Employee Attrition Project

DDS ANALYTICS
TALENT MANAGEMENT SOLUTIONS FOR FORTUNE 100 COMPANIES

Presented By: Mike Flores
March 9, 2025

AGENDA

- Executive Summary
- Attrition Cost Analysis
- Classification Modeling
- Exploratory Data Analysis (EDA)
- Questions

Executive Summary

DDS Analytics is pleased to present an initial analysis, predictive modeling strategy, and key findings to **assist Frito Lay in understanding and mitigating employee attrition**. Our approach leverages advanced analytics to identify key drivers of attrition, predict potential turnover, and estimate the financial impact of our intervention strategies.

Objective

Frito Lay aims to gain insight into employee attrition and proactively implement measures to retain key talent.

- **Identifying top factors** contributing to employee turnover.
- **Building predictive models** to anticipate attrition.
- **Estimating the financial implications** of attrition and intervention strategies.

Predictive Model Development

We developed two classification models to predict employee attrition, **k-Nearest Neighbor**, and **Naïve Bayer**. Our best-performing model achieves a high degree of accuracy in predicting attrition risk, enabling targeted retention efforts.

By utilizing our predictive model, Frito Lay can focus retention strategies on employees most likely to leave, optimizing the use of incentives and reducing turnover-related expenses. Our preliminary cost-benefit analysis suggests that implementing these targeted interventions can generate substantial cost savings.

Attrition Cost Analysis

Cost Analysis – KNN Model Predictions

- **Cost of Attrition = $10,000**: Cost incurred when an employee leaves the company (e.g., hiring, training, lost productivity).
- **Cost of Retention Intervention = $2,500:** Cost associated with attempting to retain an employee (e.g., bonuses, promotions, PTO, reskilling, additional support).

The formula used:

$$\text{Total Cost} = (\text{Cost of Attrition} \times \text{False Negatives}) + (\text{Cost of Retention Intervention} \times \text{False Positives})$$

$$\text{Total Cost} = (\$10{,}000 \times 16) + (\$2{,}500 \times 38)$$

Total Cost = $255,000

5

Classification Modeling

KNN & Naïve Bayer Models

KNN & Naïve Bayer Model Statistics

KNN Model

		ACTUALS: No (No Attrition)	ACTUALS: Yes (Attrition)
PREDICTED	No (No Attrition)	182 (True Negative)	16 (False Negative)
	Yes (Attrition)	38 (False Positive)	25 (True Positive)

KNN Confusion Matrix

- **Accuracy** = model **correctly predicted 79.3% of all cases** (both attrition and non-attrition).
- **Sensitivity (25 TP)** = model **correctly identified 82.7% of employees who actually left** (TP) but, might be misclassifying employees who will stay (FP)
- **Specificity (182 TN)** = only **correctly identified 61.0% of employees who actually stayed** (TN).

Naïve Bayer Model

		ACTUALS: 0 (No Attrition)	ACTUALS: 1 (Attrition)
PREDICTED	0 (No Attrition)	168 (True Negative)	14 (False Positive)
	1 (Attrition)	57 (False Negative)	22 (True Positive)

Naïve Bayer Confusion Matrix

- **Accuracy** = model **correctly predicted 72.8% of all cases** (both attrition and non-attrition).
- **Sensitivity (22 TP)** = model **correctly identified 74.7% of employees who actually left** (TP) but, might be misclassifying employees who will stay (FP)
- **Specificity (168 TN)** = **only correctly identified 61.1% of employees who actually stayed** (TN).

7

Modeling Comparison

Comparison of Confusion Matrix:

Metric	KNN Model	Naïve Bayer Model
Accuracy	79.3%	72.8%
Sensitivity	82.7%	74.7%
Specificity	61.0%	61.1%

Cost Modeling Rationale: **Applied KNN Model** predictions to Cost Model.

- **Accuracy: KNN model** has **higher accuracy** (79.3% vs. 72.8%).
- **Sensitivity: KNN model** has **better sensitivity** (82.7% vs. 74.7%), meaning it correctly identifies more cases of non-attrition.
- **Specificity: Naïve Bayer model** is **slightly better in specificity** (61.1% vs. 61.0%), meaning it makes slightly fewer false positive predictions, **but these differences are minimal**.

Exploratory Data Analysis (EDA)

Data Dictionary

Variable	Description	Levels within Variables
ID	Employee's unique company ID	
Age	Employee's age in years	
Attrition	Indicates whether the employee has left the company	Yes = Left company, No = Still employed
BusinessTravel	Employee's frequency of business travel	Non-Travel, Travel_Rarely, Travel_Frequently
DailyRate	Employee's daily salary rate of pay (USD)	
Department	Employee's related company department	Human Resources, Research & Development, Sales
DistanceFromHome	Employee's distance traveled from home to workplace (miles)	
Education	Employee's highest level of education achieved	1 = Below College, 2 = College, 3 = Bachelor, 4 = Master, 5 = Doctor
EducationField	Employee's field of education	Human Resources, Life Sciences, Marketing, Medical Sciences, Other, Technical Degree
EmployeeCount	A constant value of 1 for all records, possibly used as a placeholder	
EmployeeNumber	Unique identifier assigned to each employee	
EnvironmentSatisfaction	Employee's satisfaction with their work environment on a scale from 1 to 4	1 = Low, 2 = Medium, 3 = High, 4 = Very High
Gender	Employee's gender	Male, Female
HourlyRate	Employee's hourly wage (USD)	
JobInvolvement	Employee's level of job involvement on a scale from 1 to 4	1 = Low, 2 = Medium, 3 = High, 4 = Very High
JobLevel	Employee's job level within the company on a scale from 1 to 5	1 = Entry Level, 2 = Junior Level, 3 = Mid Level, 4 = Senior Level, 5 = Executive Level
JobRole	Employee's specific role within the company	Healthcare Representative, Human Resources, Laboratory Technician, Manager, Manufacturing Director, Research Director, Research Scientist, Sales Executive, Sales Representative
JobSatisfaction	Employee's satisfaction with their job on a scale from 1 to 4	1 = Low, 2 = Medium, 3 = High, 4 = Very High
MaritalStatus	Employee's marital status	Single, Married, or Divorced
MonthlyIncome	Employee's monthly salary (USD)	
MonthlyRate	Employee's monthly wage rate (USD)	
NumCompaniesWorked	Number of companies the employee has previously worked for	
Over18	Indicates whether the employee is over 18 years old	Y = Yes
OverTime	Whether the employee works overtime	Yes, No
PercentSalaryHike	Employee's percentage increase in salary	
PerformanceRating	Employee's performance rating on a scale from 1 to 4	1 = Low, 2 = Good, 3 = Excellent, 4 = Outstanding
RelationshipSatisfaction	Satisfaction with work relationships on a scale from 1 to 4	1 = Low, 2 = Medium, 3 = High, 4 = Very High
StandardHours	Standard working hours	80
StockOptionLevel	Employee's stock option level on a scale from 0 to 3	0 = None, 1 = Low, 2 = Medium, 3 = High
TotalWorkingYears	Employee's total years of work experience	
TrainingTimesLastYear	Number of training sessions attended in the past year	
WorkLifeBalance	Employee's work-life balance rating on a scale from 1 to 4	1 = Bad, 2 = Good, 3 = Better, 4 = Best
YearsAtCompany	Number of years the employee has worked at the company	
YearsInCurrentRole	Number of years in the employee's current role	
YearsSinceLastPromotion	Number of years since employee's last promotion	
YearsWithCurrManager	Number of years the employee has worked with their current manager	

