A project report on:

Climate Change: An analysis on factors affecting CO_2 Emission

Introduction

A pressing matter at hand for more than a decade has been the severe changes in the weather pattern across the globe. The rate at which CO_2 is emitted per capita has been identified as the prime indicator for such changes in the climate. There are many factors that contribute towards carbon emissions, factors such as population, GDP, and fuel consumption. An upward trend in these factors across the globe has affected the amount of CO_2 emitted annually. This project employees a visual and data driven approach towards investigating correlations or effects each of these factors has on CO_2 emissions.

Related Work

The methods or techniques for the theoretical analysis of CO_2 emissions are not brand-new but have been inspired by previous works on carbon emission and its sources. Few of such studies that perfectly synthesizes this project include, research on the deriving factors of CO2 emissions using machine learning by Shanshan Li, Yam Wing Siu, and Guoqin Zhao from the Central University of Finance and Economics, Beijing, China. Their research aims at investigating the relation between carbon emission and economic growth, industry structure, urbanization, growth rate of energy and research and development. Their research focuses on China, which contribute nearly 28% to the global emission. They employed methods like K-nearest neighbors, and other nonlinear, ensemble models and artificial neural networks to conduct a sensitivity analysis on CO_2 emission around its centroid value. Using methods their research drew conclusions that not all provinces in China should industrialize but focus on research and development to contribute towards minimizing CO_2 emission. Another research that was highly pragmatic was by Ali Acaravci and Ilhan Ozturk from Mustafa Kemel University, Turky. Their research examines the casual relationship between carbon dioxide emissions, energy consumption and economic growth using autoregressive distributed lag (ARDL). They use bounds F-test to prove a long run relationship between CO_2 emission per capita, energy consumption and GDP. Their results support the environmental Kuznets Curve hypothesis. These studies have laid the foundation for the idea to explore and analyze the correlations between CO_2 emission and the factors responsible for it.

Motivation

Though there were a lot of visualizations and data found regarding carbon emission by certain countries, there was little work found on analysis of factors contributing to it. These factors play a very important role in understanding and discovering the major sources of carbon emission and ways to reduce it. Such analysis would help people understand how much of an imminent threat climate change is and how CO_2 emission is subsidizing it. A more sophisticated analysis can be built on this project that can help governments understand the urgent need to invest towards controlling aforementioned factors, which in-turn mitigates carbon emission.

Methodology

The right and efficient method for analysis is very important for generating and understanding prime information. The solution to an ever-changing problem is to employee flexible and adaptive analysis. The effects of CO_2 emissions on the population and life expectancy and the effects of GDP on CO_2 emission and vice versa is one such problem. This project emphasizes analyzing this problem through exploiting correlations. In the process of exploiting relations between attributes, we learn by how much one attribute of the variable we wish to study is influencing another one of its attributes. The process of analysis is only as good as the data. This project utilizes three different datasets, each describing a factor, contributing, or affecting the CO_2 emissions. The first of the datasets is a GDP data set named "GDP_data" containing country names, their alpha codes, GDP indicator, its code and a particular country's GDP in dollars.

	A	B	C	D	E
1	Country	Code	Indicator	Indicator_Code	GDP_Millions
2	Afghanistan	AFG	GDP (current US$)	NY.GDP.MKTP.CD	18353881130
3	Albania	ALB	GDP (current US$)	NY.GDP.MKTP.CD	15147020535
4	Algeria	DZA	GDP (current US$)	NY.GDP.MKTP.CD	1.75E+11
5	Angola	AGO	GDP (current US$)	NY.GDP.MKTP.CD	1.01E+11
6	Antigua and Barbuda	ATG	GDP (current US$)	NY.GDP.MKTP.CD	1605351852
7	Argentina	ARG	GDP (current US$)	NY.GDP.MKTP.CD	5.18E+11
8	Armenia	ARM	GDP (current US$)	NY.GDP.MKTP.CD	12457941907
9	Aruba	ABW	GDP (current US$)	NY.GDP.MKTP.CD	3202188607
10	Australia	AUS	GDP (current US$)	NY.GDP.MKTP.CD	1.43E+12
11	Austria	AUT	GDP (current US$)	NY.GDP.MKTP.CD	4.55E+11
12	Azerbaijan	AZE	GDP (current US$)	NY.GDP.MKTP.CD	47112941176
13	Bahamas, The	BHS	GDP (current US$)	NY.GDP.MKTP.CD	13022100000
14	Bahrain	BHR	GDP (current US$)	NY.GDP.MKTP.CD	37652500000
15	Bangladesh	BGD	GDP (current US$)	NY.GDP.MKTP.CD	2.74E+11
16	Barbados	BRB	GDP (current US$)	NY.GDP.MKTP.CD	5086500000
17	Belarus	BLR	GDP (current US$)	NY.GDP.MKTP.CD	60031262269
18	Belgium	BEL	GDP (current US$)	NY.GDP.MKTP.CD	5.43E+11
19	Belize	BLZ	GDP (current US$)	NY.GDP.MKTP.CD	1915899787
20	Benin	BEN	GDP (current US$)	NY.GDP.MKTP.CD	14250987026
21	Bermuda	BMU	GDP (current US$)	NY.GDP.MKTP.CD	7224329000
22	Bhutan	BTN	GDP (current US$)	NY.GDP.MKTP.CD	2446866405
23	Bolivia	BOL	GDP (current US$)	NY.GDP.MKTP.CD	40287647931
24	Bosnia and Herzegovina	BIH	GDP (current US$)	NY.GDP.MKTP.CD	20183510561
25	Botswana	BWA	GDP (current US$)	NY.GDP.MKTP.CD	18663265549
26	Brazil	BRA	GDP (current US$)	NY.GDP.MKTP.CD	1.92E+12

The second dataset, named "CO2Emission_Ldata" consists of CO_2 emissions values in metric ton, life expectancy, population, country name and their alpha codes. The dataset also has a column named "YearlyChange" containing both negative and positive values, wherein negative values indicate a reduction in carbon emission.

