##### Problem 1 #####

The datasetUSArrests.csvcontains statistics in arrests per 100,000 residents for assault and murder, in each of the 50 US states, in 1973. Also given is the percentage of the population living in urban areas:https://github.com/bforoura/IDS/tree/main/HW2

Use the pre-processing techniques at your disposal to prepare the dataset for analysis.

Excel

1. Address all the missing values.
2. Look for outliers and smooth noisy data, if any.
3. Plot Murder rates for all 50 states
4. Plot histogram of assaults
5. Plot Murder rate vs. Assault rate.
6. Prepare the dataset to establish a relation between an urban population category and a crime type.Hint: Convert the urban population percentage into categories, for example, small (<50%), medium (<60%), large (<70%), and extra-large (70% and above) urban population.

MySQL

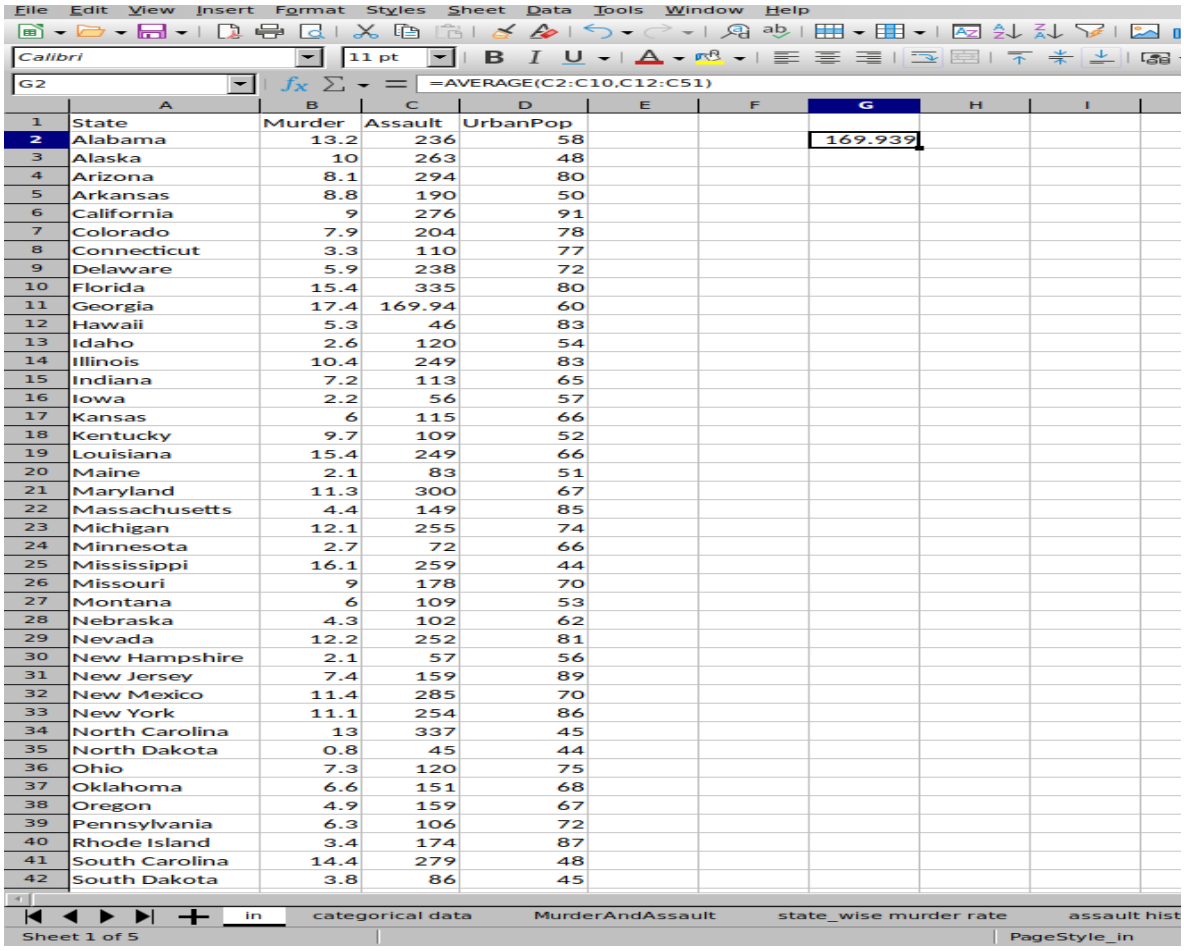
1. Import the original CSV file into MySQL and create the tableUSArrests.
2. Use SQL to replace all missing values in a column by the average.Hint:Use theUpdatecommand.
3. Findmin,max,mean, andvarianceof all numeric attributes in SQL.
4. Use SQL to answer the following questions:

    -- Which state has the maximum murder rate?

    -- List of states in ascending order of urban population percentages.

    -- How many states have higher murder rates than Arizona? List those states.

Ans:




| | A | B | C | D | E | F | G | H | I |
|---|---|---|---|---|---|---|---|---|---|
| 1 | State | Murder | Assault | UrbanPop | | | | | |
| 2 | Alabama | 13.2 | 236 | 58 | | | 169.939 | | |
| 3 | Alaska | 10 | 263 | 48 | | | | | |
| 4 | Arizona | 8.1 | 294 | 80 | | | | | |
| 5 | Arkansas | 8.8 | 190 | 50 | | | | | |
| 6 | California | 9 | 276 | 91 | | | | | |
| 7 | Colorado | 7.9 | 204 | 78 | | | | | |
| 8 | Connecticut | 3.3 | 110 | 77 | | | | | |
| 9 | Delaware | 5.9 | 238 | 72 | | | | | |
| 10 | Florida | 15.4 | 335 | 80 | | | | | |
| 11 | Georgia | 17.4 | 169.94 | 60 | | | | | |
| 12 | Hawaii | 5.3 | 46 | 83 | | | | | |
| 13 | Idaho | 2.6 | 120 | 54 | | | | | |
| 14 | Illinois | 10.4 | 249 | 83 | | | | | |
| 15 | Indiana | 7.2 | 113 | 65 | | | | | |
| 16 | Iowa | 2.2 | 56 | 57 | | | | | |
| 17 | Kansas | 6 | 115 | 66 | | | | | |
| 18 | Kentucky | 9.7 | 109 | 52 | | | | | |
| 19 | Louisiana | 15.4 | 249 | 66 | | | | | |
| 20 | Maine | 2.1 | 83 | 51 | | | | | |
| 21 | Maryland | 11.3 | 300 | 67 | | | | | |
| 22 | Massachusetts | 4.4 | 149 | 85 | | | | | |
| 23 | Michigan | 12.1 | 255 | 74 | | | | | |
| 24 | Minnesota | 2.7 | 72 | 66 | | | | | |
| 25 | Mississippi | 16.1 | 259 | 44 | | | | | |
| 26 | Missouri | 9 | 178 | 70 | | | | | |
| 27 | Montana | 6 | 109 | 53 | | | | | |
| 28 | Nebraska | 4.3 | 102 | 62 | | | | | |
| 29 | Nevada | 12.2 | 252 | 81 | | | | | |
| 30 | New Hampshire | 2.1 | 57 | 56 | | | | | |
| 31 | New Jersey | 7.4 | 159 | 89 | | | | | |
| 32 | New Mexico | 11.4 | 285 | 70 | | | | | |
| 33 | New York | 11.1 | 254 | 86 | | | | | |
| 34 | North Carolina | 13 | 337 | 45 | | | | | |
| 35 | North Dakota | 0.8 | 45 | 44 | | | | | |
| 36 | Ohio | 7.3 | 120 | 75 | | | | | |
| 37 | Oklahoma | 6.6 | 151 | 68 | | | | | |
| 38 | Oregon | 4.9 | 159 | 67 | | | | | |
| 39 | Pennsylvania | 6.3 | 106 | 72 | | | | | |
| 40 | Rhode Island | 3.4 | 174 | 87 | | | | | |
| 41 | South Carolina | 14.4 | 279 | 48 | | | | | |
| 42 | South Dakota | 3.8 | 86 | 45 | | | | | |

Fig 1.1.1: Address missing value using =AVERAGE(C2:C10,C12:C51)

