Do Wealthier High Schools Have More Sport Teams?

American Statistical Association (ASA) Statistics Project Competition

May 2019

I. Introduction

Dubbed Operation Varsity Blues, the college admissions scandal that broke out in March of 2019 garnered national attention and caused public indignation. Fifty parents were charged with cheating the college admissions process by increasing their children's chances of acceptance to elite colleges through means of falsifying athletic credentials and standardized test scores. While the news sparked heated discussions at my high school, it also piqued my interest in investigating the inequalities between wealth of schools and extracurricular opportunities offered.

Since the admission scandal involved multiple cases of forged athletic ability, this observational study will research the relationship between the wealth of schools and the number of athletic teams offered. The scope of the study is limited to public high schools in Minnesota. Public high schools are funded by federal and state money and property taxes from local governments ("Financing Education in Minnesota"). Property tax is derived from the assessed home value. Therefore, the wealth of a public school is related to its surrounding home values. This study attempts to examine the association between the median home value of a public high school's location and the number of athletic teams that school offers.

II. Statistical Question

In Minnesota, is there a relationship between the median home value at a public high school's location and the number of athletic teams?

III. Data Collection and Sampling Methodology

The name of every public high school and its enrollment are collected from the Minnesota Department of Education's Data Center. Charter schools and non-conventional public

schools are excluded because their funding is not directly associated with the property tax, leaving a population of 301 public high schools ("A Primer on Minnesota Charter Schools").

I decided to use stratified random sampling since its samples give more precise estimates than simple random samples of the same size if the strata are chosen wisely. First, I considered dividing all schools into three subgroups as large, medium, or small by using enrollment. The stratum should contain schools that share a common characteristic thought to be associated with variables being measured, namely median home value but not enrollment. Since there is a considerable difference in median home value between city and countryside, a high school's community environment is more suited to be the strata criterion. Therefore, each school is classified into one of three strata (urban, suburban, or rural) based on its location. Each stratum is assigned proportional representation in the sample size of thirty (n =30) depending on its share of total enrollment as seen in Table 1. For instance, rural schools enroll 95,404 students or 40.3% of total enrollment at public schools, so $30 \times 40.3\% = 12$ rural schools are chosen for the sample.

Next, random samples are chosen from each stratum. To randomly choose which high schools are included in the sample, each high school is numbered. A random number generator selects the appropriate amount of schools in each stratum, ignoring repeats.

Stratum	# of Schools	Enrollment	% of Total	# of Samples
Rural	214	95,404	40.3	12
Suburb	60	103,877	43.9	13
Urban	27	37,564	15.9	5
Total	301	236,845	100	30

Table 1: Stratification of all public high schools by community environment

To gather data on the number of athletic teams at the selected high schools, I visited their school websites and student handbooks, recording all the athletics listed. In attempt to maintain consistency, all levels of athletics available (e.g. junior varsity, adapted sports, and recreational sports) are included in the count.

To gather the median home values, I first considered using the 2010 U.S. Census. This appeared outdated because it was administered nine years ago. Thereafter, I used Zillow's home value index, which gives an estimated median home value in a certain geographic region. The zip code of the high school's location was entered for geographic area (see an example in Figure 1).

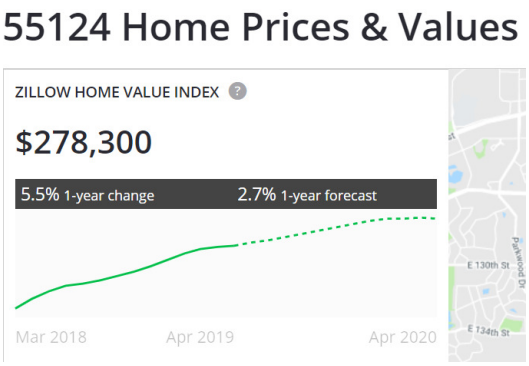


Figure 1: Screenshot of Apple Valley Senior High School's median home value, zip code 55124

IV. Linear Regression Significance Test

Hypotheses:

Let β_1 = the slope of the regression line relating y = *number of athletic teams* to x = *median home value (in thousands $).*

H_0: $\beta_1 = 0$

H_a: $\beta_1 \neq 0$

$\alpha = 0.05$

Conditions:

1. Linear – In Figure 2, the scatterplot of *number of athletic teams* vs. *median home value (in thousands $)* shows a linear pattern by the Straight Enough Condition. Also, the residual plot in Figure 3.1 shows no curved patterns.

2. Independent – Sampling is done without replacement. The sample size of 30 < 10% of all public high schools in this study (301).

