A

HOMEWORK ASSIGNMENT REPORT

ON

"HYPOTHESIS TESTING"

SUBMITTED BY:

ADARSH KESIREDDY

LIU CHIA-YU

VIRAJ BHAKTA

ZEQIAN FENG

SUBMITTED TO:

Dr. SCOTT KING

FALL 2020

Table of Contents

ABSTRACT 5
INTRODUCTION 6
Testing a Hypothesis: 7
Step 1: State your null and alternate hypothesis 7
Step 2: Collect data 7
Step 3: Perform a statistical test 7
Step 4: Determine whether to accept or refute the null hypothesis 7
Step 5: Present your findings 7
Team Roles 8
DATASET DESCRIPTION: 10
PACKAGES: 10
WORKFLOW: 11
DATA ACQUIRATION: 11
DATA PROCESSING: 11
DATA ANALYSIS: 11
HYPOTHESIS 1 11
Null and Alternative Hypothesis 11
Data 12
Statistical Test 12
Calculate Hospitalization Rate 12
Calculate Fatality Rate 12
Result 12
Conclusion 13
HYPOTHESIS 2 13
Null and Alterative Hypothesis 13
Data 14
Statistical Test 14

Calculate ICU Rate 14
Calculate Fatality Rate 14
Result 14
Conclusion 14
HYPOTHESIS 3 15
Null and Alternative Hypothesis 15
Data 15
Statistical Test 16
Calculate Transmission Rate 16
Result 16
Conclusion 17
HYPOTHESIS 4 17
Null and Alternative Hypothesis 17
Data 17
Statistical Test 18
Calculate Transmission Rate: 18
Result 18
Conclusion 18
HYPOTHESIS 5: 19
Null and Alternative Hypothesis: 19
Data 19
Statistical Test 19
Calculate Population Density 19
Calculate Transmission Rate 20
Result 20
Conclusion 21
HYPOTHESIS 6 21
Null and Alternative Hypothesis 21
Data: 22
Statistical Test: 22
Category 2 kinds of rural counties: 22
Infection Rate: 22
RESULT: 22

Conclusion 23
HYPOTHESIS 7: 23
Null and Alternative Hypothesis 23
Data 24
Statistical Test 24
Calculate Infection Rate: 24
Result 24
Conclusion 26
HYPOTHESIS 8: 27
Null and Alternative Hypothesis 27
Data: 27
Statistical Test: 27
Category 2 kinds of rural counties: 27
Infection Rate: 27
RESULT: 27
Conclusion 29
CONCLUSION 30
REFERENCES 30

ABSTRACT

This document demonstrates work done by Group 4, for the data analysis assignment. Initially, eight statements were presented. The team developed null and alternative hypothesis. Various csv, and excel files were imported into python and perform various types of statistical test. These files contain Coronavirus Disease 2019 (Covid-19) information for the state of Texas. Moreover, the file contained various data related to Texas and United States. The statistical test determines if the hypothesis is to be accepted or rejected. Overall, few null hypotheses were rejected. If null hypothesis is rejected then alternative hypothesis should be accepted, is one among the many assumptions made.

INTRODUCTION

Hypothesis testing is a systematic method for using statistics to analyze our theories about the universe. Scientists most frequently use it to test concrete predictions, called hypotheses, which originate from theories.

In hypothesis testing, there are 5 main steps:

1. As a null (Ho) and alternative (Ha) hypothesis, state your research hypothesis.
2. Gather knowledge in a way designed to test the hypothesis.
3. Perform a statistical test that is suitable.
4. Decide whether it is endorsed or denied by the null hypothesis.
5. Present the findings and outcomes.

The method you will use when testing a hypothesis will often follow some version of these measures, although the exact specifics details might vary.

Testing a Hypothesis:

Step 1: State your null and alternate hypothesis

It is important to restate it as a null (Ho) and alternate (Ha) hypothesis after establishing your initial research hypothesis (the prediction that you want to investigate) so that you can mathematically test it.

Typically, the alternative hypothesis is your original hypothesis that predicts a relationship between variables. A prediction with no relationship between the variables you are interested is known as null hypothesis.

Step 2: Collect data

It is important to carry out sampling and collect data in a way that is structured to validate the hypothesis for a statistical test to be true. You should not draw statistical inferences about the population about which you are interested if the data is not representative.

Step 3: Perform a statistical test

There are several statistical tests available, but all of them are focused on the comparison of variance within a group (how the data is distributed within a category) versus variance between groups (how different the categories are from each other).

If the difference between groups is high enough that there is little to no overlap between groups, then your statistical test will reflect that by showing a low *p*-value. This suggests it is impossible that by chance the variations between these groups came about.

Alternatively, if there were high variance within the group and low variance between groups, then with a high p-value, your statistical test would represent that. This means that any disparity you quantify between groups is likely to be due to chance. Your statistical test selection will be dependent on the type of data you have obtained.

Step 4: Determine whether to accept or refute the null hypothesis

Based on the findings of the statistical test, you will have to determine whether to accept or refute the null hypothesis. In most cases, to direct your decision, you will use the p-value provided by your statistical test. And in most situations, if the null hypothesis is correct, the cutoff for refuting the null hypothesis will be 0.05-that is, if there is a less than 5 percent probability that you might see these outcomes.

Step 5: Present your findings

The findings of hypothesis testing will be presented in the results.

Team Roles

Following are the various hypotheses which are needed to be tested by using various statistical tests.

1. A higher hospitalization rate gives a higher fatality rate
2. A higher ICU rate gives a higher fatality rate
3. As restriction lifted with Open Texas the rate of infection increased, with a slight delay
4. More testing leads to lower transmission later
5. Higher population density means higher transmission rate
6. Rural counties neighboring urban areas had higher rates than rural counties not near urban areas
7. Travel between regions affected spread
8. Rural counties had less spread (low population, as opposed to population density)

Roles	**Members**	**Responsibility**
Data Acquiring	Viraj Bhakta and Zeqian Feng	• Sort out the various data needed for this assignment from the group discussion. – Viraj • Accurately and completely collect the data needed for the job. If there is data that cannot be found, discuss how to replace it. – Viraj • Classify all collected data to clarify the specific purpose of the data. – Feng
Data Processing	Zeqian Feng, Liu Chia-Yu and Adarsh Kesireddy	• Data cleaning: The available data has lot of missing values (from Assignment 1). Data processing takes the team decision on how to handle the missing values and cleans the data. – Adarsh and Liu • Clean data will be passed to analyst and visualization team. Before passing on the data, processing team will make sure the data is clean up to their abilities. – Liu and Feng

