Bits N' Bytes

Predictive Analysis
COVID-19 & Future Pandemics

Our team selected **SDG 11** and created two **models** as our own **solutions** using **indicators** we put together towards **analyzing the effect** of COVID-19 on the pursuit of sustainable cities and communities. Both solutions **visualize** their results as a way to inform the general public.

Solution 1

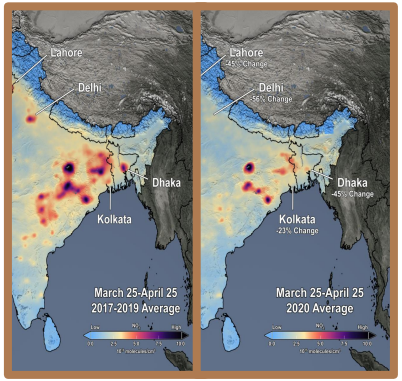
1. Utilize data from the **global level** in order to predict and visualize **pandemic growth and pollution** levels over time.
2. Uses the **nearest neighbor** approach based on **user inputed key factors** to accurately **match a country in the future** with a country during Covid-19, so if another pandemic happens, they can get an **estimation** of how the pandemic would affect their country.
3. Uses the programming language "**Julia**" and a python script to create **interactive visualizations** of pandemic cases over time and NO2 pollution levels over time.

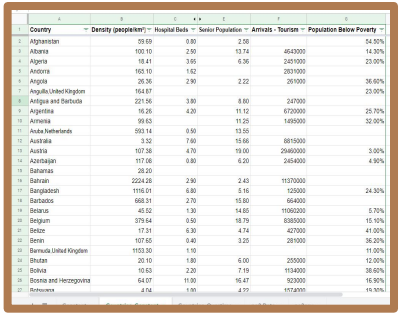
Solution 2

1. Solution 2 shows COVID-19's impact on **sustainability** by **providing geospatial information** on future predicted regions
2. Utilizes data at the **county level** to provide spatial patterns at the **national/country scale**
3. Accurately **predicts emerging Coronavirus hotspots** in the **short-term**
4. Uses **linear regression** to form optimized functions for each county's new COVID-19 cases per day.
5. Predicts hotspots by finding counties with the **most rapidly growing rates**
6. Choropleth map **visually indicates** nodal regions with high case rates

Model 1: Long-Term Global Pandemic Tracker Process

1. Before gathering our datasets and creating a model, we decided to get an understanding on **root causes** of the virus as well as some drastic **effects**. Below is an example of the virus affecting **air pollution** in India. Therefore, we decided to make a model returning such trends.

