Cal Colistra Github link: https://github.com/CalColistra/IDS/tree/main/Project2

Introduction to Data Science

Project 2

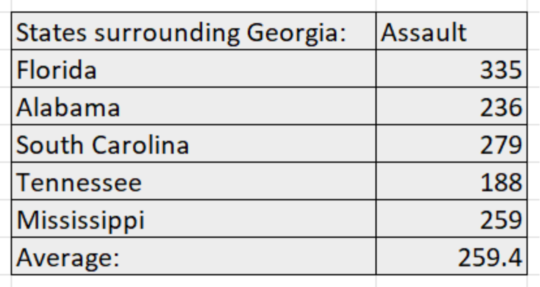
2/13/2022

Purpose of Project 1:

To analyze data containing statistics in arrests per 100,000 residents. The data shows assault and murder rates for each of the 50 US states in 1973. Also to analyze how the percentage of population that lives in urban areas may have an effect on assault and murder rates.

Methodology for Problem 1:

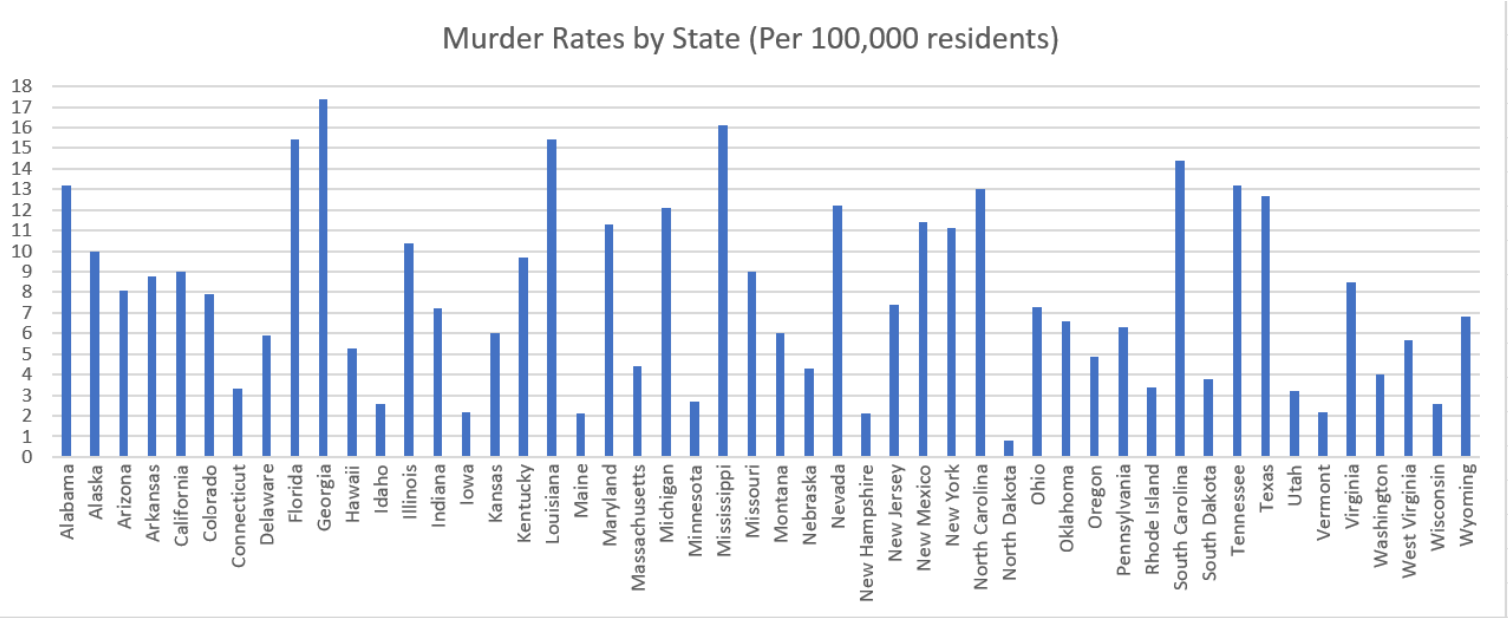
The data was provided by my professor, Dr. Forouraghi. The method used to collect the data is unknown. To cleanse the data, there was a missing value that needed to be replaced (Assault in Georgia). I decided to replace this missing value with the average assaults of 5 states that closely surround Georgia (Florida, Alabama, South Carolina, Tennessee, and Mississippi). The chart here shows how I calculated this, also note that I rounded the 259.4 to a 259 to match the decimal format of the rest of the values. I did not see any noisy data or outliers that needed to be addressed. The model was built to plot a bar graph of murder rates for all 50 states, a histogram of assaults, a scatter plot of murder rates vs. assault rates, and two bar graphs that form a relationship between urban population percentage and assault and murder rates.



