**Deaths of Despair: A comparison of suicides per 100,000 amongst age groups based on Kaggle suicide data (1985 – 2016)**

|  |  |
| --- | --- |
| Assignment submission for 7COM1079 - [Team Research and Development Project](https://herts.instructure.com/courses/76820) by GROUP 49 | |
| Purvish Himanshukumar Jariwala |  |
| Z Mohammed Ayyaz Khan |  |
| Diwakar Ranjan | M.Sc. Advanced Computer Science (dr18abc@herts.ac.uk) |
| Vikrant Sharma |  |
| Rahul Soni | M.Sc. Artificial Intelligence & Robotics (rs20acs@herts.ac.uk) |

**Abstract**

Stories about suicide that appear in the news tend to focus on celebrities who have taken their own lives and on clusters of deaths among students. They miss the bigger picture: that, at a global level, suicide has declined by just 8% since its peak in 2009 (139211 suicides in absolute terms) almost returning to its previous low 127156 suicides of in 2006, for our dataset.

**Objective**

This study aims to assess if there is a difference in the proportion of suicides per 100,000 amongst age groups.

**Method**

Data from Kaggle [1] was examined using R programming language to create visualisations and understand the suicides per 100,000 amongst age groups and genders. Plots were created to further aid in examining the suicides per 100,000 trends over demographics.

We further used chi-square test to further answer our hypothesis.

**Results**

Nearly 129776 people took their own lives in 2015 up 27 percentage points from 1985, and suicide is an important part of the rising number of “deaths of despair” described by Anne Case and Angus Deaton, economists [2]. This cause of death has increased the overall mortality rate amongst male over female. Both absolute levels of suicide per 100,000 people and recent increases are particularly high amongst those aged 75 years and over. The male suicide rate remains almost four times that for women.

**Implications**

On average suicide rate per 100k increases with age (section 4.1.3) and suicide rate per 100k in men is 3 times higher to those in women (section 4.1.4). We further examined suicide rate per 100k across both genders and age groups and notice that there is evidence to reject the null hypothesis.

**Research Question**

Is there a difference in proportion of suicide rate per 100k amongst the age groups?

# **PROJECT OVERVIEW**

# Key Insights:

* Suicide rates per 100k are decreasing globally to an average of 11.23 (section 2.1)
* Suicide rate per 100k in men is 3 times higher than women
* Globally suicide rate per 100k increases with age
* Children in 5 -14 age group contributes to about 1% of all suicides across all age groups
* Suicide rate per 100k remains high in 75 and over age group for every year in the record set

# Introduction:

Suicide regularly gets from profound sensations of sadness. Casualties ordinarily have the powerlessness to see answers for issues or to adapt to testing life conditions which at that point drives them to consider suicide to be the lone arrangement. As indicated by World Health Organization Suicide is a significant medical condition worldwide and is a main source of death. More than 800,000 individuals die consistently from suicides which is assessed to be at a pace of one individual like clockwork. However, suicide is preventable when ideal, powerful intercessions are actualized at public, civil and individual levels. With regards to a nation's pay level, suicide does not simply happen in major league salary nations, however it is a worldwide marvel in all locales of the world.

This research will be focused on Is there a difference in a proportion of suicide rate/100k amongst the age groups and supporting hypothesis as:

H0: There is no difference in the proportion of suicide rate/100k amongst age groups.

H1: There is a difference in the proportion of suicide rate/100k amongst age groups.

# Data:

The information to be utilized in this task was gathered by extracting data from an online information base (Kaggle.com).

