## **Phase-3 Submission – Data Analytics**

**Student Name:** A. Andro Jerslin Jebina   
 **Register Number:** 212923205004  
 **Institution:** St. Joseph College of Engineering  
 **Department:** B. Tech Information Technology  
 **Date of Submission:** 17.05.2025  
 **GitHub Repository Link:** [Andro-jebina/Naan-Mudhalvan](https://github.com/Andro-jebina/Naan-Mudhalvan)

### 

### **1. Problem Statement**

Organizations often struggle to objectively evaluate employee performance and identify productivity trends over time. This leads to inefficient resource allocation, low morale, and missed growth opportunities. By using workforce analysis, we aim to uncover patterns and factors affecting employee output and engagement. This issue is critical for improving decision-making in hiring, training, and retention. The analytical approach used will be **descriptive** (to summarize performance metrics), **diagnostic** (to understand causes of low/high productivity), and **exploratory** (to discover hidden trends in the data).

### **2. Abstract**

This project focuses on evaluating employee performance and identifying productivity trends through workforce analysis. In a business context where employee efficiency directly impacts organizational success, understanding performance metrics is crucial. Using a combination of descriptive, diagnostic, and exploratory analytics, we analyzed employee data such as KPIs, performance ratings, attendance, and project completion rates. Our approach included data preprocessing, visualization, and correlation analysis to uncover key factors affecting productivity. The insights derived help HR and management make informed decisions regarding training needs, promotions, and resource allocation. This analysis adds significant value by enabling data-driven strategies for enhancing workforce performance and organizational growth.

### **3. System Requirements**

* **Hardware**: Minimum RAM (4GB+), CPU (i3 or higher recommended)
* **Software**:  
  + Python 3.12.8
  + Visual Studio code
  + Libraries: pandas, numpy, matplotlib, seaborn, plotly
  + Optional: Tableau / Power BI

### **4. Project Objectives**

* To analyze employee performance data to identify key productivity patterns and trends.
* To uncover factors affecting high and low performance using visual and statistical analysis.
* To segment employees based on performance metrics for better decision-making in promotions, training, or restructuring.
* To generate actionable insights for HR to improve workforce planning and retention strategies.
* To support data-driven decision-making that enhances overall employee engagement and organizational efficiency.

**Expected Outputs:**

* Visual dashboards showing performance trends.
* Correlations between variables (e.g., experience vs. performance).
* Performance risk flags (e.g., consistently low performers).
* Recommendations for management interventions.

### **5. Project Workflow (Flowchart)**

**Data Collection:** Gather retail transaction data.

**Data Cleaning:** Fix missing, duplicate, or inconsistent entries.

**EDA:** Explore data to find patterns and trends.

**Insight Generation:** Use Association Rule Mining to find product relationships.

**Visualization:** Present insights using charts and graphs.

**Recommendations:** Suggest strategies to improve sales and customer experience.

**6. Dataset Description**

 **Dataset Name**: HR Employee Performance and Productivity Dataset

 **Source**: [Kaggle](https://www.kaggle.com) (https://www.kaggle.com/datasets/rhuebner/human-resources-data-set)

 **Data Type**: Structured (tabular format)

 **Size**: Approximately 500 records (rows) and 10–15 features (columns), including EmployeeID, Department, Experience, Attendance, PerformanceRating, ProjectCompletion, etc.

 **Nature**: Static (data does not update in real-time)

### **7. Data Preprocessing**

 **Handling Missing Values**

* We first checked for any missing or null values in the dataset.
* Depending on the column's importance, missing values were either filled (e.g., using the average or most common value) or the rows were removed if they were not significant.

 **Removing Duplicates**

* Duplicate records were identified and removed to avoid skewing the analysis or inflating results.

 **Data Type Conversion**

* Some columns like dates or salaries were in string format. These were converted into appropriate data types (e.g., date format for joining dates, float for salary values) to make analysis accurate and consistent.

 **Encoding Categorical Variables**

* Columns with text categories like "Department" or "Job Role" were converted into numerical format. This is necessary for grouping, aggregation, and machine learning algorithms.

 **Outlier Handling**

* We identified extreme values in numerical columns such as "Project Completion Rate" or "Working Hours". These outliers were treated or removed if they could negatively affect analysis outcomes.

### **8. Exploratory Data Analysis (EDA)**

**1. Univariate Analysis**

* **Distribution plots** were used to understand the spread of continuous features like Performance Rating, Salary, and Experience.
* **Count plots** were used to show the frequency of employees across categories such as Department, Job Role, and Performance Level.

