文本

中度可信度描述已自动生成

**Exploring the Relationship Between Hospitalization and Mortality:**

**A Cross-National Comparative Study Based on COVID-19 Data**

**Lecturer: Brenda Mullally**

**Student: Haopeng Liang**



**Project Report**

Data Science in Practice

# Introduction

## Background

Since the outbreak of COVID-19 at the end of 2019, medical systems around the world have faced unprecedented challenges. Countries' medical resources, hospital bed availability, and medical staff's response capabilities greatly affect the mortality and recovery rates of the epidemic. Because COVID-19 can cause severe respiratory symptoms, a large number of patients require hospitalization, including treatment in general wards and intensive care units. Therefore, hospital processing capacity has become an important indicator for evaluating the response of public health systems in various countries to the epidemic.

## Purpose

The main objective of this study was to analyze the relationship between COVID-19 hospitalization rates and mortality as a window to observe and evaluate the handling capacity of hospitals in different countries during the outbreak. A deeper understanding of the relationship between these two indicators can provide insight into how countries' health systems are actually performing at the height of a pandemic. In addition, this research may also help us understand how to allocate and use healthcare resources more effectively during similar public health crises in the future.

## Research Questions and Hypotheses

This study will explore the following research questions:

* What is the relationship between hospitalization rates and mortality during the COVID-19 outbreak?
* Does this relationship vary significantly between countries, and do these differences reflect the capacity of hospitals in each country?

Based on these questions, this study proposes the following hypotheses:

* **Hypothesis 1**: There is a positive correlation between hospitalization rates and mortality, that is, countries with higher hospitalization rates also have relatively higher mortality rates.
* **Hypothesis 2**: There is a negative correlation between the number of hospital beds and mortality, i.e. countries with more beds have lower COVID-19 mortality rates.

## Data Sources and Variables

The analysis in this study is based on two main datasets selected from ***Our World in Data***. The first dataset (***covid-hospitaliztions.csv***) includes daily hospitalizations and ICU use for COVID-19 for each country, while the second dataset (***owid-covid-data.csv***) provides detailed statistics on COVID-19 cases and deaths from the beginning of the outbreak to the present day. Specific data include country name, ISO code, observation date, cumulative and new cases, cumulative and new deaths, number of patients in hospital, number of ICU patients, and number of hospital beds per 1,000 people. In addition, the dataset also provides demographic information and health care resources for each country, such as the number of hospital beds and the number of health care personnel, which provides important background information for this study.

Key variables include:

* **iso\_code**: The ISO code for the country, which identifies the country analysed.
* **entity**: Full name of country.
* **indicator** & **value**: Indicates the flow of people in hospitals and ICU wards in hospitals.
* **continent**: Indicates the continent in which the country is located.
* **date**: Date of data recording.
* **total\_deaths**: Cumulative number of deaths.
* **new\_deaths**: New number of deaths.
* **hosp\_patients**: Number of current inpatients.
* **icu\_patients**: Number of patients in the ICU.
* **hospital\_beds\_per\_thousand**: Hospital beds per thousand people.
* **population**: Indicates the total population of the country.

These variables will be used to construct analytical models to explore the relationship between hospitalization rates and mortality and their reflection to hospital handling capacity.

## Significance of the Study

Through comprehensive analysis of these data, this study hopes to reveal the potential association between hospitalization rates and mortality and explore the role of medical resource allocation in controlling COVID-19 mortality. This research can not only provide decision support for public health policymakers, but also enhance public awareness of the importance of epidemic response and resource allocation.

# Data Management

## Data Merging and Cleaning

First I select all the columns I need to use in both csv data files and write them all in an array, then selectively read only those columns in the csv file, reducing the performance loss of the computer:

1. hosp\_cols = ['entity', 'date', 'iso\_code', 'indicator', 'value']
2. death\_cols = ['date', 'iso\_code', 'continent', 'total\_deaths', 'new\_deaths', 'hosp\_patients',
3. 'icu\_patients', 'hospital\_beds\_per\_thousand', 'population']
4. hosp\_data = pd.read\_csv(hosp\_data\_path, usecols=hosp\_cols)
5. death\_data = pd.read\_csv(death\_data\_path, usecols=death\_cols)

Then delete row data with null values, because null data will greatly affect the analysis, and it is completely unnecessary to use filling. Therefore, after deleting the data, merge the two data (hosp\_data and death\_data) into one New csv file, which facilitates subsequent re-reading and visual observation.

1. merged\_data.to\_csv(clean\_merged\_data\_path, index=False)
2. *# Consolidation of data*
3. merged\_data = pd.merge(hosp\_data, death\_data, on=['date', 'iso\_code'], how='inner')
4. *# Handling Missing Values - Here we choose to remove rows containing missing values*
5. merged\_data.dropna(inplace=True)
6. return merged\_data

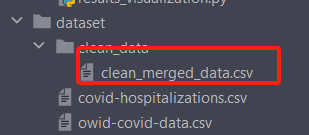


FIG 1 New csv file

Finally, the newly created data set is successfully stored.

## Data Management Description

### Load dataset and recode the categorical variables

First, use pd.read to read the data and obtain the data type data in DataFrame format. Then, read the year, month, and day from the date variable column and put them into the new variable column.

1. covid\_data = pd.read\_csv(datapath)
2. *# divide the date object into three columns of year, month, and day data*
3. covid\_data['date'] = pd.to\_datetime(covid\_data['date'])
4. covid\_data["year"] = covid\_data['date'].apply(lambda x: x.year)
5. covid\_data["month"] = covid\_data['date'].apply(lambda x: x.month)
6. covid\_data["day"] = covid\_data['date'].apply(lambda x: x.day)

The purpose of this is to analyze the annual average or monthly average respectively in subsequent analysis, or you can use day as the content of the x-axis.

Since the missing data was removed during the previous data cleansing process, there will be no missing data now.

I have also considered recoding and classifying the values in variables, but it is mainly useful for drawing map-related data. However, it is not necessary to display the map data in this analysis, so there is no idea to classify and recoding variables.

### Create Secondary variables

Next, I created a custom auxiliary variable called mortality\_rate, which represents the proportion of the total number of deaths due to covid-19 to the total number of people in the country and then multiplied by 1000. This is because The ratio data obtained directly will be very small. For the convenience of observation, here I multiply it by one thousand.

1. covid\_data['mortality\_rate'] = (covid\_data['total\_deaths'] / covid\_data['population']) \* 1000

### Group the variables

I also used pd.qcut to divide the quartile (0 to 3) for the number of total\_deaths obtained for each country with the latest date. Nine countries had quartiles of 3, while only three countries had quartiles of 0.

1. covid\_data['deaths\_quartile'] = pd.qcut(covid\_data['total\_deaths'], 4, labels=False)

### Run Frequencies

Finally I ran the frequency profile on some variables and the results are as follows:

#### Number of entries present on each continent

1. print("Number of entries present on each continent")
2. print(covid\_data['continent'].value\_counts())

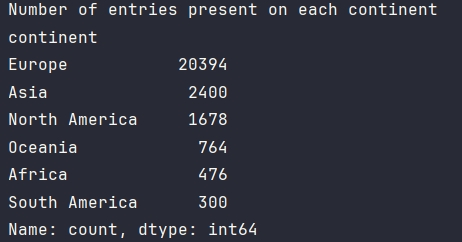


FIG 2 Result of total entries in continent

|  |  |
| --- | --- |
| Number of entries present on each continent | |
| continent | |
| Europe | 20394 |
| Asia | 2400 |
| North America | 1678 |
| Oceania | 764 |
| Africa | 476 |
| South America | 300 |
| Name: count | dtype: int64 |

#### Number of entries corresponding to each type of indicator

1. print("Number of entries corresponding to each type of indicator")
2. print(covid\_data['indicator'].value\_counts())

