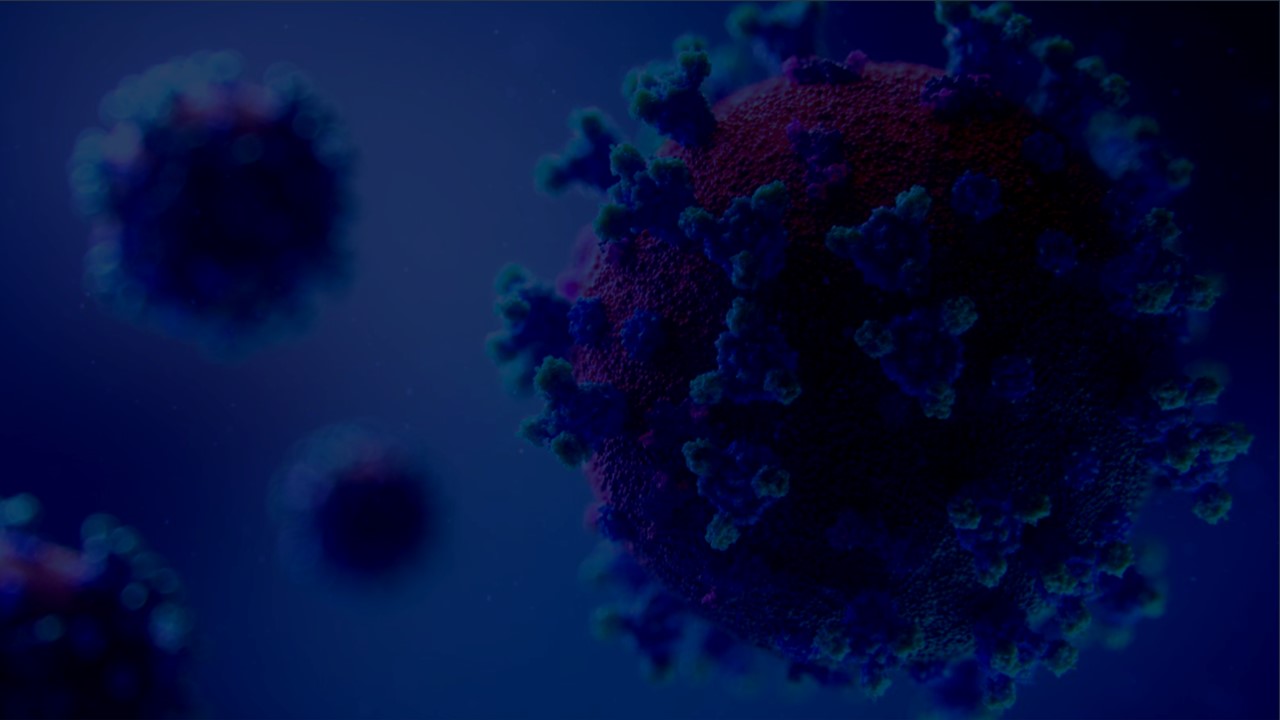
|  |
| --- |
| **COVID 19 US**  January 2020 – June 2021 |
| DANA 4800-001  Project Part II |



**Team 6**

Priti Ranjan Sahoo, Andrea Chan, Valeria Velez

Table of Contents

**[1.](#_Toc100219593)****[INTRODUCTION](#_Toc100219593)** [2](#_Toc100219593)

[2.](#_Toc100219594) **[MISSING VALUES AND ACCURACY](#_Toc100219594)** [3](#_Toc100219594)

**[3.](#_Toc100219595)****[ANALYSIS OF FEATURES](#_Toc100219595)** [4](#_Toc100219595)

**[4.](#_Toc100219596)****[PREDICTIVE MODEL](#_Toc100219596)** [11](#_Toc100219596)

**[5.](#_Toc100219597)****[SUMMARY](#_Toc100219597)** [11](#_Toc100219597)

**[6.](#_Toc100219598)****[REFERENCES](#_Toc100219598)** [12](#_Toc100219598)

**[7.](#_Toc100219599)****[APPENDIX 1 – TABLEAU LINK](#_Toc100219599)** [13](#_Toc100219599)

**[8.](#_Toc100219600)****[APPENDIX 2 – SAS CODE](#_Toc100219600)** [14](#_Toc100219600)

# **INTRODUCTION**

The aim of the second part of the project is to conduct a further analysis on the features available in the Girhub dataset, identifying the possible relationship with Covid 19 cases and deaths in the United States, and to create an accurate prediction model on the Covid cases.

The available data contains in total 3 Covid waves in the country, from January 2020 to June 2021, where a significant increase is observed in each of the Covid 19 waves. In addition, under the administration of President Biden in 2021 a decrease in cases is observed, including an increase of the fully vaccinated residents, in line with his government ambitious goal of 100 million people in his first 100 days of administration[[1]](#footnote-0).

Chart, line chart

Description automatically generated

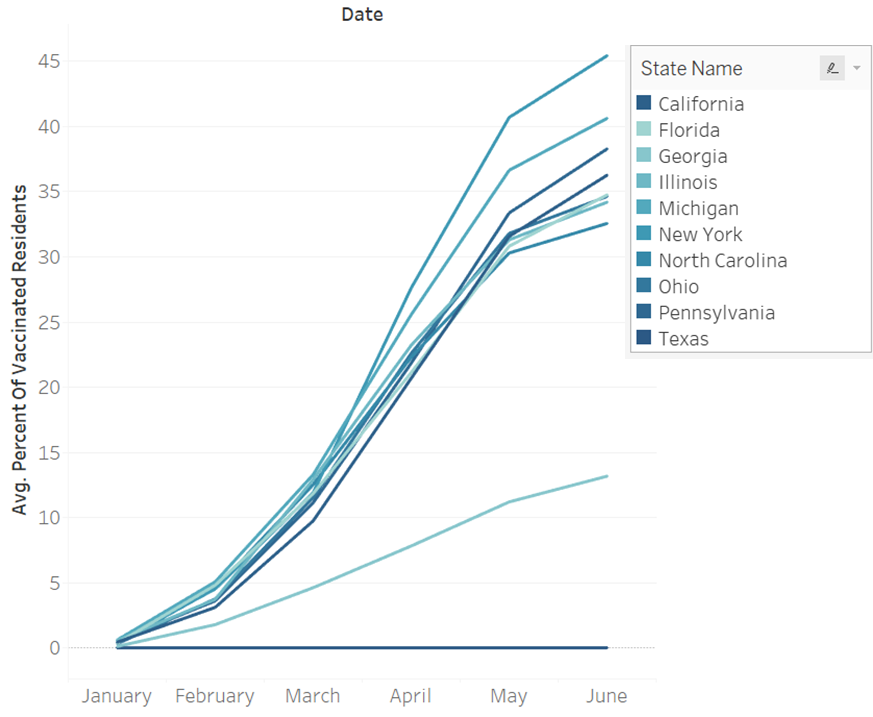
# **MISSING VALUES AND ACCURACY**

It was found that the GitHub dataset contained in total 1.961 counties from 3.142 total counties in the United States, meaning that 62.3% is missing. However, the data was considered representative since there is just missing 11% of the total US Population (total population census 2020: 331.449.281, dataset: 294.167.800). The counties with missing data are clearly defined on the map:

Map

Description automatically generated

Apparently, there were many uncounted deaths and covid cases in the US during the pandemic, mainly in the rural areas with less coverage of health care and professionals[[2]](#footnote-1). For example, it was found in the dataset that Texas, the second most populous state in the country, did not report any fully vaccinated residents, which can be seen in the graph below.



