

**SQL-Mongo Project – IBM HR Analytics Attrition**

BUAN 6320

May 2, 2020

**Project Report**

Arush Rao

# **Contents**

[**Relational Data Model** 5](#_Toc39351247)

[**Assumptions/Notes About Data Entities and Relationships :** 5](#_Toc39351248)

[**Entity-Relationship Diagram:** 7](#_Toc39351249)

[**Physical MySQL Database** 8](#_Toc39351250)

[**Assumptions/Notes About Data Set :** 8](#_Toc39351251)

[**Screen shot of Physical Database objects :** 9](#_Toc39351252)

[**Data in the Database:** 9](#_Toc39351253)

[**SQL Queries** 25](#_Toc39351254)

[**SQL Query 1:** 25](#_Toc39351255)

[**Question:** 25](#_Toc39351256)

[**Notes:** 25](#_Toc39351257)

[**Translation:** 25](#_Toc39351258)

[**Cleanup:** 25](#_Toc39351259)

[**Screenshot:** 25](#_Toc39351260)

[**SQL Query 2:** 26](#_Toc39351261)

[**Question:** 26](#_Toc39351262)

[**Notes:** 26](#_Toc39351263)

[**Translation:** 26](#_Toc39351264)

[**Cleanup:** 26](#_Toc39351265)

[**Screenshot:** 26](#_Toc39351266)

[**SQL Query 3:** 27](#_Toc39351267)

[**Question:** 27](#_Toc39351268)

[**Notes:** 27](#_Toc39351269)

[**Translation:** 27](#_Toc39351270)

[**Cleanup:** 27](#_Toc39351271)

[**Screenshot:** 27](#_Toc39351272)

[**SQL Query 4:** 28](#_Toc39351273)

[**Question:** 28](#_Toc39351274)

[**Notes:** 28](#_Toc39351275)

[**Translation:** 28](#_Toc39351276)

[**Cleanup:** 28](#_Toc39351277)

[**Screenshot:** 28](#_Toc39351278)

[**SQL Query 5:** 29](#_Toc39351279)

[**Question:** 29](#_Toc39351280)

[**Notes:** 29](#_Toc39351281)

[**Translation:** 29](#_Toc39351282)

[**Cleanup:** 29](#_Toc39351283)

[**Screenshot:** 29](#_Toc39351284)

[**SQL Query 6:** 30](#_Toc39351285)

[**Question:** 30](#_Toc39351286)

[**Notes:** 30](#_Toc39351287)

[**Translation:** 30](#_Toc39351288)

[**Cleanup:** 30](#_Toc39351289)

[**Screenshot:** 30](#_Toc39351290)

[**Data Review for MongoDB** 31](#_Toc39351291)

[**Assumptions/Notes About Data Collections, Attributes and Relationships between Collections:** 31](#_Toc39351292)

[**Physical Mongo Database** 31](#_Toc39351293)

[**Assumptions/Notes About Data Set** 31](#_Toc39351294)

[**Screen shot of Physical Database objects (Database, Collections and Attributes) :** 32](#_Toc39351295)

[**Data in the Database:** 33](#_Toc39351296)

[**MongoDB Queries/Code** 34](#_Toc39351297)

[**Query 1** 34](#_Toc39351298)

[**Question 1:** 34](#_Toc39351299)

[**Notes/Comments About MongoDB Query/Code and Results:** 34](#_Toc39351300)

[**Translation – Query:** 34](#_Toc39351301)

[**Screen Shot of MongoDB Query/Code and Results :** 34](#_Toc39351302)

[**Query 2** 35](#_Toc39351303)

[**Question 2:** 35](#_Toc39351304)

[**Notes/Comments About MongoDB Query/Code and Results:** 35](#_Toc39351305)

[**Translation – Query:** 35](#_Toc39351306)

[**Screen Shot of MongoDB Query/Code and Results :** 35](#_Toc39351307)

[**Query 3** 36](#_Toc39351308)

[**Question 3:** 36](#_Toc39351309)

[**Notes/Comments About MongoDB Query/Code and Results:** 36](#_Toc39351310)

[**Translation – Query:** 36](#_Toc39351311)

[**Screen Shot of MongoDB Query/Code and Results :** 36](#_Toc39351312)

# **Relational Data Model**

## 

### **Assumptions/Notes About Data Entities and Relationships :**

1. **Employee - EmpSalary**

An Employee can have multiple salaries. One salary can belong to only one employee. They have an identifying ‘one to many’ relationship. (From, To and Employee\_ID form a composite Primary Key for the Salary Table.)

1. **Employee - EmpReview**

An Employee can have one to many Reviews. A review can belong to only one employee. They have an identifying ‘one to many’ relationship. (From, To and Employee\_ID form a composite Primary Key for the Review Table.)

1. **Employee – EmpDepartment**

An Employee can be part of only one department (Dept\_ID). A department can have one to many employees. They have a non-identifying ‘one to many’ relationship.

1. **Employee – Job Level**

An Employee can have only one Job Level (Job\_Level\_ID). A Job can belong to one to many employees. They have a non-identifying ‘one to many’ relationship.

1. **Employee – Travel Master**

An Employee can have only one Travel ID associated to them. A Travel ID can be associated to one to many employees. They have a non-identifying ‘one to many’ relationship.

1. **Employee – Education Level Master**

An Employee can have only one education level (EduLevel\_ID) associated to them. An education level can be associated to one to many employees. They have a non-identifying ‘one to many’ relationship.

1. **Employee – Standard Hours**

An Employee can have only one 'Standard hours' value (StdHours\_ID) associated to them. A 'Standard Hours' value can be associated to one to many employees. They have a non-identifying ‘one to many’ relationship.

1. **Employee – Job Involvement**

An Employee can have one Job Involvement ID (Involvement\_ID) associated to them. A Job Involvement ID can be associated to one to many employees. They have a non-identifying ‘one to many’ relationship.

1. **Employee – Education Field**

An Employee can have only one Education Field (EduField\_ID) associated to them. An Education Field can be associated to one to many employees. They have a non-identifying ‘one to many’ relationship.

1. **Employee – Work Life Balance**

An Employee can have only one Work Life Balance ID (Balance\_ID) associated to them. A Work Life Balance ID can be associated to one to many employees. They have a non-identifying ‘one to many’ relationship.

1. **Employee – Rel Satisfaction**

An Employee can have only one Relationship Satisfaction ID (Rel\_ID) associated to them. A Relationship Satisfaction ID can be associated to one to many employees. They have a non-identifying ‘one to many’ relationship.

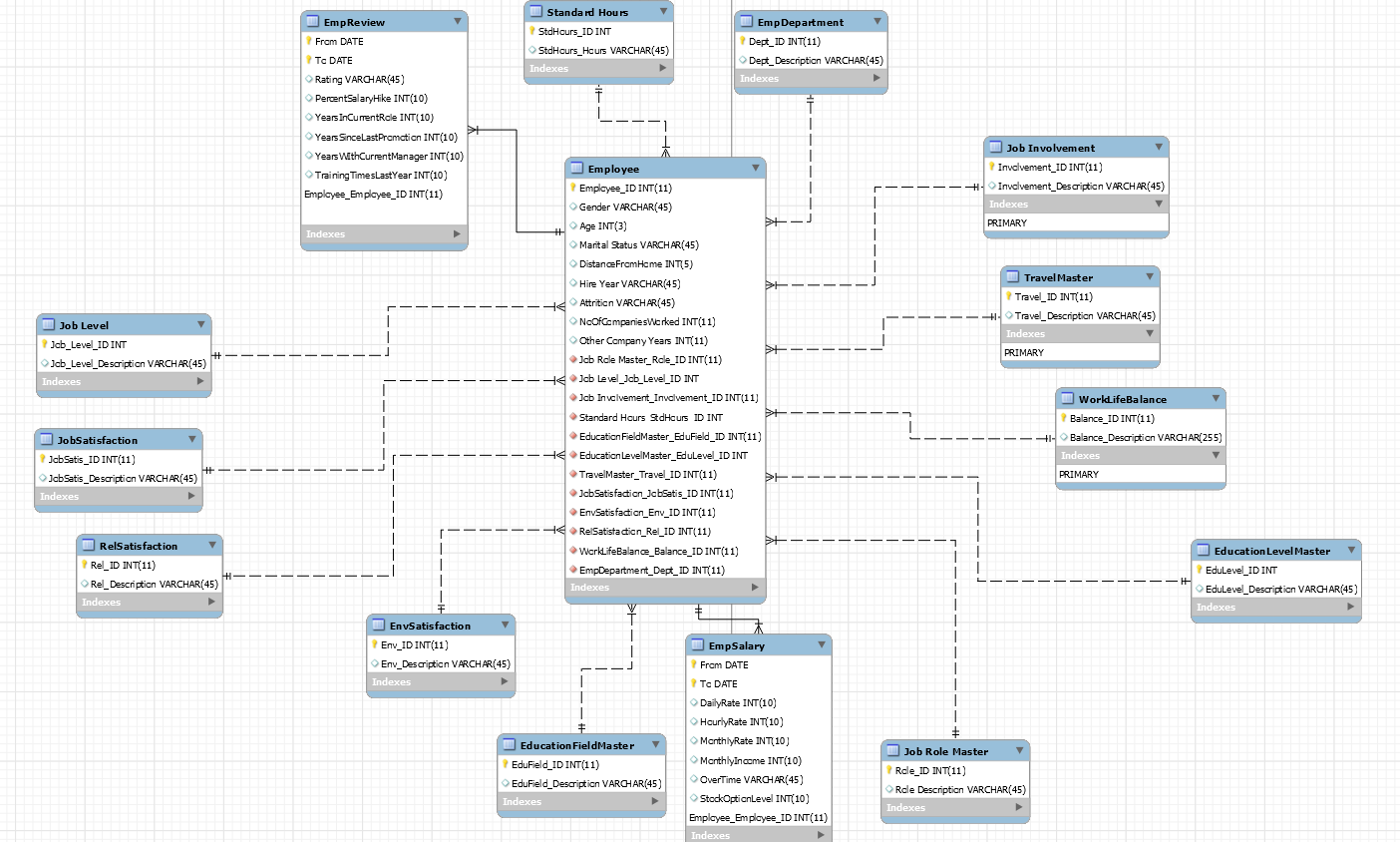
1. **Employee – Env Satisfaction**

An Employee can have only one Environment Satisfaction ID (Env\_ID) associated to them. An Environment Satisfaction ID can be associated to one to many employees. They have a non-identifying ‘one to many’ relationship.

1. **Employee – Job Satisfaction**

An Employee can have only one Job Satisfaction ID (JobSatis\_ID) associated to them. A Job Satisfaction ID can be associated to one to many employees. They have a non-identifying ‘one to many’ relationship.

### **Entity-Relationship Diagram:**



# **Physical MySQL Database**

## 

### **Assumptions/Notes About Data Set :**

1. Salary and Performance are always period related. Though the Salary and Performance details are employee related, they need to have a period. So, the 'From' and 'To' date have been included in the Salary and Performance tables. We assume the salary and performance to be for the previous year i.e. 01/01/2019 – 12/31/2019.
2. The employee table includes the fields of Total Working Years and Years at Company. However, these values are static and would remain the same if we were to use this database the next year which would render it unusable. We introduce 2 columns called ‘Hire Year’ and ‘No. of Years worked at Other companies’ to tackle this issue. We assume that all calculations are to be made considering the current year as 2020. This would enable us to make calculations at any given time in the future as well.
3. Since the data set does not include the significance of the Job Levels, we assume the Job Levels to be the following:

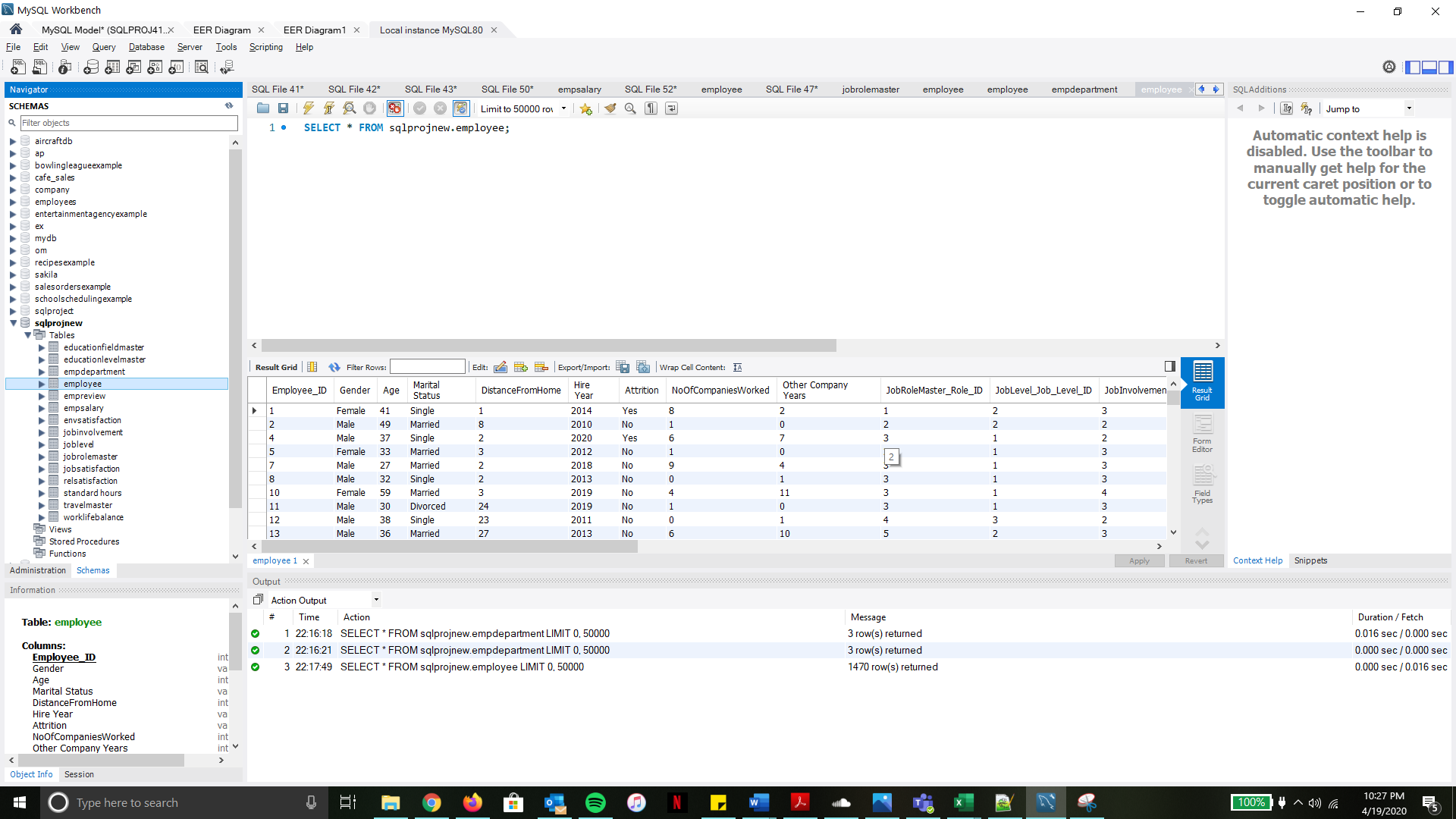
|  |  |
| --- | --- |
| Job Level | Significance |
| 1 | Staff |
| 2 | Senior Staff |
| 3 | Managers & Advisors |
| 4 | Middle Management |
| 5 | Executive Management |

1. Although the ‘Standard hours’ value is same for all employees i.e. 80, we create a separate table assuming the standard hours may change or if there are employees working across different countries, these values may vary.

