Code First Girls Data Nanodegree

Project Report

Team Members: Sophie Owen **[Solo Project]**

Introduction

Project Goal

The goal of the project was to determine whether there is a correlation between climate conditions of a country and its Economic Indicators. This can then be used as a tool to:

1. Persuade world leaders to take climate change more seriosly. The more temperatures and climate conditions fluctuate, the more of a problem this could be for them politically end economically.
2. Inform business owners where in the world and in what climates their companies may thrive the most (if they are dependent on that).

The project aims to be used as a tool for informing people and for confirming assumptions written about in scientific papers, as well as predicting the effect of Climate Change on the Economic stability of countries around the world.

Project Requirements

Project requirements which were met include:

- Using more than one API to collect and fetch data
- Use of key scientific packages: Numpy, Matplotlib and Pandas

Optional Additions

As well as the required features, the following were also included:

- Urllib package to access APIs which are commonly used to quickly fetch data required
- Refactoring through use of functions to increase the efficiency of the code and keep it looking tidy
- Use of seaborn to further visualise data
- Use of folium to also visualise data
- Scipy to build models using linear regression and assess the quality of the models
- Plotting logarithmic lines of best fit to visualise trends in data
- Use of subplots and heatmaps

The Aims and Objectives of the Project Code

1. To identify links between economic indicators of countries and their climates
2. To identify correlations between economic indicators themselves
3. To access and process data on:
    a. Average temperatures of countries and the maximum and minimum temperatures they reach throughout the year
    b. Economic indicators of countries such as GDP, Population and Unemployment.
4. Exploratory Analysis of the data, looking at links between economic indicators and other economic indicators for each country
5. To visualise all findings in graphical form

Roadmap of the report

1. **Introduction:** Aims, objectives and roadmap of the report.
2. **Background:** Context of the analysis.
3. **Steps Specifications:** The process followed in the project to identify suitable data sources, access data, pre-process and analyse. Overview of the project plan.

4. **Implementation and Execution:** The modules and packages that were used and why and where they were implemented. Achievements, challenges and implementation approach.
5. **Results Reporting:** Findings from the analysis.
6. **Conclusion:** Review of the project.

Background

The book 'Prisoners of Geography', suggests that the geographical properties of countries have a huge impact on the politics of the country, including likelihood of going to war and relationships with other countries. For example, India and China have a long border between them and have had many reasons to go to war previously. However, they have only had one, one-month long war in history. It is likely this is caused by the presence of one of the largest mountain ranges in the world between them – the Himalayas. On the flipside, Russia and Ukraine have one of the flattest plains of land along their borders, which is perhaps why they have had such conflicts which have so easily led to fights. Flat land means no difficult terrain to traverse to launch an attack on the other country.

It is because of this strong link between land properties and politics that the current topic is being explored: the link between climate and economic success. Many articles have been published on supposed links between higher temperatures impacting economic growth. The propose the following[1]:

1. Higher temperatures reduce economic growth in poorer countries
2. Higher temperatures reduce growth rates
3. Higher temperatures impact agriculture, industry and political stability

However, many papers which exist on this topic are based on assumptions and basic statistical analysis. This project builds on this with more in-depth analysis.

The findings of this project will be used as a tool for governments and business owners to plan with more justification. It will help climate-sensitive companies, by informing them where their business is likely to be most successful. It will do this by assessing GDP per Capita and the effect climate has on this. It can also be used as a tool to persuade governments into thinking more seriously about climate change. Global temperatures are increasing, and the findings of this report may show that higher temperatures are linked to lower GDP per capita and potentially higher unemployment and smaller populations. This is not optimal for governments.

Steps Specifications

Framing Questions

When the topic was decided, initial questions were generated to help determine what data needed to be collected and how it needed to be processed. However, it was decided these would be confirmed once pre-processing had been completed as this would then enable initial analysis to be done and initial insights to be found. It was also realised that more questions would come up as analysis was done, so the questions were kept as topics to be explored and were open to change. The questions changed most drastically at the start when searching for data sources, as the original aim was to find data sources on crop growth and explore how this is impacted by climate. However, while searching for data the current topic was explored and found to be more interesting with data more readily available.

Once the data had been processed and was ready for analysis, the topic was researched further, and the most important economic indicators identified. These could then be compared to climate.

Gathering Data

Gathering data began by searching for reliable data sources online. The IMF was used to provide data on economic indicators, using the World Economic Outlook (WEO) report and data source. This contains economic data on every country in the world. This database was downloaded from the IMF website as a .csv

[1] (Dell, et al., 2011)

file and imported into the Python script. Once the data had been pre-processed (more detail on this below) a list of the country names was extracted. This was then used to access weather data.

To get weather data the search started on 'RapidAPI' for weather APIs as this has a lot of helpful data. Reviews and recommendations were used to identify the best API and various APIs were tried out, including DarkSky, Open Weather Map and Stormcloud. However, many of these only provide realtime weather data and very few provide historical data without a subscription to the service.

Meteostat API has historical weather data from as long ago as 1980, up to the present day. This requires a subscription to make calls to the API a certain number of times and a certain frequency but is the most reliable and one of the only ones which provide accurate and free historical data.

To access the data for a variety of locations, the weather station ID is required. Therefore, another endpoint of the Meteostat API was used, 'nearby', to get the station ID of the closest station to given coordinates. To get the given coordinates another API was used using the 'urllib' package. This was called 'Open Street Map' and provides the coordinates for a given country or city name. The country names were extracted from the 'Country' column in the WEO dataframe. It was decided Capital cities should be used to get the station ID as this would be the economic centre of the country. Another .csv file was downloaded from online which was a table of country names, Capital cities and their continents. This was imported into the python script as a dataframe and a new column was added which concatenated the Capital city names with the Country, to create a list of accurate locations which could be passed through the Open Weather Map API.

Pre-Processing

Exploratory analysis was extremely helpful when pre-processing all sets of data. Pre-processing steps for each data source imported and accessed were required to go through data cleaning, then conversion into a dataframe which could be accessed using functions with little additional input from the user.

Data cleaning was the first step and involved:

- Identifying null or NaN values and replacing or removing them
- Changing the type of data in the rows of the dataframes, i.e. float or string, and ensuring they were as required for processing. Many numerical values were found to contain commas which prevented them from being converted to floats. Therefore the commas needed to be removed.
- NaN values were removed and a maximum requirement was set for countries and their variables to be included in further analysis.

Exploratory analysis included:

- Heat mapping based on correlation coefficients
- Plotting graphs to visualise correlations
- Applying the shape, info, describe and size methods to data to understand it more clearly

In-Depth Analysis

As there was a huge amount of data to look at, it was important to start by researching relationships internally within a data frame and between data frames. It was determined whether certain economic indicators affected other economic indicators, e.g., Population and its effect on GDP per capita, so the link between temperature and GDP wasn't biased. The links between minimum, maximum and average temperatures and the range between minimum and maximum was explored, indicating the extremity and fluctuations of climate.

Correlations were identified using heatmaps and inbuilt libraries. These were useful to gain initial insights into potential relationships that could be explored.

Models were built to show the strength of the relationship between different variables and to predict the GDP per capita of other countries not included in the analysis, based on their temperatures. Testing and Training sets were helpful to build and validate models.

Graphs were plotted to visualise the results and identify patterns of interest.