There are sum of 27,820 cases where each case speaks to a nation and the suicide rate inside an age-group gathering of males or females for the year somewhere in the range of 1987 and 2016 alongside the nation's GDP at that point. As referenced before, the research will be focused about finding any connection between Suicide rate/100k and age-group, in this way these are the two factors which will be the fundamental concentration in which suicide rate per 100k is a dependent variable and age-group is an independent variable

This research is an observational investigation since the members are seen with no sort of interference. Therefore, the objective is to check whether is there any difference in the proportion of suicide rate per 100k amongst age-groups. The number of inhabitants in interest is all people ages 5 and up who ended it all. The socioeconomics from this data comes from nations around the globe so we can sum up our decisions to the population internationally. In any case, since the Research is observational the discoveries cannot be utilized to set up causal connections, just **Comparison of proportions**. For likely inclination, we need to accept that each nation similarly revealed all suicide occasions, in any case our decisions might be inaccurate. Suicide is viewed as a negative fascination for a nation's economy so on the off chance that numbers go unreported, at that point the outcomes will reflect error.

# Restrictions or Limitations and Assumptions

1. The data in this Research depends on reports up to 2015 excluding 2016
2. The significance level (α) is set to 0.05 and is expected that there is a homogeneity of fluctuation in the data

# **PROJECT MANAGEMENT AND ORGANIZATION PLAN**

# Task Management & Team Meetings:

The team met every alternative day at 20:00 (GMT) with their task status, and which Trello card will they pick next. Any impediments were also brought in the team meetings so that risks of delivering the research can identified and mitigated. Microsoft Teams was used as a medium of communication due to ongoing Covid-19 restrictions [3].

Trello Board [4] was used to manage the team tasks and to create visibility of the work involved. As the team is comprised on 5 members we agreed on the Work in Progress (WIP) limit of 5. WIP limit of 4 for all cards being actively worked on – marked as “Doing” on Trello Board and WIP limit of 1 that required peer review – marked as “Review” on Trello Board.

Further labels were added to every card to visualise which stage of the research does it corresponds to, so that goal of the research is kept in mind at all times during the development. Due date was added in every card as outlined in the coursework specification guide [5]

‘R’ programming language used from sourcing and cleaning data, creating visualisation and perform further analysis to test hypothesis.

# Storage:

We used bitbucket [6] as a version control repository to keep a record of up to date work shared between all resources in the team.

Every member has their own branch where they pushed their code to be later reviewed in the team meetings. Peer review was conducted during the team meetings and necessary amends were suggested (if any). Following successful peer review, the code was *merged* into the master branch which was later *pulled* by the team members so that they can work on the latest copy of the code.

# **DATA COLLECTION AND PROCESSING DOCUMENTATION**

# Data Collection Methods:

Every team member brought their own interests and their chosen dataset for team discussions. After much deliberation suicide\_dataset was finalised by the team members downloaded from Kaggle as it fulfils all the criteria [7] we set out in our almost daily meeting using *Microsoft Teams*. We explain these criteria as

1. Dataset should be easy to *understand* i.e. individual/team should not spend a long time (more than 2 days) to investigate terminology/column name/data in a column
2. It should to *feasible* for the team to implement given limited understanding of the course/R and help in framing a research question and hypothesis
3. Dataset should help in understanding the *concept* of Correlation, Comparison of means, or comparison of proportion and draw visualisation
4. The answer to the RQ can be easily *summarised* in a report

# Data Processing & Cleaning Methods:

Following steps were taken to ensure better data quality.



The above steps are explained in detail below

1. The suicide dataset is in CSV format and we imported the dataset in R. Total of 27820 observations were recorded across 2 genders (male, female), 6 age groups (5-14 years, 15-24 years, 25-34 years, 35-54 years, 55-74 years, 75+ years) and 101 countries. The data can be understood as

<create table here>

1. Next, we visualised the number of records for every year for the period 1985 – 2016 and removed all records for 2016 as suicides were not reported by all countries in the dataset.
2. We expect there are at least 12 records (6 age groups and 2 genders) for every year reported by each country in the dataset.
   1. we noted that there are only 30 countries that have reported suicides for all the years over the 1985 – 2015 period.
   2. we further examined that these countries account ~40% of the total observations reported in the data set (11076/27820).
   3. we, therefore, considered this data as viable enough to further carry our assessment and answer the research question.
3. Finally, we removed any columns that we do not need in our study. This includes

