# Introduction

Asian American comprises a panoply of differing and unique intersectional identities, histories, and experiences, yet Asian Americans are essentialized as a group and broadly stereotyped as the model minority, which shapes information to college access and campus resources (Museus & Truong, 2009; Palmer & Maramba, 2015; Poon & Byrd, 2013). Scholars and researchers have frequently called for the disaggregation of Asian American data to ensure that ethnic minorities are supported within the Asian American racial category (Museus & Truong, 2009).

In the pursuit of data disaggregation, I hope to map the different experiences of Asian American ethnic groups across Philadelphia and their access to higher education utilizing K-clustering and Moran’s I. K-cluster analysis will be used to show how different Asian American groups are different distributed across Philadelphia. Moran’s I will be used to show how different resources are distributed across Philadelphia; for example, tutoring services, higher education institutions, and conditions of k-12 schools. Overall, I argue that Asian Americans have a variety of different educational resources, and as such public policy should better distribute these resources.

# Literature Review

## Asian American Geographic Differences

Liu's (2018) ethnography highlighted highlights the differences in Asian American perspectives, they built a framework that includes three parts. One part of that framework is the discourse on Xi Jinping's "Chinese Dream," which emphasizes class advancement and ethnic empowerment through market liberalization and expansion of, specifically, Chinese capital. The Chinese Dream can be applied to the Chinese immigrants' movement that moved into ethnoburbs following Japanese and Mexican Americans, which fundamentally reshaped the ethnoburb (Cheng, 2013).

Race as a geographic cultural construct shapes how students experienced being Asian American. In Chan’s (2017), study on geographic differences in being Asian American the theme of race as a social identity was broken updivided into four subthemes: (1) distancing racial identity, (2) the strategic use of racial identity, (3) shifting experiences of race and racial identity to describe the importance of students' hometowns and high schools, and (4) how they now felt about their racial context. This theme described how some students felt closer to their identity, while other students felt more distanced because they no longer were the only Asian American identifying person in their hometown.

The literature on Asian American geographic differences is not extensive, but the literature that does exist indicates that there are differences in access to locations and socialization. Patterns of immigration and discourse also shaped where Asian Americans could move to, and the environment that Asian Americans occupy also shape their own ideas perceptions of self. Geography plays an important role in shaping the Asian American experience; however, very little has been done to explicitly understand geography and its relationship to college access for Asian American populations.

## College Access Frameworks

Postsecondary institutions in the United States are critical for developing a workforce and providing individual opportunities for development. Although the need for postsecondary education is evident, the need is often mismatched by various factors (Dache et al., 2021). Perna (2006) offers a conceptual model of higher education which encompasses four layers: (1) habitus, (2) school and community context, (3) higher education context, (4) social, economic and policy context. Although this framework does consider context, it does not explicitly address the geographic context (Turley, 2009). Turley argues that college choice must be situated in the geographic context and found that high school seniors had a wide range of colleges within commuting distance, zip code had a small but significant increase in the odds of applying to college. Finally, schools that are more conveniently accessible had higher application rates.

# Framework

Hillman (2016) builds upon the geographic distance of higher education institutions to argue for the existence of education deserts — places where there are no educational opportunities. These education deserts show that place shapes the decision-making process in deciding whether to attend and where to attend college. The idea of applying the term geography of opportunity is to show that there are unequal opportunities to higher education. Like food deserts, education deserts are constrained along the lines of race and class. Individual choices to go to college are shaped by their geographic context, which constrains the options of school context, community context, and their social habitus. Dache-Gerbino (2016) argues that geographic context is critically important using a Critical Geographic College Access (CGCA) framework to visually show how urban development and modernization failed black communities and that locations of colleges are not just coincidences but socially constructed around a history of residential segregation.

Although geographic analysis considers the way residents and communities can see and access higher education (Dache-Gerbino, 2016; Turley, 2009), Colleges and universities must recognize and reach out to these communities. Jaquette and Salazar (2018) found that college recruiters strategically select high schools for recruiting, typically picking high schools that are whiter and wealthier. Even when schools primarily made of students of color performed well on tests, colleges would still visit the predominately white high schools. This proposed study challenges college access from a student perspective to the responsibility of higher education institutions. College access is not just about what school students visually see and are conveniently close to, but what schools have taken the time to reach out and recruit.

