Unemployment Rate and School Shooting Casualties

Group 4: Jay Park, Ryan Huett, Lorenzo Murillo IV
EE 381 Sec 02: Probability and Statistics of Computing
December 16, 2021

Contributions

- Jay Park (33%)
    - Wrote the program to merge the input data
    - Wrote the Analysis and Findings
- Ryan Huett (33%)
    - Interpreted data output
    - Wrote conclusion of our hypothesis
- Lorenzo Murillo IV (33%)
    - Wrote the program to compute the needed values for the regression.
    - Wrote the Project Outline along with the formulas
    - Produced the graphs and code output

Introduction: We perform a linear regression analysis that examines unemployment percentage rates as well as the number of casualties in school shootings in the last 21 years with data provided by Kaggle.com and the Bureau of Labor Statistics. We hypothesize that there is a positive relationship between the two, specifically that higher unemployment rates contribute to a decrease in mental health, increasing the likelihood and frequency of school shootings.

Objectives:

- To apply a linear regression model to determine the relationship between two variables.
- To analyze a socioeconomic effect on society and determine if higher unemployment rates can be attributed to an increase or decrease in school shootings. We will then make a conclusion on whether the given data set supports a linear regression model.

Problem Description: School shootings have become more frequent as of recent, with almost 200 casualties in 2021 alone. In this project, we attempt to determine if our data regarding unemployment can be linked to an increase or decrease in school shootings. In our analysis we will use the rate of unemployment as our independent variable and the casualties of a school shooting as our dependent variable. The data set will be compiled in a java program and then interpreted by using a linear regression model. We seek to determine if there is any true relationship between these two social occurrences. We hypothesize that there is a positive relationship between unemployment rate and school shootings where b1 > 0.

Key Variables:

$n = sample\ size$

$$\bar{x} = \frac{\Sigma x_i}{n}$$

$$\bar{y} = \frac{\Sigma y_i}{n}$$

$$\Sigma x_i \Rightarrow \sum_{i=0}^{n} x_i = (x_1 + x_2 + \ldots + x_n)$$

$$\Sigma y_i \Rightarrow \sum_{i=0}^{n} y_i = (y_1 + y_2 + \ldots + y_n)$$

$$\Sigma x_i^2 \Rightarrow \sum_{i=0}^{n} (x_i^2) = (x_1^2 + x_2^2 + \ldots + x_n^2)$$

$$\Sigma y_i^2 \Rightarrow \sum_{i=0}^{n} (y_i^2) = (y_1^2 + y_2^2 + \ldots + y_n^2)$$

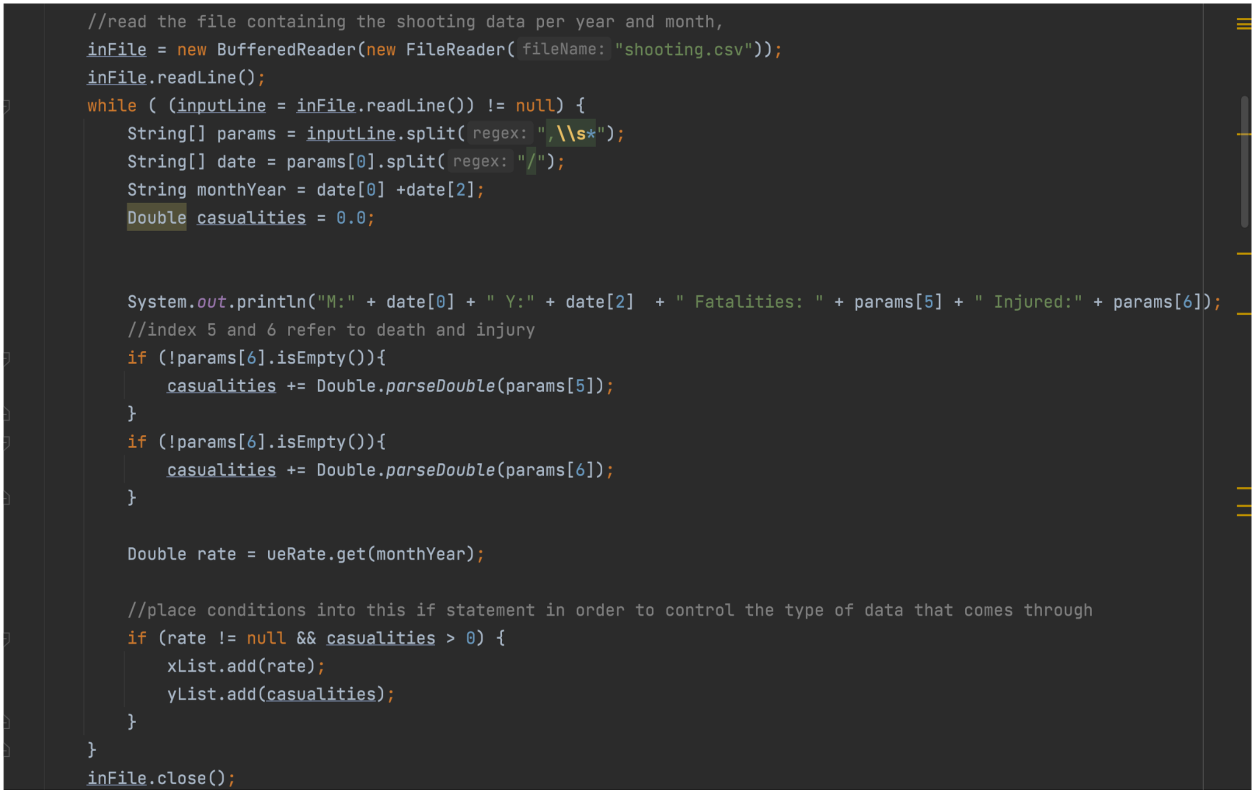
$$\Sigma x_i y_i \Rightarrow \sum_{i=0}^{n} x_i y_i = (x_1 y_1 + x_2 y_2 + \ldots + x_n y_n)$$