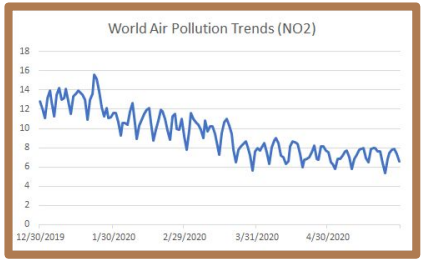


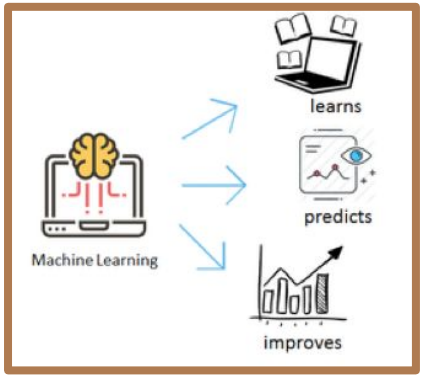


Country	Density (people/km²)	Hospital Beds	Senior Population	Arrivals - Tourism	Population Below Poverty
Afghanistan	59.69	0.80	2.58		54.50%
Albania	100.10	2.90	13.74	4543000	14.30%
Algeria	18.41	1.90	6.36	2491000	23.00%
Andorra	165.10	1.62		2831000	
Angola	26.26	2.90	2.22	261000	36.60%
Anguilla,United Kingdom	164.87				23.00%
Antigua and Barbuda	221.56	3.80	8.60	2477000	
Argentina	16.26	4.20	11.12	6720000	25.70%
Armenia	99.63		11.25	1495000	32.00%
Aruba,Netherlands	593.14	0.50	13.55		
Australia	3.32	7.60	15.66	8815000	
Austria	107.56	4.70	19.00	29460000	3.00%
Azerbaijan	117.06	0.80	6.20	2454000	4.90%
Bahamas	39.20				
Bahrain	2224.28	2.90	2.43	11370000	
Bangladesh	1116.01	0.80	5.16	125000	24.30%
Barbados	668.31	2.70	15.80	664000	
Belarus	45.52	1.30	14.66	11060200	5.70%
Belgium	379.64	0.50	18.79	8380000	15.10%
Belize	17.51	0.30	4.74	427000	41.00%
Benin	107.65	0.40	3.25	297000	36.20%
Bermuda,United Kingdom	1153.30	1.10			11.00%
Bhutan	20.10	1.80	6.00	255000	12.00%
Bolivia	10.63	2.20	7.19	1134000	38.60%
Bosnia and Herzegovina	64.07	11.00	16.47	923000	16.90%
Botswana	4.04	1.80	4.23	1572000	19.30%

2. Our next step was to gather **datasets** of the root causes. Based on extensive research and deduction, we outlined them to be: **Population Density, Hospital Beds, Senior Population, Travel, Poverty**. Using numerous sources, we **gathered this data** as shown above and **organized** it based on our needs.

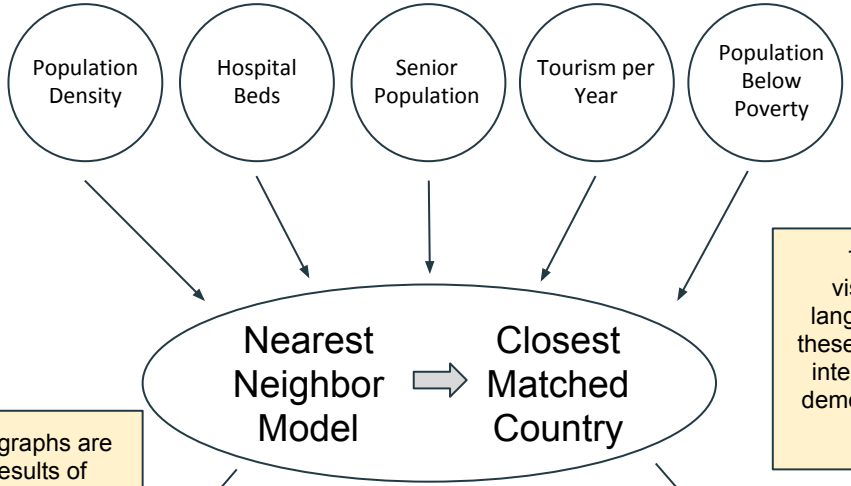
3. Next, we decided to gather the **datasets** for each country's **COVID-19 cases** and **air pollution trends** from January 1 to May 30 on a daily basis. Using **Excel's Pivot Table** Technology, we created two spreadsheets. Here's a graph to show **air pollution trends** in our data.





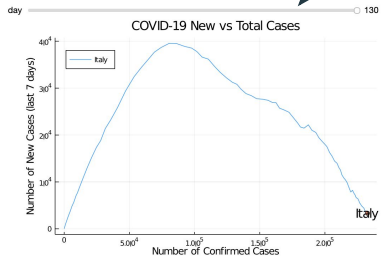
4. Our last step was to import these datasets into our **Julia & Python** code and use **Machine Learning Algorithms** to create a model. Therefore, when we receive any user input (A Future Country's Characteristics), we would to match with our database and **return corresponding visuals**. Our code is shown above.

