Weather-Disease Relationship

1st Yihao Zhao
Team08 - DataHeros
Champaign, USA
yihao4@illinois.edu

2nd Chirag Gupta
Team08 - DataHeros
Champaign, USA
chiragg4@illinois.edu

3rd Jaqueline Ortiz
Team08 - DataHeros
Champaign, USA
jortiz71@illinois.edu

***Abstract*—The goal of this project is to investigate the relationship between climate and disease rates. We have collected and combined disease data from CDC and climate data from NOAA to clean and preform data modeling using linear, random forests, and gradient boosting models. Utilizing data visualization to assist our findings, and ultimately created an interactive shiny application dashboard to generate an explorable explanation of the three models we built. We have also created multiple graphs to provide context of the combination of datasets of climate and disease.**

Based on the analysis, it appears that the data is not sufficient to draw conclusions about the relationship between climate and human illness though we consider it might be the reason for insufficient and inappropriate data.

I. Introduction

We are addressing the relationship between changes in weather and disease rates in the USA to identify possible reasons why people get sick, and to explore whether or not there is some relationship between climate change and human illness. We aim to draw insights from trends in illnesses and weather to draw conclusions on if they are correlated, and how so.

Our motivation to this project was the major uncertainty regarding this topic based on the various sources we looked at. We wanted to use various analyses to draw conclusions regarding this and make regression models to model any such collinearity between the two datasets.

The three Models we decided to implement had various climate parameters like year, month, minimum temperature, precipitation, Palmer Hydrological Drought Index, Palmer Modified Drought Index, Palmer Drought Severity index, and Palmer Z index as the predictor variables based on Feature Extraction and decided to go with Total Deaths as the response variable.

References

[1] Jonathan A. Patz, David Engelberg, and John Last. The Effects of Changing Weather on Public Health Annual Review of Public Health Vol. 21:271-307 (Volume publication date May 2000) https://doi.org/10.1146/annurev.publhealth.21.1.271
In this work, Extreme weather was studied to draw parallels between weather and disease using various analyses. This relates to our current project because the author provides samples and ways in which we can analyze the same and work on finding correlation between our variables.

[2] Sheehan MC. 2021 Climate and Health Review – Uncharted Territory: Extreme Weather Events and Morbidity. International Journal of Health Services. 2022;52(2):189-200. doi:10.1177/00207314221082452
In this work, "2021 Climate and Health Review – Uncharted Territory: Extreme Weather Events and Morbidity" data on cost, mortality, and displacement for 30 extreme weather events were summarized from three annual reports (Sheehan, 2022). These events were studied to raise awareness on all the effects of climate change, including cost, in hopes to raise urgency over the issue. This relates to our current project because the author provided many examples of extreme weather events associated with climate change that provided background on ways temperature change influences weather events. Sheehan used a case study to explore the impact of climate change on mortality, but in our project we are using a model-based design to predict how weather change influences mortality.

[3] Rohr, Jason R., et al. "Frontiers in Climate Change–Disease Research." Trends in Ecology & Evolution, Elsevier Current Trends, 12 Apr. 2011.
This article mentions that although there is more and more public attention attracted by the notion of climate change will generally increase the human diseases, there are some opposite voices as well due to many confounders; for example, the climate change could also damage the parasite extinctions, and which might also have a significant effect on human and even all creatures' diseases. To our project, this article suggests that to better link the data to our fitted models, we would need to pay some close attention on addressing confounding variables and context dependencies with consideration of community-level interactions and functional traits.

[4] C.H. & Fellingham, and S.A. Wyndham. Climate and Disease — South African Medical Journal. https://journals.co.za/doi/abs/10.10520/AJA20785135_18566.
This article discusses various relationships between mortality rate for total death and climate changes in temperature by seasons with the confounders of different ages and races of people, areas. By performing various data visualizations to seasonal temperature changes with different explanatory variables, The author indicates that the death rate tends to be higher in Winter than Summer with older people, and seasonal variations in air temperature have a significant effect due to the death rate. To our projects, we can draw some insights of different graphs from this article to better assist us plotting useful and clear graphs to better understand the relationship and effect behind this dataset.

