**BUAN 6320.S01 DATABASE FOUNDATIONS FOR BUSINESS ANALYTICS.**

**SEMESTER PROJECT PHASE-2.**

**DONE BY: VIRAJA NELAGI.**

# Contents

# Relational Data Model

## Assumptions/Notes About Data Entities and Relationships

- Emp_no serves as the primary key for multiple tables and as a foreign key.

- The field of education's varied value is influenced by the Education and Education field columns. The employee's level of education is represented by their education, and their area of specialization within that field is represented by their education field and there are six various education fields in this database.

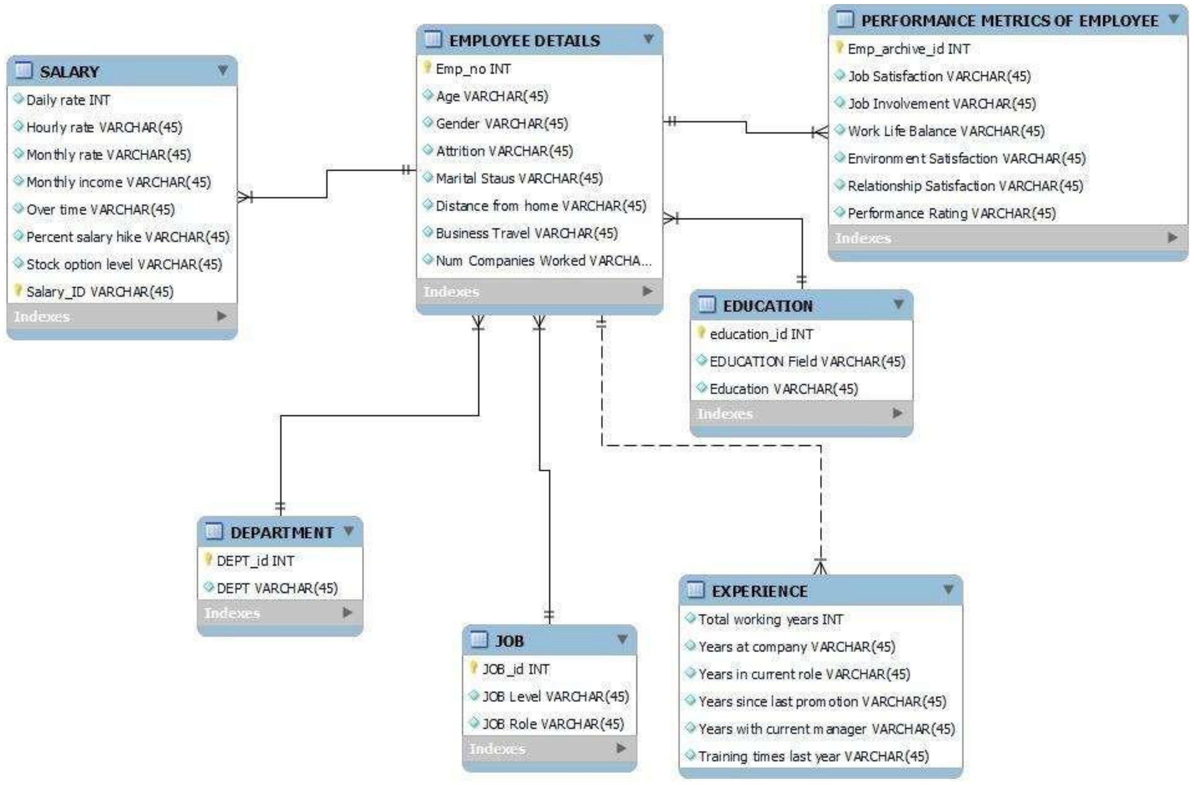
- A many-to-many link between JobLevel and JobRole. At each employment level, there are numerous job roles and there are nine different job roles at different levels in this database.

- The pay rates for various employee categories are represented by the hourly rate, overtime, daily rate, monthly rate, and monthly income. It also covers whether the worker has put in extra time.

- Dep_Id serves as primary key to department table.In the same way Education_ID ,Salary_ID and Job_ID serves as primary keys in their respective education ,salary and job tables.

- All the fields are completely dependent on the primary key of the respective table. There are no partial dependencies, duplicate data, and no transitive dependencies. Hence the model is in 3NF.

## Entity-Relationship Diagram



SALARY
Daily rate INT
Hourly rate VARCHAR(45)
Monthly rate VARCHAR(45)
Monthly income VARCHAR(45)
Over time VARCHAR(45)
Percent salary hike VARCHAR(45)
Stock option level VARCHAR(45)
Salary_ID VARCHAR(45)
Indexes
EMPLOYEE DETAILS
Emp_no INT
Age VARCHAR(45)
Gender VARCHAR(45)
Attrition VARCHAR(45)
Marital Staus VARCHAR(45)
Distance from home VARCHAR(45)
Business Travel VARCHAR(45)
Num Companies Worked VARCHA...
Indexes
PERFORMANCE METRICS OF EMPLOYEE
Emp_archive_id INT
Job Satisfaction VARCHAR(45)
Job Involvement VARCHAR(45)
Work Life Balance VARCHAR(45)
Environment Satisfaction VARCHAR(45)
Relationship Satisfaction VARCHAR(45)
Performance Rating VARCHAR(45)
Indexes
EDUCATION
education_id INT
EDUCATION Field VARCHAR(45)
Education VARCHAR(45)
Indexes
DEPARTMENT
DEPT_id INT
DEPT VARCHAR(45)
Indexes
JOB
JOB_id INT
JOB Level VARCHAR(45)
JOB Role VARCHAR(45)
Indexes
EXPERIENCE
Total working years INT
Years at company VARCHAR(45)
Years in current role VARCHAR(45)
Years since last promotion VARCHAR(45)
Years with current manager VARCHAR(45)
Training times last year VARCHAR(45)

# Physical MySQL Database

## Assumptions/Notes About Data Set

- Since the EmployeeCount column was not helping the model, it was eliminated. Its value remains constant at 1.

- Data in the Over-18 column is unnecessary. The age column can be used to deduce information from the over-18 column. Thus, the column was eliminated.

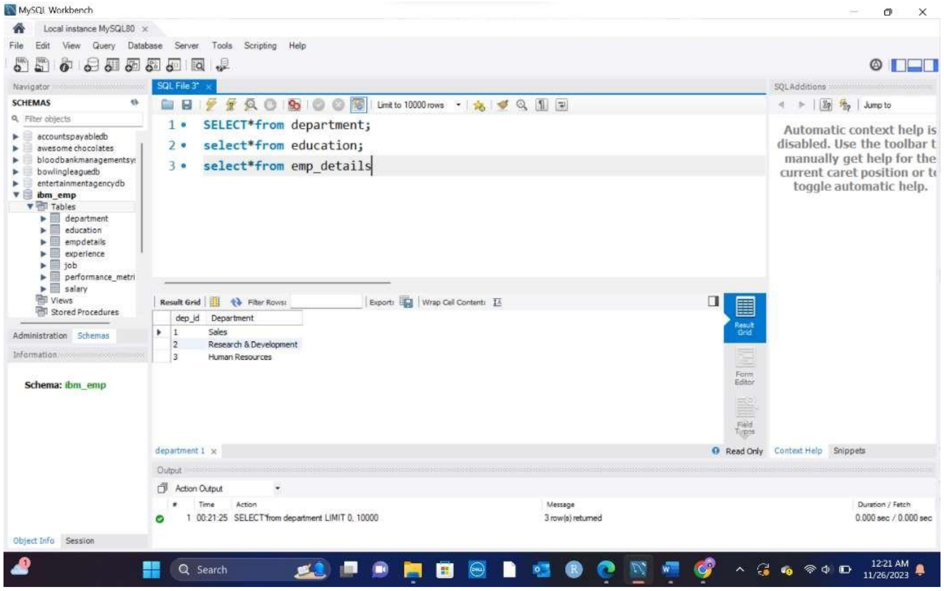
- The StandardHours column was eliminated as every employee's value was the same.

## Data in the Database

| Table Name | Primary Key | Foreign Key | # of Rows in Table |
|---|---|---|---|
| Department | Department_ID | | 3 |
| Education | Education_ID | | 30 |
| Employeedetails | Emp_no | Education_ID<br>Job_ID<br>DepartmentID | 1470 |
| Experience | | Emp_no | 1259 |
| _metrics_employee | Emp_archive ID | ➢ Emp_no | 1470 |
| Salary | Salary ID | ➢ Emp_no | 1470 |
| Job | Job ID | | 26 |

