Timeline of Historical Pandemics

Team members:

- Nandhini Nallathambi
- Patricia Colomer
- Razvan Radu
- Siyuan Liang
- Salma Abdirahman

1. Introduction

A pandemic is a global disease outbreak. It differs from an outbreak or epidemic because it affects a wider geographical area, infects a greater number of people than an epidemic and is often caused by a new virus or a strain of virus that has not circulated among people for a long time. Humans usually have little to no immunity against it. Pandemic causes much higher numbers of deaths than epidemics and often create social disruption, economic loss, and general hardship.

Throughout history there have been a number of pandemic diseases that have severely affected the lives of citizens worldwide.

In this project, we analyse the worst pandemics that have occurred through history.

2. Data Cleaning

a. Extraction

The dataset is comprised of 9 CSV files from Kaggle.com, compiled by the CDC (Centres for Disease Control and Prevention), which provides a record of several major historical disease outbreaks. The files include data on aspects such as the type of disease/outbreak, the location and timing of such outbreaks, its death toll and a number of other information.

Link to dataset: https://www.kaggle.com/datasets/thedevastator/a-comprehensive-history-of-major-disease-outbrea

Files:

df_1.csv - This file contains the details of all major outbreaks with the death toll of over a million.

df_2.csv - This file contains the details of all outbreaks that have happened since the ancient times.

df_4.csv - This file contains the list of all natural disasters by the estimated number of deaths.

df_5.csv - This is a subset of df_4.

df_11.csv - This is a copy of df_5.

df_16.csv - This file contains the epidemics categorised by historical periods.

df_22.csv - This is a subset of df_16.

df_24.csv - This file contains the history of medicine.

df_25.csv - This file contains public health notes.

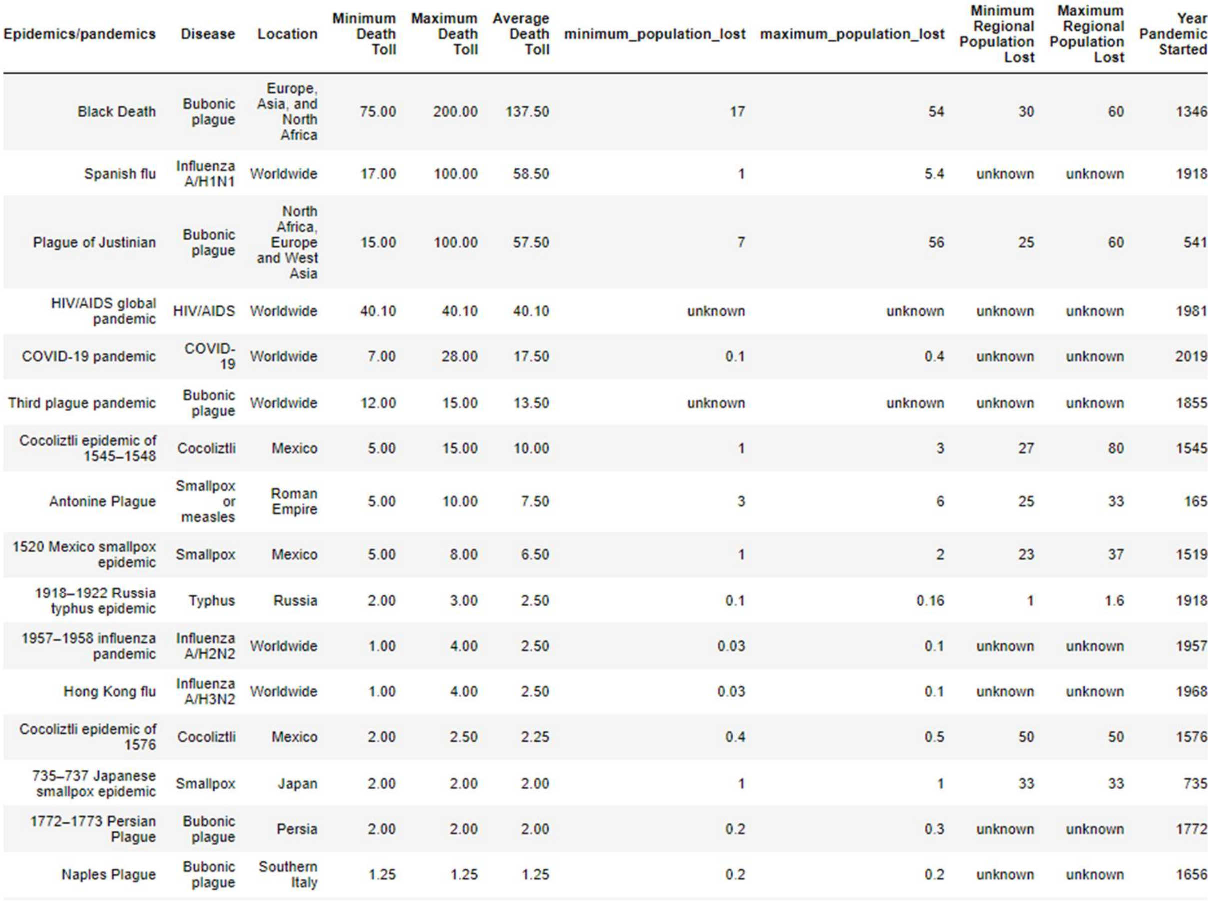
As a group we had decided to drop files that either were copies of other files or showed no relevant data. This included files: df_5, df_11.csv, df_22.csv and df_25.csv.

