Understanding Critical Factors of Minority and Transferred Students in U.S. Schools

Cathleen Li, Carol Li, Jinxiang Ma, and Jiayi Shen.

Author contributions

Cathleen and Carol contributed question 1, background, and dataset.

Jinxiang Ma and Jiayi Shen contributed question 2.

Abstract, aims, methods, and discussion are written by all four authors.

Abstract

- In this project, we will explore how minority and transferred students affect school environment such as racial tension, violent incidents, and also how school location influence transferred students rate. We will also explore how urbanacity,violent incidents, school size and crime level influence number of transferred students;
- The aims of the project are to answer if the presence of minority students aggravates racial tension. Also, we aim to explore possible factors that influence number of transferred students.
- From the analysis, we have found that the presence of minority students both aggravate the level of racial tension and increase violent incidents. The higher the minority percentage is, the larger the amount of violent incidents occur and the larger the higher level racial tensions (level 5, 8, 10) is. We also conclude that generally the increase of violent influences means the increase in number of transferred students. Also, urbanicty will lead to more transferred studnets: with larger school sizes in city, more transferred students we have; smaller school sizes in town, less transferred students we have. Moreover, we also find that an association between crime level and number of transferred students.

---

1. Introduction

1.1 Background

The National Education Statistics Center (NCES) is the main federal entity that collects and analyzes education-related data in the United States and other countries. NCES performs the duties of Congress to collect, organize, analyze, and report complete statistical data on the state of education in the United States; conduct and publish reports; review and report on international education activities.

The data our group use for this project is called the School Survey on Crime and Safety (SSOCS), which is the main source of NCES school-level crime and security data. SSOCS is a sample survey of approximately 4,800 public primary and secondary schools. It aims to provide estimates of school crime, discipline, disorder, programs and policies, and is large enough to provide a national estimate of all public schools. SSOCS is administered in the spring of the school year so that the principal can report the most complete information possible. Specifically, we use the data set from the 2005-06 school year. In the raw data, the SSOCS questionnaire requires principals to report on various topics related to crime and safety, including school practices and plans, parent and community involvement in schools, school security personnel, and so on.

In this project, two topics we plan to explore are:

1. Does the presence of minority students aggravate racial tension?
2. What factors can affect the number of transfer students?

1.2 Aims

The first question aims to answer if the presence of minority students aggravates racial tension. Since racial tension has always been an issue in the society, and the tragic death of George Floyd really provoked people into the "Black Lives Matter" movement. Moreover, since the pandemic, "Asian hate crime" has been happening quite frequently. Since school violence is a topic that many people concern about now, we feel like it is important to see if there is a relationship between the presence of minority students and racial tension.

The second question aims to explore the relationship between variables: urbanicity, number of transferred students, school sizses, and crime rates. Since we believe that these variables are important signs of whether a school can provide a safe and good environment for students, we then uses statistical methods: simple linear regression, mutiple linear regression to study the dataset. We are intersted to see whether crime level, school size, and urbanicity impact a students' transfer decision or not. Through the progress, we also look at other variables like school sizes and urbanicity altogether.

2. Materials and methods

2.1 Datasets

Basic information

General description:

The data are school survey on crime and safety that gathered responses from approximately 4,800 public elementary and secondary schools from 2005–06 school year. The survey is administrated during the spring of the school year for the principals to report the most complete information possible.

Source:

We use school-level data on crime and safety for the U.S. Department of Education, National Center for Education Statistics (NCES).

https://nces.ed.gov/surveys/ssocs/index.asp?FType=1 (https://nces.ed.gov/surveys/ssocs/index.asp?FType=1)

Collection methods:

The data are collected from the surveys distributed to approximately 4,800 public elementary and secondary school principals in the U.S. during 2005–06 school year. School principals are in charge of filling out the survey each year.

Sampling design and scope of inference:

The relevant population is all U.S. elementary and secondary schools. The sampling frame is all U.S. elementary and secondary school principals recorded. The sampling mechanism is random sample since all school principals have equal likelihood of being selected to participate in this survey. The scope of inference is broad since the sample is said to be representative of the population. The sample could be used to safely extrapolate all U.S. elementary and secondary schools' crime and safety situations.

Data semantics and structure

Units and observations:

Observational units are distinct 4800 public elementary and secondary schools from 2005 to 2006 school year.

