How does the trend of COVID19 cases relate to deaths?

BY ZAHRA ADAHMAN, MBS

SEPTEMBER 30TH, 2020.

Introduction

- Coronavirus disease (COVID19) was identified as a novel coronavirus following reports of an unknown respiratory infections sometime fatal, emerging from Wuhan, China.
- The COVID19 virus rapidly led to a pandemic, as it transmitted by air and contact, hence the swift spread across earth (2). Almost ever country has been inflicted by the virus.
- The **fatality of this disease is important to understand and track**. As of moment, September 30TH, 2020 over 33.7 million people has been confirmed infected and 1.01 million dead from the disease. In united States of America alone, only 7 million has been infected and 206,000 people dead from the disease.
- This project aims to understand the relationship between the number of total COVID19 case and the total number of deaths related to COVID19.

Data

- COVID19 data was obtained in July 2020 from Amazon Web Services (AWS) data exchange provided by IHME is an independent population health research center at UW Medicine, part of the University of Washington.
  - The .csv file provides more COVID19 data from countries around the world.
  - However, the file contains more detailed information of the United States of America (USA) and Canada by State/Province and county.

```
COVID19NA.dtypes
```

COUNTY_NAME	object
PEOPLE_POSITIVE_CASES_COUNT	int64
PROVINCE_STATE_NAME	object
REPORT_DATE	object
CONTINENT_NAME	object
DATA_SOURCE_NAME	object
PEOPLE_DEATH_NEW_COUNT	int64
COUNTRY_ALPHA_3_CODE	object
COUNTRY_SHORT_NAME	object
PEOPLE_POSITIVE_NEW_CASES_COUNT	int64
PEOPLE_DEATH_COUNT	int64

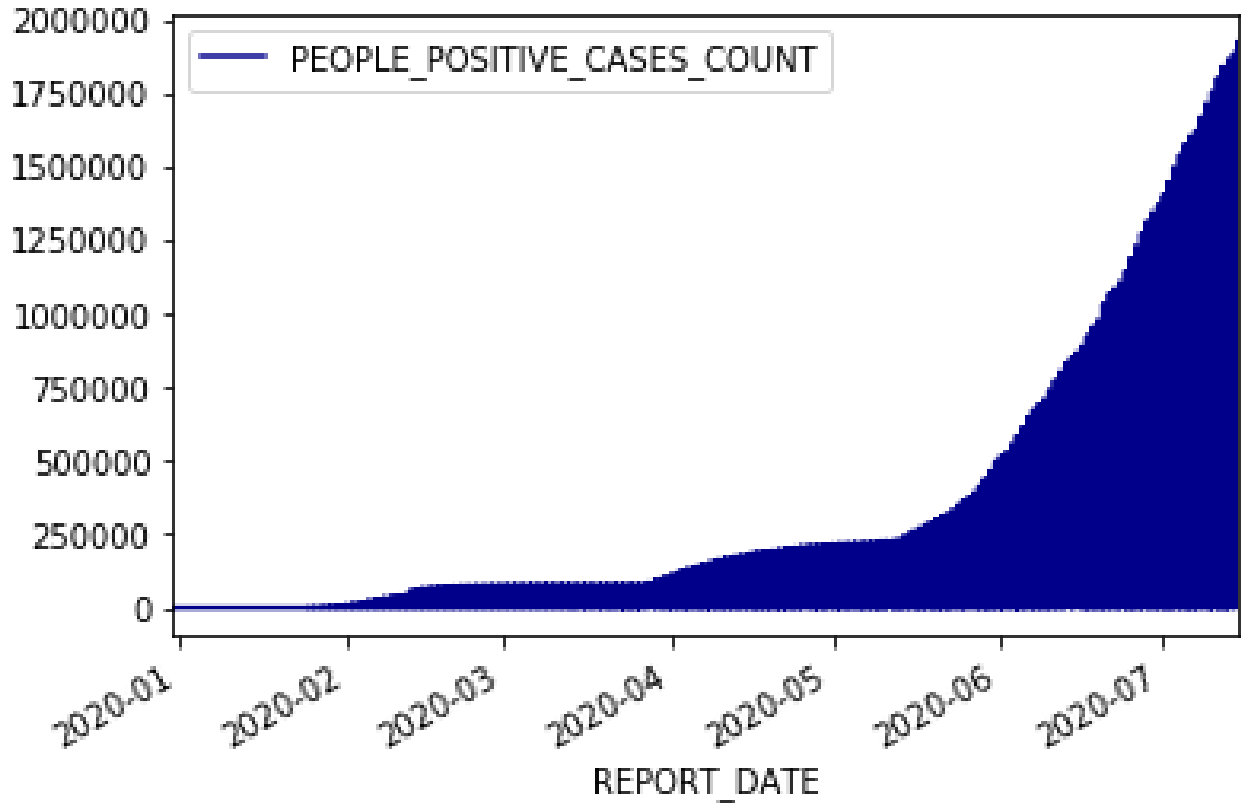
```
dtype: object
```

Table 1. Variables of the COVID19 data

Methodology

- Python 3 programming language
  - Pandas library
  - Matplotlib package - plots
  - Seahorse package - linear regression
  - SciKit learn package – linear regression and modeling, R^2 score.
- Variables – Total positive COVID19 case count and death count.

