

THE IMPACT OF WEATHER & SOCIO-ECO FACTORS ON THE SPREAD OF COVID-19

Higher Diploma in Science in Data Analytics

Muhammad Shakir Aleem

Student Id: 10534253

[10534253@mydbs.ie](mailto:10534253@mydbs.ie)

Final Report

Supervisor Name: Obinna Izima

Submission date: 24/08/2021

Abstract:

On the month of December, 2019 first Covid-19 case reported in China. From the city of Wuhan, it started spreading across the globe. In January 2020, WHO declared Covid-19 as pandemic. The virus spread through social interaction and through air as well. So far, those who has weak health, live in huge population, don’t have good medical facilities etc. are the main reasons of its spread.

This project assess whether weather condition do also have any influence and link with fast spread of Covid-19. The second part of the project is also to find and trend between social-economic factors and number of covid-19 cases.

Acknowledgement:

I would like to thanks my final project supervisor and class fellow in helping in selection of right dataset for my idea of project. My supervisor guided me in initial stage what quality of dataset should I have. My class fellows also helped my coming up with this particular project idea.

Table of Contents

[1. Introduction: 5](#_Toc80705267)

[1.1. Purpose of project: 5](#_Toc80705268)

[2. Background: 5](#_Toc80705269)

[3. Requirements Specification and Design 6](#_Toc80705270)

[3.1. Project/Business requirements: 6](#_Toc80705271)

[3.2. Outcome of project: 6](#_Toc80705272)

[3.3. Information requirements: 6](#_Toc80705273)

[3.4. Specification and Design: 6](#_Toc80705274)

[3.5. Predictive models selection: 9](#_Toc80705275)

[3.5.1. Reason of selecting particular prediction models: 9](#_Toc80705276)

[4. Implementation: 10](#_Toc80705277)

[4.1. Dataset: 10](#_Toc80705278)

[4.2. Importing and inspecting the data: 10](#_Toc80705279)

[4.3. Data preparation: 12](#_Toc80705280)

[4.3.1. Conversation & filtering: 12](#_Toc80705281)

[4.3.2. Data Visualization: 12](#_Toc80705282)

[4.3.3. Removing outliers 15](#_Toc80705283)

[4.3.4. Scaling & standardization: 20](#_Toc80705284)

[4.4. Data Modelling: 20](#_Toc80705285)

[4.4.1. Regression Modelling 20](#_Toc80705286)

[4.4.2. Classification modelling: 21](#_Toc80705287)

[4.5. Social-Economic factors vs Covid-19 spread: 23](#_Toc80705288)

[4.5.1. Data preparation: 23](#_Toc80705289)

[5. Testing & Results: 30](#_Toc80705290)

[5.1. Multi-liner regression Result 30](#_Toc80705291)

[5.2. Liner regression Result 34](#_Toc80705292)

[Simple linear regression evaluation: 34](#_Toc80705293)

[5.3. LASSO result 34](#_Toc80705294)

[LASSO evaluation result 35](#_Toc80705295)

[5.4. Overall Regression result: 36](#_Toc80705296)

[5.5. Logistic regression Result 36](#_Toc80705297)

[5.6. Random forest Result 36](#_Toc80705298)

[6. Conclusion and Future Work: 38](#_Toc80705299)

[6.1. Limitation 38](#_Toc80705300)

[6.2. Future work 38](#_Toc80705301)

[Appendix: 38](#_Toc80705302)

[Bibliography 39](#_Toc80705303)

# Introduction:

Many analytic studies have been conducted on COVID-19 and its pattern and factors of spread since pandemic was started. In January 2020 WHO declared COVID-19 as pandemic. After that most countries enforced different measures to prevent spread of virus in their country based on demographics, social, activity, and travel data found in the reviewed studies (Alsunaidi, et al., 2021). Several sentiment analyses also have been done using big data. Mostly, case studies based on geographical location, travel pattern of countries population, precautions taken by nation and government and medical facilities available on individual countries etc. (Arman Behnam, 2021). According to many case studies when pandemic was started, most common factors impacted in spread of virus like climatology parameters, tobacco smoking, and air pollution.

## Purpose of project:

Some researchers have been initiated on how virus is transfer through different medium in environment (Hadi Eslami, 2020). Big data and ML models have been built based on several data which were collected from different source like Hospital, social media, countries health care system etc. (Alsunaidi, et al., 2021). The researches should be done to assess whether environmental factors and social-economic factors of different location across globe could influence spread of virus among people, therefore, governments and organizations could necessary action to prevent Covid-19 infection and dead from it.

Relationship between Covid-19 cases/death and environmental features as well as socio-eco features will be shown by visualization. Data modelling will also be conducted and results of modelling will be evaluated in this report.

# Background:

Since Covid-19 pandemic started there have been many factors identified and studies have been conducted based on data availability. If you see the trend of number of cases across the globe and observe different countries’ Covid-19 statistic then you might notice that usually countries which have cold and harsh weather, despite having better medical facility their percentage of covid-19 cases with respect to their populations are higher than the countries which are located at or near equator.

Usually, most part of the world low cases of Covid-19 reported during summer as compare to winter. Countries who have averagely high temperature or their weather is not too harsh like Australia, UAE, Saudi Arabia, Thailand, Pakistan, Malaysia etc. they reported less covid-19 cases as compare to Europe and north America (Control, 2021). We do see some outlier in hotter region in globe that countries might have higher number of cases despite they have warm weather.

To analyze whether temperature and other weather factors could impact covid-19 spread or death study should be conducted using weather data of each country and its covid-19 statistic.

# Requirements Specification and Design

## Project/Business requirements:

The purpose of this project is to experiment and examine any correlation between weather parameters and covid-19 cases and related death. Apply the basic principles of data modelling to predict number of cases and death using environmental factors like temperature, wind pressure, precipitation, fog etc.

The second part of the project is to visualize the social-economic factors against the total cases of each country e.g. life expectancy score vs total cases, sex ratio, median age etc. to assess tend of Covid-19 and social-eco factors. The study could benefit different countries and governments to cope with the increasing covid-19 cases.

The core outputs would contain the different visualization before data preparation and after preparation. Use prepared data to build predictive models.

## Outcome of project:

* Weather features are correlated with daily cases and daily death or not?
* Experiment weather features could help in predicting daily cases and deaths.
* Apply up to two labels(classes) of severity to the target variables i.e. low cases/low death high cases/high death and run classification models to develop a predictive model and score each severity for belonging either of the newly recoded classes
* Find whether correlation ship exist between socio-eco factors and target variables. Visualize the available countries socio-eco factors vs daily cases and death and see is there any trend present between them.

## Information requirements:

Dataset has been acquired from kaggle. Dataset folder contains individual countries files, each file contained covid-19 states and other features related to weather and social economic. Individual countries’ weather records are combined in one file to temperature dataset. The main dataset contain 10 continuous variables and 3 categorical variables. The main dataset is a subset version of other dataset in folder which contain information about each country. There are two target/independent variables selected for this project i.e. daily cases and daily death.

The link of kaggle dataset is [here](https://www.kaggle.com/aestheteaman01/covcsd-covid19-countries-statistical-dataset).

## Specification and Design:

Dataset folder contains the temperature dataset which is combination of all the countries 3 months’ data i.e. January 2020 to March 2020 covid-19 daily cases and death record with the temperature and other socioeconomic indicators e.g. life expectancy, male/female ratio, lungs patient etc.. Original data is normalized and scaled to use for modeling. Before normalizing/standardization, data exploration visualization is done on original data.

After standardization, different modelling techniques have been applied to dataset to predict number of daily cases and daily death based on weather’s variables. There are two dependent/ target columns have been chosen for this project.

**Design:**

Final report

Data preparation

Data Source (socio-eco data)

Visualization

(Variable vs Covid-19 cases)

Final result/ model

Accept/Reject null hypothesis

Model /evaluate result

Split data and train data

Data preparation

**Modelling**

Data visualisation

Date transformation

Filtering Data

Data source

## Predictive models selection:

To find the most important features of dataset which have strong correlation with Covid-19 cases and death and to predict daily cases and death based on environmental factor following regression models have been selected.

