Name: Tang Ho Chun

Background and Environment Setup

Given a dataset related to COVID-19 to perform data analysis and study. Finally derive a set of insights and suggestions, and propose a model for predicting new cases percentage.

The notebook reads the data resource online via google oauth.

Number of records: 3864
Number of features: 90
Number of numeric features (including labels): 88

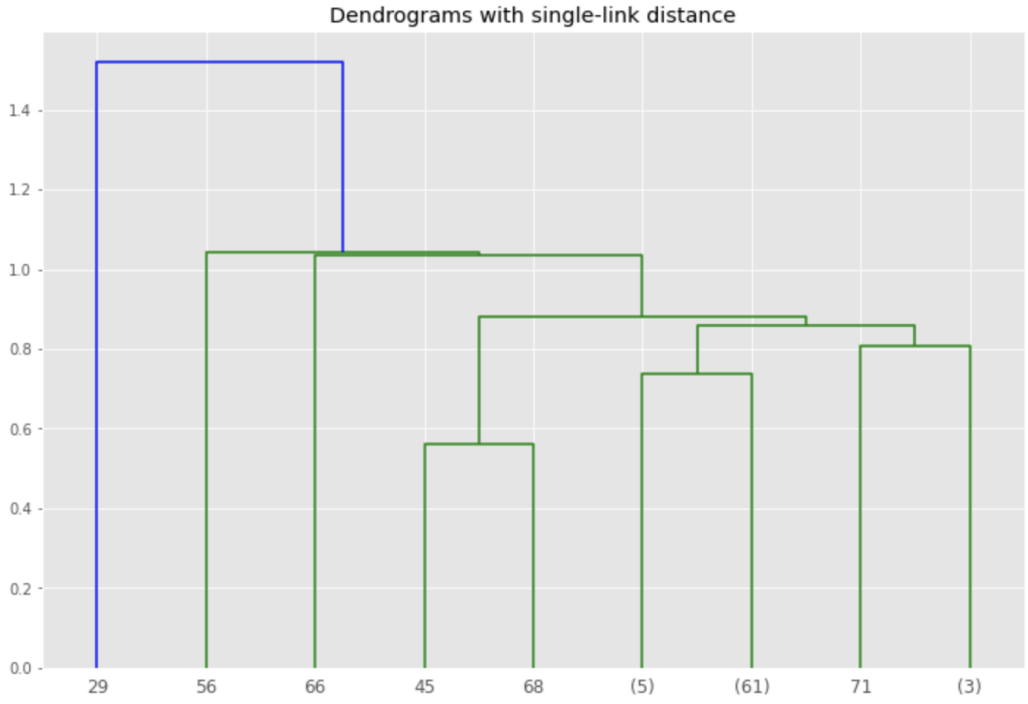
Task 1 — Country Data

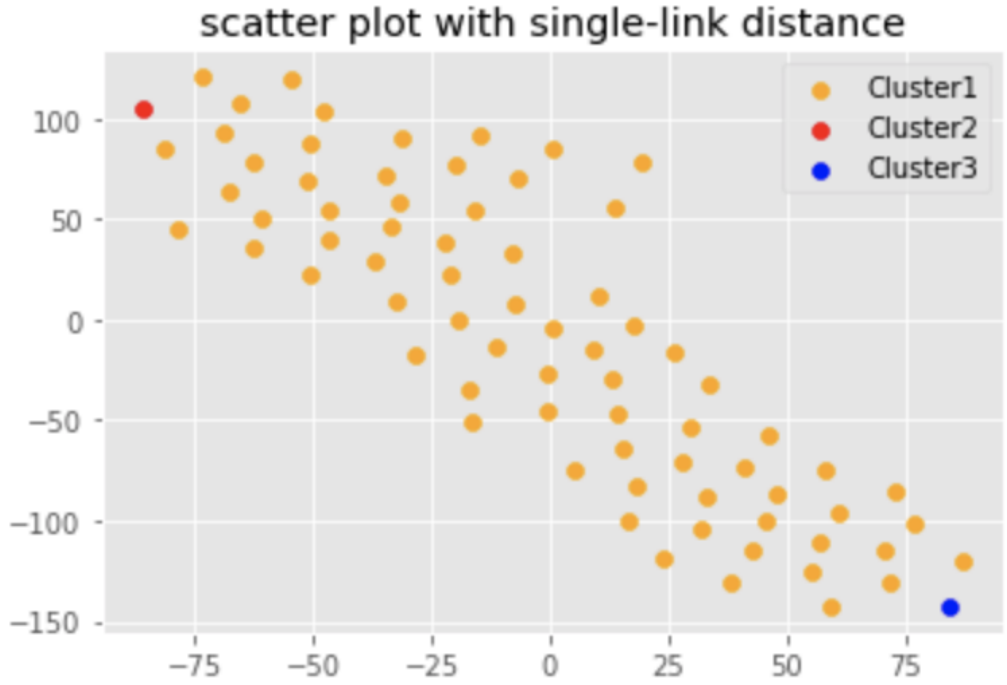
1a. Data preprocess

(1) Remove duplicate:
    # data before removing duplicate: 3864
    # data after removing duplicate: 107
(2) Imputation:
    # Total NAs before imputation: 76
    - pop_density: 1
    - safe_water: 36
    - safe_san: 38
    # Total NAs after imputation: 0
(3) Standardize: mean = 0, std = 1
(4) Remove outliers:
    # data before remove outliers: 107
    IQR based removal:
    - Consider outliers if 1.5 IQR away from first and third quartile
    - # data after removal: 75
    Z-score based removal:
    - Consider outliers if z-score absolute value >= 1.5
    - # data after removal: 76
    Use IQR based removal for later analysis