* HDI for year
* gdp\_for\_year ($)
* gdp\_per\_capita ($)
* generation

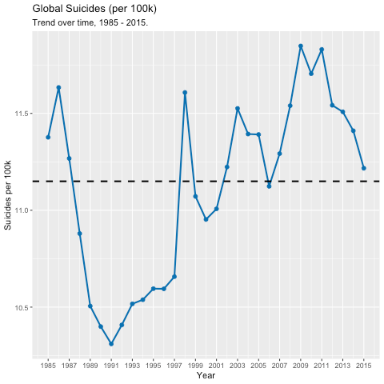
No further data enrichment was made to the original data set.

1. **VISUALISATION**

# Global Analysis

* + 1. Global Trend

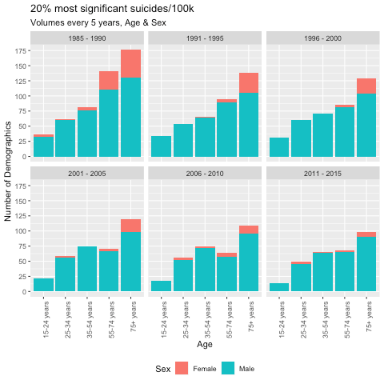
The dashed line is the **global average suicide rate** from 1985 - 2015: **11.5 suicides** (per 100k, per year). There was a spike in suicide rate per 100k in 2009 that is the largest in the dataset. This could be related to the global slowdown of 2009, but we cannot be certain as the dataset does not list the cause of suicides.



* + 1. Top 20% highest risk between 1985 – 2016

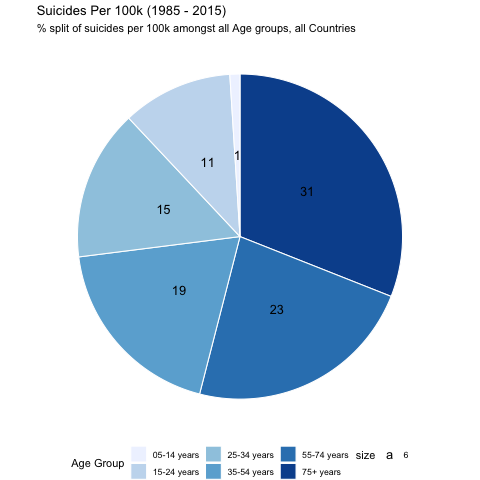
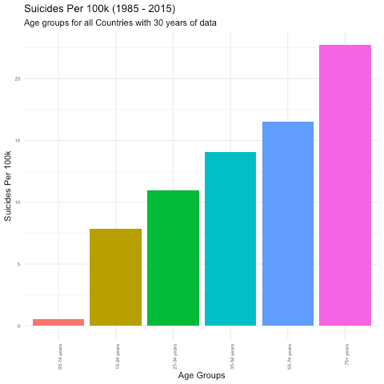
We were interested to understand how suicide rate per 100k has changed every 5 years. There are 30 country data in our filtered data set (see section 3.2) and we were interested in the 20% highest risk suicides per 100k and compare the demographics between 1985 and 2015.

We define a demographic as a year in a particular country for some combination of sex and age.