Testing

As the project was an individual project, time was extremely limited and there was no time to write code to officially test the functions. Given more time, tests would be written o test all functions such as

However, as the code was written the functions were tested briefly. For example, when writing a function to access a specific variable of a particular country, a couple of WEO Subject Codes and countries were used to test the output from this function. Outputs were regularly printed to check they looked as expected and contained the data required, and the whole code was run from the top to test the order was correct.

Data Sources

Economic Indicators

Data on Economic Indicators was found on the *International Monetary Fund* website. Every six months, in April and October, they release reports called the 'World Economic Outlook' (WEO) which details economic data on every country, year by year. They release a full report on findings as well as the database of economic data, which can be downloaded. This was the preferred option as it could then be imported into the Python script for simple processing and analysis.

Economic indicators featured in the report include GDP Per Capita (in PPP, USD and local currency), Unemployment rate (as a percentage of the total work force), Population, Inflation, Government Debt, and more.

Originally the Economic Indicator data was attempted to be accessed through an API, but when searching for data sources online databases were found to be produced by the more reliable sources, such as the IMF.

Weather Data

Meteostat was chosen as the API to access weather data. The aim was to test many APIs out then decide which one to used based on which was the most successful. Meteostat was chosen as it is one of the only APIs that provides easily accessible historical weather data. It outputs a JSON object with two properties: meta and data. The meta output contains information on when the API was accessed and debugging information. The data output contains the weather information required to analyse the data. A key is required to access the data.

The station ID is required to access weather data for a specific location, and Meteostat provides an endpoint to access this from location coordinates. Another API was therefore required to find the coordinates of a specific location. Open Weather Map was used for this. It takes in the location name and outputs the coordinates of that location.

Implementation and Execution

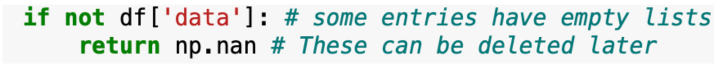
Development Approach and Team Member Roles

The project was completed as a solo project. This was due to previous commitments and travel over December and the New Year which would have made it challenging to work as part of a team. This presented many challenges, with time to complete it being the most difficult.

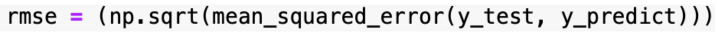
However, the project was planned in the same was as it would if it were a team project. It was split into stages, with each part being completed one after the other with slight overlap and iteration. The main stages were; finding data sources and defining the questions; accessing and processing the data; exploratory analysis and data cleaning; conversion of dataframes into files that can be accessed at any time without making more calls to the API; and detailed data analysis.

Tools and Libraries

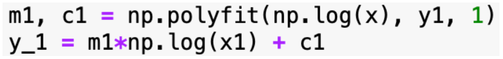
Numpy: Used for mathematical equations, data cleaning, curve fitting and finding correlation coefficients.



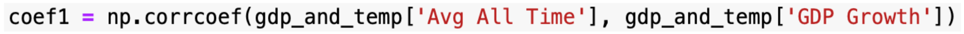
```
if not df['data']: # some entries have empty lists
    return np.nan # These can be deleted later
```



```
rmse = (np.sqrt(mean_squared_error(y_test, y_predict)))
```



```
m1, c1 = np.polyfit(np.log(x), y1, 1)
y_1 = m1*np.log(x1) + c1
```



```
coef1 = np.corrcoef(gdp_and_temp['Avg All Time'], gdp_and_temp['GDP Growth'])
```

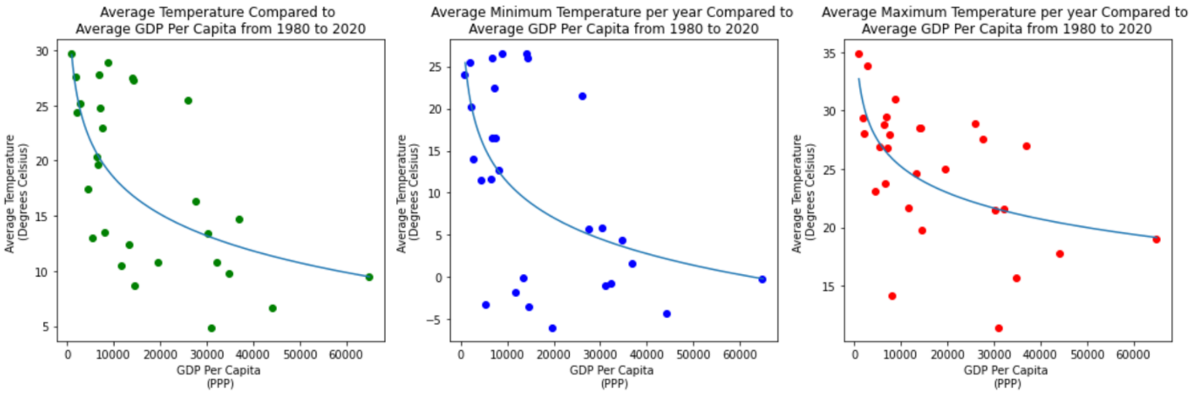
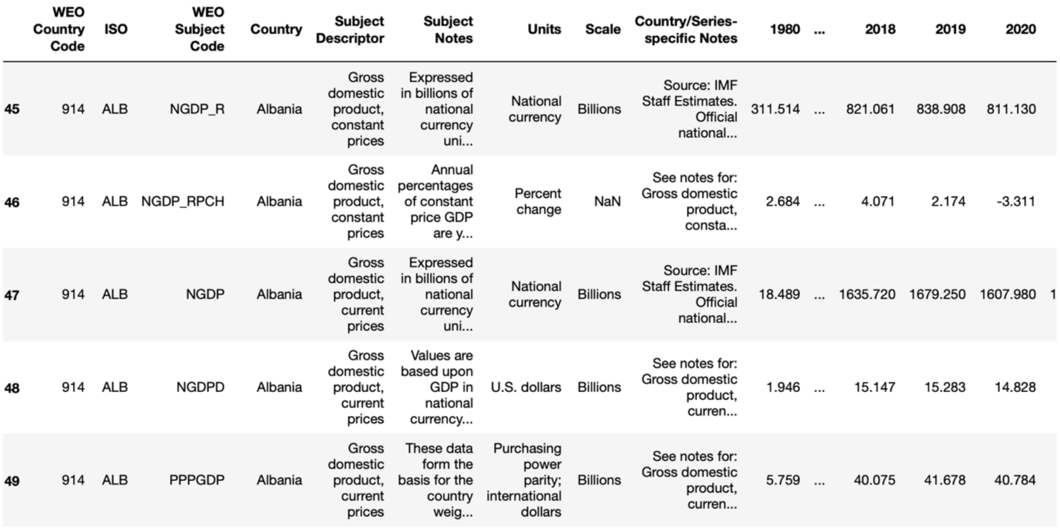


Figure 1: Graphs showing the relationship between average, minimum and maximum temperatures from 1980 to 2020, compared to GDP Per Capita. Lines of best fit are shown with an logarithmic relatiosnhip between the variables.

Pandas: The most used module in the project. Used for importing data, converting to data frames, organising data and combining data from different data sources, data cleaning, converting to .csv files and doing data analysis.