1b. Hierarchical clustering

(i) Single-link distance

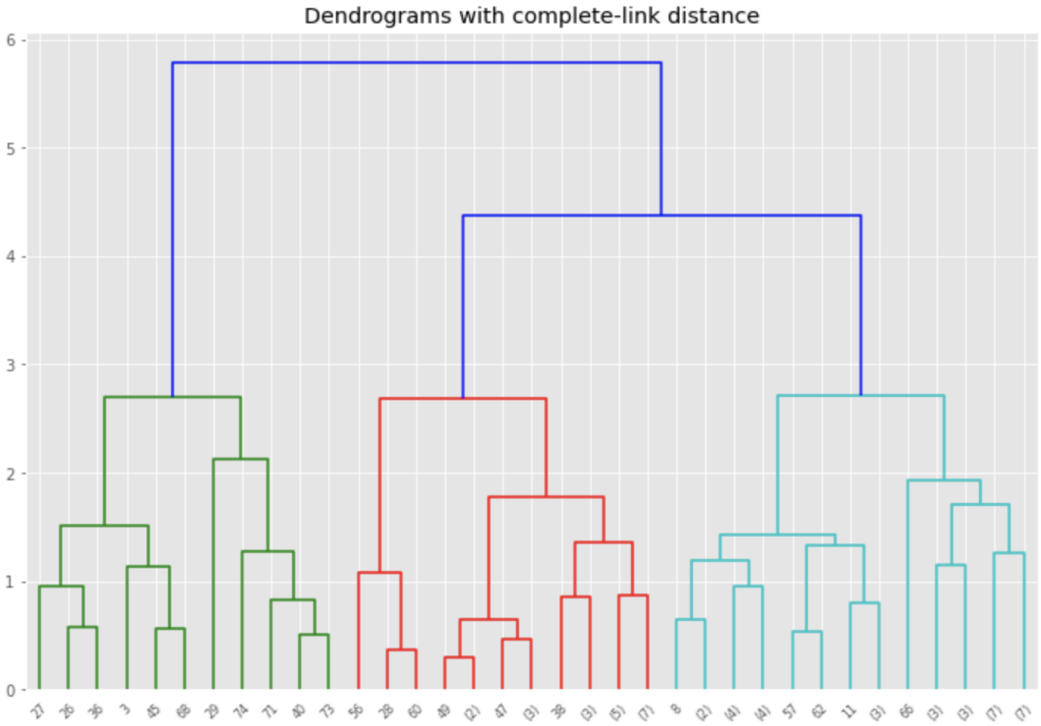


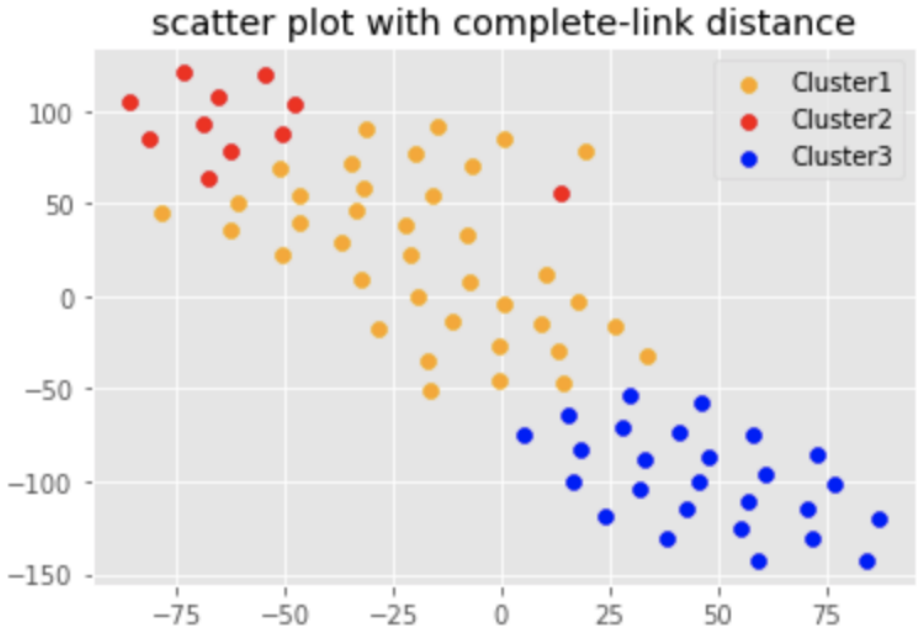



```
For single-link distance:

Number of elements in cluster 0 is 73
Number of elements in cluster 1 is 1
Number of elements in cluster 2 is 1
```

(ii) Complete-link distance

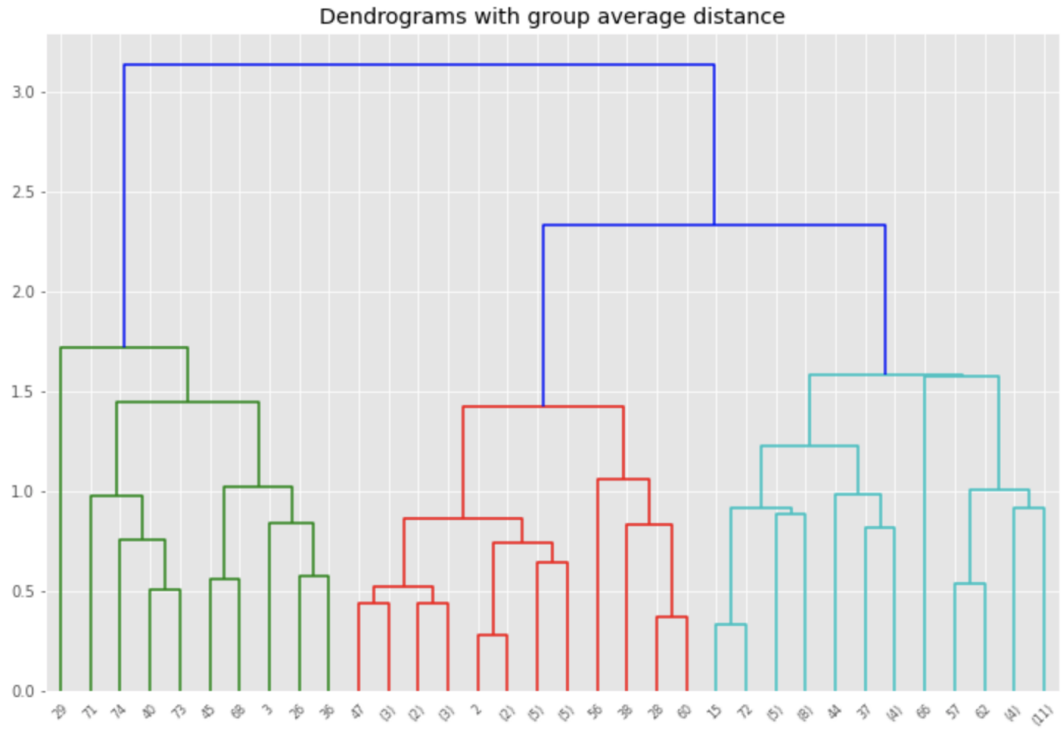


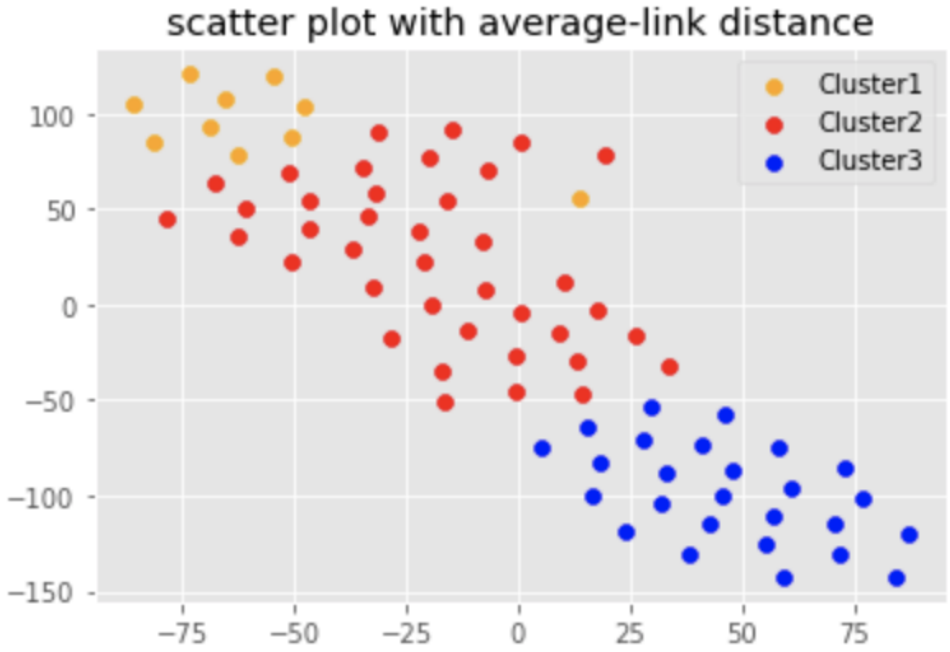



```
For complete-link distance:

Number of elements in cluster 0 is 38
Number of elements in cluster 1 is 11
Number of elements in cluster 2 is 26
```

(iii) Average-link distance






```
For average-link distance:

Number of elements in cluster 0 is 10
Number of elements in cluster 1 is 39
Number of elements in cluster 2 is 26
```

1c. Clustering Validity Measures

Davies–Bouldin Score Result:

- single-link: 0.5137742665128471
- complete-link: 0.819269481106398
- average-link: 0.7733178860340701

For the Davies–Bouldin metric, the lower the value the farther the clusters are apart, so single-link distance is the best of 3 under this metric.

