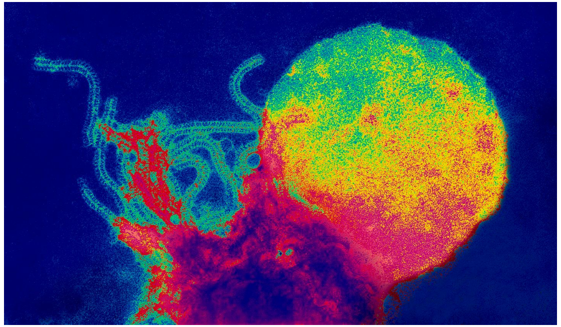
Preventable Diseases

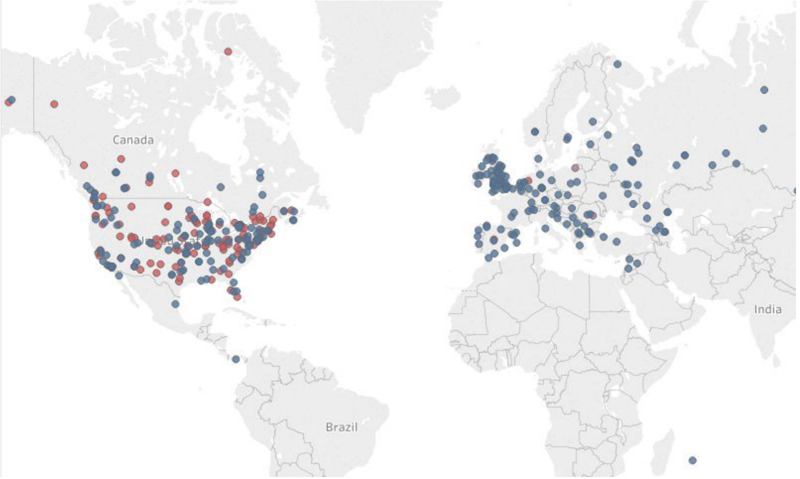
By: Javier Gonzalez Compte

What is it about?



Measles Virus

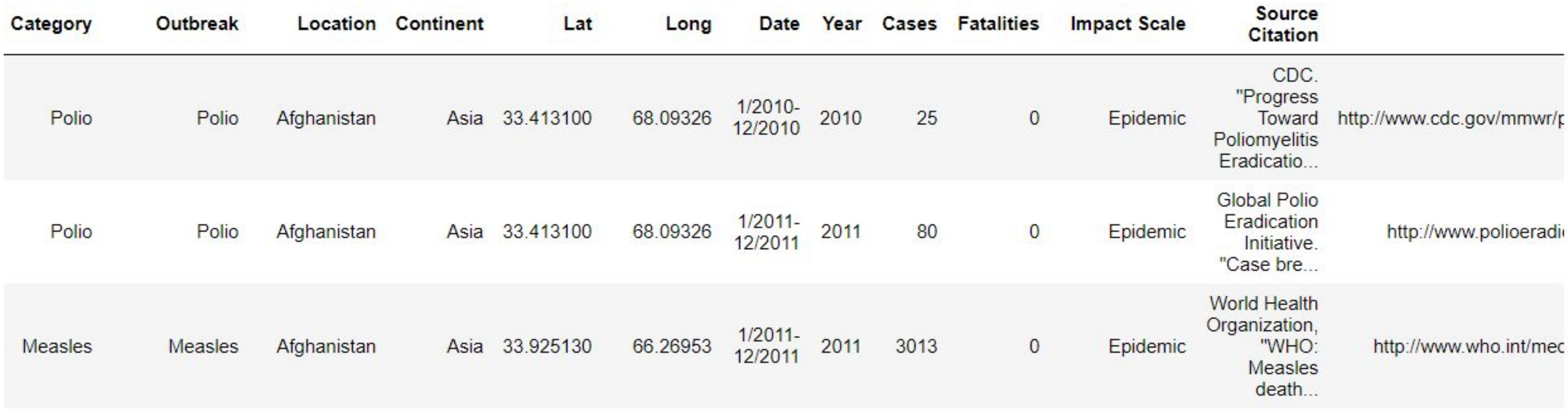
- Preventable Diseases that are curable such as Polio, Measles, Mumps, Whooping Cough
- Get Hotspots
- Predict Type of Disease
- Predict Cases



