# The Effect of Key Demographic Indicators on the Presence of Secondary Quality of Life Factors in New York City

By Alex Holme

## ABSTRACT

The study utilized zip code census data and New York City government data to examine the relationship between key demographic metrics (median income, population, and total income) and the presence of secondary quality of life amenities. These amenities included flu vaccination sites, licensed businesses, libraries, farmers markets, and businesses started through city programs. The study aimed to determine if wealthier communities in New York City have better access to these amenities, reflecting a broader perception in the US that higher wealth correlates with improved quality of life.

The findings revealed a notable correlation between both population size and total income with the availability of amenities. No correlation between median income and the presence of these amenities was established. This indicates amenities are at least partially democratized as certain amenities most closely tracked population. Other metrics most closely tracked total income revealing certain amenities are more abundant in high income areas. Many key demographic factors such as race, age, and leisure time were not considered in this study, these require further research to come to a more complete understanding of the distribution of amenities.

## 1. INTRODUCTION

In the US, what zip code you live in can have a huge impact on core quality of life factors like access to education, employment, safe housing and a healthy environment. In addition to these key quality of life factors, living location can have a big impact on secondarily important amenities that can impact quality of life like access to public parks, libraries, high quality produce, and proximity to businesses. One advantage advertised by high population density cities like New York is a large amount publicly available amenities to all people regardless of income.

The following study seeks to investigate the impact primary metrics of a community including per capita income, population size, and total income impact community access to quality-of-life boosting amenities in New York City. How much is the presence of these amenities driven by income and how much is driven by overall population?

## 2. DATASET

The following study combines census data to determine key metrics of zip codes in the US including population and median income, measured in dollars per person (including those not in the labour force, such as children or the elderly). The dataset was published on Kaggle and had already been cleaned filtered to only include New York City zip codes. No data cleaning was required, beyond dropping unnecessary columns. From median income and population, total income for each zip code was calculated by multiplying the two values.

Data on amenity distribution by zip code was taken directly from the NYC.gov website. Data on the number of flu vaccination sites, number of actively licensed businesses, number of libraries, number of farmers markets, and number business that were started through city business acceleration programs were considered. All New York City zip codes are five digits, all numeric, and start with the number 10 or 11. Data was filtered to only include zip codes that met the described criteria. Any data with null values entered for a metric was not considered. A Python

Pandas Group By function was performed to find the sum of amenities present in zip codes. After this was performed, each metric was investigated for plausibility. The data presented in this study only includes datasets that were determined contain plausible sums for zip codes.

With the secondary amenity data cleaned and checked for plausibility, it was merged with the primary census metrics. This merge was performed with a Pandas Merge function, where only zip codes contained in the census dataset was considered. If a zip code was only present in secondary amenity data, it was considered a data entry error and dropped. Where no secondary amenity data was present for a zip code, its value was assumed to be 0. For example, if a zip code was not contained in the library dataset, it was assumed that zip code did not have a library.

