ISSN: 0067-2904

Data Analysis of COVID-19 Exploration on the 'Peduli-lindungi' Application in Indonesia

Rossi Passarella[1*], Kemahyanto Exaudi[2], Gulfi Oktariani[3] and Xosya Salassa[4]
[1] *Intelligent System Research Group, Faculty of Computer Science, Sriwijaya University, Indonesia*
[2] *Strategic Pervasive Computing Intelligent Embedded System Research Group, Sriwijaya University, Indonesia*
[3] *Data Analyst, Palm Oil Fund Management Agency, BPDPKS, Indonesia*
[4] *Data engineer, Enterprise Digital Technology Services (EDTS), Indonesia*

Received: 8/6/2023 Accepted: 26/1/2024 Published: 28/2/2025

Abstract

The peduli-lindungi (Care-Protect) application is an official application used by the Indonesian government to monitor crowds of people who are active in closed and public spaces. The purpose of this application is to facilitate actions that can be taken by the government to prevent the spread of COVID-19. One of the actions is user tracing. In this study, an analysis was carried out on data originating from the Peduli-lindungi Application, which is confidential. Some of the data fields used are still protected by the government. To get accurate data analysis results, preprocessing techniques are carried out on Peduli-lindungi data to produce new datasets as observational data material. The results obtained show that 4073 data points were successfully observed based on the population identification number (NIK) or that there was a reduction in the number of observation data points by 2.83% from the Peduli-lindungi source data. The exploratory data analysis (EDA) method and data visualization techniques using Scimago Graphica and Orange Data Mining were successfully applied to this study. The analysis results obtained show that as many as 13% of the number of people based on NIK do not comply with government regulations on traveling or entering public places, as indicated by the color status in the Peduli-lindungi application, namely red and black. As many as 60% of them do their activities by entering shopping centers, namely malls. Based on the results of the observations and data analysis that have been carried out in this study, it can be concluded that the MALL location is a cluster that must be the center of attention for the government.

Keywords: covid-19, Peduli-lindungi app, Exploratory data analysis, user color, malls.

1. Introduction

Corona Virus Disease 19 (COVID-19) is a highly contagious virus that has plagued the world [1], including Indonesia, for almost 2 years. The presence of this virus makes some of the activities of the world community disrupted [2] or paralyzed. The impact of this virus requires every community to make new habits in activities both indoors and outdoors. The Indonesian government officially stated that this virus entered Indonesia in March 2020 [3]. One of the first residents to claim to be infected was from the West Java province of Depok City [4]. However, the government is optimizing the control of this virus by releasing a new application called Peduli-lindungi. Peduli-lindungi is an application developed by the government to conduct tracing [5] to stop or reduce the spread of the COVID-19 virus in Indonesia [6].

*Email: passarella.rossi@gmail.com

This application can be used by the community to provide location information voluntarily when visiting or entering a public place. With public awareness of this application, it can facilitate the tracking and clustering process if there are people who are confirmed positive. Peduli-lindungi also provides statistical information such as the spread of COVID-19 and the risk of transmission, information and quotas for COVID-19 referral hospitals, vaccine registration, QR scans for screening, digital passports for flight eligibility, travel history, and telemedicine, which can be used to check one's own health and also allows consultation with general practitioners and specialists. The spread of COVID-19 in Indonesia can be controlled using Peduli-lindungi. Each user is categorized based on a different color ID (green, yellow, red, and black) [7] to determine their health condition and track record [8]. The green color indicates that the user can do activities in public places with the criteria of having vaccinated the advanced dose (booster), not being a COVID-19 patient or close contact, having negative PCR test results, and having recovered from COVID-19 for less than 90 days. The yellow color indicates that users cannot move in public places regulated by local government regulations with the criteria of having received a complete vaccine, not in close contact with COVID-19 patients, and recovering from COVID-19 for less than 90 days. Users who cannot travel to public places, have not received a vaccine, are not COVID-19 patients, or are in close contact are categorized in red. While blacks are categorized as users who have been positive for COVID-19 for less than 10 days or have had a history of contact with positive cases for less than 14 days.

Centralized data from Peduli-Lingungi can be used to assess the effectiveness of a regulation that has been enforced against the spread of COVID-19 in Indonesia. So that in this study, the results presented are insights obtained from the data that has been provided so that it can be a benchmark for making decisions and government regulations that need to be updated.

2. Related Work

Researchers from all over the world frequently study human activity tracking, particularly to ascertain human behavior [9], habits [10], social security (crime) [11], and protection (risk of disease spread) [12]. The data generated from tracking can produce large-scale data. Several studies have been carried out to analyze large data sets in order to extract data patterns and insights. One of the methods used is K-means clustering [13]. Some researchers use GPS (the Global Positioning System) and the Internet of Things (IoT). A cell phone is one of the instruments that many people do not understand they may use for tracking. It is unlawful to trace a mobile phone if the owner does not consent. To legitimize tracking, we must create an app that can be downloaded from Google Play and install it intentionally and without restriction. The step of creating an application and requiring it to be activated was taken by the Indonesian government to control the spread of the virus.

