7COM1079-0901-2024 - Team Research and Development Project

Final report title: HEALTH - COVID CASES and AGE in Malaysia Covid Cases

Group ID: 172 A

Dataset number: DS046

Prepared by: Ranpal Reddy Mula [23102715]

S N Venkata Lankalapalli [23095065]

Srikanth Reddy Poreddy [23088077]

Sai Teja Talasila [23024034]

Jaya Raju Vangapati [22070311]

University of Hertfordshire

Hatfield, 2024

Table of Contents

Introduction

Problem statement and research motivation

The increasing number of COVID-19 cases among the 18-29-year-old population in Malaysia draws concerns toward public health and safety. Understanding the relation of the absolute number of cases in this age group to their population size in different states is very important (Loo and Letchumanan, 2021). This study aims to uncover patterns that will help in the targeted interventions and resource allocation The study will provide valuable insight into the management of future outbreaks and improvement of health strategies for young adults, towards better health outcomes in this demographic.

The data set

The dataset contains weekly COVID-19 case counts by different age groups within states. It includes the absolute case counts for the ages 0-4, 5-11, 12-17, 18-29, 30-39, 40-49, 50-59, 60-69, 70-79, and 80+. It also contains their respective population sizes. It also gives the percentages of cases for each age group. The structured data will help to analyse trends and correlations of case numbers with population size in the diverse demographics in Malaysia.

Research question

RQ: Is there a significant correlation between the absolute number of COVID-19 cases in the age group 18-29 years (abs_18_29) and the population size of this age group (capita_18_29) across different states in Malaysia?

The analysis will use Spearman's Rho or Kendall's Tau tests to answer the above research question. This is a non-parametric test for correlation that do not assume normality. Data on COVID-19 cases and population sizes for the 18-29 age group will be collected. These tests will help in the identification and measurement of the strength of the relationship between these variables across different states.

Null hypothesis and alternative hypothesis (H0/H1)

The null and alternate hypothesis is mentioned below:

1. Null Hypothesis (H0): There is no association between the absolute number of COVID-19 cases within the age group 18-29 years (abs_18_29) and the population size for the same age group (capita_18_29) in different states in Malaysia. This hypothesis will suggest that increases or decreases in case numbers are independent of the change in population size within Malaysia.

2. Alternative Hypothesis (H1): There is a relationship between the absolute number of COVID-19 cases in the age group 18-29 years, abs_18_29, and the population size of this age group, capita_18_29, across different states in Malaysia. This hypothesis shows that as the population increases or decreases, the number of COVID-19 cases will similarly change.

Background research

Research papers

A review of demographic characteristics and COVID-19 case numbers highlights the relationship of how the absolute number of cases of COVID-19 among different age groups and the population size of this group within Malaysia.

In this context, (Okoli, Neilson and Abou-Setta, 2022) used a dataset at the country level between January 2020 and August 2021. This study established the significant associations of the various country-level characteristics with COVID-19 cases. The author found that case numbers were positively correlated with the male and female ratio. It was also found that the proportion of individuals aged 60 and above, countries with higher universal health coverage indices, and significant tourism activity had higher number of covid cases. Higher economic status, however, correlated negatively with case numbers and mortality rates.

Another study by (Chen et al., 2020) reviewed the emigration of populations from Wuhan in the early stages of the outbreak. The author found that 59.91% of all confirmed cases in China were located in Hubei province, with a 0.943 correlation coefficient between provincial case numbers and emigration from Wuhan, which strongly points out the role of population movement in the distribution of cases.

A similar study by (Wong and Li, 2020) focused on population density in U.S. counties. It shows that it accounted for 57% of variation in infection cases. These studies have shown that demographic factors are very important in understanding COVID-19 dynamics. This also lays a foundation to explore similar correlations among the young adult population in Malaysia.

