PROJECT TITLE

Covid 19 analysis across the globe

TEAM MEMBERS

Dinesh Tadepalli
Sai Sandeep Gollamudi
Nishanth Bharadwaj Boppa
Sunny Poosala
Naveen Paritala

DETAILS OF TEAM MEMBERS

Student Id's & Email addresses of the team members are listed below

Dinesh Tadepalli, 11512545 – dineshtadepalli@my.unt.edu

Sai Sandeep Gollamudi, 11512548 – saisandeepgollamudi@my.unt.edu

Nishanth Bharadwaj Boppa, 11532558 – nishanthbharadwajboppa@my.unt.edu

Sunny Poosala, 11520296 – sunnypoosala@my.unt.edu

Naveen Paritala, 11535603 – naveenparitala@my.unt.edu

ABSTRACT

Covid 19 pandemic has caused a lot of stress across various communities, and countries. No one possibly imagined a pandemic at the start of 2019. Something that started in one country, spread rapidly across the globe.

Though Covid 19 affected almost every country, there are some countries that recorded significantly higher death rates than the rest of the countries, there are various factors in consideration for this difference.

We as a team, for this project, want to verify if the GDP per Capita, Cardiovascular death rates, and Human Development Index affected the mean value of Covid deaths per million between countries.

DATA DESCRIPTION AND STATISTICAL TESTS

- We are working on the dataset chosen from Kaggle. The dataset consists of daily covid cases, and deaths along with various indexes of the country, for example, extreme poverty, population density, life expectancy, and various other parameters.
- The dataset consists of 67 attributes and more than one lakh fifty thousand records.
- For this particular project, we decided to focus only on certain parameters of the country like ‘GDP per Capita’, ‘Human Development Index’, ‘Cardiovascular death rate’ along with ‘total cases’, ‘total death’, and ‘death per million’ caused by covid.

EXPLORATORY DATA ANALYSIS

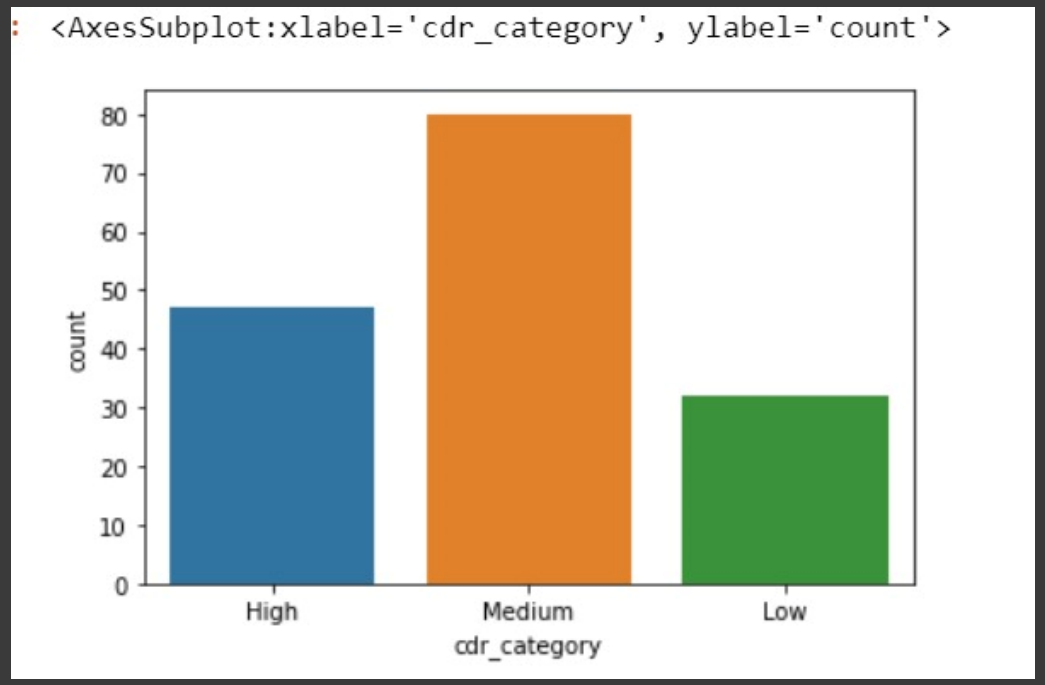
- As per our requirement, we don't require the daily count of covid cases and death from early 2020, So we took the last record from each country in the dataset which included total cases and total deaths that happened in the county due to Covid up to that point in time.
- As we are only focusing on particular attributes for this project, we have selected them and moved them into a new data frame and then removed the null values from the records.
- The final shape of the data frame is (159,9), 159 rows with 9 columns.

