1. INTRODUCTION

Provide an overview of your project relative to your specific goals. Explain why this project was important to you. Explain your hypothesis (or hypotheses) you set out to test. What hunch(es) did you have about the data?

1.1 Project overview

By calculating case_fatality_rate and fatality_rate, and various analyses including correlation and geo heatmap analysis, I I confirmed some hypothesized correlations of these rates with some socioeconomical factors.

1.2 Explanation of importance

I hope to learn the differences in overall incidence rate (overall incidences per 100 thousand individuals), overall fatality rate (overall fatalities per 100 thousand individuals), and case fatality rate (overall fatalities per overall incidences) of the COVID-19 pandemic between/among different EU/EEA countries.

This data may reveal part of the differences of capabilities of these countries in meeting the public health challenges presented by the COVID-19 pandemic. Overall incidence rate may be related to the capabilities in preventive medicine, and overall fatality rate may be related to the capabilities both in preventive medicine and in therapeutic medicine. Case fatality rate may be related to the capabilities more in therapeutic medicine than in preventive medicine. However, "comparisons between countries should be done with caution" as warned by ECDC at https://www.ecdc.europa.eu/sites/default/files/documents/Description-and-disclaimer_daily_reporting.pdf.

1.3 Hypothesis

These rates of interest, are hypothesized to negatively correlate with gdp_pps, healthExp_pi and positively correlate with L_pneumDeath, and rreg (a new variable created in this project). popDen is thought to positively correlate to incidence_rate. and incidence rate correlate to fatality rate.

2. DATA EXPLAINED

Provide a link to your data source if applicable. Give a brief over view of the steps needed to clean the data. Provide a data dictionary to the final data used in your analysis.

2.1 Data sources

Data preparation

Generation or conversion of variables (Please also refer to 2.3 Data dictionary):

country is renamed from countriesAndTerritories;

(daily) cases are summed over the time span until 7/10/2022 to provide (total) cases for each country;

(daily) deaths are summed over the time span until 7/10/2022 to provide (total) deaths for each country;

case_fatality_rate is the ratio between total deaths and total cases;

pop2020 is renamed from popData2020;

incidence_rate, or overall incidence rate, is the sum of the daily incidence rate, or the sum of the ratio between the number of newly reported (daily) cases and pop2020 of the country multiplied by 100 thousand;

fatality_rate, or overall fatality rate, is the sum of the daily fatality rate, or the sum of the ratio between the number of newly reported (daily) deaths and pop2020 of the country multiplied by 100 thousand;

popDen, or population density, is the ratio between pop2020 and area of the country;

L_pneumDeath is the logarithm in the base of 10 of pneumDeath, or the death rate of pneumonia of the country standardized to the European population;

rreg is the reciprocal of the ratio between healthExp_pi to gdp_pps multiplied by 100 or the ratio between gdp_pps to healthExp_pi multiplied by 100 (10000/euro);

Data cleaning:

Daily incidence_rate and fatality_rate have high variations but they are valid data according to the data sources. Since this project is using overall incidence_rate and fatality_rate, both normally distributed, for further analysis, the 22 EU countries with no or less than 6 days of missing (daily) cases and deaths by early this July are selected and missing values are relaced by zero on those days.

2.2 Data dictionary

Variable, Definition, Datatype, Use (yes/no), Justification, Inclusion in final data;

geoId, 2-letter code of the country, string/object, yes, used as country id for analysis, included in the final data;

country, Name of the country, string/object, yes, used as country name for analysis, included in the final data;

case_fatality_rate, Ratio between total deaths and total cases or ratio between overall fatality_rate and overall incidence_rate of the country, float64, yes, used for analysis (important parameter for COVID-19), included in the final data;
incidence_rate, Overall incidence_rate = sum of the incidence_rate, or sum of the ratio between the number of newly reported (daily) cases and pop2020 of the country multiplied by 100 thousand), float64, yes, used for calculation and analysis (important parameter for COVID-19), included in the final data;
fatality_rate, Overall fatality_rate = sum of the fatality_rate, or sum of the ratio between the number of newly reported (daily) deaths and pop2020 of the country multiplied by 100 thousand), float64, yes, used for calculation and analysis (important parameter for COVID-19), included in the final data;
popDen, Population density or ratio between the population and the area of the country, float64, yes, used for analysis (expected to correlate to important parameter for COVID-19), included in the final data;
gdp_pps, Gross domestic product per capita in purchasing power standards (PPS) or ratio between the level of gross domestic product (GDP), expressed in purchasing power standards, and total population (Volume indices of real expenditure per capita (in PPS_EU27_2020=100)) of the country, int32, yes, used for calculation and anaylysis (expected to correlate to important parameter for COVID-19), included in the final data;
L_pneumDeath, Logarithm of death rate of pneumonia of the country standardized to the European population, float64, yes, used for analysis (expected to correlate to important parameter for COVID-19), included in the final data;

healthExp_pi (in euro per inhabitant), Annual healthcare expenditure per inhabitant of the country, float64, yes, used for calculation and anaylysis (expected to correlate to important parameter for COVID-19), included in the final data;

rreg, Reciprocal of the ratio between healthExp_pi to gdp_pps multiplied by 100 (10000/euro), float64, yes, used for analysis (expected to correlate to important parameter for COVID-19), included in the final data;

