# The Road to Vision Zero

## Traffic Crashes and Poverty in New York City


TRANSPORTATION ALTERNATIVES

AZAVEA SUMMER of MAPS

# Table of Contents

## The Road to Vision Zero
Traffic Crashes and Poverty in New York City



*Prepared By*

TRANSPORTATION ALTERNATIVES

AZAVEA SUMMER of MAPS



# Executive Summary

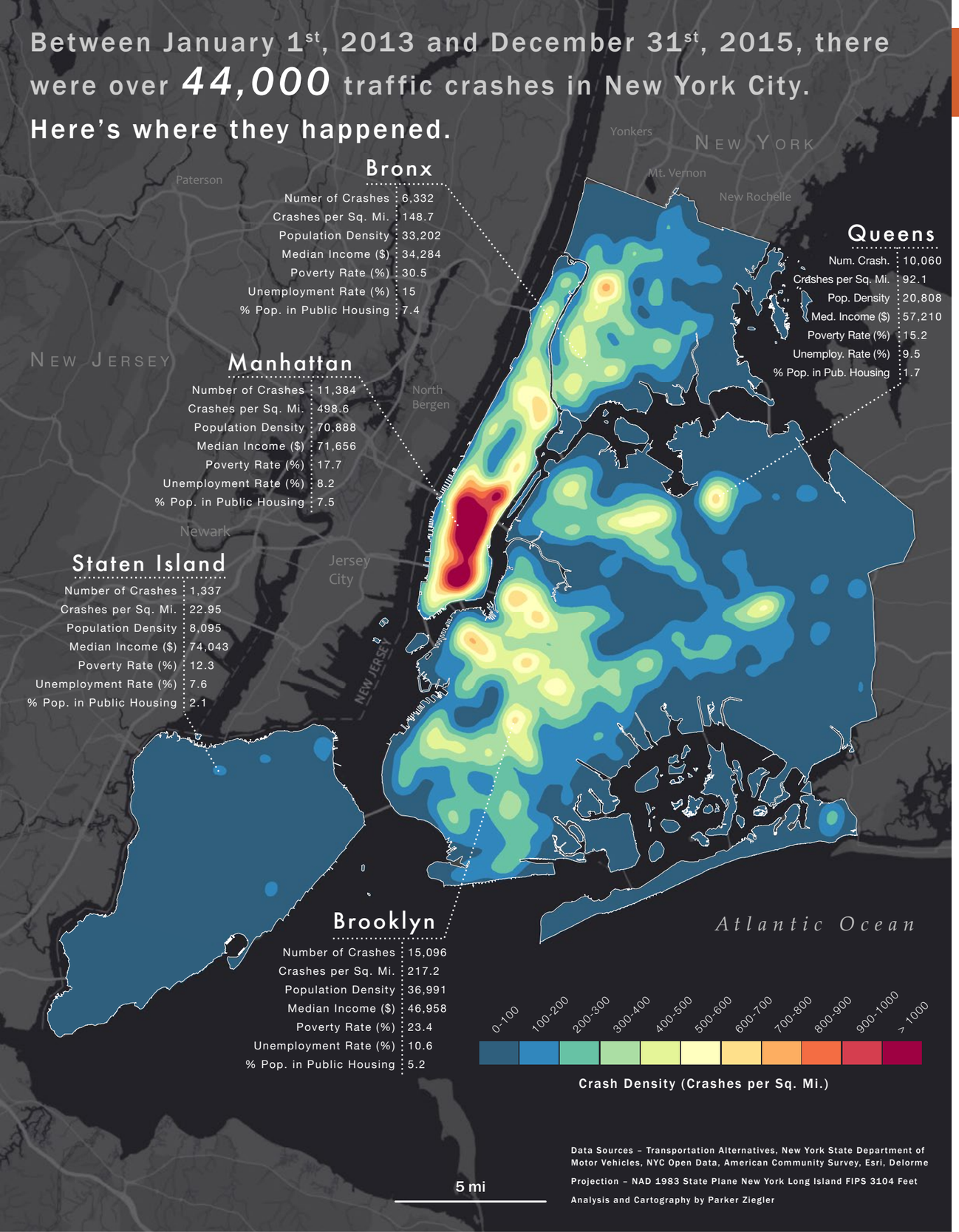
Between January 1, 2013 and December 31, 2015, there were over **44,000** traffic crashes in New York City involving motorists, cyclists, and pedestrians. **Poor urban design**, **a dense built environment**, and **a lack of diversity in transportation infrastructure** continue to make traversing the city dangerous and difficult for residents. However, not all New Yorkers are affected in the same way. Those living in poorer parts of the city are often more susceptible to the negative effects of poor urban planning and city disinvestment. Within this three year time span, **54%** of crashes occurred in city council districts where the poverty rate is **above 15%***. **56%** of crashes occurred in city council districts where the median income is **less than $51,000***. **61%** of crashes occurred in city council districts where the population density is greater than **38,000 people per square mile***. These findings suggest a relationship between poverty and traffic violence in New York City.

In 2014, after significant organizing efforts from community organizations, local businesses, and individuals affected by traffic violence, Mayor Bill de Blasio released the City's **Vision Zero Action Plan**[1], an initiative to end injurious and fatal traffic crashes on New York City streets. The initiative outlined a 63-step programmatic approach to achieving Vision Zero, including **ramping up traffic monitoring initiatives**, **redesigning arterial streets**, **reducing speed limits**, and **increasing penalties for aggressive driving**. However, the Action Plan made no specific mention of examining correlations between traffic crashes and measures of poverty. This report, prepared through a collaboration between Transportation Alternatives and Azavea's Summer of Maps fellowship, seeks to address this gap in the literature through geospatial and statistical analyses of New York City crash data.

The report is divided into four sections – 1) key findings and citywide trends, 2) a discussion of trends within each of New York City's boroughs, 2) directions for future research, and 4) methodology. Within each section are accompanying maps, tables, and graphics designed to help users visualize and understand the data. We believe that data-driven approaches to exploring traffic crashes can help to uncover areas of the city that demonstrate strong relationships between poverty, poor transportation planning, and traffic violence.

At the heart of the Vision Zero design initiative is the belief that **no loss of life is acceptable on our streets.** With appropriate transportation engineering, collaborative urban planning, and active research, traffic crashes can be prevented in New York City's poorest communities.

* Each of these values represents the city median value.




# Key Findings and Citywide Trends

* **Note:** Statistical analysis of citywide trends was performed using data aggregated by Census tracts. Tracts in Manhattan were excluded from the citywide analysis due to the large number of outliers among this dataset; they are treated separately in the **Manhattan** section. For more information, please read the **Methodology** section.

The results of our analysis show that there are statistically significant correlations ($p < 0.05$) between specific measures of poverty and the density of traffic crashes in New York City. Among the socioeconomic and demographic variables surveyed, **population density** displayed the strongest significant correlation (0.530) with crash density, followed closely by **family poverty rate** (0.390), the **number of people in poverty** (0.364), **median income** (-0.363), and **individual poverty rate** (0.288). The **unemployment rate** and the **percent of the population living in public housing** also display significant correlations with crash density, but these were weak in comparison (0.145 and 0.061, respectively).

The sign of each correlation – positive or negative – provides information on the nature of the relationship between each independent variable and the crash density. Population density, family poverty rate, number of people in poverty, individual poverty rate, unemployment rate, and percent of the population living in public housing are all **positively** correlated with crash density, suggesting that an increase in these factors is associated with an increase in the crash density. Alternatively, median income is **negatively** correlated with crash density, suggesting that an increase in median income is associated with a decrease in crash density.

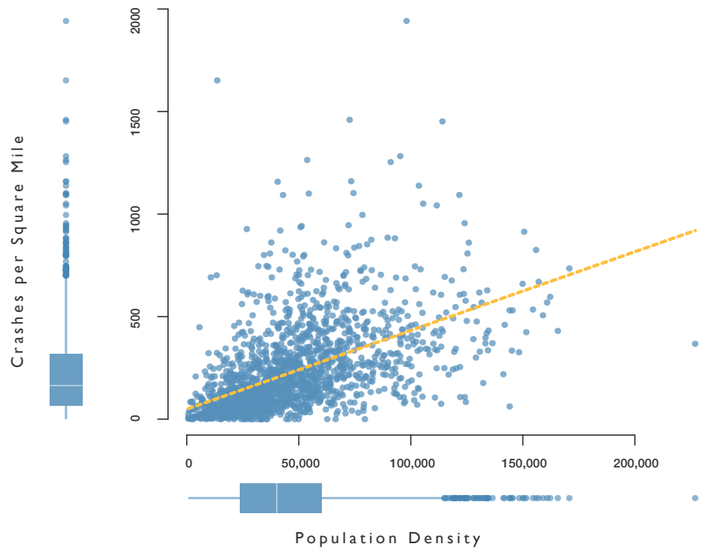



**Figure 1 – Relating Population Density and Crash Density.** Tracts with higher population densities are strongly associated with higher crash densities. They also tend to be poorer; for example, population density and family poverty rate have a statistically significant correlation of 0.485.

