Group C
Report -Week 4
Data set: CO2 emission
Date: 2022 Nov 5
Sanjina Poudel(jina.poul@gmail.com)

Objective:

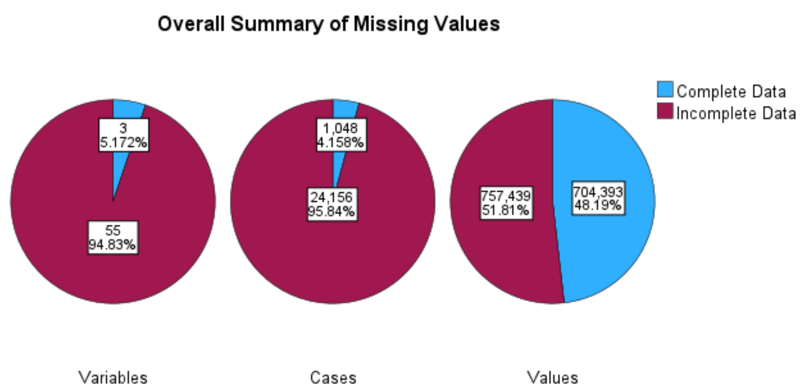
The main objective of this project is to know and explore this dataset. The study includes missing data analysis, hypothesis testing and correlation.

Methodology:

This study is done using SPSS and python.

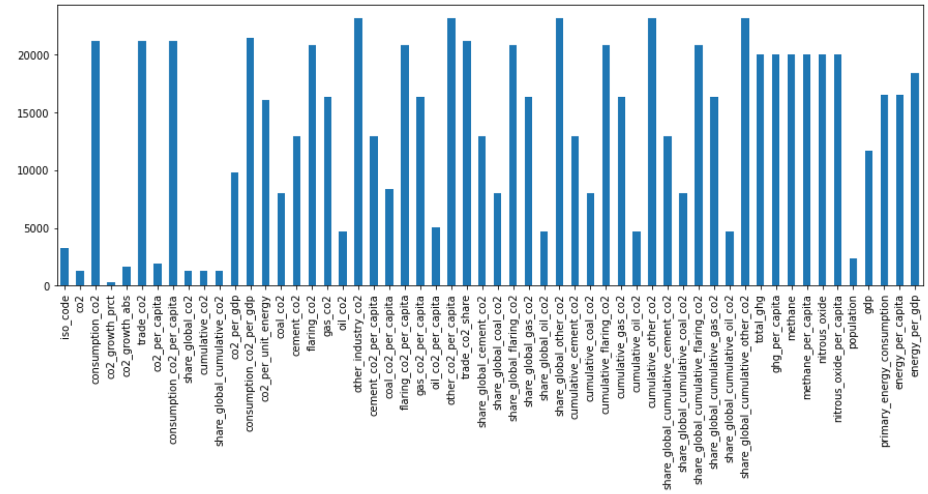
Dataset:

The dataset contains 58 variables and almost 25000 observations. In this study we choose some variables for further analysis. The figure below shows the summary of the original dataset.



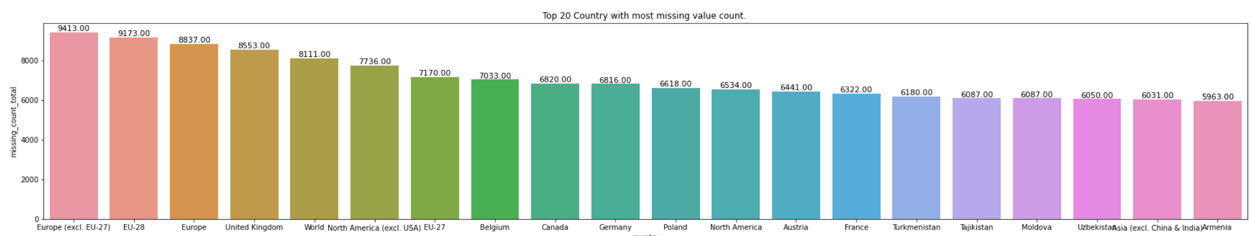
Only 5.172% of variables , 4.158% of cases in total we have only 48.19% of the overall dataset have complete data.

Frequency of missing values for each variables (Ref: report.xls `sheet_name='mv_frequency_variable'`):



The figure above shows that most of the variables have a higher number of missing values.Out of 58 variables 13 have (10>=missing % <=50) , 20 have (20>=missing % <=80) and 20 have (15>=missing % <=100).

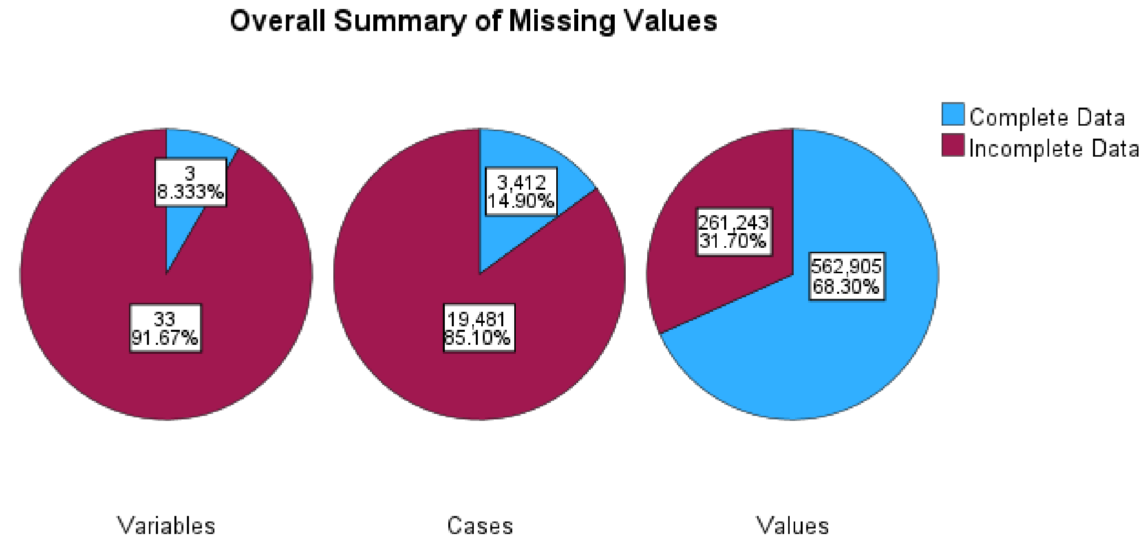
Figure below shows the top 20 countries with the most missing data.



Europe(excl.EU-27)have most of the missing values 9413. Data have been extracted for the missing value based on countries(ref: `sheet_name='mv_by_each_country'`). Report of the percentage of missing value of every country for each variable is reported in (ref:`'mv_variable_by_each_country_percetage')`

Selection of variables:

From the analysis , it has been found that the variables have a lot of missing data. So, We decided to remove the variables that have 70-100 % of missing data. After removing variables that have 70-100 percent of missing value , figure below shows the overall summary of missing value.

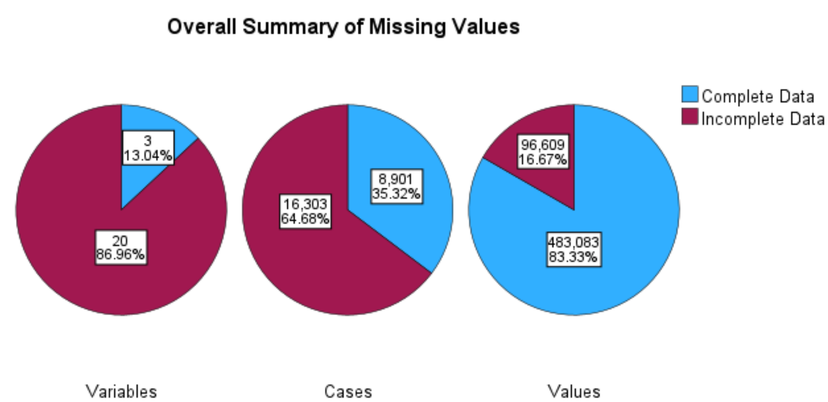


Still, We didn’t get a satisfactory result. Based on research, it has been found that the missing value upto 50% can be considered for the large dataset. So, We remove the variables that have more than 50 percent of missing data.

Final dataset for the study:

The original dataset contains the annual fossil co2 emission for all countries since 1750 broken down by fuel or process: coal, oil, gas, cement and other industries. But in our dataset, the dataset contains information about co2 , co2 from fuel and co2 from coal.

The dataset now contains 23 variables and 25204 data.



Explanation about variables:
Co2: Annual national or regional Co2 emission per year .

Co2_growth_prct: Growth percentage of co2 emission compared to last year.
Co2_growth_abs: Growth value of co2 emission compared to last year.
Co2_per_capita: Measure the average annual emissions per person for a country or region.It is calculated by dividing the total annual emissions of the country or region by its total population. *Co2_per_capita = Co2/population.*
Share_global_co2: This metric is calculated by dividing a country or region's emissions by the global emissions in any given year.
Cumulative_co2: Sum of annual emissions from 1750 onwards.
Share_global_cimulative_co2: This metric is calculated by dividing a country or region's cumulative emissions by the global cumulative emissions in any given year.
Co2_per_gdp: It is used to measure how carbon-intensive a country's economy is by dividing a country or region's annual CO_2 emissions by its total annual gross domestic product (GDP).

