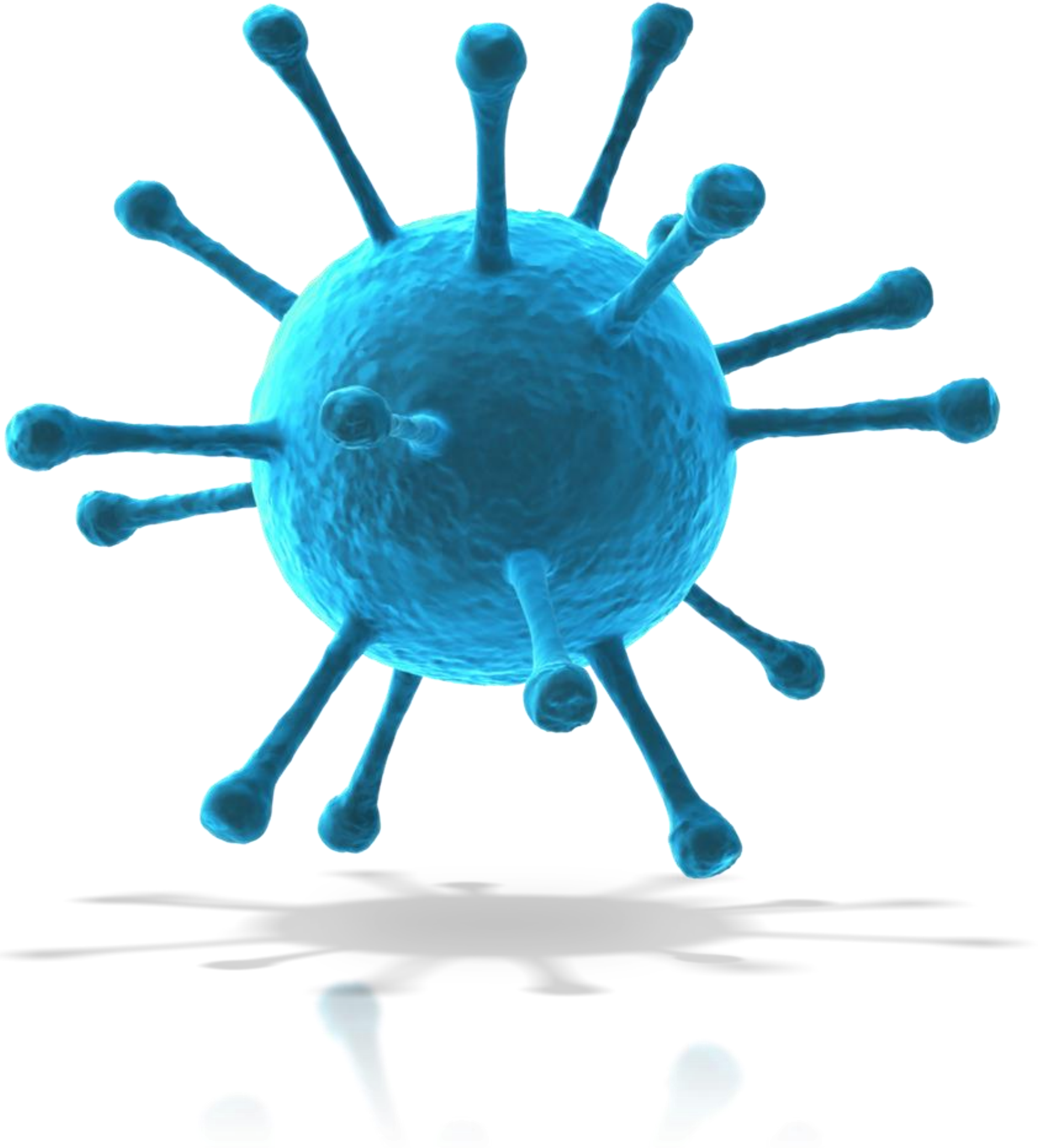
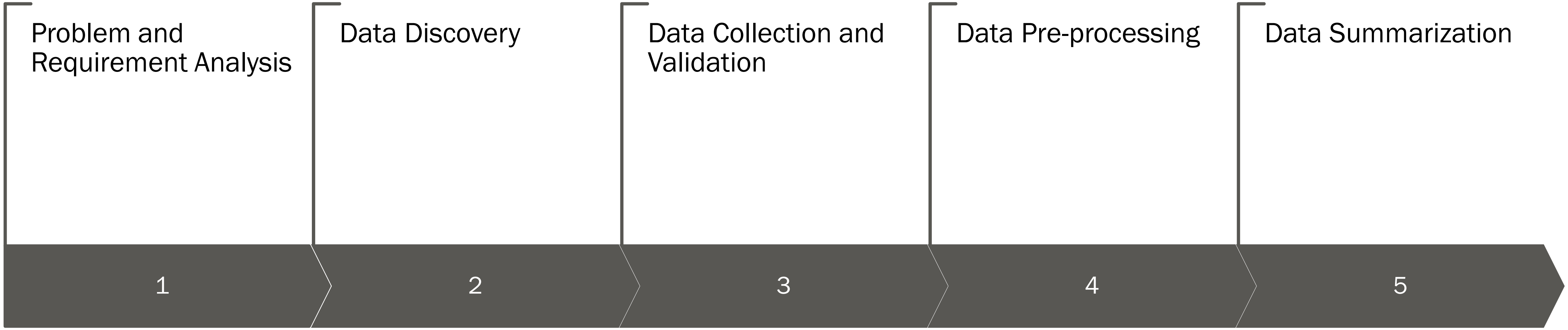


