Caiden Puma
DATA0200 – Group 2

Scaffold Activity 2 White Paper

COLAB LINK: DATA 0200 Scaffold Activity 2.ipynb

Question 1:

The expanded data dictionary includes the summary statistics for the variables that we have narrowed the dataframe down to. In Scaffold 1, we excluded observations that has n/a values as we need values in order for the district data to be useful so no variables have missing values. Most all of the data is numerical, with the only categorical variable being the district name which are all unique. As such, the frequency of each category is not included and all statistics pertain to the numerical nature of the variables:

Variable Name	Number of Observations	Data Type	N	N Missing	Mean	Standard Deviation	Minimum	Maximum
District Name	290	Categorical	276	0	n/a	n/a	n/a	n/a
Total number of buses	290	Discrete	276	0	39.90942	88.102146	1	1271
(1) Chronic Absenteeism Rate	290	Continuous	276	0	31.388406	11.811945	2.3	75.7
(2) Total High School Completers	290	Discrete	276	0	1081.626812	2695.202746	12	40198
Enrolled in College	290	Discrete	276	0	668.826087	1547.326687	3	22065
College Going Rate	290	Continuous	276	0	56.906522	14.307207	18.9	88.4
Enrolled In-State	290	Discrete	276	0	602.826087	1426.736961	0	20524
Enrolled Out-of-State	290	Discrete	276	0	65.98913	143.002924	0	1539
Not Enrolled in College	290	Discrete	276	0	412.800725	1182.072787	6	18133
(3) Total Student Cohort	290	Discrete	276	0	1164.213768	2917.917514	11	43582
Regular HS Diploma Graduates	290	Discrete	276	0	1035.086957	2561.990093	9	38084
Adult Ed. HS Diploma (Count)	290	Discrete	276	0	0.228261	1.24553	0	14
GED Completer (Count)	290	Discrete	276	0	0.894928	3.815734	0	48
Dropout (Count)	290	Discrete	276	0	75.644928	228.729767	0	3378
Still Enrolled (Count)	290	Discrete	276	0	33.297101	109.448496	0	1338

Question 2:

*All graphs/tables from the output of the colab can be found at the end of the document.

For the univariate analysis, we began by graphing all of the numerical variables as histograms. This included all variables other than the districts, which was our only categorical variable. We can see that the distributions for college going rate and chronic absenteeism are fairly normal, with more observations in teh center of the distributions and less on the edges. All of the other distributions seem to have mostly observations on the left side of the distribution, making them pretty skewed (the data isnt evenly spread out). These are also confirmed in looking at the data dictionary when we look at the mean (average) and standard deviations (realted to the spread of the data) when compared to the range of each of the variables. We then graphed box plots for each of the variables individually, again confirming the distributions that we saw in the histograms. The distributions for college going rate and chronic absenteeism have large quartiles whereas the others are fairly compressed, each with quite a few outliers. Finally, we create some pie charts pertaining to the college going and graduation rates which showed about 60 percent of student attended college, of which about 90 percent stayed in state. For the graduation rates, nearly 90 percent gradduated with the next largest group being drop outs at 6.5 percent followed by still enrolled, other, and GED, respectively.

For the bivariate analysis, we began by computing a correlation matrix to give us an idea of how all of the variables were associated with one another (with higher values indicated more of an association). The strongest correlations (>0.9) were between total number of bussses and total student cohort, total number of busses and total high school completers, total high school completers and enrolled in college, total high school completers and total student cohort, enrolled in college and enrolled in state, as well as total student cohort and regular high school diploma graduates. Those with moderate correlations (0.5-0.9) included those like total number of busses and drop out rate, enrolled out of state and GED completers, and total student cohort and still enrolled, among others. There were also a few interesting negative correlations which we must note, particularly in chronic absenteeism and college going rate as well as chronic absenteeism and enrolled out of state. We also plotted scatter plots for the number of busses against chronic absenteeism, college going rate, and high school completers which did not seem to show any significant relationships.