## Screen shot of Physical Database objects.

**Department Table:**

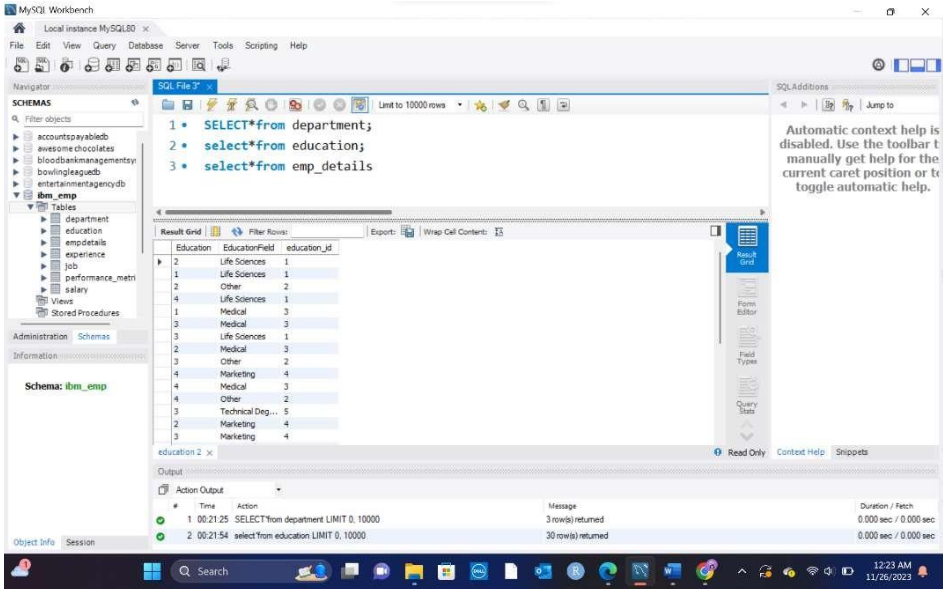


```sql
SELECT*from department;
select*from education;
select*from emp_details
```

| dep_id | Department |
|---|---|
| 1 | Sales |
| 2 | Research & Development |
| 3 | Human Resources |

```
#  Time      Action                                    Message             Duration / Fetch
1  00:21:25  SELECT*from department LIMIT 0, 10000     3 row(s) returned   0.000 sec / 0.000 sec
```

## Education table:

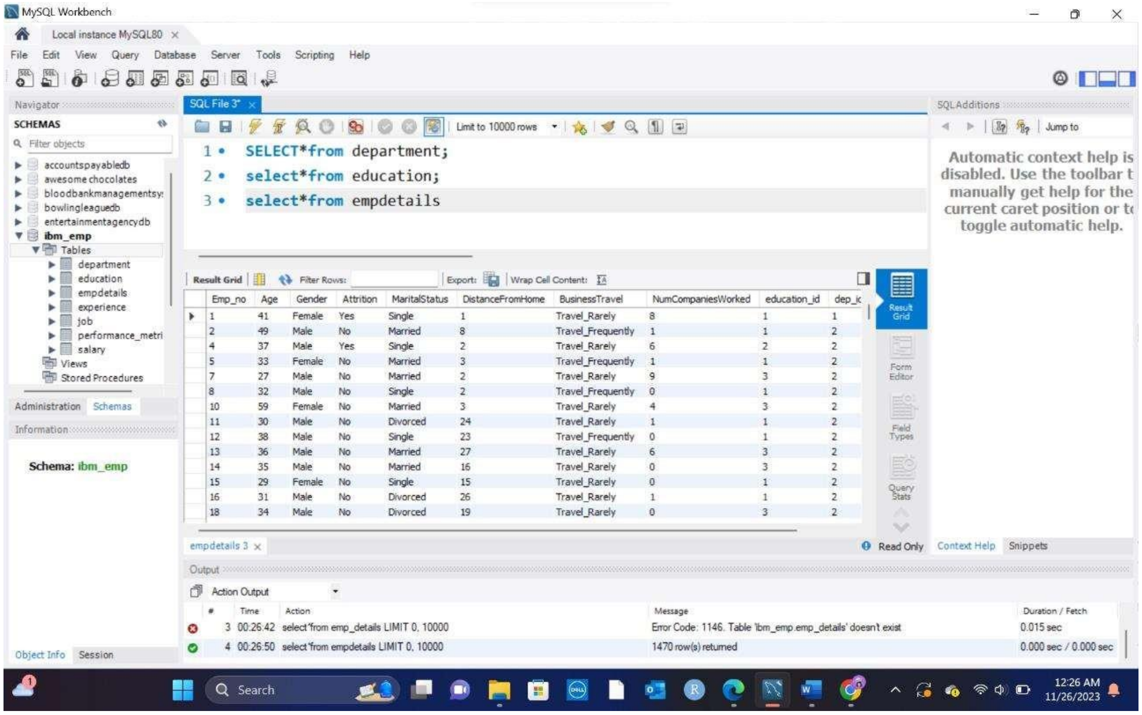


```sql
SELECT*from department;
select*from education;
select*from emp_details
```

| Education | EducationField | education_id |
|---|---|---|
| 2 | Life Sciences | 1 |
| 1 | Life Sciences | 1 |
| 2 | Other | 2 |
| 4 | Life Sciences | 1 |
| 1 | Medical | 3 |
| 3 | Medical | 3 |
| 3 | Life Sciences | 1 |
| 2 | Medical | 3 |
| 3 | Other | 2 |
| 4 | Marketing | 4 |
| 4 | Medical | 3 |
| 4 | Other | 2 |
| 3 | Technical Deg... | 5 |
| 2 | Marketing | 4 |
| 3 | Marketing | 4 |

```
#  Time      Action                                    Message             Duration / Fetch
1  00:21:25  SELECT*from department LIMIT 0, 10000     3 row(s) returned   0.000 sec / 0.000 sec
2  00:21:54  select*from education LIMIT 0, 10000      30 row(s) returned  0.000 sec / 0.000 sec
```

## Employeedetails table:

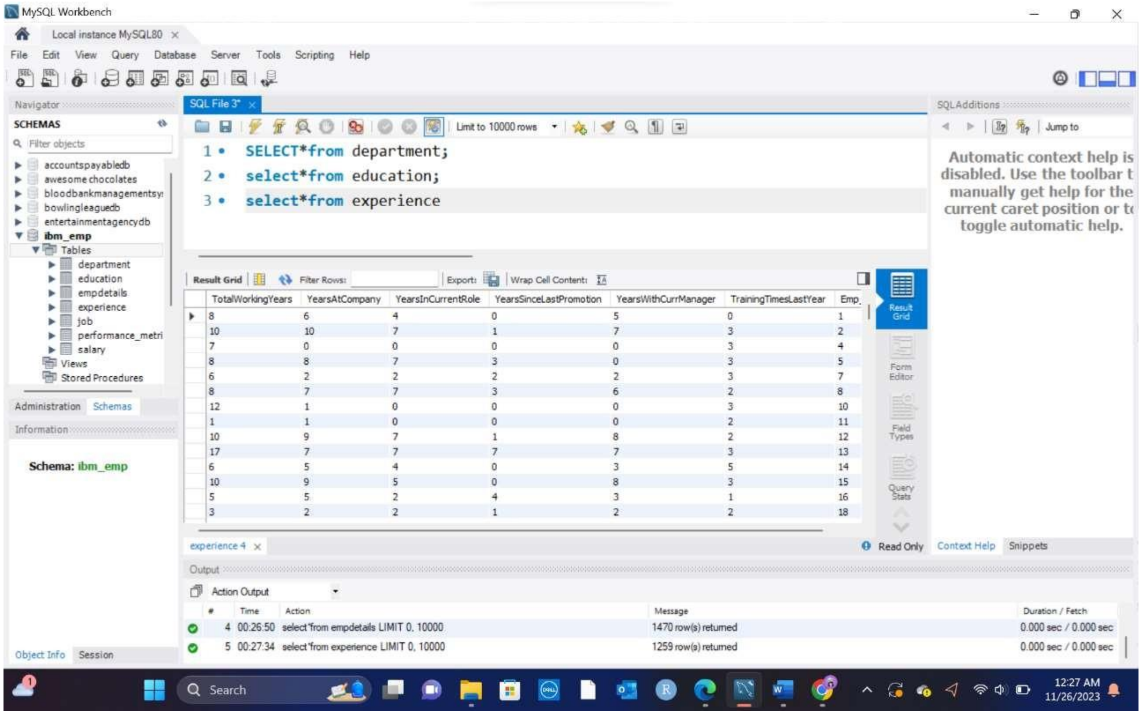


```
SELECT*from department;
select*from education;
select*from empdetails
```

| Emp_no | Age | Gender | Attrition | MaritalStatus | DistanceFromHome | BusinessTravel | NumCompaniesWorked | education_id | dep_ic |
|---|---|---|---|---|---|---|---|---|---|
| 1 | 41 | Female | Yes | Single | 1 | Travel_Rarely | 8 | 1 | 1 |
| 2 | 49 | Male | No | Married | 8 | Travel_Frequently | 1 | 1 | 2 |
| 4 | 37 | Male | Yes | Single | 2 | Travel_Rarely | 6 | 2 | 2 |
| 5 | 33 | Female | No | Married | 3 | Travel_Frequently | 1 | 1 | 2 |
| 7 | 27 | Male | No | Married | 2 | Travel_Rarely | 9 | 3 | 2 |
| 8 | 32 | Male | No | Single | 2 | Travel_Frequently | 0 | 1 | 2 |
| 10 | 59 | Female | No | Married | 3 | Travel_Rarely | 4 | 3 | 2 |
| 11 | 30 | Male | No | Divorced | 24 | Travel_Rarely | 1 | 1 | 2 |
| 12 | 38 | Male | No | Single | 23 | Travel_Frequently | 0 | 1 | 2 |
| 13 | 36 | Male | No | Married | 27 | Travel_Rarely | 6 | 3 | 2 |
| 14 | 35 | Male | No | Married | 16 | Travel_Rarely | 0 | 3 | 2 |
| 15 | 29 | Female | No | Single | 15 | Travel_Rarely | 0 | 1 | 2 |
| 16 | 31 | Male | No | Divorced | 26 | Travel_Rarely | 1 | 1 | 2 |
| 18 | 34 | Male | No | Divorced | 19 | Travel_Rarely | 0 | 3 | 2 |

```
#  Time      Action                                     Message                                                  Duration / Fetch
3  00:26:42  select*from emp_details LIMIT 0, 10000     Error Code: 1146. Table 'ibm_emp.emp_details' doesn't exist   0.015 sec
4  00:26:50  select*from empdetails LIMIT 0, 10000      1470 row(s) returned                                     0.000 sec / 0.000 sec
```

