

Zach Makki

Intro to Data Science

Homework #2

02/14/2022

Problem #1

### 1. Purpose

Address missing values and outliers in the USA Arrests dataset and find the relationship between murder rate and assault rate as well as additional information about each attribute in the dataset.

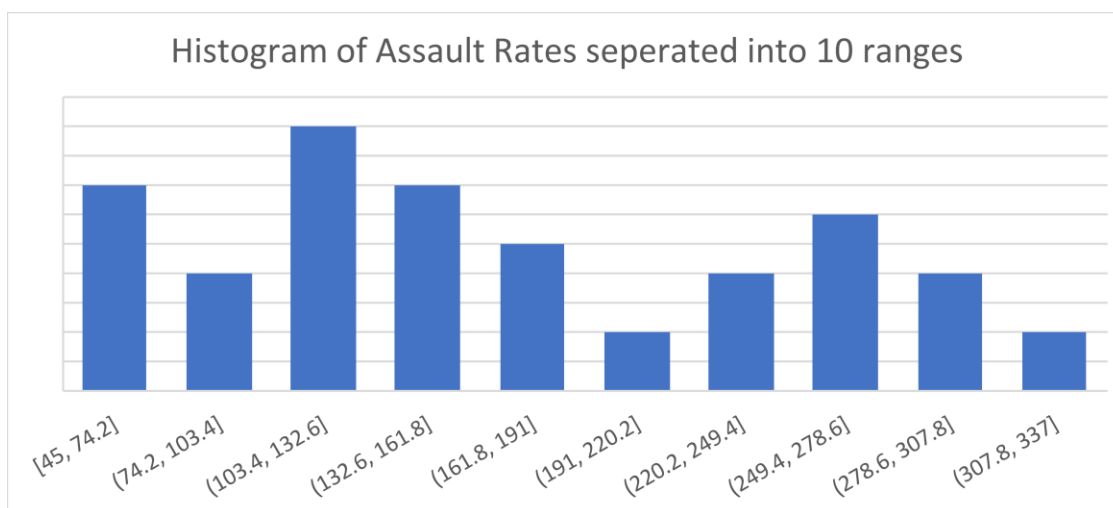
### 2. Methodology

For the dataset two different tools were used to clean and visualize the data, excel and MySQL. This section will first review the work done in excel. To address missing values the average of any column which contained a missing value was calculated and the missing value was then replaced with its columns average. Outliers were established by plotting each attributes first individually, then in comparison with each other. An excel if condition was used to populate the values of the UrbanPop\_Cat created column. The if condition logic is as follows if the urban population percent for a state is between 50 and 60% the state is assigned the category "mid". If the percent is between 60 and 70 then it is assigned category "high". All values above 70 percent are categorized as "very-high" and all other values (those less then 50) are categorized as "low".