Variable descriptions:

Provide a table of variable descriptions. If your dataset is large and you'll only work with a subset of the total available variables, limit your attention to the variables that you'll work with. Here's a template you can work with:

Name	Variable description	Type	Units of measurement
C0562	Crime where school located	catogory	No
C0570	# of students transferred to school	Numeric	number of students
C0374	How often student racial tensions	catogory	No
VIOINC06	Total number of violent incidents recorded	Numeric	number of incidents
FR_LOC4	Urbanicity - from 03-04 CCD (School)	catogory	No
FR_CATMN	Recoded % minority student enrollment in school - based on 03-04 CCD frame variables (School)	catogory	No
FR_SIZE	School size categories - based on 03-04 CCD frame variables (School)	Numeric	school sizes

2.2 Methods

The first question studies the relationship between minority and racial tension via comparing the percentage of minority, racial tensions, and the number of violent incidents happened. Changes in racial tensions are identified for different level of minority percentage as well as changes in the number of violent incidents, derived from several line and scatter plots. Furthermore, under the bar chart, there is strong signal in the percentage of minority about each racial tensions, based on sample sizes of each racial tensions in each level of minority percentage.

The second question studies the relationship between number of transferred students and number of violent incidents via simple linear regression. We also explored the relationship between the number of transferred students and other catogorical variables: school sizes and urbanicity through mutiple linear regession. Specifically, we explore the data first by scatterplots and tables to analyze their relationships. Further, we use formula in simple linear regression (SLR) and mutiple linear regression (MLR) to explore their coefficeints, stardard errors.....

3. Results

3.1 Minority and Disharmony

Main Question: Does the presence of minority students aggravate racial tension and increase violent incidents?

3.1.1 Minority and Racial Tensions

In this section, we mainly focus on analyzing if there is a relationship between minority percentage and the level of racial tensions.

Racial tensions	count
on occassion	1626
Never	838
at least once a month	137
at least once a week	97
daily	26

Table 1: Distribution of response to the racial tensions choice

From the table above, we can see there are 5 levels of racial tensions recorded in this survey, and the choice "on occassion" is the most popular one. "Never" is the second most popular choice, and "daily" is the least popular.

We assign numerical numbers to different levels of racial tensions depending on how often tensions occur. We think it would be reasonable to assign "never" to 0, "at least once a month" to 3, "at least once a week" to 8, and "daily" to 10.

Percent minority	Racial tensions mean	Racial tensions min	Racial tensions max
20% - 50%	2.91831	0	10
5% - 20%	2.57064	0	10
50% +	2.4831	0	10
<5%	1.74292	0	10
Do Not Know	1.96053	0	8

Table 2: Racial tensions mean, max and min for each minority density group

From the table above, we can see all minority density groups have the same racial tensions min and max except "Do not know" group that only has 8 as their maximum tension level. All groups seem to have different racial tensions mean within the range of 1.74 to 3.

In order to show the racial tensions mean for each group more visually, we decide to display a plot for it.

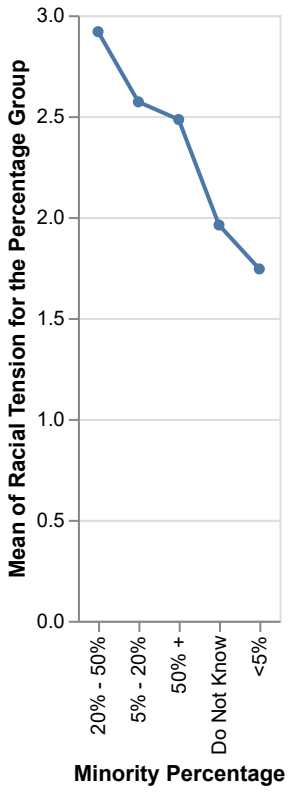

Figure 1: line plot of minority percentage(x axis) and mean of racial tensions (y axis) in decreasing order.

Based on the table, we have several findings:

1. The highest mean of racial tensions happens among minority percentage 20% to 50%, and the second highest is 5% to 20%.
2. The lowest mean happens among <5% percentage.
3. Surprisingly, the highest minority density of 50%+ rank only the 3rd for mean racial tension.

