John Salt

Intro to Data Science

Project 2

February 15, 2022

<div align="center">US Arrests and Child Mortality</div>

Purpose

The purpose of this project is to use various tools such as Excel and SQL to analyze these data sets.

Methodology

By using excel to view and manipulate the data we are able to view statistics based on the data and insert missing values, fix outliers, and generate graphs of the data.

Conclusion and Summary

Problem 1:

First, I started by adding the missing value for Georgia's assault rate. I did this by replacing the missing value with the average of all other states' assault rates. Next, I plotted the data to look for outliers in the set. Doing this I compared the Urban Population percentage to both murder and assault rates to look for unusual values. Doing this, I noticed that Hawaii's Urban population was much higher and by contrast, the Assault rate was low in comparison to other states. In order to clean this, I again replaced the value with the average. Other than that I didn't notice any other outlying values.

Murder Rates by State
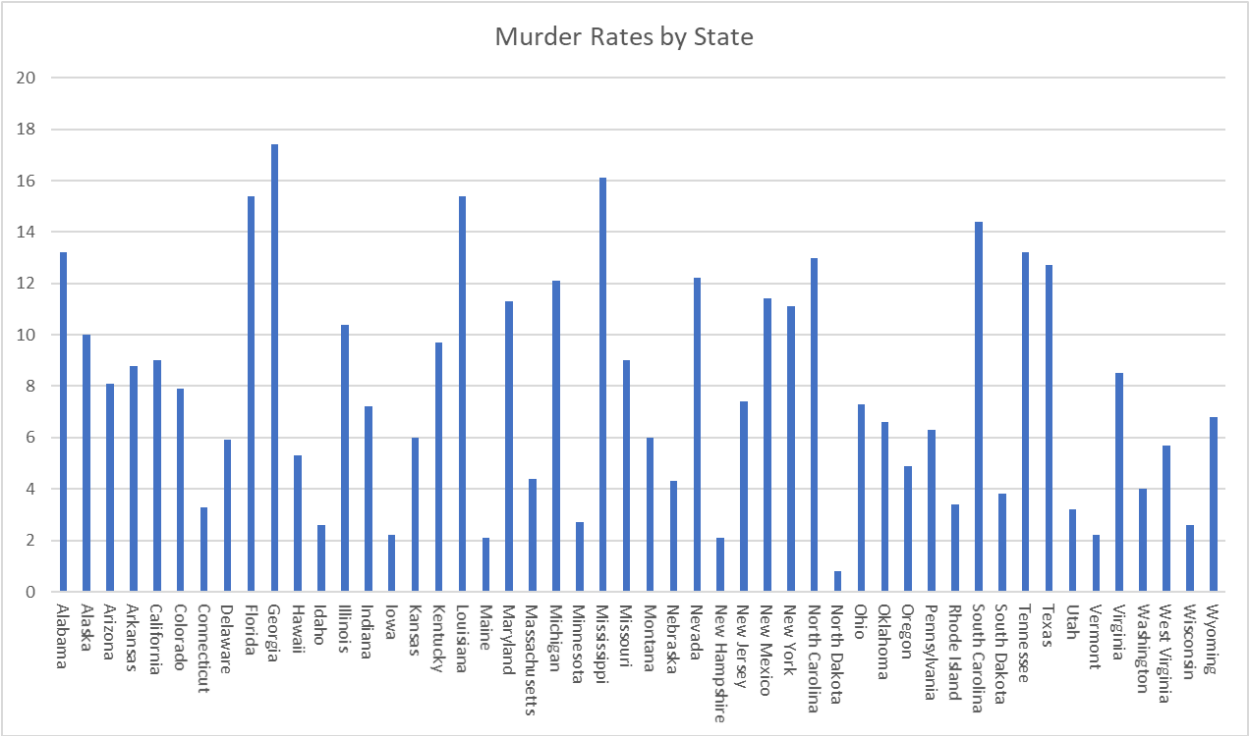
First, I plotted the murder rate by state using a bar graph.
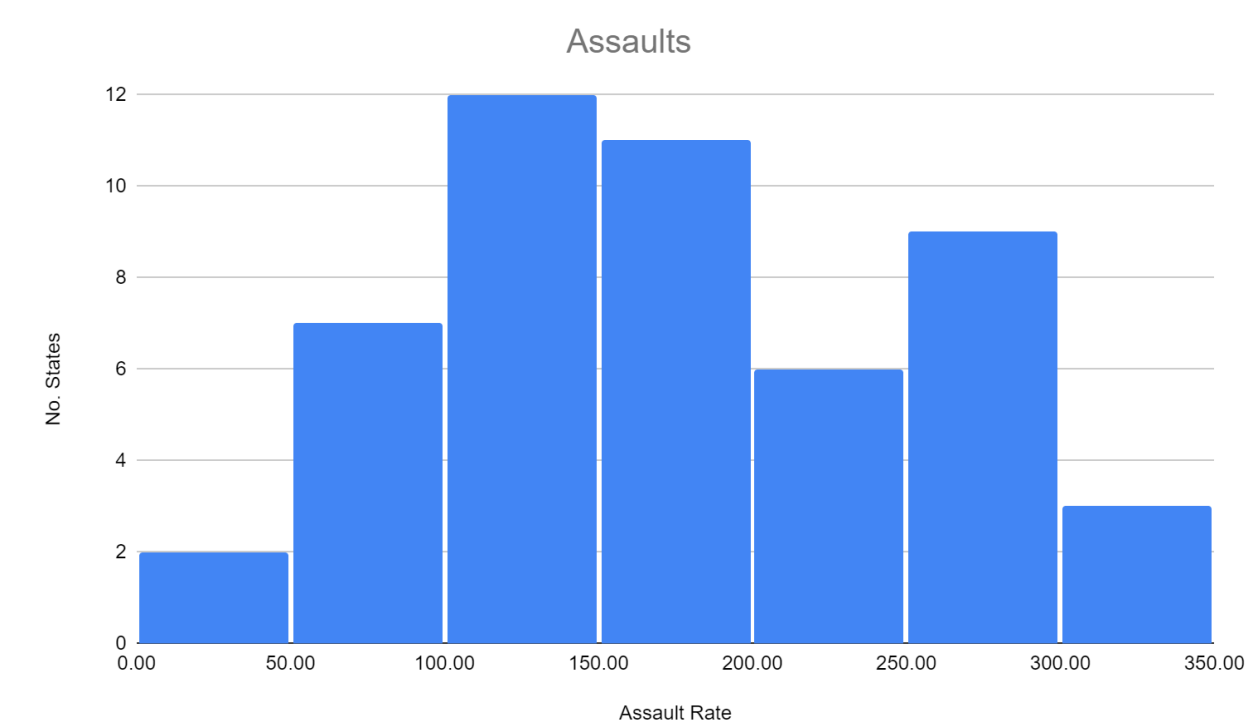
Assaults

