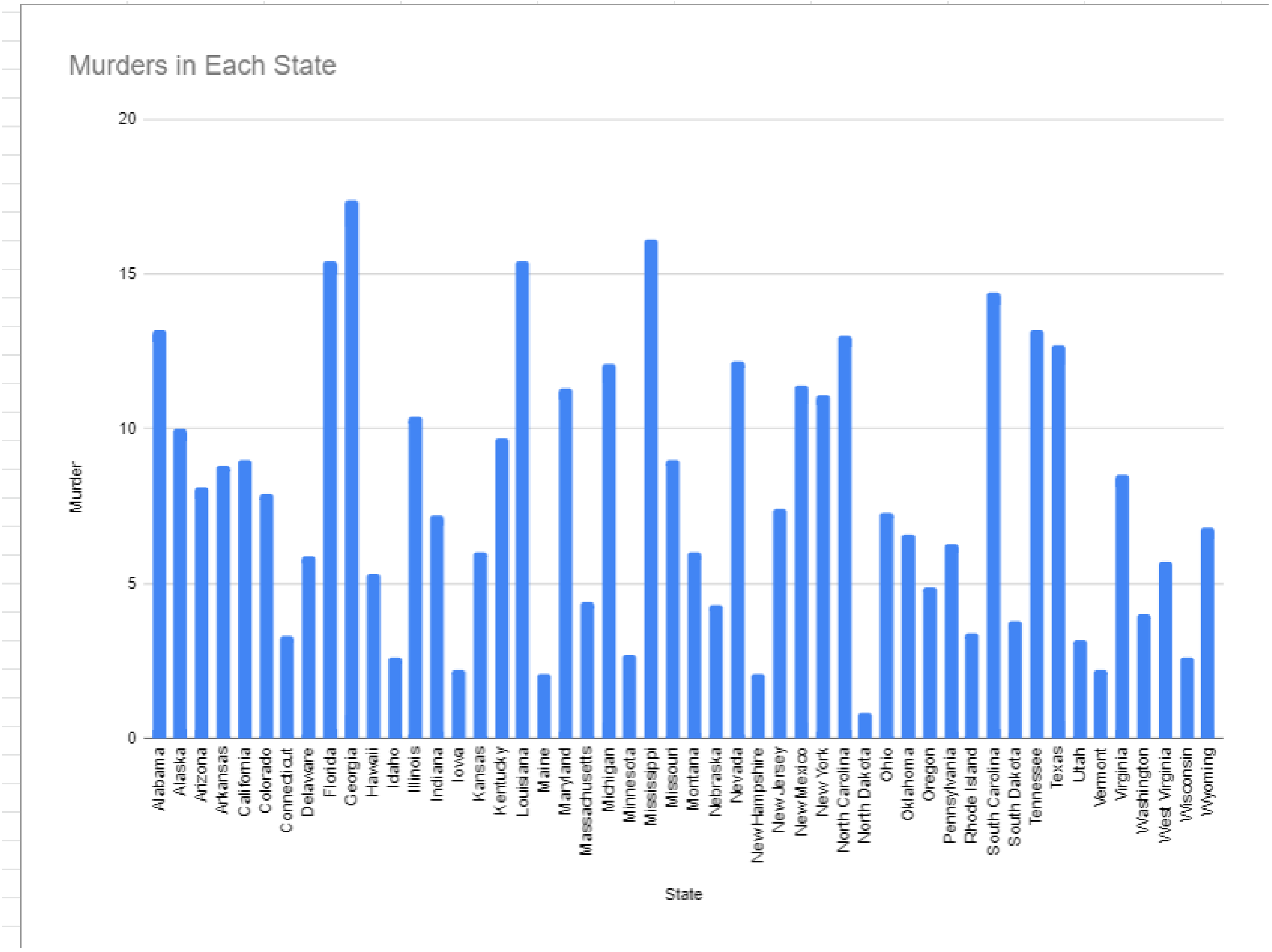
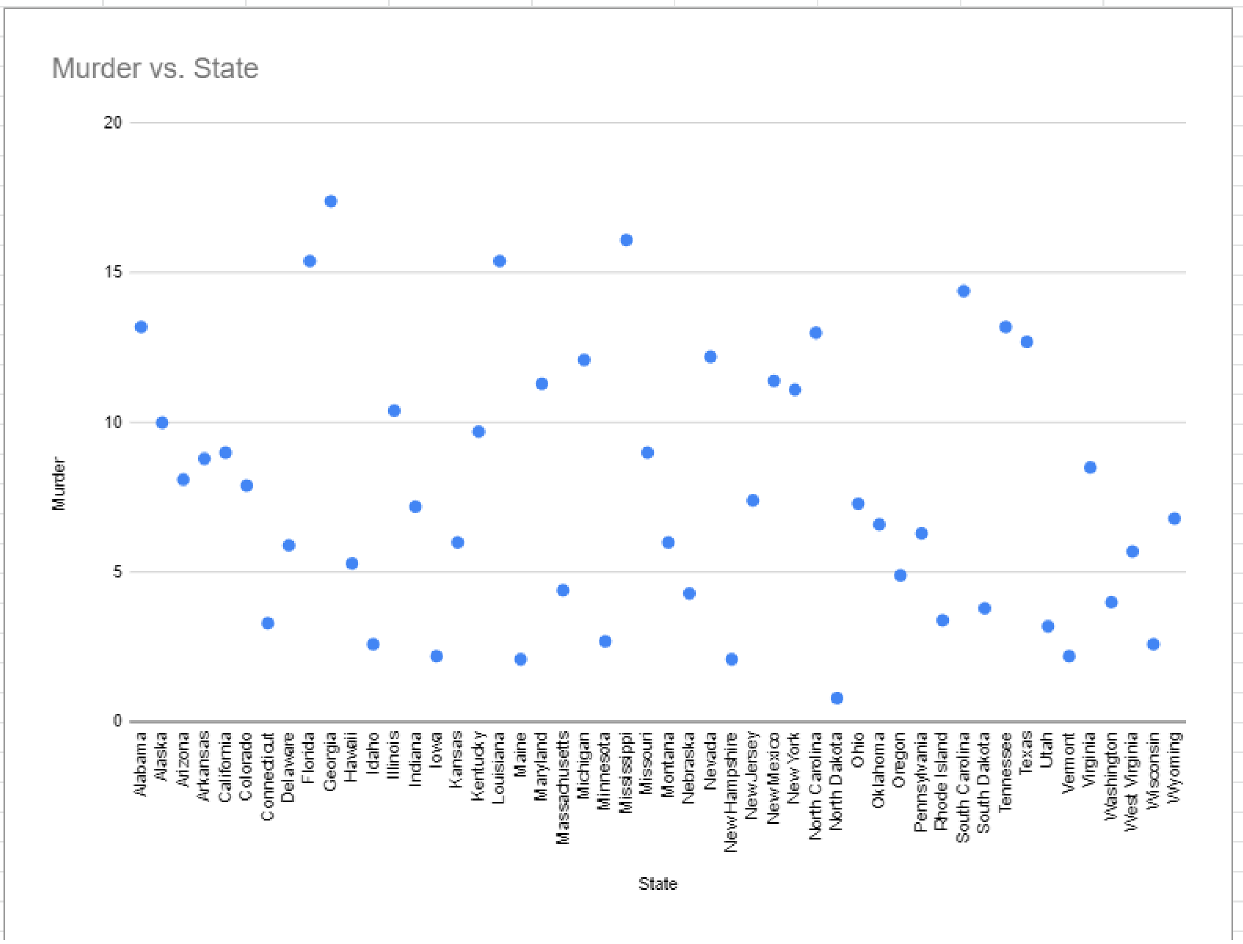
Patrick O’Boyle
CSC 346 Introduction to Data Science
Project 2
2/7/2022
Repo: https://github.com/oboyle3/patoboyle-dataScienceFun/tree/main/IDS/HW2