CASE ANALYSIS OF NOVEL CORONA VIRUS COVID 19

PRIYA PAYYADI SURENDRAN

ID:1533656

TOPICS DISCUSSED



1. Problem and Requirement Analysis
2. Data Discovery
3. Data Collection and Validation
4. Data Pre-processing
5. Data Summarization

PROBLEM AND REQUIREMENT ANALYSIS

1. Application Background of the proposed data mining

- Topic chosen is the analysis of the cases of COVID-19 in New Zealand from 26th February 2020 to 1st August 2020 which can be applied to medical analysis to get a better insight on how to manage health crisis of this virus outbreak.
- How the virus spread pattern is.
- Health care providers, organizations, government can anticipate the resources and facilities to manage and tackle the situation.

2. Stakeholders of the research

- People
- Health care Organization
- Hospitals
- Government(Federal, state & Local)
- Employers
- Financial Institutions(Banks and Insurance company)

3. Data mining analytics tasks

- **Numeric Prediction:** Daily cases reported (confirmed or probable). This can be used to compare and arrive at the conclusions like probability of COVID-19 being completely eradicated or probability of increases in cases reported.
- **Association:** Between age, sex and cases being reported. This can be used to analyse whether special preventive measures among males and females of various age group can be taken or if already taken will help them to recover or fight against the virus.

DATA DISCOVERY

1. Operational Procedures

- Proper identification of a problem or an opportunity that can be analysed (COVID-19 case analysis of New Zealand)
- Find in detail the association of the problem or what exactly should be analysed and why exactly it is happening (study the pattern of the virus COVID-19 in different age group and sex)
- What data to be collected, how to be collected and from where to be collected(data from Ministry of Health New Zealand from the start of the outbreak of the virus on 26th February 2020 till 1st August 2020 of different age group and sex)
- The data collected and prepared is valid and useful for analysing the problem or opportunity.

2. Data in Demand

- Age group of the person infected
- Sex of the person infected
- Date on which case was reported
- Geographical location
- Recent travel details
- Status of the case
- Deceased status
- Details of health centre where the person is admitted(resources available or not)
- Region wise count of cases per day
- Recovery status of the person infected
- Recovery modes of the person infected(hospitalised or home remedies)
- Lockdown status

