

Data analyzing towards unsafe water death percentage

Name: Tianle Pan (tpan501)

UoA ID:383296356

GitHub username: tpan501

GitHub Repository: <https://github.com/tpan501/iteration-4>

Content

[1. Situation understanding 2](#_Toc147568441)

[1.1. Business/situation objectives 2](#_Toc147568442)

[1.2. Assess the situation 2](#_Toc147568443)

[1.2.1. Resource inventory 3](#_Toc147568444)

[1.2.2. Requirements, Assumptions, Constraints 3](#_Toc147568445)

[1.2.3. Risks and contingencies 4](#_Toc147568446)

[1.3. Determining the data mining goal 5](#_Toc147568447)

[1.4. Data project plan 5](#_Toc147568448)

[2. Data understanding 6](#_Toc147568449)

[2.1. Initial data collection 6](#_Toc147568450)

[2.2. Describe the data 9](#_Toc147568451)

[2.3. Explore the data 12](#_Toc147568452)

[2.3.1. Pre-settings 12](#_Toc147568453)

[2.3.2. Data visualization 14](#_Toc147568454)

[2.4. Verify data quality 28](#_Toc147568455)

[3. Data preparation 32](#_Toc147568456)

[3.1. Select the data 32](#_Toc147568457)

[3.2. Integrate various data sources 33](#_Toc147568458)

[3.3. Clean the data 35](#_Toc147568459)

[3.3.1. Missing value 35](#_Toc147568460)

[3.3.2. Outliers and extremes 39](#_Toc147568461)

[3.4. Construct the data 42](#_Toc147568462)

[3.5. Format the data as required 44](#_Toc147568463)

[4. Data transformation 45](#_Toc147568464)

[4.1. Reduce the data 45](#_Toc147568465)

[4.2. Project the data 46](#_Toc147568466)

[5. Data mining Method 47](#_Toc147568467)

[5.1. Data mining method candidate 47](#_Toc147568468)

[5.2. Data mining method selection 48](#_Toc147568469)

[6. Data mining algorithm 48](#_Toc147568470)

[6.1. Exploratory analysis of data mining algorithm 48](#_Toc147568471)

[6.1.1. K-Nearest Neighbors 48](#_Toc147568472)

[6.1.2. Decision tree 49](#_Toc147568473)

[6.1.3. Random Forest 49](#_Toc147568474)

[6.2. Selecting data mining algorithm 49](#_Toc147568475)

[6.3. Build model with the algorithm 49](#_Toc147568476)

[7. Data mining 51](#_Toc147568477)

[7.1. Create and justify test design 51](#_Toc147568478)

[7.2. Conduct data mining 52](#_Toc147568479)

[7.3. Search for patterns 53](#_Toc147568480)

[7.3.1. Pattern 1 55](#_Toc147568481)

[7.3.2. Pattern 2 56](#_Toc147568482)

[7.3.3. Pattern 3 56](#_Toc147568483)

[8. Interpretation 57](#_Toc147568484)

[8.1. Study and discuss the mined pattern 57](#_Toc147568485)

[8.1.1. Interpret pattern 1 57](#_Toc147568486)

[8.1.2. Interpret pattern 2 57](#_Toc147568487)

[8.1.3. Interpret pattern 3 57](#_Toc147568488)

[8.2. Visualize the data, results, models and patterns 58](#_Toc147568489)

[8.3. Interpret the results, models and patterns 59](#_Toc147568490)

[8.4. Assessing and evaluate results, models and patterns 60](#_Toc147568491)

[8.4.1. Decision making review 61](#_Toc147568492)

[8.5. Iterations 62](#_Toc147568493)

[Reference: 64](#_Toc147568494)

# Situation understanding

## Business/situation objectives

Good health and well-being have always been prioritized by the United Nations (UN) and they are keeping making great effort on it. However, there is one type of source of diseases which is continuing set people throughout the world in danger, and that is the water safety. According to the 17 Sustainable Development Goal (SDGs) from the Department of Economic and Social Affairs Sustainable Development (UNDESA), 2.2 million people are still having difficulty to get access to the safely managed water sources (UNDESA, 2023). Additionally, the number of people who died from using or drinking unsafe water reached 1.23 million, which is as serious as the global road crashes death. Therefore, it is very necessary for the data scientists to analyze the relevant data to find out the connection between certain factors and the occurrence of high death share from unsafe water usage. Why the unsafe water death share is bigger in some region? As the UN has set the well-being and clean water and sanitation as the SDGs, finding the reasons behind the data and figuring out the actions to benefit people’s well-being and life become a necessity.

In this iteration, I will use nine step data mining process to probe into the residents’ death from unsafe water. The objectives will be the same as the previous iteration, which is I will use a range of variables will be considered and analyzed to:

* Discover the underlying reasons behind the high death share of the regions and,
* Predict the death risk shifting in the next coming years.

This project will be considered as a success if:

* A clear correlation between the high death percentage and potential variables is found out and,
* The number of global deaths from unsafe water can be reduced by 5% in the next year.

## Assess the situation

Currently I am going to collect the data from the data set online and analyze the data by myself. There are some pre-requisites which need to be clarified before transiting to the next section.

### Resource inventory

**Data.**  Due to the reason that I am going to analyze the reasons behind the issues which is threatening human being’s right of life, there are different institution and organizations which can provide the data related to the situation of hazardous water. Apart from the departments under the UN, such as World Health Organization, a range of data set can be found in some third-party organization’s website. Those organizations provide the data set that they have collected to the public, motivating more people to make contribution to make the world a better place. To optimize my data mining accuracy, I will keep using the a data set recording the trend of death percentage of unsafe water sources within 20 years from a non-profit organization called “Our World In Data” (Ritchie & Roser, 2021) as I did in the iteration 3. On their website more data sets which focus on different domains and perspectives can be found such as the percentage of different countries’ people getting access to safe water source. Therefore, I will combine various data sets together and analyze them as a whole.

**Hardware.**  The main data mining process will be implemented on a HP ZHAN laptop 2022 which carries Gen Intel(R) Core(TM) i7-1260P and 32G memory. Because of the limited computation capacity of the CPU, the efficiency of the data mining process may be impacted negatively. Therefore, if the data mining process is not effective and efficient enough, actions need to be taken to change the situation, such as shifting the construction of the data set.

**Software.** This data mining process will be proceeded by using Jupyter Notebook. Jupyter is an open-source project, which is specialized in data analysis and mining. Jupyter notebook is one of the software belong to this project, and it is a web-application supporting people to do live coding, data visualization and equations. Different from Spyder and other IDE, Jupyter notebook is an interactive computing environment (Project Jupyter, n.d.). “Jupyter” is derived from three programming languages which are Julia, Python and R. For this iteration, I will mainly use Python to do the programming and the rest of data mining process, and several libraries that are supported by Python like pandas and pyspark may be used to contribute to the data mining process.

It is notable that for this iteration the uploading process will be done by using EC2 instance and GitHub. For this iteration, the Jupyter Notebook will be used on the EC2 instance. After finishing every modification, I need to use git bash code to push the changes to my GitHub repository.

### Requirements, Assumptions, Constraints

In order to finish the data mining process and carry out the final results and prediction smoothly, I understand that it will require me to go through the 9-steps data mining process from the beginning to the end. I need to collect the related data sets as many as I can that can show the potential effect on the death share of the people who died from unsafe water, and try my best to clean the data sets to make sure it does not have null and empty values to interfere the data mining.

There are also technical requirements for this iteration. I was required to use Python language on Jupyter Notebook to do the data mining with the library called pyspark, which is specialized to process a large amount of data parallelly.

For this iteration, I assumed that there are 4 types of possible phenomena which may happen in the end. The first one is I will be not able to find enough meaningful data to deduce the patterns of the phenomena. The second condition is there is no clear pattern for them, and the third one is the patterns and results may not be special enough to make contribution to the solutions to the problems. Meanwhile, the results and predictions will be relatively simple with high accuracy.

Those three assumptions to some extent express the constraints of the data mining process this time, which is that the analysis process may not cover all the factors that can lead to the death by using the unsafe water. Instead, this iteration will only provide a reference to remind people of the importance of several variables as they are having larger weight in causing this problem. However, this result may be too one-sided or short-sighted to find out the real connections and relationships.

### Risks and contingencies

|  |  |
| --- | --- |
| Risk | Contingency plan |
| There is not enough variables in the data set for data analyst to deduce patterns and correct results. | Do the brainstorm and collect more data from more data sets by all means. |
| As a new beginner with only Java language experience, even though I had done the iteration 3 with Python on Spyder IDE, using Jupyter Notebook to do data mining can be still challenging and may cause error during the process. Furthermore, the way to use pyspark is different from the one to use pandas. | More learning related to Python and pyspark needs to be done. The instructions from the provided Github repository should be fully made use of. |
| Data set is not able to carry out a meaningful result or pattern for people to work on plans to take actions. | The reason why the data mining process is not working smoothly needs to be elucidated. |
| I may encounter different kinds of error or bugs when I am committing changes and pushing them to my GitHub repository. Furthermore, I may also meet AWS related errors or incorrect operation. | I should get more familiar with the GitHub processing sequence and also get more information about AWS and EC2 instance. |

## Determining the data mining goal

The objective of this data mining process will be:

* Managing to choose an algorithm to create a model which can reflect the patterns behind the occurrence of death share of using unsafe water by using data from 2000 to 2019.
* Discover patterns that results in a country’s risk on death percentage caused by unsafe water.
* Using the information and output provided by the model, seizing the most important factors that cause the death.

## Data project plan

|  |  |  |  |
| --- | --- | --- | --- |
| Name / Title | Type | Start Date | End Date |
| iteration 4 - BDAS | project | 2023/10/2 | 2023/10/13 |
| First Group | group | 2023/10/2 | 2023/10/13 |
| Business understanding | task | 2023/10/2 | 2023/10/4 |
| Data understanding | task | 2023/10/4 | 2023/10/4 |
| Data preparation | task | 2023/10/5 | 2023/10/7 |
| Data transformation | task | 2023/10/6 | 2023/10/8 |
| Data mining method | task | 2023/10/9 | 2023/10/10 |
| Data mining algorithm | task | 2023/10/9 | 2023/10/10 |
| Data mining | task | 2023/10/10 | 2023/10/12 |
| Interpretation | task | 2023/10/12 | 2023/10/13 |

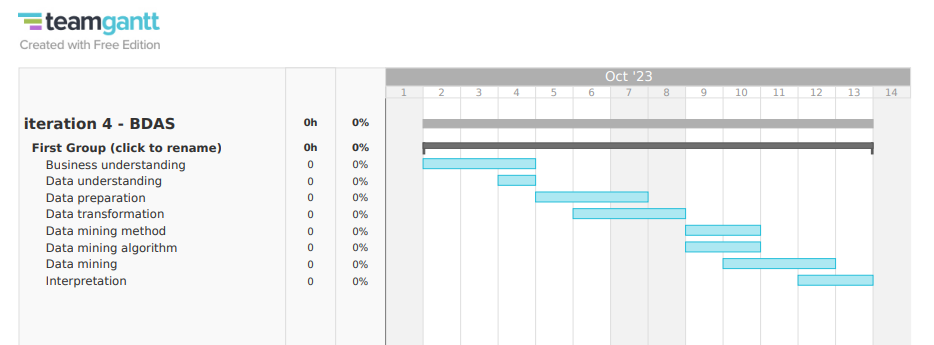


Figure Team Gantt time schedule

# Data understanding

## Initial data collection

The data will be mainly gathered from the NGO “Our World In Data” website, where there are plenty of statistics resources related to 17 SDG goals. “Our World In Data” is an organization where the team devotes themselves to discovering the issues and helping reforming the world. Their objective is preserving the knowledge in the form of data in the database, making people easier to get access to knowledge, and contributing to the solutions to the problems (Roser, n.d.). Most data sets will come from the “clean water” section in the website <https://ourworldindata.org/water-access>.



Figure Our World In Data homepage

Currently, as mentioned in section 1.2.2., I need to improve the data sets variety and diversity by finding more suitable data sets to support the data mining process. Fortunately, there are plenty of data sets recoding different data from various domains in “Our World In Data” website, which can be useful for this research, and the critical thing is which ones are worthy enough to collect.

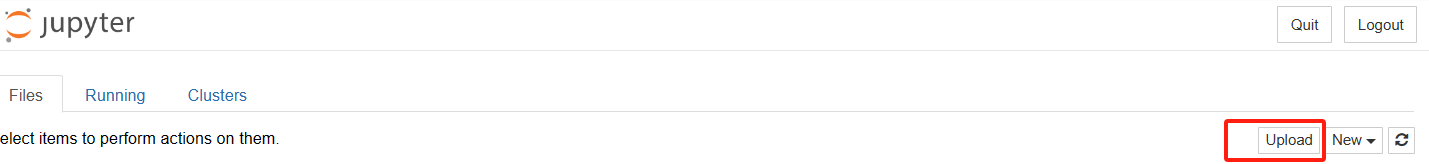
Therefore, I need to make a hypothesis here to make it more convenient to choose the data set I need. The assumption is “the higher death share is related with the local safe or clean water’s availability to the residents there, and government support also plays important role in it”.

Based on the hypothesis, I have collected some more data sets from the Internet. Apart from the death percentage and different water safe level, the extra added figures and tables are:

1. death-rates-unsafe-water.csv
2. deaths-from-diarrheal-diseases-who.csv
3. historical-gov-spending-gdp.csv
4. improved-water-sources-vs-gdp-per-capita.csv
5. number-without-improved-water.csv
6. number-without-safe-drinking-water.csv
7. people-practicing-open-defecation-of-population.csv
8. physicians-per-1000-people.csv
9. public-healthcare-spending-share-gdp.csv
10. public-health-expenditure-share-GDP-OWID.csv
11. water-bodies-good-water-quality.csv
12. healthcare-expenditure-vs-gdp
13. urban-improved-water-access-vs-rural-water-access

In iteration 4, although there were other data sets which may have relation to the symptoms, most of them did not contain enough data rows to support the accuracy of the final result. After selecting, a new data set is added into the data mining process, which is a file recording the percentages about people’s accessibility to safely-drinking water in both urban and rural area. Those data sets cover different domains which may have relationship with the large death percentage of some certain regions. However, because there are more than 10 data sets which need to be analyze as a whole, data analyst should carefully merge them and do the data cleaning and selection before entering the next stage.

To display the data sets, I shall begin the first step in the Jupyter Notebook, which is importing the data sets into the editor. Before that, I need to upload the data sets that I may require into the Jupyter Notebook so I am able to find the path when I want to use them in the data mining process.



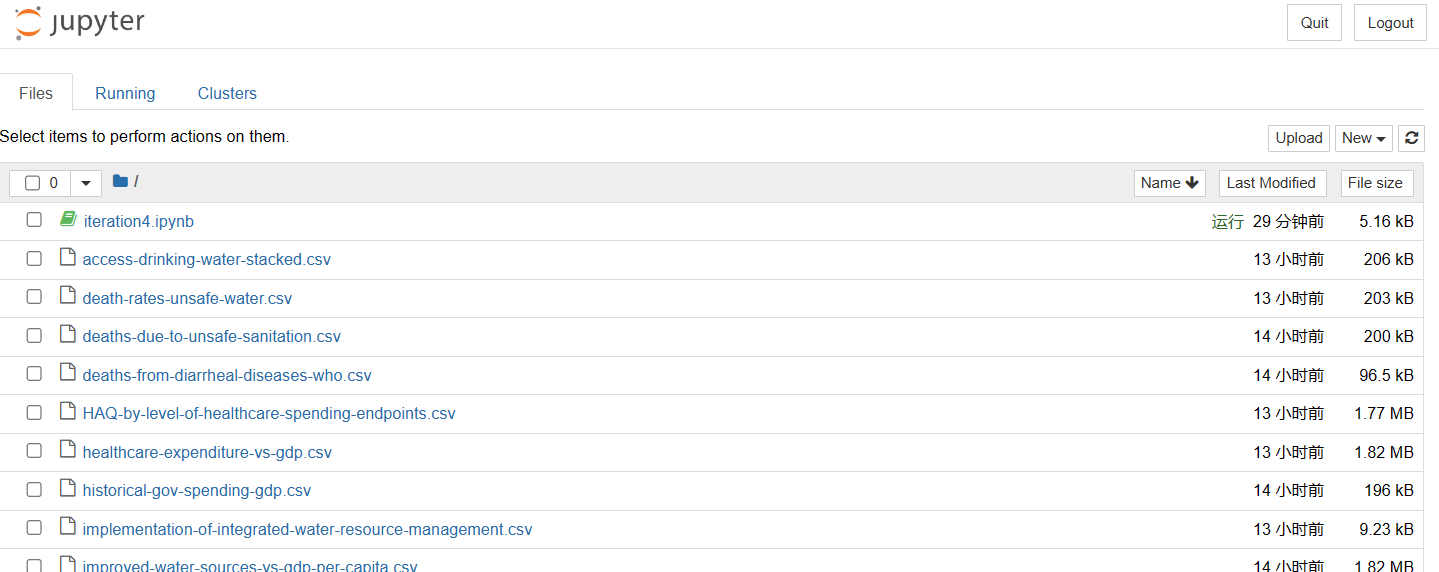


Figure Upload data sets

For this iteration, I was required to use pyspark, so the spark library should be imported at the beginning of the data mining process.

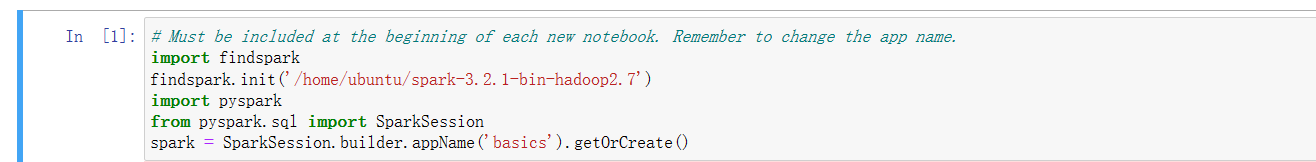


Figure Import pyspark

Then I need to import the files which were uploaded by me into the Jupyter Notebook.

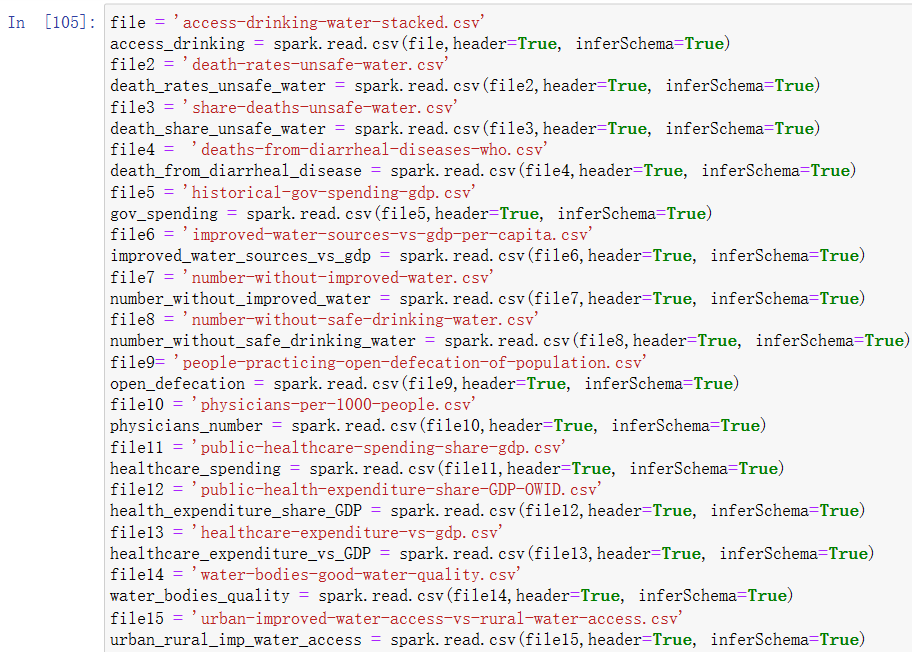
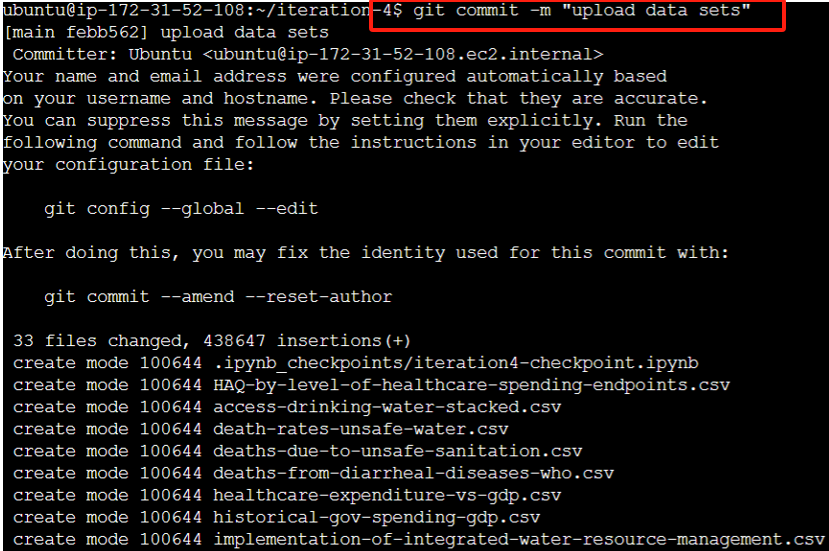
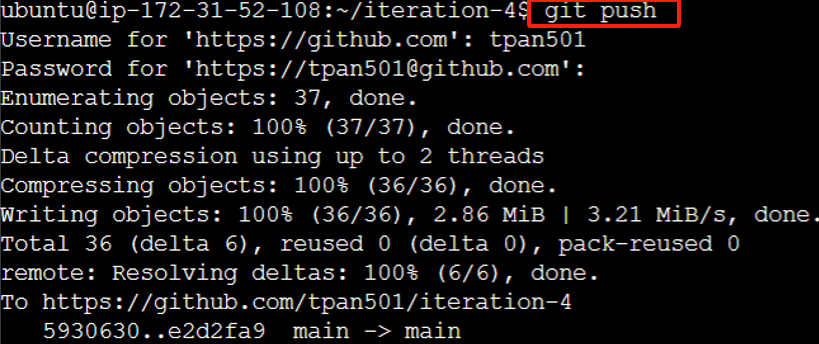


Figure Import data sets

Subsequently, I will commit this change and push it into my GitHub repository.





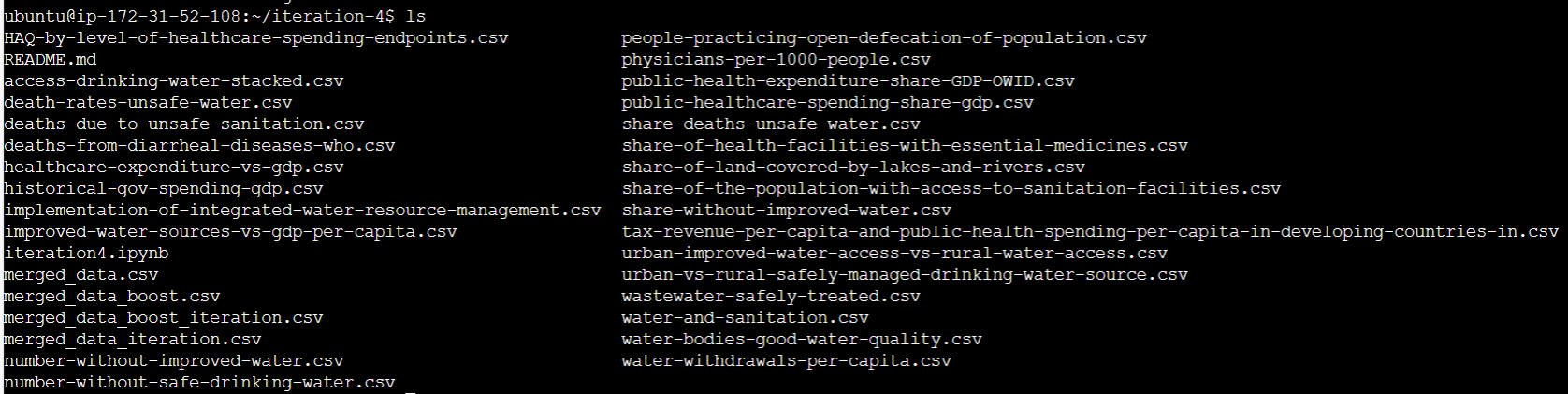


Figure Commit and push changes to GitHub

The repository after I pushed the changes is shown below.

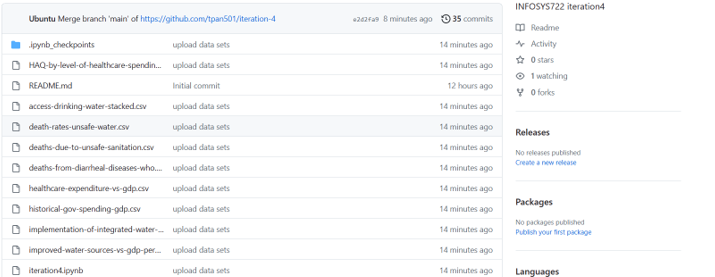


Figure Uploaded data sets in GitHub repository

## Describe the data

At this moment I have got 13 data bases which most of them talk about unsafe water death from different perspectives. All the data sets are in csv format, and the records inside them varies alongside different documents. For most of them, they have recorded around 4000 pieces of data, but on the other hand, it can be viewed that the data sets with less than 300 records.

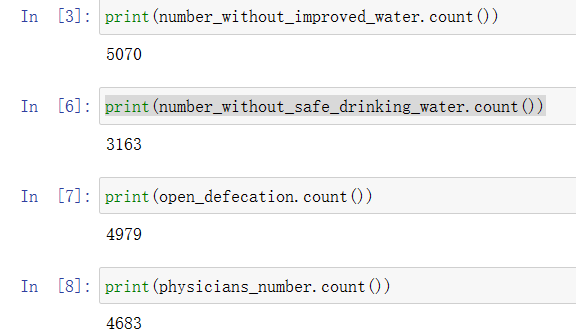




Figure The difference in the volume of data sets

In this section, I will display all the data sets I’ve obtained in my laptop and introduced the basic information about the variables that are involved in those data sets. It is noted that the duplicated fields will be filtered.

**share\_death\_unsafe\_water.csv:** This data set talks about the percentage of people from different countries around world who died from unsafe water from 2000 to 2019.

**Entity:** represents the countries.

**Code:** represents country’s code which is decided by ISO 3166 (International Organization for Standarization, 2023).

**Year:** represents the time spot that the data is about.

**Share of total deaths that are from all causes attributed to unsafe water source, in both sexes aged age-standardized:** represents share of total deaths.

**access-drinking-water-accessed.csv:**

**wat\_sm:** represents “safely managed drinking water”. It means the water source should be available from the premises, at any time if people needed, and it should be containing no pollution (World Health Organization, 2023).

**wat\_bas\_minus\_sm:** represents “basic drinking water source”. Basic water service means the water source that does not meet the “safely managed” criteria but can be achieved within 30 minutes round-trip time (World Health Organization, 2023).

**wat\_lim:** represents “limited water source”. It refers to the water source that requires more than 30 minutes to achieve water (World Health Organization, 2023).

**wat\_unimp:** represents “unimproved water source”. It refers to the water source that fails to implement protection actions from being polluted.

**wat\_sur:** represents “no access or surface water only”. It refers to the phenomenon where people get water only from rivers, lakes and other unprocessed natural resources.

**death-rates-unsafe-water.csv**

**Deaths that are from all causes attributed to unsafe water source per 100,000 people, in both sexes aged age-standardized:** represents the number of death among 100,000 people because of unsafe water.

**deaths-from-diarrheal-diseases-who.csv**

**death\_count - Cause: Diarrhoeal diseases - Sex: Both sexes - Age\_group: ALLAges:** represents the death amount from diarrheal dieases.

**historical-gov-spending-gdp.csv:**

**Government Expenditure (IMF based on Mauro et al. (2015)):** represents the percentage of annual expenditure for a country as a share of GDP.

**improved-water-sources-vs-gdp-per-capita.csv:**

**GDP per capita, PPP (constant 2017 international $):** records the GDP per capita of a country.

**Population (historical estimates):** records the number of population of a country.

**Continent:** records the continent name of the country.

**number-without-improved-water.csv:**

**wat\_imp\_number\_without:** records the number of people who cannot get access to the improved water source.

**number-without-safe-drinking-water.csv**

**wat\_sm\_number\_without:** records the number of people who cannot get access to the safely managed water source.

**people-practicing-open-defecation-of-population.csv:**

**san\_od:** represents the share of people who dispose their excrement in the outside area, such as forests and bodies of water (Ritchie & Roser, 2021).

