Group 11

Harsha Satyavolu

Karthik Mahanth

Sairam Venkataraman

Vivek Narasimhamoorthy

**SQL-Mongo Project – IBM Employee Attrition**

BUAN 6320

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Activity** | **Harsha Satyavolu** | **Karthik Mahanth** | **Sairam**  **Venkataraman** | **Vivek Narasimha** |
| Prepared Data Model and Created Physical DB | X |  |  |  |
| Loaded Data into Database | X |  |  |  |
| Wrote SQL Queries |  |  |  |  |
| Prepared Mongo Database |  |  |  |  |
| Loaded data into Mongo DB |  |  |  |  |
| Wrote Mongo Queries |  |  |  |  |
| Prepared Report |  | X |  |  |
| Reviewed Report |  |  | X | X |

Contents

Relational Data Model 4

Assumptions/Notes About Data Entities and Relationships 4

Entity-Relationship Diagram 6

Physical MySQL Database 6

Assumptions/Notes About Data Set 6

Screen shot of Physical Database objects 7

Data in the Database 15

SQL Queries 17

SQL Query 1 17

Question 17

Notes/Comments About SQL Query and Results (Include # of Rows in Result) 17

Translation 17

Screen Shot of SQL Query and Results 17

SQL Query 2 18

Question 18

Notes/Comments About SQL Query and Results (Include # of Rows in Result) 18

Translation 18

Screen Shot of SQL Query and Results 18

SQL Query 3 19

Question 19

Notes/Comments About SQL Query and Results (Include # of Rows in Result) 19

Translation 19

Screen Shot of SQL Query and Results 19

SQL Query 4 20

Question 20

Notes/Comments About SQL Query and Results (Include # of Rows in Result) 20

Translation 20

Screen Shot of SQL Query and Results 20

SQL Query 5 21

Question 21

Notes/Comments About SQL Query and Results (Include # of Rows in Result) 21

Translation 21

Screen Shot of SQL Query and Results 21

SQL Query 6 22

Question 22

Notes/Comments About SQL Query and Results (Include # of Rows in Result) 22

Translation 22

Screen Shot of SQL Query and Results 22

Data Review for MongoDB 23

Assumptions/Notes About Data Collections, Attributes and Relationships between Collections 23

Physical Mongo Database 24

Assumptions/Notes About Data Set 24

Screen shot of Physical Database objects (Database, Collections and Attributes) 24

Data in the Database 24

MongoDB Queries/Code 25

Mongo Query 1 25

Question 25

Notes/Comments About MongoDB Query/Code and Results (Include # of Documents in Result) 25

Translation 25

Screen Shot of MongoDB Query/Code and Results 25

Mongo Query 2 26

Question 26

Notes/Comments About MongoDB Query/Code and Results (Include # of Documents in Result) 26

Translation 26

Screen Shot of MongoDB Query/Code and Results 26

Mongo Query 3 27

Question 27

Notes/Comments About MongoDB Query/Code and Results (Include # of Documents in Result) 27

Translation 27

Screen Shot of MongoDB Query/Code and Results 27

# Relational Data Model

## Assumptions/Notes About Data Entities and Relationships

In our Data Model, Employee Table acts as a primary entity that consists of all the primary details of an employee and the rest of the tables of this data model involve various aspects of employee such as the salary, performance, project. Below are some of the primary entities, relationships and assumptions.

**Key Entities and Relationships:**

1. Employee table is the primary entity of this relation with Employee Number as its primary key and it consists of all the attributes of an employee such as the employee Name, Gender, Age, Employment Type, Marital Status, Stock Option. It also consists of employee's company related details such as Total Work Experience, Current Experience in that Company, Years with current Manager, Business Travel and Attrition details. There are also additional attributes such as Distance from Home, Number of Companies worked, Over Time details. Apart from the dataset columns, we have included Employment Type in our data Model which tells about whether an employee is a Full-Time or Contract employee.
2. There are various foreign keys in the Employee Table such as Department ID, Manager ID, Project ID, Job Role ID, Job Level ID, Education Level ID and Education Field ID. This implies that there are relationships between Employee Table and many of the above foreign Key tables.
3. Department Table is uniquely identified by Department ID and its other columns are Department Name and the Dept Location. Each employee belongs to only one department and a department can have many employees which determines a one-to-many relationship between Department and an Employee.
4. Manager entity consists of the Manager ID as the primary identifier, Manager name and the project he manages. Each employee has only one manager and many employees work under a manager. Therefore, there exists a one-to-many relationship between manager and Employee.
5. Project entity consists of the Project ID as the primary key, project Location and the Manager of that Project. Each Project is managed by a manager and a manager manages only one project as per our assumption. Each Project has many employees and an employee works for only one project. There exists a one to one relationship between Project and Manager tables, one-to-many relationship with Project and Employee.
6. Job Level consists of attributes such as Job Level ID and Job Level Description. Job Role table consists of a Job Role ID and its description. Each employee belongs to only one Job Level and Job Role and both tables can consist of many employees. Therefore, there exists a one-to-many relationship between both Job Level and Job Role with employees.
7. There are two tables regarding the education Level and Education fields. Education Level table consists of the Education Level ID and its description. Education Field consists of the Field ID and the different types of Educational fields.
8. Rating Levels entity consists if the Rating ID as its primary key and various rating Levels. Emp Performance has an evaluation ID as the primary key, date of review, Percent Hike, Job Involvement as the other columns and Rating ID acts as foreign key in this table.
9. Each employee’s performance is identified by the evaluation ID and each Evaluation ID belongs to only one employee. Therefore, there exists a one to one relation between the Employee table and the Emp Performance table.
10. There are 4 different Employee Satisfaction tables such as the Relationship Satisfaction, Work Life Balance, Job satisfaction and the Environment Satisfaction. Each table consists of an ID and its own categories of the Satisfaction Levels.
11. Employee Feedback table is uniquely identified by Feedback ID and all the other columns are the foreign keys from Satisfaction Levels tables and Employee table. Each Satisfaction Level can be chosen by any number of employees and an employee can choose only one satisfaction level of one type. Therefore, there exists many to one relationship between the Employee Feedback table and all the satisfaction tables such as the Relationship Satisfaction, Work Life Balance, Job satisfaction and the Environment Satisfaction.

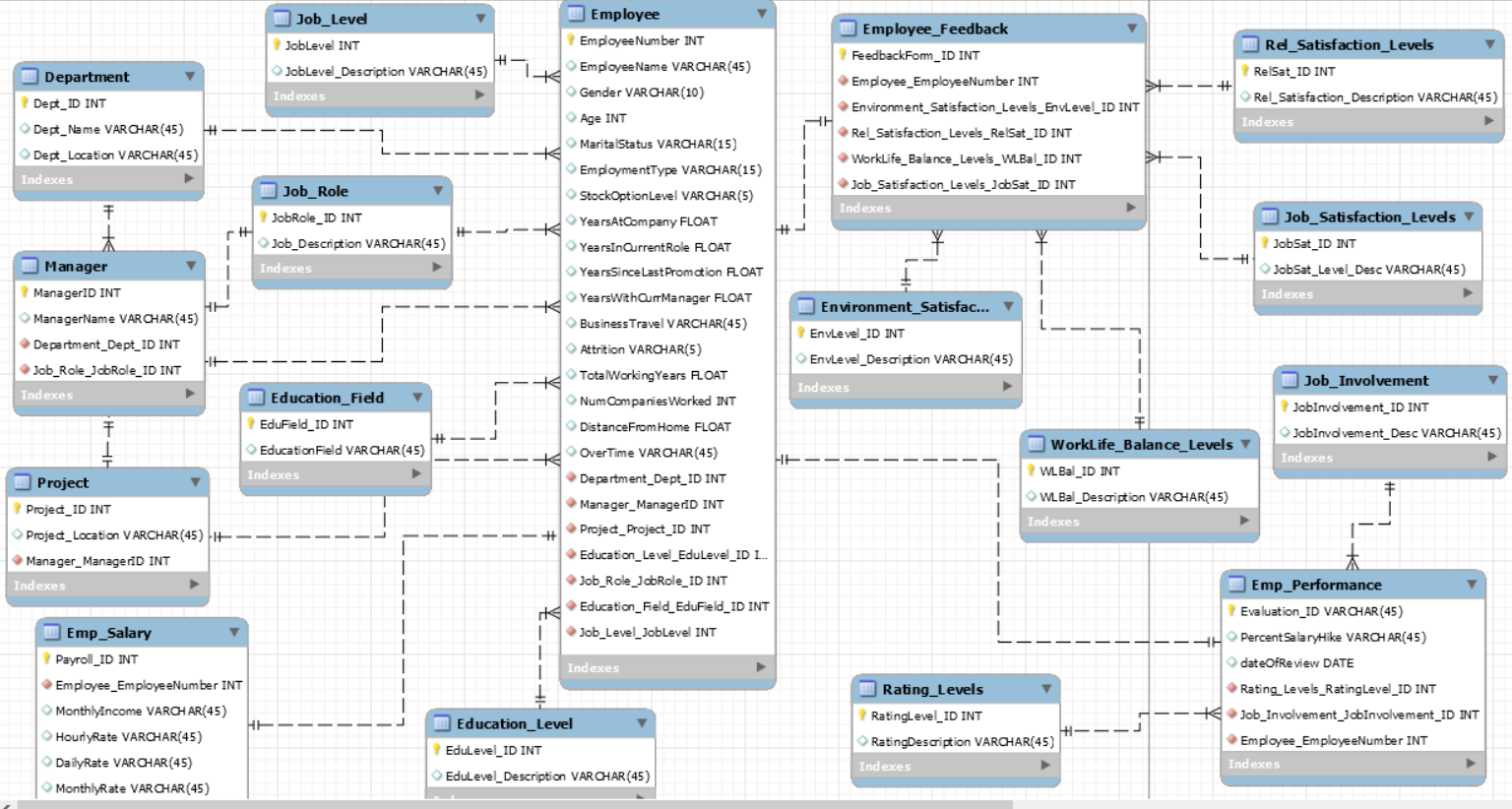
**Assumptions of our Data Model:**

1. We have added one more extra column ‘Employment Type’ in the Employee table apart from the dataset. It describes whether an employee is a Full-Time or a contract. Assuming that all the employees in the given dataset as Full-Time employees.
2. Assuming that HourlyRate, Daily Rate and MonthlyRate are for Contract and Monthly Income is for full time employees. Company offers both full time and contract roles for employees and employees has the right to choose between both.
3. Each manager doesn’t have any specific manager ID, it’s the same as the employeeID.
4. Assuming that Manager Role is the highest of all the job Roles.
5. Each Manager manages only one project and a project has only one manager as per our assumptions.
6. Manager has a fixed Job Role ID 5006. Therefore, there exists a one to one relationship between Job Role Entity and Manager entity.
7. Assuming that employee belongs to only one education Level. Although he has multiple degrees, the highest qualification is considered to be his educational Level. There exists a one-to-many relationship between Education Level and Employee.
8. Assuming that each employee has a project and it is a one-to-many relation between Project and Employee.
9. Although an Employee has studied multiple degrees in various Educational Fields, the latest one is considered.
10. Assuming that Employee Feedback is populated by Employee and there exists a one-to-one relationship between Employee Feedback and Employee.
11. Assuming that employee’s performance throughout the year is evaluated by Evaluation ID and one employee is tracked through only one Evaluation ID.
12. Assuming that Employee fills the feedback Survey form only once in a company and is tracked by a Feedback Form ID.

**Normalization Form:**

1. We have created every table with a unique primary key as mentioned below in the ER diagram.
2. Every attribute defined in the table is unique and has no multipart-values, thus exhibiting the first normal form.
3. Every non-prime attribute in every table is fully dependent on the primary key and not dependent on any of the non-key columns which satisfies the 2NF and 3NF normal forms.
4. The partial dependency with every value is taken care of through dividing the data into multiple tables.
5. All keys in all tables are related with only appropriate foreign keys and relations are specified in a meaningful manner.