3. Data Available

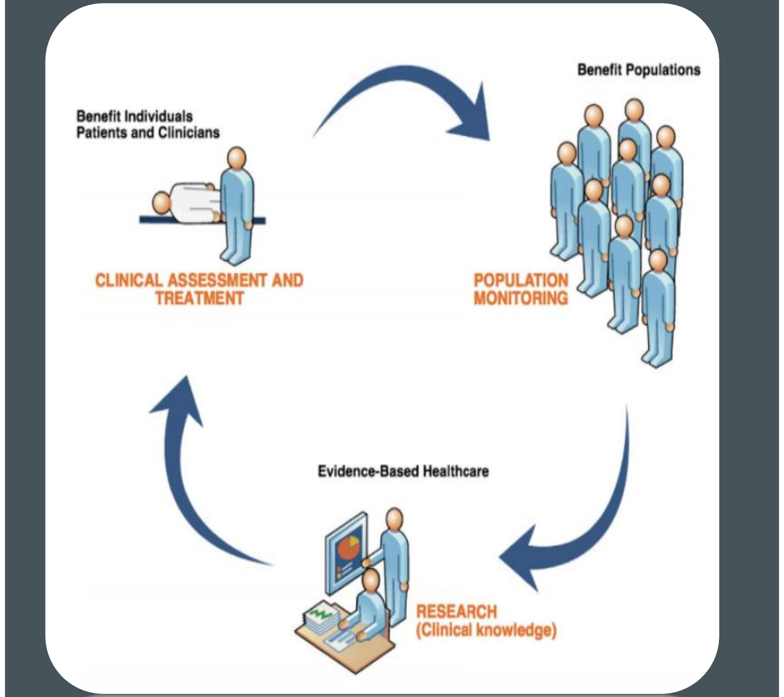
- Age group of the person infected
- Sex of the person infected
- Date on which case was reported
- Recent travel details like last country visited, flight number, date of departure and arrival
- Status of the case (Confirmed or probable)
- Deceased status
- Region wise (DHB)count of cases per day

4. Gap between data Available and data in demand

- Recovery mode(hospitalized or not, if hospitalised how? etc)
- Recovery status
- Details of health centre where the person is admitted(resources available or not)

DATA COLLECTION & VALIDATION

1. Software setup for data validation



- The software used is SNOMED CT by the SNOMED International as a part of global effort to manage this pandemic. SNOMED CT is used to record, share, integrate and analyse COVID-19 data.
- SNOMED CT can be used to capture COVID-19 related data for data elements like Provider & Facility Details, Patient Demographics, Clinical Assessments, Tests & Investigations, Prevention, Treatment & Education.
- There are SNOMED codes for each data element, and these are stored in the patient’s record by health care providers and these are stored by each DHBs in a local database. Later, Ministry of Health collects these data and publishes it for general people after processing it.

2. Hardware setup for data validation

- The hardware used to collect COVID-19 data are swab collection kits and testing equipment of laboratories.
- The swabs are mainly collected from upper respiratory parts like nose and throat and are called nasopharyngeal and oropharyngeal swabs.
- RT-PCR(Reverse Transcription Polymerase Chain Reaction) technique is used to test these swab samples

3. Requirements for Data Collection

- Data collected is related to the research
- Data collected must be valid.
- The data collected should be the data in demand and there should not be a gap between data in demand and its availability
- The data gathered should be useful for the stakeholders
- Data collected to analyse the problem or the opportunity under the research must answer all the research questions.

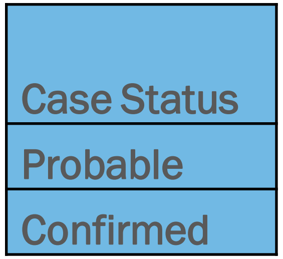
4. Data Validation

- Make sure that the gathered data helps the researchers to tackle the problem or opportunity under question.
- Data validated will be valid and clean.
- Convert the raw data by using data transformation and data pre-processing methods like cleaning, integrate , transform

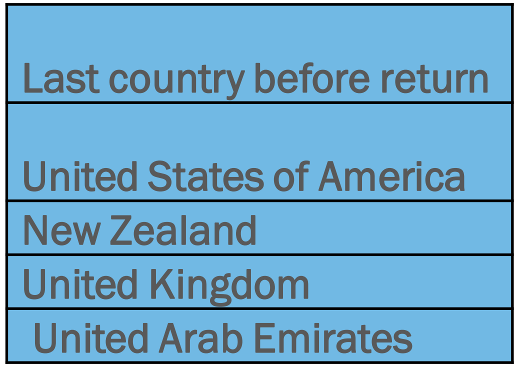
DATA PRE-PROCESSING

1. Data Integration from different sources

- Attributes must have same name.