Excel

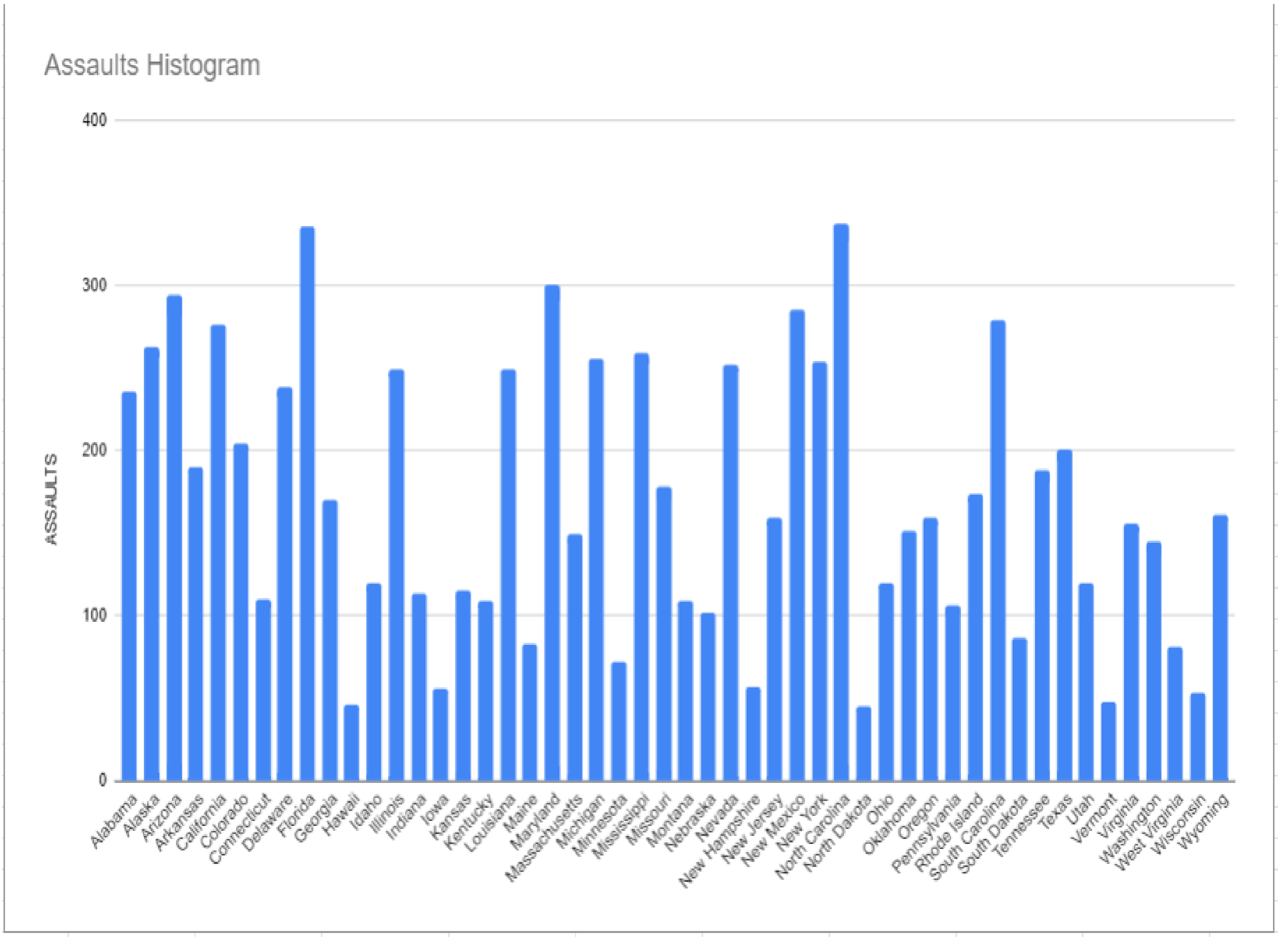
A. One missing value that I found was in the assault category for the state of Georgia. I took the average of all the other values in assault using google sheets and ended up with 169.93 as a replacement.
B. For the murder category, I didn't find any outliers. The range of min to max in that set is 0.8 to 17.4. The assault category varied very much with the smallest being 45 and the highest 337. The urban population had no outliers. I didn't find any noisy data with the set given. I don't think there are any data entry problems or faulty collections that were given.
C. This is the plotted graph for the murder rates of all 50 states:



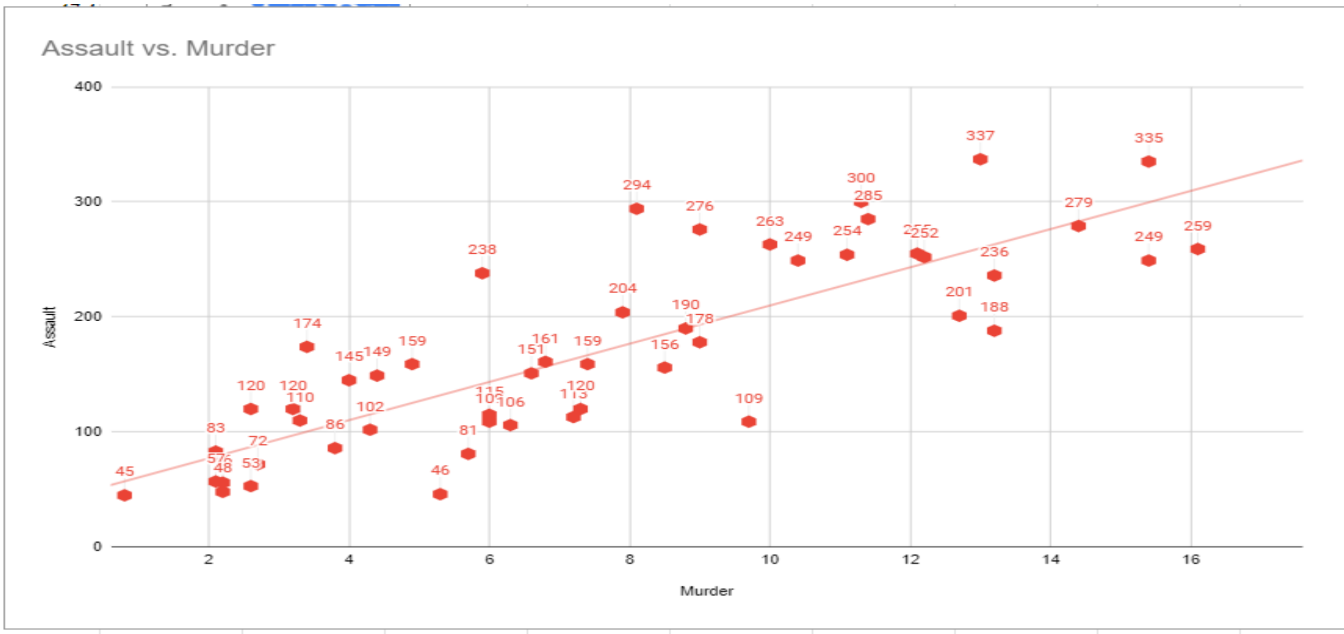
This is the same graph but in scatter plot form:


Murder vs. State
20
15
10
5
0
Murder
Alabama
Alaska
Arizona
Arkansas
California
Colorado
Connecticut
Delaware
Florida
Georgia
Hawaii
Idaho
Illinois
Indiana
Iowa
Kansas
Kentucky
Louisiana
Maine
Maryland
Massachusetts
Michigan
Minnesota
Mississippi
Missouri
Montana
Nebraska
Nevada
NewHampshire
NewJersey
NewMexico
New York
North Carolina
North Dakota
Ohio
Oklahoma
Oregon
Pennsylvania
Rhode Island
South Carolina
South Dakota
Tennessee
Texas
Utah
Vermont
Virginia
Washington
West Virginia
Wisconsin
Wyoming
State

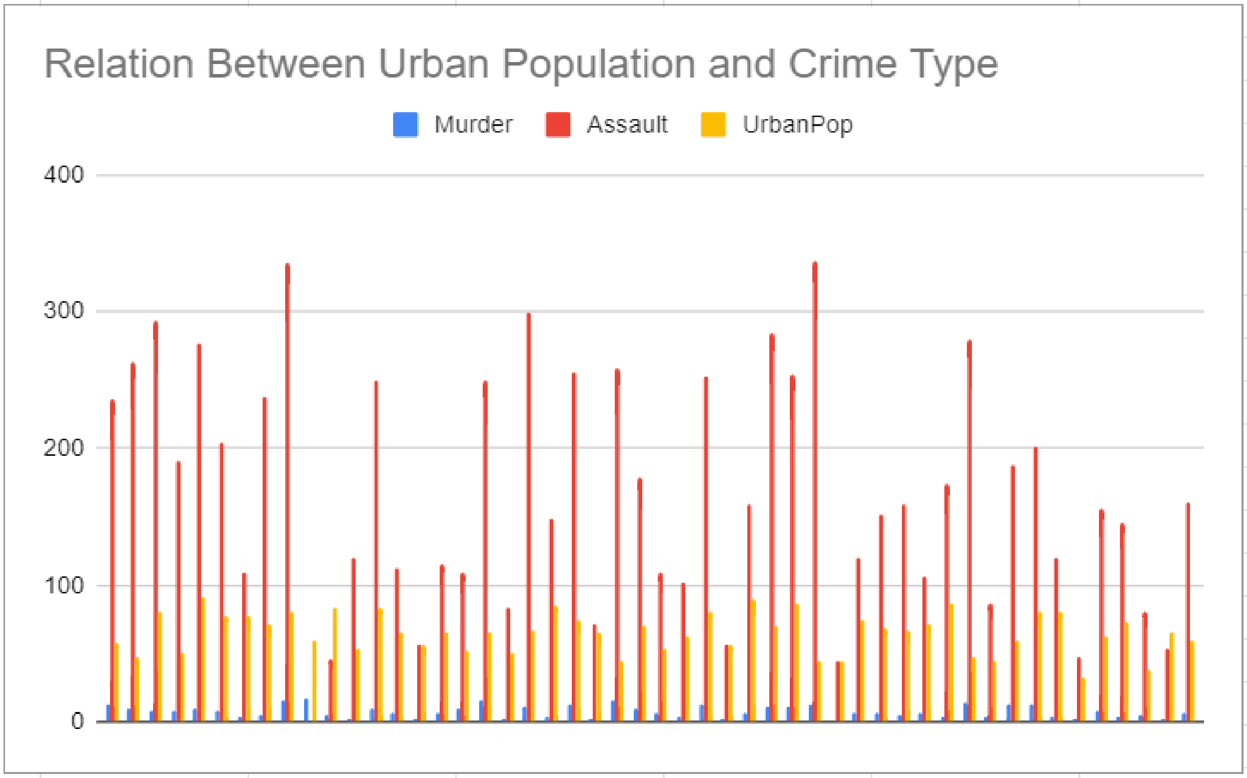
D. This is the assaults histogram:



E. This is the murder vs the assault rate:

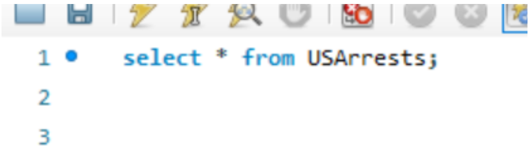


F. This is the relation between an urban population category and a crime type.

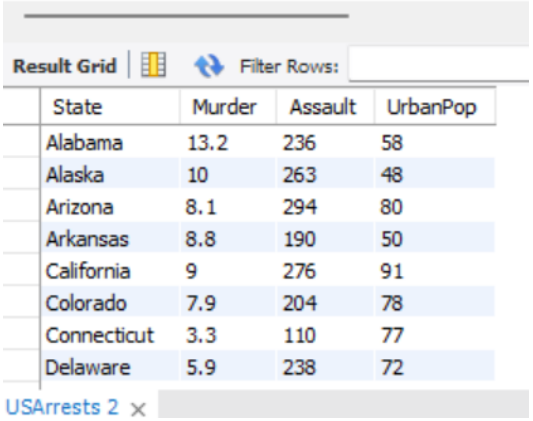


<u>MYSQL-</u>

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

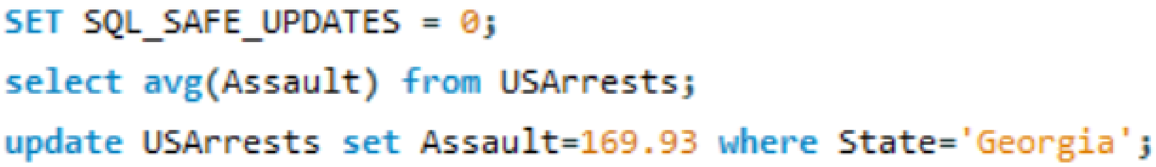


```
select * from USArrests;
```



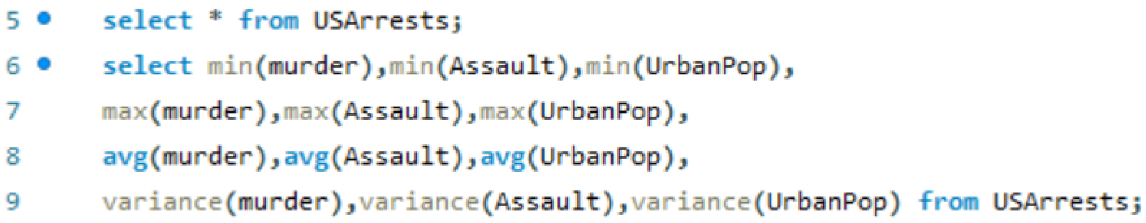
State	Murder	Assault	UrbanPop
Alabama	13.2	236	58
Alaska	10	263	48
Arizona	8.1	294	80
Arkansas	8.8	190	50
California	9	276	91
Colorado	7.9	204	78
Connecticut	3.3	110	77
Delaware	5.9	238	72

B. The average to be replaced is 169.93



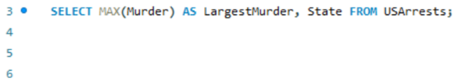
```
SET SQL_SAFE_UPDATES = 0;
select avg(Assault) from USArrests;
update USArrests set Assault=169.93 where State='Georgia';
```

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

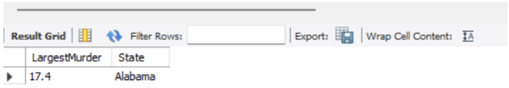


```
select * from USArrests;
select min(murder),min(Assault),min(UrbanPop),
max(murder),max(Assault),max(UrbanPop),
avg(murder),avg(Assault),avg(UrbanPop),
variance(murder),variance(Assault),variance(UrbanPop) from USArrests;
```

D. Alabama had the maxmimim murder rate

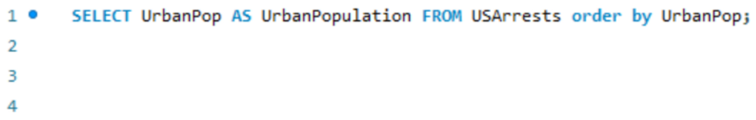


```
SELECT MAX(Murder) AS LargestMurder, State FROM USArrests;
```

