CMPT 459 FALL 2020
Milestone 1 Report

Leo Chen 301276105
Gerland Lok 301260310
Parsa Alamzadeh 301316272
Simon Fraser University

1.1 Exploratory Data Analysis

We did exploratory analysis in this section to get an idea of our data. We plotted global concentration of covid using the longitude and latitude values of the individual dataset.
We also plotted important basic stats for each country such as the confirmed cases, deaths, incidence rate, and case-fatality ratio. We printed the number of empty values in each attribute. The plot world_map.jpg includes the number of confirmed cases and fatality rate for given coordinates (based on the location dataset), the radius of the points corresponds to the number of confirmed cases and the color represents the fatality rate at that given point. Further, the gender_graph.jpg visualizes the number of cases for each gender for different age groups. There are more visualizations available in the graphs directory. The number of missing values are also available in the Milestone_1.ipynb file.

1.2 Data Cleaning and Imputing missing values

We did age conversion. For age ranges we used the mean of the upper and lower bound to represent the age of the patient and for ages that we had an age and above, we use the base age for representing the patient's age. For patients that their age were represented in months, we converted those into years. For dates, they were represented as “dd.mm.yyyy”, and some were represented as “dd.mm.yyyy - dd.mm.yyyy”; for ranged dates we used the initial date and for the rest we converted them to datetime object.

1.3 Dealing with outliers

We determine outliers by first plotting our data and discovering any data points that seem irregular. For location data, we dropped the values that had empty longitude and latitude. We found that these rows were usually cruise ships or unknown locations. For individual cases we also dropped datasets with irregular locations by checking their latitude/longitude and country. For some countries we came across rows with state/province set to “Recovered”.

1.4 Transformation and US State Data Aggregation

We aggregated the counties by first using a Weighted Mean Centre formula that takes in the confirmed cases, longitude and latitude data of every county of a specific state. This would give us the average centre of all confirmed cases of COVID-19 of each state in the United States represented by a single latitude and longitude coordinate. For the confirmed cases, deaths, recoveries, and active cases metrics, we summed up all of these values from the individual

counties for each state. Incidence Rate is the measure of how many confirmed cases of COVID-19 there are per 100k residents. We decided to recalculate this value for the state level. To do this, we reverse engineered the Incidence Rate formula to extract the population of each country. To find the population of each county, we used the formula:

population = (confirmed cases x 100k) / Incidence Rate of County

We then summed up the population values for each county in the state to find the total population of that state. We already know the total number of confirmed cases per state as it was calculated prior. So we can use the following formula to find the state Incidence Rates:

State Incidence Rate = (Total Confirmed Cases x 100k) / State population.

Finally, for the Case Fatality Ratio, we also decided to recalculate this value at the state level. To do so, we used the Total confirmed cases and Total deaths values we found in the earlier steps and plugged them into the following formula to determine the state Case Fatality Ratio:

Case Fatality Ratio = (Total Deaths / Total Confirmed Cases) x 100

1.5 Joining the Cases and Location Dataset

The joining of the Individual Cases and Location Datasets we chose to join based on “Country” and “Province” pairs. This was done in four phases. Firstly, we extracted the Individual Cases that had known Country and Province attributes and joined those cases with the corresponding data for those Country and Province pairs in the Locations dataset. This became our working data frame. We then joined US cases with the US State level aggregates we’ve compiled in 1.4 and appended the resulting data frame to our working data frame. Finally, to handle the individual cases which only had a known Country attribute (Province value is empty), we decided to aggregate entries in the Locations dataset to the Country level. For instance, this would give us aggregate values for countries such as Australia and the United Kingdom for attributes such as Confirmed Cases, Deaths, Latitude, Incidence Rate and etc. This would be similar to what we did in 1.4 except we’re aggregating to the Country level rather than State level. Once we’ve got our dataframe with these aggregated values for every country, we merged again