Process of data import into MySQL workbench:

The following tables were imported into R as R dataframe objects using the following code

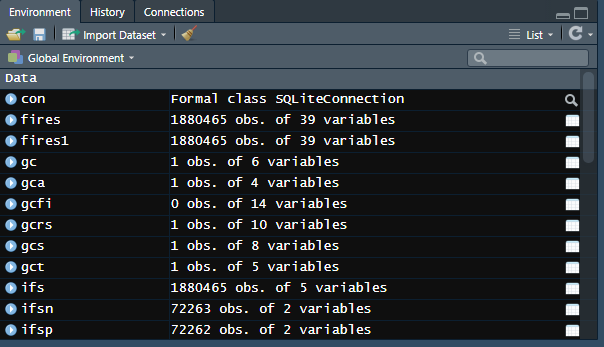
***con <- dbConnect(SQLite(),'FPA\_FOD\_20170508.sqlite')***

***dbListTables(con)***

***fires <- dbReadTable(con,'Fires')***

1. Fires
2. geometry\_columns\_statistics
3. geometry\_columns\_time
4. idx\_fires\_shape\_rowid
5. idx\_fires\_shape\_parent
6. idx\_fires\_shape\_node
7. idx\_fires\_shape
8. geometry\_columns\_statistics
9. geometry\_columns\_auth
10. geometry\_columns
11. geom\_cols\_ref\_sys
12. geometry\_columns\_statistics
13. geometry\_columns\_time
14. idx\_fires\_shape
15. idx\_fires\_shape\_node
16. idx\_fires\_shape\_parent
17. idx\_fires\_shape\_rowid
18. NWCG\_UnitIDActive\_20170109
19. spatialite\_history
20. spatial\_ref\_sys\_aux
21. spatial\_ref\_sys\_all
22. spatial\_ref\_sys
23. sqlite\_sequence
24. vector\_layers
25. vector\_layers\_auth
26. vector\_layers\_statistics

**Data loaded into R dataframes:**



**Null values for each table: (Sparse data)**

|  |  |  |
| --- | --- | --- |
| **Fires** | | |
| **Column Name** | **Count** | **Percentage** |
| OBJECTID | 0 | 0.00 |
| FOD\_ID | 0 | 0.00 |
| FPA\_ID | 0 | 0.00 |
| SOURCE\_SYSTEM\_TYPE | 0 | 0.00 |
| SOURCE\_SYSTEM | 0 | 0.00 |
| NWCG\_REPORTING\_AGENCY | 0 | 0.00 |
| NWCG\_REPORTING\_UNIT\_ID | 0 | 0.00 |
| NWCG\_REPORTING\_UNIT\_NAME | 0 | 0.00 |
| SOURCE\_REPORTING\_UNIT | 0 | 0.00 |
| SOURCE\_REPORTING\_UNIT\_NAME | 0 | 0.00 |
| LOCAL\_FIRE\_REPORT\_ID | 1459286 | 77.60 |
| LOCAL\_INCIDENT\_ID | 820821 | 43.65 |
| FIRE\_CODE | 1555636 | 82.73 |
| FIRE\_NAME | 957189 | 50.90 |
| ICS\_209\_INCIDENT\_NUMBER | 1854748 | 98.63 |
| ICS\_209\_NAME | 1854748 | 98.63 |
| MTBS\_ID | 1869462 | 99.41 |
| MTBS\_FIRE\_NAME | 1869462 | 99.41 |
| COMPLEX\_NAME | 1875282 | 99.72 |
| FIRE\_YEAR | 0 | 0.00 |
| DISCOVERY\_DATE | 0 | 0.00 |
| DISCOVERY\_DOY | 0 | 0.00 |
| DISCOVERY\_TIME | 882638 | 46.94 |
| STAT\_CAUSE\_CODE | 0 | 0.00 |
| STAT\_CAUSE\_DESCR | 0 | 0.00 |
| CONT\_DATE | 891531 | 47.41 |
| CONT\_DOY | 891531 | 47.41 |
| CONT\_TIME | 972173 | 51.70 |
| FIRE\_SIZE | 0 | 0.00 |
| FIRE\_SIZE\_CLASS | 0 | 0.00 |
| LATITUDE | 0 | 0.00 |
| LONGITUDE | 0 | 0.00 |
| OWNER\_CODE | 0 | 0.00 |
| OWNER\_DESCR | 0 | 0.00 |
| STATE | 0 | 0.00 |
| COUNTY | 678148 | 36.06 |
| FIPS\_CODE | 678148 | 36.06 |
| FIPS\_NAME | 678148 | 36.06 |
| Shape | 0 | 0.00 |

|  |  |  |
| --- | --- | --- |
| **NWCG\_UnitIDActive\_20170109** | | |
| **Column Name** | **Count** | **Percentage** |
| OBJECTID | 0 | 0.00 |
| UnitId | 0 | 0.00 |
| GeographicArea | 0 | 0.00 |
| Gacc | 0 | 0.00 |
| WildlandRole | 0 | 0.00 |
| UnitType | 0 | 0.00 |
| Department | 350 | 5.97 |
| Agency | 398 | 6.78 |
| Parent | 5867 | 100.00 |
| Country | 0 | 0.00 |
| State | 0 | 0.00 |
| Code | 0 | 0.00 |
| Name | 0 | 0.00 |

|  |  |  |
| --- | --- | --- |
| **idx\_fires\_shape** | | |
| **Column Name** | **Count** | **Percentage** |
| Pkid | 0 | 0 |
| Xmin | 0 | 0 |
| Xmax | 0 | 0 |
| Ymin | 0 | 0 |
| Ymax | 0 | 0 |

**Problems with data loading:**

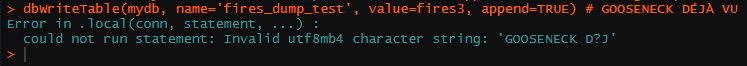
**Buffer pool size:**

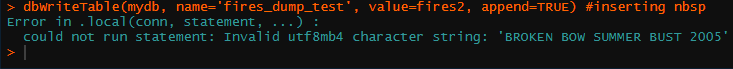
Increasing the buffer\_pool\_size in my.ini file from 8MB to 1024MB solved the problem of

***Error Code: 1206. The total number of locks exceeds the lock table size***

**Bad Data:**

There were problems with inserting data into MySQL as the original dataset contained characters that did not comply with the utf8mb4 charset of MySQL





The dataset was then exported to a text file and the following python script was run to identify ASCII characters beyond 126

***with open("C:/Users/Varun/Desktop/MSBA/DB/Assignment 5/my\_data\_old.txt") as f:***

***while True:***

***c = ord(f.read(1))***

***if not c:***

***print("End of file")***

***break***

***elif(c>126):***

***print(c)***

The following characters were identified.

|  |
| --- |
| **ASCII** |
| 160 |
| 176 |
| 191 |
| 192 |
| 198 |
| 201 |
| 209 |
| 241 |

Two options were considered:

1. Changing the charset from utf8mb4 to latin1 as that would allow characters like À and É
2. Replacing all occurrences of the problematic characters with other characters

Option 1 did not work.

Option 2 was followed.

The following R code was used to replace all occurrences of problematic ASCII values with its replacement

***fires1$COMPLEX\_NAME <- gsub("\u00C0", "A", fires1$COMPLEX\_NAME, fixed = TRUE)***

The transformation was applied to all the columns for each and every problematic character.

**For fires table:**

|  |  |  |  |
| --- | --- | --- | --- |
| **ASCII Value** | **Character** | **Encoding** | **Replacement** |
| 160 |  | \u00A0 |  |
| 176 | ° | \u00B0 | \* |
| 191 | ¿ | \u00BF | ' |
| 192 | À | \u00C0 | A |
| 198 | Æ | \u00C6 | AE |
| 201 | É | \u00C9 | E |
| 209 | Ñ | \u00D1 | N |
| 241 | ñ | \u00F1 | N |

ASCII value 160, which is a non breaking space has been replaced with a regular space.

|  |  |  |  |
| --- | --- | --- | --- |
| **ASCII Value** | **OBJECTID** | **Before** | **After** |
| 160 | 2752 | BROKEN BOW SUMMER BUST 2005 COMPLEX | BROKEN BOW SUMMER BUST 2005 COMPLEX |
| 176 | 1425157 | 13° FIRE | 13\* FIRE |
| 191 | 1676592 | IT WASN¿T THE NEIGHBOR | IT WASN'T THE NEIGHBOR |
| 192 | 1815095 | GOOSENECK DÉJÀ VU | GOOSENECK DÉJÀ VU |
| 198 | 242709 | DREWRYÆS BLUFF OBS | DREWRYAES BLUFF OBS |
| 201 | 1815095 | GOOSENECK DÉJÀ VU | GOOSENECK DÉJÀ VU |
| 209 | 1852022 | GARCEÑO LOOP | GARCENO LOOP |
| 241 | 1529483 | Salineño Vol. Fire Dept | Salineno Vol. Fire Dept |

