Report on assignment2 (individual)

[t] Yichen Chen

yche3494

510138903

1. Your contribution to the group work. Describe non-visualization tasks you carried out towards the group achieving its task

For Data processing part, firstly I get "confirmed_cases_table1_location.csv" from data.nsw.gov.au. Then I use python3 in jupyter notebook to do some Data processing. I apply .isna () and .sum() function to count how many rows we have is none. Then I got this result:

```
data_confirmed = pd.read_csv("confirmed_cases_table1_location.csv")
num_na = data_confirmed.isna().sum()
print(num_na)
```

```
notification_date      0
postcode               0
lhd_2010_code        964
lhd_2010_name        964
lga_code19           964
lga_name19           964
dtype: int64
```

The output showed that there are 964 rows are not complete. Then I use .dropna() to delete these rows. After that I use .loc to extract "notification_date" column, which will be a useful column for our work.

```
variables = ['notification_date']
data_confirmed = data_confirmed.loc[:, variables]
data_confirmed.head()
```

	notification_date
0	2020-01-25
1	2020-01-25
2	2020-01-25
3	2020-01-27
4	2020-03-01

Then I get "pcr_testing_table1_location_agg.csv" from data.nsw.gov.au. Then use .head() function to take a glimpse:

	test_date	postcode	lhd_2010_code	lhd_2010_name	lga_code19	lga_name19	test_count
0	2020-01-01	2038	X700	Sydney	14170	Inner West (A)	1
1	2020-01-01	2039	X700	Sydney	14170	Inner West (A)	1
2	2020-01-01	2040	X700	Sydney	14170	Inner West (A)	2
3	2020-01-01	2041	X700	Sydney	14170	Inner West (A)	1
4	2020-01-01	2069	X760	Northern Sydney	14500	Ku-ring-gai (A)	1

Then I extract "test_date" and "test_count" columns and take a glimpse:

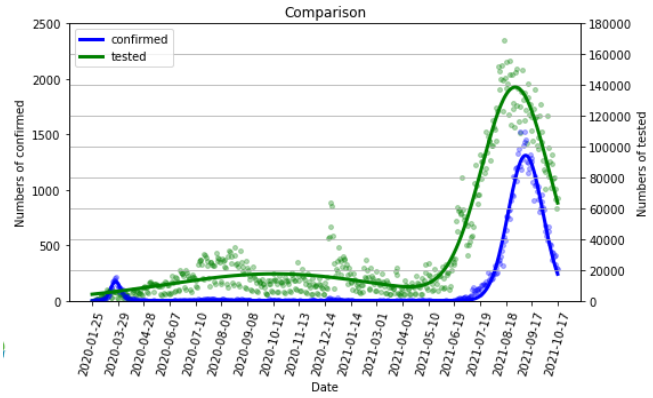
	test_date	test_count
0	2020-01-01	1
1	2020-01-01	1
2	2020-01-01	2
3	2020-01-01	1
4	2020-01-01	1
5	2020-01-01	1
6	2020-01-01	1
7	2020-01-01	1
8	2020-01-01	1
9	2020-01-01	2

We can see that it is very messy. We need to group the sum of "test_count" by "test_date":

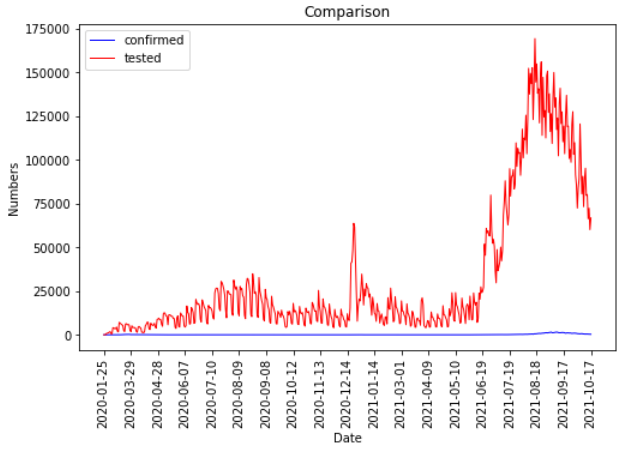
```
data_tested = data_tested.groupby(['test_date'],as_index=False).agg({'test_count': 'sum'})
data_tested.head(10)
```

	test_date	test_count
0	2020-01-01	23
1	2020-01-02	1
2	2020-01-03	1
3	2020-01-04	3
4	2020-01-08	1
5	2020-01-10	1
6	2020-01-14	1
7	2020-01-15	1
8	2020-01-20	1
9	2020-01-22	4