3. Materials and Methods

The data source used in this study is the Peduli-Lindungi database, which has an extension format (.csv). The database provided has limited access from the Ministry of Finance in collaboration with the Ministry of Health and is intended for the 2022 Ministry of Finance Hackathon competition [14]. The challenge faced in this study is to process data with limited information so that it becomes the best analysis as a result of this study. The data used has been downloaded since September 12, 2022. This database file has a time span from February 10 to August 10, 2022, and contains information about the activities of Indonesian people while in the visited environment. The data recorded in the form of a population identification number (NIK) is 1,597,179 (4,085 NIK). Based on the observation results of the initial NIK data, it was found that there were 12 disguised population variables, namely several NIKs,

COVID-19 status, gender, QR Peduli-lindungi check-in time, QR Peduli-lindungi checkout time, province code, province name, district code, district name, outlet category, outlet sub category, and account color status Peduli-lindungi.

Based on the findings of 12 disguised variables from the given database, this study requires an initial investigation of the data used to find patterns and anomalies based on statistical tests using exploratory data analysis (EDA) techniques [15]. So that, by applying this EDA technique, we can check the assumptions that are disguised and then represent them in the form of a graph. The initial investigative test process using the EDA technique in this study is shown in Figure 1.

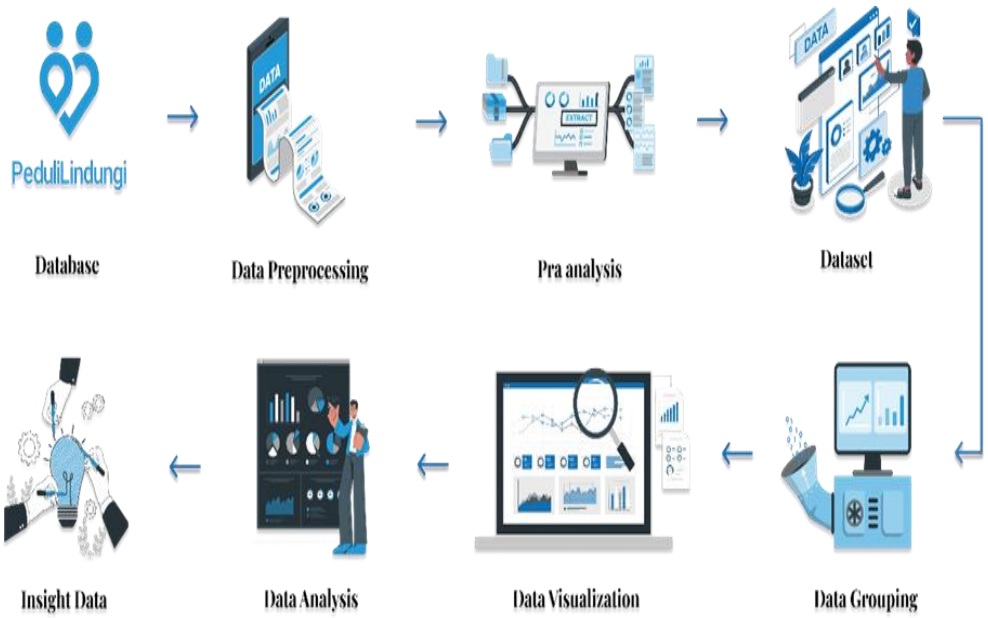

Figure 1: Initial data investigation process of the study based on the EDA technique [16].

Figure 1 explains that in the first stage, the initial observation of data that has been downloaded from Peduli-lindungi is carried out. The results of observation data are preprocessed to produce structured data variables according to the needs of this study. At this stage, we remove observation data that contains empty values in the variables NIK hashed, province_name, city_name, user_color_status, and check-in time. If any empty or null data is found in these five variables, the observation data in the raw data will be automatically removed. The drop-value method was used to handle missing data on the five variables. This method was chosen because the observation data is really important, and it would be unethical to fill in missing data in the analysis of the spread of this disease. The results of this preprocessing are referred to as a new database. This new database is the source of data for this study. Cleaning up empty or missing data on generated variables, such as the population's gender, is the first step in this preprocessing technique. After the data cleaning process is carried out, the missing value is obtained for as many as 5,022 data points. In the sub-category variables of public places visited, such as tourist attractions, restaurants, malls, fast food restaurants, stations, and others, there are missing values of 11,482 data points. As for duplicate data from the resident's NIK variable category, check-in date category, outlet category, and outlet sub-category, there are 1,127,383 data points from 2,433 duplicate NIKs. Based on the results of the data cleaning inventory that has been carried out, there are 453,292 data points with a total of 4,073 NIKs that can be used in this study.

The next stage is pre-analysis, which involves filtering variables from the data that has been investigated by cleaning the data. This filtering process is to form a new dataset, which is the

source of data for analysis in this study. This pre-analysis stage resulted in 453,292 datasets. The variables formed from the dataset are grouped based on the color status used in the Peduli-lindungi application. Black color status has 6 data from 6 NIK, red has 79 data from 78 NIK, yellow has 1,291 data from 158 NIK, and green has 245,979 data from 2,625 NIK. Residents who experienced a change in color status for 6 months or various color statuses amounted to 205,937 data points from 1,206 NIK.

