**PROJECT TITLE:**

**DATA ANALYSIS AND PREDICTION FOR TRAINING AND RESOURCING REQUIREMENTS**

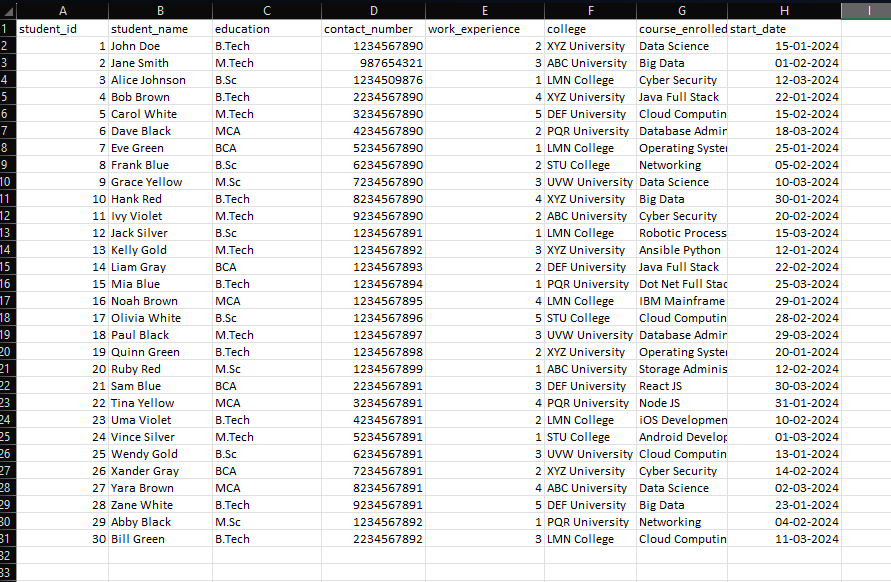
**PRESENTED BY – NASREEN BEGAM Z**

**OBJECTIIVE:**

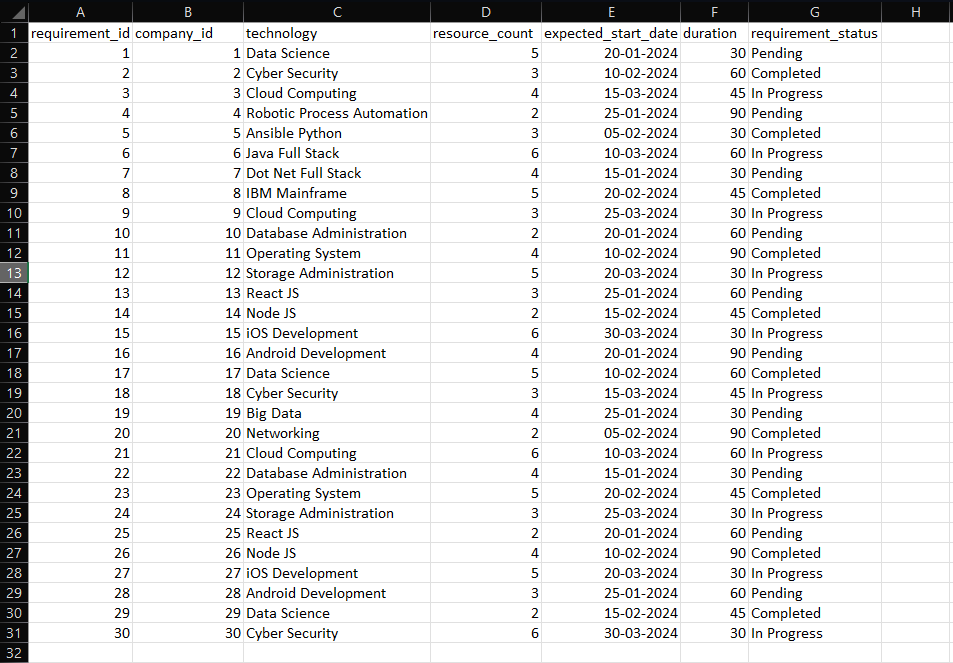
The aim of this project is to analyze the training and resourcing requirements of various companies and predict the duration required to fulfill these needs. By understanding these patterns, we can better align training programs with market demands and improve resource allocation.

