Data Loading & Cleaning and Preprocessing import pandas as pd In [1]: import numpy as np df = pd.read\_csv(r"C:\Users\rizwa\OneDrive\Desktop\Intern\Final project\HR-Employee-Attrition.csv") print("Data Information:") print(df.info()) # Check for missing values print("\nMissing Values:") print(df.isnull().sum()) # Handle missing values (if any) df = df.dropna() # Display basic statistics after handling missing values print("\nSummary Statistics:") print(df.describe()) # Check for duplicates print("\nDuplicate Rows:") print(df.duplicated().sum()) # Handle duplicates (if any) # Example: Drop duplicate rows df = df.drop\_duplicates() # Confirm changes print("\nData Information After Cleaning:") print(df.info()) Data Information: <class 'pandas.core.frame.DataFrame'> RangeIndex: 1470 entries, 0 to 1469 Data columns (total 35 columns): Column # Non-Null Count Dtype ----------- - -1470 non-null int64 0 Age 1 Attrition 1470 non-null object BusinessTravel 1470 non-null object DailyRate 1470 non-null 3 int64 Department 1470 non-null object DistanceFromHome 1470 non-null int64 Education 1470 non-null EducationField 1470 non-null object EmployeeCount 1470 non-null 8 int64 1470 non-null 9 EmployeeNumber int64 10 EnvironmentSatisfaction 1470 non-null int64 11 Gender 1470 non-null object 1470 non-null 12 HourlyRate int64 1470 non-null 13 JobInvolvement int64 JobLevel 1470 non-null int64 14 JobRole 1470 non-null 15 object JobSatisfaction 1470 non-null int64 16 17 MaritalStatus 1470 non-null object MonthlyIncome 1470 non-null 18 int64 1470 non-null MonthlyRate 19 int64 20 NumCompaniesWorked 1470 non-null int64 21 0ver18 1470 non-null object 22 OverTime 1470 non-null object 1470 non-null 23 PercentSalaryHike 24 PerformanceRating 1470 non-null int64 25 RelationshipSatisfaction 1470 non-null int64 1470 non-null 26 StandardHours int64 27 StockOptionLevel 1470 non-null int64 28 TotalWorkingYears 1470 non-null int64 29 TrainingTimesLastYear 1470 non-null int64 WorkLifeBalance 1470 non-null int64 1470 non-null int64 31 YearsAtCompany YearsInCurrentRole 1470 non-null 1470 non-null 33 YearsSinceLastPromotion 34 YearsWithCurrManager 1470 non-null dtypes: int64(26), object(9) memory usage: 402.1+ KB Missing Values: 0 Age Attrition 0 BusinessTravel 0 DailyRate 0 Department 0  ${\tt DistanceFromHome}$ 0 Education 0 EducationField 0 EmployeeCount 0 EmployeeNumber 0 EnvironmentSatisfaction 0 Gender 0 HourlyRate 0 JobInvolvement JobLevel JobRole JobSatisfaction 0 MaritalStatus 0 MonthlyIncome 0 MonthlyRate 0 NumCompaniesWorked 0 0ver18 0 **OverTime** 0 PercentSalaryHike 0 PerformanceRating 0 RelationshipSatisfaction 0 StandardHours 0 StockOptionLevel TotalWorkingYears 0 TrainingTimesLastYear 0 WorkLifeBalance YearsAtCompany 0 YearsInCurrentRole 0 YearsSinceLastPromotion YearsWithCurrManager dtype: int64 **Summary Statistics:** Age DailyRate DistanceFromHome Education EmployeeCount \ 1470.000000 1470.000000 1470.000000 1470.000000 1470.0 count 36.923810 802.485714 9.192517 2.912925 1.0 mean 9.135373 std 403.509100 8.106864 1.024165 0.0 18.000000 102.000000 1.000000 