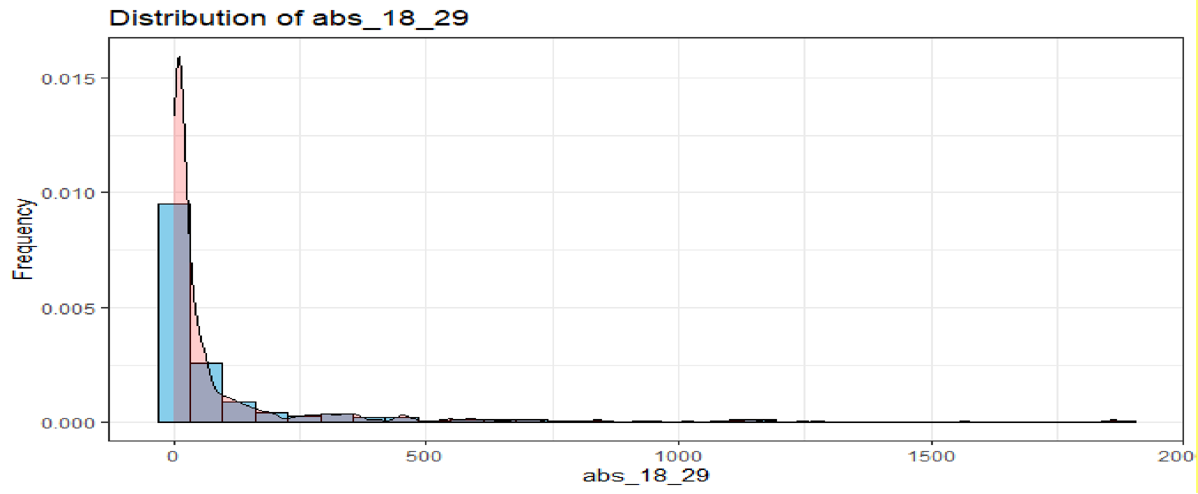
Why RQ is of interest (research gap and future directions according to the literature)

The research question on the relationship between COVID-19 cases in the 18-29 age group and population size in Malaysia addresses an important gap in the already existing literature. Some studies have been conducted on demographic factors influencing COVID-19 outcomes. But these studies have not focused on young adults, especially in the age group 18-25. Understanding this relationship is important because it helps inform targeted public health strategies and interventions for this demographic. Future research should focus on the socioeconomic factors, mental health, and behavioural responses during the pandemic on young adults. These ideas will help to form approaches to improve strategies for any future health crises among the young populations.

Visualisation

Appropriate plot for the RQ

~~~
# Visualisation

```{r}

ggplot(df, aes(x = abs_18_29)) +

  geom_histogram(bins = 30, aes(y = ..density..), fill = "skyblue", color = "black") +

  geom_density(alpha = .2, fill = "red") +

  labs(title = "Distribution of abs_18_29", x = "abs_18_29", y = "Frequency") +

  theme_bw()

