# Data Analysis on World Health Organization dataset

We took open data from World health organization website for our analysis of death due to unnatural causes worldwide. These are broadly classified as Transport accidents, Falls, Accidental drowning and submersion, Exposure to smoke, fire and flames, fire and flames, Accidental poisoning by and exposure to noxious substances, Intentional self-harm(Suicide), Assault and all other external causes- Total and both sexes. We realized we needed additional data like population, GDP Ranking, Income level of the countries and got those information from Wikipedia and combined them with our data thru excel as pythonic code would have taken another week of our time.

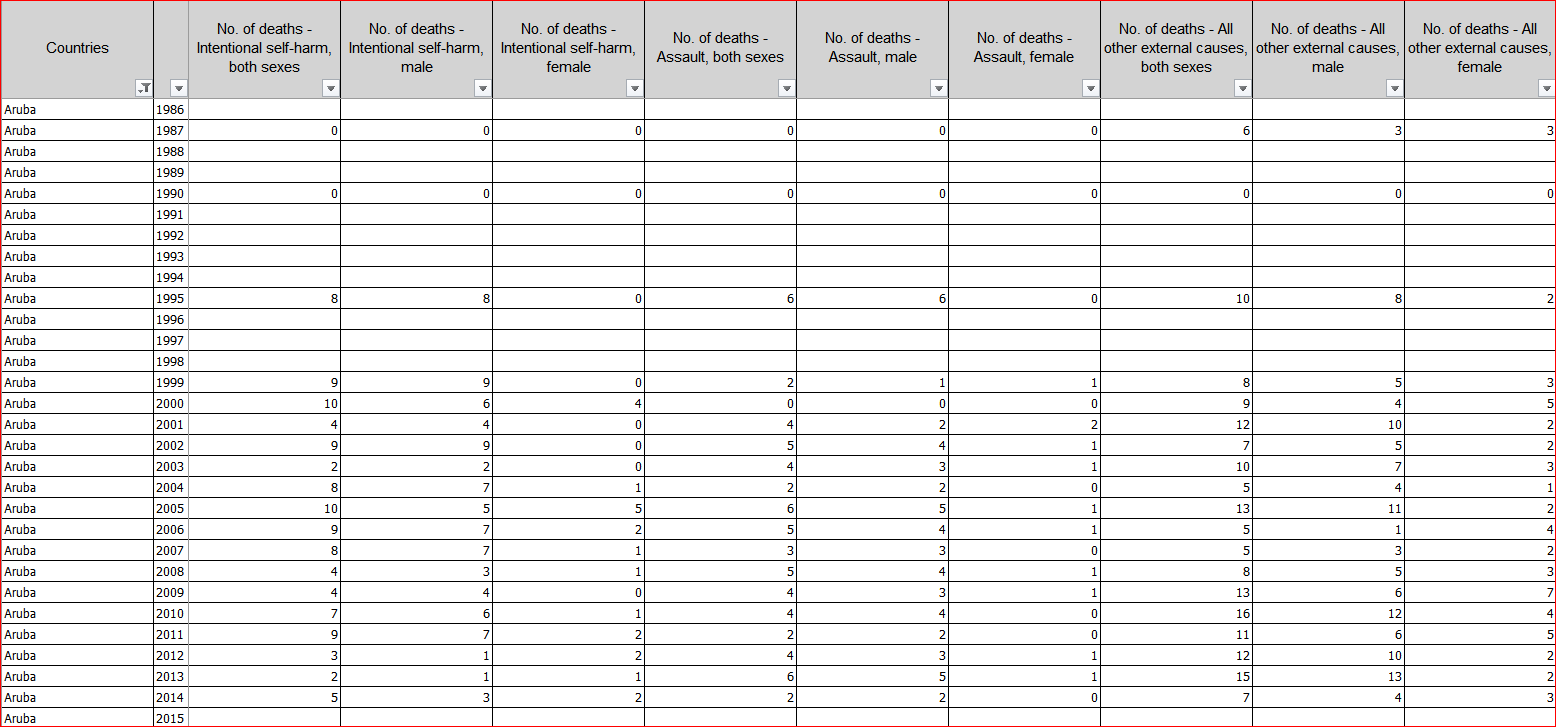
**Data preparation/Challenges**

The data came from multiple sources and multiple format. We tried to combine them thru python, but due to limited time frame, decided to combine them thru Excel and then clean them with python. The data we got from WHO data website was from year 1985 thru 2015. When we analyzed the data, it was missing information for few years for most of the countries.

