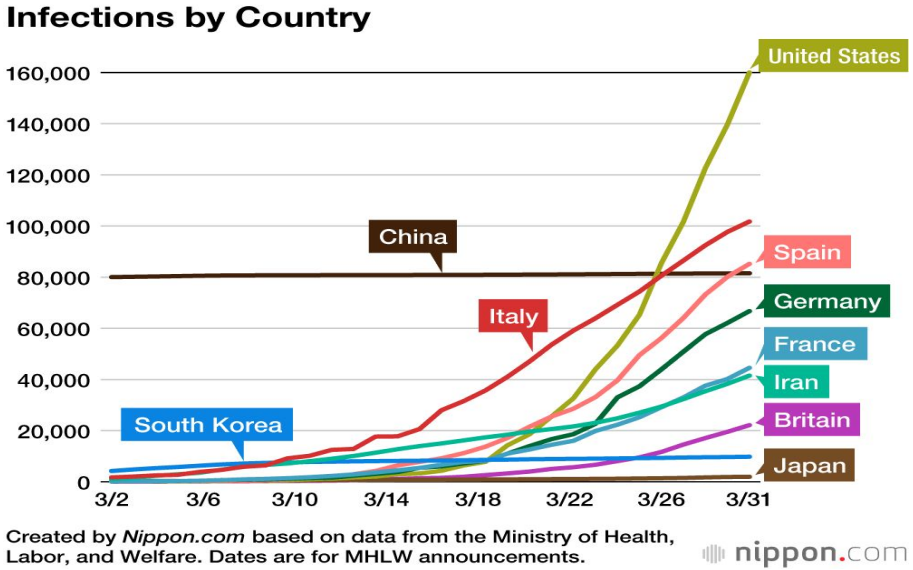
PROBLEM STATEMENT - **To predict daily covid19 cases at each county level for US public health client.**

Coronavirus disease 2019 (Covid 2019) is a fast spread infectious disease, which is caused by severe acute respiratory syndrome coronavirus 2 (SARS -CoV-2). It was first reported in Wuhan, China and since then its been spreading all over the world at very fast rates.



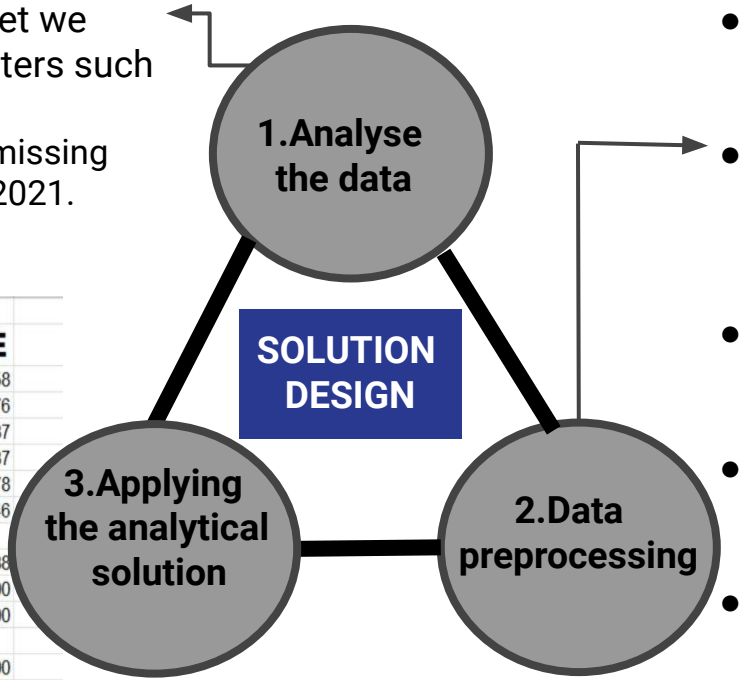
The daily new cases in county will help to-

- The future predictions of the confirmed cases will help in planning better of medical facilities.
- It will also be helpful in vaccine distribution drive.
- It will help the government to propose various public health interventions such as lockdown, social distancing, closing of schools etc. to slow down the spread of Covid-19.

