Analysis and Predictive Modeling of COVID-19 and its Effects

Aman Aryan
20111009

Deeksha Arora
20111017

Shruti Sharma
20111061

Tamal Deep Maity
20111068

Disclaimer

The work carried out in this project has not been re-used from any other course project at IITK or elsewhere. It is a purely novel idea carried out as a course project for CS771 under the excellent guidance of Dr. Piyush Rai. All the references are duly acknowledged.

1 Introduction

COVID-19 is an ailment caused by the novel Coronavirus which is biologically known as Severe Acute Respiratory Syndrome Corona Virus-2 (SARS-CoV-2, formerly called 2019-nCoV). It was first identified amid a widespread outbreak of respiratory illness cases in Wuhan City, China. Wuhan was the epicenter of the spread of the virus.

Coronavirus-2019 outbreak was reported by World Health Organization (WHO) on 31st December, 2019 and has spread around the globe since months now and has caused serious health, social and financial issues. Regions across the world applied restrictive measures such as closing public places, schools, universities, quarantining individuals etc to contain the spread of COVID-19. The economy of major countries went downhill.

```
Total confirmed cases till 12/8/20 are 68225313
Total deaths till 12/8/20 are 1556822
Therefore, mortality rate is 2.281883265233243
```

2 Problem Statement and Motivation

With each passing day, the effect of COVID-19 is getting more and more severe. There are continuous strains and ill-effects of COVID on healthcare status and social life of people. It has recently been reported to infect a few animal species as well. Therefore, it becomes essential to deploy technology to model its future trend and be geared up to control the pandemic.

Tracking the active corona cases as they bulge up helps governments and common people be cautious about their future steps. They may then decide and form policies for better control over the virus. It will lead to less havoc and lower fatality rate.

In this project, we aim to develop a probabilistic predictor for confirmed coronavirus cases in the future 1 to 30 day window. Our objective is to predict the GDP of a particular region and correlate it to the predicted COVID-19 cases and other indirect features on which GDP depends. Some of them

are percentage of unemployed people, age group of population, mobile cellular subscriptions, which indicates expenditure of population.

3 Literature Review

There has been profound research and studies on predicting the outbreak of epidemics. Different kinds of prediction algorithms are used for different type of available data.

The SIR (Susceptible, Infected, Recovered) family of models are among the most popular ones to learn the dynamics of COVID-19 like epidemics. The SIRD (D = Death) model is an extension of the SIR model, and is used to demonstrate the differences of infectious rate at different countries. Fuzzy based model are used in time-variant systems.

There have been Markov switching models in literature to model the spread of diseases. Neural Networks have been used to predict HIV infections. Markov chain Monte Carlo methods and other kinds of Bayes filters like Particle filters and Kalman filters have also been used.

4 Novelty of this work

This projects aims to establish a connection between outbreak of COVID-19 in a country and the extent of adverse effects it had on the financial, social and personal state of that country's residents. To begin with, this work correlates variance of Gross Domestic Product and trade in each financial quarter as a function of pandemic's outbreak over that period. We propose a relation between both and a potential impact of the latter on the former.

It models the spread of virus as decrease in capital and expenditure. This leads to reduced demand in the markets, a part of which is induced by imposition of countrywide lockdowns.

We aim to predict the GDP of a country in Quarter 3 of 2020 when there were enough coronavirus infections and it had created its strong foothold over the globe. Through these projections, we classify whether a country will face growth or recession in the future financial quarters.

5 Dataset

- All reported experiments are based on online COVID-19 data repository maintained by John Hopkins University - `https://github.com/CSSEGISandData/COVID-19` - for time-period of 22 January 2020 to 8 December 2020.
- Dataset for GDP Projection: [1]
  - Gross Domestic Product(Real Index) - IMF Data
  - Total Population (2019) (Worldbank)
  - Unemployment ratio (Worldbank)
  - Unemployment ratio [FOR INDIA]
  - Mobile cellular subscriptions (Worldbank)
  - Population age 65 and above data (Worldbank)

6 Methodology

- The dataset is filtered and structured into time-series format to decipher the trends.
- An adaptive online Kalman filter provides us estimates of 1-30 day prediction for each region.

[1] Links for dataset in Section 11

- The results of this predictor are used in projecting the GDP of a region.
- A variety of ML models like SVR, Linear Regressor, Ridge Regressor, Random Forest Regressor, Multi Layer Perceptron regressor, Stochastic Gradient Descent are trained using the predictions of COVID cases and other influencing parameters. Random Forest Regressor was reported to have the least mean squared error out of all the models as seen below.

