Effects of environmental parameters and their interactions on the spreading of SARS-CoV-2

Introduction : -

In 2019 The world suffered the SARS-CoV-2-related pandemic with a high number of deaths and hospitalization. The effect of atmospheric parameters on the amount of hospital admissions (temperature, solar radiation, particulate matter,NO2, relative humidity and wind speed) is studied through about 2 years .

Dataset : -

I started to create our data depending on The environment parameters which increase transmission of disease through 2 years. The Dataset consists of 100,000 records in different Countries .

Columns : -

Date : The Date of the study which has been recorded starting from 01-11-2019 to 31-12-2020 , using Faker Library to get real date between this range .

Country : The Country of the study where has been recorded using Faker Library to get real Countries all over the world .

Temperature: The Temperature which has been recorded starting from -10 to 40 using Random Library .

Humidity: The Humidity which has been recorded starting from 20 to 90 using Random Library .

NO2: The NO2 which has been recorded starting from 10 to 25 using Random Library .

PM(particulate_matters): The PM which has been recorded starting from 5 to 55 using Random Library .

Solar Radiation: The SR which has been recorded starting from 100 to 280 using Random Library .

Wind Speed: The Wind Speed which has been recorded starting from 0 to 15 using Random Library .

Impact Score : Describe the impact of disease transmission through equation .

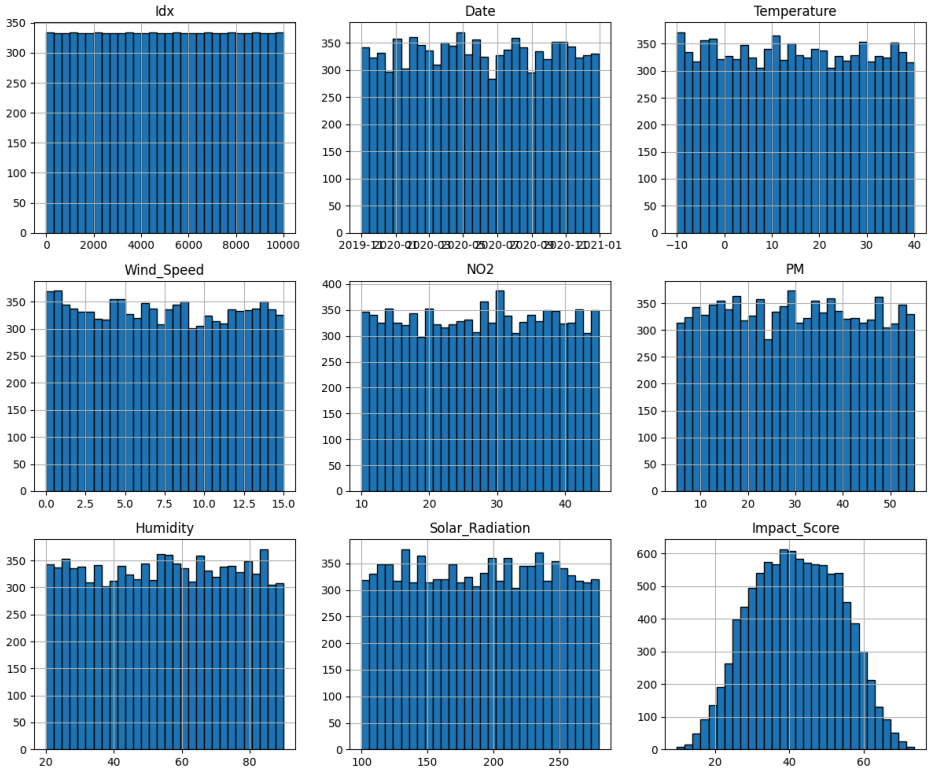
EDA: -

- **Investigate Dataset and display its properties , display maximum , minimum , Avg , Counting ,Mean and std for each column so we could have better statistics .**

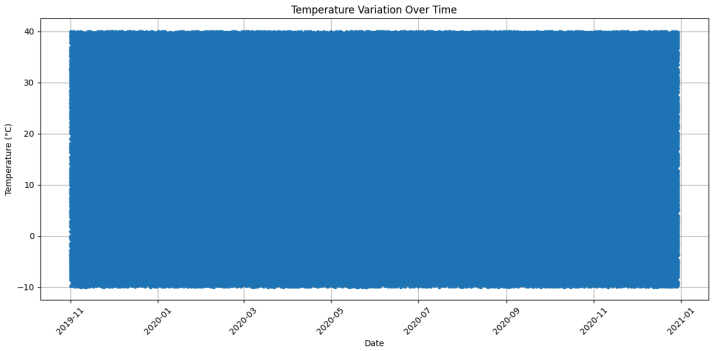
- **Display info of dataset so we could know the type of each Column and size of our dataset .**