## Experience Table:

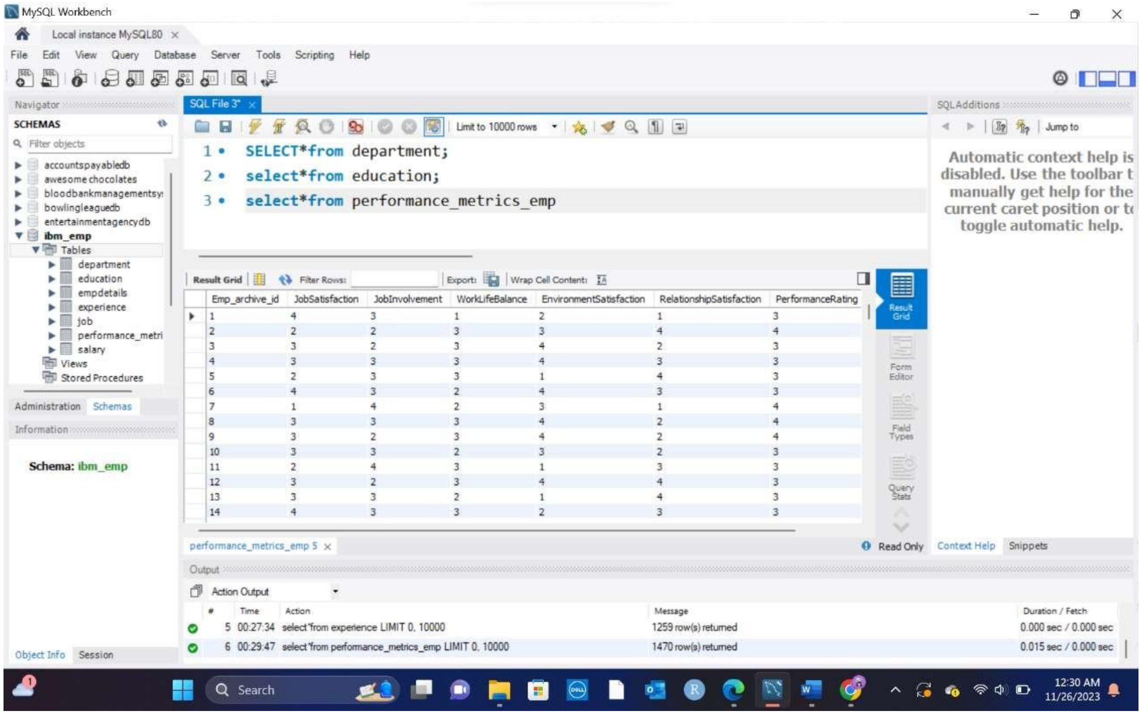


```
SELECT*from department;
select*from education;
select*from experience
```

| TotalWorkingYears | YearsAtCompany | YearsInCurrentRole | YearsSinceLastPromotion | YearsWithCurrManager | TrainingTimesLastYear | Emp_ |
|---|---|---|---|---|---|---|
| 8 | 6 | 4 | 0 | 5 | 0 | 1 |
| 10 | 10 | 7 | 1 | 7 | 3 | 2 |
| 7 | 0 | 0 | 0 | 0 | 3 | 4 |
| 8 | 8 | 7 | 3 | 0 | 3 | 5 |
| 6 | 2 | 2 | 2 | 2 | 3 | 7 |
| 8 | 7 | 7 | 3 | 6 | 2 | 8 |
| 12 | 1 | 0 | 0 | 0 | 3 | 10 |
| 1 | 1 | 0 | 0 | 0 | 2 | 11 |
| 10 | 9 | 7 | 1 | 8 | 2 | 12 |
| 17 | 7 | 7 | 7 | 7 | 3 | 13 |
| 6 | 5 | 4 | 0 | 3 | 5 | 14 |
| 10 | 9 | 5 | 0 | 8 | 3 | 15 |
| 5 | 5 | 2 | 4 | 3 | 1 | 16 |
| 3 | 2 | 2 | 1 | 2 | 2 | 18 |

```
#  Time      Action                                     Message                 Duration / Fetch
4  00:26:50  select*from empdetails LIMIT 0, 10000      1470 row(s) returned    0.000 sec / 0.000 sec
5  00:27:34  select*from experience LIMIT 0, 10000      1259 row(s) returned    0.000 sec / 0.000 sec
```

## Performance_metrics_emp table:

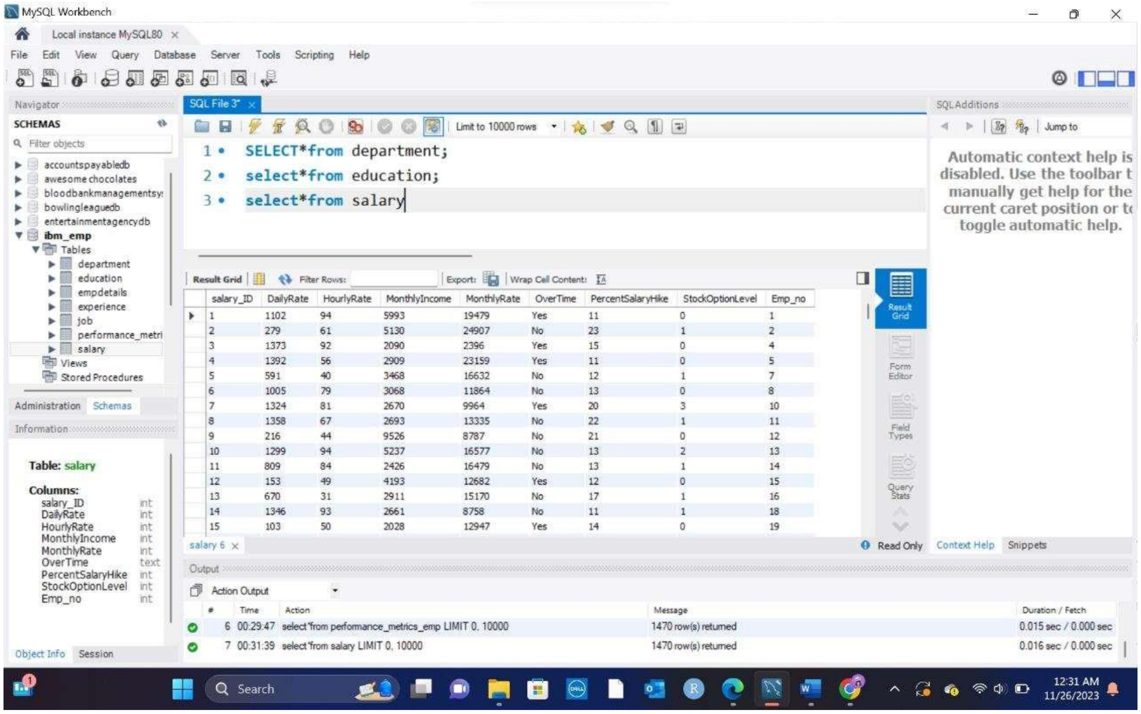


```
SELECT*from department;
select*from education;
select*from performance_metrics_emp
```

| Emp_archive_id | JobSatisfaction | JobInvolvement | WorkLifeBalance | EnvironmentSatisfaction | RelationshipSatisfaction | PerformanceRating |
|---|---|---|---|---|---|---|
| 1 | 4 | 3 | 1 | 2 | 1 | 3 |
| 2 | 2 | 2 | 3 | 3 | 4 | 4 |
| 3 | 3 | 2 | 3 | 4 | 2 | 3 |
| 4 | 3 | 3 | 3 | 4 | 3 | 3 |
| 5 | 2 | 3 | 3 | 1 | 4 | 3 |
| 6 | 4 | 3 | 2 | 4 | 3 | 3 |
| 7 | 1 | 4 | 2 | 3 | 1 | 4 |
| 8 | 3 | 3 | 3 | 4 | 2 | 4 |
| 9 | 3 | 2 | 3 | 4 | 2 | 4 |
| 10 | 3 | 3 | 2 | 3 | 2 | 3 |
| 11 | 2 | 4 | 3 | 1 | 3 | 3 |
| 12 | 3 | 2 | 3 | 4 | 4 | 3 |
| 13 | 3 | 3 | 2 | 1 | 4 | 3 |
| 14 | 4 | 3 | 3 | 2 | 3 | 3 |

```
#  Time      Action                                               Message             Duration / Fetch
5  00:27:34  select*from experience LIMIT 0, 10000                1259 row(s) returned  0.000 sec / 0.000 sec
6  00:29:47  select*from performance_metrics_emp LIMIT 0, 10000   1470 row(s) returned  0.015 sec / 0.000 sec
```