Silhouette Score Result:

- single-link: 0.10201944334301274
- complete-link: 0.4138670479526472
- average-link: 0.42968049502721356

For the Silhouette score, it measures the closeness of intra-cluster elements with respect to alternatives. The closer the value to 1 is better, so average-link distance is the best of 3 under this evaluation.

1d. Visualization

(i) Three country clusters:

Group 1:

Countries: 'Bolivia', 'Guatemala', "Lao People's Democratic Republic", 'Peru', 'Cambodia', 'Myanmar', 'Nicaragua', 'Nepal', 'Senegal', 'Mali'

	index	pop_total	pop_density	GDP	basic_water	safe_water	basic_san	safe_san
count	10.000000	1.000000e+01	10.000000	10.000000	10.000000	6.000000	10.000000	5.000000
mean	231.100000	2.094376e+07	74.055834	6468.767884	84.978949	37.852119	62.543983	32.784351
std	278.956568	1.428360e+07	61.137676	3379.900334	6.105321	16.762400	11.095004	17.040602
min	3.000000	6.545502e+06	10.480146	2423.828765	78.260830	16.081866	39.335420	18.709404
25%	51.000000	1.270892e+07	26.392651	3811.473443	80.889581	26.195000	59.599459	21.455765
50%	72.000000	1.654528e+07	67.982832	5493.235881	81.917973	38.797002	63.193140	22.938271
75%	297.000000	2.637104e+07	89.629179	8784.341511	90.548966	51.288046	72.023001	42.764723
max	767.000000	5.404542e+07	195.939107	13380.364420	94.190581	55.990782	74.459410	58.053590

This group is the poorest (Lowest GDP) with very poor hygiene measures.

Group 2:

Countries: 'Kyrgyzstan', 'Ukraine', 'Libyan Arab Jamahiriya', 'Dominican Republic', 'Egypt', 'Oman', 'Costa Rica', 'Colombia', 'Uzbekistan', 'Morocco', 'South Africa', 'El Salvador', 'Vietnam', 'Azerbaijan', 'Algeria', 'Moldova, Republic of', 'Paraguay', 'Thailand', 'Bosnia and Herzegovina', 'Ecuador', 'Jordan', 'Venezuela', 'Sri Lanka', 'Serbia', 'Albania', 'Tunisia', 'Croatia', 'Turkey', 'Honduras', 'Panama', 'Uruguay', 'Kazakhstan', 'Romania', 'Bulgaria', 'Chile', 'Puerto Rico', 'Argentina', 'Belarus', 'Iraq'

	index	pop_total	pop_density	GDP	basic_water	safe_water	basic_san	safe_san
count	39.000000	3.900000e+01	39.000000	39.000000	39.000000	24.000000	39.000000	22.000000
mean	89.205128	2.464838e+07	95.322751	17489.428637	95.572848	81.524968	92.347198	49.728553
std	85.608404	2.666615e+07	91.084068	8003.124202	3.982459	12.728658	6.657334	22.541345
min	0.000000	2.657637e+06	3.795632	5470.811536	85.522116	58.833327	75.747098	16.986489
25%	19.000000	6.455226e+06	38.857146	11833.434120	93.673354	72.225044	87.787643	27.719643
50%	64.000000	1.073896e+07	77.029671	15643.731450	96.483971	85.378852	94.258505	50.239961
75%	120.500000	3.789078e+07	101.742866	22700.898870	99.007799	92.943843	97.453477	67.644944
max	322.000000	1.003881e+08	360.017362	35948.191960	100.000000	98.639170	100.000001	80.554925

This group is average in terms of wealthiness and cleanliness.

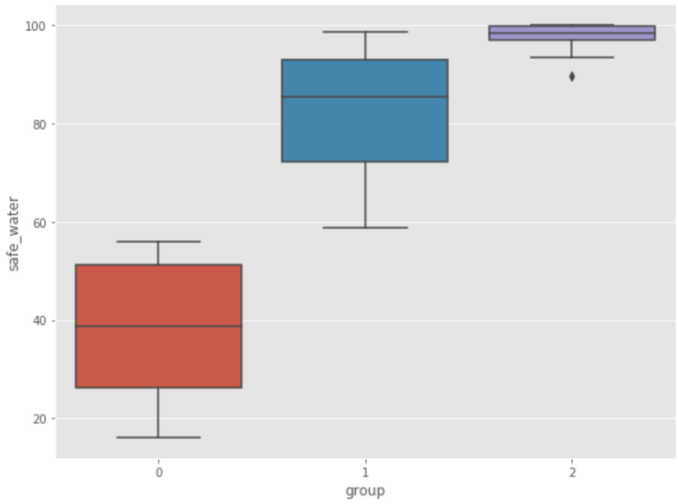
Group 3:
Countries: 'Italy', 'Canada', 'Austria', 'Czech Republic', 'Malaysia', 'New Zealand', 'United Kingdom', 'Slovakia', 'Kuwait', 'Poland', 'Switzerland', 'Greece', 'Finland', 'Portugal', 'Sweden', 'Norway', 'Germany', 'Hungary', 'Slovenia', 'Saudi Arabia', 'Australia', 'Ireland', 'France', 'United Arab Emirates', 'Spain', 'Denmark'

	index	pop_total	pop_density	GDP	basic_water	safe_water	basic_san	safe_san
count	26.000000	2.600000e+01	26.000000	26.000000	26.000000	23.000000	26.000000	26.000000
mean	71.461538	2.341429e+07	108.033904	49811.061718	99.605637	97.800511	98.929310	90.460826
std	69.135363	2.378461e+07	77.052477	13900.266243	0.859682	2.598203	1.691049	7.606282
min	1.000000	2.087946e+06	3.247871	29525.577360	96.695939	89.572762	91.245181	75.639872
25%	25.250000	5.594874e+06	36.399485	41045.950625	99.740445	97.028748	98.778477	83.457854
50%	54.000000	1.027744e+07	107.554727	49171.831160	99.999998	98.441489	99.255525	93.315917
75%	106.250000	3.675908e+07	137.145556	55992.944715	100.000000	99.793867	99.820011	96.535138
max	297.000000	8.313280e+07	274.708982	88240.901030	100.000005	100.000000	100.000000	100.000000

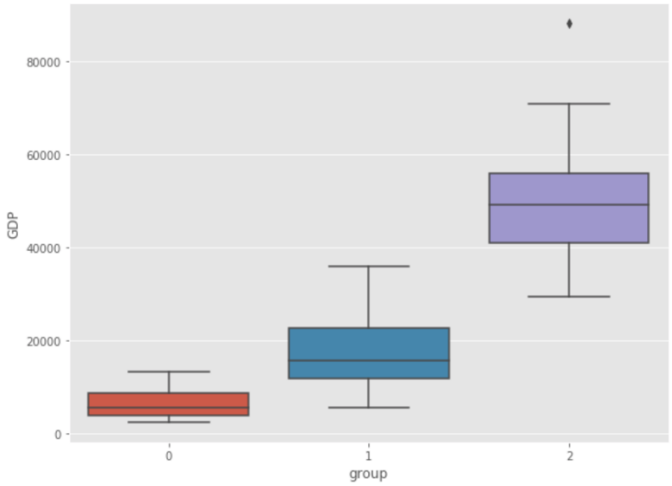
This group is the most wealthy (Highest GDP) and the most clean.

