Robert Cheadle

6.1: Sourcing Open Data

Data Source:

Data Set: Life Expectancy at Birth, Total (Years)

Source: "Life Expectancy at Birth, Total (Years)", Published online at OurWorldInData.org. Retrieved from: 'https://datacatalog.worldbank.org/search/dataset/0037712/World-Development-Indicators' [Online Resource]

Data Collection:

"Life Expectancy at Birth, Total (Years)" is an external and . The data since 1961, which will be used in this analysis is available in the World Development Indicators (WDI) published by the World Bank. This data is collected on an annual basis by the United Nations Population Division. This data can be considered trustworthy.

Content:

The contents of this data set include location data describe by 'Entity'(country) and 'Code'(country code). Date data described as 'Year' and quantitative data 'Life expectancy at birth, total (years).'

Life expectancy at birth is highly sensitive to the rate of death in the first few years of life. This analysis will only be considering life expectancy at birth not at different age groups. Also, it is important to note that period and cohort life expectancy estimates are statistical measures, and they do not consider any person-specific factors such as lifestyle choices.

Relevance:

This data set is relevant to the project hypothesis and objective presented because life expectancy will give insight into how a countries gross domestic product (GDP), population, carbon-dioxide emissions from production overtime may or may not show significant correlation or other relationships.

Data Set: CO_2 and Greenhouse Gas Emissions

Source: Hannah Ritchie, Max Roser and Pablo Rosado (2020) - "CO_2 and Greenhouse Gas Emissions". Published online at OurWorldInData.org. Retrieved from: 'https://ourworldindata.org/co2-and-other-greenhouse-gas-emissions' [Online Resource]

Data Collection:

This dataset is a collection of key metrics maintained by Our World in Data. It is updated regularly and includes data on CO2 emissions (annual, per capita, cumulative, and consumption-based), other greenhouse gases, energy mix, and other relevant metrics. This dataset is a compilation of data from other data sources such as the Global Carbon Project and BP.

Content:

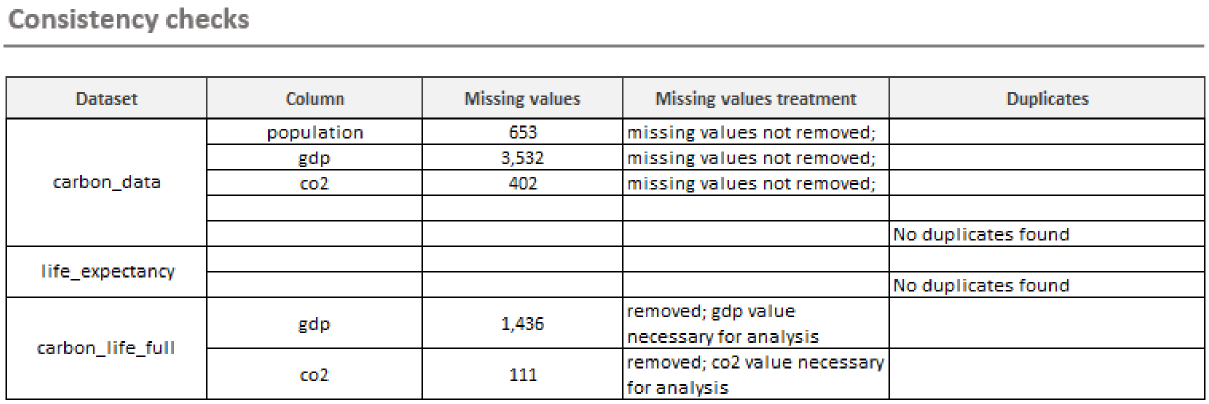
The contents of this data set include location data described as ‘country’ and ‘iso_code,’ a 3-letter country code. As well as quantitative data relevant to this project in the form of population, GDP, and co2 (annual total production-based emissions), measured in million tons.

Limitations of this data are we do not know the data collection and quantitative methods of the original sources. As well as any potential bias those sources may have in their collection and measurements.