b. Transformation

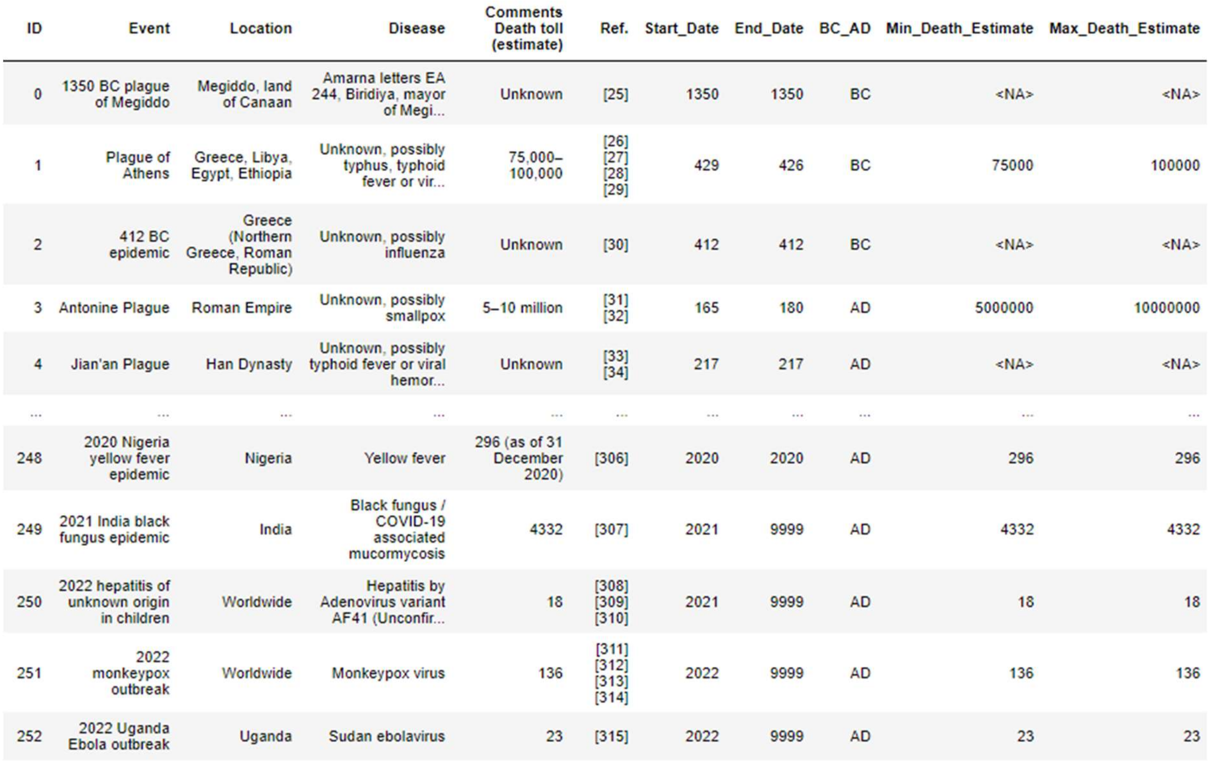
A clean dataset was created comprising the data of the relevant files. Jupyter Notebook was used for this purpose. We did the following:

