Achievement 6

Exercise 6.1 **Dan Franklin**

In this project I'd like to look at relationships between emissions and life expectancy. I am interested to see if there is a correlation between the two. Increased emissions over the years can suggest a country is developing and therefore such things such as education and medical care would likely improve, so I'd expect to see life expectancy increasing, and I wonder if it gets to a point where the emissions begin to negatively affect quality of life and we see the increase in life expectancy slow or even decrease.

I have two datasets both sourced on Kaggle.

Primary dataset :

Life Expectancy - https://www.kaggle.com/datasets/ulrikthygepedersen/life-expectancy

Secondary dataset :

Emissions by Country - https://www.kaggle.com/datasets/thedevastator/global-fossil-co2-emissions-by-country-2002-2022

Data Sourcing - Life Expectancy	
Data source	https://www.kaggle.com/datasets/ulrikthygepedersen/life-expectancy https://data.nasdaq.com/databases/WB/data World Bank provides data from hundreds of countries and regions around the world, from multiple categories such as finance, economy, energy, education, health, poverty, agriculture, employment, population, land use, foreign aid, climate change, government expenditures, literacy, mortality, and patents. This is open data that we've sourced and made freely available to you. Create a free Nasdaq Data Link account to get access to this and all open data on Nasdaq Data Link. You will also get API access and sample data access to most premium data products. License Attribution 4.0 International (CC BY 4.0)
Data collection	It is compiled from various sources, including national statistical agencies, the World Health Organization (WHO), and other relevant data sources.
Contents	The life expectancy per country dataset provides a comprehensive overview of the life expectancy of each country. The dataset includes information on life expectancy at birth, covering all countries in the world. **country_code:** Country Code **country_name:** Country Name **year:** Year (1960 – 2020) **value:** Life Expectancy per country per year
Data relevance	Provides the necessary info on life expectancy across the globe Complied from various official sources including national statistics agencies and WHO Includes recent data up to the year 2020 which you can argue is the most up to date as possible before the pandemic had an affect on the numbers

Data Sourcing - Emissions by Country		
Data source	https://www.kaggle.com/datasets/thedevastator/global-fossil-co2-emissions-by-country-2002-2022 https://zenodo.org/record/7215364 Andrew, Robbie M., & Peters, Glen P. (2022). The Global Carbon Project's fossil CO2 emissions dataset (2022v27) [Data set]. Zenodo. https://doi.org/10.5281/zenodo.7215364 **License: CC0 1.0 Universal (CC0 1.0) - Public Domain Dedication** No Copyright - You can copy, modify, distribute and perform the work, even for commercial purposes, all without asking permission. See Other Information.	
Data collection	The Global Carbon Project (GCP) has been publishing estimates of global and national fossil CO_2 emissions since 2001. In the first instance these were simple re-publications of data from another source, but over subsequent years refinements have been made in response to feedback and identification of inaccuracies.	
Contents	**Country**	The name of the country. (String)
	ISO 3166-1 alpha-3	The three-letter code for the country. (String)
	Year	The year of the data. (Integer)
	Total	The total amount of CO2 emissions for the country in the given year. (Float)
	Coal	The amount of CO2 emissions from coal for the country in the given year. (Float)
	Oil	The amount of CO2 emissions from oil for the country in the given year. (Float)
	Gas	The amount of CO2 emissions from gas for the country in the given year. (Float)
	Cement	The amount of CO2 emissions from cement production for the country in the given year. (Float)
	Flaring	The amount of CO2 emissions from flaring operations for the country in the given year. (Float)
	Other	The amount of CO2 emissions from other sources for the country in the given year. (Float)
	Per Capita	The amount of CO2 emissions per capita for the country in the given year. (Float)
Data relevance	Provides geographic and annual statistics of gas emissions Has detailed statistics for different types of gases The most up-to-date available data and has potential for creating a forecast. Was published in 2022	

