# EDUCATION PROJECT: CLEANING DATA

## Introduction

My goal in putting together this project was twofold. First, from a technical point of view, I wanted to spend some time understanding data cleaning techniques in Python. Secondly, from the point of view of educational policy, I wanted to spend some time exploring differences in high school education between two cities: Las Vegas, NV and Milwaukee, WI. I selected these two cities because while they are similarly sized (each around 600,000 people), they are considered to be, respectively, the least and most segregated metropolitan areas in the United States.

I was interested in answering three questions.

(1) Do the types of schools (public, charter, alternative, or private) attended by high school students differ between these two cities?
(2) How do the racial distributions of the different types of schools compare to the racial distributions of the city populations as a whole, and of the populations of high school aged people in these cities.
(3) How does access to Advanced Placement (AP) testing[1] in these two cities compare with each other and with the American high school population as a whole.

In order to answer these questions, I used primarily data from the following sources:

(1) The Civil Rights Data Collection,
(2) The National Center for Education Statistics (NCES) (also here),
(3) The College Board, and
(4) The United States Census Bureau Fact Finder.

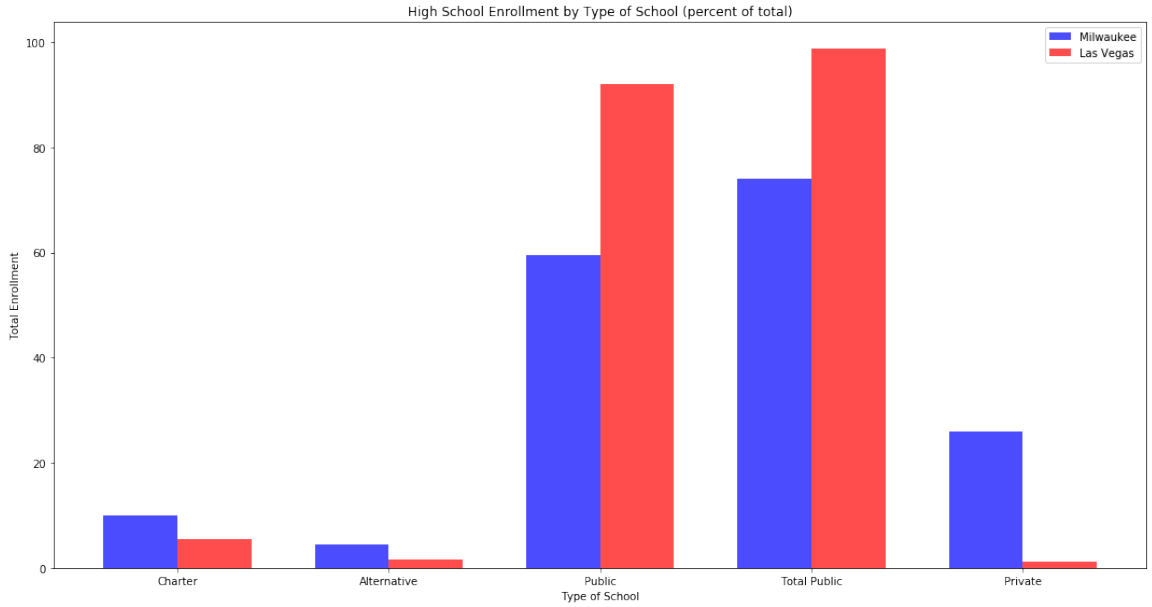
## Question 1:

In order to answer this question, I first needed to solve a minor problem with my data. To be precise, the Milwaukee Public School District includes those public schools that are located within the city of Milwaukee's boundaries. By contrast, the schools in Las Vegas belong the Clark County School District, which also includes suburban and rural schools. As a result, direct comparing schools from these two districts seemed unlikely to be particularly edifying. Thus, I began by filtering schools based on their location code, that is, a description of the area in which the school was located, retaining only those schools found in large urban areas.

