Course: CAPP 30122
Team: Apichat Sukonthasakorn (apichat)
Charisma Lambert (charisml)
Ambert Avila (aravila)

Project: Does school performance correlate to the socioeconomic indicators of an area?

---

## Abstract:

Using data from City of Chicago and ZIP Atlas, we examined how the socioeconomic indicators of a given area impacts school performance metrics. The socioeconomic indicators of interest were: median household income, poverty rate, and unemployment rate. We defined an area by its zip code and, looking only at High Schools, we examined the data of seven performance areas:

- **student attainment:** based on how well the school performed on standardized tests; schools are rated on a scale of ‘far below expectations’, ‘below average’, ‘average’, ‘above average’, ‘met expectations’, and ‘far above expectations’, which we hard coded numerically on a scale of 1 - 6.
- **culture/climate:** based on student and teacher responses to the My Voice, My School ‘5Essentials’ survey; schools are rated on a scale of ‘not yet organized’, ‘partially organized’, ‘moderately organized’, ‘organized’, and ‘well organized’, which we hard coded numerically on a scale of 1-5; for ‘not enough data’ we labeled None.
- **Mobility rate:** the percentage of students who experienced at least one transfer in or out of the school, excluding graduates (Illinois Report Card).
- **Chronic truancy:** a student who is absent from school without valid cause for 5% or more days at any time of the school year (CPS Comprehensive Policy on Attendance).
- **11th grade SAT score:** the average SAT score of 11th grade students at the school.
- **Drop out rate:** the percent of students enrolled in grades 9-12 at any time during a school year who dropped out during that year (CPS Office of Accountability).
- **Suspension rate:** the removal of a student from their regular educational schedule for in-school or out-of-school consequences (CPS Student Code of Conduct).

We cleaned the data to show the average of each performance area by zip code in order to run a regression analysis of the socioeconomic indicators on the school performance areas of all schools in that area. In our research we found that median household income has a statistically significant relationship with SAT score, poverty has

a statistically significant relationship SAT score and suspension rate, and unemployment has a statistically significant relationship with chronic truancy and suspension rate.

To showcase our findings, we built a web application to display socioeconomic indicators by zip and the relationship between socioeconomic indicators and statistically significant school performance metrics. Our hope is that our project pushes for more transparency in data reporting from CPS. We encountered some schools who did not have data points for all of the performance metrics that we were examining and felt that this gives an incomplete picture of how schools are performing. We would also hope that our project serves as a baseline to build upon with more data or external factors that allows the district to examine educational outcomes from a more holistic lens. Students are not just what we see in front of us, but a culmination of their daily life experiences. Our project is a starting point, but there are so many other factors that can allow school boards and policy makers to make more informed decisions for supporting student development.

**<u>Interface:</u>**

Download our repository locally by cloning it to your local network. After cloning, ensure your local environment is set up to run our code by navigating to the folder where you've cloned and run 1) "poetry install" and 2) "poetry shell" in the command line.

Finally, run alias vsapp='PYTHONPATH=$(pwd) python3 coded_school/Visualization/app.py', then vsapp. Copy the generated url (http://127.0.0.1:8050/) into your webpage to interact with our application and findings.

**<u>Structure:</u>**

```
Coded-school-Chicago
↳.gitignore
↳README.md
↳poetry.lock
↳pyproject.toml
↳coded_school
   ↳__init__.py
   ↳ Data
        ↳__init__.py
        ↳Step 1: income_data.py
        ↳Step 2: api_school_data.py
                    ↳merged_data.csv
        ↳Step 3: analysis.py
```

↳**Visualization:**

↳__init__.py<br>
↳ Step 4: app.py

**Roles:**

| Name | Module | Tasks |
|---|---|---|
| Apichat (apichat) | Visualization/ | • **app.py:** This script features a Dash application used for visualizing data.<br>• It starts with importing all the gathered data, including financial and social indicators by zipcode, summarized information on public high school performances by zipcode, and comprehensive data containing school performance names and locations.<br>• Following that, I define a function to conduct regression analysis. This function takes a list of attributes representing the independent variables we wish to analyze for their impact on a dependent variable. It returns the coefficient of the analysis, the predicted value of the model, the standard error, the p-value, R-square value, and F-statistics.<br>• Next, I proceeded to the visualization section, which comprises three distinct visualizations: 1) A summarized table showcasing the results of the regression analysis. 2) A choropleth map illustrating the |