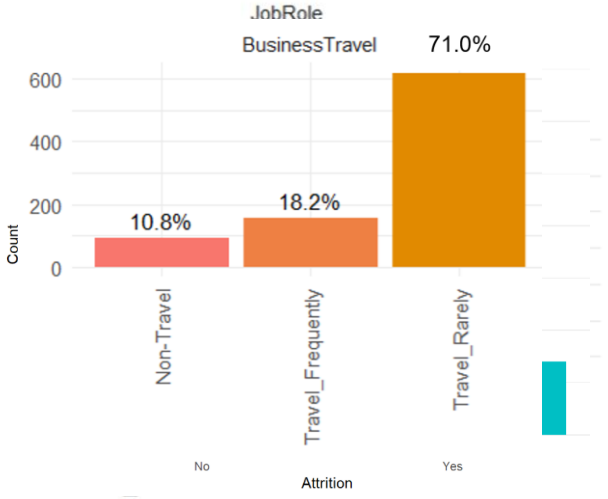
Frito Lay Data Sample

Variable	Description	Levels within Variables
ID	Employee's unique company ID	
Age	Employee's age in years	
Attrition	Indicates whether the employee has left the company	Yes = Left company, No = Still employed
BusinessTravel	Employee's frequency of business travel	Non-Travel, Travel_Rarely, Travel_Frequently

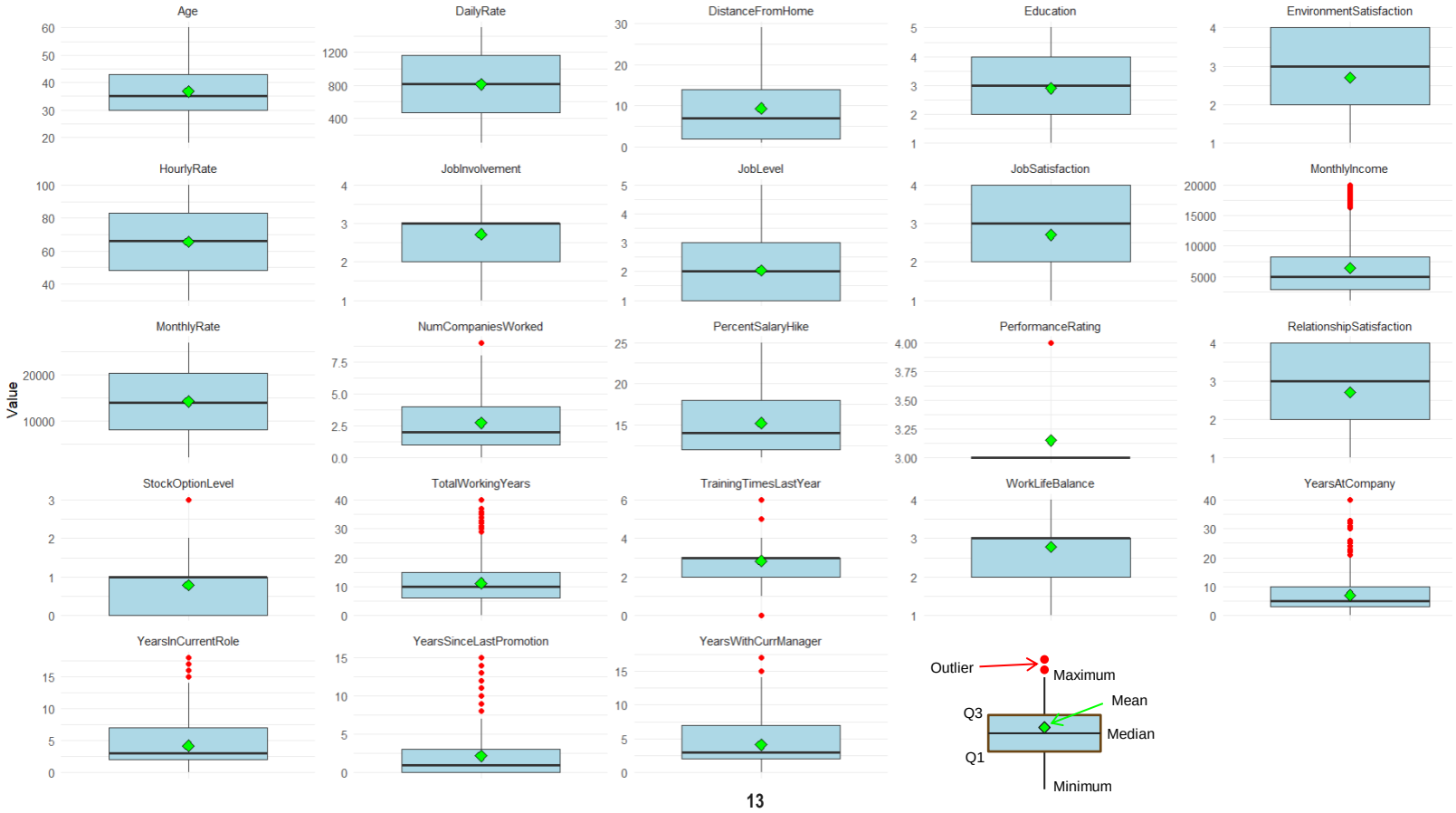
- Dataset has 870 rows and 36 columns.
- The 36 columns consist of; 27 – Quantitative & 9 – Qualitative (Categorical) data types
- Removed non-informative columns, e.g., little to no relationship to Attrition
- Verified that no missing data was present across all columns.
- Retained all outliers as they were valid numerical values.

