Group #:2

Aashritha Sagar Veesamsetty

Hothal Vimalbhai Patel

Shubham Agarwal

Sachin Thomas

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

BUAN 6320

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Activity** | **Member 1** | **Member 2** | **Member 3** | **Member 4** |
| Prepared Data Model and Created Physical DB | x | x | x | x |
| Loaded Data into Database | x | x | x | x |
| Wrote SQL Queries | x | x | x | x |
| Prepared Mongo Database | x | x | X | x |
| Loaded data into Mongo DB | x | x | X | x |
| Wrote Mongo Queries | x | x | x | x |
| Prepared Report | x | x | X | x |
| Reviewed Report | x | x | x | x |

Contents

Relational Data Model 4

Assumptions/Notes About Data Entities and Relationships 4

Entity-Relationship Diagram 6

Physical MySQL Database 8

Assumptions/Notes About Data Set 8

Screen shot of Physical Database objects 8

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 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 5 23

Question 23

Notes/Comments About SQL Query and Results (Include # of Rows in Result) 23

Translation 23

Screen Shot of SQL Query and Results 23

SQL Query 6 25

Question 25

Notes/Comments About SQL Query and Results (Include # of Rows in Result) 25

Translation 25

Screen Shot of SQL Query and Results 25

Data Review for MongoDB 27

Assumptions/Notes About Data Collections, Attributes and Relationships between Collections 27

Physical Mongo Database 28

Assumptions/Notes About Data Set 28

Screen shot of Physical Database objects (Database, Collections and Attributes) 28

Data in the Database 28

MongoDB Queries/Code 29

Mongo Query 1 29

Question 29

Notes/Comments About MongoDB Query/Code and Results (Include # of Documents in Result) 29

Translation 29

Screen Shot of MongoDB Query/Code and Results 29

Mongo Query 2 30

Question 30

Notes/Comments About MongoDB Query/Code and Results (Include # of Documents in Result) 30

Translation 30

Screen Shot of MongoDB Query/Code and Results 30

Mongo Query 3 31

Question 31

Notes/Comments About MongoDB Query/Code and Results (Include # of Documents in Result) 31

Translation 31

Screen Shot of MongoDB Query/Code and Results 31

# Relational Data Model

## Assumptions

* Education and EducationField are assumed to be constant in the hope that the employee has no plans for further education.
* Job Level and Job role can change for an employee over time. So we assume that there might be multiple records for an employee in these categories
* Age and Over18 are related fields which can be set as a calculated field. We assume that the age gets added on in the table over the years rather than having multiple records for the same employee's age
* Hourly Rate, Standard Hours, Overtime, Daily Rate, Monthly Rate and MonthlyIncome can change over time with a company. The employee can have multiple records under related fields.
* Each Employee can have multiple records under the SalaryInfo, JobDetails, Performance, Satisfaction tables. Logic behind the assumption being : an employee can have changes in these details while working with a company over the years
* Employee profile table has information that does not require multiple records for the employee. These details can be updated as a calculated field over time rather than having multiple records unlike other tables
* Education\_Details, EnvironmentSatisfaction\_Details, JobInvolvement\_Details, JobSatisfaction\_Details, PerformanceRating\_Details, RelationshipSatisfaction\_Details, WorkLifeBalance\_Details has description as to what each integer means with a description
* Primary keys had to be manually added to some of the derived tables to ensure 3NF structure of the database. This was necessary since the data did not have multiple records for an employee.
* The following field was removed from the tables since it is not critical to the 3NF integrity
* Employee count

Notes About Data Entities and Relationships

Following are the Table descriptions and relationships:

1. **Employee\_PersonalInfo :** This table has the personal details of the employee and most of the data being captured is constant. The data also has some information about education, and experience as well.
2. **Employee\_WorkProfile :** This table stores the data regarding the employee’s time with the firm. Also, the table has the employee’s attrition status.
3. **Employee\_JobDetails :** The table lists the details about the current department and job role associated with the employee, in addition to the Business travel category.
4. **Employee\_Performance :** This table has details relevant to the performance reviews for the employee.
5. **Employee\_SalaryInfo :** The table has details regarding the employees salary information in addition to the standard hours of work and stock options.
6. **Employee\_Satisfaction :** The table has a record of details regarding the employee’s general satisfaction regarding work and the firm. Also, it has details with regards to the Work Life balance of the employee.
7. **Education\_Details :** Table has the description associated with the field Education available in the Employee\_PersonalInfo table.
8. **JobInvolvement\_Details :** Description of JobInvolvment levels given in the dataset
9. **PerformanceRating\_Details :** Description of PerformanceRating numbers in the dataset
10. **JobSatisfaction\_Details :** Description of JobSatisfaction as given in the dataset
11. **RelationshipSatisfaction\_Details :** Description of RelationshipSatisfaction numbers in the dataset
12. **EnvironmentSatisfaction\_Details :** EnvironmentSatisfaction data description as give in the dataset
13. **WorkLifeBalance\_Details :** Description of WorkLifeBalance as given in dataset

Reasons for the Model to be in 3NF

The model presented in the 3NF form. It is said to be in 3NF since it satisfies the following conditions:

·         Each table contains only a single value

·         Each record is unique

·         Primary key is defined for each table that is Unique and Not Null

·         No transitive and functional dependencies

# Entity-Relationship Diagram

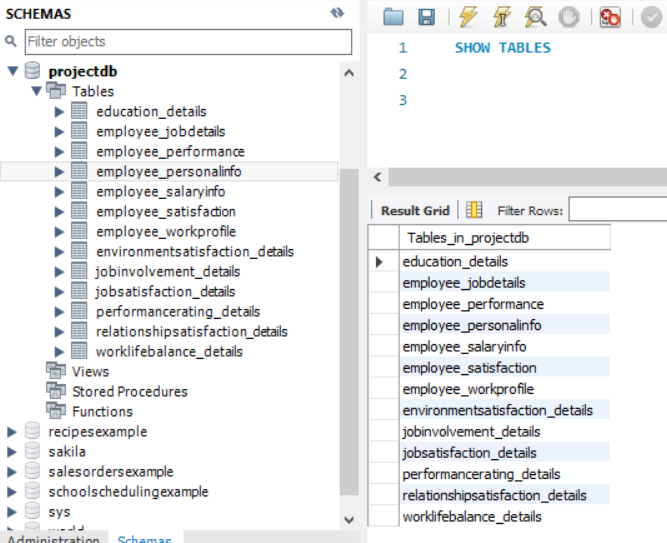
# 

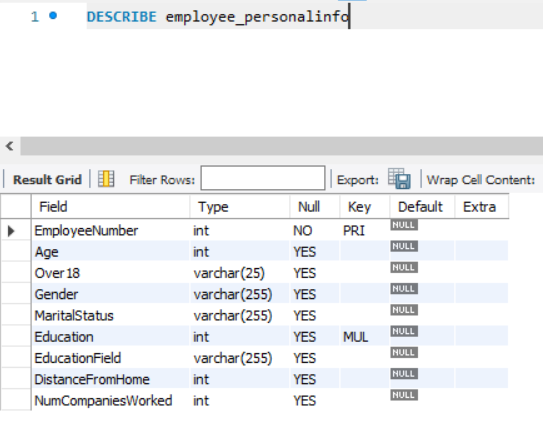
# Physical MySQL Database

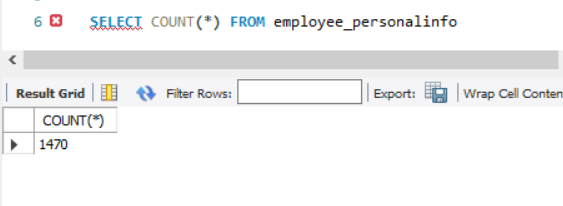
## 

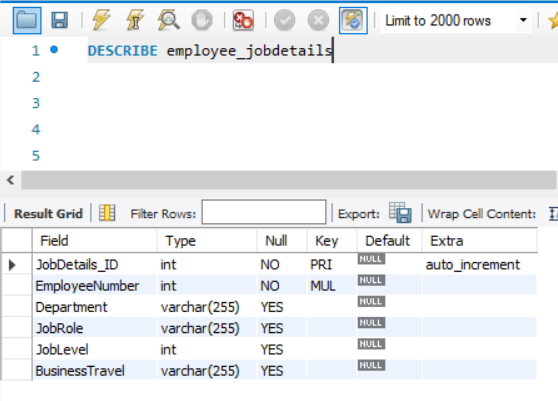
## Screen shot of Physical Database objects

