# 

# 

# 

# **Group 3 Project Report**

**COVID-19 IMPACT ON SCHOOL EDUCATION**

**MIS 6380.5D1**

**Prof. Naser Islam**

**Ashwin Manjunath Bharadwaj, Parth J Kakadiya, Kavya Prakash Hathwar, Apoorva Rani, Saketh Sadu**

**Table of Contents**

|  |  |
| --- | --- |
| **Content** | **Page Number** |
| ***Executive Summary*** | 2 |
| ***Data Description*** | 3 |
| ***Data Cleaning*** | 4 |
| ***General Introduction*** | 6 |
| ***Insights and findings*** | 7 |
| *Hypothesis 1* | 7 |
| *Hypothesis 2* | 8 |
| *Hypothesis 3* | 9 |
| *Hypothesis 4* | 10 |
| *Hypothesis 5* | 11 |
| ***Conclusion*** | 12 |
| ***References*** | 13 |

**Executive Summary**

The COVID-19 pandemic has forced human resources departments to ask new questions. In these challenging times, analytics have been critical to helping us make decisions confidently and quickly. Covid has challenged various day to day activities among which school education is one of them. The purpose of the project is to utilize factors associated with student attendance in the state of Texas and to bring forward insightful information on how covid impacted school education trends. This topic was targeted specifically due to the abnormalities in the current academic school year.

The COVID-19 pandemic manifests itself in various ways around the world in terms of cases, hospitalizations, and fatalities, wreaking havoc on the global economy, social, sociological, health, and most crucially, educational sectors. The goal of this project is to discover factors associated with student attendance in Texas and to prevent COVID -19 in schools by implementing methods such as immunizations for teachers, staff and eligible children, the use of masks and physical separation, and screening testing.

**Tools Utilized**

Tableau was chosen as the visualization tool by the team since it has been the market leader in data visualization for the past six years. Since it runs on Big Query, the tool is extremely quick and can handle a large amount of data. Tableau automatically detects the device you're using to view the report and adjusts the dashboards accordingly without the developer having to do additional steps.

**Insights and Findings**

1. There is no direct correlation between the absent percentage increase in schools of the county with that of the covid increase percentage in the respective county.

2. The absentees' frequency in rural area schools was higher when compared to schools in urban and suburban areas as the rate of covid case increase was higher for rural areas.

3. Non-Metro areas had a higher average absent percentage and lesser average attendance percentage compared to Metro areas.

4. There is no correlation between average present percentage and total covid cases in schools.

5. Covid-19 affected Elementary school children more when compared to High School and Middle school children because of less immunity.

**Data Description**

Our team retrieved the initial/original datasets namely Grade Level Dataset, School Level Dataset, and District Level Dataset from Texas School (Freedom of Information Act); U.S. Census on Kaggle.

**Daily Student Attendance**

The vast majority of the information used in this project was gathered from Texas school districts, public census data, and public COVID 19 data. An email was sent to 40 school districts around the state of Texas requesting student attendance data (FOIA). 19 of those districts provided the needed data in their response. Different kinds of data were gathered because of the original message issued to districts being ambiguous. The "daily" records of student attendance and a "summary" of student attendance records for the first few weeks of this academic school year were the main differences between the data obtained. This data is with respect to 11 schools.

Variables used from school district datasets included dates, weekdays, school name, school type, district, absent percentage increase, covid percentage increase and grade level.

**School Information, County Description, Metropolitan vs. Non-Metropolitan**

The dataset "Current Schools" provided by the Texas Education Agency, has the following variables: school name, school zip, district number, school type, and school zip. The second dataset, "District Type," included characteristics of each school district, such as whether it was a big urban region, an independent town, or a rural location. A district's district number, description provided by the Texas Education Agency (TEA), and National Center for Education Statistics were among the characteristics from the "District Type" dataset that were chosen for usage. Selected variables from this dataset include county name and metro areas.