Not included in the final data but useful:

cases, Total cases = sum of the number of newly reported (daily) cases of the country, float64, yes, used for calculation of incidence_rate and case_fatality_rate, not included in the final data;

deaths, Total deaths = sum of the number of newly reported (daily) deaths of the country, float64, yes, used for calculation of fatality_rate and case_fatality_rate, not included in the final data;

pop2020, Eurostat 2020 data of the country, int64, yes, used for calculation(daily), not included in the final data;

area (in square kilometre), Area of the country, float64, yes, used for calculation, not included in the final data;

pneumDeath, Death rate of pneumonia of the country standardized to the European population, float64, yes, used for calculation and preliminary analysis (expected to correlate to important parameter for COVID-19), not included in the final data;

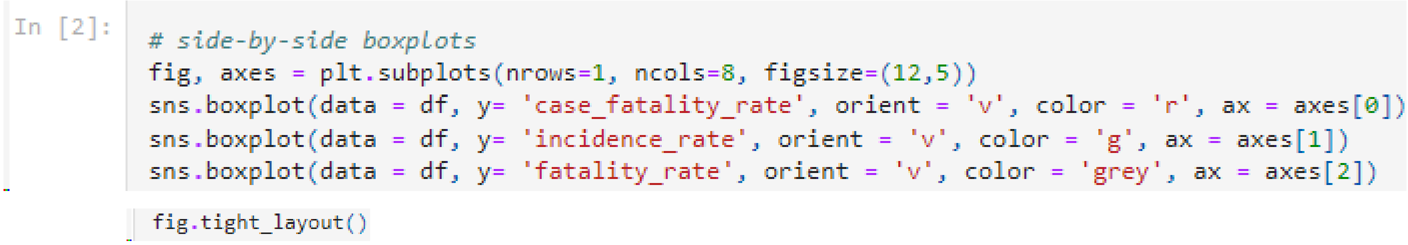
In [1]:

```
# general preparation
import pandas as pd
import numpy as np
import seaborn as sns
import matplotlib.pyplot as plt

# set up notebook to display multiple output in one cell
from IPython.core.interactiveshell import InteractiveShell
InteractiveShell.ast_node_interactivity = "all"

# read in final data
df = pd.read_csv('covid_report_data.csv')
```

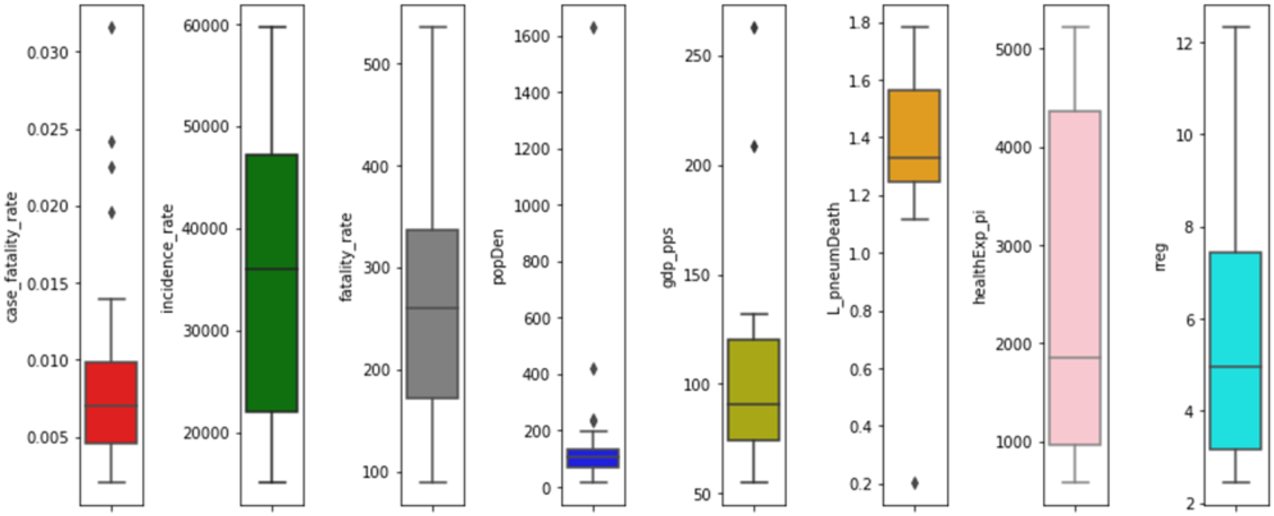
3. RESULTS

Read in your final data. Include code that shows your most important findings. Select at least five analytical processes that highlight your data and your findings. Remember that what you are showing in code is related to your hypotheses/hunches. This section is a combination of code and a summary of findings. Note that all findings should be clearly written in Markdown cells and should be adjacent to the code cells being explained.

