Team 16: Predict high-prone areas for future pandemics from past covid data.

Divya Shrivastava, Mansi Malegaonkar, Mukta Bisht, Raghavendra Srinivas, & Vasavi Govindaraju

Introduction:

The research topic will explore the effects of population density, weather (temperature, dew point, etc.), medical assistance, urban/rural, and healthy lifestyle (based on CDC data) on COVID-19 cases and use it to predict areas prone to higher cases and plan resource allocation.

Problem Statement:

We will use the COVID-19 dataset to identify potential features that could help accurately predict high-prone areas for future pandemics or epidemics with similar characteristics to COVID-19, which could impact public health, the economy, or the healthcare system. Predictions will help with resource allocation, intervention strategies, or policy decisions in the areas that may need the most. We want to point out that predicting future pandemics and epidemics is a challenging task that requires a multidisciplinary approach. We need to clearly understand where future pandemics will spread the most in different counties/regions in the US. We must also consider ethical and societal implications when using data to make decisions that affect people's lives. While machine learning can provide valuable insights, it is just one of many tools available to policymakers and medical authorities.

Business Justification:

Effective planning of resources and personnel (specifically medical front-line workers) is crucial to mitigate the pandemic and control transmission. Better resource management eventually leads to lower expenses overall for an organization since the cost is proactively predetermined and doesn't require reassessment/addition of resources for the future, which leads to an efficient workflow. Government officials and medical authorities can employ this prediction for proactive planning to reduce costs and have adequate resources and facilities. Our initial hypothesis is that there is a correlation between covid cases and weather, as well as covid cases and population density. We will add or remove variables during our analysis and evaluate model performance.

Literature Survey:

According to the study conducted by McKinsey for hidden costs during covid for every million people that seek treatment, the US health system will incur roughly $5.3 billion in direct costs, and every deferred treatment could make the condition 9% costlier to treat. Based on research, we have found that forecasting COVID-19 cases plays a crucial role in planning and supplying resources effectively. A study on COVID-19 death rates for six categories of urbanicity, ranging from the most urban to the most rural, showed that death rates were higher in urban areas (Curtin and Heron, 2022). Research has also demonstrated the influence of weather conditions and population density on COVID-19 by utilizing time-series data that did not include demographic and access to health providers in the region (Bhimala, Patra, Mopuri, and Mutheneni,2021). A study suggested that the virus spread is higher in colder weather and showed the negative impact of temperature and wind speed on the spread of the virus (Ganslmeier, Furceri, and Ostry, 2021). We aim to include these various factors to build a model to predict similar events in the future and help plan resources and access to medical providers.

Data and Variables:

The dataset is downloaded from Kaggle and consists of various columns at the county level in the timeframe from January to December 2020. The dataset has over 500K rows with 227 columns. It has multiple features like covid, health, population, housing, weather, crime, race, causes of death, and socio-economic factors. A great diversity of information is available from all columns that are potential predictors for the number of cases in a county/state. From our literature survey papers, we narrowed it down to a subset to build our hypotheses and study its relation to covid cases. We chose the dependent variable as 'cases', and the list of independent variables is shown in Table 1 below.

date	percent_adults_with_obesity	percent_rural
county	percent_physically_inactive	num_below_poverty
FIPS	num_primary_care_physician	num_age_65_and_older
total_population	primary_care_physicians_rate	num_minorities
population_density_per_sqmi	percent_adults_with_diabete	mean_temp
num_deaths	percent_65_and_over	max_temp
percent_fair_or_poor_health	percent_female	min_temp
average_number_of_physically_unhealthy_day	num_rural	

Table 1 – Starting features

The original number of observations/records is 790,331. After cleaning the data and dropping the null values, the observations decreased to 721,450, a 9% reduction. We checked the sanity of the data by filtering out data points to remove odd/out-of-range values to ensure we have no value out of the expected range for that column. For example, we had a few criteria in place for some of our features:

- The temperature columns cannot exceed 150 degrees.
- The percentage columns could only be between 0 and 100.
- The number of minorities, total physicians, rural population, deaths, and age greater than 65 must all be less than the county's total population.

The sanity check further reduced our data count to 628,036: a 20% reduction in the observations from the original dataset. We further confirmed that there was no duplication of the record counts. We split our data into 80% training and 20% testing, subsequently used in all our models.

Approach and Methodology:

During our Exploratory Data Analysis (EDA), using the plots of the geographic distribution of covid cases, we ascertained that the North-East region (especially New York) was the most impacted, followed by Florida. When we attempted to find the outliers, we found that New York data seemed like an outlier, as it was at a different proportion than others, meaning that the number of cases was significantly higher than in other states. Still, we decided to keep those data points because New York has a higher total population than other states. So, when we take the proportion of cases to the entire population, the percentage of cases in New York is like that in other states.

