Project Report

# GitHub URL

<https://github.com/PBlack1979/UCDPA_patrickblack.git>

# Abstract

This project sought to provide an initial analysis of health system preparedness for dealing with Covid-19 across European Union (EU) countries. To do this, annual health expenditure per capita was analysed for the 3 year period of 2017-19 and compared with cases and deaths due to Covid-19 up until present day. Jupyter notebooks was used to write Python code to import, prepare, analyse and visualise several datasets and provide insight into the data.

# Introduction

This project analyses three years of health expenditure data in EU countries alongside Covid-19 confirmed cases and deaths data. Given the continued prevalence of Covid-19 and impact that it has had on the whole world, further understanding is required to identify what level of nation state health expenditure (per annum) would ensure a nation states health system is adequately prepared to effectively manage further outbreaks of similar infectious diseases. Whilst many government policies (eg. Health, Migration, Economy) across nation states may have differed significantly during the two years in which the world was most directly impacted by Covid-19 (ie. 2020 and 2021) this project specifically considers health expenditure per annum in the three years of 2017-2019, in order to assess nation states health systems preparedness for dealing with a significant adverse health event ie. a global pandemic.

# Dataset

I used the following datasets for this project;

**Health Expenditure**

**Source:** Organisation for Economic Co-operation and Development (OECD)

**Link:** <https://data.oecd.org/healthres/health-spending.htm>

**Description:** The health expenditure data is a measure of all expenditure on health care

goods and services for a large range of countries from the 1970’s up to current day. I chose this source as the OECD is a reputable intergovernmental organisation which produces datasets to support government policy making.

**European Union Population**

**Source:** Eurostat

**Link:** <https://ec.europa.eu/eurostat/databrowser/view/tps00001/default/table?lang=en>

**Description:** The population data indicates the number of people residing in the

respective EU countries on 1st January 2022.This data was selected as Eurostat is the European Commission’s primary organization with responsibility for producing statistics and EU population data was required in order to calculate per capita rates of Covid-19 in EU countries.

**27 Countries of the European Union**

**Source:** European Union

**Link:** <https://european-union.europa.eu/principles-countries-history/country-profiles_en>

**Description:** I prepared an EXCEL spreadsheet that detailed the 27 nation states of the ,

EU. This was required to enable analysis across all the EU countries.

**Covid-19 Data**

**Source:** Postman

**Link:** <https://documenter.getpostman.com/view/10808728/SzS8rjbc>

**Description:** This source was chosen as it provided current Covid-19 data that was

available for a large number of countries, including all EU countries, via an Application Programming Interface (API).

**Current Exchange Data**

**Source:** GigHub account

**Link:** <https://github.com/fawazahmed0/currency-api#readme>

**Description:** This source was chosen as it provided current exchange rate data for

converting USD, which Health Expenditure data was obtained for, into Euro and was available to import via an API.

# Implementation Process

I used Jupyter Notebooks to programme in Python – the data file is split into the five relevant ‘milestone’ sections as outlined below.

**1) Data**

1a): I imported NumPy, Pandas, Requests, Matplotlib and Seaborn into Python in

order to be able to complete all programming codes.

1b) & 1c): Multiple real-world datasets were imported. I imported two types of flat file, a

CSV document and two EXCEL document by using the relevant read code (pd.read\_csv & pd.read\_excel). Whilst to import data from the two separate API sources, I had to use code from the Requests module to send a GET request to retrieve the data, then parsed it using json() and finally converted it to a dataframe using pd.DataFrame.

**2) Importing (including Data File Storing & Manipulating)**

The Python code print() was used extensively throughout many steps below and for each dataset the code info(), head() and shape were used to get a better understanding of the structures of the data sources. Furthermore;

2a): **Health Expenditure Data:** drop\_duplicates() and count() to identify that data for 53 nation states was contained across 52 years.

2d): **Covid-19 Data:** sum() to get the total deaths across all countries and max() to obtain the highest deaths in one country. I created and stored several additional dataframes that filtered the data based on logical operators (== or <100) and also used sort\_values() to view sorted data. I also used NumPy aggregations (np.sum, np.min, np.max) for confirmed cases of Covid-19.

2e): **Exchange Rate Data:** sort\_index() was used to view the data in a descending alphabetical order.

**3) Preparation (includes Dictionary/Lists, Conditional Statements, Looping & Groupby)**

3a): **EU 27 Data:** To group the data I entered a LIST of the twenty Euro

countries. Using a FOR loop along with a conditional statement (IF and ELSE), I created a new column of data (using append()) onto this dataset that indicated whether each country used the Euro or not. I then created several further dataframes from which I validated (using count(), groupby() & get\_group()) that the 20 Euro countries were successfully identified in the updated dataset.