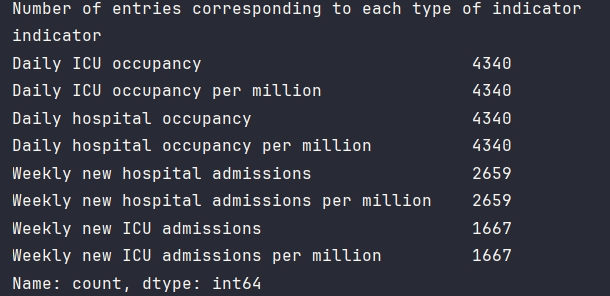


FIG 3 Result of entries in indicator

|  |  |
| --- | --- |
| Number of entries corresponding to each type of indicator | |
| indicator | |
| Daily ICU occupancy | 4340 |
| Daily ICU occupancy per million | 4340 |
| Daily hospital occupancy | 4340 |
| Daily hospital occupancy per million | 4340 |
| Weekly new hospital admissions | 2659 |
| Weekly new hospital admissions per million | 2659 |
| Weekly new ICU admissions | 1667 |
| Weekly new ICU admissions per million | 1667 |
| Name: count | dtype: int64 |

#### Number of null values in each column

1. print("Number of null values in each column")
2. print(covid\_data.isnull().sum())

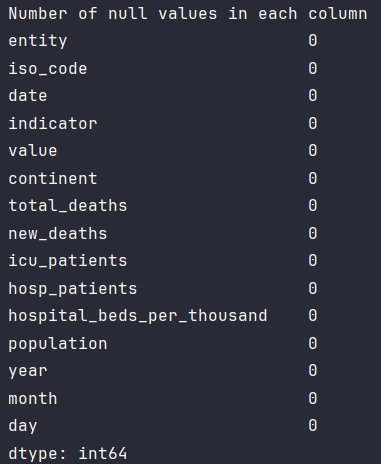


FIG 4 Result of null values in colemn

|  |  |
| --- | --- |
| Number of null values in each column | |
| entity | 0 |
| iso\_code | 0 |
| date | 0 |
| indicator | 0 |
| value | 0 |
| continent | 0 |
| total\_deaths | 0 |
| new\_deaths | 0 |
| icu\_patients | 0 |
| hosp\_patients | 0 |
| hospital\_beds\_per\_thousand | 0 |
| population | 0 |
| year | 0 |
| month | 0 |
| day | 0 |
| dtype: int64 | |

#### Names of participating countries in each continent

1. print("Names of participating countries in each continent")
2. print(covid\_data.groupby('continent')['entity'].nunique())



FIG 5 Result of participated countries in continent

|  |  |
| --- | --- |
| Names of participating countries in each continent | |
| continent | |
| Africa | 1 |
| Asia | 3 |
| Europe | 23 |
| North America | 2 |
| Oceania | 1 |
| South America | 1 |
| Name: entity | dtype: int64 |

#### Name of all countries participating in the statistical analysis

1. print("Name of all countries participating in the statistical analysis")
2. print(list(covid\_data['entity'].unique()))

|  |
| --- |
| Name of all countries participating in the statistical analysis |
| ['Australia', 'Austria', 'Belgium', 'Bolivia', 'Bulgaria', 'Canada', 'Cyprus', 'Czechia', 'Denmark', 'Estonia', 'Finland', 'France', 'Ireland', 'Israel', 'Italy', 'Lithuania', 'Luxembourg', 'Malaysia', 'Netherlands', 'Portugal', 'Romania', 'Serbia', 'Slovakia', 'Slovenia', 'South Africa', 'South Korea', 'Spain', 'Sweden', 'Switzerland', 'United Kingdom', 'United States'] |

# Descriptive Statistics and Graphing

## Select the variables of interest

First of all, according to the theme of my report, I think the most important variable is total\_deaths, which most directly reflects the number of deaths in a country during covid-19, so that we can sense whether the country has created a negative impact on its existing population. Reasonable medical resource conditions will be the basic numerical variable throughout this report.

In addition, through the columns selected above, I decided to use continent and entity (country name) as my categorical variables in the general statistical description, so that I can intuitively see the differences between countries or continents for different values.

## Create a single variable chart

### Histogram

With the above analysis, I first present the statistics of the data of total\_deaths in the form of a histogram of frequency distribution:

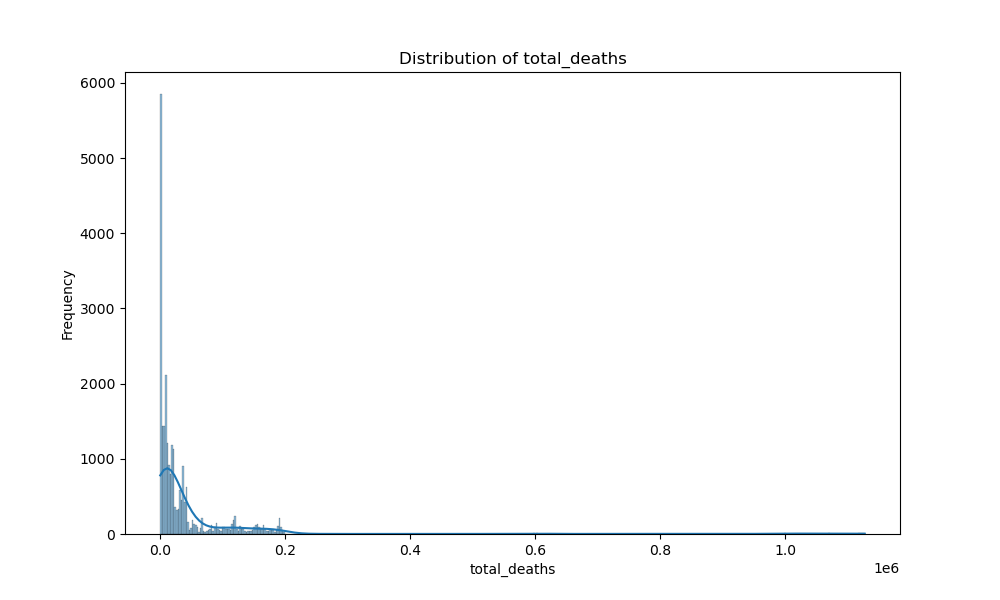


FIG 6 Histogram Distribution of total\_deaths

As you can see from this histogram, the vast majority of total deaths are concentrated between 0 and 0.2 times the sixth power of ten, which is between 0 and 200,000, while some values slightly above 200,000 also need some attention, considering that they may be in more populous countries, such as the United States.

Then do the same frequency distribution histogram statistics for the new\_deaths data:

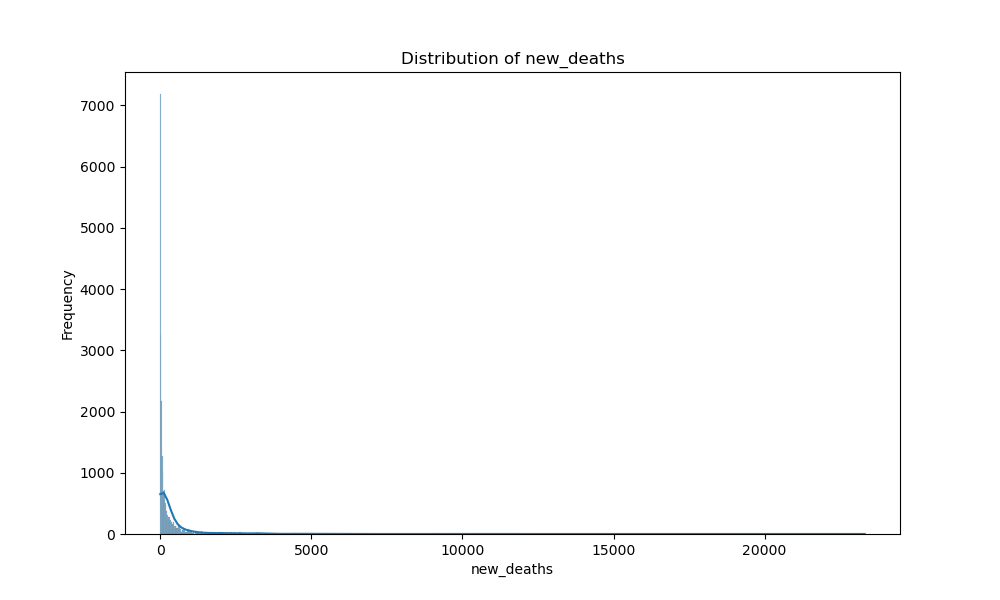


FIG 7 Histogram distribution of new\_deaths

From this frequency distribution chart of new death data, we can see that the data is mainly concentrated between 0 and 2000, which is a relatively normal range. However, there are also very few data that are very large, even exceeding 20,000 new deaths. Quantity, this data is still very scary.

