Team 7: Jagriti, Mehran, Yin Htoo Htoo Naing

Student ID: 100378830, 100364596, 100366895

DANA-4800-001

Instructor: Monica

10 Apr 2022

**Index**

1. Analysis 1- Health factors....................................................... 2
2. Introduction............................................................. 2
3. List of Variable........................................................ 2
4. Variables Chosen for Final Analysis....................... 2
5. Percent of Insured Residents.................................... 3
6. T-test for Insured Residents.................................... 5
7. Ventilators Capacity Ratio....................................... 7
8. T-test for Ventilators Capacity Ratio...................... 8
9. Conclusion................................................................ 9
10. Summary…………………………………………... 9
11. Analysis 2 – Climatic Factors……………………………....... 10
12. Report……………………………………………… 10
13. T-test………………………………………………. 10
14. Discussion and Conclusion………………………… 14
15. Prediction…………………………………………… 14
16. Analysis 3 – Sociological Features……………………………. 16
17. T-test………………………………………………… 16
18. Z-test………………………………………………… 21
19. Recommendations and Conclusion…………………. 23
20. Bibliography…………………………………………………… 24
21. Appendix………………………………………………………. 25

**Analysis 1**

**Health Factors**

**Introduction**

The data set provided to us is a non-uniform dataset in terms of numerous factors from state to state or state to county or county to county. To make a conclusive result, I tried to take the top five states having maximum number of Covid-19 confirmed cases as well as the states having least number of Covid-19 confirmed cases. But that combined analysis breaks the pattern of uniformity for highest confirmed cases states. Thus, that idea was dropped, and this report now conducts analysis over the top five states (California, Florida, New York, Texas and Illinois) having Covid-19 cases.

While working on this analysis, it was initially expected that health factors should play a significant role in controlling Covid-19 deaths, but the results of analysis showed a different picture. As Covid-19 outbreak at accelerating pace. Thus, I was hard for the government to provide eminent hospital services to all the patients. This could be the reason these services have ridiculously small significance on Covid-19 deaths. Even the result for patients with any other health ailment has not shown much correlation with target variables. Therefore, this analysis will show that there are some insignificant variables in Covid-19 USA dataset.

**List of Variable**

There are six main variables that are considered in this group, and these are given below:

1. Percent of Smokers
2. Percent of Diabetes
3. Percent of Insured People
4. Hospital Bed Ratio
5. Ventilators Capacity Ratio
6. ICU Bed ratio

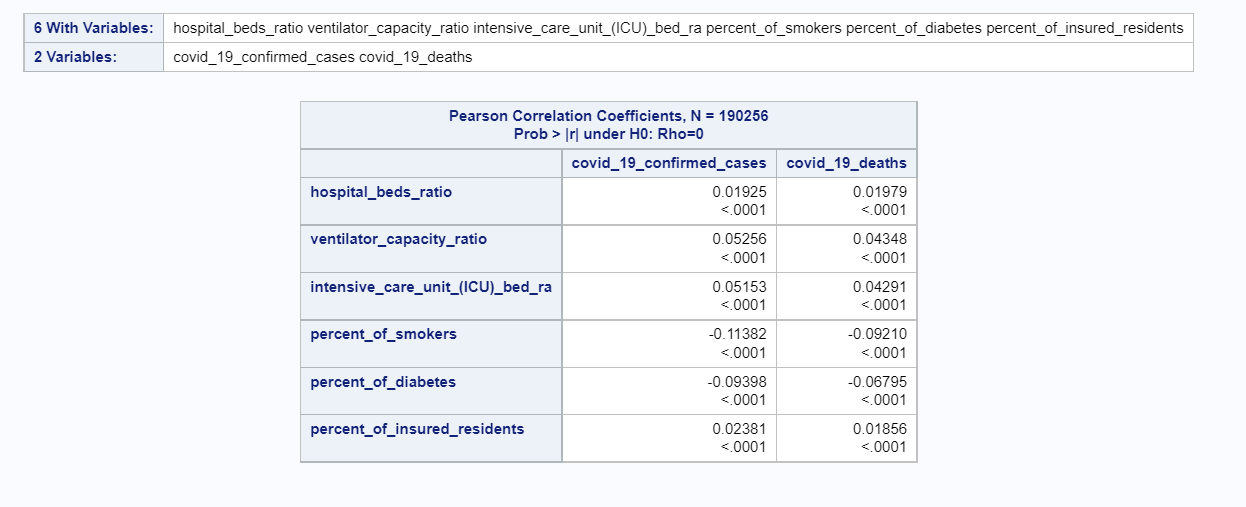
**Variable Chosen for final analysis**

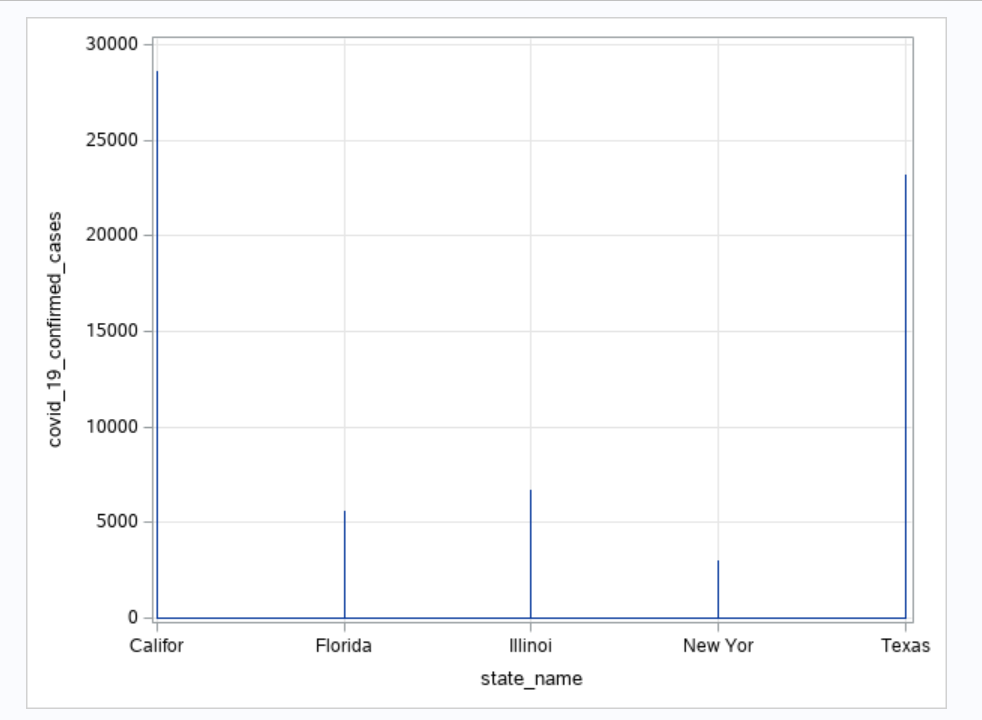
For the final analysis, this report will detail only two variables. These variables are listed below:

1. Percent of Insured Residents
2. Ventilators Capacity Ratio

The reason behind choosing these two variables can be verified by following table of correlation among target variables and analysis variables. Correlation coefficients have exceptionally low values and for percent of smokers and percent of diabetes, it is negative. These were hard to justify for negative coefficients. Hence, the percent of insured residents is taken into consideration.

Among hospital bed ratio, ICU beds ration and ventilators capacity ratio, ventilators were the most prominent requirement to treat critical patients and these are thus chosen for further analysis.

