Group #:wed4pm\_group18

Group Members:

Madhu Prakash

Nikhil Narayane

Priya Sambhavi

**SQL-Mongo Project – Spatial Data of US Wildfires**

BUAN 6320

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Activity** | **Madhu**  **Prakash** | **Priya**  **Sambhavi** | **Nikhil**  **Narayane** |  |
| Prepared Data Model and Created Physical DB | x | x | x |  |
| Loaded Data into Database | x |  |  |  |
| Wrote SQL Queries | x |  | x |  |
| Prepared Mongo Database | x | x | x |  |
| Loaded data into Mongo DB | x | x | x |  |
| Wrote Mongo Queries |  | x | x |  |
| Prepared Report | x | x | x |  |
| Reviewed Report | x | x | x |  |

Contents

Data Model

Assumptions/Notes About Data Entities and Relationships

Entity-Relationship Diagram

Physical Database

Assumptions/Notes About Data Set

Screen shot of Physical Database objects

Data in the Database

SQL Queries

Query 1

Question

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

Translation

Screen Shot of SQL Query and Results

Query 2

Question

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

Translation

Screen Shot of SQL Query and Results

Query 3

Question

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

Translation

Screen Shot of SQL Query and Results

Query 4

Question

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

Translation

Screen Shot of SQL Query and Results

Query 5

Question

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

Translation

Screen Shot of SQL Query and Results

Query 6

Question

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

Translation

Screen Shot of SQL Query and Results

Data Review for MongoDB

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

Physical Mongo Database

Assumptions/Notes About Data Set

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

Data in the Database

MongoDB Queries/Code

Query 1

Question

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

Translation

Screen Shot of MongoDB Query/Code and Results

Query 2

Question

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

Translation

Screen Shot of MongoDB Query/Code and Results

Query 3

Question

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

Translation

Screen Shot of MongoDB Query/Code and Results

Query 4

Question

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

Translation

Screen Shot of MongoDB Query/Code and Results

Query 5

Question

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

Translation

Screen Shot of MongoDB Query/Code and Results

Query 6

Question

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

Translation

Screen Shot of MongoDB Query/Code and Results

# Data Model

## Assumptions/Notes About Data Entities and Relationships

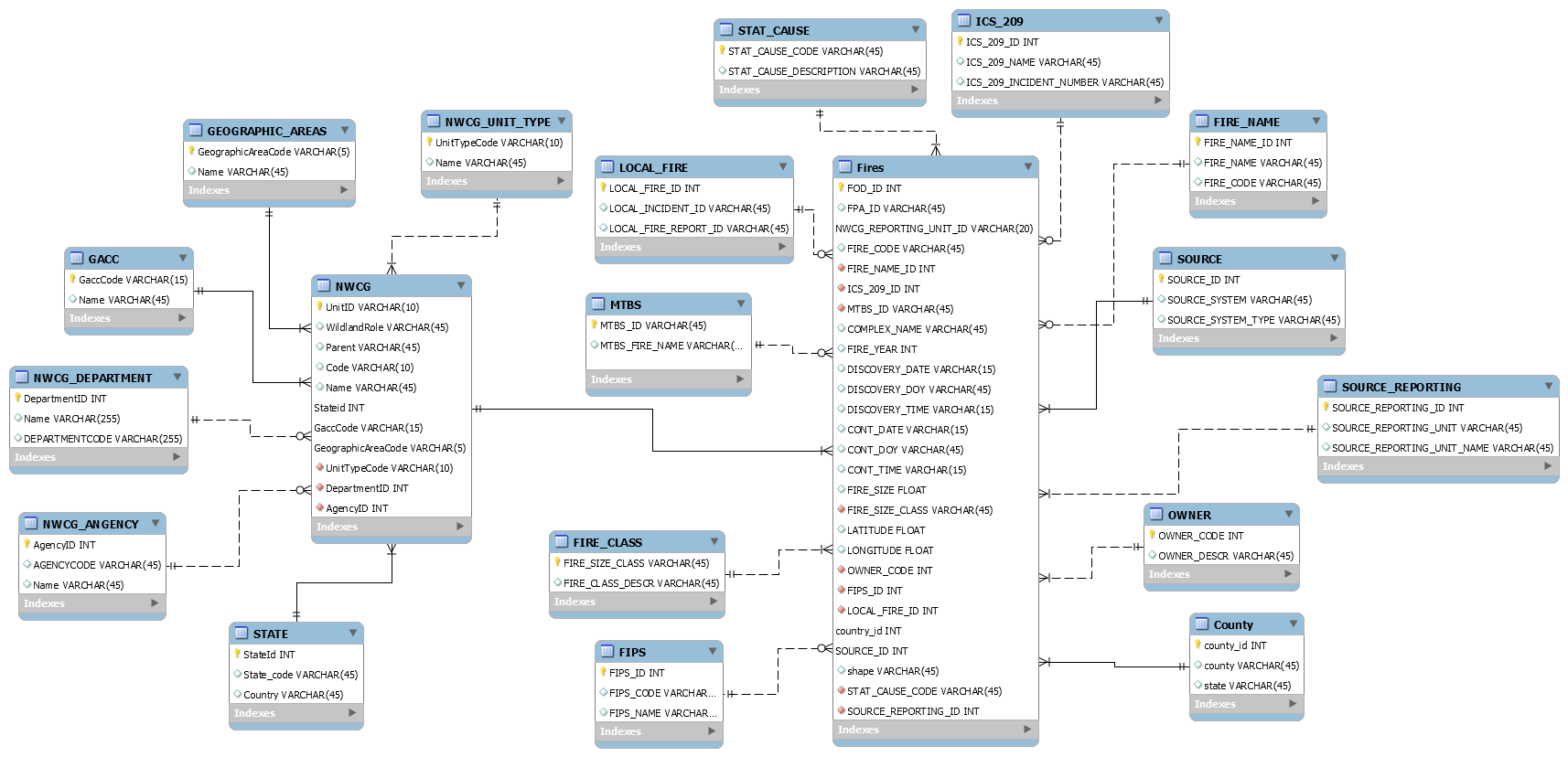
Assumptions :-

1. Relationship with nwcg and fire table is identifying because the record was prepared by nwcg and no record can exist without the nwcg.
2. The source records were drawn from source systems and source system contains its reporting id. So the fire table cannot exit without the source table so it is identifying relationship
3. The geographic area, gacc and state table are identifying because there cannot be an agency without physical space allotted to it.
4. All other tables are non-identifying because the child table can stand alone without ant other tables key
5. Forest has firefighting agencies that takes care of the fire in the forest
6. Agency is divided into departments and units and those units handle the fire
7. Units are diving into unit types so the unittype has a different table
8. The local forest station maintains different registers for different types of incidents and maintains one register containing all the incidents.
9. The burn severity of the fire is measured after the fire problem has been handled by the NWCG.
10. All the agencies mentioned in the table were fully functional at the time of fire.
11. All the major firefighting agencies have authorisation all over the US and thus handle fire emergencies through ICS.
12. Causes for fire are correctly determined and thus categorised into fire name and fire code.
13. Fire records drawn from the source have source reporting ID and name
14. Source is divided into source system and source system type.
15. The fires are measured and categorized depending on fire size.

Reasons why the data model is in 3NF:

1. No multi-part or Multi-valued fields
2. It does not have any [non-prime attribute](https://en.wikipedia.org/wiki/Non-prime_attribute) that has functional dependency
3. No Transitive dependency - every non-key column is functionally dependent only on primary key and nothing else

## Entity-Relationship Diagram:



# Physical Database

## Assumptions/Notes About Data Set:

Empty data:

In NWCG table, the parent column is all NULL.

In both NWCG and FIRES tables, the column OBJECT\_ID has been removed.

Bad data:

In FIRES table, shape column is NaN and that has been changed to varchar.

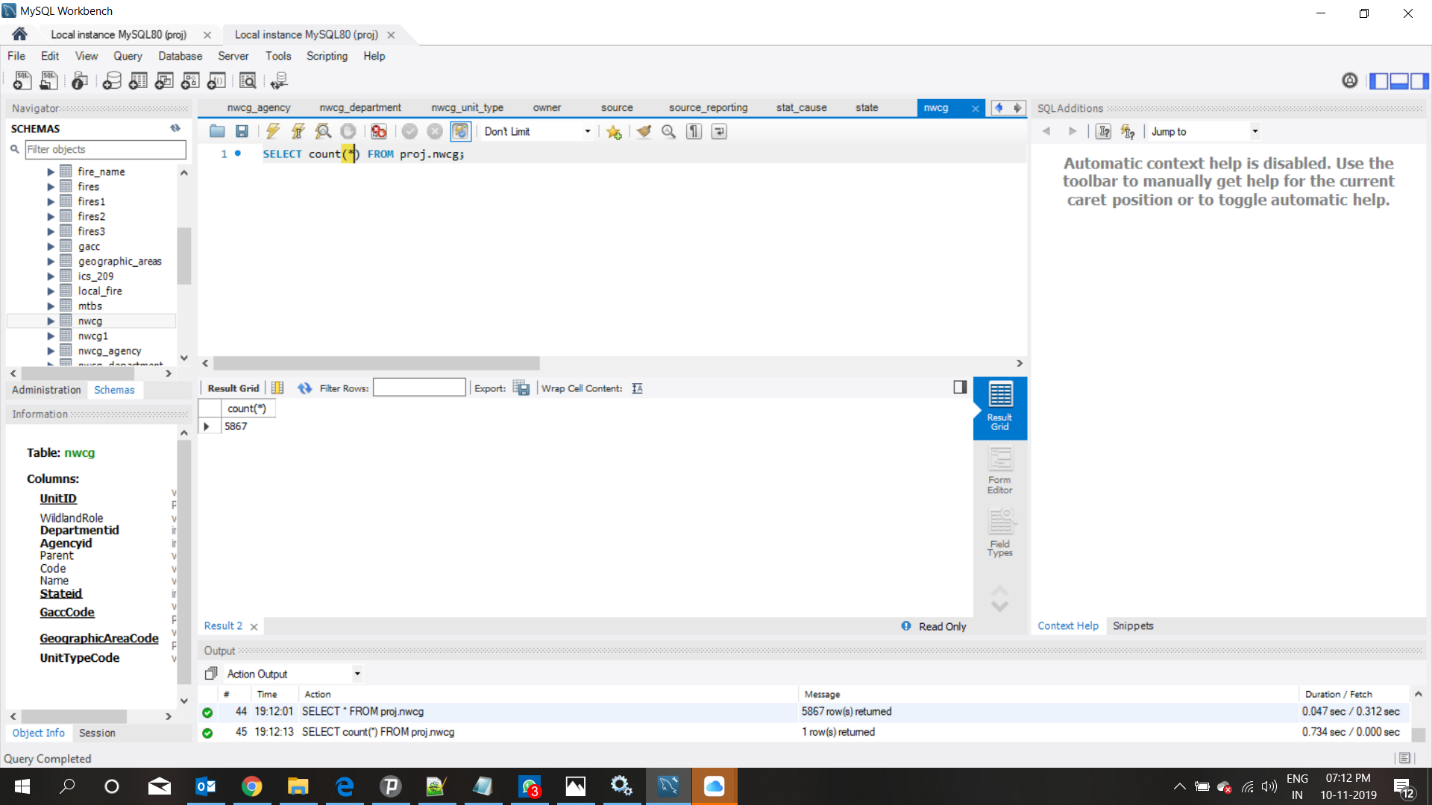
Sparse data: The columns mentioned below are not completely filled.

* In ICS\_209 table, the columns ICS\_209\_NAME and ICS\_209\_INCIDENT\_NUMBER.
* In MTBS table, the columns MTBS\_ID, MTBS\_FIRE\_NAME.
* In FIRES table, the columns COMPLEX\_NAME, CONT\_DATE, CONT\_TIME, CONT\_DOY, DISCOVERY\_TIME.

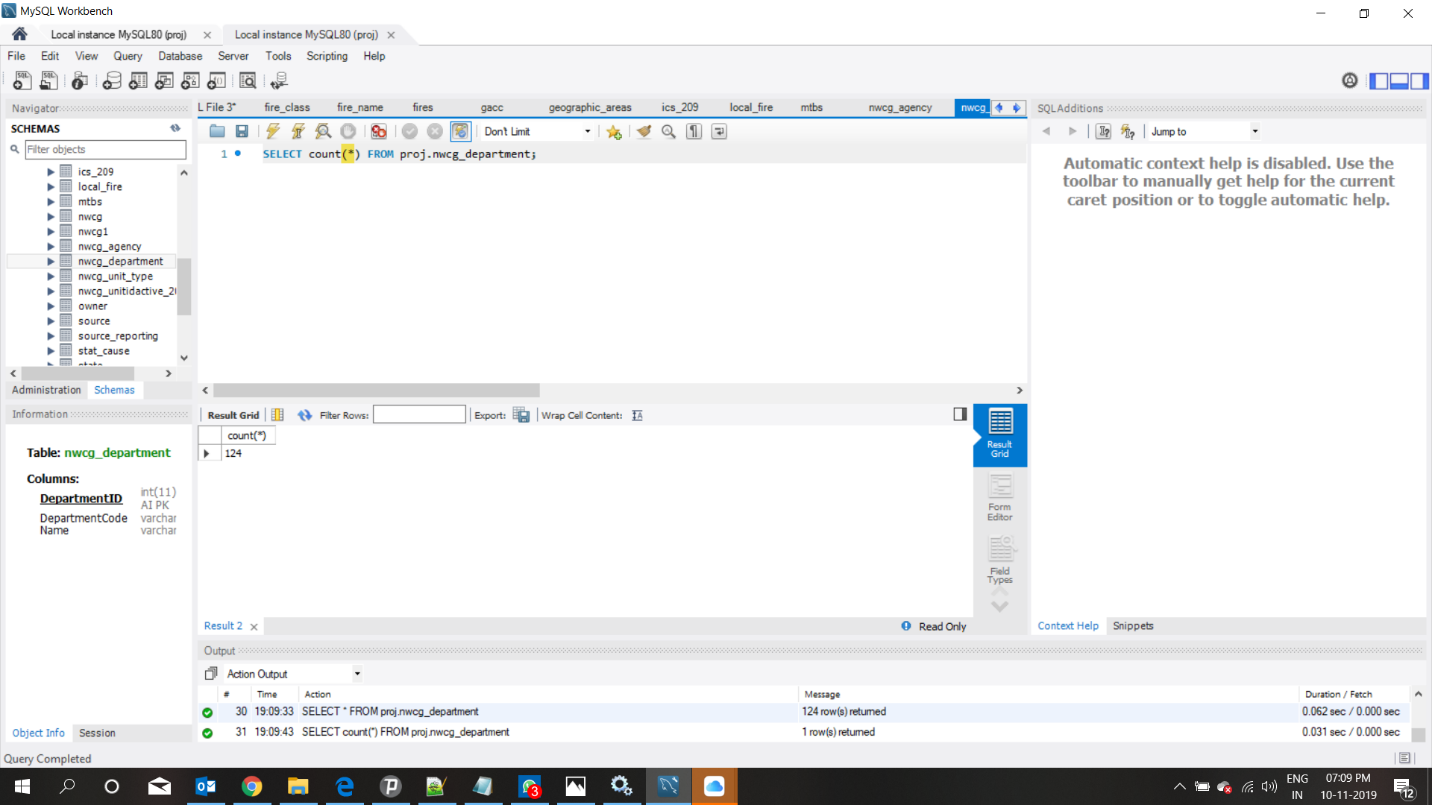
The columns DISCOVERY\_DATE, DISCOVERY\_TIME, CONT\_TIME, CONT\_DATE from the FIRES table have been converted to date-time format from Julian format.