Formula	Description
$S_{xy} = \sum x_i y_i - (n \cdot \overline{y} \bullet \overline{x})$	Covariance of X and Y: Sum of the product of the difference between x and its means and the difference between y and its means.
$S_{xx} = \sum x_i^{2} - (n \cdot \overline{x}^{2})$	Sum of Squares: Sum of the squares of the difference between each x and the mean x value.
$\beta_1 = \frac{S_{xy}}{S_{xx}}$	Regression Coefficient: slope of linear regression model
$\beta_0 = \overline{y} - (\beta_1 \bullet \overline{x})$	Regression Constant: Intercept of linear regression equation
$SSE = \frac{RSS}{\sqrt{S_{xx}}} \Longrightarrow \frac{\sum[y_i - \left(\beta_0 + (\beta_1 \bullet x_i)\right))]^{2}}{\sqrt{S_{xx}}}$	Sum of Squares Error: Measure of how much variation in y is left unexplained by model
$SSR = [(\beta_0 + \left(\beta_1 \cdot \sum x_i\right)) - \overline{y}]^{2}$	Sum of Squares Residuals: Additional amount of explained variability in y due to the regression model.
$SST = [\sum y_i - \overline{y}]^{2}$	Maximum Sum of Squares: Max sum of errors for data
$r = \frac{S_{xy}}{\sqrt{S_{xx}}} \bullet \boxed{\sqrt{\sum y_i^{2} - \left(n \cdot \overline{y}^{2}\right)}}$	Coefficient of Correlation: Relation between x and y values
$r^{2} = \frac{SSR}{SST}$	Coefficient of Determination: Square of correlation (r) between predicted y scores and actual y scores
$var = \frac{SSE}{n-2}$	Variance: Measure of how far observed values are from the average of predicted values
$se\left(\beta_1\right) = \frac{\sqrt{var}}{\sqrt{S_{xx}}}$	Standard Error of Beta: An estimate of the standard deviation of the sampling distribution of beta.
$y = \beta_0 + \beta_1 x_i$	Linear Regression Equation: Used to demonstrate the between a scalar response and one or more explanatory variables.

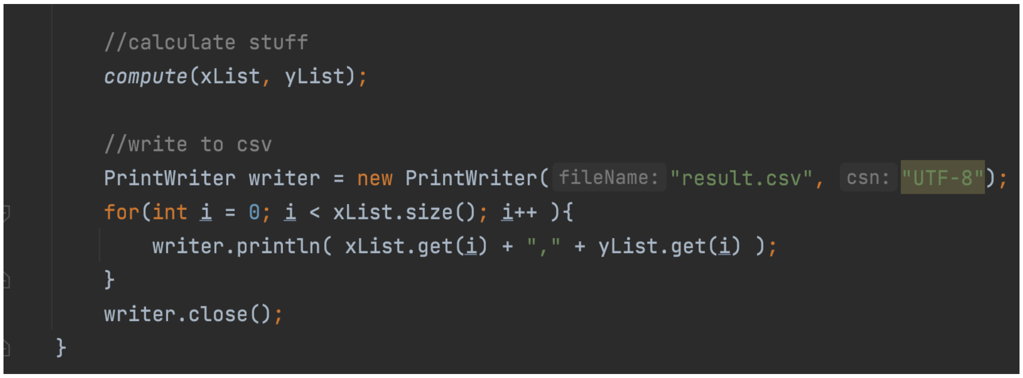
Data Used:

- Unemployment Rate from the Bureau of Labor Statistics
  - Link: https://www.bls.gov/charts/employment-situhttps://www.kaggle.com/ecodan/us-school-shootings-dataset/data?select=pah_wikp_combo.csvation/civilian-unemployment-rate.htm
  - Includes unemployment rates from Nov 2001 - Nov 2021 for different gender and race demographics
- School Shootings since
  - Link: https://www.kaggle.com/ecodan/us-school-shootings-dataset/data?select=pah_wikp_combo.csv
  - Includes each individual school shooting instance with the location, education level, number of deaths, number of injuries, and a description about the event

