Data analysis on death by disease

Asifur Rahman

Contents

Problem Overview

- Disease can change the picture of a country or world. As we have seen in 2021 because of covid out gdp went down and now things are getting normal and gdp becoming normal also.

  So my main objective of this project figure out if there relationship predicting gdp with the death number caused by disease. Also i will build a time series model to predict future gdp in coming years.

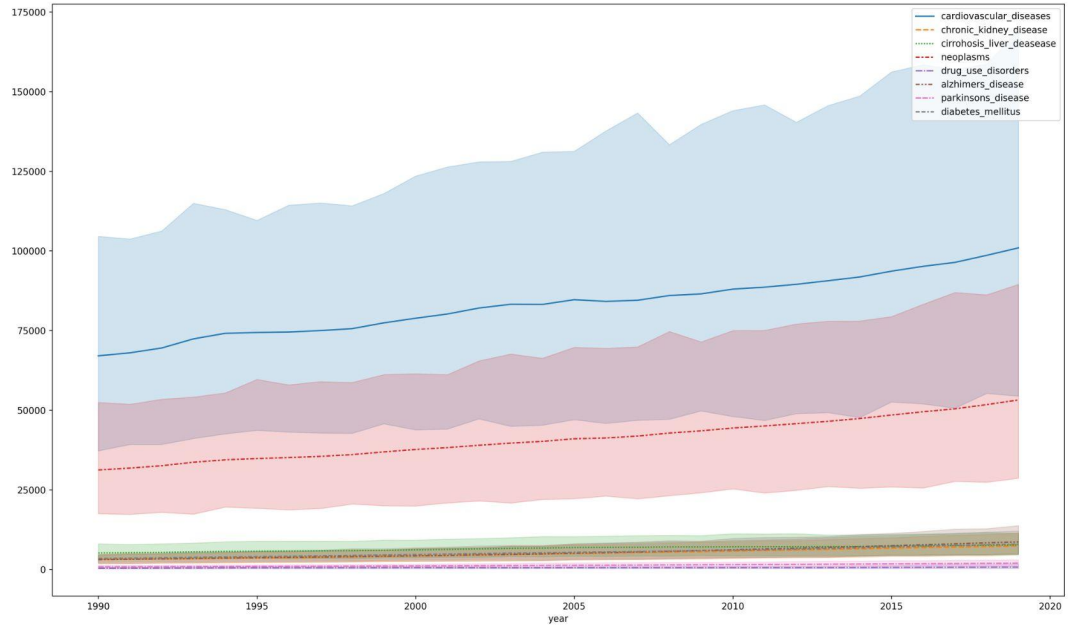
Data Overview

4 datasets

1. 30 year data of death that cause by disease for all countries
2. 71 years gdp data of United States (from 1950 to 2021)
3. 30 years cancer death by type data for all countries
4. 30 years population data of all countries

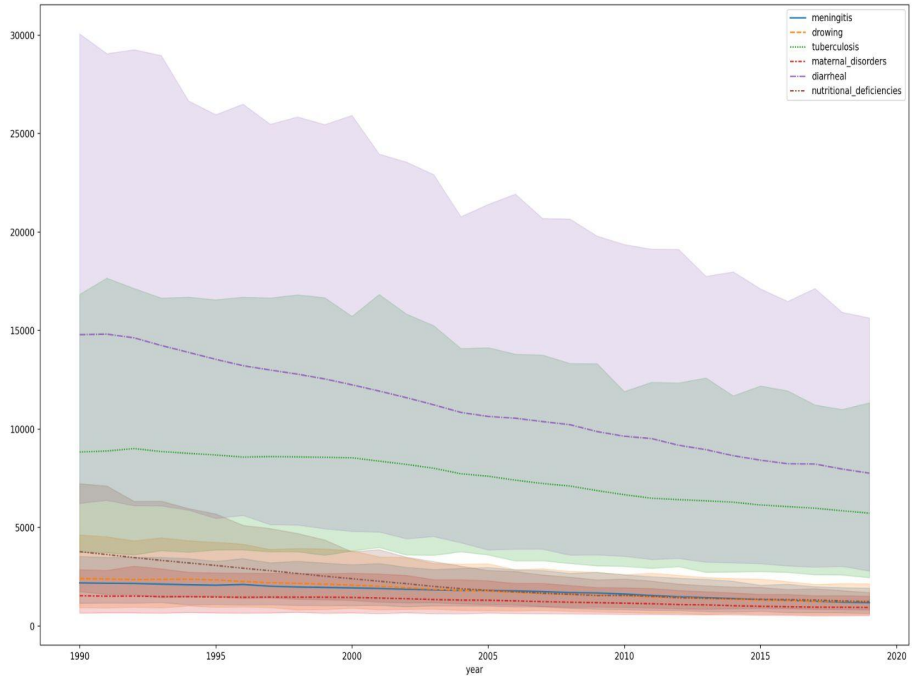
Exploratory Data Analysis

- From 1990 to 2019 death caused by disease like cancer, cardiovascular disease, liver, alzheimer's, drug use disorder, diabetes have increased.

