**Summary:**

There is a direct relationship between the total amount of income from families with the number of playgrounds available in a zip code. However, there is a weak inverse relationship between the weighted average income of a family and the availability of playgrounds. This is in line with the idea that playgrounds are readily available in populated residential areas usually of a lower to middle socioeconomic class. This could potentially be because these areas might receive more government funding for public projects like parks and playgrounds or are in demand for families to move to. What is evident is that every county in the city has a different relationship with playgrounds because of residential marked areas, government policy and public funding for low income areas.

**Approach:**

Given the data constraints, playground availability will be defined as the number of playgrounds in a zip code. This approach does create some limitations which will be discussed in question 2 of the conceptual questions.

Keeping the limitations in mind, the investigation will take the following steps:

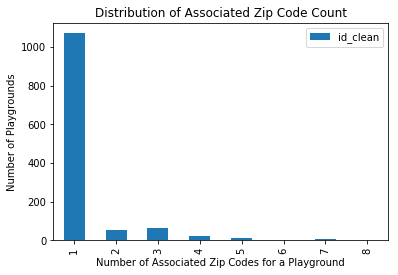
1. The number of playgrounds that are present in every zip code will be counted.
2. One value for every zip code will be determined to represent the AGI of that zip code.
3. These two numbers are joined by zip code as well as grouped at a higher level by the NYC county the zip code is in.
4. The best way to determine a relationship is to see how well one variable can be used to predict the other. For this reason, R^2 values (correlation, in this case, as there are only two variables) will be calculated.
5. The relationship will also be visualized using scatter plots and using geographic heat maps to show the density of parks and wealth across the city.

**Project Description:**

1. Explore and Load the Data

*Playground Data:*

The first step to cleaning the playground data is to associate each playground with at least one zip code. These zip codes are the ones closest to the playground making this playground “available” to all the residents of that zip code. Most playgrounds in NYC are part of a park and have a park ID along with a unique playground ID. Using the parks directory, this park ID can be used to obtain a zip code for a playground. Many of these parks, however, are associated with multiple zip codes and thus one playground could be counted toward more than one zip code. This fits the definition of availability as that park is close to the residents of that zip code. 14% of playgrounds are associated with more than one zip code. Playgrounds that are part of a school have a school ID which cannot be used to connect to the parks file and thus a zip code. The file does provide the coordinates to these playgrounds. In order to obtain a zip code, the coordinates can be reverse geolocated using the Google Maps API. 18% of all playgrounds fall in this category. While it may be true that the school services multiple zip codes, the reverse geolocation process will only give one zip code, creating the assumption that school playgrounds only service one zip code. 1% of playgrounds will not be associated with a zip code because of incomplete data and will be dropped from the analysis. The chart below shows the distribution of zip code counts for a given a park.

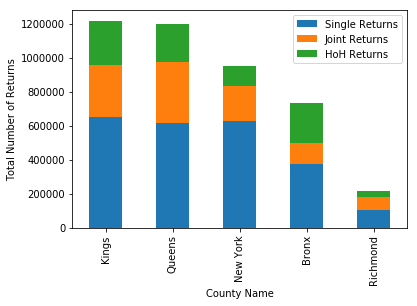
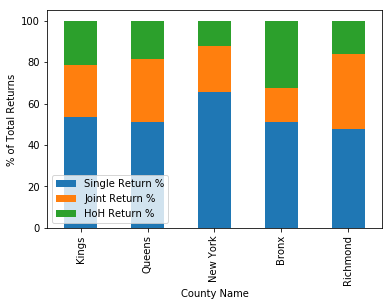


For a portion of this investigation, parks (and playgrounds) that are associated with 3 or more zip codes will be considered outliers and ignored in determining the relationship. The reason for this is that these parks, such as Central Park, are massive in size and therefore readily available to more people outside of the zip code it is associated with. 9% of parks fall in this category.

