Covid-19 Impact

Group 5

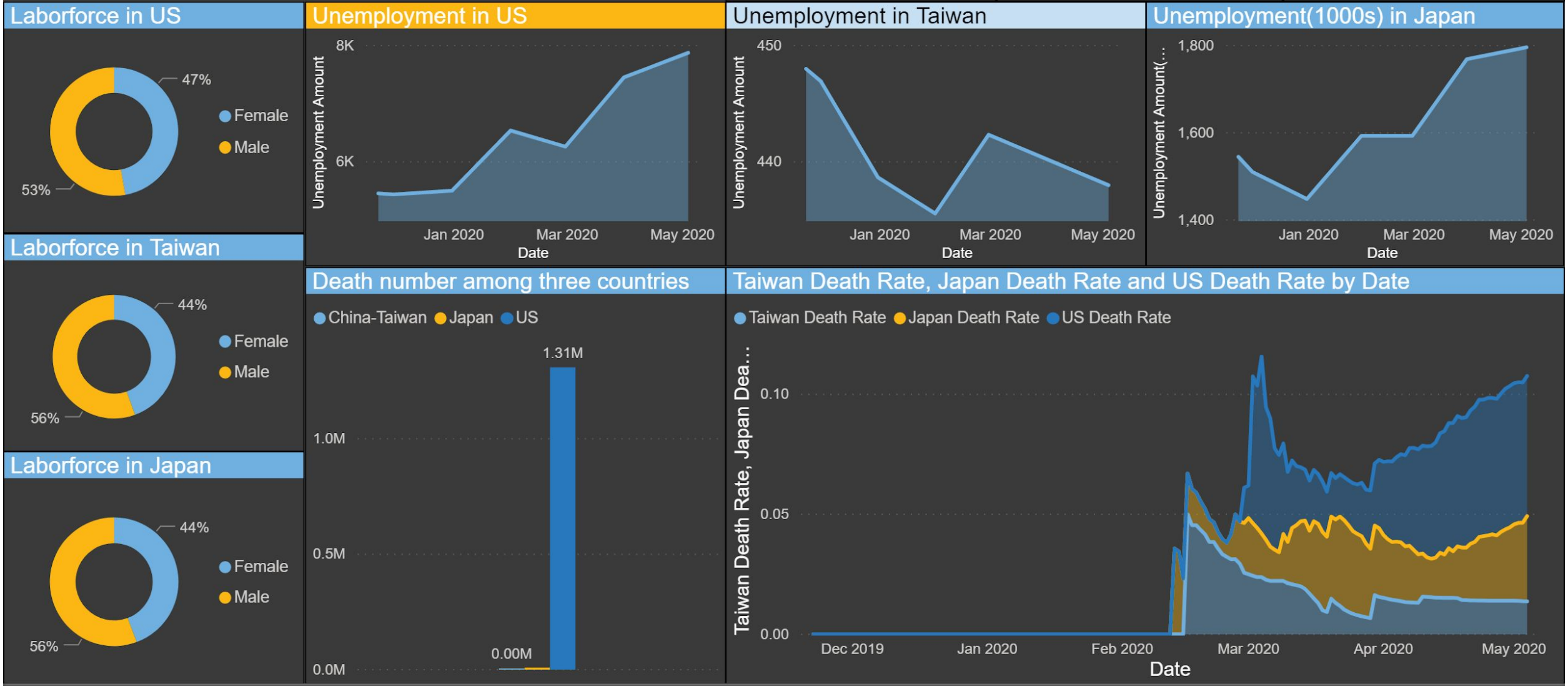
CONTENTS

INTRODUCTION	Introduction of Covid-19 situation
DATA SELECTION	Select five features to generate clusters and combine with COVID-19 data to do the prediction
CLUSTERING	Make comparison between HCA and K-Means model
PREDICTION	Multi-regression to predict unemployed number

Covid-19 Data Analysis

Data collected from US, China-Taiwan and Japan from November 2019 to May 2020

US Avg. Death Rate	Taiwan Avg. Death Rate	Japan Avg. Death Rate
5.20%	1.42%	2.59%



Data Selection

Procedure

Cluster

- Select GDP per capita, income per capita, industry percent, inflation rate and life expectancy as five main features to cluster countries

Prediction

- Select U.S, Taiwan and Japan as three main countries to predict the unemployment based on confirmed cases and death number

Data Source: John Hopkins https://github.com/CSSEGISandData/COVID-19
International Labour Organization https://ilostat.ilo.org/), https://covid19.healthdata.org/united-states-of-america
MIT prediction model https://www.covidanalytics.io/projections
Gapminder https://www.gapminder.org/data/

Cluster – Data Preprocessing

gdp_per_capita_yearly_growth

income_per_person

industry_percent_of_gdp

inflation_annual_percent

life_expectancy_years

Merge Files

	country	gdp_per_cap	...	inflation	life_exp
0	Afghanistan	3.02	...	0.792	63.7
1	Albania	5.03	...	0.948	78.3
2	Algeria	2.63	...	7.560	77.9
3	Andorra	NaN	...	0.896	NaN
4	Angola	3.46	...	34.800	64.6
..	...	...	...	...	...
180	Venezuela	-0.56	...	NaN	75.2
181	Vietnam	4.90	...	3.400	74.6
182	Yemen	1.28	...	47.200	68.1
183	Zambia	2.89	...	9.330	63.7
184	Zimbabwe	2.87	...	28.000	61.7