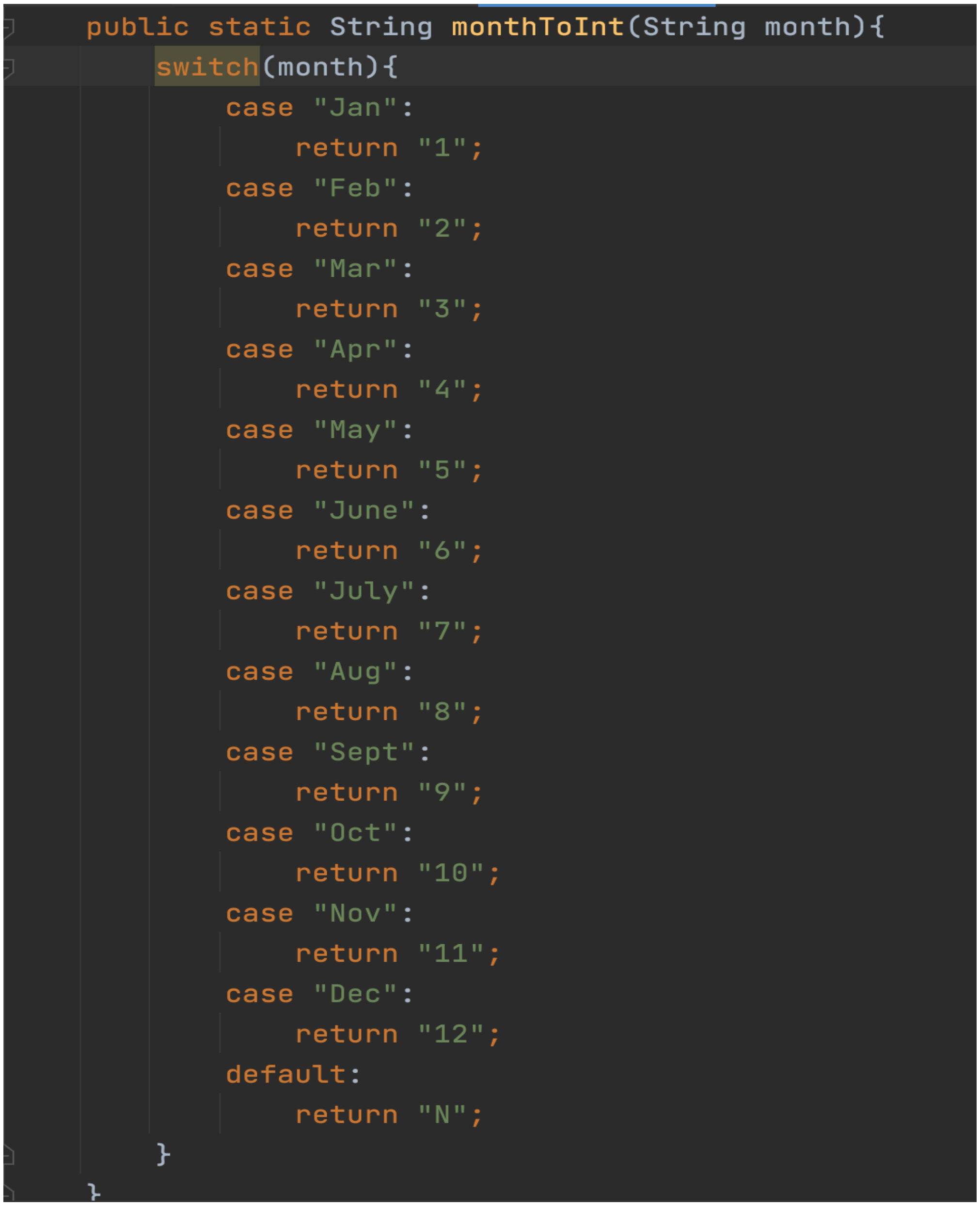
Java Code: We perform the majority of the analysis in Java. We take two datasets, the one with the compiled shooting data from Kaggle, and the unemployment rates for different age groups and races from the Bureau of Labor Statistics. We match the unemployment rate for a specific month and year, to the month and year of the shooting incident. We use the helper function “monthToInt” to aid in the conversion. For each incident, we then plot the unemployment rate against the number of casualties (injuries + death). We also generate a csv file containing this data. Afterwards, we apply the formulas using the data to acquire our values. The instructions for the computation are stored in the function “compute()” which performs all of the formulas that were listed above on the data. The console output shows the values we obtain after computation. We can also apply conditions to our data. The data that we retrieved from Kaggle was stratified. It included each incident along with the number of deaths, number of injuries, education level, area, state, and city.



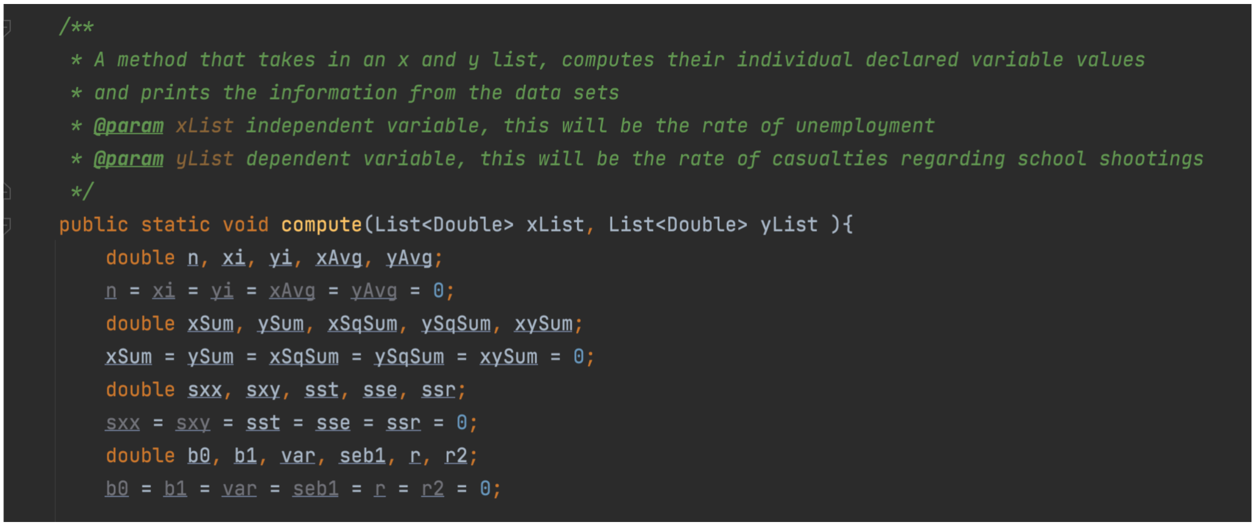
```
    //read the file containing the shooting data per year and month,
    inFile = new BufferedReader(new FileReader( fileName: "shooting.csv"));
    inFile.readLine();
    while ( (inputLine = inFile.readLine()) != null) {
        String[] params = inputLine.split( regex: ",\\s*");
        String[] date = params[0].split( regex: "/");
        String monthYear = date[0] +date[2];
        Double casualities = 0.0;


        System.out.println("M:" + date[0] + " Y:" + date[2]  + " Fatalities: " + params[5] + " Injured:" + params[6]);
        //index 5 and 6 refer to death and injury
        if (!params[6].isEmpty()){
            casualities += Double.parseDouble(params[5]);
        }
        if (!params[6].isEmpty()){
            casualities += Double.parseDouble(params[6]);
        }

        Double rate = ueRate.get(monthYear);

        //place conditions into this if statement in order to control the type of data that comes through
        if (rate != null && casualities > 0) {
            xList.add(rate);
            yList.add(casualities);
        }
    }
    inFile.close();
```