Coal_co2: Annual emission of CO2 from coal.
Oil_co2: Annual emission of CO2 from oil.

coal_co2_per_Capita: Measure the average annual emissions of Co2 from coal per person for a country or region.
oil_co2_per_Capita:Measure the average annual emissions of Co2 from oil per person for a country or region

Share_global_coal_co2: Global share of carbon dioxide emitted from coal.
Share_global_oil_co2: Global share of carbon dioxide emitted from oil.
Cumulative_coal_co2: Sum of annual emissions of carbon dioxide emitted from coal from 1750 onwards.
Cumulative_oil_co2: Sum of annual emissions of carbon dioxide emitted from oil from 1750 onwards.
Share_global_cumulative_coal_co2: Global share of cumulative emission of carbon dioxide emitted from coal
Share_global_cumulative_oil_co2: Global share of cumulative emission of carbon dioxide emitted from coal
Population: Population of the region or a country in that year
Gdp: Gross domestic product

Missing value imputation:

Co2 = Imputed by Co2_per_capita*Population
Co2_per_capita= Imputed by Co2 / Population
Co2_per_gdp=Imputed by Co2 / gdp
Population= Imputed by Co2_per_capita*Co2
coal_co2_per_Capita=Imputed by coal_co2/population
oil_co2_per_Capita=Imputed by oil_co2/population

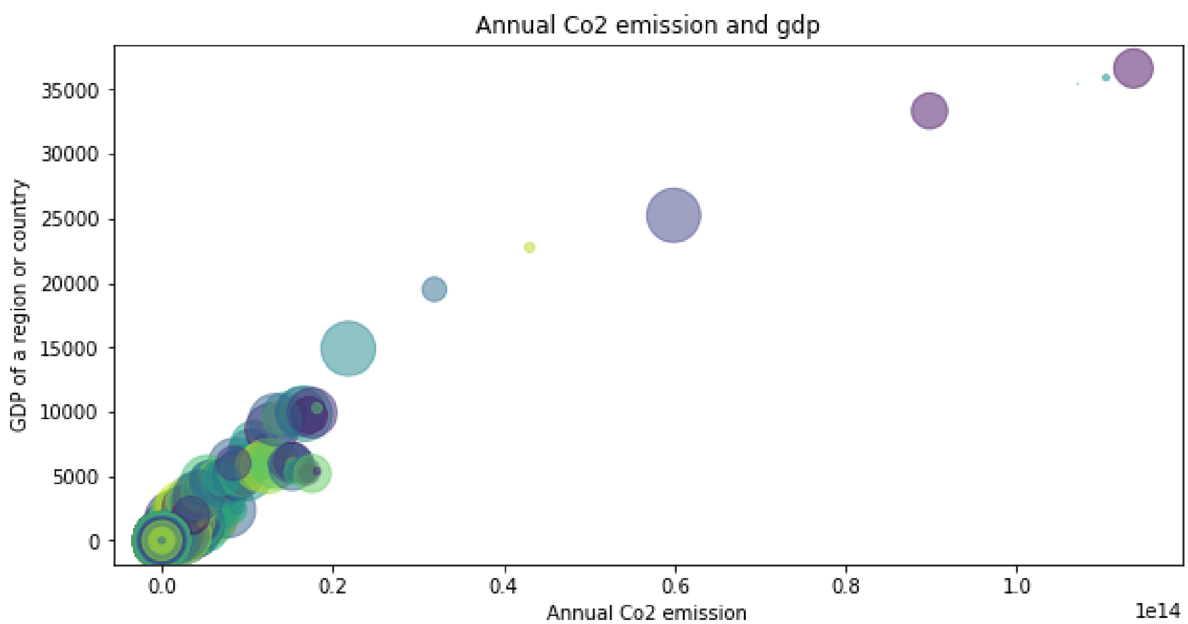
Normal Distribution of the variables

Histogram with normality curve is computed for all variables**(ref: histogram.spv)**. Based on histogram and normality curve we can find normal distribution for most of the variables. Also from the observation it is found that the variables have extreme values. But it is difficult to say exactly. So, skewness and kurtosis values are also computed**(ref: Appendix B)**. From these values, we find that not any variable follows the normality distribution.

Research question:

Based on whole data

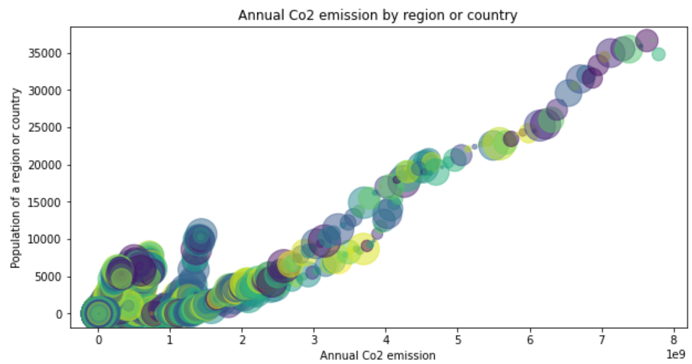
1. Is there any relation between gdp and co2 emission?



<Figure size 432x288 with 0 Axes>

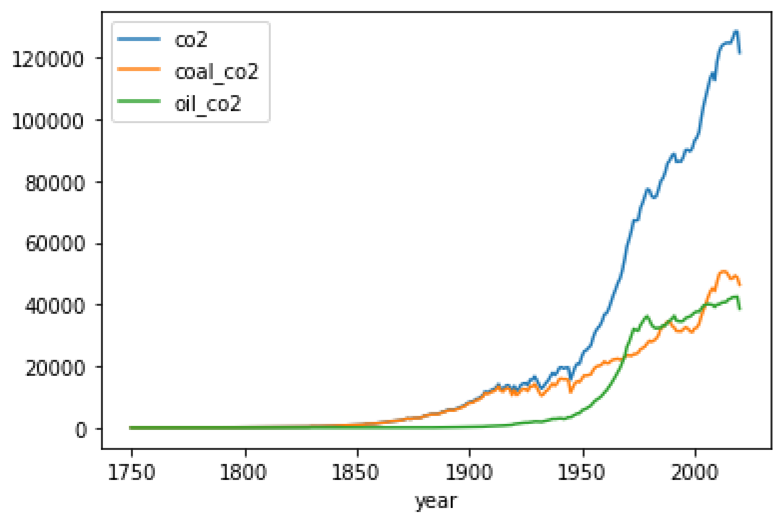
From the scatter plot above, it shows that the plot follows the nearly straight line. It is found that as gdp increases, the annual emission of co2 also increases. But some value shows the extreme point, it may be because our dataset contains the dataset by region too which have the sum of value over many countries.

2. Is there any relation between population and co2 emission ?

    a. 

From the scatter plot above we can see the annual co2 emission and population have some direct dependencies. The minimum annual emission is 0 while maximum is 36702.503. The emission increases as the population increases.

3. Carbon emission by year

    a. 

From the plot above it shows that there is a rapid increase in emission from 1950. It shows that there are more industry revolutions during that year. It also shows that in the beginning the consumption of coal was more than that of oil.

Research question- Hypothesis testing

Is there any significant difference in the share of global co2 in 2000 and 2005?

For this, annual emission of co2 is selected only for 2000 and 2005. Since, we are taking only some samples first let's check the distribution of data if it is normally distributed or not. The sample size is 231. So, computing the z scores.

Skewness: share_global_co2_2000 :9.654/0.160 =60.3375
Kurtosis: share_global_co2_2000:112.71/0.319=353.32
Skewness: share_global_co2_2005 :9.684/0.160=60.525
Kurtosis: share_global_co2_2005:112.181/0.319=31.66

The distribution of global share of Co2 in 2000(mean = 1.5952, sd=0.20) and 2005(mean = 1.5998, sd=0.30) is not normally distributed based on skewness and kurtosis value.

We should now apply non parametric test.Wilcoxon test

Null hypothesis: The central tendencies of the distribution of global share of co2 in 2000 and 2005 are the same.In other words, the distribution of share global of co2 is same across the categories in the year.

Alternative hypothesis: The central tendencies of the distribution of global share of co2 in 2000 and 2005 are not equal

Hypothesis Test Summary

	Null Hypothesis	Test	Sig.^{a,b}	Decision
1	The median of differences between share_global_co2_2000 and share_global_co2_2005 equals 0.	Related-Samples Wilcoxon Signed Rank Test	.003	Reject the null hypothesis.

a. The significance level is .050.
b. Asymptotic significance is displayed.

Related-Samples Wilcoxon Signed Rank Test Summary

Total N	231
Test Statistic	1484.500
Standard Error	268.937
Standardized Test Statistic	-2.958
Asymptotic Sig.(2-sided test)	.003

The p-value(0.003) is less than the significance level(0.05), the decision is to reject the null hypothesis. We have enough evidence to conclude that the difference between the population medians is statistically significant.

Check the co2 from coal consumption between different continents, if they exhibit a similar pattern for consumption of coal ?

The annual co2 from coal (coal_co2) consumption for continents Asia, Europe and Africa is considered for the hypothesis testing.

First check the normal distribution for each continent to decide which method to use for hypothesis testing.Descriptive statistics, histogram with normal curve and normal q-q plot is plotted for each continent.

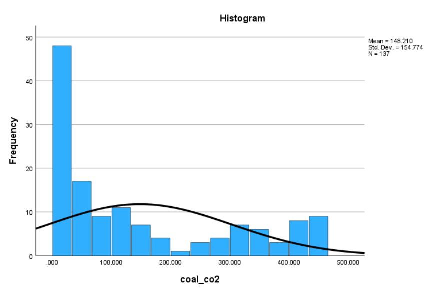
Africa:

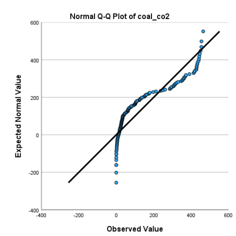
Statistics

coal_co2

N	Valid	137
	Missing	0
Mean		148.20960
Median		78.90200
Mode		.048[a]
Skewness		.831
Std. Error of Skewness		.207
Kurtosis		-.848
Std. Error of Kurtosis		.411

a. Multiple modes exist. The smallest value is shown





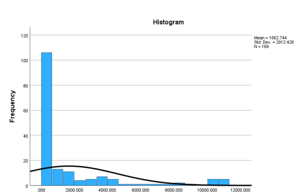
The sample size is 137. So, computing the z scores(Skewness: coal_co2=4.01   Kurtosis: coal_co2:=2.06). The distribution of Co2 from coal for Africa (mean = 148.209) is not normally distributed based on skewness and kurtosis value.

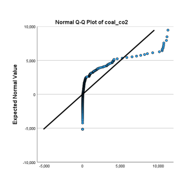
Asia:

Statistics

coal_co2

N	Valid	169
	Missing	22
Mean		1662.74417
Median		218.75000
Mode		.002
Skewness		2.166
Std. Error of Skewness		.187
Kurtosis		3.771
Std. Error of Kurtosis		.371





The sample size is 169. So, computing the z scores(Skewness: coal_co2:=11.58   Kurtosis: coal_co2:=10.16). The distribution of Co2 from coal for Asia(mean = 1662.74) is not normally distributed based on observation.

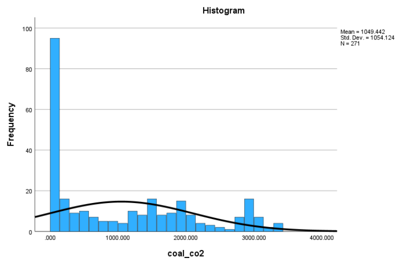
Europe:

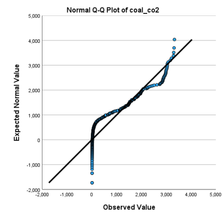
Statistics

coal_co2

N	Valid	271
	Missing	0
Mean		1049.44153
Median		707.37700
Mode		9.351[a]
Skewness		.627
Std. Error of Skewness		.148
Kurtosis		-.946
Std. Error of Kurtosis		.295

a. Multiple modes exist. The smallest value is shown





Skewness: coal_co2:=6.756   Kurtosis: coal_co2:=3.206

A distribution is called approximate normal if skewness or kurtosis (of the data are between 1 and 1. Although this is a less reliable method in the small to moderate sample size (i.e n 300 because it can not adjust the standard error (as the sample size increases, the standard error decreases) .To overcome this problem, a z test is applied for the normality test using skewness and kurtosis A. medium sized samples 50 n 300 at absolute z value±3.29 conclude the distribution of the sample is normal.