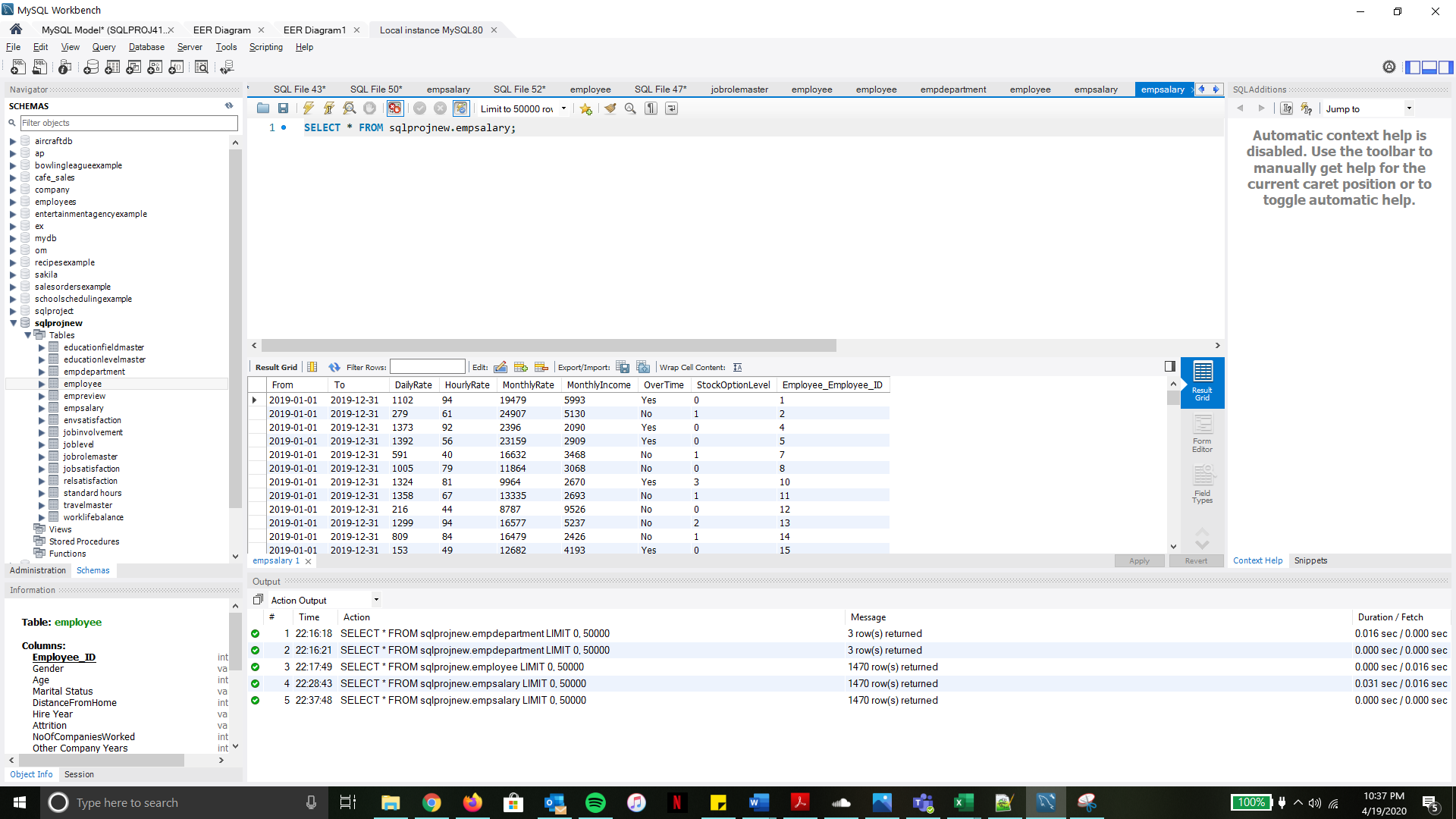
### **Screen shot of Physical Database objects :**

### **Data in the Database:**

|  |  |  |  |
| --- | --- | --- | --- |
| **Table Name** | **Primary Key** | **Foreign Key** | **# of Rows in Table** |
| Employee | Employee\_ID | JobRoleMaster\_RoleID  JobLevel\_Job\_Level\_ID  JobInvolvement\_Involvement\_ID  StandardHours\_StdHours\_ID  EducationFieldMaster\_EduField\_ID  EducationLevelMaster\_EduLevel\_ID  TravelMasterTravel\_ID  JobSatisfaction\_JobSatis\_ID  EnvSatisfaction\_Env\_ID  RelSatisfaction\_Rel\_ID  WorkLifeBalance\_Balance\_ID  EmpDepartment\_Dept\_ID | 1470 |



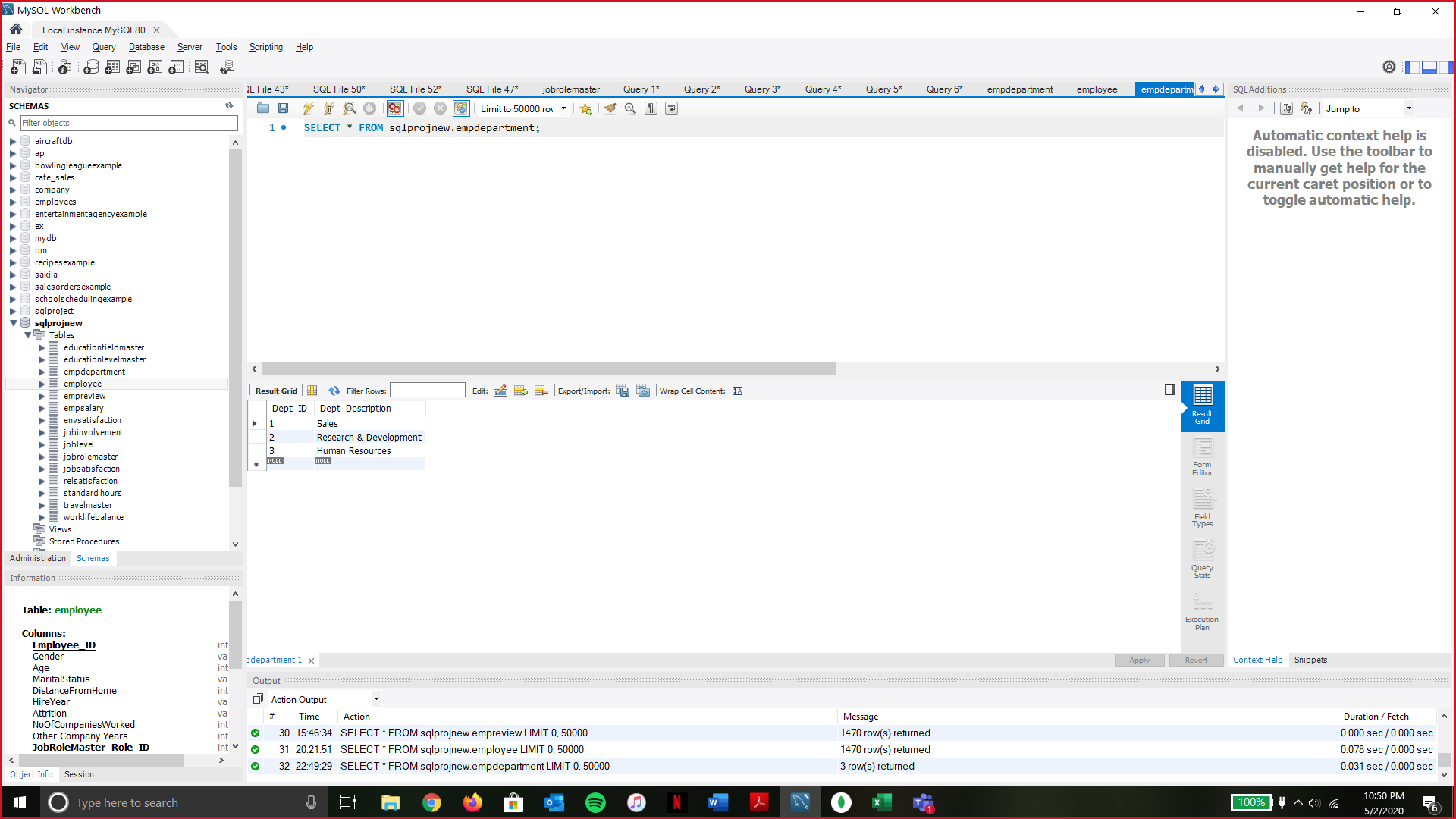
|  |  |  |  |
| --- | --- | --- | --- |
| **Table Name** | **Primary Key** | **Foreign Key** | **# of Rows in Table** |
| EmpSalary | From  To  Employee\_ID | N/A | 1470 |



|  |  |  |  |
| --- | --- | --- | --- |
| **Table Name** | **Primary Key** | **Foreign Key** | **# of Rows in Table** |
| EmpReview | From  To  Employee\_ID | N/A | 1470 |



|  |  |  |  |
| --- | --- | --- | --- |
| **Table Name** | **Primary Key** | **Foreign Key** | **# of Rows in Table** |
| EmpDepartment | Dept\_ID | N/A | 3 |

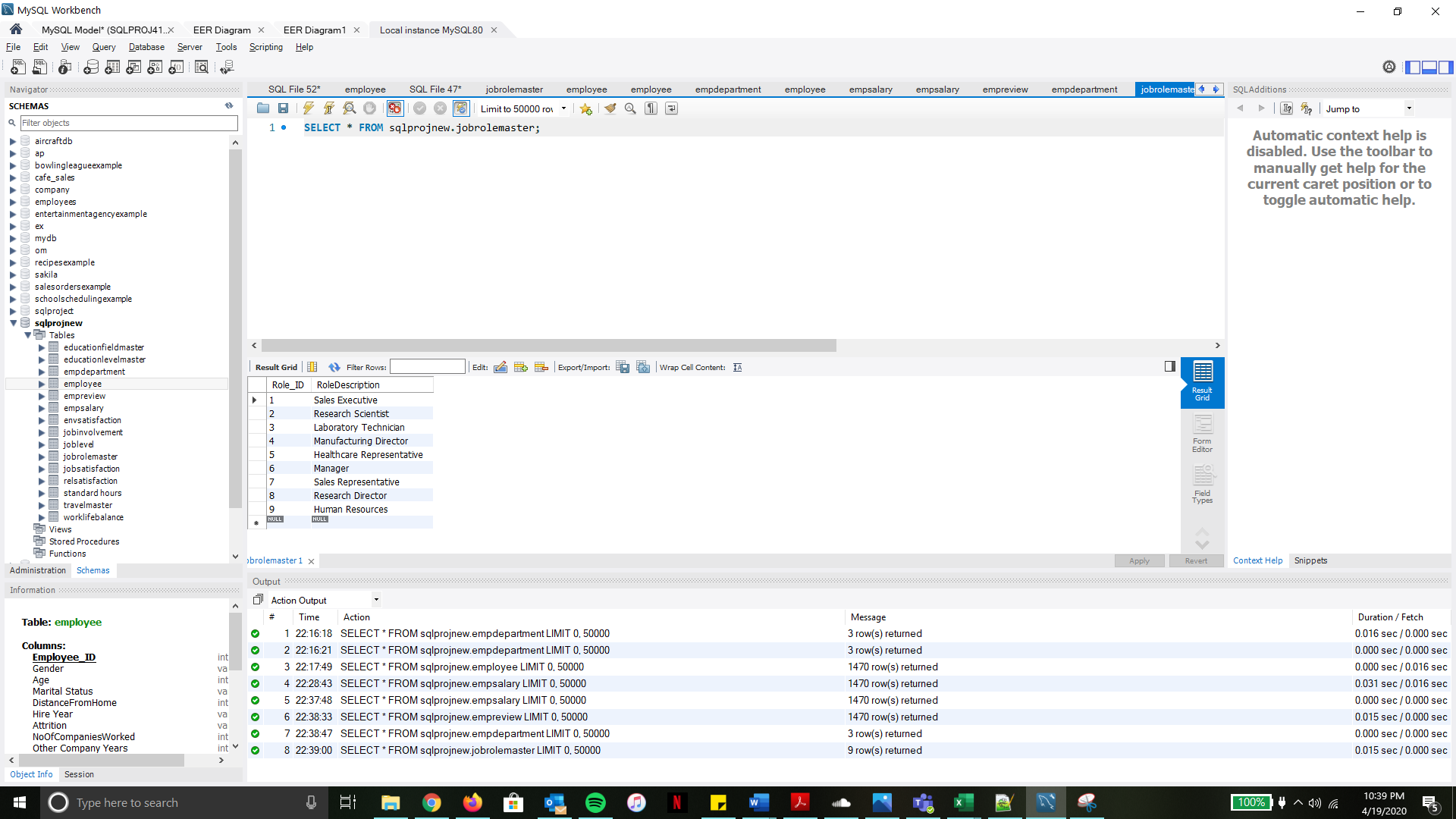


\*Change made: Dept\_ID:1 was changed from Human Resources to Sales.