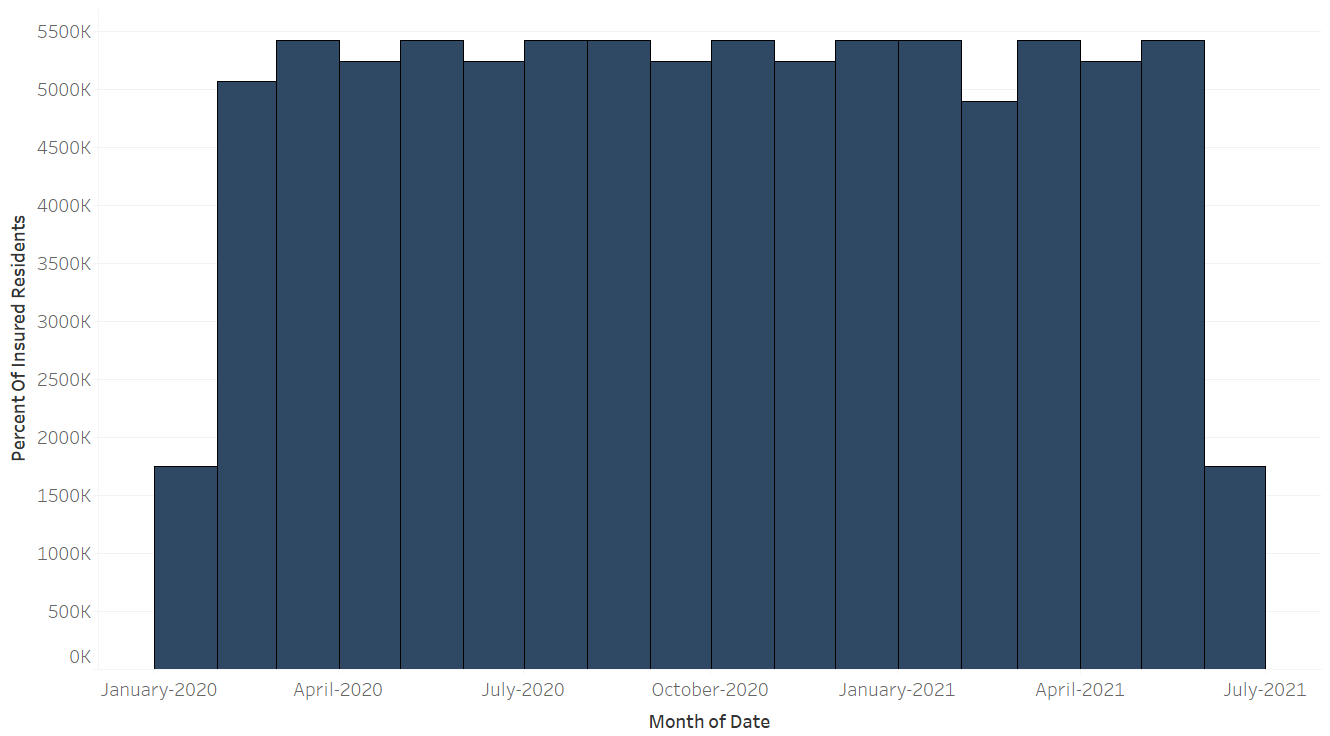
The chosen states according to Covid-19 confirmed cases are shown in the following graph. The graph also shows relative variations among the states.



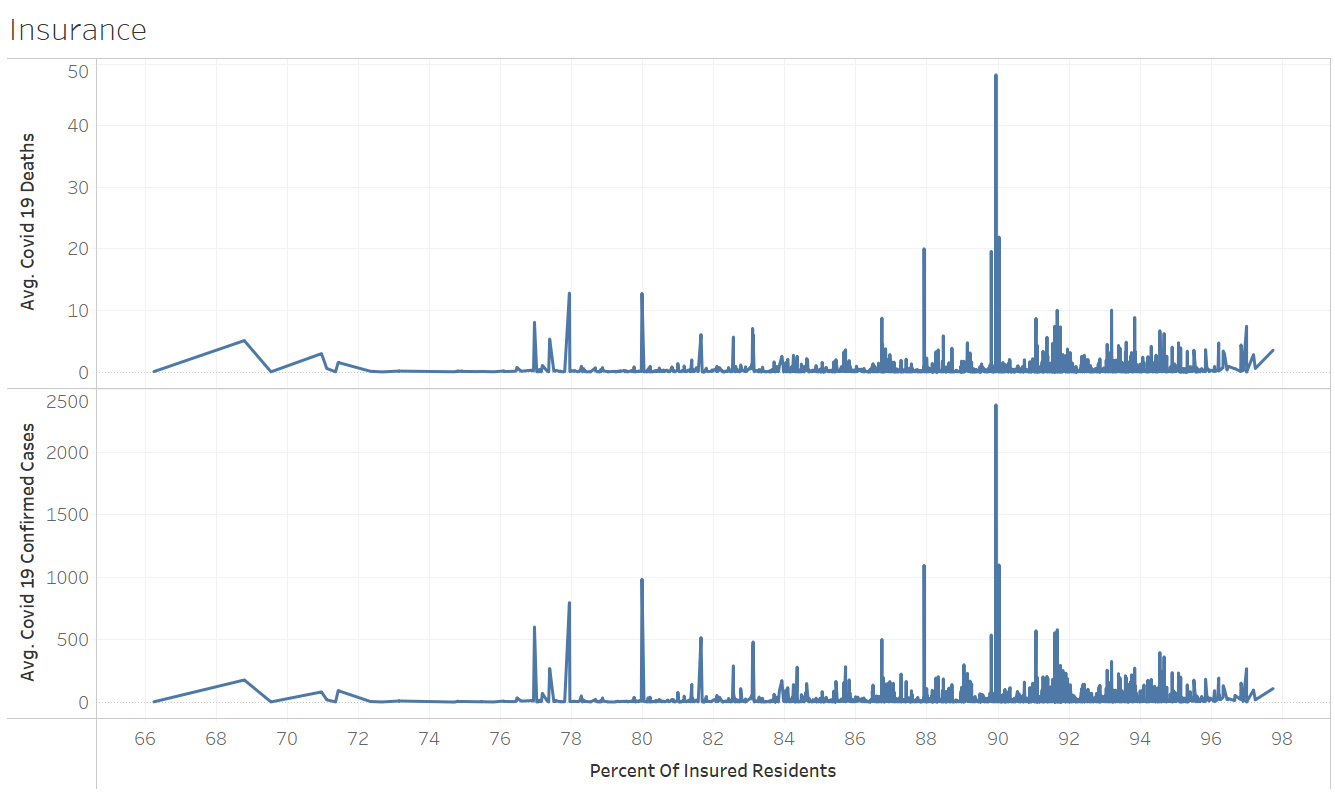
**Analysis**

**Percent of Insured Residents:** This variable is chosen because it can be assumed that people having insurance were able to approach advanced treatment and get cured from Covid-19.

The graph for insured residents throughout this time period is shown as:



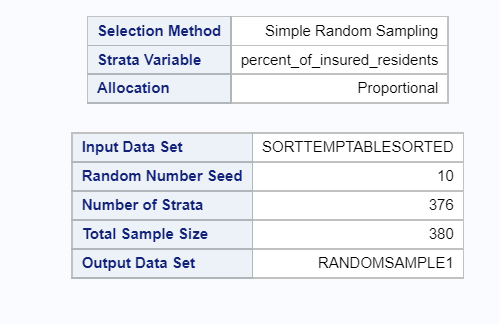
This is clarified that this variable is almost constant throughout this period.

