BDM300 Project

Name: Ran Arino (ID: 153073200)

1. Description of the data set

All the data is collected from the website and belongs to a Tokyo in Japan. The data set is about COVID-19, the hospitalized situation, some population transfer data (from highway and airport), and some social data.

The original data set is composed of 30 rows and 16 columns; each row shows each month between March 2020 and August 2022, and the column is like the following.

(1) ***date*** (chr):
Monthly date from March 2020 to August 2022.

(2) ***num_positives*** (num; binary):
Whether the number of COVID test positives in Tokyo is greater than the one of the previous month or not; “1” = increase, “0” = equal or decrease.
Source: https://catalog.data.metro.tokyo.lg.jp/en/dataset/t000001d0000000011

(3) ***num_tests*** (num; binary):
whether the number of total Covid-19 tests in Tokyo is greater than the one of the previous month or not; “1” = increase, “0” = equal or decrease.
Source: https://catalog.data.metro.tokyo.lg.jp/en/dataset/t000010d0000000086

(4) ***per_positives*** (num; binary):
Whter the percentage of the positives in Tokyo (number of positives / number of tests in a month * 100) is greater than the one of the previous month or not; “1” = increase, “0” = equal or decrease.

(5) ***num_consult*** (num; binary): Whether the number of consultations at the novel coronavirus call center in Tokyo is greater than the one of the previous month or not; “1” = increase, “0” = equal or decrease.
Source: https://catalog.data.metro.tokyo.lg.jp/en/dataset/t000010d0000000071

(6) **num_hospital_patients** (num; binary):
Whether the number of hospitalized patients in Tokyo is greater than the one of the previous

month or not; “1” = increase, “0” = equal or decrease.

Source: https://catalog.data.metro.tokyo.lg.jp/en/dataset/t000010d0000000092

(7) ***minor_moderate_sym*** (num; binary):

Whether the number of patients whose symptoms are minor or moderate in Tokyo is greater than the one of the previous month or not; “1” = increase, “0” = equal or decrease.

Source: https://catalog.data.metro.tokyo.lg.jp/en/dataset/t000010d0000000092

(8) ***severe_sym*** (num; binary):

Whether the number of patients whose symptoms are severe in Tokyo is greater than the one of the previous month or not; “1” = increase, “0” = equal or decrease.

Source: https://catalog.data.metro.tokyo.lg.jp/en/dataset/t000010d0000000092

(9) ***per_severe*** (num; binary):

Whether the percentage of severe patients (number of severe / number of hospitalized patients * 100) in Tokyo is greater than the one of the previous month or not; “1” = increase, “0” = equal or decrease.

(10) ***highway_traffic*** (num; binary):

Whether the number of traffic on the Metropolitan Expressway is greater than the one of the previous month or not; “1” = increase, “0” = equal or decrease.

Source: https://www.shutoko.co.jp/company/database/traficdata/

(11) ***air_domestic_passenger*** (num; binary):

Whether the number of domestic passengers in Haneda Airport is greater than the one of the previous month or not; “1” = increase, “0” = equal or decrease.

Source: https://www.tokyo-airport-bldg.co.jp/result/

(12) ***air_foreign_passemger*** (num; binary):

Whether the number of foreign travelers in the international airport in Haneda is greater than the one of the previous month or not; “1” = increase, “0” = equal or decrease.

Source: https://www.tokyo-airport-bldg.co.jp/result/

(13) ***air_total_passenger*** (num; binary):

Whether the total number of passengers in Haneda Airport is greater than the one of the previous month or not; “1” = increase, “0” = equal or decrease.

Source: https://www.tokyo-airport-bldg.co.jp/result/

(14) ***restriction_period*** (num; binary):

Whether Tokyo was under the restriction (either State of Emergency Declaration or Semi-

emergency Coronavirus Measures) in each month or not; “1” = under any restriction or “0” = no restriction.

(15) **unemployment_rate** (num; binary):

The unemployment rate – the unemployed are people of working age who are without work, are available for work – is greater than the one of the previous month or not; “1” = increase, “0” = equal or decrease.

Source: https://data.oecd.org/unemp/unemployment-rate.htm

(16) ***telework_rate***:

The percentage of teleworking companies in Tokyo is greater than the one of the previous month or not; “1” = increase, “0” = equal or decrease.