1.000000 1.0 min 25% 30.000000 465.000000 2.000000 2.000000 1.0 50% 36.000000 802.000000 7.000000 3.000000 1.0 43.000000 14.000000 75% 1157.000000 4.000000 1.0 max 60.000000 1499.000000 29.000000 5.000000 1.0 EmployeeNumber EnvironmentSatisfaction HourlyRate JobInvolvement \ 1470.000000 1470.000000 1470.000000 1470.000000 count mean 1024.865306 2.721769 65.891156 2.729932 602.024335 1.093082 20.329428 0.711561 std 1.000000 1.000000 30.000000 1.000000 min 2.000000 25% 491.250000 2.000000 48.000000 50% 1020.500000 3.000000 66.000000 3.000000 75% 1555.750000 4.000000 83.750000 3.000000 2068.000000 4.000000 100.000000 4.000000 max JobLevel ... RelationshipSatisfaction StandardHours \ 1470.0 count 1470.000000 ... 1470.000000 80.0 mean 2.063946 ... 2.712245 1.106940 ... std 1.081209 0.0 min 1.000000 1.000000 0.08 25% 1.000000 2.000000 80.0 2.000000 ... 50% 3.000000 80.0 75% 4.000000 3.000000 80.0 max 5.000000 4.000000 TotalWorkingYears TrainingTimesLastYear \ StockOptionLevel 1470.000000 count 1470.000000 1470.000000 0.793878 11.279592 2.799320 mean 1.289271 0.852077 7.780782 std 0.000000 0.000000 0.000000 min 2.000000 25% 0.000000 6.000000 50% 3.000000 1.000000 10.000000 75% 15.000000 3.000000 1.000000 3.000000 6.000000 max 40.000000 WorkLifeBalance YearsAtCompany YearsInCurrentRole \ 1470.000000 1470.000000 count 1470.000000 2.761224 7.008163 4.229252 mean 0.706476 6.126525 3.623137 std min 1.000000 0.00000 0.000000 25% 2.000000 3.000000 2.000000 50% 3.000000 5.000000 3.000000 75% 3.000000 9.000000 7.000000 max 4.000000 40.000000 18.000000 YearsSinceLastPromotion YearsWithCurrManager 1470.000000 count 1470.000000 2.187755 4.123129 mean 3.222430 3.568136 std min 0.000000 0.000000 25% 0.000000 2.000000 50% 1.000000 3.000000 75% 3.000000 7.000000 15.000000 17.000000 max [8 rows x 26 columns] Duplicate Rows: Data Information After Cleaning: <class 'pandas.core.frame.DataFrame'> Int64Index: 1470 entries, 0 to 1469 Data columns (total 35 columns): # Column Non-Null Count Dtype 0 Age 1470 non-null int64 Attrition 1470 non-null object 1 2 BusinessTravel 1470 non-null object 3 DailyRate 1470 non-null int64 4 Department 1470 non-null object 5 DistanceFromHome 1470 non-null int64 6 Education 1470 non-null int64 1470 non-null EducationField object 8 EmployeeCount 1470 non-null int64 9 EmployeeNumber 1470 non-null int64 10 EnvironmentSatisfaction 1470 non-null int64 11 Gender 1470 non-null object HourlyRate 12 1470 non-null int64 13 JobInvolvement 1470 non-null int64 14 JobLevel 1470 non-null int64 15 JobRole 1470 non-null object JobSatisfaction 1470 non-null MaritalStatus 1470 non-null object MonthlyIncome 1470 non-null MonthlyRate 19 1470 non-null int64 20 NumCompaniesWorked 1470 non-null int64 21 0ver18 1470 non-null object 22 **OverTime** 1470 non-null object 23 PercentSalaryHike 1470 non-null int64 24 PerformanceRating 1470 non-null int64 25 RelationshipSatisfaction 1470 non-null int64 26 1470 non-null StandardHours int64 27 int64 StockOptionLevel 1470 non-null 28 TotalWorkingYears 1470 non-null int64 29 TrainingTimesLastYear 1470 non-null int64 30 