**MASTER IN HEALTH DATA COLLECTIONS**

**Intake : MHIA124**

**HIA 322 GROUP PROJECT**

**TITLE :**  
**EXPLORING THE ASSOCIATION OF MYSEJAHTERA APPS EFFECTIVENESS WITH COVID-19 CASES IN MALAYSIA (Jan 2020 – June 2022)**

KHAIRUL IZWAN BIN MOHAMED @ HASHIM 00000047728

CHOW TAK KUAN 00000044142

AN JIFENG 00000047312

**CONTENT**

|  |  |  |
| --- | --- | --- |
|  |  | **Page** |
| **1.** | Introduction   * 1. Data Source   2. Objective | **3** |
| **2.** | Methodology   * 1. Data Collection   2. Data preprocessing   3. Descriptive analysis   4. Statistical analysis | **4**  **5**  **5**  **6** |
| **3.** | Results   * 1. Daily check-ins with daily new Covid-19 cases   2. Daily check-ins with unvaccinated cases   3. Daily check-ins with cluster cases |  |
| **4.** | Discussion   * 1. Descriptive Analysis   2. Statistical Analysis |  |
| **5.** | Conclusion | **13** |
| **6.** | References | **13** |
| **7.** | Appendix   * 1. Table 1.0   2. Table 2.0   3. Table 3.0   4. Table 4.0   5. Figure 1   6. Figure 2 |  |
|  |  |  |

**1.0 Introduction**

The COVID-19 pandemic has affected the whole wide world. It brought challenges that required countries to adapt quickly with creative solutions to manage and control the spread of the novel virus. In Malaysia, the MySejahtera app has been one of the key tools in this effort. It was launched in April 2020, designed to help the public and authorities to track COVID-19 cases, facilitate contact tracing, manage vaccinations, and allow check-ins at locations through QR codes. These features made it essential for monitoring and managing daily activities during the pandemic.

As the pandemic emerge into national crisis, the implementation of the Movement Control Order (MCO) phases across Malaysia further highlighted the role of MySejahtera app. Starting in March 2020, the MCO was introduced in phases, enforcing travel restrictions, limiting businesses, and controlling gatherings to minimize the virus’s spread. The phases of MCO are as follows:

1. MCO: 18th March 2020 – 3rd May 2020
2. Conditional MCO: 4th May 2020 – 9th June 2020
3. Recovery MCO: 10th June 2020 – 31st March 2021
4. MCO by states: 13th January 2021 – 31st May 2021
5. Total Lockdown: 1st June 2021 – 28th June 2021

During this time, MySejahtera’s check-in function became a critical source of data, helping authorities and health officials monitor population movements and identify potential hotspots and clusters identification. This link between mobility data and rising or falling COVID-19 cases during various MCO phases provides an opportunity to study the patterns and impacts of these measures.

This study aims to explore the relationship between MySejahtera check-ins and the number of COVID-19 cases reported, with a focus on how these patterns were influenced by the different MCO phases. By looking into the data, this research hopes to reveal how public movement and interactions contributed to the spread of the virus. It also seeks to evaluate whether digital check-in systems like MySejahtera can serve as effective tools for managing and predicting outbreaks in the future.

By understanding how MySejahtera data ties into case trends and MCO policies, this study could offer valuable insights for future public health planning. Identifying key trends, areas at higher risk, and patterns in case surges may help authorities respond more effectively and plan targeted interventions in the future. A descriptive analysis of the data has been collected and analyzed. The attributes of the data are listed below:

1. date: yyyy-mm-dd format; data correct as of 2359hrs on that date
2. checkins: number of checkins at all locations registered on MySejahtera
3. unique\_ind: number of unique accounts which checked in
4. unique\_loc: number of unique premises checked into
5. i: in the time density file, checkins are aggregated by half-hour buckets, giving 48 in total; bucket i corresponds to the ith half-hour slot of the day. for instance, i = 0 corresponds to 0000 - 0029; i = 31 corresponds to 1500 - 1529.
6. cases\_new: cases reported in the 24h since the last report
7. cases\_import: imported cases reported in the 24h since the last report
8. cases\_active: Covid+ individuals who have not recovered or died
9. cases\_recovered recovered cases reported in the 24h since the last report
10. cases\_cluster: number of cases attributable to clusters; the difference between cases\_new and the sum of cases attributable to clusters is the number of sporadic cases
11. cluster\_x: cases attributable to clusters under category x; possible values for x are import, religious, community, highRisk, education, detentionCentre, and workplace
12. cases\_agecat: cases falling into one of 4 age categories, i.e. child (0-11), adolescent (12-17), adult (18-59), elderly (60+); note that the sum of cases by age may not equal the total cases for that day, as some cases are registered without ages or with unverifiable age data
13. cases\_pvax: number of partially-vaccinated individuals who tested positive for Covid (perfect subset of cases\_new), where "partially vaccinated" is defined as receiving at least 1 dose of a 2-dose vaccine at least 1 day prior to testing positive, or receiving the Cansino vaccine between 1-27 days before testing positive
14. cases\_fvax: number of fully-vaccinated who tested positive for Covid (perfect subset of cases\_new), where "fully vaccinated" is defined as receiving the 2nd dose of a 2-dose vaccine at least 14 days prior to testing positive, or receiving the Cansino vaccine at least 28 days before testing positive
15. casual\_contacts: number of casual contacts identified and notified by CPRC's automated contact tracing system
16. hide\_large: number of large hotspots identified by CPRC's hotspot identification system
17. hide\_small: number of small hotspots identified by CPRC's hotspot identification system
    1. **Data source**

The data for this study was obtained from the official **COVID-19 Malaysia GitHub repository**, which previously and currently still maintained by the Ministry of Health (MOH) Malaysia. This repository provides comprehensive and publicly accessible datasets on COVID-19 cases, testing, vaccinations, and related statistics. The data includes daily updates on the number of reported cases, recoveries, deaths, and other metrics, as well as information categorized by states and districts.

For this analysis, the dataset utilized was last accessed on 8th December 2024. The specific files used include “cases\_malaysia.csv”, “cases\_state.csv”, “checkin\_malaysia.csv”, “checkin\_state.csv”, “trace\_malaysia.csv” and “clusters.csv”. These CSV files provide detailed information regarding MySejahtera total check-in per day for the whole country and states, and detailed information regarding Covid-19 cases daily update.

The GitHub repository is available at: https://github.com/MoH-Malaysia/. All data was cleaned, processed, and analysed in accordance with ethical standards, and no personal or identifiable information was used.

* 1. **Objective**

**Research question**:

* What is the check-in pattern recorded via MySejahtera app associated with the number of Covid-19 daily new cases in Malaysia?
* How is the MCO affect the reporting of new daily Covid-19 cases in the whole country and the states?
* How effective is the MySejahtera app identify the geolocation of new cluster or new hotspot of outbreak?
* Can check in data trend by MySejahtera app being used as early indicator for prevention of contagious cases?
* How can the check-ins data collected be improved the current MySejahtera app to better predict and manage future pandemic?

**Objective**:

* To determine the association of daily check-in data by MySejahtera app with reported of daily new Covid-19 cases
* To determine the association of daily check-in data by MySejahtera app with reported daily new unvaccinated Covid-19 cases
* To determine the significant association of detecting new hotspot or cluster of Covid-19 cases via MySejahtera check-in pattern

**2.0 Methodology**

The study aims to fulfill the objectives, which mainly to examine the relationship of daily MySejahtera app’s usage and the Covid-19 trend cases. The data collected and use was during the period of 1st December 2020 till 11th June 2022. During this period few MCOs occurred throughout the country and states, with the surge of cases as well.

**2.1 Data collection**

Usage of MySejahtera app became crucial nationwide. The check-ins data were collected upon entering facilities and places were obtained from official Ministry of Health, Malaysia thru reliable data source.

**2.2 Data preprocessing**

**Data integration**

Upon gathering the data collected from the data source, there were 6 comma separated value (CSV) files intended to use for this study. Initially all the CSV files were standardized in terms of the date format. Then, we identify the period of dates consists of row to use as the data. Unintended data was excluded from the main merged CSV files. Two merged CSV files created focusing on national check-ins data usage with Covid-19 trends and states check-ins data as well. We align the data points from different sources to ensure consistency in the merged data files created.