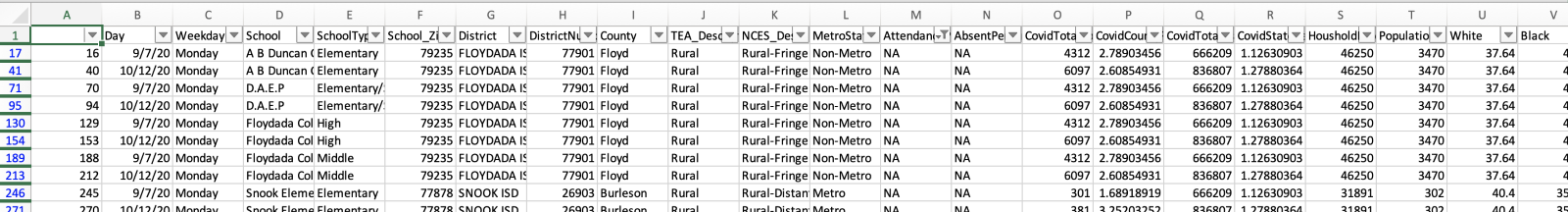
**COVID-19 and Demographics Data**

Past academic school years have seen noticeable differences in student attendance, therefore live COVID-19 data was obtained from the New York Times to investigate any connections. Data from this dataset are accessible in three forms and are updated daily (country, state, and county). Both COVID-19 cases by state and COVID-19 cases by county were chosen as variables from this dataset.

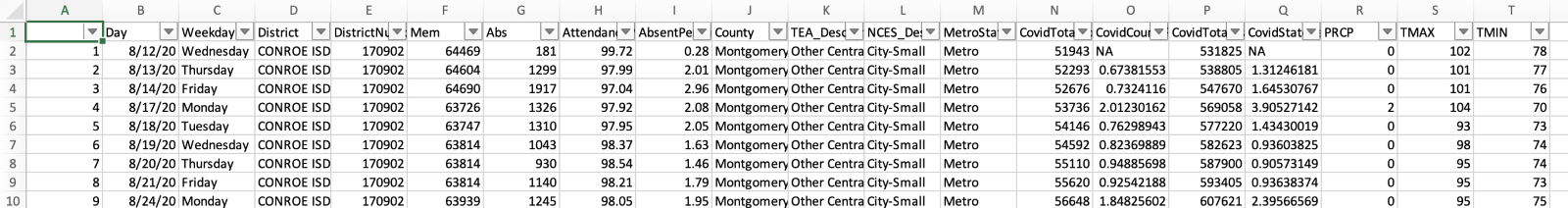
## 

**Data Cleaning**

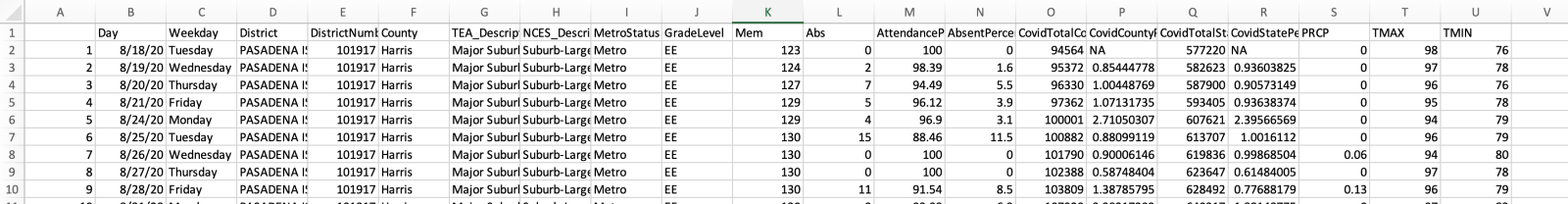
In this project we used three different files, School level data, Grade level data and District level data to prove our hypothesis. After reviewing the data files our team started working on data cleansing by removing duplicate columns, checking for missing values, incorrect data, and unwanted columns. From the School level data table, NA values were removed from the files where that data was not dependent on other files. All files were joined at the county level.