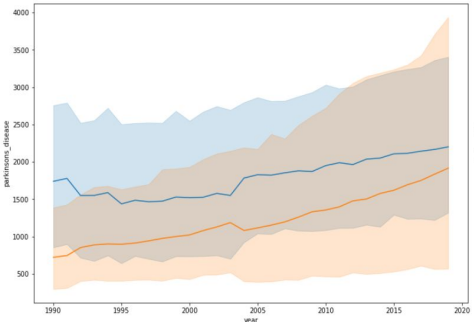


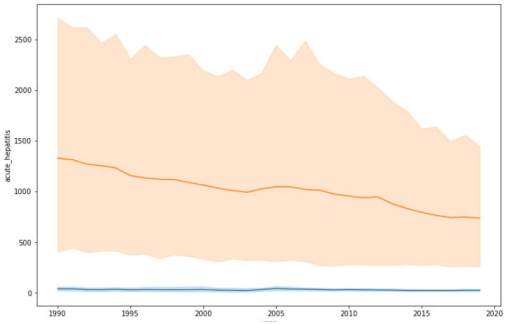
Death by decease

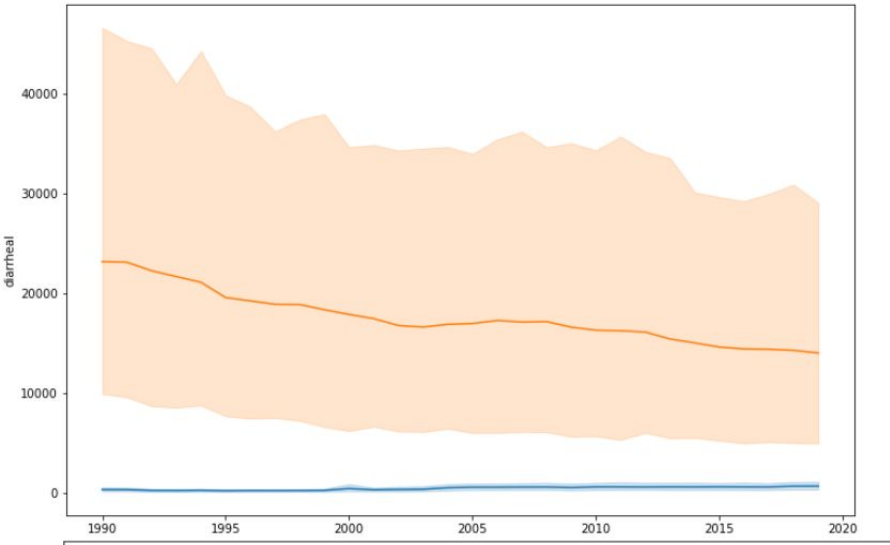
- meningitis disease, drowing, dirhea, maternal disorder, nutritional deficiencies death rate have been decreased in last 30 years
-

