Group 4
Aaron Le
Anita Shen
Anita Tam
Emy Tang
Hsin-Pei (Angela) Chou
CEO: Ian Glynn
Jae Chung
CCO: Novely Joshi
Tanay Agarwal

# DP FINAL

## UC Berkeley IISE Alumni Mentorship Program Database System

IISE
UC Berkeley Chapter

Berkeley
UNIVERSITY OF CALIFORNIA

IEOR 115 | Fall 2021

# CLIENT: IISE

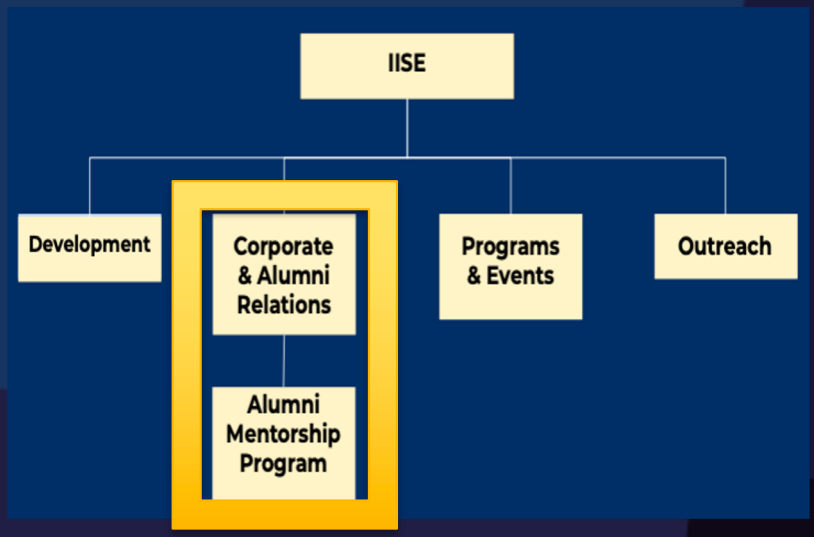
**UC Berkeley Institute of Industrial and Systems Engineers**
***Largest IEOR & ORMS student organization at Cal***

IISE offers academic and professional resources to its members.

Client Goals: Create Centralized Database for the Alumni Mentorship Program and provide insights on its benefits to members

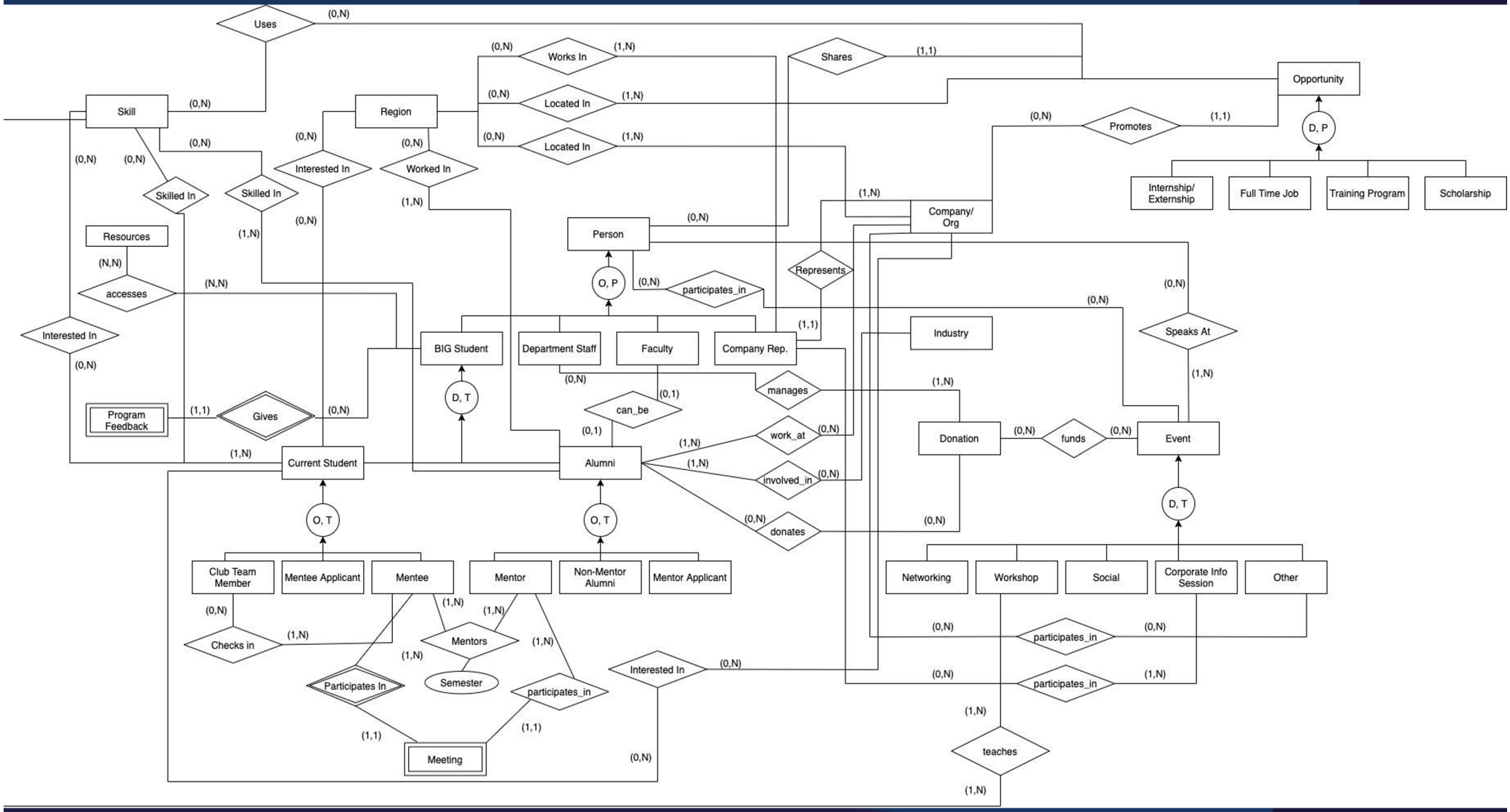
Client Contact: Shaan Sheth, Corporate and Alumni Relations team