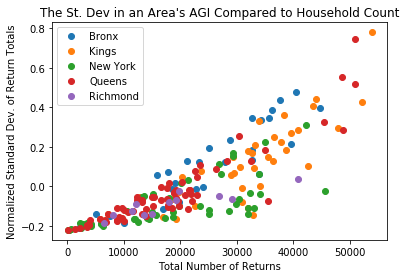
Using a NYC government location directory, playgrounds will be connected to a county. As playgrounds are going to be grouped into zip codes, it is possible for one playground to fall into different counties. For example, playground M037-05B is associated with 5 different zip codes. 4 of these zip codes are Manhattan but one of these zip codes is in the Bronx. 14 (1%) playgrounds fall into two different counties

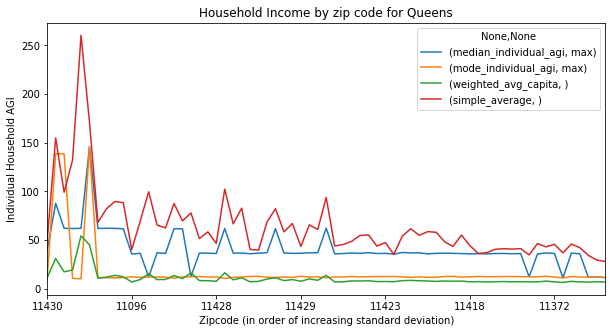
*Income Data:*

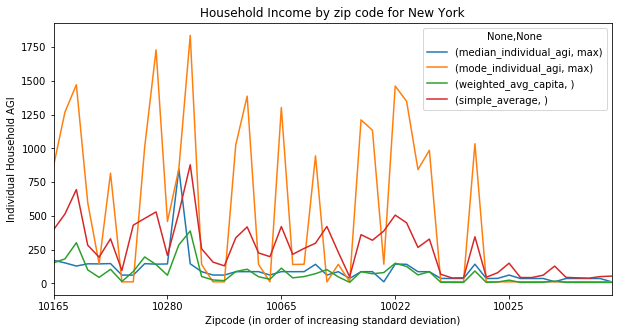
The income level data is at a zip code level. For every zip code in NYC, two types of numbers are provided. The first is the total adjusted gross income reported and the second is the number of tax returns. The number of returns splits into different types of return: the number of single returns, joint returns, and head of household (HoH) returns. It is assumed that each of these returns represent one household and the income coming from each household will be used to measure the relationship between household income and playground availability. In the beginning of the investigation, the total number of returns column will be used to represent the zip code as opposed to adding the three different categories of returns. Summing up the single, joint and HoH yields a number that tends to be four percent off than the total number of returns.

The chart below shows the breakdown of each of these returns across the different counties in NYC. As shown below, the household diversity of each county is very different, especially Manhattan where over 60% of returns filed are single returns. It is for this reason that the relationship between household income and playground availability should also be analyzed at a county level as opposed to all of NYC.

In the income file, the numbers above are split between one of six categories for every zip code. Each category represents a different range of household adjust gross income level (AGI) where 1 represents a range from 0-$25K and then moving up in $25K increments until level 5. This level represents a range of $100K to $200K and the 6th range represents households above $200K. Therefore, looking at the total number of returns from each category for a given zip code gives insight into the population distribution. In fact, comparing the different standard deviations of these six numbers across every area can be a metric to measure the economic homogeneity of each zip code. An interesting point to note is that as the population of a zip code increases, the standard deviation increases as well suggesting that the zip codes become less economically diverse with increasing population.



Lastly, every zip code needs to be represented by one income number. In this case, four different metrics are calculated (described in detail in question 4 of this section). The four include a median AGI, mode AGI, weighted average AGI and a simple average AGI. The metrics tend to be far apart for areas with a more homogenized populace (lower std. dev). The gap in the metrics shrinks, however as the standard deviation increases and an AGI range becomes more heavily weighted. The zip codes in Queens are a clear example of this as shown in the chart below. Simple average is usually consistently the highest household AGI as this metric calculation overweighs larger incomes.

The above trend shown in Queens is consistent with all other counties except for Manhattan where the relationship between standard deviation and population is also opposite to that of the other counties as shown in the scatter chart above. It is most likely because the New York tends to have very economically diverse areas compared to the other counties as shown in the plot above. Given this deviation, it is likely New York will present a much different relationship then the other counties.