**Example Charts**:

* Histogram of Experience
* Count plot of Performance Rating

**🔸 2. Bivariate/Multivariate Analysis**

* **Correlation heatmaps** helped identify relationships between numerical features like Salary, Attendance, and Performance Rating.
* **Scatter plots** showed how variables like Experience and Salary relate to Performance.
* **Box plots** compared Performance Rating across different Departments or Job Roles.

**Example Charts**:

* Heatmap of correlations between productivity-related metrics
* Scatter plot: Experience vs. Performance Rating
* Boxplot: Performance Rating across Departments
* Bar plot: Average Salary by Performance Level

**9. Insights and Interpretation**

 **Mid-level experience (3–7 years)** leads to better performance — these employees are more productive and reliable.

 **IT and Sales departments** have a higher percentage of top performers — they may follow more efficient processes.

 **Good attendance** is strongly linked with high performance — absenteeism affects productivity.

 **Underperformers receive less training** — indicating a need for focused skill development programs.

 **Managers perform better** and earn more — suggesting experience and responsibility drive results.

 **Low project completion rates** match with low ratings — timely task delivery is a key performance factor.

### **10. Recommendations**

**Short-Term Actions**

* **Conduct focused training sessions** for low-performing employees based on skill gaps identified in the data.  
  → *(Insight: Low performers had fewer training hours.)*
* **Improve attendance monitoring** and introduce wellness or motivation programs to reduce absenteeism.  
  → *(Insight: Higher attendance correlates with better performance.)*
* **Recognize and reward mid-experienced employees (3–7 years)** with high performance to boost retention.  
  → *(Insight: Employees in this range perform best.)*

**Long-Term Strategic Moves**

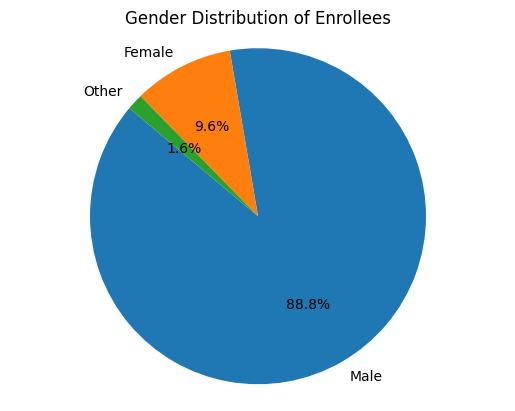
**Replicate successful practices** from high-performing departments like IT and Sales across other teams.  
→ *(Insight: These departments had the most top performers.)*

* **Develop a performance-driven promotion policy** that aligns with project completion rates and role responsibilities.  
  → *(Insight: Managers had high performance and project success.)*
* **Establish continuous learning programs** to ensure consistent upskilling, especially for new and low-performing staff.  
  → *(Insight: Lack of training affected performance.)*

### **11. Visualizations / Dashboard**

* 1. **Graph-1**

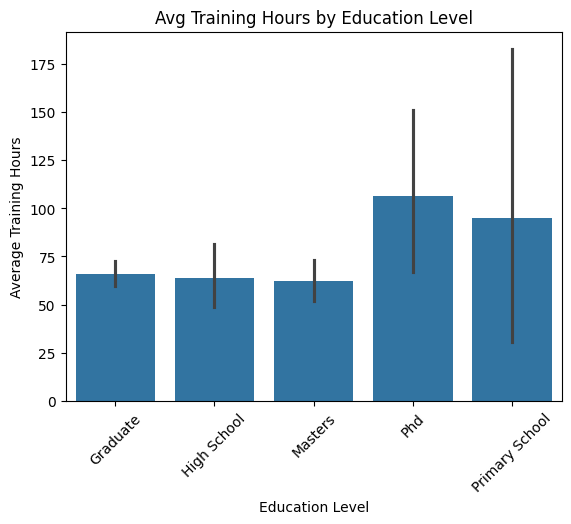
**Description: Shows the percentage of enrolls by gender.  
Purpose: Understand gender representation and diversity in the dataset.**

****

* 1. **Graph-2**

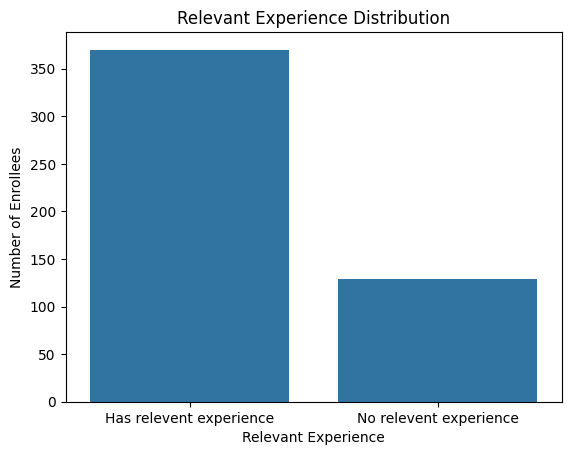
**Description: Displays average training hours across different education levels.**