During COVID-19, many groups debated the effects of temperature on the spread of COVID-19 because, during the summer, the viruses are supposedly less prevalent than in winter. However, a comparison of minimum temperature geo-distribution (Chart 3) and COVID-19 geo-distribution (Chart 2) does not support that argument. Our initial analysis found a strong correlation between total population (and population density) to the COVID-19 cases spread, which was apparent in the NY case (Chart 1).

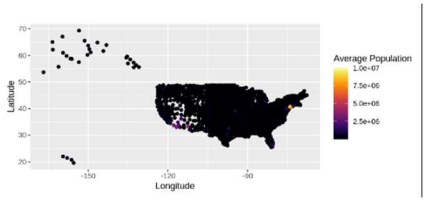

Chart 1: Avg. population distribution in US

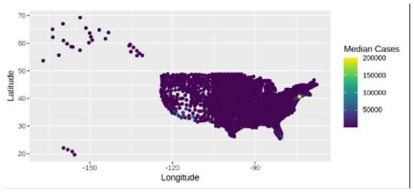

Chart 2: Median cases in US

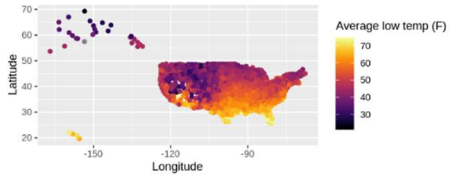

Chart 3: Average min temperatures across US

For the county-level data, as you can see from the correlation matrix below in Table 2, num_primary_care_physicians, num_age_65_and_older, total_population, and num_minorities fields showed multicollinearity. As a result, we dropped these columns.

A matrix: 13 × 13 of type dbl

	percent_adults_with_obesity	percent_physically_inactive	num_below_poverty	total_population	population_density_per_sqmi	percent_adults_with_diabetes	percent_fair_or_poor_health	percent_female	min_temp	average_number_of_physically_unhealthy_days	num_rural	cases	deaths
percent_adults_with_obesity	1.00000000	0.60333075	-0.17855745	-0.23749161	-0.21251270	0.54838541	0.45753291	0.04151063	0.13341134	0.43075485	-0.03668102	-0.15283856	-0.11374004
percent_physically_inactive	0.60333075	1.00000000	-0.15593472	-0.21755357	-0.16568316	0.60004478	0.58929898	0.03907936	0.23524755	0.54472303	-0.11021468	-0.11873207	-0.06462251
num_below_poverty	-0.17855745	-0.15593472	1.00000000	0.97413907	0.53398452	-0.11053256	-0.01134599	0.10315135	0.05079869	-0.04950472	0.17713080	0.75076822	0.66844550
total_population	-0.23749161	-0.21755357	0.97413907	1.00000000	0.52672365	-0.15366628	-0.08463853	0.10916544	0.04154442	-0.12127423	0.21096585	0.74332360	0.62855434
population_density_per_sqmi	-0.21251270	-0.16568316	0.53398452	0.52672365	1.00000000	-0.13846939	-0.07359643	0.12901473	0.02795880	-0.11196987	-0.06271751	0.46388744	0.60137945
percent_adults_with_diabetes	0.54838541	0.60004478	-0.11053256	-0.15366628	-0.13846939	1.00000000	0.50627745	0.12846110	0.19002279	0.50651361	-0.02293196	-0.09794871	-0.05917341
percent_fair_or_poor_health	0.45753291	0.58929898	-0.01134599	-0.08463853	-0.07359643	0.50627745	1.00000000	0.03636193	0.26994954	0.88351982	-0.12125572	-0.02172428	-0.02010631
percent_female	0.04151063	0.03907936	0.10315135	0.10916544	0.12901473	0.12846110	0.03636193	1.00000000	0.07285419	0.07976948	0.17491703	0.07728884	0.06414952
min_temp	0.13341134	0.23524755	0.05079869	0.04154442	0.02795880	0.19002279	0.26994954	0.07285419	1.00000000	0.17991296	0.01635573	0.02612634	0.02571328
average_number_of_physically_unhealthy_days	0.43075485	0.54472303	-0.04950472	-0.12127423	-0.11196987	0.50651361	0.88351982	0.07976948	0.17991296	1.00000000	-0.01331075	-0.06649378	-0.04660730
num_rural	-0.03668102	-0.11021468	0.17713080	0.21096585	-0.06271751	-0.02293196	-0.12125572	0.17491703	0.01635573	-0.01331075	1.00000000	0.09903685	0.02821225
cases	-0.15283856	-0.11873207	0.75076822	0.74332360	0.46388744	-0.09794871	-0.02172428	0.07728884	0.02612634	-0.06649378	0.09903685	1.00000000	0.78243815
deaths	-0.11374004	-0.06462251	0.66844550	0.62855434	0.60137945	-0.05917341	-0.02010631	0.06414952	0.02571328	-0.04660730	0.02821225	0.78243815	1.00000000