Source:

2021-04 and 2022-08:

https://www.metro.tokyo.lg.jp/tosei/hodohappyo/press/2022/09/12/04.html

2020-03, 2020-04, 2020-12 & between 2021-01 and 2021-03 (mean of 2 rates; first and second half). Between 2020-05 and 2020-11 is projected from the graph by myself, which means that it’s not precise data.

https://www.metro.tokyo.lg.jp/tosei/hodohappyo/press/2021/12/09/06.html

3. Analysis Approach and Business Problems/Questions (hypothesis)

The data set includes data about COVID-19, hospital situations, human transfer data, and some social data. Considering those data, I define the following hypothesis or business problems before exploring data:

(1) The number of COVID-19 test positives will relate to the number of consultations at the novel coronavirus call center – a possible indicator to show human anxiety against COVID-19.
-> both data could have a positive association.

(2) The more people reach the novel coronavirus call center for the purpose of the consultation, the more the COVID-19 tests could be taken. In other words, humans would like to confirm their safety by proving that they aren't infected.
-> both data could have a positive association.

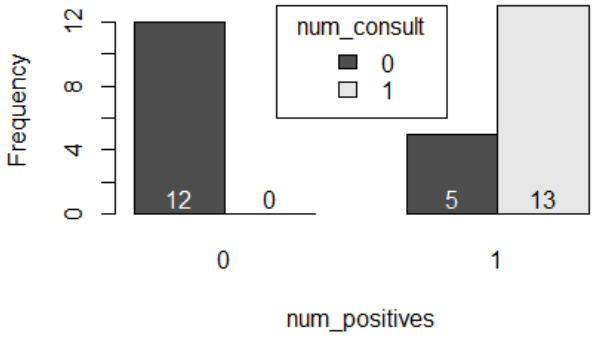
(3) If the percentage of the severe rate against the number of total hospitalized patients increases, the people would have more anxiety and reach COVID-19 call center.
-> both data could have a positive association (occurs at the same time).

(4) If Tokyo is under any restriction period,
   a. the volume of Metropolitan Expressway traffic will decrease. (Restrict on moving)
   b. the number of domestic passengers in Haneda airport will decrease. (Restrict on moving)
   c. the unemployment rate will increase. (Restrict on the business hours)
   d. the percentage of teleworking companies in Tokyo will increase (Restrict on in-person work)

(5) The variable "restriction_period" will not be associated with the following variable: num_positives and num_hospital_patients.
-> It is because the Government of Japan or the Tokyo metropolitan Government may announce the state of the emergency (restriction) when those two numbers will increase, governments will deregulate the restriction when those two numbers will decline sufficiently (monthly base). That's why the monthly based increase and decrease of above two numbers may occur independently regardless of the restriction period.

4. Analyze the dataset

Hypothesis (1):
***num_positives* and *num_consult* have a positive relationship; the increase in test positives may cause a rise in consultations at the novel coronavirus call center.**

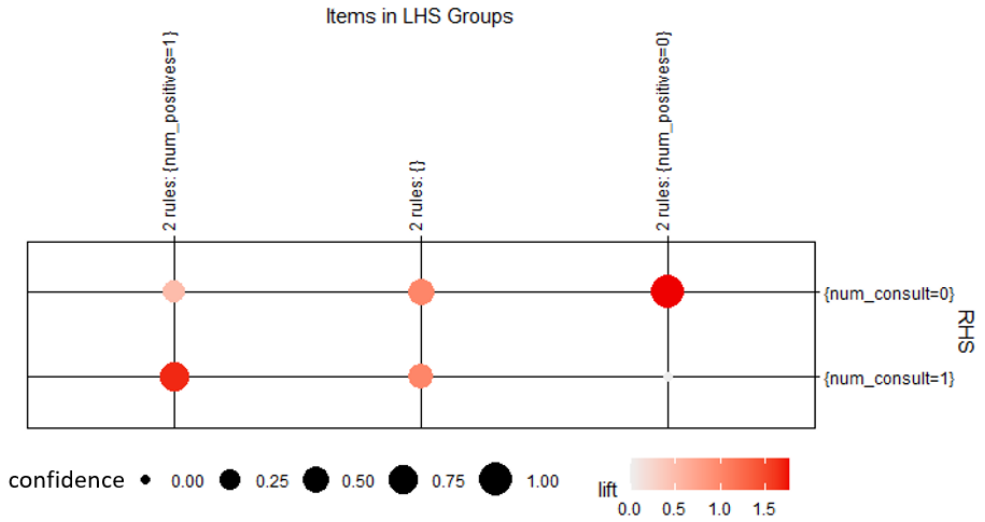


The above graph shows the frequency of *num_consult* given the condition of *num_positives*. As shown in the graph, when the number of test positives compared to the previous month declines (*num_positives* = 0), the number of consultations at the novel coronavirus call center simultaneously declines (*num_consult* = 0) as well, whose occasion is 100% (12 out of 12 times).