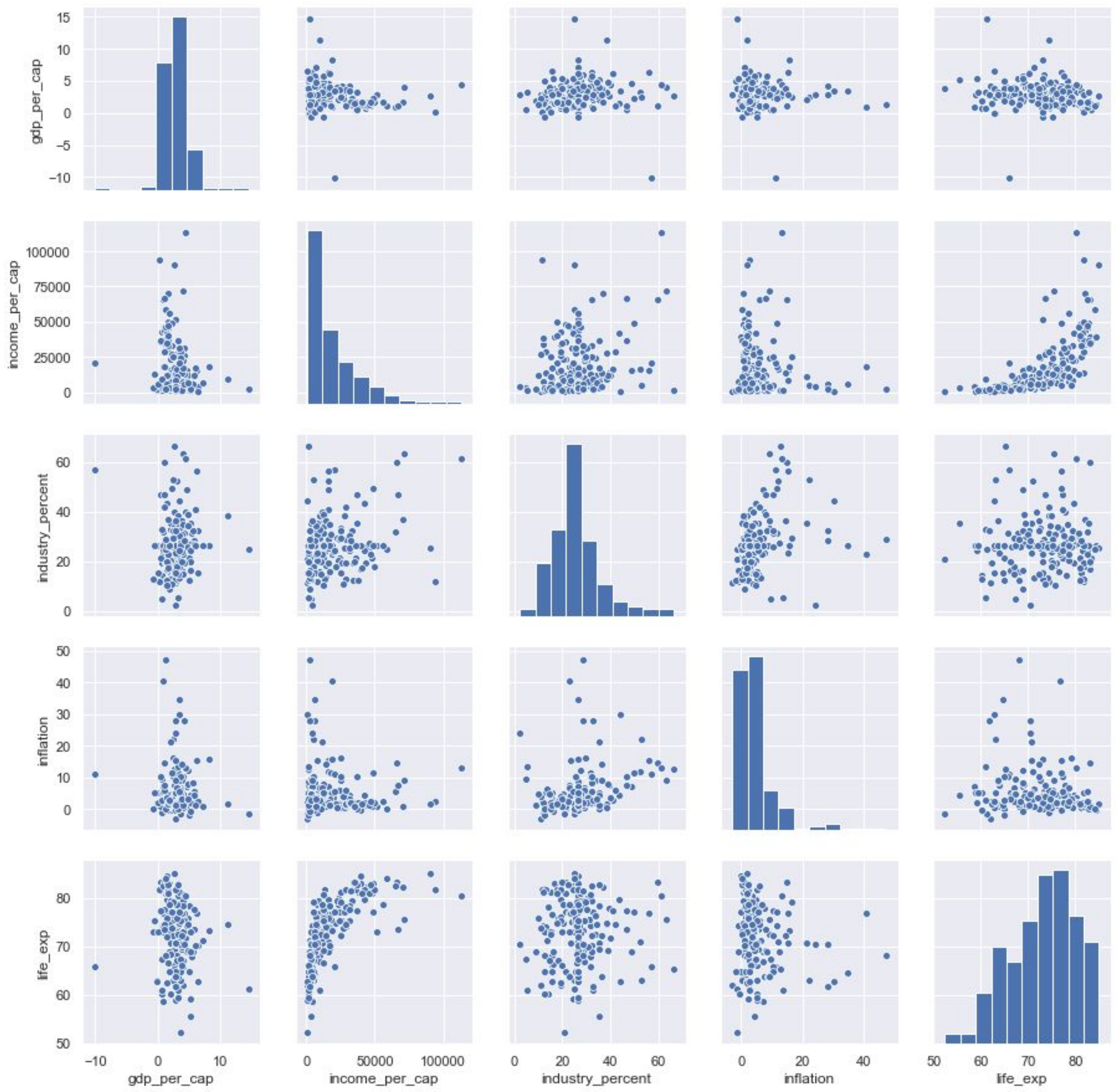
```python
# merge all colums into one dataset
merged_inner1 = pd.merge(gdp_df, income_df, on='country')
merged_inner2 = pd.merge(merged_inner1, industry_percent_df, on='country')
merged_inner3 = pd.merge(merged_inner2, inflation_df, on='country')
merged_total= pd.merge(merged_inner3, life_exp_df, on='country')
```