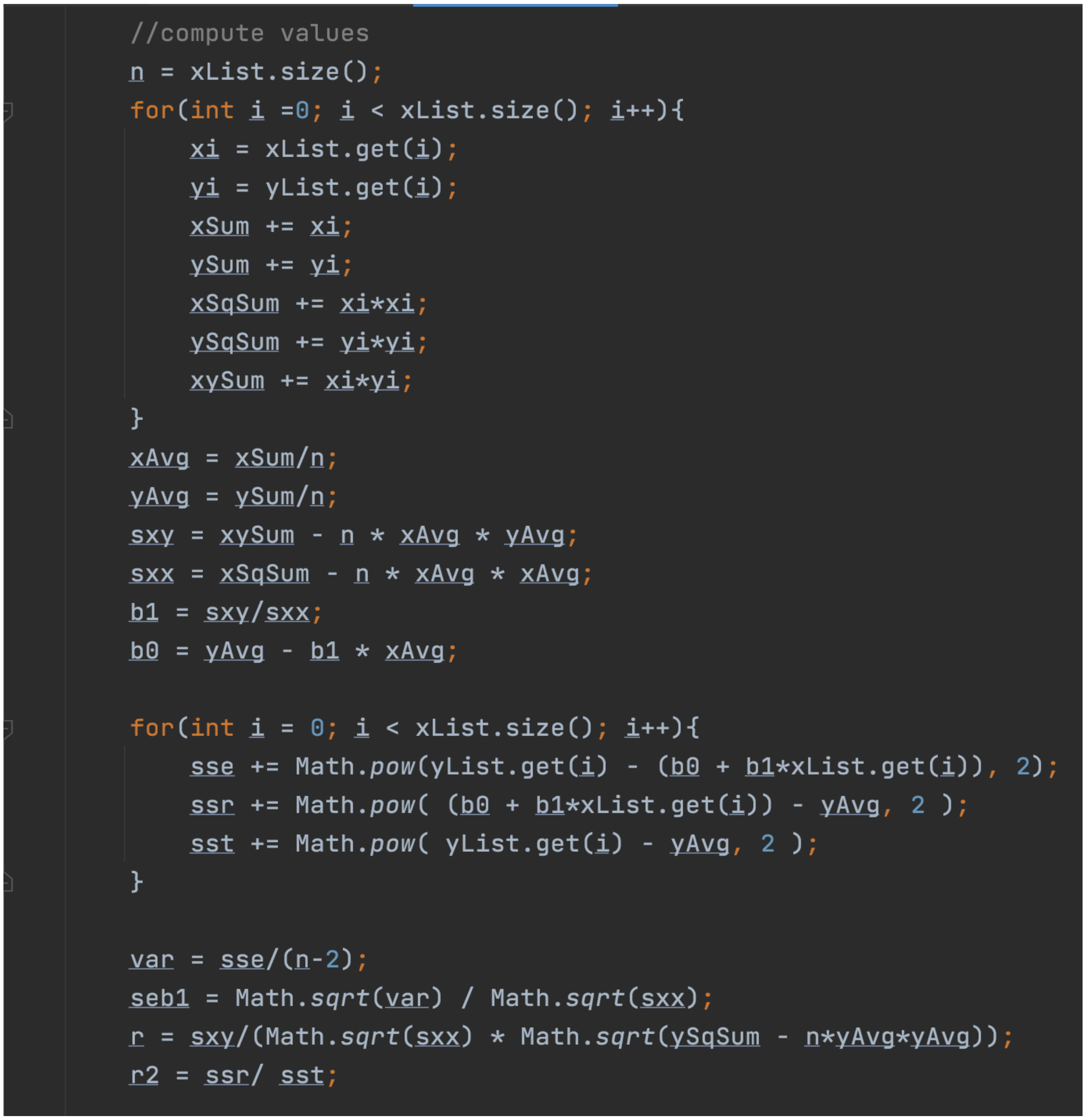
```
        //calculate stuff
        compute(xList, yList);

        //write to csv
        PrintWriter writer = new PrintWriter( fileName: "result.csv",  csn: "UTF-8");
        for(int i = 0; i < xList.size(); i++ ){
            writer.println( xList.get(i) + "," + yList.get(i) );
        }
        writer.close();
    }
```



