Name: Hari Kishor Chintada

Title: A7 Project Report

Assigned County: Allegheny County from Pennsylvania

A picture containing text

Description automatically generated

**A7: Project Report**

Contents

[1. Introduction 3](#_Toc90395598)

[2. Background/Related Work 3](#_Toc90395599)

[3. Methodology 4](#_Toc90395600)

[4. Findings 8](#_Toc90395601)

[5. Discussion/Implications 8](#_Toc90395602)

[6. Limitations 9](#_Toc90395603)

[7. Conclusion 9](#_Toc90395604)

[8. References 10](#_Toc90395605)

[9. Data Sources 10](#_Toc90395606)

# Introduction

Can we get an accurate total number of Covid 19 infected people? I think it is impossible as we can only get infection status for those who are tested. Apart from false positives, if the test result is negative today that doesn’t mean the person is not infected. Depending on the individual immune system it may take a few days to show symptoms, so test results depend on the test window. So testing IS very crucial to understanding the spread of the pandemic.

Using Covid19 confirmed cases data we can observe cases are increased between Nov 2020 & March2021 and at peak around Dec 2020 even though mask mandate is in place. We have many articles and debates about masking policy, some support mandatory, and some questions about it. Does masking really reduce infections? Based on the Pennsylvania Allegheny county survey result 76% people followed the masking policy, still why did we observe an increase in confirmed cases? Most of us might have the same question, So I would like to check if it is related to the number of tests? Is confirmed cases increased due to increase in the number of tests? Why are more people tested during this period?

This problem is human centered as the data is related to people Covid19 PCR tests & results. This analysis helps humans to understand the pandemic and take necessary actions. For example, if a surge in testing ultimately results in more positive cases, then people can take more precautions a few days earlier than after test results are published. Normally it takes a few days to get test results.

This analysis will help if anyone has any questions about whether more tests resulted covid cases increasing.

# Background/Related Work

I would like to understand how Covid-19 is progressing and progress of community transmission so that everyone can take appropriate actions to suppress spread. Number of confirmed cases tells the development of a pandemic, but the confirmed cases are based on tests. In this exercise, I am trying to explore the relation between tests and confirmed cases and what caused the surge in tests?

Hypothesis: Covid confirmed cases increased due to more PCR Tests

Research question:

* Why are more people tested at some duration?  Is there any relation between test volume and infection rate, does high confirmed cases are due to more tests?
* Is there any relation between death rate, test rate and infection rate?

# Methodology

Since PCR Tests and confirmed cases, dataset is time series data, we can plot visuals and use correlation to identify the relation between these elements.

* Time Series Analysis: Since data is by day and we have two time series variables I am planning to use time series plots to identify the patterns. These visualizations can help to understand the patterns more effectively than raw data.

o   Plot number of tests/test coverage rate and confirmed case/infection rate as time series

o   Observe the visuals and identify trends/patterns

* + Interpret the result
* Correlation: Since we have two continuous variables, I choose to calculate correlation to quantify the associations between two variables using following steps:

o   finds the association between number of tests/test coverage rate and confirmed case/infection rate

o   Calculate correlation coefficient

* + Interpret the result
* Regression: Use regression to understand the association between two variables

o   Use Simple Linear Regression

o   independent variable Number of tests/Test coverage rate and dependent variable confirmed cases/infection rate

o   Interpret the result

**Implementation:** Let’s look at types of correlation to understand the relationship between variables.

* **Positive Correlation**: Two variables are positively correlated if one variable increases the other one increases and if one variable decreases the other one decreases. Example: Cake size and Price, Circle radius and area.
* **Negative Correlation**: Two variables are negatively correlated if one variable increases the other one increases and if one variable decreases the other one decreases. Example: Planets average temperature and distance from the Sun.
* **No Correlation:** Mo correlation means there is no relationship between variables. Example: Animal size and lifespan.

 We can easily understand the correlation using the below plots.

Chart, scatter chart

Description automatically generated

As a first step, we can plot Covid PCR Tests and Confirmed cases and visualize the relation to find the correlation between these two variables.

Chart, histogram

Description automatically generated

Above plot shows confirmed cases are increased with increase in PCR Tests and decreased when tests are low except for few intervals. Now let’s check if we can prove the correlation between these two variables using statistical methods. We can use correlation coefficient to determine the relationship, it is a statistical measure of the strength of the relationship between the relative movements of two variables. The values range between -1.0 and 1.0.

I am using TWO methods, Pearson correlation and Spearman correlation to calculate correlation coefficient. In general Pearson correlation evaluates linear relationship between two variables and Spearman correlation evaluates the monotonic relationship.

Pearson correlation coefficient is calculated using the following formula, though we use standard python library to get value directly.

