Covid-19 In Malaysia

By: Mona Bandov
Date: December 8, 2023

https://www.cnn.com/2021/01/12/asia/malaysia-state-of-emergency-covid-intl-hnk/index.html

Introduction:

The analysis focuses on understanding regional disparities in COVID-19 incidence rates across different states (13 states) in Malaysia and explores the impact of the pandemic on various regions in terms of cases and recoveries. It aims to uncover significant variations in how states have been affected. Additionally, the analysis takes a closer look at the correlation between population density in different states and the rate of COVID-19 transmission. By examining how states with diverse population densities have managed the spread of the virus, the analysis aims to uncover insights into the relationship between population density and the pandemic's impact in Malaysia between January 25, 2020 to January 25, 2023.

Research Questions:

- Regional Disparities:
  - How do COVID-19 incidence rates vary across different states in Malaysia?
  - Are there significant disparities in the impact of the pandemic on various regions in terms of case new cases and case recovery?
- Population Density and Transmission
  - Is there a correlation between population density in different states and the rate of COVID-19 transmission?

Data Collection and Cleaning:

The datasets collected include:

I. Covid Cases by State: https://covid-19.samsam123.name.my/api.html
II. Population of Malaysia: https://documenter.getpostman.com/view/16605343/Tzm8GG7u
III. Malaysia ShapeFile: https://cartographyvectors.com/map/1477-malaysia-with-regions

The state level Covid-19 case data and population data were extracted from an API. There were 14 HTTP request calls (13 states and 1 population data). Upon extraction and cleaning, the data was converted into a CSV file for data analysis. All CSV state files were merged to get the total COVID-19 in Malaysia cases between January 25, 2020 to January 25, 2023. The shapefile that contains all the coordinates for the States in Malaysia are stored in. These coordinates were merged with the final CSV file.

More information on the data cleaning, processing and results can be found*: results.txt, raw.txt, and processed.txt.*

The final (cleaned) static data set: used for graphing.

final_covid_data.csv → Covid Cases by State + Population of Malaysia
final_rate_data.geojson → Covid Cases by State + Population of Malaysia + Malaysia Shapefile

There are 13 columns and 16,455 rows.

A codebook can be found in *proccessed.txt*

Note: During the cleaning process, there was a lot of data manipulation and direct changes to the CSV files. When running the document, load static data from data > processed > *final_covid_data.csv* and *final_rate_data.geojson*.

Data Exploration:

- ➢ Found in *graphing.ipynb*:
    - ○ Summary statistics
    - ○ Looking for NA's
    - ○ Dimensions

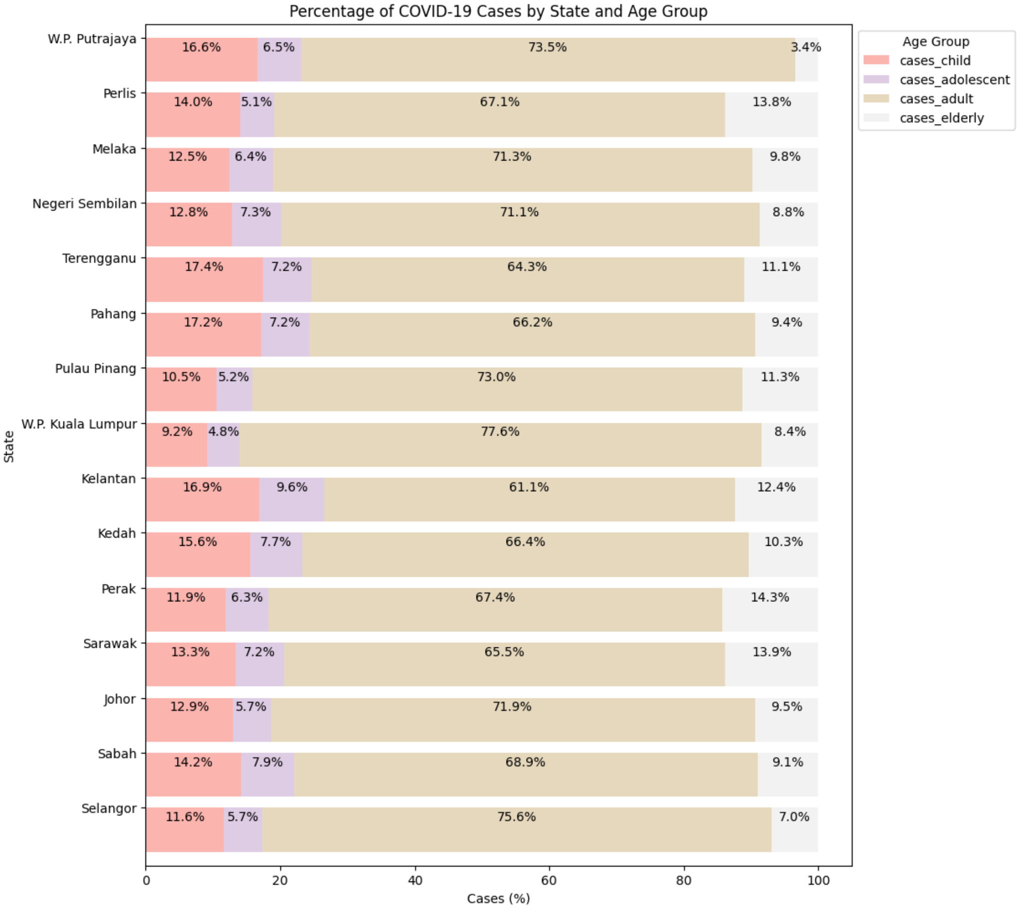
Challenges:

The API that belonged to the author had been broken, and therefore it did not capture real-time changes. The author was contacted through Linkedin and he provided the source of the data. Luckily, this prompted him to fix his API for future users. The original API had all the data, but did not include separate state data. Extracting state data was long because it took 20 minutes (20 minutes x 13 states = 260 minutes) for each dataset to be processed and exported into a static CSV file. Computer processing was slow. The shapefile slowed down the computer.

Data Analysis 1: [Static Bar Graph]

How does the distribution of COVID-19 cases across different age groups vary among states in Malaysia?

Fig. 1



Age is classified as : Child → 12 and under | Adolescent → 13-17 | Adult → 18-59 | Elderly → 60+

Findings: The states are ranked (bottom to top) from most populated state to the least populated state. The most populated state is Selangor with 11.6% of the new covid cases are children, 5.7% are adolescents, 75.6% are adults and 7% are the elderly. The least populated state is W.P. Putrajaya with 16.6% of the new covid cases being children, 6.5% being adolescent, 73.5% being adults, and 3.4% being elderly. The least populated state has the lowest percentage of new cases infecting the elder population.

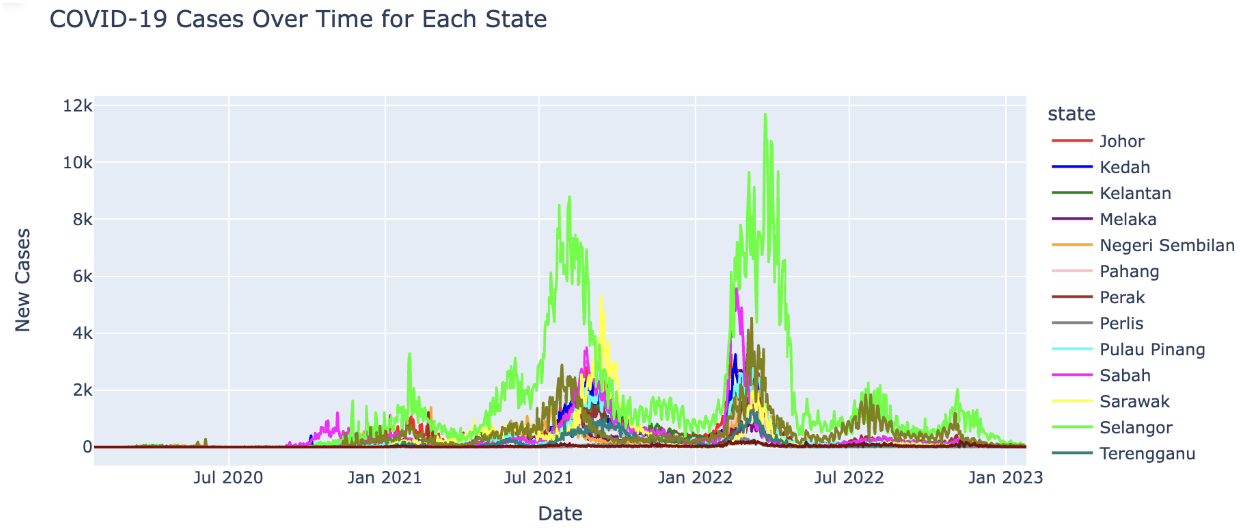
Data Analysis 2: [Interactive Graph]
How does COVID-19 incidence rates vary across different states in Malaysia?