ggsave("histogram_abs_18_29.png", width=8, height=6)
~~~



The chart used in this analysis is a histogram that shows the spread of absolute COVID-19 cases in the age group 18-29 years. This plot displays the frequency of case counts across different ranges., It is very easy to evaluate patterns, skewness, and possible outliers in the data for further analysis.

Useful information for the data understanding

The histogram of "abs_18_29" shows a right-skewed distribution. It peaking near zero with a long tail extending to higher values. Most of the observations cluster around the low counts. This further indicates that there is a high frequency of low values and a low frequency for high values. The plot also compares different distributions of the variable.

Analysis

Statistical test used to test the hypotheses and output

The Spearman's Rho test is selected to determine the relationship between 'abs_18_29' and 'capita_18_29'. It is appropriate for non-parametric data and allows for rank-based analysis without an assumption of normality. The Spearman rho for this analysis is approximately 0.676. This shows a moderate to strong positive correlation. The test statistic (S) is 4,857,102, with a p-value less than 2.2e-16, indicating the variables are statistically related at the 0.05 level.

The null hypothesis is rejected /not rejected based on the p-value

The null hypothesis, H0, is rejected. This means that there is a statistically significant relationship between the absolute number of COVID-19 cases in the age group 18-29 years ('abs_18_29') and the population size of the same age group ('capita_18_29'). The Spearman correlation coefficient of about 0.676 and a p-value less than 2.2e-16. This shows an increase in the population size of young adults, the cases of COVID-19 also tend to rise. This highlights the importance of considering demographic factors in public health strategies which are aimed at managing COVID-19 within this age group.

Evaluation – group's experience at 7COM1079

What went well

Our group worked well in collaboration to analyse the COVID-19 data. It harnessed our different skills of research, data analysis, and visualization. The communication helped sharing ideas and feedback that greatly enhanced the quality of our findings. Also, the statistical tests like Spearman's Rho gave strong insights into how case numbers are linked with population size. Then the possibility of creating clear visualizations also helped in analysing complex information.

Points for improvement

While our group did a good job overall, we could have done better at the first stage of planning to divide the tasks more effectively. Some of the group members felt burdened by their tasks because roles were not clearly defined. Also, we faced some problems with the interpretation of data that could have been avoided if it was discussed earlier. This will also help future projects. As by enhancing our time management skills, we would have enough time for revisions and for doing practice presentations that will make the final product very polished.

Group’s time management

Our group managed time reasonably well. But we faced some pressures arising from overlapping deadlines. We could do better with setting clearer milestones and regular check-ins to make sure that everything is managed according to the deadlines. This would enhance our efficiency and reduce stress if we allocate specific time slots for each phase of the project.

Project’s overall judgement

Overall, the project went well. It then provided very important insights on the trends of COVID-19 among young adults in Malaysia. By Combining rigorous analysis with effective communication resulted in a compelling presentation. Though there were minor challenges, our collaborative efforts showed the growth in both individual skills and group dynamics. This helped setting a good foundation for future projects.

Comment on the GitHub log output **(50 words)**

The GitHub log output shows that the project was successfully uploaded. The commit history is very clean and organized. This is good version control practice. As it helps to easily tracks changes and collaborate on the code. But, the lack of detailed commit messages limits understanding of specific updates made throughout the project.

Conclusions

Results explained

These results indicate the moderate-to-strong positive relation between the absolute number of COVID-19 cases within the age group 18-29 and their population size. The Spearman correlation coefficient of about 0.676 suggest that increasing population sizes are associated with higher numbers of cases. The p-value is also statistically significant. It also shows that association is real, which targeted public health interventions taking the population dynamics into consideration is a necessity in the management of COVID-19 among the young adult population in Malaysia.

Interpretation of the results

The analysis shows a strong correlation between the absolute number of COVID-19 cases in the 18-29 age group and their population size. This also suggests that with an increase in the young adult population, the COVID-19 cases rise among the young adult population. These findings have very important implications for public health strategies. It also stresses on the importance of targeted interventions toward young adults. So, understanding these dynamics will help provide valuable insights for policies that will address specific challenges that this age group faces in the time of health crises. This will also help in the prevention of pandemic and healthcare resource allocation.

Reasons and/or implications for future work, limitations of your study

The Future studies should focus on reviewing other factors that affect COVID-19 cases in young adults. These factors include socioeconomic status and behavioural patterns. Then the limitations of the study include the focus on a single age group and potential inaccuracies in data. So, addressing these gaps will improve knowledge and public health strategies for wider populations.

Reference list

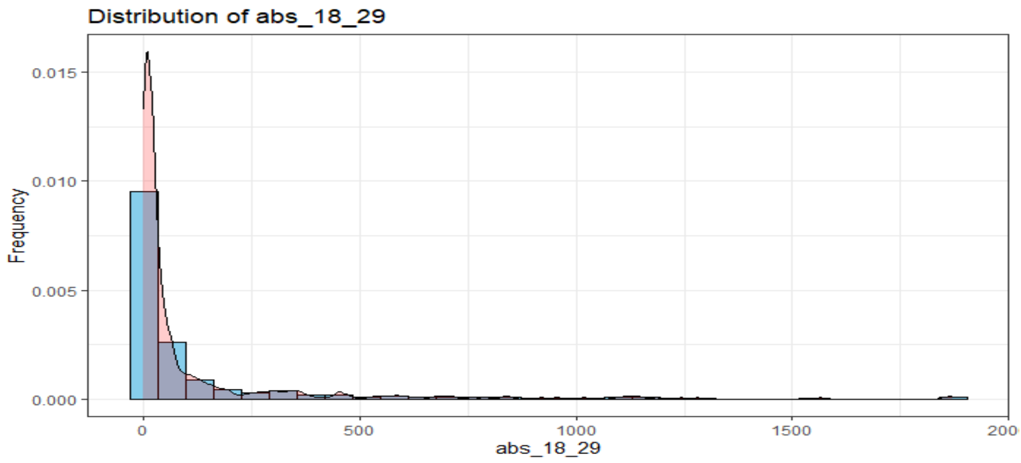
- Chen, Z.-L., Zhang, Q., Lu, Y., Guo, Z.-M., Zhang, X., Zhang, W.-J., Guo, C., Liao, C.-H., Li, Q.-L., Han, X.-H. and Lu, J.-H. (2020). Distribution of the COVID-19 epidemic and correlation with population emigration from Wuhan, China. Chinese Medical Journal, 133(9), pp.1044–1050. doi:https://doi.org/10.1097/cm9.0000000000000782.
- Loo, K.-Y. and Letchumanan, V. (2021). COVID-19: Malaysia's fight against this deadly virus. Progress In Microbes & Molecular Biology, 4(1). doi:https://doi.org/10.36877/pmmb.a0000204.
- Okoli, G.N., Neilson, C.J. and Abou-Setta, A.M. (2022). Correlation between country-level numbers of COVID-19 cases and mortalities, and country-level characteristics: A global study. Scandinavian Journal of Public Health, p.140349482210989. doi:https://doi.org/10.1177/14034948221098925.
- Wong, D.W.S. and Li, Y. (2020). Spreading of COVID-19: Density matters. PLOS ONE, 15(12), p.e0242398. doi:https://doi.org/10.1371/journal.pone.0242398.

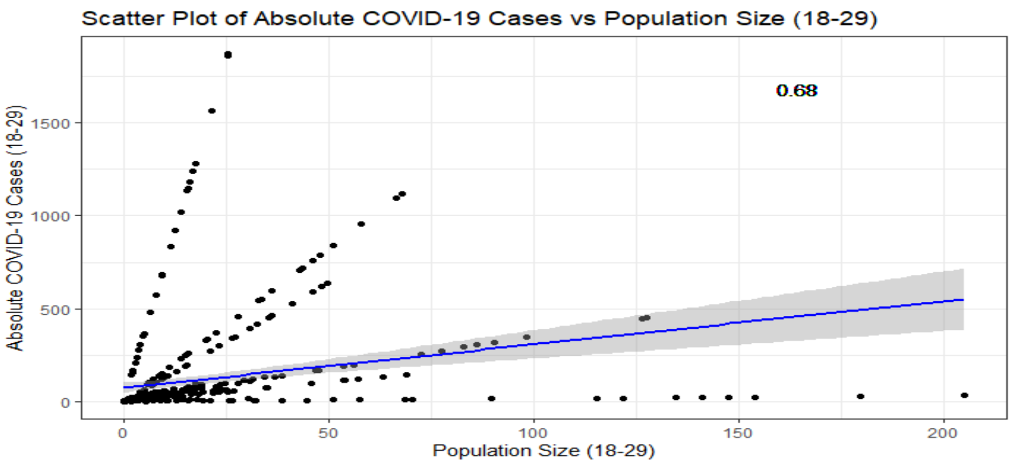
Appendices

A. R code used for analysis and visualisation

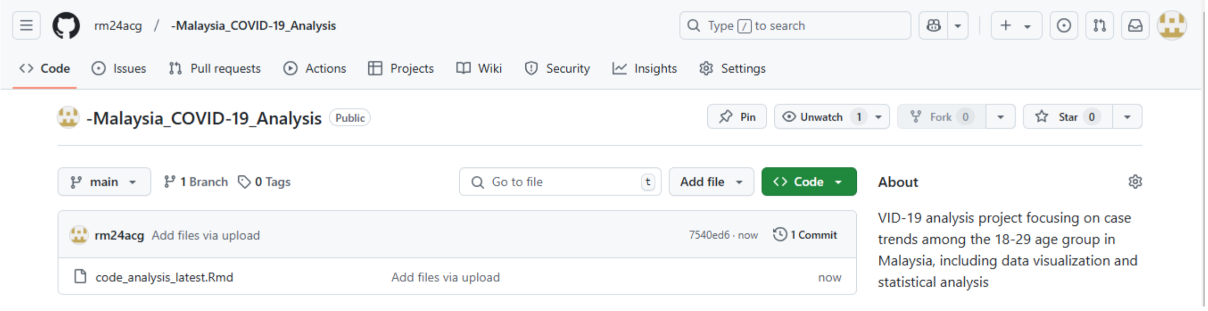
~~~
---
title: "R Notebook"
output: html_notebook
---