* Multi-linear regression
* Linear regression
* Lasso regression

In second part, target variables in dataset assigned labels/classes to run classification models in order to predict which environmental factor play important role in high and low daily cases and death. Following are the models used for this purpose.

* Logistic regression
* Random forest

In third part, a separate dataset is used to visualize trend between socio-eco factors and ratio of total cases of countries and total population.

### Reason of selecting particular prediction models:

**Multi-linear regression:** The key purpose of selecting Multi-linear regression is to see which features are playing significant role in prediction of daily cases and deaths. It allows you to find relationship between dependent and independent features which is key goal of the project.

**Linear regression:** It is commonly used to predict relationship between continuous independent variables and predicted variable. This method models data with linear combination of the explanatory/predictor variables (attributes) (Satyavishnumolakala, 2020).

**Lasso regression:** Like multi and simple linear regression methods Lasso also key to build and select variables that can be applied to different regression types (Peter, 2017).

**Logistic regression:** It is selected to predict whether particular weather condition cause high number of cases/death or low number of cases/death. It is good model for labelled data. It is easier to implement, interpret, and very efficient to train (Rout, 2020).

**Random Forest:** Random forest generated low correlation decision tree to predict label variable. It is also efficient than decision tree. It can reduce overfitting, bias and overall variance.

**Tools & techniques:**

Regression and classification techniques have been applied to environment dataset to make predictive models. Then, socio-economic indicators dataset used to visualize socio-eco vs covid-19 relationship in order to extract the insight of relationship between different factors and covid-19 spread.

# Implementation:

## Dataset:

Following are the few rows of temperature dataset.

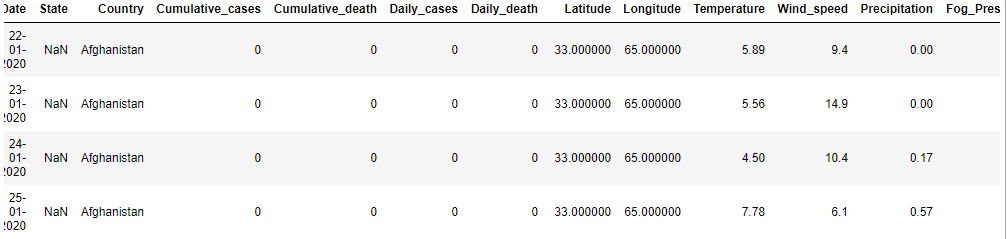


Fig 4.1- original dataset

## Importing and inspecting the data:

Dataset folder contains excel file for each country which has record of weather, global location coordinate, life expectancy, population records. Following are some inside information of dataset.

**Dataset information:**

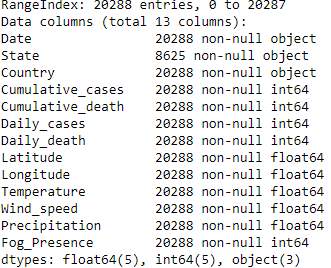


Fig 4.2- Dataset’s columns data type and count

**Dataset description:**

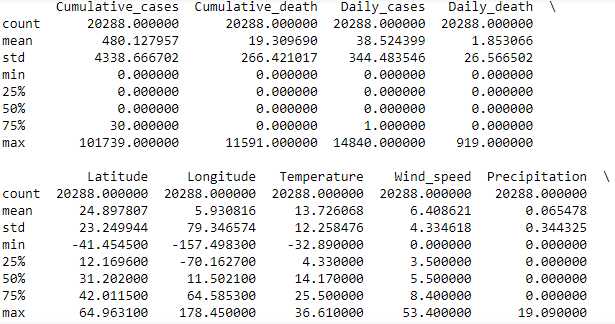


Fig 4.3- Dataset’s columns description

**N/A and empty value check:**

As you can see from fig 4.4 null values in only one column and no NA value found in dataset. Therefore, dataset doesn’t required cleaning,

**Null row count**

Date 0

State 11663

Country 0

Cumulative\_cases 0

Cumulative\_death 0

Daily\_cases 0

Daily\_death 0

Latitude 0

Longitude 0

Temperature 0

Wind\_speed 0

Precipitation 0

Fog\_Presence 0

dtype: int64

Out[15]:

**N/A value count**

Date False

State True

Country False

Cumulative\_cases False

Cumulative\_death False

Daily\_cases False

Daily\_death False

Latitude False

Longitude False

Temperature False

Wind\_speed False

Precipitation False

Fog\_Presence False

dtype: bool

Fig 4.4- Null value column(s)

## Data preparation:

Data in dataset need to be cleaned and prepare for modelling. Different scaling and transformation technique will be applied to prepare data for predictive modelling. “State” column from dataset is dropped because it is not continuous variable and may not be playing part in daily cases and deaths of covid19. We are keeping country and date columns only for visualization and modelling purpose.

Dataset also contains “Cumulative\_cases” and “Cumulative\_death” columns which are also drop from dataframe because they are not required for modelling according to objectives.

### Conversation & filtering:

First, rows which have daily cases value zero are filtered out from the dataset. The reason of filtration of zero daily cases is to avoid overfitting and under fitting of models.

After filtering rows that contains zeros in daily cases column then conversion is applied on date column. Date of each row has been converted to weekdays. This will help in group of cases by weekdays. It helps us to understanding on which most of the cases covid-19 and death have been reported.

Following are the number of each weekday’s daily cases and daily death records.

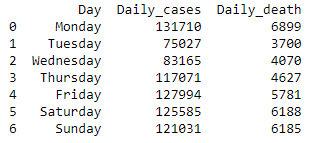


Fig 4.5- Number of Covid-19 cases and death each weekday. Dates converted into weekdays

### Data Visualization:

**Weekdays vs cases:**

Fig. 4.6 graph shows how many COVID-19 cases reported on each day of week during 3 months. In data preparation, dates are converted to days then numbers of cases were counted of each day of week to visualize graph.

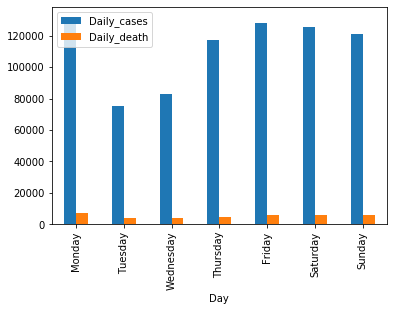


Figure 4.6 weekdays vs total covid cases

It is clear from graph that numbers of cases reported are very high on Monday. The justification for this could be people meet during working day and socialize during working day and as we know it usually takes 4-5 days to covid symptoms appear. Authorities need to focus on those days when covid-19 cases get spiked.

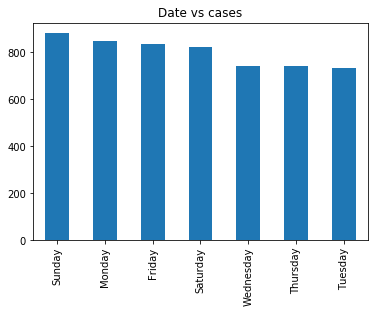


Fig 4.7- Bar chart to show distribution of cases among each weekday.

Fig4.7 graph shows that number of records/rows is almost equally distributed among each weekday i.e. total record in dataset for Monday is similar to total number of record of Tuesday.

**Scattered plot:**

Fig 4.8 is the scatter plot of all continuous variables. The plot shows the correlations between each variable with rest of the other continuous variables. Scattered plot shows that variables are not normally distributed.