**We observe suicides per 100k for the 35 – 54 years age group has largely remained flat since 1991 at about 70 suicides per year. Although the suicides per 100k has decreased for 75+ years age group we remain concerned it being high throughout 30-year period.**

* + 1. By Age Group

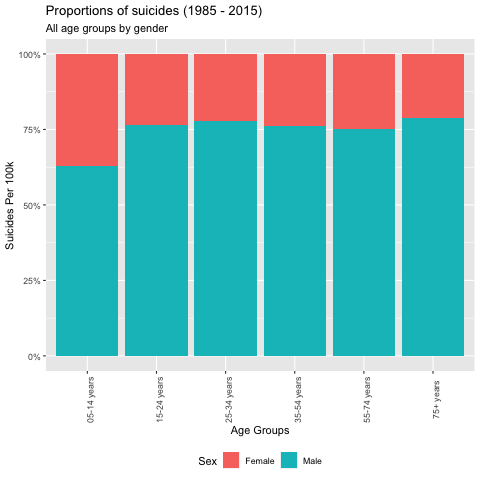
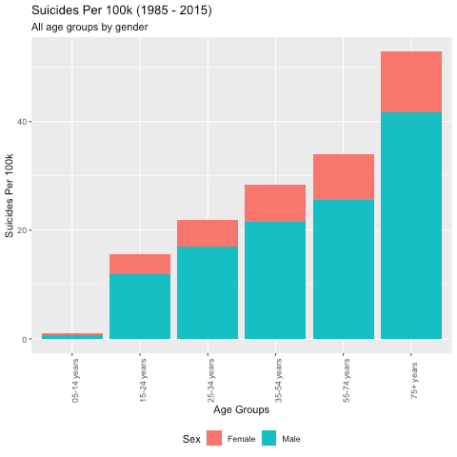


**The above bar plot shows the suicide rate per 100k for all age groups with 30 years data.**

**We notice that suicide rate per 100k increases with rise in age. We also detect and remain concerned about the high suicide rate per 100k for 75+ age, which further needs to be examined. Therefore, the ratio indicates suicide rate per 100k is highest in the 75+ age group and minimum in the 5-14 age group.**

**The pie chart, also confirms that suicides rate per 100k is about 31% for those in 75 years and over age group. We are anxious that suicide rate per 100k in children (5 – 14 years) is 1% and recommend further studies on what causes children to take of such extreme measures. A study [8] suggests factors there are broad risks factors that makes it difficult for parents and health-workers to spot children at risk of suicides.**

* + 1. By Age Group and Gender



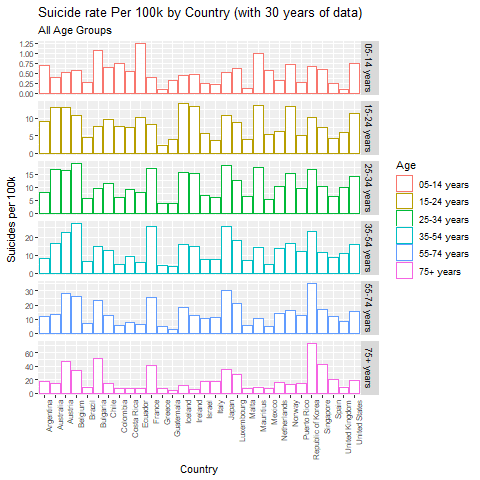
**The bar plot and stacked bar plot above shows the suicide rate per 100k for all age groups and all genders (male and female).**

**We observe that suicide rate per 100k in males is about 3 times that of females between 1985 and 2015.**

**While suicide rate per 100k amongst females for all age categories remains the same, we inspect that females in the age group 5 – 14 years, have higher suicide rate per 100k over males compared to all other age groups at about 37%.**