EXPLORATORY DATA ANALYSIS

```
[65] df_stat.describe()
```

	total_cases	total_deaths	total_deaths_per_million	cardiovasc_death_rate	diabetes_prevalence	gdp_per_capita	hospital_beds_per_thousand	human_development_index
count	1.590000e+02	159.000000	159.000000	159.00000	159.000000	159.000000	159.000000	159.000000
mean	2.616505e+06	36605.603774	1184.565063	257.31722	8.046415	19775.640157	2.940132	0.739038
std	7.902384e+06	107426.892689	1233.228738	120.09498	4.029152	19863.316059	2.338385	0.145133
min	2.757000e+03	5.000000	3.101000	79.37000	0.990000	661.240000	0.100000	0.394000
25%	6.319350e+04	799.000000	169.584000	164.00150	5.425000	5456.520500	1.300000	0.633000
50%	4.815120e+05	5381.000000	819.656000	240.20800	7.110000	13593.877000	2.300000	0.759000
75%	1.934836e+06	20982.500000	1924.617000	329.78850	10.080000	27827.371500	3.850500	0.851000
max	7.826944e+07	931741.000000	6264.019000	724.41700	22.660000	116935.600000	13.050000	0.957000

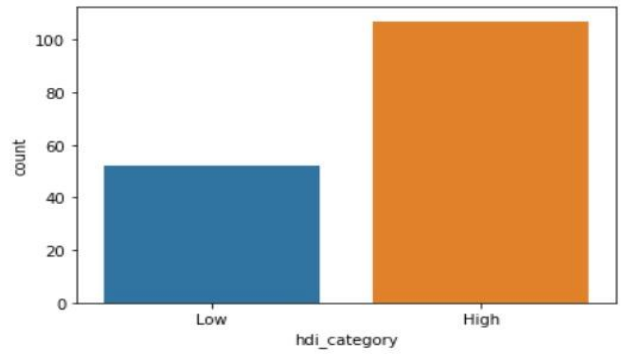
VISUALIZATIONS



Cardiovascular death rate category count visualization

VISUALIZATIONS

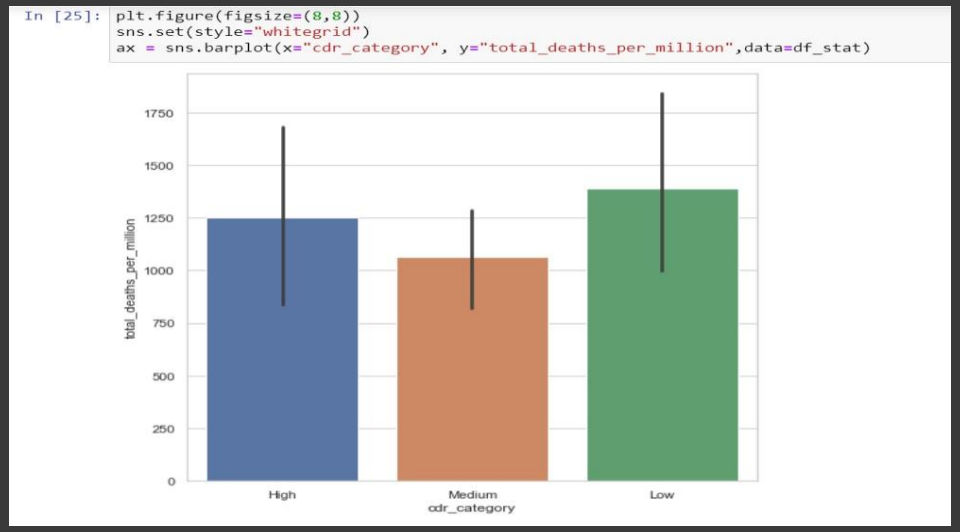
```
In [22]: sns.countplot(df_stat.hdi_category)

C:\anaconda\lib\site-packages\seaborn\_decorators.py:36: FutureWar
ersion 0.12, the only valid positional argument will be `data`, an
sult in an error or misinterpretation.
  warnings.warn(

Out[22]: <AxesSubplot:xlabel='hdi_category', ylabel='count'>
```