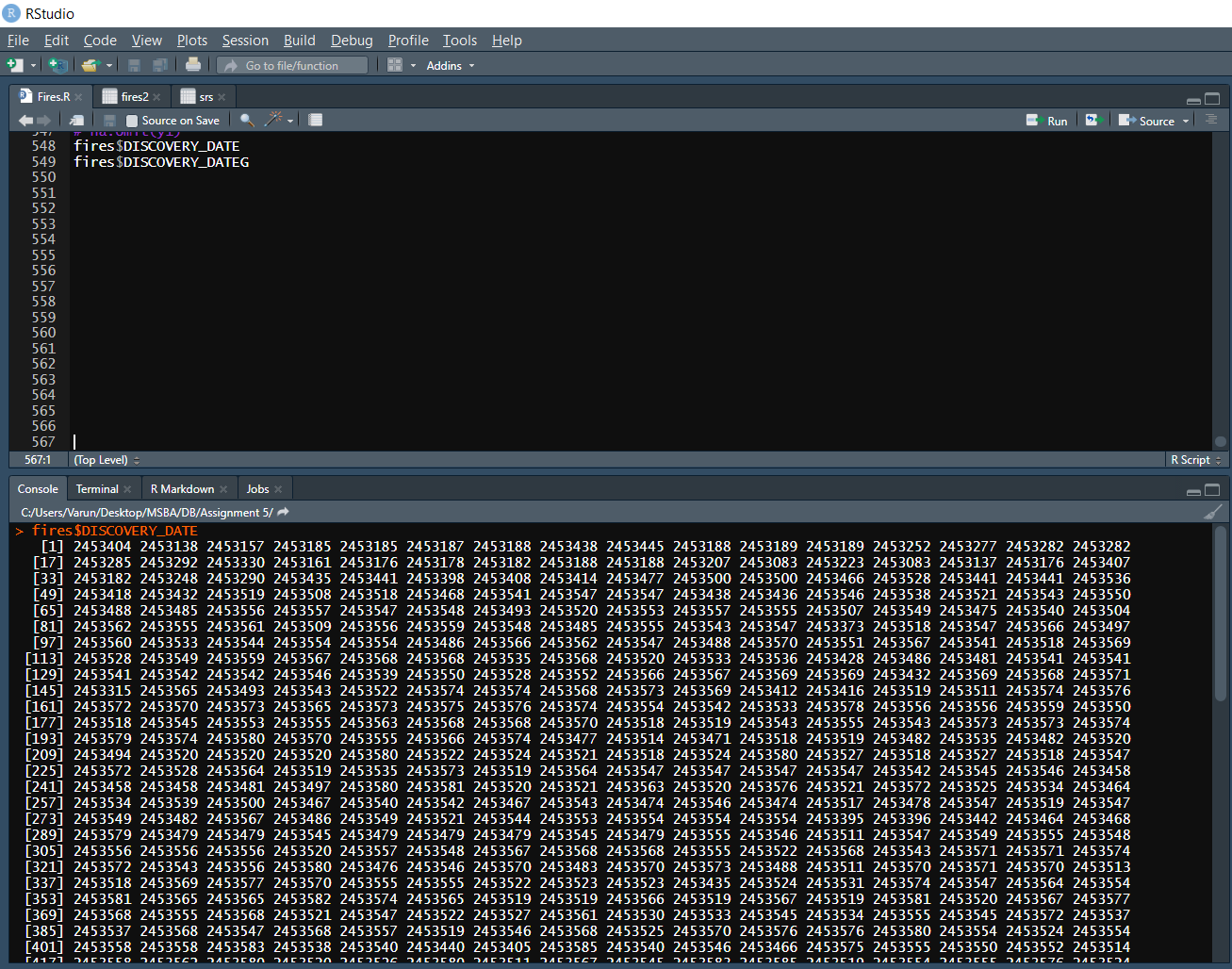
**Conversion of date format:**

The dates given in the columns DISCOVERY\_DATE and CONT\_DATE were in the Julian date format. They have been converted to Gregorian date by using the following R code.

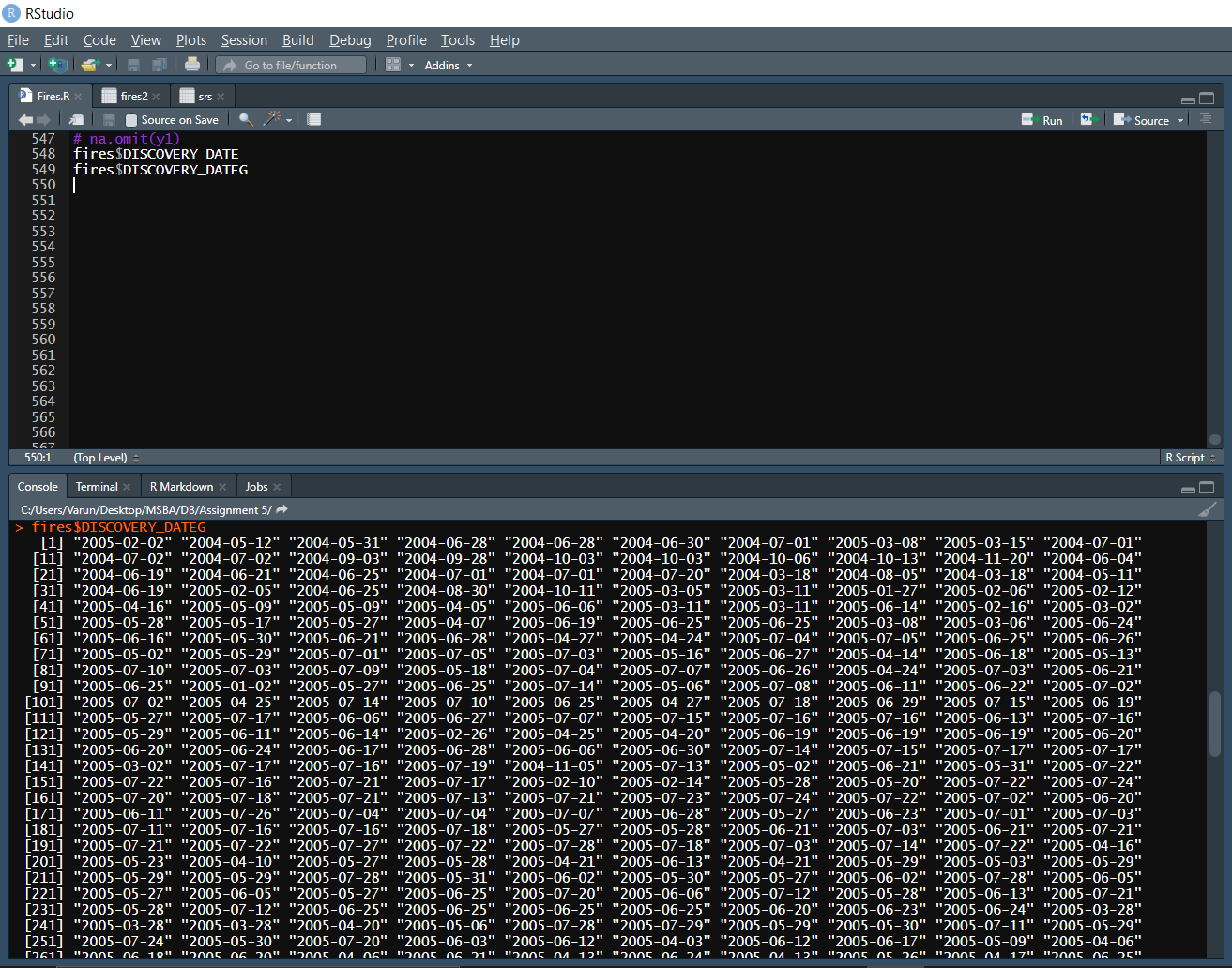
**fires$CONT\_DATEG <- as.Date(fires$CONT\_DATE - 2458014.5, origin = '2015-08-13')**

**fires$DISCOVERY\_DATEG <- as.Date(fires$DISCOVERY\_DATE - 2458014.5, origin = '2015-08-13')**

**Before conversion:**



**After conversion:**



**Shapes column:**

We were unable to obtain information from the shapes column. This column contains BLOB values.

**Inconsistency between values present in** **nwcgunitid column in fires table and** **UNITID in nwcgunitidactive table:**

The following values were present in the nwcgunitid column in fires table and not in UNITID in nwcgunitidactive table

|  |
| --- |
| USVTSFM |
| USPRIITF |
| USOHSFM |
| USNDSFM |
| USMOSFM |
| USMASFM |
| USKSSFM |
| USKSKSX |
| USKSJA409 |
| USINSFM |
| USILSFM |
| USIASFM |
| USHINPS |
| USHIHIS |
| USHICNTY |
| USDESFM |
| USCODDQ |

Loading rows containing these values will cause the following problem

***Error Code: 1452. Cannot add or update a child row: a foreign key constraint fails***

This has affected 34757 rows.

The workaround to load the data into the database is to set the FOREIGN\_KEY\_CHECKS variable to 0 and setting it back to 1 after inserting.

**Loading data:**

The following code was run in R to load the entire dataset into MySQL

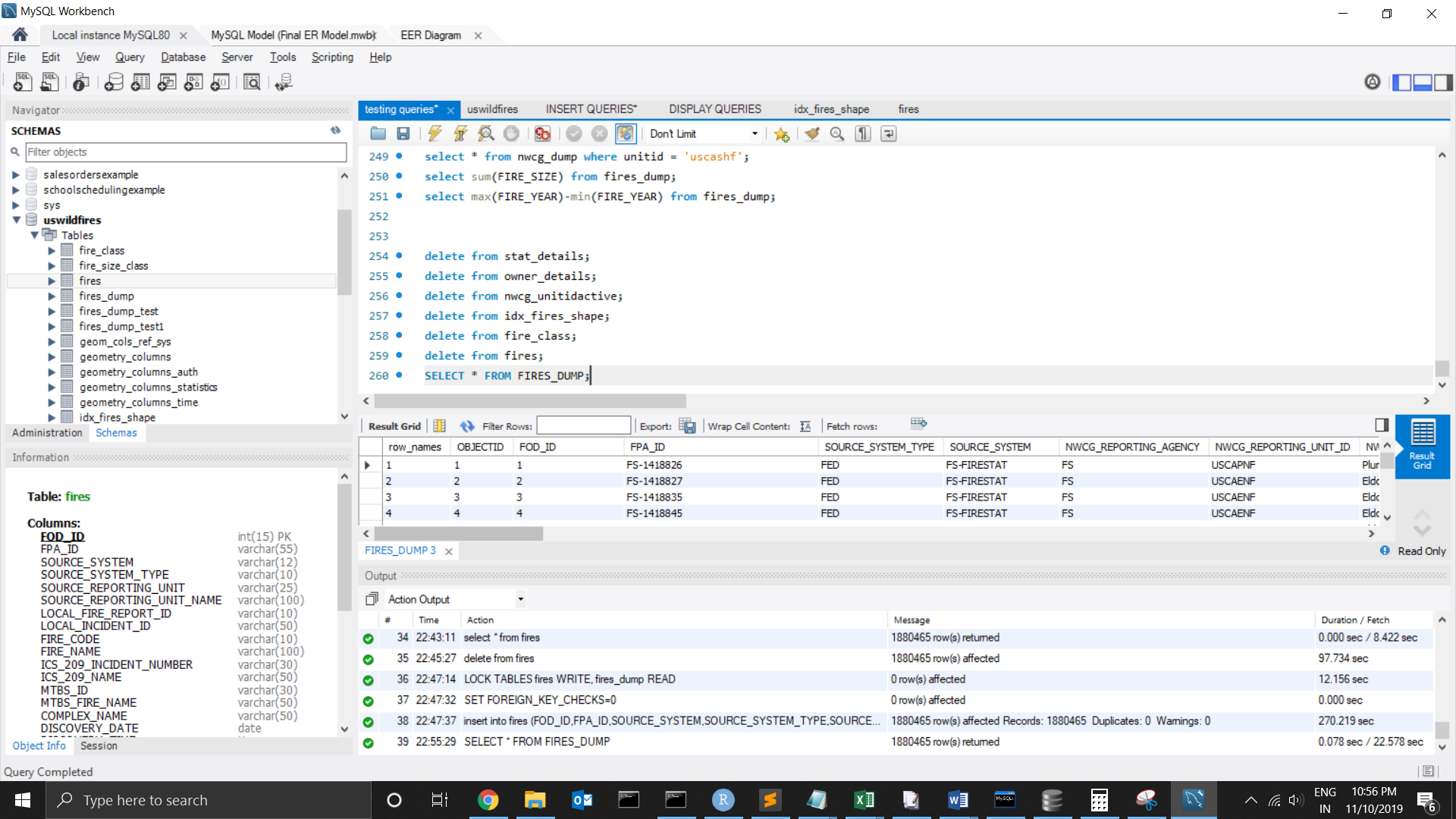
**mydb = dbConnect(MySQL(), user='root', password='root', dbname='uswildfires', host='localhost')**