# By Country

* + 1. Per Age Group



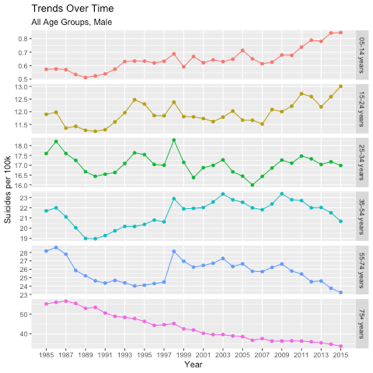
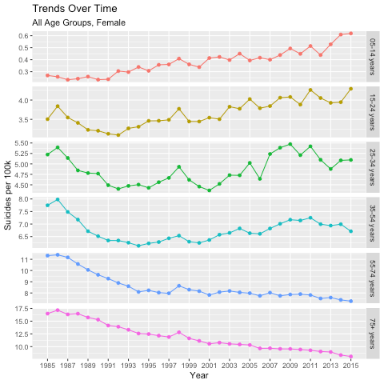
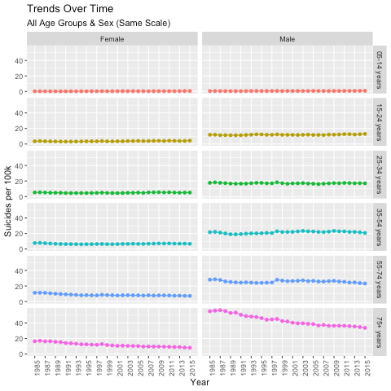
**The bar plot shows the suicide rate per 100k by the countries with 30 years of data for all age groups.**

**We observe Ecuador, Iceland, Japan, Belgium has the highest suicide rate per 100k and UK, Greece and Guatemala have the lowest suicide rate per 100k in the age groups 05-14 years, 15-24 years, 25-34, years and 35-54 years respectively. Similarly, the Republic of Korea has the highest suicide rate per 100k and Guatemala have the lowest suicide rate per 100k in the age groups 55-74 years and 75+ years.**

**We also notice in the country Republic of Korea, the suicide rate per 100k increases with rising in age as in the country United States, Argentina, and Guatemala, the suicide rate per 100k decreases with rising in age.**

# Trend Plots

* + 1. By Age Group & Gender



We share the following insights:

* + - * + Globally the likelihood of suicide increases with age
        + Although the suicide rate in 5 – 14 years remains small and it has doubled in the past 30 years for our filtered dataset for both males and females
        + Suicide rate per 100k for 75+ year age group for both genders is linearly decreasing and has dropped by about 83% since 1988
        + Suicide rate per 100k for 15 – 24 years age group for both genders is highest in 2015 than any time in the 30 years of history

1. **ANALYSIS**

Data Analysis is the cycle of methodically applying statistical and additionally intelligent strategies to portray and outline, consolidate and assess data which can also help in testing our hypothesis.

This statistical analysis will especially focus on the suicide and age-group which is done with the help of proportion test in R.

Proportion test is used to check our hypothesis and based on that findings are as follow

**Table 1.** Total population (in 100k) and suicide rate per 100k (for all age groups)

|  |  |  |
| --- | --- | --- |
| **Age Group** | **Suicides per 100k** | **Total Population (in ‘000,000)** |
| 05-14 years | 0.515137072 | 55148.42857 |
| 15-24 years | 7.869480444 | 55974.8516 |
| 25-34 years | 10.97714016 | 54335.09925 |
| 35-54 years | 14.08981603 | 91890.12811 |
| 55-74 years | 16.51480639 | 56337.14245 |
| 75+ years | 22.70232293 | 17917.72592 |

In Table 1, the column “suicideper100k” is suicide rate per 100k population of the total population for a given age group. In addition, the column “total population (in 100k)” denotes the total population (male and female) for a given age group.

By looking at the Table 1 and plots in visualization, we can say that there is a difference in the proportion of the suicide rate amongst age groups. The table suggests suicide rate is reasonably increasing with the age.

We need to find whether the difference is significant or not? For that, we are using Chi-squared test. To demonstrate the test in R, we are using a prop.test function.