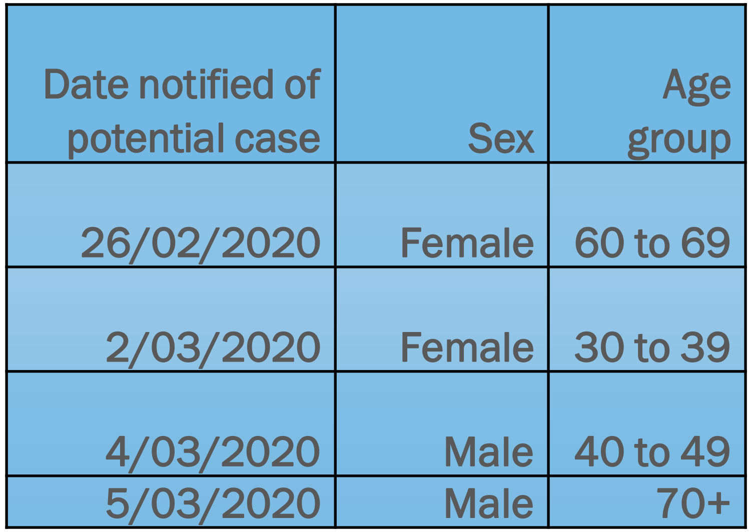


Case Status
Probable
Confirmed



Last country before return
United States of America
New Zealand
United Kingdom
United Arab Emirates

- Attributes must have same representation (Date format, age group categorization)



Date notified of potential case	Sex	Age group
26/02/2020	Female	60 to 69
2/03/2020	Female	30 to 39
4/03/2020	Male	40 to 49
5/03/2020	Male	70+

