Title: Reported National mortality due to respiratory illness.

Source: data.cdc.gov (https://data.cdc.gov/NCHS/Provisional-COVID-19-Death-Counts-by-Week-Ending-D/r8kw-7aab)

Information:

As we all know the year 2020 has been a year like no other, from everyone being grounded to everyone wearing masks. All in large part due to a horrible virus that wreaked havoc worldwide. In general, all respiratory illness are in need of immediate medical intervention. In some of the respiratory illness they can display some similar symptoms to each other making it difficult to figure out what's going on. In this particular data I grew some interest in reviewing the data largely because I wanted to see if there are patterns, shifts, trends, or occurrence that I can link between these mortality types, states, and time frame where they are occurring and any other interesting findings I can come across.

Hypothesis: States that have higher numbers of deaths due to influenza also have the highest number of deaths due to Covid19.

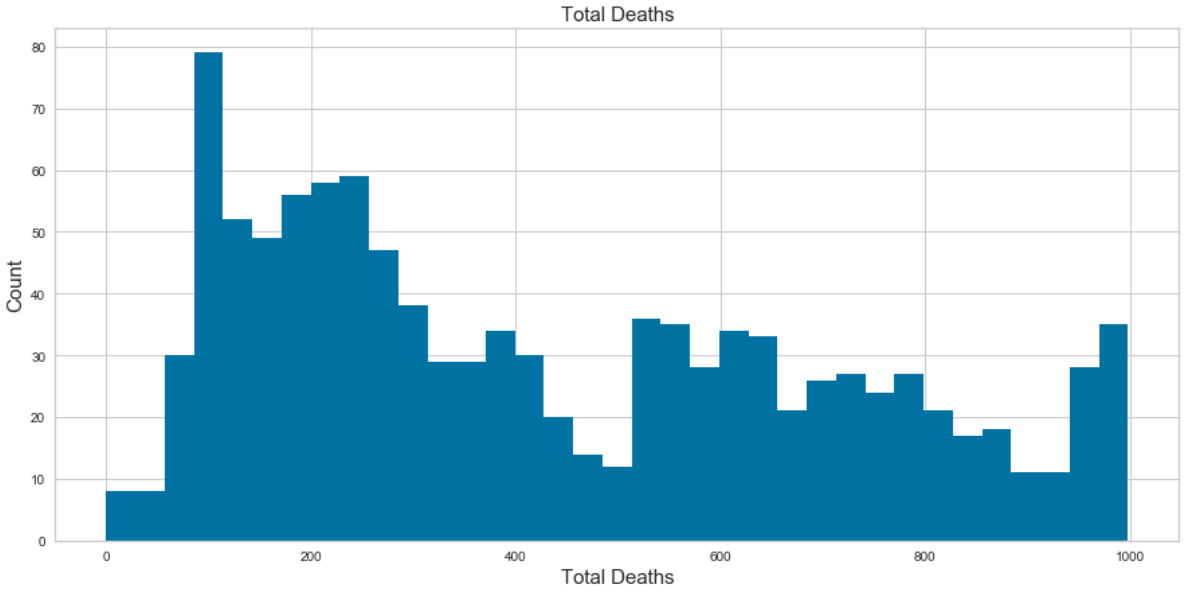
Features in the data:

1) Data as of
2) Start week
3) End week
4) Group
5) State
6) Indicator
7) Covid 19 deaths
8) Total Deaths
9) Percent of expected deaths
10) Pneumonia deaths
11) Pneumonia and covid deaths
12) Influenza deaths
13) Pneumonia, influenza, or covid 19 deaths
14) Footnotes