[5] Rohr, Jason R., et al. "Frontiers in Climate Change–Disease Research." Trends in Ecology & Evolution, Elsevier Current Trends, 12 Apr. 2011 https://www.sciencedirect.com/science/article/pii/S0169534711000711?casa_token=L4q
ueGJIvr-qsjHp50tqNDcFmtdEfihdrOSBMmDHI.
This article mentions that although there is more and more public attention attracted by the notion of climate change will generally increase the human diseases, there are some opposite voices as well due to many confounders; for example, the climate change could also damage the parasite extinctions, and which might also have a significant effect on human and even all creatures' diseases. To our project, this article suggests that to better link the data to our fitted models, we would need to pay some close attention on addressing confounding variables and context dependencies with consideration of community-level interactions and functional traits.

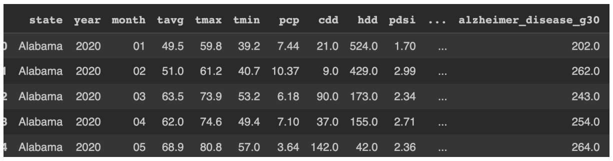
[6] Colwell RR, Patz JA. Climate, Infectious Disease and Health: An Interdisciplinary Perspective
https://www.ncbi.nlm.nih.gov/books/NBK559442/ doi: 10.1128/AAM-Col.20Jun.1997
This article mentions that climate change and human health reviewed the potential effects of global climate change on human diseases. The report found that there is a clear link between the incidence of vector borne diseases such as malaria, hantavirus, dengue, and viral encephalitis and climate and weather factors. Other infectious diseases, such as

campylobacteriosis and cholera, have cyclical patterns of occurrence that suggest a link to climate, but the mechanisms behind this link are not yet well understood. This report focuses on the connections between weather variability and the incidence of infectious diseases, and does not discuss the potential effects of long-term climate change on disease transmission. For our project, we can focus more on finding the long-term trends or patterns of climate change on disease since it is what is missing in the report because of the insufficient database. This suggests that we need to check on the limitation and availability of the data sets we use for the project.

[7] Prillaman, McKenzie. "Climate Change Is Making Hundreds of Diseases Much Worse." Nature News. Nature Publishing Group, August 12, 2022 https://www.nature.com/articles/d41586-022-02167-z#: :text=Increases%20in%20temperature%20and%20rainfall,waterborne% A study published in Nature Climate Change has found that climate change has exacerbated more than 200 infectious diseases and dozens of non-transmissible conditions. Camilo Mora, a data scientist at the University of Hawaii at Mānoa, and his colleagues examined 77,000 research papers, reports and books for evidence of how climate change has affected infectious diseases, finding that climate hazards bring people and disease-causing organisms closer together, leading to a rise in cases. Global warming can also make some conditions more severe and affect how well people fight off infections. The study quantifies the many ways in which climate change affects human diseases, according to Mora. For our project, we can learn the usage of figures in this article to demonstrate the relationships between climate change and various diseases.

II. Data

The data being used is the Climate data from 2020-2022 in the USA, and the notified diseases in the USA from 2020-2022
(Climate data from NOAA)
https://www.ncdc.noaa.gov/cdo-web/datasets
(Notified diseases by CDC)
https://data.cdc.gov/NNDSS/NNDSS-Weekly-Data/x9gk-5huc
Climate Data has 1728 records and 12 features while Disease Data has 8046 records and 35 features. The combined full dataset has 1728 records and 28 features.
The Climate Data updated monthly and the Disease Data updated weekly
For Climate Data, it can be imported directly by its url link using the Pandas function 'read.csv'; the Disease data would need to use the web-scraping method.