Replace Missing Value

	gdp_per_cap	income_per_cap	...	inflation	life_exp
country			...		
Afghanistan	3.020000	1740.0	...	0.792000	63.700000
Albania	5.030000	12300.0	...	0.948000	78.300000
Algeria	2.630000	13900.0	...	7.560000	77.900000
Andorra	2.913517	51500.0	...	0.896000	72.996703
Angola	3.460000	5730.0	...	34.800000	64.600000
...	...	...	...	...	...
Venezuela	-0.560000	12500.0	...	5.288489	75.200000
Vietnam	4.900000	6610.0	...	3.400000	74.600000
Yemen	1.280000	2360.0	...	47.200000	68.100000
Zambia	2.890000	3740.0	...	9.330000	63.700000
Zimbabwe	2.870000	2620.0	...	28.000000	61.700000

```python
from sklearn.preprocessing import Imputer
# replce nan value to mean
imputer = Imputer(missing_values=np.nan, strategy='mean')
```

Cluster – Pairplot



Show relationships between variables

Cluster - Dimension Reduction

2D

```
n_clusters = 2, silhouette score 0.390532
n_clusters = 3, silhouette score 0.407605
n_clusters = 4, silhouette score 0.395579
n_clusters = 5, silhouette score 0.417256
n_clusters = 6, silhouette score 0.373466
n_clusters = 7, silhouette score 0.369369
n_clusters = 8, silhouette score 0.377828
```

3D

```
n_clusters = 2, silhouette score 0.324499
n_clusters = 3, silhouette score 0.311090
n_clusters = 4, silhouette score 0.330543
n_clusters = 5, silhouette score 0.314995
n_clusters = 6, silhouette score 0.346927
n_clusters = 7, silhouette score 0.342908
n_clusters = 8, silhouette score 0.362966
```

4D

```
n_clusters = 2, silhouette score 0.325129
n_clusters = 3, silhouette score 0.290669
n_clusters = 4, silhouette score 0.333702
n_clusters = 5, silhouette score 0.351264
n_clusters = 6, silhouette score 0.336181
n_clusters = 7, silhouette score 0.304214
n_clusters = 8, silhouette score 0.304703
```

After two, three and four dimensional comparison, we can see that the model performs best when n = 5 and d = 2

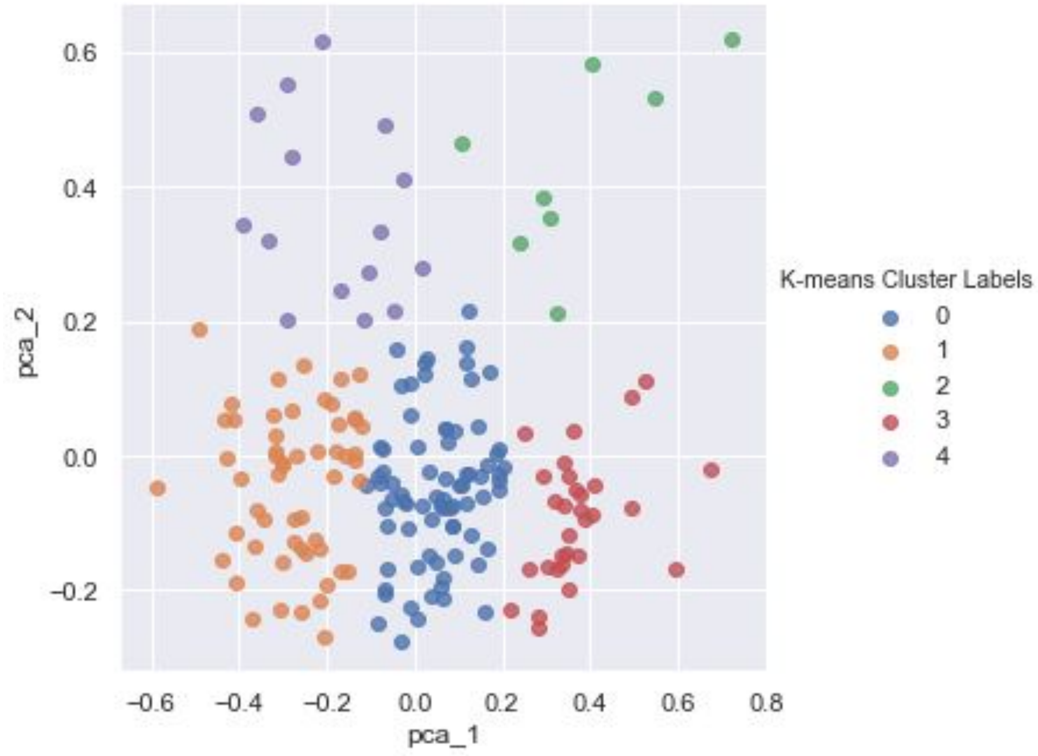
Cluster - Model Comparison

HCA		K-Means	
n clusters	silhouette score	n clusters	silhouette score
2	0.324499	2	0.390532
3	0.311090	3	0.407605
4	0.330543	4	0.395579
5	0.314995	**5**	**0.417256**
6	0.346927	6	0.373466
7	0.342908	7	0.369369
8	0.362966	8	0.377828

To determine the best k values, 2-8 clusters are tested, we can see from the table when model is K-Means and n = 5 the result is the best, thus I choose K-Means model and divide countries in 5 clusters

