ASSIGNMENT: DEFINING A DATA-DRIVEN BUSINESS CASE

STUDENT: ANASTASIA HASWANI

COURSE: CEBD 1150

INSTRUCTOR: ADRIÀN GONSÀLEZ SÀNCHEZ

EXPLORING AND ANALYZING COVID-19 CONSIDERING VARIOUS FACTORS IN ORDER TO IMPROVE EPIDEMIOLOGICAL TECHNOLOGIES AND PROTOCOLS

REASONING: Current situation that looks almost unreal shows us that unfortunately there is a lot of room for improvement in the epidemical protocols all over the world. Nevertheless, we can use it to learn how political decisions, social behavior, and other factors influence the dynamics of the pandemic. It will be key to new solutions to protect us in the future.

VALUE: The importance of this project is clear. Machine learning will permit bringing the epidemiological science to the next level. The wave of new ideas, IT applications, medical appliances, and protocols will appear. Each country could participate in this initiative as well as benefit of it.

OBJECTIVE: Using various datasets from all over the world explore and evaluate the influence of different factors on the dynamics of new cases, the number of deaths, and other epidemiological characteristics. Based on the obtained solutions, create a list of ideas and recommendations that will greatly ease the difficulties with future epidemics, and bring the epidemiological public policies, and epidemiology in general to a new level.

OUTCOME:

- Official documentation with the specific directions of an upgrade of the national epidemiological preparation.
- A list of guidelines and policies to be performed in case of a new epidemic.
- APIs to be used by the government, medical stuff, and population.

DATA: The main data is taken from the Humanitarian Data Exchange. https://data.humdata.org/dataset/novel-coronavirus-2019-ncov-cases. Novel Corona Virus (COVID-19) epidemiological data since 22 January 2020. The data is compiled by the Johns Hopkins University Center for Systems Science and Engineering (JHU CCSE) from various sources including the World Health Organization (WHO), DXY.cn. Pneumonia. 2020, BNO News, National Health Commission of the People's Republic of China (NHC), China CDC (CCDC), Hong Kong Department of Health, Macau Government, Taiwan CDC, US CDC,

Government of Canada, Australia Government Department of Health, European Centre for Disease Prevention and Control (ECDC), Ministry of Health Singapore (MOH).

Fields available in the data include Province/State, Country/Region, Last Update, Confirmed, Suspected, Recovered, Deaths. On 23/03/2020, a new data structure was released. The current resources for the latest time series data are:

- time_series_covid19_confirmed_global_narrow.csv
- time_series_covid19_deaths_global_narrow.csv
- time_series_covid19_recovered_global_narrow.csv

Each of the three datasets contains the number of cases for every day in each country. Below are the fields:

Province/State
Country/Region
Lat
Long
Date
Value
ISO 3166-1 Alpha 3-Codes
Region Code
Sub-region Code
Intermediate Region Code

Here are the first 5 rows of the confirmed cases. As you can see, the data has been downloaded as of April 20, 2020.

	Province/State	Country/Region	Lat	Long	Date	Value	ISO 3166-1 Alpha 3-Codes	Region Code	Sub-region Code	Intermediate Region Code
1	NaN	Afghanistan	33.0	65.0	2020-04-20	1026	AFG	142	34	NaN
2	NaN	Afghanistan	33.0	65.0	2020-04-19	996	AFG	142	34	NaN
3	NaN	Afghanistan	33.0	65.0	2020-04-18	933	AFG	142	34	NaN
4	NaN	Afghanistan	33.0	65.0	2020-04-17	906	AFG	142	34	NaN
5	NaN	Afghanistan	33.0	65.0	2020-04-16	840	AFG	142	34	NaN

Potentially for our project, all the fields can be used as future analyses will be various and complex. On top of that, a lot of external data will have to be found and adapted to our needs. Demographics, location datasets, social, political, educational, and weather data – all of them and more will be needed to help to perform such a high scale analytics.

EDA: In order to better understand our basic data, to get general information and develop the principal directions for our future studies, we conducted a small preliminary analysis[1].

