.

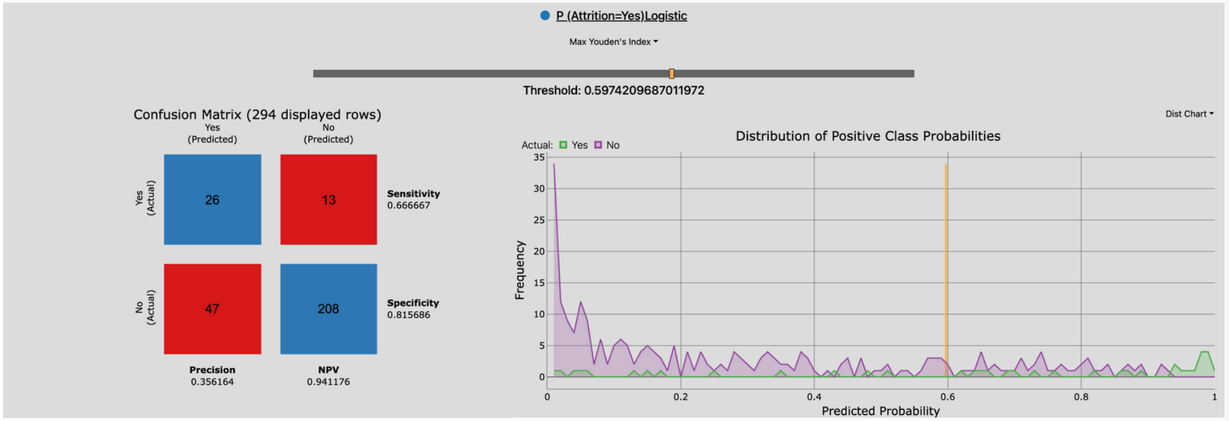
We will use it as the threshold for the models performance as it optimizes the trade-off between sensitivity and specificity, ensuring a balancing.

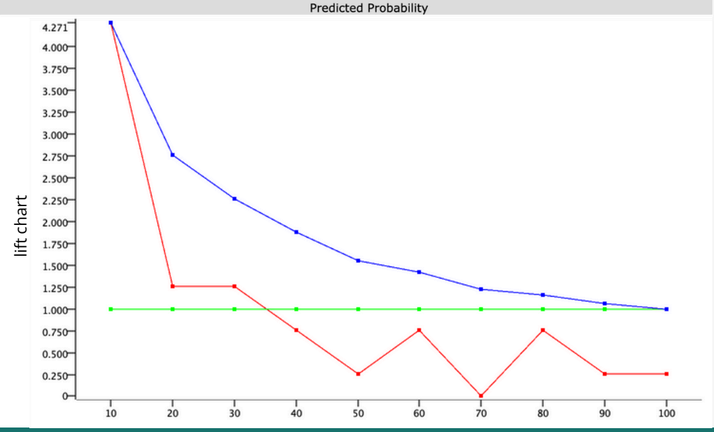
LOGISTIC REGRESSION

PERFORMANCE







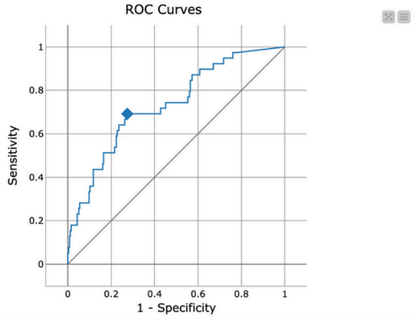


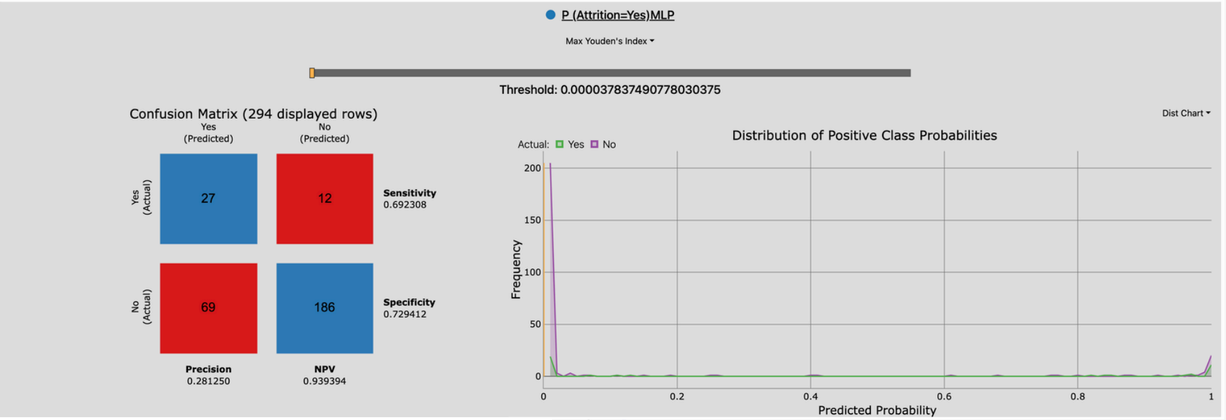
The logistic regression model shows strong predictive performance with a **well-optimized** threshold for balancing sensitivity and specificity, as illustrated by the ROC curve. The confusion matrix reveals that while the model captures a reasonable number of employees likely to leave, there are still missed true positives (false negatives) and some overpredictions (false positives), which slightly impacts precision.

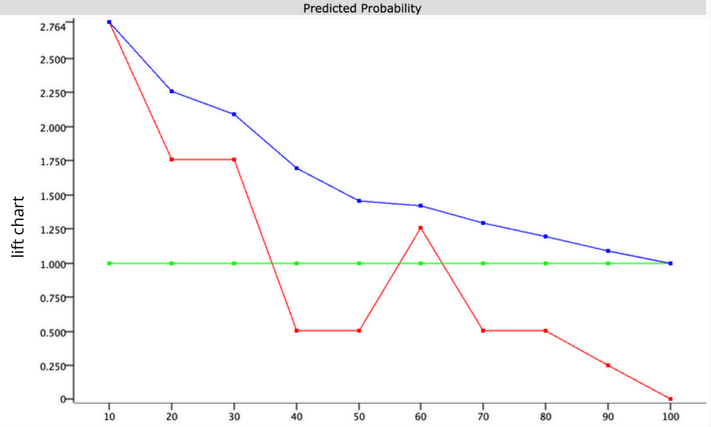
The scoring metrics demonstrate a **trade-off:** high specificity suggests strong accuracy in identifying employees who will stay, while lower sensitivity indicates some limitations in catching all potential leavers. The probability distribution highlights a clear **separation**, with most negative cases concentrated near **zero** and positive cases spread closer to the **threshold**. The lift chart confirms the model's ability to **prioritize high-risk** employees effectively for targeted retention strategies.

MLP

PERFORMANCE





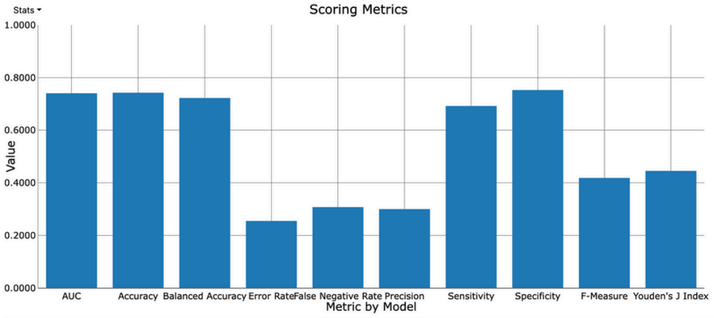


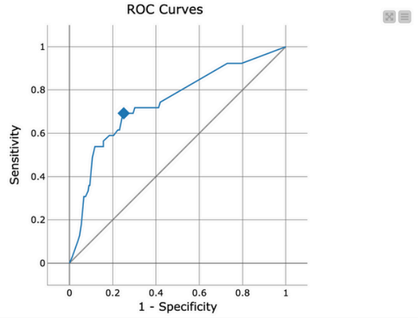
The MLP model demonstrates weaker performance compared to other models in predicting employee attrition. The ROC curve highlights **suboptimal** discriminative power, with **low separation** between positive and negative cases. The confusion matrix shows a high number of false positives (69) and false negatives (12), leading to poor precision. This overprediction tendency undermines the model's reliability for guiding retention strategies.

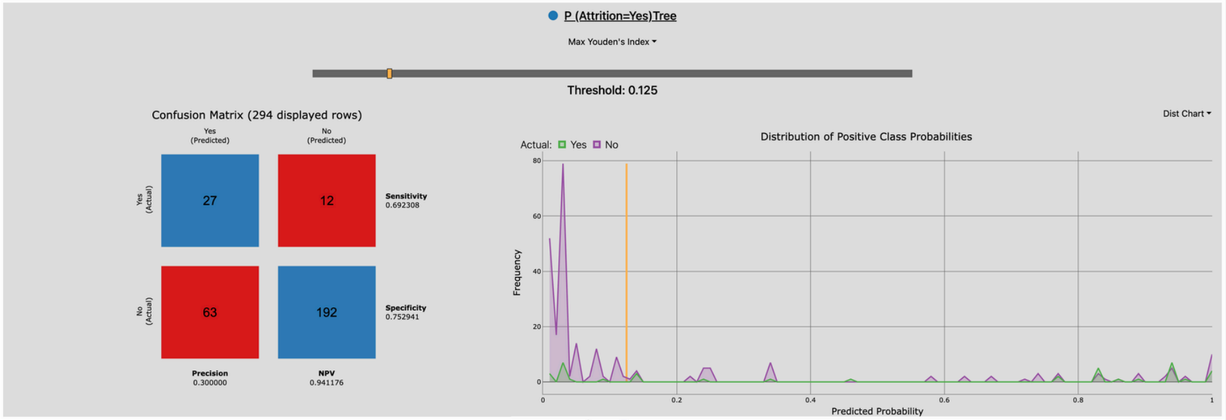
The scoring metrics indicate **moderate sensitivity** but **poor specificity**, showing the model slightly favors identifying potential leavers while struggling with accurate classification of those who stay. The probability distribution reflects **limited prediction confidence**, with most probabilities clustered near zero. The lift chart further confirms the model's **inefficiency**, offering little improvement over random selection in prioritizing at-risk employees. This suggests the MLP requires significant **tuning** or **reconsideration**.