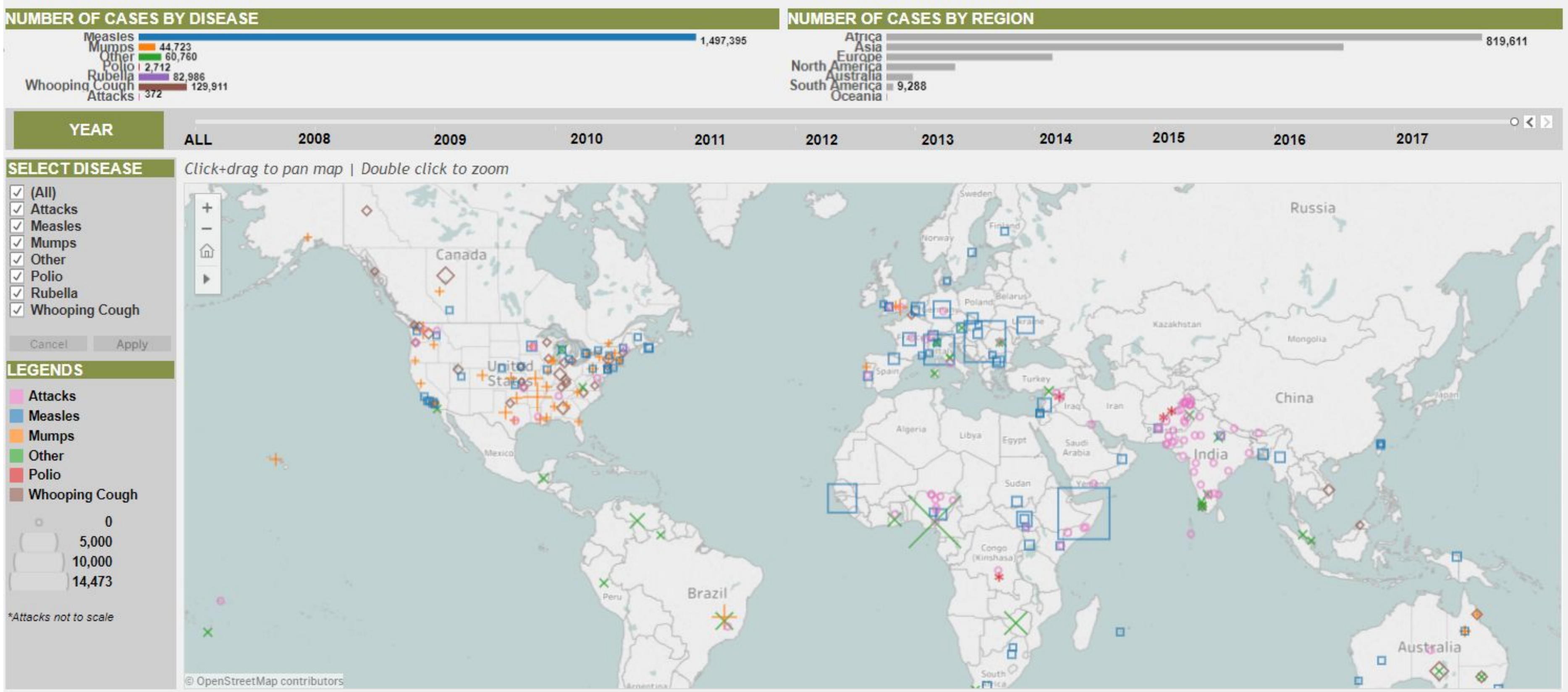
Made in Tableau

Data

- The Global Health Program at the Council on Foreign Relations has been tracking reports by news media, governments, and the global health community on these outbreaks since the fall of 2008.



Category	Outbreak	Location	Continent	Lat	Long	Date	Year	Cases	Fatalities	Impact Scale	Source Citation	
Polio	Polio	Afghanistan	Asia	33.413100	68.09326	1/2010-12/2010	2010	25	0	Epidemic	CDC. "Progress Toward Poliomyelitis Eradicatio...	http://www.cdc.gov/mmwr/p
Polio	Polio	Afghanistan	Asia	33.413100	68.09326	1/2011-12/2011	2011	80	0	Epidemic	Global Polio Eradication Initiative. "Case bre...	http://www.polioeradi
Measles	Measles	Afghanistan	Asia	33.925130	66.26953	1/2011-12/2011	2011	3013	0	Epidemic	World Health Organization, "WHO: Measles death...	http://www.who.int/mec



NUMBER OF CASES BY DISEASE

Measles	1,497,395
Mumps	44,723
Other	60,760
Polio	2,712
Rubella	82,986
Whooping Cough	129,911
Attacks	372

NUMBER OF CASES BY REGION

Africa	819,611
Asia	
Europe	
North America	
Australia	
South America	9,288
Oceania	

YEAR

ALL 2008 2009 2010 2011 2012 2013 2014 2015 2016 2017

SELECT DISEASE

- ☑ (All)
- ☑ Attacks
- ☑ Measles
- ☑ Mumps
- ☑ Other
- ☑ Polio
- ☑ Rubella
- ☑ Whooping Cough