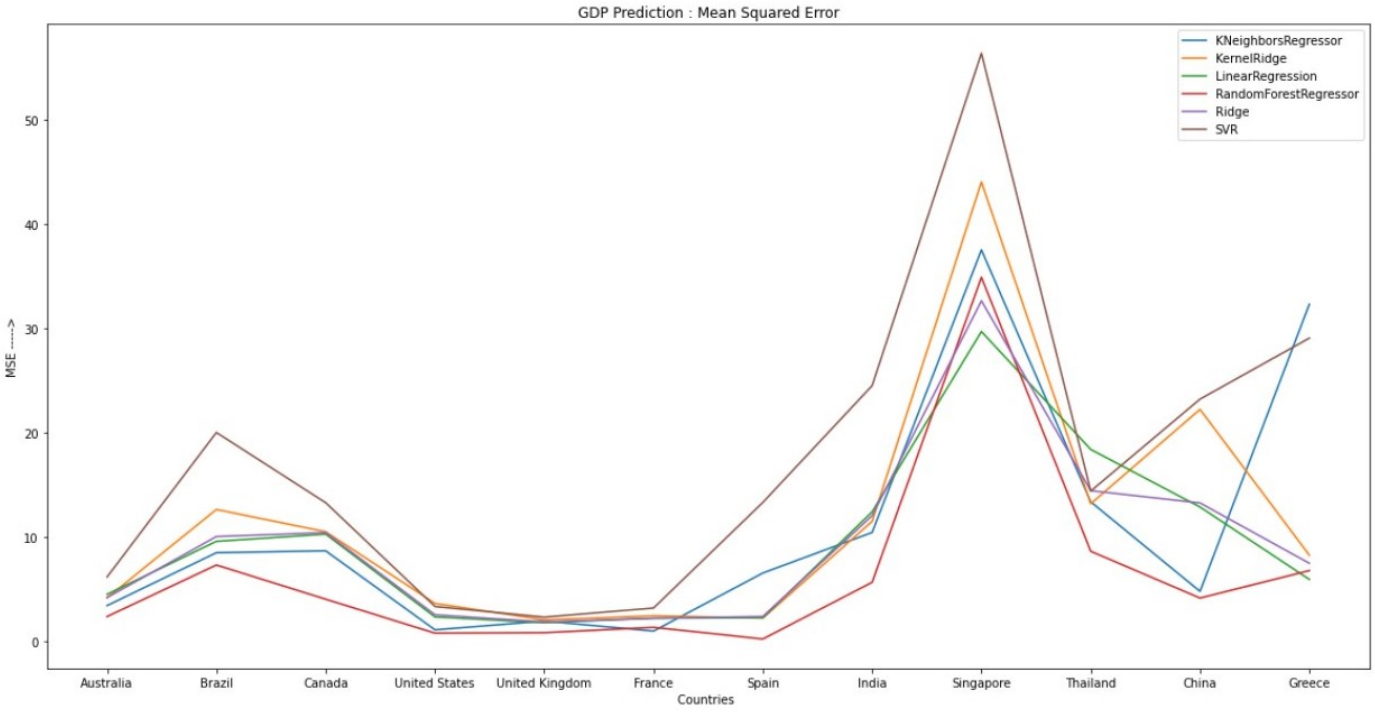

Figure 1: MSE Scores for various GDP predictor models

6.1 Features

6.1.1 Features for COVID cases Predictor

- Actual Confirmed Cases
- One day change
- Three day change
- Five day change
- Seven day change
- One day change rate (100*(One day change) / (Cases one day ago))
- Three day change rate
- Five day change rate
- Seven day change rate
- Infected rate
- Kalman Prediction

6.1.2 Features for training GDP Predictor

All the below data for each country.

- Total population

- Unemployment ratio
- Mobile cellular subscriptions per 100 people
- Population above 65 years of age
- Predicted COVID-19 cases for 2020 Quarter 3 (output from our other model)

The idea behind considering these features is that we actually wanted to capture features that don't affect GDP in a very direct fashion as well as those which do. For example, the number of mobile cellular subscriptions might be indicating a more digitally enhanced nation and by accounts of it, a more robust economy. This is in contrast to a feature like Unemployment Ratio that directly affects the GDP of a country.