**physicians-per-1000-people.csv:**

**Physicians (per 1,000 people):** refers to the number of physicians among 1000 people of a region.

**public-healthcare-spending-share-gdp.csv:**

**Domestic general government health expenditure (% of GDP):** represents the share of healthcare spending as a proportion of GDP.

**public-health-expenditure-share-GDP-OWID.csv:**

**public\_health\_expenditure\_pc\_gdp:** represents the amount of expenditure that government spend on every person for healthcare development.

**water-bodies-good-water-quality.csv:**

**6.3.2 - Proportion of bodies of water with good ambient water quality (%) - EN\_H2O\_WBAMBQ:** represents the overall good water quality proportion.

**6.3.2 - Proportion of river water bodies with good ambient water quality (%) - EN\_H2O\_RVAMBQ:** represents the good water quality in river.

**6.3.2 - Proportion of groundwater bodies with good ambient water quality (%) - EN\_H2O\_GRAMBQ:** represents the good water quality in groundwater

**6.3.2 - Proportion of open water bodies with good ambient water quality (%) - EN\_H2O\_OPAMBQ:** represents the good water quality in open water.

**healthcare-expenditure-vs-gdp.csv:**

**Current health expenditure per capita, PPP (current international $):** represents GDP per capita of the country.

**urban-improved-water-access-vs-rural-water-access.csv:**

**wat\_imp\_urban:** represents the percentage of people’s accessibility in urban area to the safely-drinking water

**wat\_imp\_rural:** represents the percentage of people’s accessibility in rural area to the safely-drinking water

According to the description of all the variables, it can be noticed there are some duplicated fields which may have the same functionality, such as “population”, “GDP per capita”. Additionally, the death share and the death rate also show the similar data. Therefore, data analyst needs to figure out which fields should remain in the final data set.

## Explore the data

### Pre-settings

Causing death from the unsafe water should have a range of potential factors affect the situation in a country of region. It can be the percentage of people who can reach safe water sources, the level of how the government focus on this issue and the level the healthcare services in this country or other factors. In this section, I will be visualizing the data distribution based on the data I have collected.

To perform the data visualization, I will use the show() method in Jupyter notebook and display the output first.

The death\_share\_unsafe\_water data set is visualized as below. However, it can be noticed that pyspark did not recognize my column names correctly. Instead, it automatically made up some default column names like “c0”, “c1”. I should add some more parameters to let pyspark to scan my variable names properly.

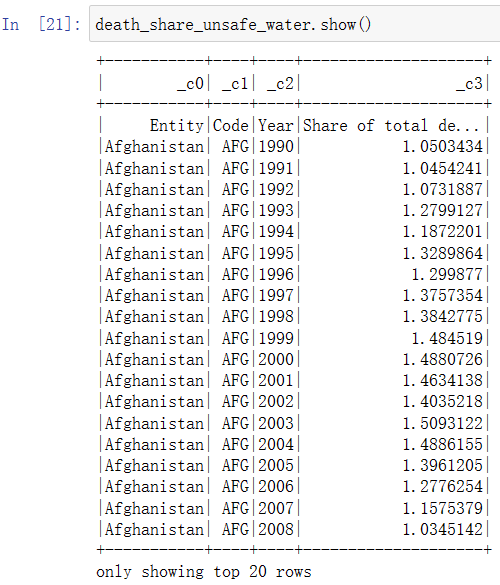


Figure Default display with pyspark

I add “header = true” for pyspark to work properly so it will not be the obstacle for data mining process. Now the show() method displays the results properly.

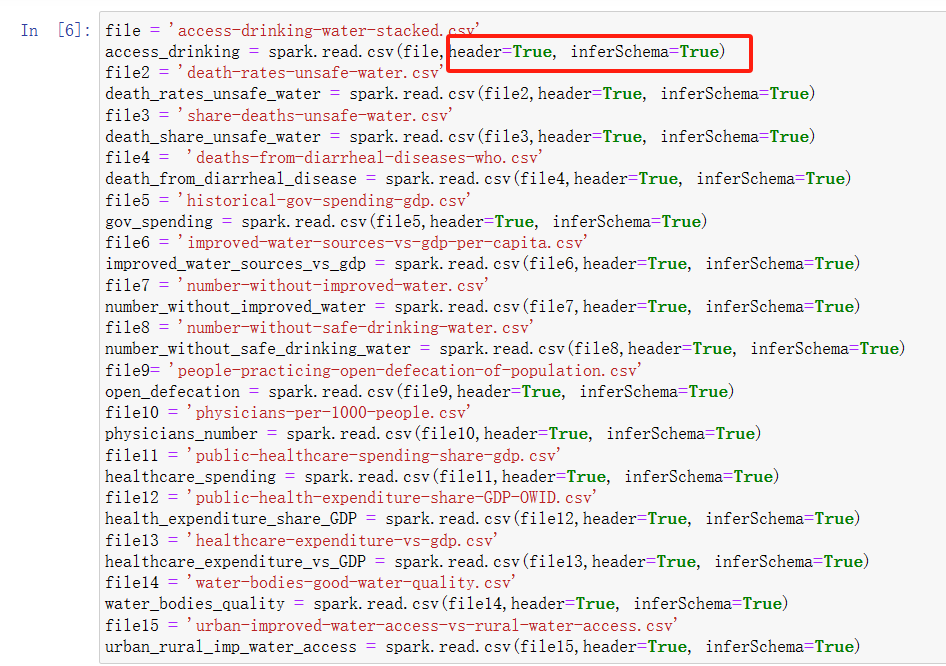


Figure Adjust metrics when importing

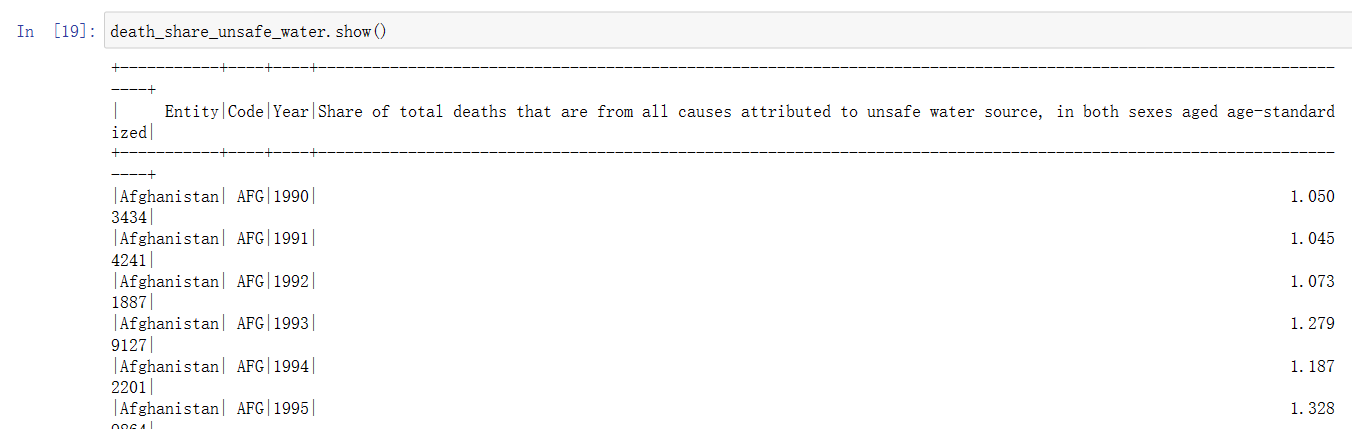


Figure death\_share\_unsafe\_water data set preview

Unfortunately, in Jupyter Notebook, pyspark seems to not have a clearer way to do the data visualization like pandas can do, since pandas support matplotlib and seaborn and other libraries that can support better data visualization. Therefore, I will transfer the pyspark data frames into pandas data frames for this section to achieve the better effects.

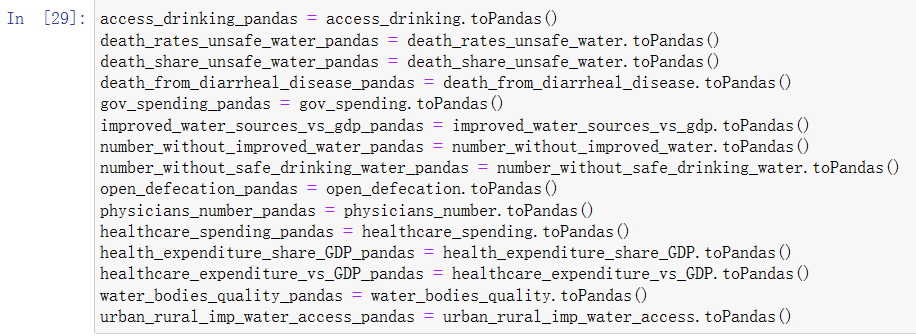


Figure Transform to pandas data frame

Because the data set contains about 200 countries’ data, it is not possible and not practicable to visualize all the countries data in one picture. Therefore, I decided to pick out the 5 countries that have the highest death share， 5 countries with the lowest death share, and 5 countries whose death share is the closest to the median death share according to the most recent year in the data set to create data visualization. For this data mining process, the most recent year will be set as 2019, and countries selected for visualization all come from a typical data set “death\_share\_unsafe\_water”.

It should be noted that many data sets that I’ve provided also includes the data about a particular region, such as African Region (WHO), European Region (WHO) and Region of the Americas (WHO). Those regions will be filtered because I want to pay attention to the country level, and it can avoid hassle when it comes to the GDP or spending related variables.

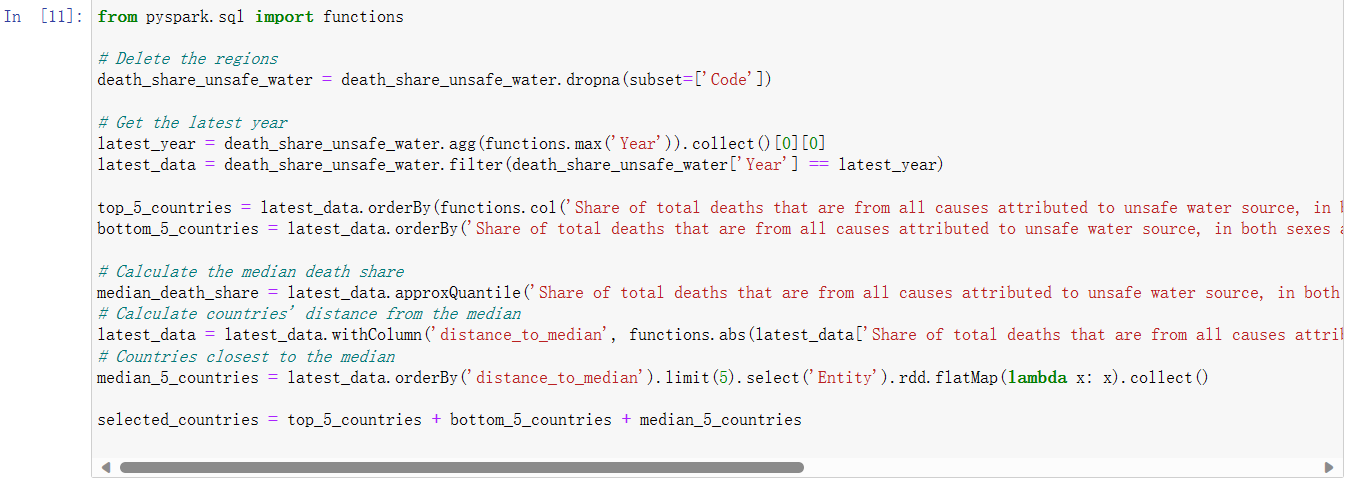


Figure Selecting countries for data visualization

After this, I can print the country list out and show which countries will be used as the example to do the visualization.

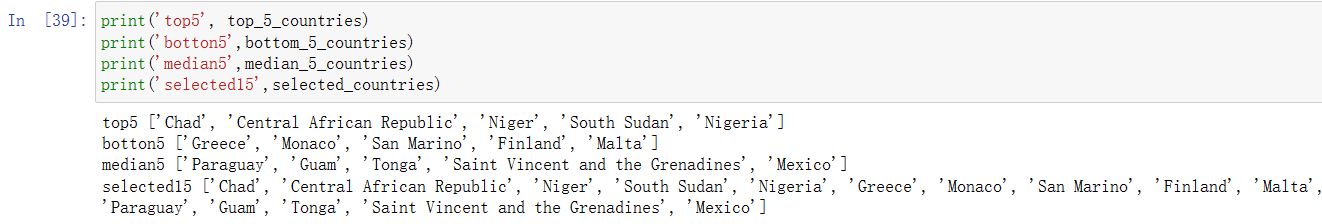


Figure Countries that were seleted

### Data visualization

First of all, the matplotlib library will be imported.



Figure Import matplotlib

Because most data sets I have collected contains “Entity (countries)” attribute and “Year” attribute, it is better for me to do visualization in a form of line chart. In order to draw the graph with both year, countries and the typical attributes that we are focusing on, I added 15 countries into an array and iterate them though the loop, so that all the 15 countries’ data can be shown on one line graph. The x-axis will be the time, and y-axis will be the typical (or target) data.

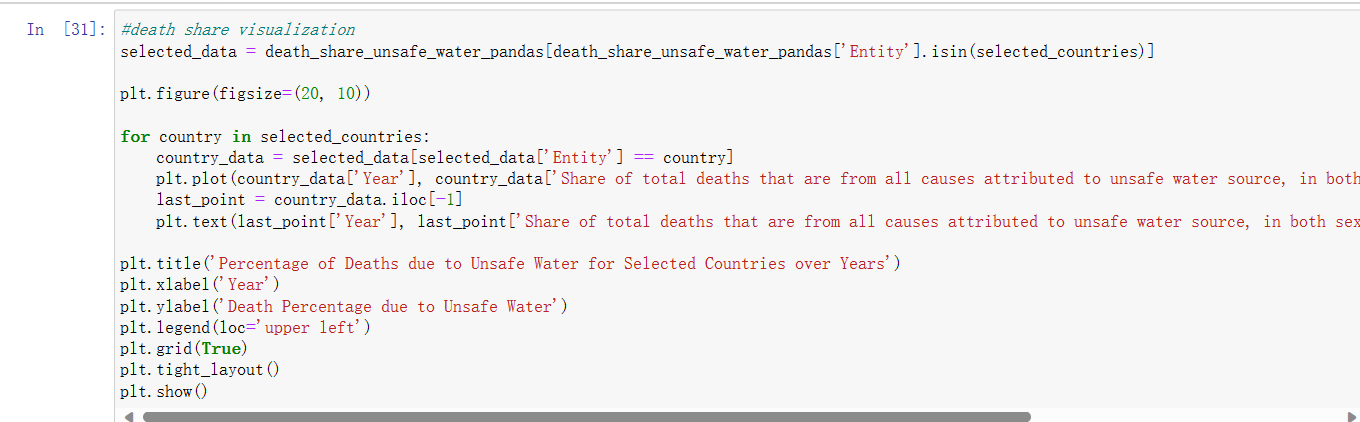


Figure Data visualization codes for death\_share

The distribution of death share of unsafe water is showed below. It can be viewed that the distance between and top 5 and bottom 5 countries is very huge. Even the data of countries with the median death share looks relatively low, compared with the top 5 countries.

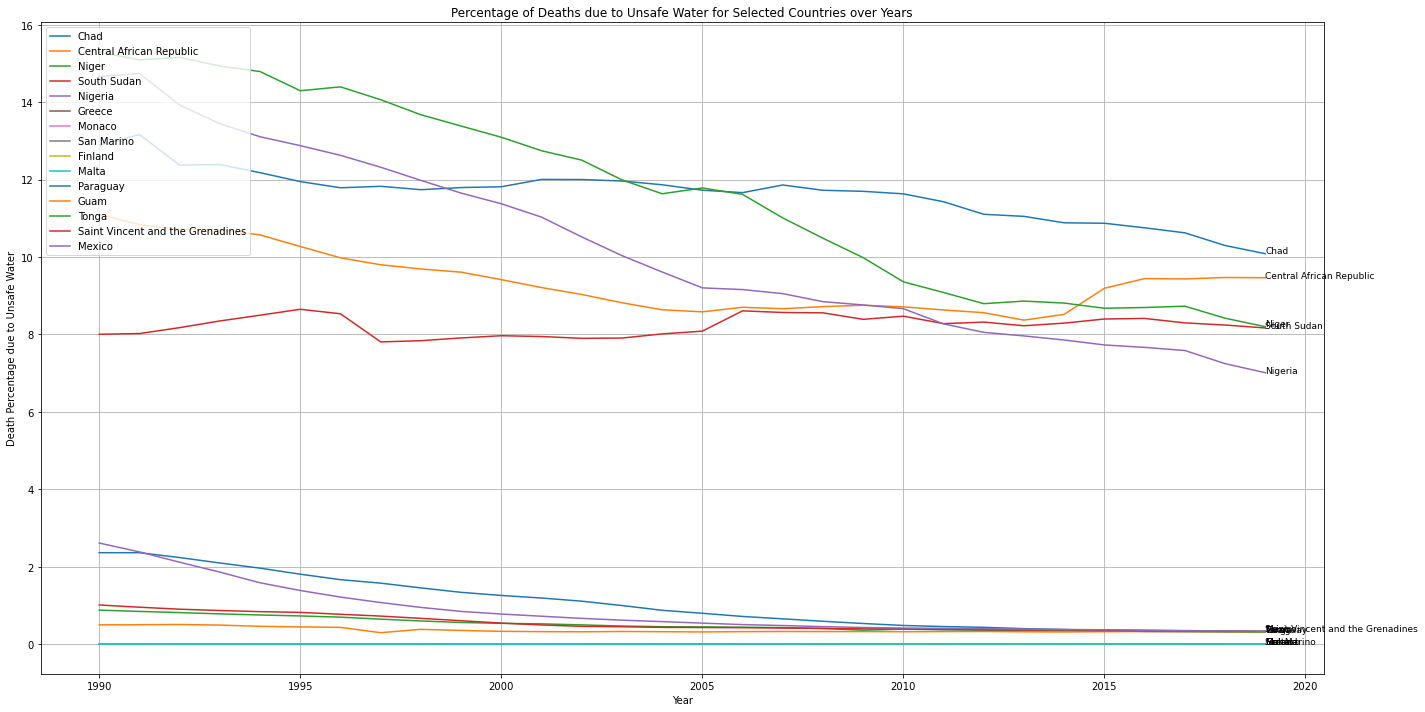


Figure Percentage of death to unsafe water for selected countries over years

The rate of death of unsafe water distribution is showed below. It does not have an obvious difference from the percentage graph, but only the ranking of the top 5 countries have shifted a little bit.

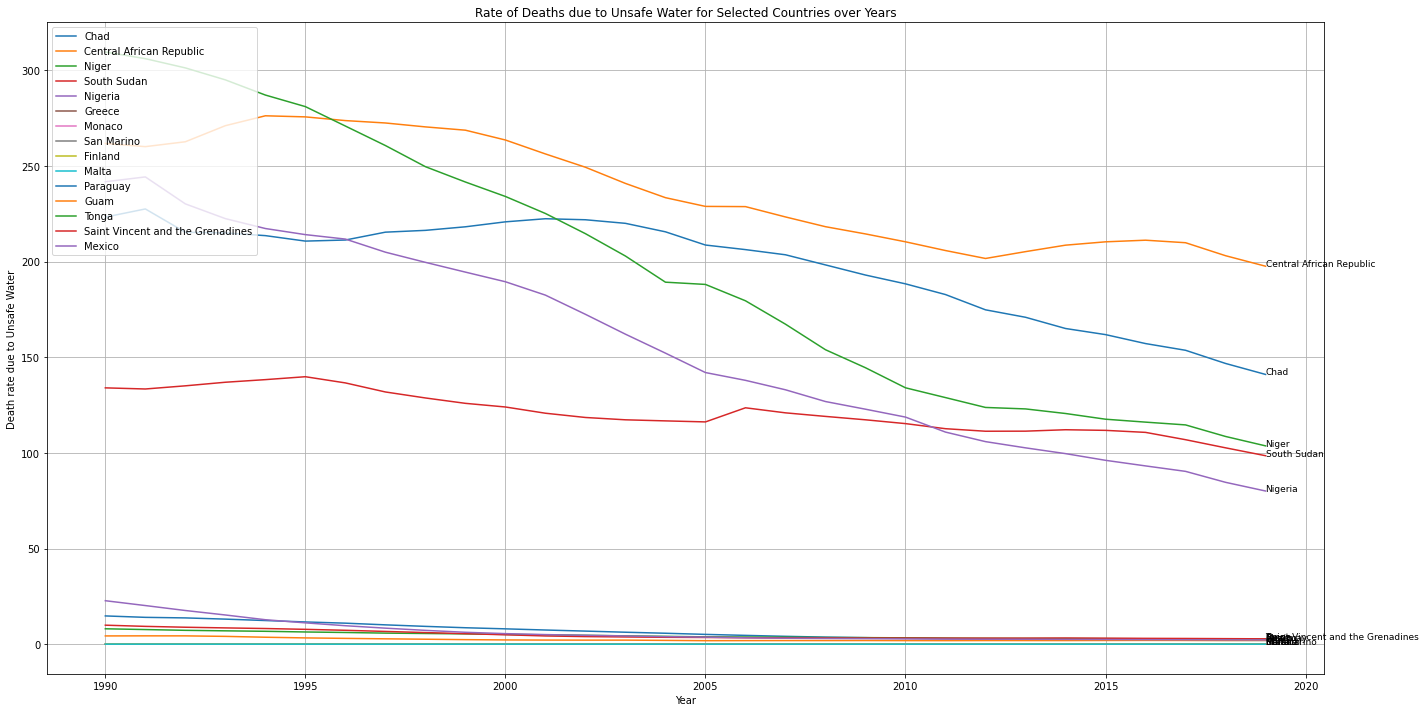


Figure Rate of death to unsafe water for selected countries over years

When I was trying visualize the percentage of people’s accessibility to different water types, an error occurred showing there is NaN value in the data set.

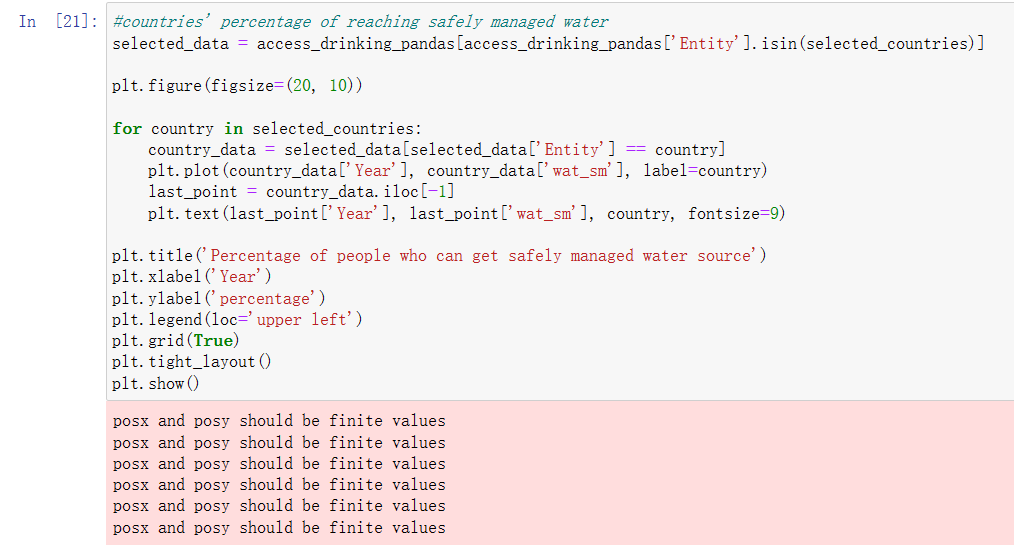


Figure Errors when there are NaN values

Therefore, I need to figure out whether the data cell is a null value or not before doing the visualization.

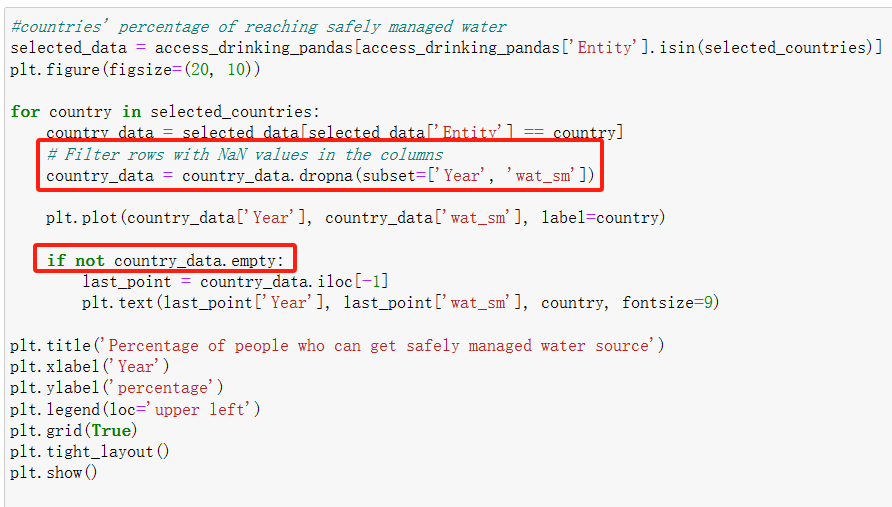


Figure Remove the NaN value rows

Now I can display the percentage of people who can get safely managed, basic, limited, unimproved, surface water source from different countries respectively.