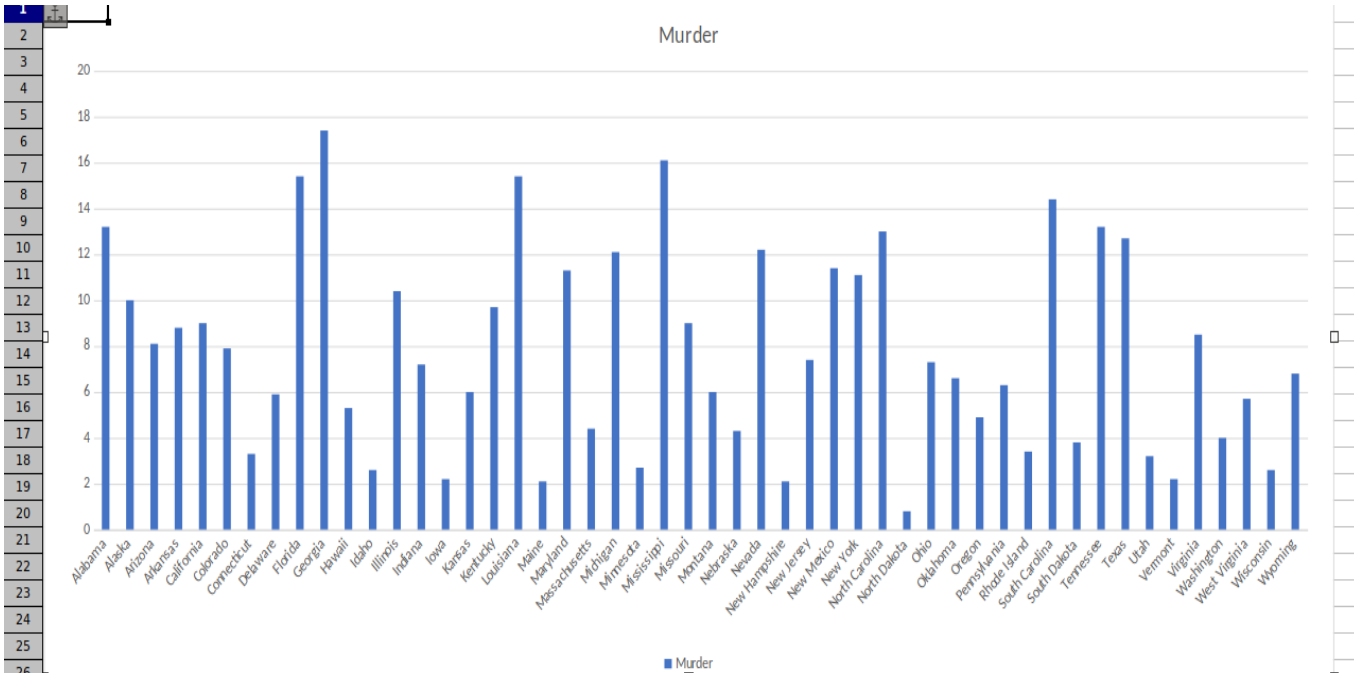



Fig 1.1.2: Plot Murder rates for all 50 states

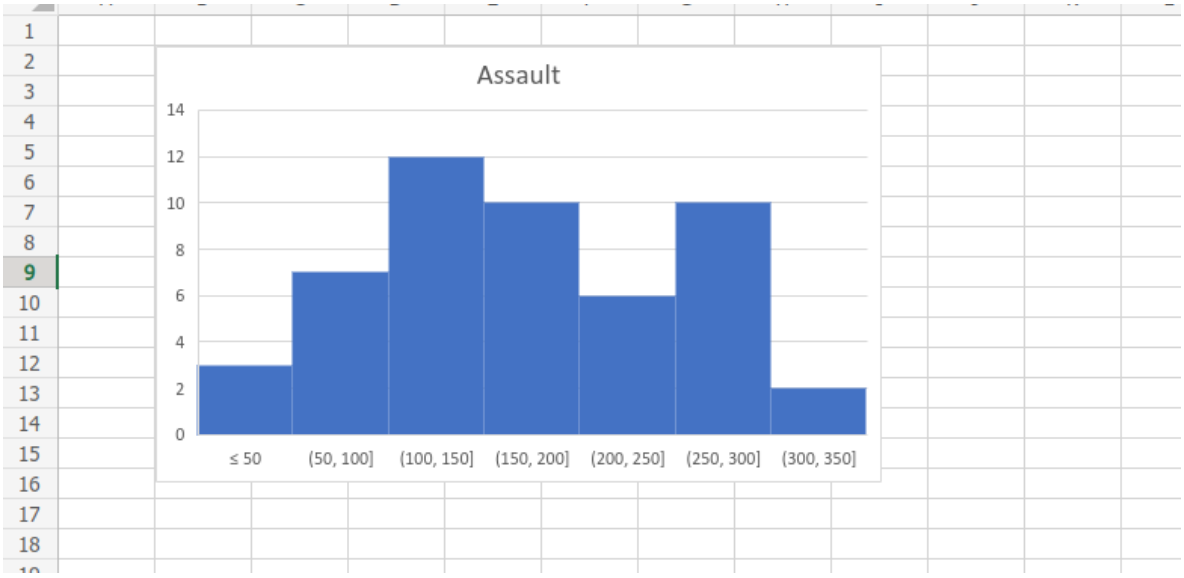



Fig 1.1.3: Plot histogram of assaults

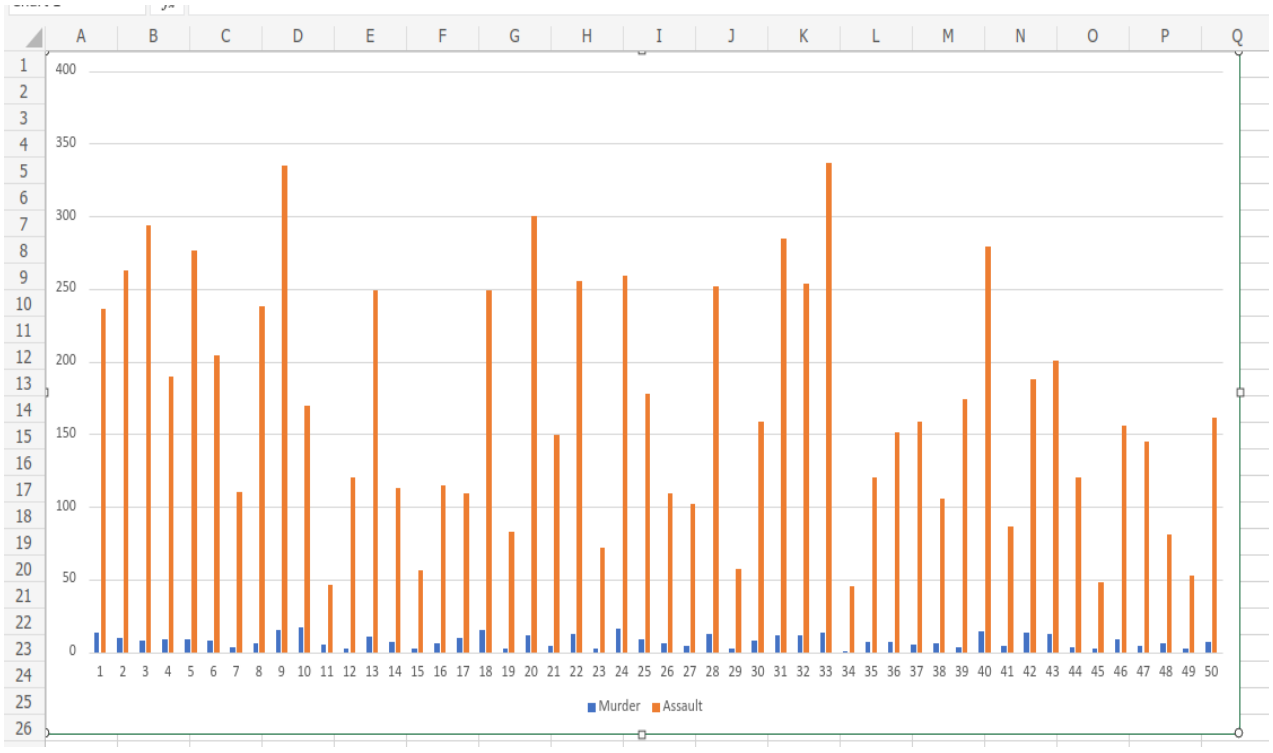



Fig 1.1.4: Plot Murder rate vs. Assault rate.

F14 =IF(AND(B14<5,C14<100), "least",IF(AND(B14<10,C14<150), "less", IF(AND(B14<15,C14<200),"more","most")))

| | A | B | C | D | E | F | G |
|---|---|---|---|---|---|---|---|
| 1 | State | Murder | Assault | UrbanPop | Urban_cat | crime_type | |
| 2 | Alabama | 13.2 | 236 | 58 | medium | most | |
| 3 | Alaska | 10 | 263 | 48 | small | most | |
| 4 | Arizona | 8.1 | 294 | 80 | Extra large | most | |
| 5 | Arkansas | 8.8 | 190 | 50 | medium | more | |
| 6 | California | 9 | 276 | 91 | Extra large | most | |
| 7 | Colorado | 7.9 | 204 | 78 | Extra large | most | |
| 8 | Connecticu | 3.3 | 110 | 77 | Extra large | less | |
| 9 | Delaware | 5.9 | 238 | 72 | Extra large | most | |
| 10 | Florida | 15.4 | 335 | 80 | Extra large | most | |
| 11 | Georgia | 17.4 | 169.9388 | 60 | large | most | |
| 12 | Hawaii | 5.3 | 46 | 83 | Extra large | less | |
| 13 | Idaho | 2.6 | 120 | 54 | medium | less | |
| 14 | Illinois | 10.4 | 249 | 83 | Extra large | most | |
| 15 | Indiana | 7.2 | 113 | 65 | large | less | |
| 16 | Iowa | 2.2 | 56 | 57 | medium | least | |
| 17 | Kansas | 6 | 115 | 66 | large | less | |
| 18 | Kentucky | 9.7 | 109 | 52 | medium | less | |
| 19 | Louisiana | 15.4 | 249 | 66 | large | most | |
| 20 | Maine | 2.1 | 83 | 51 | medium | least | |
| 21 | Maryland | 11.3 | 300 | 67 | large | most | |
| 22 | Massachu | 4.4 | 149 | 85 | Extra large | less | |
| 23 | Michigan | 12.1 | 255 | 74 | Extra large | most | |
| 24 | Minnesota | 2.7 | 72 | 66 | large | least | |
| 25 | Mississipp | 16.1 | 259 | 44 | small | most | |
| 26 | Missouri | 9 | 178 | 70 | Extra large | more | |
| 27 | Montana | 6 | 109 | 53 | medium | less | |
| 28 | Nebraska | 4.3 | 102 | 62 | large | less | |
| 29 | Nevada | 12.2 | 252 | 81 | Extra large | most | |
| 30 | New Hamp | 2.1 | 57 | 56 | medium | least | |
| 31 | New Jerse | 7.4 | 159 | 89 | Extra large | more | |
| 32 | New Mexi | 11.4 | 285 | 70 | Extra large | most | |
| 33 | New York | 11.1 | 254 | 86 | Extra large | most | |
| 34 | North Car | 13 | 337 | 45 | small | most | |
| 35 | North Dak | 0.8 | 45 | 44 | small | least | |
| 36 | Ohio | 7.3 | 120 | 75 | Extra large | less | |

Fig 1.1.5: urban population category and a crime type

#Formula:

1. =IF(D2<50, "small",IF(D2<60, "medium", IF(D2<70,"large","Extra large")))

2. =IF(B2<5 AND C2<100, "least",IF(B2<10 AND C2<150, "less", IF(B2<15 AND C2<200,"more","most")))

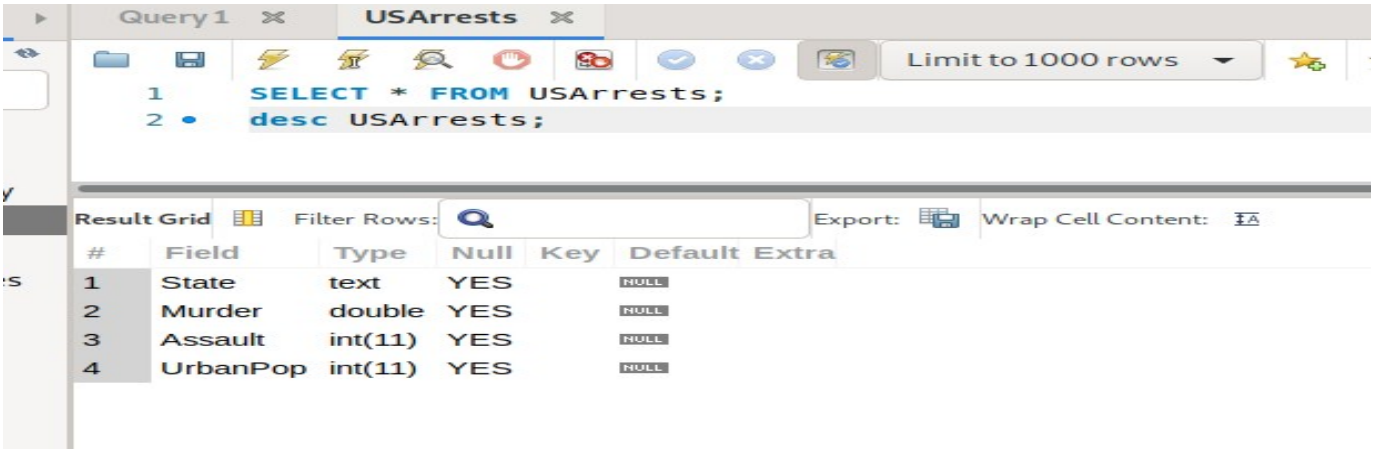



Fig 1.2.2: mysql describe table USArrests.

Command:  desc USArrests;

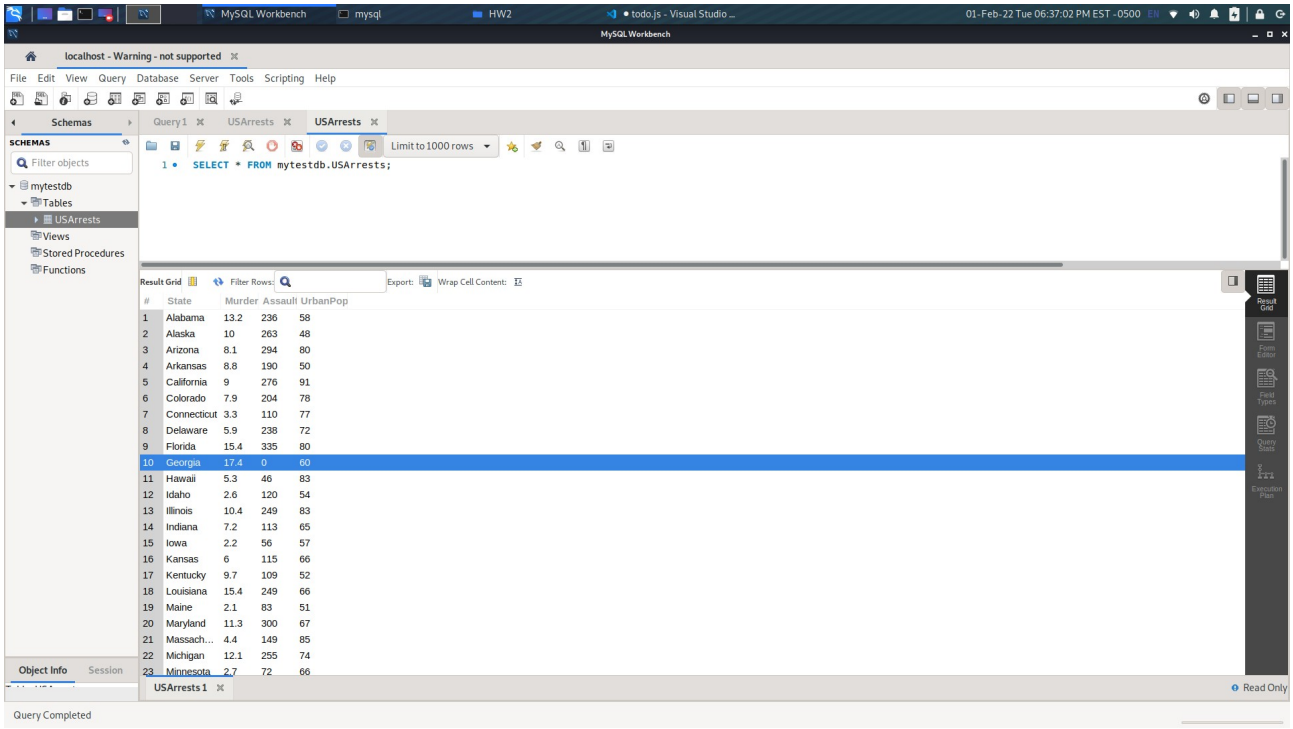



Fig 1.2.2: mysql import data and select projection.