6.2 Data Preprocessing

- Fixed region names in John Hopkins dataset.
- Considered regions for which reliable population data is available.
- For features like Unemployment ratio and Population above 65 years of age, the yearly data had to be converted to quarterly data.
- For quarterly values of COVID-19 cases before 2020, all values were filled as 0. We used spline interpolation to fill up these values. Spline interpolation is a form of interpolation where the interpolant is a special type of piecewise polynomial called a spline. Spline interpolation is often preferred over polynomial interpolation because the interpolation error can be made small even when using low degree polynomials for the spline.
- Interpolation is performed only if the number of data-points is more than 0. Otherwise the country is not taken into analysis.

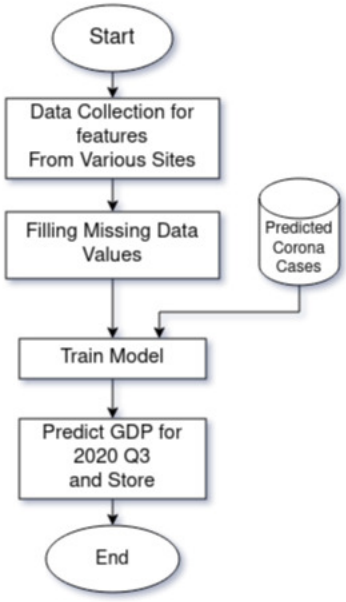

Figure 2: Flowchart of the work

7 Tools/ Softwares Used

Libraries/Frameworks:

- scikit-learn

- matplotlib
- tensorflow
- keras
- scipy

7.1 Algorithm for COVID cases prediction using Kalman Filter [1]

Algorithm 1: Kalman Prediction

Result: (Actual Confirmed Cases, Kalman Prediction)
initialization: matrices X,C,A,Q,B,J,M
for *1 to T* **do**
Update X_i and C_i as
$X_i = A_{i-1}X_{i-1} + B_iJ_i$,
$C_i = A_{i-1}C_{i-1}A_{i-1}^T + Q_{i-1}$
Parameters Updates
$M_i = Y_i - G_iX_i$
$S_i = G_iC_iG_i^T + R_i$
$K_i = C_iG_i^TS_i^{-1}$
$X_i = X_i + K_iM_i$
$C_i = C_i - K_iS_iK_i^T$
end

7.2 Architecture of Neural Net for COVID cases Prediction

- An input layer, 3 hidden layers, An output layer
- Activation function - LeakyRELU
- Optimizer - Adam
- Learning Rate - 0.001
- Cost function - MSE
- Epochs - 1000

7.3 Parameters in Random Forest Regressor for GDP Projection

- Number of estimators - 4
- Loss function - MSE

The entire project is run on Google Colaboratory. It is built on Python3. No external dependencies/permissions are required.