For programming part, I think a lot of code examples in this article can already illustrate. For deriving visualization application part, I import matplotlib.pyplot module to derive visualization, and you will see particular visualization I did in next part. For Literature survey, [2] I check the literature and tutorials about multi-Gaussian curve fitting on this website, then I use double Gaussian fitting to fit scatter plot to make the picture easier to read:




```
plt.figure(figsize=(8,5))
l1, = plt.plot(index,y_confirmed.date_count,color='blue',linewidth=1,linestyle='solid')
l2, = plt.plot(index,y_tested.test_count,color='red',linewidth=1,linestyle='solid')
plt.title("Comparison")
plt.xlabel("Date")
plt.ylabel("Numbers")
plt.xticks(index[0:len(index):30],
           rotation = 90)
plt.legend(handles=[l1,l2],labels=['confirmed','tested'],loc="upper left")
plt.savefig("3.jpg",bbox_inches = 'tight')
plt.show()
```



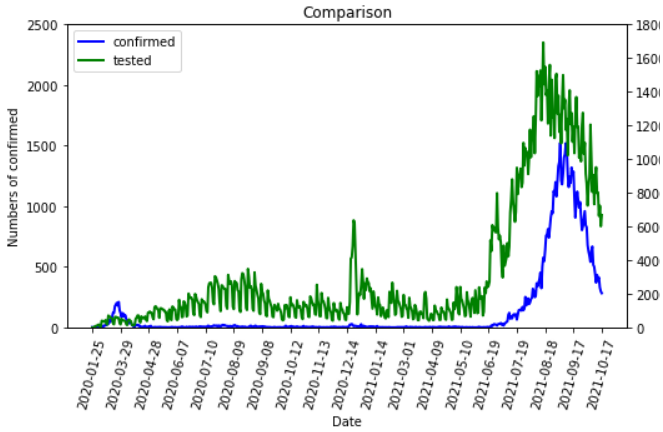
Another thing I did about Literature survey is to find the definition of KPI and KGI to quantify the effect of our measures. With the help of this website [3]http://www.woshipm.com/data-analysis/590642.html and [4] COMP5048 week8 lecture's content about KGI and KPI, I define KGI is to completely control the COVID-19 and KPI is to make the confirmed numbers of current month lower than previous month. And I also read the document of matplotlib[5] to help to establish impressive charts. For Project Management part, as a team member, I participated in every group meeting to discuss the division of tasks and their respective responsibilities.

We can see that it's not ideal, for the numbers of confirmed is relatively too small compared with test numbers. So I try to redesign it by adding another y-axis for numbers of confirmed cases:

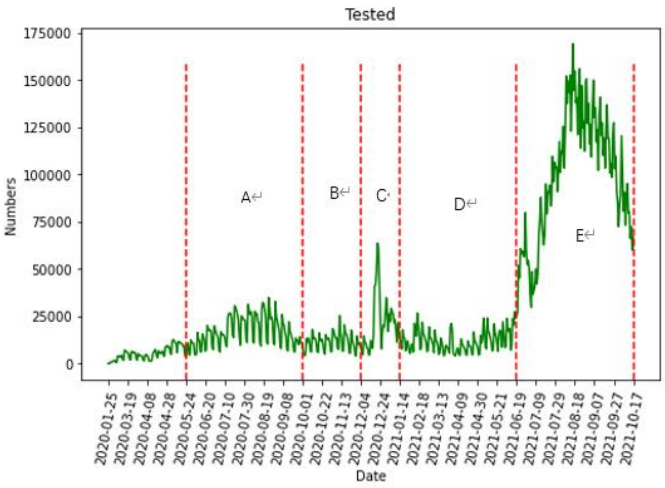
```
fig, ax1 = plt.subplots(figsize=(8, 5))



ax1.set_xlabel('Date')
ax1.set_ylabel('Numbers of confirmed')
ax1.set_xticklabels("?", rotation = 75)
ax1.set_ylim(0, 2500)
l1, = ax1.plot(index[::1],y_confirmed.date_count[::1],color='blue',linewidth=2,linestyle='
ax1.tick_params(axis='y')

ax2 = ax1.twinx()
ax2.set_ylim(0, 180000)
ax2.set_ylabel('Numbers of tested')
l2, = ax2.plot(index[::1],y_tested.test_count[::1],color='green',linewidth=2,linestyle='s
ax2.tick_params(axis='y')

#ax1.axvline("2021-09-13", linestyle='—', color='orange')

plt.title("Comparison")
plt.xticks(index[0:len(index):30])
plt.legend(handles=[l1,l2],labels=['confirmed','tested'],loc="upper left")



fig.tight_layout()
plt.show()
```