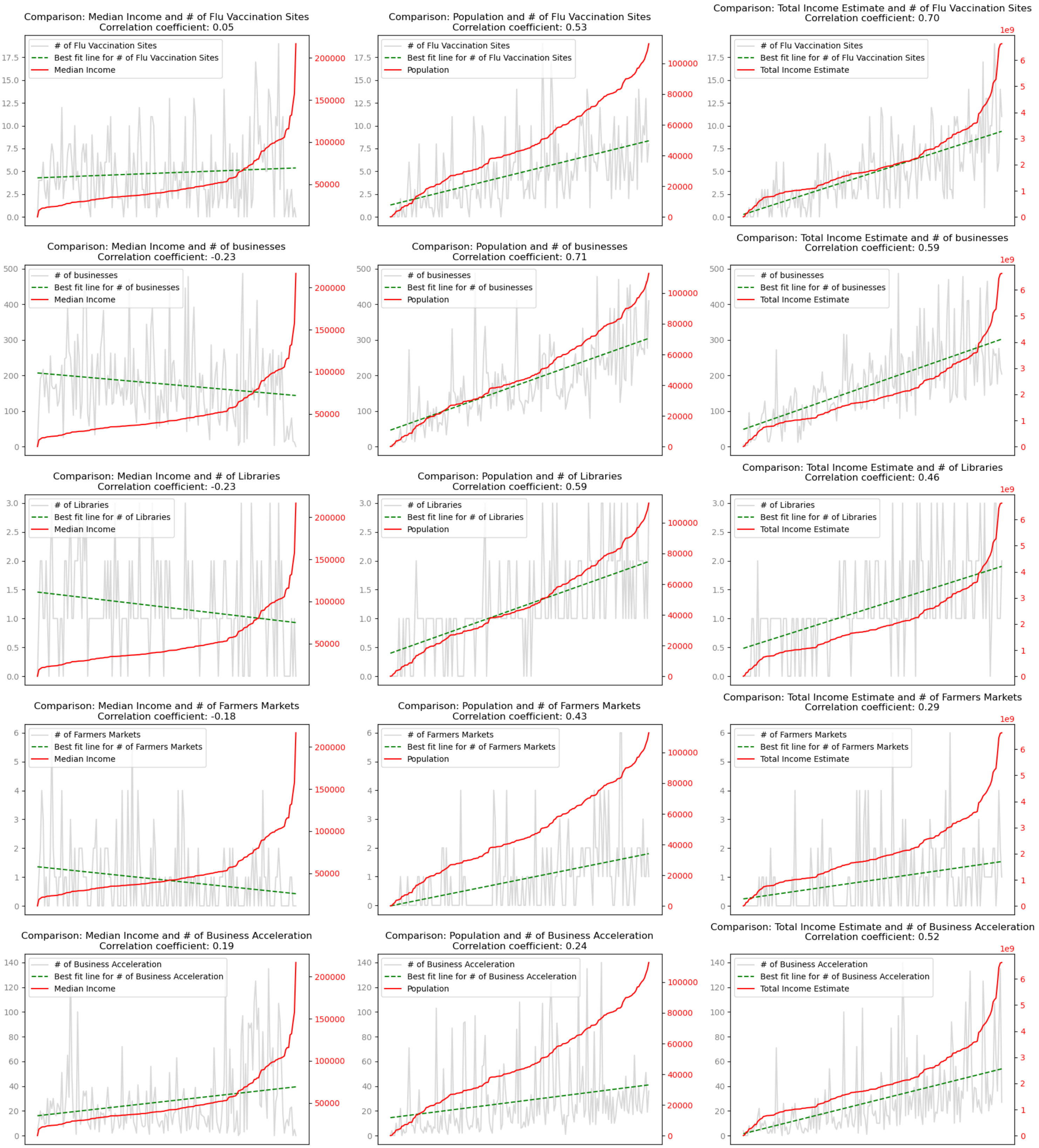
All plots presented in this study are plotted in ascending order of zip codes by their primary metric. For example, any plot featuring population was sorted such that the lowest population zip code was plotted on the far left of the graph and the highest population to the right. The y-values, the secondary amenity metric, considered in that plot would then also be plotted from left to right in ascending order of its primary metric. Plots are also presented with a best fit line of the secondary metric to help understand their general trend. This was calculated using the polyfit and poly1d function included in the Scipy library in python.

## 3. **RESULTS**

The impact key metrics play in the presence of secondary amenities was primarily evaluated on the correlation coefficient observed. The correlation coefficient presented in this study was calculated using the pearsonr function (See equation 1) contained in the Python Scipy library. This function returns a value from -1 to 1, with -1 being a perfectly inverse correlation, 1 being a perfect correlation, and 0 representing no correlation relationship observed between the two metrics considered. The relationship between all primary and secondary metrics is presented in figure 1 (see next page).

$$r = \frac{n(\sum xy) - (\sum x)(\sum y)}{\sqrt{[n\sum x^2 - (\sum x)^2][n\sum y^2 - (\sum y)^2]}}$$

*Equation 1: SciPy Pearsonr Correlation Function, where n is the number of data points, x and y and the individual points. r is the coefficient returned and presented in this study*



Comparison: Median Income and # of Flu Vaccination Sites
Correlation coefficient: 0.05

Comparison: Population and # of Flu Vaccination Sites
Correlation coefficient: 0.53

Comparison: Total Income Estimate and # of Flu Vaccination Sites
Correlation coefficient: 0.70

Comparison: Median Income and # of businesses
Correlation coefficient: -0.23

Comparison: Population and # of businesses
Correlation coefficient: 0.71

Comparison: Total Income Estimate and # of businesses
Correlation coefficient: 0.59

Comparison: Median Income and # of Libraries
Correlation coefficient: -0.23

Comparison: Population and # of Libraries
Correlation coefficient: 0.59

Comparison: Total Income Estimate and # of Libraries
Correlation coefficient: 0.46

Comparison: Median Income and # of Farmers Markets
Correlation coefficient: -0.18

Comparison: Population and # of Farmers Markets
Correlation coefficient: 0.43

Comparison: Total Income Estimate and # of Farmers Markets
Correlation coefficient: 0.29

Comparison: Median Income and # of Business Acceleration
Correlation coefficient: 0.19

Comparison: Population and # of Business Acceleration
Correlation coefficient: 0.24

Comparison: Total Income Estimate and # of Business Acceleration
Correlation coefficient: 0.52

*Figure 1: Correlation plots of all primary and secondary metrics considered in this study. Plots contain best fit line and observed correlation coefficients.*

The strongest correlation observed between all relationships was the correlation between population and the total number of businesses, 0.71. As well as the correlation present between the number of flu vaccination sites, and total income, 0.70. In general, a reasonably strong correlation was observed between population as well as total income with amenities with mean correlations of 0.50 and 0.51 achieved, respectively.