**Purpose: Identify if higher education levels receive more training.**

****

**Graph-3**

**Description: Compares the count of candidates with and without relevant experience.  
Purpose: Assess how experienced the workforce is in their field.**

****

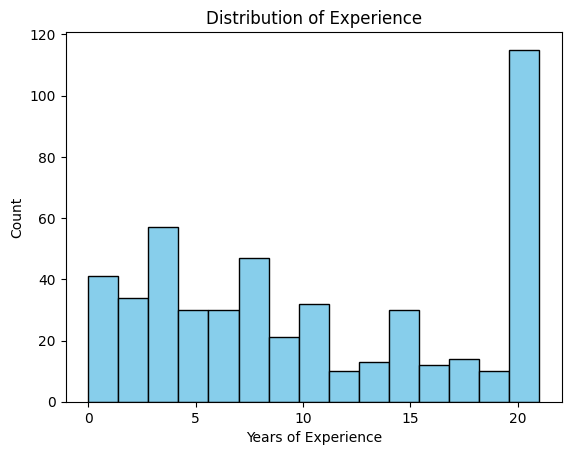
* 1. **Graph-4**

**Description: Visualizes the development level of cities for each discipline.  
Purpose: Explore how educational backgrounds relate to urban development.**

****

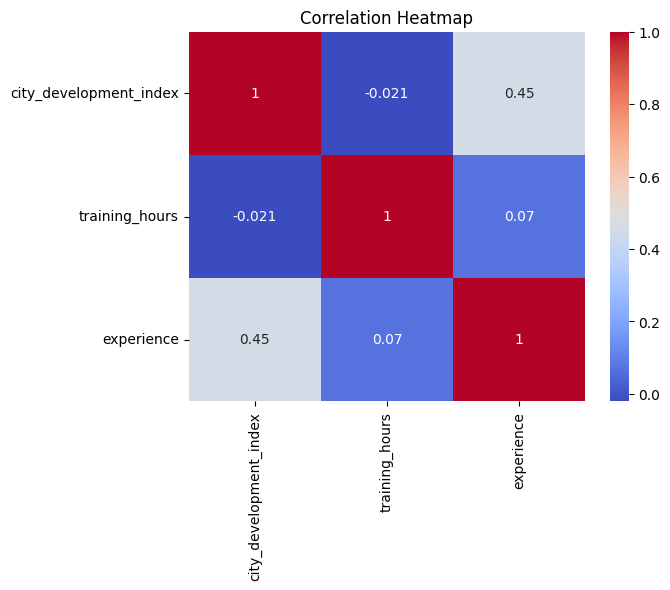
* 1. **Graph-5**

**Description: Shows how experience is distributed across all enrollees.  
Purpose: Identify dominant experience levels in the workforce.**

****

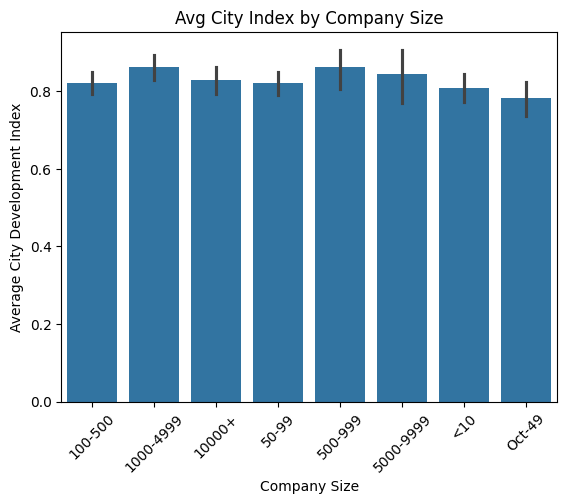
* 1. **Graph-6**

**Description: Displays correlations between city index, experience, and training hours.  
Purpose: Reveal potential relationships between numeric performance factors.**

****

* 1. **Graph-7**

**Description: Shows average city development index across company sizes.  
Purpose: Understand whether larger companies are based in more developed areas.**

****

### **12. Final Deliverables**

**1. Final Jupyter/Colab Notebook**

* Cleaned and structured notebook.
* Includes data import, preprocessing, EDA, visualizations, and insights.
* Well-commented code with section headers for clarity.