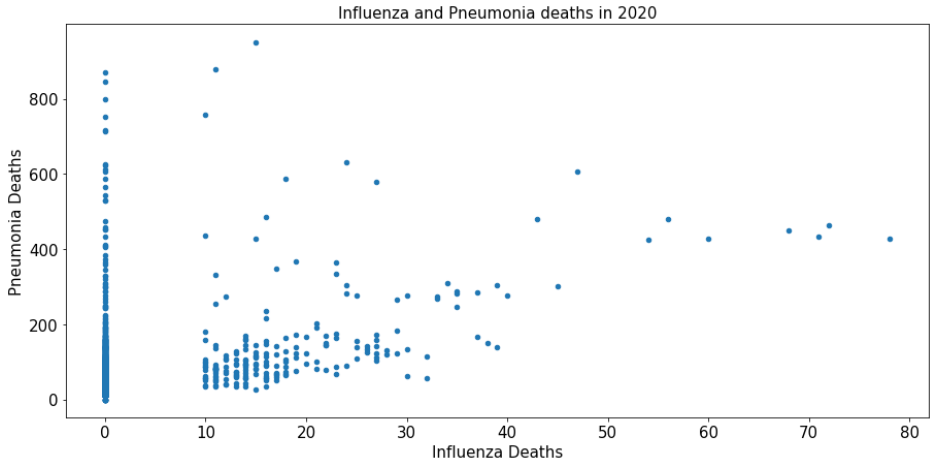
Step by step:

1) Load data
2) Check dimensions of data
3) Scan through data for glaring problems
4) Get descriptive and summary information on data
5) Get necessary cleaning and modifications done on the data
6) Generate and review different types of graphs for insight

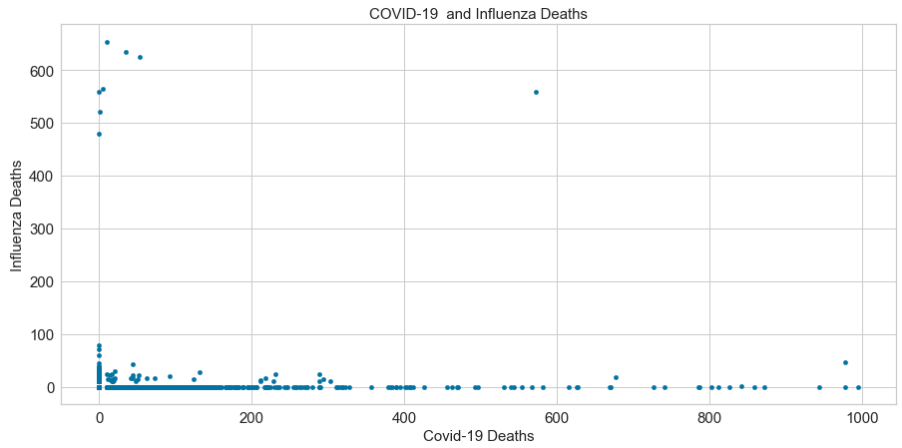
- Dimensions of data are: (1944, 14)



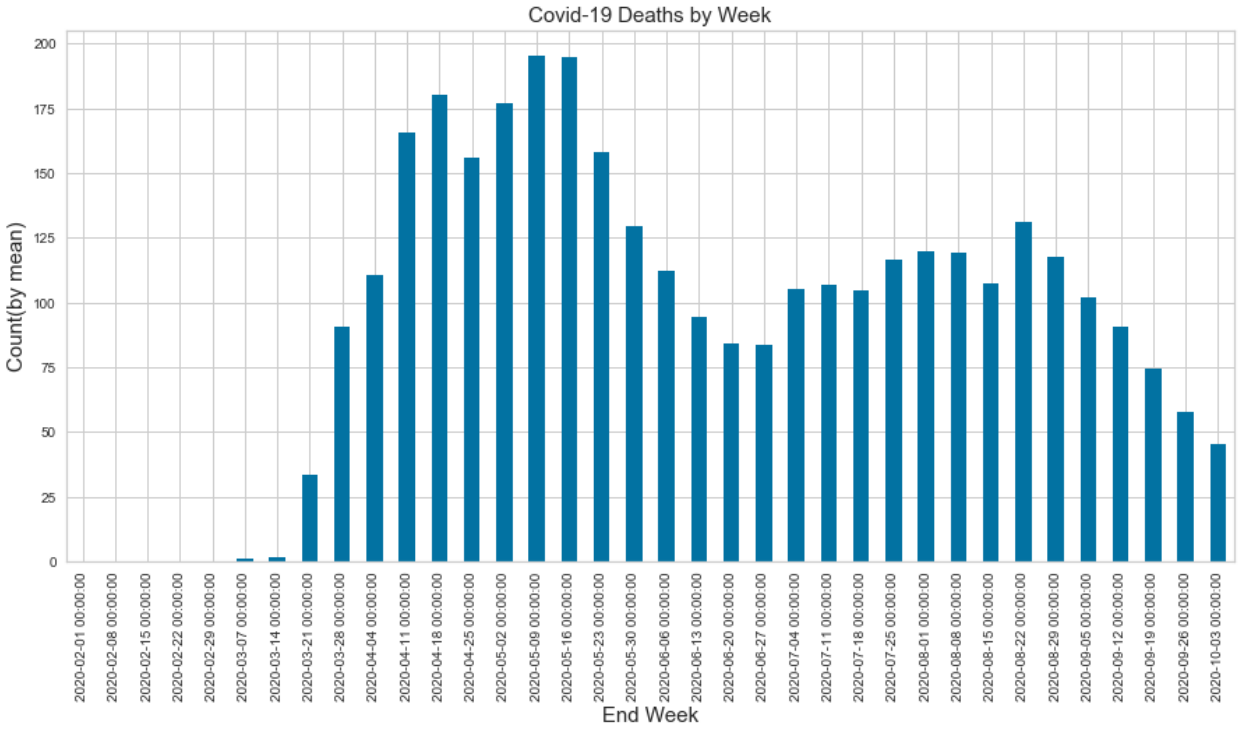
- Distribution of total deaths is not a uniform distribution.