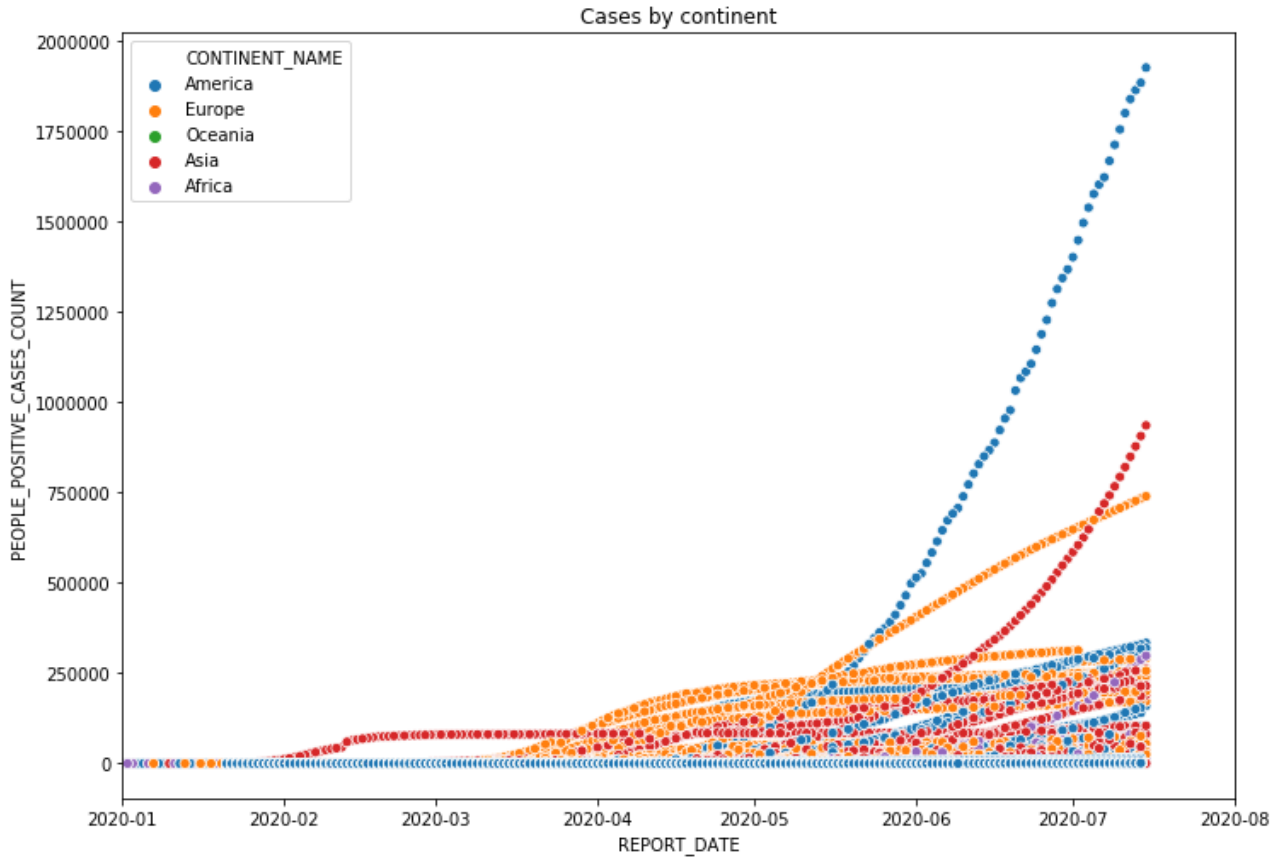
Data Visualization



- The world-wide total positive COVID19 case count trend increased from January 2020 to July 2020.

Fig 1. Total positive COVID19 case by date from wrangled ‘COVID19_World.csv’ data.