How this can be done ?

- Machine learning models have been used to predict the number of confirmed cases of Covid-19 of each county of the state.
- From the dataset provided, we developed a set of algorithm that uses predictive analysis to identify the number of confirmed COVID-19 cases.
- It predicts the data set for 15 days for 47130 counties in the range (1001-56045) and 51 states in the range (1-56).

- Upon analysing the data set we found that certain parameters such as "S_D_dly_new_ test" and "S_D_cummulative_test" is missing from date 1/2/2021 - 12/2/2021.



- Univariate/bivariate/multivariate analysis of the variables is done first.
- Then correlation analysis of the features is done and also p-values(using t-test) of the variables are found.
- Then splitting the model into training and test sets on the basis of time.
- After that, we apply feature scaling, to remove variables which are highly correlated with.
- And the variables with lowest p-value is considered to be the most important parameter.
- Then we apply multiple linear regression model to predict the missing values after standardizing the values of each column.

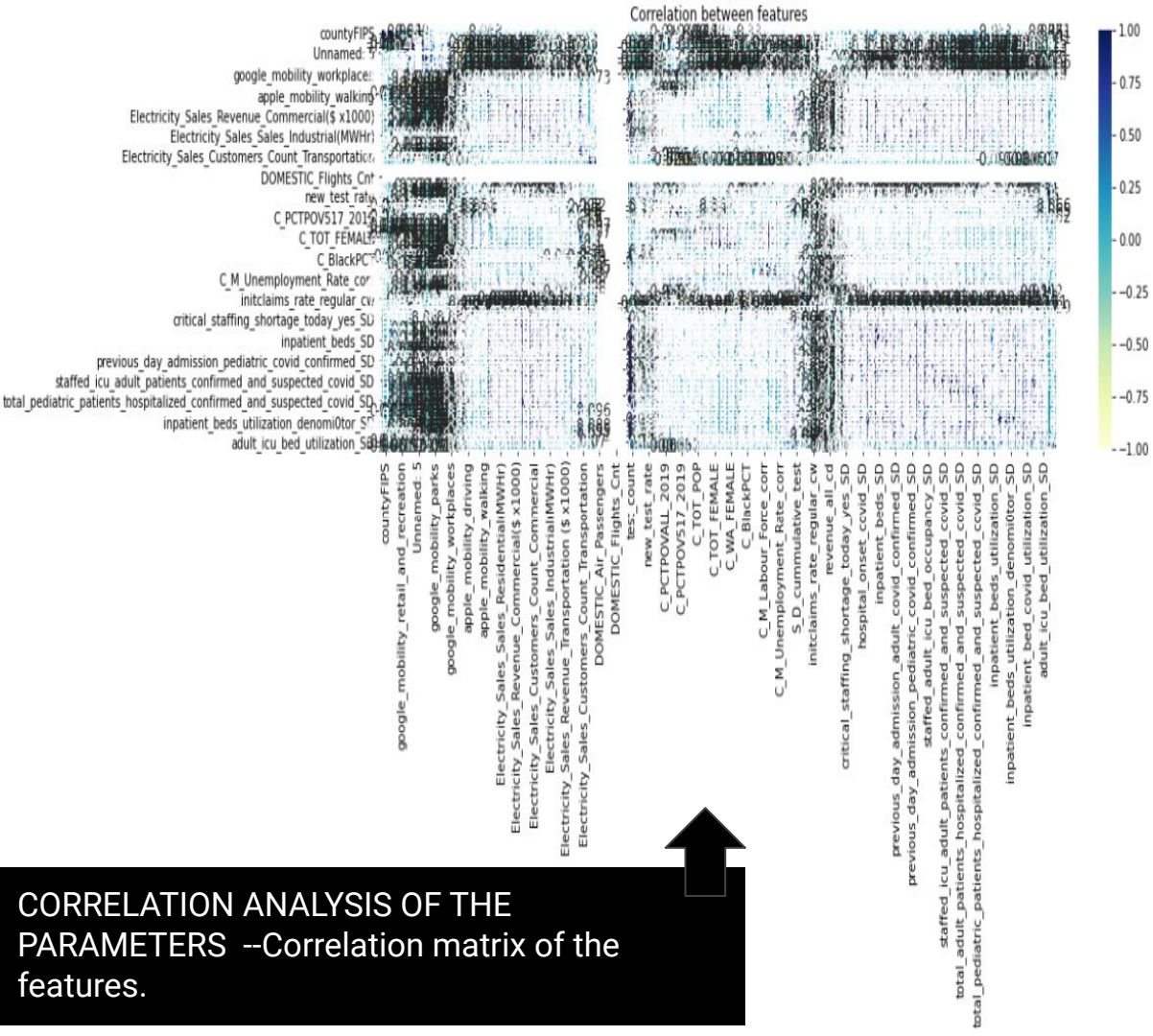
VARIABLES		P-VALUE
Google_mobility	1.retail and recreation	2.40E-258
	2.grocery_and_pharmacy	6.10E-176
	3.parks	6.10E-287
	4.transit-stations	1.46E-87
	5.workplaces	3.60E-278
	6.residential	6.30E-246
Apple_mobility	1.driving	4.20E-288
	2.transit-stations	0.00E+00
	3.walking	0.00E+00
Electricity_supply	1.Residential	0.00E+00
	2.Commercial	0.00E+00
	3.Industrial	0.00E+00
	4.Transportations	0.00E+00
yoy seated diner data		2.12E-49
inticlaims_count_regular_cw		1.04E-11
inticlaims_rate_regular_cw		2.05E-19
medhinic		2.20E-254
merchants_all_cd		0
revenue_all_cd		4.20E-274
spend_all_cd		0.583832