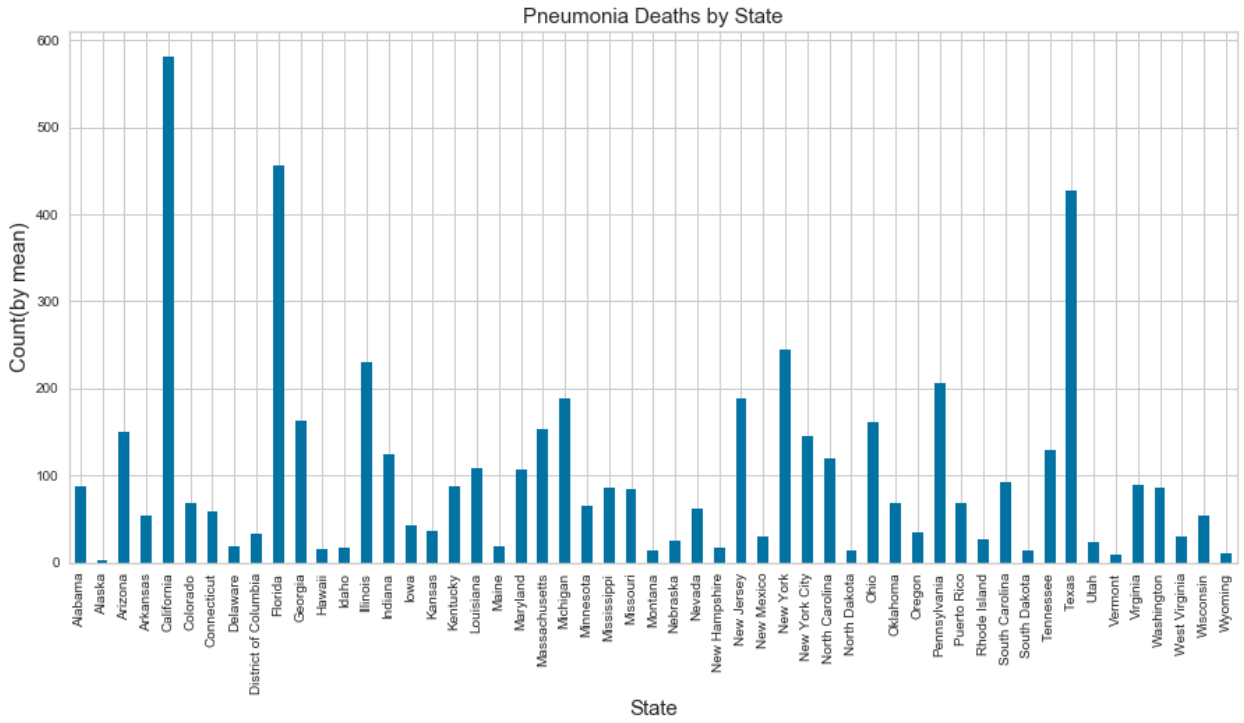
- Appears to be somewhat of a relationship between pneumonia and flu deaths.