- Selected relevant columns and dropped the ones that had little interest for this analysis
- Renamed the columns to give them more relevant titles
- Extracted the data from columns that listed several items so it would be more readable and easier to manipulate
- Created new columns to store the split columns
- Merged data-frames to show more comprehensive information
- Created new columns showing averages of deaths and duration of epidemics

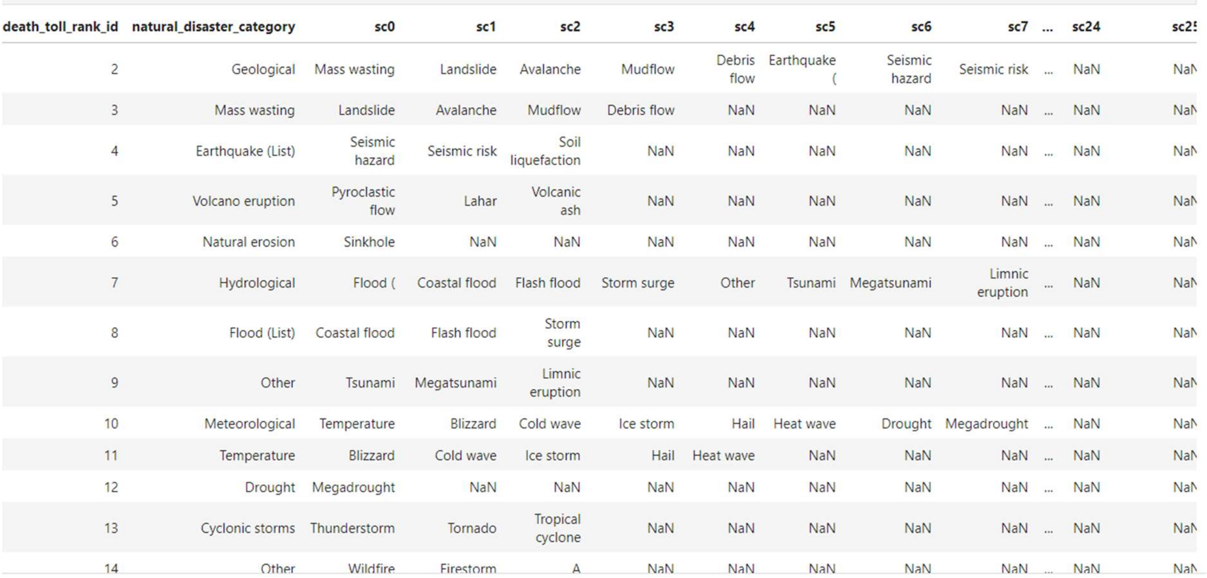
Bellow, the final data-frames are shown:



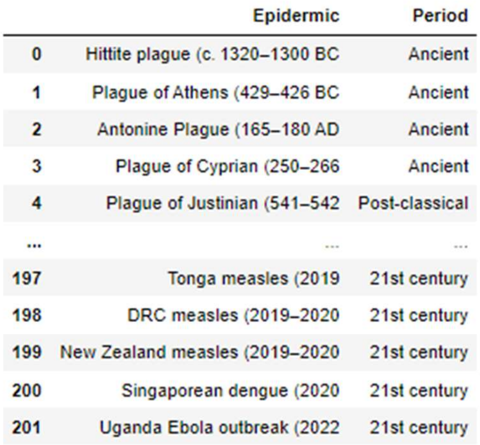
Epidemics/pandemics	Disease	Location	Minimum Death Toll	Maximum Death Toll	Average Death Toll	minimum_population_lost	maximum_population_lost	Minimum Regional Population Lost	Maximum Regional Population Lost	Year Pandemic Started
Black Death	Bubonic plague	Europe, Asia, and North Africa	75.00	200.00	137.50	17	54	30	60	1346
Spanish flu	Influenza A/H1N1	Worldwide	17.00	100.00	58.50	1	5.4	unknown	unknown	1918
Plague of Justinian	Bubonic plague	North Africa, Europe and West Asia	15.00	100.00	57.50	7	56	25	60	541
HIV/AIDS global pandemic	HIV/AIDS	Worldwide	40.10	40.10	40.10	unknown	unknown	unknown	unknown	1981
COVID-19 pandemic	COVID-19	Worldwide	7.00	28.00	17.50	0.1	0.4	unknown	unknown	2019
Third plague pandemic	Bubonic plague	Worldwide	12.00	15.00	13.50	unknown	unknown	unknown	unknown	1855
Cocoliztli epidemic of 1545–1548	Cocoliztli	Mexico	5.00	15.00	10.00	1	3	27	80	1545
Antonine Plague	Smallpox or measles	Roman Empire	5.00	10.00	7.50	3	6	25	33	165
1520 Mexico smallpox epidemic	Smallpox	Mexico	5.00	8.00	6.50	1	2	23	37	1519
1918–1922 Russia typhus epidemic	Typhus	Russia	2.00	3.00	2.50	0.1	0.16	1	1.6	1918
1957–1958 influenza pandemic	Influenza A/H2N2	Worldwide	1.00	4.00	2.50	0.03	0.1	unknown	unknown	1957
Hong Kong flu	Influenza A/H3N2	Worldwide	1.00	4.00	2.50	0.03	0.1	unknown	unknown	1968
Cocoliztli epidemic of 1576	Cocoliztli	Mexico	2.00	2.50	2.25	0.4	0.5	50	50	1576
735–737 Japanese smallpox epidemic	Smallpox	Japan	2.00	2.00	2.00	1	1	33	33	735
1772–1773 Persian Plague	Bubonic plague	Persia	2.00	2.00	2.00	0.2	0.3	unknown	unknown	1772
Naples Plague	Bubonic plague	Southern Italy	1.25	1.25	1.25	0.2	0.2	unknown	unknown	1656