Sql: select * from  USArrests;

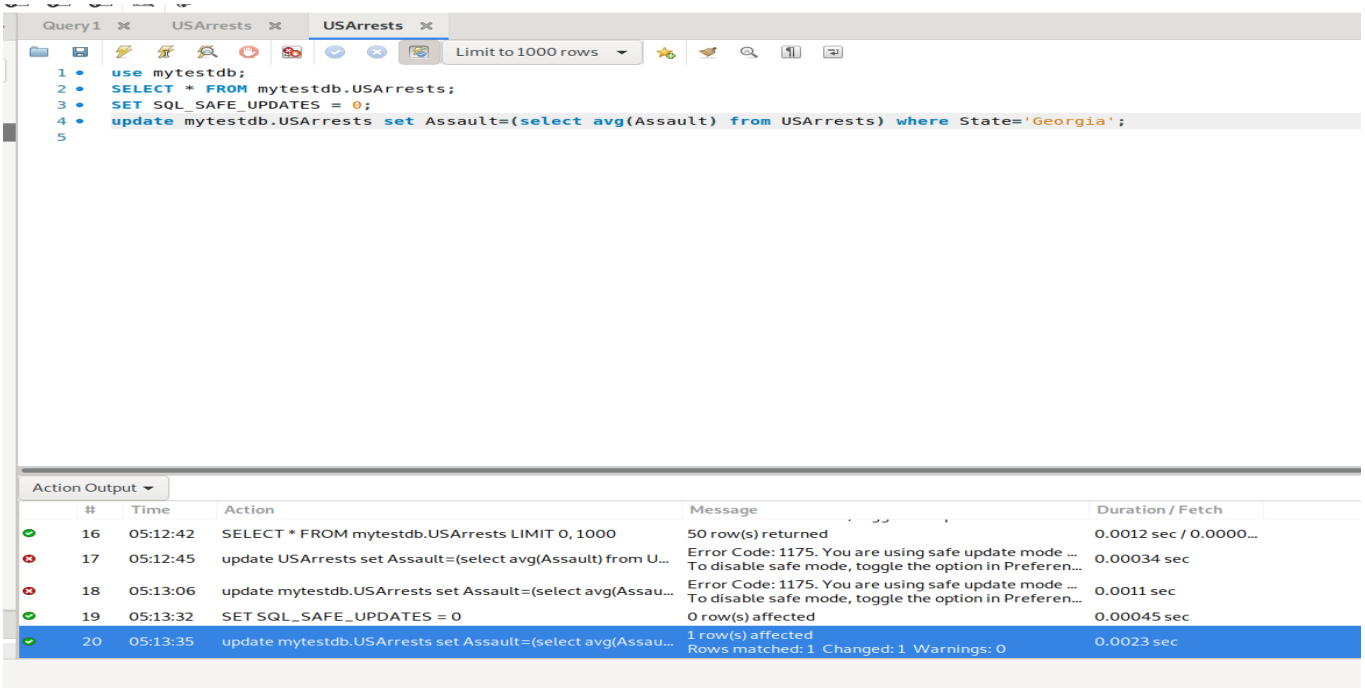
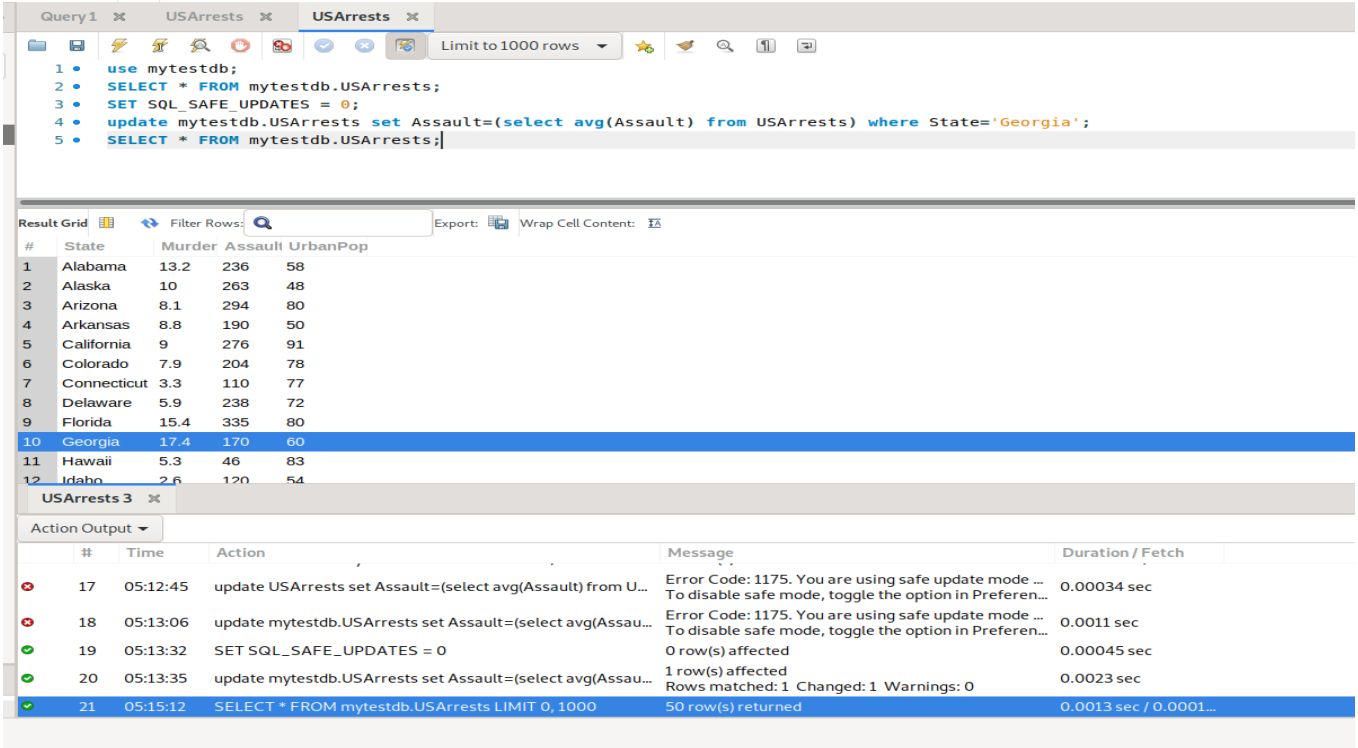



Fig 1.2.3 : address missing command.

Sql: update  USArrests set assault=(select avg(assault) from USArrests) where state=’Georgia’;




Sql:Fig 1.2.4: Again selection.

Sql: select * from  USArrests;

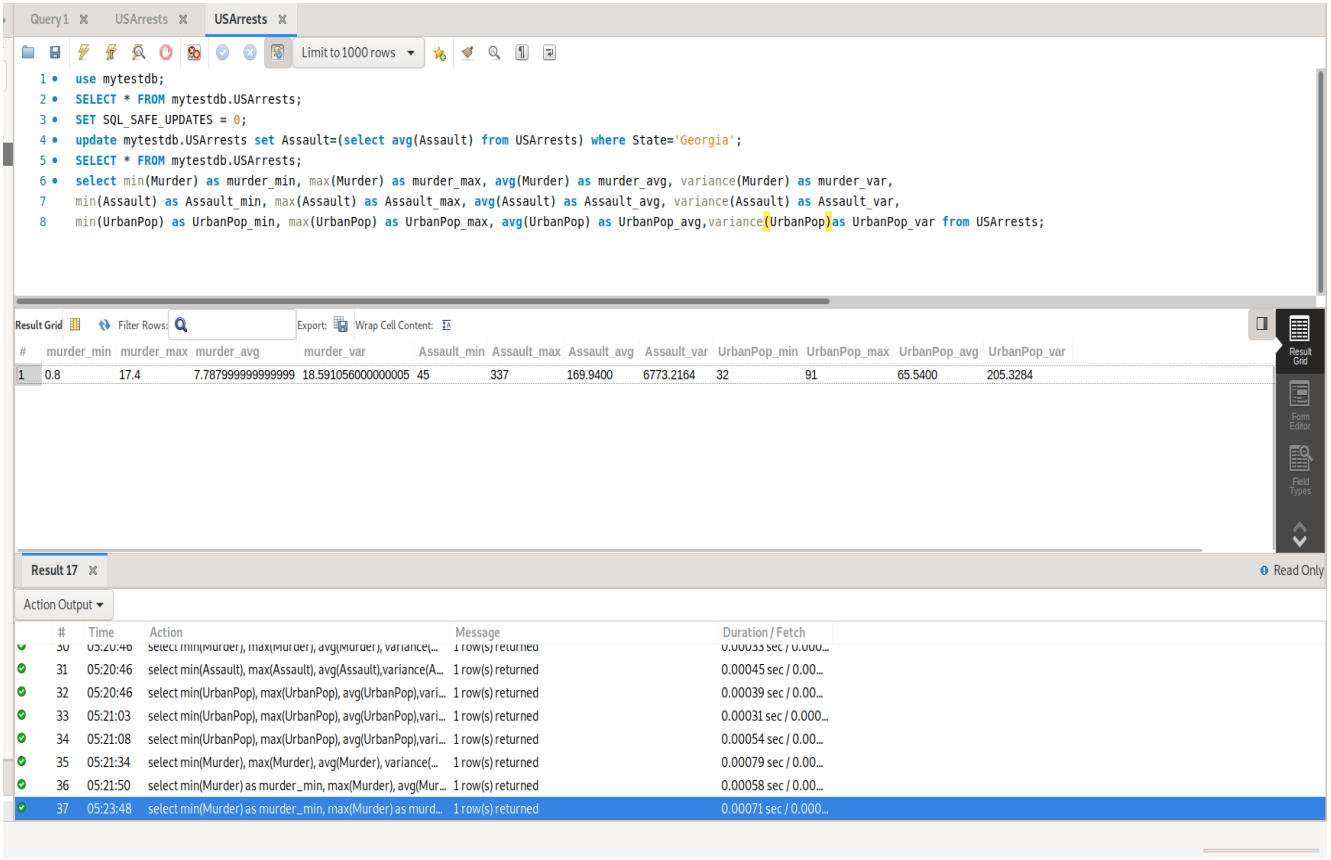



Fig: 1.2.5: min max, avg

sql: select min(Murder) as murder_min, max(Murder) as murder_max, avg(Murder) as murder_avg, variance(Murder) as murder_var, min(Assault) as Assault_min, max(Assault) as Assault_max, avg(Assault) as Assault_avg, variance(Assault) as Assault_var, min(UrbanPop) as UrbanPop_min, max(UrbanPop) as UrbanPop_max, avg(urbanPop) as UrbanPop_avg, variance(UrbanPop) as UrbanPop_var from USArrests;