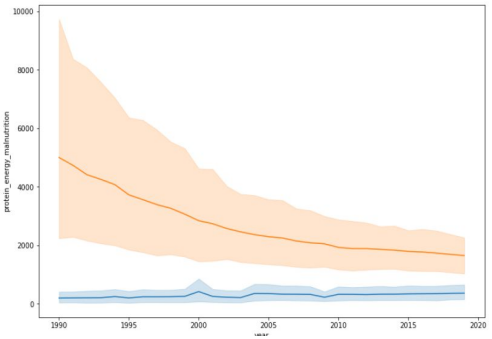


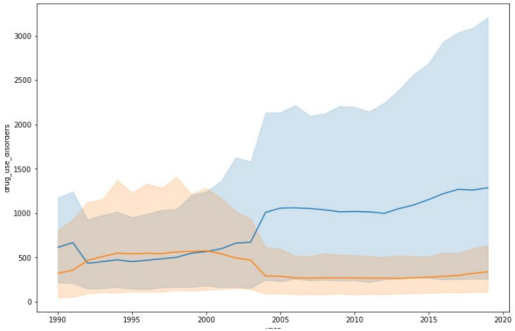
Developed vs undeveloped

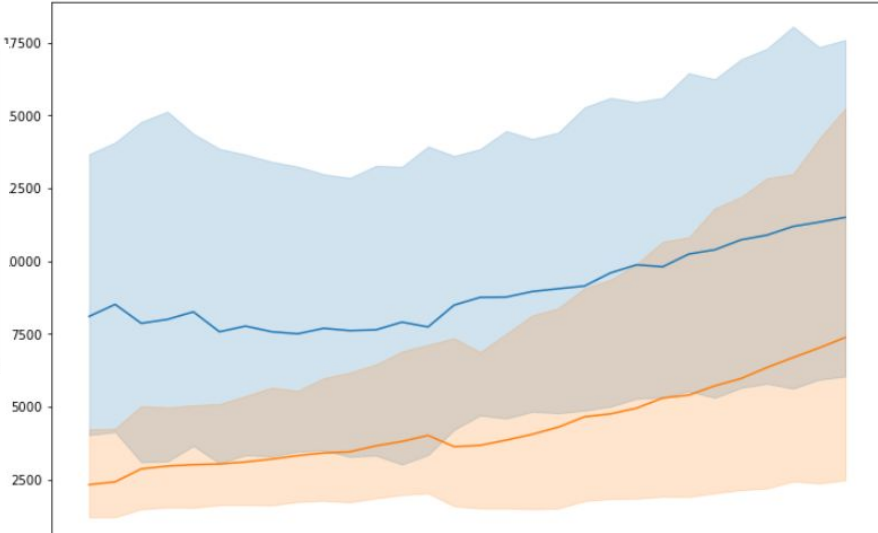




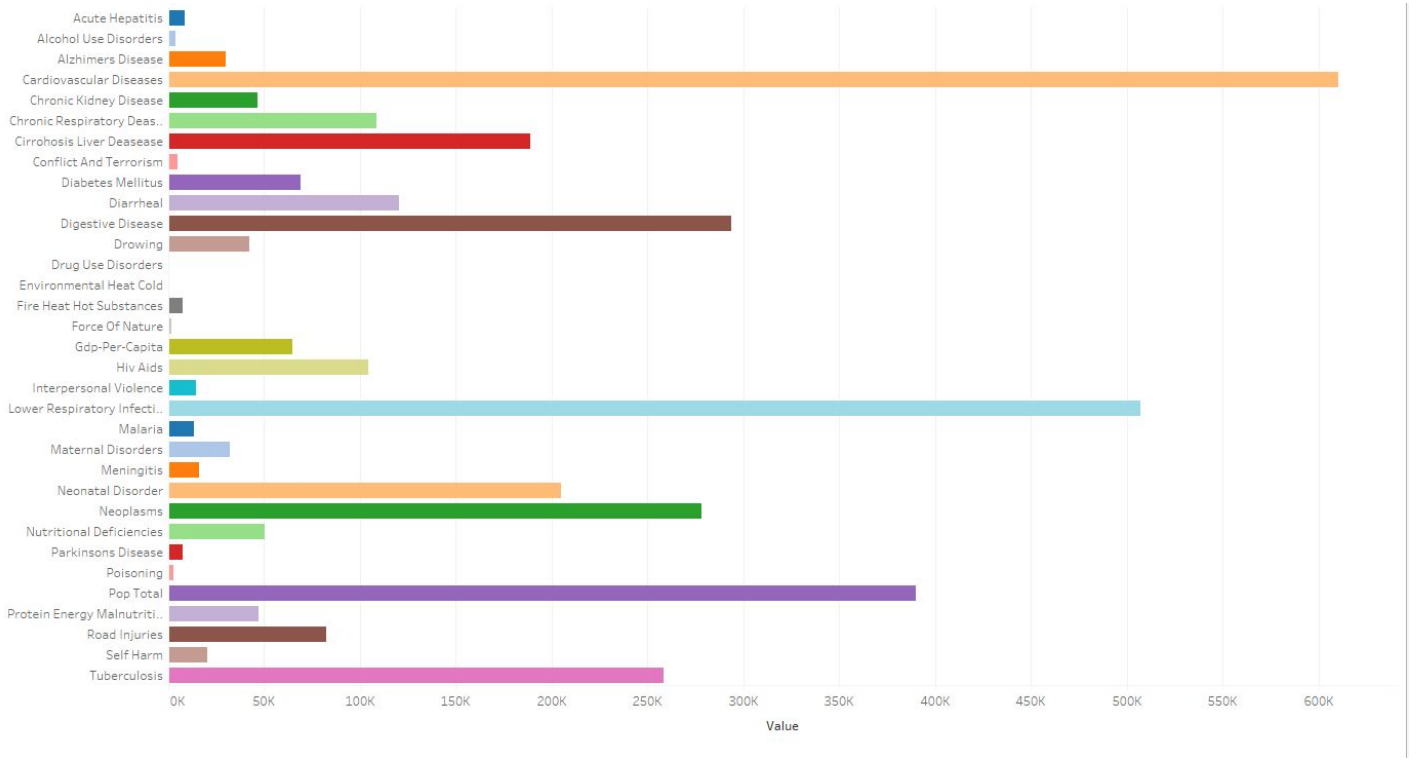




