Department of Industrial and Management Engineering,
Indian Institute of Technology, Kanpur

MBA652A – Statistical Modelling for Business Analytics

Project 3: Prediction of Covid19 Cases Location Wise and Number of Resulting Fatalities for Future Dates

GUIDED BY:

DEVLINA CHATTERJEE

PhD (IISc Bangalore)

SUBMITTED BY:

GROUP 03:

PRADNESH LACHAKE (19114007)

ASHISH UBANA (18114005)

SAURABH GUPTA (19114010)

VIVEK PRAJAPAT (19114019)

TABLE OF CONTENTS:

ACKNOWLEDGEMENT:

We are highly indebted to Prof. Devlina Chatterjee, for her guidance and continuous support in completing this project. It is because of the knowledge and skills acquired during the course work, along with her comprehensive style of teaching, that we are able to understand the subject in a better way and are able to complete this modelling project successfully.

INTRODUCTION

This project aims to predict the cumulative number of confirmed COVID19 cases in various locations across the world, as well as the number of resulting fatalities, for future dates. The data we collected from Johns Hopkins CSSE has both cross section and time series features as confirmed cases and fatalities vary location wise as well as time wise. Hence, we decided to use panel regression to predict the output.

Objective:

The objective of our project is as follows:

- Prediction of cumulative number of confirmed COVID19 cases in various locations across the world for future dates.
- Prediction of cumulative number of fatalities due to COVID19 in various locations across the world for future dates.

Methodology:

The methodology for this project is as follows:

- Summary of data and data visualization.
- Panel data analysis by building Pooled regression, Fixed effect regression, and Random effect regression models for both dependent variables, confirmed cases and fatalities.
- Testing for panel effects to choose one model over the other by using LM test and Hausman test.

Economic theory:

If we predict effects of COVID19 on our society in the future beforehand then:

- Government will get time to coordinate ways to enhance our health care delivery system capacity to respond to an increase in cases.
- We can do rapid assessment of the likely efficacy of school closures, travel bans, bans on mass gatherings of various sizes, and other social distancing approaches.
- Accordingly we can decide methods to control the spread in communities, barriers to compliance and how these vary among different populations.

DATA

Source:

The data for this project is taken from Johns Hopkins CSSE.
https://github.com/CSSEGISandData/COVID-19/tree/master/csse_covid_19_data/csse_covid_19_time_series

Description:

Our data set is a Panel Data set. A panel data set, also called longitudinal data set, is one that studies the same parameters at different points in time. It has a total of 35995 observations studied over 115 days from 22 January 2020 to 15 May 2020 for 313 unique regions of total 184 Countries.
The balance in the dataset is established as there are no missing values.

Variables:

Sr. No.	Variable	Description
1	Id	Unique row identifier
2	Province_State	Unique region name of a specific country
3	Country_Region	Country name
4	Date	Date
5	ConfirmedCases	Cumulative number of confirmed cases
6	Fatalities	Cumulative number of fatalities

Dependent Variables and Independent Variables:

Dependent Variables:
The dependent variable is the one being tested and measured and whose value may be affected by the change of other independent variables. In our analysis, we have following dependent variables:

1. **ConfirmedCases**: This variable contains the cumulative number of confirmed cases on a particular date and at a particular location. There is no missing data.
2. **Fatalities**: This variable contains the cumulative number of fatalities on a particular date and at a particular location. There is no missing data.