# Methods

To understand the geographic context of the data, I am pulling from five different sources of data, (1) U.S Census Bureau data (TidyCensus), (2) Integrated Postsecondary Education Data System (IPEDS) database, (3) SafeGraph, (4) Carnegie classification, and (5) Open Trip Planner (OTP) - SEPTA Metro and Bus GTFS files. I will be utilizing the U.S. Census bureau’s five year American Community Survey (ACS) data using the Tidycensus package in R. The U.S. Census Bureau provides 5-year estimates from 2011 by census track, I plan to use this data to provide a context to the importance of Asian Americans to the Philadelphia region, and more specifically the growth of specific Asian ethnic groups in Philadelphia county. IPEDS data is collected by the National Center for Education statistics (NCES), and it is required that higher education institutions submit their data to receive federal funding. SafeGraph is company that collects data point data to track how many people come in and out of a space. With this they also have data based on the North American Industry Classification Systems (NAICS). I requested data for all the education related industries in Philadelphia. This dataset has been coded to separate different institutions from each other. Carnegie classifications are ways that higher education institutions are organized, this can vary by variable. This data is not included in the IPEDS dataset and needs to manually be added. Lastly I am using OTP, to predict estimated travel times using SEPTA rail and bus times. The current proposal does not include any data from this section, because more data exploration needs to be done to find what locations are important to understand transit. OTP is time consuming in running the code, and to ensure efficient use of resources I want to make sure I have my location and goals in mind before use.

## Spatial Autocorrelation Analysis

To begin this analysis, I first determined if broadly Asian populations were spatially autocorrelated in Philadelphia using Moran’s I. Moran’s I has been widely used to test for spatial autocorrelation or spatial dependencies and its value determines the strength of autocorrelation indicating how clustered values are. Values that are closer to 1 indicate strong positive autocorrelation, while values closer to -1 indicate negative autocorrelation that being how repelled values are. Conversely, if values are positively autocorrelated they are spatially clustered. If the value is close to 0, then there is no spatial autocorrelation, indicating a random pattern. Unlike Pearson's correlation coefficient, Moran's I does not always lie within -1 and 1 and can be situated beyond each of these values. Moran's I can be found through the following formula where is the mean of the variable, is the variable value at particular location i, is the variable value at location j, is the weight indexing location of i relative to j, and finally, n is the number of points or areal units. The formula for spatial autocorrelation shares similarities with the correlation coefficient. For example, in Pearson's correlation coefficient where the difference from point i is subtracted from the average of all the y values is captured in spatial autocorrelation by taking the weighted value i and j. The difference is then multiplied from the actual value of x at the point i from the average of x.

To use Moran's I, a weight matrix needs to be created: rook or queen matrices. Rook neighbors describe when a polygon is in contact with other polygons that share the same sides. In contrast, queen neighbors describe when a polygon shares a side or a vertex with another polygon. The decision of the type of neighbor to use will impact how weight matrices are created. Matrices are created by forming an n by n table to summarize the spatial relationships in the data. This can be done by using continguity-based measures or distance-based measures. Distance-based measures measure if polygons are within a certain distance of each other by indicating a one which is yes, or zero which is no. Contiguity-based proximity measurements capture if polygons are sharing boundaries and are captured by the weight matrix with either a one where they share a rook or queen neighbor (whichever is selected), or zero which is no they do not share a rook or queen neighbor.

For the initial study I used queen matrices to include all shared vertices and line segments. Once value I has been found, a significance test must be performed on the value to determine its significance. To test the significance of I, the variable in question is shuffled to different geographic locations on the map. Moran's I is then calculated in the new geographic location. This process is completed 999 times in addition to the first permutation to yield 1000 different possible outcomes for Moran's I. Each of these values are then placed in descending order.

In this instance, the null hypothesis states that there is no spatial autocorrelation. The alternate hypothesis is that there is spatial autocorrelation. Simultaneously, the pseudo P-value finds the likelihood of a value as large as the observed value using the 999 possible outcomes. The pseudo p-value is obtained by taking the rank and dividing by 1000; for example, if the rank of the actual case were ranked number 1, then the pseudo p-value would be , which would indicate a statistically significant relationship because the value is less than 0.05. In this case, this would determine that the I value is statistically significant.

## K-Cluster Analysis

Rather than create a model that predicts Asian American college attendance by ethnic group, I will argue that Asian American populations in Philadelphia have a variety of different conditions which shape their access to higher education. To do this, I will utilize K-cluster analysis to create two different models. The first model will use indicators of college access such as economic outcomes and population education status. The second model will use broad Asian ethnic categories to show where different Asian American populations are distributed across Philadelphia.

There are a variety of methods to determine the most effective way to select the amount of clusters in a group. In R there is a package called NbClust which has up to 30 methods for choosing the optimal number of clusters. One particularly useful method is the Scree Plot. In using the Scree plot, the key is to identify the bend in the “elbow” of the plot where the drop in the within group Sum of Squared Errors becomes very small.