3b-1): **Health Expenditure Data – Identifying Average Annual Expenditure:** I

combined logical (==) and Boolean (&) operators along with sort\_values() to manipulate this dataset. Having identified that 2017, 2018 & 2019 all contained data for 46 countries, I then used isin() to filter the dataset and used drop(columns) to only keep relevant data. To obtain several statistics, I created multiple extra dataframes that used groupby() and performed either a single calculation (mean()) or used the agg method for multiple calculations (agg([max,min])).

3b-2): **Health Expenditure Data – Preparation of DataFrame for Merging:** I

combined logical (==) and Boolean (&) operators to create a dataframe which contained filtered data that displayed the metric I required (USD per capita). I then created several variables (Filter\_2017, Filter\_2018 & Filter\_2019) which were applied to additional dataframes. To keep only relevant data, drop(columns) was applied to these dataframes aswell as using rename(columns) to format column headers. I then used merge(how=left) to combine data from 3 separate dataframes into one wider dataframe that displayed the expenditure across several columns. Two further new calculated columns were created to summarise the data. Following this, apply(str) was used to change a data type to a string to enable a message for each country to be displayed in the table that contained text and numerical values. Finally, I used merge(how=right) to merge Health Expenditure dataframe with the EU 27 dataframe and hence extract health expenditure data for the 27 countries of the EU only.

3c): **EU Population Data – Removing Duplicates & Replace Missing Values:**

I used groupby(), count() and drop\_duplicates() to identify that data was provided for only 22 of the 27 EU countries from the imported data source. Using merge(how=right), the EU Population data file was merged with the EU 27 data file and isna() used to confirm that there were missing values. The code fillna(0) was to replace missing values. To improve display of the dataset set\_index to index the data and sort\_index() then used to display which countries did or did not use the Euro.

3d): **Covid-19 Data:** I used drop(columns), set\_index and sort-values to

manipulate the dataset into a dataframe that was suitable for merging. I then used merge(how=right) to combine this dataframe with the EU 27 dataframe whilst also used melt to reshape the dataframe.

3e): **Exchange Rate Data:** I used loc to slice the created dataframe to display

the relevant Euro to USD exchange rate, which I then entered as a float into a new column on the ultimate dataframe created for health expenditure data

**4) Analysis**

4a): **EU Population Data – Updating Missing Data:** a DICTIONARY was

created that contained population data for the missing countries and pd.DataFrame was used to convert this into a dataframe. Following the created and manipulation of several further dataframes, several NumPy functions (np.max & np.min) using the agg() method were calculated and a custom function was defined to calculate the difference between the maximum and minimum populations within the EU.

4b): **Health Expenditure Data:** after formatting the dataframe using round(), several calculations including NumPy functions on the data were completed (max(), min(), np.mean, np.median). To further analyse the data rank() and quantile() were used and a FOR loop was created (that included conditional statements IF, ELIF & ELSE) to generate a new column in the dataframe that indicated the country level of health expenditure.

4c): **Covid-19 Data:** the dataframe had several new columns calculated and

added as this enabled more comparable ‘per capita’ figures to used for analysis. The agg() method was used twice to calculate several NumPy functions (np.min & np.max) and another custom function was created, this time to calculate the percentage difference between the country with the lowest Covid-19 deaths per 10,000 people and the highest.

# Results

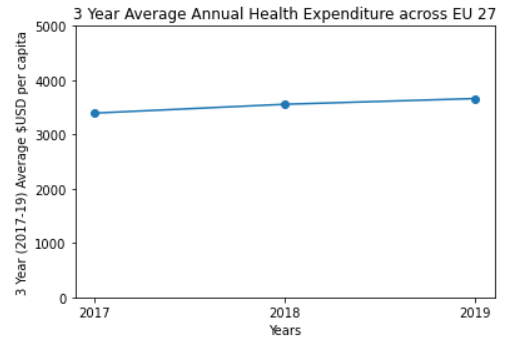
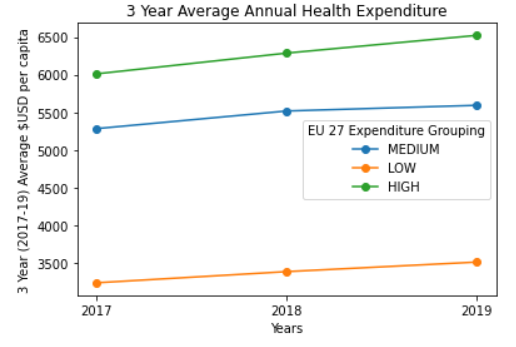
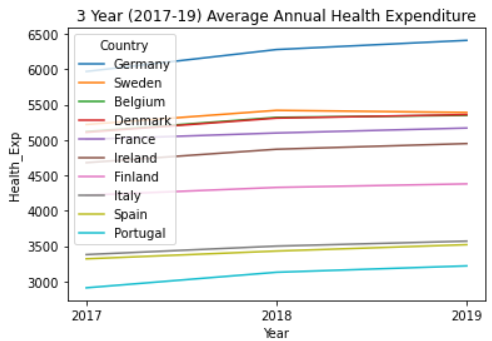
**5) Visualisation**