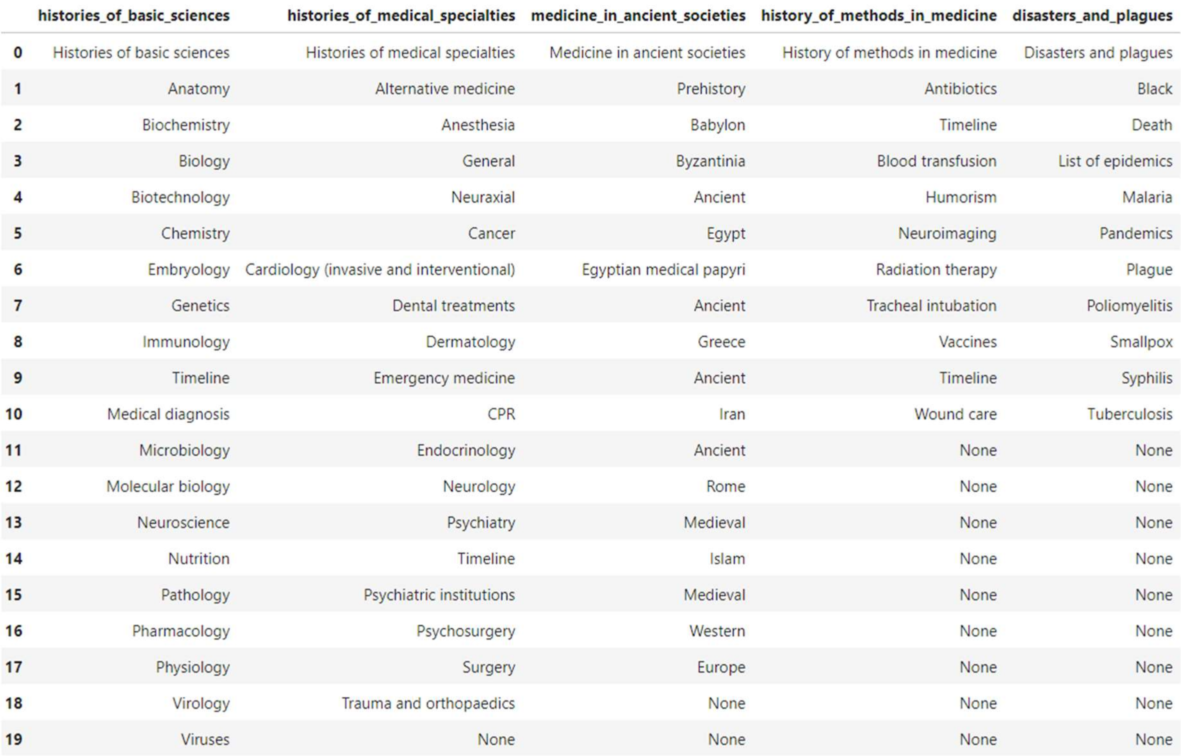
ID	Event	Location	Disease	Comments Death toll (estimate)	Ref.	Start_Date	End_Date	BC_AD	Min_Death_Estimate	Max_Death_Estimate
0	1350 BC plague of Megiddo	Megiddo, land of Canaan	Amarna letters EA 244, Biridiya, mayor of Megi...	Unknown	[25]	1350	1350	BC	<NA>	<NA>
1	Plague of Athens	Greece, Libya, Egypt, Ethiopia	Unknown, possibly typhus, typhoid fever or vir...	75,000–100,000	[26] [27] [28] [29]	429	426	BC	75000	100000
2	412 BC epidemic	Greece (Northern Greece, Roman Republic)	Unknown, possibly influenza	Unknown	[30]	412	412	BC	<NA>	<NA>
3	Antonine Plague	Roman Empire	Unknown, possibly smallpox	5–10 million	[31] [32]	165	180	AD	5000000	10000000
4	Jian'an Plague	Han Dynasty	Unknown, possibly typhoid fever or viral hemor...	Unknown	[33] [34]	217	217	AD	<NA>	<NA>
...	...	...	...	...	...	...	...	...	...	...
248	2020 Nigeria yellow fever epidemic	Nigeria	Yellow fever	296 (as of 31 December 2020)	[306]	2020	2020	AD	296	296
249	2021 India black fungus epidemic	India	Black fungus / COVID-19 associated mucormycosis	4332	[307]	2021	9999	AD	4332	4332
250	2022 hepatitis of unknown origin in children	Worldwide	Hepatitis by Adenovirus variant AF41 (Unconfir...	18	[308] [309] [310]	2021	9999	AD	18	18
251	2022 monkeypox outbreak	Worldwide	Monkeypox virus	136	[311] [312] [313] [314]	2022	9999	AD	136	136
252	2022 Uganda Ebola outbreak	Uganda	Sudan ebolavirus	23	[315]	2022	9999	AD	23	23



