UNIVERSITI SAINS MALAYSIA

Rancangan Kokurikulum

WMU 122 Data Science Literacy

Semester II, Academic Session: 2021/2022

Final Report

G08

Analysis of Covid-19 Patients

Names	Matric No	USM Email Addresses	Module	Role	School
NURUL AIN BINTI JALIL	157619	aienjalil01@student.usm.my	Data Analytical Dashboard	Leader	Physics
NURUL AIN IRDINA BINTI NOROL AZHAR	158513	ainirdina@student.usm.my	Data Analytical Dashboard	Member	Physics
KELLY CHONG HO YEE	157371	kellychong@student.usm.my	Data Preprocessing & Preparation	Member	Physics
ANG KAI YEN	156797	kaiyen@student.usm.my	Data Cleaning	Member	Management

TABLE OF CONTENTS

INTRODUCTION

How many persons of each age group have recovered from COVID-19 in Malaysia? What clusters in Malaysia have been impacted by COVID-19? In Malaysia, how many individuals are partially vaccinated, completely immunised, or receive a booster vaccine every day? Such concerns have significant public health consequences. Since the discovery of a new coronavirus in China in late December 2019, the virus has swiftly spread to become an unprecedented pandemic, affecting the whole world. Until December 3rd, 2021, the World Health Organisation announced that COVID-19 has spread to practically every country in the world, with 263,563,622 confirmed cases and 5,232,562 deaths. On January 25, 2020, Malaysia's first three cases of COVID19 were verified as imported cases. Since then, the Malaysian government has been working hard to reduce and limit the disease's spread through a variety of strategies and collaboration efforts, including the implementation of a series of movement control orders (MCO), a vaccination programme, and the development of a national recovery plan. The total number of COVID-19 cases in Malaysia was 53,633,304 from June 1, 2021 to May 31, 2022. Is data reporting from Malaysia's national and subnational administrations detailed enough to address such questions? In this study, we focus on data information by recording and analysing the reporting of surveillance data, total of cases of each group, clusters that were impacted, sums of recovery, new and active cases, and vaccination monitoring data during the second wave of COVID-19. Age distribution for recovery cases, new cases, and active cases following immunisation stratified by vaccine, which is how many people in Malaysia were fully, partly, and additionally boosted daily.

This unusual COVID-19 pandemic has posed a significant challenge to policymakers throughout the world in terms of developing successful solutions depending on local disease spread, healthcare system resources, economic and political variables, and public perception. It's critical to understand how illness spreads in cluster infections. It not only aids in assessing the present pandemic scenario, but it also provides a basis for establishing effective and strategic prevention and control measures, such as vaccine planning, targeted screening for high-risk populations, and strategic health education and promotion campaigns. The description of COVID-19 clusters in Malaysia has lacked evidence up to this point. This study aims to analyse COVID-19 in Malaysia in terms of the number of new cases, active cases, recovery cases, cases in each age group, cluster types, and vaccination data such as partially vaccinated, completely vaccinated, and booster for daily use in Malaysia. The findings of this study might be used by governments in Malaysia to develop COVID-19 preventive and control efforts. The COVID-19 data must be reported for the following reasons. It allows health experts to follow the spread of COVID-19 throughout age groups, the type of clusters that are impacted, vaccination data monitoring, and the number of cases reported daily. It also allows researchers to gain new insights, and scrutinise the data to understand the rationale behind the policies put forth by the government.

DATASET

The data used in this analysis consists of the daily and cumulative incidence (confirmed cases) of COVID-19 Malaysia. For the cases which include new cases, recoveries and active cases were taken from the year 2021 to 2022. The data for Italy was obtained from where the raw data was sourced from the Ministry of Health Malaysia (MoH) website named COVIDNOW. The starting dates for sets of data indicate the dates on which the cases were confirmed in this country, however, it should be noted that in some cases were not confirmed until after 5p.m in the evening every day. In the remainder of this section, we provide a simple exploratory analysis of the incidence data.

In September 2021, The Ministry of Health (MoH) launched a simplified platform that provide all daily information related to Covid-19 including more granular data that was previously unavailable in a bid to increase transparency. The website covidnow.moh.gov.my still employs mostly the same indicators used from the previous reporting practice, but with the addition of new and detailed information like the number of daily tests and patients treated at low-risk treatment and quarantine centres (PKRC).The ministry had also tweaked some of the denominators used previously, like reporting the number of coronavirus-related deaths based on the time of death instead of reporting only when deaths are confirmed, which had caused backlogs that drove the daily numbers up.

It was via this spirit of collaboration with the open data community that COVIDNOW was born. A group of innovative developers worked with us at MoH, 100% pro bono, to build a site that would help Malaysians monitor the epidemic. This work was led by the Crisis Preparedness and Response Centre (CPRC), in conjunction with CPRC Hospital and the National Public Health Laboratory (NPHL). MOH will also make more granular data available on the open sharing platform Github, a continuation of past practice. Malaysia was also recognised as the first few countries to share its data on the website, now considered a go-to open-source platform that gives scientists, epidemiologists and public health experts crucial information to understand the virus and how to respond to it.

Data collection method is done through an app called MySejahtera. MySejahtera is a mobile application developed by the Government of Malaysia to facilitate contact tracing efforts in response to the COVID-19 pandemic in Malaysia. The main goal is quick identification of persons who may have come into close contact with anyone who has tested positive for COVID-19.

The ethical appropriateness during data collection is three principles which are time limitation, use restriction and security.

- ✓ Time Limitation

All measures shall be temporary in nature and limited in scope. If governments and health systems expand monitoring and surveillance powers then such powers should be time bound, and only continue for as long as necessary to address the current pandemic. Measures should be fully withdrawn at the earliest moment after the epidemic has ended locally. There are legitimate concerns that digital proximity tracking will be unnecessary yet will remain in place.

- ✓ Restriction

The sale and use of data for commercial purposes or advertising activities should be strictly prohibited. Recognizing that governments may have existing data protection laws and frameworks already in place, the sharing of data with government departments, agencies or third parties that are not involved in the public health response should be prohibited. The sharing of data with law enforcement or immigration departments or agencies should also be prohibited.

- ✓ Security

Privacy and the principles included in this guidance. Every effort should be made to ensure high security, including encryption, of any personal data or health data collected and of any devices, applications, servers, networks, or services involved in collection, transmission, processing, and storage. Applications should be subject to third-party audits and penetration testing, and developers should publish full details about their security protocols.

DATA PREPROCESSING AND PREPARATION

Finding quality data, data cleaning and data transformation are all processes in the data preparation process. Finding the proper data is the first step in the data preparation process. It is critical to discover each dataset after gathering the data. This step is discovering the data in preparation for the data quality review, which will guarantee that it is of good quality, reliable, and relevant to our project. Following the creation of an appropriate data collection, a data cleaning procedure is required to eliminate any unnecessary data, fill in missing values, and remove extraneous data. Transforming data is changing the format or value inputs in order to achieve a certain result or make the data more understandable to a diverse audience.