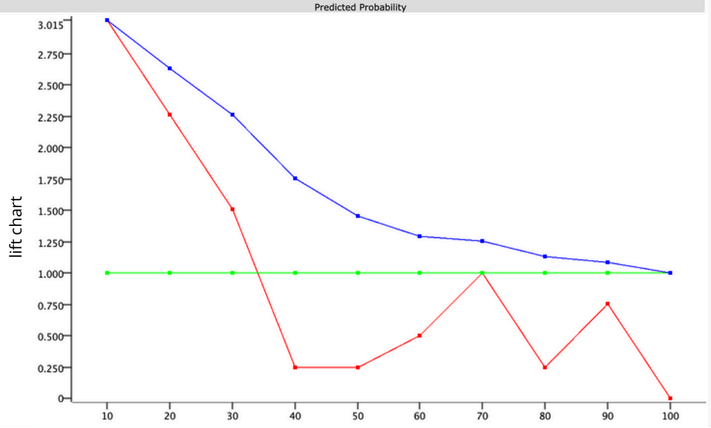
DECISION TREE

PERFORMANCE







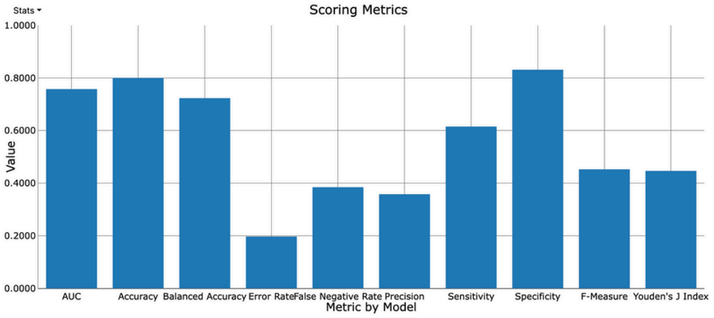


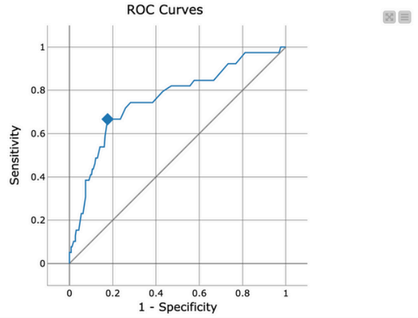
The Decision Tree model demonstrates moderate performance in predicting employee attrition. The ROC curve illustrates a reasonable **trade-off** between sensitivity and specificity, with the threshold optimized at Youden's Index. The confusion matrix highlights a balanced performance, though false positives (63) remain a notable issue, reducing precision.

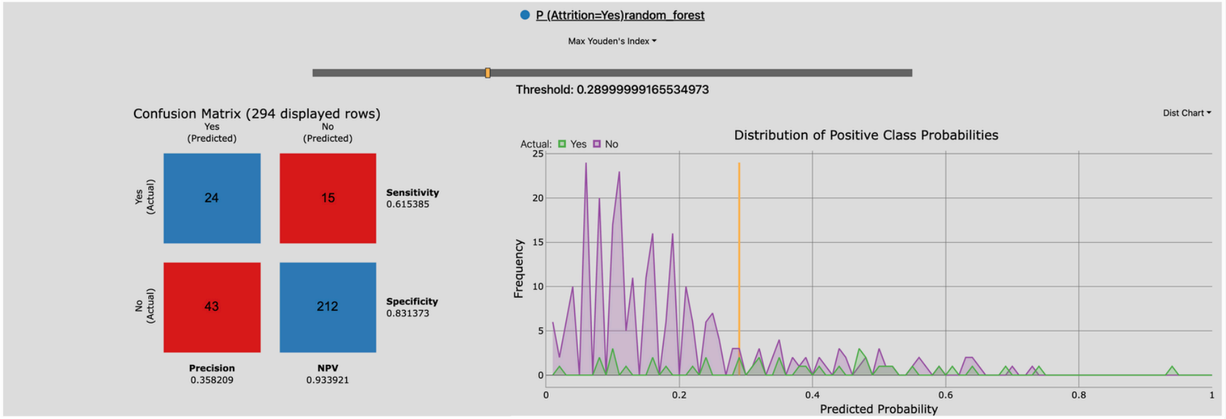
Scoring metrics reveal higher sensitivity than specificity, indicating the model is more effective at identifying employees likely to leave while struggling to accurately classify those who will stay. The probability distribution shows that although **class separation** is evident, many predicted probabilities cluster near the threshold, reflecting uncertainty in some classifications. The lift chart supports these findings, with **moderate improvement** over random selection. While actionable, the model requires tuning to enhance overall precision and reliability.

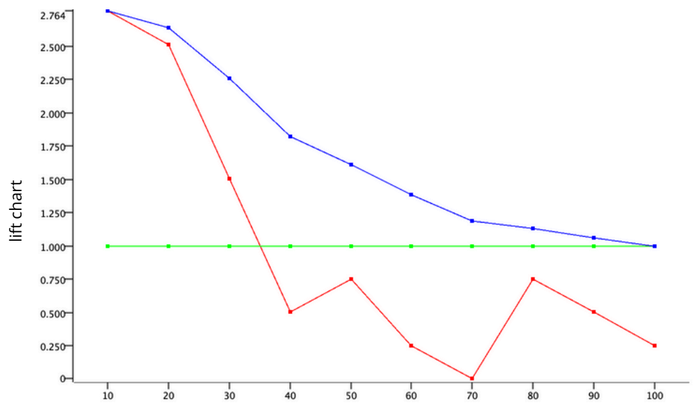
RANDOM FOREST

PERFORMANCE









The Random Forest model demonstrates stronger performance compared to the MLP. The ROC curve highlights good discriminative ability, supported by a high AUC value. The confusion matrix indicates a better balance between **false positives** and **false negatives**, though false positives remain relatively high, slightly impacting precision.

Scoring metrics reveal **higher specificity** than sensitivity, indicating the model is more effective at identifying employees who will stay rather than those likely to leave. The probability distribution chart shows clear **separation**, with predicted probabilities concentrated closer to the **optimal threshold**. While the model effectively classifies many cases, further tuning could enhance its sensitivity, making it more reliable for identifying employees at risk of attrition.

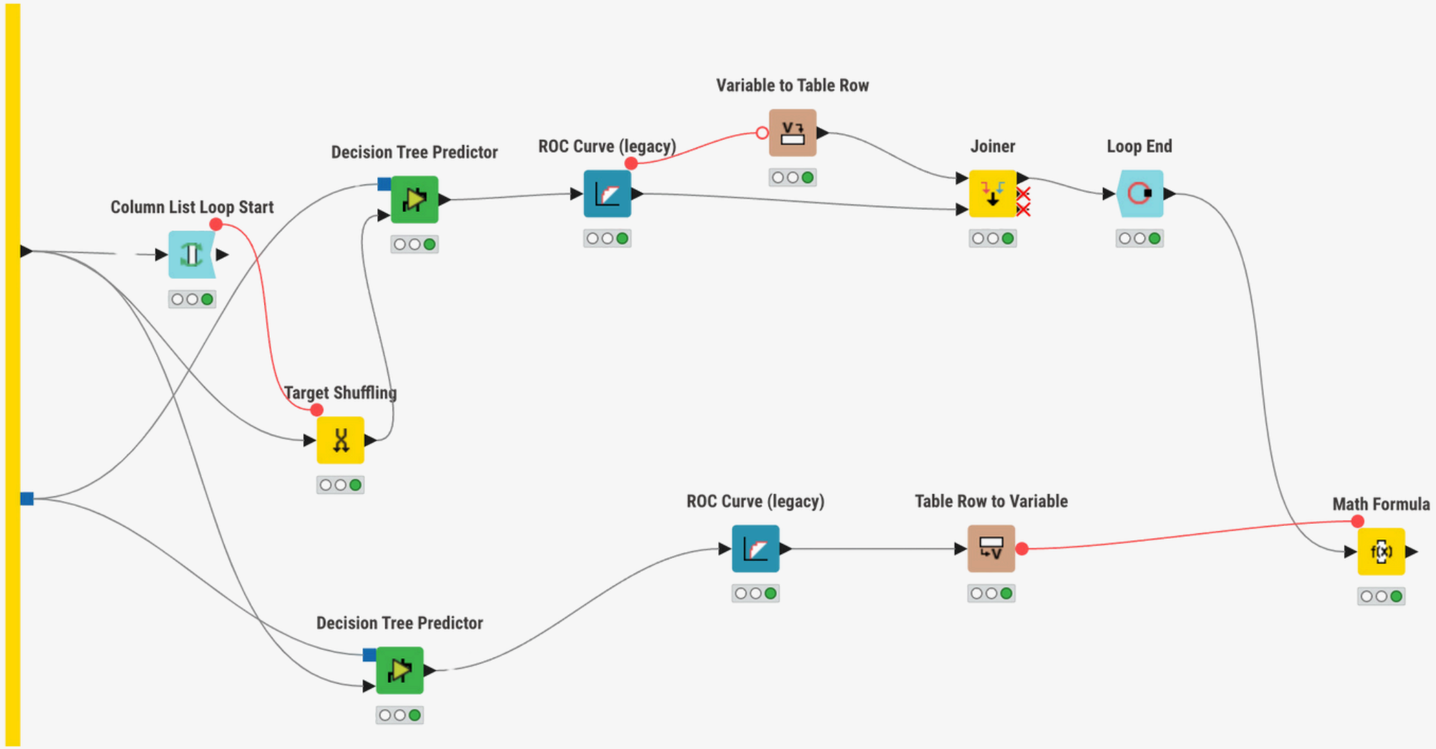
04 FEATURE OPTIMIZATION

FEATURE IMPORTANCE

FEATURE IMPORTANCE

To get an idea of which **features have the greatest impact on the AUC**, and are therefore considerable as the most informative for our model, we implemented a Feature Importance metanode for each model.

