Factors Affecting Students' Attendance Rate in Texas

Introduction

Students' attendance in school is a strong predictor of a student's career outcomes. While it is important that students should attend classes regularly, it is also important to understand what the potential factors could be affecting student attendance rate. With the recent covid-19 pandemic, many of the schools' attendance rate decreased. The Education Department wants to analyze what factors have contributed to the absent rate of students in the 2020-2021 academic year because there might be various factors like the demographics, covid cases, weather conditions, weekends, etc. that might affect student's willingness to attend the classes.

Identifying the appropriate factors can help the Education Department to plan, implement, and manage the activities carried out for students in the upcoming years. If the attendance rate is affected by some serious factors like extreme weather conditions or high covid cases, then the Education Department may grant permission to the schools to conduct online classes. However, if the attendance rate is low even during normal conditions, say normal weather, then the Education Department must take an initiative to conduct a deeper survey and address more specific issues or factors affecting the attendance of students in schools.

The scope of this project is limited to the Texas state, however, on a broader scale the results of this data may help the Education Department of other states make similar analysis and plan out the education activities for students in their states, considering Texas data analysis as a reference.

Data Set Explanation

The dataset that we are using has been downloaded from Kaggle. Vast majority of the data in this project was collected from school districts, public census information, and public health center data from across Texas. The focus behind this data collection was on "daily attendance," so the data that was sent back by 11 school districts was selected. The 11 school districts that sent data were (1) Conroe ISD, (2) Cypress-Fairbanks ISD, (3) Floydada ISD, (4) Fort Worth ISD, (5) Pasadena ISD, (6) Snook ISD, (7) Socorro ISD, (8) Klein ISD, (9) Garland ISD, (10) Dallas ISD, and (11) Katy ISD.

However, there were slight discrepancies in this dataset, three school districts sent daily attendance data that included student grade levels, but other school districts did not include this information. The timeline of this collected data is from August 2020 to December 2020, which was the 2020-2021 academic school year. Each row of the dataset is the data of every day attendance from various schools in Texas. This dataset has 19 variables and their description is given in the following table:

Table 1

Variable	Variable sample example	Description
Day	8/12/2020	Calendar day, month and year
Weekday	Wednesday	Day of the week
District	CONROE ISD	Name of District where the specific school is present
DistrictNumber	170902	Zip number assigned to specific district
Members	64469	The number of students enrolled that day
Absent	181	Number of students absent
AttendancePercent	99.72%	Total attendance of students
AbsentPercent	0.28%	Percentage of absent students

County	Montgomery	Name of the county where the school is located
TEA_Description	Other Central City	TEA (Texas Education Agency) description states whether the school is located in rural, suburban or other central city location
NCES_Description	City-Small	National Center of Education Statistics (NCES) description states whether the city, suburb, or rural region is small, large or distant
MetroStatus	Metro	Metro Status gives the information whether the school is located in metropolitan or non-metropolitan region
CovidTotalCountyCases	51943	Total number of covid cases on that specific day in that county
CovidCountyPercentIncrease	0.6738	This shows the covid percent increase rate in that county, as compared to the previous covid cases numbers
CovidTotalStateCases	531825	Total number of covid cases on that specific day in the state - Texas
CovidStatePercentIncrease	1.3124	This shows the covid percent increase rate in Texas, as compared to the previous covid cases numbers
PRCP	0	Precipitation for that day
Temperature Max	102	Maximum temperature in Fahrenheit for that day
Temperature Min	78	Minimum temperature in Fahrenheit for that day

Key Variables

Among all the 19 variables in the dataset, we have chosen to perform analysis on 6 key variables. These 6 key variables are shown in **Table 1**:

Table 1

Key Variables	Description
Absent	Number of students absent
District	Name of District where the specific school is present
CovidTotalCountyCases	Total number of covid cases on that specific day in the county where that school is present
Temperature Max	Maximum temperature in Fahrenheit for that day
Temperature Min	Minimum temperature in Fahrenheit for that day
Weekday	Day of the week

Descriptives of Key Variables

- From the 6 key variables, 2 variables were categorical: 'District' and 'Weekday'.
  - 'District' variable categories included CONROE ISD, FLOYDADA ISD, PASADENA ISD, SNOOK ISD, SOCORRO ISD, KLEIN ISD, GARLAND ISD, DALLAS ISD, FORT WORTH ISD, and KATY ISD.
  - 'Weekday' had categories as Monday, Tuesday, Wednesday, Thursday, and Friday.
  - For remaining 4 key variables, the below **Table 2** defines the name of the variable, classification of variable as Ratio or Interval is done, and their mean and standard deviation are calculated using the Excel Descriptive Statistics Tool in the Analysis Toolpak.