Higher population areas were observed to correlate more with the presence of libraries and the total number of businesses licensed in that zip code achieving correlation coefficients of 0.59 and 0.71 respectively. These are higher than the scores observed between total income, with scores of 0.46 and 0.59 for libraries and businesses.

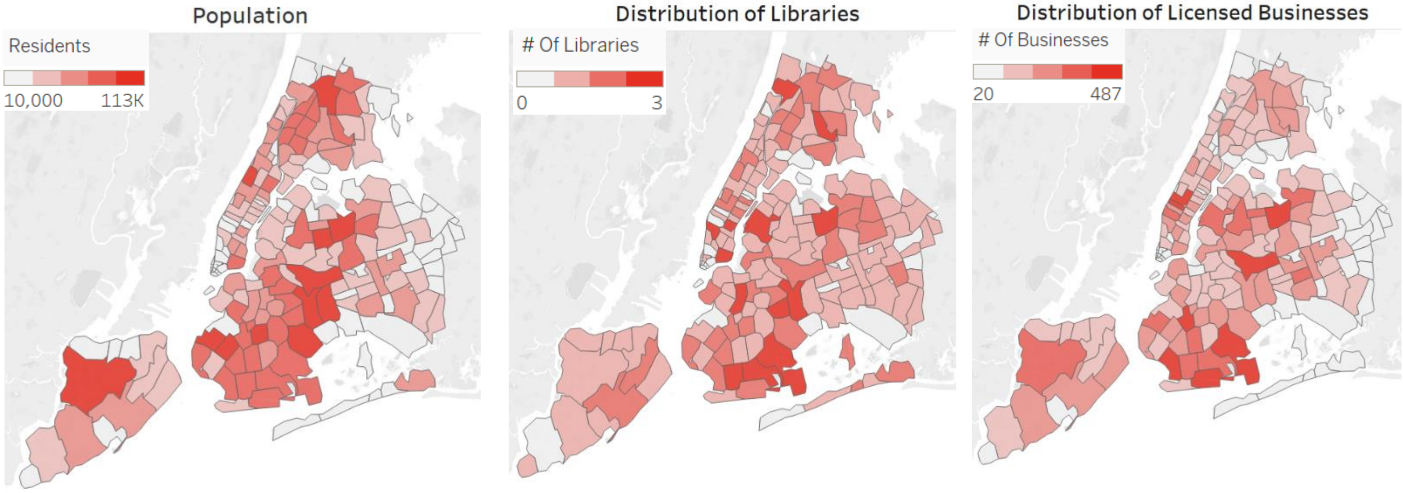


*Figure 2: Distribution of population and amenities with larger correlation coefficients*

Higher total income seemed to be a better indicator of the number of flu vaccination sites and businesses that were founded through city business acceleration programs, with coefficients of 0.70 and 0.52 observed. These are higher than the coefficients observed by population data, which had coefficients of 0.53 and 0.24 respectively. It can be concluded then that presence certain secondary amenities are more heavily correlated with population while some are more heavily correlated with total income. One does not appear to drive the presence of secondary amenities noticeably more than the other.