By **comparing** the model's performance with the original data to its performance when a specific feature is randomized (shuffled), we can investigate the change in AUC. Features that cause a bigger drop in AUC when shuffled are considered more informative. This helped us to identify the key drivers of the models' predictions.

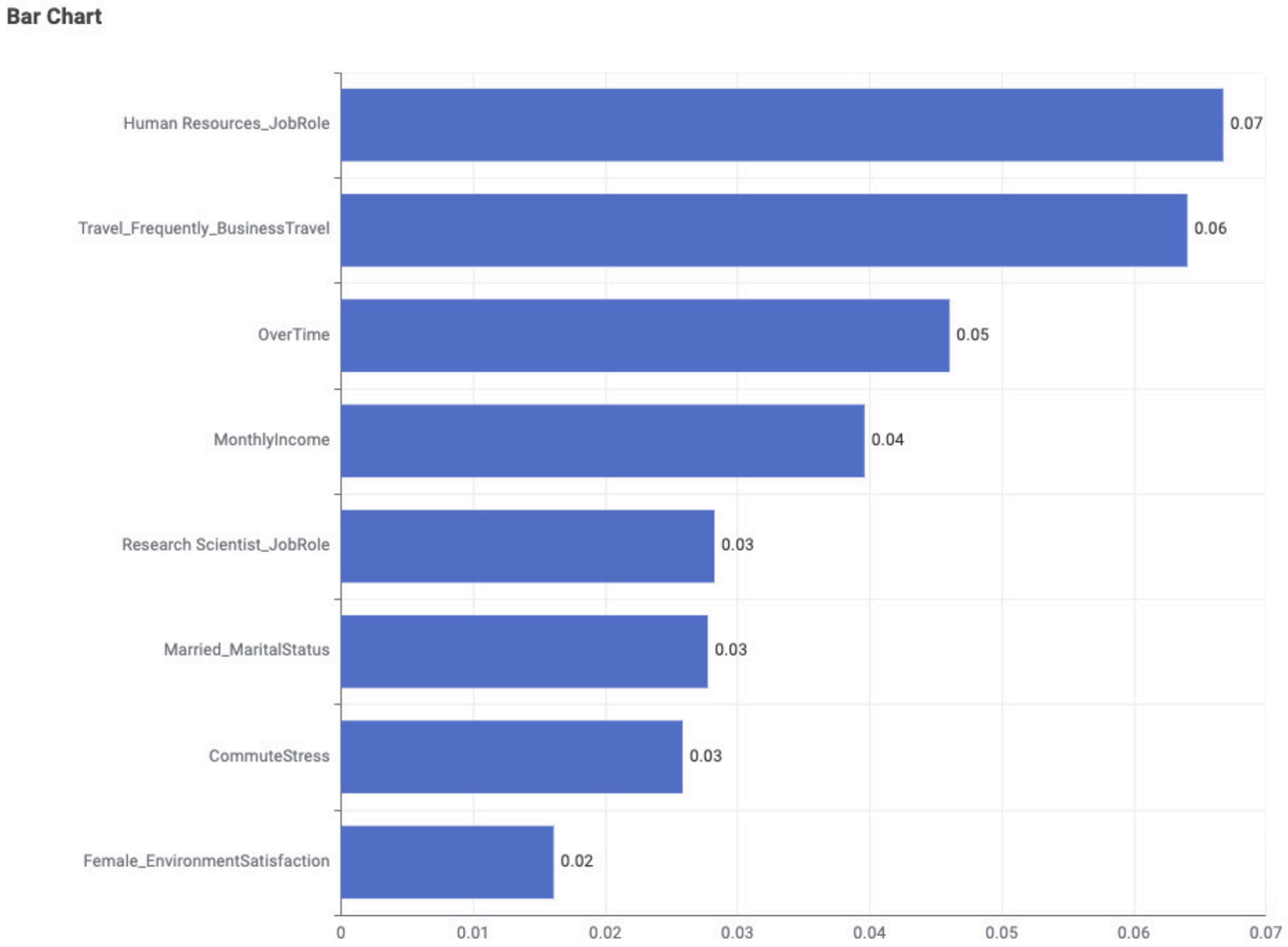


LOGISTIC

FEATURE IMPORTANCE

Here we see the **features** ranked by the importance scores obtained by the model.

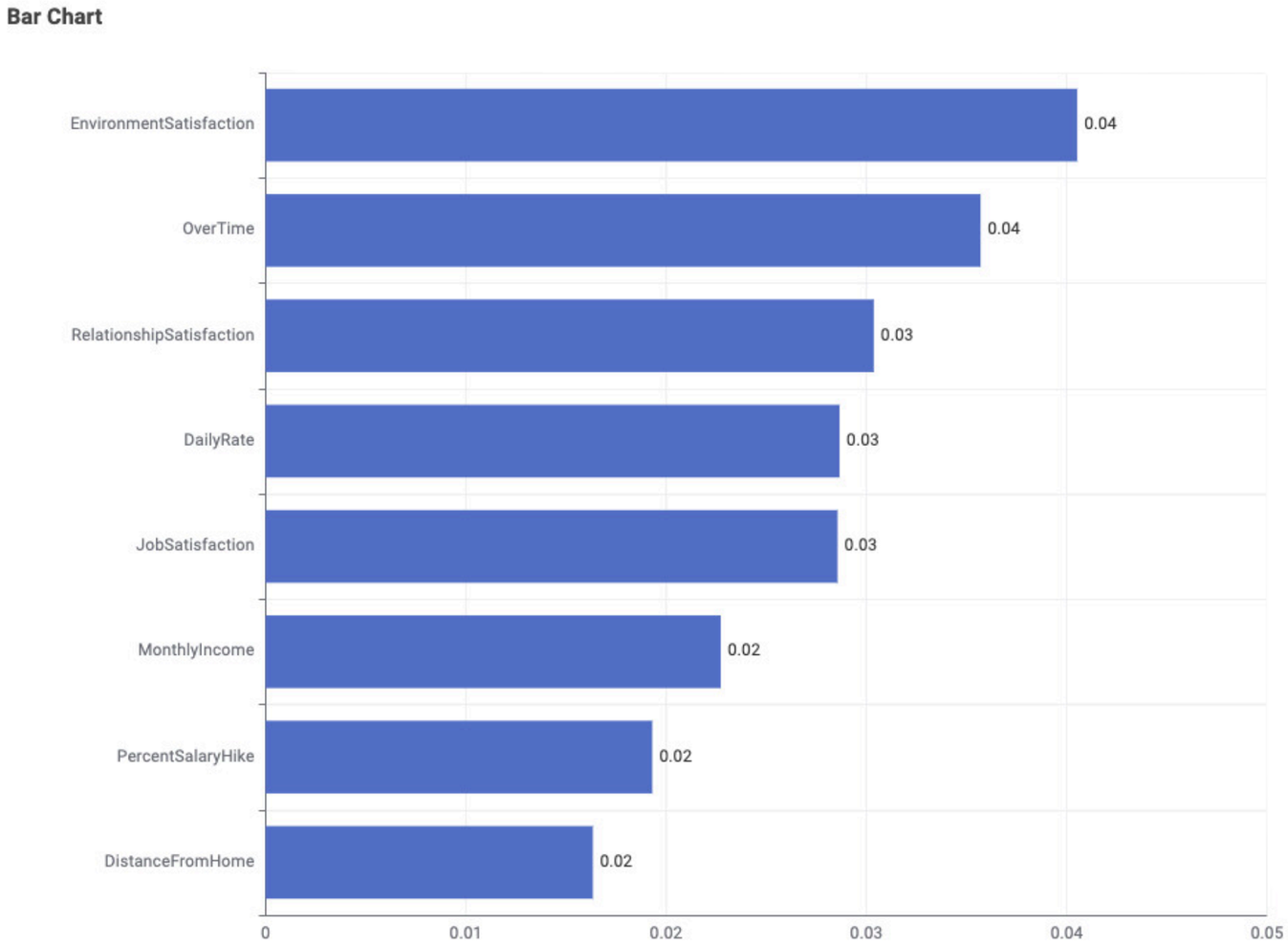
The graph shows that in the Logistic regression model the only features with a significant gap in their scores, are **HumanResources_JobRole** and **Travel_Frequently_BusinessTravel**.
The other features, however, do not have a big enough gap and therefore a significant importance rank cannot be built among them.