3. Normal – There is no skewness. Though the dotplot of residuals (Figure 3.3) appears to have a low outlier, there are no outliers as shown by the boxplot (Figure 3.2). A low outlier is defined as less than $Q1 - 1.5 \times IQR = -7.43 - 1.5 \times (15.28) = -30.35$, which is smaller than the minimum residual value (-27.33).

4. Equal SD – The residual plot (Figure 3.1) has no pattern and shows a fairly equal amount of scatter around the "residual = 0" line for all values of x.

5. Random – Stratified random sampling method is used to select the schools for the sample.

Calculations:

$$t = \frac{b_1 - \beta_1}{SE_b} = \frac{0.1015464 - 0}{0.014716} = 6.900$$

$$P - value = 2 \times (8.421 \times 10^{-8}) = 1.684 \times 10^{-7}$$

Conclusion:

The P-value < 0.05, so we reject the null hypothesis. There is convincing evidence to conclude a linear relationship between *median home value (in thousands $)* and *number of athletic teams* at public high schools in Minnesota. The true slope of the regression line is positive.

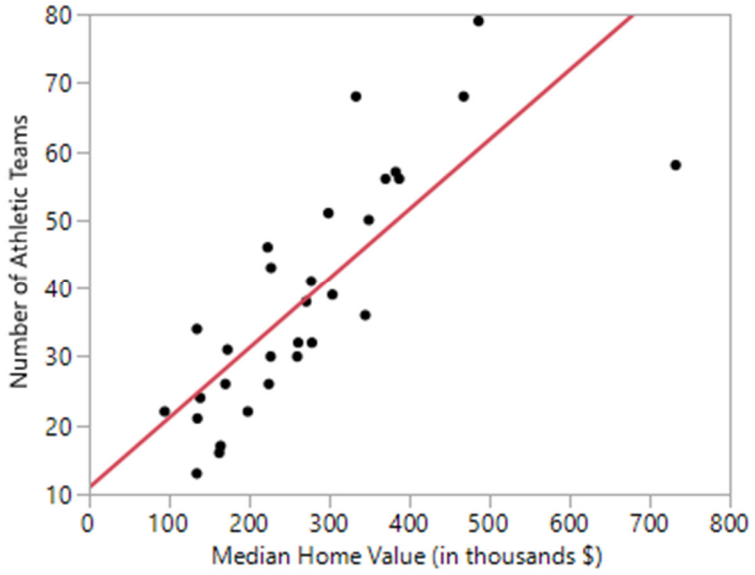


Figure 2: Scatterplot of *Number of Athletic Teams* vs. *Median Home Value (in thousands $)* with least-squares regression line

Summary of Fit

RSquare	0.629703
RSquare Adj	0.616478
Root Mean Square Error	10.53353
Mean of Response	38.73333
Observations (or Sum Wgts)	30

Linear Fit

# of Athletic Teams = 10.935347 + 0.1015464*Median Home Value (in thousands)

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	10.935347	4.46399	2.45	0.0208*
Median Home Value (in thousands)	0.1015464	0.014716	6.90	<.0001*

Table 2

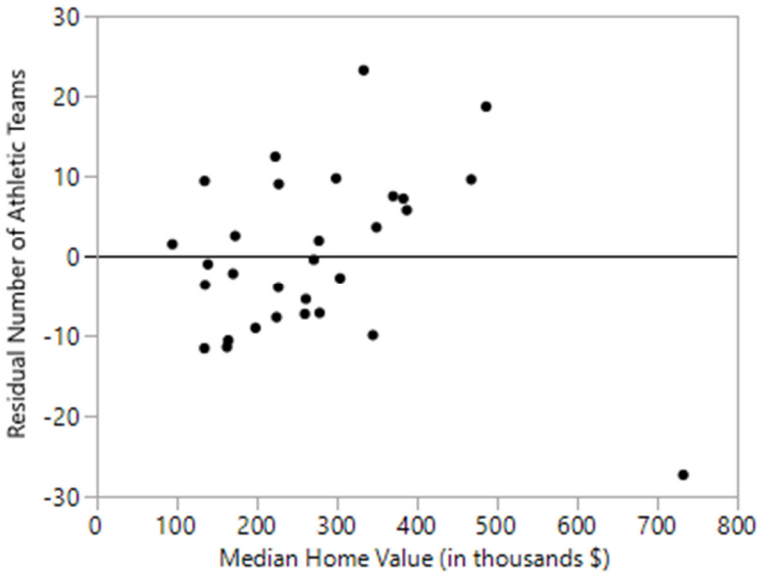


Figure 3.1: Residual plot for the relationship between *Number of Athletic Teams* and *Median Home Value (in thousands $)*