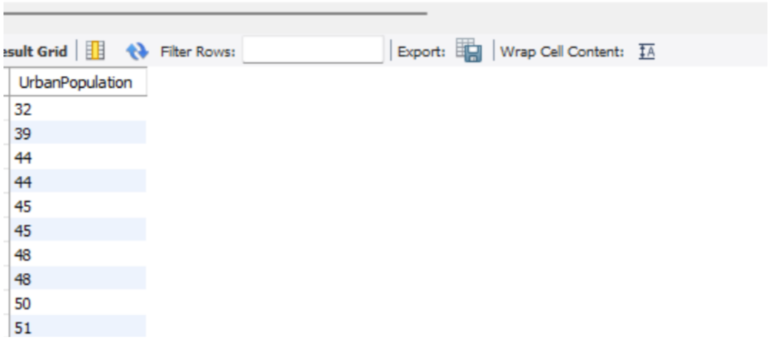


LargestMurder	State
17.4	Alabama

List of states in ascending order of urban population percentages

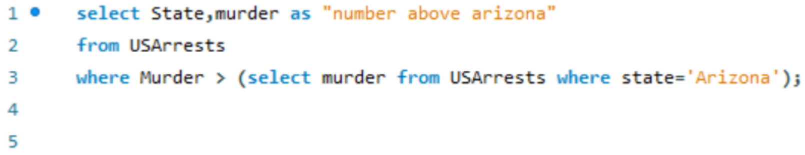


```sql
SELECT UrbanPop AS UrbanPopulation FROM USArrests order by UrbanPop;
```

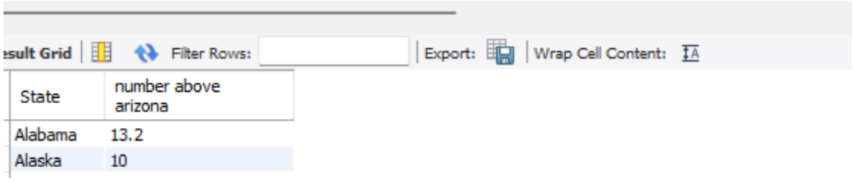


UrbanPopulation
32
39
44
44
45
45
48
48
50
51

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



```sql
select State,murder as "number above arizona"
from USArrests
where Murder > (select murder from USArrests where state='Arizona');
```



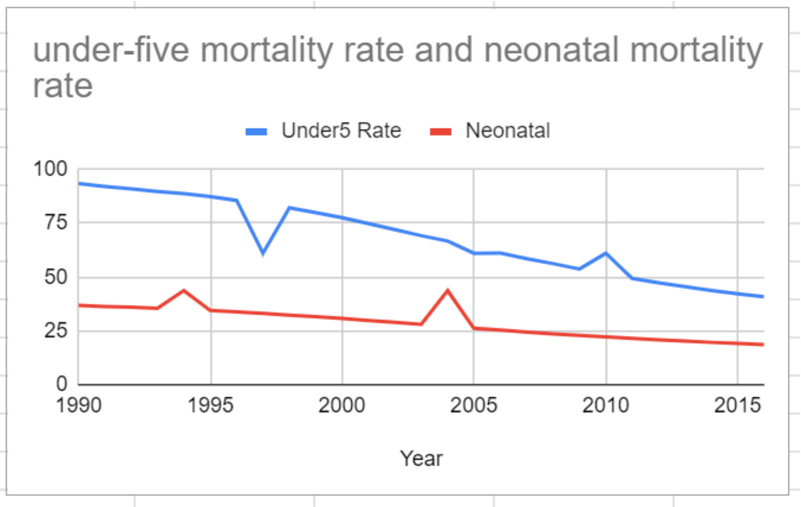
State	number above arizona
Alabama	13.2
Alaska	10

Problem 2:

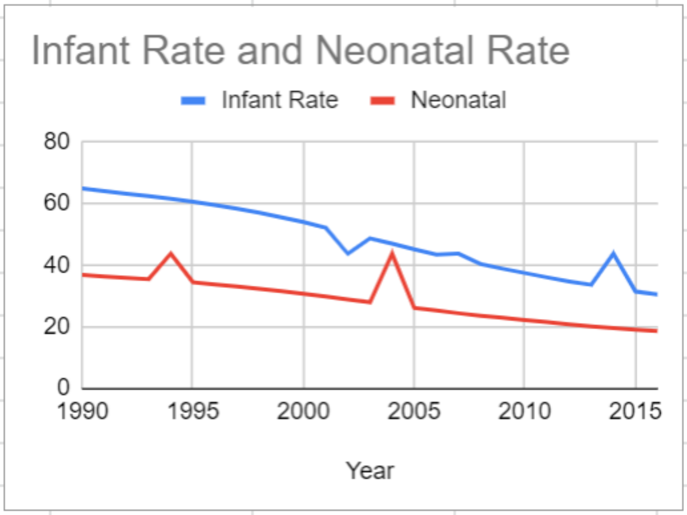
Excel -

A. To address the missing values I took the averages of each column and replaced those values. I didn't find any major outliers that I could remove. There also doesn't seem like there are any inconsistencies that need to be resolved in the data set.