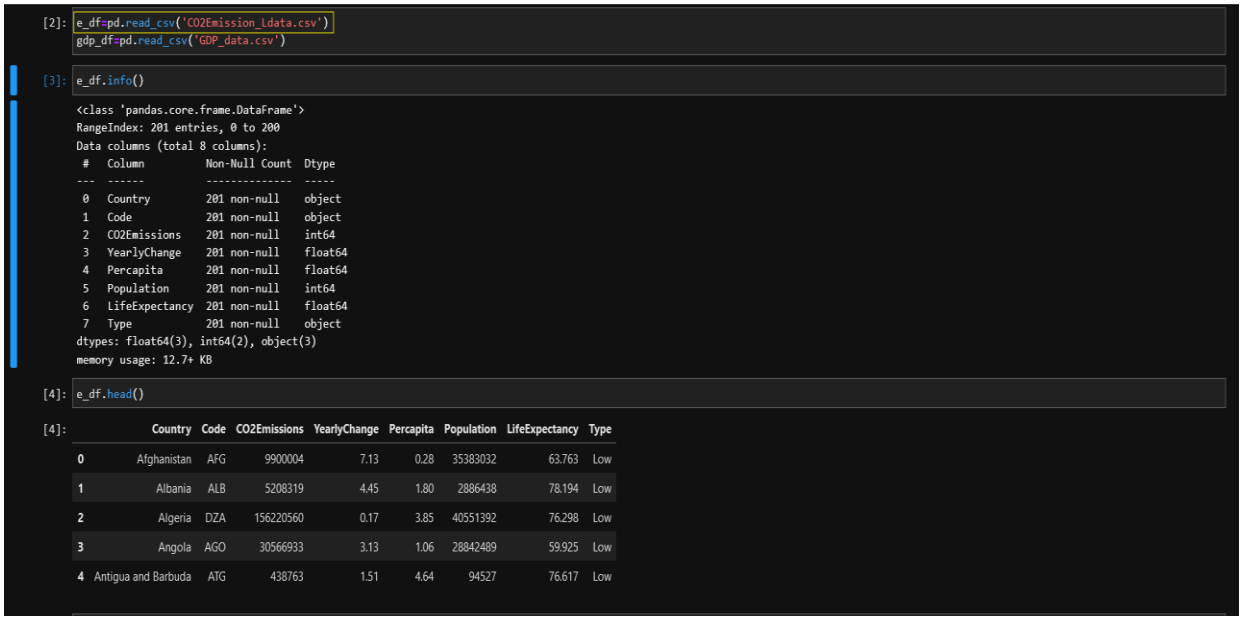
	A	B	C	D	E	F	G	H
1	Country	Code	CO2Emissions	YearlyChange	Percapita	Population	LifeExpectancy	Type
2	Afghanistan	AFG	9900004	7.13	0.28	35383032	63.763	Low
3	Albania	ALB	5208319	4.45	1.8	2886438	78.194	Low
4	Algeria	DZA	156220560	0.17	3.85	40551392	76.298	Low
5	Angola	AGO	30566933	3.13	1.06	28842489	59.925	Low
6	Antigua and Barbuda	ATG	438763	1.51	4.64	94527	76.617	Low
7	Argentina	ARG	200708270	0.16	4.61	43508460	76.221	High
8	Armenia	ARM	4597845	3.06	1.57	2936143	74.64	Low
9	Aruba	ABW	286871	1.51	2.74	104872	75.868	Low
10	Australia	AUS	414988700	-0.98	17.1	24262712	82.959	High
11	Austria	AUT	73764112	1.54	8.43	8747301	81.258	Low
12	Azerbaijan	AZE	33614235	-0.41	3.45	9736043	72.493	Low
13	Bahamas	BHS	4404247	1.51	11.65	377930	73.329	Low
14	Bahrain	BHR	24458384	2.5	17.15	1425792	76.899	Low
15	Bangladesh	BGD	74476230	4.5	0.47	157977153	71.785	Low
16	Barbados	BRB	1541447	1.88	5.39	285796	78.888	Low
17	Belarus	BLR	62655669	4.9	6.63	9445643	74.031	Low

The final data set used contains data on the extent of use of different types of fuels by each country and the respective year. The columns in this dataset are “Year”, “Country”, “Solid Fuel”, “Liquid Fuel”, “Gas Fuel”, “Gas Flaring”, “Per Capital” and “Bunker fuels” and a column holding the aggregate of the amount of usage of different fuels mentioned above, called “Total”. The dataset uses metric ton as a measurement for fuel usage.