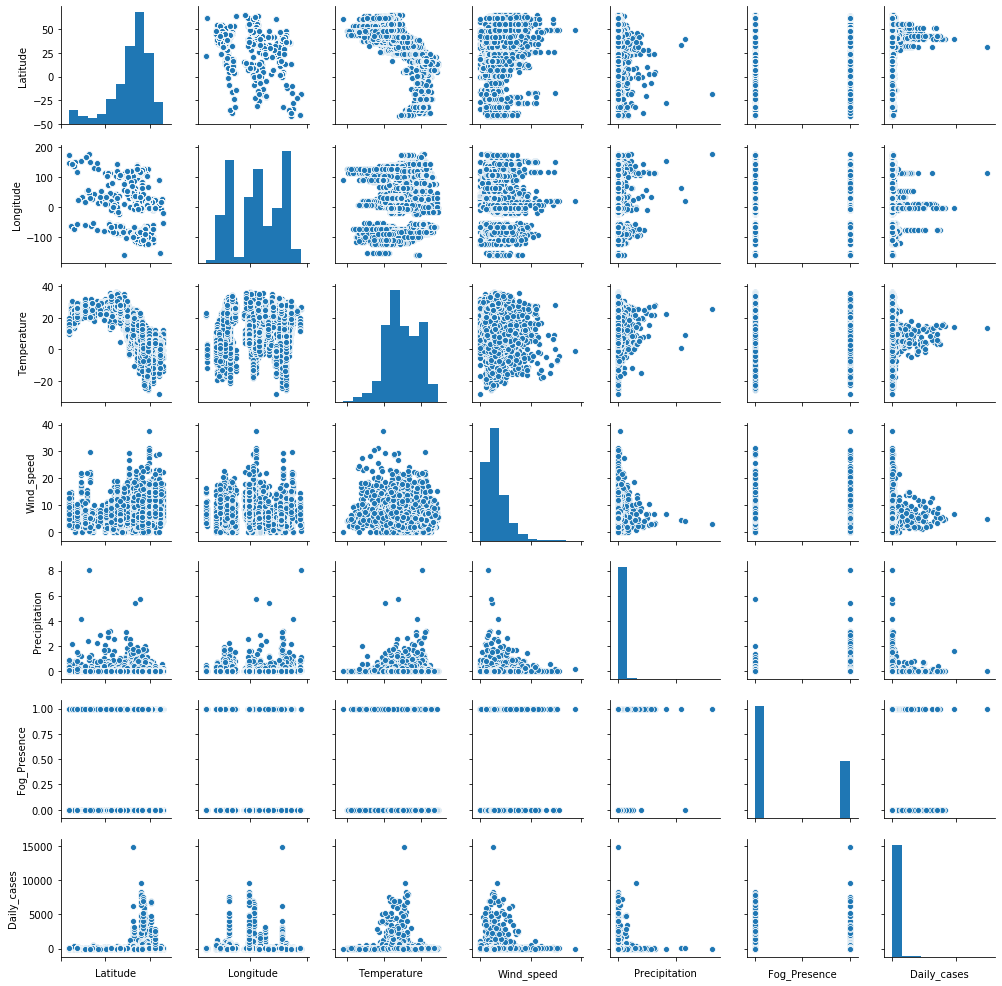


Fig 4.8- scatter plot (correlation fig) of weather’s factors

### Removing outliers

As you can see in fig-4.8 that there are many outliers in the dataset. Fig-4.9 is the scatter plot after removing few outliers from dataset.

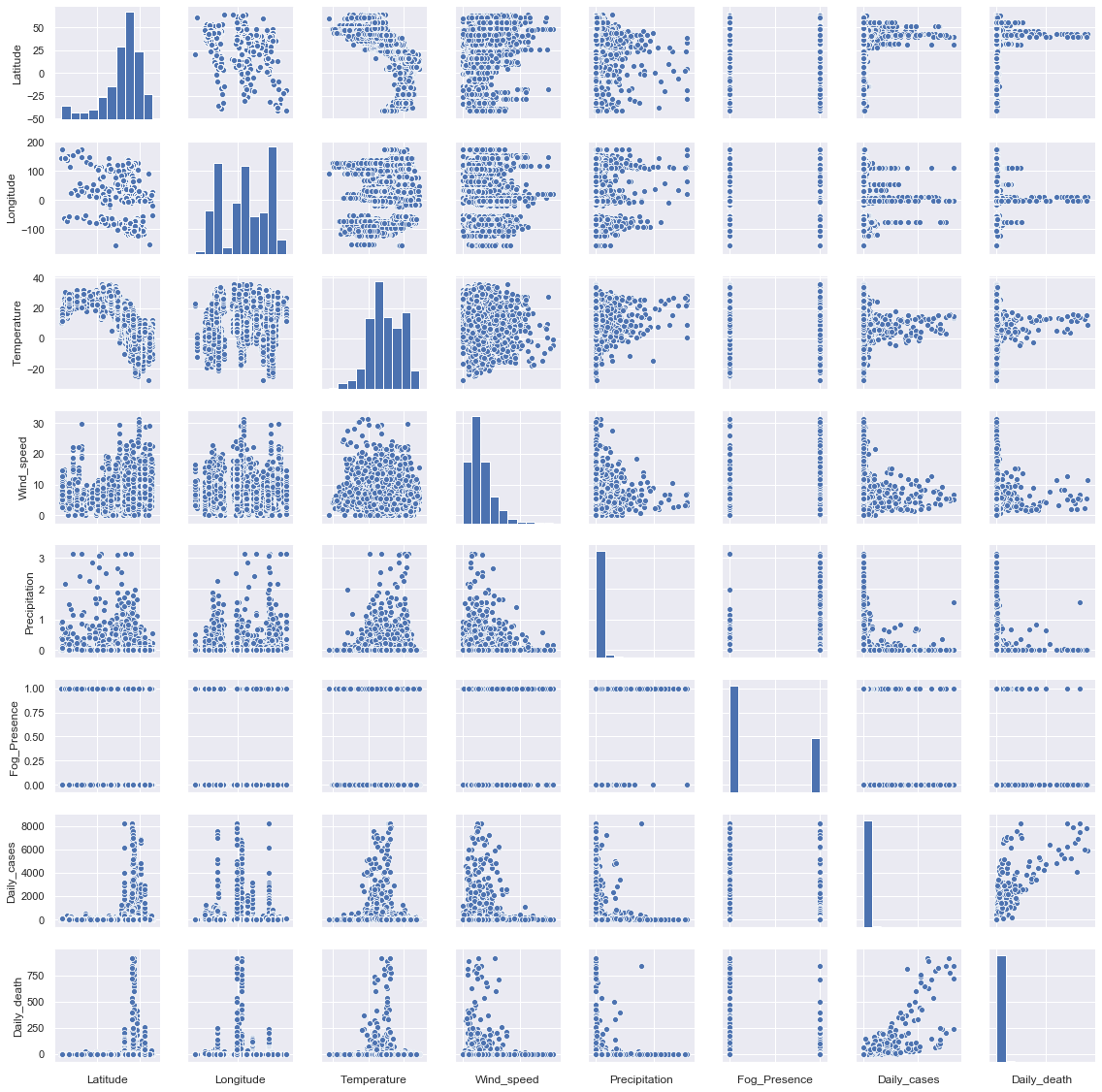
****

Fig 4.9- scatter plot after removing outliers

**Scattered plot with transformed data:**

Log transformation is applied to make data normally distributed. Transformed data then used to generate scattered plot to evaluate whether variables are normally distributed or not.