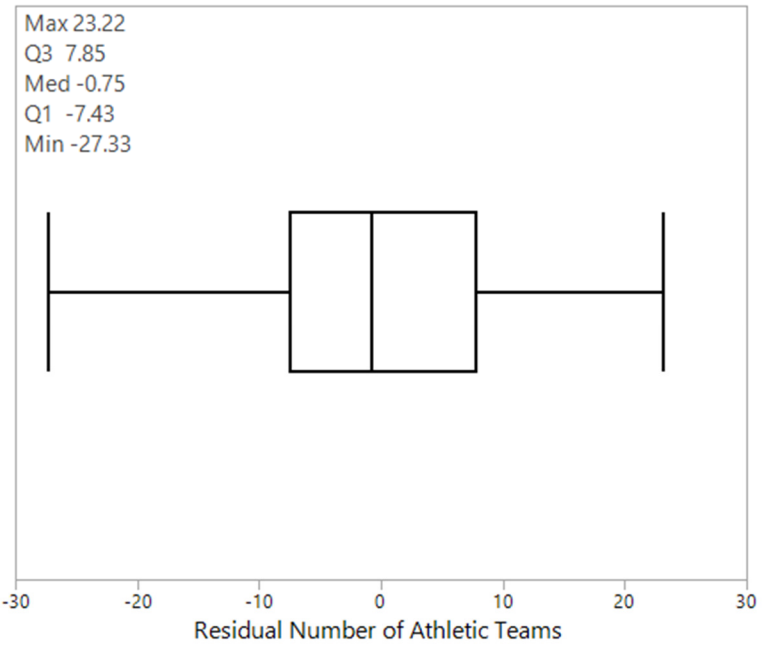


Figure 3.2: Boxplot of residuals with of Number of Athletic Teams with 5 number summary

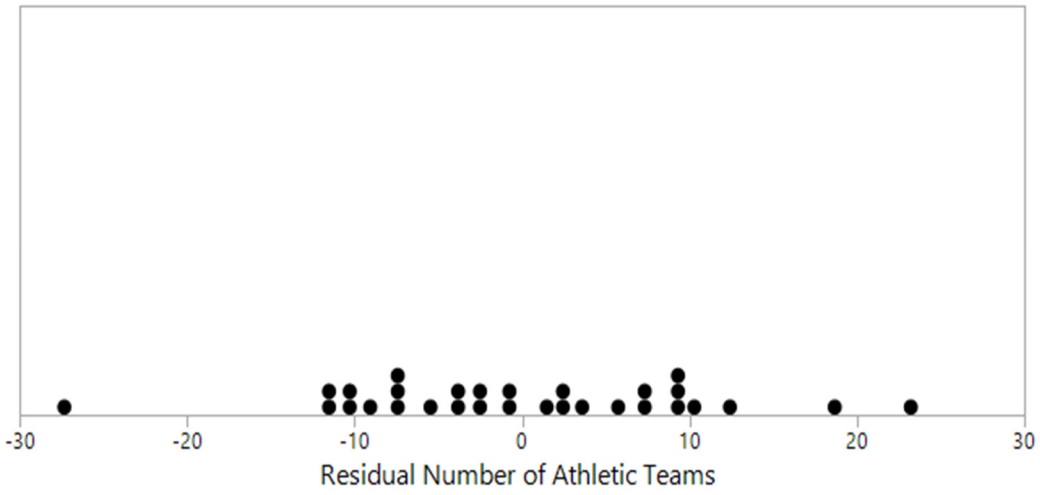


Figure 3.3: Dotplot of residuals

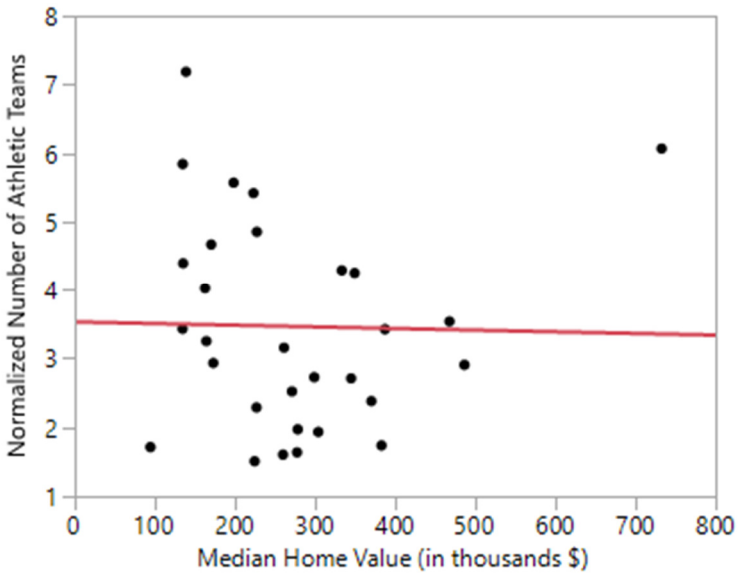
V. Graphical Analysis of Normalized Scatterplot

From the above linear regression significance test (r value: 0.79), there appears to be a positive strong correlation between the number of athletic teams and the median home value of the school location.