Table 2

Key Variables	Measurement scale	Mean	Standard Deviation
Covid cases	Ratio	80726.35	53592.87
Maximum temperature	Interval	84.06	11.81
Minimum temperature	Interval	62.99	12.32
Number of absentees	Ratio	3313.78	3195.95

Statistical Analysis Approach

1. Correlation

The first statistical analysis conducted on the data was to find the correlation between
1. 'Absent' students and 'Covid Total County Cases', and
2. 'Absent' students and weather (i.e. Minimum and Maximum Temperature).

For the purposes of this course project, a high correlation was defined as having a correlation coefficient with absolute value between 0.6-1, a moderate correlation as being between 0.4-0.6, and a low correlation between 0-0.4. The alpha used to evaluate the *P-value* was 0.05.

1. We found that the correlation value for 'Absent' students and 'Covid Total County Cases' was **0.366**, which indicated low and positive correlation. The individual values were plotted on a scatter plot, as seen in **Figure 1**. In addition, a trendline was added to show the correlation between the two variables.

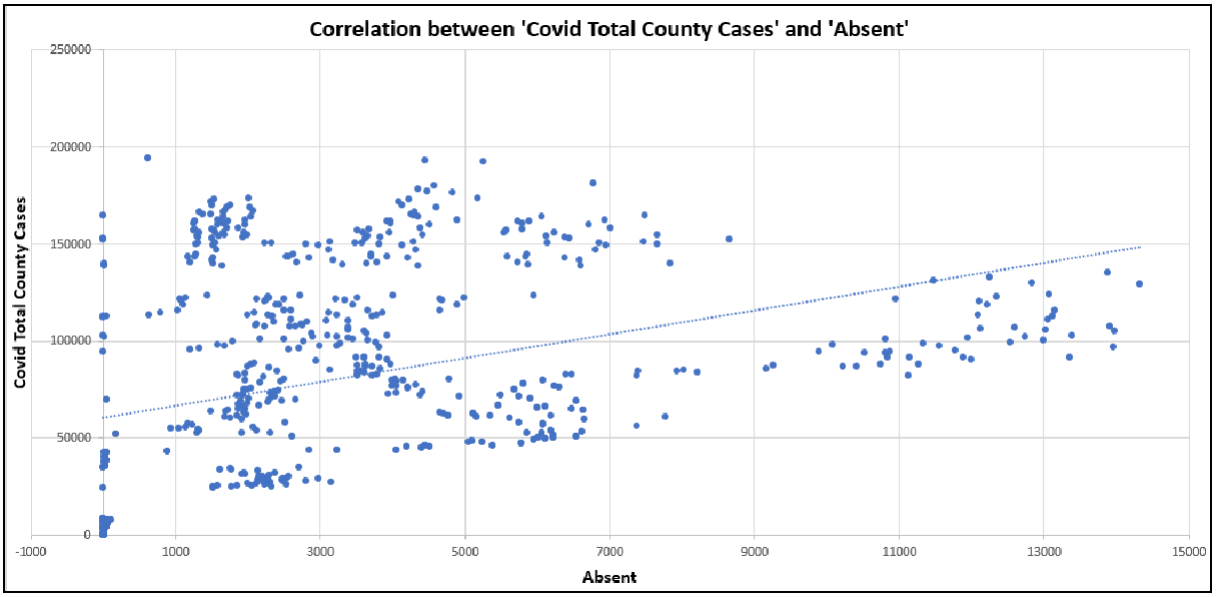

Figure 1

2. For 'Absent' and 'Temperature minimum', the correlation value was **-0.183**, indicating negative correlation. Also, for 'Absent' and 'Temperature maximum', the correlation value was **-0.242**, which indicated negative correlation.
   Thus, we discarded the *Weather* factor in our further analysis.