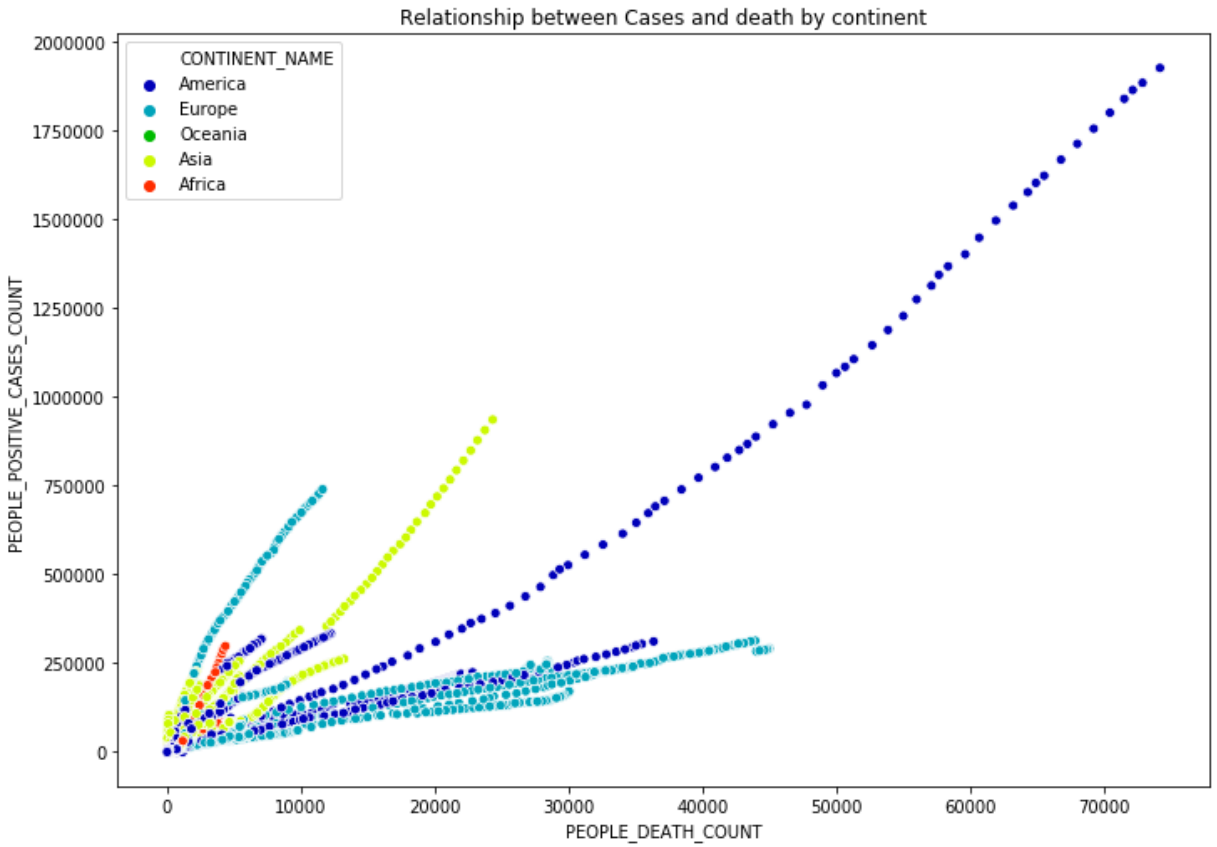
Data Visualization



- The total positive COVID19 case count in the Continent of Americas (include North and South Americas) trend increased from January 2020 to July 2020 **at the highest rate.**

Fig 2. Total positive COVID19 case count by continent from wrangled 'COVID19_World.csv' data.

Data Visualization



- There's a linear relationship between the variables, total positive COVID19 case count and death count.

Fig 3. Relationship between positive COVID19 case count and death count by continent.

Data Modeling-World COVID19

Splitting into training and testing data.

- ‘COVID19_World.csv’ Data was split at the 80:20 ratio based on the Pareto principle.

