Naresh Kalluri (10755799)
Intro to Data Science (CSC 346 D01 Spring 2022).
Assignment – 1 (Project 2).
Date: 2-15-2022.

GitHub Repository: https://github.com/nk755799/IDS

Project 2:

Problem 1:

The dataset **USArrests.csv** contains 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 (Links to an external site.)

Excel:

Purpose of the Project:

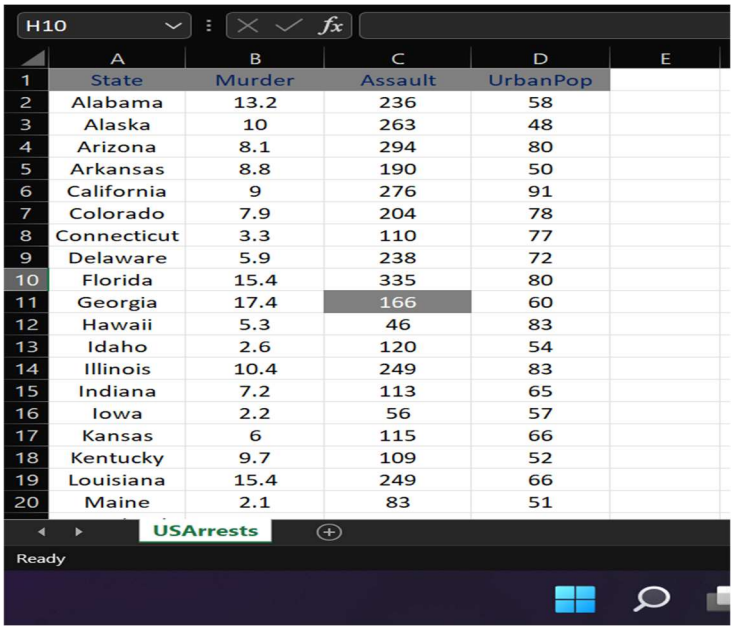
To prepare the dataset for analysis by using the pre-processing Techniques.

Operations:

A. By seeing the data, it can prove that there are some missing values in the data. The one missing value I found on that data is that Georgia doesn't have the Assault value.

By using the Mean(average) I can get that value, so I took the whole column for Mean and I got the value = 166. so, I placed 166 as the Assault value for the state of Georgia.

Result:

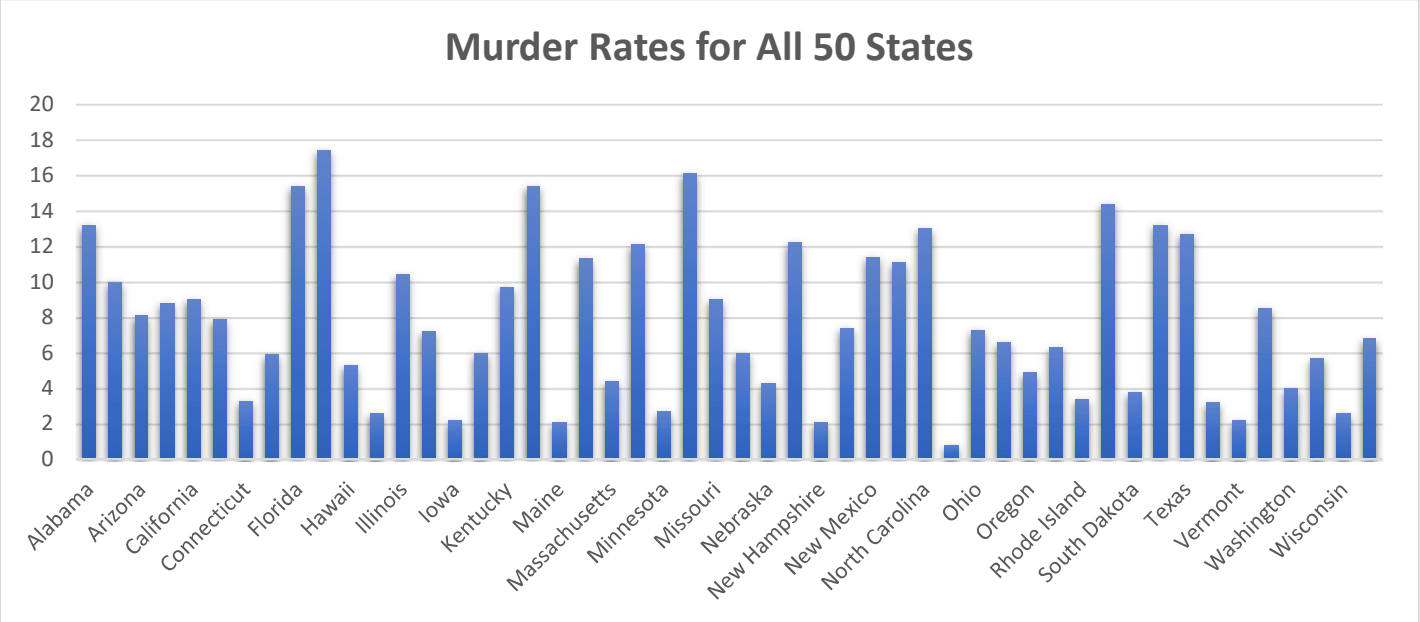


	A	B	C	D	E
1	State	Murder	Assault	UrbanPop	
2	Alabama	13.2	236	58	
3	Alaska	10	263	48	
4	Arizona	8.1	294	80	
5	Arkansas	8.8	190	50	
6	California	9	276	91	
7	Colorado	7.9	204	78	
8	Connecticut	3.3	110	77	
9	Delaware	5.9	238	72	
10	Florida	15.4	335	80	
11	Georgia	17.4	166	60	
12	Hawaii	5.3	46	83	
13	Idaho	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	

B. There is only one missing value is there and I fixed that too.

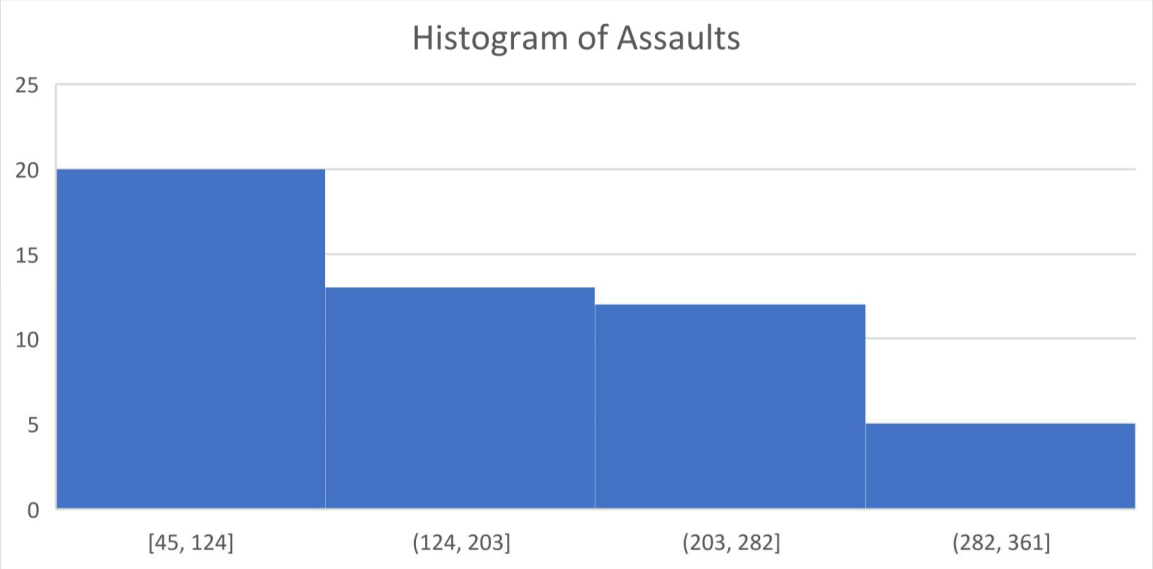
C. Plot Murder rates for all 50 states.

Result: For clearly view Please refer to my GitHub repository HW2 Folder



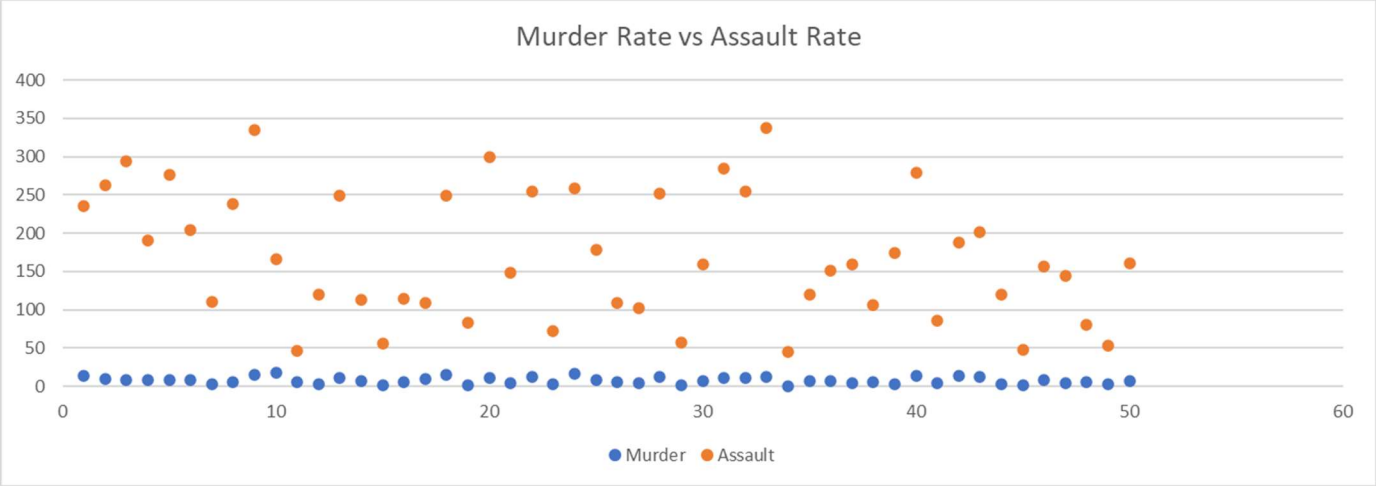
D. Plot histogram of assaults:

Result:



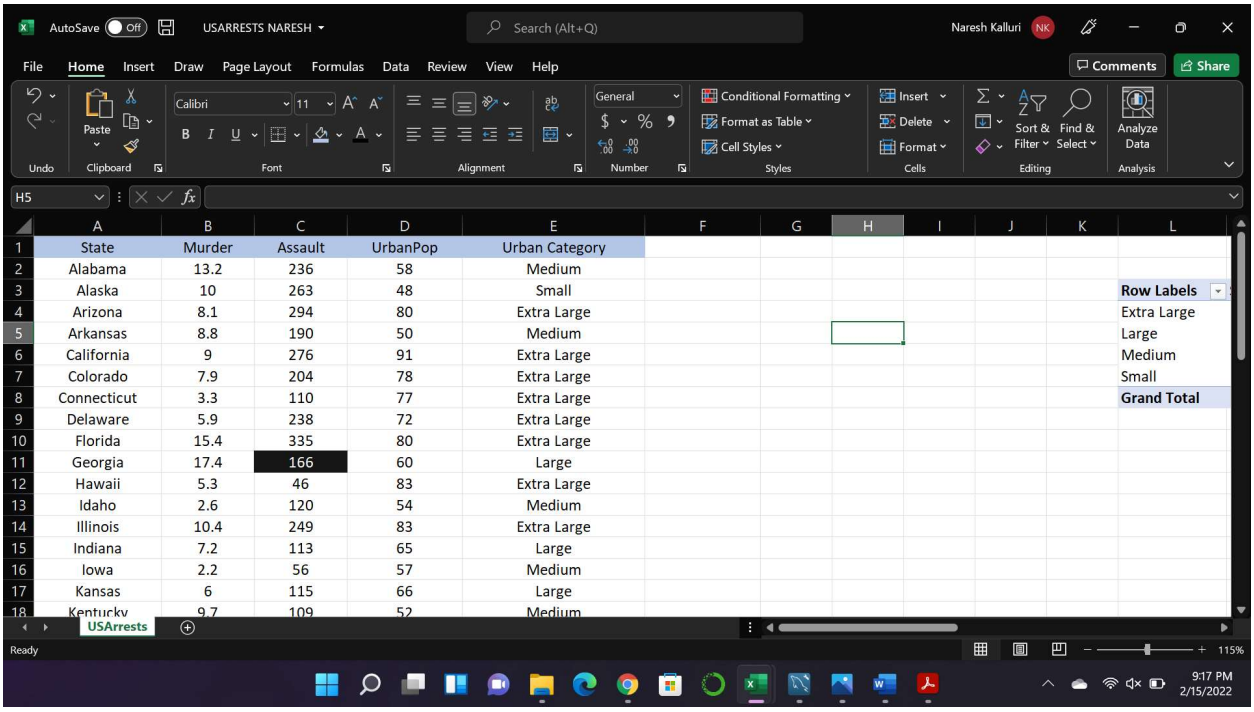
E. Plot Murder rate vs. Assault rate.

Result:



F. Prepare the dataset to establish a relationship between an urban population category and a crime type.

I converted the urban population into categories like Small, Medium, Large and Extra Large and got a relation.



