Exploratory Data Analysis

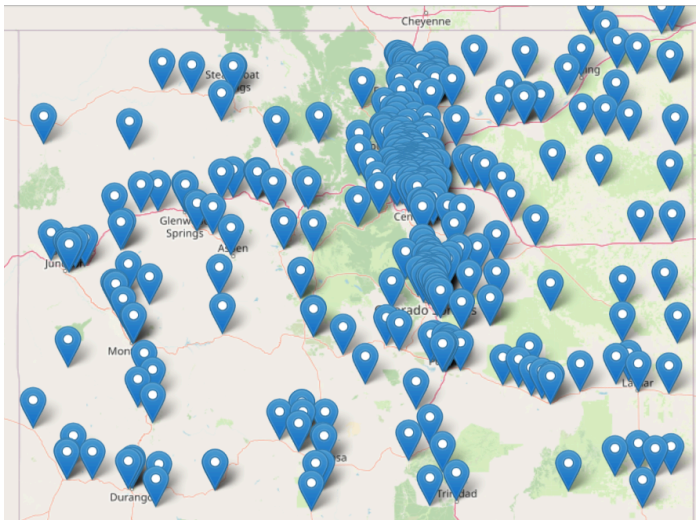
Melody Goldanloo, Jenna Ramsey-Rutledge, Byron Selvage

Introduction:

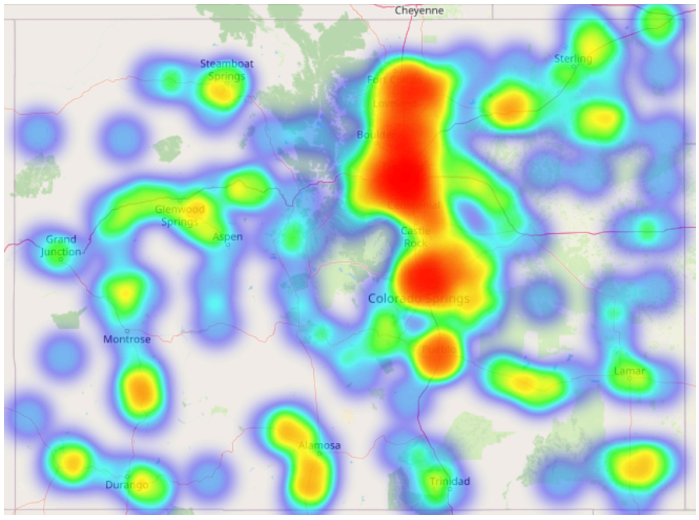
This project aims to develop a model to predict student outcomes, such as graduation rates, from publicly available education data. Our cleaned dataset focuses on high schools in Colorado and includes features such as enrollment, title I eligibility, teacher salaries, and total school expenditure.

Feature Exploration:

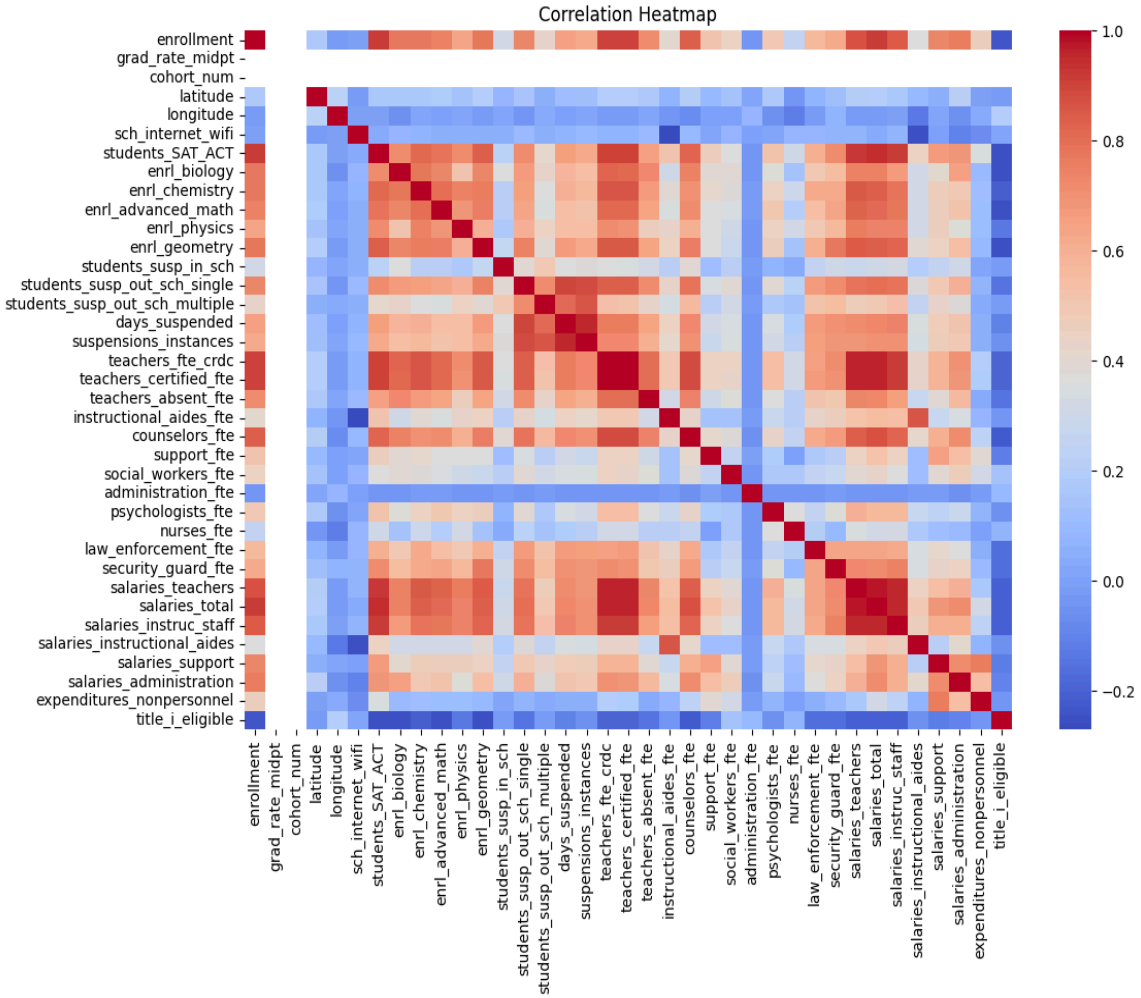
To begin data exploration, we used the folium library to create an interactive map of all the locations of the schools in our dataset.



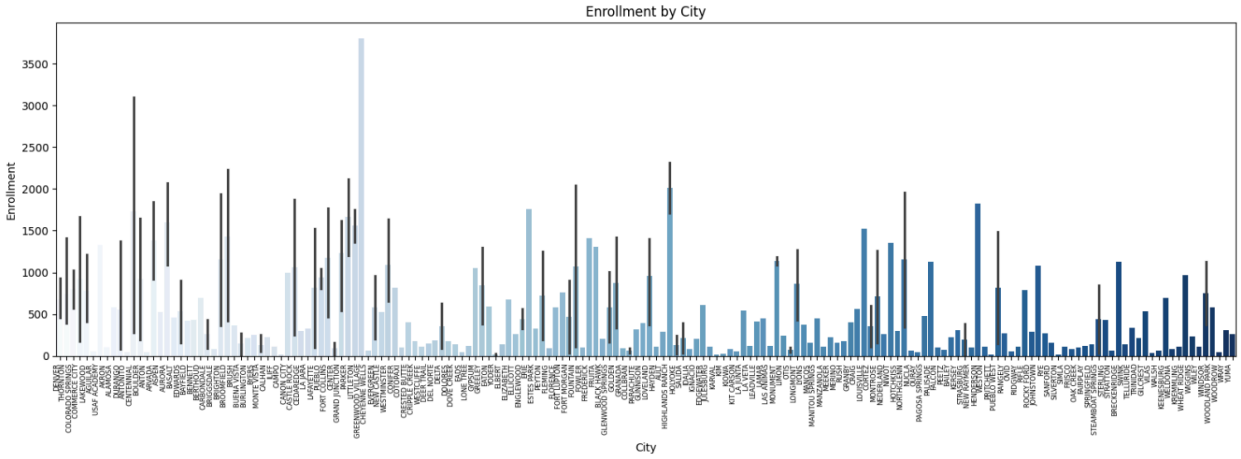
This visualization was very useful in showing what spread we had across the entirety of Colorado. We also experimented with folium to produce more maps such as a heat map showing graduation rates.