### Bar Chart

Then I divided the final total death data according to continent and drew a bar chart:

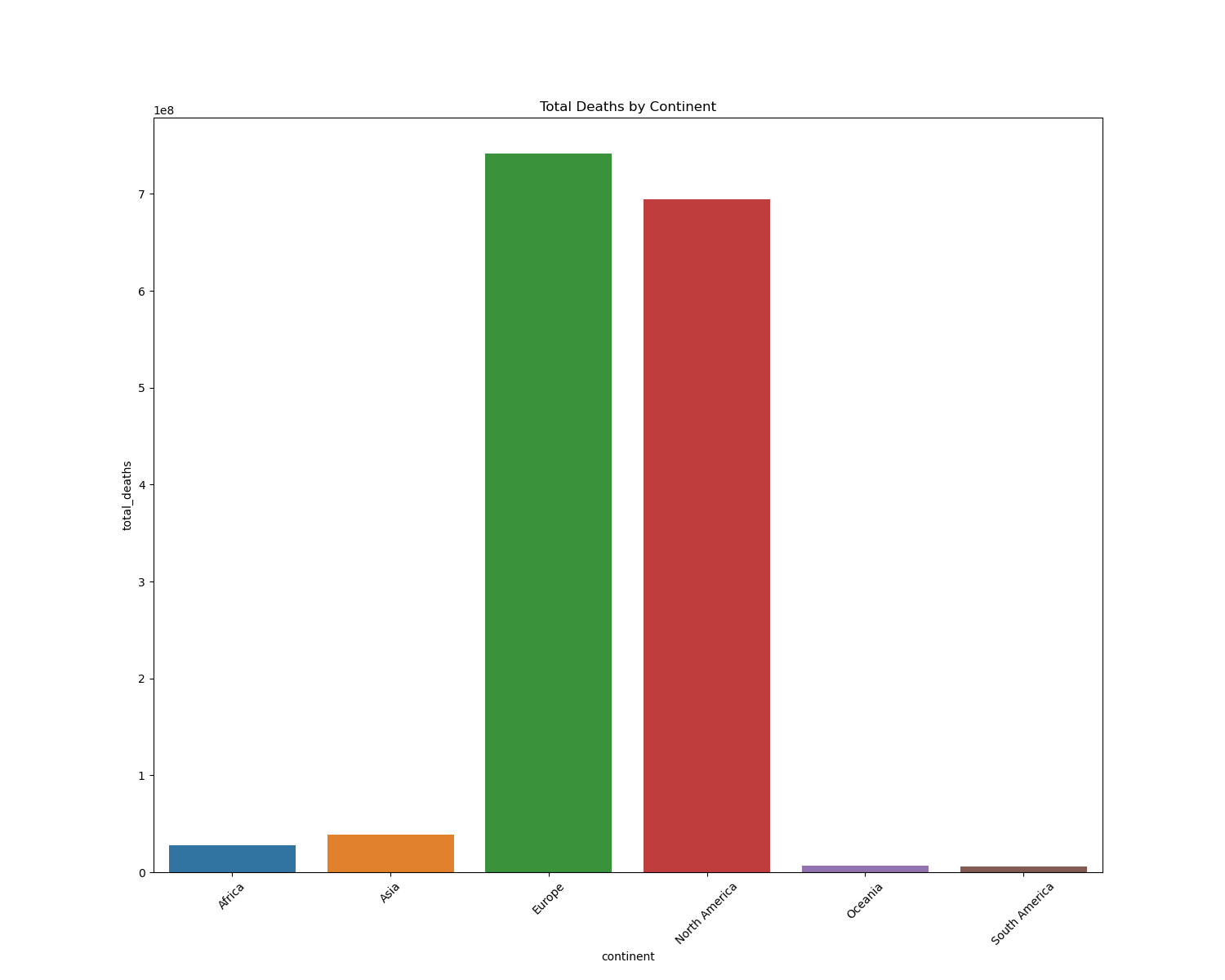


FIG 8 Bar Total deaths by continent

It is easy to see from this bar chart that the total number of deaths in Europe and North America will be very high, this is because in Europe, the number of countries participating in the statistical survey is the largest, and then in North America, although the number of countries is small, but the population is very large, so it will result in the situation shown in the chart.

## Determine Type of Variable

Numerical and categorical variables have been analysed and classified, and the other two types of variables will be classified and analysed in the following:

### Explanatory variables

Explanatory variables mean that the variables that cause direct results can be explained. My topic here is the relationship between medical resources and death, so death is the result, while medical resources are the interpretation. Then icu\_patients, hosp\_patients, hosp\_patients, hosp\_patients. hospital\_beds\_per\_thousand and the relationship between different indicators and their corresponding values are all my explanatory variables.

### Response variables

The response variable is more like a dependent variable, what kind of results will be caused by the change of the explanatory variable, so here new\_deaths and total\_deaths are my response variables.

## Preliminary data relationship diagram

In addition to the above analysis, I also conducted a preliminary analysis of the relationship between some data variables. First, I used a heat map to form a preliminary association matrix, and used colors to observe the differences between several variables.

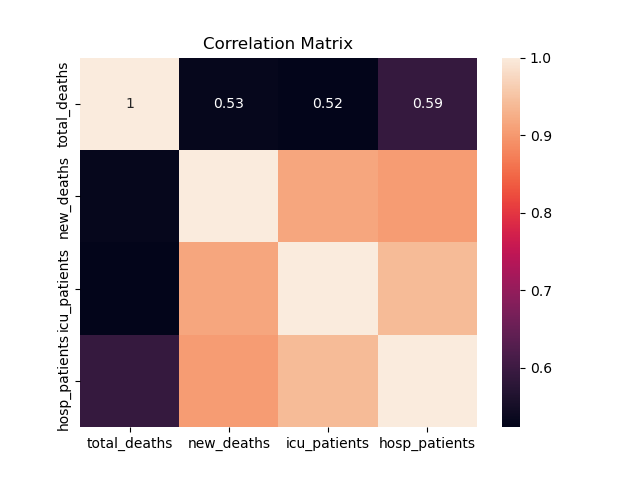
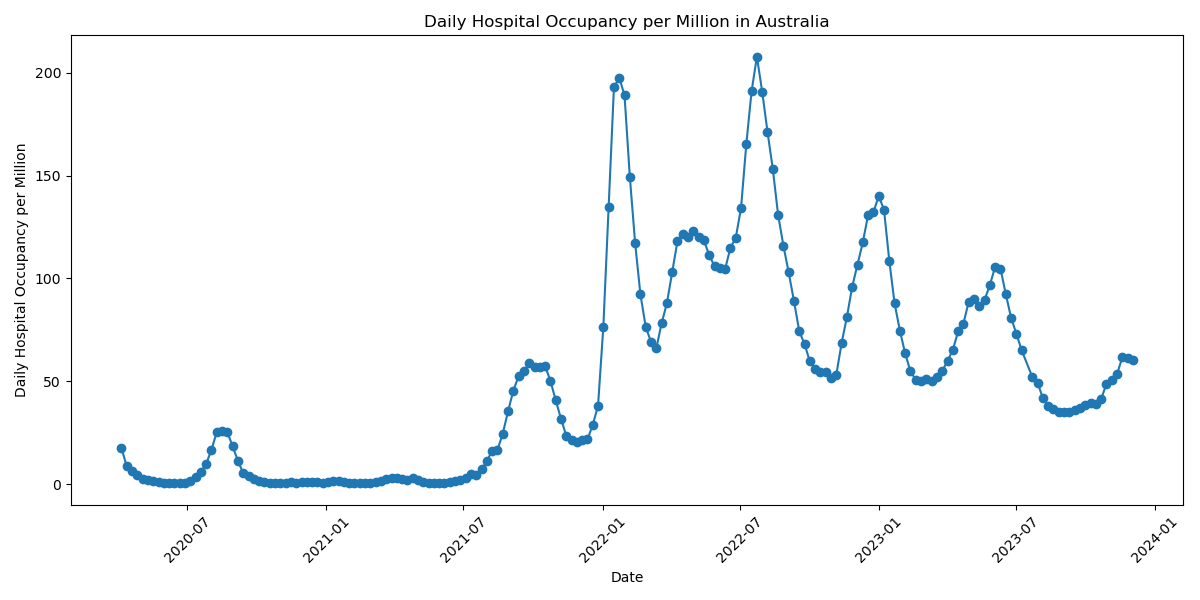


