COMP5048 Assignment 2: Individual Report

***Author**: Xing Xing, **Student ID:** 500390560, **UniKey:** XXIN7882, **Date:** 4 November 2021*

I. Introduction

Since 31 December 2019, when the first novel Covid-19 infection was reported in China, how to effectively stop the spread of the virus has become the most urgent concern of Australia and NSW governments for the past 2 years. Till now, NSW government successively applied 4 strategies to stop the spreading of coronavirus in NSW. More specifically these strategies ordered by implementation date are:

- Enforcing good hygienic procedure,
- Encouraging Covid-19 testing,
- Related travel and gathering restrictions,
- Encouraging vaccination,

For this assignment, the main objective is to find out which method or combination of methods have the highest effectiveness by applying appropriate data analysis and visualisations. Base on related data analysis and visualisation in. the group report, the combination of encouraging vaccination, encouraging testing, and enforcing good hygienic procedure is the most efficient strategy. The content below will summarise and demonstrate author's personal contributions to the group project base on non-visualisation contribution and visualisation contribution aspects. For non-visualisation aspect, a responsibility of project management, document management, and data pre-processing were token. For visualisation aspect, good hygienic line chart and interactive visualisation were created with related analysis by author.

II. Non-visualisation contributions

A. *Project management*

The author's effort on project management could be summarized as ensuring the final deliverable will be completed on time with expected quality. 2 project management methods were applied to attain a successful group report which are creating project schedule and continuous quality monitoring. Project scheduling meant to assure the project could be finished on time. Continuous quality monitoring is utilised to guarantee the high quality of final deliverable.

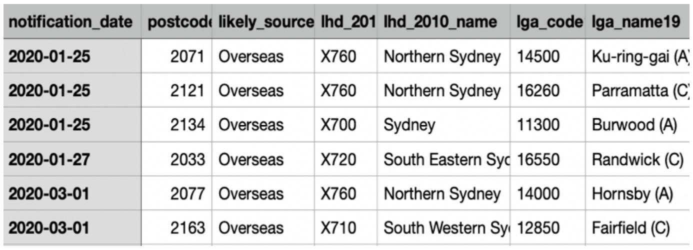
On 12 September 2021, by discussing with the members in group RETUT05-02, a clear project schedule was created and splitted the final deliverable into smaller tasks and assigned to each member. More specifically, the final deliverable was defined as a high-quality report consist of 5 sections include related data visualisation namely introduction and conclusion, enforcing good hygienic process analysis, encouraging testing analysis, travel restriction analysis, encouraging vaccination analysis, and strategies comparation. Since the final report was expected to be completed within 4 weeks, three progress check points are designed as follow: the first 2 weeks were scheduled to complete related preparations such as data collecting, literature survey, and data pre-processing. The third week was scheduled as finishing the individual strategy analysis and visualisation. The last week before deadline was expected to finish the introduction/conclusion and overall strategies comparison. Based on the three check points as described above, continuous quality monitoring by each check points were applied to ensure the quality in stage and ensuring the overall quality. By quality monitoring, the potential problems of literal analysis and visualisation diagrams were detected. With efficient communication with original contributor, the issue could be fixed before next check point. With this method, the quality of final deliverable meets the overall expectation.

B. *Document management*

The author's effort on document management could be concluded as integrating group members' individual work into the final deliverable with decent format. To achieving this, an open google doc was used to record each member's individual work and data visualisation. At each check point, all members work will be formatted as 10pt Times New Roman and export to Word document from google doc. Furthermore, at the last week, final version of group report was also ensured to satisfy the format specification according to assignment instruction.