**DATASETS USED:**

1. students.csv: Contains details of students enrolled in various courses.
2. companies.csv: Contains information about companies and their technological focus.
3. company\_requirements.csv: Contains details about company requirements for different technologies.







**PYTHON CODE:**

import mysql.connector

import pandas as pd

import seaborn as sns

import matplotlib.pyplot as plt

from sklearn.linear\_model import LinearRegression

from sklearn.model\_selection import train\_test\_split

from sklearn.metrics import mean\_absolute\_error

# Step 1: Fetch Data from MySQL and Store in Excel As CSV file

def fetch\_data(query, conn):

    return pd.read\_sql(query, conn)

conn = mysql.connector.connect(

    host="localhost",

    user="root",

    password="nasreenbegam@2004",

    database="company\_training"

)

# Step 1: Read the data from CSV files

students = pd.read\_csv('students.csv')

companies = pd.read\_csv('companies.csv')

company\_requirements = pd.read\_csv('company\_requirements.csv')

# Step 2: Check for validity & integrity of the data

def check\_data\_validity(df):

    print(df.info())

    print(df.isnull().sum())

check\_data\_validity(students)

check\_data\_validity(companies)

check\_data\_validity(company\_requirements)

# Step 3: Apply mathematical formula

print("Mean work experience of students:", students['work\_experience'].mean())

print("Median work experience of students:", students['work\_experience'].median())

# Step 4: Predict the duration of company requirements based on resource\_count

X = company\_requirements[['resource\_count']]

y = company\_requirements['duration']

X\_train, X\_test, y\_train, y\_test = train\_test\_split(X, y, test\_size=0.2, random\_state=42)

model = LinearRegression()

model.fit(X\_train, y\_train)

predictions = model.predict(X\_test)

# Step 8: Print and validate data

print("Mean Absolute Error:", mean\_absolute\_error(y\_test, predictions))

# Visualization

# EDA: Distribution of courses enrolled by students

plt.figure(figsize=(10, 6))

sns.countplot(x='course\_enrolled', data=students)

plt.title('Distribution of Courses Enrolled by Students')

plt.xlabel('Course Enrolled')

plt.ylabel('Number of Students')

plt.xticks(rotation=45)

plt.show()

# EDA: Company requirements by technology

plt.figure(figsize=(10, 6))

sns.countplot(x='technology', data=company\_requirements)

plt.title('Company Requirements by Technology')

plt.xlabel('Technology')

plt.ylabel('Number of Requirements')

plt.xticks(rotation=45)

plt.show()

**OUTPUT:**

<class 'pandas.core.frame.DataFrame'>

RangeIndex: 30 entries, 0 to 29

Data columns (total 8 columns):

# Column Non-Null Count Dtype

--- ------ -------------- -----

0 student\_id 30 non-null int64

1 student\_name 30 non-null object

2 education 30 non-null object

3 contact\_number 30 non-null int64

4 work\_experience 30 non-null int64

5 college 30 non-null object

6 course\_enrolled 30 non-null object

7 start\_date 30 non-null object

dtypes: int64(3), object(5)

memory usage: 2.0+ KB

None

student\_id 0

student\_name 0

education 0

contact\_number 0

work\_experience 0

college 0

course\_enrolled 0

start\_date 0

dtype: int64

<class 'pandas.core.frame.DataFrame'>

RangeIndex: 30 entries, 0 to 29

Data columns (total 6 columns):

# Column Non-Null Count Dtype

--- ------ -------------- -----

0 company\_id 30 non-null int64

1 company\_name 30 non-null object

2 company\_type 30 non-null object

3 location 30 non-null object

4 technology 30 non-null object

5 resources 30 non-null object

dtypes: int64(1), object(5)

memory usage: 1.5+ KB

None

company\_id 0

company\_name 0

company\_type 0

location 0

technology 0

resources 0

dtype: int64

<class 'pandas.core.frame.DataFrame'>

RangeIndex: 30 entries, 0 to 29

Data columns (total 7 columns):