Furthermore, when the number of test positives increases, the number of consultations increases by around 72% of chance (13 out of 18 times).

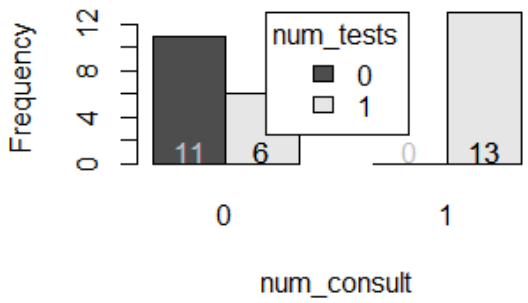
From the aspect of the association rules, the following rows are matched in terms of assessing hypothesis (1):

rules$lhs	rules$rhs	support	confidence	coverage	lift	count
{num_positives=0}	{num_consult=1}	0.00	0.00	NaN	0.00	0
{num_positives=1}	{num_consult=0}	16.67	27.78	0.6	0.4901961	5
{num_positives=0}	{num_consult=0}	40.00	100	0.4	1.7647059	12
{num_positives=1}	{num_consult=1}	43.34	72.22	0.6	1.6666667	13



Considering that a higher lift value (greater than 1.5, which is shown in red color in an interactive graph) of both "{*num_positives*=0} => {*num_consult*=0}" and "{*num_positives*=1} => {*num_consult*=1}" and lower lift value (less than 0.5) of others, we may conclude that the number of test positives and the number of consultants is positively correlated. For example, the probability of two events at the same time – the increase in COVID test positives and the raise in consulting cases – is 1.67 times greater than the case where we assume that two events occur independently with no association at all. Also, the 0.49 lift value (under a situation where the increase in the number of test positives and a decrease in the number of consultants by monthly base) would decline the possibility that num_positives and num_consultants are negatively correlated.

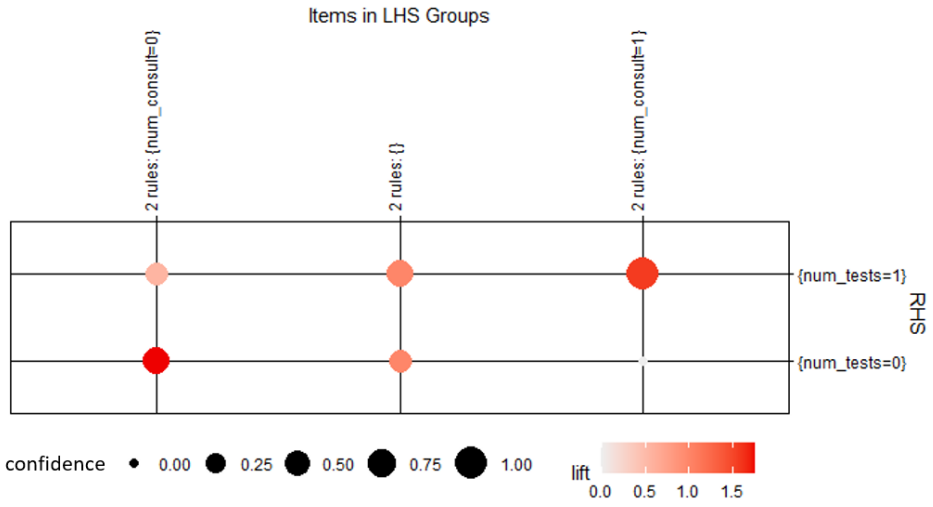
Hypothesis (2): the more consulting cases at the novel coronavirus call center are, the more the COVID test cases will be taken, which means that two events are positively correlated.



The above graph shows the frequency of *num_tests* given the condition of *num_consult*. As shown in the graph, when the number of consulting cases compared to the previous month increases (*num_consult*= 1), the number of tests simultaneously increases (*num_tests* = 1) as well, whose occasion is 100% (13 out of 13 times). Furthermore, when the number of consulting cases decreases, the number of tests decreases by around 65% of probability (11 out of 17 times), which might not be significantly higher rate.

