The Characteristics of Covid

Meng Chen
New York University
New York, United Stated
mc8451@nyu.edu

Xiao Ma
New York University
New York, United Stated
xm2074@nyu.edu

Yiran Ma
New York University
New York, United Stated
ym2360@nyu.edu

ABSTRACT

In this project, we explore the potential factors that may influence the number of deaths caused by the covid-19.

ACM Reference Format:
Meng Chen, Xiao Ma, and Yiran Ma. 2022. The Characteristics of Covid. In *Proceedings of ACM Conference (Conference'17).* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

At the end of 2019, the sudden outbreak of coronavirus quickly spread rapidly to most countries around the world. The covid seriously threatens people's life and hinders the development of the economy.

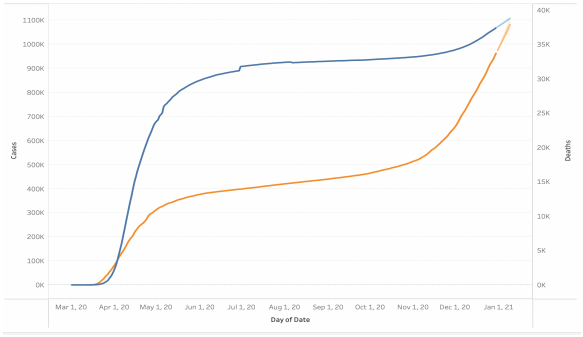
There are lots of cases that have happened around us. We even caught the covid-19 and also felt the damage of covid to the economy. The inflationary effects of price rises and some relatives even went bankrupt. All those situations inspire us to do some analysis about the covid. We want to study it and analyze the trend of covid and some relative reasons that affect the covid. Finally, we want to predict the trend of covid and hope to help control and fight the virus effectively.

2 METHODOLOGY

There are three modules in this project. We use Linear Regression Module and Multiple Linear Regression Module to calculate the relationships between cases, locations and deaths.

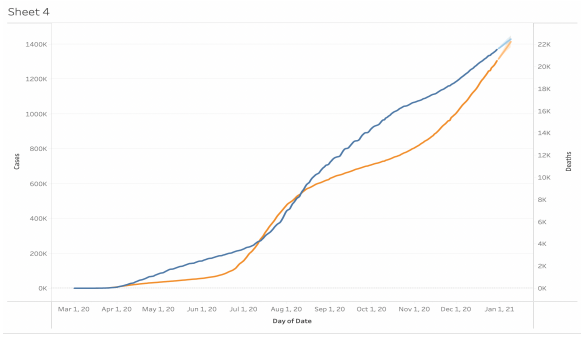
2.1 Regression Module

[1]For this part, we use data from New York Times to explore relations between cases, locations and deaths caused by the covid-19. At first, we need to be familar with the data and use Tableau[2] to visualize it.



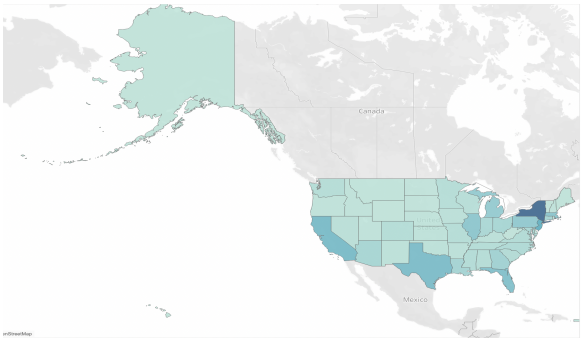
pic2.0 sum of cases and deaths in New York

Pic2 shows the sum of cases and deaths in New York in 2020. The sum of deaths and cases keep growing all the time.



pic2.1 sum of cases and deaths in Florida

Pic2.1 shows the sum of cases and deaths in Florida in 2020. It's similar with the pic2.0 but with different growth rate.