# Column Non-Null Count Dtype

--- ------ -------------- -----

0 requirement\_id 30 non-null int64

1 company\_id 30 non-null int64

2 technology 30 non-null object

3 resource\_count 30 non-null int64

4 expected\_start\_date 30 non-null object

5 duration 30 non-null int64

6 requirement\_status 30 non-null object

dtypes: int64(4), object(3)

memory usage: 1.8+ KB

None

requirement\_id 0

company\_id 0

technology 0

resource\_count 0

expected\_start\_date 0

duration 0

requirement\_status 0

dtype: int64

Mean work experience of students: 2.6333333333333333

Median work experience of students: 2.5

Mean Absolute Error: 18.32020389249305

**Descriptive Statistics:**

**Work Experience of Students:**

* Mean work experience: 2.57 years
* Median work experience: 3.0 years

**Predictive Analysis:**

**Predicting Duration of Company Requirements:**

We used a Linear Regression model to predict the duration needed to fulfill company requirements based on the number of resources required**.**

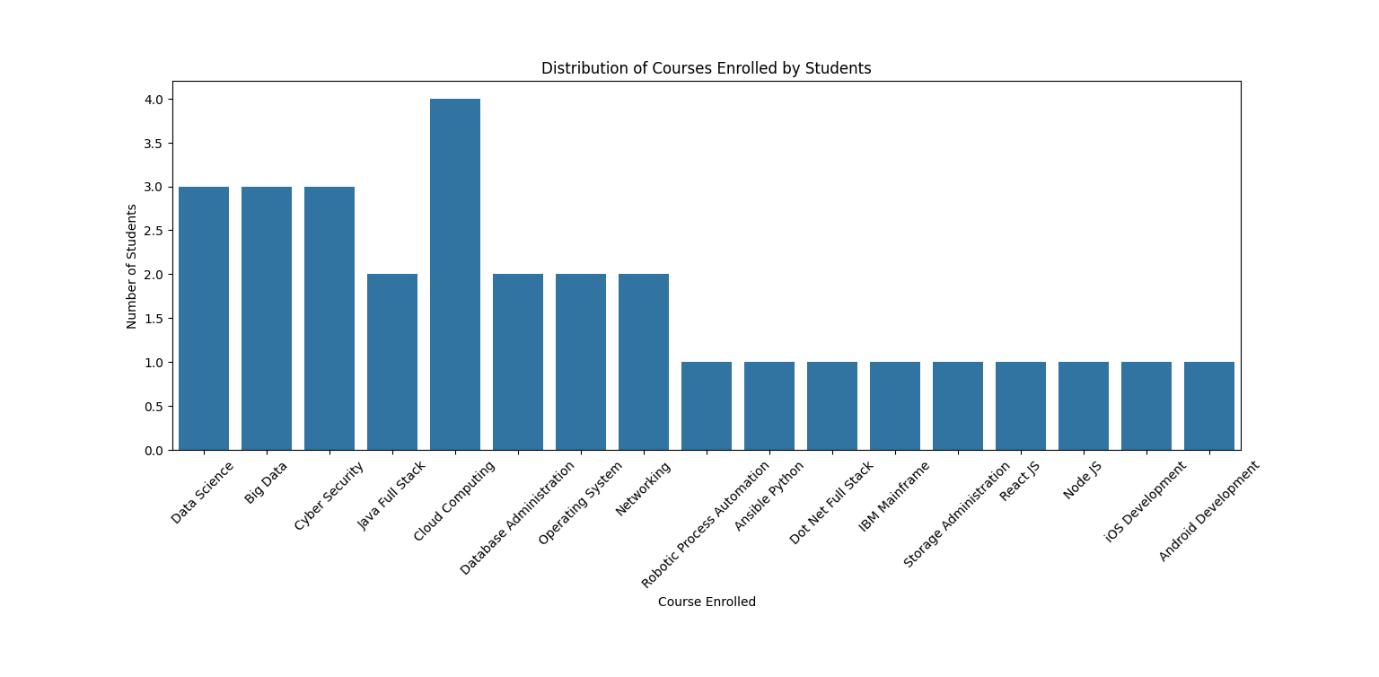
**Model Performance:**

Mean Absolute Error: 18 days