Human development index category count visualization.

VISUALIZATIONS

```
In [25]: plt.figure(figsize=(8,8))
         sns.set(style="whitegrid")
         ax = sns.barplot(x="cdr_category", y="total_deaths_per_million",data=df_stat)
```



Total Covid deaths per million in different CDR category countries

INFERENCES

As part of the project, we are drawing three inferences.

- We wanted to check if there is a difference in the mean of “deaths per million” due to covid between high and low “human development index” countries.
- For the second inference, we wanted to verify if there is a difference in the mean value of “deaths per million” between high, medium, and low “cardiovascular death rate category” countries.
- For the third inference, we wanted to check if there is an association between the “GDP per Capita” and Covid’s “Deaths per million” value of the country.

STATISTICAL TESTS

- For inference 1, we are using the **Mann-Whitney U test** to verify if there is a difference in the mean value of "Total deaths per million" between high and low HDI countries

- Coming to inference 2, we are using the **Kruskal-Wallis test** to verify if there is a difference in the mean value of "Total deaths per million" between high, medium, and low CDR countries.

- For inference 3, we are using **Spearman's Rho** to verify the correlation between "GDP per Capita" and "Total deaths per million" of the country

INFERENCE 1

The null hypothesis for inference 1 is

H0: There is no difference in the mean value of “Total deaths per million” between high and low HDI (human development index) countries.

Whereas the alternate hypothesis is

H1: There is a difference in the mean value of “Total deaths per million” between high and low HDI (human development index) countries.

INFERENCE 1

What is Human Development Index?

The Human Development index can be defined as the average of achievements in the basic human development features health, standard of living, and knowledge of the people. Each country is rated between 0 and 1.

MANN-WHITNEY U TEST

When we tested inference 1 with the Mann-Whitney U test we got P value less than 0.05 which is less than the significance value we have taken (0.05).

So, we **rejected the null hypothesis** for inference 1.

INFERENCE 2

For inference 2 the null hypothesis is

H0: There is no difference in the mean value of “Total deaths per million” between high, medium, and low CDR(Cardiovascular death rate) category countries.

And the alternate hypothesis is

H1: There is a difference in the mean value of “Total deaths per million” in between high, medium, and low CDR(Cardiovascular death rate) category countries.

INFERENCE 2

What is Cardiovascular death rate?

Cardiovascular death rate is defined as the number of deaths due to cardiovascular diseases per one lakh individuals in the country.

KRUSKAL-WALLIS TEST

When we tested inference 2 with the Kruskal-Wallis test we got p value more than 0.05.

So, we **fail to reject null hypothesis** for inference 2

INFERENCE 3

For inference 3, the null hypothesis is

H0: There is no association between “GDP per Capita” of the country and the COVID “Deaths per million” value.

the alternate hypothesis is

H1: There is an association between “GDP per Capita” of the country and the COVID “Deaths per million” value.

INFERENCE 3

What is GDP per Capita?

GDP per Capita is one of the widely used measurements for the living standards in the country. It is calculated by dividing the total value generated by the economy of the country in that particular year by its population. A High GDP per Capita is generally associated with high household incomes. A high household income means generally having the ability to spend money on a better health care and also higher GDP per Capita also means the government has better ability to provide assistance and health care to its population.