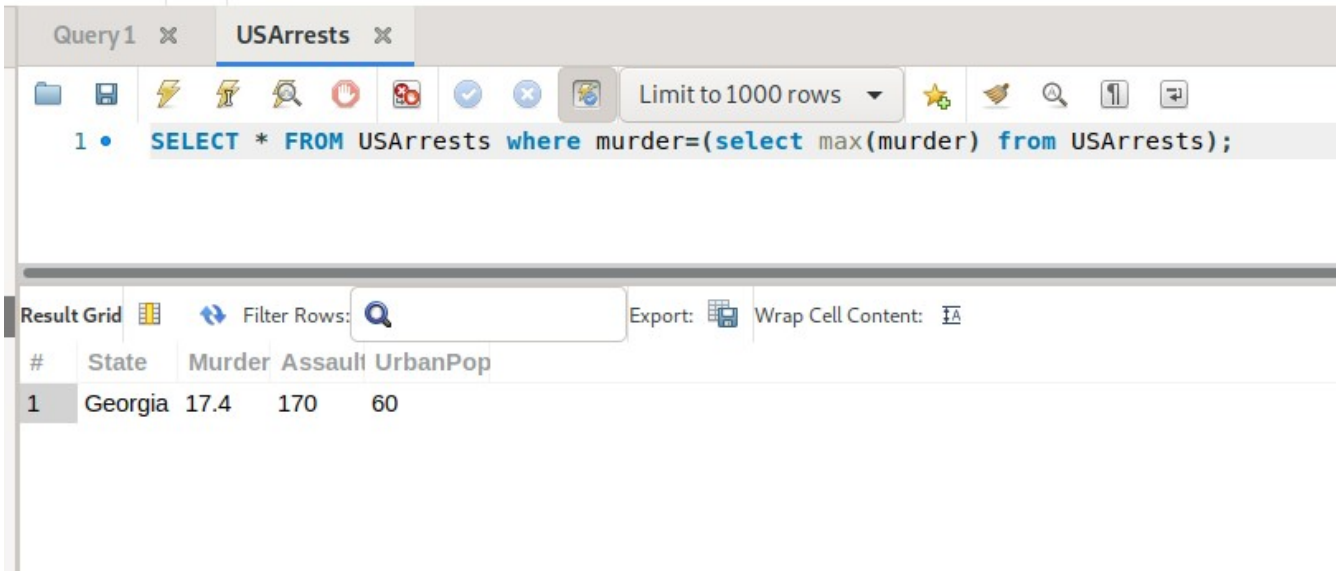



Fig 1.2.6 : Which state has the maximum murder rate?

sql: select * from USArrests where murder=(select max(murder) from USArrests);

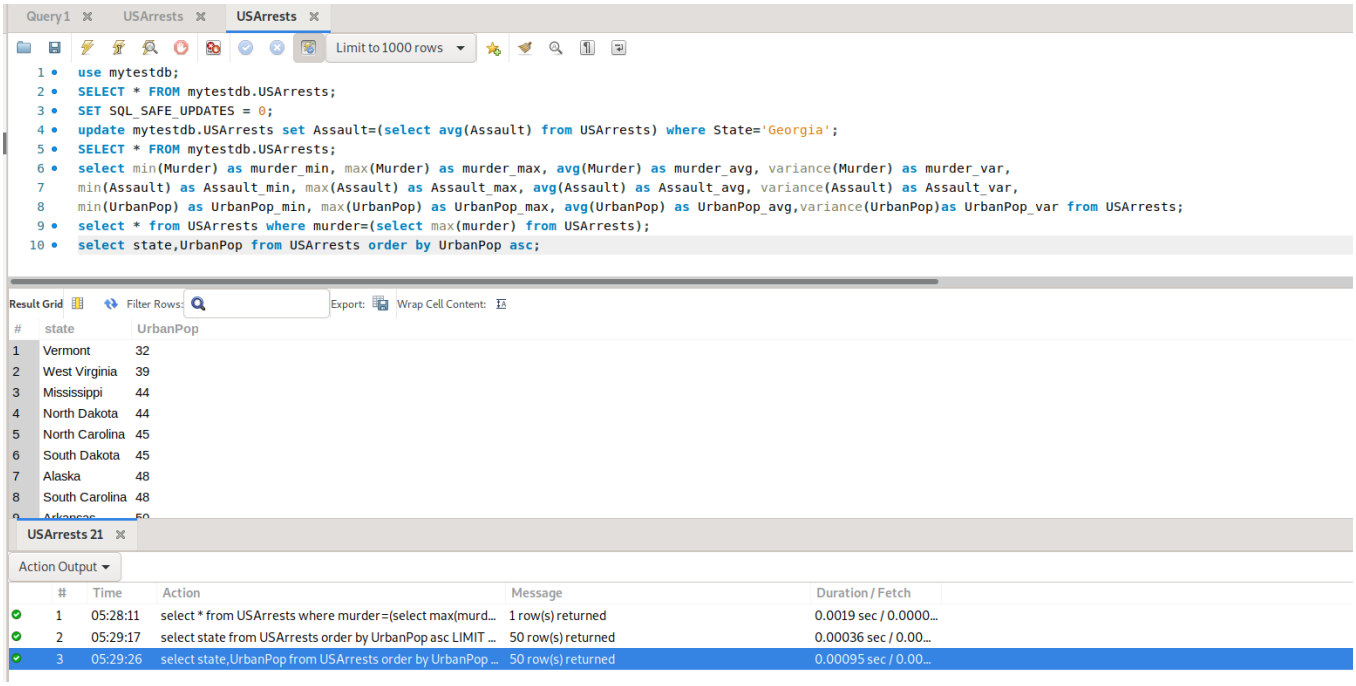



Fig 1.2.7 List of states in ascending order of urban population percentages.

Sql: select state,urbanpop from USArrests order by urbanpop asc;

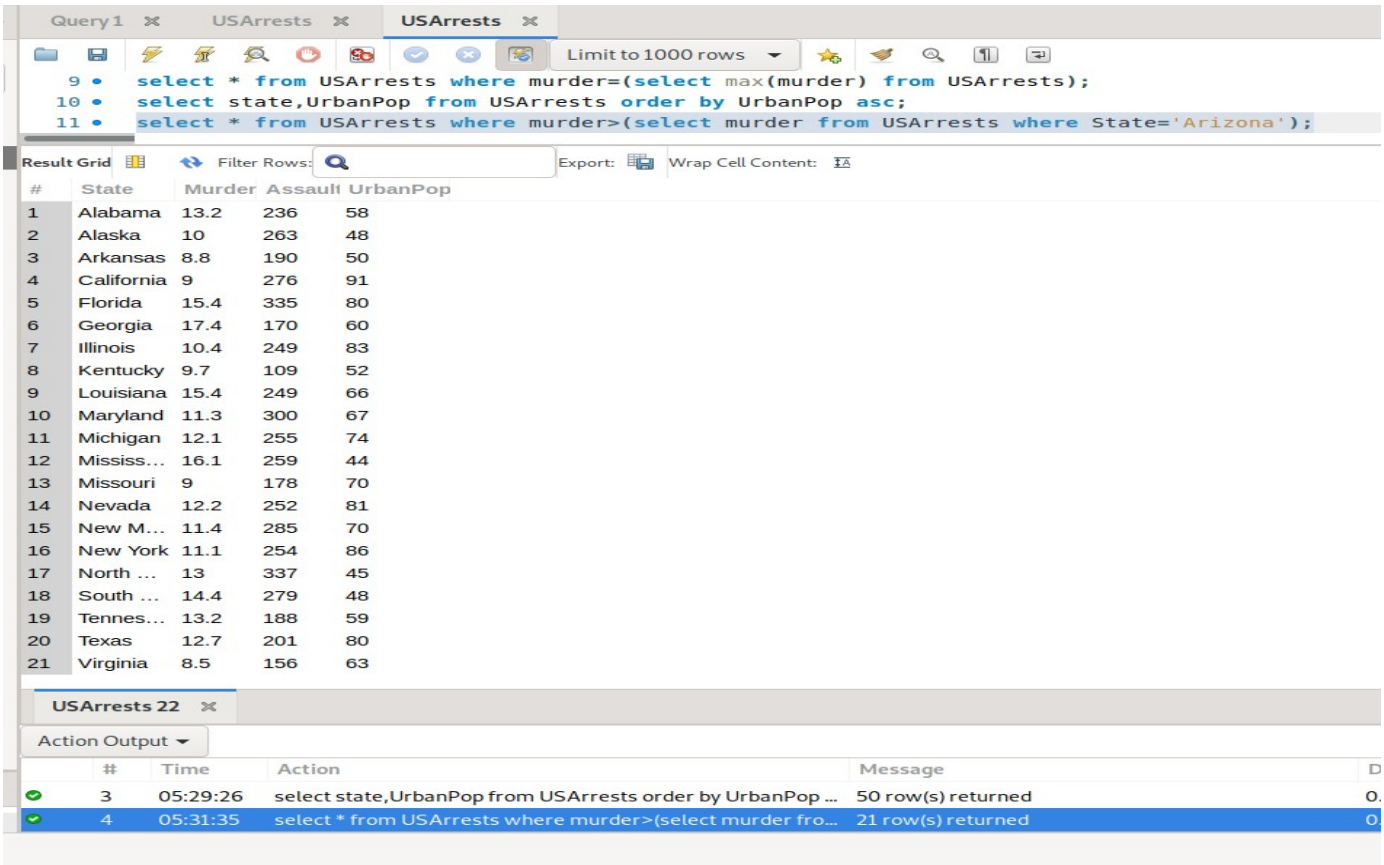



Fig 1.2.8 How many states have higher murder rates than Arizona? List those states.

Sql: select state from USArrests where murder>(select murder from USArrests where state='Arizona');

####### Problem 2 ######

. This problem's dataset involveschild mortality ratesand is inspired by data collected from UNICEF. The original dataset is available here: https://github.com/bforoura/IDS/tree/main/HW2

According to the report, the world has achieved substantial success in reducing child mortality during the last few decades. According to the UNICEF report, globally the under-five age mortality rate has decreased from 93 deaths per 1,000 live births in 1990 to less than 50 in 2016.

However, the dataset has a number of missing instances, which need to be fixed before a clear progress on child mortality can be explained from the year of 1990 to 2016. Use this dataset to complete the following tasks:

Excel

1. Address all the missing values using the techniques at your disposal.
2. Prepare the dataset to establish the following relations (graphs in Excel):

--Under-five mortality rateandneonatal mortality rate.

--Infant mortality rateandneonatal mortality rate.

--Yearandinfant mortality rate.

Hint: You may think of converting the mortality rates into five-point Likert scale values. You may count the year before this dataset (i.e., 1989) as the starting point of this program, to assess the progress we have made as the years have passed.

MySQL

1. Import the original data set into MySQL
2. Use SQL to fill in the missing values in each column using the medians.
3. Answer the following queries in SQL:

-- Display the entire table.

-- Which years have the lowest and highest infant mortality years, respectively?

-- In what years the neonatal mortality rates were above average?

-- Display the sorted infant mortality rates in descending order.

-- Displaymin,max,mean,variance, andstandard deviationfor each mortality rate.

-- Add a new column calledAbove-Five Mortality Rateand populate it with appropriate values.Hint:UseAlter Table Add Column.

-- Display the entire table again.