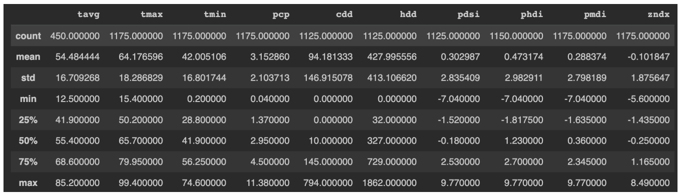


	state	year	month	tavg	tmax	tmin	pcp	cdd	hdd	pdsi	...	alzheimer_disease_g30
0	Alabama	2020	01	49.5	59.8	39.2	7.44	21.0	524.0	1.70	...	202.0
1	Alabama	2020	02	51.0	61.2	40.7	10.37	9.0	429.0	2.99	...	262.0
2	Alabama	2020	03	63.5	73.9	53.2	6.18	90.0	173.0	2.34	...	243.0
3	Alabama	2020	04	62.0	74.6	49.4	7.10	37.0	155.0	2.71	...	254.0
4	Alabama	2020	05	68.9	80.8	57.0	3.64	142.0	42.0	2.36	...	264.0

Fig. 1. Example of full data.

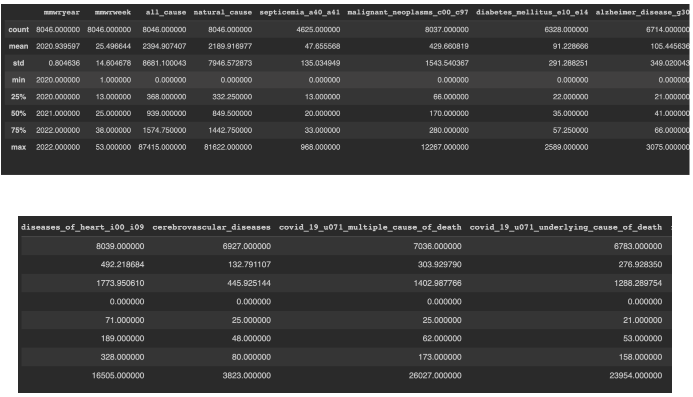
III. Exploratory Data Analysis

We performed some exploratory data analysis on the data: The sizes of the data are as follows (Climate, disease and full respectively)



	tavg	tmax	tmin	pcp	cdd	hdd	pdsi	phdi	pmdi	zndx
count	450.000000	1175.000000	1175.000000	1175.000000	1125.000000	1125.000000	1125.000000	1150.000000	1175.000000	1175.000000
mean	54.484444	64.176596	42.005106	3.152860	94.181333	427.995556	0.302987	0.473174	0.288374	-0.101847
std	16.709268	18.286829	16.801744	2.103713	146.915078	413.106620	2.835409	2.982911	2.798189	1.875647
min	12.500000	15.400000	0.200000	0.040000	0.000000	0.000000	-7.040000	-7.040000	-7.040000	-5.600000
25%	41.900000	50.200000	28.800000	1.370000	0.000000	32.000000	-1.520000	-1.817500	-1.635000	-1.435000
50%	55.400000	65.700000	41.900000	2.950000	10.000000	327.000000	-0.180000	1.230000	0.360000	-0.250000
75%	68.600000	79.950000	56.250000	4.500000	145.000000	729.000000	2.530000	2.700000	2.345000	1.165000
max	85.200000	99.400000	74.600000	11.380000	794.000000	1862.000000	9.770000	9.770000	9.770000	8.490000

Fig. 2. Numerical summary of climate data.



	mmwryear	mmwrweek	all_cause	natural_cause	septicemia_a40_a41	malignant_neoplasms_c00_c97	diabetes_mellitus_e10_e14	alzheimer_disease_g30
count	8046.000000	8046.000000	8046.000000	8046.000000	4625.000000	8037.000000	6328.000000	6714.000000
mean	2020.939597	25.496644	2394.907407	2189.916977	47.655568	429.660819	91.228666	105.445636
std	0.804636	14.604678	8681.100043	7946.572873	135.034949	1543.540367	291.288251	349.020043
min	2020.000000	1.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
25%	2020.000000	13.000000	368.000000	332.250000	13.000000	66.000000	22.000000	21.000000
50%	2021.000000	25.000000	939.000000	849.500000	20.000000	170.000000	35.000000	41.000000
75%	2022.000000	38.000000	1574.750000	1442.750000	33.000000	280.000000	57.250000	66.000000
max	2022.000000	53.000000	87415.000000	81622.000000	968.000000	12267.000000	2589.000000	3075.000000