Next, I used a histogram to plot the assault rate for the states. This graph shows the number of states with assault rates falling within the ranges on the horizontal axis. After this I used a line graph to plot Murder vs. Assault rates.
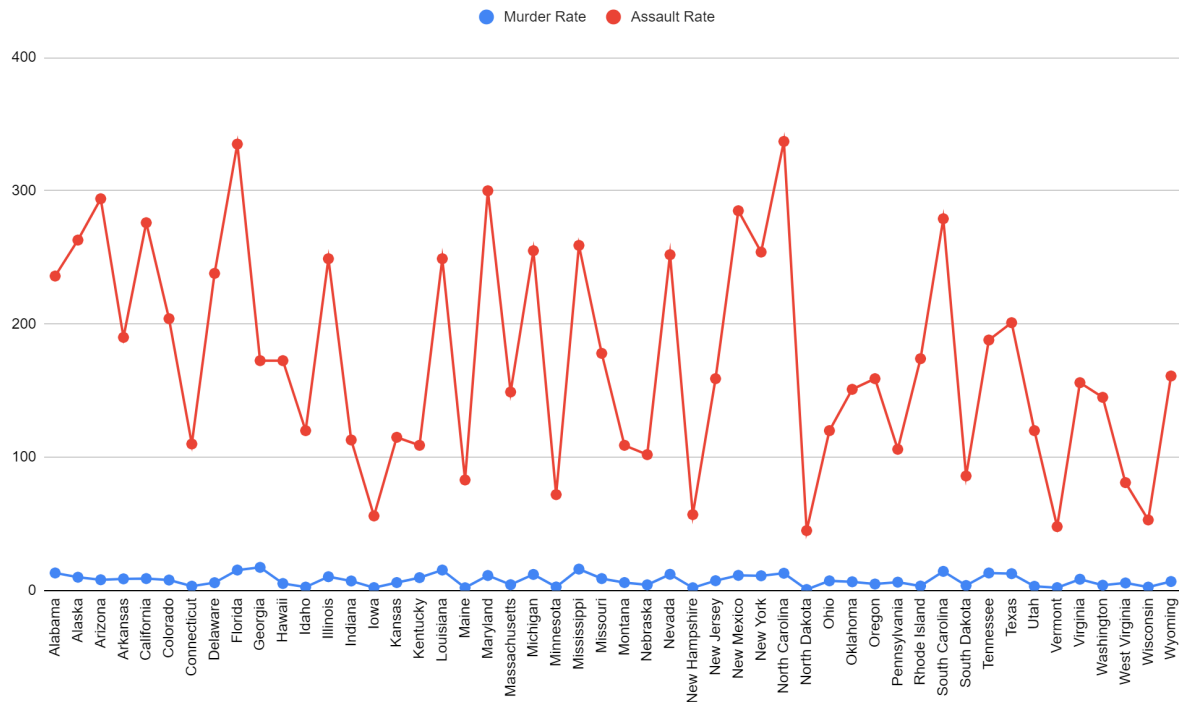
SQL Analysis:

Using SQL, I was able to collect the min, max, average, and variance for each column and produced this table:

	Min	Max	Avg	Variance
Assault	45	337	169.88	6,773.39
Murder	0.8	17.4	7.788	18.59
UrbanPop	32	91	65.54	205.33

1. **Which state has the maximum murder rate?**

   Georgia, 17.4

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

STATE	UrbanPop	STATE	UrbanPop
Vermont	32	Wisconsin	66
West Virginia	39	Maryland	67
Mississippi	44	Oregon	67
North Dakota	44	Oklahoma	68
North Carolina	45	Missouri	70
South Dakota	45	New Mexico	70
Alaska	48	Delaware	72
South Carolina	48	Pennsylvania	72
Arkansas	50	Washington	73
Maine	51	Michigan	74
Kentucky	52	Ohio	75
Montana	53	Connecticut	77
Idaho	54	Colorado	78
New Hampshire	56	Arizona	80
Iowa	57	Florida	80
Alabama	58	Texas	80
Tennessee	59	Utah	80
Georgia	60	Nevada	81
Wyoming	60	Hawaii	83
Nebraska	62	Illinois	83
Virginia	63	Massachusetts	85
Indiana	65	New York	86
Kansas	66	Rhode Island	87
Louisiana	66	New Jersey	89
Minnesota	66	California	91

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

There are 21 states whose murder rate is higher than Arizona. Those states are:

Alabama, Alaska, Arkansas, California, Florida, Georgia, Illinois, Kentucky, Louisiana,

Maryland, Michigan, Mississippi, Missouri, Nevada, New Mexico, New York, North Carolina,

South Carolina, Tennessee, Texas, Virginia

Problem 2:

  To start with this dataset I replaced the various missing values with the average of the year before and the year after. This was done to ensure that the data maintains a decreasing trend. Next, I examined the data looking for outliers, which I did not, and then I started plotting the different values.
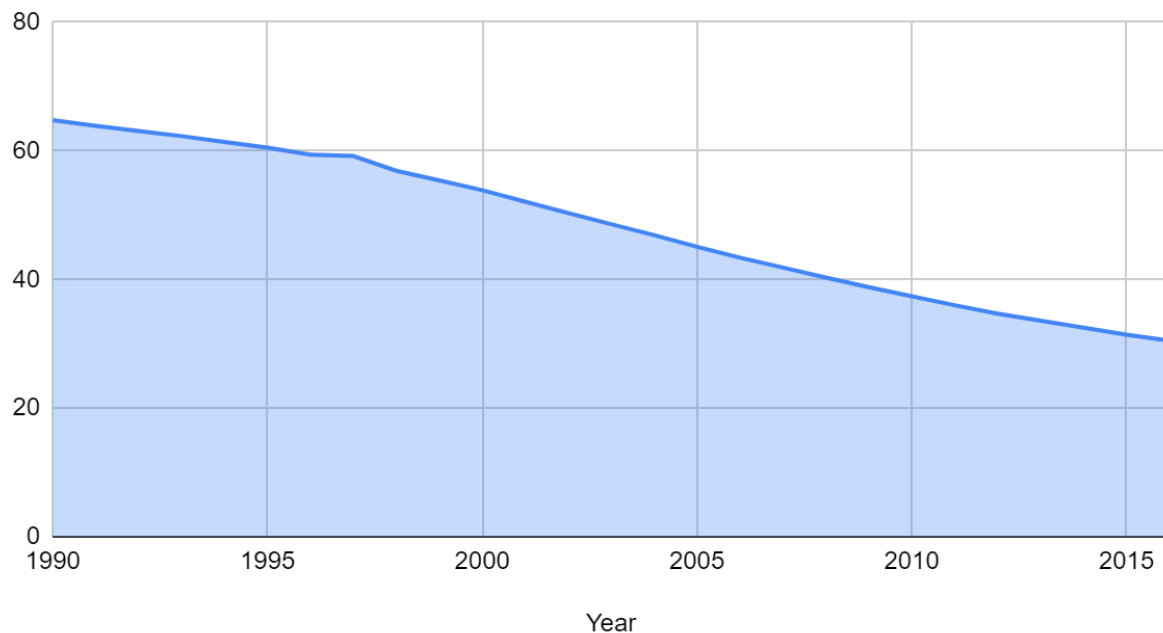

