University of Pittsburgh
Department of Economics
April 25, 2025

MQE Capstone for the Pittsburgh Parks Conservancy:

Measuring and Mapping Community Investment Need

A guide to assist the Pittsburgh Parks Conservancy in their effort to prioritize public investment needs on a park-by-park basis using analyses and visualizations of data characterizing the demographics, environment, public health, crime, and other attributes of Pittsburgh parks and their surrounding walkshed areas.

Authors

Lap Pham

Patrick Swain

José de los Ríos

Maxwell Snodgrass

Acknowledgements

Throughout this project, Professor Randall Walsh was an invaluable resource as an advisor, from broad discussions on the scale of the analysis to guidance on complex technical processes to ensure the robustness of our results. Notably, he recommended the technique of areal interpolation for attaching third-party data to walksheds, which became the cornerstone of our geospatial data analysis. Additionally, Professor David Huffman provided counsel in periodic check-ins during the process of creating the project.

We would also like to acknowledge the 2019 report performed by Interface Studio LLC for the Pittsburgh Parks Conservancy, from which we derived spatial data including shapefiles of parks and walksheds that became an integral part of our analyses and improved the efficiency of our workflow. Additionally, 2022 data on air quality across North America by the Atmospheric Composition Analysis Group at Washington University was a key part of our environmental analysis.

Finally, our primary touchpoint within the Pittsburgh Parks Conservancy was Ross Chapman, the Conservancy's Chief of Operations. His breadth of knowledge on Pittsburgh's public spaces from his experience at the Conservancy and as director of the City's Department of Parks and Recreation lent ample expertise to defining the scope of our project and its underlying motivations.

Outline

I. Introduction

As candidates for a Master of Science in Quantitative Economics at the University of Pittsburgh, we have developed an analytical acumen that enables us to harness data to bolster public utility. For our capstone project, we dedicated the spring semester to analyzing data for the Pittsburgh Parks Conservancy and uncovering patterns and disparities affecting Pittsburgh's public parks and their surrounding communities, or "walksheds." After rigorous quantitative analyses of statistics describing relevant factors including demographic data, environmental metrics, public health outcomes, and crime rates, we produced this document and the accompanying datasets and visualizations as a guide for the Conservancy to quantify dimensions of community need on a park-by-park basis. This project is a resource to enable the Conservancy to make data-driven decisions in its efforts to allocate park budgets, secure public and private investment, and engage stakeholders across the communities of its footprint effectively and equitably.

We analyzed data from an array of public sector sources from federal bodies like the U.S. Census Bureau, the Center for Disease Control and Prevention (CDC), the U.S. Forest Service, and the Environmental Protection Agency (EPA) to local entities like the Pittsburgh Water and Sewer Authority (PWSA) and the Pittsburgh Police Department (PPD). We also cross-referenced private data from the Pew Research Center and Washington University in St. Louis to corroborate our findings. We targeted variables that offered descriptions of the population and environment of the walksheds, such that we could concisely characterize the demographics, environment, health, and crime of a given walkshed. After downloading and cleaning data into concise datasets preserving selected relevant variables, we used the technique of areal interpolation to attach data from geographic identifiers like Census tracts to the walksheds. After attachment and additional calculations scaling and standardizing the data, we produced datasets that displayed selected variables as raw counts, percentages, and z-scores.

We employed a technical toolkit that included R packages sf, ggplot2, and tmap, geographic information systems software QGIS, and Python packages seaborn and matplotlib to analyze and visualize our cleaned and attached datasets. We developed indexes to model relative racial diversity of walksheds, analyzed the correlations between selected relevant variables, segmented parks based on relationships between tree canopy, acreage, sports facilities, and other factors, and broke down violent and nonviolent crime in walksheds and parks over time.

II. Data and Methodology

A "walkshed" refers to the area where a typical resident lives within a five-minute walk of a given park. We let the walksheds represent the surrounding community of a park – and hence, we sought to synthesize and transform data to analytically describe those communities. With in-depth profiles of the parks and their respective walksheds, detailing their demographic characteristics, environmental factors, and other descriptive statistics, we could robustly answer comparative questions about patterns and disparities in the urban communities where the Pittsburgh Parks Conservancy operates.

We selected an array of metrics that offered broad characterizations of the parks and walksheds, falling into four main categories: socioeconomic demographics, ecology and pollution, public health, and crime. We drew from a variety of data sources to obtain raw datasets to begin our analysis – most of them were public, including data exported from the U.S. Census Bureau, the Pittsburgh Water and Sewer Authority, the U.S. Forest Service, and reports from other public sector data providers. The diversity of our sources made data cleaning a lengthy stage of the process, with significant time and energy dedicated towards selecting relevant columns, standardizing numerical identifiers, and merging smaller tables into large datasets for wider analysis of the trends and relationships between selected variables.

The shapefiles of the walksheds, giving us precise geospatial data on the size and shape of walksheds, came from a 2019 report conducted for the Conservancy by Interface Studio LLC. Using QGIS, we separated parks from their walksheds. To attach data collected at the Census tract level and other geographic identifiers to the walksheds, we employed a technique called areal interpolation using R's spatial features package. This process uses geometric information of two intersecting polygons and estimates the data within the intersection – effectively, it allows us to predict the characteristics of a walkshed based on the areas like Census tracts, whose characteristics we know, with which it intersects.

After interpolation, we summarized all the intersections for a given walkshed to create datasets with a single row describing each park's walkshed. For variables measured in counts, we transformed them into proportions using total walkshed population and scaled them with z-scores. This is a measurement that describes how many standard deviations a datapoint lies away from the average, giving us a way to analyze walksheds relative to each other.

II. Data and Methodology → A. Census

We sourced demographic data from the 2020 decennial Census and the 2023 American Community Survey. Our analysis focused on four main factors: race and ethnicity, age, poverty, and vacancy. After downloading four large .csv files from the U.S. Census Bureau specified to Census tracts in Allegheny County, PA, we cleaned the datasets in R to eliminate extraneous columns, standardize numerical tract identifiers, and combine them into a single file grouped by tract.

The raw data represented counts of column variables per tract – for instance, we could see the number of residents between 20 and 24 years of age, were below the poverty line, or were Hispanic/Latino in a given Census tract. We performed aerial interpolation with the spatial features package in R using a shapefile of parks and their walksheds. This produced a dataset that contained the interpolated counts of each variable in the polygons representing the intersection of each Census tract and walkshed. Summarizing by walkshed and weighting by population, we created a dataset that showed the counts of each variable for selected relevant variables in each walkshed.

We used midpoints to transform the many age columns into a single median age column and converted counts into percentages by dividing each count by the total population of the walksheds. Using those new variables, we standardized the values using z-scores to create an index of each variable and assigned each z-score a rank. Hence, we produced datasets of raw counts, percentages, scaled index scores, and ranks of each park walkshed for the characteristics of total population, Hispanic/Latino, White, Black, Asian, occupied parcels, vacant parcels, residents below the poverty line, and age variables, among several others.

Using R packages including ggplot2, we visualized the indexed data using bar graphs (exported as .png files), choropleth maps (exported as interactive web-based .html files), and other methods for selected variables to compare walksheds to each other through rankings and illuminate geospatial patterns that reveal socioeconomic disparities and trends across Pittsburgh neighborhoods.

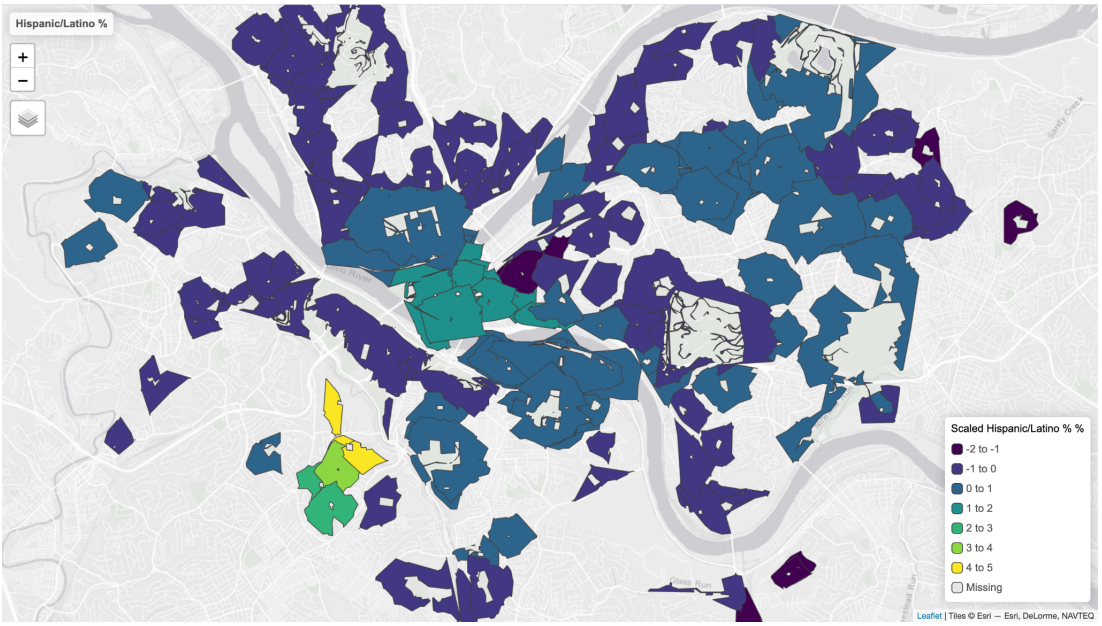
II. Data and Methodology → A. Census → 1. Race

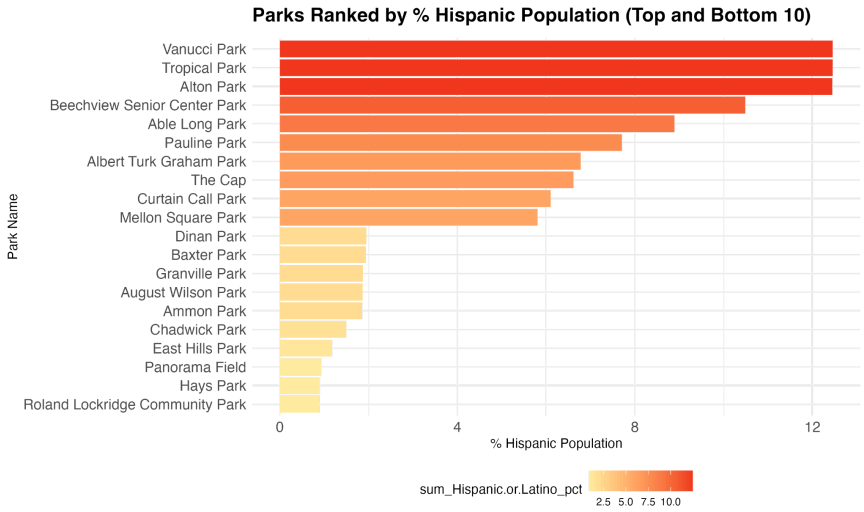
We broke down race and ethnicity by four main categories – Hispanic/Latino, Black, White, and Asian. The raw datasets included counts for other races including Native Hawaiian/Pacific Islander and American Indian, but in almost all cases, the counts were too low for any meaningful analysis or visualization. We also preserved a column that described the count of residents who identified with two or more races. We visualized those four racial categories using interactive geospatial .html maps (created with the R package tmap) and bar graphs showing the walksheds with the top and bottom 10 ranked percentages (created with the R package ggplot2).

Hispanic/Latino Population

The first ethnicity we analyzed was Hispanic/Latino. This variable was unique in that the Census recognizes Hispanic/Latino not as a racial category, but as an ethnic identifier, such that anyone of any race can identify as Hispanic/Latino without necessarily falling into the "two or more races" column. A map of the Hispanic/Latino index on walksheds in Pittsburgh shows that there are a relatively low count of Hispanic/Latino residents across the city apart from a major cluster in Beechview and a smaller concentration in Downtown and the Bluff. Likewise, the walksheds with the highest proportion of Hispanic/Latino residents were Vanucci, Tropical, and Alton Parks – all within a few blocks in Beechview. The lowest proportion appeared in areas with especially high concentrations of Black residents such as the Hill District and Homewood.

Percent Hispanic/Latino population by walkshed

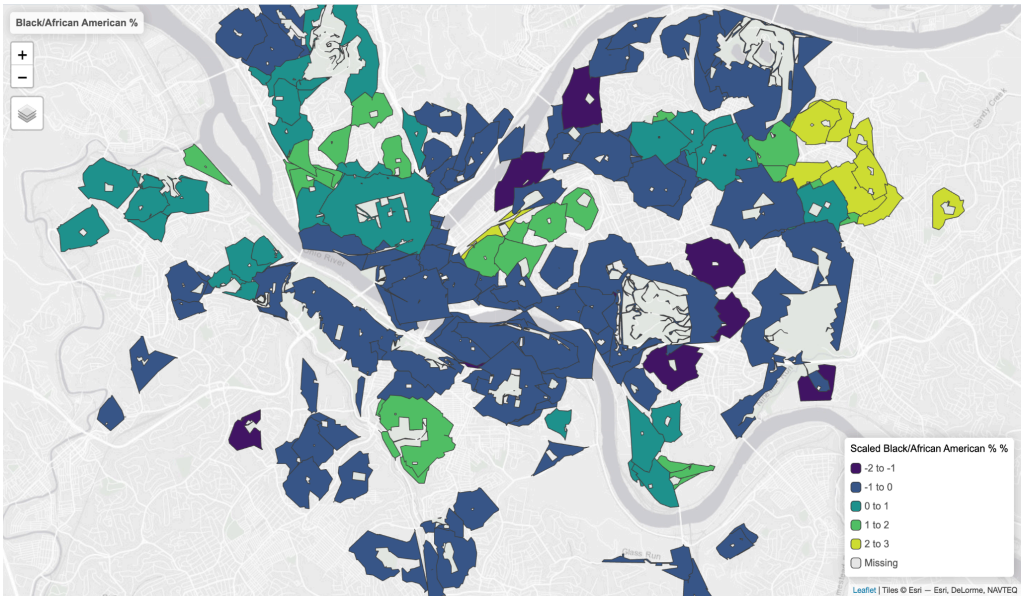


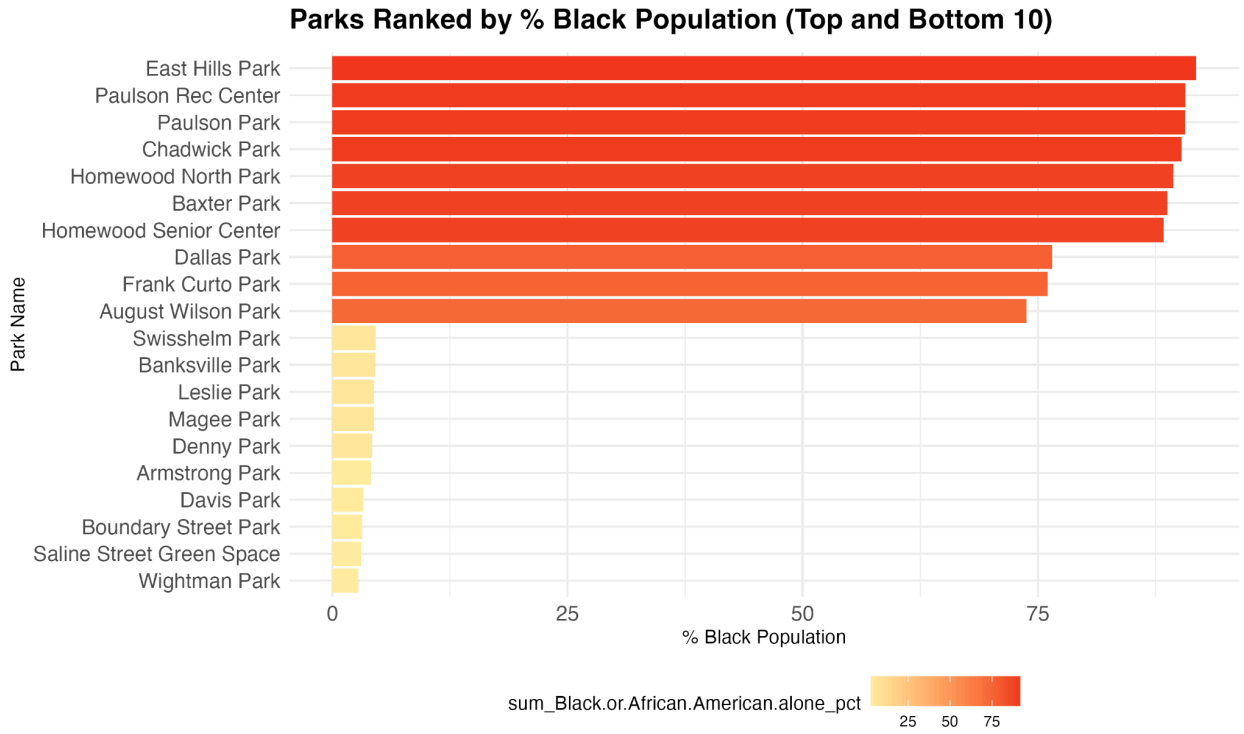


Black/African American Population

Similarly, we found clusters of Black residents in a few key park walksheds concentrated in a handful of Pittsburgh neighborhoods – namely Homewood, the Hill District, the North Side, and Knoxville. Walksheds with lower proportions appeared in largely White neighborhoods in Lawrenceville, Greenfield, the Strip District, and Swisshelm Park and the largely White and Asian neighborhood of Squirrel Hill. The walksheds whose Black population proportions were the highest were overwhelmingly in northeast Pittsburgh, with the top eight walksheds in the predominantly Black neighborhood of Homewood, followed by several parks in historically Black communities of the Hill District.

Percent Black population by walkshed

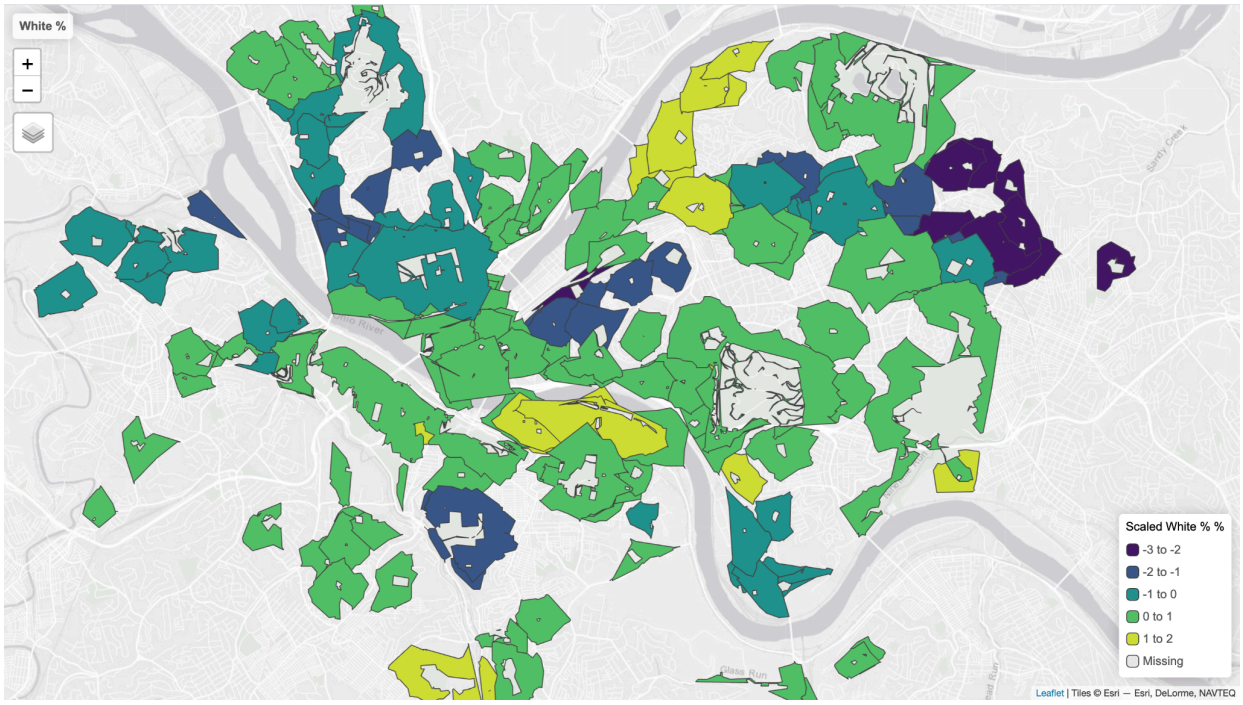


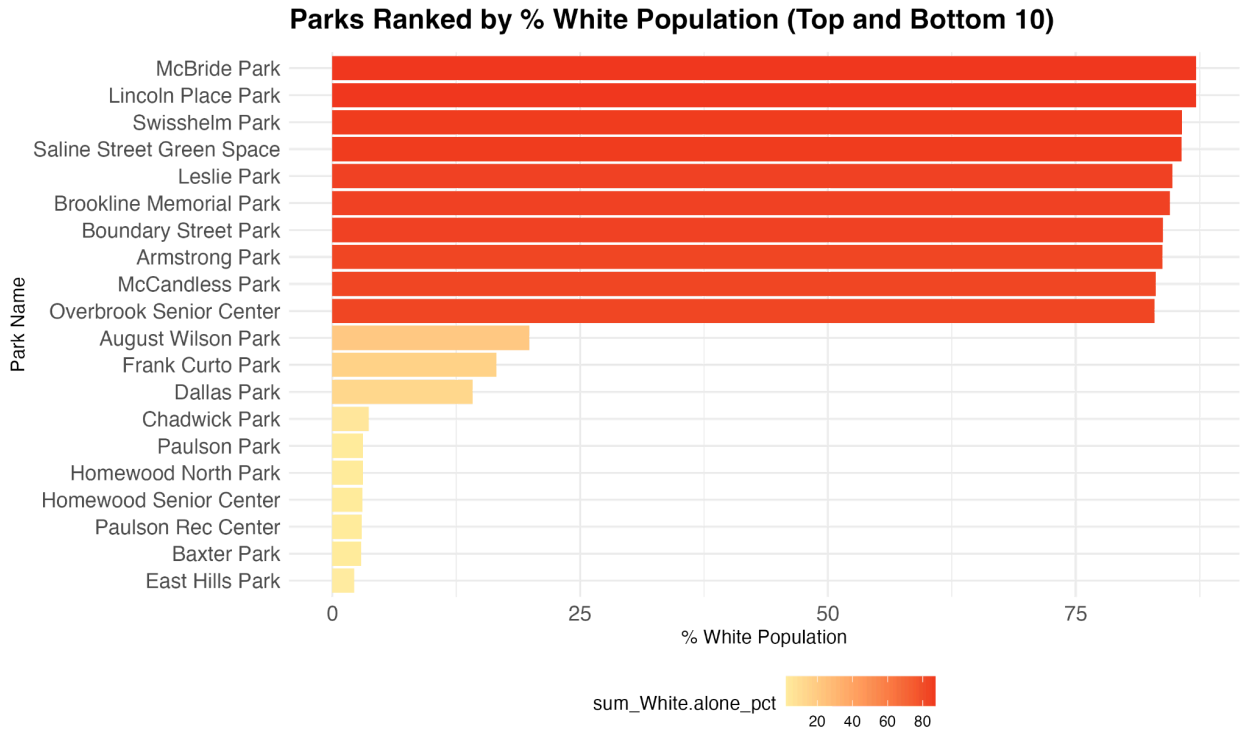


White Population

White residents, the largest ethnic group in Pittsburgh, are distinctly common among the four racial categories we visualized – unlike the three racial minority groups in our dataset, White residents had high proportions in most walksheds apart from a few key neighborhoods. Visually, there seems to be a strong negative association between walksheds with high White and Black population proportions – the only park walksheds with very low White proportions appeared in Homewood, Lincoln-Lemington-Belmar, Knoxville, the Hill District, Garfield, and parts of the North Side. As we found through visualizations of the Black population, most of those areas fall into walksheds with the largest concentrations of Black residents, and the walksheds with the lowest White population (overwhelmingly in northeast Pittsburgh) are nearly identical to the top walksheds for the Black population. The opposite is not as stark – many of the walksheds with the highest White population proportions are lower on the list of Black population, but several notable walksheds in the latter also have large concentrations of Asian residents.