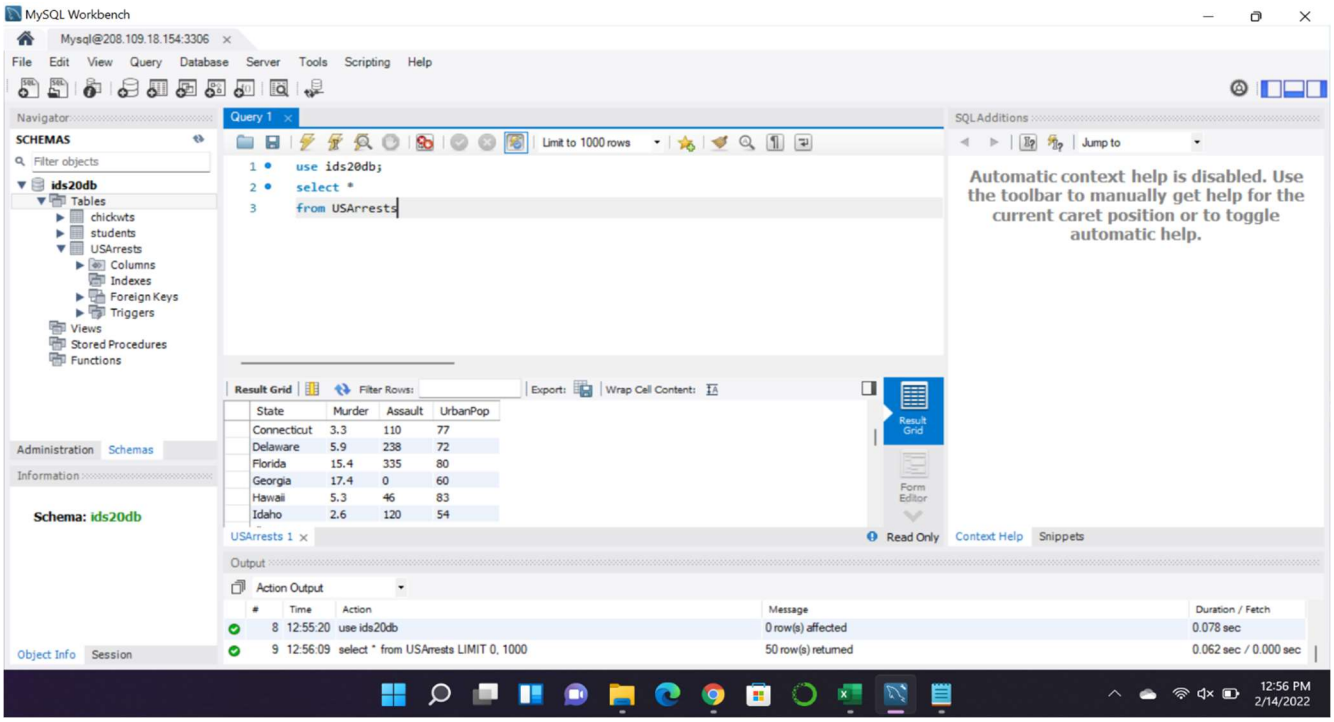
	A	B	C	D	E
1	State	Murder	Assault	UrbanPop	Urban Category
2	Alabama	13.2	236	58	Medium
3	Alaska	10	263	48	Small
4	Arizona	8.1	294	80	Extra Large
5	Arkansas	8.8	190	50	Medium
6	California	9	276	91	Extra Large
7	Colorado	7.9	204	78	Extra Large
8	Connecticut	3.3	110	77	Extra Large
9	Delaware	5.9	238	72	Extra Large
10	Florida	15.4	335	80	Extra Large
11	Georgia	17.4	166	60	Large
12	Hawaii	5.3	46	83	Extra Large
13	Idaho	2.6	120	54	Medium
14	Illinois	10.4	249	83	Extra Large
15	Indiana	7.2	113	65	Large
16	Iowa	2.2	56	57	Medium
17	Kansas	6	115	66	Large
18	Kentucky	9.7	109	52	Medium

Row Labels	Sum of Murder	Sum of Assault	Count of UrbanPop
Extra Large	169.8	4150	21
Large	93.7	1797	12
Medium	59.9	1148	9
Small	66	1398	8
Grand Total	**389.4**	**8493**	**50**

MySQL:

A. Import the original CSV file into MySQL and create the table USArrests.

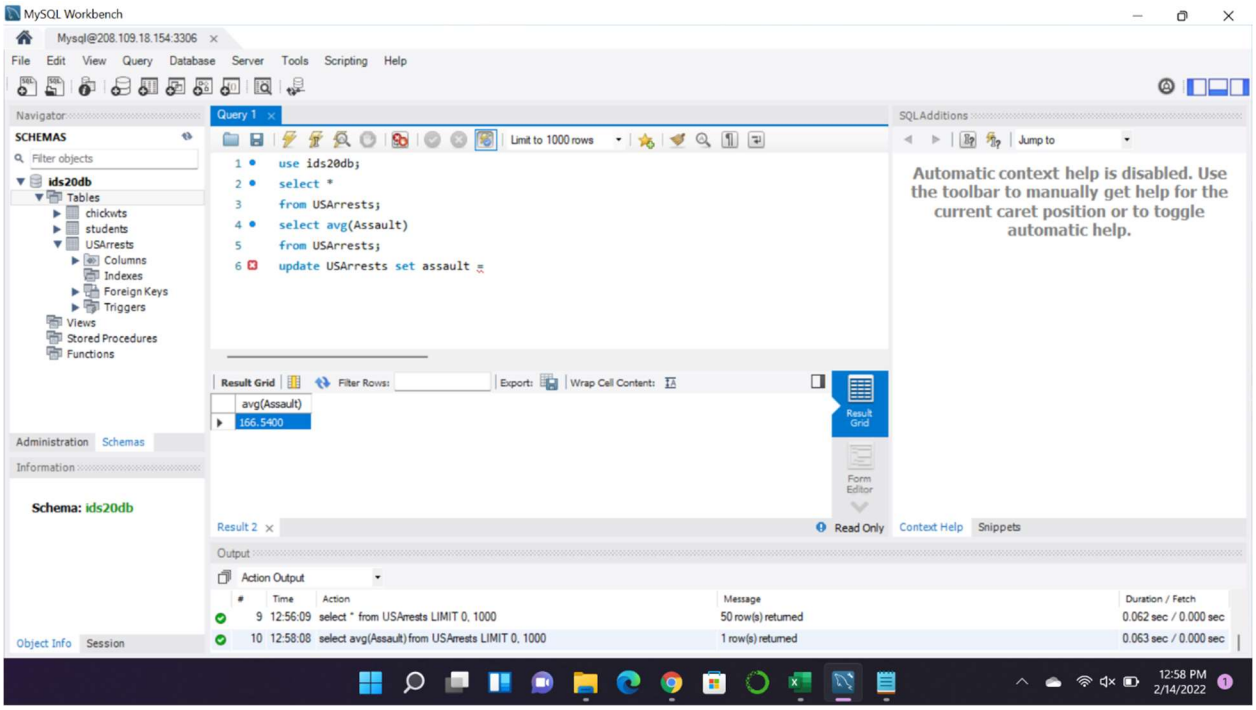
Result: Imported the original CSV file into MySQL workspace and created the table USArrests.



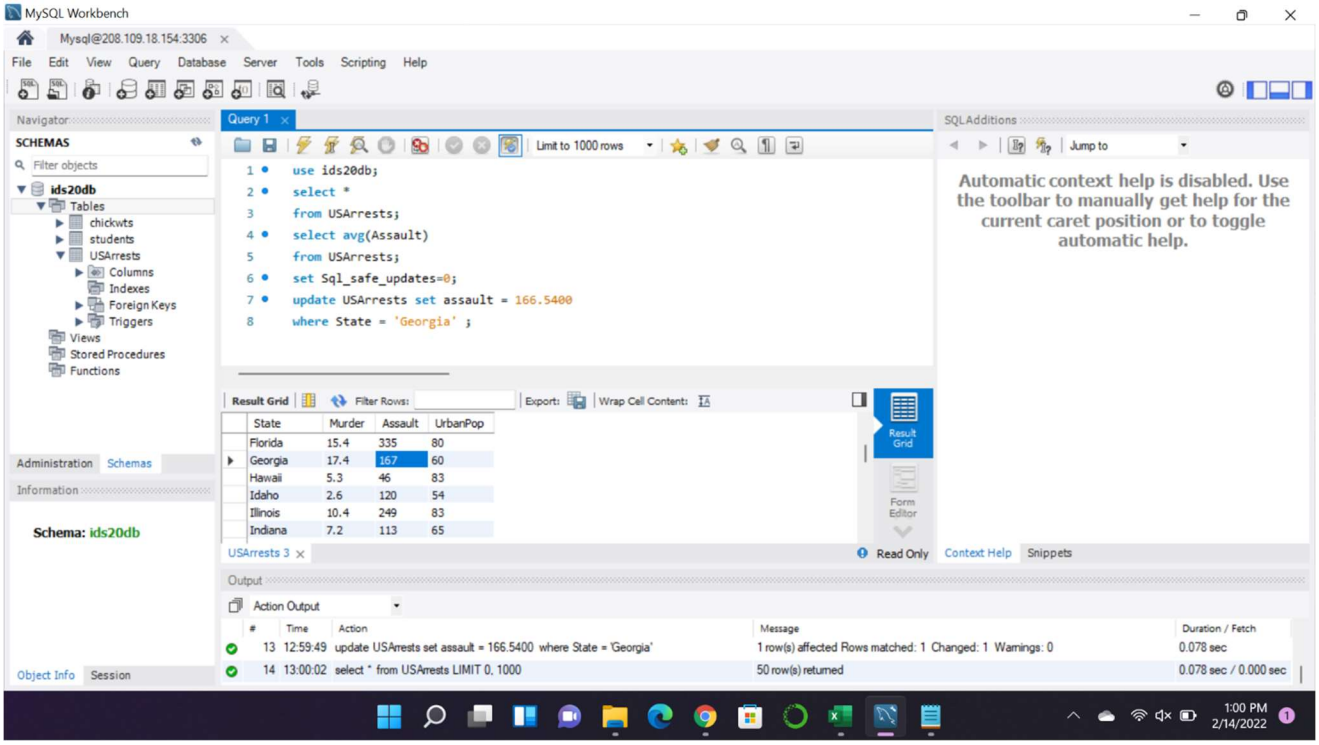
B. Use SQL to replace all missing values in a column by the average. Hint: Use the Update command.

Result:

First I took the average method to get the value of the missing Assault value of the state of Georgia.

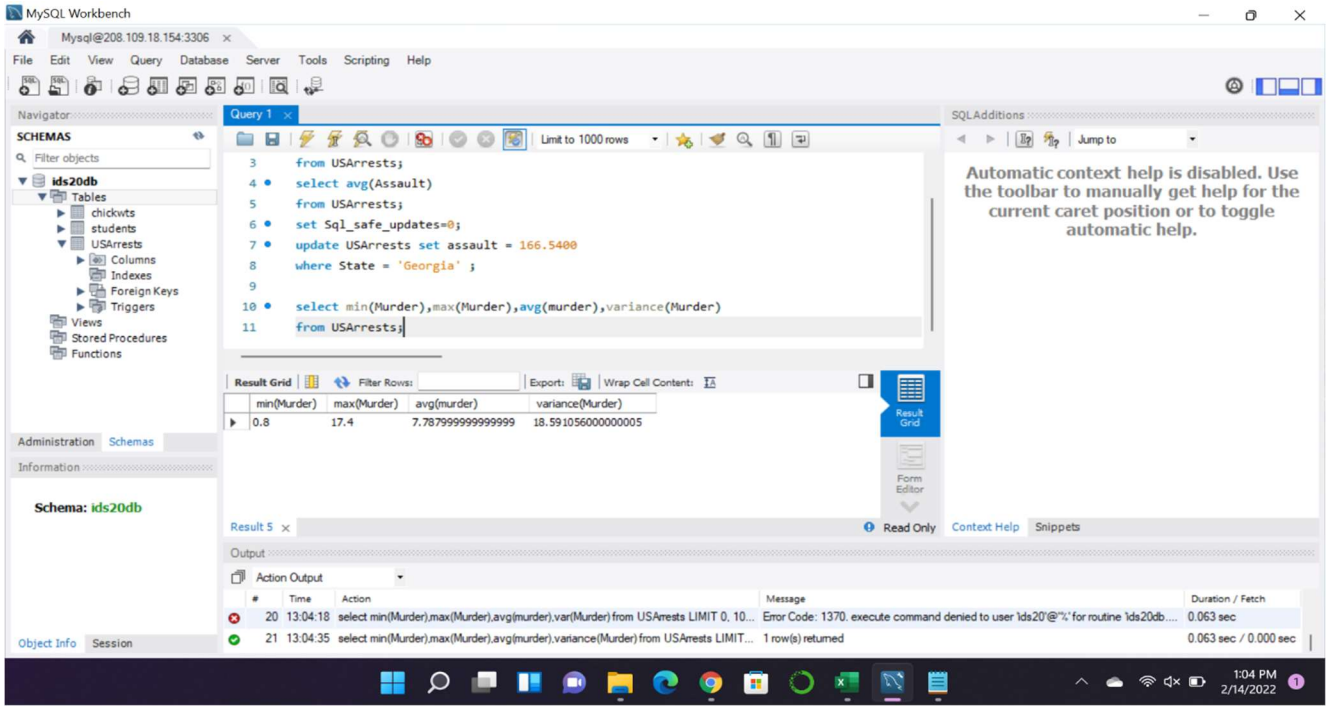


Next used the Update command to update the Georgia Assault Value with an average of column ’166.54’