MLP

FEATURE IMPORTANCE

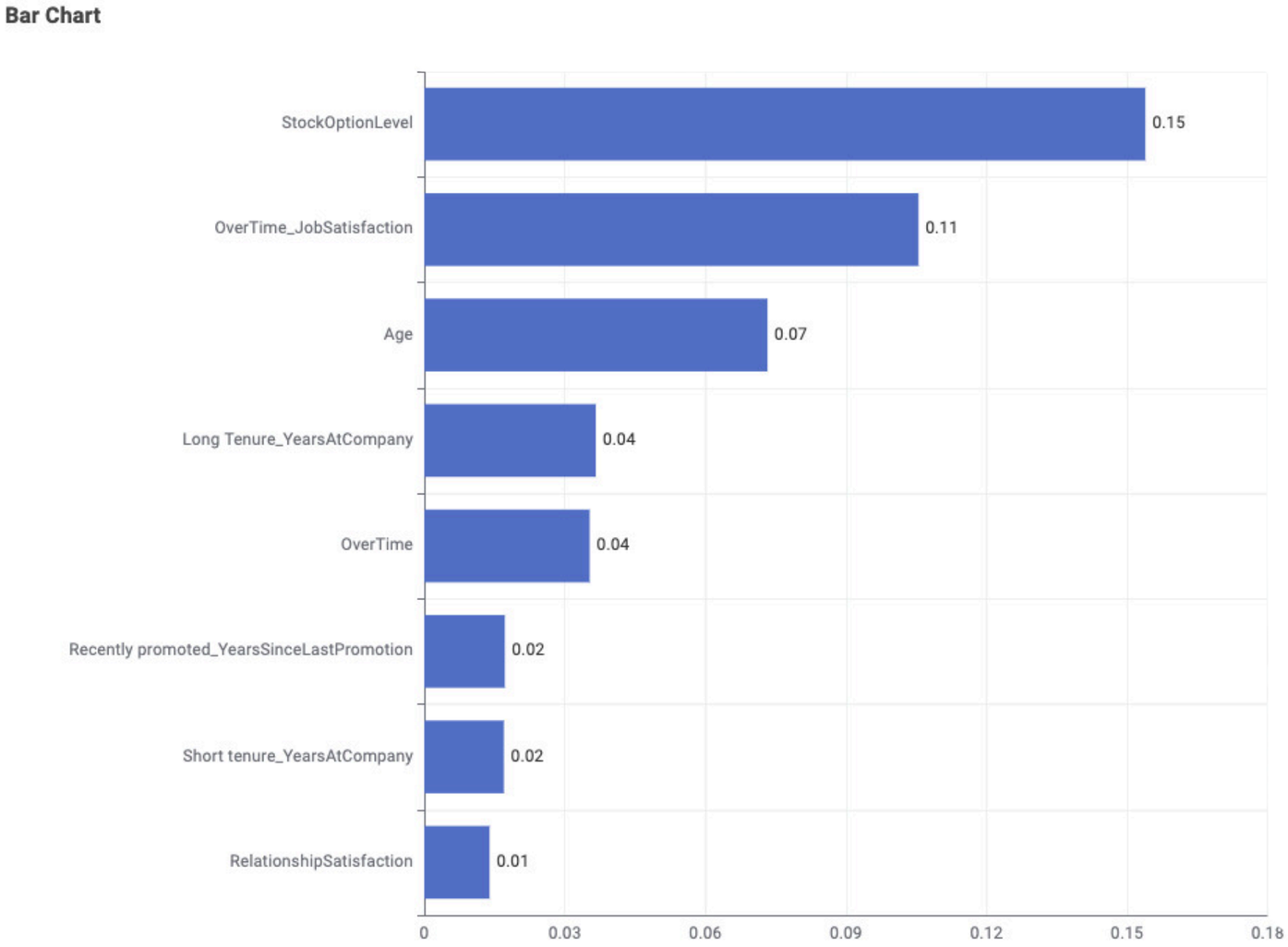
In the Multi-Layer Perceptron the only two features with a significant difference with the others are **EnviromentSatisfaction** and **OverTime**. Moreover, RelationshipSatisfaction, DailyRate and JobSatisfaction have similar values but the difference is too little to create a robust and reliable importance ranking.



DECISION TREE

FEATURE IMPORTANCE

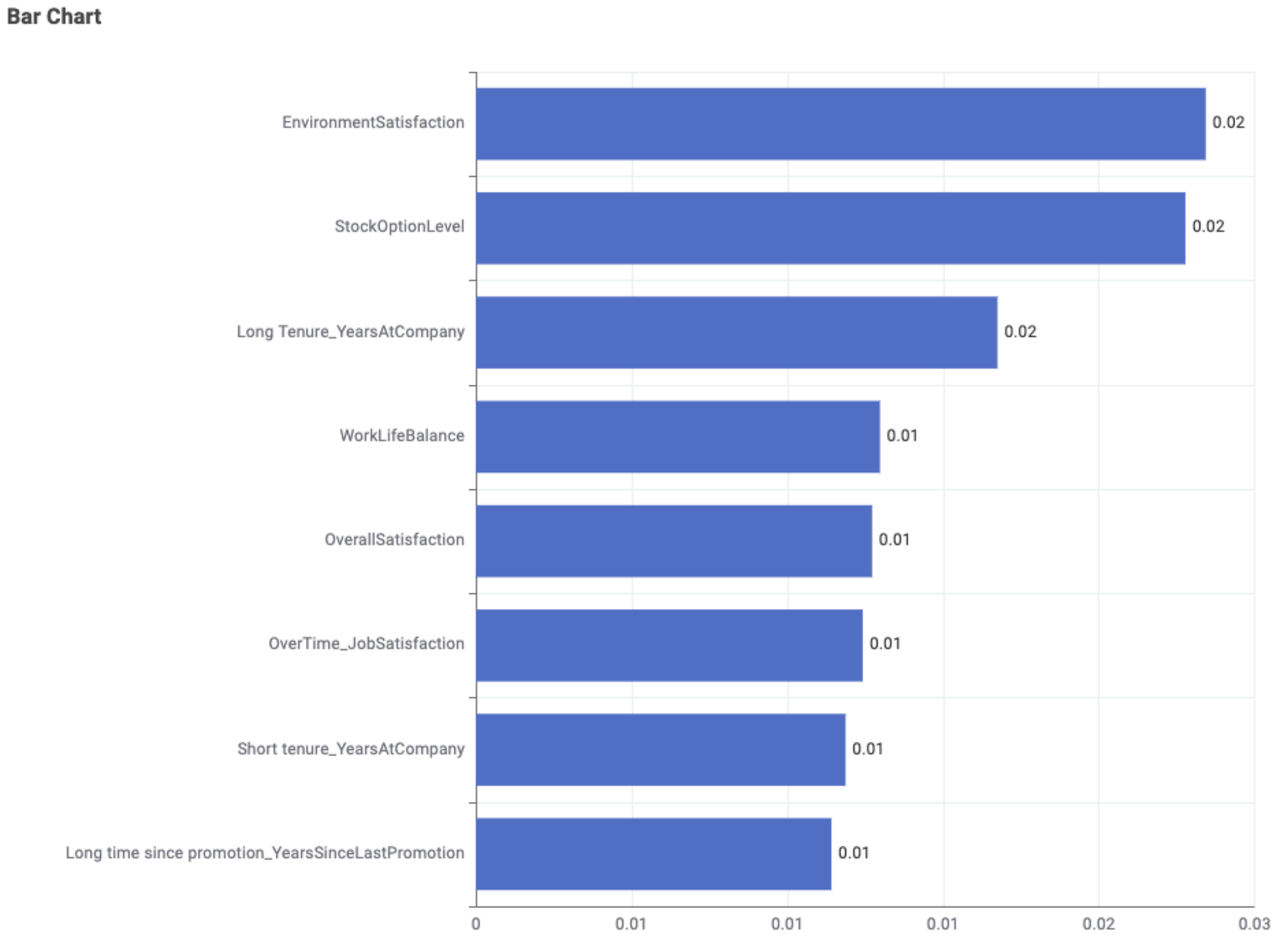
In the Decision tree we see that the top tree most important feature for the AUC score of the model are clearly **StockOptionLevel**, **OverTime_JobSatisfaction** and **Age**.



RANDOM FOREST

FEATURE IMPORTANCE

The bar chart shows EnviromentSatisfaction, StockOptionLevel displaying a higher importance score. However it is key to notice that the difference between all the values presented in the chart are **too little to be significant.**



05 MODELS COMPARISON

OVERALL

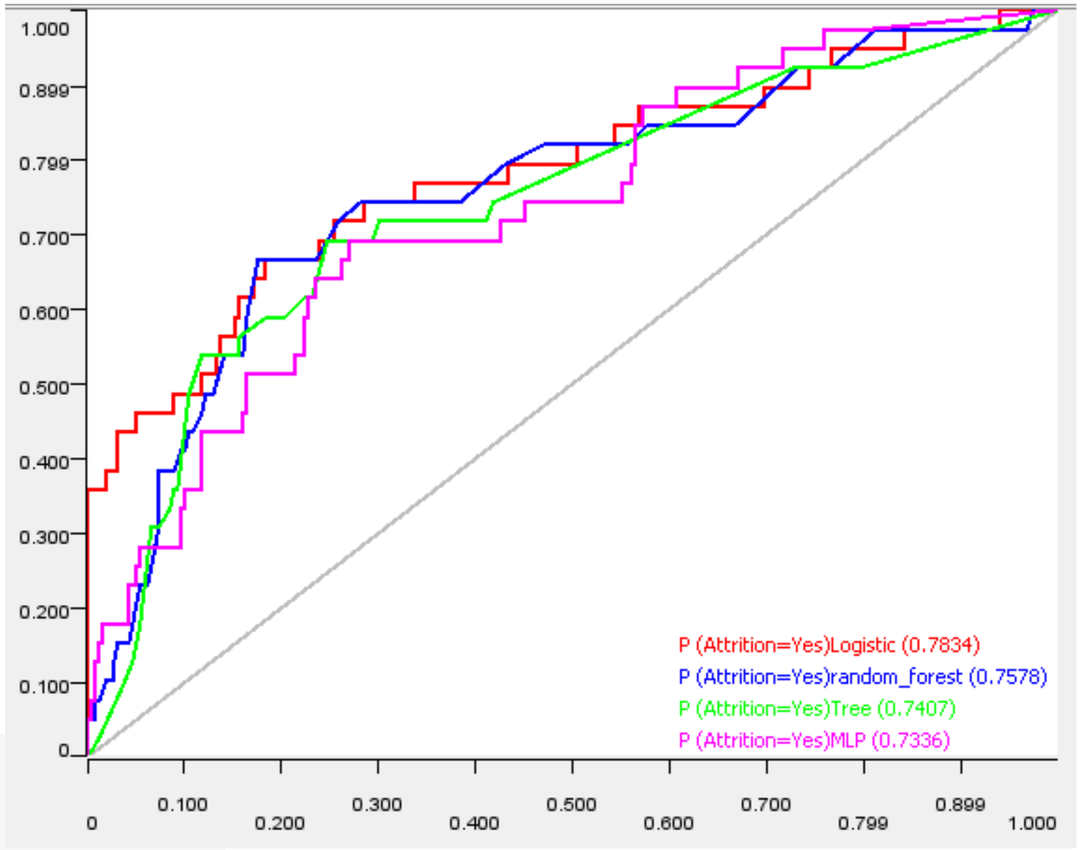
MODEL COMPARISON

Model Name	Accuracy	Recall	Precision	Sensitivity	Specificity	AUC	total costs
Logistic	0.796	0.667	0.356	0.667	0.816	0.783	668800
RANDOM FOREST	0.803	0.615	0.358	0.615	0.831	0.758	767200
DECISION TREE	0.745	0.692	0.3	0.692	0.753	0.741	625200
MLP	0.724	0.692	0.281	0.692	0.729	0.734	627600