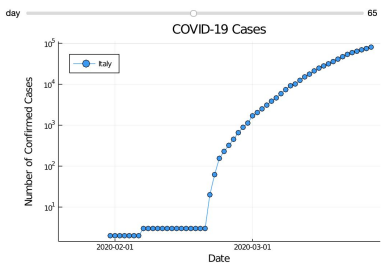
Model 1: Long-Term Global Pandemic Tracker Process

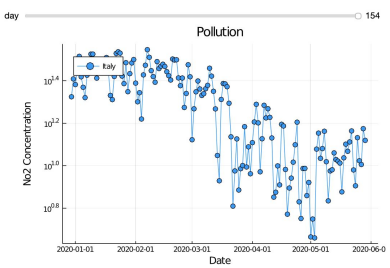


The Julia visualization language allows these graphs to be interactive!! The demo shows how.

These graphs are the results of Italy-similar factors.

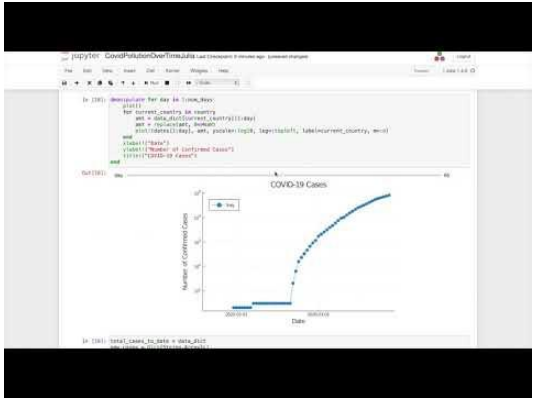






EXAMPLE!

All the user (Ex: A future country) has to do is plug in their characteristics, and the program returns a country with similar characteristics and numerous visuals such as **Predicted Cases & Air Pollution Trends** the future country must be aware of.



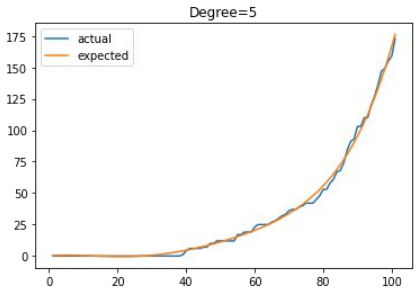
Model 2: Short-Term COVID-19 National Predictor

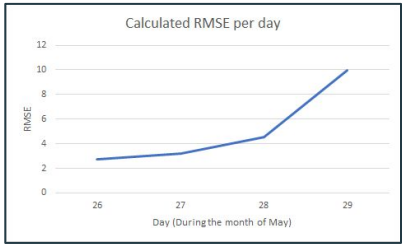
1. Before we began our data analysis, we utilized **satellite imagery** of light pollution from **NASA** to identify metropolitan hubs that could become potential hotspots of COVID-19. This gives us a preliminary visualization of the distribution of cases over the US.

2. Through research, we acquired a list of **COVID-19 cases per day** for each county in the US between 1/22/20 and 5/29/20.

	A	B	C	D	E	F	G	H
1	countyFIPS	County Na	State	stateFIPS	5/20/2020	5/21/2020	5/22/2020	5/23/2020
2	1001	Autauga C	AL	1	136	147	149	155
3	1003	Baldwin C	AL	1	270	270	271	273
4	1005	Barbour C	AL	1	96	100	104	105
5	1007	Bibb Coun	AL	1	52	52	55	58
6	1009	Blount Co	AL	1	47	48	49	49
7	1011	Bullock Co	AL	1	64	71	89	105
8	1013	Butler Cou	AL	1	312	321	329	335
9	1015	Calhoun C	AL	1	136	136	137	138
10	1017	Chambers	AL	1	330	330	330	330
11	1019	Cherokee	AL	1	30	31	33	33
12	1021	Chilton Co	AL	1	83	84	85	86
13	1023	Choctaw C	AL	1	129	133	135	140
14	1025	Clarke Co	AL	1	89	91	92	97
15	1027	Clay Count	AL	1	27	27	27	27
16	1029	Cleburne C	AL	1	13	13	13	13
17	1031	Coffee Co	AL	1	184	184	189	196
18	1033	Colbert Co	AL	1	110	112	117	125
19	1035	Conecuh C	AL	1	22	23	24	25
20	1037	Coosa Cou	AL	1	33	33	33	33
21	1039	Covington	AL	1	59	62	63	63
22	1041	Crenshaw	AL	1	52	52	53	58
23	1043	Cullman C	AL	1	71	73	73	74
24	1045	Dale Coun	AL	1	71	76	77	81
25	1047	Dallas Cou	AL	1	172	179	182	191
26	1049	DeKalb Co	AL	1	206	209	209	216
27	1051	Elmore Co	AL	1	226	238	242	255
28	1053	Escambia	AL	1	39	39	39	40
29	1055	Etowah Cc	AL	1	223	225	225	228