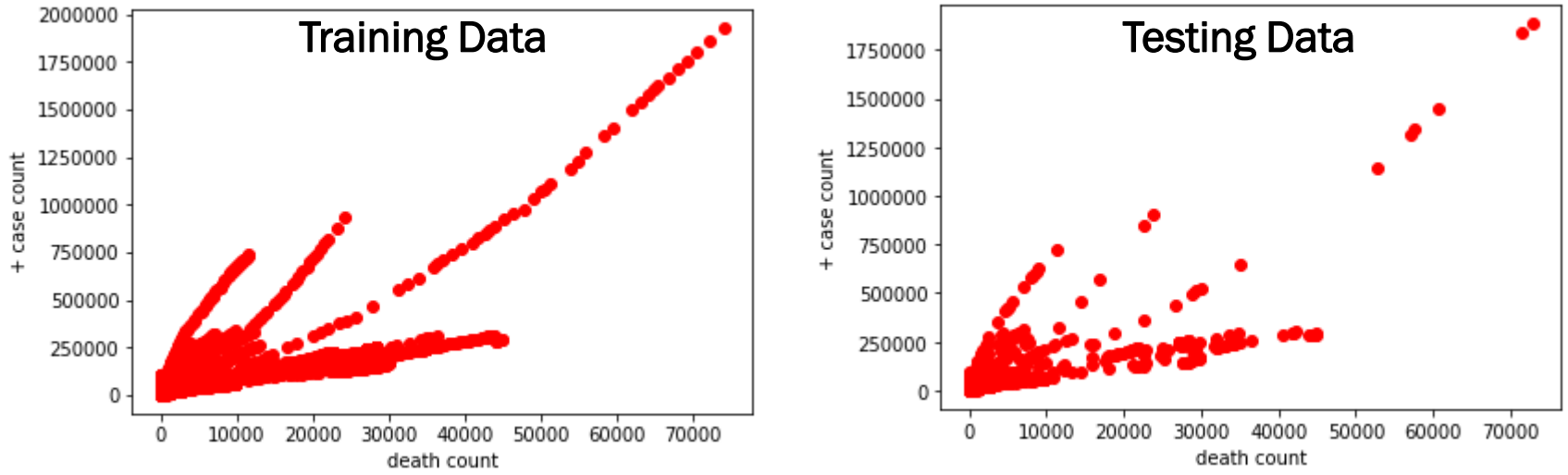


Fig 4. Training and testing data showing the relationship between positive COVID case count and death count by continent.

Data Modeling-World COVID19

Running linear regression analysis on training data.

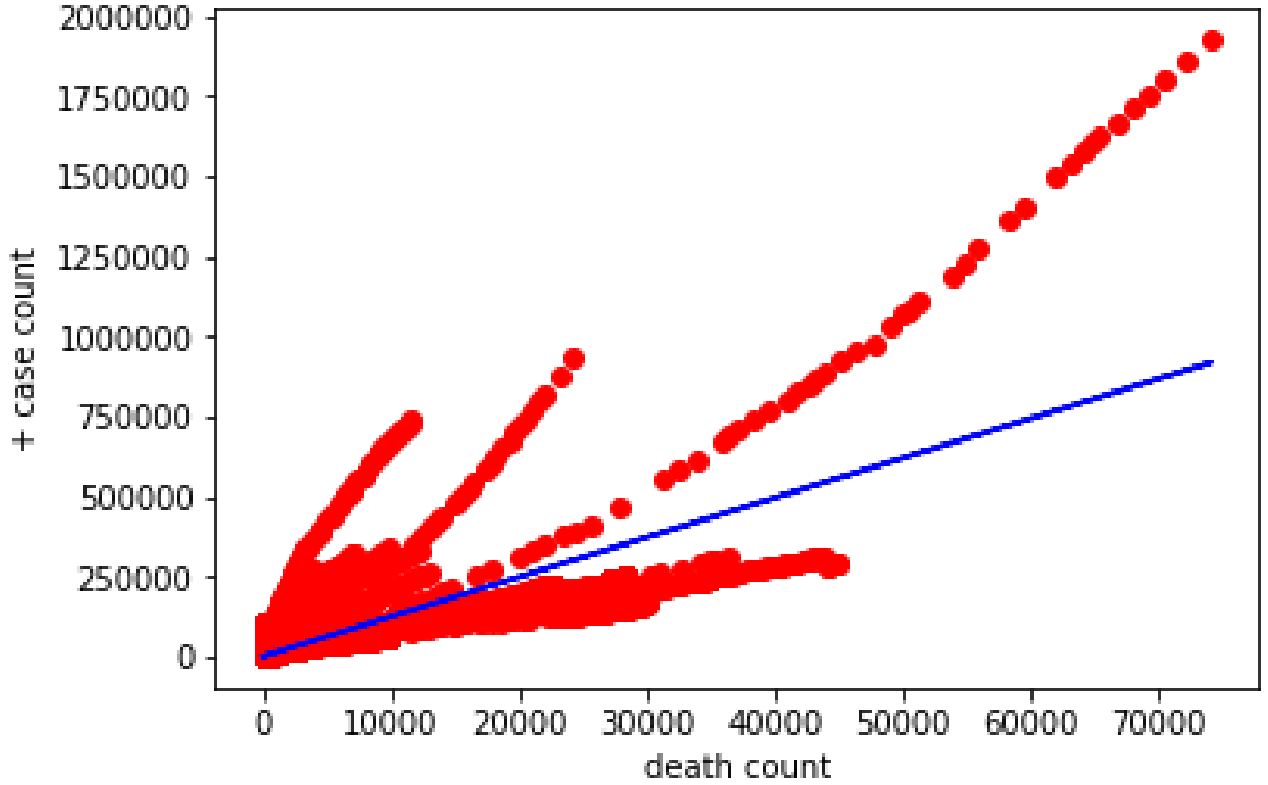


Fig 5. Fitting linear regression line on the training data of ‘COVID19_World.csv’.

Data Modeling-World COVID19

Running linear regression model on testing data.

