CA682 Data management and visualisation

Name	Akash Bhargava
Student Number	18210613
Programme	MSc in Computing (Data Analytic)
Module Code	CA682
Assignment Title	Data Visualisation
Submission date	16-10-2018
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: Akash Bhargava Date: 16-12-2018

Introduction:

Data Visualization is a field where we try to understand complicated data with the help of many visual representations be it graph, charts or any other forms of visualization. This helps in the process of data analysis which ultimately helps in data analytics.

In a world of progress, everything comes at a cost. In this visualization assignment I have tried to implement and analyse the suicide rates of the countries around the world. Before the implementation of this visualization, I was under the impression that the countries with good progress rates i.e. America, Russia, China and European countries will be impacted by lesser suicides compared to developing and underdeveloped countries.
But the results obtained were quite astonishing.

We will begin this report by discussing the dataset that was used for this visualization.
Next, we will discuss all the processes (cleaning, pre-processing and processing) and tools involved to create the visualization.
Towards the end of this report we will discuss the findings and inferences from the created visualization.

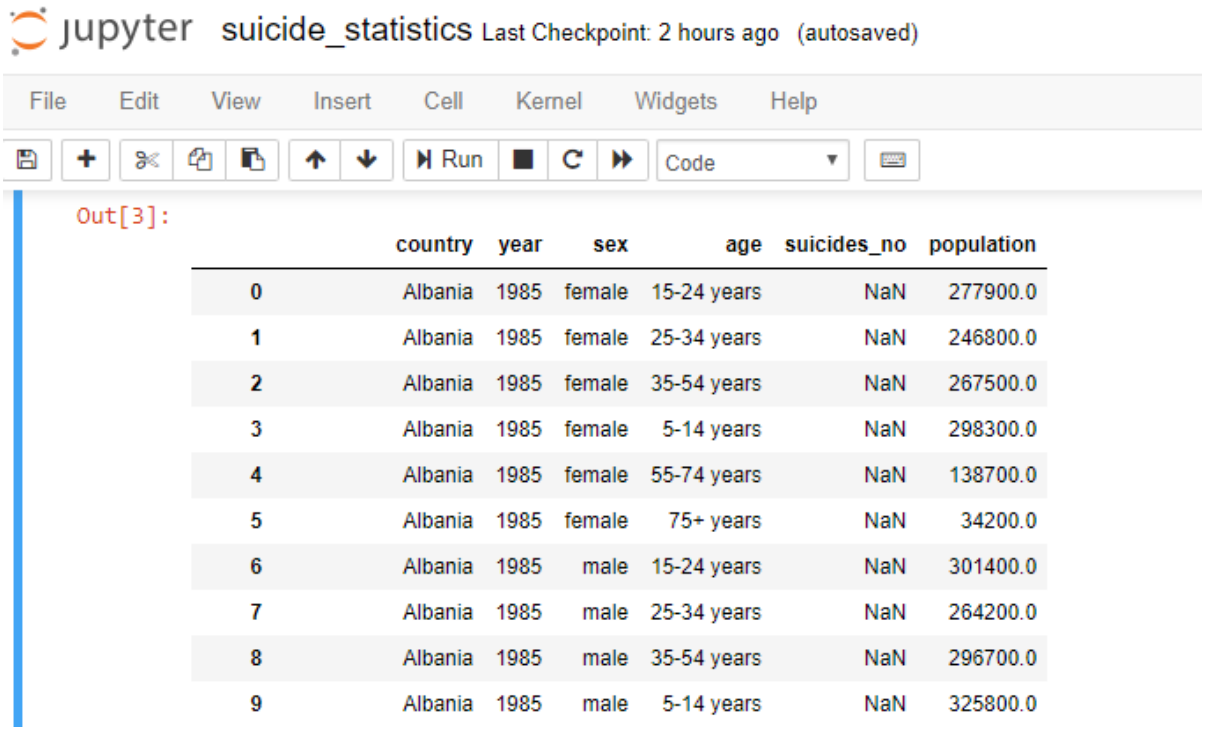
Data Set:

For this visualization I took the dataset of World Health Organisation. The data set is called who_suicide_statistics. It comprises of following attributes:

1. Country
   It comprises most of the countries round the world.
2. Year 1979-2016
   The data is divided yearly from 1979-2016
3. Sex M F
   The dataset is divided between males and females
4. Age: Age group
   The data is categorised based on the age groups typically 5-14 years, 15-24 years, 25-34 years, 35-54 years, 55-74 years and 75+ years.
5. suicides_no Number of suicides
   It gave the number of suicides for that category i.e. Country, Year, Sex and Age.
6. population Number of all living people
   It gave the population count for that category i.e. Country, year, sex and age.