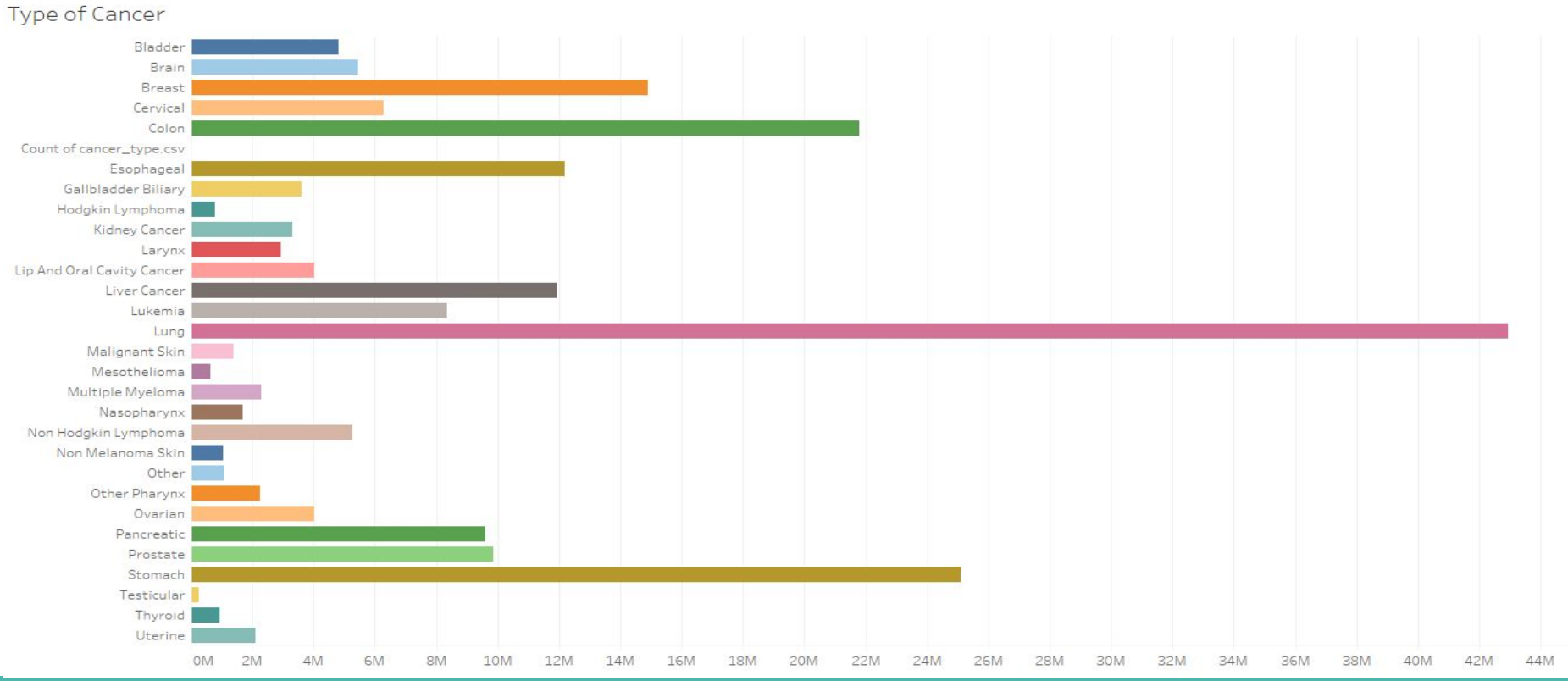




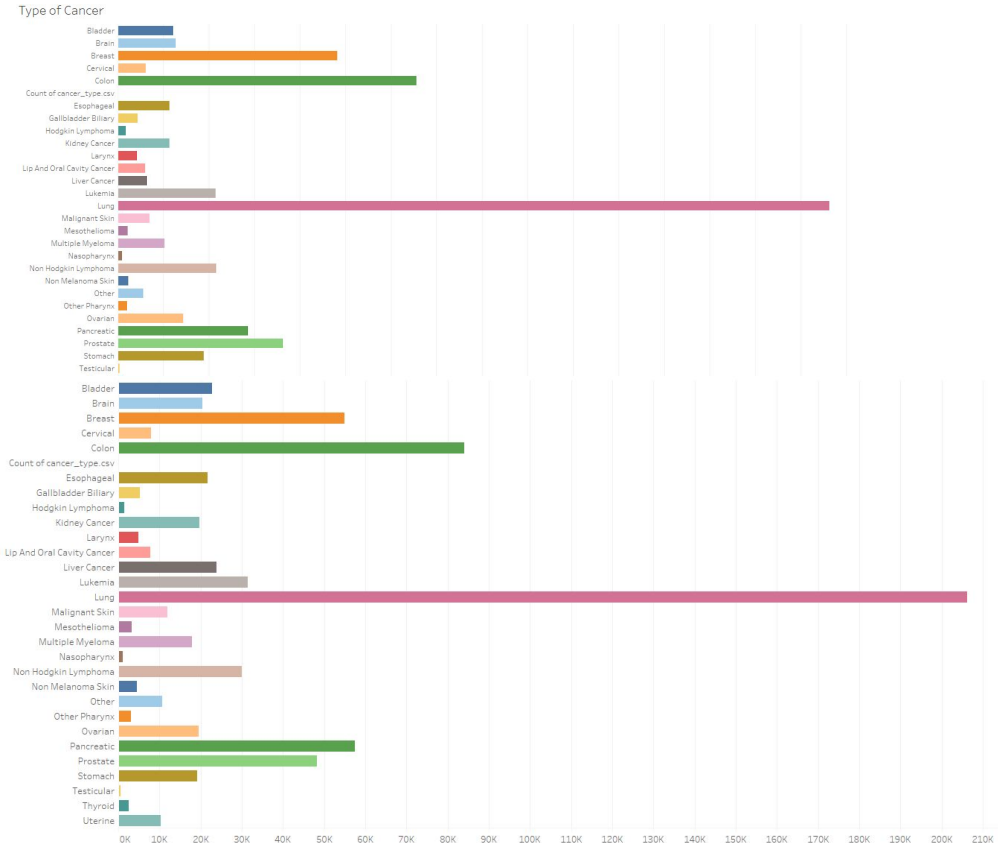
Sum of death by disease in last 30 years