Ans:

| | A | B | C | D | E |
|---|---|---|---|---|---|
| 1 | Year | Under-five mor | Infant mortality | Neonatal mortality rate | |
| 2 | 1990 | 93.4 | 64.8 | 36.8 | |
| 3 | 1991 | 92.1 | 63.9 | 36.3 | |
| 4 | 1992 | 90.9 | 63.1 | 35.9 | |
| 5 | 1993 | 89.7 | 62.3 | 35.4 | |
| 6 | 1994 | 88.7 | 61.4 | 39.24 | |
| 7 | 1995 | 87.3 | 60.5 | 34.4 | |
| 8 | 1996 | 85.6 | 59.4 | 33.7 | |
| 9 | 1997 | 100.54 | 58.2 | 33.1 | |
| 10 | 1998 | 82.1 | 56.9 | 32.3 | |
| 11 | 1999 | 79.9 | 55.4 | 31.5 | |
| 12 | 2000 | 77.5 | 53.9 | 30.7 | |
| 13 | 2001 | 74.8 | 52.1 | 29.8 | |
| 14 | 2002 | 72 | 72.69 | 28.9 | |
| 15 | 2003 | 69.2 | 48.6 | 28 | |
| 16 | 2004 | 66.7 | 46.9 | 38.4 | |
| 17 | 2005 | 103.09 | 45.1 | 26.1 | |
| 18 | 2006 | 61.1 | 43.4 | 25.3 | |
| 19 | 2007 | 58.5 | 71.06 | 24.4 | |
| 20 | 2008 | 56.2 | 40.3 | 23.6 | |
| 21 | 2009 | 53.7 | 38.8 | 22.9 | |
| 22 | 2010 | 105.64 | 37.4 | 22.2 | |
| 23 | 2011 | 49.3 | 36 | 21.5 | |
| 24 | 2012 | 47.3 | 34.7 | 20.8 | |
| 25 | 2013 | 45.5 | 33.6 | 20.2 | |
| 26 | 2014 | 43.7 | 69.44 | 19.6 | |
| 27 | 2015 | 42.2 | 31.4 | 19.1 | |
| 28 | 2016 | 40.8 | 30.5 | 18.6 | |
| 29 | | | | | |

Fig: 2.1.1 Address all missing data

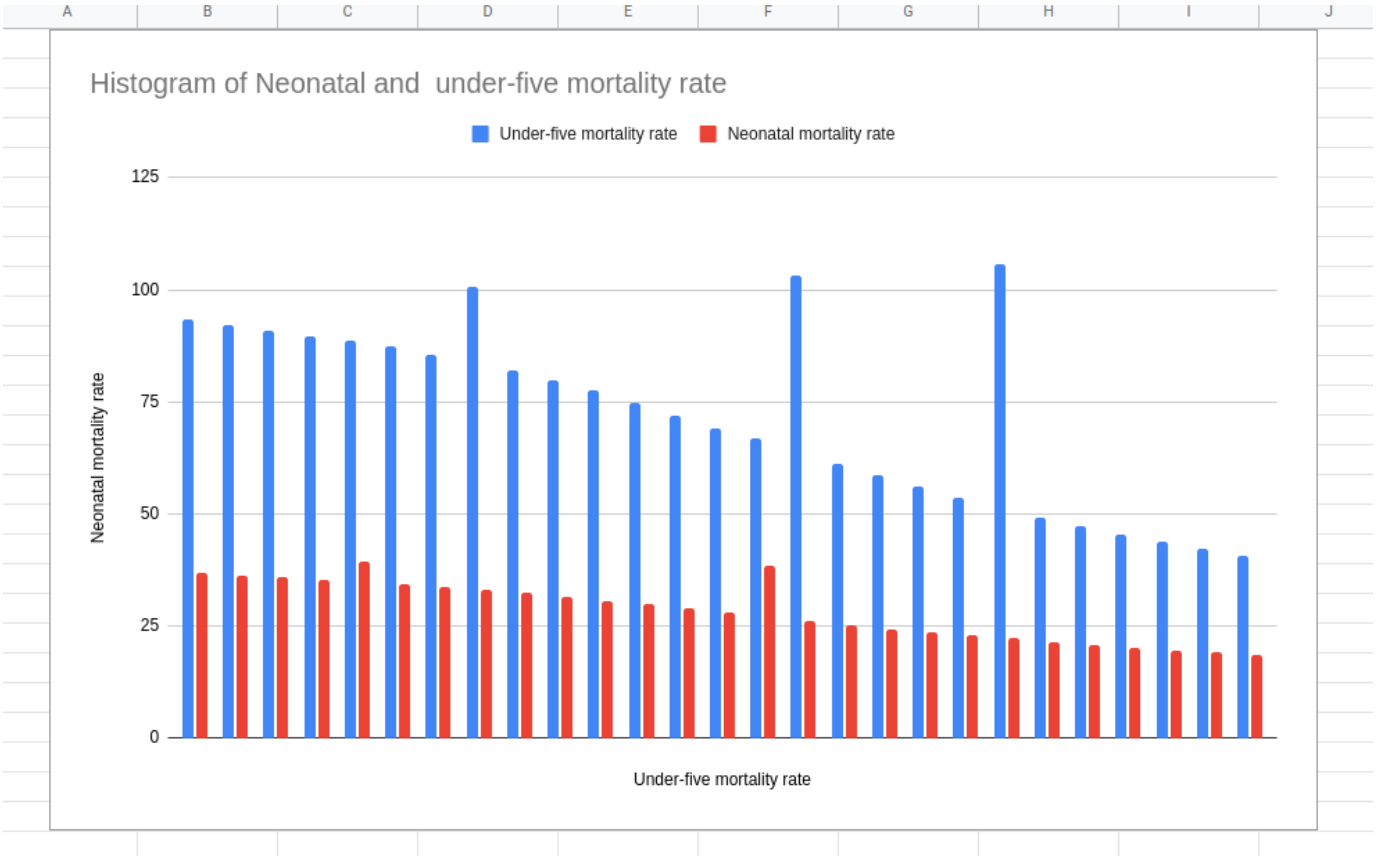



Fig 2.1.2 chart for Under-five mortality rate and neonatal mortality rate.

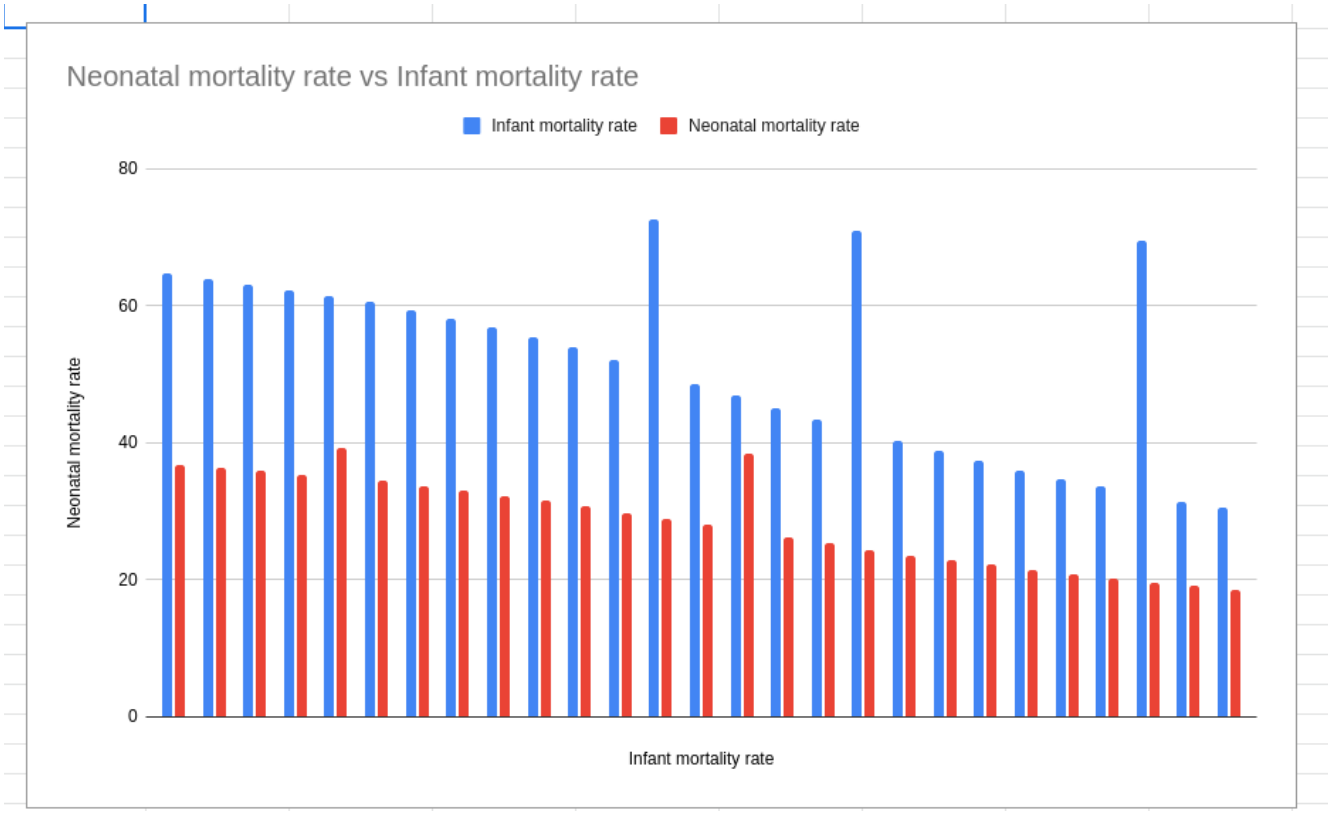



Fig 2.1.3: chart for Infant mortality rate and neonatal mortality rate.

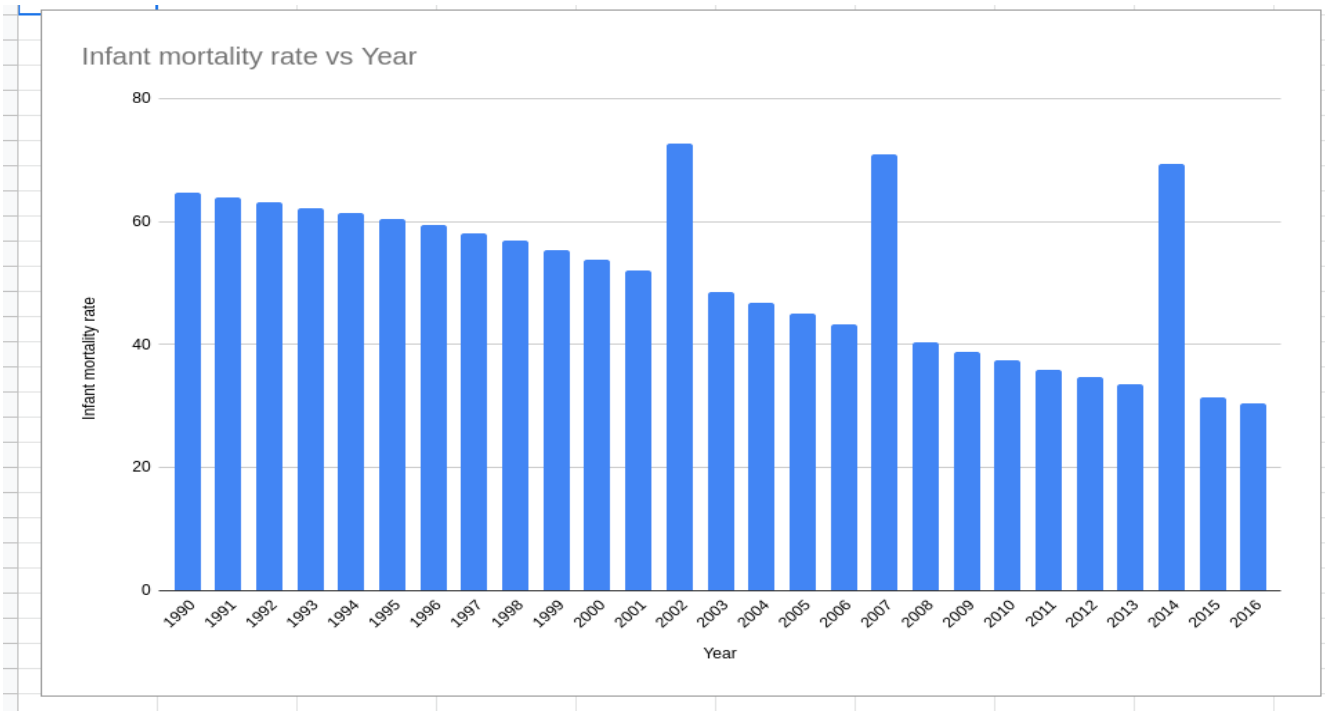



Fig 2.1.4 chart for --Year and infant mortality rate.