Cancel Apply

LEGENDS

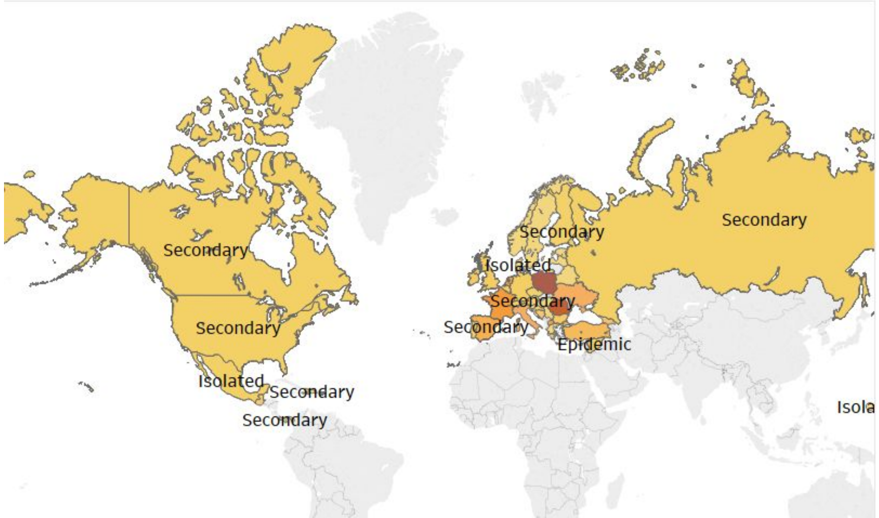
- Attacks
- Measles
- Mumps
- Other
- Polio
- Whooping Cough

○	0
▢	5,000
▢	10,000
▢	14,473

*Attacks not to scale

Click+drag to pan map | Double click to zoom

What we learn from the Data



Made in Tableau

1587 Data entries with 13 categorizations

Europe & North America: 721 entries

Africa : 275 entries

Asia : 436 entries

Australia: 120 entries

Focus : North America and Europe

Outbreaks(Diseases):

Measles : 186 for Europe and 181 for North America

Whooping Cough: 7 for Europe and 160 for North America:

Impact Scale:

Epidemic 122 for Europe and 95 for North America

Cluster: 97 for Europe and 179 for North America

Fatalities:

North America had an average of 14% Europe had a 20%

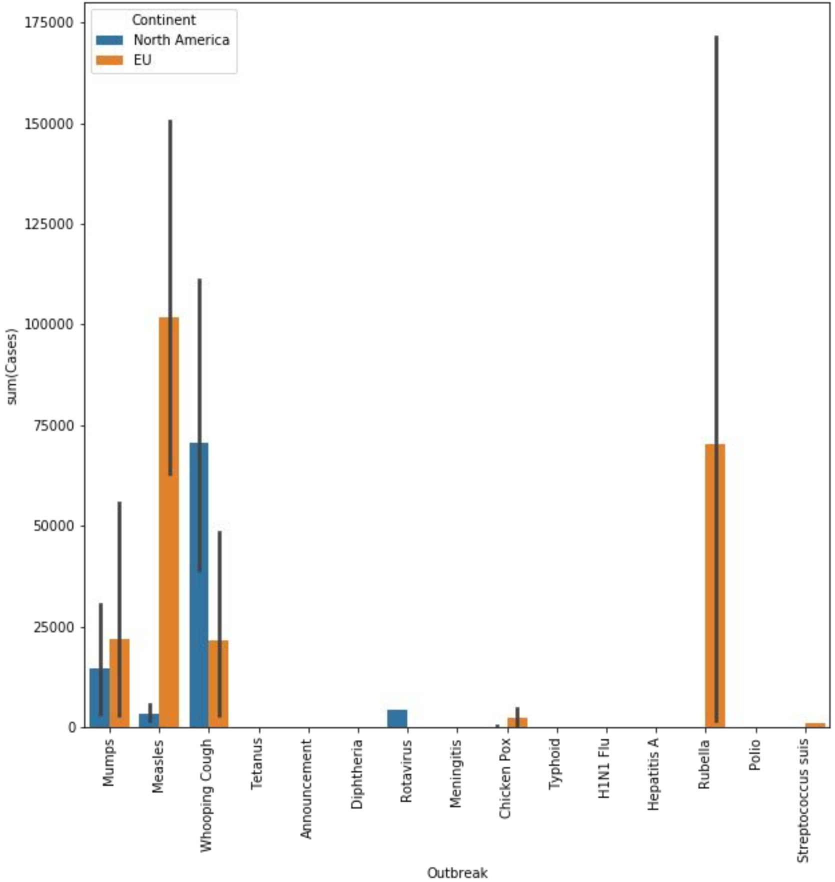
Distinct Occ.: Europe & NA 687

Cases:

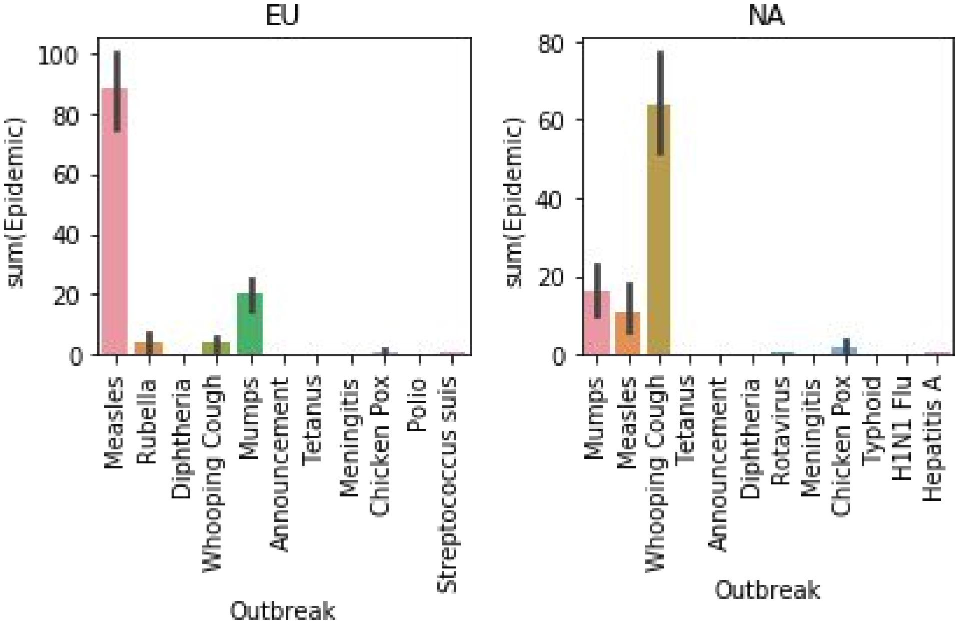
Europe : 226,456

North America: 93,810

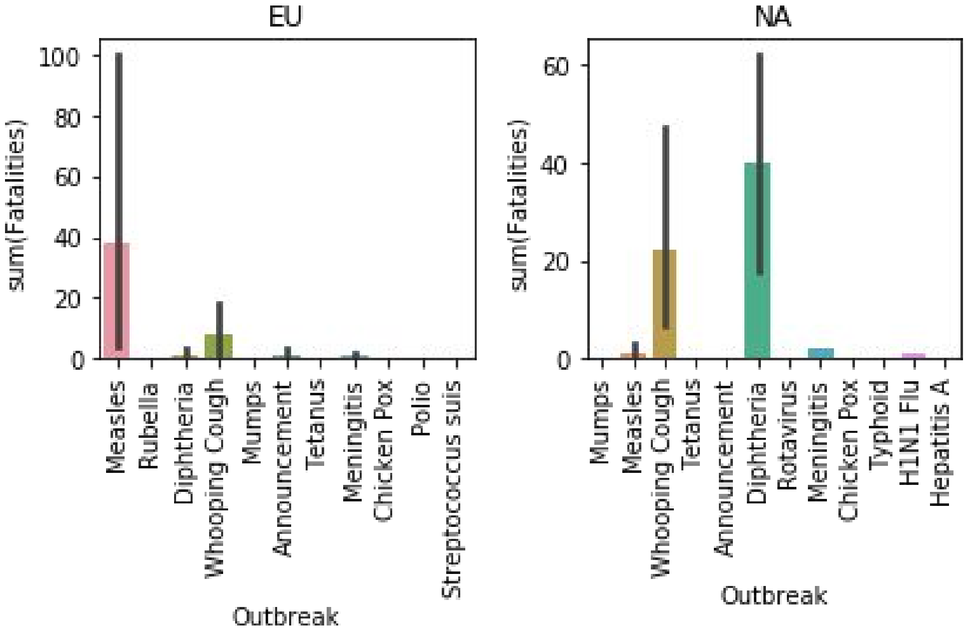
Cases in North America and Europe



Epidemics in North America and Europe



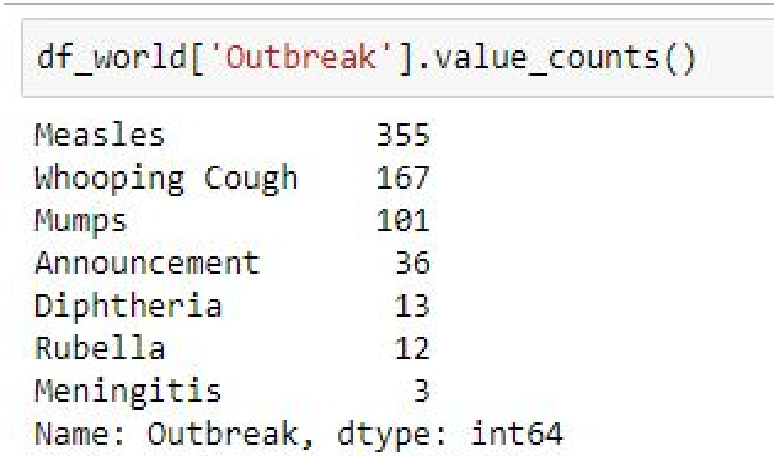
Most Fatal Diseases



Mislabeled Data

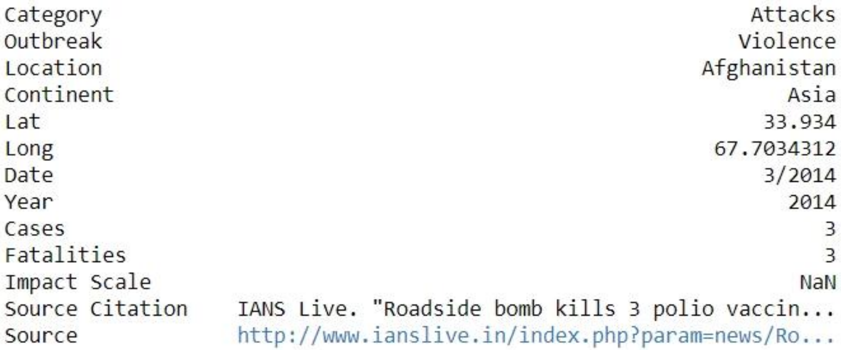
Lots of Mislabeled data:

- Dropped Violence since it affected Fatalities
- Dropped Diseases with one occurrence



```
df_world['Outbreak'].value_counts()
```

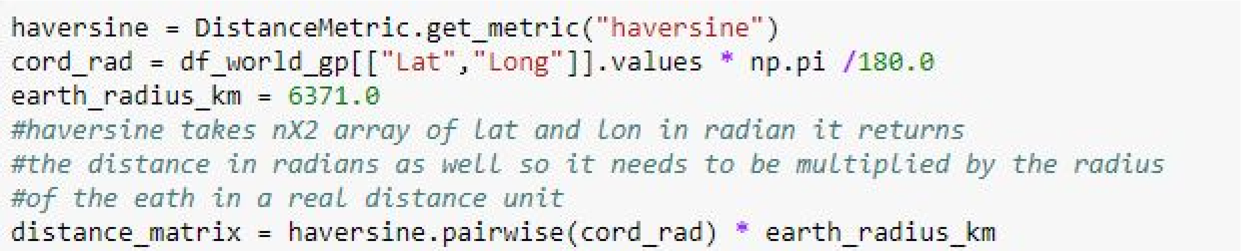
```
Measles           355
Whooping Cough    167
Mumps             101
Announcement       36
Diphtheria         13
Rubella            12
Meningitis          3
Name: Outbreak, dtype: int64
```



```
Category                                                  Attacks
Outbreak                                                 Violence
Location                                              Afghanistan
Continent                                                    Asia
Lat                                                        33.934
Long                                                   67.7034312
Date                                                       3/2014
Year                                                         2014
Cases                                                           3
Fatalities                                                      3
Impact Scale                                                  NaN
Source Citation    IANS Live. "Roadside bomb kills 3 polio vaccin...
Source             http://www.ianslive.in/index.php?param=news/Ro...
```

Hotspot Clustering Using DBSCAN

- Lat and Long Coordinates converted to Radians
- Used Haversine and Earth Radius (6371.0 km)
- DBS model parameters epsilon 190 km, minimum samples 3



```
haversine = DistanceMetric.get_metric("haversine")
cord_rad = df_world_gp[["Lat","Long"]].values * np.pi /180.0
earth_radius_km = 6371.0
#haversine takes nX2 array of lat and lon in radian it returns
#the distance in radians as well so it needs to be multiplied by the radius
#of the eath in a real distance unit
distance_matrix = haversine.pairwise(cord_rad) * earth_radius_km
```