diseases_of_heart_i00_i09	cerebrovascular_diseases	covid_19_u071_multiple_cause_of_death	covid_19_u071_underlying_cause_of_death
8039.000000	6927.000000	7036.000000	6783.000000
492.218684	132.791107	303.929790	276.928350
1773.950610	445.925144	1402.987766	1288.289754
0.000000	0.000000	0.000000	0.000000
71.000000	25.000000	25.000000	21.000000
189.000000	48.000000	62.000000	53.000000
328.000000	80.000000	173.000000	158.000000
16505.000000	3823.000000	26027.000000	23954.000000

Fig. 3. Numerical summary of disease data.

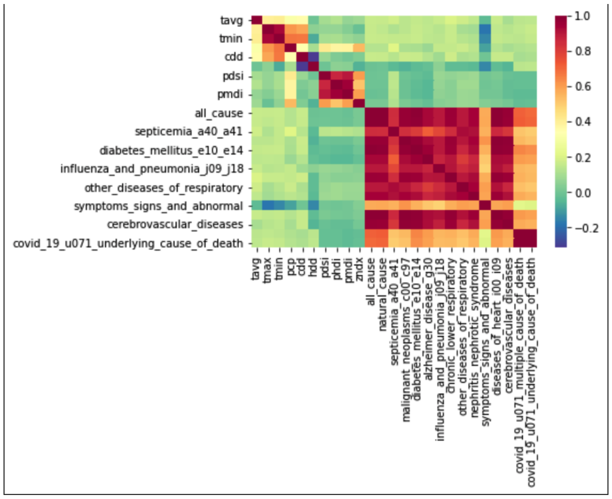


Fig. 4. Heatmap of data.

Figure 4 and figure 5 show the heatmap and correlation and seem to indicate that there exists little to no correlation between climate and disease.

Figure 6 shows the boxplot of all deaths and shows us potential outliers.
Figure 7 shows that the correlation between deaths and average temperature is near zero as the regression line is straight and has almost zero slope.
Figures 8 and 9 are the histogram and boxplot for average temperature (one of our predictors) for further analysis.
Figure 10 is the overarching description of the entire dataset.

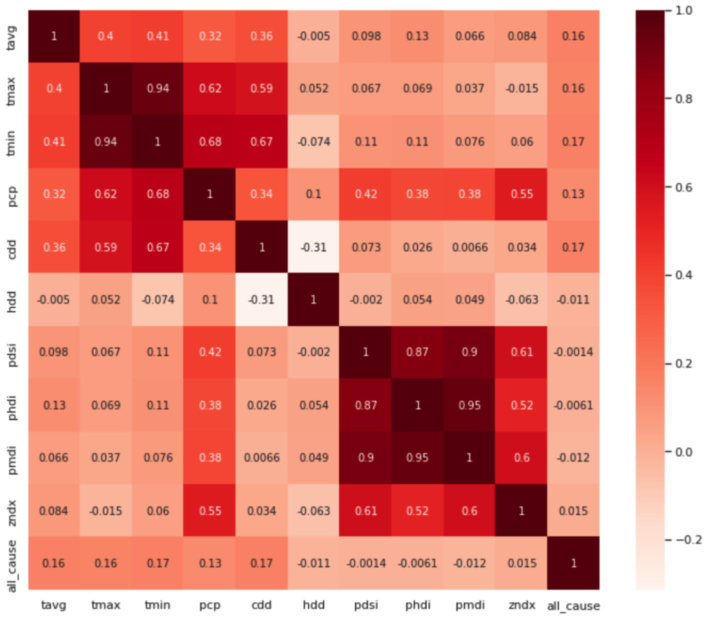


Fig. 5. Correlation matrix of data.