C. *Data processing*

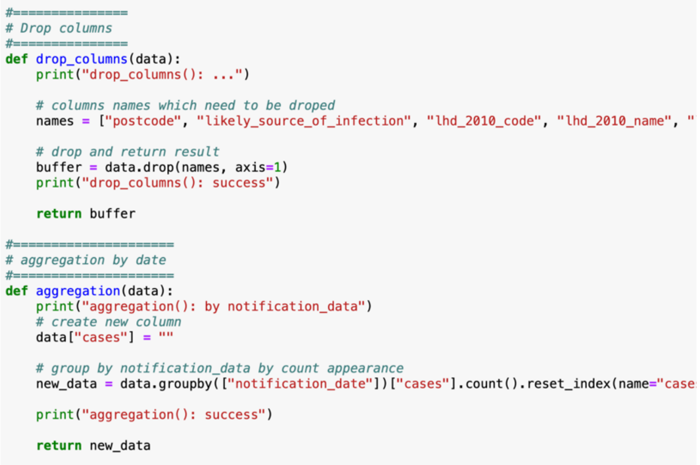
Data processing and cleaning means applying related processing technics on raw datasets and produce cleaned dataset which is more supportive for further data visualisation and analysis. Data pre-processing and cleaning is an important preparation step for exploratory data analysis and visualisation. For this assignment, related analysis and data visualisations are built based on processed dataset from Data NSW [2] which was listed in the assignment description. In this dataset, there are 69016 rows and 7 columns which indicated that there are 7 variables. For all variables (columns) in this dataset are 1 ordinal date data and 6 nominal string data. Notification date could be considered as ordinal data since date has clear order and could be sorted. The rest columns are all treated as nominal categorical data since it only provides classification information. The origin data was stored in csv format, the figure below shows the first 6 rows of the original data:



notification_date	postcod	likely_source	lhd_201	lhd_2010_name	lga_code	lga_name19
2020-01-25	2071	Overseas	X760	Northern Sydney	14500	Ku-ring-gai (A
2020-01-25	2121	Overseas	X760	Northern Sydney	16260	Parramatta (C
2020-01-25	2134	Overseas	X700	Sydney	11300	Burwood (A)
2020-01-27	2033	Overseas	X720	South Eastern Syd	16550	Randwick (C)
2020-03-01	2077	Overseas	X760	Northern Sydney	14000	Hornsby (A)
2020-03-01	2163	Overseas	X710	South Western Sy	12850	Fairfield (C)

Fig [1]: First 6 rows of covid data in NSW from Data NSW

By observing the first 6 rows of original data in Fig [1], it is obvious to find out that all covid infection cases are recorded individually as each row in the dataset. Interestingly, there are no id column to unique each record. In this dataset, an appropriate statistical method could be applied to summarize the number of people infected for each day. More specifically, an aggregation on notification date which count the number of rows grouped by exact date could bring up the expected result. Related python code is shown below:



```
#===============
# Drop columns
#===============
def drop_columns(data):
    print("drop_columns(): ...")

    # columns names which need to be droped
    names = ["postcode", "likely_source_of_infection", "lhd_2010_code", "lhd_2010_name", "

    # drop and return result
    buffer = data.drop(names, axis=1)
    print("drop_columns(): success")

    return buffer

#=====================
# aggregation by date
#=====================
def aggregation(data):
    print("aggregation(): by notification_data")
    # create new column
    data["cases"] = ""

    # group by notification_data by count appearance
    new_data = data.groupby(["notification_date"])["cases"].count().reset_index(name="case

    print("aggregation(): success")

    return new_data
```

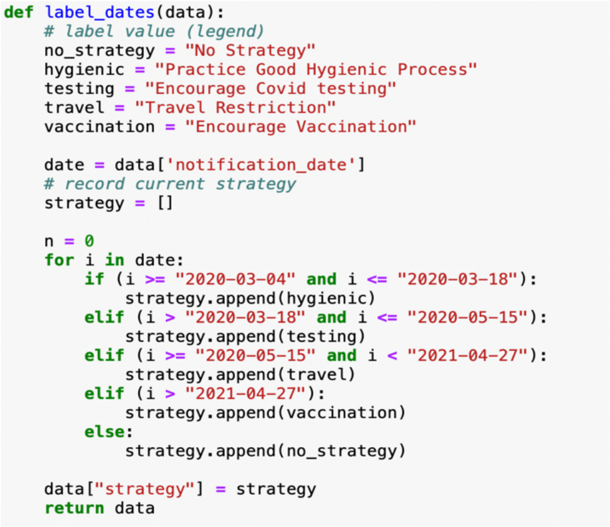
Fig [2]: Drop columns and aggregation

In figure 2 there are two functions are listed which are *drop_column()* and *aggregation()*. *Drop_columns()* only kept notification_date column for further aggregation. *Aggregation()* aggregated number of infections by each day and created a new column namely cases as a new quantitative integer ratio variable since a negative cases number is meaningless and inappropriate.