Since,the sample is medium sized and does not satisfy the condition of z value requirement . So, it is concluded that the three continents Asia, Africa and Europe do not have normally distributed co2 emission from coal data.

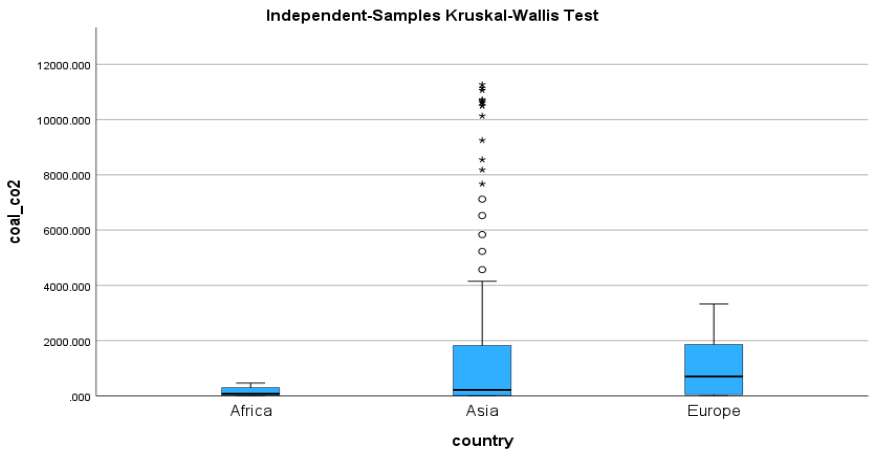
Independent -samples kruskal wallis test
H0 : There is no difference between countries emission of co2 from coal
H1: There is a difference between countries emission of co2 from coal

Hypothesis Test Summary

	Null Hypothesis	Test	Sig.[a,b]	Decision
1	The distribution of coal_co2 is the same across categories of country.	Independent-Samples Kruskal-Wallis Test	<.001	Reject the null hypothesis.

a. The significance level is .050.
b. Asymptotic significance is displayed.

The p-value is less than the significance level. So, a decision is made to reject the null hypothesis.So, The distribution of co2 emission from coal is not the same across different continents.



The boxplot for different continents shows that Asia has some extreme outliers.And their median does not lie in the same line. From the visualization; also we can say that the distribution is not equivalent.

Pairwise Comparisons of country

Sample 1-Sample 2	Test Statistic	Std. Error	Std. Test Statistic	Sig.	Adj. Sig.[a]
Africa-Asia	-91.087	19.165	-4.753	<.001	.000
Africa-Europe	-127.446	17.476	-7.293	<.001	.000
Asia-Europe	-36.359	16.340	-2.225	.026	.078

Each row tests the null hypothesis that the Sample 1 and Sample 2 distributions are the same.
Asymptotic significances (2-sided tests) are displayed. The significance level is .050.
a. Significance values have been adjusted by the Bonferroni correction for multiple tests.

From the pairwise comparisons of each country, we could not find any pair following a similar distribution.

Conclusion:

1. Data is not normally distributed.
2. There is a direct relationship between population and co2 emission, gdp and co2 emission.
3. There is a significant difference in the share of global co2 in 2000 and 2005
4. There is a significant difference in the co2 emission from coal in different continents(Asia, Europe and Africa).

Appendix A:

Case Processing Summary

	Cases					
	Included		Excluded		Total	
	N	Percent	N	Percent	N	Percent
year * country	25204	100.0%	0	0.0%	25204	100.0%
co2 * country	23949	95.0%	1255	5.0%	25204	100.0%
consumption_co2 * country	3976	15.8%	21228	84.2%	25204	100.0%
co2_growth_prct * country	24931	98.9%	273	1.1%	25204	100.0%
co2_growth_abs * country	23585	93.6%	1619	6.4%	25204	100.0%
trade_co2 * country	3976	15.8%	21228	84.2%	25204	100.0%
co2_per_capita * country	23307	92.5%	1897	7.5%	25204	100.0%
consumption_co2_per_capita * country	3976	15.8%	21228	84.2%	25204	100.0%
share_global_co2 * country	23949	95.0%	1255	5.0%	25204	100.0%
cumulative_co2 * country	23949	95.0%	1255	5.0%	25204	100.0%
share_global_cumulative_co2 * country	23949	95.0%	1255	5.0%	25204	100.0%
co2_per_gdp * country	15389	61.1%	9815	38.9%	25204	100.0%
consumption_co2_per_gdp * country	3761	14.9%	21443	85.1%	25204	100.0%
co2_per_unit_energy * country	9141	36.3%	16063	63.7%	25204	100.0%
coal_co2 * country	17188	68.2%	8016	31.8%	25204	100.0%
cement_co2 * country	12248	48.6%	12956	51.4%	25204	100.0%
flaring_co2 * country	4382	17.4%	20822	82.6%	25204	100.0%
gas_co2 * country	8845	35.1%	16359	64.9%	25204	100.0%
oil_co2 * country	20539	81.5%	4665	18.5%	25204	100.0%
other_industry_co2 * country	1999	7.9%	23205	92.1%	25204	100.0%
cement_co2_per_capita * country	12218	48.5%	12986	51.5%	25204	100.0%
coal_co2_per_capita * country	16860	66.9%	8344	33.1%	25204	100.0%
flaring_co2_per_capita * country	4381	17.4%	20823	82.6%	25204	100.0%
gas_co2_per_capita * country	8835	35.1%	16369	64.9%	25204	100.0%
oil_co2_per_capita * country	20181	80.1%	5023	19.9%	25204	100.0%
other_co2_per_capita * country	1999	7.9%	23205	92.1%	25204	100.0%
trade_co2_share * country	3976	15.8%	21228	84.2%	25204	100.0%
share_global_cement_co2 * country	12248	48.6%	12956	51.4%	25204	100.0%
share_global_coal_co2 * country	17188	68.2%	8016	31.8%	25204	100.0%
share_global_flaring_co2 * country	4382	17.4%	20822	82.6%	25204	100.0%
share_global_gas_co2 * country	8845	35.1%	16359	64.9%	25204	100.0%
share_global_oil_co2 * country	20539	81.5%	4665	18.5%	25204	100.0%
share_global_other_co2 * country	1999	7.9%	23205	92.1%	25204	100.0%
cumulative_cement_co2 * country	12248	48.6%	12956	51.4%	25204	100.0%
cumulative_coal_co2 * country	17188	68.2%	8016	31.8%	25204	100.0%
cumulative_flaring_co2 * country	4382	17.4%	20822	82.6%	25204	100.0%
cumulative_gas_co2 * country	8845	35.1%	16359	64.9%	25204	100.0%
cumulative_oil_co2 * country	20539	81.5%	4665	18.5%	25204	100.0%
cumulative_other_co2 * country	1999	7.9%	23205	92.1%	25204	100.0%
share_global_cumulative_cement_co2 * country	12248	48.6%	12956	51.4%	25204	100.0%
share_global_cumulative_coal_co2 * country	17188	68.2%	8016	31.8%	25204	100.0%
share_global_cumulative_flaring_co2 * country	4382	17.4%	20822	82.6%	25204	100.0%
share_global_cumulative_gas_co2 * country	8845	35.1%	16359	64.9%	25204	100.0%
share_global_cumulative_oil_co2 * country	20539	81.5%	4665	18.5%	25204	100.0%
share_global_cumulative_other_co2 * country	1999	7.9%	23205	92.1%	25204	100.0%
total_ghg * country	5208	20.7%	19996	79.3%	25204	100.0%
ghg_per_capita * country	5155	20.5%	20049	79.5%	25204	100.0%
methane * country	5211	20.7%	19993	79.3%	25204	100.0%
methane_per_capita * country	5157	20.5%	20047	79.5%	25204	100.0%
nitrous_oxide * country	5211	20.7%	19993	79.3%	25204	100.0%
nitrous_oxide_per_capita * country	5157	20.5%	20047	79.5%	25204	100.0%
population * country	22878	90.8%	2326	9.2%	25204	100.0%
gdp * country	13538	53.7%	11666	46.3%	25204	100.0%
primary_energy_consumption * country	8690	34.5%	16514	65.5%	25204	100.0%
energy_per_capita * country	8681	34.4%	16523	65.6%	25204	100.0%
energy_per_gdp * country	6803	27.0%	18401	73.0%	25204	100.0%

Appendix B:

Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation	Skewness		Kurtosis	
	Statistic	Statistic	Statistic	Statistic	Statistic	Statistic	Std. Error	Statistic	Std. Error
co2	23949	.000	36702.503	267.86194	1521.680894	13.598	.016	240.989	.032
co2_growth_prct	24931	-99.64	102318.51	21.0986	702.56599	126.420	.016	18085.891	.031
co2_growth_abs	23585	-1895.244	1736.258	5.14689	55.259760	6.362	.016	235.661	.032
co2_per_capita	23307	.000	748.639	4.17081	14.912201	27.079	.016	1027.565	.032
share_global_co2	23949	.00	100.00	4.9840	17.70499	4.445	.016	19.502	.032
cumulative_co2	23949	.000	1696524.177	10357.10451	61206.025253	13.886	.016	267.049	.032
share_global_cumulative_co2	23949	.00	100.00	5.1265	18.48008	4.299	.016	17.908	.032
co2_per_gdp	15389	.000	7.776	.42179	.483864	4.714	.020	44.190	.039
coal_co2	17188	.000	15062.902	175.35817	786.106838	10.396	.019	141.638	.037
oil_co2	20539	.000	12229.642	106.25438	602.683622	12.398	.017	192.921	.034
coal_co2_per_capita	16860	.000	34.184	1.55152	2.552112	3.729	.019	25.534	.038
oil_co2_per_capita	20181	.000	748.639	2.63550	15.129275	29.902	.017	1130.214	.034
share_global_coal_co2	17188	.00	100.00	6.9898	20.76197	3.609	.019	12.297	.037
share_global_oil_co2	20539	.00	100.00	2.9935	12.01868	6.256	.017	42.682	.034
cumulative_coal_co2	17188	.000	788362.044	8791.76718	39131.720328	9.331	.019	121.902	.037
cumulative_oil_co2	20539	.000	592621.162	3296.58423	21645.263301	15.003	.017	297.267	.034
share_global_cumulative_coal_co2	17188	.00	100.00	7.2118	21.63650	3.488	.019	11.222	.037
share_global_cumulative_oil_co2	20539	.00	100.00	3.0016	12.13410	6.152	.017	41.118	.034
population	22878	1490.0	7794798725.0	70723221.101	379585833.96	11.381	.016	162.393	.032
gdp	13538	55432000.000	1.136E+14	2.87709E+11	2.180094E+12	37.065	.021	1717.214	.042
Valid N (listwise)	8901								

get missing value statistics by country group

Code by Sanjina Poudel - jina.poul@gmail.com

Dataset: Co2 Emission

```
import matplotlib.pyplot as plt
import pandas as pd

df1 = pd.read_csv('owid-co2-data.csv')
```

```
df1.head()
```

	iso_code	country	year	co2	consumption_co2	co2_growth_prct	co2_growth_abs	trade_co2	co2_per_capita	consumption_co
0	AFG	Afghanistan	1949	0.015	NaN	NaN	NaN	NaN	0.002	
1	AFG	Afghanistan	1950	0.084	NaN	475.0	0.070	NaN	0.011	
2	AFG	Afghanistan	1951	0.092	NaN	8.7	0.007	NaN	0.012	
3	AFG	Afghanistan	1952	0.092	NaN	0.0	0.000	NaN	0.012	
4	AFG	Afghanistan	1953	0.106	NaN	16.0	0.015	NaN	0.013	