Independent Variables:
An independent variable is a variable that is changed or varied in an experiment to examine the effect on dependent variables. In our analysis, we have following independent variables:

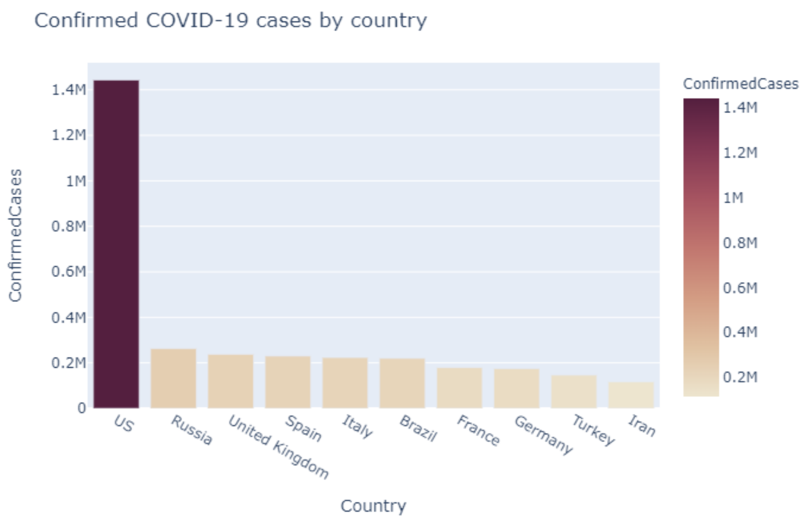
1. **Date**: This variable contains dates of total 115 days from 22 January 2020 to 15 May 2020. Hence, 115 observations are there for each unique region.

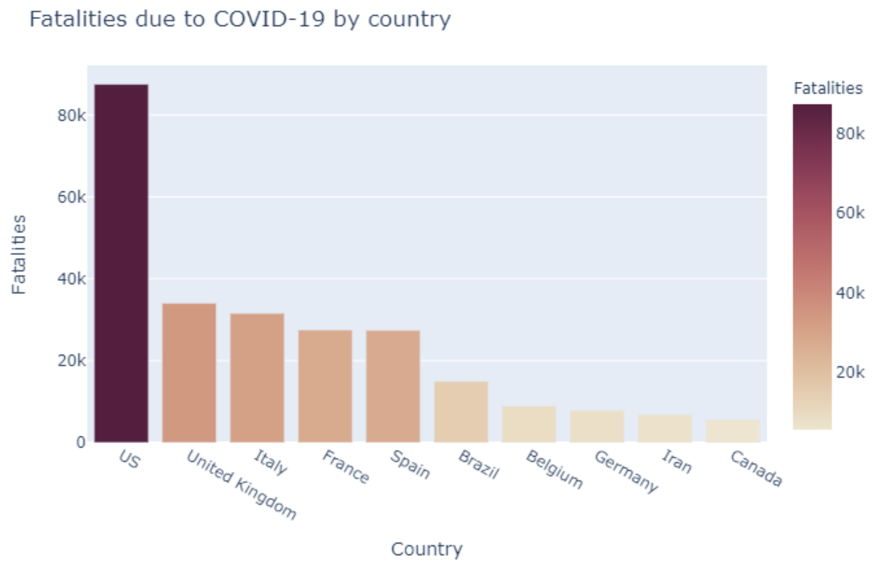
2. **Unique_region**: This column is created to show the unique regions across which data is collected. It is basically the combination of Province_State and Country_Region columns of the original dataset. Total 313 unique regions under 184 countries are there in the dataset. Because, some countries have more than one province state. Following is the list of countries with more than one province state:

Country	**No. of regions**
US	54.0
China	33.0
Canada	12.0
France	11.0
United Kingdom	11.0
Australia	8.0
Netherlands	5.0
Denmark	3.0