FIG 9 Heat Diagram Crrelation Matrix

Thus, it can be guessed that the number of icu patients and the number of hospital patients have a close relationship with the number of new deaths, that is, the correlation may be strong, but it may only be close in value, and the other relationship is not weak, I guess it may be negative correlation.

Here's an example of hospital resource use per million Australians for all recorded time periods, using Australia as an example.

FIG 10 Line Chart in AUS daily per million

It can be seen from this picture that at the peak, about 200 people per million people in Australia occupied hospital resources every day, and the peak occurred around February and August of 2022.

# Divide Subsets and Perform Statistics

In this section, I will first visualize countries with higher and lower total deaths and total deaths as a proportion of the country's total population; and then visualize countries with higher and lower allocations of medical resources.

It should also be emphasized that this so-called proportional data is actually the ratio of the total number of deaths to the total population of the country multiplied by one thousand.

## Total Deaths Situation

Here, I use a bar chart to describe and classify these ten countries selected for each project, in order to observe the problems of quantity and proportion, in which the blue one is the highest value and proportion, and the red one is the lowest.

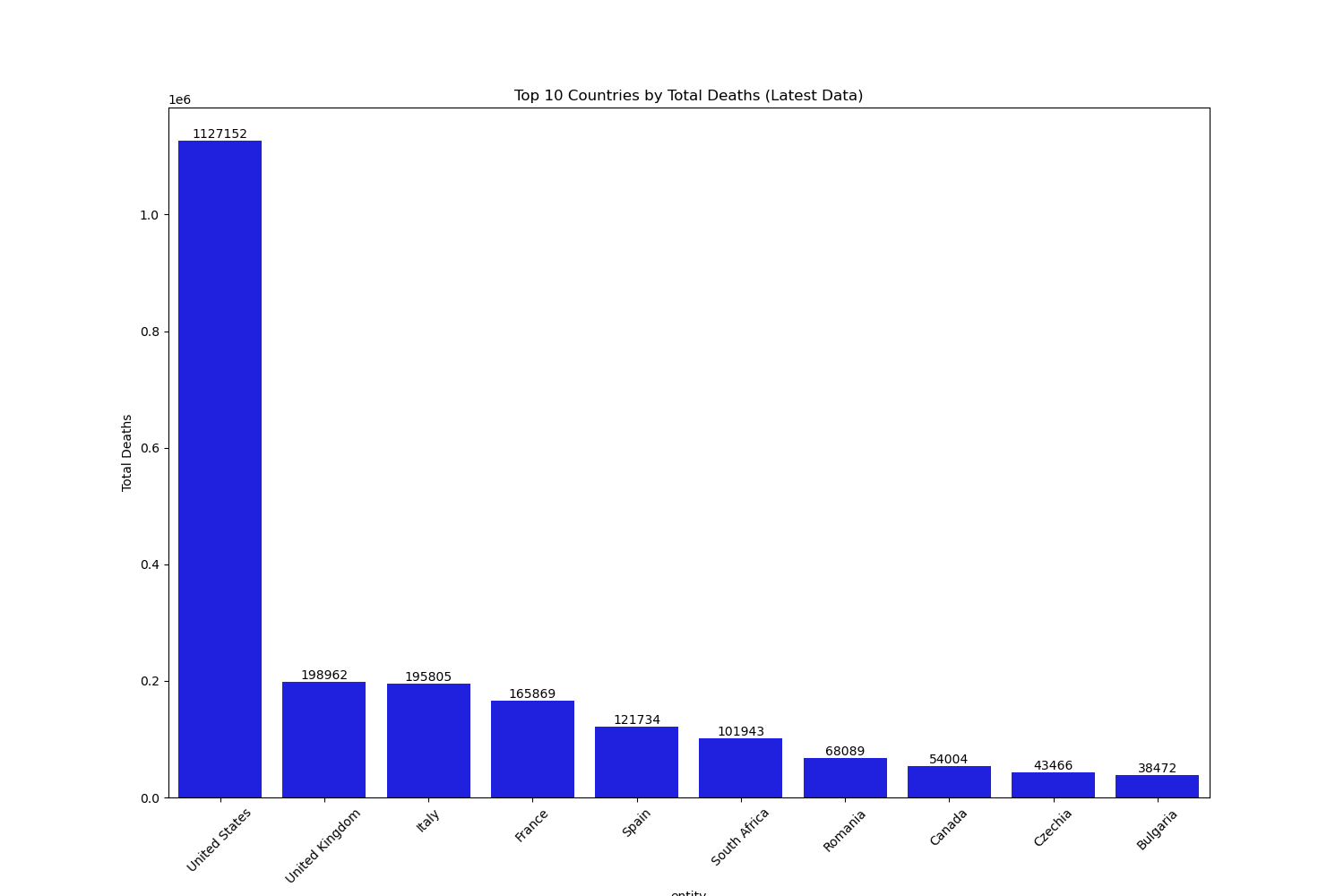


FIG 11 Top10 total deaths number

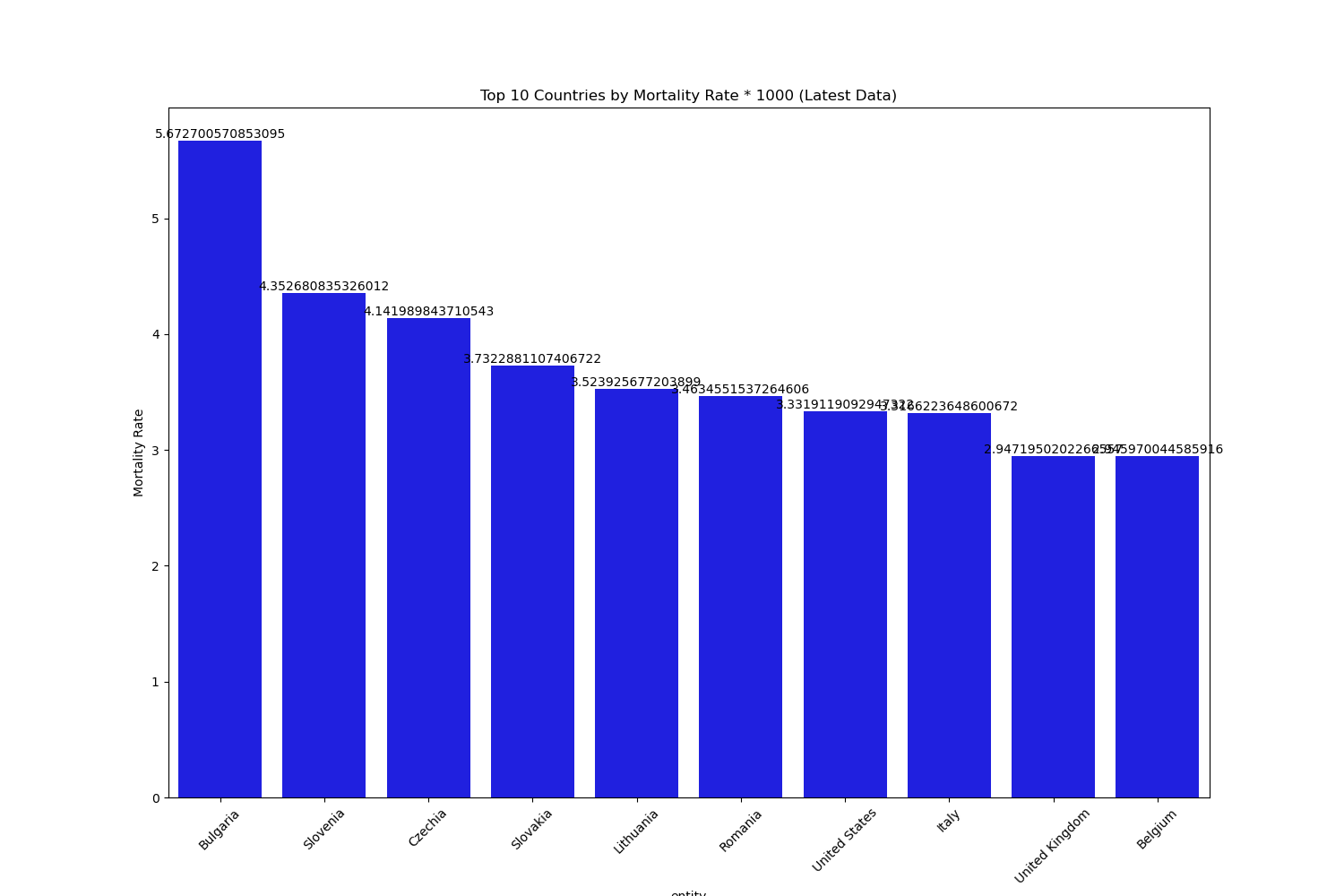


FIG 12 Top10 total deaths percentage

In terms of the total number of deaths, the United States of America is in a very high position with more than one million deaths, and the gap with the United Kingdom is more than 900,000, making it the country with the highest number of deaths due to Covid-19 in the world. In the section with the highest percentage of deaths, it can be seen that Bulgaria became the country with the highest number of deaths, through the calculation of this percentage can be seen to reach more than 0.567 per cent of the total population of the country, in addition to the old developed countries, such as Italy, the United Kingdom, and so on.