From the aspect of the association rules, the following rows are matched in terms of assessing hypothesis (2):

rules$lhs	rules$rhs	support	confidence	coverage	lift	count
{num_consult=0}	{num_tests=1}	20.00	35.29	0.5572	0.5572755	6
{num_consult=1}	{num_tests=0}	0.00	0.00	NaN	0.00	0
{num_consult=0}	{num_tests=0}	36.67	64.71	0.5667	1.7647059	11
{num_consult=1}	{num_tests=1}	43.34	100.00	0.4333	1.5789474	13

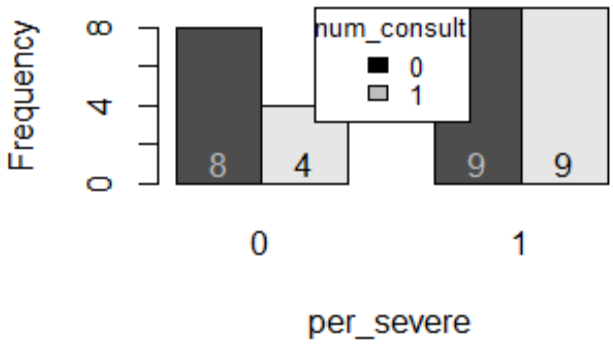


From the above table and graph, we can understand that the probability that both the number of consulting cases and tests decrease at the same time in a certain month is 1.76 times higher than the case where both are assumed to occur independently. Although the confidence rate may be moderate rate (64.71%), the lift value would be high enough to support the positive correlation between the two

numbers. The probability that both numbers increase simultaneously is 1.58 times higher than by any chance (the confidence is also 100%, which is shown in a bigger size of the circle in an interactive graph). Furthermore, the probability that the number of consulting cases decreases and the number of tests increases at the same time in a certain month is 0.56 times higher than by any chance, which indicates that its probability declines by around half. This outcome would support the rejection of a possible negative correlation between the two numbers.

Considering those observations, we can conclude that the number of consulting cases at the COVID call center and the number of COVID tests would have a positive correlation.

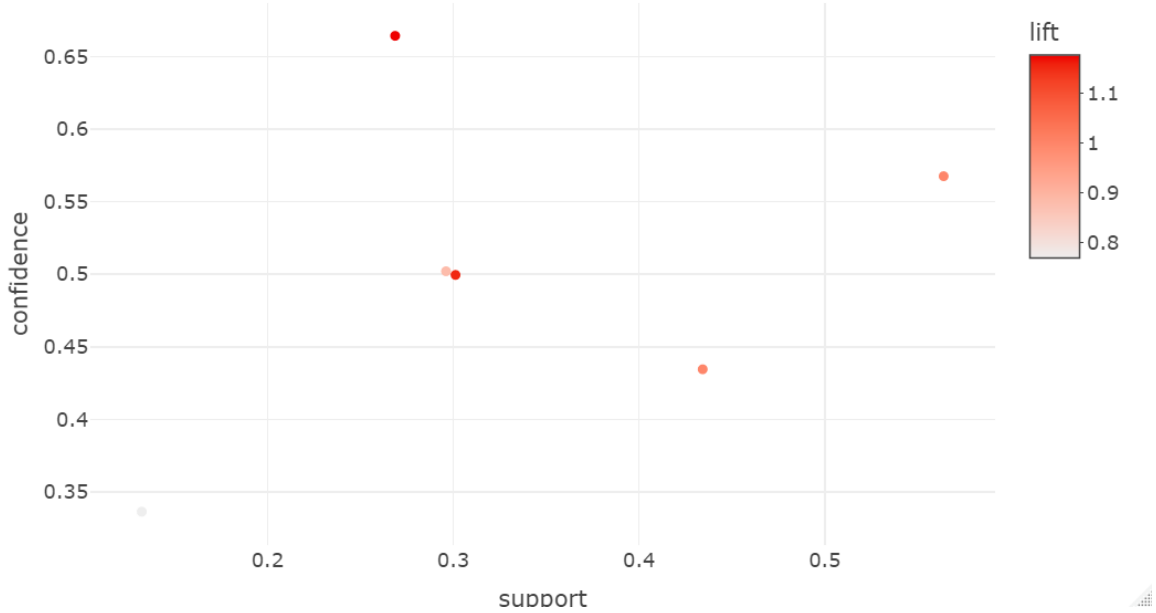
Hypothesis (3): the higher the percentage of patients whose symptoms are severe is, the greater the number of consulting cases at the COVID call center; *per_severe* and *num_consult* would be positively correlated.



From the above graph, it is likely that there is not a significant difference in the number of consulting cases by whether the percentage of severe symptoms increases or decreases. One of the reasons is that the number of consulting cases is the same regardless of the up and down percentage of severe symptoms in hospitals. When the percentage of severe symptoms decreases, the number of consulting cases tends to decrease as well, but it may not be a significant difference.