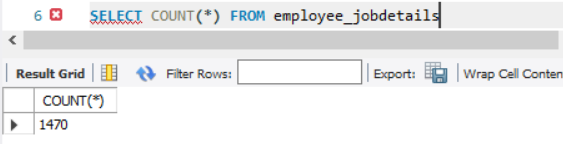
Tables in the Database



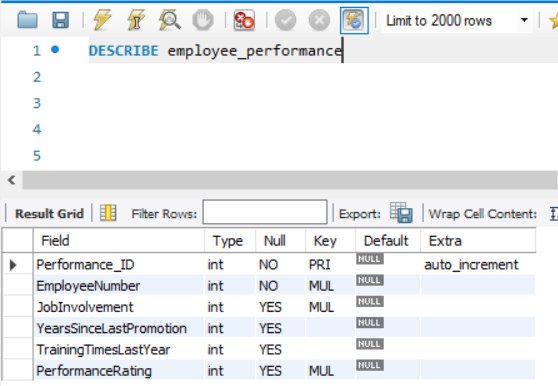
1. Employee\_PersonalInfo  


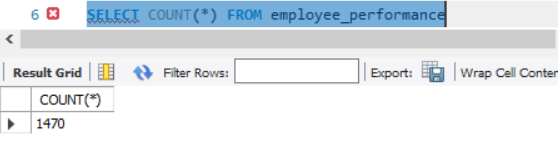
  
  
2. Employee\_JobDetails



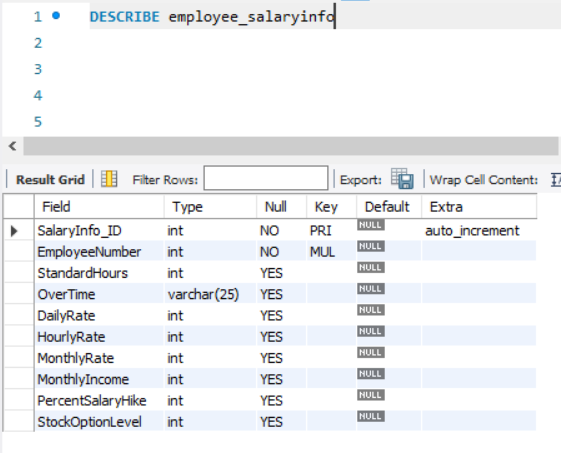


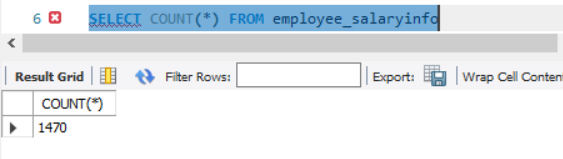
3. Employee Performance



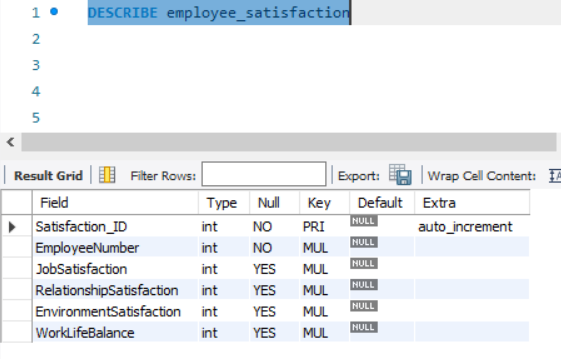


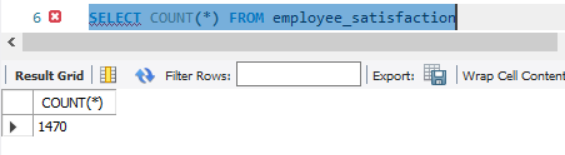
4. Employee\_SalaryInfo



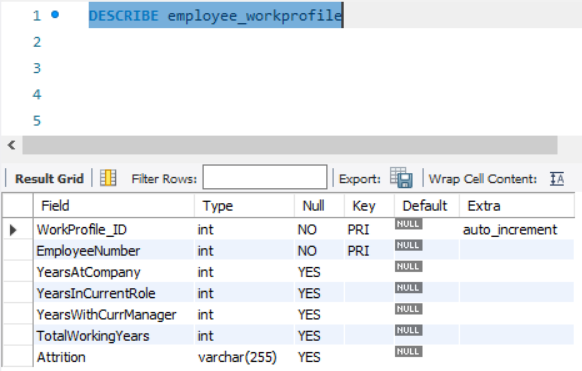


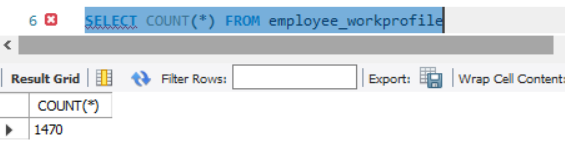
5. Employee\_Satisfaction



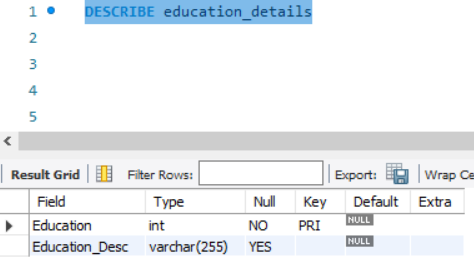


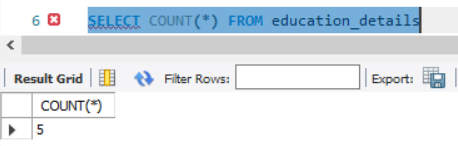
6. Employee\_WorkProfile



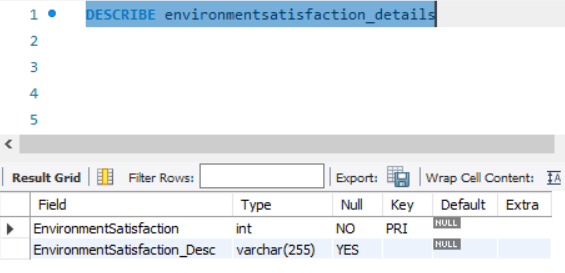


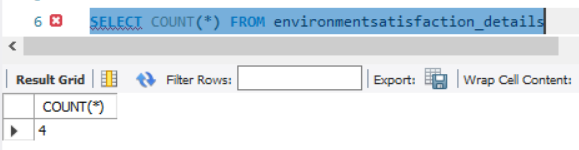
7. Education\_Details



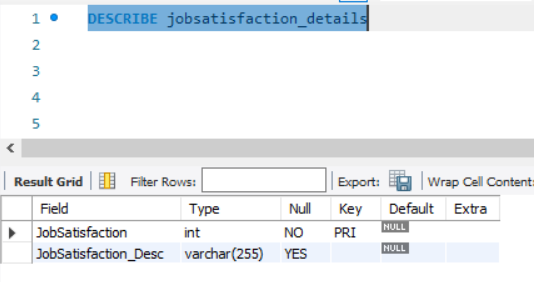


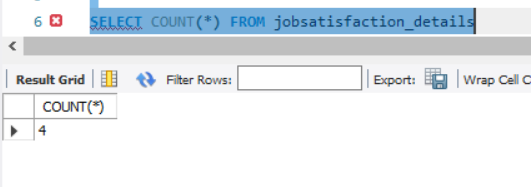
8. EnvironmentSatisfaction\_Details



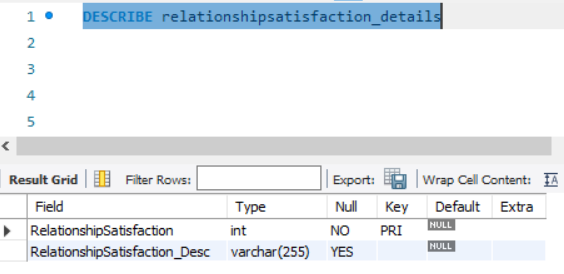


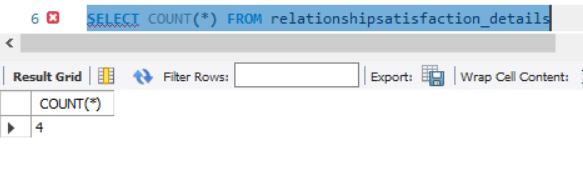
9. JobSatisfaction\_Details



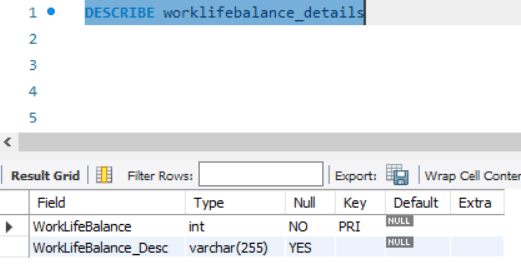


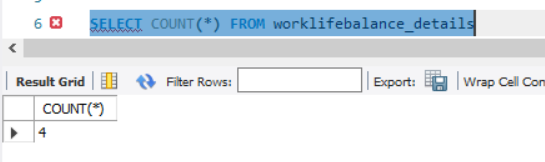
10. RelationshipSatisfaction\_Details



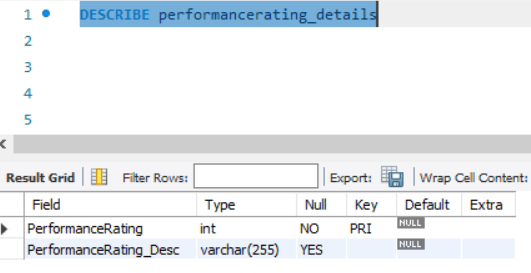


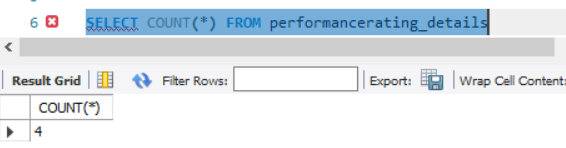
11. WorkLifeBalance\_Details



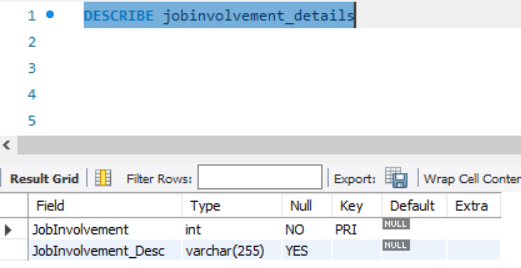


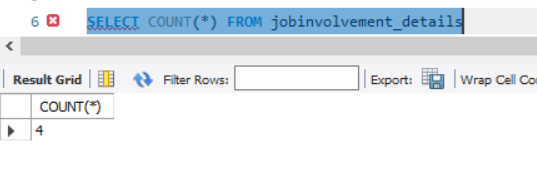
12. PerformanceRating\_Details





13.  JobInvolvement\_Details





## 

Data in the Database

|  |  |  |  |
| --- | --- | --- | --- |
| **Table Name** | **Primary Key** | **Foreign Key** | **# of Rows in Table** |
| Employee job details | Jobdetails\_id INT | Employeenumber INT | 1470 |
| Education details | Education INT | Education INT | 5 |
| Employee personal information | Employeenumber INT | NA | 1470 |
| Employee salary information | Salaryinfo\_id INT | Employeenumber INT | 1470 |
| Employee satisfaction | Satisfaction\_id INT | Employeenumber INT | 1470 |
| Job satisfaction details | Jobsatisfaction INT | Jobsatisfaction INT | 4 |
| Relationship satisfaction details | Relationshipsatisfaction INT | Relationshipsatisfaction INT | 4 |
| Work-life balance details | Worklifebalance INT | Worklifebalance INT | 4 |