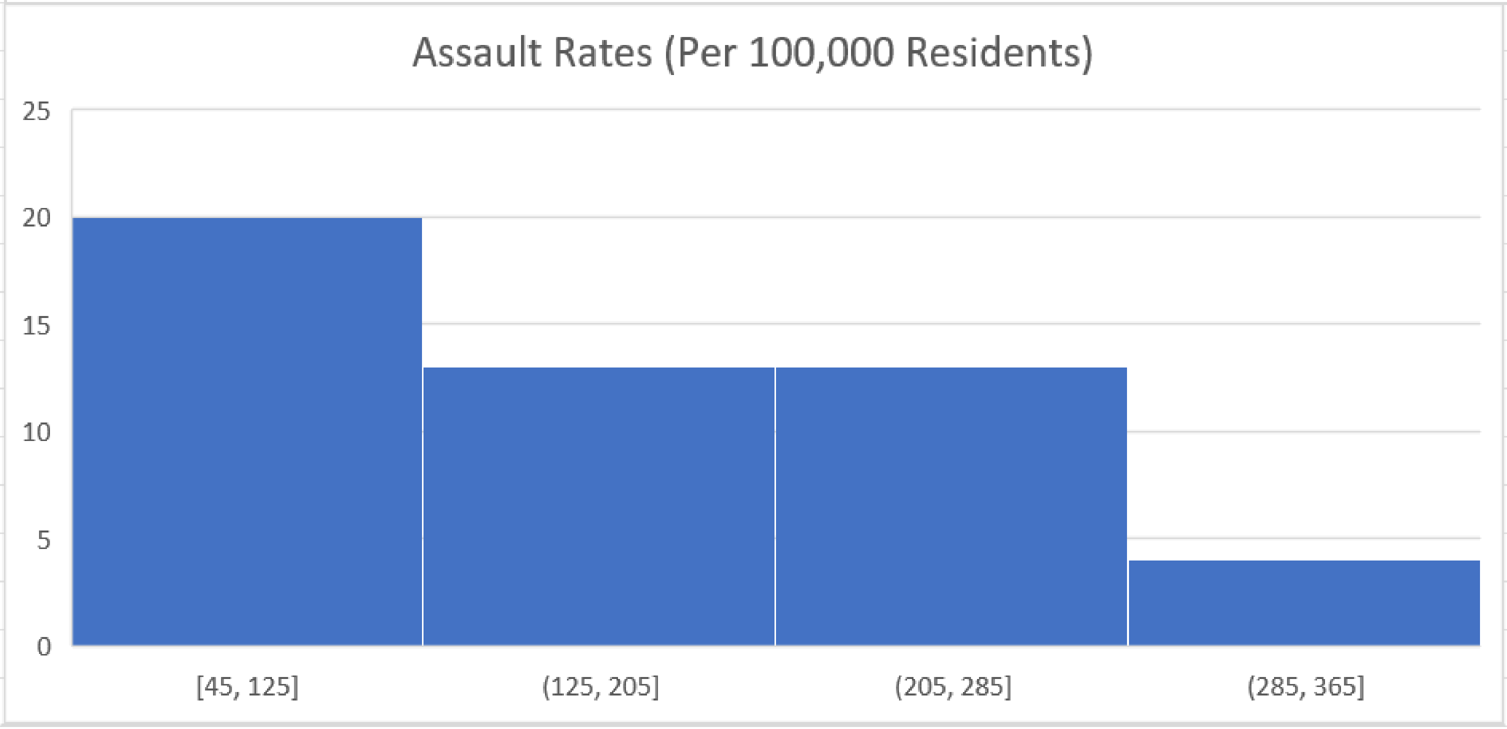
States surrounding Georgia:	Assault
Florida	335
Alabama	236
South Carolina	279
Tennessee	188
Mississippi	259
Average:	259.4