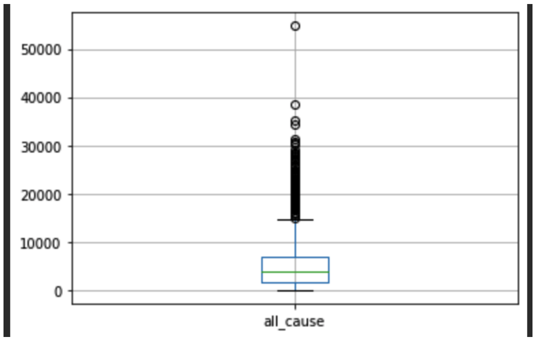


Fig. 6. Boxplot of all deaths

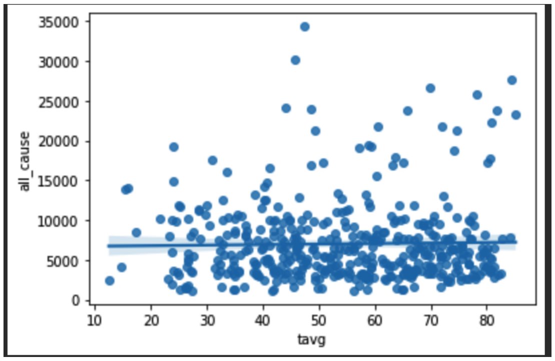


Fig. 7. Scatterplot of average temperature with all deaths with Linear regression

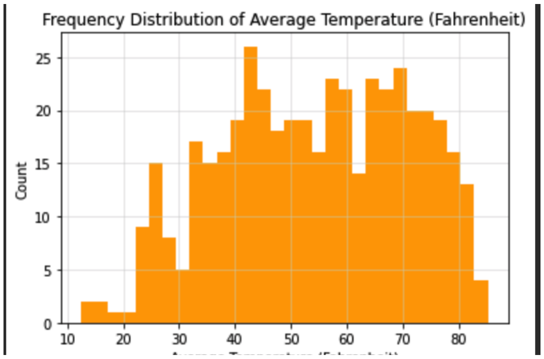


Fig. 8. Histogram of average temperature

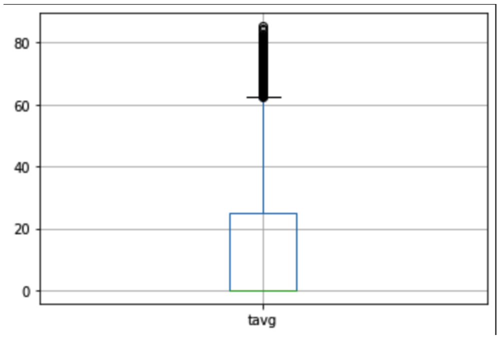
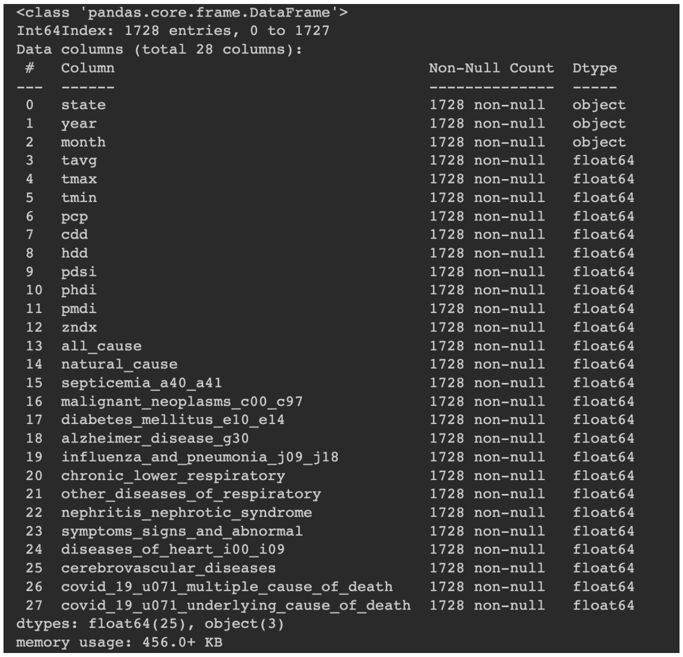