1. Completeness of data

One assumption surrounding this data is that it encompasses all of NYC. This makes it a really powerful data set for this investigation since it does not represent just a segment of the population and a relationship can be found amongst the whole city. However, this assumption needs to be verified.

The first thing to check is if all the NYC zip codes are present in the income file. The income file does not include NYC zip codes below a certain population. In total, the income file ignores 120 NYC zip codes with 100 of them being in Manhattan. These are zip codes that incorporate just a few blocks. However, three of these ignored zip codes are also present in the parks data and so these particular parks need to be ignored.

The next thing to show is the population that is represented in the income file is at least complete for these counties. Single returns represent one individual but joint returns would need to be doubled and summed with HoH returns and the number in the dependents column to get the total population that is represented in this data. NYC census data can be used to verify the numbers present in the income table but there are few drawbacks to this. Firstly, the last census was taken in 2010, while this data set is from 2016, therefore the numbers are going to be different. Second, a chunk of the NYC population might not even be included in the income tax data**,** as a portion of this population might not file returns. This includes undocumented residents in the city. There is an estimated 500K residents in the city but it is unknown what percentage of this group are filing returns.

Last, it is important to check if all the playgrounds and parks are accounted for. The dataset only includes government playgrounds that is funded and provided by the city. If the business case of this investigation is only related to playgrounds, then using this set of playgrounds is acceptable. However, if the business purpose of this investigation involves searching for trends such as the availability of land in the city, location trends of parents with disposable income, etc. then the investigation needs to expand past parks and include all publicly accessible recreation spaces such as parks, community gardens, etc.

1. Household Income vs Availability of Playgrounds

The investigation begins by first looking at the correlations between the absolute numbers of total returns and total AGI in the zip code. This gives a relationship between the overall household population of the zip code with the number of playgrounds available. The table below shows there is a fairly high correlation between the total number of returns and the number of playgrounds. This makes sense as more populous zip codes will have more playgrounds.

Next is to determine a metric to describe the family living in these zip codes. In this investigation, the 4 metrics that are used are listed below:

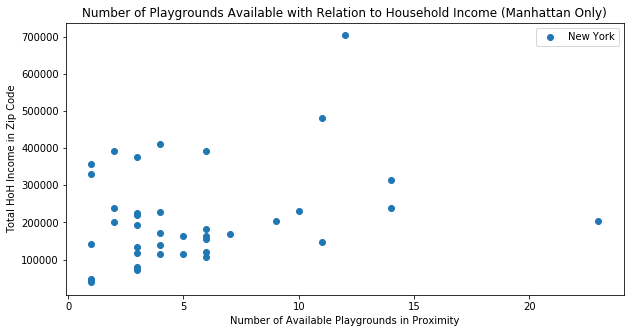
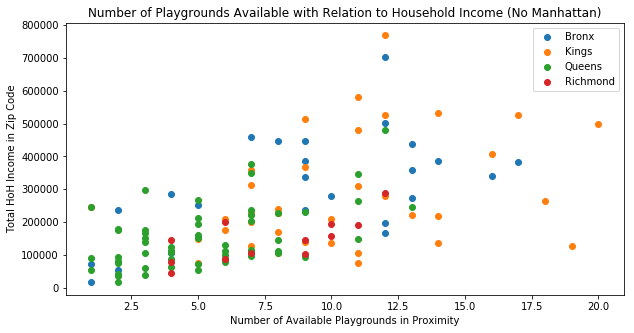
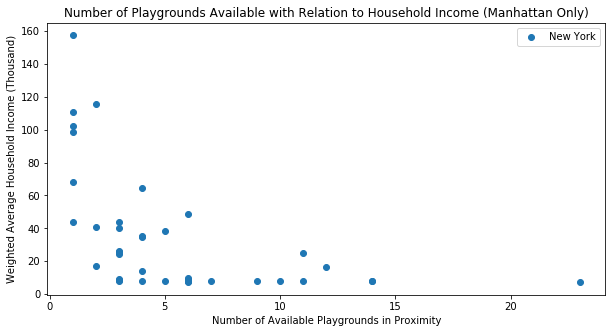
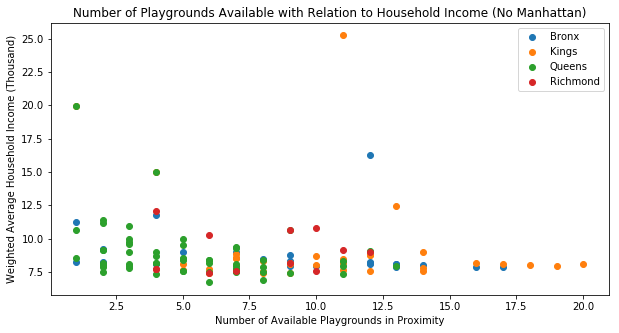
1. Simple Average (per capita) = total AGI/total number of returns
2. Weighted Average (per capita) where the weight is determined by the total number of returns across each of the 6 income categories
3. Mode Income = the simple average income of the category with the most returns for that category
4. Median Income = the simple average income of the category with the median number of returns