| Environment satisfaction details | Environmentsatisfaction INT | Environmentsatisfaction INT | 4 |
| Employee work profile | Workprofile\_id INT  EmployeeNumber INT | Employeenumber INT | 1470 |
| Employee performance | Performance\_id INT | Employeenumber INT | 1470 |
| Job involvement details | Jobinvolvement INT | Jobinvolvement INT | 4 |
| Performance rating details | Performancerating INT | Performancerating INT | 4 |

# SQL Queries

## SQL Query 1

### Question

*Assume that the company has several branch offices around the country and employee morale is down. The company feels that not enough employees are traveling frequently between offices and that more employees should travel frequently to the branch offices to improve morale amongst each other. Is the company correct in feeling this way?*

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

For this question, it is assumed that the employees who travel frequently have a higher employee morale. As employee morale is a qualitative measure of satisfaction, we considered the 4 measures of satisfaction in the data to make sure all variables are considered and given equal weight. We considered the summation of the scores for job satisfaction, relationship satisfaction, environment satisfaction and work life balance as a true depiction of employee morale which is then divided by a count of 4.

The aggregate scores of the 3 groups who travel rarely, frequently and do not travel were compared to confirm if there’s truth to the statement. From the output we have concluded that there is a very marginal difference in the scores for employee morale between those who travel frequently and those who do not travel at all. In fact, those who do not travel at all have scored higher with a 2.77 score in the employee morale score.

Thus, we’d like to say that the company could be wrong in thinking that travelling frequently could improve employee’s morale.

# of Rows in Result: 3

### Translation

Select travel category and average of summation of job satisfaction, relationship satisfaction, environment satisfaction, work life balance ratings named as morale

from employee job details table joined with employee satisfaction table on employee number