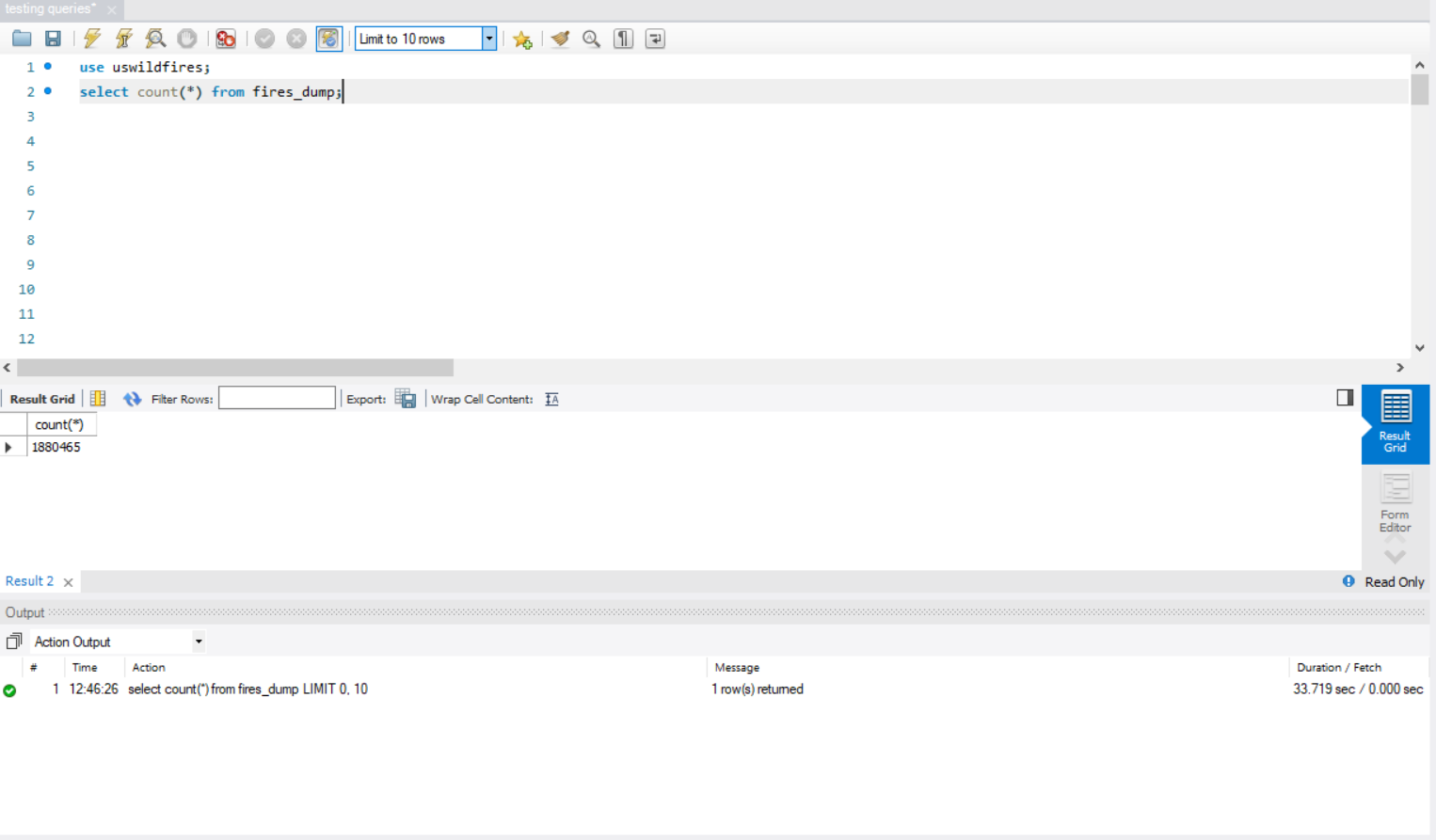
**dbWriteTable(mydb, name='fires\_dump', value=fires1, overwrite=TRUE)**



The dbWriteTable returns **TRUE**, meaning all the rows were inserted successfully

**Verifying in MySQL**





**Column Lengths:**

The following R code was used to determine the column lengths for the tables fires and nwcg

***fires\_col\_len <- lapply(fires, function(x) max(nchar(x,type = "chars", allowNA = FALSE, keepNA = FALSE)))***

***nwcg\_col\_len <- lapply(nwcg, function(x) max(nchar(x,type = "chars", allowNA = FALSE, keepNA = FALSE)))***

***ifs\_col\_len <- lapply(ifs, function(x) max(nchar(x,type = "chars", allowNA = FALSE, keepNA = FALSE)))***

|  |  |
| --- | --- |
| **FIRES** | |
| **Column** | **Length** |
| OBJECTID | 7 |
| FOD\_ID | 9 |
| FPA\_ID | 49 |
| SOURCE\_SYSTEM\_TYPE | 9 |
| SOURCE\_SYSTEM | 11 |
| NWCG\_REPORTING\_AGENCY | 6 |
| NWCG\_REPORTING\_UNIT\_ID | 9 |
| NWCG\_REPORTING\_UNIT\_NAME | 79 |
| SOURCE\_REPORTING\_UNIT | 21 |
| SOURCE\_REPORTING\_UNIT\_NAME | 74 |
| LOCAL\_FIRE\_REPORT\_ID | 6 |
| LOCAL\_INCIDENT\_ID | 28 |
| FIRE\_CODE | 6 |
| FIRE\_NAME | 70 |
| ICS\_209\_INCIDENT\_NUMBER | 19 |
| ICS\_209\_NAME | 37 |
| MTBS\_ID | 29 |
| MTBS\_FIRE\_NAME | 49 |
| COMPLEX\_NAME | 43 |
| FIRE\_YEAR | 4 |
| DISCOVERY\_DATE | 9 |
| DISCOVERY\_DOY | 3 |
| DISCOVERY\_TIME | 4 |
| STAT\_CAUSE\_CODE | 2 |
| STAT\_CAUSE\_DESCR | 17 |
| CONT\_DATE | 9 |
| CONT\_DOY | 3 |
| CONT\_TIME | 4 |
| FIRE\_SIZE | 9 |
| FIRE\_SIZE\_CLASS | 1 |
| LATITUDE | 11 |
| LONGITUDE | 13 |
| OWNER\_CODE | 2 |
| OWNER\_DESCR | 21 |
| STATE | 2 |
| COUNTY | 50 |
| FIPS\_CODE | 3 |
| FIPS\_NAME | 31 |
| Shape | 369 |

|  |  |
| --- | --- |
| **NWCG** | |
| **Column** | **Length** |
| OBJECTID | 4 |
| UnitId | 8 |
| GeographicArea | 2 |
| Gacc | 8 |
| WildlandRole | 28 |
| UnitType | 49 |
| Department | 5 |
| Agency | 5 |
| Parent | NA |
| Country | 2 |
| State | 2 |
| Code | 4 |
| Name | 90 |

|  |  |
| --- | --- |
| **idx\_fires\_shape** | |
| **Column** | **Length** |
| pkid | 7 |
| xmin | 17 |
| xmax | 17 |
| ymin | 16 |
| ymax | 16 |

**Data Loading:**

**TRANSLATION:** insert into stat\_details table into the following columns Stat\_Cause\_Code, Stat\_Cause\_Descr the distinct values from the columns STAT\_CAUSE\_CODE, STAT\_CAUSE\_DESCR table

**CLEAN UP:**

**QUERY:**

-- inserting stat\_details table

SET @@AUTOCOMMIT=0;