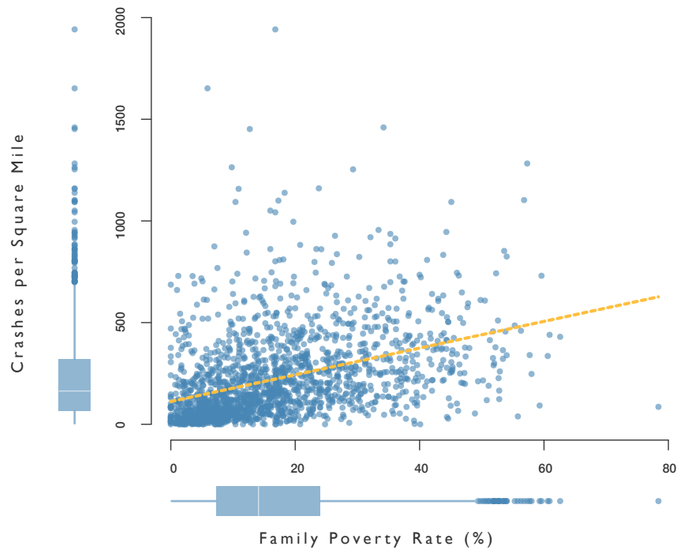



**Figure 2. Relating Family Poverty Rate and Crash Density.** Tracts with higher family poverty rates are strongly associated with higher crash densities. Higher family poverty rates are also strongly correlated with higher unemployment rates (0.782), lower median incomes (-0.884), and a greater percentage of the population living in public housing (0.578).

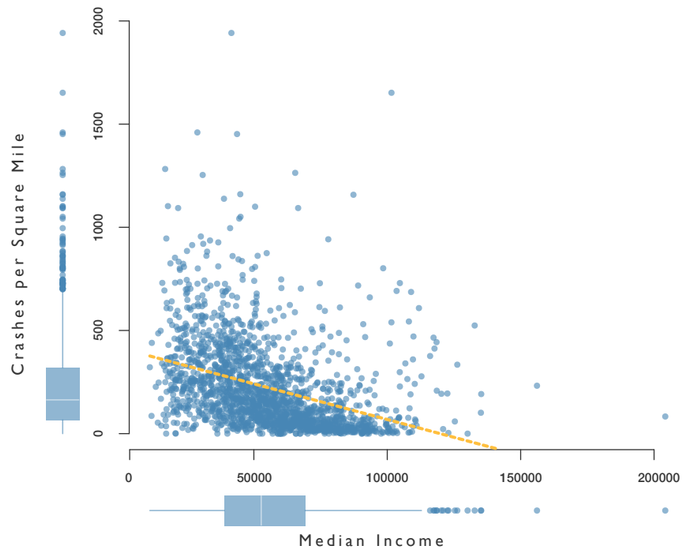



**Figure 3. Relating Median Income and Crash Density.** Increases in median income are strongly associated with decreases in crash density in every borough except Manhattan. Higher median incomes are also associated with lower poverty rates (-0.543), lower unemployment rates (-0.760), and fewer people living in public housing (-0.447).

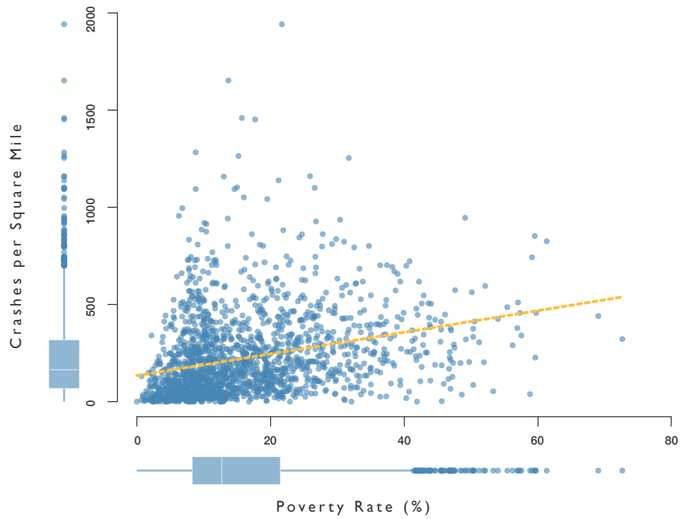



**Figure 4. Relating Poverty Rate and Crash Density.** Increases in individual poverty rate are strongly associated with increases in crash density across New York City. Higher poverty rates are also associated with higher population densities (0.443), lower median incomes (-0.543), and more people living in public housing (0.39).

**Census tracts** with **higher population densities**, higher **individual** and **family poverty rates**, and **lower median incomes** tend to have **higher crash densities**.

**Spatial variables** were also included in the analysis to uncover potential geographic clustering of crashes. **Mean distance to Central Park** was used as an approximation for distance to the city center, and displayed a statistically significant negative correlation with crash density (-0.361). This suggests that the parts of each borough that are closer to the city center are associated with higher crash densities.

In addition to examining relationships between crash density and the observed variables, we also felt it was important to look at relationships between the variables themselves. This process can help in variable selection for regression analysis (see **Methodology**). For example, we observed a strong, statistically significant, negative correlation between **median income** and the **percent of people who use public transportation to commute to work** (-0.425) and a strong, statistically significant, positive correlation between **family poverty rate** and **the percent of people who use public transportation to commute to work** or **walk to work** (0.395 and 0.369, respectively). This suggests that poorer people tend to rely more on alternatives to driving in their daily commute, which can make them more vulnerable to traffic violence. Unfortunately, Census data on transportation does not include any additional information on transportation experiences other than commuting. In this sense, the data is biased against poorer people who suffer from higher unemployment – their experiences of moving about the city are not captured by these statistics.

Ultimately, these findings suggest that poorer areas of New York City dispropotionately suffer from a higher density of traffic crashes. **Figure 5** provides a list of the top five socioeconomic correlates with crash density found in this analysis. **Figure 6** visualizes all statistically significant correlations between every combination of variables surveyed, with the relative strength and sign of each correlation represented by the size and hue of circles, respectively.



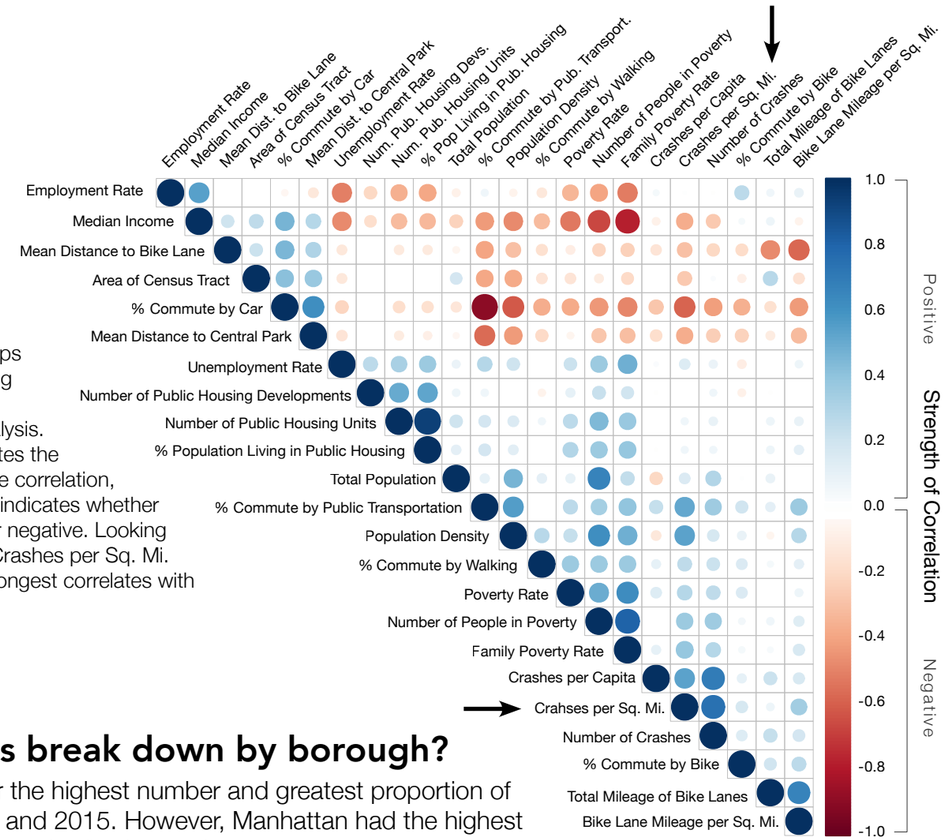
| Variable | Coefficient of Correlation (r) | Significance ($p < 0.05$) |
|---|---|---|
| **Population Density** | 0.530 | $2.22 \times 10^{-16}$ |
| **Family Poverty Rate (%)** | 0.390 | $2.22 \times 10^{-16}$ |
| **Median Income ($)** | -0.363 | $2.22 \times 10^{-16}$ |
| **Poverty Rate (%)** | 0.288 | $2.22 \times 10^{-16}$ |
| **Unemployment Rate (%)** | 0.145 | $2.22 \times 10^{-16}$ |

**Figure 5. The top five socioeconomic correlates with crash density.**

Correlations range from 0 – 1, with 0 representing no relationship and 1 representing perfect correlation. These five variables all display significant correlations with crash density; however, the strength of correlation between population density and crash density, for example, is much stronger than that between unemployment rate and crash density. To learn more about how these values were calculated, see **Methodology.**