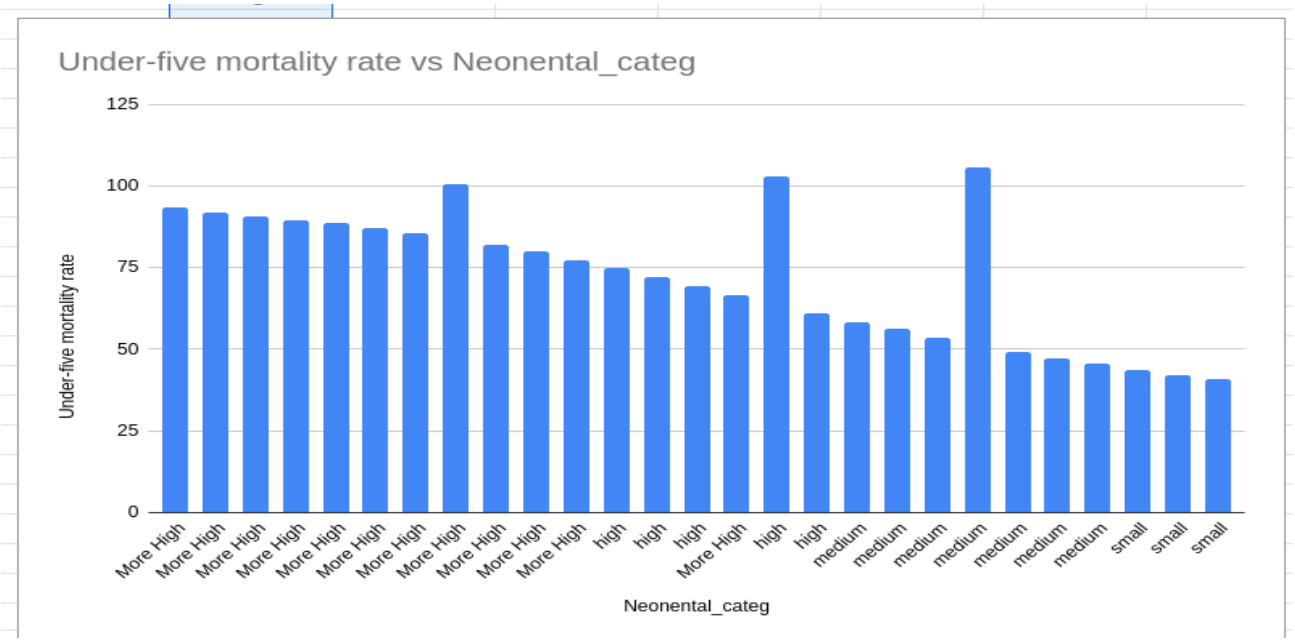



Fig 2.1.4 chart for Under-five mortality rate and neonatal mortality rate using categorical

formula: =IF(D2<20, "small",IF(D2<25, "medium", IF(D2<30,"high","More High")))

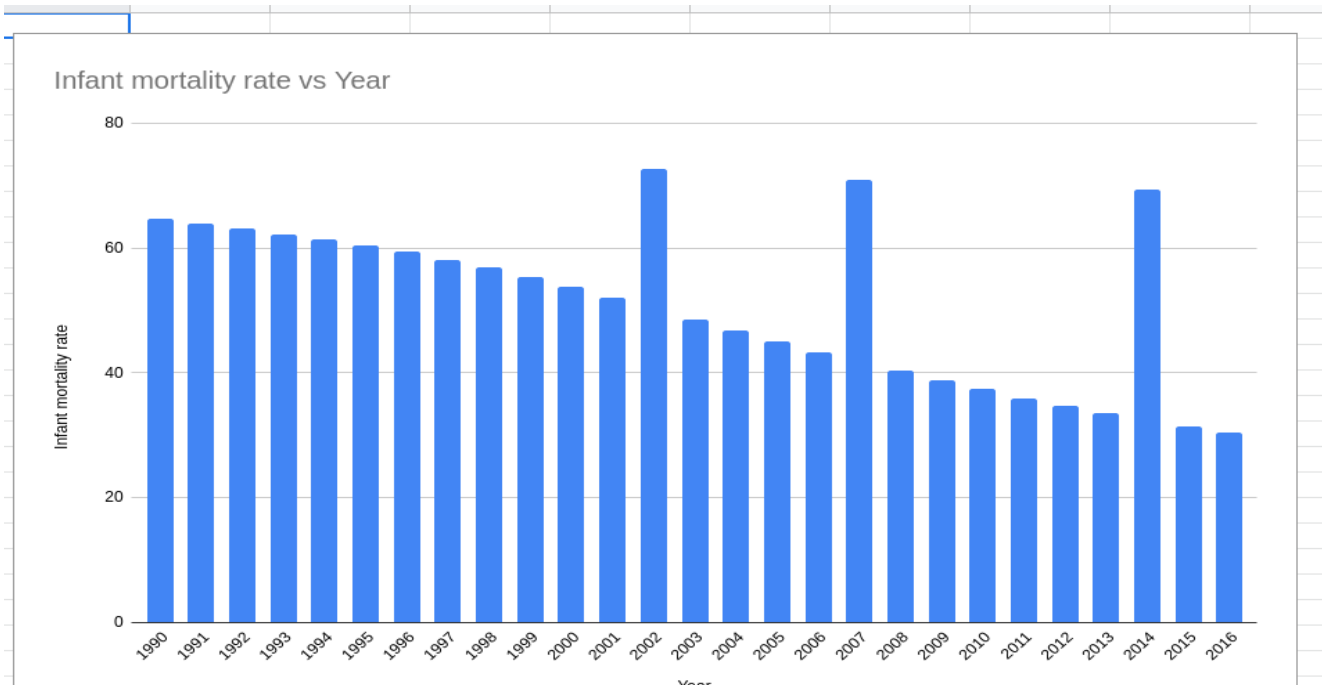



Fig 2.1.5 --Infant mortality rate and neonatal mortality rate.

formula: =IF(D2<20, "small",IF(D2<25, "medium", IF(D2<30,"high","More High")))

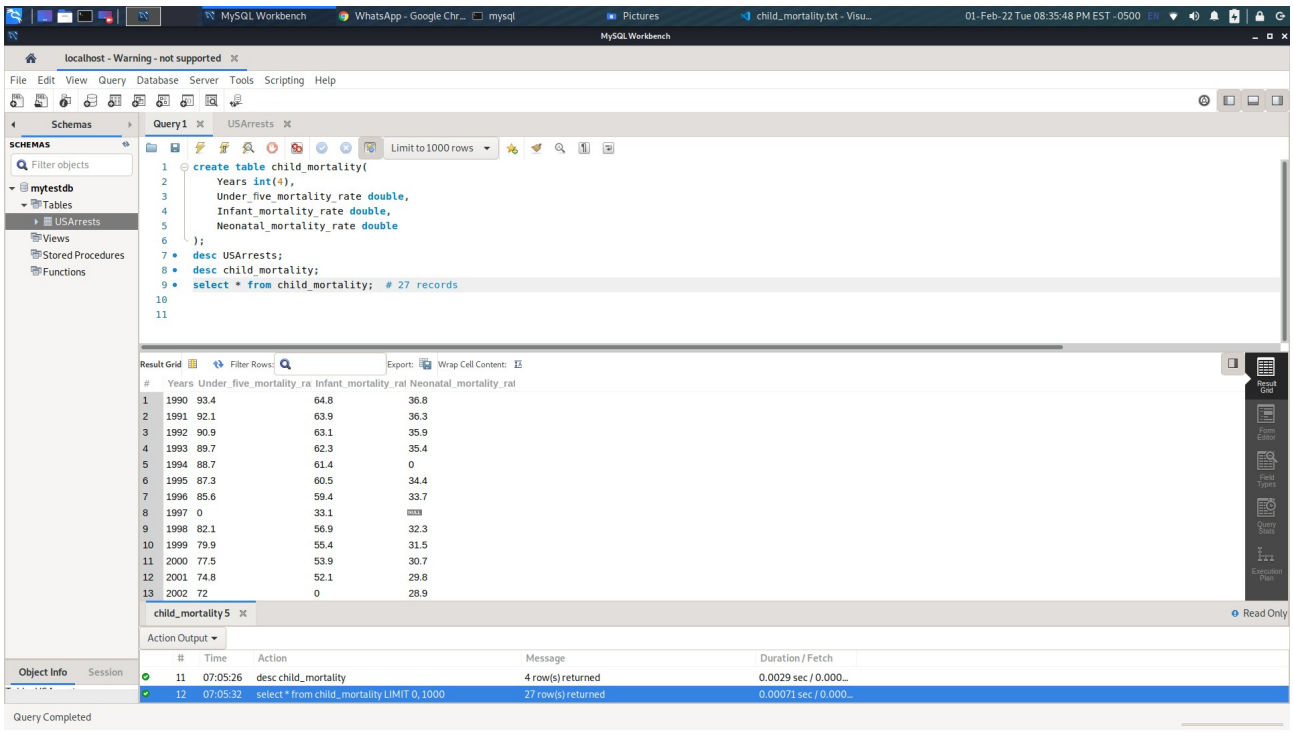



Fig 2.2.1 : mysql import dataset and projection.

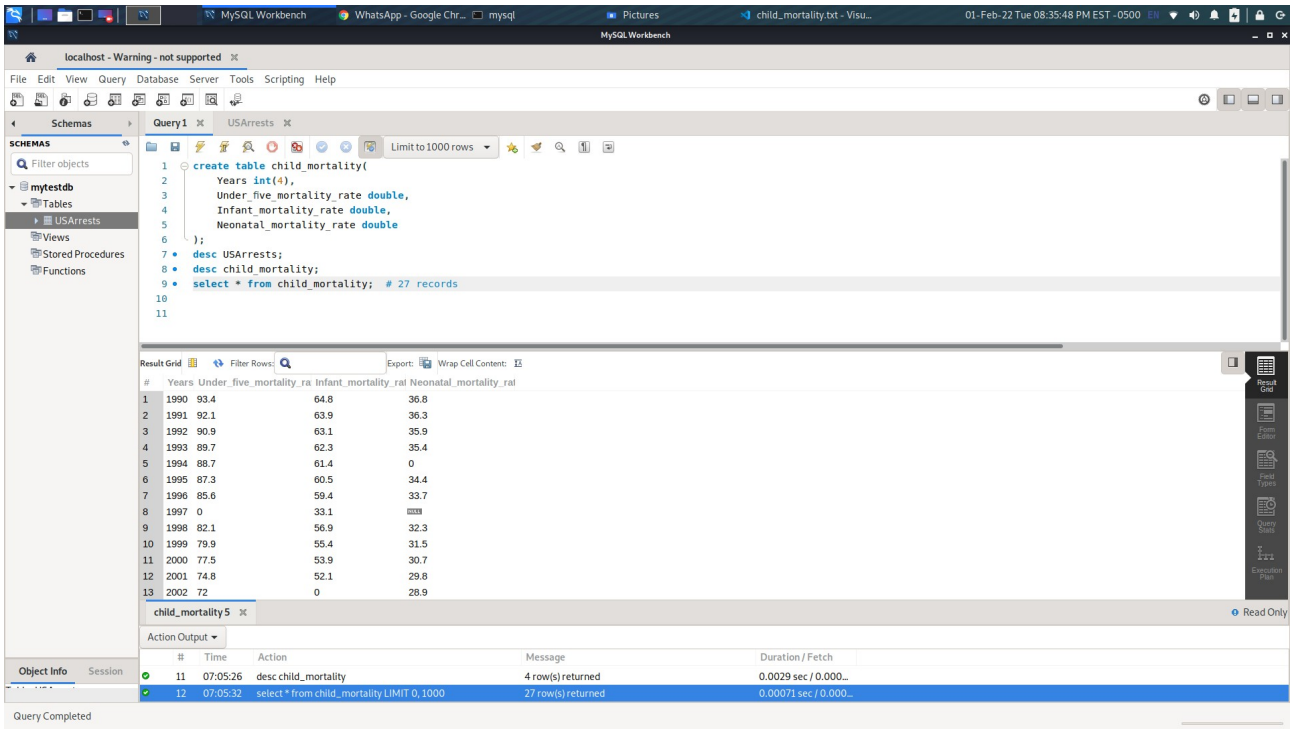



Fig 2.2.2 get median value

```
9
10  SET @rowindex := -1;
11  update child_mortality set Under_five_mortality_rate=(
12      SELECT
13         AVG(d.Under_five_mortality_rate) as Median
14      FROM
15         (SELECT @rowindex:=@rowindex + 1 AS rowindex,
16                 child_mortality.Under_five_mortality_rate AS Under_five_mortality_rate
17          FROM child_mortality
18          ORDER BY child_mortality.Under_five_mortality_rate) AS d
19      WHERE
20      d.rowindex IN (FLOOR(@rowindex / 2), CEIL(@rowindex / 2))
21  ) where Under_five_mortality_rate=0;
```