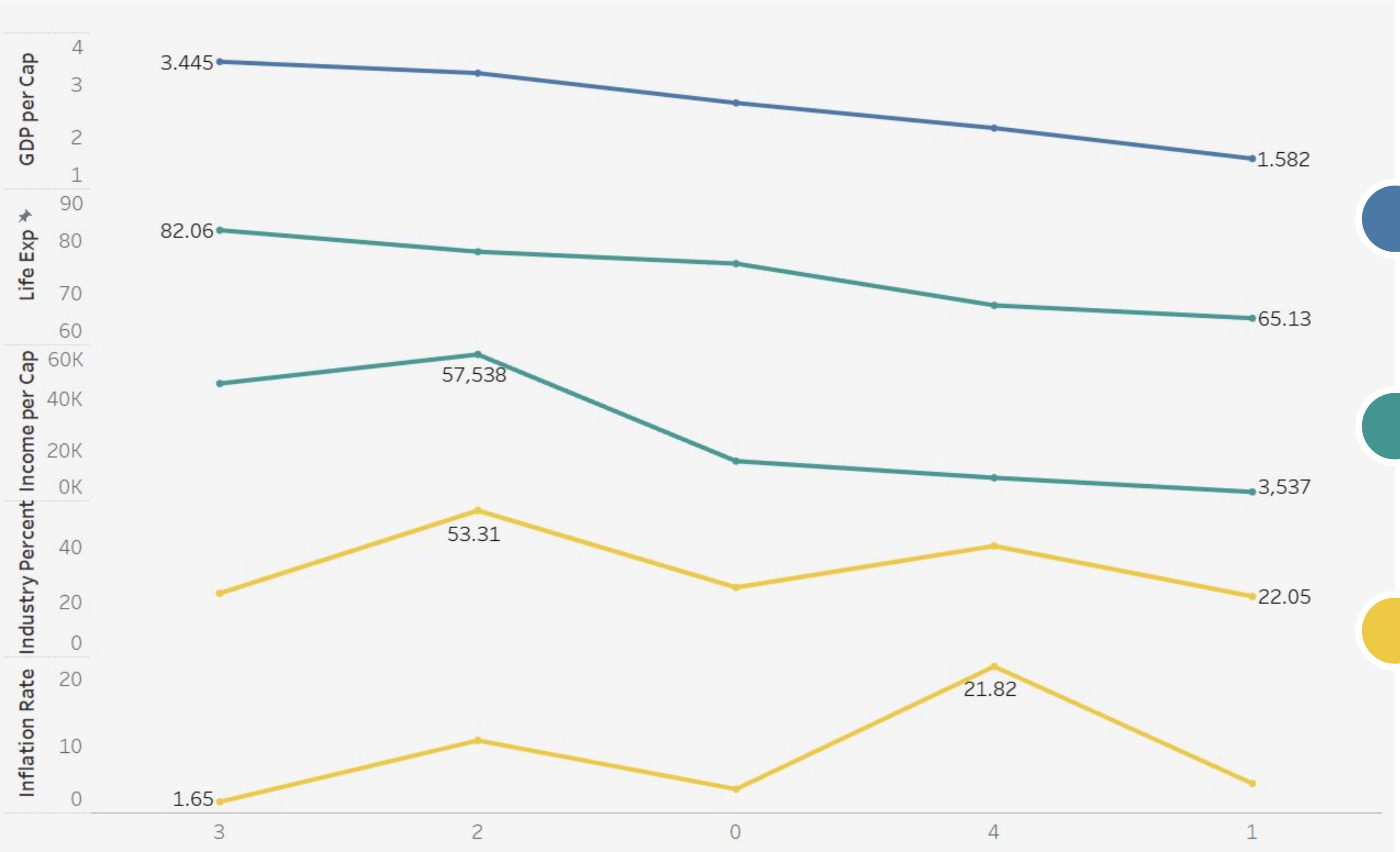
Cluster – Result

```python
# tol: stop when the distance between centers of two adjacent clusters is
# less than 0.0001
# max_iter: stop when iteration reach to 500
clus_kmeans = KMeans(n_clusters=5, tol=0.0001, max_iter=500)
# fit to input data
kmeans =clus_kmeans.fit(X_pca)

# get cluster assignments of input data and print first ten results
df_country['K-means Cluster Labels'] = kmeans.labels_
print(df_country[:10])

# visualize the group of countries set with the cluster labels displayed
X_pca_frame['K-means Cluster Labels'] = kmeans.labels_
sns.lmplot(x='pca_1', y='pca_2',
           hue="K-means Cluster Labels", data=X_pca_frame, fit_reg=False)

# save the results in csv file
df_country.to_csv(r'kmeans_cluster.csv', index = False)
```



- Above is the code snippet used to build cluster model
- Finally 5 clusters are established
- Scatter Plots to show cluster results

Cluster – Result



Cluster 3

Developed Country

Cluster 0 and 2

Developing Country

Cluster 1 and 4

Undeveloped Country

Prediction - Overview

Random Forest

- Higher accuracy
- Better at handling missing values while maintaining accuracy
- Low bias due to Bagging & Ensembling

Linear Regression

- Helpful to use if direct linear relationship between Covid cases & Unemployment exists