The correlations of these figures with the total number of playgrounds is listed below:

Looking at this table, the relationship between an individual household income changes as compared to using absolute. This shows that there is an inverse relationship and it is more geared to where someone lives, while using absolute shows more the relation of population to the number of parks available. The weak inverse relationship suggests that all households with higher net worth tend to live in areas that do not have playgrounds. Part of this reason is that areas of lower income levels tend to have more space and government funding for public projects. This is evident in the Bronx which has a much higher inverse relationship then the other counties.

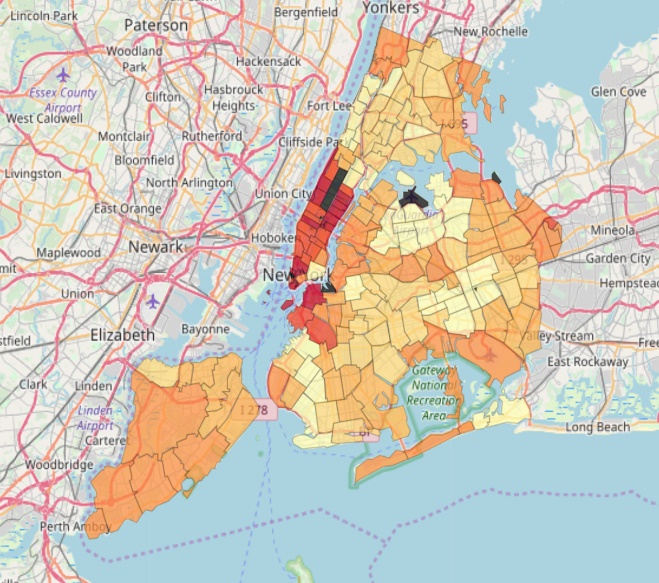
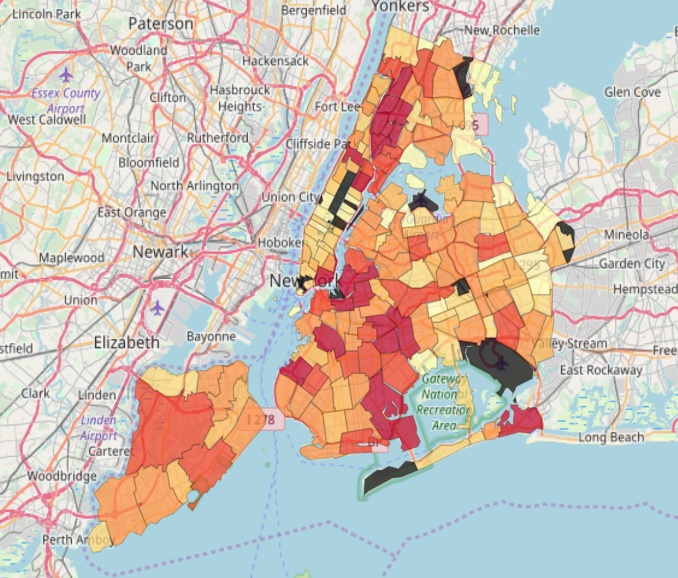
As shown in the standard deviation here, using simple average provides the most variance in correlation. Ideally, this high variance could be considered positive as it means the simple average provides much more information. However, as shown in section 2, the simple average consistently calculates an AGI number higher than the others. The reason for this is that the simple average is extremely sensitive to the large AGI numbers present in the 6th category of AGI (over $200K). Thus, the next best option would then be to use the Median AGI or the Weighted Average AGI. However, looking at the charts comparing the 4 metrics (in question 1 of this section), the Median AGI also tends to be fairly close in value to the Simple Average AGI. The reason for this is that the median value is actually determined by dividing the total amount for the median AGI category with the number of returns for that category meaning variability gets lost. The median AGI is not truly indicative of the range it came from. If all individual returns were present, the median AGI would be the best option, however for this case the weighted average option is the best solution overall. The weighted AGI per household is calculated by dividing over the total returns twice thus eliminating the effect of large dollar amounts of large populations However, it could be the case that the individual characteristics of every county make it that a different metric might be suitable to determine the relationship of that area.

