1.1
We used the panda's library to obtain the data and manipulate it such that the class label is consistent across all records, which is necessary for data mining algorithms that we will use in the future.

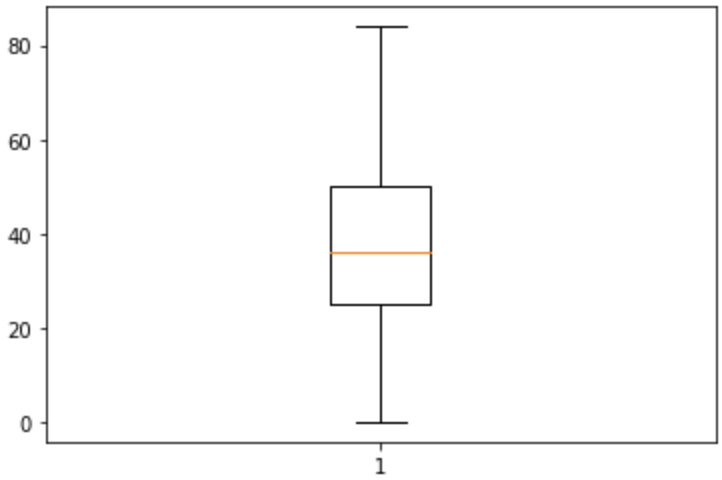
1.2
This is a classification task

1.3
Again, we went through the pandas library and utilized Jupyter Notebook to construct visual representations of the data for the tasks specified. We generated and analyzed histograms and scatter plots, as well as a table displaying the amount of NaN values for each column. For certain histograms, we used a log-based graph to better visualize the data, as using a regular histogram schema was not very useful due to the scale of data from specific countries (e.g., India).

As a part of 1.5, we did some outlier visualization. We used boxplots to verify that the age attribute had some outliers, and neutralized them in 1.5

1.4
We removed missing age/date values from test and train CSV. There is no reasonable way to impute either of these attributes, and as there are so many rows where both are missing, we decided that any attempts at imputations would lead to non-sensical data. For example, it wouldn't make any sense to take the "average age" of people in India and assigning this to the age of any missing age attributes. As the number of missing fields in sex was a lot, and removing all the null ones would have led to reduction of important data. We decided to randomize the sex, as probability of getting covid is same in all sex. One thing that we did not tackle was the impartial distribution of sex in Indian population.
Since age, date and sex may be relevant to the disease, rather than removing the age, date and sex attributes, we removed the rows containing missing age and date values and filled sex with random values. On the other hand, the additional information attribute was entirely removed, as it is unlikely to be relevant for the data mining task and most of the information it provides is very variable and inconsistent. Sometimes it's just an acronym, and most of the time it's the empty string. It would also require making a word dictionary and significantly increasing the number of attributes to transform it into a suitable format, which would not be worth the effort given the lack of utility of the attribute.
From locations.csv we removed blank rows of Incident_Rate, Case_Fatality_Ratio. We were able to remove them because there were not many missing values to dilute the data.

1.5
We used some data visualization to make the outliers apparent. We decided to remove the rows that contained these outliers.
As we can see in this image



Age above 85 was removed as an outlier.

1.6

Since some provinces are missing from the location dataset, we decided to join the cases to the location dataset by country. We first updated the location dataset: For each country, we aggregated the relevant attributes (Deaths, Recovered, Active, Incident Rate, Case Fatality) using either sum or average functions and discarded irrelevant features (latitude / longitude – these are already captured on a per-case basis in the cases training set).

Problem: Some countries only kept track of how many confirmed cases and deaths occurred, and did not keep track of the recoveries, e.g., Belgium, United States, Serbia. A solution we can implement is to simply subtract Confirmed – Deaths to obtain an approximation of the amount of recovered (keeping in mind that some of the confirmed may be active cases, and may potentially die – however, this is a relatively small proportion). A better solution may be to simply remove the "recovered" column all together, as the Case_Fatality column should provide the data mining algorithm with all the information it needs, making Recovered redundant. (see 1.7)

1.7

Columns to remove from datasets:

- Additional information (Justification in 1.4)
- Recovered (Justification in 1.6)
- Province (Not always filled out, and we think that having both Country and Latitude/Longitude will be sufficient for providing the data mining algorithm with geographical information and cultural information.
- Outcome – We replaced it in 1.1
- Source – Intuitively irrelevant to classifying outcome groups.
- Active – Unlikely to be relevant, and possibly made redundant by date_confirmation, which provides more information regarding time.

- Confirmed – Case_Fatality is Deaths / Confirmed. This is much more relevant to the task of classifying outcome groups, making both Confirmed and Deaths irrelevant.
- Deaths - See above

Columns to keep in final dataset:

- Country – Different countries have different methods of stopping the spread. Certain races may be more or less susceptible due to biological differences.
- Incident_Rate
- Case_Fatality - Gives metric for likelihood that any given person who has coronavirus will die from it. Very relevant.
- Age – Older people are more likely to die from COVID
- Latitude/Longitude – Provides geographical information
- Chronic_disease_binary – Provides useful information about the patients health