Task A

1. Data description

6 data files are loaded into the dataframe and used to conduct data analysis. Data are collected from 1st January, 2020 to 31st March, 2022. Daily change and cumulative counts at both state level and national level, deaths and population information is collected from these datasets. Specific variables include confimed cases, deaths, tests, positive tests, recovered cases, hospitalised cases, cases in ICU, cases on a ventilator and population. The other data files are not used since the data in those files are covered in these 6 data files.

2. Statistics (Q1)

(1) National level

By 31th March, a total of 4322777 positive Covid-19 cases are confirmed in Australia. The recover rate (recovered cases/confirmed cases) is 7.2% and death rate (deaths/confirmed cases) is 0.14%. The number of cases in ICU is 90 and the number of cases on a ventilator is 27 in total. 56359439 people in Australia get vaccinated.

On average, 5424 people is confirmed as positive, 8 people die after being confirmed as positive and 391 people recover every day.

(2) State level

New South Wales and Victoria appear to experience most serious outbreaks of Covid-19, having much more confirmed and deaths cases than the other states. New South Wales and Victoria appear to experience most serious outbreaks of Covid-19, having much more confirmed and deaths cases than the other states. The recover rate of Victoria and Australian Capital Territory is significantly higher than the other states. Victoria reports the highest death rate in the country.

(3) Deaths

The average age of death cases is 76. The number of deaths increase with age. People at 80-89 age group seem to be most vulnerable to Covid-19 since most death cases appear in this age group.

(4) Population

The total population of Australia is 25459470. Among all states, New South Wales have the largest population and Victoria ranks the second. Northern Territory has the smallest population.

3. Visulisation and insights

3.1 Box plot

A boxplot is a standardised method of visualising data distribution using a five summary. Several boxplots are created to display age distribution in different categories.

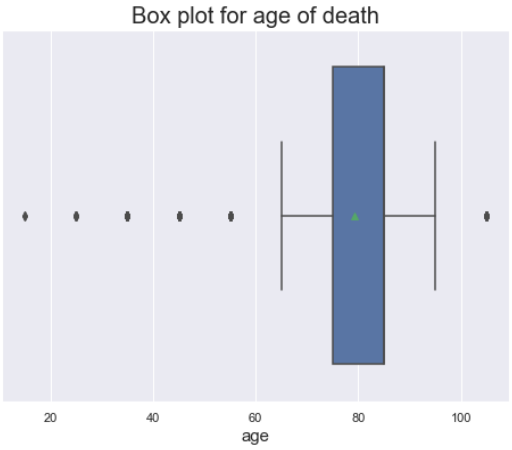
(1) Data cleaning for ‘age’ and ‘gender’

It is observed that there are a huge amount of missing values in deaths dataset (Galarnyk, 2022). There are 5916 missing values for ‘age’ variable, 3878 missing values for ‘age_bracket’ variable, 3923 missing values for ‘age_bracket’ variable and 3738 missing values for ‘dd’ variable. ‘dd’ is the first age number of ‘age_bracket’. Despite huge missing values, age and gender are worth investigating since they are important demographic features. People within the same age group or gender group may have similar patterns after testing positive to Covid-19.

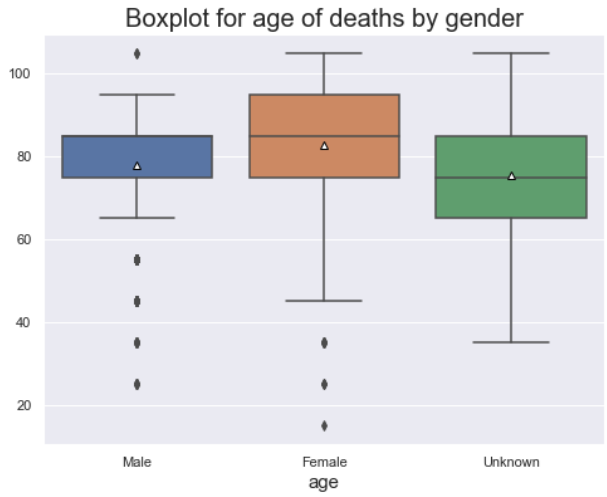
Since ‘dd’ has much more non-null observations than ‘age’, this variable is used to approximate the true age. To make the value of this variable closer to the true age, I add 5 to each value in ‘dd’ column, which is the middle number of corresponding age bracket. Observations with abnormal values in ‘dd’ such as ‘Not reporte’, ‘Unknow’ and ‘0-’ are excluded. After data cleaning, ‘dd’ is renamed as ‘age’ and replaces the original ‘age’ column.

In ‘gender’ variable, there are abnormal gender types in the gender variable such as Male* and Female* and several data of gender are unknown. Since the cause of these abnormal gender types is not certain, they are classified as ‘Unknown’ to avoid misleading.