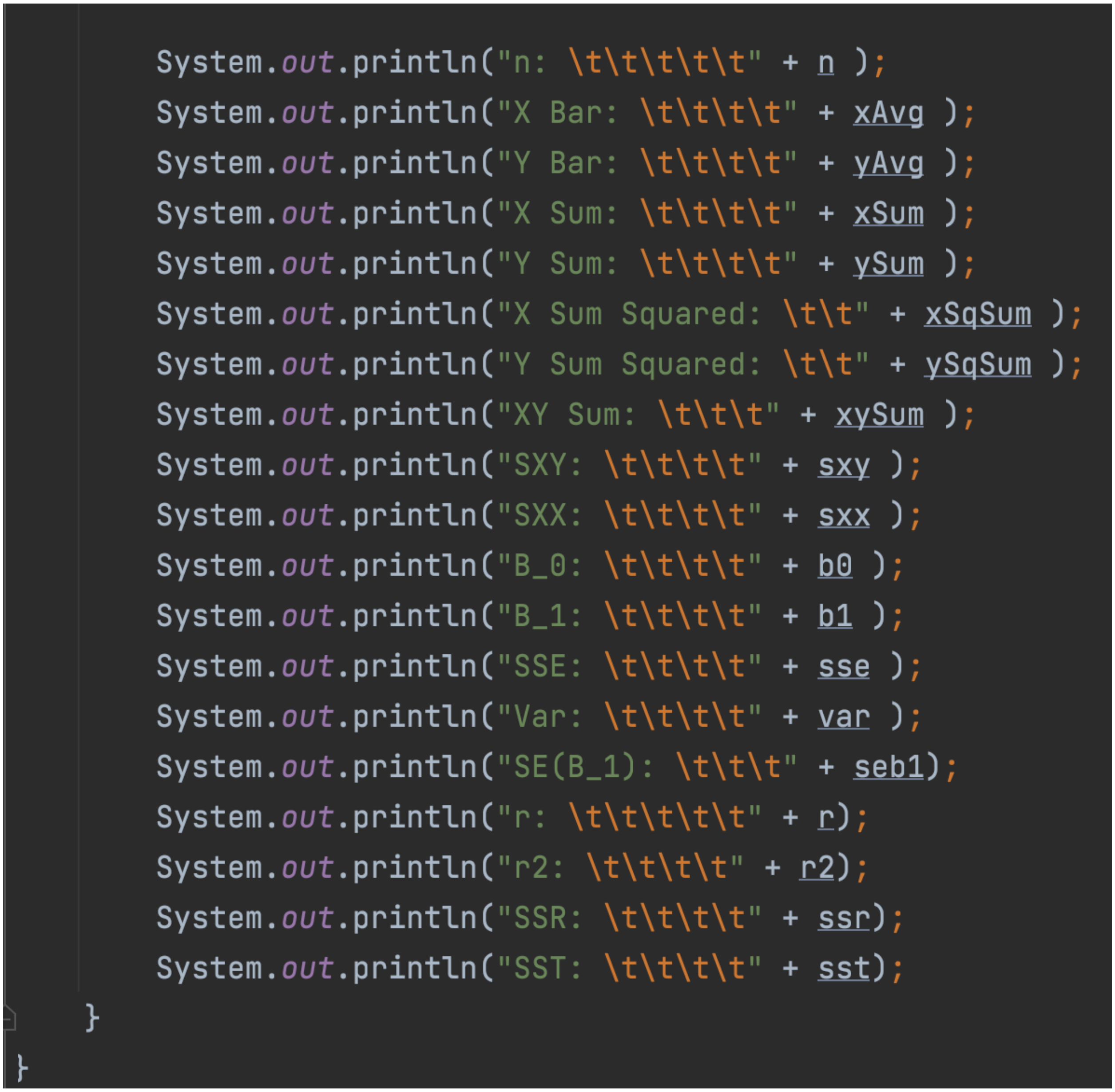
```java
    public static String monthToInt(String month){
        switch(month){
            case "Jan":
                return "1";
            case "Feb":
                return "2";
            case "Mar":
                return "3";
            case "Apr":
                return "4";
            case "May":
                return "5";
            case "June":
                return "6";
            case "July":
                return "7";
            case "Aug":
                return "8";
            case "Sept":
                return "9";
            case "Oct":
                return "10";
            case "Nov":
                return "11";
            case "Dec":
                return "12";
            default:
                return "N";
        }
    }
```



```
/**
 * A method that takes in an x and y list, computes their individual declared variable values
 * and prints the information from the data sets
 * @param xList independent variable, this will be the rate of unemployment
 * @param yList dependent variable, this will be the rate of casualties regarding school shootings
 */
public static void compute(List<Double> xList, List<Double> yList ){
    double n, xi, yi, xAvg, yAvg;
    n = xi = yi = xAvg = yAvg = 0;
    double xSum, ySum, xSqSum, ySqSum, xySum;
    xSum = ySum = xSqSum = ySqSum = xySum = 0;
    double sxx, sxy, sst, sse, ssr;
    sxx = sxy = sst = sse = ssr = 0;
    double b0, b1, var, seb1, r, r2;
    b0 = b1 = var = seb1 = r = r2 = 0;
```