2. Your contribution to the visualization-related tasks. Describe what you did to make contributions towards "visualization-related" tasks

The part I am responsible for is the "Encouraging test" part. Firstly I tried to visualize both confirmed cases and test numbers in one line chart with single y-axis, the result is shown below:

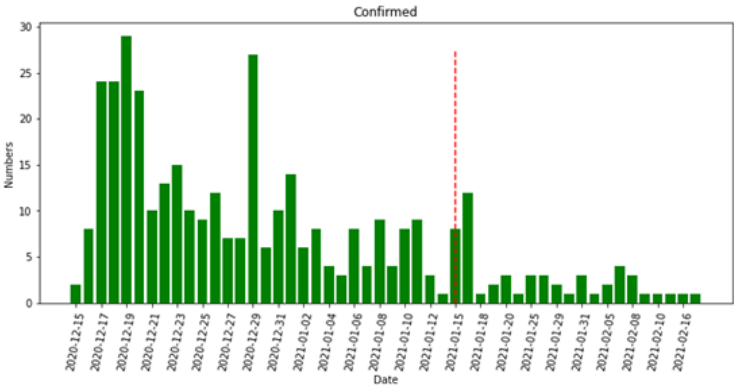


But for some reason, this chart is also useful for other members' work, even better than my own part. So, for my own part, I create two new charts. First one is to find the time period that government encourage covid-19 test:



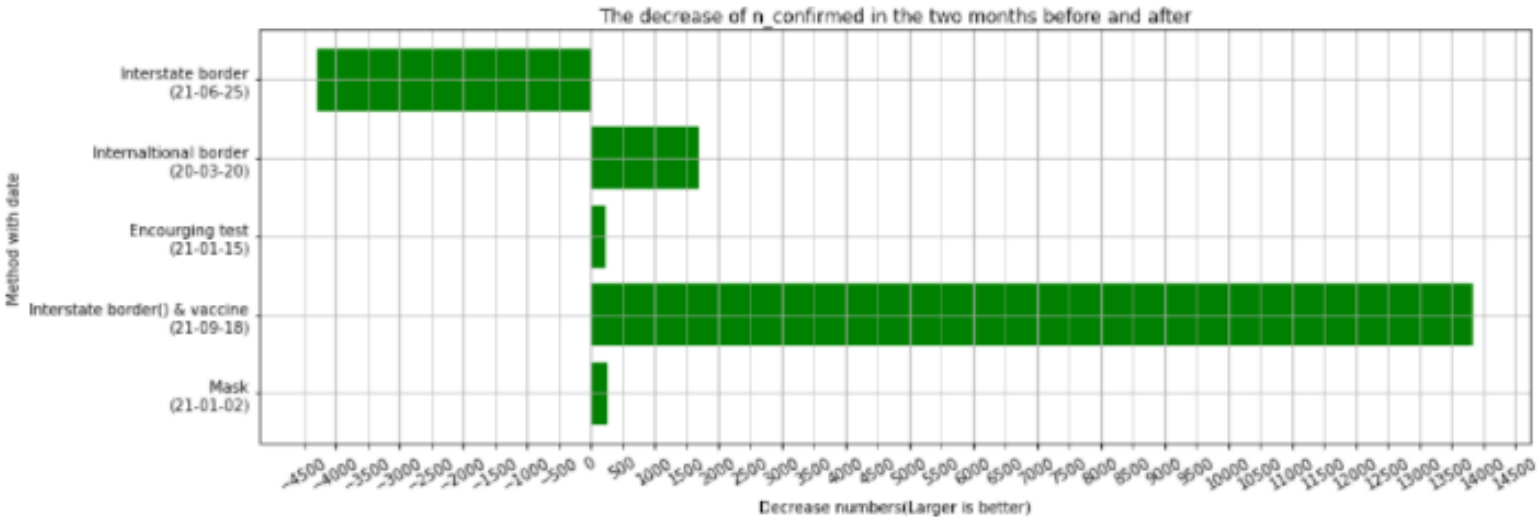
We can see they are A, C and E which is apparently greater than B and D.