- The table highlights the performance of four models across the selected metrics.
- **Random Forest stands out** with the highest accuracy and specificity, making it a strong performer, though it comes with the highest total cost.
- Logistic Regression offers a good **balance**, performing well in specificity and overall predictive power while keeping costs relatively low.
- The Decision Tree and MLP models show similar **recall**, but the Decision Tree edges ahead in accuracy and precision, also achieving the lowest cost.
- Meanwhile, **MLP lags slightly behind the others in most metrics**, particularly in precision and overall predictive ability.

ROC CURVES

MODEL COMPARISON



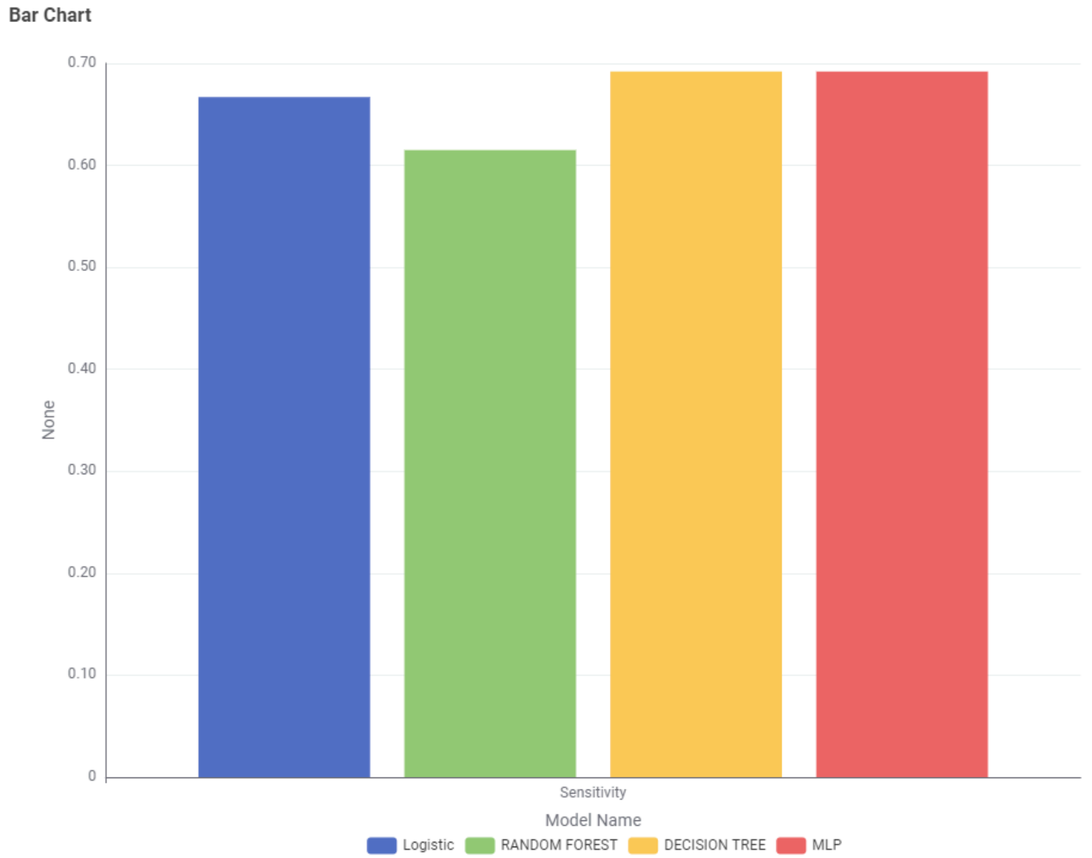
- **Logistic Regression** achieves the highest AUC (0.7834), indicating the best performance among the models.
- **Random Forest** follows with an AUC of 0.7578, showing slightly lower predictive capability.
- The **Decision Tree** model has an AUC of 0.7407, performing moderately well.
- The **Multi-Layer Perceptron** (MLP) has the lowest AUC (0.7336), indicating relatively weaker performance.

The ROC curves demonstrate that Logistic Regression maintains **superior discrimination ability** across various thresholds, while the MLP lags behind the others.

SENSITIVITY

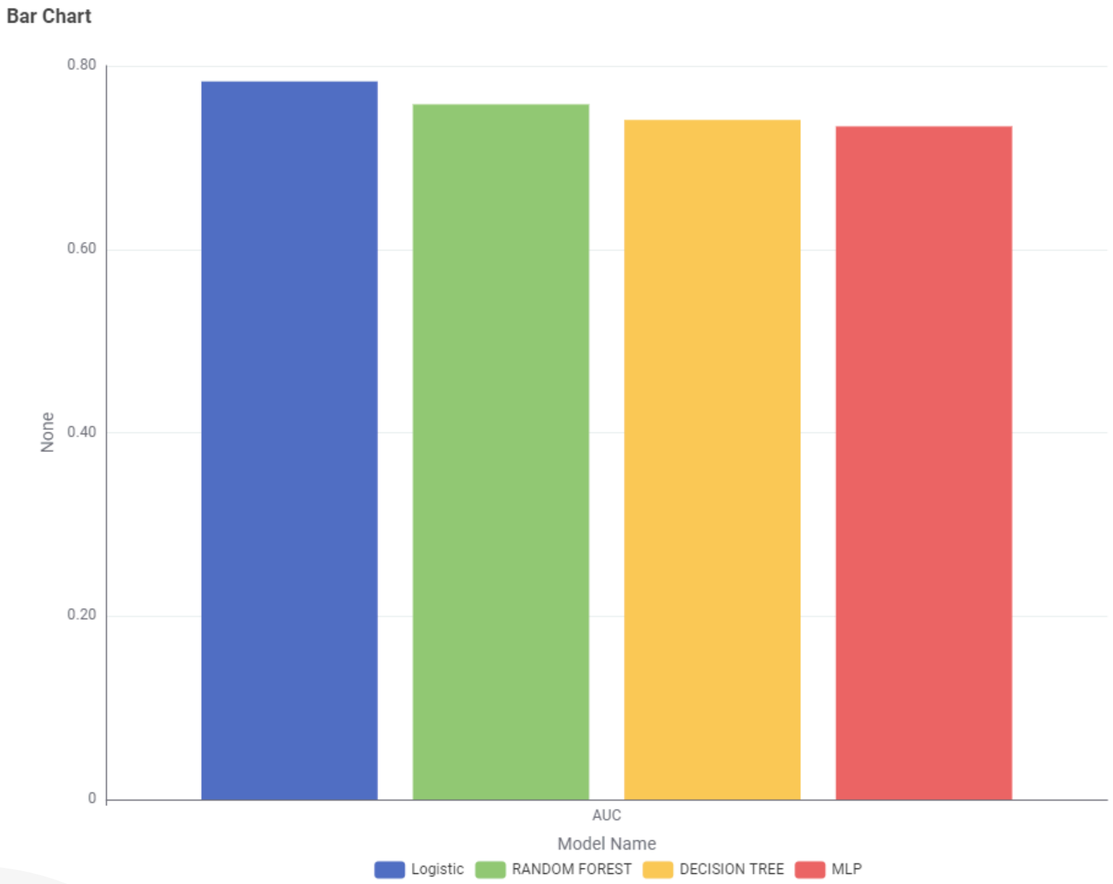
MODEL COMPARISON

The sensitivity metric is plotted on the y-axis, while the model names are categorized on the x-axis. The **MLP and Decision Tree** models exhibit the highest sensitivity values, closely followed by Logistic Regression, while Random Forest has the lowest among the four. Despite **slight variations**, all models demonstrate relatively similar performance, with sensitivities around 0.65 to 0.7, indicating consistent ability to correctly identify positive instances across the models.