Text

Description automatically generated

Spearman correlation coefficient is calculated using the following formula.

Graphical user interface, text, application

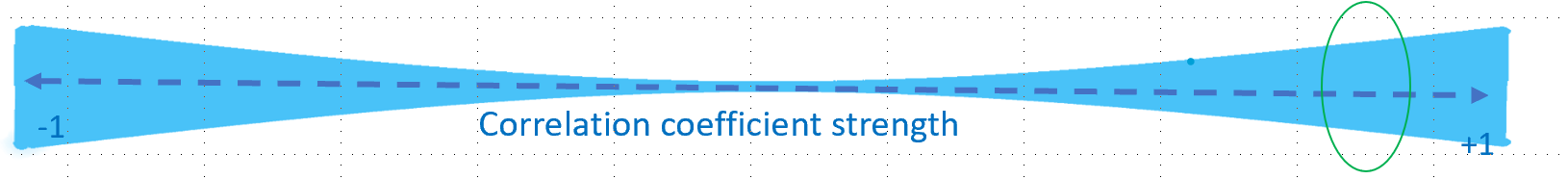
Description automatically generated

Now let’s check the result using our data:

Chart, scatter chart

Description automatically generated

|  |  |  |
| --- | --- | --- |
|  | Pearson correlation | Spearman correlation |
| Correlation coefficient | 0.780 | 0.844 |
| p-Value | 0.000 | 0.000 |



How can we interpret the results?

Both Pearson & Spearman methods returned coefficient value ~0.8. So, there is a strong positive correlation between PCR tests and confirmed cases.  Also, p-values are ~0.0, that is the probability that the correlation between PCR Tests and confirmed cases in the sample data occurred by chance, which is very low. So, we can conclude that correlation between these two variables is statistically significant.

Chart, line chart

Description automatically generated

Chart, scatter chart

Description automatically generated

Above results show p-value is significantly lower when we use covid confirmed cases rate than the previous result when we use confirmed case numbers.

# Findings

Let’s look at what observations we can find using our analysis results. I am breaking it into TWO parts, relation between PCR Tests and confirmed cases and Confirmed cases per 1000 PCR Tests. Time series and correlation coefficient shows confirmed cases are increased when more PCR tests are performed. But when we check confirmed case rate, i.e., confirmed cases per 1000 PCR tests, we see they both are not strongly correlated.

So, correlation between volume is not always the right way to find relation between two variables. We must understand the data elements first and use the right data for analysis.

# Discussion/Implications

Why are your findings important or interesting; How could  future research build on this study?

This section should include a thoughtful reflection that describes the specific ways that human centered data science principles informed your decision-making in this project.

By looking at PCR Test and confirmed cases trends it is obvious that we see more cases when the tests are more and less cases when tests are down. It is easy to convince people who don't have any statistical knowledge. But is it the right way to perform this analysis? This is the reason we need an appropriate review process to validate data analysis methods and result interpretation.

# Limitations

In this section I am listing some of the assumptions and few factors which are outside of our controls.

* Covid19 symptoms may appear at different duration depends on individual immunization
* Multiple sources - As data is collected from multiple different sources the quality might not be the same across
* Reporting date inconsistency – There might be delay in data reporting, some teams may work on only weekdays so weekend data might be reported on next week day. If we consider test result, does all labs submitted data on same basis, depends on staff availability, some might be submitted at the end of the day and others may be afternoon.
* Population – We have yearly population, so we are considering same population for the whole year
* Test/Confirmed case count – Do everyone followed same rules? Do they counted number of tests performed or number of people tested? Even if requester provided clear guidelines to report each person only once, there is practical problems like It is hard to get this info, Same person may go to different facilities and reported by each If there is no centralized real time system to check if the person is already reported.
* Test incorrect results – Similar to any other results, Covid-19 test results may also contain some wrong results
* Manual errors – There could be some mistakes while entering data by humans
* Date inconsistency – Date range is not the same across all data sources, so for some period we have some unknown data elements.
* Not considering other external factors which might influence surge in test, like test drives by government or mandatory tests using contact tracing.

# Conclusion

Our Hypothesis “Cases are up due to more testing” is wrong.  Though we observe an increase in confirmed cases volume when we perform more tests, the rate is not increased in the same proportionate. You may wonder why we see increase/decrease in confirmed cases? There are external factors which might impact increase/decrease in covid confirmed cases. For example, new variants, vaccination/herd immunity, mask policy, social distancing, work from home, travel restrictions and lock downs might influence covid virus exposure and spread.