For data pre-processing and preparation, we extract, transform and load data using Power query, a data transformation and data preparation engine. Power Query has a graphical interface for fetching data from data sources and a Power Query editor for implementing transformations. We remove the rows of data that we do not need to analyse. We filter the dates to one year, from June 2021 to May 2022, so that we can analyse it more easily on a month-by-month or quarter-by-quarter basis. In this report, we used two datasets for the analysis. The first dataset includes confirmed cases and the second dataset includes the number of people who have been vaccinated. The data set, which comprised confirmed cases and the number of vaccines, the data was clean in terms of missing or conflicting numbers.

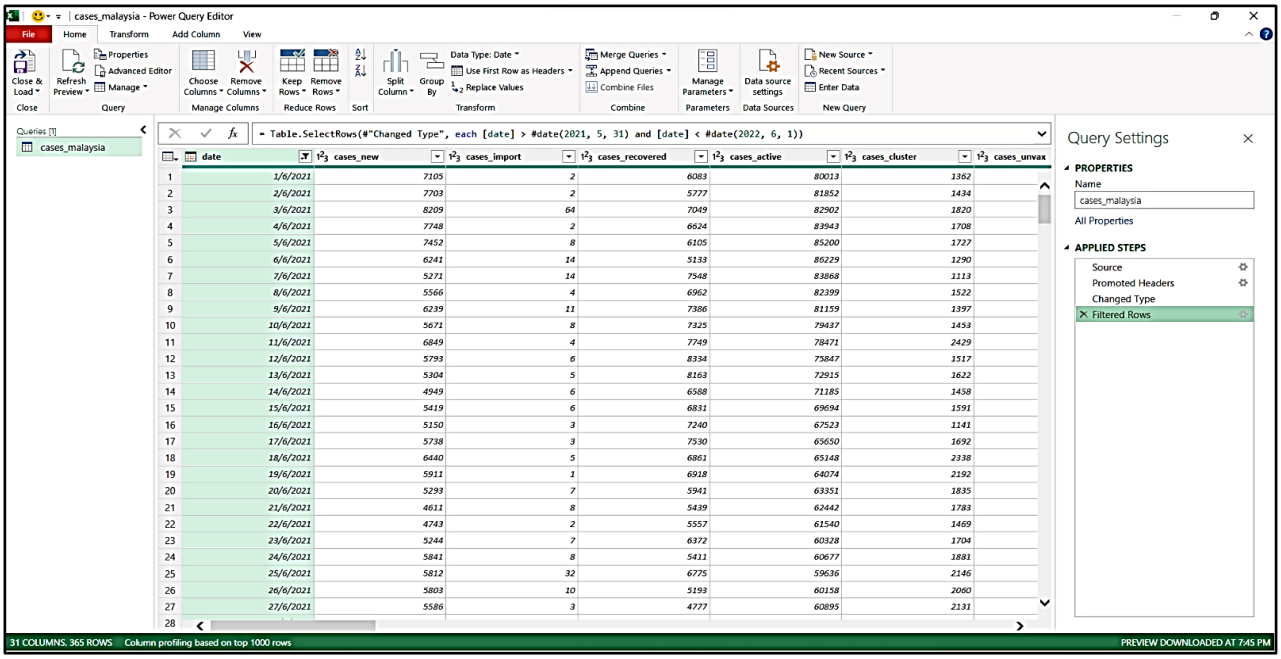


Figure 1: Data set of confirmed cases from 1st June 2021 to 31st May 2022

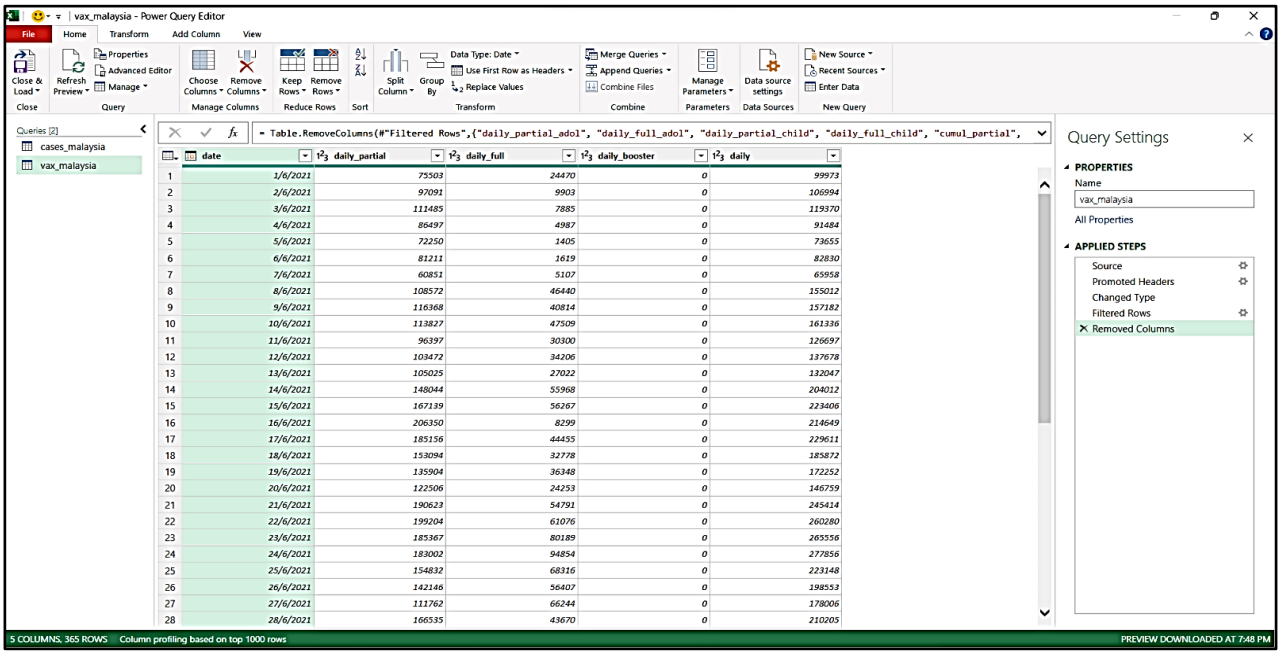