To contextualize these findings, we must consider that the outliers and distributions are likely due to some districts having larger populations. For example, the largest county is going to need a higher number of busses, have a higher number of students and thus graduates, college attendees, etc. than the smallest county which has a fraction of the population and area. The strong correlations between total student cohort and variables like total high school completers, regular HS diploma graduates, and college enrollment reflect the skew we see in the distributions. Large school districts inherently produce more graduates and college-goers, which is what we would expect. We see that college-going rate and chronic absenteeism have distributions centered around the mean, with larger interquartile ranges which aligns with the weaker correlations these variables have with variables like the total student count. In summary,

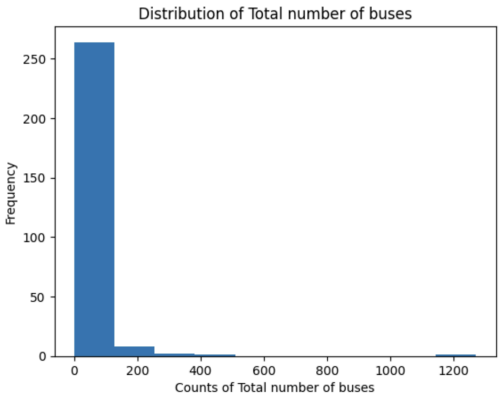
we see large districts presenting as outliers and driving factors like the mean and standard deviations up, skewed distributions in metrics related to total student cohort due to district size, and more normal distributions in college going rate and absenteeism resulting in lower correlations.

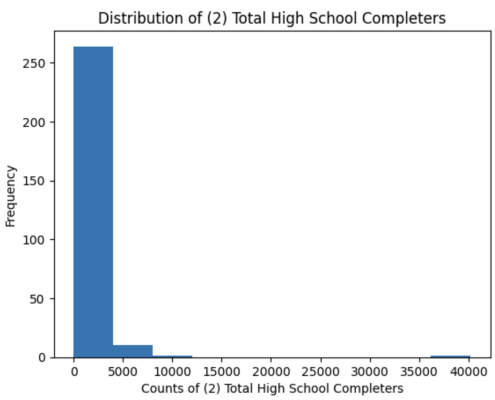
Question 3:

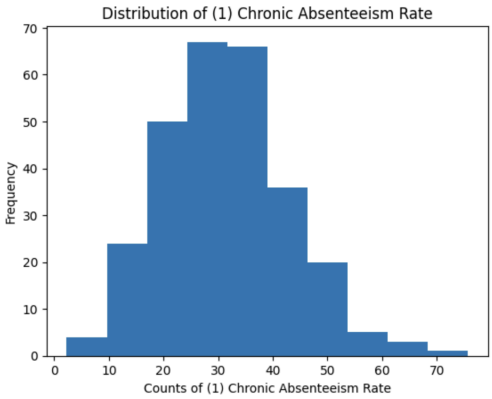
We chose these bivariate analyses because the correlation matrix gives a general idea of which variables are associated across all of the variables. This allowed us to more thoroughly get a picture of what our data looked like as a whole, seeing as we had only done univariate analysis up to that point. As for the scatter plots, we chose the variables against the busses because it is the main question for out project – seeing how the size of bus fleets affects student performance by these metrics (graduation rate, college going rate, and absenteeism). This was the first step, and not much other processing was done beyond simplifying the amount of variables we had from the first scaffold activity and changing around some data types to more easily work with the data. There will definitely be more exploration in the future, but we wanted to get a good idea of what our dataset was and familiarize ourselves with its intricacies for this first round of analysis.

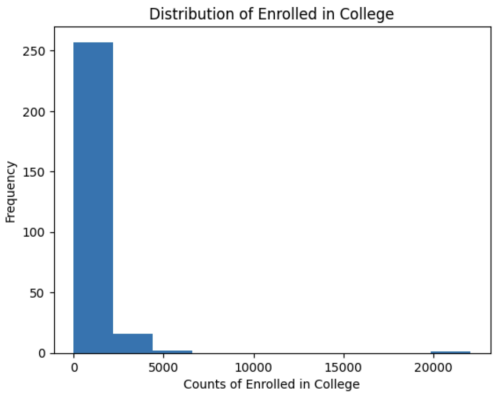
Most of the limitations seem to liee in the varying districted sizes. This skewed our data quite a bit and made our analyses a bit less generalizable. I think that It may be worth while to take the time to normalize the data based on district size in terms of how many total students that are and re-run the analyses to determine if new relationships arise as a result. Additionally, there was missing data that the team decided to exclude which I feel may be reintroduced back into the dataframe as there are no n/a values in it currently. Even if a district does not have a value for one variable, it may be useful to have the data for the variables it does have entries for.