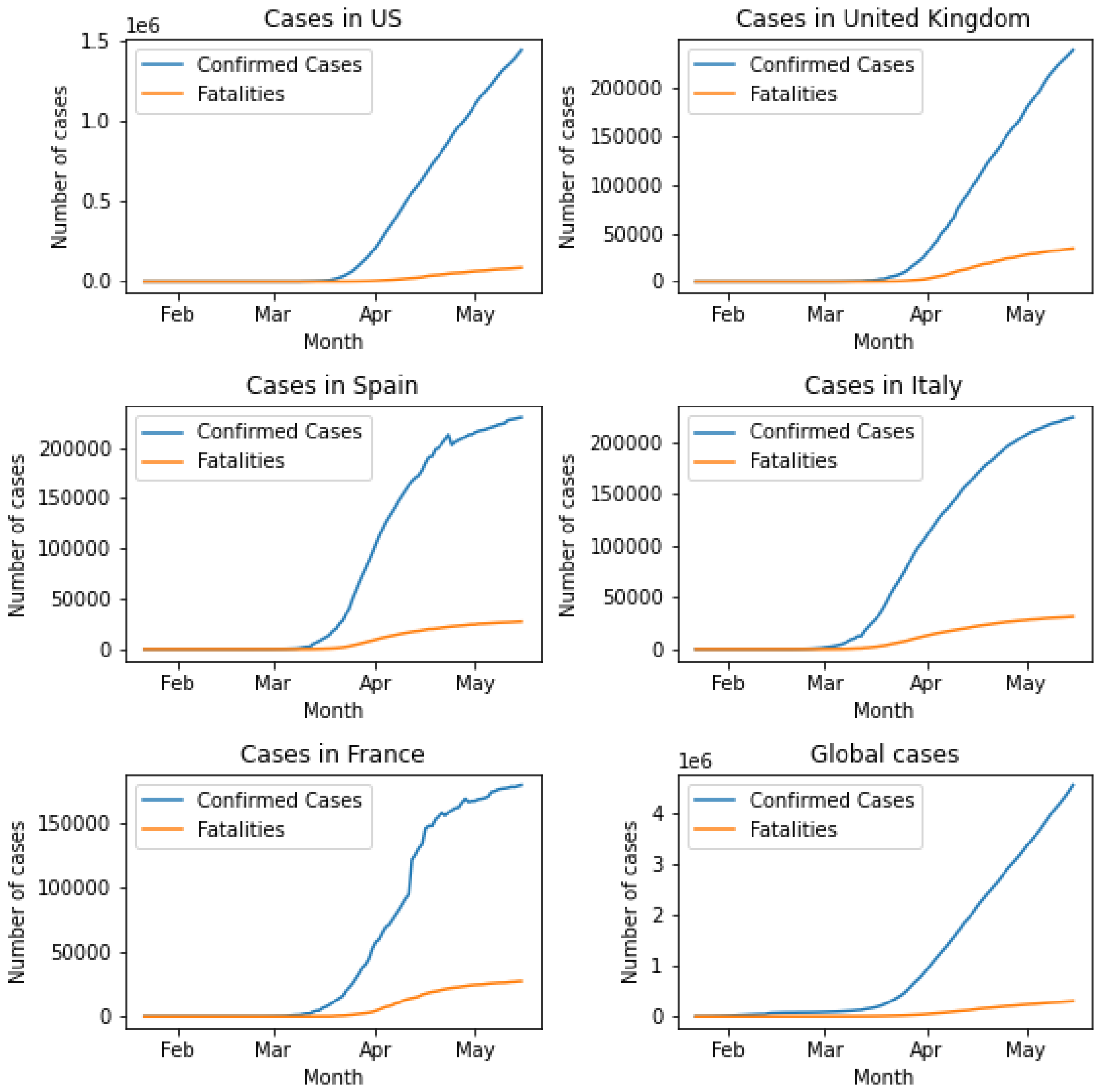
Distribution and Visualisation of the Data:

Top 10 countries with highest number of confirmed cases		Top 10 countries with highest number of fatalities	
Country	Confirmed Cases	Country	Fatalities
US	1442653	US	87525
Russia	262843	United Kingdom	34078
United Kingdom	238005	Italy	31610
Spain	230183	France	27532
Italy	223885	Spain	27459
Brazil	220291	Brazil	14962
France	179630	Belgium	8959
Germany	175233	Germany	7897
Turkey	146457	Iran	6902
Iran	116635	Canada	5679





Let's have a look at the data distribution of confirmed cases and fatalities over time for highly affected countries separately and also over entire globe:



PANEL DATA ANALYSIS

Panel data sets or a longitudinal data set, is a set of cross-sectional data collected for the same parameters across various years. For the analysis of the panel dataset in hand, we have employed the following regression techniques:

1. Pooled Regression
2. Fixed Effect Regression
   a. Entity Fixed Effects
   b. Time Fixed Effects
   c. Entity and Time Fixed Effects
3. Random Effect Regression

A. Confirmed Cases Prediction

1. Pooled Regression Model

The pooled regression model ignores any differences over entities or time and treats each data as a separate entity.

```
                          PooledOLS Estimation Summary
================================================================================
Dep. Variable:                      y   R-squared:                        0.9993
Estimator:                  PooledOLS   R-squared (Between):              0.9998
No. Observations:               35682   R-squared (Within):               0.9990
Date:                Fri, Jun 12 2020   R-squared (Overall):              0.9993
Time:                        15:13:54   Log-likelihood                -2.723e+05
Cov. Estimator:             Clustered
                                        F-statistic:                   5.421e+07
Entities:                         313   P-value                           0.0000
Avg Obs:                       114.00   Distribution:                F(1,35681)
Min Obs:                       114.00
Max Obs:                       114.00   F-statistic (robust):          1.421e+05
                                        P-value                           0.0000
Time periods:                     114   Distribution:                F(1,35681)
Avg Obs:                       313.00
Min Obs:                       313.00
Max Obs:                       313.00

                             Parameter Estimates
==============================================================================
            Parameter  Std. Err.     T-stat    P-value    Lower CI    Upper CI
------------------------------------------------------------------------------
x              1.0228     0.0027     376.97     0.0000      1.0175      1.0281
==============================================================================
```

2. Fixed Effect Regression Models

a. Entity Fixed Effects Model

In entity fixed models, each country is considered as a separate entity.

```
                          PanelOLS Estimation Summary
================================================================================
Dep. Variable:                      y   R-squared:                        0.9990
Estimator:                   PanelOLS   R-squared (Between):              0.9997
No. Observations:               35682   R-squared (Within):               0.9990
Date:                Fri, Jun 12 2020   R-squared (Overall):              0.9993
Time:                        15:11:22   Log-likelihood                -2.698e+05
Cov. Estimator:             Clustered
                                        F-statistic:                   3.669e+07
Entities:                         313   P-value                           0.0000
Avg Obs:                       114.00   Distribution:                 F(1,35368)
Min Obs:                       114.00
Max Obs:                       114.00   F-statistic (robust):          7.945e+04
                                        P-value                           0.0000
Time periods:                     114   Distribution:                 F(1,35368)
Avg Obs:                       313.00
Min Obs:                       313.00
Max Obs:                       313.00

                             Parameter Estimates
==============================================================================
            Parameter  Std. Err.     T-stat    P-value    Lower CI    Upper CI
------------------------------------------------------------------------------
x              1.0180     0.0036     281.86     0.0000      1.0110      1.0251
==============================================================================

F-test for Poolability: 15.253
P-value: 0.0000
Distribution: F(312,35368)

Included effects: Entity
```

b. Time Fixed Effects Model

In time fixed model, variations across the time are taken into consideration but not across entities.

```
                          PanelOLS Estimation Summary
================================================================================
Dep. Variable:                      y   R-squared:                        0.9993
Estimator:                   PanelOLS   R-squared (Between):              0.9998
No. Observations:               35682   R-squared (Within):               0.9990
Date:                Fri, Jun 12 2020   R-squared (Overall):              0.9993
Time:                        15:16:00   Log-likelihood                -2.718e+05
Cov. Estimator:             Clustered
                                        F-statistic:                   5.036e+07
Entities:                         313   P-value                           0.0000
Avg Obs:                       114.00   Distribution:                 F(1,35567)
Min Obs:                       114.00
Max Obs:                       114.00   F-statistic (robust):          1.402e+05
                                        P-value                           0.0000
Time periods:                     114   Distribution:                 F(1,35567)
Avg Obs:                       313.00
Min Obs:                       313.00
Max Obs:                       313.00

                             Parameter Estimates
==============================================================================
            Parameter  Std. Err.     T-stat    P-value    Lower CI    Upper CI
------------------------------------------------------------------------------
x              1.0223     0.0027     374.43     0.0000      1.0169      1.0276
==============================================================================

F-test for Poolability: 5.2816
P-value: 0.0000
Distribution: F(113,35567)

Included effects: Time
```