Summary for Problem 1:

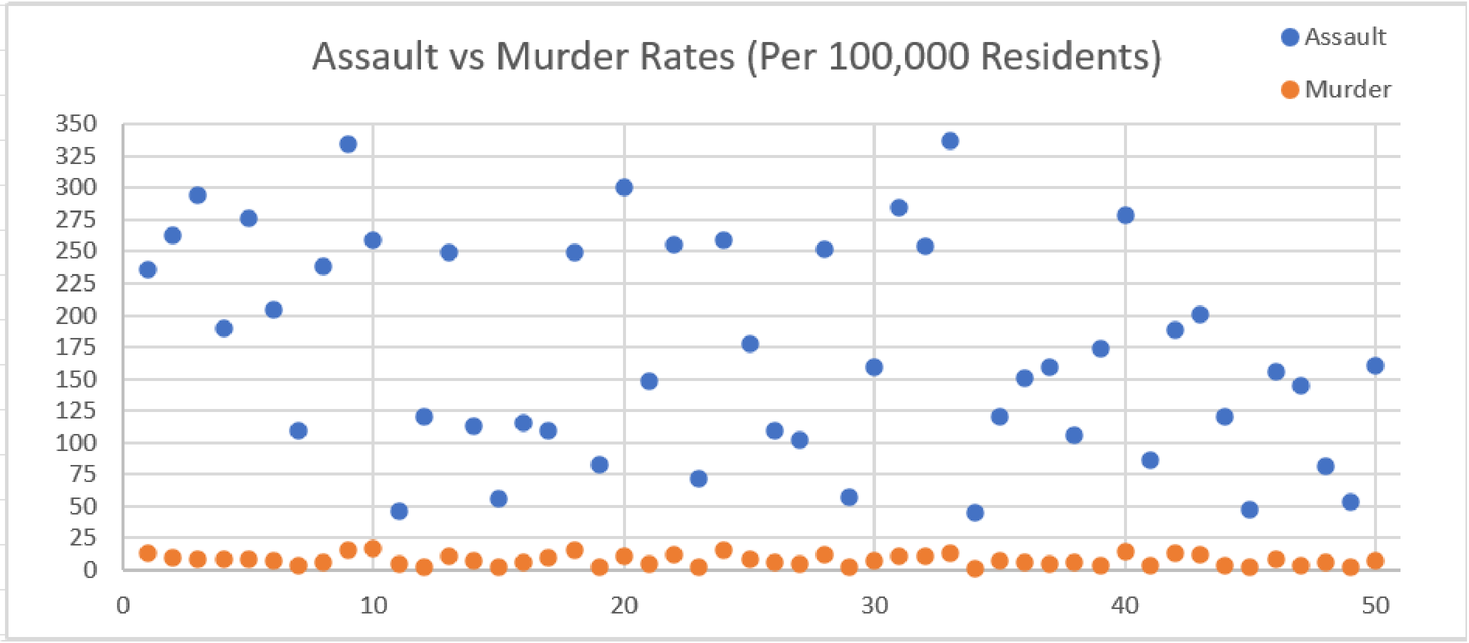
Bar graph of murder rates for all 50 states:



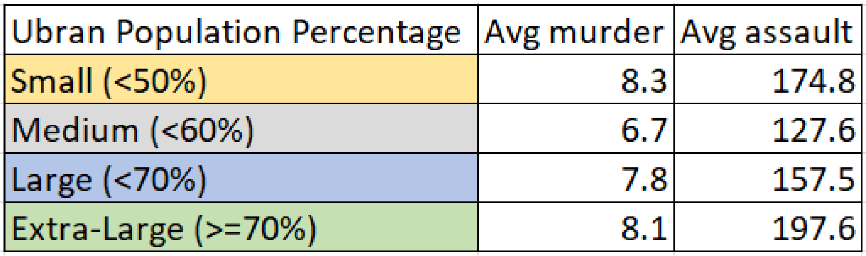
Histogram of assaults: Note that the x-axis shows groups of states in which their assault rate lies in the designated ranges. The y-axis shows the amount of states that lay within these designated ranges.



