Analysis of MLS teams:

City demographics, Attendance, and Club Success

GES 486

Sam Weaver

Introduction:

One of my favorite things in life is to watch soccer. From playing for almost 20 years of my life, to finding professional clubs to support, soccer has been a very important aspect of my life. Despite the presence of Major League Soccer, I have always struggled to follow the sport in America and have instead looked elsewhere in Europe. This is due to many factors including, availability, skill level, and personal interest. For this project I only focused on the first two reasons. Major League Soccer was first founded in 1996 and is extremely young compared to other leagues around the world. At launch the league had high hopes for success in America and looked to increase national interest in the sport. This was partially successful, as general interest has grown, but the American fanbase is not comparable to leagues in Europe and South America. Soccer has simply been played longer in other countries and has become a significant part of cultures, specifically in Central and South America. Statistics published by the Statista Research Department showed that in 2020, 30.4% of MLS players identified as Latino or Hispanic. With increasing levels of immigration in the States, I hypothesize that cities with larger Hispanic and Latino populations will have greater MLS attendance rates and in turn, more successful teams over time. I also

believe that lower median household incomes affects populations ability to attend games.

Methods:

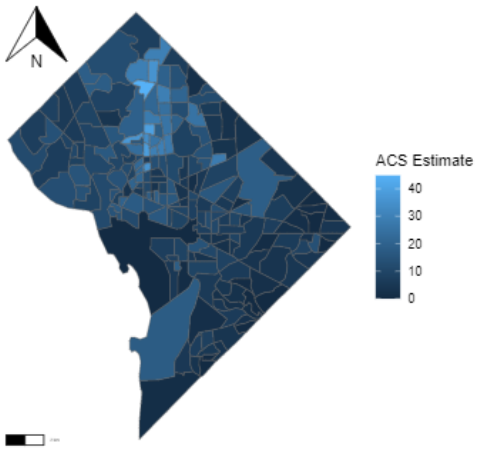
To start my Analysis, I gathered general data on the MLS from https://github.com/gavinr/usa-soccer. This included a .csv file of current MLS teams, their stadium coordinates, capacity, when the team was founded, and a link to the team website. For my purposes, I only needed the team name, coordinates, and stadium capacity. Using RStudio, I imported the .csv file and narrowed the data set to only include my significant data. I also used data from https://www.goal.com/en-us/news/mls-cup-all-winners-list/blt15fd45904a53e495 to add a field for the total number of MLS Championships each club had won. Using this information and ggplot2, I created a map of the USA that showed the location of each MLS club, and how many trophies they had. The goal of this was to visualize how widespread the MLS is and to determine which cities had the most championships. To further analysis of individual teams, I obtained the 2021 average attendance records for each team and created a .csv file to import into RStudio. This is also where I added the stadium capacity information from the first data set. This allowed me to create a field for the average percentage of the stadium, average / total capacity, that each team filled for the entire 2021 season. With these data sets, I picked the 2 cities with the most success, based on championships, and the city with the lowest percent attendance to analyze and compare the Hispanic populations and the median household income using 1-year ACS estimates from the US Census Bureau. I pulled the total populations and

split the data into White, Black, Native, Asian, HIPI, and Hispanic in order to get a percent total Hispanic population. From this analysis, I created a set of maps comparing median household income to percent Hispanic population, and created graphs to determine the correlation between the data.