3.1 Boxplots for data validation

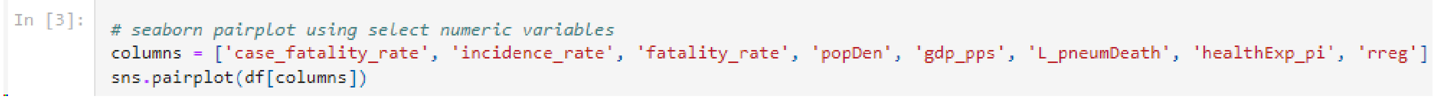


```
In [2]:  # side-by-side boxplots
         fig, axes = plt.subplots(nrows=1, ncols=8, figsize=(12,5))
         sns.boxplot(data = df, y= 'case_fatality_rate', orient = 'v', color = 'r', ax = axes[0])
         sns.boxplot(data = df, y= 'incidence_rate', orient = 'v', color = 'g', ax = axes[1])
         sns.boxplot(data = df, y= 'fatality_rate', orient = 'v', color = 'grey', ax = axes[2])
          fig.tight_layout()
```

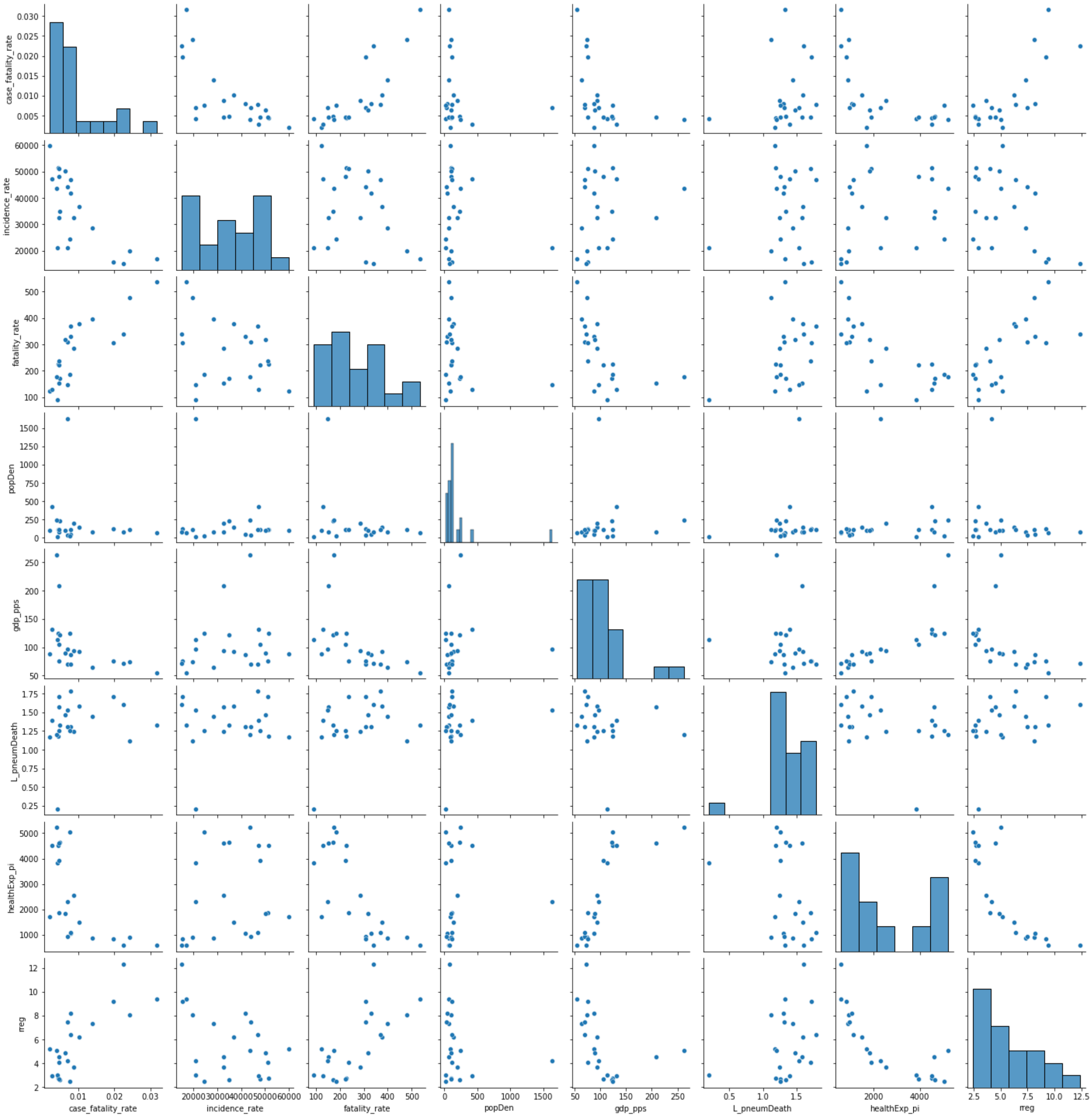


3.2 Seaborn pairplot for data distribution and correlation

```
'case_fatality_rate': right-skewed distribution with outliers at the high value;
    negative correlation with 'incidence_rate', as expected from mathematical relationship;
    positive correlation with 'fatality_rate', as expected from mathematical relationship;
    seemingly no correlation with 'popDen' (broad distribution over similar 'popDen'); possible
    negative correlation with 'gdp_pps', as expected;
    seemingly no correlation with 'L_pneumDeath' (broad distribution over similar 'L_pneumDeath'), unexpected;
    possible negative correlation with 'healthExp_pi', as expected;
    possible positive correlation with 'rreg', as expected.........................
```



```
In [3]:  # seaborn pairplot using select numeric variables
         columns = ['case_fatality_rate', 'incidence_rate', 'fatality_rate', 'popDen', 'gdp_pps', 'L_pneumDeath', 'healthExp_pi', 'rreg']
         sns.pairplot(df[columns])
```



