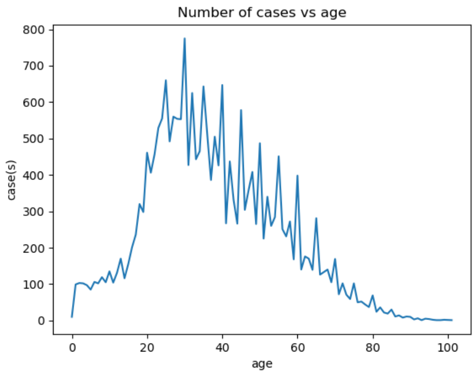
Yuhang Yuan (301357634)
Sunny Yang (301351209)
Zhe Liu (301449316)

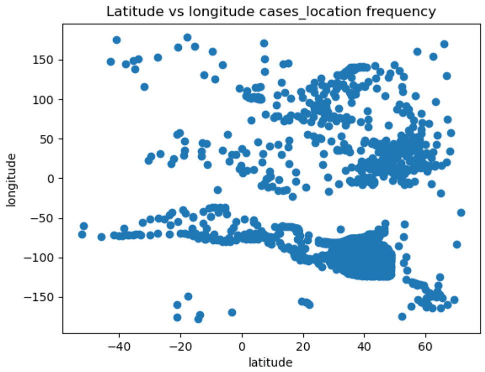
CMPT459 Milestone 1

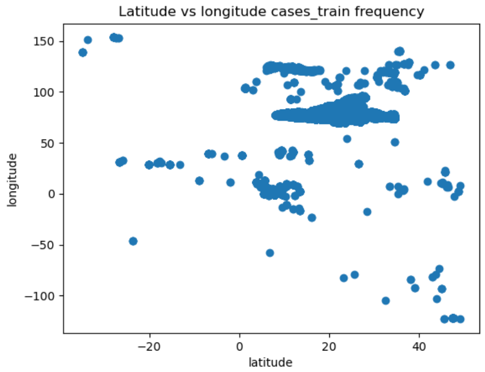
1.2 Outcome labels

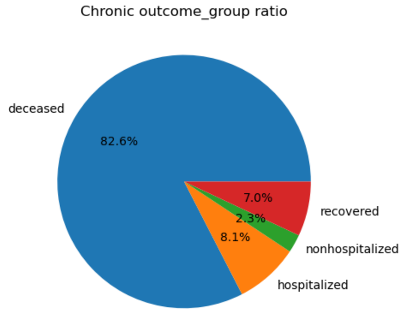
- Within the training data set, there exist four main categories of possible values for the **outcome_group** column; "hospitalized, nonhospitalized, deceased, recovered." However, due to the messiness of real-world data, these outcomes are expressed in multiple methods.
    - Within the training data, we used the data mining task of *preprocessing*, and more specifically the method of "word clustering" which groups similar corresponding outcomes into their respective **outcome_group**
    - In order to predict outcome_group labels in the test data, the mining task of *classification* would be used.