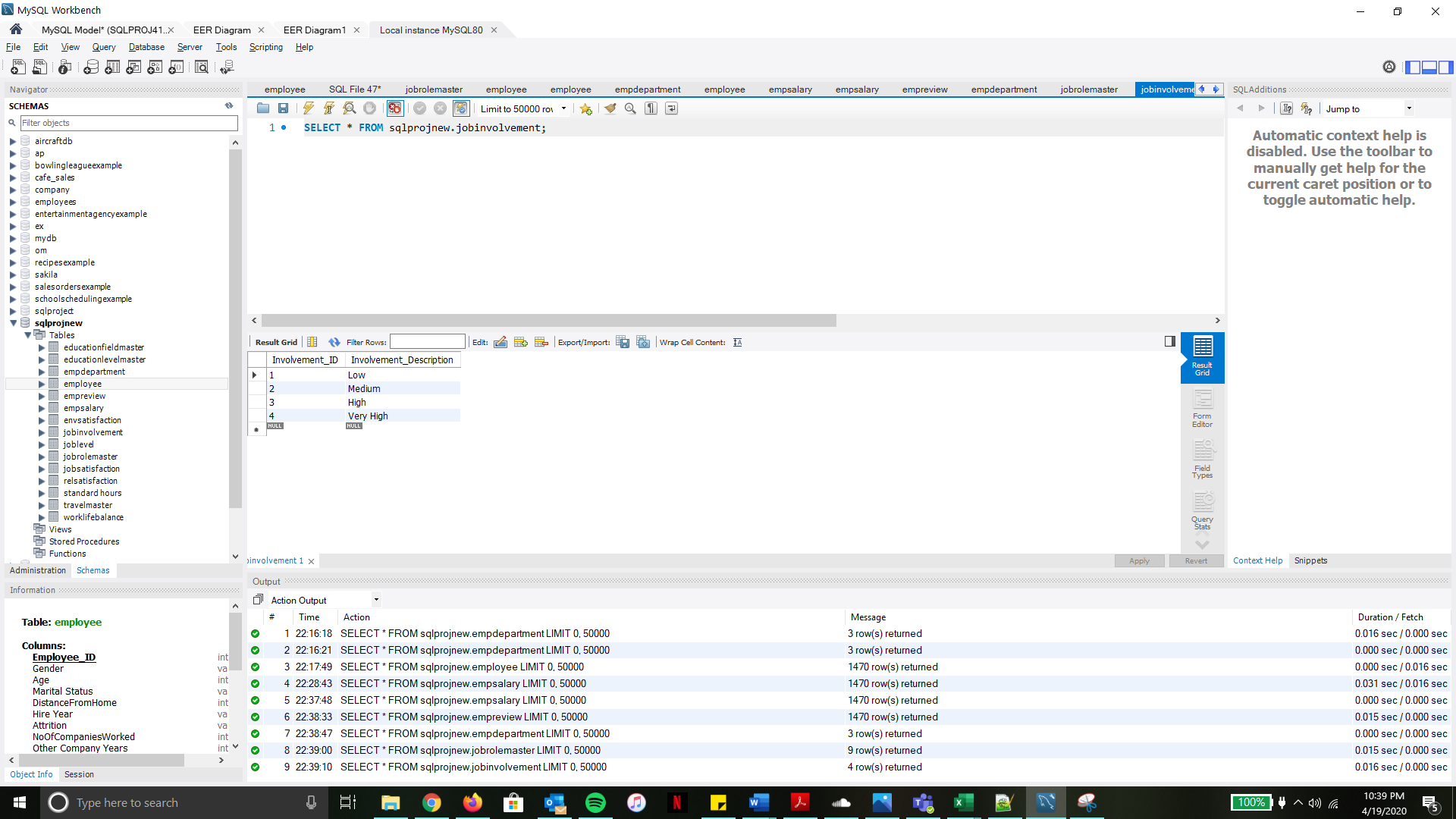
Dept\_ID: 3 was changed from Sales to Human Resources.

This was done as the mapping in the employee table was done based on these values.

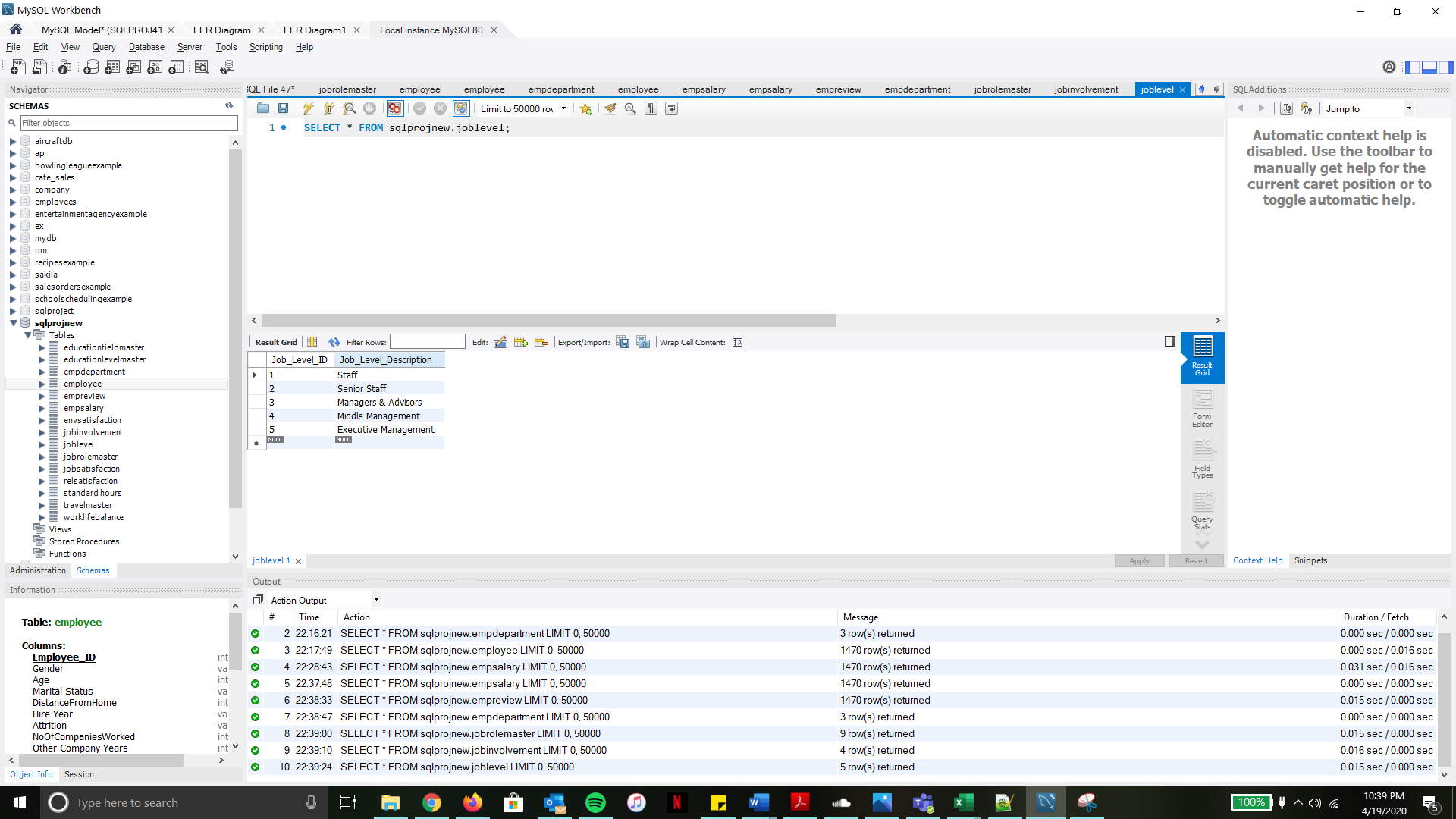
|  |  |  |  |
| --- | --- | --- | --- |
| **Table Name** | **Primary Key** | **Foreign Key** | **# of Rows in Table** |
| JobRoleMaster | Role\_ID | N/A | 9 |



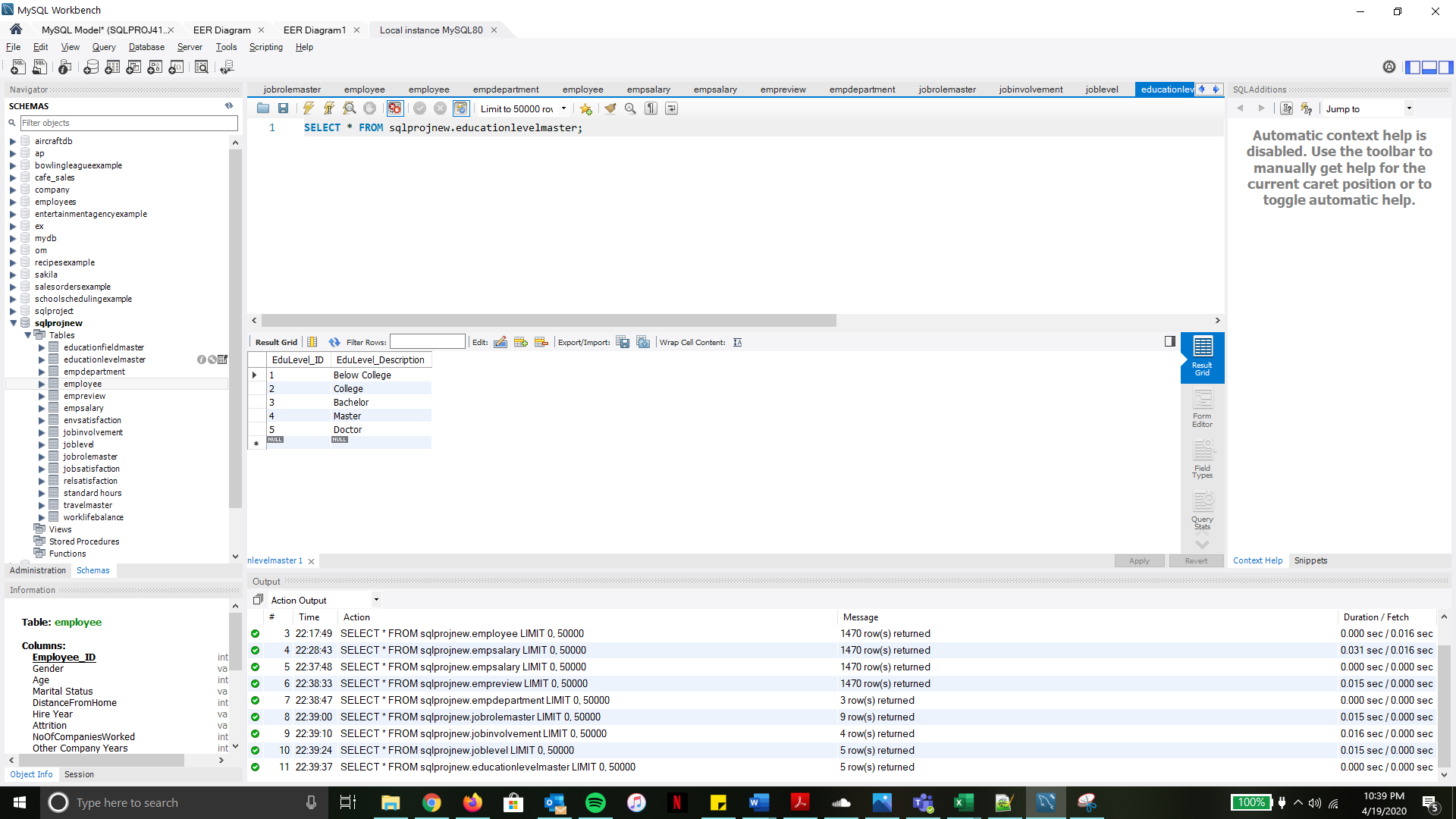
|  |  |  |  |
| --- | --- | --- | --- |
| **Table Name** | **Primary Key** | **Foreign Key** | **# of Rows in Table** |
| JobInvolvement | Involvement\_ID | N/A | 4 |



