Group 49

Kevin Cleary
James O'Neill

Predicting a given country's suicide rates for a specified year

Our Project aims to predict a country's suicide rates for a given year given a number of socio-economic factors:

- Country's profitability
- Country's Healthcare stats
- Employment Rates

We hope to use our findings to determine what factors can impact the rise/fall of suicide rates from year to year and identify positive change that can be taken to combat a rise in these rates.

We plan to use the following datasets to train and test our data/algorithm:

- Research on Suicide rates in each country spanning from years 1985-2016 (World Health Organisation)
- Dataset including stats on various socioeconomic fields (WorldBank.org World Development Indicators)

Group 49

Kevin Cleary
James O'Neill

Understanding Our Dataset

Understanding our Dataset

From the World Health Organisation.

- Important Columns
  - Country
  - Year
  - Age Range
  - Suicides per 100k Population
  - GDP for Year

Note..

- Human Development Index not present for a good few rows.

- Generation column not necessary

- Use Suicides/100K population to get the rate of suicide as opposed to the number of suicides.

country	year	sex	age	suicides_n	population	suicides/100k pop	country-year	HDI for year	gdp_for_year ($)	gdp_per_capita ($)	generation
Albania	1987	male	15-24 years	21	312900	6.71	Albania1987		2,156,624,900	796	Generation X
Albania	1987	male	35-54 years	16	308000	5.19	Albania1987		2,156,624,900	796	Silent
Albania	1987	female	15-24 years	14	289700	4.83	Albania1987		2,156,624,900	796	Generation X
Albania	1987	male	75+ years	1	21800	4.59	Albania1987		2,156,624,900	796	G.I. Generation
Albania	1987	male	25-34 years	9	274300	3.28	Albania1987		2,156,624,900	796	Boomers
Albania	1987	female	75+ years	1	35600	2.81	Albania1987		2,156,624,900	796	G.I. Generation
Albania	1987	female	35-54 years	6	278800	2.15	Albania1987		2,156,624,900	796	Silent
Albania	1987	female	25-34 years	4	257200	1.56	Albania1987		2,156,624,900	796	Boomers
Albania	1987	male	55-74 years	1	137500	0.73	Albania1987		2,156,624,900	796	G.I. Generation
Albania	1987	female	5-14 years	0	311000	0	Albania1987		2,156,624,900	796	Generation X
Albania	1987	female	55-74 years	0	144600	0	Albania1987		2,156,624,900	796	G.I. Generation
Albania	1987	male	5-14 years	0	338200	0	Albania1987		2,156,624,900	796	Generation X
Albania	1988	female	75+ years	2	36400	5.49	Albania1988		2,126,000,000	769	G.I. Generation
Albania	1988	male	15-24 years	17	319200	5.33	Albania1988		2,126,000,000	769	Generation X
Albania	1988	male	75+ years	1	22300	4.48	Albania1988		2,126,000,000	769	G.I. Generation
Albania	1988	male	35-54 years	14	314100	4.46	Albania1988		2,126,000,000	769	Silent
Albania	1988	male	55-74 years	4	140200	2.85	Albania1988		2,126,000,000	769	G.I. Generation
Albania	1988	female	15-24 years	8	295600	2.71	Albania1988		2,126,000,000	769	Generation X
Albania	1988	female	55-74 years	3	147500	2.03	Albania1988		2,126,000,000	769	G.I. Generation
Albania	1988	female	25-34 years	5	262400	1.91	Albania1988		2,126,000,000	769	Boomers
Albania	1988	male	25-34 years	5	279900	1.79	Albania1988		2,126,000,000	769	Boomers
Albania	1988	female	35-54 years	4	284500	1.41	Albania1988		2,126,000,000	769	Silent
Albania	1988	female	5-14 years	0	317200	0	Albania1988		2,126,000,000	769	Generation X
Albania	1988	male	5-14 years	0	345000	0	Albania1988		2,126,000,000	769	Generation X
Albania	1989	male	75+ years	2	22500	8.89	Albania1989		2,335,124,988	833	G.I. Generation
Albania	1989	male	25-34 years	18	283600	6.35	Albania1989		2,335,124,988	833	Boomers