Another data processing related tasks is adding labels for each row to represent the latest Covid-19 controlling strategy was activated by NSW government for each day. More specifically, create another column to indicate the activation period for all 4 strategies. According to Lupton's documentary [3], activation time of these strategies could be concluded as follows:

- Urge good hygienic process: from 04/03/2020.
- Encourage covid testing: from 18/03/2020.
- Gathering restriction: from 15/05/2020.
- Encourage vaccination: from 27/04/2021 till now.

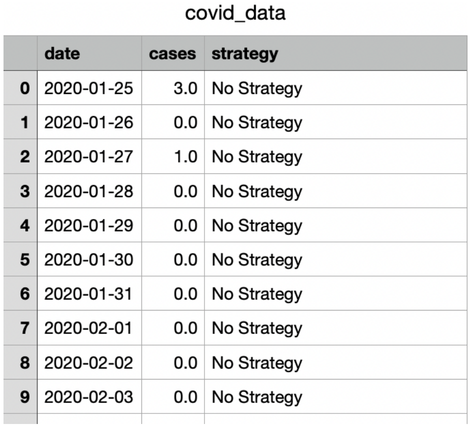
Base on the period which shown above, a new categorical nominal string column which named strategy were added to the dataset by executing the following python code:



```
def label_dates(data):
    # label value (legend)
    no_strategy = "No Strategy"
    hygienic = "Practice Good Hygienic Process"
    testing = "Encourage Covid testing"
    travel = "Travel Restriction"
    vaccination = "Encourage Vaccination"

    date = data['notification_date']
    # record current strategy
    strategy = []

    n = 0
    for i in date:
        if (i >= "2020-03-04" and i <= "2020-03-18"):
            strategy.append(hygienic)
        elif (i > "2020-03-18" and i <= "2020-05-15"):
            strategy.append(testing)
        elif (i >= "2020-05-15" and i < "2021-04-27"):
            strategy.append(travel)
        elif (i > "2021-04-27"):
            strategy.append(vaccination)
        else:
            strategy.append(no_strategy)

    data["strategy"] = strategy
    return data
```

Fig [3]: Add periodical labels to the dataset.

After the processing above, a new dataset which contain 3 columns was created and saved locally for further analysis and data visualisation. Also, this dataset was shared to all members in RETUT05-02. Since this categorical label column was created, adding colour to the visualisation would be more convenient and efficient. The first 9 rows of processed clean dataset with three columns are shown below (the first unnamed column is an automatic id column when using python3 pandas library to unique each row in dataset):



covid_data

	date	cases	strategy
0	2020-01-25	3.0	No Strategy
1	2020-01-26	0.0	No Strategy
2	2020-01-27	1.0	No Strategy
3	2020-01-28	0.0	No Strategy
4	2020-01-29	0.0	No Strategy
5	2020-01-30	0.0	No Strategy
6	2020-01-31	0.0	No Strategy
7	2020-02-01	0.0	No Strategy
8	2020-02-02	0.0	No Strategy
9	2020-02-03	0.0	No Strategy

Fig [4]: First 9 rows of processed clean dataset

In the processed clean dataset, there are 3 variables namely: date, cases, strategy. Date is an ordinal data represent each day from 2020-01-25 to 2021-10-10 which could be ordered and compared with other variable with the same data type. Cases is a ratio variable represent the new Covid-19 infection cases in NSW for each day which could not have negative value in this field. Strategy is a nominal categorical variable which represent the latest activated strategy for each day which will be used for colouring for visualisation diagrams.

III. VISUALISATION CONTRIBUTIONS

In the visualisation related tasks of this assignment, the author was responsible for implementing enforcing good hygienic process visualisation and interactive visualisation. Following content will summarize author's personal contribution, design process, and implementation process for related diagrams.