From the aspect of the association rules, the following rows are matched in terms of assessing hypothesis (3):

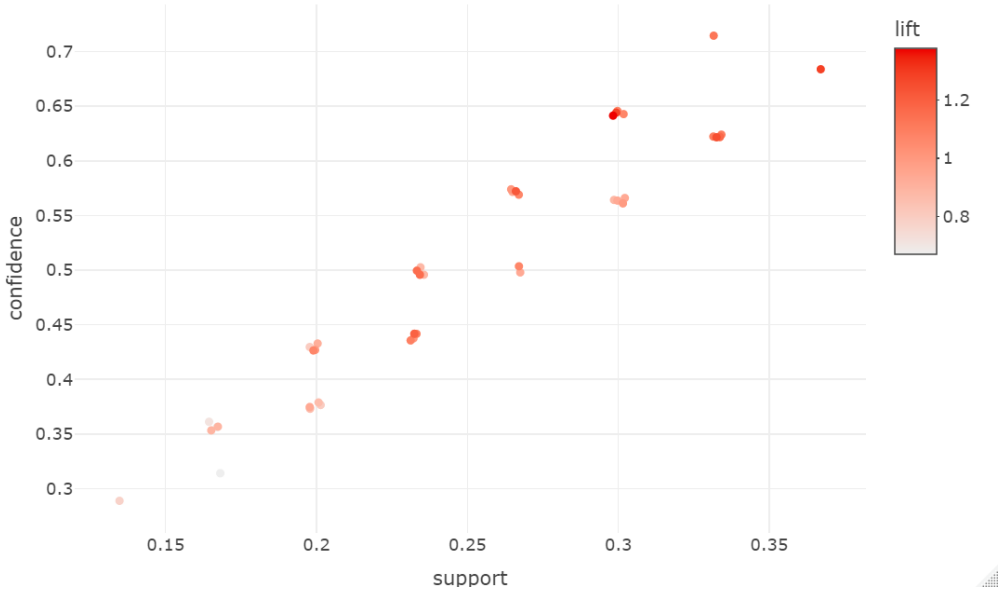
rules$lhs	rules$rhs	support	confidence	coverage	lift	count
{per_severe=0}	{num_consult=1}	13.34	33.34	0.4	0.7692308	4
{per_severe =0}	{num_consult=0}	26.67	66.67	0.4	1.1764706	8
{per_severe =1}	{num_consult=1}	30.00	50.00	0.6	1.1538462	9
{per_severe =1}	{num_consult=0}	30.00	50.00	0.6	0.8823529	9



From the above table and graph, we can see that the confidence and the lift value are not high enough to support a positive correlation of them when both the percentage of severe symptoms and the number of consulting cases increase or decrease compared to the previous month. If we focus on the lift value, the occurrence of the increase in severe symptoms' percentage and both consulting cases is 1.18 times greater than the assumption that both are independent. This ratio is close to 1, which shows that the two variables are independent.

Thus, even if the percentage of severe symptoms against the capacity of patients in the hospital increases and decreases, the number of consulting cases may not be affected; both have neither positive nor negative correlation.

Hypothesis (4): If the restriction (either State of Emergency Declaration or Semi-emergency Coronavirus Measures) starts or remains in Tokyo on any day in a month, a) the volume of Metropolitan Expressway traffic will decrease, b) the number of domestic passengers in Haneda airport will decrease, c) unemployment rate will increase, and d) the percentage of teleworking companies will increase.



This graph shows the support, confidence, and lift of each association rules (all ruls$lhs are either {restriction_period = 0} or {restriction_period = 1}). There are a lot of data points that have higher confidence and lift value, so we can expect to find many positive or negative relationships among variables. The below table focuses more on my hypotheses:

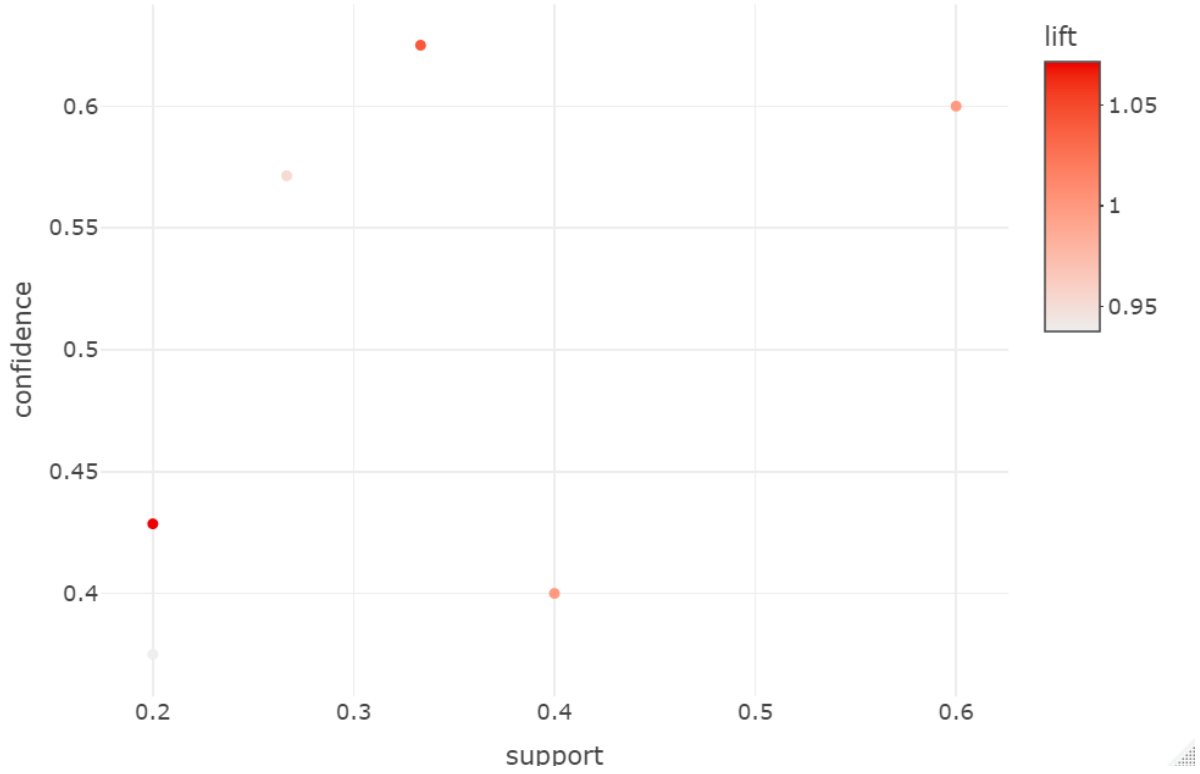
rules$lhs	**rules$rhs**	**support**	**confidence**	**coverage**	**lift**	**count**
{restriction_period =0}	{highway_traffic=1}	33.34	62.50	0.533	1.1718759	10
{restriction_period =1}	{highway_traffic=0}	26.67	57.14	0.467	1.2244898	8
{restriction_period =0}	{air_domestic_passenger=1}	36.67	68.75	0.533	1.2890625	11
{restriction_period =1}	{air_domestic_passenger=0}	30.00	64.29	0.467	1.3775510	9
{restriction_period =0}	{unemployment_rate=0}	30.00	56.25	0.533	0.9375000	9
{restriction_period =1}	{unemployment_rate=1}	16.67	35.71	0.467	0.8928571	5
{restriction_period =0}	{telework_rate=0}	33.33	62.50	0.533	1.1029412	10
{restriction_period =1}	{telework_rate=1}	23.33	50.00	0.467	1.1538462	7

a) In terms of the relationship between the restriction period and the volume of highway traffic, both may be negatively correlated but the confidence and the lift value are not high enough to support the hypothesis. For example, the probability that no restriction erupts on any day in Tokyo in a month and the increase in highway traffic is 1.17 times higher than the case where both events occur independently. Although the confidence is relatively high (62.50%), the lift value should be close to 1.5 at least in order to support the hypothesis. Thus, I conclude that the restriction period and the high traffic volume don't have a significant negative correlation.
b) As for the relationship between the restriction period and the number of domestic passengers in Haneda Airport, we may conclude that both could be negatively correlated if we take into consideration the outcome of the confidence and the lift value. For example, the probability that the no restriction erupts on any day in Tokyo in a month and the increase in domestic passengers in Haneda airport is 1.29 times higher than the case where both occur independently. Also, its confidence is relatively high (68.17%), which will support their negative association. Thus, it is likely that the restriction period and the number of domestic airplane passengers have a negative relationship.
c) I expected that both the invocation of restriction and the increase in the unemployment rate would occur at the same time, but the association rule doesn't support my hypothesis. For example, the probability that Tokyo is under a restriction period on any day in a month in Tokyo and the increase in the unemployment rate is 0.89 higher than in the case where both occur independently. Also, their confidence is a low rate, so I can't obtain evidence to support their positive relationship. One of the possibilities of causing this outcome could be the error of data extraction. More specifically, the announced unemployment rate per month is likely to be based on the previous month's data (the unemployment rate which is disclosed in October could reflect the data in September). Thus, an error in data integration or time lag might be caused in the pair of *restriction_period* and *unemployment_rate* data.
d) The last one is the positive relationship between the going-out restriction and the percentage of teleworking (remote working) in Tokyo. Considering that the lift values are 1.10 (*restriction_period = teleworking_rate = 0*) and 1.15 (*restriction_period = teleworking_rate = 1*), we can't assert that there is a positive association between the restriction and teleworking. Also, both confidence is close to 50%, which could support that *restriction_period* and

teleworking_rate occur independently. Thus, it may be difficult to assert that the restriction period and the teleworking rate have a positive correlation.