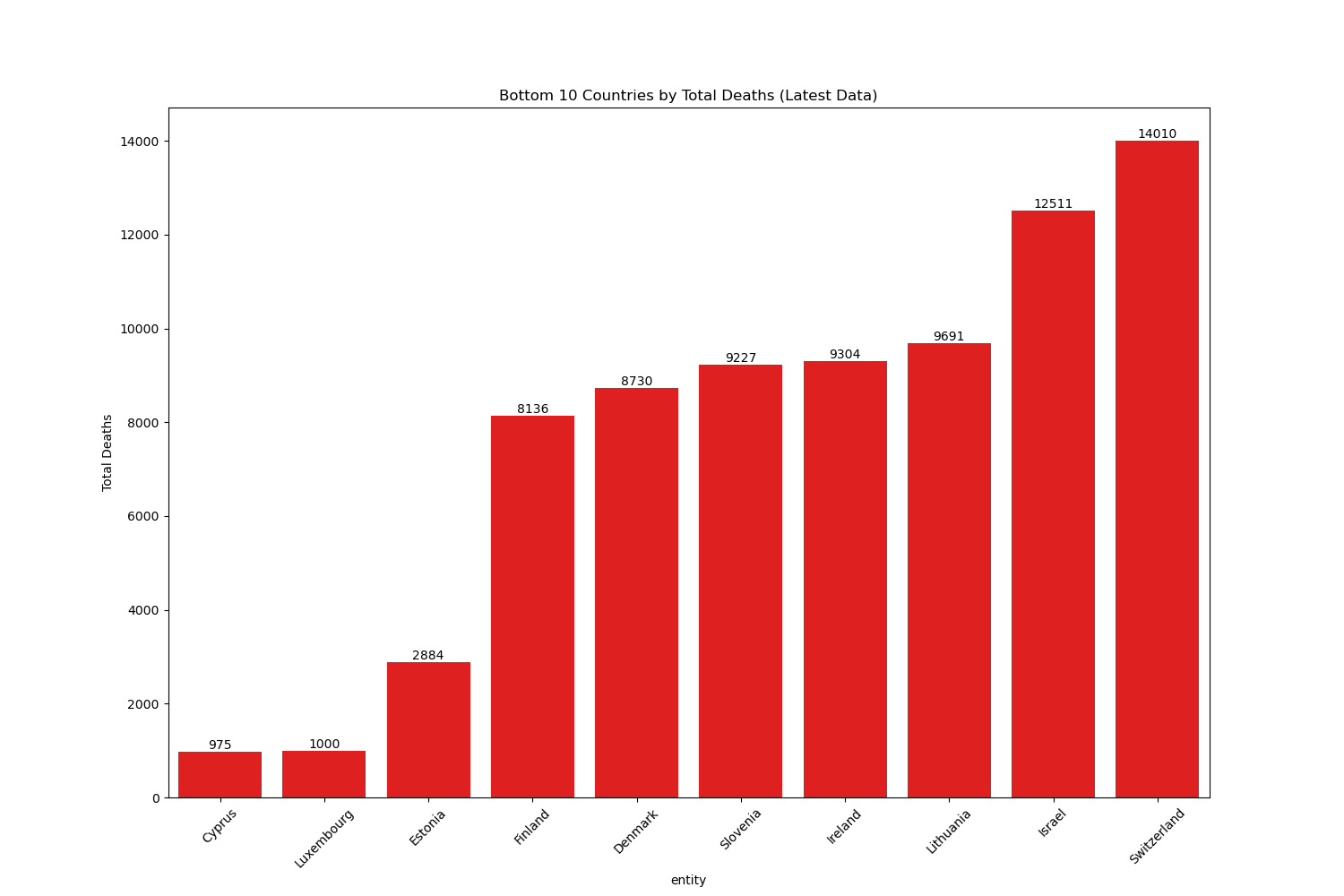


FIG 13 Bottom10 total deaths number

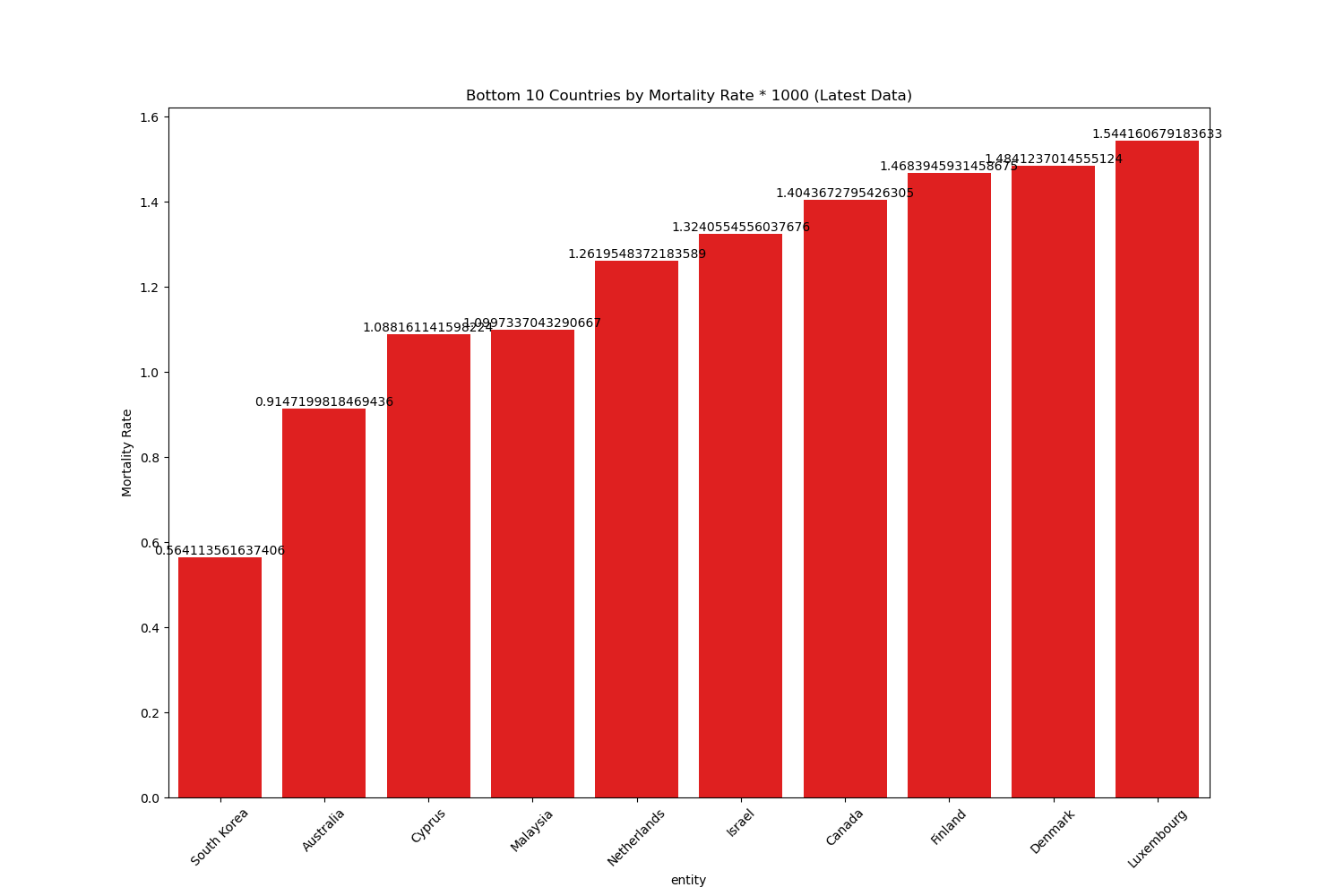


FIG 14 Bottom10 total deaths percentage

Among the countries with the lowest death toll, Cyprus ranks first, with only 975 people losing their lives due to the epidemic, followed by Luxembourg with 1,000 people, and there are still thousands of people in other countries who have lost their lives due to the epidemic. Death from covid-19.

In proportional data, South Korea from Asia occupies the top spot with a mortality rate of 5.64 per 10,000, followed by Australia, but it is also very close to 1 per 1,000.

## Medical Resources Situation

A similar step is taken here, first showing the total number, then showing the proportion, again selecting some countries from the maximum and minimum values to show and describe, where the variables include the number of beds per 1,000 people, the total number of hospital patients during the record period, the total number of icu patients and the respective descriptions of hospital and ICU occupancy per million people. And provide their maximum minimum, median and mean values.