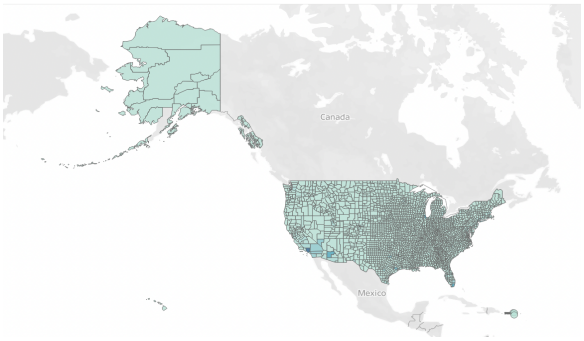


pic2.2 map of deaths through all statues

Conference'17, July 2017, Washington, DC, USA

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

Pic2.2 is the map that indicates how many deaths through all the states in the United States in 2020. We use tableau to generate the longitudes and latitudes of each status.



pic2.3 map of cases through all statues

Pic2.3 is similar with pic2.2. But it indicates the sum of cases of each states.

2.1.1 Linear Regression. We use Linear Regression[3] to explore the relation between cases and deaths in California. The input is the cases of the covid-19 of California in 2020 and the output the deaths caused by the covid-19 of California in 2020.

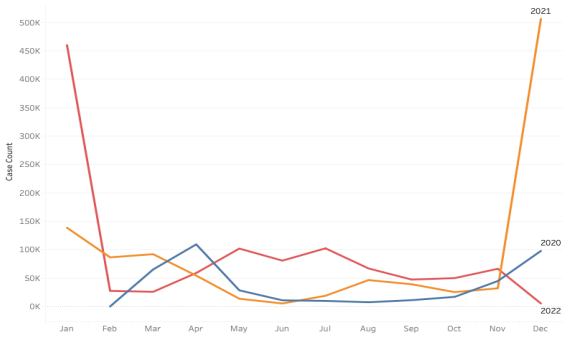
2.1.2 Multiple Linear Regression. We use Multiple Linear Regression[4] to explore the relation between cases, fips and deaths in California. The input is the cases of the covid-19 of California in 2020 and the output the the deaths caused by the covid-19 of California in 2020.

2.2 SVM Module

[5]For this part, we use data from New York Health data to find the relationship between the season or the temperature and the count of cases of coivd-19 through some visualizations. Also, we explore the relationship between the quantity of hospitalized and deaths, making a model with SVM and evaluating it.
Firstly, we use Tableau and python to visualize it. We show the change of deaths varying with months and the change of deaths for each season, including the total three years(2020-2022) data and the comparison of each year.

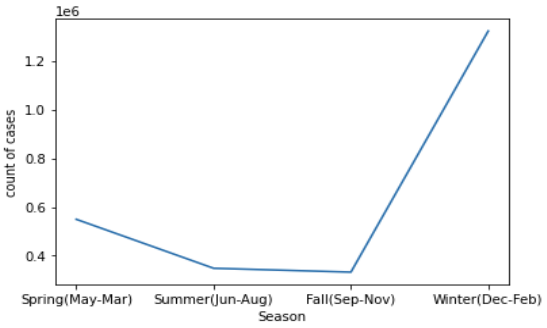
pic2.4 total sum of cases in 2020-2022 per month

Pic2.4 shows the total sum of cases in 2020-2022 for each month. It is obvious that in winter, that is December, January, and February, the sum of cases far exceeds other seasons.



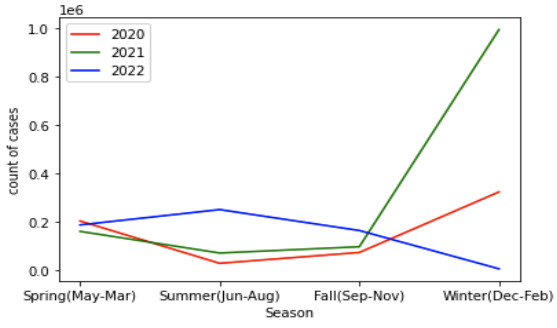
pic2.5 sum of cases per month for each year: 2020-2022

Pic2.5 shows the sum of cases per month per year. And the data for December 2022 is not complete. Apart from that and the start of covid-19 in 2020, the quantity in winter is obvious large compared with other months.



pic2.6 total sum of cases in 2020-2022 per season

Pic2.6 shows the total sum of cases in 2020-2022 per season. After group by season, the difference becomes more obvious.