The results of the data grouping process are then used to determine data patterns through the data visualization stage. The goal is to make it easier for researchers to convey information effectively and clearly so as to produce conclusions in the decision-making process. The output of the data visualization process is used as input data for the data analysis stage to produce data output that can solve problems and provide insights into the data of the Peduli-Lindungi application.

4. Result and Discussion

4.1 Data pre-processing and pre-analysis results

This stage describes the results of the implementation of the method used to obtain insight data and drive data from the Peduli-lindungi data used. The database used has the properties shown in Table 1. As for the structured data analysis process in this study, Python software is used by applying the Pandas open source library [17]. Orange data mining is used to classify and analyze protected data [18] and visualize the resulting data using the Scimago Graphica application [19].

Table 1: Properties of a Given Database

Parameter	Description
Database downloaded	September 12, 2022 at 9:03 pm
Data Format	Microsoft Excel Comma Separated Values File (.csv)
Data size	287 MB (301.417.298 bytes)
Created	September 18, 2022, 12:08:51
Modified	September 06, 2022, 14:03:52
Number of observation data (rows)	1597179
Number of variables (columns)	12
Features data	Categorical = 7; Numeric = 4 ; Meta attributes : = 1
Number of Unique NIK_Hashed	4085

The purpose of this stage is to see the distribution of data, including the mean, median, dispersion, minimum value, and maximum value of the data, and the most important is the missing value observation. The distribution of data from observations that have been made based on gender is represented by red for men and blue for women, as shown in Figure 2.

Figure 2 describes in general how the database received provides important initial information, namely that the database has a data time span of 6 months (February 10 to August 10, 2022). The province that appears the most is DKI Jakarta, precisely in the city of Administration, South Jakarta. The highest outlet category is in shopping centers, with a sub-outlet category, namely malls. While the highest color data for ID Peduli-lindungi shows the color green.

Figure 2: Statistical features of the Peduli-lindungi database

By knowing the basic data of the given database, this study gets some important things that help in determining the scope of analysis to get data insight and data-driven. Data value abnormalities are produced in the form of data that shows the observation results of gender have empty values (null).

The amount of data that has empty values for gender amounts to 5,022 observations, or 0.314%. The total distributed database is only 12 NIK_hashed. Referring to the rules imposed, NIK_hashed is a unique meta-attribute code where one NIK_hashed represents one person. Based on the visualization data generated from Scimago Graphica, databases that have empty values in the gender variable can be seen in Figure 3. From the distribution visualization of Figure 3, an insight into these 12 people is obtained, where for 6 months there is a 41.67% change in color status from green to yellow, and within this percentage there is a movement from green to yellow and black as much as 20% (1 in 5 people). This group of data has gone through an improvement process by referring to the NIK_hashed against the previous data that has a value for gender. However, what these 12 people found in the database was that the gender variable was unknown or empty (null).

In addition, this study also sets aside data that has an empty value in the outlet_sub_category variable because if this data is empty, there is no comparison when filling in the data. It is different if the outlet_category variable has an empty value; then it can be filled in by referring to the outlet_sub_category. Thus, the amount of data found in the outlet_sub_category variable that is empty is 11,482 data observations. The total data eliminated in this analysis is 16,504 data observations, or 1.033% of the initial database.
At this stage, the database is eliminated from 1,597,179 observations to 1,580,675 data observations, with the number of NIK_Hashed as many as 4,073 people. The number of NIK_hashed was reduced from 4,085 people to 4,073 people. The next step is to look for data duplication that refers to the date or day of the checkin_timestamps variable. It is assumed that it is not a matter of how many times people enter and exit the same building or place, but based on the day or date recorded where the person is. Thus, the checkin_timestamps and checkout_timestamps variables are not the focus of our study, even though a lot of checkout_timestamps data shows NaN (not a number) values.