AUC

MODEL COMPARISON

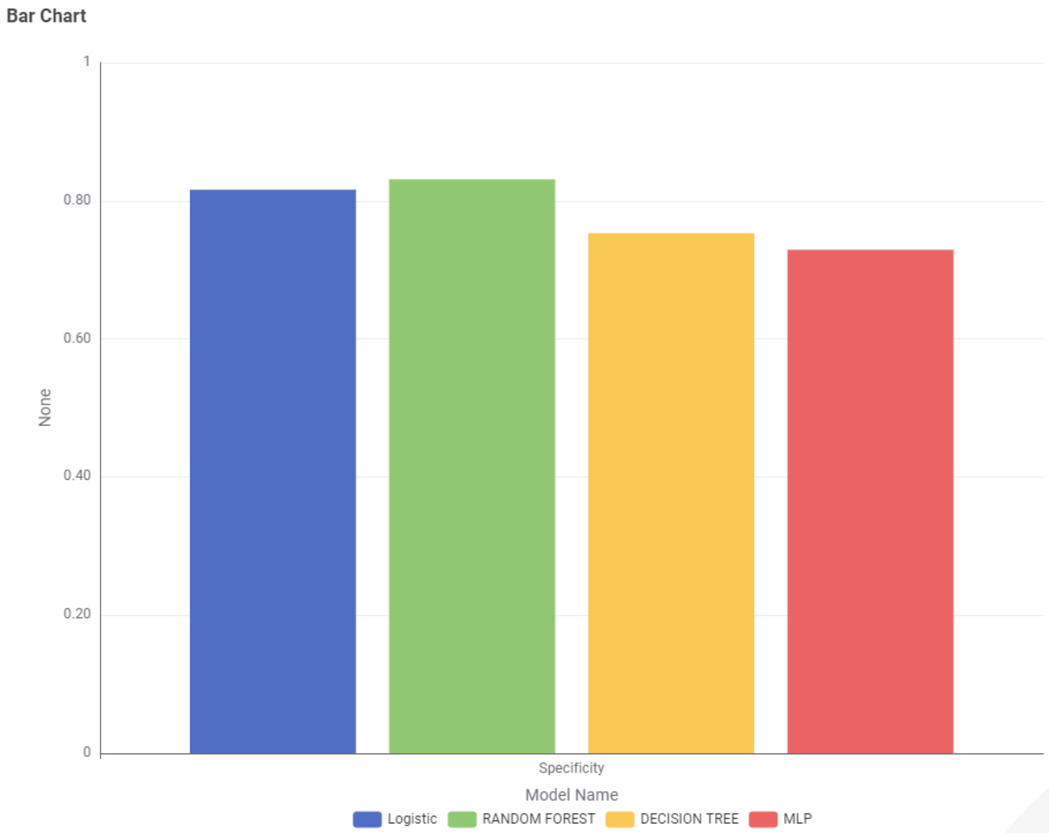


AUC, displayed on the y-axis, quantifies the ability of the models to distinguish between classes, with higher values indicating better discriminatory performance. The results show **comparable performance** across the models, with AUC values closely clustered around 0.75–0.8. The Logistic Regression and Random Forest models slightly outperform MLP and Decision Tree, suggesting marginally better classification capability. This uniformity implies **robust and competitive model behavior.**

SPECIFICITY

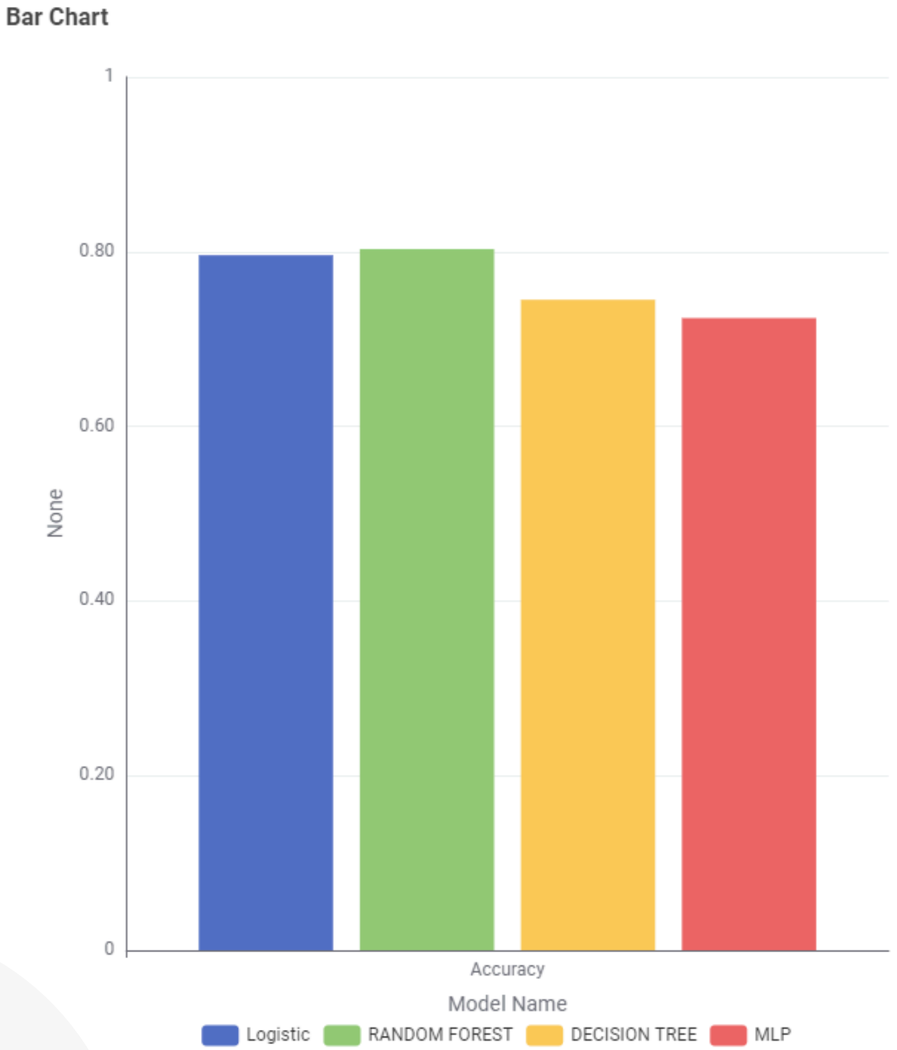
MODEL COMPARISON

Specificity, represented on the y-axis, indicates the models' ability to correctly identify negative instances. The **Random Forest model achieves the highest** specificity, slightly above the other models, which show relatively close performance. Logistic Regression follows closely, while Decision Tree and MLP exhibit slightly **lower values**, though the overall range remains high, around 0.75 to 0.85. These results highlight the models' effectiveness in minimizing false positive rates.



ACCURACY

MODEL COMPARISON

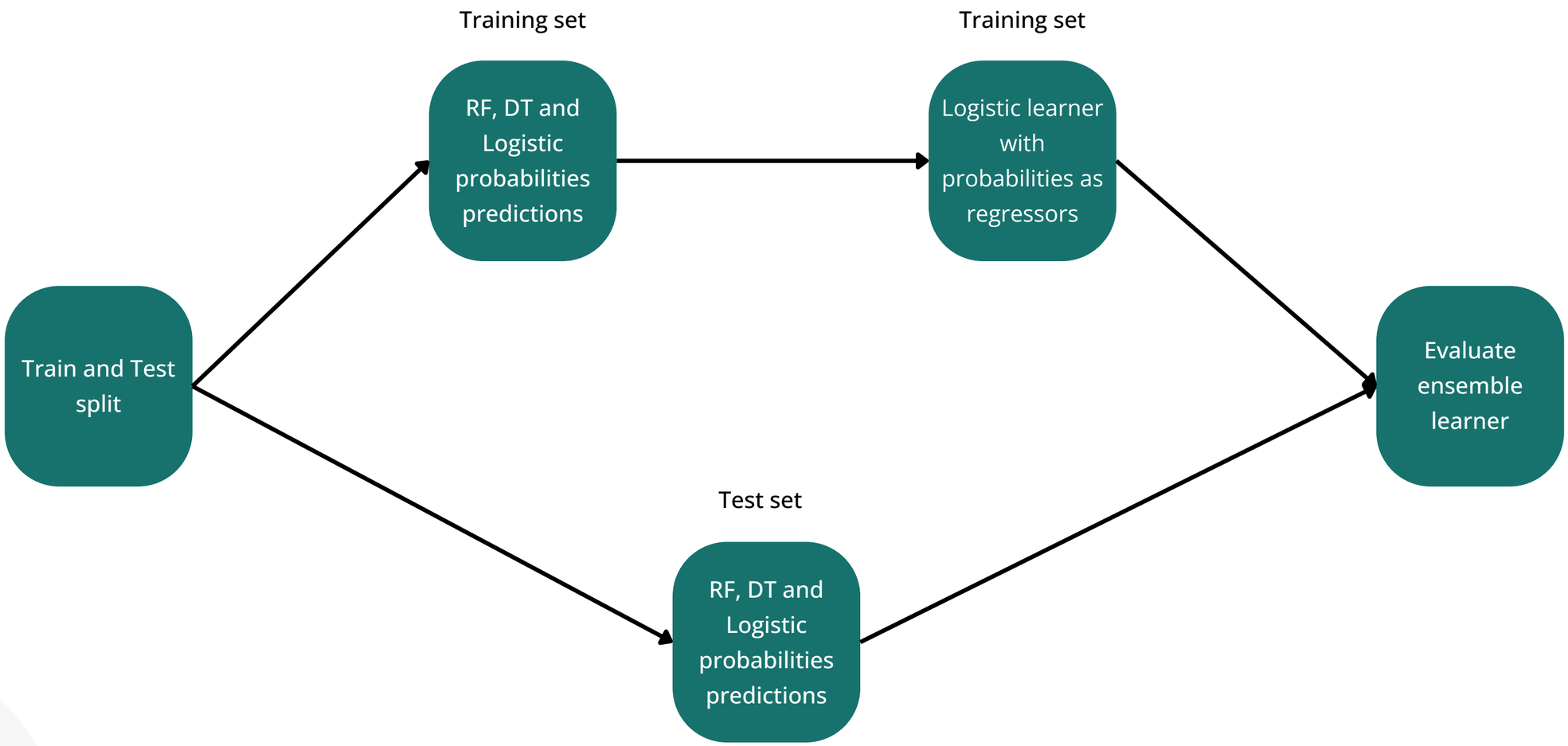


Accuracy, represented on the y-axis, measures the proportion of **correctly classified instances** among all instances. Logistic Regression and Random Forest achieve the highest accuracy, closely followed by Decision Tree and MLP, which have slightly lower but comparable values. The **range of accuracy is relatively narrow**, around 0.75 to 0.85, indicating that all models perform **similarly** in their ability to correctly classify instances in the dataset.