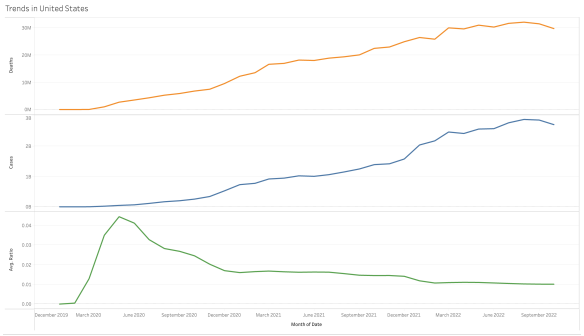
pic2.7 sum of cases per season for each year: 2020-2022

Pic2.7 shows the total sum of cases per season for there years 2020-2022. Comparing the three years' data, we also find that the trend of cases is decreasing.[6]

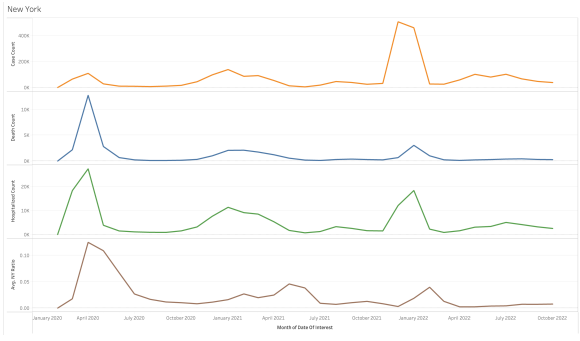
2.2.1 SVM. [7]We use SVM to explore the relationship between the count of hospitalized people and deaths in the USA. The input is the count of hospitalized people with covid-19 and the output is the total count of deaths. The data contains three-year data from February 2020 to December 2022.
In addition to improving the accuracy of the data, we use Grid Search to find the best parameters and use it to build the model.

2.3 Decision Tree Regressor

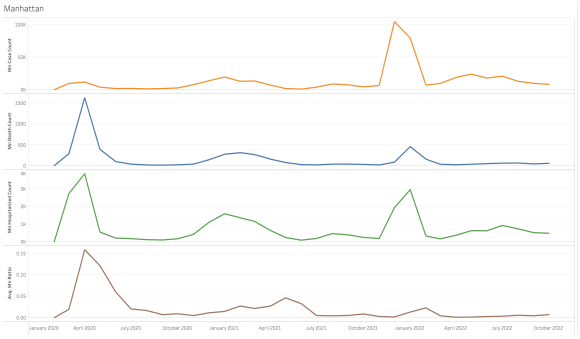
In this section, we mainly focus on the trends of the Covid-19. To investigate fully, we not only take the data from the whole United States, we also take the data from New York and Manhattan to see the Covid status around us. We will study the trends of cases, deaths, death ratio and hospitalized count based on the time from 2020 to 2022 and from New York to Manhattan.



pic2.8 shows the trends of cases, deaths, death ratio and hospitalized count in the United States from 2020-2022



pic2.8 shows the trends of cases, deaths, death ratio and hospitalized count in New York from 2020-2022



pic2.8 shows the trends of cases, deaths, death ratio and hospitalized count in the Manhattan from 2020-2022

2.3.1 Decision Tree Regressor. We use the tree regresion to see the detailed relation between cases, probable cases, hospitalized count and death count. The inputs here are the data for New York City from New York City Health Site, ranging from 2020 to 2022. The output is the exact death count give all the input data.