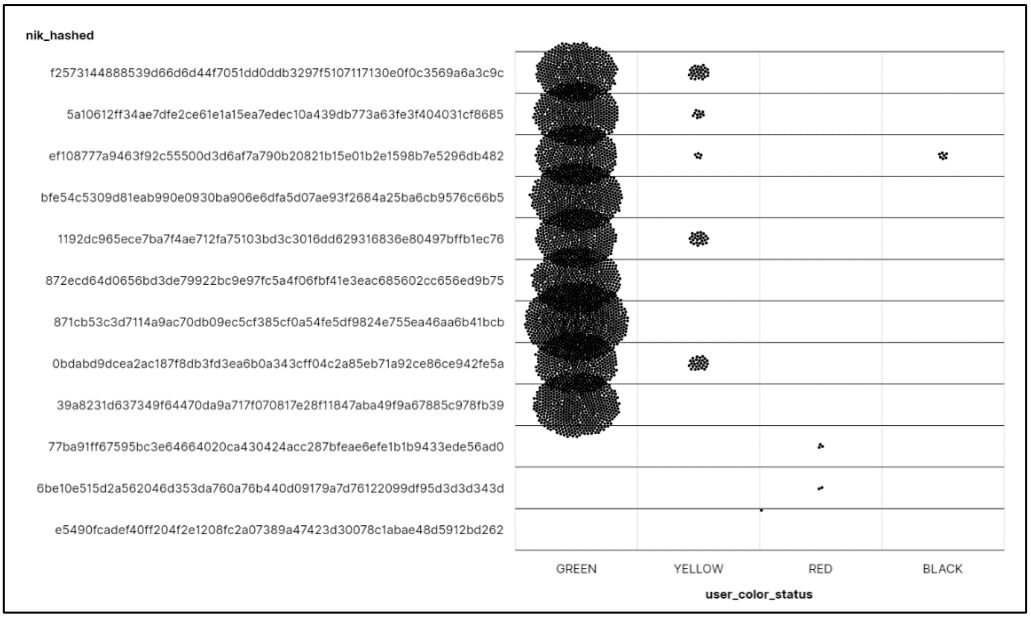

Figure 3: Distribution of databases with no values on the gender variable

Based on the results of the analysis that has been carried out using the Pandas library, we found 71,683 observations from 2,433 people who were detected tapping or recorded entering a place on the same date, outlet_category, and outlet_sub_category. Therefore, the total database of 1,580,675 observations was subtracted from 71,683 observations, resulting in 453,292 observations of 4,073 people. We refer to this observation data as the dataset.

4.2 Data grouping

Before the data visualization process of the resulting dataset is carried out, the first stage is to perform statistical features on the dataset to see if there is any missing data. The results indicate that there is no missing data. All the data variables that are the focus of this study show a value of 0 (0%). The statistical feature results shown in Figure 4 indicate that the red and blue colors refer to the gender of the users of the Peduli-lindungi application. The red color indicates a male, and the blue color indicates a female. The percentage of these two genders is 69% male and 31% female. Meanwhile, the grouping of data based on the color of users who care about protecting the dataset is shown in Table 2. This grouping is very important to do to further focus the purpose of data observation Peduli-lindungi in finding data insight and data-driven. The MIX, or a combination of four colors, is based on the facts found where there are several NIKs (users) who experience color changes obtained from six months of observation data. Meanwhile, the black, red, yellow, and green color groups did not show any color changes during the six months of observation data. Therefore, in this study, the red and black color groups and the MIX data group, especially the red and black colors, are discussed. In the following explanation, only these 3 data groups are discussed, while the yellow and green color groups are ignored.

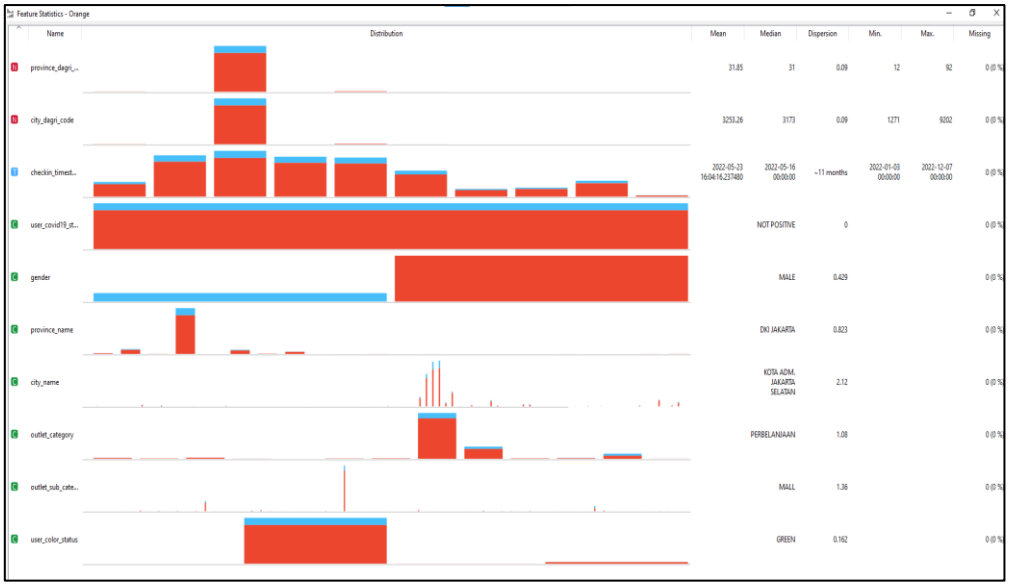

Figure 4: Feature Statistics of the Final Dataset

Table 2: Data distribution of datasets based on the user_color_status variable

User_color_Status	Total Observations	Number of NIK
Black (FIX)	6	6
Red (FIX)	79	78
Yellow (FIX)	1.291	158
MIX (Changing color)	205.937	1.206
Green (FIX)	245.979	2.625

4.3 Data Visualization and data analysis

The dataset that has been formed is then visualized based on the distribution of provinces, cities, gender, outlet categories, and outlet subcategories. This visualization is done to help the analysis process and gain data insight.

4.3.1 Dataset visualization

In general, the distribution of the used Peduli-lindungi data is spread across 22 provinces, with the highest order of observation data in DKI Jakarta province with 2,142 data observations. Banten province with 1,658 observations and West Java province with 1,304 observations. Based on this distribution data, it can be concluded that this data represents 59.46% of the total provinces in Indonesia included in the data set provided by the Ministry of Health in collaboration with the Ministry of Finance.