	WEO Country Code	ISO	WEO Subject Code	Country	Subject Descriptor	Subject Notes	Units	Scale	Country/Series-specific Notes	1980	...	2018	2019	2020	
45	914	ALB	NGDP_R	Albania	Gross domestic product, constant prices	Expressed in billions of national currency uni...	National currency	Billions	Source: IMF Staff Estimates. Official national...	311.514	...	821.061	838.908	811.130	
46	914	ALB	NGDP_RPCH	Albania	Gross domestic product, constant prices	Annual percentages of constant price GDP are y...	Percent change	NaN	See notes for: Gross domestic product, consta...	2.684	...	4.071	2.174	-3.311	
47	914	ALB	NGDP	Albania	Gross domestic product, current prices	Expressed in billions of national currency uni...	National currency	Billions	Source: IMF Staff Estimates. Official national...	18.489	...	1635.720	1679.250	1607.980	1
48	914	ALB	NGDPD	Albania	Gross domestic product, current prices	Values are based upon GDP in national currency...	U.S. dollars	Billions	See notes for: Gross domestic product, curren...	1.946	...	15.147	15.283	14.828	
49	914	ALB	PPPGDP	Albania	Gross domestic product, current prices	These data form the basis for the country weig...	Purchasing power parity; international dollars	Billions	See notes for: Gross domestic product, curren...	5.759	...	40.075	41.678	40.784	

Figure 2: An example of a pandas dataframe used in the analysis. This is showing the WEO data when first loaded into the source code.

Matplotlib: Used for visualising data. Plotting bar charts, scatter and line graphs, histograms and formatting graphs.

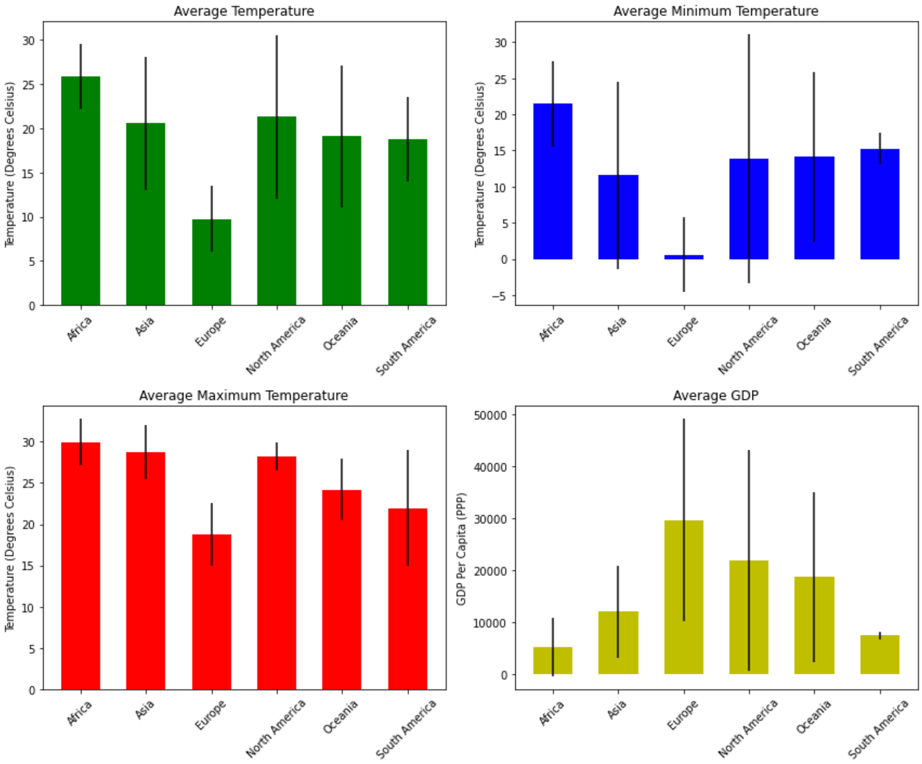


Figure 3: Bar Charts showing the average, maximum and minimum temperatures from 1980 to 2020 by Continent. The yellow bar chart shows the average GDP per Capita by continent in this time frame. Standard deviations are quite high.

Requests: Used to make API requests to the Meteostat API and Open Street Map. Creates a request to pull data from the API in JSON format to then process using pandas.

Urllib: This is a URL fetcher which can use various protocols. It was used for fetching coordinates of locations around the world, as this is a simple and common request. Using Urllib in this case is simpler than a full API request and uses much fewer lines of code.

Datetime: Used to convert the indexes of DataFrames to datetime format, so they could be cross-correlated and combined easily.

OS: Used to create folder and save files to these folders. Doing this meant that when processing the data later, the API request did not need to be made again, only the file from the folder needed accessing.

Folium: Visualises data that's been created in Python in an interactive map. This was used to visualise the spread of weather location data available and whether it was a fair representation of weather around the world.

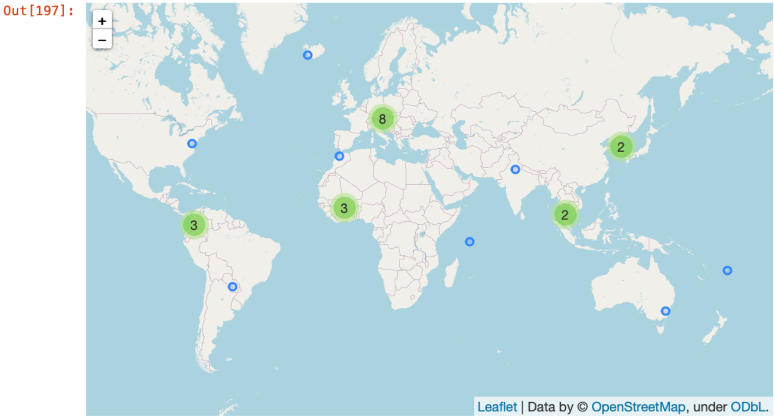


Figure 4: The world map created using folium. This pin points coordinates of the countries used for detailed analysis to visualise the spread of data.

Seaborn: Used to build heatmaps of correlations between variables within a pandas data frame. Useful during exploratory analysis to identify relationships between variables that could be explored in more detail.

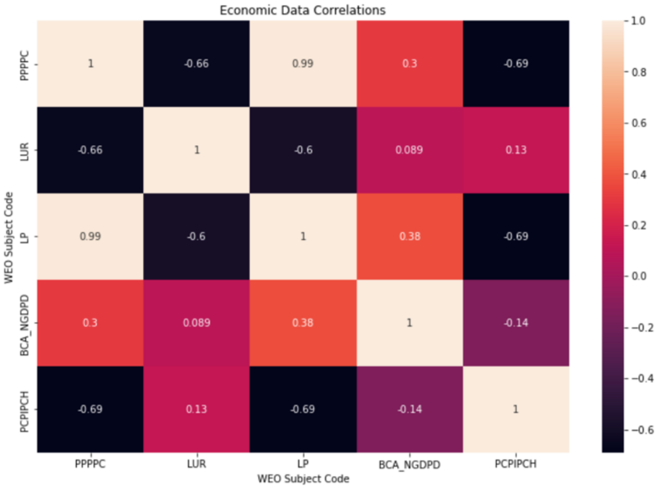
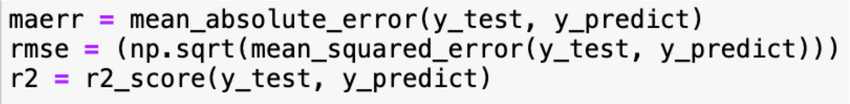
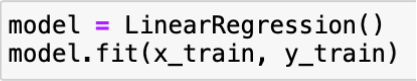