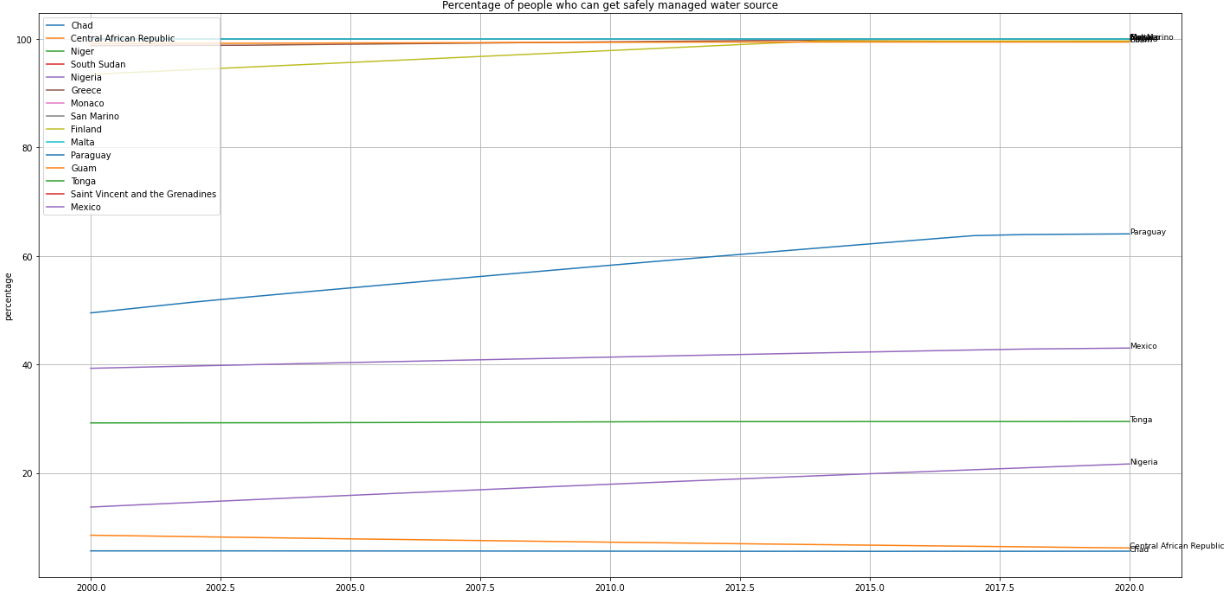


Figure Percentage of people who can get safely managed water source

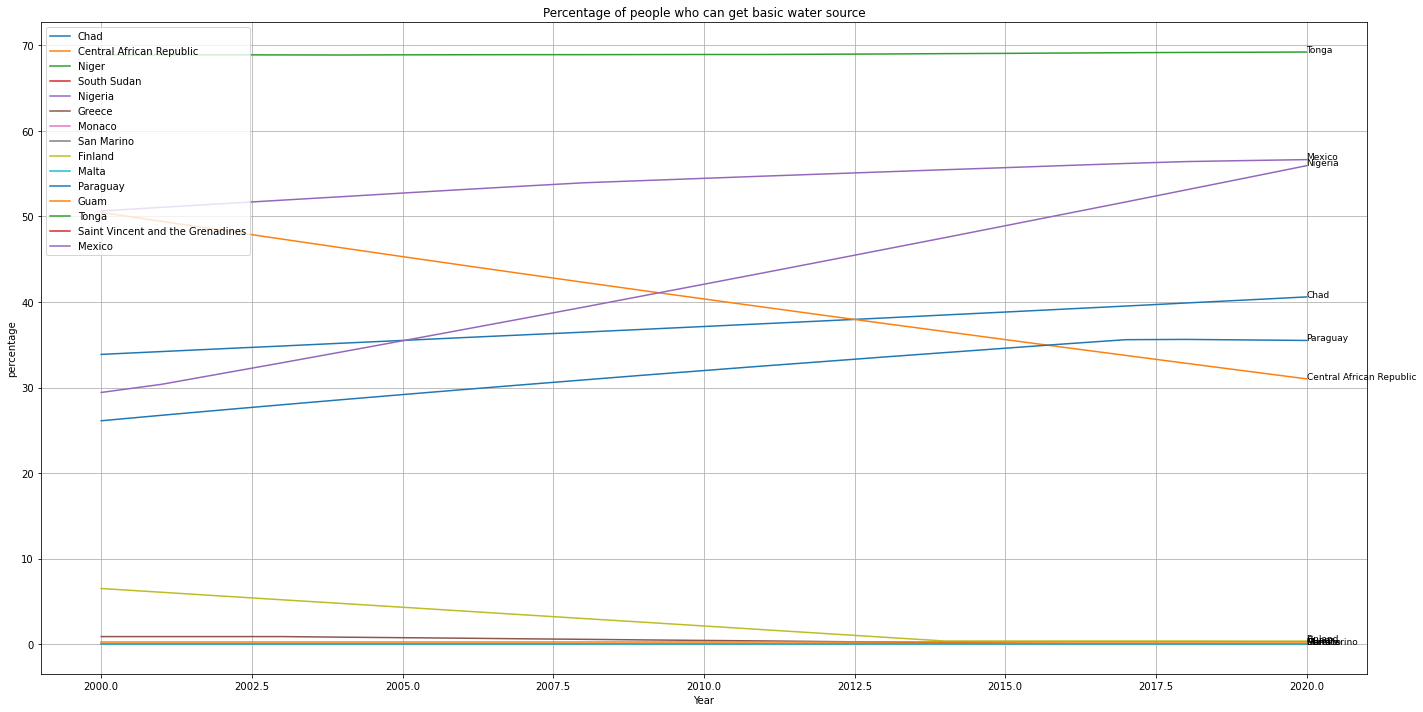


Figure Percentage of people who can get basic water source

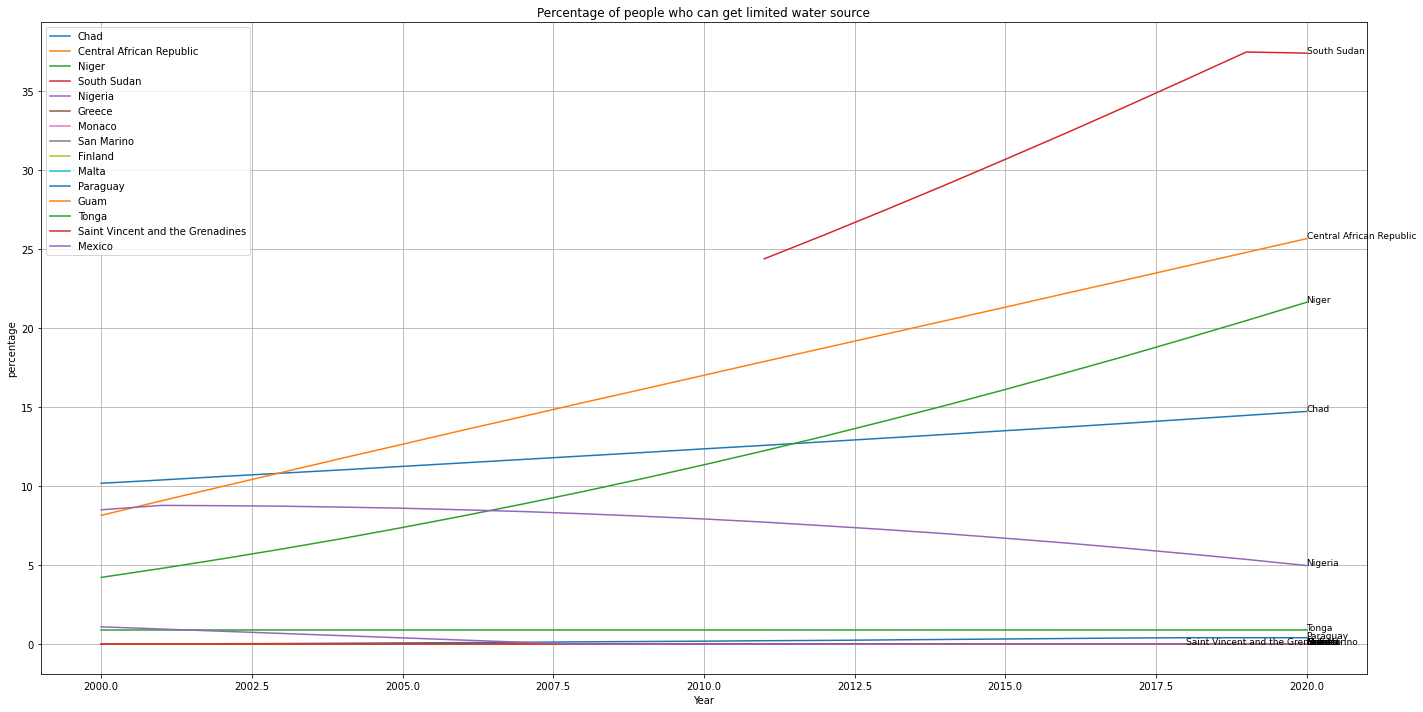


Figure Percentage of people who can get limited water source

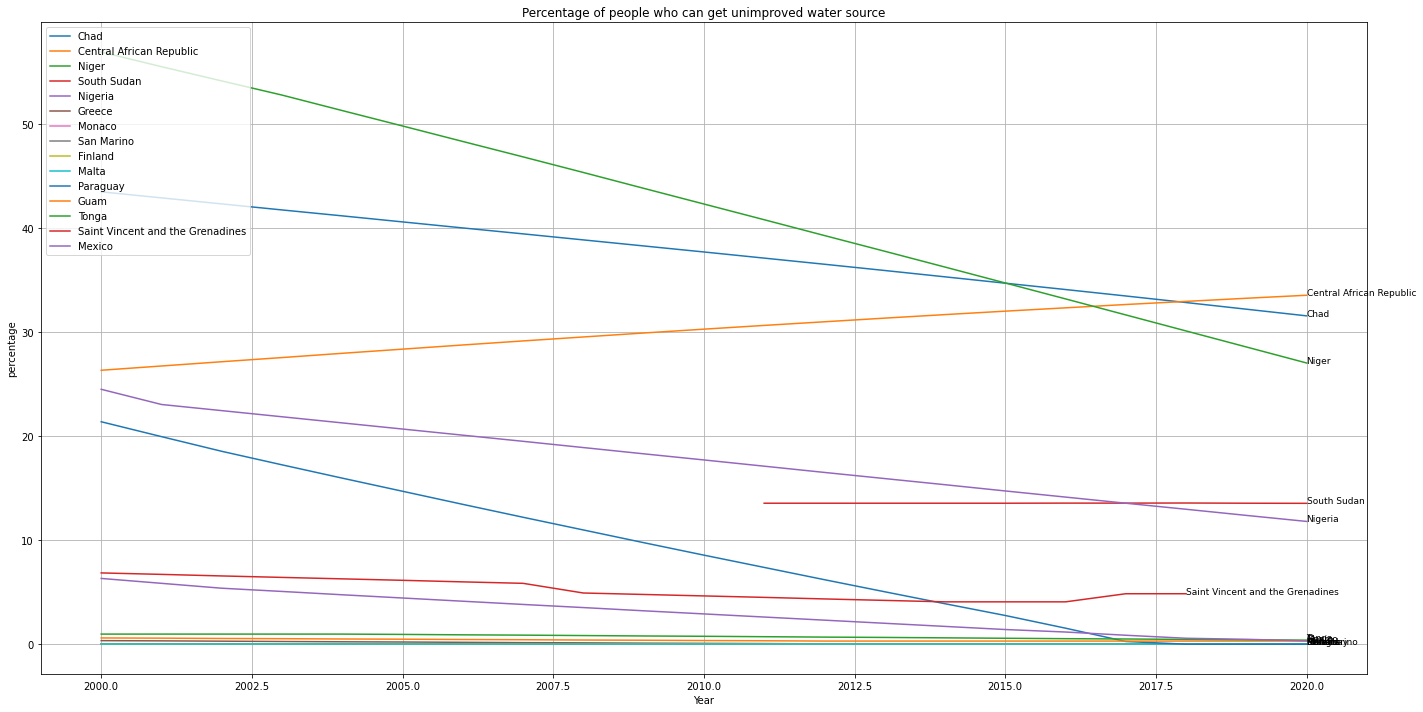


Figure Percentage of people who can get unimproved water source

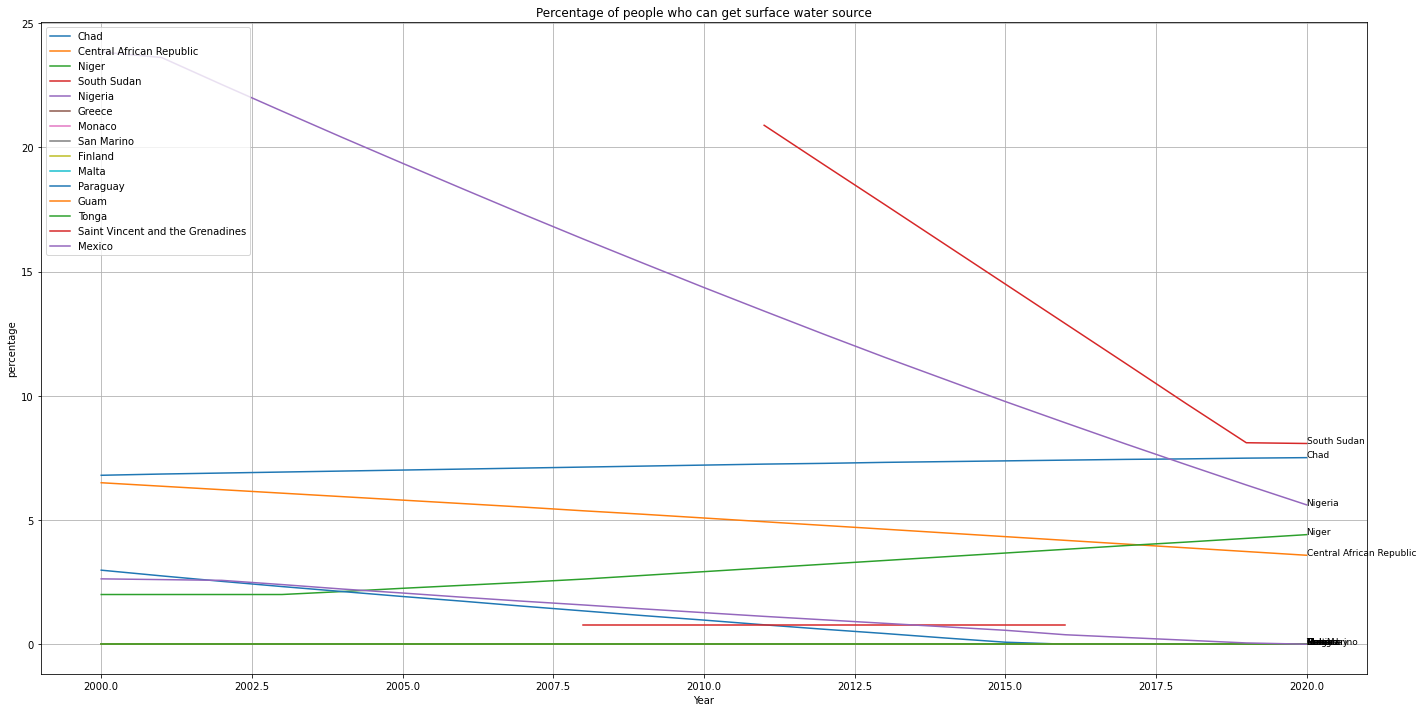


Figure Percentage of people who can get surface water source

The distribution of percentage of death from diarrheal diseases in each country is showed below. It should be informed that Mocano and San Marino’s data from the bottom 5 countries and Guam’s data from the median countries is not available in the data set, which caused out-of-bound error. Therefore, if statement is set in the loop to check if the selected country exists in the current data set.



Figure Use if statement to skip the unavailable records

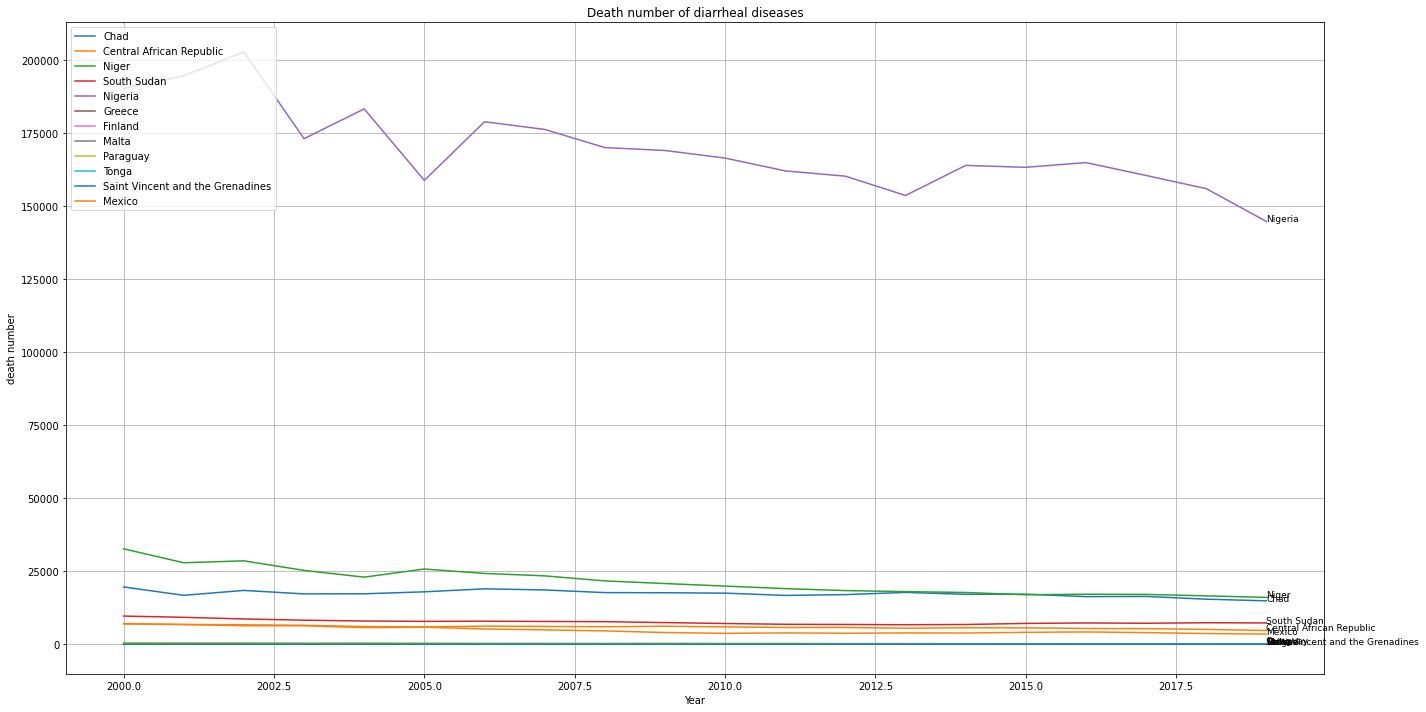


Figure Death number of diarrheal diseases

The historical expenditure percentage relative to GDP of each country is shown below. It can be viewed that this data set’s latest data is the data from 2011.

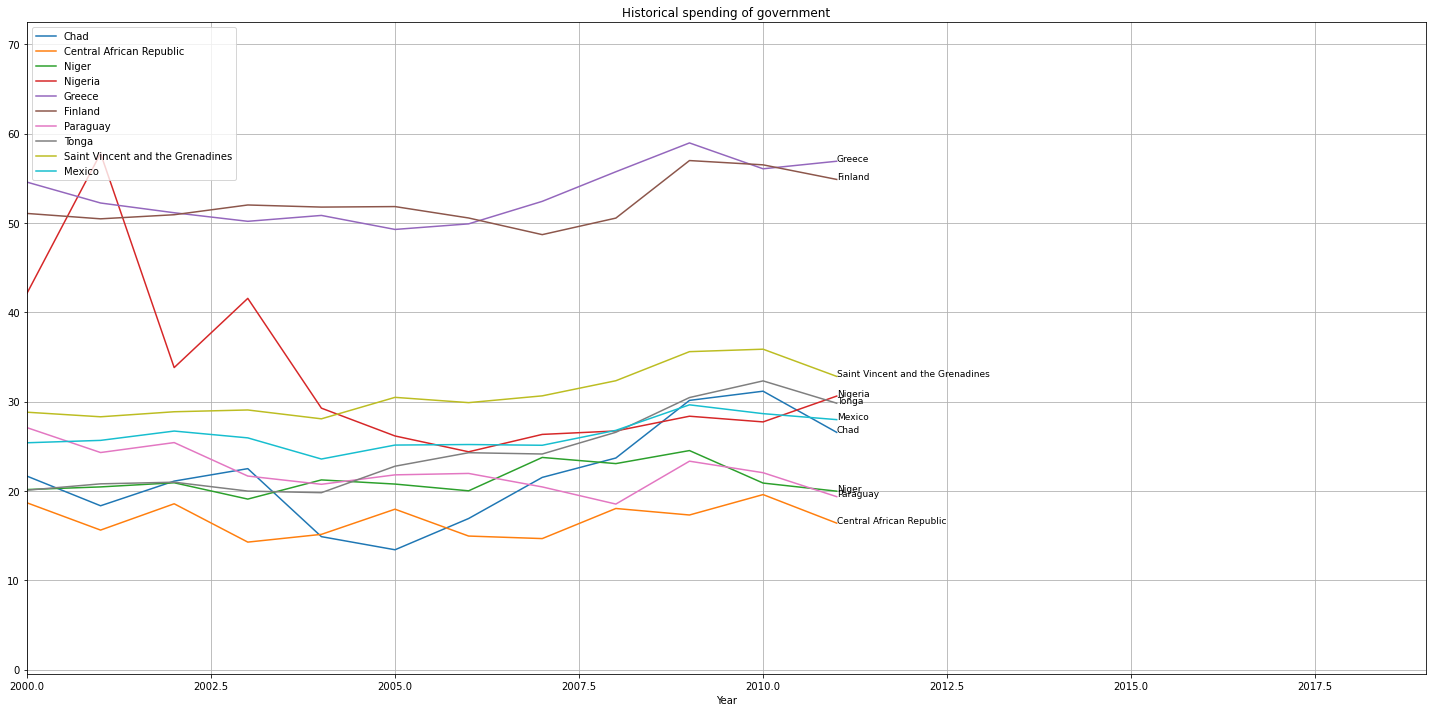


Figure Historical spending of government

The GDP per capita of each country is shown below.

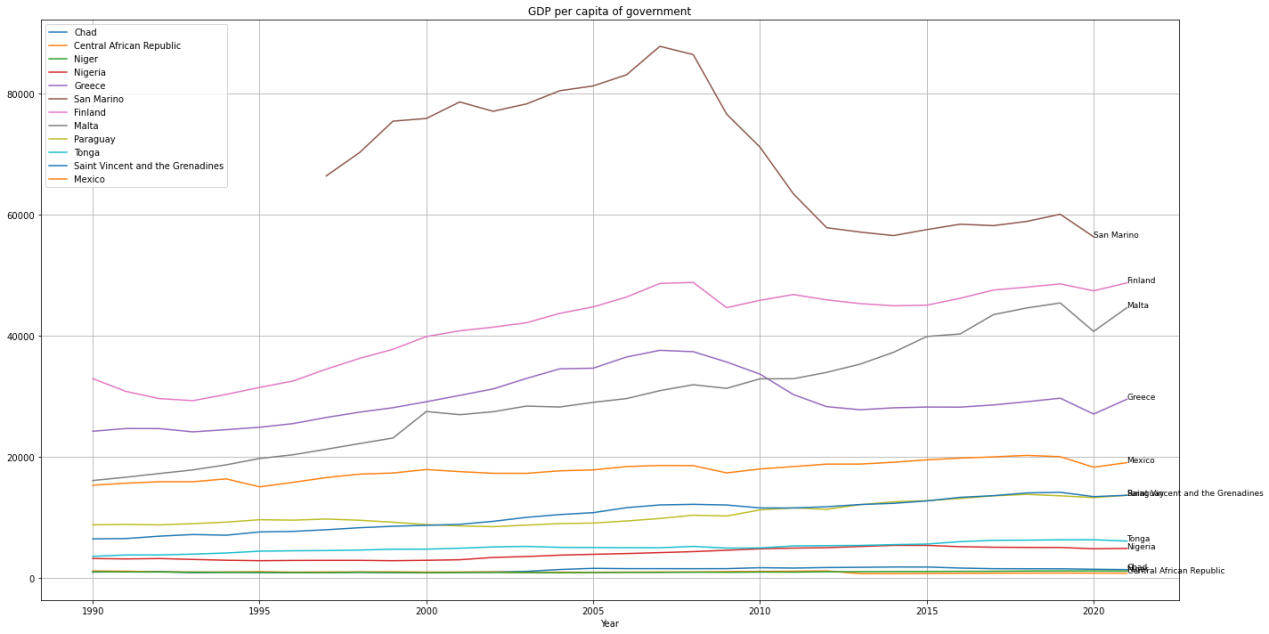


Figure GDP per capita of government

The population of each country in different years is shown below. Please note that matplotlib may automatically use the scientific notation, so the figure written at the y-axis should be multiple by 8.

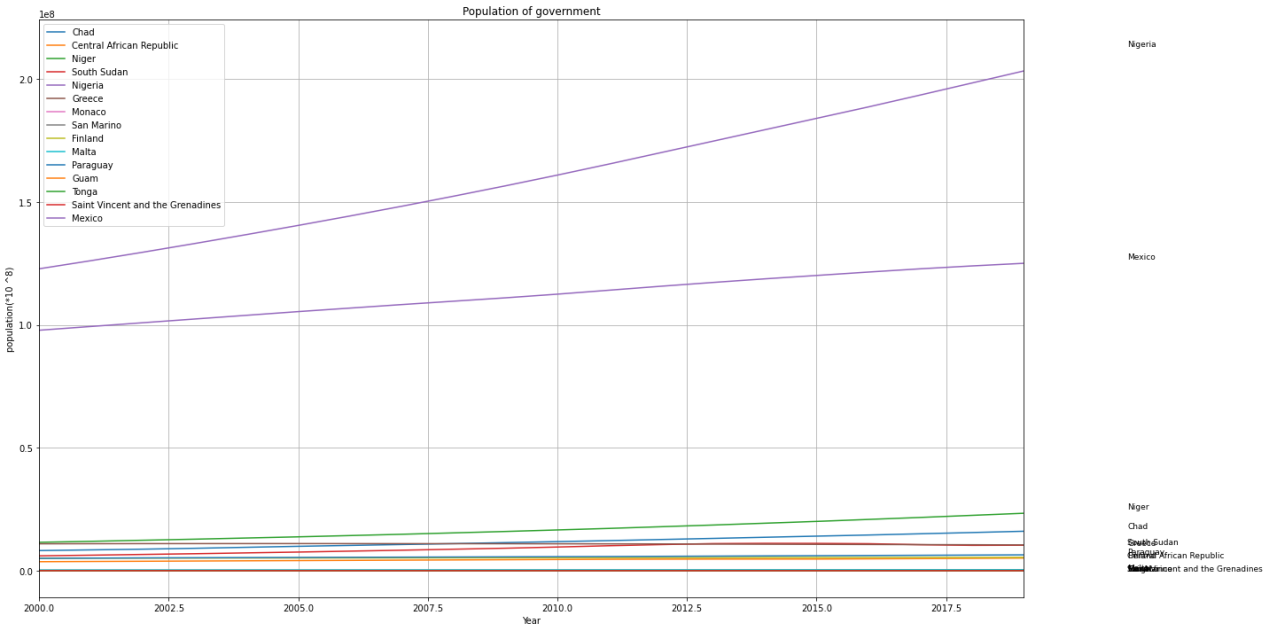


Figure Population of country

The number of people who are unable to get improved water source is shown below.

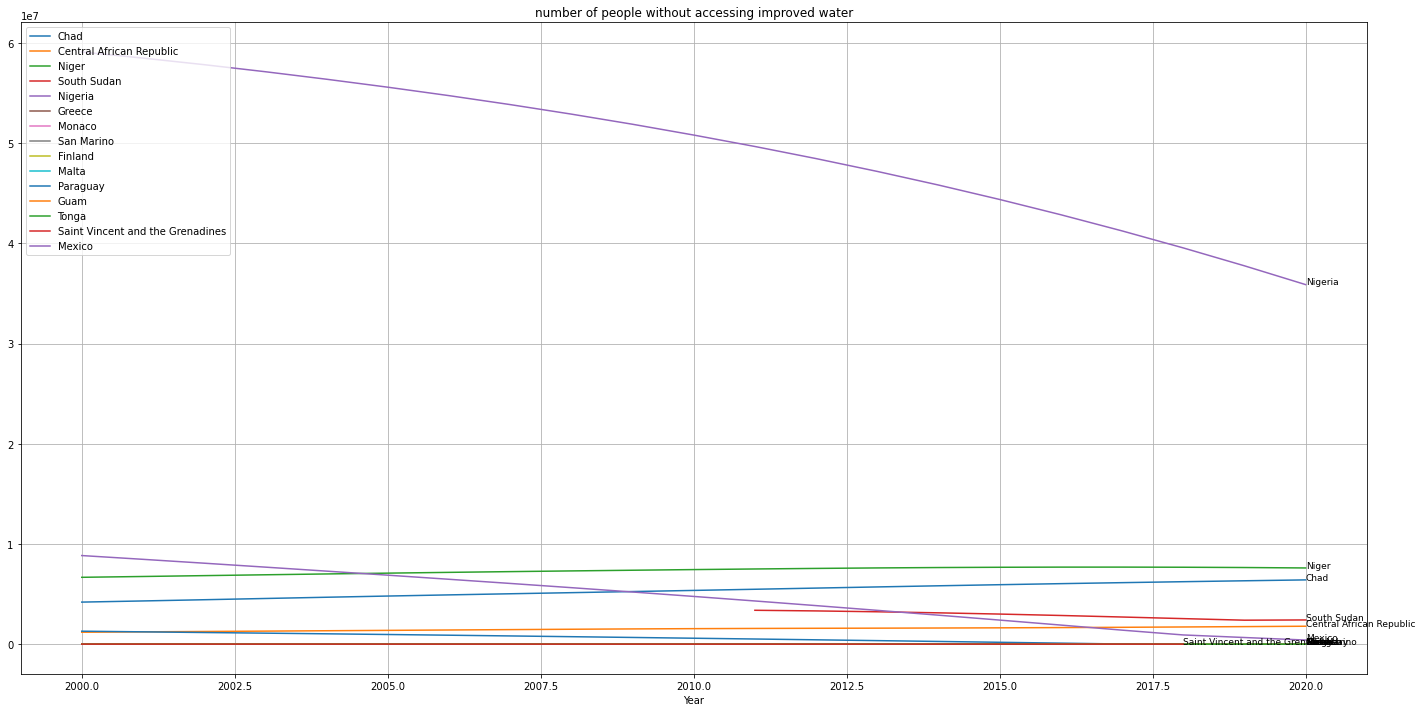


Figure Number of people without accessing improved water

The number of people who cannot get safe drinking water from each country is shown below.