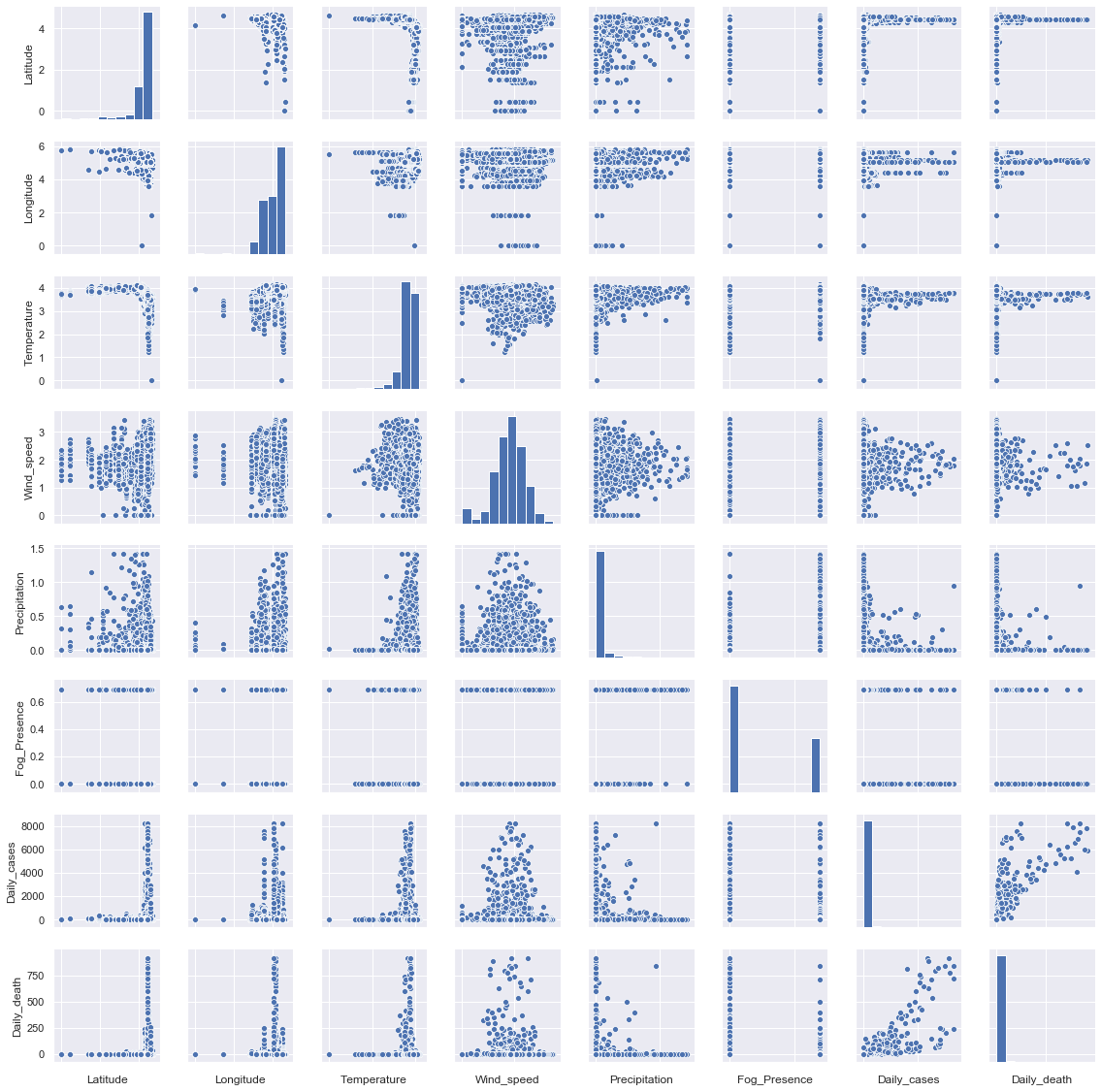


Fig 4.10- scatter plot after log transformation

*Note: As you can see after applying log transform data become normally distrusted slightly but it hasn’t improved too much.*

**Heatmap:**

The next figure is the headmap of all continuous variables except “Daily\_death” variable. The heatmap is generated to visualize the correlation between daily covid-19 cases with other weather continuous variables.

**Deaily cases heatmap:**

The figure 4.11 show heat maps which indicate that “latitude” is most directly correlated variable with daily cases. Whereas “temparature” is most negative correlated variable with the daily cases.

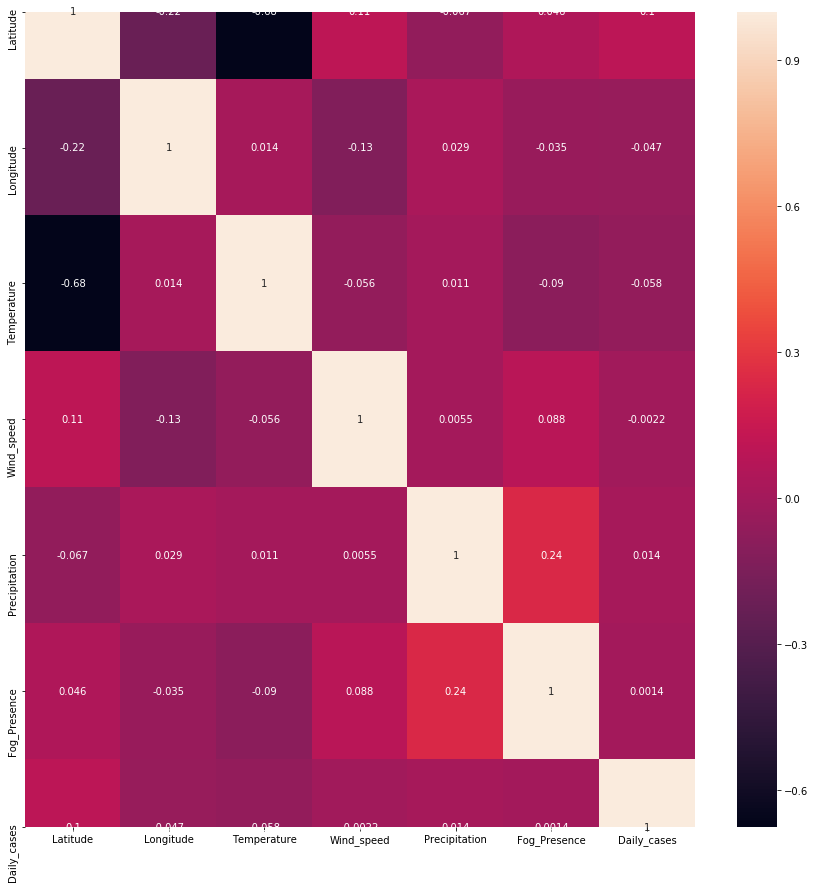


Fig 4.11- Heat map of number of daily cases and weather features

**Daily death heatmap:**

Fig-4.12 graph visualized the correlation between daily death and other continuous variable. The fog presence is most negatively correlated with daily death number and most positively correlated latitude variable.

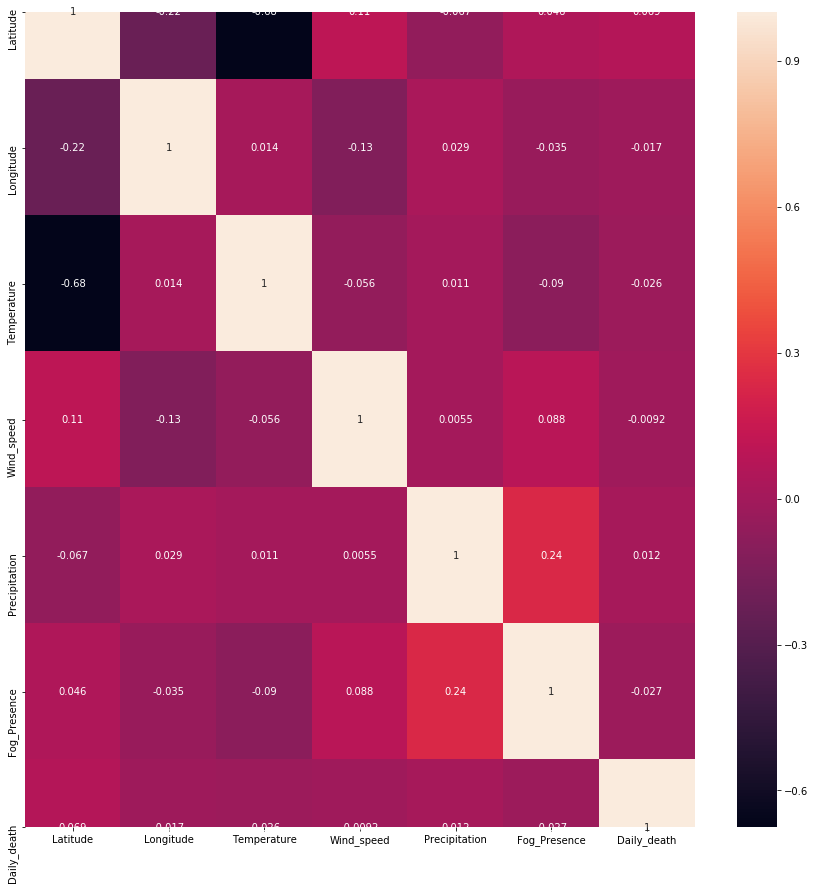


Fig 4.12- Heat map of number of daily death and weather features

### Scaling & standardization:

So far there are three different techniques have been selected to scale and transform data for modeling. Following are the techniques we have used.