8 Results

8.1 Prediction of Confirmed cases using Kalman Filter

	0	confirmed	date	Country/Region
9768315		0	2020-12-09	India

Figure 3: Prediction table for countrywise confirmed Cases - Short Term

Actual cases on 2020-12-09 in India = 9771292

	Country/Region	rmse	mae
33	China	73.924522	8.080997

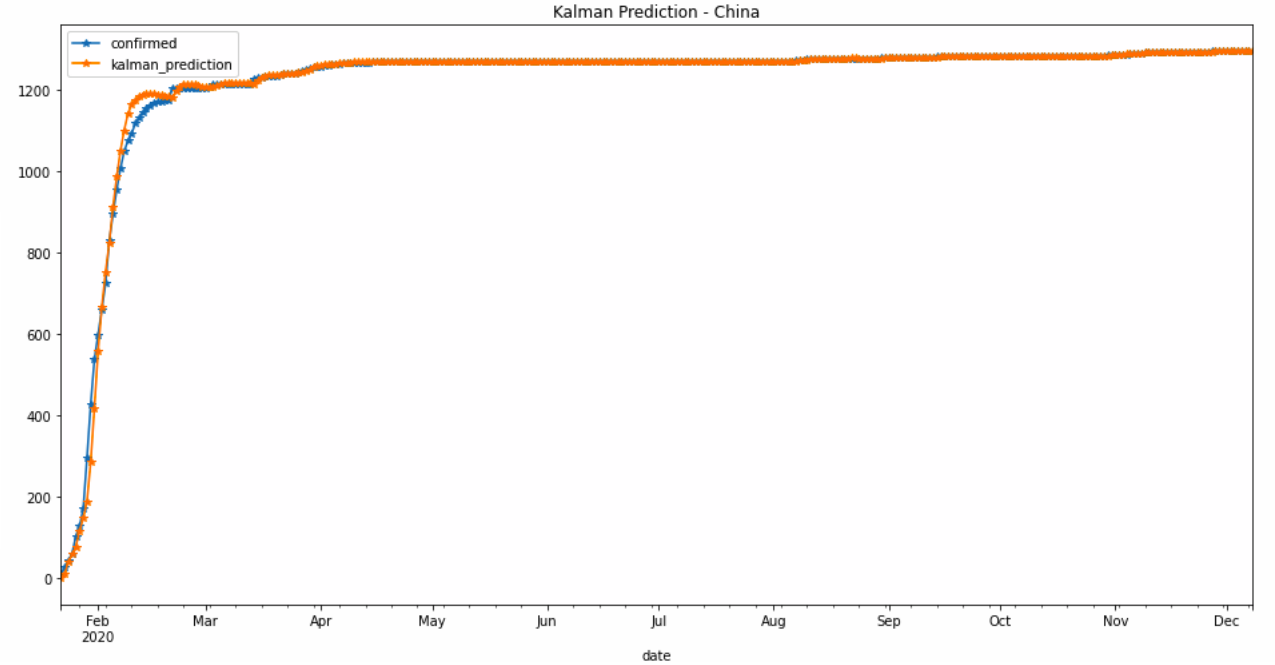

Figure 4: Prediction curve for China's confirmed Cases - Short Term

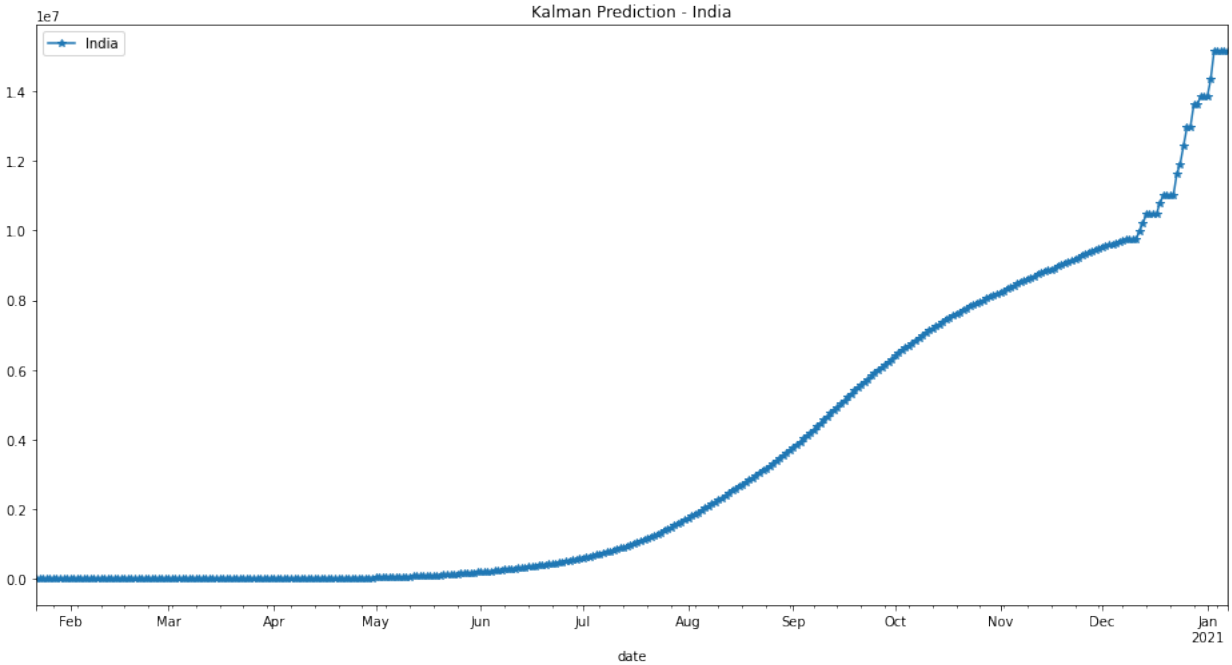

Figure 5: Prediction curve for countrywise confirmed Cases - Long Term

A sharp rise in coronavirus cases from July, roundabout whence the lockdown was lifted. The prediction trend shows inconsistencies for Jan-2021, raising concerns on using Kalman Prediction as reliable metric for modeling long-term trends.

