CA682 Data management and visualisation

Name	Aruna Bellgutte Ramesh
Programme	MCM (Cloud Computing)
Module Code	CA682
Assignment Title	Data Visualisation
Submission date	16th December 2018
Module coordinator	Suzanne Little

I declare that this material, which I now submit for assessment, is entirely my own work and has not been taken from the work of others, save and to the extent that such work has been cited and acknowledged within the text of my work. I understand that plagiarism, collusion, and copying are grave and serious offences in the university and accept the penalties that would be imposed should I engage in plagiarism, collusion or copying. I have read and understood the Assignment Regulations set out in the module documentation. I have identified and included the source of all facts, ideas, opinions, and viewpoints of others in the assignment references. Direct quotations from books, journal articles, internet sources, module text, or any other source whatsoever are acknowledged and the source cited are identified in the assignment references. This assignment, or any part of it, has not been previously submitted by me or any other person for assessment on this or any other course of study.

I have read and understood the referencing guidelines found recommended in the assignment guidelines.

Name: **Aruna Bellgutte Ramesh** Date: **16th December 2018**

Visualization of Suicide Causes in India (for the period from 2003 to 2012)

Introduction

It is sad to note that despite the advent of technology, increasing mental health care centres, improved medicines and health care institutes, the number of deaths due to suicide is only increasing. It seems important to understand why people commit suicide in order to curb the increasing deaths.

India is the second most populated country and seventh largest country by area, analysing their suicide statistics can bring more insight to the causes of suicides.

Visualization of the data was done with **D3 JS (v5)**. For front end architecture, AngularJS was used. Since the data was pulled from database rather than from CSV, MySQL was used for DB, and Node JS with Express was used for middleware config.

Questions

The following questions were answered through visualization of the data:

1. What are the major reasons for suicide?
2. Do the reasons for suicide differ for men and women? If so, can the dataset available answer how different reasons affect / influence men and women?
3. Do the reasons for suicide change with age? If so, what are the main reasons for suicide for different age groups?
4. Which parts / states of India have high number of suicides?

Dataset

The suicide dataset of India was taken from Indian government website at https://data.gov.in/catalog/stateut-wise-distribution-suicides-causes. Here, only the dataset for the years from 2001-2012 available at the link https://data.gov.in/resources/stateut-wise-distribution-suicides-causes-during-2001-2012 was downloaded. The csv file was downloaded from the given location and was imported into MySQL Database after cleaning and processing the data.

Dataset Summary

India consists of 29 States and 7 Union Territories (UTs). Data for all these states and UTs for the years from 2001 to 2012 are provided in the CSV file. It contains 12769 rows of data with 16 columns. Major columns include the "Cause", "Total Male", "Total Female" and "Grand Total". However, there are also columns for different age groups of male and female too which sum up to give "Total Male" and "Total Female" and finally the sum of these two as "Grand Total". There are about **26 reasons/causes for suicide** as given by dataset.

It was important to note that, under the STATE/UT column, there were non-state and non-UT values like TOTAL (STATES), TOTAL (UTs) and TOTAL (ALL INDIA). These rows give summation of suicide deaths totalled for all states alone, all UTs alone and total of states and UTs, respectively. In the “Cause” column, there were non-cause values like “Total” and “Total Illness”. “Total” row gives the total suicides and “Total Illness” gives the number of suicides whose reason / cause was due to illness. Another point noted was that “Total Illness” was the summation of deaths whose causes / reasons were Illness (Aids/STD), Cancer, Paralysis, Insanity/Mental Illness, Other Prolonged Illness. These rows helped me greatly in visualising the data. **Note: For ease, I’ve considered the data only from 2003 to 2012.**

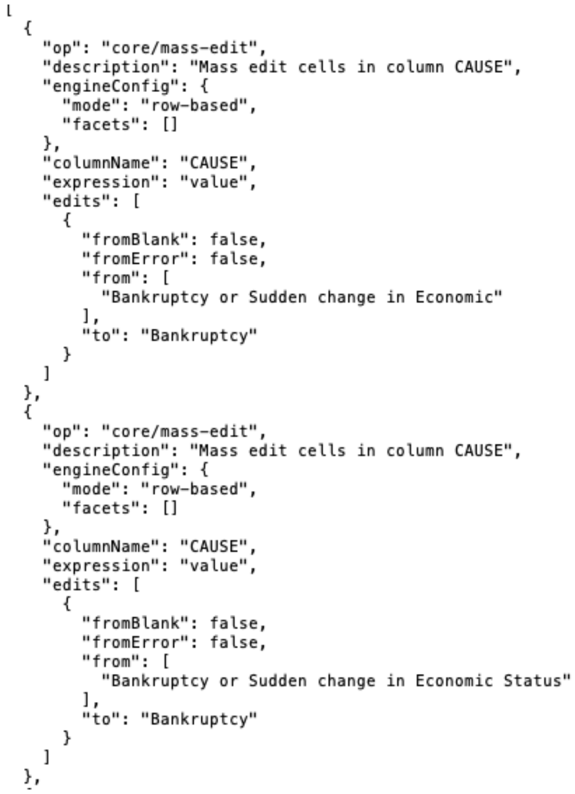
Process

Steps followed for data visualization are as follows:

1. Data was downloaded and analysed. It is very important to **analyse** the data and its domain. This helps in understanding the visualizations better.