Table 2 – Correlation matrix of all features

For the weather-related attributes, we observed that only temperature was not providing us with much information, so we added the following weather-related factors to see if they give us some info: 'sea_level_pressure', 'visibility', 'wind_speed', 'max_wind_speed', 'wind_gust', 'precipitation', 'fog', 'rain', 'snow', 'hail', 'thunder', 'tornado'. From the correlation matrix of Table 3, we didn't see multicollinearity as an issue with these variables, and the correlation coefficient is very low for all of them, indicating that there is not much of a relationship between weather factors provided in the dataset and the number of cases by itself.

	min_temp	sea_level_pressure	visibility	wind_speed	max_wind_speed	wind_gust
min_temp	1.000000000	-0.287582235	0.0426868554	-0.122291384	-0.093677780	-0.074971424
sea_level_pressure	-0.287582235	1.000000000	0.0745524641	-0.195664190	-0.249184947	-0.268235583
visibility	0.042686855	0.074552464	1.0000000000	0.044427347	0.009567511	-0.009670061
wind_speed	-0.122291384	-0.195664190	0.0444273466	1.000000000	0.694405565	0.605178319
max_wind_speed	-0.093677780	-0.249184947	0.0095675113	0.694405565	1.000000000	0.893651245
wind_gust	-0.074971424	-0.268235583	-0.0096700607	0.605178319	0.893651245	1.000000000
precipitation	0.100923893	-0.110208474	-0.2952091060	0.014630866	0.068163882	0.099512221
fog	0.008194383	-0.023560206	-0.4817103826	-0.135089588	-0.019801702	-0.002034860
rain	0.161684680	-0.189804378	-0.2923245880	-0.023328734	0.145931790	0.211810577
snow	-0.325583319	0.034712303	-0.2606462145	0.136015010	0.086477357	0.080198135
hail	0.001907667	-0.024660613	-0.0080577536	0.011969926	0.039951375	0.047082188
thunder	0.285791737	-0.127751932	-0.1056472144	-0.104597131	0.167191115	0.228266790
tornado	0.005913296	-0.002886908	0.0006833717	-0.002666056	-0.001533911	-0.003167761
cases	0.047222121	0.030795102	-0.0023331813	-0.002879050	-0.018279307	-0.016924725

	precipitation	fog	rain	snow	hail	thunder
min_temp	0.100923893	0.008194383	0.1616846802	-0.325583319	0.0019076668	0.285791737
sea_level_pressure	-0.110208474	-0.023560206	-0.1898043783	0.034712303	-0.0246606128	-0.127751932
visibility	-0.295209106	-0.481710383	-0.2923245880	-0.260646215	-0.0080577536	-0.105647214
wind_speed	0.014630866	-0.135089588	-0.0233287337	0.136015010	0.0119699259	-0.104597131
max_wind_speed	0.068163882	-0.019801702	0.1459317904	0.086477357	0.0399513749	0.167191115
wind_gust	0.099512221	-0.002034860	0.2118105772	0.080198135	0.0470821880	0.228266790
precipitation	1.000000000	0.109641204	0.3068462056	0.006961773	0.0150376504	0.228232247
fog	0.109641204	1.000000000	0.1128301143	0.142463297	0.0132841838	0.128707485
rain	0.306846206	0.112830114	1.0000000000	0.036417491	0.0341953531	0.472826001
snow	0.006961773	0.142463297	0.0364174909	1.000000000	0.0092607871	-0.060734835
hail	0.015037650	0.013284184	0.0341953531	0.009260787	1.0000000000	0.059934730
thunder	0.228232247	0.128707485	0.4728260007	-0.060734835	0.0599347299	1.000000000
tornado	-0.001830286	-0.001646825	0.0069773520	-0.001139202	-0.0001675734	0.003296580
cases	0.006608352	-0.009171983	0.0003767263	-0.011169561	-0.0011135408	-0.006011816

	tornado	cases

Table 3 - Correlation matrix for weather related features

Since the data has a lot of inconsistencies, we aggregated it at the state level to see if it gives us further insights. For this purpose, we have taken most factors' averages and the median for cases at the state level.