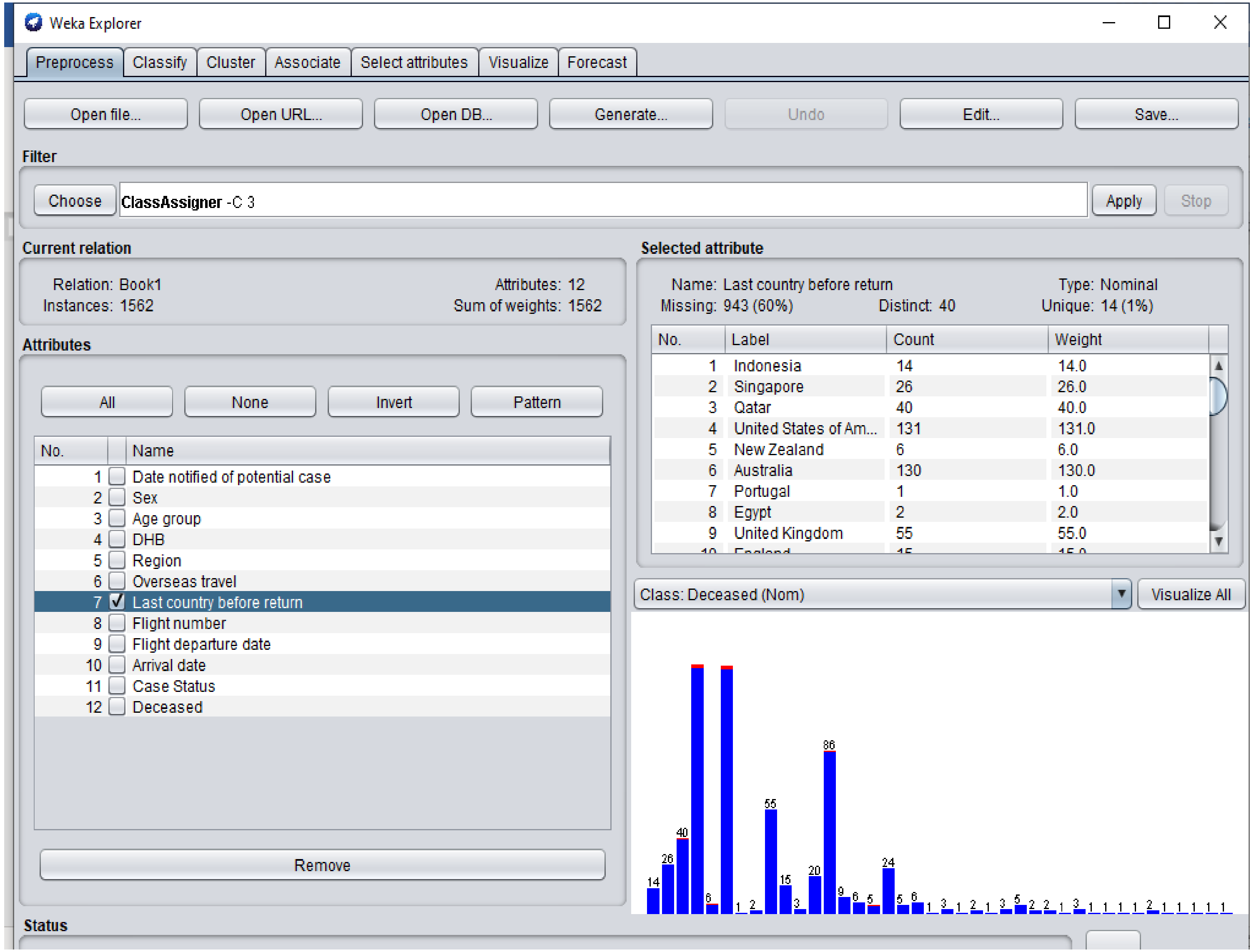
2. Data Cleaning

- Real world data are dirty, noisy, inconsistent and incomplete
- Missing values: ignored, using a constant like “unknown” to fill in or impute the missing values using attribute mean or decision tree
- Noisy data: Clustering, detecting the suspicious values by a combined inspection by human and computer or smoothening it out by fitting the data in to regression functions.

2. Data Cleaning

- My dataset contained many missing values : attributes like ”Overseas Travel”, ”Last country before return”, ”Flight number”, ”Flight Departure date” and “Arrival Date” .
- I replaced the missing values with the constant value “unknown” which seems to be more apt for the research.
- The attribute “Last country before return” to New Zealand has 943 missing values and I used the filter “ReplaceMissingValuesWithUserConstant” and substituted it with constant value ”Unknown”.

2. Data Cleaning



Filter

ClassAssigner -C 3

Current relation

Relation: Book1
Instances: 1562
Attributes: 12
Sum of weights: 1562

Attributes

No.		Name
1	☐	Date notified of potential case
2	☐	Sex
3	☐	Age group
4	☐	DHB
5	☐	Region
6	☐	Overseas travel
7	☑	Last country before return
8	☐	Flight number
9	☐	Flight departure date
10	☐	Arrival date
11	☐	Case Status
12	☐	Deceased