Percent White population by walkshed

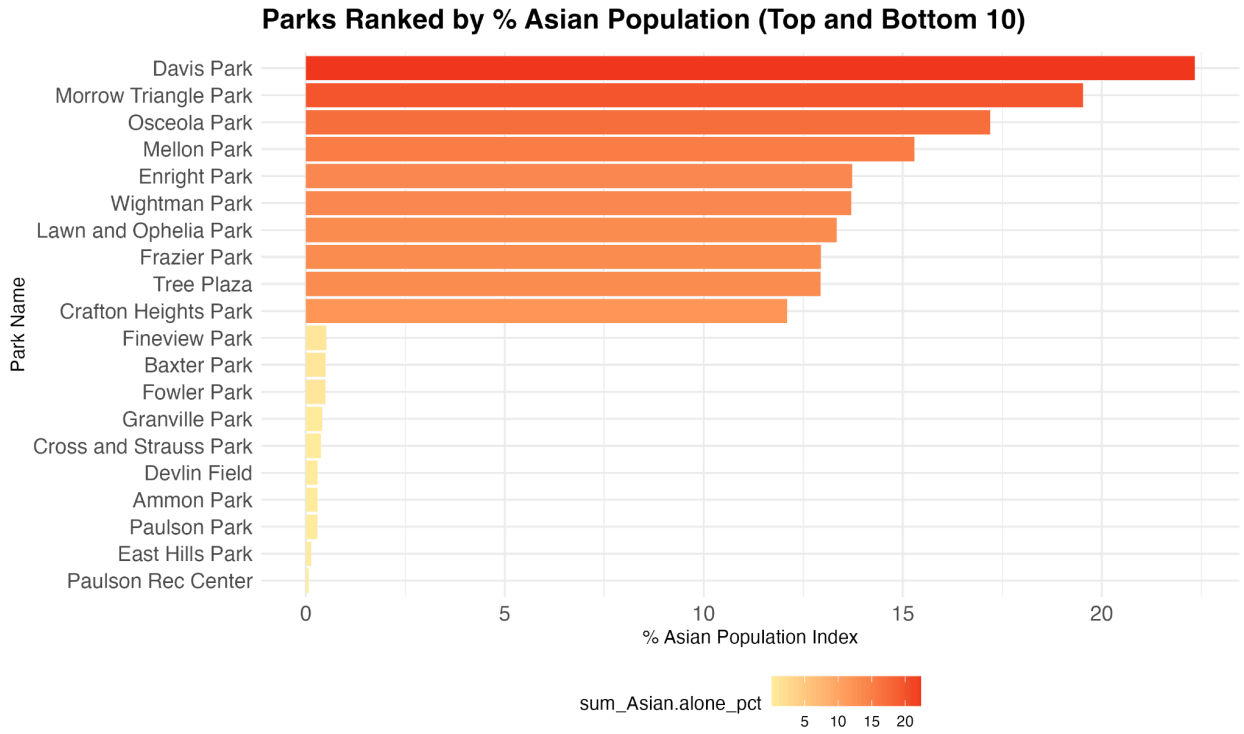




Asian Population

The Census racial category “Asian” is very broad, encompassing residents from the Indian subcontinent to east Asia and Oceania. We found very low proportions of Asian residents in almost all neighborhoods of Pittsburgh apart from several clusters across the East End and smaller concentrations in the West End and Oakland. The highest proportions by far appeared in the corridor from East Liberty to Squirrel Hill, namely the walkshed of Davis Park in Squirrel Hill South. Walksheds in Bloomfield, Shadyside, Point Breeze, and Squirrel Hill North including Morrow Triangle, Osceola, Mellon, Enright, and Wightman also had higher proportions. Several Oakland walksheds including Lawn and Ophelia Park and Frazier Park also appeared in the top 10 Asian population walksheds. The walksheds with the three lowest Asian populations mirror the top three walksheds for the Black population.

Percent Asian population by walkshed

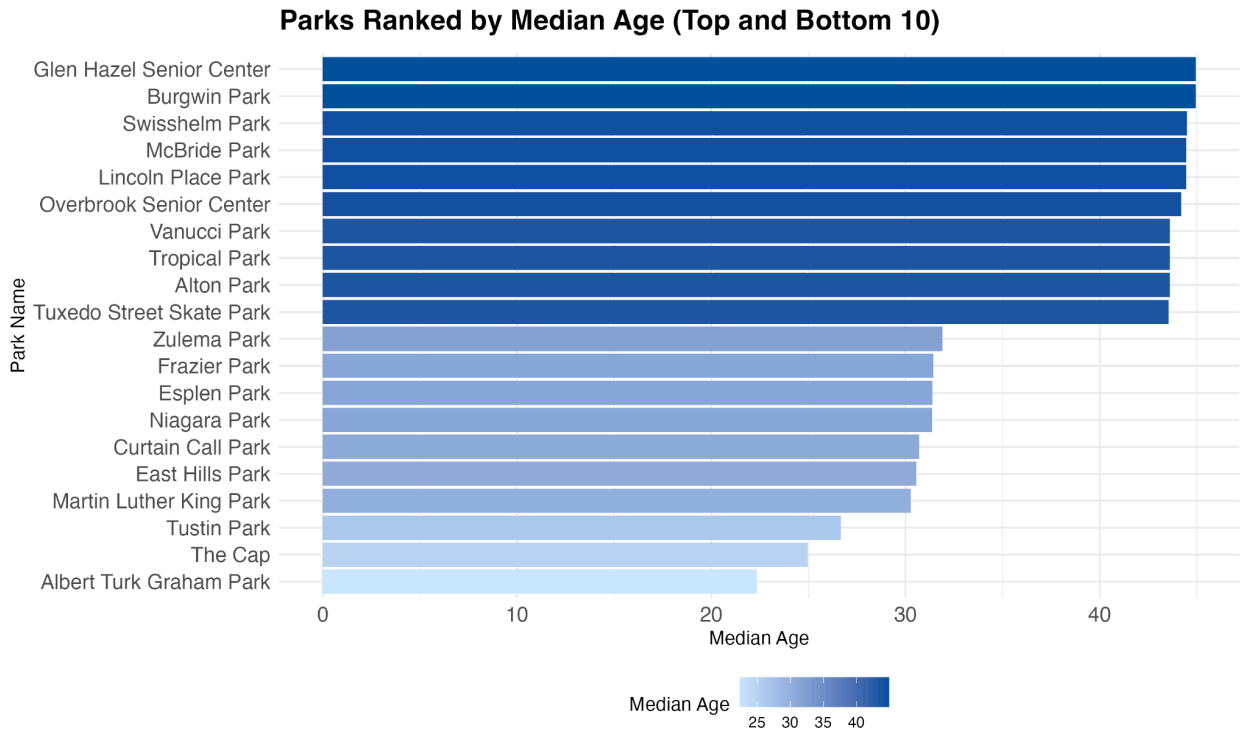


The geospatial distributions of racial categories across park walksheds reveal significant socioeconomic patterns and disparities. For instance, the strong negative association between more White and Black walksheds reflects the legacy of past and present de facto racial and economic segregation in Pittsburgh, a phenomena common in many American urban centers. Black, Hispanic/Latino, and Asian residents are overwhelmingly clustered in a few walksheds and neighborhoods. With surface-level analysis, park walksheds do not appear to be very racially diverse – there were no parks whose index scores appeared in the top half of all four racial population indexes. In the Analysis section, we used more robust techniques to analyze and visualize the racial diversity of the walksheds and the relationships between racial categories and other demographic data.

II. Data and Methodology → A. Census → 2. Age

After summarizing age columns into a single median age metric, we could characterize the middle value of ages in each walkshed, and thus, the age of a typical resident. This ranged from around 45 for the Glen Hazel Senior Center and Burgwin Park's walksheds in Hazelwood to about 22 for Albert Turk Graham Park in Crawford-Roberts at the junction of the Hill District and the Bluff.

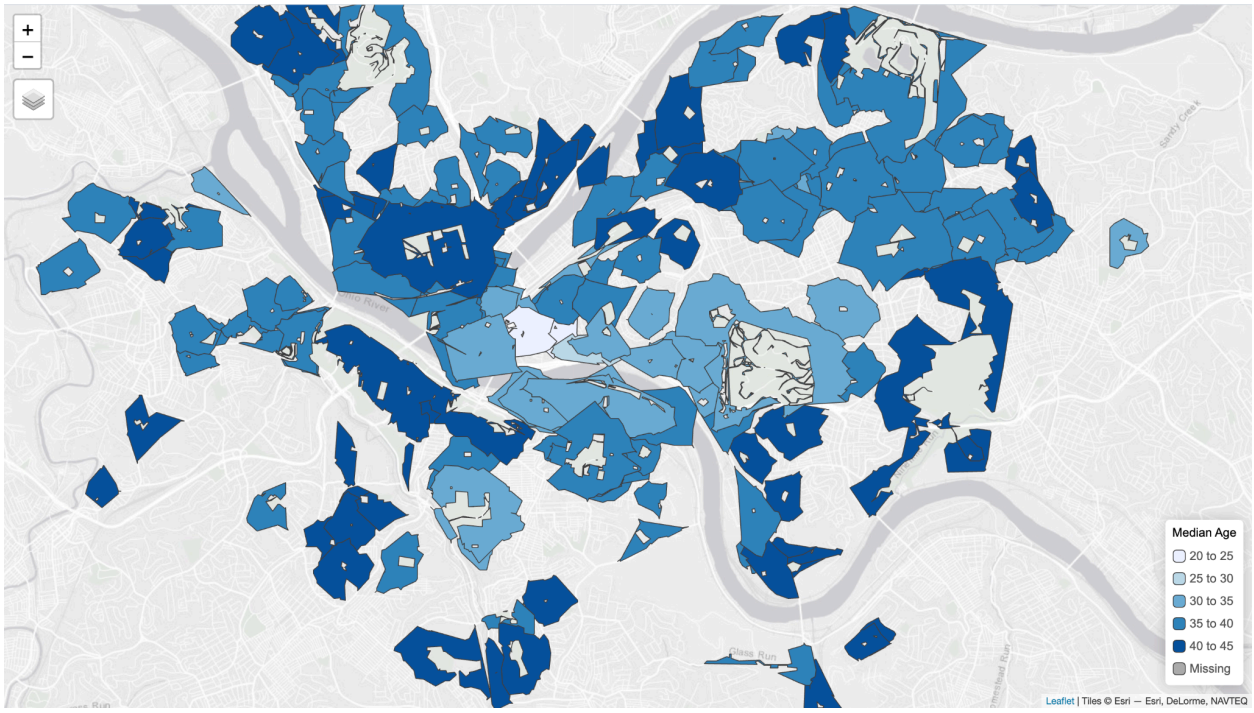
The locations of universities are likely a significant factor in this spatial distribution – many of the youngest walksheds are located in Oakland, Squirrel Hill, the South Side Flats, and Downtown around the "studentified" neighborhoods near the University of Pittsburgh, Duquesne University, Carnegie Mellon University, and others. The four youngest walksheds – Albert Turk Graham, the Cap, Tustin, and Martin Luther King – are all within half a mile of Duquesne's campus. Zulema, Frazier, and Niagara Parks in Oakland near Pitt also appeared in the top 10 youngest walksheds.



On the other end of the median age range, two of the oldest walksheds belonged to senior centers – Glen Hazel and Overbrook. There were a few distinct clusters of older residents. The walksheds surrounding Frick Park – most notably Swisshelm Park – Hazelwood and Glen Hazel, several areas of the North Side, Lawrenceville, Mount Washington, and the South Hills all had median ages above 40.

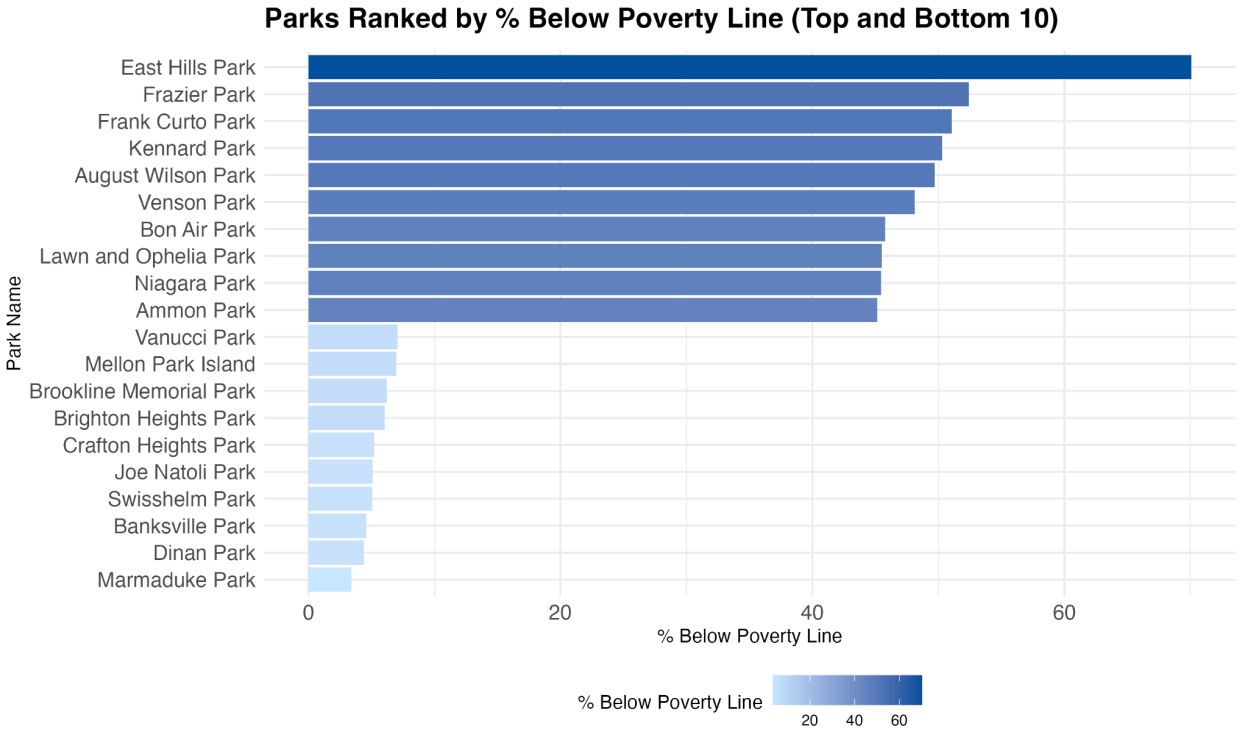
Interestingly, the three significantly Hispanic/Latino walksheds we identified through our racial data visualization – Vanucci, Tropical, and Alton in Beechview – not only appear in the top 10 oldest walksheds, but appear in the same order as their Hispanic/Latino population ranks. Since the Beechview walksheds have such high Hispanic/Latino proportions compared to the next most Hispanic/Latino walkshed, we should not treat them as representative of the greater relationship between age and Hispanic/Latino population, but the pattern reveals intriguing characteristics of the Beechview community nonetheless.

Median age by walkshed



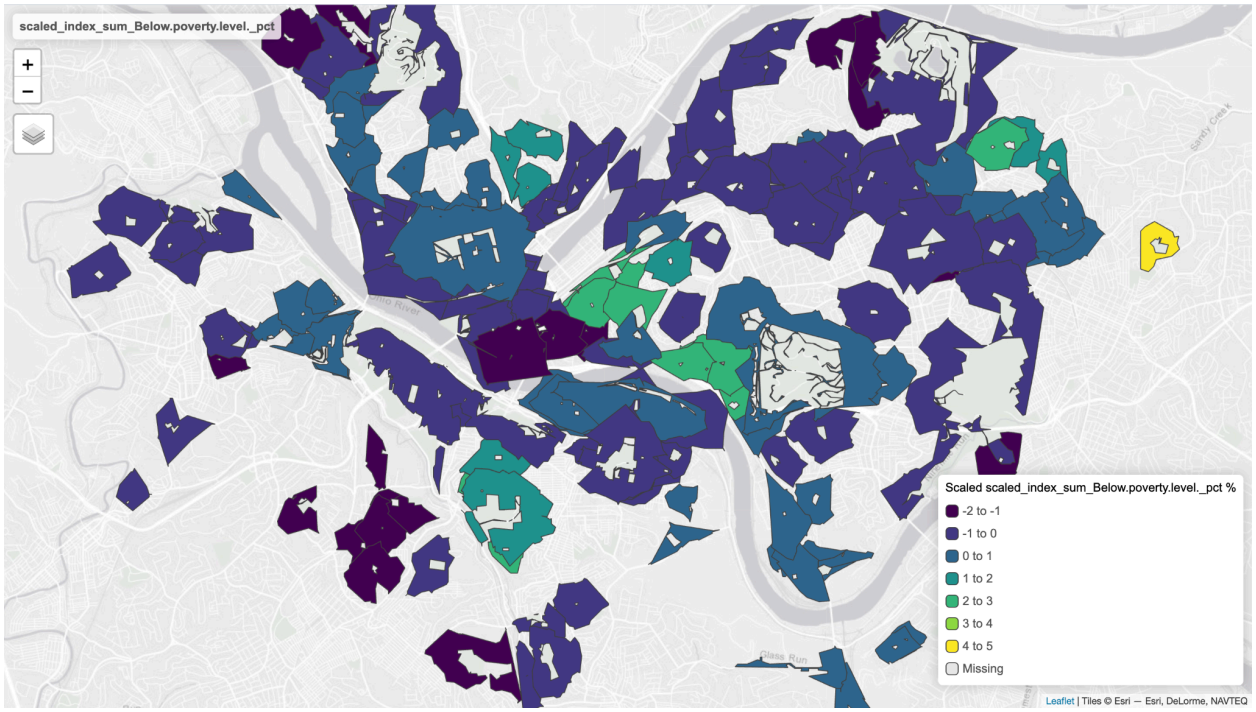
II. Data and Methodology → A. Census → 3. Poverty

The raw Census datasets contained a multitude of columns related to income and poverty, but the most concise was residents below the poverty line. In the cleaned dataset, we preserved counts of residents for whom the poverty status was known and the count of residents below the poverty line. The Census Bureau does not have a single threshold for poverty, rather a series of thresholds for each individual based on age, size of household unit, and number of children. For instance, in the 2020 Census, the poverty line for one person under 65 years with no children was $13,465. Thus, the count represented the number of people in each tract who had been individually determined to experience poverty via the respective threshold according to their demographic characteristics.



Similar to the race indexes, we summarized the interpolated poverty counts by park, divided them by walkshed population, standardized them into an index with z-scores, and ranked the z-scores. We produced a single column that contained scaled values describing the relative poverty of each walkshed.

Percent below poverty level by walkshed



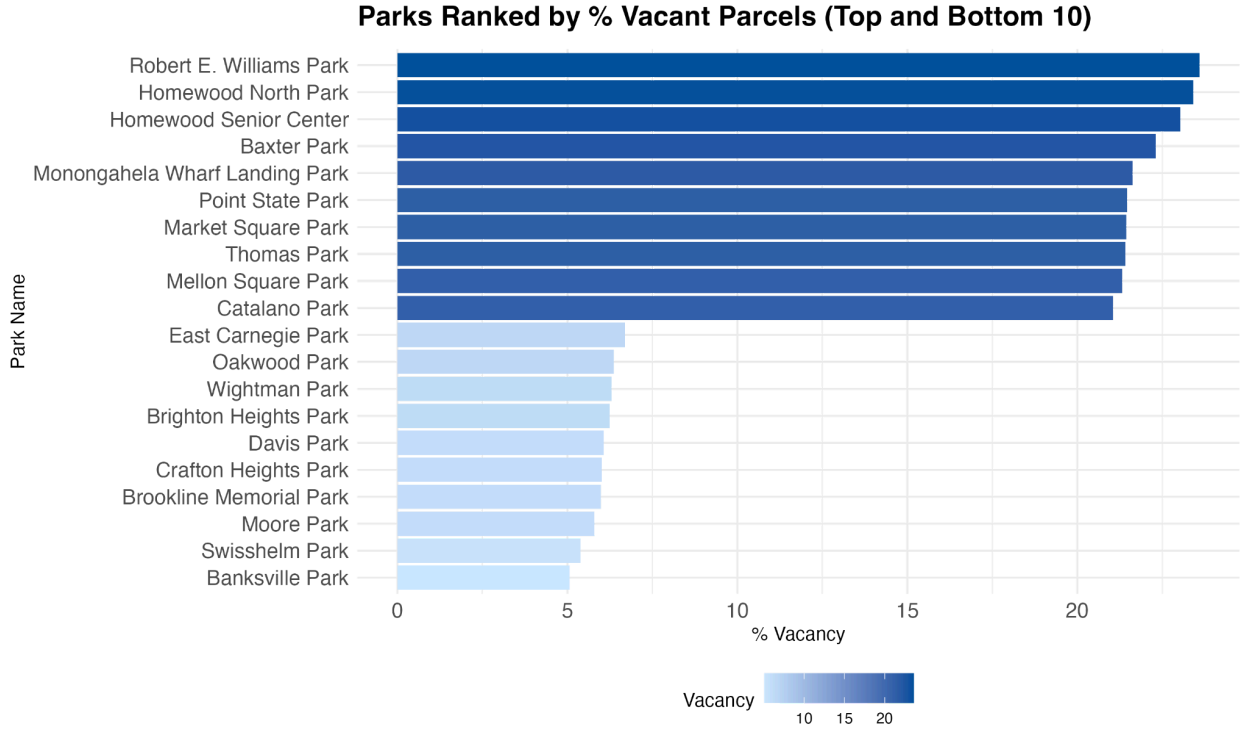
The walkshed that stands out the most from this ranking is East Hills Park – it is extremely poor relative to the rest of the walksheds, with a percent below the poverty line far above the second poorest walkshed of Frazier Park. The interpolated counts suggested out of the roughly 646 residents of the East Hills Park walkshed, an estimated 70% lived below the poverty line. No other walkshed comes close to that proportion, but poverty tends to be concentrated in walksheds in the Hill District, Oakland, Knoxville, and Homewood.

The cluster of residents living below the poverty line in Oakland in the walksheds of Niagara, Lawn and Ophelia, and Frazier Parks could be less indicative of socioeconomic disparities – as could be the case in Homewood and other economically depressed communities – but the large student population in the University of Pittsburgh's footprint, many of whom have little to no annual income as full-time university students.

The smallest proportions of residents below the poverty line appeared in the walkshed of Marmaduke Park, a playground in the affluent Brighton Heights neighborhood. Downtown, Stanton Heights, Swisshelm Park, and a few pockets of the South Hills made up the rest of the bottom 10 walksheds in poverty.

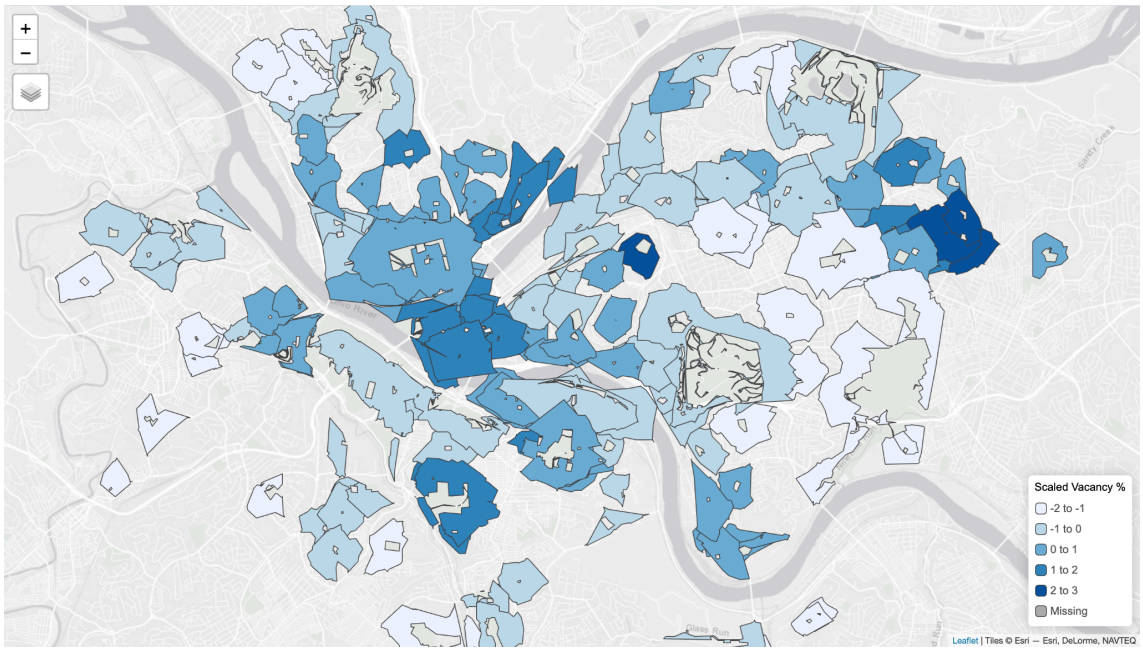
II. Data and Methodology → A. Census → 4. Vacancy

After cleaning the Census data, we preserved three columns: total parcels, count of occupied parcels, and count of vacant parcels. These described the number of property parcels that were legally considered vacant – i.e. at the time of the Census record, there was nobody permanently living in the parcel or it was occupied entirely by individuals who maintained full-time residences elsewhere. After summarizing across walksheds, dividing the count of vacant parcels by total parcels to obtain a proportion, standardizing via z-score, and ranking the walksheds, we had a column describing parks by scaled vacancy.



The walkshed with the highest rate of vacancy was Robert E. Williams Park in the Upper Hill District, while the following three were in Homewood and the remainder of the top 10, apart from Catalano Parklet in Troy Hill, were in Downtown Pittsburgh. The Hill District and Homewood have historically experienced economic disinvestment, leading to deteriorating residential infrastructure that renders many property parcels uninhabitable. The culprit of Downtown's high vacancy rates is likely the high density of commercially-zoned parcels, hotels, and office buildings with no legal residents.

Vacancy rate by walkshed



Vacancy rates were lowest in Banksville Park's walkshed in the South Hills, at just 5% unoccupied parcels. Many of the walksheds with the least vacancy are clustered in the corridor from Friendship to Swisshelm Park, encompassing areas of Shadyside, Point Breeze, Squirrel Hill, and assorted walksheds surrounding Frick Park. Another noticeable hotspot of low vacancy rates is the heavily residential South Hills, home to Banksville Park, as well as several other walksheds from Crafton Heights to Brookline Memorial Parks. As we will see in the Analysis section, there seems to be some correlation between walksheds' proportion of residents below the poverty line and unoccupied parcels – the high-vacancy walksheds in the Hill District and Homewood are also among the poorest on average – but that association is not statistically strong due to the high vacancy and low poverty of areas like Downtown Pittsburgh and Troy Hill.

After cleaning, indexing, and visualizing Census data related to race, age, poverty, and vacancy, we considered other factors to statistically describe the parks and their walksheds – namely the environment, public health, and crime – and further analyze these metrics and the relationships between them.