NWCG\_REPORTING\_AGENCY has 3 records more than agency column in nwcg table so those three records have been added to agency column form nwcg

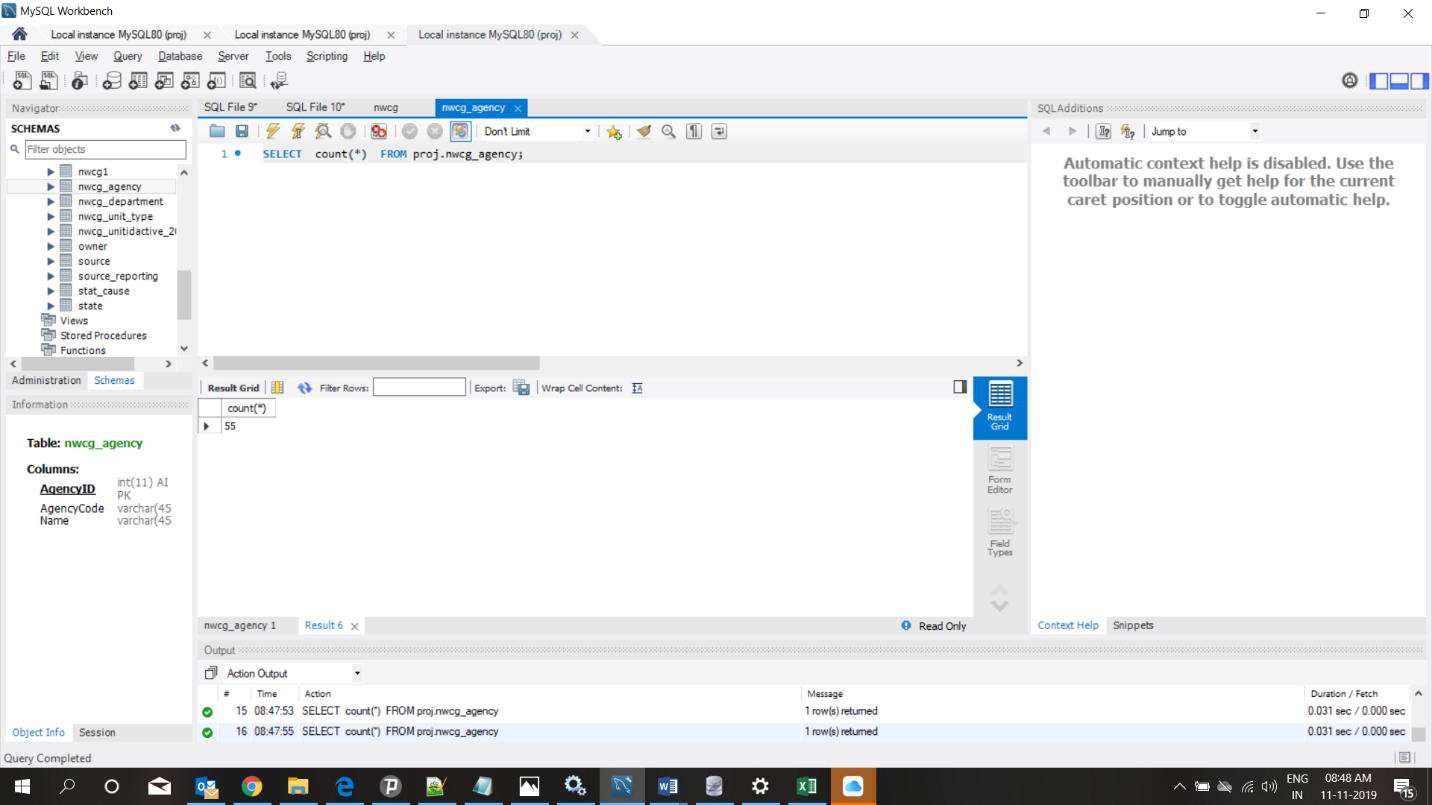
## Screen shots of Physical Database objects:

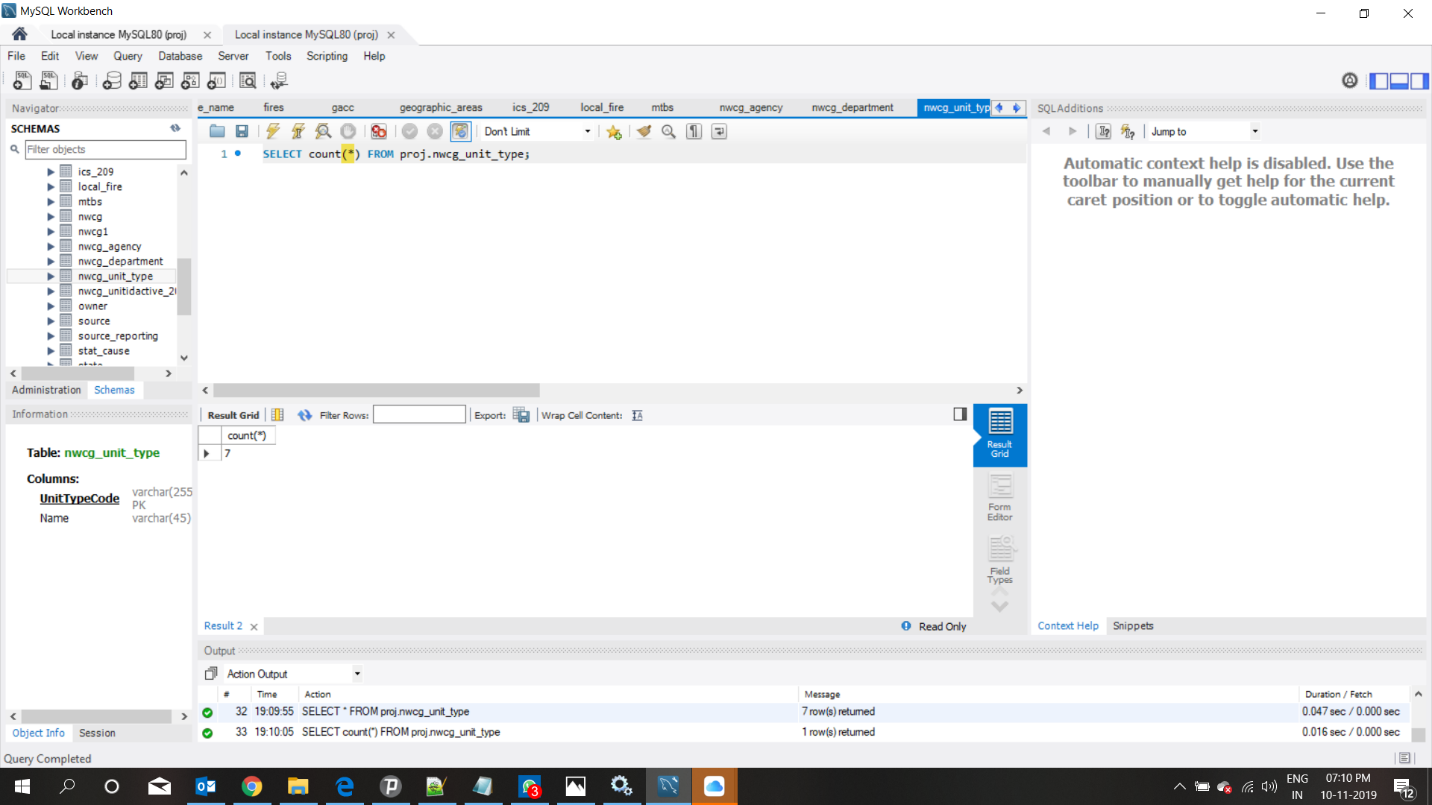
NWCG:

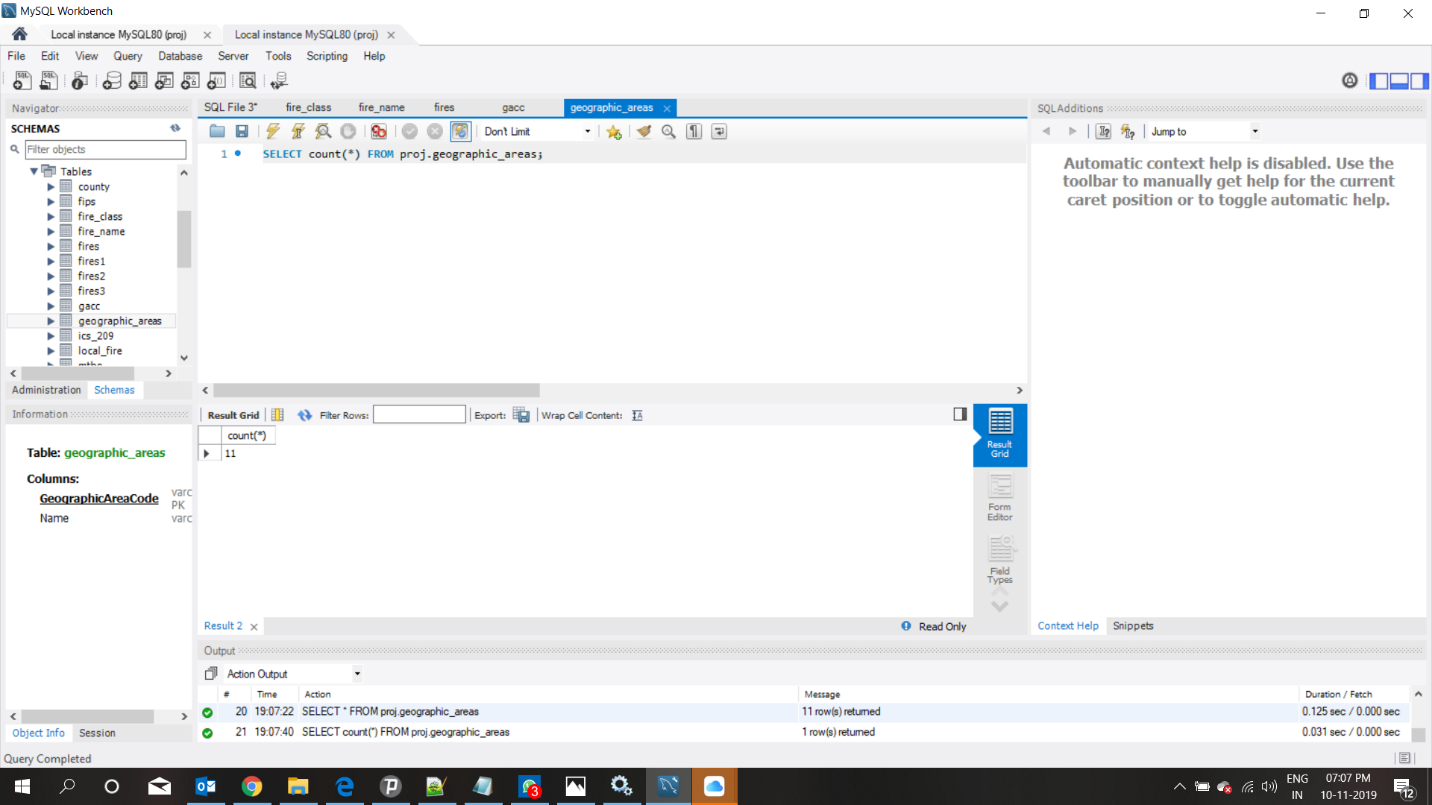
NWCG\_DEPARTMENT:



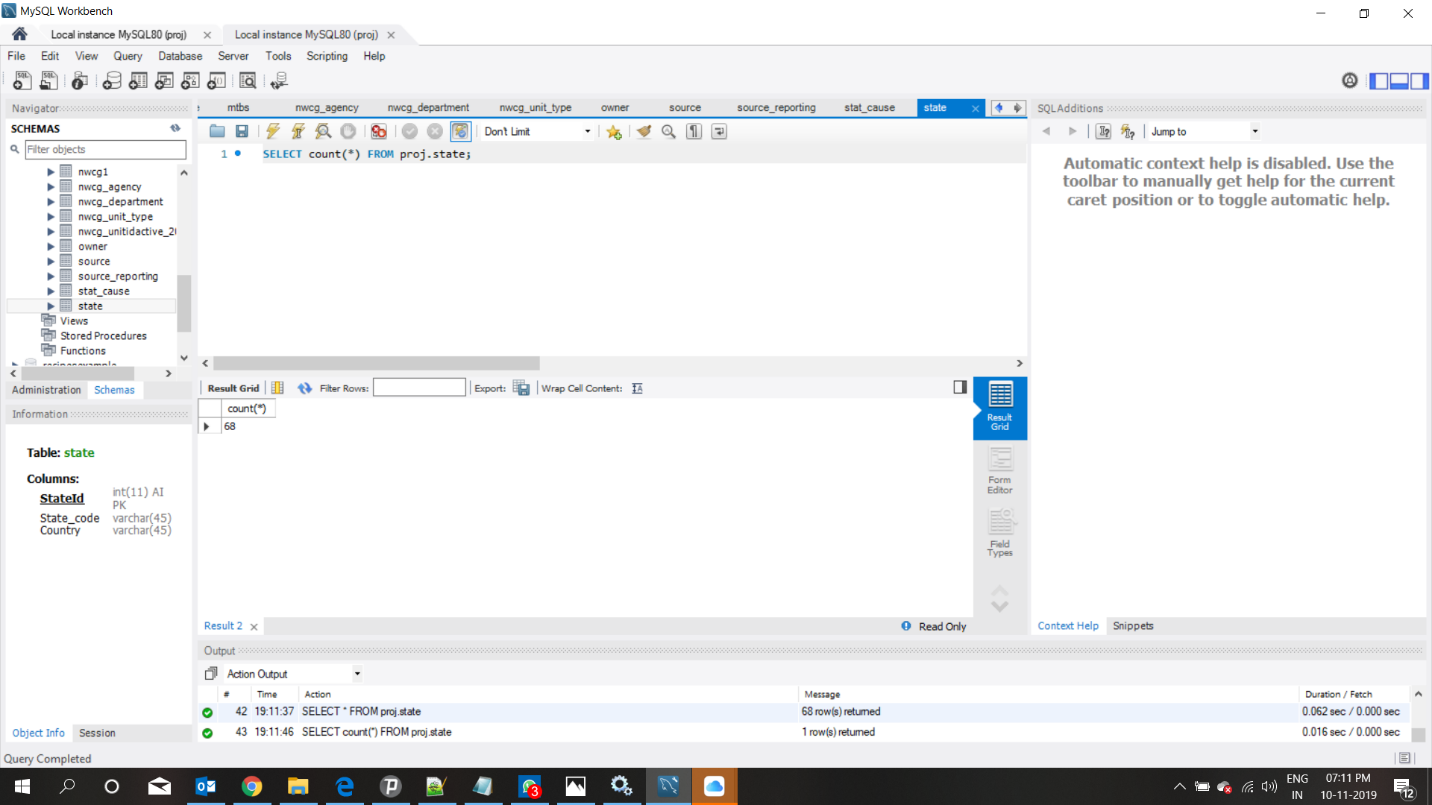
NWCG\_Agency:



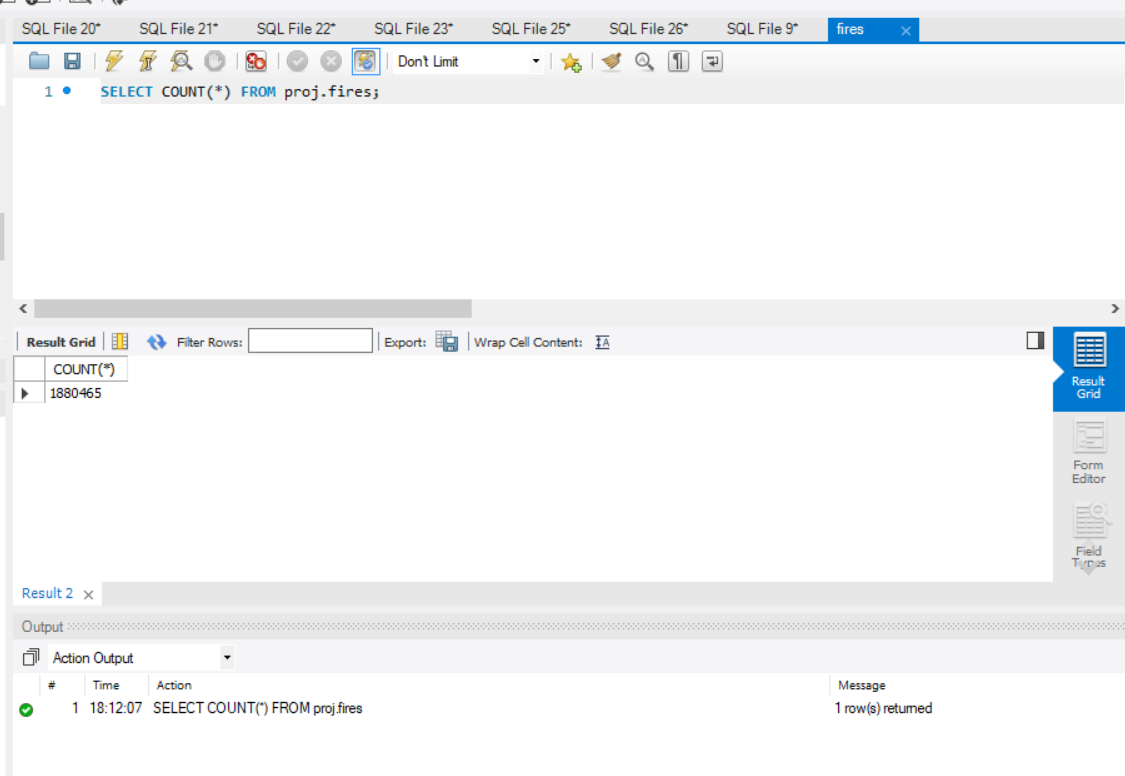
NWCG\_UNIT\_TYPE: 

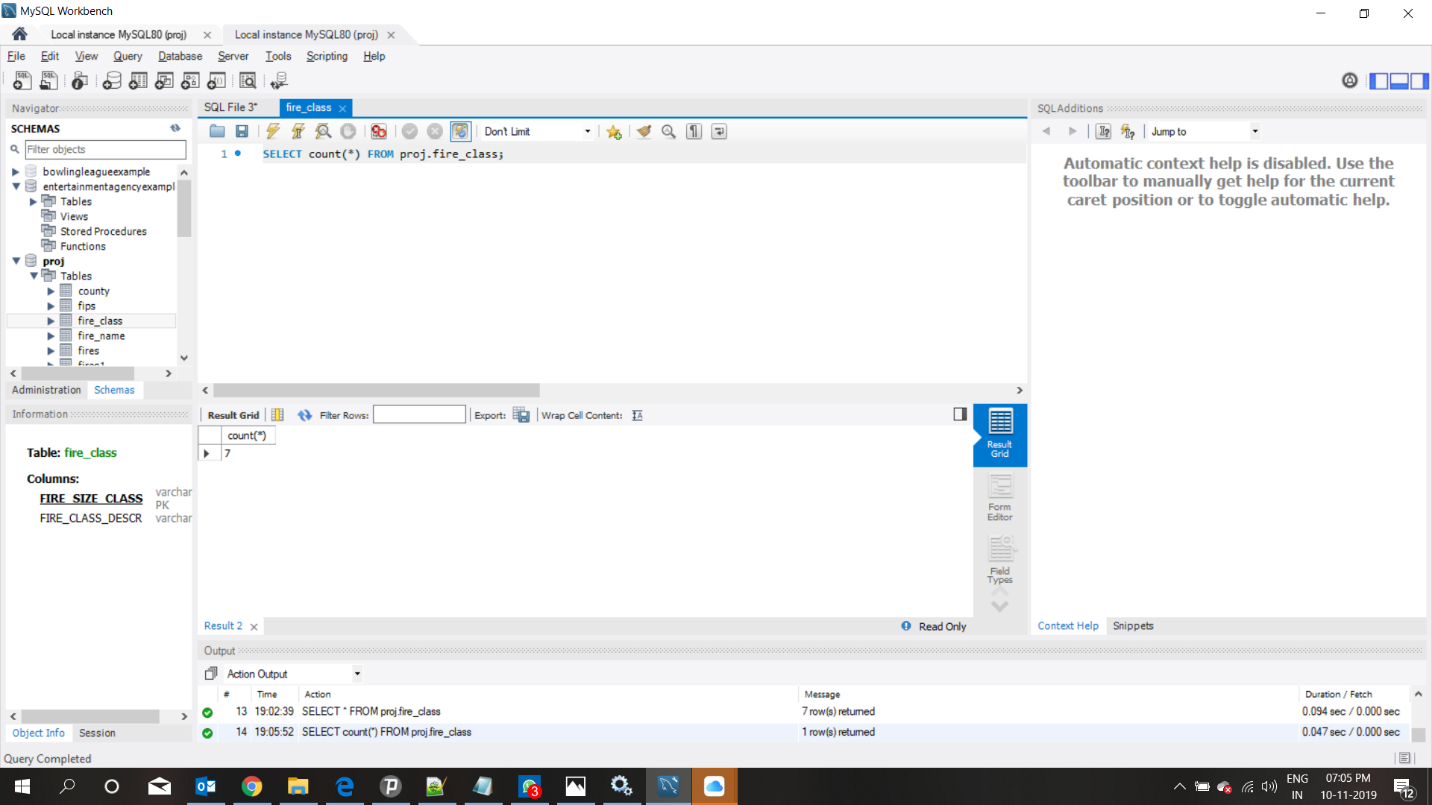
GEOGRAPHIC\_AREAS: 

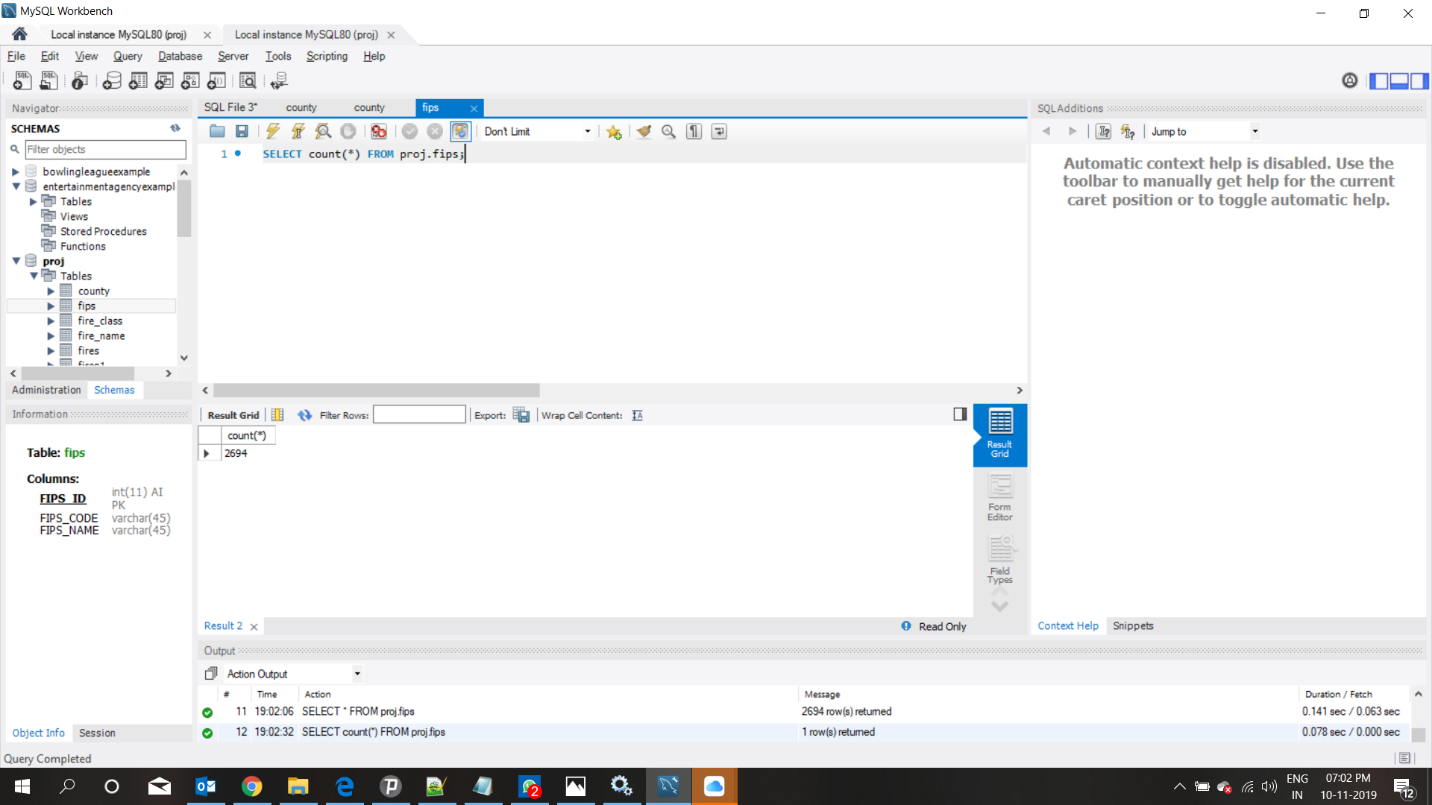
GACC: 

