

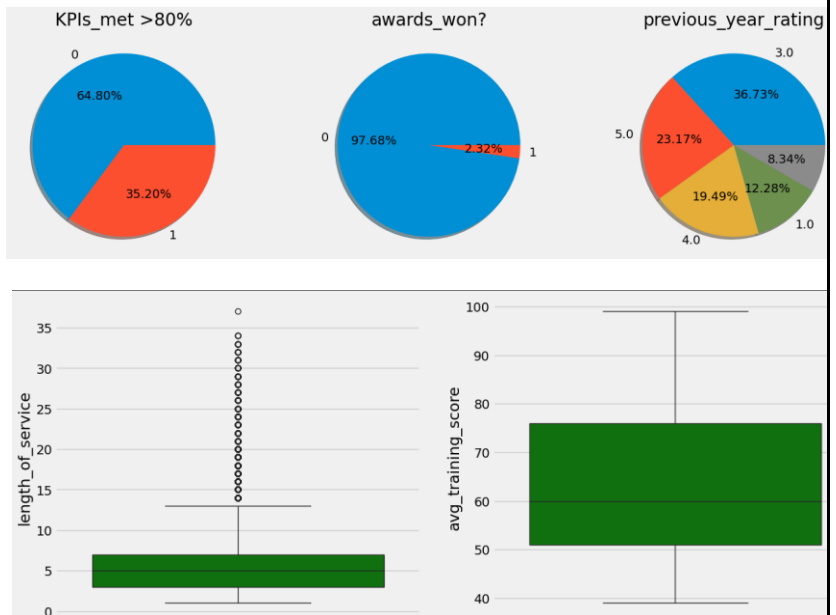
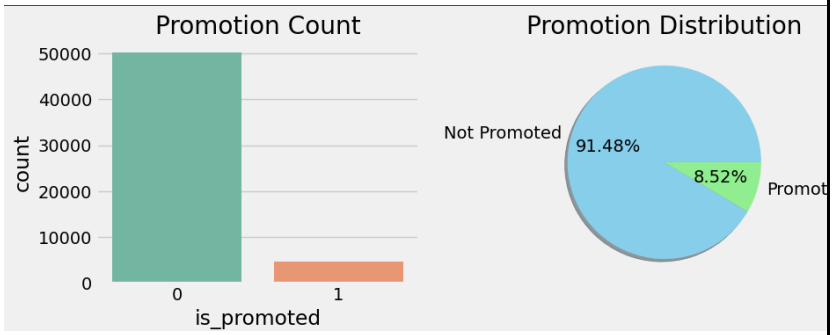
## Data Collection and Preprocessing Phase

Date	25 June 2025
Team ID	SWTID1750155746
Project Title	Human Resource Management: Predicting Employee Promotions using Machine Learning
Maximum Marks	6 Marks

### Data Exploration and Preprocessing Report

Dataset variables will be statistically analyzed to identify patterns and outliers, with Python employed for preprocessing tasks like normalization and feature engineering. Data cleaning will address missing values and outliers, ensuring quality for subsequent analysis and modeling, and forming a strong foundation for insights and predictions.