II. Data and Methodology → B. Environment

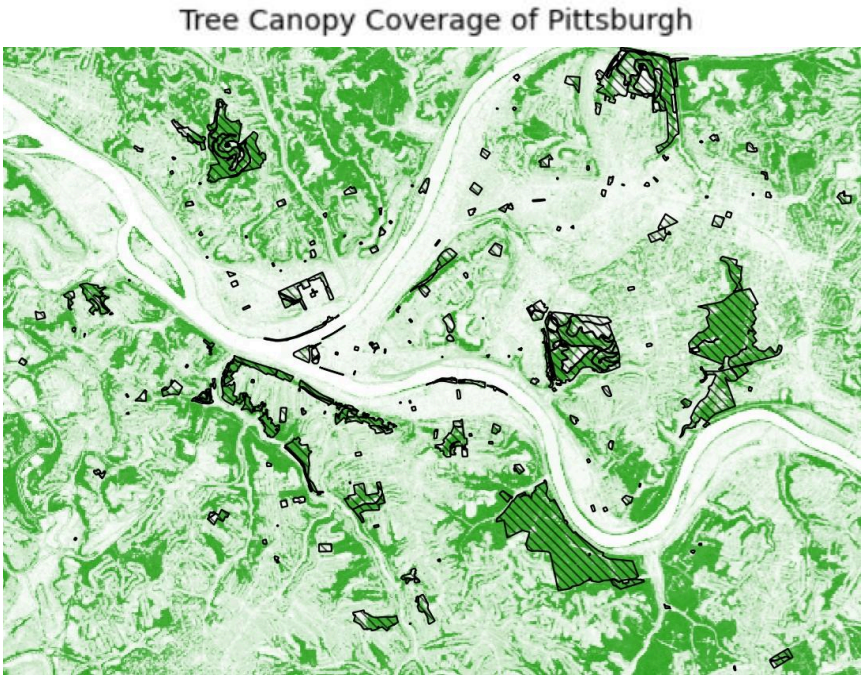
For our analysis of the parks' environmental data, we used three variables: tree canopy coverage, pollution, and sewershed priority. The tree canopy coverage data was sourced from the U.S. Department of Agriculture Forest Service's 2021 Conterminous United States (CONUS) Tree Canopy Cover (TCC) dataset. This report uses satellite imagery to calculate the tree canopy coverage percentage of 30 × 30 meter areas across the conterminous United States. The pollution data was sourced from a study of annual North American pm2.5 concentrations completed by Washington University in St. Louis. This particular dataset measures the pm2.5 concentrations in $\mu g/m^3$ within $0.01^\circ \times 0.01^\circ$ areas in 2022. The sewershed priority data was sourced from the 2024 updated dataset describing combined sewershed priority ratings completed by the Pittsburgh Water and Sewer Authority (PWSA). Priority of a sewershed is measured based on the need for intervention in mitigating overflowing of sewers and storm drains. Specific tracts of land within Pittsburgh's city limits were given one of three ratings by the PWSA: high priority, secondary priority, and low priority.

Each dataset was attached to individual parks and their walksheds through the use of geospatial analysis software QGIS. We loaded the tree canopy and pollution data as raster files, and used a technique called zonal averaging to calculate the average tree canopy coverage and pollution levels for each park and walkshed. We loaded the sewershed priority rating data as a vector layer, and we subsequently used aerial interpolation to calculate an aggregated score of priority. Two additional variables were calculated for sewershed priority measuring the percentage of each park contained within a secondary priority and high priority sewershed boundary.

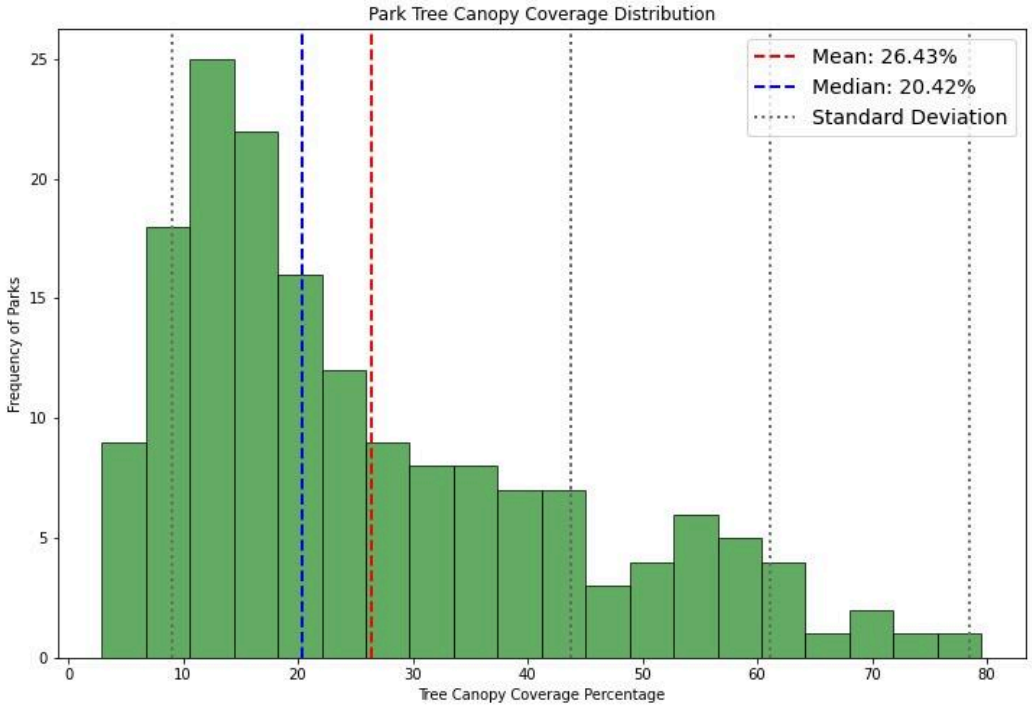
Using Python packages including seaborn and matplotlib, we visualized the environmental data using bar graphs, scatterplots, and other methods to highlight specific parks with a need for environmental intervention.

II. Data and Methodology → B. Environment → 1. Tree Canopy

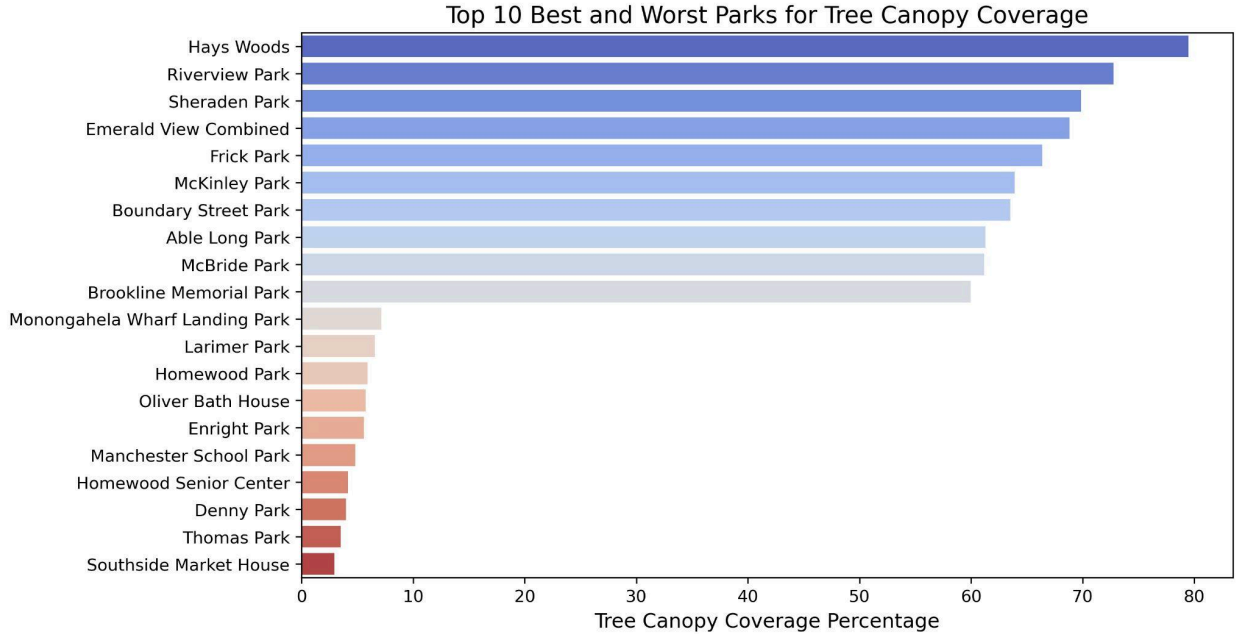
We used the technique of zonal averaging to calculate the average percentage of tree canopy coverage within each park. This technique involves averaging all of the tree canopy coverage percentages (each being calculated in 30 × 30 meter areas) contained within each park and walkshed boundary. The variables of "Tree_Canopy_Park" and "Tree_Canopy_Walkshed" represent average percentages and serve as an accurate indicator for tree canopy distribution among the entire park or walkshed.



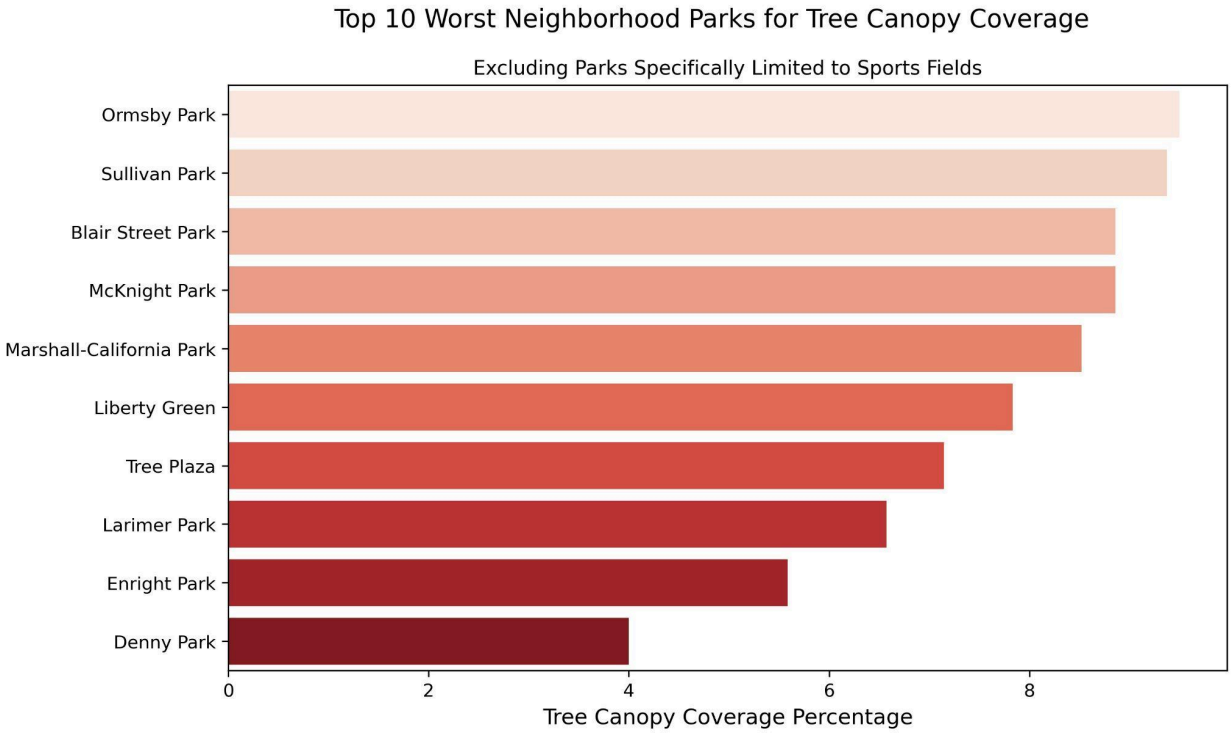
This graphic shows each Pittsburgh park layered on top of a raster layer sourced from the U.S. Department of Agriculture Forest Service's calculation of tree canopy coverage percentages in the conterminous United States. Areas with darker green colors represent a higher percentage of tree canopy coverage. Through calculating the zonal tree canopy coverage averages of each park, we were able to compute the mean and median average for all parks in Pittsburgh and rank each park from most to least tree canopy coverage. The mean and median tree canopy coverage percentage across all parks within Pittsburgh is 26.43% and 20.42%, respectively. This indicates a right skew in the distribution across all parks shown in the histogram below.



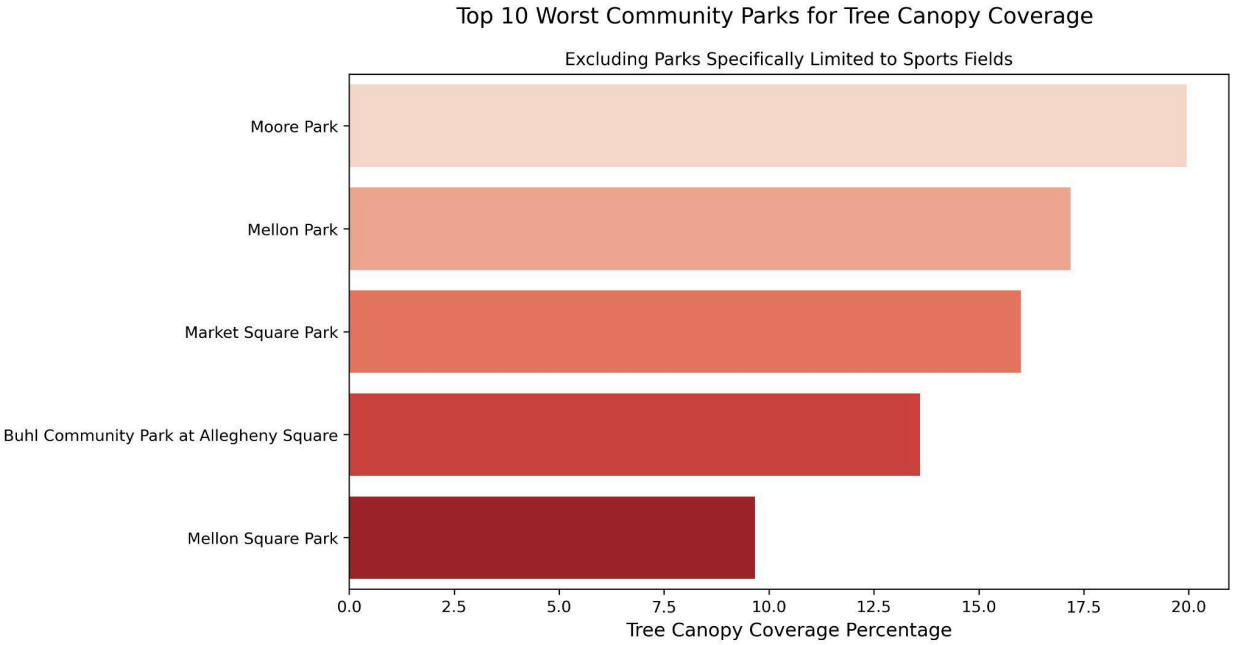
Parks in Pittsburgh significantly vary in their tree canopy coverage with a standard deviation of 17.37% and a range of 76.57%. There is a high concentration of parks below the mean, especially those with around 10%-20% coverage, and a less concentrated and more dispersed collection of parks that exceed the mean. Below is a bar graph that showcases the top 10 best and worst parks for tree canopy coverage.



On the high side of the extreme is Hays Woods, Riverview Park, and Sheraden Park. Parks on the extreme low end include Southside Market House, Thomas Park, and Denny Park. It is important to analyze the underlying reasons of why these parks are on the extreme ends of the dataset. For example, Southside Market House is a historic building listed as a park and should have an expected tree canopy coverage of nearly zero. Additionally, parks that are specifically limited to sports fields, like Manchester School Park and Homewood Park, are expected to have low tree canopy. On the other side of the spectrum, Hays Woods is a largely undeveloped tract of land that lies within city boundaries and it only became a park in 2023, and is therefore expected to have high tree canopy coverage. Three neighborhood parks in the bottom 10 that are not buildings and are not specifically limited to sports fields are Denny Park, Enright Park, and Larimer Park. Below is a bar graph representing the bottom 10 parks for tree canopy coverage in parks categorized as neighborhood parks. Neighborhood parks tend to have less tree canopy coverage on average than community parks. All parks specifically limited to sports fields were manually excluded from this bar graph. This includes parks where the entirety of its area comprises a sports field and parking lot. No intervention on tree canopy would be viable in these parks which include Manchester Field and Homewood Park.



While community parks have a higher tree canopy coverage on average than neighborhood parks, some of these parks still struggle with low tree canopy. Below is a bar graph representing the bottom five community parks for tree canopy coverage.



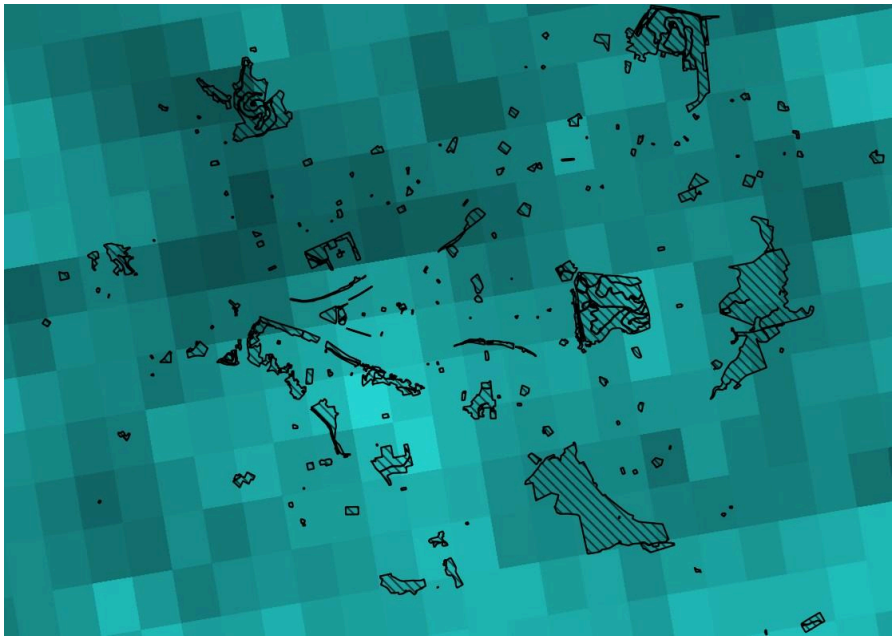
A notable point is that park acreage is moderately positively correlated with park tree canopy coverage. Smaller parks tend to have lower tree canopy coverage while larger parks have higher tree canopy coverage. We will explore this relationship further in the analysis section of this report.

We also calculated tree canopy coverage percentages for each walkshed. This allowed us to see the differences between a park's tree canopy and its surrounding area's tree canopy. We used the same method of zonal averaging to calculate each walkshed's tree canopy coverage percentage.

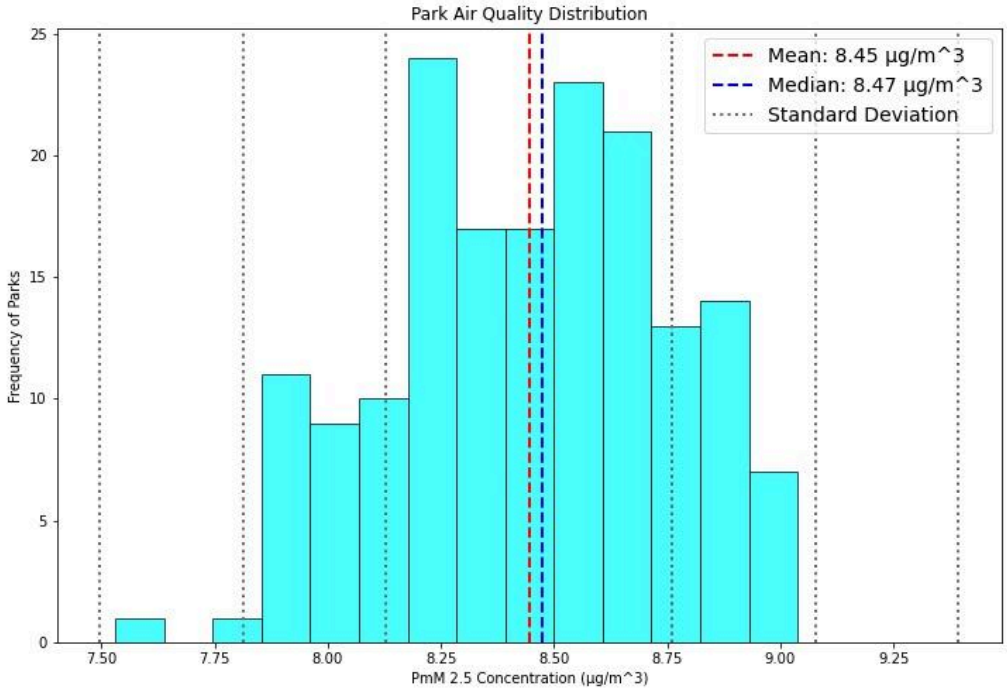
II. Data and Methodology → B. Environment → 2. Pollution

Pollution of a park in this report is measured by the average pm2.5 concentration measured in micrograms per cubic meter (μg/m^3). We used the technique of zonal averaging to calculate the average pollution within each park. This technique involves averaging all of the pm2.5 concentration levels (each being calculated in 0.01° × 0.01° or ~1.11 × .844 km areas) contained within each park and walkshed boundary. The variables of "Pollution_Park" and "Pollution_Walkshed" represent average pm2.5 concentrations and serve as an accurate indicator for pollution levels across the entire park or walkshed.

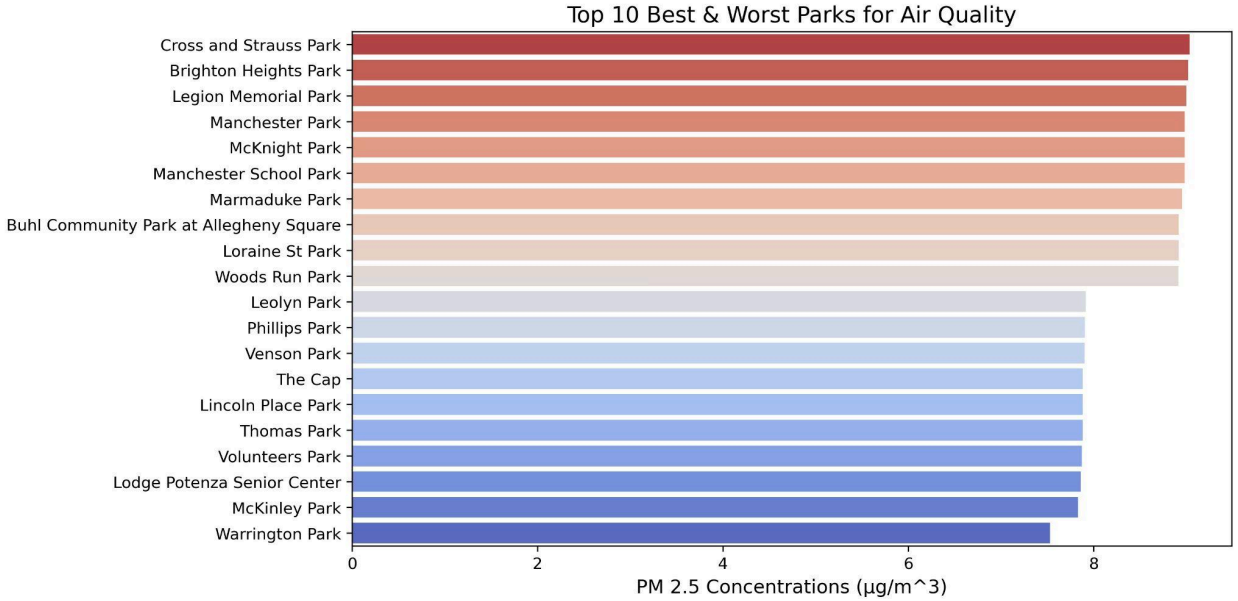
PM2.5 Concentration Levels of Pittsburgh



This graphic showcases each Pittsburgh park layered on top of a raster layer sourced from the Washington University in St. Louis' calculation of pm2.5 concentrations in North America. Areas with darker blue colors represent a higher concentration of pollution. Through calculating the zonal pollution averages of each park, we were able to compute the mean and median pollution levels for each park in Pittsburgh and rank each park from best to worst in terms of their air quality. The mean and median pm2.5 concentration levels across all parks within Pittsburgh is 8.45 μg/m^3 and 8.47 μg/m^3 respectively. This indicates a relatively normal distribution across all parks shown in the histogram below.



Parks in Pittsburgh vary slightly in their pollution levels with a standard deviation of .31 µg/m^3 and a range of 1.51 µg/m^3. All parks are within two standard deviations of the mean with the exception of Warrington Park. Warrington Park is a statistical outlier in low pollution levels. As shown in the raster layer graphic, the area of Mt. Washington has the least amount of pollution and therefore parks close-by like Warrington Park, Lodge Potenza Senior Center, Venson Park, and McKinley Park are among the parks with the best air quality. Below is a bar graph that showcases the top 10 best and worst parks for air quality.