3.3 Correlation matrix and heatmap to confirm correlations

```
Confirmed correlations (absolute value of correlation coefficients [0-0.25): no; [0.25-0.50): weak; [0.5-
0.75): moderate; [0.75-1]: strong) for:

'case_fatality_rate':
    moderate negative correlation with 'incidence_rate', as expected;
    strong positive correlation with 'fatality_rate', as expected;
    no correlation with 'popDen';
    weak negative correlation with 'gdp_pps', as expected;
    no correlation with 'L_pneumDeath', unexpected;
    moderate negative correlation with 'healthExp_pi', as expected;
    strong positive correlation with 'rreg', as expected. …………………….
```

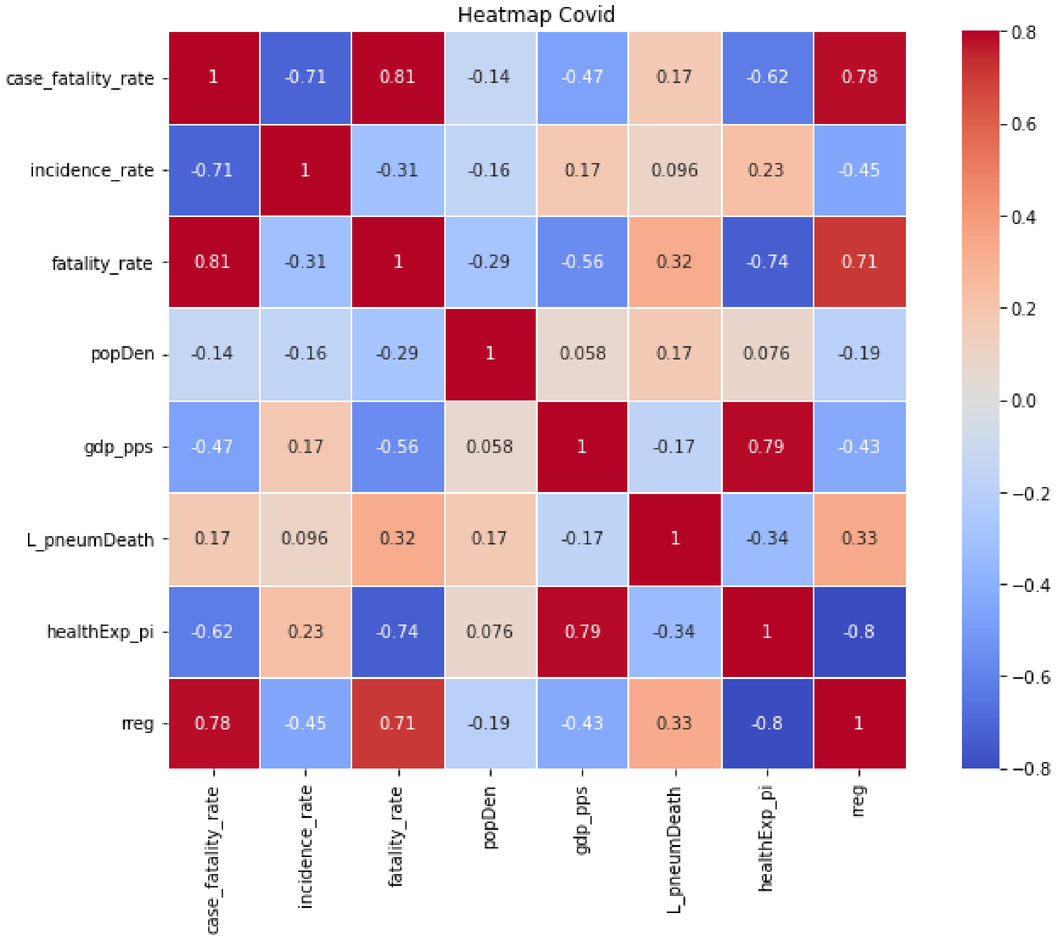
In [4]:

```
# create a data matrix using select numeric variables
df_corr = df[columns]
df_corr.corr()

# create a heat map using all numeric variables.
# https://matplotlib.org/3.1.1/gallery/color/colormap_reference.html
corrmat = df_corr.corr()
f, ax = plt.subplots(figsize=(12, 8))
sns.heatmap(corrmat, vmax=.8, square=True, annot=True, cmap='coolwarm', linewidths=.5 )
plt.title('Heatmap Covid')
plt.savefig('Correlation Heat Map Covid')

# explain how the visual cues of the heatmap represent the correlactions.
print('The red color represents positive correlation and the blue color represents negative correlation.')
print('Darker color indicates higher absolute value of correlation coefficient.')
```