2. Data was **cleaned** with the help of :

Fig(1) Sample of Open Refine changes



```
[
  {
    "op": "core/mass-edit",
    "description": "Mass edit cells in column CAUSE",
    "engineConfig": {
      "mode": "row-based",
      "facets": []
    },
    "columnName": "CAUSE",
    "expression": "value",
    "edits": [
      {
        "fromBlank": false,
        "fromError": false,
        "from": [
          "Bankruptcy or Sudden change in Economic"
        ],
        "to": "Bankruptcy"
      }
    ]
  },
  {
    "op": "core/mass-edit",
    "description": "Mass edit cells in column CAUSE",
    "engineConfig": {
      "mode": "row-based",
      "facets": []
    },
    "columnName": "CAUSE",
    "expression": "value",
    "edits": [
      {
        "fromBlank": false,
        "fromError": false,
        "from": [
          "Bankruptcy or Sudden change in Economic Status"
        ],
        "to": "Bankruptcy"
      }
    ]
  },
```

a) **Open Refine:** This tool was used for spell check, space indentation and clustering of related groups. Some of the cause names were really big and verbose and had scope for reduction. Such rows were edited.
Example: “**Bankruptcy or Sudden change in Economic Status**” could be reduced to a simple “**Bankruptcy**”. See Fig(1) for a part of the open refine changes.

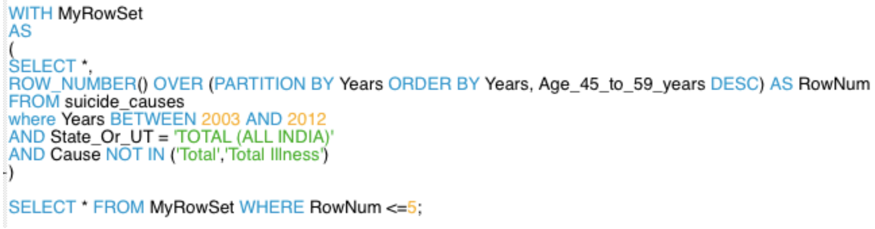
b) **Excel:** Since the CSV data was required to be imported to MySQL DB for data extraction, some of the mischievous column names had to be renamed.
Example: **STATE/UT** was a column name that might have troubled while fetching the records from that column as “/” acts like an arithmetic expression in queries. Hence, it was renamed to **State_Or_UT.**

3. Dataset was **formatted** to a small extent. Since there were designated columns, for male and female, under each age group, to find total count of a particular age group

meant fetching male under that age group and female under that age group and adding the two values. This would simply waste process time and it seemed efficient to rather add the two columns and store it in a new column on excel sheet for easy access.

4. Data was **imported** into **MySQL DB** (8.0.13). MySQL work bench was used to try out a couple of queries to understand the data better. Final queries were written to fetch the required data for the intended visualization. Example: SQL query to fetch top five reasons for suicide of various ages was constructed. See Fig(2).

Fig(2) SQL query to fetch top five reasons for suicide of various ages



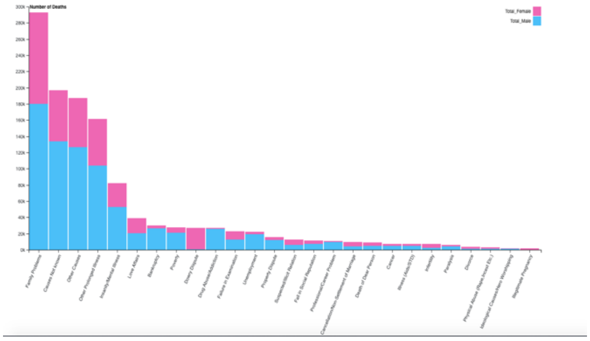
```
WITH MyRowSet
AS
(
SELECT *,
ROW_NUMBER() OVER (PARTITION BY Years ORDER BY Years, Age_45_to_59_years DESC) AS RowNum
FROM suicide_causes
where Years BETWEEN 2003 AND 2012
AND State_Or_UT = 'TOTAL (ALL INDIA)'
AND Cause NOT IN ('Total','Total Illness')
-)

SELECT * FROM MyRowSet WHERE RowNum <=5;
```

5. Setup **middleware** via **NodeJS** and connected to local MySQL server and the required schema.

6. Setup **frontend** architecture via **AngularJS.** Used **bootstrap** library to build responsive web pages. All the frontend libraries use CDN link rather than the local downloaded versions of the same.