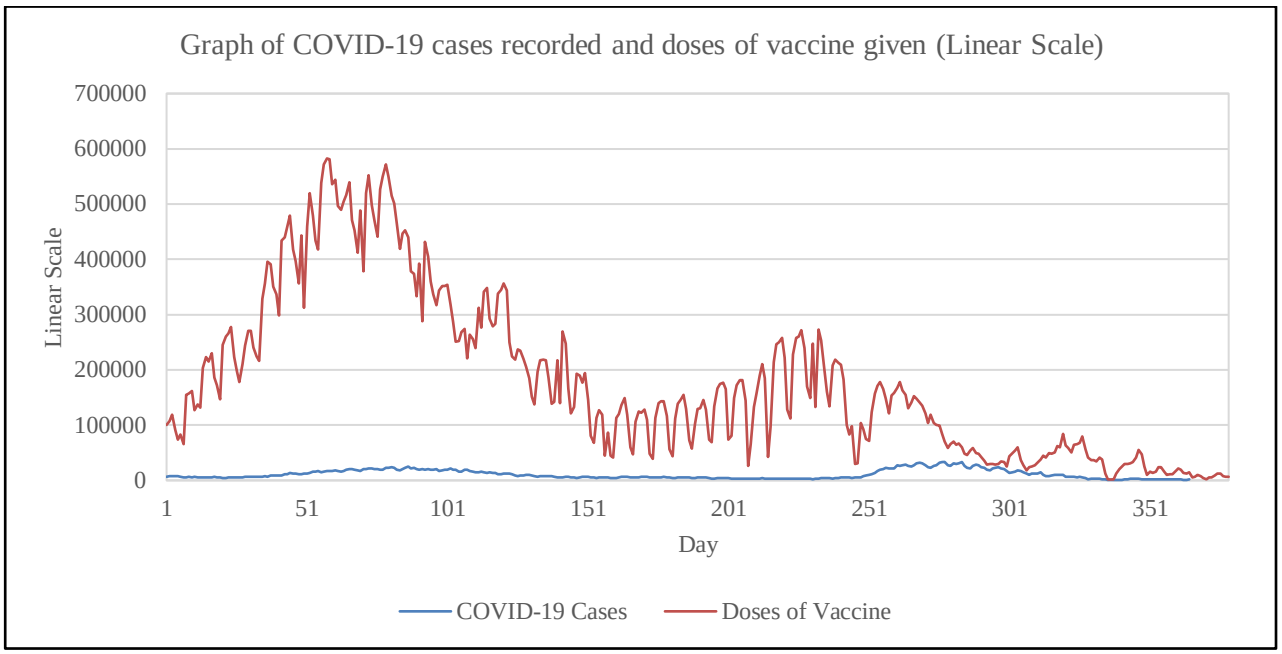
Figure 2: Data set of number of doses of vaccination from 1st June 2021 to 31st May 2022

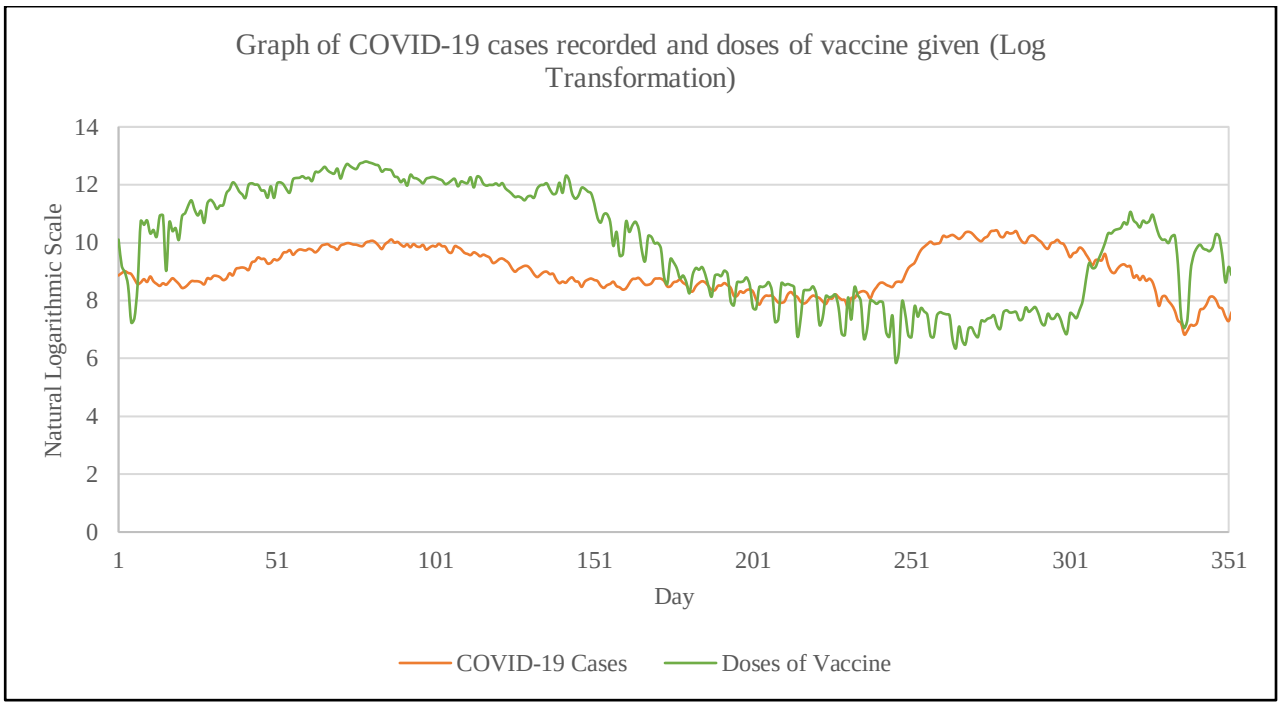
DATA CLEANING

Natural Log Transformation

The vertical axis of our charts is shown using a natural logarithmic scale. Log scales are particularly suited to display trends in relative rates of change, like a virus spreading in this case. By comparing the slopes of two lines, a log scale allows us to compare epidemics at a very early stage with those that are much more advanced, even though they have very different absolute numbers of cases or deaths.

On a natural log scale, an epidemic looks like a steep diagonal line that flattens towards a horizontal line as its rate of growth slows. On the more familiar linear scale, the same data looks like a hockey stick shooting upwards, which gives a better sense of the overall size of its epidemic.