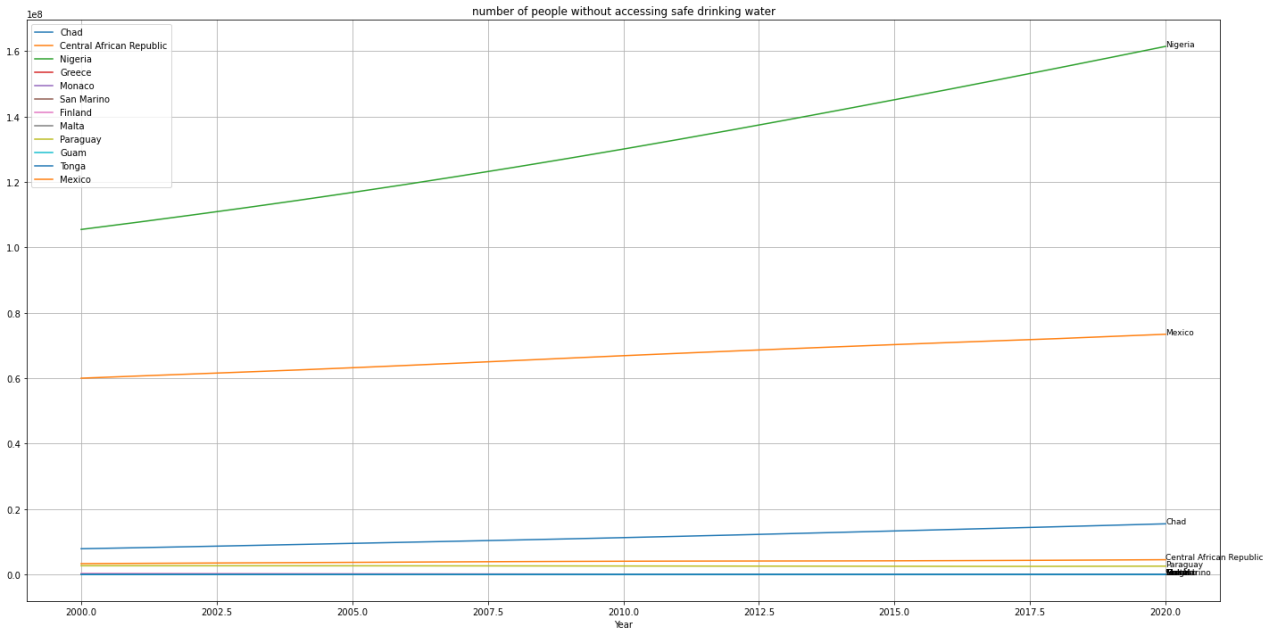


Figure Number of people without accessing safe drinking water

The percentage of people doing open-defecation in each country in shown below. We can see that the trend of the proportion is going lower every year but there are countries still having a relatively high percentage in open-defecation in the public.

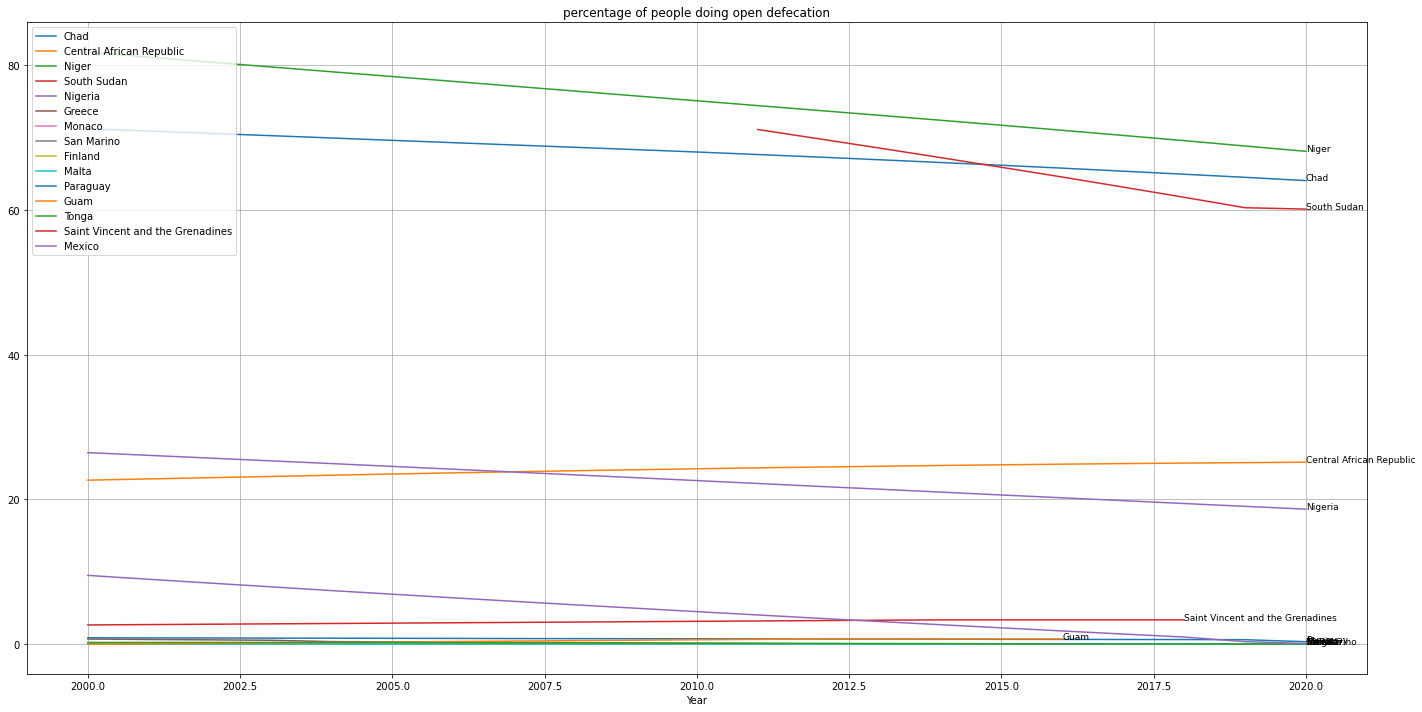


Figure Percentage of people doing open defecation

This number of physicians every 1000 people of each country is shown below. It is noticed that the data in this set has missing value as well. The general number of doctors is rising every year, which represents the effort of the UN and each country these years.

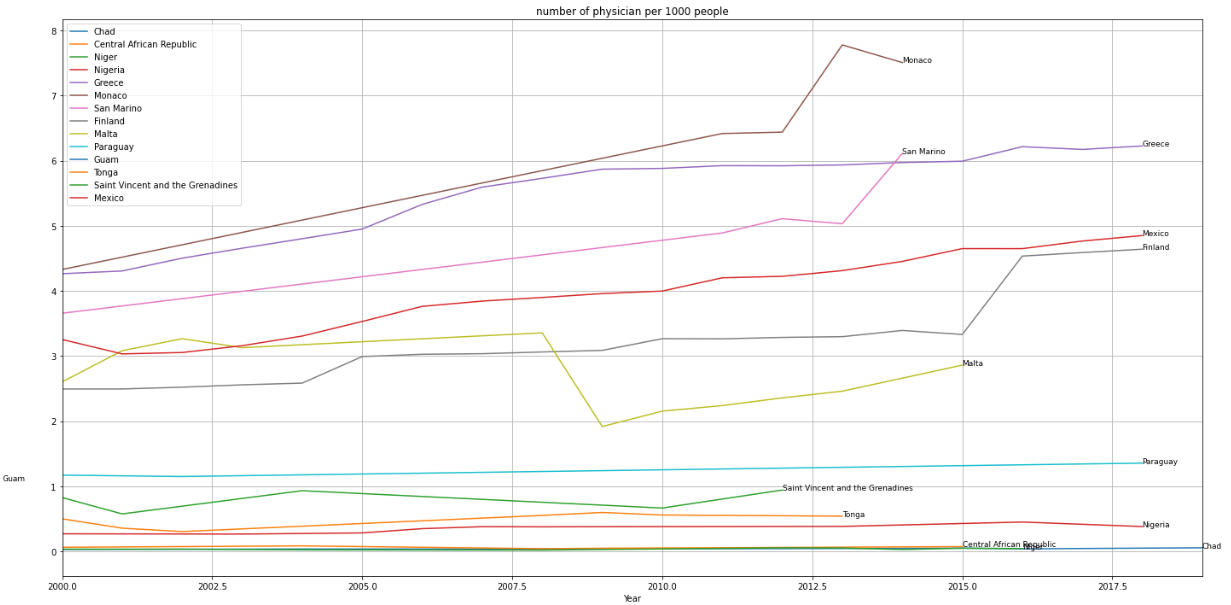


Figure Number of physicians per 1000 people

The share of healthcare spending of each government as a percentage of GDP is shown below.

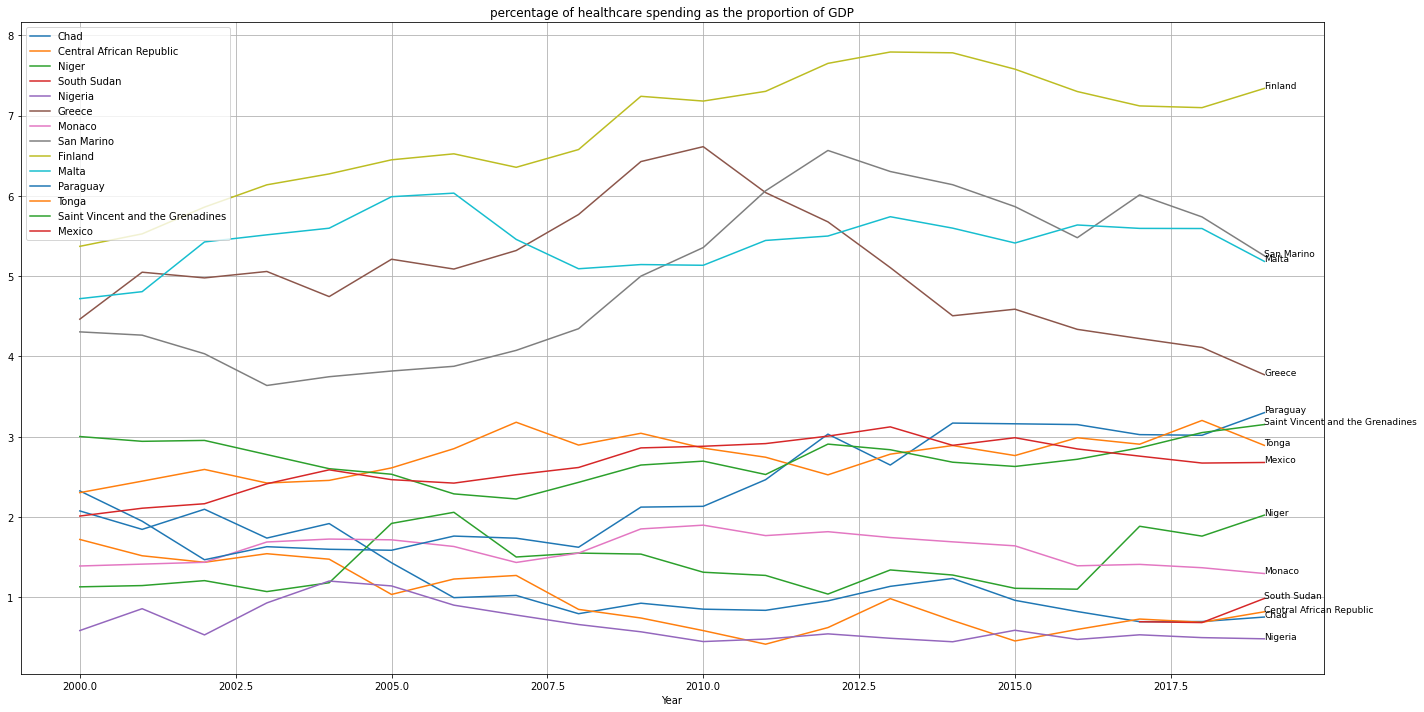


Figure Percentage of healthcare spending in the proportion of GDP

The distribution of healthcare spending per capita is shown below.

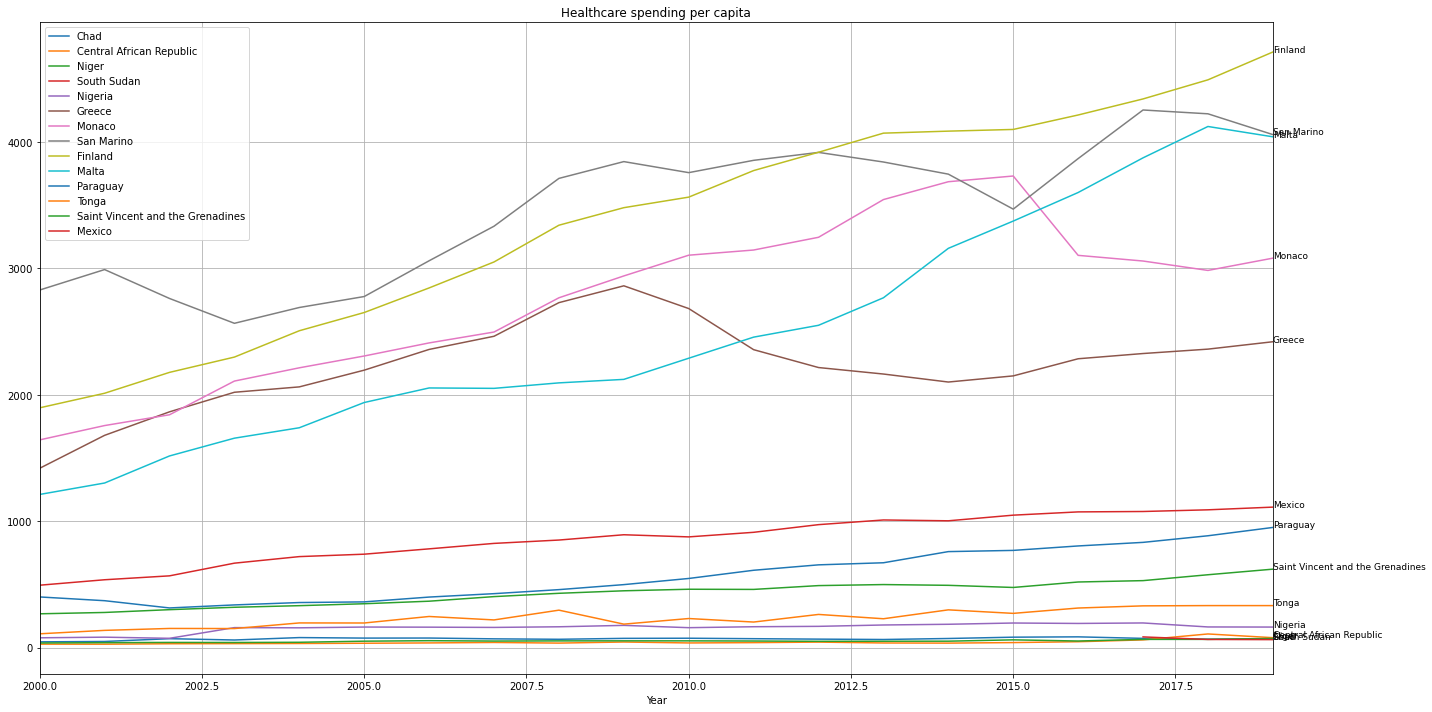


Figure Healthcare spending per capita

The percentage which represents the overall water quality is shown below. It is noticed that this data set only has the data for 2017 and 2020. In order to make it convenient for the visualization, we will use 2020’s data. This data set does not have all the countries’ data that I have set before. Among all the 15 countries I’ve set for data visualization, this data set only has data for 6 of them in 2020.

It needs to be mentioned that seaborn library was used to show the bar charts.

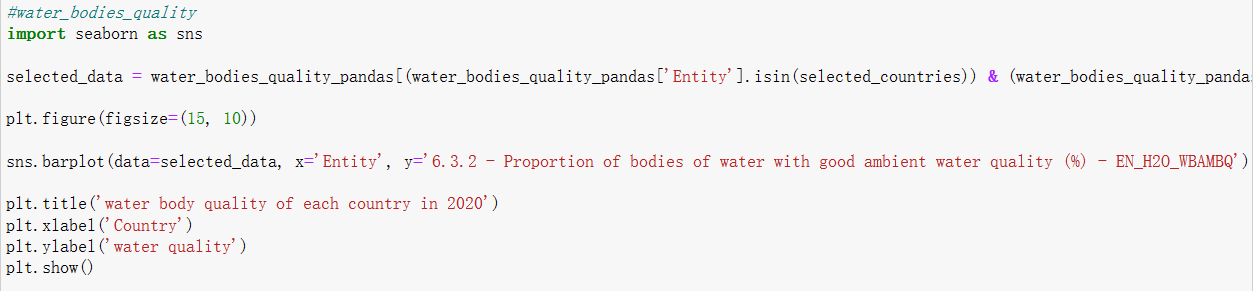


Figure Use seaborn for visualization

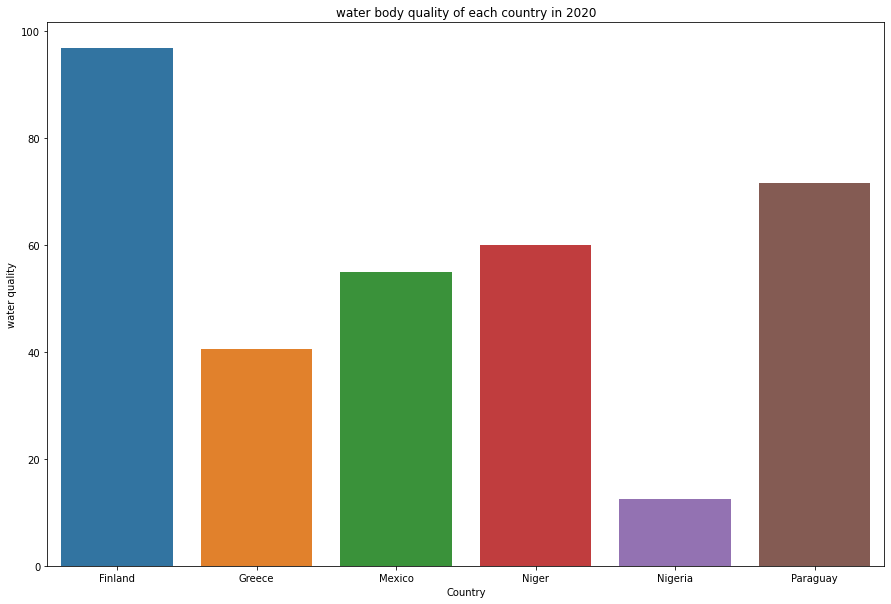


Figure Water body quality of each country in 2020

The water quality (river) of the countries is showed below:

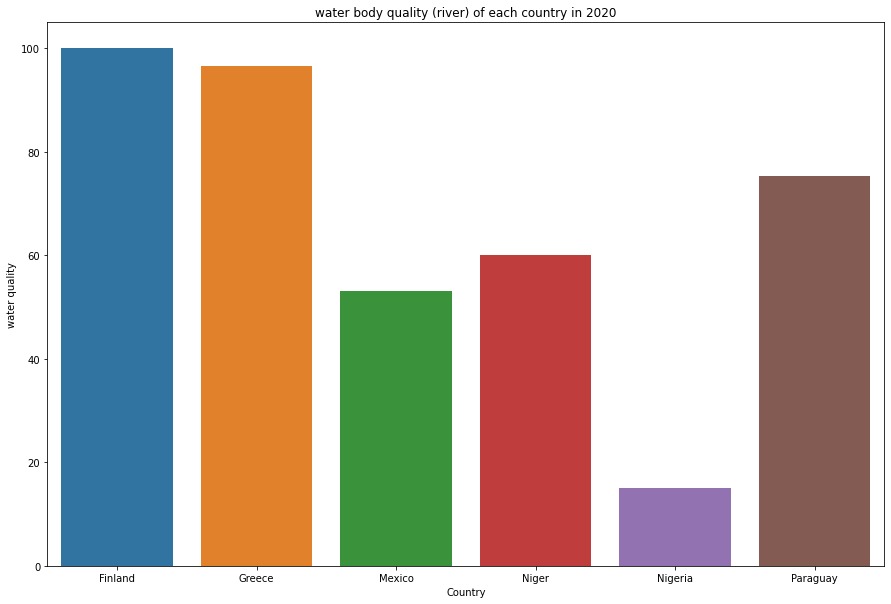


Figure Water quality (river) of each country in 2020

The proportion of water quality (ground water) is showed below. It can be seen that only one data set out of 15 countries has data for this attribute.

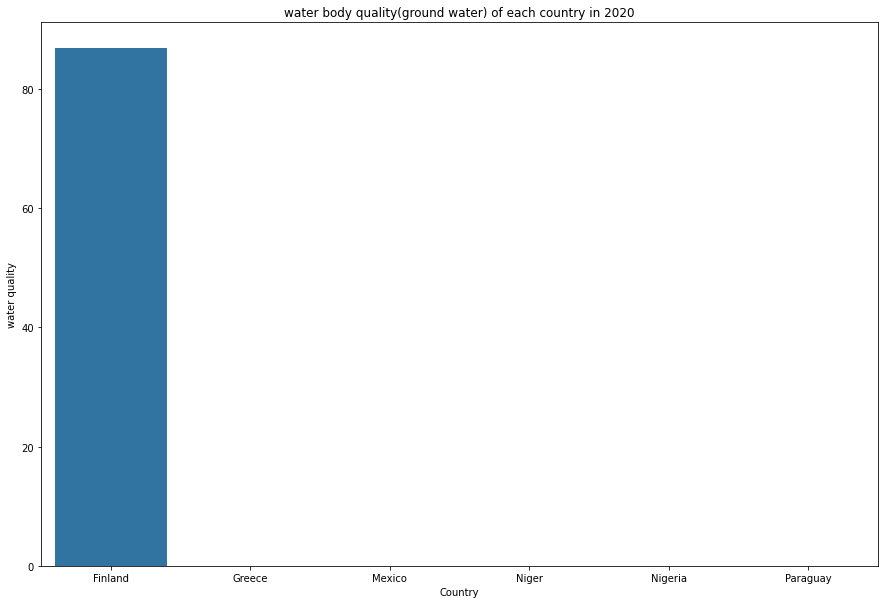


Figure Water body quality (ground water) of each country in 2020

The water quality of surface water is showed below.

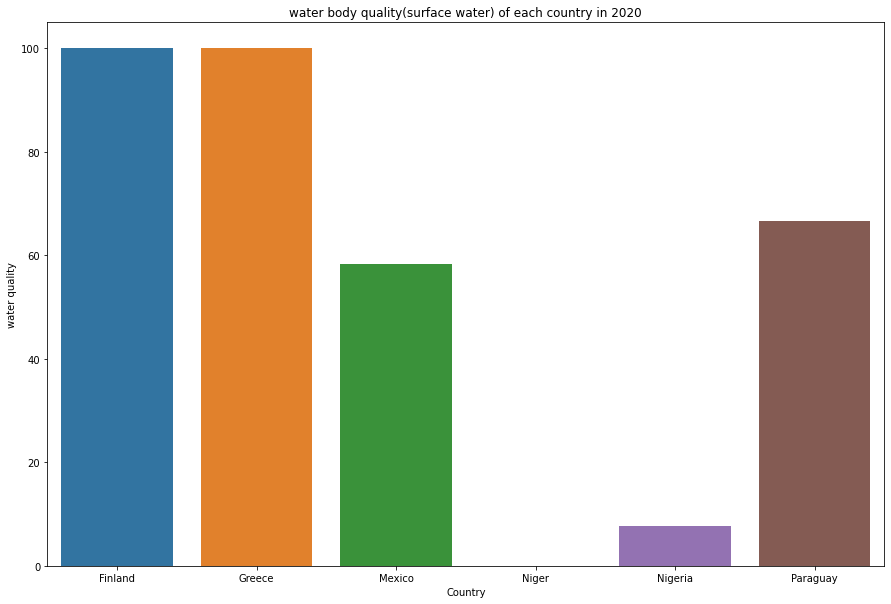


Figure Water body quality (surface water) of each country in 2020

The people from urban area accessibility to safely-drinking water is shown below.

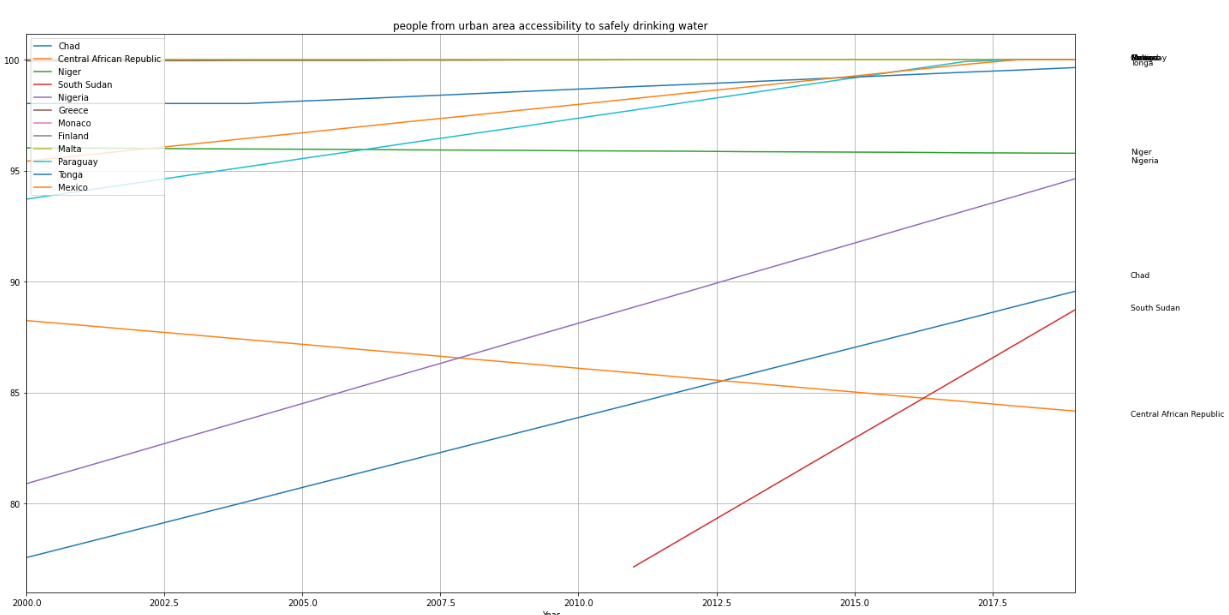


Figure Percentage of people from urban area accessibility to safe drinking water

The people from rural area accessibility to safely-drinking water is shown below.

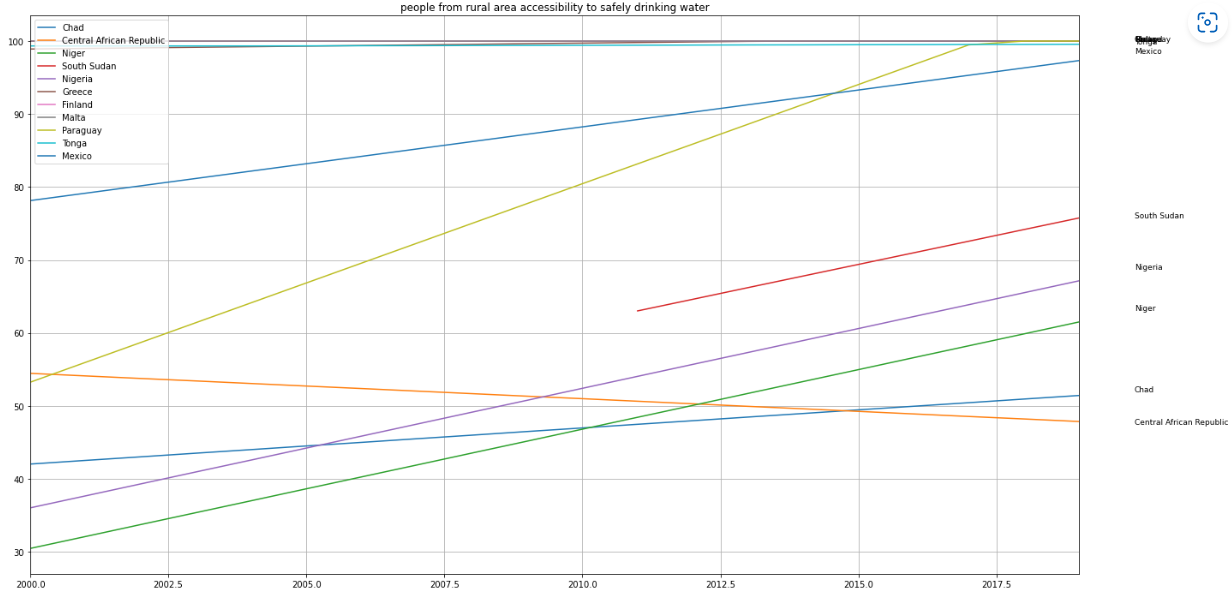
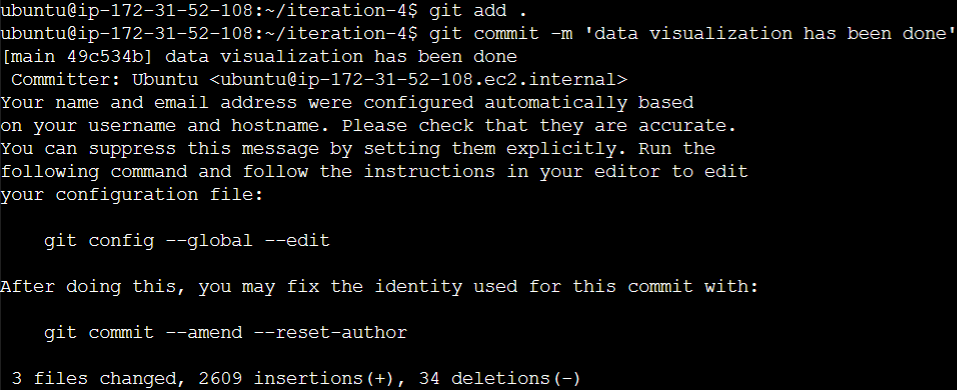


Figure Percentage of people from rural area accessibility to safe drinking water

After exploring all the data set so far, it can be concluded that the overall data set has got plenty of variables that I can probe into, but on the other hand, there are also many fields that contains a lot of missing values for all the countries around the world and all the time spot between 2000 to 2019. Furthermore, some of the variables share the same functionality, such as the percentage data and the numeric or rate-related data. Therefore, I need to pay attention to select the data carefully in the subsequent stages.