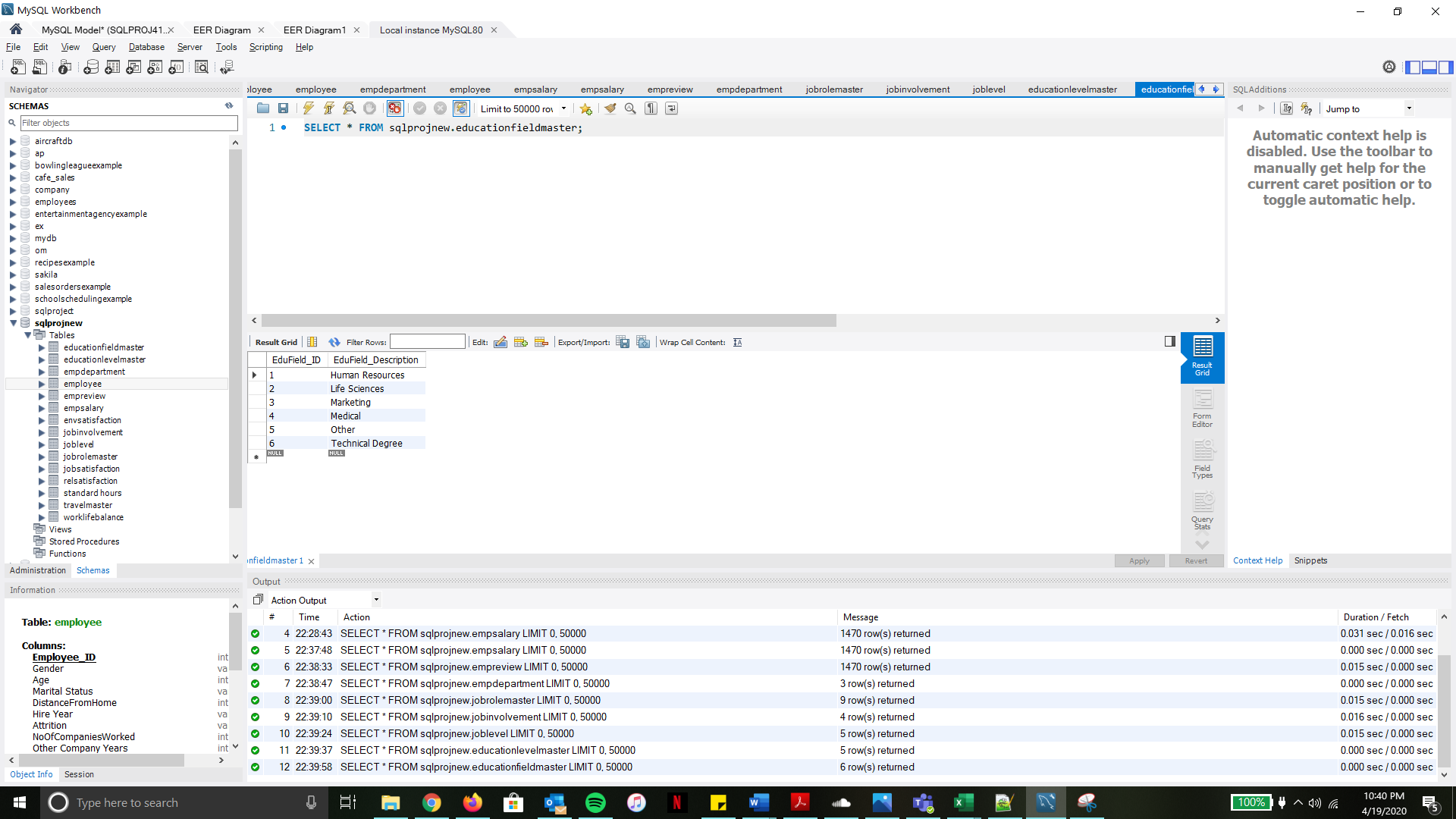
|  |  |  |  |
| --- | --- | --- | --- |
| **Table Name** | **Primary Key** | **Foreign Key** | **# of Rows in Table** |
| JobLevel | Job\_Level\_ID | N/A | 5 |



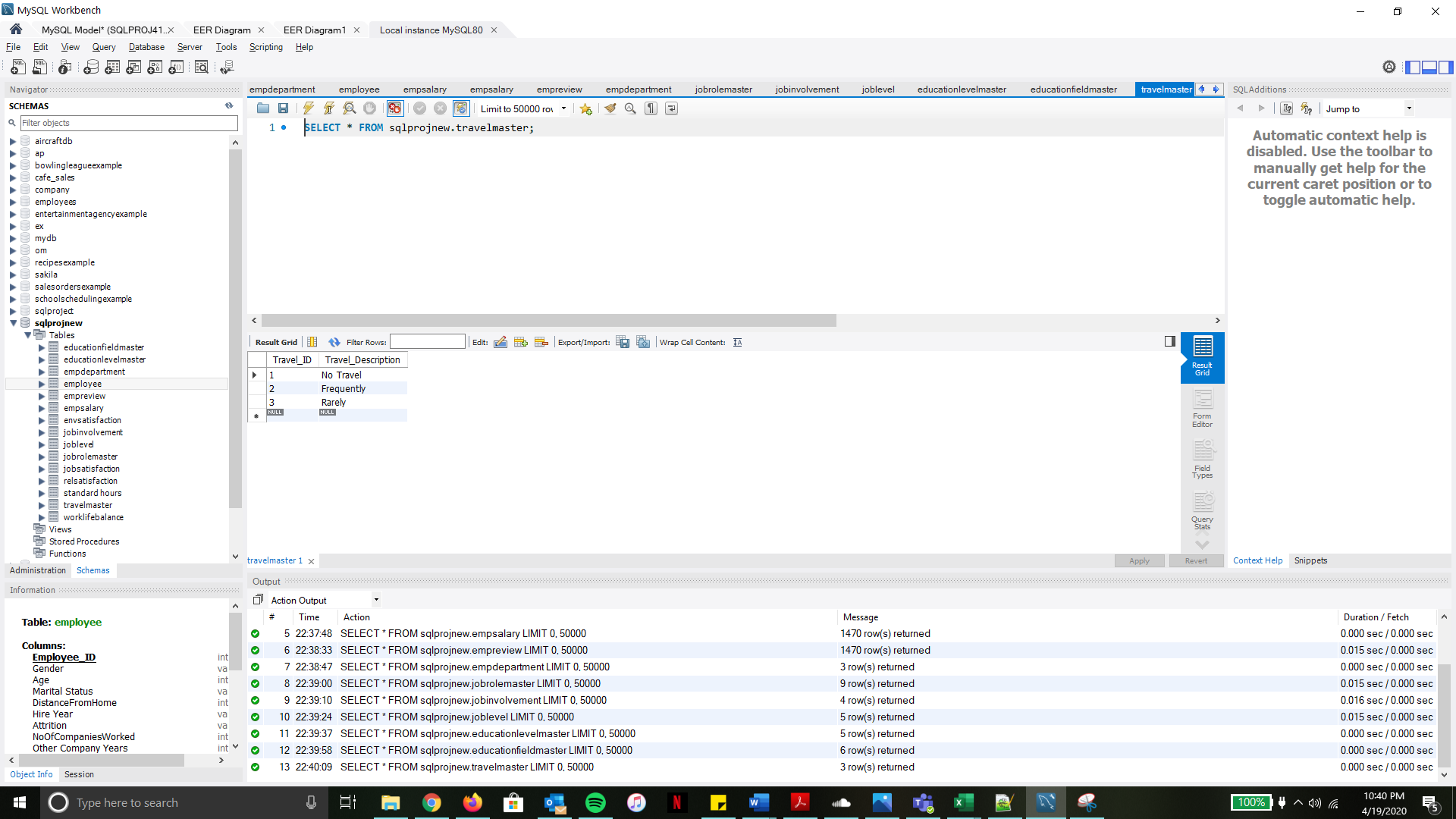
|  |  |  |  |
| --- | --- | --- | --- |
| **Table Name** | **Primary Key** | **Foreign Key** | **# of Rows in Table** |
| EducationLevelMaster | EduLevel\_ID | N/A | 5 |



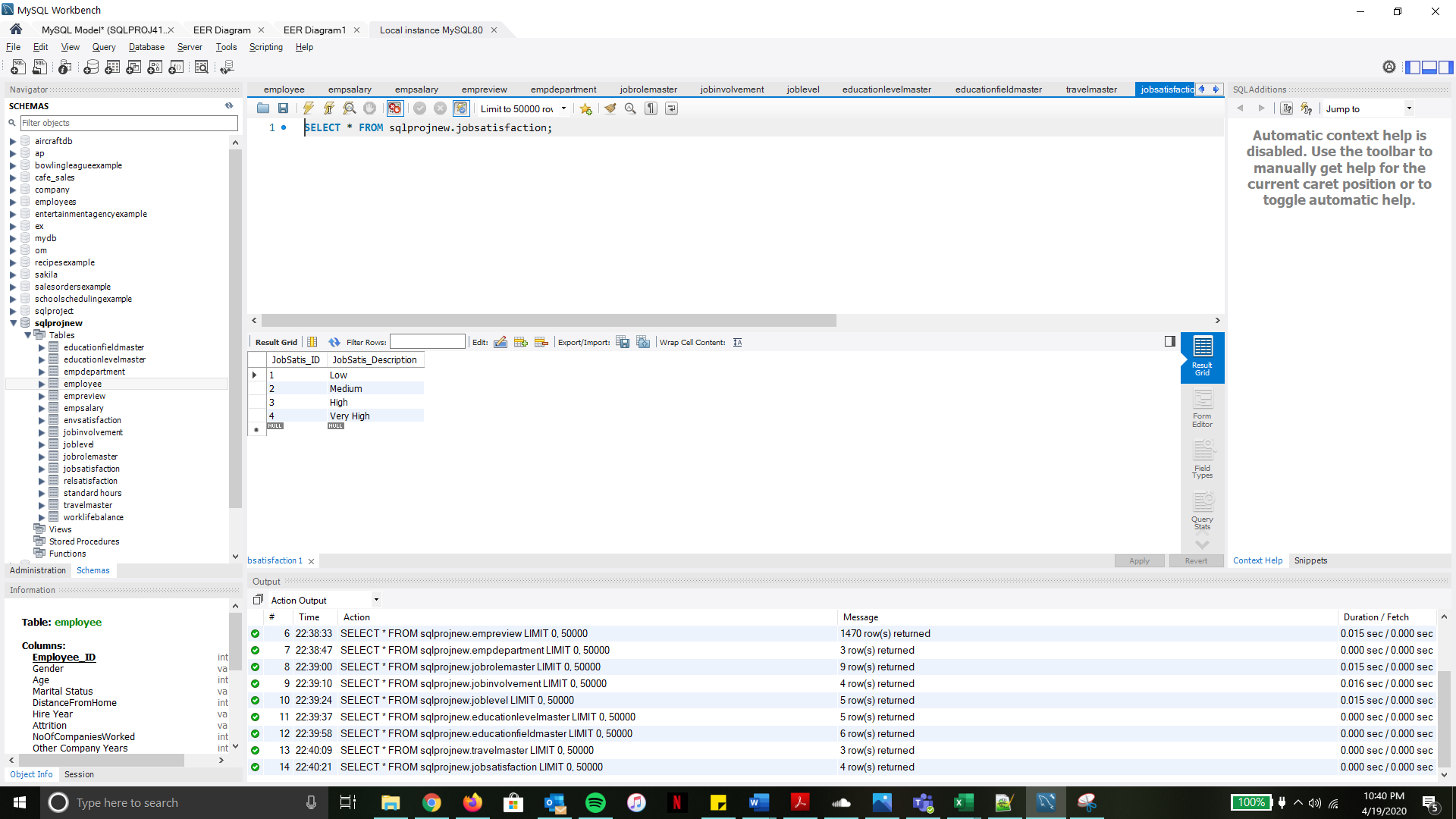
|  |  |  |  |
| --- | --- | --- | --- |
| **Table Name** | **Primary Key** | **Foreign Key** | **# of Rows in Table** |
| EducationFieldMaster | EduField\_ID | N/A | 6 |



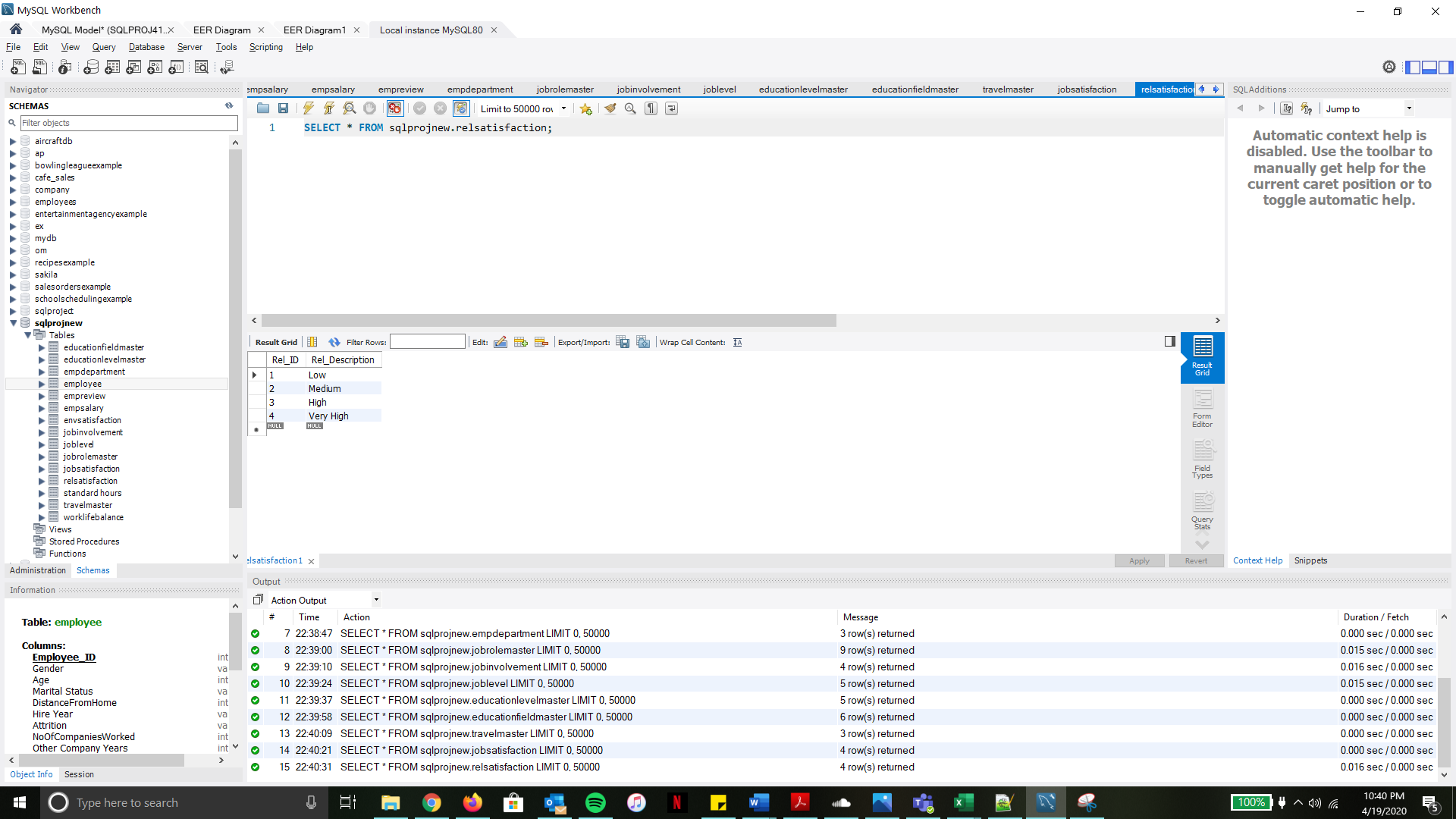
|  |  |  |  |
| --- | --- | --- | --- |
| **Table Name** | **Primary Key** | **Foreign Key** | **# of Rows in Table** |
| TravelMaster | Travel\_ID | N/A | 3 |