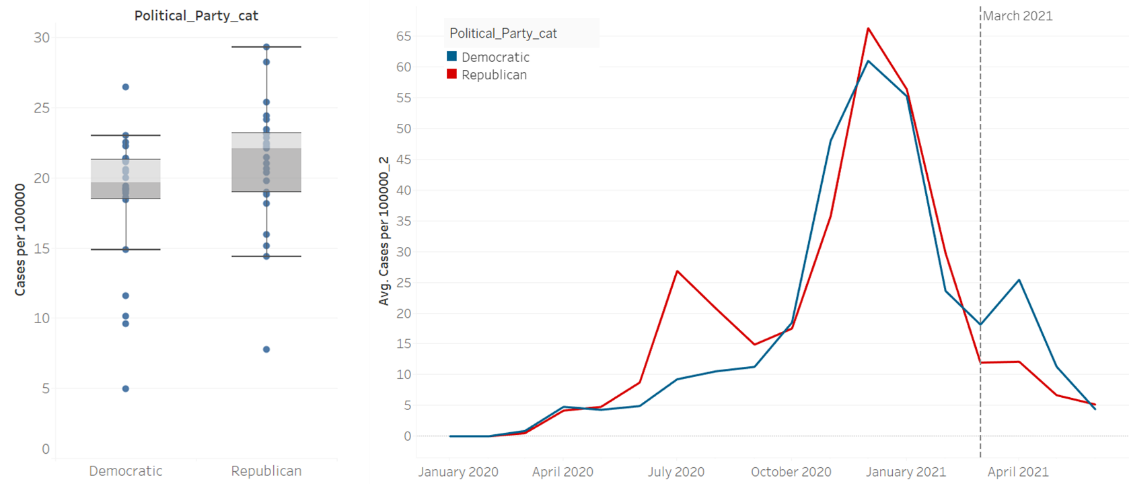
# **ANALYSIS OF FEATURES**

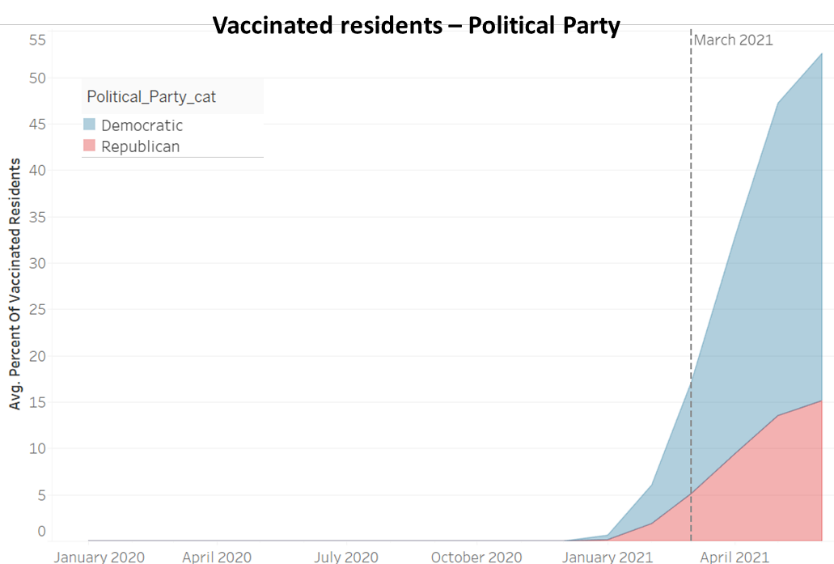
**3.1 POLITICAL PARTY**

In order to compare state by state regardless their population, an index was created consist in the cases per 100.000 inhabitants, according to the following calculation:

([Covid cases]/[total population])\*100.000

It could be seen that the republican counties have higher whiskers in the boxplot than the democratic ones, meaning that there are more counties with higher covid cases per 100.000 in the republican’s, which is also clear in the 3 waves. This could be related to the difference in mitigation efforts and vaccine uptakes. The graphic also shows that after march 2021 the behavior started to change showing more Covid cases for the democratic states.



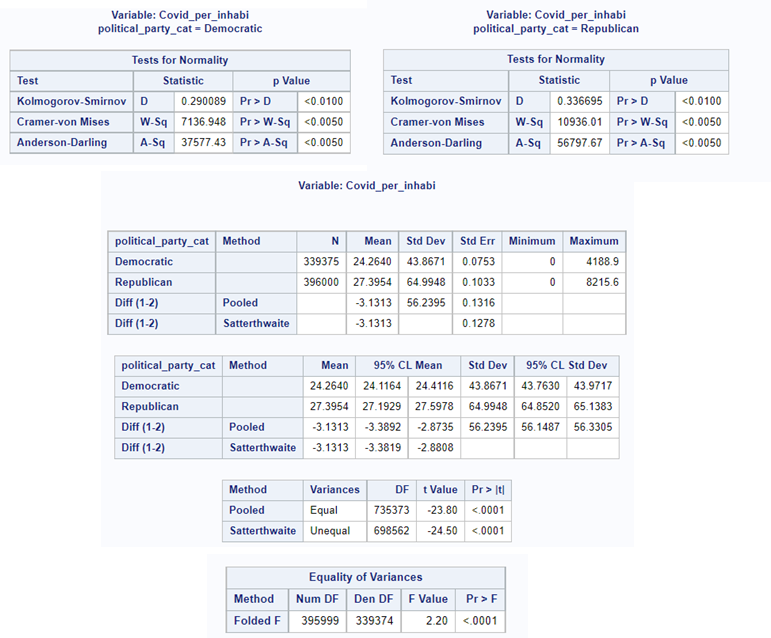


T-TEST

To validate the difference between the mean of the two political parties, a t-test were conducted using SAS, according with the following hypotheses:

**H0:** The mean of the cases per 100.000 inhabitants for both political parties do not differ.

**Ha:** There is a difference between the mean of the cases, according to the political party.



In this case, in the table Equality of Variances is possible to see that the P-value Pr > F is less than .05 meaning that the variables are unequal then we should read the Satterthwaite section of the results. Obtaining a T-Value is -24.50 and P-Value less than 0.0001 (less than the alpha value .05). In this case, the null hypothesis is rejected and the mean of the covid per 100.000 inhabitant for both political parties are significantly different.

We can also see that the standard deviation for the republican counties are higher (std dev = 64.99) than the democratic counties (std dev = 43.86).

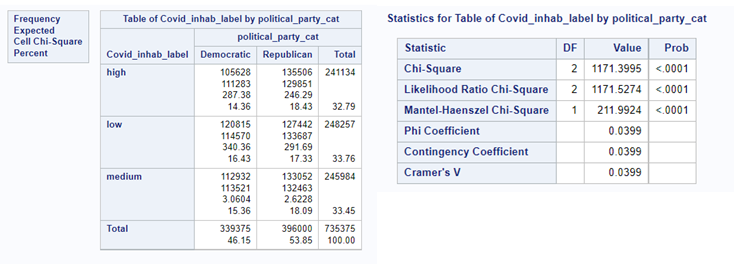
CHI-SQUARE

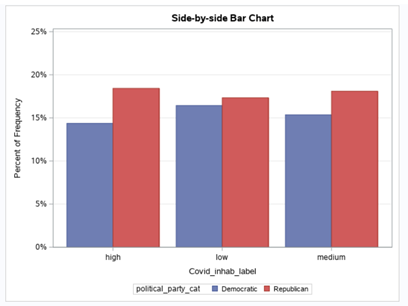
In order to understand if political party are related with Covid cases per 100.000 inhabitants, a chi-square was conducted using a binning technique on the Covid cases rating:

Graphical user interface, text, application

Description automatically generated

The result of the Chi-Square analysis are showed below:

****

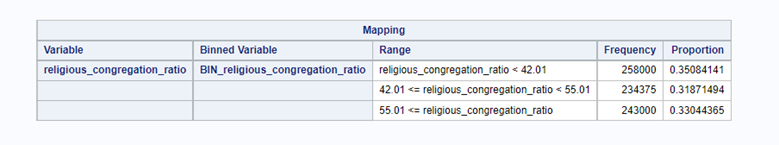
****

Since P-Value is less than .05 we can reject the null hypotheses and conclude that political party are related to Covid cases.