Out[4]:

	case_fatality_rate	incidence_rate	fatality_rate	popDen	gdp_pps	L_pneumDeath	healthExp_pi	rreg
case_fatality_rate	1.000000	-0.710206	0.808063	-0.137006	-0.468682	0.168729	-0.622046	0.781757
incidence_rate	-0.710206	1.000000	-0.313502	-0.158528	0.174623	0.095860	0.227121	-0.450789
fatality_rate	0.808063	-0.313502	1.000000	-0.288966	-0.564582	0.321013	-0.739114	0.709414
popDen	-0.137006	-0.158528	-0.288966	1.000000	0.057888	0.168318	0.075930	-0.192700
gdp_pps	-0.468682	0.174623	-0.564582	0.057888	1.000000	-0.165874	0.789839	-0.428296
L_pneumDeath	0.168729	0.095860	0.321013	0.168318	-0.165874	1.000000	-0.340597	0.329921
healthExp_pi	-0.622046	0.227121	-0.739114	0.075930	0.789839	-0.340597	1.000000	-0.802516
rreg	0.781757	-0.450789	0.709414	-0.192700	-0.428296	0.329921	-0.802516	1.000000

```
The red color represents positive correlation and the blue color represents negative correlation.
Darker color indicates higher absolute value of correlation coefficient.
```



3.4 Ranked categories to confirm correlations

3.4.1 Ranked CFR categories to support categorical correlations

With increased rank in case_fatality_rate (CFR), or cfrRank, cfrCat 1-3 categories' means and other descriptive statistics show consistently decreasing incidence_rate, increasing fatality_rate (and frRank), decreasing gdp_pps, increasing L_pneumDeath, decreasing healthExp_pi, and increasing rreg, supporting both confirmed and some expected but unconfirmed correlations, except for no correlation with popDen, which has the unexpected, lowest popDen corresponding to the highest CFR.

In [5]:

```python
# get rankings in case_fatality_rate and fatality_rate, respectively
df['cfrRank'] = df['case_fatality_rate'].rank()
df['frRank'] = df['fatality_rate'].rank()

from pandas import Categorical
# divide 22 countries into three groups (8, 8, 6)
top_cat = 8
low_cat = 16

# assign category
df['cfrCat'] = Categorical(np.where(df['cfrRank'] <= top_cat, 1, 2))
df['cfrCat'] = Categorical(np.where(df['cfrRank'] > low_cat, 3, df['cfrCat']))
df['frCat'] = Categorical(np.where(df['frRank'] <= top_cat, 1, 2))
df['frCat'] = Categorical(np.where(df['frRank'] > low_cat, 3, df['frCat']))

pd.set_option('display.max_rows', 500)
df.groupby('cfrCat').mean()
df.cfrCat.value_counts()
df.groupby('cfrCat').describe().transpose()
```

Out[5]:

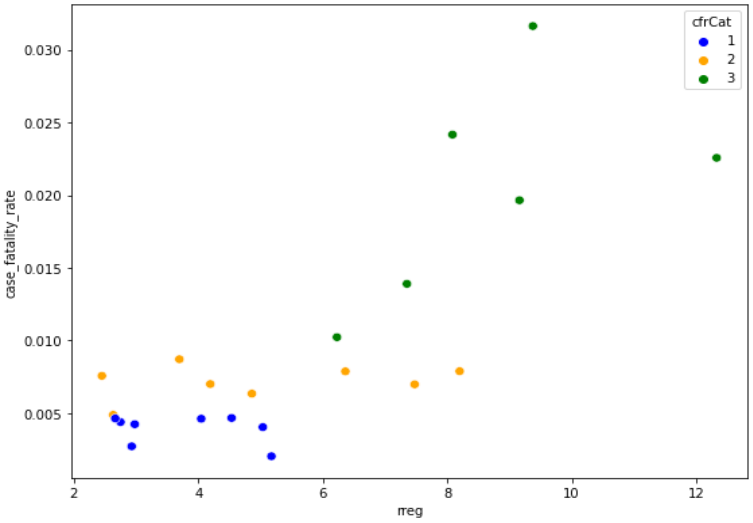
	case_fatality_rate	incidence_rate	fatality_rate	popDen	gdp_pps	L_pneumDeath	healthExp_pi	rreg	cfrRank	frRank
cfrCat										
1	0.003923	44335.000000	169.125000	146.000000	138.875000	1.210387	3773.967500	3.764528	4.5	6.000000
2	0.007160	36983.250000	264.000000	296.500000	94.000000	1.404277	2432.256250	4.983977	12.5	11.625000
3	0.020353	22142.166667	405.833333	96.333333	72.333333	1.463758	879.796667	8.755986	19.5	18.666667

3.4.2 Ranked CFR categories to confirm exemplary categorical correlations

Both scatterplots and boxplots by cfrCat support categorical negative correlation between healthExp_pi and case_fatality_rate and positive correlation between rreg and case_fatality_rate.