In addition, we aggregated the data at the state level by date and aggregated factors using sum and average. We found that multi-collinearity was not an issue in this set of data. We verified the correlation matrix on this dataset and saw total_deaths (derived from deaths) and total_population (derived from the population) are the most correlated fields, followed by counties_reported (derived from fips with cases on the day), count_rural (derived from num_rural), and avg_min_temp (derived from min_temp). We decided not to use total_deaths in this case because total_deaths have a more casual relationship than total_cases.

	total_cases	total_deaths	total_pop	counties_reported	avg_min_temp	avg_age_65_and_older	avg_phy_unh	count_rural	avg_percent_adults_with_diabetes	avg_population_density_per_sqmi
total_cases	1.00000	0.76761	0.67182	0.44060	0.07512	0	-0.00535	0.05969	0.02334	0.11305
total_deaths	0.76761	1.00000	0.60663	0.29421	0.04845	0	-0.07529	0.14457	-0.03002	0.22511
total_pop	0.67182	0.60663	1.00000	0.65406	0.13508	0	-0.01407	0.14957	-0.00432	0.20973
counties_reported	0.44060	0.29421	0.65406	1.00000	0.13394	0	0.11939	-0.14363	0.18124	-0.09329
avg_min_temp	0.07512	0.04845	0.13508	0.13394	1.00000	0	0.20361	0.03170	0.36623	0.02335
avg_age_65_and_older	0.00000	0.00000	0.00000	0.00000	0.00000	1	0.00000	0.00000	0.00000	0.00000
avg_phy_unh	-0.00535	-0.07529	-0.01407	0.11939	0.20361	0	1.00000	0.02444	0.67176	-0.18948
count_rural	0.05969	0.14457	0.14957	-0.14363	0.03170	0	0.02444	1.00000	0.08232	-0.01561
avg_percent_adults_with_diabetes	0.02334	-0.03002	-0.00432	0.18124	0.36623	0	0.67176	0.08232	1.00000	-0.13855
avg_population_density_per_sqmi	0.11305	0.22511	0.20973	-0.09329	0.02335	0	-0.18948	-0.01561	-0.13855	1.00000

Table 4 – Correlation Matrix at state level

Overview of Modeling

For predicting total cases and high-prone areas, we tried models like Linear Regression, Log-linear regression, Log-log, GAM, GAM+PCA, and Random Forest, and below are our results.

Linear:

We built a linear model using date, total_pop, counties_reported, count_rural, avg_min_temp, and state on the rolled-up state data. From the model summary, we got that R2 for the model is .64, which tells us the model explains 64% of the variability in the data, and all the variables other than counties reported were significant. We observed that the std error for the model was relatively high. Heteroscedasticity is an issue when plotting graphs to see the model's fit (Chart 4). The normality assumption holds overall (Chart 5), but the data has very heavy tails indicating a concentration of most values around the tails.

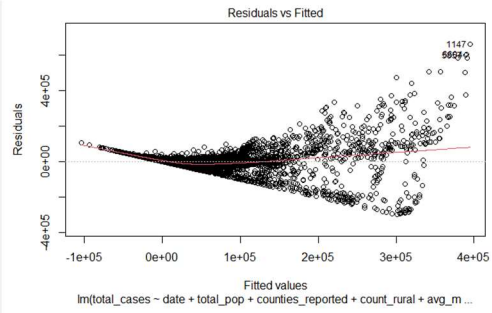

Chart 4 – Linear Regression – Residual vs fitted plot

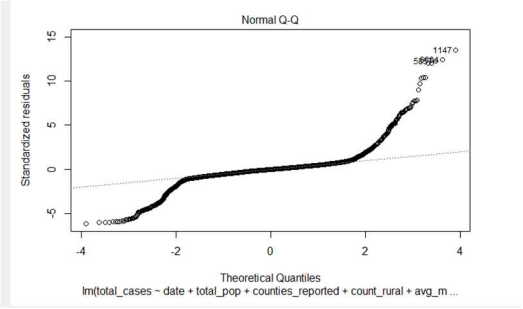

Chart 5 - Linear Regression – QQ plot

Since we saw heteroskedasticity, we tried to perform a box-cox transformation resulting in a recommendation for a log transformation, so we started with a log-linear model.

Log-Linear:

For the Log-linear model, we transformed the total_cases into a log while keeping all the predictors from above. After this transformation, we saw that most predictors, including "counties_reported", were significant. We also noticed that R2 improved to 0.77, meaning applying log transformation to the dependent variable increased the variability explained by the model to 77% of the data. We also saw that the Std error had dropped considerably after this, as seen in the summary table (Table 6)

When we plotted graphs to see the model's fit, we noticed that the heteroscedasticity was much better after the transformation, but we still saw signs of non-constant variance. The log transformation has improved the normality assumption, but we saw a heavy left tail.