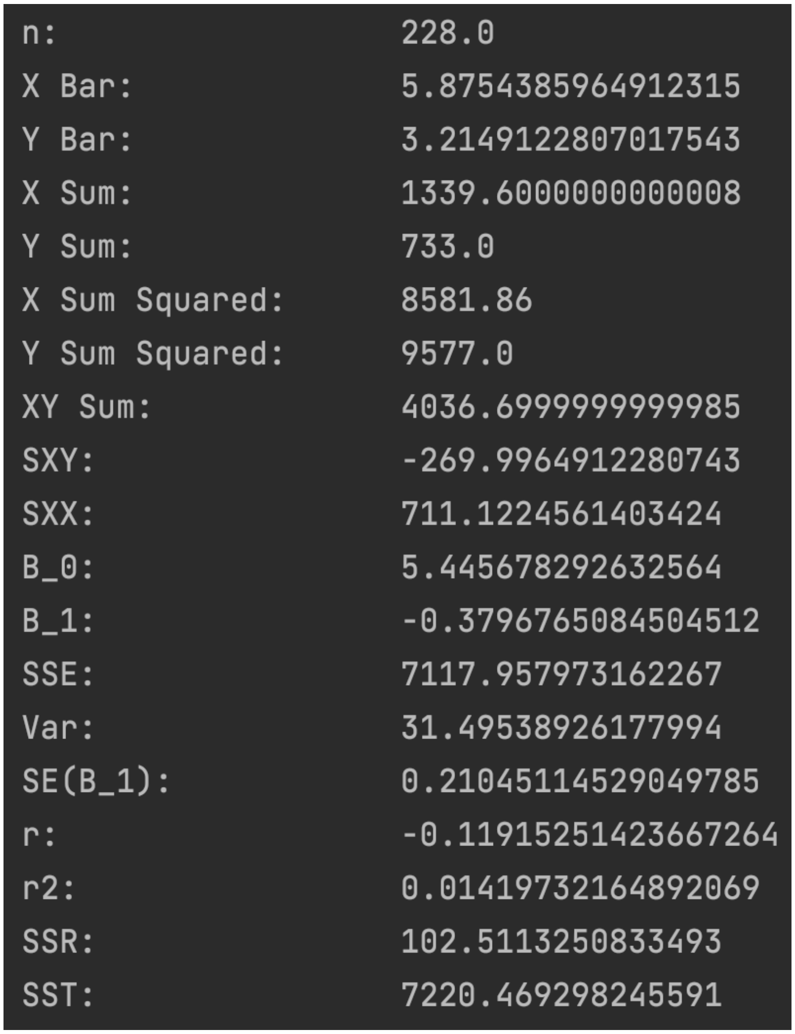


```
    //compute values
    n = xList.size();
    for(int i =0; i < xList.size(); i++){
        xi = xList.get(i);
        yi = yList.get(i);
        xSum += xi;
        ySum += yi;
        xSqSum += xi*xi;
        ySqSum += yi*yi;
        xySum += xi*yi;
    }
    xAvg = xSum/n;
    yAvg = ySum/n;
    sxy = xySum - n * xAvg * yAvg;
    sxx = xSqSum - n * xAvg * xAvg;
    b1 = sxy/sxx;
    b0 = yAvg - b1 * xAvg;

    for(int i = 0; i < xList.size(); i++){
        sse += Math.pow(yList.get(i) - (b0 + b1*xList.get(i)), 2);
        ssr += Math.pow( (b0 + b1*xList.get(i)) - yAvg, 2 );
        sst += Math.pow( yList.get(i) - yAvg, 2 );
    }

    var = sse/(n-2);
    seb1 = Math.sqrt(var) / Math.sqrt(sxx);
    r = sxy/(Math.sqrt(sxx) * Math.sqrt(ySqSum - n*yAvg*yAvg));
    r2 = ssr/ sst;
```



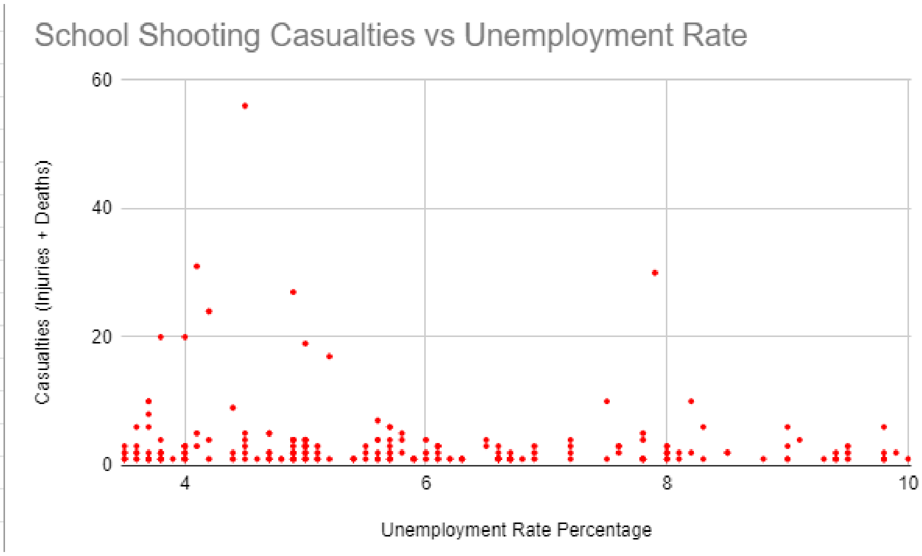
```
        System.out.println("n: \t\t\t\t\t" + n );
        System.out.println("X Bar: \t\t\t\t" + xAvg );
        System.out.println("Y Bar: \t\t\t\t" + yAvg );
        System.out.println("X Sum: \t\t\t\t" + xSum );
        System.out.println("Y Sum: \t\t\t\t" + ySum );
        System.out.println("X Sum Squared: \t\t" + xSqSum );
        System.out.println("Y Sum Squared: \t\t" + ySqSum );
        System.out.println("XY Sum: \t\t\t" + xySum );
        System.out.println("SXY: \t\t\t\t" + sxy );
        System.out.println("SXX: \t\t\t\t" + sxx );
        System.out.println("B_0: \t\t\t\t" + b0 );
        System.out.println("B_1: \t\t\t\t" + b1 );
        System.out.println("SSE: \t\t\t\t" + sse );
        System.out.println("Var: \t\t\t\t" + var );
        System.out.println("SE(B_1): \t\t\t" + seb1);
        System.out.println("r: \t\t\t\t\t" + r);
        System.out.println("r2: \t\t\t\t" + r2);
        System.out.println("SSR: \t\t\t\t" + ssr);
        System.out.println("SST: \t\t\t\t" + sst);
    }
}
```