Data Analyst	Adarsh Kesireddy, Liu Chia-Yu, Viraj Bhakta and Zeqian Feng	• Draw relationships and conclusion that help in proposed hypothesis from the obtained data. – teamwork during team meeting • Perform tests - Liu. • Pass information to visualization and paperwork – Liu, Adarsh and Viraj • Evaluation: quality of analysis. – teamwork during team meeting.
Visualization	Adarsh Kesireddy and Liu Chia-Yu	• Get result from analyst – Liu • Make them visually appeal – Liu and Adarsh • Clean chart that are readable – Liu and Adarsh • Code evaluation: quality of charts that are clean and readable. – Liu and present in team meeting
Paperwork and Team Organizing	Adarsh Kesireddy, Viraj Bhakta and Liu Chia-Yu	• Paperwork would be initializing and doing most of the report for the group. Teammates will be involved with corrections in the report. – Viraj • Presentation: Team inputs will be taken while preparing the presentation during the meeting and presentation will be done by. – Liu • Maintaining of weekly (meeting) report will be performed. – Adarsh • Meetings will be organized and maintained –Adarsh
GitHub	Adarsh Kesireddy	• Create GitHub repository. – Adarsh • Provide access to all teammates. – Adarsh

		• Monitor repository. – Adarsh

DATASET DESCRIPTION:

In addition to the latest coronavirus disease 2019 (COVID 19), which is causing an outbreak of respiratory disease worldwide, the Texas Department of State Health Services (DSHS) is working closely with the Centers for Disease Control and Prevention (CDC). All data comes from the coronavirus dashboard of the Texas Department of State Health Services (DSHS '). Several Excel files comprise the underlying data currently displayed on the COVID-19 DSHS Dashboard. On several tabs, data is shown and includes daily and cumulative case and fatality data, approximate recoveries, statewide hospital data. There is regular overwriting of records. The excel files that are included in the assignment is as follows:

Cases: Cases over Time by County

Fatalities: Fatalities over Time by County

Active Cases: Estimated Active Cases over Time by County

Total Tests: Cumulative Tests over Time by County

Hospitalization: Combined Hospital Data over Time by Trauma Service Area (TSA)

Population Density: Population Density by Counties in Texas

Rural and Urban Counties: PHR_MSA_County_masterlist

PACKAGES:

Import pandas

Pandas is a software package for python. It is a must to learn for data-science and dedicatedly written for Python language. It offers intuitive data structures as a simple, demonstrative, and customizable platform. With this incredible kit, you can easily manipulate any form of data, such as structured or time series data.

Import matplotlib.pyplot

A Python library that uses Python Script to write 2-dimensional graphs and plots is Matplotlib. Mathematical or science applications often involve a representation of more than a single axis. This library lets us build several plots at a time. However, you can use Matplotlib to even manipulate various characteristics of figures.

Import plotly

The Plotly Python library is an open-source interactive plot library that supports over 40 different types of plots covering a broad variety of statistical, economic, geographical, scientific and 3-dimensional use cases. Built on top of the Plotly JavaScript library (plotly.js), Plotly allows Python users to build beautiful web-based interactive visualizations that can be viewed in Jupyter notebooks, stored in standalone HTML files, or used as part of web applications that are solely Python-built using Dash.

WORKFLOW:

DATA ACQUIRATION:

- To deal with collecting information from various repositories for relevant dataset to enable for analyst to test the hypothesis.
- This task will be mainly be performed in the beginning of the project.
- Require a lot of work in gathering proper information that will be relevant for the analysis.

DATA PROCESSING:

- On identified datasets, data cleaning will be carried out, such as finding missing values or incorrect data or replacing the data with other values and making the datasets consistent with other datasets.
- To draw relationships between these datasets, cleaned data is supplied to the Data Analyst.

DATA ANALYSIS:

- Analysis of county data provided by the data processor to infer patterns and relationships in order to draw stronger conclusions that could support our proposed hypothesis.
- To prove our hypothesis, Statistical Software Packages will be used for statistical analysis.

HYPOTHESIS 1

The given statement is:

“A higher hospitalization rate gives a higher fatality rate”

As per the statement, as hospitalization rate increases then fatality rate will also increase.

Null and Alternative Hypothesis

First, a null and alternative hypothesis are generated. Null and alternative hypothesis are as shown below:

Null Hypothesis	Alternative Hypothesis
Statement: Hospitalization has no impact on fatality	Statement: Hospitalization has impact on fatality

Testing method: Fatality Rate = ((slope)*(Hospitalization rate)) + intercept If slope is zero, then the statement is true.	Testing method: Fatality Rate = ((slope)*(Hospitalization rate)) + intercept If slope is not zero, then the statement is true
Statistical Test: Linear regression t-test	Statistical Test: Linear regression t-test

Data

Once hypothesis was generated, data from hospitalization rate (CombinedHospitalDataoverTimebyTSA.xlsx) and fatality (TexasCOVID19DailyCountyFatalityCountData.xlsx) are imported into Python. The date ranges considered for this hypothesis is: 04/12 – 09/21. The data from the files are imported as list.

Statistical Test

Linear regression t-test was performed on the data.

Calculate Hospitalization Rate

Directly acquire the number in cumulative form from a sheet named "COVID Hospitalizations (%)"

Calculate Fatality Rate

1. Collect the fatality cases by county from a sheet "Fatalities by County"
2. Sum all fatality cases up by TSA area. This is considered as fatality rate for the problem.

Result

Up on performing t-test, we got a slope values as 0.337, and confidences level was 99.9% for Corpus Christi TSA county. The below image shows the slop and confidence level for Corpus Christi TSA county:

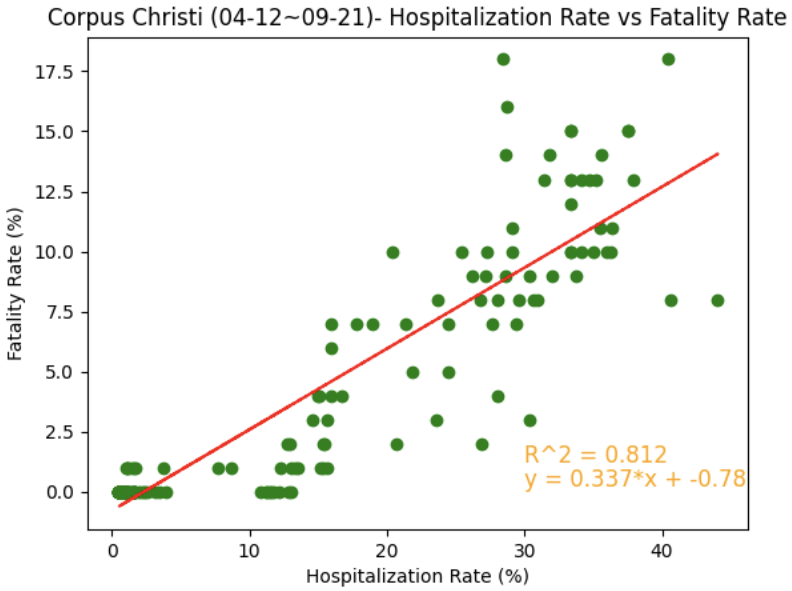

Figure 1 Corpus Christi linear regression hospitalization rate and fatality rate

Around 20 TSA counties were used for the analysis. Listing out those 20 counties is a tedious task, so below map was generated to show the fatality rate compared to hospitalization of the TSA counties. The white areas in the image are due to grouping issues.