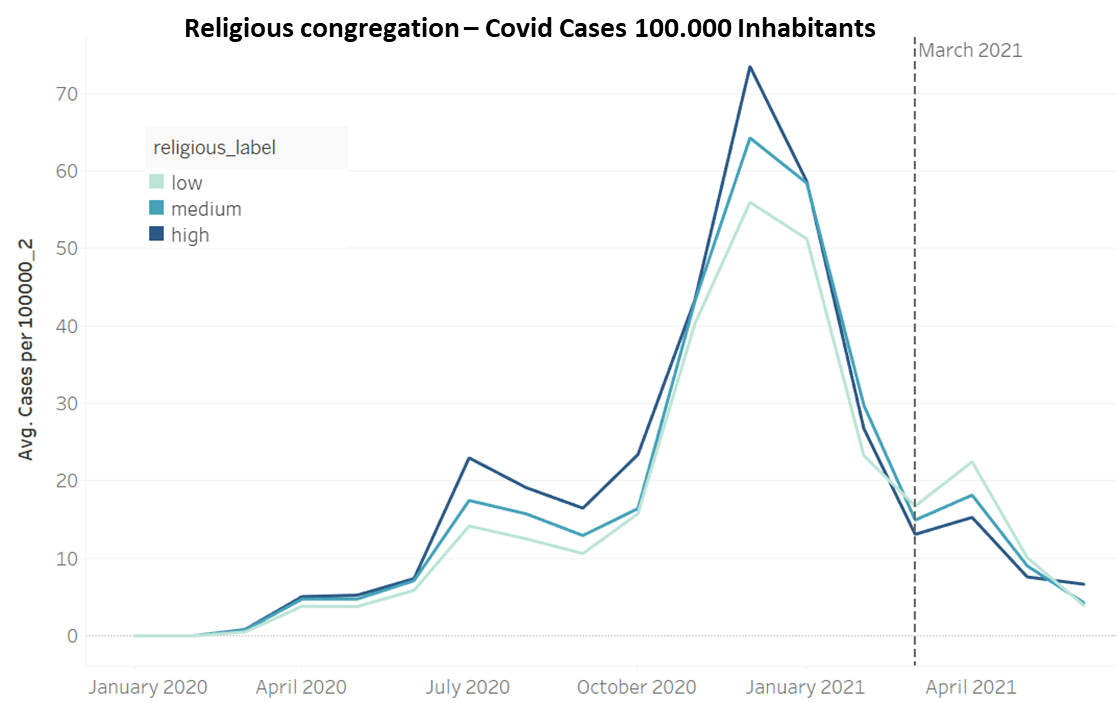
Checking the independence of both events with the following formula, being A equal to political party (Democratic) and B the level of Covid cases per inhabitants (High), It is possible to see that there is a small difference between P(A) and P(A/B) meaning that both events are not independent:

**3.2 RELIGIOUS CONGREGATION RATIO**

A binning technique with equal frequency was applied to the religious ratio according to the table below:

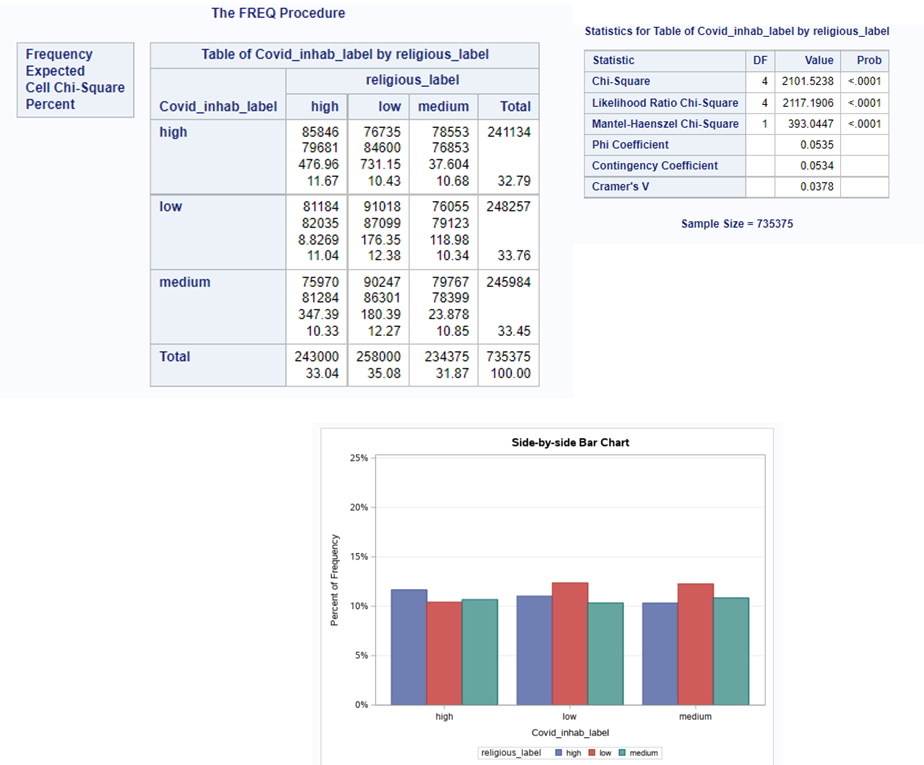


According to the graph below, until March 2021 it is possible to say that the less the religious congregation ratio, the less the covid cases per 100.000 inhabitants. This behavior could be explained for the religious gathering that are related with the spread of covid cases.



CHI-SQUARE:

The chi-square help us to reject the null hypotheses, showing that there is a correlation between religious congregation ratio and Covid cases per 100.000 inhabitants, with a p-value less than .05.



The boxplot also shows that the median is getting higher for Covid cases per 100.000 inhabitants:

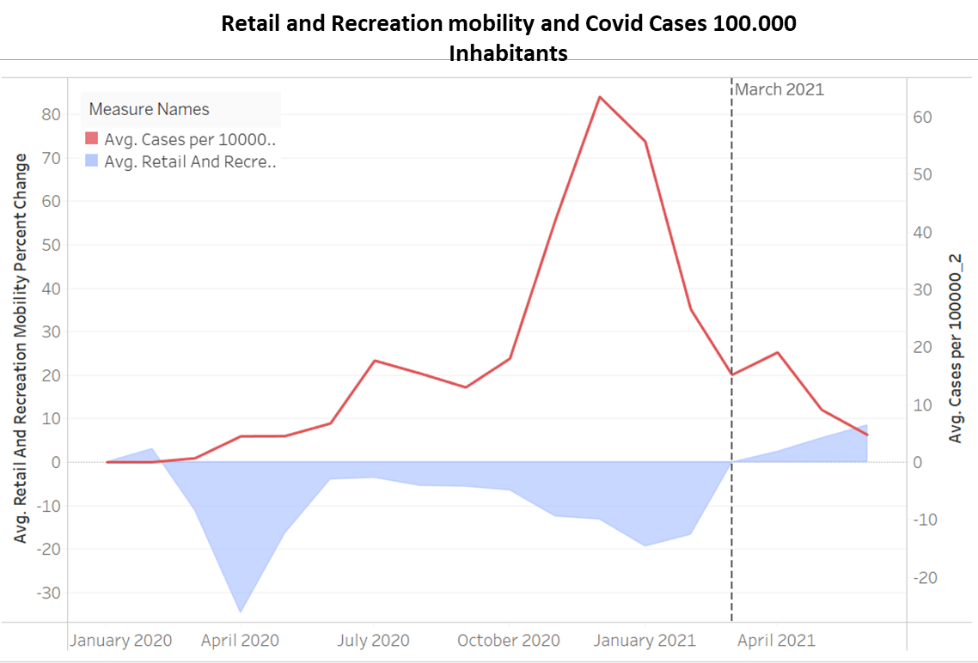
* Low religious ratio – Median 17.47
* Medium religious ratio – Median 20.35
* High religious ratio – Median 21.99

**Chart, box and whisker chart

Description automatically generated**

**3.3 RETAIL AND RECREATION MOBILITY**

The retail and recreation mobility feature is calculated with the change in mobility compare to the baseline. The baseline corresponds to the average mobility in the period of January 3 until February 6, 2020[[3]](#footnote-2). Apparently, retail and recreation mobility feature has a has a negative correlation with the Covid cases per 100.000 inhabitants rate. The change in the behavior with politics and religion matches with loosened of the social distance after March 2021.