Fig. 9. Boxplot of average temperature



```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 1728 entries, 0 to 1727
Data columns (total 28 columns):
 #   Column                                Non-Null Count  Dtype
---  ------                                --------------  -----
 0   state                                 1728 non-null   object
 1   year                                  1728 non-null   object
 2   month                                 1728 non-null   object
 3   tavg                                  1728 non-null   float64
 4   tmax                                  1728 non-null   float64
 5   tmin                                  1728 non-null   float64
 6   pcp                                   1728 non-null   float64
 7   cdd                                   1728 non-null   float64
 8   hdd                                   1728 non-null   float64
 9   pdsi                                  1728 non-null   float64
 10  phdi                                  1728 non-null   float64
 11  pmdi                                  1728 non-null   float64
 12  zndx                                  1728 non-null   float64
 13  all_cause                             1728 non-null   float64
 14  natural_cause                         1728 non-null   float64
 15  septicemia_a40_a41                    1728 non-null   float64
 16  malignant_neoplasms_c00_c97           1728 non-null   float64
 17  diabetes_mellitus_e10_e14             1728 non-null   float64
 18  alzheimer_disease_g30                 1728 non-null   float64
 19  influenza_and_pneumonia_j09_j18       1728 non-null   float64
 20  chronic_lower_respiratory             1728 non-null   float64
 21  other_diseases_of_respiratory         1728 non-null   float64
 22  nephritis_nephrotic_syndrome          1728 non-null   float64
 23  symptoms_signs_and_abnormal           1728 non-null   float64
 24  diseases_of_heart_i00_i09             1728 non-null   float64
 25  cerebrovascular_diseases              1728 non-null   float64
 26  covid_19_u071_multiple_cause_of_death 1728 non-null   float64
 27  covid_19_u071_underlying_cause_of_death 1728 non-null   float64
dtypes: float64(25), object(3)
memory usage: 456.0+ KB
```

Fig. 10. Overarching description

IV. Method

We first used Recursive Feature Elimination to remove all features but 8. The predictor variables used are year, month, minimum temperature, precipitation, Palmer Hydrological Drought Index, Palmer Modified Drought Index, Palmer Drought Severity index, and Palmer Z index. The response variable is all deaths caused during that particular month.

The first model we used is a linear model. Even though all the assumptions were not satisfied, we decided to go with this first to see how it performed. The slopes were [674.22, -82.8, 27.75, 6.09, 2.35, 141.92, 5.58, 0.44, -223.55, -125.09, 269.2, -17.06] and the intercept was -1357962.47 on the test set. This assumes that there is a linear relationship between the predictor variables and the response variable.

The second model we tried was a Random Forest Regression model with 8 predictor variables and 100 estimators. This constructs a multitude of decision trees at training time and then takes a combination of the best ones for results.

The third model we tried was a Gradient Boosting Regression model with 8 predictor variables, learning rate 0.1, alpha 0.9, and 100 estimators. This is based on a group of machine learning algorithms that combine many weak learning models together to create a strong predictive model.

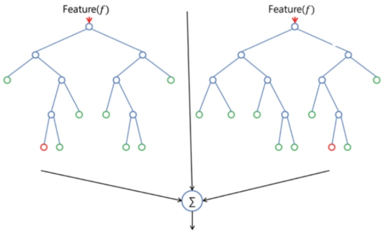


Fig. 11. Random Forest Regression

These two models seem to be a better approach as shown in the following picture.

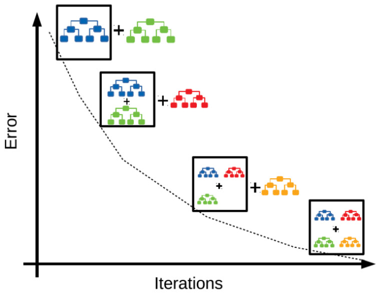


Fig. 12. Gradient Boosting

The assumptions of a linear relationship are not satisfied completely by the predictors and response, and decision trees are a better way to predict these relationships.