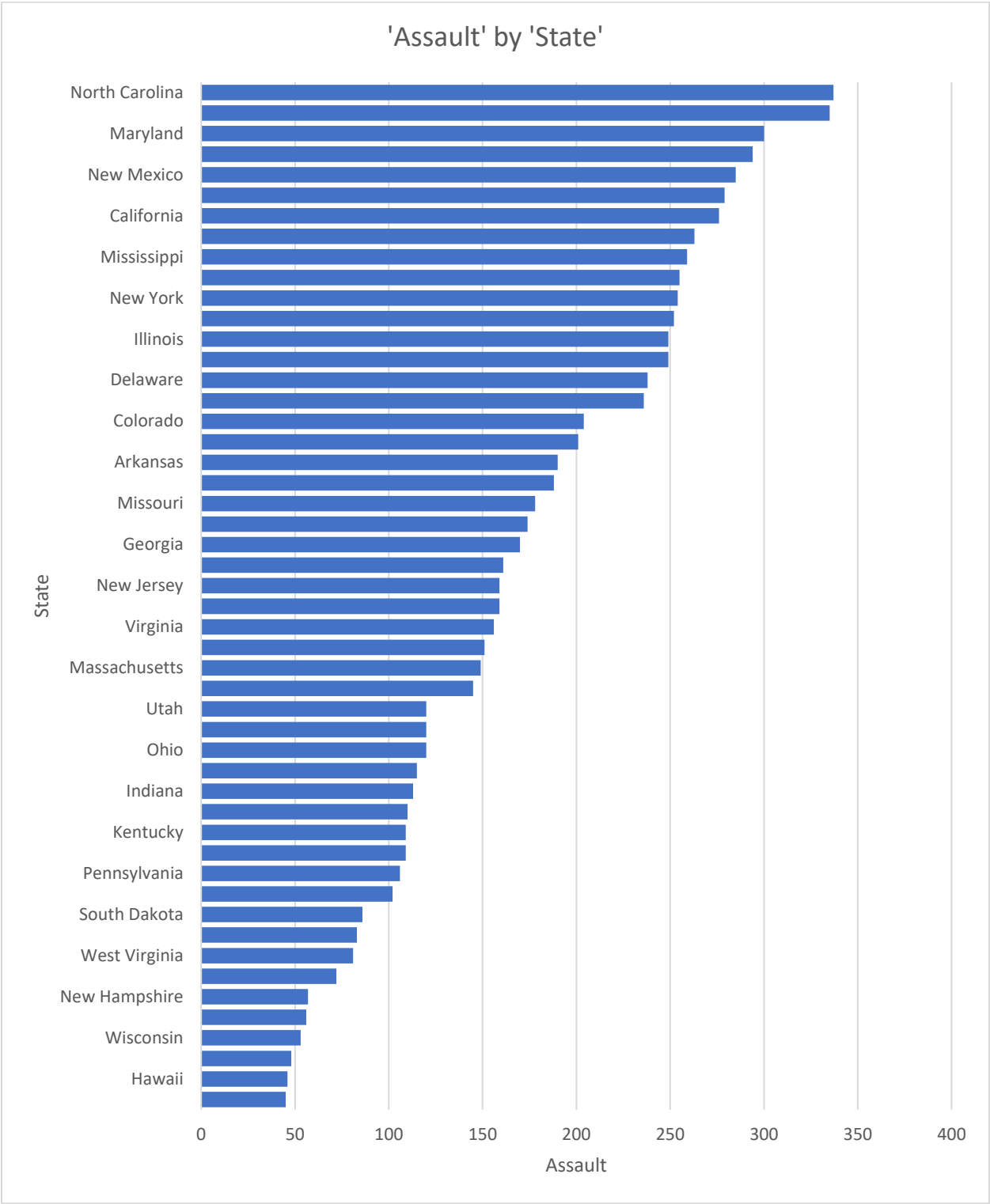
In SQL the column averages were calculated and used to fill all missing values. Queries were then used to get each attributes minimum values, maximum value, mean, and variance. Queries were also executed to establish which states had the highest murder rate, list the states in order of urban population percentage, and determine which states had a higher murder rate then Arizona.