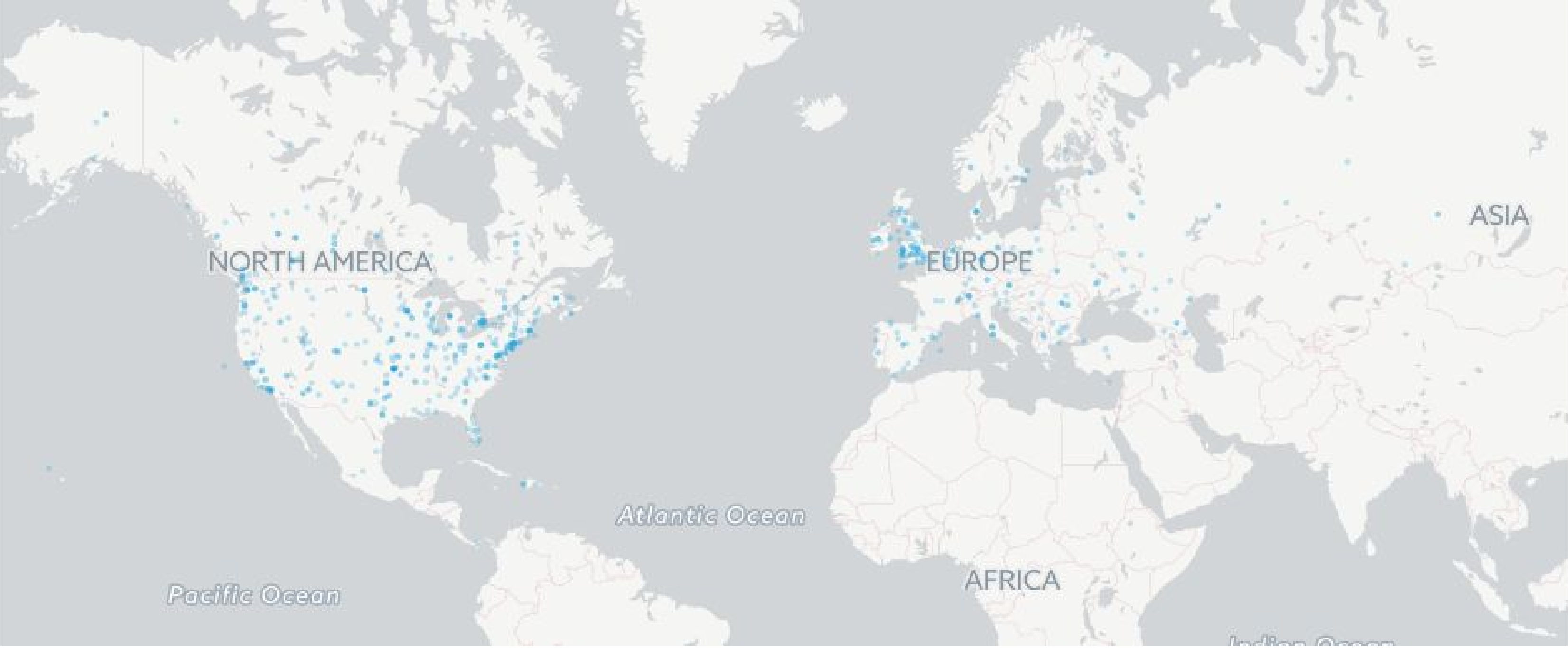
Results

holoviews.org

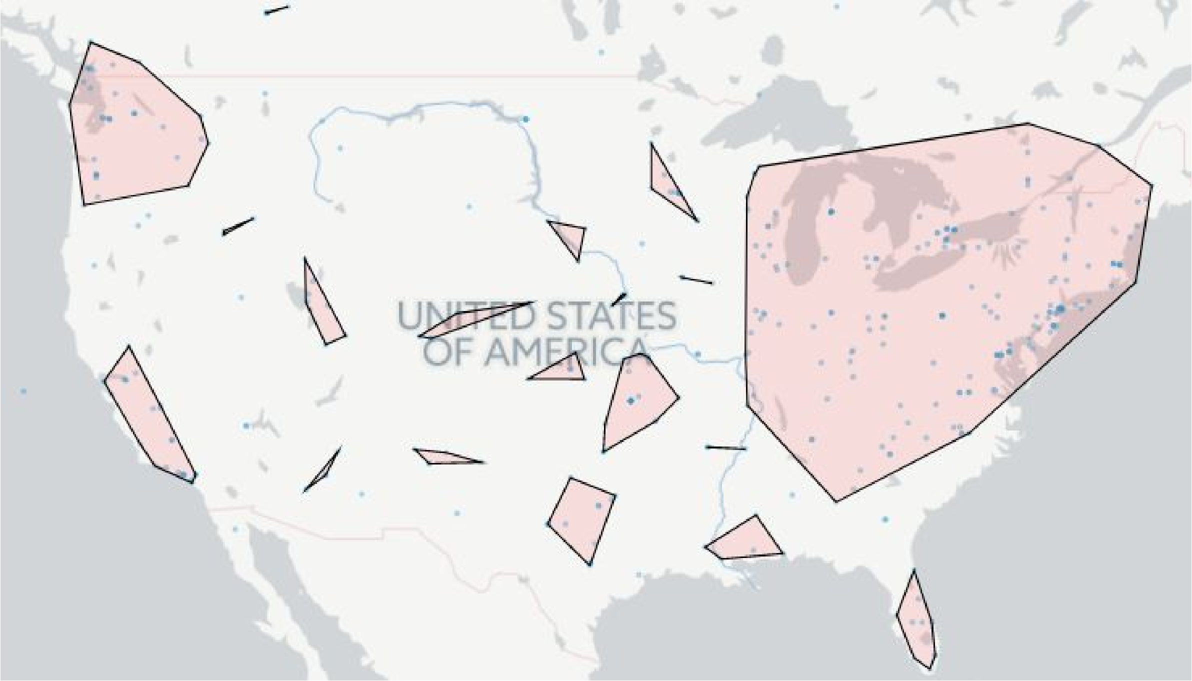
- 45 Clusters
- Silhouette Coefficient of 0.020511
- Mapping the clusters Holoviews with Bokeh, geoviews and geopandas was used