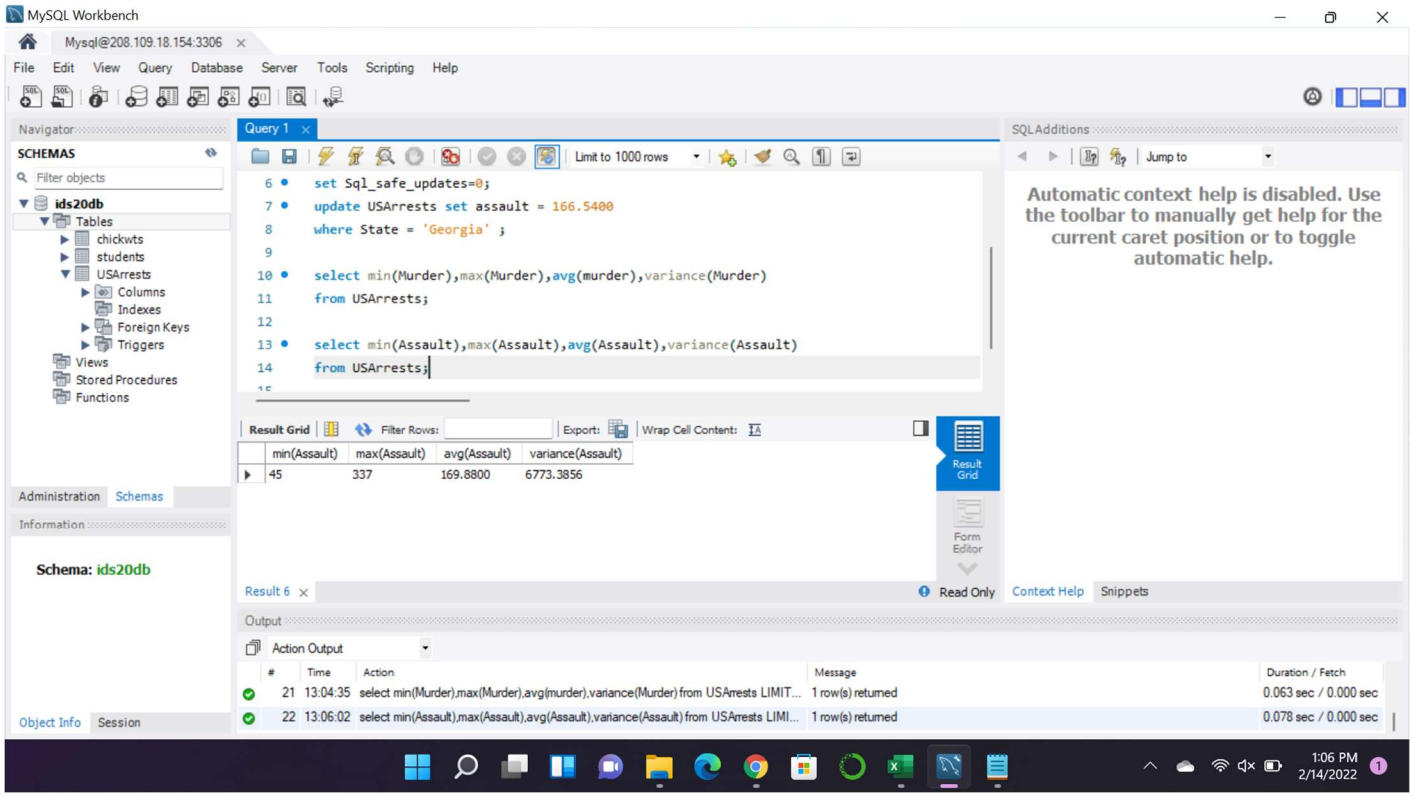


C. Find min, max, mean, and variance of all numeric attributes in SQL.

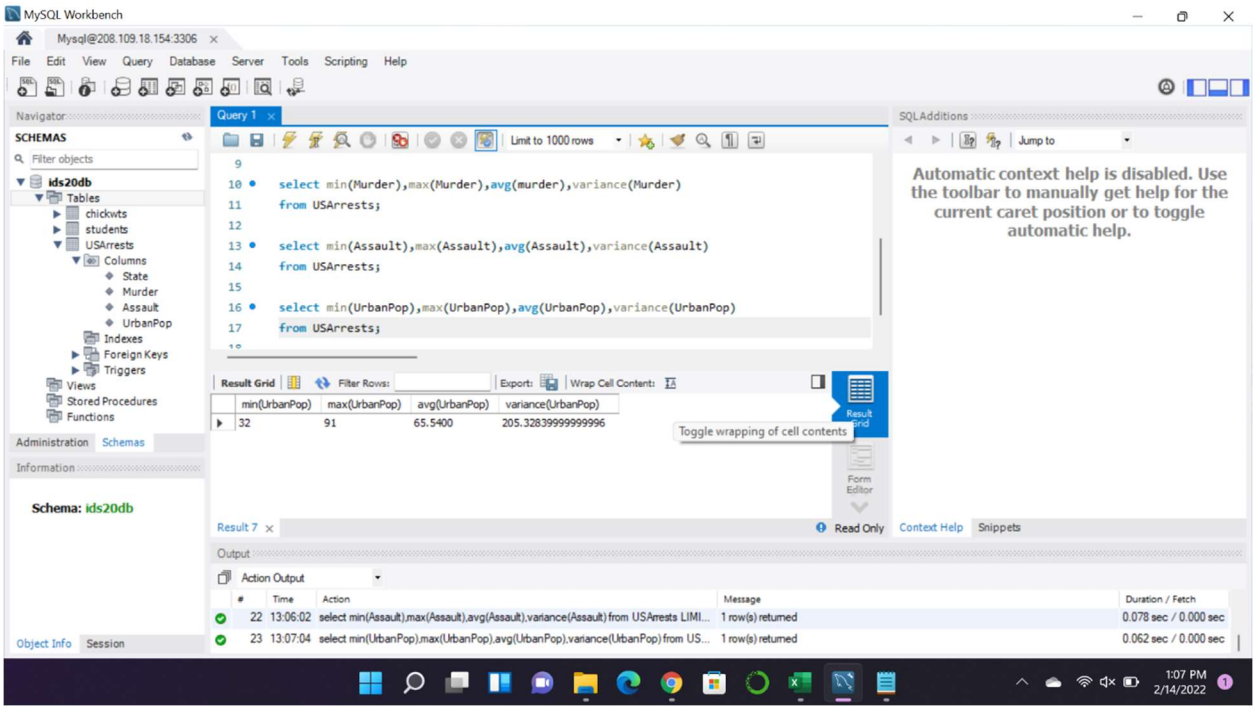
Result: For Murder:



For Assaults:

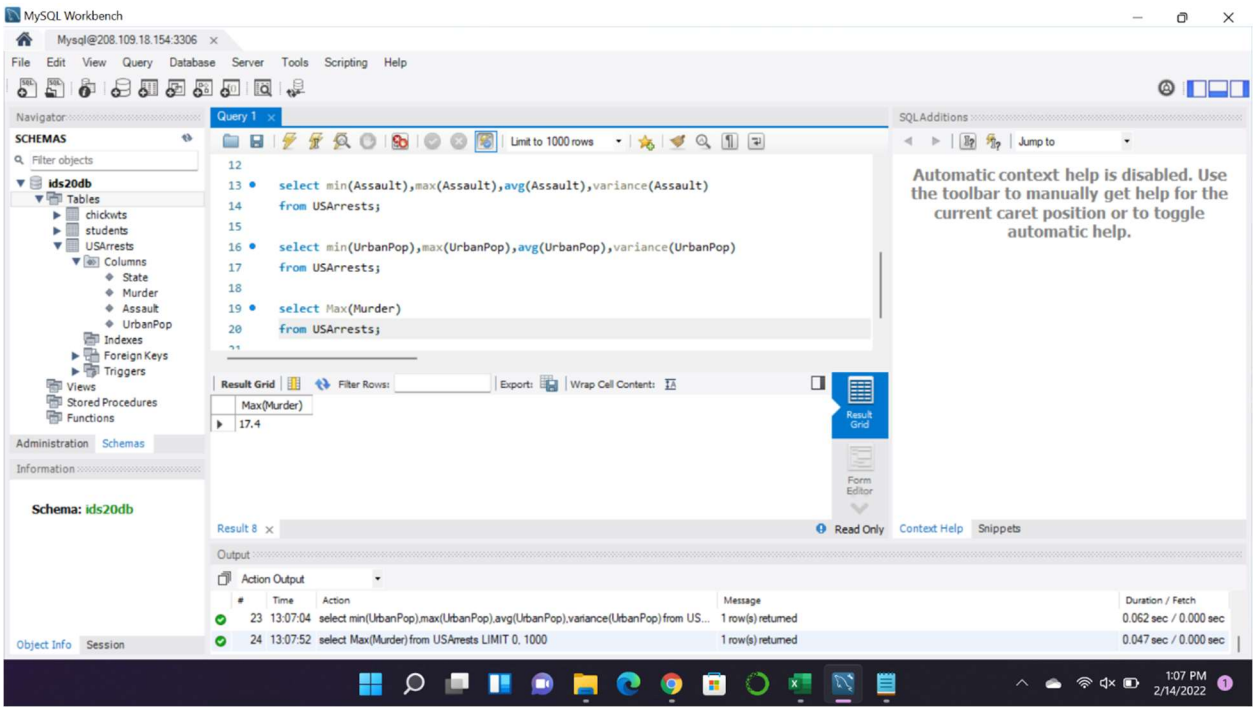


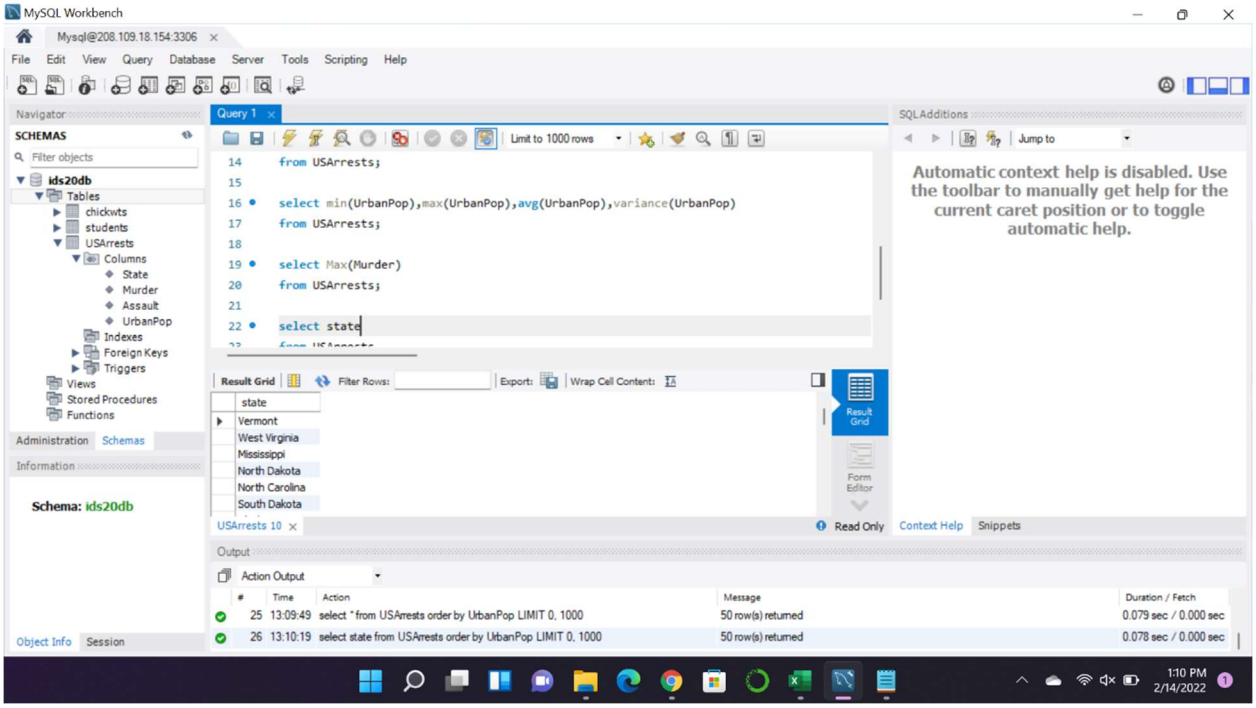
For Urban Population

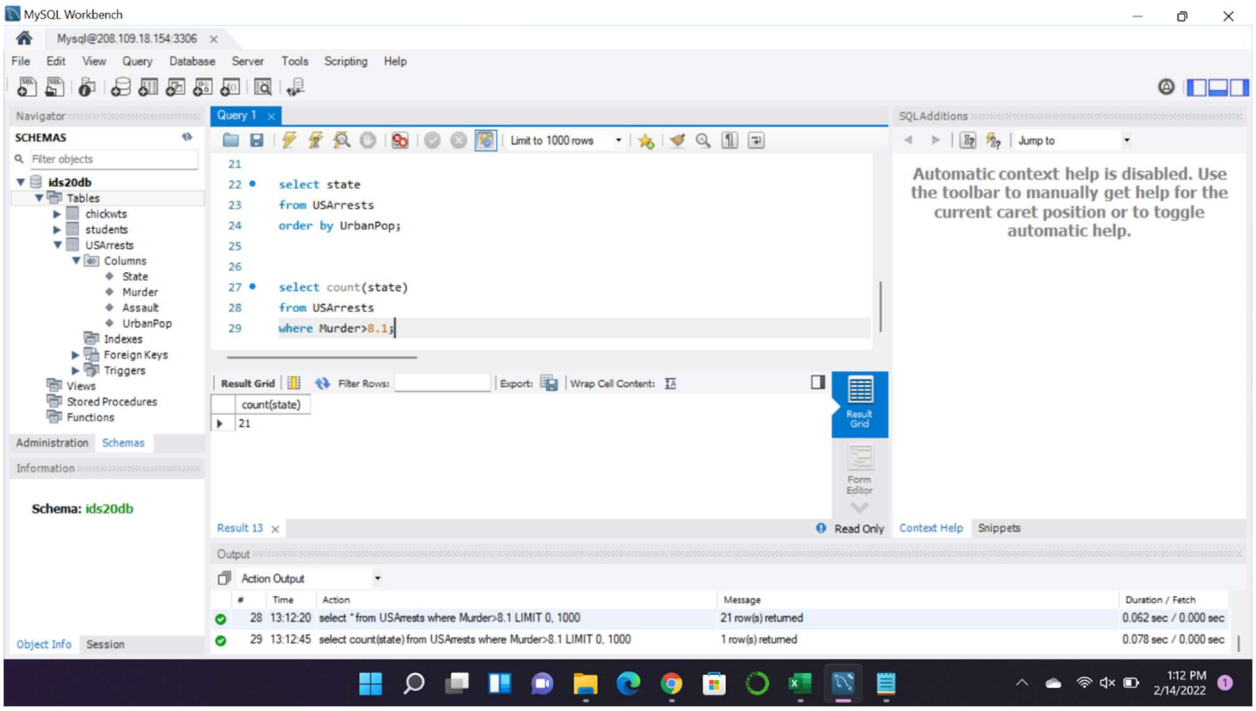


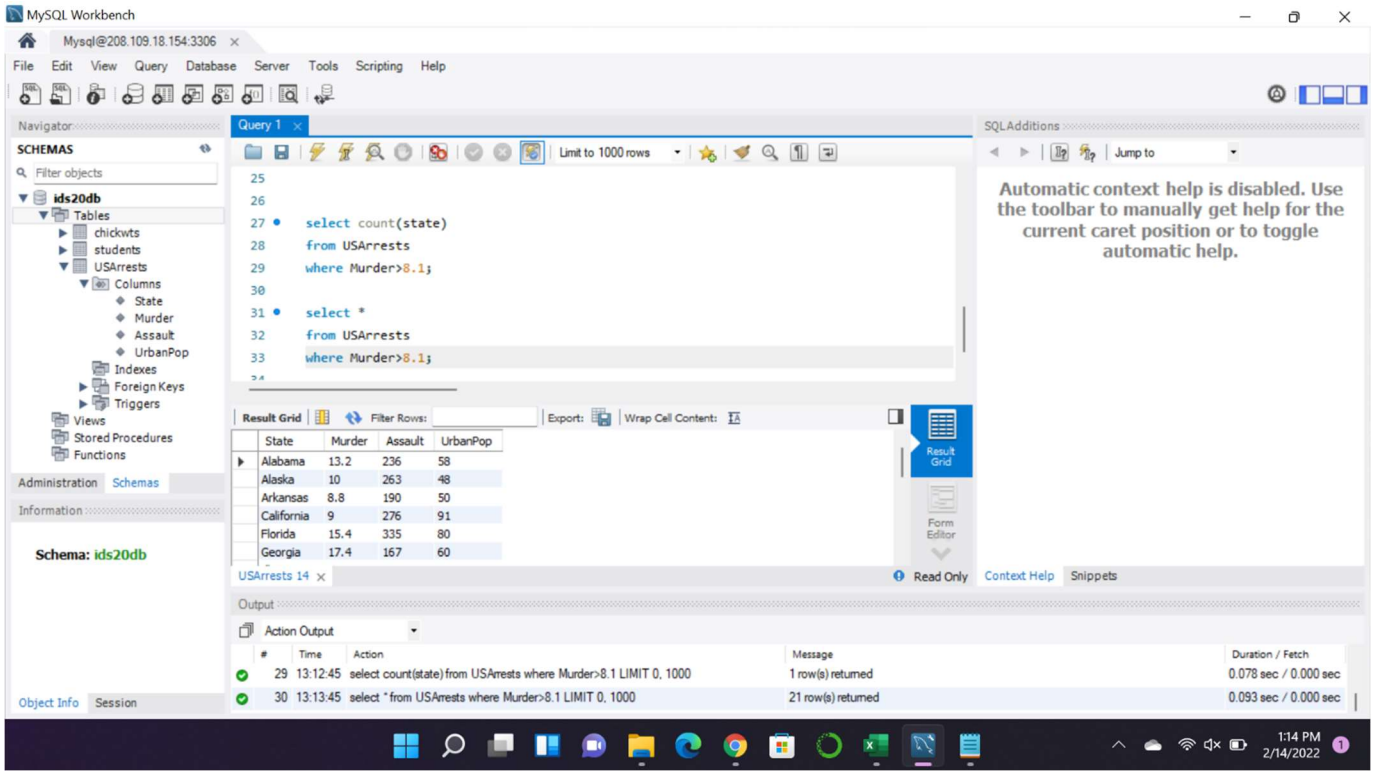
D. Which state has the maximum murder rate?

Result:




MySQL Workbench
Mysql@208.109.18.154:3306
File Edit View Query Database Server Tools Scripting Help
Navigator
SCHEMAS
Filter objects
ids20db
Tables
chickwts
students
USArrests
Columns
State
Murder
Assault
UrbanPop
Indexes
Foreign Keys
Triggers
Views
Stored Procedures
Functions
Administration Schemas
Information
Schema: ids20db
Object Info Session
Query 1
Limit to 1000 rows
14 from USArrests;
15
16 select min(UrbanPop),max(UrbanPop),avg(UrbanPop),variance(UrbanPop)
17 from USArrests;
18
19 select Max(Murder)
20 from USArrests;
21
22 select state
Result Grid Filter Rows: Export: Wrap Cell Content:
state
Vermont
West Virginia
Mississippi
North Dakota
North Carolina
South Dakota
USArrests 10
Read Only
SQLAdditions
Jump to
Automatic context help is disabled. Use the toolbar to manually get help for the current caret position or to toggle automatic help.
Context Help Snippets
Output
Action Output
Time Action Message Duration / Fetch
25 13:09:49 select * from USArrests order by UrbanPop LIMIT 0, 1000 50 row(s) returned 0.079 sec / 0.000 sec
26 13:10:19 select state from USArrests order by UrbanPop LIMIT 0, 1000 50 row(s) returned 0.078 sec / 0.000 sec
1:10 PM
2/14/2022


MySQL Workbench
Mysql@208.109.18.154:3306
File Edit View Query Database Server Tools Scripting Help
Navigator
SCHEMAS
Filter objects
ids20db
Tables
chickwts
students
USArrests
Columns
State
Murder
Assault
UrbanPop
Indexes
Foreign Keys
Triggers
Views
Stored Procedures
Functions
Administration Schemas
Information
Schema: ids20db
Object Info Session
Query 1
Limit to 1000 rows
21
22 select state
23 from USArrests
24 order by UrbanPop;
25
26
27 select count(state)
28 from USArrests
29 where Murder>8.1;
Result Grid Filter Rows: Export: Wrap Cell Content:
count(state)
21
Result 13
Read Only
SQLAdditions
Jump to
Automatic context help is disabled. Use the toolbar to manually get help for the current caret position or to toggle automatic help.
Context Help Snippets
Output
Action Output
Time Action Message Duration / Fetch
28 13:12:20 select * from USArrests where Murder>8.1 LIMIT 0, 1000 21 row(s) returned 0.062 sec / 0.000 sec
29 13:12:45 select count(state) from USArrests where Murder>8.1 LIMIT 0, 1000 1 row(s) returned 0.078 sec / 0.000 sec
1:12 PM
2/14/2022


MySQL Workbench
Mysql@208.109.18.154:3306
File Edit View Query Database Server Tools Scripting Help
Navigator
SCHEMAS
Filter objects
ids20db
Tables
chickwts
students
USArrests
Columns
State
Murder
Assault
UrbanPop
Indexes
Foreign Keys
Triggers
Views
Stored Procedures
Functions
Administration Schemas
Information
Schema: ids20db
Object Info Session
Query 1
Limit to 1000 rows
25
26
27 select count(state)
28 from USArrests
29 where Murder>8.1;
30
31 select *
32 from USArrests
33 where Murder>8.1;
Result Grid Filter Rows: Export: Wrap Cell Content:
State Murder Assault UrbanPop
Alabama 13.2 236 58
Alaska 10 263 48
Arkansas 8.8 190 50
California 9 276 91
Florida 15.4 335 80
Georgia 17.4 167 60
Result Grid
Form Editor
USArrests 14
Read Only
SQLAdditions
Jump to
Automatic context help is disabled. Use the toolbar to manually get help for the current caret position or to toggle automatic help.
Context Help Snippets
Output
Action Output
Time Action Message Duration / Fetch
29 13:12:45 select count(state) from USArrests where Murder>8.1 LIMIT 0, 1000 1 row(s) returned 0.078 sec / 0.000 sec
30 13:13:45 select * from USArrests where Murder>8.1 LIMIT 0, 1000 21 row(s) returned 0.093 sec / 0.000 sec
1:14 PM
2/14/2022

Problem 2:

Excel:

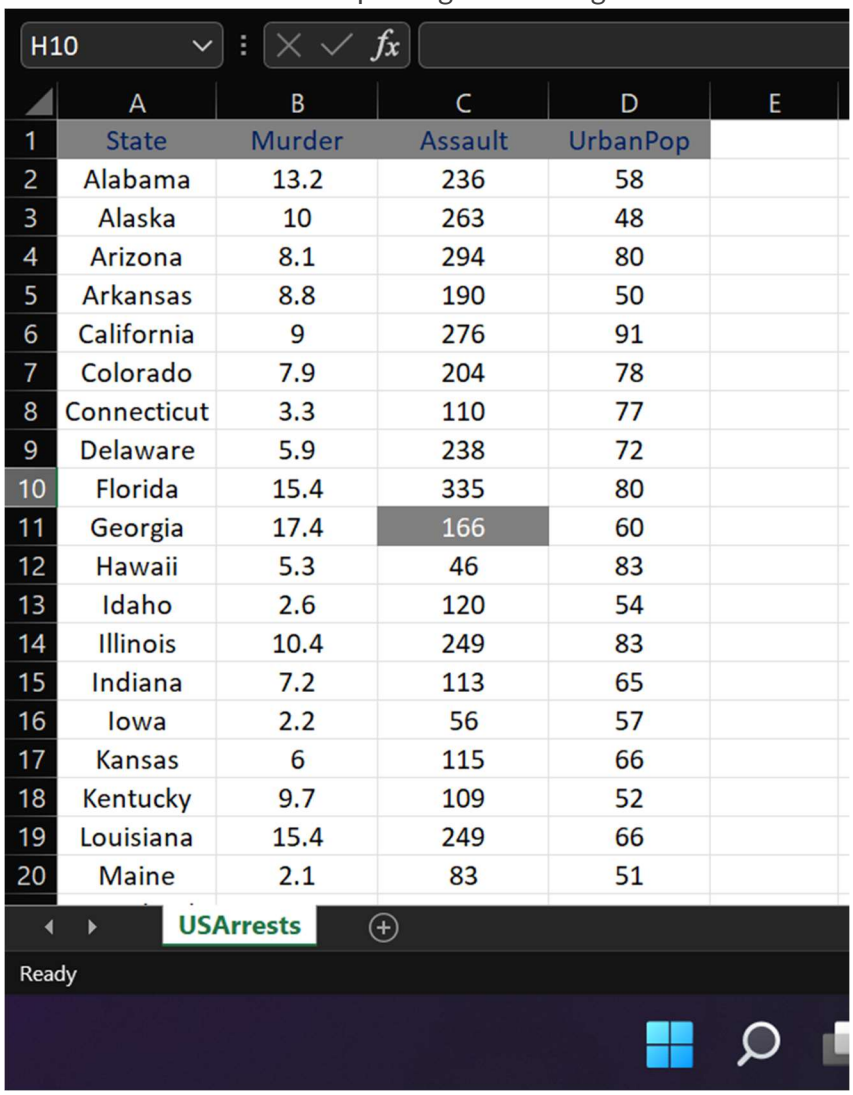
Purpose of the Project:

To prepare the dataset for analysis by using the pre-processing Techniques.

Operations:

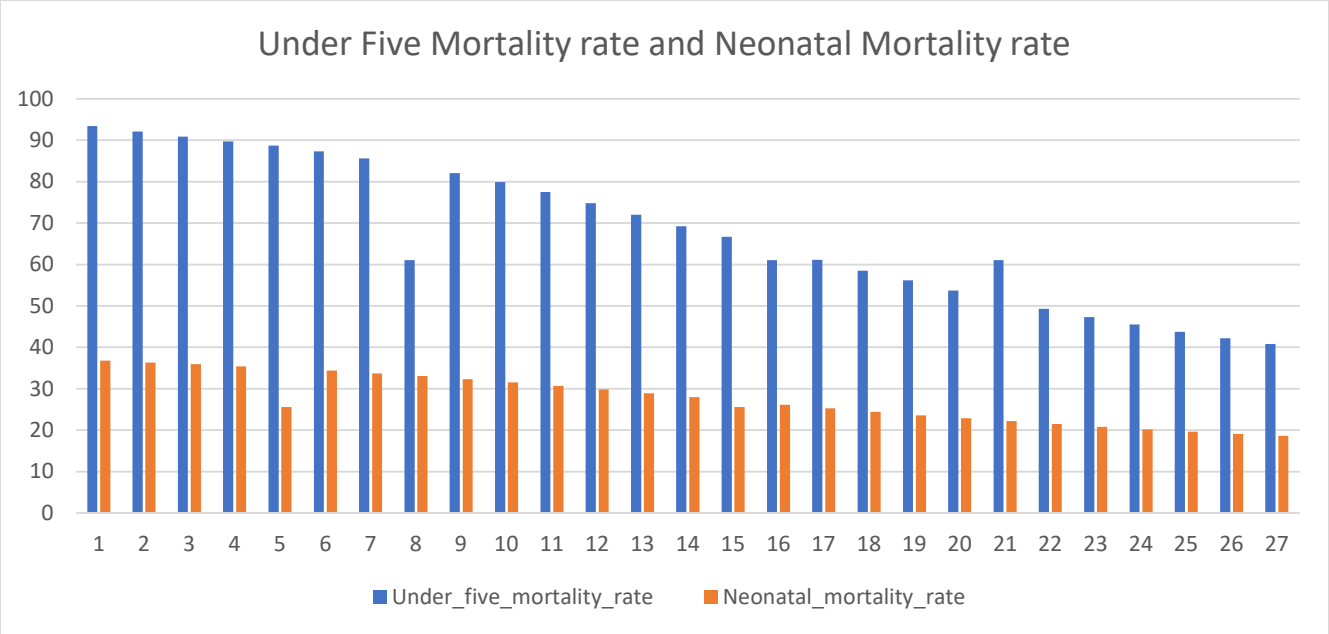
- First took the Child mortality rate file in the professor Github Repository https://github.com/bforoura/IDS/tree/main/HW2.
- Imported the data into the Excel worksheet but I found several missing values in the data, and I applied the average method for every respective column to get the Mean values and I placed those values in place of the missing value for their respective columns.