After this section, the current process will be committed and pushed onto my GitHub repository.



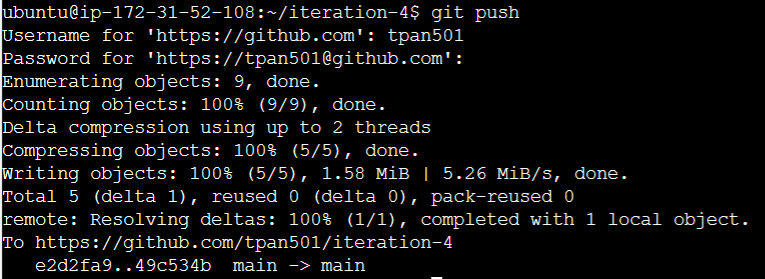


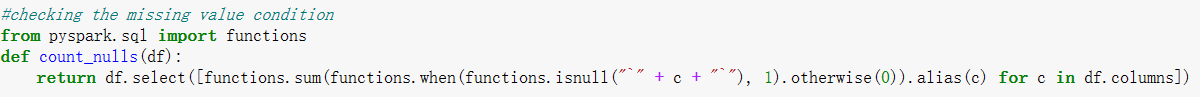


Figure Commit and push changes to GitHub repository

## Verify data quality

This section will be mainly checking the data’s quality, whether it has any missing values, extremes and outliers.

Missing values. With pyspark library, the accumulation of missing values will be mainly done by the following function. I will loop through all the columns in the data set, and check if the current cell is null value or not. If the current cell is null, it will return 1, otherwise 0. I will sum the count up and show it in a new data frame with the same column names.



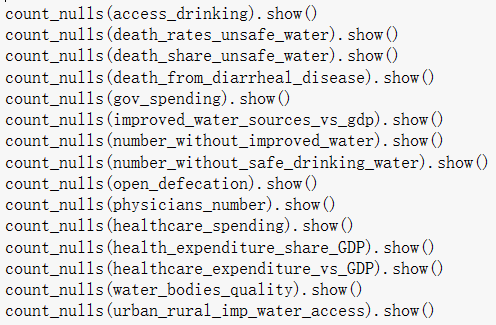
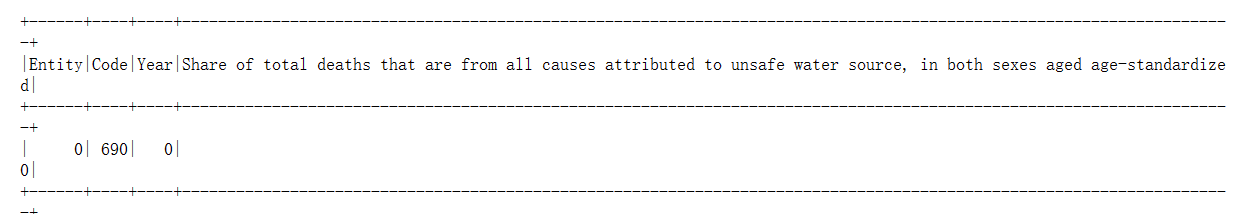


Figure Define a function for showing null values count for each variable

The data for death share and rate of unsafe water’s missing value condition is fine. Both of the data set’s have almost perfect completion. There are some missing value for the death\_rate\_unsafe\_water in the “Code” variables, but this is not relevant to the people’s death so it can be neglected.



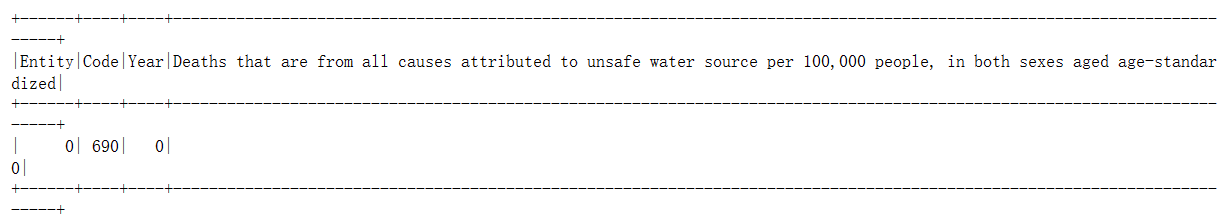


Figure Missing values of the data sets

The data for people achieving different safe level of water sources do has over 1000 missing values for “wat\_bas\_minus\_sm”, “wat\_sm”, which is about 20% of the total records. “wat\_sur” also has 92 missing values, which is relatively fewer.

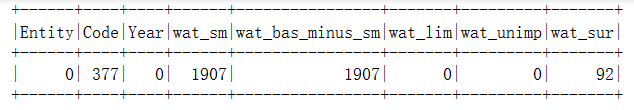


Figure Missing values of the data sets

deaths-from-diarrheal-diseases-who.csv is a highly-completed data set, which only has 120 missing values for the “Code”.

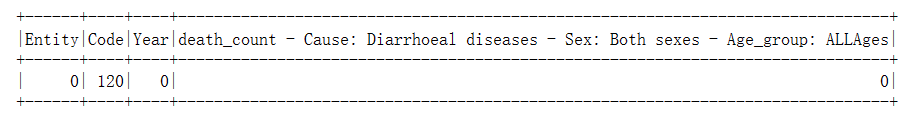


Figure Missing values of the data sets

Government historical spending data set also have completed all the cells. However, as mentioned in the section 2.3, its data has not been updated since 2011.

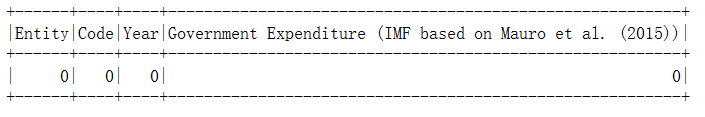


Figure Missing values of the data sets

Differing from others, “improved\_water\_sources\_vs\_gdp” has plenty of missing values. This is mainly because the time span in the data set is very long, and the data of most countries started from 18 centuries, where there might not be the concept of GDP. Furthermore, there are also some countries’ “Year” attributes are “-10000”, which could be the outliers.

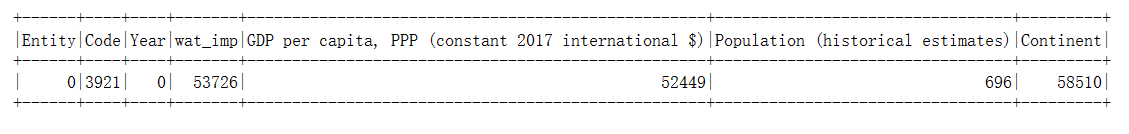


Figure Missing values of the data sets

number-without-improved-water.csv and number-without-safe-drinking-water.csv has high completion.

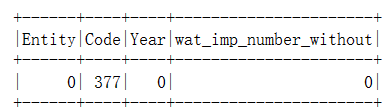
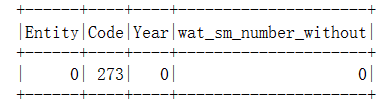
 

Figure Missing values of the data sets

people-practicing-open-defecation-of-population.csv has high completion.

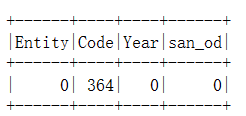


Figure Missing values of the data sets

physicians-per-1000-people.csv has high completion.

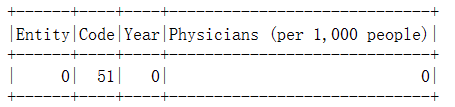


Figure Missing values of the data sets

public-healthcare-spending-share-gdp.csv has high completion.

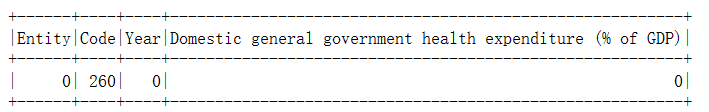


Figure Missing values of the data sets

public-health-expenditure-share-GDP-OWID.csv’s completion is perfect.

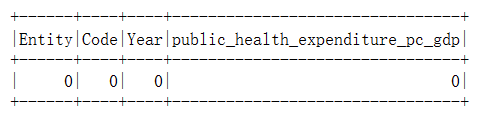


Figure Missing values of the data sets

There are a lot of missing values for the “healthcare-expenditure-vs-GDP” in Code, Population, health expenditure per capita, and other variables. The main reason is still the special way of recording data. Like “improved\_water\_sources\_vs\_gdp.isnull().sum()”, it recorded data from “-10000” to 2021, so there are lot of data is unavailable.

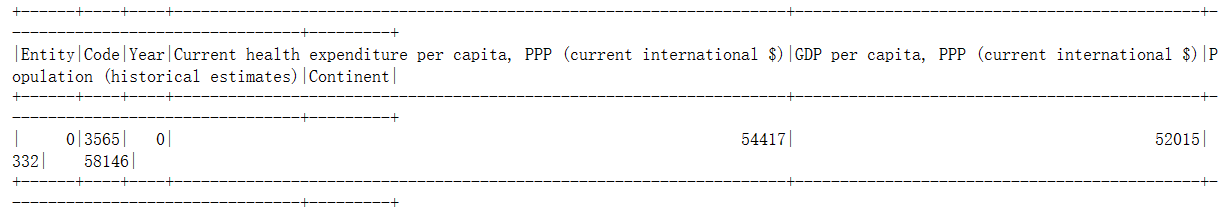


Figure Missing values of the data sets

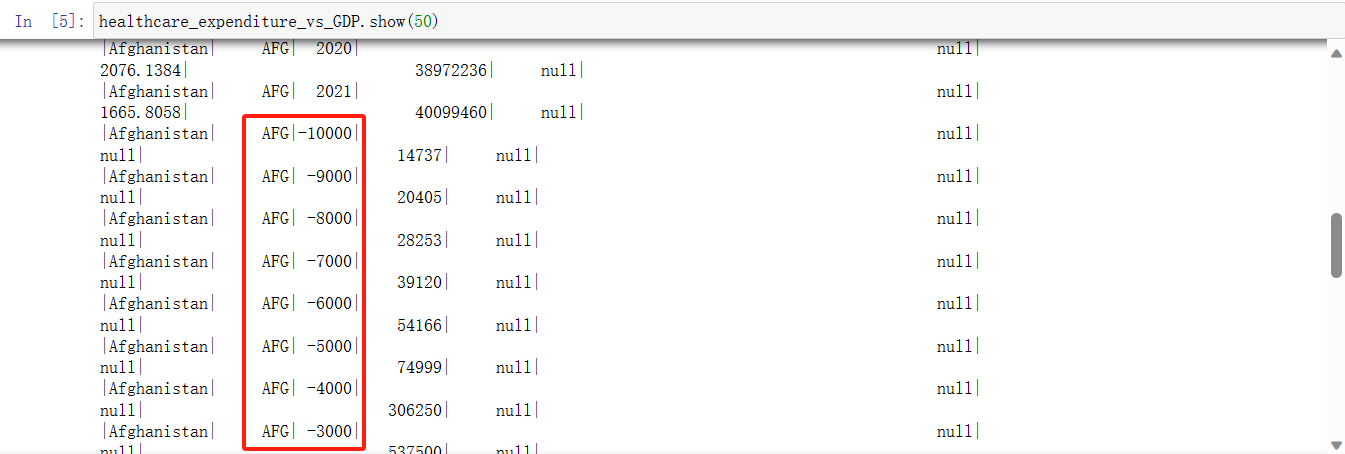


Figure Huge volume of missing values because of NaN

“water-bodies-good-water-quality.csv” has a fine completion.

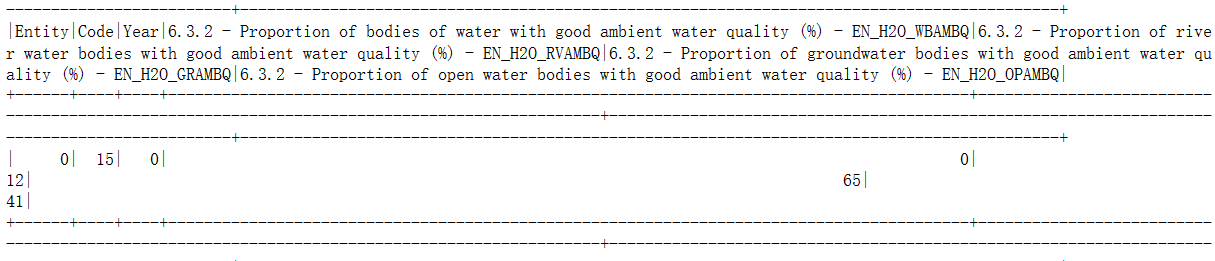


Figure Missing values of the data sets

“urban-improved-water-access-vs-rural-water-access.csv” contains plenty of null values. After viewing the construction of this data set, I found the reason of this is the same as the “improved\_water\_sources\_vs\_gdp”, which is the data set covers a wide range from -10000 till 21st century. Similarly, we cannot expect people to record down the data of water quality 1000 centuries ago.

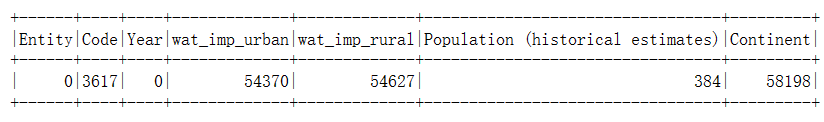


Figure Missing values of the data sets

At this moment, from the perspective of checking missing values, most data sets are already well-completed. There are some data sets which has data recorded for years before 2000 having missing values, but the data from 2000 to 2019 is relatively unabridged.

If we view all the raw data generally, it can be found that the death share of people who die from unsafe water was declining every year from 2000 - 2019 and many metrics which are related to the death of unsafe water were shifting in a linear way.

After checking the data quality, the process will be pushed onto the GitHub repository.

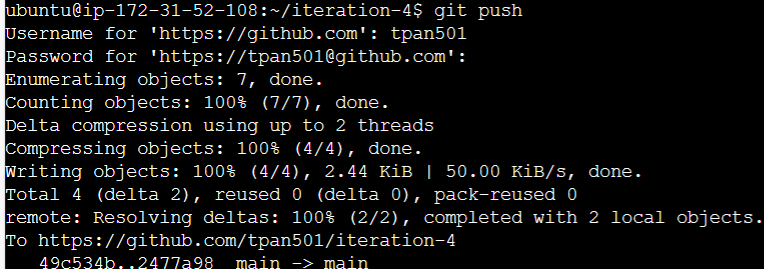


Figure Uploaded data sets in GitHub repository

# Data preparation

## Select the data

With all those data sets in hand, it is very necessary to pick out the ones which are really useful to the data mining process. The dataset collected contain data from different time period. Most data sets are from 2000 - 2019, and some others’ time spans vary dramatically (for example, for 16th centuries to now or even -10000 to now). I finally decided to select the data set from 2000 - 2019 as the foundation of the data sets construction, because the dataset’s integration and completion is the highest. Therefore, other datasets may be modified or removed in the subsequent sections.

As mentioned in the previous section, some of the data sets share the same functionality, so I should begin the data selection by removing the redundant data sets.

The first question is whether I should keep the data sets which obtain the percentage data or the data sets with the numeric data. For example, death share and death rate both represent the degree of suffering that unsafe water has brought to the country. Barr’s article (2022) discussed the difference between using percentage and counts. She claimed that proportion numbers will take the context into account, which means we will not see the numbers only but a wider picture instead (Barr, 2022). On the other hand, counts or numbers are easier to understand, but they cannot reflect the over trend because it does not take background information into account (Barr, 2022). Therefore, according to our business objectives, using percentage should be a better method to indicate the changes along with the time elapse. Therefore, the number related data sets should be put aside for this data mining process, such as “death-rates-unsafe-water.csv” and “healthcare-expenditure-vs-gdp.csv”.

Another question is whether we should keep those data sets which are lack of enough data. Although the missing value level of each data sets is acceptable, some data sets do not have enough records for some specific years, which can be useless to reach the data mining goal (mentioned in section 1.3). For example, “historical-gov-spending-gdp.csv” data is not available after 2011, which also need to be removed. Additionally, although I really want to analyze the relationships between water bodies quality and death, but “water-bodies-good-water-quality.csv”, only having the data in 2017 and 2020.

Duplicated data sets should also be considered to be removed. In the data set I’ve collected, there are three data sets related government spending: “public-healthcare-spending-share-gdp.csv”, “public-health-expenditure-share-GDP-OWID.csv”, “healthcare-expendi-ture-vs-gdp.csv”. All of them are about government spending on healthcare, but I finally selected “public-healthcare-spending-share-gdp.csv” because it has more data volume than others during 2000-2019.

There are also some other data sets, like “water-bodies-good-water-quality.csv”, only having the data in 2017 and 2020. However, I really want to keep them in the data set and use them as a reference data to check the relationships between water bodies quality and death. This part will be reviewed again in the subsequent sections.

## Integrate various data sources

In this section we will combine the rest of the data sets together, so that it will be more convenient for us to do the data cleaning and data algorithm selecting in the subsequent stages.

The specific implementation of merging the data sets is inner joining the datasets which share the same variables. For the data sets like “share\_death\_unsafe\_water.csv”, “access-drinking-water-accessed.csv”, they all have “Entity”, “Year” and “Code”, so these keys will be used for inner join in the merging process. Almost all the data sets can be combined in this way. The figure below shows the data sets which contain the information from 2000 to 2019, so we can inner join them together. Fortunately, pyspark supports the similar merging techniques as pandas does, with a little bit modification on the codes. By using “join()” method, I was able to merged the data sets up.

For the new data set which is added for this iteration, since this data set includes a large amount of data, I will use inner join to merge it with other data sets.

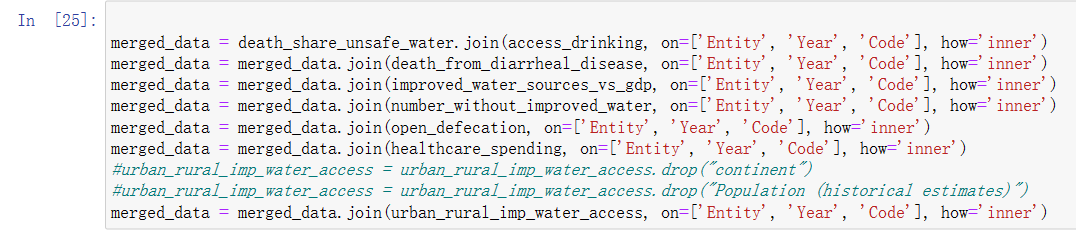


Figure Merging datasets

There is also another point which need to be noticed that there are some data sets do not have as big volume as the previous merged ones, although their dataset completion is high. For example, “physicians-per-1000-people.csv” does not record data for all the years between 2000 to 2019, and some countries do have updated data to the latest version but some others’ latest data is for 2015. For these data sets I chose to use “left join” to merge them with the final data set, and I will be determining whether to keep them or not according to the missing value proportion and my business objectives consideration.

I mentioned the importance of clean water availability impact in the hypothesis in section 2.1, so the combination of the water quality is necessary. However, this data set only contains data for only 2017 and 2020, so I have to use “left join” to merge those two data sets without discarding too much data.

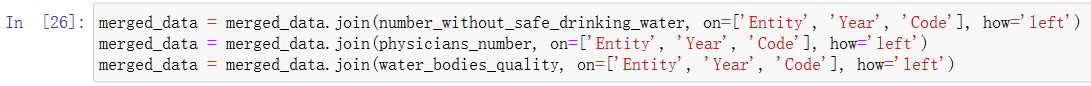


Figure Merging datasets (left)

Finally, we can export this merged data set. Under pyspark library, when I tried to export merged data set as a csv file, I found that pyspark could only export the file as a folder with the data set in there.

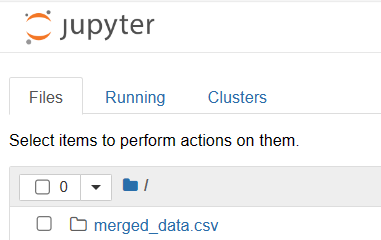


Figure Export merged data set (pysark)

However, I prefer the way from iteration 3 that is exporting the csv file directly, which is much clearer, so I transformed the data set to a pandas data frame so I could do exportation in the previous way.



Figure Transfer to pandas data frame

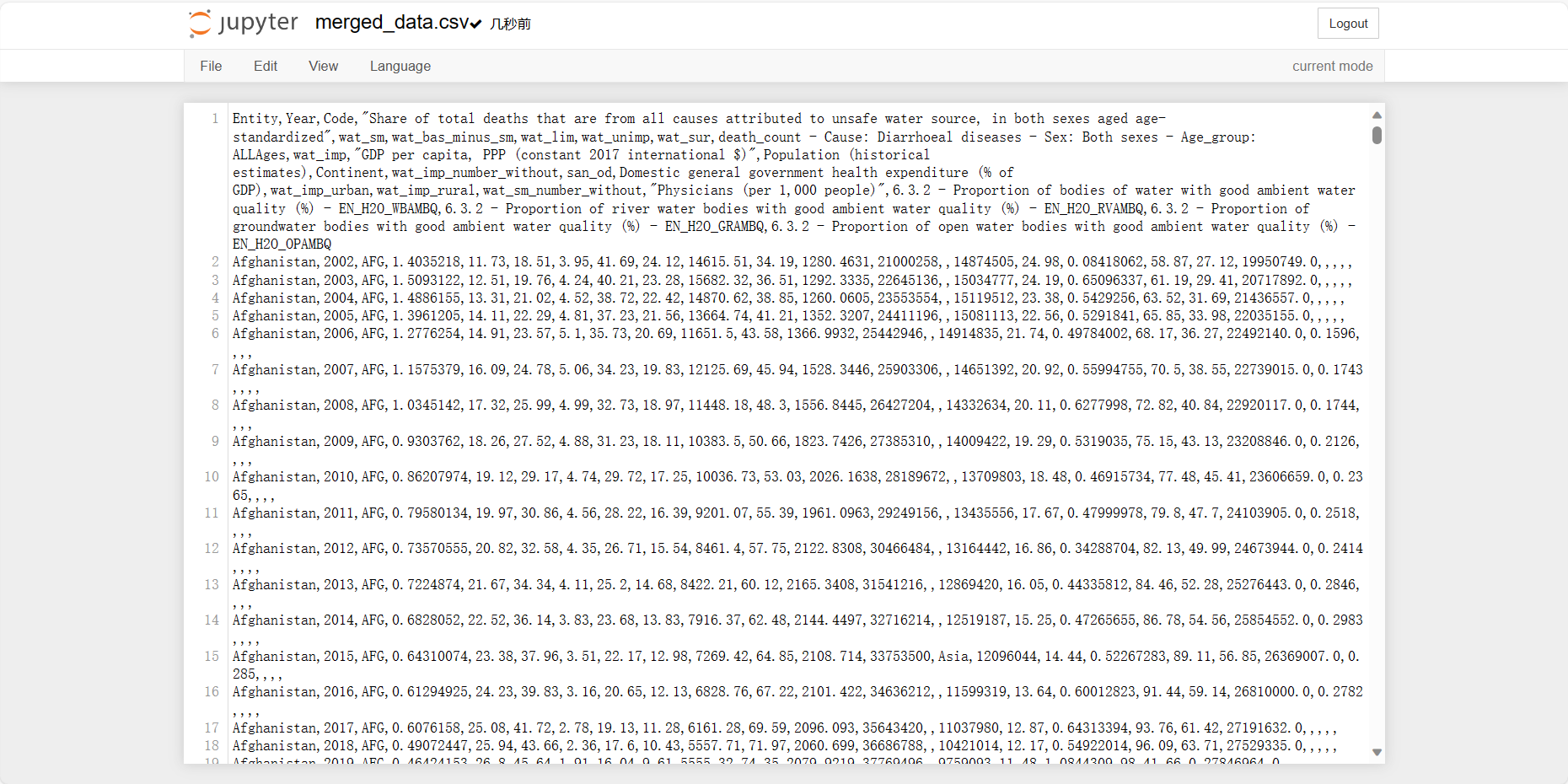


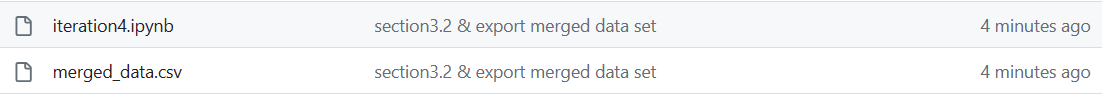
Figure Transfer to pandas data frame

The count of rows is shown below, which is 3470 rows.



Figure Number of rows of merged data set

After this step, I pushed the changes to the GitHub repository.



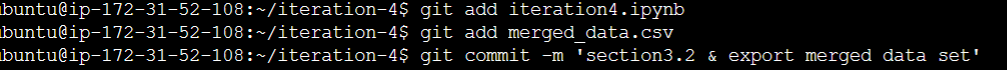


Figure Uploaded data sets in GitHub repository

## Clean the data

As mentioned in the iteration 3, noise shows up when data does not provide informative function for data mining process (de Amorim & Hennig, 2015). Excessive noise the data set will have a negative impact on the accuracy of the final results and forecast (Gupta & Gupta, 2019). As elucidated above, there would be some changes made to the data set due to the wrong data imputation and missing values, which leaded to outliers, extremes and duplicate data in the data set, and this was where the noise coming from. Data miners should discover the noise in the data and clean it before starting analyzing.

### Missing value

To do the data cleaning, the quickest way is deleting all the rows with missing values in them. There is an instruction talking about how to remove the rows with missing values to reach the data cleaning objective, which is na.drop().

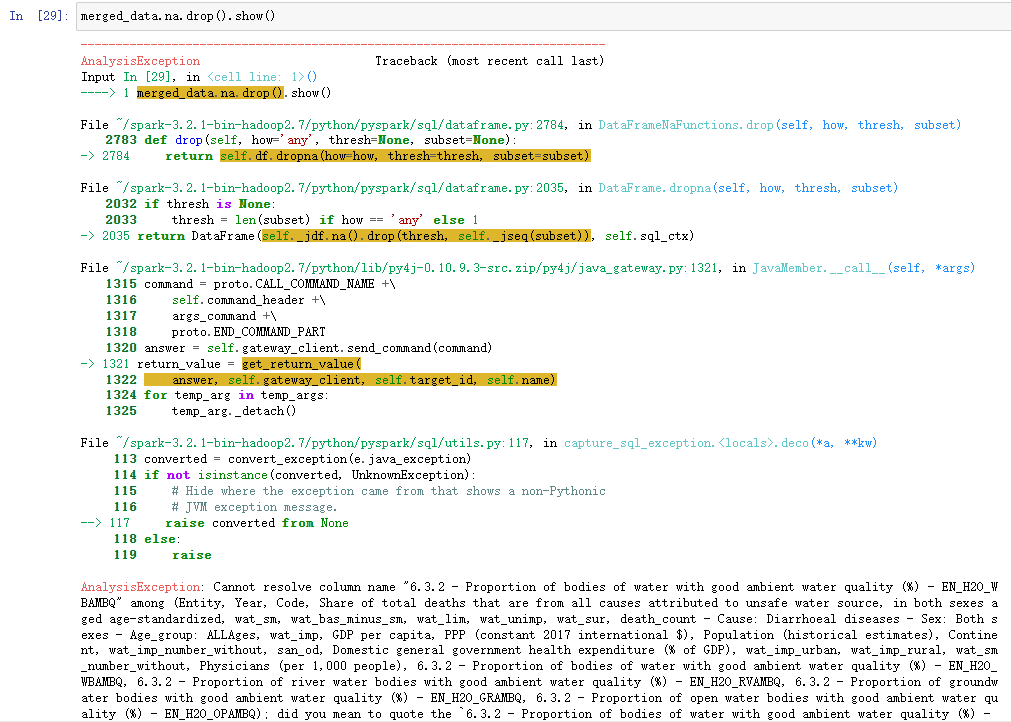


Figure Errors occurred when pyspark failed to recognize the column names

However, it looked like pyspark did not recognize my variables in a correct way, so I have to rename the variables related to the water quality (because they are too long and including different symbols).