Figure 5: Heatmap using Seaborn showing the relationship between economic indicators of Australia

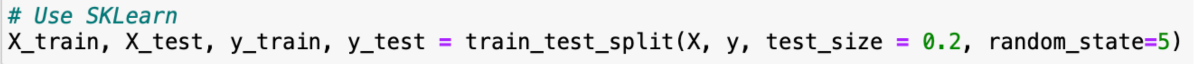
Sklearn: A machine learning library used to find correlations between sets of data and build models to predict other values. Then used to test the models that have been built.



```
maerr = mean_absolute_error(y_test, y_predict)
rmse = (np.sqrt(mean_squared_error(y_test, y_predict)))
r2 = r2_score(y_test, y_predict)
```



```
model = LinearRegression()
model.fit(x_train, y_train)
```



```
# Use SKLearn
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size = 0.2, random_state=5)
```

Agile development

As it was an individual project, making use of all team based agile elements was difficult. However, they were made use of where possible. For example, the whole project utilised an iterative approach. The code was initially written in comment form to provide an overall structure and to assess the flow. Then the code was built up from here. As each section was complete, it would be iterated on to improve the efficiency and simplify the statements.

When the next stage of code was written, the previous stage would then be reviewed again to identify areas for improvement, such as data frame structures and variable names. The code was then reviewed again at the very end. Where functions could be written to replace loops, they were. An example of one of the efficiency improvements was in switching from for loops to lambda functions. This meant that many lines of code could be replaced with just one, also speeding up the process.

Refactoring was used by rewriting snippets of the code in a new file and testing this before replacing the section in the main code where it was required.

Implementation challenges

Many challenges came up during the project, the main few of which are outlined below.

There were many challenges with the project, and it was a big learning curb. The first challenge was in the processing of the data. As the data came from multiple sources from different providers, ensuring the indexes of each data frame were consistent was key. There were some discrepancies between the country names, for example in one dataset the country was called “The Bahamas” and in the other “Bahamas”. This was solved early on as soon as the weather data was converted into a dataframe, to prevent coming up with problems later down the line such as an empty dataset when searching the WEO dataframe for the wrong country name.

Missing data and NaN values presented some problems in both datasets. There was a lot of missing data in the weather API calls which meant precipitation, humidity and more variables could not be analysed. This was tackled by counting the NaN values per country and selecting only those with 10 or fewer NaN values, representing less than 2% of the dataset. Missing values in the WEO spreadsheet were identified using the heatmap ‘Seaborn’ Module as described above. Fortunately, these corresponded with the missing values from the weather dataset and so would not be used anyway. To assess the spread of data from different countries, the locations with enough data points were plotted on a map using the ‘Folium’ module, showing there was a reasonable spread between locations to get a good picture of the properties of different countries around the world.

It was also difficult deciding which weather variables would be the most insightful to compare economic data against. Therefore, they were first compared against one another then all of them were compared against specific economic variables. Originally the weather data was plotted as an average value per year from 1980 to 2020, however there was such a tiny difference in temperature across this time (Figure 6) that it was decided it should be averaged across the whole 40 years, as this was more comparable. The GDP data was cross-correlated from each country, with correlations that had an R value of 0.8 and above, so it was decided this should also be averaged across the 40 years for ease of comparison and analysis.

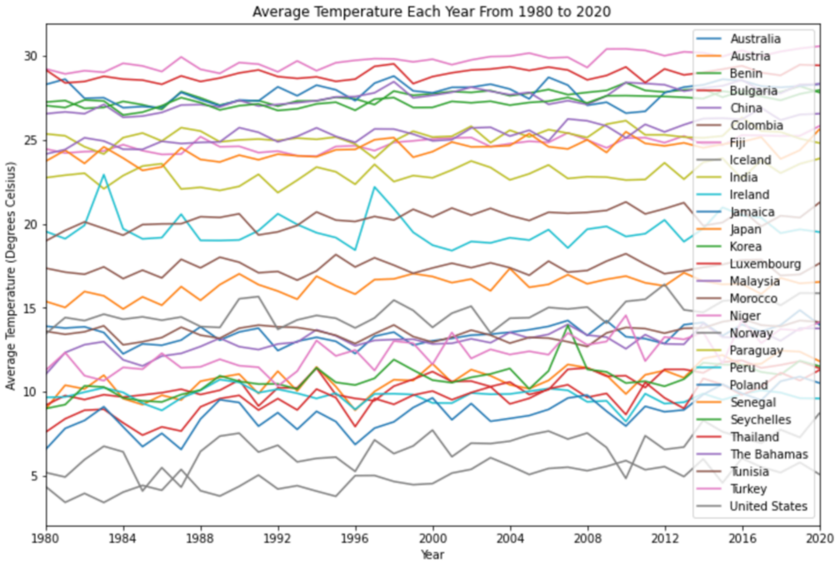


Figure 6: Average Temperature per year of Filtered Countries

Another challenge that came up repeatedly was working with datetime. All data was originally collected to give a value per year, so date columns were converted to 4-digit years. However, when converting this back to datetime after formatting, it only recognised it as minutes and seconds. Therefore, the numbers ’01-01’ were added to the end so it could be converted to datetime more effectively.

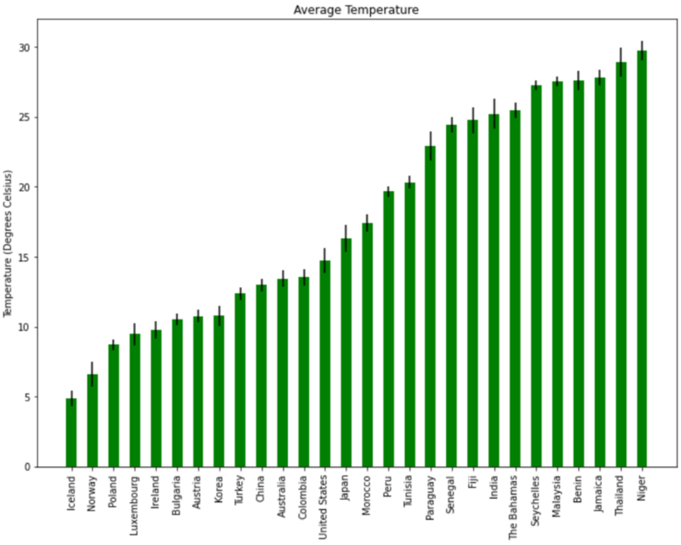
Result Reporting

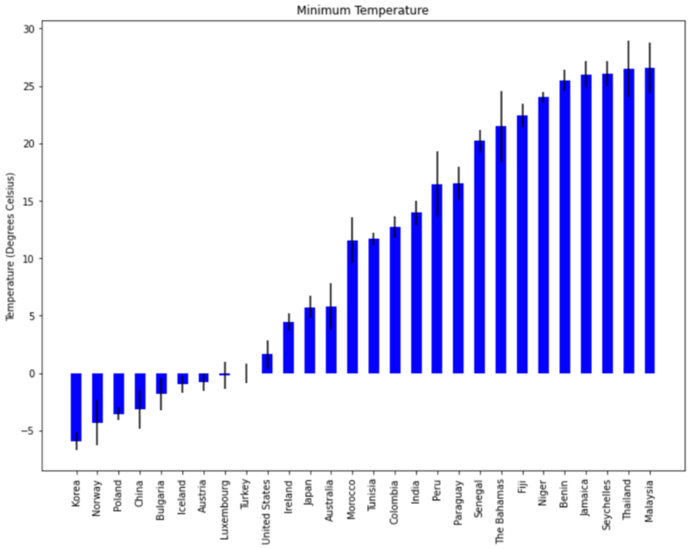
The weather data and WEO data was first analysed for contextual analysis on the data, then they were compared to one another to answer the questions posed.