Selected Boxplots:

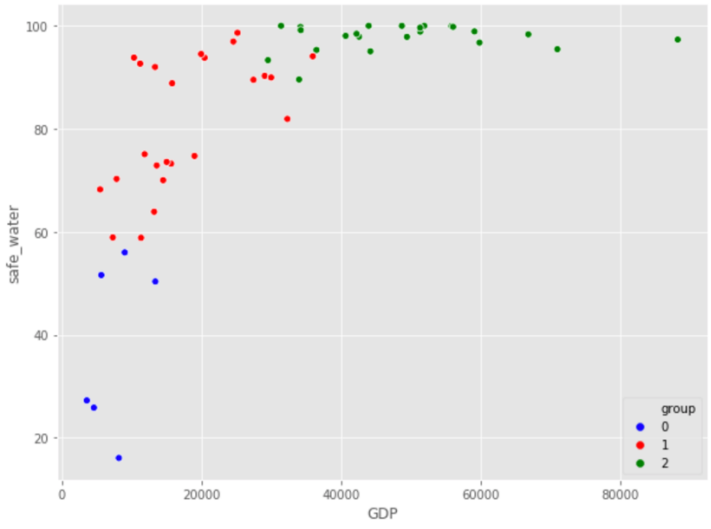
- safe_water



- GDP



(iii) scatter plot with two significant attributes

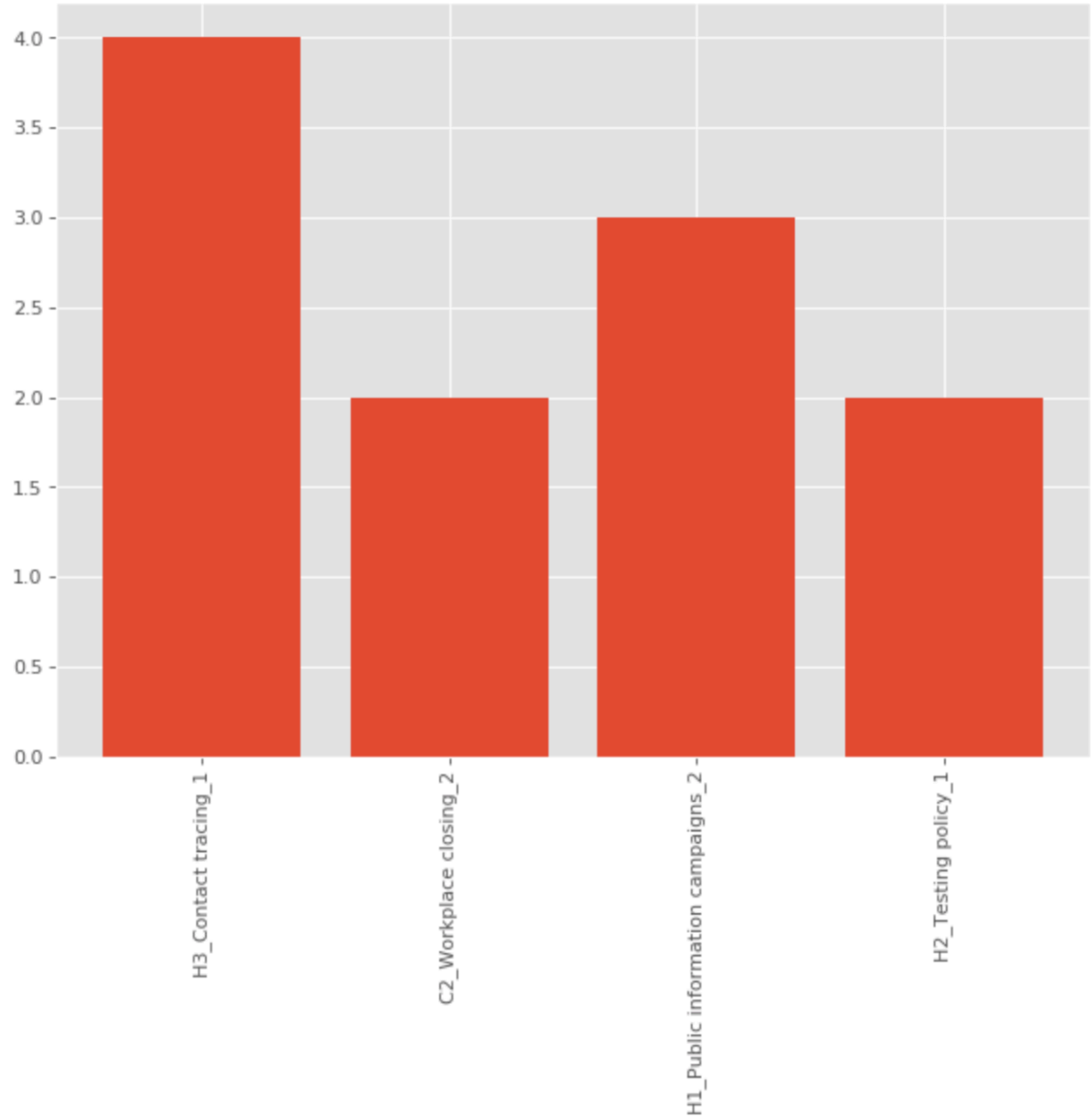


Task 2 — Policies Data:

Association Rule Mining:

Group 1:
Top 5 Association Rule:

	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction
703	(H3_Contact tracing_1, C2_Workplace closing_2)	(minimum_50)	0.351792	0.550489	0.228013	0.648148	1.177405	0.034356	1.277559
4912	(H1_Public information campaigns_2, H3_Contact...	(minimum_50)	0.351792	0.550489	0.228013	0.648148	1.177405	0.034356	1.277559
142	(H3_Contact tracing_1)	(minimum_50)	0.397394	0.550489	0.250814	0.631148	1.146522	0.032053	1.218675
1539	(H1_Public information campaigns_2, H3_Contact...	(minimum_50)	0.397394	0.550489	0.250814	0.631148	1.146522	0.032053	1.218675
141	(H2_Testing policy_1)	(minimum_50)	0.573290	0.550489	0.358306	0.625000	1.135355	0.042717	1.198697