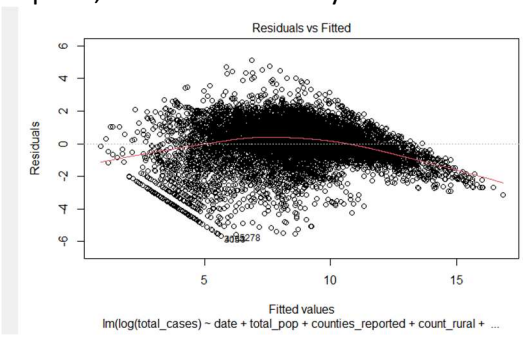

Chart 6 – Log Linear Model – Residual vs fitted plot

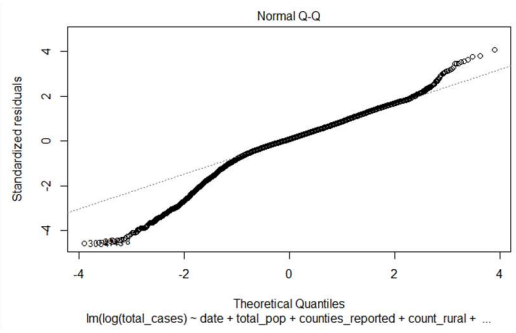

Chart 7 – Log Linear Model – QQ plot

We checked for the VIF from the above model; the counties_reported and count_rural had a VIF of a little higher even though within range, so we decided to try other models.

Log-Log:

We built a log–log model by taking the log of counties_reported and log(count_rural + 1) and observed that the R2 of the model improved to 0.858, which says that the model explains 85.8% variability in the data. We also noticed a reduction in Std error to 0.9859, with all the variables being significant.

When we plotted graphs to see the model's fit, we noticed that heteroscedasticity is no longer an issue with the model, and the fitted vs. Residual is almost a straight line which is what we want. Although we still have a heavier tail, the normality assumption holds; overall, the model looks like a good fit.

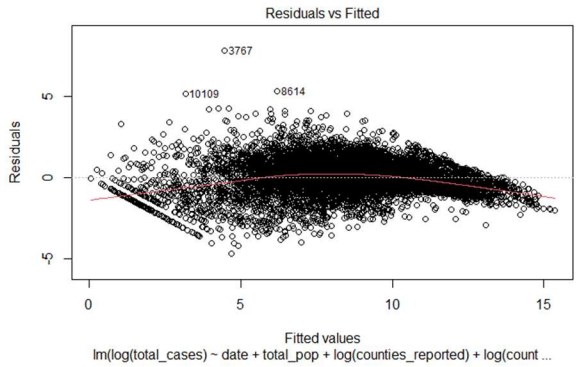

Chart 8 – Log Log Model – Residual vs fitted plot

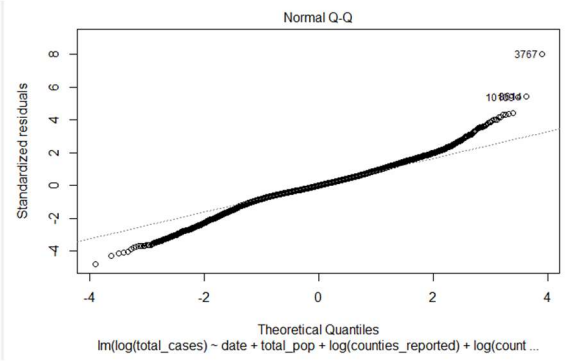

Chart 9 - Log Log Model – QQ plot

GAM:

We built a General Additive Model (GAM) to account for non-linear relationships between variables. In all cases, the results were inferior. For example, even after removing columns with multicollinearity, the VIF values for all the features were incredibly high, indicating that there was still multicollinearity between the features. From another correlation matrix, we saw a high correlation between avg_pcp (avg number of primary care physicians), avg_65_old (avg number of people 65 or older), and avg_num_below_poverty. We decided to remove these two columns: avg_num_below_poverty and avg_pcp. When we reran the GAM models, the VIF values were still incredibly high but lower than the previous run. Additionally, the R2 values were very high.

When we created a model with only the numeric values as predictors (avg_65_old, counties_reported, total_pop, avg_min_temp, avg_phy_unh, count_rural), the R-squared value was 0.584, which means that the model explains the variance in the response variable, "total_cases". We plotted residual vs. fitted for the GAM model, which shows the relationship between the predictor variable and the response variable after accounting for the nonlinear effects of other variables included in the model. We found that some predictors were not affected much by the s() function, whereas others were, as shown in the charts below. The s() function applies a smoothing function to the predictor variable, which allows us to capture nonlinear relationships. Below are snapshots of plots for two of the predictor variables used in the model, avg_min_temp (chart 10) and counties_reported (chart 11). Chart 10 shows that the data points are mostly close to 0 on the y-axis, indicating that the response variable does not vary much with the smooth effect of the predictor variable. Chart 10, on the other hand, shows a nonlinear relationship and indicates that it is well-fitted for this predictor variable. In the end, we decided this was not a great model to use due to the poor results we were obtaining.

