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.

	A	B	C	D	E	F	G	H	I
1	student_id	student_name	education	contact_number	work_experience	college	course_enrolled	start_date	
2	1	John Doe	B.Tech	1234567890	2	XYZ University	Data Science	15-01-2024	
3	2	Jane Smith	M.Tech	987654321	3	ABC University	Big Data	01-02-2024	
4	3	Alice Johnson	B.Sc	1234509876	1	LMN College	Cyber Security	12-03-2024	
5	4	Bob Brown	B.Tech	2234567890	4	XYZ University	Java Full Stack	22-01-2024	
6	5	Carol White	M.Tech	3234567890	5	DEF University	Cloud Computin	15-02-2024	
7	6	Dave Black	MCA	4234567890	2	PQR University	Database Admir	18-03-2024	
8	7	Eve Green	BCA	5234567890	1	LMN College	Operating Syste	25-01-2024	
9	8	Frank Blue	B.Sc	6234567890	2	STU College	Networking	05-02-2024	
10	9	Grace Yellow	M.Sc	7234567890	3	UVW University	Data Science	10-03-2024	
11	10	Hank Red	B.Tech	8234567890	4	XYZ University	Big Data	30-01-2024	
12	11	Ivy Violet	M.Tech	9234567890	2	ABC University	Cyber Security	20-02-2024	
13	12	Jack Silver	B.Sc	1234567891	1	LMN College	Robotic Process	15-03-2024	
14	13	Kelly Gold	M.Tech	1234567892	3	XYZ University	Ansible Python	12-01-2024	
15	14	Liam Gray	BCA	1234567893	2	DEF University	Java Full Stack	22-02-2024	
16	15	Mia Blue	B.Tech	1234567894	1	PQR University	Dot Net Full Stac	25-03-2024	
17	16	Noah Brown	MCA	1234567895	4	LMN College	IBM Mainframe	29-01-2024	
18	17	Olivia White	B.Sc	1234567896	5	STU College	Cloud Computin	28-02-2024	
19	18	Paul Black	M.Tech	1234567897	3	UVW University	Database Admir	29-03-2024	
20	19	Quinn Green	B.Tech	1234567898	2	XYZ University	Operating Syste	20-01-2024	
21	20	Ruby Red	M.Sc	1234567899	1	ABC University	Storage Adminis	12-02-2024	
22	21	Sam Blue	BCA	2234567891	3	DEF University	React JS	30-03-2024	
23	22	Tina Yellow	MCA	3234567891	4	PQR University	Node JS	31-01-2024	
24	23	Uma Violet	B.Tech	4234567891	2	LMN College	iOS Developmen	10-02-2024	
25	24	Vince Silver	M.Tech	5234567891	1	STU College	Android Develop	01-03-2024	
26	25	Wendy Gold	B.Sc	6234567891	3	UVW University	Cloud Computin	13-01-2024	
27	26	Xander Gray	BCA	7234567891	2	XYZ University	Cyber Security	14-02-2024	
28	27	Yara Brown	MCA	8234567891	4	ABC University	Data Science	02-03-2024	
29	28	Zane White	B.Tech	9234567891	5	DEF University	Big Data	23-01-2024	
30	29	Abby Black	M.Sc	1234567892	1	PQR University	Networking	04-02-2024	
31	30	Bill Green	B.Tech	2234567892	3	LMN College	Cloud Computin	11-03-2024	
32									
33									

