CMPT459 Milestone1

ZiZe (Max) Zhang: 301301987
Siyuan (Leon) Wu: 301313026
LiZhou Ding: 301326630
Covid19 case dataset

1.1 Exploratory Data Analysis

The case dataset NaN data calculation:
{'age': 296874, 'sex': 293734, 'province': 6568, 'country': 24, 'latitude': 2, 'longitude': 2, 'date_confirmation': 462, 'additional_information': 522969, 'source': 209191, 'outcome': 0}
The location dataset NaN data calculation:
{'Province_State': 168, 'Country_Region': 0, 'Last_Update': 0, 'latitude': 80, 'longitude': 80, 'Confirmed': 0, 'Deaths': 0, 'Recovered': 0, 'Active': 2, 'Combined_Key': 0, 'Incidence_Rate': 80, 'Case-Fatality_Ratio': 48}

1.2 Data cleaning and Imputing missing values

Cleaning case dataset:

Actions performed on str type in ‘age’ imputation

1. For ages with dash sign ‘-’, drop the sign and use the mean value of the two ages
2. For ages with plus or minus sign, remove the plus or minus sign
3. Convert all forms of string numbers to integers
4. For ages with month value, we add the integer result of the month divided by 12 to the age value
5. Drop the age data that have unknown characters in them

Action performed on nan data imputation

1. Around 20 percent of the data are missing age or sex, therefore it would have a large impact if simply remove those data
2. For nan data in age attribute, replace nan data with the median value obtained from the age attribute
3. For nan data in sex attributes, calculate the ratio of the two genders and assign the gender according to the ratio.

Cleaning location dataset:

numerical data:

- Active: using the formula ‘ Active cases = total cases - total recovered - total deaths’ to evaluate the missing Active value. If the value is negative or non-feasible then add 0 to the missing index
- Incidence_Rate: using the formula ‘Incidence Rate = cases per 100,000 persons’ to evaluate the missing Incidence_Rate. Since there is no missing value for confirmed cases. Non-feasible values will be eliminated in the outlier section.
- Case-Fatality_Ratio: follow the formula ‘Case-Fatality Ratio (%) = Number recorded deaths / Number cases’ to fill in the missing value. Check if the Number cases is greater than zero, otherwise fill 0 in the missing index

- Latitude and Longitude: Since the missing value of latitude and longitude are 81, which are small enough to neglect, dropna() was applied to remove the missing value

categorical data:

- Province_State: Since there are no extra evidences that support for this value, the optimal way would be removing the value
- Last _Updated: After the group's discussion, we all think that the last updated value would not influence the outcome. Hence the column of last updated value is dropped.

1.3 Outlier processing

For both data, the first step we take to deal with outliers is to remove data that are clearly wrong. For example, negative age entries or negative active cases entry will be removed during this process. If we demonstrate the data in different attributes of the location dataset, many consecutive data points would be marked as outliers, and this may be caused by some locations having extremely high active cases. Therefore, outlier detection using IQR scores will lead to massive data loss. To remove data entry errors while keeping correct values as much as possible, the top and the bottom 10 percent of the data will be removed. By having this method, skewness will be fixed back to the normal range of 1 to -1.

A separate dataset "cleaned_cases.csv"will be created for joining

1.4 Transformation

In order to aggregate the information in location dataset from city level to province level in the US, the method groupby() is used. After applying groupby().sum(), we will get a full set of summed up 'Active', 'Deathes', and 'Confirmed' cases from all the cities in each province. For 'Incidence rate' and 'case fatality ratio', the percentage will be divided by the number of cities in their own province. Last of all, the name of 'US' will be replaced by 'United Stated' to help prepare for the joining in 1.5.

Two separate dataset will be created:

- temp.csv is stored the grouped data
- location_aggregated is the dataset that created for joining

1.5 Joining Attributes

For the joining process, we chose to use the province and country as the join attributes. One reason is that using province and country makes more sense for data analyzing as the data will be most likely to be analyzed province by province or country by country. Furthermore, from the accuracy perspective, two locations with close latitude and longitude may have very different death rates or active cases. The reason for that may be because they belong to different countries or provinces which are having different pandemic policies. Therefore, using country and provinces is best suitable for joining in this case.