5 rows × 58 columns

```
df_country = df1.groupby('country').apply(lambda x: x.isna().sum())
df_country.head()
```

country	iso_code	country	year	co2	consumption_co2	co2_growth_prct	co2_growth_abs	trade_co2	co2_per_capita	consumpti
Afghanistan	0	0	0	0	72	1	1	72	0	
Africa	137	0	0	0	107	1	1	107	0	
Albania	0	0	0	0	58	1	1	58	0	
Algeria	0	0	0	0	105	1	1	105	0	

```
print(len(df_country))
```

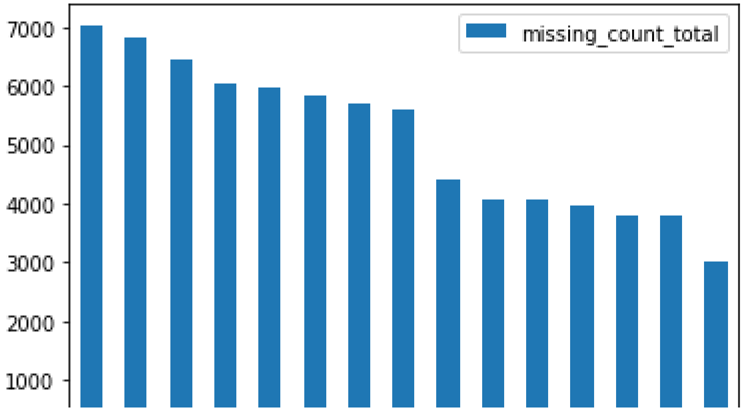
```
51
```

```
df_country['missing_count_total'] = df_country.sum(axis=1, numeric_only= True)
```

```
df_country = df1.groupby('country').apply(lambda x: x.isna().sum())
df_country['missing_count_total'] = df_country.sum(axis=1, numeric_only= True)
df_country = df_country.sort_values(by=['missing_count_total'], ascending = False)
new_df = df_country[['missing_count_total']].copy()
new_df['country'] = new_df.index
```

```
df_country = df_country.sort_values(by=['missing_count_total'], ascending = False)
new_df = df_country[['missing_count_total']].copy()
new_df['country'] = new_df.index
dff = new_df.head(15)
plt.figure(figsize=(30,5))
dff.plot(kind='bar')
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x7f948c43fe50>
<Figure size 2160x360 with 0 Axes>
```




```
total_data = len(df1)
new_df['percent']=(new_df['missing_count_total'] / total_data) * 100
```

```
dff2 = new_df.tail(50)
plt.figure(figsize=(30,5))
dff2.plot(x='country', y='percent',kind='bar')
```

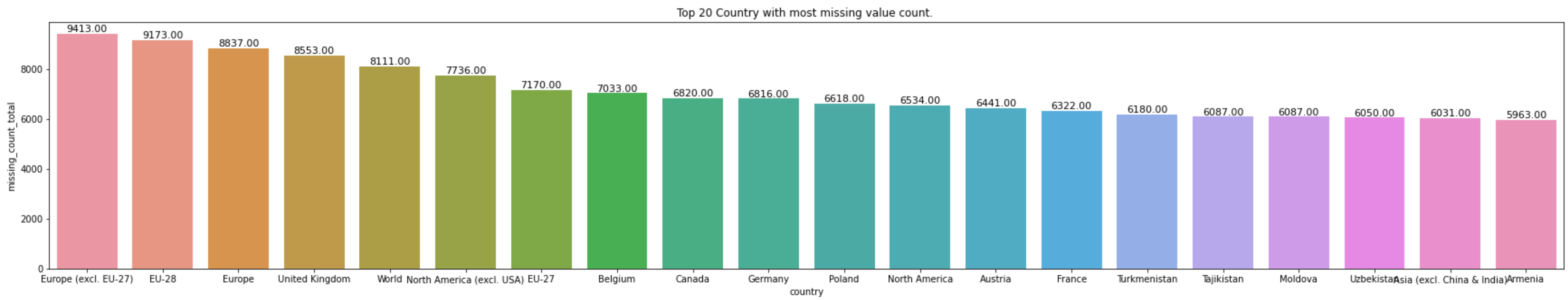
```
<matplotlib.axes._subplots.AxesSubplot at 0x7f948c30cc10>
<Figure size 2160x360 with 0 Axes>
```


```
new_df.tail(10)
```

	missing_count_total	country
country		
Tuvalu	1142	Tuvalu
Andorra	1142	Andorra
United Arab Emirates	1137	United Arab Emirates
Oman	1113	Oman
Wallis and Futuna	902	Wallis and Futuna
Liechtenstein	877	Liechtenstein
Lesotho	821	Lesotho
Timor	743	Timor
Namibia	651	Namibia
Kosovo	465	Kosovo

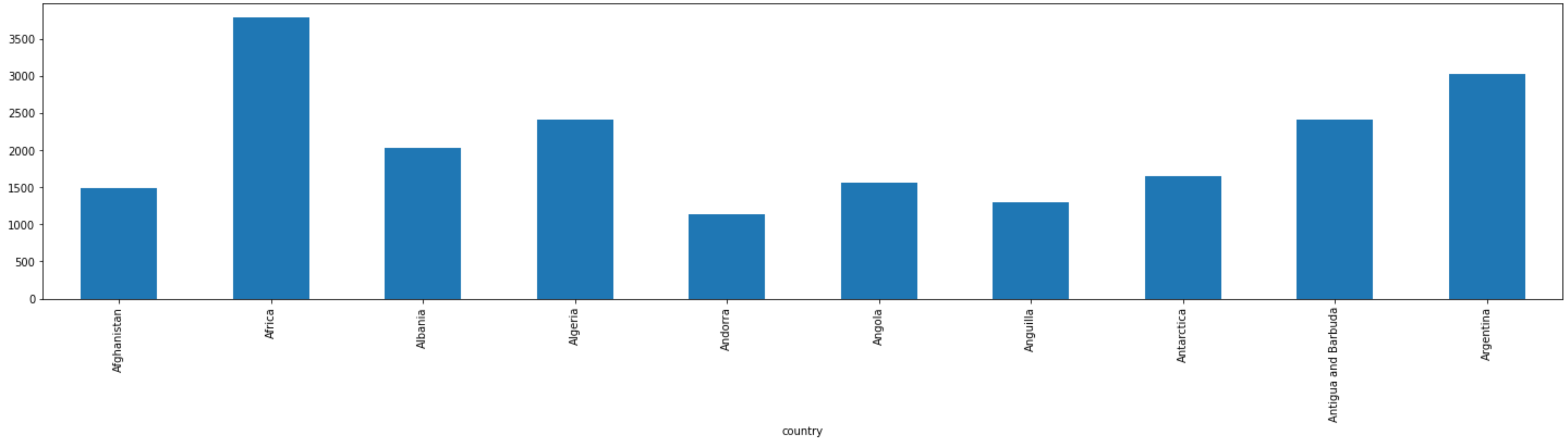
```
# Importing libraries for dataframe creation
# and graph plotting
import numpy as np
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt

df = new_df.head(20)
# Defining the plot size
plt.figure(figsize=(30, 5))

# Defining the values for x-axis, y-axis
# and from which dataframe the values are to be picked
plots = sns.barplot(x='country', y='missing_count_total', data=df)
```

```
# Iterrating over the bars one-by-one
for bar in plots.patches:

    # Using Matplotlib's annotate function and
    # passing the coordinates where the annotation shall be done
    plots.annotate(format(bar.get_height(), '.2f'),
                   (bar.get_x() + bar.get_width() / 2,
                    bar.get_height()), ha='center', va='center',
                   size=11, xytext=(0, 5),
                   textcoords='offset points')


# Setting the title for the graph
plt.title("Top 20 Country with most missing value count. ")

# Finally showing the plot
plt.show()
```




```
plt.figure(figsize=(25,5))

print(len(df_country))
dff = df_country.head(10)
dff.sum(axis=1). plot(kind='bar')
#.plot(kind='bar')
```

```
244
<matplotlib.axes._subplots.AxesSubplot at 0x7faa320cbb50>
```



Get missing value for each variable and save in csv

```
#df_count =df.isna().sum()
df_csv = pd.read_csv('owid-co2-data.csv')
total_rows = len(df_csv)
total_columns = df_csv.shape[1]
print("total_rows: ", total_rows)
print("total_columns: ", total_columns)
```

```
total_rows:  25204
total_columns:  58
```

```
df_missing_variable_count= df_csv.isna().sum()
#df_missing_variable_count= df_csv.isna().sum()[df_csv.isna().sum()>0]
df_missing_variable_count=df_missing_variable_count.sort_values(ascending=False)
df_missing_variable=pd.DataFrame({'feature':df_missing_variable_count.index, 'missing_number':df_missing_variable_count.values})
#only 2 variables do not have missing value
```

```
df_missing_variable['percent']= (df_missing_variable['missing_number'] / total_rows) * 100
```

```
print("total columns that have missing values : ", len(df_missing_variable))
```

```
total columns that have missing values :  58
```

```
df_missing_variable.head()
```