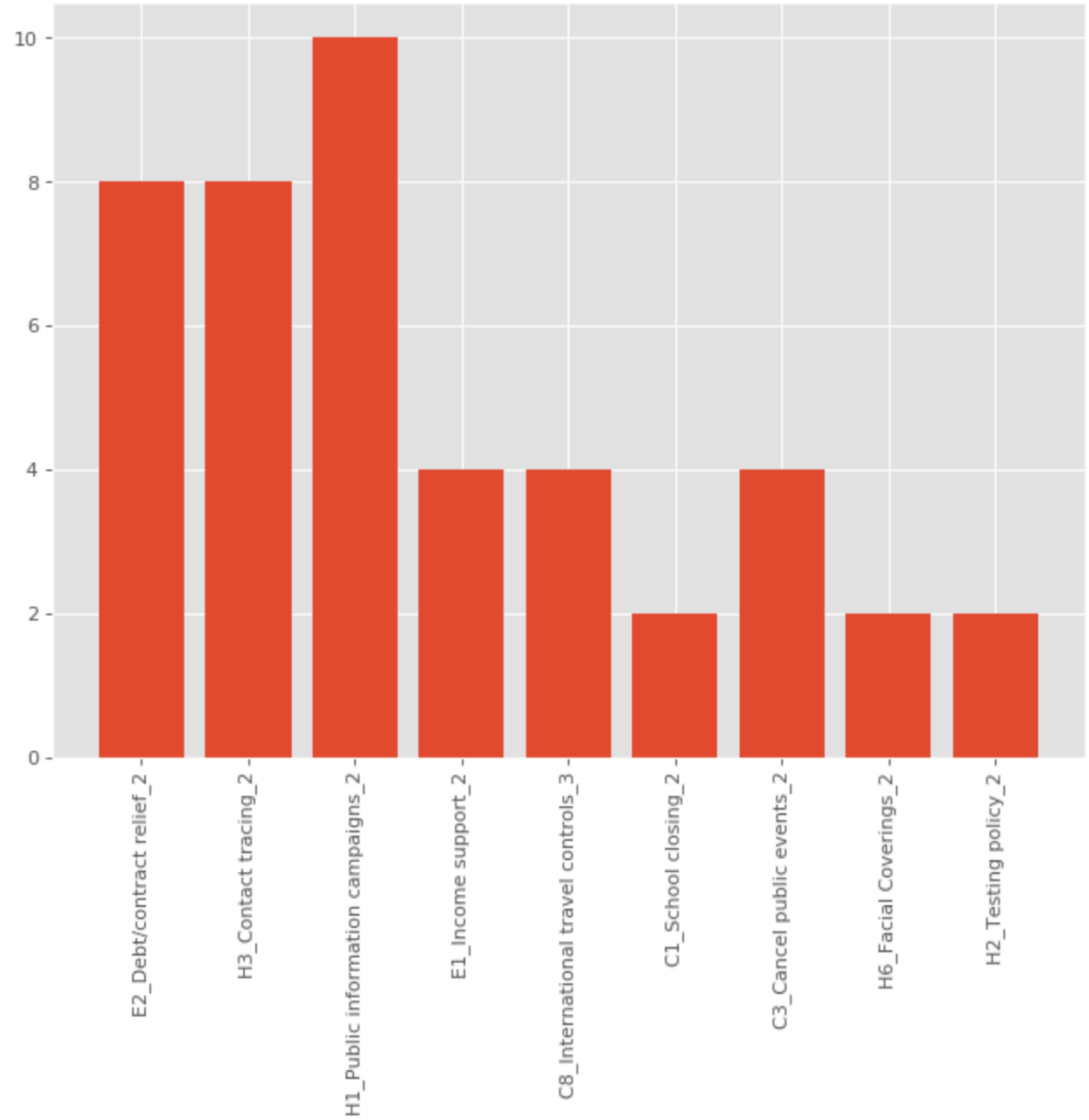


Group 2:

No Association Rule fit the criteria

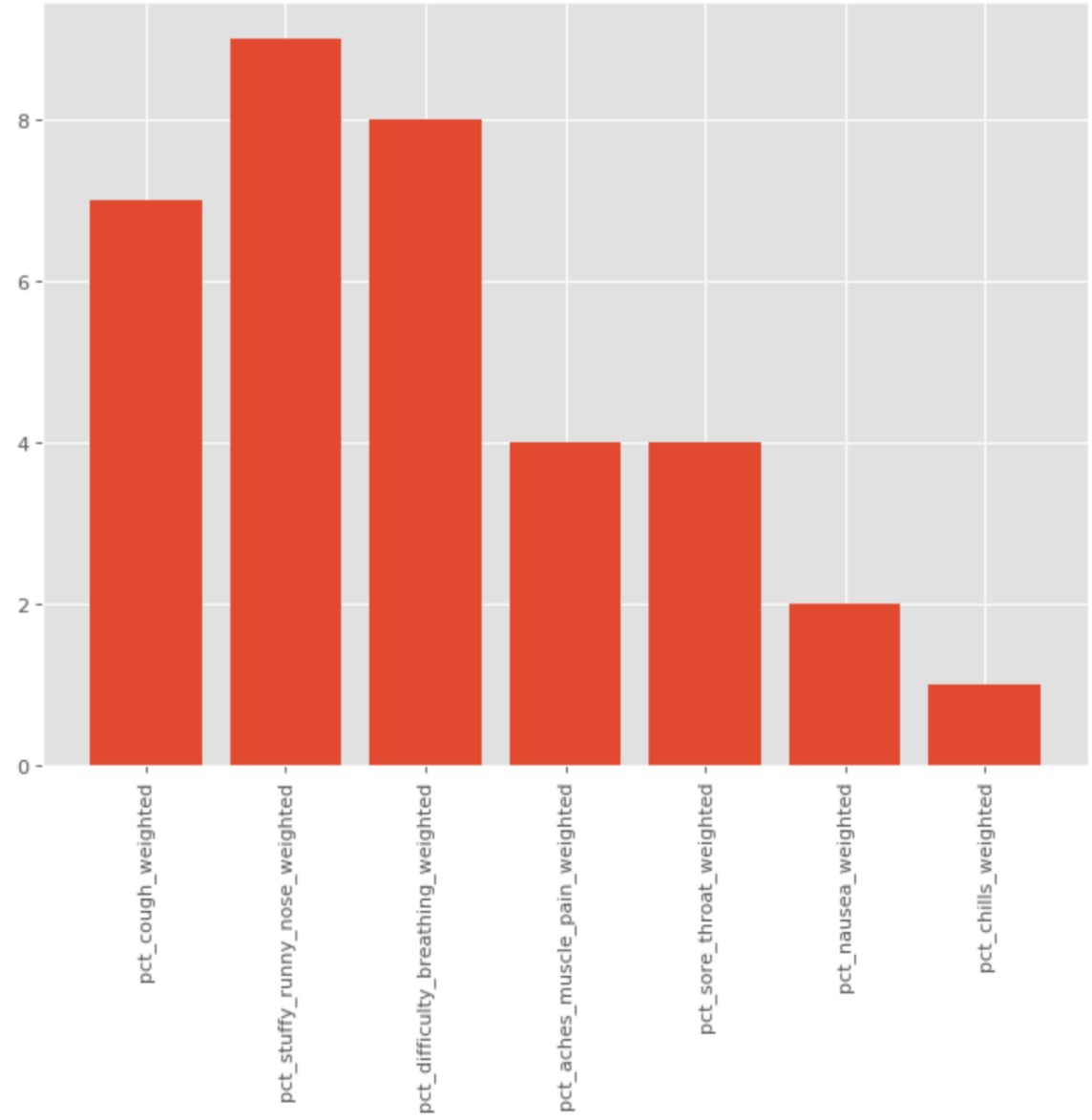
Group 3:

	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction
364	(E2_Debt/contract relief_2, H3_Contact tracing_2)	(minimum_50)	0.410569	0.541667	0.285569	0.695545	1.284082	0.063178	1.505420
580	(E2_Debt/contract relief_2, H3_Contact tracing...	(minimum_50)	0.410569	0.541667	0.285569	0.695545	1.284082	0.063178	1.505420
553	(E2_Debt/contract relief_2, H1_Public informat...	(minimum_50)	0.341463	0.541667	0.225610	0.660714	1.219780	0.040650	1.350877
322	(E2_Debt/contract relief_2, E1_Income support_2)	(minimum_50)	0.341463	0.541667	0.225610	0.660714	1.219780	0.040650	1.350877
544	(H3_Contact tracing_2, H1_Public information c...	(minimum_50)	0.394309	0.541667	0.259146	0.657216	1.213323	0.045562	1.337093