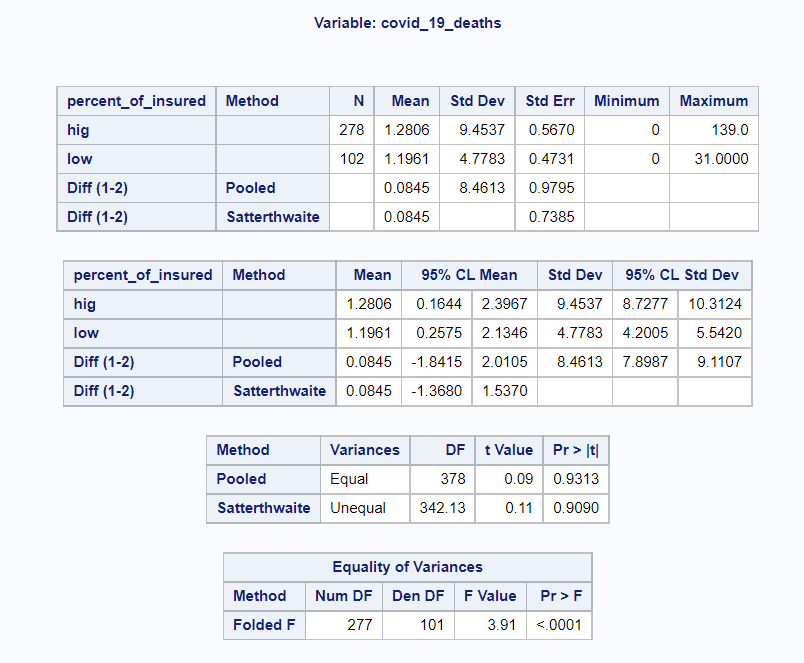


These two graphs are consistent with each other as these show similar spikes like when 90% of the insured residents got Covid-19 and they eventually died also. The association of Covid-19 confirmed cases was very small as observed by calculating correlation coefficient. But these graphs also counter the assumption that insured people can get better treatment from advanced as they were died with similar ratio of confirmed cases among these states.

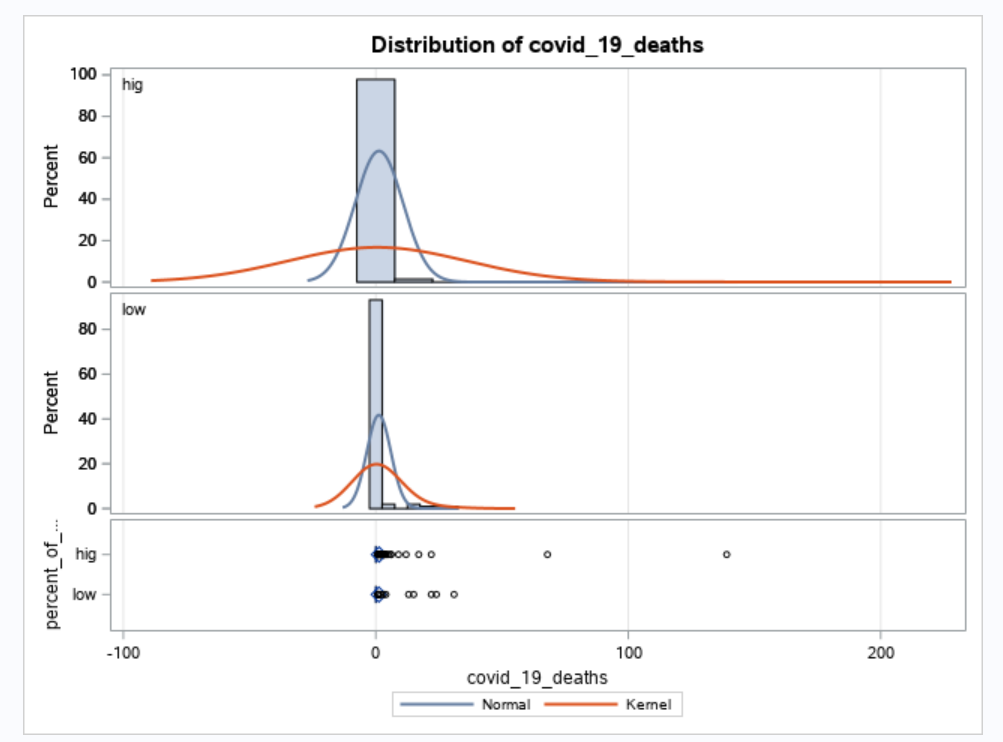
**T-test**

The samples have been stratified from the sample set of these 5 states and t-test is operated on these sample versus Covid-19 deaths. To make the consistency, t-test has been done for the sample only, not with the population dataset.





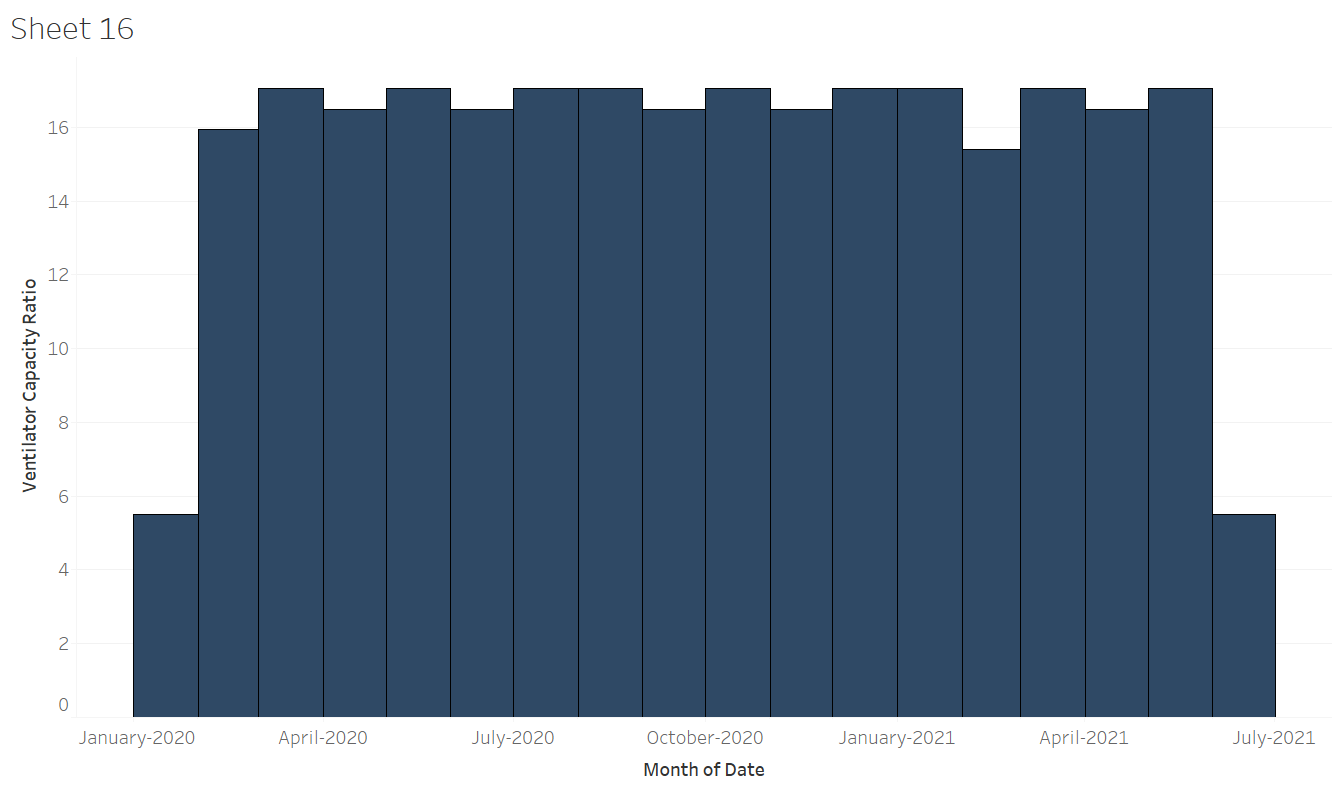
As the means for these sample sets are different, thus **Folded F test** is significant to be studied. The t value is very small as well as the p-value. But p-value is less than 0.05, thus we can reject null hypothesis. And these two samples have different variance.