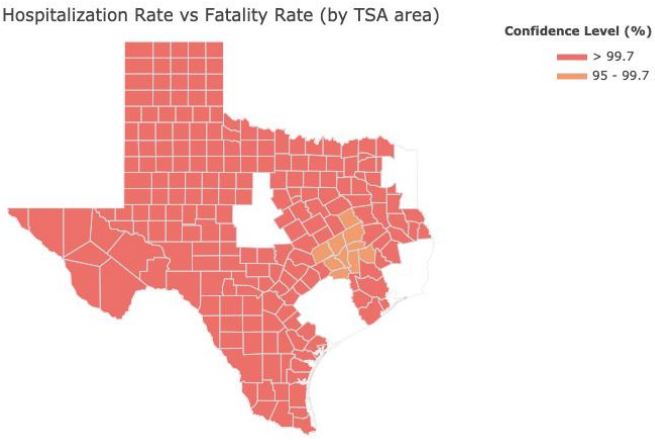

Figure 2 Hospitalization rate compared with Fatality rate using TSA counties

Conclusion

From the above, linear regression t-test the slope for none of each counites was zero. Moreover, the confidence level for most of the TSA counties is more than 95%. Thus, null hypothesis cannot be true. With the initial assumption, we accept the alternative hypothesis. "Hospitalization does have impact on fatality".

HYPOTHESIS 2

The statement given is:

"A higher ICU rate gives a higher fatality rate"

As per the statement when ICU rate increases this will in turn increase fatality rate.

Null and Alterative Hypothesis

The null and alternative hypothesis of the statement are as shown below:

Null Hypothesis	Alternative Hypothesis
Statement: ICU rate and fatality rate are not corelated.	Statement: ICU rate and fatality rate are related
Method: Fatality rate = ((slope)*(ICU rate)) + intercept If slope turns out to be zero, then the hypothesis will be accepted	Method: Fatality rate = ((slope)*(ICU rate)) + intercept If slope of the method is not zero, then the hypothesis will be accepted
Testing Method: Linear Regression t- test	Testing Method: Linear Regression t-test

Data

For the analysis, data from ICU rate (CombinedHospitalDataoverTimebyTSA.xlsx) and data from fatality rate (TexasCOVID19DailyCountyFatalityCountData.xlsx) are used. All the data modification was performed in Python and list was used to store the data.

Statistical Test

Linear regression t-test was performed for the data. To perform the test first the calculation mentioned below should be performed on the data.

Calculate ICU Rate

1. Collect two kinds of data from sheets “Beds Available” and “Beds Occupied”. Please do keep in mind that the data is from ICU.
2. ICU Rate = “Beds Occupied” / (“Beds Available” + “Beds Occupied”)

Calculate Fatality Rate

1. Collect the fatality cases by county from a sheet “Fatalities by County”
2. Sum all fatality cases up by TSA area. This is considered as fatality rate for the problem.

After the data is cleaned as per requirements, t-test was performed. Python did provide us with p-value up on passing x and y values for the requirements.

Result

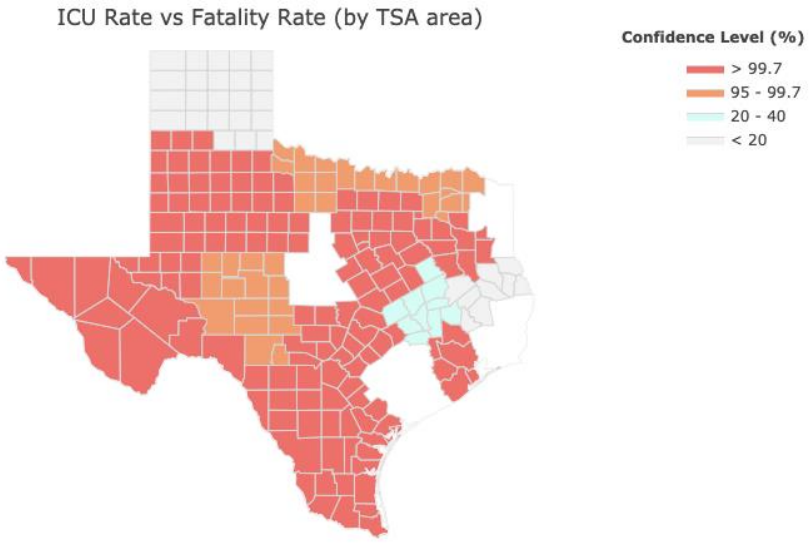

Figure 3 Texas ICU rate and fatality rate per TSA county map

First, we did check the slope values for all the TSA counites. Up on quick check (using code), it was found that none of the slopes were zero. Similarly, we can see the Confidence Levels in more than half of TSA areas are greater than 95% as shown in the figure above.

Conclusion

Linear regression t-test was performed on the data. Similar, to the above analysis none of the slope values were zero and the confidence levels for most of the counites were more than 95%.

Thus, we accept alternative hypothesis. “ICU rate and the hospitalization rate are corelated to each other.”

HYPOTHESIS 3

The statement is:

“As restriction lifted with open Texas the rate of infection increased, with a slight delay”

As per this statement when Texas lift the infection rate also increased.

Null and Alternative Hypothesis

The null and alternative hypothesis are as below:

Null Hypothesis	Alternative Hypothesis
Statement: Lift in restriction has no effect on the transmission rate	Statement: Lift in restriction has effect on the transmission rate
Method: Rate of transmission 1 (R1) = transmission rate in range (03/26 – 05/21) Rate of transmission 2 (R2) = transmission rate in range (05/21 – 07/25) If both rate of transmission are equal, then the statement is true.	Method: Rate of transmission 1 (R1) = transmission rate in range (03/26 – 05/21) Rate of transmission 2 (R2) = transmission rate in range (05/21 – 07/25) If both rate of transmission are not equal, then the statement is true.
Statistical Test: Chi-squared test; Paired t-test	Statistical Test: Chi-squared test; Paired t-test

Data

For the analysis, transmission rate is calculated. The calculation is explained in the next section. The data for transmission rate is obtained from

(TexasCOVID19DailyCountyCaseCountData.xlsx). The date from the calculations were given in the previous section.

Statistical Test

Chi-squared test was performed on the data. This test was selected since it is used for the analysis over the same comparison. Best example would be comparison of the data of a particular county over two period of times. Thus, making this test very suitable for our analysis.