Group results by travel category

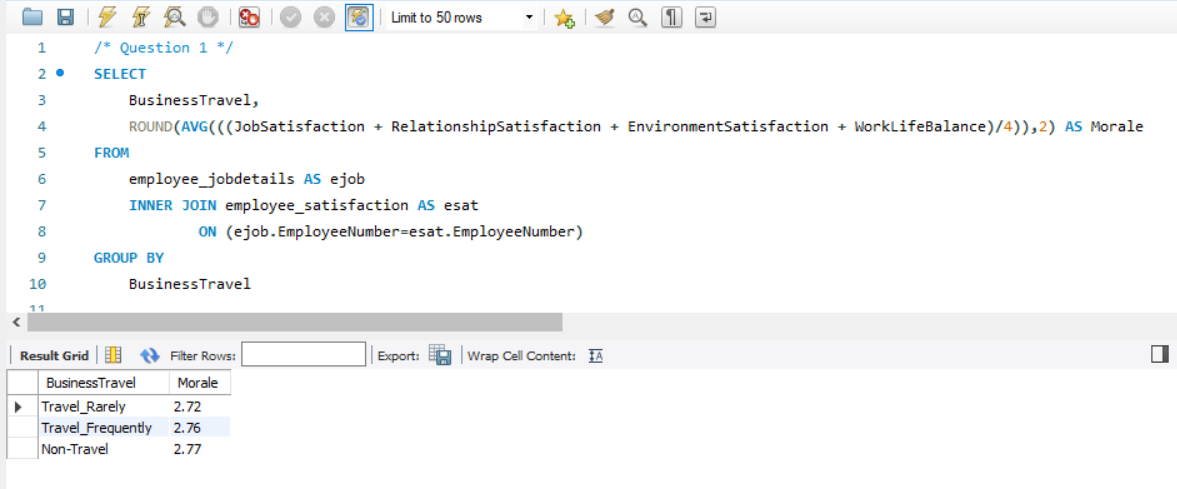
### Clean-up

Select business travel, avg satisfaction as morale

From employee\_jobdetails inner join employee\_satisfaction on employee number

Group By business travel

### Screen Shot of SQL Query and Results



## SQL Query 2

### Question

*Which department's employee is the most likely to have the shortest commute between home and work?*

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

For this question, the column *distance from home* is interpreted to be represented by the number of miles travelled from the employee’s residence. Firstly, the average distance travelled for each department is taken as a collective measure. We interpreted the question as the probability of a department's employee to travel the least to work. We have observed that there’s no single employee in the database with the least number of miles travelled.

We used the distance travelled for each employee grouped by the department they work in. From the output below, we observe that the average distance travelled by the HR employees is less compared to the other employees in the R&D and Sales department.

# of Rows in Result: 3

### Translation

Select department and average of distance from home as the average distance

From employee job details table inner joined with employee personal info table on the employee number

Group results by department

Order results by average distance from home

### Clean-up

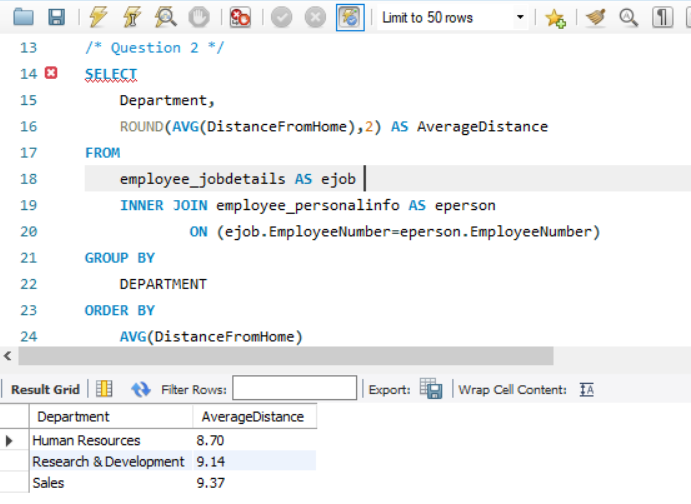
Select department, avg (distance from home) as average distance

From employee\_jobdetails inner join employee\_personalinfo on employeenumber

Group by department

Order by avg (distance from home)

### Screen Shot of SQL Query and Results



## SQL Query 3

### Question

*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 SQL Query and Results (Include # of Rows in Result)

To determine if this candidate has potential to work in Sales with an academic history in medicine, we chose to compare the performance of existing employees within the Sales department by their educational fields. The aggregate scores for performance rating of these employees were used as a measurement of success in their jobs.

After comparing the scores of the average performance of current Sales department employees with different academic backgrounds, it is safe to say that Sales employees with a background in Medicine have done just as well as the others. Sales department employees with a Medicine background have in fact outperformed most educational backgrounds marginally in performance.

# of Rows in Result: 5

### Translation

Select department, education field, count of employee numbers as count of employees, average of performance ratings as average performance rating

From employee job details table inner joined with employee personal info table on employee number, inner joined with employee performance table on employee number

Where department is sales

Group results by department and education field

Order results by average of performance rating in descending order

### Clean-up

Select department, education field, count (employeenumber) as countofemployees, avg (performance rating) as averageperformancerating

From employee\_jobdetails inner join employee\_personalinfo on employeenumber

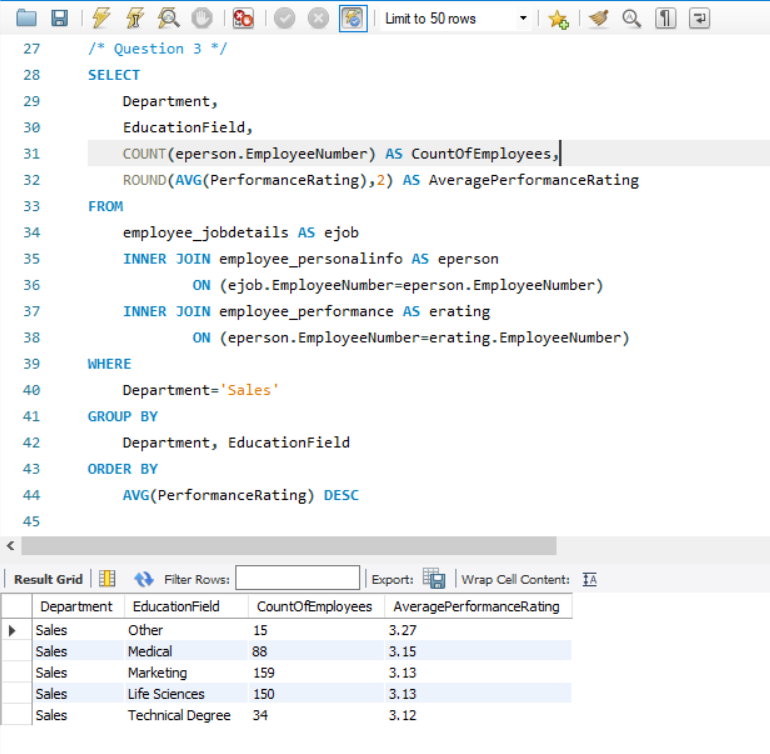
inner join employee\_performance on employee number

Where department = 'sales'

Group by department, education field

Order by avg (performance rating) desc

### Screen Shot of SQL Query and Results



## SQL Query 4

### Question

*The HR department feels they have the highest job satisfaction while the Research & Development department feels their department has the highest environment satisfaction. Who is right?*

### 

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

The two bases of distinction in this problem statement were the 2 satisfaction variables. In the first case, the aggregate scores of the *job satisfaction* variable was used to determine which department has the highest score and in the second case, *environment satisfaction* was used to determine the highest scoring department.