### 3. Data

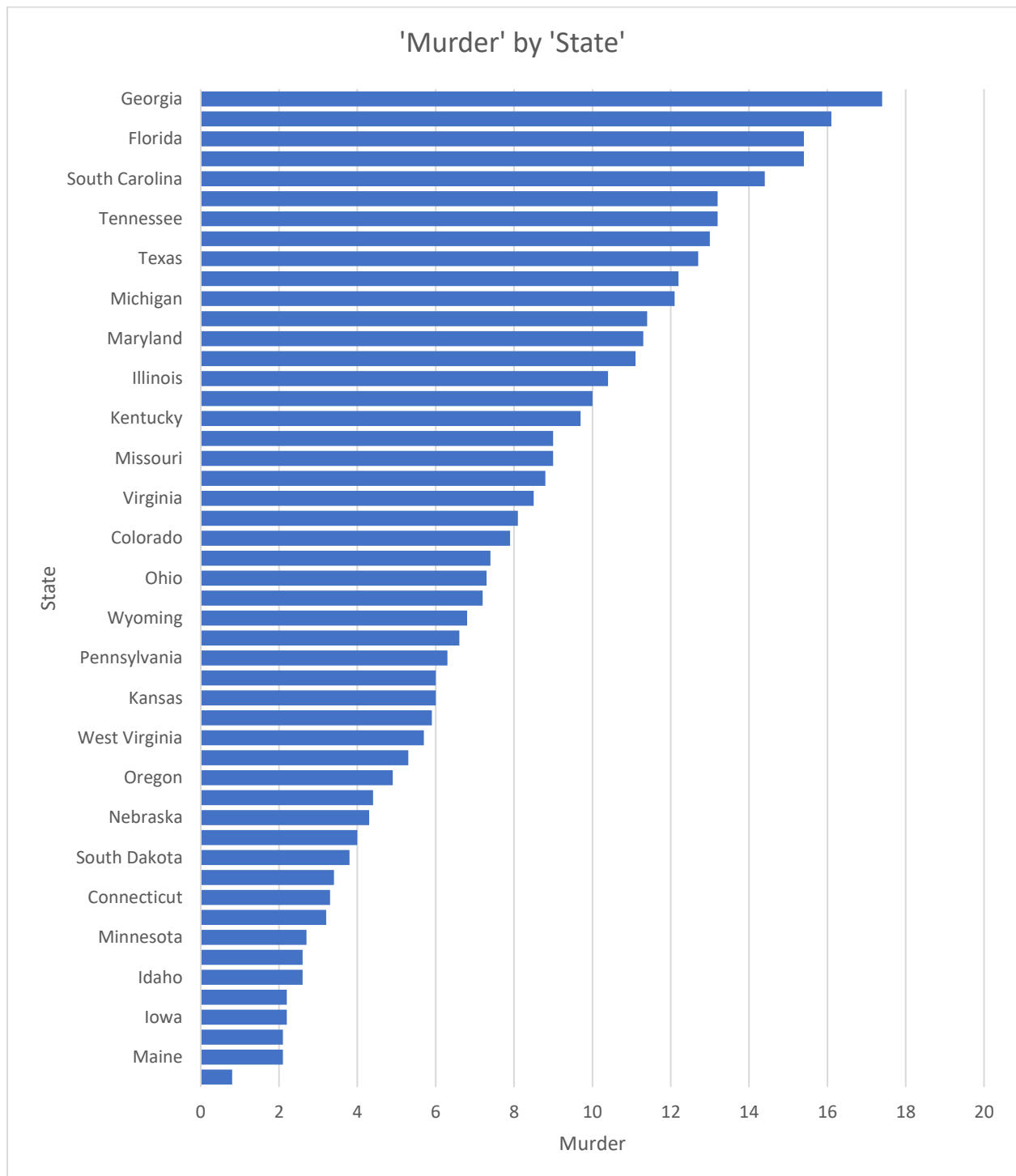
Assault Rate Data Distribution



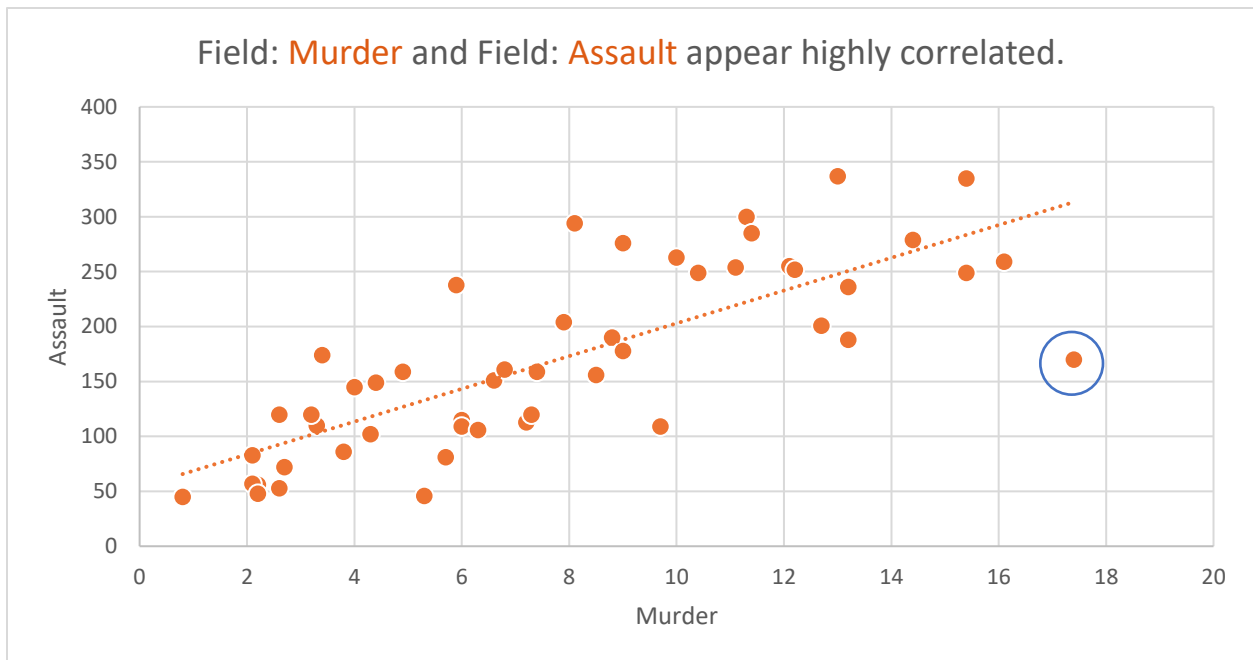
States sorted by Assault Rate (Due to size of chart not all state names pictured, to see all view chart in excel)