Prediction - Liberties with the data

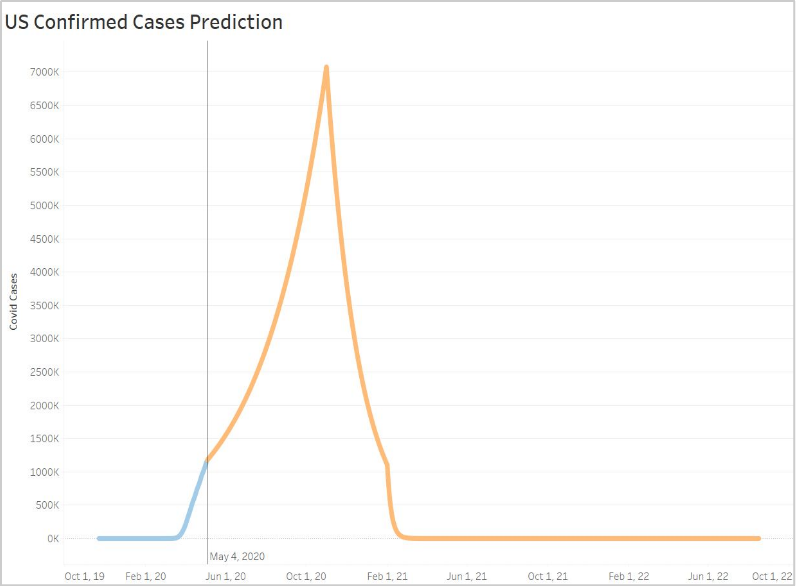
Covid cases in the future cannot be predicted with any real accuracy.

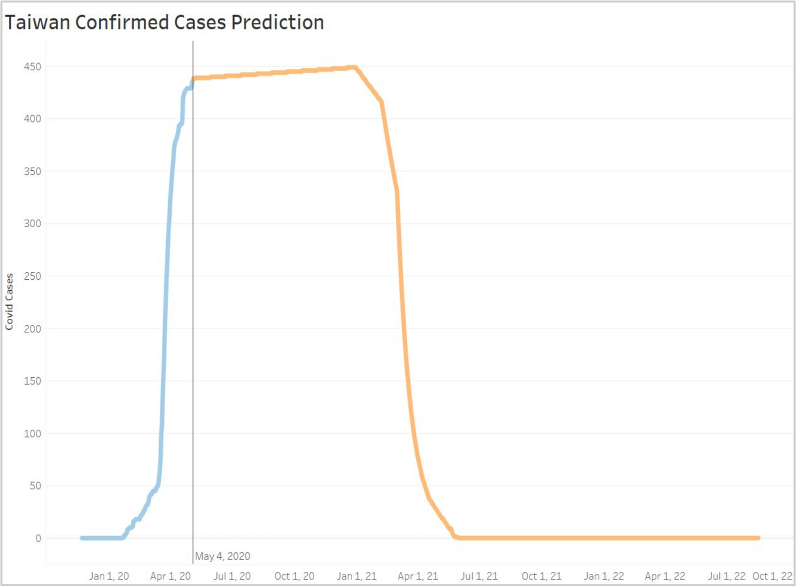
Too many variables to consider such as:

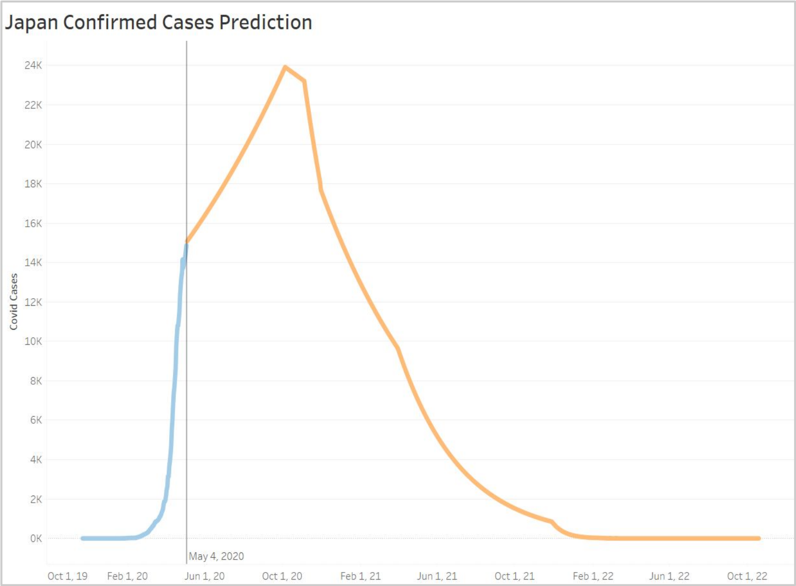
- Economic factors
- Political Factors
- Medical research progress
- Many more

Covid-19 case for each country were manually predicted based on information available and educated guesses