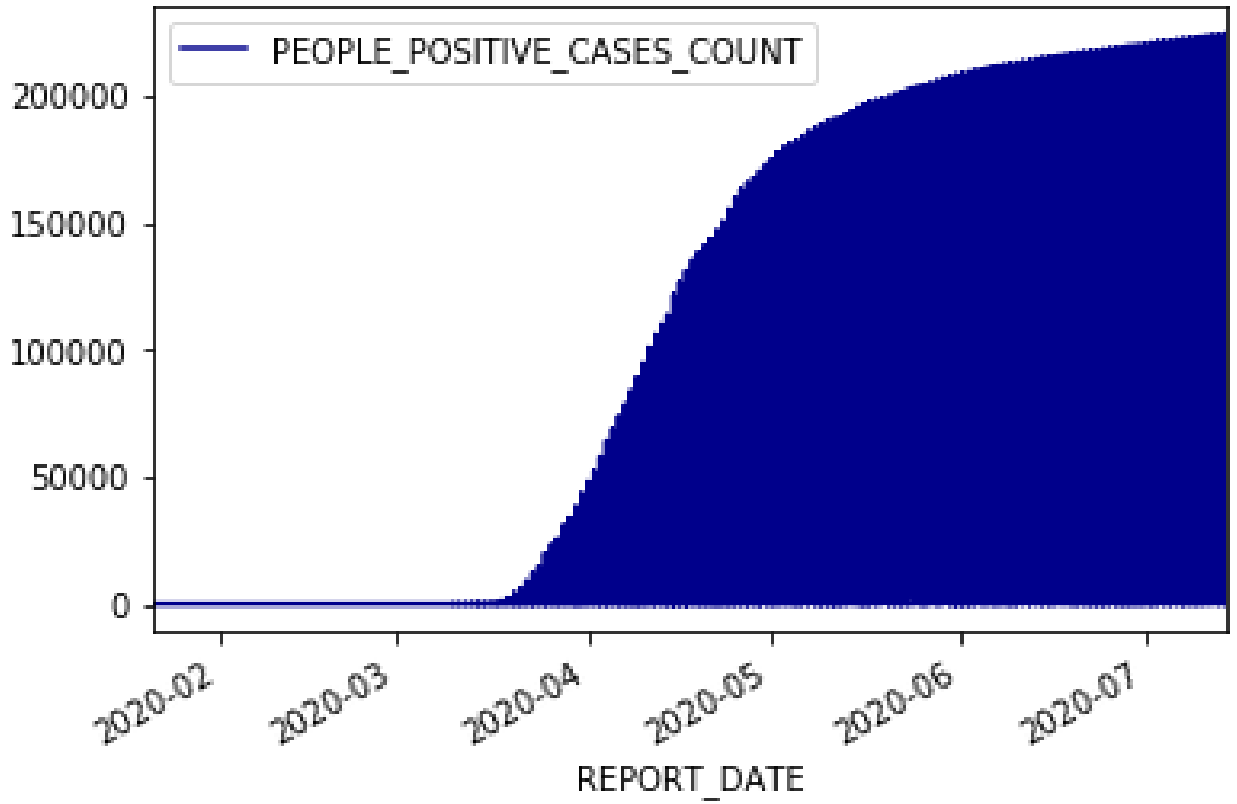
```
The mean absolute error: 812.19
The residual sum of squares (MSE): 89608618.52
The R2-score: 0.42
```

The statistical power of the linear regression model **is low with the R-squared (the coefficient of determination) of 0.42** (which is far from 1).

Next step will be to focus on using this linear regression model on just data from specific countries instead of the whole world e.g United States, Canada, Italy etc.

This is because, several factors (known and unknown) has led to different pattern which is why there is a burst like positive slope between the linear relationship between the death and total cases count variables.

Data Visualization-USA COVID19



- A new dataframe was created from the COVID19_World.csv to contain only data from USA.
- The total positive COVID19 case count trend in United States of America increased from January 2020 to July 2020.

Fig 6. Total positive COVID19 case in USA by date from wrangled data.

Data Visualization-USA COVID19

- The total positive COVID19 case count in New York State increased from January 2020 to July 2020 **at the highest rate.**

