Milestone 1 report

1.1 Exploratory Data Analysis

First, we looked at the attributes of each csv file. Next, we listed how much data is missing in each column of each csv file. We integrated the data and drew diagrams of some key information. They are: Heat map of the world; Ranking of countries with the most cases; Ranking of countries with the highest incidence rate; Ranking of countries with the highest mortality rate; Number of male and female cases in various age groups; Pie chart which compares female cases and male cases; etc.

1.2 Data cleaning and Imputing missing values

In this part, we unify the age format that appears in the data into a specific value. At the same time, we convert the date information that appears in the data into a standard date format. We also impute those missing values in age column with mean value of the age.

1.3 Dealing with outliers

In the dataset, there are outliers in the longitude and latitude columns. Our processing method is to remove the values with latitude greater than 90 and less than -90, and at the same time remove the values with longitude greater than 180 and less than -180. At the same time, we remove rows which age is smaller than 0 or larger than 110.

1.4 Transformation

First, we extract the data with the value of US in all country columns from the data set, and store it in a new dataframe, and then separate the data for each state to calculate the relevant data for each state. Then create a new dataframe, save the data of each state after the calculation in it.

1.5 Joining the cases and location dataset

We use "province, country" as the standard to integrate the two sets of data. Delete the meaningless columns and delete completely duplicate rows. And recalculate the data of each attribute.

1.6 Outcome labels

These labels record the medical treatment status of these patients, respectively: hospitalized, non-hospitalized, deceased and recovered. The type of the data mining task is classification.