	A	B	C	D	E	F	G	H	I
1	Year	Country	Total	Solid Fuel	Liquid Fuel	Gas Fuel	Gas Flaring	Per Capita	Bunker fuels
2	2007	Afghanistan	620	204	327	84	0	0.02	9
3	2009	Afghanistan	1846	419	1349	74	0	0.07	9
4	2008	Afghanistan	1147	294	767	81	0	0.04	9
5	2010	Afghanistan	2308	627	1601	74	0	0.08	9
6	2011	Afghanistan	3338	1174	2075	84	0	0.12	9
7	2012	Afghanistan	2933	1000	1844	84	0	0.1	9
8	2000	Afghanistan	211	1	136	61	6	0.01	4
9	2001	Afghanistan	223	19	134	57	6	0.01	4
10	2006	Afghanistan	450	44	310	90	0	0.02	9
11	2013	Afghanistan	2731	1075	1568	81	0	0.09	9
12	2002	Afghanistan	292	15	120	149	0	0.01	4
13	2005	Afghanistan	362	29	235	90	0	0.01	9
14	2003	Afghanistan	326	25	164	127	0	0.01	7
15	2004	Afghanistan	259	25	162	62	0	0.01	0
16	2014	Afghanistan	2675	1194	1393	74	0	0.08	9
17	2001	Albania	879	22	853	4	0	0.29	37
18	2002	Albania	1023	23	993	7	0	0.33	37
19	2000	Albania	824	19	776	6	0	0.27	34
20	2005	Albania	1160	20	1068	6	0	0.37	49
21	2006	Albania	1063	20	967	6	0	0.34	38
22	2004	Albania	1136	23	1026	9	0	0.36	41
23	2003	Albania	1171	23	1062	7	0	0.38	40
24	2007	Albania	1071	20	922	9	0	0.34	17
25	2008	Albania	1193	28	1036	4	0	0.38	7
26	2009	Albania	1194	106	932	5	0	0.41	7
27	2010	Albania	1254	117	953	7	0	0.43	7
28	2011	Albania	1429	146	1030	8	0	0.5	7

Each of these datasets have collected from different sources. These sources include worldbank.org (GDP_data), Kaggle and OECD.org (CO2Emission_Ldata and Fossil_data). The objective of analyzing data is to summarize its most important characteristics. This is done through an approach called EDA, which expands to Exploratory Data Analysis. EDA assists in getting a better understanding of the data and the problem statement and discovering patterns. The very first step

in EDA is to get the basic understanding of the datasets being used, like the number of rows and columns, the data types of column entries etc.

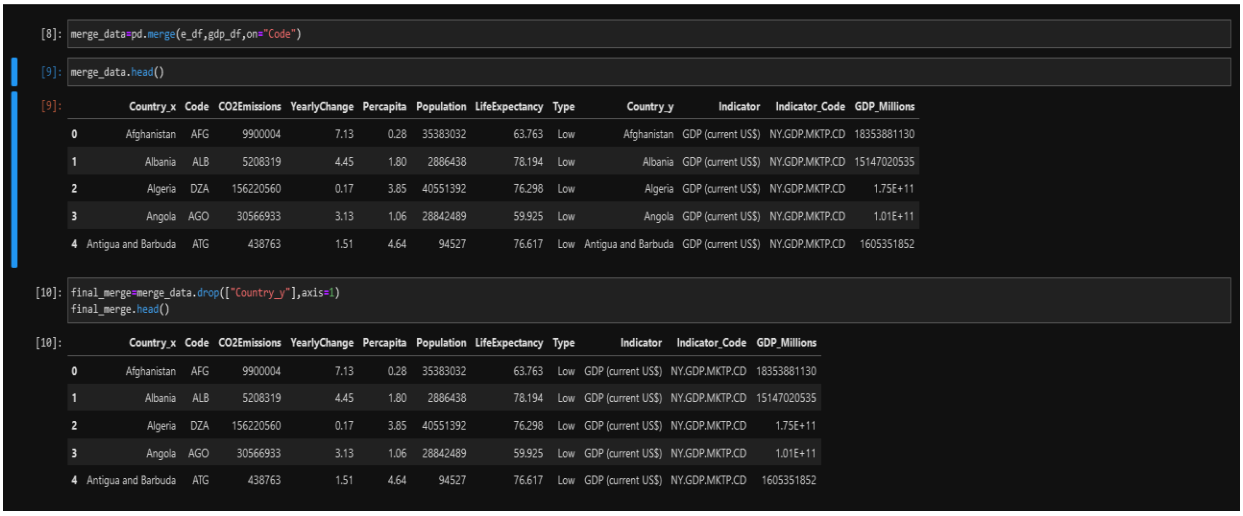


```
[2]: e_df=pd.read_csv('CO2Emission_Ldata.csv')
     gdp_df=pd.read_csv('GDP_data.csv')

[3]: e_df.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 201 entries, 0 to 200
Data columns (total 8 columns):
 #   Column          Non-Null Count  Dtype
---  ------          --------------  -----
 0   Country         201 non-null    object
 1   Code            201 non-null    object
 2   CO2Emissions    201 non-null    int64
 3   YearlyChange    201 non-null    float64
 4   Percapita       201 non-null    float64
 5   Population      201 non-null    int64
 6   LifeExpectancy  201 non-null    float64
 7   Type            201 non-null    object
dtypes: float64(3), int64(2), object(3)
memory usage: 12.7+ KB

[4]: e_df.head()
```

[4]:

	Country	Code	CO2Emissions	YearlyChange	Percapita	Population	LifeExpectancy	Type
0	Afghanistan	AFG	9900004	7.13	0.28	35383032	63.763	Low
1	Albania	ALB	5208319	4.45	1.80	2886438	78.194	Low
2	Algeria	DZA	156220560	0.17	3.85	40551392	76.298	Low
3	Angola	AGO	30566933	3.13	1.06	28842489	59.925	Low
4	Antigua and Barbuda	ATG	438763	1.51	4.64	94527	76.617	Low

To reach closer to solving a problem statement, all datasets that directly relate to the problem must be merged into a single table. This allows for easy handling and analysis of important data.



```
[8]: merge_data=pd.merge(e_df,gdp_df,on="Code")

[9]: merge_data.head()
```

[9]:

	Country_x	Code	CO2Emissions	YearlyChange	Percapita	Population	LifeExpectancy	Type	Country_y	Indicator	Indicator_Code	GDP_Millions
0	Afghanistan	AFG	9900004	7.13	0.28	35383032	63.763	Low	Afghanistan	GDP (current US$)	NY.GDP.MKTP.CD	18353881130
1	Albania	ALB	5208319	4.45	1.80	2886438	78.194	Low	Albania	GDP (current US$)	NY.GDP.MKTP.CD	15147020535
2	Algeria	DZA	156220560	0.17	3.85	40551392	76.298	Low	Algeria	GDP (current US$)	NY.GDP.MKTP.CD	1.75E+11
3	Angola	AGO	30566933	3.13	1.06	28842489	59.925	Low	Angola	GDP (current US$)	NY.GDP.MKTP.CD	1.01E+11
4	Antigua and Barbuda	ATG	438763	1.51	4.64	94527	76.617	Low	Antigua and Barbuda	GDP (current US$)	NY.GDP.MKTP.CD	1605351852

```
[10]: final_merge=merge_data.drop(["Country_y"],axis=1)
      final_merge.head()
```

[10]:

	Country_x	Code	CO2Emissions	YearlyChange	Percapita	Population	LifeExpectancy	Type	Indicator	Indicator_Code	GDP_Millions
0	Afghanistan	AFG	9900004	7.13	0.28	35383032	63.763	Low	GDP (current US$)	NY.GDP.MKTP.CD	18353881130
1	Albania	ALB	5208319	4.45	1.80	2886438	78.194	Low	GDP (current US$)	NY.GDP.MKTP.CD	15147020535
2	Algeria	DZA	156220560	0.17	3.85	40551392	76.298	Low	GDP (current US$)	NY.GDP.MKTP.CD	1.75E+11
3	Angola	AGO	30566933	3.13	1.06	28842489	59.925	Low	GDP (current US$)	NY.GDP.MKTP.CD	1.01E+11
4	Antigua and Barbuda	ATG	438763	1.51	4.64	94527	76.617	Low	GDP (current US$)	NY.GDP.MKTP.CD	1605351852

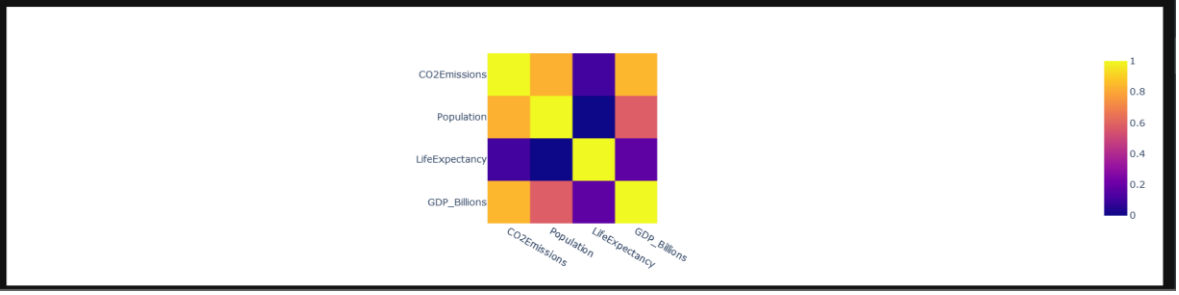
When exploiting the relations among the contributing factors, EDA can be done using a correlation matrix and a correlation heatmap. A correlation matrix is a simple table displaying the correlation values which is a measure that demonstrates a liner relationship between two variables. These correlations can be visualized using a heatmap, wherein differentiating heat (color) represents a certain value. The closer the correlation value is to 1.0 the stronger the liner relationship.

```
[23]: rv=data_final[["CO2Emissions","Population","LifeExpectancy","GDP_Billions"]]
      corr=rv.corr()
      corr
```

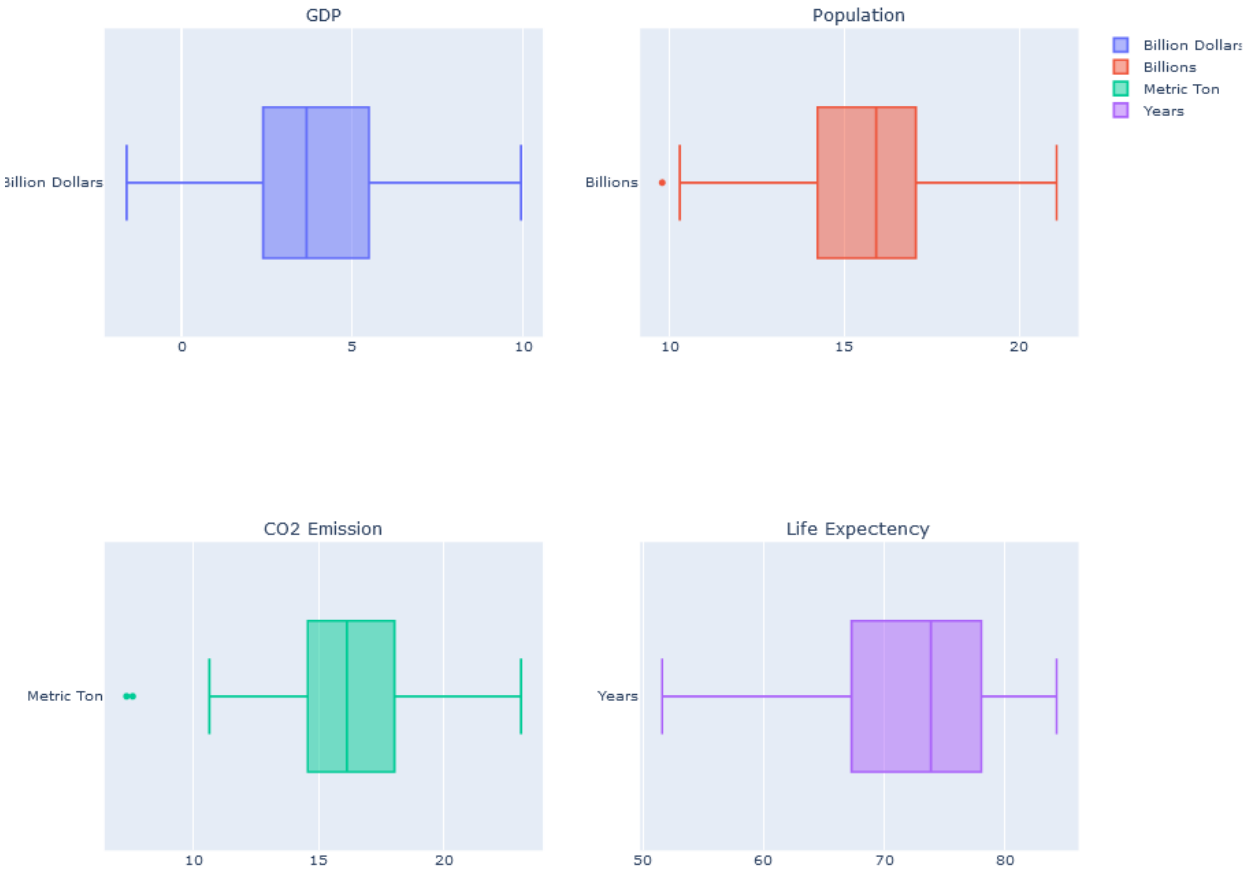
[23]:

	CO2Emissions	Population	LifeExpectancy	GDP_Billions
CO2Emissions	1.000000	0.825281	0.104171	0.840963
Population	0.825281	1.000000	-0.000216	0.579781
LifeExpectancy	0.104171	-0.000216	1.000000	0.172531
GDP_Billions	0.840963	0.579781	0.172531	1.000000

```
[24]: fig = px.imshow(rv.corr())
      fig.show()
```



The next step in the EDA process is to identify the outliers and decide upon whether to treat them. The most efficient way to do so is by plotting box plots, which consists of an inter quartile range with 25th and 75th (Q1 and Q2) percentile of the data as its bounds and which is turned within the bounds of the maximum and minimum of the given data. Any observations lying outside of these bounds is an outlier.



The datasets used, contains values with very high range (GDP and CO_2 emission), that can drastically differentiate between records due to their nature. This can cause the data to skew, which can distort its understanding. This problem is tackled by converting such values to their logarithmic form. As seen in the above boxplot, there exist certain outliers, whose effect on the data is negligible and hence can be disregarded.

Visualizing a relationship is highly pragmatic to understand the effect one attribute is having on another. The same has been done to prove a hypothesis that increase in GDP contributes to an increase in CO_2 emission. Nations having a high GDP indicated that the nation is industrialized, has better exports and has been invested in for developments all of which contribute towards carbon emission.

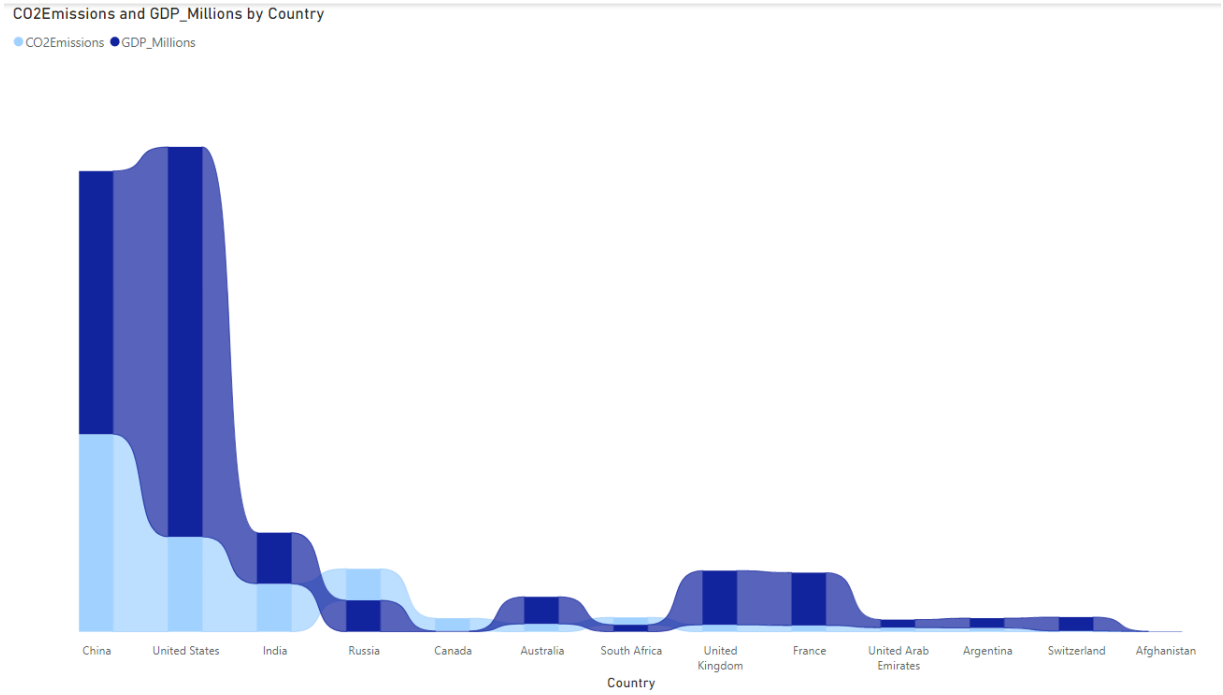
The visualization below proves the above hypothesis true. The visualization is a scatter plot depicting a simple yet significant relation between GDP and carbon emission. Though the plot it can be understood that there is an upward trend in carbon emission with increasing GDP.