- **Check if there is any missing value in each column .**

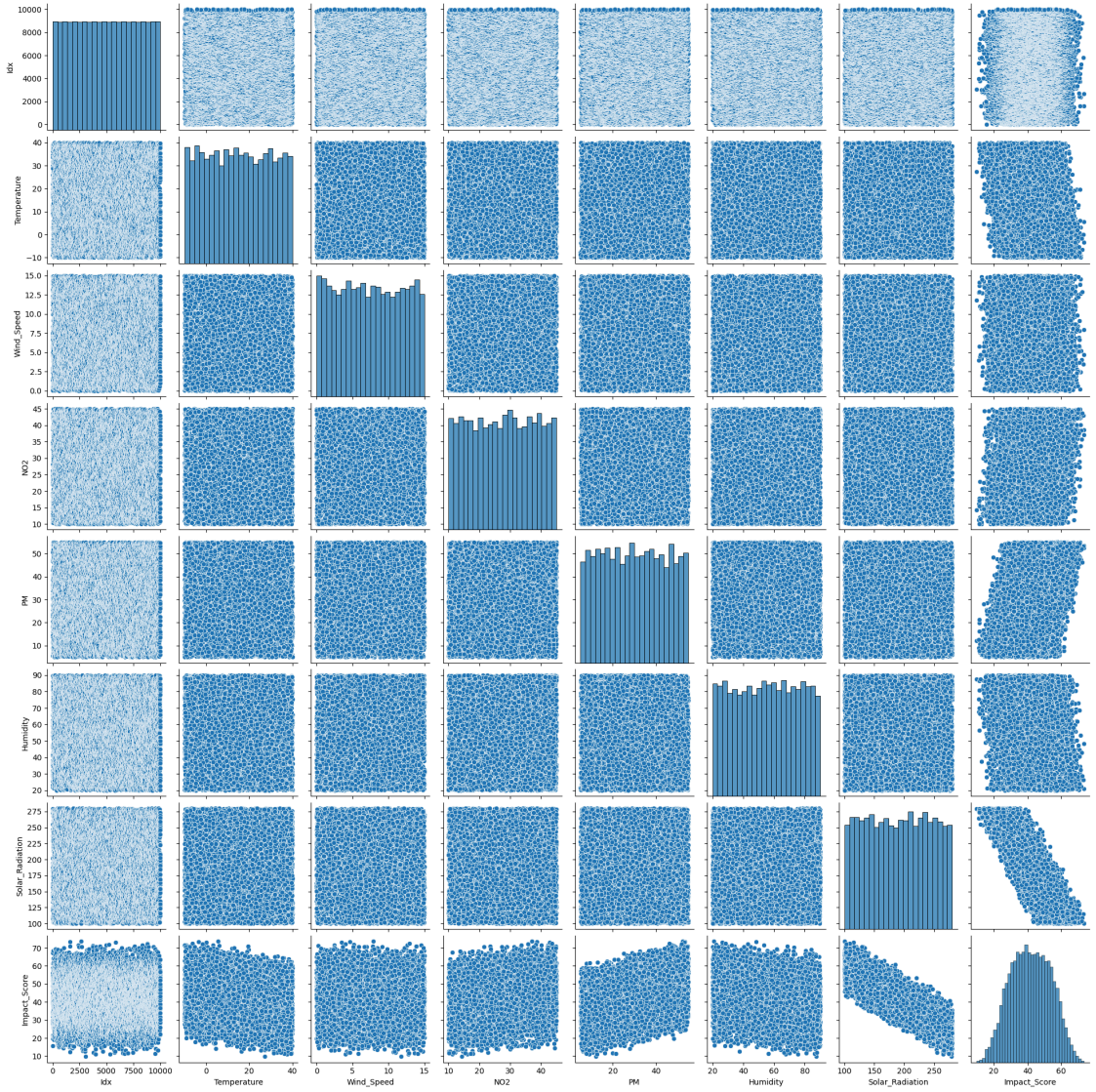
- **Visualize each numerical value in the dataset so we could know the frequency of each value in our dataset as shown in fig 1 .**