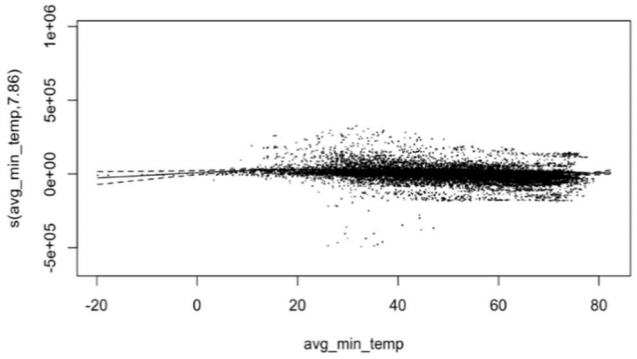

Chart 10 - Residual vs. Fitted

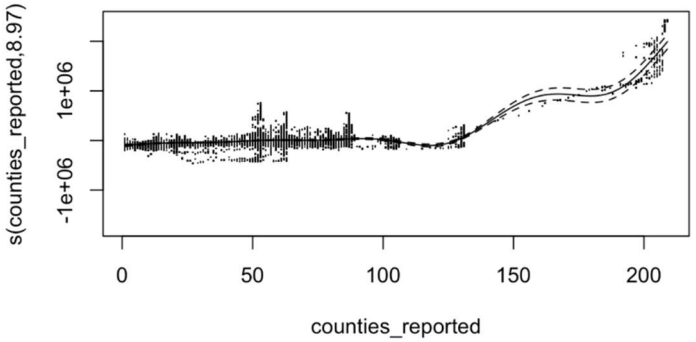

Chart 11 – Residual vs. Fitted

GAM + PCA:

Since GAM was not a very useful model, we performed Principal Component Analysis (PCA) on the data and then tried GAM by selecting the columns total_population, average_number_of_physically_unhealthy_days cases, min_temp, date, num_rural to accomplish this. First, we scaled the predictors and performed PCA. Before we created the GAM model, we checked the variance explained by each principal component using the Scree plot. We found that three principal components explained approximately 80% of the variance, so we used those three in our GAM model, with our response variable being the "cases" column. One thing to note is that, unlike the data we used for the GAM model above, we used

the original county-level data for this model. Unfortunately, after looking at the model, we found that the R-squared value was low, with a value of 0.383. As a result, we decided to move forward with other models instead. Chart 12 is a graph of variance proportions; Variance proportions show the graph we plotted for the variance proportion of each principal component and the summary below in Fig1 for the GAM model on the top four principal components.

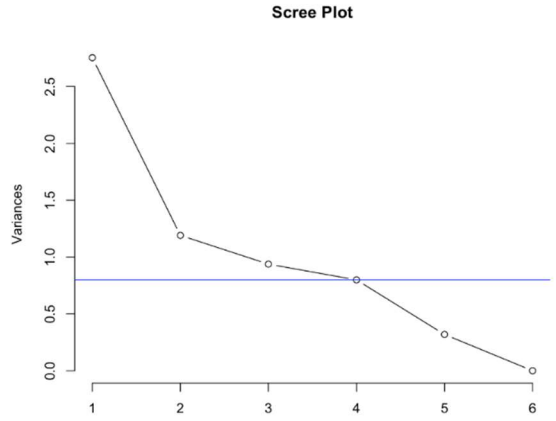

Chart 12 - Variance proportions

```
Family: gaussian
Link function: identity

Formula:
cases ~ s(PC1) + s(PC2) + s(PC3)

Parametric coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 0.0003708  0.0010088   0.368    0.713

Approximate significance of smooth terms:
        edf Ref.df     F p-value
s(PC1) 8.994  9.000 40822  <2e-16 ***
s(PC2) 8.960  8.999 40794  <2e-16 ***
s(PC3) 8.973  9.000 40619  <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

R-sq.(adj) =  0.383   Deviance explained = 38.3%
GCV = 0.61456  Scale est. = 0.61454   n = 603888
```

Fig1- Summary for GAM + PCA model

Random Forest:

We explored and fit the random forest model using the "ranger" package. In this model, we kept the data at the county level instead of aggregating it at the state level. Furthermore, unlike the other models, we decided that our response variable would be the "fips" code. For the random forest model, we considered 3 variables at each split, set the number of trees to 500, used the "permutation" method to compute variable importance, and set the seed value to 123. We predicted the test data from this model and got an RMSE value of 124.9852 (significantly higher than the log-log model). This high value indicated that the prediction values fell far from the measured true values. We also plotted a bar graph looking at variable importance. The top three were "num_age_65_and_older", "num_rural", and "total_population", respectively. When we compared the absolute difference between the predicted and true values, more than 80% of the records had a difference more significant than 5,000, so we did the same check with the predictions for the log-log model and found the results for that model to be better by 28.31%. Chart 13 shows the plot for variable importance from the random forest model. Using the variable importance, we built a weighted risk score and identified the top 10 counties at high risk, which is a first step towards planning resources and medical supplies.