5a) – c): I used a combination of logical and Boolean operators, groupby(), creation

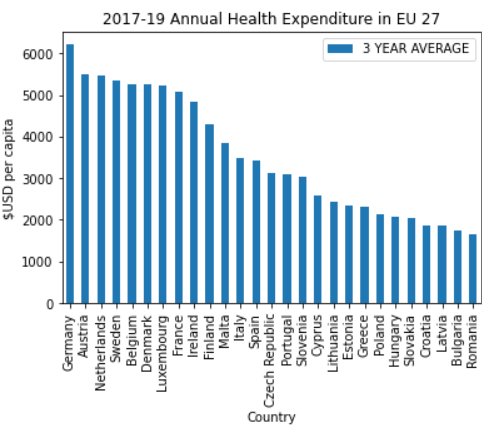
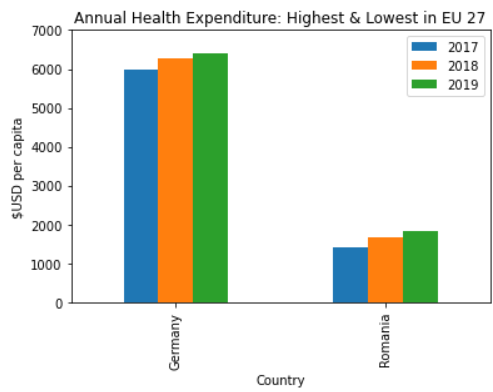
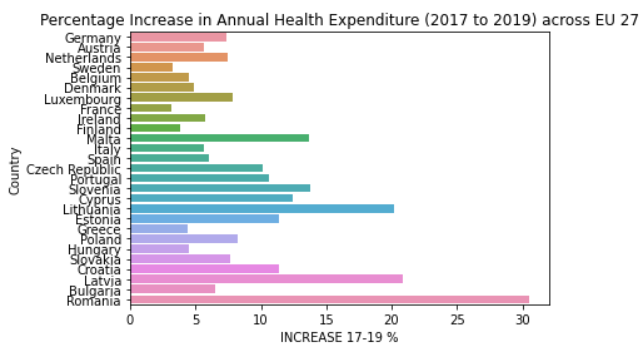
of new columns, sort\_values and merge() to manipulate the created dataframes so that the data was prepared to visualise in a variety of charts.

5d) – 5f): I created over 20 charts using both Matplotlib and Seaborn, including

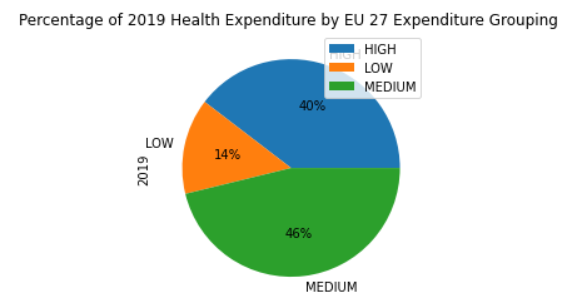
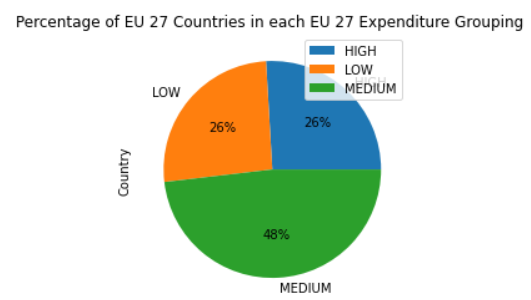
**Line Chart**: plt.plot & sns.lineplot