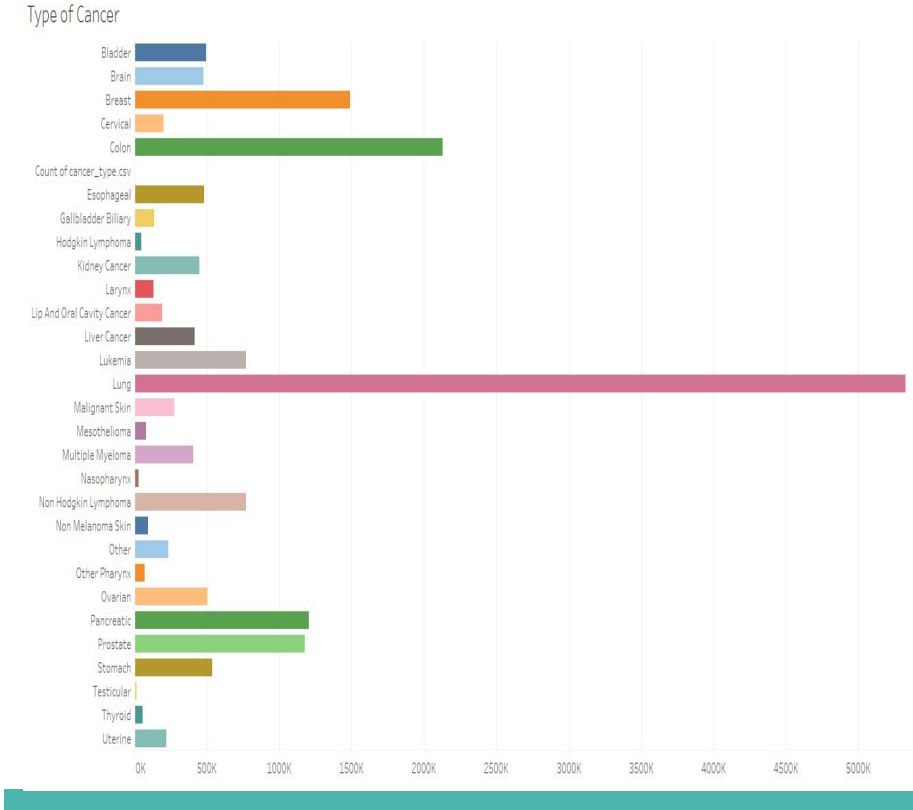


Sum of death by type of cancer



Death rate by type of cancer in USA





cancer

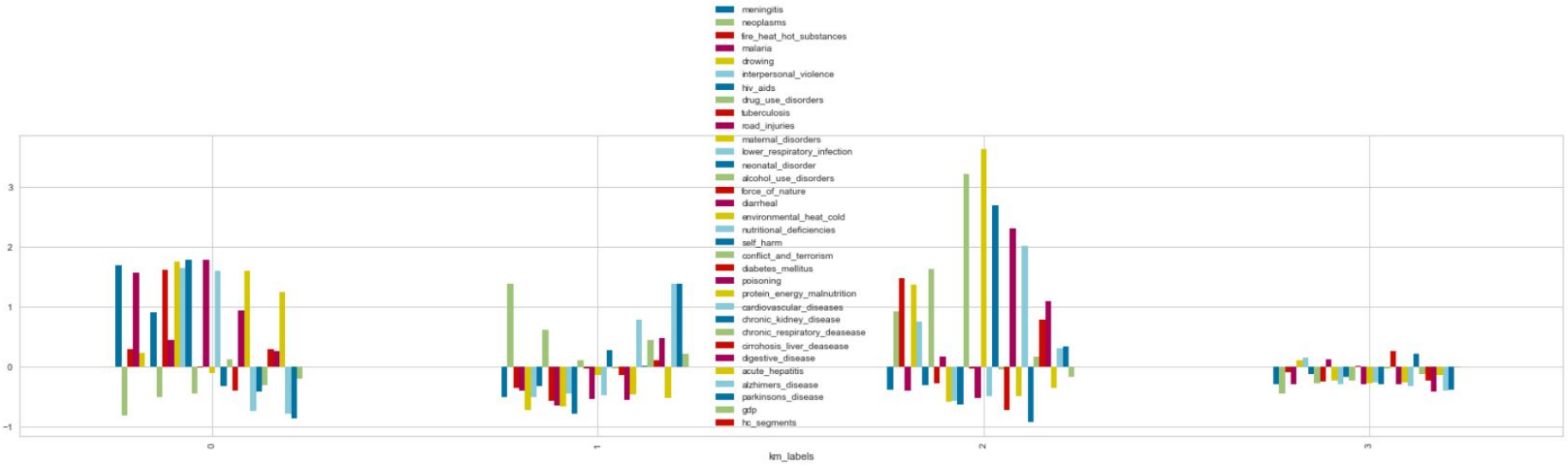
- Cancer rate in Australia is about 579 per 100000 people
- USA cancer rate is 571 per 100000 and death rate is about 183 people per 100k

death rate by disease

country	disease	amount
Bulgaria	cardiovascular_diseases	11.302586
Lesotho	hiv_aids	5.078255
Hungary	neoplasms	3.540913
Fiji	diabetes_mellitus	1.938482
Chad	diarrheal	1.904574
Central African Republic	tuberculosis	1.820255
Mali	neonatal_disorder	1.628125
Sierra Leone	malaria	1.499666
Nepal	chronic_respiratory_deasease	1.426123
Somalia	lower_respiratory_infection	1.343720
Japan	alzhimers_disease	1.299650
Japan	alzhimers_disease	1.299650
Afghanistan	PopTotal	1.000000
Mauritius	chronic_kidney_disease	0.992566
Mali	nutritional_deficiencies	0.756181
Romania	digestive_disease	0.755296
Mali	protein_energy_malnutrition	0.739377
Central African Republic	road_injuries	0.641606
Afghanistan	conflict_and_terrorism	0.638637
Egypt	cirrohosis_liver_deasease	0.623932
El Salvador	interpersonal_violence	0.472159
Lesotho	self_harm	0.346427
Niger	meningitis	0.333429
Belarus	alcohol_use_disorders	0.314472
Chad	maternal_disorders	0.203108
United States	drug_use_disorders	0.199708
Solomon Islands	drowing	0.181719
Germany	parkinsons_disease	0.154160
Bahamas	force_of_nature	0.128374

clustering

Countries can be divided within 4 different clusters.