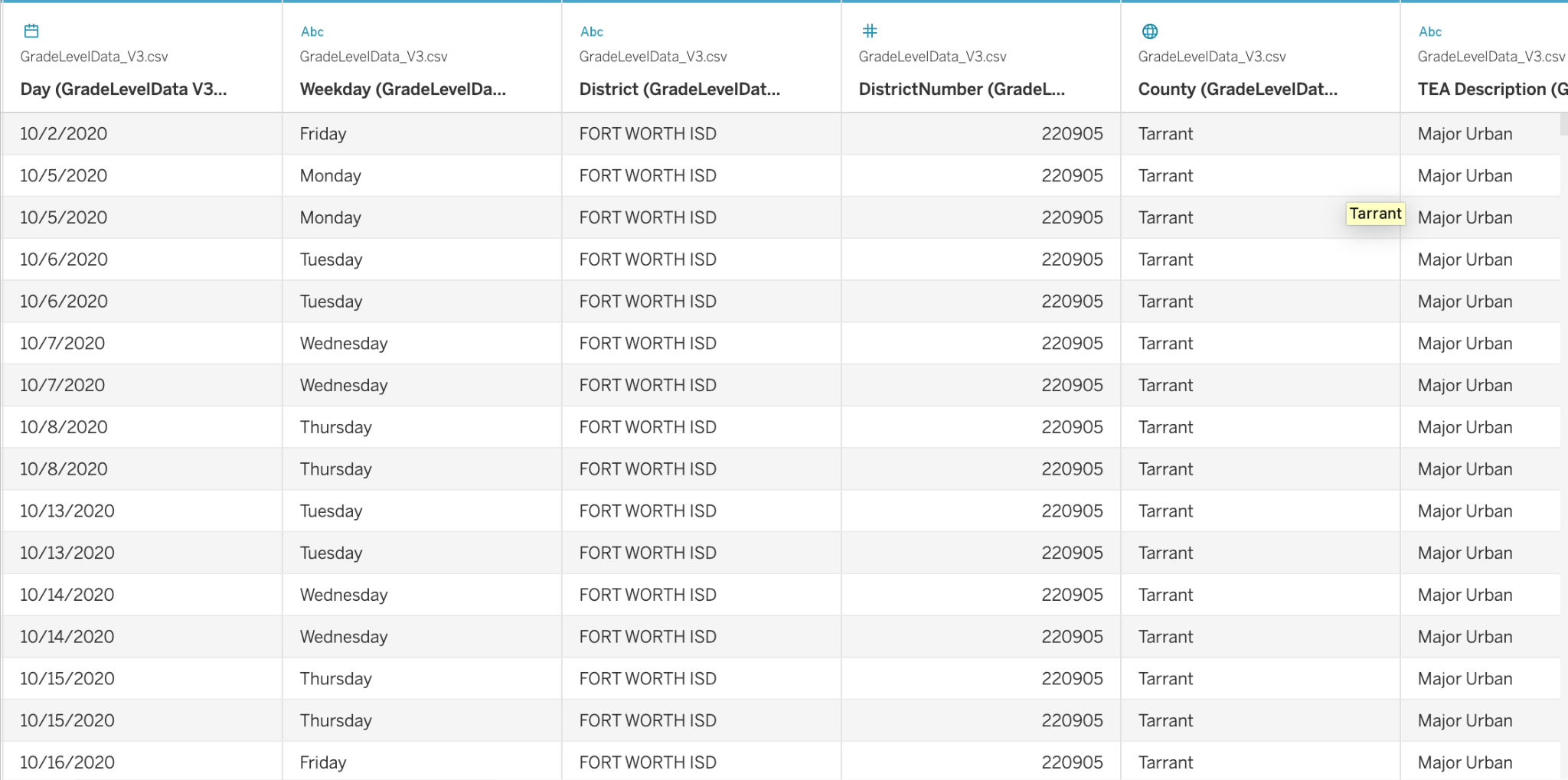
From district level data, we removed some null values from rows where there is no data other than the school's name or county name which helps us to prove our hypothesis with clear results. In a few columns we added average in attendance and absences where the value was null, or the cell was empty.



From Grade level data, we filtered out the column where we have two values showing city and rural or rural and sub rural together. It was filtered only to keep city, rural, major urban and suburban areas which helped us divide the school based on the area in which area of county schools are located. Null values were replaced with the average of the respective column to provide better vision in the dashboard.



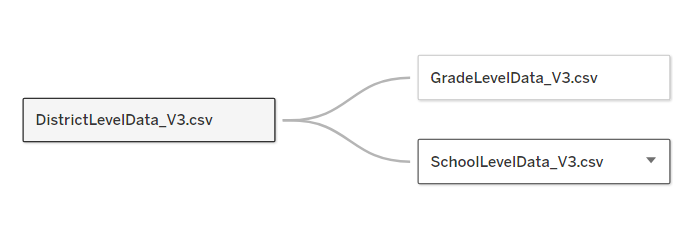
Following the addition of all data files to the tableau data source, we discovered flaws in a few files that did not pertain to our hypothesis and also lacked any connections to other data sources, so we eliminated those files from the table. We categorized the city, metro, district, and county based on our data. In addition, we had few columns which we did not need for our project but we kept these columns as it had relations with other tables while joining these three files.



**General Introduction**

This project's goal is to use variables related to student attendance in the state of Texas and to present meaningful data on how covid affected attendance trends in schooling. Due to the anomalies in the current school year, this topic was chosen specifically. To achieve this, our team chose datasets specific to Daily student attendance, School Information, County Description, Metropolitan vs. Non-Metropolitan areas data at Grade, State and District Levels.

Following data cleaning, the data model for the datasets imported into Tableau is displayed below:



The following hypotheses were analyzed with the three datasets.

Hypothesis 1: There could be a direct correlation between the CovidCountyPercentIncrease and the AbsentPercent of the schools in that county.