bokeh.pydata.org

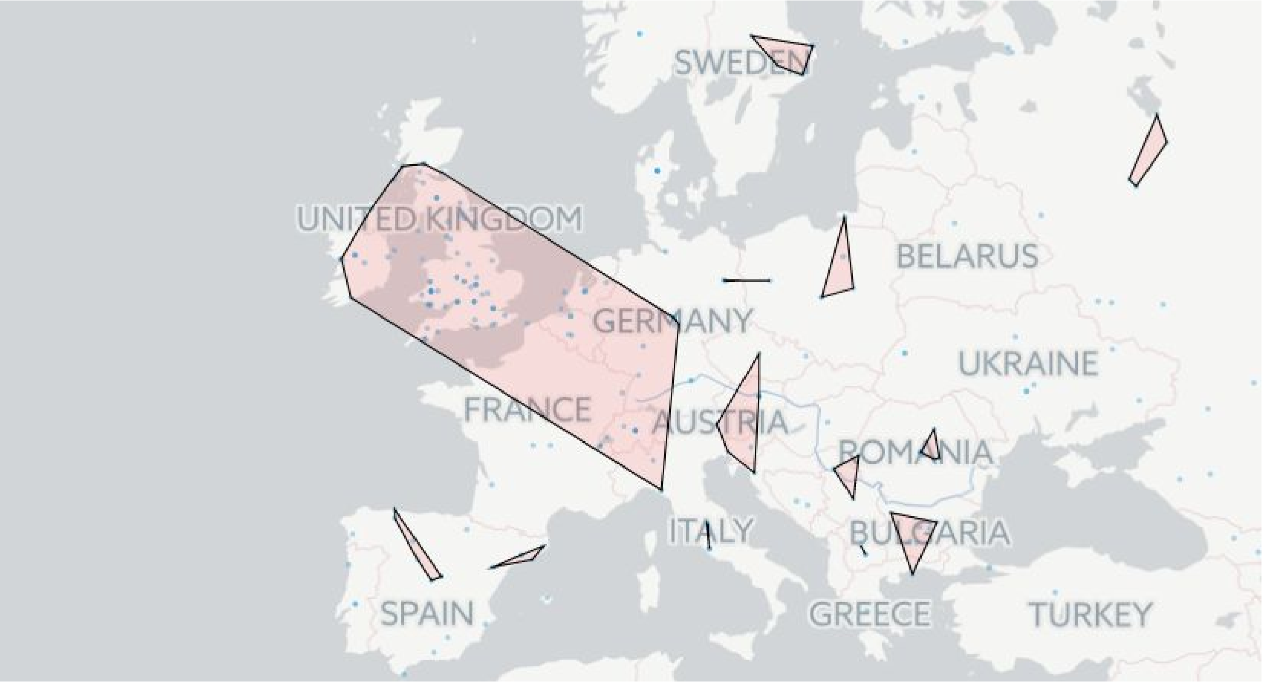
Before Clustering



North America

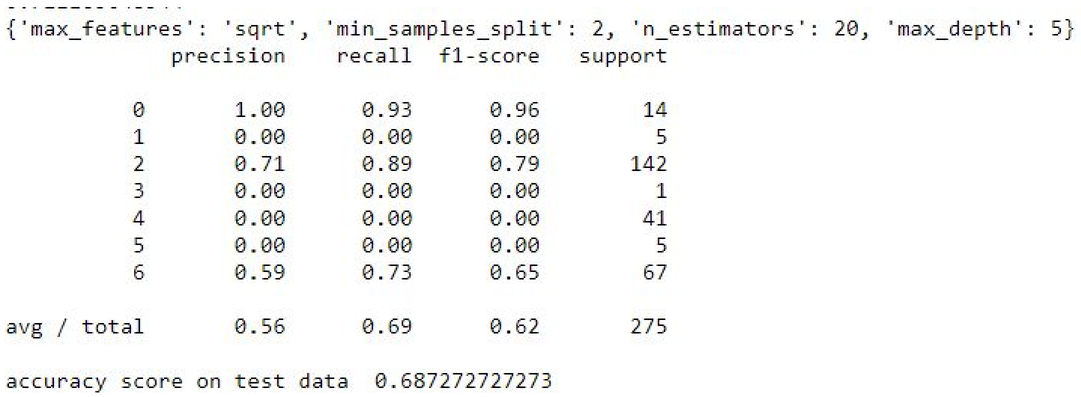


Europe


SWEDEN
UNITED KINGDOM
BELARUS
GERMANY
UKRAINE
FRANCE
AUSTRIA
ROMANIA
ITALY
BULGARIA
SPAIN
GREECE
TURKEY

Classifier on Diseases

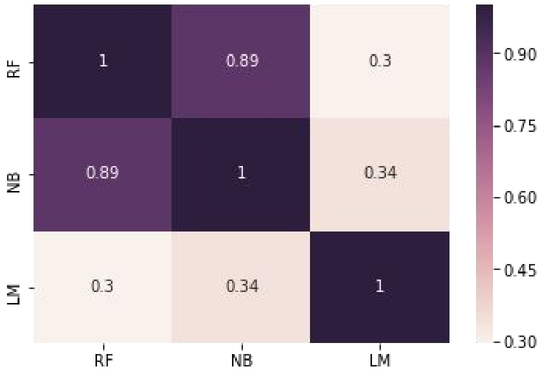
- Random Forrest
- Features : Cases, Fatalities, Impact Scale, Cluster labels, Continent
- Baseline 51.67%
- Accuracy score 68.72% on test data 71.11% on training data



```
{'max_features': 'sqrt', 'min_samples_split': 2, 'n_estimators': 20, 'max_depth': 5}
             precision    recall  f1-score   support

          0       1.00      0.93      0.96        14
          1       0.00      0.00      0.00         5
          2       0.71      0.89      0.79       142
          3       0.00      0.00      0.00         1
          4       0.00      0.00      0.00        41
          5       0.00      0.00      0.00         5
          6       0.59      0.73      0.65        67

avg / total       0.56      0.69      0.62       275

accuracy score on test data  0.687272727273
```

Regression on Cases

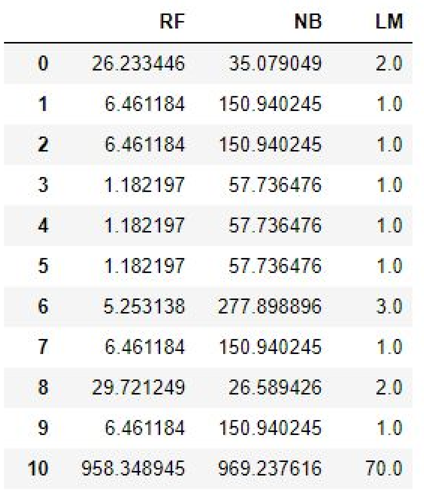