Age Group	Child	Adolescent	Adult	Elderly
Jun	21652	11392	124143	15596
Jul	46552	23169	253662	24755
Aug	87654	46416	432151	46673
Sep	77984	42030	322839	48846
Oct	38950	14885	141988	28048
Nov	27191	6025	106037	20929
Dec	19406	5418	84501	15298
Jan	16009	11057	74385	10660
Feb	77560	40528	408066	44069
Mar	95237	31354	552031	75710
Apr	33621	19717	167832	21796
May	5540	4358	42726	5755

Table 1: COVID-19 cases by age group

Clusters	Import	Religious	Community	High Risk	Education	Detention Centre	Workplace
Jun	55	1373	7529	971	586	2124	28251
Jul	12	170	14728	2442	766	4279	40731
Aug	0	508	16763	1805	565	3024	38086
Sep	24	106	11175	1319	847	1090	23702
Oct	0	124	3376	918	806	94	6145
Nov	0	92	476	363	1079	53	2679
Dec	34	89	296	384	1023	86	2151
Jan	125	196	262	238	6890	160	1941
Feb	7	17	203	730	9451	519	3134
Mar	0	0	75	1124	3191	555	2653
Apr	0	0	1	185	3197	0	126
May	0	0	11	23	1	6	0

Table 2: COVID-19 cases by clusters

Vaccination Status	Unvaccinated	Partially Vaccinated	Fully Vaccinated	Booster
Jun	169059	7159	3404	0
Jul	281269	63065	16959	0
Aug	344633	194842	93507	0
Sep	195131	101605	202705	0
Oct	65441	20350	139995	161
Nov	39843	2725	116100	2472
Dec	28674	963	90449	5218
Jan	23199	629	72472	16372
Feb	102444	4630	255754	209150
Mar	110021	20397	203924	424841
Apr	34244	11539	50628	149674
May	5811	1226	11681	39788

Table 3: COVID-19 cases by vaccination status

Vaccination Status	Partially Vaccinated	Fully Vaccinated	Booster
Jun	4013082	1247968	0
Jul	8244804	4566906	0
Aug	6015561	8418551	0
Sep	3827692	5566374	251
Oct	1489407	3912670	319300
Nov	174646	669820	2161967
Dec	102542	162373	3845814
Jan	53053	80480	5728178
Feb	943020	38232	2651682
Mar	501878	53674	1119929
Apr	263043	924668	214994
May	211268	344608	61292

Table 4: Vaccination Status

Vaccination Status	New Cases	Recovered Cases
Jun	179622	192642
Jul	361293	231958
Aug	632982	547086
Sep	499441	593100
Oct	225947	319938
Nov	161140	162443
Dec	125304	148174
Jan	112672	98625
Feb	571978	332561
Mar	759183	843518
Apr	246085	406869
May	58506	80731