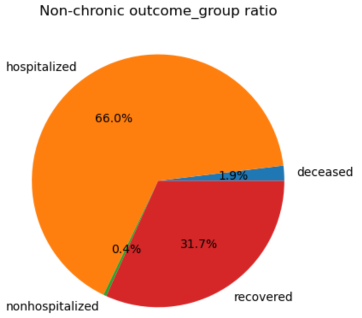
1.3 Exploratory Data Analysis

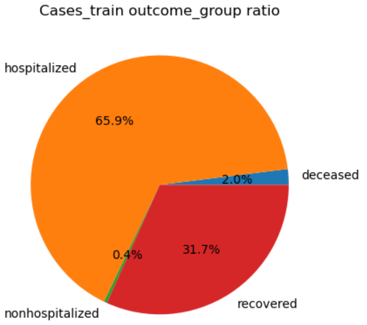












1.4 Data cleaning and imputing missing values

- **AGE**: Within the attribute of age exist entries with various formats. Our approach is to treat all data initially as a float. This method would preserve the accuracy of values in the event of calculations. Namely, the average of intervals such as 20-29 would be taken and imputed as the value for that row. Upon finishing, all floats will be rounded and converted to a standard integer.
- **SEX:** Null values are dropped and not imputed due to the binary and categorical nature of the attribute. It seems unprofessional to randomly assign either male/female to the missing values.
- **Province/Country:** Imputed missing values with a placeholder value ("unspecified"). The reason being rows with missing **Country** could be due to the special nature (politically) of the location such as "Taiwan." While rows with missing **Provinces** are due to the fact that some countries simply did not record at the sublevel of provinces. Overall, the missing values do not account for a significant portion of the dataset, however, we felt that it was necessary to conserve these values as dropping would imply neglecting data gathered from an entire country/region.
- **Latitude /Longitude:** Geo-coordinates provide precision when in location data. Imputing through approximation would defeat its intention for precision. For instance, what coordinates would be used if a value originated from **country** = Canada, which covers 9.985 million km^2 .Therefore missing values within Latitude and Longitude are dropped.
- **Source and Additional_Information:** Nulls within these two attributes are imputed with a placeholder value of ("none") to conserve the attributes for future use.

- **Recovery/Active:** Data from the US are missing, and since they make up the majority of the rows in the location dataset, it is unrealistic to impute or drop. Thus we filled with a placeholder value of -1 to maintain the attributes for possible future use
- **Case_Fatality_Ratio:** The validity of missing values is first checked, cases such as (confirmed = 0) but (death = some value other than 0) are dropped. Followed by imputation through computation of "death" divided by "Confirmed" of the respective row. Values that continue to be null due to the scenarios such as 0 death divided 0 confirmed, would be imputed to 0 instead

1.5 Dealing with outliers

- Potential outlier containing attributes

Attributes	**Identification method**	**Response**
Dataset: Training/test		
Age	68-95-99 rule, where any value ±3*standard deviation from the mean is considered an outlier. With a mean of 37.69 and STD= 17.34; bounds were set at 0 ≤ age < 89.7 The 0 lower bound is due to humans being age = 0 upon birth	Outlier values were **removed** as an age of over 89.71 is already above the average life expectancy of humans in 2022. In the cases that such humans do exist, they should only represent a tiny percentage of datasets. Thus, can be neglected.
Latitude/ Longitude Lat/ Long_	Set up bounds of -90° ≤ Latitude ≤ 90° and -180° ≤ Longitude ≤ 180°, out of bound values implies the location does not exist on Earth	Since values were of geographical coordinates which represent real world locations, minimal room available for the existence of outliers. Outliers are **removed**.
Location		
Death, Recovered, Active	Confirmed ≥ Death, Recovered, Active	Due to being realistic records of the human population, it does not seem justified to use statistical methods to check for outliers. A more linear approach as seen in the "identification method" is utilized for the appearance of outliers. Outliers are **Removed**
Case_Fatality ratio	68-95-99 rule, setting bounds as 0 ≤ CFR < mean(2.02) + 3*STD(1.51) Lower bound = 0 due to no such thing as negative fatality rate	Outliers that breach the lower bound are **removed** as negative fatality rate is impossible. Outliers breaching the upper bound are **kept** due to the possibility of data coming from concentrated communities.

1.6 Joining the cases and location dataset (province/country)

- Joined through left join
  - Train X location: 20587 rows
  - Test X location: 10197 rows

1.7 feature selection selecting

Selected:

- **Province/country:** Represents Outcome group for locations all over the world. Based on the location, safety measures for covid might be different and thus have varying effects on the outcomes
- **Age, Sex, Chronic diseases:** According to known research, these three attributes do have a direct correlation with the effects of covid on patients and thus are kept in our prediction model
- **Case_fatality_rate, outcome_group:** Quantifiable, referenceable and probabilistic method of representing covid outcome stats as opposed to arbitrary numbers seen within confirmed/death/recovered/ active

Discarded attributes:

- **Latitude/ Longitude:** We think the geological location is too precise and unnecessary for determining outcomes. Outcomes distribution between two points such as Vancouver and Richmond should not vary much as the population of the two can easily traverse between each other. The precision would only lead to extra computation
- **Source/addition information**: mostly empty values and has no apparent link to outcomes. May have usage but require an additional machine learning algorithm to recognize keywords
- **Date_confirmation/ Last_update:** Realistically, only validates data's ability to accurately represent the status of the current world. However, does not have a direct influence on the outcome results.
- **Combined_key:** Redundant information seen within province/country
- **Confirmed/ death/recovery/active:** arbitrary values that could mean anything without context. Can only be useful when combined/compared with other values and attributes such as "case_fatality_rate"
- **Incident_rate**: tells infection rate of covid within specified populations but is not detailed to differentiate the subset outcomes we are predicting such as hospitalized/ hospitalized/ deceased/ recovered.