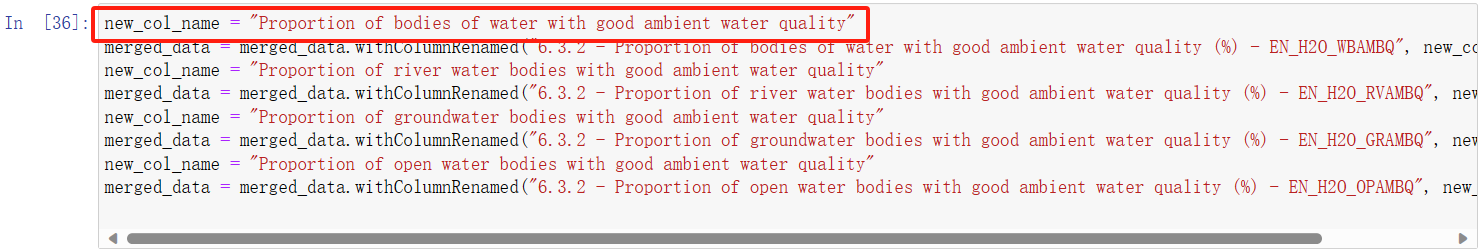


Figure Rename the columns

However, the result after dropping rows made me surprised. There is no data left in the data set after cleaning.

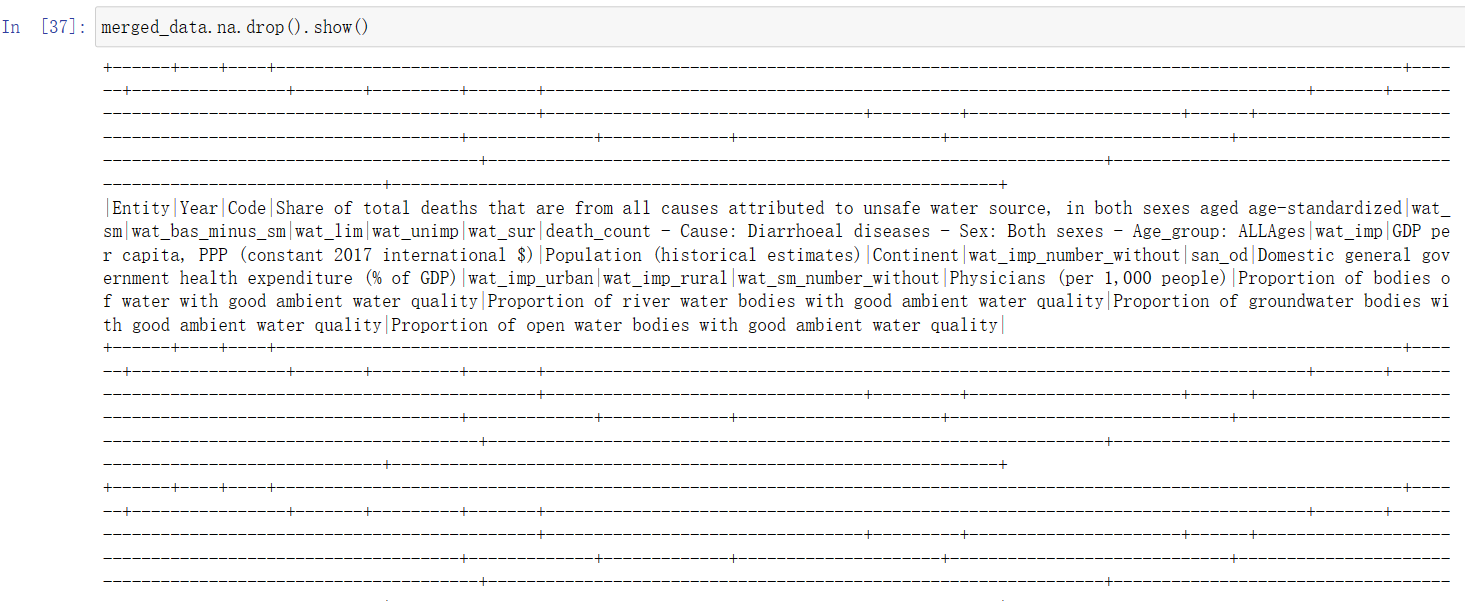


Figure cleaned data set

Clearly there is no shortcut to do the data cleaning, so I have to find my old way used in iteration 3. First, I need to find which variable needs to be cleaned. Compared with pyspark, it is much more convenient to view the proportion of missing values in each variable of the data set.

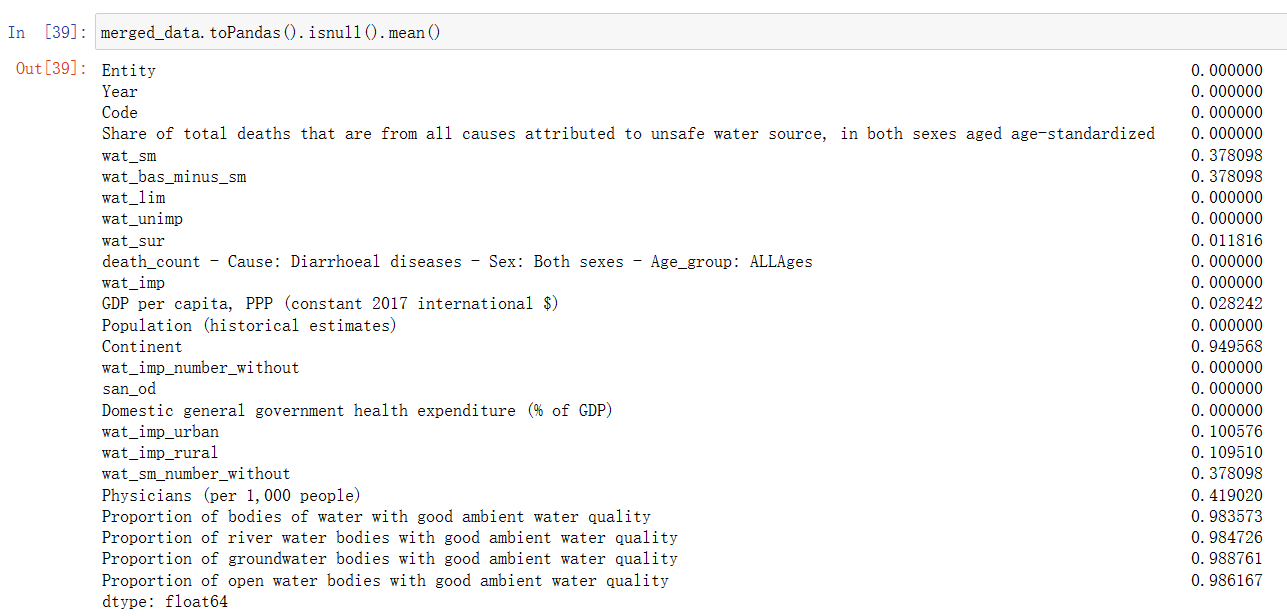
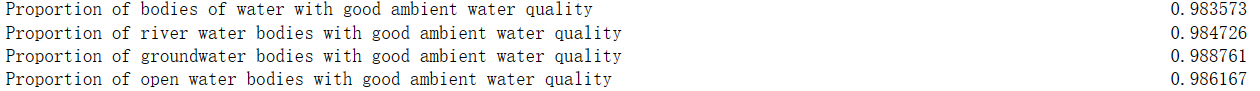


Figure Share of missing values

Because the original data sets were showing a good completion level, there is not many data sets containing too much missing values. “wat\_sm” and “wat\_bas\_minus\_sm” do have some missing values, but 37% is not a big proportion.

For the “wat\_sm\_number\_without” and “physicians (per 1,000 people), their missing value percentage is not over 50%, which does not need to be filter out of the data mining process. “Continent” is over 50% because of the way of recording, but this column is not very relevant with the death by unsafe water, so this field can be removed.

The proportion of water quality’s missing value is around 98%, and this is probably why the data set was empty after dropping rows with missing value. This is what I expected, but after contemplating the feasibility of using this data for data mining process, I chose to discard these columns.



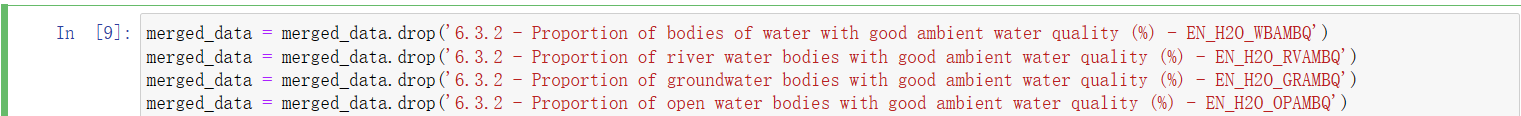


Figure Remove water quality columns

As mentioned before, “Continent” needs to be removed, so a drop() method is used to do this.



Figure Dropped column

Now I started to work on cleaning “wat\_sm” and “wat\_bas\_minus\_sm”’. By examining the merged data sets I had found that the reason causing missing values for these two columns is usually because the countries do not have any records for these two variables. This problem mainly derived from the merging process. Because it is very hard to impute values for cells without any reference values, I have to delete all the data related to the countries where “wat\_sm” and “wat\_bas\_minus\_sm” data is not available.



Figure Drop rows in columns with NaN value

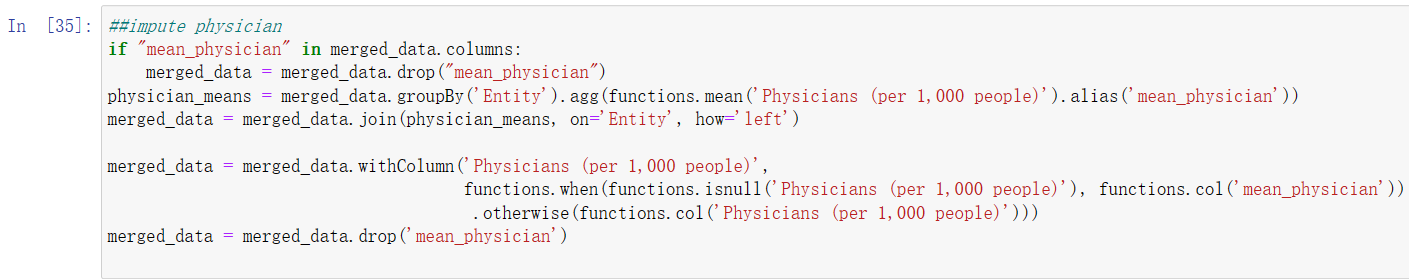
After this operation, the missing value of these two fields is cleared.



Figure Share of missing value

“Wat\_sur”, “GDP per capita, PPP (constant 2017 international $)” and “Physicians (per 1,000 people)” has relatively tiny percentage of missing value, and the missing happens randomly, so I decide to use mean value imputation according to the other values of the same country.





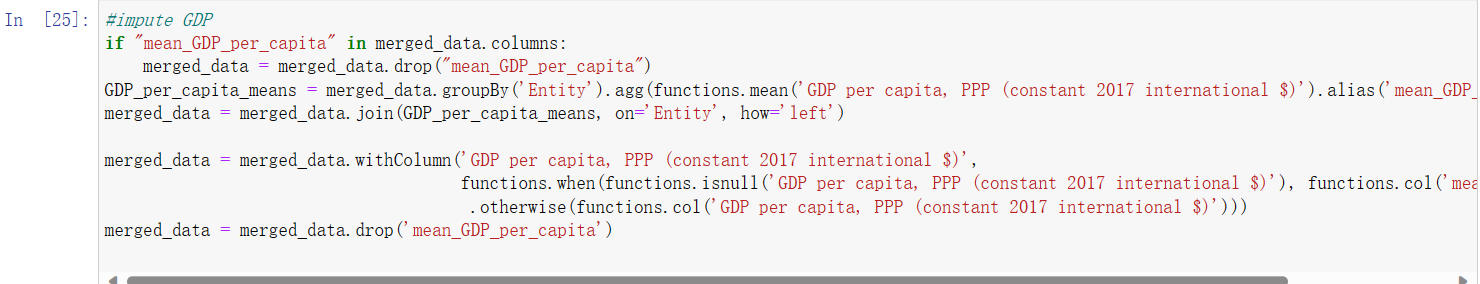


Figure Data imputation

After imputation, the missing value is cleared.



Figure Share of missing value

Lastly in this sub-section, I need to handle the missing value problem for the new data set about the people’s accessibility to safe water in both urban and rural areas. The missing value percentage is around 8.1% and 9.5%, which is quite low, but we still need to make it as clean as possible to ensure the quality.

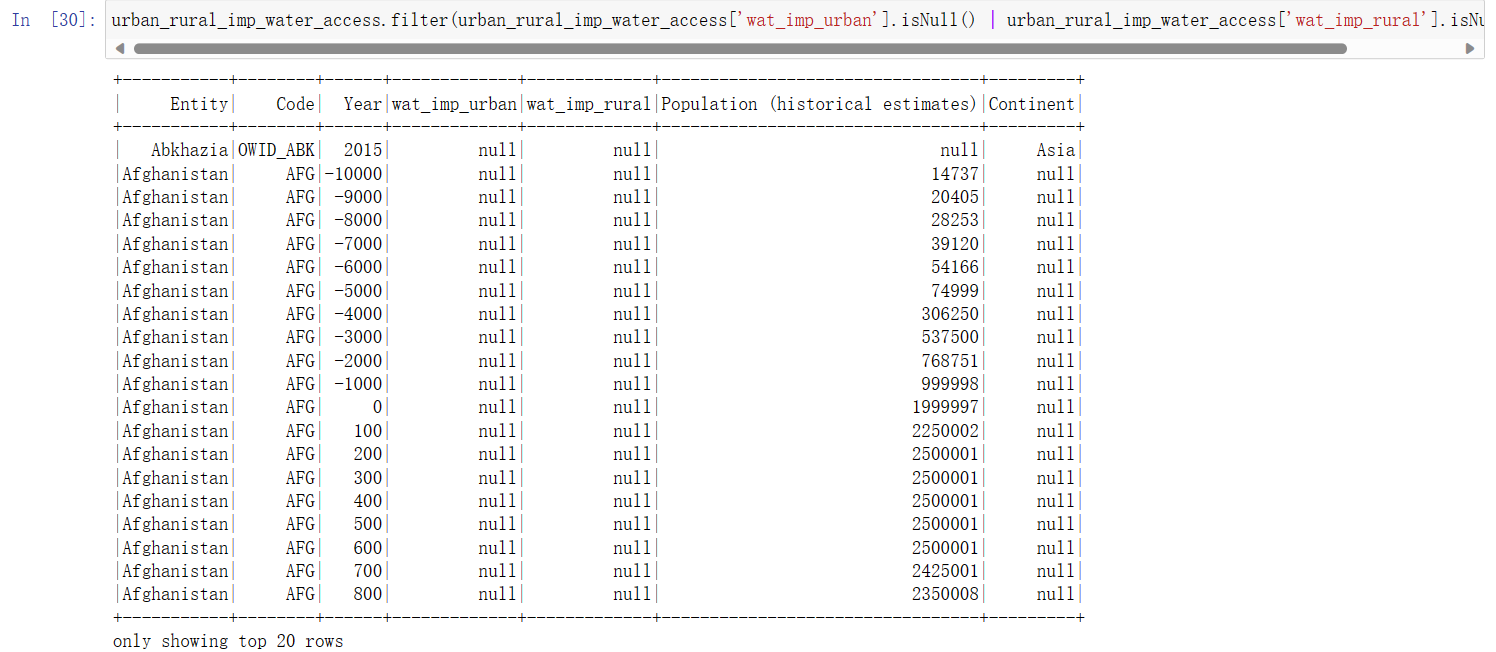


Figure Missing value of accessibility to clean water in urban and rural proportion

It can be found that the null value usually comes from the data collected in ancient time, so I just need to remove the rows.



Figure Drop rows with null value

Now all the variables are clean and the number of the rest records is 1951.

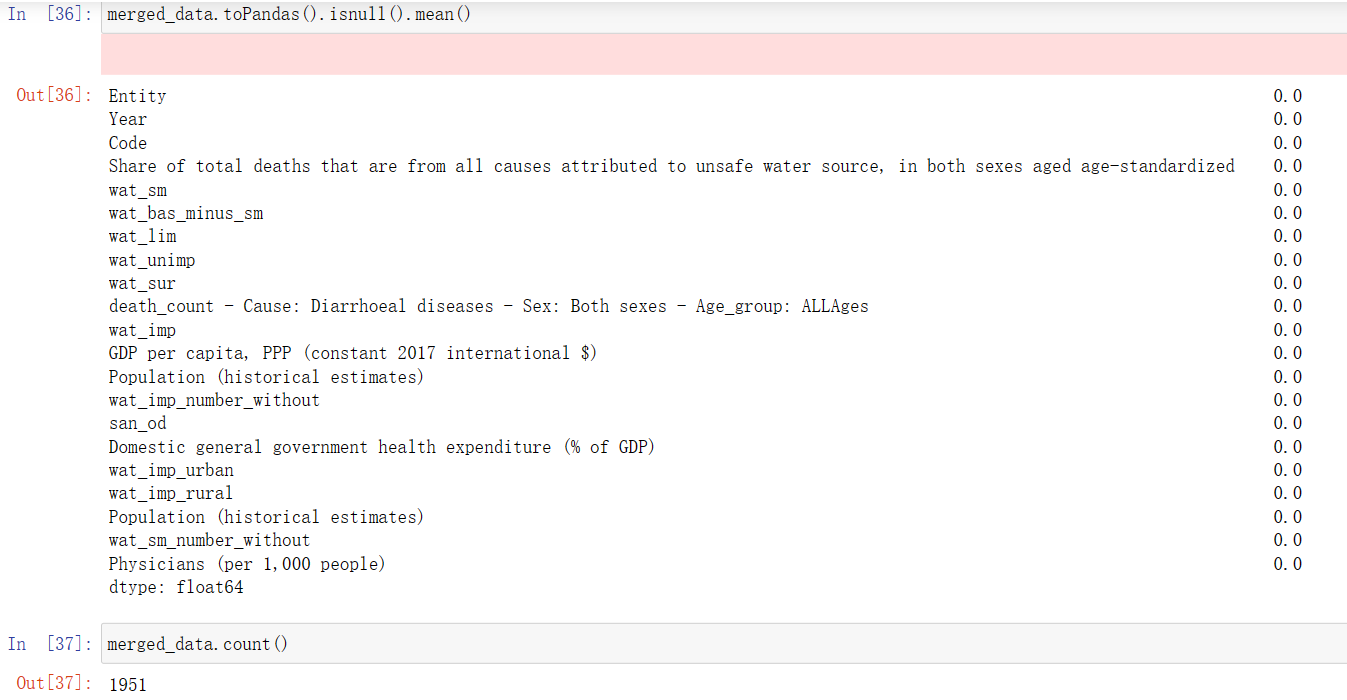


Figure Cleaned data set

The cleaned data set will be exported so I can use it for the next sections.



After this section, the changes will be pushed to GitHub repository.

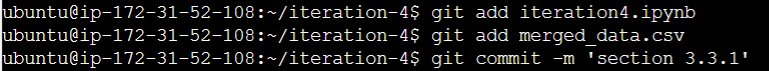


Figure Uploaded data sets in GitHub repository

### Outliers and extremes

In this section we are going to check whether the merged data set has any outliers and extremes or not. According to Hays (2023), 99.7% of data are following the standard deviations and they should be within the first 3 standard deviations of the mean. This will be used as the tool to discover the outliers and extremes of each variable.

Currently there are 1951 pieces of records in the current cleaned data set.

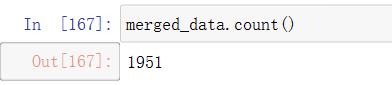
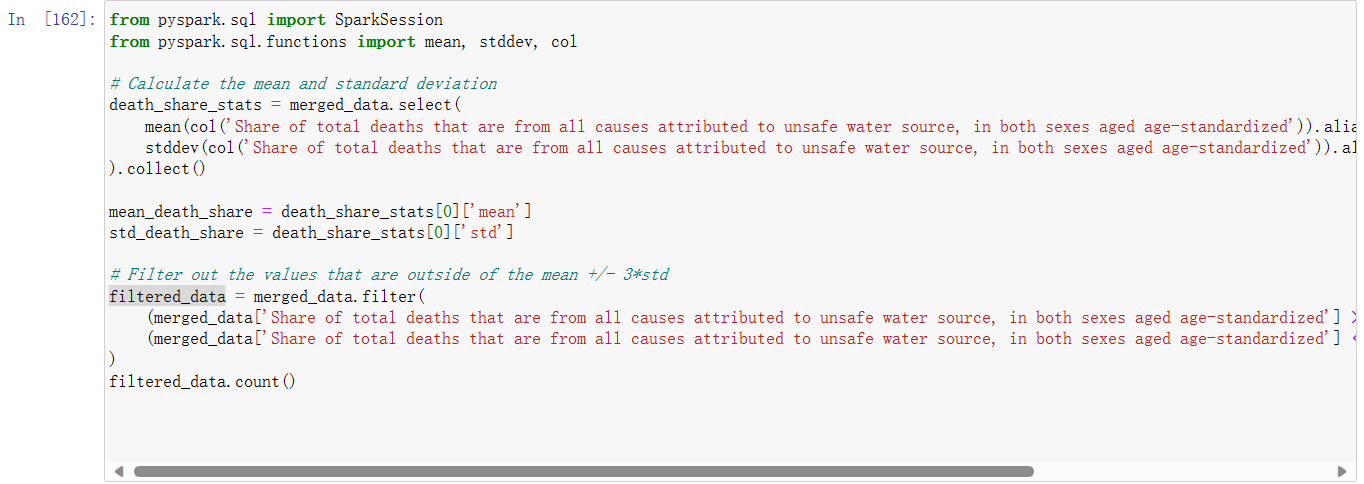
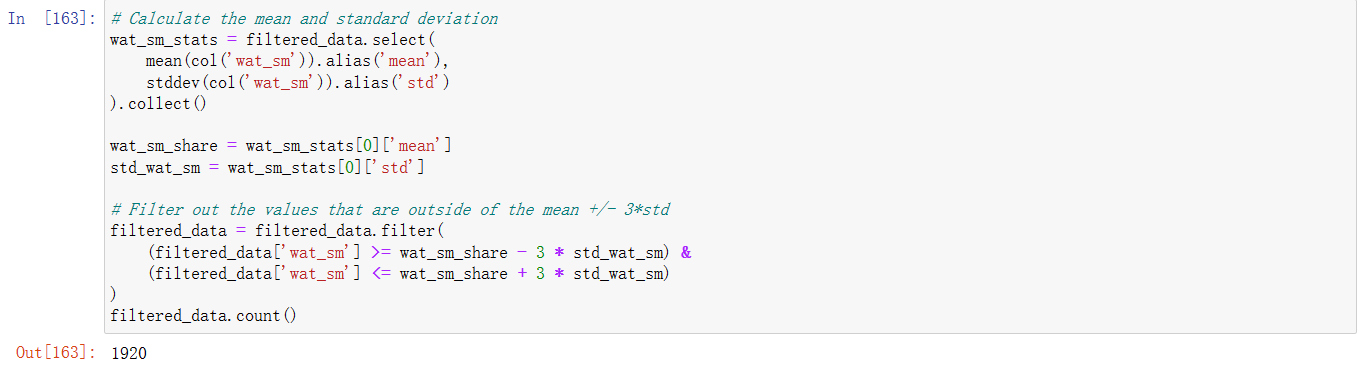


Figure merged data set count

I used “mean”, “stddev” and “col” from pyspark.sql.functions to calculate the mean and standard deviation of each variable, and then I will compare the value of each cell with the mean and standard deviation. If the value is out of the range I set, it means it could be an outlier and it can be removed.

For death share by unsafe water, it does have a few outliers and they are removed from the data set. After that, “wat\_sm”, “wat\_bas\_minus\_sm” did not show they have any outliers and extremes.





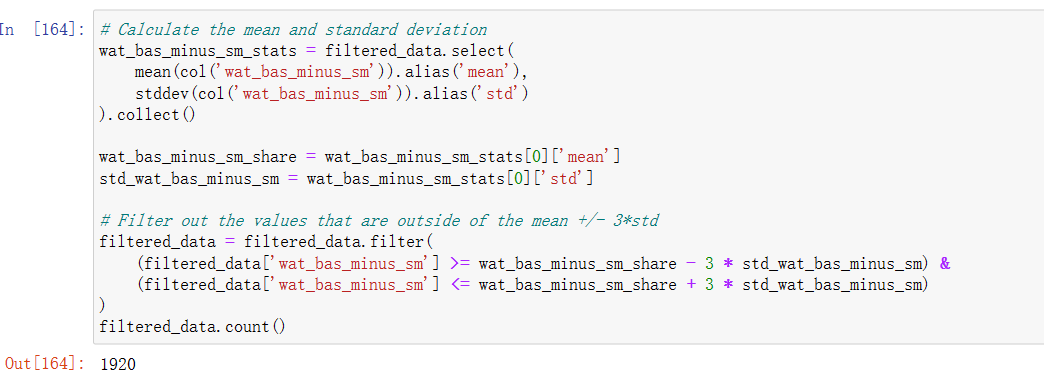


Figure Use standard deviation to discover outliers

“wat\_lim” has some outliers (1920 – 1868 = 52) and they are filtered.

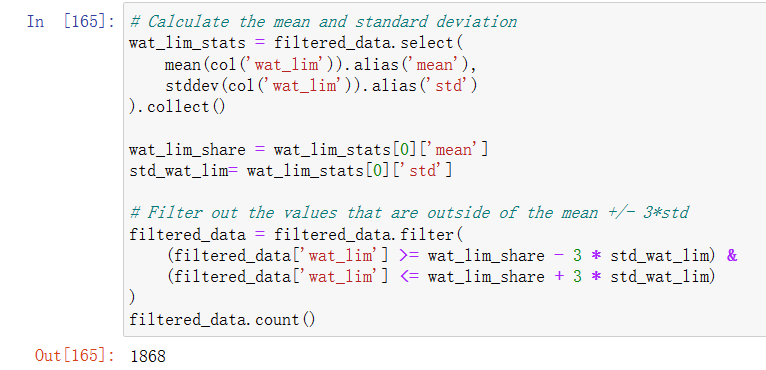


Figure Use standard deviation to discover outliers

Due to the variety of variables of this data sets, I chose to create an array and let the rest of variables to loop through the elements and filter all the outliers and extremes out. When the loop cycle finished, there were 1434 records left.



Figure Use standard deviation to discover outliers

Now the cleaning process has finished, so I committed the changes and pushed them to the GitHub repository.



Figure Uploaded data sets in GitHub repository

## Construct the data

In order to view the pattern behind the rising death percentage caused by unsafe water, I decided to regroup “Share of total deaths that are from all causes attributed to unsafe water source, in both sexes aged age-standardized”. This time I will use the strategy from iteration 3 and divide the value of this variable into 3 groups.

By using describe() method in Spyder, I can use the calculated medium (50%) percentage, 25% percentile and 75% percentile to make them as the boundaries. However, describe() in pyspark data frame shows none of these figures, but I can still use the “approxQuantile” to achieve my goal.

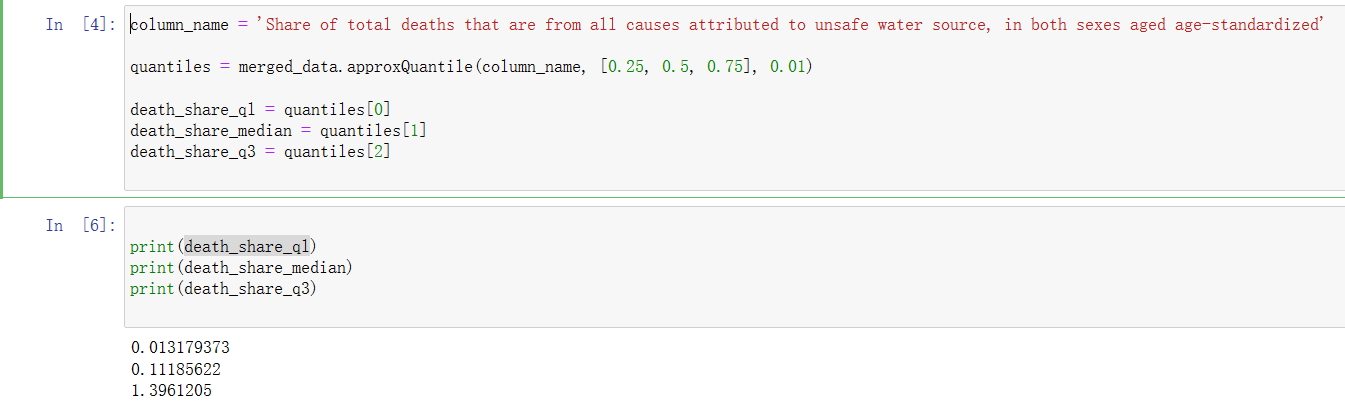


Figure Calculate 25%, 50%, 75% percentile

The countries whose death share is below 25% will be regarded low-risk, countries with 25% - 75% will be regarded as medium risk and the rest countries are in high-risk. This variable will become the target variable of this data mining process, and I will analyze the reason why the country is “high risk”, “medium risk” and “low risk” in this year.

