

Data Collection and Preprocessing Phase

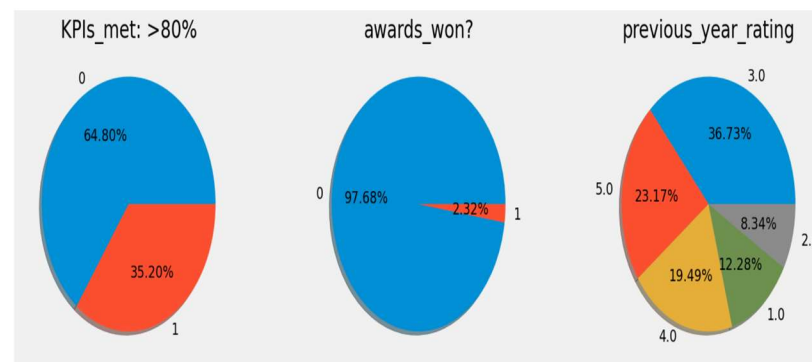
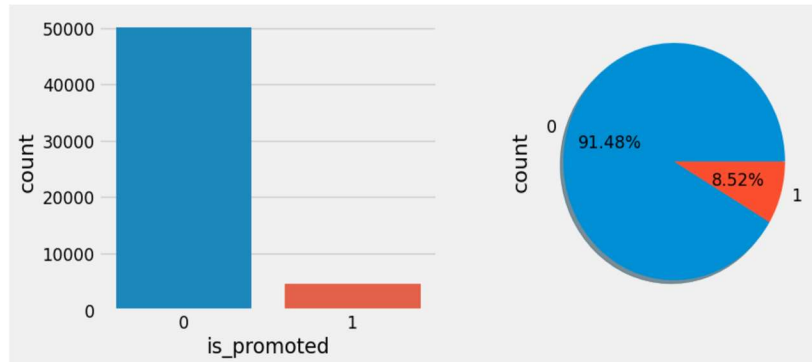
Date	15 July 2024
Team ID	-
Project Title	Human Resource Management: Predicting Employee Promotions Using Machine Learning
Maximum Marks	6 Marks

Data Exploration and Preprocessing Report

Identifies data sources, assesses quality issues like missing values and duplicates, and implements resolution plans to ensure accurate and reliable analysis.