States sorted by Murder Rate before outlier smoothed (Due to size of chart not all state names pictured, to see all view chart in excel)

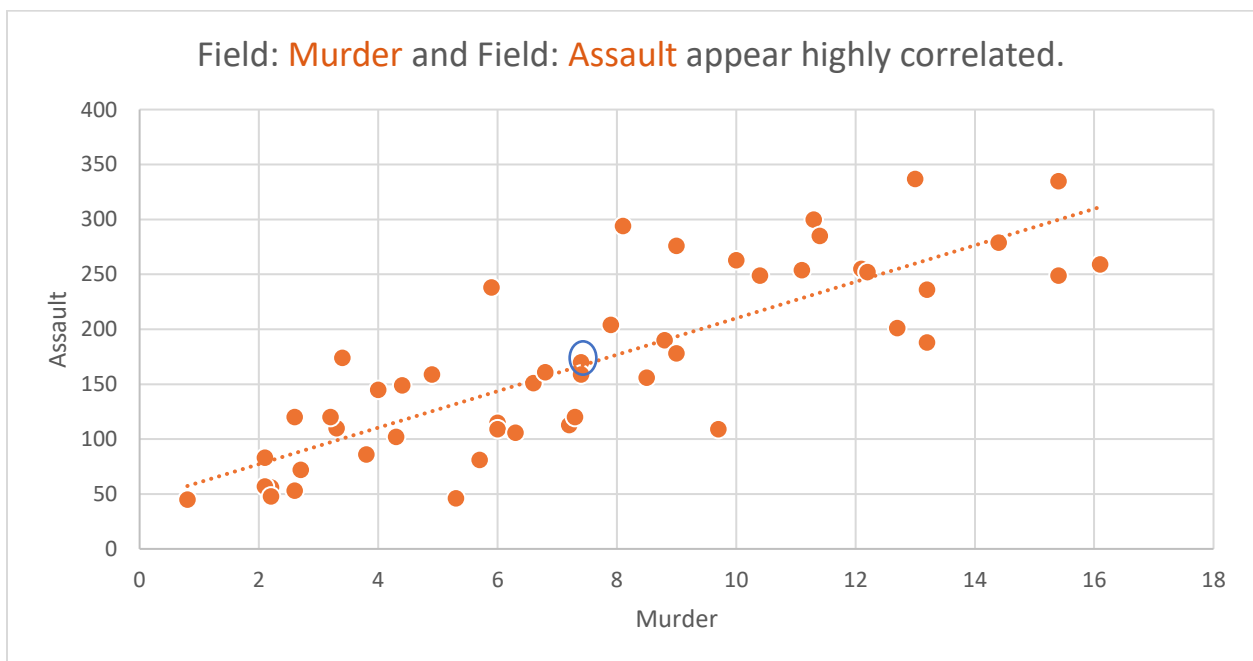


Scatter Plot of Murder rate vs Assault rate before outlier smoothed



The outlier in this data point is highlighted in red, its Georgia based on the murder rate. What's interesting about this data point is if you remove the first digit from the murder rate it falls almost perfectly on the linear correlation line between Assault rate and Murder rate. Based on this I will act as though the Georgia Murder rate, which was listed as the highest in the country, is a typo and remove the first digit, adjusting it from 17.4 to 7.4.