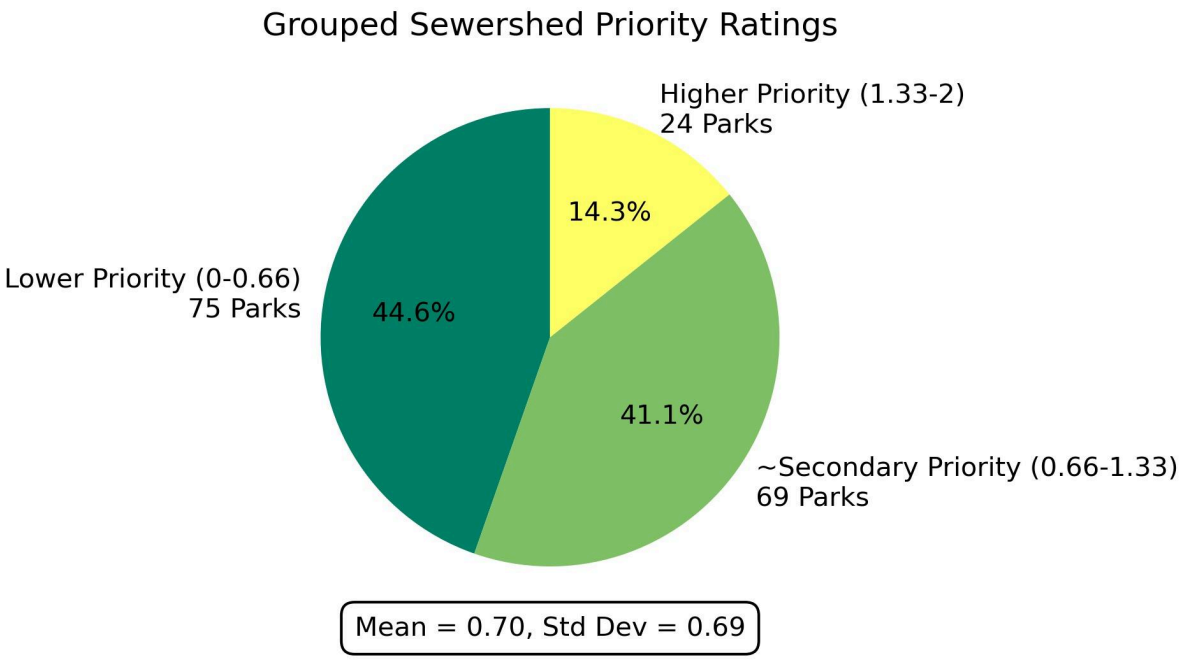
As shown in the raster layer graphic, parks on the North Side are within the area with the worst pollution levels in Pittsburgh's city limits. All 10 of the parks with the worst air quality are located north of the Allegheny River.

Pittsburgh as a whole is a city that is below average in terms of their air quality. As of February 7, 2024, the United States Environmental Protection Agency has set "the level of the primary (health-based) annual pm2.5 standard at 9.0 micrograms per cubic meter to provide increased public health protection, consistent with the available health science." The top three parks for worst air quality in Pittsburgh barely exceed this new standard of 9.0 $\mu g/m^3$.

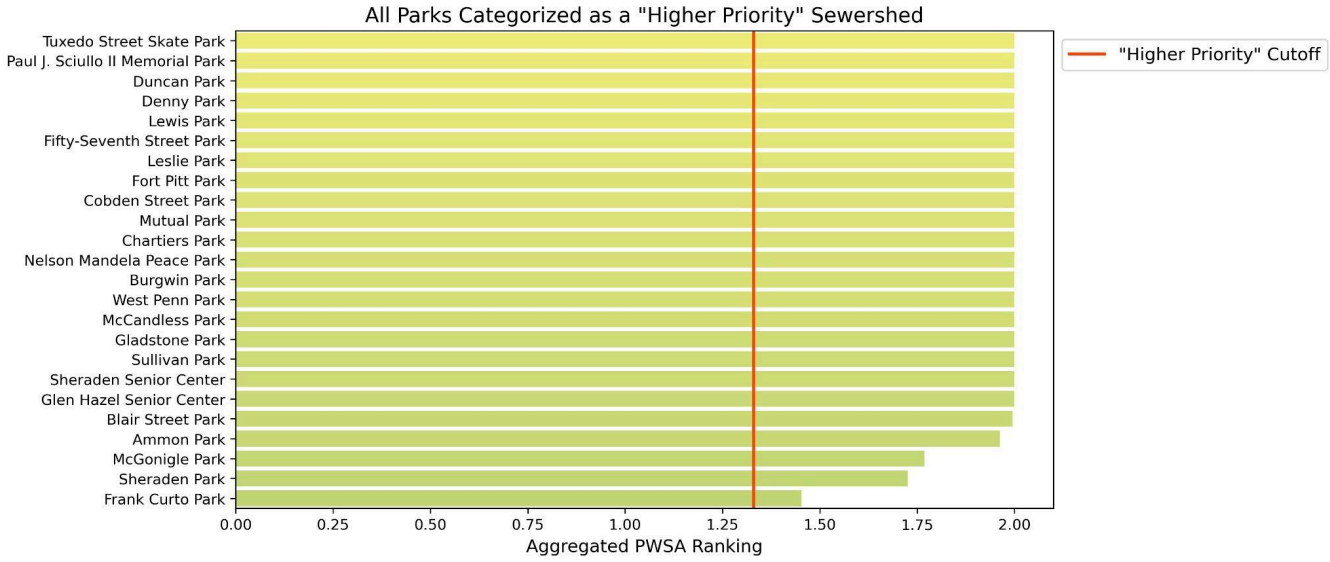
We also calculated pollution levels for each walkshed. This allowed us to see the differences between a park's pollution level and its surrounding area's pollution level. We used the same method of zonal averaging to calculate each walkshed's pm2.5 concentration average.

II. Data and Methodology → B. Environment → 3. Sewershed Priority

We used aerial interpolation to calculate the aggregated sewershed priority score within each park. The Pittsburgh Water and Sewer Authority (PWSA) has rated tracts of land within Pittsburgh based on their sewershed's need for intervention. The PWSA assigns a 2 to areas with high priority, a 1 to areas with secondary priority, and a 0 to areas with low priority. Through aerial interpolation, sewershed priority scores are attached to the parks and are weighted depending on how much of the park intersects with the boundaries created by the PWSA. The resulting weighted average of all intersections represent the final sewershed priority score, with the minimum of 0 being low priority and the maximum of 2 being high priority. Below is a pie chart representing three groups of sewershed ratings. Parks with an aggregated priority score of 0-.66 are labelled as "lower priority," .66-1.33 are labelled as "~secondary priority," and 1.33-2 are "higher priority."



The mean of .70 reveals that parks on average in Pittsburgh are located within areas with lower priority ratings. However, there are 24 parks, making up 14.3% of all parks, that exceed an aggregated score of 1.33 and can be labelled as "higher priority." Below is a bar graph that shows all of the parks that fall within this category of "higher priority."



Among the 24 parks that are categorized as “higher priority,” 19 of them have a priority rating of exactly 2. This means they are located fully within a PWSA sewershed boundary with the “high priority” rating. 5 of the parks have a rating greater than 1.33 but less than 2. This means the park intersects with a “high priority” boundary and one of the lower rated sewershed priority boundaries. However, all of these parks intersect the most with the “high priority” boundaries.

Two additional variables were calculated for explaining the parks’ sewershed priorities. These two variables, labelled as “percent_rank1” and “percent_rank2,” measure the percentage of the area of the park that exists within the sewershed priority boundaries of “secondary priority” and “high priority” respectively. These variables are especially useful in interpreting the aggregated priority rating. They show exactly what comprises each rating for a park by giving exact percentages of the parks’ intersected areas with all three types of sewershed rating boundaries.

II. Data and Methodology → C. Crime

We began the development of the crime data by collecting Pittsburgh police blotter data from the Western Pennsylvania Regional Data Center. We then combined the crime data with the walkshed and park spatial data provided by Interface to produce charts and statistics to provide insights into the relationships between parks and different crime statistics.

The police blotter data came with crime hierarchy descriptions, the location of the crime, and the date and time of the report. The crimes were each associated with a longitude and latitude point, making attachment of each crime to a park or walkshed simple after the data was cleaned. We cleaned the data by removing irrelevant columns from the dataset, as well as removing missing values. Due to the size of the dataset, this removal did not affect the quality of the data. Furthermore, we created crime datasets through organizing crimes by year in order to understand the changes in crimes over the period of 2016 to 2023. We selected these years due to the lack of accessible data from years preceding them.

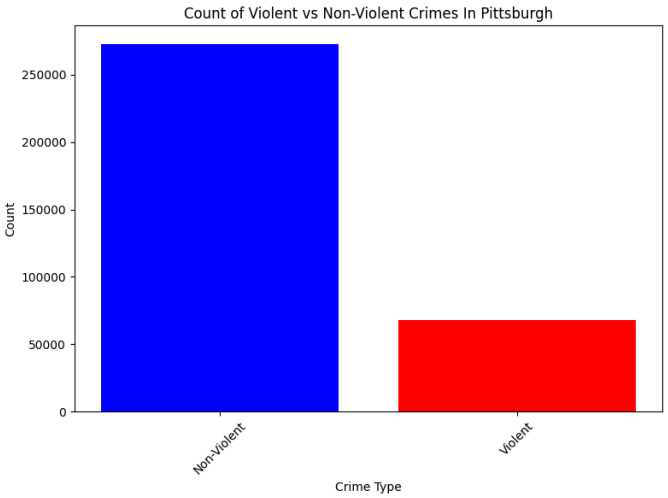
Using the walksheds and park geospatial polygons defined previously, we were able to use QGIS to attach each individual crime in the dataset to the parks. To ensure that the conclusions drawn were comparable, we computed the total counts of crimes, the crimes per acre, the violent crime count, the violent crime count per acre, and the nonviolent crime count and the nonviolent crime count per acre in each park and walkshed respectively.

Using these counts, we computed z-scores for each of the crimes counts per acre to create a method for ranking each park. To maintain interpretability, the visualizations we will provide in the analysis section will be using crime counts per acre, violent crime counts per acre, and nonviolent crime counts per acre.

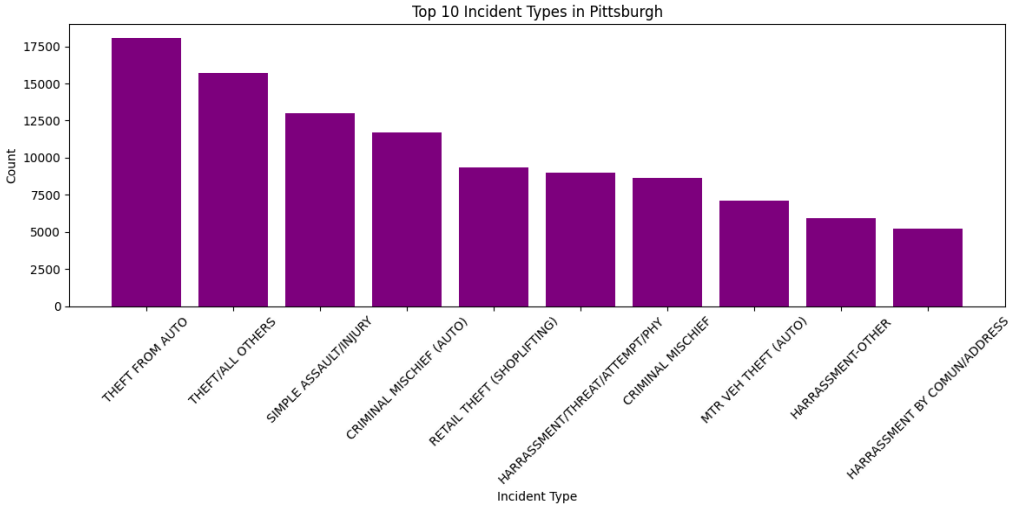
II. Data and Methodology → C. Crime → 1. Violent Crime & Nonviolent Crime

When preparing the data for analysis, we split the data into violent and nonviolent crimes. The police blotter crime data contained an Incident Hierarchy Description column containing types of crimes in an interpretable manner. To isolate the violent crime data from the nonviolent crime we analyzed each Incident Hierarchy Description and decided whether the crime was violent.

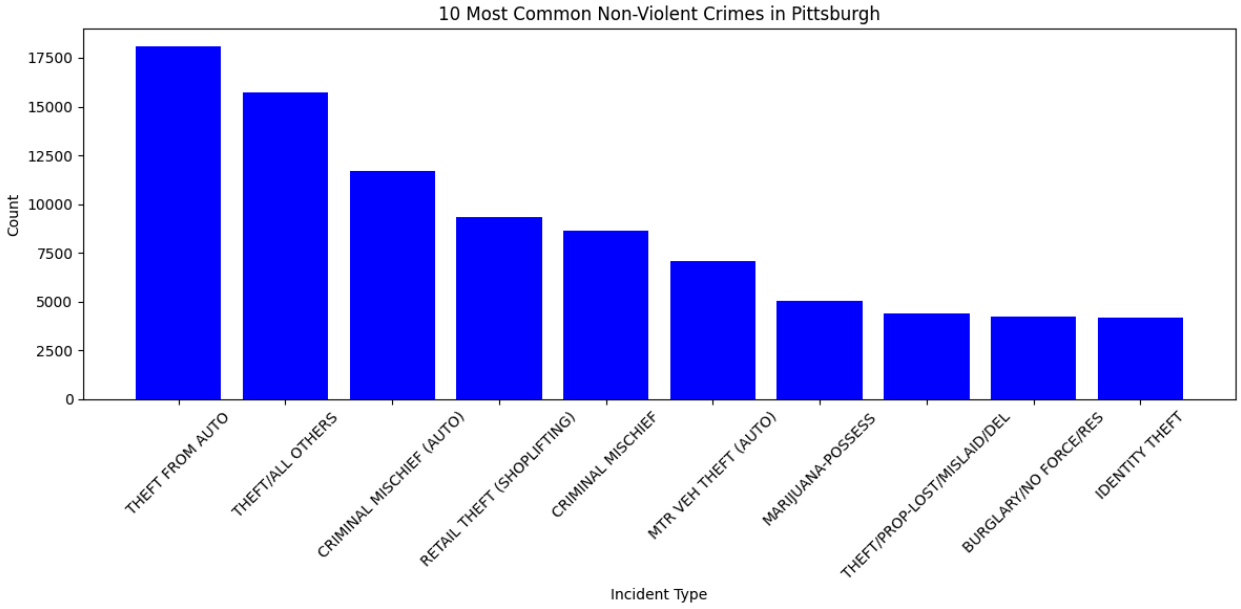
After defining these variables, we plotted the counts of violent and nonviolent crimes in order to understand the structure of the data.

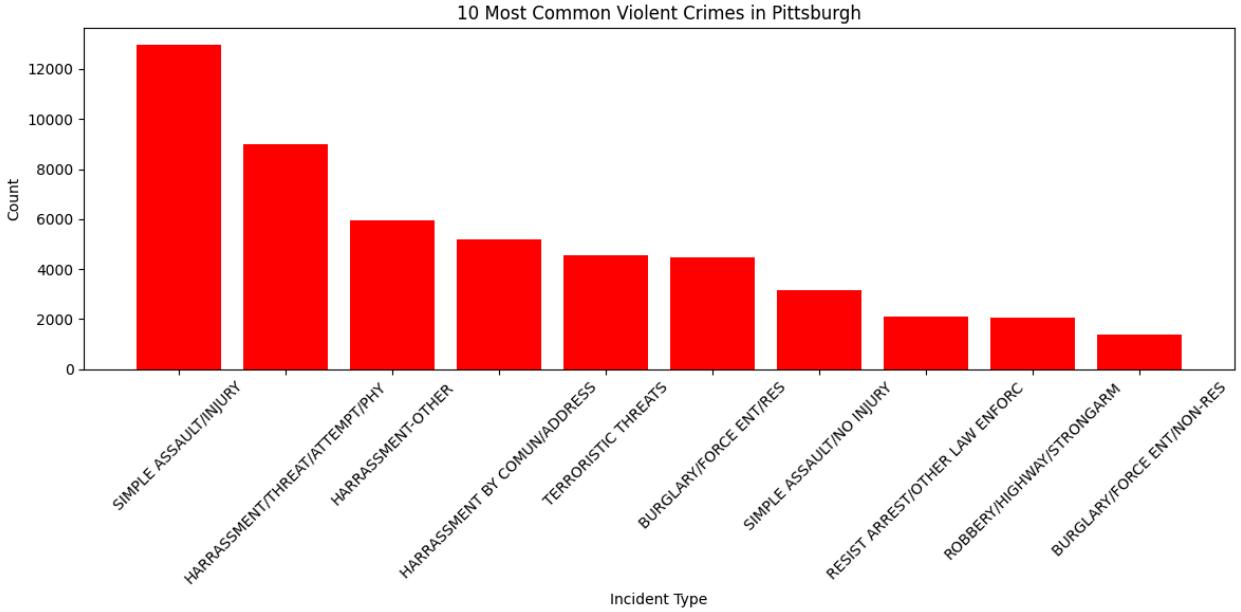


The distribution shows that nonviolent crime largely, and accurately, outweighs violent crime, ensuring that the data is not misrepresenting the true citywide distribution. After confirming this, we analyzed counts of each crime hierarchy item.



According to the Pew Research Center (2024), the most common crimes in 2024 in the United States were larceny theft, motor theft, burglary, and assault. With many of these crimes represented in the figure above, we can safely assume that the data represented a plausible distribution of these types of crimes. To further understand the distribution of the data, we looked at the 10 most common violent and nonviolent crimes in the total dataset too.





II. Data and Methodology → D. Public Health

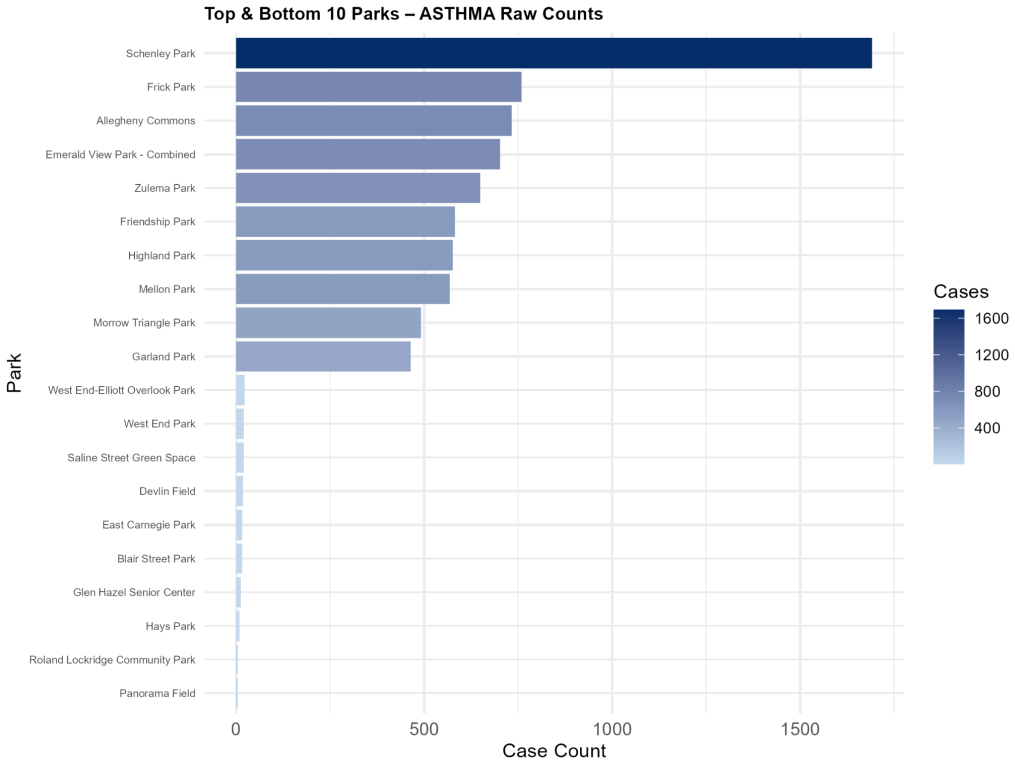
We obtained health metrics for this project from the CDC's PLACES database, which provides small area estimates of population health outcomes at the census tract level. Specifically, the dataset utilized, "PLACES: Census Tract Data 2023 Release," reports crude prevalence rates as percentages for various chronic conditions across U.S. Census tracts. The focus for this analysis was on four public health indicators: asthma, obesity, depression, and diabetes. These metrics were reported as crude prevalence among adults aged 18 years or older, alongside tract-level population estimates.

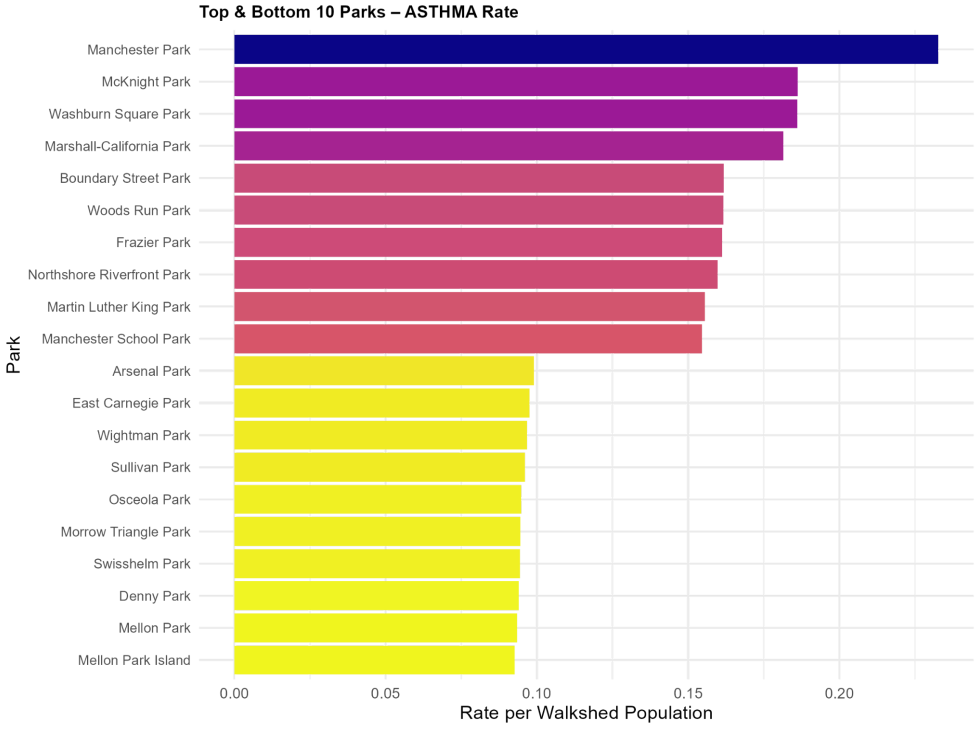
Spatial data for park walksheds were provided by the Interface project files, but the dataset included outdated disease metrics from prior CDC data. The dataset also included polygon geometries for each park's walkshed, and other metrics identifying individual park sites.

To integrate health data with walkshed geography, we first joined Census tract-level health prevalence rates with TIGER/Line shapefiles, which provided the official geographic boundaries for all U.S. Census tracts. This step ensured that each row of CDC health data was linked to a valid spatial polygon, enabling geographic operations. Once we applied the CDC health data with spatial geometry, we overlaid it onto the walkshed polygons using spatial joins in R's spatial features package. This allowed us to estimate the health burden within the area accessible by foot around each park. For walksheds that intersected multiple Census tracts, we calculated weighted population values based on the proportion of the Census tract that overlaps with the walkshed.

To enable comparison across diseases and parks, we calculated disease prevalence rates for each health metric, along with z-scores. This facilitated the identification of parks associated with relatively higher or lower health burdens. We later used the standardized scores for ranking parks and visualizing disparities.

II. Data and Methodology → D. Public Health → 1. Asthma

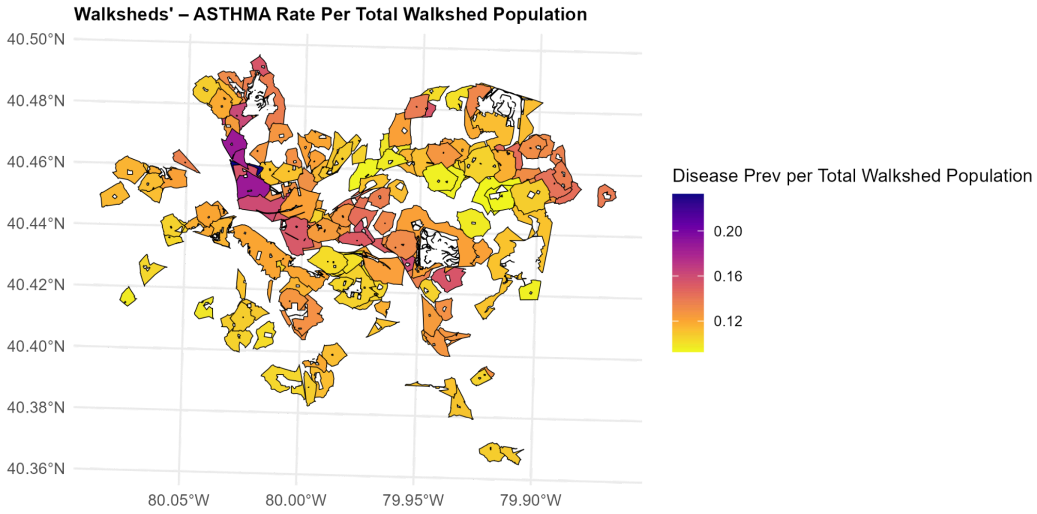




The two complementary bar graphs display the distribution of asthma burden across park walksheds in Pittsburgh. One shows relative burden – proportion of local population with asthma – and the other shows absolute burden – total estimated number of individuals with asthma in each walkshed.