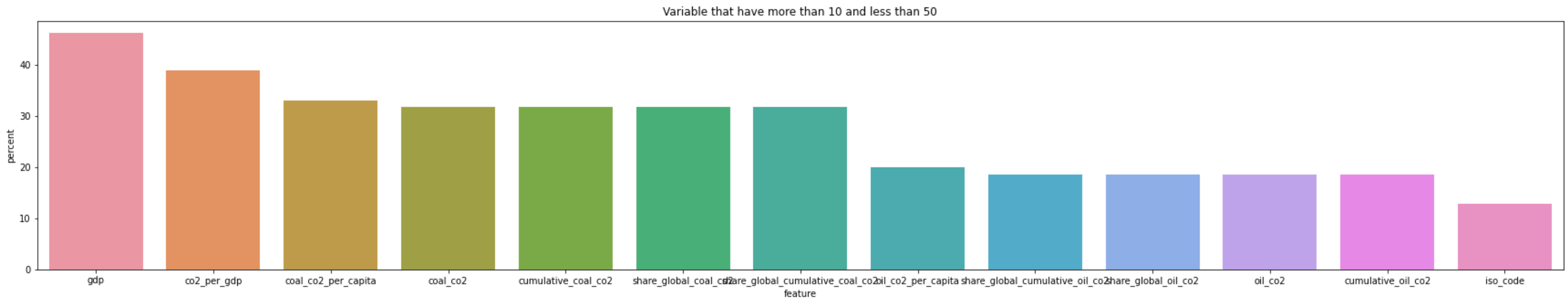
	feature	missing_number	percent
0	share_global_cumulative_other_co2	23205	92.068719
1	share_global_other_co2	23205	92.068719
2	cumulative_other_co2	23205	92.068719
3	other_co2_per_capita	23205	92.068719
4	other_industry_co2	23205	92.068719

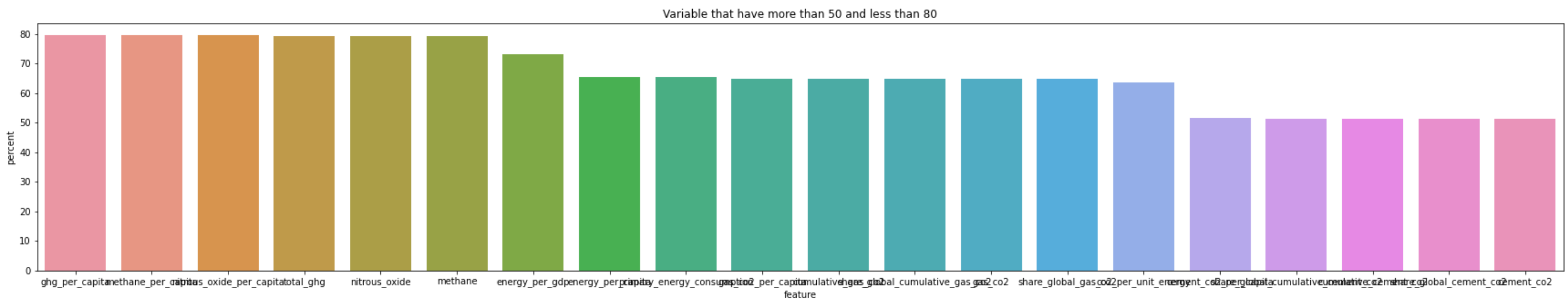
```
#filter features that have missing percentage more than 10 percentage
df_missing_more_than_10_less_than_50 =df_missing_variable.loc[(df_missing_variable['percent'] >= 10) & (df_missing_variable['percent']
df_missing_more_than_50_less_than_80 =df_missing_variable.loc[(df_missing_variable['percent'] > 50) & (df_missing_variable['percent']
df_missing_more_than_80_less_than_100 =df_missing_variable.loc[(df_missing_variable['percent'] > 80) & (df_missing_variable['percent']
print(len(df_missing_more_than_10_less_than_50))
print(len(df_missing_more_than_50_less_than_80))
print(len(df_missing_more_than_80_less_than_100))
```

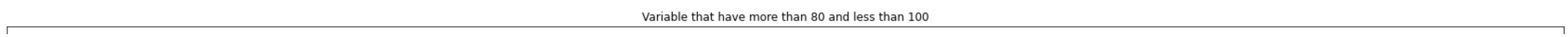
```
13
20
15
```

```
import matplotlib.pyplot as plt
plt.figure(figsize=(30,5))
```

```
#df_missing_more_than_10.plot(x='feature',y='percent', kind='bar')
sns.barplot(x='feature', y='percent', data=df_missing_more_than_10_less_than_50)
plt.title('Variable that have more than 10 and less than 50')
plt.show()

plt.figure(figsize=(30,5))
sns.barplot(x='feature', y='percent', data=df_missing_more_than_50_less_than_80)
plt.title('Variable that have more than 50 and less than 80')
plt.show()

plt.figure(figsize=(30,5))
sns.barplot(x='feature', y='percent', data=df_missing_more_than_80_less_than_100)
plt.title('Variable that have more than 80 and less than 100')
plt.show()
```







Choose countries that have less than 10 percentage of missing values

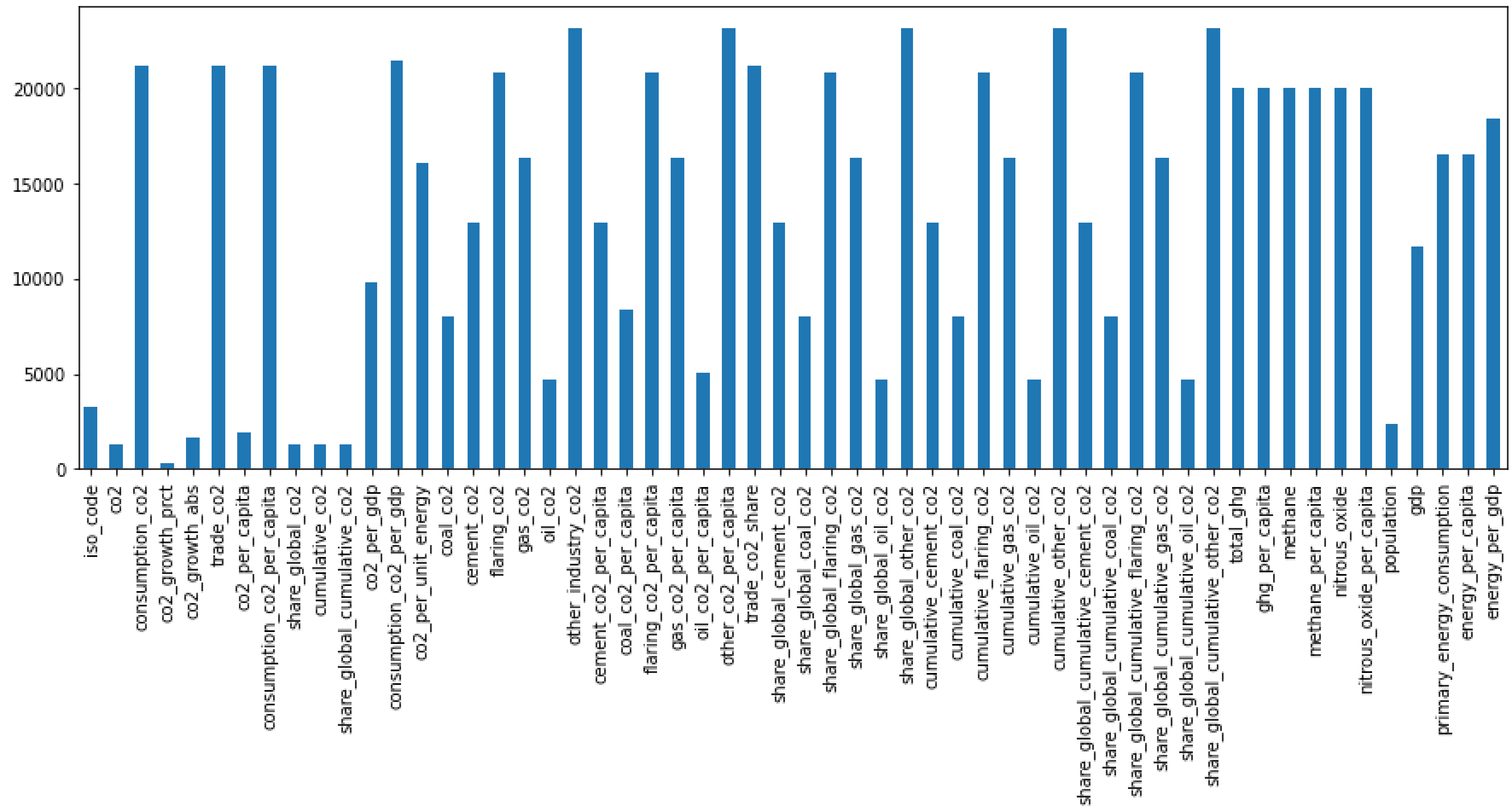

```
df_missing_variable.head()
```

```
df_count =df.isna().sum()
df_missing= df.isna().sum()[df.isna().sum()>0]



sorted = df_missing.sort_values()


df_miss=pd.DataFrame({'feature':sorted.index, 'missing number':sorted.values})
df_miss.head()
```

	feature	missing number
0	co2_growth_prct	273
1	co2	1255
2	share_global_co2	1255

```
import matplotlib.pyplot as plt
plt.figure(figsize=(15,5))
df.isna().sum()[df.isna().sum()>0].plot(kind='bar')
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x7fb0c5309750>
```




```
df_country = df.groupby('country').apply(lambda x: x.isna().sum())
```

```
df_country.head()
```

country	iso_code	country	year	co2	consumption_co2	co2_growth_prct	co2_growth_abs	trade_co2	co2_per_capita	consumpti
Afghanistan	0	0	0	0	72	1	1	72	0	
Africa	137	0	0	0	107	1	1	107	0	
Albania	0	0	0	0	58	1	1	58	0	
Algeria	0	0	0	0	105	1	1	105	0	
Andorra	0	0	0	0	31	1	1	31	0	