**Data cleaning**

As the merged files came from different format, we addressed missing values and duplicated data. All the processes were done via phyton codes, run either from Spyder software or Jupyter Notebook via anaconda prompt. We found that missing values were reported at the Covid-19 cluster data, probably the data was not recorded in the initial phase before the case surge. All missing values were already filled with appropriate values. 559 data rows of check-ins data were taken for this study.

**Ethical considerations**

In this study, data privacy of all MySejahtera app user were not affected. No private data were exposed. The data source also mentioned that the data was available for data processing. Necessary credit already mentioned in this study document.

**Tools and Software**

Data analysis and calculation was made using python codes via Spyder and Jupyter Notebook. We did employ certain library in the process for statistical analysis and visualizations such as Matplotlib, Tableau or statsmodel.

**2.3 Descriptive Analysis (data in Malaysia)**

**Jifeng part**

* **- words**
* **- chart/histogram**
* **- table**

**2.4 Descriptive Analysis (data for states in Malaysia)**

**Dr Chow part**

* **- words**
* **- chart/histogram**
* **- table**

**2.3 Statistical Analysis**

By using the merged data, we are keen to assess the strength and direction of relationship between the MySejahtera check-ins with daily new Covid-19 cases, unvaccinated cases, and reported clusters trend. We have applied the Pearson Correlation method via python codes. The correlation coefficient with significance value were generated.

To explore predictive relationships among the said data, we have used the Ordinary Least Squares (OLS) logistic regression model. The details of the model as below:

* Dependant variables: Daily new Covid-19 cases, daily unvaccinated Covid-19 cases, number of clusters daily report
* Independent variables: Daily MySejahtera check-ins

All the generated report and tables consists of significance value. We take p-value<0.05 as significant relationship and association among the variables. Regression model was created based on the following equation:

**3.0 Results**

**3.1 MySejahtera Apps Daily Check-in usage and Daily new Covid-19 cases**

The dataset dated from Jan 2020 till June 2022 comprises from a lot of numbers as it records all the check-ins done by the Malaysians during Covid period, albeit pre, during or post lockdown time. One of the objectives of this study is to look into the correlation of this check-ins data and new covid cases reported.

Based on the dataset, there is significant correlation between daily check-in and new Covid-19 cases day to day. It was both verified by Pearson Correlation method (p-value = 0.0001) and Ordinary Least Squares (OLS) logistic regression model (p-value = 0.0001). However, both methods showed there is very minimal changes in the new cases recorded daily affected by the check-ins. Only 2.6% variation was found in Table 1.0. Thus, even though the correlation among them is statistically significant, but the overall performance to explain the changes is very weak. The same phenomenon also being reflected in Covid-19 daily active cases, as per Table 2.0. Furthermore, the scatter plot in Figure 1.0 also shown only slight upward of the regression line which indicated weak positive relationship among daily Covid-19 cases and check-in apps usage.

**Table 1.0 : OLS regression table for association of dependant variable (new daily Covid-19 cases) and independent variable (daily MySejahtera Check-ins)**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Variable** | **Coefficient** | **Std. Error** | **t-statistic** | **p-value** | **95% CI** |
| **Intercept** | 4738.05 | 899.85 | 5.265 | 0.000 | [2970.54, 6505.56] |
| **MySejahtera Apps Daily Check-ins** | 0.0002 | 0.00004 | 3.861 | 0.000 | [0.00008, 0.0003] |