Current Data System: Spreadsheets on applicants, member skills, mentorship pairings, events, attendance, etc




Berkeley
UNIVERSITY OF CALIFORNIA

# EER Diagram


Uses
Skill
Region
Works In
Located In
Located In
Shares
Opportunity
Promotes
D, P
Internship/ Externship
Full Time Job
Training Program
Scholarship
Interested In
Worked In
Skilled In
Skilled In
Resources
accesses
Interested In
Person
O, P
participates_in
Represents
Company/ Org
Industry
Speaks At
BIG Student
Department Staff
Faculty
Company Rep.
D, T
manages
can_be
Program Feedback
Gives
Current Student
Alumni
work_at
involved_in
donates
Donation
funds
Event
D, T
O, T
O, T
Club Team Member
Mentee Applicant
Mentee
Mentor
Non-Mentor Alumni
Mentor Applicant
Networking
Workshop
Social
Corporate Info Session
Other
Checks in
Mentors
Semester
Participates In
participates_in
Interested In
participates_in
participates_in
Meeting
teaches

# EER Diagram → Relational Schema

1. Person(PID, Person_name, Phone, Email)
    a. BIG_STUDENT(PID[1], LinkedIn, Years_of_Club_Involvement, Graduation_Year)
        i. Current_Student(PID[1], Club_Team_Member, Mentee_Applicant, Mentee)
        dom(Club_Team_Member) = dom(Mentee_Applicant) = dom(Mentee) = {0,1}
        i. Alumni(PID[1], Mentor, Non-Mentor_Alumni, Mentor_Applicant, Is_Faculty)
        dom(Mentor) = dom(Non-Mentor_Alumni) = dom(Mentor_Applicant) = dom(Is_Faculty) = {0,1}
    a. Department_Staff(PID[1])
    b. Faculty(PID[1], Is_Alumni)
    dom(Is_Alumni) = {0,1}
    c. Company_Rep(PID[1], Company_Legal_Entity_ID[9])

2. Opportunity(OID, URL, OName, Expiration_date, Description, Shared_by_PID[1], Promoted_by_Company_Org_Legal_entity_ID[9])
    a. Internship_Externship(OID[2], Is_paid, Wage, Duration, Start_date)
    b. Full_Time_Job(OID[2], Role, Start_date)
    c. Training_Program (OID[2], Duration)
    d. Scholarship (OID[2], Award_amount)

3. Event(Event_Date, Event Name, Location, Description)
    a. Networking_Event(Event_Date[3], Event Name[3], Networking_Event_Format)
    b. Workshop_Event(Event_Date[3], Event Name[3])
    c. Social_Event(Event_Date[3], Event Name[3])
    d. Corporate_Info_Session(Event_Date[3], Event Name[3], Company_Org_Legal_entity_ID[9])
    e. Other_Event(Event_Date[3], Event Name[3], Event_type)
        dom(Networking_Event_Format) = {mixer, other, null}
        dom{Event_type} = varchar(45)