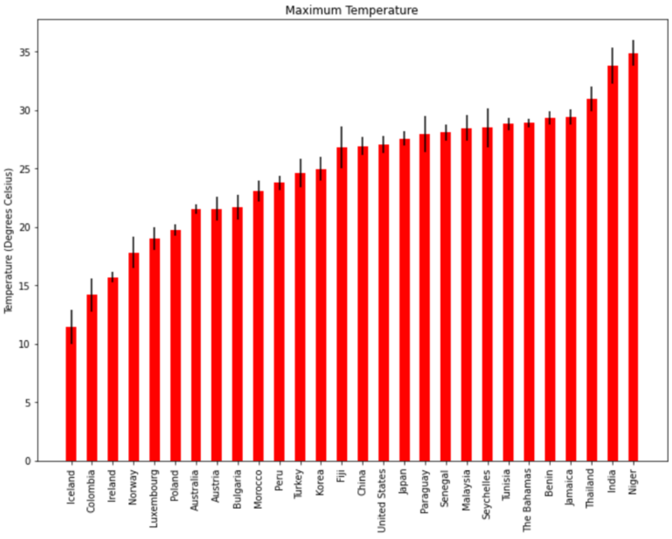
Weather Data

1. Iceland is the coldest country of countries analysed based on its average temperature each year. Similarly, Norway has low temperatures of around -4 at its coldest to 17 degrees at its hottest throughout the year. Niger has the hottest average temperature and the hottest maximum temperature per month, of around 34 degrees celsius. Its climate generally stays hot as the minimum temperature is around 25 degrees Celsius. Malaysia also has a hot climate and it’s average, minimum and maximum temperatures all fall at around 25 degrees, suggesting it stays at this temperature all year round. However, some of the standard deviations are fairly high which makes the minimum and

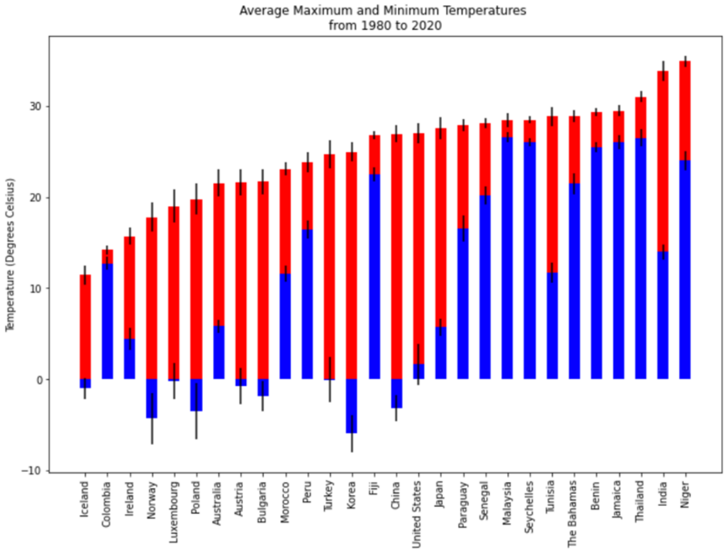
maximum temperature less reliable. This could be due to the higher presence of NaN values in these variables compared to the average temperature, or shows higher fluctuations in the extremities which balance each other out during the year, resulting in a consistent average temperature.

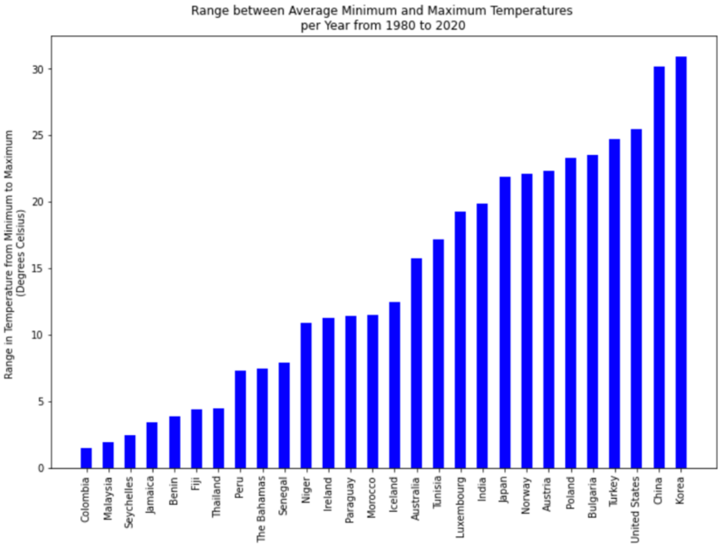






2. Korea shows one of the highest ranges between the minimum and maximum, from -6 to 25 degrees Celsius. China is not far behind, with a range in temperature between -4 and 27 degrees C. This





shows it has a more extreme climate than some of the other countries. Colombia and Malaysia have the smallest range in climates, with a change of just 1-2 degrees Celsius throughout the year for both. While Iceland has a fairly big range in temperature, it remains one of the coldest climates, with a maximum of 12 degrees Celsius and a minimum of -1. The US has a large range in temperature. A pattern is seen where some very large countries which fall closer the equator and span far away from it have the largest difference between maximum and minimum temperatures.

Economic Data

1. All countries have similar trends in change in GDP. When one country has an increase in GDP, it is likely all the other countries do as well. This is shown by the heatmap below, where the correlations between GDP per capita across countries are all above 0.75, showing a strong trend.

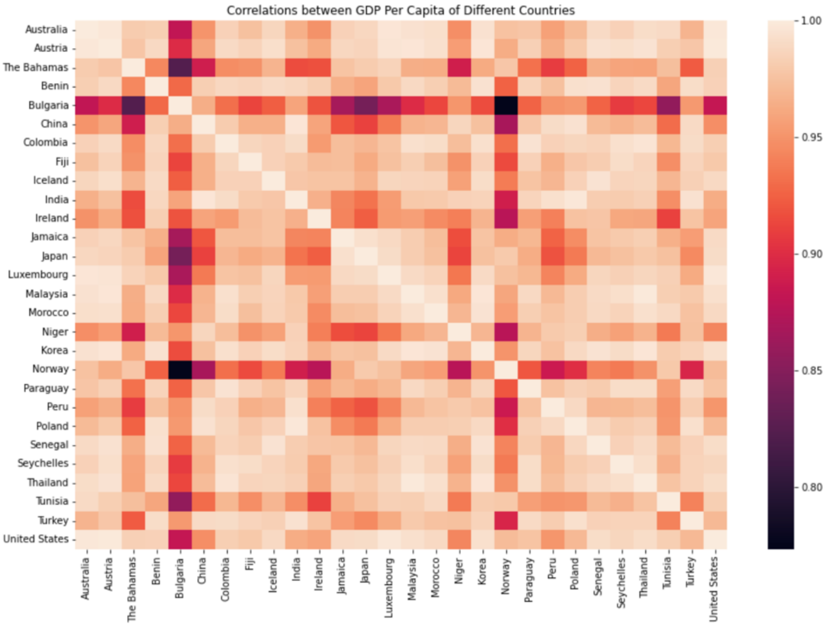


Figure 7: Seaborn Heatmap visualising the relationship between the GDP of various countries

Australia was taken as an initial example to compare Economic variables against itself.