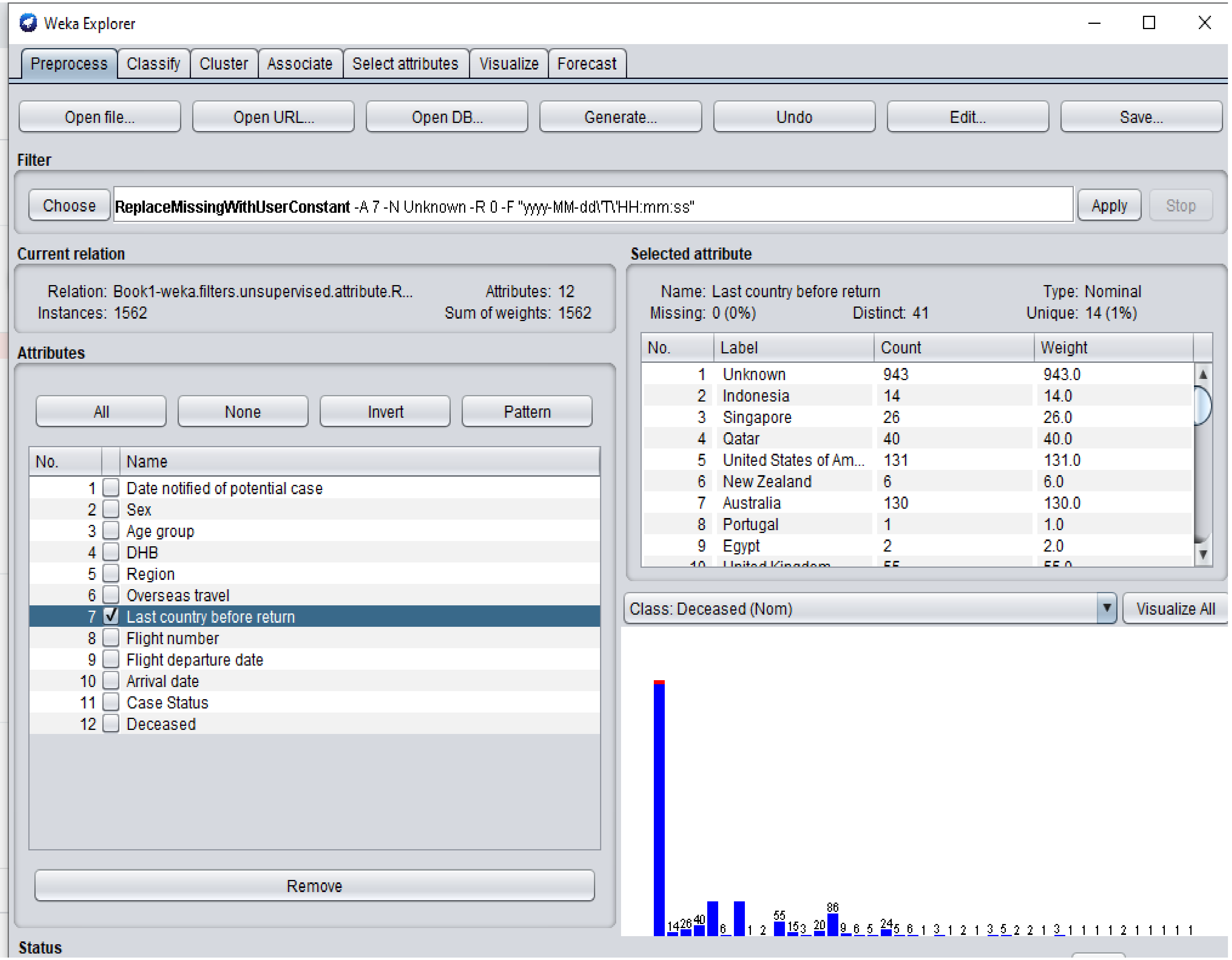
Selected attribute

Name: Last country before return
Missing: 943 (60%)
Distinct: 40
Type: Nominal
Unique: 14 (1%)

No.	Label	Count	Weight
1	Indonesia	14	14.0
2	Singapore	26	26.0
3	Qatar	40	40.0
4	United States of Am...	131	131.0
5	New Zealand	6	6.0
6	Australia	130	130.0
7	Portugal	1	1.0
8	Egypt	2	2.0
9	United Kingdom	55	55.0
10	England	15	15.0

Class: Deceased (Nom)

2. Data Cleaning



Filter

ReplaceMissingWithUserConstant -A 7 -N Unknown -R 0 -F "yyyy-MM-dd\'T\'HH:mm:ss"

Current relation

Relation: Book1-weka.filters.unsupervised.attribute.R...   Attributes: 12
Instances: 1562   Sum of weights: 1562

Attributes

No.		Name
1	☐	Date notified of potential case
2	☐	Sex
3	☐	Age group
4	☐	DHB
5	☐	Region
6	☐	Overseas travel
7	☑	Last country before return
8	☐	Flight number
9	☐	Flight departure date
10	☐	Arrival date
11	☐	Case Status
12	☐	Deceased

Selected attribute

Name: Last country before return   Type: Nominal
Missing: 0 (0%)   Distinct: 41   Unique: 14 (1%)

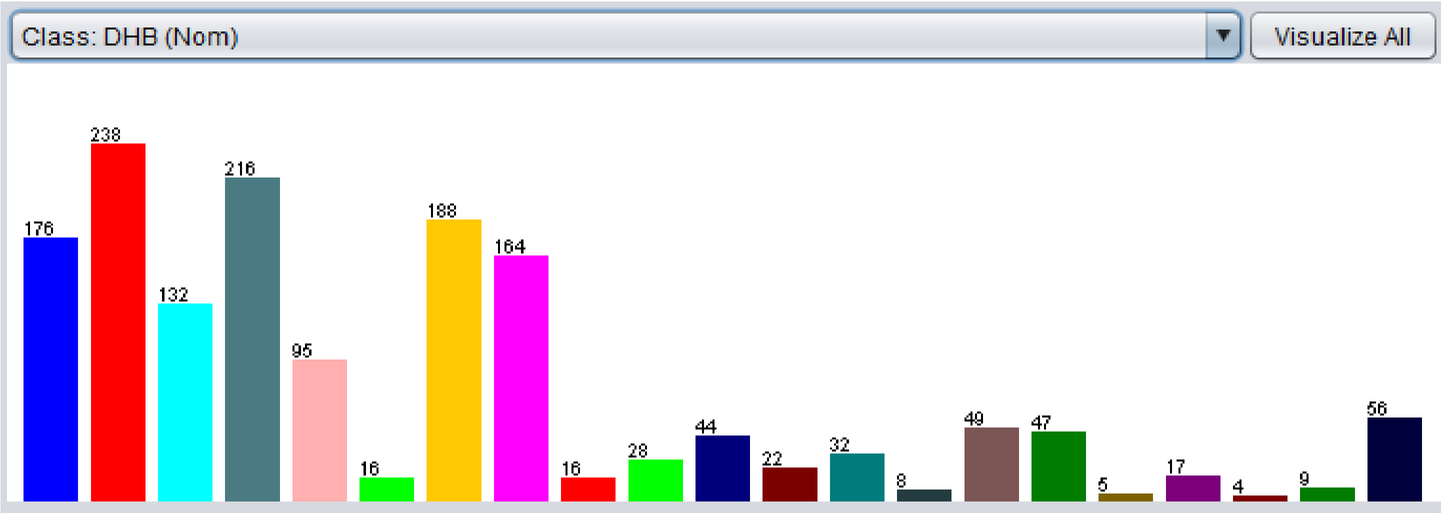
No.	Label	Count	Weight
1	Unknown	943	943.0
2	Indonesia	14	14.0
3	Singapore	26	26.0
4	Qatar	40	40.0
5	United States of Am...	131	131.0
6	New Zealand	6	6.0
7	Australia	130	130.0
8	Portugal	1	1.0
9	Egypt	2	2.0
10	United Kingdom	55	55.0