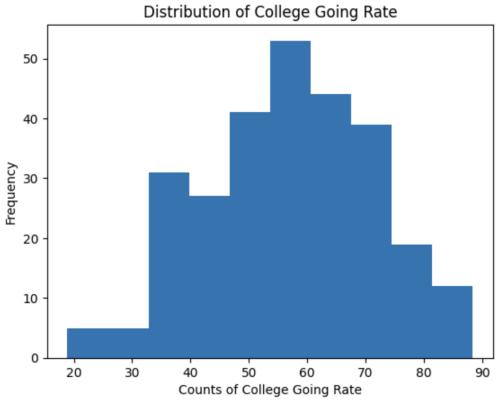
In implementing these changes, I think that the data will show slightly relationship between the bus fleets and absenteeism as the absenteeism rates do seem to be very sighly skewed right which suggests that some districts have higher absenteeism rates, particularly in the few outliers seen in the box plot. These may not be causal though, seeing it is may have to do with factors like population density, access to public transit, etc. As for the relationship between fleet size and graduation/college-going rate, I think the relationship will be fairly slight or even negligible as the college going rate seems to have a slight left-skew which may suggest that a few districts have lower rates. The graduation rate is likely highly related to college going rate, but needs to be converted to a rate to reveal any true relationship. However, I dont think these will be entirely related to the size of fleet and may better be described by access to resources, quality of teachers, socioeconomic status of the student population, etc. which will ultimately manifest as a relationship with fleet size/lower absenteeism in the data.

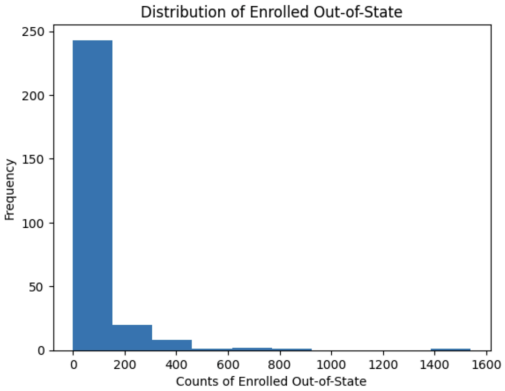

Distribution of Total number of buses
Frequency
Counts of Total number of buses

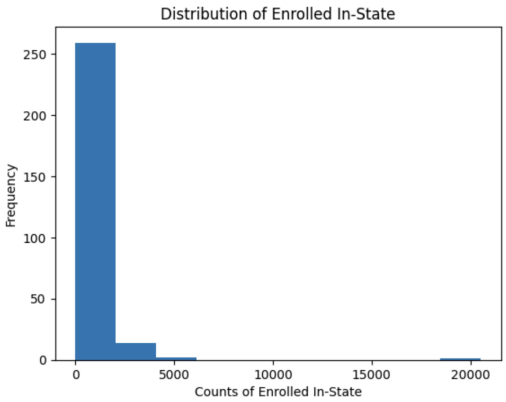

Distribution of (2) Total High School Completers
Frequency
Counts of (2) Total High School Completers

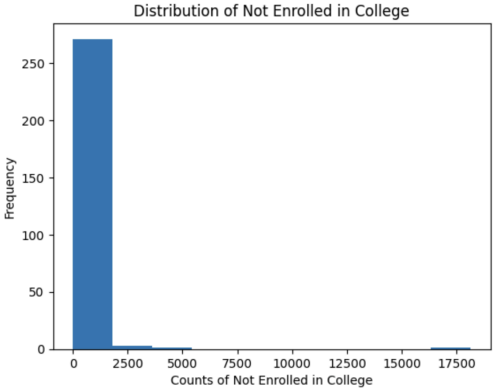

Distribution of (1) Chronic Absenteeism Rate
Frequency
Counts of (1) Chronic Absenteeism Rate

