Milestone 1 :

Clémence Barsi, Juliette Challot, Pauline Conti.

Dataset :

We have chosen two different datasets both on cardio-vascular diseases. Additionally, we used two other datasets to process the previous ones.

The first dataset collects 70'000 records of patients with 11 features and one target (if the patient has a cardio-vascular disease). The features are age, sex, height, weight, systolic blood pressure, diastolic blood pressure, cholesterol, glucose, smoking, alcohol usage, and physical activity. The dataset values were collected at the medical examination of the patients. As ages were given in days, we had to convert the ages in years for better understanding. After some pre-processing, we observed that most of the records concerned people were above 39 years old with only 4 records of younger people. Hence, with this dataset we'll focus on people older than 39. Of those records, about half (34'979 records) had a cardio-vascular disease.

The second dataset is a compilation of mortality data from the World Health Organization. It collects data reported annually by the civil registrations of the member states, only registering medically-certified deaths (diagnosis by lay people are not included). It contains 792'103'115 records of death of which 103'526'649 are due to cardiovascular disease (~13% of all deaths). This dataset records the sex, country, year, cause, and age of death. Records span from 1988 to 2019. To process this dataset, two additional datasets were needed : one being the country name associated with its code and the second associating the cause of death to its code. We will need to see if each year and each country has enough records. Some countries are missing as they do not report to WHO and the latest years of some countries are also missing as WHO takes at least a year to proof-check the data submitted. As there was a lot of data, we had to manually merge five .csv files containing the mortality records of the tenth revision of ICD (classification of diseases, the tenth being the most recent one).

Problematic :

What can increase the probability of having a cardiovascular disease and how the proportion of death due to cardiovascular diseases has evolved throughout the years in the world?

Cardio-Vascular diseases are the first cause of death worldwide, the World Health Organisation estimates that they are responsible for about 31% of all global deaths. However, upon preprocessing and data exploration, they seem to only make up 13% of the WHO dataset that we have. It is hard to know why that is for now but we will keep this fact in mind when creating our visualizations so as to not provide incomplete insights. Using our visualizations we'd like to raise awareness towards those diseases. Indeed, we often don't realize how frequent and diverse the symptoms can be.

Hence we will also focus a part of the visualization on how some symptoms leading to those illnesses can affect our body. We'd like to identify if there are factors that are commonly observed in people suffering from those and if we can prevent them by being cautious on our smoking or alcohol consumption for example.

We also want to see if people coming from certain countries are more frequently touched by cardio-vascular diseases and if the death rate associated with these diseases varies.

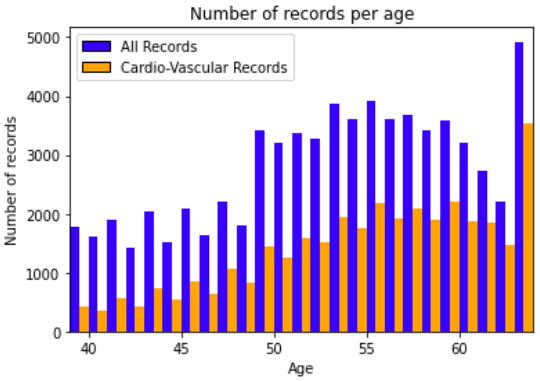
Finally we want to see if the deaths by cardio-vascular diseases have increased throughout the years and what part of population is more subject to it, through the lens of age and gender.

We are addressing our visualizations to everybody as anybody can be or has a relative that might be more subject to a cardiovascular disease. More awareness on these diseases and common symptoms can help young people be more cautious of their health meanwhile providing at risk people with partial answers and solutions on some conditions they might start to notice.

Exploratory Data Analysis :

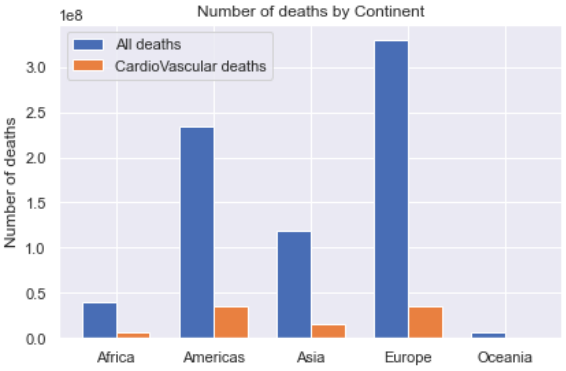
Our visualizations aim to address two axes. First, the evolution of cardiovascular diseases (CVD) throughout the years and in the world. Second, the parameters that might influence the occurrence of a CVD.