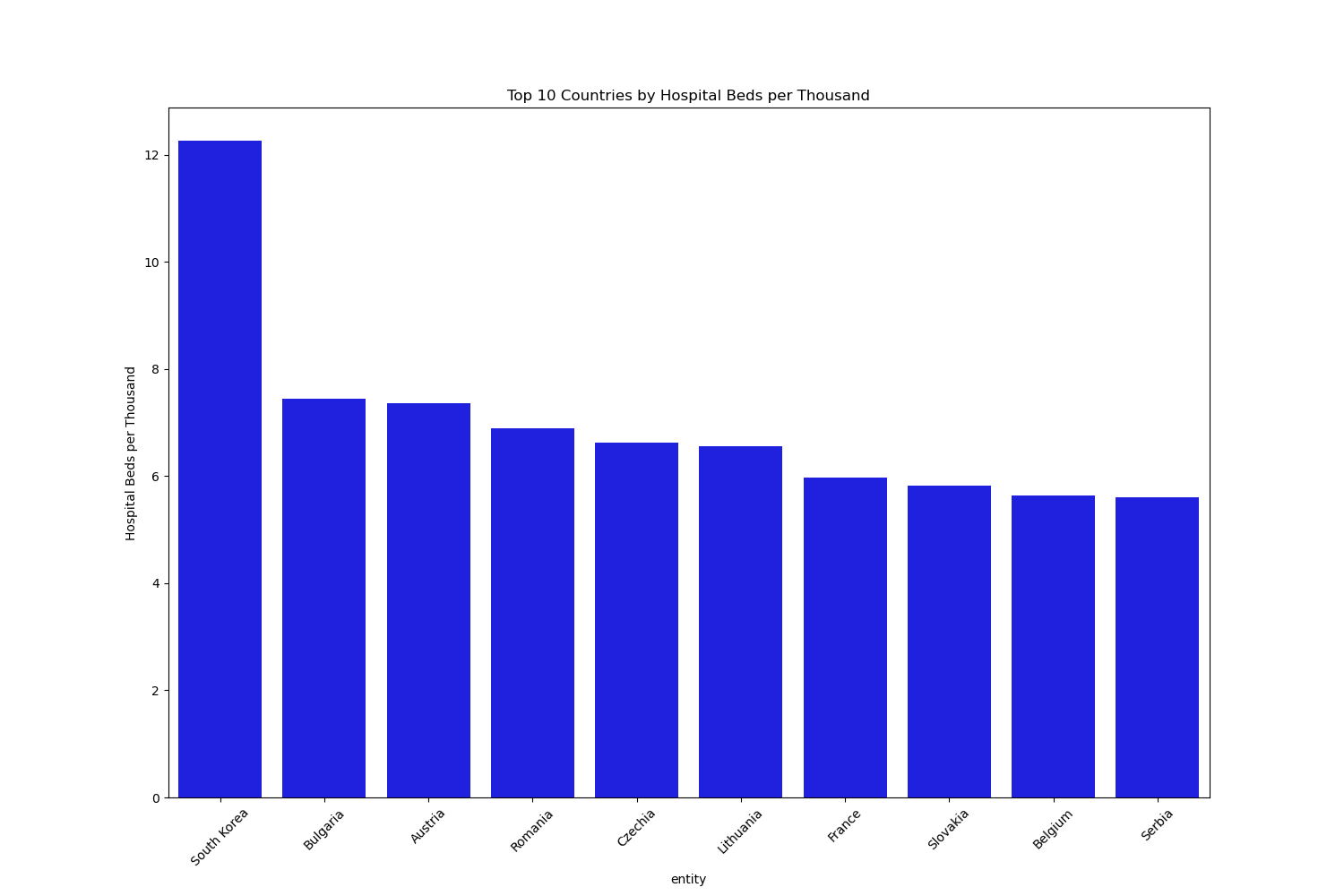


FIG 15 Top10 hosp beds per 1000

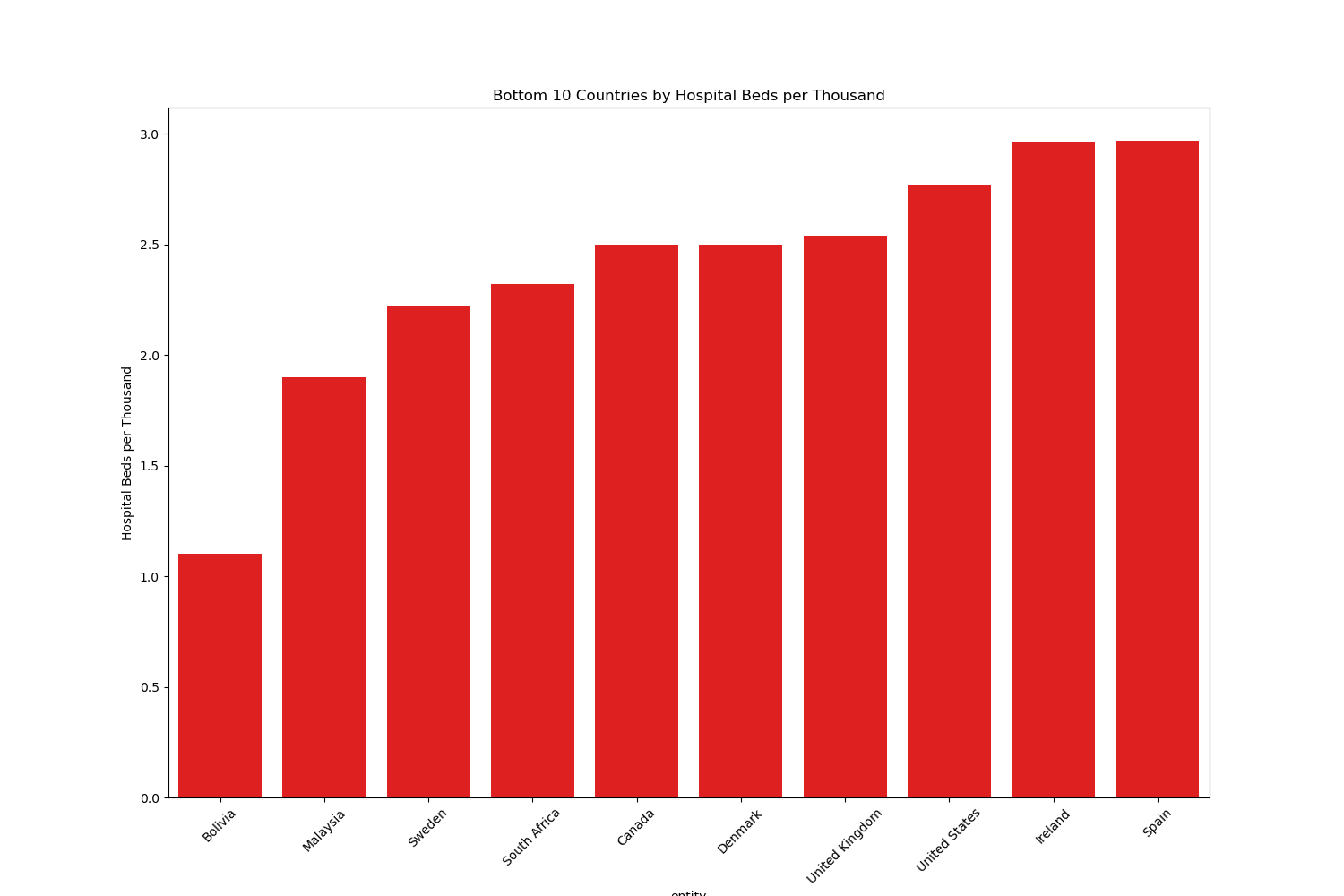


FIG 16 Bottom10 hosp beds per 1000

Looking at these two graphs, South Korea has the highest number of hospital beds per capita, over 12 beds per 1,000 people, and to some extent can be considered to have the most adequate hospital healthcare resources of any of these countries, and looking back it can be seen that South Korea does indeed also have the smallest percentage of deaths due to covid-19, but looking at Bulgaria, the second place country, this country surprisingly ranks first amongst these countries in the percentage of deaths due to covid-19, despite having about 7.5 beds per 1,000 people. Despite having about 7.5 hospital beds per 1,000 people, Bulgaria