Data Overview – Attrition Frequency

- **Attrition Rate**: ***140 employees (16%) left the company***, making retention a key concern.
- **Business Travel**: ***158 employees travel frequently, which may contribute to attrition*** compared to the 618 who rarely travel.
- **Overtime**: ***252 employees work overtime, which could impact job satisfaction and burnout,*** while 618 do not.
- **Education Field**: ***Life Sciences (40%) and Medical (31%) dominate,*** showing potential industry-specific retention patterns.
- **Job Roles & Leadership**: ***33% (287) of employees hold leadership roles, which may indicate retention challenges*** at different career levels.



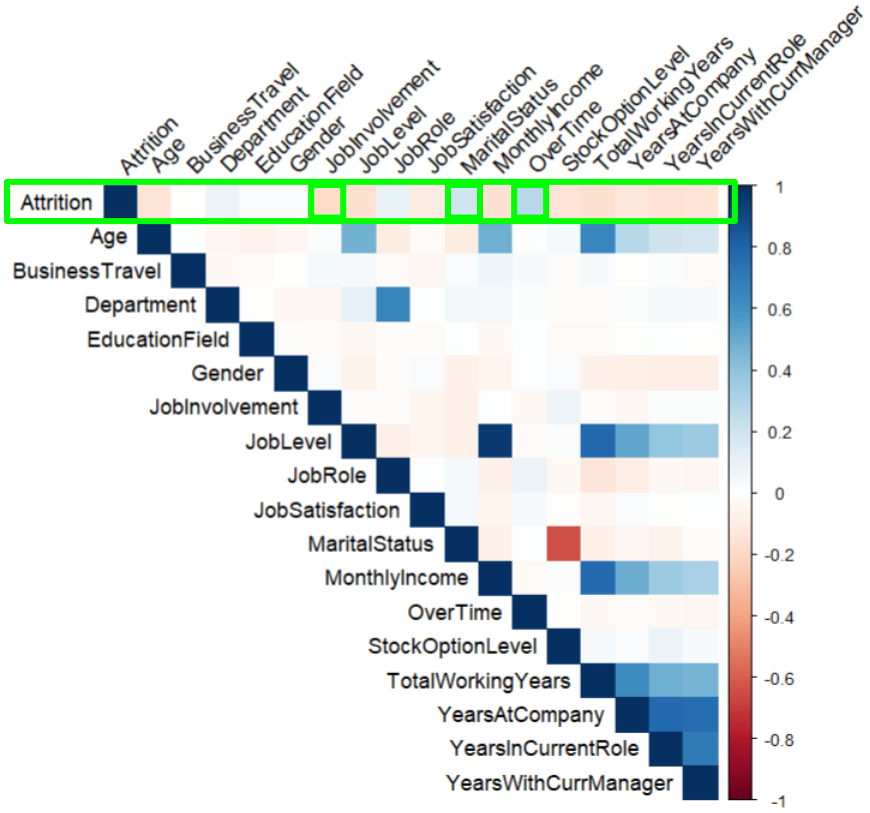
Data Overview – Quantitative Variable Statistical Summary



Data Overview – Quantitative Variable Statistical Summary

- **Years at Company**: Tenure ranges from **0 to 40 years**, averaging **6.96 years**. **50% have stayed 5 years or less**, while **25% have been here for 3 years or less**.
- **Job Satisfaction**: **Scores range from 1 to 4, averaging 2.71. Half rate it 3 or lower, and 25% report low satisfaction (2 or less), affecting retention.**
- **Work-Life Balance**: **Ratings span 1 to 4, averaging 2.78. Half rate it 3 or lower, suggesting work-life challenges that may drive attrition.**
- **Years Since Last Promotion**: **Time since last promotion ranges from 0 to 15 years, averaging 2.17 years. 25% have waited at least 3 years, which may impact motivation.**
- **Num Companies Worked**: **Employees have worked at 0 to 9 companies, averaging 2.73. 25% have worked at 4+ companies, indicating higher mobility.**

Correlation Between Variables – Relationships to Attrition



Correlation Hierarchy

Rank	Category	Correlation
1	OverTime	0.2720
2	MaritalStatus	0.1970
3	JobInvolvement	0.1878
4	TotalWorkingYears	0.1672
5	JobLevel	0.1621
6	YearsInCurrentRole	0.1562
7	MonthlyIncome	0.1549
8	Age	0.1494
9	StockOptionLevel	0.1487
10	YearsWithCurrManager	0.1468
11	YearsAtCompany	0.1288
12	JobSatisfaction	0.1075
13	JobRole	0.0905
14	Department	0.0870
15	EducationField	0.0261
16	Gender	0.0253
17	BusinessTravel	0.0061

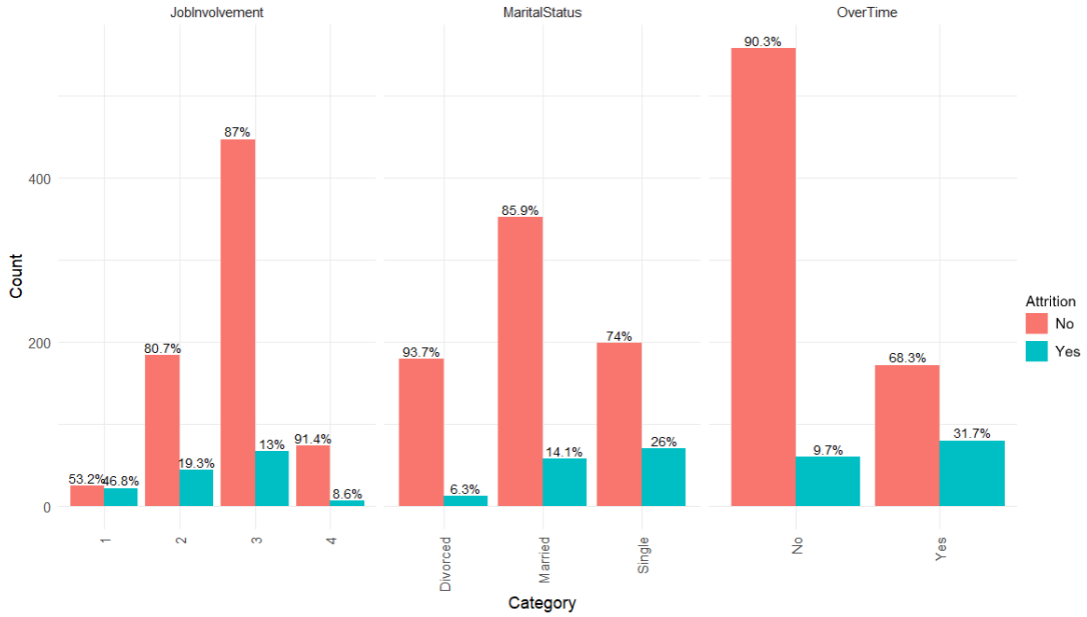
Reading the Correlation Matrix

- Attrition impact from the variables
- Visualize the strength of relationships
- Darker the color the stronger the relationship to Attrition.
- No color, little to no relationship to Attrition.