3.4.3 Ranked FR categories to support categorical correlations

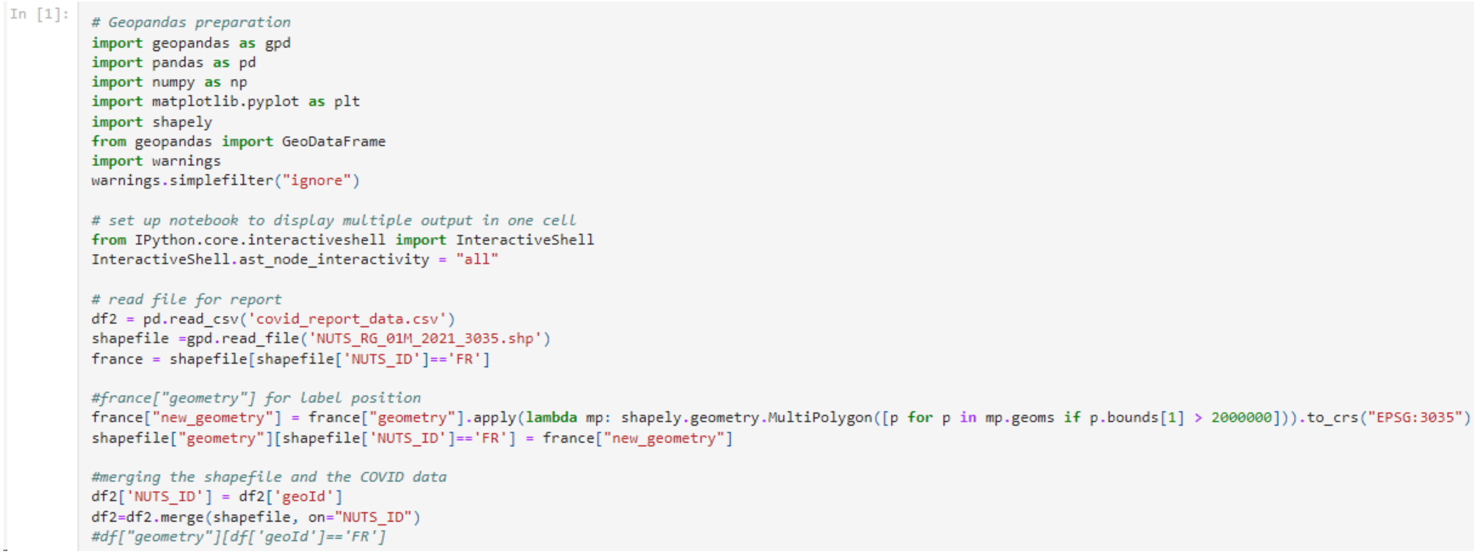
With increased rank in fatality_rate (FR), orfrRank, frCat 1-3 categories' means and other descriptive statistics show consistently increasing fatality_rate (and frRank), decreasing popDen, decreasing gdp_pps, increasing L_pneumDeath, decreasing healthExp_pi, and increasing rreg, supporting both confirmed and some expected but unconfirmed correlations, except for no correlation with incidence_rate, which has the unexpected, lowest incidence_rate corresponding to the highest FR.....................................



3.4.4 Ranked FR categories to confirm exemplary categorical correlations

Both scatterplots and boxplots by frCat support categorical positive correlation between case_fatality_rate and fatality_rate and negative correlation between gdp_pps and fatality_rat

3.5 Geo heatmap to support confirmed correlations



In [1]:

```
# Geopandas preparation
import geopandas as gpd
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import shapely
from geopandas import GeoDataFrame
import warnings
warnings.simplefilter("ignore")

# set up notebook to display multiple output in one cell
from IPython.core.interactiveshell import InteractiveShell
InteractiveShell.ast_node_interactivity = "all"

# read file for report
df2 = pd.read_csv('covid_report_data.csv')
shapefile =gpd.read_file('NUTS_RG_01M_2021_3035.shp')
france = shapefile[shapefile['NUTS_ID']=='FR']

#france["geometry"] for label position
france["new_geometry"] = france["geometry"].apply(lambda mp: shapely.geometry.MultiPolygon([p for p in mp.geoms if p.bounds[1] > 2000000])).to_crs("EPSG:3035")
shapefile["geometry"][shapefile['NUTS_ID']=='FR'] = france["new_geometry"]

#merging the shapefile and the COVID data
df2['NUTS_ID'] = df2['geoId']
df2=df2.merge(shapefile, on="NUTS_ID")
#df["geometry"][df['geoId']=='FR']
```

3.5.1 Geo heatmaps to reveal countries' ranking in variables

Geo heat maps generally support that the high-ranked countries in CFR and FR are those countries which joined EU more recently, with low gdp_pps, low healthExp_pi, and high rreg.