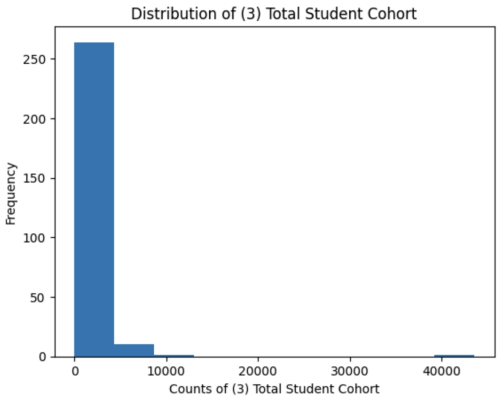

Distribution of Enrolled in College
Frequency
Counts of Enrolled in College

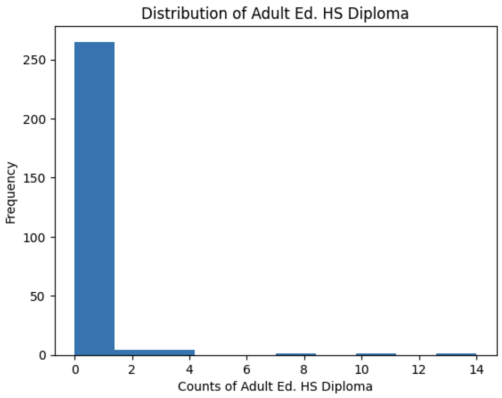

Distribution of College Going Rate
Frequency
Counts of College Going Rate

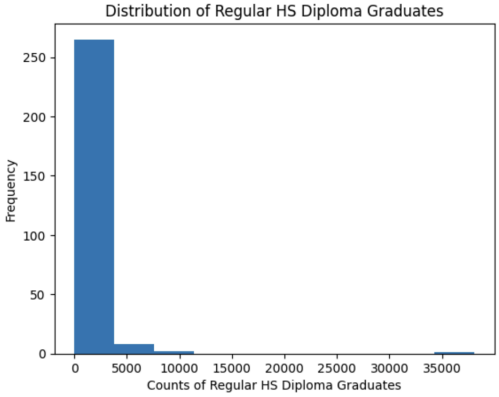

Distribution of Enrolled Out-of-State
Frequency
Counts of Enrolled Out-of-State

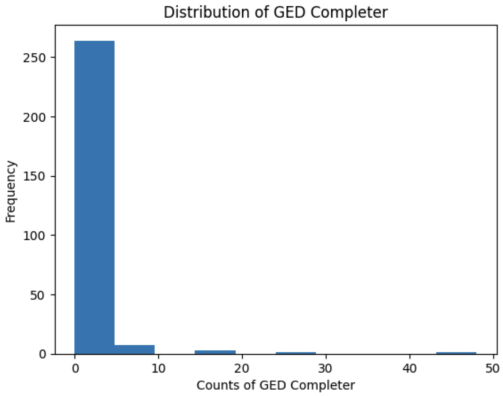

Distribution of Enrolled In-State
Frequency
Counts of Enrolled In-State

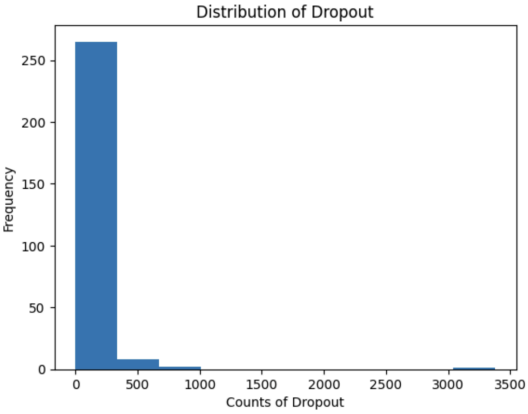

Distribution of Not Enrolled in College
Frequency
Counts of Not Enrolled in College

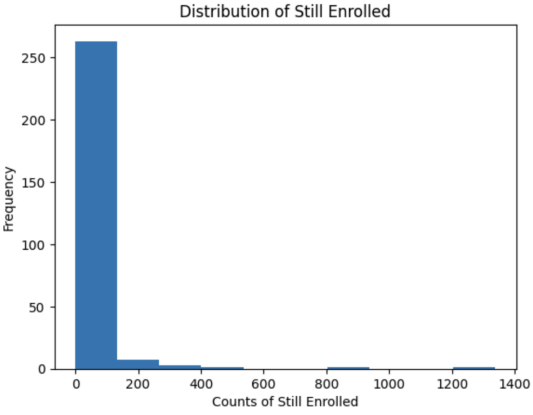

Distribution of (3) Total Student Cohort
Frequency
Counts of (3) Total Student Cohort