Data Overview – Attrition vs. Highest Correlations

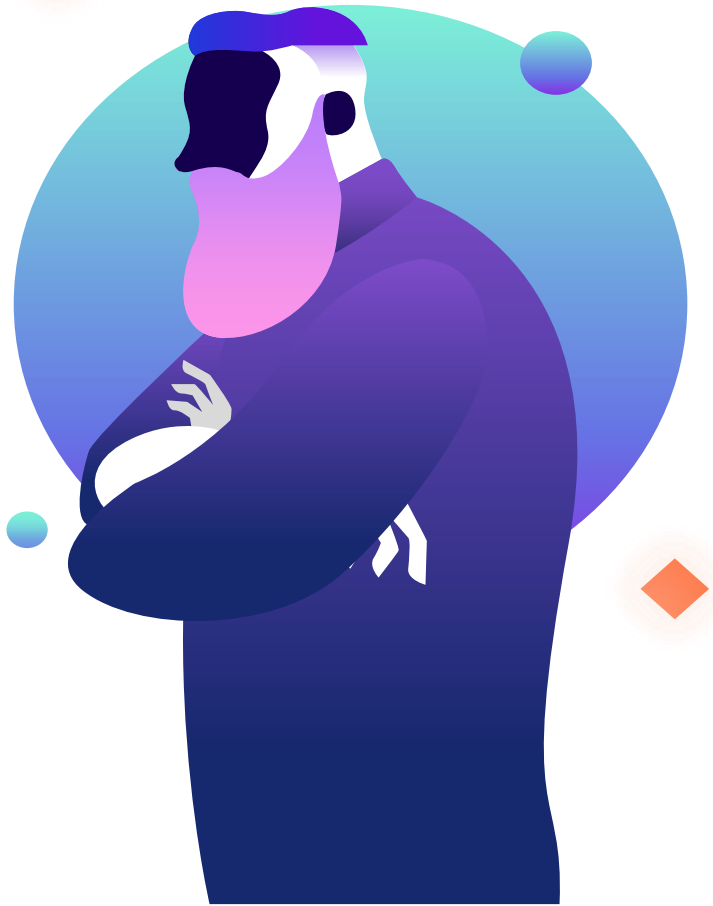
Reading the Correlation Matrix

- Attrition impact from the variables
- Visualize the strength of relationships
- Darker the color the stronger the relationship to Attrition.
- No color, little to no relationship to Attrition.



QUESTIONS?

DDS ANALYTICS
TALENT MANAGEMENT SOLUTIONS
FOR FORTUNE 100 COMPANIES



DDS ANALYTICS
TALENT MANAGEMENT SOLUTIONS FOR FORTUNE 100 COMPANIES

Thank You

Contact Info:

Mike Flores
972-439-6307
mikeflores@mail.smu.edu

Appendix

Solution Approach

- **Data** – Collect and acquire relevant data from various sources.
- **Wrangle** – Clean, preprocess, and transform the data for analysis.
- **Explore** – Conduct exploratory data analysis (EDA) to identify patterns and insights.
- **Model** – Build, train, and evaluate predictive models using statistical and machine learning techniques.
- **Report** – Communicate findings and insights through visualizations, reports, and presentations.

Data Overview

Sample Data: First 5 Rows:

ID	Attrition	BusinessTravel	DailyRate	Department	DistanceFromHome	Education	EducationField	EmployeeCount	EmployeeNumber	EnvironmentSatisfaction	Gender	HourlyRate	Age	JobInvolvement
1	No	Travel_Rarely	117	Sales	13	4	Life Sciences	1	859	2	Male	73	32	3
2	No	Travel_Rarely	1308	Research & Development	14	3	Medical	1	1128	3	Male	44	40	2
3	No	Travel_Frequently	200	Research & Development	18	2	Life Sciences	1	1412	3	Male	60	35	3
4	No	Travel_Rarely	801	Sales	1	4	Marketing	1	2016	3	Female	48	32	3
5	No	Travel_Frequently	567	Research & Development	2	1	Technical Degree	1	1646	1	Female	32	24	3
6	No	Travel_Frequently	294	Research & Development	10	2	Life Sciences	1	733	4	Male	32	27	3

ID	JobLevel	JobRole	JobSatisfaction	MaritalStatus	MonthlyIncome	MonthlyRate	NumCompaniesWorked	Over18	OverTime	PercentSalaryHike	PerformanceRating	RelationshipSatisfaction
1	2	Sales Executive	4	Divorced	4403	9250	2	Y	No	11	3	3
2	5	Research Director	3	Single	19626	17544	1	Y	No	14	3	1
3	3	Manufacturing Director	4	Single	9362	19944	2	Y	No	11	3	3
4	3	Sales Executive	4	Married	10422	24032	1	Y	No	19	3	3
5	1	Research Scientist	4	Single	3760	17218	1	Y	Yes	13	3	3
6	3	Manufacturing Director	1	Divorced	8793	4809	1	Y	No	21	4	3

ID	StandardHours	StockOptionLevel	TotalWorkingYears	TrainingTimesLastYear	WorkLifeBalance	YearsAtCompany	YearsInCurrentRole	YearsSinceLastPromotion	YearsWithCurrManager
1	80	1	8	3	2	5	2	0	3
2	80	0	21	2	4	20	7	4	9
3	80	0	10	2	3	2	2	2	2
4	80	2	14	3	3	14	10	5	7
5	80	0	6	2	3	6	3	1	3
6	80	2	9	4	2	9	7	1	7

- Dataset has 870 rows and 36 columns.
- The 36 columns consist of; 27 – Quantitative & 9 – Qualitative (Categorical) data types
- There is not any missing data in any of the columns.
- There are several columns which add little (no) value to the solution: ID, EmployeeCount, EmployeeNumber, Over18, StandardHours

Data Overview – Statistical Summary

Quantitative Variables - 5 Number Summaries

5 Number Summary	Age	DailyRate	DistanceFromHome	Education	EnvironmentSatisfaction	HourlyRate	JobInvolvement	JobLevel	JobSatisfaction	MonthlyIncome	MonthlyRate	NumCompaniesWorked	PercentSalaryHike
Min Value:	18	$103.00	1	1	1	$30.00	1	1	1	$1,081.00	$2,094.00	0	11%
1st Quartile (Q1):	30	$472.50	2	2	2	$48.00	2	1	2	$2,840.00	$8,092.00	1	12%
Median (Q2):	35	$817.50	7	3	3	$66.00	3	2	3	$4,946.00	$14,074.00	2	14%
Mean :36.83	36.83	$815.20	9.339	2.901	2.701	$65.61	2.723	2.039	2.709	$6,390.00	$14,326.00	2.728	15%
3rd Quartile (Q3):	43	$1,165.80	14	4	4	$83.00	3	3	4	$8,182.00	$20,456.00	4	18%
Max Value:	60	$1,499.00	29	5	4	$100.00	4	5	4	$19,999.00	$26,997.00	9	25%