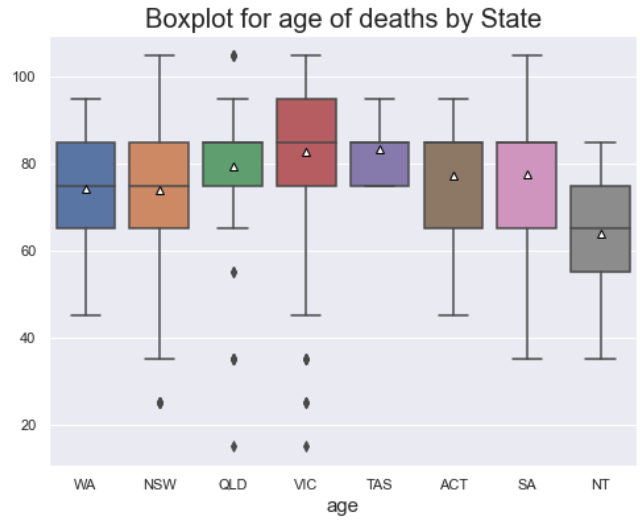
(2) Box plot for age of deaths
Box plot can show the distribution of ages of death cases. The box plot is comparatively short. This indicates that overall deaths happen among the elder group. The middle box represents the middle 50% of ages for death. Outliers can be detected from this plot. I decide not to remove these outliers since it is valid that people at younger age or older age can die after getting Covid-19.



(3) Box plot for age of deaths by gender
The box plot below is generated to analyse the age distribution of deaths in different gender groups. The box plot shows that female group has a wider distribution of age than male group and it’s average age is higher than that of male group. The age of ‘Unknown’ group spreads from around 40 to 120.



(4) Box plot for age of deaths by state
The age distribution of deaths is also analysed by state. New South Wales and South Australia have the widest age distribution among the states. Tasmania has the shortest age distribution, followed by Queensland. In terms of average age of deaths, Victoria and Tasmania have the highest mean value and Northern Territory has the lowest mean value among the states.

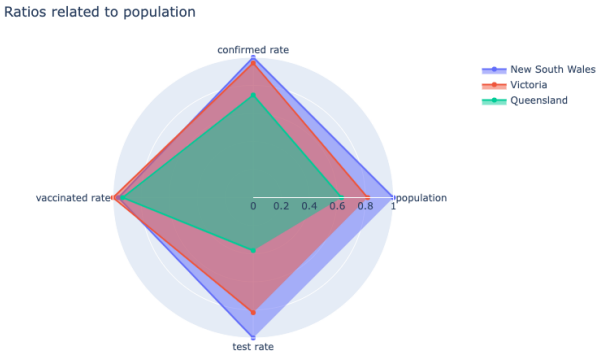


3.2 Radar plot

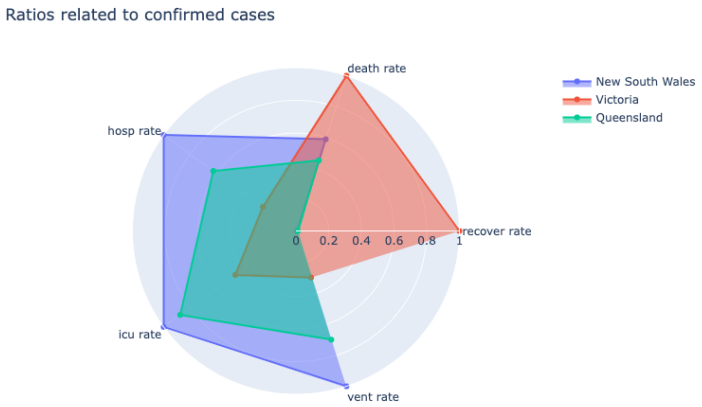
(1) Calculate ratios
Two sets of ratios are calculated to make comparison among states. The first set includes 'population', 'confirmed rate', 'vaccines rate' and 'test rate' which are obtained by dividing the number of corresponding cases by population. Since each state has huge difference in population which will affect the number of cases, it is more reasonable to compare ratios. The second set includes 'recover rate', 'death rate', 'hosp rate', 'icu rate' and 'vent rate'. These ratios are obtained by dividing the number of corresponding cases by the number of confirmed cases.

(2) Compare three states using the ratios related to population
New South Wales, Victoria and Queensland are selected to make comparison since these are most populated states which could be affected by Covid-19 significantly. Given that different ratios may have a large difference in scale, the result cannot be presented well when putting them in the same radar plot. Thus, the scale is adjusted by dividing each value in a certain ratio category by the maximum number in that ratio category. As a result, each value will be ranged between 0 and 1.

