

# INFS3208 Project by Shristi Gupta

## Suicide Statistics Data Analysis




### Proposal

This project is a 'Type II Project: Big Data Focused – Big Data Queries'. It revolves around the analysis of Suicide Statistics from 1979 to 2016 provided by the World Health Organisation (WHO). The link to the dataset used is <https://www.kaggle.com/datasets/szamil/who-suicide-statistics?resource=download>. The dataset is divided by country, year, age groups, sex and it provides the number of suicides for each division as well as the population. Suicides have become a major cause of deaths of individuals. The numbers are increasing as ever. Suicides are becoming a major concern worldwide and thus this project focuses on more than 110 countries. The overall objective of this project was to analyse the number of suicides in different countries, by different sex, amongst different age groups and by different years ranging from 1979 to 2016.

This project consists of different queries written using Spark SQL. I've imported all the necessary python modules such as pyspark and matplotlib followed by the creation of a spark session in a Jupyter Notebook. Data exploration and pre-processing includes loading the dataset from Hadoop Distributed File System (HDFS), getting a summary of the data, looking at the schema, counting duplicate rows (which resulted in null) and dropping null values. The queries include finding the top 10 countries with the most suicides which resulted in Russian Federation being at the top with a total suicide count of 1500992. The query finding the top 10 years leads to the conclusion that suicides were at their peak from 1995 to 2004 in the total range of 1979 to 2016. The leading year was 2002 and thus I wrote a query finding the top 5 countries with the most suicides in this year. Yet again, the Russian Federation gave us the highest number. When comparing the distribution of the number of suicides among males and females, one finds that almost more than 75% of the total number of suicides were by males. This prominently shows the mental health condition of males as compared to females. An interesting revelation found was that Russian Federation again comes at the top when finding countries where males died but when it comes to females, Japan emerges at the top. Ukraine emerges in the top 5 countries where males died while Republic of Korea emerges when finding the top 5 countries where females died. The query finding the total number of suicides by age groups resulted in 35-54 years having most number of suicides and 5-14 years having the least. While comparing the total number of suicides versus population index, we come across another interesting conclusion. One would expect Russian Federation to be at the top but Lithuania and Hungary come before it. This owes to the massive population of Russia. I also found the total number of suicides in Australia over those years as a personal interest and it came to be 80279 which is far less than many other countries. Along with the queries, I've also provided visualisation of the results found in the form of bar chart and pie chart.

As the name suggests, Big Data is big. Along with that, it's fairly modern thus making it obvious that traditional computing has limitations when it comes to big data. Traditional computing is unable to deal with the enormous size of big data. The efficiency of the traditional data storage systems decreases when it comes to big data and it becomes slow.

It is unable to keep up with the demands of modern data when looking for trends. Each operation on big data takes a long time when using traditional computing. This makes it highly unsuitable for carrying out queries on huge datasets containing tens of thousands of rows. Cloud computing services such as Hadoop help this project as it can easily store a huge dataset over a number of worker nodes. This leads to increased storage as well as more processing power. Hadoop is cost effective and scalable, thus making it suitable for managing and processing big data. Apache Spark optimises the execution of queries for analysis of big data. This project uses Docker, Apache Spark (PySpark), HDFS and a VM on GCP. Docker Compose shares multi-container applications. Using a single Docker Compose YML file, I've defined all the services being used in this project and I can run all of them using a single command. I've used two Spark worker nodes along with the master node. PySpark helped me in using Apache Spark with Python. It allowed me to create and work on data frames. Moreover, it provides SQL libraries that make it very easy to run SQL queries on the dataset. This project uses HDFS to store the dataset over three worker nodes. The major advantages that it offers is that is fast and open-source.

Compute Engine	
1 x	  
Region: Iowa	
108,631 total hours per month	
Provisioning model: Regular	
Instance type: n1-standard-2	USD 10.32
Operating System / Software: Free	
Estimated Component Cost: USD 10.32 per 1 month	
Persistent Disk (Accompanying)	
1 x boot disk	
Product accompanying: Compute Engine	
Zonal standard PD: 50 GiB	USD 0.80
USD 0.80	
A portion of your estimate fits within the <a href="#">Persistent disk free tier</a> .	
Total Estimated Cost: USD 11.12 per 1 month	