P-values for various variables wrt to "S_D_dly_new_ test" and S_D_cummulative_test" is given to find the significance between the parameters.

Lowest p-value - 4.20E-288
Variable -"Apple_mobility_driving "

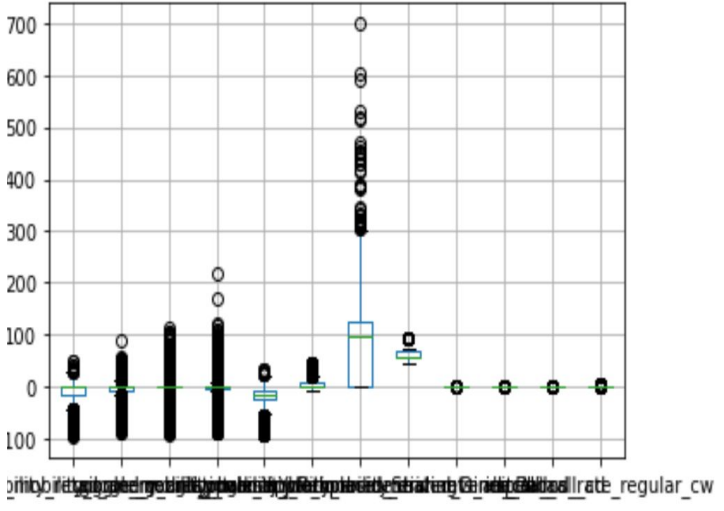
$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_i X_i$$

Y : Dependent variable
β_0 : Intercept
β_i : Slope for X_i
X = Independent variable