06 EXTRA LEARNER

TRAINING

ESNEMBLE



TRAINING

ESNEMBLE

We thought was interesting to explore the performances of an **ensemble learner** which leverages the strengths of base models—Logistic Regression, Random Forest, and Decision Tree—by combining their prediction probabilities using a meta-logistic regression approach. Instead of directly relying on the outputs of individual models, the learner takes their **probabilities as input features** and uses **logistic regression to optimize the final predictions.** We belived this technique would imporve performance by capturing complementary patterns and reducing individual model biases.

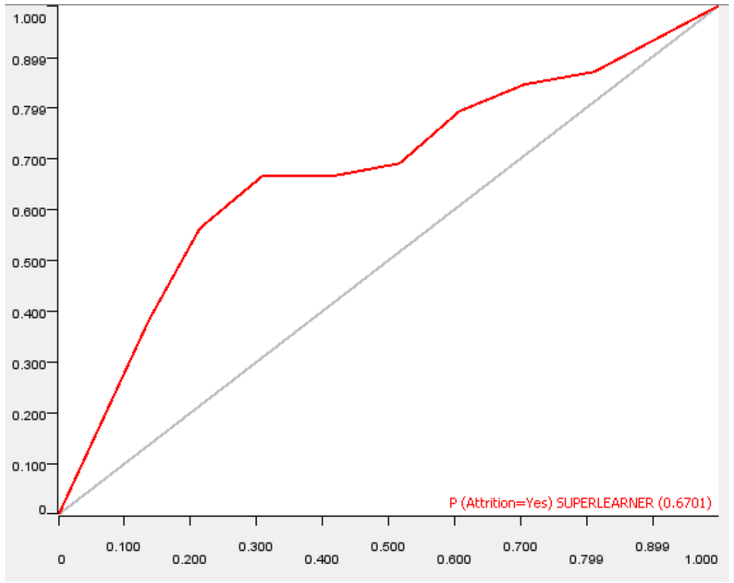
The training of this learner follows the following steps to **avoid any data leakage**:

1. Split train and test
2. Train:
    a. Train our base modeles on the training set
    b. Predict probabilities for training set probabilities
    c. Train a logistic learner with these probabilities as independent variables and Attrition as target
3. Test:
    a. make probability predictions with base models
    b. Output probability predictions from the metalearner based on base models predictions

RESULTS

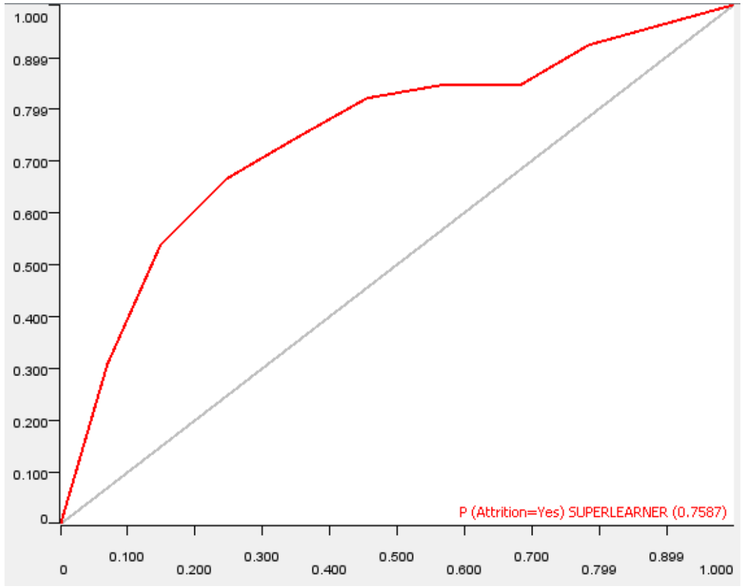
ESNEMBLE

ROC curve ensemble



After Laplace regularization

ROC curve ensemble



The ensemble learner initially underperformed: the final logistic regression struggled to generalize based on the base models' probabilities. By plotting the training error we have immediately seen how the problem was **overfitting**, therefore we applied regularization to the logistic learner to reduce the variance of our final ensemble.

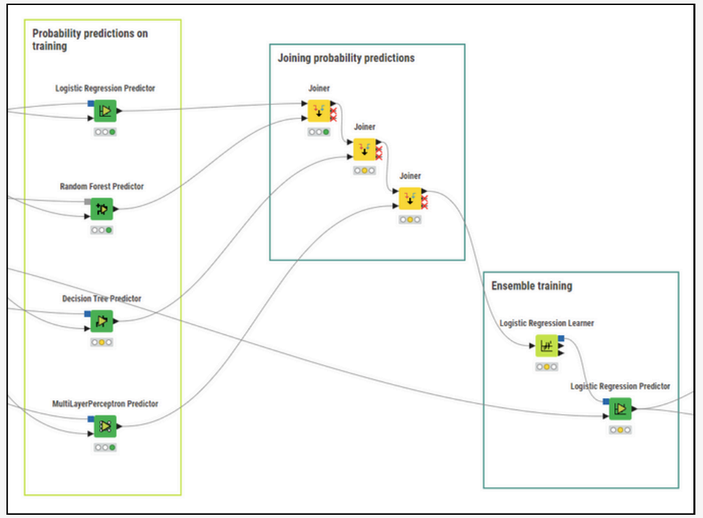
Specifically, applying the **Laplace regularization**, the results become comparable to the ones achieved with the base models. Still the training accuracy is very high, then probably further fine tuning of regularization would have benefitted performances

CONCLUSIONS

ESNEMBLE

Although the ensemble achieved the **second-highest AUC score** among the evaluated models, the improvement over simpler approaches was marginal. Given the considerable computational cost and complexity involved in training and maintaining such a combined ensemble, **its use is difficult to justify**. The added effort and resources required for training do not yield a substantial advantage in performance, making it **less practical** compared to more efficient and straightforward models that deliver comparable results.

Recall *Number (dou...*	Precision *Number (dou...*	Sensitivity *Number (dou...*	Specificity *Number (dou...*	F-Measure *Number (dou...*	AUC *Number (dou...*
0.667	0.366	0.667	0.824	0.473	0.759



MANAGERIAL IMPLICATIONS

01 KEY DRIVERS

02 POSSIBLE APPLICATIONS

01 KEY DRIVERS

KEY DRIVERS

This section focuses on the **detailed analysis** of the **most critical features** identified in our study, derived from the Logistic Regression model, which emerged as the best-performing algorithm in our evaluation. These features—**Human Resources (Job Role)**, **Travel Frequently (Business Travel)**, and **Overtime**—demonstrate the strongest influence on employee attrition. By examining these drivers, we aim to uncover actionable insights that can help address and reduce turnover effectively.

Human Resources (Job Role)

Travel Frequently (Business Travel)

Over Time

To address the high attrition in Human Resources, some solutions could be:

Role-Specific Interventions:

Provide tailored support for HR employees, such as mental health resources and stress management programs.
Encourage professional development and leadership training for HR roles to enhance engagement and satisfaction.

Feedback Loops:

HR employees may have unique insights into organizational gaps. Establish dedicated feedback mechanisms to ensure their voices are heard.

Fair Workload Distribution:

Address potential overwork concerns by evaluating workload distribution across HR functions, especially during periods of organizational change.

KEY DRIVERS

HUMAN RESOURCES (JOB ROLE)

KEY DRIVERS

BUSINESS TRAVEL: TRAVEL FREQUENTLY

BusinessTravel_TravelFrequently was one of the variables identified as **most predictive** in our logistic regression model. We also observe that the predicted coefficient for this variable is positive, indicating that employee travelling for work are **more likely to leave the company** compared to the reference group, i.e. non-travel.
This suggests that frequent travel may contribute to **work-life imbalance, stress, and career disengagement**, increasing the likelihood of employees leaving the company.

To address the high attrition risk associated with frequent business travel, the following interventions could be considered:

1. **Flexible Work Arrangements**: Offer remote work options or flexible schedules to reduce travel strain and improve work-life balance.
2. **Increased Travel Support**: Provide travel perks, allowances, and upgraded accommodations to make frequent travel more comfortable and less stressful.
3. **Travel Rotation or Limits**: Introduce policies that limit the frequency or duration of travel, or implement a travel rotation system to prevent employees from being consistently away from home.

Employees who work overtime frequently are likely experiencing **higher work-related stress**, **reduced work-life balance**, and possible **burnout**. These factors can contribute to higher attrition rates, as employees might leave to **seek better work-life balance.**

Possible solutions are:

- **Implement Overtime Pay or Time Off**: Offer fair compensation or additional leave to employees who work overtime to maintain motivation and reduce dissatisfaction.
- **Redistribute Workload**: Adjust team responsibilities to prevent over-reliance on a few employees for excessive overtime.
- **Promote Work-Life Balance**: Encourage flexible schedules, monitor overtime patterns, and foster a culture that prioritizes employee well-being.