A. Visual Application selection

The application which was used to realise related visualisation is Tableau Desktop version-2021.2. Tableau is a data analysis and visualisation tool with mature and stable user interface which is user friendly and be able to realise various visualisation methods. Also, Tableau provides many interfaces to faultlessly connect with various type of data source such as: MySQL, json, MongoDB, CSV, etc. In this assignment, all data are stored in csv file which generated by python pandas in data processing step which described earlier in this report. Tableau provides perfect real-time connection with CSV file. Hence, since Tableau have such advantages and suitable with the related visualisation tasks, it was selected as the visualisation application for creating related data visualisation diagrams.

B. Good hygienic process visualisation

1) Determine visualisation method

The purpose of good hygienic process visualisation diagram is bringing the readers an obvious comparison of new Covid-19 infection cases for each day before and after good hygiene policy was enforced by NSW government. According to the content from USYD COMP5048 week 5 lecture and lecture slides [6], a line chart is appropriate to be used to ideally stand out the trend of data by connecting points with continuous line. Also, colours with high chromatic aberration could make the comparation of different part of data more obviously. By personal preference, blue and purple will be used to represent the two parts of data divided by the good hygienic policy releasee date. Hence, a coloured line chart will be used for this visualisation.

2) Determine axes arrangement and visual variables

Since the purpose of this diagram is demonstrate the number of new cases trend difference before and after good hygiene policy applied in NSW, date, cases, and label which are the variables in new clean dataset were selected as visual variables in this diagram. Date is an ordinal variable and will be arranged on horizontal axis as the time flow. Cases is a ratio variable which demonstrate the number of cases for corresponding date and will be arranged as vertical axis to reflect the changing by time. Label which indicated the policy activation status is a nominal categorical variable will be used for colouring. According to the lecture slides about axis [7] and colours [8], a smaller range of axis which could be enough to explain what was happening for this policy and concentrate the readers' attention. For x-axis (date), according to the NSW government notification [4], the first Covid-19 case in NSW was reported on 21 January 2020, the start of axis will be set at 21 January 2020. Since good hygienic process policy was released on 4 March 2020 [5] and individually operated until April 2020. The end of x-axis will be set as 31 March 2020. For y-axis, which is cases, since the peak of the cases in this x-axis is 213 cases/day and the lowest point is 0 cases at the beginning period of x-axis, the range of y-axis will be selected as 0 to 220 cases/day. As discussed earlier, colour will represent the label variable. Purple will use to represent the data when hygiene policy was not activated. Blue will use to represent the data when hygiene policy was activated from 4 March 2020. The detailed related visualisation is shown below [9]:

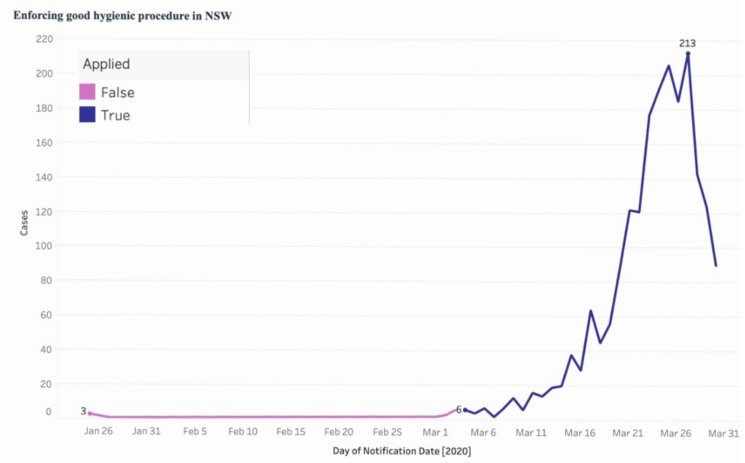


Fig [5]: Enforcing good hygienic procedure in NSW [9]

In this diagram, the legend is added inside the body since the left top area is relatively blank. And a trend of cases changing before and after the policy released could be relatively observed and understood.

C. Interactive visualisation

In this assignment, author implemented an interactive data visualisation in Tableau dashboard to let audiences interactively browse and understand the data. The implementation application which was selected to be used is Tableau since it provides a dashboard platform which have decent in-built interactive functionalities. So, Tableau dashboard will be utilised for building the interactive visualisations.