8.2 Prediction of Confirmed and Death cases using Neural Network

Date	Confirmed(Predicted)	Confirmed(Actual)	Deaths(Predicted)	Deaths(Actual)
03 Dec	47368000	65220892	1700700	1506259
04 Dec	47877000	65899236	1714870	1518669
05 Dec	48392000	66539967	1729140	1528867
06 Dec	48912000	67073662	1743540	1536055
07 Dec	49438000	67591271	1758060	1544532
08 Dec	49970000	68225723	1772700	1556834
09 Dec	50507000	68894596	1787460	1569374
10 Dec	51050000	69592554	1802350	1581856
11 Dec	51599000	71081574	1817370	1594777
12 Dec	52153000	71704885	1832510	1605017

Figure 6: Prediction Table for Global Confirmed and Death Cases

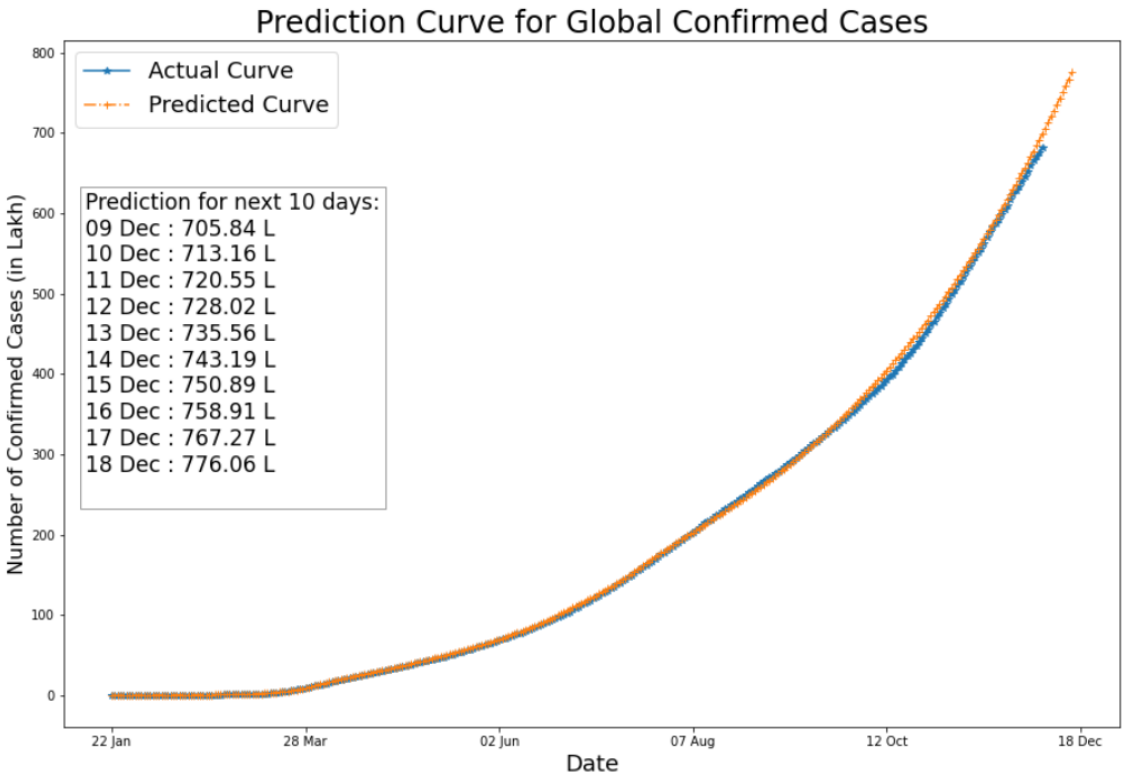

Figure 7: Prediction Curve for Global Confirmed Cases

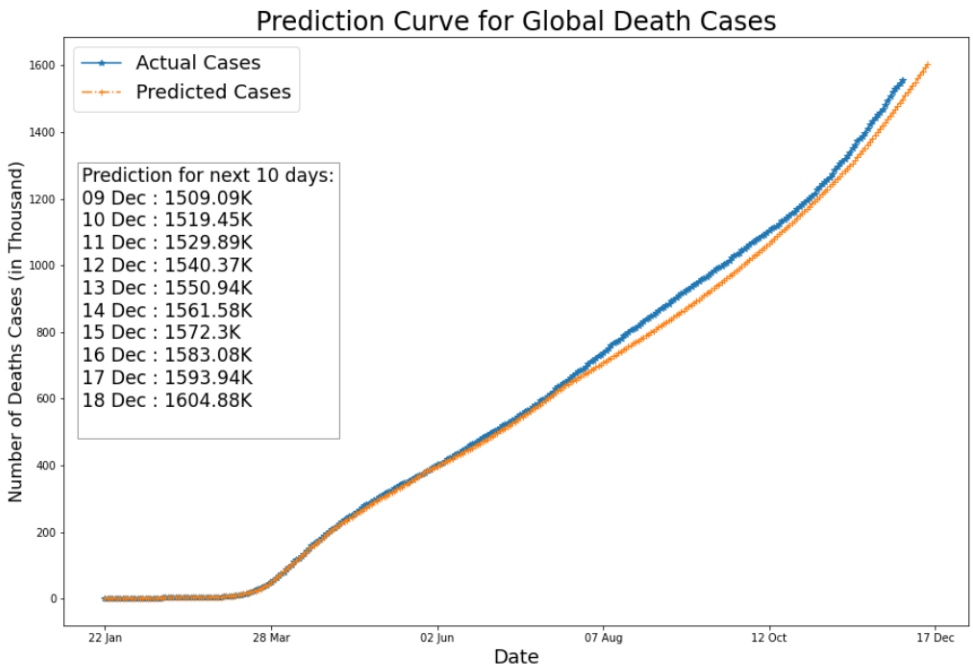

Figure 8: Prediction Curve for Global Death Cases

8.3 GDP Projection

Year/Quarter	2020Q1	2020Q2	2020Q3_Predicted
Australia	3.342654	-5.848295	0.312897
Brazil	-1.500000	-9.600000	-3.016573
Canada	1.432935	-11.300000	-3.917316
United States	2.006274	-5.000000	0.105028
United Kingdom	0.645599	-13.549770	-4.747919
France	-3.200245	-15.594570	-6.518146
Spain	-2.642289	-20.632573	-7.573589
India	-22.569823	-7.000000	1.366490
Singapore	-0.800000	-13.200000	-3.519417
Thailand	-1.328246	-14.488241	-4.469328
China	-6.800000	3.200000	6.084559
Greece	-2.742718	-17.112160	-7.683863

Figure 9: Prediction of percent change in GDP of Countries

9 Observations/ Insights

- Kalman prediction and time-dependent variables like day-wise changes are correlated to the target i.e. confirmed cases.
- The 1-day Kalman prediction is very accurate and powerful while a longer period prediction is more challenging but provides a future trend.
- Kalman filter results in large mean average error for long term predictions.