Confirmed Cases Prediction



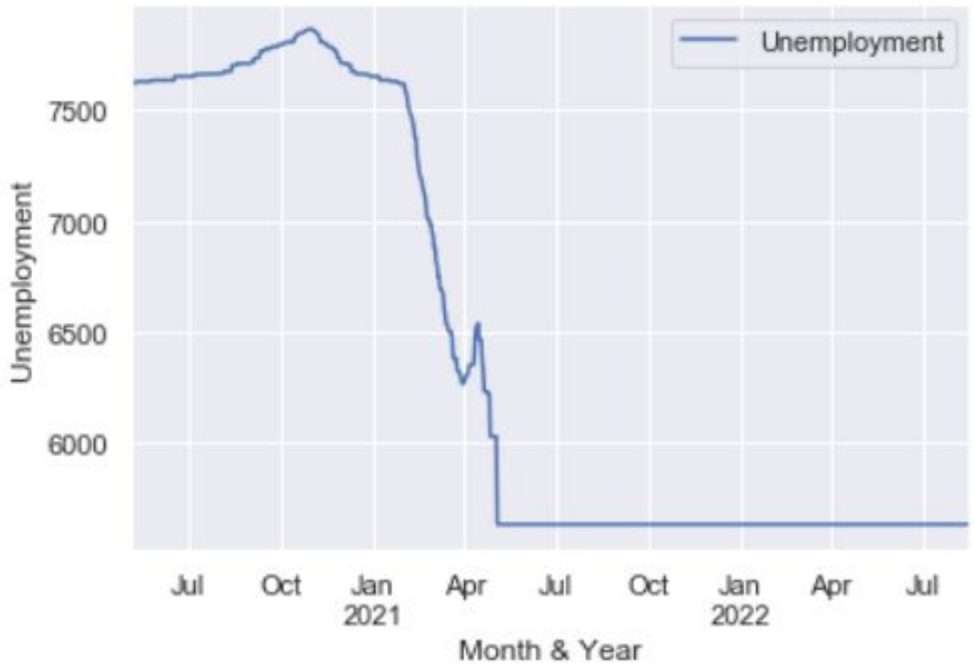


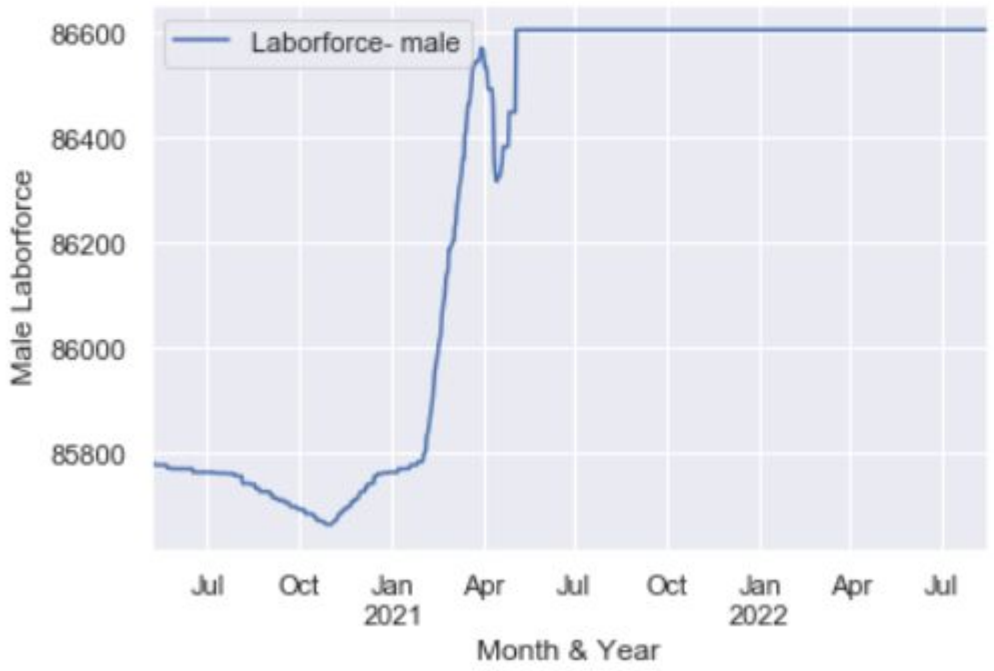


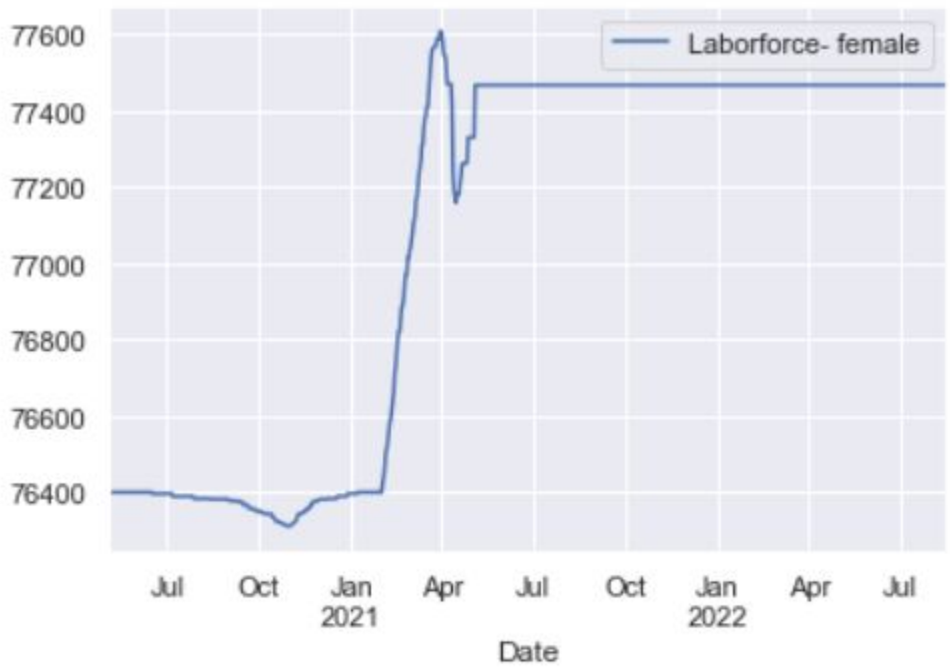
Prediction - Results

Model/Country	US	Japan	Taiwan
RF F1 Score	96%	87%	72%
Linear Regression R Squared	75%	22%	13%