Therefore, minority percentage does not seem to affect racial tensions if we consider the degrees of racial tensions through the their means. However, it would be biased to only consider racial tension means since there may be extreme highs and lows that drive the mean. Therefore, we will analysis this racial disharmony though another factor-- violence incidents.

3.1.2 Minority and Violence Incidents

In this section, we mainly focus on the relationship between the minority percentage and the number of violent incidents. Importantly, we neglect the group of unknown percentage of minority, because it contains data from different level of minority percentage, whose analysis has no reference value. Table2 is the summary of violent incidents grouped by the minority percentage.

Percent_minority	('Violent_incidents', 'min')	('Violent_incidents', 'max')	('Violent_incidents', 'mean')	('Violent_incidents', 'std')
20% - 50%	0	258	28.0091	35.769
5% - 20%	0	188	19.5967	23.8851
50% +	0	588	41.1414	55.6408
<5%	0	240	14.5033	20.9288
Do Not Know	0	487	30.2237	61.3607

Table 2: The summary of violent incidents for each minority percentage, including minimums, maximums, averages, and standard deviations.

Based on the table, we have several findings:

1. For all level of percent minority, the minimum of violent incidents is 0;
2. The maximum of violent incidents is highest for more than 50 percents minority, and that is lowest for 5 to 20 percents minority;
3. The average violent incidents is highest for more than 50 percents minority, and that is lowest for less than5 percents minority;
4. The standard deviation is lowest for less than 5 percents minority.

Next, we use a line plot to show the trend. Figure2 was performed on the minority percentage and the number of violent incidents to identify whether there is any relationship between these two variables.

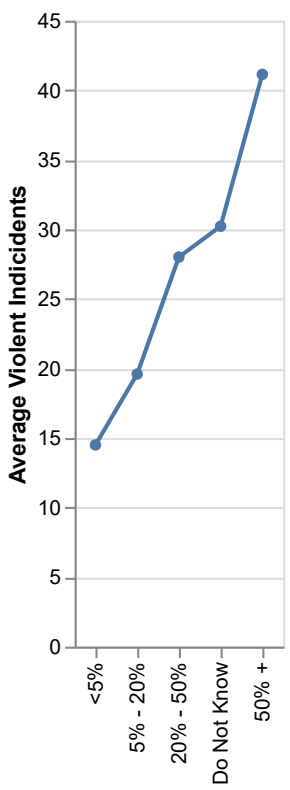

Figure 2: line plot of minority percentage(x axis) and average number of violent incidents (y axis) in ascending order.

The trend of average number of violent incidents is noticeably, that as more average number of incidents happened, the percentage of minority increases as well.

Last, we use a scatter plot to show the distribution. Figure3 was performed on the minority percentage and the number of violent incidents to identify how number of violent incidents distributes for each level of minority percentage.

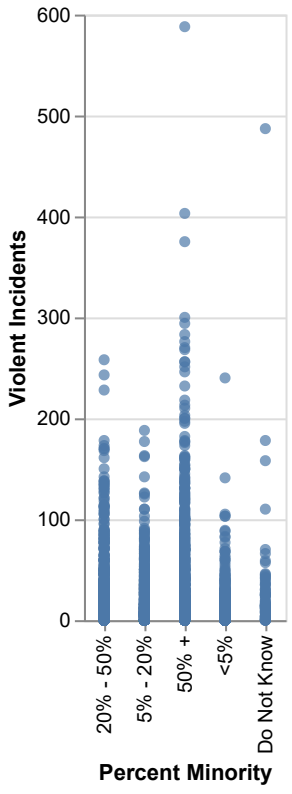

Figure 3: scatter plot of minority percentage(x axis) and number of violent incidents (y axis).

The distribution of number of violent incidents is largest for more than 50 percent minority, which mainly focuses from 0 to 300; and the distribution is the smallest for less than 5 percent minority, which is from 0 to 100 (except the "Do Not Know" group). Moreover, it is significant that the range of distribution increases as the percentage of minority increases.

Therefore, we could conclude that minority percentage has positive relationship with number of violent incidents, that as number of violent incidents increases, minority percentage increases. At this point, we wonder if the relationship between minority percentage and violence incidents has any connection to the relationship between minority percentage and the racial tension levels.