Scatter Plot of Murder rate vs Assault rate after outlier smoothed (adjusted point circled)



### SQL Table for Min, Max, Mean, Variance of Murder rate

	min(murder)	max(murder)	avg(murder)	variance(murder)
▶	0.8	17.4	7.787999999999999	18.591056000000005

### SQL Table for Min, Max, Mean, Variance of Assault rate

	min(Assault)	max(Assault)	avg(Assault)	variance(Assault)
▶	45	337	169.8800	6773.3856

### SQL Table for Min, Max, Mean, Variance of Urban Pop

	min(UrbanPop)	max(UrbanPop)	avg(UrbanPop)	variance(UrbanPop)
▶	32	91	65.5400	205.32839999999996

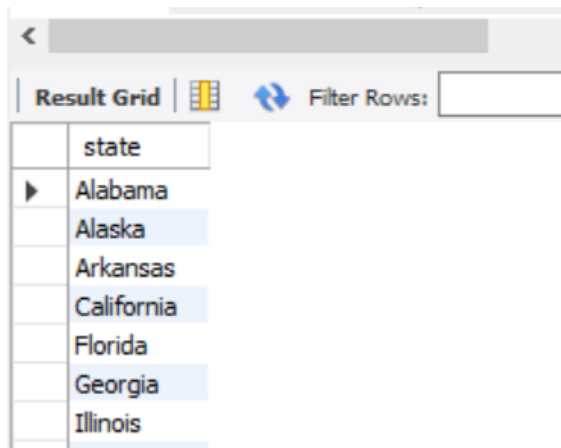
### State with the highest Murder rate

	state
▶	Georgia

### States Listed by Urban Pop Size (ascending)

	state
▶	Vermont
	West Virginia
	Mississippi
	North Dakota
	North Carolina
	South Dakota
	Alaska

States with a higher Murder rate than Arizona



The screenshot shows a web-based data interface. At the top, there is a search bar with a left arrow and a 'Filter Rows:' button with a double arrow icon. Below this is a table with a single column labeled 'state'. The table contains the following states: Alabama, Alaska, Arkansas, California, Florida, Georgia, and Illinois. The states are listed in a light blue background, and the first state, Alabama, has a small right-pointing triangle next to it.

state
Alabama
Alaska
Arkansas
California
Florida
Georgia
Illinois

#### 4. Conclusions

In conclusion after examining the data using both excel and MySQL we can establish that there is, in this dataset, a linear relationship between murder rate and assault rate. From this linear relationship we were able to identify and smooth one outlier in the dataset. When examining the distribution of the assault rate we can also see that this dataset is not distributed in a bell-shaped fashion and the data may need to be normalized further before performing analytics with it.

## Problem #2

### 1. Purpose

Examine the child mortality dataset to see if relationships can be established between the different mortality rates listed in the dataset.