Table 5: New and recovered COVID-19 cases

DASHBOARD

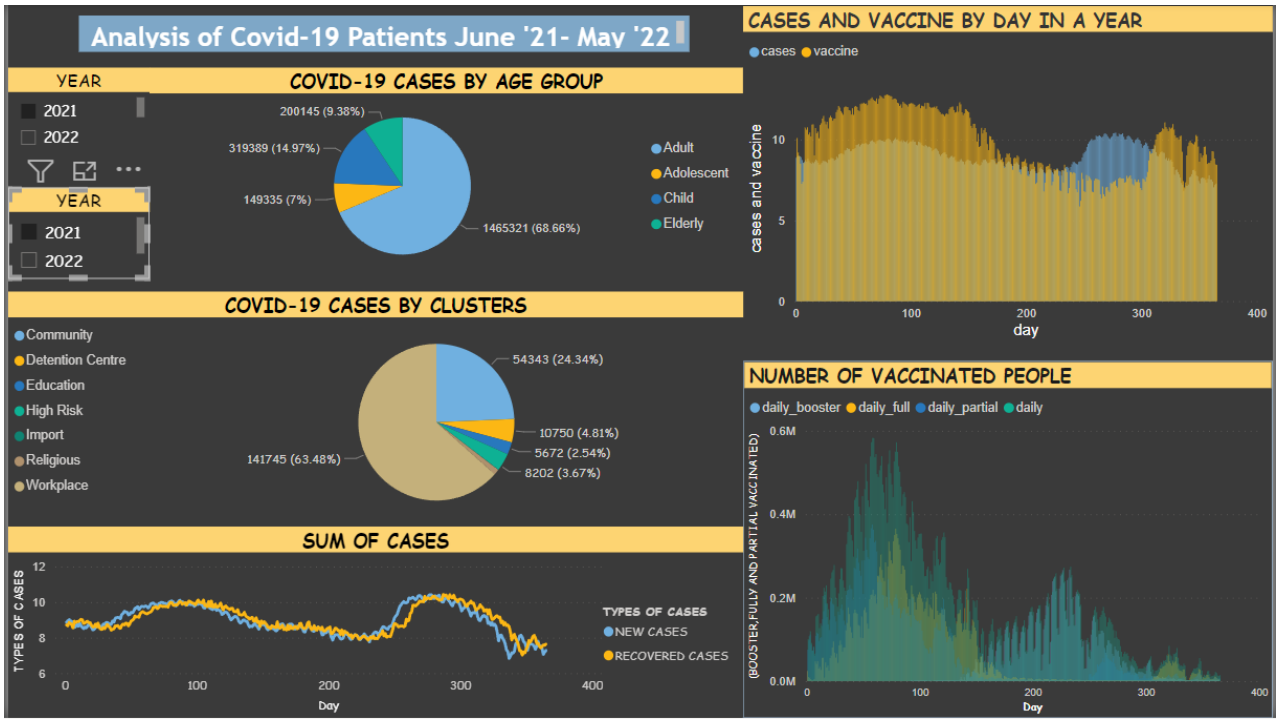


Figure 3: Dashboard of COVID-19 in 2021

