

|  |
| --- |
| **HANOI UNIVERSITY OF SCIENCE AND TECHNOLOGY SCHOOL OF INFORMATION AND COMMUNICATION TECHNOLOGY**  **DEPARTMENT OF SOFTWARE ENGINEERING**  -\* |
| INTRODUCTION TO DATA SCIENCE  Prof. Muriel Visani |
| PROJECT: Analysis the correlation between the number of COVID cases/deaths in South East Asia and multiple factors |
| Students: |
| **Do Quoc An – 20194414**  **Ta Viet Cuong – 20194422**  **Nguyen Hoang Dang – 20194423**  **Ha Vu Thanh Dat – 20194424**  Hanoi, Dec 2021 Member of group    * Đỗ Quốc An – StudentID: 20194414   Email: an.dq194414@sis.hust.edu.vn   * Tạ Việt Cường – StudentID: 20194422   Email: [cuong.tv194422@sis.hust.edu.vn](mailto:cuong.tv194422@sis.hust.edu.vn)   * Nguyễn Hoàng Đăng – StudentID: 20194423   Email: [dang.nh194423@sis.hust.edu.vn](mailto:dang.nh194423@sis.hust.edu.vn)   * Hà Vũ Thành Đạt – StudentID: 20194424   Email: [dat.hvt194424@sis.hust.edu.vn](mailto:dat.hvt194424@sis.hust.edu.vn) |
| Description of Problem This project attempts to find the factors that affect the number of cases and deaths of coronavirus disase 2019 (COVID-19) patients in South East Asia. Countries were grouped based on population. On the overall, median age of the country, and average temperature are negatively related with the number of deaths from the virus. On the other hand, population density is positively related with the deaths due to COVID for low populated countries.  We have mentioned the recovery and stringency index in previous submittion but the recovery index is redundant with other variable and this index we can not crawl it from web, so we just use stringency index in this project.  Conclusion, we have six exploratory variables: GDP, average temparature, population density, median age, average rainfall and stringency index and our main variables are number of COVID case and deaths . Data ScrapingData Cleaning After scraping, we load data from csv file by pandas libarary  Table  Description automatically generated  *Figure 1: Average rainfall of each country*  Firstly, we check NA values for all data |
| **Text  Description automatically generated with medium confidence**  *Figure 2: Example NA value of country name*  COVID-19 dataset is a time-series dataset of entire world about new case, death,… we filter South East Asia area from it |
| **Graphical user interface, table  Description automatically generated with medium confidence** |
| *Figure 3: COVID-19 data of South East Asia*  Cause the dataset includes many index, we just need some important index such as total\_case, new\_case, total\_death,…  Table  Description automatically generated with low confidence  *Figure 4: Necessary index*  We merge all country and index into a dataset to easy visualize |
| **Table  Description automatically generated** |

*Figure 5: Dataset after preprocessing*

# Exploratory Data Analysis

Number of Report Cases

Table

Description automatically generated with medium confidence

The distribution of new cases below illustrate a possitive skew graph. We can easily see that new case of COVID dramatically increase from June to September due to Delta variant. Moreover, confirmed case of Indonesia hit the highest point and take faraway the remaining country of South East Asia.

Chart, line chart

Description automatically generated

Number of Report Deaths

Table

Description automatically generated

The distribution of Death cases is quite similar to new cases with possitive skewness. However, death cases of Indonesia is fourth time to other country, while new cases is just twice.

Chart, line chart

Description automatically generated

Spearman Heatmap

Chart

Description automatically generated

Cause we have many index to visualize, we can not display pairplot in the report, so we will highlight some index that have relation to each other.

Chart, scatter chart

Description automatically generated

We can observe from the graph above that we can divided South East Asia into 3 group based on total case vs GDP:

* Group 1 (Myanmar, Timor Leste, Laos, Campuchia): This group include developing countries that has lowest GDP and cases.
* Group 2 (Vietnam, Malaysia, Thailand, Philippines, Indonesia): Also containing developing countries but this group has stronger economy and bigger area.
* Group 3 (Singapore, Brunei): This group can be outlier cause two countries has smallest area (just by a city of other country) and their GDP is too high.

Chart, scatter chart

Description automatically generated

The figure above show that the median age in range 22.5 to 30 has most of cases, signifying citizen in this age has most affectd by COVID-19.

# Visualization

Chart, line chart

Description automatically generated

The figure above give information about new cases, deaths, vaccination and stringency index by time series of Vietnam. We can observe from the graph that since June 2021, new case and death grow enormously, and in this time, Vietnam start vaccination to citizen

* The stringency index reach a peak when the case rise sharply in July and go to downtrend when cases decrease and vaccination increase
* In November, the vacination hit the highest point and the stringency is lowest, we can see new case grow slightly but deaths is not, cause inviduals has vaccinated.

Chart, pie chart

Description automatically generated

The chart represent the proportion of case in South East Asia. It can be clearly seen that countries in Group 2 has most cases, more than 90%, while Indonesia take one-third of this group.