3.1.3 Minority, Racial Tensions, and Violence Incidents

In this section, we focus on the relationship between the minority percentage rate, racial tension levels, and the number of violent incidents and seek to find the relationship between these three variables if there is any. We think the best way to determine the relationship is through visualization.

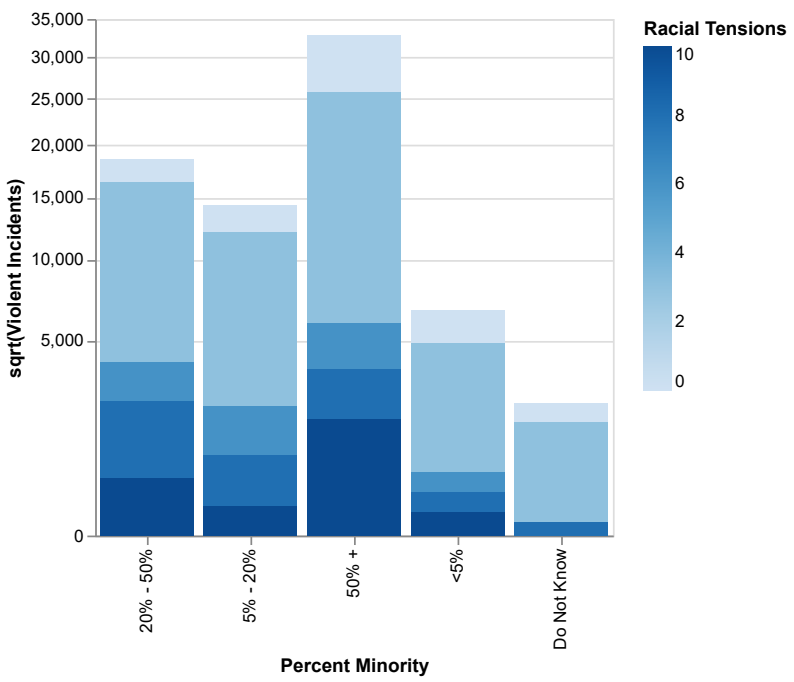

Figure 5: bar plot of minority percentage (x axis) and square root of total number of violent incidents (y axis), colored by racial tension levels.

From the bar plot above, we could see 50%+ minority percentage has the highest violent incidents and racial tensions. Surprisingly, it also has the largest racial tension level 0 (never). For the other minority groups, the amount of violent incidents seem to the second lowest, and the racial tensions levels seem to be much smaller than other groups except the "Do Not Know" group. "Do Not Know" group is the only group that does not have racial tension level 10 (daily). Most of violent incidents has 3 or 5 as their racial tension levels.

This is an interesting discovery because from this graph, we could conclude that minority percentage influences both the amount of violent incidents and racial tensions. The higher the minority percentage is, the larger the amount of violent incidents occur. Moreover, the higher the minority percentage is, the larger the higher level racial tensions (level 5,8,10) might occur. Because of the lack of information, the "Do Not Know" group reports the smallest amount of violent incidents and does not include any racial tension 10 (daily).

3.2 Transfer Students

In this section, we want to explore possible factors that affacts the number of transfer students.