2. Multiple Regression Analysis

Using the Regression tool in Excel Data Analysis Toolpak, we ran regression analysis on various variables.
For regression analysis, we considered the Null and alternative hypothesis as follows:

- H0: The variable does not have influence on the absent rate of students
- H1: The variable has influence on the absent rate of students

❖ First, we considered only the *County Total Covid Cases* for regression analysis. See **Table 3.**

Table 3

Regression Statistics	
Multiple R	0.366218182
R Square	**0.134115757**
Adjusted R Square	**0.132807775**
Standard Error	2976.166407
Observations	664

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	908222755.2	9.1E+08	102.536375	1.66989E-22
Residual	662	5863709012	8857566		
Total	663	6771931767			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	1550.795841	208.9303883	7.42255	3.5425E-13	1140.549757	1961.041925	1140.549757	1961.041925
CovidTotalCountyCases	0.021838981	0.002156718	10.126	1.6699E-22	0.017604149	0.026073814	0.017604149	0.026073814

Considering only the Covid cases factor, we inferred that the Adjusted R Square value was **0.13** and the p-value was less than 0.05. Thus, Null hypothesis is rejected and alternate hypothesis is accepted stating that *Covid County Total Cases* has influence on the absent rate of students.

❖ The second multiple regression was run on the *Districts* variable. The districts included in the *Districts* column were CONROE ISD, FLOYDADA ISD, PASADENA ISD, SNOOK ISD, SOCORRO ISD, KLEIN ISD, GARLAND ISD, DALLAS ISD, FORT WORTH ISD, and KATY ISD. These districts were encoded into a data matrix of 0's and 1's using a one-hot encoding method. A zero-base variable was not used; each district was compared against the others. The output of the regression returned both positive and negative coefficients.
See **Table 4**.

Table 4

Regression Statistics								
Multiple R	0.890002142							
R Square	**0.792103812**							
Adjusted R Square	**0.788920103**							
Standard Error	1468.327589							
Observations	664							
ANOVA								
	df	*SS*	*MS*	*F*	*Significance F*			
Regression	10	5364072968	5.36E+08	248.7990735	3.6473E-215			
Residual	653	1407858799	2155986					
Total	663	6771931767						
	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	5999.85	232.1629767	25.84327	5.6184E-102	5543.973968	6455.726032	5543.973968	6455.726032
CONROE ISD	-4058.39717	307.5363835	-13.1965	2.03514E-35	-4662.276687	-3454.517653	-4662.276687	-3454.517653
FLOYDADA ISD	-5964.138136	300.7354017	-19.8318	7.19116E-69	-6554.663221	-5373.61305	-6554.663221	-5373.61305
PASADENA ISD	-4137.766667	299.7211141	-13.8054	3.18084E-38	-4726.300093	-3549.23324	-4726.300093	-3549.23324
SNOOK ISD	-5989.646296	306.3091783	-19.5543	2.23936E-67	-6591.116069	-5388.176523	-6591.116069	-5388.176523
SOCORRO ISD	-3841.584694	312.8887345	-12.2778	2.52978E-31	-4455.974106	-3227.195282	-4455.974106	-3227.195282
KLEIN ISD	-4603.197826	317.4411861	-14.5009	1.64062E-41	-5226.526448	-3979.869204	-5226.526448	-3979.869204
GARLAND ISD	-2284.801613	297.7811512	-7.67275	6.14606E-14	-2869.525722	-1700.077504	-2869.525722	-1700.077504
DALLAS ISD	5130.538889	306.3091783	16.74954	1.18443E-52	4529.069116	5732.008662	4529.069116	5732.008662
FORT WORTH ISD	-1948.6	268.0787142	-7.26876	1.03981E-12	-2475.0003	-1422.1997	-2475.0003	-1422.1997
KATY ISD	-2185.685821	293.3914483	-7.44973	2.97424E-13	-2761.790294	-1609.581348	-2761.790294	-1609.581348

Considering only the Districts variable, we can see that the Adjusted R Square value is **0.78** which is greater than the Covid cases regression statistics.
Furthermore, all the p-values are less than 0.05 which means that all these districts are significant predictors of student's absent rates.
As the p-value is less than 0.05, we reject the null hypothesis and accept the alternative hypothesis stating that the 'Districts' variable has an influence on the absent rate of students.