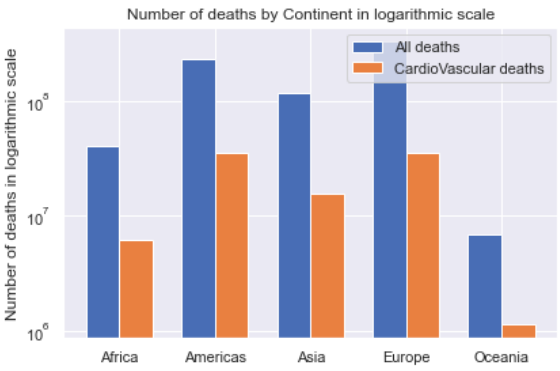
The following figure shows the number of records contained in the first dataset by age. We compare all records to the ones associated with CVD. With ages ranging from 39 to 64 years old, the two metrics follow similar trends, slight increase with the patient's age. However, it also seems that the proportion of cardio-vascular records increases with the years, going from approximately 25% around 40 years old to 75% for 63 years old patients. This is in line with what we know of most cardiovascular diseases, as age is often a significant factor in risk.



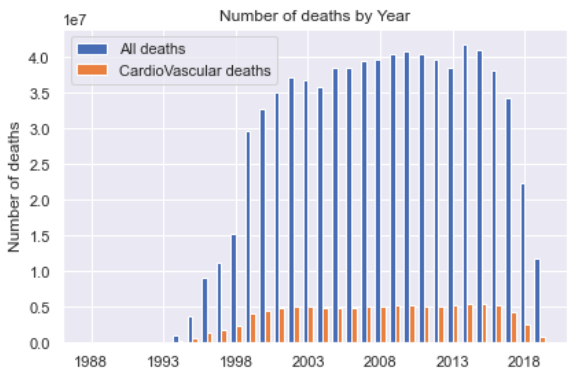
The following figures are based on the WHO Mortality database, spanning from 1988 to 2019. They represent the number of deaths from all causes and CVD, by continent. We note two things:

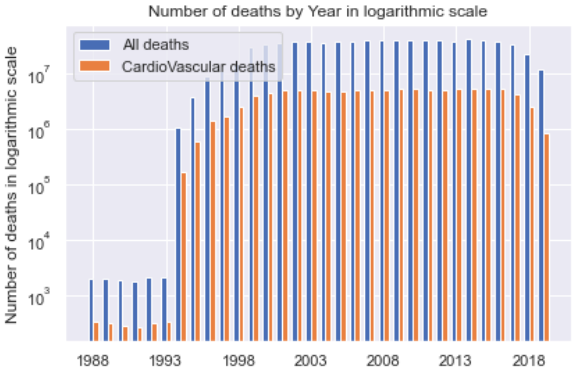
- First, we have much more records from America and Europe than from Asia, Africa and Oceania. As this does not match the actual disparities in population, Europe and America might be over-represented in the dataset. We have to take this into consideration, making sure that this skewness does not influence the visualization's message and conclusions.
- Second, the ratio of CVD deaths to all deaths seems constant in logarithmic scale, meaning that it is smaller for continents with many records than it is for less represented continents. This can be surprising as we expect a bigger percentage of deaths to be from CVD in those continents than in Africa, Asia and Oceania, but we can imagine that a wider range of causes are documented in eastern countries than in developing countries.





These last graphs, also based on the WHO data, show the number of deaths by year, in general and from CVDs. The CVD deaths seem to be constant with the years (number of deaths & proportion of overall deaths). There is a slight decrease since 2018, its origin might be interesting to investigate, along with the significant rise in total number of deaths from 1994.