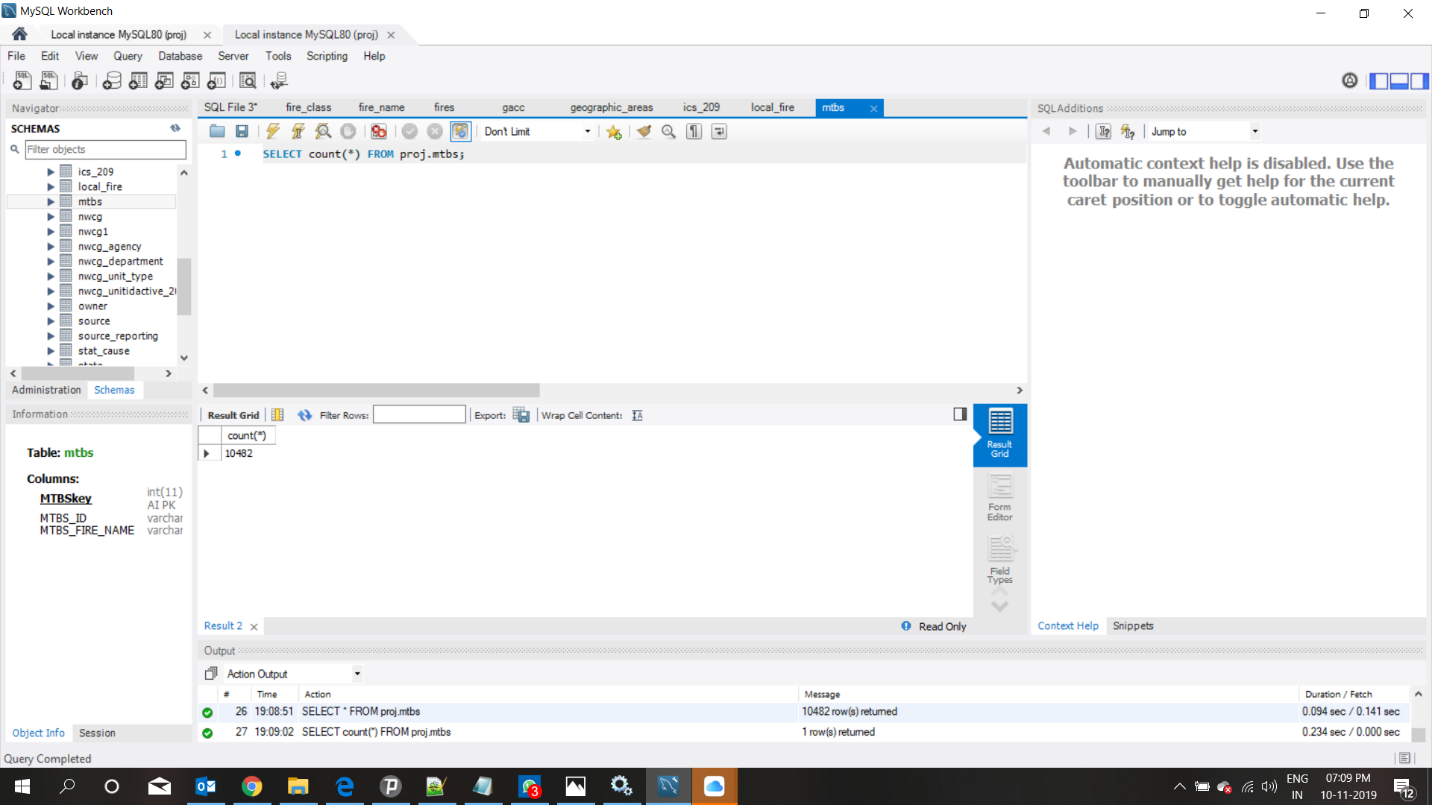
STATE: 

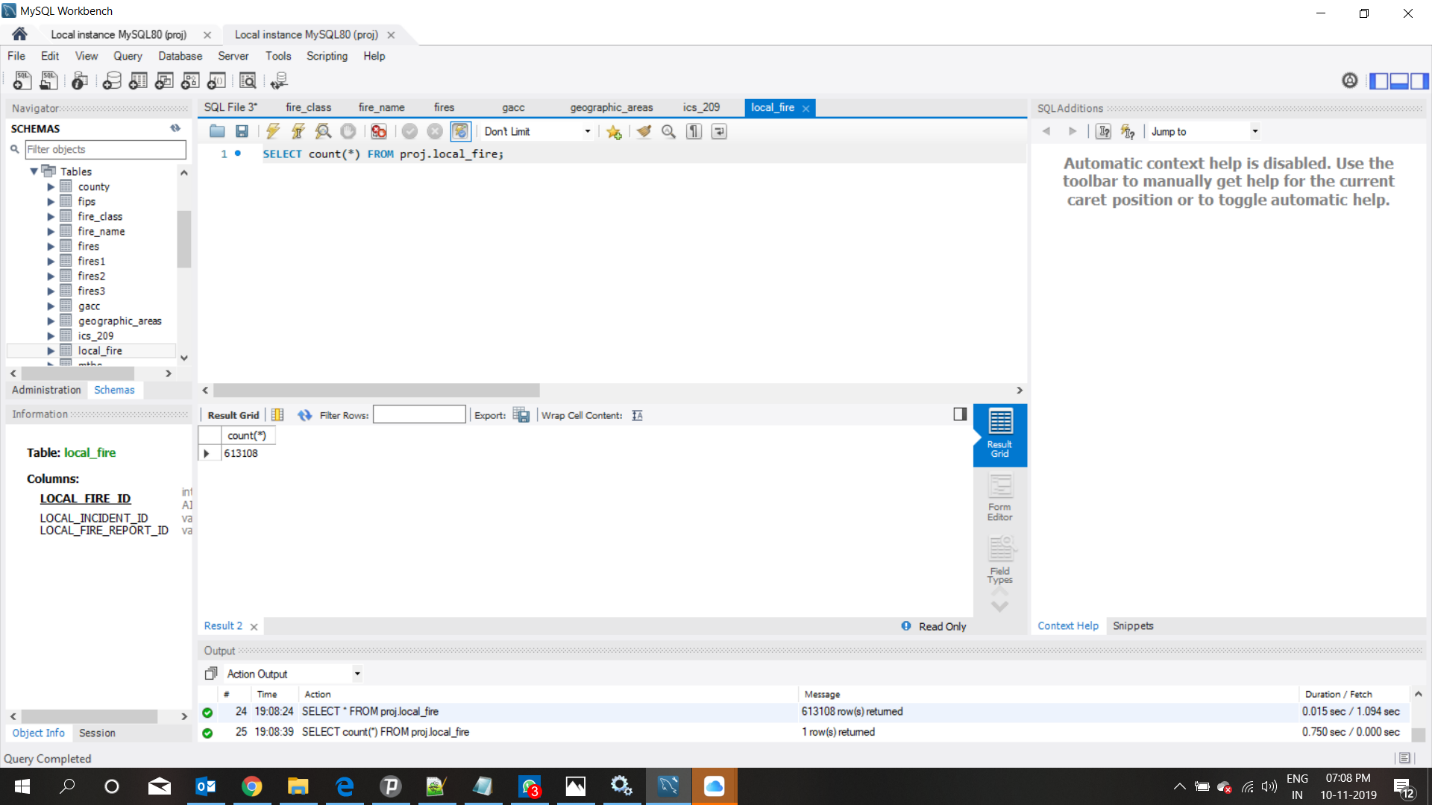
FIRES:

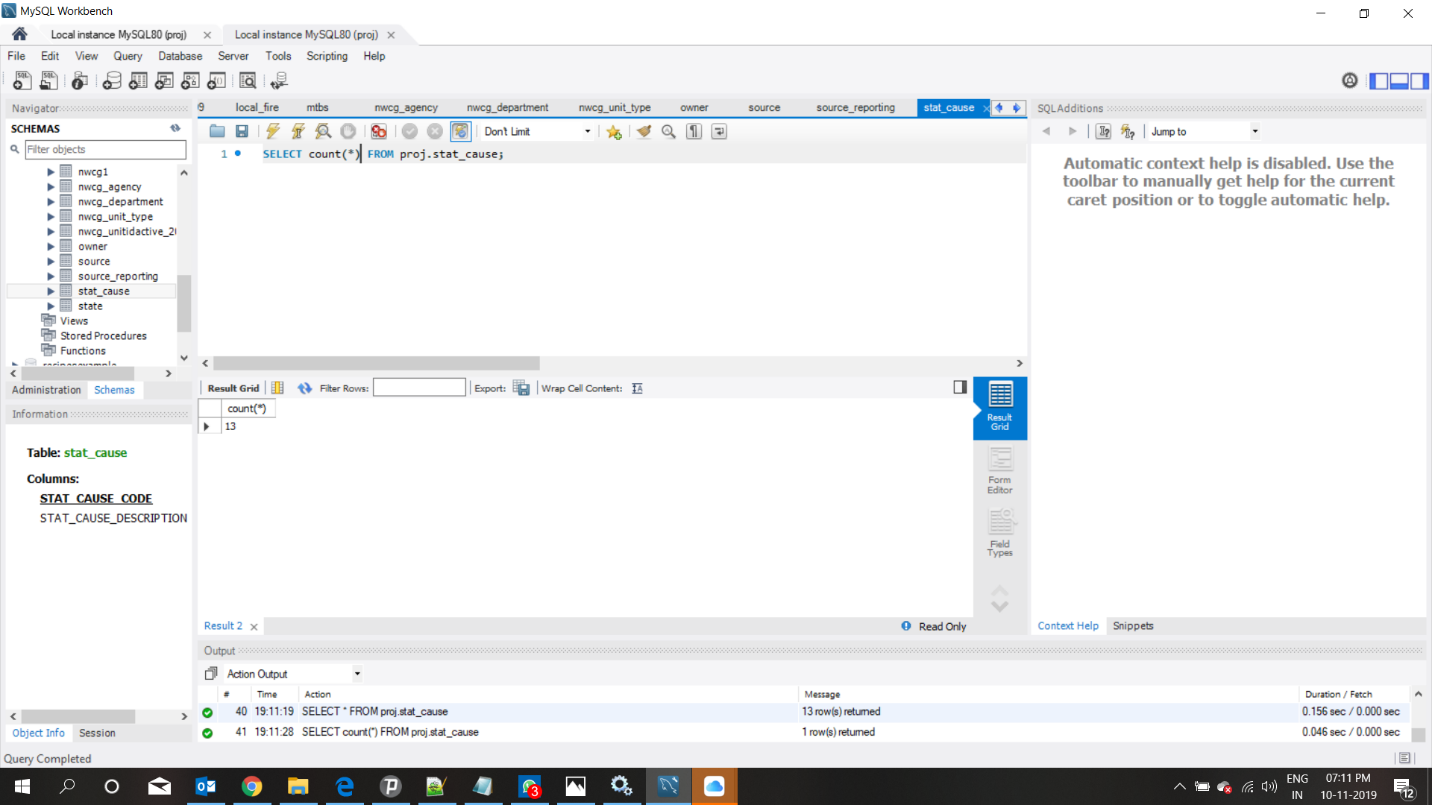


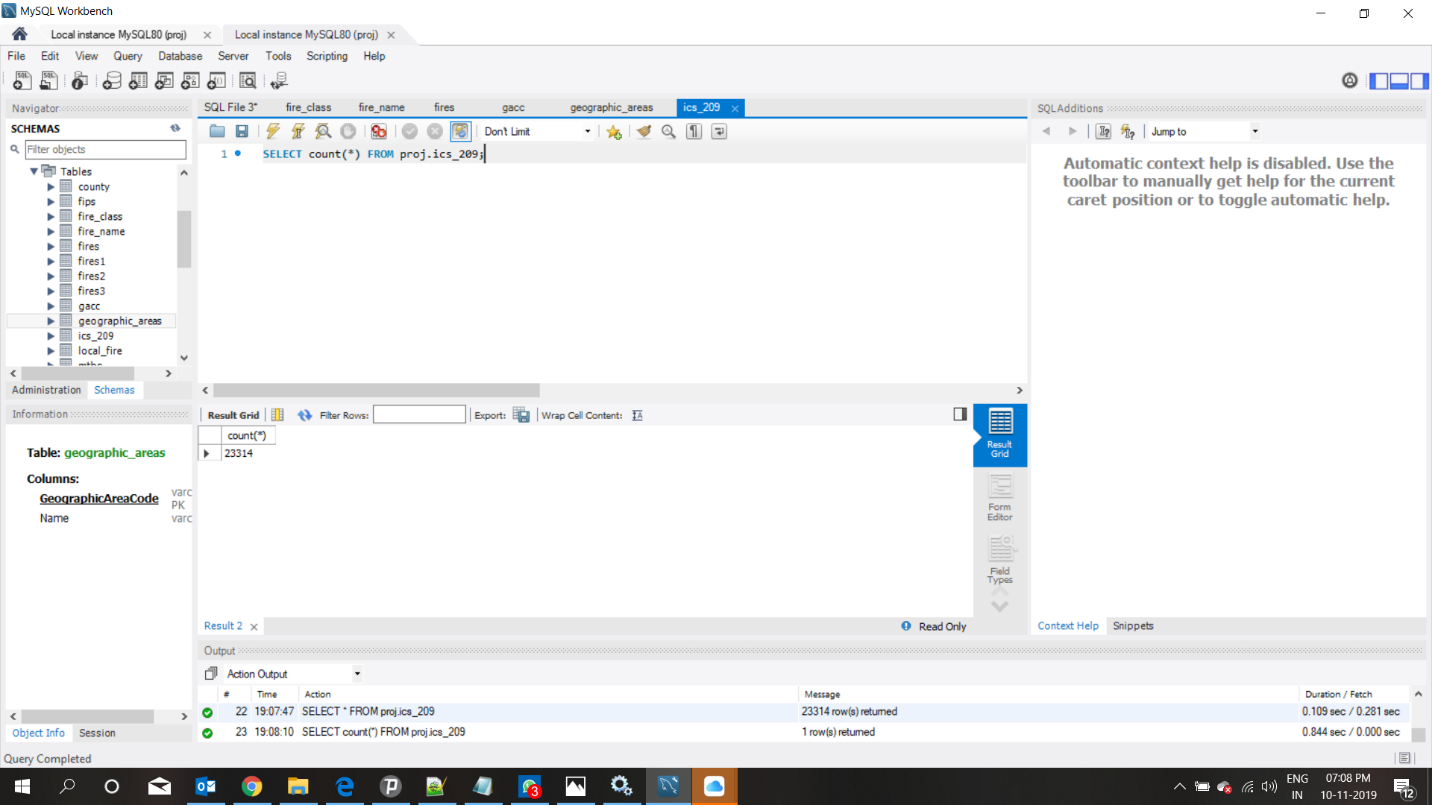
FIRE\_CLASS: 