*Log transformation*: This transformation is used for normalized data. If data are in several orders it can be normalized using this transformation. In temperature dataset of covid-19 continuous variables are in different ranges. Therefore, log transform is used for make them normalized (Wicklin, 2011).

*MinMaxScaling*: This method of scaling your dataset is used to transform your data between range of 0 and 1. This is commonly used in data analytic modeling (Brownlee, 2020).

*StandardScaling*: it is another technique like MinMaxScaling. In this method, we get series of standardized data whose mean is 0 and standard deviation is 1 (Brownlee, 2020).

## Data Modelling:

Dataframe contains countries weather records from January to march of 2020 with the date, countries’ names, states, covid-19 daily cases and daily deaths records columns. All records belong to different countries.

The Null Hypothesis H0 states; that none of the variables influence Daily cases and daily deaths: **H0:μ1=μ0**

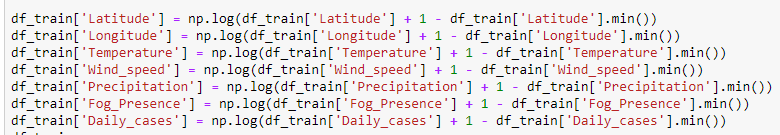
### Regression Modelling

Multilinear regression, linear regression and LASSO models have been chosen for regression modelling.

#### Multi-linear regression

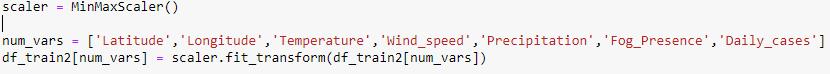
Multilinear regression has been applied to create predictive model for daily cases and daily death. Both log transformation and MinMaxScaling methods have been applied separately and model is evaluated based on each transformations.

**Log transformation**



***Note****: same process repeated for “Daily death” as target column.*

**MinMaxScale**



***Note****: same process repeated for “Daily death” as target column.*

First modelling is done for all continuous variables and “daily cases” is target variable. After that same continuous variables are features variable and “daily death” is the target variable.

Train data and test sets split into 70% and 30% ratio and random state is set to 100.

#### Linear regression

The second model, which is run on dataset, is linear regression. Like multilinear regression linear regression also run for both target/dependent variables i.e. daily cases and daily death.

No transform or scaling has been applied before running linear regression. Columns of main dataset are renamed then modeling is applied

#### LASSO regression

Third model is LASSO, which is run on both “daily cases” and “daily death” column. “StandardScaler” method is used to transform data. Main dataset is divided into train and test sets with ratio 70 % to 30% respectively.

### Classification modelling:

To test whether environmental feature could help in prediction of severity of COVID-19 cases, main dataset’s target variables are divided into two types i.e. high cases and low cases and high death and low death. Experiment conducted to find the link between whether particular feature(s) cause high cases or death by covid-19.

#### Data preparation:

First we divide particular number of Covid-19 cases to low and high. A new column is added in dataset name “cases\_severity”. A “Low” band is applied to those row which contains daily cases less than 1000 and if daily Covid-19 cases more than 1000 each day then it allocates “high” to that row.

Similarly, another column “death\_severity” has been added in dataset and “low” & “high” bands applied where daily deaths are less than 100 and more than “100” respectively.

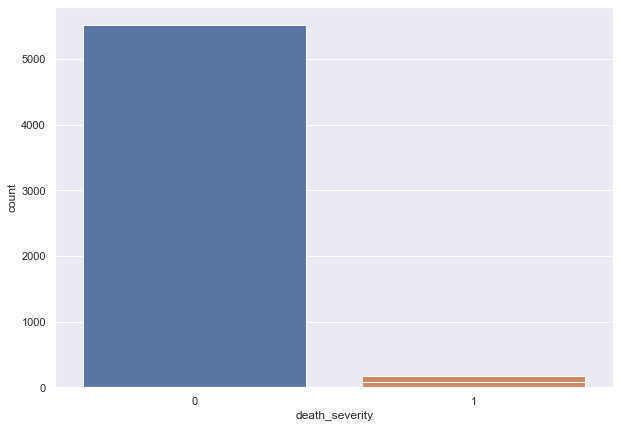


Figure 4.13 Count of severities after allocating band

From figure 4.13 you can see the after applying classes to target variables, dataset become unbalanced. There are very few days when deaths and cases exceeded to “High severity”.

**Balancing target variable:**

To run classification model we need to balance target variables. There could be over fit and under fit issues if we run classification models on original dataset. For this purpose, SMOTE method is used to balance data.

Figure 4.14 represent balanced daily cases classed and figure 4.15 show balanced daily death classes. Bar charts show High and low classes of Daily cases and daily death target variables.

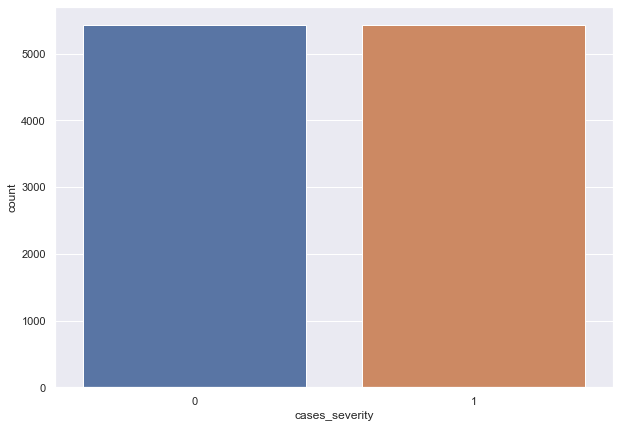


Figure 4.14 Balanced daily cases target variable

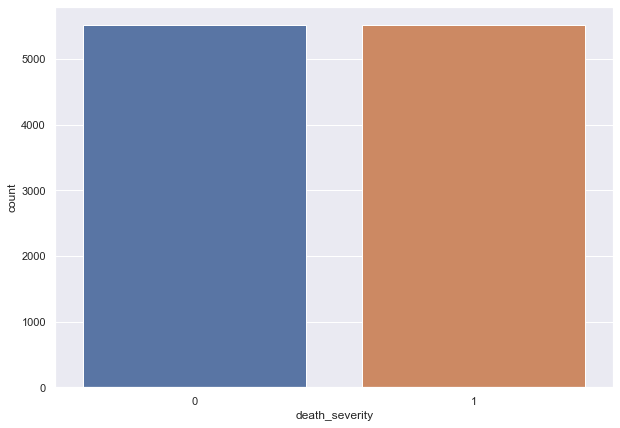


Figure 4.15 Balanced daily death target variable

#### Logistic regression

Logistic regression is run separately to predict low or high daily cases and daily death. Dataset is divided into train and test sets with ratio 70% and 30% respectively. For scaling dataset “MinMaxScaler” method is applied before running logistic regression.

#### Random forest classification

Random forest is run as secondary classification model to predict and check which environmental feature(s) impacting most daily cases or deaths severities. Similar to previous models, dataset is into train and test sets with 70% to 30%.