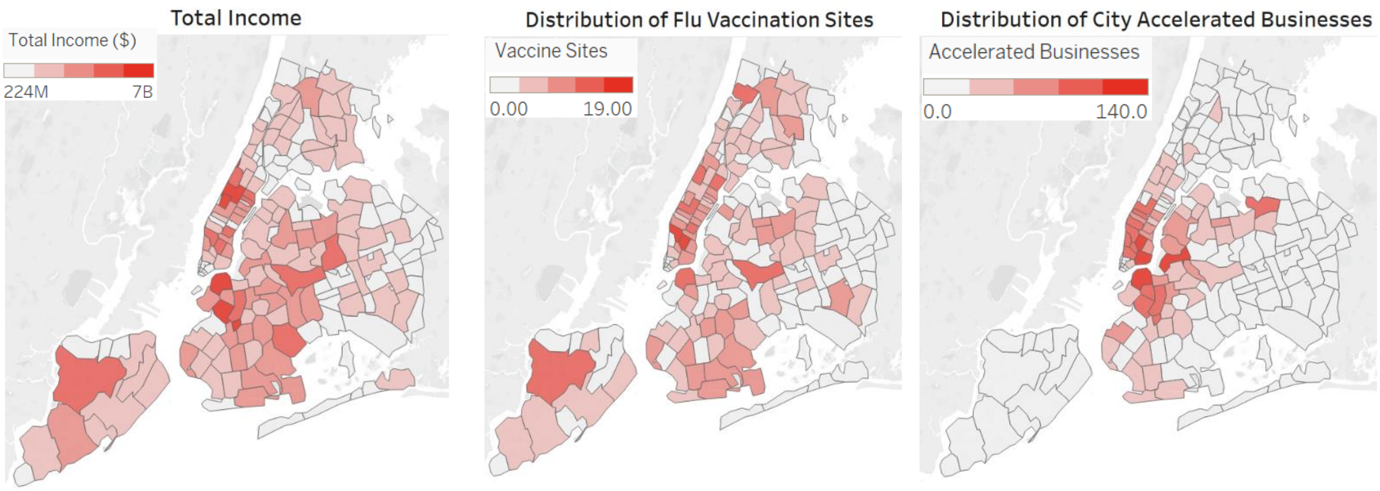


*Figure 3: Distribution of Total Income and amenities with larger correlation coefficients*

Across all secondary metrics considered in this study, almost no correlation with median income was observed, with a mean across all metrics of -0.08. This paired with the stronger correlation with total income would seem to indicate wealthy areas with low populations do not have large amounts of amenities present. From the limited dataset considered in this study it would appear population is a better indicator of how well serviced a community is than median income.

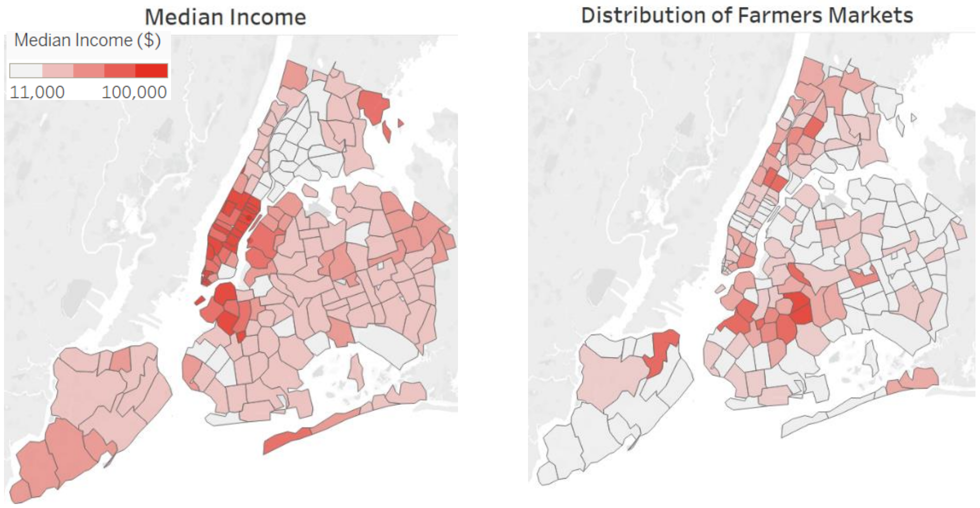


*Figure 4: Distribution of median income and the final amenity farmers markets. Little to no correlation was observed with amenities and median income.*

## 4. **CONCLUSION**

A study was conducted using zip code census data and New York City government published data on the presence of amenities in the different zip codes of New York. The study sought to investigate if key metrics of a zip code, including median income, population, and total income effect the presence of these auxiliary secondary quality of life amenities. The general perception in the US is that higher wealth communities have access to more resources and an improved quality of life. This study sought to see if the same phenomenon was observed in New York City as well.

The secondary metrics chosen for examination in this study were the number of flu vaccination sites, number of actively licensed businesses, number of libraries, number of farmers markets, and number business that were started through city business acceleration programs. These were selected as they are key benefits to living in a city but are often not considered or studied as heavily as more traditional quality of life metrics like access to education, hospitals, employment opportunities, and a healthy environment. A Pearsonr correlation was calculated between these secondary metrics and the key primary demographic indicators considered in this study. The strength of the correlation was the primary factor in determining how demographics impacted the presence of secondary amenities.

No correlation was established between median income and the presence of secondary quality of life amenities. An observable correlation between population and amenities was present. This can serve as an argument that amenities are more democratized and are more present with increases in population than income. However, this is not the full story as a similarly strong correlation between the total income of a zip code and the amenities present was also observed. It would appear certain metrics, like the number of total businesses within a zip code, are more driven by total population, with an observed correlation coefficient of 0.71 compared to 0.59 for total income. While other amenities like the presence of flu vaccination sites was more heavily correlated with total income, with an observed correlation coefficient of 0.70 compared to 0.53 for total population.

Based on the metrics considered in this study, it would appear both population size and total income play a role in the quality of life experienced by a community. Both demographic indicators are significantly more correlated with the presence of amenities than median income. Further investigation is required to paint a more complete picture of how demographic shifts effect quality of life. Factors such as race, age, and leisure time available for residents could all provide more insights into why amenities are distributed how they are in New York City. The effect of these demographic indicators warrants further study in the future.