Berkeley
UNIVERSITY OF CALIFORNIA

# EER Diagram → Relational Schema

4. Skill(Skill_name, Is_technical)
5. Resources(Resource_ID, Resource_URL, Resource_type, Resource_name)
6. Program_Feedback(Feedback_ID, Big_Student_PID[1a], Semester, Program_rating, Additional_comments)
7. Region(Region_ID, Region_name)
8. Meeting(Meeting_number, Mentee_PID[1ai], Mentor_PID[1aii], Rating, Meeting_type)
9. Company_Org(Legal_entity_ID, Involvement_type_in_club, Company_name)
10. Industry(SIC, Industry_name)
11. Donation(Donation_ID, Amount, Date)
12. Hiring_Locations(Legal_entity_ID[9], City)
13. Current_Student_Interested_In_Skill(Skill_Name[4], Current_Student_PID[1ai], Level_of_Interest)

Berkeley
UNIVERSITY OF CALIFORNIA

# EER Diagram → Relational Schema

14. Big_Student_Accesses_Resources(Resource_ID[5], Big_Student_PID[5])
15. Alumni_has_Skill(Skill_Name[4], Alumni_PID[1a], Proficiency_Level, Semester)
16. Current_Student_has_Skill(Skill_Name[4], Current_Student_PID[1a], Semester, Start_proficiency, End_proficiency)
17. Opportunity_Uses_Skill(Skill_Name[4], OID[2])
18. Current_Student_Interested_In_Region(Region_ID[7], Current_Student_PID[1ai])
19. Alumni_Worked_In_Region(Region_ID[7], Alumni_PID[1aii])
20. Company_Rep_Works_In_Region(Region_ID[7], Company_Rep_PID[1d])
21. Opportunity_Located_In_Region(Region_ID[7], OID[2])
22. Company_Org_Located_In_Region(Region_ID[7], Company_Org_Legal_Entity_ID[9])
23. Person_Attends_Event(PID[1], Event_Name[3], Event_Date[3])
24. DepStaff_Manages_Donation(DepStaff_PID[1b], DID[11])

Berkeley
UNIVERSITY OF CALIFORNIA

# EER Diagram → Relational Schema

25. Alumni_Works_At_Company(Alumni_PID[1aii], Company_Org_Legal_Entity_ID[9], Job_Title)
26. Alumni_Involved_In_Industry(Alumni_PID[1aii], SIC[10])
27. Alumni_donates_Donation(Alumni_PID[1aii], DID[11])
28. Donation_Funds_Event(Event_Name[3], Event_Date[3], DID[11], Amount)
29. Person_Speaks_At_Event(PID[1], Event_Name[3], Event_Date[3])
30. CompanyRep_Participates_In_Infosession(Company_Rep_PID[1d], Infosession_Name[3d], Infosession_Date[3d])
31. Workshop_teaches_Skill(Workshop_Date[3b], Workshop_Name[3b], Skill_name[4])
32. Student_Interested_In_Company(Current_Student_PID[1ai], Company_Org_Legal_Entity_ID[9])
33. Mentorship(Mentor_PID[1aii], Mentee_PID[1ai], Semester)
34. Club_Mentee_Check_In(Club_Team_Member_PID[1ai], Mentee_PID[1ai], Description)
35. Mentee_Applicant_Prefers_Mentor_Applicant(Mentee_PID[1ai], Mentor_PID[1aii])
36. Major(Big_Student_PID[1a], Major_name)
37. Hobby(Big_Student_PID[1a], Hobby_name)

Berkeley
UNIVERSITY OF CALIFORNIA

# Normalization: Example 1

3. Event(Event__Date, Event__Name, Location, Description)

Violates 2NF because of a partial dependency:
{Event_Name} -> {Description}

How to normalize:

3. Event(Event__Date, Event__Name, Location)

Event_Description(Event__Name[3], Description)

# Normalization: Example 2

2a. Internship_Externship(OID[2], Is_paid, Wage, Duration, Start_date)

Violates 3NF because of a transitive dependency:

{OID} -> {Wage}

{Wage} -> {Is_paid}

How to normalize:

2a. Internship_Externship(OID[2], Wage, Duration, Start_date)

Berkeley
UNIVERSITY OF CALIFORNIA

# Normalization: Example 3

Modified Relation:

1a. BIG_STUDENT(PID[1], LinkedIn, Years_of_Club_Involvement, Major, Graduation_Year)

Violates 1NF because of a multivalued attribute: one student/alumni could pursue a double major