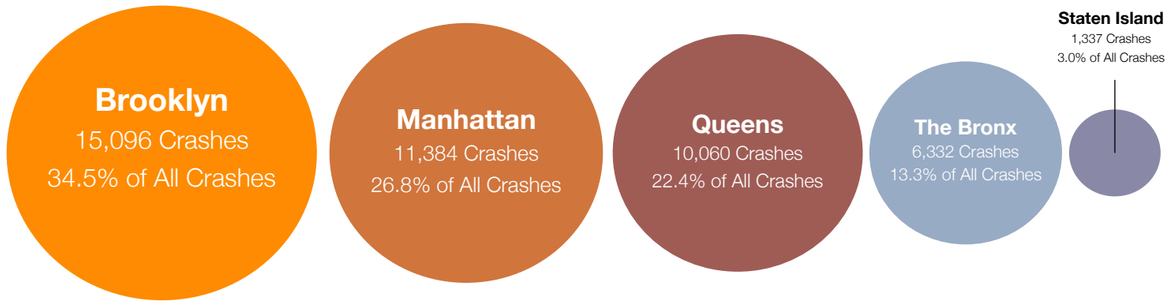
**Figure 6. All significant correlations ($p < 0.05$) between surveyed variables.**

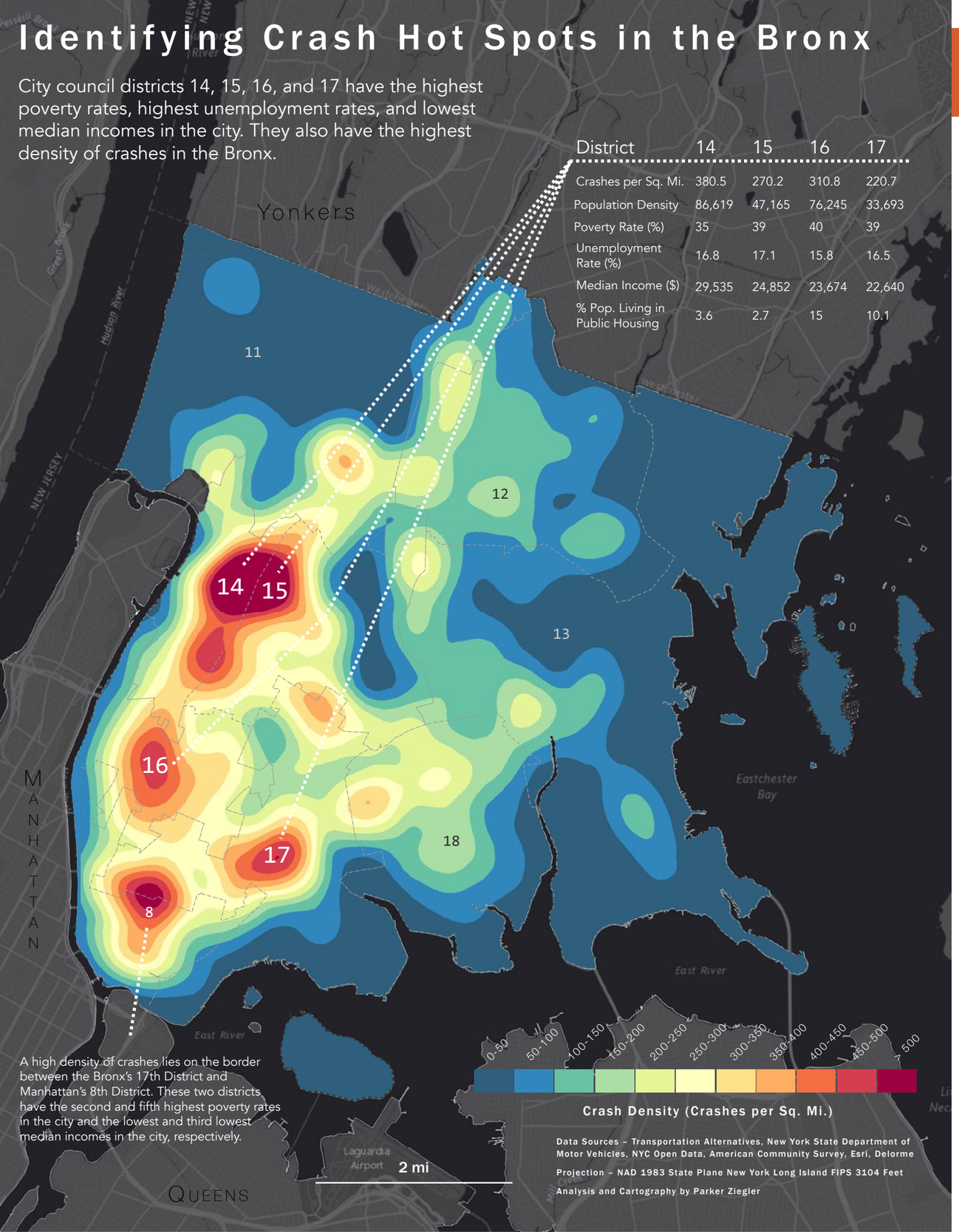
The correlogram at right helps to pick out particularly strong correlations among every variable surveyed in the analysis. The size of the circle connotes the relative strength (0 - 1) of the correlation, while the color (blue or red) indicates whether the relationship is postive or negative. Looking at the row and column for Crashes per Sq. Mi. can help us pinpoint the strongest correlates with this measure.




## How do crashes break down by borough?

Brooklyn accounted for the highest number and greatest proportion of crashes between 2013 and 2015. However, Manhattan had the highest density of crashes iamong the five boroughs.






# Traffic Crashes and Poverty in The Bronx

**13.3%** of all crashes that took place in New York City between 2013 and 2015 happened in the Bronx. The borough has both the **third highest crash density** (148.7 crashes per square mile) and **third highest per capita crash rate** (0.448 crashes per person) in the city, ranking just behind Manhattan and Brooklyn. The Bronx is also the most impoverished of the city's boroughs, with the **highest poverty rate** (30.5%), **family poverty rate** (28%), and **unemployment rate** (15%) in the city. It also has the city's **lowest median income** at $34,284.

Spatially, crashes in the Bronx are clustered in the poorer city council districts closer to Manhattan. **Districts 14**, **15**, **16**, and **17** have the four highest crash densities in the borough at **380.5**, **270.2**, **310.8**, and **220.7** crashes per square mile, respectively. They are also home to the four highest poverty rates, the four highest family poverty rates, the four highest unemployment rates, and the four lowest median incomes in the entire city. In contrast, the Bronx's other city council districts (11, 12, 13, and 18) have markedly lower crash densities. The average crash density among these districts is 106.35 crashes per square mile, 177% less than the average of 295.55 in districts 14, 15, 16, and 17. Similarly the average median income among districts 11, 12, 13, and 18 ($47,996) is 90.6% greater than that among districts 14, 15, 16, and 17 ($25,175). These relationships suggest that poorer communities in the Bronx are disproportionately affected by traffic violence.

Looking at the strength of correlation between different socioeconomic variables and crash density in the Bronx, we find that many of the relationships present at the city level are even more pronounced within the borough. There are strong, statistically significant relationships between **crash density** and **population density** (0.538), **family poverty rate** (0.525), **median income** (-0.515), **individual poverty rate** (0.292), and **unemployment rate** (0.272), all of which suggest a clear correlation between poorer communities and traffic violence in the borough. In fact, these statistical and spatial relationships are stronger here than in any other borough in the city, making the Bronx a particularly important borough for outreach.