Section	Description																																																																																																																																																
Data Overview	<div><div><u>Dimensions:</u> 54808 rows × 14 columns</div><div><u>Descriptive statistics:</u></div></div>																																																																																																																																																
	<table><tr><th></th><th>employee_id</th><th>department</th><th>region</th><th>education</th><th>gender</th><th>recruitment_channel</th><th>no_of_trainings</th><th>age</th><th>previous_year_rating</th><th>length_of_service</th><th>KPIs_met >80%</th></tr><tr><td>count</td><td>54808.000000</td><td>54808</td><td>54808</td><td>52399</td><td>54808</td><td>54808</td><td>54808.000000</td><td>54808.000000</td><td>50684.000000</td><td>54808.000000</td><td>54808.000000</td></tr><tr><td>unique</td><td>NaN</td><td>9</td><td>34</td><td>3</td><td>2</td><td>3</td><td>NaN</td><td>NaN</td><td>NaN</td><td>NaN</td><td>NaN</td></tr><tr><td>top</td><td>NaN</td><td>Sales & Marketing</td><td>region_2</td><td>Bachelor's</td><td>m</td><td>other</td><td>NaN</td><td>NaN</td><td>NaN</td><td>NaN</td><td>NaN</td></tr><tr><td>freq</td><td>NaN</td><td>16840</td><td>12343</td><td>36669</td><td>38496</td><td>30446</td><td>NaN</td><td>NaN</td><td>NaN</td><td>NaN</td><td>NaN</td></tr><tr><td>mean</td><td>39195.830627</td><td>NaN</td><td>NaN</td><td>NaN</td><td>NaN</td><td>NaN</td><td>1.253011</td><td>34.803915</td><td>3.329256</td><td>5.865512</td><td>0.351974</td></tr><tr><td>std</td><td>22586.581449</td><td>NaN</td><td>NaN</td><td>NaN</td><td>NaN</td><td>NaN</td><td>0.609264</td><td>7.660169</td><td>1.259993</td><td>4.265094</td><td>0.477590</td></tr><tr><td>min</td><td>1.000000</td><td>NaN</td><td>NaN</td><td>NaN</td><td>NaN</td><td>NaN</td><td>1.000000</td><td>20.000000</td><td>1.000000</td><td>1.000000</td><td>0.000000</td></tr><tr><td>25%</td><td>19669.750000</td><td>NaN</td><td>NaN</td><td>NaN</td><td>NaN</td><td>NaN</td><td>1.000000</td><td>29.000000</td><td>3.000000</td><td>3.000000</td><td>0.000000</td></tr><tr><td>50%</td><td>39225.500000</td><td>NaN</td><td>NaN</td><td>NaN</td><td>NaN</td><td>NaN</td><td>1.000000</td><td>33.000000</td><td>3.000000</td><td>5.000000</td><td>0.000000</td></tr><tr><td>75%</td><td>58730.500000</td><td>NaN</td><td>NaN</td><td>NaN</td><td>NaN</td><td>NaN</td><td>1.000000</td><td>39.000000</td><td>4.000000</td><td>7.000000</td><td>1.000000</td></tr><tr><td>max</td><td>78298.000000</td><td>NaN</td><td>NaN</td><td>NaN</td><td>NaN</td><td>NaN</td><td>10.000000</td><td>60.000000</td><td>5.000000</td><td>37.000000</td><td>1.000000</td></tr></table>		employee_id	department	region	education	gender	recruitment_channel	no_of_trainings	age	previous_year_rating	length_of_service	KPIs_met >80%	count	54808.000000	54808	54808	52399	54808	54808	54808.000000	54808.000000	50684.000000	54808.000000	54808.000000	unique	NaN	9	34	3	2	3	NaN	NaN	NaN	NaN	NaN	top	NaN	Sales & Marketing	region_2	Bachelor's	m	other	NaN	NaN	NaN	NaN	NaN	freq	NaN	16840	12343	36669	38496	30446	NaN	NaN	NaN	NaN	NaN	mean	39195.830627	NaN	NaN	NaN	NaN	NaN	1.253011	34.803915	3.329256	5.865512	0.351974	std	22586.581449	NaN	NaN	NaN	NaN	NaN	0.609264	7.660169	1.259993	4.265094	0.477590	min	1.000000	NaN	NaN	NaN	NaN	NaN	1.000000	20.000000	1.000000	1.000000	0.000000	25%	19669.750000	NaN	NaN	NaN	NaN	NaN	1.000000	29.000000	3.000000	3.000000	0.000000	50%	39225.500000	NaN	NaN	NaN	NaN	NaN	1.000000	33.000000	3.000000	5.000000	0.000000	75%	58730.500000	NaN	NaN	NaN	NaN	NaN	1.000000	39.000000	4.000000	7.000000	1.000000	max	78298.000000	NaN	NaN	NaN	NaN	NaN	10.000000	60.000000	5.000000	37.000000	1.000000
		employee_id	department	region	education	gender	recruitment_channel	no_of_trainings	age	previous_year_rating	length_of_service	KPIs_met >80%																																																																																																																																					
	count	54808.000000	54808	54808	52399	54808	54808	54808.000000	54808.000000	50684.000000	54808.000000	54808.000000																																																																																																																																					
	unique	NaN	9	34	3	2	3	NaN	NaN	NaN	NaN	NaN																																																																																																																																					
	top	NaN	Sales & Marketing	region_2	Bachelor's	m	other	NaN	NaN	NaN	NaN	NaN																																																																																																																																					
	freq	NaN	16840	12343	36669	38496	30446	NaN	NaN	NaN	NaN	NaN																																																																																																																																					
	mean	39195.830627	NaN	NaN	NaN	NaN	NaN	1.253011	34.803915	3.329256	5.865512	0.351974																																																																																																																																					
	std	22586.581449	NaN	NaN	NaN	NaN	NaN	0.609264	7.660169	1.259993	4.265094	0.477590																																																																																																																																					
	min	1.000000	NaN	NaN	NaN	NaN	NaN	1.000000	20.000000	1.000000	1.000000	0.000000																																																																																																																																					
25%	19669.750000	NaN	NaN	NaN	NaN	NaN	1.000000	29.000000	3.000000	3.000000	0.000000																																																																																																																																						
50%	39225.500000	NaN	NaN	NaN	NaN	NaN	1.000000	33.000000	3.000000	5.000000	0.000000																																																																																																																																						
75%	58730.500000	NaN	NaN	NaN	NaN	NaN	1.000000	39.000000	4.000000	7.000000	1.000000																																																																																																																																						
max	78298.000000	NaN	NaN	NaN	NaN	NaN	10.000000	60.000000	5.000000	37.000000	1.000000																																																																																																																																						