Hypothesis 2: The possibility of having lower frequency of absentees in rural area schools is higher when compared to urban and suburban areas.

Hypothesis 3: Chances of average absentee percentage in non-metro areas is higher than that of metro areas since metro areas took better precautions.

Hypothesis 4: The Odds of having County’s with higher covid cases have lower present percentage in school from August 2020 to November 2020.

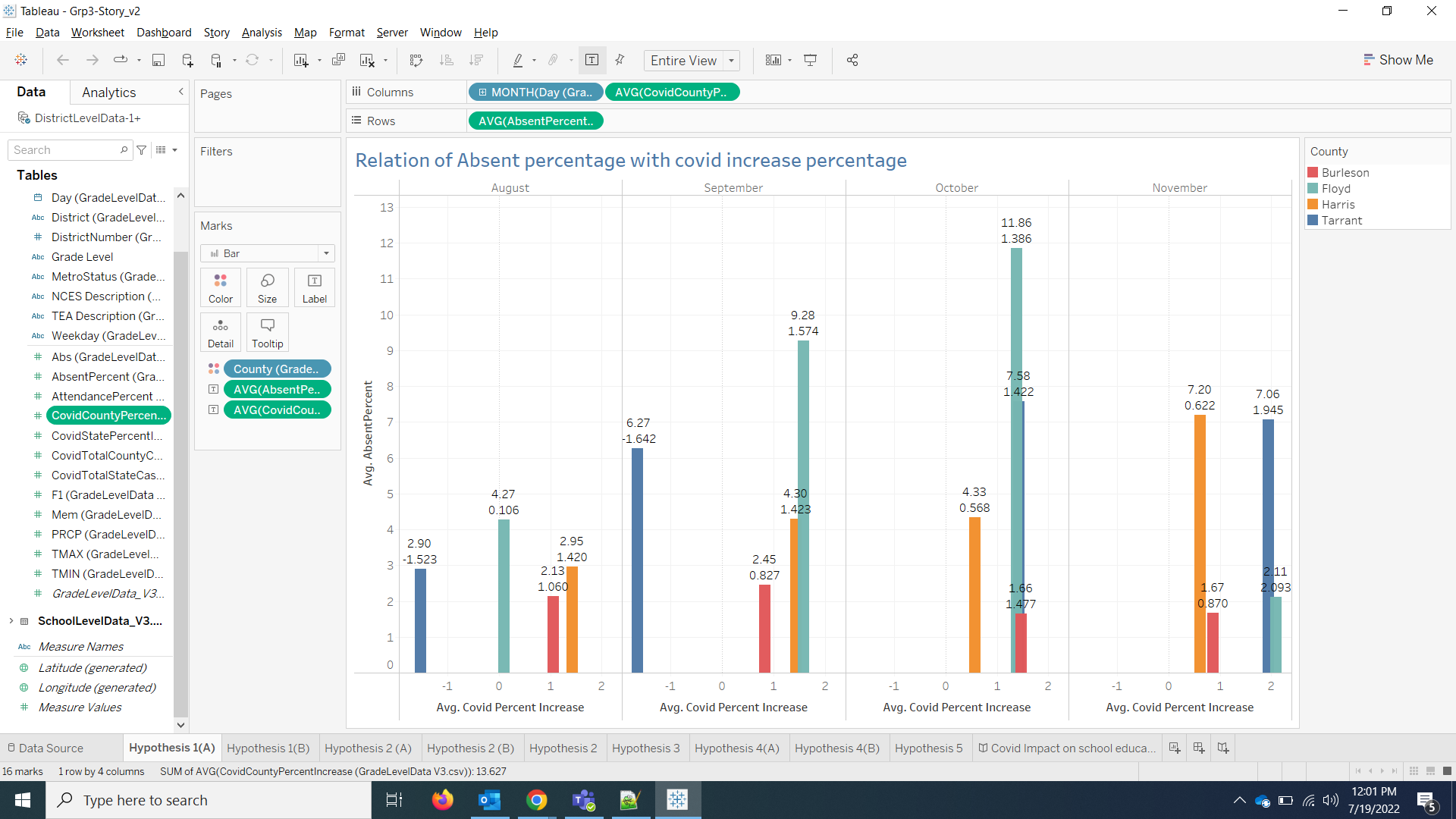
Hypothesis 5: Covid-19 may have affected Elementary school children more when compared to High School and Middle school because of less immunity.