### 2. Methodology

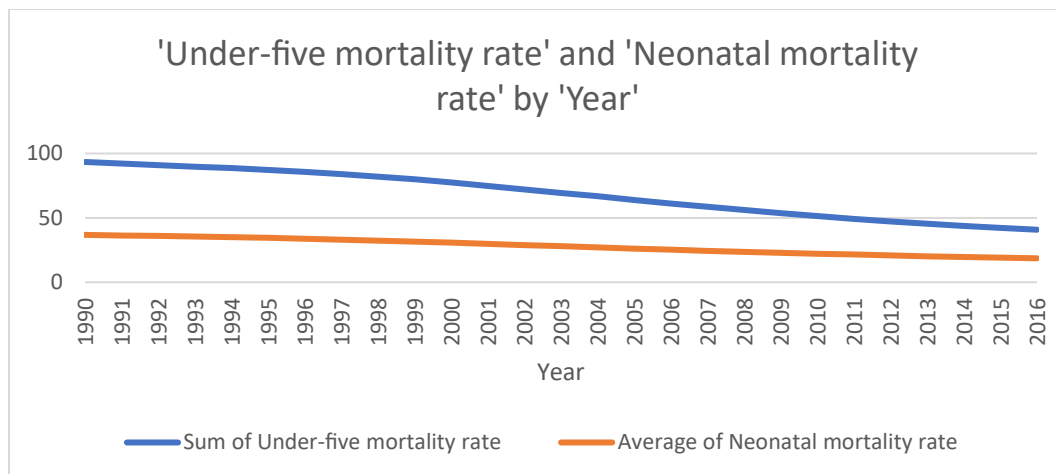
The child mortality dataset needed some work done on it before any analysis could be performed. When first downloaded the dataset had a mix of formats. The header line was comma delimited and the data was space delimited, this presented challenges for excels data import functionality. Due to the small number of attributes in this dataset I was able to manually remove the commas in the header line and add quotations around each column name as to not break the space delimited format. These changes allowed excel to easily import the data into a table. Once the data was imported the numeric columns were converted from text to numeric and the missing values were ready to be filled. To fill the missing values column averages were first used, but upon examining the resulting data column averages were deemed to be not a good method of filling missing values. The missing values were then replaced with excels linear fill tool, which examines the values directly above and below missing values and defines a step count that can be used to fill the missing values between them. This tool is very useful for linear datasets, which this dataset appears to be. Once the missing values were filled using this technique graphs were created to visualize the relationship between Under-five mortality rate and neonatal mortality rate, Infant mortality rate and neonatal mortality rate, and Year and infant mortality rate. After the graphing was complete new columns were created categorizing each of the mortality rates listed in the dataset. The logic of the Infant\_Cat column is as follows, if the mortality rate is less than 45 then it is assigned category "Low Mortality", if the rate is between 45 and 70 it is assigned category "Medium Mortality", and if the rate is above 70 then it is assigned category "High Mortality". The same three categories are used for the other two mortality rates in the dataset, with the only difference being the number ranges used to assign the categories.

The excel document was then saved as a csv file and uploaded to MySQL. First all missing values were replaced with the average of their respective columns. Then queries were used to find out the years with the highest and lowest infant mortality rates, in which years neonatal mortality was above average, and the infant mortality rates sorted in descending order. Numeric analysis was also done on the attributes calculating each attributes minimum value, maximum value, variance, and standard deviation. A new column was added to the database called "Above-Five Mortality Rate" and populated with values based on the opposite values of the "Under-Five Mortality Rate". This was done by subtracting the "Under-Five Mortality Rate" column from 100 for each year in the dataset. The resulting data was then added to the "Above-Five Mortality Rate" column.

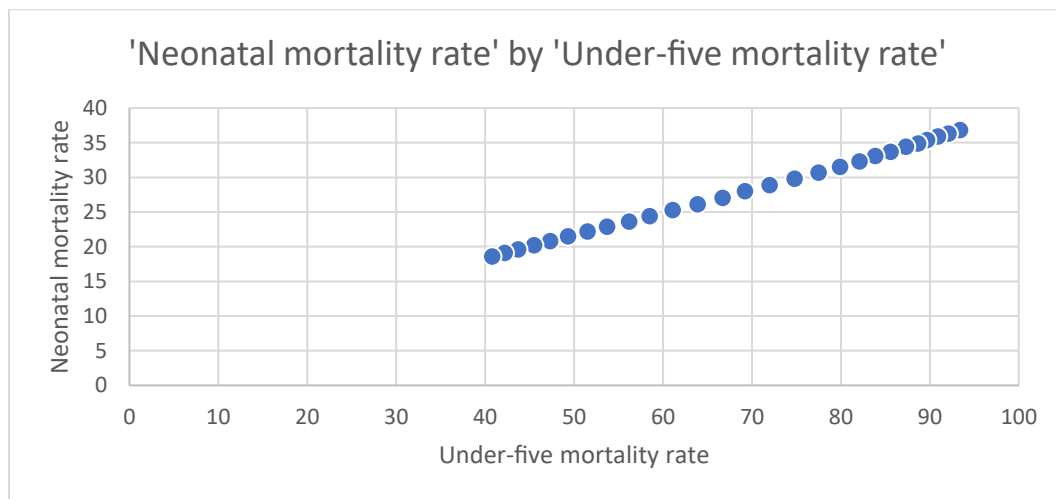
The child mortality table was then exported using MySQL workbenches Data Export Wizard to a JSON file on my laptop. I then used the website [https://www.oxygenxml.com/xml\\_json\\_converter.html](https://www.oxygenxml.com/xml_json_converter.html) to convert the JSON file to an XML file. I then saved the XML file to my laptop as well.