In considering the choice of type of high schools in the two cities, I obtained the following results. Milwaukee has a much higher percentage of students in private schools (26% vs 1.1%) and charter schools (10% vs 5.4%) than does Las Vegas. Not surprisingly, they have a much smaller percentage of students in public high schools (59.5% vs 92%). It is worth noting here

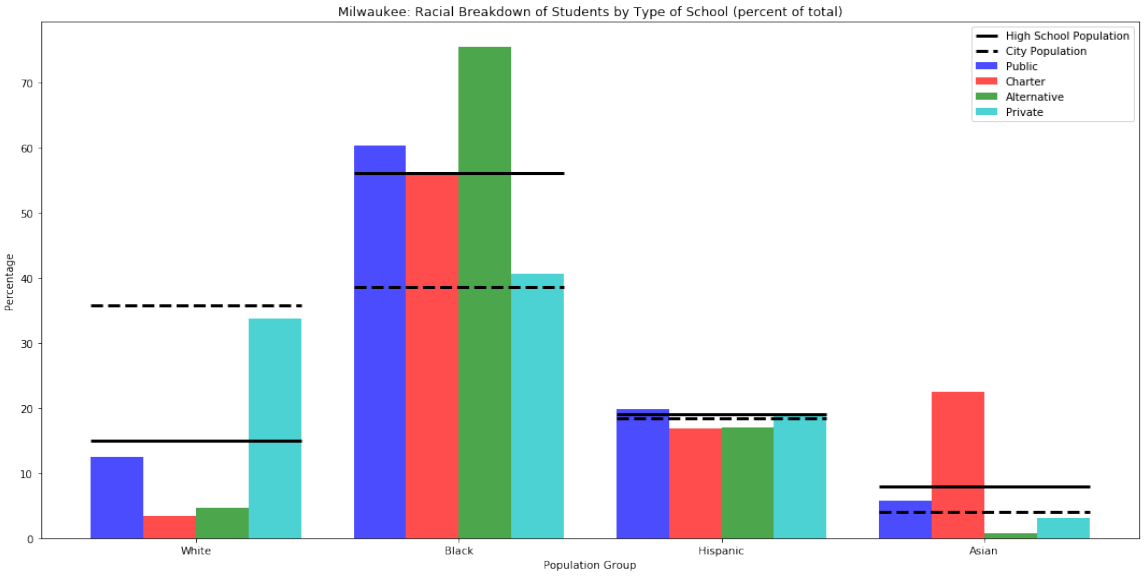
[1]These tests allow students to earn university credit while still in high school.

that Wisconsin has a robust voucher system[2], as well as significant support for charter schools at the state level.



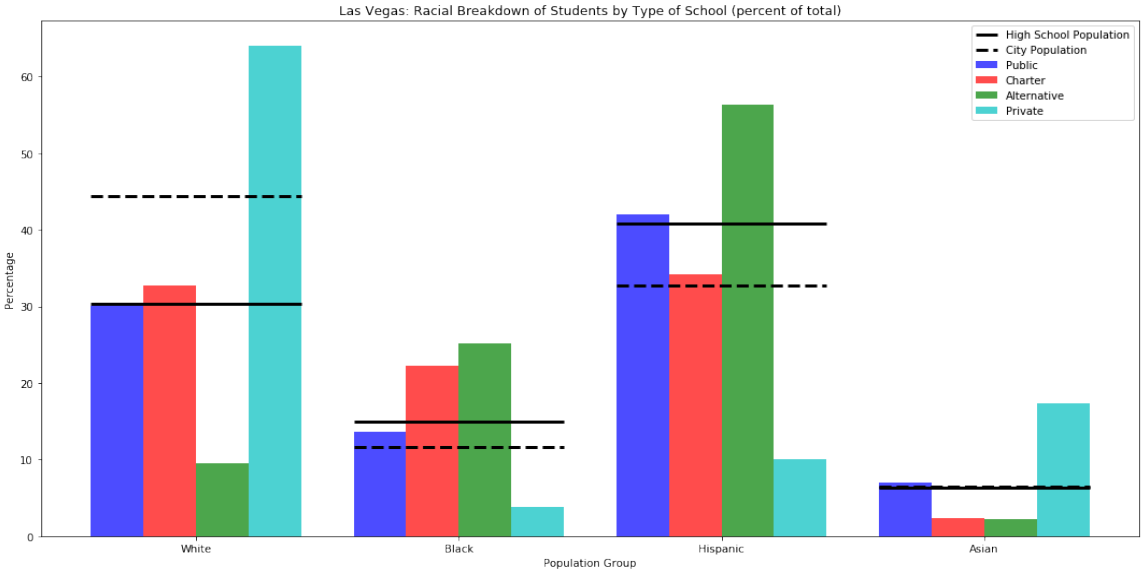

## Question 2:

In comparing the high school populations of the various types of schools for the two cities I found the following results. In Milwaukee, we the city population as a whole is much whiter than the high school population. In addition, private schools are disproportionately white, while black students are over represented in alternative schools[3]. Hispanic student representation in all types of schools hues fairly closely to their representation in the high school population. Finally, Asian students are significantly over represented in the charter schools.