Under five mortality rate and Neonatal mortality rate
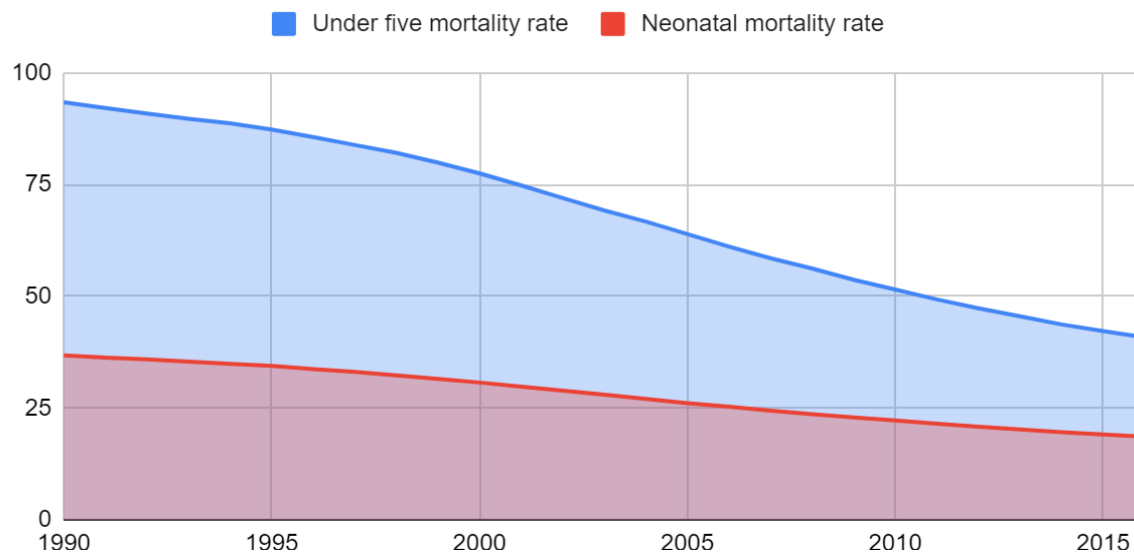
Infant Mortality Rate and Neonatal Mortality Rate
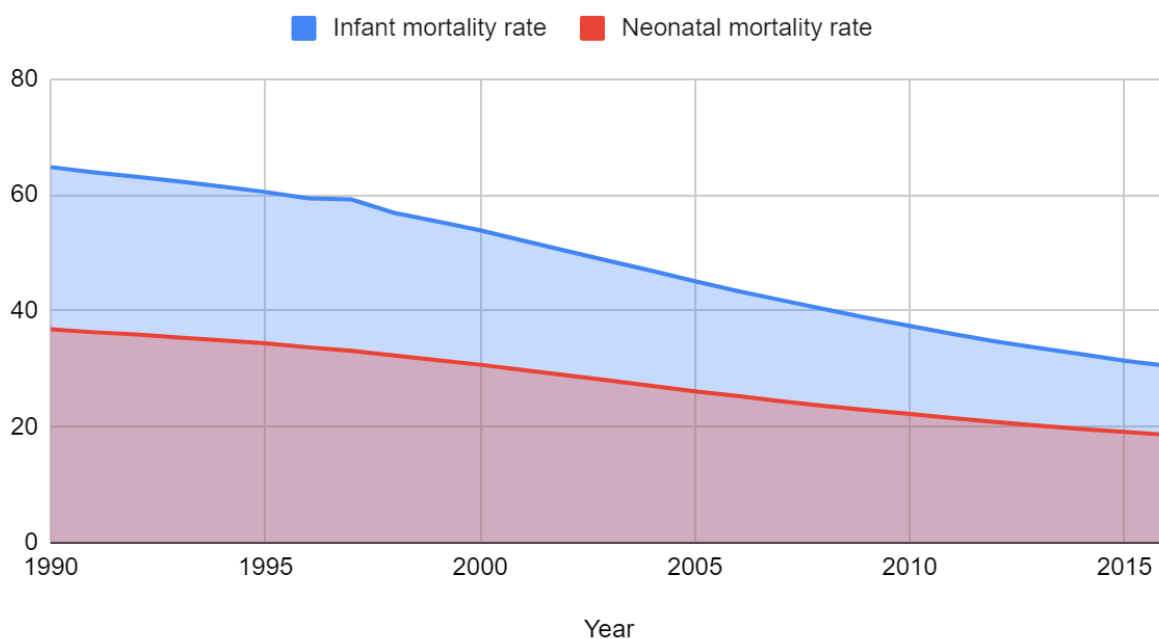
Infant Mortality Rate by Year