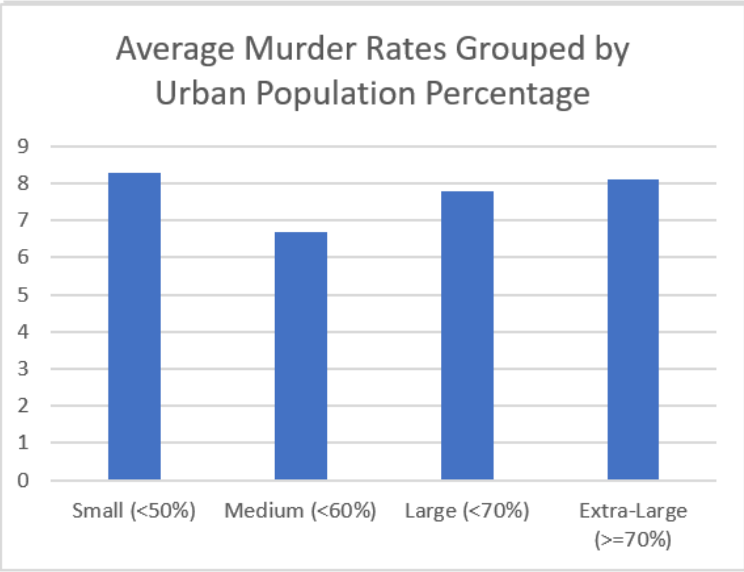
Scatter plot of murder rates vs. assault rates:

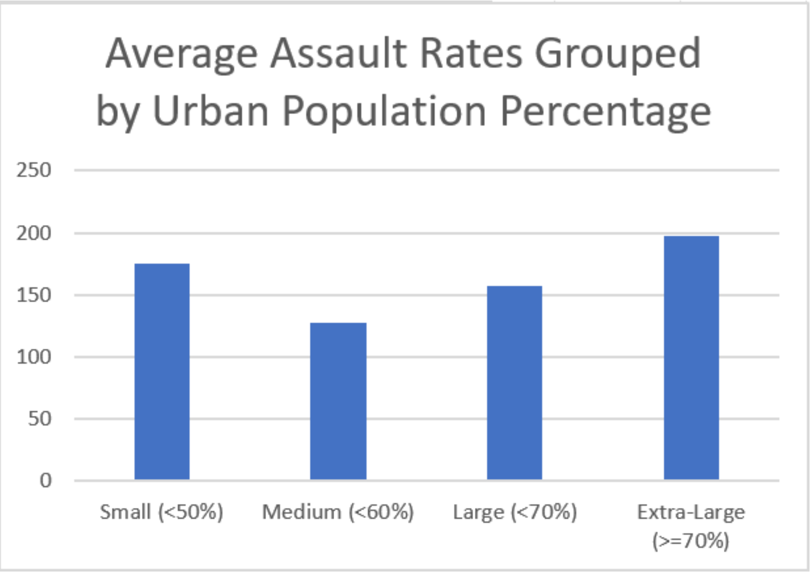


Two bar graphs that form a relationship between urban population percentage and assault and murder rates:



Ubran Population Percentage	Avg murder	Avg assault
Small (<50%)	8.3	174.8
Medium (<60%)	6.7	127.6
Large (<70%)	7.8	157.5
Extra-Large (>=70%)	8.1	197.6





Conclusion for Problem 1:

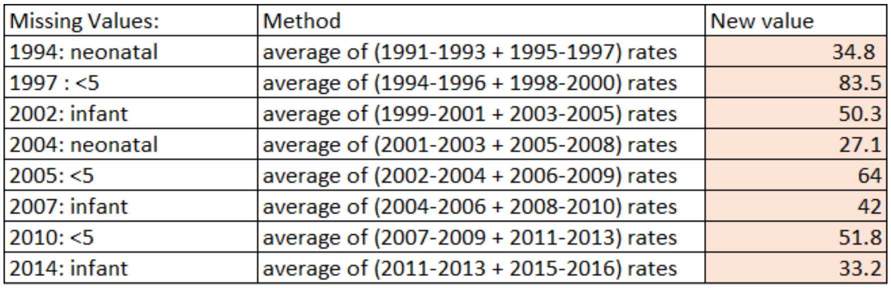
Based on the **Bar graph of murder rates for all 50 states** it can be noted that Georgia has the highest murder rate of all 50 states and North Dakota has the lowest. Based on the **Histogram of assaults** it can be noted that most states have assault rates in the range of 45 to 125. The **Scatter plot of murder rates vs. assault rates** shows how, across the US, murder rates are consistently lower than assault rates. Lastly, the **Two bar graphs that form a relationship between urban population percentage and assault and murder rates** shows how there is not a clear relationship between urban population percentage and murder/assault rates. This is because there is not much a difference in assault and murder rates when they are sorted into urban population percentages. For example, the average murder rates for states which have small urban population percentage is not very much different from states with a large or extra large urban population percentage. This is very similar for assault rates because the average assault rates for states with extra-large urban population percentages is not much different that states with small urban population percentages.

Purpose of Problem 2:

To analyze data regarding child mortality rates from 1990 to 2016. The data consists of three categories: children under five years old, infants, and neonatal. Also there were a number of missing values that needed to be replaced.

Methodology for Problem 2:

The data was provided by my professor, Dr. Forouraghi and it was inspired by data collected from UNICEF. In order to cleanse the data I had to replace the missing values shown in this chart below. For each missing value I decided to take the average of the rates 3 years before and 3 years after the year of the missing value. For example, to replace the missing value in 1994 for neonatal rates, I took the average of rates in years 1991-1993 and 1995-1997