Hypothesis (5): Even if Tokyo is under restriction or not, the increase and decrease in the number of test positives and of patients in hospitals can occur independently.

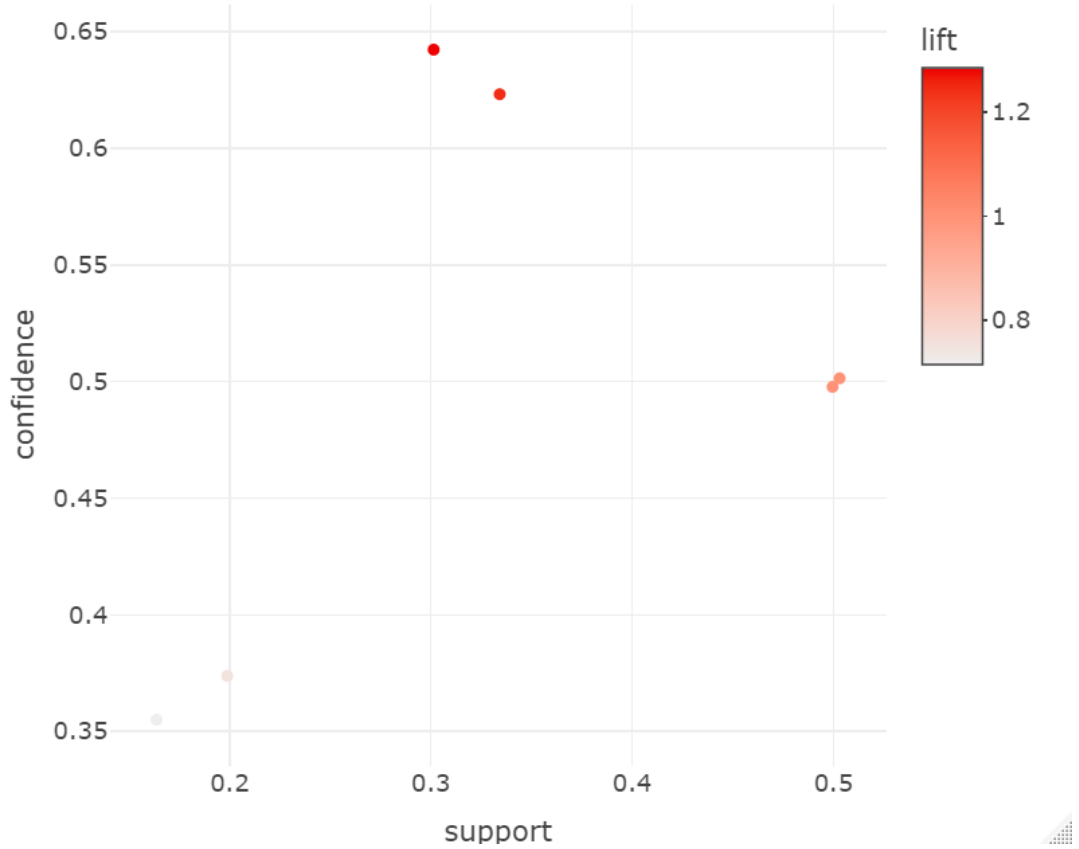
rules$lhs	rules$rhs	support	confidence	coverage	lift	count
{restriction_period =0}	{num_positives=0}	20.00	37.50	0.533	0.9375000	6
{restriction_period =1}	{num_positives=1}	26.67	57.14	0.467	0.9523810	8
{restriction_period =0}	{num_positives=1}	33.33	62.50	0.533	1.0416667	10
{restriction_period =1}	{num_positives=0}	20.00	42.85	0.467	1.0714286	6



The above table shows the association rules related to the *restriction_period* and *num_positives.* If we focus on the lift value, all values are close to 1, which means that two events will occur independently

with neither positive nor negative correlation. Also, each confidence is not too high from the graph, so it is likely that *restriction_period* and *num_positives* have no correlation.