This is the static version, the interactive version is called *interactive_new_cases_by_state.html*.

Interactive Features:

- ➢ Select States
- ➢ Zoom In and Out
- ➢ Take Screenshots of Output

Fig. 2

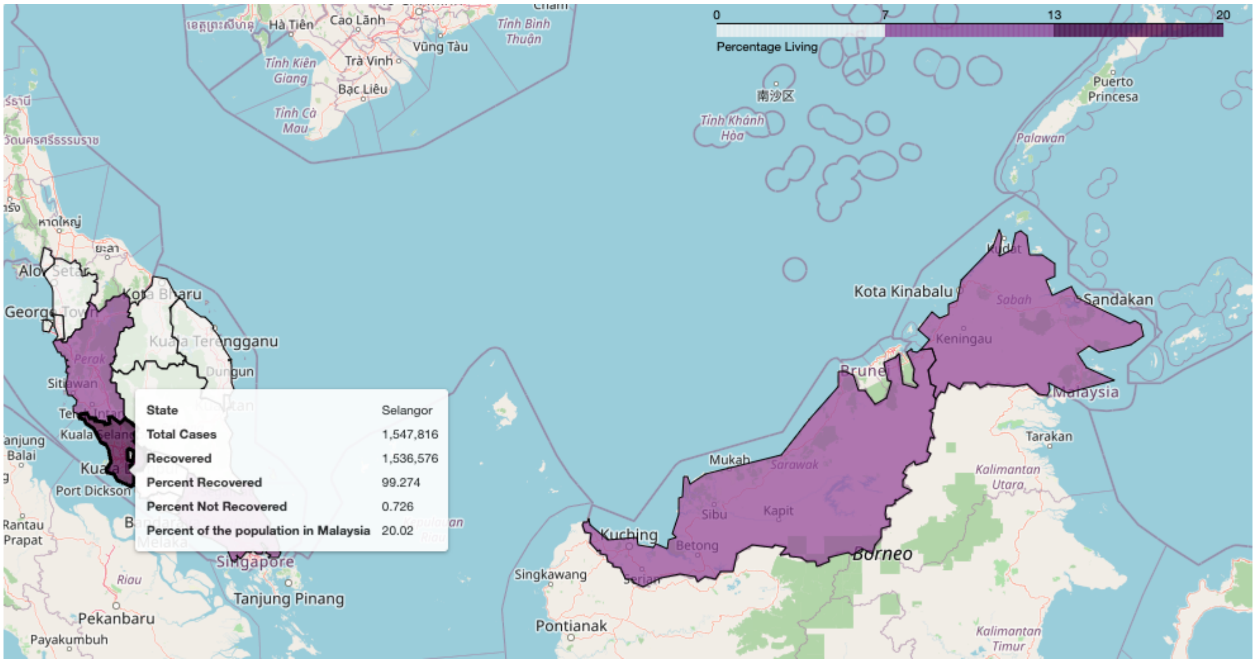


Findings: Since this is an interactive plot (plotly), the states can be selected to view the relationship of COVID-19 cases over time for each state. Between January 25, 2020 to January 25, 2023, there seems to be 2 major peaks. The first major peak August 6, 2021 with reported 8792 new cases and the second peak March 24, 2022 with reported 11,622 new cases. These two peaks happen in Selangor. After September 6, 2020, the number of reported new Covid cases started to increase rapidly. This could be due to several reasons: testing availability, people avoiding the hospital, or mis-information around Covid. By May 2023, the number of new cases started to drop. This could be due to lower testing rates.