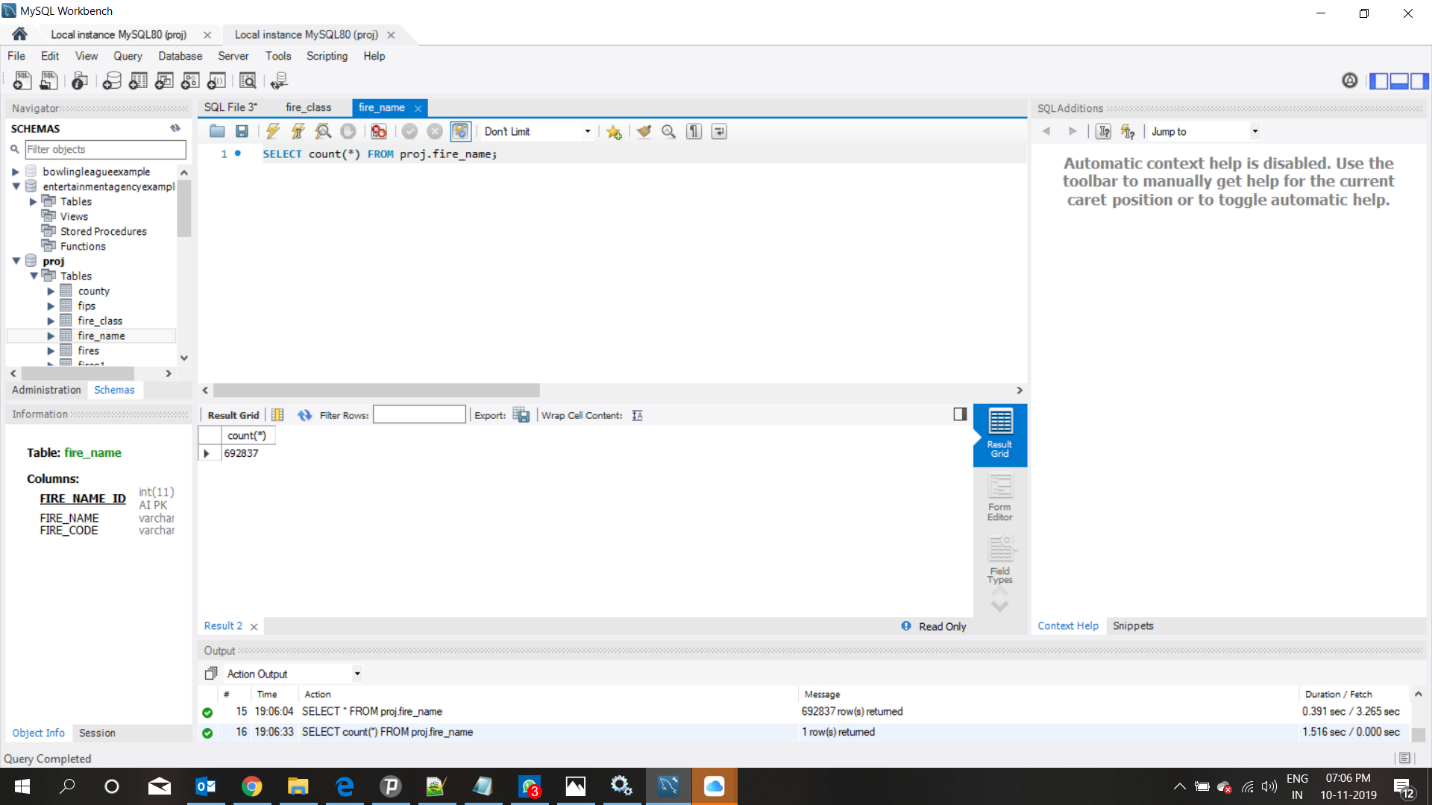
FIPS: 

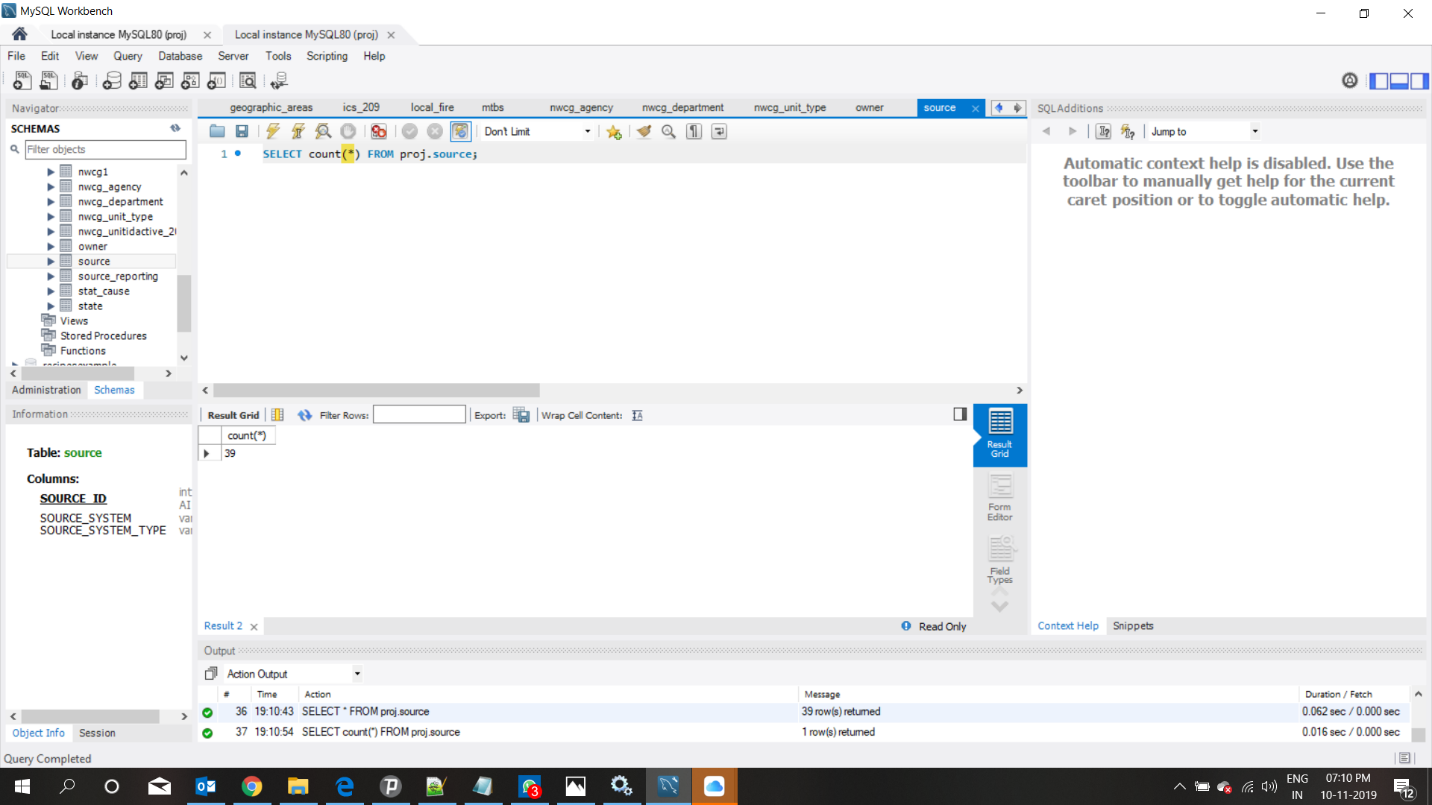
MTBS: 

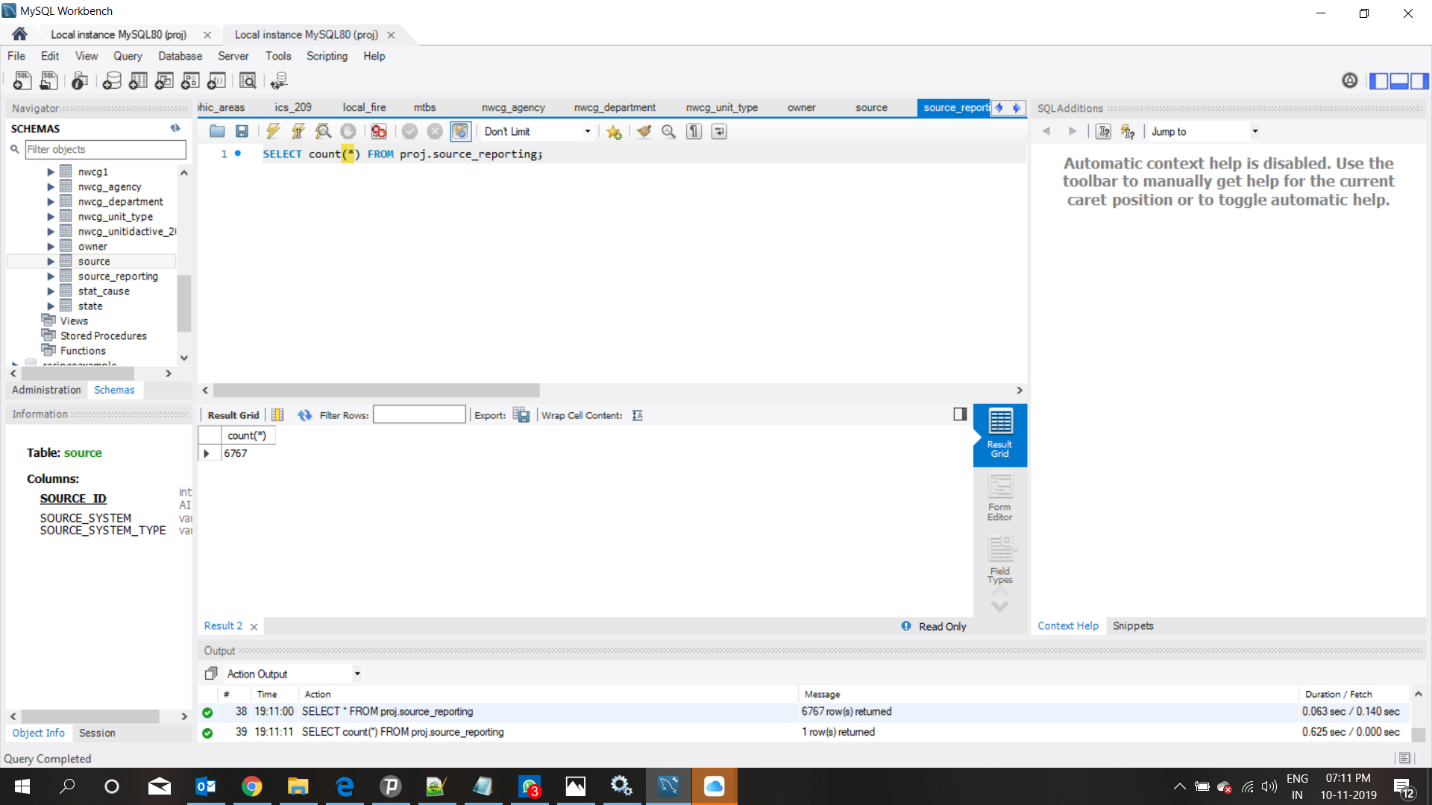
LOCAL\_FIRE: 

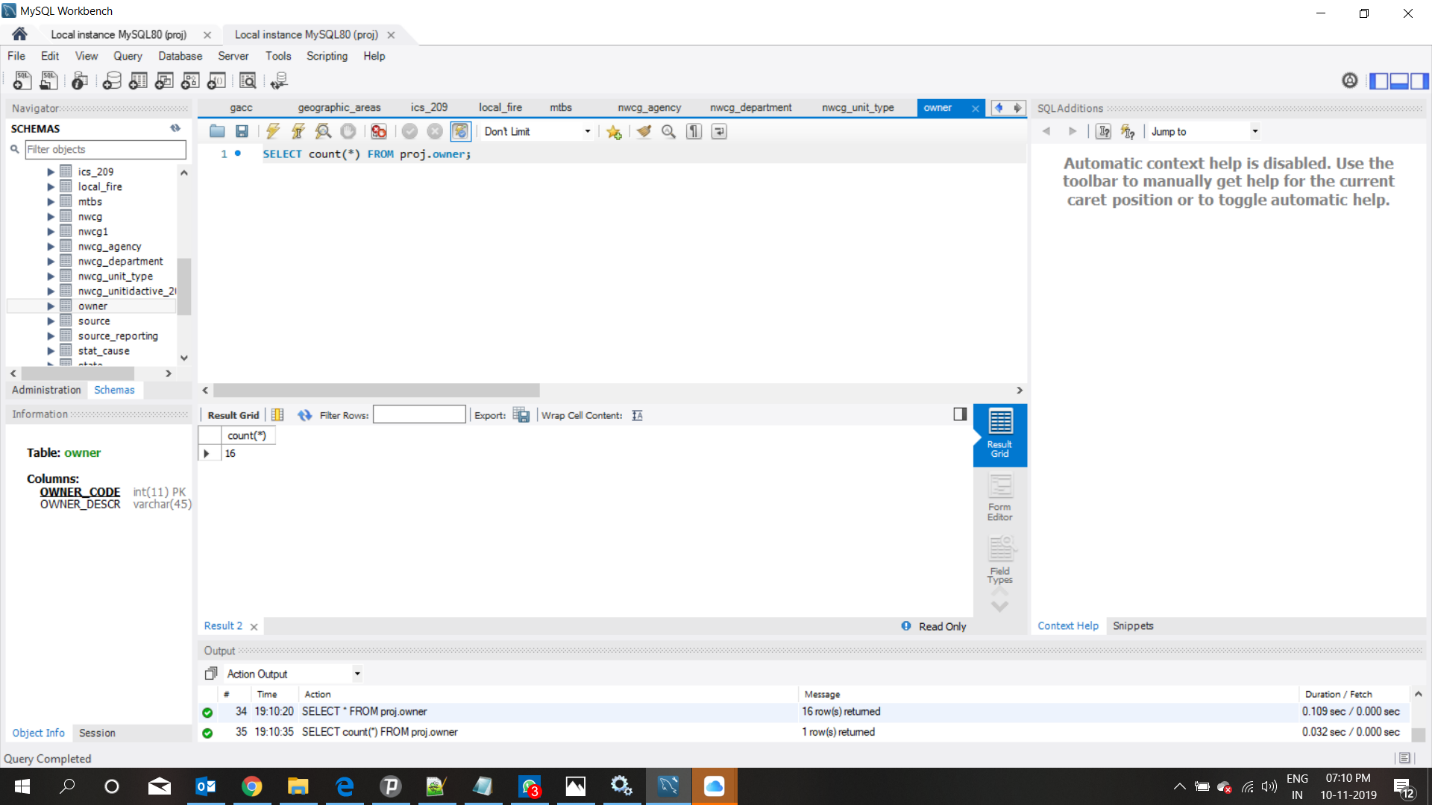
STAT\_CAUSE: 