One assumption that was made in the previous correlation calculations was using the total returns and not distinguishing between single, joint and household returns. Below are the correlation comparisons after comparing HoH returns.



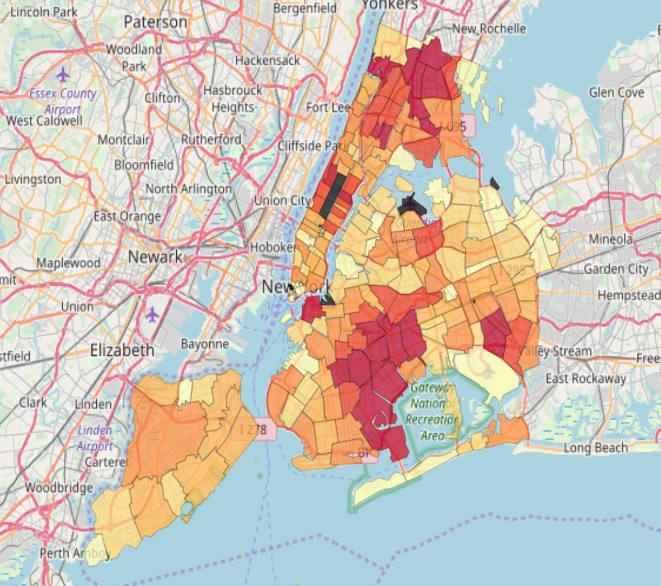
For a majority of the NYC counties and the city as a whole, using HoH income strengthened the direct relationship between the absolute amount of HoH income there is in a given area but at the same time strengthened the inverse relationship an individual family has in that same area. However, for some of these counties such as Staten Island (Richmond) this is not the case. For areas like Staten Island, the number of people and the number of returns was the best indicator.

In summary, there seems to be a relationship between parks and how much money there is from families living in that area. However, there is a weaker inverse relationship between the average income of one these families and the number of playgrounds available to them. This means that areas with a large number of playgrounds have a lot of money coming from those families but there are a lot of families in that area to create that number. This is consistent with the regressions from the beginning of this analysis that showed that larger populations have more playgrounds in them. Part of this reason could be because of because of real estate trends where middle income families want a suburban life but want to live near the city so they live in open areas which are likely to have parks and playgrounds. However, it is certain that every county behaves differently when it comes to the relationship between playgrounds and income. Some counties are residential and thus have more space, while some might get different funding for schools in their area. The next step of this investigation is to into account these county specific characteristics.