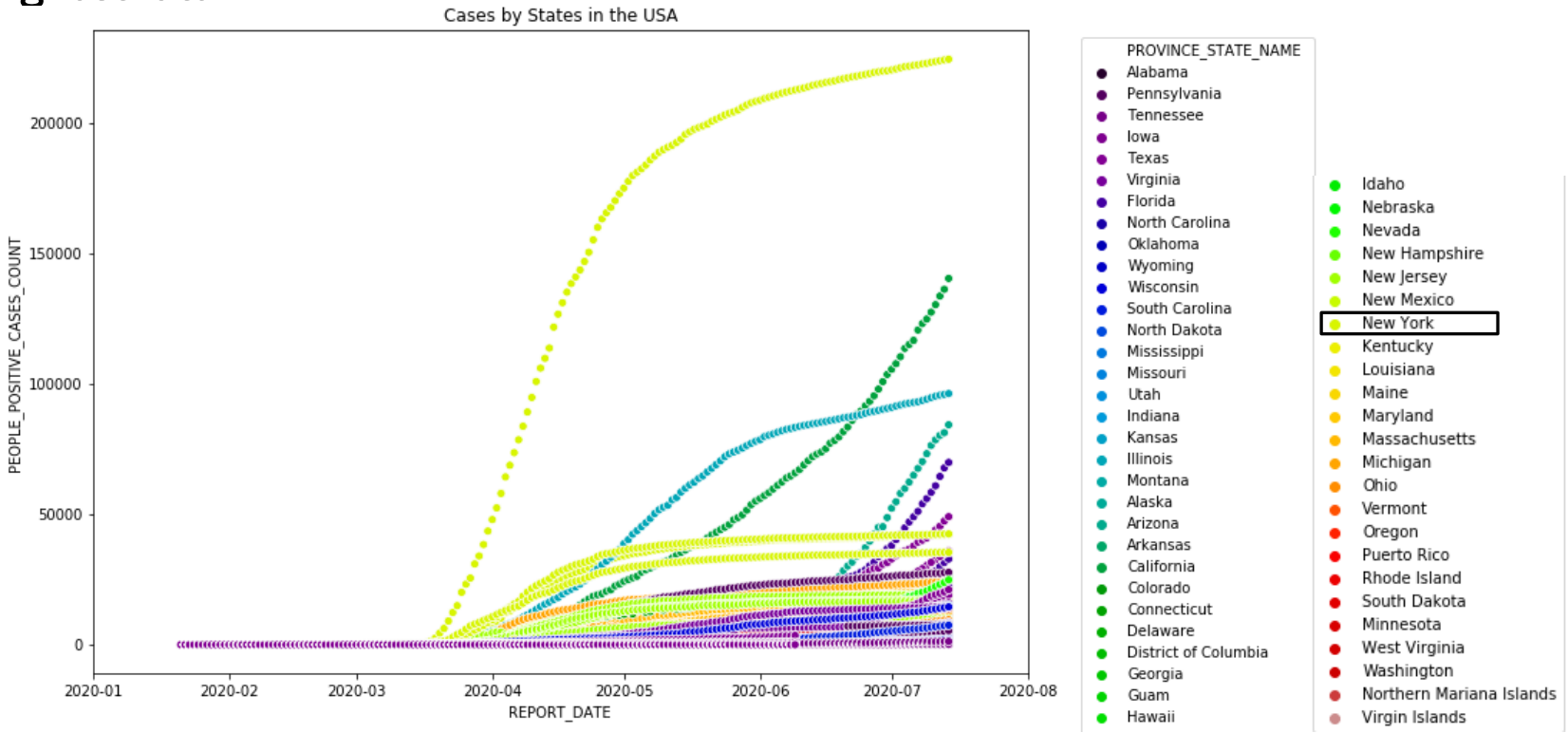


Fig 7. Total positive COVID19 case count by states in USA from wrangled 'COVID19_World.csv' data.

Data Visualization-USA COVID19

- There's a linear relationship between the variables, total positive COVID19 case count and death count.