# Loading Libraries

```{r}
library(ggplot2)
library(tidyverse)
library(readr)
```
# Loading Datatset
```{r}
df <- read_csv("data/cases_age.csv")
head(df)
```
# Data Cleaning

```{r}
colSums(is.na(df))
```


```{r}
df <- na.omit(df)
head(df)
colSums(is.na(df))
```


# Visualisation
```{r}
ggplot(df, aes(x = abs_18_29)) +
  geom_histogram(bins = 30, aes(y = ..density..), fill = "skyblue", color = "black") +
~~~

~~~
  geom_density(alpha = .2, fill = "red") +
  labs(title = "Distribution of abs_18_29", x = "abs_18_29", y = "Frequency") +
  theme_bw()
ggsave("histogram_abs_18_29.png", width=8, height=6)
```


# Checking Correlation through spearman's test
```{r}
correlation_test <- cor.test(df$abs_18_29, df$capita_18_29, method = "spearman")
print(correlation_test)
```


# Scatter plot
```{r}
ggplot(df, aes(x = capita_18_29, y = abs_18_29)) +
  geom_point() +
  geom_smooth(method = "lm", se = TRUE, color = "blue") +
  labs(title = "Scatter Plot of Absolute COVID-19 Cases vs Population Size (18-29)",
      x = "Population Size (18-29)",
      y = "Absolute COVID-19 Cases (18-29)") +
      theme_bw() +
  geom_text(aes(label = sprintf("%.2f", cor(df$abs_18_29, df$capita_18_29, method =
"spearman"))), x = 0.8 * max(df$capita_18_29), y = 0.9 * max(df$abs_18_29))
```
~~~





B. GitHub log output.



rm24acg Add files via upload 7540ed6 · now 1 Commit

code_analysis_latest.Rmd Add files via upload now

VID-19 analysis project focusing on case trends among the 18-29 age group in Malaysia, including data visualization and statistical analysis