death_toll_rank_id	natural_disaster_category	sc0	sc1	sc2	sc3	sc4	sc5	sc6	sc7	...	sc24	sc25
2	Geological	Mass wasting	Landslide	Avalanche	Mudflow	Debris flow	Earthquake (	Seismic hazard	Seismic risk	...	NaN	NaN
3	Mass wasting	Landslide	Avalanche	Mudflow	Debris flow	NaN	NaN	NaN	NaN	...	NaN	NaN
4	Earthquake (List)	Seismic hazard	Seismic risk	Soil liquefaction	NaN	NaN	NaN	NaN	NaN	...	NaN	NaN
5	Volcano eruption	Pyroclastic flow	Lahar	Volcanic ash	NaN	NaN	NaN	NaN	NaN	...	NaN	NaN
6	Natural erosion	Sinkhole	NaN	NaN	NaN	NaN	NaN	NaN	NaN	...	NaN	NaN
7	Hydrological	Flood (	Coastal flood	Flash flood	Storm surge	Other	Tsunami	Megatsunami	Limnic eruption	...	NaN	NaN
8	Flood (List)	Coastal flood	Flash flood	Storm surge	NaN	NaN	NaN	NaN	NaN	...	NaN	NaN
9	Other	Tsunami	Megatsunami	Limnic eruption	NaN	NaN	NaN	NaN	NaN	...	NaN	NaN
10	Meteorological	Temperature	Blizzard	Cold wave	Ice storm	Hail	Heat wave	Drought	Megadrought	...	NaN	NaN
11	Temperature	Blizzard	Cold wave	Ice storm	Hail	Heat wave	NaN	NaN	NaN	...	NaN	NaN
12	Drought	Megadrought	NaN	NaN	NaN	NaN	NaN	NaN	NaN	...	NaN	NaN
13	Cyclonic storms	Thunderstorm	Tornado	Tropical cyclone	NaN	NaN	NaN	NaN	NaN	...	NaN	NaN
14	Other	Wildfire	Firestorm	A	NaN	NaN	NaN	NaN	NaN	...	NaN	NaN



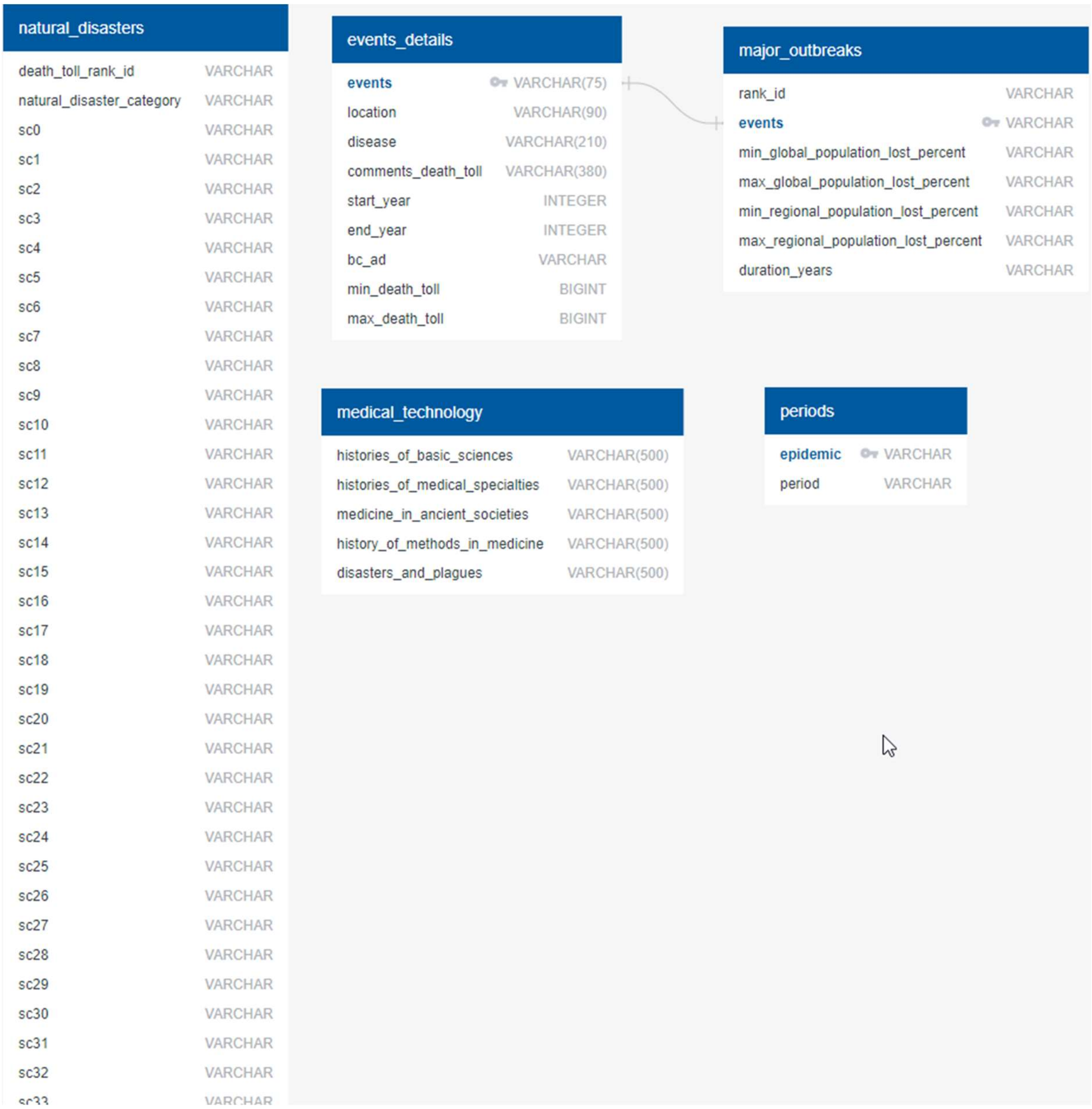
	Epidermic	Period
0	Hittite plague (c. 1320–1300 BC	Ancient
1	Plague of Athens (429–426 BC	Ancient
2	Antonine Plague (165–180 AD	Ancient
3	Plague of Cyprian (250–266	Ancient
4	Plague of Justinian (541–542	Post-classical
...	...	...
197	Tonga measles (2019	21st century
198	DRC measles (2019–2020	21st century
199	New Zealand measles (2019–2020	21st century
200	Singaporean dengue (2020	21st century
201	Uganda Ebola outbreak (2022	21st century



