**Report on the Relational analysis between the Carbon dioxide emission with GDP per capita and Power consumption of the country.**

**Description of the Problem:**

According to IEA statistics in 2017, we still get about 80% of our energy demand by burning fossil fuels. Fossil fuel is carbon-rich resources so burning it causes emission of carbon dioxide and other carbon substituents. Carbon emissions trap heat in the atmosphere and drive climate change as well as oceans are becoming more acidic than previously by absorbing as much as a quarter of all man-made carbon emissions.

Gross domestic products (GDP) is a monetary measure of the market value of all the final goods and services produced in a specific period, often annually. GDP per capita at purchasing power parity (PPP) is arguably more useful when comparing differences in living standards between nations.

We need to analyse the relationship between GDP/capita (PPP) with Carbon emissions. So could it be possible to determine whether both are relational or not and how are they relate to each other?

Studying these features could tell us whether a correlation between carbon emission and GDP per capita is positive or negative. Our targeted audience could be UG and PG student who would have been studying Social Science and it might help the environmental NGO's who are working across this domain. I hope they could find key findings from this analysis report.

**Description of the data:**

So to find a relationship between Carbon emissions by all countries and GDP/capita we need to the analysis report on carbon emissions of all countries and the second one is GDP per capita according to purchasing power parity. Since lots of institutions, organizations and countries havebeen studying and analysing these two-parameter. So I think we could not have any problem to collect the relevant data.

I have searched to get the data and came across the World Bank website. This website has Metadata where I find the data about GDP per capita according to purchasing power parity where I get data in excel format of all countries GDP per capita from 1990 to 2018. Similarly, I get carbon dioxide emissions data on the same website.

So the features to analyse the relation could be annual carbon emission in MtCO2e and GDP per capita according to purchasing power parity. Also could be more features required to study the problem thoroughly.

**Methodology:**

**Data Understanding:**

**Brief description of the data:**

As I search for the required data of Carbon emission and GDP per capita of all countries. I find that I cannot get this data in a single tabular format. I need to get both tabular from the World Bank data. Merging of this data requires to bring it in a single table. I get carbon emission per MtCO2e\* from World Bank metadata site from 1960 to 2017 for all countries similar to that GDP per capita data from 1990 to 2017.

**1. Data Processing and Preparation:**

As I proceed to analyze the data which could be required to do relational analysis. I notice that there are some missing values for a certain year in both carbon emission and GDP per capita data. As we know that while doing any statistical analysis this missing value (row value/observation) of certain features (column value) could be a problem in the certain machine learning algorithm.

So before doing the analysis, I need to either fill this missing value using method which is provided by pandas library such as bfill, ffill and imputation by mean and other methods or drop certain observation which has missing values.

Since each method has pros and cons and also filling of missing values depend on the problem statement and implementation prepared algorithmic model. I decide to choose later procedure because of the vulnerability of my model. Since my prime audience would be NGO's and PG/UG student I think that drop such missing observation could not much hamper model.

Also I there is a mismatch in feature column (i.e. yearly data); have to drop carbon emission data column from 1960 to 1989. As this data is separate I have to merge it to bring it in single tabular format using pandas merge function and to optimise the merged data I use inner join method to merge them on the name of the country. After merging of the data I get 157 observation for emission and GDP per capita from 1990 to 2017.

**Exploratory Data Analysis (EDA):** As we know that before doing the analysis we need to check how data is spread over the range, how a feature of the dataframe is related to each other. So to do EDA I use the seaborn library. To visualise the analysis seaborn library has very handful method and function without much codings such as pairplot, heatmap, histogram and many more. From which I use the heatmap to know the correlation between the feature; also use pairplot to visualise the data and boxplot to know about outliers.

While doing correlation between the feature I get that CO2 emission and GDP per capita are only 6-7 percent relational so from this I assert that there may be a large number of outliers present in this data so to know how the data is skewed whether it is positively skewed or negatively I plot boxplot and also calculate skewness using skew function of pandas library. From this, I get that data do not follow thumb rule for skewness from this I get that outliers in the data have a greater effect. So to minimise the effect of outliers on it and to follow the thumb rule of skewness I take log transform of the data using numpy library function "log10" means log with base 10. After transforming it and calculating again skewness of the feature we get that GDP per capita is negatively skewed while CO2 emission is positively skewed.

But outliers effect were minimised as you see boxplot for each feature.

**Inferential Statistical Testing**: Before the use of the data to model the data using any algorithm we need to test our hypothesis which is called as hypothesis testing. There are different hypothesis testing for a different category such as categorical observation, continuous observation, etc.

Since my model has a continuous observation I use Pearson correlation coefficient to know about features; I get that before transforming the data there is nearly no relation between the CO2 emission and GDP per capita. To cope up with hypothesis transformation give us a "moderate correlation" between this feature with around 43%.

After doing the above procedure I think data is prepared to use for modelling the model out of it.

**Modelling:**

After completing the data preparation procedure we use this prepared data to build a model from it. I choose to do simple regression analysis on the 2014 data from the dataframe with a feature as CO2\_emission, GDP\_per\_capita and Power\_consumption. But this regression model has negative R^2 value that means our model could not relate this feature efficiently. From this model, I analyse that we need to pre-process further the data to optimise the model metrics. To do that we need to scale the data that relevant to each other. For that, I use a scikit-learn pre-processing method and use different scaling parameter such as standard scalar, min-max scaling, normalization, etc. After preprocessing and again building the regression model still we could not optimise it efficiently. So I decide to change the model and use ordinary least square (OLS) model to build the model and this time we get a better model with R^2 score equal to 0.83 - 0.84. That means 83 – 84% our model will predict the result with unknown input. Also, I model an SVR model with this data using cross-validation and further optimise the model with better quality than previous models.

**Result:**

I assume that there may be a relationship between the CO2 emission by the country and the respective GDP per capita. So Inferential Statistical testing as Pearson Correlation tells that there is a positive 42% relation between these two variable. Along with this OLS model also tells a similar story with 84% accuracy. And also Support Vector Regression (SVR) tells that our model is efficiently predicting the value for input and model is the best fit with an RMSE value of .11.

**Discussion:**

After building the model and presenting it with one the NGO's who is working in the environmental domain; we discussed the viability of the model we come to know that this model there could be a chance to refine with more relevant data pre-processing method and fine parameter tuning.

But still, it will be implementable with this result. And along with this refinement of a model can be done using feedback from the user who will be going to use it.

**Conclusion:**

After discussing with a potential user of this model I analyse that I need to better pre-processed the data and use another method to fill the missing value of the data. Also, I could further do the hyperparameter tuning of the model to get better with the model.

**Glossary:**

GDP - gross domestic product

SVR - support vector regression

MtCO2e - Metric Ton of Carbon dioxide equivalence

RMSE - root mean square error

IEA - International Energy Agency