- Number of countries in datasets: 185
- Time period: from 2020-01-22 to 2020-04-20
- Total number of cases: 2,472,258
- Total number of deaths: 169,985

[1] You can view the code here

- Total number of recovered: 645,738
- First 5 countries with the highest number of cases:

US	784326
Spain	200210
Italy	181228
France	156480
Germany	147065

- First 5 countries with the highest number of deaths:

US	42094
Italy	24114
Spain	20852
France	20292
United Kingdom	16550

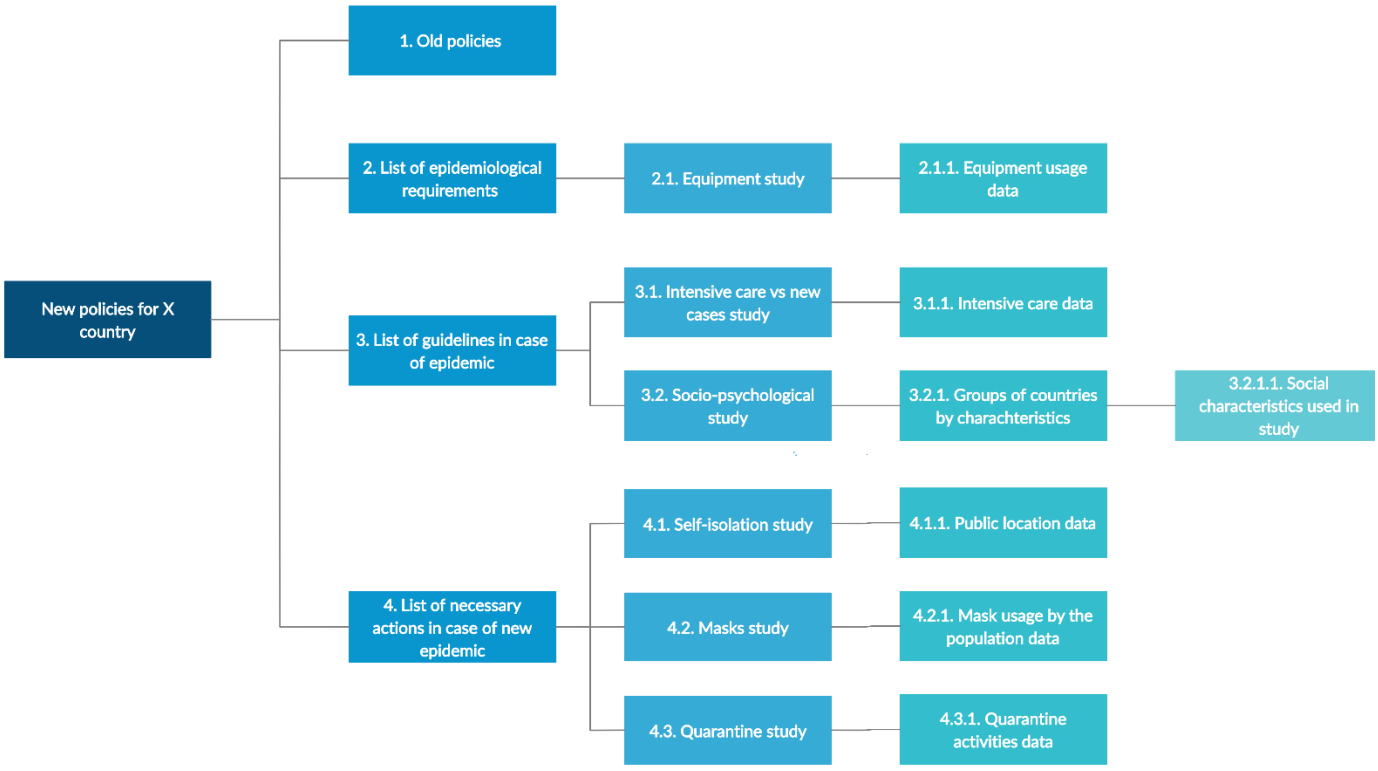
- First 20 countries with the highest treatment success rate (highest difference between percentage of recovered people and mortality):

Country	Confirmed	Recovered	Deaths	Mortality	Cured_Rate	Success
Diamond Princess	712	644	13	1.83	90.45	88.62
Cambodia	122	107	0	0	87.7	87.7
China	83817	77745	4636	5.53	92.76	87.22
Saint Lucia	15	13	0	0	86.67	86.67
Brunei	138	116	1	0.72	84.06	83.33
Vietnam	268	214	0	0	79.85	79.85
Iceland	1773	1362	10	0.56	76.82	76.25
Korea, South	10674	8114	236	2.21	76.02	73.81
Mauritania	7	6	1	14.29	85.71	71.43
Thailand	2792	1999	47	1.68	71.6	69.91
Austria	14795	10631	470	3.18	71.86	68.68
Uganda	56	38	0	0	67.86	67.86
New Zealand	1440	974	12	0.83	67.64	66.81
Liechtenstein	81	55	1	1.23	67.9	66.67
Mauritius	328	224	9	2.74	68.29	65.55
Iran	83505	59273	5209	6.24	70.98	64.74
Jordan	425	282	7	1.65	66.35	64.71
Australia	6547	4124	67	1.02	62.99	61.97
Switzerland	27944	18600	1429	5.11	66.56	61.45
Iraq	1574	1043	82	5.21	66.26	61.05