From the output it is evident that the Sales team has scored higher in job satisfaction and not the HR. The R&D is right in believing that they are the most satisfied with the work environment.

# of Rows in Result: 3

### Translation

Select department, average of job satisfaction as average job satistaction, average of environment satisfaction as average environment satisfaction

From employee job details table inner joined with employee satisfaction table on employee number

Group results by department

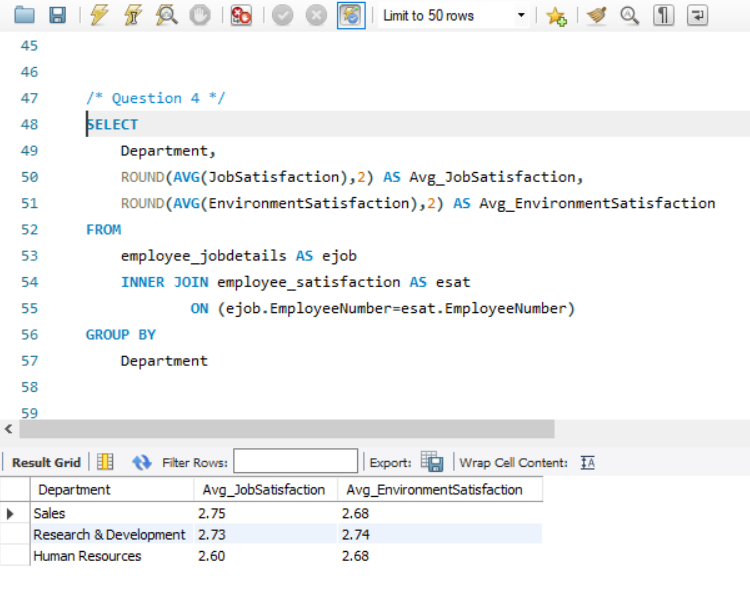
### Clean-up

Select department, avg (job satisfaction) as avg\_jobsatistaction, avg (environment satisfaction) as avg\_environmentsatisfaction

From employee\_jobdetails inner join employee\_satisfaction on employee number

Group by department

### Screen Shot of SQL Query and Results



## SQL Query 5

### Question

*An employee in the 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?*

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

For this problem we used the average daily rate as a basis for differentiation in the wage earned by both the male and female population of the company. We used the daily rate as it would be devoid of any inherent bias. But during our analysis we came across that the job level of these employees could play a role in the difference in wage earned. Hence, output consists of employee daily rate comparisons by the department and by job-level for each gender.

* From the output below we can tell that the women of the HR department are all paid more than the male employees except for level 4 employees.
* Whereas in the R&D department, the claim remains true with male employees earning more than the females. The females in level 5 are the only ones earning more than men but with a very small difference in average pay.
* In the Sales department, women seem to be earning more than their male colleagues with only an exception for level 2 employees.

In conclusion, the claim that females are paid less than males in the company does not hold true in all cases.

# of Rows in Result: 30

### Translation

Select department, job level, gender, average of daily rate as average daily rate

From employee job details table inner joined with employee personal info table on employee number, inner joined with employee salary info table on employee number

Group results by department, job level, and gender

Order results by department, job level, and gender

### Clean-up

Select department, job level, gender. avg (daily rate) as avg\_dailyrate

From employee\_jobdetails inner join employee\_personalinfo on employee number

inner join employee\_salaryinfo on employee number

Group by department, job level, gender

Order by department, job level, gender

### Screen Shot of SQL Query and Results

## SQL Query 6

### Question

*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?*

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

For this question we chose the variables as the marital status of men and their average performance to indicate if there is any truth to the statement made by the magazine company. We considered the average performance of men who were married, divorced and single and compared them using a tabular representation and grouping them by the department they’re working for.

We observed that married men do not necessarily have a higher performance rating compared to single and divorced men. When we group the data by the department, we observe that the case varies for men working in each department.  For instance, in Human Resources, men who are divorced outperformed single and married men wrt performance. In Research & Development, married men scored higher. However, the difference in performance rating is only marginal and there is no significant difference in performance rating scores.

# of Rows in Result: 9

### Translation

Select marital status, average of performance rating as average performance rating

From employee job details table inner joined with employee personal info table on employee number, inner joined with employee performance table on employee number

Group results by marital status

Order results by average of performance ratings in descending order

### Clean-up

Select marital status, avg (performance rating) as averageperformancerating

From employee\_jobdetails inner join employee\_personalinfo on employee number inner join employee\_performance on employee number