Example Code Output : These are the results of the computations using the data and the formulas above. In this case we use the total unemployment rate vs the number of casualties per incident.



```
n:                   228.0
X Bar:               5.8754385964912315
Y Bar:               3.2149122807017543
X Sum:               1339.6000000000008
Y Sum:               733.0
X Sum Squared:       8581.86
Y Sum Squared:       9577.0
XY Sum:              4036.6999999999985
SXY:                 -269.9964912280743
SXX:                 711.1224561403424
B_0:                 5.445678292632564
B_1:                 -0.3796765084504512
SSE:                 7117.957973162267
Var:                 31.49538926177994
SE(B_1):             0.21045114529049785
r:                   -0.11915251423667264
r2:                  0.01419732164892069
SSR:                 102.5113250833493
SST:                 7220.469298245591
```

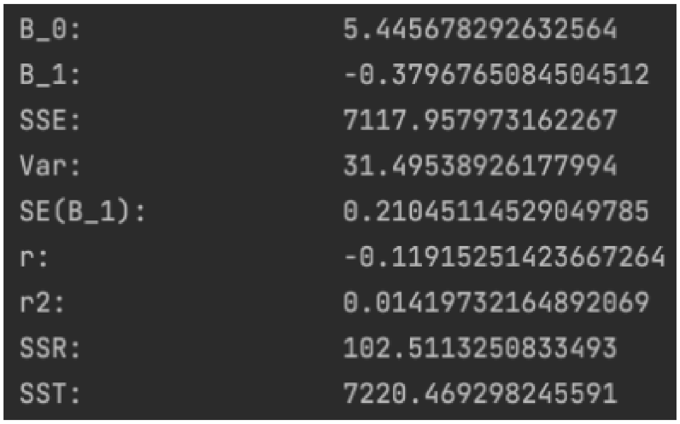
Findings/Analysis: We started the analysis by first analyzing the relationship between the total unemployment rate and the total number of casualties for each school shooting incident between Nov 2001 - Nov 2021. After running our java code on the input csv files with the specified parameters, we get the following graph.



Plot of 228 data points

We also produce a csv file which can be found here: https://docs.google.com/spreadsheets/d/1T6aoHC4Y4lR33VohvhULpqnmXZiSvNfB7tI8bZktBm0/edit?usp=sharing

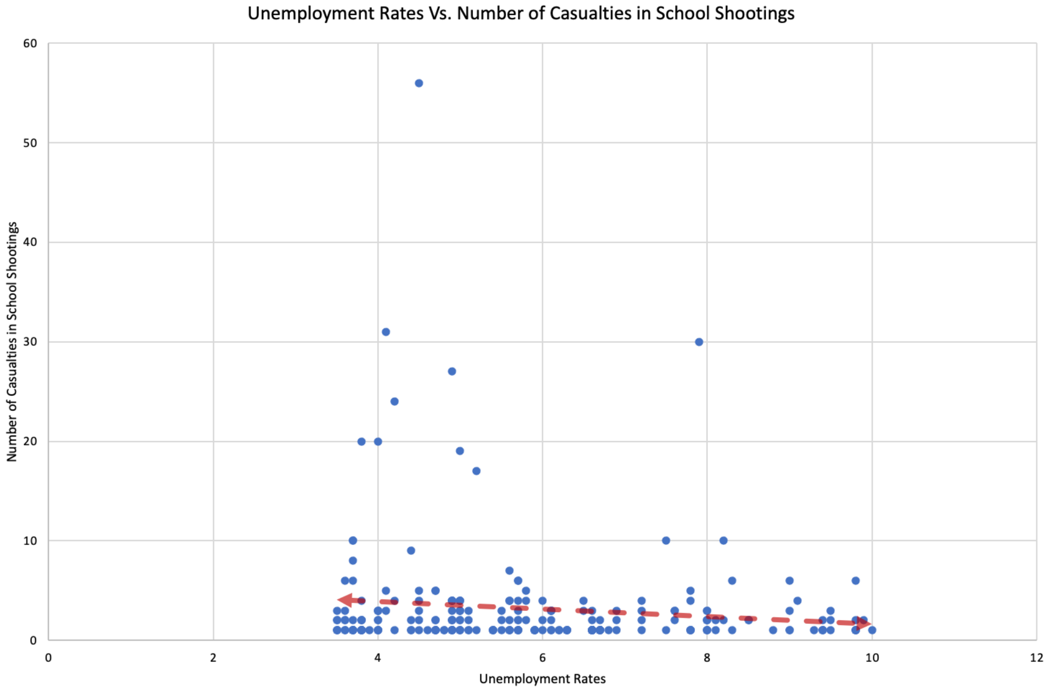
We also computed the equation using the linear regression model which we can see in the console output of our program. We also computed values for the SSE, Var, Standard Error, coefficient of correlation, and coefficient of determination.



```
B_0:                    5.445678292632564
B_1:                    -0.3796765084504512
SSE:                    7117.957973162267
Var:                    31.49538926177994
SE(B_1):                0.21045114529049785
r:                      -0.11915251423667264
r2:                     0.01419732164892069
SSR:                    102.5113250833493
SST:                    7220.469298245591
```