Homoscedasticity means that the residuals have equal or almost equal variance across the regression line. By plotting the error terms with predicted terms we can that the observations are not evenly distributed in the plot, so which means the residuals does not have constant variance at every level of x. This is one of the key assumptions that was violated. For the Normality of residuals, The residual terms are pretty much normally distributed for the number of test points we took. In which case, it meets the condition of Normality

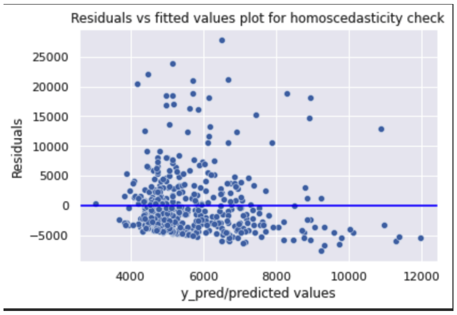


Fig. 13. Homoscedasticity check

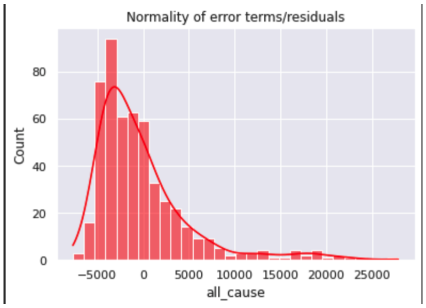


Fig. 14. Normality of error

There could have been alternative approaches taken:
Alternative models could have been tried with different parameters.
Different datasets could have been chosen.
These models could have been made with specific diseases as the response rather than all deaths.
All these methods might have changed our results.

We will be using RMSE and R-squared scores to assess our models as they can provide a quantifiable measuere to show how well our predictor variables explain our response

V. Results/Discussion

A data point first goes through the cleaning stage. NaN's are replaced with zero and it is converted to monthly and combined. Recursive feature extraction is applied and then the necessary parameters are chosen. Then this point is passed through the various different models to get a predicted number of deaths which is then compared to the actual number of deaths to give scores.

Linear Model : We have a very high Root Mean Squared Error and low R-Squared score. This signifies that only

MAE	MSE	RMSE	Rsquared
3672.48	27164157.81	5211.92	0.05

TABLE I
RESULTS FOR THE LINEAR MODEL

around 5 percent of the variance of the response variable is explained by the variance of the predictor variables. This model is therefore not good enough to predict deaths and our initial assumptions were incorrect.

MAE	MSE	RMSE	Rsquared
3178.68	20723400.54	4552.3	0.28

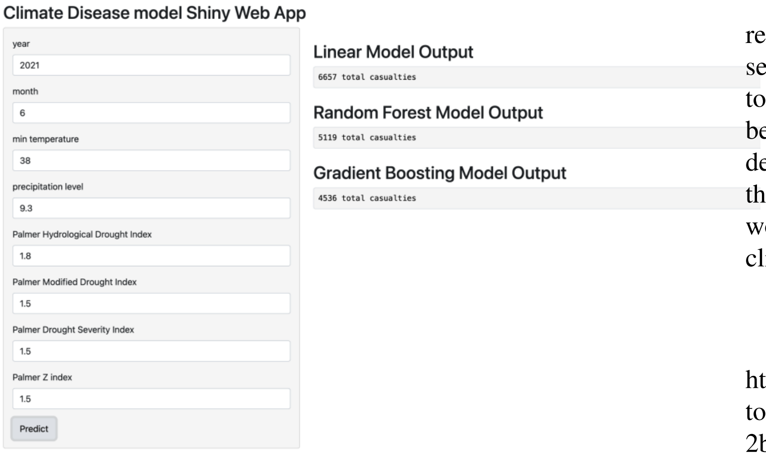
TABLE II
RESULTS FOR THE RANDOM FOREST MODEL

Random Forest Regressor : We have a high Root Mean Squared Error and low R-Squared score. This signifies that around 28 percent of the variance of the response variable is explained by the variance of the predictor variables. This model is therefore still not good enough to predict deaths and seems to indicate that the given predictors aren't enough to predict total deaths.