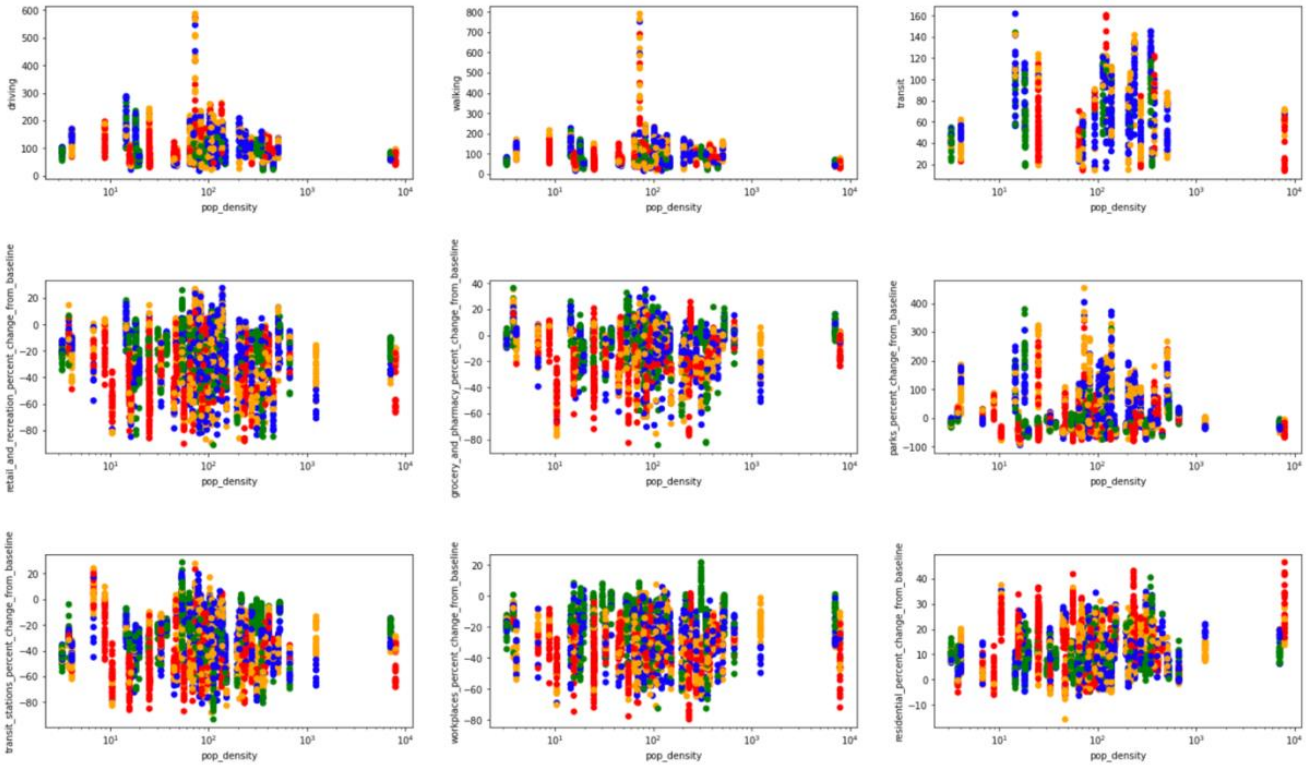
Task 3 — Symptoms Data

	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction
2921	(pct_cough_weighted, pct_stuffy_runny_nose_wei...	(total_cases_percentages)	0.307195	0.5	0.220238	0.716933	1.433867	0.066641	1.766369
625	(pct_stuffy_runny_nose_weighted, pct_difficult...	(total_cases_percentages)	0.355331	0.5	0.247930	0.697742	1.395484	0.070264	1.654217
3268	(pct_cough_weighted, pct_aches_muscle_pain_wei...	(total_cases_percentages)	0.289337	0.5	0.201605	0.696780	1.393560	0.056936	1.648968
3333	(pct_cough_weighted, pct_stuffy_runny_nose_wei...	(total_cases_percentages)	0.297101	0.5	0.205745	0.692509	1.385017	0.057195	1.626062
816	(pct_stuffy_runny_nose_weighted, pct_aches_mus...	(total_cases_percentages)	0.336698	0.5	0.231625	0.687932	1.375865	0.063276	1.602217



Task 4 — Symptoms Data

4a. Create scatter plot against log pop_density



4b. Find highest correlation attribute in each division

For group 0:

New_cases_percentages	1.000000
Workplaces_percent_change_from_baseline	0.423692
Retail_and_recreation_percent_change_from_baseline	0.409629
grocery_and_pharmacy_percent_change_from_baseline	0.335942
residential_percent_change_from_baseline	0.333803
transit_stations_percent_change_from_baseline	0.226935

For group 1:

New_cases_percentages	1.000000
Workplaces_percent_change_from_baseline	0.430647
Transit_stations_percent_change_from_baseline	0.340053
Retail_and_recreation_percent_change_from_baseline	0.296959
Grocery_and_pharmacy_percent_change_from_baseline	0.286697
Residential_percent_change_from_baseline	0.244358

For group 2:

New_cases_percentages	1.000000
Workplaces_percent_change_from_baseline	0.239121
Driving	0.095914
Residential_percent_change_from_baseline	0.080657
Parks_percent_change_from_baseline	0.077761
Transit_stations_percent_change_from_baseline	0.063155

For group 3:

New_cases_percentages	1.000000
Workplaces_percent_change_from_baseline	0.271165
Residential_percent_change_from_baseline	0.256093
Grocery_and_pharmacy_percent_change_from_baseline	0.247136
Retail_and_recreation_percent_change_from_baseline	0.209459
Transit_stations_percent_change_from_baseline	0.130293

For group 4:

New_cases_percentages	1.000000
Workplaces_percent_change_from_baseline	0.340643
Transit_stations_percent_change_from_baseline	0.333826
Transit	0.294173
Residential_percent_change_from_baseline	0.276341
Retail_and_recreation_percent_change_from_baseline	0.195812

Task 5 — Additional Analysis and Insights

For parts 5b and 5c, further analysis was performed by using apriori to generate frequent itemsets, then association rules function was used to find the rules and their confidence and lift ratio values.

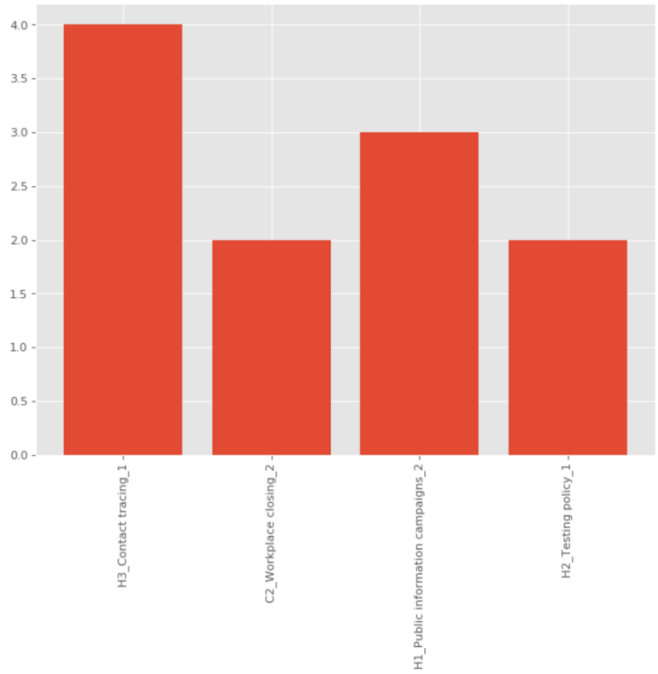
5a. Task 2 extension: seeking policy attributes related to higher new cases percentages

By using the generated policy data in task 2, frequent itemsets and related association rules were found for each cluster produced in task 1.