1) Determine visualisation method

The purpose of this interactive diagram is providing dynamic filters and moveable x-axis for readers to initiatively explore the specific information that interested them. For example, if user would like to only see the cases changing during vaccination strategy applied, by changing the filter, the data will be refined down to only demonstrate the vaccination period data and hide the other parts. According to the content from USYD COMP5048 week 5 lecture and lecture slides [6], since this diagram shall contain nearly all data for all strategies and regulations, coloured line chart could effectively separate the data under different strategies. Hence, an interactive-coloured line chart will be used as the framework to implement this diagram.

and colours [8],

2) Determine axes arrangement and visual variables

Since the purpose of this diagram is to demonstrate the infection cases changing during any strategy activation period of data. Similar with the diagram which discussed earlier, the x-axis is appropriate to be arranged as date which is an ordinal times series data. For y-axis, the cases number per day should be arranged which is a ratio numeric data. For several strategy labels, should be using different colours to separate data into different period based on strategy activation time interval. According to the lecture slides about axis and visual variables [7], x-axis will be set a range from January 2020 to October 2021 which included all date variables since all data should be included. For y-axis, since the lowest value and highest value are 0 and 1527 respectively, the y-axis range should be set from 0 to 1600 cases per day. For strategies which stored 'strategy' variable, various colour should be added to the line to demonstrate the different strategy activation period. More specifically, red will represent when there is no Covid prevention strategy applied in NSW. Cyan will represent the period after good hygiene policy was applied. Blue will represent the period when encouraging test is applied. Green will represent the period when travel and gathering restriction was activated. Orange will represent the vaccination encouraging policy after it is released. There should be two filtering function which are filtering data by ticking in the box which contain all possible strategy period data and customise the date interval by sliding the time bar. More specific details of this diagram is shown below [10]:

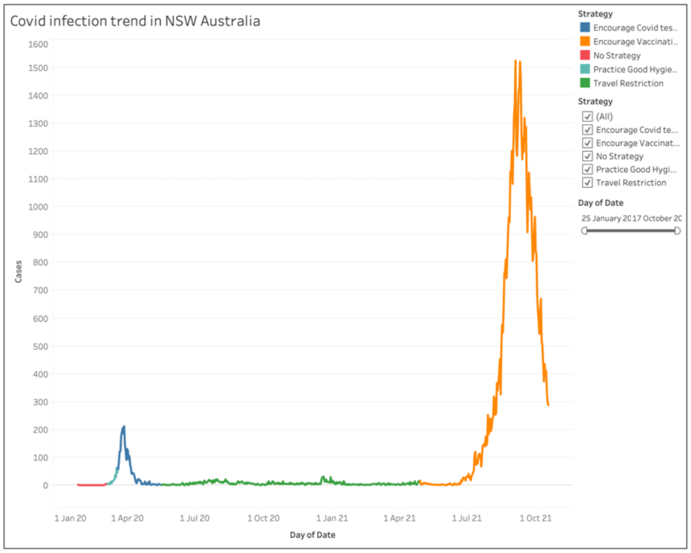


Fig [6]: interactive visualisation [10]

The diagram from above is an overview of this interactive visualisation which created by author. The control box which is the interactive interface between users and diagram could be found at the right top of the diagram. More filtered diagram example will be shown below:

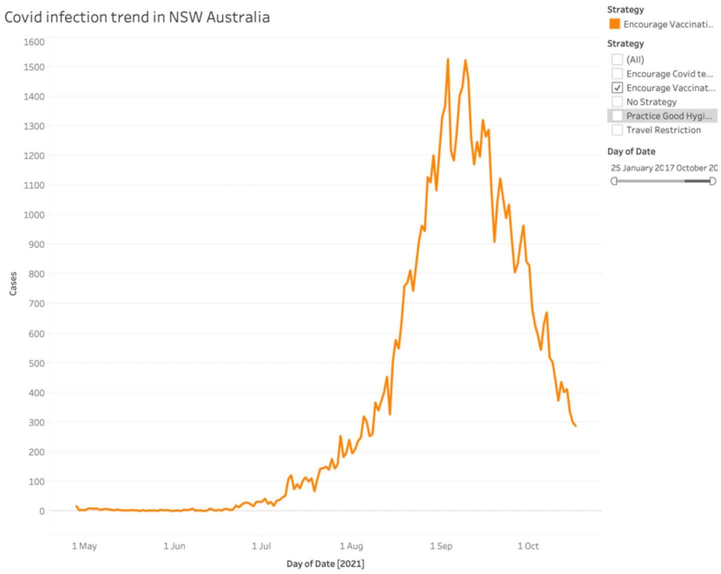


Fig [7]: interactive visualisation after strategy filter applied [10]