Paired t-test, on the other hand also does the same time of analysis. The team was not sure, or we were not able to find difference between those two tests. Making us to apply both the tests on the data.

Calculate Transmission Rate

1. Calculate daily cases for each county from sheets “Total Tests Received”
2. Collect each county’s population from a file “2018_txpopest_county.csv”
3. Transmission Rate = daily cases / population * 1000000 / period

Result

The transmission rate for all the counties were calculated as explained in the previous section. Upon completion, the data sets were passed to the Chi-squared test. The confidence value was obtained. The map below shows the confidence level of all counties in Texas for chi-squared test.

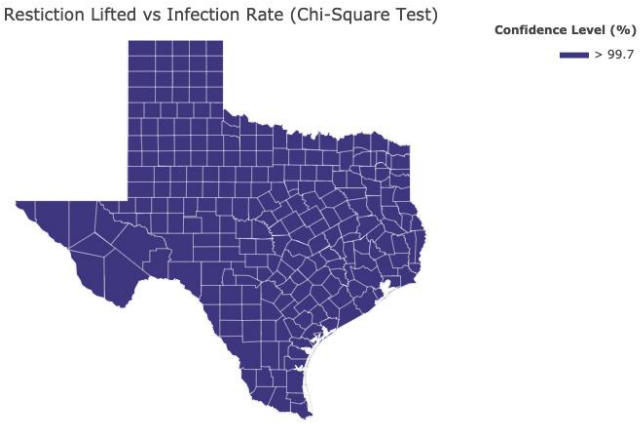

Figure 4 Chi-squared test performed on counties in Texas over before and after lifting the restrictions

The same data set were passed to paired t-test, and the confidence level is obtained. The map below shows the confidence level of all counties in Texas for paired t-test.

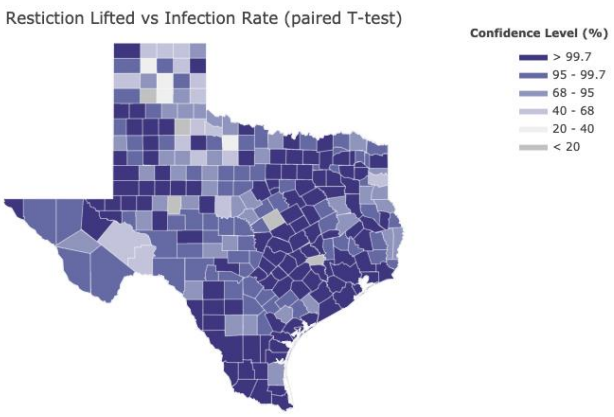

Figure 5 Paired T-test performed on all counites in Texas before and after lifting the restrictions

Conclusion

From the results obtained from above, we can clearly see the confidence level is more than 95% for most of the counites. Thus, making over alternative hypothesis true: "Lift in restriction has effect on the transmission rate"

HYPOTHESIS 4

The given statement is:

"More testing leads to lower transmission."

As per this statement, more testing or increase in testing has led to lower transmission.

Null and Alternative Hypothesis

The null and alternative hypothesis for the statement is as below:

Null Hypothesis	Alternative Hypothesis
Statement: Testing has no impact on transmission rate	Statement: Testing did lower transmission rate
Method: Transmission rate = ((slope)*(Testing)) + intercept If slope turns out to be zero, then the hypothesis will be accepted	Method: Transmission rate = ((slope)*(Testing rate)) + intercept If slope is not zero, then the hypothesis will be accepted
Statistical Method: Linear Regression t-test	Statistical Method: Linear Regression t-test

Data

To perform the statistical test, data from testing (TexasCOVID-19CumulativeTestsOverTimebyCounty.xlsx) was obtained. Moreover, data from transmission

rate (TexasCOVID19DailyCountyCaseCountData.xlsx) was obtained. The date range applied was 04/21 – 09/11.

Statistical Test

Linear regression t-test was performed on the data. For this analysis we are using all counties in Texas.

Calculate Transmission Rate:

1. Calculate daily cases for each county from sheets "Total Tests Received"
2. Collect each county's population from a file "2018_txpopest_county.csv"
3. Transmission Rate = daily cases / population * 1000000

Result

The data for the analysis is extracted into Python. The unrequired information is removed from the lists. Transmission rate is calculated for all the counites in Texas. The data of transmission rate and testing for all the counties is passed on to t-test. The confidence level is obtained. Below is the map showing confidence level for all counites in Texas.

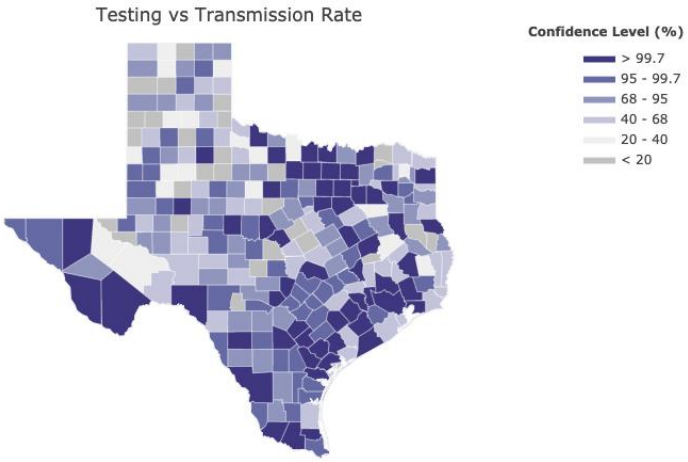

Figure 6 Texas map showing the confidence level for testing in comparison with transmission rate

Conclusion

Most counties in Texas have slope value more than zero. Moreover, the confidence level is more than 95% in most of the counties. Alternative hypothesis is selected. Thus, ": Testing did lower transmission rate"

HYPOTHESIS 5:

Given statement:

"Higher population density means higher transmission rate"

As per the given statement, the density of the population has effect on transmission rate. It states that both are directly proportional to each other.

Null and Alternative Hypothesis:

The null and alternative hypothesis for the given statement is:

Null Hypothesis	Alternative Hypothesis
Statement: Population density has no impact on transmission rate	Statement: Higher population density has higher transmission rate
Method: Transmission rate = ((slope)*(Population Density)) + intercept If slope is zero, then the hypothesis will be accepted	Method: Transmission rate = ((slope)*(Population Density)) + intercept If slope is not equal zero, then the hypothesis will be accepted
Statistical Method: Linear Regression t-test	Statistical Method: Linear Regression t-test

Data

The data required for this analysis are population density (Population Density by Counties in Texas.xlsx) and transmission rate (TexasCOVID19DailyCountyCaseCountData.xlsx). The calculation of transmission rate is explained in the next section. The date range used for the calculate is 05/21 – 07/25.

Statistical Test

Linear regression t-test was performed on the data.