3.2.1 Exploratory analysis

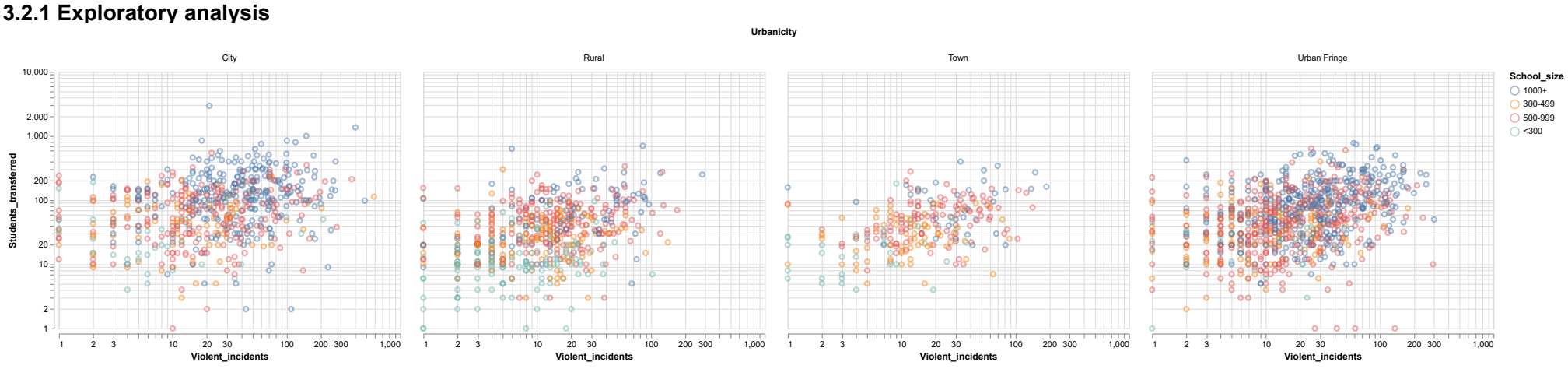

Figure 1: Distribution between Students transfer and violent incidents

From the above scatterplots, we first explore variable of violent incidents against response variable students_transferred. The plot is colored by school_size. Also, the plots are facet by urbanicity. Thus, we can observe the relation between violent incidents and students_transferred by urbanicity and school sizes.

1. We observe that, in general, with more violent incidents in x axis, there are more transferred students in y-axis considering school sizes and urbanicity.
2. For example, in school sizes 1000+, we can see blue points are increase in a linear form in city and urban fringe. Rural and town do not have a lot school sizes with siz "1000+"; however, we can still find out that blue points in rural scatterplot and town scatterplot increase in a linear form either.
3. The exploratory analysis give us a preview of what the relationship between violent incidents and transferred students look like ;

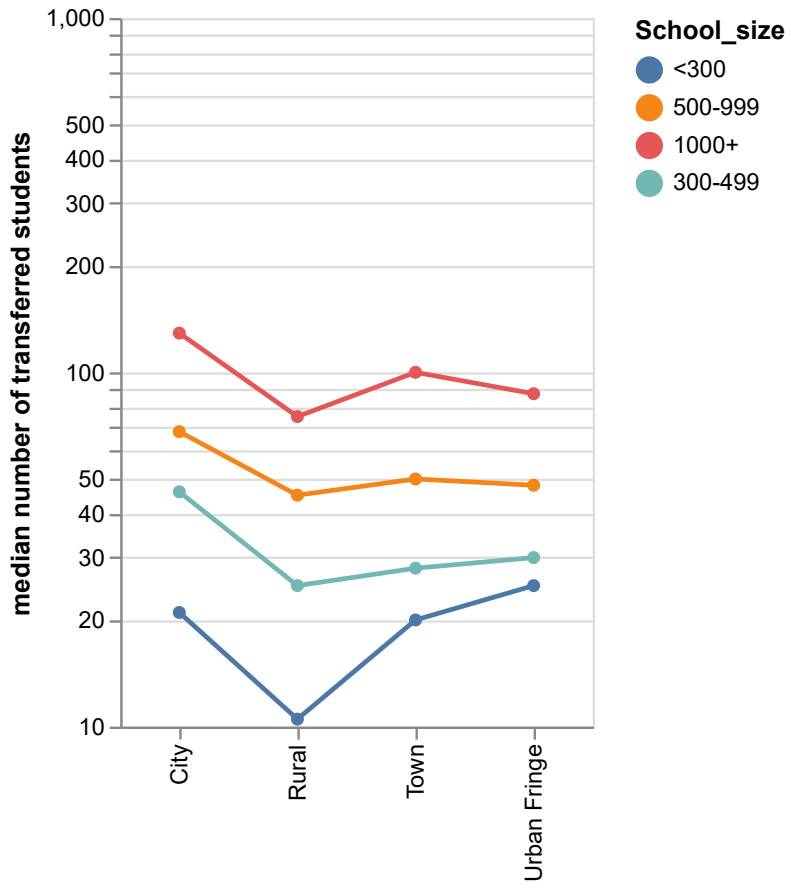

Figure 2: points and line plot between urbanicity, school size and median number of transferred students

From the above line and point plot, we explore the relationship between urbanicity and median number of transferred students. The plot is colored by school sizes.