PEARSON TEST

To validate the independence of these 2 variables a Pearson test was conducted with the following hyphoteses:

**H0:** There is not relationship between the retail and recreation mobility and the Covid cases per 100.000 inhabitants

Table

Description automatically generated

The test shows a negative and week relationship with a r-coefficient of -0.11 between both variables. Since P-Value is less than .05 the null hypotheses is rejected, meaning that there is a relation between retail and recreation mobility and Covid cases per 100.000 inhabitants.

**3.4 VIRUS PRESSURE**

Chart, line chart

Description automatically generated

For virus pressure of COVID-19 from the research of World Health Organization (WHO), it spreads between people in several ways:

* Current situation proves that virus spreads mainly close contact with each other e.g., at a conversational distance.
* The virus spread through small liquid particles from an infected people’s nose and mouth when they speak, cough, sneeze, or breath. Virus particles may also transmit through droplet transmission by the eyes, mouth, or nose.
* Also, the virus can spread via poor ventilation systems or people stay indoor for a long period of time because the aerosols remain suspended in the air.
* Finally, people may infect by touching the infectious surfaces or objects that have been contaminated by the virus when touching their eyes, nose, or mouth.[[4]](#footnote-3)

Moreover, in some counties with high levels of pollution measures more in danger of die from COVID-19 and where it needs to carry out the exercise of social distancing, states by Francesca Dominici, a professor of biostatistics at Harvard University’s T.H. Chan School of Public Health and co-author of the study.[[5]](#footnote-4)

We also do understand, the closer or crowd counties and population, the higher COVID cases increased. That is cause and effect in some sense.

**3.5 MEATPLANTS**

Chart, bar chart, histogram

Description automatically generated

Amongst to our data and information on hand, we find that there are a significant correlation and relationship between meat plants and COVID-19 cases by Chi-square. It did show the dynamic connection in between. In the meantime, we also found there are studies and research regarding the strong relationship between meat plants and COVID-19 cases.

Based on the statistic of North American Meat Institute in Fiscal Year 2018, there were 835 inspected meat plants federally.[[6]](#footnote-5)

From the public analysis report from Reuters in Jan 2022, it is nearly 90% of the meat plants declared COVID-19 confirmed cases in 2020 and early 2021.[[7]](#footnote-6) Data from the top U.S. meatpackers (based on the number of employees) Tyson Foods, JBS, Cargill, Smithfield Foods and National Beef made public in October indicated 59,000 cases of COVID-19 among their workers from March 2020 to February 2021. This figure were around 3 times higher than previous estimates.[[8]](#footnote-7)

According to the research from Proceedings of the National Academy of Sciences of the United States of America (PNAS), published on November 19, 2020. The result states a strong positive relationship between meat plant processing and local society of COVID-19 transmissions, which may accelerate the spread of virus into the surrounding population. PNAS estimates the total excess COVID-19 cases associated with meat plants approximately to be 236,000 to 310,000 (6 to 8% of all US cases) as of July 2021. The correlation is detected primarily among large meat processing facilities and large meatpacking plants. Furthermore, they further find evidence that many meat plants were shut down or temporary closure of high-risk meat plants to diminish the rates and to halt the spread of COVID-19 case growth. Whilst the smaller, decentralized meat plants do not arise to contribute to the transmission.[[9]](#footnote-8)

PNAS has provided more detailed on Meat plants and COVID-19 transmissions on county-level per thousand. Here is the result PNAS issued to the public.

Chart, line chart

Description automatically generated[[10]](#footnote-9)

On the other hand, the report by the federal Government Accountability Office (GAO) outbreak among meat plant workers were relatively higher compared with other manufacturing or packaging workers. In the initial stage of pandemic, employees work close together and not even report their illnesses because of fear of loss their job. Also, making social distancing in the manufacturing or packaging sectors is basically unfeasible to execute. This would lead more crowded working conditions and increase risk of COVID-19 transmission.

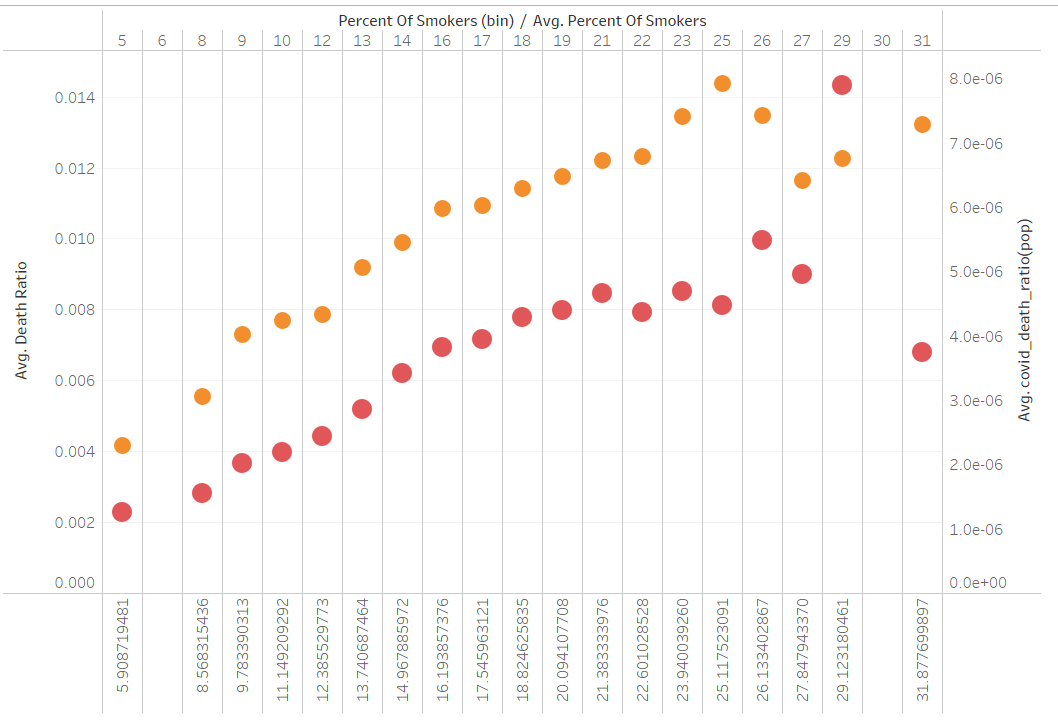
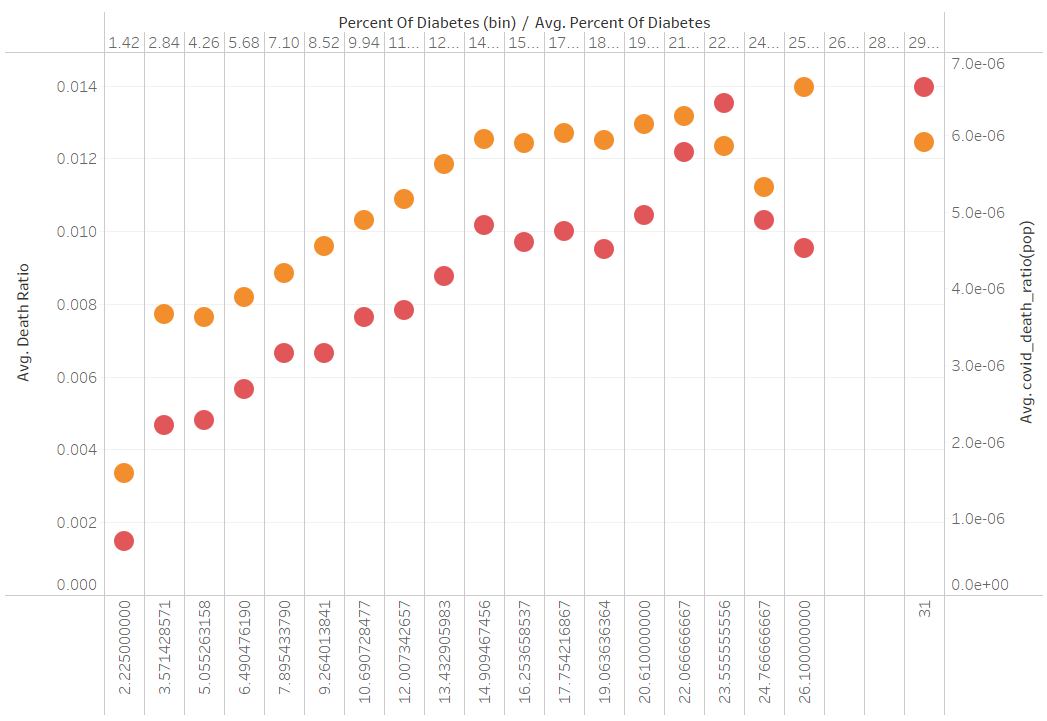
Diagram