5 Number Summary	PerformanceRating	RelationshipSatisfaction	StockOptionLevel	TotalWorkingYears	TrainingTimesLastYear	WorkLifeBalance	YearsAtCompany	YearsInCurrentRole	YearsSinceLastPromotion	YearsWithCurrManager
Min Value:	3	1	0	0	0	1	0	0	0	0
1st Quartile (Q1):	3	2	0	6	2	2	3	2	0	2
Median (Q2):	3	3	1	10	3	3	5	3	1	3
Mean :36.83	3.152	2.707	0.7839	11.05	2.832	2.782	6.962	4.205	2.169	4.14
3rd Quartile (Q3):	3	4	1	15	3	3	10	7	3	7
Max Value:	4	4	3	40	6	4	40	18	15	17

- **Min Value (Minimum)** – The smallest data point in the dataset.
- **First Quartile** (Q1/ 25th Percentile) – The median of the lower half of the dataset. 25% of the data falls below this value.
- **Second Quartile** (Q2/ Median / 50th Percentile) – The middle value of the dataset, splitting it into two equal halves. 50% or half of the data falls below this value.
- **Third Quartile** (Q3/ 75th Percentile) – The median of the upper half of the data. 75% of the data falls below this value.
- **Max Value** (Maximum) – The largest data point in the dataset.