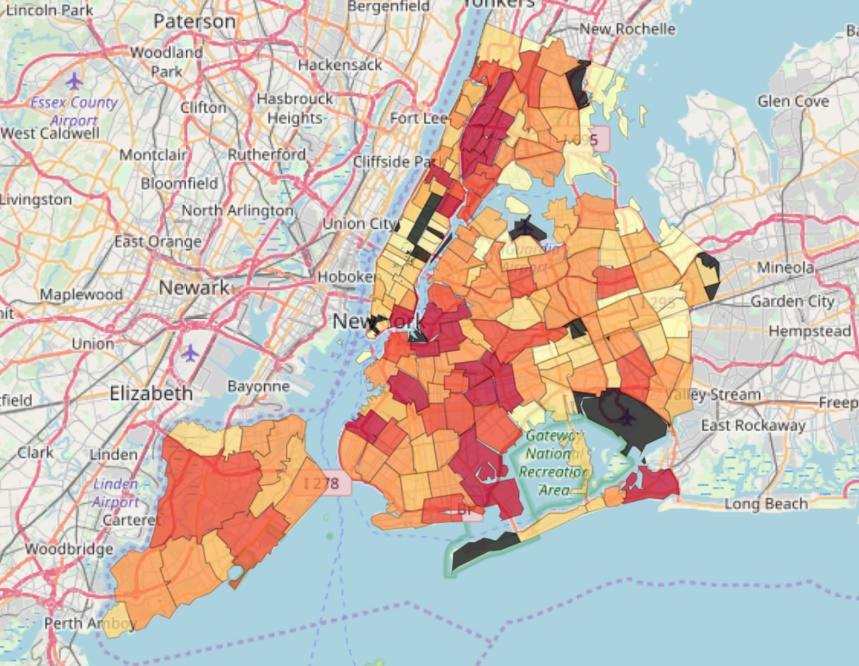
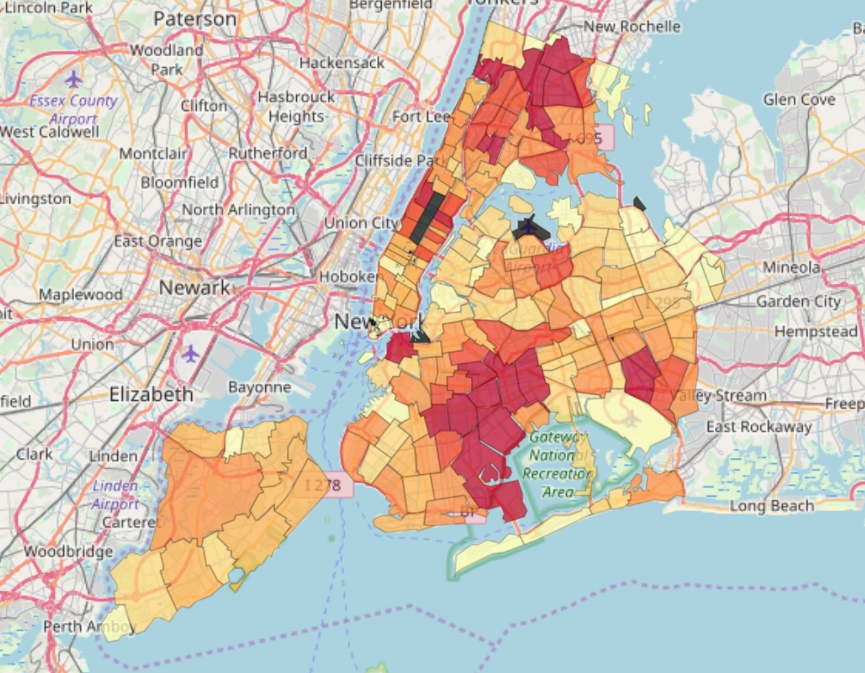


Weighted Average Household Income

Total Household Income Distribution



Park Distribution



Weighted Average Income (HoH) per Family

Total HoH Income in a Zip Code

Number of Parks Distribution in NYC

**Conceptual Questions:**

1. What are some avenues you explored which did yield insights you used in the main write-up? What would you do differently if you could start from the beginning?

The two most helpful avenues were the maps, and the use of HoH returns. The maps provided an interesting visual and context to understand the spread of playgrounds and wealth throughout the city especially given the geographic nature of this problem. Seeing the correlations strengthen after inputting the HoH return values helped strengthen the case for the relationship and showed that this relationship can be determined with better data and more features.