LOCK TABLES stat\_details WRITE, fires\_dump READ;

insert into stat\_details

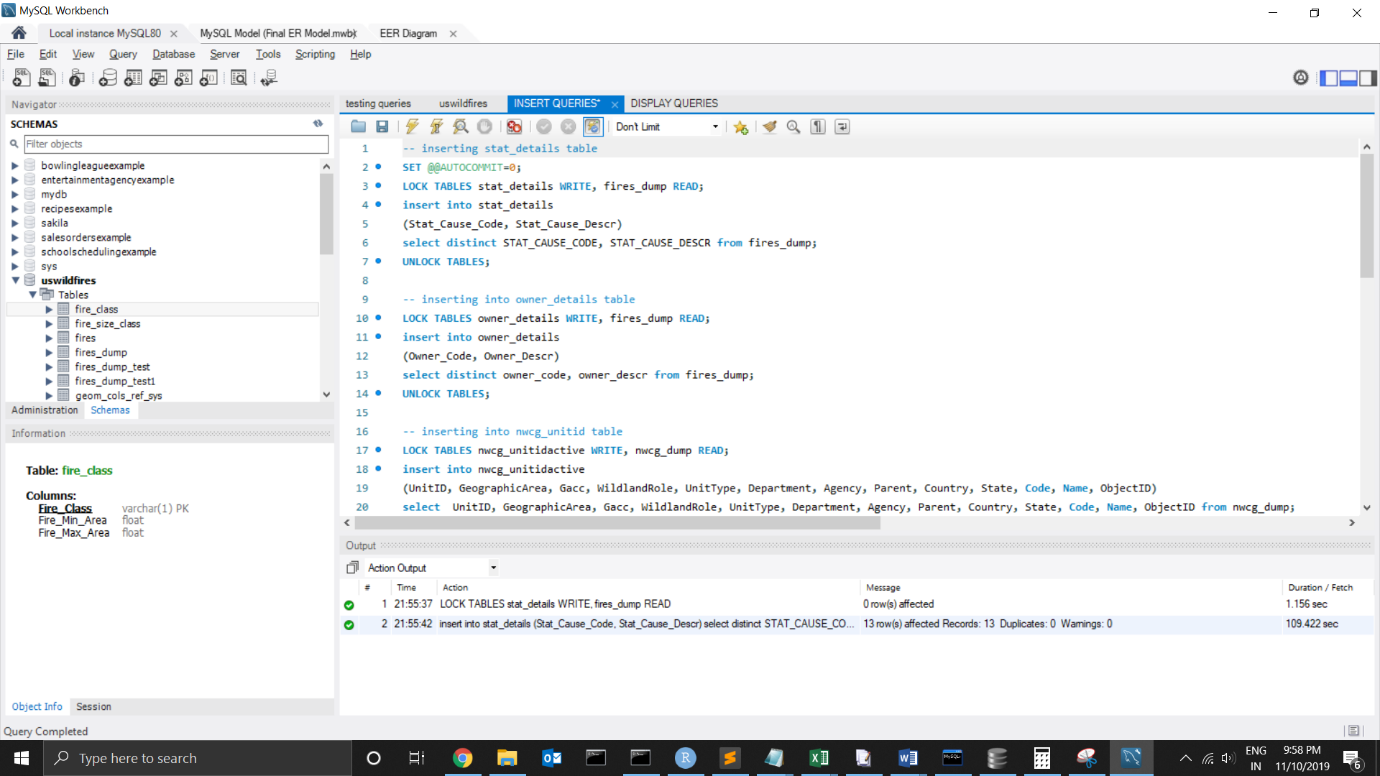
(Stat\_Cause\_Code, Stat\_Cause\_Descr)

select distinct STAT\_CAUSE\_CODE, STAT\_CAUSE\_DESCR from fires\_dump;

-- select \* from fires\_dump;

UNLOCK TABLES;

**SCREENSHOT:**



**TRANSLATION:** insert into owner\_details table into the following columns Owner\_Code, Owner\_Descr the distinct values of owner\_code, owner\_descr columns from fires\_dump

**CLEAN UP:**

**QUERY:**

LOCK TABLES owner\_details WRITE, fires\_dump READ;

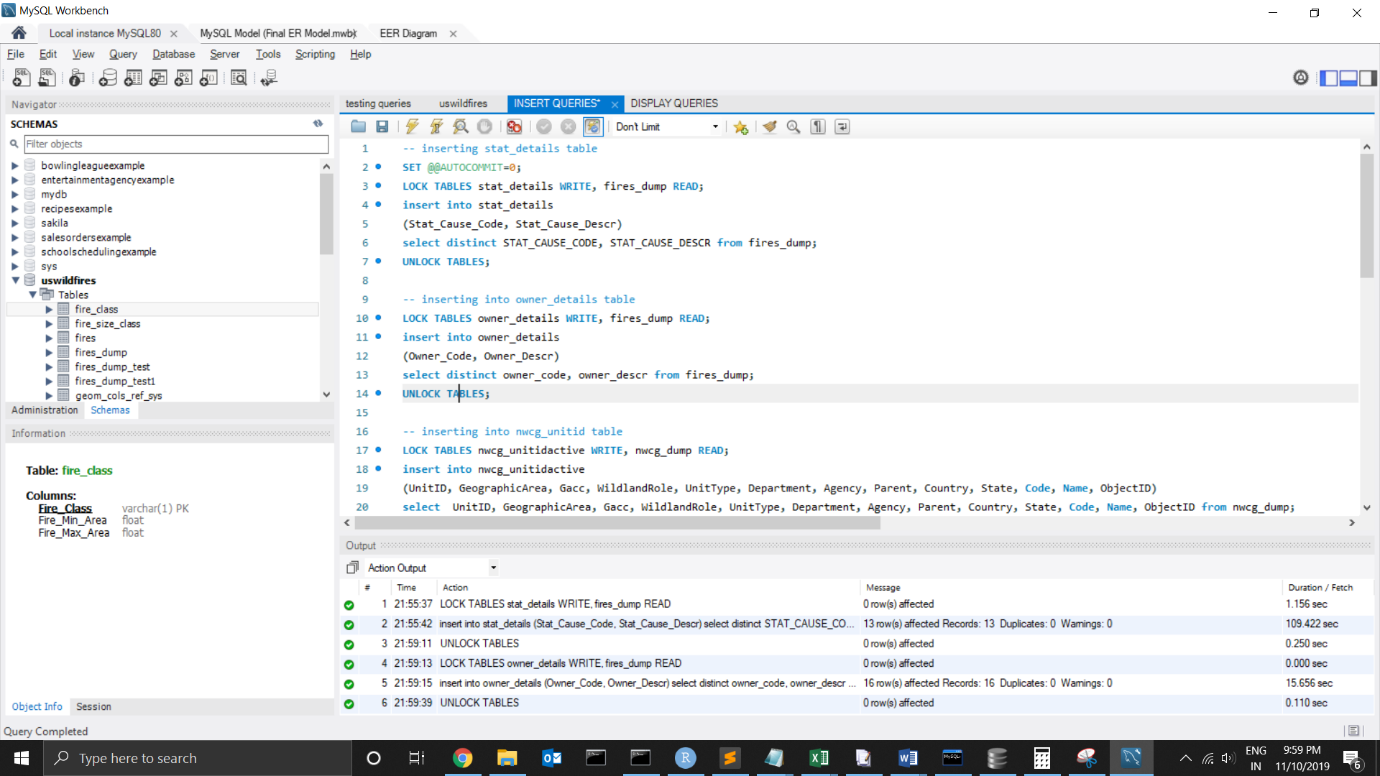
insert into owner\_details

(Owner\_Code, Owner\_Descr)

select distinct owner\_code, owner\_descr from fires\_dump;

UNLOCK TABLES;

**SCREENSHOT:**



**TRANSLATION:** insert into nwcg\_unitidactive table into the following columns UnitID, GeographicArea, Gacc, WildlandRole, UnitType, Department, Agency, Parent, Country, State, Code, Name, ObjectID the values from the columns of UnitID, GeographicArea, Gacc, WildlandRole, UnitType, Department, Agency, Parent, Country, State, Code, Name, ObjectID from nwcg\_dump table

**CLEAN UP:**

**QUERY:**

LOCK TABLES nwcg\_unitidactive WRITE, nwcg\_dump READ;

insert into nwcg\_unitidactive

(UnitID, GeographicArea, Gacc, WildlandRole, UnitType, Department, Agency, Parent, Country, State, Code, Name, ObjectID)

select UnitID, GeographicArea, Gacc, WildlandRole, UnitType, Department, Agency, Parent, Country, State, Code, Name, ObjectID from nwcg\_dump;

UNLOCK TABLES;

**SCREENSHOT:**



**TRANSLATION**: insert into idx\_fires\_shape table into the following columns pkid, xmin, xmax, ymin, ymax the values of pkid, xmin, xmax, ymin, ymax from idx\_fires\_shape\_dump

**CLEAN UP:**

**QUERY:**

LOCK TABLES idx\_fires\_shape WRITE, idx\_fires\_shape\_dump READ;

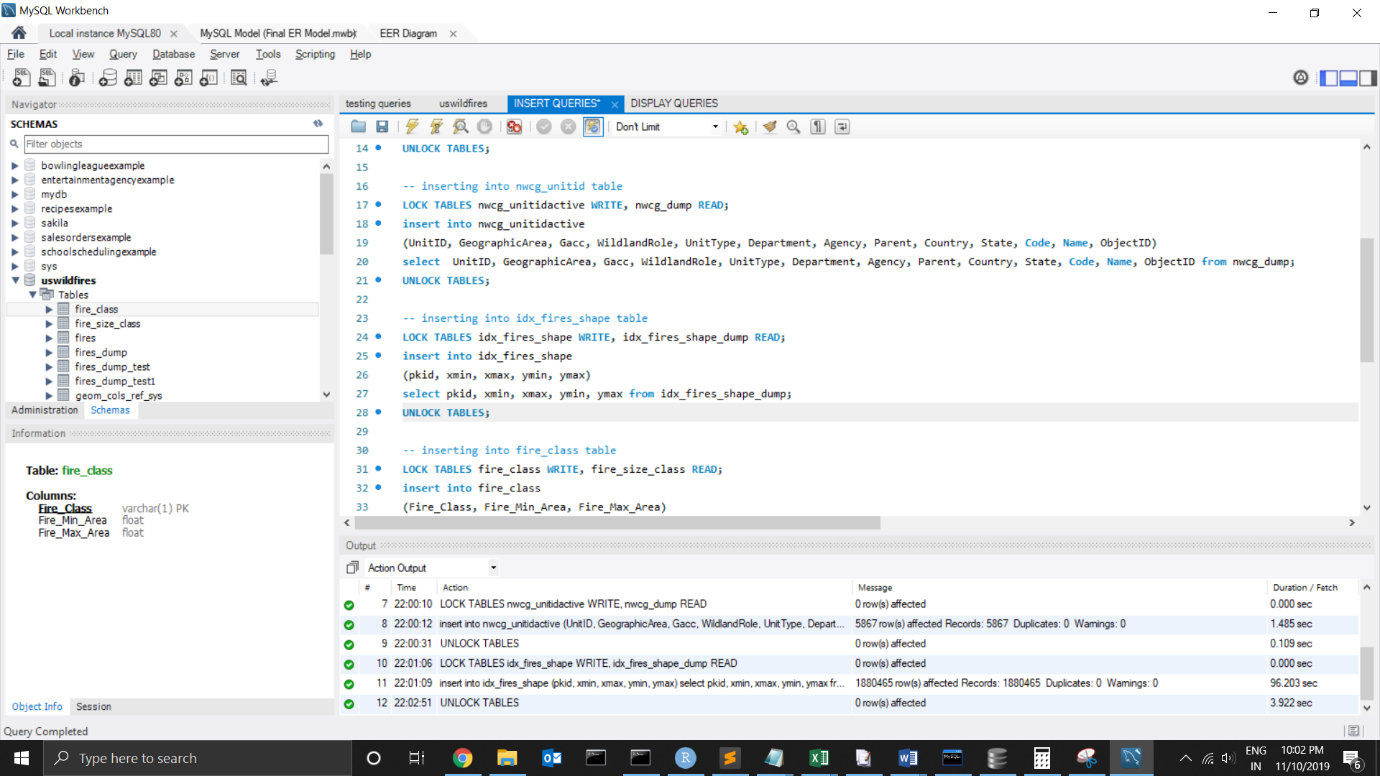
insert into idx\_fires\_shape

(pkid, xmin, xmax, ymin, ymax)

select pkid, xmin, xmax, ymin, ymax from idx\_fires\_shape\_dump;