3.5.2 Geo heatmaps to reveal countries' ranking in variables

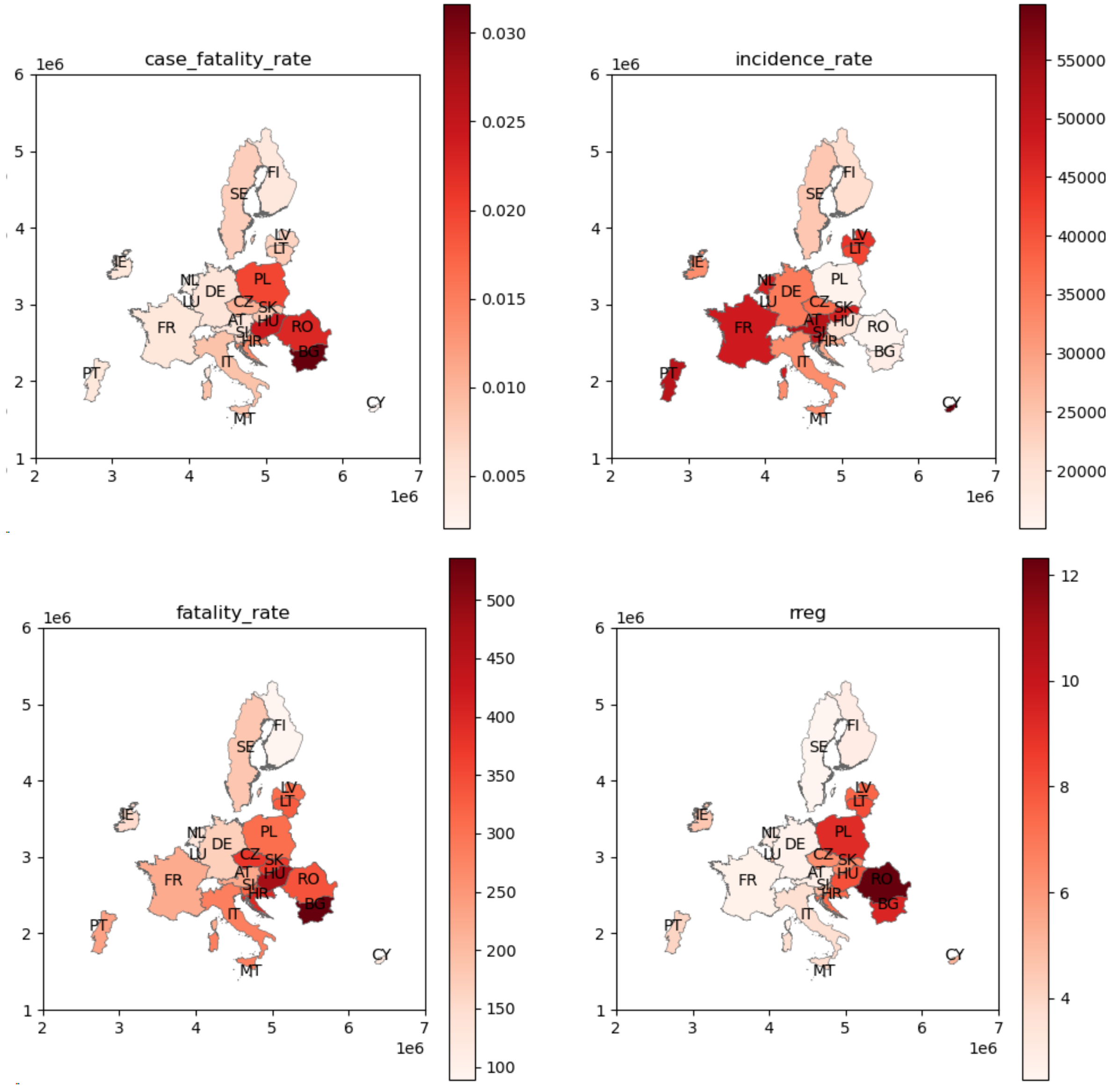
Geo heat maps generally support that the high-ranked countries in CFR and FR are those countries which joined EU more recently, with low gdp_pps, low healthExp_pi, and high rreg.