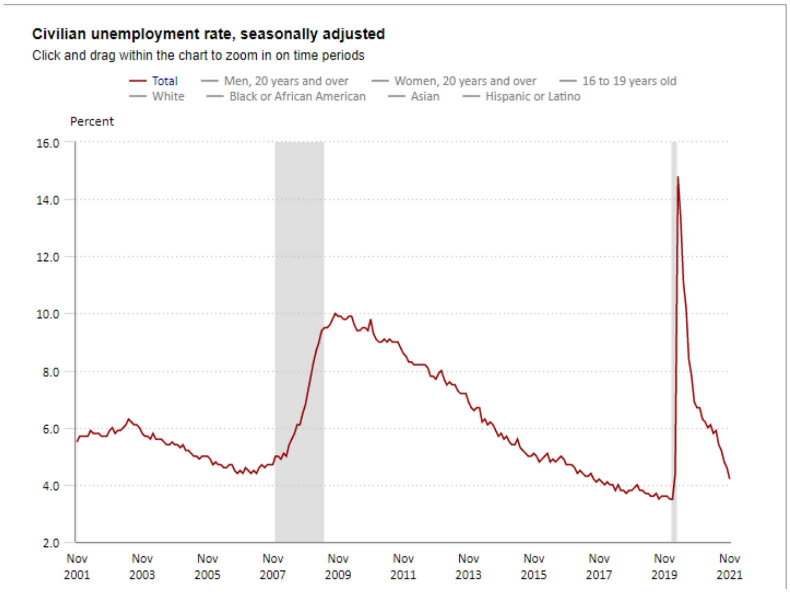
The output from our program. The equation would be y = -0.379 + 5.4456

As we can see, the fitted line using the data is y = -0.379 + 5.4456. This indicates that there is a negative relationship between the unemployment rate for a given month and the number of school shootings. The following graph shows the fitted line among the data points



Graph of results that uses plots and displays a linear trend sloping downward:
Unemployment Rates are independent, Number of Casualties in a School Shooting is dependent
Data spans a 21 year timeline. The total number of points is 228.

So our initial hypothesis that there would be a positive increase in school shootings with an increase in the unemployment rate is not supported by this sample. We can also see that the computed value for the coefficient of correlation is close to 0, indicating there is little relation between our data. One interesting thing to note is that there is a larger number of school shooting incidents between the 2 - 6 percent unemployment rate. However, this does not indicate that having an unemployment rate within that range increases the likelihood of school shootings. We can see from the graph for our unemployment data that over 50% of the points are within that range.

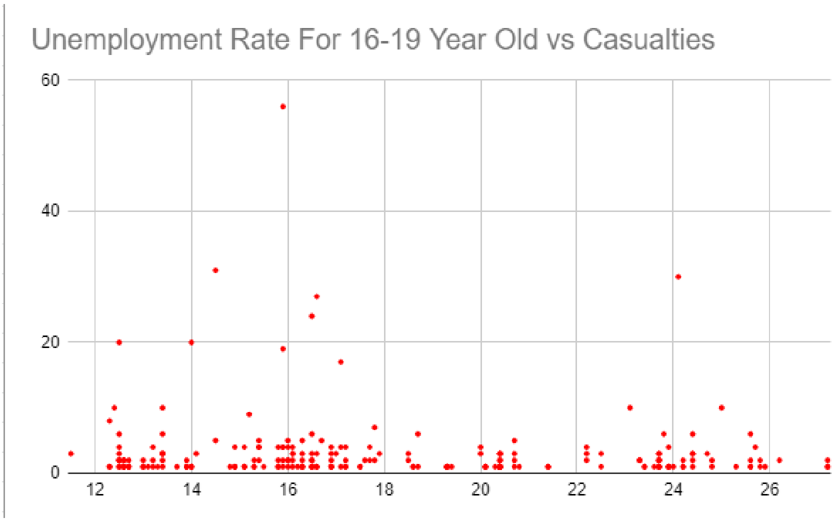


Graph of unemployment rates

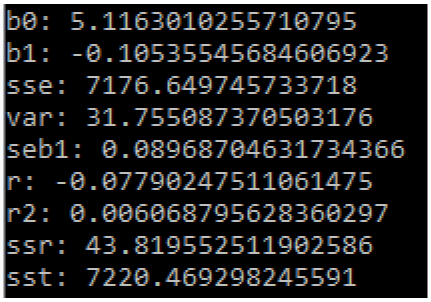
So the reason why there are a larger number of data points within that range is because it is more likely for the unemployment rate to sit within the 2-6% range and it increases the likelihood that a school shooting will occur during that time.

We also attempted to see how the age ranges for unemployment rates may affect the values we compute. We restrict the age range to 16-19 year olds to see if this may affect the number of school shootings.

The follow link is to the table containing the data: https://docs.google.com/spreadsheets/d/1P9AXL3fTTpj7aEcl3gsAej6FZs5ZW7GlmXKw75idgTQ/edit?usp=sharing



Similar graph but with the age range for unemployment restricted to 16-19



```
b0: 5.1163010255710795
b1: -0.10535545684606923
sse: 7176.649745733718
var: 31.755087370503176
seb1: 0.08968704631734366
r: -0.07790247511061475
r2: 0.006068795628360297
ssr: 43.819552511902586
sst: 7220.469298245591
```

This is the resulting values that we could when performing the regression

Very similar to the analysis done without the age restrictions on unemployment rate, the slope of the regression line is close to 0. The coefficient of correlation is also near 0 indicating that there is little correlation between our data.

Conclusion:

Looking at our covariance, we have a large negative number, which tells us that there is little similar behavior between socioeconomic status and the rate of school shootings. This is backed up by the Sum of Squares and it's error, residual and maximum, which all have large outputs, implying there's a large amount of variation between our data. We can see a trend with the coefficient of correlation showing our two variables having a negative correlation, meaning that when we have more school shootings, we can see an ever so slight increase in unemployment. The problem with this trend, and our initial hypothesis, is that the coefficient of determination has a low value, implying we have a low prediction rate for this trend, as well as a low level of correlation with variance. In conclusion, if we look at our output we can clearly see that we have a low prediction rate and little similar behavior between socioeconomic effect and school shootings. Using the standard error of beta, we can see that we have a low error rate based on our observed values, so we are sure that we have a good regression model.