WorkLifeBalance 1470 non-null int64 1470 non-null 31 YearsAtCompany int64 YearsInCurrentRole 1470 non-null 33 YearsSinceLastPromotion 1470 non-null int64 34 YearsWithCurrManager 1470 non-null int64 dtypes: int64(26), object(9) memory usage: 413.4+ KB None Data Analysis and Modeling (Analyzed the cleaned HR data to identify patterns, trends, and correlations, EDA: Exploratory Data Analysis) import pandas as pd In [11]: import matplotlib.pyplot as plt import seaborn as sns # Visualize the distribution of Age plt.figure(figsize=(10, 6)) sns.histplot(df['Age'], bins=30, kde=True) plt.title('Distribution of Age') plt.show() Distribution of Age 140 120 100 Count 60 40 20 20 40 Age # Visualize the attrition count plt.figure(figsize=(8, 5)) sns.countplot(x='Attrition', data=df) plt.title('Attrition Count') plt.show() Attrition Count 1200 1000 800 count 600 400 200 Yes No Attrition In [13]: # Explore the correlation matrix correlation\_matrix = df.corr() plt.figure(figsize=(12, 10)) sns.heatmap(correlation\_matrix, annot=True, cmap='coolwarm', linewidths=0.5) plt.title('Correlation Matrix') plt.show() C:\Users\rizwa\AppData\Local\Temp\ipykernel\_1516\4129031465.py:2: FutureWarning: The default value of numeric\_only in DataFrame.corr is deprecated. In a future version, it will def ault to False. Select only valid columns or specify the value of numeric\_only to silence this warning. correlation\_matrix = df.corr() Correlation Matrix 1.0 03<mark>0.68</mark>0.0<del>2</del>0.02<mark>0.310.210.22</mark> 0.2 .010.01).0240.03<mark>0.51</mark>.004<mark>0.5</mark>0.028**0.3**.0066011905 030.0010.21 DailyRate -.011 .0501.0108.0201.0406.0001.0301.00707.0302.0308.002.3000040707 0442.0105003050348.0184009090348.02 DistanceFromHome -.001070051 0405004160307.02017009150190.010.01 036.0105030100688066300-\$070107.02-\$0.02 \$0.044.0207006 Education -0.210.0107021 0410.02017.0101.0420.1-0.0101.0916.021 01**(0.15**().02/5009080690.06().05(4.06 EmployeeCount -- 0.8 EmployeeNumber - 0.0-0.0501030.04 06pl.01040240.03bl.0-101004840-009009 .01¢x030500•¢0901¢9x04¢5x01050103k00•†0301-†30.02¢0.07 1 -0.95.00089900209088093038.01-0.03-20.008007 EnvironmentSatisfaction -0.010.018.016.02 00-01-400-2070109:0208001050108.01-16.00 .0-61.0902.30908.500406.0-20.02(4.02-10.02 HourlyRate - .024.026.030.01 0 90**8 9 0 19** 30 4 3 **1 - 0** . 0 1 d2 . 0 2 **0** . 0 1 d5 . 0 1 d5 0 1 <del>0</del> 5 . 0 1 d7 . 0 2 d2 0 3 0 2002 0 0 450 50 145 . 0 145 . 0 20 10 0 0 807 0 20 14 0 2 JobInvolvement -0.030.04050080804 - 0.6 .0009004020208.01<mark>3 1-0</mark>.001<mark>0995</mark>0.040.140.0305.0201.02 JobLevel -0.510.000800530.1 01-0.780.01@03-0.530.390.350.38 0-40500-4080741.0-201001<mark>91-0</mark>.0 **0**-70200-405 60.02 00-20801 01+0.602.00-5080-1090603.800-2030148.02 JobSatisfaction -.0009030.00-0701 0-D500-63016.01 5.96.00721 0.03:0.150.0207.0107.02 .020203 0.510.360.340.34 MonthlyIncome - 0.5 0070010.09 MonthlyRate -. 028.0302020.02 .01.03.036.01.05.01.0.004.0006.4035 1 0.0308.0906.40909.800 .034.0205001050048.024.00030034603 NumCompaniesWorked - 0.30.038.02 9.1 00 tb 30 1 ck 0 2 ck 0 1 tb . 