Relevance:

This data set is relevant to the project hypothesis and objective presented because a countries gross domestic product (GDP), population, carbon-dioxide emissions from production overtime may or may not show significant correlation to a country’s life expectancy at time of birth.

Data Profile:

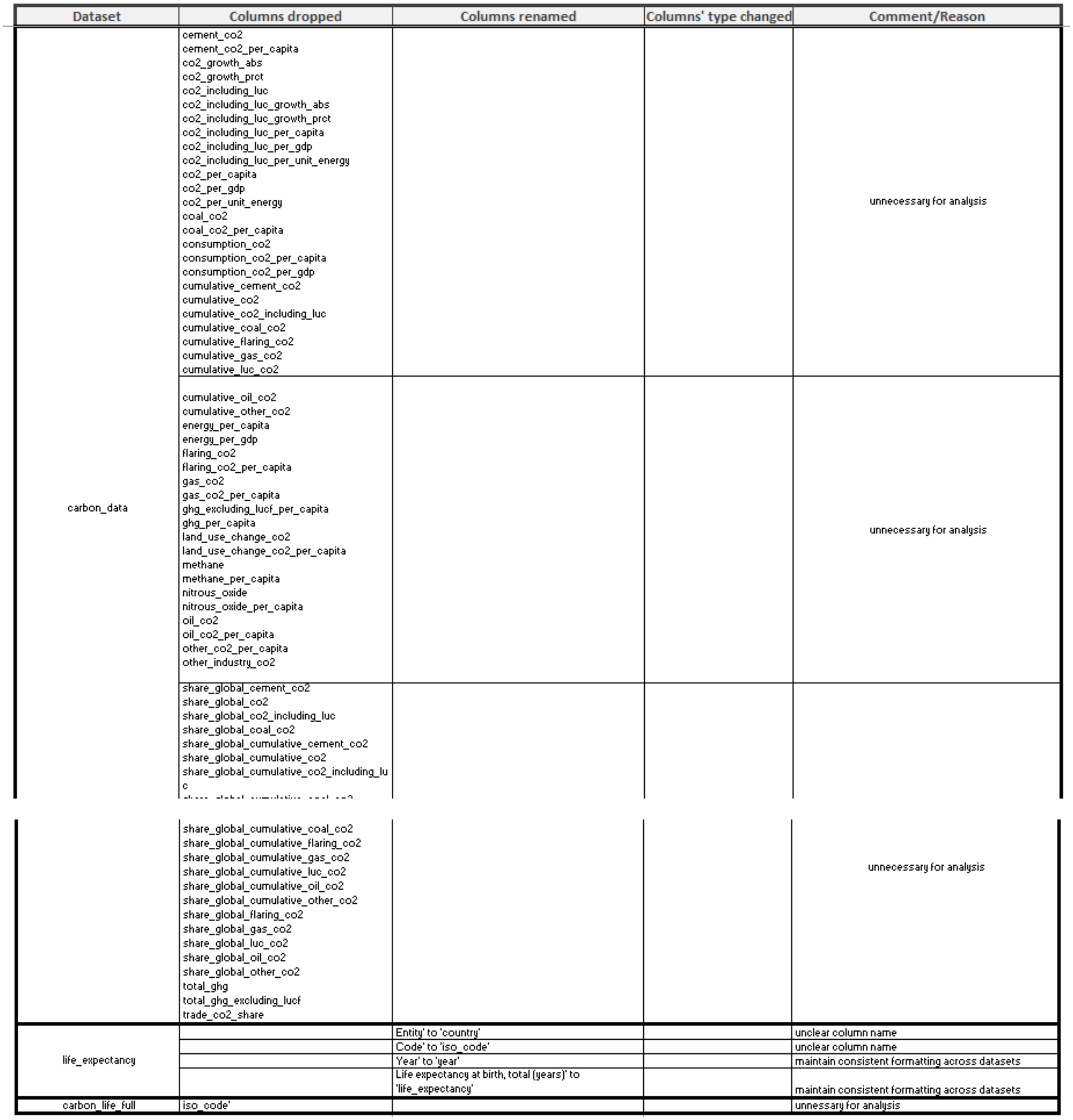


Consistency checks

Dataset	Column	Missing values	Missing values treatment	Duplicates
carbon_data	population	653	missing values not removed;	
	gdp	3,532	missing values not removed;	
	co2	402	missing values not removed;	
				No duplicates found
life_expectancy				
				No duplicates found
carbon_life_full	gdp	1,436	removed; gdp value necessary for analysis	
	co2	111	removed; co2 value necessary for analysis	

- The above table displays consistency checks performed on raw and cleaned datasets.

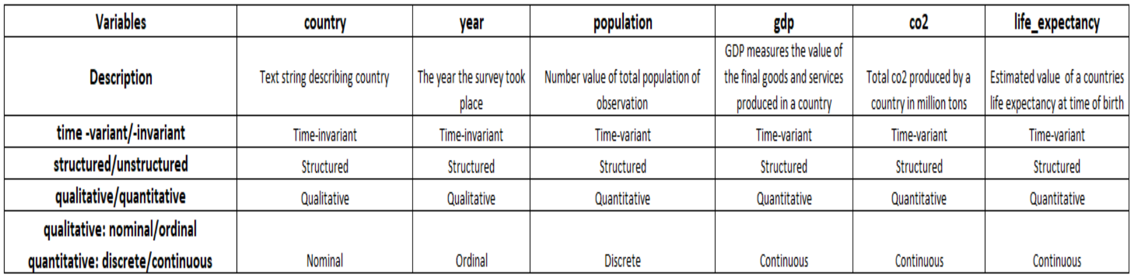
Wrangling steps



Dataset	Columns dropped	Columns renamed	Columns' type changed	Comment/Reason
carbon_data	cement_co2 cement_co2_per_capita co2_growth_abs co2_growth_prct co2_including_luc co2_including_luc_growth_abs co2_including_luc_growth_prct co2_including_luc_per_capita co2_including_luc_per_gdp co2_including_luc_per_unit_energy co2_per_capita co2_per_gdp co2_per_unit_energy coal_co2 coal_co2_per_capita consumption_co2 consumption_co2_per_capita consumption_co2_per_gdp cumulative_cement_co2 cumulative_co2 cumulative_co2_including_luc cumulative_coal_co2 cumulative_flaring_co2 cumulative_gas_co2 cumulative_luc_co2			unnecessary for analysis