Unemployment Prediction, US



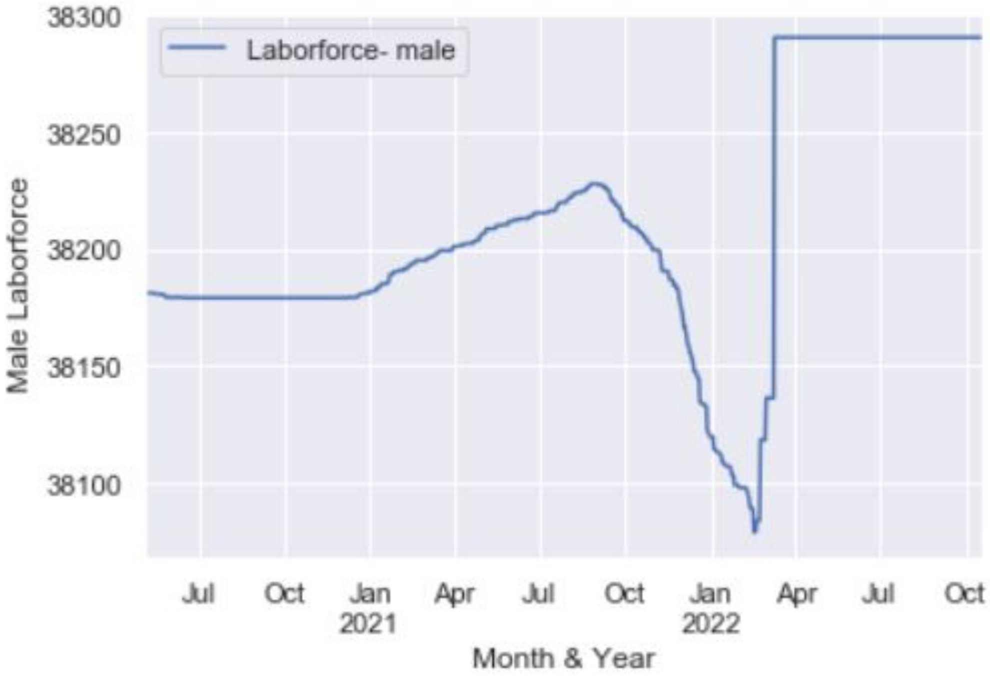




Random Forest F1 Score = 96%

Linear Regression R squared = 75%

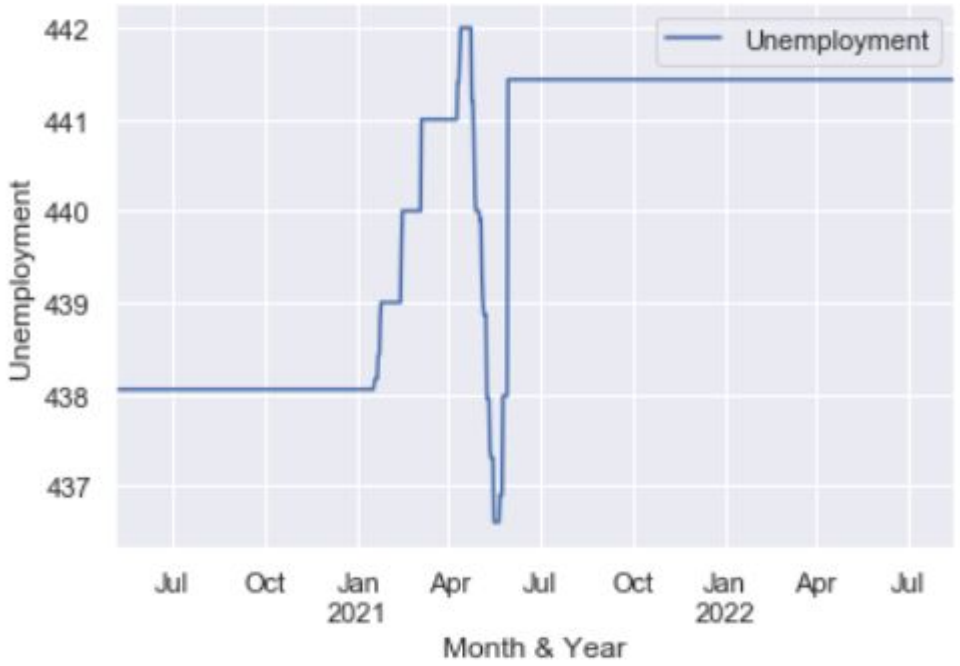
Unemployment Prediction, Japan

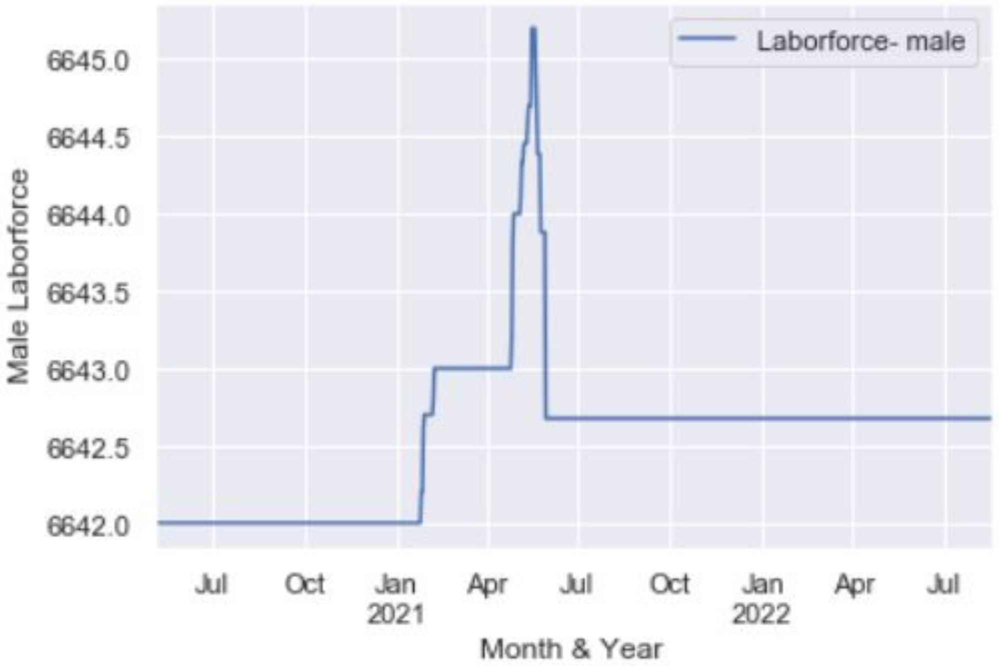


Random Forest F1 Score = 87%