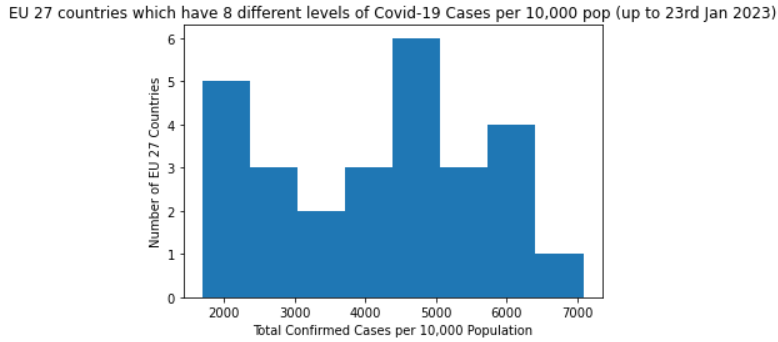
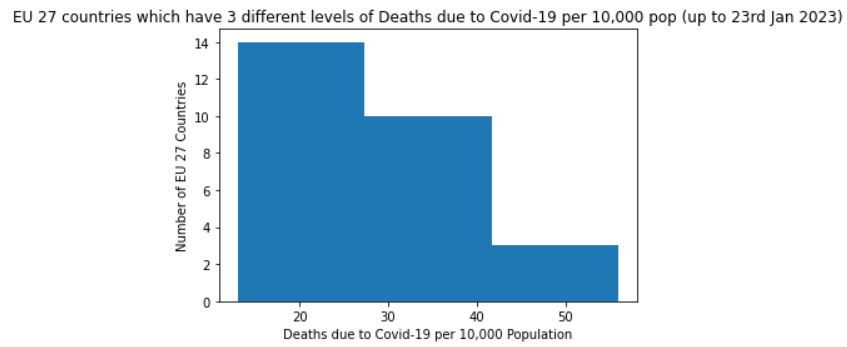
**Bar Chart:** plt.bar, plot(kind=bar) & sns.barplot

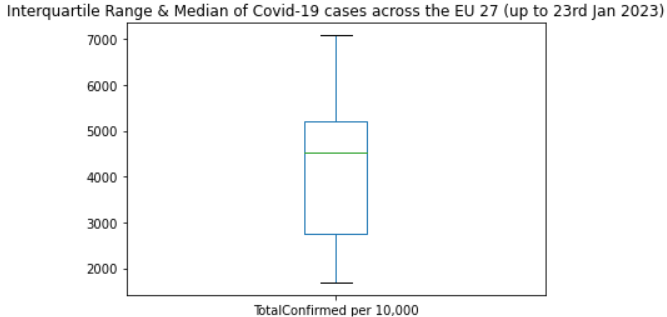
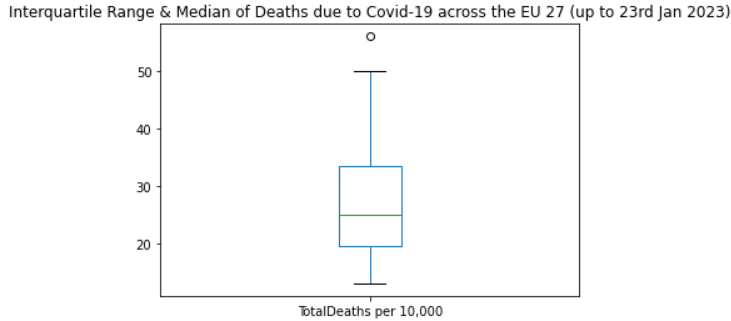
**Pie Chart:** plot(kind=pie)