In these graphs, I plotted Under-five vs Neonatal mortality rate, Infant vs Neonatal mortality rate, and the infant mortality rate by year.

SQL Analysis:

To analyze this data set I first found the years with the lowest and highest of each mortality rate. In all instances the year with the lowest was 2016 and the year with the highest was 1990. Next, all years between 2004 and 1990 (inclusive) had neonatal mortality rates that were higher than average. Below are the images showing the tables after displaying all the content, displaying the data after sorting by descending infant mortality rate, and finally the table after adding a row for above five mortality rates. Notice that in the second table the values aren't in order by year as using the median as a filler for missing values leads to inaccurate values.

year	infant mortality rate
1996	59.4
1997	59.2
1998	56.9
1999	55.4
2000	53.9
2001	52.1
2014	48.6
2007	48.6
2003	48.6
2002	48.6
2004	46.9
2005	45.1
2006	43.4
2008	40.3
2009	38.8
2010	37.4
2011	36
2012	34.7
2013	33.6
2015	31.4
2016	30.5

Year	Under five mortality rate	Infant mortality rate	Neonatal mortality rate
1990	93.4	64.8	36.8
1991	92.1	63.9	36.3
1992	90.9	63.1	35.9
1993	89.7	62.3	35.4
1994	88.7	61.4	28
1995	87.3	60.5	34.4
1996	85.6	59.4	33.7
1997	69.2	59.2	33.1
1998	82.1	56.9	32.3
1999	79.9	55.4	31.5
2000	77.5	53.9	30.7
2001	74.8	52.1	29.8
2002	72	48.6	28.9
2003	69.2	48.6	28
2004	66.7	46.9	28
2005	69.2	45.1	26.1
2006	61.1	43.4	25.3

(Note: "92.1" appears as a marked value beside the 1992 Under five mortality rate.)

Year	Under five mortality rate	Infant mortality rate	Neonatal mortality rate	Above Five Mortality Rate
1995	87.3	60.5	34.4	99912.7
1996	85.6	59.4	33.7	99914.4
1997	69.2	59.2	33.1	99930.8
1998	82.1	82.1	32.3	99917.9
1999	79.9	55.4	31.5	99920.1
2000	77.5	53.9	30.7	99922.5
2001	74.8	52.1	29.8	99925.2
2002	72	48.6	28.9	99928
2003	69.2	48.6	28	99930.8
2004	66.7	46.9	28	99933.3
2005	69.2	45.1	26.1	99930.8
2006	61.1	43.4	25.3	99938.9
2007	58.5	48.6	24.4	99941.5
2008	56.2	40.3	23.6	99943.8
2009	53.7	38.8	22.9	99946.3
2010	69.2	37.4	22.2	99930.8