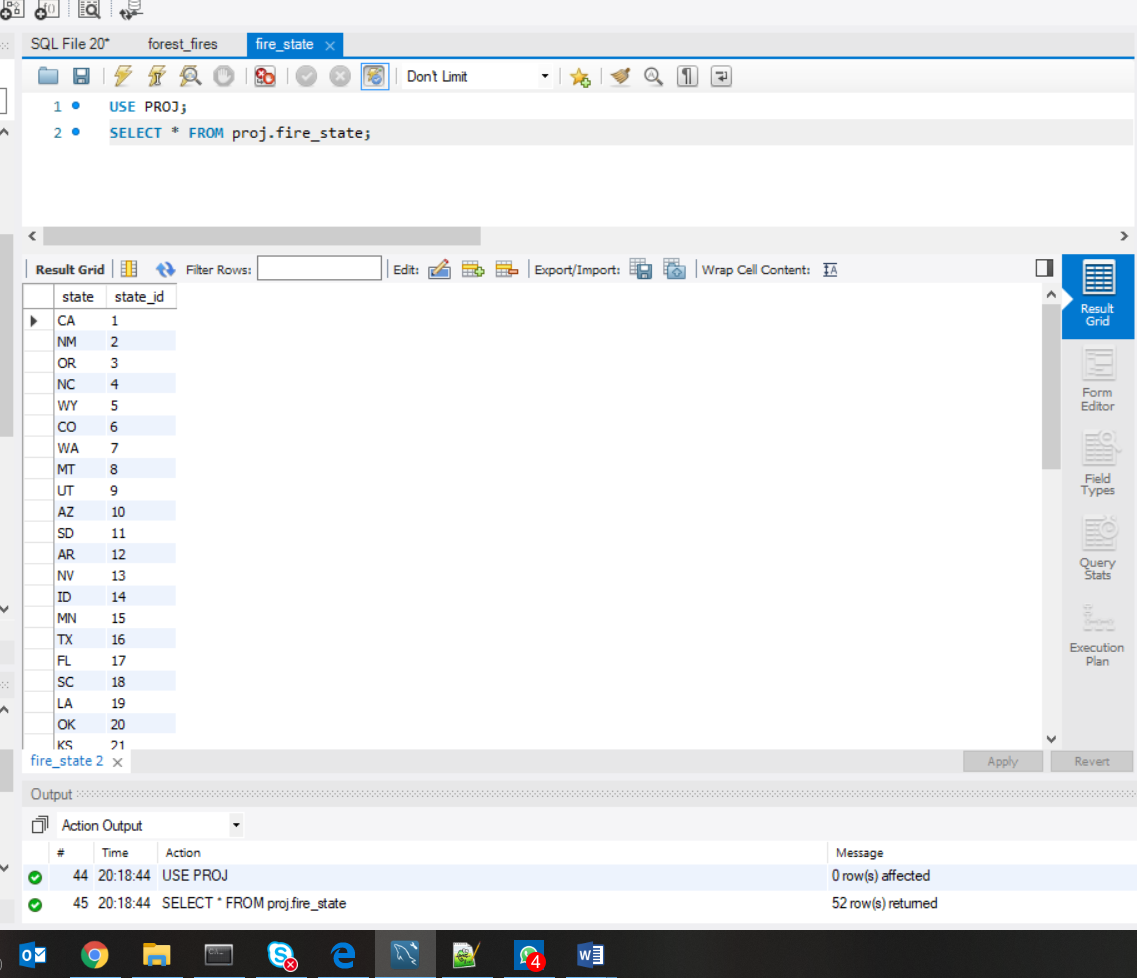
ICS\_209: 

FIRE\_NAME: 

SOURCE: 

SOURCE\_REPORTING: 

OWNER: 

FIRE\_STATE: 

## Data in the Database

|  |  |  |  |
| --- | --- | --- | --- |
| **Table Name** | **Primary Key** | **Foreign Key** | **# of Rows in Table** |
| NWCG  NWCG\_DEPARTMENT  NWCG\_ANGENCY  NWCG\_UNIT\_TYPE  GEOGRAPHIC\_AREAS  GACC  STATE  FIRES  FIRE\_CLASS  FIPS  MTBS  LOCAL\_FIRE  STAT\_CAUSE  ICS\_209  FIRE\_NAME  SOURCE  SOURCE\_REPORTING  OWNER  FIRE\_STATE | UnitID  DepartmentID  AgencyID  UnitTypeCode  GeographicAreaCode  GaccCode  StateId  FOD\_ID  FIRE\_SIZE\_CLASS  FIPS\_ID  MTBS\_ID  LOCAL\_FIRE\_ID  STAT\_CAUSE\_CODE  ICS\_209\_INT  FIRE\_NAME\_ID  SOURCE\_ID  SOURCE\_REPORTING\_ID  OWNER\_CODE  STATE\_ID | UnitTypeCode  DepartmentID  AgencyID  Staeteid  Gacccode  Geographicareacode  FIRE\_NAME\_ID  ICS\_209\_ID  MTBS\_ID  FIRE\_SIZE\_CLASS  OWNER\_CODE  FIPS\_ID  LOCAL\_FIRE\_ID  STAT\_CAUSE\_CODE  SOURCE\_REPORTING\_ID  COUNTY\_ID  SOURCE\_ID | 5867  124  55  7  11  12  68  1880465  7  2694  10482  613108  13  23314  682937  39  6767  16  52 |

# SQL Queries

## Query 1

### Question

A leading beverage company has announced a billion-dollar fund for removing debris from forests, rivers and mountains in the US. All states are interested. Which 2 states have the least chance to win a share of the fund?

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

The two states which has the least number of fire causes because of debris burning has the least chance to win the fund.

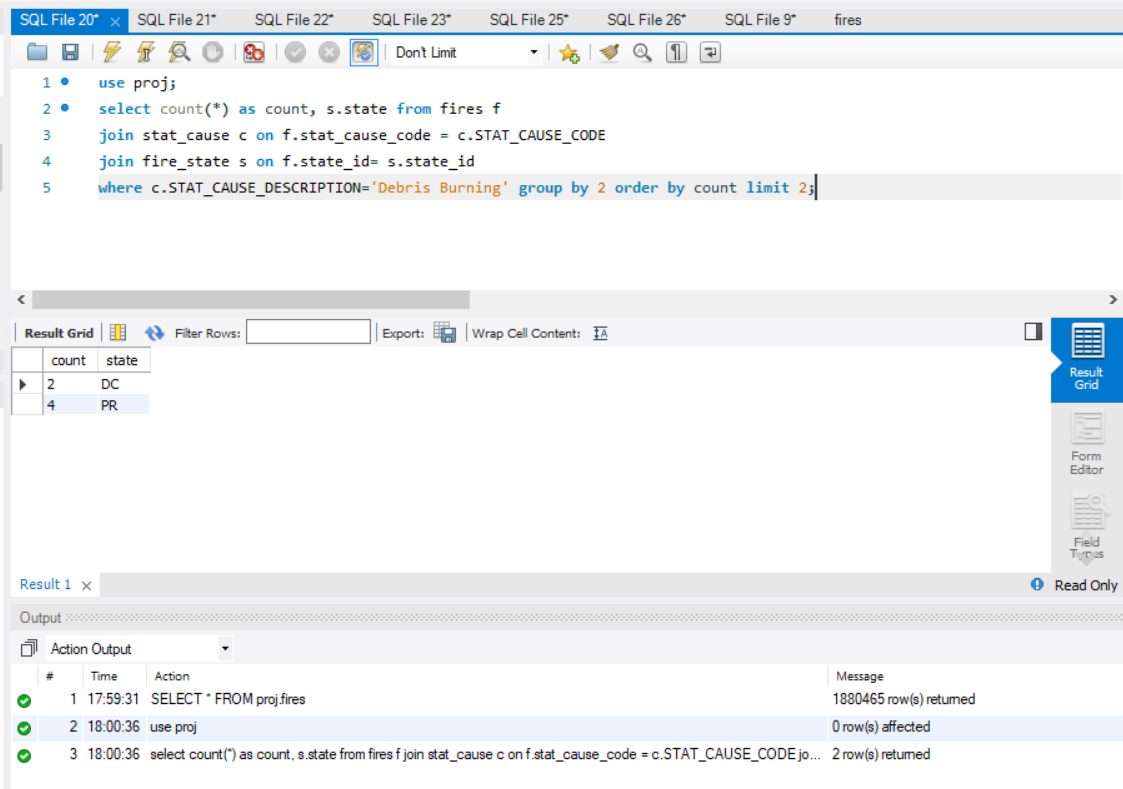
Result- DC and PR have the least chance of winning this as they contain least no of fires because of debris

### Translation

SELECT COUNT(\*) ,STATES FROM FIRES TABLE JOINING WITH STAT\_CAUSE ON MATCHING STAT\_CAUSE\_CODE JOIN FIRE\_STATE ON STATE\_ID WHERE STAT CAUSE DESCRIPTION IS DEBRIS BURNING AND GROUP IT BY STATES ORDER BY COUNT IN ASCENDING LIMIT TO 2

### Screen Shot of SQL Query and Results

Number of Rows=2



## Query 2

### Question

One of the reporting agencies has suggested that children be banned from its forests unless there is one adult for every 4 children in a group visiting a forest. Name top 5 forests where this would be the least appropriate.

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

Find the 5 forest where the fires caused by children are least

Assumption: nwcg has one unit for each forest and the forest unit is named with the naming convention “national forest”

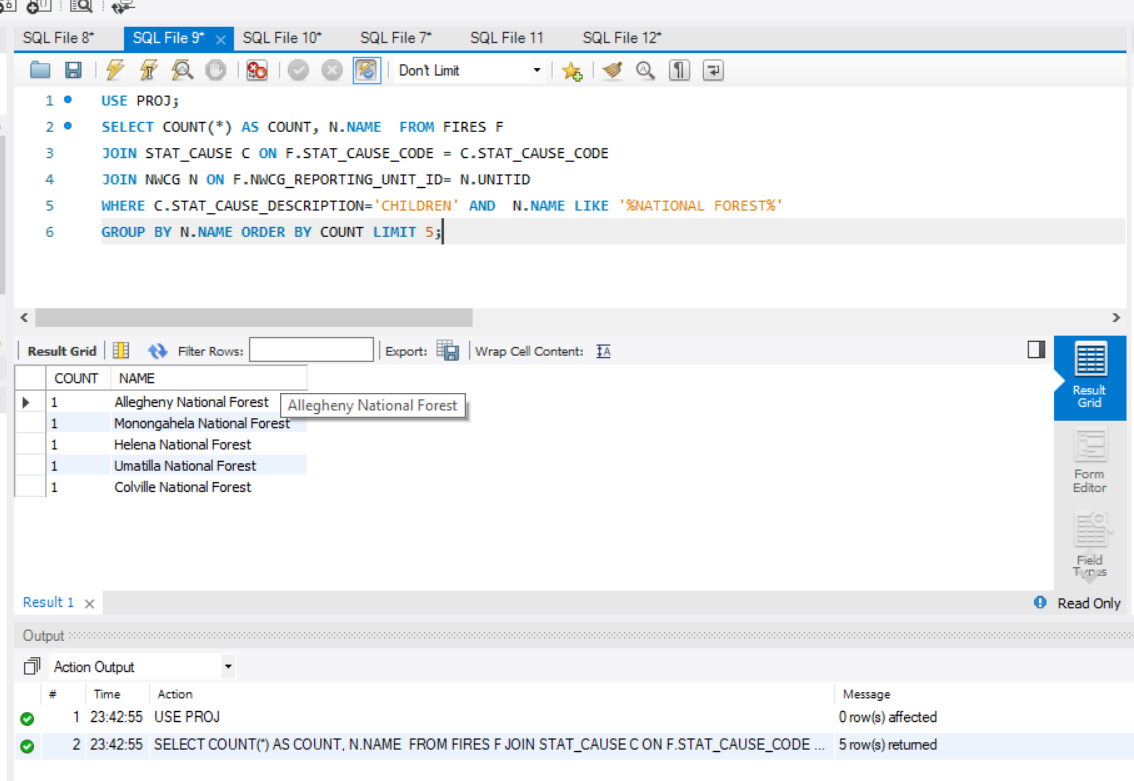
Result: There are many forests which has least equal no of fires caused by children but we are limiting it 5.

### Translation

SELECT COUNT(\*) , NAME FROM FIRES JOINIG WITH STAT\_CAUSE ON MATCHING STAT\_CAUSE\_CODE JOIN NWCG ON MATCHING UNITID WHERE STAT\_CAUSE DESCRIPTION IS CHILDREN AND NAME IS LIKE “NATIONAL FOREST” GROUP THEM BY NAME ORDER BY COUNT AND LIMIT TO 5

### Screen Shot of SQL Query and Results

Number of Rows-5



## Query 3

### Question

One advocacy group says human actions and nature are equally to blame for most wildfires. Write a query that can help determine the truth of this statement.