## Salary table:



```
SELECT*from department;
select*from education;
select*from salary
```

| salary_ID | DailyRate | HourlyRate | MonthlyIncome | MonthlyRate | OverTime | PercentSalaryHike | StockOptionLevel | Emp_no |
|---|---|---|---|---|---|---|---|---|
| 1 | 1102 | 94 | 5993 | 19479 | Yes | 11 | 0 | 1 |
| 2 | 279 | 61 | 5130 | 24907 | No | 23 | 1 | 2 |
| 3 | 1373 | 92 | 2090 | 2396 | Yes | 15 | 0 | 4 |
| 4 | 1392 | 56 | 2909 | 23159 | Yes | 11 | 0 | 5 |
| 5 | 591 | 40 | 3468 | 16632 | No | 12 | 1 | 7 |
| 6 | 1005 | 79 | 3068 | 11864 | No | 13 | 0 | 8 |
| 7 | 1324 | 81 | 2670 | 9964 | Yes | 20 | 3 | 10 |
| 8 | 1358 | 67 | 2693 | 13335 | No | 22 | 1 | 11 |
| 9 | 216 | 44 | 9526 | 8787 | No | 21 | 0 | 12 |
| 10 | 1299 | 94 | 5237 | 16577 | No | 13 | 2 | 13 |
| 11 | 809 | 84 | 2426 | 16479 | No | 13 | 1 | 14 |
| 12 | 153 | 49 | 4193 | 12682 | Yes | 12 | 0 | 15 |
| 13 | 670 | 31 | 2911 | 15170 | No | 17 | 1 | 16 |
| 14 | 1346 | 93 | 2661 | 8758 | No | 11 | 1 | 18 |
| 15 | 103 | 50 | 2028 | 12947 | Yes | 14 | 0 | 19 |

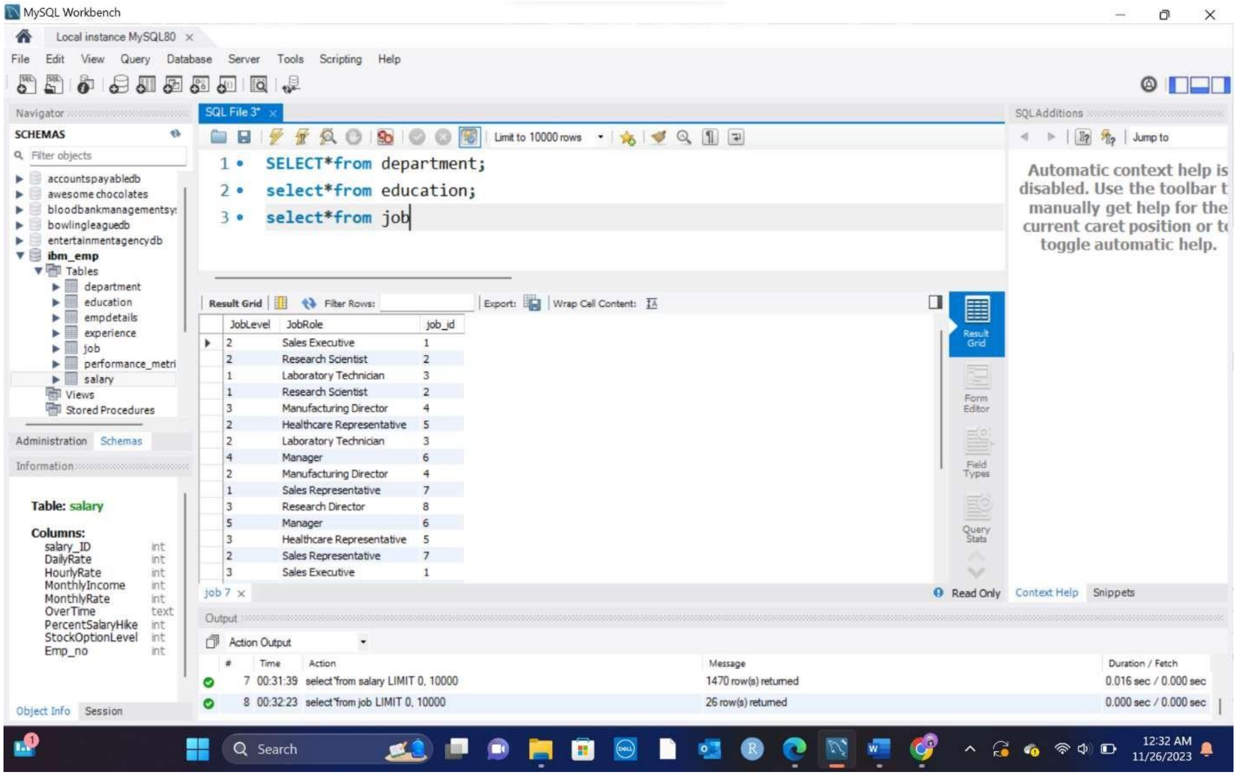
Table: salary

Columns:

| | |
|---|---|
| salary_ID | int |
| DailyRate | int |
| HourlyRate | int |
| MonthlyIncome | int |
| MonthlyRate | int |
| OverTime | text |
| PercentSalaryHike | int |
| StockOptionLevel | int |
| Emp_no | int |

```
#  Time      Action                                               Message             Duration / Fetch
6  00:29:47  select*from performance_metrics_emp LIMIT 0, 10000   1470 row(s) returned  0.015 sec / 0.000 sec
7  00:31:39  select*from salary LIMIT 0, 10000                    1470 row(s) returned  0.016 sec / 0.000 sec
```

## Job Table:



```
SELECT*from department;
select*from education;
select*from job
```

| JobLevel | JobRole | job_id |
|---|---|---|
| 2 | Sales Executive | 1 |
| 2 | Research Scientist | 2 |
| 1 | Laboratory Technician | 3 |
| 1 | Research Scientist | 2 |
| 3 | Manufacturing Director | 4 |
| 2 | Healthcare Representative | 5 |
| 2 | Laboratory Technician | 3 |
| 4 | Manager | 6 |
| 2 | Manufacturing Director | 4 |
| 1 | Sales Representative | 7 |
| 3 | Research Director | 8 |
| 5 | Manager | 6 |
| 3 | Healthcare Representative | 5 |
| 2 | Sales Representative | 7 |
| 3 | Sales Executive | 1 |

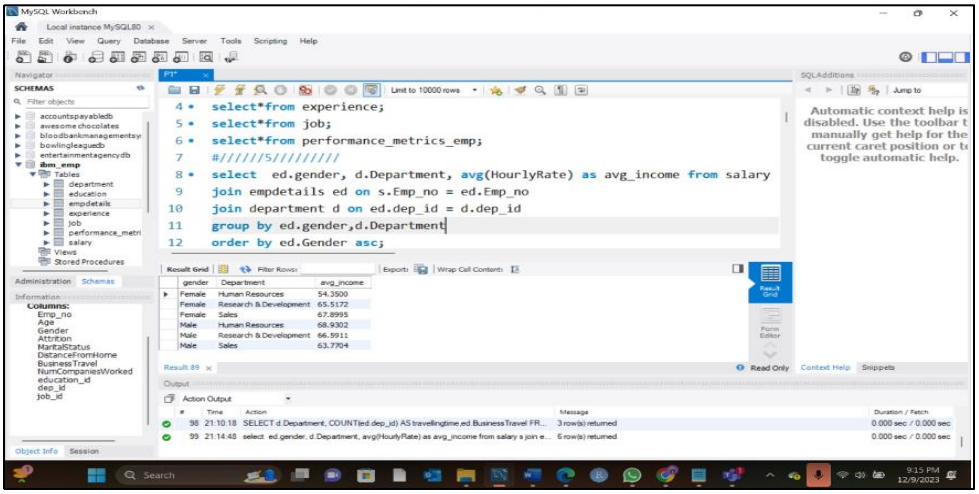
Table: salary

Columns:

| | |
|---|---|
| salary_ID | int |
| DailyRate | int |
| HourlyRate | int |
| MonthlyIncome | int |
| MonthlyRate | int |
| OverTime | text |
| PercentSalaryHike | int |
| StockOptionLevel | int |
| Emp_no | int |

| # | Time | Action | Message | Duration / Fetch |
|---|---|---|---|---|
| 7 | 00:31:39 | select*from salary LIMIT 0, 10000 | 1470 row(s) returned | 0.016 sec / 0.000 sec |
| 8 | 00:32:23 | select*from job LIMIT 0, 10000 | 26 row(s) returned | 0.000 sec / 0.000 sec |

In Phase 2, we imported our data into MongoDB and wrote 3 queries, 5 queries from SQL.



```
select*from experience;
select*from job;
select*from performance_metrics_emp;
#//////5/////////
select  ed.gender, d.Department, avg(HourlyRate) as avg_income from salary
join empdetails ed on s.Emp_no = ed.Emp_no
join department d on ed.dep_id = d.dep_id
group by ed.gender,d.Department
order by ed.Gender asc;
```

| gender | Department | avg_income |
|---|---|---|
| Female | Human Resources | 54.3500 |
| Female | Research & Development | 65.5172 |
| Female | Sales | 67.8995 |
| Male | Human Resources | 68.9302 |
| Male | Research & Development | 66.5911 |
| Male | Sales | 63.7704 |

| # | Time | Action | Message | Duration / Fetch |
|---|---|---|---|---|
| 98 | 21:10:18 | SELECT d.Department, COUNT(ed.dep_id) AS travellingtime,ed.BusinessTravel FR... | 3 row(s) returned | 0.000 sec / 0.000 sec |
| 99 | 21:14:48 | select  ed.gender, d.Department, avg(HourlyRate) as avg_income from salary s join e... | 6 row(s) returned | 0.000 sec / 0.000 sec |