Calculate Population Density

1. First the land area data was obtained per county
2. The population of the county was obtained

3. Ratio of the first two data would give us population density per county

Calculate Transmission Rate

1. Calculate daily cases for each county from sheets “Total Tests Received”
2. Collect each county’s population from a file “2018_txpopest_county.csv”
3. Transmission Rate = daily cases / population * 1000000 / period

Result

The data obtained from the sources is processed as mentioned in the above section. The map below shows the population density of all counites with in Texas. Similarly, a plot showing the transmission rate of the counites in Texas is also presented.

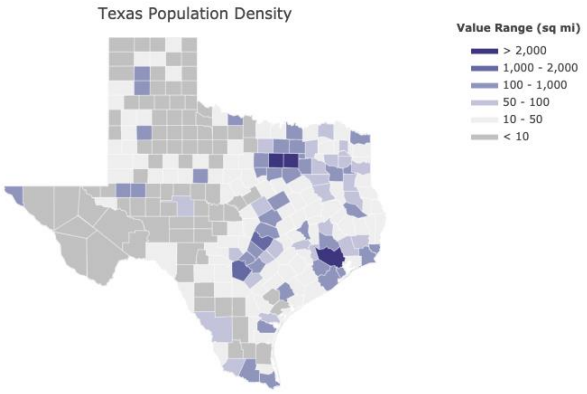

Figure 7 Texas map showing population density

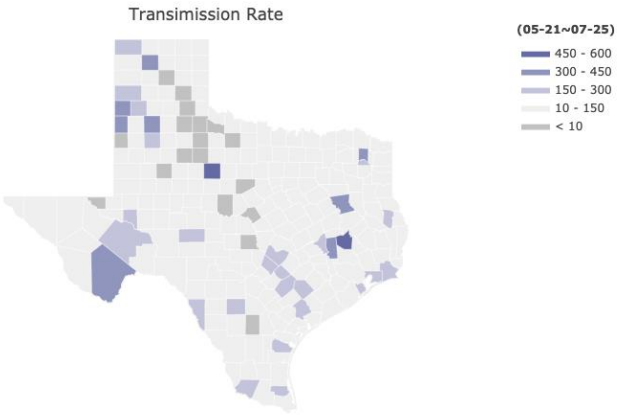

Figure 8 Transmission rate in all the counites in Texas

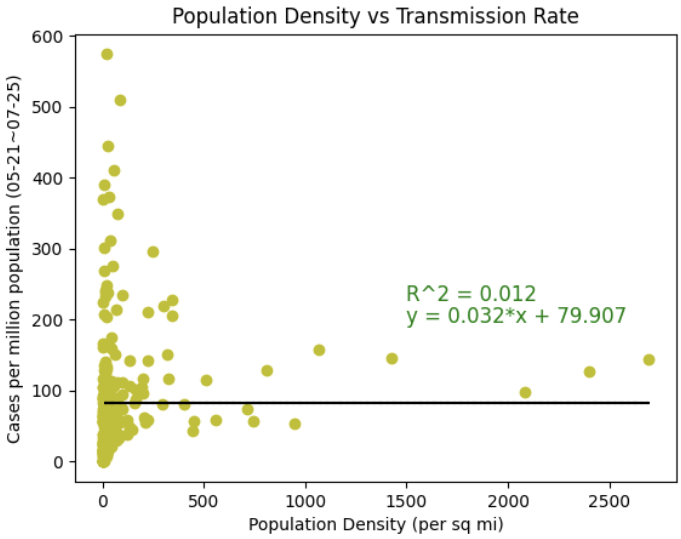

Figure 9 Linear relationship for all the counties in Texas

Conclusion

From the comparison in linearity of the tests in all counites none of the slope was zero. The overall confidence level was 91.49%. Thus, we are accepting null hypothesis: "Population density has no impact on transmission rate"

HYPOTHESIS 6

The given statement is:

"Rural counties neighboring urban areas had higher rates than rural counties not near urban areas"

As per the given statement, rural counties near urban areas have higher rates than they are not near urban areas.

Null and Alternative Hypothesis

First, a null and alternative hypothesis are generated. Null and alternative hypothesis are as shown below:

Null Hypothesis	Alternative Hypothesis
Statement: Rural counties near urban counties do not have higher rates than rural counties not near urban counties.	Statement: Rural counties near urban counties have higher rates than rural counties not near urban counties
Testing method: R1=Rate average rural counties neighboring urban R2=Rate average rural counties neighboring urban R1 - R2 <=0.	Testing method: R1 - R2 > 0 If difference of R1 and R2 is greater than 0 then statement is true.

If difference of R1 and R2 is less than equal to 0 then statement is true.	
Statistical Test: Linear Regression t-test	Statistical Test: Linear Regression t-test

Data:

The data needed for the analysis has been taken from urban counties and rural counties (PHR_MSA_County_masterlist.xlsx) are imported into Python. The calculation of rural rate and urban rate is explained in next section. The date ranges considered for this hypothesis is: 05/21 – 07/25.

Statistical Test:

t test was performed on the data.

Category 2 kinds of rural counties:

1. Collect adjacent counties for each county in Texas from a file "county_adjacency.txt"
2. If a rural county has one or more urban counties adjacent, it is categorized as "counties neighboring urban areas." On the other hand, if one rural county only has rural counties adjacent, it is categorized as "rural counties not near urban areas."

Infection Rate:

1. Calculate daily cases for each county from sheets "Total Tests Received"
2. Collect each county's population from a file "2018_txpopest_county.csv"
3. Infection Rate = daily cases / population * 1000000

RESULT:

Rural counties neighboring urban areas did not have higher rates than rural counties not near urban areas.