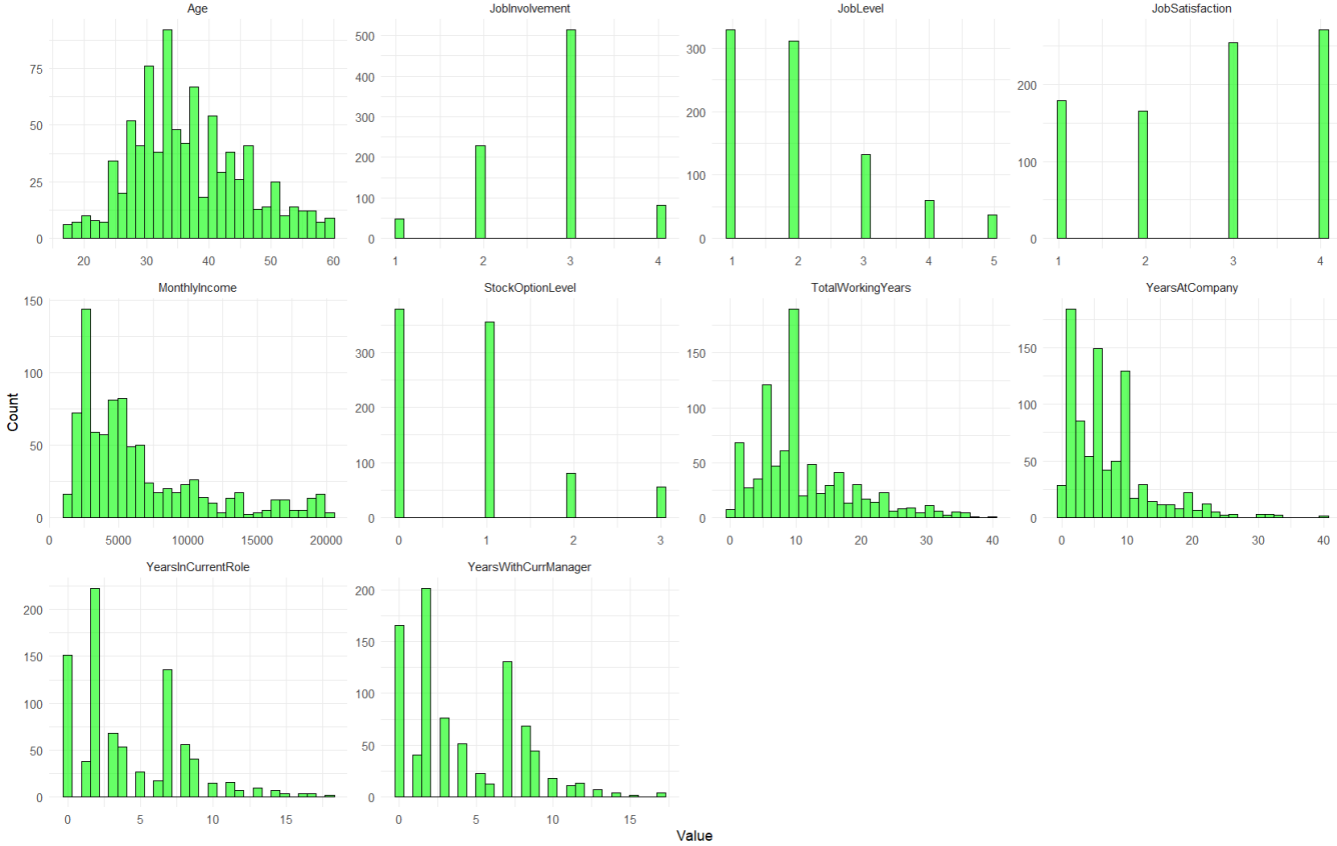
Data Overview – Statistical Summary

Categorical Variables

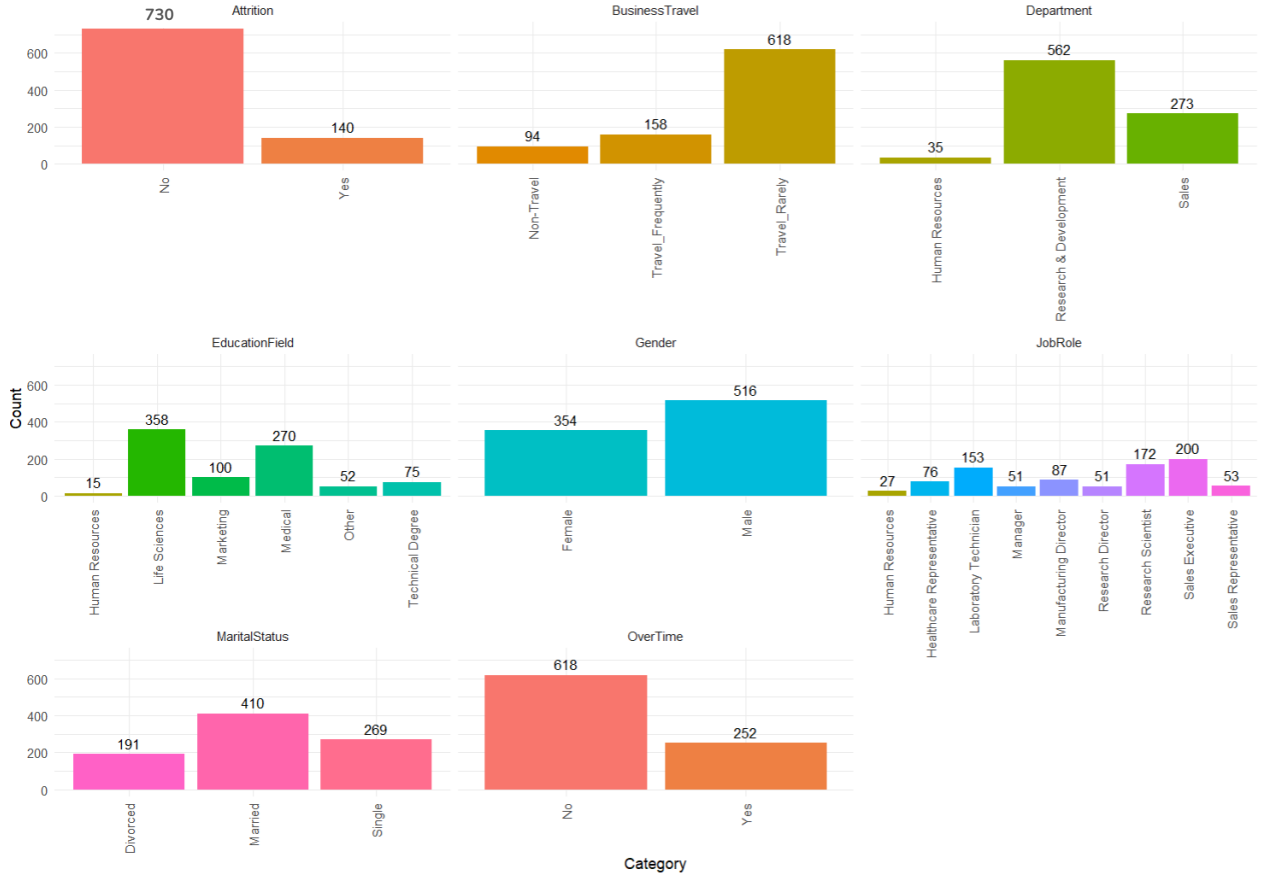
Attrition	BusinessTravel	Department	EducationField	Gender	JobRole	MaritalStatus	OverTime
No: 730 (84%)	Non-Travel: 94	Human Resources: 35	Human Resources: 15	Female: 354	Sales Executive: 200	Divorced: 191	No: 618
Yes: 140 (16%)	Travel_Frequently: 158	Research & Development: 562	Life Sciences: 358	Male: 516	Research Scientist: 172	Married: 410	Yes: 252
	Travel_Rarely: 618	Sales: 273	Marketing: 100		Laboratory Technician: 153	Single: 269	
			Medical: 270		Manufacturing Director: 87		
			Other: 52		Healthcare Representative: 76		
			Technical Degree: 75		Sales Representative: 53		

- **140 (16%) Employees** left the company.
- Majority, **618**, of employees rarely travel.
- Human resources staff makes up less than **4%** of company.
- **40%** of company **is educated in Life Sciences**, **31% is educated in Medical**.
- Almost **60%** of company is Male.
- **33%** (287) of **employees are leadership roles**.
- **47%** (410) of employees are **Married**.
- Majority, **618**, of employees do not work overtime.

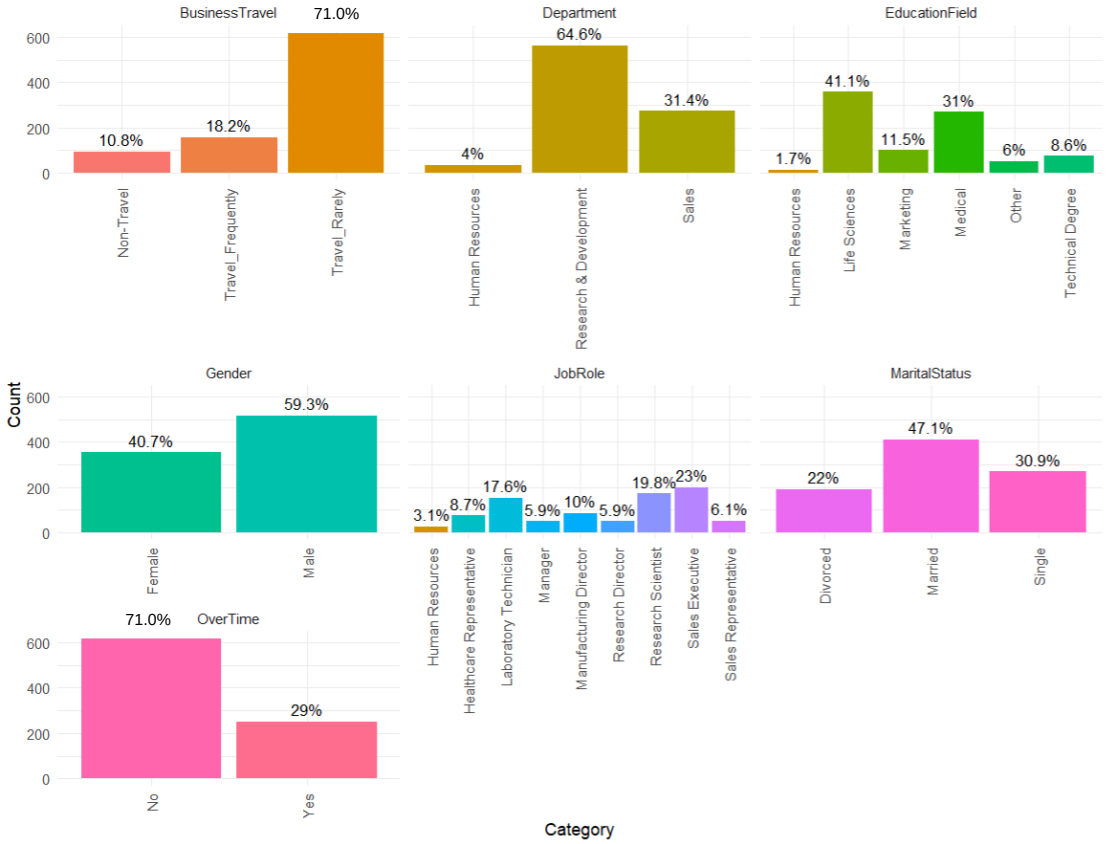
Data Overview – Quantitative Variable Frequency