	cumulative_oil_co2 cumulative_other_co2 energy_per_capita energy_per_gdp flaring_co2 flaring_co2_per_capita gas_co2 gas_co2_per_capita ghg_excluding_lucf_per_capita ghg_per_capita land_use_change_co2 land_use_change_co2_per_capita methane methane_per_capita nitrous_oxide nitrous_oxide_per_capita oil_co2 oil_co2_per_capita other_co2_per_capita other_industry_co2			unnecessary for analysis
	share_global_cement_co2 share_global_co2 share_global_co2_including_luc share_global_coal_co2 share_global_cumulative_cement_co2 share_global_cumulative_co2 share_global_cumulative_co2_including_luc share_global_cumulative_coal_co2 share_global_cumulative_flaring_co2 share_global_cumulative_gas_co2 share_global_cumulative_luc_co2 share_global_cumulative_oil_co2 share_global_cumulative_other_co2 share_global_flaring_co2 share_global_gas_co2 share_global_luc_co2 share_global_oil_co2 share_global_other_co2 total_ghg total_ghg_excluding_lucf trade_co2_share			unnecessary for analysis
life_expectancy		Entity' to 'country'		unclear column name
		Code' to 'iso_code'		unclear column name
		Year' to 'year'		maintain consistent formatting across datasets
		Life expectancy at birth, total (years)' to 'life_expectancy'		maintain consistent formatting across datasets
carbon_life_full	iso_code'			unnessary for analysis

- The above table shows all columns dropped from raw and cleaned datasets, also includes column name changes.