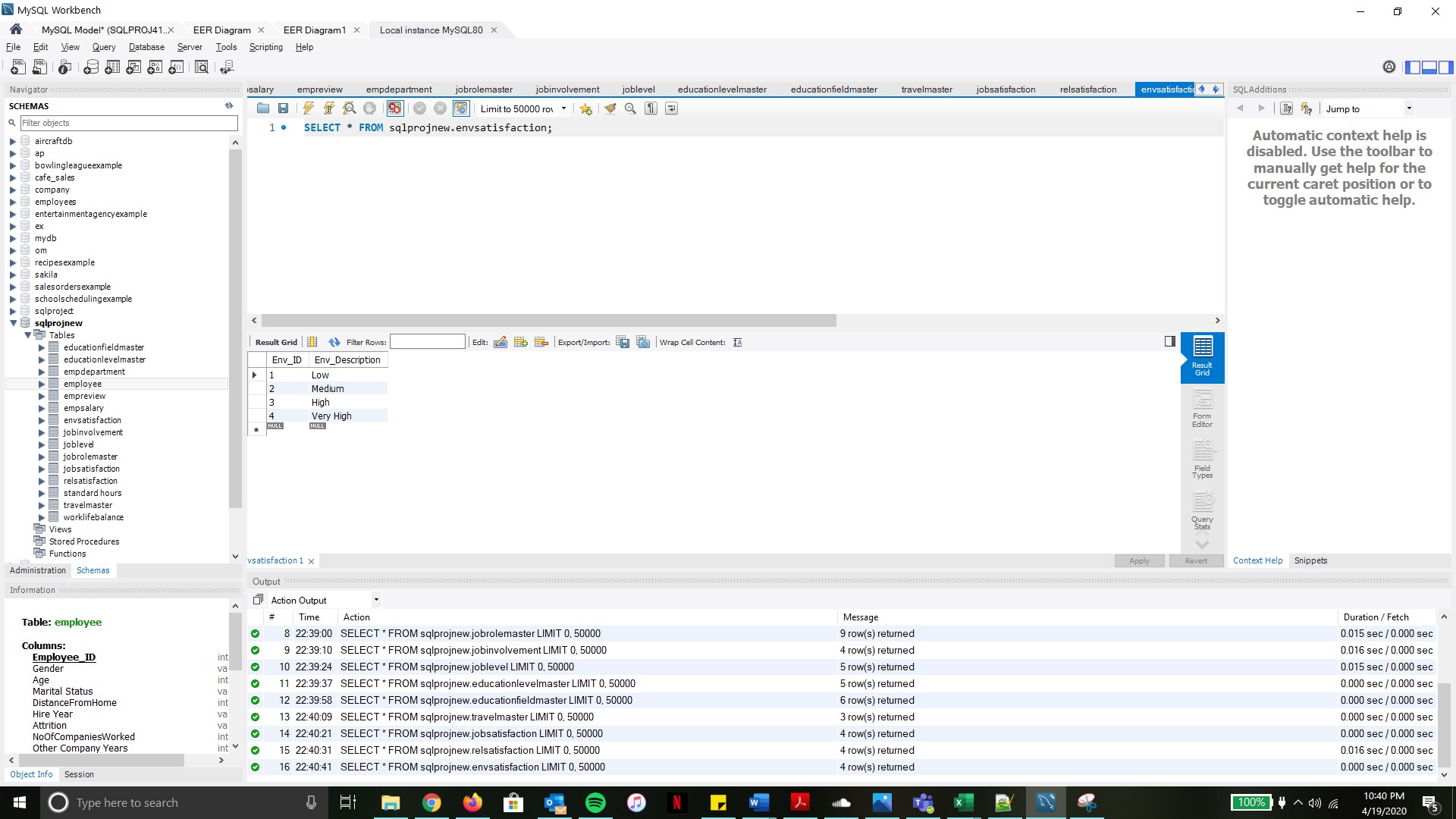
|  |  |  |  |
| --- | --- | --- | --- |
| **Table Name** | **Primary Key** | **Foreign Key** | **# of Rows in Table** |
| JobSatisfaction | JobSatis\_ID | N/A | 4 |



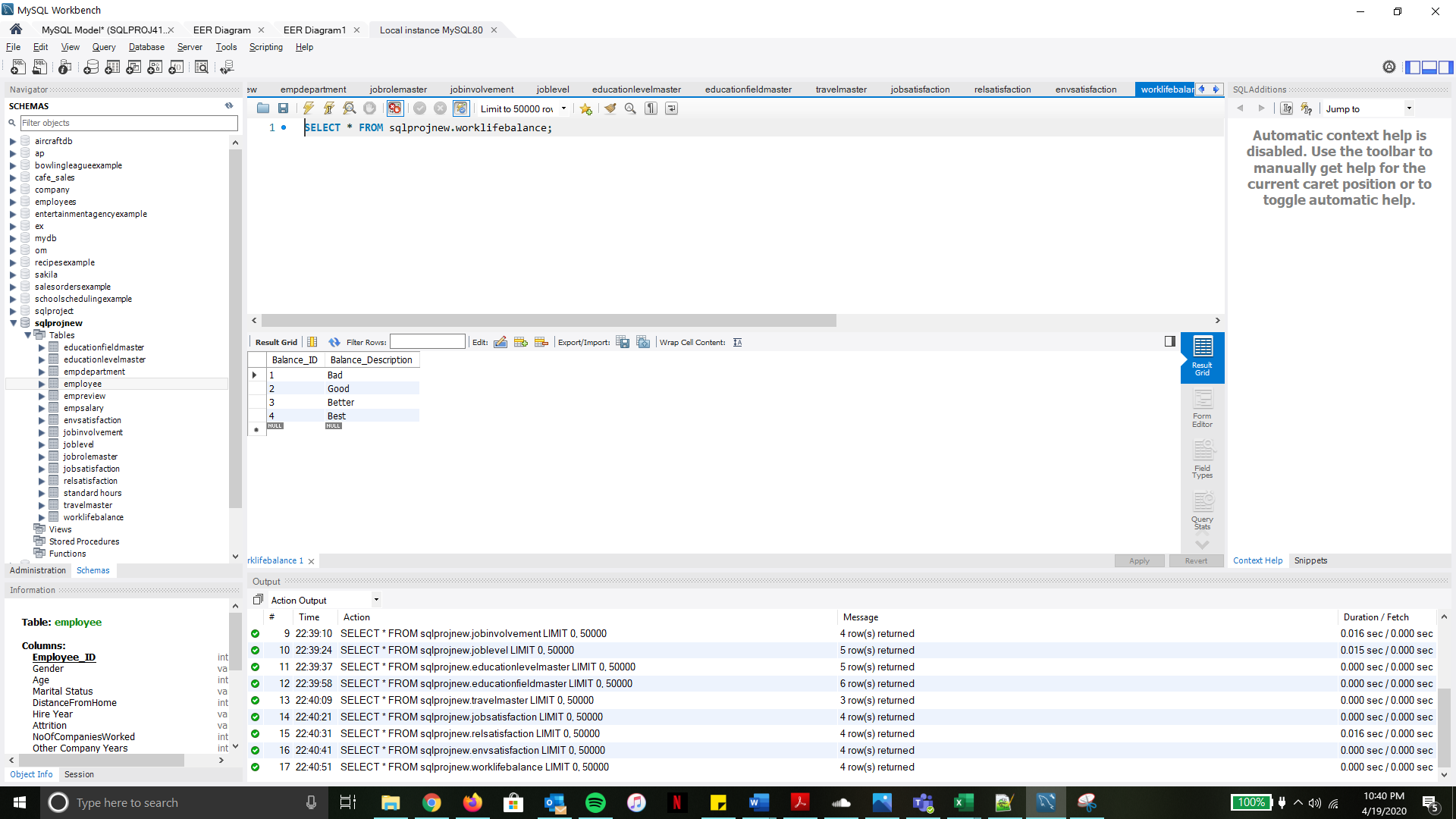
|  |  |  |  |
| --- | --- | --- | --- |
| **Table Name** | **Primary Key** | **Foreign Key** | **# of Rows in Table** |
| RelSatsifaction | Rel\_ID | N/A | 4 |



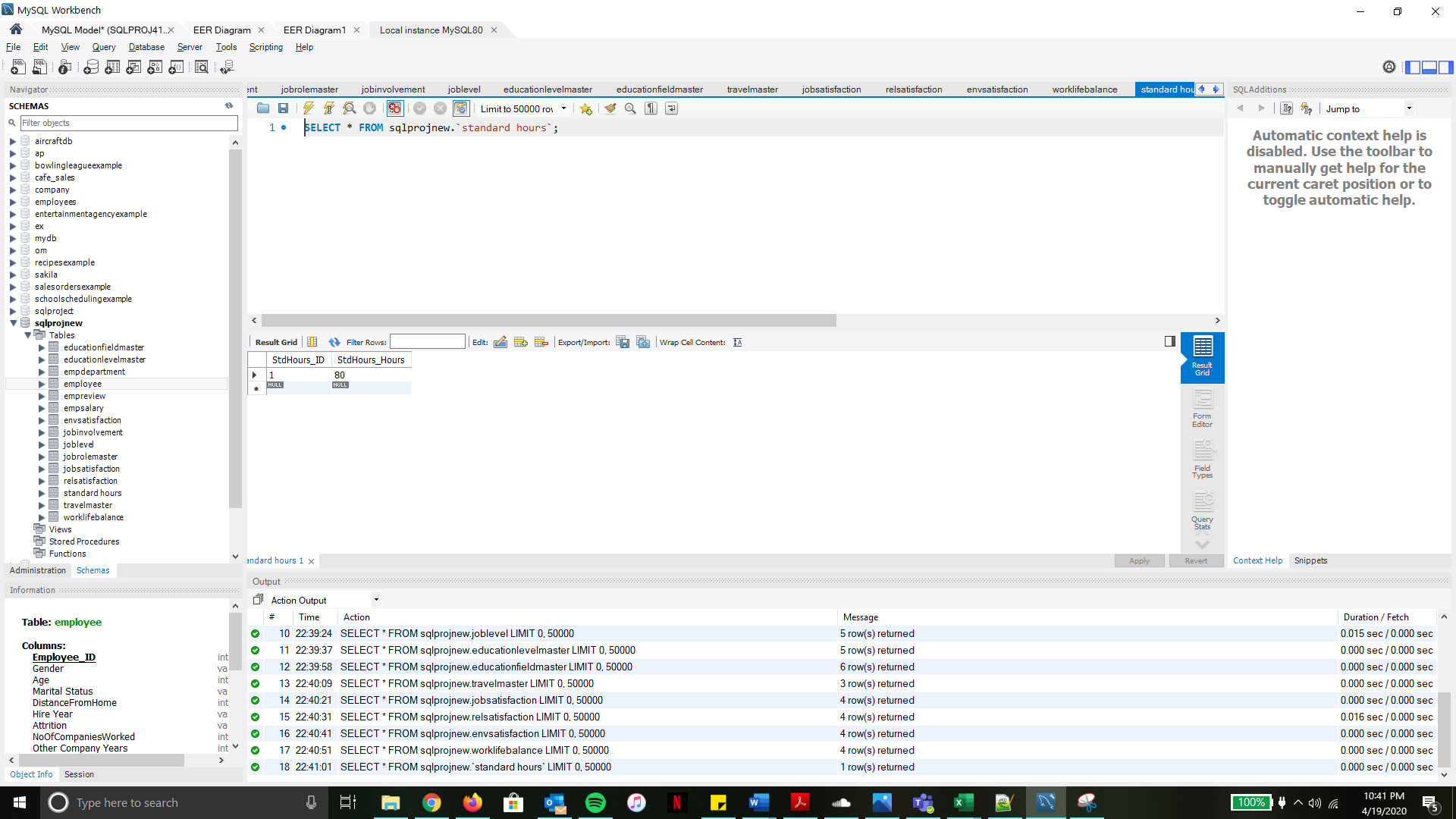
|  |  |  |  |
| --- | --- | --- | --- |
| **Table Name** | **Primary Key** | **Foreign Key** | **# of Rows in Table** |
| EnvSatisfaction | Env\_ID | N/A | 4 |



|  |  |  |  |
| --- | --- | --- | --- |
| **Table Name** | **Primary Key** | **Foreign Key** | **# of Rows in Table** |
| WorkLifeBalance | Balance\_ID | N/A | 4 |



|  |  |  |  |
| --- | --- | --- | --- |
| **Table Name** | **Primary Key** | **Foreign Key** | **# of Rows in Table** |
| StandardHours | StdHours\_ID | N/A | 1 |



# **SQL Queries**

### **SQL Query 1:**

**Question:** A new research scientist who loves to travel joins the firm and is told by HR that his job role is one of the top two roles in terms of employees that travel frequently. Is HR right in saying so? Why or Why not?

**Notes:** In order to find out if the HR is right in saying so, we need to find the list of all employees who travel frequently and group them based on their Job role. This would help us in understanding the Job roles which have employees travelling frequently. On querying the database, it returns **9 rows** grouped by Job Role with the number of employees in each Job Role who travel frequently. We see that Sales Executive is the Job Role with the most employees(59) who travel frequently followed by Research Scientist(54). Since HR says that Research Scientist is one of the two job roles with employees travelling frequently, they are right in saying so.

**Translation:** Select list of all employees who travel frequently and group them by Job Role

**Cleanup:** SELECT RoleDescription, count(RoleDescription) , TravelDescription

FROM Employee

JOIN Travelmaster ON TravelID

JOIN JobRoleMaster ON ROLEID

WHERE TravelID = 2 (TravelFrequently)

GROUP BY RoleDescription

#### **Screenshot:**



### **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 least?

**Notes:** To find the department’s employees who will gain the least, we need to group all employees based on the department they belong to and then find the average distance from home of these employees. On querying the database, it returns **3 rows** as there are only 3 departments along with the average distance from work of these employees. Based on the result, we can say that employees belonging to the ‘**Human Resources’** department stand to gain the least with an increase in the per mile compensation as the average distance from home of these employees is the least.