4.3.2 Black color group (User Color Black)

The black category is categorized as users who are detected scanning the Peduli-lindungi application at a place and date and detected in the dataset or 6-month time span only in black (no color change). The visualization data shows only 6 users (hashed NIK) with 2 detected city locations, namely Tangerang district (50%), and Bekasi city (50%), where users enter a shopping building in the form of a motel in each city. Here something is interesting, namely that all 6 users are female. In addition, the visits of these 6 users were only detected in February and were not detected again until August 2022, based on the dataset used. Figure 5 shows the data visualization of users Peduli-lindungi in black in the form of a plot of links between variables over 6 months, and no color change was detected.

Based on the color meaning rules in the Peduli-lindungi application, black is positive for COVID-19 for less than 10 days and/or has a contact history with positive cases for less than 14 days. However, the dataset used does not provide the specific name of the mall detected, so we cannot trace this user interacting with other users because it can cause the analysis to be biased.

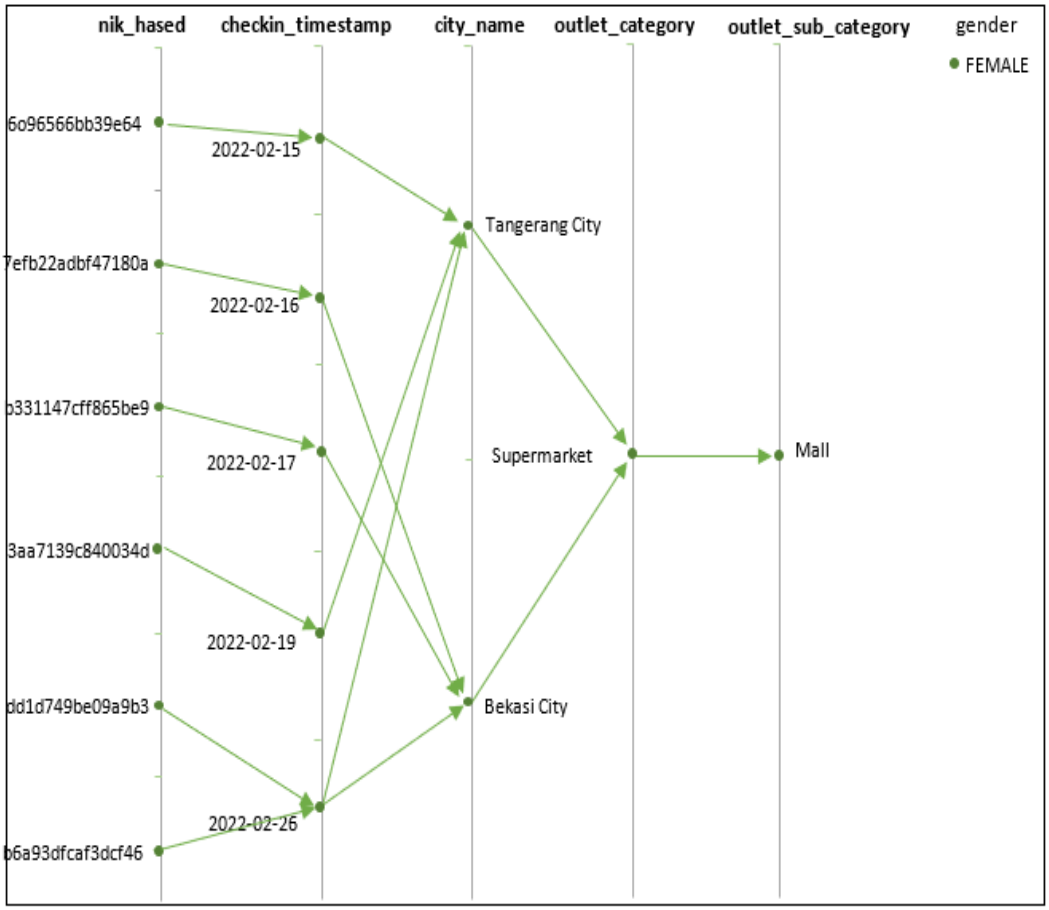

Figure 5: Visualization of black Peduli-lindungi user data over a 6-month period that has no detectable color change

4.3.3 Red color group (User Color Red)

The distribution of the movement of the red-colored Peduli-lindungi group within 6 months shows that 99.17% gather or enter shopping centers in the form of malls in Bekasi City and Tangerang district. It can be assumed that the mall is a meeting point for users who care about protecting themselves, just as users who have black skin do. The distribution of data in this group based on gender is as follows: 34 women and 44 men were detected. If we look deeper into Tangerang district, 47 users in this group were detected, with a distribution of 19 people being female and the remaining 28 people being male. As for the mall in Bekasi City, 31 people were detected, with a distribution of 15 people being female and 16 people being male. When viewed in terms of time distribution, the red user data in this group began to be detected from February 11 to July 24, 2022.

4.3.4 Color group changing (User Color Changing (MIX))

The next group to be analyzed is the combined color group (MIX). This group is categorized as a group that has experienced a color change in the Peduli-lindungi application during the 6 months of observation data in the dataset. The distribution of 1,206 people in this group is in 14 provinces.