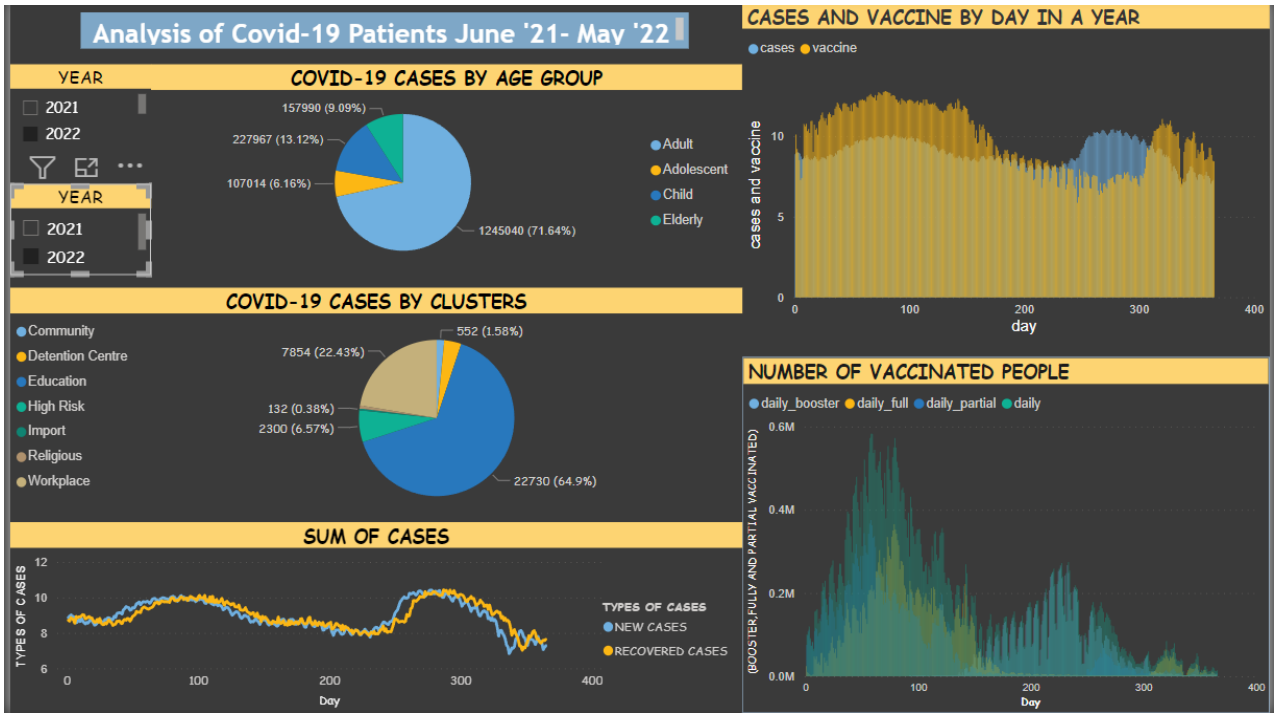
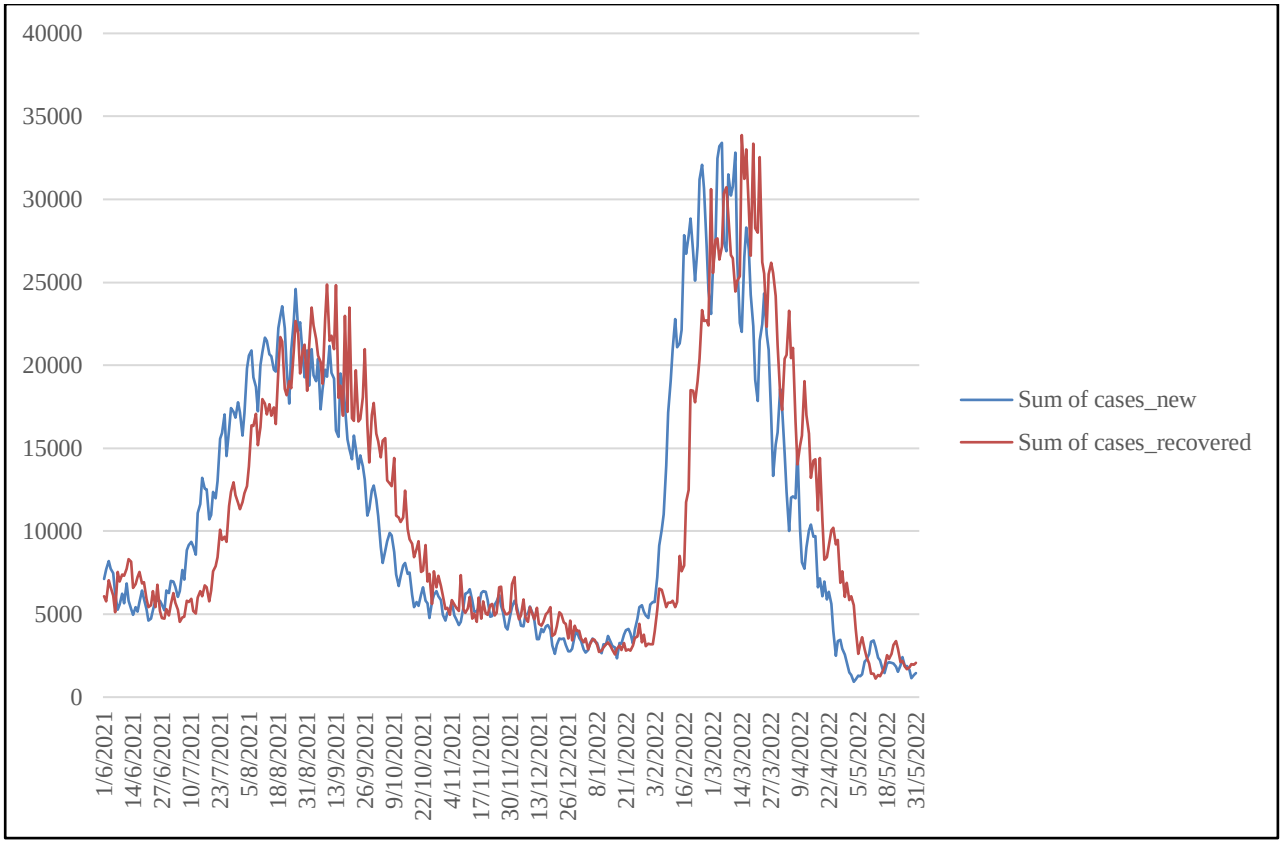
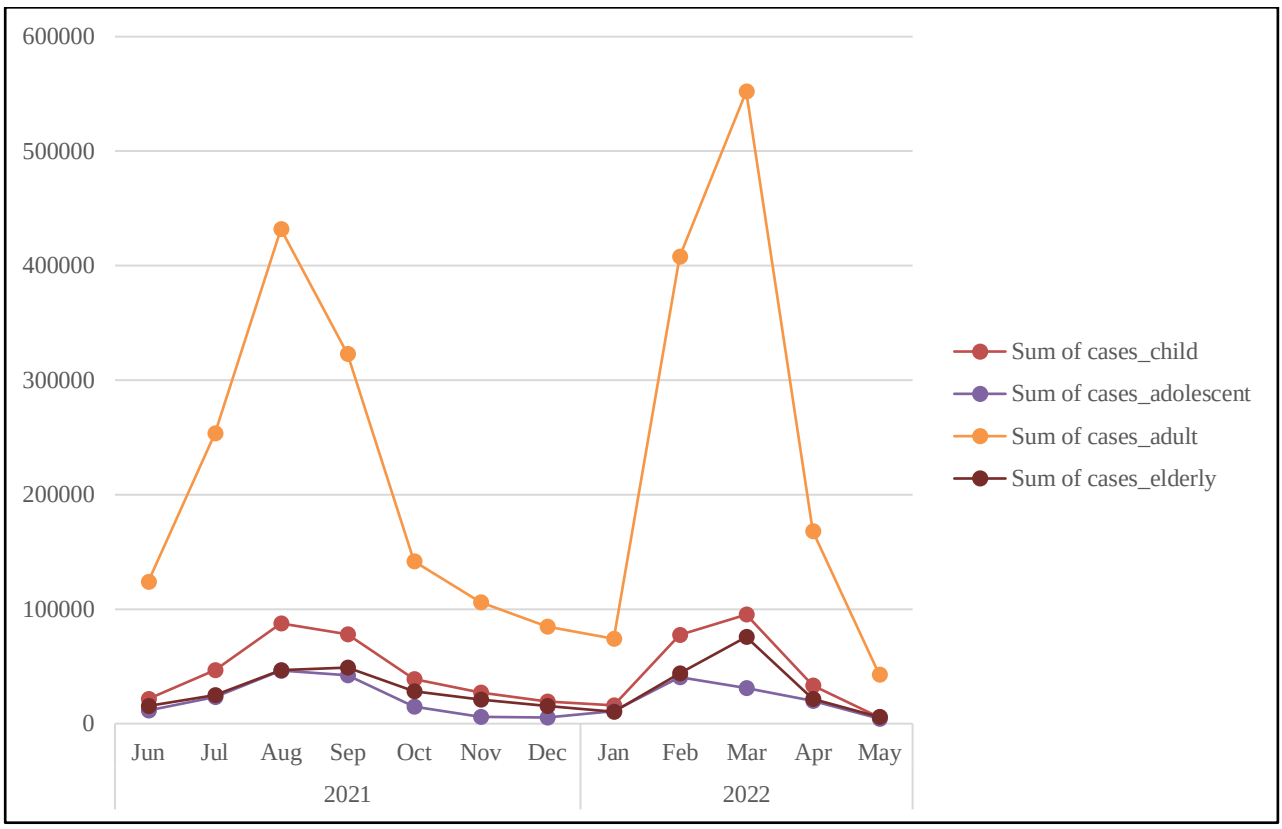