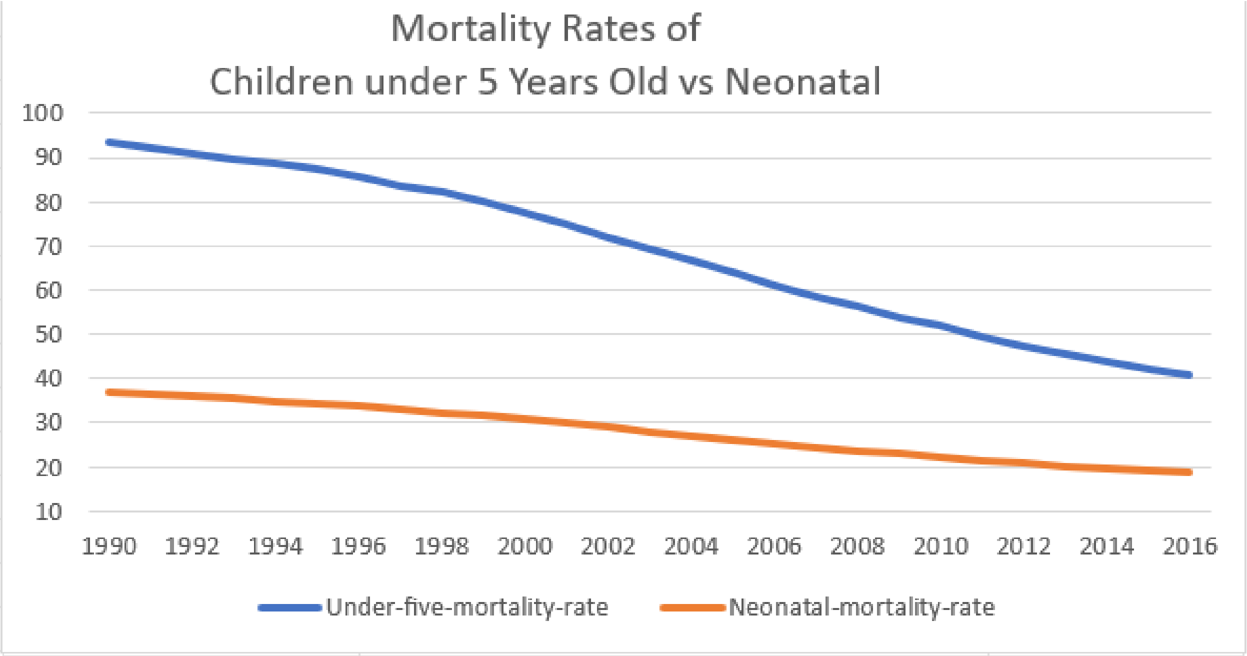


Missing Values:	Method	New value
1994: neonatal	average of (1991-1993 + 1995-1997) rates	34.8
1997 : <5	average of (1994-1996 + 1998-2000) rates	83.5
2002: infant	average of (1999-2001 + 2003-2005) rates	50.3
2004: neonatal	average of (2001-2003 + 2005-2008) rates	27.1
2005: <5	average of (2002-2004 + 2006-2009) rates	64
2007: infant	average of (2004-2006 + 2008-2010) rates	42
2010: <5	average of (2007-2009 + 2011-2013) rates	51.8
2014: infant	average of (2011-2013 + 2015-2016) rates	33.2

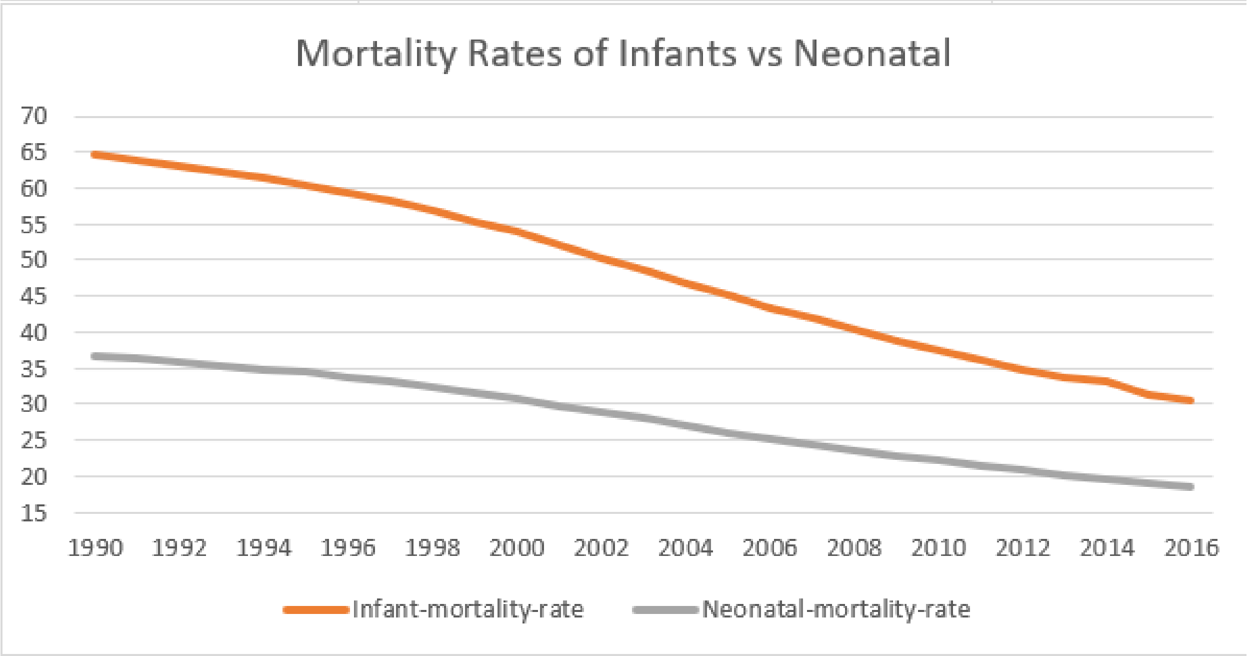
and inserted the new value in 1994. I did not see any noisy data or outliers that needed to be addressed. The model was built to show the relationship between children under five years old and neonatal via a line graph, the relationship between infants and neonatal mortality rates via a line graph, and another relationship between the years and infant mortality rates via a line graph.