1. Missing data

Some countries had data from 195, but missing several years in recent years, some were missing both at the beginning period and ending period.

For e.g. Cuba.



We used forward fill method to populate the data for future years and followed by backward fill to populate the beginning years as our first country in the set ‘Albania’ didn’t have data for first few years. We encountered a problem with that approach as the data of one country used to prefill with another. We added the condition to check for the country and make sure we apply **fillna** to same country for the given set of years. We looped thru each country one by one to apply this function.

1. Column heading were long and sentence like. We created a metadata file which would check for strings in the spreadsheet and assign the column name dynamically. Code would check for Text1 and Text2 in the header and assign the column name from corresponding row in column directory.

|  |  |  |
| --- | --- | --- |
| Cols | Text1 | Text2 |
| Countries | Countries |  |
| Year | Year |  |
| IncomeLevel | Income | Level |
| TotalPop | Total | Population |
| PopMale | Population | Male |
| PopFemale | Population | Female |
| TransportAccAll | Transport accidents | both |
| TransportAccMale | Transport accidents | , male |
| TransportAccFemale | Transport accidents | female |
| FallsAll | Falls | both |
| FallsMale | Falls | , male |
| FallsFemale | Falls | female |
| DrowningAll | Accidental drowning | both |
| DrowningMale | Accidental drowning | , male |

1. When the data was read from the excel file, Year 1985 would appear as 1985/12/31 by default. We used the pandas datetimeindex.shrfttime function to format the year as %Y(CCYY). Added the year as index.

We created 2 spreadsheets one with absolute values and another with per 100 k numbers.

Tools used:

Excel spreadsheets

Jupiter Notebook

numpy–scientific computing

pandas –data frames

matplotlib–2D plotting

scikit-learn – KNN, regression Analysis

Math

Datetime

Algorithms

Pre-processing

Performance evaluation

And more …

**Analysis**

**Objective**

Insurance companies selling Global products need to access countries mortality rate, risk factors, overall income level while coming up with a rate table to adequately cover for the risk or to find out if the market is viable to do Business. Also, another advantage of accurate prediction for the travel minded people is to access the safety threat each country poses.

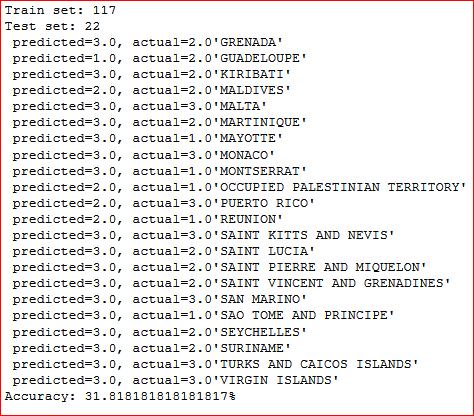
**Classifiers: KNN model:**

We tried KNN model with the intention of finding the unknown(covariance) from a group of

variables i.e. types of death.

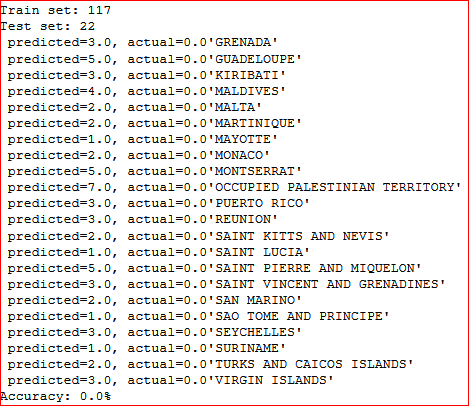
In the first test, we used the All types of death per 100k and tried to predict the income level of the country. We used the data from 2012 and used the 117 observations as Test dataset and 22 as training. The maximum accuracy we got with various combination was 49%.