UNLOCK TABLES;

**SCREENSHOT:**



**TRANSLATION:** insert into fire\_class table into the following columns Fire\_Class, Fire\_Min\_Area, Fire\_Max\_Area the values of FIRE\_SIZE\_CLASS, MIN, MAX from fire\_size\_class

**CLEAN UP:**

**QUERY:**

LOCK TABLES fire\_class WRITE, fire\_size\_class READ;

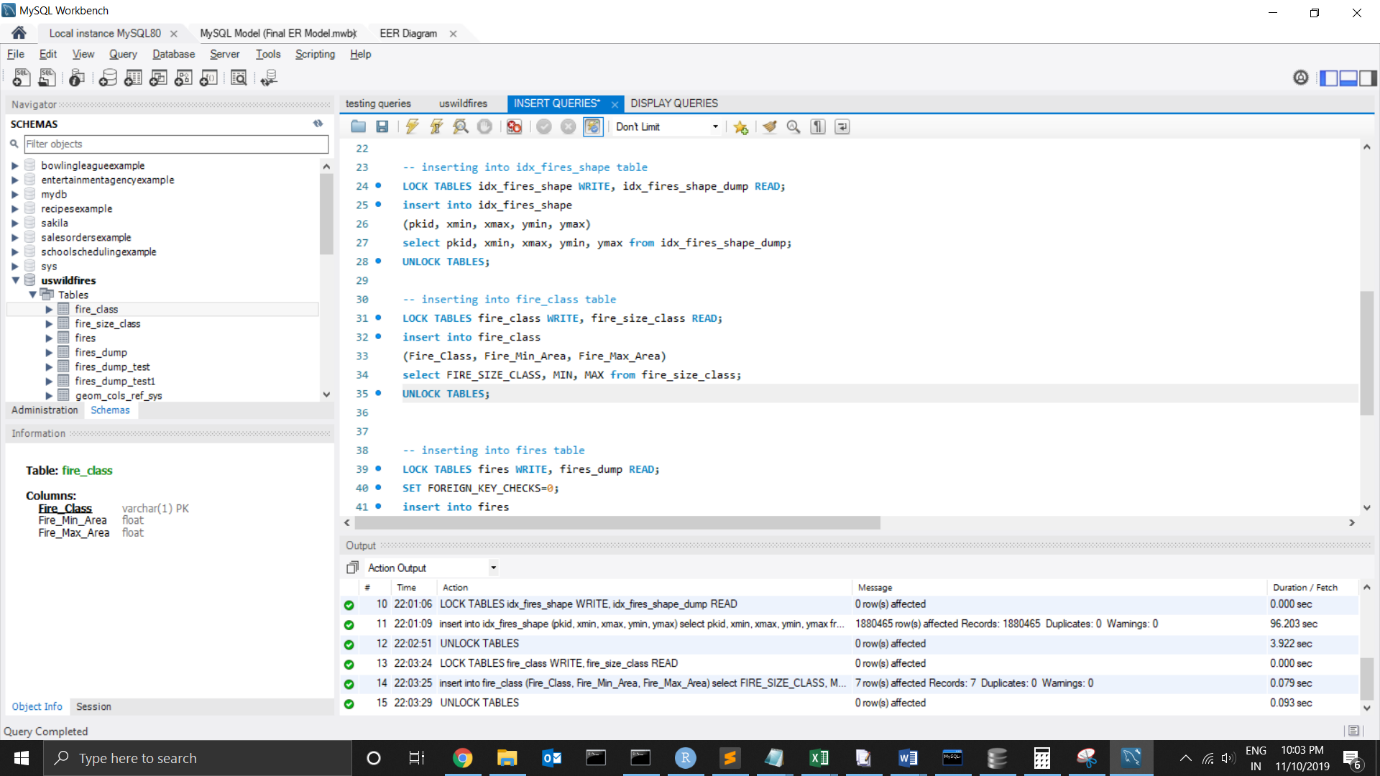
insert into fire\_class

(Fire\_Class, Fire\_Min\_Area, Fire\_Max\_Area)

select FIRE\_SIZE\_CLASS, MIN, MAX from fire\_size\_class;

UNLOCK TABLES;

**SCREENSHOT:**



**TRANSLATION:** insert into fires table into the following columns OBJECTID,FOD\_ID,FPA\_ID,SOURCE\_SYSTEM,SOURCE\_SYSTEM\_TYPE,SOURCE\_REPORTING\_UNIT,SOURCE\_REPORTING\_UNIT\_NAME, LOCAL\_FIRE\_REPORT\_ID,

LOCAL\_INCIDENT\_ID,FIRE\_CODE, FIRE\_NAME,ICS\_209\_INCIDENT\_NUMBER,ICS\_209\_NAME,MTBS\_ID,MTBS\_FIRE\_NAME,COMPLEX\_NAME,

DISCOVERY\_DATE,DISCOVERY\_TIME,CONT\_DATE,CONT\_TIME,FIRE\_SIZE,LATITUDE,LONGITUDE,STATE,

COUNTY,FIPS\_CODE,FIPS\_NAME,Stat\_Details\_Stat\_Cause\_Code ,Owner\_Details\_Owner\_Code,NWCG\_UnitIDActive\_UnitID

,idx\_fires\_shape\_pkid the values of OBJECTID,FOD\_ID,FPA\_ID,SOURCE\_SYSTEM,SOURCE\_SYSTEM\_TYPE,SOURCE\_REPORTING\_UNIT,SOURCE\_REPORTING\_UNIT\_NAME, LOCAL\_FIRE\_REPORT\_ID,

LOCAL\_INCIDENT\_ID,FIRE\_CODE, FIRE\_NAME,ICS\_209\_INCIDENT\_NUMBER,ICS\_209\_NAME,MTBS\_ID,MTBS\_FIRE\_NAME,COMPLEX\_NAME,

DISCOVERY\_DATE,DISCOVERY\_TIME,CONT\_DATE,CONT\_TIME,FIRE\_SIZE,LATITUDE,LONGITUDE,STATE,

COUNTY,FIPS\_CODE,FIPS\_NAME,Stat\_Cause\_Code ,Owner\_Code,NWCG\_REPORTING\_UNIT\_ID,

OBJECTID FROM FIRES\_DUMP

**CLEAN UP:**

**QUERY:**

LOCK TABLES fires WRITE, fires\_dump READ;

SET FOREIGN\_KEY\_CHECKS=0;

insert into fires

(FOD\_ID,FPA\_ID,SOURCE\_SYSTEM,SOURCE\_SYSTEM\_TYPE,SOURCE\_REPORTING\_UNIT,SOURCE\_REPORTING\_UNIT\_NAME, LOCAL\_FIRE\_REPORT\_ID,

LOCAL\_INCIDENT\_ID,FIRE\_CODE, FIRE\_NAME,ICS\_209\_INCIDENT\_NUMBER,ICS\_209\_NAME,MTBS\_ID,MTBS\_FIRE\_NAME,COMPLEX\_NAME,

DISCOVERY\_DATE,DISCOVERY\_TIME,CONT\_DATE,CONT\_TIME,FIRE\_SIZE,LATITUDE,LONGITUDE,STATE,

COUNTY,FIPS\_CODE,FIPS\_NAME,Stat\_Details\_Stat\_Cause\_Code ,Owner\_Details\_Owner\_Code,NWCG\_UnitIDActive\_UnitID

,idx\_fires\_shape\_pkid)

select FOD\_ID,FPA\_ID,SOURCE\_SYSTEM,SOURCE\_SYSTEM\_TYPE,SOURCE\_REPORTING\_UNIT,SOURCE\_REPORTING\_UNIT\_NAME, LOCAL\_FIRE\_REPORT\_ID,

LOCAL\_INCIDENT\_ID,FIRE\_CODE, FIRE\_NAME,ICS\_209\_INCIDENT\_NUMBER,ICS\_209\_NAME,MTBS\_ID,MTBS\_FIRE\_NAME,COMPLEX\_NAME,