Result

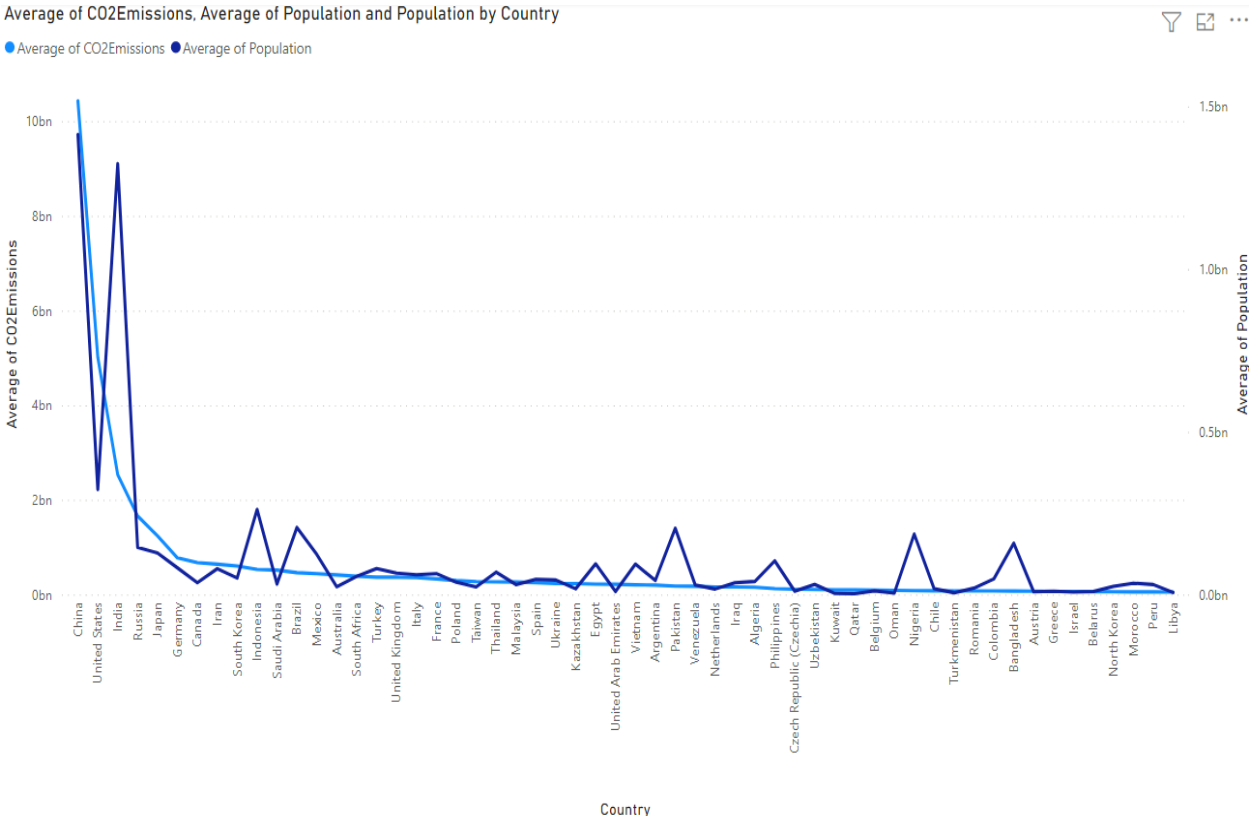
Through data visualization data can be transformed into a visual representation. This project emphasizes analyzing factors affecting carbon emissions and, in the process, we test the hypothesis of whether GDP in any way is affecting CO_2 emission. The visualization produced as a result of the analysis perfectly answers this hypothesis. PowerBI by Microsoft was used to create

visualization to understand the relations between the factors. The set of visualizations is done for the data that is merged using the “GDP_data” and “CO2Emissions_Ldata”.