SUBPROJECTS: Our project must be divided into subprojects. Each subproject constitutes a whole direction, a separate subject of study with its subtasks, timeline and role distribution. Below are a few possible subprojects:

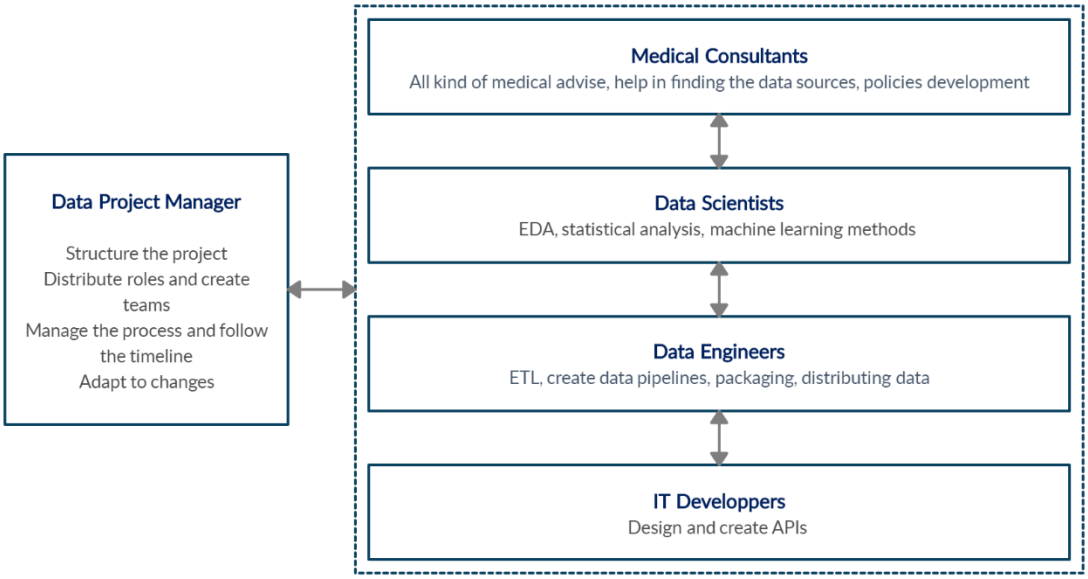
- How the level of compliance with self-isolation rules in different countries affects the spread of the virus?
- How the timing of quarantine actions affects the rate of spread of the virus and the number of cases?
- How does the widespread wearing of masks in public places affect the number of patients?
- How many masks, ventilators and other important equipment is going to be needed as a reserve for a specific country?
- Analysis of the interaction between the number of confirmed cases and the number of intensive care patients in order to find out a realistic picture of the epidemiological situation in the country.
- Analysis of the influence of socio-psychological characteristics of the population on the spread of the virus.

WBS: In order to organize the workflow of each subproject we would need to create a WBS for each project. Below is a draft of WBS for creating new policies for a specific country. It only includes the analysis we mentioned above as an example.



Based on the diagram above we can now clearly see the main tasks to be completed as well as their sequence. Each task will be divided into smaller tasks. The next step would be developing the timeline.

PROJECT STRUCTURE: To perform such a big project we need a number of people. The distribution of responsibilities would depend on the complexity of a specific subproject. Now, we can highlight the general roles, their interrelation and responsibilities:



Let's show an example of role distribution in specific situation. For this we will take one of the subprojects from WBS above:

We will now specify few subtasks for each of the point and assign the roles:

2.1.1.	Equipment usage data	
	Find the data resources	Medical Consultant Data Scientist
	ETL	Data Engineer
2.1.	Equipment Study	
	Data wrangling	Data Scientist

	What are the most important and needed equipment that we need to have in reserve?	Medical Consultant
	The trend of equipment usage during covid-19 (personal protection, ventilators)	Data Scientist
	Number of masks and other equipment used for each state	Data Scientist
	What is the percentage of equipment that local production can provide?	Data Scientist
2	List of epidemiological requirements	
	Number of equipment needed as reserve in each state	Medical Consultant Data Scientist

PRODUCT BACKLOG: There are a huge number of possible derivative products that can be created based on completed research or from the ideas born in the process of work.

The first and largest product may be a powerful analytical API that will analyze, visualize, and predict some information in the event of future epidemics.

Some smaller IT products may be created for the general public during a pandemic:

- A digital test linked with an application to enable the access to public places.

- An application that can determine the proximity of an infected person.

- An application that gives a signal when the distance to the person near you is less than a certain fixed number.

- The public temperature scanning checkpoints.