## Social-Economic factors vs Covid-19 spread:

The second part of this project is to find the correlation ship between some socio-eco features and ratio of covid-19 and population ratio of few countries. The reason of considering ratio is to see how many percentage of population got infected by Covid-19 and then find the relationship between social-economic factors with percentage of infected population.

### Data preparation:

#### Handling null:

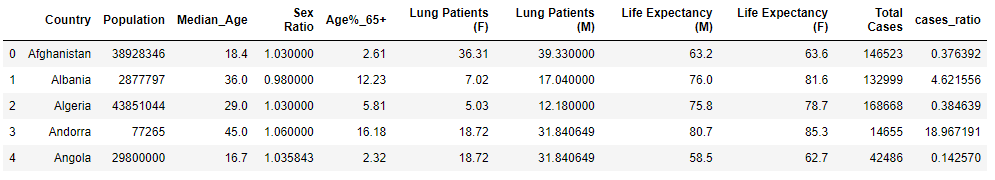
Socio-eco dataset contains null values. Therefore, we replace particular column’s null/not applicable value with mean value of that column. Data is, then, written to separate file.

#### Population and Covid-19 cases ratio column:

New column is added into dataframe which contain each country’s population to covid-19 cases ratio. Following is the formula used to create target column of each country.

**('Total Cases'/ 'Population')\*100**

#### Final dataset:



#### Features vs ratio visualization:

From different type of visualization, relationship between Covid-29 cases ratio and individual feature of socio-eco will be found.

**% of Age above 65 ratio vs covid-19:**

Graph shows that ratio of population which are above 65 year of age. The dataset is sorted by %ofAge in ascending order then line graph is plotted with sorted dataset.

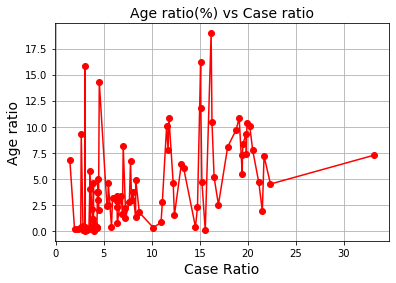


Figure 4.16 Percentage of age above 65 vs Covid cases

**Evaluation:**

Graph slightly indicates that there is linear relationship between age above then 65 and covid-19 infected percentage. Mean, more above 65 age people are, the more cases are reported.

**Life expectancy and covid-19 ratio:**

Fig-4.17 is the life expectancy of male and female in each country vs covid-19 ratio in that country plot. Each country has different life expectancy scores of male and female. CYAN plot represent “Male” while YELLOW represent female numbers.

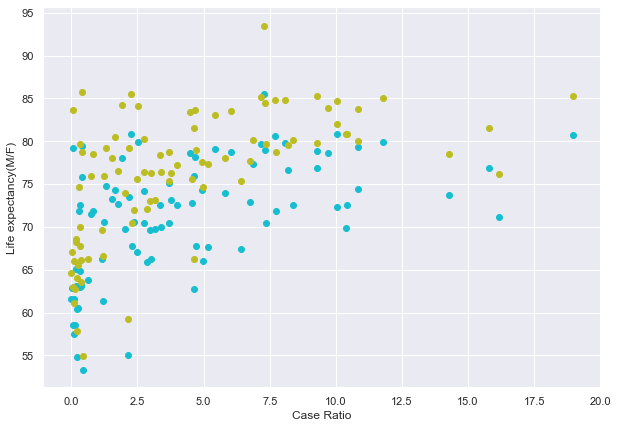


Figure 4.17 Life expectancy (m/f) vs covid cases ratio

**Evaluation:**

It also shows that there could be strong positive relationship between number of life expectancy and covid-19 spread.

**Lung patient vs Covid-19 ratio:**

Lung patient of male and female in each country could be significant factor to get infected by covid-19 virus. Blue plot represents numbers of male lung patients and orange represent female lung patients.

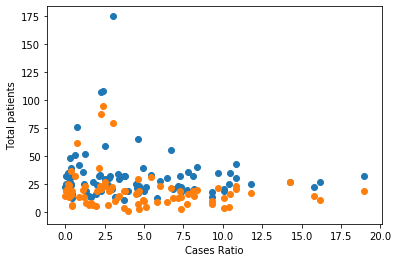


Figure 4.18 lung patient (m/f) vs covid cases ratio

**Evaluation:**

It seems from fig-4.18 graph that if in a country number of male lung patients high then they get infected with virus easily whereas this relationship with number of lung patient in female is weak. Mean, Virus is male oriented.

**Median age vs COVID-19 cases ratio:**

Fig-4.19 graph is a plot of median age of each country and their Covid-19 ratio. Whether high and low median age could be cause of high and low spread of virus cases.

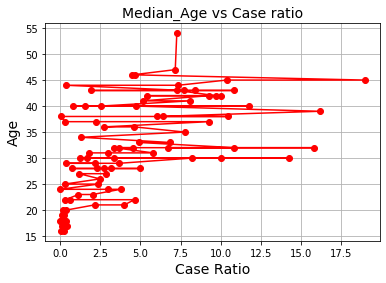


Figure 4.19 Median age vs covid cases ratio

**Evaluation:**

From fig-4.19 graph it can be visualized that if median age in a country is higher than number of people get infected by virus is also high. There is linear correlation between them.

**Sex ratio vs covid-19 case:**

Relationship between sex ratio in country and ratio of covid-19 spread also evaluated in fig-4.20 graph. Dataset is sorted by sex ration in ascending order then line graph generated by dataset.

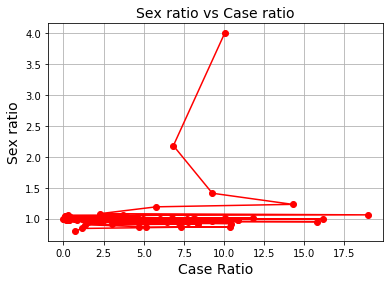


Figure 4.20 sex ratio vs covid cases ratio

**Evaluation:**

It is clear from graph that there is no such linear relationship between sex and covid-19 ratios. Sex ratio is steady mostly across all countries in dataset while ratios of spread of covid-19 are different.

**Heat map of socio-eco vs covid-19 ratio:**

Figure-4.21 is heat map of socio-eco dataset. It shows the correlation between variables of dataset. The purpose of populating correlation chart is to give knowledge which feature(s) value(s) impact high and low ratio of Covid-19 cases.

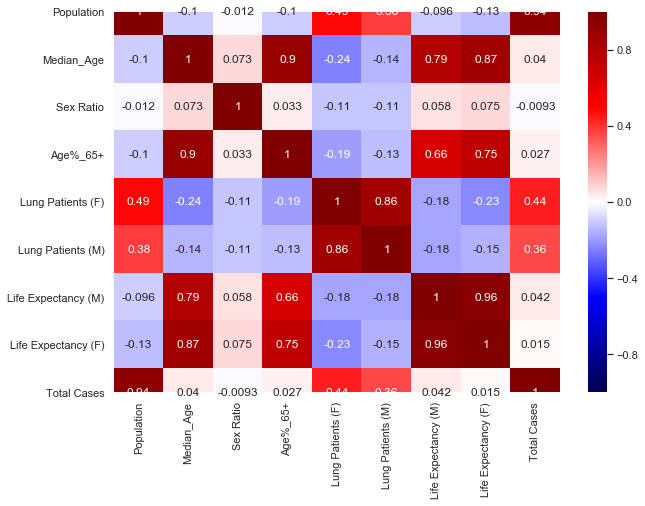


Figure 4.21 Heat map of socio-eco features

**Evaluation:**

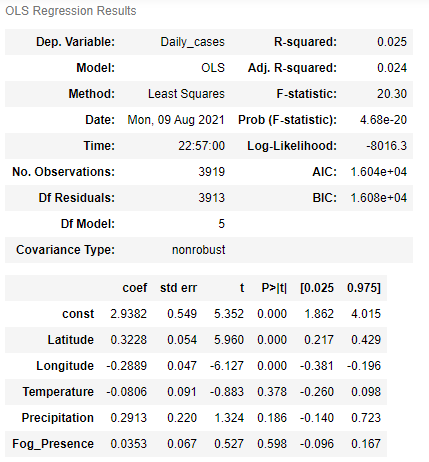
Heat map shows that those countries which have high population suffer from high number of cases of COVID-19. After population, high number of male and female lung patients in a country likely to have high spread of covid-19. Interestingly life expectancy of male has more influence on covid-19 cases. Higher the life expectancy the more covid-19 patient reported but if country has high number of female high life expectancy is not likely to link high cases. Median age has similar kind of correlation like male life expectancy of a country. Female life expectancy and percentage of 65 year above age have very weak positive relation with covid-19 cases. Sex ratio has negative weak correlation ship with covid-19 spread.