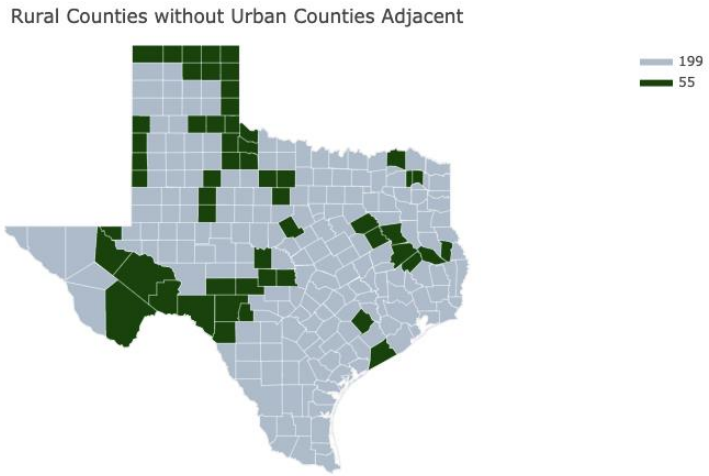

Figure 10 Texas map with rural counites having no urban county adjacent

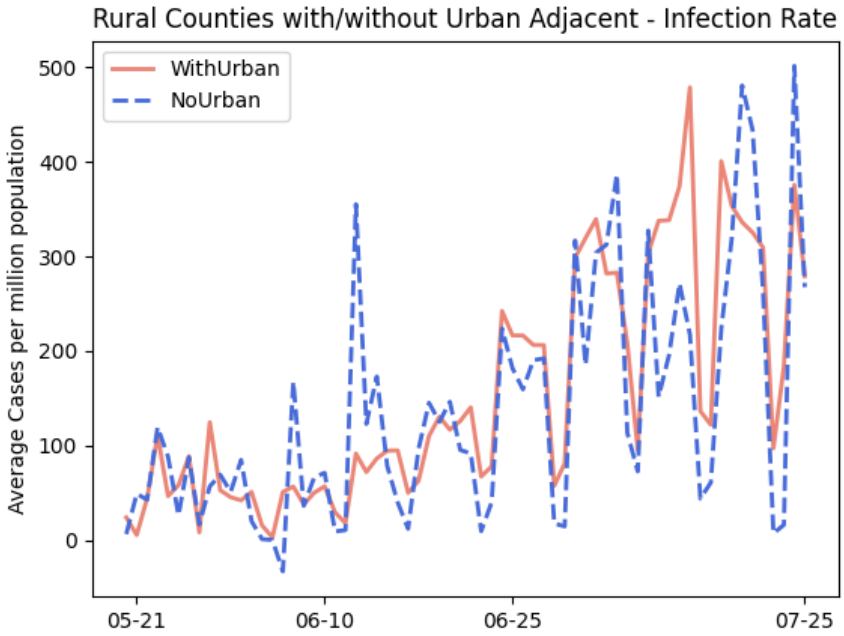

Figure 11 Rural counties with and with our urban adjacent per infection rate

Conclusion

From the above graph we can see that t test was performed on data and we can see that rural counties near urban counties did not have higher rate than not near urban counties. So that we can accept the null hypothesis.

HYPOTHESIS 7:

The given statement is:

"Travel between regions affected spread"

As per the statement, Travel between the regions affected in the increase of spreading the virus.

Null and Alternative Hypothesis

First, a null and alternative hypothesis are generated. Null and alternative hypothesis are as shown below:

Null Hypothesis	Alternative Hypothesis
Statement: Travel between the region do not affected spread.	Statement: Travel between the region affected spread.
Testing method: μ: infection rate in a period	Testing method: At least one of the travel data is different from others

$\mu_{Residential} = \mu_{Grocery\&Pharmary} = \mu_{Transit} = \mu_{Rereatioin} = \mu_{Work} = \mu_{Park}$	
Statistical Test: ANOVA	Statistical Test: ANOVA

Data

Data needed for the hypothesis is collected from mobility data (https://www.google.com/covid19/mobility/) and are imported into Python. The calculation of infection rate is explained further. The date ranges considered for this hypothesis is: 05/01 – 05/30.

Statistical Test

ANOVA was performed on the data.

Calculate Infection Rate:

1. Calculate daily cases for each county from sheets "Total Tests Received"
2. Collect each county's population from a file "2018_txpopest_county.csv"
3. Infection Rate = daily cases / population * 1000000

Result

Travel between the region affected spread.

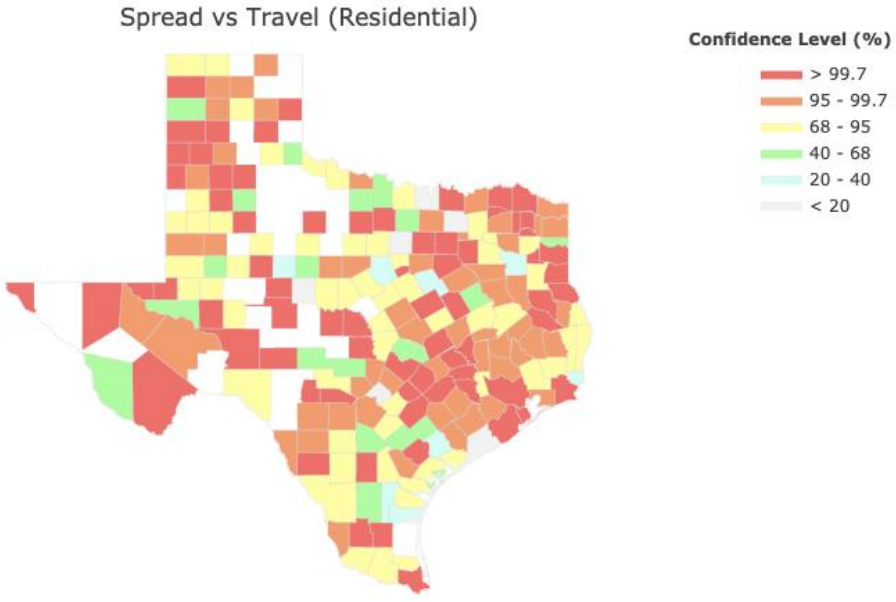

Figure 12 Texas map showing spread of virus corresponding to travel wrt residential

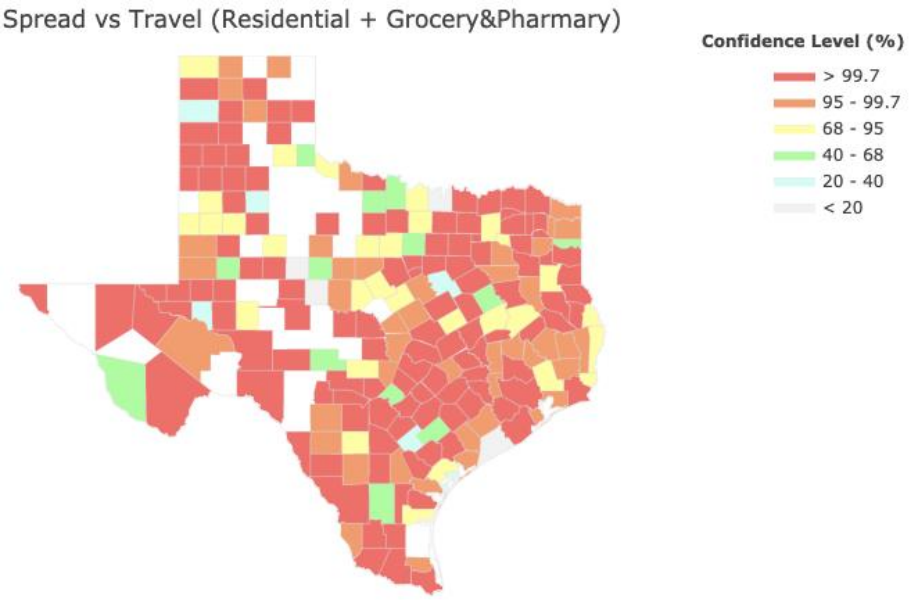

Figure 13 Texas map showing spread of virus corresponding to travel wrt residential and grocery