Results:

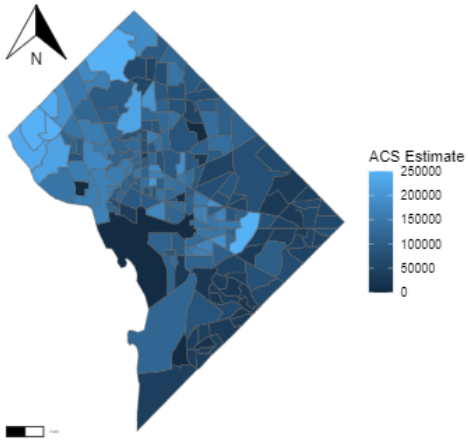
From my initial map and data tables of MLS teams, you can see that LA Galaxy and DC United have the most championships in MLS history with 5 and 4, respectively. Both of these teams have been a part of the MLS since it was founded in 1996, and have played every year since. Los Angeles has since added an additional team, LAFC joined in 2018, who has since won one championship. From the attendance data, New York City FC had the lowest attendance percentage in 2021 and, on average, filled 36% of their stadium. It is to be noted that LA Galaxy had a 54% attendance rate and DC United had 65%. LAFC was the highest on the list, filling 88% of their stadium capacity. From this information I pulled 1-year ACS data for total population and median household income for Los Angeles, Washington DC, and the Bronx NY. I chose to only look at the Bronx rather than all of New York City because that is where New York City FC's stadium is located. Using the census data for 2021, I used ggplot to create interactive maps at a tract level. This allowed me to compare individual census tracts and analyze city trends for both Hispanic populations and mhhi.

DC Hispanic Population



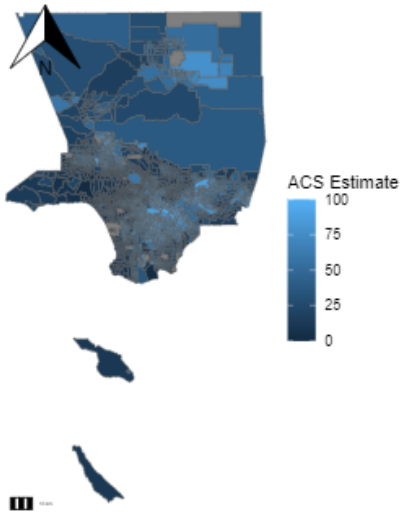
Data from 2021 1-year ACS, US Census Bureau

DC Minimum Household Income

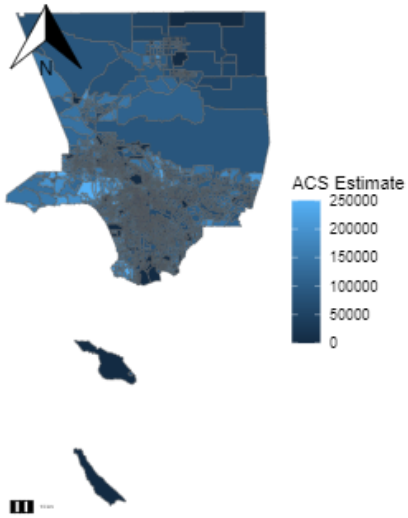


Data from 2021 1-year ACS, US Census Bureau

LA Hispanic Population

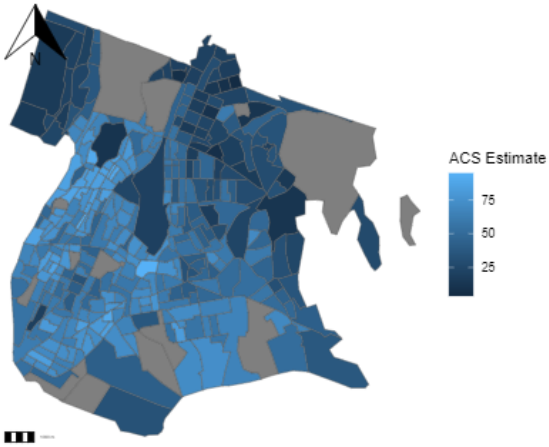


LA Minimum Household Income



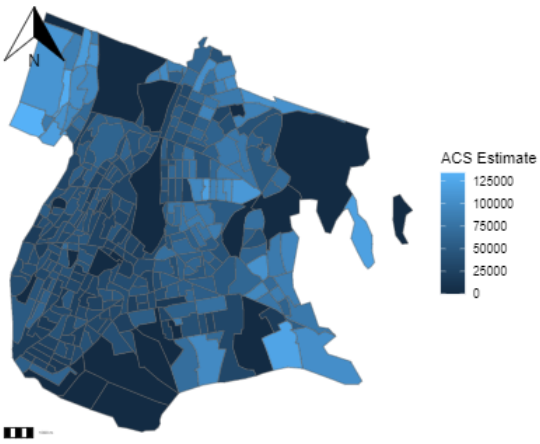
Data from 2021 1-year ACS, US Census Bureau