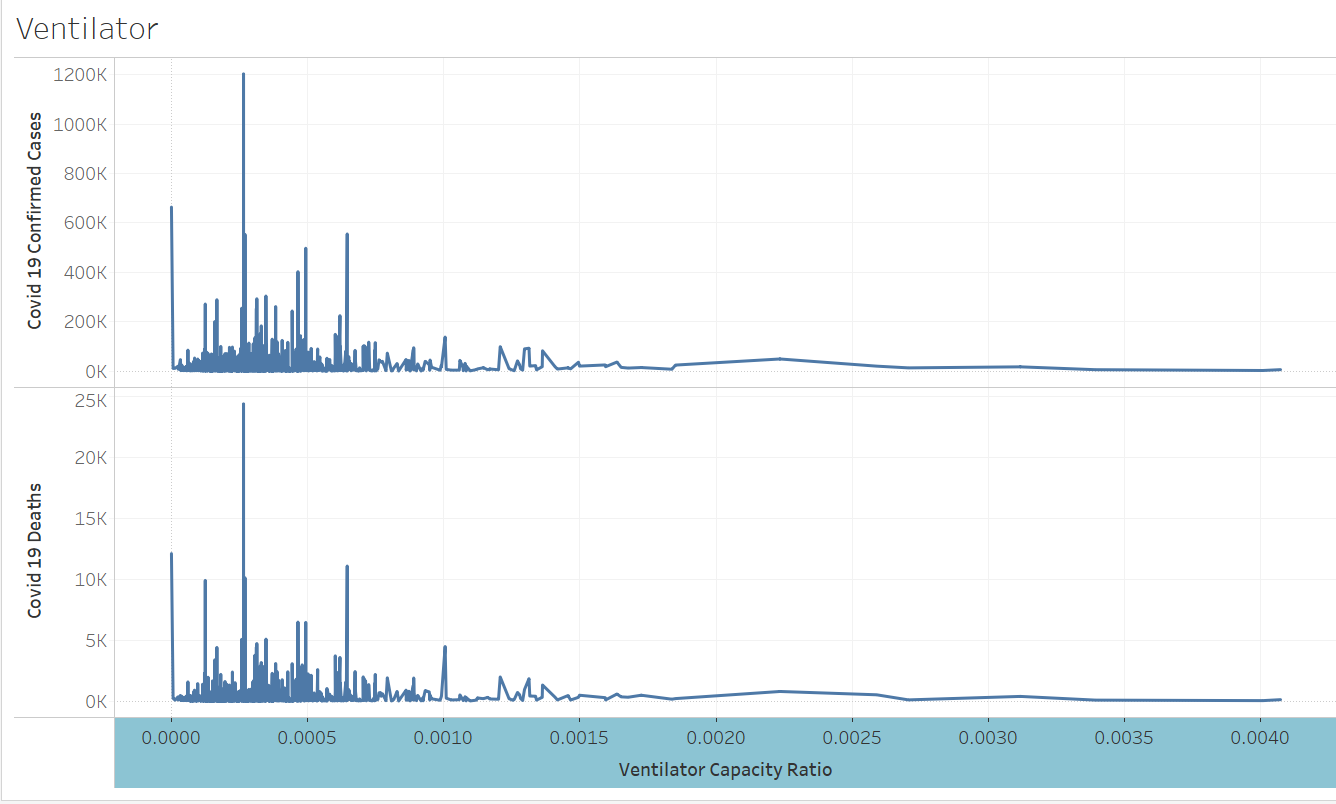
Also, the distribution plots among high as well as low strata for insured residents versus Covid-19 deaths has shown that these samples are right skewed not normal distributed. Thus, two-way t-test describes that mean of sample falls in upper tail of normal distribution.

**Ventilators Capacity Ratio:** Ventilators were hard to manage during Covid-19 pandemic and people were dying because of lack of ventilators in the hospital. This analysis will try to find some significant relationship between these two variables.

The following graph shows that ventilators ratio was not increasing by significant numbers, or this was almost stable variable throughout this period.



By observing the upcoming graph, it is again clear that the association of Covid-19 confirmed cases and Covid-19 deaths were consistent as both plots have similar trends. But because of low correlation coefficient, this is hard to conclude the trend among these variables with ventilators ratio.

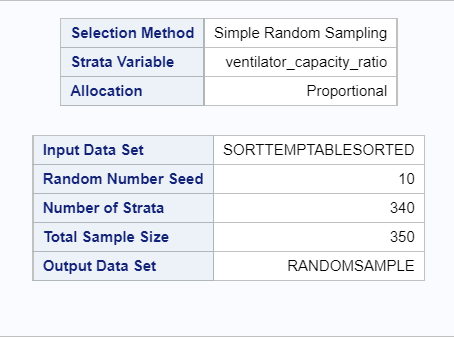


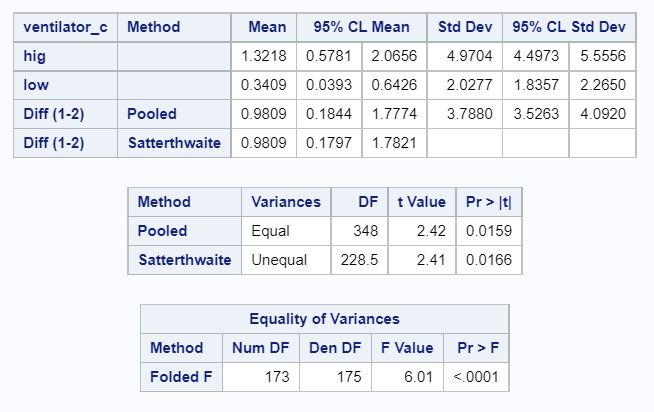
Also, it could be genuine to infer that as during initial period of pandemic, Covid-19 confirmed cases were shot up and the ration of ventilators were low, and more people were dying. Yet the ratio increased by a small fraction and Covid-19 cases were reduced by that time and consequently deaths were reduced.

**T test**

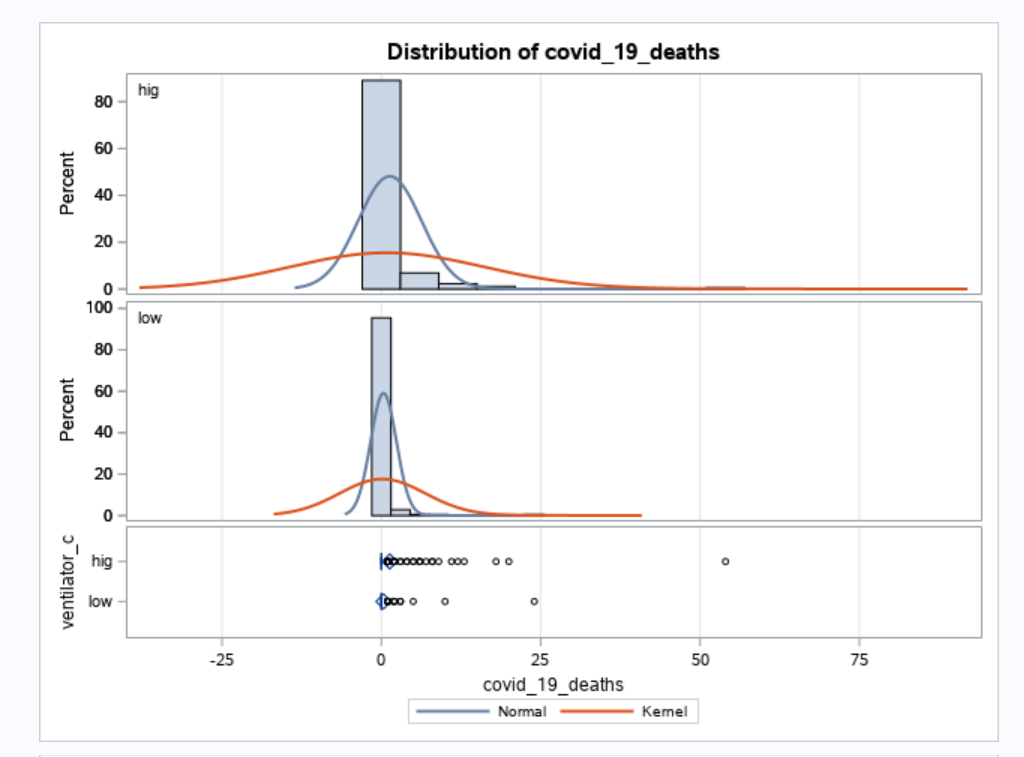
T- test has been performed to understand the trend of pattern between Covid-19 deaths with ventilators capacity ratio. The results are shown below:

This analysis is conducted with a similar process to the above. High and Low strata have been taken from the sample set.