**Insights and Findings**

**Hypothesis 1**

It would be wise to think that, the increase in covid increase percentage in the county would affect the absent percentage of the schools in the county. A bar graph is developed to depict this correlation and, after the visualization of the data for the Hypothesis, it depicts the other. Data is depicted in the bar graph as the bars and the width of the bins defines the average covid increase percentage and the values of the measures depict the average increase in the absent percentage of the schools in the county with respective of the months in the year 2020. Hue being the county names with the color Red for Burleson, Light Teal for Floyd, Orange for Harris, and Blue for Tarrant County respectively.

As we can see in the visualization that, in the month of August for Floyd County, the average covid increase percentage was 0.106 and the average absent percentage was 4.27, and the covid increase percentage increased gradually in September and October, and the absent percentage also increased. However, in the month of November, the covid increase percentage has increased from 1.386 to 2.093, but the absent percentage decreased from 11.86 to 2.11. This variation is almost similar in all the counties. The absent percentage may increase initially with the increase in covid percentage, but later, the same is decreased though, the covid percentage is increased, and vice-versa.

Therefore, we can conclude that the covid increase percentage has no effect on the absent increase percentage of schools in the county and hence we can reject this hypothesis.

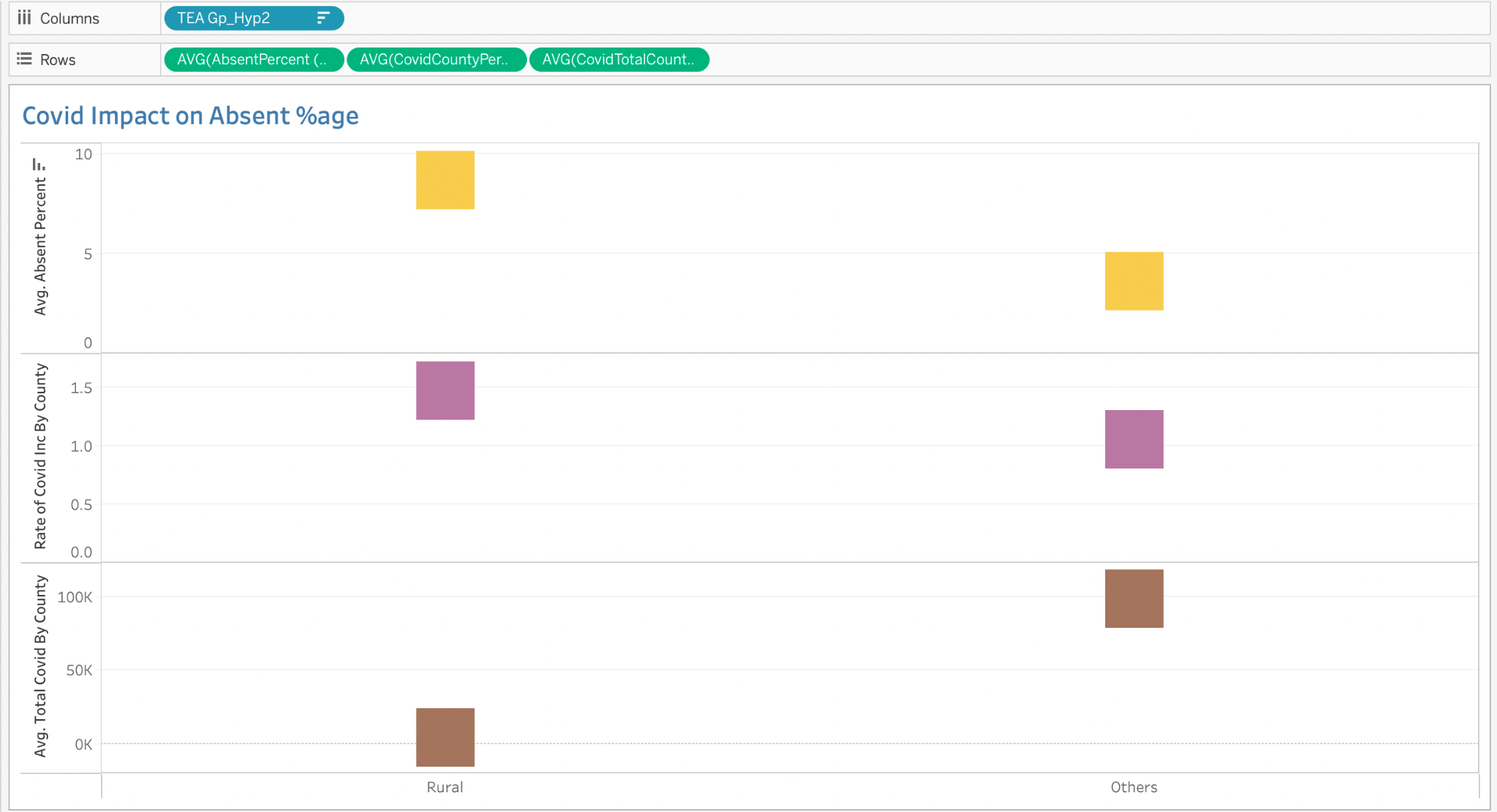
**Hypothesis 2**

Hypothesis 2 stated that there was a possibility of having a lower frequency of absentee in schools from rural areas when compared to urban and suburban areas during Covid. Data from schools were used to confirm if this holds true for the surveyed counties under the banner of Urban, Suburban and Rural areas.