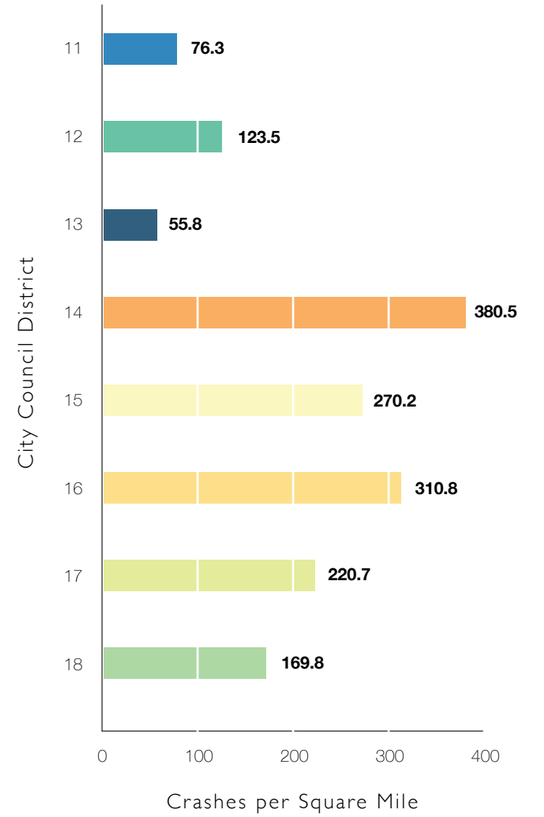
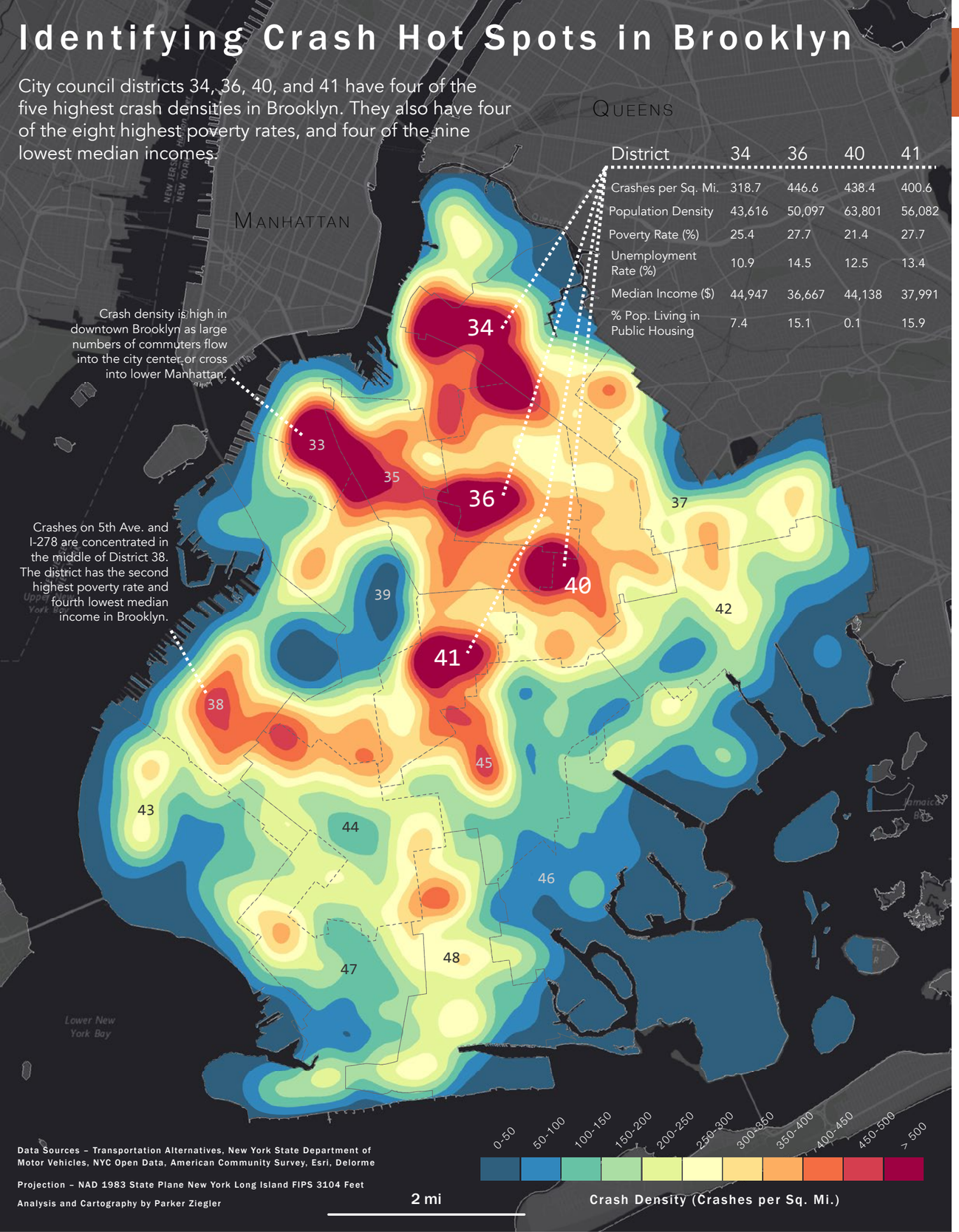



**Figure 8. Distribution of Crash Density Values in the Bronx.** Crash densities are markedly higher in city council districts 14, 15, 16, and 17 when compared to other districts in the Bronx.




# Traffic Crashes and Poverty in Brooklyn

**34.5%** of all crashes that took place in New York City between 2013 and 2015 happened in Brooklyn. The borough was home to **the greatest number of crashes** in this time period – **15,096** in total. At **217.2** crashes per square mile, its crash density is the second highest in the city behind Manhattan. Its per capita crash rate is also second at **0.587** crashes per person. Economically, Brooklyn is an extremely diverse borough with median incomes ranging from the mid-$30,000s in Bedford Stuyvesant, Brownsville, and East New York to the low-$90,000s in downtown Brooklyn and Park Slope. Rapid, aggressive gentrification in Brooklyn has further increased wealth disparity and raised both the population density and the commuter density in the borough.

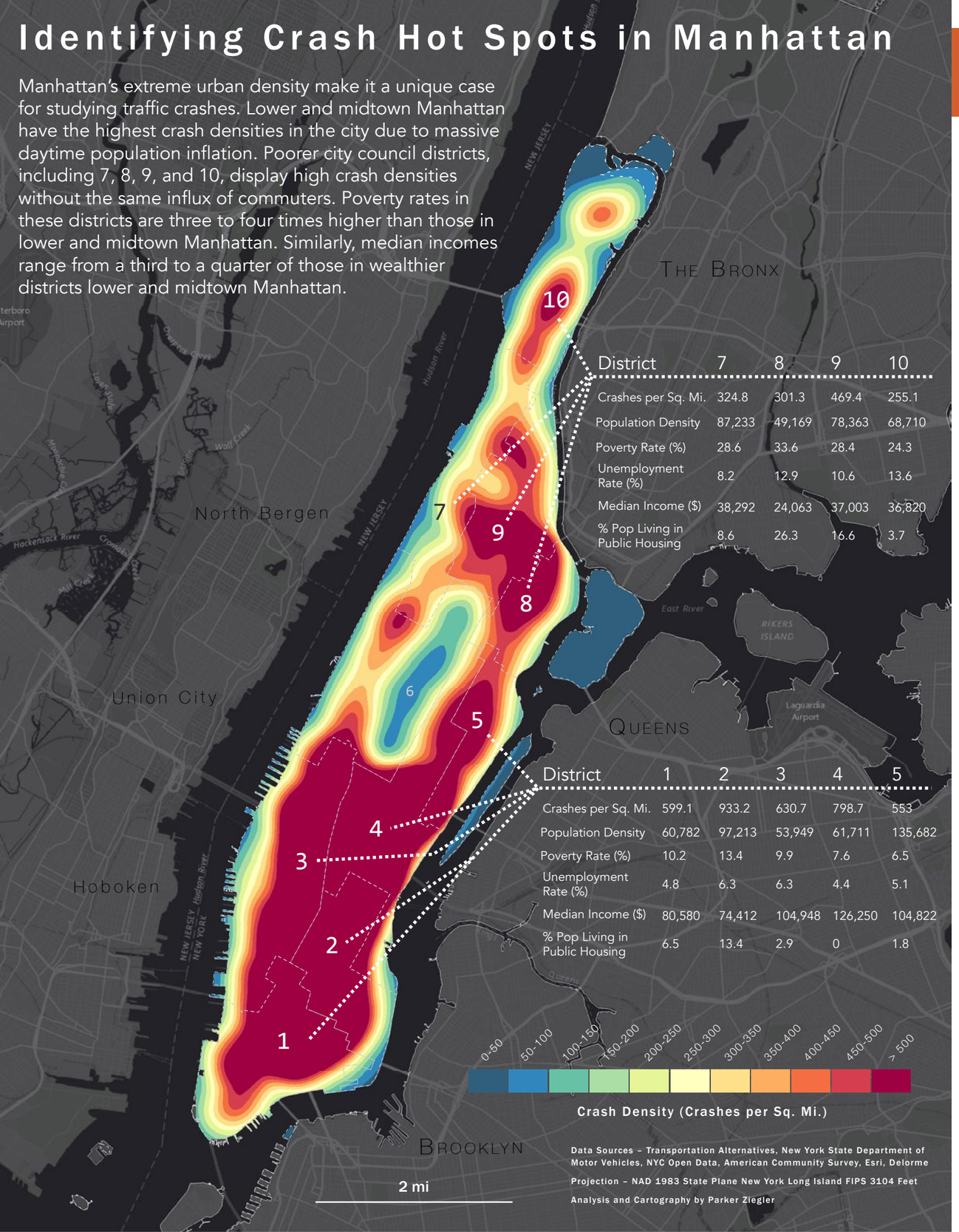
Spatially, crashes in Brooklyn are concentrated in a few major areas. **Districts 34**, **36**, **40**, and **41**, corresponding roughly to Bushwick, Bedford Stuyvesant, Brownsville, and Flatbush, have four of the five highest crash densities in the borough at 318 .7, 446.6, 438.4, and 400.6 crashes per square mile, respectively. They also display four of the six highest population densities, four of the eight highest poverty rates, and four of the nine lowest median incomes in the borough. This area of Brooklyn is also densely populated, with each of these districts having over 43,000 people in each square mile. As they continue to change in the midst of gentrification and redevelopment, additions of even more commuters further crowds streets in these districts.

Downtown Brooklyn also represents a major hotspot of crash activity, with a crash density of **293** crashes per square mile in **district 33**. Crashes here are likely due to the large daytime swell in commuters entering lower Manhattan. In fact, Brooklyn has the second largest daytime population decline of any county in the United States, only trailing its northern neighbor, Queens.[2] With over 500,000 people leaving the borough each day to reach Manhattan, many of them passing through downtown Brooklyn, the probability of crashes in this area is higher than other parts of the borough. Finally, **district 38** represents an interesting case for Brooklyn. Although the population density is the third lowest in the borough, there is a noticeable hotspot in the center of the district corresponding to the Sunset Park neighborhood. The poverty rate here, at 30.3%, is the second highest in Brooklyn, while the median income of $39,122 is the fourth lowest in the borough.