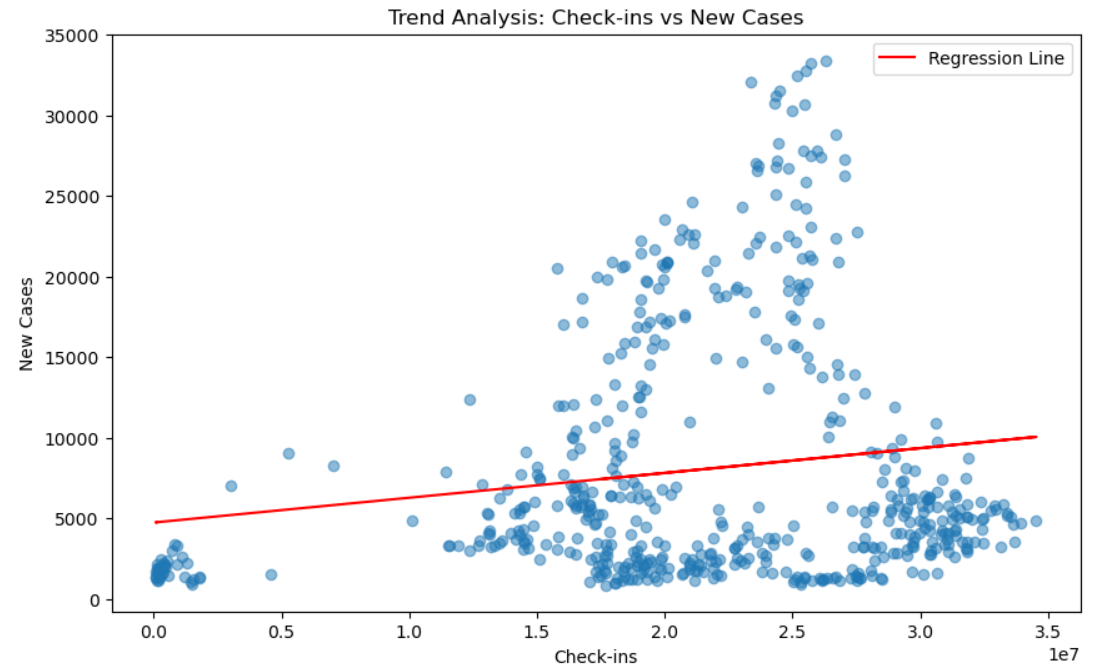


Figure 1: Scatter plot of trend analysis – Daily MySejahtera check-ins and new Covid-19 cases

**3.2 MySejahtera Apps Daily Check in and Daily new Covid-19 unvaccinated cases**

On the other angle, the apps which intended to detect, assist and boost the Covid-19 vaccination program in Malaysia was also checked for statistically related to new daily unvaccinated cases. The dataset did not reveal a good relation between daily check-ins of MySejahtera apps and the daily unvaccinated cases. Table 3.0 showed that there was only 0.5% of the variation in unvaccinated cases daily explainable by the daily check-ins (p-value >0.05).