MAE	MSE	RMSE	Rsquared
3338.97	23639617.43	4862.06	0.18

TABLE III
RESULTS FOR THE GRADIENT BOOSTING MODEL

Gradient Boosting Regressor : We have a high Root Mean Squared Error and low R-Squared score. This signifies that only around 18 percent of the variance of the response variable is explained by the variance of the predictor variables. This model is therefore still not good enough to predict deaths and seems to indicate that the given predictors aren't enough to predict total deaths.



Climate Disease model Shiny Web App

year: 2021
month: 6
min temperature: 38
precipitation level: 9.3
Palmer Hydrological Drought Index: 1.8
Palmer Modified Drought Index: 1.5
Palmer Drought Severity Index: 1.5
Palmer Z index: 1.5

Linear Model Output
```
6657 total casualties
```

Random Forest Model Output
```
5119 total casualties
```

Gradient Boosting Model Output
```
4536 total casualties
```

Fig. 15. Shiny app interface

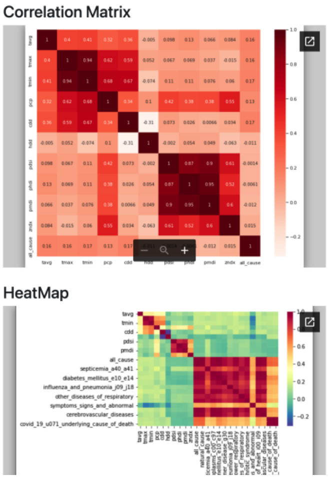

Correlation Matrix
tavg
tmax
tmin
pcp
cdd
hdd
pdsi
phdi
pmdi
zndx
all_cause
HeatMap
tavg
tmin
cdd
pdsi
pmdi
all_cause
septicemia_a40_a41
diabetes_mellitus_e10_e14
influenza_and_pneumonia_j09_j18
other_diseases_of_respiratory
symptoms_signs_and_abnormal
cerebrovascular_diseases
covid_19_u071_underlying_cause_of_death

Fig. 16. Shiny app interface - EDA

Our Shiny application is pretty simple to use. Decide on the predictor parameters on the left side. When a person clicks on the predict button, the app inputs the parameters into three different models and outputs the predicted response variable based on the inputs. The models are trained as soon as the app is started.

Below this, there are a couple of plots based on our data analysis that showcase more information about the predictors and the response.

VI. CONCLUSION AND FUTURE WORK

Based on our analyses through various models, we can conclude that the 8 climate and disease variables are insignificant in predicting mortality rate. We used 3 models which all indicated low correlation between our predictor and response variables.

This project allowed us to apply our newly developed web scraping knowledge. Additionally, we spent a lot of time exploring different models, and building our statistical background.

In the future, we would be interested in exploring these relationships with a different disease data source, as there seems to be no plausible connection between weather and total deaths based on our models. Another possibility would be to manipulate the current disease data set to turn disease death count columns into proportion columns by dividing the column by "allcause", the total death count column. It would be interesting to apply our current models to a different climate-disease data set, hopefully yielding stronger models.

VII. APPENDIX

The timeline of work:
https://online.officetimeline.com/shareable-link?
token=8WawCb3CjPoF2gHOHbBgPx92v2glJmQN%
2bTbcPOU0c8bKsyBuBc6bEFfG3r7Etik357US86pEqdCqeQiFYgFdPTA4
The Shiny app code can be found on :
https://drive.google.com/file/d/
1X5eLV0d9nHPqU322tUk5Y8Fojlk3oN3v/view?usp=sharing

The ipynb file can be found on :
https://colab.research.google.com/drive/1hwBbqW375CjPQCZ33FGXQXBheXZB8-1F?usp=sharing

VIII. Contribution

jortiz71 - 33 % ; chiragg4 - 33 % ; yihao4 - 34 %

yihao4 Built linear regression model and related works, associated in data clearing and data appearance as well as the project writing, etc.

chiragg4 made Web Script of climate data, data clearing and built Random Forest regression and Gradient Boosting Regression models; associated in building shiny dashboard and project writing, etc.

jortiz71 Data clearing and merging datasets, doing the exploratory data analysis and building the shiny app dashboard as well as the project writing, etc.

Everyone worked on the progress report together, doing their parts on time.