3. For each county, we used **polynomial regression** to predict the expected number of cases for a future date. To achieve an optimal function, we strove to minimize the the **Root Mean Squared Error (RMSE)** between the actual and expected predictions. For each county, we tested polynomial functions from degrees 0 to 20, and chose the one with the smallest RMSE.





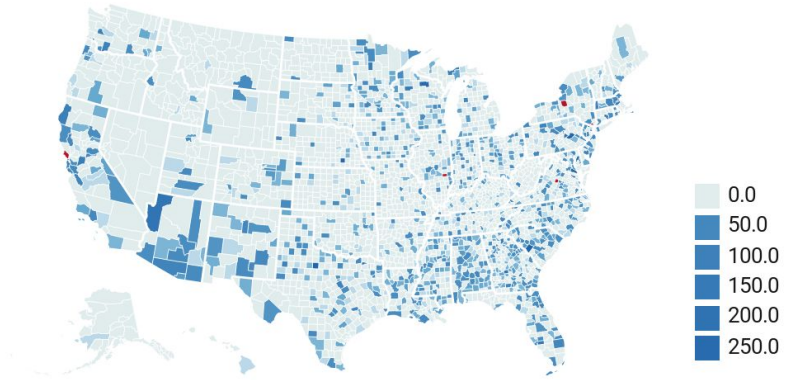
4. To test the accuracy of our model, we used it to **predict the expected number of cases** for each county between May 26th and 29th and compared the results with the actual number of cases. As the number of days increased, the **typical deviation** between predicted and actual began to increase. The line-graph shows the association between number of days and the deviation between predicted and actual.

Model 2: Short-Term COVID-19 National Tracker

5. By computing the instantaneous rate of change for the county at a given future date, we can obtain the predicted future rate of new cases per day in each county. Using data from May 25th, we estimated the rate of new cases on May 28th and **graphed the results on a choropleth to display emerging Coronavirus hotspots in the US at the county level**. The actual number of new cases on May 28th typically varied from the predicted value by 4.53 cases. The choropleth graphically displays the distribution of predicted cases on the day. The county with the highest predicted rate of new cases is Onondaga County in New York with 1771 predicted new cases on May 28th. Some factors that lead to this are its high population density of 571.37 people per square mile, whereas unaffected regions in Nevada average 28 people per square mile. Using the data on the choropleth allows local municipalities and governments to predict the rise in cases in their communities and accordingly enforce lockdowns, stricter social distancing rules, and order more Personal Protective Equipment. COVID-19 has **severely affected the sustainability of cities and communities**, but through the usage of this model, leadership at any level can predict the future cases and adequately prepare their cities and communities.

Expected Rate of Covid-19 Cases per Day

The choropleth map displays the distribution of predicted rates of Covid-19 cases per day in each county for May 28th.



Darker shades of blue indicate higher predicted rates of Covid-19 cases in the given counties. Counties in red have predicted rates exceeding 400 cases per day.

Map: Bits N' Bytes • Source: USAFacts • Created with Datawrapper