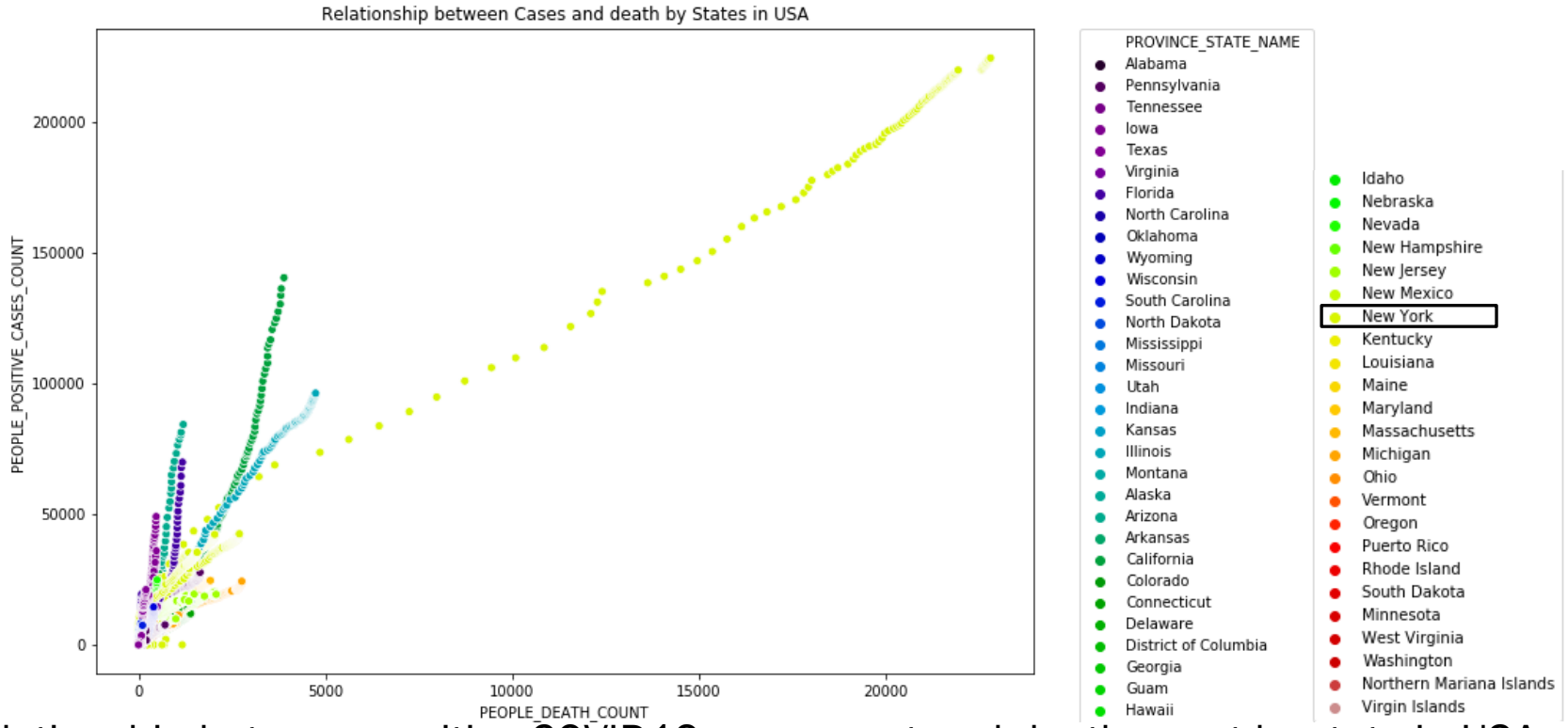


Fig 8. Relationship between positive COVID19 case count and death count by state in USA.

Data Modeling-USA COVID19

Splitting into training and testing data.

- The USA COVID19 data was split at the 80:20 ratio based on the Pareto principle.

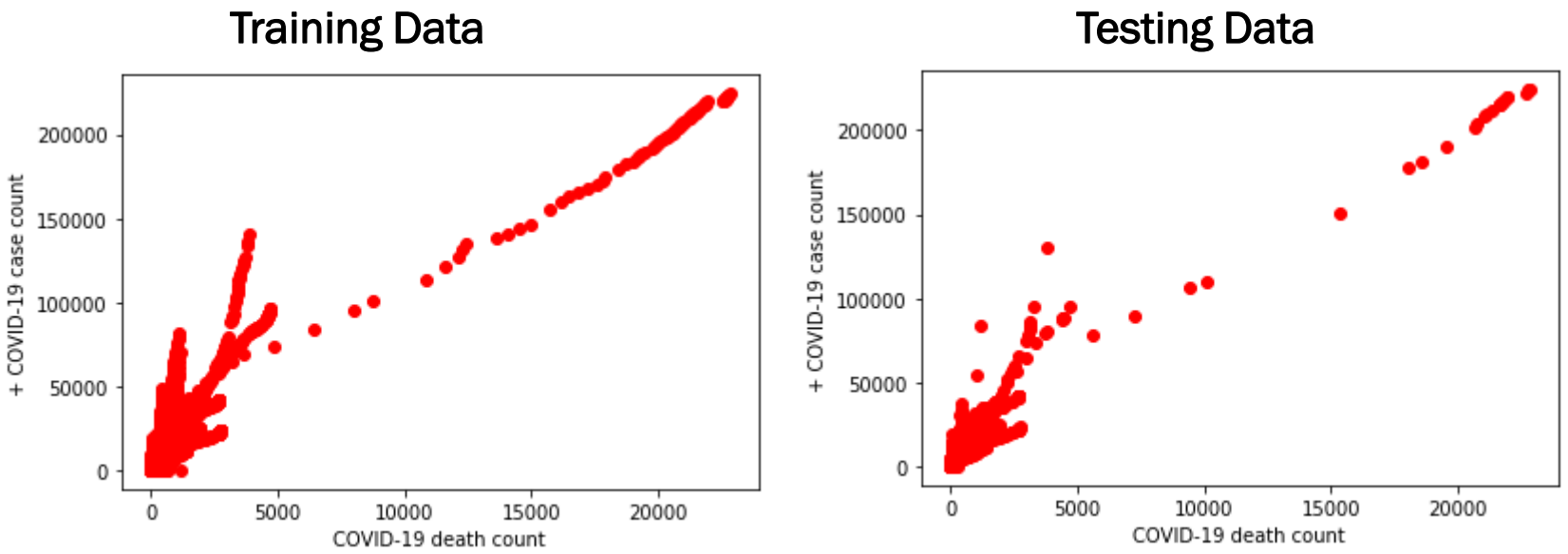


Fig 9. Training and testing data showing the relationship between positive COVID case count and death count by state in USA.