2. There is an extremely strong correlation between GDP per capita and Population from 1980 to 2020. This suggests that the bigger the population, the more money they bring in per person. The R-value for this correlation is 0.99. There is also a mild negative correlation between GDP per Capita and Inflation. If GDP is increasing and inflation is decreasing, this is unexpected, therefore there may be another variable which is causing both these variables to change in this way. However, if inflation increases, this reduces the purchasing power of money, which reduces demand and consumption which causes a decrease in GDP.

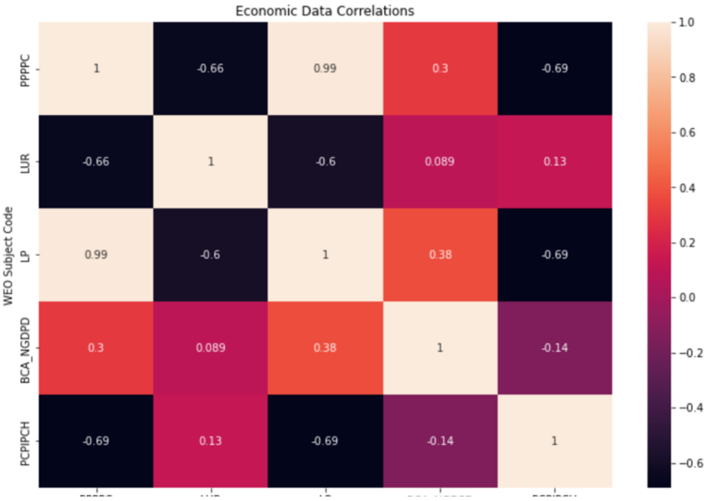


Figure 8: Heatmap showing the relationship between economic indicators of Australia.

3. GDP per capita, Population, Unemployment Rate and Inflation have the strongest correlation to one another in Australia. Population, Unemployment rate and Inflation can be taken and used to predict GDP per capita, with an R^2 value of 0.98. This is extremely strong, almost unrealistically so. It is likely this is because population has been increasing for the past 20 years and so has GDP. However, if population were to decrease, the question remains would GDP continue to increase.? There are not enough changes in the pattern of population to be confident that this has an impact on GDP.

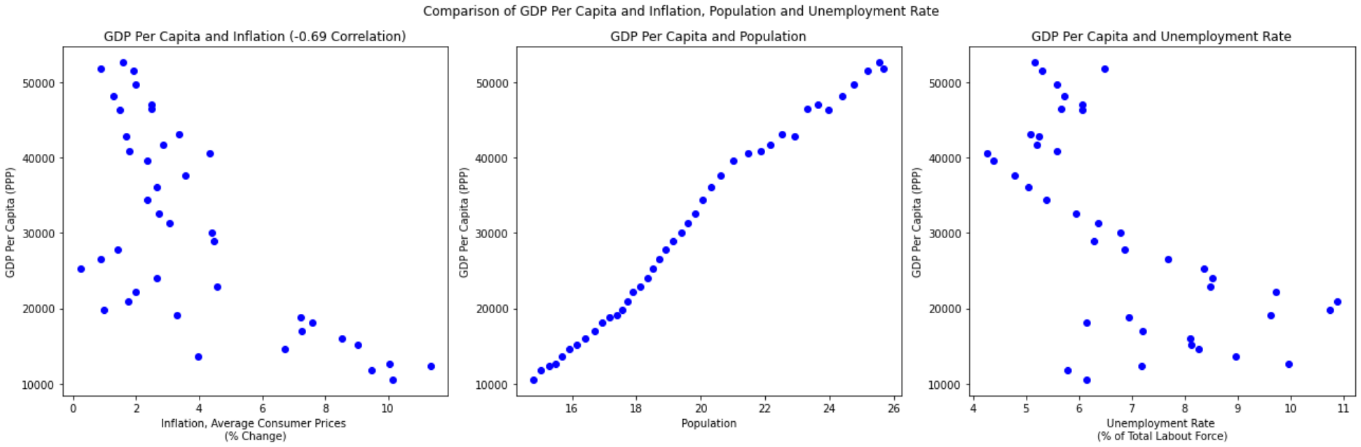


Figure 9: Scatter graphs showing the relationship between economic indicators and their effect on GDP per Capita

4. When Inflation and Unemployment Rate are used to predict GDP per Capita, the model built has an R^2 value of 0.59. This shows that Population in the model has a very high weighting.

5. Western countries in North America and Europe have the highest GDP per capita, particularly Luxembourg and Norway, closely followed by the US and Ireland. Despite China's reputation for industry, it is generally considered cheaper to manufacture there by Western Companies and so their GDP is one of the lowest out of the countries analysed. However, it has decreased a huge amount from 1980 to 2020, so the average GDP is less representative.

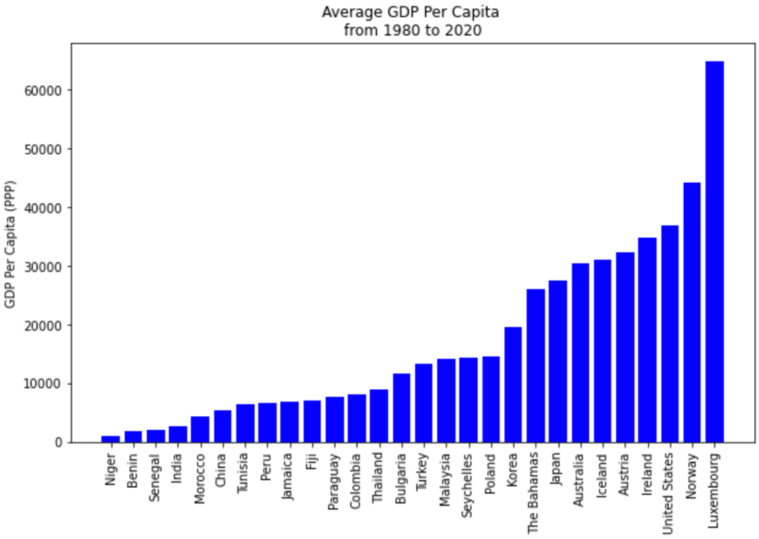
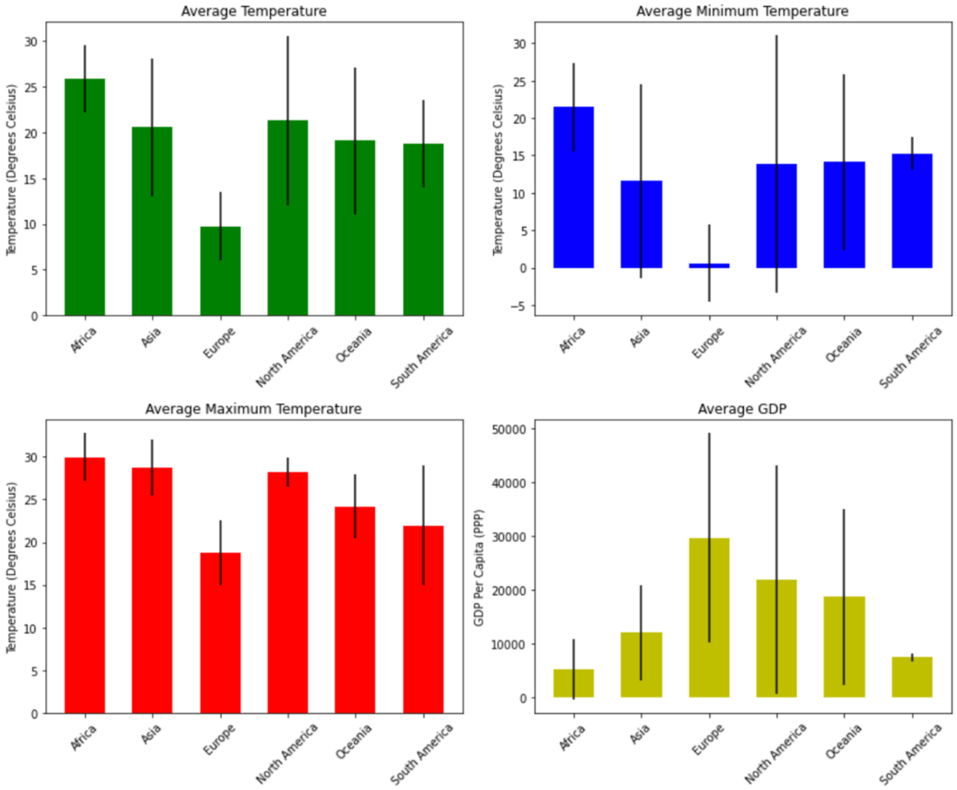