**Income level Ranking: 1 Low Income to 3 High Income**



**Global Peace Ranking: 1 being Safest to 7 Unsafe**

Second test was to see if we can predict the Global peace ranking. We used the ranking from Wikipedia for 2012. We split the countries into 7 batches with 1 being safe to 7 extremely unsafe. Even though most of the countries the data existed, there were several countries it was missing. We used the missing countries as test dataset.



Obviously thru KNN, the accuracy would be 0% as the missing countries began with zero or none values. On a closer look, it came close for several countries like Monaco, which was ranked as 2. Upon google search the result came back as Monaco is indeed safest.

“The safest place in Europe - **Monaco**. If **safety** is your number one priority in a destination, then **Monaco** is for you. Security for this tiny constitutional monarchy is provided by an armed national police force consisting of 515 men and women, giving it the reputation of being the "safest square mile" in Europe” 🡺Apr 16, 2017

<https://www.worldnomads.com/travel-safety/europe/monaco/safest-spot-in-europe-monaco>

Maldives, a tourist destination was ranked among the countries in category 4.

### *“*[*The real situation in Maldives for tourists after Riots. Is it safe ?*](https://www.dreamingofmaldives.com/maldives-blog/the-situation-in-maldives-for-tourists-after-riots-is-it-safe/)

https://www.dreamingofmaldives.com › Home › News

*You have probably read and heard that more violence occurred on the* Maldives *following the coup of Tuesday, February 7,* 2012*. Since then, the most enticing titles broke out in the media: I quote: “the violence spread to other islands,” “anarchy won several islands of the archipelago” and other explosive slogans .”*

Montserrat and Guadeloupe, both was predicted to be 5 ranking category and again a simple google search revealed that these countries were not safe after all.

*“Travel to the Exclusion Zone on the south end of the island is generally not permitted, for* ***safety*** *reasons. The* ***Montserrat*** *Volcano Observatory publishes current risk assessments and exclusion zone limits.* ***Montserrat*** *is generally a* ***safe*** *place. However in recent years, violent crime has increased.*

*Guadeloupe is a beautiful French Caribbean island great for visiting and relaxing on the beach. However, like most places, Guadeloupe comes with a few warnings and dangers to avoid on your trip.”*

Same with Palestine as unsafe with 7.

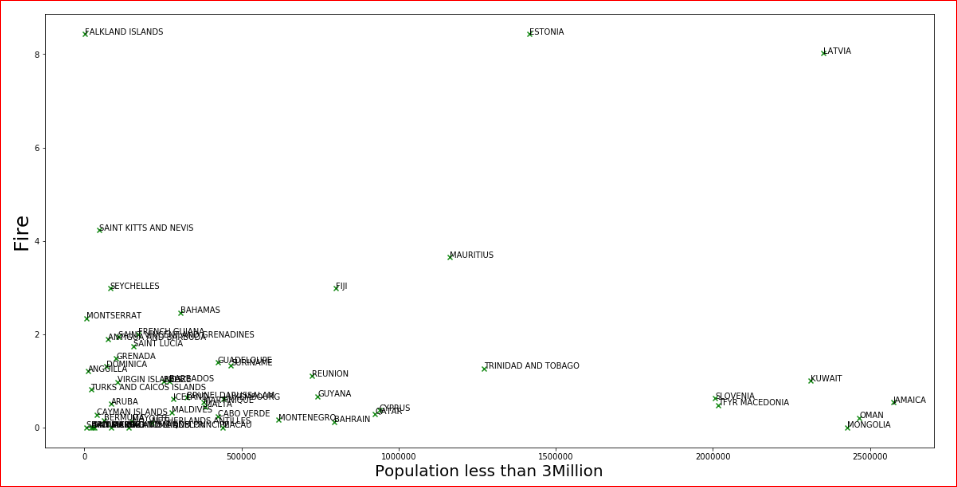
But whereas Saint Pierre and Miquelon with unsafe ranking of 5 seems to be safe according to google search. Hence even these predictions with KNN is a hit and miss.