Cluster

1. Cluster 3

   Turkey, Azerbaijan, Cyprus, Lebanon

2. Cluster 1

   Iceland, Greece, Ireland, Italy

3. Cluster 0

   Mali, Lesotho, Nigerm Guinea

4. Cluster 2

   Russia, Kazakhstan, Belarus, Greenland

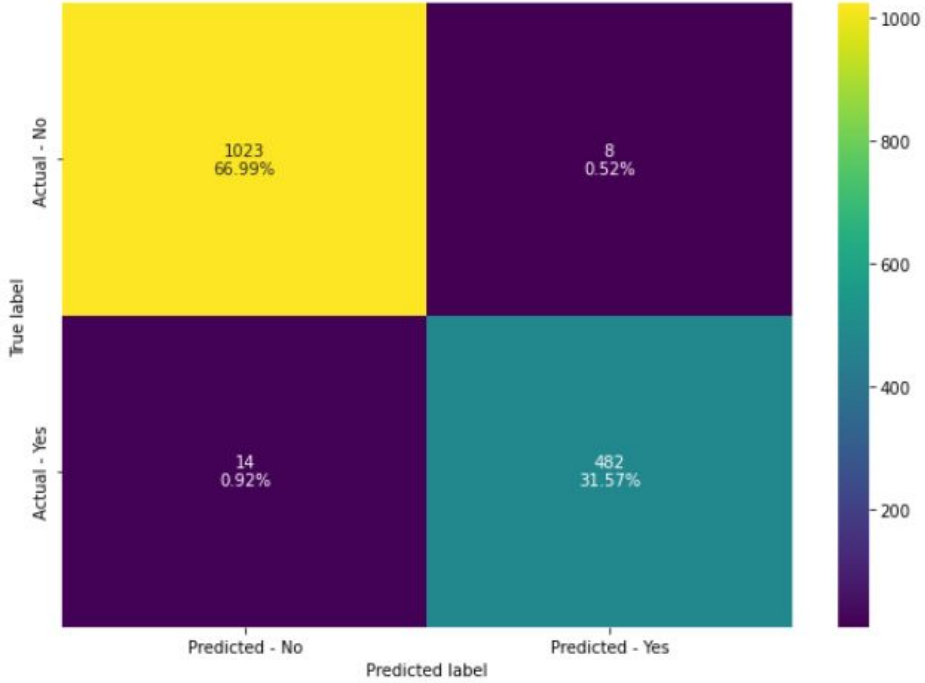
Predicting gdp by death rate

Random forest have score better than all other model. Its have 100 percent accuracy on train and in test almost 99 percent. Also precision and recall score have score close to 100.

	model	train_score	test_score	recall_score_train	recall_score_test	precision_train	precision_test	cross_val_score
0	Logistic_Regression	0.835486	0.814669	0.548443	0.479839	0.908309	0.904943	0.827619
1	Tuned_logreg	0.863841	0.844794	0.659170	0.598790	0.893318	0.886567	0.859064
2	knn	0.918866	0.863130	0.826990	0.717742	0.914833	0.837647	0.845593
3	tuned_knn	1.000000	0.849378	1.000000	0.665323	1.000000	0.837563	0.852892
4	Decision_Tree	1.000000	0.971185	1.000000	0.949597	1.000000	0.961224	0.969681
5	tuned_decision_tree	0.729927	0.719057	0.925606	0.915323	0.549846	0.539834	0.728521
6	bagged_d_tree	0.999719	0.979699	0.999135	0.951613	1.000000	0.985386	0.975009
7	tuned_bagged_d_tree	0.996070	0.979699	0.992215	0.955645	0.995660	0.981366	0.974450
8	random_forest	1.000000	0.985593	1.000000	0.971774	1.000000	0.983673	0.983154
9	tuned_random_forest	0.999439	0.983628	0.998270	0.961694	1.000000	0.987578	0.982873
10	ada_boost	0.948344	0.922069	0.906574	0.864919	0.932384	0.891892	0.934581
11	tuned_ada_boost	0.742841	0.744597	0.892734	0.891129	0.565789	0.568123	0.759401
12	svc	0.765020	0.754420	0.294118	0.266129	0.941828	0.923077	0.760245
13	Gradient_Boost	0.989613	0.967256	0.976644	0.935484	0.991220	0.962656	0.963781
14	tuned_gradient_boost	0.992139	0.967256	0.981834	0.935484	0.993870	0.962656	0.965185