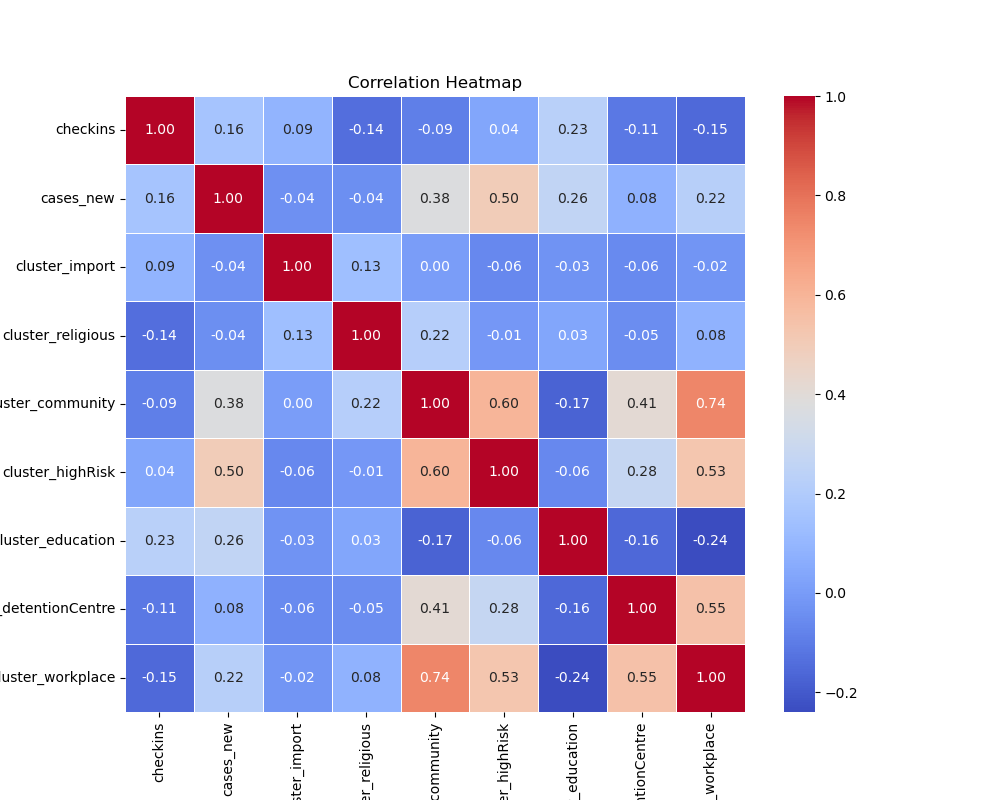


Figure 2: Correation heatmap of daily MySejahtera check-ins

**Table 2.0: OLS regression table for association of dependant variable (daily Covid-19 active cases) and independent variable (daily MySejahtera Check-ins)**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Variable** | **Coefficient** | **Std. Error** | **t-statistic** | **p-value** | **95% CI** |
| **Intercept** | 56590 | 10000 | 5.659 | 0.000 | [36900, 76200] |
| **MySejahtera Apps Daily Check-ins** | 0.0017 | 0.0004 | 3.840 | 0.000 | [0.001, 0.003] |

**Table 3.0: OLS regression table for association of dependant variable (daily Covid-19 unvaccinated cases) and independent variable (daily MySejahtera Check-ins)**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Variable** | **Coefficient** | **Std. Error** | **t-statistic** | **p-value** | **95% CI** |
| **Intercept** | 3961.01 | 367.58 | 10.776 | 0.000 | [3238.99, 4683.02] |
| **MySejahtera Apps Daily Check-ins** | -0.000 | 0.000 | -1.632 | 0.103 | [-0.0000585, 0.0000054] |

**3.3 MySejahtera Apps Daily Check-in with cluster of Covid-19 cases**

There are few clusters of emerging Covid-19 cases during the said duration of endemic time in Malaysia. The heatmap shown as Figure 2.0 that the nearest correlation with were cluster cases within the community (of the infected cases), followed by high risk cluster and cluster at Covid-19 detention centre. Further correlation using OLS regression with multiple independent variables also revealed that there were positive and great significant of new daily Covid-19 cases with cluster of cases in the community, high risk group and education. These clusters were associated with higher new Covid-19 cases daily (p-value<0.05), as per Table 4.0.

**Table 4.0 : OLS regressions table for association of dependant variable (daily Covid-19 new cases) with independent variable (daily MySejahtera Check-ins, clusters of Covid-19 cases)**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Dependant Variable** | **R-squared** | **F-statistic** | **p-value** | **Log-likelihood** | **AIC/BIC** |
| New Covid-19 cases daily | 0.384 | 42.90 | 0.000 | -5659 | 11340/11380 |
|  |  |  |  |  |  |
| **Independent Variable** | **Coefficient** | **Std. Error** | **t-statistic** | **p-value** | **95% CI** |
| **Intercept** | 3961.01 | 367.58 | 10.776 | 0.000 | [3238.99, 4683.02] |
| **MySejahtera Apps Daily Check-ins** | 0.00007 | 0.00003 | 2.050 | 0.041 | [0.000003, 0.000138] |
| **Cluster import** | -16.38 | 80.28 | -0.204 | 0.838 | [-174.06, 141.31] |
| **Cluster Religious** | -16.61 | 5.93 | -2.801 | 0.005 | [-28.26, -4.96] |
| **Cluster community** | 13.75 | 2.40 | 5.725 | 0.000 | [9.03, 18.47] |
| **Cluster high-risk** | 103.26 | 11.37 | 9.081 | 0.000 | [80.92, 125.59] |
| **Cluster education** | 18.54 | 2.23 | 8.308 | 0.000 | [14.16, 22.92] |
| **Cluster detention centre** | -6.56 | 4.58 | -1.434 | 0.152 | [-15.55, 2.43] |
| **Cluster workplace** | -1.65 | 0.88 | -1.872 | 0.062 | [-3.39, 0.08] |
|  |  |  |  |  |  |

**4.0 Discussion**

**Descriptive Analysis**

**Statistical Analysis**

Few practical concerns and highlights can be learn and unlearn after the Covid-19 pandemic in Malaysia. In this study, it clearly shows that certain steps can be taken to mitigate the spread of the disease. The arrival of super app such as MySejahtera app have a great potential towards the advantage of controlling the disease – Covid-19.