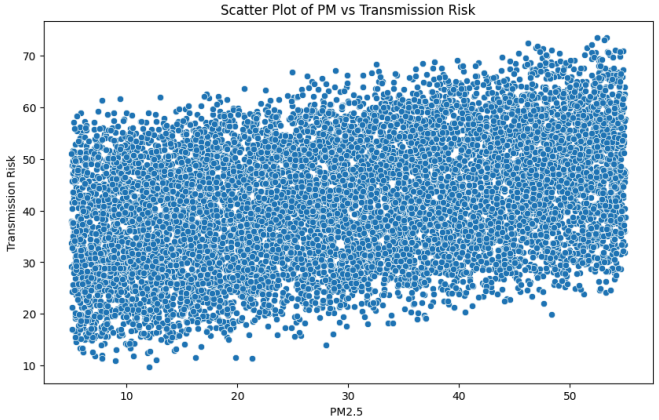
- **Visualize any parameter which affects on transmission(Ex Temperature) through the date of the whole study as shown in fig 2.**



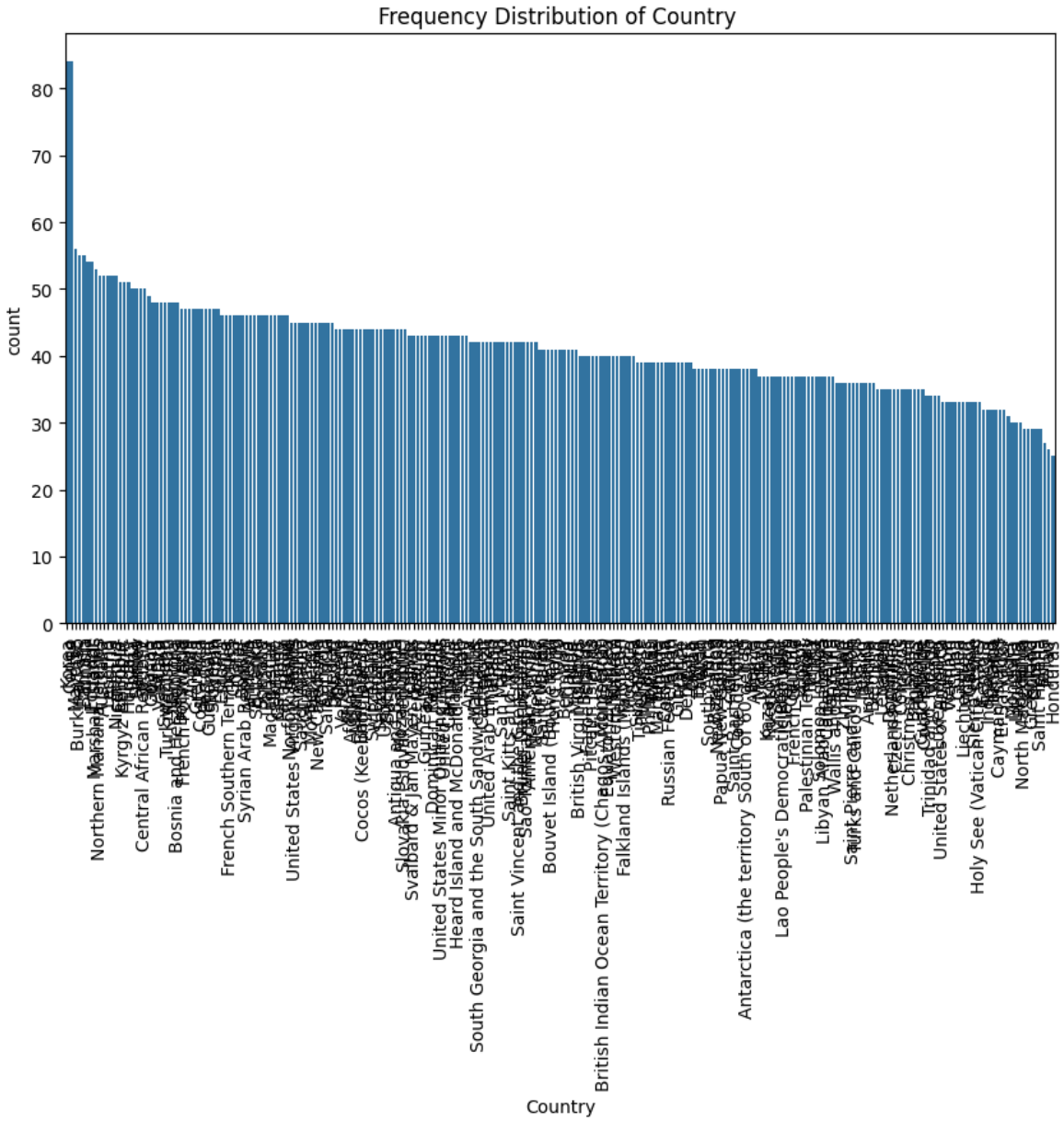
- **Visualize each numerical value in the dataset related to others value so we could know the relationship between them as shown in fig 3**