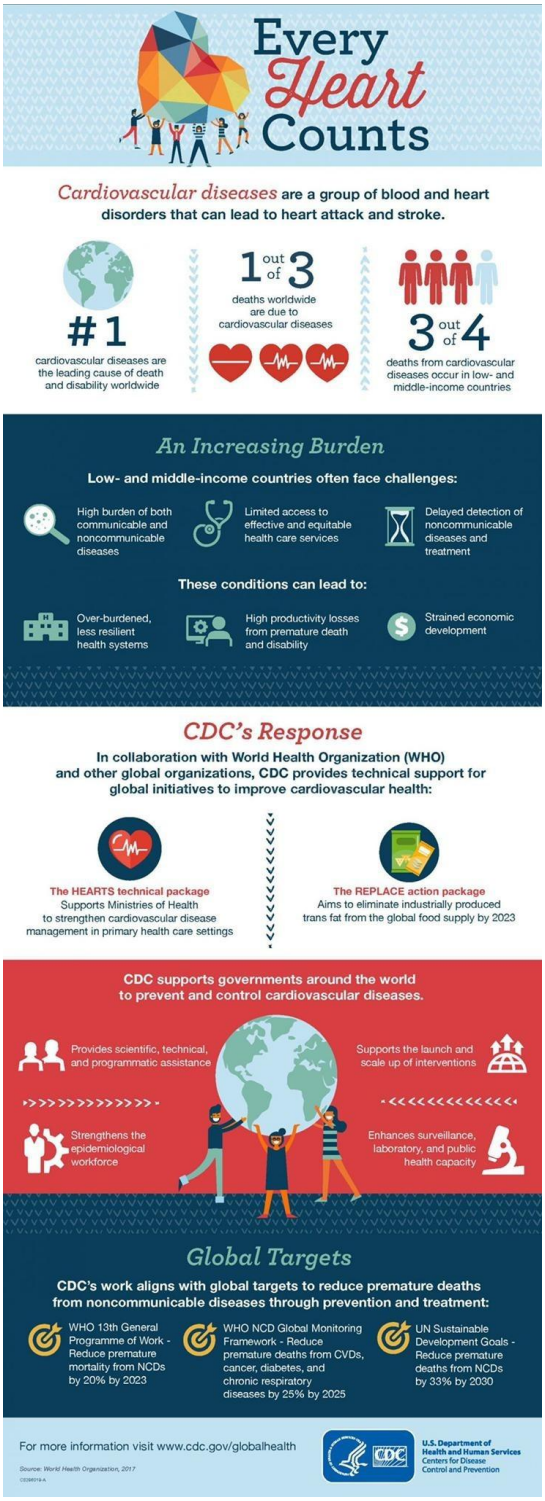
Related work :

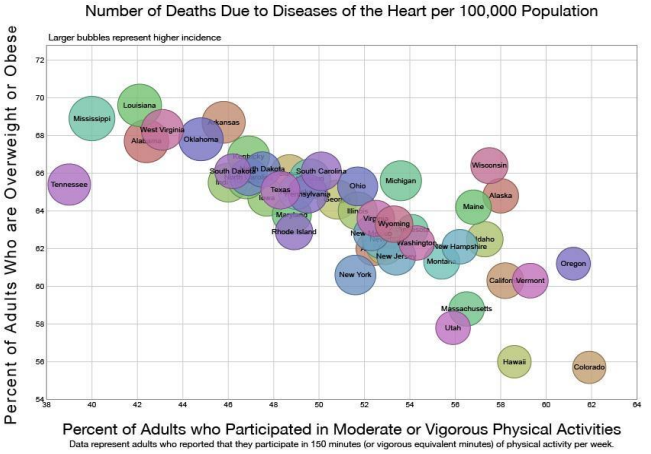
The first dataset we selected is quite recent and has been used mainly to develop machine learning algorithms. Indeed we found several scientific papers using it for early prediction of CVDs using different algorithms (Random Forest, Naïve Bayes, KNN, Logistic Regression…) to predict whether or not a patient suffers from a CVD based on symptoms and individual features such as gender, cholesterol, alcohol intake... Early prediction is key as it could drastically reduce the mortality rate of these diseases. Most studies available showed very basic visualizations to quickly illustrate the type of data used for developing their prediction algorithms but didn't use the data as an end to raise awareness and highlight the characteristic symptoms and/or profile of patients through efficient visualizations.