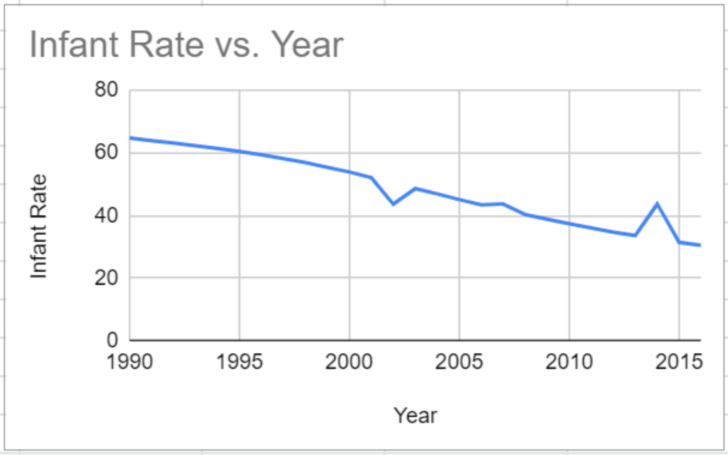
B. Here is the graph for the relation of under-five mortality rate and neonatal mortality rate.



Here is the graph for Infant mortality rate and neonatal mortality rate.

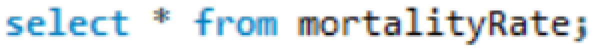


Here is the graph for the year and infant mortality rate.

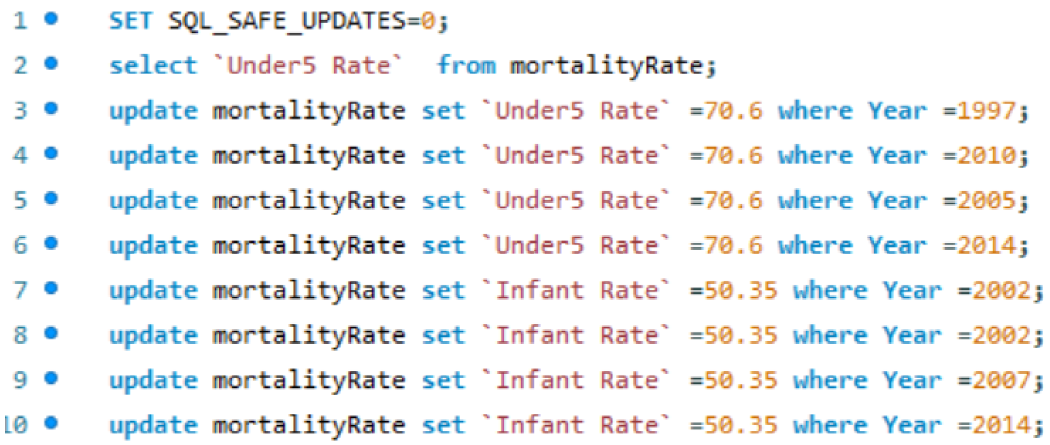


MYSQL -

A. Import the original dataset to mysql.

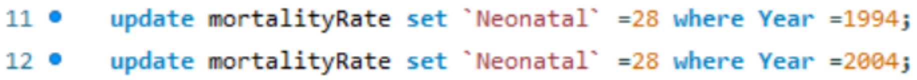


```
select * from mortalityRate;
```



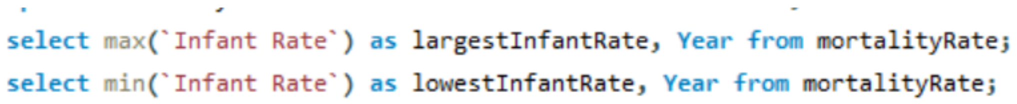
```
SET SQL_SAFE_UPDATES=0;
select `Under5 Rate`  from mortalityRate;
update mortalityRate set `Under5 Rate` =70.6 where Year =1997;
update mortalityRate set `Under5 Rate` =70.6 where Year =2010;
update mortalityRate set `Under5 Rate` =70.6 where Year =2005;
update mortalityRate set `Under5 Rate` =70.6 where Year =2014;
update mortalityRate set `Infant Rate` =50.35 where Year =2002;
update mortalityRate set `Infant Rate` =50.35 where Year =2002;
update mortalityRate set `Infant Rate` =50.35 where Year =2007;
update mortalityRate set `Infant Rate` =50.35 where Year =2014;
```

B.