Interpretation of data analysis results and summarizing it in simple text is needed so that all humans can understand. It is important to review data analysis and results, good if we have an expert review committee to ensure the method used for analysis is appropriate one and data is not manipulated (ex: take partial data) to get desired outcome. Also sharing the code and data can help others to reproduce or replicate the analysis as we gather new data points.

# References

* [Trump said Covid-19 testing 'creates more cases.' We did the math - STAT (statnews.com)](https://www.statnews.com/2020/07/20/trump-said-more-covid19-testing-creates-more-cases-we-did-the-math/)
* <https://bousteadtravel.com/covid-19-screening/>
* <https://timesofindia.indiatimes.com/travel/travel-news/latvia-imposes-nearly-monthlong-lockdown-as-covid-19-cases-surge/as87252044.cms>
* <https://www.bbc.com/news/world-52103747>

# Data Sources

In this analysis I used covid confirmed cases, mask mandate information, PCR test & covid deaths data. Describing the data source and schema here.

* Covid confirmed cases: Aggregate COVID-19 case counts and rates by date of report. Counts include both confirmed and probable cases, and a person is counted as a case only once.
  + Data Source: [COVID-19 Aggregate Cases Current Daily County Health | PA Open Data Portal](https://data.pa.gov/Covid-19/COVID-19-Aggregate-Cases-Current-Daily-County-Heal/j72v-r42c)
  + License: [Public Domain U.S. Government](https://www.usa.gov/government-works)
  + Attributes:
    - Jurisdiction: Pennsylvania County of residence
    - Date: Case reported date
    - New Cases: Number of new confirmed and probable cases first reported to the Department of Health on that date
    - 7-day Average New Cases: Rolling 7-day average of new confirmed and probable cases
    - Cumulative cases: Cumulative confirmed and probable cases reported through that date
    - Population (2019): Most current population data available
    - New Case Rate: Number of new confirmed and probable cases reported that date per 100,000 population
* Covid test volume: Daily aggregates of COVID-19 PCR test results reported to the Department of Health. It includes tests with positive, negative, and inconclusive results. If the same test result (identified by having the same patient, collection date, and result) was reported more than once (either accidentally or reported by both the ordering facility and the performing lab), it is counted only once in this data. Only tests performed on Pennsylvania residents are included.
  + Data Source: [COVID-19 PCR Test Counts March 2020 - Current Statewide Health | PA Open Data Portal](https://data.pa.gov/Covid-19/COVID-19-PCR-Test-Counts-March-2020-Current-Statew/r6ti-va88/data)
  + License: Public Domain U.S. Government
  + Attributes:
    - Date: The date the test result was reported to department of Health
    - New PCR Tests: Number of deduplicated PCR test results for Pennsylvania residents reported to the department of Health
* Covid death volume:
  + Data Source: [COVID-19 Aggregate Death Data Current Daily County Health | PA Open Data Portal](https://data.pa.gov/Covid-19/COVID-19-Aggregate-Death-Data-Current-Daily-County/fbgu-sqgp/data)
  + License: [Public Domain U.S. Government](https://www.usa.gov/government-works)
  + Attributes: aggregate death data. Data set has additional columns like 7 day average, cumulative numbers, rate. I am not using those.
    - County Name:
    - Date of Death:
    - New Deaths:
    - 2019 Population:
    - Date: Vaccination Date as reported to PA-SIIS, the Pennsylvania Statewide Immunization Information System
    - County Name: Pennsylvania county name
    - Partially Covered: counts of individuals who received a vaccine dose that provides partial coverage against the disease (Dose-1 for some providers)
    - Fully Covered: counts of individuals who received a vaccine dose that provides full coverage against the disease (Dose-2 for some providers)
* Masking policy information- [masking mandates by county](https://data.cdc.gov/Policy-Surveillance/U-S-State-and-Territorial-Public-Mask-Mandates-Fro/62d6-pm5i)
  + Citation: Data can be cited as: CDC, COVID-19 Community Intervention & Critical Populations Task Force, Monitoring & Evaluation Team, Mitigation Policy Analysis Unit, the CDC, Center for State, Tribal, Local, and Territorial Support, Public Health Law Program, and Max Gakh, Assistant Professor, School of Public Health, University of Nevada, Las Vegas, “U.S. State and Territorial Orders Requiring Masks in Public,” (August 15, 2021).
* Mask compliance survey - [mask compliance survey](https://github.com/nytimes/covid-19-data/tree/master/mask-use)
  + Citation: “The New York Times and Dynata”

There are no ethical considerations for the data set I am using, data is aggregated by day, and it doesn’t include any individual or demographic (like Gender, Ethnicity, Age group) information. In fact, I explored various sources to get daily aggregation by demographic so that I can slice the data and perform in depth analysis. But couldn’t find any such data set, CDC only published aggregated snapshot data than daily volume.