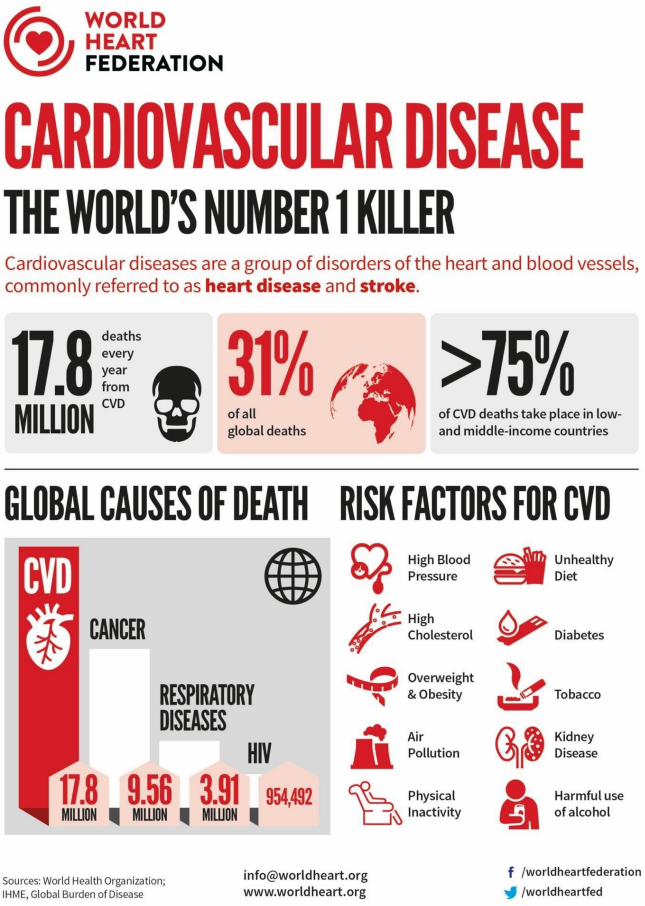
The second dataset will allow us to show parameters such as the evolution of the number of deaths caused by CVDs over the years and the local variations. Again we found really interesting papers using similar datasets but they were quite old (more than 10 years old) thus using previous versions of the dataset we chose and they didn't necessarily target CVDs as this dataset contains the mortality data for all the ICD (International Classification of Diseases) categories.

Our approach really targets awareness as well as an historical visualization of CVDs mortality, symptoms and profiles around the world. The goal is to make a concrete and user friendly representation of all the data at our disposal so that people quickly get access to this crucial information. Being able to identify symptoms and being aware of aggravating factors can be key in getting early diagnosis as well as reducing the number of CVDs altogether. Both aspects are key to lower the mortality of such conditions.

During our research, we came across a few awareness campaigns among which :







The one from the World Heart Federation is already quite informative but we would like to go further by making animated visualizations. We will also give more precise data on which risk factors are the most redundant as well as the opportunity for visitors to see which parts of the world are the most impacted by CVDs. However we will use these campaigns as reference when making our visualizations so as to give accurate information to the reader.
We haven't used any of these datasets before.

References

Link to first dataset : https://www.kaggle.com/sulianova/cardiovascular-disease-dataset
Link to the second dataset and the additional ones :
https://www.who.int/data/data-collection-tools/who-mortality-database
Papers on CVDs :
https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=8550857
http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.206.3899&rep=rep1&type=pdf
https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=4493524
https://www.researchgate.net/profile/Vikas-Chaurasia-2/publication/259474824_Early_Prediction_of_Heart_Diseases_Using_Data_Mining_Techniques/links/0c96052bfd32153b24000000/Early-Prediction-of-Heart-Diseases-Using-Data-Mining-Techniques.pdf
https://www.aimsjournal.org/uploads/78/6557_pdf.pdf
Papers using previous versions of the WHO mortality dataset :
https://www.ahajournals.org/doi/pdf/10.1161/CIRCULATIONAHA.115.018931
https://gh.bmj.com/content/bmjgh/2/2/e000298.full.pdf
Mortality from ischaemic heart disease by country, region, and age: Statistics from World Health Organisation and United Nations
Every Heart Counts Campaign: https://www.cdc.gov/globalhealth/infographics/every-heart-counts.html
Top right graph:
https://digitalsplashmedia.com/2014/03/visualizing-heart-disease-correlations-in-the-us-and-eu/
World Heart Federation Campaign:
https://www.world-heart-federation.org/resources/cardiovascular-disease-infographic/