Because of different means of two stratified samples, **Folded F test** is being used. The t value is again as small as 2.41 and this is a two-way test and distribution is right skewed, but p-value is less than 0.05. So, this can be concluded that these two strata have different variance from each other.



Again, for this association, t-test depicts that the mean value lies in upper tail and hence the sample dataset is not normally distributed according to this analysis.

**Conclusion:** It can be concluded that the analysis fails to completely justify the association of target variables with health factors variables. Only ventilators capacity ratio somewhat justifies the relationship. Hence, these variables can be considered less important for the analysis purpose because the analysis variables were almost stable with time period.

**Summary:**

* Top five states according to Covid-19 Confirmed Cases are chosen.
* Analysis variables are chosen according to the correlation coefficients.
* Percent of Insured Residents were hardly varying throughout this time period.
* Covid- 19 Confirmed Cases were consistent with Covid-19 deaths and both variables are not justifying any relationship with analysis variable.
* T-test is conducted on two different means of the samples and it proved that sample have different variances.
* Ventilators Capacity Ratio was almost constant throughout this period.
* Covid- 19 Confirmed Cases were consistent with Covid-19 deaths and both variables are not justifying any relationship with Ventilators Ratio.
* T-test has been conducted on different strata of mean values and justifies that mean lies on upper tail and samples have different means.

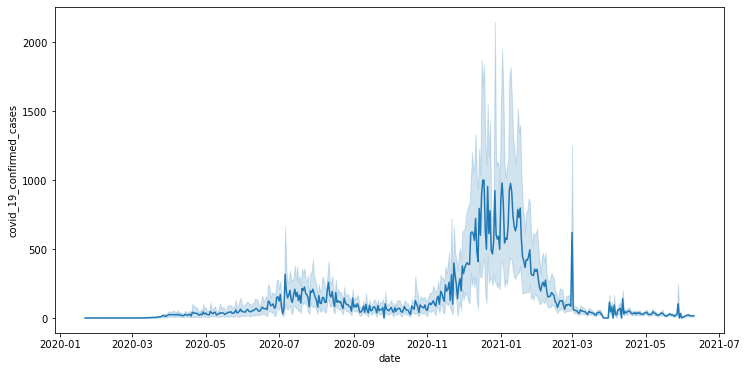
**Analysis 2**

# **Report**

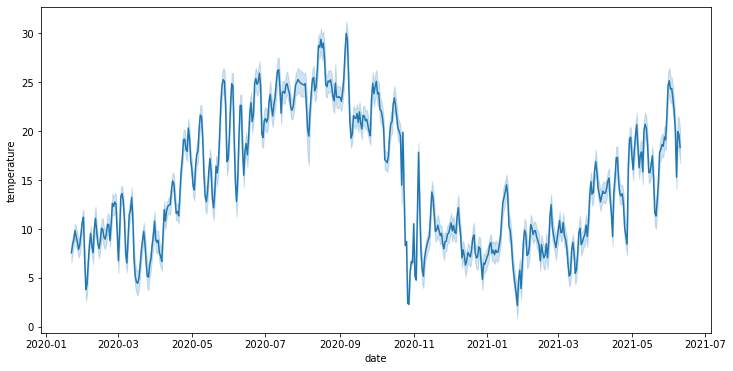
In this second part, t-test and prediction of covid\_19\_confimed\_cases versus of climatic features (temperature and precipitation) are analyses.

# **T-Test**

In the descriptive analysis it was shown that for the for the highest ten states, there is a weak correlation between temperature and covid\_19\_confimed\_cases (or covid\_19\_deaths). If we choose California state (with the most covid\_19\_confimed\_cases) and plot covid\_19\_confimed\_case versus date as shown in the following, we can find that in the winter season (from 2020-11 to 2021-3) there is much more cases of covid\_19\_confirmed\_cases than other seasons:



Therefore, it seems that the mean value of covid\_19\_confimed\_cases of winter period is different from the mean value of covid\_19\_confimed\_cases of other seasons. To do t-test, we have to define a new categorical variable for the temperature variable having two states: low and high. If we plot temperature versus date for the whole data, we will find:



From this picture we can choose the threshold value for the temperature to suppurate temperature into two different group. Value of 15 degrees is selected. So, through all the data of California observations if the temperature has got below 15 degrees, we define ‘low’ label and upper that 15 degrees we will have ‘high’ label for the new categorical variable called temp\_cat. The next step is sampling. Using *stratified sampling* with ‘temp\_cat’ as the strata with a sample size of 5000 samples per strata in SAS, our new data is ready for t-test analysis. Now, with the defined variable of temp\_cat with exactly two categories and the continuous dependent variable of covid\_19\_confirmed\_cases, we can do t-test analysis as follows:

At first, we define null and alternative hypothesis as:

The null hypothesis (H0) and alternative hypothesis (H1) of the Independent Samples t-test can be expressed as:

H0: µ1 = µ2 ("the two population means are equal")  
H1: µ1 ≠ µ2 ("the two population means are not equal")

where µ1 and µ2 are the population means for group 1 (low temperature population) and group 2 (high temperature population), respectively.

Before we perform our hypothesis tests, we should decide on a significance level (denoted α). The significance level is the threshold we will use to decide whether a test result is significant. For this project, let's use α = 0.05, or 5%.

Now, we look at descriptive statistics and graphs to get an idea of the differences in the groups' distributions, means, and variances. We compare the means and standard deviations of covid\_19\_confirmed\_cases for the low and high temp\_cat categorical variable, as well as graphs that compare the distribution of covid\_19\_confirmed\_cases for low versus high temp\_cat

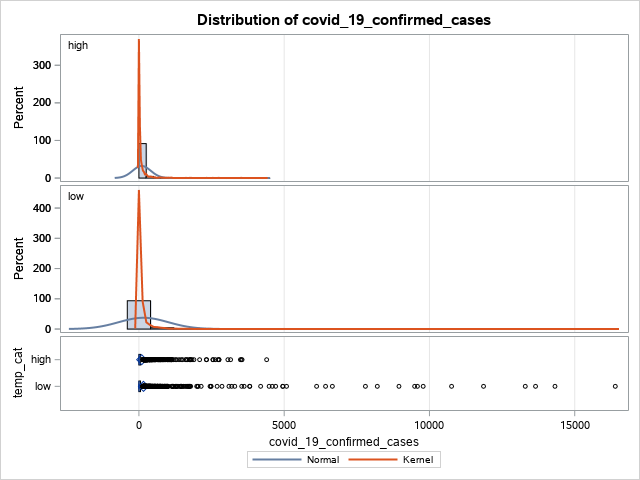
**The MEANS Procedure**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Analysis Variable: covid\_19\_confirmed\_cases** | | | | | | |
| **temp\_cat** | **N Obs** | **N** | **Mean** | **Std Dev** | **Minimum** | **Maximum** |
| high | 2193 | 2193 | 95.0889 | 309.528 | 0 | 4396.00 |
| low | 2807 | 2807 | 157.2034 | 855.570 | 0 | 16396.00 |

From this table, we can see that:

* There are 2193 high and 2807 low (totally 5000 samples).
* The mean covid\_19\_confimed\_case for high category is 95, while the mean covid\_19\_confimed\_case for low is 157. This is a difference of 62 persons.
* The standard deviation for high category of confimed\_case is around 310 persons, while the standard deviation of the low category of confimed\_case is near 856 persons.

Looking at the graphs from the PROC TTEST output in SAS:



The top two panels in the histogram show the distribution of the covid\_19\_confimed\_cases for high and low categories, respectively. We can see that the distribution of covid\_19\_confimed\_cases for both these two categories is approximately symmetric, but the data range for the low category is much larger. In the bottom panel, we see comparative boxplots of the same data. It is clear that low category has much more outliers than high category.

Four tables appear in the PROC TTEST output.