Description automatically generated A picture containing person, people, scene, crowd

Description automatically generated**3.6**

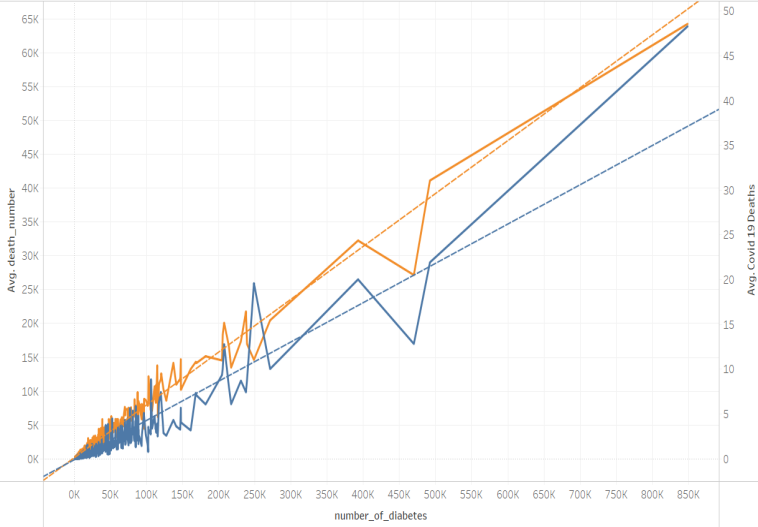
**3.6 PERCENT OF SMOKER AND DIABETES**

There are two important variable called percent of smoker and percent of diabetes. The parameter percent of smoker states the number of adult smoker in percentage per total population of each county. Similarly the percent of diabetes states the percent of adult diabetes per total population of each county. Upon analysis we found that the percent of smoker and diabetes have high correlation with the death ratio (parameter in our dataset which states total number of death per total population of the county).



Here to analysis properly , we have converted number of Covid-19 death per county to Covid-19 death ratio (apple to apple comparison) . The Covid-19 death ratio is calculated by : *(Total number of Covid-19 death / total population of county )\*100*.Then for comparing we have averaged the Covid-19 death ratio as well as death ratio. We have also binned the percent of smoker/ diabetes to its nearest percentage.

From the graph above it is evident that, more the percentage of smoker/diabetes , more is the death ratio as well as Covid-19 death ratio. This is because , the smoking habit and diabetes are the key factors in Covid-19 severity and death as well as death in general. Hence they are positively correlated. The following graph illustrate the same about their correlation.



R2-Score of Avg.death\_number (Percent of Smoker) : 0.96 (p-value <0.001)

R2-Score of Avg.death\_number (Percent of Diabetes) : 0.96 (p-value <0.001)

R2-Score of Avg.Covid-19\_death (Percent of Smoker) : 0.85 (p-value <0.001)

R2-Score of Avg.Covid-19\_death (Percent of Diabetes) : 0.88 (p-value <0.001)

# **PREDICTIVE MODEL**

As we were already aware of the problem being time series problem, we decided to design a prediction model based on SARIMAX model (Seasonal Auto Regression Integrated Moving Average model With Exogenous Variable) .

To start designing , we have to sample our dataset for the model to learn and test. For our sample, we filtered the top 10 populated counties. Because of the wide variety of demographic and political difference among different counties , we wanted to focus on the model which will help us with the prediction of a particular type of demographic (population heavy counties). Moreover, as population is correlated with number of cases, we assumed top populated counties will give us more covid cases observation to train.

The Top 10 counties were :

* 'Los Angeles County',
* 'Orange County',
* 'Cook County',
* 'Harris County',
* 'Maricopa County',
* 'Montgomery County',
* 'San Diego County',
* 'Jefferson County',
* 'Clark County',
* 'Dallas County'

The feature engineering performed were as follows :

* no\_of\_death = death\_ratio/total\_population (per counties)
* no\_of\_immigrant = immigrant\_student\_ratio/total\_population (per counties)

Our final filtered data-frame contain the following columns :

* 'total\_population'
* 'number\_of\_meat\_plants'

Independent Variable

* 'no\_of\_death'
* 'daily\_state\_test','virus\_pressure'
* 'no\_of\_student\_immigrants'
* 'airport\_distance'

Response Variable

* 'covid\_19\_confirmed\_cases'

We also notice that the dates were not continuous i.e not all the date observations were present within the data-frame time period. Hence we had to resample the date columns into **daily** frequency with **mean** aggregation function for rest of the independent variables.

Then we used seasonal decompose library from time series statsmodel . (from statsmodels.tsa.seasonal import seasonal\_decompose). We noticed there is a significant cyclic seasonality factor in weekly pattern.

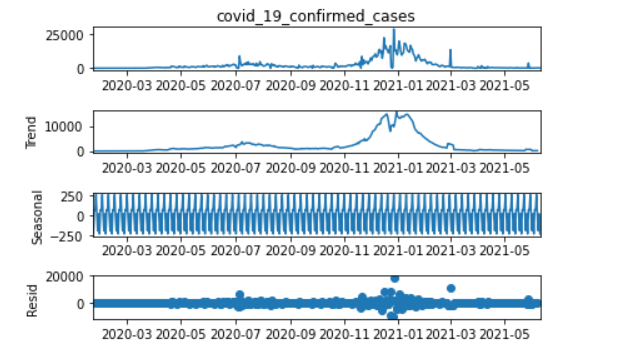
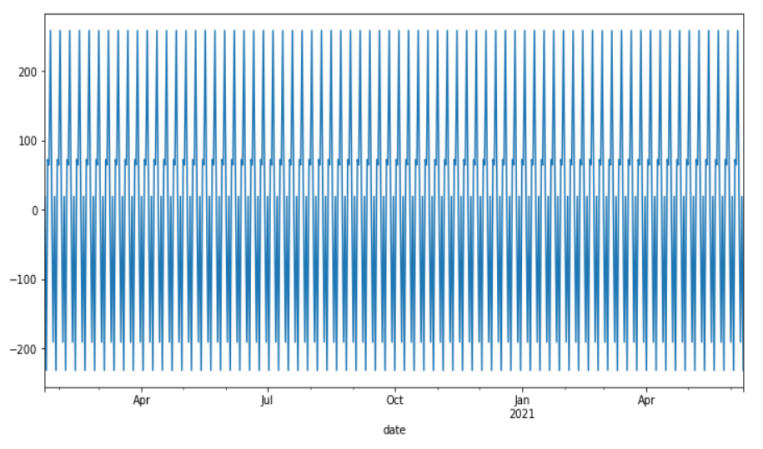
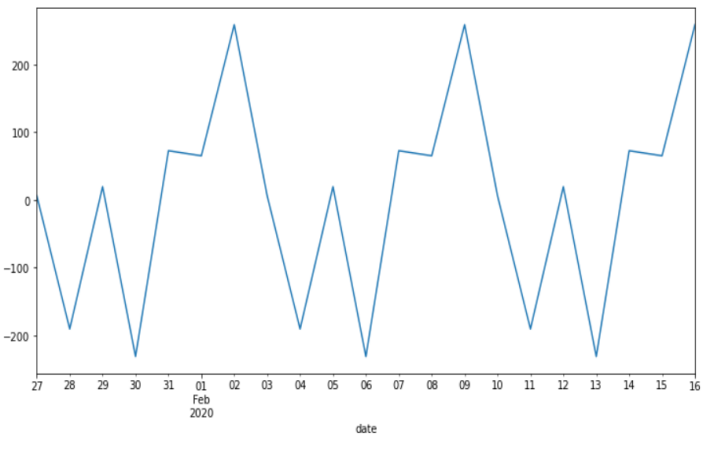