  Below are the result after updating the missing values of the data

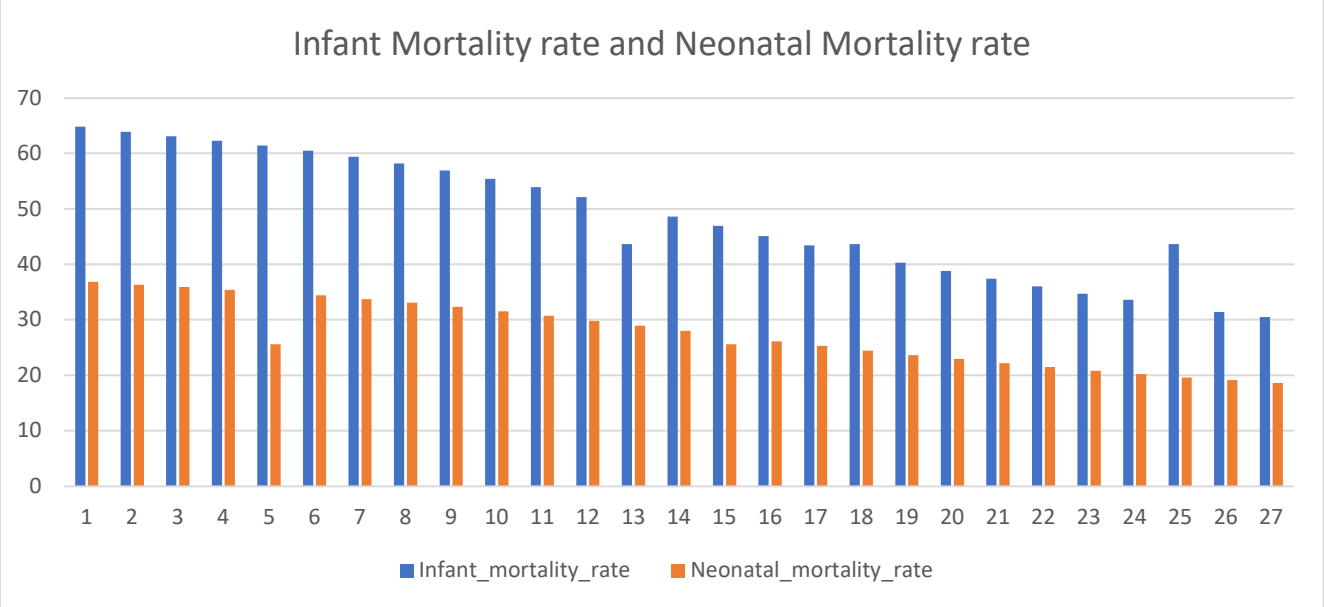


	A	B	C	D	E
1	State	Murder	Assault	UrbanPop	
2	Alabama	13.2	236	58	
3	Alaska	10	263	48	
4	Arizona	8.1	294	80	
5	Arkansas	8.8	190	50	
6	California	9	276	91	
7	Colorado	7.9	204	78	
8	Connecticut	3.3	110	77	
9	Delaware	5.9	238	72	
10	Florida	15.4	335	80	
11	Georgia	17.4	166	60	
12	Hawaii	5.3	46	83	
13	Idaho	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	

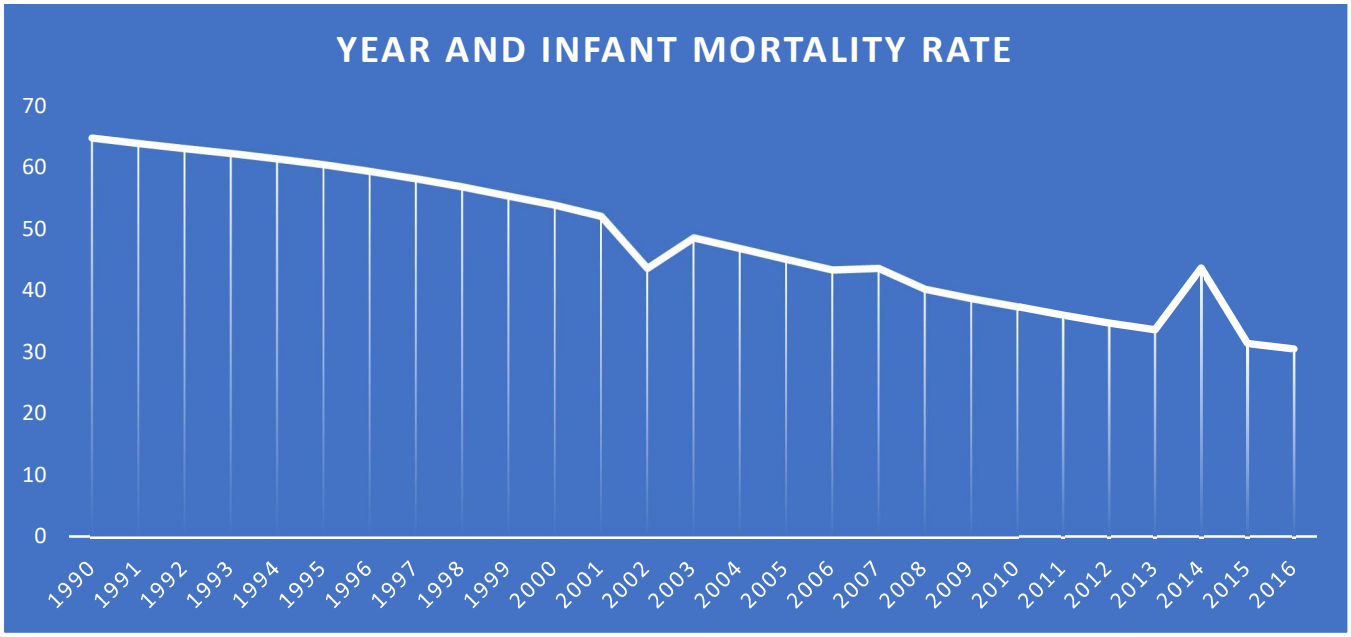
B. Under-five mortality rate and neonatal mortality rate.



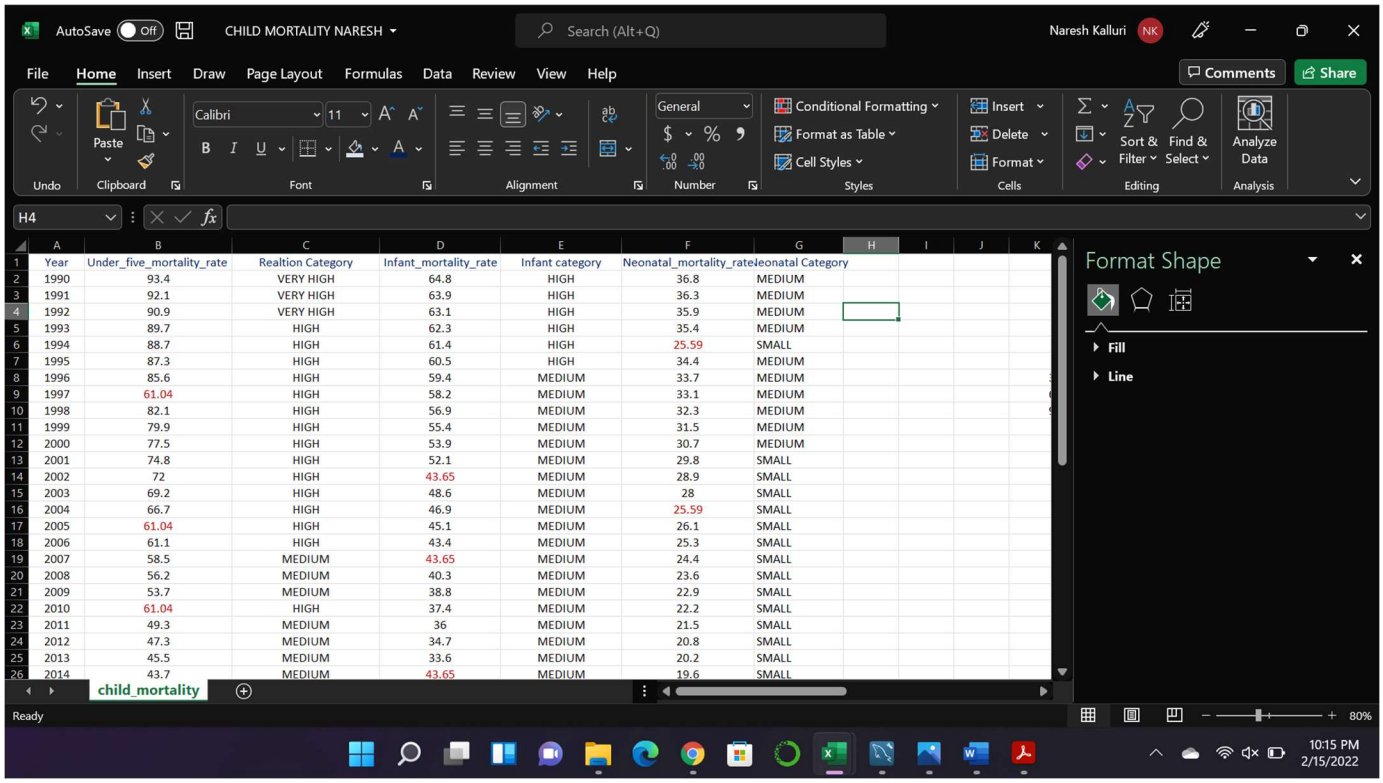
Infant mortality rate and neonatal mortality rate.



Year and infant mortality rate.



Categoric relation between every attribute in the data.

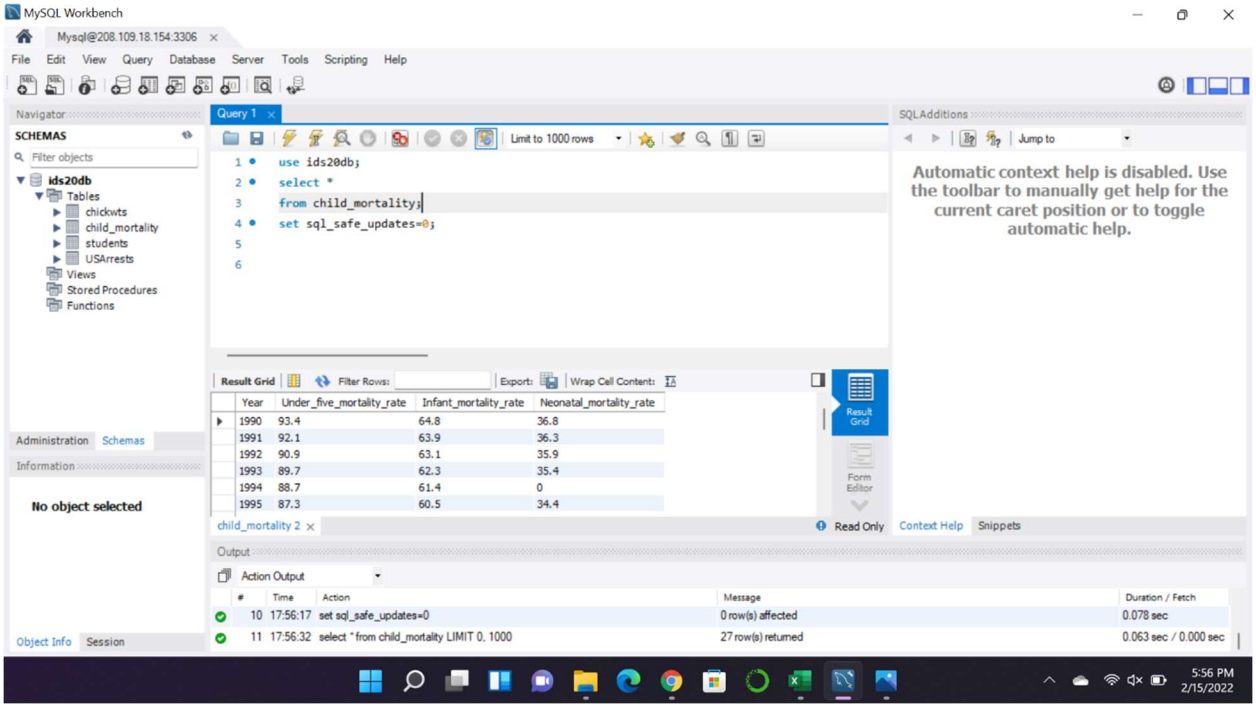


	A	B	C	D	E	F	G
1	Year	Under_five_mortality_rate	Realtion Category	Infant_mortality_rate	Infant category	Neonatal_mortality_rate	Neonatal Category
2	1990	93.4	VERY HIGH	64.8	HIGH	36.8	MEDIUM
3	1991	92.1	VERY HIGH	63.9	HIGH	36.3	MEDIUM
4	1992	90.9	VERY HIGH	63.1	HIGH	35.9	MEDIUM
5	1993	89.7	HIGH	62.3	HIGH	35.4	MEDIUM
6	1994	88.7	HIGH	61.4	HIGH	25.59	SMALL
7	1995	87.3	HIGH	60.5	HIGH	34.4	MEDIUM
8	1996	85.6	HIGH	59.4	MEDIUM	33.7	MEDIUM
9	1997	61.04	HIGH	58.2	MEDIUM	33.1	MEDIUM
10	1998	82.1	HIGH	56.9	MEDIUM	32.3	MEDIUM
11	1999	79.9	HIGH	55.4	MEDIUM	31.5	MEDIUM
12	2000	77.5	HIGH	53.9	MEDIUM	30.7	MEDIUM
13	2001	74.8	HIGH	52.1	MEDIUM	29.8	SMALL
14	2002	72	HIGH	43.65	MEDIUM	28.9	SMALL
15	2003	69.2	HIGH	48.6	MEDIUM	28	SMALL
16	2004	66.7	HIGH	46.9	MEDIUM	25.59	SMALL
17	2005	61.04	HIGH	45.1	MEDIUM	26.1	SMALL
18	2006	61.1	HIGH	43.4	MEDIUM	25.3	SMALL
19	2007	58.5	MEDIUM	43.65	MEDIUM	24.4	SMALL
20	2008	56.2	MEDIUM	40.3	MEDIUM	23.6	SMALL
21	2009	53.7	MEDIUM	38.8	MEDIUM	22.9	SMALL
22	2010	61.04	HIGH	37.4	MEDIUM	22.2	SMALL
23	2011	49.3	MEDIUM	36	MEDIUM	21.5	SMALL
24	2012	47.3	MEDIUM	34.7	MEDIUM	20.8	SMALL
25	2013	45.5	MEDIUM	33.6	MEDIUM	20.2	SMALL
26	2014	43.7	MEDIUM	43.65	MEDIUM	19.6	SMALL