Figure 4: Dashboard of COVID-19 in 2022

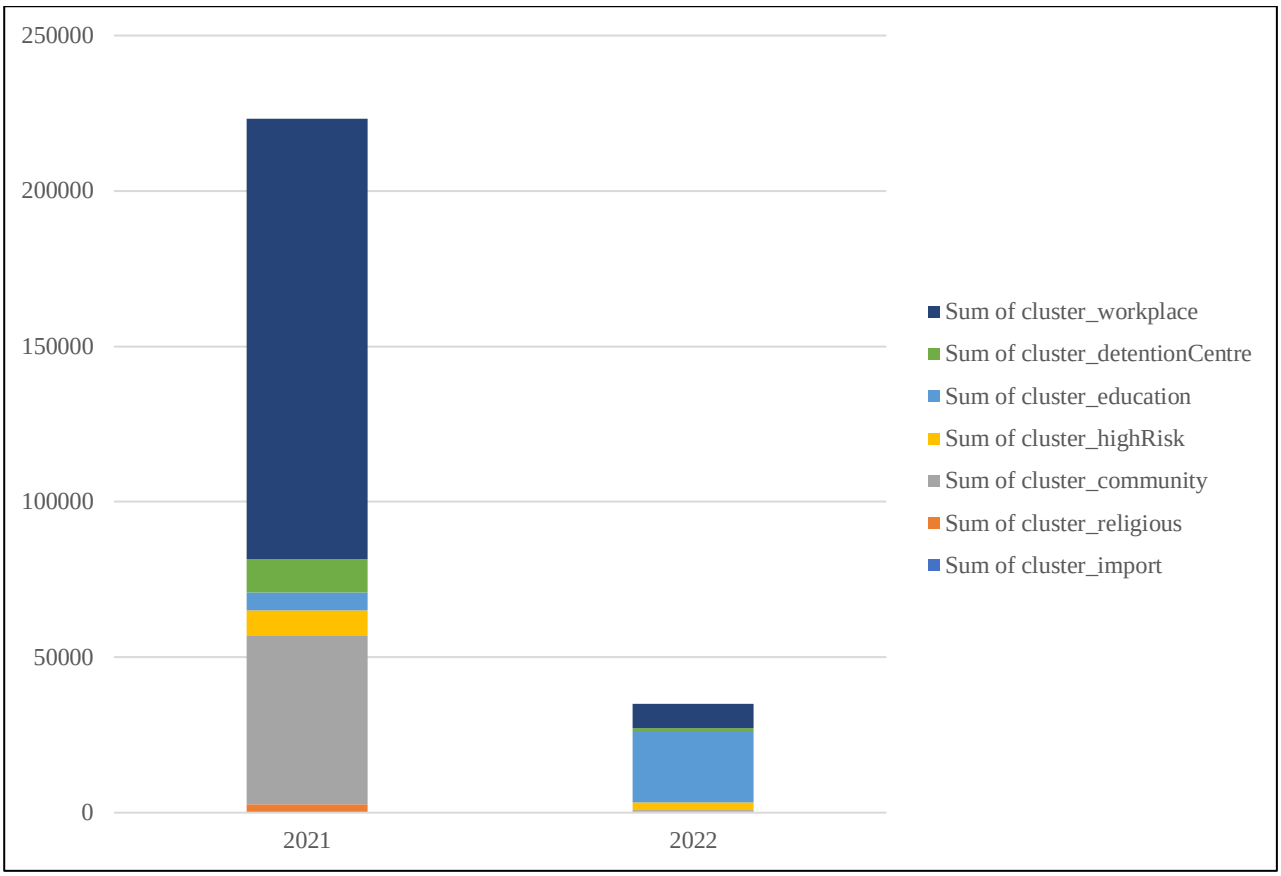
PROPOSED SOLUTION



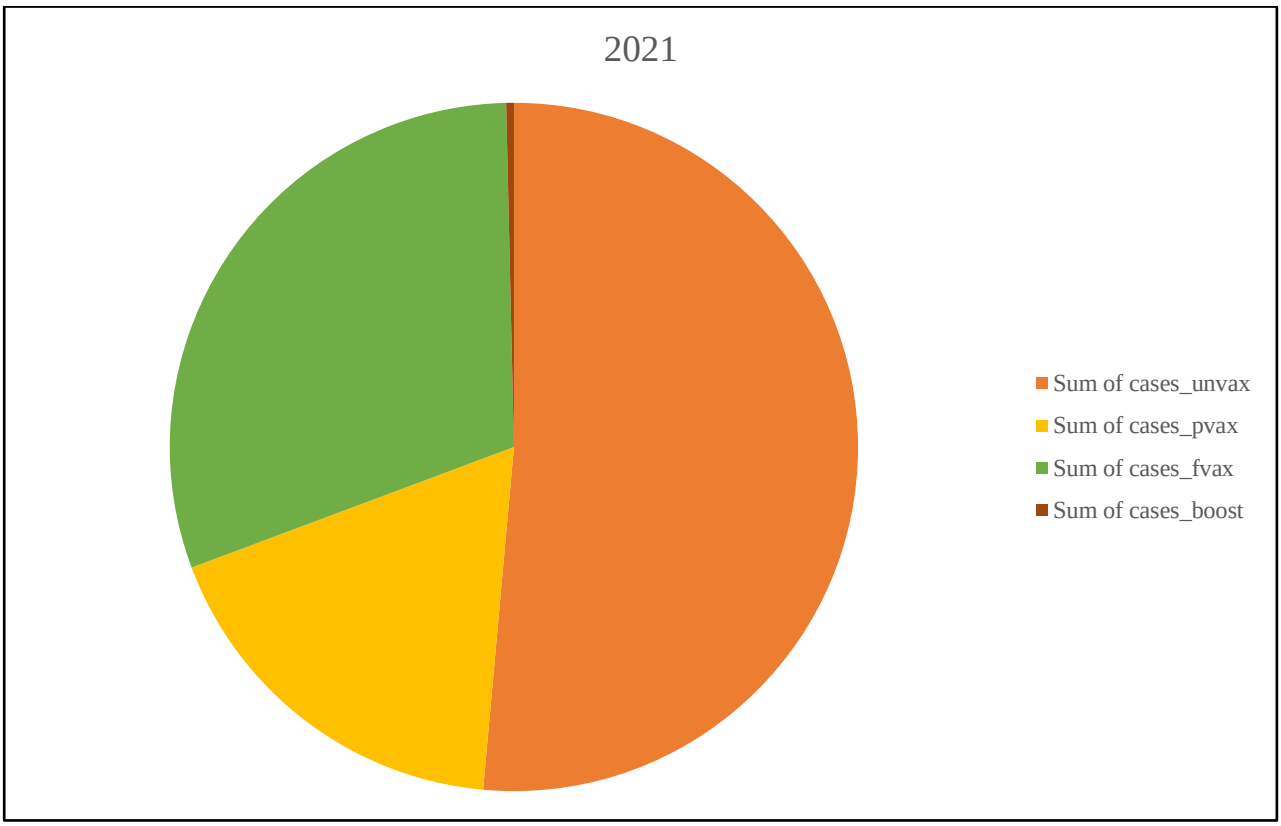
Graph 1: COVID-19 cases from June 2021 until May 2022