First Average of Covid total cases was mapped against Targeted Areas where it showed that Urban areas had maximum number of covid count among three targeted areas. Second Average of Absent Percent was mapped against Targeted Areas where it displayed those Rural areas had a higher number of absent students from schools.

To understand why absent in schools from rural area had higher count despite having lower number of covid cases when compared to other targeted areas, visualization was made among Average of Absent Percent, Average of Covid total cases and Rate of Covid Increase Percentage with help of square shape and each column on y axis distinguished using color mark.

It was concluded that assumption made in hypothesis 2 about having lower frequency of absentee in schools from rural areas does not hold true and the reason for that was due to higher rate of covid count increase in rural areas when compared to urban and suburban areas.

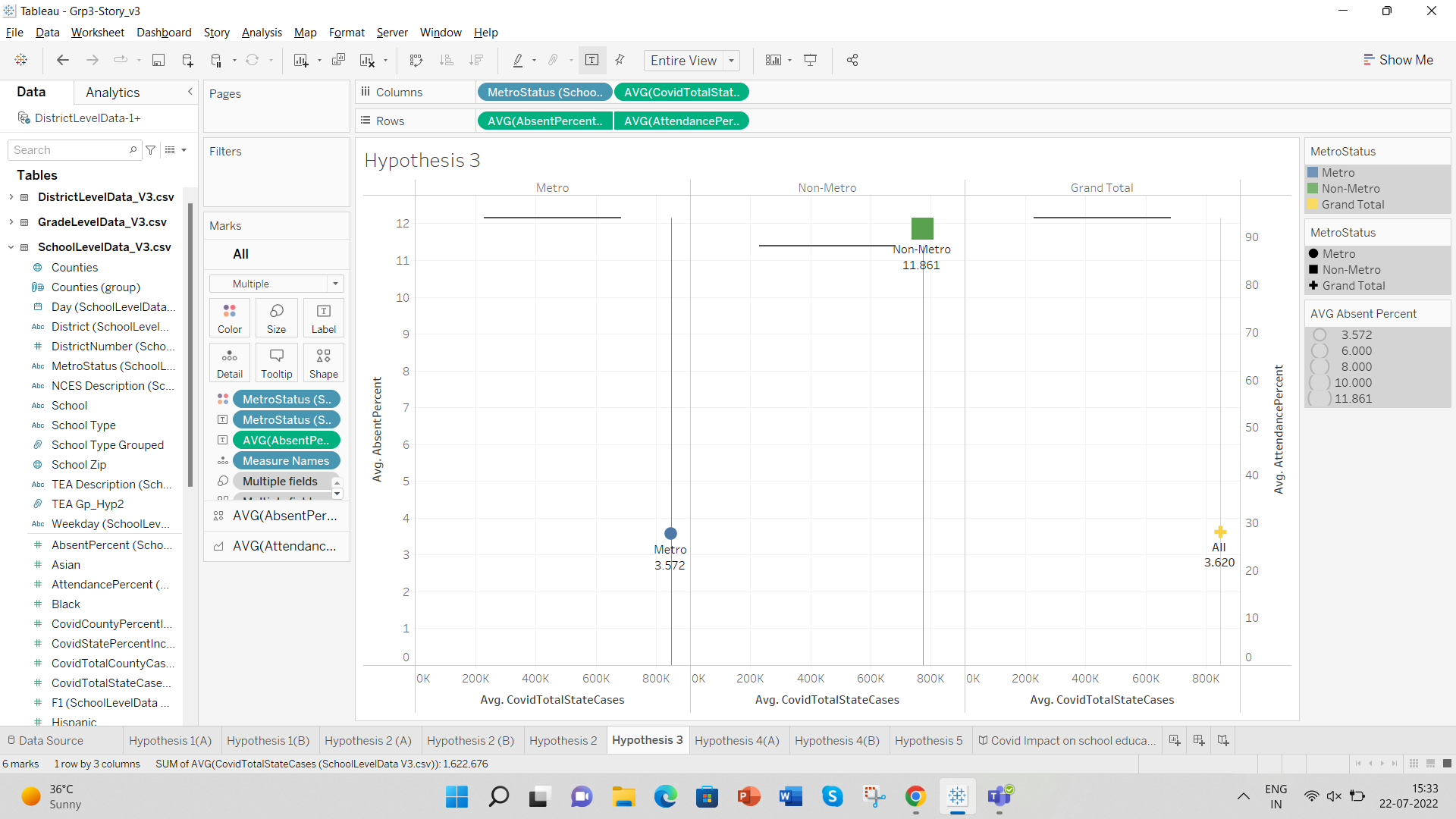