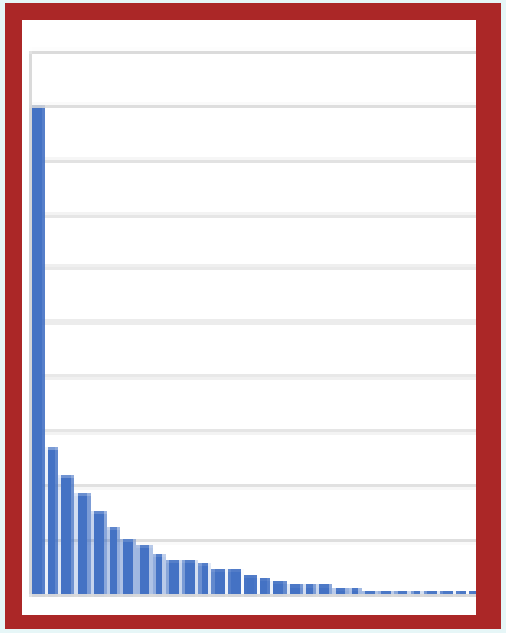
Group 49

Kevin Cleary
James O'Neill

Finding the Central Tendency

For the 'Suicides/100K pop' field

- Mean: 12.8161
- Trimmed Mean (2%): 11.29141
- Median: 5.99
- Mode: 0
- Midrange: 113.94
- (Max = 224.97, Min = 0)



This basic histogram of our training dataset shows a violent negative skew.

- this is largely due to the abundance of 0 values in the data.
- Further study of the dataset showed us that round 15% of the values for this attribute were 0.
- Accountable for the fall in the trimmed mean, the much lower median and a useless Mode.
- High midrange value due to extreme outliers in the upper-end of the scale.

Group 49

Kevin Cleary
James O'Neill

Data Preparation

Rise/Fall in Rates

- The Initial Data is not very telling of what we want; to see the rise/fall of these rates from year to year
- Need to compare the rates from year to year
- Also decided to group together all age/gender groups for each country year.
- This eliminated abundance of 0s

  (15% of population = 0 originally, now down to roughly 1.6%)

Group 49

Kevin Cleary
James O'Neill

Data Preparation

```
while ((line = br.readLine()) != null) {

    // use comma as separator
    String[] row = line.split(cvsSplitBy);

    groupedSNo += Float.valueOf(row[4]);
    groupedPop += Float.valueOf(row[5]);
    groupedSRate += Float.valueOf(row[6]);

    if ((lineCount % 12) == 0) {
        System.out.println(currentCountry);
        changeInSRate = groupedSRate - lastYearSRate;
        if (currentCountry.equals(row[0])) {
            if (changeInSRate == 0) {
                trend = "Remain";
            } else if (changeInSRate > 0) {
                trend = "Increase";
            } else {
                trend = "Decrease";
            }
            changeInSRateStr = df.format(changeInSRate);
        } else { currentCountry = row[0]; trend = "N/A"; changeInSRateStr = "N/A"; }
        rowString = String.join(", ", row[0], row[1], String.valueOf(groupedSNo), String.valueOf(groupedPop), String.valueOf(df.format(groupedSRate)), trend, changeInSRateStr);
        System.out.println("Row " + lineCount + " values are: {" + rowString + "}");
        writer.append(rowString);
        writer.append("\n");

        lastYearSRate = groupedSRate;
        groupedSNo = 0;
        groupedPop = 0;
        groupedSRate = 0;
        changeInSRate = 0;
    }
```

Calculating the change between this year and the last for each country

Grouped into Categories: The trend will either increase, decrease or remain the same

Group 49

Kevin Cleary
James O'Neill

Data Preparation

Rise/Fall in Rates

- The Initial Data is not very telling of what we want; to see the rise/fall of these rates from year to year
- Need to compare the rates from year to year
- Also decided to group together all age/gender groups for each country year.
- This eliminated abundance of 0s