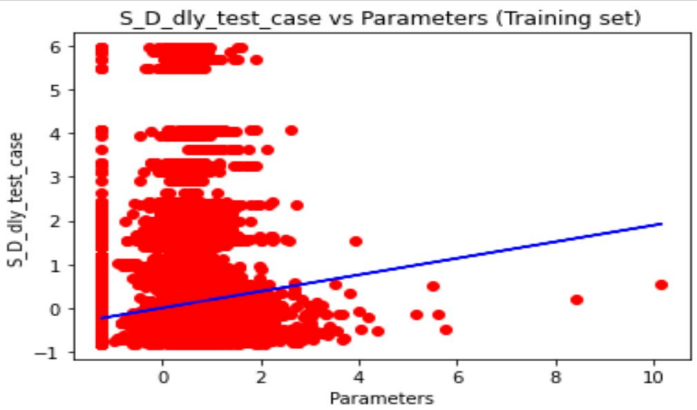
CORRELATION ANALYSIS OF THE PARAMETERS --Correlation matrix of the features.

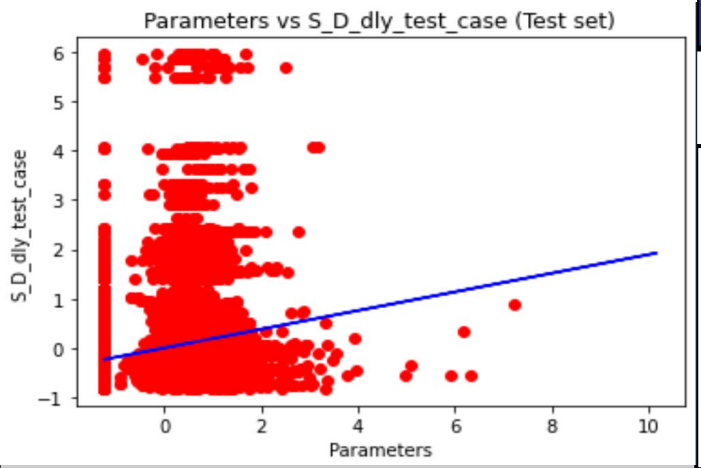
Boxplot parameters of various parameters such as Google_mobility, apple_mobility, yoy_seated_diner_data, inticlaims , revenues collected, spending , new shops open



FOR VALIDATION WE CALCULATED -

1. Mean Absolute error -> 0.683500%
2. Mean Squared Error -> 0.956616%
3. Root mean squared error -> 0.978%





ANALYTICAL APPROACH

SEPARATING THE PARAMETERS ON THE BASIS OF IMPORTANCE AND CORRELATION TEST BETWEEN THE FEATURES

- Patients who have tested positive the previous day(adult/pediatric/icu and non-icu patients included).)
- Patients who have tested positive with the onset of covid 19.
- Total tests conducted on that day
- Increase patients from the shortage of medical staffs.
- Previous day suspected patients.

Most useful parameters /top drivers of the model

Parameters where we assume certain probability

1.Google mobility data
2. Apple mobility data
3.ELECTRICITY CUSTOMERS COUNT
   - Residential - approx 0%
   - Industrial - approx 0.00001%(infected workers)
   - Transportation - approx 0.00005% (infected in public transportation)
   - Commercial - approx 0.00005% (cases from supermarkets, shops,malls etc)
4.YOY Seated diner data
   - Here we assume the chance of getting corona positive is (0.0000025/no. of days).

- Merchants_all_cd
- Revenue_all_cd
- Spend_all_cd
- C_TOT_MALE
- C_TOT_FEMALE
- C_MinorityPCT
- C_BlackPCT
- C_HispanicPCT

Parameters of least importance

→ MATHEMATICAL FORMULAS AND CALCULATIONS

$test = no.of\ confirmed\ tests * f$ (Laboratory confirmed tests conducted county wise)

$onset = no.of\ patients\ admitted * f$ (Patients with the onset of symptoms)

$previous\ day\ adult\ patients = previous\ day\ suspected\ adult\ patients * p * f$

. (Previous day adult who were suspected)