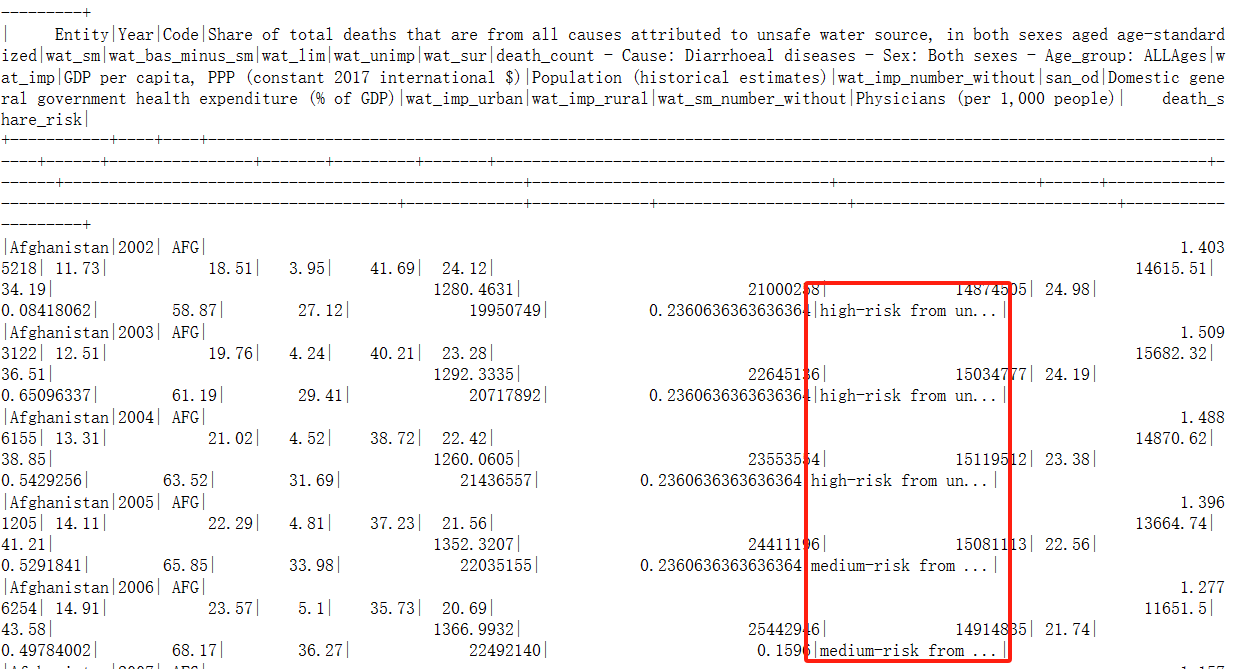


Figure Construct new column

GDP per capita can be re-constructed into a new and clearer field. According to Principles of Macroeconomics 2e, countries can be regrouped into low income, middle income and high-income countries (Steven et al., 2017), and it has multiple standards to do this. Currently “GDP per capita, PPP (constant 2017 international $)” can be used for regrouping. As the book quoted from the World bank regulation (Steven et al., 2017), $1,025 or less GDP per capita is low-income country, $1,025 - $12.475 is middle-income country and what have higher GDP per capita are high-income countries.

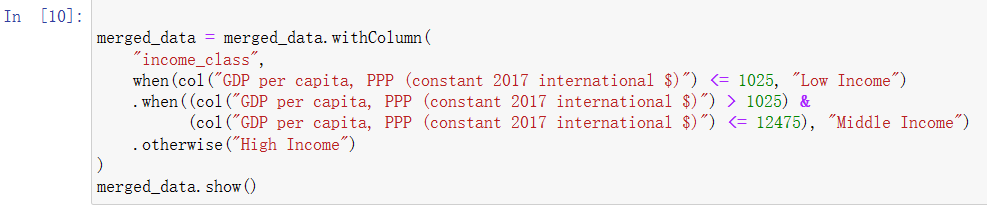


Figure construct new column for GDP per capita

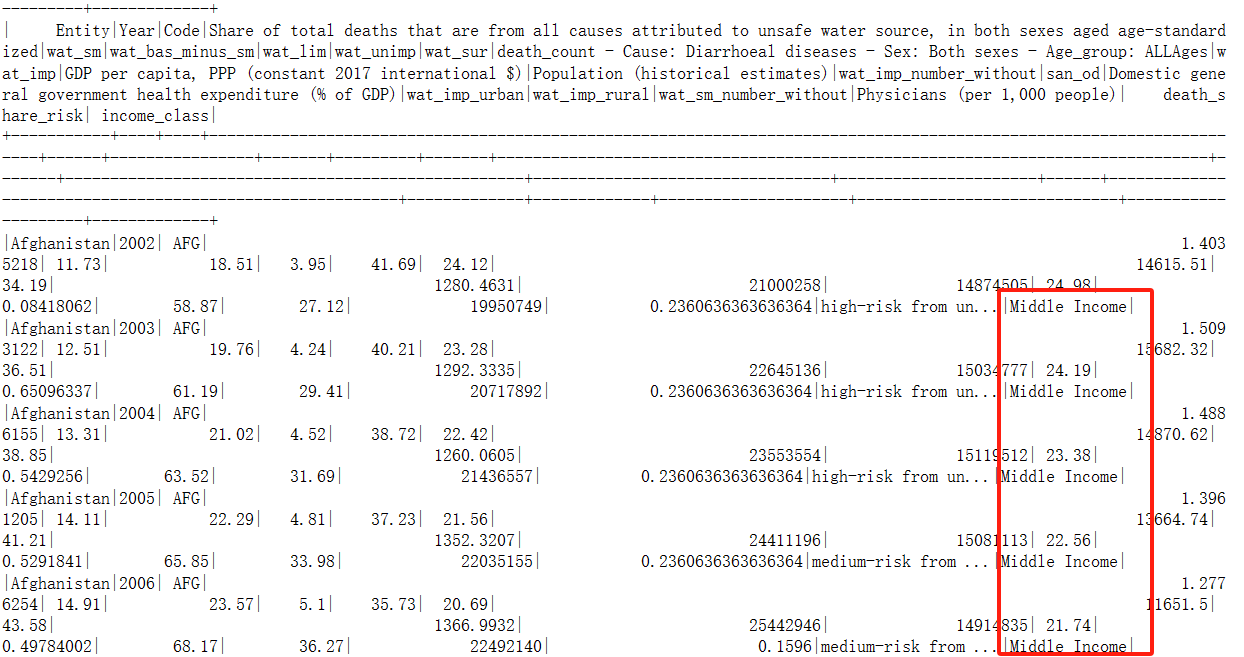


Figure Construct new column

## Format the data as required

The current data contains a range of variables, most of them are in percentage form, but there are also some of them are in pure numeric form. Therefore, in this section I will try to unify the format of this data set by transforming some of the data’s format to percentage form.

After examining the variables, there are a few of them can be transformed: “death\_count - Cause: Diarrhoeal diseases - Sex: Both sexes - Age\_group: ALLAges”, “wat\_imp\_number\_without”, “wat\_sm\_number\_without” and “Physicians (per 1,000 people)”.

The reformatting is basically done by using count divided by population.

To transform physician count per 1000 people to percentage, I need to calculate the total number of physicians in the country and then achieve the percentage of the physicians in the country.

After doing this, I removed all the pure numeric columns to keep the data set tidy.



Figure Reformat the counts to percentages

This processed data set will be exported with the name of “merged\_data\_prepared”, so we can know it has done the data preparation stage. The changes will be pushed to GitHub.

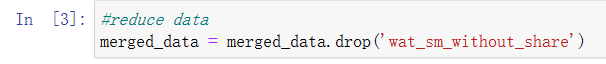


Figure Uploaded data sets in GitHub repository

# Data transformation

## Reduce the data

After data cleaning, the whole data set still needs to be examined carefully and determine whether I should adjust the structure or not. Currently, we still have two fields whose function may overlap with each other, which are “wat\_imp\_without\_share” and “wat\_sm\_without\_share”. Those fields both describe a country’s situation of people who cannot achieve good water source, and the difference is “improved water source” is more general than “safe-drinking water source”. Therefore, I decided to remove the “wat\_sm\_without\_share” columns and kept the general one. The reason is that too analyze the death share caused by unsafe water, if improved water is proved to have negative relationship with the death, safe managed water source should have the same impact.



Although it is not apparent in the data set, it can be viewed that some countries do not obtain the data for some early years like 2000, 2001. From the value distribution of “Year”, early years data of some countries are not as completed as the one of others. It is also noticeable that the count of data from each year does not vary a lot in this data set, and only the data recorded before 2002 has relatively smaller number. Therefore, I decided to filter the rows out whose value in the “Year” column is equal to or less than 2002, so I can try my best to keep the same volume of data for countries.

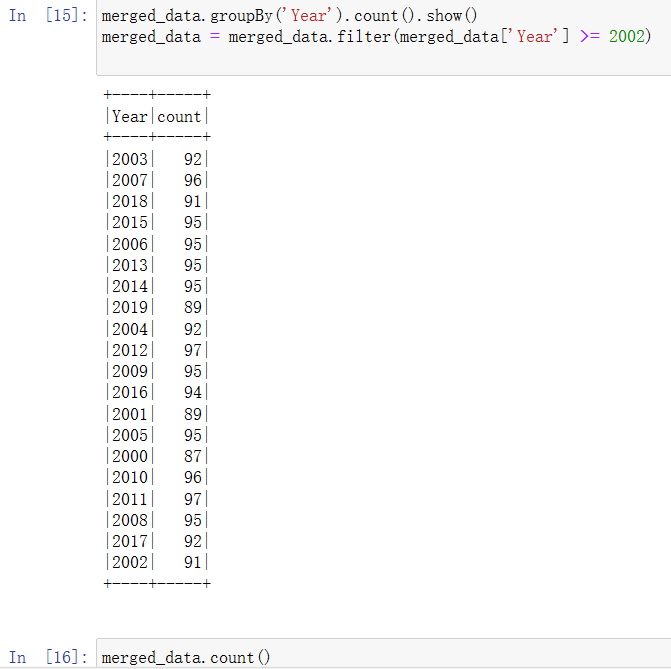


Figure Reduce data by selecting

## Project the data

The goal of data mining of unsafe water death is to figure out the affecting factor behind the death share of people from unsafe water, so whether a region has high, medium or lower risk of death from unsafe water will be the target variable of this research. Currently the distribution of our target variable is:

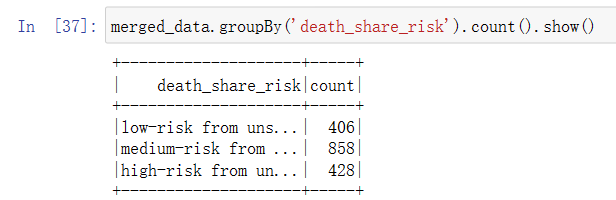


Figure Current distribution of target variable

It is very found that the distribution is quite even, but high-risk party and low-risk party still has smaller proportion than medium-risk party. Therefore, I decided to boost the samples in the high-risk and low-risk party to increase the accuracy of the subsequent machine learning and data mining process.

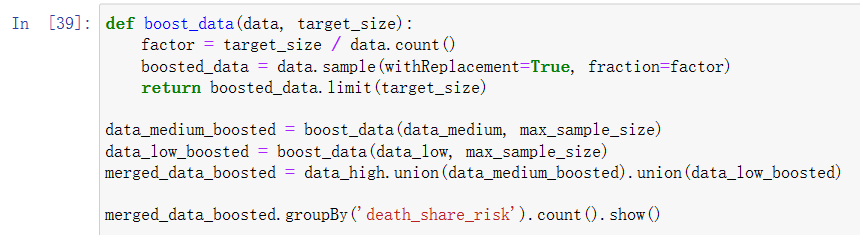


Figure Balance the data set (boosting)

The balanced distribution is shown as below.

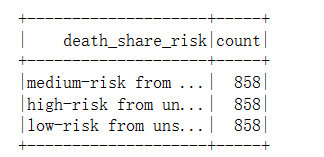


Figure Balanced target variable

The changes will be pushed to the GitHub.

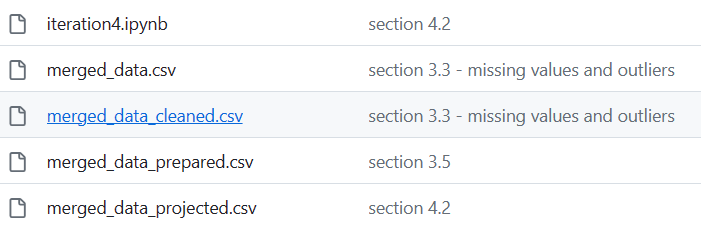


Figure Uploaded data sets in GitHub repository

# Data mining Method

## Data mining method candidate

The data mining objectives, which was stated in section 1.1, is discovering what causes the higher death share in some countries because of unsafe water. Therefore, there are several choices for me to decide which method should I use for data analysis.

Classification is a supervised learning method (Delua, 2021) that is used to predict discrete labels or classes of data. In classification, data values will be divided into different classes, which can help us to build connections between different variables, so we can make a prediction of the future (Sharma, 2022). Classification is widely used in the way of “binary classification” when there are only two values in the target variables, and it also compatible with target variables with 3 types of value. In my context, the merged data set can be divided into three parts according to the value of the target variable I have set in the section 3.4. Basing on this, I can identify which variables are casting significant impact on the target variable.

On the other hand, clustering is another data mining method that gather data points together by analyzing their similarity (Sharma, 2022). It will separate all the data into different groups, and it will label each group with their different features. The data points are similar in one group but totally different from the ones in another cluster. Clustering is widely used in biological domain and business domain to identity different demographic groups (Sharma, 2022). In my context, I can separate the whole data set into different clusters, and each cluster indicate different features, and I can conduct summaries about which cluster has the higher risk of death from the unsafe water and what are the characteristics of the cluster.

The main difference of these two methods is classification categorizes data points according to their class labels, but clustering categorizes data points by their similarity (Bisht, 2021).

## Data mining method selection

According to the business/situation objective, I am focusing on discovering the factors resulting in the higher risk of death share caused by unsafe water. Therefore, for the binary distribution of the target variables, it will be better for me to use classification method for data mining, because I am going to predict an object’s “death\_share\_risk” attribute, which is ternary value, based on the characteristics of input data.

# Data mining algorithm

For classification method, there are a few algorithms that can be used for this data mining process, such as Decision tree, Random Forest and other algorithms. I need to mention due to the reason that I am analyzing the target variables within 3 variables, so I cannot use logistic regression because it prefers binary values.

## Exploratory analysis of data mining algorithm

### K-Nearest Neighbors

KNN is a supervised learning classifier, which relies on the closeness between data points for classification or prediction. It can be applied to regression or classification problems, but is usually more used for classification problems (IBM, 2023). When a new observation is available, it finds the k closest observations to that observation in the training set and predicts the most frequent class of these k observations (IBM, 2023). In this current situation, I am working on a classification problem because the values from the target variable is “discrete” instead of “continuous”.

### Decision tree

Decision tree was once used in the iteration 2 data mining process. This algorithm will display the classification in a form of tree. Each node shows a condition and it will be divided into two or multiple sub-nodes. The dividing process will keep continuing until the last condition (Sharma, 2022). When it comes to the last node (leaf node), it will provide us with a result for prediction and relation.

### Random Forest

Random forest is based on decision tree, and it is able to combine multiple decision trees’ computation to achieve the final result from various sub-sets (Sharma, 2022), and this makes it less overfitting and more accurate.

## Selecting data mining algorithm

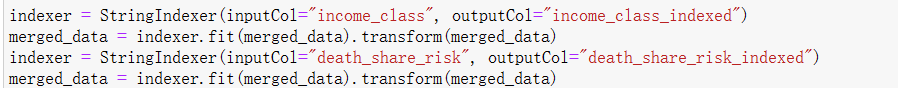
Random forest is an ensemble method (IBM, 2023), and it can aggregate the results of decision tree and generate the most recognized result from them. To some extent, this means random forest algorithm can decrease the overfitting of the machine learning process and increase the accuracy of the model output. Compared with Random Forest, K-NN’s result can make people feel easier to interpret, and meanwhile its accuracy is high enough to achieve the data mining goal. Additionally, K-NN does not require too much adjustment on the algorithm’s parameter. However, with pyspark library data frame, I cannot install K-NN library from the third party in an online Jupyter Notebook, so it is not realistic for me to use K-NN for this data mining process.

On the other hand, Decision tree sometimes may get over-complexed if there are too many circumstances.

Therefore, Random Forest will be used as the data mining algorithm for this iteration.

## Build model with the algorithm

Before everything start, I need to go through a few extra steps, otherwise the algorithm will not function correctly. In section 3.5 I had changed the values’ format from some variables in the data set. For example, I constructed new field “death\_share\_risk” by classifying all the data points whose death share was over the median amount as “high risk from unsafe water. Additionally, I also set “income\_class” variables according to the GDP per capita of each record. Those new values are in String type which can not work properly if I do not change them to numeric type, because “VectorAssembler” does not support String type value. Therefore, I need to transform them into numeric types first.



Then, I need to combine multiple numeric columns into a single vector column. When I was choosing the input variables for algorithm implementation, I filtered the “Entity”, “Code” out because they were not relevant to the target variables. 'Share of total deaths that are from all causes attributed to unsafe water source, in both sexes aged age-standardized' and reformatted field “death\_share\_risk” and its index-transformed column were eliminated as well because they are the target fields. Leaving them in the input variables will lead to misunderstanding when I am analyzing the patterns.



Figure Input variables

Because the algorithms I chose belong to classification, I divided the data set into two parts to use them as training data set and testing data set., and the proportion was set as 80% and 20% respectively (iteration 3 was 70% to 30%). Random state was set 40 so I could get different partitions for every trial.

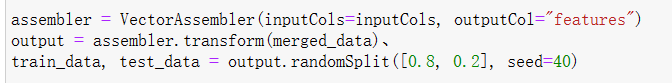


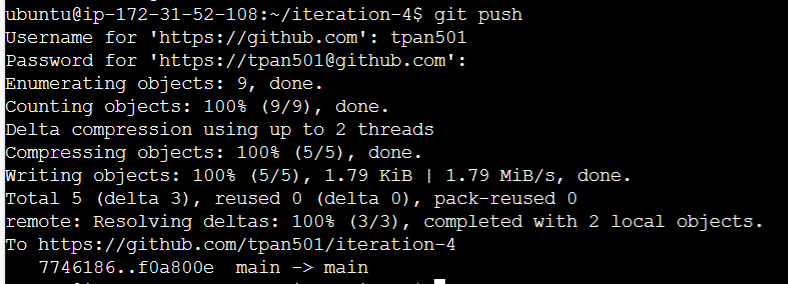
Figure Algorithm setting

The random forest implementation code is shown below.



Figure Implement Random Forest

The changes had been pushed to the GitHub.



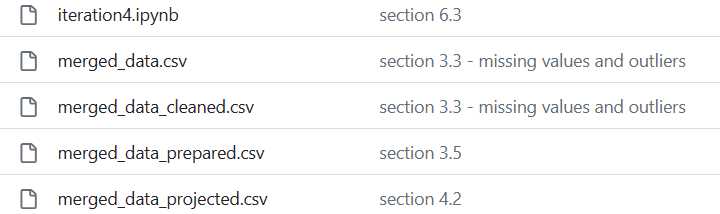


Figure Uploaded data sets in GitHub repository

# Data mining

## Create and justify test design

As illustrated in the section 6.3., before creating a model I need to determine the partition for the training data and testing data. The proportion for training and testing data set is 80% and 20% of the whole data set, and random state was set as 40 to generate different partition each time. Random state can make the division random.

The reason of setting training set and testing set is to guarantee the best result and avoid the over fitting (Gholamy et al., 2018). Training set is for models to do learning and parameter optimizing, while the testing set is for testing and comparing with the current data (Gillis, 2022). The variance in your parameter estimations increases with smaller training data, and with fewer data points for testing, the volatility of your performance metric rises (Sachin, 2023).

There should be a certain ratio for the volume of training set and testing set, and what is mostly used is 80/20 (Sachin, 2023). One of the reasons that advocate 80/20 is Pareto Principle, which is also called 80/20 rule, where 20% of elements are responsible for 80% of results.



Figure Setting training and testing partition

Currently the number of data records in training set is 2040 and the one of data in testing records is 534.

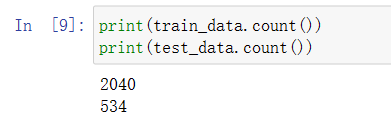


Figure Volume of training and testing data set

## Conduct data mining

A part of conduction of random forest algorithm has been displayed in section 6.3. I generate a new assembler object and assign its value with VectorAssembler, where the parameter is the input variables that I had set in section 6. 3.. After transformation I got a new data frame “output” with an extra column named “feature”.



Figure Conduct data mining

Then, I imported RandomForestClassifier from the pyspark.ml.classification library and created a classifier object. For the Random Forest metric, I set number of tree to 300 because I want to achieve more accurate result from more decision trees in the random forest, and “random\_state” was set to 40 to make the computation random, avoiding the radical situation happening.

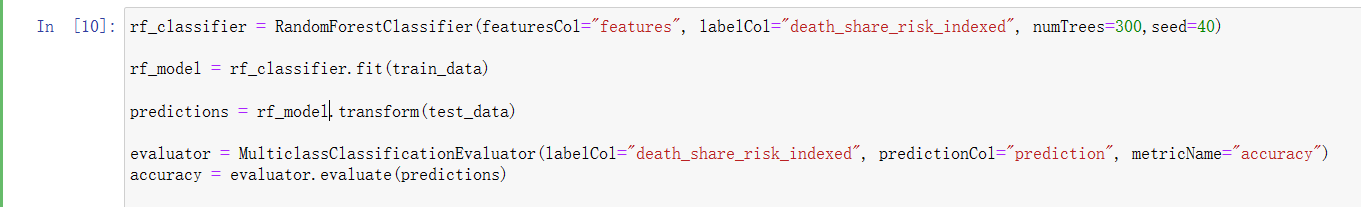


Figure Conduct Random Forest

## Search for patterns

In this section, I will be looking for the patterns from the result of random forest. For this iteration, due to the limitation of pyspark and Jupyter notebook, I was informed that the shap library and scikit-learn library are not available in the Jupyter Notebook. Therefore, it will be difficult for me to explain the relationships between some typical variables and the target variable. Instead, I had to use the “featureImportances” attribute to check the how important of each variable to my target variable, and the result is shown below.

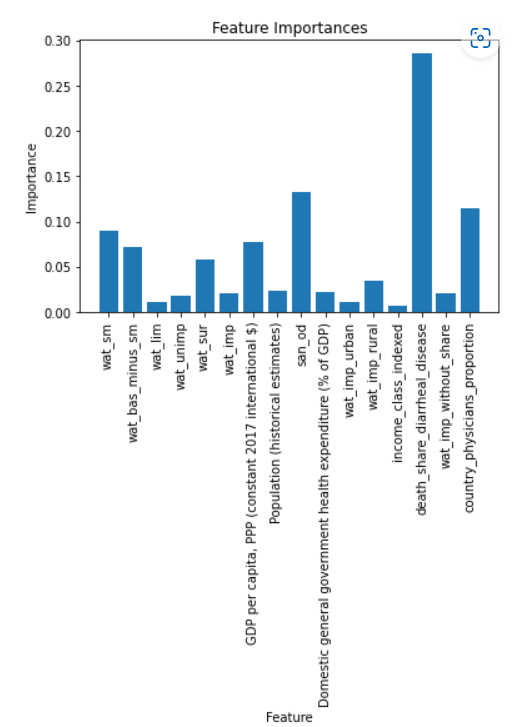


Figure Feature importance visualization

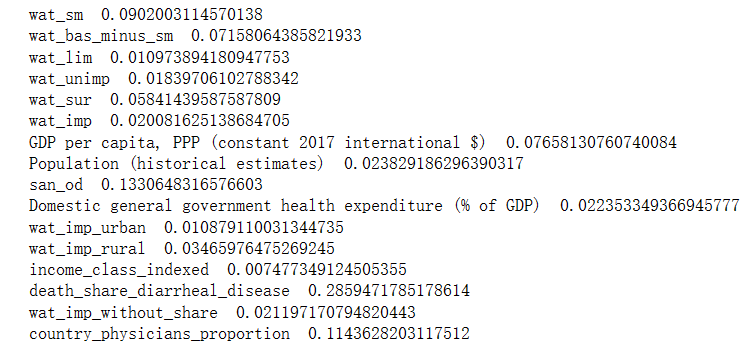


Figure Feature importance visualization

Third related variable is about how many people are physicians or doctors in the country, and the proportion of people who can get safely drinking water ranked after it. Therefore, the most important factors which cause death share rising might not be the availability of water source or intrastation.

In order to carry out a clear understanding of the relationships between different factors and the target variable, I drew a heat map with all the variables on it, which showed the correlation among different fields.

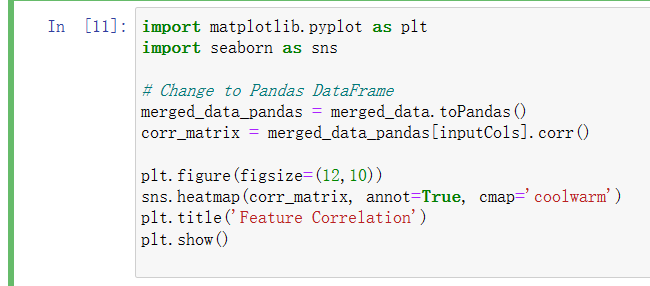


Figure Generate heat map by seaborn and matplotlib

The generated heat map is shown below.

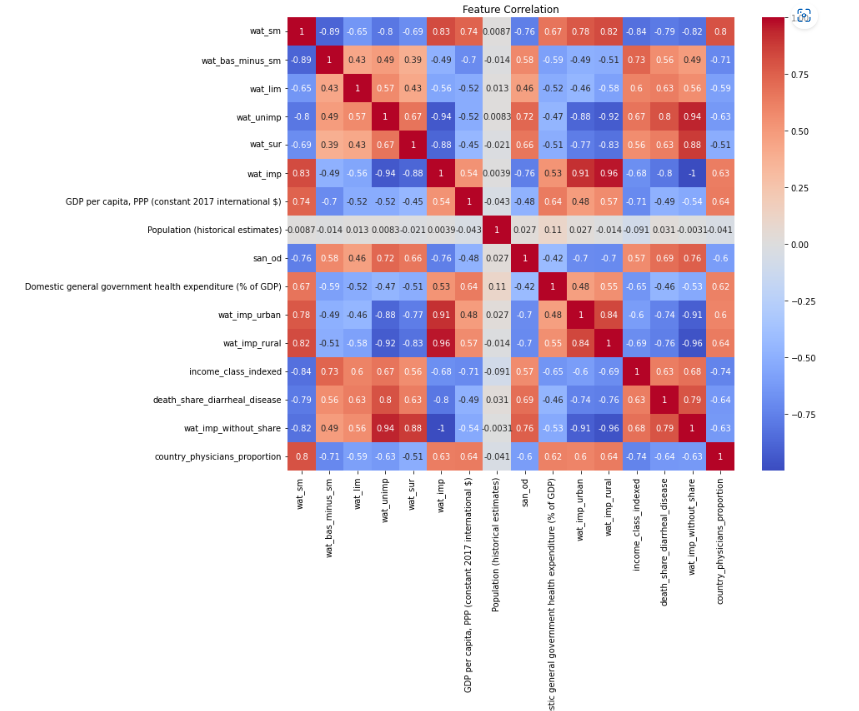


Figure Heat map

### Pattern 1

The most important factor which can impact the death share most is the proportion of people who died because of diarrheal diseases in the country. The correlation of it reached nearly 30%. According to the heat map we can discover the fact that the proportion of people who die from diarrheal disease is significantly overlapping with the people whose accessibility to unimproved water is high. Apart from that, the correlation of “death\_share\_diarrheal\_disease” is also high with “san\_od”, “income\_level” and “wat\_imp\_without\_share”.