c. Entity & Time Fixed Effects Model

In this model, variations across both the time and countries are considered. Through this both the effects that vary across the countries but remain same over time and the effects that vary across the time and remain same across the countries are handled.

```
                          PanelOLS Estimation Summary
================================================================================
Dep. Variable:                      y   R-squared:                        0.9990
Estimator:                   PanelOLS   R-squared (Between):              0.9997
No. Observations:               35682   R-squared (Within):               0.9990
Date:                Fri, Jun 12 2020   R-squared (Overall):              0.9993
Time:                        15:11:24   Log-likelihood                -2.695e+05
Cov. Estimator:             Clustered
                                        F-statistic:                   3.392e+07
Entities:                         313   P-value                           0.0000
Avg Obs:                       114.00   Distribution:                F(1,35255)
Min Obs:                       114.00
Max Obs:                       114.00   F-statistic (robust):          7.604e+04
                                        P-value                           0.0000
Time periods:                     114   Distribution:                F(1,35255)
Avg Obs:                       313.00
Min Obs:                       313.00
Max Obs:                       313.00

                             Parameter Estimates
==============================================================================
            Parameter  Std. Err.     T-stat    P-value    Lower CI    Upper CI
------------------------------------------------------------------------------
x              1.0176     0.0037     275.75     0.0000      1.0103      1.0248
==============================================================================

F-test for Poolability: 13.181
P-value: 0.0000
Distribution: F(425,35255)

Included effects: Entity, Time
```

3. Random Effect Regression Model

In this the entities are chosen at random, hence the effect of not including the entity would not be correlated with the dependent variable.

```
                        RandomEffects Estimation Summary
================================================================================
Dep. Variable:                      y   R-squared:                        0.9991
Estimator:              RandomEffects   R-squared (Between):              0.9998
No. Observations:               35682   R-squared (Within):               0.9990
Date:                Fri, Jun 12 2020   R-squared (Overall):              0.9993
Time:                        15:11:24   Log-likelihood                -2.701e+05
Cov. Estimator:            Unadjusted
                                        F-statistic:                   3.871e+07
Entities:                         313   P-value                           0.0000
Avg Obs:                       114.00   Distribution:                 F(1,35680)
Min Obs:                       114.00
Max Obs:                       114.00   F-statistic (robust):          3.871e+07
                                        P-value                           0.0000
Time periods:                     114   Distribution:                 F(1,35680)
Avg Obs:                       313.00
Min Obs:                       313.00
Max Obs:                       313.00

                             Parameter Estimates
==============================================================================
            Parameter  Std. Err.     T-stat    P-value    Lower CI    Upper CI
------------------------------------------------------------------------------
Intercept      60.105     7.7804     7.7252     0.0000      44.855      75.355
x              1.0187     0.0002     6222.1     0.0000      1.0184      1.0190
==============================================================================
```