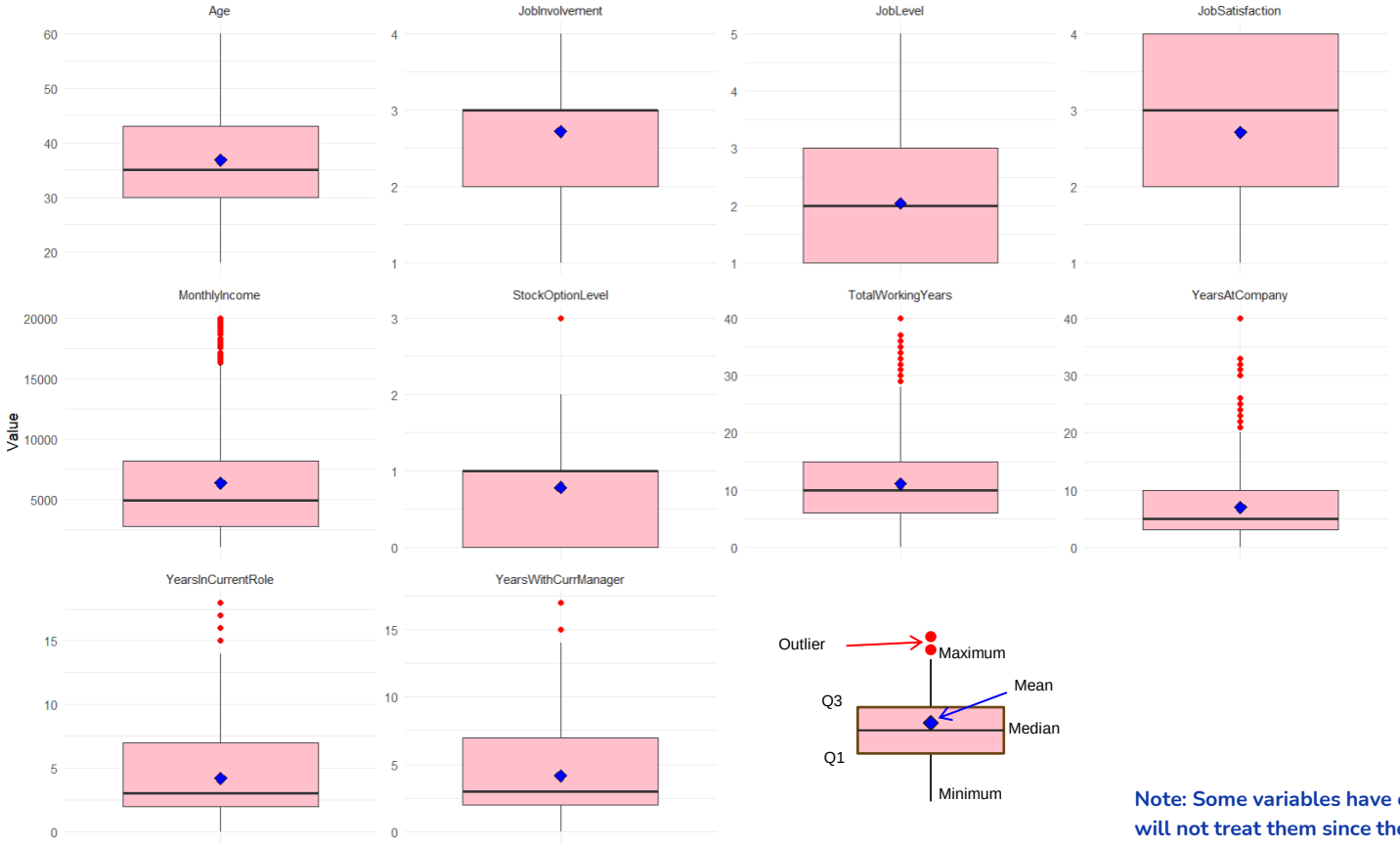
Data Overview – Categorical Variable Frequency