❖ The third multiple regression was run by considering the *Districts* variable as well as the *Covid Total County Cases.*
Since both the variables influence the absent rate of students, we further wanted to determine that during the covid, which districts were significant predictors of student's absence rate.
See **Table 5** for regression analysis.

Table 5

Regression Statistics	
Multiple R	0.90053746
R Square	**0.81096772**
Adjusted R Square	**0.80777852**
Standard Error	1401.20112
Observations	664

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	11	5491818058	5E+08	254.286031	2.237E-227
Residual	652	1280113709	1963365		
Total	663	6771931767			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	2038.86162	538.7217518	3.78463	0.00016813	981.022696	3096.70055	981.022696	3096.70055
CONROE ISD	-2017.73549	387.4680599	-5.2075	2.5696E-07	-2778.57129	-1256.89969	-2778.57129	-1256.89969
FLOYDADA ISD	-2152.04595	552.910689	-3.8922	0.00010955	-3237.74641	-1066.3455	-3237.74641	-1066.3455
PASADENA ISD	-3919.82425	287.2923372	-13.644	1.8168E-37	-4483.95409	-3355.6944	-4483.95409	-3355.6944
SNOOK ISD	-2038.51338	570.4220023	-3.5737	0.00037796	-3158.59921	-918.427554	-3158.59921	-918.427554
SOCORRO ISD	-846.808198	476.4410251	-1.7774	0.07597478	-1782.35212	88.7357278	-1782.35212	88.7357278
KLEIN ISD	-4698.22261	303.1579367	-15.498	2.3752E-46	-5293.50628	-4102.93893	-5293.50628	-4102.93893
GARLAND ISD	-783.609334	339.6872341	-2.3069	0.02137541	-1450.62227	-116.596395	-1450.62227	-116.596395
DALLAS ISD	6312.20783	326.9612213	19.3057	5.0353E-66	5670.18381	6954.23186	5670.18381	6954.23186
FORT WORTH ISD	557.311756	402.4413435	1.38483	0.16657896	-232.925722	1347.54924	-232.925722	1347.54924
KATY ISD	-2163.98703	279.9916073	-7.7288	4.1204E-14	-2713.78109	-1614.19297	-2713.78109	-1614.19297
CovidTotalCountyCases	0.02740871	0.00339795	8.06625	3.4842E-15	0.02073647	0.03408096	0.02073647	0.03408096

From the above regression analysis, we can infer that the Adjusted R Square value is **0.80**, which is even greater than the previous regression statistics.
We can see that all the variables except the Districts 'Socorro ISD' and 'Fort worth ISD', have p-value < 0.05
Hence, all the variables except the two districts - 'Socorro ISD' and 'Fort worth ISD', are significant predictors for the absent rate of students.

We also performed regression analysis on the *Weekdays* variable to check if this variable has any influence on the attendance rate. But the p-value for all weekdays was greater than 0.05, which means that *Weekdays* do not have an influence on the student's attendance rate.

We selected '*Districts' and 'Covid Total County Cases'* variables for multiple regression analysis for following reasons:

1. The Districts data will help us understand whether the districts lie in the rural, urban, suburban or city. This can help us analyze the pattern of attendance in rural, urban and/or central regions.
   Interestingly, we found that the absent rate of students is high in the DALLAS ISD District and lowest in FLOYDADA ISD and SNOOK ISD. [See **Figure 2**].