Bronx Hispanic Population



Data from 2021 1-year ACS, US Census Bureau

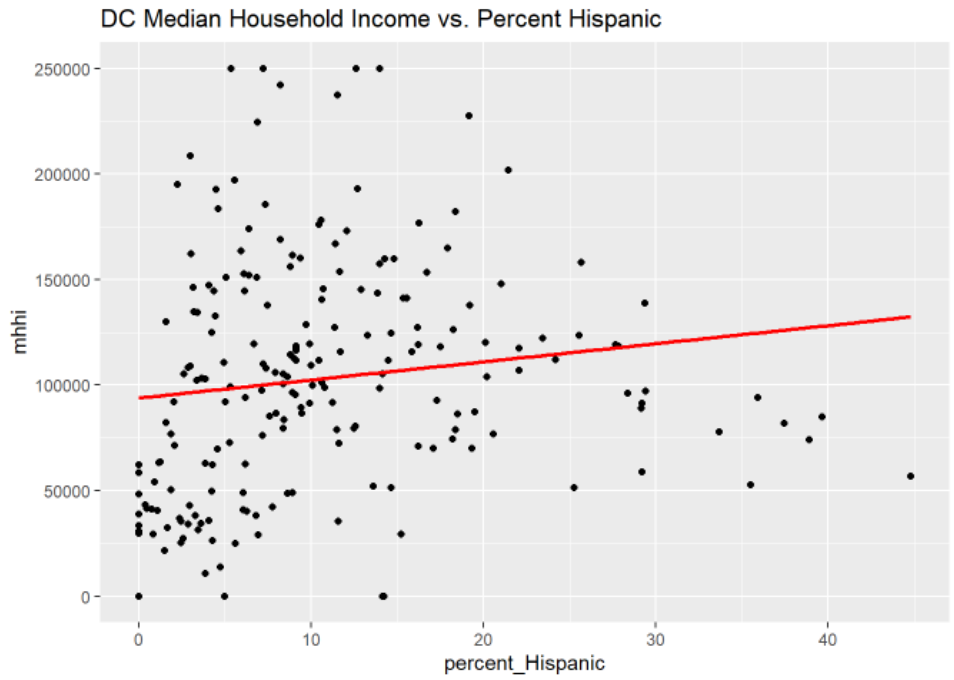
Bronx Minimum Household Income

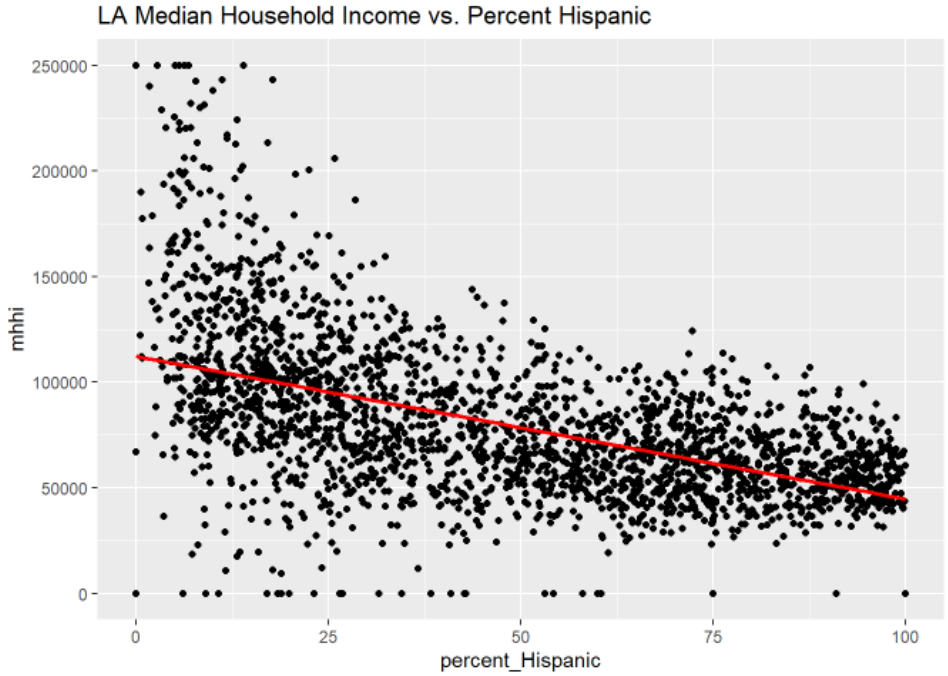


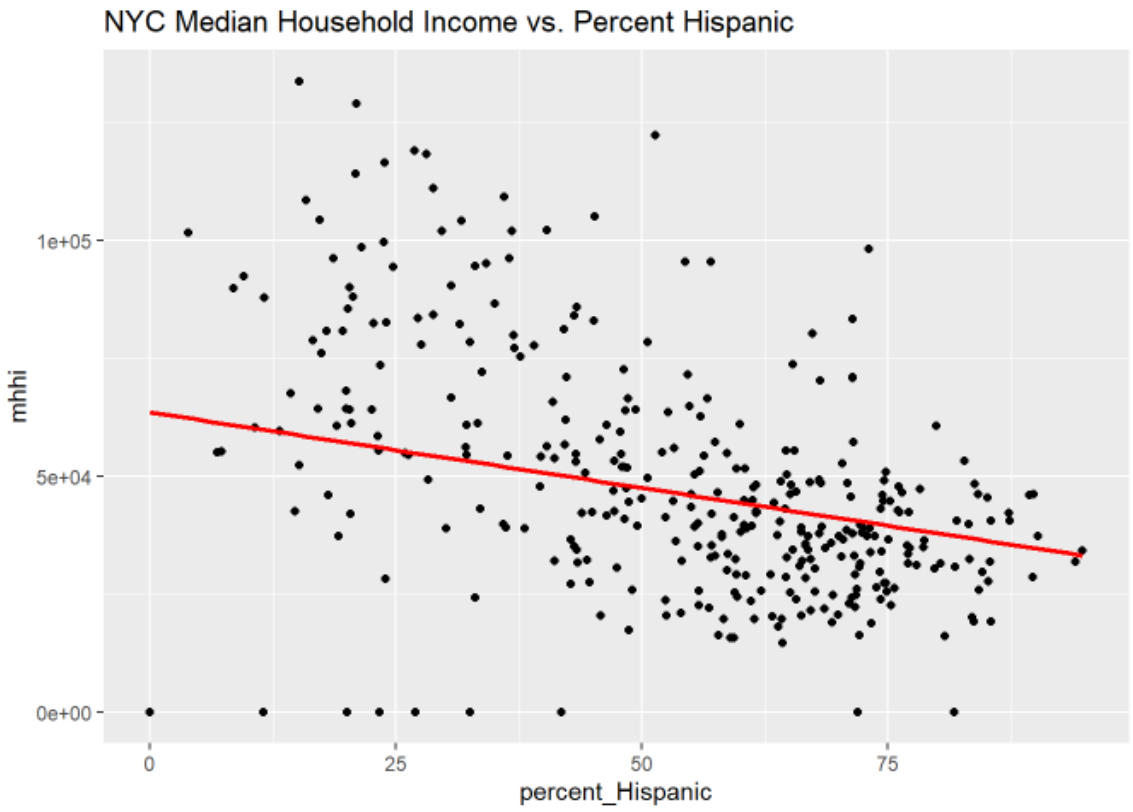
Data from 2021 1-year ACS, US Census Bureau

Looking at the generated plots, there seems to be a clear inverse relationship between Hispanic population and median household income. That is as the Hispanic population increases, light blue regions on the left plots, the median household income decreases, dark regions on the right. At first glance, this is not as noticeable for the DC plots, as the overall median household income is very high.