“There is little crime in Saint Pierre and Miquelon and this destination should be considered one of the safest possible in North America***. “***

**Classifiers: Scatter graph**

Below are the Scatter plots for 2 categories of country. One for smaller countries and another for larger countries with population more than 100M. This considers the mean population vs fire death per 100 K mean as coordinates. We started with scatter plot of 2011, 2012 and since they looked similar took the mean value for last 30 years and some countries stand out compared to the others. Again, cursory inspection on google corroborates the findings on the plot.

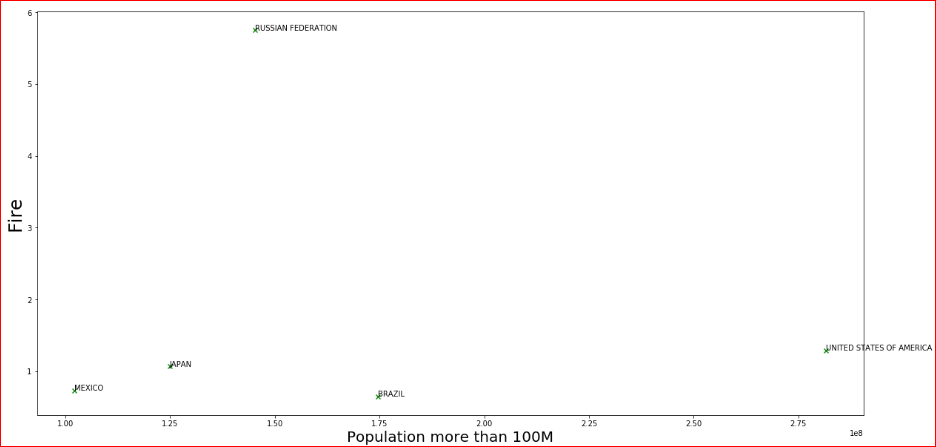
**Fire vs Population graph for countries with less than 3M population**



### “[Killer fire shows Latvia is one of the deadliest countries in Europe](https://www.baltictimes.com/news/articles/17402/)

https://www.baltictimes.com/news/articles/17402/

Feb 28, 2007 - GAPING HOLES: The Alsunga fire highlighted Latvia's poor fire safety practices and high rate of fire-related deaths. RIGA - Latvia's notorious fire safety record made international headlines last week when a blaze ripped through the attic of a three-story convalescent home for disabled people, taking 25 ...”

**Fire vs Population graph for countries with less than 100M population** 

### *“*[*Russia has the most severe fire season in a decade – Wildfire Today*](http://wildfiretoday.com/2012/09/14/russia-has-the-most-severe-fire-season-in-a-decade/)

wildfiretoday.com/2012/09/14/russia-has-the-most-severe-fire-season-in-a-decade/

*Sep 14, 2012 -* Russia *has the most severe* fire *season in a decade. Satellite photo,* fires *in* Russia*, September 11,* 2012*. NASA (click to enlarge). The photo above, taken by NASA's MODIS (Moderate Resolution Imaging Spectroradiometer) satellite on September 11, got my attention due to it's artistic nature. I have seen a ...”*

**Conclusion**

Above analysis shows how relatively easy in python to do the data analysis which helps Insurance company or Travelers/Websites to put the information about Countries safety in terms of types of death and how well they need to be prepared to mitigate the risk and in some circumstances, avoid the risk. There is always some sort of concern about missing data and how it affects the data quality.

About using Python for data analysis, even though Python data frames can provide similar functionality like Excel, it has a steep learning curve in terms of gathering data and combining them from multiple sources. Often it can be frustrating in the beginning to do simple tasks in python, which can be done with a click of few buttons on Excel. However, in terms of cleaning and aggregating the data, once the model is built on Python, it can be easily reusable and notebooks can be shared easily among peers for them to reuse. Amount of help available for various issues we faced and the plugin supports makes python one of leading contender for Data Analysis.