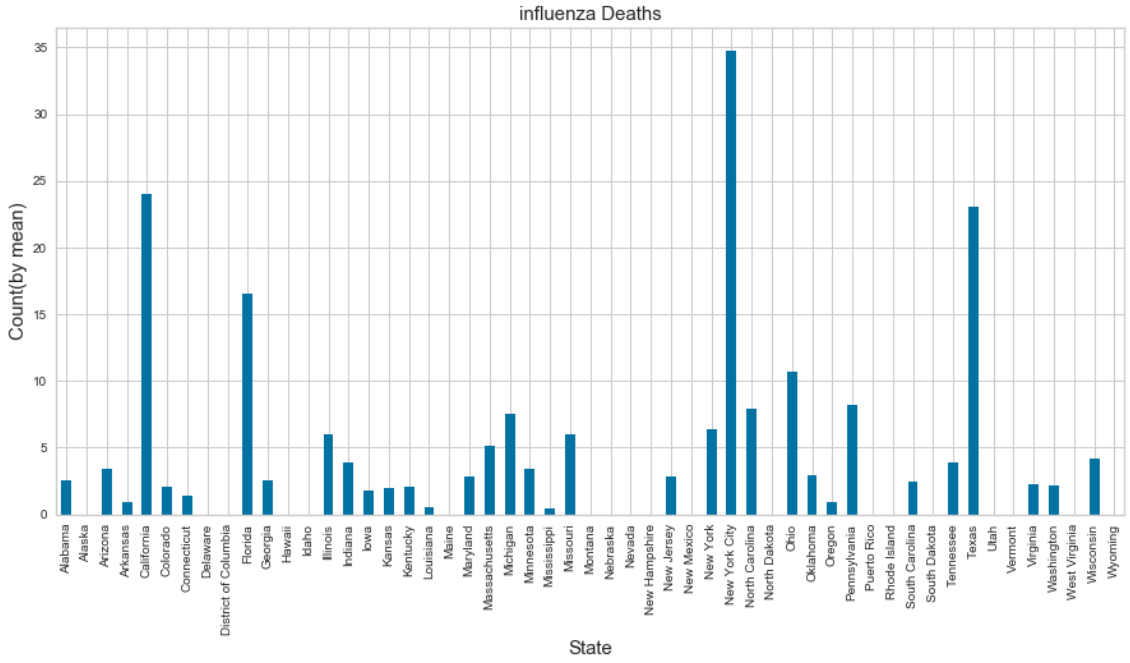
- There appears to be not much of relationship between these variables.