Looking at the strength of correlation between different socioeconomic variables and crash density in Brooklyn, we find that they do not differ considerably from that observed in the city as a whole. There remain statistically significant correlations between **crash density** and **population density** (0.361), **mean distance to Central Park** (-0.340), **individual poverty rate** (0,242), **family poverty rate** (0.215), **median income** (-0.173), and **unemployment rate** (0.100); however they are weaker in some cases than the city level correlations and considerably weaker than those observed in the Bronx.

# Identifying Crash Hot Spots in Manhattan

Manhattan's extreme urban density make it a unique case for studying traffic crashes. Lower and midtown Manhattan have the highest crash densities in the city due to massive daytime population inflation. Poorer city council districts, including 7, 8, 9, and 10, display high crash densities without the same influx of commuters. Poverty rates in these districts are three to four times higher than those in lower and midtown Manhattan. Similarly, median incomes range from a third to a quarter of those in wealthier districts lower and midtown Manhattan.




Data Sources – Transportation Alternatives, New York State Department of Motor Vehicles, NYC Open Data, American Community Survey, Esri, Delorme

Projection – NAD 1983 State Plane New York Long Island FIPS 3104 Feet

Analysis and Cartography by Parker Ziegler

# Traffic Crashes and Poverty in Manhattan

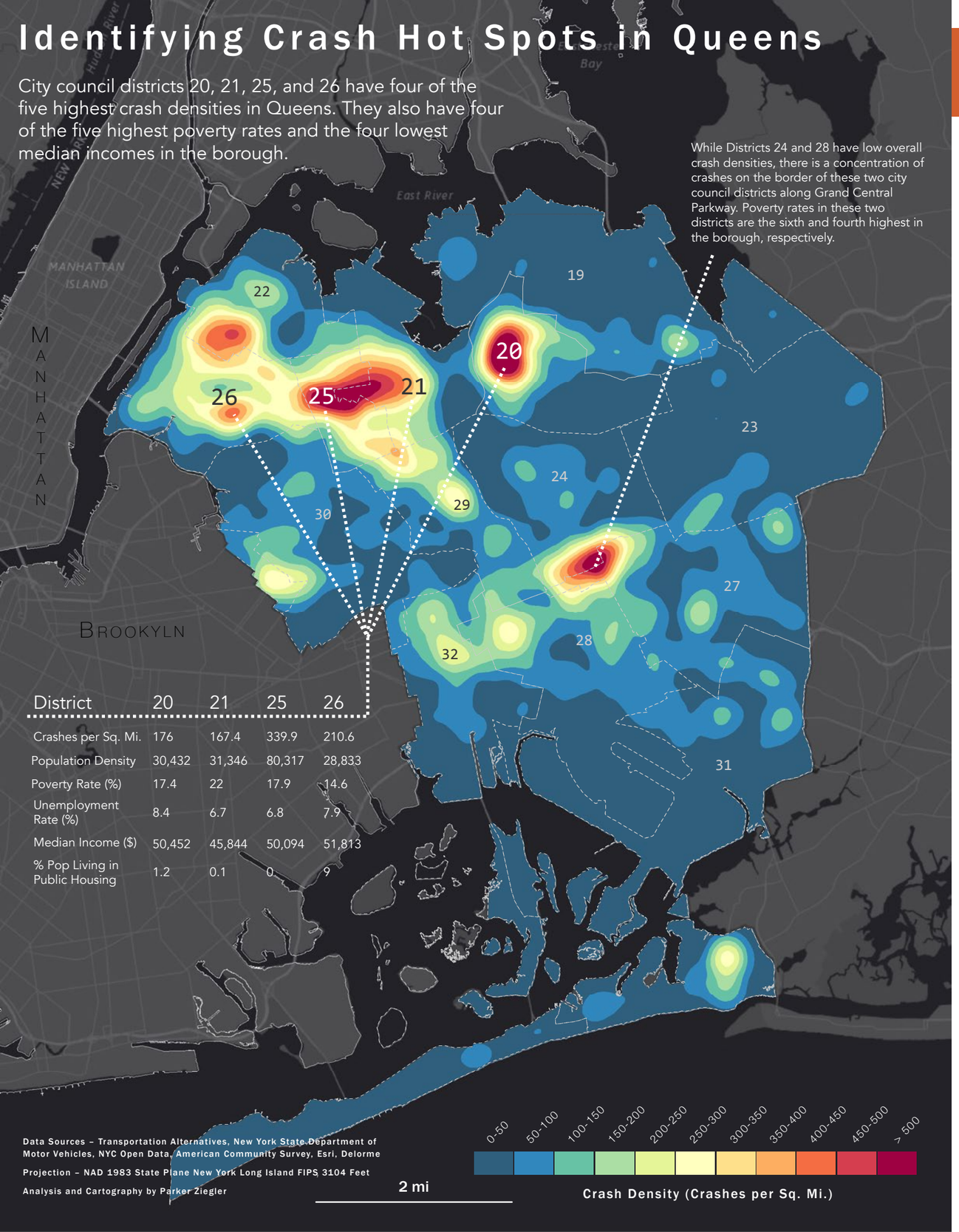
**26.8%** of all crashes that took place in New York City between 2013 and 2015 happened in Manhattan. While the borough was second in the city by number of crashes – **11,384** in total – it has by far the highest crash density among the five boroughs. At **498.6** crashes per square mile, its crash density is roughly **130% greater** than the second highest, Brooklyn (217.2). Similarly, the per capita crash rate of **0.703** is higher than any other borough, although by a smaller margin (19% greater than Brooklyn's 0.587).

Manhattan is the most densely populated county in the country.[3] Among its city council districts, only District 8 has under 50,000 people per square mile. Most districts fall somewhere between 60,000 and 90,000 people per square mile, with the 5th district taking highest in the city at 135,681 people per square mile. Manhattan also displays extreme economic variability among its city council districts. The difference between the maximum and minimum median incomes of Manhattan districts is $102,187, between the fourth (maximum) and eighth (minimum) districts. Similarly poverty rates range from 6.5% in the fifth district to 33.6% in the eighth district. This level of extreme economic disparity within the borough makes theorizing relationships between traffic crashes and poverty more complex.

Within the borough and the city as a whole, **districts 1 – 5** display the highest crash densities at **599.1**, **933.2**, **630.7**, **798.7**, and **553.0** crashes per square mile, respectively. These districts, comprising Lower and Midtown Manhattan as well as the Upper East Side, are some of the wealthiest in the city; the average median income among them is $98,256 while unemployment rests at just 5.4%. They also experience extreme increases in their daytime population, with most of Manhattan's 1,500,000 daily commuters heading to work in Lower and Midtown Manhattan.[2] This near doubling of daytime population likely contributes to the high crash densities observed in this part of the city.

**Districts 7 – 10** also display high crash densities in comparison with the rest of the city at **324.8**, **301.3**, **469.4**, and **255.1** crashes per square mile. Comprising Harlem, Washington Heights, and parts of the South Bronx, these districts are some of the poorest in the city. The average median income in these districts is just $34,045 with average unemployment at 11.3% and the average poverty rate at 25.1%. Unlike districts 1 – 5, districts 7 – 10 do not experience the same magnitude of daytime population increase, suggesting that higher crash densities here may be more connected to extreme population densities and poor transportation planning.

Looking at the strength of correlation between different socioeconomic variables and crash density in Manhattan, we find that relationships between crash density and poverty tend to behave inversely compared to other boroughs. For example, increases in population density, individual poverty rate, family poverty rate, and unemployment rate are all associated with decreases in the crash density (-0.135, -0.147, -0.185, -0.249, respectively), while increases in median income are associated with increases in the crash density (0.190). These results suggest that connections between traffic crashes and poverty in Manhattan operate according to different mechanics in the wealthier and poorer parts of the borough, and that crashes are being caused by distinct factors in these areas.




# Traffic Crashes and Poverty in Queens

**22.4%** of all crashes that took place in New York City between 2013 and 2015 happened in Queens. With **10,060** crashes in this time period, the borough had the third highest number of crashes among the five boroughs. However, being the largest borough by area in the city, Queens had a relatively low crash density – **92.1** crashes per square mile, higher only than Staten Island. Similarly, the per capita crash rate of **0.441** was the fourth highest in the city. Economically, Queens is less stratified than other New York boroughs. Median incomes among city council districts range from a minimum in the low $50,000s to a maximum in the low $70,000s. Poverty rates tend to hover below 12% while unemployment rates fall under 10% except in three districts. However, Queens displays highly variable population density across the borough. More suburban districts have less than half the number of people per square mile (13,328 on average) as the more urban districts closer to Manhattan (29,819 on average). **District 25** is a sharp outlier with 80,317 people per square mile.

Spatially, crashes in Queens are clustered in the denser districts closer to Manhattan. **Districts 20**, **21**, **25**, and **26** have four of the five highest crash densities borough. They also have four of the five highest population densities, four of the five highest poverty rates, and the four lowest median incomes in the borough, all of which suggest a relationship between crash density and poverty. Crashes in these denser districts are also likely a result of the massive daytime population decrease Queens experiences as commuters head to work in Manhattan. In fact, Queens has the highest daytime population decrease of any county in the country, losing approximately 600,000 people every morning.[2] Many of them move through these more urban districts. One outlier detected in the analysis is a high concentration of crashes on the border of **districts 24** and **28**, near Grand Central Parkway. Poverty rates in these two districts are the fourth and sixth highest in the borough, respectively.