	histories_of_basic_sciences	histories_of_medical_specialties	medicine_in_ancient_societies	history_of_methods_in_medicine	disasters_and_plagues
0	Histories of basic sciences	Histories of medical specialties	Medicine in ancient societies	History of methods in medicine	Disasters and plagues
1	Anatomy	Alternative medicine	Prehistory	Antibiotics	Black
2	Biochemistry	Anesthesia	Babylon	Timeline	Death
3	Biology	General	Byzantinia	Blood transfusion	List of epidemics
4	Biotechnology	Neuraxial	Ancient	Humorism	Malaria
5	Chemistry	Cancer	Egypt	Neuroimaging	Pandemics
6	Embryology	Cardiology (invasive and interventional)	Egyptian medical papyri	Radiation therapy	Plague
7	Genetics	Dental treatments	Ancient	Tracheal intubation	Poliomyelitis
8	Immunology	Dermatology	Greece	Vaccines	Smallpox
9	Timeline	Emergency medicine	Ancient	Timeline	Syphilis
10	Medical diagnosis	CPR	Iran	Wound care	Tuberculosis
11	Microbiology	Endocrinology	Ancient	None	None
12	Molecular biology	Neurology	Rome	None	None
13	Neuroscience	Psychiatry	Medieval	None	None
14	Nutrition	Timeline	Islam	None	None
15	Pathology	Psychiatric institutions	Medieval	None	None
16	Pharmacology	Psychosurgery	Western	None	None
17	Physiology	Surgery	Europe	None	None
18	Virology	Trauma and orthopaedics	None	None	None
19	Viruses	None	None	None	None

3. Loading data to database

We created the following ERD to represent the schema of the database.