# Testing & Results:

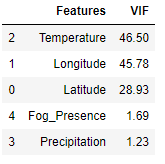
## Multi-liner regression Result

While using train dataset, there are four different results evaluated using multilinear regression i.e. log transform for daily cases and death and MinMaxScale for daily cases and death

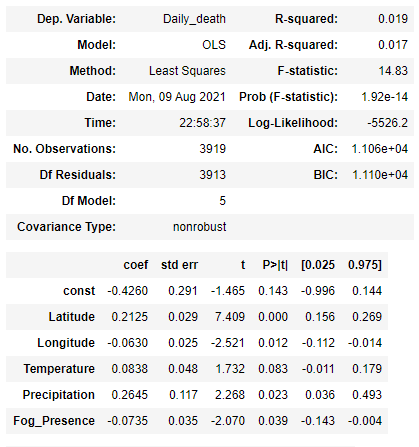
**“Log Transformation” and “daily cases”:**



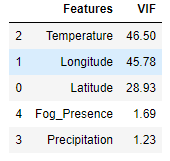
*VIF*



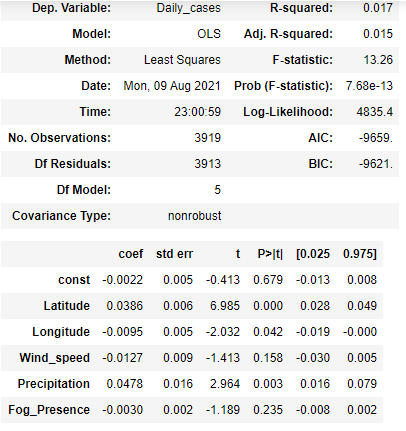
**“Log Transformation” and “daily death”:**



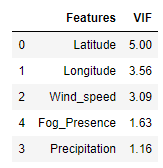
VIF



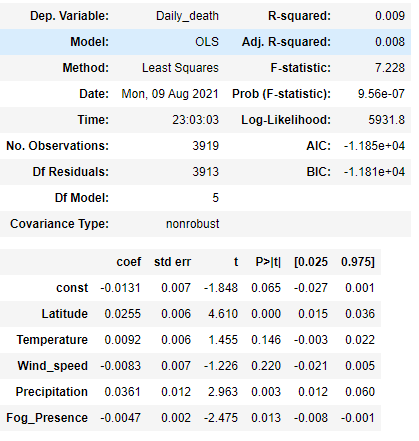
**“MinMaxScale” and “daily cases”:**



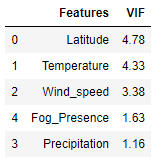
VIF



**“MinMaxScale” and “daily death”:**



VIF



***Note:* As you can see log transform results are better than MinMaxScaler transformation results.**

#### Multi-liner regression evaluation result:

Following table illustrate the scores of Multi-linear regression method for both dependent variables i.e. daily case and daily death. Multi linear regression models evaluated using log transformation as well as MinMaxScaler data transformation.

|  |  |  |  |
| --- | --- | --- | --- |
| **Target variable** | **Transformation method** | **Feature variable** | **Score** |
| Daily cases | Log transform | Latitude,  Longitude, Temperature, Precipitation, Fog\_Presence | 0.0129 |
| Daily cases | MinMaxScaler | Latitude,  Longitude,  Wind\_speed,  Fog\_Presence,  Precipitation | 0.0055 |
| Daily Death | Log transform | Temperature,  Longitude,  Latitude,  Fog\_Presence,  Precipitation | 0.012 |
| Daily Death | MinMaxScaler | Latitude,  Temperature,  Wind\_speed,  Fog\_Presence,  Precipitation | 0.003 |

## Liner regression Result

The model score when “daily cases” as target column is “0.01217942348510792”

The intercept value when “daily cases” scenario is “[50.46510316]”

Coefficients are:

[[ 3.60687238 -0.28825498 1.3806063 -3.51007481 101.71597917

-21.12532413]]

The model score when “daily death” as target column is “0.007281832331929361”

The intercept value when “daily cases” scenario is “[-0.01688669]”

Coefficients are:

[[2.33347516e-01 2.81383458e-03 1.53265362e-01 -1.52593421e-01

2.85253703e+00 -4.68842488e+00]]

|  |  |  |
| --- | --- | --- |
| **Error** | **Daily death error** | **Daily cases** |
| MAE | 11.632046168505786 | 203.8356469550698 |
| MSE | 1874.690015625511 | 242288.92083921688 |
| MAE square | 3.410578568000712 | 14.277102190398084 |

### Simple linear regression evaluation:

For daily cases according to coefficient values

An additional 1 unit in Latitude, Longitude, Temperature, Wind\_speed, Precipitation, Fog\_Presence are associated with an increase in Daily\_cases of [[3.6, 0.28, 1.3, -3.5, 101.7 and -21.1]] widgets

## LASSO result

Following are the result for target column “daily cases”

The alpha of lasso is “3.612”

The MAE of Lasso is 199 with standard deviation 36.21

R ^ 2 of Lasso model for daily cases is 0.010

Following are the coefficient of variables

[('X1', 63.078648818690816),

('X2', -12.199999312526304),

('X3', 0.0),

('X4', -13.416623618270977),

('X5', 18.284967703400554),

('X6', -0.7169282093478736)]

Following are the artifacts for target column “daily death”

Alpha is 0.001

R^2 of daily death is 0.005

The MAE of Lasso is 13.046 with standard deviation 3.174

Coefficients of columns are

[('X1', 4.833342366081169),

('X2', 0.17941793941332618),

('X3', 1.7315854855139525),

('X4', -1.325680453751225),

('X5', 2.0334654716533547),

('X6', -1.4841980324738442)]

### LASSO evaluation result

From above result it is evaluated that similar to multi and simple liner regression, LASSO is also not performing well on the dataset. The prediction score is only 1%.

You can also see ALPHA value is high which mean coefficient of few features tend to zero.

“X3” represent “**Temperature**” and “X6” represent “**Fog**” and there coefficient zero means that these variables are not playing role in prediction of **daily cases**. **Latitude** and **precipitation** are most significant features in prediction of covid-19 cases

For **daily death** the alpha value is 0.001 which means that most of the coefficients of variables are not tend to zero.

Except “**longitude**” all variables have non-zero coefficient mean they are significant features to predict daily death but there prediction score almost zero percent. Most significant features are latitude and precipitation.

## Overall Regression result:

Results of regression models show that models have very poor performance. R square-value of linear regression and scores of LASSO are very low but on the other hand linear regression coefficient is not 0 so we cannot reject ***null hypothesis***. Therefore, regression predictive model is not possible.

From multilinear regression result it is evaluated that following are the significant variables in spread of Covid-19 cases and death.

**For death:**

**Temperature** and **latitude** are the significant variables that play role in prediction of death by covid-19

**For daily cases:**

**Longitude**, **latitude** and **temperature** are the significant variables that play role in prediction of Covid-19 daily cases.

## Logistic regression Result

Following result extracted from logistic regression on train and test dataset.

**Accuracy of Logistic regression classifier on training daily cases set: 0.65**

**Accuracy of Logistic regression classifier on test daily cases set: 0.64**

**Accuracy of Logistic regression classifier on training daily death set: 0.72**

**Accuracy of Logistic regression classifier on test daily death set: 0.73**

#### Logistic regression evaluation

From train and test regression result it can be concluded that results are much closed and there is very less room of error.