B. Fatalities Prediction

We can also determine all models' performance for prediction of fatalities in the same way we did for confirmed cases prediction. Following is the summary of time fixed effect model:

```
                          PanelOLS Estimation Summary
================================================================================
Dep. Variable:             Fatalities   R-squared:                        0.9994
Estimator:                   PanelOLS   R-squared (Between):              0.9999
No. Observations:               35682   R-squared (Within):               0.9991
Date:                Fri, Jun 12 2020   R-squared (Overall):              0.9994
Time:                        15:39:08   Log-likelihood                -1.863e+05
Cov. Estimator:             Clustered
                                        F-statistic:                   2.911e+07
Entities:                         313   P-value                           0.0000
Avg Obs:                       114.00   Distribution:                F(2,35566)
Min Obs:                       114.00
Max Obs:                       114.00   F-statistic (robust):          1.881e+05
                                        P-value                           0.0000
Time periods:                     114   Distribution:                F(2,35566)
Avg Obs:                       313.00
Min Obs:                       313.00
Max Obs:                       313.00

                             Parameter Estimates
==============================================================================
            Parameter  Std. Err.     T-stat    P-value    Lower CI    Upper CI
------------------------------------------------------------------------------
Fata_lag       1.0098     0.0051     199.80     0.0000      0.9999      1.0197
x              0.0013     0.0004     2.9452     0.0032      0.0004      0.0022
==============================================================================

F-test for Poolability: 6.1684
P-value: 0.0000
Distribution: F(113,35566)

Included effects: Time
```

TESTING FOR PANEL EFFECT

1. LM Test

To decide whether the OLS Pooled model is better or the Fixed Effects Model, we use the LM Test.

The null hypothesis is that OLS is better than the fixed effects model.

Since p-value is less than 0.05, we reject null hypothesis, thus we can say that the fixed effects model should be chosen over the OLS pooling model.

2. Hausman Test

Test for choosing between Fixed Effect Model or Random Effect Model.

To decide whether the Fixed effect model is better or the Random effect model, we use the Hausman Test.

The null hypothesis is that Fixed Effect model is better than Random Effect model.

Since the p-value is larger than 0.05, we fail to reject the null hypothesis, thus we can say that fixed effect model is better than the random effects model.

CONCLUSION

After building several models and carrying out various tests, we have found that Fixed Effects Regression model is better than Pooled OLS or Random Effect Model and that the Time Fixed Effect model has a higher R2 value and is better able to explain the dependable variable.

Even after all these regressions, we can't guarantee that the models in here will be able to perfectly capture the relation of the dependable variables as there might be omitted variable bias. There are several omitted variables such as the population of the region, average population age of the region, steps taken by the government to stop the spread, etc. variables may have a significant impact on the active COVID 19 cases and ultimately on the fatalities in that region.

REFERENCES

Dataset: https://github.com/CSSEGISandData/COVID-19/tree/master/csse_covid_19_data/csse_covid_19_time_series

Linear model documentation: https://bashtage.github.io/linearmodels/doc/panel/models.html