Data Modeling-USA COVID19

Running linear regression analysis on training data.

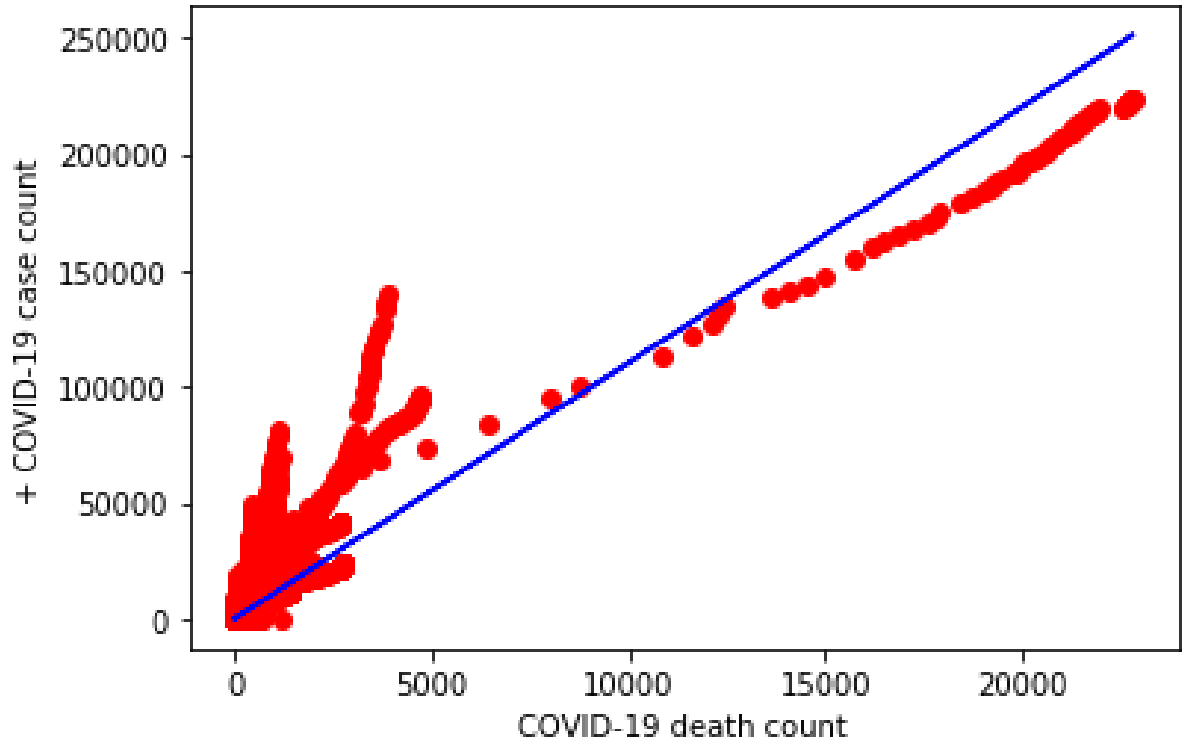


Fig 10. Fitting linear regression line on the USA COVID19 training data.

Data Modeling-USA COVID19

Running linear regression model on testing data.

```
The mean absolute error: 229.43
The residual sum of squares (MSE): 1130264.08
The R2-score: 0.90
```

The statistical power of the linear relationship between positive COVID-19 cases and COVID-19 deaths, the **R-squared, the coefficient of determination**, generated is **0.90.**

This indicates that **90% of the COVID-19 data in the USA** and **the positive (+) direction** of the slope line explains that there is a **positive correlation** between the number of COVID-19 death count and the number COVID-19 total case count.

Conclusion and discussion

- Considering the world COVID19 data to determine the linear relationship between the number of COVID-19 death count and the number COVID-19 total case count isn't the best approach. This is because there are different elements that influence this relationship in different countries and continents.

- In the USA, there are both unknown and known factors in the country that may have caused the number of COVID-19 total case count increases as the number of COVID-19 death count increases between January and July, 2020. This model doesn't address it.
- The linear model only shows a **positive correlation** between the number of COVID-19 death count and the number COVID-19 total case count.
- The next step will be applying this linear regression model to determine the relationship between the variables in different countries. The R^2 score can serve as a guage for risk of fatality in different levels such as state, province, or country level.

References

1. Novel coronavirus structure reveals targets for vaccines and treatments. *NIH Research Matters. Webpage, accessed Sept. 21, 2020. Link.*

2. Santarpia JL, et al. Aerosol and surface contamination of SARS-CoV-2 observed in quarantine and isolation care. Sci Rep. 2020 Jul 29;10(1):12732.