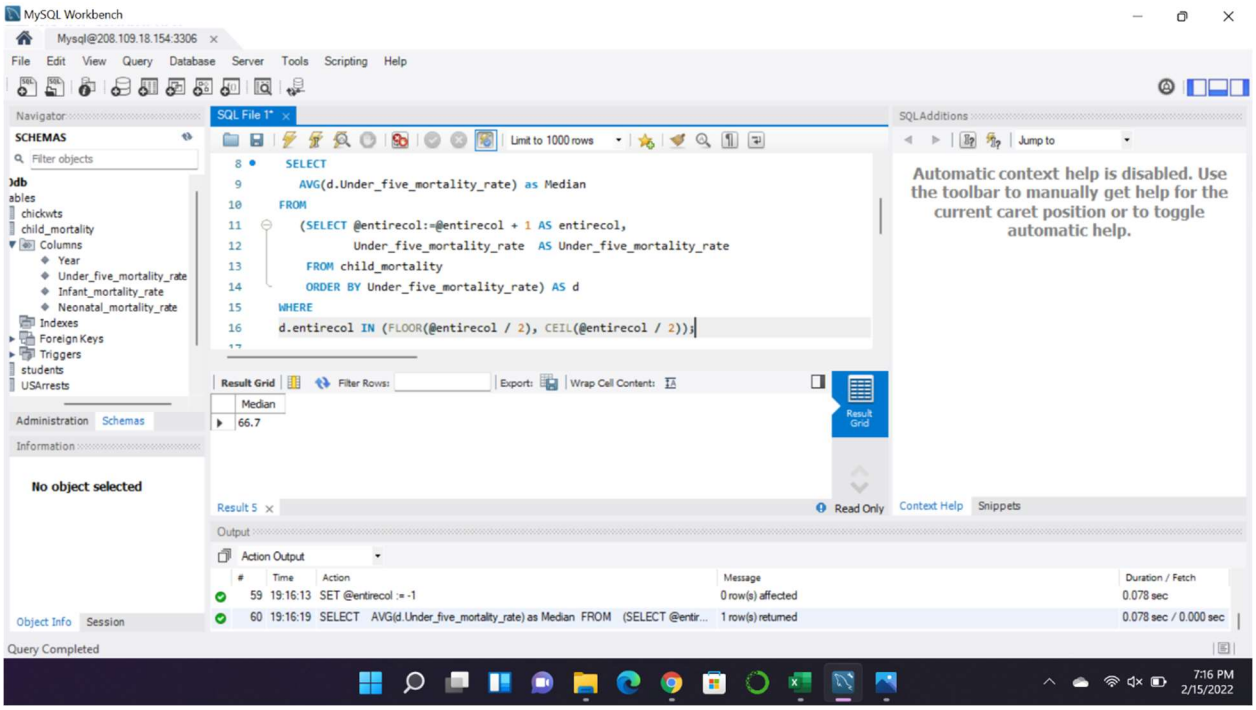
MySQL:

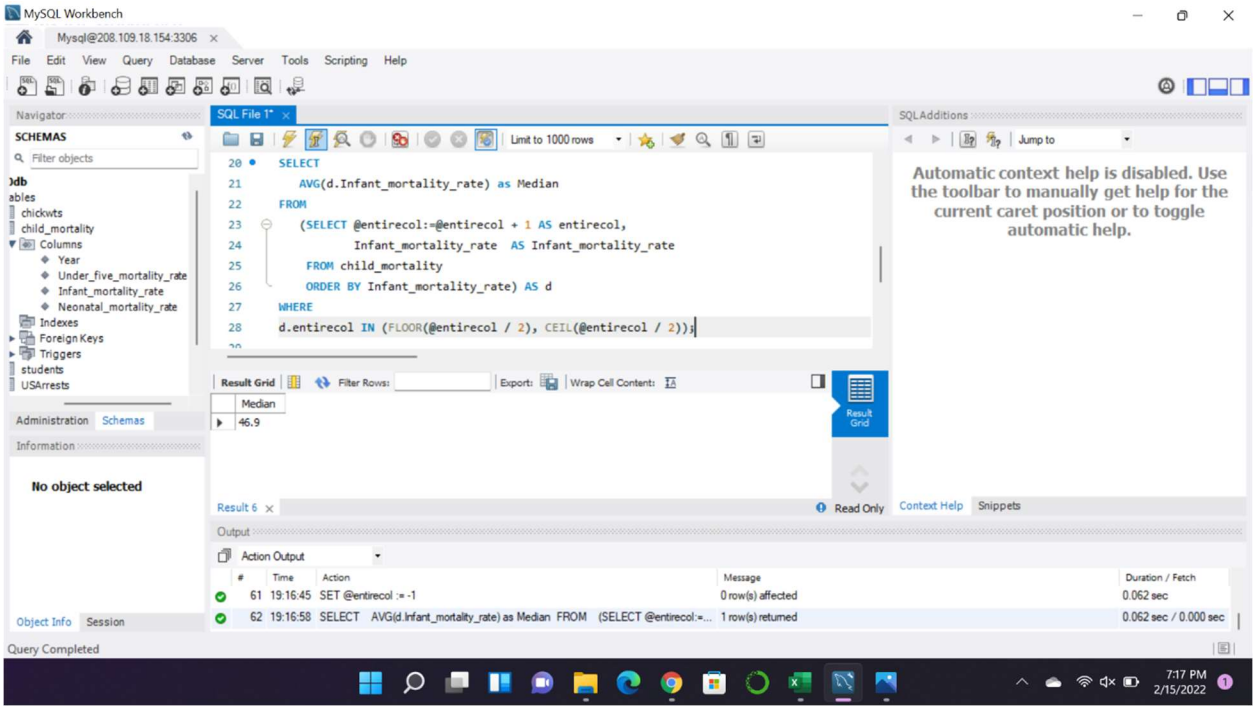
A. Import the original data set into MySQL:

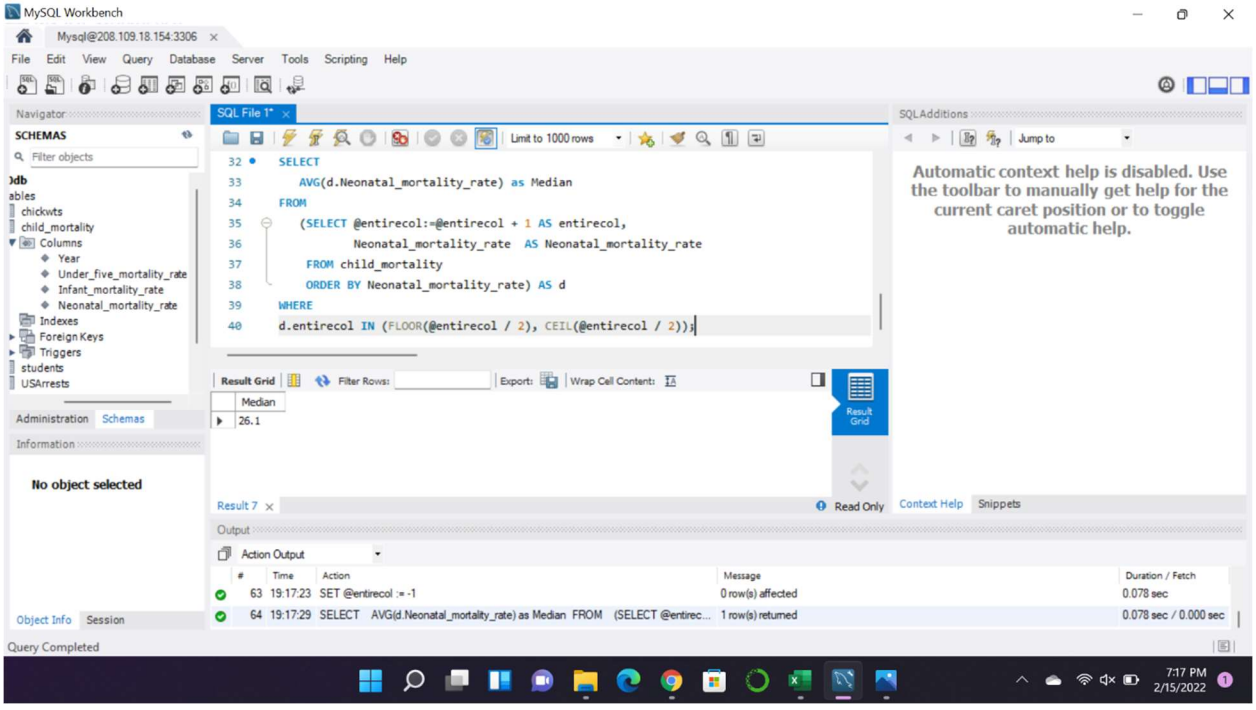
Result: Imported the original data set into the MySQL workspace successfully.



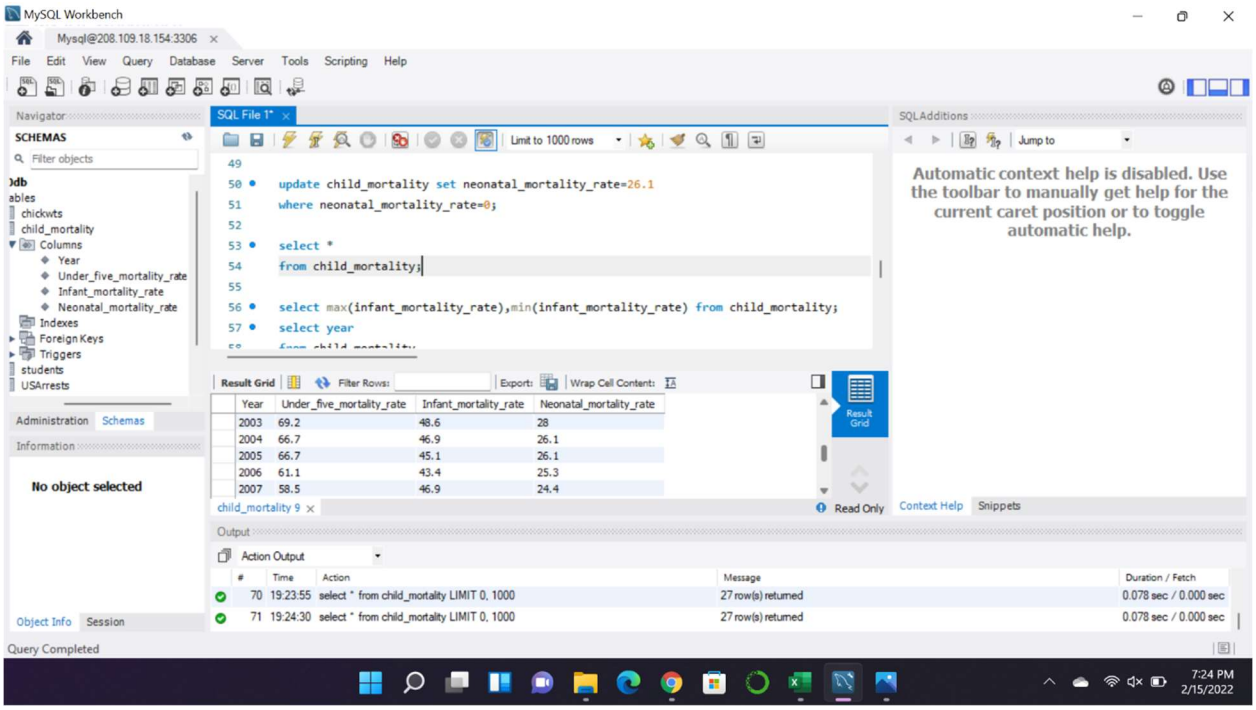
B. Use SQL to fill in the missing values in each column using the medians.



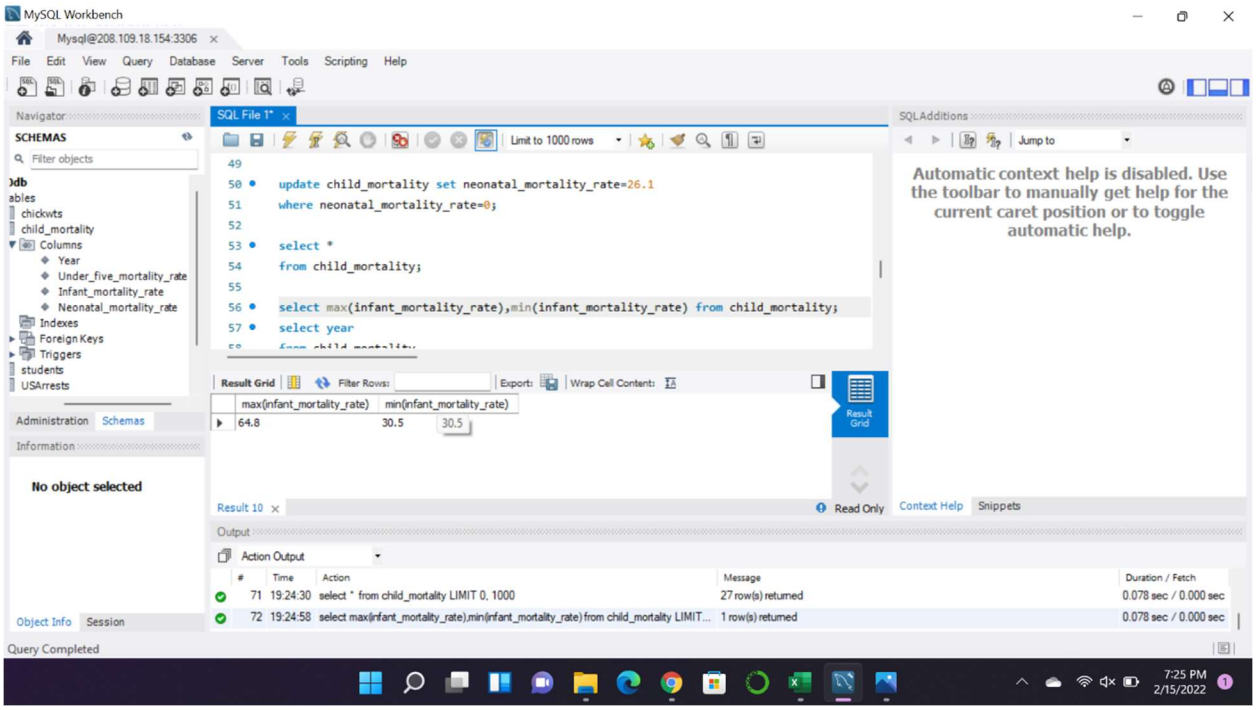

MySQL Workbench
Mysql@208.109.18.154:3306
File Edit View Query Database Server Tools Scripting Help
Navigator
SCHEMAS
Filter objects
db
ables
chickwts
child_mortality
Columns
Year
Under_five_mortality_rate
Infant_mortality_rate
Neonatal_mortality_rate
Indexes
Foreign Keys
Triggers
students
USArrests
Administration Schemas
Information
No object selected
Object Info Session
SQL File 1*
Limit to 1000 rows
20 SELECT
21 AVG(d.Infant_mortality_rate) as Median
22 FROM
23 (SELECT @entirecol:=@entirecol + 1 AS entirecol,
24 Infant_mortality_rate AS Infant_mortality_rate
25 FROM child_mortality
26 ORDER BY Infant_mortality_rate) AS d
27 WHERE
28 d.entirecol IN (FLOOR(@entirecol / 2), CEIL(@entirecol / 2));
Result Grid
Filter Rows:
Export:
Wrap Cell Content:
Median
46.9
Result 6
Read Only
SQLAdditions
Jump to
Automatic context help is disabled. Use the toolbar to manually get help for the current caret position or to toggle automatic help.
Context Help Snippets
Output
Action Output
Time Action Message Duration / Fetch
61 19:16:45 SET @entirecol := -1 0 row(s) affected 0.062 sec
62 19:16:58 SELECT AVG(d.Infant_mortality_rate) as Median FROM (SELECT @entirecol:=... 1 row(s) returned 0.062 sec / 0.000 sec
Query Completed
7:17 PM
2/15/2022