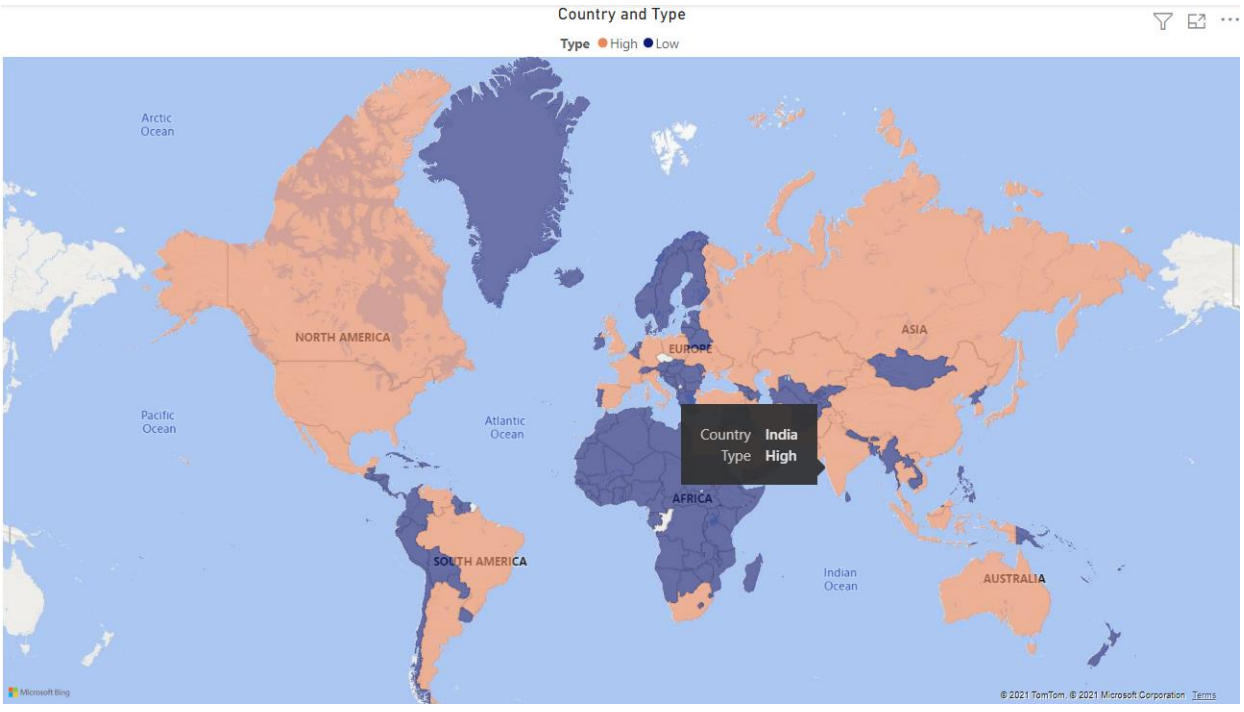
The first visualization perfectly answers the hypothesis and proves it correct. Nations with higher GDP do have high carbon emissions. High GDP indicates that a particular nation is industrialized, has a higher work force, better transportation, high investments for developments etc. All of which contribute towards high carbon emissions. A ribbon chart is used because it is lot similar to a stacked-column chart, but it ranks and displays change in values for different categories.



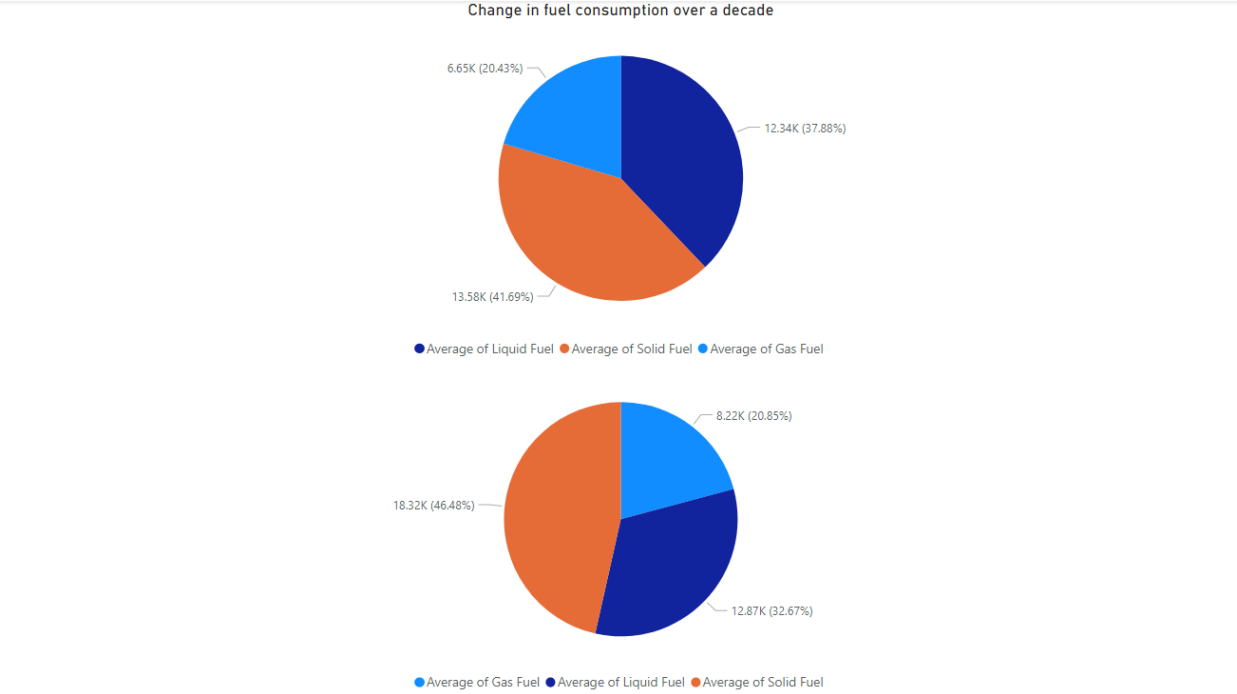
The second visualization is used to test the hypothesis of whether population effects carbon emission. A multi-line chart is used where in population is plot against carbon emission for each country. The reason for using a multi-line chart is its ability to capture trends and changes. Even the smallest of changes can be represented.



The third visualization, though very simple, encompasses very useful information. It describes and marks all those countries that lie on the extreme ends of the co2 emission spectrum. This is done so by adding a new column to the final dataset called “Type” where in each country is labelled high or low based the amount of emission. If the CO2 emission for a country id greater than or equal to the average of world’s carbon emission it is labeled as high, else low.