Figure : The whole seasonal decompose image with trend , seasonal and residual plot



Left fig : seasonal pattern , right fig : zoomed seasonal pattern

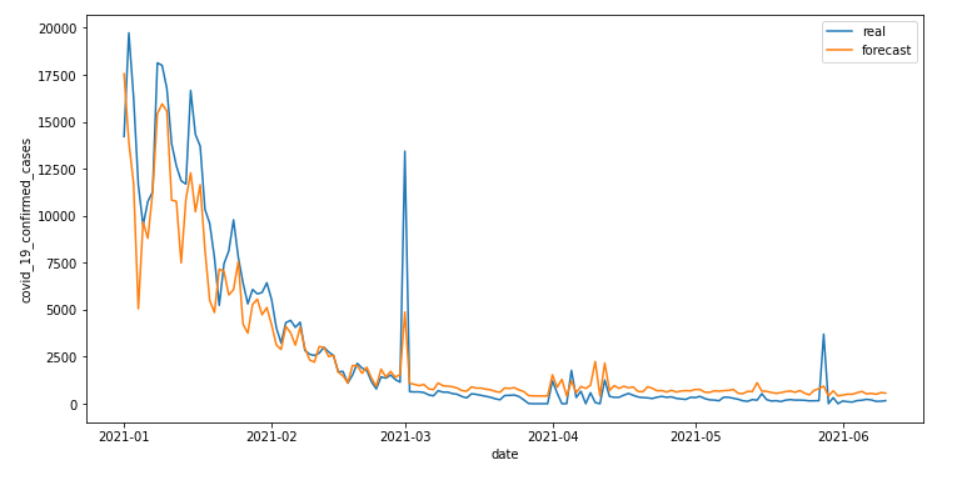
To train our model , we have split our data set into train and test set. We have selected the 1st wave, 2nd wave and rise of 3rd wave to learn and train, and used the rest of 3rd to evaluate our model. (date range :22-01-2020 to 31-12-2020 for train, 01-01-2021 to rest for test).

For our model configuration, we used a powerful parameter hyper-tuning model called autoarima, which gave us the combination of p,d,q values (p:Auto regression element, d:Differentiating, q: Moving Average element). The same also provided us with seasonal P,D,Q values too. In our case, it was all 0 with m = 7 (seasonal pattern). It was because Virus pressure was the only variable which was partial to current covid 19 cases. It was the only significant predicting factor for our time series model. Hence the other elements were negligible.

We finally used SARIMAX model with auto arima configured p,d,q values. For our exogenous variable we used the following variables:

* 'total\_population',
* 'number\_of\_meat\_plants'
* 'no\_of\_death'
* 'virus\_pressure'
* 'no\_of\_student\_immigrants'
* 'airport\_distance'

To evaluate , we used the most populated county ‘Los Angeles County’. We tried to predict for the entire test period.



Here we see, the forecast is very similar to real value , but a slightly before time . This is because virus pressure is the covid 19 cases itself but noted on previous time period which spread and impacted the neighbouring counties. We also recommend to predict on 7 days advance, as the model might not be accurate after 7 days.We tried to evaluate our model for 7 days forecast using root mean squared error metrics.

RMSE : 4017.90

Dependent Mean : 13346.42

Hence Los Angeles County prediction has only 30% prediction error. Similarly Orange County has 20% and Cook County has 18% prediction error.

# **SUMMARY**

* Until March 2021, the behavior of Covid cases per 100.000 inhabitants seems to reverse its trend related with some features like political party and religious congregation. This situation matched with the loosened of the social distancing in that month, showing a increase in mobility in retail and recreation locations.
* Most Pearson and Chi-Square test showed weak relationship between the different features with the Covid Cases, for this reason was valuable to understand the variables through other techniques and graphs.
* Percent of Smoker/ Diabetes have linear correlation with number of covid deaths and death in general.
* Virus pressure has the highest Pearson correlation with covid cases and the forecasting factor for time series model.
* We have a significant cyclic seasonal pattern in our data . Our covid cases have a strong weekly pattern which was derived from statsmodel’s seasonal decompose method.
* After parameter hyper-tuning, we get the configuration of p,d,q and P,D,Q as 0,0,0 for our SARIMAX model with m = 7.
* After using SARIMAX model with exogenous variable , we get a forecast with prediction error of 18-30 % for top three counties over the period of 7 days.

# **REFERENCES**

Haratian, Arezoo, Hadi Fazelinia, Zeinab Maleki, Pouria Ramazi, Hao Wang, Mark A. Lewis, Russell Greiner, and David Wishart. 2021. “Dataset of COVID-19 Outbreak and Potential Predictive Features in the USA.” Data in Brief 38 (October): 107360. https://doi.org/10.1016/j.dib.2021.107360.

Kravitz, Dillon Bergin, Betsy Ladyzhets, Mohar Chatterjee and Derek. n.d. “The US Is Undercounting COVID Deaths, Researchers Say. Now They Have a Tool to Figure out Why.” USA TODAY. <https://www.usatoday.com/story/news/nation/2021/12/09/covid-death-count-america-cdc/6435124001/>.

BBC News. 2020. “Covid-19 in the US: Is This Coronavirus Wave the Worst Yet?,” November 18, 2020, sec. US & Canada. <https://www.bbc.com/news/world-us-canada-54966531>.

# **APPENDIX 1 – TABLEAU LINK**

The following link contains the presentation related to our analysis in Tableau Public:

<https://public.tableau.com/app/profile/valeria.velez2119/viz/Tableau_Covid_Presentation/Covid19_USA?publish=yes>

# **APPENDIX 2 – SAS CODE**

LIBNAME mylib "/home/u60671031/Langara\_vvv/Covid";

/\* Creating global statements \*/

%LET file\_1 = '/home/u60671031/Langara\_vvv/Inputs/Covid\_clean.csv';

/\* Importing files \*/

proc import datafile = &file\_1 dbms = csv out = mylib.covid19;

getnames = yes;

option validvarname=V7;

guessingrows = max;

run;

proc contents data=mylib.covid19;

run;

/\* Creating a new column about the cases per 100.000 inhabitants \*/

proc sql;

create table mylib.covid3 as

Select \*, (Covid\_19\_Confirmed\_Cases/Total\_Population)\*100000 as Covid\_per\_inhabi

from mylib.covid19

where date >= "01Jun2020"d;

quit;

/\* Creating a new column about the cases per 100.000 inhabitants \*/

proc sql;

create table mylib.covid3 as

Select \*, (Covid\_19\_Deaths/Total\_Population)\*100000 as Covid\_deaths\_inhabi

from mylib.covid19;

quit;

/\* POLITICS \*/

data mylib.covid3;

set mylib.covid3;

format political\_party\_cat $char19.;

if political\_party = 0 then political\_party\_cat = 'Republican';

if political\_party = 1 then political\_party\_cat = 'Democratic';

run;

/\* T-Test political parties \*/

ods noproctitle;

ods graphics / imagemap=on;

/\* Test for normality \*/

proc univariate data=MYLIB.COVID3 normal mu0=0;

ods select TestsForNormality;

class political\_party\_cat;

var Covid\_per\_inhabi;

run;

/\* t test \*/

proc ttest data=MYLIB.COVID3 sides=2 h0=0 plots(showh0);

class political\_party\_cat;

var Covid\_per\_inhabi;

run;

/\* Obtaining categorical dataset\*/

proc hpbin data=mylib.covid3 output=delete numbin=3 pseudo\_quantile;

input Covid\_per\_inhabi;

run;

data mylib.covid3;

set mylib.covid3;

format Covid\_inhab\_label $char12.;

if Covid\_per\_inhabi < 4.9293649705 then Covid\_inhab\_label = "low";

if Covid\_per\_inhabi >= 4.9293649705 and Covid\_per\_inhabi < 23.003703196 then Covid\_inhab\_label = "medium";

if Covid\_per\_inhabi >= 23.003703196 then Covid\_inhab\_label = "high";

drop Watch\_time\_h;

run;