DISCOVERY\_DATE,DISCOVERY\_TIME,CONT\_DATE,CONT\_TIME,FIRE\_SIZE,LATITUDE,LONGITUDE,STATE,

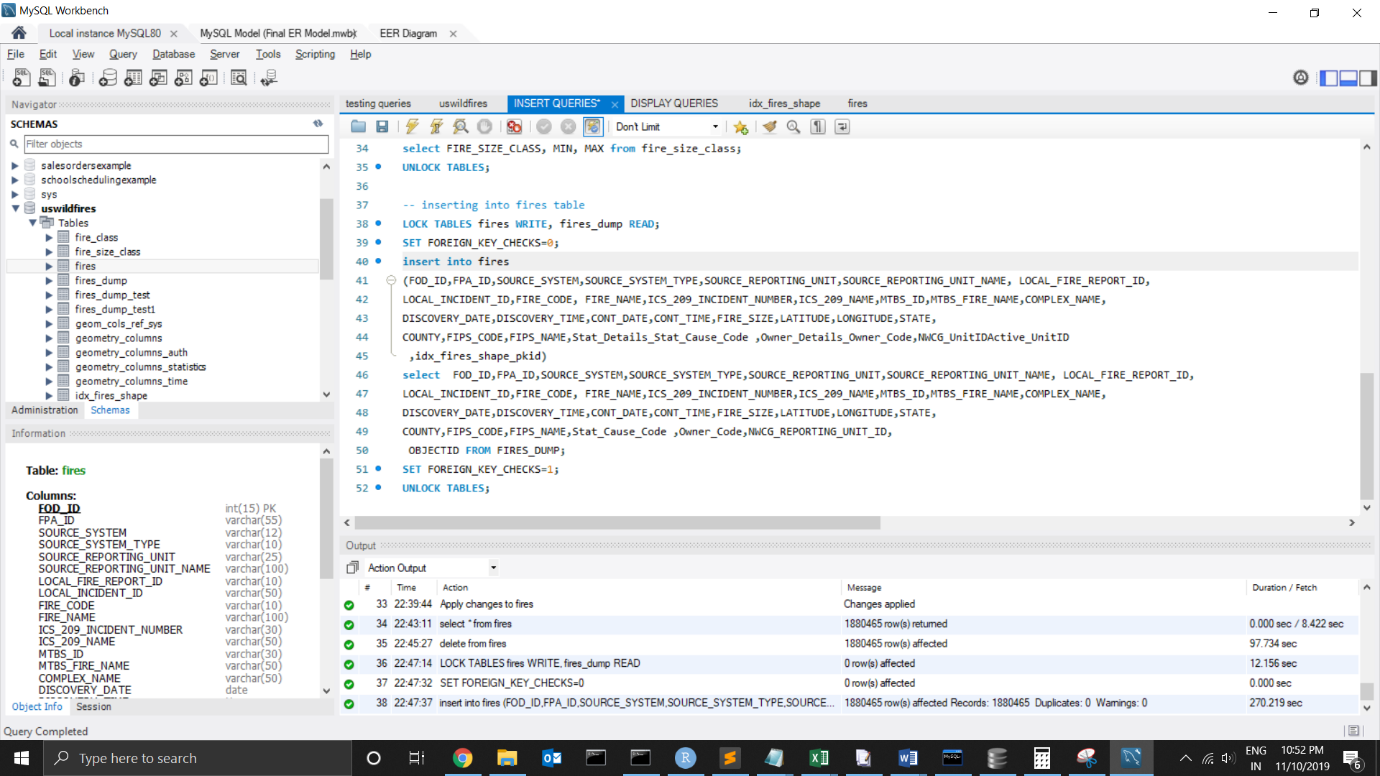
COUNTY,FIPS\_CODE,FIPS\_NAME,Stat\_Cause\_Code ,Owner\_Code,NWCG\_REPORTING\_UNIT\_ID,

OBJECTID FROM FIRES\_DUMP;

SET FOREIGN\_KEY\_CHECKS=1;

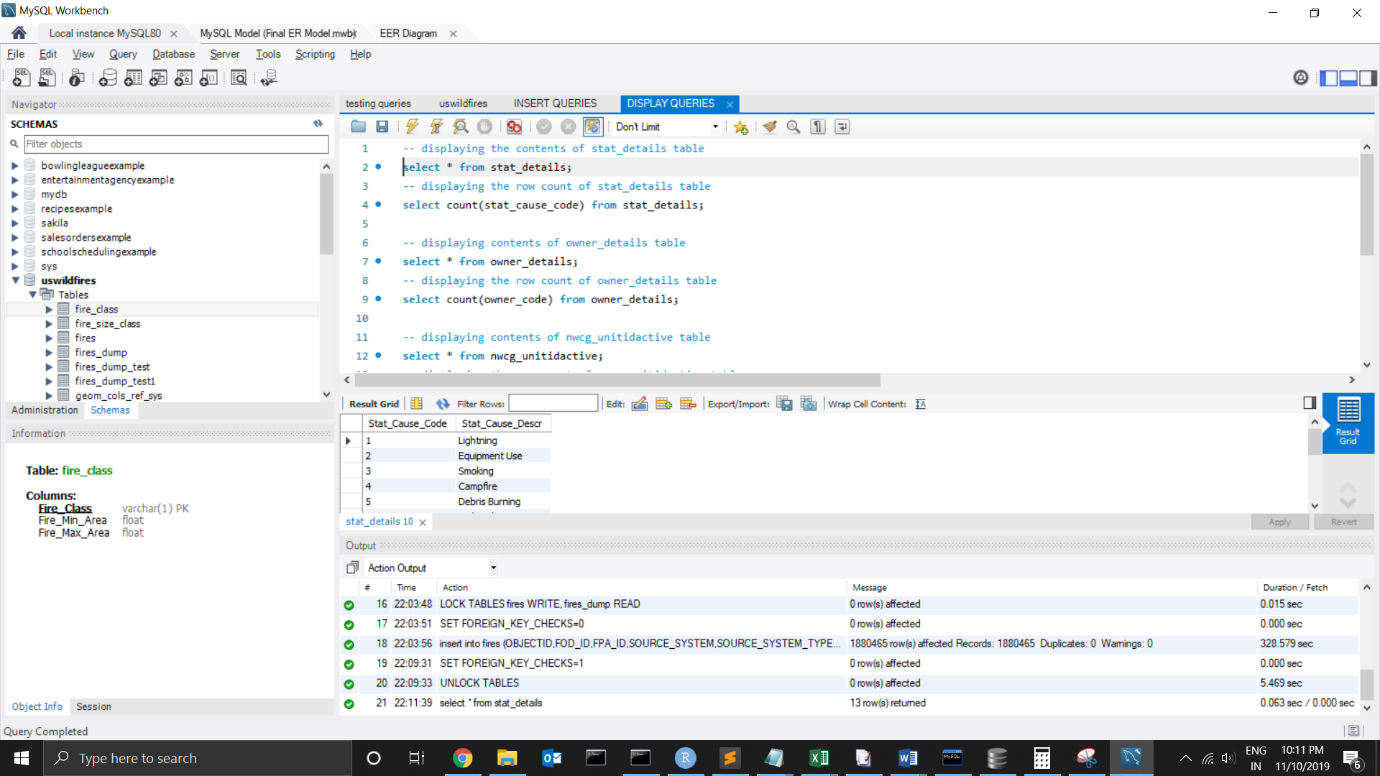
UNLOCK TABLES;