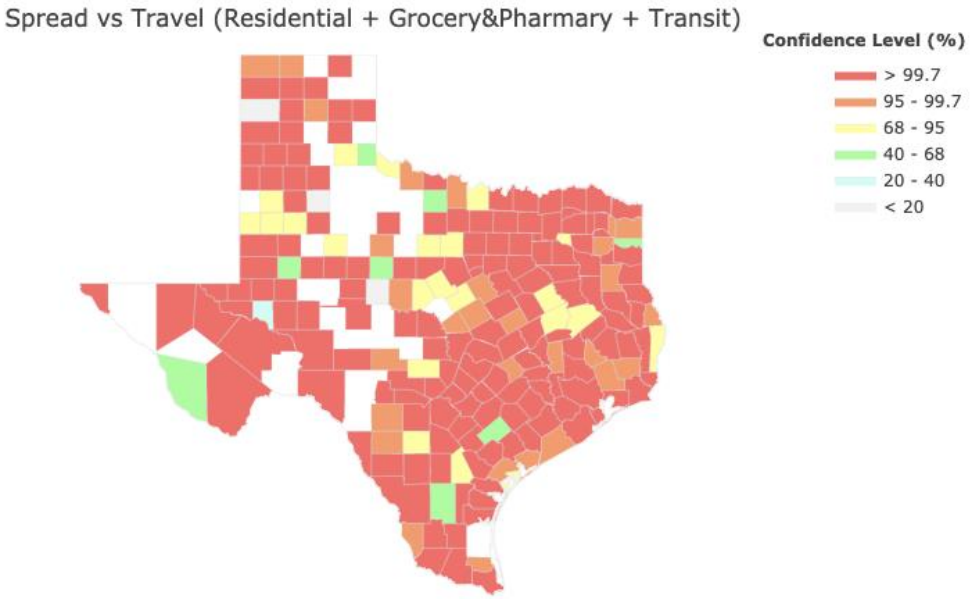

Figure 14 Texas map showing spread of virus corresponding to travel wrt residential, grocery and transit

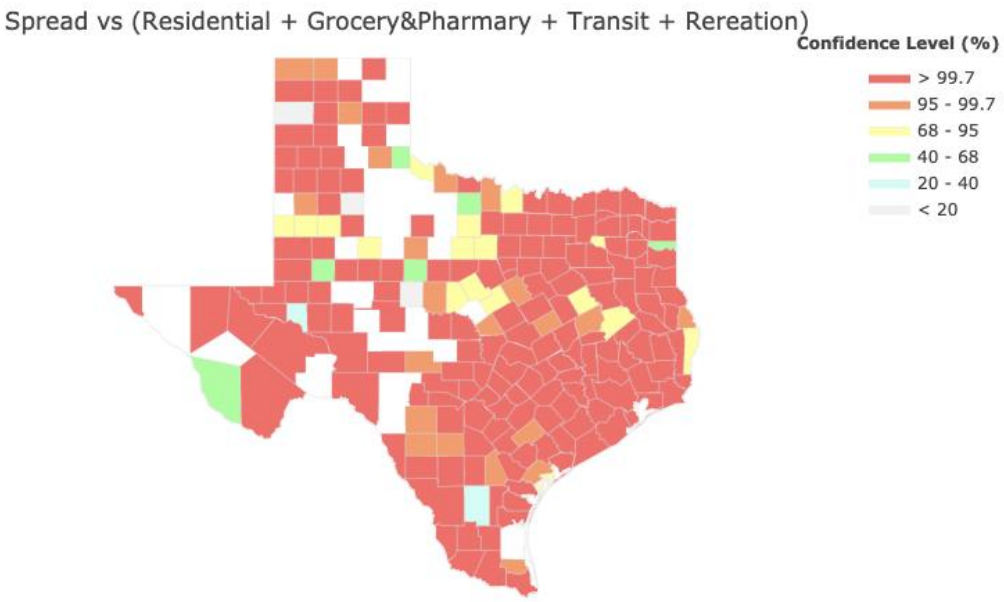

Figure 15 Texas map showing spread of virus corresponding to travel wrt residential, grocery, transit and recreation

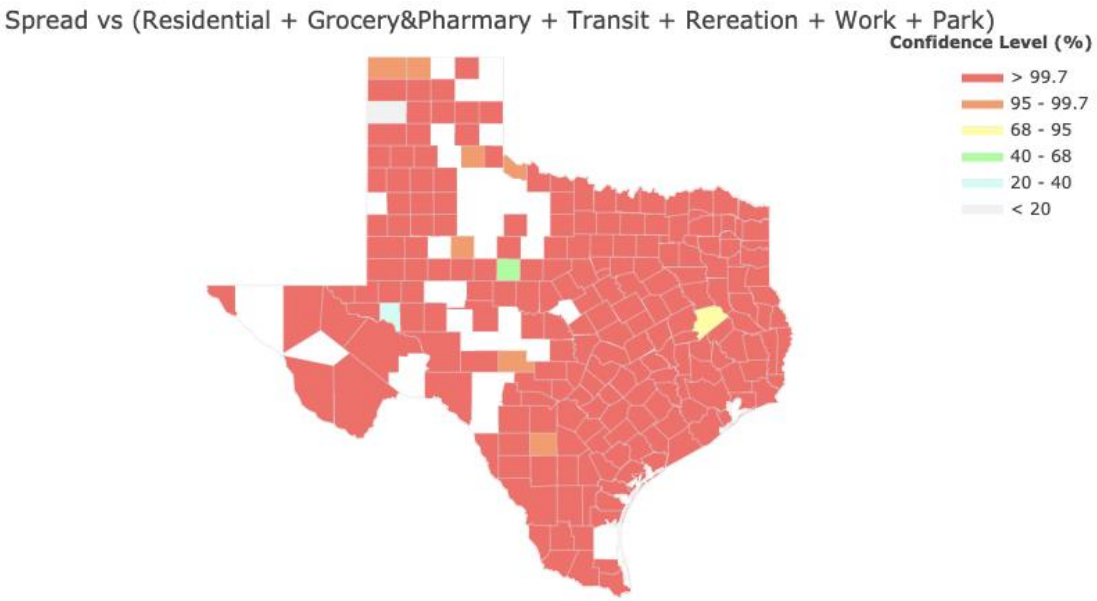

Figure 16 Texas map showing spread of virus corresponding to travel wrt residential, grocery, transit, recreation, work and park

Conclusion

From the above, Anova test that was performed on each county we can see that the confidence level for most of the counties is more than 95%. Thus, null hypothesis cannot be true.