The axis and visual variables arrangement of mine: "The number of tests is ratio data, because the number of tests can be true 0, and the date is ordered data. They are homogeneous. We represent the date via x -axis and using y-axis to represent the number of tests. We can observe the trend in a period of the line chart to easily determine whether in this period the number of tests increased and when it started to decrease." [6] Then I design this bar chart to qualitatively judge whether the measure is effective:



And the reason of my arrangement is "Confirmed numbers are ratio data, because it can be true zero, and date is internal data, they are homogeneous. We used x axis to represent date, y axis to represent numbers of confirmed, determine whether the numbers of testing are high or low via the height of columns in different sides of red dotted line." [6] Which is same as my group report.

Another chart I designed is this one:



The establishment of this bar chart is to summarize whether a measure is useful or not, in a quantitative way. I designed it as a bar chart because it can be clearly to determine whether the decrease of n_confirmed in the two months before and after are positive or not, If not, I can conclude that this method did not achieve the KPI. And another benefit of this bar chart is that I can immediately find the best method by compare the height or length of each bar, seems like that the combination of lockdown and vaccine is the best one. My arrangement: “Decrease numbers are internal data, because it does not have true zero, the decrease numbers can be negative, which means n_confirmed increase after month. We represent these decrease numbers by the horizontal axis. They are homogeneous. As for methods (or combination), they are nominal data and heterogeneous. We arrange them on the vertical axis. Bar chart can clearly distinguish the performance of a particular method via the length of bar, and another reason why we use bar chart is that the vertical axis represents heterogeneous data.” [6]

3. For my contributions with respect to what I’ve learnt

1. Data Types in week 2 lectures. I divide the data into four types: nominal, ordinal, interval and ratio to help deriving visualization. For example, it’s not appropriate to arrange particular methods like mask, test, vaccine as line chart axis, because they are heterogeneous. I did contributions in “data types” direction for every chart I designed.[7]

2.Exploratory data analysis and visualization design in week 5 , 6 and 7 lectures. The major types of chart we applied for this assignment are line chart and bar chart. I combine their respective advantages to arrange a specific visualization method for each specific problem. As for specific decision, I have mentioned in detail for every single visualizations above.[8]

3. The definition of KGI and KPI in week 8 lecture. I use these metrices to quantify the performance of every method. My set KGI as completely control the COVID19, and KPI “Decrease of confirmed cases monthly before and after”. As the result, the combination of vaccine and lockdown seems like the best one in KPI. I also did visualization for this KPI, which is mentioned above in part 2.[4]

References

[1] data.nsw.gov.au

[2] Yuancccc，高斯曲线拟合，https://blog.csdn.net/Yuancccc/article/details/85307980

[3] Kuan Li, 解析 KGI、CSF、KPI——数据分析的一种思路 http://www.woshipm.com/data-analysis/590642.html

[4] Masahiro Takatsuka, USYD-COMP5048-WEEK8-Lecture

[5] https://matplotlib.org/stable/contents.html

[6] Group RETUT01-04 of USYD COMP5048, Assignment2 report

[7] Masahiro Takatsuka, USYD-COMP5048-WEEK2-Lecture

[8] Masahiro Takatsuka, USYD-COMP5048-WEEK5,6,7-Lecture