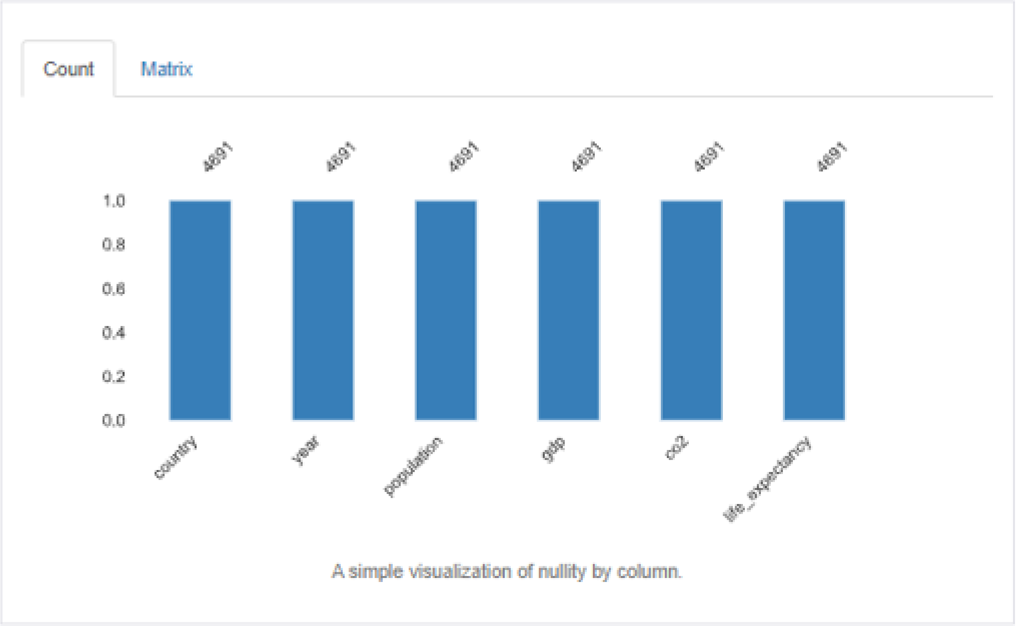
Summary Statistics



Variables	country	year	population	gdp	co2	life_expectancy
Description	Text string describing country	The year the survey took place	Number value of total population of observation	GDP measures the value of the final goods and services produced in a country	Total co2 produced by a country in million tons	Estimated value of a countries life expectancy at time of birth
time -variant/-invariant	Time-invariant	Time-invariant	Time-variant	Time-variant	Time-variant	Time-variant
structured/unstructured	Structured	Structured	Structured	Structured	Structured	Structured
qualitative/quantitative	Qualitative	Qualitative	Quantitative	Quantitative	Quantitative	Quantitative
qualitative: nominal/ordinal quantitative: discrete/continuous	Nominal	Ordinal	Discrete	Continuous	Continuous	Continuous

- The above table includes information about each variable in the dataset and the type of data each variable consists of.

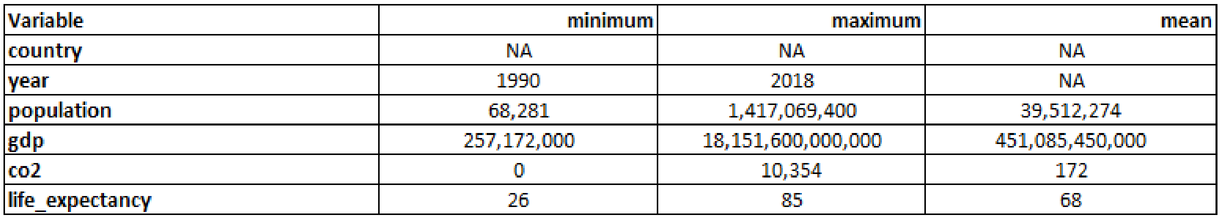
Counts Expected from the variables



A simple visualization of nullity by column.