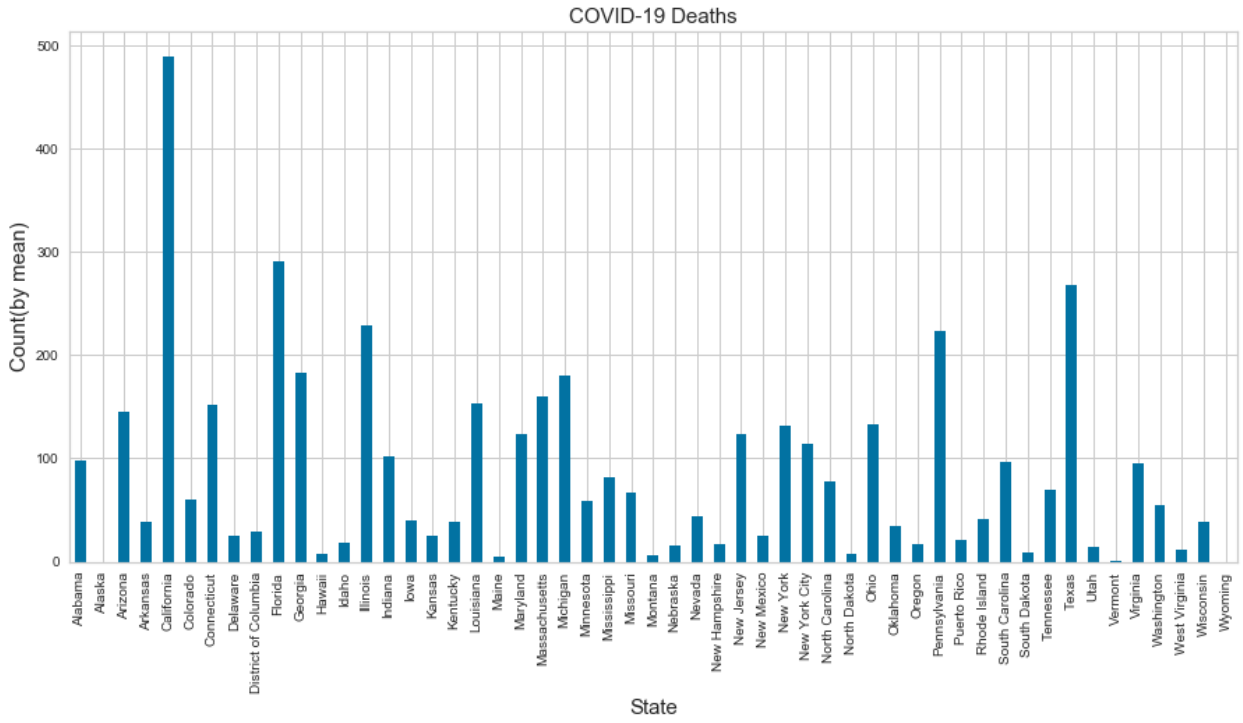
- Numbers of Covid-19 deaths appear to be decreasing.



- Appears California leads the pack with most reported pneumonia death cases.



- Appears NYC leads the groups in most flu deaths.



- Appears California leads the pack in Covid-19 deaths.

Part II: Original Analysis Case Study

Step by step (Dimension and feature reduction):

1) View data

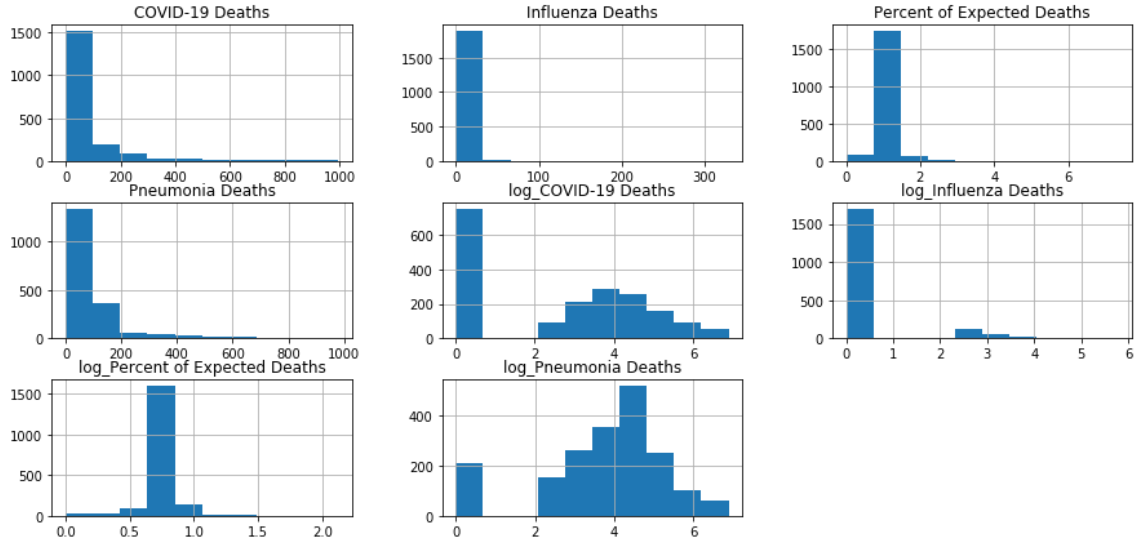
	Data as of	Start week	End Week	Group	State	Indicator	COVID-19 Deaths	Total Deaths	Percent of Expected Deaths	Pneumonia Deaths	Pneumonia and COVID-19 Deaths	Influenza Deaths	Pneumonia, Influenza, or COVID-19 Deaths	Footnote
1939	10/9/2020	9/5/2020	2020-09-05	By week	Puerto Rico	Week-ending	42.0	435.0	0.82	68.0	28.0	NaN	82.0	One or more data cells have counts between 1–9...
1940	10/9/2020	9/12/2020	2020-09-12	By week	Puerto Rico	Week-ending	65.0	358.0	0.64	69.0	44.0	0.0	90.0	NaN
1941	10/9/2020	9/19/2020	2020-09-19	By week	Puerto Rico	Week-ending	44.0	181.0	0.31	41.0	28.0	0.0	57.0	NaN
1942	10/9/2020	9/26/2020	2020-09-26	By week	Puerto Rico	Week-ending	35.0	85.0	0.13	32.0	28.0	0.0	39.0	NaN
1943	10/9/2020	10/3/2020	2020-10-03	By week	Puerto Rico	Week-ending	25.0	48.0	0.08	23.0	19.0	0.0	29.0	NaN