**Translation:** Calculate average distance from home of employees based on their department.

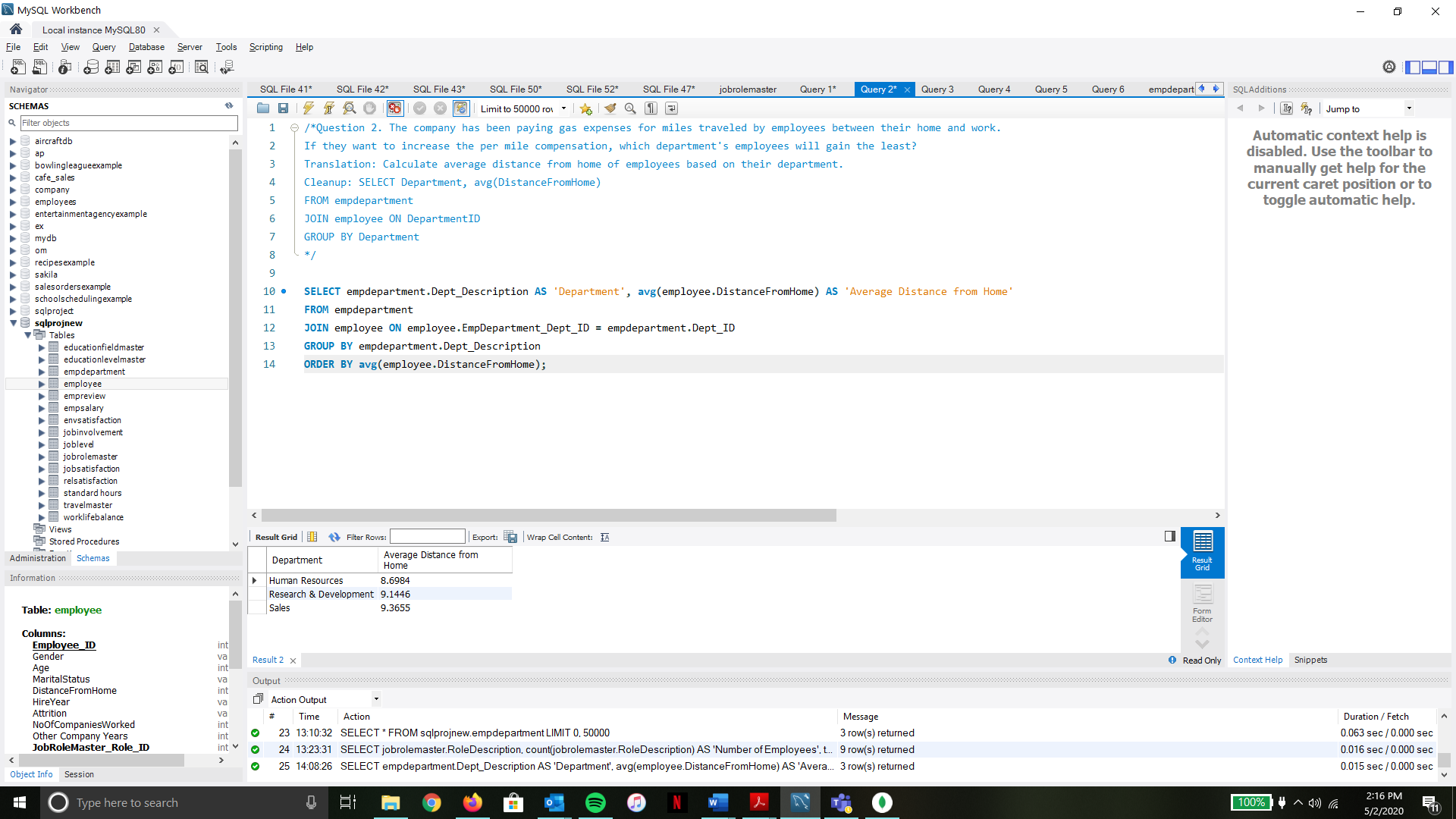
**Cleanup:** SELECT Department, avg(DistanceFromHome)

FROM empdepartment

JOIN employee ON DepartmentID

GROUP BY Department

#### **Screenshot:**



### **SQL Query 3:**

**Question:** A new employee with a Marketing degree wants to work in HR.

Do you believe the company might be able to give him a chance to work in HR? Why or Why not?

**Notes:** To verify this, we need to check if there are any existing employees in the HR department with an education in the field of Marketing. On querying the database, it returns **0 rows** indicating that there are currently no employees in the HR department who also have an education in the Marketing Field. Hence, we can conclude that the company will not be able to give the new employee with a marketing degree, a chance to work in HR.

**Translation:** Select list of employees working in HR Department with an education in the Marketing field.

**Cleanup:** SELECT empdartment.Dept\_Description, educationfieldmaster.EduField\_Description,

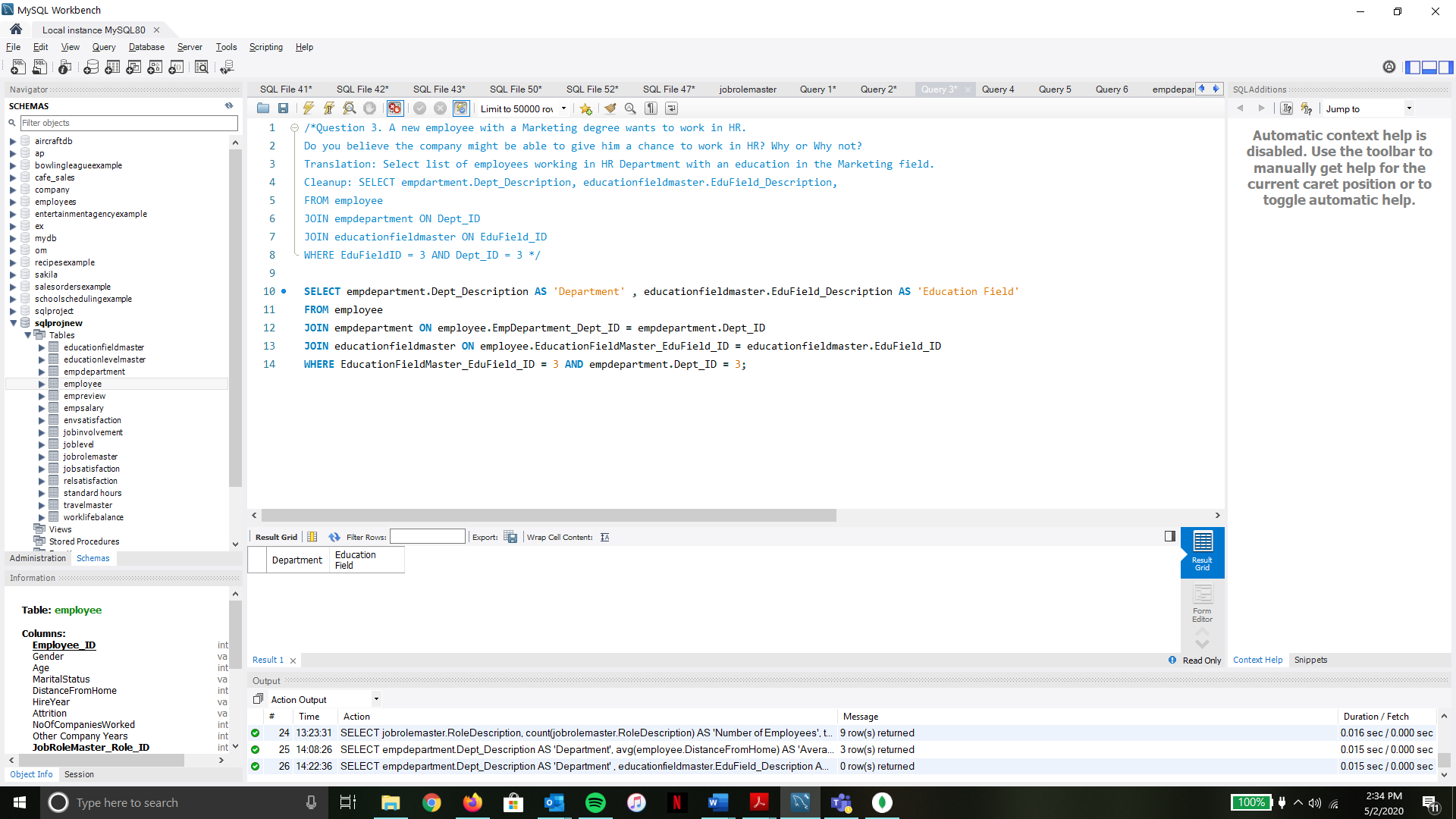
FROM employee

JOIN empdepartment ON Dept\_ID

JOIN educationfieldmaster ON EduField\_ID

WHERE EduFieldID = 3 AND Dept\_ID = 3

#### **Screenshot:**



### **SQL Query 4:**

**Question:** Sales feels that their environment satisfaction score is higher than HR,

but HR job satisfaction score is lower than Research & Development. Are they right?

**Notes:** To verify this, we need to check the environment satisfaction rate and the job satisfaction rate of the employees based on their department. On querying the database, it returns **3 rows**, as there are only 3 departments along with the average environment satisfaction and job satisfaction score. We see that the Sales department’s environment satisfaction score is lower than the HR department’s environment satisfaction score, however, the HR job satisfaction score is lower than the Research & development department’s job satisfaction score. Therefore, Sales is wrong about their assumption of the Environment satisfaction sore, but right about their assumption of the Job Satisfaction score.

**Translation:** Select list of Average Environment Satisfaction and Job Satisfaction scores of all employees grouped by Department

**Cleanup:** SELECT empdepartment.Dept\_Description, avg(envsatisfaction.Env\_ID) , avg(JobSatisfaction\_JobSatis\_ID)

FROM empdaprtment

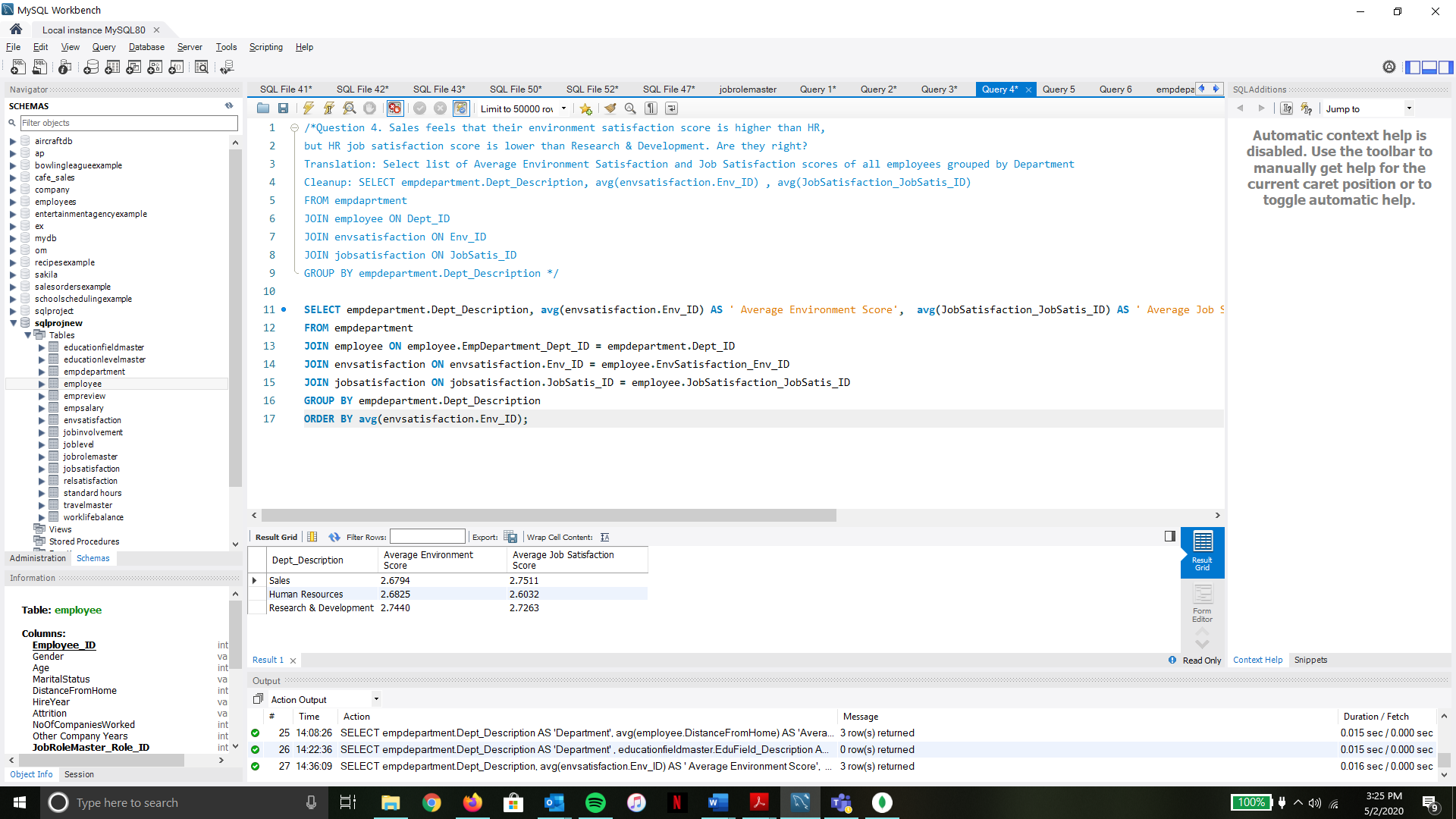
JOIN employee ON Dept\_ID

JOIN envsatisfaction ON Env\_ID

JOIN jobsatisfaction ON JobSatis\_ID

GROUP BY empdepartment.Dept\_Description

#### **Screenshot:**



### **SQL Query 5:**

**Question:** An employee from Life Sciences education field working in Research & Development department has complained to HR that employees with her educational background are paid more in Sales department than in Research & Development. What insight can you provide to prove or disprove that statement?