ranks first among these countries in terms of the proportion of deaths, which can cause some confusion from here on out. Bolivia, with an average of one bed per 1,000 inhabitants, is not in the top ten in terms of the number of deaths and the percentage of deaths.

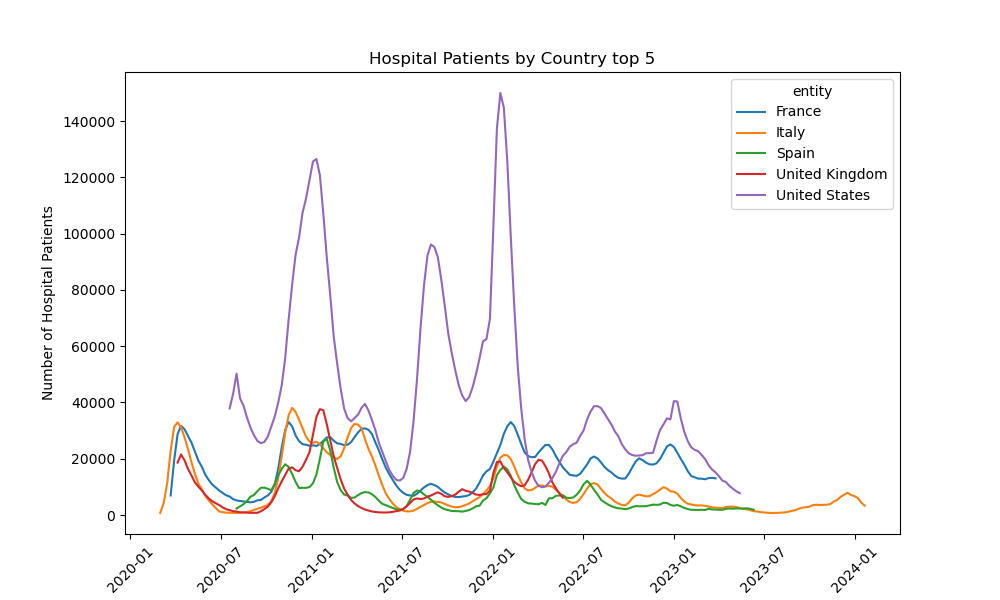


FIG 17 Top5 hosp patients

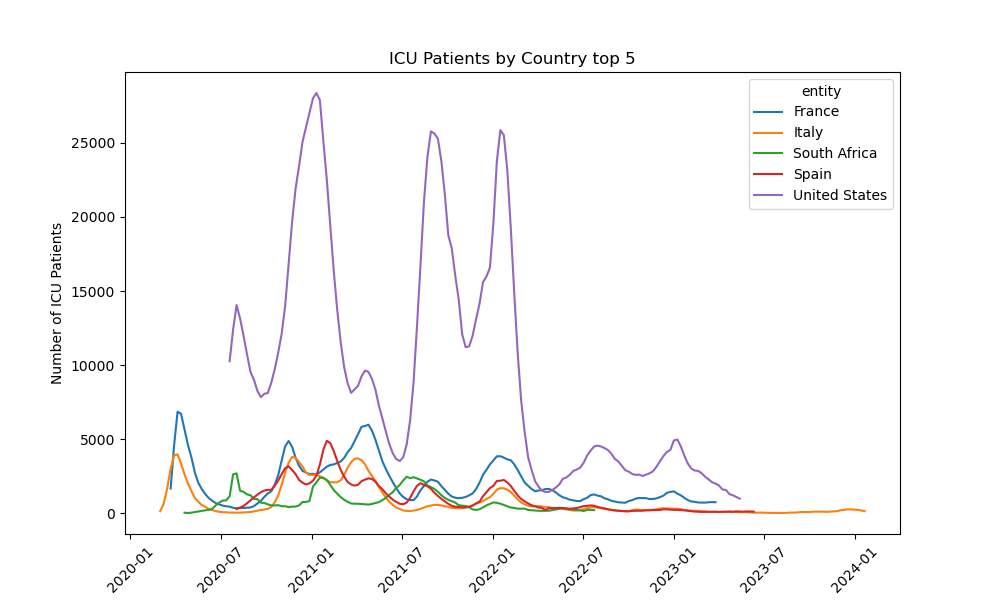


FIG 18 Top5 icu patients

Among the top five countries, the number of hospital resources used and the number of ICU resources used are basically proportional except for the United Kingdom and South Africa. It can be seen that the United States was in a very leading position almost throughout the entire period because of its population. Others Although countries fluctuate slightly, the differences are not too big.