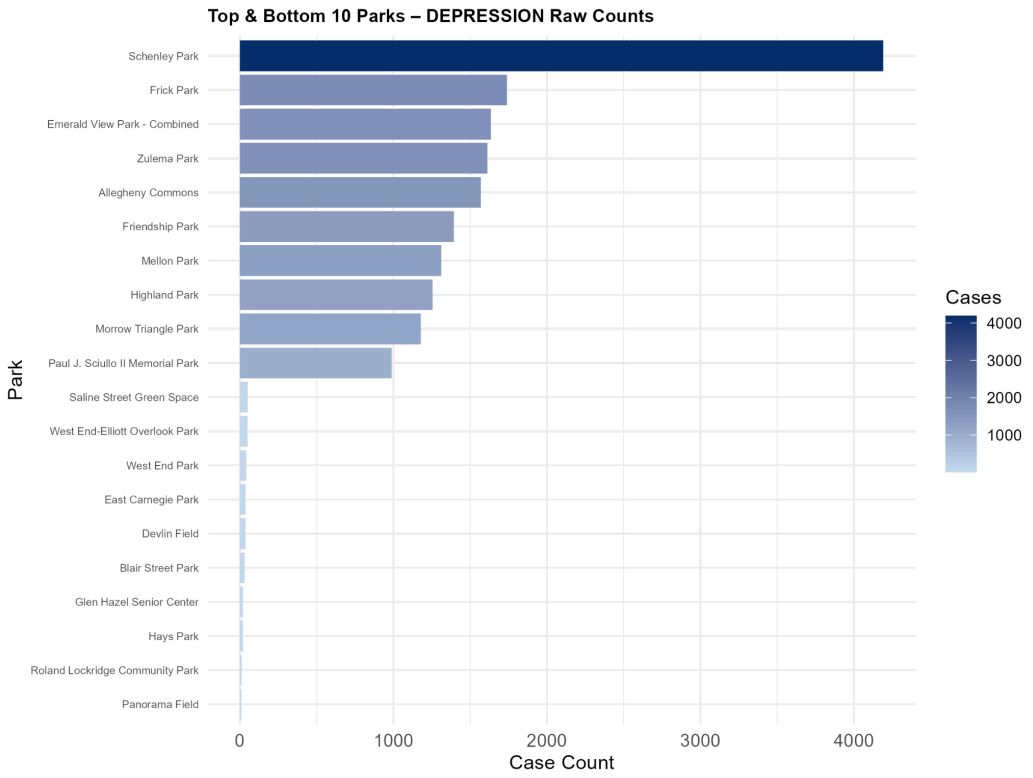
The first chart, raw counts, highlights the top 10 and bottom 10 parks by absolute burden by estimating the total number of people with asthma per walkshed. Larger parks such as Schenley Park, Frick Park, and Allegheny Commons serve high population areas and thus account for the largest absolute number of asthma cases.

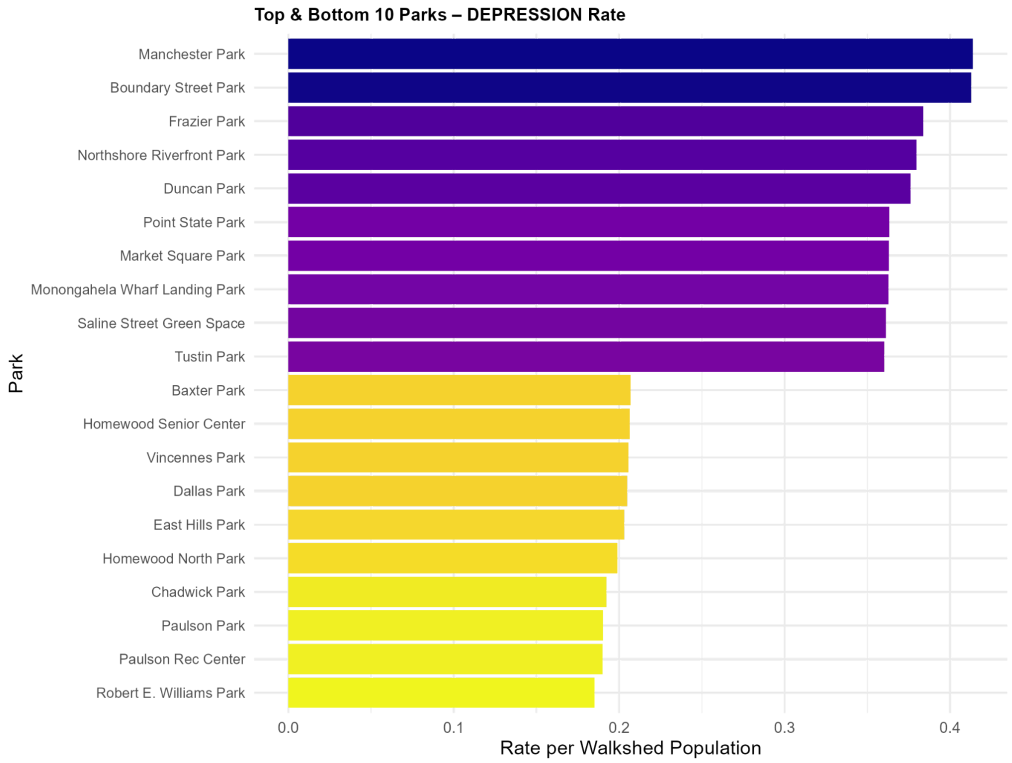
The second chart highlights the top 10 and bottom 10 parks by the proportion of residents within each walkshed reporting asthma. This metric identifies neighborhoods where asthma rates are disproportionately high, even if the overall number of residents is low. Manchester Park, McKnight Park, and Washburn Square have the highest asthma prevalence relative to their local populations. Manchester Park stands out with a significantly greater asthma rate than all other parks.



The choropleth map provides a geographic image displaying the relative asthma rates. The North Side contains the largest burden of asthma. Two parks, Morrow Triangle Park and Mellon Park, fall into both high-rate and high-total categories. This indicates there may be a high disease prevalence rate and an overall high total number of affected individuals. It may suggest that residents here suffer from poor air quality, socioeconomic disadvantages, or both.

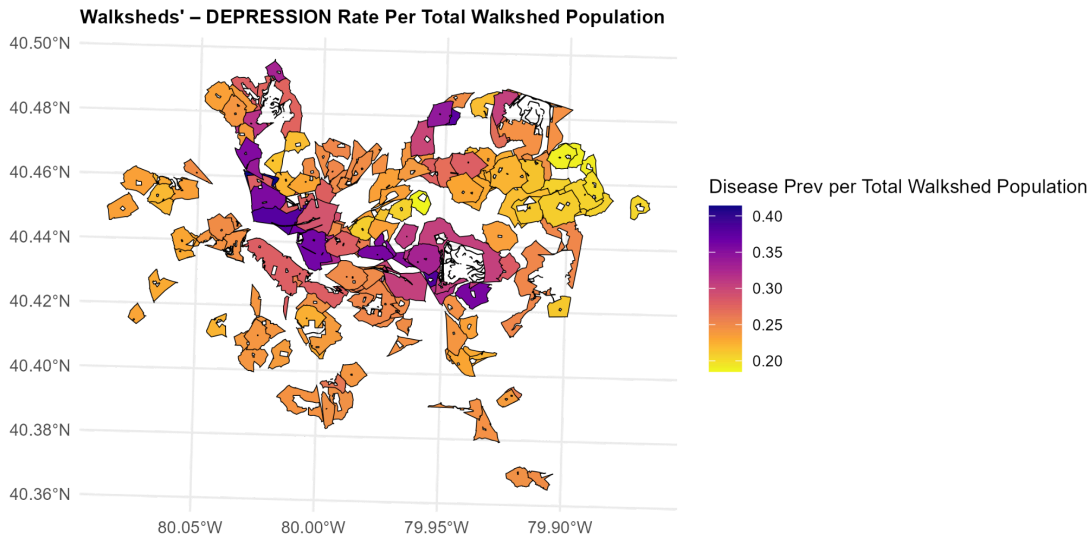
II. Data and Methodology → D. Public Health → 2. Depression





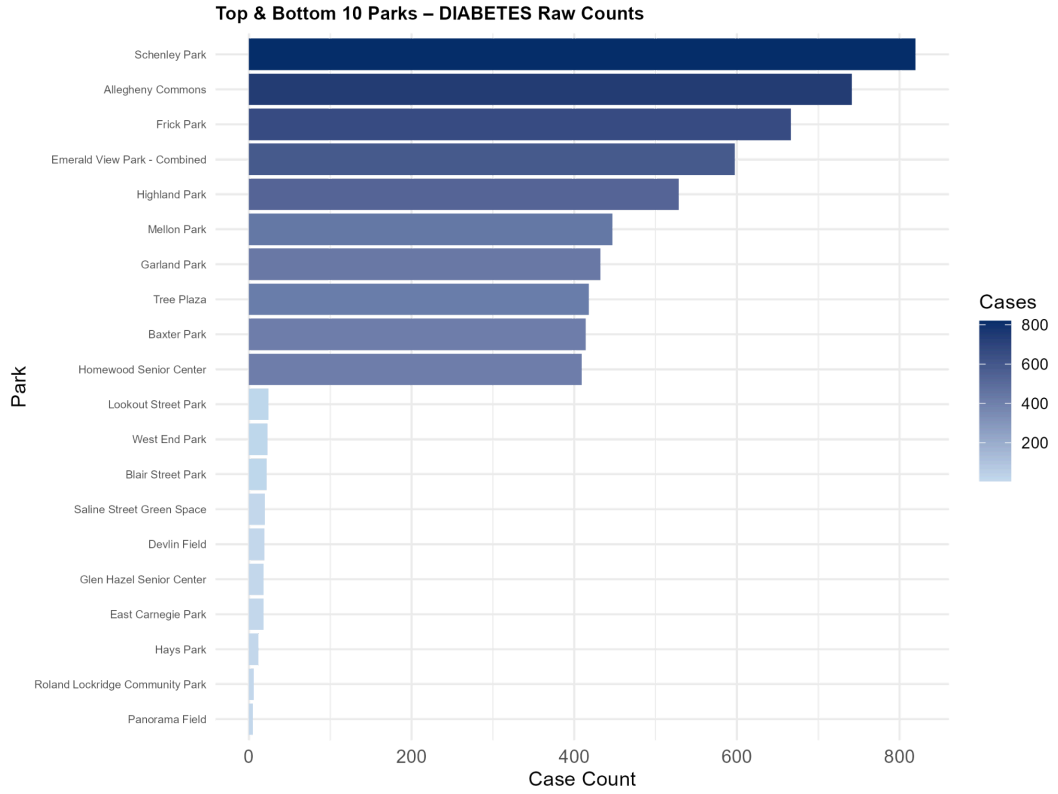
We summarized our analysis of depression prevalence across Pittsburgh park walksheds in two bar graphs. The first graph shows absolute burden – the total estimated number of individuals with depression in each walkshed – and the second graph shows relative burden – the proportion of the population in the walkshed with depression.

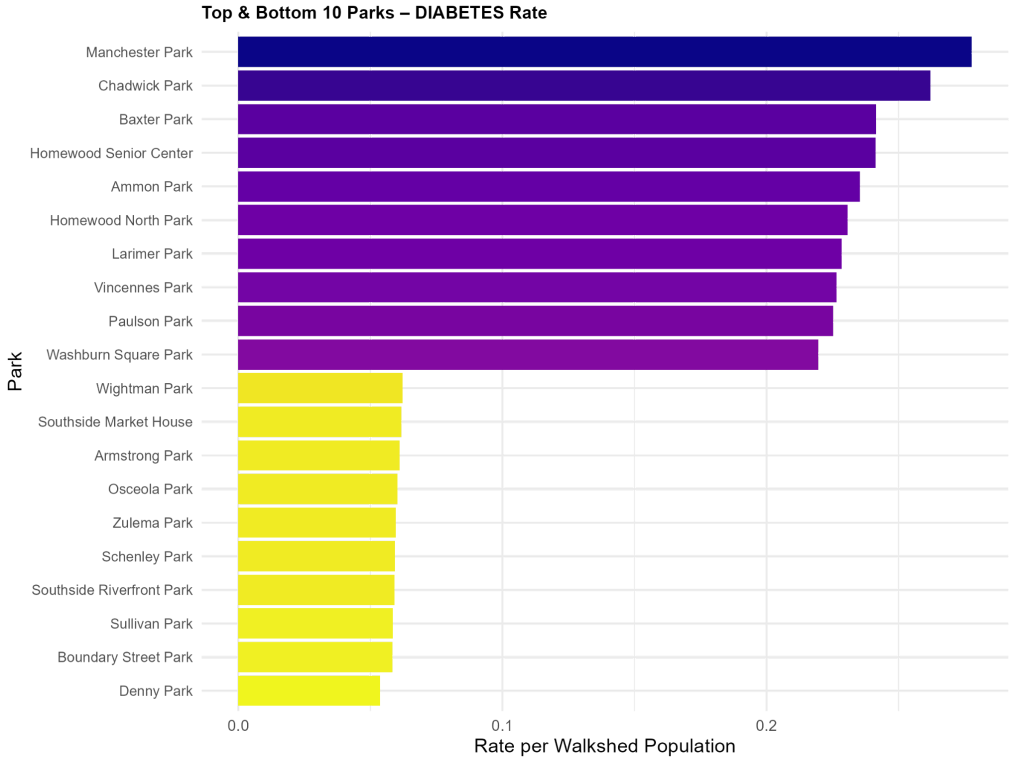
Absolute burden is greatest in larger regional parks including Schenley, Frick, and Emerald View Parks, which serve dense neighborhoods with high total populations. Central and South Oakland are home to a large off-campus student population and a rapidly growing underserved community, as highlighted by the Allegheny County Community Need Index in 2021.



Relative burden is highest around smaller urban parks such as Manchester Park, Boundary Street Park, and Frazier Park. The choropleth map spatially illustrates the concentration of relative burden in the North Side, Downtown, and South Oakland. It is worth noting that both Duncan Park and McCandless Park exhibit disproportionately high relative rates of depression compared to surrounding parks.

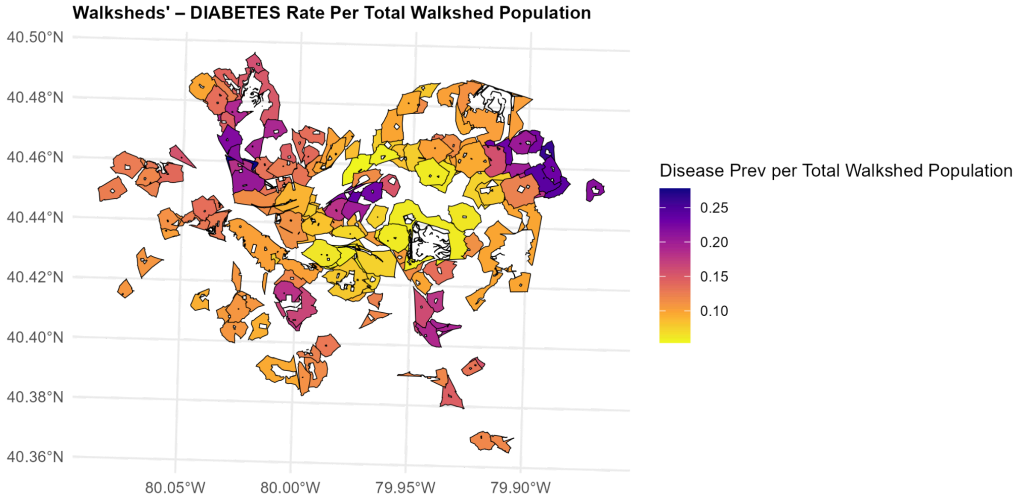
II. Data and Methodology → D. Public Health → 3. Diabetes





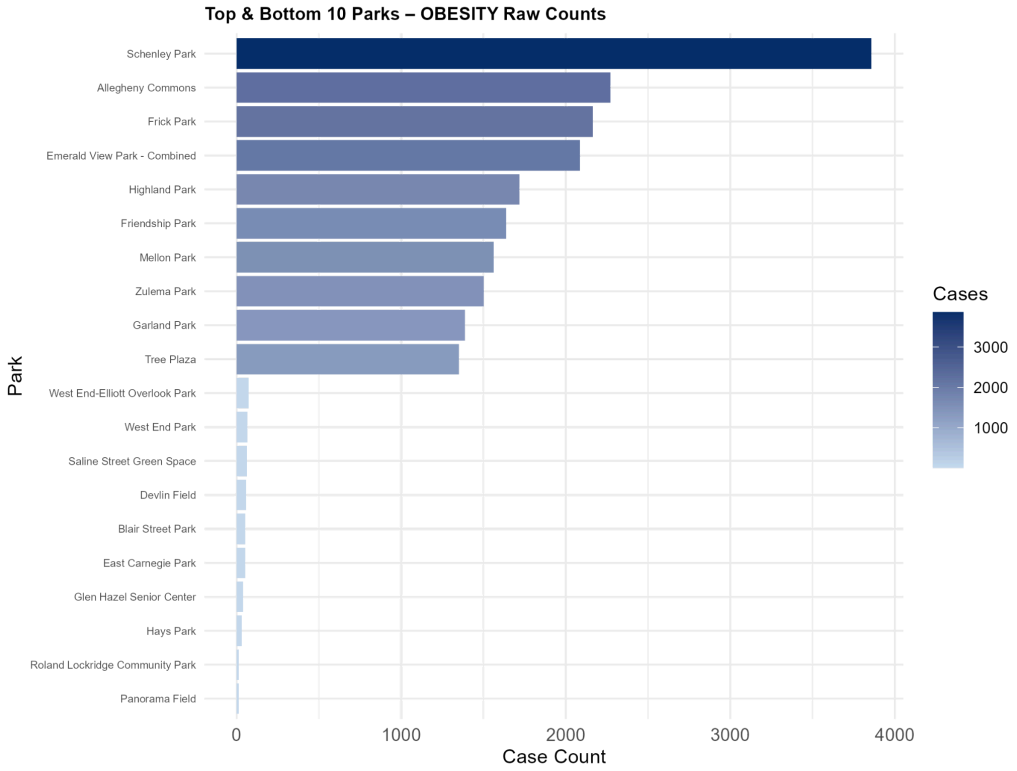
Our analysis of diabetes prevalence follows the presentation of asthma and depression, with two graphs displaying identical metrics. The first shows absolute burden, the raw estimate of walkshed residents with diabetes. The other shows relative burden, the proportion of residents each walkshed with diabetes.

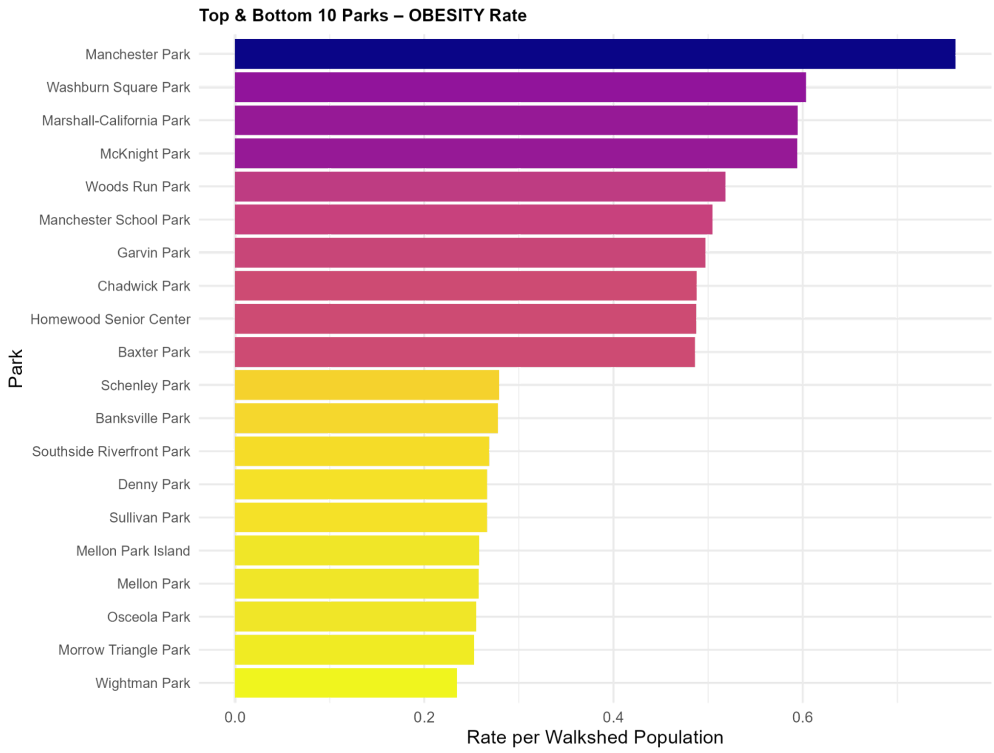
Absolute burden, summarized by the raw counts bar graph, is highest around larger parks including Schenley Park, Frick Park, and Allegheny Commons, similar parks to high-ranking walksheds for other absolute metrics. However, these numbers reflect population density more than community risk for diabetes. The parks that rank at the top for absolute burden are not the same as those for relative burden – and neither are the neighborhoods they represent.



Relative burden is highest around Chadwick Park, Baxter Park, and the Homewood Senior Center. These parks are located in historically underserved communities, where residents may face structural barriers to managing chronic health conditions – including limited access to health food, preventative care, and safe physical activity spaces. The choropleth map of diabetes rates reinforces this idea, highlighting distinct clusters of elevated prevalence of diabetes in the Hill District, Homewood, and East Hills.

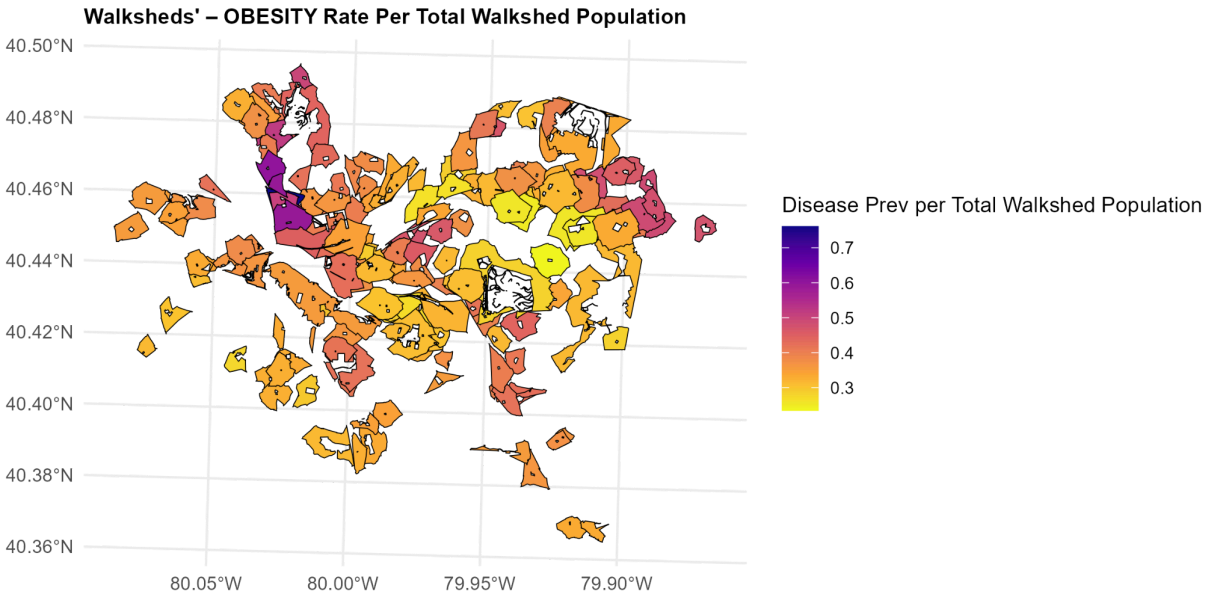
II. Data and Methodology → D. Public Health → 4. Obesity





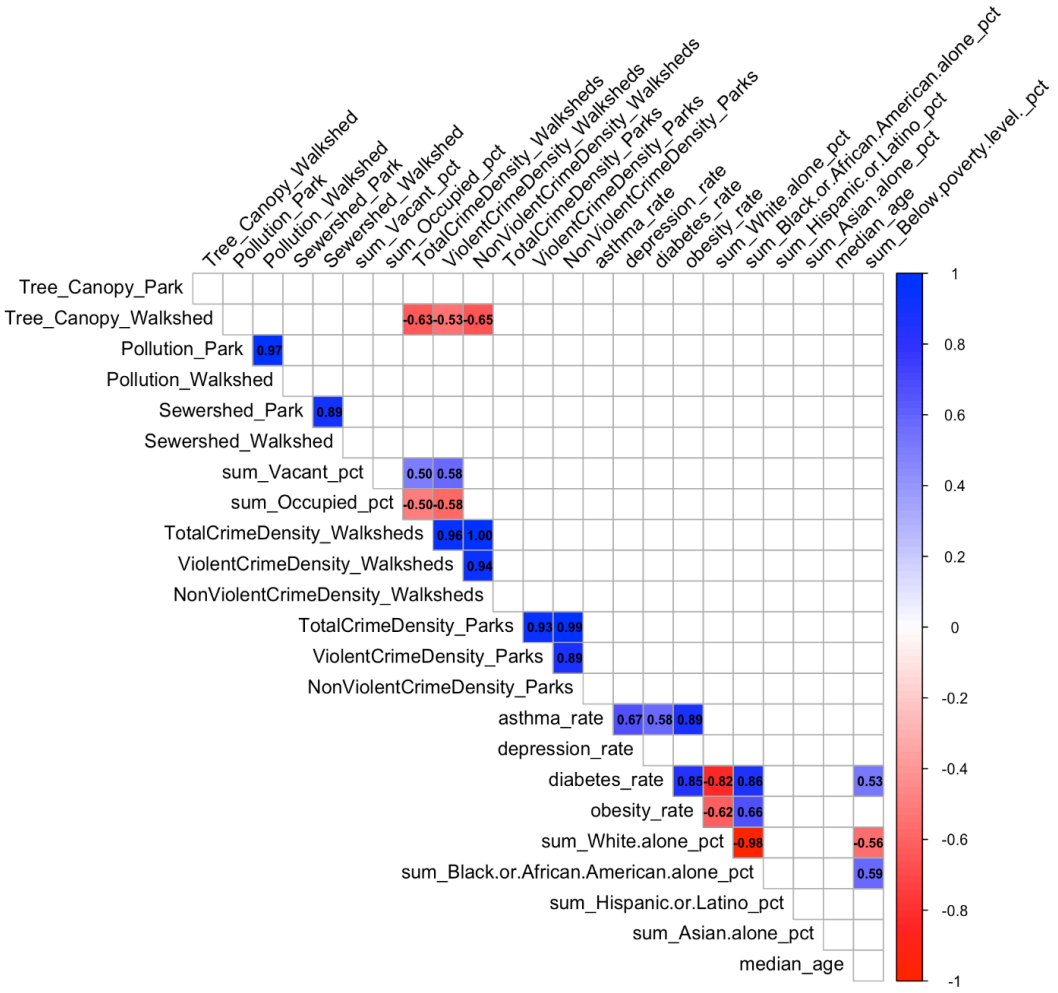
Similar to our analysis of other public health metrics, we displayed obesity prevalence across walksheds in two bar graphs, the first showing the absolute burden and the second showing the relative burden.