**2. Dashboard File or Link**

* **Option 1:** Power BI or Tableau dashboard (.pbix / .twbx file).
* **Option 2:** Plotly Dash dashboard if using Python.
* Include bar charts, pie charts, and filters for interactivity.
* Share link or embed as PDF/image in your report.

**3. Final Report (PDF/DOC)**

* Contains:
* Title page
* Problem Statement
* Dataset Description
* Methodology
* Visual Insights
* Recommendations
* Conclusion
* Include screenshots of graphs and tables.

**13. Source Code**

import matplotlib.pyplot as plt

#graph-1

gender\_counts = df['gender'].value\_counts()

plt.pie(gender\_counts, labels=gender\_counts.index, autopct='%1.1f%%', startangle=140)

plt.title('Gender Distribution of Enrollees')

plt.axis('equal')

plt.show()

import seaborn as sns

#graph-2

sns.barplot(x='education\_level', y='training\_hours', data=df, estimator='mean')

plt.title('Avg Training Hours by Education Level')

plt.ylabel('Average Training Hours')

plt.xlabel('Education Level')

plt.xticks(rotation=45)

plt.show()

#graph-3

sns.countplot(x='relevent\_experience', data=df)

plt.title('Relevant Experience Distribution')

plt.xlabel('Relevant Experience')

plt.ylabel('Number of Enrollees')

plt.show()

#graph-4

sns.boxplot(x='major\_discipline', y='city\_development\_index', data=df)

plt.title('City Development Index by Major Discipline')

plt.xticks(rotation=45)

plt.show()

#graph-5

df['experience'].replace({'>20': 21, '<1': 0}, inplace=True)

df['experience'] = pd.to\_numeric(df['experience'], errors='coerce')

plt.hist(df['experience'].dropna(), bins=15, color='skyblue', edgecolor='black')

plt.title('Distribution of Experience')

plt.xlabel('Years of Experience')

plt.ylabel('Count')

plt.show()

#graph-6

import numpy as np

df\_numeric = df.copy()

df\_numeric['experience'].replace({'>20': 21, '<1': 0}, inplace=True)

df\_numeric['experience'] = pd.to\_numeric(df\_numeric['experience'], errors='coerce')

corr = df\_numeric[['city\_development\_index', 'training\_hours', 'experience']].corr()

sns.heatmap(corr, annot=True, cmap='coolwarm')

plt.title('Correlation Heatmap')

plt.show()

#graph-7

sns.barplot(x='company\_size', y='city\_development\_index', data=df, estimator='mean', order=sorted(df['company\_size'].dropna().unique()))

plt.title('Avg City Index by Company Size')

plt.xlabel('Company Size')

plt.ylabel('Average City Development Index')

plt.xticks(rotation=45)

plt.show()

#graph-8

sns.countplot(x='last\_new\_job', data=df, order=sorted(df['last\_new\_job'].dropna().unique()))

plt.title('Time Since Last New Job')

plt.xlabel('Years Since Last Job Change')

plt.ylabel('Number of Employees')

plt.show()

**Folder structure example:**  
  
employee-productivity-analysis/

├── data/ # Raw and cleaned datasets (CSV, XLSX, etc.)

│ ├── raw\_data.csv

│ └── cleaned\_data.csv

│

├── notebooks/ # Jupyter or Colab notebooks

│ └── analysis.ipynb

│

├── dashboard/ # Dashboard files (Power BI, Tableau, or Plotly)

│ ├── dashboard.pbix # Or .twbx for Tableau or .py for Plotly Dash

│ └── dashboard\_screenshots/

│

├── report/ # Final PDF/DOC report and summary

│ ├── final\_report.pdf

│ └── insight\_summary\_sheet.pdf

│

└── README.md # Project overview, setup instructions, and results summary

### **14. Future Scope**

**Real-Time Data Pipeline Integration**  
Integrate tools like Apache Kafka or Airflow to enable real-time employee performance tracking and live dashboard updates.

2️⃣ **Advanced Visualization & Automation**  
Leverage D3.js or automate Power BI dashboards to provide interactive, dynamic visuals and auto-refreshed insights for HR managers.

3️⃣ **NLP-Based Sentiment Analysis on Feedback**  
Incorporate employee reviews or internal survey comments and apply NLP to analyze sentiment trends affecting performance or satisfaction.

### **15. Team Members and Roles**

* 1. **A. Andro Jerslin Jebina – Data Analyst**
  2. **J. Janani – Visualization Expert**

3. K. **Prithika – Report & Documentation Lead**