2) Needed to find out what to do with null values. Reviewed the data and assessed the goal I was trying to achieve, I decided to insert zeros for all null values.
3) Now I faced a dilemma, dealing with only 14 features I had to make a decision based on my goal and data on what I was going to do in regards to dimensionality (reduction or addition?)
4) Started by reducing some nice features that were not of value for this analysis and ended up with only 5 features.

	Pneumonia Deaths	Influenza Deaths	COVID-19 Deaths	Percent of Expected Deaths	State
36	56.0	14.0	0.0	0.94	Alabama
37	61.0	10.0	0.0	1.01	Alabama
38	76.0	0.0	0.0	1.01	Alabama
39	68.0	0.0	0.0	1.01	Alabama
40	63.0	14.0	0.0	1.12	Alabama
...	...	...	...	...	...
1939	68.0	0.0	42.0	0.82	Puerto Rico
1940	69.0	0.0	65.0	0.64	Puerto Rico
1941	41.0	0.0	44.0	0.31	Puerto Rico
1942	32.0	0.0	35.0	0.13	Puerto Rico
1943	23.0	0.0	25.0	0.08	Puerto Rico

1908 rows × 5 columns

5) Noticed some really skewed data and small values, and the best course of action to take in this situation was to apply a feature transformation to this data. I applied log transformation (log(1+x).)



6) Had one nominal categorical feature representing 50 states and territories. Decided to take the 50 states and territory, break them down into 5 region categories (West, South, Midwest, Northeast, US territory).
7) Transformed categorical data into numerical data using one hot encoding (n-1). Data is now containing 9 total features and is ready for model evaluation.

	log_Influenza Deaths	log_Pneumonia Deaths	log_COVID-19 Deaths	log_Percent of Expected Deaths	Region_Midwest	Region_North-East	Region_South	Region_US Territory	Region_West
36	2.708050	4.043051	0.0	0.662688	0	0	1	0	0
37	2.397895	4.127134	0.0	0.698135	0	0	1	0	0
38	0.000000	4.343805	0.0	0.698135	0	0	1	0	0
39	0.000000	4.234107	0.0	0.698135	0	0	1	0	0
40	2.708050	4.158883	0.0	0.751416	0	0	1	0	0
41	2.639057	4.077537	0.0	0.698135	0	0	1	0	0
42	0.000000	4.127134	0.0	0.693147	0	0	1	0	0