4.3.5 Data analysis and data insight

These 2 colors explain why people who cannot travel to public places, have not received vaccines, and are also not close contacts of COVID-19 in the Peduli-Lindungi application get red. The last status is black, where this community has been positive for COVID-19 for less than 10 days or has had a history of contact with positive cases for less than 14 days. The distribution of red and black data in this group amounted to 174 people (NIK_hashed), with a distribution in 9 provinces in 22 cities. DKI Jakarta and the Administrative City of South Jakarta, with a total of 124 observations, are the highest points of red and black distribution; the most sub-category outlets are malls, followed by private office buildings and stations with an order of 298, 75, and 25 observations. The details of the black and red user data are shown in Table 3 and Table 4.

Table 3 shows the highest number of outlets detecting users concerned with black protection: malls, followed by private office buildings and fast food restaurants. While in clinics and hospitals, it is possible that the user aims to do treatment or other purposes. Whereas malls, private offices, and fast food restaurants are, based on the rules, prohibited areas.

The same thing with the distribution of users in red: this group was detected to do a lot of activities in the mall, followed by office buildings and stations. Most malls are in the Administrative City of West Jakarta, while most private offices are in the Administrative City of South Jakarta. The means of transportation used by many users are found in red, namely at stations in the Tangerang district, while the location of the factory, refinery, or power plant is found in the administrative city of South Jakarta. The distribution of data, in this case, is outlined in Table 4. To see deeper data or special insight data on black data taken from Table 3, the tabulation is obtained as in Table 5. In Table 5, it can be concluded that these 11 people (NIK_hashed) experienced a color change from green to black and then to green again. The average duration of these 11 people disappearing or not doing activities during their black status is 8 days. In addition, based on this data, it was found that the average NIK_hashed that entered the previously detected mall was 54.54%.

Table 3: Distribution of outlets in the subcategory Peduli-lindungi black

	Private Office	Clinic	Mall	Hospital	Fast Food Restaurant
Buleleng Regency	1				
Tangerang Regency			1		
Administrative City of West Jakarta				1	
Administrative City of Central Jakarta		1	1		
Administrative City of South Jakarta			1		
Bekasi City			2		
Bogor City					1
Semarang City			2		
South Tangerang City				1	

Table 4: Distribution of outlets in the red-colored Protect subcategory

Region	Bank	Cinema	Private Office	SOE office	Mall	Factory/ Refinery/ Power Station	Restaurant	Hospital	Fast Food Restaurant	Station	Supermarket	Tenant Mall
Bekasi Regency						2						
Bogor Regency										1		
Gresik Regency						1						
Karawang Regency						1						
Sleman Regency					4							
Tangerang Regency					8					8		
Administrative City of West Jakarta		1	5		88						4	
Administrative City of Central Jakarta			15	3	18					3		
Administrative City of South Jakarta			53		59	9				1		1
Administrative City of East Jakarta					5					4		
Administrative City of North Jakarta	1				15							
Balikpapan City						1						
Bandung City					2							
Bekasi City					23							
Depok City					19				1	1		
Medan City					1							
Semarang City			1		14			1				
Surabaya City					34		1					
Tangerang City					1					1		
South Tangerang City										5		

Table 5: Insight data of users who care about protecting black

Black Status in Peduli-Lindungi					First detect		Before		After (changing color detecting)		Duration	Changing color
First Date in dataset	Code	Gender	Freq	City Name	Check in timestamp	Location	Check in timestamp	Location	Check in timestamp	Location		
2/10/2022	Case 1	Male	1	Tangerang Regency	2/10/2022	Mall	-	-	2/20/2022	Mall (Depok City)	10	Black to Green
2/16/2022	Case 2	Female	1	Bekasi City Administrative City of West Jakarta	2/16/2022	Mall	-	-	2/22/2022	Mall	6	Black to Green
2/10/2022	Case 3	Male	1		3/6/2022	Hospital	3/3/2022	Mall (Administrative City of South Jakarta)	-	-	-	Green to Black
2/16/2022	Case 4	Female	1	Bekasi City	2/16/2022	Mall	-	-	2/22/2022	Mall	6	Black to Green
2/10/2022	Case 5	Female	1	South Tangerang City	3/6/2022	Hospital	3/4/2022	Hospital (East Jakarta Administrative City)	3/16/2022	Mall (Administrative City of South Jakarta)	10	Black to Green
2/10/2022	Case 6	Female	1	Administrative City of South Jakarta	2/11/2022	Mall	2/10/2022	Mall	2/21/2022	Private Office	10	Black to Green
2/10/2022	Case 7	Female	1	Administrative City of Central Jakarta	2/24/2022	Mall	2/23/2022	Mall (Bekasi City)	2/28/2022	Station (Depok City)	4	Black to Green
2/10/2022	Case 8	Male	1	Administrative City of Central Jakar	2/18/2022	clinic	2/16/2022	Mall	2/28/2022	Mall	10	Black to Green

				ta								
2/10/2022	Case 9	Male	1	Bulleleng City	2/14/2022	Private Office	2/13/2022	Station (Serang City)	3/4/2022	Station (Serang City)	18	Black to Green
2/10/2022	Case 10	Male	1	Semarang City	2/18/2022	Mall	2/17/2022	Mall	2/27/2022	Mall	9	Black to Green
2/10/2022	Case 10	Male	1	Semarang City	2/20/2022	Mall	2/18/2022	Mall	2/27/2022	Mall	7	Black to Green
2/13/2022	Case 11	Male	2	Bogor city	2/13/2022	Fast Food	-	-	2/15/2022	Private Office (Administrative City of South Jakarta)	2	Black to Green

There are interesting things found in Table 5. It was found that a male in the city of Semarang was free to go in and out of the mall even though the color status of the Peduli-lindungi application was black on February 18, 2022, and returned to the mall on February 20, 2022. After February 20, 2022, the person concerned returned to activities and was detected on February 27, 2022, at the Semarang city mall. To facilitate the analysis of NIK_hashed in Table 5, it is coded with the name *Case* as the order of naming. In other words, 1 case name can represent 1 NIK or person.