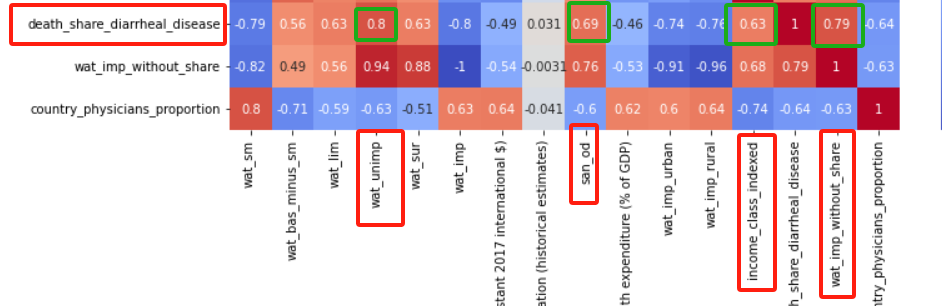


Figure Correlation of death\_share\_diarrheal\_disease

### Pattern 2

The second important factor is the proportion of people who are doing op defecation in the public area, and its feature importance also reached around 15%. “san\_od” also related with “wat\_unimp”, “death\_share\_diarrheal\_disease” and “wat\_imp\_without\_share” closely. It is noticeable that open defecation also comes along with the changing in surface water accessibility.

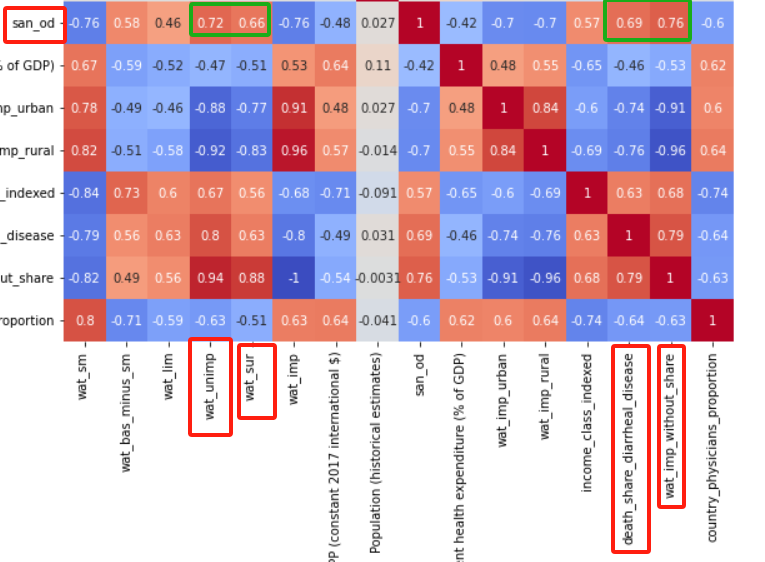


Figure Correlation of san\_od

### Pattern 3

Differing from iteration 3, physician number of a country was the most important factor of people dying from the unsafe water, but in this iteration, it had become the third important factor. According to the heat map, physician number is also closely related with the accessibility to the safely managed water, the improved water, the government health expenditure and the prevalence of safe drinking water in urban and rural areas.

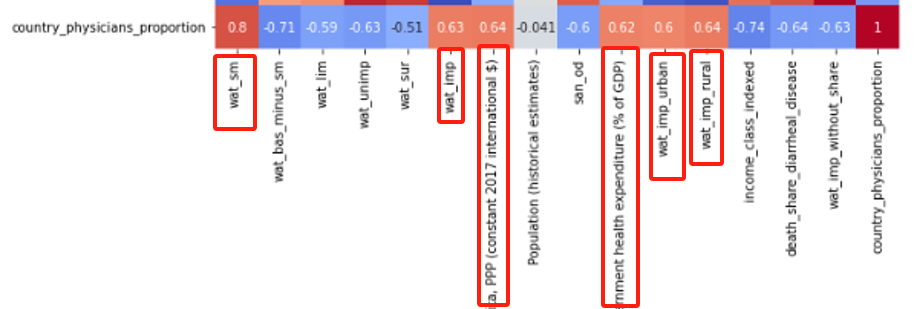


Figure Correlation of country\_physician\_proportion

# Interpretation

## Study and discuss the mined pattern

### Interpret pattern 1

In pattern 1 I have concluded that the death percentage of people who died from diarrheal disease is overlapping with the people who died from unsafe water in most cases. This phenomenon indicates that after people used, drank or took the unsafe water, they would be very likely to encounter diarrheal diseases, which was too serious to cure and caused their death at last. Additionally, we can find in the heat map, the rising of diarrheal disease comes with the rising of proportion of people doing open defecation in the public area, and low accessibility to safe managed water and lower income level. It is very possible that people in poorer countries and regions lack the approaches to achieve clean water, leading to the high prevalence of diarrheal diseases.

### Interpret pattern 2

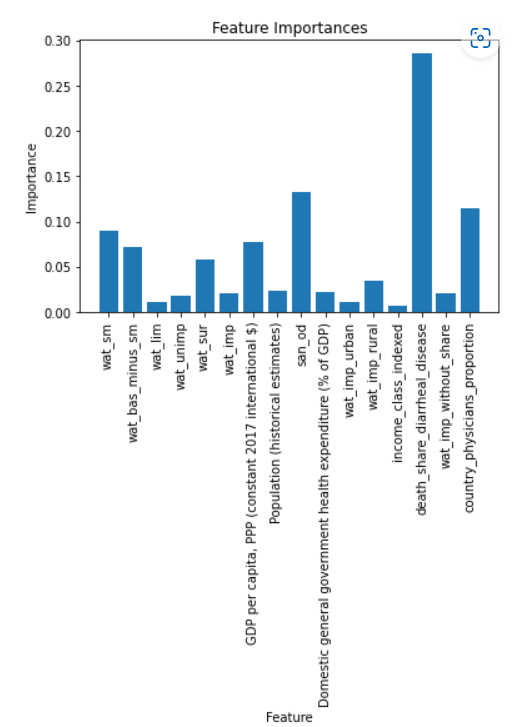
Pattern 2 is about the second importance factor, which is the proportion of people who are doing open defecation. Apart from the similar variables that are related with “san\_od”, there is one more variable “wat\_sur” is also closely related with it. Apparently, if more people are doing defecation in the public area instead of their sanitation facilities, the surface water will be polluted. Meanwhile, if people are not able to get clean water from the basic infrastructure, they have to go outside and get water from the surface water like river, stream or rain. However, due to the influence of open defecation, gaining water from outside environment will only make people suffer from diarrheal diseases, thereby they could die because of this.

### Interpret pattern 3

Physician number ranks the third position in the importance of people’s death caused by unsafe water. When the number of physicians is low, the country usually does not have a high availability of safely managed water source. On the other hand, if the number of physicians is high, it usually comes with the higher proportion of improved or safe water source accessibility, higher healthcare expenditure from the government and higher coverage of clean water in both urban and rural areas. It is worth mentioning that the higher physician’s number does not mean higher income level. Therefore, it is worth trying for those lower-income country to introduce more doctors to stop the growing death share.

## Visualize the data, results, models and patterns

In this section I will display all the output and results from the random forest model.



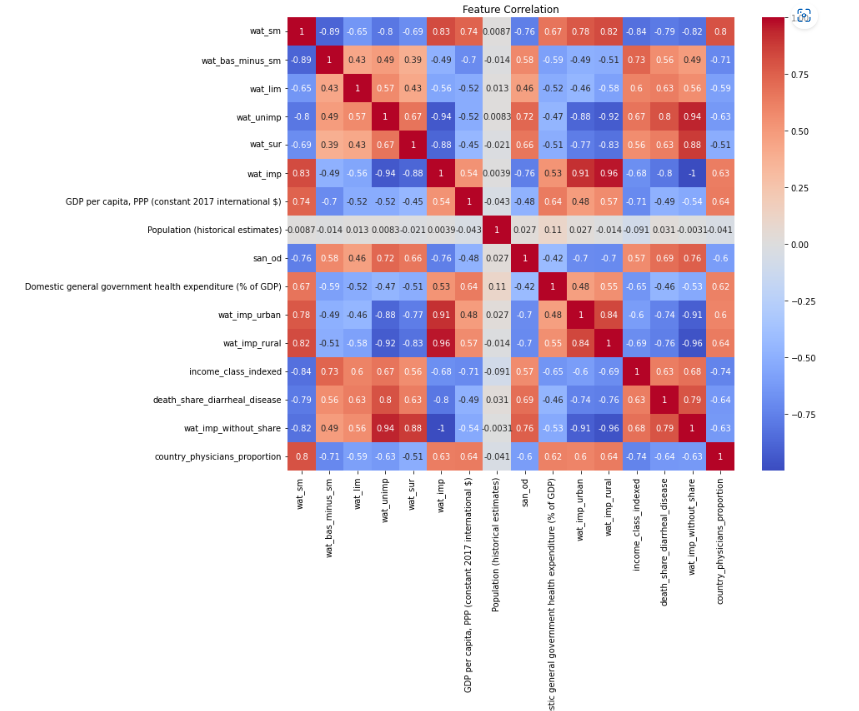


Figure Visualization of results and model

## Interpret the results, models and patterns

**Pattern 1:** “death share diarrheal disease” is the most related factor that affect the risk of death when people are using the water there. After the initial analysis in section 8.1.1, we can say that it is reasonable. According to the WHO (2023), unsafe water contains micro bacteria which can pass the diseases like “cholera, dysentery, typhoid and polio” to the people, and this will result in 505000 diarrheal diseases every year. When diarrheal disease is prevalent in one region, it means the proportion of people who can reach clean water is relatively low, because the dirty water is one of the most critical sources where the diseases come from. Meanwhile, when the death share of people having diarrheal disease is high, the income level tends to be lower, which is also fair because government need to be funded to change the situation.

**Pattern 2:** open defecation also takes the most responsibility to the rising death share of people by using unsafe water. From section 8.1.2. we can discover the relation between open defecation and surface water accessibility. It is imaginable that if people do more defecation in the public area, the excrement will get cycled in the natural environment. Because it won’t get disposed in a short time, it will be some parts of it flowing into the surface water and contaminate it. If the water is polluted microbiologically, people’s lives will be at risk (WHO, 2023).

Pattern 3: The third most important factor is “country\_physicians\_proportion”, which is reasonable, because if there are more doctors and other people who can provide health care services, more people can get help and support after they have used the unsafe water. According to Centers for Disease Control (2022) and Prevention in the US, the gems, chemicals and other hazardous substance in the water can cause people to infect various diseases like Guinea worm disease. With more physicians and doctors in the country, the percentage of people who die from unsafe water will be reduced because they can be saved from the diseases. However, physician number is in proportion to the healthcare expenditure of government, no matter which income level this country has, which indicate that a country’s policy and plan is very important to reduce the death number of people from being killed by unsafe water.

## Assessing and evaluate results, models and patterns

Before the data mining process began, I once made a hypothesis on the death share caused by unsafe water, which is “the higher death share is related with the local safe or clean water’s availability to the residents there, and government support also plays important role in it”. After analyzing the data set, it can be seen that the hypothesis may be partly correct. The availability of water source do have a great impact on people’s health, but its impact is actually limited. What impacts our target variable most is the proportion of people who died from diarrheal disease and the one of people who are doing open defecation. The number of physicians also becomes one of the most important factors. However, it is different from what I carried out in iteration 3 that the number of physician is actually closely related with the government’s expenditure in healthcare, so the fundings and budgets do matter.

Now when I came back to the model I have used for the data mining, I need to make an assessment to test my data mining model to find how well the model worked. I imported MulticlassClassificationEvaluator from pyspark.ml.evaluation. Then, I put my test data set to my trained model and generated a data frame “prediction”. Lastly, I used evaluator to test the accuracy of my model.

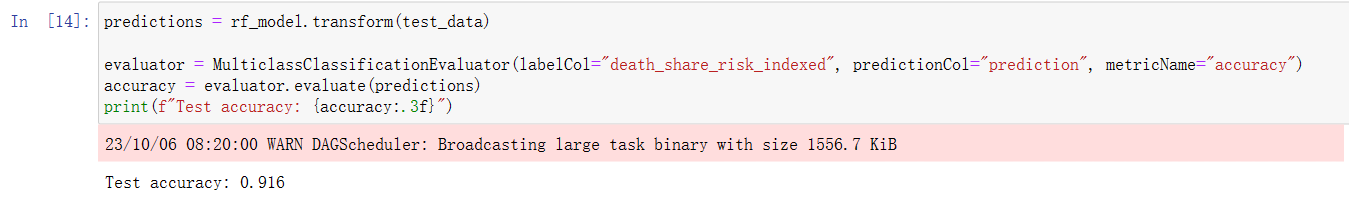


Figure Examine the score of the model

The results are shown below.

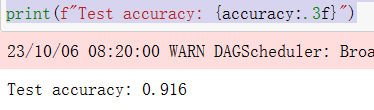
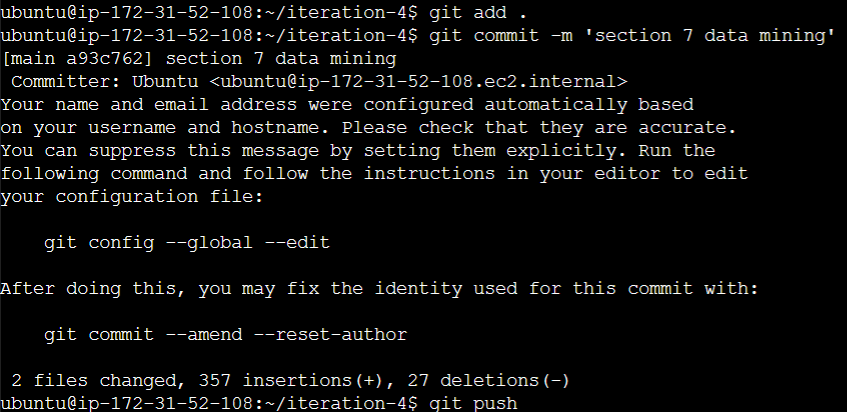


Figure The score of the model

It can be seen that this random forest performed very perfectly in the data mining process. The score was about 91% which means it can calculate the patterns and it can make predictions with very high accuracy. It can be a good thing, but it also has an underlying issue that this model or the data set is too simple, the volume of data set is too small or the model was overfitting. Therefore, I should be more careful when I am selecting the data from the data sets and choosing the data mining algorithms to guarantee a better and more comprehensive result.

After checking the model quality, the changes were committed and pushed to the GitHub.



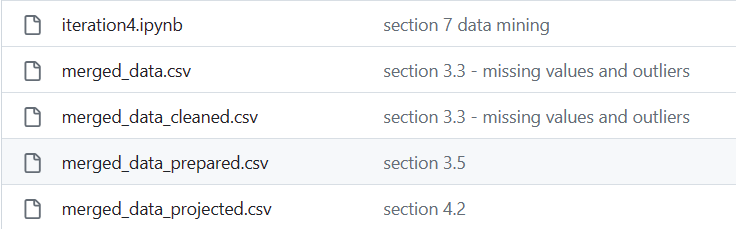


Figure Uploaded data sets in GitHub repository

### Decision making review

After choosing multiple data sets, I firstly determined my business objective according to the data sets I found, which was “discover the underlying reasons behind the high death share of the regions and predict the death risk shifting in the next coming years”, because I believed if we can make decisions to lower this percentage, there will be less people dying because of unsafe water. Secondly, to finding the potential factors of death, I looked for the data sets in the website “Our Would In Data” and picked out multiple tables which I thought were relevant to my target. Thirdly, I chose to integrated my data sets into a whole one. During this process, I filtered some variables which were either containing a large number of missing values or irrelevant to the business objectives or having duplicate functions. Fourthly, I constructed new binary data for some certain variables for the convenience of data mining process, such as “death\_share\_risks” and “income\_class”. Then, to increase the accuracy of the data mining, I reduced the rows of the data and projected the data set by balancing it. When it comes to choosing data mining method and algorithm, in order to generate an accurate result, I picked random forest in the classification method, because my target variable is a categorized variable instead of a continuous one.

## Iterations

In section 8.4. I have mentioned the score of my data mining model’s assessment, and it was shown that it had very high accuracy for testing data set. However, high accuracy also means there could be something wrong with the model or data set, or the model might be too simple.

There could be various reason which can lead to this potential problem, but I tried one more time with a different approach to test it again if my model is working properly or not.

Therefore, in the next iteration, I decide that I need to start with changing the structure of the target variable. In section 6.3., I used the “death\_share\_risk” and its indexed version “death\_share\_risk\_indexed” as target variable, which was transformed from “Share of total deaths that are from all causes attributed to unsafe water source, in both sexes aged age-standardized” in section 3.4., to use them as the target variable and input them to the Random Forest model. “death\_share\_risk” and “death\_share\_risk\_indexed” are categorized data type, but Random Forest also supported continuous data type. Therefore, in the next iteration, I will use “Share of total deaths that are from all causes attributed to unsafe water source, in both sexes aged age-standardized” as target variable to run the model.

I will still use the data set which has been projected from section 4.2., but this time the input variable will all be in numeric and continuous data type. Furthermore, it is very important that the algorithm I need to use has to be Random Forest Regression. Therefore, I need to import “RandomForestRegressor” from “pyspark.ml.regression”. For the input columns, I removed “death\_share\_risk” and “income\_level”, because both of them are not continuous data type.



Figure Setting the input variables

The result of the Random Forest feature importance is shown below.

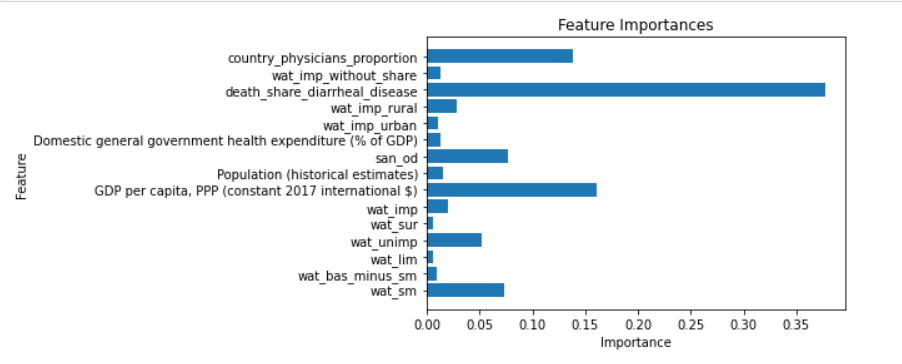


Figure Feature importance of new iteration

From the result of the model, we can find the result is different from the one we have done before. People who died from diarrheal diseases still accounts for the most part of the importance, but the second important factor become the GDP per capita of the government.

It is also necessary to examine the quality of this new model, and it showed 41.7% as the score. For Root Mean Squared Error, the smaller it is, the more performance the model has (Frost, n.d.), which is opposite to the classification model. Compared with iteration4, there is a gap between two results, and I need to find out which one is the better way to predict the future and achieve my data mining objective. In addition, it is not enough for me to figure out what is really affecting the death share caused by the unsafe water because there are still plenty of constraints for this data mining process, such as lack of data and insufficient data volume. I need to keep working on finding more data sets to support my results from the last iterations.

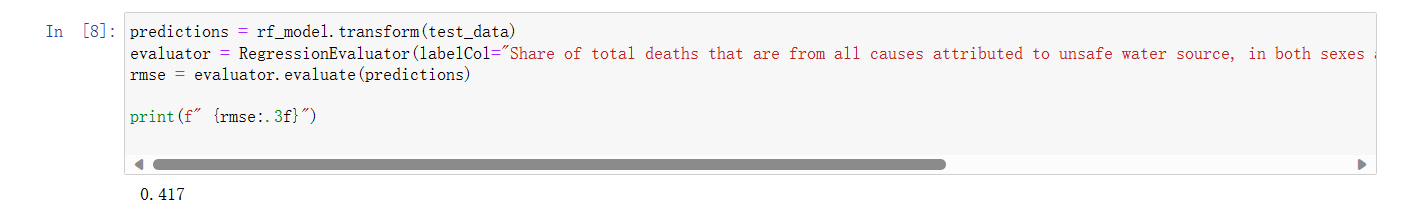
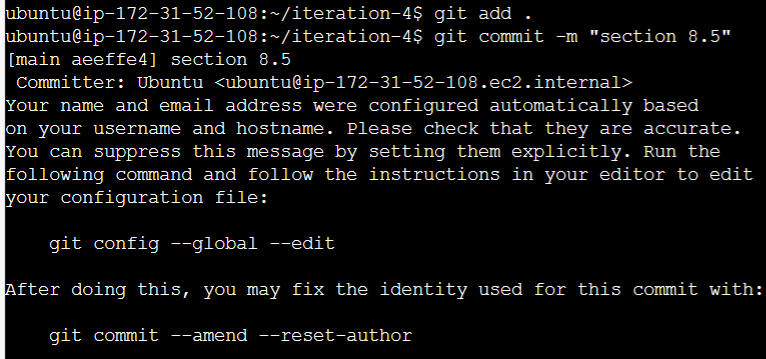


Figure Score of the model

The changes will be pushed to the GitHub.



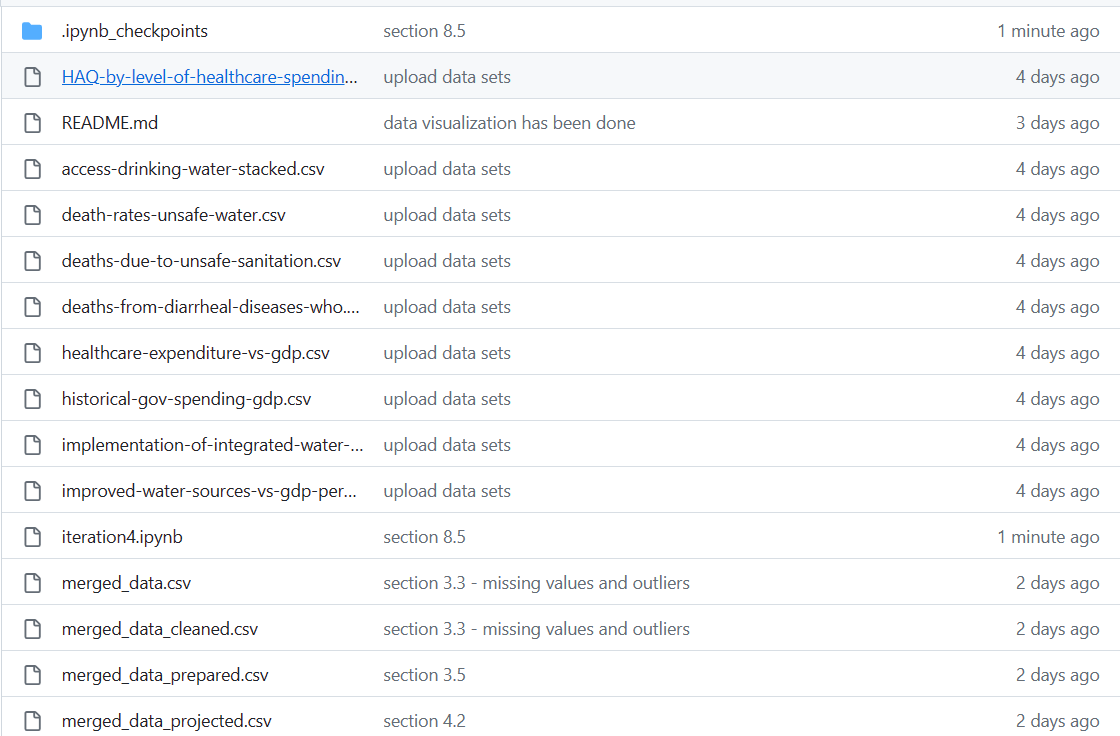


Figure Uploaded data sets in GitHub repository

# Reference List

Barr S. (2022). Are Percentages Always Better Than Counts to Measure Performance? Staceybarr. <https://www.staceybarr.com/measure-up/are-percentages-always-better-than-counts-to-measure-performance/>

Bisht A. (2021). ML | Classification vs Clustering. Geeksforgeeks. <https://www.geeks-forgeeks.org/ml-classification-vs-clustering/>

Centers for Disease Control and Prevention (2022). Disease Impact of Unsafe Water. CDC. <https://www.cdc.gov/healthywater/global/disease-impact-of-unsafe-water.html>

de Amorim, R. C., & Hennig, C. (2015). Recovering the number of clusters in data sets with noise features using feature rescaling factors. Information Sciences, 324, 126–145. https://doi.org/10.1016/j.ins.2015.06.039

Delua, J. (2021). Supervised vs. Unsupervised Learning: What’s the Difference? IBM Blog. [Supervised vs. Unsupervised Learning: What’s the Difference? - IBM Blog](https://www.ibm.com/blog/supervised-vs-unsupervised-learning/)

Department of Economic and Social Affairs Sustainable Development (2023). Ensure availability and sustainable management of water and sanitation for all. United Nations. <https://sdgs.un.org/goals/goal6>

Gholamy, A., Kreinovich, V., & Kosheleva, O. (2018). Why 70/30 or 80/20 Relation Between Training and Testing Sets: A Pedagogical Explanation. Technical Report: UTEP-CS-18-09. The University of Texas at El Paso. <https://scholarworks.utep.edu/cgi/viewcontent.cgi?article=2202&context=cs_techrep>

Gillis, A. S. (2022). What is data splitting and why is it important? TechTarget. <https://www.techtarget.com/searchenterpriseai/definition/data-splitting>

Gupta, S., & Gupta, A. (2019). Dealing with Noise Problem in Machine Learning Data-sets: A Systematic Review. Procedia Computer Science, 161, 466–474. https://doi.org/10.1016/j.procs.2019.11.146

IBM (2023). IBM Cloud Pak for Data 4.7.x. IBM. [https://www.ibm.com/docs/en/-c loud-paks/cp-data/4.7.x?topic=modeling-c50-node](https://www.ibm.com/docs/en/-cloud-paks/cp-data/4.7.x?topic=modeling-c50-node)

IBM (2023). What is the k-nearest neighbors algorithm? IBM. <https://www.ibm.com/topics/knn>

International Organization for Standardization (2023). Country Codes Collection. ISO. <https://www.iso.org/obp/ui/#search/code/>

Frost, J. (n.d.). Root Mean Square Error (RMSE). Statistics By Jim. <https://statisticsbyjim.com/regression/root-mean-square-error-rmse/>

Project Jupyter. (n.d.). Jupyter. <https://jupyter.org/>

Hays, A. (2023). Empirical Rule: Definition, Formula, Example, How It's Used. Investopedia. https://www.investopedia.com/terms/e/empirical-rule.asp

Ritchie, H. & Roser M. (2021). Clean Water and Sanitation. OurWorldInData. [https-://ourworldindata.org/clean-water-sanitation](https://ourworldindata.org/clean-water-sanitation).

Roser, M. (n.d.). About. Our World In Data. <https://ourworldindata.org/about>

Sachin (2023). What is the 80/20 Rule inMachine Learning? TechMediaToday. <https://www.techmediatoday.com/what-is-the-80-20-rule-in-machine-learning/>

Sharma, R. (2022). Classification in Data Mining Explained: Types, Classifiers & Applications [2023]. upGrad. <https://www.upgrad.com/blog/classification-in-data-mining/>

Sharma, R. (2022). Cluster Analysis in Data Mining: Applications, Methods & Requirements. upGrad. [Cluster Analysis in Data Mining: Applications, Methods & Requirements | upGrad blog](https://www.upgrad.com/blog/cluster-analysis-data-mining/)

Steven A. Greenlaw, Timothy Taylor, Eric Dodge, Cynthia Gamez, Andres Jauregui, Diane Keenan, Dan MacDonald, Amyaz Moledina, Craig Richardson, David Shapiro, & Ralph Sonenshine. (2017). Principles of Macroeconomics 2e (2nd ed.). Openstax.

World Health Organization (2023). Water supply, sanitation and hygiene monitoring. Would Heath Organization. <https://www.who.int/teams/environment-climate-change-and-health/water-sanitation-and-health/monitoring-and-evidence/wash-monitoring#:~:text=In%20order%20to%20meet%20the%20criteria%20for%20a,the%20water%20supplied%20should%20be%20free%20from%20contamination>.

World Health Organization (2023). Drinking water. Would Heath Organization.

<https://www.who.int/news-room/fact-sheets/detail/drinking-water>

I acknowledge that the submitted work is my own original work in accordance with the University of Auckland guidelines and policies on academic integrity and copyright. (See: https://www.auckland.ac.nz/en/students/forms-policies-and-guidelines/student-policies-and-guidelines/academic-integrity-copyright.htmlLinks to an external site.).

I also acknowledge that I have appropriate permission to use the data that I have utilised in this project. (For example, if the data belongs to an organisation and the data has not been published in the public domain then the data must be approved by the rights holder.) This includes permission to upload the data file to Canvas. The University of Auckland bears no responsibility for the student's misuse of data."