Figure 10: Bar chart showing the GDP Per Capita of all countries in the filtered list of countries

Weather & Economic Data

1. Europe has the lowest average, minimum and maximum temperatures but the highest GDP per Capita on average from 1980 to 2020. However, the minimum and maximum temperatures have a high standard deviation which suggests there is a large variation throughout the continent. Africa has the highest temperatures and the lowest GDP.



2. The average, minimum and maximum temperatures of each country show a negative correlation to GDP per Capita. As temperature increases, the GDP generally decreases. This could be because it is much easier to continue construction and other development work year-round in countries with milder temperatures. This relationship is more likely to be logarithmic as this curve of best fit is more in-trend with the shape of the data.

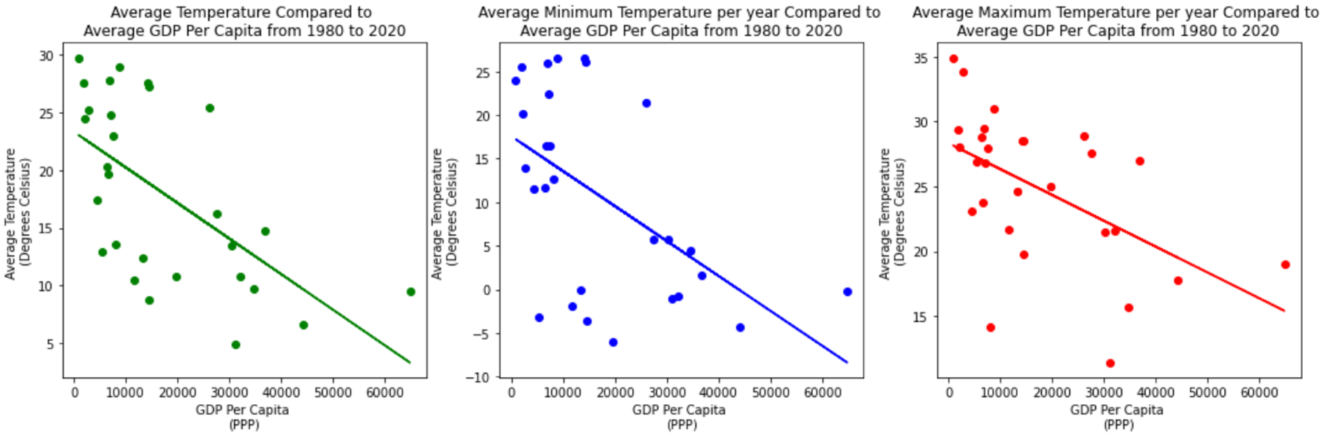


Figure 12: Scatter graphs showing the relationship between temperatures and GDP per Capita with a linear line of best fit showing a negative correlation.

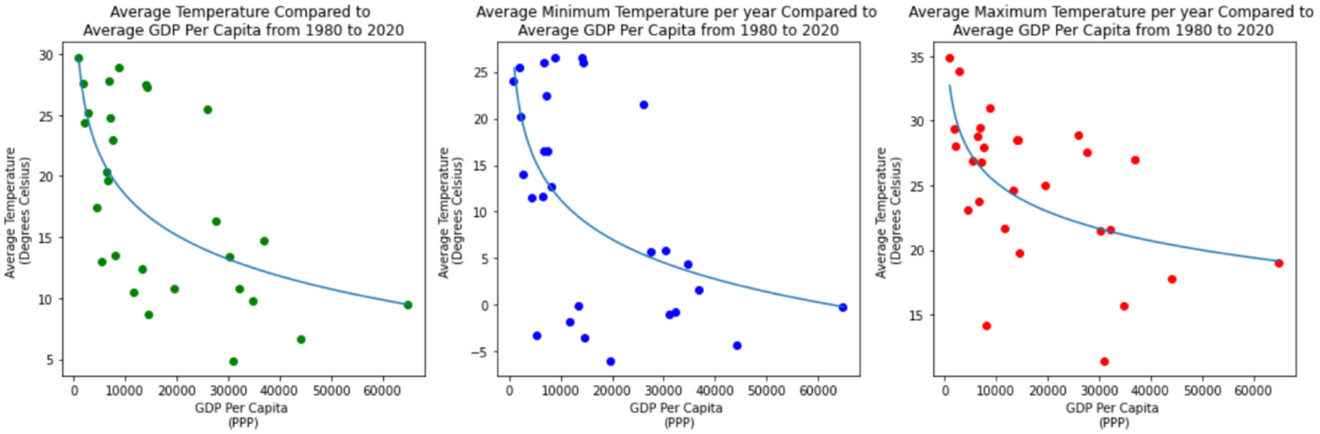


Figure 11: Scatter graphs showing the relationship between temperatures and GDP per Capita with a logarithmic line of best fit showing a decreasingly negative correlation

3. When grouped by continent, the trend in GDP and temperature is similar, showing a negative correlation between the two variables.