**Hypothesis 3**

Hypothesis 3 stated that, there were chances of having a higher average absentee percentage in non-metro areas than that of metro areas during Covid, since metro areas took better precautions. Data from schools were used to confirm whether the hypothesis can be proven for the surveyed counties under two categories, namely Metro and Non-Metro areas.

Two graphs were plotted to compare the average attendance and average absent percentage for Metro and Non-Metro areas. A third graph shows the overall statistics. The data of average attendance percentage and the data of average absent percentage were plotted against the data of average Covid total state cases. An area graph is plotted to represent average attendance percentage, whereas the average absent percentage is represented by different shapes. The size of the shapes varies with respect to the value they hold. The blue area and shape represent the data of Metro areas, the green area and shape represents the data of Non-Metro areas and the yellow area and shape represents the overall data.

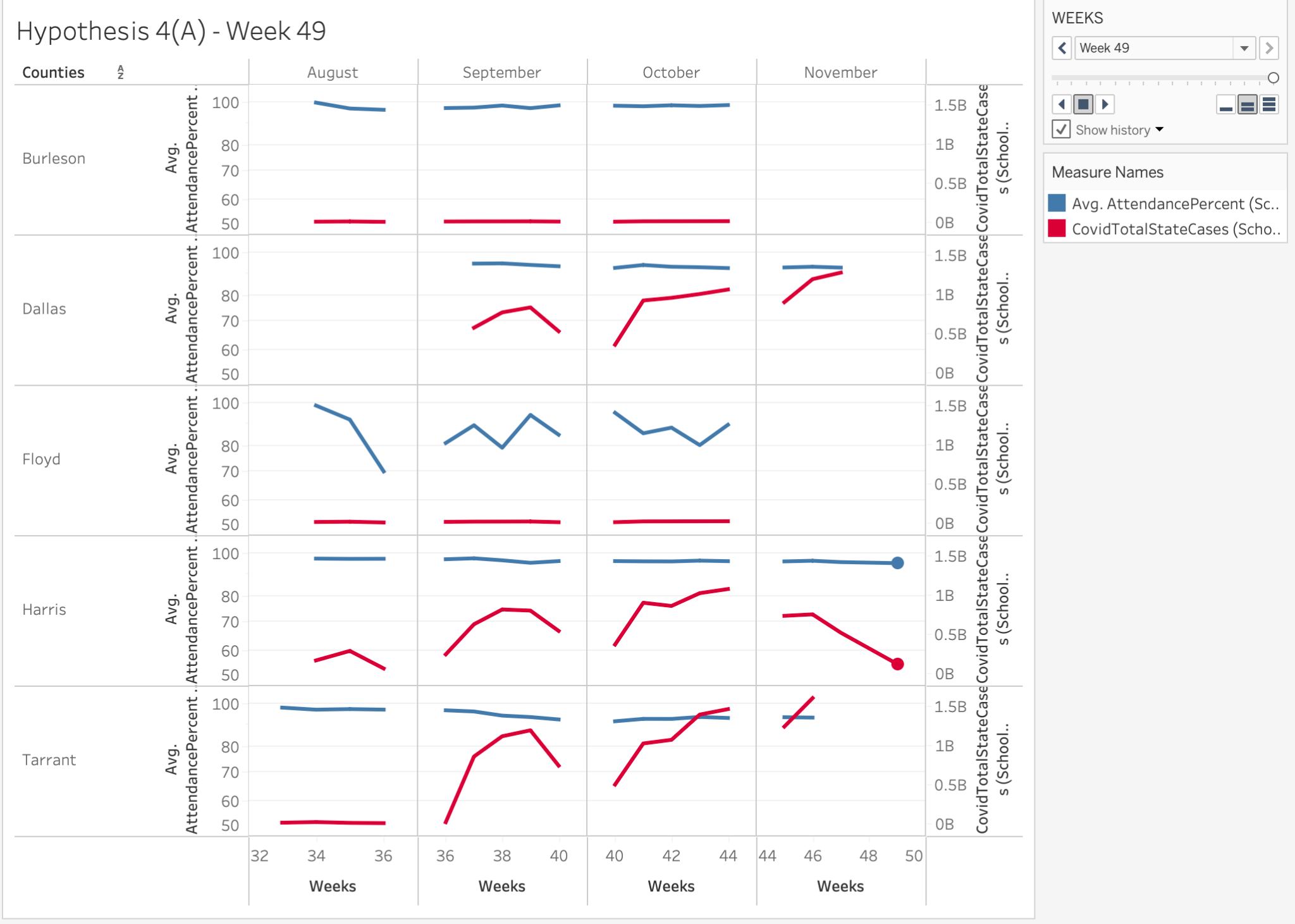
The visualizations clearly suggest that the average absent percentage in Non-Metro areas (11.861) is higher than that of Metro areas (3.572). This allows us to conclude that the proposed hypothesis can be proven successfully.