3 RESULTS

3.1 Regression Module

The results of the linear regression module are as follows:

OLS Regression Results

Dep. Variable:	deaths	R-squared:	0.915
Model:	OLS	Adj. R-squared:	0.915
Method:	Least Squares	F-statistic:	1.810e+05
Date:	Mon, 05 Dec 2022	Prob (F-statistic):	0.00
Time:	19:09:34	Log-Likelihood:	-1.1376e+05
No. Observations:	16792	AIC:	2.275e+05
Df Residuals:	16790	BIC:	2.275e+05
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-6.9520	1.694	-4.104	0.000	-10.273	-3.631
cases	0.0182	4.27e-05	425.400	0.000	0.018	0.018

Omnibus:	6414.391	Durbin-Watson:	1.810
Prob(Omnibus):	0.000	Jarque-Bera (JB):	3959305.675
Skew:	-0.301	Prob(JB):	0.00
Kurtosis:	78.223	Cond. No.	4.11e+04

pic3.1 results of the linear regression module

The R-squared is 0.915, which is closer to 1. It means 'cases' and 'deaths' have strong relations. The p valus is 0, which is less than 0.05. It indicates 'cases' are significant.

The results of the multiple linear regression module are as follows:

OLS Regression Results

Dep. Variable:	deaths	R-squared:	0.866
Model:	OLS	Adj. R-squared:	0.866
Method:	Least Squares	F-statistic:	7598.
Date:	Mon, 05 Dec 2022	Prob (F-statistic):	0.00
Time:	19:10:09	Log-Likelihood:	-14778.
No. Observations:	2354	AIC:	2.956e+04
Df Residuals:	2351	BIC:	2.958e+04
Df Model:	2		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	36.2341	7.166	5.057	0.000	22.182	50.286
cases	0.0155	0.000	105.104	0.000	0.015	0.016
fips	-0.0002	0.000	-1.350	0.177	-0.001	0.000

Omnibus:	321.106	Durbin-Watson:	2.334
Prob(Omnibus):	0.000	Jarque-Bera (JB):	2342.832
Skew:	0.416	Prob(JB):	0.00
Kurtosis:	7.816	Cond. No.	9.80e+04

pic3.2 results of the multiple linear regression module

The R-squared is 0.866, which is closer to 1. It means 'cases', 'fips' and 'deaths' have strong relations. The p value of 'cases' if 0, which indicates 'cases' has strong relation with 'deaths'. However, the p value of 'fips' is 0.177, which is larger than 0.05, so we do not reject the null hypothesis.

3.2 SVM Module

The results of the SVM module are as follows:

OLS Regression Results

Dep. Variable:	DEATH_COUNT	R-squared:	0.595
Model:	OLS	Adj. R-squared:	0.594
Method:	Least Squares	F-statistic:	1476.
Date:	Mon, 05 Dec 2022	Prob (F-statistic):	1.61e-199
Time:	17:59:04	Log-Likelihood:	-5380.0
No. Observations:	1008	AIC:	1.076e+04
Df Residuals:	1006	BIC:	1.077e+04
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-5.1881	1.925	-2.695	0.007	-8.966	-1.410
HOSPITALIZED_COUNT	0.2394	0.006	38.423	0.000	0.227	0.252

Omnibus:	511.790	Durbin-Watson:	0.074
Prob(Omnibus):	0.000	Jarque-Bera (JB):	9611.679
Skew:	1.878	Prob(JB):	0.00
Kurtosis:	17.654	Cond. No.	375.

pic3.4 OLS results of the SVM module

This picture shows the OLS evaluation result of the model. The R-squared value is 0.595, which is close to 1. That means there is a relatively strong relation between 'hospitalized count' and 'deaths'. Besides, the p-value is 0.007, which is small enough to show that 'hospitalized count' relates to 'deaths' strongly.

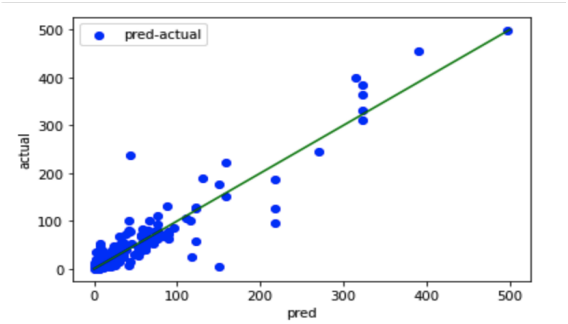
3.3 Decision Tree Regressor

Since predicting the accurate death count value is not that easy, we use a ranging method. For the predicting values, if it is within a certain range of the actual value, we will treat it to be right. That being said, we will allow a certain range of errors.