After the number of groups are selected there is an iterative 6-step process. First, K data points are randomly selected as cluster centers. Second, distances between the data points and K cluster centers is calculated. Third, each data point is assigned to a cluster whose distance from the cluster center is minimized among all cluster centers. Fourth, once all data points are assigned to a certain cluster, they are then recalculated to the new cluster center. Fifth the distance is updated between each data point and new cluster centers. Finally, if points are not reassigned the process is complete. If the data points are reassigned, then the new cluster centers need to be recalculated and the process restarts from step three of assigning each data point to a cluster whose distance from the cluster center is minimal among all cluster centers.

The goal of K-means is to minimize the within-cluster sum of squared errors, which is calculated for each cluster by computing the squared distance between each observation and the centroid of the cluster. There are some limitations to using K-means such as having to specify the number of clusters in advance, using numeric data, and being unable to handle noisy data and outliers. Other problems also include clusters differing in size, density, and non-globular shapes.

# Findings and Analysis

## Summary statistics

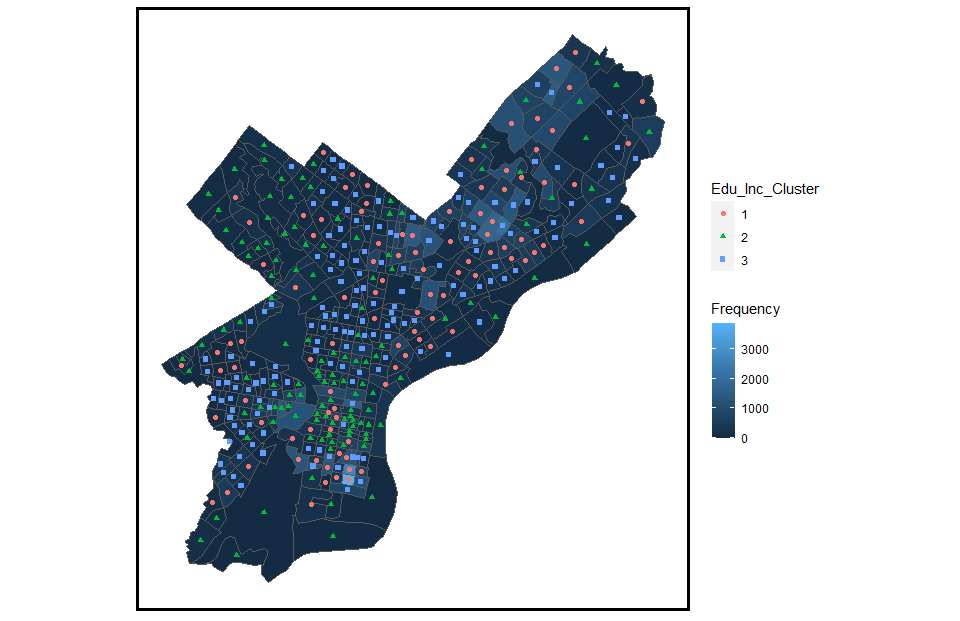
Based on the data above, I have compiled three sets of data: (1) enrollment of higher education by race, (2) census tract data on Philadelphia, and (3) geographic higher education context of Philadelphia. The findings from overall enrollment of higher education show that from 2003 to 2020, there has been an overall decline in percentage of white students who have enrolled in higher education. This percent is a breakdown of racial category divided by the total population, and as such indicates that of the total population there has been an increased enrollment of non-white students in higher education. Upon closer examination of non-white enrollment in higher education, in 2020 Asian or Pacific Islanders made up 15% of total enrollment while Black non-Hispanic students made up 11.6%. In 2003 Asian or Pacific Islander students only made up 9% of the total population. Although there is an increase in Asian or Pacific Islander populations in the United States, it is unclear which ethnic groups are represented in this population.

To try and understand the ethnic identities represented in the dataset I looked through the census data on Philadelphia to understand the breakdown of frequencies of racial groups. In 2019, Philadelphia primarily consisted of 665,333 black people, 642,060 white people, followed by 114,315 Asian people. Of the 114,315 Asian people 37,588 of them identified as Chinese, and 23, 443 identified as Indian. The Chinese and Indian population in Philadelphia have grown the most in comparison to other ethnic groups in Philadelphia. In 2011, there were 26,494 Chinese people in comparison to now 37,588 over the course of 8 year. The increase in Chinese and Indian populations may be related to globalization factors; however, before making this assertion I will need to find more literature to support this argument.