*Awareness of Covid-19 cases and MySejahtera app*

There was significant correlation between daily check-ins data by MySejahtera app recorded and new Covid-19 cases reported during the study period. For every additional check-in recorded in the data, there is estimated average increase of 0.00002 in a new daily Covid-19 cases. It could probably due to increase awareness and compliance to MySejahtera app in daily routine of Malaysians during that period. The detection of new Covid-19 cases can be seen and reported via the app. It also showed the public compliance and confidence towards the government policies. South Korea also used a detecting tracking app during the surge of Covid-19 cases4. The implications are valuable for policy makers such as Ministry of Health to encourage safety and self-coping behaviour.

The small increase in relation to daily check-ins but indeed a significant relations (p-value<0.05) may be contributed by lagged effect of cases. Due to increase mobility of public during and after the MCOs, individuals move more and check-in more frequently. This will lead to increase exposure risk to Covid-19 infection, leading to rise in detected cases after certain incubation period. The incubation period known for Covid-19 infection is around 2-14 days after first contact5. It is also depending on the variant that individuals exposed to. Hence, the rise or surge of cases may differ because it will undergo incubation period first. Patient to patient may differ as unique defensive immune system in every individual.

Other contributing factors, such as public health interventions towards Covid-19, government policies, raised vaccination rates and demographic dynamic also lead to a very small coefficient change (0.00002).

*Unvaccinated – special population*

It was challenging to get the whole public to adhere to a new policy – such as vaccination program. When vaccination introduced to the world, it was brought in to Malaysia by the Health minister on February 20212. The program establish more with the help of MySejahtera app. The data of vaccinated versus unvaccinated also obtained from this app. One of the reason the association of check-ins with unvaccinated Covid-19 cases may be attributed by the dynamics of unvaccinated population itself. Many reasons why this population refused for vaccination, or they not even captured by the apps itself. They also have reduced access to testing facilites. Given the lack of statistical significance, MySejahtera check-in data may not be a reliable tools for tracking the unvaccinated individuals. In future, government should tackle this issue with different method to better understand this type of group dynamics. **Journal ya**

*Cluster detection*

Direct impact of public mobility and high usage of MySejahtera app check-ins correlates with higher transmissions of Covid-19 infection during the study period. High risk clusters group, community clusters and education clusters group have significant coefficient and impact on daily reported Covid-19 cases. On the other hand, religious clusters have negative coefficient, maybe indicated by reduced religious activity in the community. Policy maker can use this data to mitigate new infection if ever to happen again. Focus on high risk population, education planning can reduce the spread of an infection, such as Covid-19. Closing the religious temple or mosque can also reduce the rate of infection via touch and airborne0.

**Conclusion**

There are many ways of tackling the spread of Covid-19 infection. Using MySejahtera check-in alone can reveal a lot of information that is useful. Increasing awareness among the public but adherence to a specific policy of using a simple app such as MySejahtera can give many advantages. Future improvement can be done by incorporate other demographic variables to enhance model efficacy.

**Reference**

1. https://github.com/MoH-Malaysia
2. https://en.wikipedia.org/wiki/COVID-19\_vaccination\_in\_Malaysia#:~:text=of%20Medical%20Biology.-,February,a%20period%20of%20two%20weeks.
3. Morawska, L., & Cao, J. (2020). Airborne transmission of SARS-CoV-2: The world should face the reality. *Environment international*, *139*, 105730.
4. Kim, H. (2021). COVID-19 apps as a digital intervention policy: a longitudinal panel data analysis in South Korea. *Health Policy*, *125*(11), 1430-1440.
5. How long is the Covid-19 Incubation period?, WebMD source website, article written by Sonya Collins, Even Starkman. https://www.webmd.com/covid/coronavirus-incubation-period