/\* Chi-Square analysis \*/

proc freq data = mylib.covid3;

tables Covid\_inhab\_label\*political\_party\_cat/chisq expected NOCOL NOROW cellchi2;

run;

/\* SIDE-BY-SIDE Bar Charts \*/

ods graphics / reset width=6.4in height=4.8in imagemap;

proc glm data = mylib.covid3 PLOTS(MAXPOINTS=NONE);

proc sgplot data=mylib.covid3;

title 'Side-by-side Bar Chart';

vbar Covid\_inhab\_label / group=political\_party\_cat groupdisplay=cluster stat=percent;

xaxis;

yaxis max=0.25 grid;

run;

ods graphics / reset;

/\* RELIGIOUS CONGREGATION \*/

/\* Correlation Numerical variables \*/

PROC CORR DATA=mylib.covid3 plots=matrix(histogram) PLOTS(MAXPOINTS=NONE);

VAR Covid\_per\_inhabi religious\_congregation\_ratio;

RUN;

ods graphics on;

proc hpbin data=mylib.covid3 output=delete numbin=3 pseudo\_quantile;

input religious\_congregation\_ratio;

run;

data mylib.covid3;

set mylib.covid3;

format religious\_label $char12.;

if religious\_congregation\_ratio < 42.01 then religious\_label = "low";

if religious\_congregation\_ratio >= 42.01 and religious\_congregation\_ratio < 55.01 then religious\_label = "medium";

if religious\_congregation\_ratio >= 55.01 then religious\_label = "high";

run;

/\* Chi-Square analysis \*/

proc freq data = mylib.covid3;

tables Covid\_inhab\_label\*religious\_label/chisq expected NOCOL NOROW cellchi2;

run;

/\* SIDE-BY-SIDE Bar Charts \*/

ods graphics / reset width=6.4in height=4.8in imagemap;

proc glm data = mylib.covid3 PLOTS(MAXPOINTS=NONE);

proc sgplot data=mylib.covid3;

title 'Side-by-side Bar Chart';

vbar Covid\_inhab\_label / group=religious\_label groupdisplay=cluster stat=percent;

xaxis;

yaxis max=0.25 grid;

run;

ods graphics / reset;

/\* RETAIL AND RECREATION MOBILITY \*/

/\* Correlation Numerical variables \*/

PROC CORR DATA=mylib.covid3 plots=matrix(histogram) PLOTS(MAXPOINTS=NONE);

VAR Covid\_per\_inhabi retail\_and\_recreation\_mobility\_p;

RUN;

ods graphics on;

/\* VACCINATION \*/

Data mylib.covid3\_2021;

set mylib.covid3;

where date >= "01Feb2021"d;

run;

/\* Correlation Numerical variables \*/

PROC CORR DATA=mylib.covid3\_2021 plots=matrix(histogram) PLOTS(MAXPOINTS=NONE);

VAR covid\_19\_deaths percent\_of\_vaccinated\_residents;

RUN;

ods graphics on;

/\* Correlation Numerical variables \*/

PROC CORR DATA=mylib.covid3\_2021 plots=matrix(histogram) PLOTS(MAXPOINTS=NONE);

VAR covid\_19\_confirmed\_cases percent\_of\_vaccinated\_residents;

RUN;

ods graphics on;

/\* HOUSING DENSITY \*/

/\* Correlation Numerical variables \*/

PROC CORR DATA=mylib.covid3 plots=matrix(histogram) PLOTS(MAXPOINTS=NONE);

VAR covid\_19\_deaths housing\_density;

RUN;

ods graphics on;

/\* Correlation Numerical variables \*/

PROC CORR DATA=mylib.covid3 plots=matrix(histogram) PLOTS(MAXPOINTS=NONE);

VAR covid\_19\_confirmed\_cases housing\_density;

RUN;

ods graphics on;

/\* percent\_of\_insured\_residents \*/

/\* Correlation Numerical variables \*/

PROC CORR DATA=mylib.covid3 plots=matrix(histogram) PLOTS(MAXPOINTS=NONE);

VAR covid\_19\_deaths percent\_of\_insured\_residents;

RUN;

ods graphics on;

/\* Correlation Numerical variables \*/

PROC CORR DATA=mylib.covid3 plots=matrix(histogram) PLOTS(MAXPOINTS=NONE);

VAR covid\_19\_confirmed\_cases percent\_of\_insured\_residents;

RUN;

ods graphics on;

/\* retail\_and\_recreation\_mobility\_percent\_change \*/

/\* Correlation Numerical variables \*/

PROC CORR DATA=mylib.covid3 plots=matrix(histogram) PLOTS(MAXPOINTS=NONE);

VAR covid\_19\_deaths retail\_and\_recreation\_mobility\_p;

RUN;

ods graphics on;

/\* Correlation Numerical variables \*/

PROC CORR DATA=mylib.covid3 plots=matrix(histogram) PLOTS(MAXPOINTS=NONE);

VAR covid\_19\_confirmed\_cases retail\_and\_recreation\_mobility\_p;

RUN;

ods graphics on;

/\* GDP VARIABLE \*/

/\* Correlation Numerical variables \*/

PROC CORR DATA=mylib.covid3 plots=matrix(histogram) PLOTS(MAXPOINTS=NONE);

VAR Covid\_per\_inhabi gdp\_per\_capita;

RUN;

ods graphics on;

ods graphics / reset width=5.8in height=4in imagemap;

proc sgplot data=mylib.covid3;

reg x=Covid\_per\_inhabi y=gdp\_per\_capita / nomarkers;

scatter x=Covid\_per\_inhabi y=gdp\_per\_capita;;

xaxis integer;

yaxis grid;

run;

/\* MEATPLANTS \*/

proc hpbin data=mylib.covid3 output=delete numbin=3 pseudo\_quantile;

input number\_of\_meat\_plants;

run;

PROC CORR DATA=mylib.covid3 plots=matrix(histogram) PLOTS(MAXPOINTS=NONE);

VAR Covid\_19\_confirmed\_cases number\_of\_meat\_plants;

RUN;

ods graphics on;

# **APPENDIX 2 – PYTHON CODE**

#!/usr/bin/env python

# coding: utf-8

# In[47]:

from pyforest import \*

import warnings

from statsmodels.tsa.statespace.sarimax import SARIMAX

from statsmodels.tsa.seasonal import seasonal\_decompose

warnings.filterwarnings('ignore')

lazy\_imports()

# In[48]:

# read dataset

df\_covid = pd.read\_csv('dataset/imputed-data.csv',header = 0)

df\_covid.head()

# In[49]:

# df\_informations

df\_covid.info()

# In[50]:

# date as date format

df\_covid['date'] = pd.to\_datetime(df\_covid['date'])

df\_covid['date'].head()

# In[51]:

# plt.figure(figsize = (12,6))

# sns.lineplot(x = 'date', y = 'covid\_19\_confirmed\_cases', data = df\_covid)

# sns.lineplot(x = 'date', y = 'temperature', data = df\_covid)

# In[52]:

# top high POPULATED counties

high\_case = df\_covid[df\_covid['date'] == '2020-02-01'].groupby(['county\_name'])[

'total\_population'].sum().sort\_values(ascending = False)

high\_case\_10 = list(high\_case.index[:10])

high\_case\_10

# In[53]:

# top low POPULATED counties

low\_case = df\_covid[df\_covid['date'] == '2020-02-01'].groupby(['county\_name'])[