1 4 tb . 0 5 tb . 1 5 tb . 0 1 8 1 - 0 . 0 - 1 tb . 0 1 ck 0 5 .030.240.060600894.1-20.0901.03-10.1 - 0.4 PercentSalaryHike -003 602 30.040.01 .0163.0-35200-90101-67.03|5.020.0-207006|41.01 007050-20100-05/200-3030-30600-10502 02:01 00350607.0 p60026003940365.0 1ds 02 00019004.102-10.02 0.02/0.601,00-2020209.02010032080-1070099801 PerformanceRating RelationshipSatisfaction -.05040008066009 .cox10007170010303@4.02@1.01@20201500@110530.0~@1.031 0.405.020400215.020.01-0.0105.0015000 StandardHours -060200344.050.0202.014.0101.005040340.001.000150030504 StockOptionLevel -.03&040.046.01 0.010.0101004010161.0501.0141.02 0.2 TotalWorkingYears -0.68.0150040.15 0.02**0.77**0.02 **0.24**0.0 1200 60702 .0-10|40|902|.70|902|.30|05|05|78 1-0.03600 0.630.46 0.4 0.46 TrainingTimesLastYear -0.002003250307.02 .02**4**.0-10900-18501.15.0-10800-1580121200-1150-16600-15201013002 0040100**0.**02**8 1 0**.0120.0**5** 00**8**9002 0.01d).02d800-44601d503-4b.01d803dL0ed80ed8400d8602d6.02 WorkLifeBalance -0.020.03000000 .000001<u>6</u>.020.02<mark>0.58</mark>.003<mark>0.51</mark>0.02<del>4</del>0.120.0880030401 YearsAtCompany -0.310.084009506 01,0.69,003,6012 1 - 0.0 0084018.024008<mark>0.34</mark>.002<mark>0.36</mark>).0148.049100115036.01 1 0.550.71 05 0.46.0050050.7 YearsInCurrentRole -0.20,00990190.06 YearsSinceLastPromotion -0.22 YearsWithCurrManager - 0.20.026014.00 JobLevel TotalWorkingYears WorkLifeBalance YearsInCurrentRole DailyRate Jobinvolvement JobSatisfaction StandardHours StockOptionLevel YearsSinceLastPromotion YearsWithCurrManager DistanceFromHome Education EmployeeCount EnvironmentSatisfaction HourlyRate NumCompaniesWorked TrainingTimesLastYear YearsAtCompany EmployeeNumber MonthlyIncome MonthlyRate RelationshipSatisfaction PercentSalaryHike PerformanceRating In [15]: # Explore relationships between specific variables plt.figure(figsize=(12, 6)) sns.boxplot(x='JobRole', y='MonthlyIncome', data=df) plt.title('Monthly Income by Job Role') plt.xticks(rotation=45, ha='right') plt.show() Monthly Income by Job Role 20000 17500 15000 12500 MonthlyIncome 10000 7500 5000 2500 JobRole ### Trends in Salary Over Time ### plt.figure(figsize=(12, 6)) sns.lineplot(data=df, x='YearsAtCompany', y='MonthlyIncome', hue='Attrition') plt.title('Trends in Salary Over Time') plt.xlabel('Years at Company') plt.ylabel('Monthly Income') plt.show() Trends in Salary Over Time 20000 Attrition Yes No 17500 15000 Monthly Income 12500 10000 7500 5000 2500 10 35 0 5 15 20 25 30 40 Years at Company ### Patterns in Performance Ratings ### plt.figure(figsize=(10, 6)) sns.countplot(data=df, x='PerformanceRating', hue='Attrition') plt.title('Patterns in Performance Ratings') plt.xlabel('Performance Rating') plt.ylabel('Count') plt.show() Patterns in Performance Ratings Attrition 1000 Yes No 800 600 400 200 Performance Rating In [9]: ### Trends in Turnover Over Time ### plt.figure(figsize=(12, 6)) sns.lineplot(data=df, x='YearsAtCompany', y='Attrition', estimator='mean', ci=None) plt.title('Trends in Turnover Over Time') plt.xlabel('Years at Company') plt.ylabel('Attrition