	log_Influenza Deaths	log_Pneumonia Deaths	log_COVID-19 Deaths	log_Percent of Expected Deaths	Region_Midwest	Region_North-East	Region_South	Region_US Territory	Region_West
log_Influenza Deaths	0.889075	0.298344	-0.606772	0.010387	0.008384	0.005863	-0.000174	-0.006114	-0.015387
log_Pneumonia Deaths	0.298344	2.599108	2.094220	0.063584	-0.062616	-0.056142	0.155068	0.008443	-0.041196
log_COVID-19 Deaths	-0.606772	2.094220	4.930244	0.083346	-0.019578	-0.082385	0.146501	-0.014149	-0.028063
log_Percent of Expected Deaths	0.010387	0.063584	0.083346	0.019250	-0.003001	0.001499	-0.000319	-0.001448	0.000316
Region_Midwest	0.008384	-0.062616	-0.019578	-0.003001	0.194477	-0.044879	-0.084772	-0.004987	-0.044879
Region_North-East	0.005863	-0.056142	-0.082385	0.001499	-0.044879	0.141049	-0.054496	-0.003206	-0.028851
Region_South	-0.000174	0.155068	0.146501	-0.000319	-0.084772	-0.054496	0.217985	-0.006055	-0.054496
Region_US Territory	-0.006114	0.008443	-0.014149	-0.001448	-0.004987	-0.003206	-0.006055	0.018522	-0.003206
Region_West	-0.015387	-0.041196	-0.028063	0.000316	-0.044879	-0.028851	-0.054496	-0.003206	0.141049

- **Covariance of data**

	log_Influenza Deaths	log_Pneumonia Deaths	log_COVID-19 Deaths	log_Percent of Expected Deaths	Region_Midwest	Region_North-East	Region_South	Region_US Territory	Region_West
log_Influenza Deaths	1.000000	0.196262	-0.289816	0.079399	0.020163	0.016556	-0.000396	-0.047648	-0.043451
log_Pneumonia Deaths	0.196262	1.000000	0.585027	0.284266	-0.088072	-0.092723	0.206014	0.038479	-0.068039
log_COVID-19 Deaths	-0.289816	0.585027	1.000000	0.270546	-0.019994	-0.098793	0.141316	-0.046823	-0.033652
log_Percent of Expected Deaths	0.079399	0.284266	0.270546	1.000000	-0.049048	0.028765	-0.004924	-0.076691	0.006062
Region_Midwest	0.020163	-0.088072	-0.019994	-0.049048	1.000000	-0.270973	-0.411723	-0.083086	-0.270973
Region_North-East	0.016556	-0.092723	-0.098793	0.028765	-0.270973	1.000000	-0.310791	-0.062718	-0.204545
Region_South	-0.000396	0.206014	0.141316	-0.004924	-0.411723	-0.310791	1.000000	-0.095295	-0.310791
Region_US Territory	-0.047648	0.038479	-0.046823	-0.076691	-0.083086	-0.062718	-0.095295	1.000000	-0.062718
Region_West	-0.043451	-0.068039	-0.033652	0.006062	-0.270973	-0.204545	-0.310791	-0.062718	1.000000

- **Pearson Correlation**

	log_Influenza Deaths	log_Pneumonia Deaths	log_COVID-19 Deaths	log_Percent of Expected Deaths	Region_Midwest	Region_North-East	Region_South	Region_US Territory	Region_West
log_Influenza Deaths	1.000000	0.248860	-0.276031	-0.159468	0.024773	0.023060	0.003567	-0.048400	-0.050272
log_Pneumonia Deaths	0.248860	1.000000	0.660392	0.356710	-0.095345	-0.074147	0.241616	0.021205	-0.107274
log_COVID-19 Deaths	-0.276031	0.660392	1.000000	0.542788	-0.026115	-0.096977	0.166621	-0.058185	-0.049676
log_Percent of Expected Deaths	-0.159468	0.356710	0.542788	1.000000	-0.048647	-0.083629	0.112968	-0.085270	0.036266
Region_Midwest	0.024773	-0.095345	-0.026115	-0.048647	1.000000	-0.270973	-0.411723	-0.083086	-0.270973
Region_North-East	0.023060	-0.074147	-0.096977	-0.083629	-0.270973	1.000000	-0.310791	-0.062718	-0.204545
Region_South	0.003567	0.241616	0.166621	0.112968	-0.411723	-0.310791	1.000000	-0.095295	-0.310791
Region_US Territory	-0.048400	0.021205	-0.058185	-0.085270	-0.083086	-0.062718	-0.095295	1.000000	-0.062718
Region_West	-0.050272	-0.107274	-0.049676	0.036266	-0.270973	-0.204545	-0.310791	-0.062718	1.000000