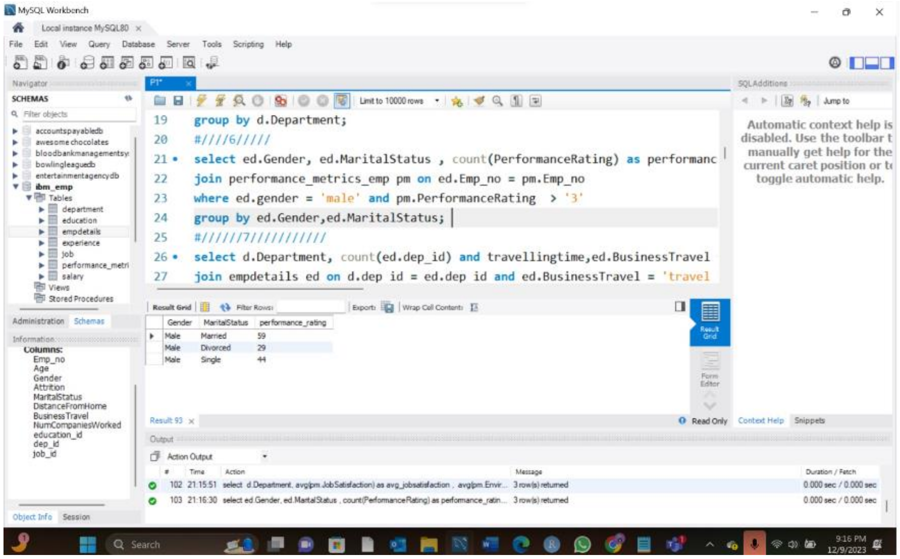
**5) An employee in Sales department has complained to HR saying that females are paid less than males in the company, in all departments. What insight can you provide to prove or disprove that statement?**

**Result: Except in sales department, females are paid less in other departments of Human Resources and Research and Development. Hence, we disapprove the statement.**

**Translation**: Select gender and average of HourlyRate from salary joined on tables with  empdetails table on s.Emp_no = ed.Emp_no joined with  department ~~table~~ on ed.dep_id = d.dep_id. Group~~ed~~ by department and gender. Order by gender in ascending.

**Cleanup**: Select gender ~~and~~ average of hourly rate from salary join~~ed on tables~~ empdetails, salary  matched ~~with~~ empno ~~in~~

salary ~~to~~ empno ~~in~~ emp_details ~~and~~ dep_id ~~in~~ empdetails ~~to~~ dep_id in department. Group~~ed~~ by department ~~and~~ gender, ~~sorted in ascending~~ order ~~on~~ gender.



```
group by d.Department;
#////6/////
select ed.Gender, ed.MaritalStatus , count(PerformanceRating) as performanc
join performance_metrics_emp pm on ed.Emp_no = pm.Emp_no
where ed.gender = 'male' and pm.PerformanceRating  > '3'
group by ed.Gender,ed.MaritalStatus;
#//////7///////////
select d.Department, count(ed.dep_id) and travellingtime,ed.BusinessTravel
join empdetails ed on d.dep id = ed.dep id and ed.BusinessTravel = 'travel
```

| Gender | MaritalStatus | performance_rating |
|---|---|---|
| Male | Married | 59 |
| Male | Divorced | 29 |
| Male | Single | 44 |

```
#    Time      Action                                                                             Message              Duration / Fetch
102  21:15:51  select  d.Department, avg(pm.JobSatisfaction) as avg_jobsatisfaction ,  avg(pm.Envir...  3 row(s) returned    0.000 sec / 0.000 sec
103  21:16:30  select ed.Gender, ed.MaritalStatus , count(PerformanceRating) as performance_ratin...    3 row(s) returned    0.000 sec / 0.000 sec
```

**6)**

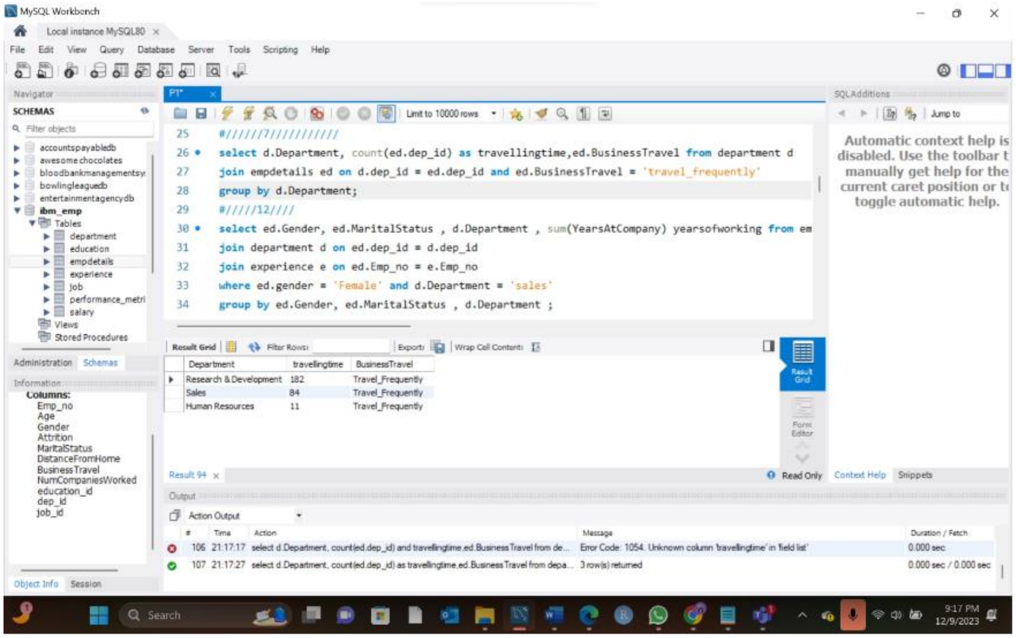
**A press article in a business magazine has said that at this company, married men have higher performance ratings than divorced or single men. What initial finding can you obtain from the data to help articulate the company's response in this regard?**

**Result: Married men have higher performance rating than Divorced and Single set of males. Hence, The press article's claim is correct.**

**Translation**: Select Gender, MaritalStatus, count of PerformanceReporting joined table performance_metrics_emp on ed.Emp_no=pm.Emp_no where ed.gender=male and pm.PerformanceRating > 3. Group by Gender and MaritalStatus.

**Cleanup**:  Select Gender, MaritalStatus, count ~~of~~ PerformanceReporting joined ~~table~~ performance_metrics_emp on ed.Emp_no=pm.Emp_no where ed.gender=male and pm.PerformanceRating > 3. Group by Gender and MaritalStatus.

7)



```
#//////7/////////////
select d.Department, count(ed.dep_id) as travellingtime,ed.BusinessTravel from department d
join empdetails ed on d.dep_id = ed.dep_id and ed.BusinessTravel = 'travel_frequently'
group by d.Department;
#/////12////
select ed.Gender, ed.MaritalStatus , d.Department , sum(YearsAtCompany) yearsofworking from em
join department d on ed.dep_id = d.dep_id
join experience e on ed.Emp_no = e.Emp_no
where ed.gender = 'Female' and d.Department = 'sales'
group by ed.Gender, ed.MaritalStatus , d.Department ;
```

| Department | travellingtime | BusinessTravel |
|---|---|---|
| Research & Development | 182 | Travel_Frequently |
| Sales | 84 | Travel_Frequently |
| Human Resources | 11 | Travel_Frequently |

```
#    Time      Action                                                                  Message                                                       Duration / Fetch
106  21:17:17  select d.Department, count(ed.dep_id) and travellingtime,ed.BusinessTravel from de...  Error Code: 1054. Unknown column 'travellingtime' in 'field list'  0.000 sec
107  21:17:27  select d.Department, count(ed.dep_id) as travellingtime,ed.BusinessTravel from depa...  3 row(s) returned                                             0.000 sec / 0.000 sec
```

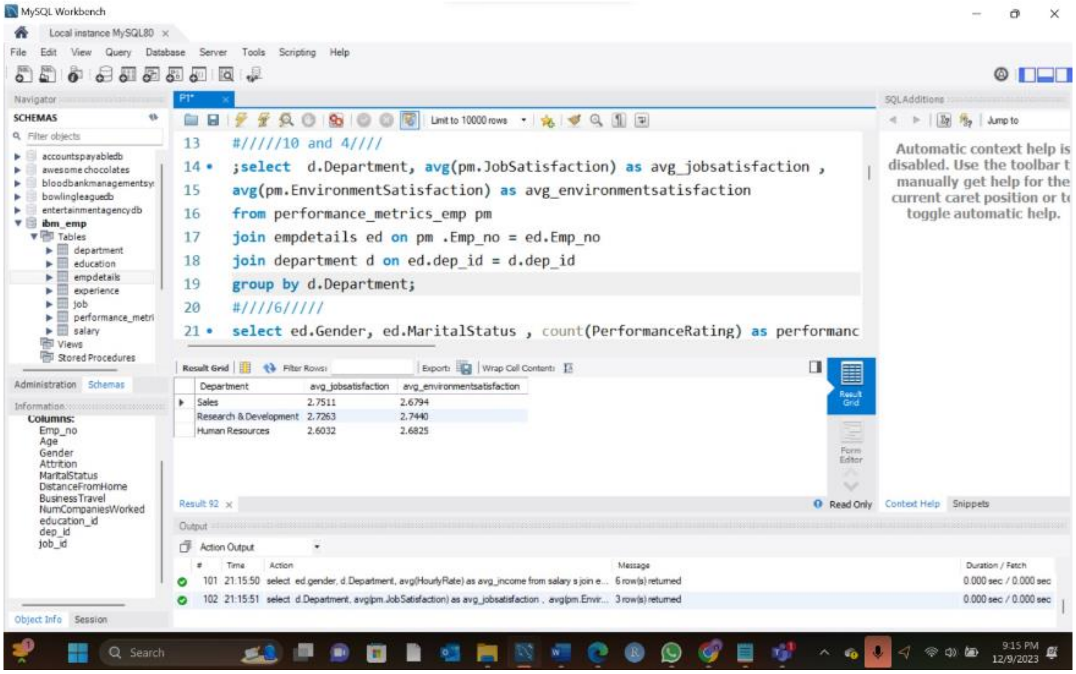
**If the company wants to cut travel costs, which department should the company focus on?**

**Result: The company should focus on Research and Development department to cut down their costs as the amount of people who travel frequently are higher than other departments.**

**Translation**: Select Department, count of dep_id from department joined table empdetails on dep_id = ed.dep_id and ed.BusinessTravel = ‘travel_frequently’. Grouped by department.