Linear Regression R squared = 22%

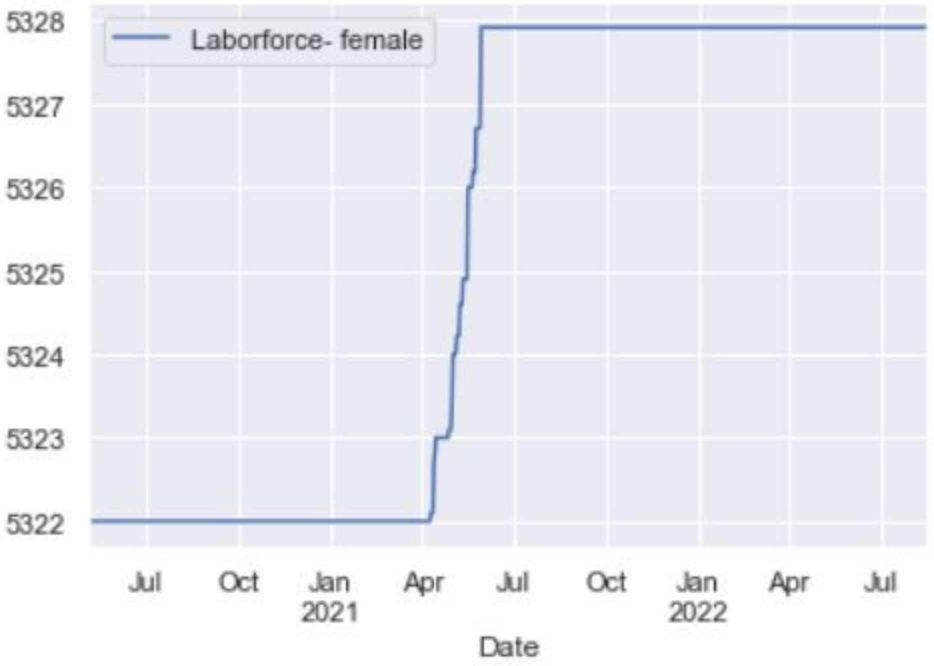
Unemployment Prediction, Taiwan





Random Forest F1 Score = 72%

Linear Regression R squared = 13%



THANKS

Group 5

Menghe Dou Parashara, Praharsh Tianyu Wei Ahmed, Awadh Srivastava, Ullas