The following radar chart compares three states using the ratios related to population. From the chart, we can find that these three states have similar vaccinated rate. The vaccinated rate of Victoria is slightly higher than that of New South Wales. In addition, New South Wales report the highest confirmed rate and test rate compared to the other two states.

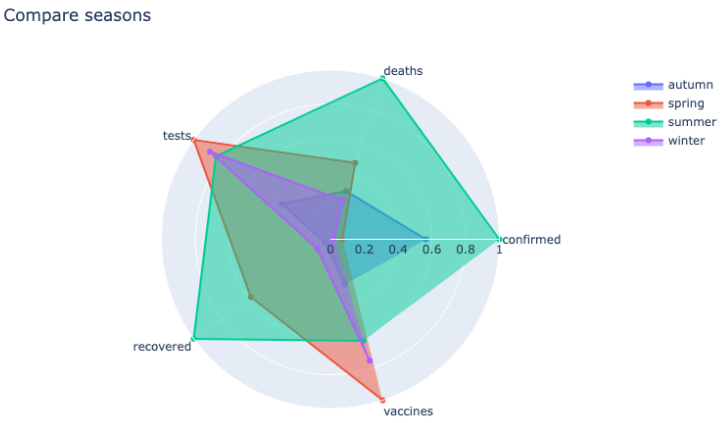


(2) Compare three states using the ratios related to confirmed cases
The radar chart below compares three states using the ratios related to confirmed cases. Victoria appears to have the highest recover rate and death rate, while the recover rate of New South Wales and Queensland is much smaller than that of Victoria. New South Wales have the highest hospitalised rate, icu rate and vent rate. With high ratio in confirmed rate, test rate, hospitalised rate, icu rate and vent rate, New South Wales government should make sure the healthcare system capacity is sufficient, since pandemic has a severe impact on healthcare system (Cleggett, 2022).



(3) Compare seasons
Variables including 'confirmed', 'deaths', 'tests', ‘recovered’ and ‘vaccines' are compared among 4 seasons to identify the effect of seasons on the Covid-19 situation. It is observed that the nubmer of recovered cases, deaths cases and confirmed cases are greatest in summer. All these ratios are smallest in autumn. The number of tests and vaccines is largest in spring.

It is reasonable that the confirmed rate is higher in summer since people are more likely to enjoy outdoor activities, which will increase the probability of infection. In this case, marketing campaigns are essential to be conducted especially in summer suggesting people to wear masks.



4. Conclusion on data quality and collection

Among the 6 data files, only ‘deaths’ dataset has missing values. There are a lot of missing values in age and gender related variables. It could be caused by the respondents’ unwillingness to fill in the personal information of their relatives who died due to Covid-19. Therefore, Australian government may take measure to encourage people to provide this kind of information. For example, when they are filling in the questions related to age and gender, the importance of gender and age information for analysis and that personal information will not be disclosed can be notified under the questions. For the ‘deaths’ dataset, information related to medical history or disease should also be collected to help confirm the cause of death.

There are also some variables containing negative values such as ‘positives’ and ‘recovered’. This issue could be caused by data collection method. Since negative number occurring in these values are not logical, the method to record values of these variables should be improved to avoid misleading.

Task B

Summary and Recommendations

As the report indicates that cases increase with age, it is important to prioritise vaccination efforts for older age groups and ensure that they have access to healthcare facilities and medical resources. Another suggestion for aging groups is to encourage the use of virtual healthcare options to reduce the need for in-person visits to healthcare facilities. This can help reduce the risk of exposure to the virus.

The box plot highlights the need for considering gender factor when addressing the COVID-19 pandemic. Given that the female group has a wider distribution of age of death than the male group, females may be at higher risk of severe illness and death from COVID-19 compared to males. Therefore, it is important to consider gender-specific factors that contribute to the difference in distribution, such as differences in healthcare-seeking behavior, lifestyle or economic factors to develop targeted and effective health policies and interventions.

New South Wales reports the highest confirmed rate, test rate, hospitalised rate, icu rate and vent rate compared to Victoria and Queensland, so it is crucial to increase testing capacity and strengthen hospital capacity. Victoria needs to maintain its high recovery rate while also working on reducing its death rate. This could be achieved through providing timely and effective medical care.

Seasonal factors should also be considered when making policy decisions related to controlling and mitigating the spread of COVID-19. Based on the observation that the number of recovered cases, deaths cases and confirmed cases are greatest in summer, the government should continue implementing measures to control the spread of COVID-19 during this season. While the warmer weather may provide some relief, it is still important to encourage people to follow public health guidelines such as wearing masks, keeping social distancing and avoiding large gatherings. Since the number of tests and vaccines is largest in spring, it may be beneficial for the government to prioritise the distribution of vaccines during this season.