5 rows × 58 columns

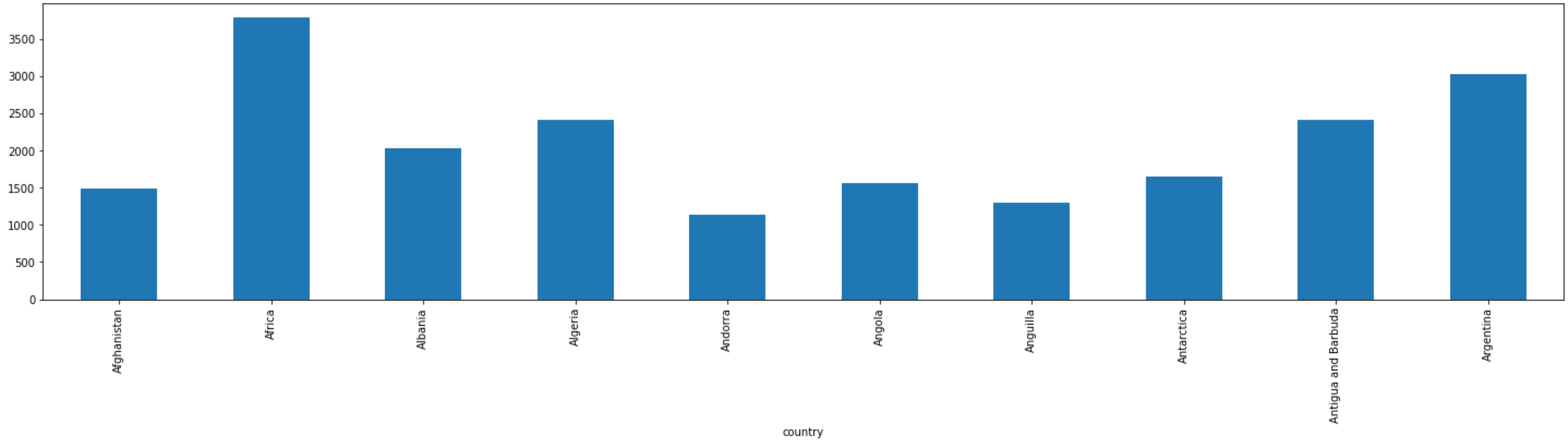
```
df_country.head()
```

country	iso_code	country	year	co2	consumption_co2	co2_growth_prct	co2_growth_abs	trade_co2	co2_per_capita	consumpti
Afghanistan	0	0	0	0	72	1	1	72	0	
Africa	137	0	0	0	107	1	1	107	0	
Albania	0	0	0	0	58	1	1	58	0	
Algeria	0	0	0	0	105	1	1	105	0	
Andorra	0	0	0	0	31	1	1	31	0	

5 rows × 58 columns

```
plt.figure(figsize=(25,5))

print(len(df_country))
dff = df_country.head(10)
dff.sum(axis=1). plot(kind='bar')
#.plot(kind='bar')
```

```
244
<matplotlib.axes._subplots.AxesSubplot at 0x7fb0bff18a50>
```



Double-click (or enter) to edit

Cleaning

```
dfwhole = pd.read_csv('owid-co2-data.csv')
print(len(dfwhole["country"].unique()))
```

```
244
```

```
total_rows = len(dfwhole)
df_missing_variable_count= dfwhole.isna().sum()
df_missing_variable_count=df_missing_variable_count.sort_values(ascending=False)
df_missing_variable=pd.DataFrame({'feature':df_missing_variable_count.index, 'missing_number':df_missing_variable_count.values})
df_missing_variable['percent']= (df_missing_variable['missing_number'] / total_rows) * 100

# Filter out teh variables that have more than 70 percentage of missing value
```

```
#filter features that have missing percentage more than 10 percentage
df_missing_more_than_70_less_than_1000 =df_missing_variable.loc[(df_missing_variable['percent'] > 70) & (df_missing_variable['percent'
print(len(df_missing_more_than_70_less_than_1000))
```

```
22
```

```
df_missing_more_than_70_less_than_1000.head()
```

	feature	missing_number	percent
0	share_global_cumulative_other_co2	23205	92.068719
1	share_global_other_co2	23205	92.068719
2	cumulative_other_co2	23205	92.068719
3	other_co2_per_capita	23205	92.068719
4	other_industry_co2	23205	92.068719

```
to_be_drop_columns = df_missing_more_than_70_less_than_1000['feature']
#Dataframe after dropping columns checking for the missing value
df2 = dfwhole.drop(to_be_drop_columns, axis=1)
df2.shape
print(len(df2["country"].unique()))
```

```
244
```

```
# Analysis missing data
total_data2 = len(df2)
df2_country = df2.groupby('country').apply(lambda x: x.isna().sum())
df2_country['missing_count_total'] = df2_country.sum(axis=1, numeric_only= True)

new_df = df2_country[['missing_count_total']].copy()
new_df['country'] = new_df.index
new_df['percent']=(new_df['missing_count_total'] / total_data2) * 100
```

```
new_df.head()
```

country	missing_count_total	country	percent
Afghanistan	263	Afghanistan	1.043485
Africa	1219	Africa	4.836534
Albania	438	Albania	1.737819
Algeria	569	Algeria	2.257578

```
variable_more_50 =df_missing_variable.loc[(df_missing_variable['percent'] > 50) & (df_missing_variable['percent'] <= 100)]
variable_less_50 =df_missing_variable.loc[(df_missing_variable['percent'] >= 0) & (df_missing_variable['percent'] <= 50)]

df_more_50 = dfwhole.drop(variable_less_50['feature'], axis=1)
df_less_50 = dfwhole.drop(variable_more_50['feature'], axis=1)

df_more_50.to_csv("co2_dataset_more_missing.csv")
df_less_50.to_csv("co2_dataset_less_missing.csv")
```

```
dfwhole.shape
```

```
(25204, 58)
```

```
df_less_50.shape
```

```
(25204, 23)
```

```
new_df.describe()
```

	missing_count_total	percent
count	244.000000	244.000000
mean	1211.795082	4.807947

```
# Drop country that have more than 10 percentage of missing value
df_country_missing_more_than_10 =new_df.loc[new_df['percent'] >10]
to_be_drop_countries = df_country_missing_more_than_10['country']
#Dataframe after dropping columns checking for the missing value
df3 = dfwhole[dfwhole['country'].isin(to_be_drop_countries)]
print(len(df3["country"].unique()))
```

```
# Drop country that have more than 10 percentage of missing value
df_country_missing_less_than_10 =new_df.loc[new_df['percent'] <= 10]
to_be_add_countries2 = df_country_missing_less_than_10['country']
df4 = dfwhole[dfwhole['country'].isin(to_be_add_countries2)]
print(len(df4["country"].unique()))
```

```
234
```

```
df5 = df2[df2['country'].isin(to_be_add_countries2)]
```

```
#convert to csv file
df5.head()
```

	iso_code	country	year	co2	co2_growth_prct	co2_growth_abs	co2_per_capita	share_global_co2	cumulative_co2	share_gl
0	AFG	Afghanistan	1949	0.015	NaN	NaN	0.002	0.0	0.015	
1	AFG	Afghanistan	1950	0.084	475.0	0.070	0.011	0.0	0.099	
2	AFG	Afghanistan	1951	0.092	8.7	0.007	0.012	0.0	0.191	
3	AFG	Afghanistan	1952	0.092	0.0	0.000	0.012	0.0	0.282	
4	AFG	Afghanistan	1953	0.106	16.0	0.015	0.013	0.0	0.388	

5 rows × 36 columns

```
df5.to_csv("co2dataset1.csv")
```

Re;ove by country

```
import matplotlib.pyplot as plt
import pandas as pd
dfwhole1 = pd.read_csv('owid-co2-data.csv')
print(len(dfwhole1["country"].unique()))
```

```
244
```

```
total_data3 = len(dfwhole1)
dfwhole1_country = dfwhole1.groupby('country').apply(lambda x: x.isna().sum())
dfwhole1_country['missing_count_total'] = dfwhole1_country.sum(axis=1, numeric_only= True)

new_dfwhole1 = dfwhole1_country[['missing_count_total']].copy()
new_dfwhole1['country'] = new_dfwhole1.index
new_dfwhole1['percent']=(new_dfwhole1['missing_count_total'] / total_data3) * 100
```

```
new_dfwhole1.describe()
```

	missing_count_total	percent
count	244.000000	244.000000
mean	3117.602459	12.369475
std	1716.200388	6.809238
min	465.000000	1.844945
25%	2010.000000	7.974925
50%	2580.500000	10.238454
75%	3838.250000	15.228734
max	9413.000000	37.347246

```
#Get countries ith more than 10 percentage of missing value
df_more_10_country =new_dfwhole1.loc[new_dfwhole1['percent'] > 10]
countries_more_than_10 = df_more_10_country['country']

df_less_10_country =new_dfwhole1.loc[new_dfwhole1['percent'] <= 10]
countries_less_than_10 = df_less_10_country['country']


print(len(countries_more_than_10))
print(len(countries_less_than_10))
```

```
125
119
```

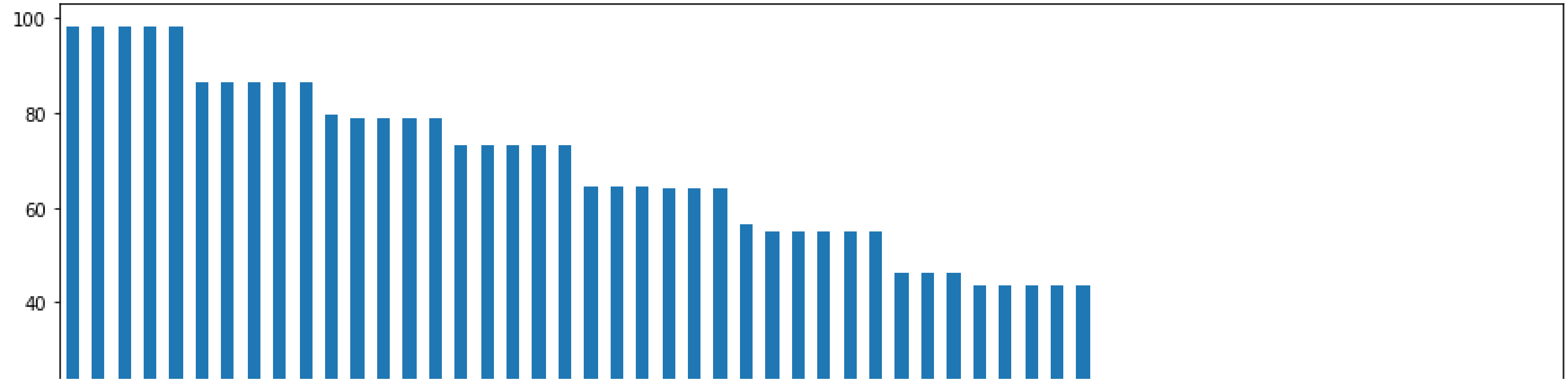
```
dfwhole2 = dfwhole1[dfwhole1['country'].isin(countries_less_than_10)]
```

```
total_rows2 = len(dfwhole2)
df_missing_variable_count2= dfwhole2.isna().sum()
df_missing_variable_count2=df_missing_variable_count2.sort_values(ascending=False)
df_missing_variable2=pd.DataFrame({'feature':df_missing_variable_count2.index, 'missing_number':df_missing_variable_count2.values})
df_missing_variable2['percent']= (df_missing_variable2['missing_number'] / total_rows2) * 100

# Filter out teh variables that have more than 70 percentage of missing value
#filter features that have missing percentage more than 10 percentage
df_missing_more_than_70_less_than_1002 =df_missing_variable2.loc[(df_missing_variable2['percent'] > 70) & (df_missing_variable2['perce


print(len(df_missing_more_than_70_less_than_1002))
```

```
20
```

```
import matplotlib.pyplot as plt
plt.figure(figsize=(15,5))
df_missing_variable2['percent'].plot(kind='bar')
```