Class: Deceased (Nom)

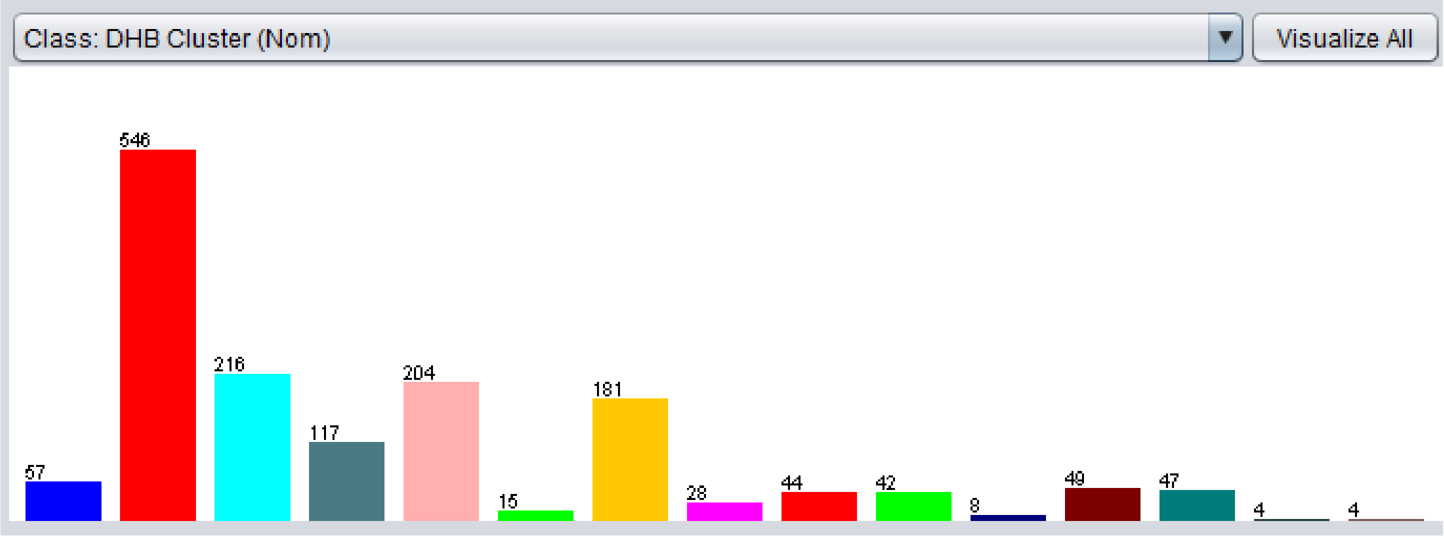
3. Data Transformation

- The final part of data pre-processing is data transformation.
- Normalization, Aggregation, Discretisation, Generalisation, Attribute construction, and smoothing are the different data transformation methods .
- The attributes in the dataset I collected is pretty much transformed.
- The age is already in categorical form and not in numerical form.
- The only attribute with most distinct values is DHB and the same can be clustered region wise to get a better understanding .