MySQL Workbench
Mysql@208.109.18.154:3306
File Edit View Query Database Server Tools Scripting Help
Navigator
SCHEMAS
Filter objects
db
ables
chickwts
child_mortality
Columns
Year
Under_five_mortality_rate
Infant_mortality_rate
Neonatal_mortality_rate
Indexes
Foreign Keys
Triggers
students
USArrests
Administration Schemas
Information
No object selected
Object Info Session
SQL File 1*
Limit to 1000 rows
32 SELECT
33 AVG(d.Neonatal_mortality_rate) as Median
34 FROM
35 (SELECT @entirecol:=@entirecol + 1 AS entirecol,
36 Neonatal_mortality_rate AS Neonatal_mortality_rate
37 FROM child_mortality
38 ORDER BY Neonatal_mortality_rate) AS d
39 WHERE
40 d.entirecol IN (FLOOR(@entirecol / 2), CEIL(@entirecol / 2));
Result Grid
Filter Rows:
Export:
Wrap Cell Content:
Median
26.1
Result 7
Read Only
SQLAdditions
Jump to
Automatic context help is disabled. Use the toolbar to manually get help for the current caret position or to toggle automatic help.
Context Help Snippets
Output
Action Output
Time Action Message Duration / Fetch
63 19:17:23 SET @entirecol := -1 0 row(s) affected 0.078 sec
64 19:17:29 SELECT AVG(d.Neonatal_mortality_rate) as Median FROM (SELECT @entirec... 1 row(s) returned 0.078 sec / 0.000 sec
Query Completed
7:17 PM
2/15/2022

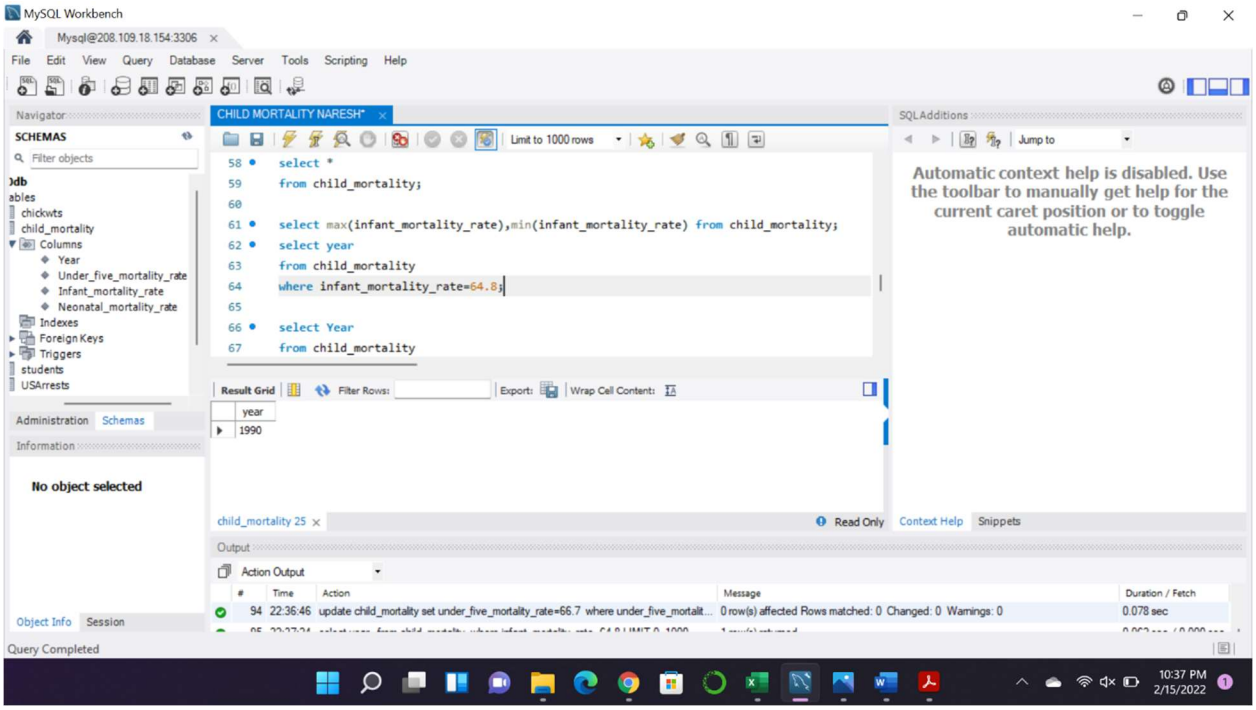
Display the entire table ;

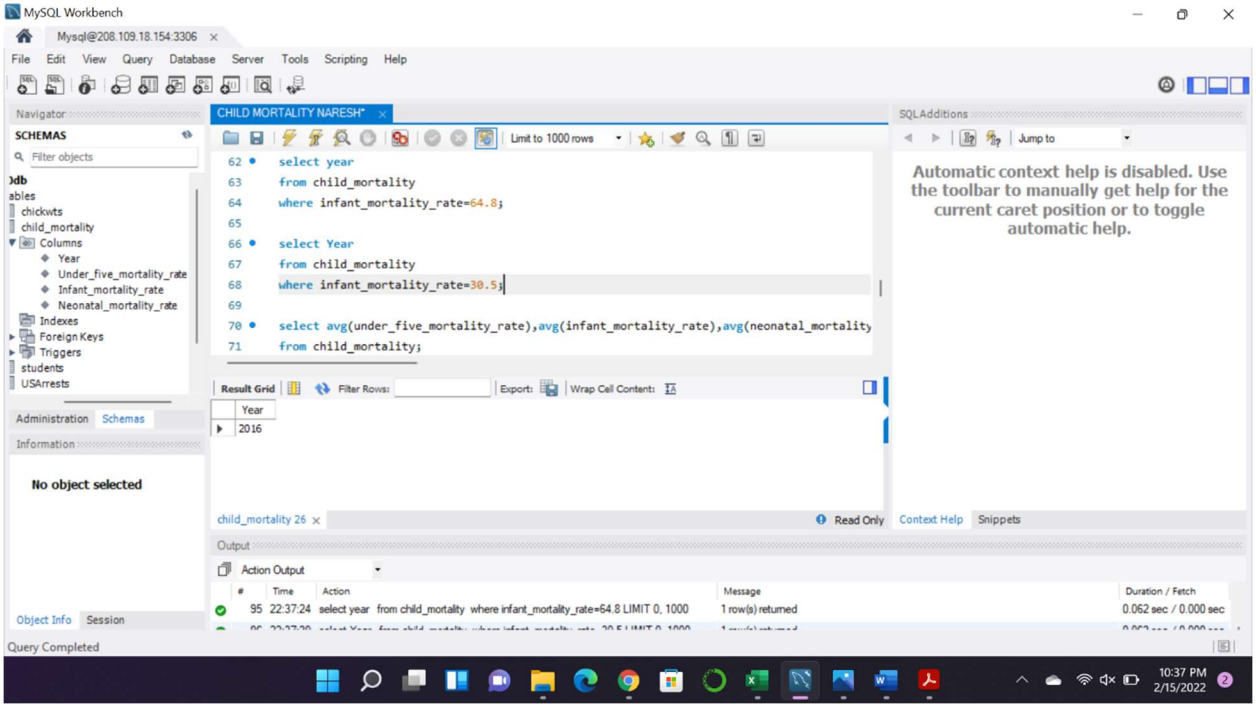


Which years have the lowest and highest infant mortality years, respectively?

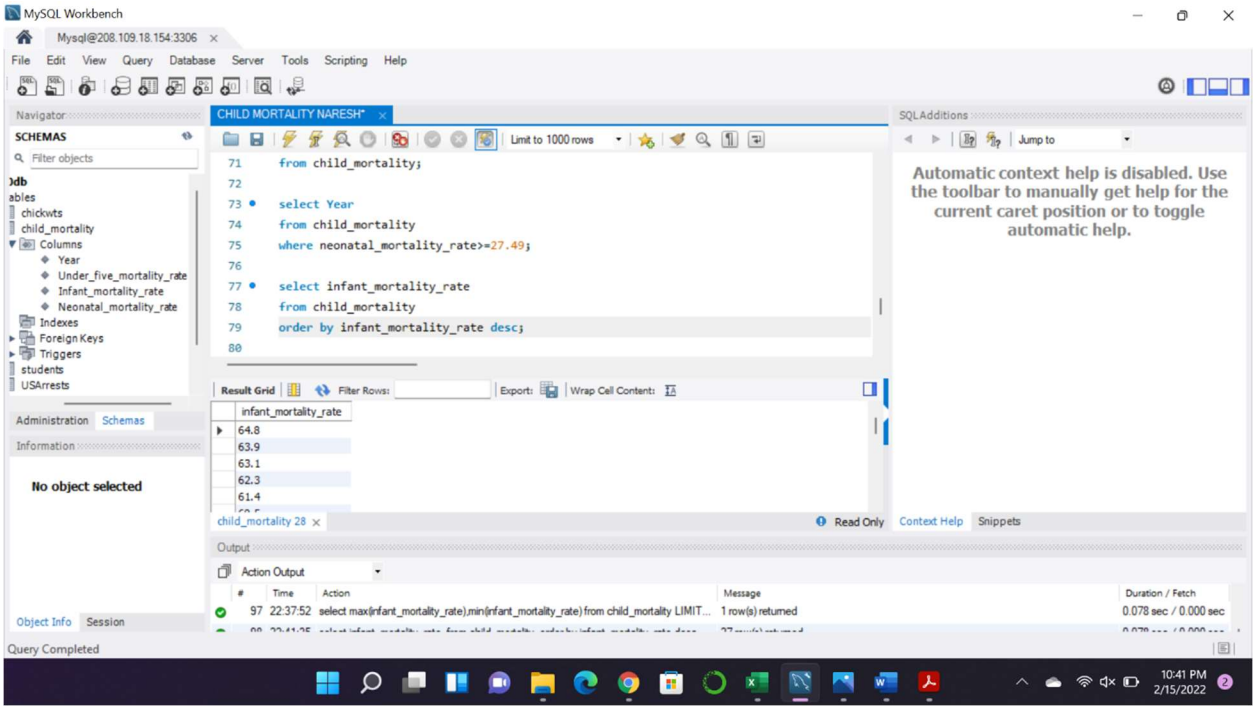


In what years the neonatal mortality rates were above average?

