Part A

Data Preprocessing Summary:

In the preprocessing phase, we conducted various transformations to prepare the dataset for further analysis. Below is a summary of the preprocessing steps:

Feature Selection:

- Univariate, Variance Threshold and Correlation-based Feature Selection are used for analysis here. Univariate feature selection methods evaluate each feature independently to determine the strength of the relationship between the feature and the target variable (In this case it might not be the most suitable one since all tuple are numeric). Variance thresholding is a simple and intuitive method that removes features with low variance. Correlation-based feature selection identifies redundant features by measuring the correlation between each pair of features.

Data Normalization:

- We applied Min-Max scaling using the **MinMaxScaler** from the **sklearn.preprocessing** module to normalize the 'recovery' and 'death' features. Normalization ensures that all features are on the same scale, preventing certain features from dominating others during analysis.

Handling Missing Values:

- This step in finished in phase 2. There was no explicit handling of missing values mentioned in the provided code snippet. If there were missing values in the 'recovery' or 'death' features, appropriate strategies such as binning method have been applied to handle them in phase 2.

Data Transformation:

- We transformed the original features 'recovery' and 'death' into their normalized counterparts, 'recovery_normalized' and 'death_normalized', respectively. This transformation ensures that these features are now scaled appropriately for further analysis.

Overall, the preprocessing steps undertaken aimed to ensure the dataset's readiness for subsequent analysis tasks. The process focused mostly on normalization because of unavailability of categorization.

Part C

Steps to do OneClassSVM:

In phase 4, our group imports the OneClassSVM machine learning algorithm from the scikit-learn library and imports the train_test_split function, which is used to split a dataset into training and testing subsets. We set x as 3 different types of vaccination data (‘partial’, ‘fully’, ‘booster’) used to train or test the machine learning; set y as ‘totalcases’ used to represent the target variable or labels that the model aims to predict. The scale of test data is defined as 30% mentioned in class. Next step, we initialize the model through setting clf instance, nu=0.01, kernel = radial basis, gamma = 0.1. After all, we run the code for training the model, making predictions, and counting outliers. The OneClassSVM detects 48 outlier data.

Analysis:

1. Some negative changed data: Before detecting outliers, we define three columns measuring the change of all three vaccination types(partial_change,fully_change,booster_change). We notice that there are some negative changes in fully_change and booster_change. We suppose they are outliers in our datasets caused by data collectors or the group of people shifting to next group of people like shifting from partial to fully, fully to booster. The one-class SVM algorithm do detect some negative changed data as outliers.

2. Booster is zero or close to zero: When fully and partially grow wildly, boosters are not active and are regarded as outliers. We think it is normal for the booster to be inactive at first, because people only achieve booster injections after multiple vaccinations.

```
The number of outlier is 48
Outlier Data:
       partial     fully   booster
0       110118         3         0
1       300249      3941         0
4       968205    111205         0
5      1081963    172339         0
10     2983627    611653        16
25    35394551   9768213      1069
28    45300763  18699268      3420
33    52563038  24851771     14789
41    57671217  27732389    376367
```

3. Large-scale variation: The initial number of booster vaccinations increased very slowly, but the subsequent number of vaccinations increase rapidly. Some intensively rising numbers may be detected as outliers. We speculate that other groups of people were move into booster as time goes by.

	partial	fully	booster
11	3922570	646557	19
16	11932958	1038690	65
17	13782804	1153405	94
21	23353297	2013619	256
22	25983643	2786610	377
29	47661261	20801744	3893
35	53960676	25774085	24200
38	55906674	26797552	160770
43	58717091	28291344	583059
47	60450747	29007074	1191885
48	62020386	29122031	2183650
62	81657327	31020758	17862699
65	82265892	31197417	18198676
68	83596095	31290709	18444581
69	84066328	31311704	18526309
71	84782244	31346409	18589083
74	85644848	31409138	18708997
82	87589077	31494992	18970730
85	88420782	31527377	19056954