<matplotlib.axes._subplots.AxesSubplot at 0x7f02ee5df6d0>




```
df_missing_more_than_70_less_than_1002.head()
```

	feature	missing_number	percent
0	share_global_cumulative_other_co2	7652	98.014602
1	other_co2_per_capita	7652	98.014602
2	share_global_other_co2	7652	98.014602
3	other_industry_co2	7652	98.014602
4	cumulative_other_co2	7652	98.014602

deal with missing value

```
import matplotlib.pyplot as plt
import pandas as pd
dff = pd.read_csv('co2_dataset_less_missing.csv')
print(len(dff["country"].unique()))
```

```
244
```

```
dn=dff.copy()
```

```
dn=dff.copy()
#Co2, co2_per_Capita, co2_per_gdp
```

```
#co2, co2_per_capita, co2_per_gdp
var1='co2'
var_per_capita = 'co2_per_capita'
var_per_gdp='co2_per_gdp'
population = 'population'

missing_indexes1 = dn[dn[var1].isnull()].index.tolist()
for ind in missing_indexes1:
  dn.loc[ind, 'co2']=dn.loc[ind, 'co2_per_capita'] * dn.loc[ind, 'population']

missing_indexes2 = dn[dn[var_per_capita].isnull()].index.tolist()
for ind in missing_indexes2:
  dn.loc[ind, 'co2_per_capita']=dn.loc[ind, 'co2'] / dn.loc[ind, 'population']

missing_indexes3 = dn[dn[var_per_gdp].isnull()].index.tolist()
for ind in missing_indexes3:
  dn.loc[ind, 'co2_per_gdp']=dn.loc[ind, 'co2'] / dn.loc[ind, 'gdp']


missing_indexes4 = dn[dn[population].isnull()].index.tolist()
for ind in missing_indexes4:
  dn.loc[ind, 'population']=dn.loc[ind, 'co2_per_capita'] * dn.loc[ind, 'population']



missing_indexes5 = dn[dn['coal_co2_per_capita'].isnull()].index.tolist()
for ind in missing_indexes5:
  dn.loc[ind, 'coal_co2_per_capita']=dn.loc[ind, 'coal_co2'] / dn.loc[ind, 'population']

missing_indexes5 = dn[dn['oil_co2_per_capita'].isnull()].index.tolist()
for ind in missing_indexes5:
  dn.loc[ind, 'oil_co2_per_capita']=dn.loc[ind, 'oil_co2'] / dn.loc[ind, 'population']


dn.to_csv("co2_dataset.csv")


len(missing_indexes5)

     5023


len(dn[dn['oil_co2_per_capita'].isnull()].index.tolist())

     4974
```

```
missing_indexes1 = dn[dn[var1].isnull()].index.tolist()
```

```
missing_indexes = dn[dn['co2_per_capita'].isnull()].index.tolist()
len(missing_indexes)
```

```
1897
```

```
for ind in missing_indexes:
  #dn.loc[ind, 'co2']=dn.loc[ind, 'co2_per_capita'] * dn.loc[ind, 'population']
  dn.loc[ind, 'co2_per_capita']=dn.loc[ind, 'co2'] / dn.loc[ind, 'population']
```

```
missing_indexes_after = dn[dn['co2_per_capita'].isnull()].index.tolist()
len(missing_indexes_after)
```

```
1848
```

Research questions

Is there any relation between gdp and co2 emission?

```
import matplotlib.pyplot as plt
import pandas as pd
dff = pd.read_csv('co2_dataset.csv')
print(len(dff["country"].unique()))
```

```
244
```

```
co2=dff['co2']
population = dff['population']
```

```
import numpy as np
N=len(co2)
```

```
colors = np.random.rand(N)
area = (30 * np.random.rand(N))**2
```

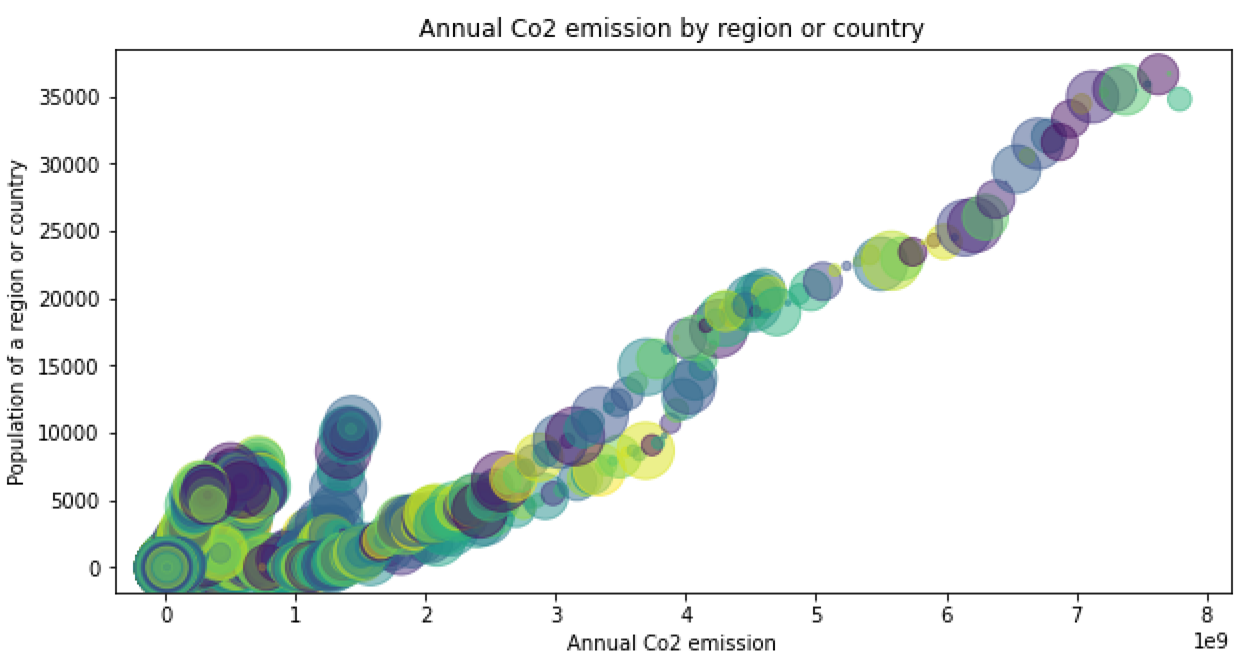
```
np.min(co2)
```

```
0.0
```

```
np.max(co2)
```

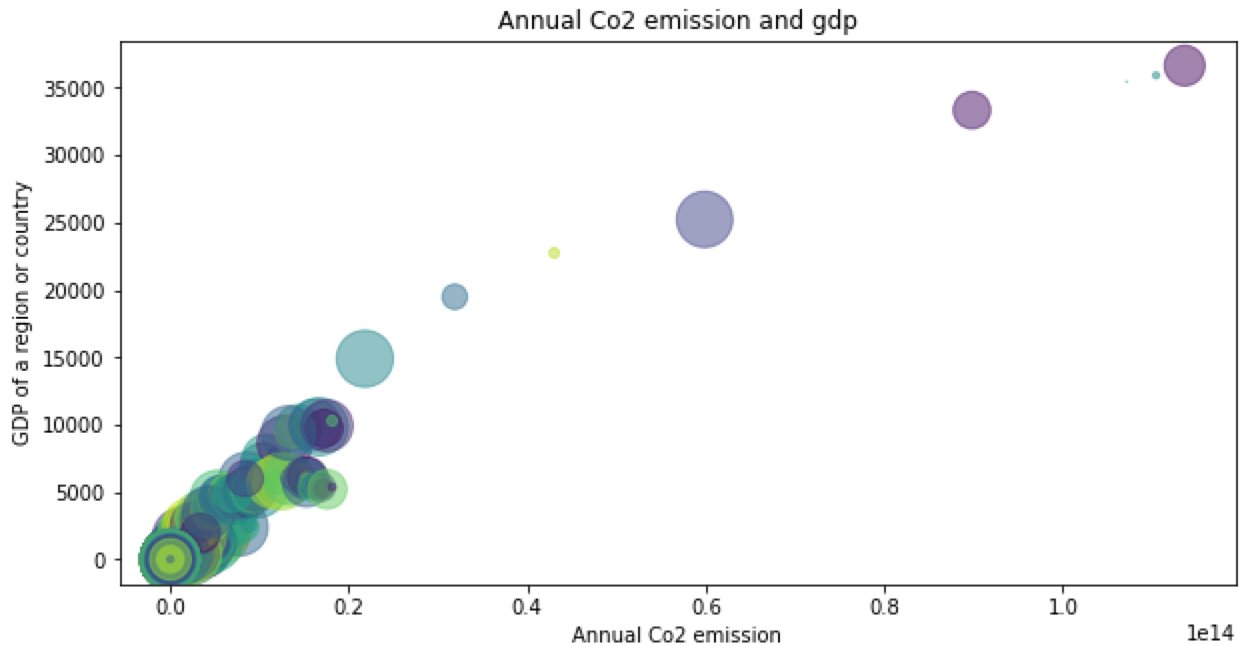
```
36702.503
```

```
plt.figure(figsize=(10,5))
plt.scatter(population, co2, s=area, c=colors, alpha = 0.5)
plt.title("Annual Co2 emission by region or country")
plt.xlabel("Annual Co2 emission")
plt.ylabel("Population of a region or country")
plt.show()
plt.savefig('Annual Co2 emission by region or country.png')
```




```
<Figure size 432x288 with 0 Axes>
```

```
gdp = dff['gdp']
plt.figure(figsize=(10,5))
plt.scatter(gdp, co2, s=area, c=colors, alpha = 0.5)
plt.title("Annual Co2 emission and gdp")
plt.xlabel("Annual Co2 emission")
plt.ylabel("GDP of a region or country")
plt.show()
plt.savefig('Annual Co2 emission and gdp.png')
```




```
<Figure size 432x288 with 0 Axes>
```

```
df_co2_sum=dff.groupby(['year'])[['co2','coal_co2','oil_co2']].sum()
```

```
df_co2_sum.plot()
plt.savefig('Annual Co2, co2bycoal co2by oil emission by year.png')
```


```
share_global_co2 = dff.loc[(dff['year'] == 2000) | (dff['year'] == 2005)]
share_global_co2.head()
```