Group by marital status

Order by avg (performance rating) desc

### Screen Shot of SQL Query and Results

# Data Review for MongoDB

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

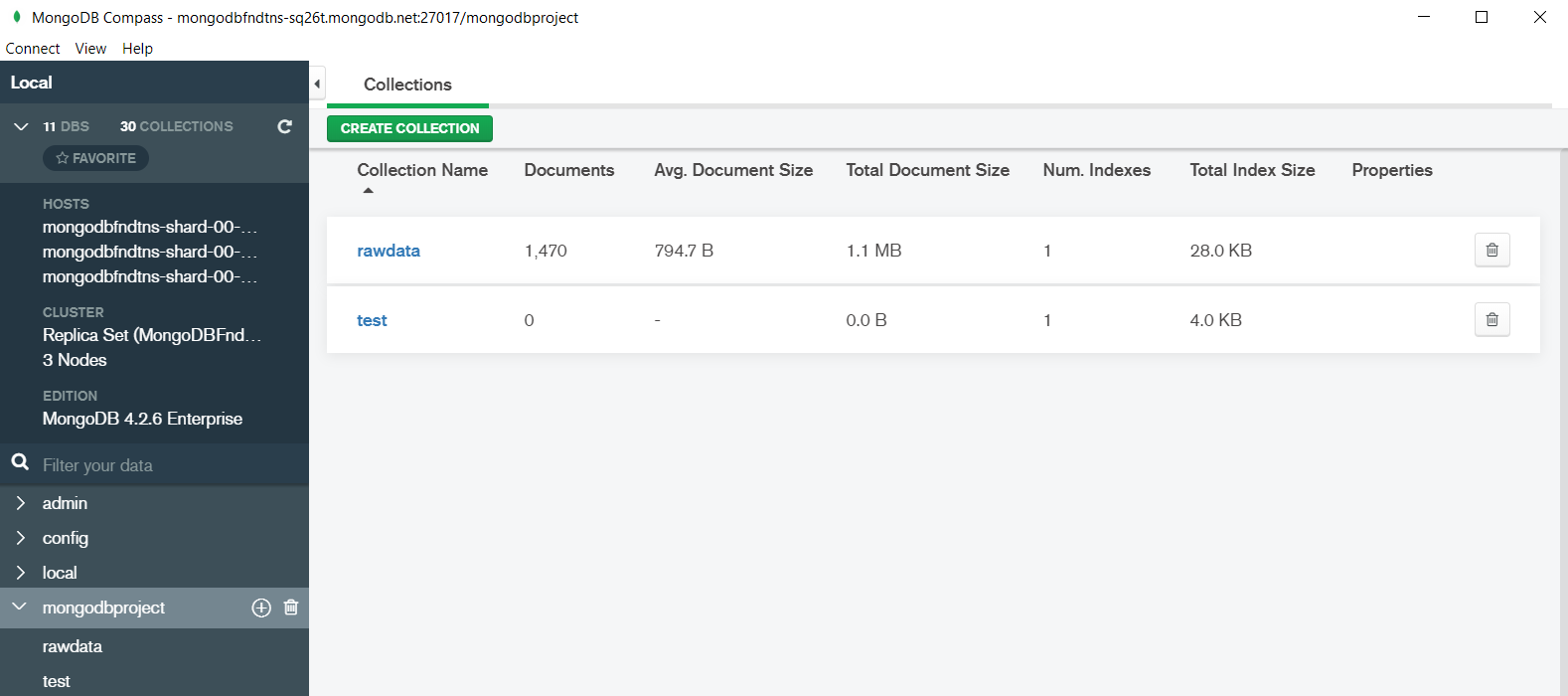
* The whole dataset has been loaded into a single collection called ‘rawdata’
* The ‘rawdata’ collection contains all the data exactly as given in the initially provided csv file for the project
* Loading Data into a single collection in mongodb did not affect the result sets from mysql since there were no duplicate records for an employee
* A single collection also enabled easy querying of the data from mongodb without data integrity issues