.

**Cleanup**: Select Department, count of dep_id from department joined ~~table~~ empdetails on dep_id = ed.dep_id and ed.BusinessTravel = ‘travel_frequently’. Group~~ed~~ by department.



```
#/////10 and 4////
;select  d.Department, avg(pm.JobSatisfaction) as avg_jobsatisfaction ,
avg(pm.EnvironmentSatisfaction) as avg_environmentsatisfaction
from performance_metrics_emp pm
join empdetails ed on pm .Emp_no = ed.Emp_no
join department d on ed.dep_id = d.dep_id
group by d.Department;
#////6/////
select ed.Gender, ed.MaritalStatus , count(PerformanceRating) as performanc
```

| Department | avg_jobsatisfaction | avg_environmentsatisfaction |
|---|---|---|
| Sales | 2.7511 | 2.6794 |
| Research & Development | 2.7263 | 2.7440 |
| Human Resources | 2.6032 | 2.6825 |

```
#    Time      Action                                                                          Message            Duration / Fetch
101  21:15:50  select  ed.gender, d.Department, avg(HourlyRate) as avg_income from salary s join e...  6 row(s) returned  0.000 sec / 0.000 sec
102  21:15:51  select  d.Department, avg(pm.JobSatisfaction) as avg_jobsatisfaction ,  avg(pm.Envir...  3 row(s) returned  0.000 sec / 0.000 sec
```

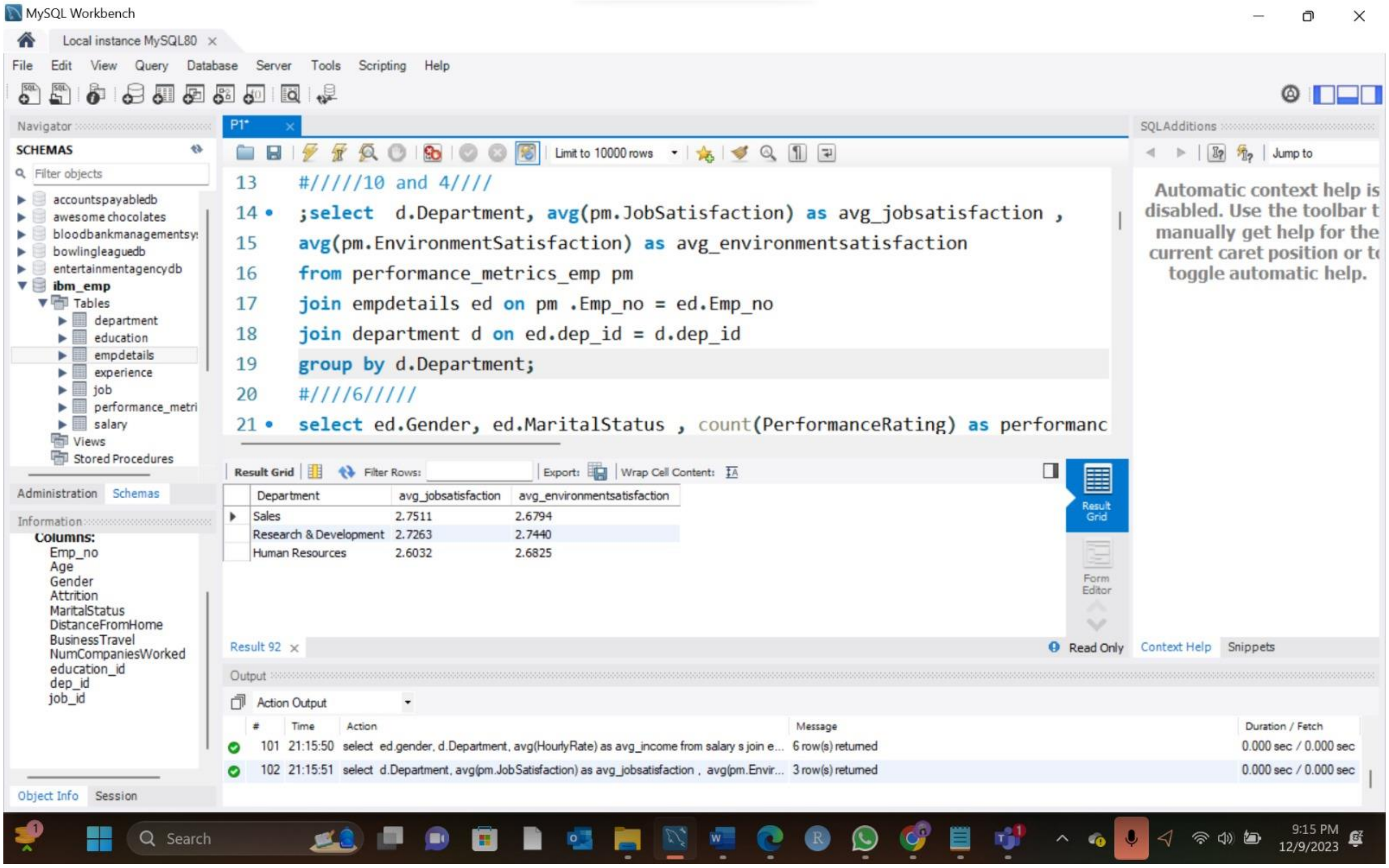
**4 – From the output of the query it is evident that the Research and Development department have more average job**

**satisfaction.**

**10- From the output of the query it is evident that the environment satisfaction score of HR is higher than sales but HR job satisfaction score is lower than Research & Development. The HR department is right.**

**Translation**: Select Department, average of JobSatisfaction, average of EnvironmentSatisfaction from performance_metrics_emp joined table empdetails on pm.Emp_no=ed.Emp_no and joined table department on ed.dep_id = d.dep_id. Grouped by department.

**Cleanup**: Select Department, average of JobSatisfaction, average of EnvironmentSatisfaction from performance_metrics_emp join~~ed~~ ~~table~~ empdetails on pm.Emp_no=ed.Emp_no and join~~ed~~ ~~table~~ department on ed.dep_id = d.dep_id. Group~~ed~~ by department.



```
#/////10 and 4////
;select  d.Department, avg(pm.JobSatisfaction) as avg_jobsatisfaction ,
avg(pm.EnvironmentSatisfaction) as avg_environmentsatisfaction
from performance_metrics_emp pm
join empdetails ed on pm .Emp_no = ed.Emp_no
join department d on ed.dep_id = d.dep_id
group by d.Department;
#////6/////
select ed.Gender, ed.MaritalStatus , count(PerformanceRating) as performanc
```

| Department | avg_jobsatisfaction | avg_environmentsatisfaction |
|---|---|---|
| Sales | 2.7511 | 2.6794 |
| Research & Development | 2.7263 | 2.7440 |
| Human Resources | 2.6032 | 2.6825 |

```
#    Time      Action                                                                          Message            Duration / Fetch
101  21:15:50  select  ed.gender, d.Department, avg(HourlyRate) as avg_income from salary s join e...  6 row(s) returned  0.000 sec / 0.000 sec
102  21:15:51  select  d.Department, avg(pm.JobSatisfaction) as avg_jobsatisfaction ,  avg(pm.Envir...  3 row(s) returned  0.000 sec / 0.000 sec
```

# Data Review for MongoDB

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

Collection Structure: MongoDB stores data in a hierarchical format, complex and nested data structures are possible. The only collection in my database with the name "ibm_emp" is called "employee."

Data organization: The original MySQL database's tables are all combined into one collection by the Employee collection. This indicates that all the attributes from the original tables are present in every document in the collection.

Attributes: The attributes Over18 and EmployeeCount are discarded while loading the dataset into the MongodB as

- Over18 can be easily found by performing logic {$gte:18}
- Employeecount is just counting the same employee alone and has no meaning.

Relationships: There is only one collection.

# Physical Mongo Database

## Assumptions/Notes About Data Set

Data completeness: It is assumed that the data set is comprehensive and includes all pertinent details regarding the employees' job descriptions, pay scales, levels of job satisfaction, and other relevant characteristics.

Data consistency: It is assumed that the data set is consistent, which means that each attribute has a set format, and that the data is constant across all records.