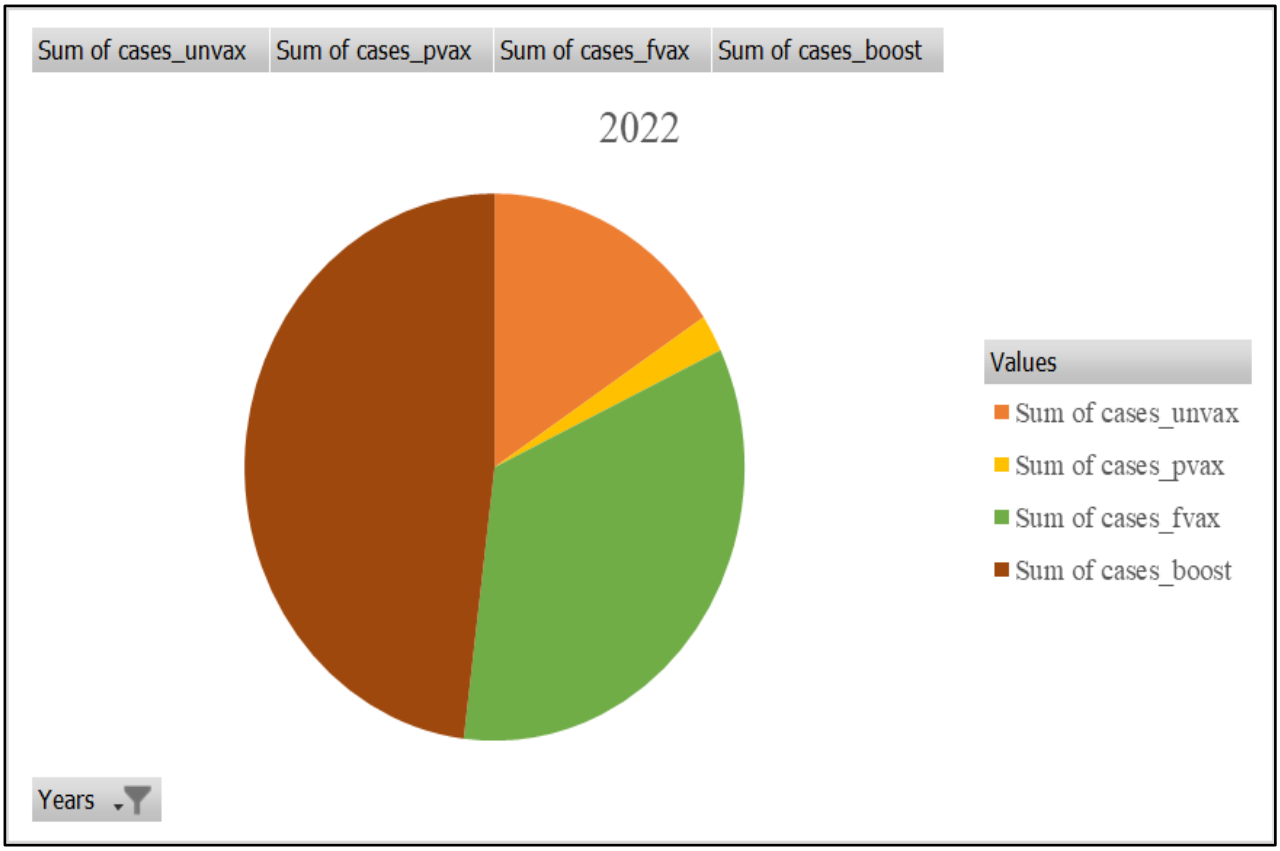
Graph 2: COVID-19 cases by age group



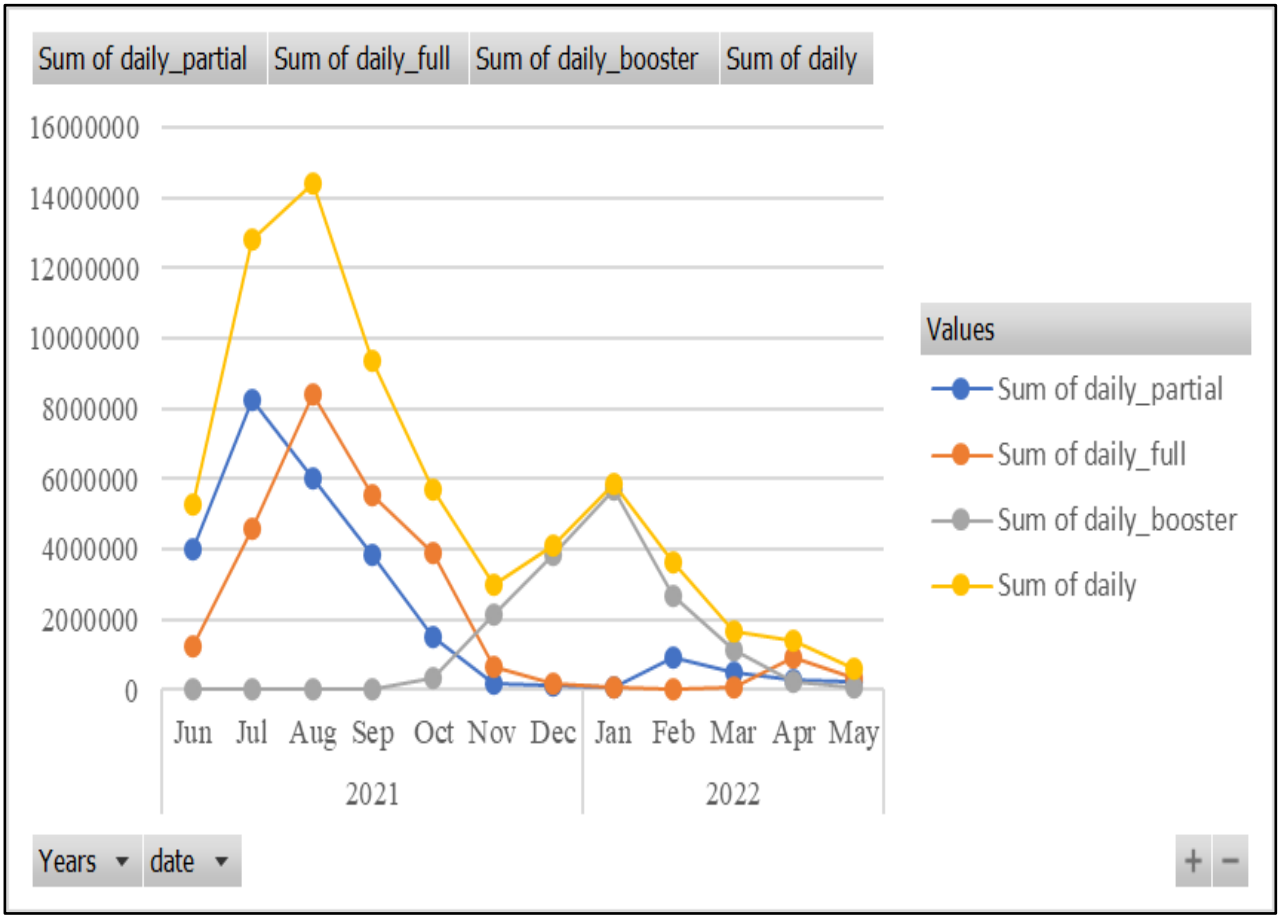
Graph 3: COVID-19 cases by clusters



Graph 4: COVID-19 cases by vaccination status in 2021



Graph 5: COVID-19 cases by vaccination status in 2022



Graph 6: Vaccination status from June 2021 until May 2022

COVID-19 cases by age group

In 2021, it is clear that the majority people affected are the adult with percentage 68.66% following by the child (14.97%), the elder (9.38%) and the adolescent (7%)**.** In 2022**,** there is no difference for the data in 2022 as the adult still has the highest number of people affected with percentage 68.66% following by the child (13.12%), the elder (9.09%) and the adolescent (6.16%)**.**

COVID-19 cases by clusters

In 2021, the workplace cluster has the highest number of people affected followed by community cluster, detention centre cluster, high risk cluster, education cluster and religious cluster. In comparison, by 2022 the Education cluster took the highest place in the number of people affected with percentage (64.9%). Followed by workplace cluster (22.43%), high risk cluster (6.57%), community cluster (1.58%) and lastly, religious cluster (0.38%).

COVID-19 cases and recovered cases

The trend of line graphs for new cases and recovered cases are quite the same. The yellow line illustrates the new cases while the blue line illustrates the new cases. In the early of June 2021 they had 8.87 percent for new cases and 8.71 percent for recovered cases. Then, in the middle of September 2021 the line went up to 9.83 percent for new cases and 9.85 percent for recovered cases. In march 2022, the trend line is at the highest peak for both which are 10.19 percent for new cases and 10.35 for recovered cases. Lastly, the line went down in May 2022 with the percentage 7.59 percent for new cases and 7.86 percent for recovered cases.

Vaccination Status

The graph illustrates trends of the vaccination status in Malaysia from 1st June 2021 until 31st May 2022. It presents four differences on how many people get vaccinated; partially, fully, booster and daily. Overall, all threes for sum of daily partial vaccinated, daily fully vaccinated and daily booster have inclined starting from June 2021. Then, in August 2021 the yellow line and the orange line steadily declined until November 2021 while the blue line has slumped since July 2021. In August 2021, the line graph for the sum of daily vaccinated and sum of daily fully vaccinated got the highest peak which are 14434112 cases and 8418551 cases. As the sum of daily fully vaccinated and sum of daily partial vaccinated began to fall from November 2021, the grey line started to rise steadily from October 2021 and started to drop down dramatically from January 2022 until May 2022. All the line graphs started to decrease continuously starting in March 2022 except for the orange line graph which increased a little bit until April 2022. The beginning of April 2022, all the graphs declined until the end of May 2022.

COVID-19 cases and vaccination status

Generally, for this graph, we can clearly see the effectiveness of the vaccine in order to control the covid-19 transmission. The percentage of new cases is maintained lower than the number of people vaccinated from June 2021 to May 2021. However, in March 2023 the percentage of new cases is higher than the number of people vaccinated. This happens because most Malaysians are already vaccinated with the percentage average 85%.

CONCLUSION

Government chooses to keep some data analytics elements hidden from the scientific community and the general public by not revealing them. One promising area for future study is to investigate what factors influence how Malaysian governments provide COVID-19 cases and vaccine monitoring data. The purpose of our project, as researchers, is to advocate for change via measurement. We demonstrated the nature and quantity of missing data across Malaysia using data cleaning and data transformation as all these processes were used for the data preparation process. Our findings give the most comprehensive and latest evidence of missing data in Malaysia's COVID-19 data reporting. Before the next wave of COVID-19, Malaysian governments should acknowledge the importance of data reporting and make it a priority. We recommend reporting the following as a starting point. First, age and cluster distribution for new, recovered and active cases. Second, vaccination coverage is divided according to the eligibility group for each dose.

The ever-changing picture of COVID-19 and its epidemiology shows no indications of slowing down. Continuous recording and analysis of data COVID-19 such as age group, types of clusters, vaccination status, sum of cases for recovery and new cases, will aid in analysing disease control locally while also permitting benchmarking efforts against other nations. Much of the COVID-19 environment remains unrecognised, particularly in middle-income countries. As the poor world continues to battle with COVID-19, further study into the disease's epidemiology is urgently needed if the pandemic is to be properly controlled. As the vaccine deployment continues in the developed world, monitoring COVID-19 epidemiology is crucial as we continue to find a way to live with this virus.