PYTHON CODE

```
# Importing necessary libraries
import pandas as pd
import matplotlib.pyplot as plt
import plotly.express as px
import warnings
warnings.filterwarnings("ignore")

# Exploring the dataset
df_train = pd.read_csv('train.csv')
df_train.shape
df_train.head()

df_1 =
df_train.fillna('NA').groupby(['Country_Region','Province_State','Date'])['ConfirmedCases'].s
um() \
                    .groupby(['Country_Region','Province_State']).max().sort_values() \
                    .groupby(['Country_Region']).sum().sort_values(ascending = False)
top10_confirmed_cases = pd.DataFrame(df_1).head(10)
top10_confirmed_cases
fig = px.bar(top10_confirmed_cases, x=top10_confirmed_cases.index, y='ConfirmedCases',
labels={'x':'Country'},
          color="ConfirmedCases", color_continuous_scale=px.colors.sequential.Brwnyl)
fig.update_layout(title_text='Confirmed COVID-19 cases by country')
fig.show()

df_2 =
df_train.fillna('NA').groupby(['Country_Region','Province_State','Date'])['Fatalities'].sum() \
                    .groupby(['Country_Region','Province_State']).max().sort_values() \
                    .groupby(['Country_Region']).sum().sort_values(ascending = False)
top10_fatalities = pd.DataFrame(df_2).head(10)
top10_fatalities
fig = px.bar(top10_fatalities, x=top10_fatalities.index, y='Fatalities', labels={'x':'Country'},
          color="Fatalities", color_continuous_scale=px.colors.sequential.Brwnyl)
fig.update_layout(title_text='Fatalities due to COVID-19 by country')
fig.show()

#Country wise distribution of data
```

```
us = df_train[df_train.Country_Region=="US"].groupby("Date")["ConfirmedCases",
"Fatalities"].sum().reset_index()
uk = df_train[df_train.Country_Region=="United
Kingdom"].groupby("Date")["ConfirmedCases", "Fatalities"].sum().reset_index()
spain = df_train[df_train.Country_Region=="Spain"].groupby("Date")["ConfirmedCases",
"Fatalities"].sum().reset_index()
italy = df_train[df_train.Country_Region=="Italy"].groupby("Date")["ConfirmedCases",
"Fatalities"].sum().reset_index()
france = df_train[df_train.Country_Region=="France"].groupby("Date")["ConfirmedCases",
"Fatalities"].sum().reset_index()
total = df_train.groupby("Date")["ConfirmedCases", "Fatalities"].sum().reset_index()
plt.figure(figsize=(8, 8))

#subplot 1
ax1 = plt.subplot(3, 2, 1)
plt.plot(us.Date, us.ConfirmedCases, label='Confirmed Cases')
plt.plot(us.Date, us.Fatalities, label='Fatalities')
plt.xlabel("Month")
plt.ylabel("Number of cases")
plt.title("Cases in US")
plt.legend()
plt.xticks(["2020-02-01","2020-03-01", "2020-04-01", "2020-05-01"])
ax1.set_xticklabels(["Feb", "Mar", "Apr", "May"])

#subplot 2
ax2 = plt.subplot(3, 2, 2)
plt.plot(uk.Date, uk.ConfirmedCases, label='Confirmed Cases')
plt.plot(uk.Date, uk.Fatalities, label='Fatalities')
plt.xlabel("Month")
plt.ylabel("Number of cases")
plt.title("Cases in United Kingdom")
plt.legend()
plt.xticks(["2020-02-01","2020-03-01", "2020-04-01", "2020-05-01"])
ax2.set_xticklabels(["Feb", "Mar", "Apr", "May"])

#subplot 3
ax3 = plt.subplot(3, 2, 3)
plt.plot(spain.Date, spain.ConfirmedCases, label='Confirmed Cases')
plt.plot(spain.Date, spain.Fatalities, label='Fatalities')
plt.xlabel("Month")
plt.ylabel("Number of cases")
plt.title("Cases in Spain")
plt.legend()
plt.xticks(["2020-02-01","2020-03-01", "2020-04-01", "2020-05-01"])
```

```
ax3.set_xticklabels(["Feb", "Mar", "Apr", "May"])

#subplot 4
ax4 = plt.subplot(3, 2, 4)
plt.plot(italy.Date, italy.ConfirmedCases, label='Confirmed Cases')
plt.plot(italy.Date, italy.Fatalities, label='Fatalities')
plt.xlabel("Month")
plt.ylabel("Number of cases")
plt.title("Cases in Italy")
plt.legend()
plt.xticks(["2020-02-01","2020-03-01", "2020-04-01", "2020-05-01"])
ax4.set_xticklabels(["Feb", "Mar", "Apr", "May"])

#subplot 5
ax5 = plt.subplot(3, 2, 5)
plt.plot(france.Date, france.ConfirmedCases, label='Confirmed Cases')