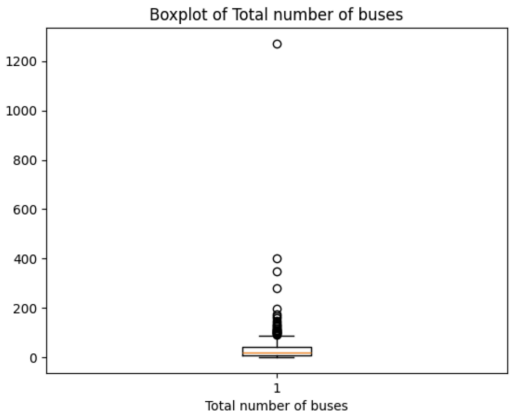

Distribution of Adult Ed. HS Diploma
Frequency
Counts of Adult Ed. HS Diploma

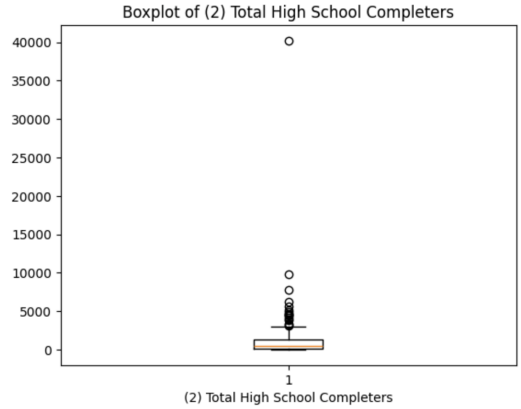

Distribution of Regular HS Diploma Graduates
Frequency
Counts of Regular HS Diploma Graduates

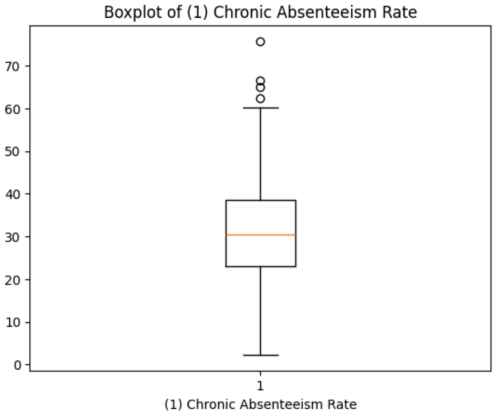

Distribution of GED Completer
Frequency
Counts of GED Completer

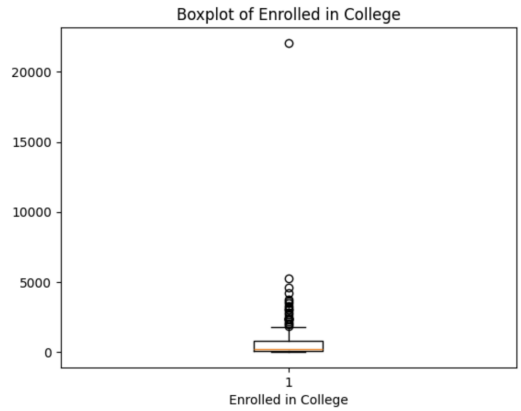

Distribution of Dropout
Frequency
Counts of Dropout

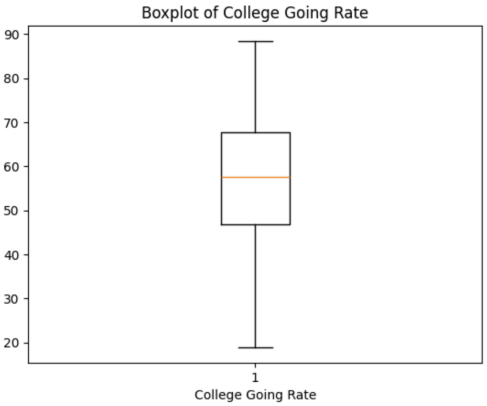

Distribution of Still Enrolled
Frequency
Counts of Still Enrolled

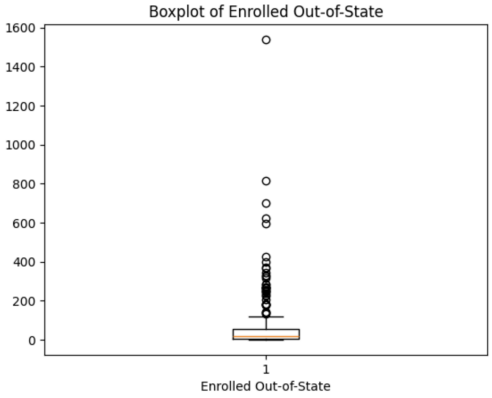

Boxplot of Total number of buses
1200
1000
800
600
400
200
0
1
Total number of buses