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

Count the number of actions by humans and count number of actions by nature that caused wildfires

Assumption: Misc and unknown/missing are assumed as not human actions. All others except lighting is called by human and lighting is caused by nature

Result - Human action-1278192, Nature-278468

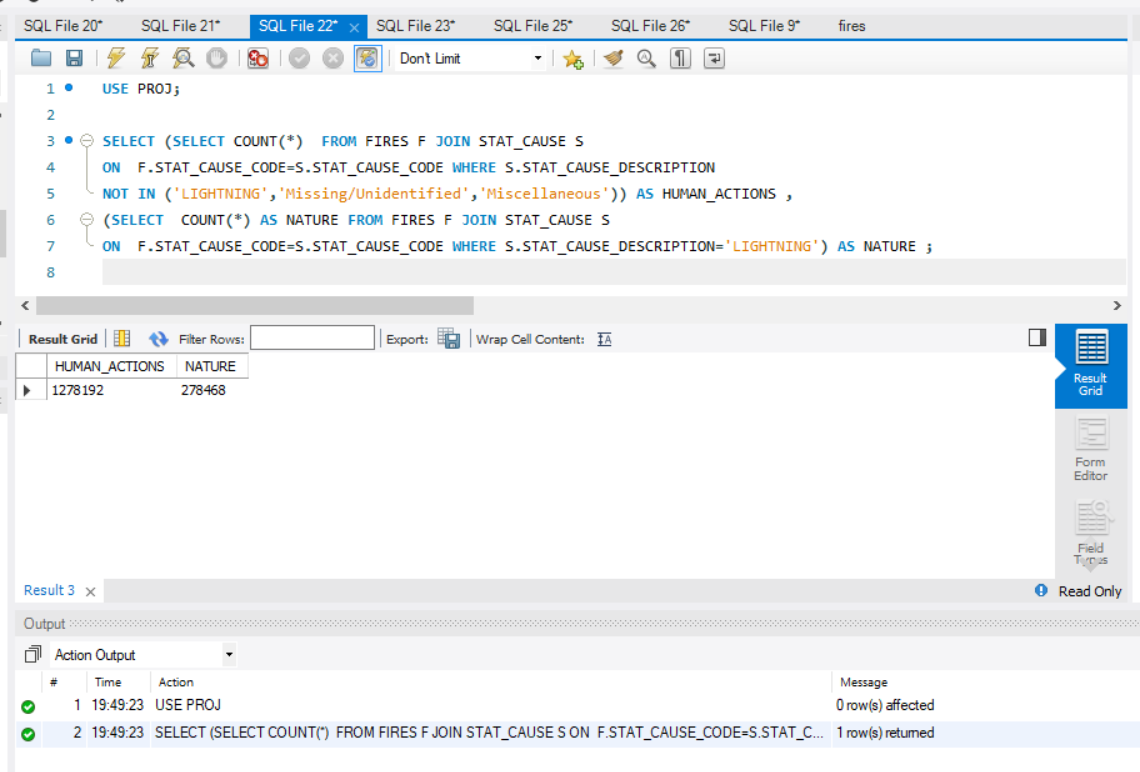
### Translation

Select (select count(\*) from fires joining stat\_cause on matching stat\_cause\_code where stat\_cause\_descr is not lighting, Missing/Unidentified,Miscellaneous) name it as humans,

(select count(\*) from fires join stat\_cause on matching stat\_cause\_code where stat\_cause\_descr is lighting) name it as nature;

### Screen Shot of SQL Query and Results

Number of rows-1



## Query 4

### Question

Which state had fires only in the second half of the calendar years?

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

Find the states which has fires only in the second half of the year that is the list of states which had no fires in first half of the year.

We have considered discovery\_doy (day of the year) so 365/2 is 183 and so we have used less than 183 in the query

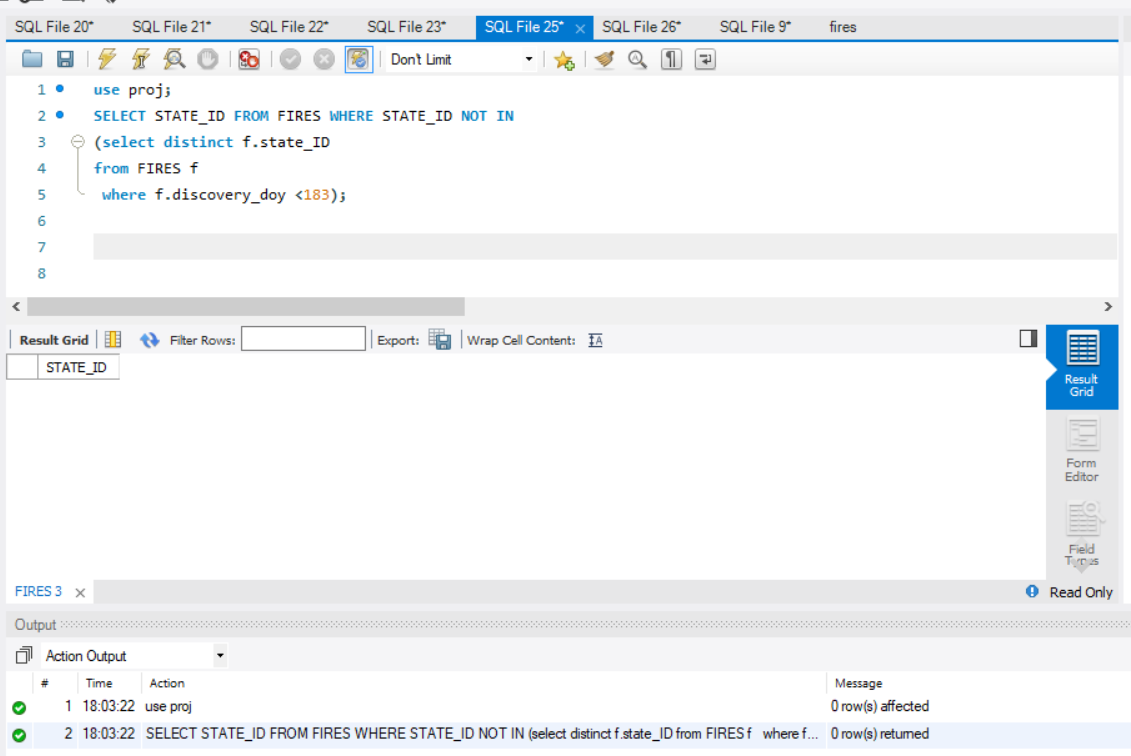
Result-none

### Translation

SELECT STATE\_ID FROM FIRES TABLE WHERE STATE\_ID IN NOT IN ( SELECT DISTINCT VALUES OF STATE\_ID FROM FIRES WHERE THE DISCOVERY\_DOY IS LESS THAN 183)

### Screen Shot of SQL Query and Results

Number of rows-0



## Query 5

### Question

Which forest had the number of fires equal to the average number of wild fires in the US?

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

Assumption: Average of US wild fires are is average of all years by no of days i.e.

2015-1992 is 23 years so 23\*365days as written is query.

There is no value equal to average so we are taking a range of 0 to 5

Result- Olympic National Forest

### Translation

SELECT COUNT(\*) , NAME FROM FIRES TABLE JOINING WITH NWCG ON MATCHING UNITID WHERE NAME IS LIKE ‘NATIONAL FOREST’ GROUP BY NAME HAVING ( COUNT –(( SELECT COUNT(\*) FROM FIRES) / (2015-1992)\*365)) BETWEEN 0 AND 5

### Screen Shot of SQL Query and Results

Number of rows-1

## 

## Query 6

### Question

What were the forests that had only one fire that lasted more than two days?

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

Find forest which have fires is more than two days and with those forests find which had only one fire all throughout the reporting data.

Assumption: nwcg has one unit for each forest and the forest unit is named with the naming convention “forest” and “wildlife”

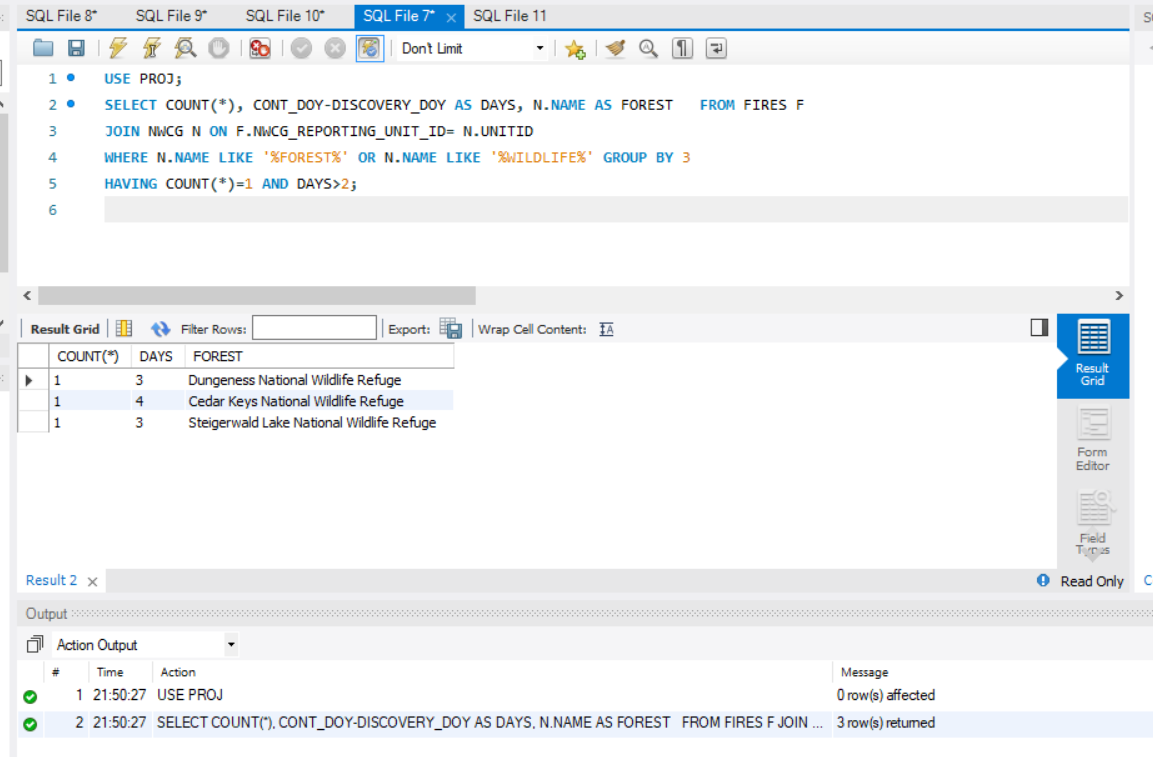
Result- 3 days with Dungeness National Wildlife Refuge,4 days with Cedar Keys National Wildlife Refuge and 3 days with Steigerwald National Wildlife Refuge

### Translation

SELECT COUNT(\*), CONTAINED\_DOY-DISCOVERY\_DOY NAMED AS DAYS, NAME FROM FIRES JOINING WITH NWCG TABLE ON MATCHING UNITID WHERE NAME IS LIKE “FOREST” OR “WILDLIFE” GROUP THEM BY NAME HAVING FIRE COUNT=1 AND DAY>2

### Screen Shot of SQL Query and Results

Number of rows-3



# Data Review for MongoDB

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

* The 2 collections in our dataset are Fires and NWCG. The fires collection have 1.8M documents and 39 fields each. The NWCG collection has 5867 documents with 13 fields each.
* Both the collections are related through the common field, unit\_id.

Assumptions:

• NWCG\_reporting\_unit\_name consists of fire stations inside and outside the forest. The names ending with “national forest” are considered to be the forest names.

• If children are supervised by an adult, it is assumed that it will prevent fire.

• The debris burning is assumed to be the cause of more debris.

# Physical Mongo Database

## Assumptions/Notes About Data Set:

Empty data:

In NWCG collection, the parent field is all NULL.

In both NWCG and FIRES collections, the field OBJECT\_ID has been removed.

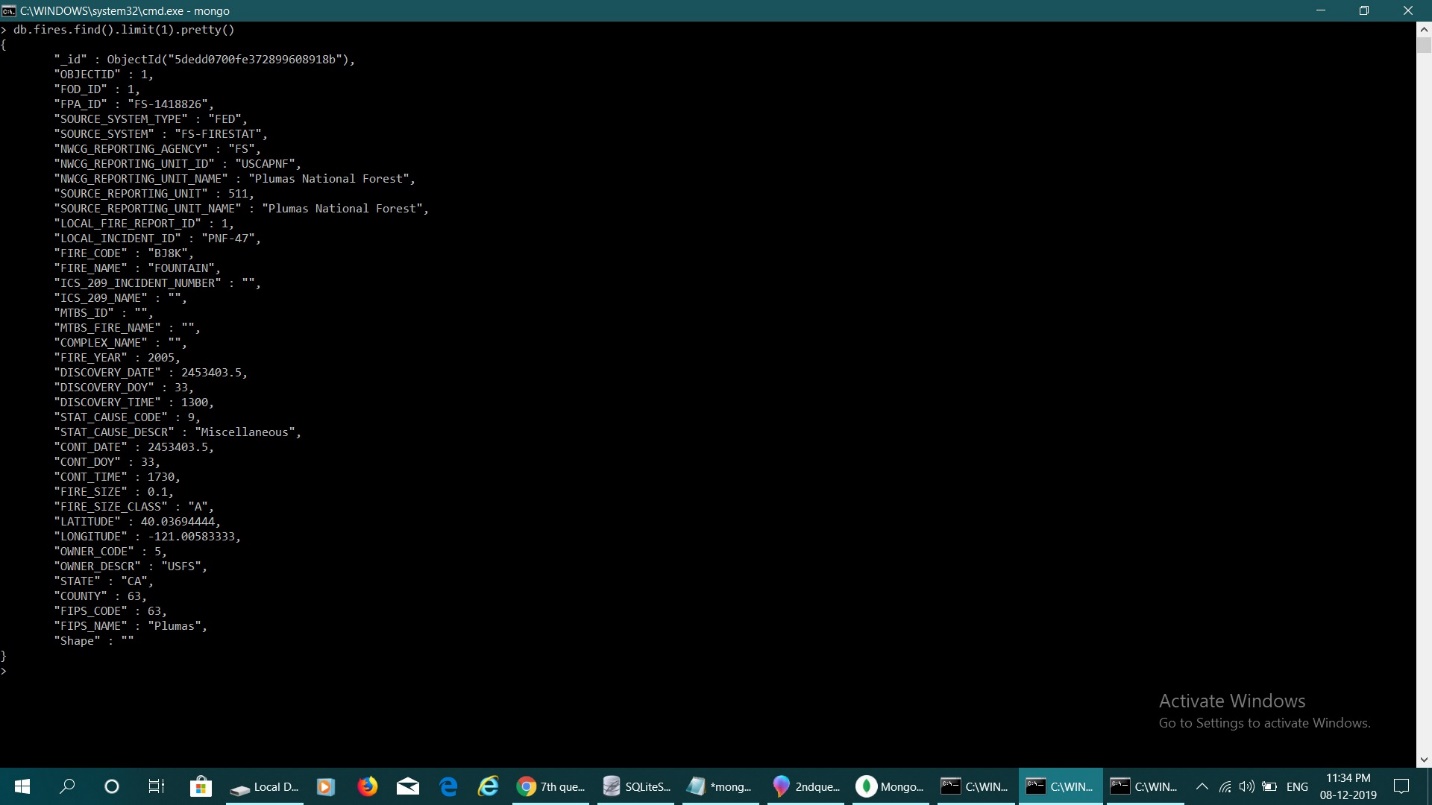
Bad data:

In FIRES collection, shape field is NaN.