Looking at the strength of correlation between different socioeconomic variables and crash density in Queens, we find that relationships between crash density and poverty tend to be even stronger than those observed across the city as a whole. There are strong, statistically significant relationships between **crash density** and **population density** (0.601), **median income** (-0.408), **individual poverty rate** (0.404), **mean distance to Central Park** (-0.400), and **family poverty rate** (0.394). These correlations all suggest a higher incidence of traffic crashes in the poorer, more densely populated districts of Queens.

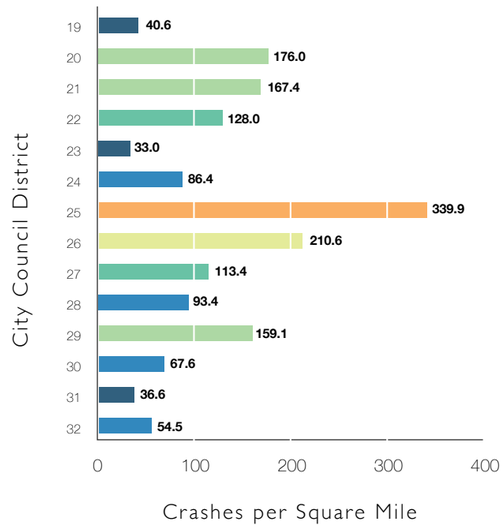
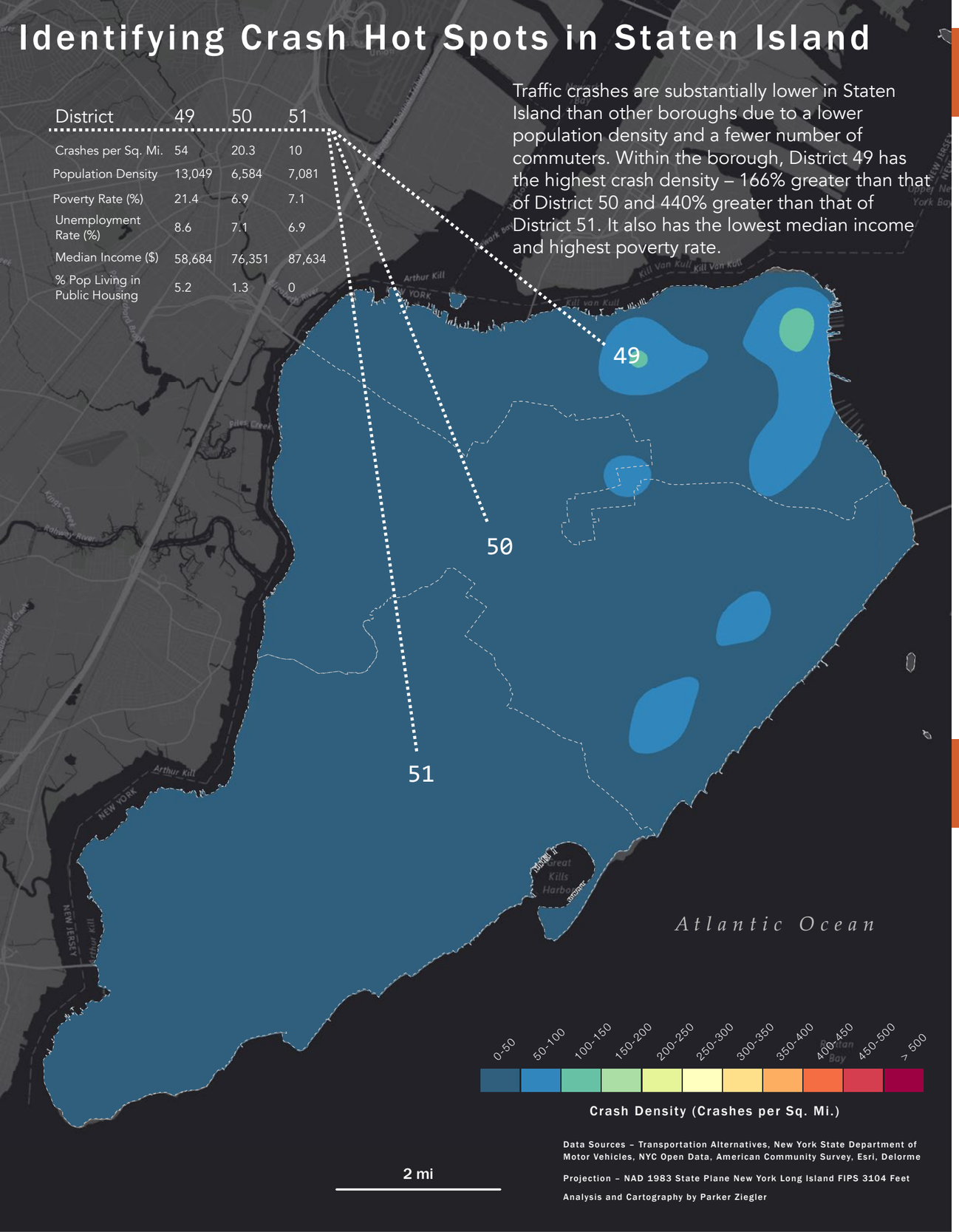



**Figure 9. Distribution of Crash Density Values in Queens.** Crash densities are more or less consistent in Queens with the exception of District 25. Its crash density of 339.9 is 61% greater than the second highest, District 26.




# Traffic Crashes and Poverty on Staten Island

**3%** of all crashes that took place in New York City between 2013 and 2015 happened in Staten Island. With just **1,337** crashes in this time period the borough had the fewest in the city. Its crash density of **22.95** crashes per square mile and per capita crash rate of **0.284** placed Staten Island as the lowest in both categories among the city's boroughs. Staten Island is also the least populous (471,522 people) and least densely populated borough (8,095 people per square mile) in the city. Economically, Staten Island has the highest median income, lowest individual poverty rate, lowest family poverty rate, and lowest unemployment rate in the city.

However, there is clear stratification among its three city council districts. **District 49**, closest to Manhattan, has a substantially higher poverty rate (21.4%) than either of its neighbors (6.9% and 7.1% for **Districts 50** and **51**, respectively). It also has a lower median income ($58,684 vs. $76,351 and $87,634), higher population density (13,048 vs. 6,584 and 7,081) and higher unemployment rate (8.6% vs. 7.1% and 6.9%). Spatially, crashes are also clustered in District 49; its crash density of **54.0** crashes per square mile is 166% greater than that of District 50 (**20.3**) and 440% greater than that of District 51 (**10.0**).

Looking at the strength of correlation between different socioeconomic variables and crash density in Staten Island, we find that relationships between crash density and poverty tend to be even stronger than those observed across the city as a whole. There are strong, statistically significant relationships between crash density and **mean distance to Central Park** (-0.490), **median income** (-0.432), **family poverty rate** (0.423), **individual poverty rate** (0.415), and **population density** (0.256). These correlations all suggest a higher incidence of traffic crashes in the poorer, more densely populated districts of Staten Island.

# Directions for Future Research

**Ultimately, this study found strong evidence for the existence of both spatial and statistical relationships between traffic crashes and poverty in New York City.** In general, areas of the city with higher population densities, lower median incomes, higher individual and family poverty rates, and higher unemployment rates had higher crash densities. This finding suggests that poorer communities in New York City are statistically more vulnerable to traffic violence.

However, continued analysis is essential to uncovering the causal factors behind this trend. What, specifically, is leading to a higher density of traffic crashes in poorer neighborhoods? Dilapidated roads, neglected repairs and maintenance of traffic signs, a lack of safe transportation options for pedestrians? Uncovering factors in the built landscape that lead to an increased probability of crashes is critical to counteracting traffic violence.

Directions for future research are many and diverse. A significant step forward would involve collecting and analyzing more comprehensive transportation data reflecting a diversity of

transportation experiences. The transportation variables in this study were limited in that they only captured commuting experiences. Another improvement could be made by normalizing the crash data by a variable other than area or population. Using normalizing variables like total daytime population or daytime traffic volume would help to temper the effect of outliers like lower and midtown Manhattan, where a massive daily influx of people dramatically increases the probability of a crash. Finally, exploring how traffic crashes and poverty are correlated, or perhaps even causal, across space requires the development of more spatially explicit regression models. The negative binomial model uncovered in this analysis (see Methodology) is an important first step, but application of additional spatial statistical techniques like geographically-weighted regression (GWR) may help to deal with the spatial autocorrelation present in most poverty data in urban areas.

As long as crashes remain a part of urban life, the work of uncovering where they happen, why they happen and how they can be prevented must continue.

# Methodology

***This study used a combination of geospatial and statistical methods for examining relationships between traffic crashes and poverty in New York City.***

## *Geospatial Methods*

The geospatial analysis for this project was carried out using ESRI's ArcGIS software. Geocoded data on traffic crashes was provided by Transportation Alternatives as a comma-separated value (CSV) file. Socioeconomic and demographic data was sourced from the U.S. Census Bureau's American Community Survey (ACS) 5-year Estimates for 2014. The 5-year Estimates were chosen because they provide the largest sample size and are considered the most reliable by the U.S. Census Bureau at fine geographic scales (the Census tract level). Spatial geometries were retrieved from NYC Open Data for boroughs, city council districts, community districts, public use microdata areas (PUMAs), police precincts, ZIP codes, and Census tracts.