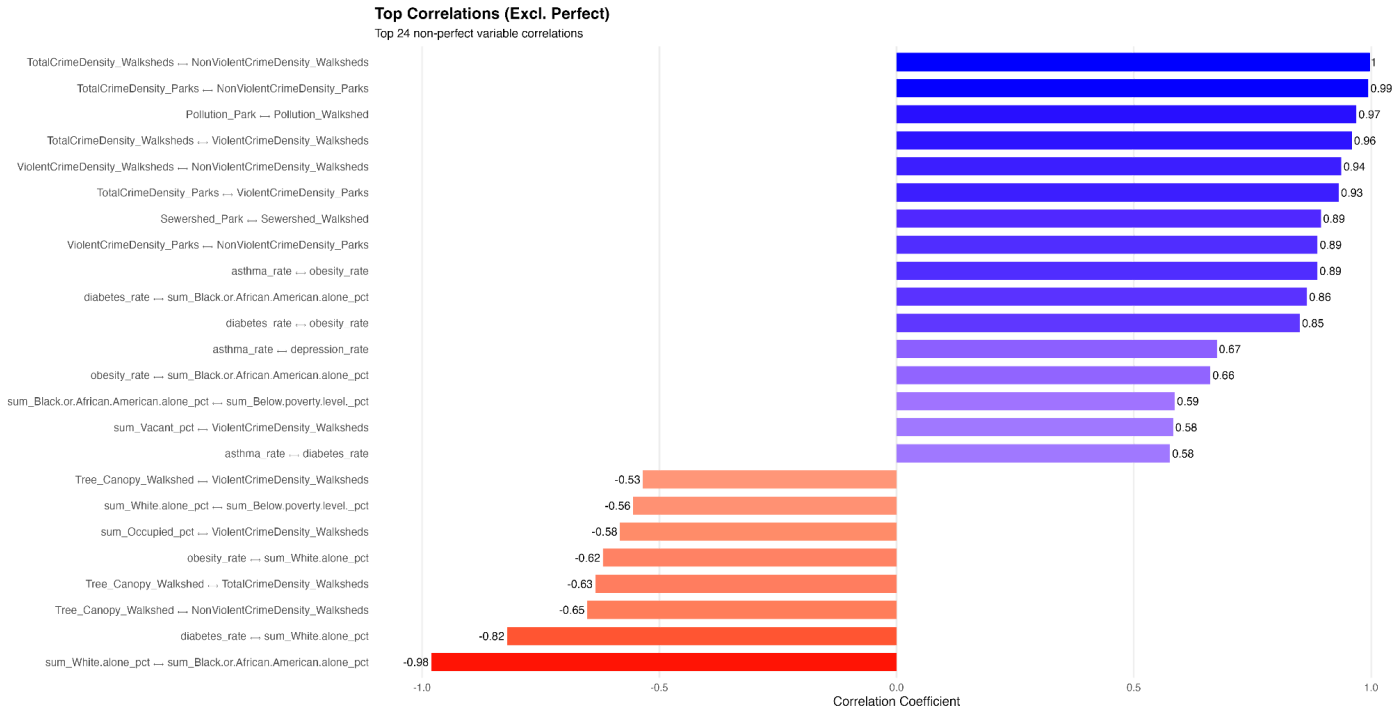
Allegheny Commons is located in the Central North Side of Pittsburgh, north of the Allegheny River. As we saw in our geospatial illustrations of asthma and depression, the relative burden of obesity is highly concentrated on the North Side. The parks that face the highest relative burden of obesity include Manchester Park, Washburn Square Park, Marshall-California Park, and McKnight Park, all of which are located on the North Side.



III. Analysis

After building a full dataset of the relevant variables, encompassing cleaned data from the Census, environment, health, and crime portions of the study, we specified population- and size-adjusted variables and created a correlation heatmap of the relationships between variables. The heatmap below contains a threshold of 0.5; hence, only stronger correlations appear. The 24 qualifying non-perfect correlations – to exclude self-evident perfect relationships like vacancy vs. occupancy – above the threshold is displayed in the bar plot below.

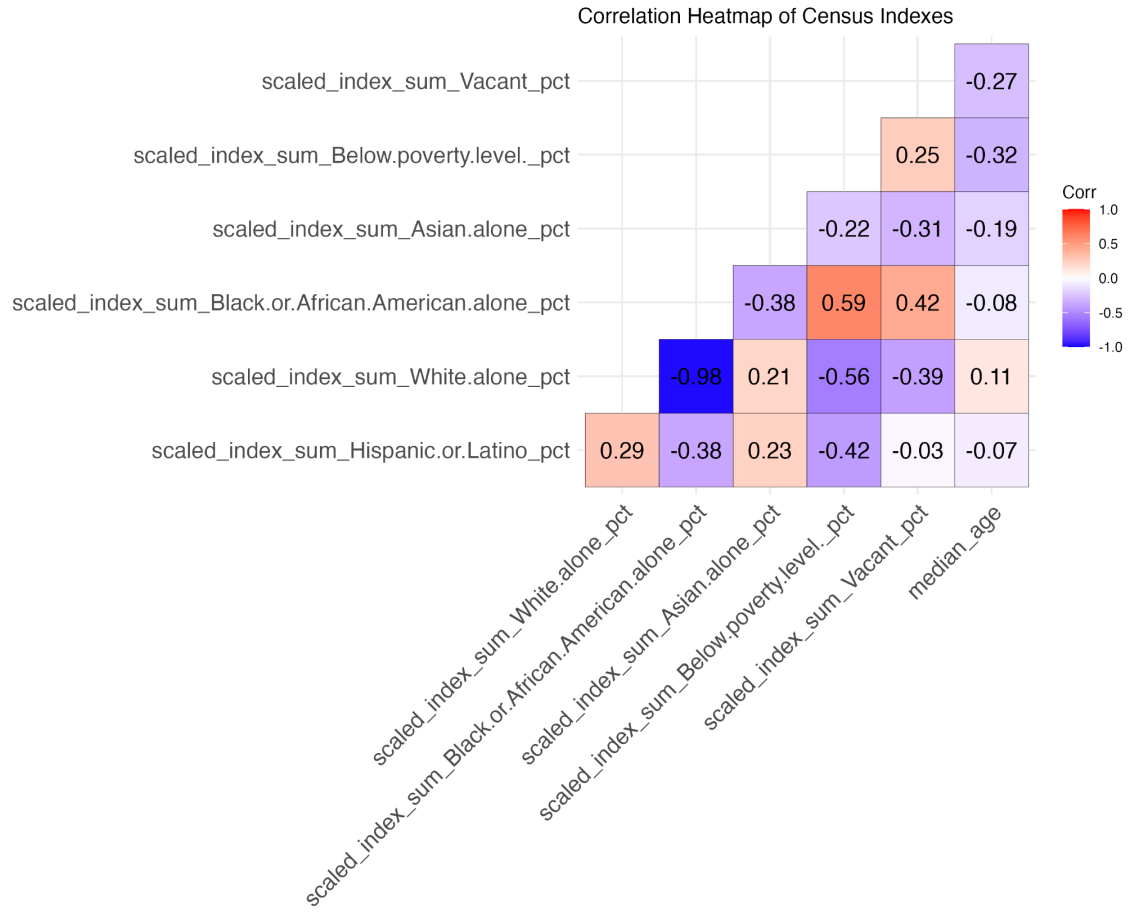




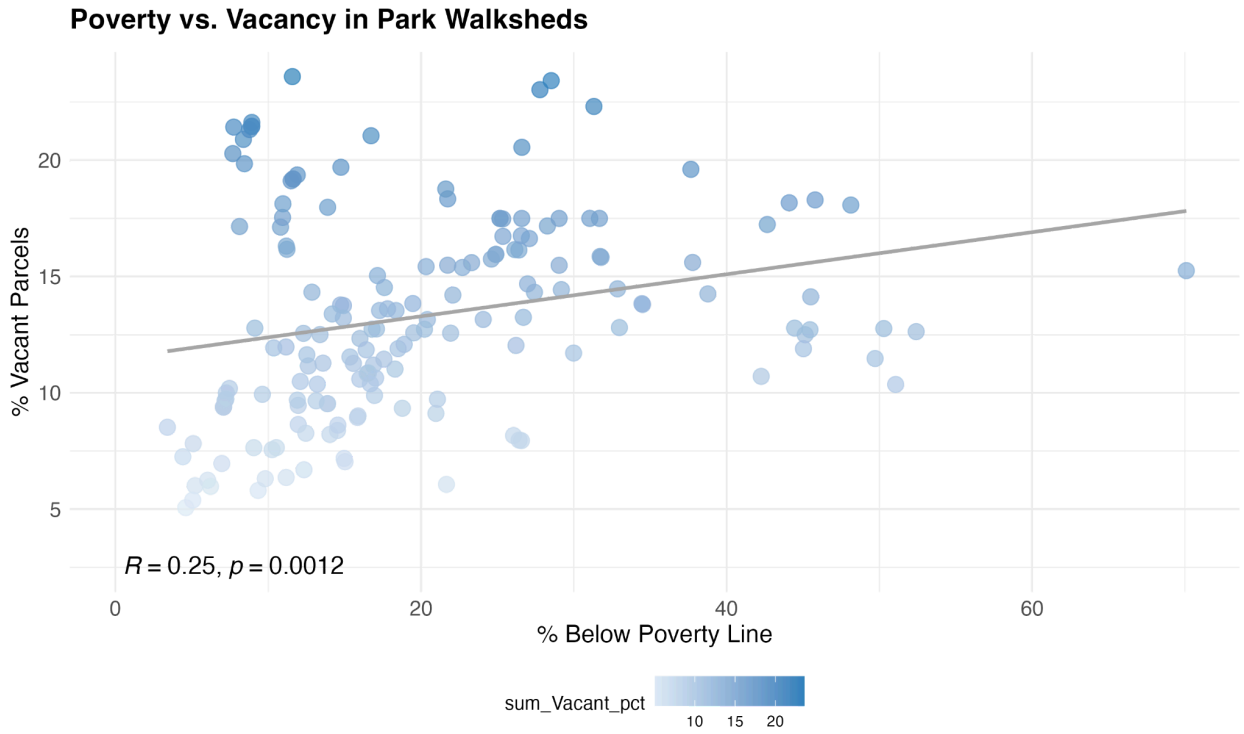
The strongest positive correlations are predictable – walksheds and parks with high rates of violent crime also have high rates of nonviolent crime. A walkshed that bears a quality – i.e. having severe pollution – likely shares that quality with the respective park. The relatively weaker correlations above the 0.5 threshold are more notable – we detected strong positive relationships between rates of asthma and obesity, diabetes and obesity, and asthma and depression. The diabetes and obesity rates were disproportionately high in walksheds with higher proportions of Black residents. Those walksheds with larger Black populations also saw higher proportions of residents under the poverty line, and the opposite was true for walksheds with larger White populations.

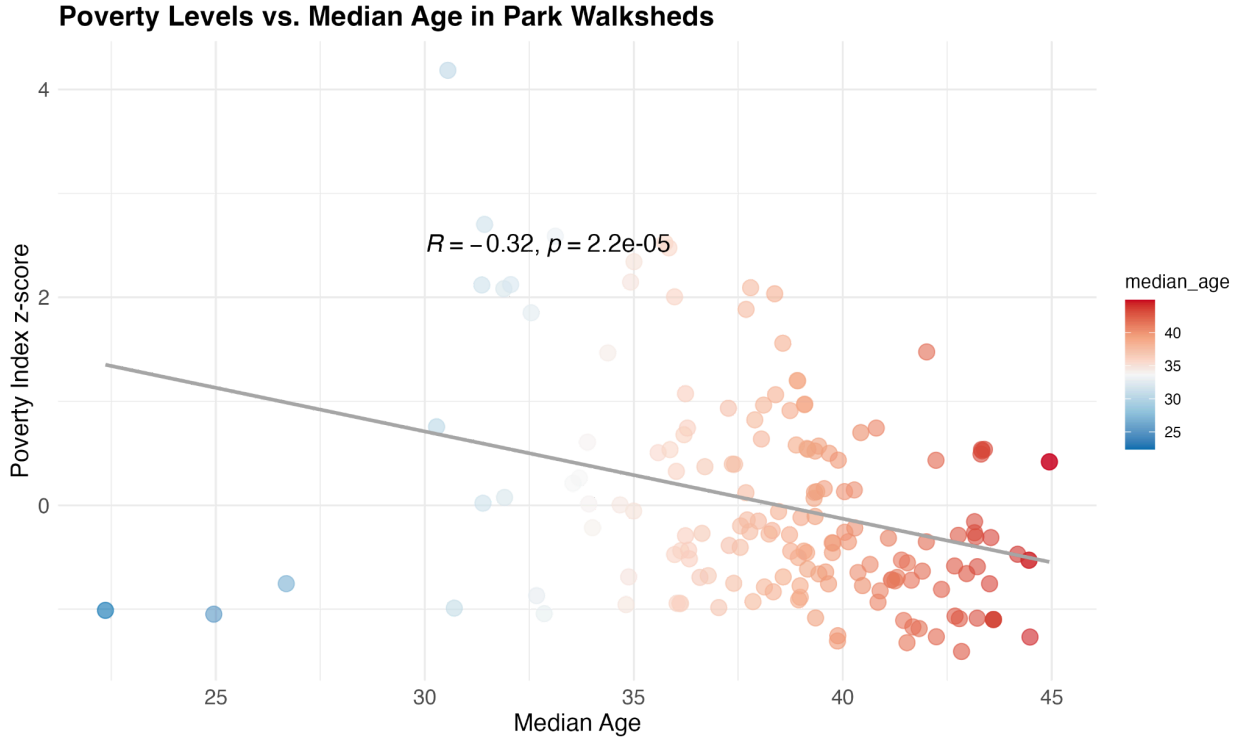
Vacancy was positively associated with violent crime, and likewise, occupancy was negatively associated with violent crime at the exact same strength. Walkshed tree canopy was negatively associated with total and nonviolent crime, and diabetes rates were strongly negatively associated with the proportion of White residents. The strongest negative association was between proportions of Black and White residents in walksheds, a predictable result considering the groups' status as by far the largest racial categories in Pittsburgh.

III. Analysis → A. Census

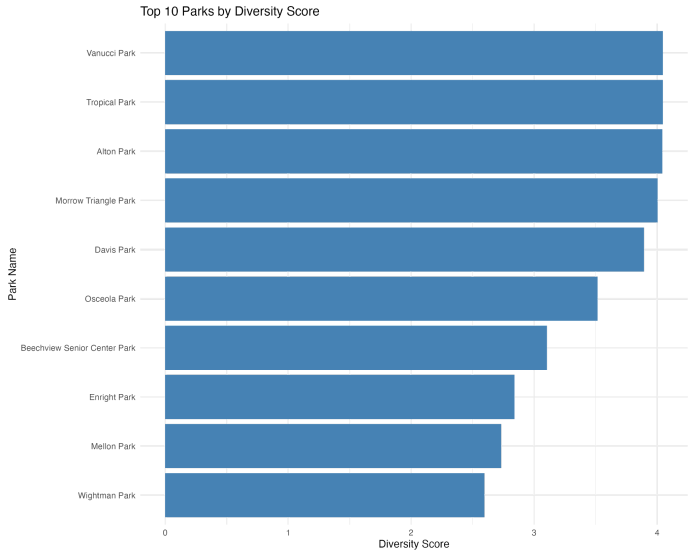


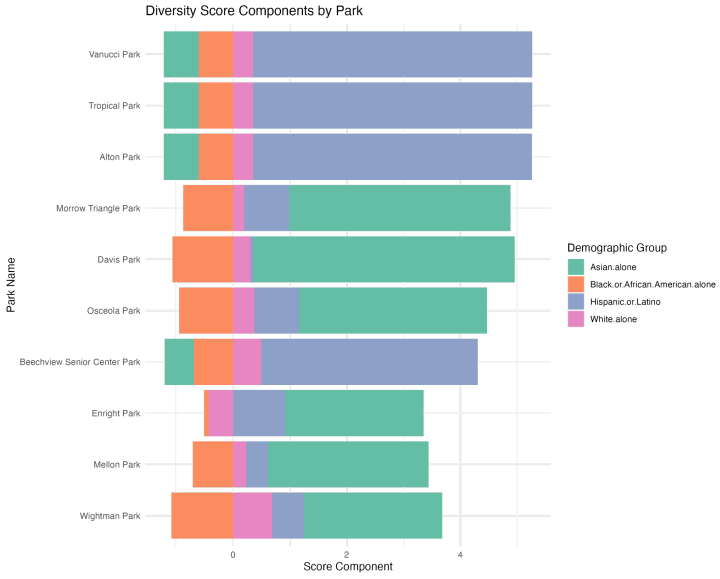
A heatmap displaying the correlation coefficients of the index values derived from Census data shows a very strong negative association between scaled values of White and Black populations in walksheds. Among the racial indexes of the walksheds, all had negative associations with poverty and vacancy apart from the Black population index. Only White residents had a positive correlation with median age, suggesting that on average, White residents are older than other racial categories in park walksheds. There is a positive relationship between poverty and vacancy and a negative relationship between poverty and age. We expanded upon these two relationships in the scatterplots below – both were relatively weak but highly significant correlations.


Poverty vs. Vacancy in Park Walksheds
% Vacant Parcels
% Below Poverty Line
R = 0.25, p = 0.0012
sum_Vacant_pct
10
15
20


Poverty Levels vs. Median Age in Park Walksheds
Poverty Index z-score
Median Age
R = −0.32, p = 2.2e-05
median_age
40
35
30
25

An intersectional approach to analyzing racial index data is considering park walksheds in terms of their ethnic diversity. The two bar plots below display "diversity scores" of the walksheds – we added the z-scores from each racial index to create a value representing the diversity of the walkshed. For instance, a walkshed with a disproportionately high number of White residents and very low relative populations of other racial categories would have a negative score – the positive White z-score plus the negative z-scores for other categories. In the first plot we displayed a ranking of most diverse walksheds by this metric, and in the second we broke down the components of each walkshed's diversity score by the positive or negative weights of each component. This process can be replicated for any given walkshed by adding the indexed values of each racial category.



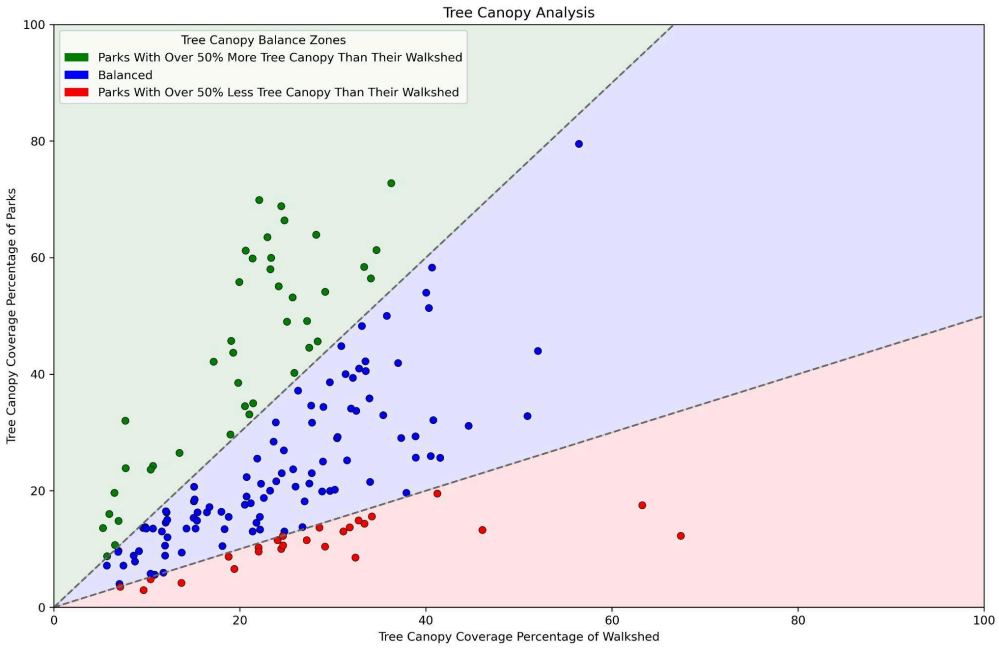


III. Analysis → B. Environmental

Among the three environmental variables included in the report, tree canopy coverage is the feature with the highest variability and the most geographically precise measurements across all parks. This is due to the highly detailed dataset offered by the U.S. Department of Agriculture Forest Service. Furthermore, improving tree canopy coverage within parks is more realistically actionable for the Pittsburgh Parks Conservancy than improving pollution levels and sewershed ratings, since the two latter variables are more subject to external forces. Therefore, this section will provide a further analysis by exploring relationships between tree canopy coverage and other variables to gain a deeper understanding of the parks that need intervention.

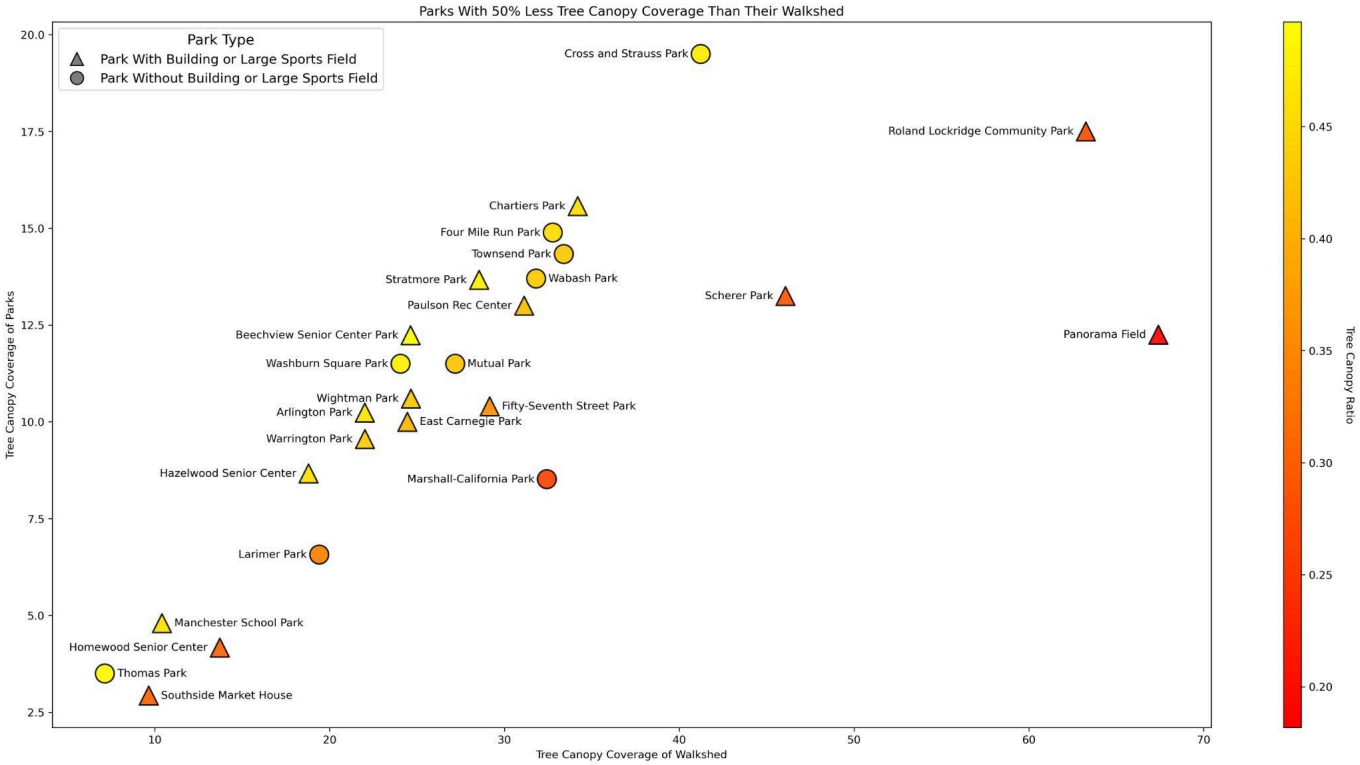
Tree Canopy: Parks vs. Walkshed Analysis

Tree canopy coverage was also calculated for each park's walkshed as it allows us to see the difference between the park's tree coverage and that of its surrounding areas. With this information, we can see which parks offer more, similar to, or less tree canopy than their surrounding areas. For this analysis, we split all parks into three categories. The first category is parks with 50% more tree canopy coverage than their walkshed. These parks excel with their coverage relative to the area of the city that they are in. The second category is parks with 50% less tree canopy coverage than their walkshed. These parks lack tree canopy relative to their surrounding areas. The final category is labelled as "balanced," as it contains all parks between the two former categories.