'total\_population'].sum().sort\_values(ascending = True)

low\_case\_10 = list(low\_case.index[:10])

low\_case\_10

# In[54]:

# feature engineering

df\_covid['median\_household\_income\_bins'] = pd.qcut(df\_covid['median\_household\_income'],q = 5).cat.codes

df\_covid['smoker\_bins'] = pd.qcut(df\_covid['percent\_of\_smokers'],q = 13).cat.codes

df\_covid['diabetes\_bins'] = pd.qcut(df\_covid['percent\_of\_diabetes'],q = 7).cat.codes

df\_covid['no\_of\_death'] = df\_covid['death\_ratio']\*df\_covid['total\_population']

df\_covid['no\_of\_student\_immigrants'] = df\_covid['immigrant\_student\_ratio']\*df\_covid['total\_population']

df\_covid['social\_distancing\_total\_grade'] = LabelEncoder().fit\_transform(df\_covid['social\_distancing\_total\_grade'])

# In[55]:

# top states filter

top\_counties = df\_covid[df\_covid['county\_name'].isin(high\_case\_10)]

top\_counties = top\_states.groupby(['county\_name','date'])[['total\_population','number\_of\_meat\_plants',

'no\_of\_death','covid\_19\_confirmed\_cases', 'daily\_state\_test','virus\_pressure',

'no\_of\_student\_immigrants','airport\_distance']].mean().reset\_index()

top\_counties.head()

# In[56]:

# create individual df for counties

county\_df = {}

county\_df = {}

for county in high\_case\_10:

county\_df[county] = pd.DataFrame(data = top\_counties[top\_counties['county\_name'] == county])

county\_df[county] = county\_df[county][['date','total\_population','number\_of\_meat\_plants',

'no\_of\_death','covid\_19\_confirmed\_cases','virus\_pressure',

'no\_of\_student\_immigrants','airport\_distance']].resample('D', on='date').mean()

# In[57]:

# testing dfs

county\_df['Los Angeles County'].head()

# In[58]:

county\_df['Los Angeles County']['covid\_19\_confirmed\_cases'].plot()

# In[59]:

# check seasonal decompose

result = seasonal\_decompose(county\_df['Los Angeles County']['covid\_19\_confirmed\_cases'],model = 'add')

result.plot();

# In[60]:

plt.figure(figsize = (12,6))

result.seasonal.plot()

# In[61]:

plt.figure(figsize = (12,6))

result.seasonal[5:26].plot()

# In[62]:

# train test split

from sklearn.model\_selection import train\_test\_split

train\_df = {}

test\_df = {}

future\_df = {}

for county in high\_case\_10:

train = county\_df[county][county\_df[county].index <= '2020-12-31']

test = county\_df[county][(county\_df[county].index > '2020-12-31')]

train\_df[county] = train

test\_df[county] = test

# In[63]:

test\_df['Los Angeles County']['covid\_19\_confirmed\_cases'].plot()

# In[64]:

# checking for each states forecast range

for county in high\_case\_10:

print(county , " : whole : ", len(county\_df[county].index))

print(county , " : train : ", len(train\_df[county].index))

print(county , " : test : ", len(test\_df[county].index))

# In[65]:

train\_df['Los Angeles County'].columns

# #### what is Auto Regression ?

# Autoregression is a time series model that uses observations from previous time steps as input to a regression equation to predict the value at the next time step. It is a very simple idea that can result in accurate forecasts on a range of time series problems

# ![image.png](attachment:image.png)

# #### what is Differentiation ?

# Differencing is a method of transforming a time series dataset. It can be used to remove the series dependence on time, so-called temporal dependence. This includes structures like trends and seasonality.

# ![image.png](attachment:image.png)

# #### what is Moving Average ?

# A moving average is defined as an average of fixed number of items in the time series which move through the series by dropping the top items of the previous averaged group and adding the next in each successive average.

# ![image.png](attachment:image.png)

# In[66]:

# auto\_arima fit #c

from pmdarima import auto\_arima

grid\_model = {}

for county in high\_case\_10:

grid\_model[county] = auto\_arima(county\_df[county]['covid\_19\_confirmed\_cases'],

exogenous = county\_df[county][['total\_population', 'number\_of\_meat\_plants', 'no\_of\_death',

'virus\_pressure', 'no\_of\_student\_immigrants',

'airport\_distance']],seasonal = True,m = 7,trace = True)

# In[67]:

for county in high\_case\_10:

print(grid\_model[county].summary())

# In[68]:

# sarimax model

ts\_model = {}

for county in high\_case\_10:

ts\_model[county] = SARIMAX(train\_df[county]['covid\_19\_confirmed\_cases'],

exog = train\_df[county].loc[:,

~train\_df[county].columns.isin(['covid\_19\_confirmed\_cases'])],

order = (0,0,0), seasonal\_order = (0,0,0,7)).fit()

print(ts\_model[county].summary())

# In[69]:

# forecast first

forecast = {}

for county in high\_case\_10:

forecast\_temp = ts\_model[county].forecast(steps = 161,exog = test\_df[county].iloc[:161,

~test\_df[county].columns.isin(['covid\_19\_confirmed\_cases'])],

typ = 'levels').rename('forecast\_sarimax')

forecast[county] = pd.DataFrame(forecast\_temp)

# In[70]:

forecast['Los Angeles County'][:7]

# In[71]:

test\_df['Los Angeles County']['covid\_19\_confirmed\_cases'][:7]

# In[72]:

#forecast plot

plt.figure(figsize = (12,6))

sns.lineplot(x = test\_df['Los Angeles County'].iloc[:161].index,

y = test\_df['Los Angeles County'].iloc[:161]['covid\_19\_confirmed\_cases'],label = 'real'

,legend = True)

sns.lineplot(x = forecast['Los Angeles County'].index, y = forecast['Los Angeles County']['forecast\_sarimax'],label = 'forecast')

# In[73]:

#forecast plot

plt.figure(figsize = (12,6))

sns.lineplot(x = test\_df['Orange County'].iloc[:161].index,

y = test\_df['Orange County'].iloc[:161]['covid\_19\_confirmed\_cases'],label = 'real'

,legend = True)

sns.lineplot(x = forecast['Orange County'].index, y = forecast['Orange County']['forecast\_sarimax'],label = 'forecast')

# In[74]:

#forecast plot

plt.figure(figsize = (12,6))

sns.lineplot(x = test\_df['Cook County'].iloc[:161].index,

y = test\_df['Cook County'].iloc[:161]['covid\_19\_confirmed\_cases'],label = 'real'

,legend = True)

sns.lineplot(x = forecast['Cook County'].index, y = forecast['Cook County']['forecast\_sarimax'],label = 'forecast')

# In[82]:

# evaluating

np.sqrt(metrics.mean\_squared\_error(test\_df['Orange County'].iloc[:7]['covid\_19\_confirmed\_cases']

,forecast['Orange County'].iloc[:7]['forecast\_sarimax']))

# In[83]:

test\_df['Orange County'].iloc[:7]['covid\_19\_confirmed\_cases'].mean() # 20 % error

# In[80]:

np.sqrt(metrics.mean\_squared\_error(test\_df['Los Angeles County'].iloc[:7]['covid\_19\_confirmed\_cases']

,forecast['Los Angeles County'].iloc[:7]['forecast\_sarimax']))

# In[81]:

test\_df['Los Angeles County'].iloc[:7]['covid\_19\_confirmed\_cases'].mean() # 30% error

# In[84]:

np.sqrt(metrics.mean\_squared\_error(test\_df['Cook County'].iloc[:7]['covid\_19\_confirmed\_cases']

,forecast['Cook County'].iloc[:7]['forecast\_sarimax']))

# In[85]:

test\_df['Cook County'].iloc[:7]['covid\_19\_confirmed\_cases'].mean() # 18% error

1. (BBC News 2020) [↑](#footnote-ref-0)
2. (Kravitz, n.d.) [↑](#footnote-ref-1)
3. (Haratian et al., 2021) [↑](#footnote-ref-2)
4. (World Health Organization, 2020) [↑](#footnote-ref-3)
5. (says, 2020) [↑](#footnote-ref-4)
6. (North American Meat Institute, 2017) [↑](#footnote-ref-5)
7. (Douglas, 2022) [↑](#footnote-ref-6)
8. (www.industryselect.com, n.d.) [↑](#footnote-ref-7)
9. (Taylor, Boulos and Almond, 2020) [↑](#footnote-ref-8)
10. (Wikipedia, 2020) [↑](#footnote-ref-9)