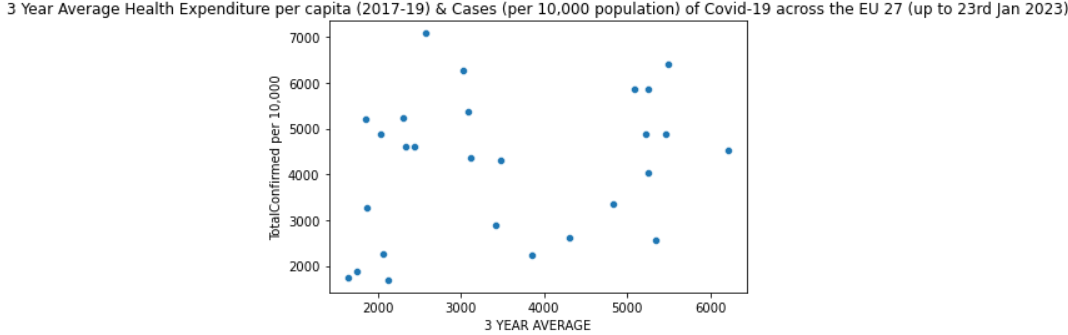
**Histogram:**plt.hist

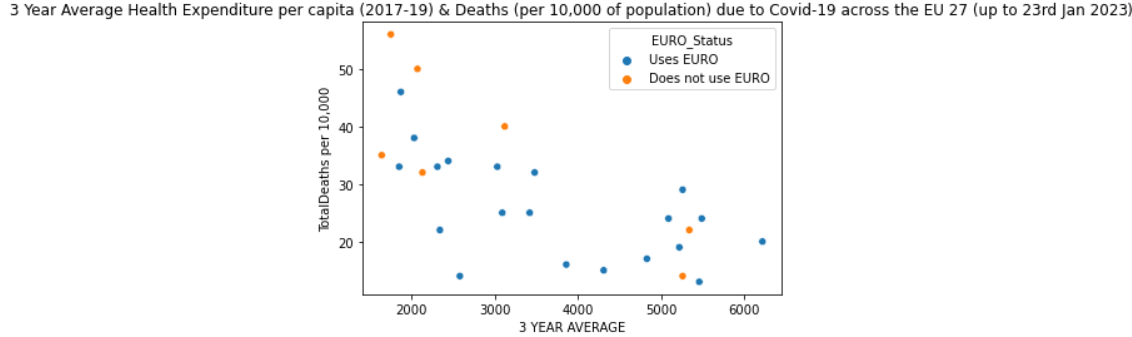
** **

**Box Plot:** plot(kind=box)

** **

**Scatter Plot:** plt.scatter & sns.scatterplot





# Insights

1. **Refer to 5d) Health Expenditure Data:** The line charts indicate that in the 3 years prior to the Covid-19 pandemic, health expenditure (USD per capita) was increasing in all EU countries. However, the bar charts indicate that those countries with annual health expenditure in the 1st quartile (categorized as “Low”) had the highest annual percentage increase in health expenditure. This may be because these countries had a low base of health expenditure from which to increase from and also EU budget redistribution funding may be helping to support these countries to significantly uplift their health expenditure to average EU levels.
2. **Refer to 5e) Health Expenditure Data:** The bar charts indicate that Germany had the highest average annual health expenditure (USD per capita) between 20217-19 whilst Romania had the lowest. Whilst the increase in annual health expenditure from 2017 to 2019 (in quantity and percentage terms) would not seem to be directly related to the level of annual health expenditure in each country, the pie charts show that the 7 countries with “High” health expenditure (26% of the 27 countries) account for 40% of total annual health expenditure across the EU 27. This would be due to the “High” health expenditure countries being stronger economically developed countries and hence being able to spend a higher amount on annual health care per person in their country.
3. **Refer to 5f) Covid-19 Data:** The bar charts indicate that the larger of the EU 27 countries had more cases and deaths due to Covid-19. The scatter plots indicate that a stronger direct correlation between size of country population and confirmed cases exists when compared to the correlation between confirmed cases and deaths due to Covid-19. As Covid-19 was an easily transmissible infectious disease, it would be logical that much larger populations would have a higher number of cases of Covid-19. However, the link between Coivd-19 cases and deaths may be less clear due to national differences in government policies (eg. Covid-19 Testing, Recording a death as due to Covid-19) which meant the data does not represent directly comparable statistics across EU members.
4. **Refer to 5g) Covid-19 Data:** The histograms and box plots indicate that there was a larger spread in the number of Covid-19 cases across the EU 27 as the interquartile range was larger. Whereas for deaths data, there would appear to be a tighter distribution of EU countries around the median. This may be explained by the differences in non-health government Covid-19 policy (eg. Bars/Restaurants or workplaces) being more divergent between the EU 27 countries (leading to more or less Covid-19 cases) whereas standards in specific health settings (eg. Hospitals) are consistent in treatment methods for a disease.
5. **Refer to 5h) Covid-19 & Health Expenditure Data:** The two scatterplots indicate a difference in the correlation between annual health expenditure and Covid-19 data (cases or deaths). No clear correlation exists when using the cases data, perhaps due to data quality issues, whereas a direct correlation is indicated between the level of annual health expenditure and deaths data. With a higher level of investment in a country’s health system (prior to Covid-19), then the country would have been better prepared to effectively manage the impact of Covid-19 and hence reduce the number of deaths per capita.

Machine Learning

In the future, this type of analysis could lead to giving insight into what the optimal level of health expenditure would be to achieve a health system that is effective in dealing with a global pandemic. More detailed health expenditure and Covid-19 data, over a longer period of time and from more sources could be used to develop a health system model which could be used for machine learning to predict future health outcomes.

Regression analysis could initially be used as numerical data for health expenditure and Covid-19 is more easily obtainable and more objective when compared to trying to categorise government policy decisions across multiple countries.