To begin the analysis, crash data and spatial geometries were reprojected into a common coordinate system (NAD 1983 StatePlane New York Long Island FIPS 3104 Feet). Crash data was then spatially aggregated within each geometry using the Spatial Join geoprocessing tool with an Intersect match option. Next, ACS data was joined to spatial geometries for which it is aggregated using a shared primary key (Geo.id2); this included Census tracts, ZIP codes, PUMAs, and boroughs (counties). For those spatial geometries by which the Census does not aggregate data, a separate method was used. Centroids were extracted from each Census tract with all attached attributes. Then, these centroids and their attributes were aggregated according to their spatial relationship with the other geometries of analysis (city council districts, community districts, and police precincts). This process effectively provided, for every geometry of interest, a count of crashes alongside all selected socioeconomic, demographic, and transportation variables. A few simple field calculations produced a crash density (crashes per square mile) and a per capita crash rate (crashes per person). These tables were then exported as CSVs to be used in further statistical analysis (see below) and provided to Transportation Alternatives.



The final piece of geospatial analysis for this project involved generating kernel densities of traffic crashes at both the city level and at the level of each borough. No population field was specified to ensure that each crash was only counted once. The search radius was specified at 0.5 miles (2640 ft.) with area units in square miles and output values as densities. The geographic method was specified as planar, as the State Plane projection preserves distance at high precision across the study region. Output rasters were set to 10 ft. resolution to provide a high level of detail. Finally, densities were classified at intervals of 100 crashes per square mile for the city level and 50 crashes per square mile for the borough level. Keeping the classification schema consistent across all five boroughs was important for ensuring direct visual comparison across the borough maps. Design work for these maps was performed using Adobe Illustrator.

## *Statistical Methods*

The statistical analysis for this project was performed using the R programing language[4] alongside RStudio, an open-source software environment for statistical computing and graphics. The statistical analysis consisted of three major phases – 1) variable selection, 2) determining correlation and significance between selected variables, and 3) developing a regression model relating traffic crashes to different measures of poverty.

Data aggregated at the Census tract level was chosen as the most appropriate for this analysis due to its large sample size and high spatial resolution. These elements of the data make it less susceptible to the influence of extreme outliers. Census tracts within Manhattan were also removed from the citywide correlation analysis and regression model. Initial exploration of the data revealed that these tracts were exercising a high degree of influence due to their inflated crash densities, a result of lower Manhattan's massive daytime population incrase. Trends consistent in the other four boroughs were often bucked by trends in Manhattan; for this reason, we chose to analyze Manhattan separately as a unique case.

### *Variable Selection*

Variables for this analysis were chosen in consultation with Transportation Alternatives. Because the project was particularly focused on exploring connections between traffic crashes and poverty, standard poverty variables like population density, median income, number of people in poverty, poverty rate, family poverty rate, employment rate, and unemployment rate were selected. Transportation Alternatives also expressed interest in looking at variables related to public housing to assess whether these residents in particular were vulnerable to a higher incidence of traffic crashes. For this reason, the number of public housing developments, number of public housing units, and the percent of the total population living in public housing were added as variables to this analysis. Data for these variables came from the New York City Housing Authority (NYCHA) via the NYC Open Data Portal.

Transportation Alternatives was also interested in looking more closely at transportation statistics to theorize potential correlations between different transportation modes, poverty, and traffic crashes. While transportation data is provided in the ACS it only reflects commuting transportation experiences, thus capturing only a proportion of all transportation events occurring in a place. This data is implicitly biased against poorer communities in which unemployment rates tend to be higher; their transportation experiences are not reflected in this data. Transportation variables included in this analysis were percent of the population commuting by car, percent of the population commuting by public transportation (excludes taxis and other ride-sharing services), percent of the population commuting by bike, percent of the population commuting by walking, the total mileage of bike lanes, and the density of bike lines (mileage of bike lanes per square mile).

Finally, two spatial variables were included in the analysis. The first, mean distance to Central Park,

was used to approximate each feature's distance from the city center. Central Park's centroid was used as the point for the city center. The second, mean distance to bike lane, was used to approximate, on average, how far a resident would have to travel to reach the closest bike lane. Both of these variables were obtained through geospatial analysis using Euclidean Distance rasters and Zonal Statistics operations with the Central Park centroid and the NYC bike lanes shapefile as inputs.

23 variables in total were included in the analysis. These are summarized in Figure 10.

### *Computing and Visualizing Correlation Between Traffic Crashes and Poverty*

An R script was developed to perform the statistical analysis and generate graphics for visualizing correlations between traffic crashes and poverty. The script first subsets the data to remove erroneous results using the dplyr[5] package. The script then uses the scatterplot function from the car[6] package to generate scatterplots relating each surveyed variable (x) to the crash density (y). The scatterplot function also generates box plots along each axis to show the spread of the minimum, maximum, first quartile, third quartile, and median for the plotted variables. These scatterplots were then exported as scalable vector graphics (SVG) files from R and polished up in Adobe illustrator.

To obtain correlation values (r) and significance values (p) for every possible combination of variables in the dataset, covariance and correlation matrices were generated using R's built-in cov and cor functions. A custom function was then defined to flatten the generated values (dependent variable, independent variable, strength of correlation, and significance of correlation) into a table structure that could be exported as a CSV. This provided the strength, sign, and significance of every possible correlation within the dataset. Finally, the corrplot function from the corrplot[7] package was used to generate a graphical display of a correlation matrix.

| Variable | Source |
|---|---|
| Number of Crashes | Transportation Alternatives |
| Total Population | American Community Survey |
| Area of Census Tract | Dervied – Geospatial Analysis |
| Crashes per Capita | Dervied – Statistical Analysis |
| Crashes per Square Mile | Dervied – Statistical Analysis |
| Population Density | Dervied – Statistical Analysis |
| Number of People in Poverty | American Community Survey |
| Poverty Rate (%) | American Community Survey |
| Family Poverty Rate (%) | American Community Survey |
| Median Income ($) | American Community Survey |
| Employment Rate (%) | American Community Survey |
| Unemployment Rate (%) | American Community Survey |
| Number of Public Housing Developments | New York City Housing Authority |
| Number of Public Housing Units | New York City Housing Authority |
| Percent of Population Living in Public Housing | Dervied – Statistical Analysis |
| Total Mileage of Bike Lanes (mi.) | NYC Open Data<br>Dervied – Geospatial Analysis |
| Mileage of Bike Lanes per Square Mile | Dervied – Geospatial Analysis |
| Percent Commute by Car | American Community Survey<br>Derived – Statistical Analysis |
| Percent Commute by Public Transportation | American Community Survey<br>Derived – Statistical Analysis |
| Percent Commute by Bike | American Community Survey<br>Derived – Statistical Analysis |
| Percent Commute by Walking | American Community Survey<br>Derived – Statistical Analysis |
| Mean Distance to Central Park (mi.) | Dervied – Geospatial Analysis |
| Mean Distance to Bike Lane (mi.) | Dervied – Geospatial Analysis |

**Figure 10. Variables selected for analysis, with the source provided.**

### *Developing a Regression Model*

The final phase of the analysis involved developing a regression model to begin to theorize causal relationships between traffic crashes and poverty in New York City. The regression model was also developed in R.

To begin regression analysis, a series of ordinary least squares (OLS) models were tested and diagnosed for potential biases. Diagnostic tests included the variance inflation factor (VIF) to assess variable multicollinearity, a QQ Plot and Shapiro-Wilks test to assess normality, a component plus residuals plot to assess linearity, and a Breush-Pagan test to assess heteroscedasticity. These tests were supplemented by a Global Validation of Linear Assumptions (GVLMA) model to ensure the model passed all assumptions of OLS regression. However, it soon became apparent that no combination of variables satisfied all OLS assumptions. Stepwise regression and Best Subsets regression were also attempted to uncover potential passing models, but neither of these methods yielded suitable results. For this reason, a different family of regression models had to be chosen.

Generalized linear models (GLM) adapt linear regression by allowing for response variables that have a non-normal distribution and relating response variables to linear models via a link function. Finding a link function that mirrors the distribution of your response variable is critical to finding a well-specified GLM. To do this, the R script employed the histDist function found in the GAMLSS[8] package to model the distribution of our response variable (in this case, Number of Crashes) against the typical distributions of four link functions – Normal, Poisson, Zero-Inflated Poisson, and Negative Binomial. The results are visualized in Figure 11.

The latter three models are all appropriate for modeling response variables that are count variables (i.e. a Number of Crashes); however, the Negative Binomial model is particularly well-suited to situations in which overdispersion is present, that is, when the observed variance of the response variable is greater than the theoretical variance. The R script checked for overdispersion in the OLS model by dividing the deviance of the model by the degrees of freedom of the residuals and found significant overdispersion to be present. This set of observations led to the selection of the Negative Binomial model as the basis for the regression model.

