Field	Dealing Method	Reason
case_month	Do nothing	Data correct
res_state	Do nothing	Data correct
state_fips_code	Do nothing	Can accurately match the res_state
res_county	Delete	Cannot match county_fips_code
county_fips_code	Do nothing	No ambiguity since res_county is deleted
age_group	Do nothing	Missing data has no negative affect
sex	Combine "Missing" and "Unknown" into "Missing"	Same meaning
race	Combine "Missing" and "Unknown" into "Missing"	Same meaning
ethnicity	Combine "Missing" and "Unknown" into "Missing"	Same meaning
case_positivity_rate	Delete	Too many missing values
exposure_yn	Combine "Missing" and "Unknown" into "Missing"	Same meaning
current_status	Do nothing	Data correct
symptom_status	Combine "Missing" and "Unknown" into "Missing"	Same meaning
hosp_yn	Combine "Missing" and "Unknown" into "Missing"	Same meaning
icu_yn	Combine "Missing" and "Unknown" into "Missing"	Same meaning
death_yn	Do nothing	Data correct
underlying_conditions_yn	Do nothing	Data correct
case_positive_specimen_interval	Take absolute value	Negative values may be caused by inverting the subtrahend and subtrahend
case_onset_interval	Take absolute value	Negative values may be caused by inverting the subtrahend and subtrahend
Process	Delete	Too many values missed(more than 80%)

Reason:

①res_county：Delete.

Here we find that the total number of counties is less than the total number of county zip codes. It is unlikely that a county will have two zip codes. Therefore, we can choose to delete one of the two. Due to the large number of zip codes, the coverage may be wider. So instead of deleting the county_zip code group, we delete this group.

②case_positivity_rate, process：Delete.

The data missing rate in this group is as high as 91.2% and 88.5%. Among the remaining data, the data distributed under different attributes does not exceed 2%. It is difficult to produce practical significance in statistics, so it is deleted. Another optional solution is to remove "Missing", keep the rest, and mark Missing as None. But this does not produce practical significance, so this option is not selected.

③case_positive_specimen_interval and case_onset_interval：Take absolute value

For the processing of this set of data, there are two general schemes: ①Take the absolute value; ②Directly discard the negative value. Negative values have to be dealt with because they are clearly not clinically meaningful. Considering that when collecting data, the "detection date" and "report date" may be reversed, and the calculation result may be negative, so based on the protection of data, I prefer to choose to keep these negative values. and correct them to absolute values.

④sex, race, ethnicity, exposure_yn, current_status, hosp_yn, icu_yn：Combine "Missing" and "Unknown" into "Missing"

Missing and Unknown express the same meaning here, just keep one of them. In addition, the option of marking all "None" was also considered. But from the perspective of easy understanding of medical statistics, it is more concise to choose Missing.

⑤Else: Do nothing

The data is correct, the options are correct, and no processing is required.