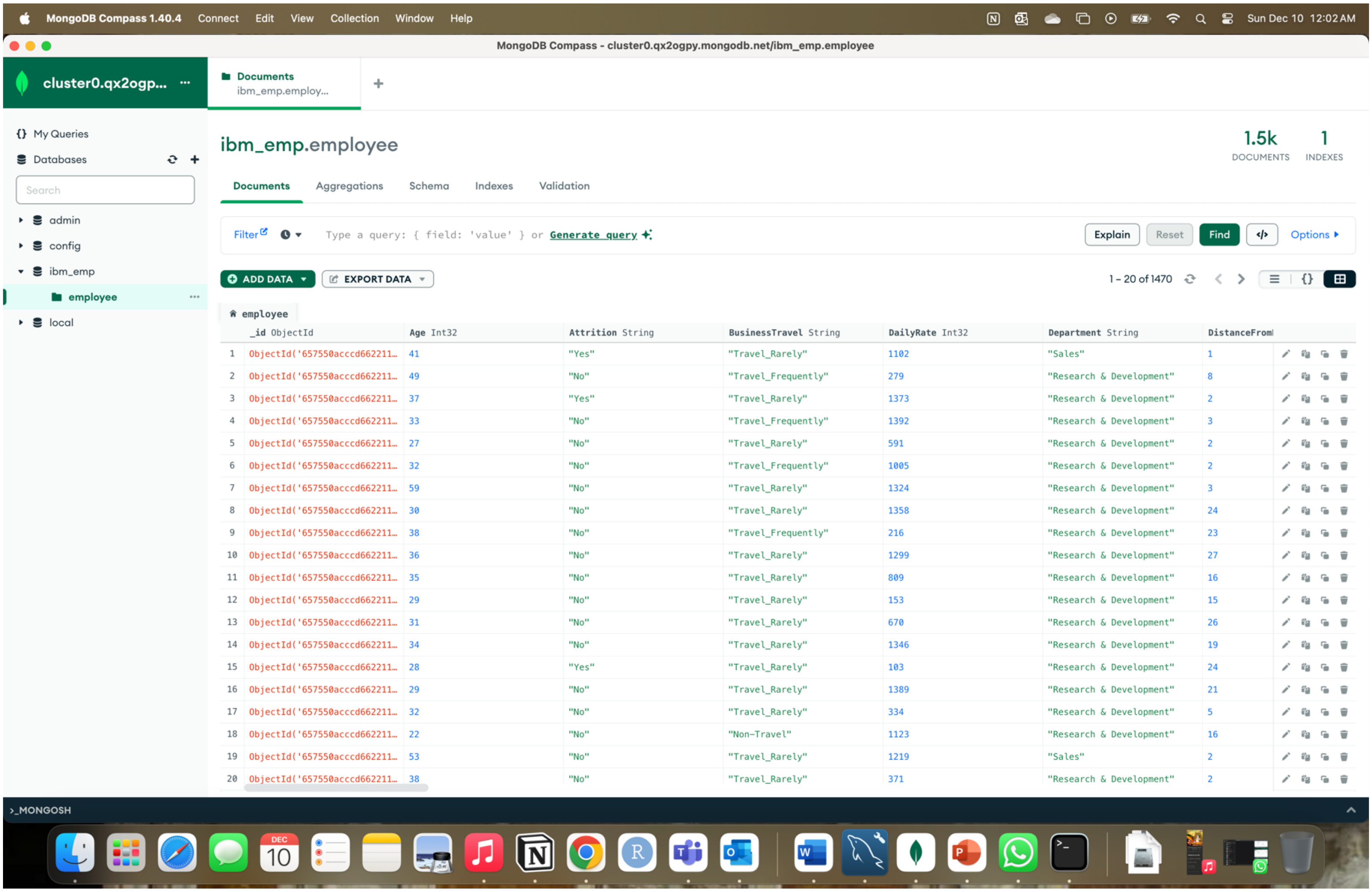
Data reliability: It is assumed that the data set is trustworthy, and that the data can be relied upon for making decisions.

Timeframe: Any analysis or inferences made from the data should consider the time period for which the data set is assumed to be representative.

Data integrity: The extraction, transformation, and loading (ETL) process are assumed to have preserved the data set's integrity, preventing any data loss or corruption.

Data confidentiality: It is assumed that the data set will be handled with the proper degree of confidentiality and that any sensitive employee information will be adequately safeguarded.

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



ibm_emp.employee

1.5k DOCUMENTS

1 INDEXES

1 – 20 of 1470

employee

| | _id ObjectId | Age Int32 | Attrition String | BusinessTravel String | DailyRate Int32 | Department String | DistanceFrom |
|---|---|---|---|---|---|---|---|
| 1 | ObjectId('657550acccd662211… | 41 | "Yes" | "Travel_Rarely" | 1102 | "Sales" | 1 |
| 2 | ObjectId('657550acccd662211… | 49 | "No" | "Travel_Frequently" | 279 | "Research & Development" | 8 |
| 3 | ObjectId('657550acccd662211… | 37 | "Yes" | "Travel_Rarely" | 1373 | "Research & Development" | 2 |
| 4 | ObjectId('657550acccd662211… | 33 | "No" | "Travel_Frequently" | 1392 | "Research & Development" | 3 |
| 5 | ObjectId('657550acccd662211… | 27 | "No" | "Travel_Rarely" | 591 | "Research & Development" | 2 |
| 6 | ObjectId('657550acccd662211… | 32 | "No" | "Travel_Frequently" | 1005 | "Research & Development" | 2 |
| 7 | ObjectId('657550acccd662211… | 59 | "No" | "Travel_Rarely" | 1324 | "Research & Development" | 3 |
| 8 | ObjectId('657550acccd662211… | 30 | "No" | "Travel_Rarely" | 1358 | "Research & Development" | 24 |
| 9 | ObjectId('657550acccd662211… | 38 | "No" | "Travel_Frequently" | 216 | "Research & Development" | 23 |
| 10 | ObjectId('657550acccd662211… | 36 | "No" | "Travel_Rarely" | 1299 | "Research & Development" | 27 |
| 11 | ObjectId('657550acccd662211… | 35 | "No" | "Travel_Rarely" | 809 | "Research & Development" | 16 |
| 12 | ObjectId('657550acccd662211… | 29 | "No" | "Travel_Rarely" | 153 | "Research & Development" | 15 |
| 13 | ObjectId('657550acccd662211… | 31 | "No" | "Travel_Rarely" | 670 | "Research & Development" | 26 |
| 14 | ObjectId('657550acccd662211… | 34 | "No" | "Travel_Rarely" | 1346 | "Research & Development" | 19 |
| 15 | ObjectId('657550acccd662211… | 28 | "Yes" | "Travel_Rarely" | 103 | "Research & Development" | 24 |
| 16 | ObjectId('657550acccd662211… | 29 | "No" | "Travel_Rarely" | 1389 | "Research & Development" | 21 |
| 17 | ObjectId('657550acccd662211… | 32 | "No" | "Travel_Rarely" | 334 | "Research & Development" | 5 |
| 18 | ObjectId('657550acccd662211… | 22 | "No" | "Non-Travel" | 1123 | "Research & Development" | 16 |
| 19 | ObjectId('657550acccd662211… | 53 | "No" | "Travel_Rarely" | 1219 | "Sales" | 2 |
| 20 | ObjectId('657550acccd662211… | 38 | "No" | "Travel_Rarely" | 371 | "Research & Development" | 2 |

## Data in the Database

| Collection Name | Relationships With Other Collections (if any) | # of Documents in Collection |
|---|---|---|
| ibm_emp | | 1470 |

# MongoDB Queries/Code

Pick 3 SQL queries and write them in MongoDB

## Question 3

A new employee from a Medical-related education field wants to work in Sales. Do you believe the company might be able to give her a chance to work in Sales? Why or why not?

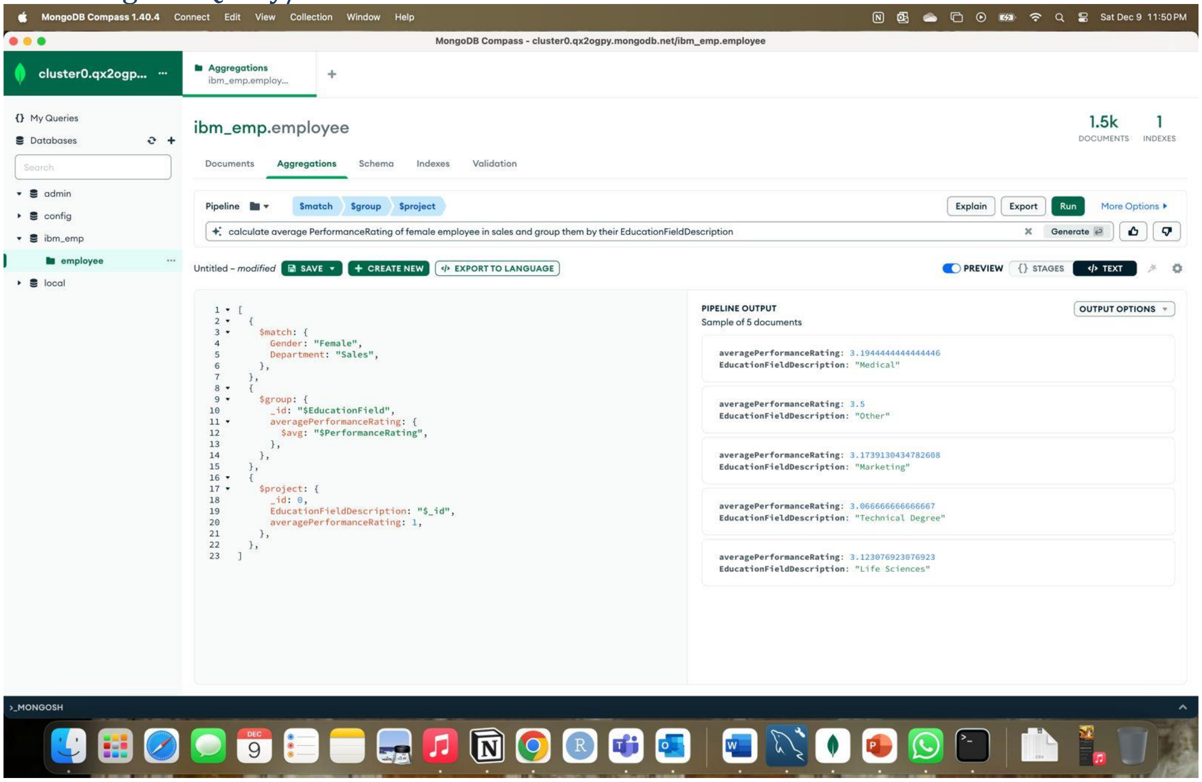
## Notes/Comments About MongoDB Query/Code and Results (Include # of Documents in Result)

Based on their educational backgrounds, sales department employees' average performance rating has been determined. The medical sector ranks second among all education fields, while other fields make up the first place. As a result, the corporation may offer her a position in sales.