HYPOTHESIS 8:

The given statement is:

"Rural counties had less spread (low population, as opposed to population density)"

As per the statement, rural counties with low population will have less spread as opposed to population density.

Null and Alternative Hypothesis

First, a null and alternative hypothesis are generated. Null and alternative hypothesis are as shown below:

Null Hypothesis	Alternative Hypothesis
Statement: Rural counties did not have less spread.	Statement: Rural counties they have less spread.
Testing method: R1=Rate average rural counties R2=Rate average urban counties R1 - R2 >= 0 If difference of R1 and R2 greater than equal to 0 then statement is true.	Testing method: R1 - R2 < 0 If difference of R1 and R2 less than 0 then statement is true.
Statistical Test: t-test	Statistical Test: t-test

Data:

Once hypothesis was generated, data required are collected from rural counties and population density (Population Density by Counties in Texas.xlsx) are imported into Python. The date ranges considered for this hypothesis is: 05/21 – 07/25.

Statistical Test:

t test was performed on data.

Category 2 kinds of rural counties:

1. collect information from "PHR_MSA_County_masterlist.xlsx"
2. If one county is Metro, we take it as an urban county. On the other hand, we take others as a rural county.

Infection Rate:

1. Calculate daily cases for each county from sheets "Total Tests Received"
2. Collect each county's population from a file "2018_txpopest_county.csv"
3. Infection Rate = daily cases / population * 1000000

RESULT:

Rural counties did not have less spread

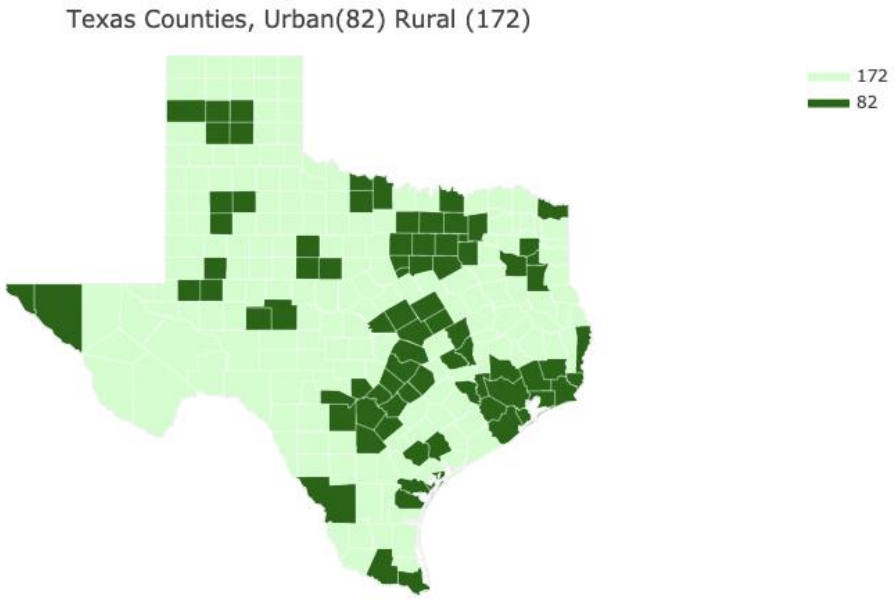

Figure 17 Texas map with rural counites having with urban county

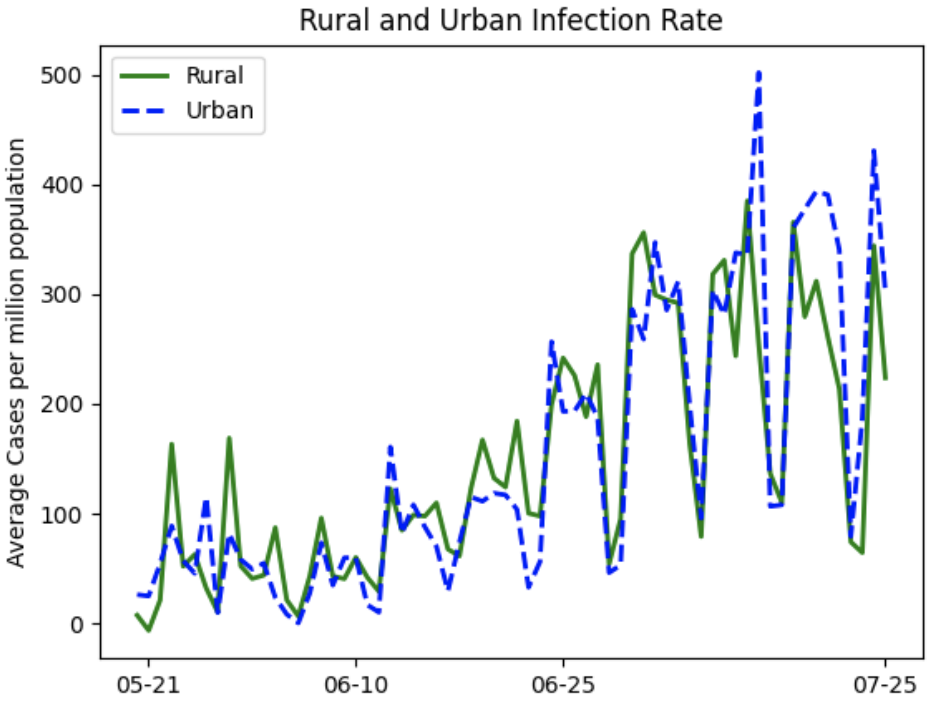

Figure 18 Rural counties with and with our urban adjacent per million population

Conclusion

Above graph, shows that t test was performed on data and we can see that rural counties as less spread as compared with urban counties on per million population. So that we can accept the null hypothesis.

CONCLUSION

As per schedule, the team was able to perform statistical tests on the data. Null and alternative hypothesis were generated. Statistical test was selected depending upon the requirement. The data was cleaned as per date range for the statement. Results from the statistical test was used to pick either null or alternative hypothesis.

Overall, we as a team were able to learn how to apply various types of statistical testing. Moreover, we were able to learn clean the data per requirement and plot them.

REFERENCES

1. **Covid 19 Data:** https://dshs.texas.gov/coronavirus/
2. **Mobility Data:** https://www.google.com/covid19/mobility/
3. **Transmission Rate:** https://www.cdc.gov/coronavirus/2019-ncov/downloads/global-covid-19/SARS-CoV-2-Transmission-Metrics.pdf
4. **T test:** https://towardsdatascience.com/inferential-statistics-series-t-test-using-numpy-2718f8f9bf2f
5. **Z test:** https://www.analyticsvidhya.com/blog/2020/06/statistics-analytics-hypothesis-testing-z-test-t-test/
6. **Chi Square Test:** https://www.tutorialspoint.com/python_data_science/python_chi_square_test.htm