To further understand if the Hispanic population had a correlation with lower median household income, I used scatter plots and a linear regression to determine the overall trend of income as the Hispanic population increases.


DC Median Household Income vs. Percent Hispanic
mhhi
250000
200000
150000
100000
50000
0
0
10
20
30
40
percent_Hispanic


LA Median Household Income vs. Percent Hispanic
mhhi
250000
200000
150000
100000
50000
0
0
25
50
75
100
percent_Hispanic



The plots for Los Angeles and the Bronx confirmed that there is a very negative correlation between the Hispanic population and median household income. The DC plot showed a slightly positive correlation but only had 206 census tracts compared to 2498 for LA, and 361 for the Bronx. The slopes of the linear regression lines for DC, LA, and Bronx were 0.137896, -0.5074496, and -0.2770464, respectively.

Conclusions:

Looking at the results, there were clear correlations between Hispanic populations and median household income for both Los Angeles and the Bronx, with a slightly positive correlation for Washington DC. This showed that Hispanic populations generally had lower incomes in these neighborhoods. I believe that this directly correlates to the attendance rates for DC United, LA Galaxy, and New York City FC.

New York City FC had the lowest attendance rate in the entire league and has only won one championship in their history. LA Galaxy has won the most championships in history, but has been competing with LAFC for fans since the team was added to the league in 2018. It was also noted that LAFC had the highest attendance rate, which indicates a reason for LA Galaxies to be much lower. I think that the Hispanic population in Washington DC having a higher median household income directly correlates with DC United's higher attendance rate. Despite LA Galaxy and DC United having the most championships, the data does not show that attendance rates have a large effect on this. If I were to conduct further research on this subject, I would look at the specific years and conduct the same research on the championship city for that year. I believe that this would better show the relationship between attendance rates, median household income, and Hispanic populations.

Reflection:

After completing this course and this final project, I feel I have learned a lot regarding Spatial Analysis and different ways to use GIS tools. Prior to this semester, I knew very little about coding in relation to GIS and had little experience with using census data. Now I feel confident in my ability to utilize and manipulate census data with code to conduct a spatial analysis. For this specific project, I think I was successful in my ability to analyze census data for multiple locations, but struggled to relate it to professional soccer teams like I had envisioned. This is partly due to the large amount of factors that go into a large scale business like a sports team. As for the class as a

whole, I feel my GIS skills have refined and it has left me wanting to continue learning in the future.

Sources:

HEIM:SPIEL. (n.d.). Major League Soccer 2021 - attendance. worldfootball.net. https://www.worldfootball.net/attendance/usa-major-league-soccer-2021/3/

Rehkemper, G (2023, January 10) USA Soccer. Github repository. https://github.com/gavinr/usa-soccer.

Statista Research Department. (2022, December 9). Share of players in the MLS in 2020 and 2021, by ethnicity. Statista. https://www.statista.com/statistics/1168049/mls-players-ethnicity/

Tolmich, R. (2023, February 21). MLS Cup Winners: Full List of Champions, from D.C. United to LAFC. Goal.com US. https://www.goal.com/en-us/news/mls-cup-all-winners-list/blt15fd45904a53e495

Walker, K. E. (n.d.). Chapter 6 Mapping Census data with r: Analyzing US census data. Chapter 6 Mapping Census data with R | Analyzing US Census Data. https://walker-data.com/census-r/mapping-census-data-with-r.html