KEY DRIVERS

OVER TIME

02 POSSIBLE APPLICATIONS

POSSIBLE APPLICATIONS

This analysis provides valuable insights into employee attrition patterns, enabling organizations to identify **key risk factors** and implement **targeted retention strategies**. By addressing these issues proactively, companies can **reduce turnover-related costs** and **improve overall employee satisfaction and engagement.**

Targeted Retention Strategies

Proactive Workfroce Planning

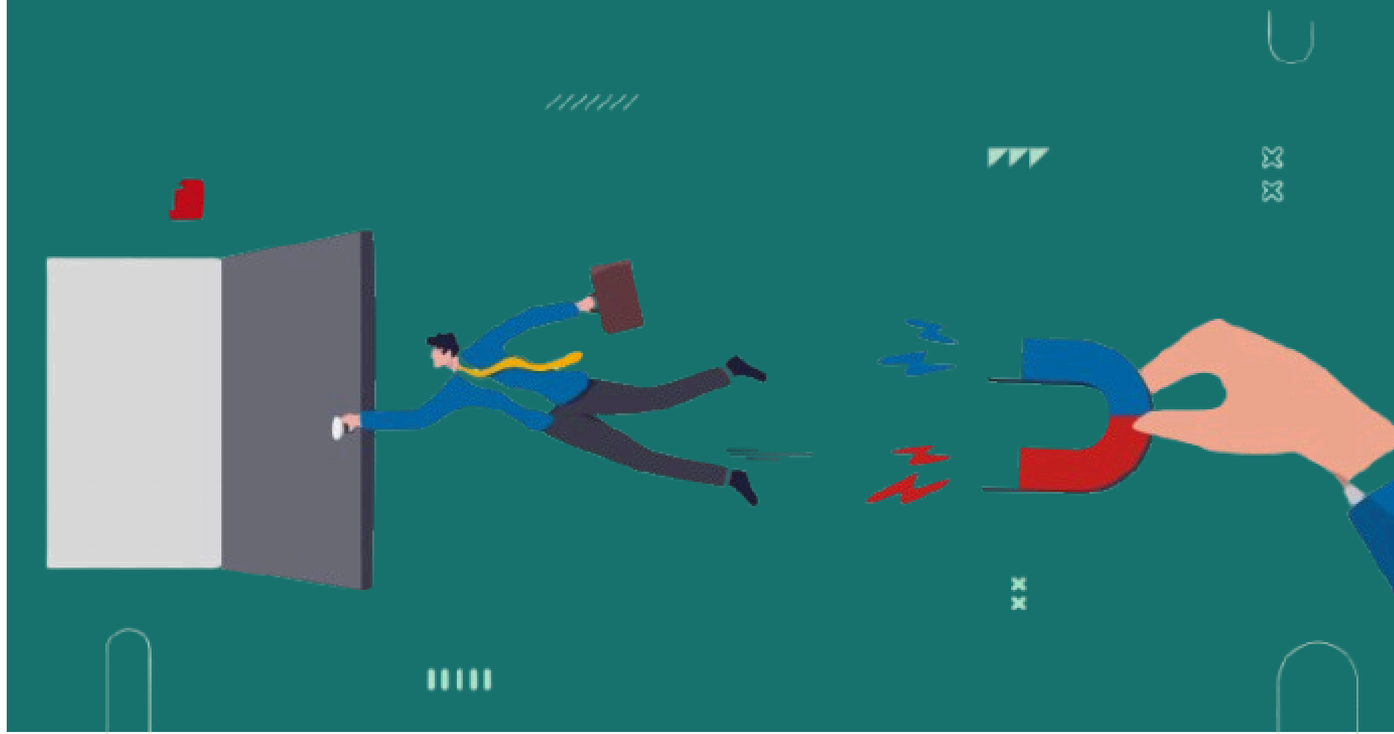
Improved Employee Experience

It is often assumed that simple, uniform approaches are the most effective in managing turnover. However, research has shown that personalized, **targeted interventions can be more cost-effective and impactful** (*Retaining Talent: Replacing Misconceptions With Evidence-Based Strategies*, Allen et al).

By identifying high-risk employees, predictive models provide valuable information that can be used by HR teams to **direct resources** to the areas where they will have the greatest impact. For instance, they could be used to create tailored retention programs to target high-performance or hard-to-replace employees with high attrition risk. By **focusing resources** on such groups, organizations can achieve higher retention rates for valuable talent while optimizing their overall cost-efficiency.

POSSIBLE APPLICATIONS

TAREGTED RETENTION STRATEGIES



POSSIBLE APPLICATIONS

TAREGTED RETENTION STRATEGIES

POSSIBLE TARGETED INTERVENTIONS

- **Flexible Work Arrangements**: Employees identified as at-risk due to poor work-life balance could be offered flexible schedules or remote work options.
- **Personalized Incentives**: Employees dissatisfied with compensation or benefits could be offered targeted salary adjustments, bonuses, or improved benefit packages.
- **Development Programs**: For employees with concerns about career progression, customised training and mentorship programs can be developed to enhance their skills and prepare them for promotions or lateral moves.
- **Enhanced Engagement**: For employees showing early signs of disengagement, initiatives such as team-building activities, recognition programs, or more frequent feedback sessions can be implemented.

These personalised retention strategies not only improve employee satisfaction but also contribute to reduced turnover rates, enhanced morale, and a more engaged workforce.

POSSIBLE APPLICATIONS

PROACTIVE WORKFORCE PLANNING

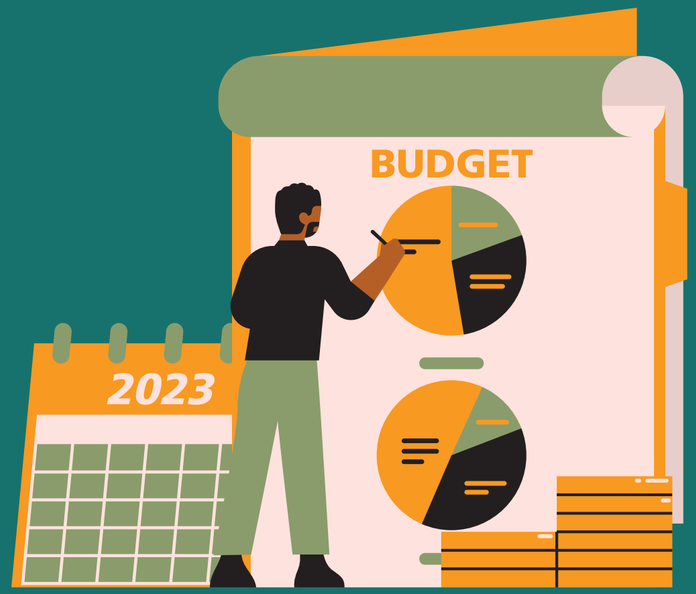
Integrating attrition prediction models into HR systems provides organizations with valuable insights to effectively **plan for workforce changes**, **minimize operational disruptions**, and ensure the organization remains **fully staffed** to meet its objectives. Two critical aspects of these predictive models include:

Anticipating Critical Vacancies

Anticipating Critical Vacancies Predictive models help HR teams identify employees who are at high risk of leaving. With this insight, organizations can **initiate succession planning -** either by **grooming internal candidates** to take over key roles or by **starting external hiring processes** in advance - to minimize the time critical positions remain vacant.

POSSIBLE APPLICATIONS

PROACTIVE WORKFORCE PLANNING



Optimized Recruitment Budgets

Forecasting attrition rates allows organizations to **allocate recruitment budgets** and resources more effectively. By predicting the number and types of employees likely to leave within a specific timeframe, HR departments can **avoid overstaffing or under-preparing for hiring needs**. This insight leads to more strategic use of recruitment funds.

These predictive capabilities give HR departments a **forward-looking perspective**, allowing them to focus on strategic workforce management and minimize the financial and operational costs of attrition.

POSSIBLE APPLICATIONS

IMPROVED EMPLOYEE EXPERIENCE

Insights from the analysis can directly **inform** changes to **workplace policies**, fostering an environment that promotes employee satisfaction and engagement:

Policy Adjustments

The analysis can inform revisions to policies related to **workload management**, vacation days, or overtime compensation, addressing **stress** and **burnout**—two key factors often linked to turnover, as addressed in *Maslach & Leiter, 2016*. **Flexible policies** tailored to employee needs can reduce attrition by fostering a healthier **work-life balance.**

Enhanced Feedback Systems

Establishing robust, continuous employee **feedback** systems allows organizations to **track job satisfaction** and **detect emerging issues**. Research (*Rosen et al., 2016*) suggests that proactive listening mechanisms significantly **reduce turnover** by creating a **responsive work environment**.

POSSIBLE APPLICATIONS

IMPROVED EMPLOYEE EXPERIENCE

Transparent Career Pathing

Lack of **career growth** is a primary driver of **attrition** *(Ng et al., 2005)*. Organizations can address this by creating clearly **defined career paths** with actionable milestones and offering regular discussions about **professional development opportunities**. This not only enhances motivation but also builds trust in organizational commitment to employees'.

Inclusive Work Culture

Identifying and addressing **workplace inequities**, such as gender pay gaps or favoritism in promotions, can build a culture of **fairness** and **inclusivity**. Studies show that **perceived fairness** in organizational practices significantly impacts employee retention (*Colquitt et al., 2001)*. Training managers on **unbiased decision-making** and establishing transparent criteria for rewards and promotions are key steps.

BIBLIOGRAPHY

- Allen, D. G., Bryant, P. C., & Vardaman, J. M. (2010). Retaining Talent: Replacing Misconceptions with Evidence-Based Strategies. Business Horizons, 53(1), 49-54.
- Maslach, C., & Leiter, M. P. (2016). Understanding the burnout experience: Recent research and its implications for psychiatry.World Psychiatry, 15(2), 103–111. https://doi.org/10.1002/wps.20311
- Rosen, C. C., Kacmar, K. M., Harris, K. J., Gavin, M. B., Hochwarter, W. A., & Ferris, G. R. (2016). Workplace politics and stress: The moderating role of resources. Journal of Vocational Behavior, 95–96, 102–113. https://doi.org/10.1016/j.jvb.2016.07.004
- Ng, T. W. H., Eby, L. T., Sorensen, K. L., & Feldman, D. C. (2005). Predictors of objective and subjective career success: A meta-analysis. Personnel Psychology, 58(2), 367–408. https://doi.org/10.1111/j.1744-6570.2005.00515.x
- Colquitt, J. A., Conlon, D. E., Wesson, M. J., Porter, C. O., & Ng, K. Y. (2001). Justice at the millennium: A meta-analytic review of 25 years of organizational justice research. Journal of Applied Psychology, 86(3), 425–445. https://doi.org/10.1037/0021-9010.86.3.425

THE TEAM

Alice Finotti

Leonardo Tonelli

Filippo Grandoni

Ludovico Panariello

Tommaso Ravasio

Federico Giorgi

THANK YOU

FOR YOUR ATTENTION

2 December 2024