- **Visualize Effect of Pm value on transmission which shows with higher of PM the transmission risk increase as shown in fig 4.**



- **Visualize the frequency distribution of countries so we could know which country has more records as shown in fig 5.**



Processing : -

1 - There is no missing value in our dataset but if there is any we could use mean value (Imputer) or drop this row or … etc .

2 - Split our dataset to input and output then drop unimportant (not effective) columns on our target .

3- Prepare the dataset by splitting input & output to train and test (80% train - 20 % test) .

4- Scaling / Normalization our input because they have different scaling (measurement) and data is ready now for training .

Model : -

- It's clear it's a regression problem which predict the transmission risk of SARS Covid 19 .

- The model as shown from equation is multiLinear regression as it worked very well as I train it Although I tried different Algorithms .

- I tried Ridge Regression ,Lasso Regression , Random Forest , Gradient Boosting and SVR .

- I made the comparison between those algorithms on different aspects (Time - Mean square error - R2 score) .

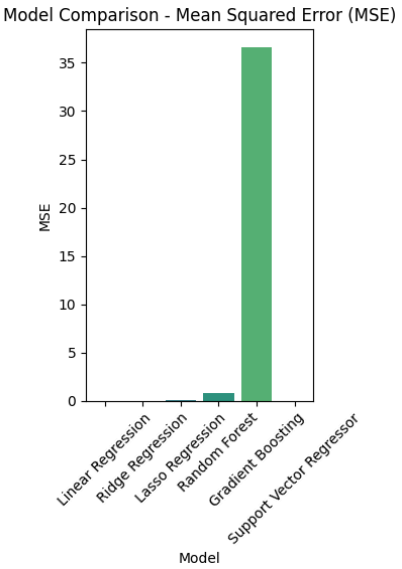
- As we see and expected Random forest has longest period and gradient Boosting has smallest R2 score between them As Shown fig 5

```
[ ] #Head of result
    results.head()
```

	MSE	R^2	Time
Linear Regression	1.165927e-28	1.000000	0.047227
Ridge Regression	2.257847e-08	1.000000	0.022910
Lasso Regression	6.003754e-02	0.999586	0.025567
Random Forest	8.086326e-01	0.994418	11.091481
Gradient Boosting	3.663204e+01	0.747136	2.659681

- All algorithms are doing fine but the most accurate score is Multi Linear regression and Ridge regression .

- Visualize the difference between Algorithms depends on MSE as shown on fig 6.



Evaluation : -

- We evaluate our model using R2 Score and MSE
- MSE : the lower it is the better performance
- R2 Score : more close to 1 make the model better

The model got 1 of R2 Score on training dataset (Which might be a Overfitting problem) but after testing it on testing dataset we still got 1 of R2 score which good indicate that there is no overfitting and 1.164*e**-18 of MSE on training dataset and 1.166*e**-10 of MSE on testing dataset which shows that the model is doing well.

Explanation for how model is strictly perfect , The data isn't real data and looks very similar to each other .

Pipeline : -

Pipeline is used to streamline the workflow of data processing, model training, and evaluation. Pipelines help ensure that the entire process is reproducible, manageable, and less error-prone .

I created simple pipeline holding the preprocessing (Standard scaler) and the regression model and test it .

Improvement : -

- We could start to focus on some countries like in Europe so we could have a good range of our parameters .
- Record more data till 2024 so we could find more accurate environmental parameters which might affect transmission .
- As the data gets increased we are gonna need more complex models which would lead us to use Dl models .
- Finding more complex relationship between the transmission impact and The environment parameter rather than the simple linear equation which would might help and be more accurate .

References : -

Effects of environmental parameters and their interactions on the spreading of SARS-CoV-2 in North Italy under different social restrictions.