COVID-19 & its IMPACTS

The Dataset for this project was collected from Kaggle and originates from Mendeley Data:The impact of Covid-19 Pandemic on the Global Economy Emphasis on Poverty Alleviation and Economic Growth. The data that has been investigated here consists of records on the impact of covid-19 on the global economy including 210 countries.

Main objective of the analysis is to focus on prediction. In this Project, Linear Regression algorithm shall be applied to find the relationship between GDP and human development index and total number of death. We will then choose the best candidate algorithm from preliminary results. The goal with this implementation is to construct a model that accurately predicts how the global economy of each country is affected.

Attributes of the data set:

Column	Non-Null values	Dtype
is0_code`	50418	object
location	50418	object
date	50418	object
total_cases	47324	float64
total_deaths	39228	float64
stringency_index	43292	float64
population	50418	int64
gdp_per_capita	44706	float64
hdi	44216	float64
Unnamed: 9	50418	object
unnamed : 10	50418	object
Unnamed: 11	50418	object
Unnamed: 12	50418	float64
Unnamed: 13	50418	object

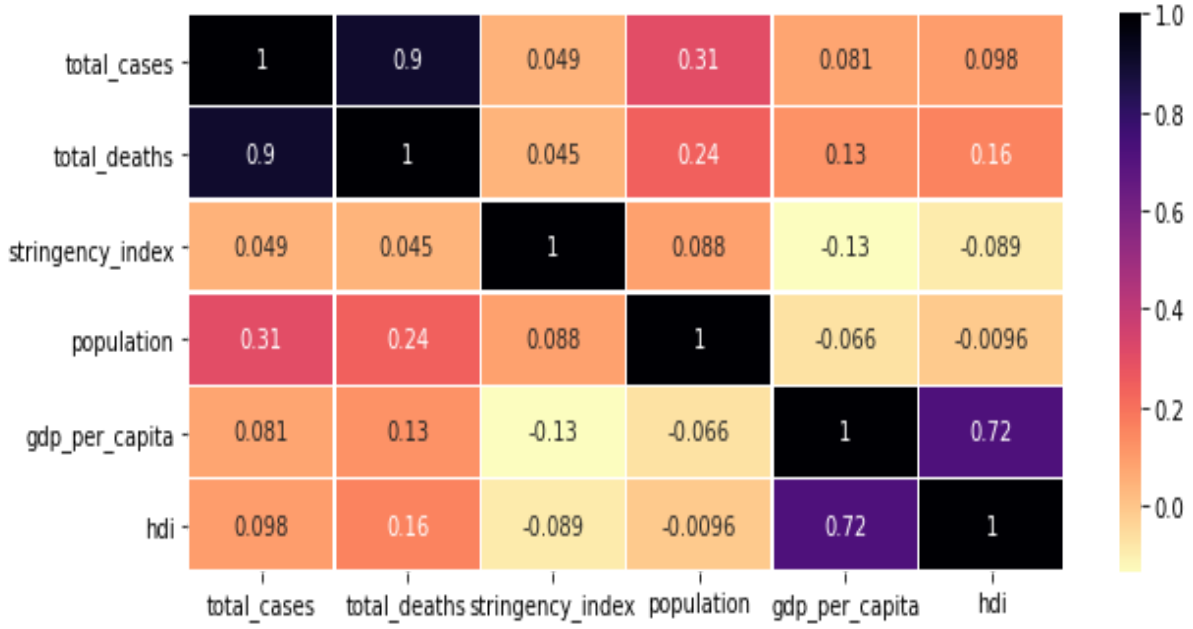
- Iso_code: country code
- Location: name of the country
- Date:
- Total_cases: number of cases of covid-19
- Total_deaths:
- Stringrncy_index: provides a computable parameter to evaluate the effectiveness of the nationwide lockdown. It is used by the Oxford COVID-19 Government Response

Tracker with a database of 17 indicators of government response such as school and workplace closings, public events, public transport, stay-at-home policies. The stringency index is a number from 0 to 100 that reflects these indicators. A higher score indicates a higher level of stringency.

- Population:
- Gdp_per_capita: Calculated by taking into account the monetary value of a nation's goods and services after a certain period of time, usually one year. It's a measure of economic activity.
- Hdi: It is a summary measure of average achievement in key dimensions of human development: a long and healthy life, being knowledgeable and having a decent standard of living. The HDI is the geometric mean of normalized indices for each of the three dimensions.

Data Exploration:

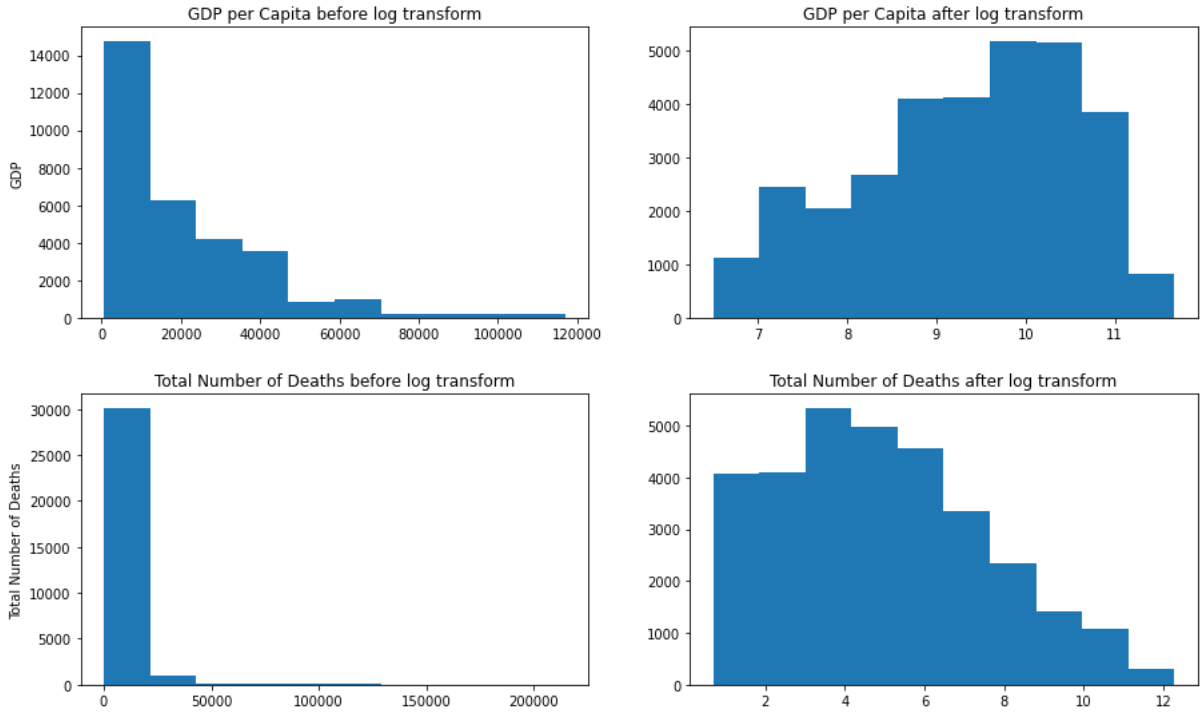
- From the table can be seen that there are irrelevant columns, which shall be removed hence like the iso, unnamed etc.
- Rest of the columns had no missing values.
- Plotted the correlation matrix to see the relation between two variables: A coefficient close to 1 means that there is a strong correlation between two variables. The diagonal line is the correlation of the variable with itself, that's why they are 1.



- Its cleary seen that HDi is strongly correlated to GDP and total deaths are correlated to number of cases.Population has also a strong relation with number of cases and deaths.
- From the heatmap, both GDP and HDI are more affected by the number of deaths than number of cases.

Feature Engineering:

- Transforming the skewed variables using log1p.

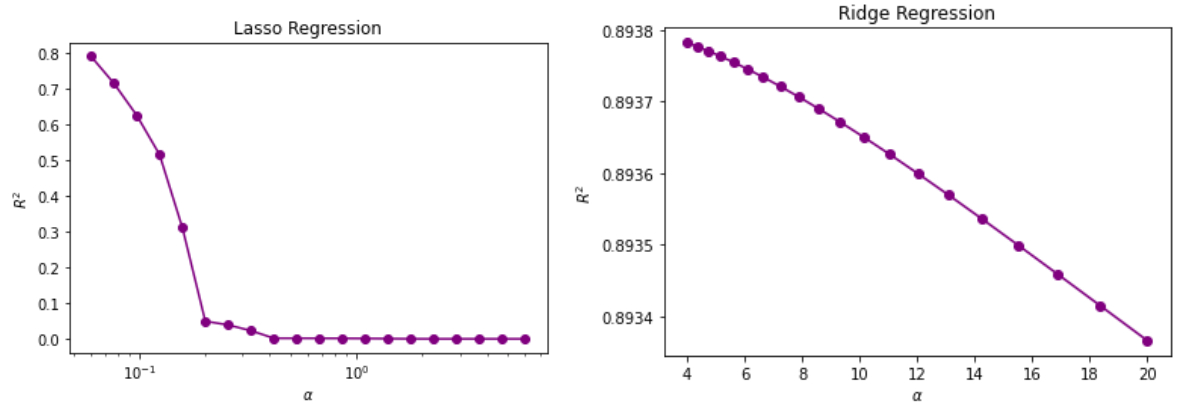


- Scalar transformation to normalizre the data. This ensures that each feature is treated equally when applying supervised learning models.
- After one-hot encoding the location column was found to be overfitting the model, hence decided to drop the location column and also its not necessary for the learning process of algorithm.

Model Training and Testing:

- Both Lasso and Ridge with proper hyperparameter tuning gave better results than plain Linear Regression.

	linear	lasso	ridge
score	0.79974	0.834581	0.89403



	Linear	Lasso	Ridge	ElasticNet
rmse	1.397396	1.396604	1.396936	1.396826

Conclusion:

Lasso gives the smallest Root-mean-squared error, however the difference in scores and errors are not significant, are actually almost identical(seen in the above table). The best candidate bseed on rmse and score results would be Lasso Regression, therefore it is recommended to LassoCV as a final model that best fits the data in terms of accuracy.

Going Forth:

Can further try to optimize lasso using GridSearchCV. To predict the effect on GDP for an individual country, we could use one-hot encoding of the countries and use that for training our models. Perhaps collecting more frequent records on specific countries would help achieve more accurate results.