| | | |
|---|---|---|
| | | distribution of financial and social information across zip codes. 3) A scatter plot indicates a regression line depicting the relationship between variables.<br>● The table displays the results of regression analysis, where inputs include school performance indices such as student attainment rating, cultural climate rating, mobility rate, chronic truancy, SAT scores for grade 11 students, dropout rate, and suspension rate. These factors serve as independent variables, while financial and social elements serve as the dependent variable. Utilizing the previously specified regression function, all relevant data is calculated. Dash_table package from plotly is used to automatically update the table based on specified values.<br>● The choropleth map illustrates the financial and social aspects within each zipcode, selectable via tick boxes. Additionally, individual school information is overlaid on the map, displaying the performance metrics for each school. The GeoJSON files incorporate zipcode boundary data sourced from the City of Chicago web portal's API. For implementation, the choropleth_mapbox package is used to generate the choropleth |

| | | |
|---|---|---|
| | | map, followed by the addition of scatter_mapbox traces to overlay school information onto the map.<br>● The scatter plot and line graph uses financial and social aspects as the Y-axis and school performance indices as the X-axis. Users can select data from the dropdown box. Plot adjustments include scaling up when the Y-axis values are large, and color changes according to the magnitude of X. |
| | Data/ | ● **Income_data.py:** This script is designed to collect financial and social indicators representing the areas within the city of Chicago. It includes functions to request and scrape data from ZIPatlast.com's tables, extracting key metrics such as the unemployment rate, poverty level among families, percentage of population enrolled in high school, and median household income in US dollars by zip code. Following data retrieval and scraping, the cleaned data is organized into a dictionary, with zip codes as keys mapped to their corresponding information based on predefined categories. Subsequently, the script proceeds to clean and format the data into suitable formats for analysis purposes. Finally, the dictionaries are converted into Pandas dataframes, ready for merging with datasets containing school information |
| Amber (aravila) | Data/ | **api_school_data.py:** |

| | | |
|---|---|---|
| | | This file collects school information through Chicago Public School's portal API and cleans the data to return a dataframe that contains zip code, student attainment rating, culture climate rating, mobility rate, chronic truancy rate, SAT scores from 11th grade, dropout rates, suspension rates, latitude and longitude and the schools name. In order to merge our educational data with our income data, a function is made to take zip codes and average the rest of the columns to the area. This is then combined with income data based on the zip code to run analysis on the factors. |
| Charisma (charisml) | Data/ | **analysis.py:**<br>This file contains the foundations of the regressions and graphs, represented on the app. The aim is to explore the relationship between the school attributes and socioeconomic indicators. This is completed with the loading of merged_data.csv, the cleaned and merged data set from api_school_data.py, and runs linear regression analysis.  I used the statsmodels library to model the relationship between factors. There is a component of data visualization that was built upon for the app that configures the data found on scatter plot graphs with hover features.<br><br>In short, the regression analysis helps to quantify the impact and the scatter plots help to visualize it.This information was built upon for the final application. |

**<u>What the project tried to accomplish and what it accomplished:</u>**

We hoped to create a Map of CPS High Schools, overlayed with performance data and socioeconomic indicators. Although we were able to do exactly that, it did come with some roadblocks.

Initially, we hoped to scrape Great Schools because it had all of the school performance metrics we were initially looking for: avg ACT / SAT score, AP participation, IB enrollment, Graduation Rate, low-income student percentage, chronic absenteeism, and student:teacher ratio. Great Schools served as the one-stop-shop for all of those data markers. We set up the structure of our project to include 3 scrapers: 1) scrape the school page for performance metrics, 2) scrape the website to get all urls off of the page and 'next page' url, and 3) scrape income data. Great Schools runs on Javascript, so we ran into a plethora of issues with getting close to the data points we wanted and ultimately were unable to get the urls necessary for running the scraper.

We pivoted to using the City of Chicago- Schools Progress Report API to acquire school data. This required us to establish new data points to examine because there was no direct alignment to what we previously hoped to examine and what the progress report assessed for. We were able to acquire data on student attainment, culture/climate, mobility rate, chronic truancy, 11th grade SAT score, drop out rate, and suspension rate. Student attainment and culture/climate data came categorically, so we had to hard code that to numerical values in order to run the regression analysis. Our data collection and analysis process was rather seamless after that.

In our final product, we created an interactive application that contains two interactive visuals and one interactive data report. The first visual is the map, which contains outlines for each zip code and one socioeconomic indicator, and points representing High Schools. Users can manipulate which socioeconomic indicator is shown on the map. The High School points can be hovered over to reveal the school's performance metrics that we were able to collect. The second visualization shows the relationship between the socioeconomic indicators and school performance metrics. This, too, can be manipulated by the user to investigate relationships of interest. Finally, we have the multiple regression analysis, where users can run the regression of their choosing, again manipulating socioeconomic indicators and school performance metrics. These three tools create the picture we were hoping to display all along.