Now, we will investigate whether enrollment size affects this correlation. I observed that the larger the high school's enrollment, the more athletic teams it has. To eliminate the impact of enrollment, I normalized the number of athletic teams as follows:

$$Normalized\ number\ of\ Athletic\ teams = \frac{Number\ of\ Atheltic\ Teams}{Enrollment} \times 100$$

The scatterplot (see Figure 4), with r = 0.074, shows a very weak correlation between *number of athletic teams per 100 students* and *median home value*. This reveals the existence of a lurking variable: school enrollment size.



Summary of Fit	
RSquare	0.000439
RSquare Adj	-0.03526
Root Mean Square Error	1.545398
Mean of Response	3.464448
Observations (or Sum Wgts)	30

Figure 4: Scatterplot of *Normalized Number of Athletic Teams* (per 100 students) *vs. Median Home Value*

VI. Conclusion

This study investigates whether there is a relationship between the number of athletic teams in a public high school and the median home value around that school in Minnesota. After performing a significance test for linear regression, I initially found convincing evidence of a linear relationship between the number of athletic teams and the median home value: the higher the median home value, meaning the wealthier the school, the more the athletic teams. However, normalizing the number of athletic teams by enrollment size revealed no correlation between the number of athletic teams per 100 students and the median home value. The lurking variable, enrollment size, therefore impacted the correlation between the number of athletic teams and the median home value found by the significance test.

VII. Error Analyses and Reflection

For some schools in the sample, the number of athletic teams may be underestimated. I collected data by visiting the school's website and included all levels of the sport in my count.

However, some schools did not specify whether there are more than one team per sport. For instance, the school may have both varsity and junior varsity teams for football, but the website only lists football. Though I tried calling some schools, there were still no firm counts. With more time, the study could be reperformed with more accurate data by contacting coaches directly for information on the number of athletic teams.

Additionally, Zillow's home value estimate relies on the availability of data in the area, so rural areas with less detailed information may have less accurate estimates. It would be more accurate if I had access to the median assessed home value of each geographic area.

VIII. Reference

[1] *A Primer on Minnesota Charter Schools*. MN Association of Charter Schools, Jan. 2017,

[2] www.mncharterschools.org/_uls/resources/A_Primer_on_Minnesota_Charter_Schools.pdf.

IX. Raw Data

High School	Community Environment	Enrollment	# of Athletic Teams	Median Home Value ($)
Southwest Senior High	Urban	1923	68	467700
Kennedy Senior High	Urban	1507	38	271100
Highland Park Senior High	Urban	1328	36	344900
Henry Senior High	Urban	1057	31	172900
Roosevelt Senior High	Urban	1014	32	261200
Minnetonka Senior High	Suburb	3280	57	382900
Burnsville High School	Suburb	2503	41	277400
Maple Grove Senior High	Suburb	2352	56	370200
Edina Senior High	Suburb	2718	79	486400
Eagan Senior High	Suburb	2016	39	303900
Mounds View Senior High	Suburb	1870	51	298900
Park Senior High	Suburb	1869	30	259900

Chanhassen High School	Suburb	1636	56	387200
Apple Valley Senior High	Suburb	1623	32	278300
Spring Lake Park Senior High	Suburb	1723	26	224500
Rogers Senior High	Suburb	1585	68	333300
Orono Senior High	Suburb	956	58	732600
Fridley Senior High	Suburb	886	43	227200
Mahtomedi Senior High	Rural	1176	50	349300
New Prague Senior High	Rural	1311	30	226800
Albert Lea Senior High	Rural	1283	22	94400
North Branch Senior High	Rural	849	46	223000
Lincoln Senior High	Rural	582	34	134700
Albany Area High School	Rural	523	17	164200
Litchfield Senior High	Rural	478	21	135200
Lewiston-Altura Secondary	Rural	398	16	162600
A.C.G.C. Secondary	Rural	379	13	134500
Aitkin Secondary School	Rural	557	26	170200
Blooming Prairie Secondary	Rural	334	24	138800
Park Rapids Senior High	Rural	395	22	198200