	Unnamed: 0	Unnamed: 0.1	iso_code	country	year	co2	co2_growth_prct	co2_growth_abs	co2_per_capita	share_global_co2
51	51	51	AFG	Afghanistan	2000	0.758	-6.40	-0.052	0.036	0.00
56	56	56	AFG	Afghanistan	2005	1.303	46.57	0.414	0.051	0.00
188	188	188	NaN	Africa	2000	886.562	6.76	56.165	1.094	3.51
193	193	193	NaN	Africa	2005	1057.342	1.99	20.656	1.155	3.57
276	276	276	ALB	Albania	2000	3.004	0.99	0.029	0.960	0.01

5 rows × 25 columns

```
dff_filter = dff[['year','country', 'share_global_co2']].copy()
```

```
share_global_co2_2000 = dff_filter.loc[(dff_filter['year'] == 2000)]
share_global_co2_2005 = dff_filter.loc[(dff_filter['year'] == 2005)]
```

```
share_globa_co2=dff[['year', 'share_global_co2']].copy()
share_globa_co2=share_globa_co2.loc[(share_globa_co2['year'] == 2005) |(share_globa_co2['year'] == 2000)]
share_globa_co2.to_csv('share_globa_co2.csv')
```

```
share_global_co2_2000.rename(columns = {'share_global_co2':'share_global_co2_2000'}, inplace = True)
share_global_co2_2005.rename(columns = {'share_global_co2':'share_global_co2_2005'}, inplace = True)
```

```
/usr/local/lib/python3.7/dist-packages/pandas/core/frame.py:5047: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html#returning-a-view-ver
  errors=errors,
```

```
share_global_co2_2000=share_global_co2_2000.drop('year', axis=1)
share_global_co2_2005=share_global_co2_2005.drop('year', axis=1)
```

```
paired_Test_global_share_co2=pd.merge(share_global_co2_2000, share_global_co2_2005, on="country")
paired_Test_global_share_co2.head()
```

	country	share_global_co2_2000	share_global_co2_2005
0	Afghanistan	0.00	0.00
1	Africa	3.51	3.57
2	Albania	0.01	0.01
3	Algeria	0.33	0.36
4	Andorra	0.00	0.00

```
print(len(paired_Test_global_share_co2))
```

```
241
```

```
paired_Test_global_share_co2.to_csv('paired_Test_global_share_co2.csv')
```

```
#data from three countries
three_continent_co2 = dff.loc[(dff['country'] == 'Africa') | (dff['country'] == 'Asia')| (dff['country'] == 'Europe')]
three_continent_co2.head()
```

	Unnamed: 0	Unnamed: 0.1	iso_code	country	year	co2	co2_growth_prct	co2_growth_abs	co2_per_capita	share_global_co2	...
72	72	72	NaN	Africa	1884	0.022	NaN	NaN	0.005	0.00	...
73	73	73	NaN	Africa	1885	0.037	66.67	0.015	0.008	0.00	...
74	74	74	NaN	Africa	1886	0.048	30.00	0.011	0.010	0.00	...
75	75	75	NaN	Africa	1887	0.048	0.00	0.000	0.010	0.00	...
76	76	76	NaN	Africa	1888	0.081	69.23	0.033	0.017	0.01	...

```
three_continent_co2.to_csv('three_continent_co2.csv')
```

```
belgium_co2 = dff.loc[(dff['country'] == 'Belgium')]
belgium_co2.head()
belgium_co2.to_csv('belgium_co2.csv')
```

```
belgium_co2.describe()
```

	Unnamed: 0	Unnamed: 0.1	year	co2	co2_growth_prct	co2_growth_abs	co2_per_capita	share_global_co2	cumul
count	219.000000	219.000000	219.000000	192.000000	218.000000	190.000000	192.000000	192.000000	
mean	2628.000000	2628.000000	1911.000000	65.328906	1.664908	0.408168	7.653427	2.409792	4
std	63.364028	63.364028	63.364028	42.010392	9.728906	6.225100	3.635492	2.086307	3
min	2519.000000	2519.000000	1802.000000	4.408000	-42.720000	-27.440000	1.136000	0.240000	
25%	2573.500000	2573.500000	1856.500000	27.161000	-2.315000	-1.545500	4.844750	0.787500	
50%	2628.000000	2628.000000	1911.000000	61.487500	0.640000	0.512500	7.957500	1.980000	2
75%	2682.500000	2682.500000	1965.500000	102.371750	5.612500	2.583250	10.567500	3.847500	6
max	2737.000000	2737.000000	2020.000000	139.787000	45.560000	25.619000	14.262000	17.390000	12

8 rows × 23 columns

```
europe_co2 = dff.loc[(dff['country'] == 'Europe')]
europe_co2.head()
```

	Unnamed: 0	Unnamed: 0.1	iso_code	country	year	co2	co2_growth_prct	co2_growth_abs	co2_per_capita	share_global_co2	...
7502	7502	7502	NaN	Europe	1750	9.351	NaN	NaN	1.007	100.0	...
7503	7503	7503	NaN	Europe	1751	9.351	0.00	0.000	NaN	100.0	...
7504	7504	7504	NaN	Europe	1752	9.354	0.04	0.004	NaN	100.0	...
7505	7505	7505	NaN	Europe	1753	9.354	0.00	0.000	NaN	100.0	...
7506	7506	7506	NaN	Europe	1754	9.358	0.04	0.004	NaN	100.0	...

5 rows × 25 columns

```
europe_co2.describe()
```

	Unnamed: 0	Unnamed: 0.1	year	co2	co2_growth_prct	co2_growth_abs	co2_per_capita	share_global_co2	cumu
mean	7637.000000	7637.000000	1885.000000	1959.984454	2.555519	18.284033	3.997257	67.561734	9

Report the percentage of missing value of every country for each variable

```
import matplotlib.pyplot as plt
import pandas as pd
dfc = pd.read_csv('owid-co2-data.csv')
dfc_country = dfc.groupby('country').apply(lambda x: x.isna().sum())
dfc_country.head()
```

country	iso_code	country	year	co2	consumption_co2	co2_growth_prct	co2_growth_abs	trade_co2	co2_per_capita	consumpti
Afghanistan	0	0	0	0	72	1	1	72	0	
Africa	137	0	0	0	107	1	1	107	0	
Albania	0	0	0	0	58	1	1	58	0	
Algeria	0	0	0	0	105	1	1	105	0	
Andorra	0	0	0	0	31	1	1	31	0	

5 rows × 58 columns

```
dfc_country.to_csv("missing_value_each_variable_by_each_country.csv")
```

```
dfc_perc=(dfc_country/len(dfc))*100
```

```
dfc_perc.head()
```

country	iso_code	country	year	co2	consumption_co2	co2_growth_prct	co2_growth_abs	trade_co2	co2_per_capita	consumpti
Afghanistan	0.000000	0.0	0.0	0.0	0.285669	0.003968	0.003968	0.285669	0.0	
Africa	0.543565	0.0	0.0	0.0	0.424536	0.003968	0.003968	0.424536	0.0	
Albania	0.000000	0.0	0.0	0.0	0.230122	0.003968	0.003968	0.230122	0.0	
Algeria	0.000000	0.0	0.0	0.0	0.416601	0.003968	0.003968	0.416601	0.0	
Andorra	0.000000	0.0	0.0	0.0	0.122996	0.003968	0.003968	0.122996	0.0	

```
dfc_country.to_csv("missing_value_each_variable_by_each_country_in_percentage.csv")
```

```
dfc_test = dfc.apply(lambda x: x.isna().sum())
dfc_test.head()
```

```
iso_code            3256
country                0
year                   0
co2                 1255
consumption_co2    21228
dtype: int64
```

```
total_data = len(dfc)
df_country = dfc.groupby('country').apply(lambda x: x.isna().sum())
df_country['missing_count_total'] = df_country.sum(axis=1, numeric_only= True)
df_country = df_country.sort_values(by=['missing_count_total'], ascending = False)
new_df = df_country[['missing_count_total']].copy()
new_df['percent']=(new_df['missing_count_total'] / total_data) * 100
```

```
new_df.head()
```

	missing_count_total	percent
country		
Europe (excl. EU-27)	9413	37.347246

```
with pd.ExcelWriter('report.xlsx') as writer:
    new_df.to_excel(writer, sheet_name='mv_by_each_country')
    dfc_country.to_excel(writer, sheet_name='mv_variable_by_each_country')
    dfc_perc.to_excel(writer, sheet_name='mv_variable_by_each_country_percetage')
    dfc_test.to_excel(writer, sheet_name='mv_frequency_variable')
```

```
/usr/local/lib/python3.7/dist-packages/openpyxl/workbook/child.py:99: UserWarning: Title is more than 31 characters. Some applic
  warnings.warn("Title is more than 31 characters. Some applications may not be able to read the file")
```

```
iso_code            3256
country                0
year                   0
co2                 1255
consumption_co2    21228
dtype: int64
```

```
dfc_country.sum()
```

```
iso_code                                 3256
country                                     0
year                                        0
co2                                      1255
consumption_co2                         21228
co2_growth_prct                           273
co2_growth_abs                           1619
trade_co2                               21228
co2_per_capita                           1897
consumption_co2_per_capita              21228
share_global_co2                         1255
cumulative_co2                           1255
share_global_cumulative_co2              1255
```

```
co2_per_gdp                              9815
consumption_co2_per_gdp                 21443
co2_per_unit_energy                     16063
coal_co2                                 8016
cement_co2                              12956
flaring_co2                             20822
gas_co2                                 16359
oil_co2                                  4665
other_industry_co2                      23205
cement_co2_per_capita                   12986
coal_co2_per_capita                      8344
flaring_co2_per_capita                  20823
gas_co2_per_capita                      16369
oil_co2_per_capita                       5023
other_co2_per_capita                    23205
trade_co2_share                         21228
share_global_cement_co2                 12956
share_global_coal_co2                    8016
share_global_flaring_co2                20822
share_global_gas_co2                    16359
share_global_oil_co2                     4665
share_global_other_co2                  23205
cumulative_cement_co2                   12956
cumulative_coal_co2                      8016
cumulative_flaring_co2                  20822
cumulative_gas_co2                      16359
cumulative_oil_co2                       4665
cumulative_other_co2                    23205
share_global_cumulative_cement_co2      12956
share_global_cumulative_coal_co2         8016
share_global_cumulative_flaring_co2     20822
share_global_cumulative_gas_co2         16359
share_global_cumulative_oil_co2          4665
share_global_cumulative_other_co2       23205
total_ghg                               19996
ghg_per_capita                          20049
methane                                 19993
methane_per_capita                      20047
nitrous_oxide                           19993
nitrous_oxide_per_capita                20047
population                               2326
gdp                                     11666
primary_energy_consumption              16514
energy_per_capita                       16523
energy_per_gdp                          18401
```