**Notes:** We need to check if all employees with an education in the Life Sciences field are paid more in the Sales department as compared to the Research and Development department. We choose Monthly income to compare how much they are being paid. On querying the database, it returns **2 rows** as we are comparing only 2 departments with the average monthly income. We see that the average monthly income of the employees in Sales is greater than the employees in Research and Development. Therefore, we can say that the employee was right in complaining as the average income proves that the statement provided by the employee was valid.

**Translation:** Select average salaries of the list of employees with an education background in Life Sciences

working in the Sales and Research & Development department.

**Cleanup:** SELECT empdepartment.Dept\_Description, educationfieldmaster.EduField\_Description, avg(empsalary.MonthlyIncome)

FROM empdepartment

JOIN employee ON Dept\_ID

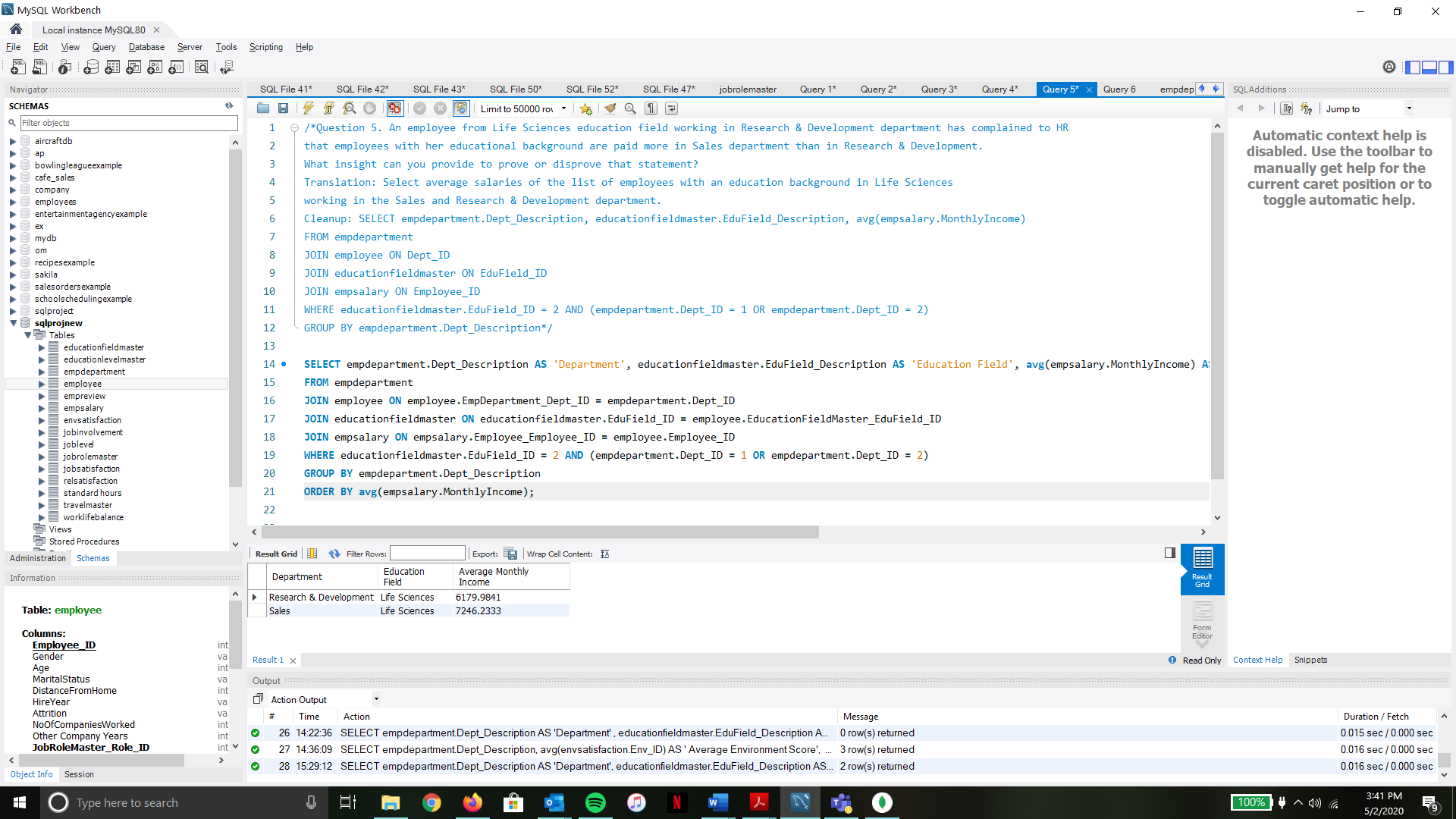
JOIN educationfieldmaster ON EduField\_ID

JOIN empsalary ON Employee\_ID

WHERE educationfieldmaster.EduField\_ID = 2 AND (empdepartment.Dept\_ID = 1 OR empdepartment.Dept\_ID = 2)

GROUP BY empdepartment.Dept\_Description

#### **Screenshot:**



### **SQL Query 6:**

**Question:** A press article in a business magazine has said that at this company, Marital status of women in Research & Development has severely affected their promotion rates. What initial finding can you obtain from the data to help articulate the company's response in this regard?

**Notes:** To check if the claims in the article are true, we need to check the years since a female employee was last promoted in the Research & Development department based on their marital status. We need to exclude the female employees who haven’t completed one year at the organization as we assume that to be eligible for a promotion, an employee needs to have completed at least a year in the organization On querying the database, **3 rows** are returned as there are 3 types of Marital Status. We see that the difference in the years of promotion based on marital status is not that great. We can say that on an average, a female employee in the Research & Development department is promoted between 2.1 years - 2.6 years. This difference is considerable enough to certainly say that their marital status has not ‘severely’ affected their promotion rates.

**Translation:** Select list of all female employees in the Research & Development department who have worked there for at least a year along with the average years since they were last promoted, grouped by their Marital Status.

**Cleanup:** SELECT employee.MaritalStatus, employee.Gender ,empdepartment.Dept\_Description , avg(empreview.YearsSinceLastPromotion)

FROM employee

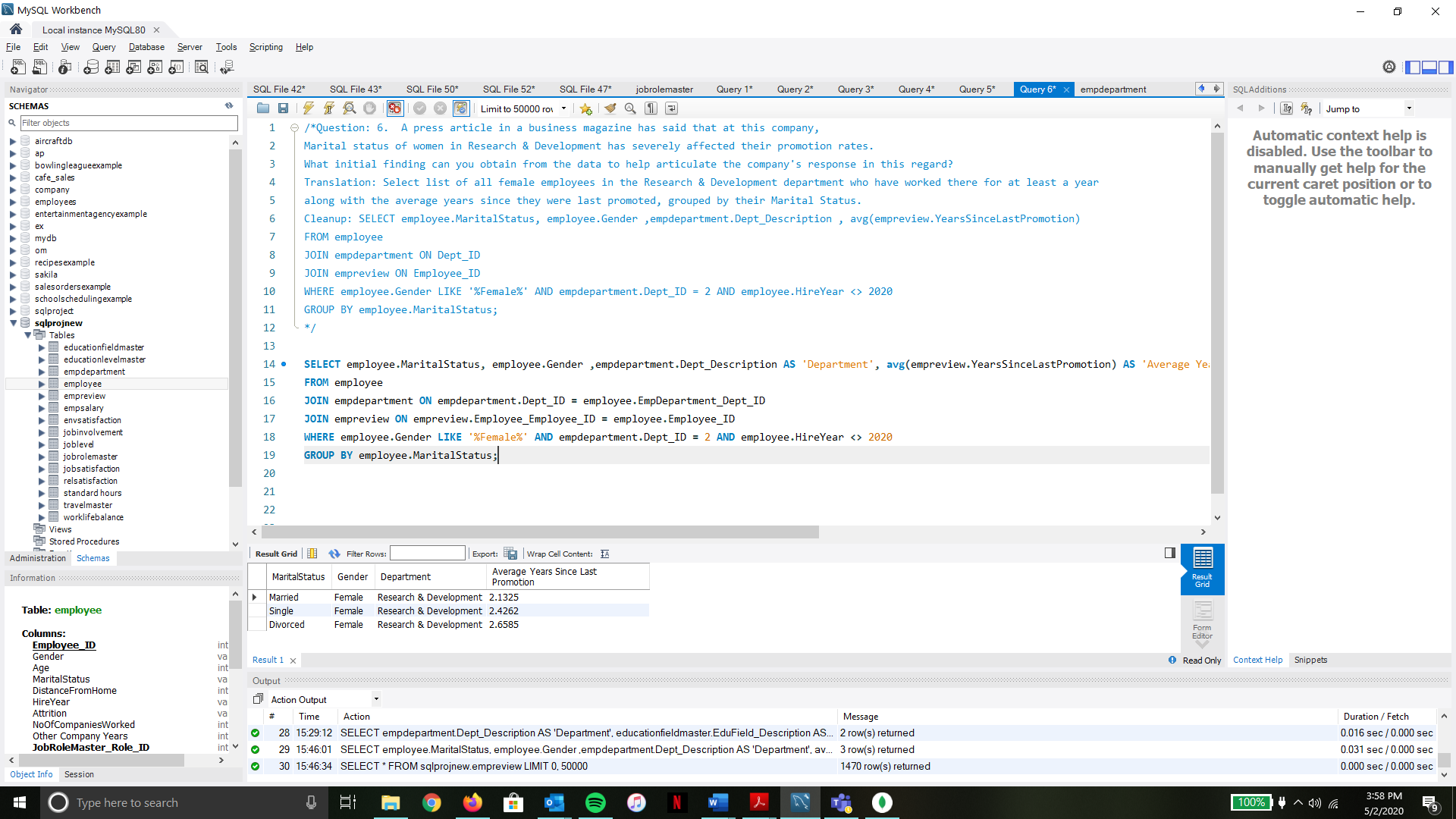
JOIN empdepartment ON Dept\_ID

JOIN empreview ON Employee\_ID

WHERE employee.Gender LIKE '%Female%' AND empdepartment.Dept\_ID = 2 AND employee.HireYear <> 2020

GROUP BY employee.MaritalStatus

#### **Screenshot:**



# **Data Review for MongoDB**

### 

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

1. We eliminate the columns **EmployeeCount** and **Over 18** as they are redundant. Every record represents data for a single employee and whether an employee is Over 18 or not will be determined by the Employee’s age.
2. Since there is only a single collection loaded, there are no relationships to be defined.
3. Since the data set does not include the significance of the Job Levels, we assume the Job Levels to be the following:

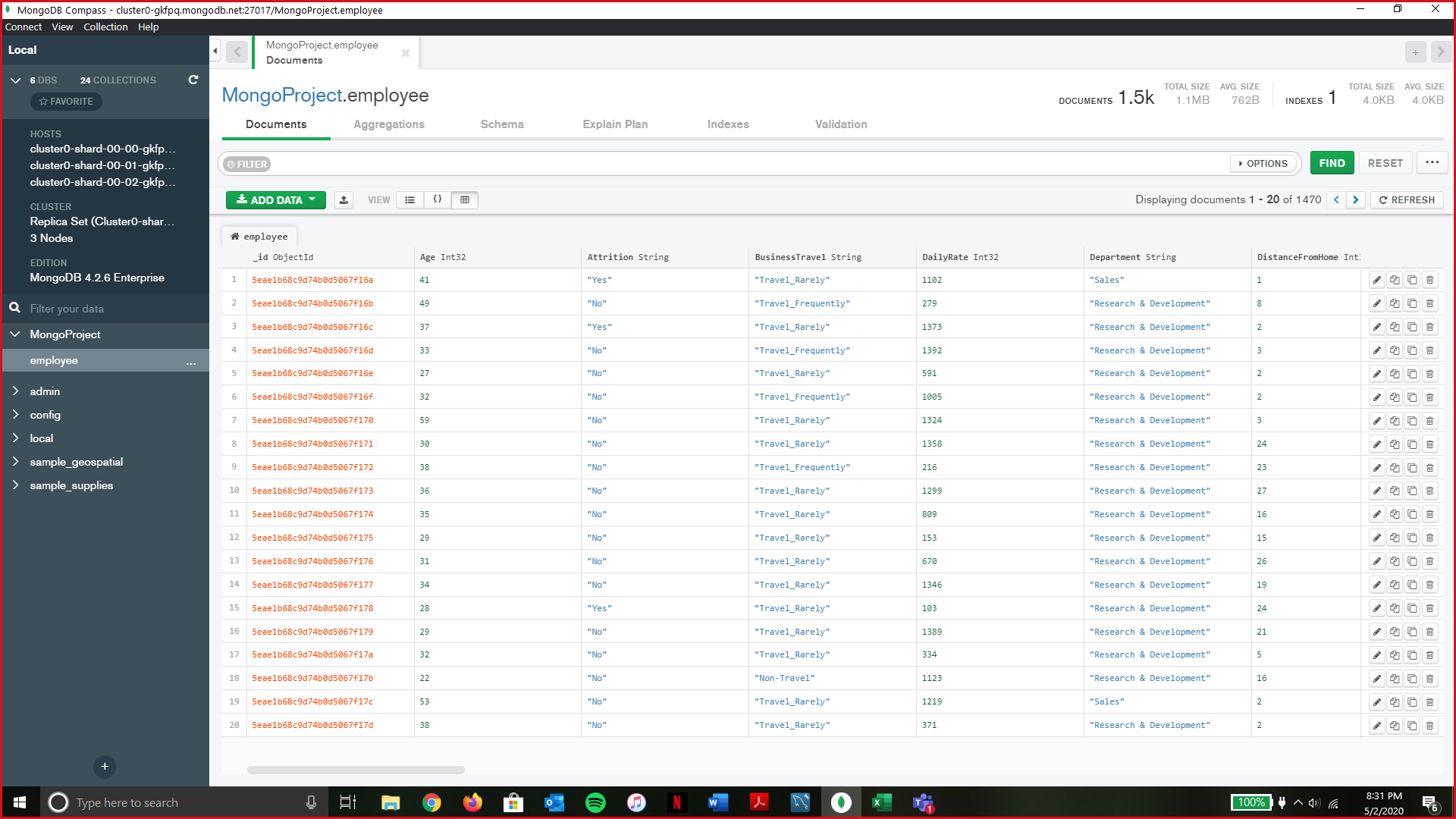
|  |  |
| --- | --- |
| Job Level | Significance |
| 1 | Staff |
| 2 | Senior Staff |
| 3 | Managers & Advisors |
| 4 | Middle Management |
| 5 | Executive Management |