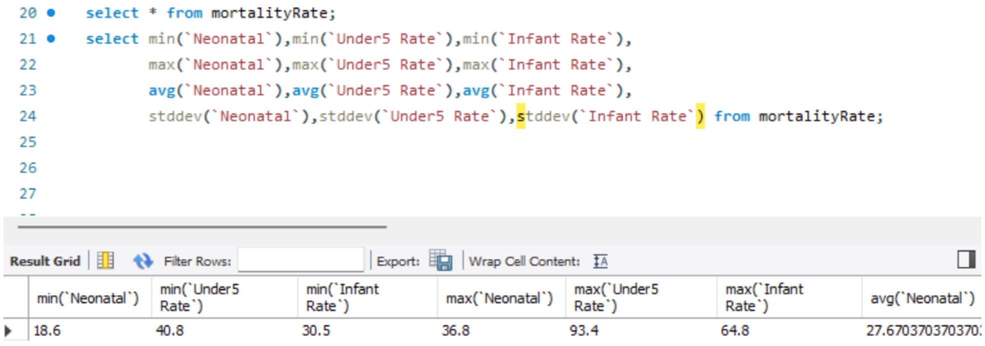
```
update mortalityRate set `Neonatal` =28 where Year =1994;
update mortalityRate set `Neonatal` =28 where Year =2004;
```

C.



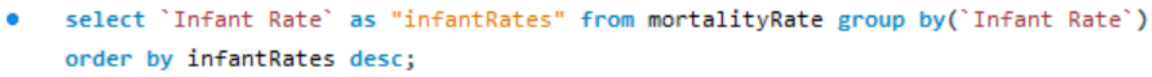
```
select max(`Infant Rate`) as largestInfantRate, Year from mortalityRate;
select min(`Infant Rate`) as lowestInfantRate, Year from mortalityRate;
```

Which years have the lowest and highest infant mortality years, respectively? 2016 is the lowest and is the highest 1990



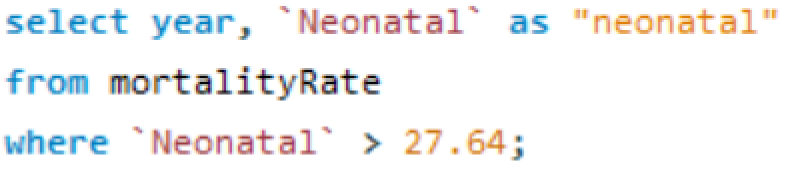
```
select * from mortalityRate;
select min(`Neonatal`),min(`Under5 Rate`),min(`Infant Rate`),
       max(`Neonatal`),max(`Under5 Rate`),max(`Infant Rate`),
       avg(`Neonatal`),avg(`Under5 Rate`),avg(`Infant Rate`),
       stddev(`Neonatal`),stddev(`Under5 Rate`),stddev(`Infant Rate`) from mortalityRate;
```

min(`Neonatal`)	min(`Under5 Rate`)	min(`Infant Rate`)	max(`Neonatal`)	max(`Under5 Rate`)	max(`Infant Rate`)	avg(`Neonatal`)
18.6	40.8	30.5	36.8	93.4	64.8	27.67037037037037



```
select `Infant Rate` as "infantRates" from mortalityRate group by(`Infant Rate`)
order by infantRates desc;
```

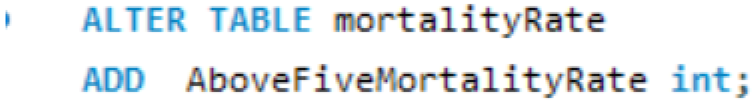
In what years the neonatal mortality rates were above average?



```
select year, `Neonatal` as "neonatal"
from mortalityRate
where `Neonatal` > 27.64;
```

Add a new column called Above-Five Mortality Rate and populate it with appropriate values. Hint: Use Alter Table Add Column.

1.



```
ALTER TABLE mortalityRate
ADD  AboveFiveMortalityRate int;
```

<u>Report</u>

1. The purpose of this assignment is to better our understanding and skills in MySQL workbench. In the first problem I first used excel to address the missing values and used the average to fill in the missing numbers. I then plotted the murder rates for the 50 states, then a histogram for the assaults. Lastly, I used excel to establish a relationship between urban population and crime. In MYSQL i first imported the csv file through the table wizard to get it into mysql. I used a mysql algorithm to get the min, max, mean, and variance.
2. In problem 2 we examined the child mortality rates. For the missing values I used the average of the comum to fill in the missing values. I then made a graph in google sheets for each of the queries in problem 2 b. I used the medians of the columns to fill in the missing data values. I used much of the same methodology as used in problem one to solve the problems in the second one.
3. The methodology I used for this assignment involved MYSQL and google sheets. I was able to implement what was asked of the assignment through MYSQL.
4. My conclusion was that murder vs what state they lived in the forst problem had so correlation and the data was very scattered. In the second problem I noted that as the years went on and with medicine getting better the rates of child mortality have gone down.
5. I used the slides from class and some websites.
   https://www.techrepublic.com/article/how-to-create-tables-and-add-data-to-mysql-database-with-mysql-workbench/
   https://www.techonthenet.com/mysql/order_by.php
   https://www.w3schools.com/sql/sql_alter.asp