By setting the range of error to be 5, we get the results as follows:

- Accuracy: 0.5498281786941581
- MSE: 552.4707903780069
- RMSE: 23.504697198177364
- MAE: 11.219931271477662
- r2: 0.8825873378629611

By analyzing the prediction values and the actual values, we can draw pictures:

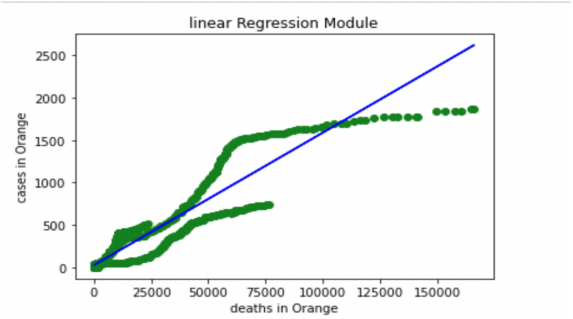


pic3.4 shows the relation between the predicted values and actual values

4 DISCUSSION

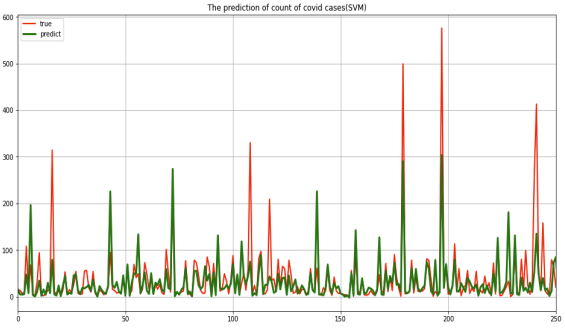
4.1 Linear Regression

For the linear regression module, since cases and deaths are not one-to-one, we can try different modules like multicollinearity modules to make it better(see pic4.1).



pic4.1 actual data and predicted data of the linear regression module

4.2 SVM



pic4.2 Real and predicted result on test data

From the picture, we can directly see the real and predicted result of the SVM model. From the model, we find one possible reason that results in lower accuracy: Due to a large number of cases in the whole country, prediction is too difficult to be exactly the same. So if we set a range to the prediction, which also shows the trend and the accuracy, this will improve the accuracy a lot.

5 CONCLUSIONS

Firstly, There are strong relationships between cases and deaths that are caused by the covid-19. The relation can be represented as a linear relation. Secondly, we still have the memory that the lower the temperature is, the fewer cases are, published by biological experts in the news. And from the visualization, we can also get this result. Besides, if there are more hospitalized people, the death will decrease. The number of hospitalized people relates to the deaths of people strongly.

6 REFERENCES

REFERENCES

[1] A. Nichols, "rd: Stata module for regression discontinuity estimation," 2016.
[2] I. Ko and H. Chang, "Interactive visualization of healthcare data using tableau," *Healthcare informatics research*, vol. 23, no. 4, pp. 349–354, 2017.
[3] S. Weisberg, *Applied linear regression*. John Wiley & Sons, 2005, vol. 528.
[4] L. E. Eberly, "Multiple linear regression," *Topics in Biostatistics*, pp. 165–187, 2007.
[5] A. Tobías and T. Molina, "Is temperature reducing the transmission of covid-19?" *Environmental research*, vol. 186, p. 109553, 2020.
[6] W. S. Byun, S. W. Heo, G. Jo, J. W. Kim, S. Kim, S. Lee, H. E. Park, and J.-H. Baek, "Is coronavirus disease (covid-19) seasonal? a critical analysis of empirical and epidemiological studies at global and local scales," *Environmental Research*, vol. 196, p. 110972, 2021.
[7] J. Gu, M. Zhu, and L. Jiang, "Housing price forecasting based on genetic algorithm and support vector machine," *Expert Systems with Applications*, vol. 38, no. 4, pp. 3383–3386, 2011.