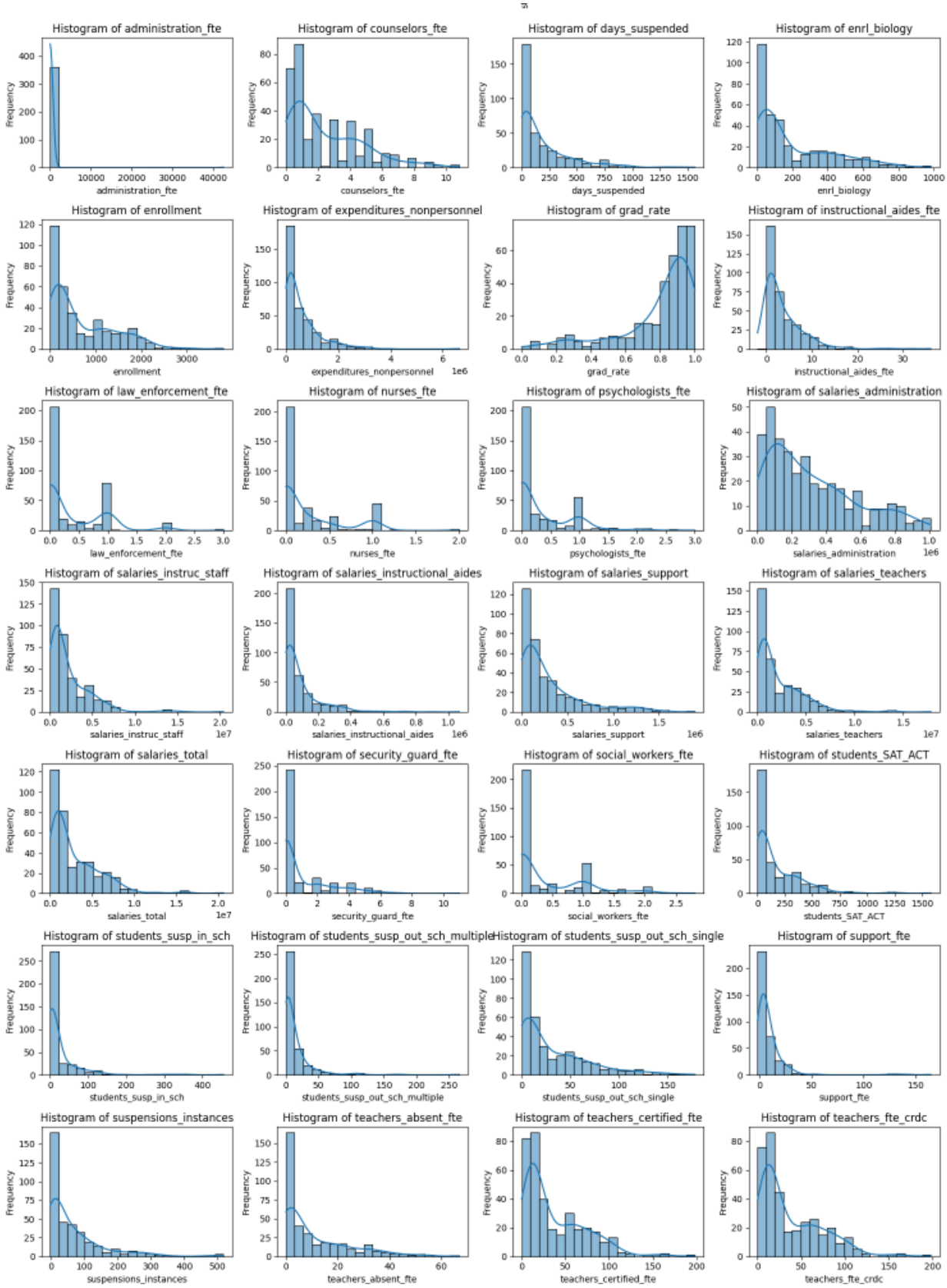
To explore the correlation between features in our dataset we used a correlation heatmap.