1. we can generally observe that in rural areas, for each category of school sizes, rural areas have least number of median number of transferred students;
2. On the other side, we can generally observe that in city, for each category of school sizes, city areas have largest number of median number of transferred students;

3.2.2 Simple Linear Regression

We want to use simple linear regression to determine whether the number of transferred students relates to the number of violent incidents. The following equation represents the simple regresion model.

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i \quad \begin{cases} i = 1, \dots, n \\ \epsilon_i \sim N\left(0, \sigma^2\right) \end{cases}$$

- y_i is the ***response variable*** , in this model, it repersents "the number of transfer students".
- x_i is the ***explanatory variable***, in this model, it repersents"the number of violent incidents".
- ϵ_i is the ***error***
- β_0 is the ***intercept***
- β_1 is the ***coefficient***
- σ^2 is the ***error variance***

First, we focus on computing the estimates. This is accomplished by using a `LinearRegression()` function and we fit the model using the `fit()` method. After that, we retrieve the coefficient estimates using `.coef_`

- Intercept estimate: $\hat{\beta}_0 = 59.27765918$
- Slope estimate: $\hat{\beta}_1 = 0.77554596$

The slope estimate indicates that, among all schools in the sample, a one-case increase in the number of violent incidents is associated with a increase of 0.776 number of transfer students.

	coefficient estimate	standard error
intercept	59.277659	2.874140
Violent_incidents	0.775546	0.054984

Table 1: Coefficient Estimates Table [SLR]

Goodness of Fit using R^2

R^2 is a statistical measure of how close the data are to the fitted regression line. It is the percentage of the response variable variation that is explain by a linear model.

$$R^2 = \frac{\text{reduction in variation}}{\text{total variation}}$$

Here, we compute R^2 using the `r2_score(...)` function, and eventually we have

$$R^2 = 0.07699304491868186$$

Only 7.69% of the variation in the number of violent incident is explained by the number of transferred student.Therefore, it is inadequate to conclude that the number of transferred student positively correlates to the number of violent incidents.

3.2.3 Multiple Linear Regression

In this section, we will use multiple linear regression to further investigate the relation of transferred students between other categorical variables. More precisely, we will model the log of transferred students as a function of urbanicity, school size and crime level.

$$\log(\text{Transferred}_i) = \beta_0 + \beta_1(\text{rural})_i + \cdots + \beta_4(\text{urban})_i + \beta_5(\text{300-499})_i + \beta_7(<300)_i + \cdots + \beta_9(\text{moderate\_crime})_i + \epsilon_i$$

	coefficient estimate	standard error
intercept	4.984034	0.081266
Rural	-0.337945	0.059062
Town	-0.145451	0.072576
Urban Fringe	-0.201611	0.050488
300 - 499	-1.058165	0.05801
500-999	-0.622436	0.045749
<300	-1.731028	0.070925
Low crime level	-0.451300	0.082800
Moderate crime level	-0.069188	0.088358
error_variance	-0.862077	NaN

Table 2: Coefficient Estimates Table[MLR]

Now look at both the estimates and standard errors for each level of each categorical variable; if some estimates are large for at least one level and the standard errors aren't too big, then estimated mean log expenditures differ according to the value of that variable when the other variables are held constant.

For example: the estimate for `Moderate crime level` is -0.069188; that means that, if urbanicity and school size are held fixed, the estimated difference in mean log transferred students between moderate crime level is -0.069188. If $\log(a) - \log(b) = -0.069188$, then $\frac{a}{b} = e^{-0.069188} \approx 0.933151$; so the estimated transferred students (not on the log scale) differ by a factor of about 1. Further, the standard error is 0.088358 , so the estimate is within 4SE of 0; the difference could well be zero. So the model suggests there is no difference in transferred studnets by crime level