Process:

1. **Selection of Data Set:**
   I used this dataset called who_suicide_statistics from kaggle.com https://www.kaggle.com/szamil/who-suicide-statistics. The main purpose was to figure out the regions with the suicide rates and try to infer some meaning out of it. As discussed above the dataset had 6 columns and 43777 rows. Below is the screenshot of the dataset.



	country	year	sex	age	suicides_no	population
0	Albania	1985	female	15-24 years	NaN	277900.0
1	Albania	1985	female	25-34 years	NaN	246800.0
2	Albania	1985	female	35-54 years	NaN	267500.0
3	Albania	1985	female	5-14 years	NaN	298300.0
4	Albania	1985	female	55-74 years	NaN	138700.0
5	Albania	1985	female	75+ years	NaN	34200.0
6	Albania	1985	male	15-24 years	NaN	301400.0
7	Albania	1985	male	25-34 years	NaN	264200.0
8	Albania	1985	male	35-54 years	NaN	296700.0
9	Albania	1985	male	5-14 years	NaN	325800.0

2. Data Cleaning:

I used Python (Jupiter Notebooks) to clean the data. Following modifications were applied:

- There were a few rows where the “suicide_no” (number of suicides) was not given. Since the data was not available and could not be simply assumed I removed those rows.
- There were a few rows where “population” was not given. Using Python, I calculated a mean population of that country using other years above the missing year and then autofill the values in the blank rows.
- After this I went on to add a new row which I called “suicide_per”. This row signified the ratio of number of suicides(suicide_no) and population. I realised that this is a better criterion to judge countries based for suicides rather than just the number of suicide itself.
- Next, I applied a group by function based on countries to get the data for every country across the complete time duration.
- Lastly, I added another column which categorised the countries based on the “suicide_per”. I created six categories based on the available data and categorised the countries accordingly.

Finally, after cleaning and processing data looked like:

:

	population	suicides_no	suicide_per	Country	Categorisation
0	8.699271e+07	1970.0	0.924752	Albania	6. Least Suicidal Countries (<1 Suicide %)
1	2.054919e+06	11.0	0.179140	Antigua and Barbuda	6. Least Suicidal Countries (<1 Suicide %)
2	1.190238e+09	93862.0	4.675511	Argentina	4. Moderate Suicidal Countries (3-6 Suicide %)
3	1.067432e+08	2422.0	1.273087	Armenia	5. Less Suicidal Countries (1-3 Suicide %)
4	1.762045e+06	120.0	1.817226	Aruba	5. Less Suicidal Countries (1-3 Suicide %)
5	6.443903e+08	80279.0	5.625449	Australia	4. Moderate Suicidal Countries (3-6 Suicide %)
6	2.873090e+08	60179.0	11.000488	Austria	2. Very High Suicidal Countries (9-12 Suicide %)
7	1.664882e+08	3366.0	0.888299	Azerbaijan	6. Least Suicidal Countries (<1 Suicide %)
8	7.715548e+06	107.0	0.439598	Bahamas	6. Least Suicidal Countries (<1 Suicide %)
9	1.675393e+07	463.0	0.467223	Bahrain	6. Least Suicidal Countries (<1 Suicide %)
10	7.590791e+06	205.0	1.031649	Barbados	5. Less Suicidal Countries (1-3 Suicide %)
11	3.164584e+08	74974.0	9.905211	Belarus	2. Very High Suicidal Countries (9-12 Suicide %)
12	3.587922e+08	75948.0	9.836461	Belgium	2. Very High Suicidal Countries (9-12 Suicide %)

Here the population represents the total population of the country over the whole time Period of the database. It should not be confused with the current population. Similarly, it's the same with suicide_no which is the aggregate value over the years.