# Physical Mongo Database

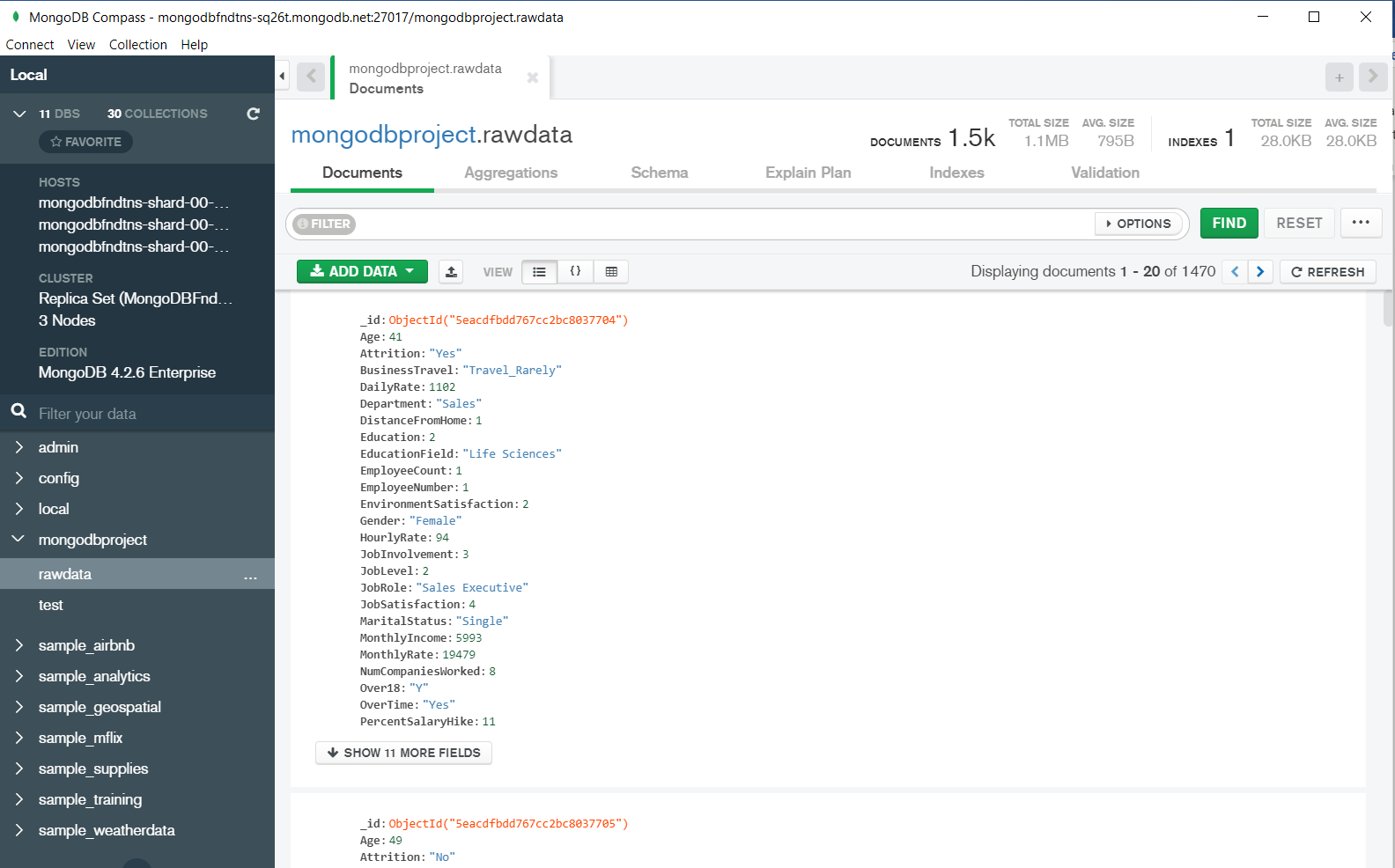
## Assumptions/Notes About Data Set

* The non-normalized data of the whole dataset is loaded into the mongodb
* No duplicate records exist for an individual employee
* For the same reason the query result from mongodb would match that of the database design implemented in mysql

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

Database and Collection 

The collection ‘rawdata’ which has the employee data has 1470 documents (rows) as seen in the above screen shot.

RawData (Collection)

## 

## Data in the Database

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

# MongoDB Queries/Code

Pick 3 SQL queries and write them in MongoDB

## Mongo Query 1

### Question

### *Which department's employee is the most likely to have the shortest commute between home and work?*

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

For this question, the column *distance from home* is interpreted to be represented by the number of miles travelled from the employee’s residence. Firstly, the average distance travelled for each department is taken as a collective measure. We interpreted the question as the probability of a department's employee to travel the least to work. We have observed that there’s no single employee in the database with the least number of miles travelled.

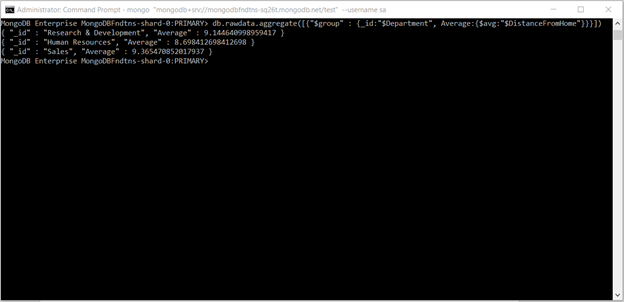
We used the distance travelled for each employee grouped by the department they work in. From the output below, we observe that the average distance travelled by the HR employees is less compared to the other employees in the R&D and Sales department.

# of Documents in Result: 3

### Translation

* Database set as mongodbproject
* Rawdata collection used
* Use aggregate function
* Group the data by “Department”
* Filter using “avg” function for “Distancefromhome”

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



## Mongo Query 2

### Question

*The HR department feels they have the highest job satisfaction while the Research & Development department feels their department has the highest environment satisfaction. Who is right?*

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

The two bases of distinction in this problem statement were the 2 satisfaction variables. In the first case, the aggregate scores of the *job satisfaction* variable was used to determine which department has the highest score and in the second case, *environment satisfaction* was used to determine the highest scoring department.

From the output it is evident that the Sales team has scored higher in job satisfaction and not the HR. The R&D is right in believing that they are the most satisfied with the work environment.

# of Documents in Result: 3

### Translation

* Database set as mongodbproject
* Rawdata collection used
* Group by “Department” set
* Filter with “avg” function for 2 fields of “JobSatisfaction” and “ EnvironmentSatisfaction”

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

## Mongo Query 3

### Question

*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?*

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

For this question we chose the variables as the marital status of men and their average performance to indicate if there is any truth to the statement made by the magazine company. We considered the average performance of men who were married, divorced and single and compared them using a tabular representation and grouping them by the department they’re working for.

We observed that married men do not necessarily have a higher performance rating compared to single and divorced men. When we group the data by the department, we observe that the case varies for men working in each department.  For instance, in Human Resources, men who are divorced outperformed single and married men wrt performance. In Research & Development, married men scored higher. However, the difference in performance rating is only marginal and there is no significant difference in performance rating scores.

# of Documents in Result: 9

### Translation

* Use mongoprojectdb as db source.
* Use rawdata as collection.
* Group the output by “MaritalStatus” and “ Department”
* Choose filter as the “avg” “PerformanceRating”
* Use “sort” to display data

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



### **Appendix**

### Mongo Queries for ease of reference

### Mongo Query1

db.rawdata.aggregate([{"$group" : {\_id:"$Department", Average:{$avg:"$DistanceFromHome"}}}])

### Mongo Query2

db.rawdata.aggregate([{

$group:{

\_id:"$Department",

AvgJobSat: { $avg: "$JobSatisfaction" },

AvgEnvSat: { $avg: "$EnvironmentSatisfaction" }}}])

### Mongo Query3

db.rawdata.aggregate([{

$group:{

\_id: {Dept:"$Department", MaritalStatus:"$MaritalStatus"},

AvgPerfRat: { $avg: "$PerformanceRating" }}},

{$sort:{"\_id.AvgPerfRat":-1}}])