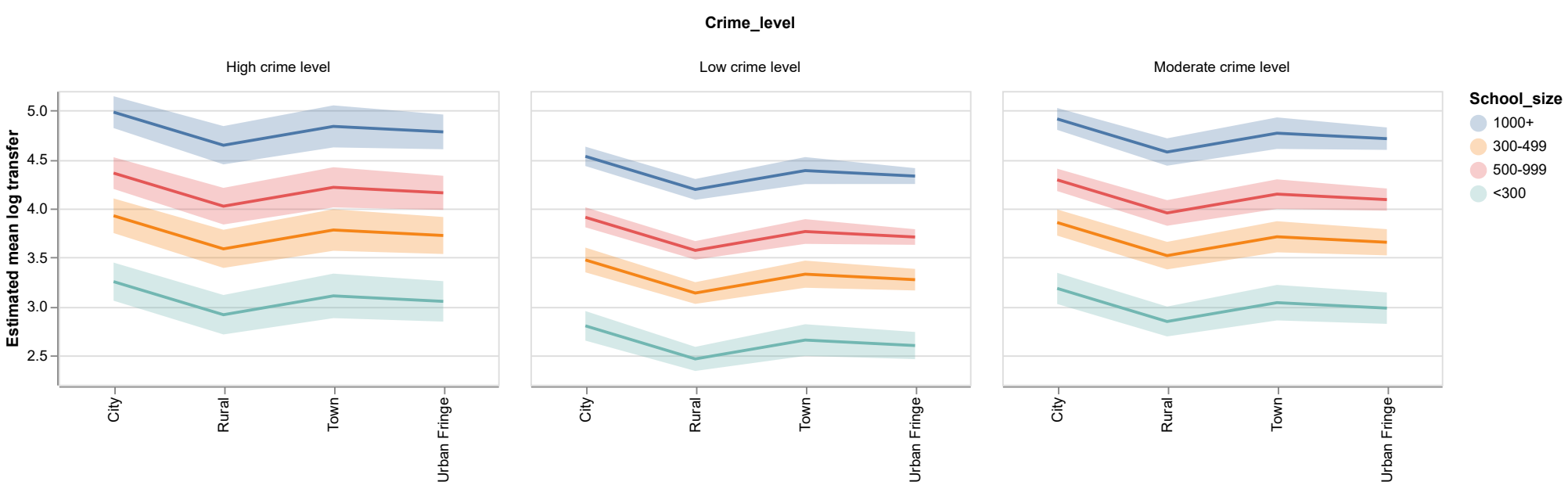

Figure 3: plot of relationship of urbanicity and estimated mean log transfer [MLR]

From the plot above, we can observe the relationship between urbanicity and estimated mean log transfer grouped by school size and faceted crime level;

1. For example, we can observe that in high crime level, in four catogory of urbanicity, school size with 1000+ always have the largest estimated mean log transfer;
2. Similarly, we can observe that in other two crime level, in four catogory of urbanicity, school size with 1000+ always have the largest estimated mean log transfer.
3. Therefore, it's adequate to conclude that school size with 1000+ always have largest estimated mean log transfer.
4. Specifically, within four catories of urbanicity, city always has largest estimated mean log transfer; rural always has the least estimated mean log transfer;
5. Finally, it is reasonable to conclude that from the four plots, they share similar trend in each crime level, comparing urbanicity against estimated mean log transfer.
6. Therefore, from the above analysis, we conclude that urbanicity and school size affect the number of transferred students.

- Students tend to transfer to colleges in big cities because there are more resources in the urban area, such as more opportunities for internships and better living quality.
- School size also influences a student's transfer decision. Large school size is often associated with abundant alumni resources. On the other hand, school size an indicator of popularity because a large school usually has more applicants.
- Surprisingly, according to the data visualization, there seemed to be an association between crime level and the number of transfer students. The mean log transfer is roughly the same for the high level and the median level of crime; however, the mean log transfer for the low crime level is lower compared with other crime levels. This is counter-intuitive and we suspect multicollinearity exists between Urbanicity and crime level.

Multicollinearity

To investigate whether multicollinearity exists between urbanicity and crime level, we slice both columns from the original data. We encode the categorical variables using `pd.getdummies()` and generate a correlation matrix using `corr()`. After that, we generate a correlation heatmap base on the correlation matrix using Seaborn.