In short, GDP per capita is often associated with the prosperity of the country.

SPEARMAN'S RHO

When we tested inference 3 with the Spearman's rho we got correlation value of 0.49 and P value less than 0.05.

So, we **rejected null hypothesis** for inference 3.

There is a corelation between GDP per Capita and total deaths per million due to covid

RESOURCES AND RELATED PROJECTS

There are a few websites from which we have understood performing statistical tests in python and have been taken for references.

https://www.geeksforgeeks.org/linear-regression-python-implementation/

The above article has information on linear regression algorithm. We referred from the above article partly for performing linear regression algorithm as part of the project.

https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.spearmanr.html

The above link leads to information on implementing Spearman r correlation test. We used the information to understand more about spearman r test.

RESOURCES AND RELATED PROJECTS

https://www.reneshbedre.com/blog/mann-whitney-u-test.html

A lot of important information has been provided on the above website regarding Mann- Whitney U test which also includes Mann-Whitney U statistical test implementation using Python. The above website has been used as reference.

https://www.scribbr.com/statistics/statistical-tests/#flowchart

Apart from the lectures, material provided by the professor Mark Albert, the above article has been used to understand on how to perform appropriate statistical test on the data.

RESOURCES AND RELATED PROJECTS

We found one article on the web named “Using excess deaths and statistics to determine COVID 19 mortalities” that is authored by Lucas Böttcher. The link can be found at

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8127858/

It is an open access article. It deals with Factors such as varied definitions of mortality, uncertainty in disease prevalence, and biased sampling complicate the quantification of fatality during an epidemic. It mainly revolves around excess deaths and mortalities and is contrasting from the project we chose to implement.

DATA SPECIFICATION

The dataset originally contained nearly one lakh sixty thousand records and 67 attributes. For this project, we have decided to draw inferences only on few attributes and as we didn't require the daily covid cases and deaths in each country from early 2020, so we have taken the final record from each country this has significantly reduced the data frame size.

At the end of exploratory data analysis our data frame size is (159,9), 159 records with 9 attributes which included Location, Total covid cases, Total covid deaths, Death per million, Cardiovascular death rate, Human Development Index, GDP per Capita of 159 countries

DATA SPECIFICATION

The Observation matrix for the selected attributes is attached below.

DATA SPECIFICATION

We have drawn three inferences based on “Human Development Index”, “Cardiovascular death rate”, and “GDP per Capita” with “Total deaths per million”.

For the linear regression model, we have given three attributes, “Human Development Index”, “Cardiovascular death rate”, and “GDP per Capita” as **features** and predicted the “Total deaths per million” attribute.

DESIGN AND MILESTONES

Tools and Frameworks

- We have used Pandas, Mat plot, and Seaborn libraries for exploratory data analysis and visualizations.
- We have used the Scikit-learn framework for creating a linear regression model.
- We have used Jupyter Notebook platform to work on this project.

Design

We have selected the required attributes from a large number of attributes available in the dataset. We have selected the final record from each country and removed null values from it. We have passed the selected attributes “GDP per Capita”, “Human Development Index”, and “Cardiovascular death rate” as inputs into the linear regression model and the predictor value is “Total deaths per million”.

DESIGN AND MILESTONES

We have divided the dataset in the ratio of 70:30 for training and testing, respectively. We have used RMSE to test the accuracy of the model

Finally, the milestones we as a team set for this project are mentioned below.

Milestone 1 Project Proposal Slides on 04/12/2022.

Milestone 2 An Internal review on project regarding statistical tests on 04/19/2022.

Milestone 3 An internal review on the possible inferences that can be drawn, review on final project report that is to be submitted and see if any improvements can be made. 04/26/2022.

Milestone 4 Final project report and record submission on 04/29/2022

REPOSITORY/ ARCHIVE

You can access the dataset, notebook file with executed code and proper documentation, along with the PPT and recording in the below repository link.

https://github.com/sandeep-gollamudi/Empirical-analysis-project

We have also attached all the above mentioned files through Canvas.