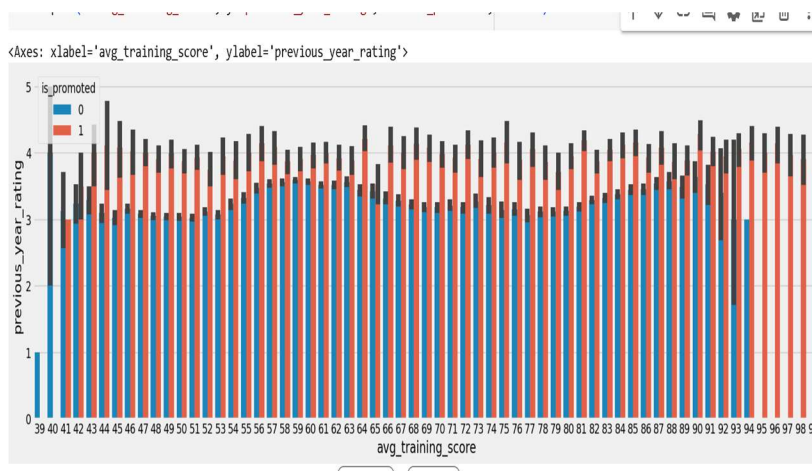
Univariate Analysis



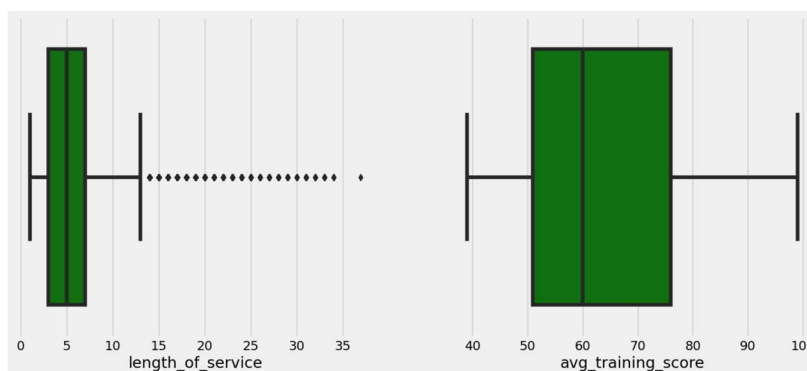
Bivariate Analysis

-

Multivariate Analysis



Outliers and Anomalies



Data Preprocessing Code Screenshots

Loading Data

```
# Reading the csv and printing its shape
df = pd.read_csv('../Dataset/emp_promotion.csv')
print('shape of train data {}'.format(df.shape))
shape of train data (54988, 14)

df.head(5)
```

	employee_id	department	region	education	gender	recruitment_channel	no_of_trainings	age	previous_year_rating	length_of_service	KPIs_met >80%	awards_won?
0	65438	Sales & Marketing	region_7	Master's & above	f	sourcing	1	35	5.0	8	1	0
1	65141	Operations	region_22	Bachelor's	m	other	1	30	5.0	4	0	0
2	7513	Sales & Marketing	region_19	Bachelor's	m	sourcing	1	34	3.0	7	0	0
3	2542	Sales & Marketing	region_23	Bachelor's	m	other	2	39	1.0	10	0	0
4	48945	Technology	region_26	Bachelor's	m	other	1	45	3.0	2	0	0

Handling Missing Data

```
# Replacing nan with mode
print(df['education'].value_counts())
df['education']=df['education'].fillna(df['education'].mode()[0])

# Replacing nan with mode
print(df['previous_year_rating'].value_counts())
df['previous_year_rating']=df['previous_year_rating'].fillna(df['previous_year_rating'].mode()[0])
```

Data Transformation

```
# Feature mapping is done on education column

df['education']=df['education'].replace(("Below Secondary", "Bachelor's", "Master's & above"),(1,2,3))

lb = LabelEncoder()
df['department']=lb.fit_transform(df['department'])
```

Feature Engineering

-

Save Processed Data	-
---------------------	---