- **Spearman Correlation**

part III: Original case study

Step 1: split data into training and testing groups (80/20)

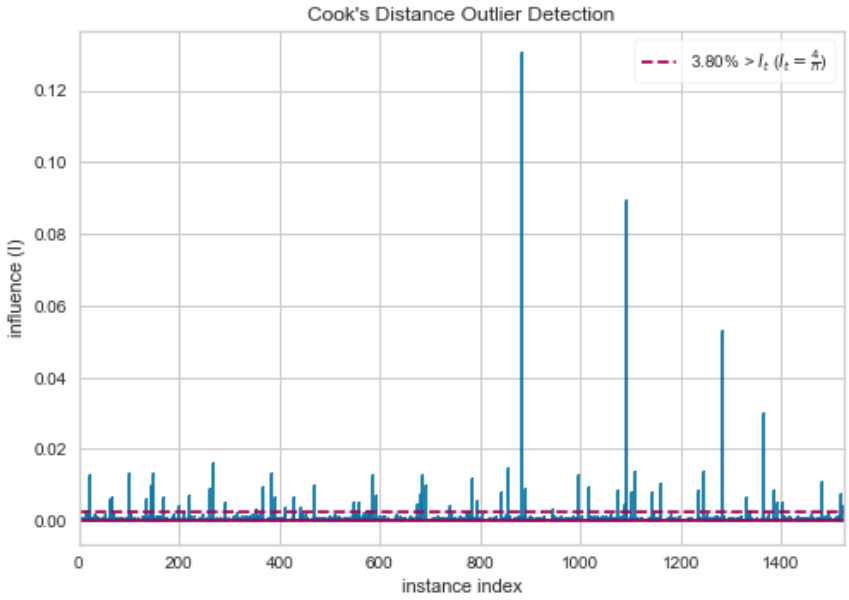
Step 2: Evaluation: in this portion I attempted to apply a regression model using multiple independent variables against cov19 mortalities. I ended up going with support vector regression as this gave me

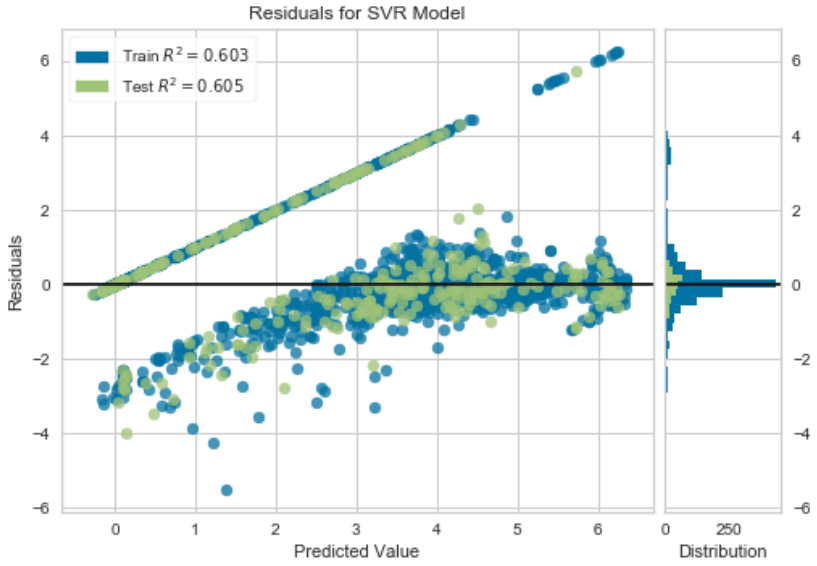
more flexibility with the data I was working with and the goal I am trying to reach, and also, I just found it really interesting.

Findings:

1) Matrices- confidence score of the model was approximately 60.5%. mean squared error value of 2.056 and a root mean square value of 1.43.

Cooks distance (outlier detection)





Residual plot (displaying q-q plot, distribution and residuals of model)

Prediction error (targets from dataset against predicted values)