Based on the visualization and analysis conducted and referring to the data summary, we focus on the red and black users. The data shows that the mall is the most dominant place in each subset of the dataset. The data details are shown in Table 6.

Table 6: Percentage of Peduli-lindungi users who entered the mall based on the NIK variable

NO	Subset dataset	Total NIK with Red	Percentage data Outlet sub category for Mall	Total NIK with Black	Percentage data Outlet sub category for Mall
1	Black	-	-	6	100%
2	Red	78	98,73%	-	-
3	MIX (changing color)	418	70 %	11	63,63 %

In addition, we also get insight from the total NIK based on the dataset of 4,073 people. 513 people were detected in red and black. These two colors are based on the rules that enforce that it is prohibited to enter public places. The percentage of violations that indicate non-compliance with government regulations is 13% of the total NIK. According to the data, 13% of the non-compliant population is concentrated on Java Island, including the capital city of Jakarta. Java Island is the most populous island in the country, with over 70% of Indonesia's population residing there. With a population density of 1.24 inhabitants per square kilometer, or 160 million people inhabiting 128,793.6 square kilometers of Java Island, it is challenging for the government to regulate the population identified by the red and black colors of the protection application ID to isolate themselves. The government's indecisiveness in controlling its population is due to the rapid spread of the virus and its deadly impact. In the future, it is crucial

for the state to address the issue of population control and provide adequate protective equipment for healthcare workers, police, and military personnel to maintain discipline.
Table 3 identifies several outlets visited by users of the peduli-lindungi application that are marked with a black color code. Private offices, malls, and fast-food restaurants are among the outlets where certain activities should not be allowed or permitted, as indicated in Table 3. The possibility of this occurrence is due to the absence of building security, which is caused by the lack of personal protective equipment and information about the spread and impact of the virus, which can result in fatalities.

5. Limitation of study

The limitations of this study must be acknowledged, as it is based on the results of data analysis from the care-protect application. The data was obtained through collaboration between the Ministry of Finance and the Ministry of Health as part of the Hackathon 2022 competition. The two ministries have filtered this data to protect the identity and COVID-19 status of application users.

The shared data has a limited duration allowed by the state to be opened, and the detailed identity of the outlet location has been removed. The limited data details may significantly impact the analysis. However, it is important to note that the Indonesian government has allowed researchers to gain insight into the process of tracking individuals through mobile applications required for entry into certain locations.

6. Conclusion

Based on what we have done in this study on the database provided by the Ministry of Health in collaboration with the Ministry of Finance, we can conclude as follows:
1. The database of the Peduli-lindungi application over a period of time has undergone several changes for confidentiality purposes. The results of the analysis carried out in this study are biased and not in accordance with the objectives for tracing. For example, the dataset provided by the ministry does not provide specific mall names, so tracing users who interact with other users cannot be done, which causes the analysis to be biased.
2. Due to the lack of detailed information in the database, the analysis carried out in this study to gain insight is limited to exploratory data analysis only.
3. The improvement of the database to become a dataset requires great attention, especially the discovery of NIK_hashed, which has no gender. The main data is the NIK and user name to get approval to use the application (registration process), and the gender is supposed to be pulled from the ministry's database, but in this case only found some data (12 NIK) with an empty gender, so this dataset was removed.
4. A large amount of repetition data and the non-compliance of users to check out of a place make it impossible to analyze how long users stay in a place.
5. There is a simplification of check-in data where if a user enters the same place (NIK_hashed, city, and outlet sub-category) several times so that there is a repetition of check-in data, it is considered to have only checked in once.
6. Locations where red and black users are mostly detected in public places where they should not be found in the dataset include as many as 513 people, or 13% of the total users analyzed.
7. The results of the data analysis show that more than 60% of the red, black, and mixed color groups are detected in shopping centers or malls.
8. The government needs to strengthen and support places such as shopping centers (malls) to enforce disciplinary rules so that the wider community feels protected. In addition,

campaigns to deal with dangerous viruses that have spread, such as COVID-19, need to be improved so that the public does not experience disobedience, shock, and fear in the future.

The learning that can be done in the future is how to cluster this application data by completing the variables in the form of job types so that it can be known which jobs are at risk of spreading the virus other than health workers, so that the government can anticipate this other than by implementing the Working from Home (WFH) program.

Acknowledgments

We are grateful to Mr. Muhammad Naufal Rachmatullah, who has been helping with the visualization of the data from the Peduli-lindungi application.

Conflict of interest

The authors declare no conflict of interest.

References

[1] H. Harapan *et al.*, "Coronavirus disease 2019 (COVID-19): A literature review," *J. Infect. Public Health*, vol. 13, no. 5, pp. 667–673, 2020, Doi: https://doi.org/10.1016 /j.jiph.2020. 03.019.