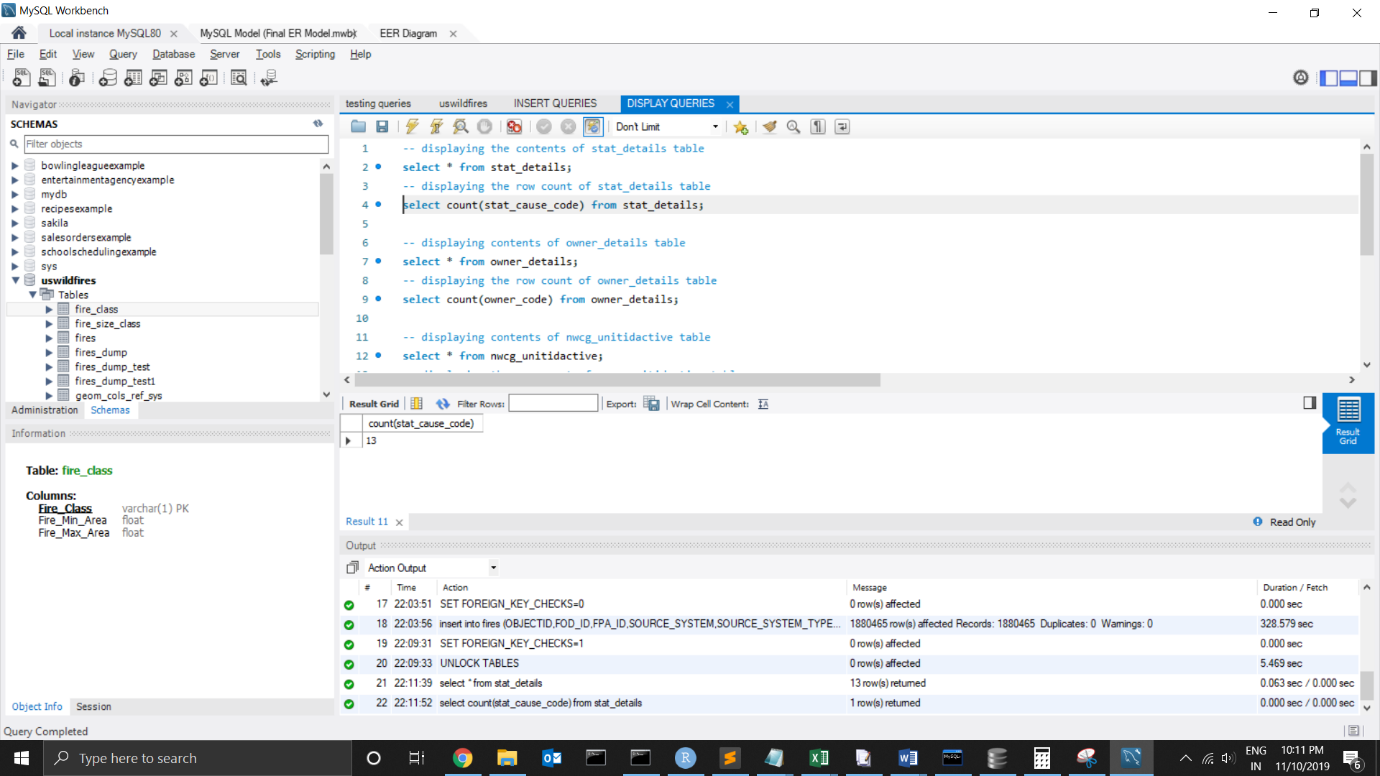
**SCREENSHOT:**



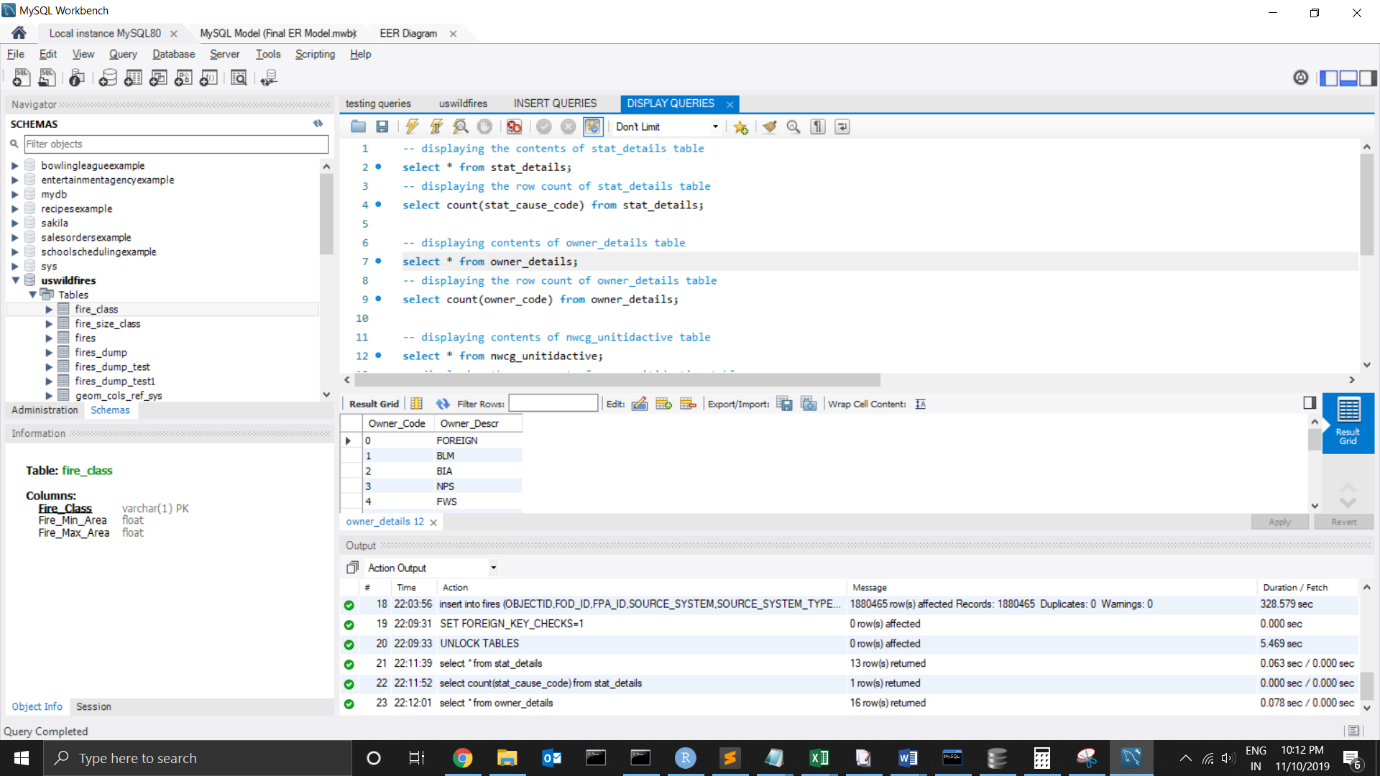
**Loaded Data:**

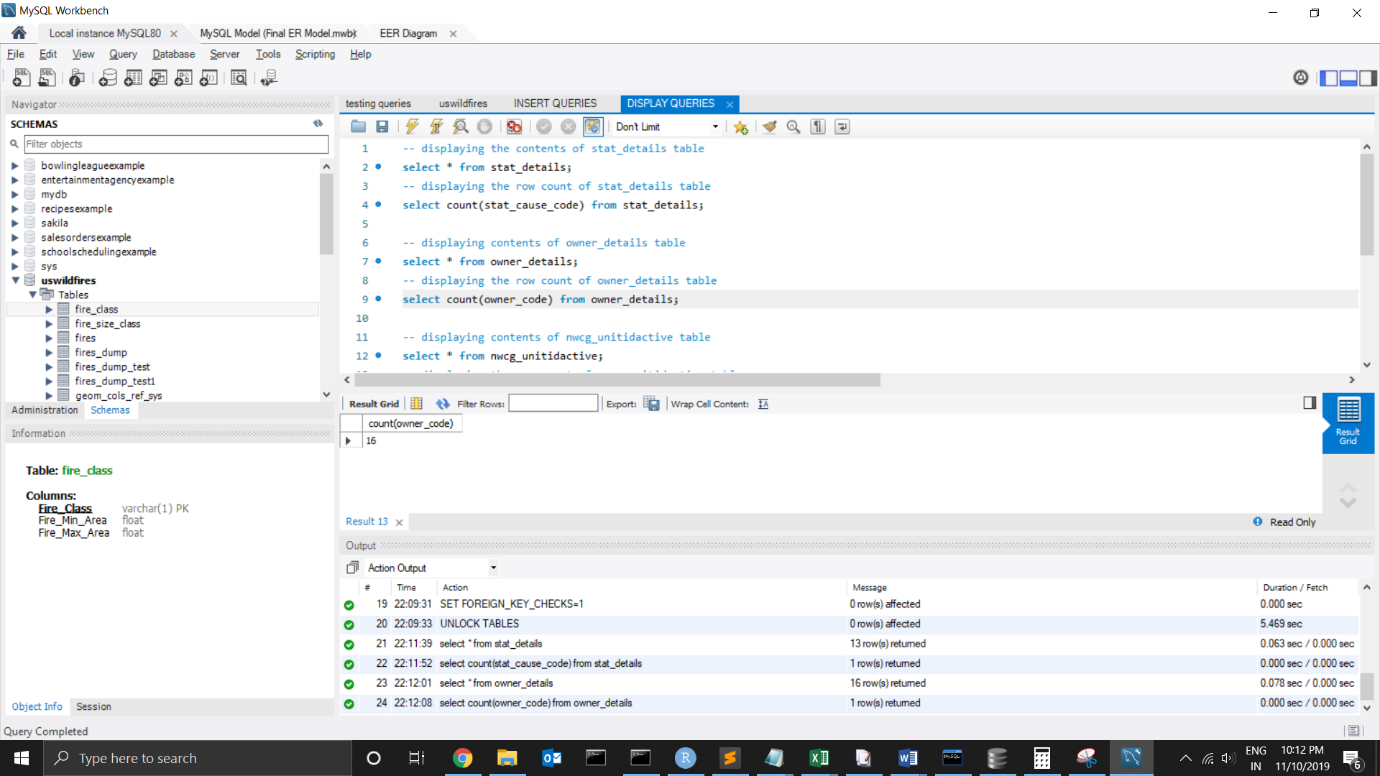
**Stat\_details:**



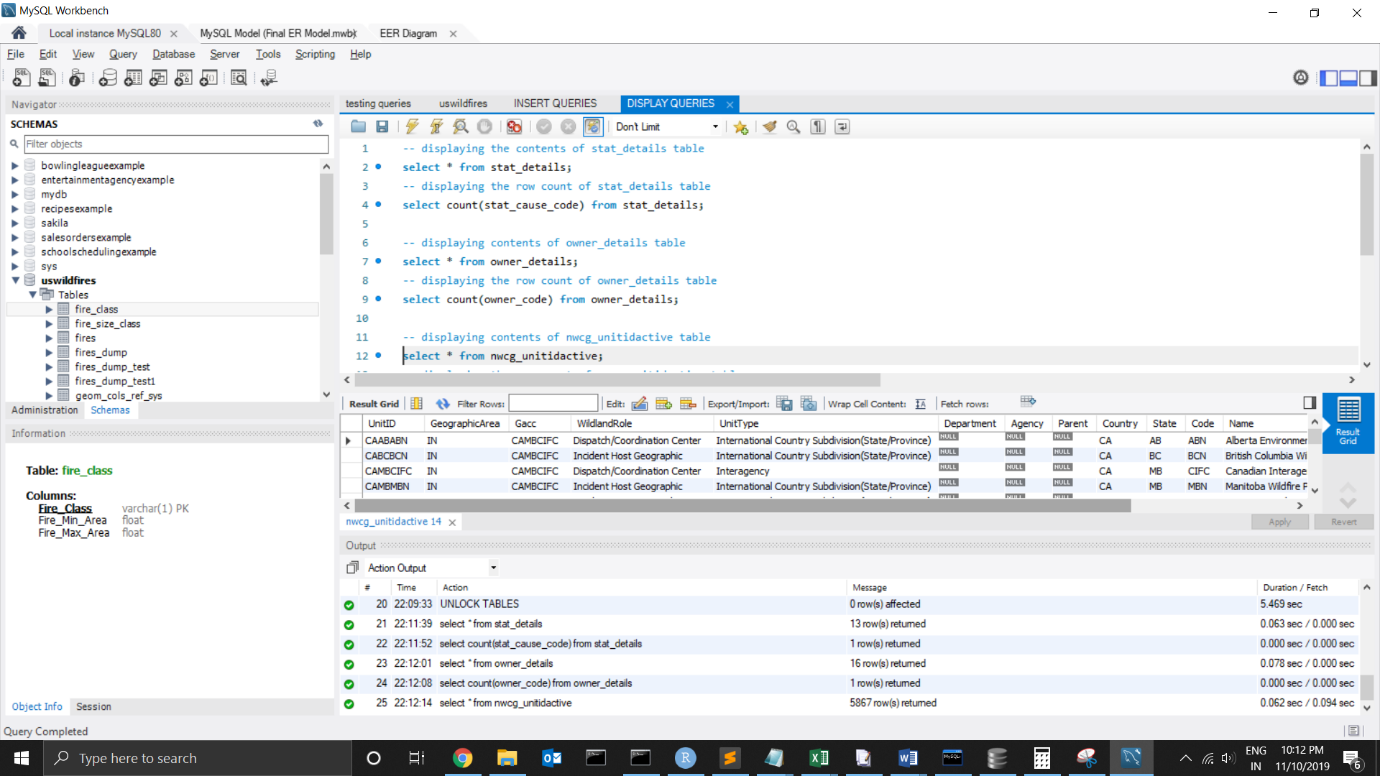