- Historical data has less effect on prediction as compared to the previous day data.
- GDP projection is consistent with actual projections for countries that provided comprehensive data about the influencers. For example, various reports in Spain show close to a 8% decrement in GDP for quarter 3 in comparison to previous quarter and 2% increment for India. Our predicted results closely follow such data.
- A decline in GDP has been projected for countries that are still under the grasp of the virus, for example India and UK. While countries that have undertaken preventive efforts early show growth in their gross domestic product.

10 Future Work

- Take testing scale into account for prediction of confirmed and death cases.
- Correlate the prediction of COVID-19 cases with other factors like healthcare, age groups etc.
- Continuous evaluation of GDP instead of quarterly evaluation.

11 External Links

- Gross Domestic Product(Real Index) - `https://data.imf.org/?sk=4c514d48-b6ba-49ed-8ab9-52b0c1a0179b`
- Total Population (2019) - `api.worldbank.org/v2/en/indicator/SP.POP.TOTL?downloadformat=csv`
- Unemployment - `api.worldbank.org/v2/en/indicator/SL.UEM.TOTL.ZS?downloadformat=csv`
- Unemployment [FOR INDIA] - `https://www.macrotrends.net/countries/IND/india/unemployment-rate`
- Mobile cellular subscriptions - `api.worldbank.org/v2/en/indicator/IT.CEL.SETS.P2?downloadformat=csv`
- Population age 65 and above data - `api.worldbank.org/v2/en/indicator/SP.POP.65UP.TO.ZS?downloadformat=csv`

Code-Link:
`https://drive.google.com/drive/folders/1MudvqEI6P1ipxaZwMAIYd-MN8ie6JuIt?usp=sharing`

Acknowledgements

[1] `https://link.springer.com/article/10.1007/s10489-020-01948-1`