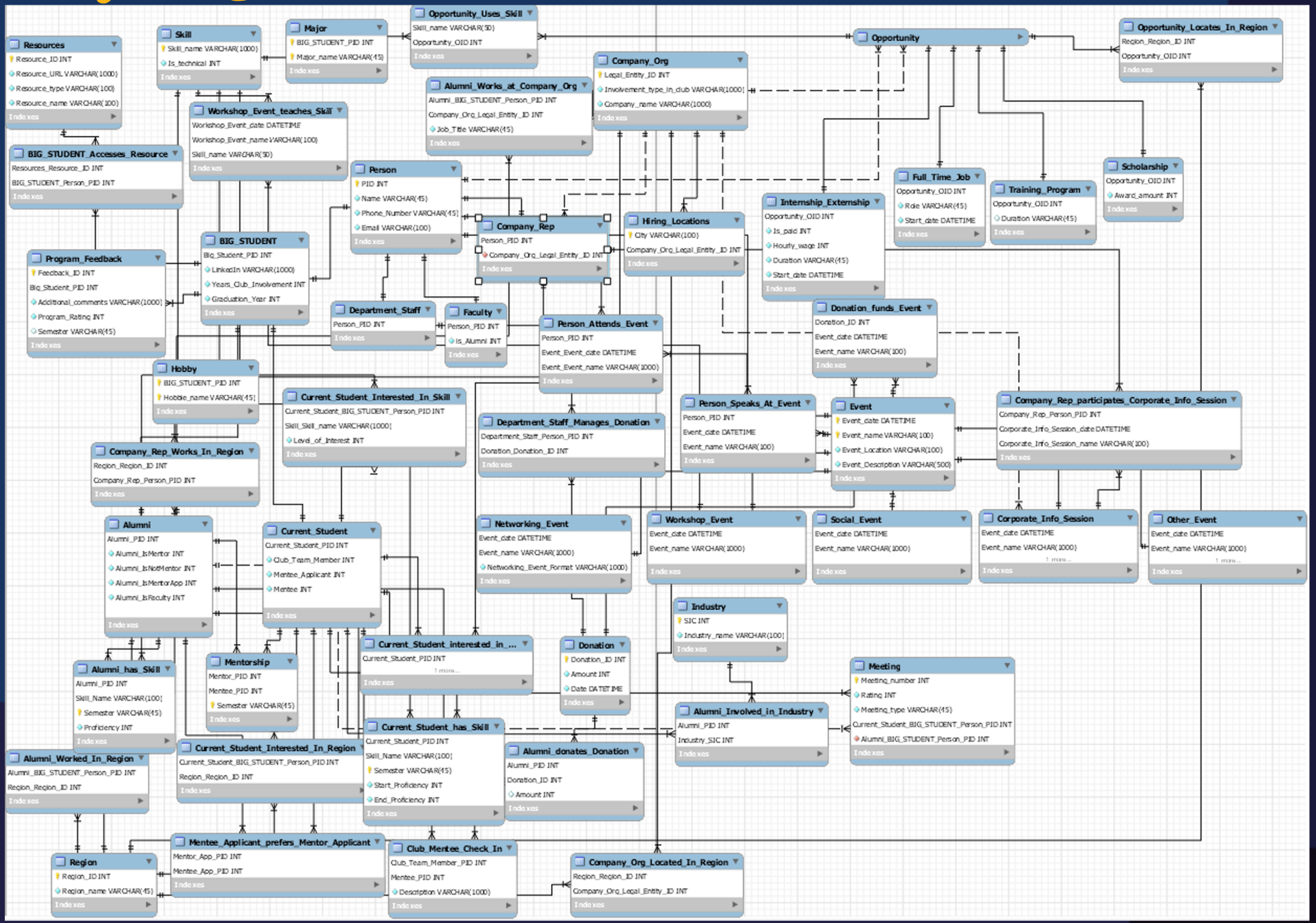
How to normalize:

1a. BIG_STUDENT(PID[1], LinkedIn, Years_of_Club_Involvement, Graduation_Year)