Data Analysis 3: [Interactive Map]

Are there significant disparities in the impact of the pandemic on various regions in terms of cases and recoveries?

Fig 3.



State	Selangor
Total Cases	1,547,816
Recovered	1,536,576
Percent Recovered	99.274
Percent Not Recovered	0.726
Percent of the population in Malaysia	20.02

Findings: This is a static image of the *recovered_rate_map.html*. The interactive map can be used to see each state's name, total cases, total recovered cases, percent recovered, and percent not recovered. Since death data is not included, the percent not recovered can be due to death or lost to follow up. According to the visualization, 1.9% (highest percent not recovered) of those total Covid cases in Melaka did not recover. About 0.54% of the people in Kelantan did not recover from Covid. W.P Kuala Lumpur, the capital, is surrounded by the most populous state Selangor (makes up 20.02% of the population in Malaysia). Their recovery rate of 0.73% is average among other states.

Data (Statistical) Analysis 4: [Scatter/Statistical Plot]
Is there a correlation between population density (total population in each data) in different states and the rate of COVID-19 transmission?

Assumptions for Linear Regression:

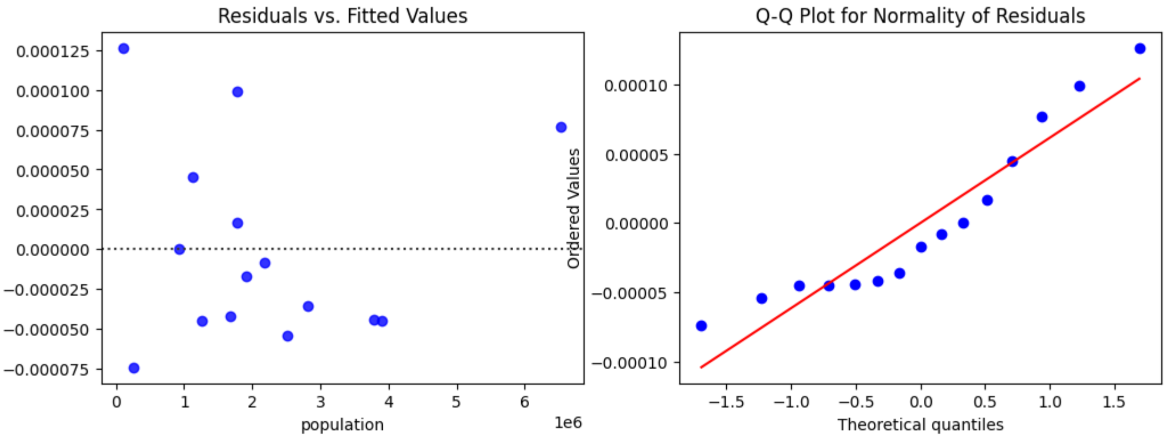
1. Linearity (violed)
2. Normality (not violated)
3. Homoscedasticity (violated)
4. Independent

Shapiro-Wilk Test for Normality:

```
Shapiro-Wilk Test for Normality:
Test Statistic: 0.8864492774009705
P-value: 0.05927490070462227
```