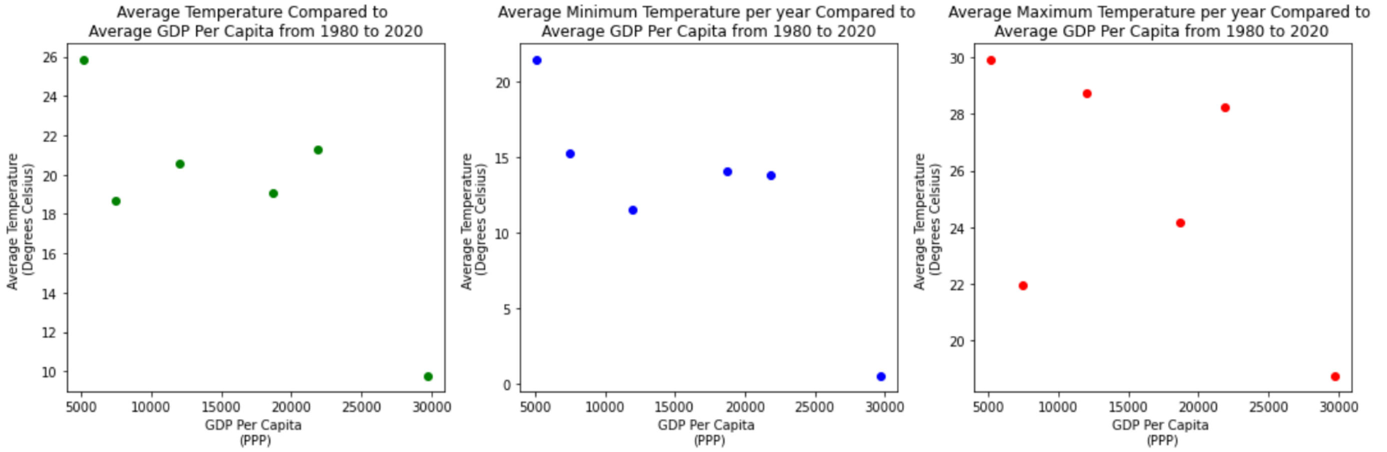


Figure 13: Scatter graphs showing the temperature and relationship to GDP per Capita grouped by continent, to identify this trend

4. There is little-to-no correlation between temperature and GDP growth (taken here as the growth from 1980 to 2020.

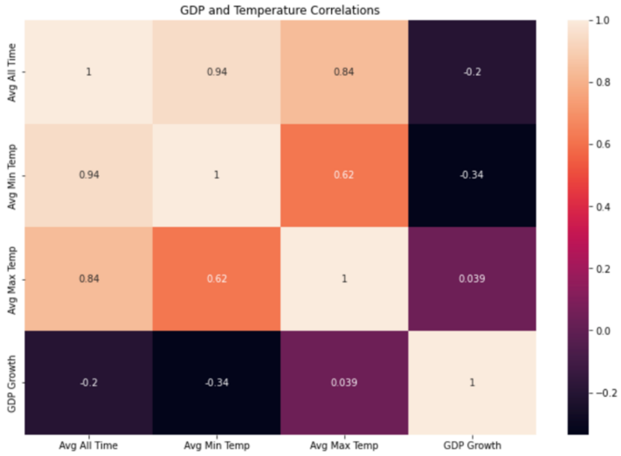
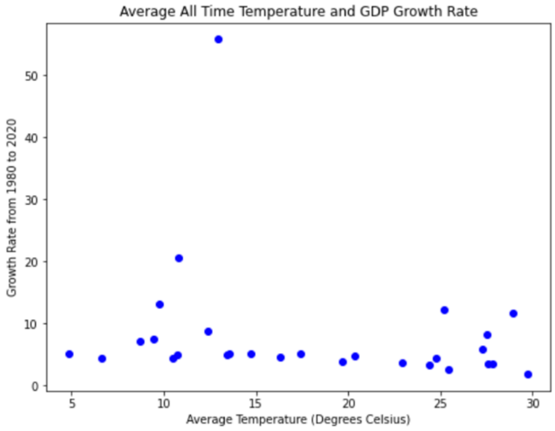
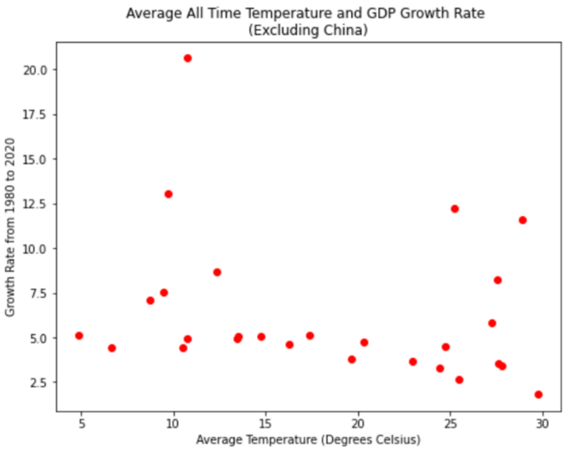


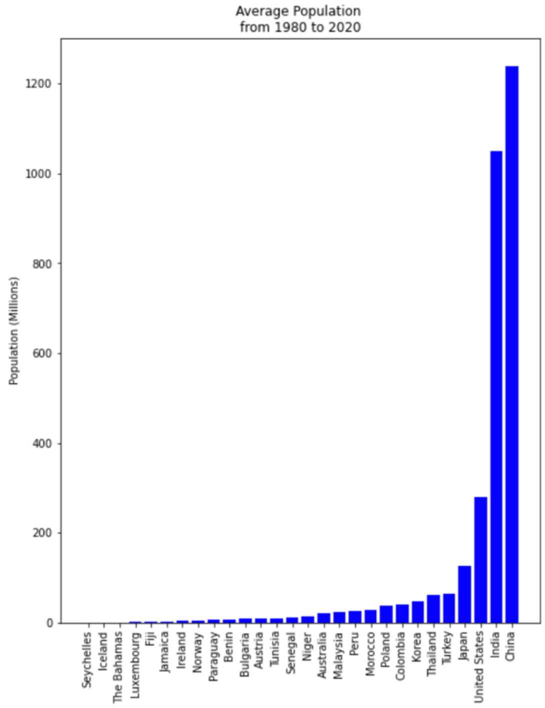
Figure 14: Heatmap showing the correlation between temperatures and growth in GDP per Capita from 1980 to 2020. There is little to no correlation here.

5. GDP per Capita growth was plotted against temperature anyway to visualise the relationship. China is an extreme outlier in this case, however it is known the data source is reliable and so this is not an anomaly. Understanding the context in the history of China's growth, it is an exception and very few countries have growth this high. Therefore, to assess the relationships between the rest of the data, China was removed, and the graph plotted again.





6. The population of China is huge in comparison to other countries and this could have an impact on the GDP per Capita, as this was shown previously to have a strong correlation between the two.



Conclusion

Out of the countries chosen there were none which have extremely cold conditions. Only 29 countries were chosen out of 195 in the world, which represents just 14.5%. The results showed a slight relationship between temperatures and GDP, and generally the colder the higher the GDP. However, these are close to an R^2 value of 0.6 and so are not strong enough to be completely confident in the results.

The objectives of the project were met and the initial key question was answered: what is the relationship between Climate and Economic Indicators? However, to progress the project further, it would be helpful to access another API (most likely paid for) that has more reliable data on weather with less NaN and missing values. This would also allow analysis of other climate conditions such as humidity and precipitation, and the impact on more economic indicator variables. It would be interesting to look more into the growth rates and small changes in weather conditions throughout years and months and the impact they have on these variables.

It is likely conditions such as natural disasters, e.g., frequency of flooding, earthquakes and droughts also have an impact on economic indicators of a country and so would be another interesting extension to the analysis.

Unfortunately, there was not time during this submission to complete all of this so the project will continue beyond the course. Being a solo project rather than a group project meant time was much more limited, having just 20% of the time available compared to groups of 5. However, it met all the requirements and achieved all the objectives specified, using a variety of modules and packages to undertake a relatively detailed data analysis project in the time available.

The course overall has been extremely educational and pushed Python skills to develop a huge amount in such a short amount of time.