**Variable: covid\_19\_confirmed\_cases**

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **temp\_cat** | **Method** | **N** | **Mean** | **Std Dev** | **Std Err** | **Min** | **Max** |
| **high** |  | 2193 | 95.089 | 309.5 | 6.6097 | 0 | 4396.0 |
| **low** |  | 2807 | 157.2 | 855.6 | 16.1486 | 0 | 16396.0 |
| **Diff (1-2)** | **Pooled** |  | -62.115 | 673.0 | 19.1816 |  |  |
| **Diff (1-2)** | **Satterthwaite** |  | -62.115 |  | 17.4489 |  |  |

The first table contains descriptive statistics for both groups, including the valid sample size (n), mean, standard deviation, standard error (s/sqrt(n)), minimum, and maximum. Much of this we already saw in the PROC MEANS output, but this table also contains the computed difference between the two means. In this case, the first mean (high) was 62 persons lower than the second mean (low). In plain English, this means that, on average, in the winter times, 62 persons people get covid\_19 more comparing to the summer times. Keep in mind that the independent samples t test is testing whether or not this difference is statistically different from zero.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **temp\_cat** | **Method** | **Mean** | **95% CL Mean** | | **Std Dev** | **95% CL Std Dev** | |
| **high** |  | 95.089 | 82.1270 | 108.1 | 309.5 | 300.6 | 319.0 |
| **low** |  | 157.2 | 125.5 | 188.9 | 855.6 | 833.8 | 878.6 |
| **Diff (1-2)** | **Pooled** | -62.115 | -99.719 | -24.510 | 673.0 | 660.1 | 686.5 |
| **Diff (1-2)** | **Satterthwaite** | -62.115 | -96.325 | -27.904 |  |  |  |

The second table contains confidence limits for the group means, confidence limits for the group standard deviations, and confidence limits for the difference in the means. We specifically want to focus on the confidence intervals for the difference in the means. Notice that there are two different confidence interval formulas for the difference. The first, Pooled, assumes that both groups have the same variance in covid\_19\_confirmed\_cases. The second, Satterthwaite, does not make this assumption (i.e., it takes into account that one group has a different variance in covid\_19\_confirmed\_cases than the other). We know from our exploratory data analysis that the athletes and non-athletes have different variances, so we should look at the Satterthwaite confidence interval. The 95% confidence interval for the difference in the covid\_19\_confirmed\_caes is (27.904, 96.325).

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Method** | **Variances** | **DF** | **t Value** | **Pr > |t|** |
| **Pooled** | Equal | 4998 | -3.24 | 0.0012 |
| **Satterthwaite** | Unequal | 3692.3 | -3.56 | 0.0004 |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Equality of Variances** | | | | |
| **Method** | **Num DF** | **Den DF** | **F Value** | **Pr > F** |
| **Folded F** | 2806 | 2192 | 7.64 | <.0001 |

Tables 3 and 4 contain the independent samples t test and Folded F test, respectively. This time, we had ample graphical evidence of unequal variances between the groups, so we can use the Folded F test to see if the difference in the variances is significant. Recall that the null hypothesis of this test is that the variances are equal; the alternative is that the variances are not equal. Because the p-value is less than alpha (.05), we reject the null hypothesis, and conclude that the variance of the covid\_19\_confirmed\_cases is different for these two groups. Because of this, we will use the Satterthwaite version of the test.

Going back to table 3, we now compare the Satterthwaite t-test's p-value (0.0004) against our chosen significance level alpha (.05). Since the p-value is smaller than alpha, we reject the null hypothesis.

# **Decision and Conclusions**

Since p=0.0004 is less than our chosen significance level α = 0.05, we can reject the null hypothesis, and conclude that the that the mean covid\_19\_confirmed\_cases for low and high category is significantly different.

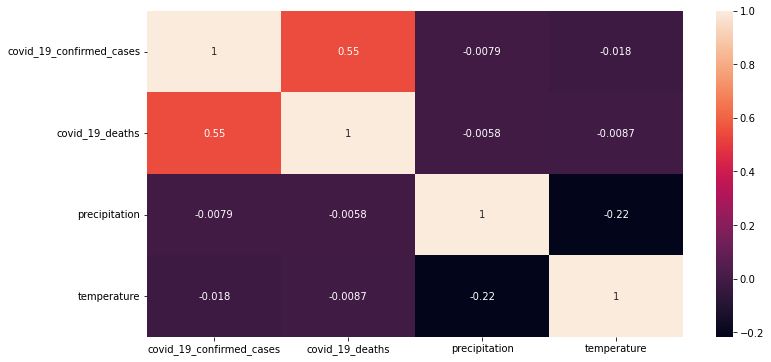
Based on the results, we can state the following:

* There was a significant difference in mean covid\_19\_confirmed\_cases between low and high.
* The average covid\_19\_confirmed\_cases for low temperature times (winter) are 62 persons more than the average covid\_19\_confirmed\_cases for high temperature times (summer). The main reason would be that in winter periods because of the cold weather, people met each other at indoor places and therefore this contagious virus would definitely did his job better!

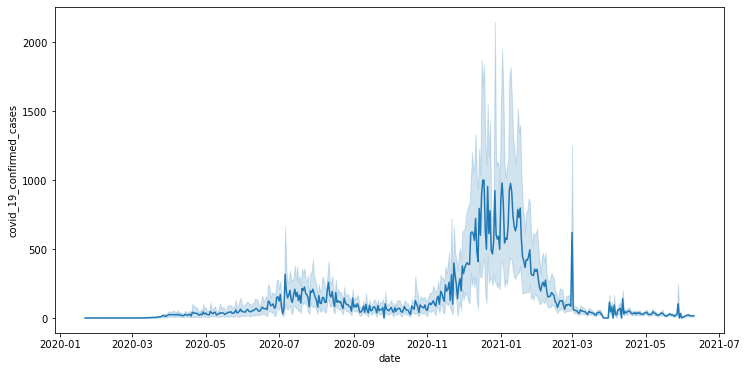
Similar tasks have been done for the other highest 9 states and approximately the same results have been extracted.

# **Prediction**

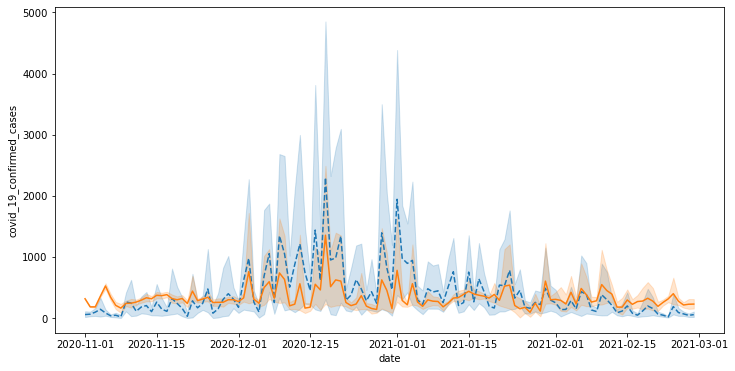
In this part, prediction process has been done to predict the covid\_19\_confirmed\_case through climatic features. At first, heatmap plot is drawn to show the correlation between our variables as follows:



This figure shows that there is strong correlation between covid\_19\_confimred\_ cases and covid\_19\_deaths. However, other variables of temperature and precipitation with covid\_19\_confimred\_ cases and covid\_19\_deaths have very weak correlations. If we plot temperature variable with date the following figure appears:



This figure shows that in the winter season (from 2020-11 to 2021-3) number of covid\_19\_confirmed cases are much more in comparison with summer periods. So, we can try regression model for prediction. All the codes in this part are written in python. The result is shown in the following figure:



The blue plot with ‘--’ linestyle is the real one and the orange plot with ‘-’ linestyle is the predicted one. To show the effectiveness of linear regression, R2-Square and root mean square indexes are calculated as follows:

R2-Square: 0.5419278790082807

Root Mean Square: 590.9065184219806

Best values for R2-Square are 1 and for the Root Mean Square is 0. So, comparing these two-pair variables depicts that approximately we can predict the covid\_19\_confimred\_ cases through climatic variables.

Similar tasks have been done for the other highest 9 states and approximately the same results have been extracted.

**Analysis 3**

# **Sociological Features**

In this report, we are going to analyze if the social distancing grades are associated with covid-19 confirmed cases.