### 3. Data

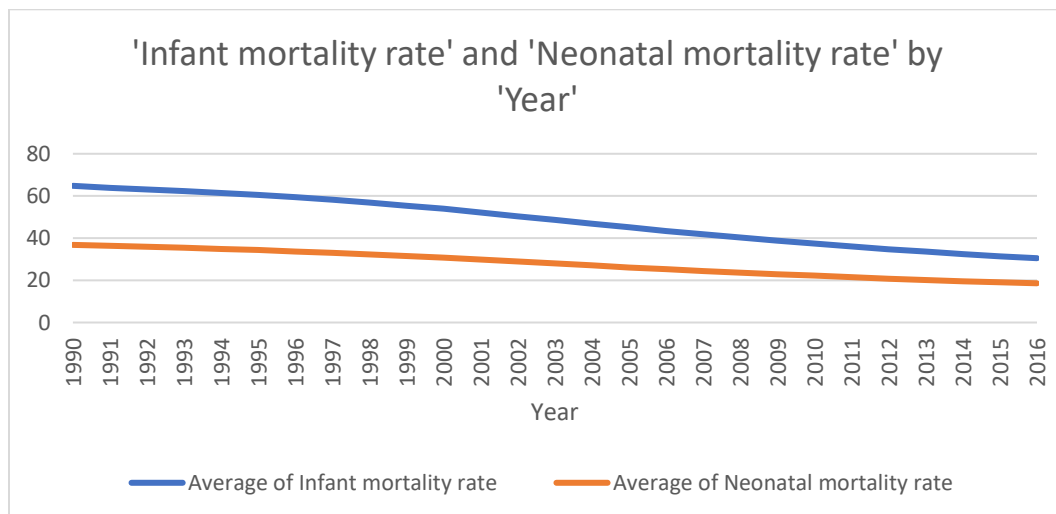
Under-Five Mortality Rate and Neonatal Mortality Rate by Year