Threshold:

support: at least 20%
confidence: at least 60%

Cluster 1:
Task5_2_policy_data_cluster1 *(CSV OUTPUT)*

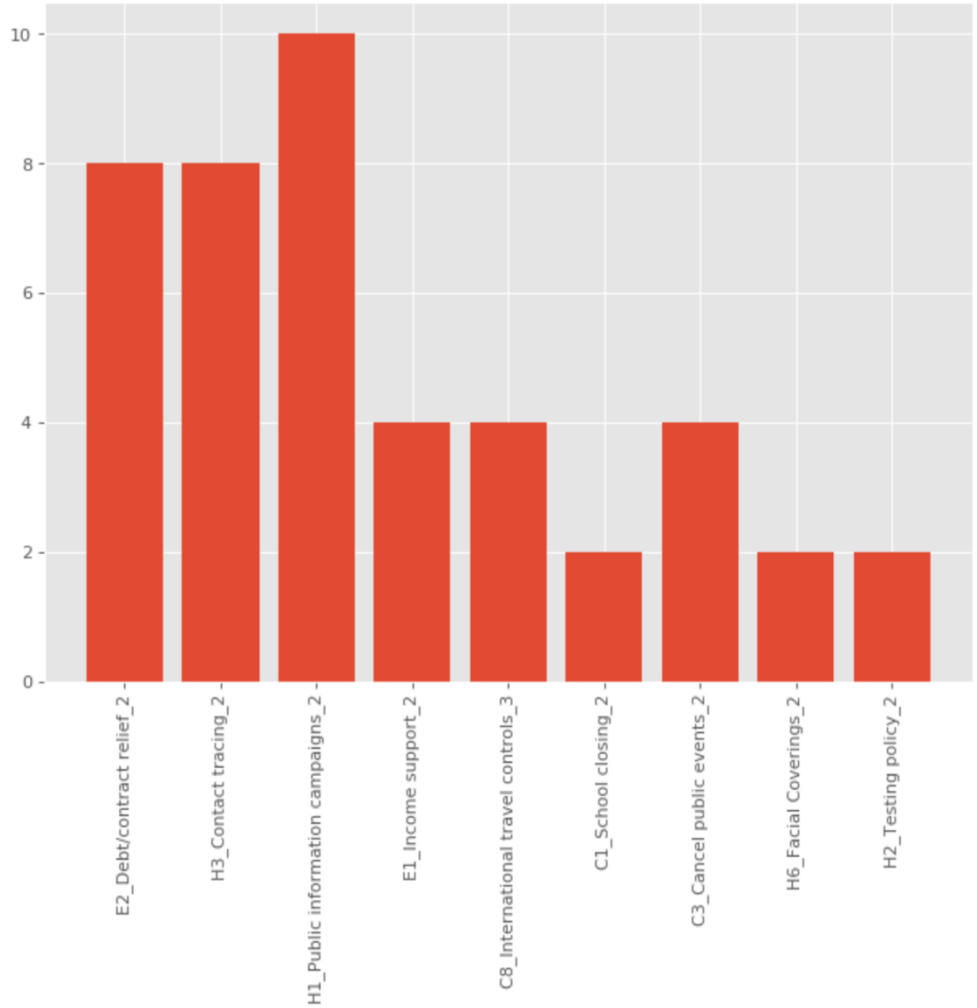


Description:
1. Contact tracing
2.Workplace closing
3. Public information campaigns
4. Testing policy

The presence of policies like work closing, public info, testing policy and contact tracing more likely implies lower new case percentages. However, there were few rules left after filtering by the threshold, hence the result was not significant.

Cluster 2:
No rules produced

Cluster 3:
Task5_2_policy_data_cluster3 *(CSV OUTPUT)*



Description:
1.School closing
2.Cancel public events
3. International travel controls
4. Income support
5. Debt/contract relief
6. Public information campaigns
7. Testing policy
8. Contact tracing
9. Facial Coverings

Compared with results in cluster 1, the policies like work closing and testing policy have no significant impacts on low new cases percentages, but public information campaigns and contact tracing still take effect to generate lower new cases percentages.

Summary:
As a result, we can conclude that contact tracing and public information campaigns are crucial attributes related to lower new cases percentages.

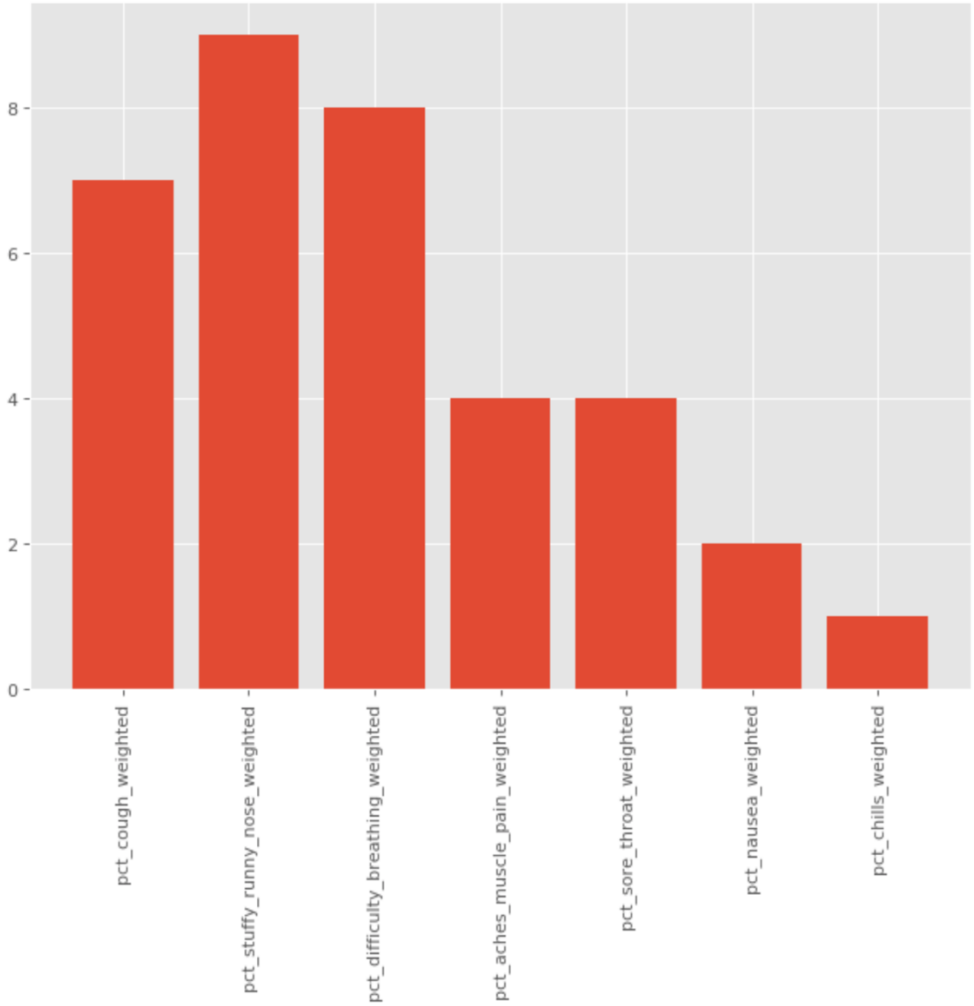
For the government, it's recommended to enhance the public information delivery about the pandemic, and distribute more resources on contact tracing .

