Ritvik Shah 301366819
Jusung Lee 301309867
Alan Thomas 301363048

Report

1.1 Exploratory Data Analysis

For casesTrain, the data was first cleaned using the cleaned data from 1.2. We first explored to see if gender had an impact on the outcome by creating a box plot and a violin plot versus age. Then we created a scatter plot on lat and lon colored by outcome to see if any regions in the world showed a trend. Next we studied the avg, std, min and max for age, latitude and longitude.
We learn that gender does not impact the outcome and that there are some hotspots for certain outcomes.
For location, the data was first cleaned from 1.2. We first made a scatter plot on confirmed deaths, recovered, active, incidence_rate, and case fatality ratio. Since they showed an exponential growth trend, we took the natural log of all of those columns to see if there were any linear relationships between these columns. We also studied the avg, std, min and max for those columns as well.
Finally we took the data from 1.5 and graphed the columns mentioned in location and plotted them against date_confirmation to see growth trends.
We learn that confirmed, deaths, recovered, active, incidence_rate, and case fatality ratio columns are exponential growth and should be logged.

1.2 Data Cleaning and Imputing Missing Values

In data cleaning we took the average of all the range ages. For example ‘10-20’ would be translated to 15. All ages that were less than a year old were bumped up to 1 years old. We parsed all strings to be numeric. All missing ages were imputed to be the average age. This ensures that those rows won't impact the age attribute since it's the average age. Missing values for sex, we created a third category “unkown”. We deleted province, country, lat and lon. We thought that a combined key of province and country represented those columns. We then deleted a small number of rows that had missing combined keys and date_confirmation.

1.3 Dealing with Outliers

The case dataset for training did not have any outliers as some of the columns like sex, province, country, etc were categorical data. The age column seemed to have data that were within the acceptable range and so did not have outliers in it. The latitude column had data within the range of -90 and 90, and the longitude columns had data within the range of -180 and 180.
The location dataset seemed more likely to have outliers as it had several columns of continuous data. We created boxplots for the columns with continuous data but the graphs were very spread out making it unclear so we logged the data in the columns to new columns and created boxplots on their log data which gave a clearer display of the outliers. Since these data grow exponentially the outliers are still valid data.

1.4 Transformation

In order to transform the information for cases in the US from a country level to a state level we iterate through all the rows that have a province located in the US and perform a set of aggregates. For each province, we add up the following columns:

- Confirmed
- Deaths
- Recovered
- Active

For the values of latitude and longitude we simply store them for each state.

In order to calculate the incidence rate we need to obtain a dataset that contains the population for each state. After we have merged the population to our existing training dataset, we calculate the incidence rate as follows:

Incidence rate = (confirmed / population) * 100000

In order to calculate the case fatality ratio we perform the following calculation:

Case fatality ratio = (deaths/confirmed) * 100

1.5 Joining the Cases and Location dataset

In order to join the training dataset and the location dataset we need to make a combined key for the training dataset. This is because we cannot use longitude/latitude values because the training dataset and the location dataset seem to have slightly different values for the same location. We also cannot use either province / country because the location dataset has a large number of missing values for the province.
We perform a left join on the dataset so that we ensure that we retain all the records in the training set, and simply use the location dataset to make the data more detailed. Moreover, performing a left join will ensure that only the rows that have the same combined key as the rows in the training dataset will be added, ensuring that there is no redundant information.
To clean up the merged dataset we drop the repeated latitude/longitude columns and both sets of province and country columns, as the combined key provides the same information.

1.6 Outcome Labels

Recovered - Recovered from covid
Hospitalized - has covid and had to be hospitalized due to severity of their health
nonHospitalized - has covid but did not have to be hospitalized due to covid not having a severe impact on their health
Deceased - died to covid
Data mining task is classification on the outcome labels