  (15% of population = 0 originally, now down to roughly 1.6%)

country	year	suicide_	population	suicide/100K pop	suicide_rate_trend	increase/decrease
Albania	1987	73	2709600	31.85	N/A	N/A
Albania	1988	63	2764300	32.46	Increase	0.61
Albania	1989	68	2803100	33.4	Increase	0.94
Albania	1992	47	2822500	18	Decrease	-15.4
Albania	1993	73	2807300	32.56	Increase	14.56
Albania	1994	50	2849300	32.18	Decrease	-0.38
Albania	1995	88	2903400	40.55	Increase	8.37
Albania	1996	89	2940200	43.62	Increase	3.07
Albania	1997	170	2977300	77.43	Increase	33.81
Albania	1998	154	3012700	66.52	Decrease	-10.91
Albania	1999	139	3029700	69.81	Increase	3.29
Albania	2000	54	2796300	30.7	Decrease	-39.11
Albania	2001	119	2799349	50.62	Increase	19.92
Albania	2002	133	2818839	62.51	Increase	11.89
Albania	2003	124	2843929	58.6	Decrease	-3.91
Albania	2004	146	2874991	65.39	Increase	6.79
Albania	2005	0	2783320	0	Decrease	-65.39
Albania	2006	0	2780176	0	Remain	0
Albania	2007	124	2770344	65.85	Increase	65.85
Albania	2008	160	2757059	71.05	Increase	5.2
Albania	2009	0	2745735	0	Decrease	-71.05
Albania	2010	96	2736025	41.66	Increase	41.66
Antigua and Barbuda	1985	0	62574	0	N/A	N/A
Antigua and Barbuda	1986	0	61270	0	Remain	0
Antigua and Barbuda	1987	0	60261	0	Remain	0
Antigua and Barbuda	1988	0	59564	0	Remain	0
Antigua and Barbuda	1989	0	59238	0	Remain	0
Antigua and Barbuda	1990	1	59334	17.24	Increase	17.24

Group 49

Kevin Cleary
James O'Neill

Finding the Central Tendency: Revised

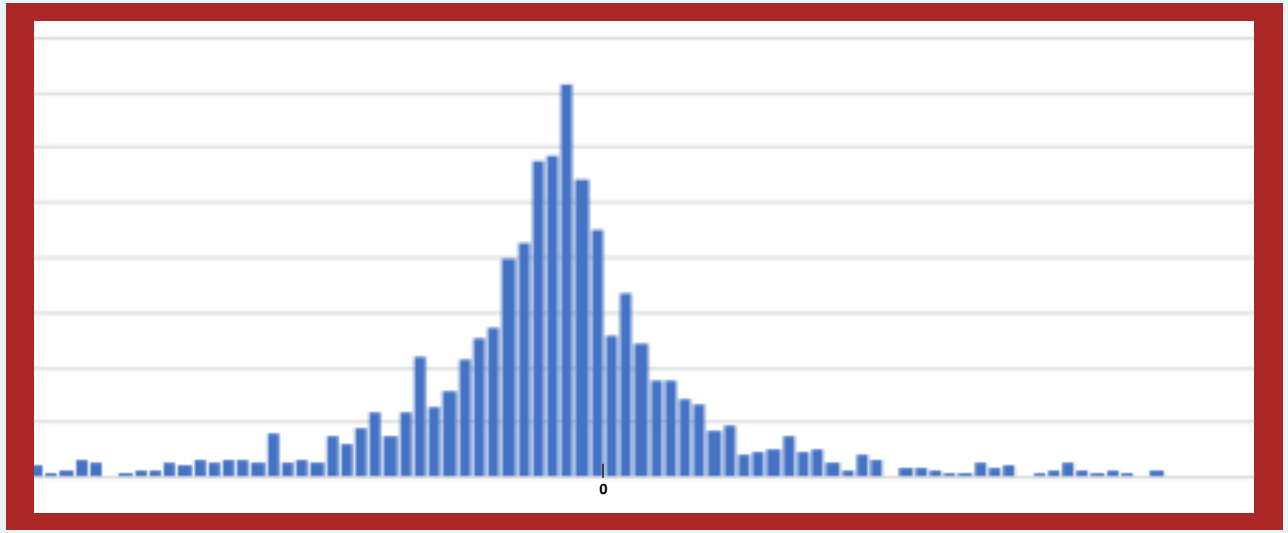
Prepared Data: rise/fall

- Mean: -0.275
- Trimmed Mean (2%): -1.595
- Median: -1.37
- Mode: 0
- Midrange: 72.34
- (Max = +422.73, Min = -278.05)

- Mode (formulaic): +0.92

(based on following formula):

$$mean - mode = 3 \times (mean - median).$$



- More of a normal sized distribution for rise/fall, slightly negatively skewed.
- Trend is most likely to be slightly decreasing.
- Extreme values still do exist, but this is not affected as much by an abundance of 0 values.