## Entity-Relationship Diagram



## Physical MySQL Database

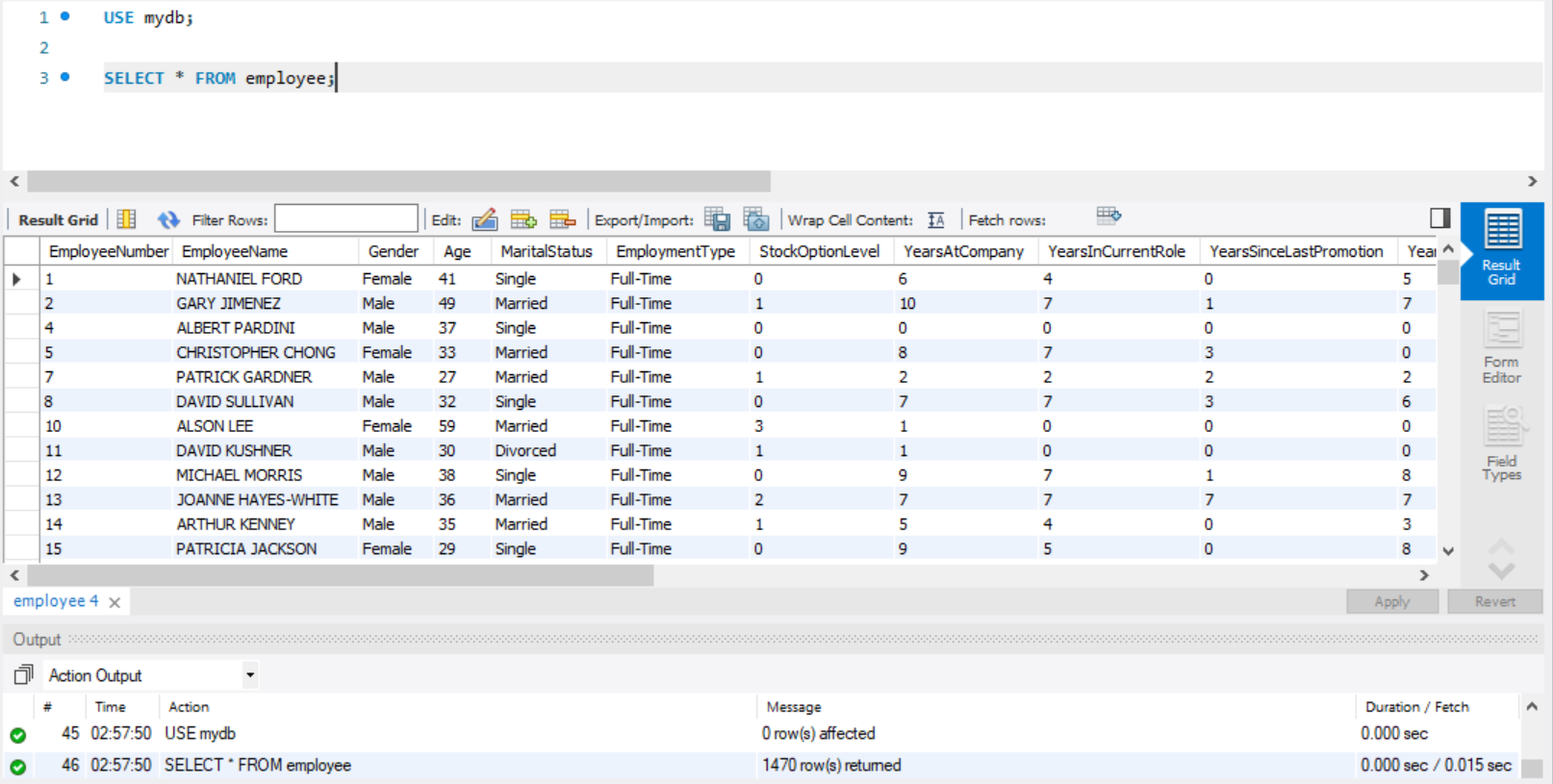
## Assumptions/Notes About Data Set

Include any assumptions made about data such as empty fields, sparse data, bad data, etc.

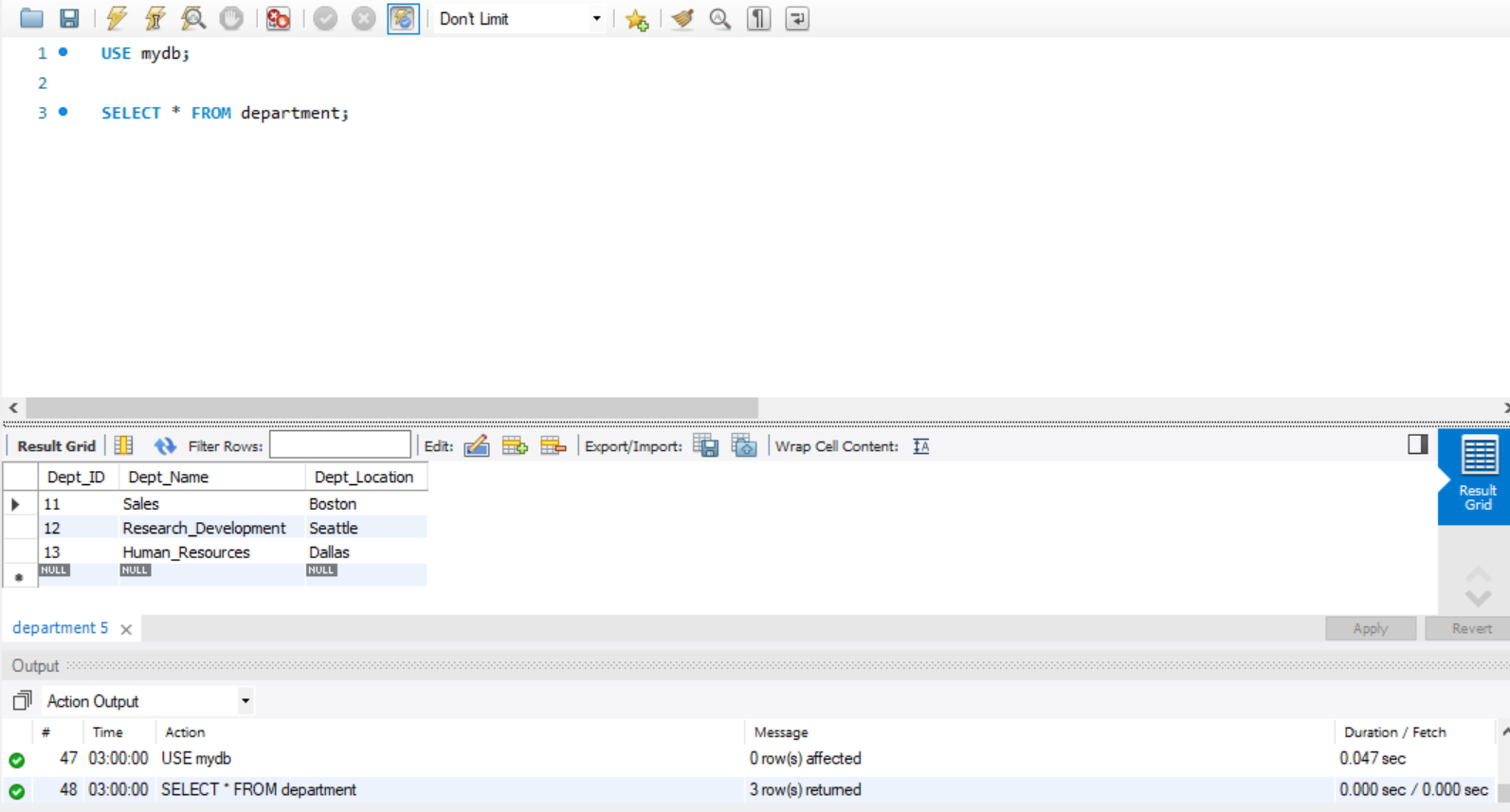
1. We assumed that Employee Number uniquely determines an employee and it is used as a reference in another table to retrieve the respective data of an employee.
2. We have removed some of the columns such as Employee Count, Over18 and Standard Hours. For these three columns, the value is constant across all the rows in the dataset and cannot be used retrieve any type of insights.
3. There are no duplicate columns or missing data in the dataset.
4. We have split the given data into different tables and established dependencies among them.

## Screen shot of Physical Database objects

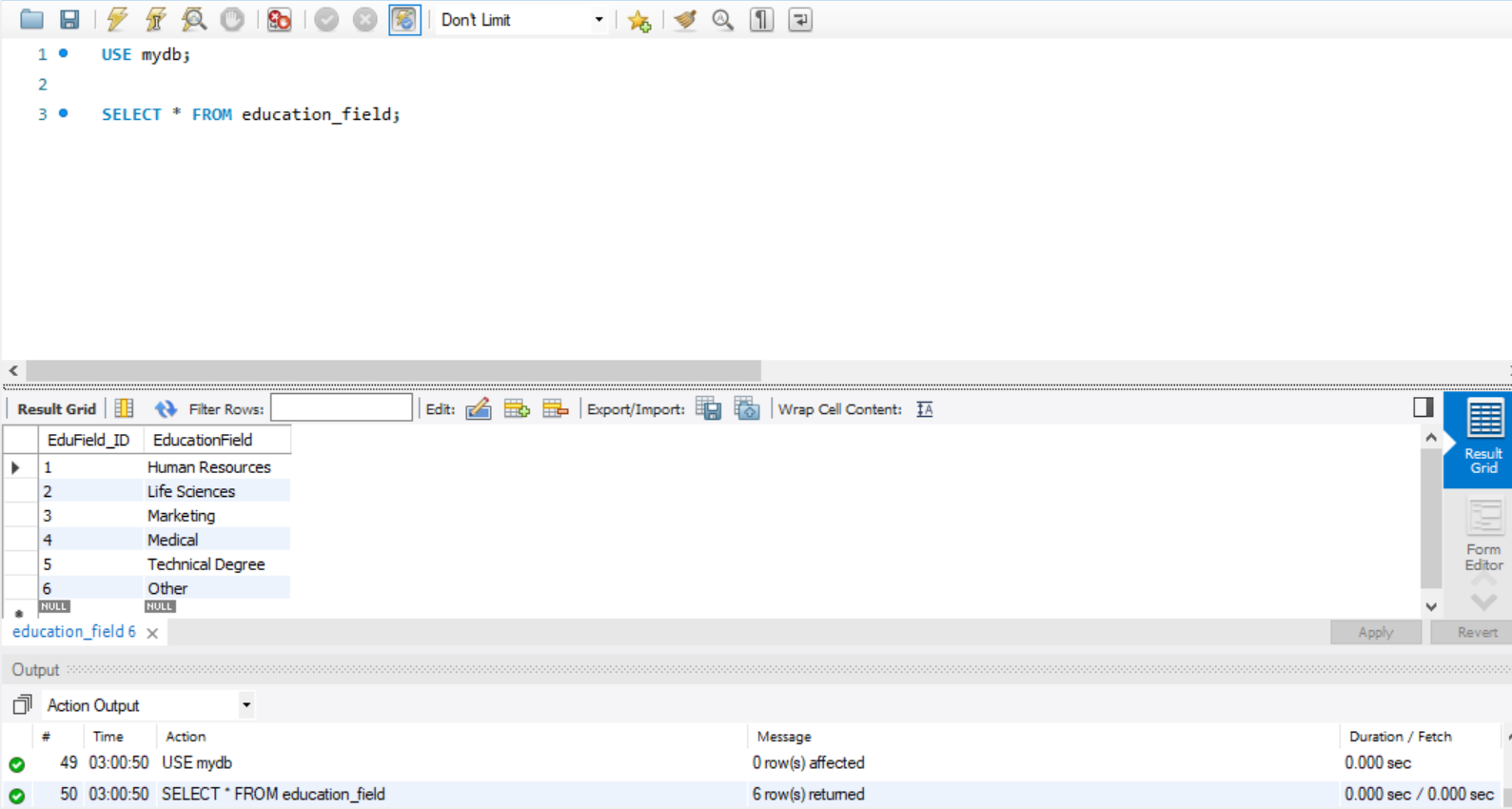
Employee Table:



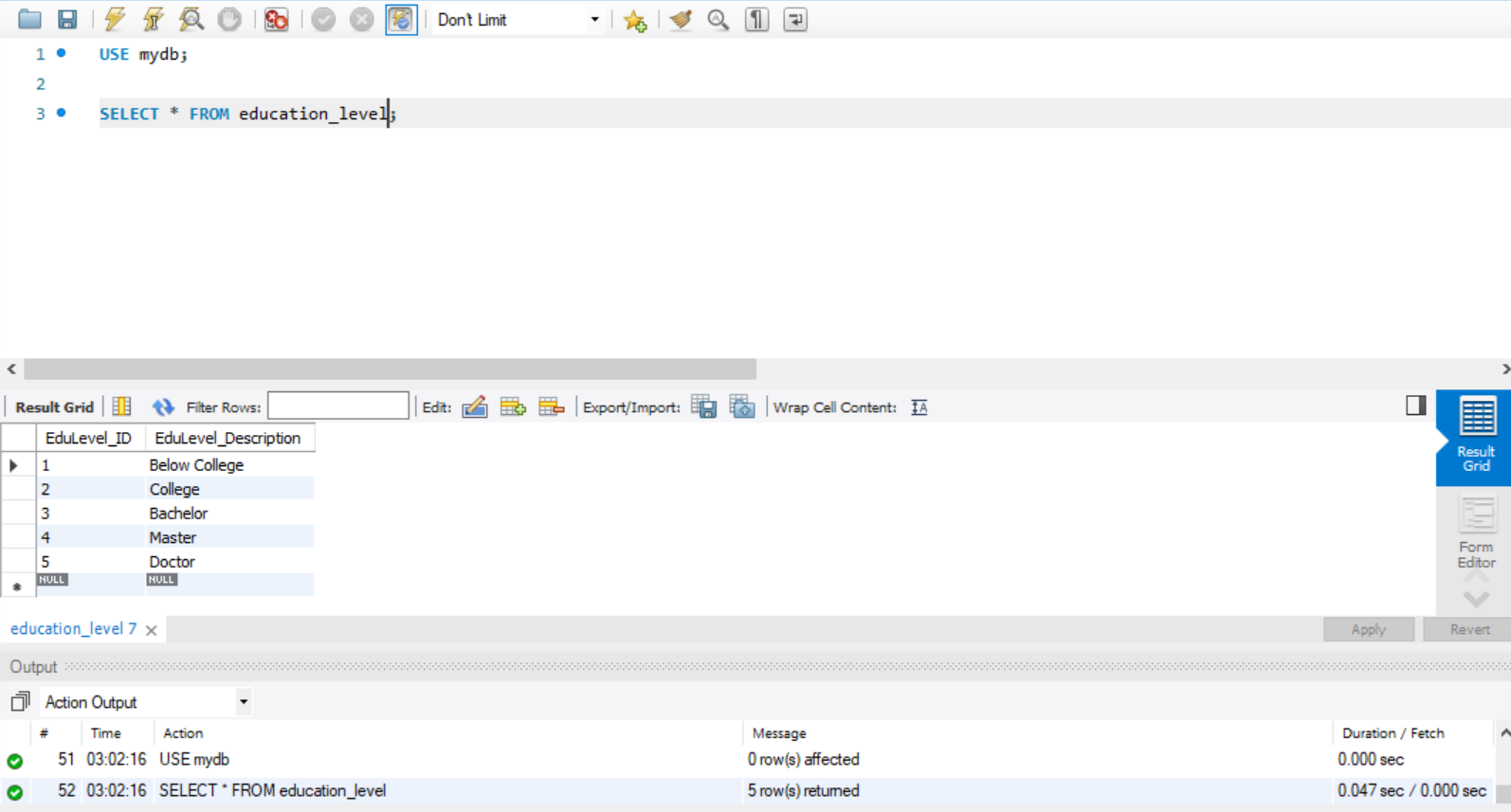
Department Table:



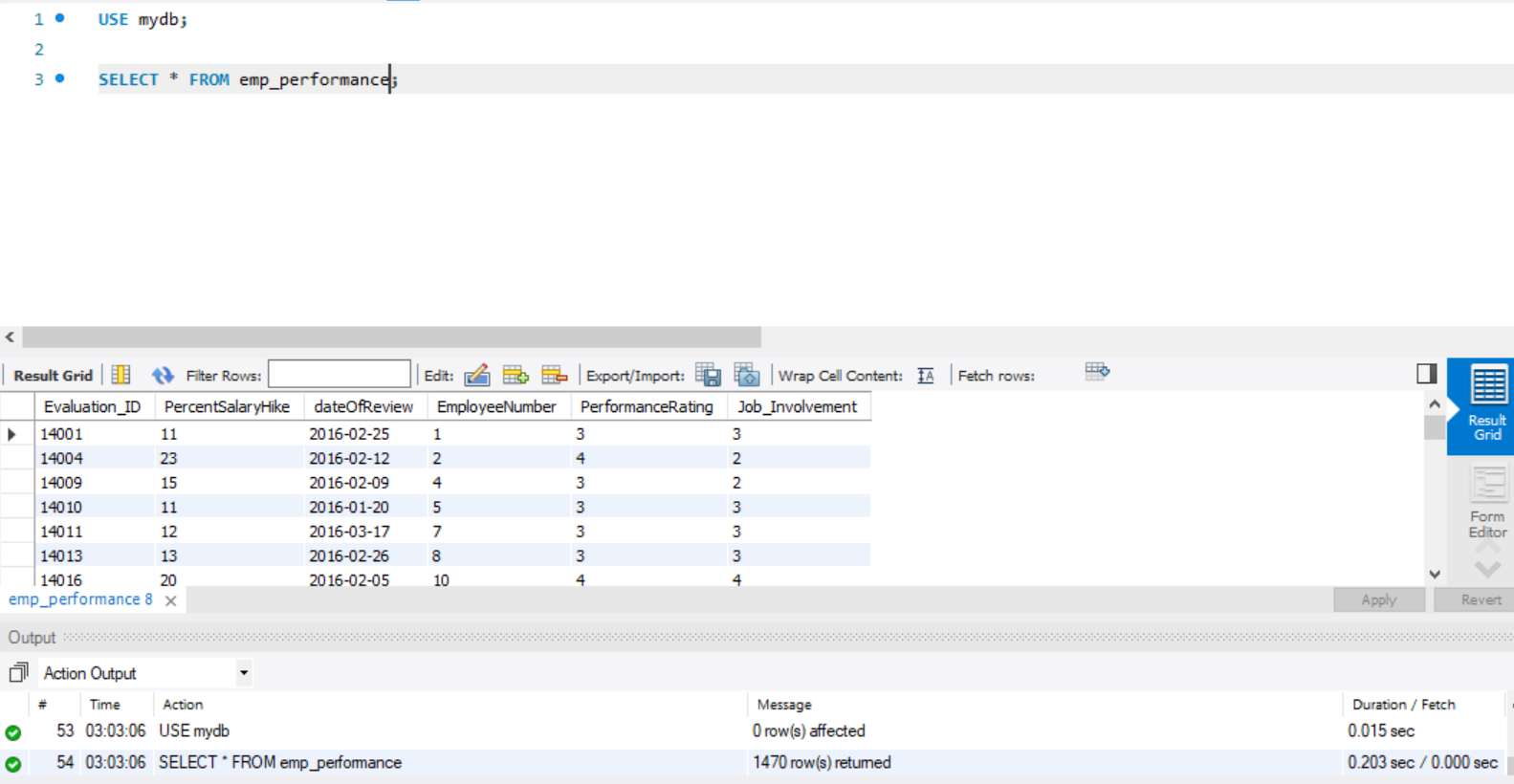
Education\_Field Table:



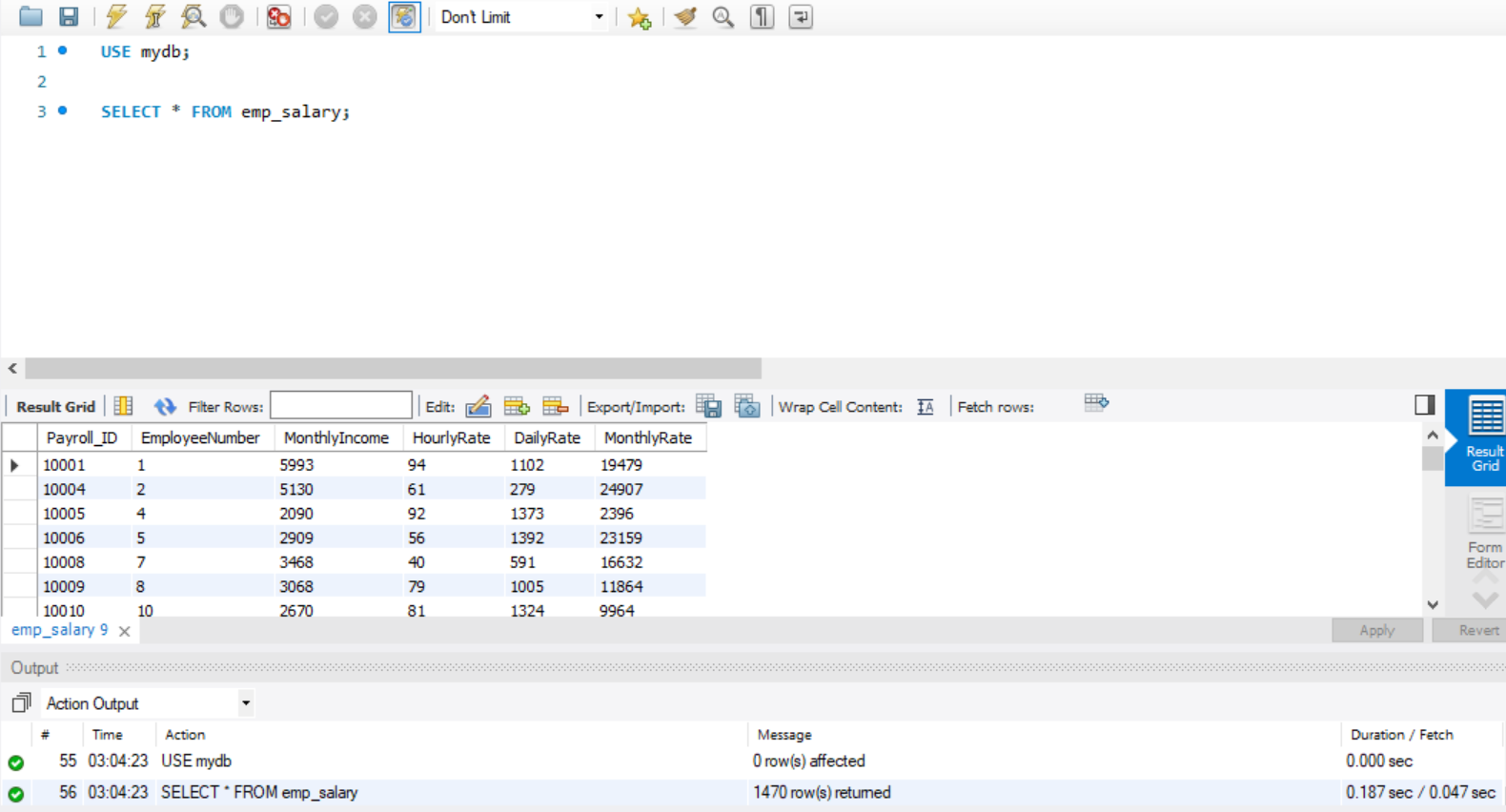
Education\_Level table:



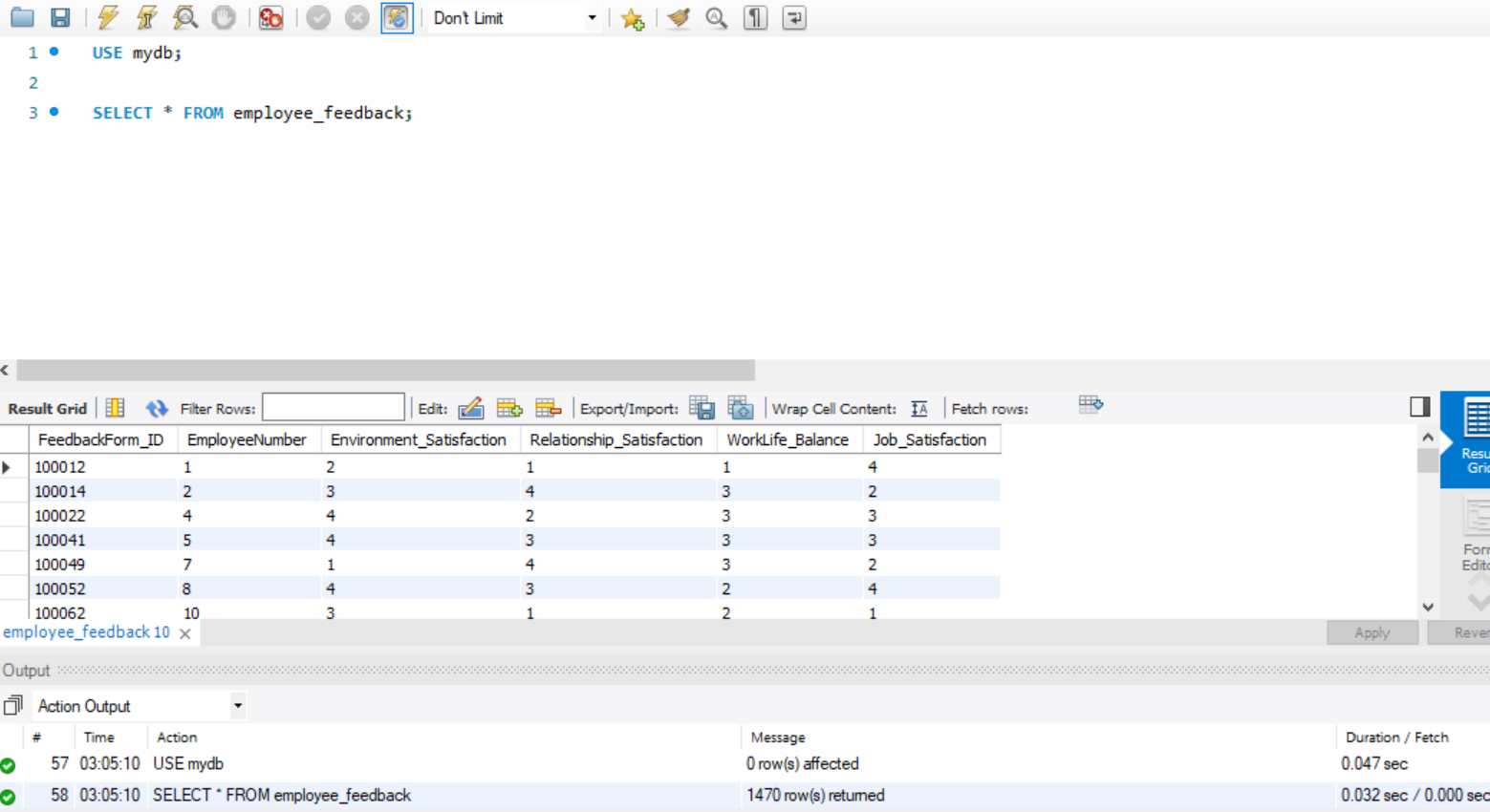
Emp\_Performance Table:



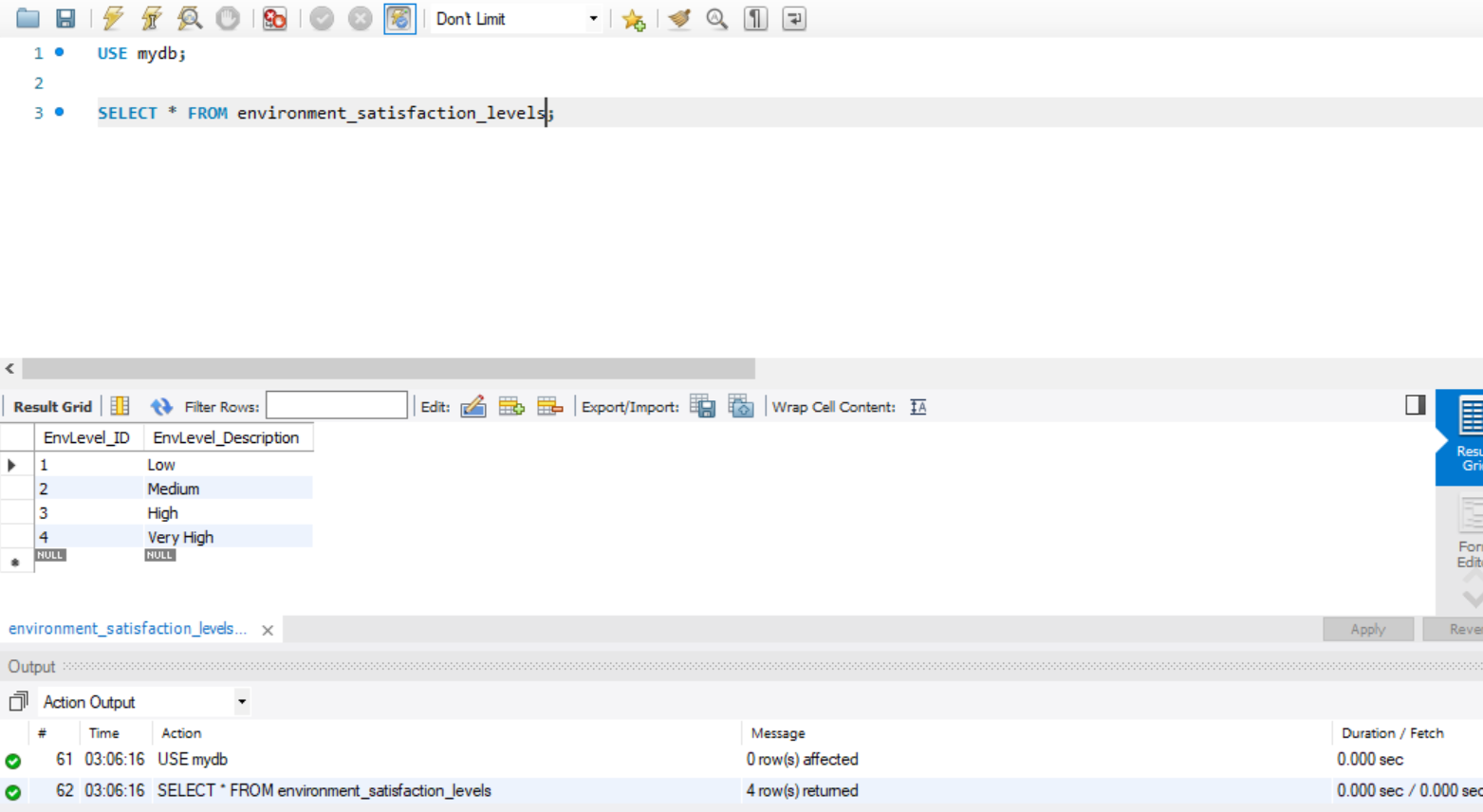
Emp\_salary Table:



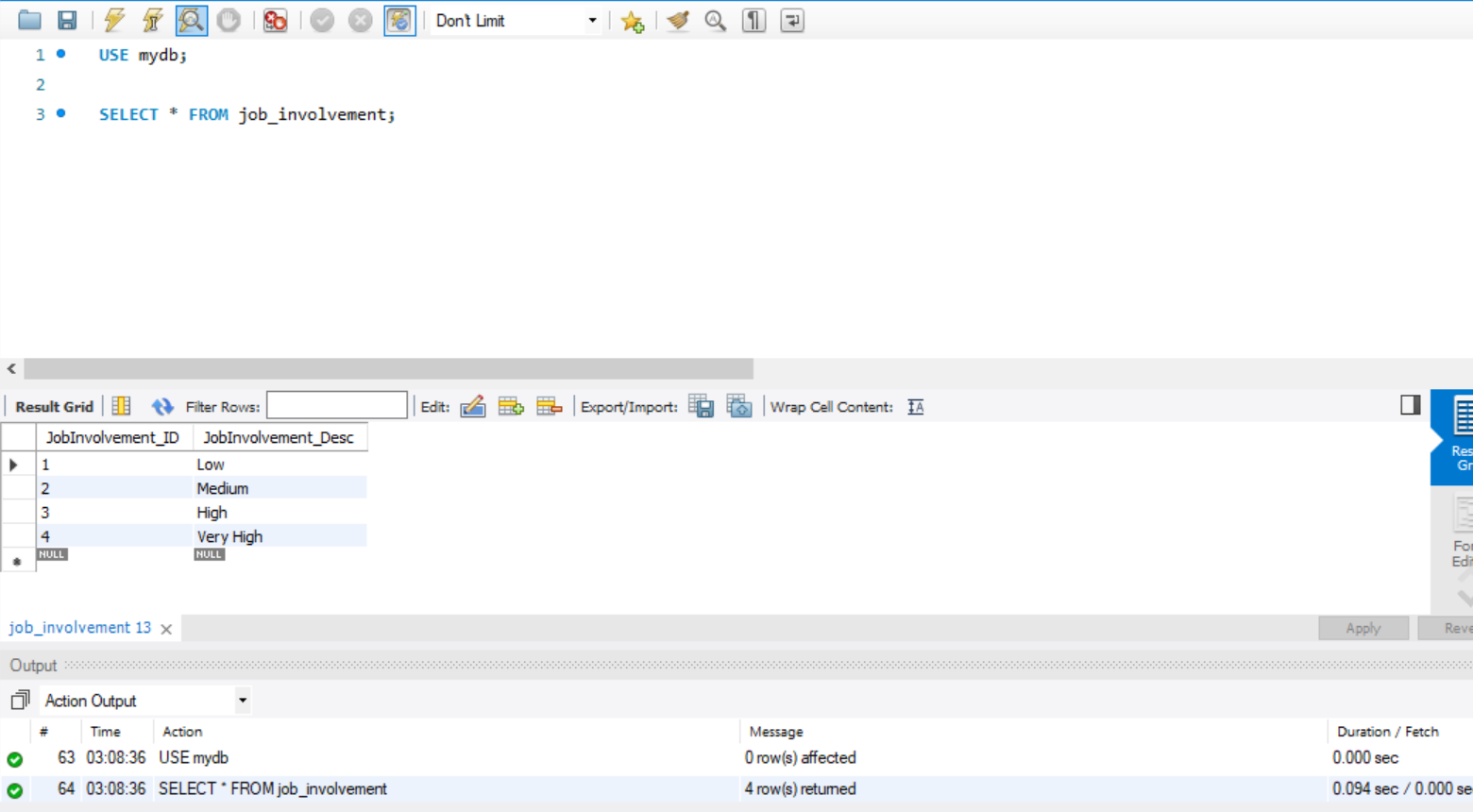
Employee\_Feedback Table:



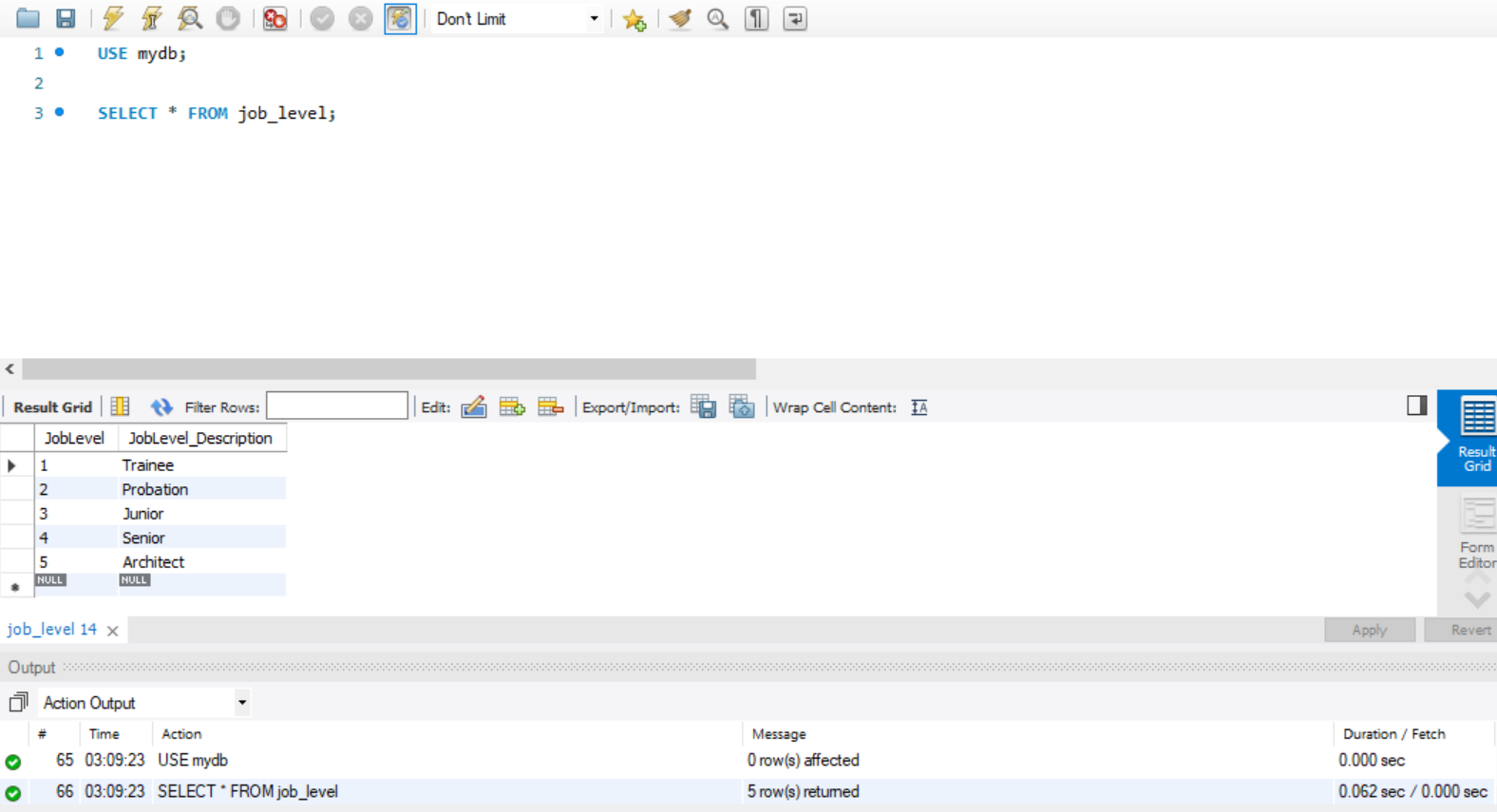
Environment\_Satisfaction\_Levels Table:



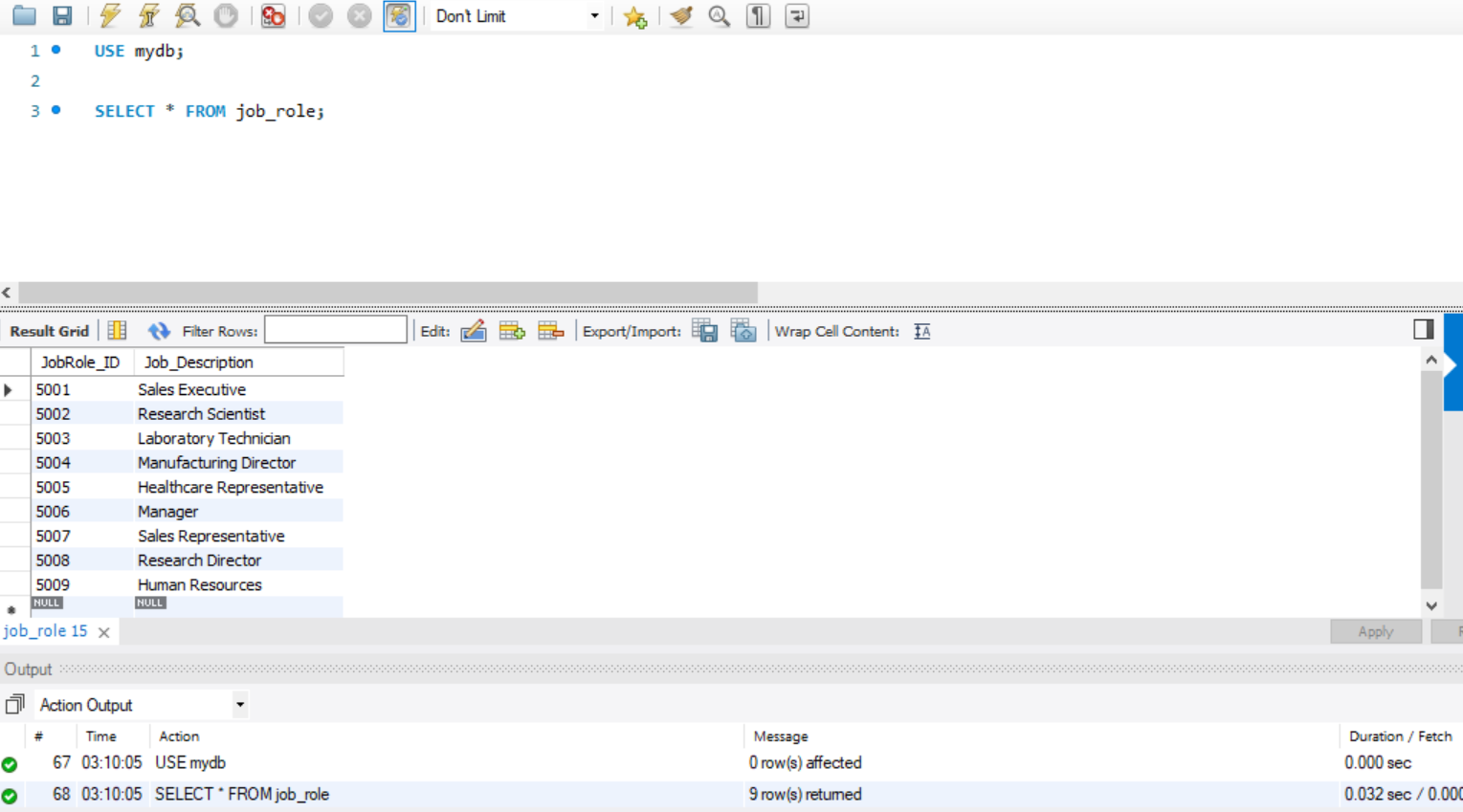
Job\_involvement Table:



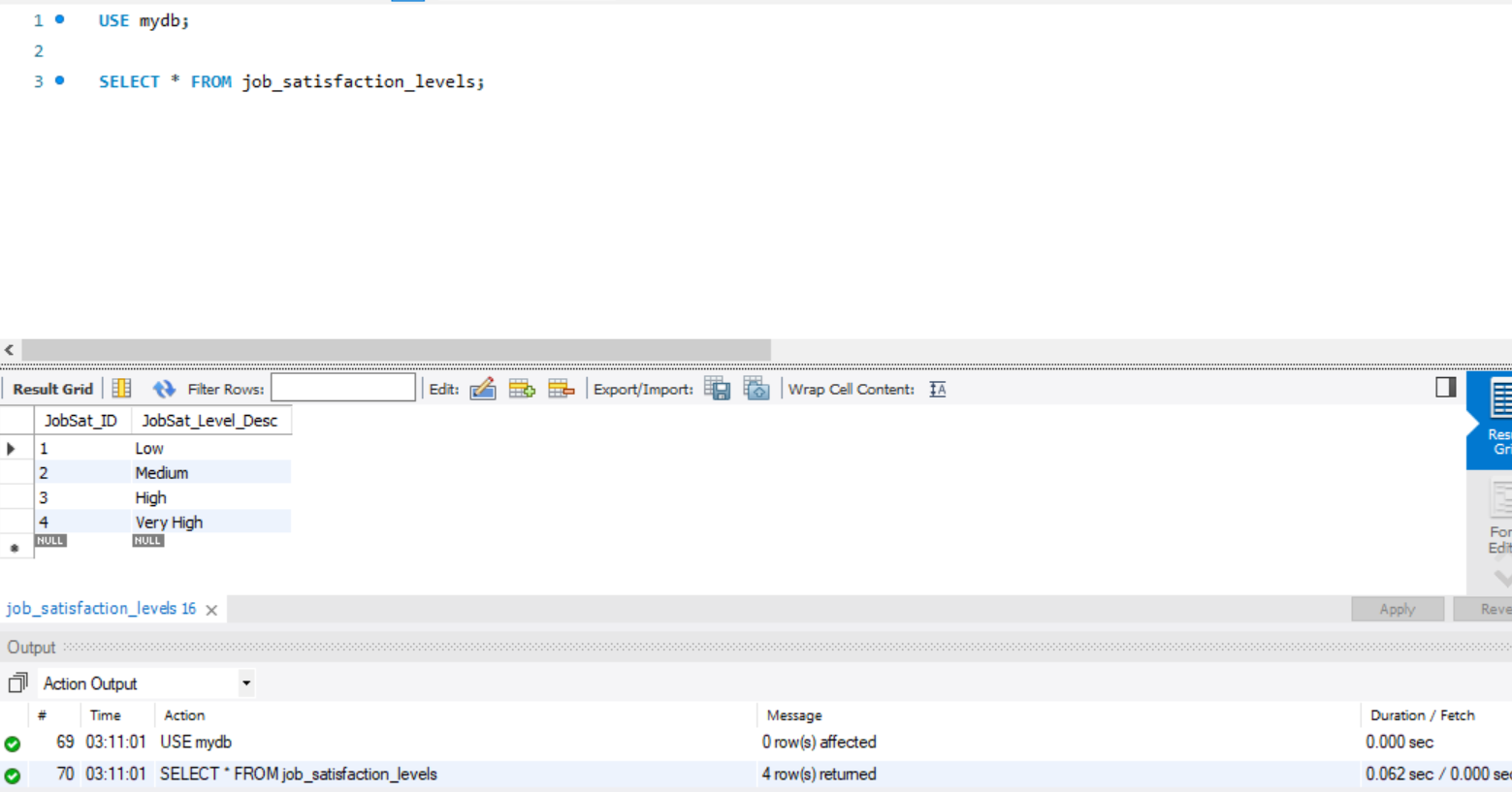
Job\_Level Table:



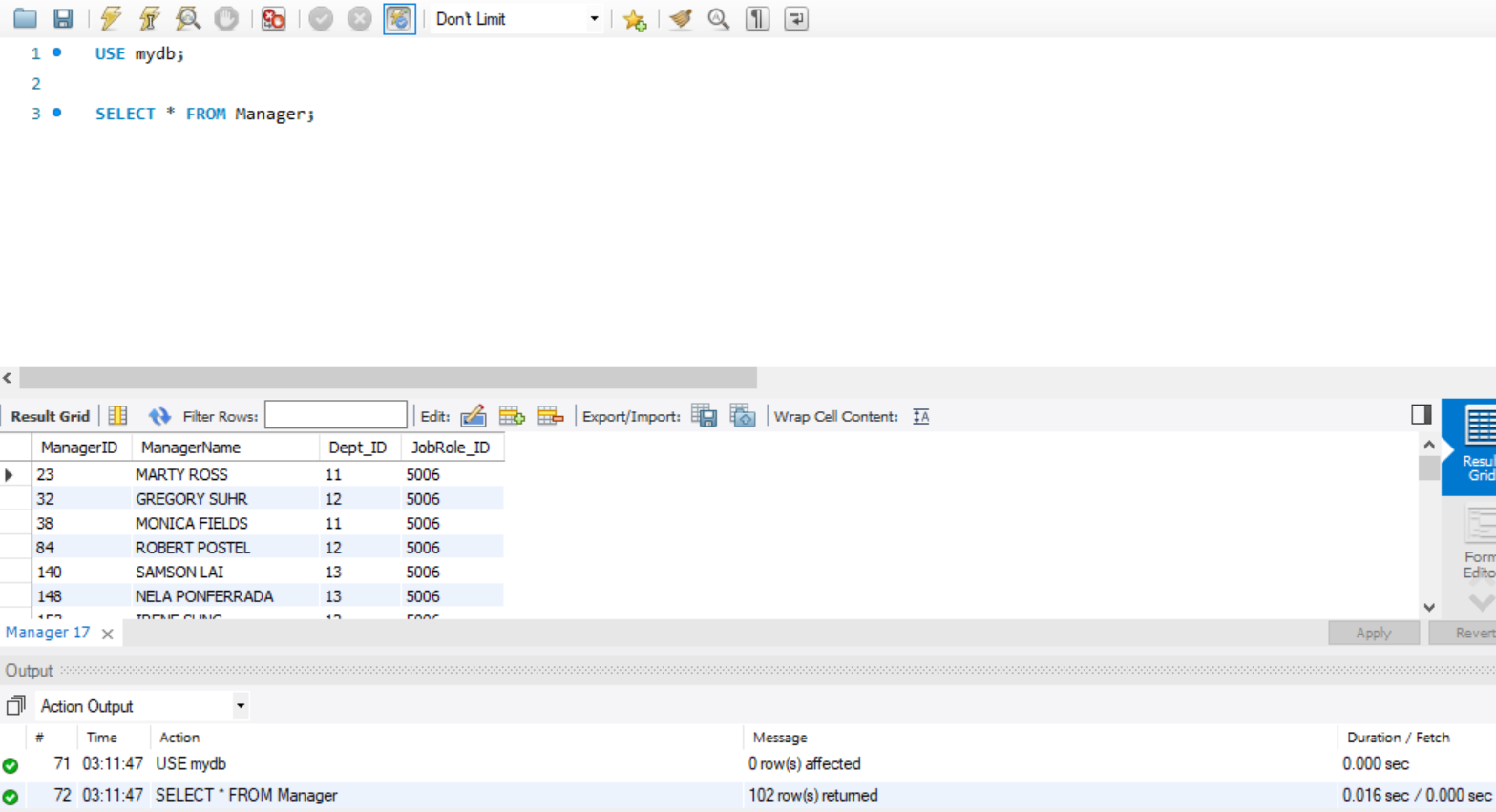
Job\_Role:



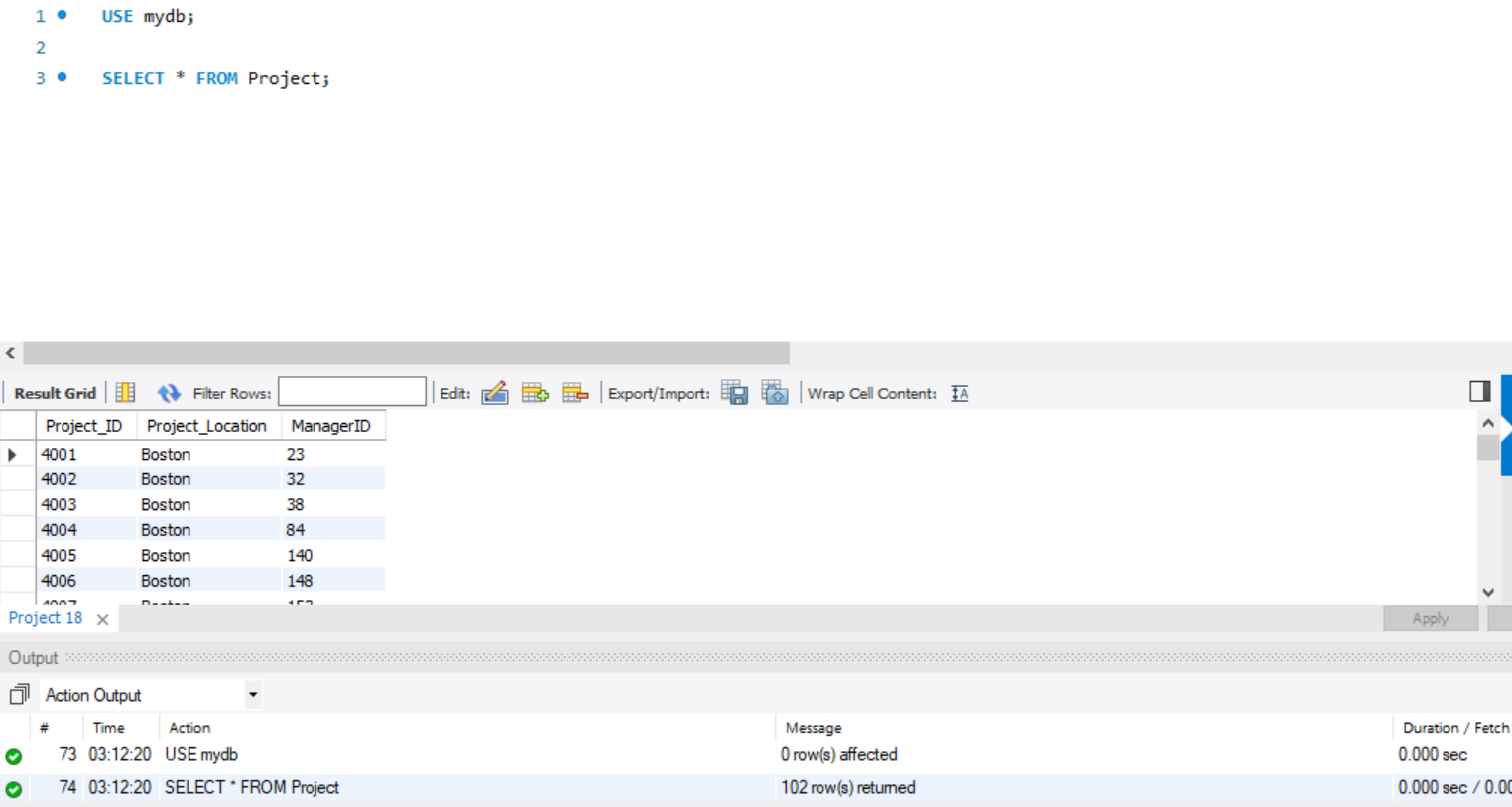
Job\_Satisfaction\_Levels Table:



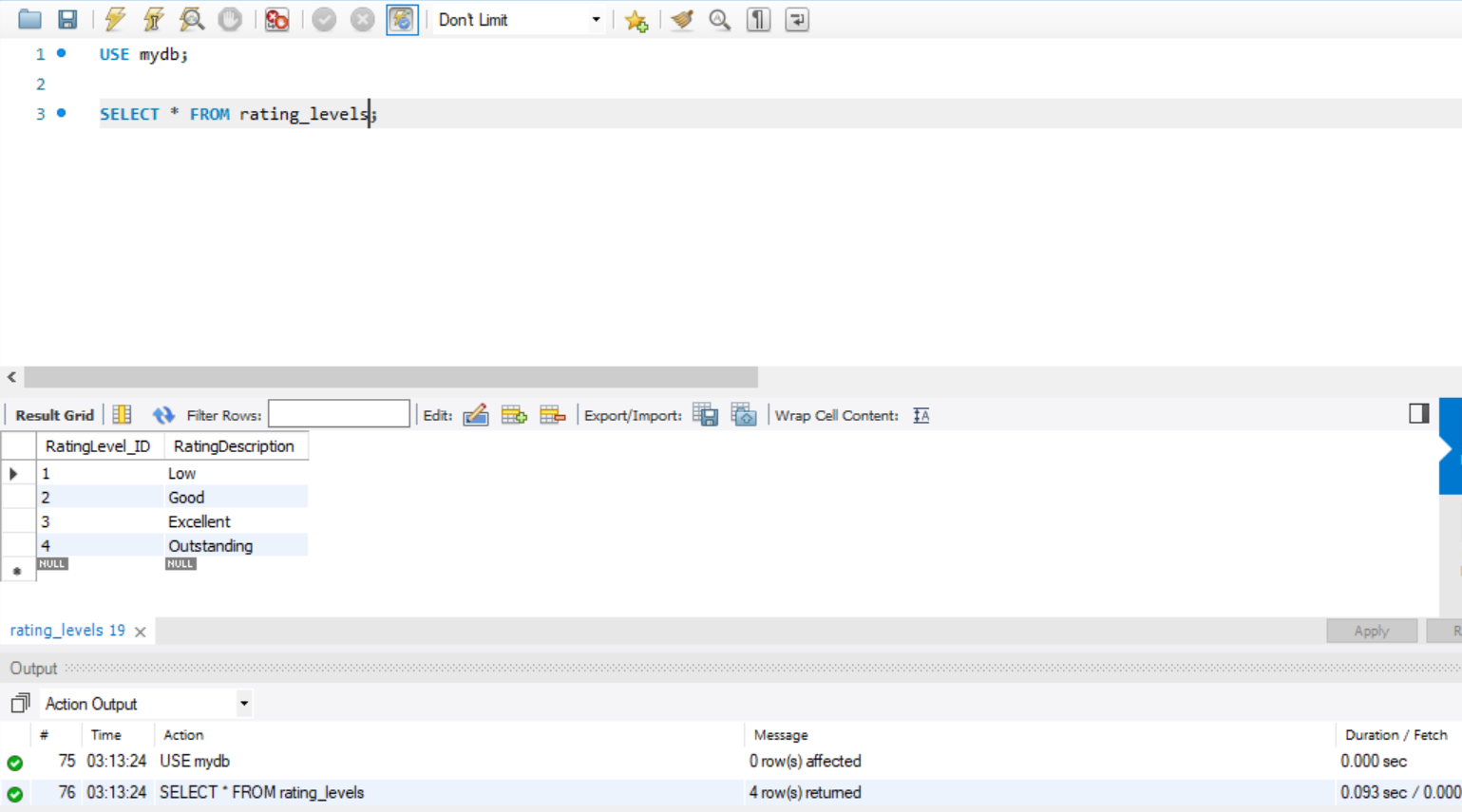
Manager Table:



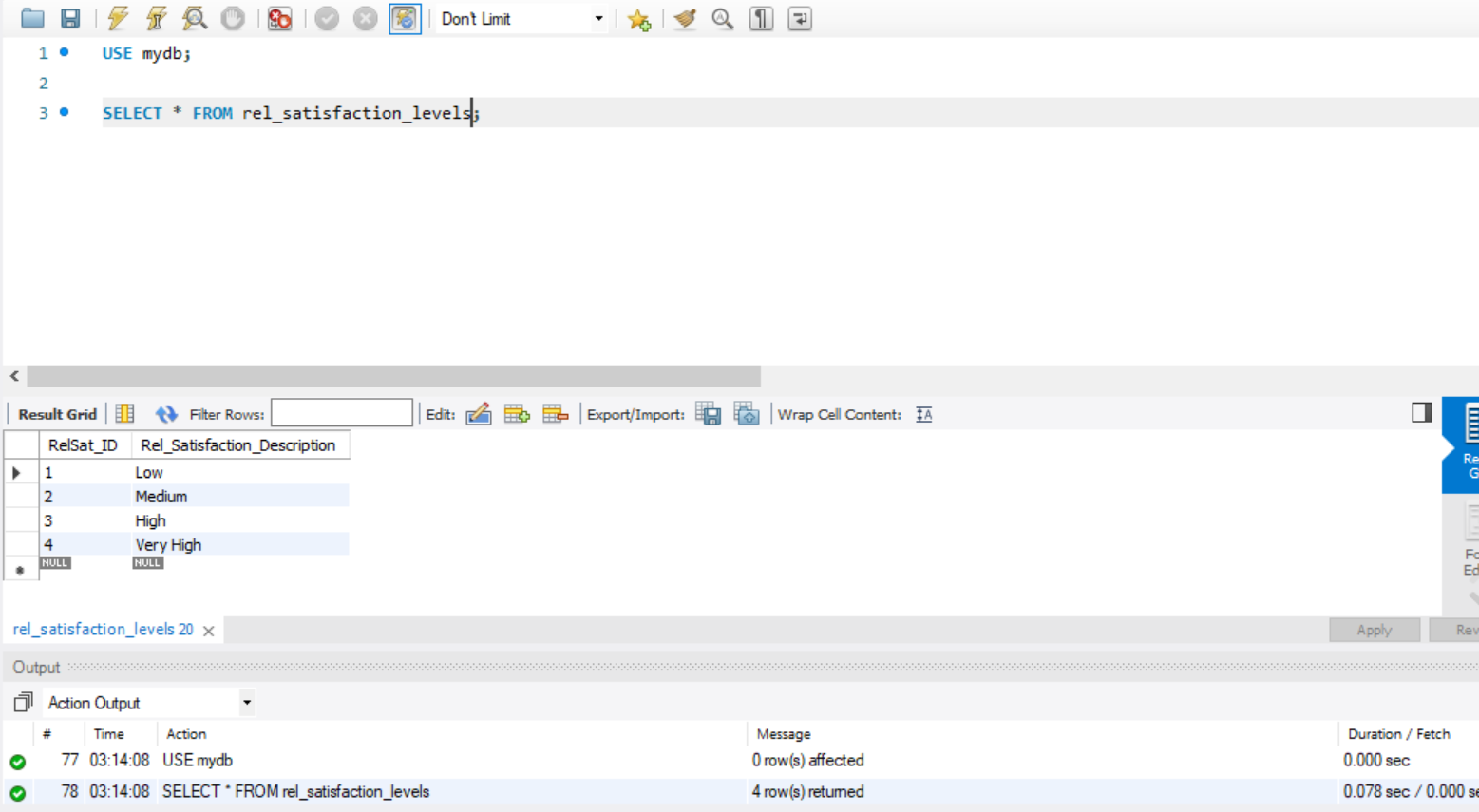
Project Table:



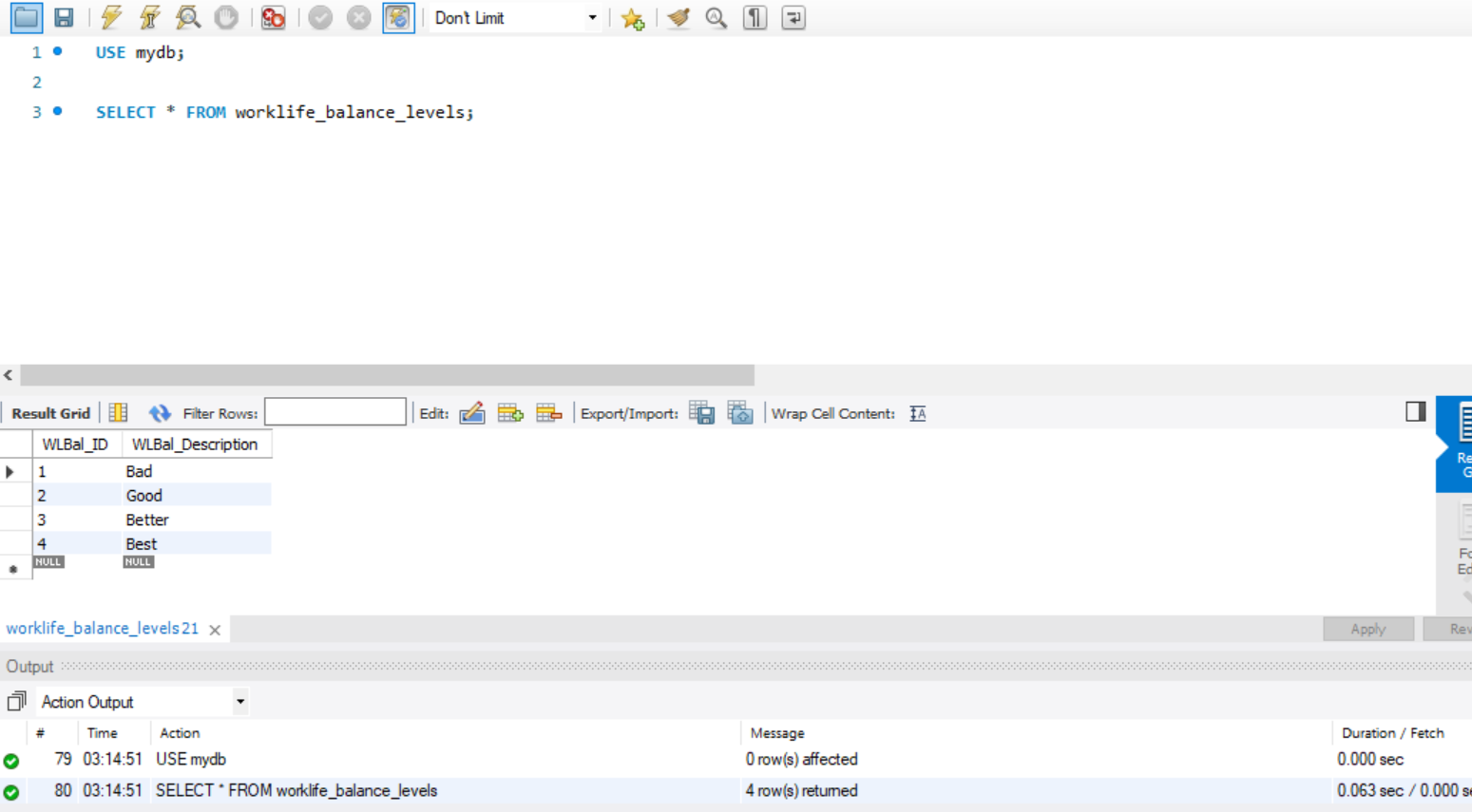
Rating\_Levels Table:



Rel\_Satisfaction\_Levels Table:



WorkLife\_Balance\_Levels Table:



## 

## Data in the Database

|  |  |  |  |
| --- | --- | --- | --- |
| **Table Name** | **Primary Key** | **Foreign Key** | **# of Rows in Table** |
| Employee | EmployeeNumber | Dept\_ID,  ManagerID,  Project\_ID,  EduLevel\_ID,  JobRole\_ID,  JobLevel\_ID,  EduField\_ID | 1470 |
| Department | Dept\_ID |  | 3 |
| Manager | ManagerID | Dept\_ID,  JobRole\_ID | 102 |
| Job\_Level | JobLevel\_ID |  | 5 |
| Job\_Role | JobRole\_ID |  | 9 |
| Project | Project\_ID | ManagerID | 102 |
| Education\_Level | EduLevel\_ID |  | 5 |
| Education\_Field | EduField\_ID |  | 6 |
| Job\_Involvement | JobInvolvement\_ID |  | 4 |
| Rel\_Satisfaction\_Levels | RelSat\_ID |  | 4 |
| WorkLife\_Balance\_Levels | WLBal\_ID |  | 4 |
| Emp\_Performance | Evaluation\_ID | EmployeeNumber,  RatingLevel\_ID,  JobInvolvement\_ID | 1470 |
| Job\_Satisfaction\_Levels | JobSat\_ID |  | 4 |
| Environment\_Satisfaction\_Levels | EnvLevel\_ID |  | 4 |
| Rating\_Levels | RatingLevel\_ID |  | 4 |
| Emp\_Salary | Payroll\_ID | EmployeeNumber | 1470 |
| Emp\_feedback | FeedbackForm\_ID | EmployeeNumber,  JobSat\_ID,  EnvLevel\_ID,  WLBal\_ID,  RelSat\_ID, | 1470 |

# SQL Queries

## SQL Query 1

### Question

If the company wants to cut travel costs, which department should the company focus on?

### Notes/Comments About SQL Query and Results (Include # of Rows in Result)

* Assuming that the company wants to cut the travel costs due to some economy crisis, we should focus more on the people who travel frequently. This incurs heavy charges on the company due to travelling, accommodation and other expenses.
* We thought Count function here could be useful for us in determining the number of people who travel frequently and we hence calculated the count of employees grouped by department who travel frequently.
* By results, we can say that, the highest number of employees who travel are from Research & Development while that of Sales and HR are 84 and 11 respectively. By this, we can say that the company can focus on the Research & Development department for cutting the travel costs.
* Results: There are 3 records in the result set which shows us the Business Travel and the count of Employees Department wise.

### Translation

Translation🡪 Select Department Name, BusinessTravel and Employee Count from Department Table inner joined with Employee table on Department ID of Department table matched with Department ID of Employee Table where Business Travel is Travel Frequently grouped by Department Name.

Query🡪

### use mydb;

### Select D.Dept\_Name, E.BusinessTravel, count(E.BusinessTravel) as EmployeeCount from Department D

### join Employee E on E.Dept\_ID = D.Dept\_ID

### where E.BusinessTravel = 'Travel\_Frequently'

### group by D.Dept\_name;

### Screen Shot of SQL Query and Results

## 

## SQL Query 2

### Question

The company has been paying gas expenses for miles traveled by employees between their home and work. If they want to increase the per mile compensation, which department’s employees will gain the most?

### Notes/Comments About SQL Query and Results (Include # of Rows in Result)

* This query is all about to check which department’s employee will gain the most when the company increases the per mile compensation.
* We assumed that Sum aggregate function could give us the best result for the employees who commute the most and hence calculated the sum of the employee’s commute distance for each of the department and arranged them in descending order to know which department’s employees commute the most.
* We have found that Research & Development’s Employees commute the most distance which is 8887, while Sales and HR takes the second and third places with 4177 and 548 units respectively.
* By the results, we can interpret that if the company wants to increase the per mile compensation, Research & Development Department’s employees gain the most since they travel more than the others.
* Result: The result 3 records with Total Distance Travelled by the employees everyday from home to work divided by the Departments.

### Translation

Translation🡪 Select the Department Name and sum of the DistanceFromHome column from Department table inner joined with the Employee table on Department ID in the Department table matching Department ID in the Employee table grouped by the Department ID and ordered by DistanceFromHome.

Query🡪

use mydb;

Select D.Dept\_Name, sum(E.DistanceFromHome) AS Total\_DistanceFromHome

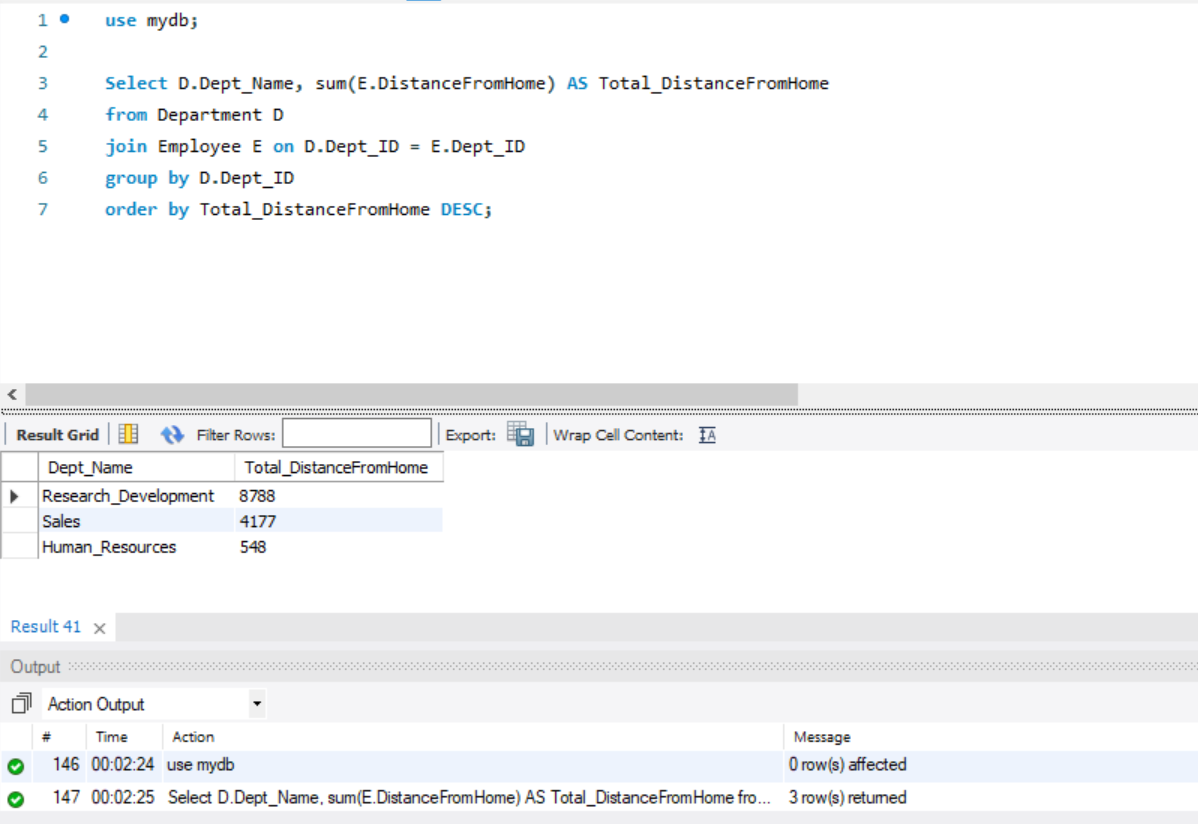
from Department D

join Employee E on D.Dept\_ID = E.Dept\_ID

group by D.Dept\_ID

order by Total\_DistanceFromHome DESC;

### Screen Shot of SQL Query and Results



## SQL Query-3

### Question

A new employee from a Life Sciences education field wants to work in Sales. Do you believe the company might be able to give him a chance to work in Sales? Why or Why not?

### Notes/Comments About SQL Query and Results (Include # of Rows in Result)

* This query is about whether a new employee from a Life Sciences education field can work in Sales and if company might be able to give a chance to work.
* We checked if there are any employees previously related to Life Sciences field working in Sales department and we also checked the performance rating of those employees if they are performing well.
* There are 150 employees at present working in Sales department with the same Medical Education background and the average performance rating of those employees is 3.12
* We believe that the company might be able to give him a chance to work in sales since there are already existing employees working in Sales with the same background and also the average performance rating is almost above ‘Excellent’ criteria.
* Result: There is 1 record displaying the number of employees working in the Sales department from Life Sciences background and also their average performance rating for the year.

### Translation

Translation🡪 Select Department Name, EducationField, Employee count, average Performance rating from Department inner joined with Employee table on matching Department ID of Department Table with Department ID of Employee table, again inner joined with education\_field table on Education Field ID in education\_field table matching with Education field ID of Employee table, inner joined again with emp\_performance table on Employee Number matching with emp\_performance and employee table where the Education Field is Life Sciences and Department Name is Sales.

Query🡪

use mydb;

select D.Dept\_Name, EF.EducationField, count(E.EmployeeNumber) as EmployeeCount,

avg(EP.PerformanceRating) as AvgPerformanceRating

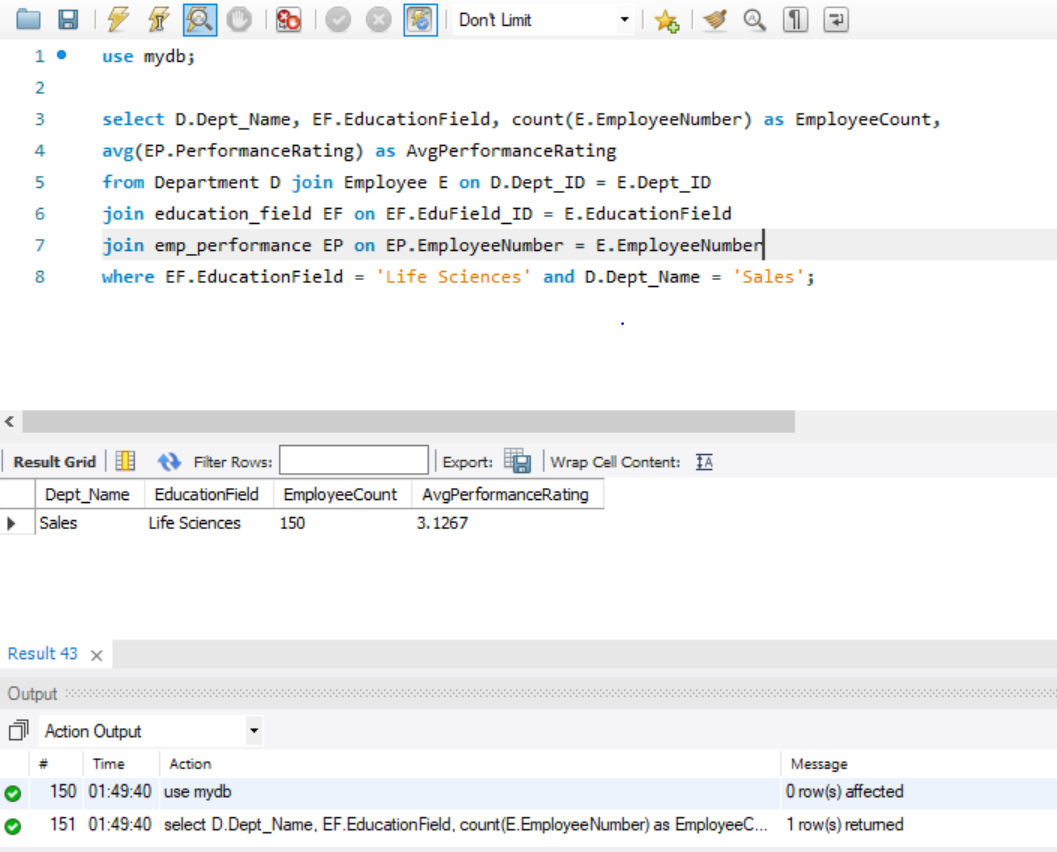
from Department D join Employee E on D.Dept\_ID = E.Dept\_ID

join education\_field EF on EF.EduField\_ID = E.EducationField

join emp\_performance EP on EP.EmployeeNumber = E.EmployeeNumber

where EF.EducationField = 'Life Sciences' and D.Dept\_Name = 'Sales';

### Screen Shot of SQL Query and Results



### SQL Query 4

### Question

HR feels that their environment satisfaction score is higher than Sales but HR job satisfaction score is lower than Research & Development. Are they right?