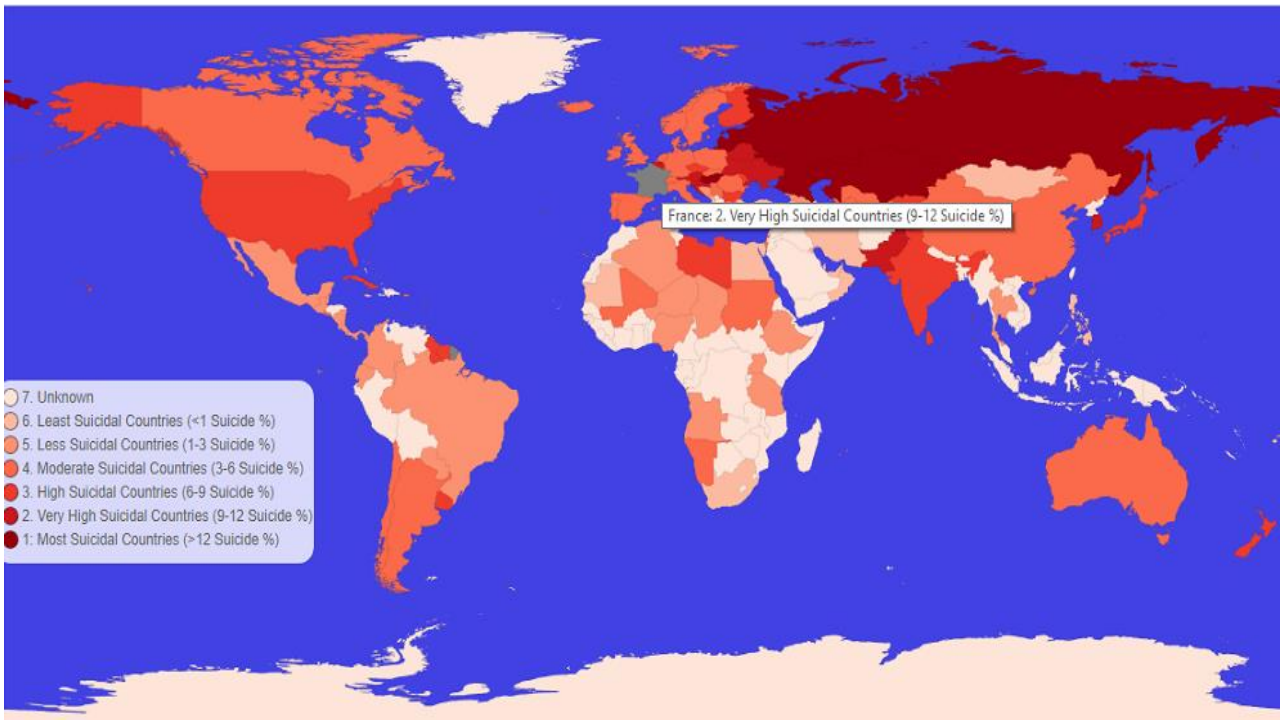
3. Selection of Chart Type:

It is a very crucial step to select the right visualization for the given data. At the beginning, I wanted to go with the standard bar chart which would have shown the "Top 10 countries" with the highest suicide percentage. But a big negative of this type of graph was to not be able to represent every country's information.
After a detailed analysis, I decided to go for a Choropleth Map. In this way I was able to represent the information of all the countries of the world without having to overload my screen with too many figures and facts.
Choropleth Map is a thematic map of a region where areas are coloured/shaded in accordance with the information provided for example population density, GDP, per-capita income and so on.
I used Choropleth map to represent the suicide percentage of every country.

I used D3.js to implement my visualization. Being a new language, it was a bit difficult initially to understand the concepts of SVG, JavaScript, JSON etc. But in the end, I was able to appreciate the flexibility of this environment. I used TopoJson library to implement the Choropleth map.

RESULTS:

As mentioned before, I wanted to represent the data of every country. Thus, the reason for selection of Choropleth map. Below is the screenshot of my visualization.



In this plot, we can see a world map which represents all the countries around the world. For the start I used a TopoJson library to implement the world map. Following are the highlights of the visualization:

- I used the Mercator projection for the world map because my aim was to represent each country clearly through a familiar map. My aim was not related to representation of the exact size of the countries. Thus, the reason for the selection.
- I used blue colour to represent the waters of the ocean. It's a common practise to represent water in blue.
- I used shades of red (schemeReds) to represent the suicide percentage, because I felt something like suicide percentage should be represented in a darker tone. There are six major categories namely Most Suicidal Countries (>12 Suicide %), Very High Suicidal Countries (9-12 Suicide %), High Suicidal Countries (6-9 Suicide %), Moderate Suicidal Countries (3-6 Suicide %), Less Suicidal Countries (1-3 Suicide %) and Least Suicidal Countries (<1 Suicide %).
  All these are based on suicide_per column of the dataset. A final category "Unknown" to represent those countries whose data is unavailable in the given dataset.
- Based on the categorisation I explained above, I created a colour legend. I added this colour legend on the left bottom of the world map because of the absence of countries in that region.
- On hovering over a country following things happen:
  - The country gets shaded in grey using hover style of CSS.
  - We get to see the name of the country and the category it belongs to.
  -

After visualizing the map, we can infer that, economy is not the only factor responsible for suicides because developed countries like France, Russia, USA etc have large suicide rates whereas countries which are considered to have a bad economic status i.e. Brazil, Turkey, Mexico have less suicide rates.

In conclusion, this visualization discovers a new question for discussion which states that what other factors might be involved that might cause in increase of suicides in these countries.

Criticism and Improvements:

The graph does not contain information of a few countries due to absence of data in the dataset. A separate category “Unknown” has been created for its representation, but it does not defy this fact.
Given more time, an improvement that I would have like to do in this visualization is categorisation i.e. I would have liked to add a panel where I could add the categories i.e. year, age-groups etc. Based on selected category the data would change. This would have given an improved reflection of the dataset.