Section	Description																																																																																																																																																																								
Data Overview	<div><div><div><div>Dimension:</div><div>54808 rows × 14 columns</div></div><div><div>Descriptive statistics:</div><table><thead><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>staff_cost_100k</th><th>awards_won?</th><th>avg_training_score</th></tr></thead><tbody><tr><td>count</td><td>54808.000000</td><td>54808</td><td>54808</td><td>52308</td><td>54808</td><td>54808</td><td>54808.000000</td><td>54808.000000</td><td>50804.000000</td><td>54808.000000</td><td>54808.000000</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><td>NaN</td><td>NaN</td></tr><tr><td>top</td><td>NaN</td><td>Sales &amp; 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><td>NaN</td><td>NaN</td></tr><tr><td>freq</td><td>NaN</td><td>10840</td><td>12343</td><td>30668</td><td>30490</td><td>30446</td><td>NaN</td><td>NaN</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>3.326256</td><td>5.865512</td><td>0.351974</td><td>0.023172</td><td>63.386750</td></tr><tr><td>std</td><td>22506.581449</td><td>NaN</td><td>NaN</td><td>NaN</td><td>NaN</td><td>NaN</td><td>0.000264</td><td>7.660169</td><td>1.256893</td><td>4.265094</td><td>0.477590</td><td>0.150450</td><td>13.371558</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><td>0.000000</td><td>39.000000</td></tr><tr><td>25%</td><td>19659.750000</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><td>0.000000</td><td>51.000000</td></tr><tr><td>66%</td><td>36225.500000</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><td>0.000000</td><td>60.000000</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><td>0.000000</td><td>76.000000</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><td>1.000000</td><td>99.000000</td></tr></tbody></table></div></div></div>		employee_id	department	region	education	gender	recruitment_channel	no_of_trainings	age	previous_year_rating	length_of_service	staff_cost_100k	awards_won?	avg_training_score	count	54808.000000	54808	54808	52308	54808	54808	54808.000000	54808.000000	50804.000000	54808.000000	54808.000000	54808.000000	54808.000000	unique	NaN	9	34	3	2	3	NaN	NaN	NaN	NaN	NaN	NaN	NaN	top	NaN	Sales & Marketing	region_2	Bachelor's	m	other	NaN	NaN	NaN	NaN	NaN	NaN	NaN	freq	NaN	10840	12343	30668	30490	30446	NaN	NaN	NaN	NaN	NaN	NaN	NaN	mean	39195.830627	NaN	NaN	NaN	NaN	NaN	1.253011	34.803915	3.326256	5.865512	0.351974	0.023172	63.386750	std	22506.581449	NaN	NaN	NaN	NaN	NaN	0.000264	7.660169	1.256893	4.265094	0.477590	0.150450	13.371558	min	1.000000	NaN	NaN	NaN	NaN	NaN	1.000000	20.000000	1.000000	1.000000	0.000000	0.000000	39.000000	25%	19659.750000	NaN	NaN	NaN	NaN	NaN	1.000000	29.000000	3.000000	3.000000	0.000000	0.000000	51.000000	66%	36225.500000	NaN	NaN	NaN	NaN	NaN	1.000000	33.000000	3.000000	5.000000	0.000000	0.000000	60.000000	75%	58730.500000	NaN	NaN	NaN	NaN	NaN	1.000000	39.000000	4.000000	7.000000	1.000000	0.000000	76.000000	max	78298.000000	NaN	NaN	NaN	NaN	NaN	10.000000	60.000000	5.000000	37.000000	1.000000	1.000000	99.000000
		employee_id	department	region	education	gender	recruitment_channel	no_of_trainings	age	previous_year_rating	length_of_service	staff_cost_100k	awards_won?	avg_training_score																																																																																																																																																											
count	54808.000000	54808	54808	52308	54808	54808	54808.000000	54808.000000	50804.000000	54808.000000	54808.000000	54808.000000	54808.000000																																																																																																																																																												
unique	NaN	9	34	3	2	3	NaN	NaN	NaN	NaN	NaN	NaN	NaN																																																																																																																																																												
top	NaN	Sales & Marketing	region_2	Bachelor's	m	other	NaN	NaN	NaN	NaN	NaN	NaN	NaN																																																																																																																																																												
freq	NaN	10840	12343	30668	30490	30446	NaN	NaN	NaN	NaN	NaN	NaN	NaN																																																																																																																																																												
mean	39195.830627	NaN	NaN	NaN	NaN	NaN	1.253011	34.803915	3.326256	5.865512	0.351974	0.023172	63.386750																																																																																																																																																												
std	22506.581449	NaN	NaN	NaN	NaN	NaN	0.000264	7.660169	1.256893	4.265094	0.477590	0.150450	13.371558																																																																																																																																																												
min	1.000000	NaN	NaN	NaN	NaN	NaN	1.000000	20.000000	1.000000	1.000000	0.000000	0.000000	39.000000																																																																																																																																																												
25%	19659.750000	NaN	NaN	NaN	NaN	NaN	1.000000	29.000000	3.000000	3.000000	0.000000	0.000000	51.000000																																																																																																																																																												
66%	36225.500000	NaN	NaN	NaN	NaN	NaN	1.000000	33.000000	3.000000	5.000000	0.000000	0.000000	60.000000																																																																																																																																																												
75%	58730.500000	NaN	NaN	NaN	NaN	NaN	1.000000	39.000000	4.000000	7.000000	1.000000	0.000000	76.000000																																																																																																																																																												
max	78298.000000	NaN	NaN	NaN	NaN	NaN	10.000000	60.000000	5.000000	37.000000	1.000000	1.000000	99.000000																																																																																																																																																												
Univariate Analysis																																																																																																																																																																									