plt.plot(france.Date, france.Fatalities, label='Fatalities')
plt.xlabel("Month")
plt.ylabel("Number of cases")
plt.title("Cases in France")
plt.legend()
plt.xticks(["2020-02-01","2020-03-01", "2020-04-01", "2020-05-01"])
ax5.set_xticklabels(["Feb", "Mar", "Apr", "May"])

#subplot 6
ax6 = plt.subplot(3, 2, 6)
plt.plot(total.Date, total.ConfirmedCases, label='Confirmed Cases')
plt.plot(total.Date, total.Fatalities, label='Fatalities')
plt.xlabel("Month")
plt.ylabel("Number of cases")
plt.title("Global cases")
plt.legend()
plt.xticks(["2020-02-01","2020-03-01", "2020-04-01", "2020-05-01"])
ax6.set_xticklabels(["Feb", "Mar", "Apr", "May"])

plt.tight_layout()
plt.savefig("Distribution.png")
plt.show()

print(f"Unique Countries = {(df_train.Country_Region.nunique())}")
print(f"Period = {len(df_train.Date.unique())} days")
print(f"From = {df_train.Date.min()}, To = {df_train.Date.max()}")
print(f"Unique Regions = {df_train.shape[0]/len(df_train.Date.unique())}")
```

```
#Countries having more than one province state
df_temp_1 = df_train.groupby("Country_Region").Date.apply(lambda x:
x.count()/len(x.unique())).reset_index()
df_temp_2 = df_temp_1.rename(columns={"Country_Region":"Country", "Date":"No. of
Regions"})
df_temp_2[df_temp_2["No. of Regions"]>1].sort_values("No. of Regions",
ascending=False).reset_index(drop=True)

#Checking for empty values in columns
print("Empty values in columns:")
print(f"Country_Region = {df_train.Country_Region.isnull().sum()}")
print(f"Date = {df_train.Date.isnull().sum()}")
print(f"ConfirmedCases = {df_train.ConfirmedCases.isnull().sum()}")
print(f"Fatalities = {df_train.Fatalities.isnull().sum()}")

#create a column containing unique regions
function = lambda row: f"{row.Province_State}.{row.Country_Region}" if
pd.notnull(row.Province_State) else row.Country_Region
df_train["Unique_Region"] = df_train.apply(function, axis=1)
df_train.sample(5)

#Drop 3columns : Id, Province_State and Country_region
df_train.drop(columns=["Id", "Province_State", "Country_Region"], inplace=True)
df_train.head()

#create y, x, fatalities, fatalities lag, Unique_Region and time varaibles
df_panel = pd.DataFrame()
df_panel['y'] = df_train.ConfirmedCases
df_panel['x'] = df_panel.y.shift(1)
df_panel['Fatalities'] = df_train.Fatalities
df_panel['Fata_lag'] = df_panel.Fatalities.shift(1)
df_panel['Unique_Region'] = df_train.Unique_Region
df_panel['time'] = df_train.Date

#Drop day1 of each region since we dont have value of x value
df_panel.drop(df_panel[df_panel.time=='2020-01-22']. index, inplace=True)

#convert datetime to interger
df_panel['time'] = pd.to_numeric(df_panel.time.str.replace('-',''))
df_panel.sample(5)

#create data as panel data
```

```
df_panel = df_panel.set_index(['Unique_Region','time'])
df_panel
#Pooled regression model
from linearmodels import PooledOLS
mod = PooledOLS(df_panel.y, df_panel.x)
res_1 = mod.fit(cov_type='clustered', cluster_entity=True)
print(res_1)

# Entity fixed effect regression
from linearmodels import PanelOLS
mod = PanelOLS(df_panel.y, df_panel.x, entity_effects=True)
res_2 = mod.fit(cov_type='clustered', cluster_entity=True)
print(res_2)

# Time fixed effect regression
mod = PanelOLS(df_panel.y, df_panel.x, time_effects=True)
res_3 = mod.fit(cov_type='clustered', cluster_entity=True)
print(res_3)

# Entity and time fixed effect regression
mod = PanelOLS(df_panel.y, df_panel.x, entity_effects=True, time_effects=True)
res_4 = mod.fit(cov_type='clustered', cluster_entity=True)
print(res_4)

# random effect regression
from linearmodels import RandomEffects
mod = RandomEffects.from_formula('y ~ 1 + x', df_panel)
res_5 = mod.fit()
print(res_5)

# Time fixed effect regression for fatalities
mod = PanelOLS.from_formula('Fatalities ~ Fata_lag + x + TimeEffects', df_panel)
res = mod.fit(cov_type='clustered', cluster_entity=True)
print(res)
```