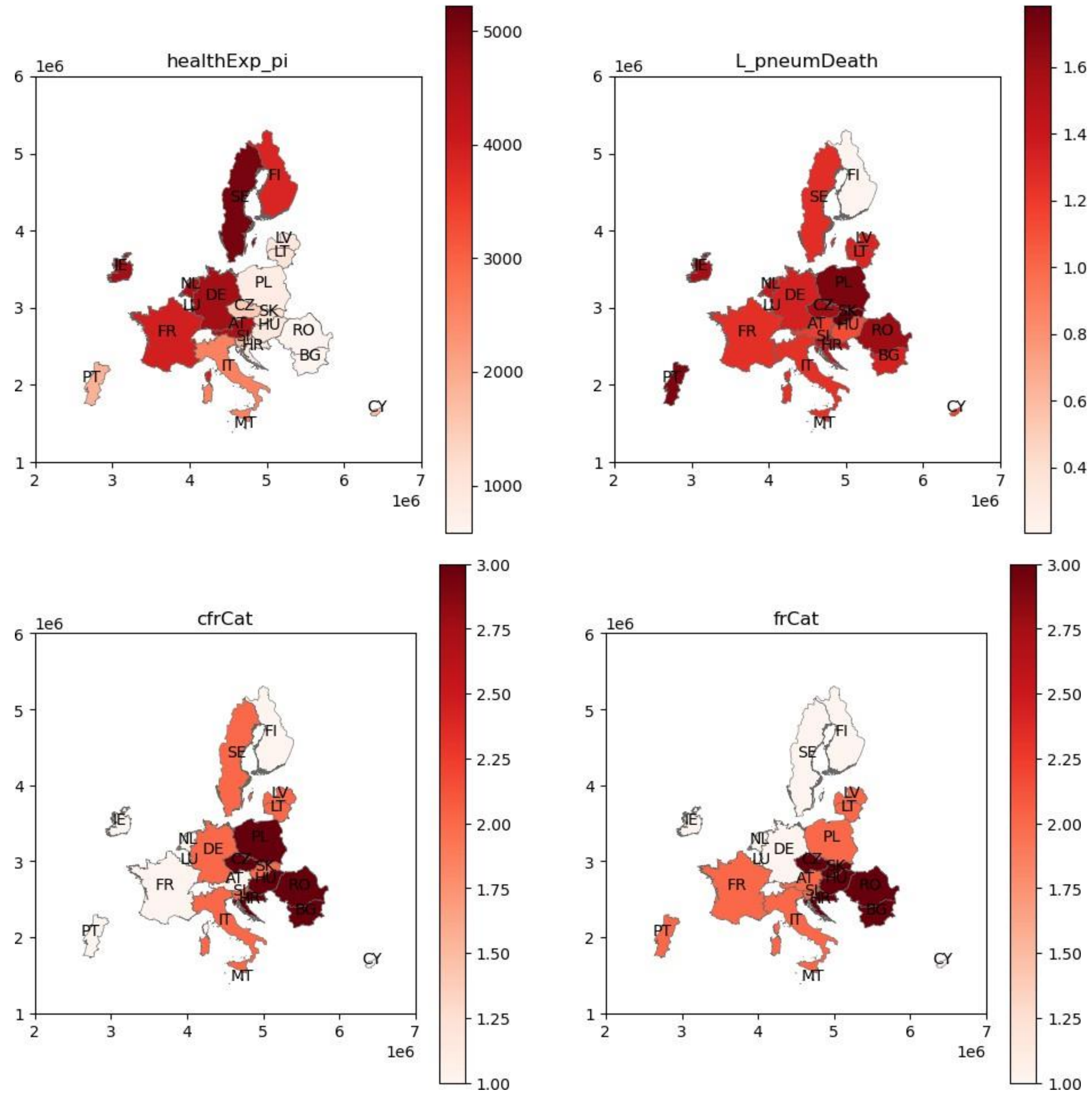
In [2]:

```
#converting the df to be compatible with geopandas plot
gdf = GeoDataFrame(df2)

#plotting the geopandas heatmap
fig, (ax1, ax2) = plt.subplots(ncols=2, figsize=(12, 6))
plt.xticks(rotation=0)
gdf.plot(column="case_fatality_rate", cmap="Reds", linewidth=0.4, ax=ax1, edgecolor=".4", legend = True)
ax1.set_title('case_fatality_rate')
ax1.set_xlim(2e6, 7e6)
ax1.set_ylim(1e6, 6e6)
#ax1.axis('off')
gdf.apply(lambda x: ax1.annotate(text=x['geoId'], xy=x.geometry.centroid.coords[0], ha='center'), axis=1);
gdf.plot(column="incidence_rate", cmap="Reds", linewidth=0.4, ax=ax2, edgecolor=".4", legend =True)
ax2.set_title('incidence_rate')
ax2.set_xlim(2e6, 7e6)
ax2.set_ylim(1e6, 6e6)
#ax2.axis('off')
gdf.apply(lambda x: ax2.annotate(text=x['geoId'], xy=x.geometry.centroid.coords[0], ha='center'), axis=1);
```

3.5.3 Geo heatmaps to support countries' confirmed correlations

With increased rank in cfrCat and frCat, these countries show consistently increasing case_fatality_rate (and cfrRank), increasing fatality_rate (and frRank), decreasing gdp_pps, increasing L_pneumDeath, decreasing healthExp_pi, and increasing rreg, supporting both confirmed and some expected but unconfirmed correlations.





SUMMARY

Provide a summary of your project. Were there any difficulties with data or the project as a whole? Provide your conclusions. What are possible next steps. Save your notebook to HTML with the naming format of lastname_Executive_Summary.HTML Upload your HTML file.

Summary: These rates of interest, CFR and FR, are confirmed to negatively correlate to gdp_pps, healthExp_pi and positively correlate with rreg (a new variable created in this project, 'regretion ratio'). popDen is thought to positively correlate to incidence_rate and incidence rate positive correlate to fatality rate, which are not confirmed in all cases.

Conclusion: in addition to low gdp_pps and healthexp_pi correlating with high fatality_rates, 'rreg' reflects health exp relative to gdp_pps. So the plan for healthcare is important.

Future: those with high incidence rate have low death rate worth more studies. popDen's unexpected neg correlation may be related to low medical care availability. More factors need to considered for futire studies.