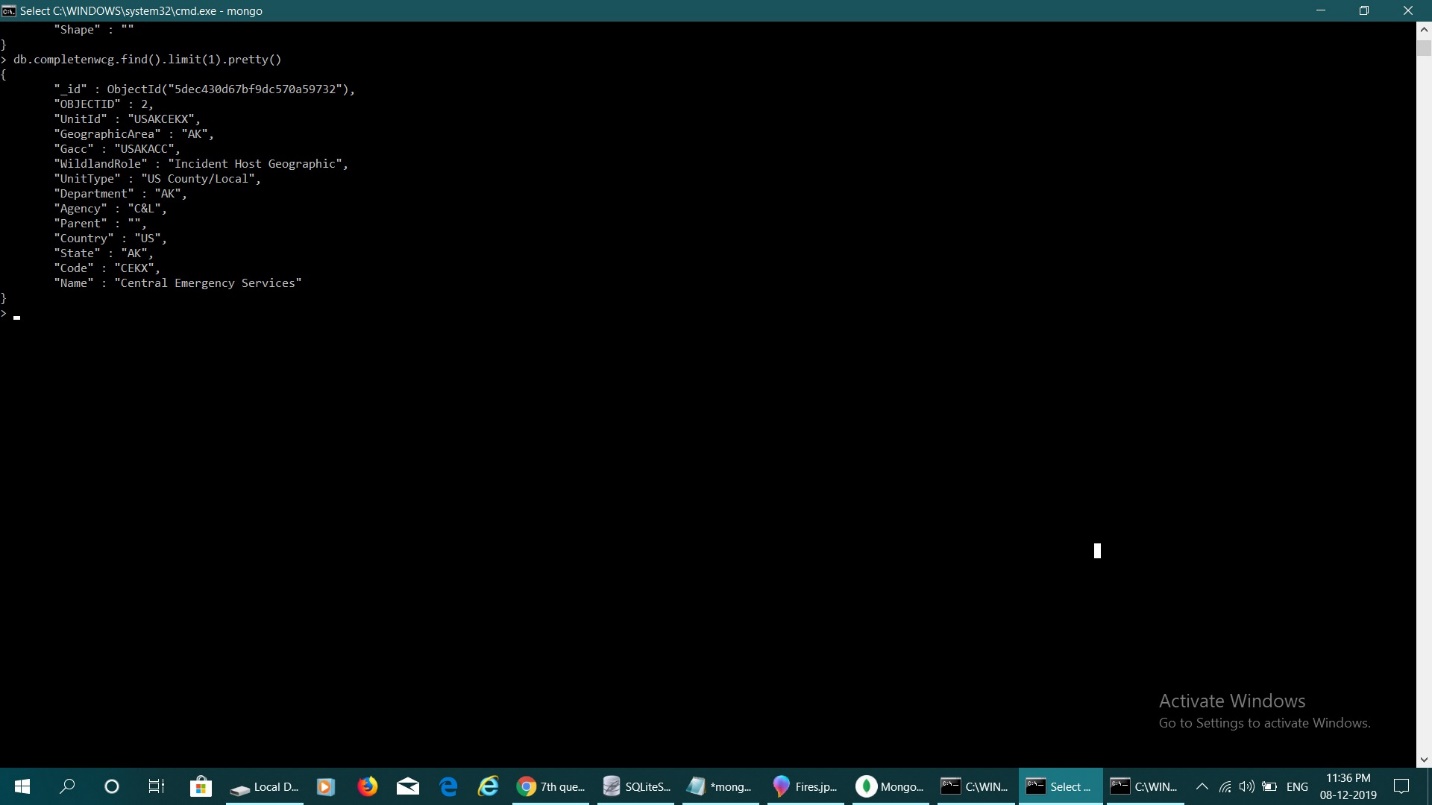
Sparse data: The fields mentioned below are not completely filled.

* In fires collection, the fields MTBS\_ID, MTBS\_FIRE\_NAME, COMPLEX\_NAME, CONT\_DATE, CONT\_TIME, CONT\_DOY, DISCOVERY\_TIME, ICS\_209\_NAME and ICS\_209\_INCIDENT\_NUMBER.

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

FIRES

## NWCG



## Data in the Database:

|  |  |  |
| --- | --- | --- |
| **Collection Name** | **Relationships With Other Collections (if any)** | **# of Documents in Collection** |
| fires | Related to NWCG through Unit\_id. | 1880465 |
| NWCG | Related to fires table through unit\_id. | 5867 |

# MongoDB Queries/Code

## Query 1

### A leading beverage company has announced a billion-dollar fund for removing debris from forests, rivers and mountains in the US. All states are interested. Which 2 states have the least chance to win a share of the fund?

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

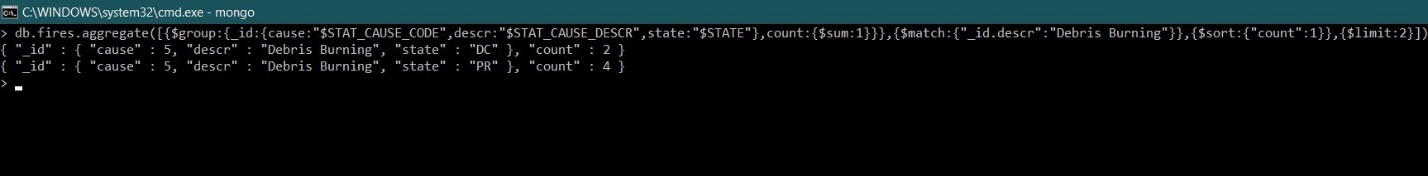
The two states which has the least number of fire causes because of debris burning has the least chance to win the fund.

Result- DC and PR have the least chance of winning this as they contain least no of fires because of debris

### Translation:

Using the aggregation pipeline, we grouped the data according to STAT\_CAUSE\_CODE, STAT\_CAUSE\_DESCR and STATE and totalled where the cause of fire is Debris burning. Then, we sorted our results in ascending order and limited the number of results to 2.

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



## Query 2

### One of the reporting agencies has suggested that children be banned from its forests unless there is one adult for every 4 children in a group visiting a forest. Name top 5 forests where this would be the least appropriate.

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

Find the 5 forest where the fires caused by children are least

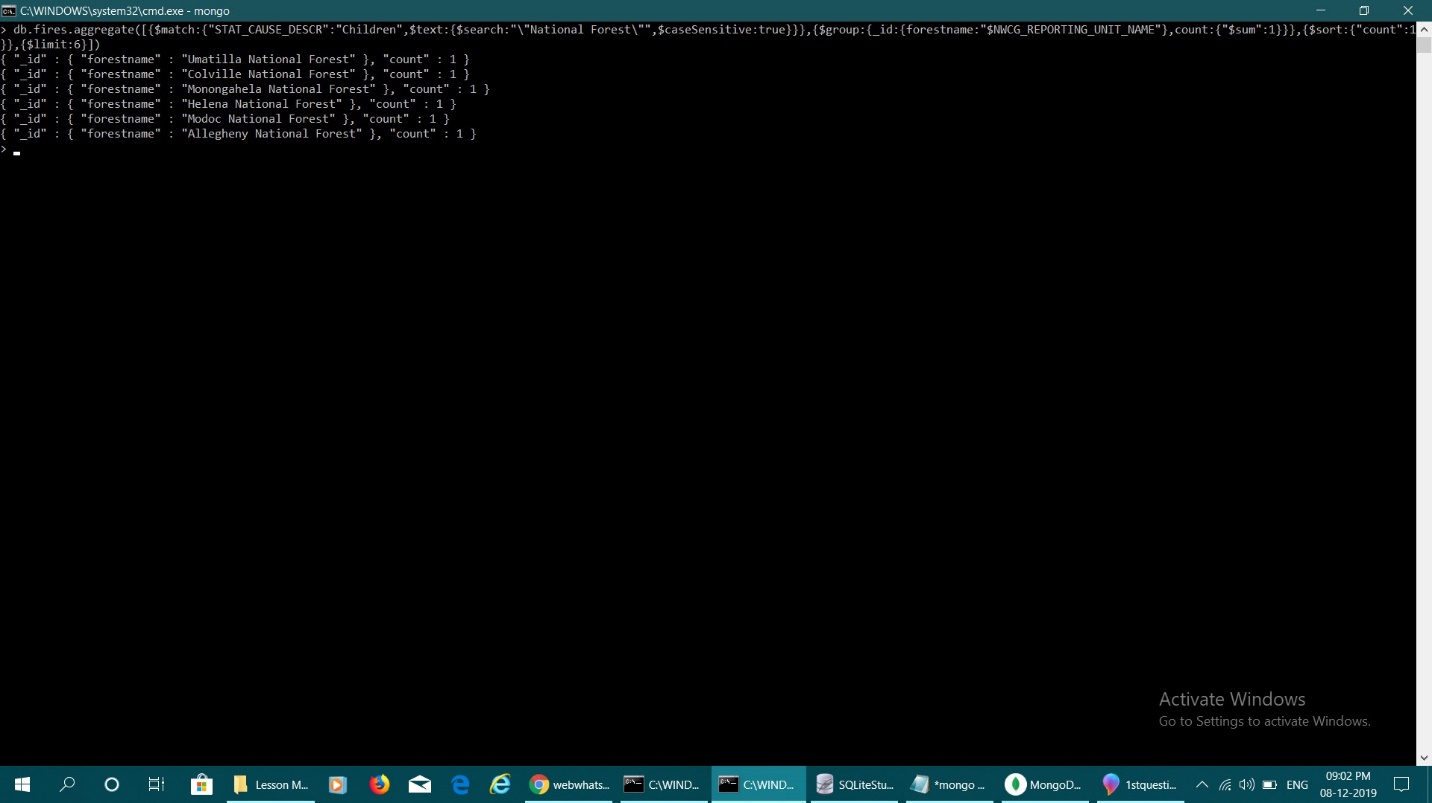
Assumption: nwcg has one unit for each forest and the forest unit is named with the naming convention “NATIONAL FOREST”

Result: There are many forests which has least equal no of fires caused by children but we are limiting it 5. The result will be different from mongo because of this .

### Translation:

Using aggregation pipeline, we filter the results using the stat\_cause\_descr as children and obtain forest name matching with national forest, group using the forest name and count, sort in ascending order with the count and limit the number of results to 6.

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



## Query 3

One advocacy group says human actions and nature are equally to blame for most wildfires. Write a query that can help determine the truth of this statement.

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

Count the number of actions by humans and count number of actions by nature that caused wildfires

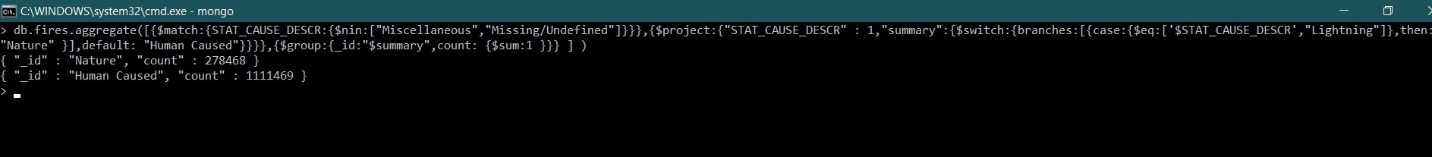
Assumption: Misc and unknown/missing are assumed as not human actions. All others except lighting is called by human and lighting is caused by nature

Result - Human action-1278192, Nature-278468

### Translation:

Match the stat\_cause\_descriptions that are not miscellaneous or missing, Project the fields Stat\_cause\_descr and the summary where lightning is shown as nature caused and all other stat\_cause\_descr are considered human caused, group using the summary.

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



## Query 4:

## Which forest had the number of fires equal to the average number of wild fires in the US?

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

Assumption: Average of US wild fires are is average of all years by no of days i.e.

2015-1992 is 23 years as written is query.

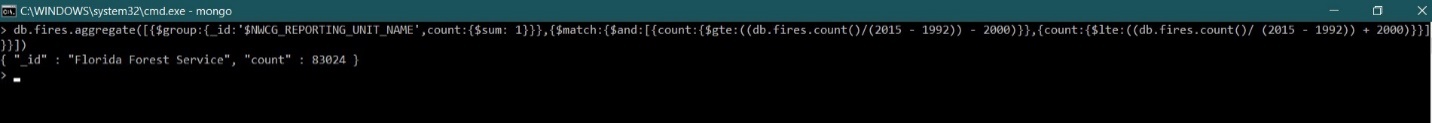
There is no value equal to average so we are taking a range

Here not taking days

### Translation:

Group the name of the forest by the NWCG reporting unit name and then count each category. This will give us the number of fires in that forest. Compare this number to the average number of fires in the forests of US and find the forest which equals the value.

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



## Query 5

### Question

Which state had fires only in the second half of the calendar years?

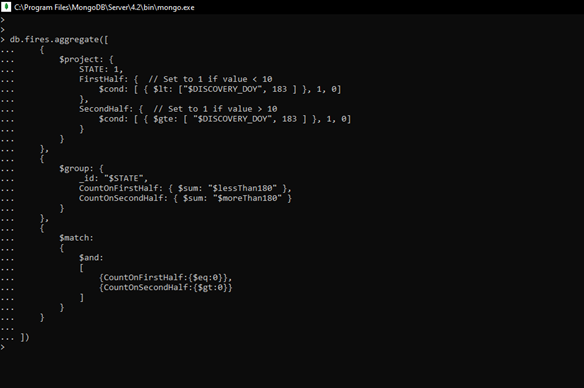
### Notes/Comments About MongoDB Query/Code and Results (Include # of Documents in Result)

Find the states which has fires only in the second half of the year that is the list of states which had no fires in first half of the year.

We have considered discovery\_doy (day of the year) so 365/2 is 183 and so we have used less than 183 in the query

Result-none

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



## Query 6

### Question

What were the forests that had only one fire that lasted more than two days?

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

Find forest which have fires is more than two days and with those forests find which had only one fire all throughout the reporting data.

Assumption: nwcg has one unit for each forest and the forest unit is named with the naming convention “forest” and “wildlife”

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

### 