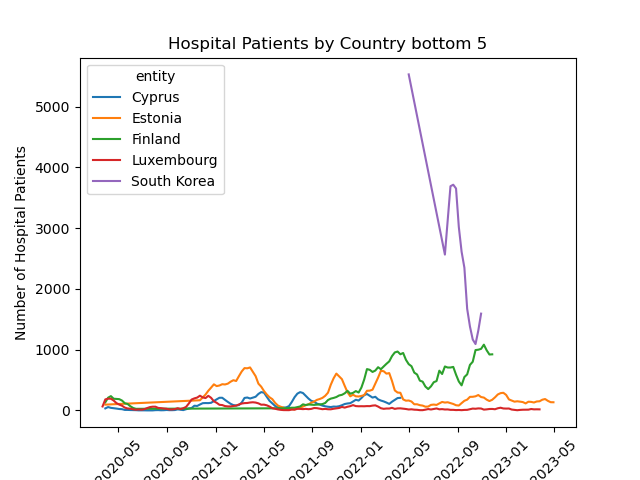


FIG 19 Bottom5 hosp patients

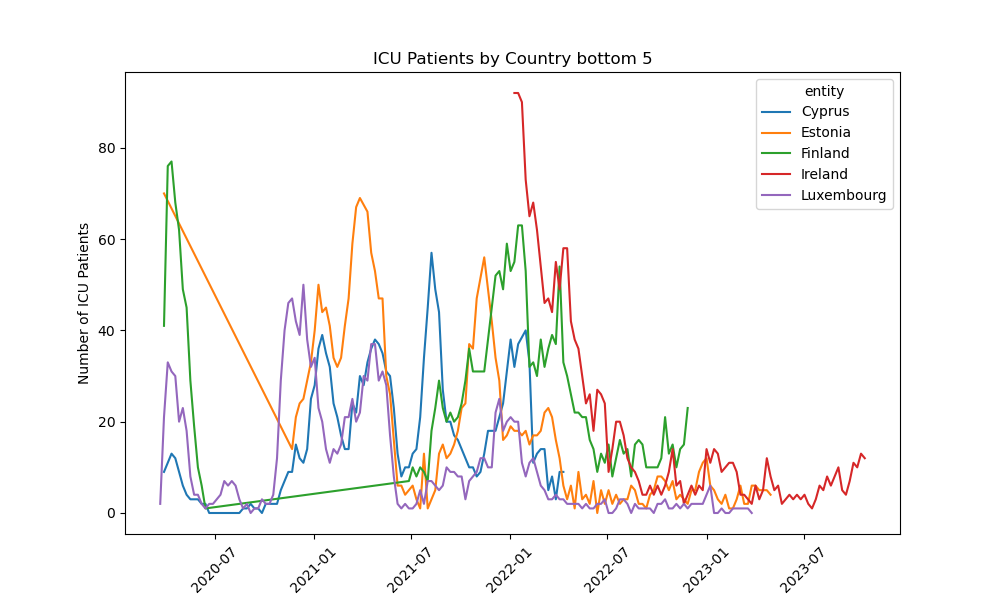


FIG 20 Bottom5 icu patients

In the last five countries that occupy the number of hospital resources and icu resources, the change is not very big, mainly South Korea, because South Korea has a population of more than 50 million, compared with the other four countries are also populous, so the number of hospital resources will be more abundant, but it can be seen that Luxembourg has been at a lower level from the beginning to the end.

For icu resource occupancy, the trends appear to be very mixed due to the small total, and the statistical significance is not significant due to the very low mortality and number of deaths in these countries.

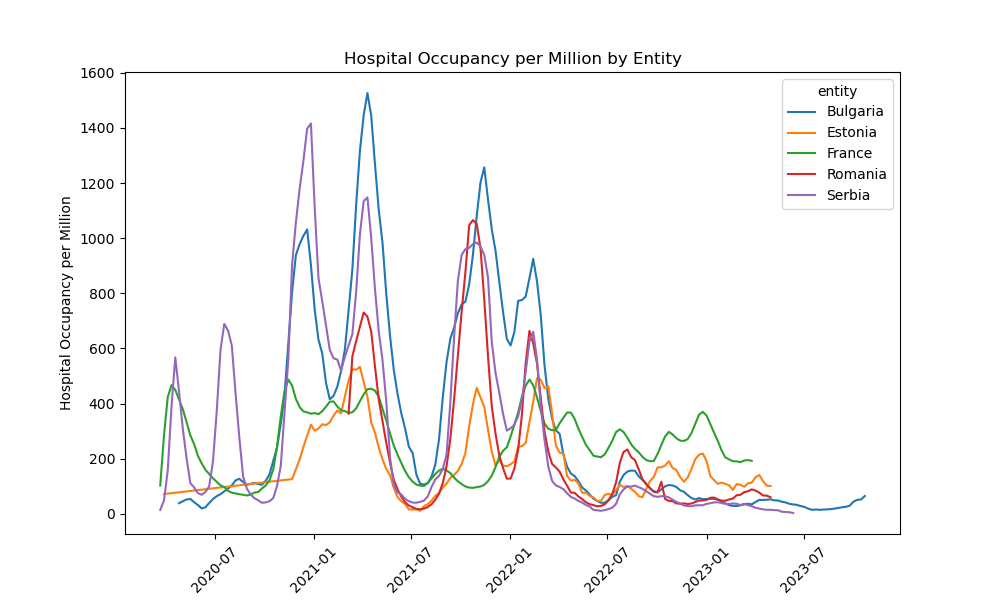


FIG 21 Hospital occupancy per million

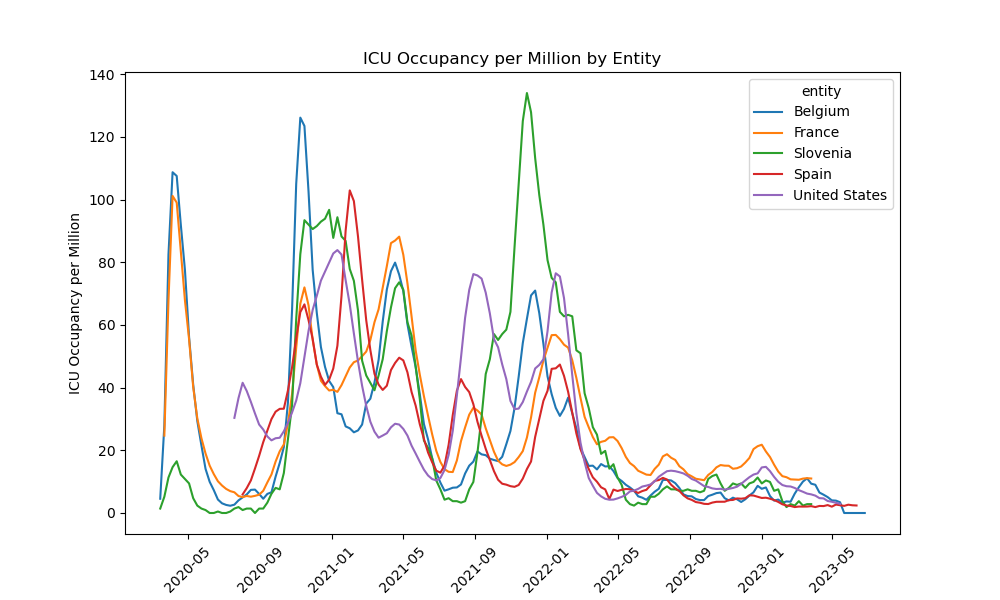


FIG 22 ICU occupancy per million

In terms of hospital occupancy and ICU occupancy per 1 million inhabitants, for reference, the countries with the highest proportion of hospital occupancy are Bulgaria, Serbia and Romania, while the countries with the highest proportion of ICU occupancy are Slovenia, France and Belgium. The above statistics show that although developed countries do not provide a lot of resources for hospitals, there are more resources for ICUs in hospitals than in developing countries.

## Variables Combinations Chart