[2] U. Napitu, Corry, and R. W. Sipayung, "Disruption and Community Life Strategy to Overcome The Covid-19 Pandemic From Social-cultural Perspective," *KnE Soc. Sci.*, vol. 8, no. 4 SE-Articles, Mar. 2023, Doi: 10.18502/kss.v8i4.12957.

[3] R. Djalante *et al.*, "Review and analysis of current responses to COVID-19 in Indonesia: Period of January to March 2020," *Prog. Disaster Sci.*, vol. 6, p. 100091, 2020, Doi: https://doi.org/10.1016/j.pdisas.2020.100091.

[4] R. Passarella and H. Veny, "Impact of Movement Restrictions During COVID-19 on Jakarta's Air Quality: A Comparative Analysis," J. Green Econ. Low-Carbon Dev., vol. 2, no. 2, pp. 91-109, 2023. https://doi.org/10.56578/jgelcd020205

[5] F. Antonius Alijoyo and E. Haerani, "The Implementation Of The Care-Protect (Pedulilindungi) Application: The Economic Impact And Constraints Faced," *Eduvest - J. Univers. Stud.*, vol. 2, no. 1, pp. 131–139, 2022, Doi: 10.36418/edv.v2i1.341.

[6] K. Kevin and R. Jayadi, "Evaluation of Pedulilindungi Applications in the Jabodetabek Region," *J. Econ. Bus. UBS*, vol. 12, no. 1, pp. 374–383, 2023, Doi: 10.52644/joeb.v12i1.138.

[7] R. Chatterjee, S. Bajwa, D. Dwivedi, R. Kanji, M. Ahammed, and R. Shaw, "COVID-19 Risk Assessment Tool : Dual application of risk communication and risk governance," *Prog. Disaster Sci.*, vol. 7, p. 100109, 2020, Doi: 10.1016/j.pdisas.2020.100109.

[8] L. Ellyanti, Y. Ruldeviyani, L. E Pradana, and A. Harjanto, "Sentiment Analysis of Twitter Users to the PeduliLindungi Using Naïve Bayes Algorithm," *RESTI (Rekayasa Sist. dan Teknol. Informasi)*, vol. 5, no. 158, pp. 414–421, 2023, Doi: 10.29207/resti.v7i2.4684.

[9] D. Monsivais, A. Ghosh, K. Bhattacharya, R. I. M. Dunbar, and K. Kaski, "Tracking urban human activity from mobile phone calling patterns," *PLOS Comput. Biol.*, vol. 13, no. 11, p. e1005824, Nov. 2017, [Online]. Available: https://doi.org/10.1371/journal.pcbi.1005824

[10] P. Smetaniuk, "Adapted Mobile Phone Use Habits (AMPUH) [Database record]," 2014. [Online]. Available: https://doi.org/10.1037/t35484-000

[11] S. Priyanka, S. Dutta, and S. Chakraborty, "Human Tracking System Based on GPS and IOT (Internet of Things)," in *International Conference on Computer Networks, Big Data and IoT (ICCBI)*, Springer International Publishing, 2020, pp. 514–521. Doi: 10.1007/978-3-030-24643-3_61.

[12] S. Kauer and S. Reid, "Mobile Tracking for Mental Health and Wellbeing," in *Encyclopedia of Mobile Phone Behavior*, IGI Global, 2015, pp. 860–870. Doi: 10.4018/978-1-4666-8239-9.ch071.

[13] Raheela zaib and OURLIS Ourabah, "Large Scale Data Using K-Means", *Mesopotamian Journal of Big Data*, vol. 2023, pp. 36–45, Feb. 2023.

[14] Anonim, "Finance and Health Hackathon 2022," *kemenkeu.go.id*, 2022. https://hackathon.kemenkeu.go.id/index.html

[15] S. Adivarekar, S. Nanwani, N. Mandal, and T. Sarode, "Implementation of Exploratory Data Analysis on Weather Data," in *2023 International Conference on Communication System, Computing and IT Applications (CSCITA)*, 2023, pp. 21–25. Doi: 10.1109/CSCITA55725.2023.10104864.

[16] E. Camizuli and E. J. Carranza, "Exploratory Data Analysis (EDA)," in *The Encyclopedia of Archaeological Sciences*, 2018, pp. 1–7. Doi: https://doi.org/10.1002/9781119188230.saseas0271.

[17] W. McKinney, "Pandas: a Foundational Python Library for Data Analysis and Statistics," 2011. [Online]. Available: https://api.semanticscholar.org/CorpusID:61539023

[18] U. Thange, V. K. Shukla, R. Punhani, and W. Grobbelaar, "Analyzing COVID-19 Dataset through Data Mining Tool 'Orange,'" in *2nd International Conference on Computation, Automation and Knowledge Management (ICCAKM)*, 2021, pp. 198–203. Doi: 10.1109/ICCAKM50778.2021.9357754.

[19] Y. HHassan-Montero, F. De-Moya-anegón, and V. P. Guerrero-Bote, "SCImago Graphica: a new tool for exploring and visually communicating data," *Prof. la Inf.*, vol. 31, no. 5, pp. 1–12, 2022, Doi: 10.3145/epi.2022.sep.02.