The first aspect to change is to understand the business context around this relationship. This relationship could potentially be valuable in examples such as understanding real estate, consumer movement or even available land trends. But if this is the case, then other sources of data need to be collected as well. In fact, when using number of parks instead of playgrounds, the relationships are actually stronger. By collecting this additional data, then the relationship between different features can help create a null hypothesis with which the significance of the relationship can be measured.

The second aspect is the definition of availability. The definition of availability was determined because of the constraints of the income data. However, given more time, it would be interesting to try to measure zip code availability by some minimum distance to the nearest playground. In order to do this, income data would have to be collected or assumed at a much smaller level then zip code.

1. What biases might you have introduced into your results as a part of your methodology? How might you reduce them with more time/information?

The definition of availability in this investigation was a huge source of bias as it introduced a high-level geographic component that placed data into high level groups. Doing this removed much of the variability the city has and generalizes the relationship between playgrounds and individuals. First each zip code is very economically diverse and encompasses a large area but gets flattened to one number. A park might be on the border of two zip codes but could be frequented by members of the other zip code more because of its proximity to the border, as was the case with school playgrounds. This does not get captured with this definition of availability. Second it introduces a geographical bias which relates all individuals to the same playgrounds even though the individual could be far away so the true nature of the relationship is not captured.

One way to avoid this is to try tweak the definition so it is not at a zip code level but on a smaller scale. The data of the playgrounds provides coordinates and distance to the park which can be used to measure availability. However, in order to fulfill this, it would require income data at a micro level which might not be possible, but a proxy metric could be used instead, such as measuring the rent prices change as distance from a playground increases.

Another source of bias is the fact that the full population might not be captured in the dataset that is assumed to represent the full city. Not capturing everyone diminishes the relationship a particular group might have with playgrounds. For example, it is very likely that a majority of undocumented workers were not captured. It is also likely that these workers are more prevalent in lower income communities of the city and so the extent of the relationship between park availability and these lower income areas is diminished. One potential fix to this is to get data on playground attendance and visits or use the new census data that will be coming soon as part of the 2020 census survey.

1. What are areas of further development of our understanding of this and related relationships?
   1. What other kinds of data would you like to consider?
   2. Are there other kinds of visualizations or models that might be enlightening?

One important piece of data that could be helpful is to get a history of parks, playgrounds and income. It would be interesting to see how a change in income is affecting the change in parks or if parks are increasing with population increase. This information can further help to understand the relationship between household income and parks. A time series regression model could be used in this instance to see how valuable a variable is in predicting the number of parks some area might have.

Second, is to account for policy, taxes and funding changes. Many of these parks are decided by board committee and changes in funding definitely change the decision of these committees and thus change how available the playgrounds might be especially in the future.

Third, is to have more geographic demographics data. More availability of better map visualizations might help find more trends with features that were not expected before, especially since the relationship behaved so differently for every county. To dive into this, it is important to understand what makes every county so different. Principal Component Analysis on these different features describing the counties can help arrive at one metric to describe a county. This metric can then be placed in the ARIMA regression to help build a better model to predict park availability.

Last, is to consider substitute activities in the area. These include parks, but also other recreational activities. There might not be a need for a playground in a certain area if there are other activities. Depending on the business case, these other activities should be considered as alternatives to the playground, especially for household families.

1. How would you go about automating your process to update when new parks are added to the data? How would you detect these updates? What changes might you have to make to your code to do this?

Currently the code brings in data directly from the website link. Assuming the metadata and the link for the most updated data does not change, the data will automatically get pushed to the code if the code is set on a schedule to run consistently. Both of these assumptions, however, are pretty unlikely especially because on the website the metadata date change and the last updated data time are exactly the same time. Therefore, the code will have to be rewritten to include tries and catches to handle an exception if the data is incorrect and not uploading correctly. Part of that try and catch can be an automatically generated email message that is sent when the file is not uploading properly. This try and catch will be able to handle changes in metadata and other technical problems with the link. If, however, the new data gets sent to a brand-new link while this old link is still maintained with the old data, then the code will be unable to recognize a change. This does seem like a very unlikely scenario but if that’s the case then the code will have to use some sort of API that directly connects to the NYC database.