36. Major(Big_Student_PID[1a], Major_name)

Berkeley
UNIVERSITY OF CALIFORNIA

# MySQL


Resources
Resource_ID INT
Resource_URL VARCHAR(1000)
Resource_type VARCHAR(100)
Resource_name VARCHAR(100)
Skill
Skill_name VARCHAR(1000)
Is_technical INT
Major
BIG_STUDENT_PID INT
Major_name VARCHAR(45)
Opportunity_Uses_Skill
Skill_name VARCHAR(50)
Opportunity_OID INT
Opportunity
Opportunity_Locates_In_Region
Region_Region_ID INT
Opportunity_OID INT
Company_Org
Legal_Entity_ID INT
Involvement_type_in_club VARCHAR(1000)
Company_name VARCHAR(1000)
Alumni_Works_at_Company_Org
Alumni_BIG_STUDENT_Person_PID INT
Company_Org_Legal_Entity_ID INT
Job_Title VARCHAR(45)
Workshop_Event_teaches_Skill
Workshop_Event_date DATETIME
Workshop_Event_name VARCHAR(100)
Skill_name VARCHAR(50)
BIG_STUDENT_Accesses_Resource
Resources_Resource_ID INT
BIG_STUDENT_Person_PID INT
Person
PID INT
Name VARCHAR(45)
Phone_Number VARCHAR(45)
Email VARCHAR(100)
Company_Rep
Person_PID INT
Company_Org_Legal_Entity_ID INT
Hiring_Locations
City VARCHAR(100)
Company_Org_Legal_Entity_ID INT
Internship_Externship
Opportunity_OID INT
Is_paid INT
Hourly_wage INT
Duration VARCHAR(45)
Start_date DATETIME
Full_Time_Job
Opportunity_OID INT
Role VARCHAR(45)
Start_date DATETIME
Training_Program
Opportunity_OID INT
Duration VARCHAR(45)
Scholarship
Opportunity_OID INT
Award_amount INT
BIG_STUDENT
Big_Student_PID INT
LinkedIn VARCHAR(1000)
Years_Club_Involvement INT
Graduation_Year INT
Program_Feedback
Feedback_ID INT
Big_Student_PID INT
Additional_comments VARCHAR(1000)
Program_Rating INT
Semester VARCHAR(45)
Department_Staff
Person_PID INT
Faculty
Person_PID INT
Is_Alumni INT
Person_Attends_Event
Person_PID INT
Event_Event_date DATETIME
Event_Event_name VARCHAR(1000)
Donation_funds_Event
Donation_ID INT
Event_date DATETIME
Event_name VARCHAR(100)
Hobby
BIG_STUDENT_PID INT
Hobbie_name VARCHAR(45)
Current_Student_Interested_In_Skill
Current_Student_BIG_STUDENT_Person_PID INT
Skill_Skill_name VARCHAR(1000)
Level_of_Interest INT
Department_Staff_Manages_Donation
Department_Staff_Person_PID INT
Donation_Donation_ID INT
Person_Speaks_At_Event
Person_PID INT
Event_date DATETIME
Event_name VARCHAR(100)
Event
Event_date DATETIME
Event_name VARCHAR(100)
Event_Location VARCHAR(100)
Event_Description VARCHAR(500)
Company_Rep_participates_Corporate_Info_Session
Company_Rep_Person_PID INT
Corporate_Info_Session_date DATETIME
Corporate_Info_Session_name VARCHAR(100)
Company_Rep_Works_In_Region
Region_Region_ID INT
Company_Rep_Person_PID INT
Alumni
Alumni_PID INT
Alumni_IsMentor INT
Alumni_IsNotMentor INT
Alumni_IsMentorApp INT
Alumni_IsFaculty INT
Current_Student
Current_Student_PID INT
Club_Team_Member INT
Mentee_Applicant INT
Mentee INT
Networking_Event
Event_date DATETIME
Event_name VARCHAR(1000)
Networking_Event_Format VARCHAR(1000)
Workshop_Event
Event_date DATETIME
Event_name VARCHAR(1000)
Social_Event
Event_date DATETIME
Event_name VARCHAR(1000)
Corporate_Info_Session
Event_date DATETIME
Event_name VARCHAR(1000)
Other_Event
Event_date DATETIME
Event_name VARCHAR(1000)
Industry
SIC INT
Industry_name VARCHAR(100)
Current_Student_interested_in_...
Current_Student_PID INT
Donation
Donation_ID INT
Amount INT
Date DATETIME
Meeting
Meeting_number INT
Rating INT
Meeting_type VARCHAR(15)
Current_Student_BIG_STUDENT_Person_PID INT
Alumni_BIG_STUDENT_Person_PID INT
Alumni_has_Skill
Alumni_PID INT
Skill_Name VARCHAR(100)
Semester VARCHAR(15)
Proficiency INT
Mentorship
Mentor_PID INT
Mentee_PID INT
Semester VARCHAR(15)
Current_Student_has_Skill
Current_Student_PID INT
Skill_Name VARCHAR(100)
Semester VARCHAR(15)
Start_Proficiency INT
End_Proficiency INT
Alumni_Involved_in_Industry
Alumni_PID INT
Industry_SIC INT
Alumni_donates_Donation
Alumni_PID INT
Donation_ID INT
Amount INT
Alumni_Worked_In_Region
Alumni_BIG_STUDENT_Person_PID INT
Region_Region_ID INT
Current_Student_Interested_In_Region
Current_Student_BIG_STUDENT_Person_PID INT
Region_Region_ID INT
Mentee_Applicant_prefers_Mentor_Applicant
Mentor_App_PID INT
Mentee_App_PID INT
Club_Mentee_Check_In
Club_Team_Member_PID INT
Mentee_PID INT
Description VARCHAR(1000)
Company_Org_Located_In_Region
Region_Region_ID INT
Company_Org_Legal_Entity_ID INT
Region
Region_ID INT
Region_name VARCHAR(45)