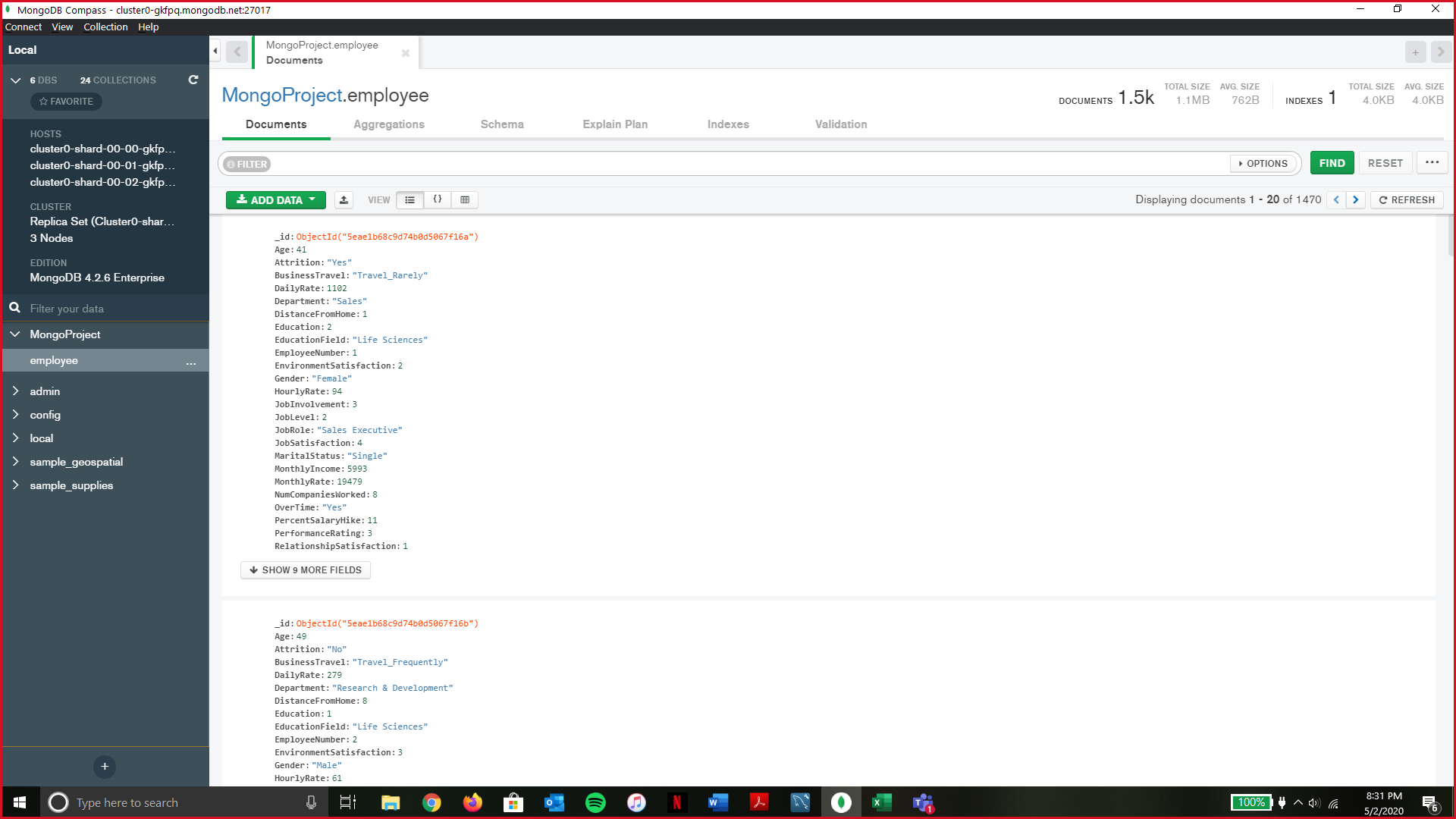
# **Physical Mongo Database**

### **Assumptions/Notes About Data Set**

The entire raw dataset which was provided is loaded into Mongo as a single collection.

### **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** |
| Employee | N/A | 1470 |

# **MongoDB Queries/Code**

### **Query 1**

**Question 1:** A new research scientist who loves to travel joins the firm and is told by HR

that his job role is one of the top two roles in terms of employees that travel frequently.

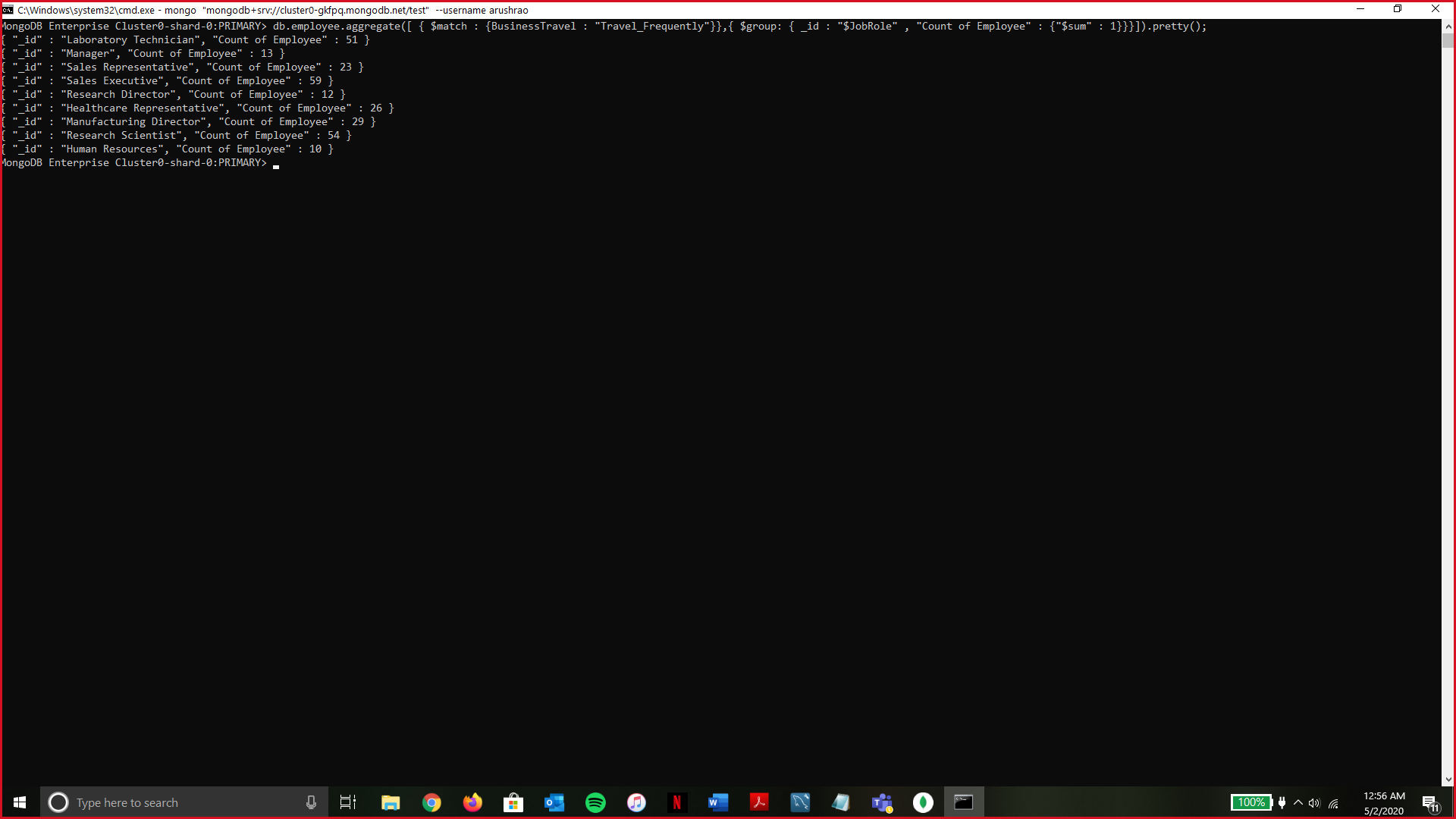
Is HR right in saying so? Why or Why not?

#### **Notes/Comments About MongoDB Query/Code and Results:**

The query returns **9 documents** which are a list of all job roles along with the count of employees in each job role who travel frequently. We see that Sales Executive is the Job Role with the most employees(59) who travel frequently followed by Research Scientist(54). Since HR says that Research Scientist is one of the two job roles with employees travelling frequently, they are right in saying so.

**Translation – Query****:** Using employee collection, utilize the aggregate function to match employees travelling frequently and group them by Job Role.

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



### **Query 2**

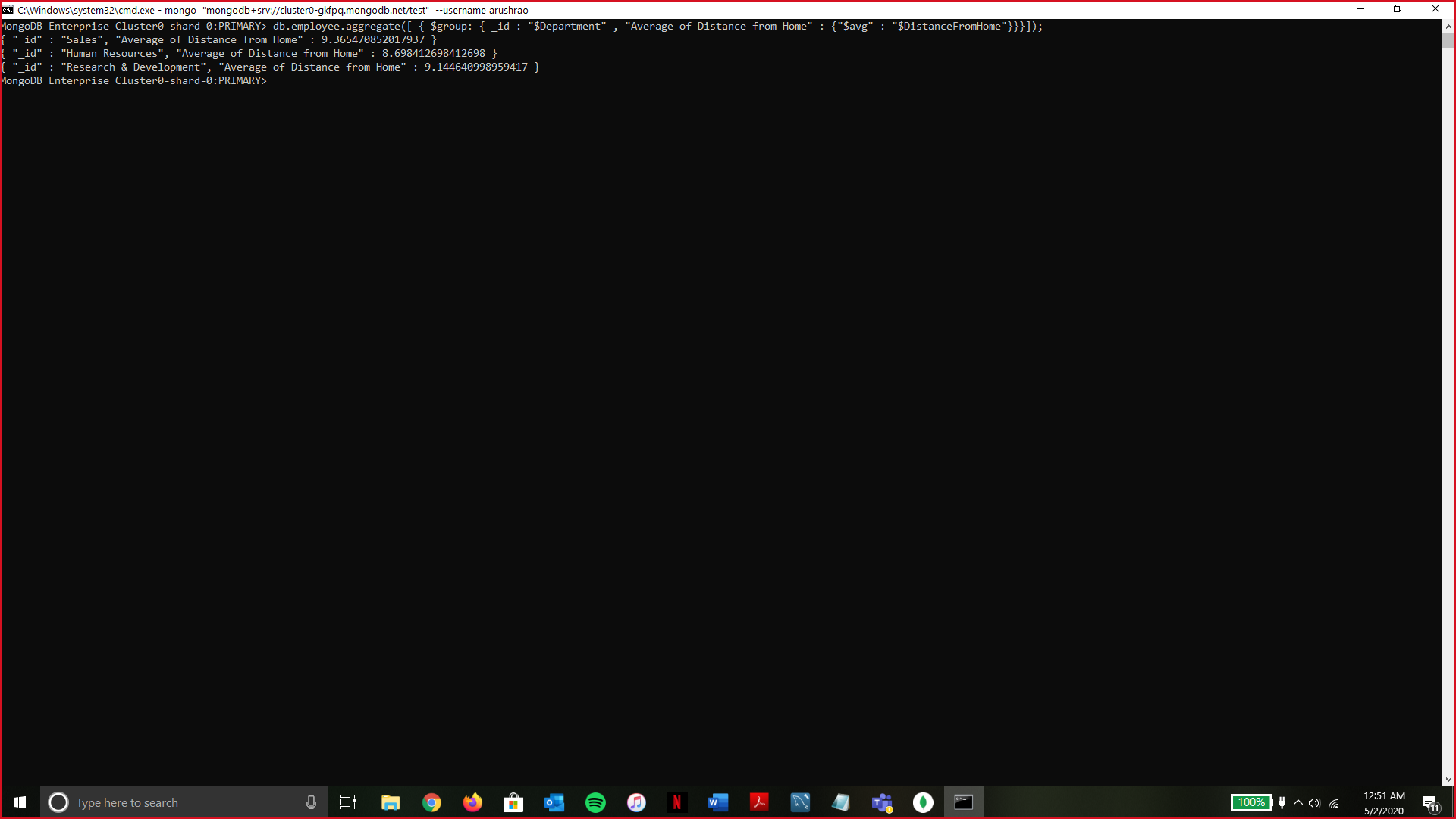
**Question 2:** The company has been paying gas expenses for miles traveled by employees between their home and work. If they want to increase the per mile compensation, which department's employees will gain the least?

#### **Notes/Comments About MongoDB Query/Code and Results:**

The query returns **3 documents** as there are only 3 departments along with the average distance from work of these employees. Based on the result, we can say that employees belonging to the ‘**Human Resources’** department stand to gain the least with an increase in the per mile compensation as the average distance from home of these employees is the least.

**Translation – Query:** Using employee collection, utilize the aggregate function to find average ‘distance from home‘ of all employees grouped by Department.

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



### **Query 3**

**Question 3:** A new employee with a Marketing degree wants to work in HR. Do you believe the company might be able to give him a chance to work in HR? Why or Why not?

#### **Notes/Comments About MongoDB Query/Code and Results:**

The query returns **0 rows** indicating that there are currently no employees in the HR department who also have an education in the Marketing Field. Hence, we can conclude that the company will not be able to give the new employee with a marketing degree, a chance to work in HR.

**Translation – Query:** Using employee collection, utilize the aggregate function to find employees working in the HR department who have an education in the field of marketing.

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