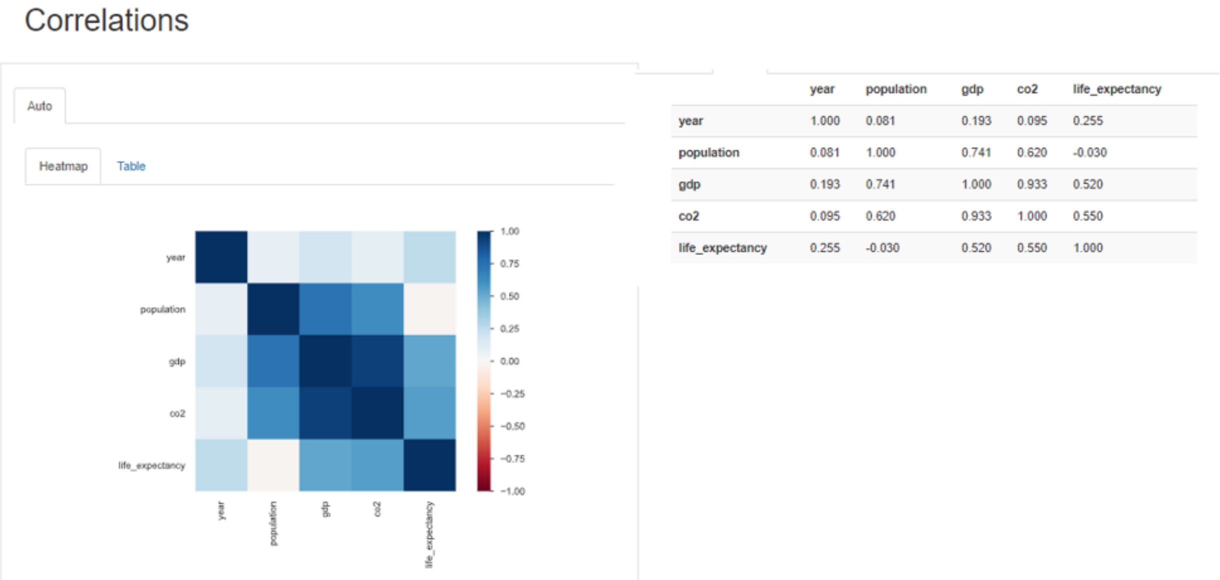
- The above table shows that the frequency of each value is consistent.

Data Accuracy



Variable	minimum	maximum	mean
country	NA	NA	NA
year	1990	2018	NA
population	68,281	1,417,069,400	39,512,274
gdp	257,172,000	18,151,600,000,000	451,085,450,000
co2	0	10,354	172
life_expectancy	26	85	68

- The above table shows some basic descriptive information about each variable found in the data set.


Correlations

	year	population	gdp	co2	life_expectancy
year	1.000	0.081	0.193	0.095	0.255
population	0.081	1.000	0.741	0.620	-0.030
gdp	0.193	0.741	1.000	0.933	0.520
co2	0.095	0.620	0.933	1.000	0.550
life_expectancy	0.255	-0.030	0.520	0.550	1.000

- The above tables show the correlation between the different variables of the data set. This information will help to determine how the variables relate to one another.

<u>Questions to explore:</u>

1. How does a country's GDP/capita effect life expectancy? Derive new column 'GDP per capita'
2. How does a country's co2(production)/capita effect life expectancy? Derive new column 'co2 per capita'
3. Do countries with high co2 production and GDP have higher life expectancy at a significant level?
4. Do countries with medium co2 production and GDP have higher life expectancy at a significant level?
5. Do countries with low co2 production and GDP have lower life expectancy at a significant level?
6. How does the rate of change from year to year for (GDP, co2, and life_expectancy) compare? Derive rate of change variables for GDP, co2, and life expectancy.
7. Does a decrease in co2 production necessarily result in decreased GDP?
8. Does increased co2/capita have any effect on a country's life expectancy?