3. Data Transformation



Class: DHB (Nom)



Class: DHB Cluster (Nom)

DATA SUMMARIZATION

1. Introduction of Dataset

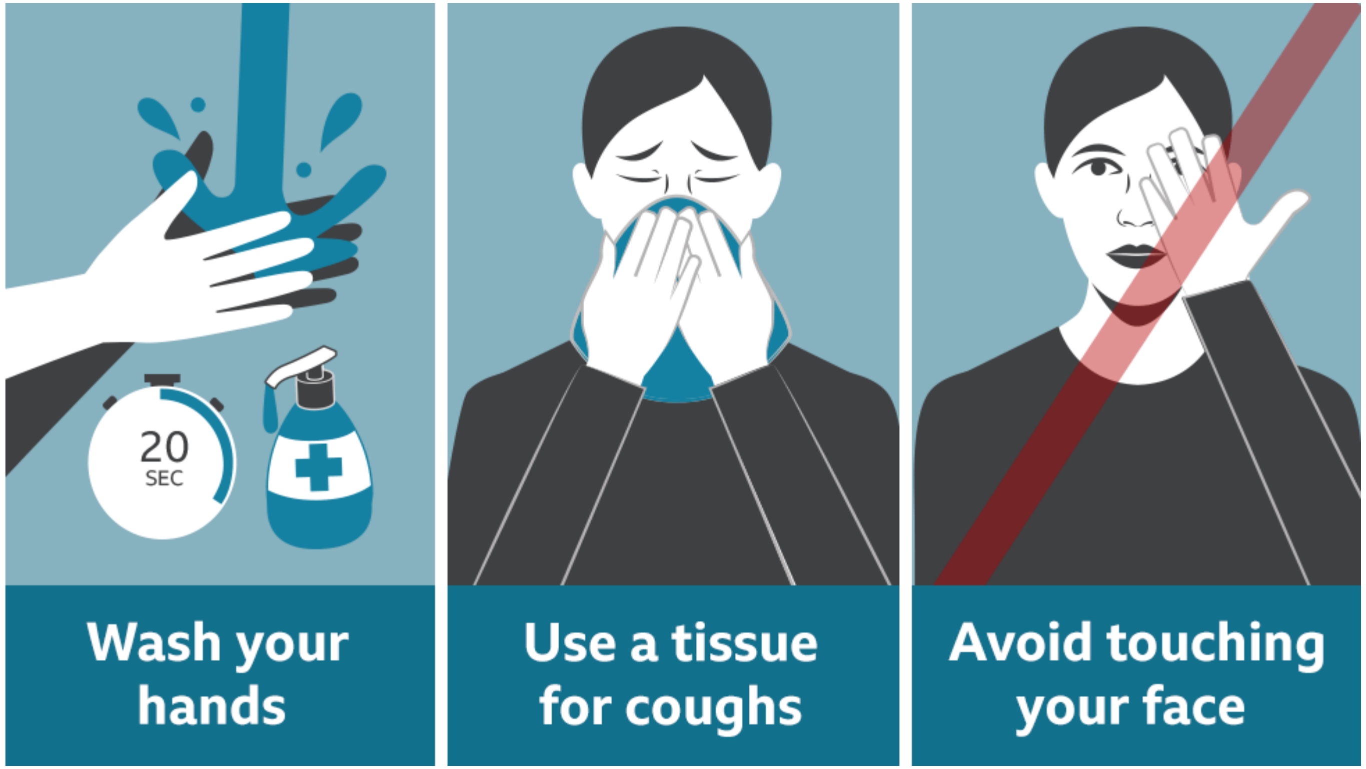
- COVID-19 is a family of viruses that got the name from their spiky crown and the name was designated by World Health Organization on 02/11/2020.
- On 11/03/2020, WHO declared it as pandemic and as a part of preventing and controlling the spread, many countries adopted lockdown.
- The dataset I chose is the collection of data elements related to COVID-19 of New Zealand from the date of virus infection being first reported i.e. 26/02/2020 to 01/08/2020 till 9am by the Ministry of Health

2. Format and statistical description of dataset

- Dataset gathered from health.govt.nz was in excel format. I converted it in to csv format to use it in weka.
- 1562 instances and 11 attributes
- All the attributes are nominal values.
- Categories like age, sex, DHB , case status, deceased or not etc
- Data arranged from the first date on which COVID-19 was reported in New Zealand till 1st August 2020.

3. Suggestion for future data collection

- The data studied was confined to New Zealand alone, may be a larger dataset either globally or continent wise.
- The dataset was already tidy and hence could not apply much of the data integration and data pre-processing. So get a more dirtier dataset.
- May be as a part of future data collection, a huge data set with more attributes that are in demand integrated from different sources can lead to a better data mining and a better research conclusion.



Questions????

THANK YOU