## Translation

Aggregate the emp_attrition collection the Database. In the first stage filter the documents using $match operator to and include only those whose Gender is set to “Female” and Department is set to “Sales”. In the second stage group the filtered documents using $group operator, the grouping key is composed by “EducationField”. Calculate the average value of PerformanceRating for each grouped documents assign it to the field avgPerformanceRating in the result.

## Screen Shot of MongoDB Query/Code and Results



calculate average PerformanceRating of female employee in sales and group them by their EducationFieldDescription

```
[
  {
    $match: {
      Gender: "Female",
      Department: "Sales",
    },
  },
  {
    $group: {
      _id: "$EducationField",
      averagePerformanceRating: {
        $avg: "$PerformanceRating",
      },
    },
  },
  {
    $project: {
      _id: 0,
      EducationFieldDescription: "$_id",
      averagePerformanceRating: 1,
    },
  },
]
```

Sample of 5 documents

```
averagePerformanceRating: 3.1944444444444446
EducationFieldDescription: "Medical"

averagePerformanceRating: 3.5
EducationFieldDescription: "Other"

averagePerformanceRating: 3.1739130434782608
EducationFieldDescription: "Marketing"

averagePerformanceRating: 3.066666666666667
EducationFieldDescription: "Technical Degree"

averagePerformanceRating: 3.123076923076923
EducationFieldDescription: "Life Sciences"
```

## Mongo Query 2

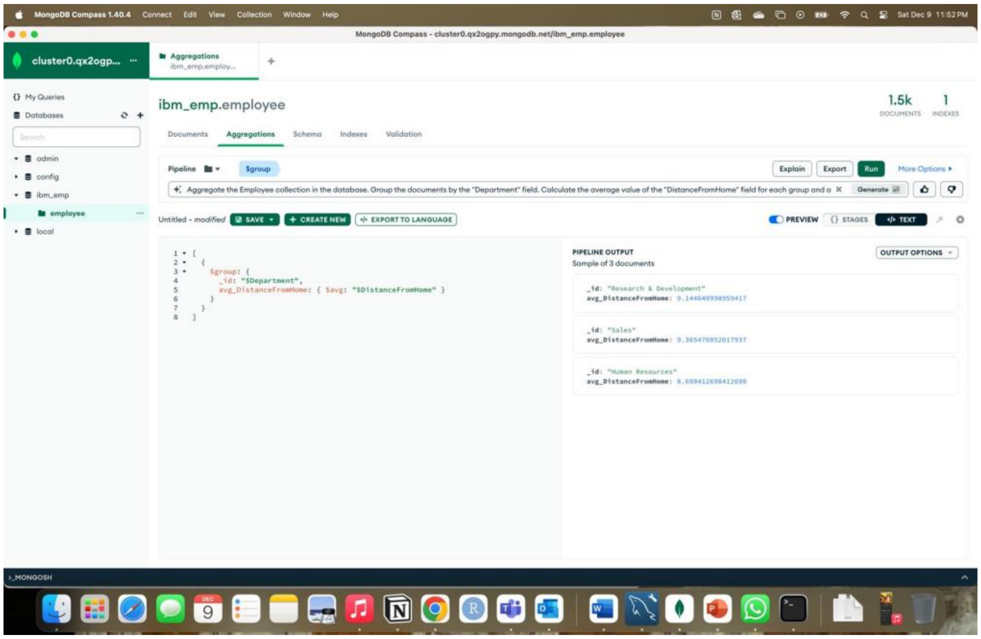
## Question 8
## Screen Shot of MongoDB Query/Code and Results

### Question 8, from the list
### Notes/Comments About MongoDB Query/Code and Results (Include # of Documents in Result)

After calculating the average distance from home for each department using the Mongo query below, the employees in the "Sales" department will receive the highest wage. The reason for this is that the sales section is the furthest away from home. The average distance from home for each department is listed in three papers in the result set.

## Translation

Aggregate the emp_attrition collection in the database. Group the documents by the "Department" field. Calculate the average value of the "DistanceFromHome" field for each group and assign it to the field "avg_DistanceFromHome" in the result.



MongoDB Compass - cluster0.qx2ogpy.mongodb.net/ibm_emp.employee

ibm_emp.employee

1.5k DOCUMENTS 1 INDEXES

Pipeline $group

Aggregate the Employee collection in the database. Group the documents by the "Department" field. Calculate the average value of the "DistanceFromHome" field for each group and a

```
[
  {
    $group: {
      _id: "$Department",
      avg_DistanceFromHome: { $avg: "$DistanceFromHome" }
    }
  }
]
```

PIPELINE OUTPUT
Sample of 3 documents

```
_id: "Research & Development"
avg_DistanceFromHome: 9.144640998959417

_id: "Sales"
avg_DistanceFromHome: 9.365470852017937

_id: "Human Resources"
avg_DistanceFromHome: 8.698412698412698
```

# Mongo Query 3

## Question 11

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

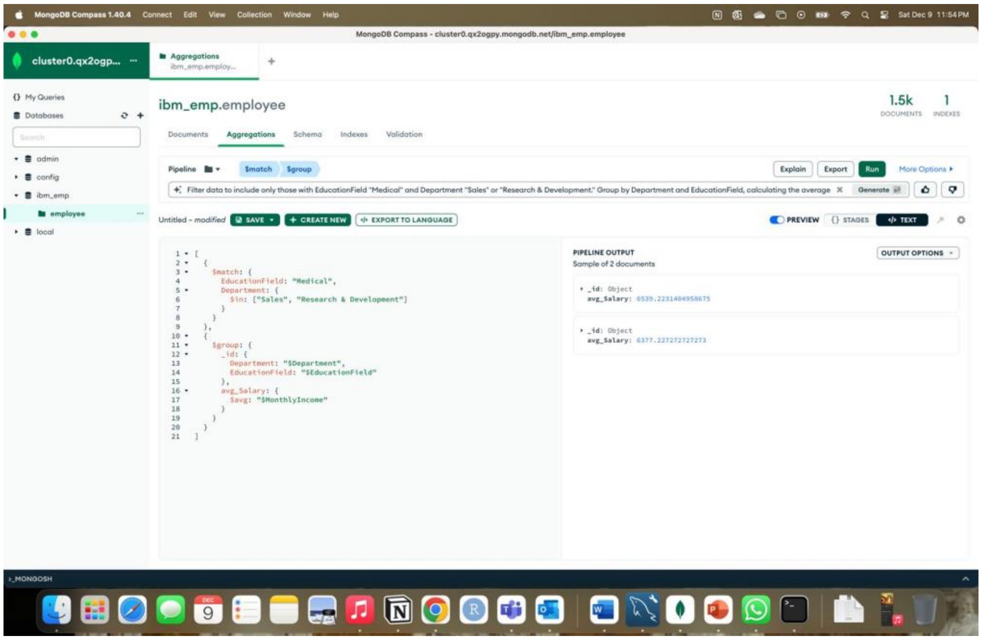
Determine the mean income of medical education field personnel working in the "Sales" and "Research & Development" divisions.

•It is evident that the Research & Development Department pays more. This can therefore be used as proof to back up the rumor.

• The outcome contains two documents: one for the medical field sales department and one for Department of Research and Development in the Medical FieldTranslation

## Translation

Aggregate the emp_attrition collection in the database. In the first stage, filter the documents using the $match operator to include only those with the EducationField set to "Medical" and the Department set. to either "Sales" or "Research & Development". In the second stage, group the filtered documents using. the $group operator. The grouping key is composed of the Department and EducationField fields. Calculate the average value of the MonthlyIncome field for each group and assign it to the field. avg_Salary in the result.



Filter data to include only those with EducationField "Medical" and Department "Sales" or "Research & Development." Group by Department and EducationField, calculating the average

```
[
  {
    $match: {
      EducationField: "Medical",
      Department: {
        $in: ["Sales", "Research & Development"]
      }
    }
  },
  {
    $group: {
      _id: {
        Department: "$Department",
        EducationField: "$EducationField"
      },
      avg_Salary: {
        $avg: "$MonthlyIncome"
      }
    }
  }
]
```

PIPELINE OUTPUT

Sample of 2 documents

```
▸ _id: Object
  avg_Salary: 6539.2231404958675

▸ _id: Object
  avg_Salary: 6377.227272727273
```