Data Understanding

Life Expectancy			
Column	**Qualitative / Quantitative**	**Discrete / Continuous**	**Nominal / Ordinal / Binary**
country_code	Qualitative	Discrete	Nominal
country_name	Qualitative	Discrete	Nominal
year	Qualitative	Discrete	Ordinal
value	Quantitative	Continuous	Nominal

Emissions by Country			
Column	**Qualitative / Quantitative**	**Discrete / Continuous**	**Nominal / Ordinal / Binary**
Country	Qualitative	Discrete	Nominal
ISO 3166-1 alpha-3	Qualitative	Discrete	Nominal
Year	Qualitative	Discrete	Ordinal
Total	Quantitative	Continuous	Nominal
Coal	Quantitative	Continuous	Nominal
Oil	Quantitative	Continuous	Nominal
Gas	Quantitative	Continuous	Nominal
Cement	Quantitative	Continuous	Nominal
Flaring	Quantitative	Continuous	Nominal
Other	Quantitative	Continuous	Nominal
Per Capita	Quantitative	Continuous	Nominal

Cleaning and consistency checks

Life Expectancy		
Problem	**Description**	**Solution**
Missing Values	-	-
Duplicates	-	-
Mixed Data Types	-	-
Outliers	Low life expectancies	I left them in as they are attributed to poor countries facing hardships. They are accurate.

Emissions by Country		
Problem	**Description**	**Solution**
Missing Values	ISO 3166-1 alpha-3 column – **1632** missing values Total column – **200** missing values Coal 41360 Oil 41387 Gas 41486 Cement 42290 Flaring 41554 Other 61484 Per Capita 44130	The rows with ISO code missing values I deleted as these represented areas rather than countries. The missing values in the Total column I replaced with 0's as that's what the missing values represented. I did the same for the missing values in the types of emissions columns and Per Capita column.
Duplicates	-	-
Mixed Data Types	-	-
Outliers	High emissions	I left them in as they reflect countries which experienced periods of increased oil production eg The Bahamas in the 1970's.

Wrangling

Life Expectancy	
Column name change	Changed the name of "value" to " Life Expectancy" to make it more understandable
Removing rows	Removed rows which contained data for values in the "country_name" column that weren't countries, totaling 50 different entries in total. Eg: 'Africa Eastern and Southern' 'Africa Western and Central' 'Arab World' 'Heavily indebted poor countries (HIPC)' 'IDA & IBRD total' 'Low & middle income' I left rows in which had outliers / low life expectancy as they were poor countries facing hardships so the figures are accurate

Emissions by Country	
Column name change	I changed the ISO 3166-1 alpha-3 column name to country_code as this is more understandable and also matches the equivalent column in the other data frame which is useful if I join them on this column
Removing rows	Removed rows which contained data for values in the "Country" column that weren't countries, totaling 9 different entries in total. Eg Antarctica. For now I have left in "International Transport" as I might use it for analysis.

Limitations and Ethics

In terms of limitations, some countries have emissions data going all the way back to the 1700's, whereas others begin much more recently, so this would make it impossible to compare them. I have kept in the old data for now as it most likely represents countries which were developing a couple of hundred years ago and are now developed countries, so it might be useful for comparison.

The data comes from reliable sources and has been used for other analyses already. One question could be the authenticity of certain countries data, perhaps they have altered the true figures, but that is speculative and the data should be take at face-value for the purpose of this analysis.

In terms of ethics I cannot see any issues as there are no PII concerns.

Questions to explore

What happens to a country's average life expectancy as emission levels increase?

When looking at emissions and life expectancy, is there a correlation?

How long does it take after the initial increase in a country's emissions (so assuming leading to improved education, healthcare etc) does it take until we see an increase in the life expectancy?

Are there any existing programs or efforts aimed at decreasing dropout rates?

Is there a decline in the rate of increasing life expectancy or even a decrease once emission levels get to a certain amount per capita? If so, what amount?

How do emissions and life expectancy compare across different parts of the world? Is it more country specific or region?

Globally, is there a trend linking life expectancy and emissions?