### Notes/Comments About SQL Query and Results (Include # of Rows in Result)

* To check which department has higher environment satisfaction in between HR and Sales, which department has higher job satisfaction in between HR, Research & Development.
* We assumed that Average function suits well for comparing the scores between two departments calculated the average rating given by the current employees on environment satisfaction and job satisfaction
* By the query results, HR has Environment Satisfaction score on an average of 2.6825 which is slightly higher compared to that of Sales 2.6794. This means HR felt right that its environment score is higher than that of Sales.
* When Job Satisfaction is taken into consideration, HR on an average has 2.6032 which is lower when compared to Research and Development’s score of 2.7263. This implies, HR’s environment score is more when compared to Sales and its Job Satisfaction score is less when compared with the Research and Development. By this we can conclude that what HR felt was right.
* Result: There are 3 records which shows the Average Environment Score and Average Job Satisfaction score Department wise.

### 

### Translation

Translation🡪 Select Department Name, Average Environment Satisfaction, Average Job Satisfaction from Department Table inner joined with Employee table on Department ID in Department table matching with Department ID of Employee table, again inner joining it with emp\_feedback table on Employee Number in feedback table matching with EmployeeNumber in Employee table and then grouped by Department Name.

Query🡪

use mydb;

Select D.Dept\_Name, avg(EF.Environment\_Satisfaction) as Avg\_Environment\_Satisfaction,

avg(EF.Job\_Satisfaction) as Avg\_Job\_Satisfaction from Department D

join Employee E on D.Dept\_ID = E.Dept\_ID

join employee\_feedback EF on E.EmployeeNumber = EF.EmployeeNumber

group by D.Dept\_Name;

### Screen Shot of SQL Query and Results:

### 

## SQL Query 5

### Question

An employee from Medical education field working in Sales department has spread a rumor saying that employees with his educational background are paid more in Research & Development than in Sales. What insight can you provide to prove or disprove that statement?

### Notes/Comments About SQL Query and Results (Include # of Rows in Result)

* To check whether Medical Education field employees are paid more in Research & Development than Sales Department.
* We have assumed MonthlyIncome column as a metric here since all the employees in the given dataset are assumed to be full time employees as mentioned in our data model.
* We assumed that Average suits well to compare the salary between the employees of two different departments and hence calculated the average Salary for each department. We rounded off the average salary column using the round function. The Job level and Job Role isn’t taken into consideration here as per our assumption and we are just comparing the average Salary of an employee.
* The average Salary of employee from Medical background working in Sales department is 6377 whereas it is 6539 in Research & Development. By this, we can conclude that the Rumor is true based on the results from the query set.
* Results: There are 3 records in the result set which shows the Education Field and the average salary of a person based on each department.

### Translation

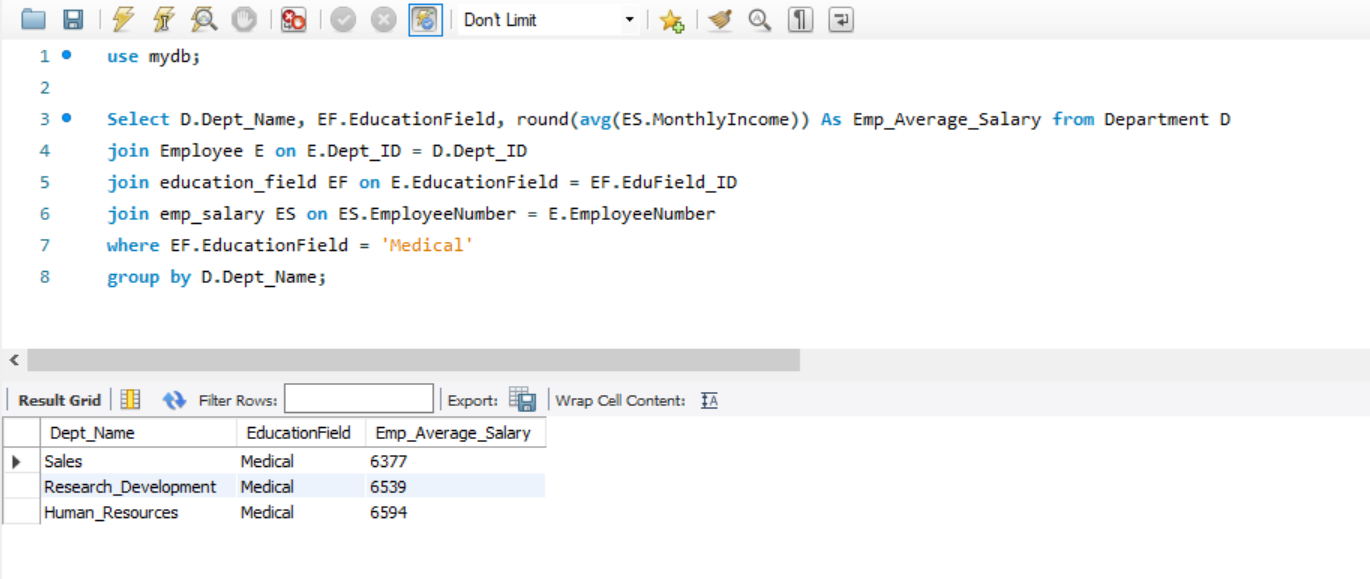
Translation🡪 Select Department Name, Education Field, Average Employee Salary from Department Table joined with Employee Table on Department ID of Department Table matching with Department ID of Employee table, again inner joining it with Education Field table on Education Field ID of Education Field table matching it with Employee table , again inner joining it with emp\_salary table on employee Number of Employee table matching it with EmployeeNumber of emp\_salary table where the Education Field is Medical and grouped by Department Name.

Query🡪

use mydb;

Select D.Dept\_Name, EF.EducationField, round(avg(ES.MonthlyIncome)) As Emp\_Average\_Salary from Department D join Employee E on E.Dept\_ID = D.Dept\_ID join education\_field EF on E.EducationField = EF.EduField\_ID join emp\_salary ES on ES.EmployeeNumber = E.EmployeeNumber where EF.EducationField = 'Medical' group by D.Dept\_Name;

### Screen Shot of SQL Query and Results



## SQL Query 6

### Question

A press article in a business magazine has said that at this company, single women in Sales have worked at the company longer than divorced or married women. What initial finding can you obtain from the data to help articulate the company's response in this regard?

### Notes/Comments About SQL Query and Results (Include # of Rows in Result)

* To check whether Single women have worked in the company longer than Married and Divorced women in Sales Department.
* We assumed that yearsAtCompany could help us in our findings. We thought that Average function could help us better in comparing the Single women to that of Married/Divorced women in the Sales Department.
* We have calculated the average stay in the company for all the women in the Sales Department. By the results, Single women work in the company on an average of 5.57 which is way less compared to that of 8.34 for Married women and 7.77 for Divorced women.
* By this we can conclude that, the press article in the magazine stated false news and, in the company, married and Divorced women work longer than the Single women.
* Results: There are 3 records with Gender, Marital Status and Average Years at Company columns which are grouped based on the Departments.

### Translation

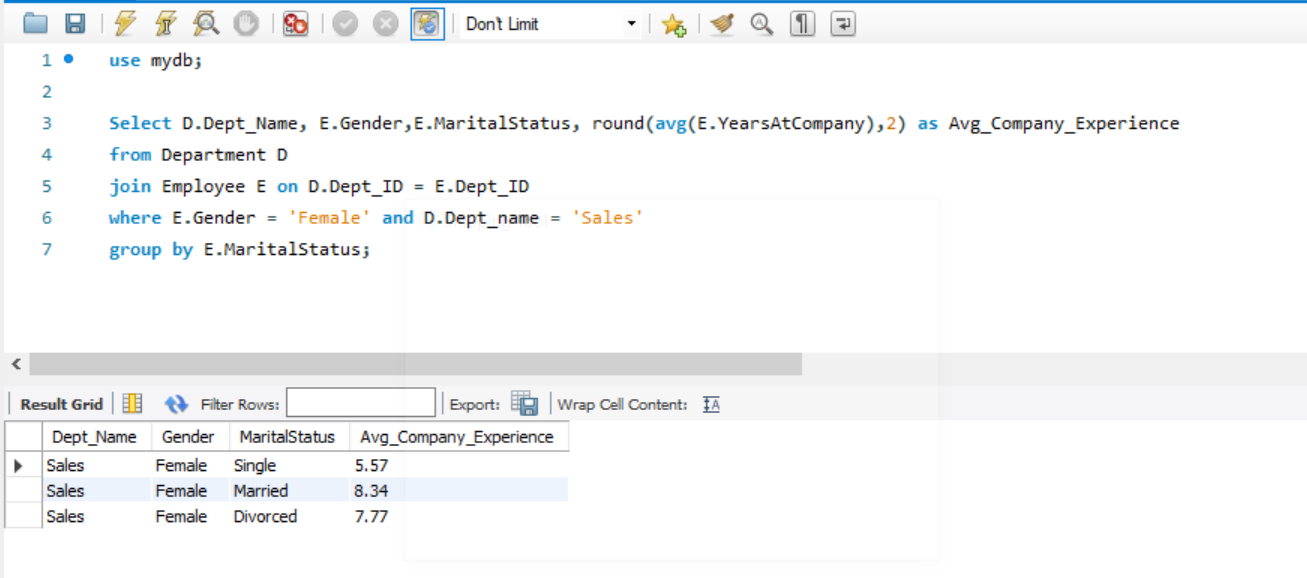
Translation🡪 Select Department Name, Gender, Marital Status and average Company Experience from Department inner join with Employee on Department ID of Department Table matched with Department ID of Employee Table where Gender is female and Department Name is Sales grouped by Marital Status.

Query🡪

use mydb;

Select D.Dept\_Name, E.Gender,E.MaritalStatus, round(avg(E.YearsAtCompany),2) as Avg\_Company\_Experience from Department D join Employee E on D.Dept\_ID = E.Dept\_ID where E.Gender = 'Female' and D.Dept\_name = 'Sales' group by E.MaritalStatus;

### Screen Shot of SQL Query and Results



# Data Review for MongoDB

## Assumptions/Notes About Data Collections, Attributes and Relationships between Collections

# Physical Mongo Database

## Assumptions/Notes About Data Set

## Screen shot of Physical Database objects (Database, Collections and Attributes)

## Data in the Database

|  |  |  |
| --- | --- | --- |
| **Collection Name** | **Relationships With Other Collections (if any)** | **# of Documents in Collection** |
|  |  |  |

# MongoDB Queries/Code

Pick 3 SQL queries and write them in MongoDB

## Mongo Query 1

### Question

### Notes/Comments About MongoDB Query/Code and Results (Include # of Documents in Result)

### Translation

### Screen Shot of MongoDB Query/Code and Results

## Mongo Query 2

### Question

### Notes/Comments About MongoDB Query/Code and Results (Include # of Documents in Result)

### Translation

### Screen Shot of MongoDB Query/Code and Results

## Mongo Query 3

### Question

### Notes/Comments About MongoDB Query/Code and Results (Include # of Documents in Result)

### Translation

### Screen Shot of MongoDB Query/Code and Results