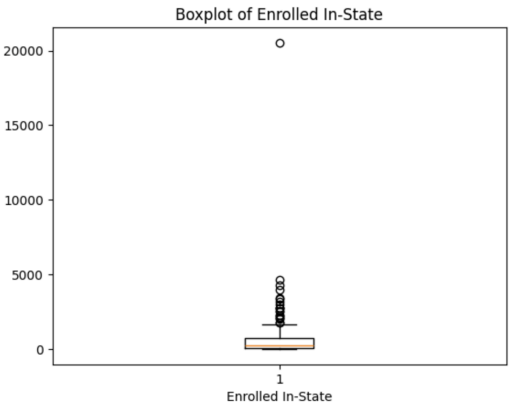

Boxplot of (2) Total High School Completers
40000
35000
30000
25000
20000
15000
10000
5000
0
1
(2) Total High School Completers

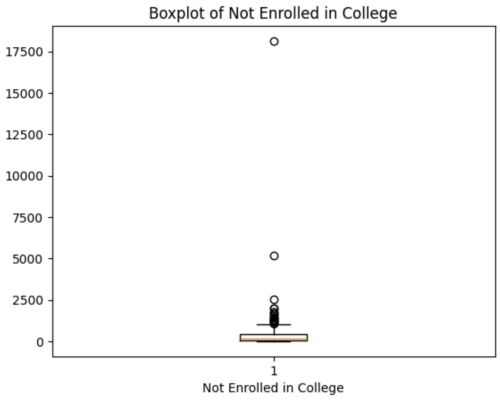

Boxplot of (1) Chronic Absenteeism Rate
70
60
50
40
30
20
10
0
1
(1) Chronic Absenteeism Rate

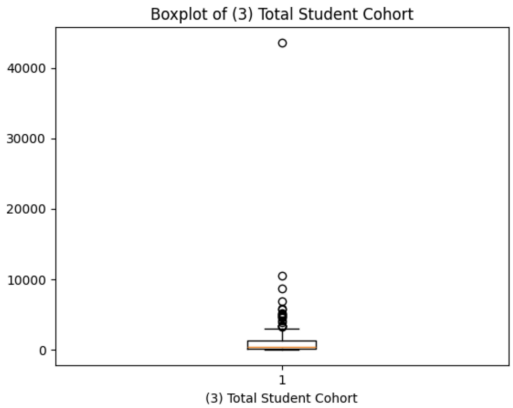

Boxplot of Enrolled in College
20000
15000
10000
5000
0
1
Enrolled in College

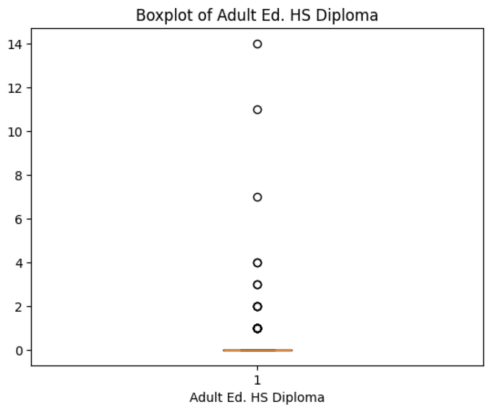

Boxplot of College Going Rate
90
80
70
60
50
40
30
20
1
College Going Rate

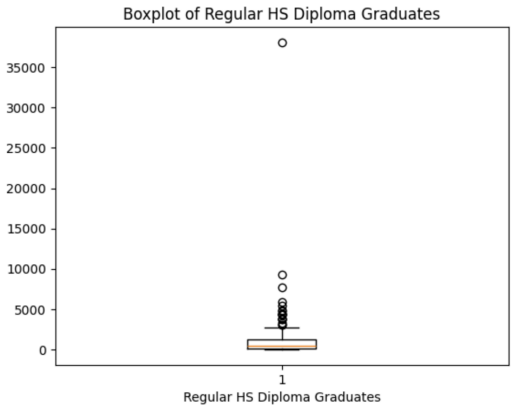

Boxplot of Enrolled Out-of-State
1600
1400
1200
1000
800
600
400
200
0
1
Enrolled Out-of-State