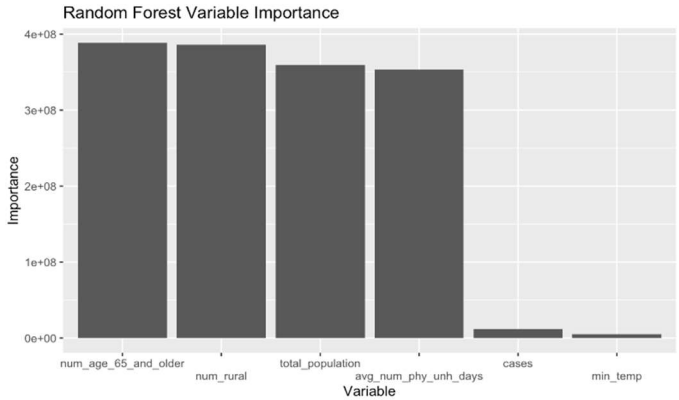

Chart 13 – Random Forest Variable Importance

	fips	county
1	04013	Maricopa
2	06037	Los Angeles
3	06059	Orange
4	06065	Riverside
5	06071	San Bernardino
6	06073	San Diego
7	12086	Miami-Dade
8	17031	Cook
9	48113	Dallas
10	48201	Harris

Table 5 – Top 10 Counties based on Risk Score

Model Summaries:

As detailed in the above section, we tried Linear, Linear-Log, GAM, GAM+PCA, and Random Forest models, and we compared them based on their R2, fit (provided above chart 4 -13), and standard error. Table 6 shows the results below, and we found that the log-log model was doing better in all three areas.

Model	R2	Standard Error	Preferred model
Linear	0.6868	48820	
Log Linear	0.7723	1.56	
Log-Log	0.858	0.972	Log-Log
GAM	0.584	N/A	
GAM+PCA	0.383	N/A	
Random Forest	0.9999147	11.1797	

Table 6 - Model Summaries

Novelty in Our Approach:

Our team came up with the concept of risk score for effectively predicting the counties and states that are high-prone to covid-like spread in the future. To develop the risk score formula, we used a correlation matrix to create the weightage for each predictor. We looked at the correlation of each predictor with num_cases, our response variable, to determine the weight. We formulated our risk score as (0.95 * total_pop + 0.62 * counties_reported + 0.26* avg_min_temp - 0.02* avg_phy_unh + 0.1* count_rural) / (10000*1.95). Using this risk score, we devised risk zones to classify the areas as high-risk, low-risk, or mid-risk. Since the risk score across all records ranged from 0.640 to 1057.157, we created each zone with a certain range to accommodate all the values in the risk score. To classify an area as “low-risk,” the risk score had to be 100 or lower, and as “mid-risk”, the risk score had to be greater than 100 and up to and including 700. Finally, to classify an area as “high-risk”, the threshold was 700. Classification is more practical than predicting the number of cases to conclude high-prone areas since this is a broader approach, and we do not need to have an exact value. Table 7 is a snapshot of our dataset after adding in risk_score and risk_zone.

state	date	total_cases	total_deaths	total_pop	counties_reported	avg_min_temp	avg_age_65_and_older	avg_phy_unh	count_rural	avg_percent_adults_with_diabetes	avg_population_density_per_sqmi	risk_score	risk_zone
Alabama	3/13/2020	3	0	749353	2	59.45	16083.97522	4.465109724	56221	13.35	377.1413242	36.79610552	Low Risk
Hawaii	5/26/2020	118	6	162456	1	63	16083.97522	3.1082267	22556	9.9	139.8637496	8.031063479	Low Risk
Kentucky	8/17/2020	1562	31	204306	4	63.025	16083.97522	5.378660143	21101	12.575	217.449372	10.06254148	Low Risk
South Carolina	10/30/2020	75521	1690	2093339	12	53.05833333	16083.97522	4.019004861	45394.83333	14.00833333	242.7919487	102.2170609	Mid Risk
Texas	6/9/2020	59582	1353	17868471	69	73.53043478	16083.97522	3.931137701	22461.97101	11.25652174	263.1395466	870.6336137	High Risk

Table 7 - Final Dataset

Things that did not work and Challenges:

While our initial hypothesis about the impact of weather on the number of covid cases, we saw a very weak relationship from the dataset, which we could not use for making predictions by itself. As a result, we used the features in conjunction with other predictors in our models.

While working with our data, we encountered a few challenges. For example, not all states had all the data for every county. The data set needed data for all the counties, so we had to aggregate the data at the state level. Because of this, we could not perform any analysis on FIPS codes since we no longer included data at the county level. Some counties also had missing data as well, so we found it best to aggregate the data at the state level. When we first tried to create a linear model at the county level, the data showed it was not a good fit; however, when we aggregated the data to the state level, we found that the results fit better.