[2]Vouchers allow students to receive public funds to pay at least part of their private school tuition.
[3]These schools often have students who are seen as problematic.

For Las Vegas, we find that the city population is once again whiter than the high school population. Here, both white and Asian students are disproportionately represented in the private school, while black and Hispanic students are over represented among the alternative school population. In addition, black students are over represented among charter school students.




## Question 3:

For AP testing, I explored questions of both access rates and success rates. This was complicated by a number of factors, largely attributable to the types of data that I had available. To begin with, the groups of students that I had data on depended on whether I was looking at AP test data for Milwaukee and Las Vegas, or at AP test data on a national basis. In the first case, I only had data on students from public, charter, and alternative high schools[4]. In the second case, my data included students from all schools, including private schools. I attempted to address this problem by using information from the NCES to estimate the total number of high school students in both publicly funded and private schools by race. This allowed me to estimate access and pass rates for groups of students nationally.

Next, in the school district data, obtained from the Civil Rights Data Collection, subgroups that had fewer than three members were simply listed as having '$\leq 2$' members, which added uncertainty to the data. In order to try to tease out values for as many of these as possible, I used the totals across all ethnic/racial groups together with the known values[5]. Similarly, I used information within each subgroup about number of students taking tests, passing tests, etc. to fill in additional values.

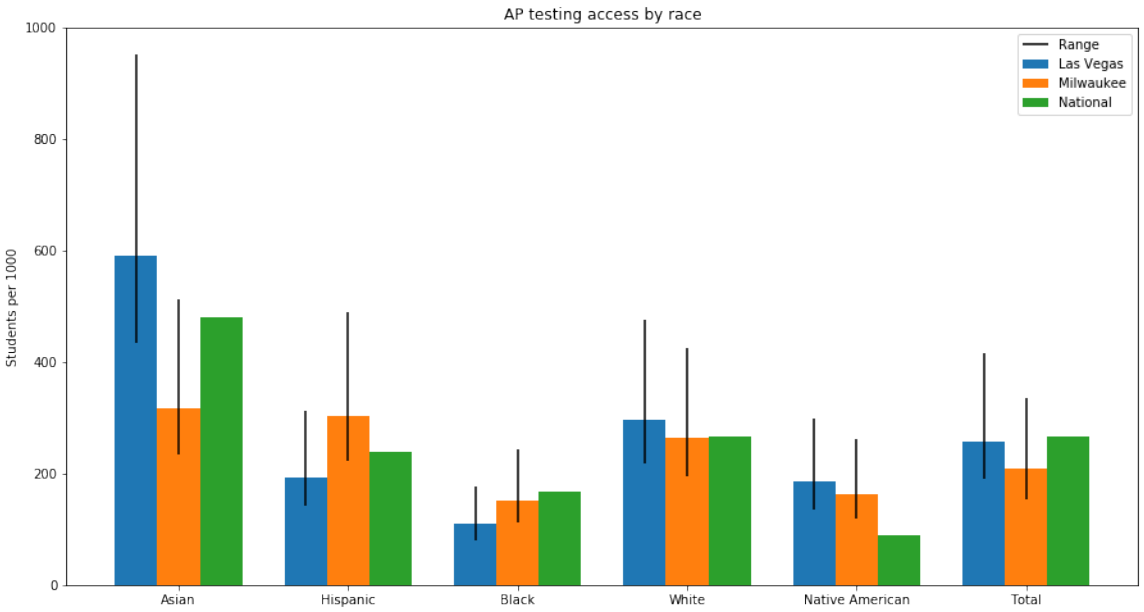
Finally, data from the College Board was broken down by class year[6], while data for Milwaukee and Las Vegas combined data for all four high school classes into a single value. Here, I assumed that the distributions of student class years among the students taking AP exams in Milwaukee or Las Vegas who were were similar to those at the national level. Furthermore, I assumed that the vast majority of students who took AP exams before their final year of high school would also take them in their final year.