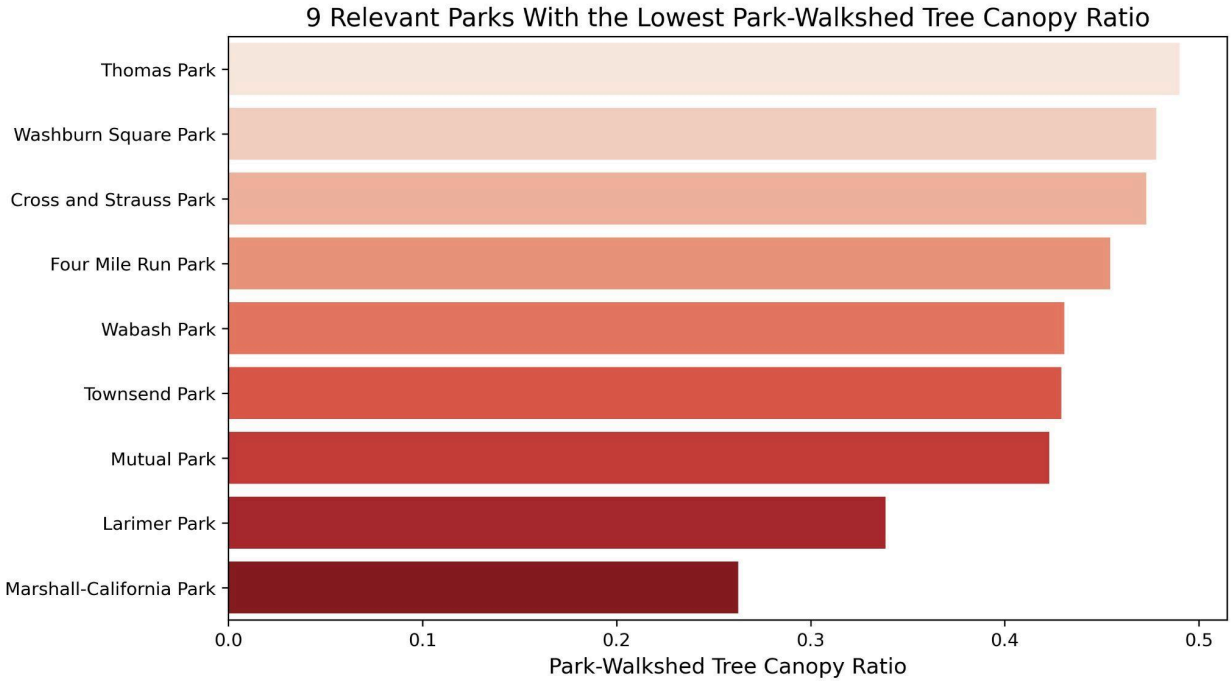


The graph above visualizes the three categories by color and separates each category with a dashed line that represents the cutoff between each category. Contained within the red category are parks that lack coverage compared to their walkshed, and they total 25 parks and account for 14.9% of the total parks. Contained within the green category are parks that excel in coverage compared to their walkshed, and they total 41 parks and account for 24.4% of the total parks. The majority of the parks are "balanced," and they total 102 parks and account for 60.7% of the total parks.

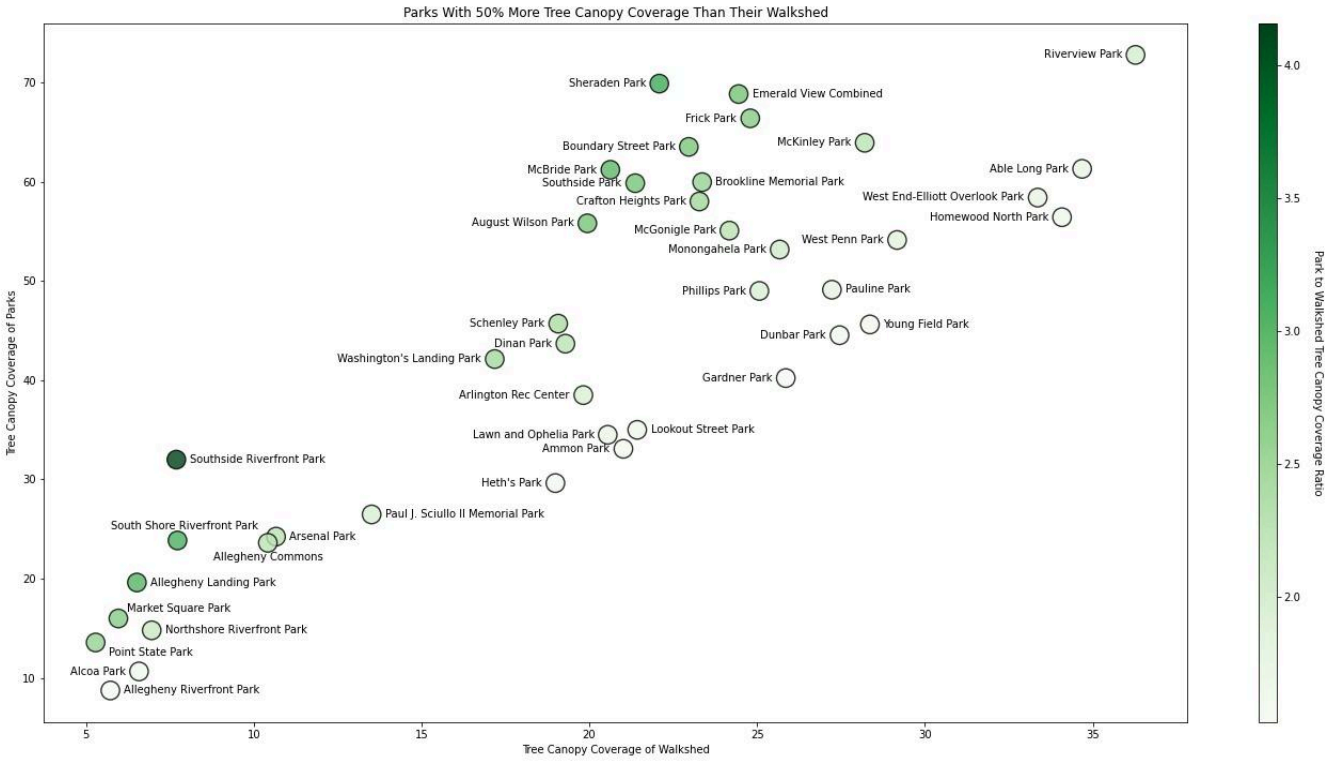
Below is a scatterplot that captures only the red section of the plot above. This section represents parks that lack in tree coverage compared to their walkshed. Each park is labelled within the graph and they are color coded by their park-walkshed tree canopy ratio. Points that are more yellow are closer to 50% less coverage than their walkshed while points that are more red are closer to around 80% less coverage than their walkshed. As mentioned in the methodology section above, parks that are buildings or contain sports fields are expected to have low coverage. Therefore, they are plotted as a triangle to showcase that they are expected to have low tree canopy. Therefore, the plot below will show two types of parks: parks that are buildings or sports fields and parks that are not.



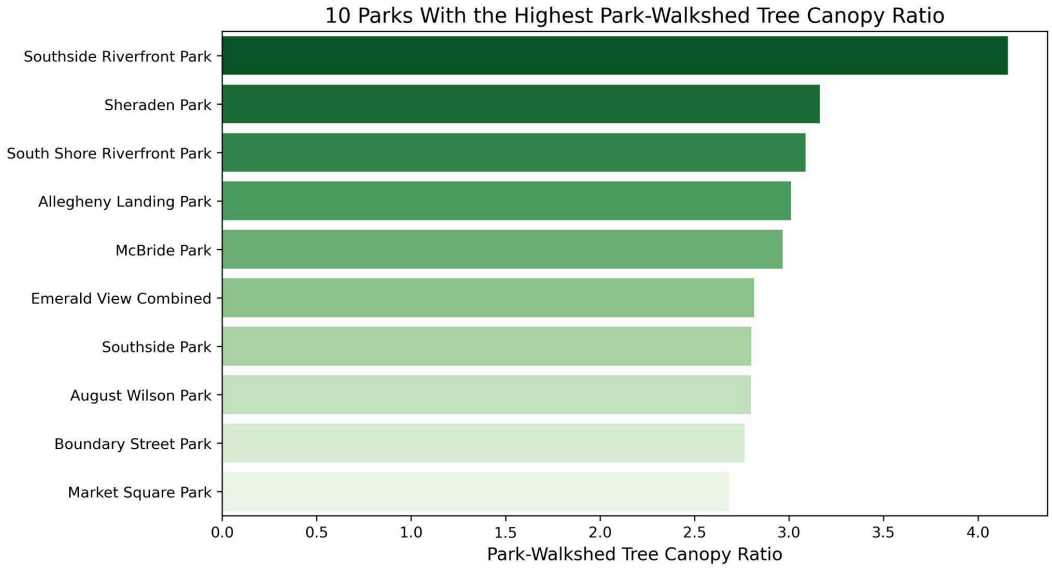
The bar graph below is a visual representation of the relevant parks with the worst park-walkshed tree canopy ratio in Pittsburgh. The bottom three park-walkshed tree canopy ratios that are not buildings or specifically limited to sports fields belong to the parks of Marshall-California Park, Larimer Park, and Mutual Park.



On the following page is a scatterplot that captures only the green zone of the scatterplot at the beginning of the environmental analysis section. This zone represents parks that excel in tree coverage compared to their walkshed. Each park is labelled within the graph and they are color coded by their park-walkshed tree canopy ratio. Points that are more green have up to four times greater coverage than their walkshed while points that are more white are closer to only around 50% more coverage than their walkshed.



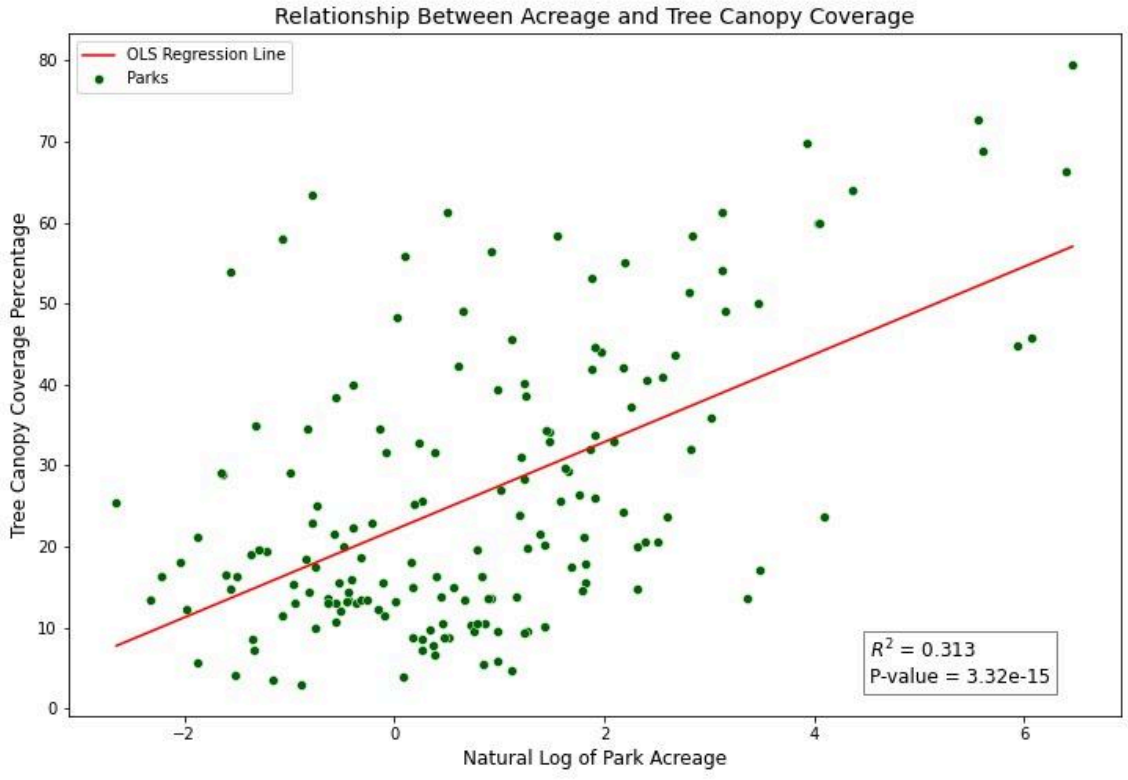
Below is a bar graph displaying the 10 parks in Pittsburgh with the best park-walkshed tree canopy ratio. The top three park-walkshed tree canopy ratios belong to the parks of Southside Riverfront Park, Sheraden Park, and South Shore Riverfront Park.



Park Tree Canopy and Acreage Relationship

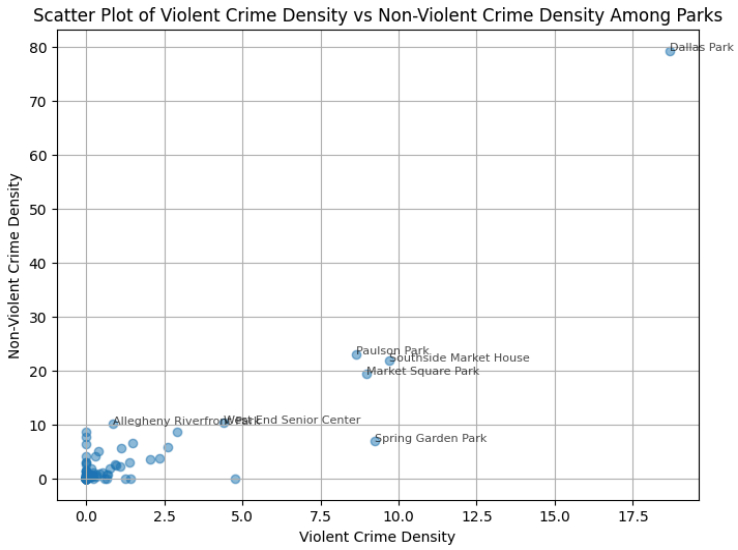
A point of notice when analyzing tree canopy coverage among parks is the general relationship between the park's coverage and their acreage. We have found that there is a positive correlation and an upward trend for the two variables. The greater the park acreage is, the more likely they will have a higher average tree canopy coverage.

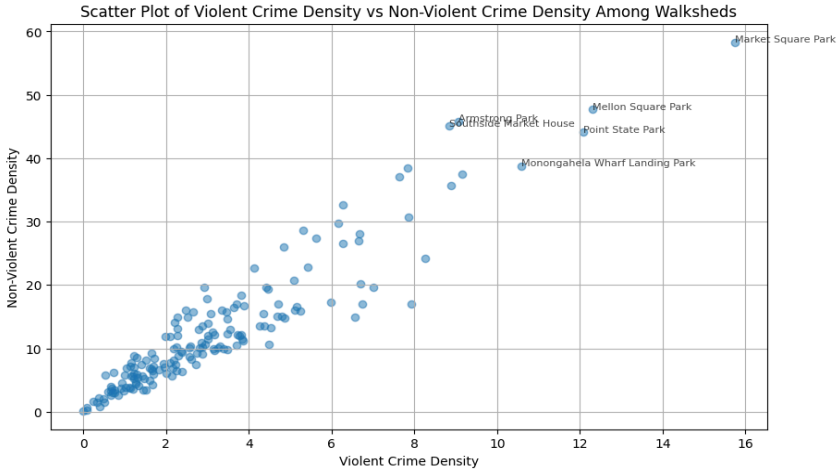
We ran a regression between the two variables to quantify the relationship. After performing a log transformation on the acreage variable to improve interpretability from raw units to percentages, we ran an ordinary least squares (OLS) regression. From the output, we can say that a 1% increase in park acreage is associated with a .054% increase in tree canopy coverage for a park. About 31% of variability in the tree canopy coverage is explained by the size of the park, indicated by an R-squared of 0.313. The full regression output is available in the environmental section of the deliverable folder.



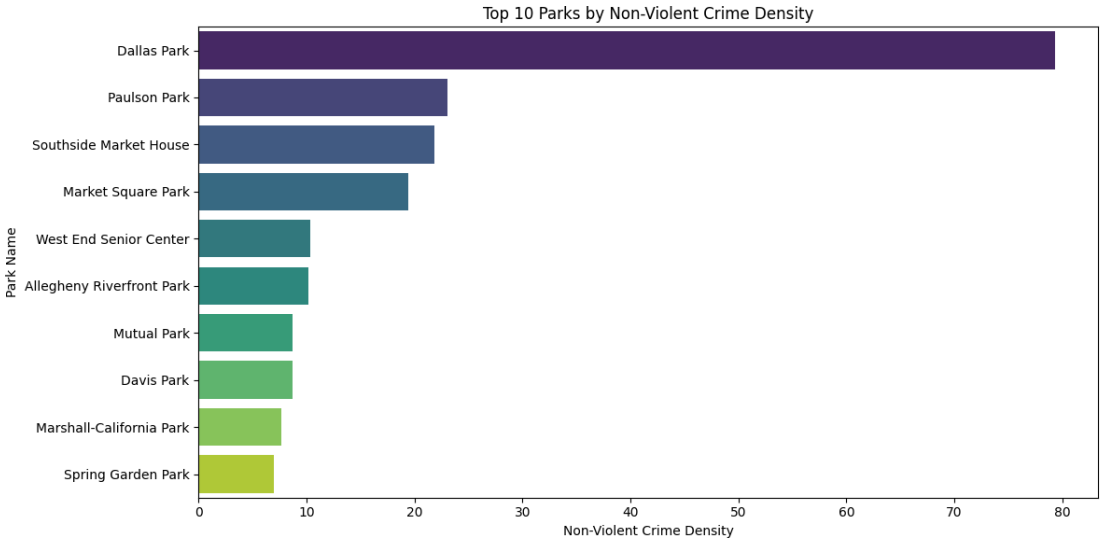
III. Analysis → C. Crime

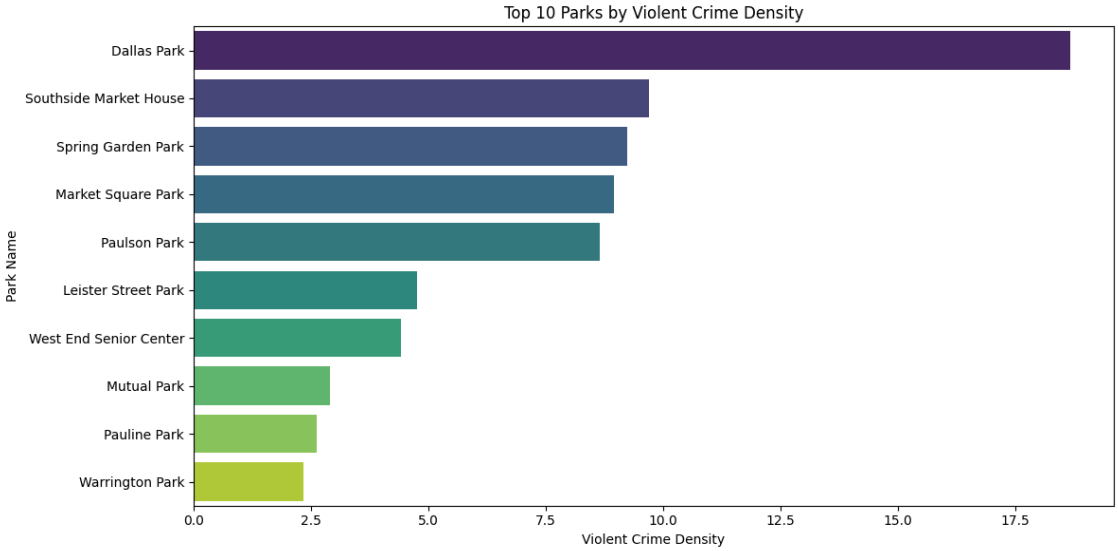
After preparing the crime data for analysis, we began to look at both the walkshed crime metrics and the park crime metrics. Additionally, we analyzed these two subjects through the total crime per acre, violent crimes per acre, and nonviolent crimes per acre category. Below are scatter plots showing the distribution of violent crimes per acre on the horizontal axis and the nonviolent crimes per acre on the vertical axis. We can see that the walkshed distribution has a much more linear relationship between the two types of crime. This could show that the parks have a variable relationship. This representation is likely due to the small amount of crimes that take place in parks compared to those within their respective surrounding walksheds.



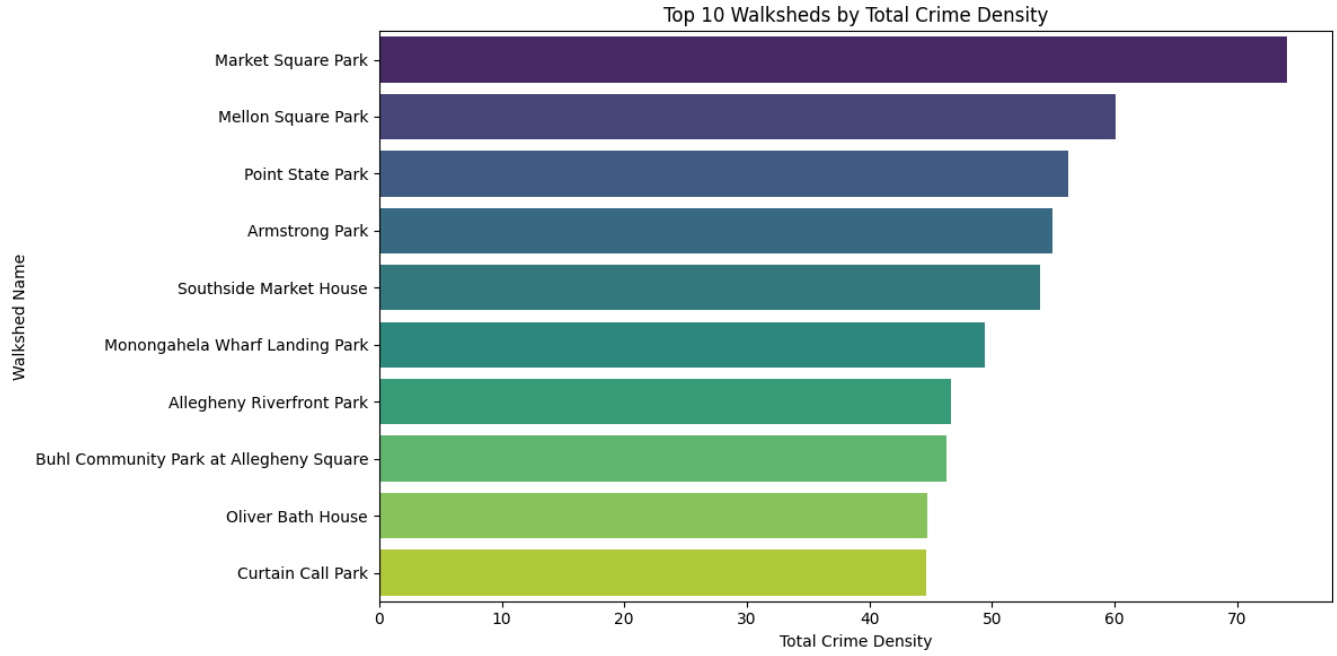


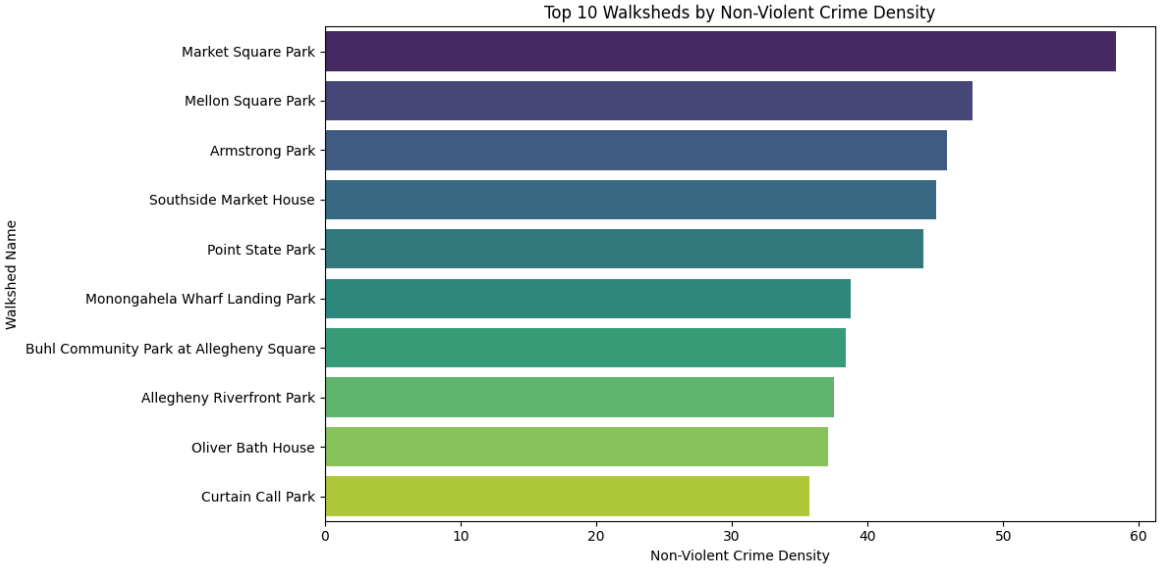
When analyzing only parks, we created three charts to display the parks with the 10 highest rates of crimes per acre, nonviolent crimes per acre and violent crimes per acre. The parks that seem to have the highest crime rates are Dallas Park, Southside Market House and Paulson Park. Warrington Park and Leicester Street Park both have high violent crime rates compared to nonviolent crime rates. Because of the larger quantity of the nonviolent crimes in the total dataset, the total crime density chart and the nonviolent crime density chart are very similar. Only Marshall-California Park is a unique park from those that make up the total crime density chart.