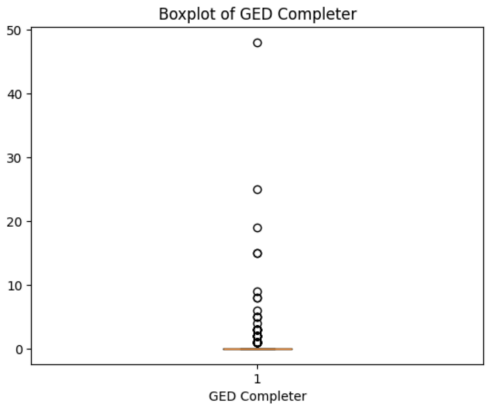

Boxplot of Enrolled In-State
20000
15000
10000
5000
0
1
Enrolled In-State

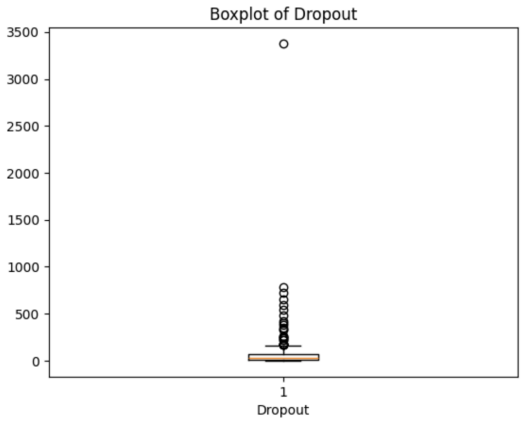

Boxplot of Not Enrolled in College
17500
15000
12500
10000
7500
5000
2500
0
1
Not Enrolled in College

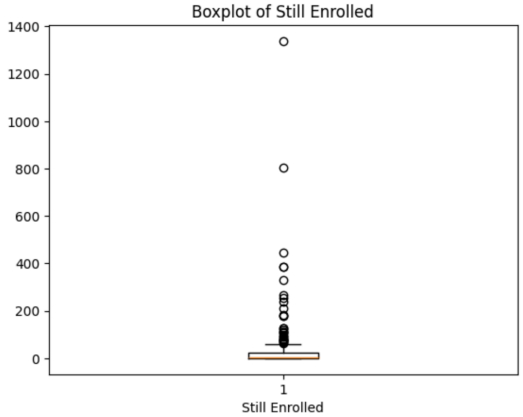

Boxplot of (3) Total Student Cohort
40000
30000
20000
10000
0
1
(3) Total Student Cohort

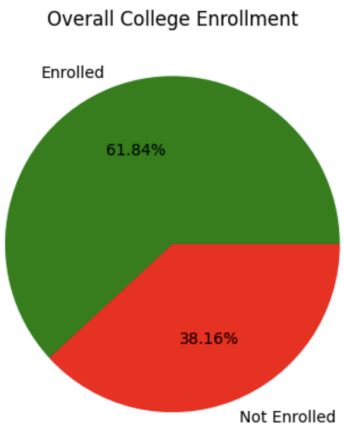

Boxplot of Adult Ed. HS Diploma
14
12
10
8
6
4
2
0
1
Adult Ed. HS Diploma

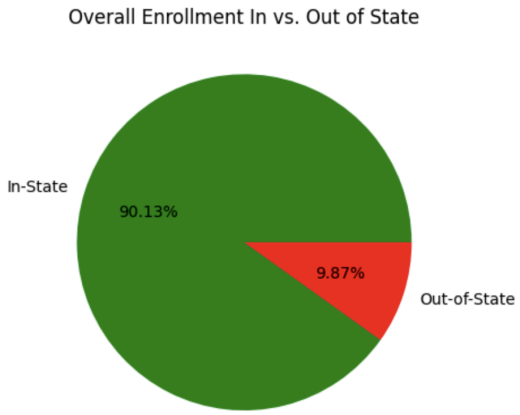

Boxplot of Regular HS Diploma Graduates
35000
30000
25000
20000
15000
10000
5000
0
1
Regular HS Diploma Graduates