$previous\ day\ pediatric\ patients = previous\ day\ suspected\ pediatric\ patients * p * f$

. (Previous day pediatric who were suspected)

$total\ adult = (confirmed\ adult\ pateints * f) + (suspected\ ad.patients * p * f)$

. (Adult Patients who were confirmed or suspected today)

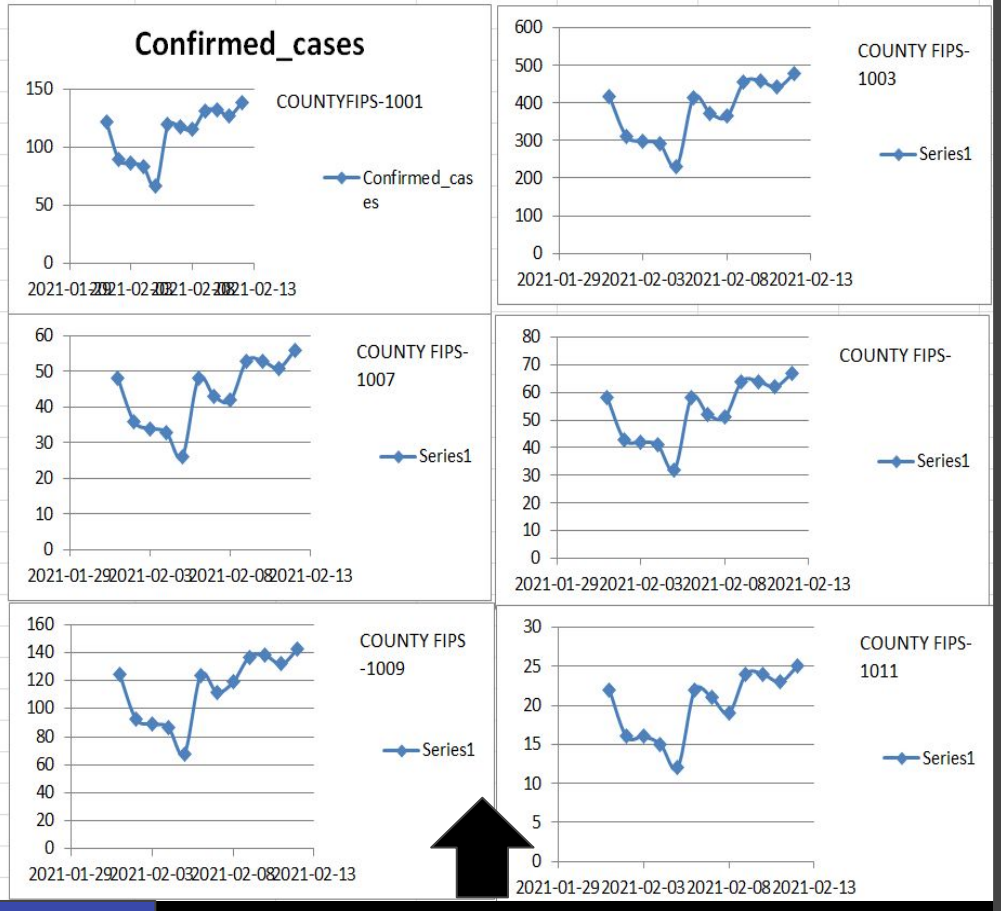
$total\ pediatric = (confirmed\ pediatirc\ patients * f) + suspected\ ped.patients * p * f)$

. (Pediatric Patients who were confirmed or suspected today)

$total\ confirmed\ cases =$
$test + onsetpatients + previous\ day\ adult + previous\ day\ pediatric + total\ adult +$
$total\ pediatric$ (Summation of all the above parameters)

$$f = \frac{population\ of\ that\ county}{total\ population\ of\ the\ state\ in\ which\ that\ county\ is\ located}$$

$$p = \frac{new\_test\_rate * total\_population\_of\_the\_state}{100000 * S\_D\_dly\_new\_test}$$