Chart, histogram

Description automatically generated

Fig: covid-19 confirmed cases with social distancing total grade

Chart, bar chart

Description automatically generated

Fig: covid-19 confirmed cases with social distancing total grade

Chart, bar chart

Description automatically generated

Fig: covid-19 confirmed cases with social distancing encounters grade

Chart, bar chart

Description automatically generated

Fig: covid-19 confirmed cases with social distancing visitation grade

Chart, bar chart

Description automatically generated

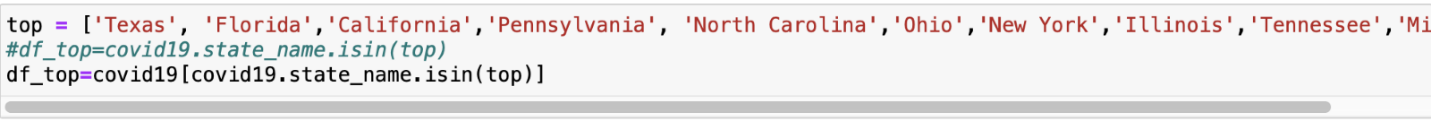
Fig: covid-19 confirmed cases with social distancing travel distance grade

# **T Test for Sociological Features**

For stratified sample: Top 10 states with highest covid confirmed cases

We are going to do T Test to determine if there is a significant difference between the means of two groups e.g., social distancing total grade A vs social distancing total grade F.

Before we do the T Test, we are going to get the sample dataset with top 10 states with highest covid confirmed cases. And the independent variables should be with 2 categories.



Text

Description automatically generated

T Test for social distancing total grade

Graphical user interface, text, application

Description automatically generated

Since p = 0.48, we accept the null hypothesis that the mean covid-19 confirmed cases of grade A is equal to the mean covid-19 confirmed cases of grade F.

T Test for social distancing visitation grade

Text

Description automatically generated

Since p < 0.05, we reject the null hypothesis that the mean covid-19 confirmed cases of grade A is equal to the mean covid-19 confirmed cases of grade F. So the means of these two groups are different from each other.

T Test for social distancing travel distance grade

Text

Description automatically generated

Since p < 0.05, we reject the null hypothesis that the mean covid-19 confirmed cases of grade A is equal to the mean covid-19 confirmed cases of grade F. So the means of these two groups are different from each other.

T Test for social distancing encounters grade

Text

Description automatically generated

Since p = 1.415, we accept the null hypothesis that the mean covid-19 confirmed cases of grade A is equal to the mean covid-19 confirmed cases of grade F.

From the above hypothesis, social distancing visitation grade and social distancing travel distance grade (A or F) can affect the covid-19 confirmed cases.

There is statistical evidence that the associated population means are significantly different.

# **Z Test for stratified sample: Top 10 states with highest covid confirmed cases**

Since the sample size is more than 30 after we do stratified sampling, we will do the Z Test to determine whether two population means are different.

Text

Description automatically generated

In this Z Test, p value= 0.00013, we reject the null hypothesis that the population means are the same and z score is 3.814.

# **Prediction on covid-19 confirmed cases**

The model we use is XGBRegressor (eXtreme Gradient Boosting Regressor).

Chart, histogram

Description automatically generated

This figure is showing the comparison of the original value and predicted values of covid-19 confirmed cases.

# **Recommendations and conclusions**

Based on the descriptive analysis we have done, the graphs are showing that the covid-19 confirmed cases are higher when the social distancing grade is poor. When there are more social encounters or more social visitation or travelling more distances, the covid-19 virus will spread to more people.

After descriptive analysis, we continue to do inferential analysis to identify if the means of the different social distancing grades are equal or not. The results of the inferential statistics (T Test and Z Test) are showing that there is a significant difference between the means of two groups of social distancing grades.

As a conclusion, Social distancing is one of the key factors to slow down the spread of covid-19 virus and prevent the new variant of virus. Individual may follow the social distancing restrictions by staying home, limiting social encounters, limiting travel, avoiding crowded areas and physically distancing from others.

# **Bibliography**

Analytics Vidhya. 2022. *XGBRegressor |Predict Future Sales using XGBRegressor*. [online] Available at: <https://www.analyticsvidhya.com/blog/2021/06/predict-future-sales-using-xgbregressor/> [Accessed 27 March 2022].

**Appendix**

**Analysis 1: SAS Codes**

%web\_drop\_table(WORK.IMPORT);

FILENAME REFFILE '/home/u60688886/covid/imputed-data.csv';

PROC IMPORT DATAFILE=REFFILE

DBMS=CSV

OUT=WORK.IMPORT;

GETNAMES=YES;

RUN;

PROC CONTENTS DATA=WORK.IMPORT; RUN;

/\* State select \*/

data work.state\_select;

set work.import;

where state\_name = 'Califor' or state\_name = 'Florida' or state\_name = 'Texas' or state\_name = 'New Yor' or state\_name = 'Illinoi';

run;

/\*Correlation \*/

ods noproctitle;

ods graphics / imagemap=on;

proc corr data=WORK.STATE\_SELECT pearson nosimple plots=none;

var covid\_19\_confirmed\_cases covid\_19\_deaths;

with hospital\_beds\_ratio ventilator\_capacity\_ratio

'intensive\_care\_unit\_(ICU)\_bed\_ra'n percent\_of\_smokers percent\_of\_diabetes

percent\_of\_insured\_residents;

run;

/\*Series plot for states \*/

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

proc sort data=WORK.STATE\_SELECT out=\_SeriesPlotTaskData;

by state\_name;

run;

proc sgplot data=\_SeriesPlotTaskData;

series x=state\_name y=covid\_19\_confirmed\_cases /;

xaxis grid;

yaxis grid;

run;

ods graphics / reset;

proc datasets library=WORK noprint;

delete \_SeriesPlotTaskData;

run;

/\* Scatter plot for insured vs confirmed \*/

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

proc sgplot data=WORK.STATE\_SELECT;

scatter x=percent\_of\_insured\_residents y=covid\_19\_confirmed\_cases /;

xaxis grid;

yaxis grid;

run;

ods graphics / reset;

/\* Scatter plot for ventilators vs confirmed \*/

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

proc sgplot data=WORK.STATE\_SELECT;

scatter x=ventilator\_capacity\_ratio y=covid\_19\_confirmed\_cases /;

xaxis grid;

yaxis grid;

run;

ods graphics / reset;

/\* Scatter plot for insured vs deaths \*/

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

proc sgplot data=WORK.STATE\_SELECT;

scatter x=percent\_of\_insured\_residents y=covid\_19\_deaths /;

xaxis grid;

yaxis grid;

run;

ods graphics / reset;

/\* Scatter plot for ventilators vs deaths \*/

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

proc sgplot data=WORK.STATE\_SELECT;

scatter x=ventilator\_capacity\_ratio y=covid\_19\_deaths /;

xaxis grid;

yaxis grid;

run;

ods graphics / reset;

/\*Stratified sampling \*/

proc sort data=WORK.STATE\_SELECT out=WORK.SORTTempTableSorted;

by ventilator\_capacity\_ratio;

run;

proc surveyselect data=WORK.SORTTempTableSorted out=work.RandomSample

method=srs sampsize=350 seed=10;

strata ventilator\_capacity\_ratio / alloc=prop;

run;

proc print data=work.RandomSample(obs=10);

title "Subset of work.RandomSample";

run;

title;

proc delete data=WORK.SORTTempTableSorted;

run;

proc sort data=WORK.STATE\_SELECT out=WORK.SORTTempTableSorted;

by percent\_of\_insured\_residents;

run;

proc surveyselect data=WORK.SORTTempTableSorted out=work.RandomSample1

method=srs sampsize=380 seed=10;

strata percent\_of\_insured\_residents / alloc=prop;

run;

proc print data=work.RandomSample1(obs=10);

title "Subset of work.RandomSample1";

run;

title;

proc delete data=WORK.SORTTempTableSorted;

run;

ods noproctitle;

ods graphics / imagemap=on;

proc means data=WORK.RandomSample chartype mean std min max n vardef=df;

var covid\_19\_deaths ventilator\_capacity\_ratio;

run;

ods noproctitle;

ods graphics / imagemap=on;

proc means data=WORK.RandomSample1 chartype mean std min max n vardef=df;