Rate') plt.show() C:\Users\rizwa\AppData\Local\Temp\ipykernel\_1516\3951006875.py:3: FutureWarning: The `ci` parameter is deprecated. Use `errorbar=None` for the same effect. sns.lineplot(data=df, x='YearsAtCompany', y='Attrition', estimator='mean', ci=None) Trends in Turnover Over Time Yes Attrition Rate No 0 10 15 20 25 30 35 40 Years at Company In [10]: ### Correlation Between Job Satisfaction and Turnover ### plt.figure(figsize=(10, 6)) sns.boxplot(data=df, x='Attrition', y='JobSatisfaction') plt.title('Correlation Between Job Satisfaction and Turnover') plt.xlabel('Attrition') plt.ylabel('Job Satisfaction') plt.show() Correlation Between Job Satisfaction and Turnover 4.0 3.5 3.0 2.5 2.0 1.5 1.0 Yes No Attrition Strategy Development and Documentation In [23]: **class** Employee: def \_\_init\_\_(self, name, position, skills, performance\_rating): self.name = name self.position = position self.skills = skills self.performance\_rating = performance\_rating self.training\_completed = False def set\_performance\_rating(self, rating): self.performance\_rating = rating def complete\_training(self): self.training\_completed = True print(f"{self.name} has completed targeted training.") def adjust\_compensation(self, new\_salary): print(f"Compensation for {self.name} adjusted to {new\_salary}.") def promote(self, new\_position): print(f"{self.name} has been promoted to {new\_position}.") self.position = new\_position class PerformanceEnhancementStrategy: def \_\_init\_\_(self): self.employees = [] def add\_employee(self, employee): self.employees.append(employee) def implement\_strategy(self): for employee in self.employees: if employee.performance\_rating < 4.0:</pre> # Offer targeted training for employees with low performance employee.complete\_training() elif 4.0 <= employee.performance\_rating < 4.5:</pre> # Consider compensation adjustment for employees with moderate performance employee.adjust\_compensation(1.05) elif employee.performance\_rating >= 4.5: # Consider career progression for high-performing employees employee.promote("Senior " + employee.position) employee1 = Employee("Ekta Agrawal", "Software Developer", ["Python", "JavaScript"], 3.8)
employee2 = Employee("Harshvardhan", "Data Analyst", ["SQL", "Excel"], 4.2) employee3 = Employee("Rizwan Siddiqui", "UX Designer", ["UI/UX", "Sketch"], 4.8) strategy = PerformanceEnhancementStrategy() strategy.add\_employee(employee1) strategy.add\_employee(employee2) strategy.add\_employee(employee3) # Implement the performance enhancement strategy strategy.implement\_strategy() # Print the results for employee in strategy.employees: print(f"{employee.name} - Position: {employee.position}, Performance Rating: {employee.performance\_rating}") Ekta Agrawal has completed targeted training. Compensation for Harshvardhan adjusted to 1.05. Rizwan Siddiqui has been promoted to Senior UX Designer. Ekta Agrawal - Position: Software Developer, Performance Rating: 3.8 Harshvardhan - Position: Data Analyst, Performance Rating: 4.2 Rizwan Siddiqui - Position: Senior UX Designer, Performance Rating: 4.8