# Fake Data Generation

- Primarily made use of the Faker library to generate the more complex "dummy" data.
  - fake.name()
  - fake.phone_number()
  - fake.email()
  - fake.text()

- Custom and basic functions were utilized to generate the rest of the data.
  - PID function, LinkedIn URL function, list comprehension, random.choice(), and np.arange().

- List of Tables Populated:
  - Club Mentee Check In
  - Program Ratings
  - Person
  - Big_Student
  - Current Student
  - Alumni
  - Alumni has Skill
  - Current Student has Skill
  - Mentorship

Berkeley
UNIVERSITY OF CALIFORNIA

# Fake Data Generation: Functions

## LinkedIn URLs:

```
def generate_linkedin():
        random_letters_num_list = ['a', 'b', 'c', 'e', 'h', 'w', '1', '2', '4', '9', 'q', 'v', '12']
        suffix = ""
        for i in np.arange(6):
            suffix += random.choice(random_letters_num_list)
        return 'www.linkedin.com/' + suffix
```

## Name, Phone Numbers, and Emails:

```
alumni_names = [fake.name() for i in np.arange(number_alumni)]
alumni_phones = [fake.phone_number() for i in np.arange(number_alumni)]
alumni_emails = [fake.email() for i in np.arange(number_alumni)]
```

## Alumni & Current Student PIDs:

```
def get_alumni_pid(num_alumni):
    return np.arange(1, num_alumni + 1)

def get_current_student_pid(num_alumni, num_current_student):
    return np.arange(num_alumni + 1, num_alumni + num_current_student)
```

Berkeley
UNIVERSITY OF CALIFORNIA

## Query 1: How do we evaluate the general mentorship ratings? How can we interpret the feedbacks from mentees?

### SQL:

```
SELECT f1.Big_Student_PID as mentor_PID, f2.Big_Student_PID as Current_Student_PID,
        f1.Program_Rating as mentor_rating, f2.Program_Rating as mentee_rating,
checkin.Description
FROM Program_Feedback f1, Program_Feedback f2, Mentorship m, Club_Mentee_Check_In as checkin
WHERE f1.Big_Student_PID < f2.Big_Student_PID,
                AND
f1.Big_Student_PID = m.Mentor_PID,
                AND
f2.Big_Student_PID = m.Mentee_PID,
                AND
checkin.Mentee_PID = f2.Big_Student_PID;
```

### Output

| Mentor_PID | Current_Student_PID | Description | Mentee_Rating | Mentor_Rating |
|---|---|---|---|---|
| 1 | 9 | I love this program so much! Definitely gained a lot of professional resources. | 5 | 4 |
| 2 | 9 | I truly disliked my experience and don't want to return again. It was not helpful. | 5 | 4 |
| 3 | 9 | I think the program was very useful and I gained a strong network from it. | 3 | 4 |
| 4 | 9 | I think the program took too much time and didn't offer as much as I thought it would have. I would join again if it had more benefits. | 6 | 5 |
| 1 | 12 | Definitely gained a lot of professional help with recruitment. | 2 | 1 |
| 6 | 12 | It was not helpful because I didn't have a good experience with my mentor. | 4 | 3 |
| 8 | 12 | Definitely gained a lot of professional help with interviews. Really liked it! | 4 | 3 |
| 1 | 14 | I developed a strong bond with my mentor and they were very helpful. | 4 | 3 |
| 8 | 14 | Definitely gained a lot of professional help with resume. | 2 | 3 |
| 3 | 11 | Definitely gained a lot of professional help with grad school. | 4 | 3 |
| 4 | 11 | The program was kinda useful, but didn't stand out very much. | 2 | 3 |
| 3 | 13 | I developed a strong bond with my mentor and they were very helpful in giving post-graduation advice! | 6 | 5 |
| 5 | 13 | The program was kinda useful, but didn't stand out very much. I wish it took less time. | 5 | 4 |
| 6 | 13 | I loved it! It was a great opportunity to connect with successful IEOR alumni. | 5 | 4 |

Berkeley
UNIVERSITY OF CALIFORNIA

# Query 1 Solution: Natural Language Processing

## NLP to Classify Positive/Negative Feedback from Mentee Checkins



Most Positive Comments:

1. I loved it! It was a great opportunity to connect with successful IEOR alumni.
2. Definitely gained a lot of professional help with interviews. Really liked it!
3. I developed a strong bond with my mentor and they were very helpful in giving post-graduation advice!