Fig 7 from above is the interactive diagram after a vaccination filer is applied. Comparing with Fig 6, a zoom effect could be observed since the diagram in Fig 7 only contain part of the data.

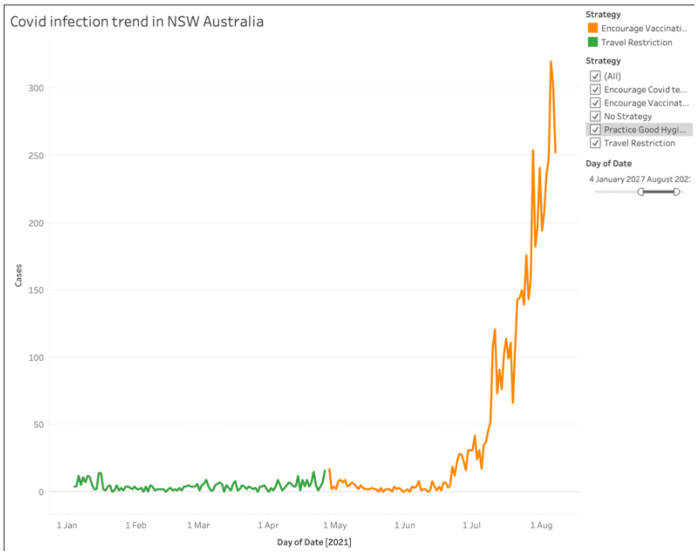


Fig [8]: interactive visualisation after date filter applied [10]

Fig 8 from above is the interactive diagram when a specific period as selected by the top right sliding bar. By observing the diagram, a period between 4 January 2021 and 7 August 2021 was selected to demonstrate to the reader.

IV. Conclusion

Overall, all members in group RETUT05-02 provided impressive effort and understanding of visual analytics. The author of this report contributed to the group by applying project management, document management, and data processing as non-visualisation contribution. Created data visualisations related to good hygiene analysis in NSW and an interactive overview visualisation as visualisation contribution.

REFERENCES

[1] "COVID Live Update: 242,019,783 Cases and 4,923,691 Deaths from the Coronavirus - Worldometer", *Worldometers.info*, 2020. [Online]. Available: https://www.worldometers.info/coronavirus/. [Accessed: 19- Oct- 2021].

[2] “NSW COVID-19 cases by likely source of infection”, NSW Government, 2020. [Online]. Available: https://data.nsw.gov.au/nsw-covid-19-data. [Accessed: 19- Oct- 2021].

[3] “Timeline of COVID-19 in Australia: the first year”, D.Lupton, 2020. [Online]. Available: https://deborahalupton.medium.com/timeline-of-covid-19-in-australia-1f7df6ca5f23. [Accessed: 19- Oct- 2021].

[4] “Coronavirus cases confirmed in NSW”, NSW Government, 2020. [online]. Available: https://www.health.nsw.gov.au/news/Pages/20200125_03.aspx. [Accessed: 5-Nov-2021]

[5] "Community urged to help prevent coronavirus", NSW Government, 2020. [Online]. Available: https://www.nsw.gov.au/news/community-urged-to-help-prevent-coronavirus. [Accessed: 5-Nov-2021].

[6] “Explotary analysis II lecture slide”, USYD COMP5048 visual analytics.

[7] “Semiology of graohs lecture slide”, USYD COMP5048 visual analytics.

[8] “Color lecture slide”, USYD COMP5048 visual analytics.

[9] “COMP5048 Assignment 2-Group report”, X.Xing, L.Sun, J.Zhang, Z.Xu. 2021.

[10] “COMP5048 Assignment 2 interactive visualisation”, X.Xing, L.Sun, 2021