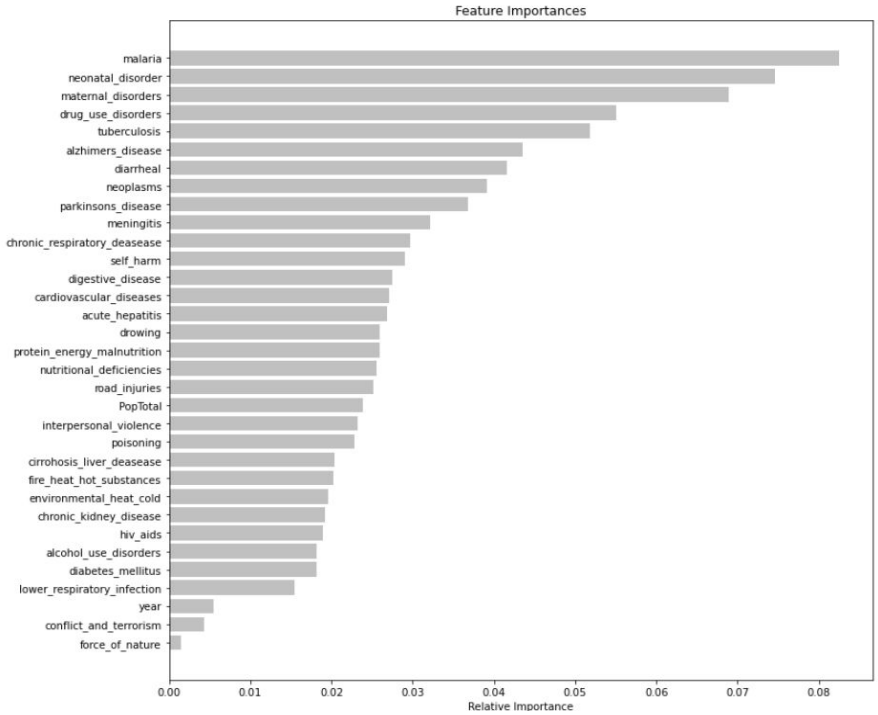
Confusion metrics

1. Base accuracy score was 31 percent

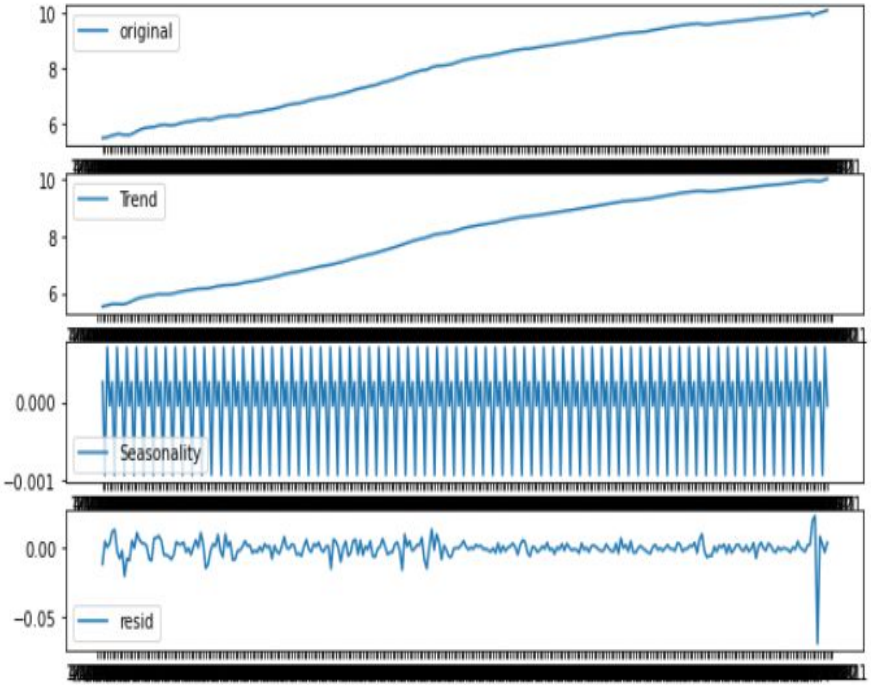


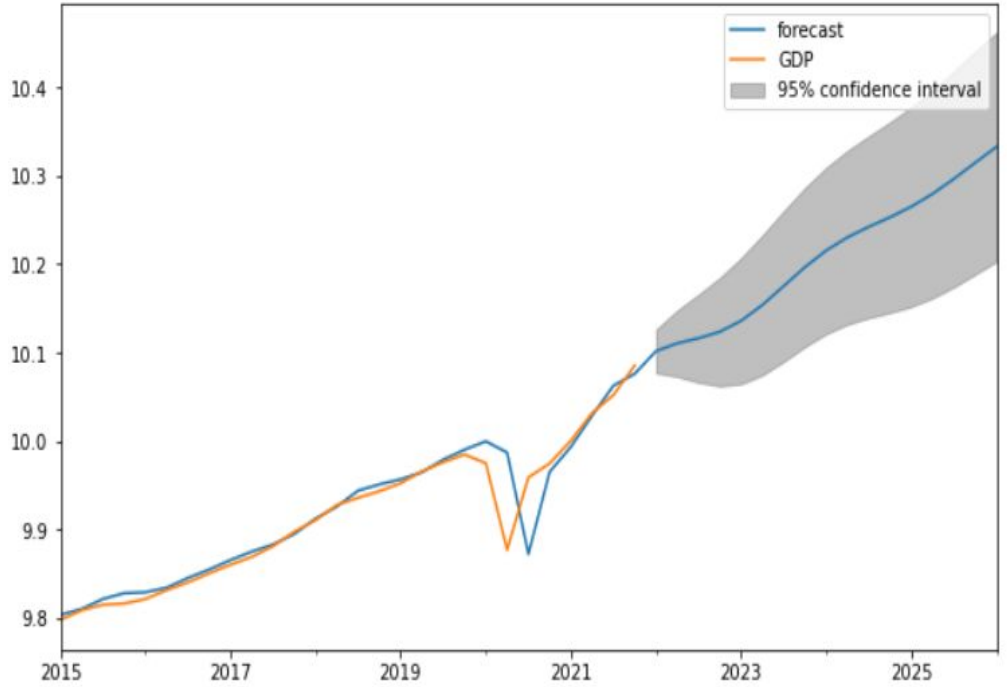
Feature importance

- These specific feature have larger effect on our model that is being used to predict gdp.



Forecasting next 4 years gdp





Conclusion

We can conclude that from cardiovascular and lower respiratory infection people are dying much more than other disease. If we compare the death rate from 1990 to 2019 we can see that some disease death rate are decreasing and some of them are increasing. The percentage of death by cancer , liver disease, cardiovascular, diabetes are increasing day by day. And death rate from disease like Malaria, diarrhea is decreasing all over the world.

Most of the developed country have high death rate from cancer, cardiovascular , alzheimer's disease , parkinson’s disease and very low in diarrhea, malaria