Most Negative Comments:

1. The program was kinda useful, but didn't stand out very much.
2. I truly disliked my experience and don't want to return again. It was not helpful.
3. I think the program took too much time and didn't offer as much as I thought it would have.

- VADER Lexicon index to capture sentiment of all Mentee feedback
- Rank based on positive and negative feedback
- Word Cloud visualization

Berkeley
UNIVERSITY OF CALIFORNIA

# Description Sentiment Analysis

```
print(''.join(open("vader_lexicon.txt").readlines()[:10]))
```

```
$:      -1.5    0.80623 [-1, -1, -1, -1, -3, -1, -3, -1, -2, -1]
%)      -0.4    1.0198  [-1, 0, -1, 0, 0, -2, -1, 2, -1, 0]
%-)     -1.5    1.43178 [-2, 0, -2, -2, -1, 2, -2, -3, -2, -3]
&-:     -0.4    1.42829 [-3, -1, 0, 0, -1, -1, -1, 2, -1, 2]
&:      -0.7    0.64031 [0, -1, -1, -1, 1, -1, -1, -1, -1, -1]
( '}{' )        1.6     0.66332 [1, 2, 2, 1, 1, 2, 2, 1, 3, 1]
(%      -0.9    0.9434  [0, 0, 1, -1, -1, -1, -2, -2, -1, -2]
('-:    2.2     1.16619 [4, 1, 4, 3, 1, 2, 3, 1, 2, 1]
(':     2.3     0.9     [1, 3, 3, 2, 2, 4, 2, 3, 1, 2]
((-:    2.1     0.53852 [2, 2, 2, 1, 2, 3, 2, 2, 3, 2]
```

```
tidyComments = undup_comments['CleanedDescription'].str.split(expand = True).stack().reset_index(level =1).rename(columns= {'level_1':'num', 0:'wor
tidyComments

merged = tidyComments.merge(sent, how = 'left', left_on = 'word', right_index = True)
merged = merged.fillna(0)
merged = merged.reset_index().groupby('index').sum()
mergedComments = pd.merge(undup_comments, merged, left_index=True, right_index=True)
mergedComments
undup_comments['polarity'] = mergedComments['polarity']
undup_comments.head()
```

| | Mentor_PID | Current_Student_PID | Description | Mentee_Rating | Mentor_Rating | CleanedDescription | polarity |
|---|---|---|---|---|---|---|---|
| **0** | 1 | 9 | I love this program so much! Definitely gained... | 5 | 4 | i love this program so much definitely gained... | 6.5 |
| **1** | 2 | 9 | I truly disliked my experience and don't want ... | 5 | 4 | i truly disliked my experience and don t want ... | 2.3 |
| **2** | 3 | 9 | I think the program was very useful and I gain... | 3 | 4 | i think the program was very useful and i gain... | 5.8 |
| **3** | 4 | 9 | I think the program took too much time and did... | 6 | 5 | i think the program took too much time and did... | 2.8 |
| **4** | 1 | 12 | Definitely gained a lot of professional help w... | 2 | 1 | definitely gained a lot of professional help w... | 5.0 |

Berkeley
UNIVERSITY OF CALIFORNIA

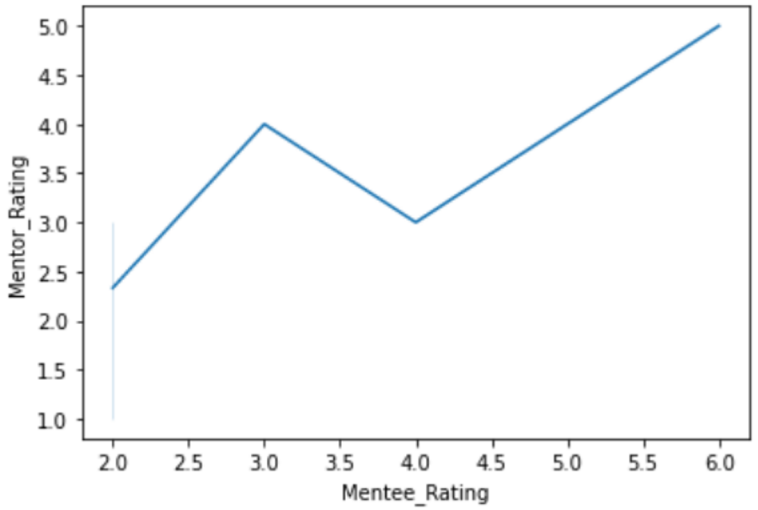
## Query 1 Solution:

## Linear Regression Model

- Fit linear line to see relationship between mentee's rating of the program and mentor's rating of the program
- Visualize trends, find correlations
- Predictive analysis for specific mentors/mentees

### Line Graph for Mentee/Mentor Program Ratings

```
sns.lineplot(data=undup_comments, x="Mentee_Rating", y="Mentor_Rating")
plt.show()
```




Mentor Rating = 1.33 * Mentee Rating -1.83
(y = 1.33 x -1.83)

Berkeley
UNIVERSITY OF CALIFORNIA

**Query 2:** **How skilled should a mentee expect to be at the end of the mentorship program? What factors contribute most to high proficiency level?**

SQL:

```
SELECT cs.Skill_name AS Skill_Name, a.Proficiency_Level AS
    Mentor_Proficiency, cs.Start_proficiency AS
    Mentee_Proficiency_Before_Mentorship, cs.End_proficiency AS
    Mentee_Proficiency_After_Mentorship
FROM Alumni_has_Skill AS a, Current_Student_has_Skill AS cs, Mentorship AS
    m
WHERE a.Alumni_PID = m.Mentor_PID AND cs.Current_Student_PID =
    m.Mentee_PID AND a.Semester = m.Semester AND cs.Semester =
    m.Semester AND a.Skill_name = cs.Skill_name;
```

Berkeley
UNIVERSITY OF CALIFORNIA

**Query 2:** How skilled should a mentee expect to be at the end of the mentorship program? What factors contribute most to high proficiency level?

```
# test train split
y = syndata['Student_End_Proficiency']
X = syndata.drop(['Student_End_Proficiency'], axis=1)
X.astype({'Mentor_PID': 'str'})

X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.3, random_state=88)
X_train.shape, X_test.shape
```

```
syndata_train = pd.concat([X_train, y_train], axis=1, join='inner')
syndata_test = pd.concat([X_test, y_test], axis=1, join='inner')
```

```
# Model
model = smf.ols(formula='Student_End_Proficiency ~ Mentor_PID + Mentor_Proficiency + Studer
                         data=syndata_train).fit()
```

**Remove insignificant feature and remodel, until all features are significant up to 95% confidence interval**

Berkeley
UNIVERSITY OF CALIFORNIA

**Query 2:** **How skilled should a mentee expect to be at the end of the mentorship program? What factors contribute most to high proficiency level?**

Result:

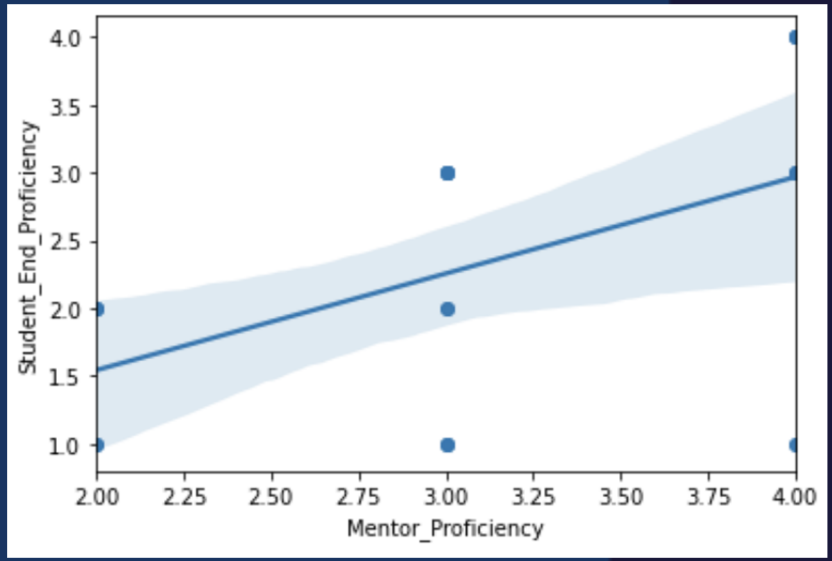
Linear Regression Line:

y = −1.5577 +
0.8930*Mentor_Proficiency +
1.1465*Student_Start_Proficiency

Training R^2 = 0.733

OSR^2 on testing data = 0.607

**Mentor_Proficiency v.s Student_End_Proficiency**




Overall the average mentee proficiency increased from 1.0 at the beginning of the program to 2.375 at the end of the program

Berkeley
UNIVERSITY OF CALIFORNIA

# Thank you!

Questions/Comments?

Berkeley
UNIVERSITY OF CALIFORNIA