After applying prop.test, the results are:

**Table 2.** Results of proportions test on sample data (all countries with 30 years of data, 1985 - 2015)

|  |  |
| --- | --- |
| X-squared = 105.75 | df = 5 |
| p-value = 0.0002286 | alternative hypothesis: two.sided |
| prop 1 | 0.000009340920 (9.340920e-06) |
| prop 2 | 0.0001405896 (1.405896e-04) |
| prop 3 | 0.0002020267 (2.020267e-04) |
| prop 4 | 0.0001533333 (1.533333e-04) |
| prop 5 | 0.0002931424 (2.931424e-04) |
| prop 6 | 0.001267031 (1.267031e-03) |

The p-value of the test is 2.2e-16, which is less than the significance level alpha 0.05 provides strong evidence against the null hypothesis, accepting that it is giving reasonable evidence to continue our focus to alternative hypothesis[9].

To verify above argument from a different perspective, we decided to run the same test (Chi-squared test) with our raw dataset.

**Table 3.** Total population and total number of suicides for the age group(for raw data)

|  |  |  |
| --- | --- | --- |
| **age\_group** | **suicidesper100k** | **total\_population (in 100k)** |
| 5-14 years | 0.622287284 | 83986.93237 |
| 15-24 years | 9.354934257 | 86429.46896 |
| 25-34 years | 13.31948569 | 84381.03587 |
| 35-54 years | 17.05731833 | 143758.8812 |
| 55-74 years | 18.83899557 | 88032.4534 |
| 75+ years | 24.52305776 | 26632.81253 |

**Table 4.** Results of proportions test on table 3.

|  |  |
| --- | --- |
| X-squared = 110.98 | df = 5 |
| p-value <  0.00000000000000022  (2.2e-16) | alternative hypothesis: two.sided |
| prop 1 | 0.000007409335 (7.409335e-06) |
| prop 2 | 0.0001082378 (1.082378e-04) |
| prop 3 | 0.0001578493 (1.578493e-04) |
| prop 4 | 0.0001186523 (1.186523e-04) |
| prop 5 | 0.0002140006 (2.140006e-04) |
| prop 6 | 0.0009207836 (9.207836e-04) |

After comparing both the tables **Table 2** and **Table 4** we observe similar results. P-value is less than the significance level (0.05) and hence we reject the null hypothesis. There is significant difference in the proportion of suicide rate per 100k amongst the age groups.

1. **CONCLUSION**
2. **REFERENCES**

[1] <https://www.kaggle.com/russellyates88/suicide-rates-overview-1985-to-2016>

[2] Anne Case and Angus Deaton, ‘Deaths of Despair and the Future of Capitalism’ [Online] <https://press.princeton.edu/books/hardcover/9780691190785/deaths-of-despair-and-the-future-of-capitalism> [Accessed: 03 Jan 2021]

[3] Covid-19 restrictions, [Online] <https://www.gov.uk/coronavirus#:~:text=(COVID%E2%80%9119)-,National%20lockdown%3A%20stay%20at%20home,spreading%20it%20without%20knowing%20it>.

[4] Team 49 Trello Board, [Online] <https://trello.com/b/lr2RoToP> [Accessed: 8 Jan 2021]

[5] Coursework specification [11 Nov 2020], [Online] <http://jnoll.nfshost.com/7COM1079-fall-20/coursework/7COM1079-fall-20-coursework.pdf> [Accessed: 8 Jan 2021]

[6] Team 49 Bitbucket Repository, [Online] <https://imrahulsoni@bitbucket.org/imrahulsoni/7com1079-group49.git> [Accessed: 8 Jan 2021]

[7] Data set selection criteria, [Online] <https://trello.com/c/0uYp2wZC> [Accessed: 8 Jan 2021]

[8] Kirsten Weir (December 2016), ‘Research on suicide overlooks young children; psychologists are working to change that’ [Online] <https://www.apa.org/monitor/2016/12/ce-corner> [Accessed: 8 Jan 2021]

[9] Statsdirect.co.uk. 2021. *P Values (Calculated Probability) And Hypothesis Testing - Statsdirect*. [online] Available at: <https://www.statsdirect.co.uk/help/basics/p_values.htm> [Accessed 8 January 2021].