Data Overview – Categorical Variable Frequency

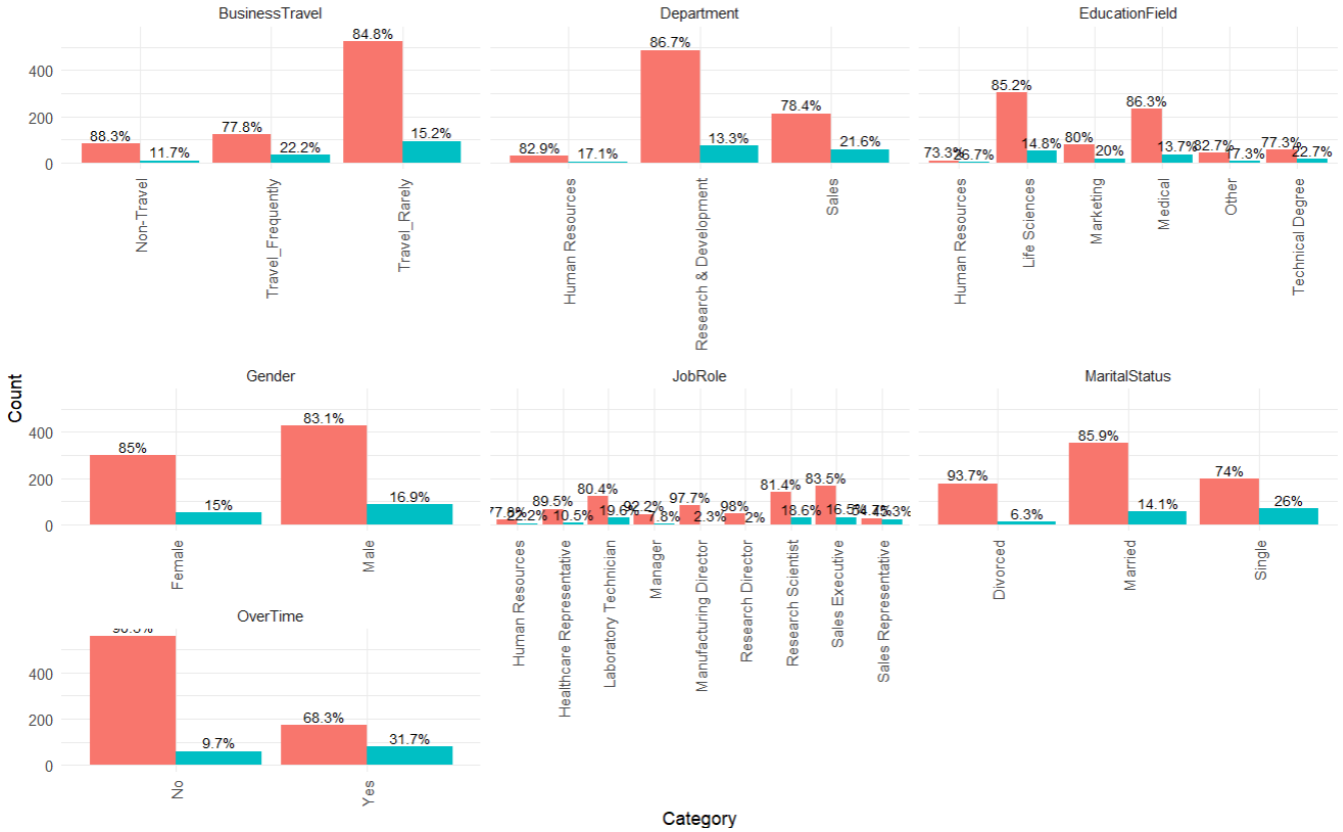


Data Overview – Quantitative Variable Statistical Summary



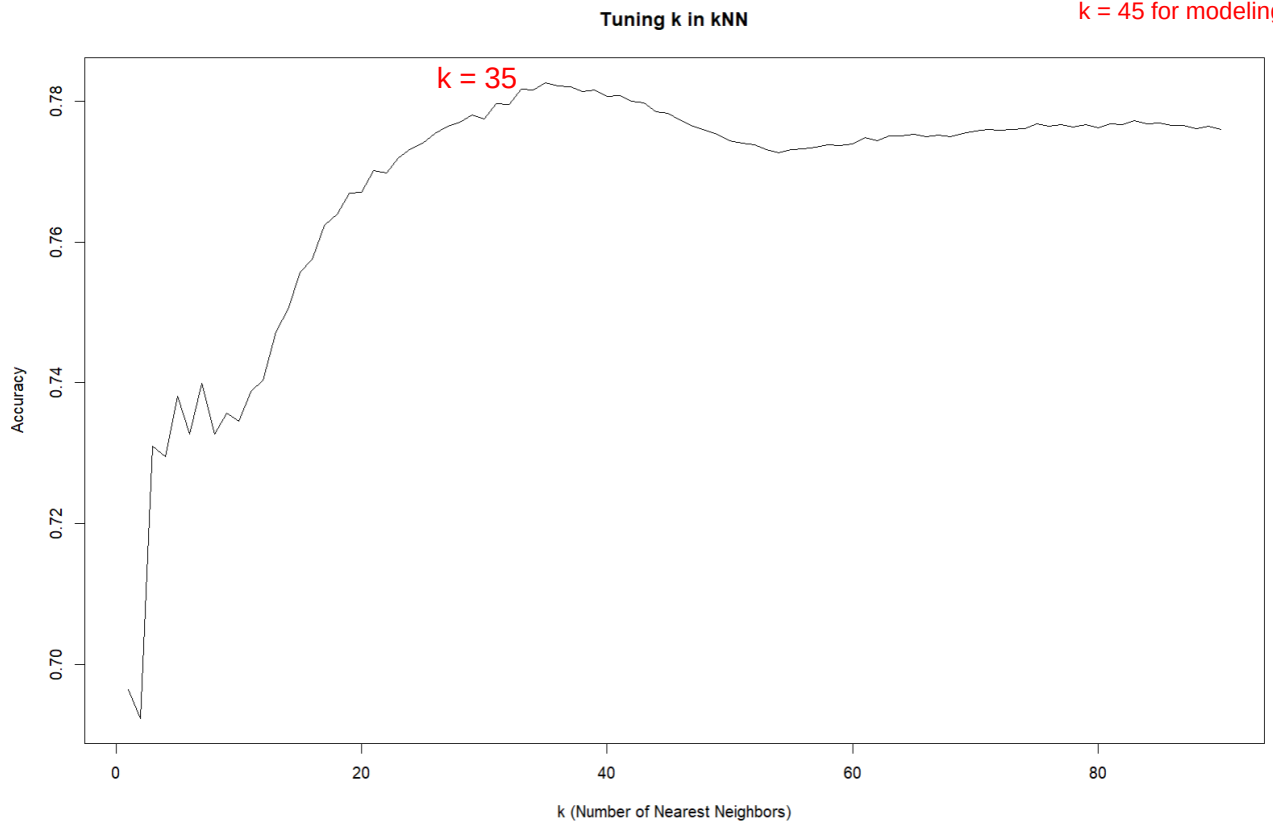
Note: Some variables have outliers. We will not treat them since they are proper values.

Data Overview – Attrition vs. Highest Correlations



kNN Hyperparameter Tuning – Best k Value

Note: k = 35 did not yield 60% Specificity. It achieved 58.5% maximum. Used k = 45 for modeling to achieve 60%.



Data Post Processing Overview

Final Dataset: First 5 Rows:

Attrition	Age	BusinessTravel	Department	EducationField	Gender	JobInvolvement	JobLevel	JobRole	JobSatisfaction	MaritalStatus	MonthlyIncome	OverTime	StockOptionLevel	TotalWorkingYears	YearsAtCompany	YearsInCurrentRole	YearsWithCurrManager
0	32	Travel_Rarely	Sales	Life Sciences	Male	3	2	Sales Executive	4	Divorced	4403	No	1	8	5	2	3
0	40	Travel_Rarely	Research & Development	Medical	Male	2	5	Research Director	3	Single	19626	No	0	21	20	7	9
0	35	Travel_Frequently	Research & Development	Life Sciences	Male	3	3	Manufacturing Director	4	Single	9362	No	0	10	2	2	2
0	32	Travel_Rarely	Sales	Marketing	Female	3	3	Sales Executive	4	Married	10422	No	2	14	14	10	7
0	24	Travel_Frequently	Research & Development	Technical Degree	Female	3	1	Research Scientist	4	Single	3760	Yes	0	6	6	3	3
0	27	Travel_Frequently	Research & Development	Life Sciences	Male	3	3	Manufacturing Director	1	Divorced	8793	No	2	9	9	7	7

- Removed non-informative columns: ID, EmployeeCount, EmployeeNumber, Over18 (only "Yes"), and StandardHours (always 80).
- Verified that no missing data was present across all columns.
- Retained all outliers as they were valid numerical values.
- Converted categorical variables into factors for analysis.
- Analyzed correlations between variables and Attrition.
- Confirmed that no missing data exists in any column.
- Dropped columns with little to no relationship to Attrition: DailyRate, DistanceFromHome, Education, EnvironmentSatisfaction, HourlyRate, MonthlyRate, NumCompaniesWorked, PercentSalaryHike, RelationshipSatisfaction, TrainingTimesLastYear, WorkLifeBalance, and YearsSinceLastPromotion.