natural_disasters	
death_toll_rank_id	VARCHAR
natural_disaster_category	VARCHAR
sc0	VARCHAR
sc1	VARCHAR
sc2	VARCHAR
sc3	VARCHAR
sc4	VARCHAR
sc5	VARCHAR
sc6	VARCHAR
sc7	VARCHAR
sc8	VARCHAR
sc9	VARCHAR
sc10	VARCHAR
sc11	VARCHAR
sc12	VARCHAR
sc13	VARCHAR
sc14	VARCHAR
sc15	VARCHAR
sc16	VARCHAR
sc17	VARCHAR
sc18	VARCHAR
sc19	VARCHAR
sc20	VARCHAR
sc21	VARCHAR
sc22	VARCHAR
sc23	VARCHAR
sc24	VARCHAR
sc25	VARCHAR
sc26	VARCHAR
sc27	VARCHAR
sc28	VARCHAR
sc29	VARCHAR
sc30	VARCHAR
sc31	VARCHAR
sc32	VARCHAR
sc33	VARCHAR

events_details	
events	VARCHAR(75)
location	VARCHAR(90)
disease	VARCHAR(210)
comments_death_toll	VARCHAR(380)
start_year	INTEGER
end_year	INTEGER
bc_ad	VARCHAR
min_death_toll	BIGINT
max_death_toll	BIGINT

major_outbreaks	
rank_id	VARCHAR
events	VARCHAR
min_global_population_lost_percent	VARCHAR
max_global_population_lost_percent	VARCHAR
min_regional_population_lost_percent	VARCHAR
max_regional_population_lost_percent	VARCHAR
duration_years	VARCHAR

medical_technology	
histories_of_basic_sciences	VARCHAR(500)
histories_of_medical_specialties	VARCHAR(500)
medicine_in_ancient_societies	VARCHAR(500)
history_of_methods_in_medicine	VARCHAR(500)
disasters_and_plagues	VARCHAR(500)

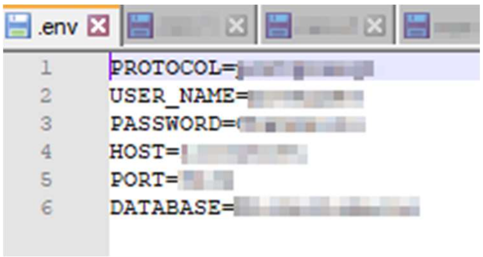
periods	
epidemic	VARCHAR
period	VARCHAR

After the clean up, we understood that the files df_2 (all outbreaks) and df_1 (major_outbreaks) had data in common and df_1 was a subset of df_2. So we created the tables **events_details** and **major_outbreaks** respectively, with events as the primary key. As events_details had the columns for location, disease, min and max death toll, start and end years, these were ignored in the major_outbreaks. We created a foreign key on events with the events_details table.

We had initially planned for **periods** table (file df_16) to be linked with the events_details and major_outbreaks tables. However, the data in the periods table was insufficient and so we made it an independent table along with **natural_disasters**(file df_4) and **medical_technologies**(file df_24) tables.

To upload the data, we made a connection to the PostgreSQL from Python using SQLAlchemy. We created a .env with the variables required for the connection string. The data from the file wass then loaded using the load_dotenv() function. We added the .env file to the .gitignore file, so that it is not uploaded to GitHub.

The .env file is as follows:



```
PROTOCOL=[redacted]
USER_NAME=[redacted]
PASSWORD=[redacted]
HOST=[redacted]
PORT=[redacted]
DATABASE=[redacted]
```

After the connection was made, we uploaded the files using to_sql() function.

The cleaned up data were uploaded into the tables under pandemics_db in PostgreSQL.

4. Conclusions

With the recent experience of COVID-19, it is clear that pandemics are not a thing of the past. Health authorities say it’s not a matter of IF a new pandemic will happen, but WHEN.

This was a major contributing factor when deciding what data we would pick and what we would base our project around. As we wanted to look into historical pandemics which much like COVID-19 did to us, greatly affected the lives of those before us.

This analysis can be used to predict future disease outbreaks by identifying patterns and trends in past outbreaks and also to develop better strategies to respond to the event. Analysing such data and reflecting on the past would be particularly useful in the in the disease surveillance and prevention by not only the World Health Organisation but that of governmental organisations such as Public Health England.