Summary for <u>Problem 2</u>:

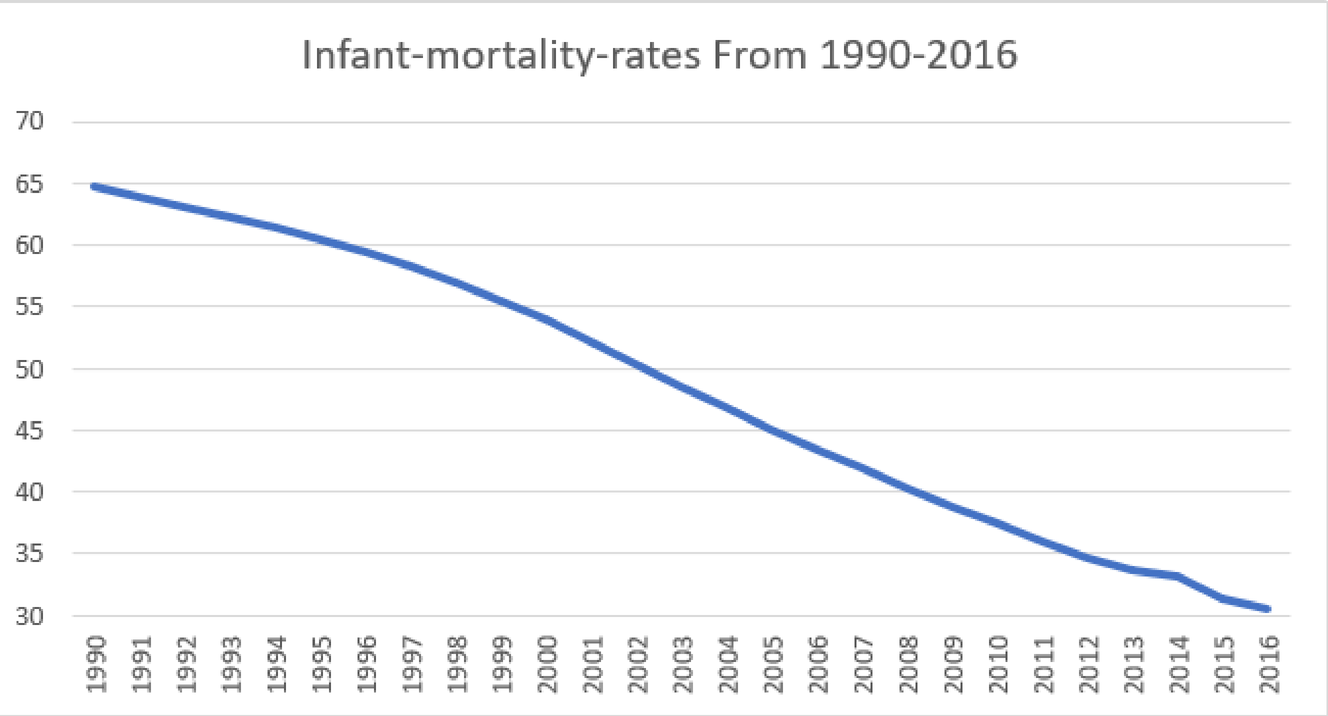
Relationship between mortality rates of children under five years old and neonatal:



Relationship between infants and neonatal mortality rates:



Relationship between the years and infant mortality rates:



Conclusion for Problem 2:

Based on the line graph showing the relationship between mortality rates of children under five years old and neonatal it can be concluded that during the years between 1990 and 2016, mortality rates for children under five years old was generally higher than the mortality rates of neonatal. It may also be important to note that although children under five had a higher mortality rate, their rate decreased more than neonatal rates decreased. In the line graph showing the relationship between infants and neonatal mortality rates, it can be seen that infant mortality rates were generally higher than neonatal mortality rates over the specified years. Lastly, in the line graph showing the relationship between the years and infant mortality rates, it can be concluded that there was a significant decrease in mortality rates because in 1990 their mortality rate was at about 65 and it steadily decreased until 2016 where the rate was about 30.

Sources:

https://sebhastian.com/mysql-median/#:~:text=To%20find%20the%20median%20value%20using%20a%20MySQL%20query%2C%20you,1%3B%20SELECT%20AVG(subq.

https://www.mysqltutorial.org/mysql-standard-deviation/

https://www.oxygenxml.com/xml_json_converter.html