[4]All high schools that are publicly funded.

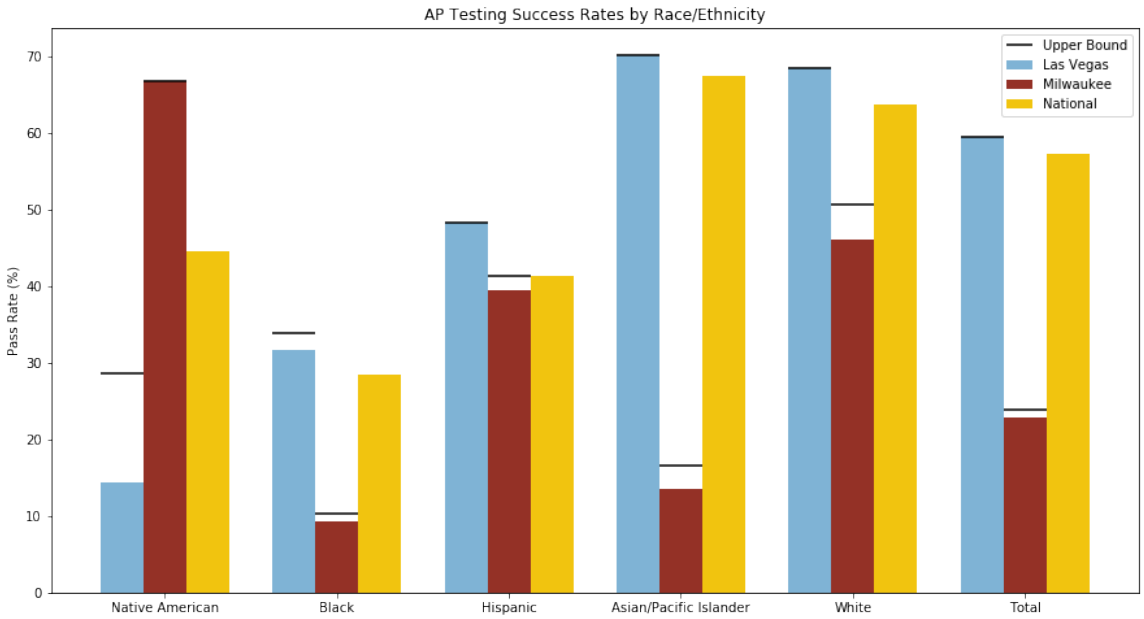
[5]For example, if there are 20 total students who took an AP exam, and we know that there are 15 black students and 5 Asian students, then no other ethnic groups can have had a student take the exam.

[6]Students can take AP exams in any year of highschool and can also take them in multiple years, if desired.

I found that certain groups of high school students in Milwaukee and Las Vegas seemed to have more access to AP tests than there analogues at the national level. In Las Vegas, Asian and Native American students took AP exams more frequently than their peers at the national level. In Milwaukee, the same was true of Hispanic and Native American students. Of course, there are also groups that take an unusually low number of AP exams relative to their cohorts at the national level, in particular Asian students in Milwaukee and black students in Las Vegas.




From the point of view of passing rates, we again find differences between the groups. Among them, Native American students in Milwaukee and white and Hispanic students in Las Vegas succeed at higher rates than their peers at the national level. By contrast, Native American students in Las Vegas as well as black, white, and Asian students in Milwaukee seem to have less success than their peers at the national level. At the level of the city as a whole, high school students in Las Vegas seem to have had marginally more success than students nationally, while high school students in Milwaukee had substantially less.



## Conclusion:

In looking at our data, we see that there are significant differences in the types of high schools chosen by the parents of high school students. In particular, for both Milwaukee and Las Vegas, we see that the segregation among school types is quite strong, though Las Vegas has seen less flight to private high schools than Milwaukee. Perhaps as a result of this, while access to AP exams for public school students of both schools is relatively similar to that at the national level, Milwaukee shows a striking lack of success among students who take an AP test.