7. Finally, the visualizations were achieved with the help of **D3 JS (v5).**

Fig(3) Stacked bar graph showing suicides for each cause segregated by gender.

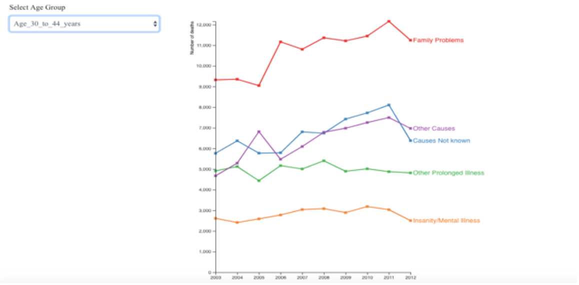


a) Stacked Bar Graph:

Reason: Multiple categories with further segregation in a compact space.

Visualization: This is a static graph indicating all the causes (categories) on x-axis and number of suicides on y-axis. The two stacks indicate the female and male deaths for a particular cause. On hovering over the stack, a pop-up title indicates the number of deaths for that stack and under that particular cause. Legend specifies the stack categories and the colours representing them. See Fig(3).

Fig(4) Multiple Line Graph showing top 5 reasons for suicide based on age group selected.

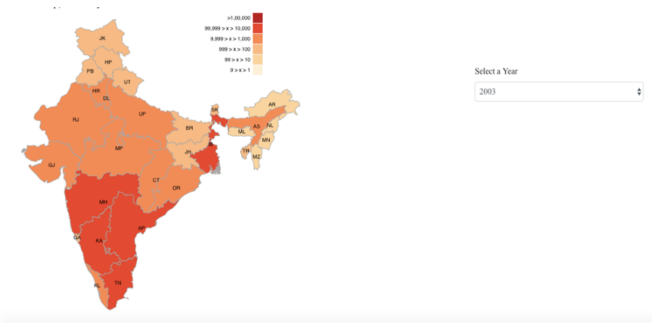


Select Age Group

Age_30_to_44_years

b) Multiple Line Graph:

Reason: Line graph is best suited for comparison over time. Used a higher version of the same to compare various categories, here causes, over time.
Visualization: This is an interactive graph that shows top 5 reasons for suicide based on age group. A dropdown is provided with various age groups to select from. Different values can be selected to see the changes in graph based on the age group selected. See Fig(4). **Note: The domain of y-axis is fixed to [0,12162] since the max number of suicides doesn't exceed this. The reason for doing this is to understand the graphs better. If the domain scales with age group, we cannot make sense of influence of one group over the other visually.**



Fig(5) Map showing number of suicides, per year, for all the states.

c) Map:

Reason: Maps are known to be the best visualization for geographic data. Since the dataset contains suicide count for various states and UTs for various years, map seemed to be the best fit for visualizing such large dataset at once.
Visualization: After studying the data, 6 ranges were chosen to show the suicide counts on map as indicated by legend. The colour gradient or saturation indicates the number of deaths. The greater the saturation, bigger is the count. This map can be studied for changes over time by changing the year from the dropdown provided alongside of map. See Fig(5).

Results

Analysis

1. Majorly, **family problem** seems to be the major reason for suicide.
2. Reasons for suicide does differ for men and women. Like **dowry dispute** seems to have impacted **more women** than men. Similarly, **professional/career problem** seem to affect **more men** than women.
3. The reasons for suicide does change with age group. Like **failure in examination** is one of the top reason for suicide for youngsters of **14 years and below** but this is not the case for other age groups.
4. Map shows that most of the **southern states of India** have greater number of suicides. It remains consistent over years. It could probably be because of larger area and hence more population and thereby more probability of suicides.

Principles Followed

1. **Trustworthy** : Inappropriate colours and fonts were not used. Chart junk eliminated.

2. **Accessible:** Appropriate legends, scale adjustments were made for easy viewing without scrolling. Even though the stacked bar chart seemed bulky, it was necessary to get the over-all understanding of all **26 categories** with gender segregation.
3. **Elegant:** maintained elegancy by adjusting font size of name of line graphs – readable and elegant.

Future Work

1. Animations could be added to make it more attractive.
2. Map visualization could be improvised using better libraries.