Confirmed cases for 15 days(1/2/2021 -15/2/2021) for counties with FIPS - 1001,1003,1007,1005,1009,1011

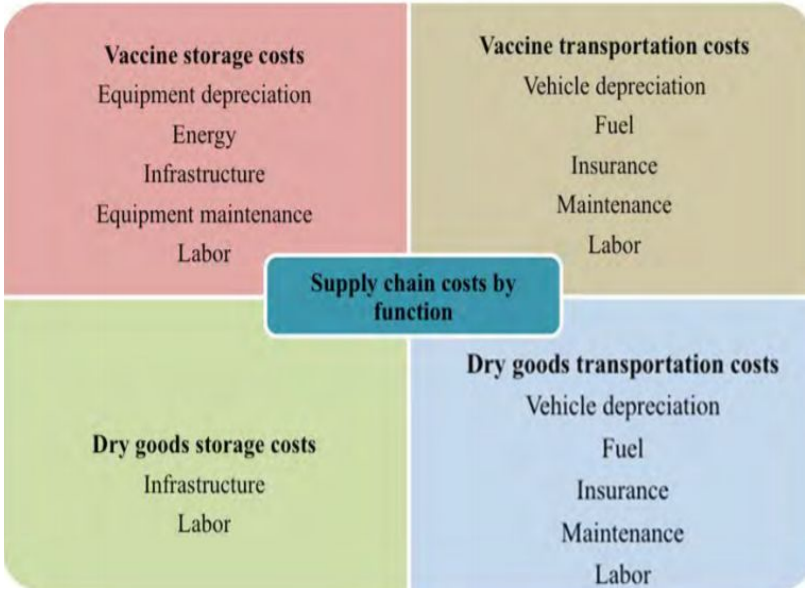
TOP 5 STATES TO BE WORST AFFECTED BY COVID-19 PANDEMIC BY 30/04/2021

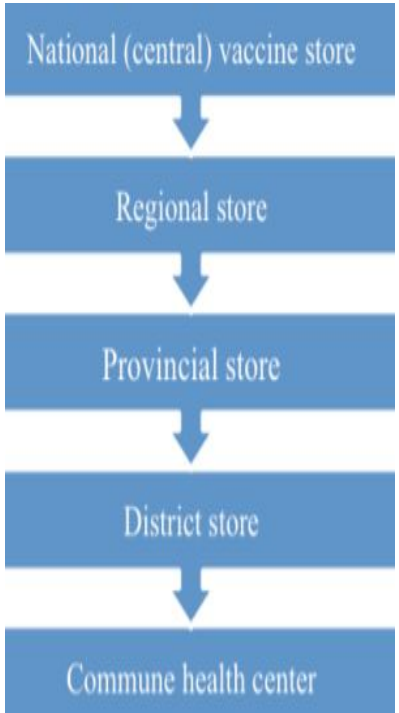
1. countyFIPS 56
2. countyFIPS 50
3. countyFIPS 25
4. countyFIPS 36
5. countyFIPS 9

ADDITIONAL VARIABLES TO IMPROVE THE MODEL PREDICTION ACCURACY

1. The percentage of people following the lockdown rules (wearing mask/social distancing etc) should be given for each county. With strict rules confirmed cases would be less.
2. The age distribution of the population for each county should also be given. The risk of getting infected with COVID-19 increases with **age**.

ADDITIONAL ANALYSIS





VACCINE DISTRIBUTION STRATEGY

1. Regional , provincial and district store should be selected in a county where labour cost is low (unemployment rate is high) and cheap electricity is available.
2. Accurate volume flow of covid vaccine should be there to avoid wastage and also the maximum capacity of vehicle to be used.
3. Vaccine should be available to all the people irrespective of their race,caste, religion, gender , salary ,status etc
4. Prioritizing mobile workforce and tracking systems should be used to minimize the number of new cases during vaccination distribution.