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

run;

/\*High-low binning \*/

proc hpbin data=work.randomsample output=randombin numbin=2 pseudo\_quantile;

input ventilator\_capacity\_ratio ;

run;

data ventilator;

set work.randomsample;

if (ventilator\_capacity\_ratio < 0.0002316286) then ventilator\_c = "low";

if (0.0002316286 <= ventilator\_capacity\_ratio) then ventilator\_c = "high";

run;

data insure;

set work.randomsample1;

if 65<(percent\_of\_insured\_residents)< 81 then percent\_of\_insured ='low';

if 81<(percent\_of\_insured\_residents) > 81 then percent\_of\_insured = 'high';

run;

/\*Ventilators t-test \*/

ods noproctitle;

ods graphics / imagemap=on;

/\* Test for normality \*/

proc univariate data=WORK.VENTILATOR normal mu0=0;

ods select TestsForNormality;

class ventilator\_c;

var covid\_19\_deaths;

run;

/\* t test \*/

proc ttest data=WORK.VENTILATOR sides=2 h0=0 plots(showh0);

class ventilator\_c;

var covid\_19\_deaths;

run;

/\* two way t-test for insured \*/

ods noproctitle;

ods graphics / imagemap=on;

/\* Test for normality \*/

proc univariate data=WORK.INSURE normal mu0=0;

ods select TestsForNormality;

class percent\_of\_insured;

var covid\_19\_deaths;

run;

/\* t test \*/

proc ttest data=WORK.INSURE sides=2 h0=0 plots(showh0);

class percent\_of\_insured;

var covid\_19\_deaths;

run;

**Analysis 2: Python Codes**

#prediction Codes:

from pyforest import \*

import warnings

warnings.filterwarnings("ignore")

import pandasql as sql

lazy\_imports()

import pandas as pd

df = pd.read\_csv('imputed-data.csv')

cases\_10\_whole=df.groupby('state\_name').sum().sort\_values(by = 'covid\_19\_confirmed\_cases', ascending = False).iloc[0:10]

cases\_10\_6\_col = df[(df['state\_name'] == 'California') | (df['state\_name'] == 'Texas') | (df['state\_name'] == 'Florida') | (df['state\_name'] == 'New York') | (df['state\_name'] == 'Illinois') | (df['state\_name'] == 'Pennsylvania') | (df['state\_name'] == 'North Carolina') | (df['state\_name'] == 'Ohio') | (df['state\_name'] == 'Michigan') | (df['state\_name'] == 'New Jersey')][['date', 'state\_name', 'covid\_19\_confirmed\_cases', 'covid\_19\_deaths', 'precipitation', 'temperature']]

cases\_10\_6\_col['state\_name'].value\_counts()

cases\_10\_6\_col.to\_csv("Cases\_data.csv")

cases\_10\_6\_col.date = pd.to\_datetime(cases\_10\_6\_col.date)

California\_data = cases\_10\_6\_col[cases\_10\_6\_col['state\_name'] == 'California']

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

Var\_Corr = California\_data.corr()

sns.heatmap(Var\_Corr, xticklabels = Var\_Corr.columns, yticklabels = Var\_Corr.columns, annot = True)

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

sns.boxplot('covid\_19\_confirmed\_cases', data=California\_data, orient='h')

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

sns.lineplot(x='date', y='covid\_19\_confirmed\_cases', data=California\_data)

California\_data[California\_data['temperature'] < -20] = -10

California\_data[California\_data['temperature'] > 50] = 40

California\_data['temperature'] = California\_data['temperature'].astype(int)

bins=[]

def create\_bins(low, width, quantity):

    for low in range(low, quantity\*width+1, width):

        bins.append((low, low+width))

create\_bins(low=min(California\_data['temperature']), width=10, quantity=6)

print(bins)

print(max(California\_data['temperature']))

bins2 = pd.IntervalIndex.from\_tuples(bins, closed="left")

California\_data['Temp\_Cat'] = pd.cut(California\_data['temperature'],bins2)

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

sns.barplot(x = 'Temp\_Cat',y = 'covid\_19\_confirmed\_cases',data = California\_data)

df\_train = pd.get\_dummies(California\_data)

df\_train[['covid\_19\_confirmed\_cases']] = df\_train[['covid\_19\_confirmed\_cases']].ffill()

df\_train[['temperature']] = df\_train[['temperature']].fillna(0)

df\_train[['covid\_19\_confirmed\_cases']] = df\_train[['covid\_19\_confirmed\_cases']].fillna(method='ffill')

from sklearn import linear\_model

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

from sklearn.preprocessing import StandardScaler

X\_train, X\_test, y\_train, y\_test = train\_test\_split( df\_train.loc[:,~df\_train.columns.isin(['covid\_19\_confirmed\_cases','date'])],

                                   df\_train['covid\_19\_confirmed\_cases'],test\_size = 0.3, shuffle = True)

sc = StandardScaler()

X\_train\_sc = sc.fit\_transform(X\_train)

X\_test\_sc = sc.transform(X\_test)

mul\_linreg = linear\_model.LinearRegression().fit(X\_train\_sc,y\_train)

pred\_lin = mul\_linreg.predict(X\_test\_sc)

metrics.r2\_score(y\_test, pred\_lin)

np.sqrt(metrics.mean\_squared\_error(y\_test, pred\_lin))

y\_test.mean()

pred\_lin.mean()

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

plot\_test = California\_data.loc[y\_test.index]

plot\_test.head()

sns.lineplot(x = plot\_test['date'],y = plot\_test['covid\_19\_confirmed\_cases'],legend = True)

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

plot\_test = California\_data.loc[y\_test.index]

plot\_test.head()

plot\_test['predicted\_covid\_19\_confirmed\_cases'] = pred\_lin

sns.lineplot(x = plot\_test['date'],y = plot\_test['covid\_19\_confirmed\_cases'],linestyle='--' ,legend = True)

sns.lineplot(x = plot\_test['date'],y = plot\_test['predicted\_covid\_19\_confirmed\_cases'],legend = True)

#preparing data for t-test

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

sns.lineplot(x='date', y='temperature', data=California\_data)

California\_data['temp\_cat'] = California\_data['temperature'].apply(lambda x: 'low' if x <= 15 else 'high')

California\_data.to\_csv("California\_data\_t\_test.csv")

**SAS Codes:**

proc import datafile='/home/u60674391/sasuser.v94/project/California\_data\_t\_test.csv' dbms=csv out=work.cal\_temp\_t\_test replace;

getnames=yes;

guessingrows=max;

run;

PROC MEANS data=work.cal\_temp\_t\_test;

   VAR covid\_19\_confirmed\_cases;

   CLASS temp\_cat;

RUN;

/\* stratified sampling with ‘temp\_cat’ as the strata with a sample size of 5000 samples per strata. \*/

proc sort data=WORK.cal\_temp\_t\_test out=WORK.sampled\_cal\_temp\_t\_test;

by temp\_cat;

run;

proc sort data=WORK.SAMPLED\_CAL\_TEMP\_T\_TEST out=WORK.SORTTempTableSorted;

by temp\_cat;

run;

proc surveyselect data=WORK.SORTTempTableSorted out=work.RandomSample

method=srs sampsize=5000 seed=13;

strata temp\_cat / alloc=prop;

run;

proc print data=work.RandomSample(obs=10);

title "Subset of work.RandomSample";

run;

title;

proc delete data=WORK.SORTTempTableSorted;

run;

PROC MEANS data=work.RandomSample;

   VAR covid\_19\_confirmed\_cases;

   CLASS temp\_cat;

RUN;

/\* t-test  \*/

PROC TTEST data=work.RandomSample;

   VAR covid\_19\_confirmed\_cases;

   CLASS temp\_cat;

RUN;

**Analysis 3-**

Covid19\_Socio.ipynb is included in submission.