## Multivariate Analysis



Outliers and Anomalies	-																																																																		
Data Preprocessing Code Screenshots																																																																			
Loading Data	<div><pre>df = pd.read_csv('emp_promotion.csv') print('Shape of train data {}'.format(df.shape))</pre><div>Shape of train data (54808, 14)</div></div> <div><pre>df.head()</pre><table><thead><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></tr></thead><tbody><tr><td>0</td><td>65438</td><td>Sales &amp; Marketing</td><td>region_7</td><td>Master's &amp; above</td><td>f</td><td>sourcing</td><td>1</td><td>35</td><td>5.0</td><td>8</td></tr><tr><td>1</td><td>65141</td><td>Operations</td><td>region_22</td><td>Bachelor's</td><td>m</td><td>other</td><td>1</td><td>30</td><td>5.0</td><td>4</td></tr><tr><td>2</td><td>7513</td><td>Sales &amp; Marketing</td><td>region_19</td><td>Bachelor's</td><td>m</td><td>sourcing</td><td>1</td><td>34</td><td>3.0</td><td>7</td></tr><tr><td>3</td><td>2542</td><td>Sales &amp; Marketing</td><td>region_23</td><td>Bachelor's</td><td>m</td><td>other</td><td>2</td><td>39</td><td>1.0</td><td>10</td></tr><tr><td>4</td><td>48945</td><td>Technology</td><td>region_26</td><td>Bachelor's</td><td>m</td><td>other</td><td>1</td><td>45</td><td>3.0</td><td>2</td></tr></tbody></table></div>		employee_id	department	region	education	gender	recruitment_channel	no_of_trainings	age	previous_year_rating	length_of_service	0	65438	Sales & Marketing	region_7	Master's & above	f	sourcing	1	35	5.0	8	1	65141	Operations	region_22	Bachelor's	m	other	1	30	5.0	4	2	7513	Sales & Marketing	region_19	Bachelor's	m	sourcing	1	34	3.0	7	3	2542	Sales & Marketing	region_23	Bachelor's	m	other	2	39	1.0	10	4	48945	Technology	region_26	Bachelor's	m	other	1	45	3.0	2
	employee_id	department	region	education	gender	recruitment_channel	no_of_trainings	age	previous_year_rating	length_of_service																																																									
0	65438	Sales & Marketing	region_7	Master's & above	f	sourcing	1	35	5.0	8																																																									
1	65141	Operations	region_22	Bachelor's	m	other	1	30	5.0	4																																																									
2	7513	Sales & Marketing	region_19	Bachelor's	m	sourcing	1	34	3.0	7																																																									
3	2542	Sales & Marketing	region_23	Bachelor's	m	other	2	39	1.0	10																																																									
4	48945	Technology	region_26	Bachelor's	m	other	1	45	3.0	2																																																									
Removing Unwanted Features	<div><pre>#Dropping unwanted features """ To predict the promotion, employee id is not required and even sex feature is also not important. For promotion region and recruitment channel is not important. So, removing employee id, sex, recruitment_channel and region"""  df = df.drop(['employee_id','gender','region','recruitment_channel'],axis=1)</pre></div>																																																																		
Handling Missing Data	<div><pre>#Replacing nan with mode print(df['education'].value_counts()) df['education'] = df['education'].fillna(df['education'].mode()[0])</pre></div> <div><pre>#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])</pre></div>																																																																		

<p>Removing Negative Data</p>	<pre>#Removing Negative Data #Finding the employee who got promoted even in poor performance. It affects the model performance.  negative = df[(df['KPIs_met &gt;80%']==0) &amp; (df['awards_won?']==0) &amp; (df['previous_year_rating']==1.0) &amp;               (df['is_promoted']==1) &amp; (df['avg_training_score']&lt;60)]  negative</pre> <pre>#Removing Negative data df.drop(index=[31860,51374],inplace=True)</pre>
<p>Handling Outliers</p>	<pre>#Handling Outliers q1 = np.quantile(df['length_of_service'],0.25) q3 = np.quantile(df['length_of_service'],0.75)  IQR = q3-q1  upperBound = (1.5*IQR)+q3 lowerBound = (1.5*IQR)-q1  print('q1 : ',q1) print('q3 : ',q3) print('IQR : ',IQR) print('Upper Bound : ',upperBound) print('Lower Bound : ',lowerBound) print('Skewed data : ',len(df[df['length_of_service']&gt;upperBound]))</pre>
<p>Handling Imbalanced Data</p>	<pre>#SMOTE for Imbalanced Data #Splitting data and resampling it  x = df.drop('is_promoted',axis=1) y = df['is_promoted'] print(x.shape) print(y.shape)  (54806, 9) (54806,)  from imblearn.over_sampling import SMOTE sm = SMOTE() x_resample, y_resample = sm.fit_resample(x,y)</pre>

Data Transformation	<pre>#Handling Categorical Values  #Feature mapping is done on education column  df['education'] = df['education'].replace(("Below Secondary","Bachelor's","Master's &amp; above"),(1,2,3))  lb = LabelEncoder() df['department'] = lb.fit_transform(df['department'])</pre>
Feature Engineering	Attached the codes in final submission.
Save Processed Data	-