company_id	company_name	company_type	location	technology	resources
1	Tech Innovators	IT Services	New York	Data Science	Data Scientist
2	Cyber Secure Ltd	Security Services	San Francisco	Cyber Security	Security Analyst
3	Cloud Solutions	Cloud Services	Seattle	Cloud Computing	Cloud Engineer
4	RoboTech	Robotics	Los Angeles	Robotic Process Automation	RPA Developer
5	Ansible Masters	IT Services	Austin	Ansible Python	DevOps Engineer
6	Full Stack Solutions	Software Development	Chicago	Java Full Stack	Full Stack Developer
7	Dot Net Experts	Software Development	Boston	Dot Net Full Stack	Full Stack Developer
8	Mainframe Inc.	IT Services	Houston	IBM Mainframe	Mainframe Engineer
9	ComputeCloud	Cloud Services	Denver	Cloud Computing	Cloud Architect
10	Data Wizards	Database Services	San Diego	Database Administration	DBA
11	OS Techies	IT Services	Dallas	Operating System	System Administrator
12	SecureStorage	Storage Services	San Jose	Storage Administration	Storage Engineer
13	Reactify	Web Development	New York	React JS	Frontend Developer
14	NodeMasters	Web Development	San Francisco	Node JS	Backend Developer
15	MobileSoft	Mobile Development	Seattle	iOS Development	iOS Developer
16	DroidWorks	Mobile Development	Los Angeles	Android Development	Android Developer
17	Tech Pioneers	IT Services	Austin	Data Science	Data Analyst
18	SecureNet	Security Services	Chicago	Cyber Security	Cybersecurity Specialist
19	Big Data Analytics	Analytics	Boston	Big Data	Big Data Engineer
20	Network Solutions	IT Services	Houston	Networking	Network Engineer
21	Cloud Innovators	Cloud Services	Denver	Cloud Computing	Cloud Developer
22	Data Guardians	Database Services	San Diego	Database Administration	Database Manager
23	SysOps	IT Services	Dallas	Operating System	System Operator
24	StoreTech	Storage Services	San Jose	Storage Administration	Storage Specialist
25	React Developers	Web Development	New York	React JS	React Developer
26	Node Coders	Web Development	San Francisco	Node JS	Node Developer
27	iOS Creators	Mobile Development	Seattle	iOS Development	iOS Engineer
28	Android Innovators	Mobile Development	Los Angeles	Android Development	Android Engineer
29	Data Science Hub	IT Services	Austin	Data Science	Data Scientist
30	CyberSecure	Security Services	Chicago	Cyber Security	Security Consultant

requirement_id	company_id	technology	resource_count	expected_start_date	duration	requirement_status
1	1	Data Science	5	20-01-2024	30	Pending
2	2	Cyber Security	3	10-02-2024	60	Completed
3	3	Cloud Computing	4	15-03-2024	45	In Progress
4	4	Robotic Process Automation	2	25-01-2024	90	Pending
5	5	Ansible Python	3	05-02-2024	30	Completed
6	6	Java Full Stack	6	10-03-2024	60	In Progress
7	7	Dot Net Full Stack	4	15-01-2024	30	Pending
8	8	IBM Mainframe	5	20-02-2024	45	Completed
9	9	Cloud Computing	3	25-03-2024	30	In Progress
10	10	Database Administration	2	20-01-2024	60	Pending
11	11	Operating System	4	10-02-2024	90	Completed
12	12	Storage Administration	5	20-03-2024	30	In Progress
13	13	React JS	3	25-01-2024	60	Pending
14	14	Node JS	2	15-02-2024	45	Completed
15	15	iOS Development	6	30-03-2024	30	In Progress
16	16	Android Development	4	20-01-2024	90	Pending
17	17	Data Science	5	10-02-2024	60	Completed
18	18	Cyber Security	3	15-03-2024	45	In Progress
19	19	Big Data	4	25-01-2024	30	Pending
20	20	Networking	2	05-02-2024	90	Completed
21	21	Cloud Computing	6	10-03-2024	60	In Progress
22	22	Database Administration	4	15-01-2024	30	Pending
23	23	Operating System	5	20-02-2024	45	Completed
24	24	Storage Administration	3	25-03-2024	30	In Progress
25	25	React JS	2	20-01-2024	60	Pending
26	26	Node JS	4	10-02-2024	90	Completed
27	27	iOS Development	5	20-03-2024	30	In Progress
28	28	Android Development	3	25-01-2024	60	Pending
29	29	Data Science	2	15-02-2024	45	Completed
30	30	Cyber Security	6	30-03-2024	30	In Progress