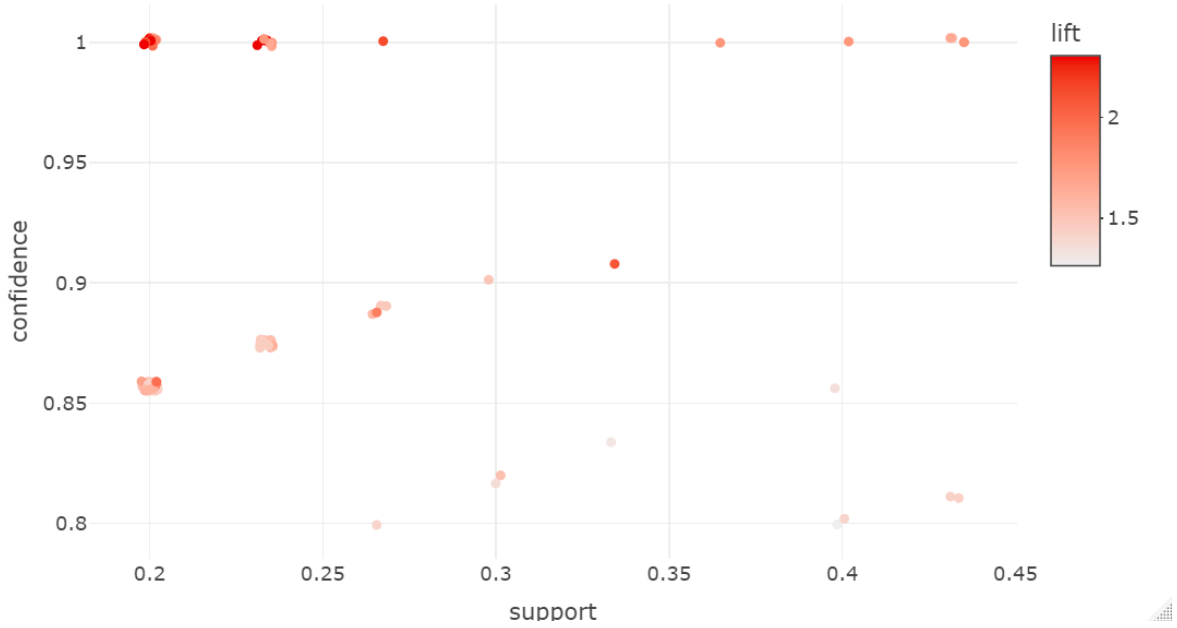
rules$lhs	rules$rhs	support	confidence	coverage	lift	count
{restriction_period =0}	{num_hospital_patients=0}	20.00	37.50	0.533	0.7500000	6
{restriction_period =1}	{num_hospital_patients=0}	30.00	64.29	0.467	1.2857143	9
{restriction_period =0}	{num_hospital_patients=0}	33.33	62.50	0.533	1.2500000	10
{restriction_period =1}	{num_hospital_patients=1}	16.67	35.71	0.467	0.7142857	6



The above table and graph shows the association rules related to the *restriction_period* and *num_hospital_patients*. Considering that the confidence and lift value of all, they might be negatively correlated. It is because the probability that Tokyo is under restriction on any day and the number of

patients in the hospital decrease in a month is 1.29 times higher than the case where both are assumed to occur independently. Also, the confidence is relatively high (64.29%). Thus, the *restriction_period* and *num_hospital_patients* may occur independently but there is a possibility that both have negative correlation.

In the end, I will run the association rule mining for the whole pattern (removed dupplication). The graph below shows the support, confidence, and lift of the whole pair of variables.



Some association patterns show a high lift value (more than 2), especially those patterns are distributed around high confidence and low support. It means that there is a clear positive or negative relationship among variables in each pattern although the probability of those patterns occurring is low against the whole.

rules$lhs	rules$rhs	support	confidence	coverage	lift	count
{severe_sym=1, highway_traffic=0}	{num_consult=1}	23.33	100.00	0.233	2.307692	7
{unemployment_rate=0, telework_rate=1}	{severe_sym=0}	20.00	100.00	0.200	1.875000	6

The above table shows the interesting association rules with high lift values, which I picked up.

The first one indicates that people are more likely to call the novel coronavirus call center in Tokyo when the number of severe symptoms in Tokyo increases and the traffic volume of highways decreases. In other words, people may feel anxiety (-> consulting about individual health) when they reduce going out as well as possible and hear about the increase in severe symptoms from the news.

The second one shows that the number of severe symptoms may decrease if the unemployment rate decreases and the rate of teleworking raises at the same time. It means that if the number of teleworkers increases with the improvement of job situations, the symptoms of infected people could ease; one of the possibilities is that the contact among people will be less due to workers won't need to use the high-crowded train (a considerable amount of people tend to push with each other and be packed in train in Japan, especially in the morning). The other possibility could be economic relief. If people can find jobs, they will obtain enough money to go to the hospital compared to the situation where they don’t have jobs. The earlier people go to the hospital and take appropriate measures, the smaller the possibility that they struggle with severe symptoms by the covid.