- Predict Cases based on Fatalities and Diseases and type of Outbreak
- Gradient Boost Regressor with Logistic Regression, Bayesian Ridge, Random Forest Regressor



Correlation heatmap of predictions of each model

Results

- R^2 score of .08897- 8% of the variance in Cases is explained
- Params on Gradient Boosting n_estimators: 100, loss: lad(least absolute deviation), max_depth:3



	RF	NB	LM
0	26.233446	35.079049	2.0
1	6.461184	150.940245	1.0
2	6.461184	150.940245	1.0
3	1.182197	57.736476	1.0
4	1.182197	57.736476	1.0
5	1.182197	57.736476	1.0
6	5.253138	277.898896	3.0
7	6.461184	150.940245	1.0
8	29.721249	26.589426	2.0
9	6.461184	150.940245	1.0
10	958.348945	969.237616	70.0

Predictions of each model
RF =RandomForest
NB = Bayes Ridge
LM=Logistic Regression

Conclusion

- SIR model on Herd Immunity create new feature of potential infections
- Able to classify more diseases with more data
- Prediction of Cases might be susceptible to how people behave against an epidemic.
- Improve Doctor-Parent communication about vaccines

Questions?

Website: Javigonscience.com
LinkedIn: linkedin.com/in/javiergonzalezcompte/

Source of Data

Council of Foreign Relations: https://www.cfr.org/interactives/GH_Vaccine_Map/#map