While working on our data and creating our models, we identified a few things to fix. We initially wanted to model a relationship between weather-related factors and the number of cases. However, there needed to be more correlation, which we did not expect to catch. We faced a few conflicts about what approach to take and what steps to take. We tackled this issue by creating a baseline from which everyone would build different models and then compare our results to see which models performed better. While comparing our models, we had trouble determining which one was the best, specifically between random forest and log-log. Since Random Forest doesn't provide an R2, we used the RMSE value to compare the models. We wanted to compare the models in another way to confirm whether our RMSE values for the models made sense.

We created a “difference” column in both models to accomplish this. This column took the absolute difference between the predicted and actual values. Since the difference looked large, we checked the percentage of records that differed less than or equal to a specific value. For example, we studied the percentage of records that differed less than or equal to 5,000. When we compared the results between random forest and log-log, we found that only 19.27% of the records from the random forest model had such a difference. In contrast, the log-log models had 57.58% of the records with such a difference, indicating that the log-log model did a better job in predicting as compared to Random Forest

Another challenge we faced was deciding whether to add risk_score into our log-log model as a predictor. We added risk_score as one of the predictors in the log-log model to see if that helped us arrive at a better model. However, we found that adding risk_score did not help us improve the variability or the model predictions, so we decided not to use risk_score as a predictor in the model. Instead, we added the risk_score as a column and decided to come up with the zones as explained in the novelty approach section. Furthermore, we decided to add this column only after creating our model, and we used it to create the risk_zone column by classifying the records into their respective zones using the scores.

Conclusion:

From analyzing all the above models, we conclude that the log-log model performs best regarding model fit, R2, and standard error. Hence, we decided to use the log-log model as our final model. From our log-log model, we concluded that although there is a positive relationship between weather and covid cases, that relationship is relatively weak. We also concluded from the dataset that the population, the count of rural areas, and the number of counties reported impact the number of cases. From our log-log model, we devised risk zones to predict high-prone areas. Our model's predictors are date, state, total_population, counties_reported, count_rural, and avg_min_temp. Our response variable was total_cases. We looked more closely at one of our predictors, count_rural. From our summary, we found that if we increase the count of rural by 1%, the total number of cases increased the most (3.56%), keeping all other factors constant.

Since we were limited with time and resources to create the best possible predictive model for high-prone areas, we have some ideas on what we would like to do, time permitting. For one, we would like to gather more extensive data, at least over one year. It would also be beneficial to ensure the new data we find is clear without gaps, like the issues we faced with the county-level data. We want to explore further the possibility of interaction terms between the features. Furthermore, we would like to build a model using the data as well as the risk_score column, and we would like to explore the use of a classification model using risk zone classification. Finally, we would like to examine datasets of health events with similar characteristics and build a model using more effective terms like R-Naught (a value calculated for diseases that can spread). Ultimately, this would provide additional quantitative methods to predict.

Overall, we concluded that COVID prediction is a complicated subject requiring more data and factors than what was available to us. Our data produced a decent model that could be used as a good starting point and can be enhanced further. Using our final log-log model, we hope to help the government and the healthcare system plan efficiently for allocating resources and staffing medical personnel.

Works Cited

Bhimala KR, Patra GK, Mopuri R, Mutheneni SR. Prediction of COVID-19 cases using the weather integrated deep learning approach for India. Transbound Emerg Dis. 2022 May;69(3):1349-1363. doi: 10.1111/tbed.14102. Epub 2021 Apr 20. PMID: 33837675; PMCID: PMC8250893. https://pubmed.ncbi.nlm.nih.gov/33837675/

Curtin SC, Heron M. COVID-19 death rates in urban and rural areas: United States, 2020. NCHS Data Brief, no 447. Hyattsville, MD: National Center for Health Statistics. 2022. DOI: https://dx.doi.org/10.15620/cdc:121523. https://www.cdc.gov/nchs/products/databriefs/db447.htm

Ganslmeier, M., Furceri, D. & Ostry, J.D. The impact of weather on COVID-19 pandemic. *Sci Rep* **11**, 22027 (2021). https://doi.org/10.1038/s41598-021-01189-3 . https://www.nature.com/articles/s41598-021-01189-3

Martins-Filho PR. Relationship between population density and COVID-19 incidence and mortality estimates: A county-level analysis. J Infect Public Health. 2021 Aug;14(8):1087-1088. doi: 10.1016/j.jiph.2021.06.018. Epub 2021 Jul 3. PMID: 34245973; PMCID: PMC8253654. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8253654/

"Understanding the Hidden Costs of COVID-19'S Potential Impact on US Healthcare." *Www.Mckinsey.Com*, 4 Sept. 2020, www.mckinsey.com/industries/healthcare/our-insights/understanding-the-hidden-costs-of-covid-19s-potential-impact-on-us-healthcare. https://www.mckinsey.com/industries/healthcare/our-insights/understanding-the-hidden-costs-of-covid-19s-potential-impact-on-us-healthcare