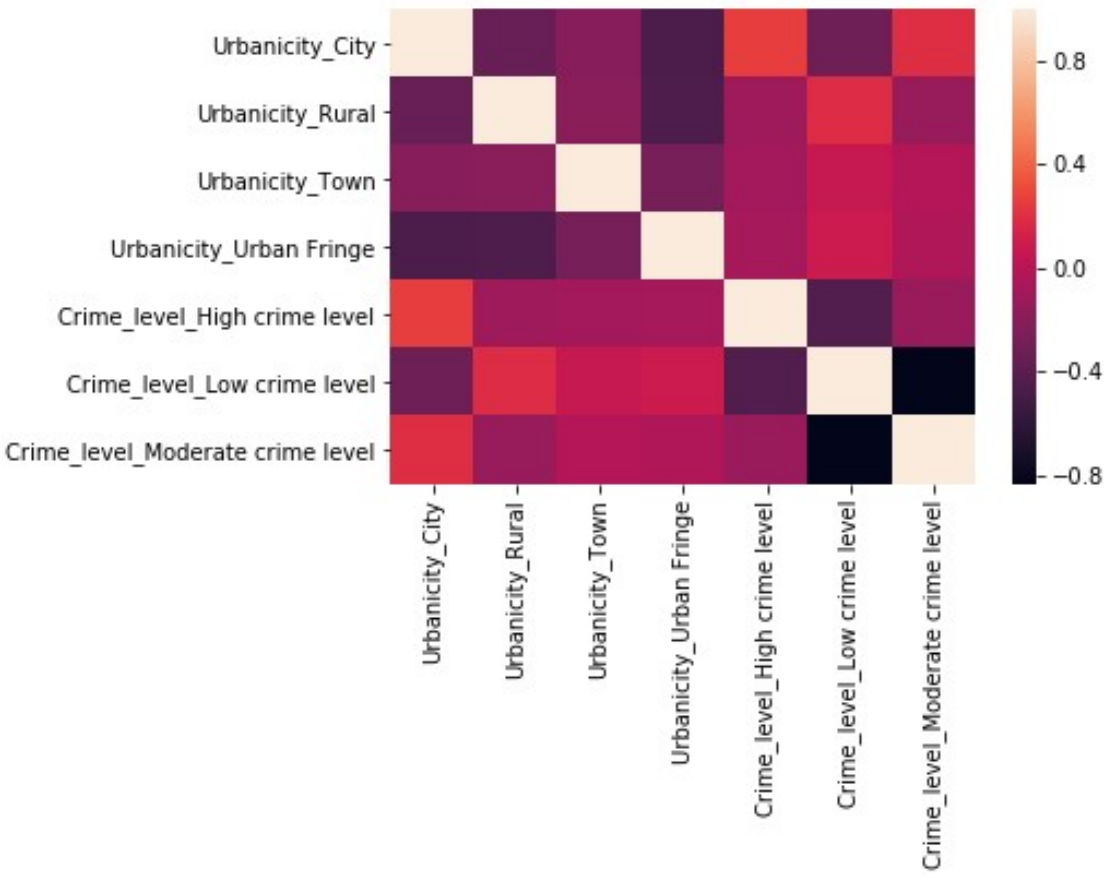

Figure 4: Seaborn Heatmap for urbanicity and crime level

Notice that city positively correlates with the high level of crime(0.4) and the moderate level of crime(0.3), whereas rural area positively correlates with the low level of crime(0.3). This result corresponds to our assumption in the previous section. Therefore, collinearity between urbanicity and crime level exists. Collinearity increase the variance of a regression fit. To address this issue, it's necessary to disregard crime level because it is redundant for our regression model.

4. Discussion

To sum up, the presence of minority students both aggravate the level of racial tension and increase violent incidents. The higher the minority percentage is, the larger the amount of violent incidents occur and the larger the higher level racial tensions (level 5, 8, 10) is. The reason of this phenomenon may be the ingrained racial discrimination towards minority group in society that has influenced the students at school. Importantly, throughout analyzing, we neglect the “Do Not Know” group because it may contain data from different level of minority percentage, whose analysis has no reference value.

In second study, we use exploratory analysis, simple linear regression and mutiple linear regression to find out possible factors that affect the number of transfer students. In particular, we analyze the relationship between the variables: school sizes, transferred students, urbanicity, and crime level. In scatterplot, we learned that school with larger school size attracts more transferred students. Also, we observe a potential connection between crime levels and urbanicity. By observing the Seaborn heatmap, we discover that there is a mild collinearity between urbanicity and crime levels.We conclude that school size and urbanicity affect the number of transfer students; however, due to the low R-square value of the SLR model and multicollinearity, there is little evidence to show that the number of transfers correlates with the violent incident or crime levels.