**Hypothesis 4**

Hypothesis 4 stated that there were chances of having counties with higher covid cases have lower present percentage in school from August 2020 to November 2020. Data from the school table were used to confirm whether the hypothesis can be proven for the surveyed counties under weeks, months, average attendance percentage and total number covid cases by county.

Data are shown in a line chart with animation where users can see how both lines are moving based on the weeks or a month. In this chart, the Red line chart represents the total number of covid cases per county per month. Green line chart represents the average attendance percentage per county based on the weeks from August 2020 to November 2020. Number on the left side on the Y axis represents the percentage and on the right side it represents the total number of covid cases. On the X-axis the bottom number represents the weeks based on the 52 weeks per year and on the top it represents the month that which weeks count on which month. From this chart we can see that when the number of covid cases are increasing, the attendance percentage is decreasing and when the number of covid cases are decreasing, the attendance percentage is increasing.



**Hypothesis 5**

According to Hypothesis 5, given their lower immunity, children in elementary school may have been more affected by Covid-19 than those in middle and high school. A bubble chart was developed to assess the effects of Covid-19 on students in different school types, such as elementary, middle, and high school students, in order to support this claim.

Data is shown in bubble charts as a collection of circles. The values of measure represent the sizes of those circles, whereas each value in the dimension field represents a circle. Elementary school is represented by color blue, High school by orange, and Middle school by color dark red. The size of the circles represents the overall number of COVID instances. The picture demonstrates that the hypothesis is valid and accurate, i.e., that the total number of COVD instances was highest among elementary school students, followed by middle school students, and finally, high school students. Therefore, we can conclude that elementary school students were most affected as clearly seen in Dallas, Floyd, Tarrant and Harris counties. Thus, Age was a crucial component that Covid Impact was supposed to take into account. The delay in vaccines for school children may have been another factor besides the lower level of immunity.

Chart, scatter chart, bubble chart

Description automatically generated

**Conclusion**

After the analysis of the three datasets namely Grade Level dataset, State Level dataset and District Level datasets on how COVID-19 impacted the education trends, we can conclude that there has been a major impact on education trends especially in terms of attendance percentages in schools due of Covid-19. Some of the conclusions from our analysis are listed below.

1. The first hypothesis, “There could be a direct correlation between the CovidCountyPercentIncrease and the AbsentPercent of the schools in that county.” - We reject this hypothesis as this is False, the absent percentage of the schools in the county is NOT affected by the increase in the Covid Percentage in the respective County.
2. The second hypothesis, “The possibility of having lower frequency of absentees in rural area schools is higher when compared to urban and suburban areas.” - We reject this hypothesis as the frequency of absentee from rural area school is higher when compared with school from urban and suburban areas because the rate of covid case increase was higher for rural areas.
3. The third hypothesis, “Chances of average absentee percentage in non-metro areas is higher than that of metro areas since metro areas took better precautions.” - We do not reject this hypothesis as Non-Metro cities had a higher average absent percentage and lesser average attendance percentage compared to Metro cities.
4. The fourth Hypothesis, “The Odds of having County’s with higher covid cases have lower present percentage in school from August 2020 to November 2020.” - we do not reject this hypothesis as the average percentage decreases while covid cases increase and the average percentage increases while covid cases decrease.
5. The fifth hypothesis, “Covid-19 affected Elementary school children more when compared to High School and Middle school children because of less immunity.” We do not reject this hypothesis as this hypothesis is true, Elementary school children are indeed affected more when compared to High School and Middle school children by Covid-19.

**References**

**https://www.kaggle.com/datasets/chrisiortiz/school-attendance-in-texas-covid-weather-ses/discussion/238939?resource=download**