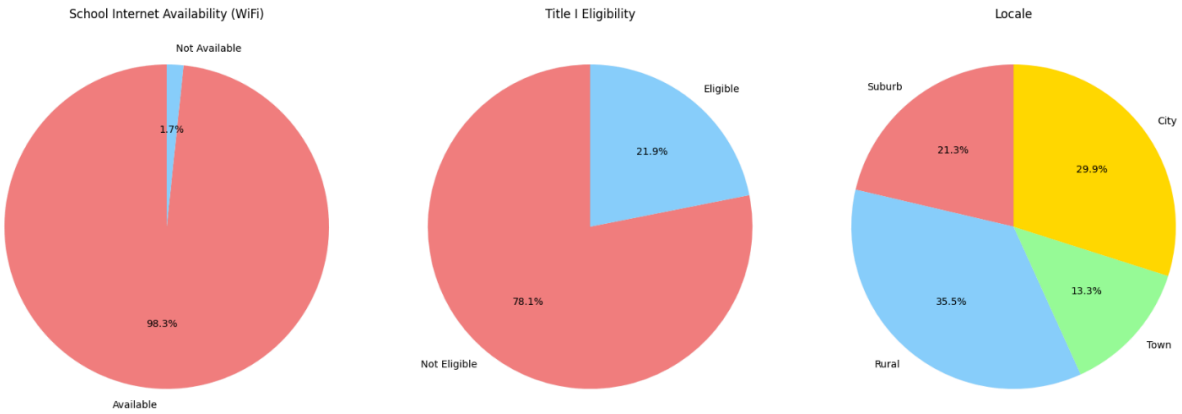
We used bar plots to explore specific relationships between features, for example, enrollment by city.



We then used histograms to explore the distributions of our continuous features.



And we used pie charts to explore the value counts of our categorical features.



Summary Data Statistics:

We used the function .describe() to generate summary statistics including the mean, median, and standard deviation of our dataset features. Shown below is the output for our first six features.

	administration_fte	counselors_fte	days_suspended	enrl_biology	enrollment	expenditures_nonpersonnel
count	361.000000	361.000000	361.000000	361.000000	361.000000	3.610000e+02
mean	122.227147	2.489418	187.770083	201.119114	700.858726	5.940978e+05
std	2236.608500	2.280215	256.061407	218.454279	703.369943	7.394952e+05
min	-2.000000	0.000000	-2.000000	1.000000	0.000000	-2.000000e+00
25%	1.500000	1.000000	16.000000	35.000000	132.000000	1.220140e+05
50%	3.180000	1.900000	78.000000	100.000000	387.000000	3.018480e+05
75%	6.000000	4.000000	269.000000	336.000000	1180.000000	8.256390e+05
max	42500.000000	10.800000	1566.000000	982.000000	3797.000000	6.618479e+06