## Random forest Result

Random forest classification Model Score for daily cases is 99.03

And the explained variance score is 0.9614319633261643

Random forest classification Model Score for daily death is 99.03

0.9614319633261643

#### Random forest evaluation

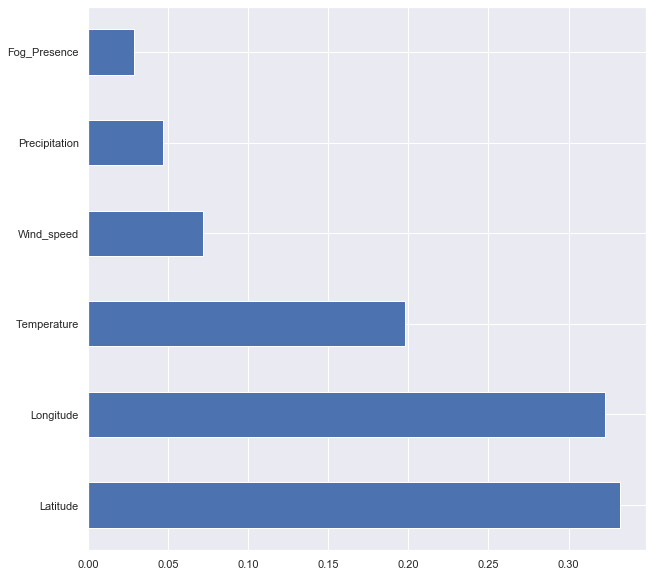


Figure 5.4 Important features to predict daily cases

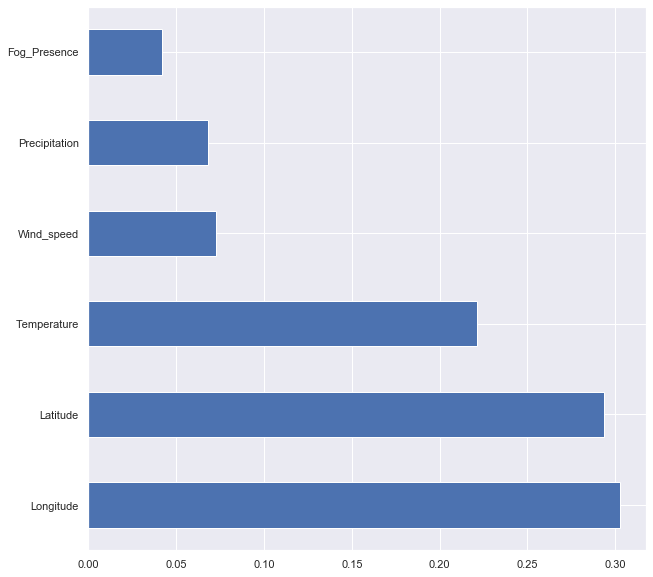


Figure 5.5 Important features to predict daily death

# Conclusion and Future Work:

It doesn’t seem that temperature and environment factors can be used for prediction of number of covid-19 cases and death. However, from project’s results we can extract hypothesis that temperature and locations i.e. longitude and latitude of countries do have some relationship with covid-19 cases and death. The higher the latitude value(either positive or negative) is, the higher the number of covid-19 which means if countries located north and south they will have relatively higher number of cases than countries located near to equator (Control, 2021). Temperature is negative correlated to covid-19 figures means that countries have low temperature have higher covid rates. Environmental factors have weak impact on daily death as compare to daily cases.

If number of covid-19 cases and death assigned severities like high and low to predict environmental factors can predict daily severity i.e. whether particular environment condition let high covid-19 daily cases/death or low then prediction models perform well.

From Socio-eco features, it is clear countries with high population likely to have high ratio of covid-19 cases to population. Visualization and heat map also show that if people of countries contain high number of lung patients, regardless of male or female patient, likely to face large number of covid-19 cases. Life expectancy of female of a country and median age figures also influence slightly covid-19 cases. Rest of the factors of socio-eco dataset i.e. male life expectancy, sex ratio and percentage of age above 65 don’t seem to have strong link with spread of covid-19.

## Limitation

Environmental dataset only contains initial stage data of Covid-19. This means that not all fully reached to countries. Hence, it may create outlier in dataset that one country is in very initial stage of covid while other gone far in this regards.

Predicting daily cases and death is very granular and predicting such granular level is not possible in medical field. Instead of predicting daily, prediction should be conducted on weekly or monthly bases.

## Future work

Large duration data of weather with more environmental features should be collected then run regression models which are used in this project to make predictions.

Build confusion matric for classification models used in this project.

The socio-eco dataset is used in this project contains very few countries data. In future, all countries’ data could be collected and regression modelling could be built to predict covid-19 cases by socio-eco factors.

Regression modelling should be applied to socio-eco dataset.

# Appendix:

*Code to transform negative and values between 0 and 1:*

df\_train3['Latitude'] = np.log(df\_train3['Latitude'] + 1 - df\_train3['Latitude'].min())

df\_train3['Longitude'] = np.log(df\_train3['Longitude'] + 1 - df\_train3['Longitude'].min())

df\_train3['Temperature'] = np.log(df\_train3['Temperature'] + 1 - df\_train3['Temperature'].min())

df\_train3['Wind\_speed'] = np.log(df\_train3['Wind\_speed'] + 1 - df\_train3['Wind\_speed'].min())

df\_train3['Precipitation'] = np.log(df\_train3['Precipitation'] + 1 - df\_train3['Precipitation'].min())

df\_train3['Fog\_Presence'] = np.log(df\_train3['Fog\_Presence'] + 1 - df\_train3['Fog\_Presence'].min())

df\_train3['Daily\_death'] = np.log(df\_train3['Daily\_death'] + 1 - df\_train3['Daily\_death'].min())

## Project Code

Project code is written using jupyter – python. Code is submitted as separate jupyter file as well as as pdf.

# Bibliography

Alsunaidi, S. J. et al., 2021. Applications of Big Data Analytics to Control. *Sensors,* Volume 2282, p. 21.

Arman Behnam, R. J., 2021. *springer.com.* [Online]   
Available at: https://link.springer.com/article/10.1007/s40808-021-01086-8  
[Accessed 11 5 2021].

Brownlee, J., 2020. *How to Use StandardScaler and MinMaxScaler Transforms in Python.* [Online]   
Available at: https://machinelearningmastery.com/standardscaler-and-minmaxscaler-transforms-in-python/  
[Accessed 11 July 2021].

Control, E. C. f. D. P. a., 2021. *COVID-19 situation update worldwide, as of week 32, updated 19 August 2021.* [Online]   
Available at: https://www.ecdc.europa.eu/en/geographical-distribution-2019-ncov-cases  
[Accessed 18 August 2021].

Hadi Eslami, M. J., 2020. *The role of environmental factors to transmission of SARS-CoV-2 (COVID-19).* [Online]   
Available at: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7226715/  
[Accessed 6 June 2021].

Peter, 2017. *What are the pros and cons of lasso regression?.* [Online]   
Available at: https://www.quora.com/What-are-the-pros-and-cons-of-lasso-regression  
[Accessed 8 August 2021].

Rout, A. R., 2020. *Advantages and Disadvantages of Logistic Regression.* [Online]   
Available at: https://www.geeksforgeeks.org/advantages-and-disadvantages-of-logistic-regression/  
[Accessed 8 August 2021].

Satyavishnumolakala, 2020. *Linear Regression -Pros & Cons.* [Online]   
Available at: https://medium.com/@satyavishnumolakala/linear-regression-pros-cons-62085314aef0  
[Accessed 09 08 2021].

Wicklin, R., 2011. *Log transformations: How to handle negative data values?.* [Online]   
Available at: https://blogs.sas.com/content/iml/2011/04/27/log-transformations-how-to-handle-negative-data-values.html  
[Accessed 5 July 2021].