Action Output

| # | Time | Action | Message | Duration / Fetch |
|---|---|---|---|---|
| 27 | 07:14:16 | desc child_mortality | 4 row(s) returned | 0.0033 sec / 0.0000... |
| 28 | 07:14:29 | select * from child_mortality LIMIT 0, 1000 | 27 row(s) returned | 0.00051 sec / 0.000... |
| 29 | 07:20:27 | desc child_mortality | 4 row(s) returned | 0.0016 sec / 0.0000... |
| 30 | 10:20:26 | select * from child_mortality LIMIT 0, 1000 | 27 row(s) returned | 0.0015 sec / 0.0000... |
| 31 | 10:22:15 | SET @rowindex := -1 | 0 row(s) affected | 0.00013 sec |
| 32 | 10:22:15 | update child_mortality set Under_five_mortality_rate=( ... | 3 row(s) affected<br>Rows matched: 3 Changed: 3 Warnings: 0 | 0.0014 sec |

Fig 2.2.3 update missing value by median

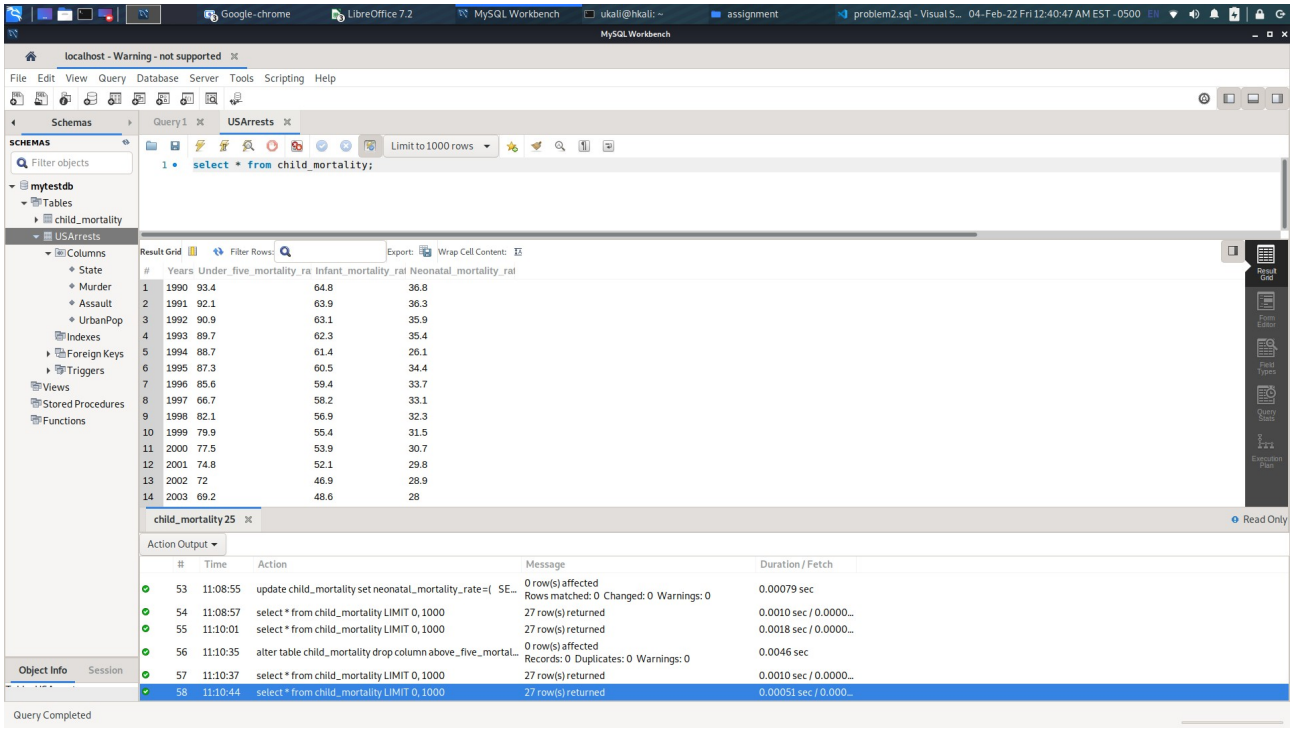



Fig: 2.2.4: projection after updating missing values

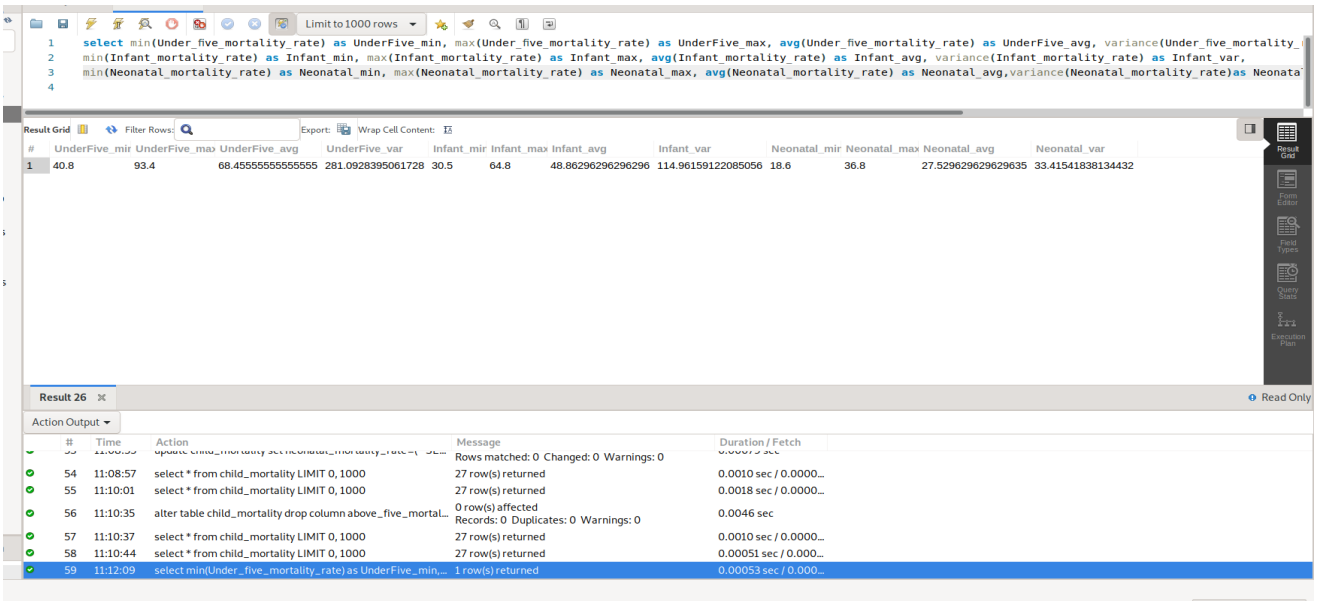



Fig: 2.2.5: Display min, max, mean, variance, and standard deviation for each mortality rate.

Sql: select min(Under_five_mortality_rate) as UnderFive_min, max(Under_five_mortality_rate) as UnderFive_max, avg(Under_five_mortality_rate) as UnderFive_avg, variance(Under_five_mortality_rate) as UnderFive_var, min(Infant_mortality_rate) as Infant_min, max(Infant_mortality_rate) as Infant_max, avg(Infant_mortality_rate) as Infant_avg, variance(Infant_mortality_rate) as Infant_var,

min(Neonatal_mortality_rate) as Neonatal_min, max(Neonatal_mortality_rate) as Neonatal_max, avg(Neonatal_mortality_rate) as Neonatal_avg,variance(Neonatal_mortality_rate)as Neonatal_var from child_mortality;

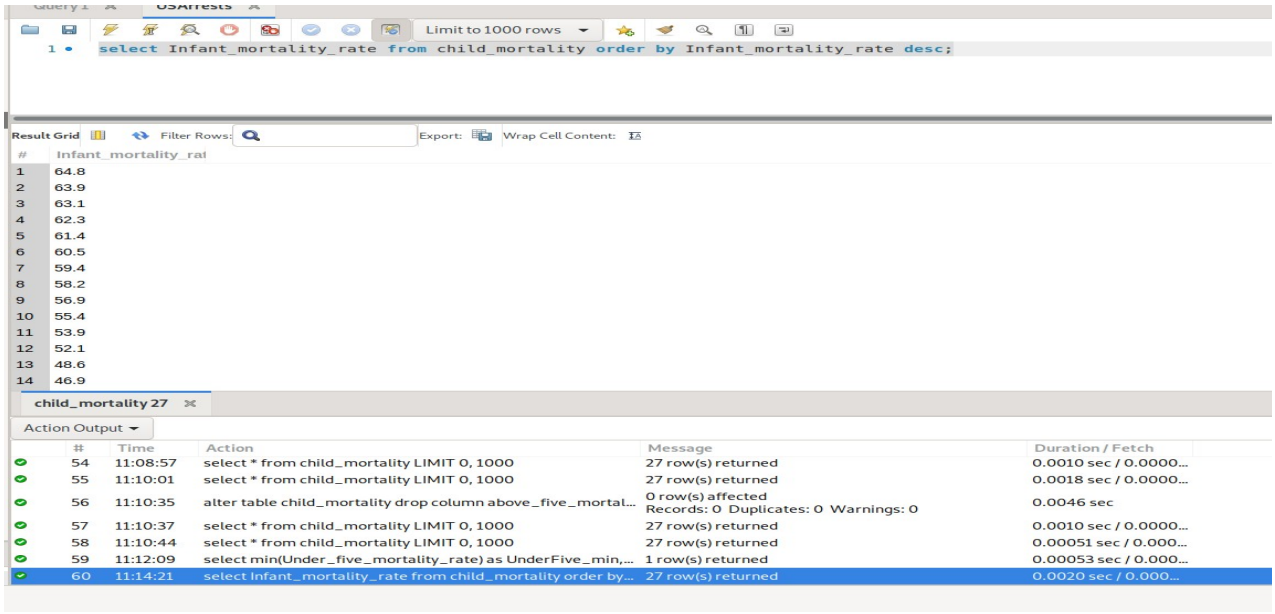



Fig 2.2.6: Display the sorted infant mortality rates in descending order.