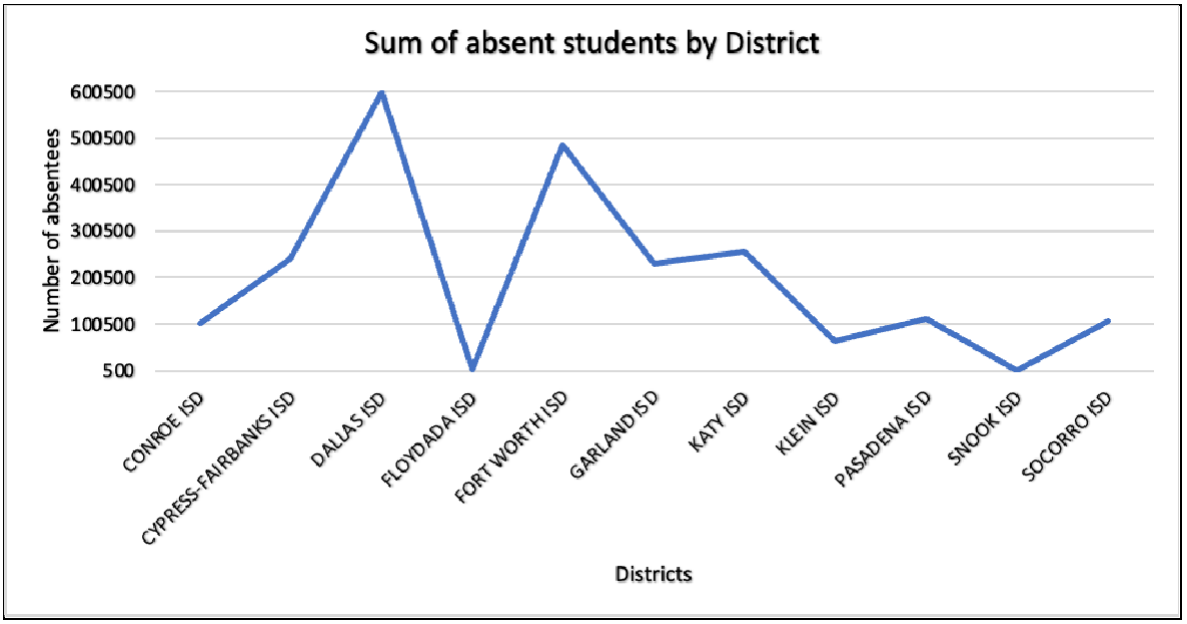

Figure 2

On investigating our dataset thoroughly, we figured out that the DALLAS ISD falls in the Major Suburban region while the FLOYDADA ISD and SNOOK ISD lie in the Rural region.
This implies that the absent rate of students is less in the Rural region as compared to other regions, and it is highest in the Major Suburban parts of the region.

2. Covid is a serious issue and students will most likely not attend in-person class for safety issues. However, we wanted to understand whether covid is the only factor affecting the attendance rate or are there other factors contributing to the attendance as well. As seen in the above analysis, we discovered that not only Covid, but also the 'Districts' factor affects the attendance rate of the students.

Other possible factors from the dataset were *'Weather'*, and *'Weekdays'* to determine whether they had any influence on the attendance rate of students. As explained above in the correlation and multiple regression part, we found that these two factors don't have any influence on the attendance rate of the students.

Results and Recommendations

- **Results**
  The results of our analysis convey that *District* and *Covid Total County Cases* factors are contributing towards the attendance rate of the students in Texas.
  - Highest number of absent students are found in the Major Suburban District region
  - Lowest number of absent students are found in the Rural Suburban District region
  - The rise in Covid Cases will mean that a greater number of students will remain absent

- **Recommendation**
  - As the pandemic is ongoing and considering that the cases may rise in future, the Education Department should grant permission to the schools for conducting online classes.
  - Considering the end of the pandemic, then it would mean that the location/regions where the schools are located, play a major role in affecting the students' attendance. Since the regions also have an impact on student's attendance rate, the Education Department should conduct a deeper survey and analyze the issues about why major students remain absent in the *Major suburban* regions as compared to the other regions.
  - Although only Texas state is considered in this dataset, it can help other states of the country to conduct a thorough survey from the beginning itself which can help them determine more specific factors for the absent rate of students, and accordingly they can plan and implement different protocols or measures as required.