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 and any encountered data quality issues:

**Feature Selection:**

* + Initially, we selected the 'recovery' and 'death' features for normalization. These features are crucial in understanding the progression and impact of COVID-19 cases.

**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.

文本

低可信度描述已自动生成

1. 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.

图形用户界面, 文本

描述已自动生成