5b. Task 3 extension: seeking symptom attributes related to higher total cases percentages

By using the generated symptom data in task 3, frequent itemsets and related association rules were found and filtered by threshold as follows:

support: at least 18%
confidence: at least 68%

Task5_3_symptom_data *(CSV OUTPUT)*



Description:

1. aches muscle pain
2. chills
3. cmnty sick
4. cough
5. difficulty breathing
6. ever tested
7. grocery outside home
8. nausea
9. sore throat
10. stuffy runny nose
11. tested recently
12. wear mask all time

Summary:

Population with people suffering from stuffy/runny nose, cough, breathing difficulty and ever tested or tested recently are more likely to have higher confirmed cases percentages.

It's recommended to take attention for people who have symptoms like runny nose, cough, breathing difficulty.

5c. Task 4 extension: Correlated feature sets

Group 4 from task 4 (highest 20% population density) will be targeted for analysis in this section, which contains Hong Kong

By using the multi-dimensional linear regression training model, we collect features by greedy selection to improve the R-squared value. We selected 4 features.

Finally, 4 attributes are found to be correlated to new cases percentages:
Residential_percent_change_from_baseline,
Workplaces_percent_change_from_baseline,
Transit_stations_percent_change_from_baseline,
Parks_percent_change_from_baseline

It was similar to the result of group 4 in part 4c:

For group 4: (number is correlation)

New_cases_percentages	1.000000
Workplaces_percent_change_from_baseline	0.340643
Transit_stations_percent_change_from_baseline	0.333826
Transit	0.294173
Residential_percent_change_from_baseline	0.276341
Retail_and_recreation_percent_change_from_baseline	0.195812

The selected attributes are a little different due to some features being correlated, and we excluded the features with less marginal benefit.

Summary:
It can be concluded that activities in workplaces, residential, transit stations are positively correlated to new cases percentages.

For group 4 -- countries with highest density, it is recommended not to have gatherings inside residential areas, stay short in transit stations, avoid exposures in workplaces like offices to minimize the chance of infection.

For the Hong Kong government, it's recommended to improve the contact tracing measures and distribute more resources on information campaigns.
Besides, the government should encourage citizens with symptoms like runny nose, cough, breathing difficulty to take coronavirus tests.

Task 6 — Model Prediction

Preprocess

Decision Tree Model

Model: Random Forest Model

Algorithm Description / Justification: Since decision trees have features locality and large variance, we can apply multiple random tree classifiers and select the prediction supported by the most classifiers, so the noise is cancelled out. Since the algorithm is based on independent random sampling, increasing the number of samples will not lead to a worse result, i.e. it does not overfit by increasing estimators.

Grid search:

```
param_grid = {
     'n_estimators': [300, 500, 1000, 2000, 3000],
     'max_depth': np.arange(5, 12, 2),
     'min_samples_split' : [2, 5, 10],
     'min_samples_leaf' : [1, 2, 4],
    }

forest = RandomForestClassifier(random_state = 10, n_jobs = -1)
forest_cv = GridSearchCV(forest, param_grid, cv = 3, verbose = True)
forest_cv.fit(X_train,y_train)
print(f'best paramter: {forest_cv.best_params_}')
print(f'score: {forest_cv.best_score_}')
```

```
Fitting 3 folds for each of 180 candidates, totalling 540 fits
```

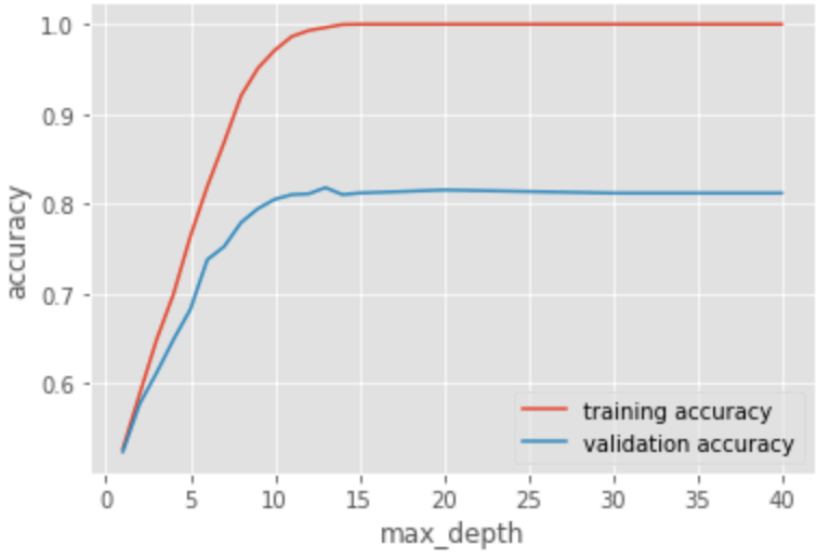
```
[Parallel(n_jobs=1)]: Using backend SequentialBackend with 1 concurrent workers.
[Parallel(n_jobs=1)]: Done 540 out of 540 | elapsed: 31.4min finished
```

```
best paramter: {'max_depth': 11, 'min_samples_leaf': 1, 'min_samples_split': 2, 'n_estimators': 3000}
score: 0.7758862822207074
```

To perform the grid search, we first set some initial parameters so as to lower the training time of the model. We set ‘bootstrap’ = ‘False’, ‘criterion’ = ‘entropy’ and ‘max_features’ = ’auto’. Then we start tuning the model. In order to tune the model, we use grid search with 3 fold cross-validation to obtain the optimal value of the hyperparameters. We will build the model bases on the best parameters calculated on grid search. As we can see, we set ‘max_depth’ = 11, min_samples_leaf’ = 1, ‘min_samples_split’ = 2, and finally the ‘n_estimators’ = 3000. Based on these parameters, the validation set of the model is 77.59%.

Fine tuning:

Max_depth is an important parameter to prevent the model overfitting the training data. Therefore, we would fine tune the model to get the best number for max_depth.



In the project, we use a for loop to set the max_depth in a list. Then, we create a random forest model using the parameter as mentioned above and set the max_depth equal to [1,2,3,4,5,6,7,8,9,10,11,12,13,14,15,20,30,40]. Therefore, the accuracy of the model with different max_depth is plotted above. We found that the accuracy of training and validation set will not change after max_depth = 15. When the max_depth = 10, the validation accuracy can achieve about 80% and the training accuracy is good enough, not exactly equal to 1. This means that the training set is not overfitted.

Adaptable parameters:

bootstrap = False, criterion = entropy and max_features = auto
Max_depth = 11
Min_samples_leaf = 1
Min_samples_split = 2
Number of estimators: 3000

Evaluation:

We further evaluate the model using classification report, as shown below:

	precision	recall	f1-score	support
0	0.85	0.91	0.88	278
1	0.75	0.77	0.76	291
2	0.80	0.70	0.74	315
3	0.88	0.91	0.89	276
accuracy			0.82	1160
macro avg	0.82	0.82	0.82	1160
weighted avg	0.82	0.82	0.82	1160

From the table above, it contains precision, which measures how many of the samples predicted as positive are actually positive; recall, which measures how many of the positive samples are captured by the positive predictions. From the table above, we can see that the precision and recall of the three classes ‘0’, ’1’, ‘2’ and ‘3’ are good enough, about 0.7 to 0.9. This shows that the model does not produce many false positives and avoid any false negatives. From the table, the f1-score column takes the precision and recall into account. The f1-score of class ‘0’ and ‘3’ is 0.88 and 0.89 respectively, while that of class ‘1’and ‘2’ is 0.76 and 0.74. This also shows that the predictive performance of the model is good enough.

Final model:
We use the parameters above to train the model. The performance of the model is as follow:
Training accuracy: 99.7%
Validation accuracy: 81.9%