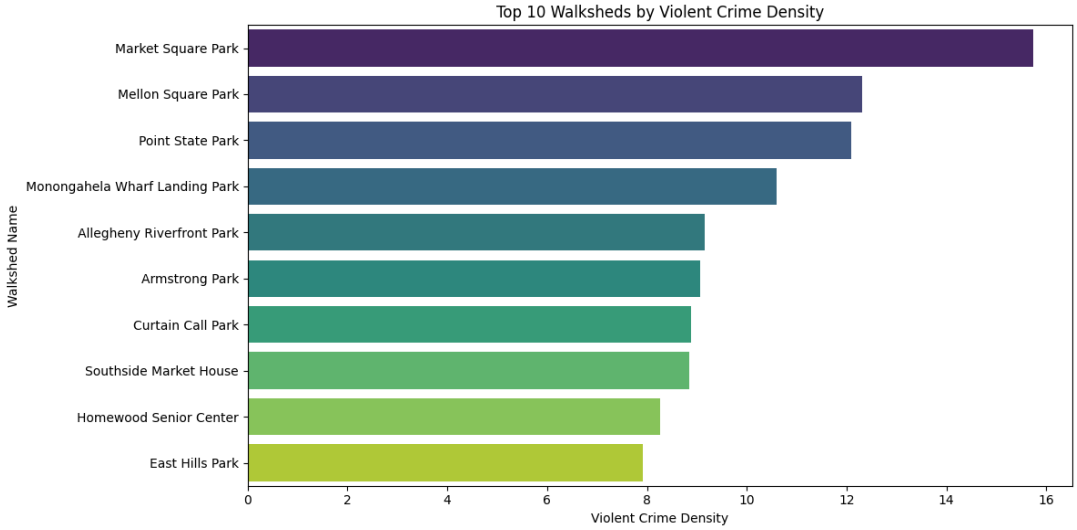




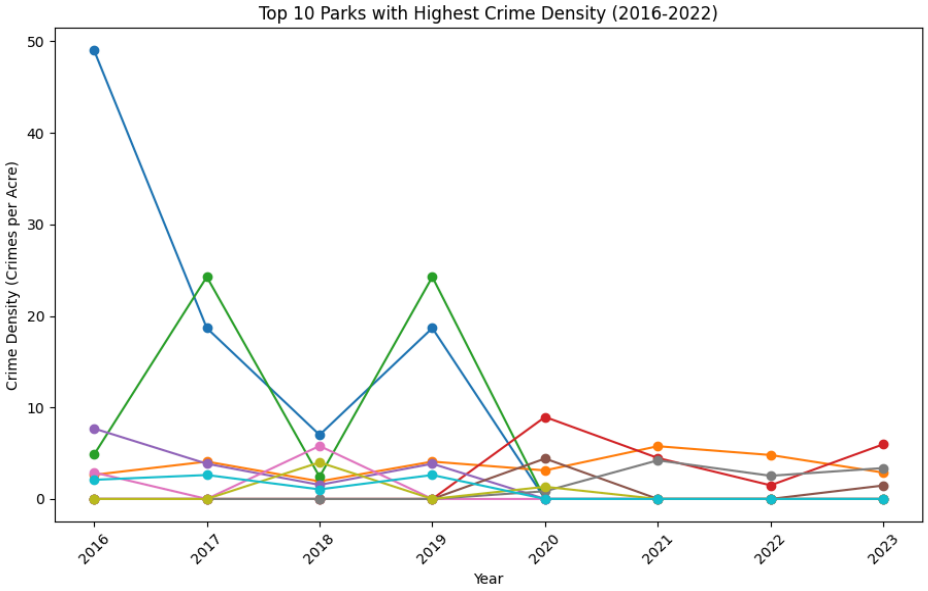
Similarly, when analyzing only walksheds, we created three charts to display the walksheds with the 10 highest rates of crimes per acre, nonviolent crimes per acre and violent crimes per acre. The walksheds that have the highest crime rates are Market Square Park, Mellon Square Park, and Point State Park. Some parks that appeared in the walkshed top 10 and park top 10 for total crime density were Market Square Park, Southside Market House, and Allegheny Riverfront Park. Homewood Senior Center and East Hills Park's walksheds both have high walkshed violent crime rates compared to total crime rates. Because of the larger quantity of the nonviolent crimes in the total dataset, the total crime density chart and the nonviolent crime density chart are the same when analyzing walkshed data, likely due to the higher counts of crimes in the walksheds than in the parks.

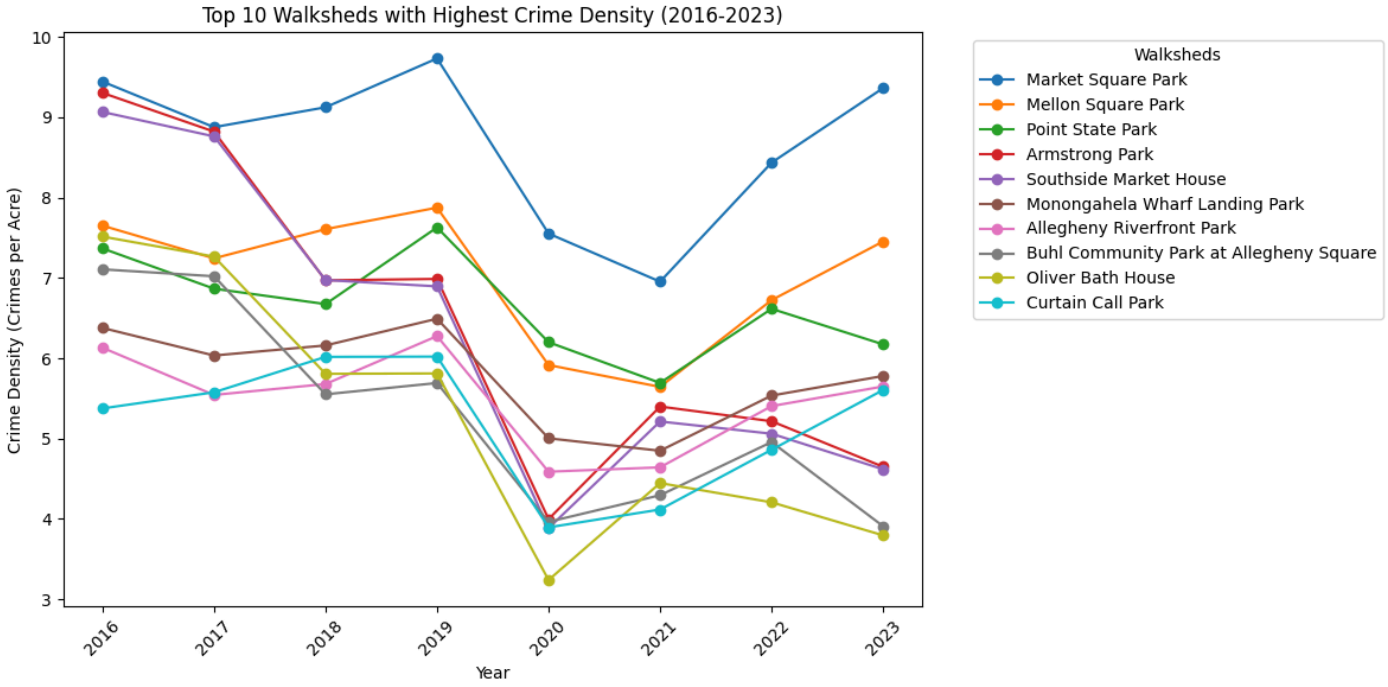




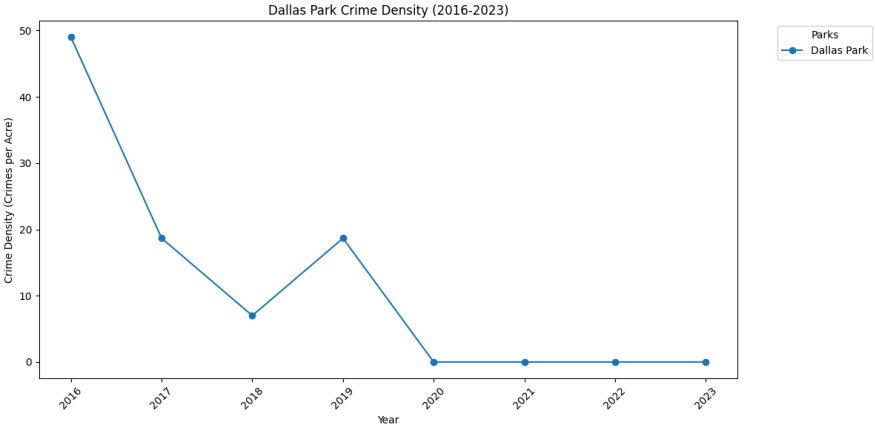


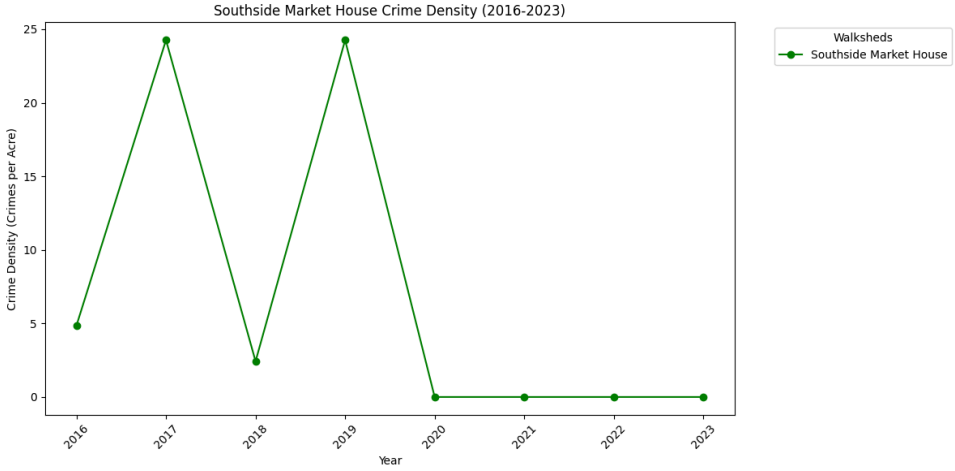
To portray the data over time, we used the 10 walksheds and parks with the highest total crimes per acre for the entire dataset and plotted each of their crimes per acre as a line plot for each year from 2016 to 2023. These plots help define the distributions of crimes over different years for each park and walkshed. An example of this analysis is that Armstrong Park had a crime per acre rate of just over 9 in 2016, but in 2020, it fell to closer to 4. The highest walkshed crime rate appeared in Market Square Park in 2019, close to 10 crimes per acre. For the parks data, we can see that there are two major spikes in the data, Dallas Park in 2016 and Southside Market House in 2017. These two counts could possibly be biasing the overall counts in these two parks.





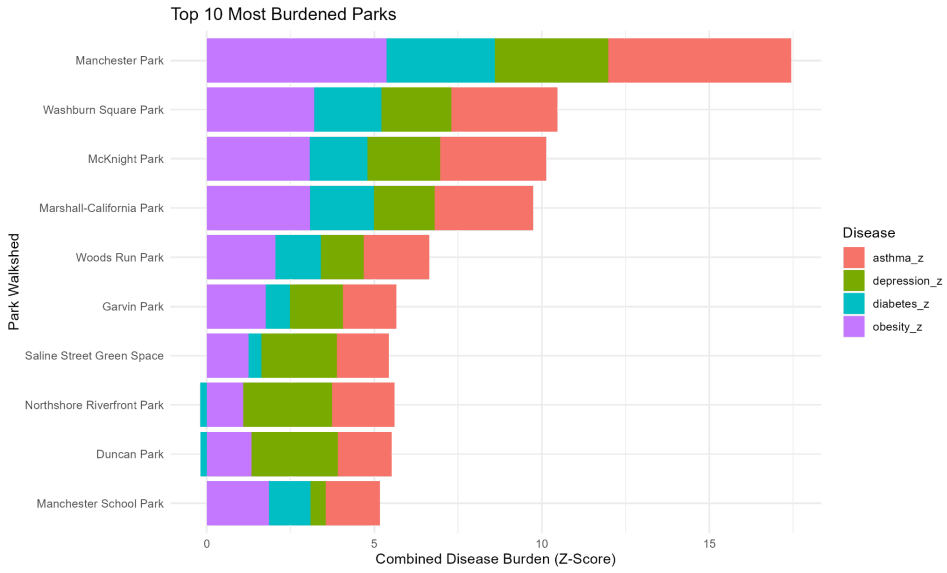
We can see, though, on the individual plot for Dallas Park, that the crime per acre does seem to be following an overall downward trend. This could show that the 80 crimes per acre may have been part of an earlier trend not captured by the data. Additionally, for the Southside Market House's individual plot, the data is much more erratic. With two major spikes, one in 2017 and one in 2019, contributing largely to the total crimes per acre, as after those years, the density levels out to around zero. This could show a possible misrepresentation of the crimes taking place in Southside Market House.





We also explored using a predictive machine learning model to attempt to explain important features contributing to crime in parks. Using a Random Forest model that uses a technique called decision trees to predict a target variable using a set of features, we targeted park total crime density and initially received an accuracy around 60%, with the strongest predictor other than direct covariates being depression rate. However, after removing directly covariate features like violent and nonviolent crime density, the accuracy plummeted to below 1%. Hence, we chose to omit this output, as the conclusions are not statistically significant enough to draw insights.

III. Analysis → D. Public Health



To identify which parks serve communities with the greatest overall health burden, we combined four disease prevalence rates and adjusted them using the statistical descriptor of z-scores. A z-score tells us how far above or below the average a given value is, on a standardized scale. For example, a z-score of 0 means the park is exactly average for that disease. By averaging the z-scores across all four diseases for each park, we created a composite health burden score that reflects population-adjusted risk. These parks have the highest combined burden, which represent communities experiencing the greatest overall health challenges after controlling for population size.

It is important to note that while these public health indicators are not a direct measure of park quality or infrastructure, they provide context about community vulnerability. Chronic health disparities often overlap with other factors included in this analysis such as tree canopy, air pollution exposure, and socioeconomic hardship. In this way, public health burden acts as an indirect but meaningful indicator for understanding where park investments may have the greatest potential impact – not only on recreation, but also on community well-being and equitable environmental access.

IV. Conclusion

The raw and processed datasets, visualizations, R code, and spatial data utilized and produced throughout this project are available in a Google Drive folder shared with the Pittsburgh Parks Conservancy. It is organized in subfolders based on the four data categories, each containing the visualizations, code, and datasets associated with our findings, along with spatial information and a master dataset named Parks_Master.csv. This file contains population- and size-scaled information for relevant variables across all parks and walksheds, and we provided a guide for interpreting the columns in the VI. Appendix section of this document.

The Conservancy can download files from the folder and conduct basic analyses of the master dataset using Microsoft Excel and other tools, according to need. We also included a bibliography of all third-party data we cited in the V. References section of this document such that future projects, along with our detailed descriptions of our workflow, can attempt to replicate our processes to accommodate new data. None of these conclusions are static – Pittsburgh is perpetually changing, and this project provides a momentary snapshot of the communities in the Conservancy's footprint.

A previous report from Interface included an index that attempted to objectively calculate and rank each park's investment need. Through dialogue with the Conservancy, we decided to forgo this metric; instead, we sought to provide the necessary cleaned and scaled data in the master dataset to allow the Conservancy, if they so desire, to construct it themselves.

The weighting of different variables related to race, crime, pollution, and disease in a formula producing a single number to describe investment need is ultimately a policy decision – a value-based, subjective judgment reserved for stakeholders, not analysts. This project empowers those stakeholders to encode their sociological and economic intuition about what fundamentally makes a park or community underinvested into concise, quantifiable, and transferable insights used to articulate investment needs to vital venues for securing funding.

Many of these analyses illuminate systemic disparities across Pittsburgh's communities, often along racial, economic, and geographic lines, that mirror patterns of historic segregation and disinvestment. We hope that the Conservancy can utilize this project to augment their strategic decision-making with data-driven robustness and help forge a more equitable, sustainable, safe, and healthy future for Pittsburgh's public spaces.

V. References

Allegheny County Department of Human Services. (2021). The Allegheny County Community Need Index: Update for 2021 with a focus on the connection between race and community need. https://www.alleghenycountyanalytics.us

Atmospheric Composition Analysis Group. (2022). *North American Regional Estimates (V4.NA.03).* Washington University of St. Louis. Retrieved January 28, 2025, from https://sites.wustl.edu/acag/datasets/surface-pm2-5/

Centers for Disease Control and Prevention. (2024). *PLACES: Local Data for Better Health, Census Tract Data 2024 Release*. U.S. Department of Health and Human Services. https://www.cdc.gov/places

City of Pittsburgh. (2023). *Police Incident Blotter (Archive)*. Western Pennsylvania Regional Data Center. https://data.wprdc.org/dataset/uniform-crime-reporting-data

Gramlich, John. (2024, April 24). *What the data says about crime in the US*. Pew Research Center. pewresearch.org/short-reads/2024/04/24/what-the-data-says-about-crime-in-the-us/

Interface Studio LLC. (2019). *Restoring Pittsburgh Parks Investment Strategy – Walksheds*. Retrieved February 3, 2025.

Pittsburgh Water and Sewer Authority. (2024). *Combined Sewersheds.* Western Pennsylvania Regional Data Center. Retrieved January 28, 2025, from https://data.wprdc.org/dataset/combined-sewershed

United States Department of Agriculture Forest Service. (2021). *The National Land Cover Database (NLCD) Tree Canopy Cover (TCC)*. U.S. Department of Agriculture. Retrieved January 28, 2025, from https://data.fs.usda.gov/geodata/rastergateway/treecanopycover/#table1

United States Environmental Protection Agency. (2025). *National Ambient Air Quality Standards (NAAQS) for PM*. Accessed April 4, 2025 from https://www.epa.gov/pm-pollution/national-ambient-air-quality-standards-naaqs-pm

U.S. Census Bureau. (2023). *American Community Survey 5-year Estimate of Age and Sex by Tract in Allegheny County, PA*. U.S. Department of Commerce. Retrieved February 3, 2025, from https://data.census.gov/

U.S. Census Bureau. (2023). *American Community Survey 5-year Estimate of Poverty by Tract in Allegheny County, PA*. U.S. Department of Commerce. Retrieved February 3, 2025, from https://data.census.gov/

U.S. Census Bureau. (2020). *Race and Ethnicity by Tract in Allegheny County, PA*. U.S. Department of Commerce. Retrieved February 3, 2025, from https://data.census.gov/

U.S. Census Bureau. (2020). *TIGER/Line Shapefile, 2020, 2010 Census, Census Tract, Allegheny County, PA* DatasetData setDataset. https://www.census.gov/geographies/mapping-files/time-series/geo/tiger-line-file.html

U.S. Census Bureau. (2020). *Vacancy by Block Group in Allegheny County, PA*. U.S. Department of Commerce. Retrieved February 3, 2025, from https://data.census.gov/

VI. Appendix

In **Deliverable → Datasets**

Parks_Master.csv: cleaned and spatially attached data including counts and proportions of demographic, environmental, health, and crime data in Allegheny County, Pennsylvania

- objectid: a unique numerical identifier for each park and their walkshed
- updatepknm: name of the park to whom the park or walkshed data corresponds
- Tree_Canopy_Park: averaged percentage of tree canopy coverage for a given park
- Tree_Canopy_Walkshed: averaged percentage of tree canopy coverage for a given walkshed
- Pollution_Park: averaged pm2.5 concentration level measured in $\mu g/m^3$ for a given park
- Pollution_Walkshed: averaged pm2.5 concentration level measured in $\mu g/m^3$ for a given walkshed
- Sewershed_Park: aggregated priority score based off PWSA ratings for a given park
- Sewershed_Walkshed: aggregated priority score based off PWSA ratings for a given walkshed
- percent_rank1: percentage of the park that falls within secondary priority sewershed boundary
- percent_rank2: percentage of the park that falls within high priority sewershed boundary
- Total_population_18plus: total population of 18+ individuals of Census 2020
- Total_asthma: est. population of individuals (age 18+) affected by asthma for a given walkshed
- Total_obesity: est. population of individuals (age 18+) affected by obesity for a given walkshed
- Total_depression: est. population of individuals (age 18+) affected by depression for a given walkshed
- Total_diabetes: est. population of individuals (age 18+) affected by diabetes for a given walkshed
- Asthma_rate: percentage of population (18+) affected by asthma within a given walkshed
- Obesity_rate: percentage of population (18+) affected by obesity within a given walkshed
- Depression_rate: percentage of population (18+) affected by depression within a given walkshed
- Diabetes_rate: percentage of population (18+) affected by diabetes within a given walkshed
- Park Acreage: total acreage of the park
- TotalCrimesInParks: number of crimes within each park polygon
- ViolentCrimesInParks: number of violent crimes within each park polygon
- NonViolentCrimeCount_Parks: number of nonviolent crimes within each park polygon
- TotalCrimeDensity_Parks: number of crimes per acre within each park polygon
- ViolentCrimeDensity_Parks: number of violent crimes per acre within each park polygon
- NonViolentCrimeDensity_Parks: number of non violent crimes per acre within each park polygon
- TotalCrimesInWalkshedDifferences: number of crimes within each walkshed
- Walkshed Acreage: total acreage of the park's walkshed
- ViolentCrimeCountsInWalksheds: number of violent crimes within each walkshed
- NonViolentCrimeCount_Walksheds: number of nonviolent crimes within each walkshed
- TotalCrimeDensity_Walksheds: number of crimes per acre within each walkshed
- ViolentCrimeDensity_Walksheds: number of violent crimes per acre within each walkshed
- NonViolentCrimeDensity_Walksheds: number of non violent crimes per acre within each walkshed
- sum_Total: count of walkshed residents
- sum_Hispanic.or.Latino: count of Hispanic/Latino residents
- sum_Not.Hispanic.or.Latino: count of non-Hispanic/Latino residents
- sum_Population.of.one.race: count of single race category residents
- sum_White.alone: count of White residents
- sum_Black.or.African.American.alone: count of Black residents
- sum_American.Indian.and.Alaska.Native.alone: count of Native American residents
- sum_Asian.alone: count of Asian residents
- sum_Native.Hawaiian.and.Other.Pacific.Islander.alone: count of Pacific Islander residents
- sum_Some.Other.Race.alone: count of residents of other races
- sum_Population.of.two.or.more.races: count of residents of two or more races
- sum_Total.parcels: count of property parcels in the walkshed
- sum_Occupied: count of occupied parcels
- sum_Vacant: count of vacant parcels

- sum_Poverty.status.determined.total: count of residents whose poverty status is known
- sum_Below.poverty.level.: count of residents below the poverty level
- sum_Under.5.years: count of residents under 5 years of age
- sum_X5.to.9.years: count of residents between 5 and 9 years of age
- sum_X10.to.14.years: count of residents between 10 and 14 years of age
- sum_X15.to.19.years: count of residents between 15 and 19 years of age
- sum_X20.to.24.years: count of residents between 20 and 24 years of age
- sum_X25.to.29.years: count of residents between 25 and 29 years of age
- sum_X30.to.34.years: count of residents between 30 and 34 years of age
- sum_X35.to.39.years: count of residents between 35 and 39 years of age
- sum_X40.to.44.years: count of residents between 40 and 44 years of age
- sum_X45.to.49.years: count of residents between 45 and 49 years of age
- sum_X50.to.54.years: count of residents between 50 and 54 years of age
- sum_X55.to.59.years: count of residents between 55 and 59 years of age
- sum_X60.to.64.years: count of residents between 60 and 64 years of age
- sum_X65.to.69.years: count of residents between 65 and 69 years of age
- sum_X70.to.74.years: count of residents between 70 and 74 years of age
- sum_X75.to.79.years: count of residents between 75 and 79 years of age
- sum_X80.to.84.years: count of residents between 80 and 84 years of age
- sum_X85.years.and.over: count of residents over 85 years of age
- sum_Total_pct: a check that %s are calculated correctly, should read 100
- sum_Hispanic.or.Latino_pct: % of Hispanic/Latino residents
- sum_Not.Hispanic.or.Latino_pct: % of non-Hispanic/Latino residents
- sum_Population.of.one.race_pct: % of residents of one race
- sum_White.alone_pct: % of White residents
- sum_Black.or.African.American.alone_pct: % of Black residents
- sum_American.Indian.and.Alaska.Native.alone_pct: % of Native American residents
- sum_Asian.alone_pct: % of Asian residents
- sum_Native.Hawaiian.and.Other.Pacific.Islander.alone_pct: % of Pacific Islander residents
- sum_Some.Other.Race.alone_pct: % of residents of other rances
- sum_Population.of.two.or.more.races_pct: % of residents of two or more races
- sum_Poverty.status.determined.total_pct: % of residents whose poverty status is known
- sum_Below.poverty.level._pct: % of residents under the poverty level
- sum_Under.5.years_pct: % of residents under 5 years of age
- sum_X5.to.9.years_pct: % of residents between 5 and 9 years of age
- sum_X10.to.14.years_pct: % of residents between 10 and 14 years of age
- sum_X15.to.19.years_pct: % of residents between 15 and 19 years of age
- sum_X20.to.24.years_pct: % of residents between 20 and 24 years of age
- sum_X25.to.29.years_pct: % of residents between 25 and 29 years of age
- sum_X30.to.34.years_pct: % of residents between 30 and 34 years of age
- sum_X35.to.39.years_pct: % of residents between 35 and 39 years of age
- sum_X40.to.44.years_pct: % of residents between 40 and 44 years of age
- sum_X45.to.49.years_pct: % of residents between 45 and 49 years of age
- sum_X50.to.54.years_pct: % of residents between 50 and 54 years of age
- sum_X55.to.59.years_pct: % of residents between 55 and 59 years of age
- sum_X60.to.64.years_pct: % of residents between 60 and 64 years of age
- sum_X65.to.69.years_pct: % of residents between 65 and 69 years of age
- sum_X70.to.74.years_pct: % of residents between 70 and 74 years of age
- sum_X75.to.79.years_pct: % of residents between 75 and 79 years of age
- sum_X80.to.84.years_pct: % of residents between 80 and 84 years of age
- sum_X85.years.and.over_pct: % of residents over 85 years of age
- sum_Total.parcels_pct: a check that %s are calculated correctly, should read 100
- sum_Vacant_pct: % of parcels that are vacant
- sum_Occupied_pct: % of parcels that are occupied