The final step in the regression analysis involved iteratively building candidate Negative Binomial models based on the best models uncovered by OLS regression using the glm.nb function in the GAMLSS package. First, all candidate variables were centered and standardized (z-scored); this allows for each variable's regression coefficient to be directly compared with others as a standard deviation from the mean. An offset parameter was also added to the candidate models to control for the propensity of larger tracts to have more crashes based simply on their size. In effect, this normalized the response variable by area. These models were then tested against the null model and compared on the basis of Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) scores. These two metrics balance goodness of fit with number of variables, penalizing models that overfit by adding additional independent predictors. The model with the lowest AIC and BIC was selected and passed through a p Chi-Square test to ensure overall model significance. Diagnositc plots were also generated to assess the fit. The regression model and all accompanying diagnostic plots are included below. Unfortunately, due to the instability of the $R^2$ statistic for Negative Binomial models, no global measure of goodness of fit was generated for this model.

## Sources


1. City of New York. (2014). *Vision Zero Action Plan 2014*. New York, NY.
2. Rudin Center for Transportation Policy and Management, Wagner School of Public Service, New York University. (2012). *The Dynamic Population of Manhattan*. New York, NY: Mitchell L. Moss and Carson Qing.
3. US Department of Commerce, Bureau of the Census. (2010). *CPH-1 Summary of Population and Housing Characteristics*. Washington, D.C.
4. R Core Team (2016). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/.
5. Hadley Wickham and Romain Francois (2016). dplyr: A Grammar of Data Manipulation. R package version 0.5.0. https://CRAN.R-project.org/package=dplyr
6. John Fox and Sanford Weisberg (2011). An {R} Companion to Applied Regression, Second Edition. Thousand Oaks CA: Sage. URL: http://socserv.socsci.mcmaster.ca/jfox/Books/Companion
7. Taiyun Wei and Viliam Simko (2016). corrplot: Visualization of a Correlation Matrix. R package version 0.77. https://CRAN.R-project.org/package=corrplot
8. Rigby R.A. and Stasinopoulos D.M. (2005). Generalized additive models for location, scale and shape,(with discussion), Appl. Statist., 54, part 3, pp 507-554.

Assessing the Fit of Different Generalized Linear Models with the Distribution of New York City Crash Data

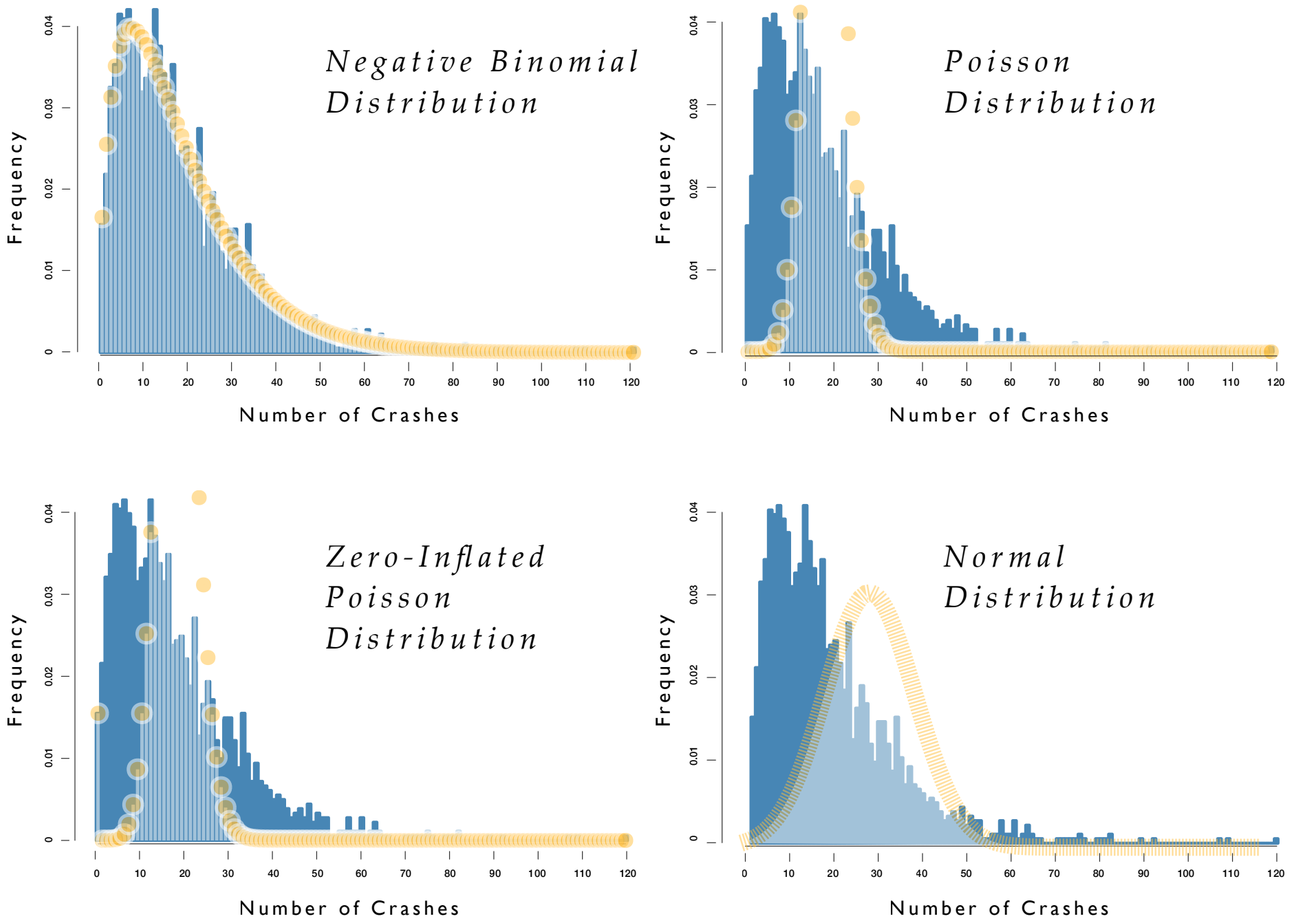



**Figure 11. Assessing the fit of different generalized linear models with the distribution of New York City crash data.** The negative binomial distribution most closely fits the data's distribution.

**Summary of Negative Binomial Regression Analysis for Variables Predicting Number of Traffic Crashes (N = 1878)**

| Variable | B (Coefficient) | SE B (Std. Error Coeff.) | β (Standardized Coefficient) |
|---|---|---|---|
| Median Income ($) | -0.000003996 | 0.000001019 | -0.09731 *** |
| Population Density (People / mi.$^2$) | 0.000002174 | 0.0000007044 | 0.06428 ** |
| Poverty Rate (%) | 0.01605 | 0.001831 | 0.18557 *** |
| Mean Distance to Central Park (mi.) | -0.06738 | 0.005722 | -0.23204 *** |
| Intercept | 3.154 | 0.1035 | 2.68786 *** |
| **AIC** | **13542** | | |
| **BIC** | **13575** | | |

* $p < 0.05$ ** $p < 0.01$ *** $p < 0.001$

**Figure 11. Resuts of the regression analysis.** Regression coefficients show the increase in **log(Number of Crashes)** for a one unit increase in the variable. For example, a 1% increase in the poverty rate results in a 0.016 increase in **log(Number of Crashes)**. Beta coefficients represent units of one standard deviation and are directly comparable. For example, Mean Distance to Central Park has the strongest pull on crash density because the absolute value of its beta coefficient is the largest.

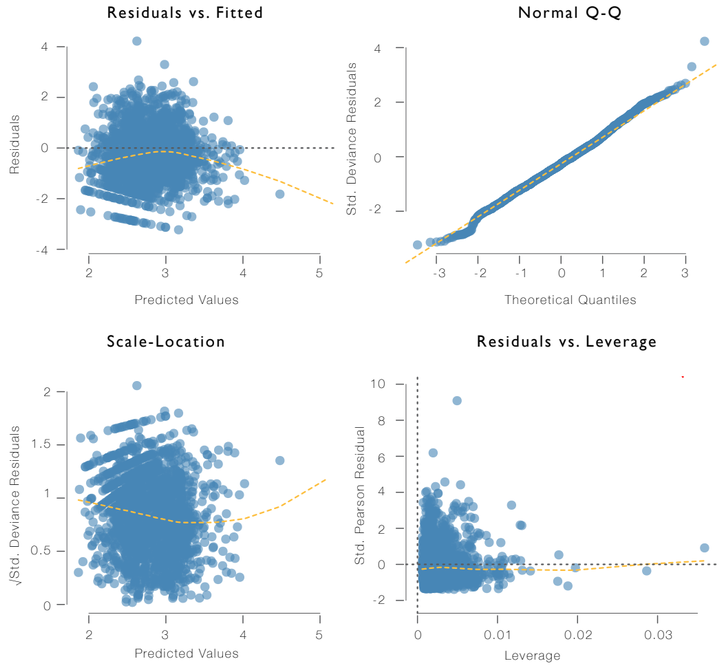



**Figure 12. Diagnostic plots of the regression analysis.** Diagnostic plots can help to assess how well a regression model fits the data. Residuals vs. Fitted produces a horizontal line in highly accurate models, which suggests that the model over and underpredicts equally; ours shows some skew at higher predicted values, suggesting it does not perform as well at higher predicted values. The Normal Q-Q plot shows that the negative binomial transform helps us meet the criteria of linearity for generalized linear regression analysis.