Scatter of Neonatal vs Under-Five

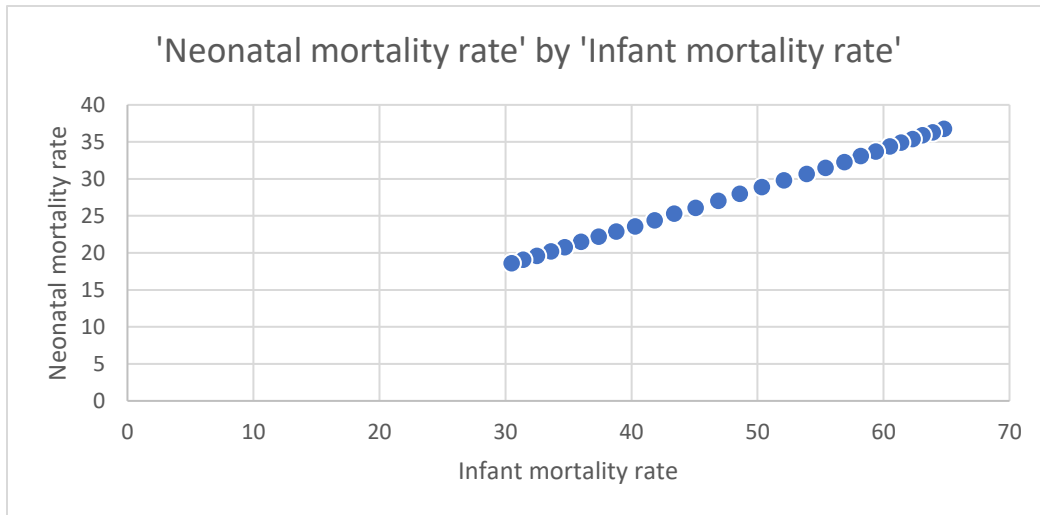


Infant and Neonatal plotted by Year

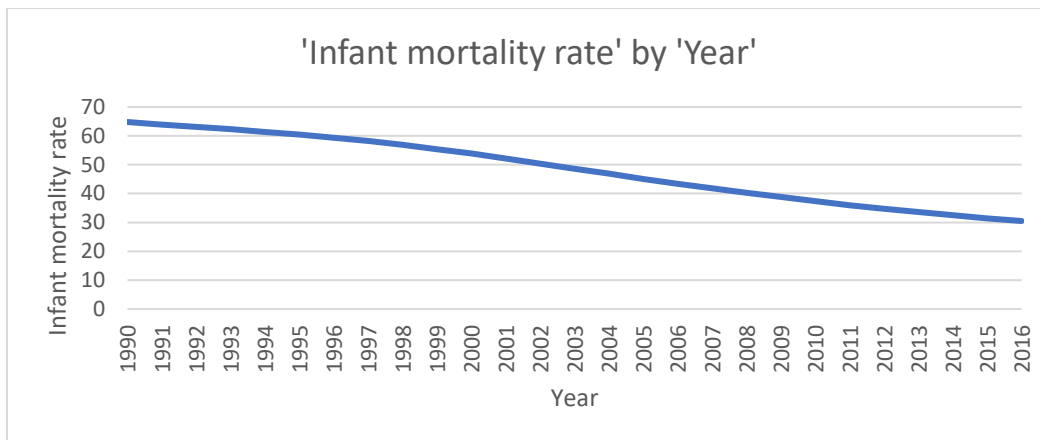




## Infant vs Neonatal Scatter



## Infant Plotted by Year



(Full Tables available in MySQL)

## SQL Table for Year with the Highest Infant Mortality Rate

Result Grid | Filter Rows:

Max number of infant mortalities
1990

## SQL Table for the Year with the Infant Lowest Mortality Rate

Result Grid | Filter Rows:

Min number of infant mortalities
2016

## SQL Table for the Number of Years the Neonatal Rate was Above Average

Result Grid | Filter Rows: | Export

Years the neonatal rate was above average
1990
1991
1992
1993
1995
1996

child\_mortality 3 x

## Full Dataset Sorted by Infant Mortality Rate Descending

Result Grid | Filter Rows: | Export: | Wrap Cell Contents: [fA](#)

Year	Under-Five mortality rate	Infant mortality rate	Neonatal mortality rate	Above-Five Mortality Rate
1990	93.4	64.8	36.8	6.6
1991	92.1	63.9	36.3	7.9
1992	90.9	63.1	35.9	9.1
1993	89.7	62.3	35.4	10.3
1994	88.7	61.4	25.5962962963	11.3
1995	87.3	60.5	34.4	12.7

## Numeric analysis for Infant Mortality Rate

Result Grid | Filter Rows: | Export: | Wrap Cell Contents: [fA](#)

min('Infant mortality rate')	max('Infant mortality rate')	avg('Infant mortality rate')	variance('Infant mortality rate')	std('Infant mortality rate')
30.5	64.8	48.50205761316...	117.4205002624939	10.836074024410035

## Numeric analysis for Neonatal Rate

Result Grid | Filter Rows: | Export: | Wrap Cell Content: [fA](#)

min('Neonatal mortality rate')	max('Neonatal mortality rate')	avg('Neonatal mortality rate')	variance('Neonatal mortality rate')	std('Neonatal mortality rate')
18.6	36.8	27.492318244170104	33.5395030492566	5.791329989670473

## Numeric analysis for Under-Five Mortality Rate

Result Grid | Filter Rows: | Export: | Wrap Cell Contents: [fA](#)

min('Under-Five mortality rate')	max('Under-five mortality rate')	avg('Under-Five mortality rate')	variance('Under-Five mortality rate')	std('Under-Five mortality rate')
40.8	93.4	67.82716049382...	286.4582472184122	16.92507746565469

#### 4. Conclusions

From examining the data and creating graphs to compare how each mortality rate is trending through the years recorded in the dataset it is clear that each mortality rate is decreasing as we move closer to the present. When the mortality rates were categorized this became even more easy to see as the oldest years in the database all have category “High Mortality” and the latest years all have the category “Low Mortality”.