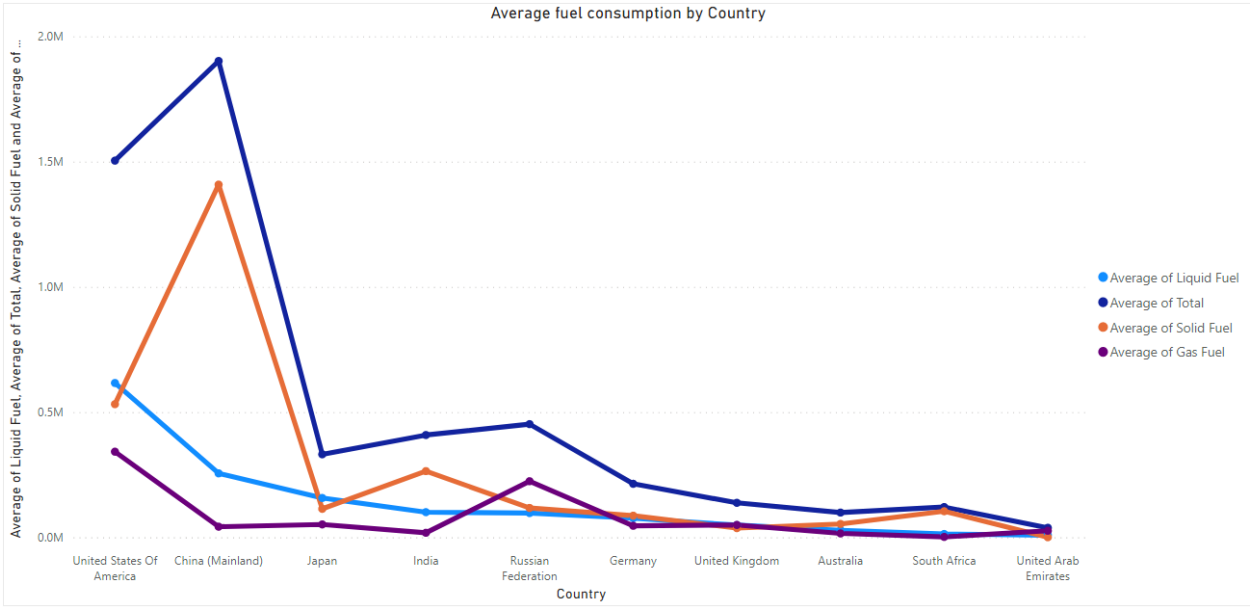


The fourth visual helps understand the transition or change in the amounts of major types of fuels used over a decade. The visualization consists of two pie charts, which are perfect to display proportions of multiple classes of data.



The change seen is drastic especially for the solid fuels which are among the heaviest contributor to carbon emission (Coal, firewood, peat, charcoal etc.), who's consumption has jumped by 5%, where has liquid and gaseous fuels have seen a 0.42% increase and a 5.2% decrease respectively.

The final visualization is a simple line chart displaying the average consumption of the above-mentioned fuel types by each country.



Microsoft PowerBI is the tool of choice for visualizing this analysis. PowerBI makes integrating spreadsheets and CSV files easy, wherein data can be directly queried and manipulated. PowerBI is highly flexible and offers a rich variety of visualizations. Apart from the advantages, PowerBI has shortcomings like inability to handle large datasets well and a crowded interface which make visualizations hard to configure. PowerBI is mostly used for business data analysis and decision making.

Discussion

Visualization improves the data quality, reduces data overload, and reduces misinterpretation. Visualization is an interactive representation of data, which amplifies cognition and visual aids help in doing so. Visualization is heavily dependent on properties of human perception and hence plays a very important role in designing visuals. But how does human perception work? Human perception works through shape and space perception. Shape perceptions include proximity, similarity, closing, continuity, and figure-background. Space perceptions include vertical predominance, parts, surface, volume, and depth. The visualization used in this project like the multi-line chart for population vs carbon emission and the average fuel consumption by major countries is categorized into shape perception, while visuals like pie chart to represent change in fuel consumption, ribbon chart to plot GDP against carbon emission and the map to categorize countries based on the amount of emission are labelled as space perception. Human perception systems understand visuals through pattern perception, sequential processing, and parallel processing.

In the first plot (Ribbon plot: CO2Emission and GDP_Millions by Country), a view or a user can easily associate GDP with carbon emission and can understand that with increase in GDP the carbon emission increases and how emission rate fluctuates in correspondence to the changes in GDP.

In second plot (multi-line chart: population vs carbon emission), the user can easily make-out the sharp increase in the carbon emission between Russia and China and understand that this is due a sharp increase in the population.

The third chart (Filled map: categorized countries based on the amount of emission) which is developed by comparing a particular country's emission rate to the world's average emission rate. This view easily make sense of the data as the two types of categories are represented using different color and these colors are used to fill each country based on the type they belong to.

The last two visuals (Pie charts: Change in fuel consumption and multi-line plot: Average fuel consumption by country) that are based on a separate dataset called "Fossil" , have a little scope for representing information that can explain the effects on carbon emission But these visuals, when compared to or are associated with the above mentioned plots help viewer get insight into the information that countries with high GDP and populations consume more fuels and their consumption of these fuels have increased over a decade, all of which contribute to high CO_2 emissions

Conclusion and Future Work

Though this project was successful in establishing the effect of Gross Domestic Product on CO_2 emission through correlation and basic visualizations, it is not sophisticated enough. This analysis can be further improved by gathering additional information like national government expenditure on solutions towards reducing carbon emission, or real time data on global temperature change with respect to carbon emission etc. Such insights can help us better understand and strongly correlate carbon emission with its factors. Another improvement would be to use Geospatial location analysis or Geographical Information System (GIS) data visualization. Using GIS, data can be directly connected to the map by integrating location data with descriptive information. GIS can use raster and grid data to represent concentration of CO_2 emission in a smaller area, in detail or in a larger area with less detail.