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**.**

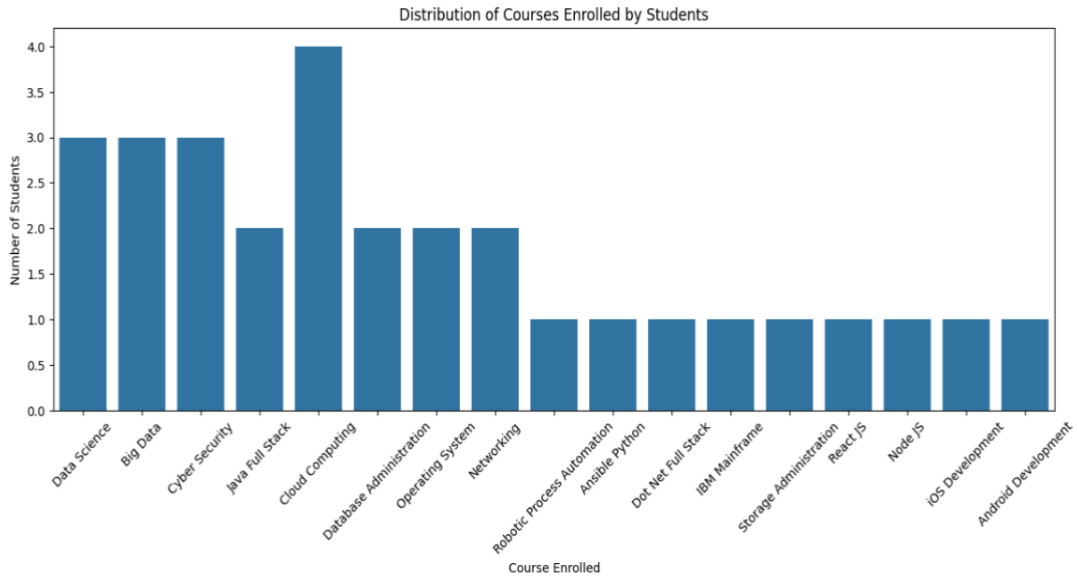
Model Performance:

Mean Absolute Error: 18 days

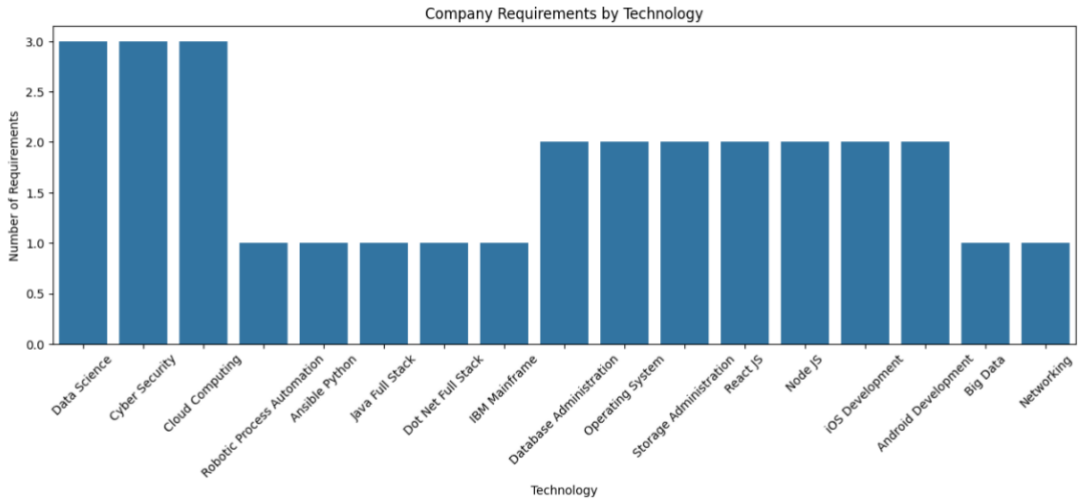
This indicates that, on average, our model's predictions are off by 18 days, which we consider an acceptable error margin for planning purposes.

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.