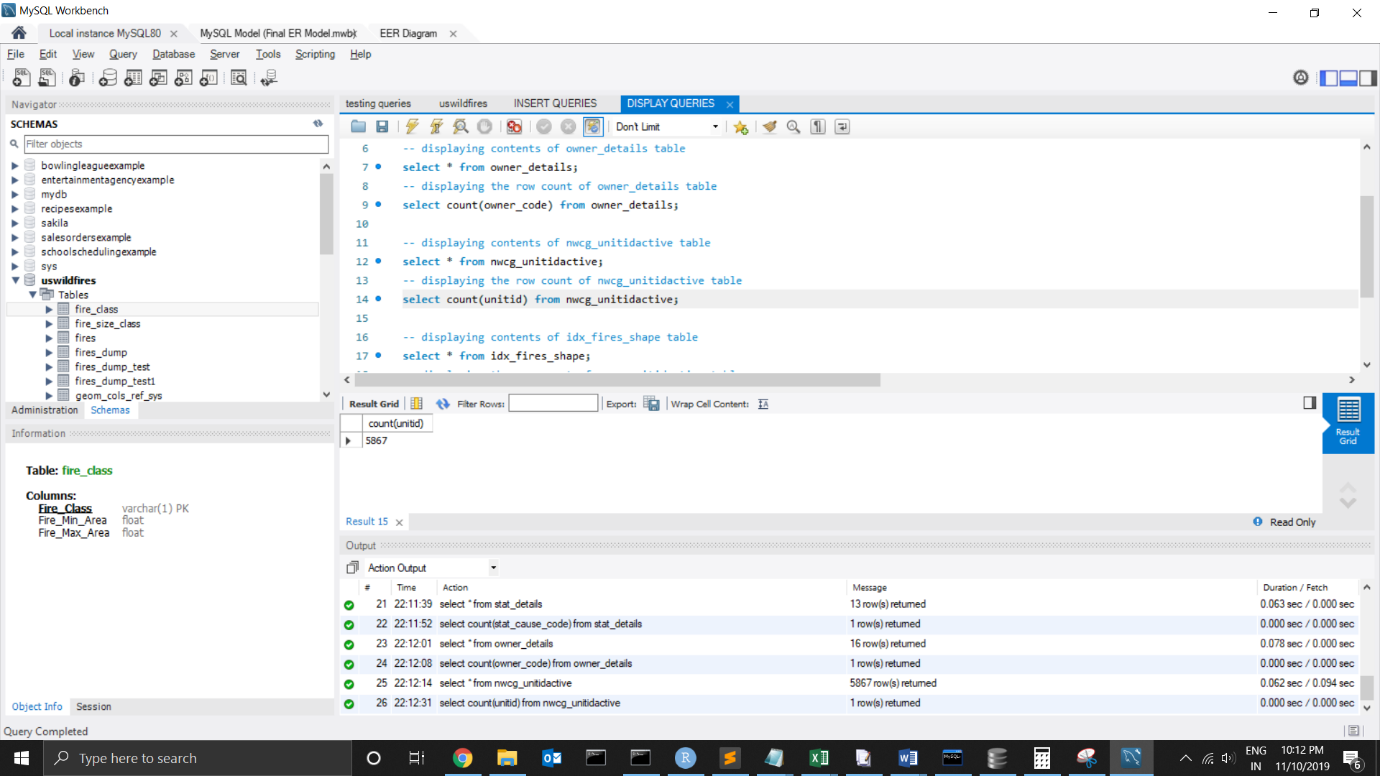
**Owner\_details:**



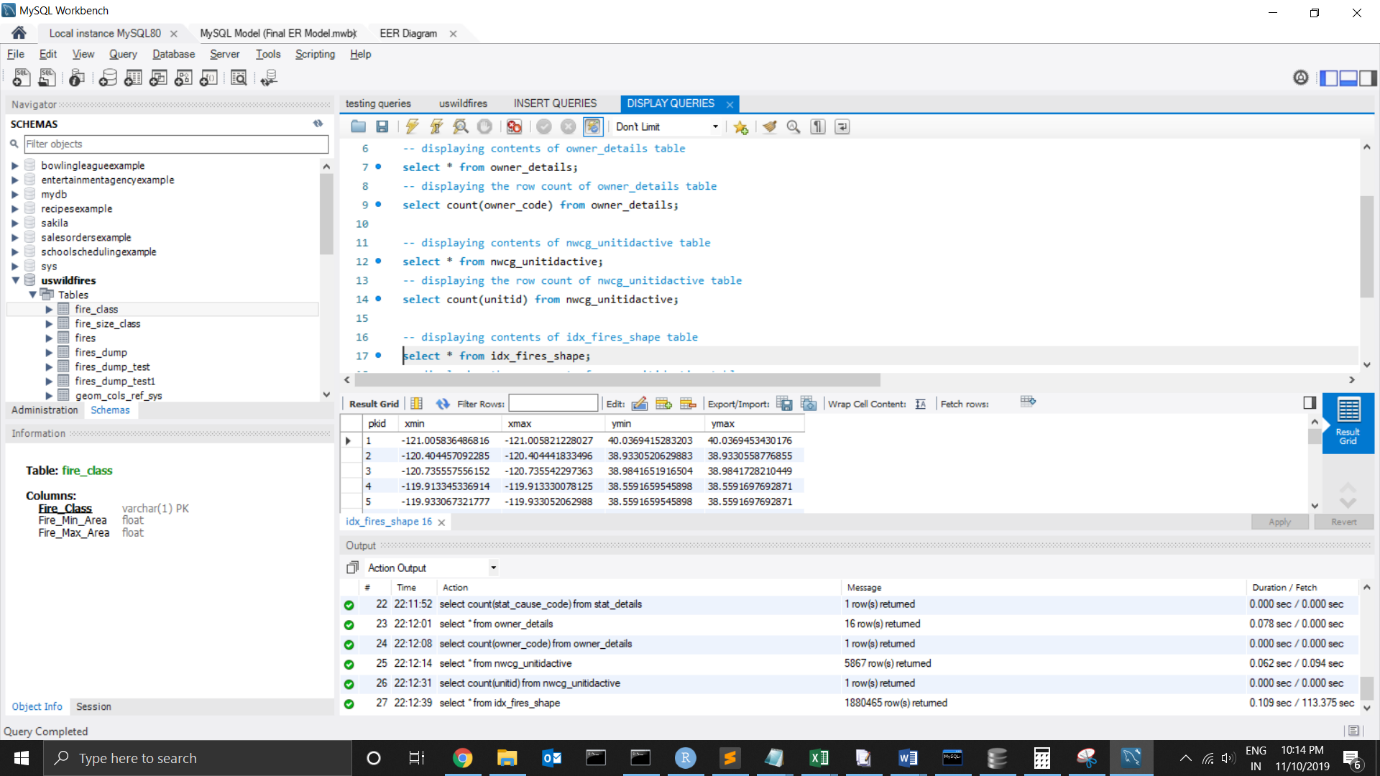


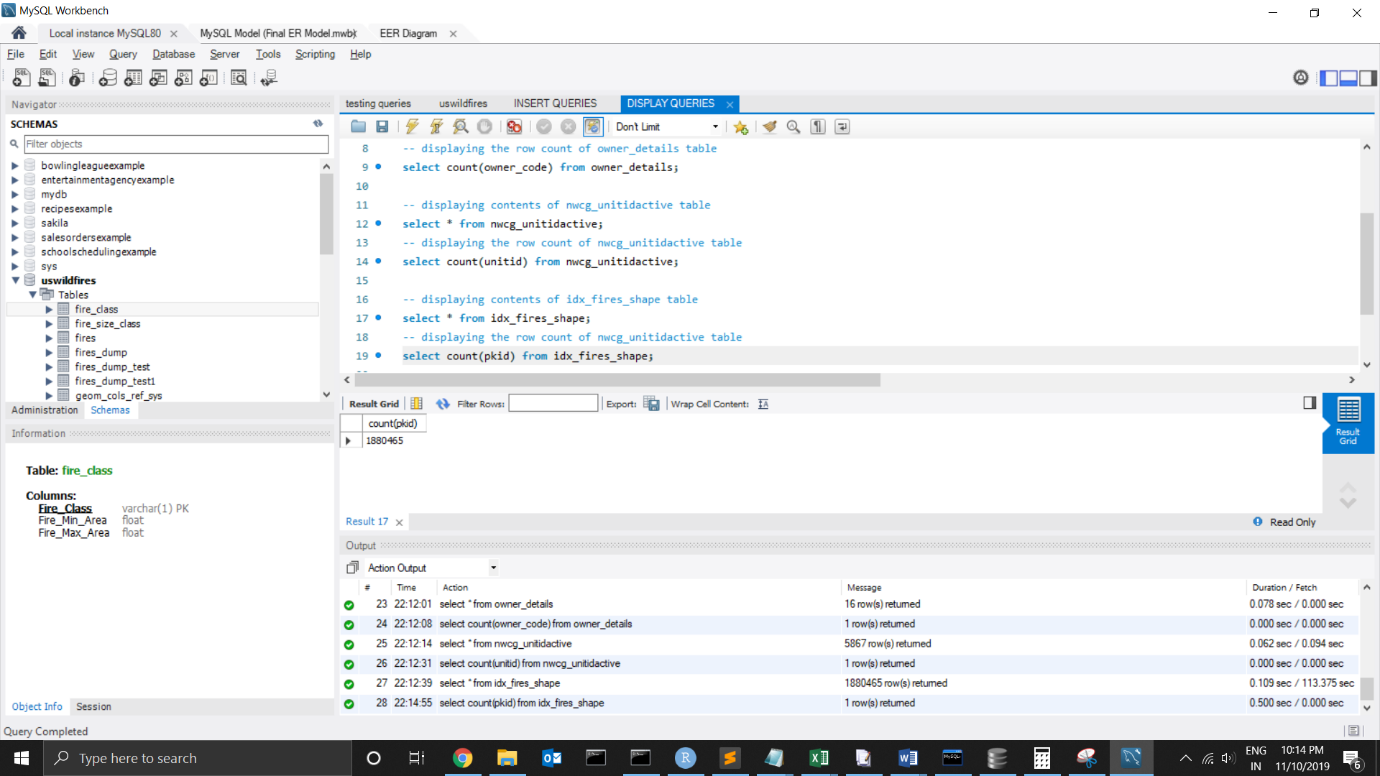
**NWCG\_UnitidActive:**





**IDX\_fires\_shape:**





**Fires:**