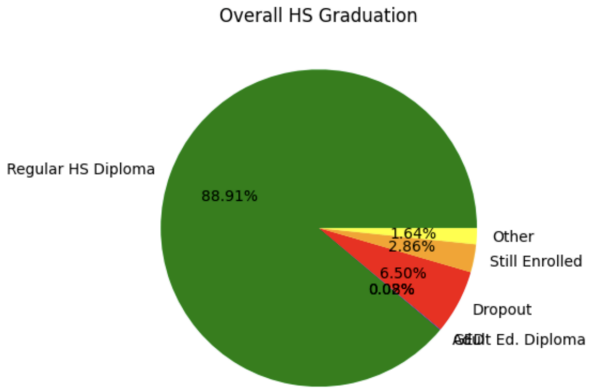

Boxplot of GED Completer
50
40
30
20
10
0
1
GED Completer


Boxplot of Dropout
3500
3000
2500
2000
1500
1000
500
0
1
Dropout


Boxplot of Still Enrolled
1400
1200
1000
800
600
400
200
0
1
Still Enrolled


Overall College Enrollment
Enrolled
61.84%
38.16%
Not Enrolled


Overall Enrollment In vs. Out of State
In-State
90.13%
9.87%
Out-of-State


Overall HS Graduation
Regular HS Diploma
88.91%
1.64%
Other
2.86%
Still Enrolled
6.50%
0.08%
Dropout
Adult Ed. Diploma

	Total number of buses	(1) Chronic Absenteeism Rate	(2) Total High School Completers	Enrolled in College	College Going Rate	Enrolled In-State	Enrolled Out-of-State	Not Enrolled in College	(3) Total Student Cohort	Regular HS Diploma Graduates	Adult Ed. HS Diploma	GED Completer	Dropout	Still Enrolled
Total number of buses	1.000000	0.056479	0.925159	0.906318	0.091206	0.907692	0.749850	0.923058	0.925784	0.926138	0.163924	0.478403	0.870927	0.735539
(1) Chronic Absenteeism Rate	0.056479	1.000000	-0.016865	-0.059270	-0.372673	-0.045837	-0.184004	0.039131	-0.010924	-0.020937	-0.024561	-0.066225	0.072582	0.047715
(2) Total High School Completers	0.925159	-0.016865	1.000000	0.990447	0.138795	0.992637	0.812617	0.983575	0.999646	0.999512	0.098514	0.450181	0.943844	0.809213
Enrolled in College	0.906318	-0.059270	0.990447	1.000000	0.221002	0.998656	0.855957	0.949290	0.988800	0.991778	0.114930	0.480862	0.908668	0.778178
College Going Rate	0.091206	-0.372673	0.138795	0.221002	1.000000	0.204929	0.346699	0.027172	0.128629	0.148804	-0.034407	0.116525	-0.015075	-0.076572
Enrolled In-State	0.907692	-0.045837	0.992637	0.998656	0.204929	1.000000	0.828009	0.956042	0.991353	0.993688	0.109851	0.469890	0.916572	0.783386
Enrolled Out-of-State	0.749850	-0.184004	0.812617	0.855957	0.346699	0.828009	1.000000	0.732376	0.807608	0.816532	0.147600	0.514703	0.686665	0.603654
Not Enrolled in College	0.923058	0.039131	0.983575	0.949290	0.027172	0.956042	0.732376	1.000000	0.984924	0.980720	0.074176	0.396995	0.962584	0.826427
(3) Total Student Cohort	0.925784	-0.010924	0.999646	0.988800	0.128629	0.991353	0.807608	0.984924	1.000000	0.998882	0.102505	0.454149	0.949872	0.819637
Regular HS Diploma Graduates	0.926138	-0.020937	0.999512	0.991778	0.148804	0.993688	0.816532	0.980720	0.998882	1.000000	0.102141	0.452489	0.935594	0.794902
Adult Ed. HS Diploma	0.163924	-0.024561	0.098514	0.114930	-0.034407	0.109851	0.147600	0.074176	0.102505	0.102141	1.000000	0.488627	0.058656	0.118711
GED Completer	0.478403	-0.066225	0.450181	0.480862	0.116525	0.469890	0.514703	0.396995	0.454149	0.452489	0.488627	1.000000	0.391695	0.435993
Dropout	0.870927	0.072582	0.943844	0.908668	-0.015075	0.916572	0.686665	0.962584	0.949872	0.935594	0.058656	0.391695	1.000000	0.885067
Still Enrolled	0.735539	0.047715	0.809213	0.778178	-0.076572	0.783386	0.603654	0.826427	0.819637	0.794902	0.118711	0.435993	0.885067	1.000000