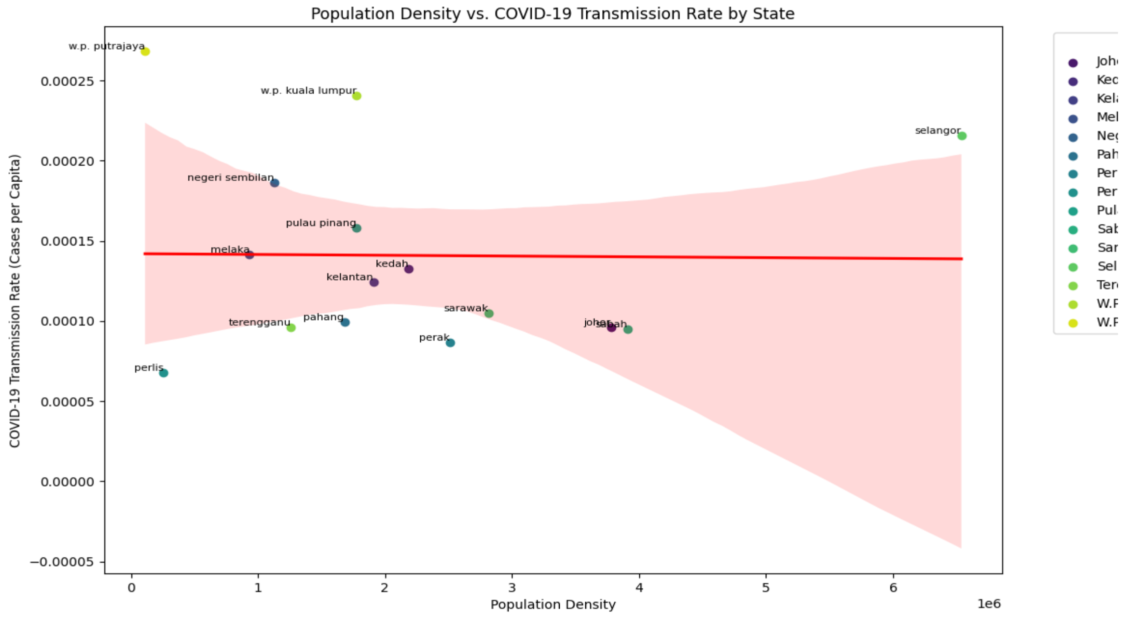
The Shapiro-Wilk test is used to validate the assumptions of normality for the residuals of a linear regression. The test statistic was 0.886 with a p-value of 0.59. Since the test statistic is closer to 1, it is more consistent with normality. The null hypothesis is that the residuals come from a normal distribution. Since the p-value (0.59) is greater than the common significance level of 0.05, we do not have sufficient evidence to reject the null hypothesis. Therefore, there is no significant departure from normality in the residuals. The normality assumption has been met. Although normality has been met, at least one of the assumptions has not been met.

Fig 4.



The residual vs. fitted graph and Q-Q plot suggest that there is a violation of assumptions. There are also only 13 data points. Non-parametric tests should be used. Spearman's Correlation is a measure of association between for nonparametric data.

Fig 5.



```
Spearman's Correlation Results:
Spearman's Correlation Coefficient: -0.2071428571428571
P-value: 0.4588428049634703
The relationship is not statistically significant.
```

The spearman's correlation coefficient between population density and COVID transmission rate is approximately -0.2071 (p-value = 0.4588). There is a weak negative correlation between the two variables. Given the p-value is greater than 0.05, the correlation is not statically significant. There is insufficient evidence to conclude that the observed correlation is different from 0. Population density is not a strong predictor of COVID-19 transmission rates. Further investigation is needed to understand the relationship between population density an COVID transmission rates.

Conclusion:

The examination of COVID-19 cases across different age groups showed patterns, with populous states like Selangor showcasing high covid cases among adults. Selangor is the state that surrounds the state of W.P. Kuala Lumpur (the capital). Meanwhile, W.P. Putrajaya, the least populated state, exhibits a lower impact on the elderly population. The interactive plot tracking incidence rates reveals high covid rates in Selangor. Moreover, an analysis of disparities in recoveries across states shows regional differences, with Melaka showcasing a higher percentage of cases not recovered. For those who did not recover, it is hard to tell if those people were lost to follow up or had died due to the Covid infection.

In tandem, an investigation into the correlation between population density and COVID-19 transmission rates introduces additional layers of understanding. Linear regression assumptions, although met for normality, reveal violations in linearity and homoscedasticity, prompting a shift to non-parametric testing. Spearman's Correlation Coefficient showed a weak negative correlation between population density (population in each state) and transmission rates, yet the lack of statistical significance implies that population density alone does not serve as a predictor for covid-rates. Although Selangor, the most populous state with the highest cases of Covid, there is statistical evidence that suggest that population density is not associated with covid rates. Further data and contextual analysis needs to understand why covid rates increase in some states as opposed to others.

Future Considerations:

My original idea for the project was to see if vaccination campaigns in Malaysia had an effect on vaccine registration and administration. However, there were many limitations to the API that prevented me from finding the right data set. Because shapefiles (lots of data) were used, more computational power is needed. I would consider looking at the area of the state versus its population. This is because bigger countries don't necessarily mean a larger population. Since I have temporal data, I could have considered looking at the data by months or by years.