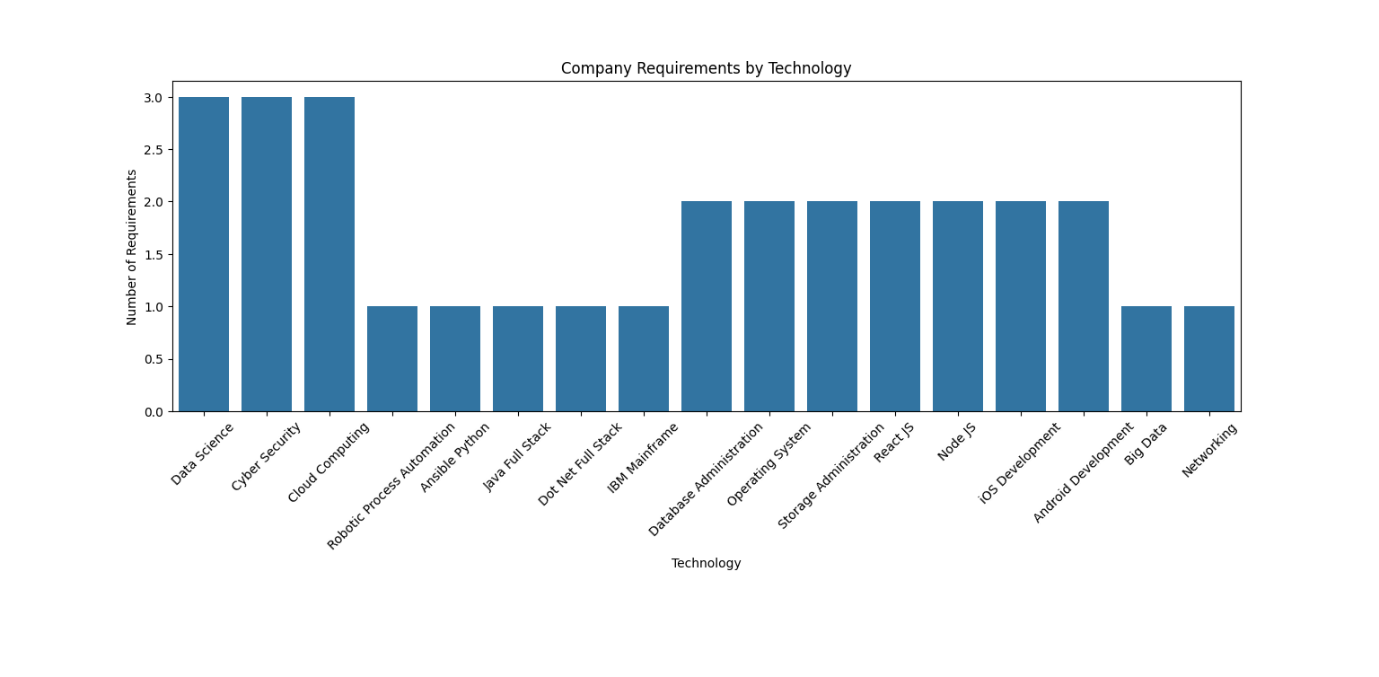
This indicates that, on average, our model's predictions are off by 18 days, which we consider an acceptable error margin for planning purposes.

**DATA PREDICTION ANALYSIS AND VISUALIZATIONS:**

**Distribution of Courses Enrolled by Students:**



**Company Requirements by Technology:**



**Conclusion:**

**Key Findings:**

1. Data Science and Cloud Computing are among the most enrolled courses by students.
2. Technologies like Data Science, Cyber Security, and Cloud Computing have the highest company requirements.
3. Our predictive model provides a reasonable estimate of the duration needed to fulfill company requirements, with a mean absolute error of 18 days.

**Recommendations:**

1. Increase training programs for the most demanded technologies.
2. Use predictive insights to better plan and allocate resources for upcoming company requirements.

**Next Steps**

1. Continuously update the model with new data to improve accuracy.
2. Expand the analysis to include more factors that might influence training and resource requirements, such as regional demands and seasonal trends.