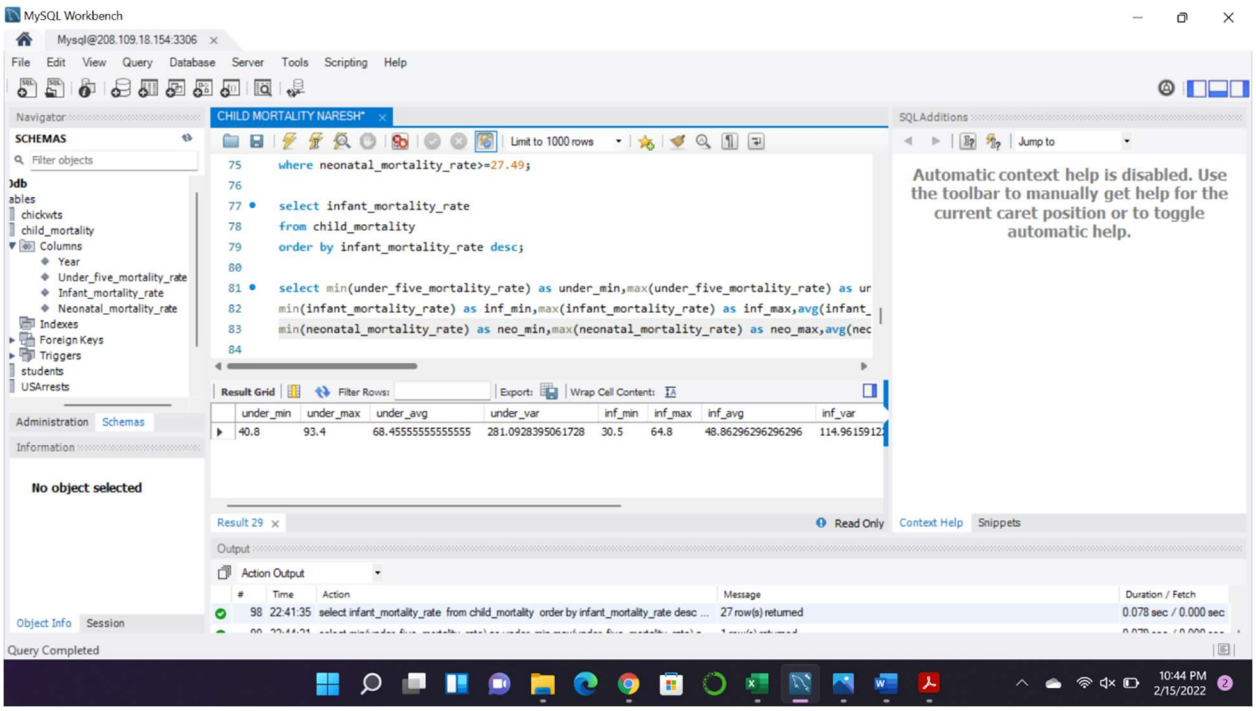




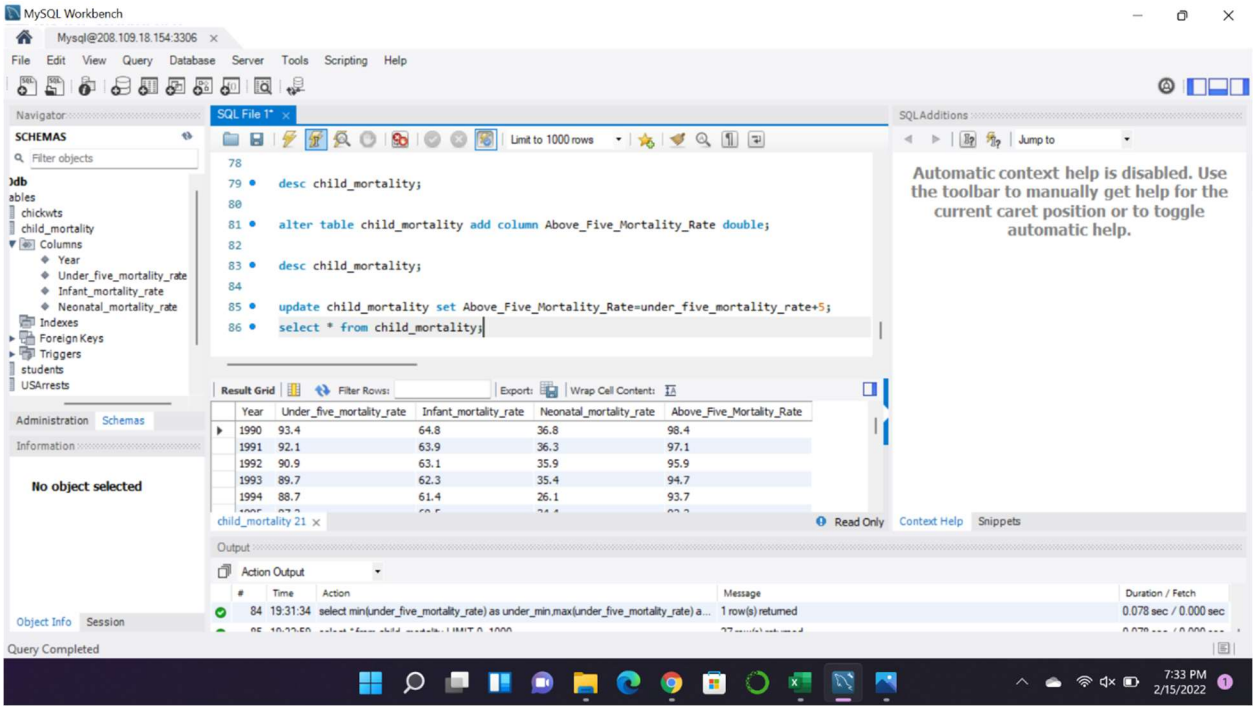
Display the sorted infant mortality rates in descending order.



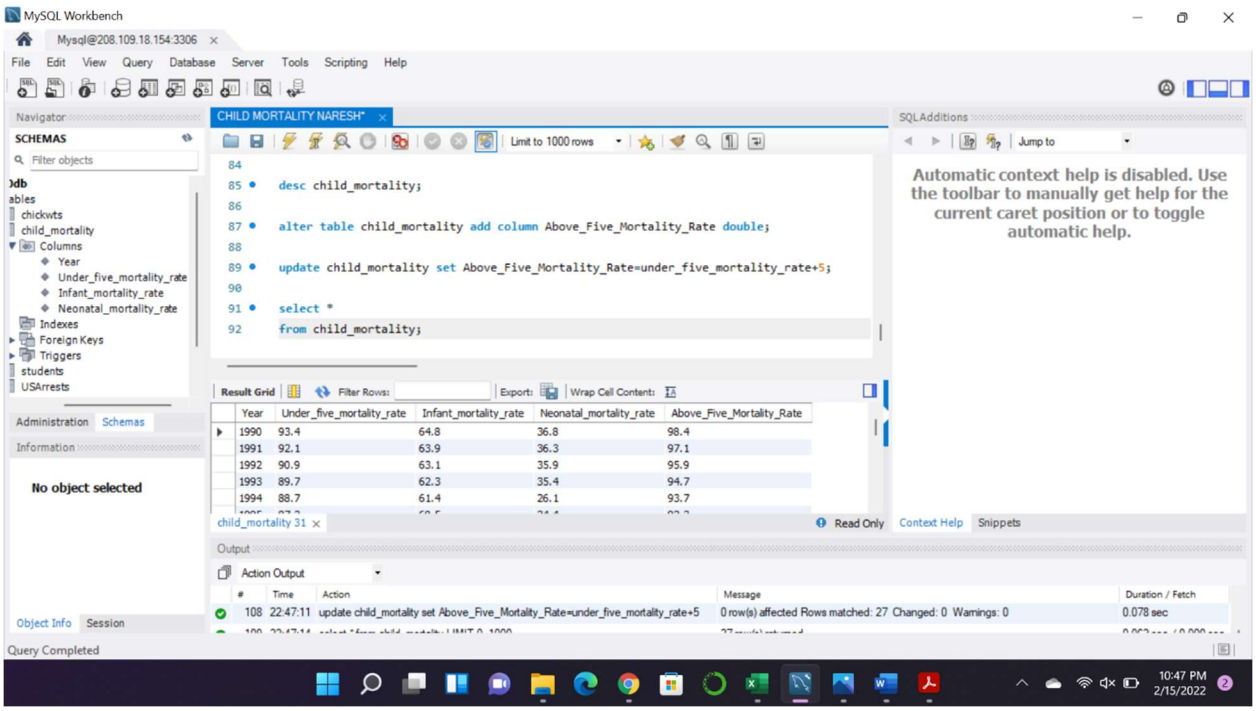
Display **min**, **max**, **mean**, **variance**, and **standard deviation** for each mortality rate.



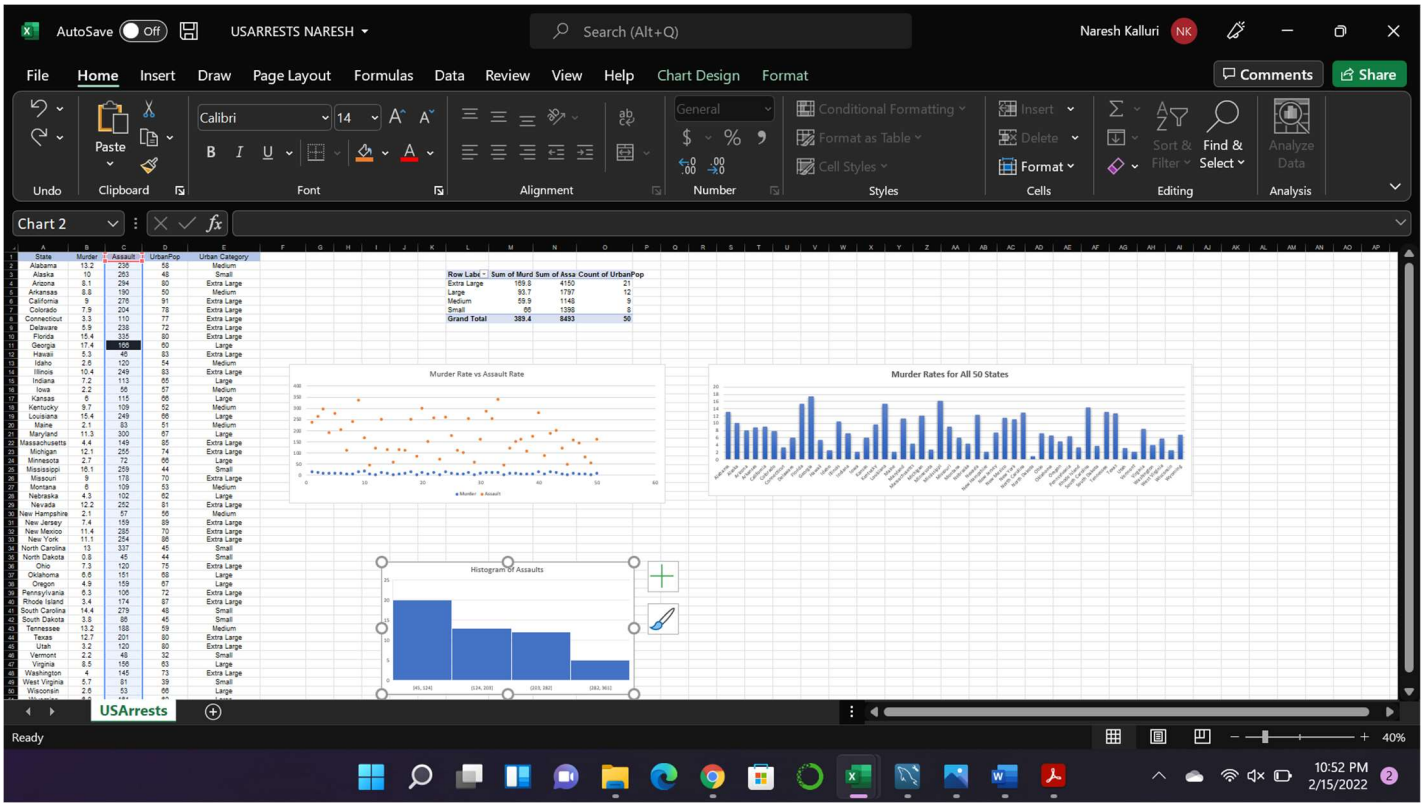
Add a new column called **Above-Five Mortality Rate** and populate it with appropriate values.



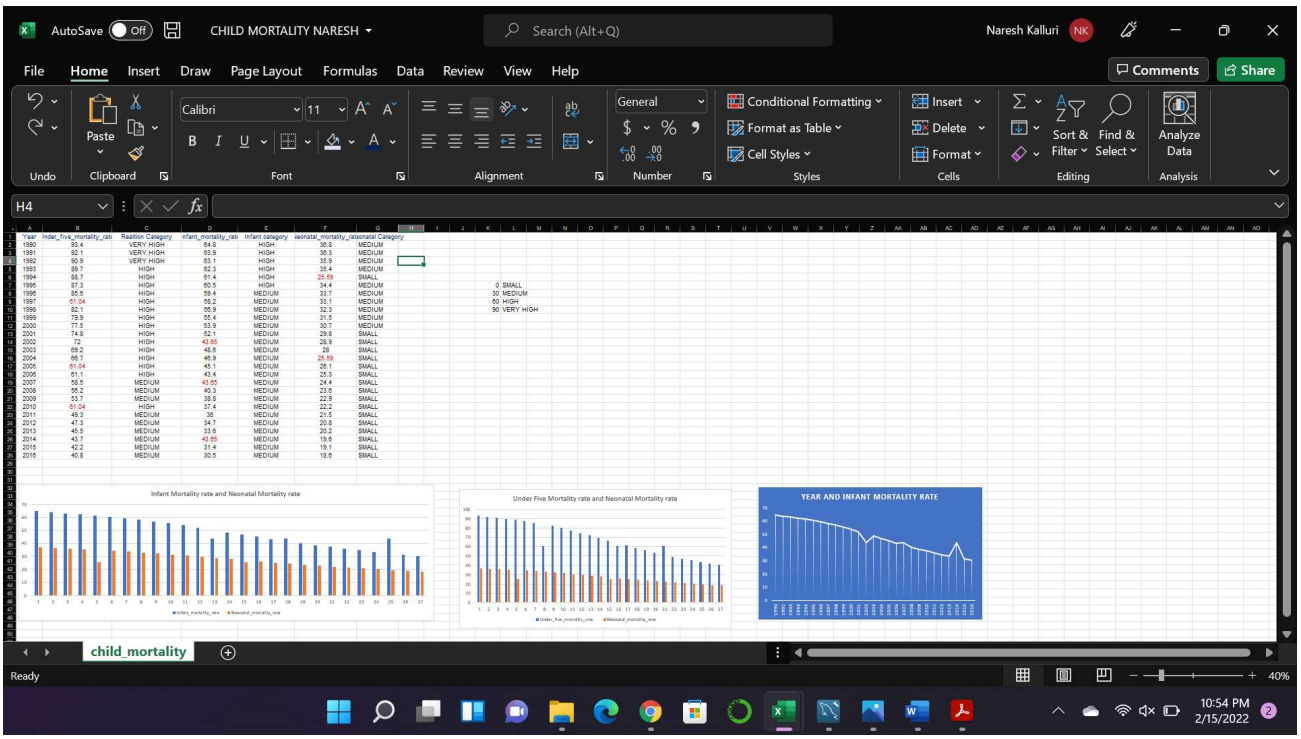
Display the entire table again.



Excel Worksheets: Problem 1:



Problem 2:



References:

https://www.geeksforgeeks.org/calculate-median-in-mysql/

Used for how to calculate Median in MySQL

I converted my cleansed data into XML and JSON formats and uploaded it to my GitHub repository including my Excel Worksheets.