Although I have found Asian ethnic group demographic frequencies, the challenge is relating this directly back to higher education enrollment. Although the ACS 5-year survey does provide this data, the data is consolidated by census tract, and it is unclear of what the experiences of individuals or households are in these census tracts. To remedy this, I am considering using Public Use Micro Sample (PUMS) data, to have a more careful look at the individuals in each census tract. PUMS data is also data collected by the US census bureau and provides data by household and individual rather than by census tract. I am continuing to explore PUMS data to understand how I can relate this back to higher education enrollment.

In addition to the demographic data, I have begun to plot the demographic data in the context of Higher Education institutions and other educational institutions. These maps have helped me identify areas in Philadelphia where the Asian American population is highest. These have shown interesting findings in that, Asian Americans are densely populated in South Philadelphia rather than the Chinatown area. From my own qualitative research, I have found many of the Southeast Asians that I have been working with located in South Philadelphia, in contrast to Central Philadelphia and Chinatown, I expected more East Asians there. To address this inconsistency, I have engineered a feature that is a ratio of ethnic group to Chinese population. What this feature will tell me are what areas primarily consist of Chinese Americans in contrast to other ethnic groups. With this I hope to be able to identify the census tracts that have a higher ratio to identify areas in the city that have a larger ratio to ethnic Chinese, then I can map their travel times to higher education institutions or other education institutions. With this I hope to be able to identify the census tracts that have a higher ratio of ethnic group that is not Chinese to Chinese in the city that have a larger ratio to ethnic Chinese, then I can map their travel times to higher education institutions or other education institutions.

## Clustering Analysis



Figure

Through clustering analysis, I made two different models, one based on economic and income variables, and the second model based on normalized frequencies of different populations of Asian Americans. Figure 1 shows the cluster on Education and income variables layered on top of the racial category of Asian by census tract. At this moment, the clustering analysis needs more fine tuning. K-cluster analysis does not seem to give clear results when there are many different variables, and as such the variables that I have will need to be consolidated into larger categories for the purposes of clustering the data. However; clustering analysis does provide a way to understand groupings and environmental factors of Asian Americans across the Philadelphia region.

# Final Deliverable

The final deliverable for this paper, will be a paper that will argue for a more nuanced understanding of Asian American college access based on a more in-depth understanding of Asian Americans, based on environmental factors of the census tract. This will be done through K-clustering analysis and Moran’s I.

Citations

Dache, A., Blue, J., Bovell, D., Miguest, D., Osifeso, S., & Tucux, F. (2021). A Calle decolonial hack: Afro-Latin theorizing of Philadelphia’s spaces of learning and resistance. *Teaching in Higher Education*, *26*(7–8), 1077–1097. https://doi.org/10.1080/13562517.2021.1940927

Dache-Gerbino, A. (2016). College Desert and Oasis: A Critical Geographic Analysis of Local College Access. *Journal of Diversity of Higher Education*, 1–21. https://doi.org/10.1037/a0021277

Hillman, N. W. (2016). Geography of College Opportunity. *American Educational Research Journal*, *53*(4), 987–1021. https://doi.org/10.3102/0002831216653204

Jaquette, O., & Salazar, K. (2018, April 13). Opinion | Colleges Recruit at Richer, Whiter High Schools. *The New York Times*. https://www.nytimes.com/interactive/2018/04/13/opinion/college-recruitment-rich-white.html, https://www.nytimes.com/interactive/2018/04/13/opinion/college-recruitment-rich-white.html

Museus, S. D., & Truong, K. A. (2009). Disaggregating Qualitative Data from Asian American College Students in Campus Racial Climate Research and Assessment. *New Directions for Institutional Research*, *Summer*(142), 17–26. https://doi.org/10.1002/ir

Palmer, R. T., & Maramba, D. C. (2015). The impact of social capital on the access, adjustment, and success of Southeast Asian American college students. *Journal of College Student Development*, *56*(1), 45–60. https://doi.org/10.1353/csd.2015.0007

Perna, L. W. (2006). Studying College Access and Choice: A proposed Conceptual Model. *Higher Education: Handbook of Theory and Research*, *XXI*, 99–157. https://doi.org/10.1016/0273-1177(93)90208-S

Poon, O. A., & Byrd, A. (2013). Beyond tiger mom anxiety: Ethnic, gender, and generational differences in Asian American college access and choices. *Journal of College Admission*, *Fall*(221), 22–31.

Turley, R. N. L. (2009). College proximity: Mapping access to opportunity. *Sociology of Education*, *82*(2), 126–146. https://doi.org/10.1177/003804070908200202