Sql: select Infant_mortality_rate from child_mortality order by Infant_mortality_rate desc;

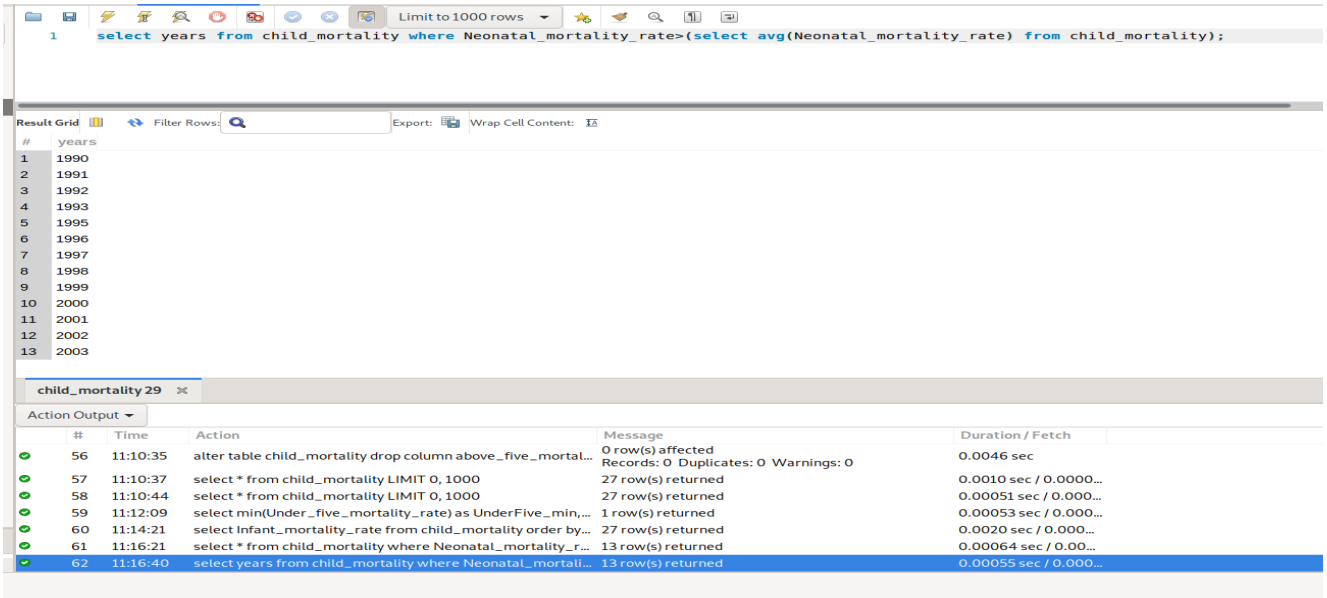



Fig: 2.2.7 - In what years the neonatal mortality rates were above average?

sql: select years from child_mortality where Neonatal_mortality_rate>(select avg(Neonatal_mortality_rate) from child_mortality);

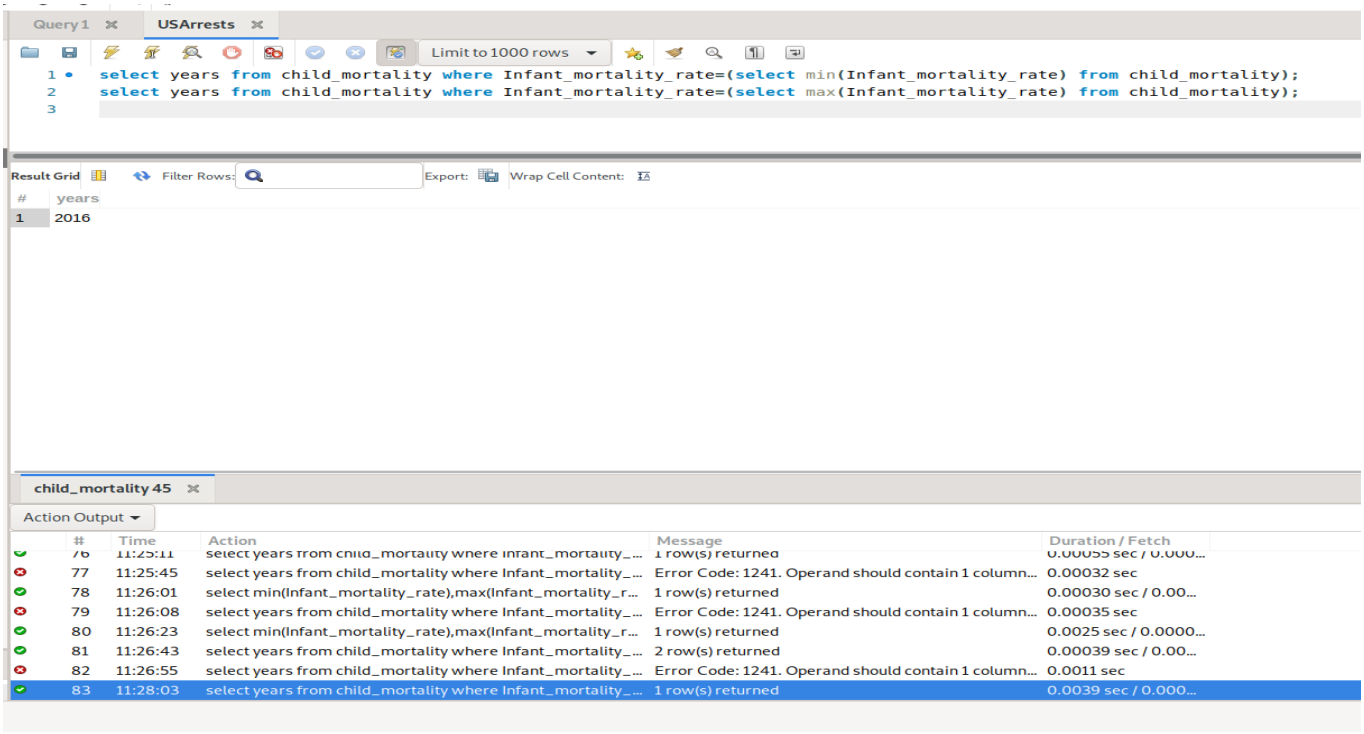



Fig 2.28 -- Which years have the lowest and highest infant mortality years, respectively?

Sql: select years from child_mortality where Infant_mortality_rate=(select min(Infant_mortality_rate) from child_mortality);

select years from child_mortality where Infant_mortality_rate=(select max(Infant_mortality_rate) from child_mortality);

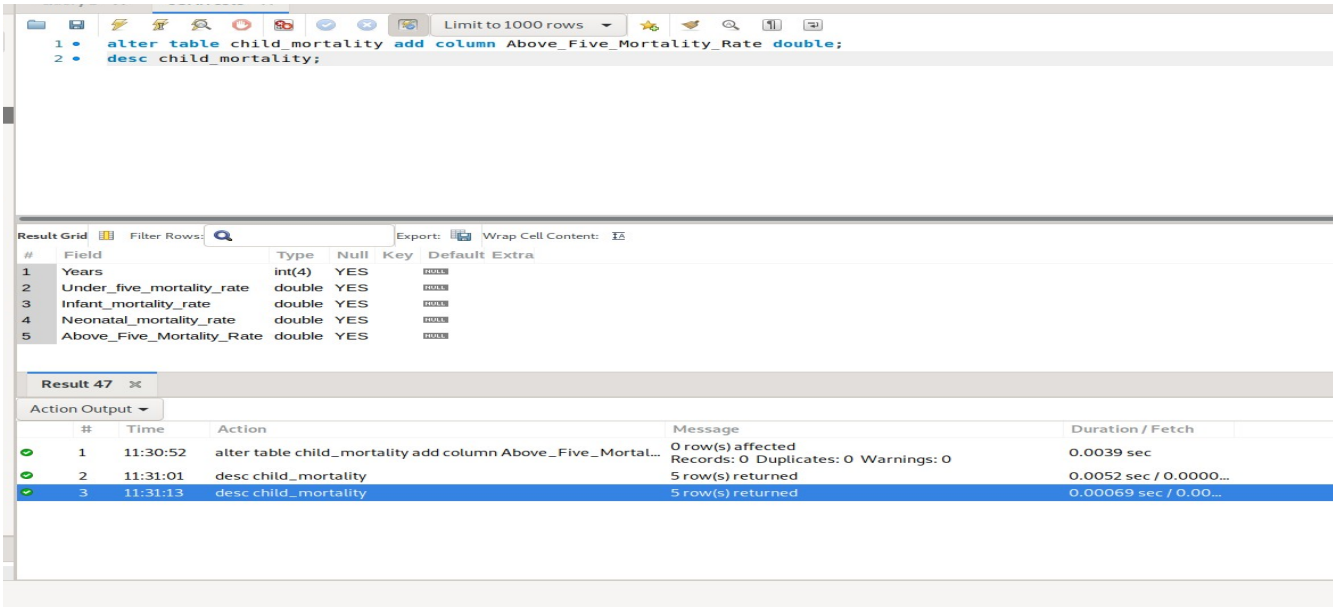



Fig 2.2.9: Add a new column called Above-Five Mortality Rate

sql: alter table child_mortality add column Above_Five_Mortality_Rate double;

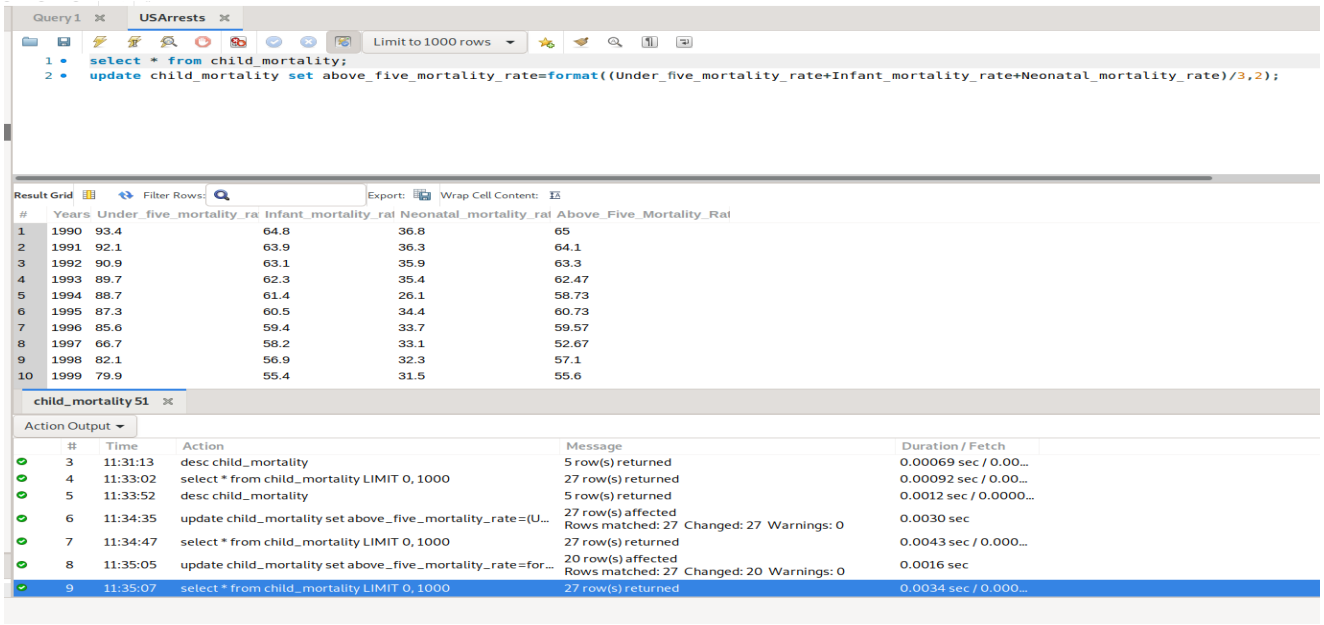



Fig: 2.3.0 Populate Above-Five Mortality Rate with appropriate values.

Sql: update child_mortality set above_five_mortality_rate=format((Under_five_mortality_rate+Infant_mortality_rate+Neonatal_mortality_rate)/3,2);

XML/JSON

Convert the cleansed data set into XML and JSON formats.

Submission Guidelines

• All the data files (CSV, XML, JSON), Excel analysis sheets, and SQL scripts must reside inIDS/HW2folder in your repository

• Summarize your findings for the problems in a PDF report (HW2.pdf) and upload it into Canvas. Thoroughly explain your data cleaning and analysis steps for each problem.

• Include the link to your publicGitHub repositoryon top of your report.