Briggs, Thomas

Phoenix Crime Analysis: Clustering and Urban Proximity

Introduction

This project aims to uncover spatial patterns in crime within the city of Phoenix, Arizona. Using geographic information systems and Python-based geoprocessing, we analyze where recent crimes are concentrated and how those locations relate to key urban infrastructure such as light rail stops and highways. Our primary tool for spatial pattern detection is the DBSCAN clustering algorithm, which helps identify high-density crime zones. Through proximity analysis, we explore whether crimes are more likely to occur near transportation corridors or further away, providing insight into urban accessibility and public safety planning.

Data

We use three main datasets:

1. Crime Data: A CSV containing over 1000 recent crime incidents in Phoenix. It includes attributes such as type of crime, date/time, address, and ZIP code. Geocoding was performed using Nominatim (OpenStreetMap).
2. Light Rail Stops: A GeoJSON file representing the locations of Valley Metro light rail stops across Phoenix.
3. Highways: A shapefile from the U.S. Census TIGER dataset, clipped to Phoenix and projected in EPSG:3857.

All data was cleaned, projected, and processed using GeoPandas and Shapely within a Python Jupyter Notebook.

Methodology

The project followed a structured workflow using Python libraries including Pandas, GeoPandas, Shapely, Scikit-learn, and Contextily. The steps were as follows:

1. Load and Clean Crime Data: The CSV file was loaded using Pandas. Missing values were handled, and records with invalid dates or coordinates were dropped.
2. Convert Tabular to Spatial Data: Using GeoPandas, the cleaned data was converted into a GeoDataFrame by creating geometry points from latitude and longitude.
3. Reproject for Analysis: To ensure accurate spatial computations (e.g., distance buffers), the data was projected from EPSG:4326 to EPSG:3857 (Web Mercator).
4. Visualize Crime Locations: Crime points were plotted on a basemap with overlaid thematic elements (e.g., MultiPoint, MultiPolygon objects) to show distribution patterns.
5. Perform Spatial Analysis: Clustering was done by DBSCAN, then applied to the spatial coordinates to identify natural groupings of crime incidents. Proximity Analysis done by distance to nearest light rail stop was computed using Shapely's spatial operations, and distance to nearest highway segment was calculated using buffer clipping and nearest neighbor queries.
6. Multi-geometry Operations: For demonstration and analysis, geometric aggregations like MultiPoint, MultiPolygon, and buffer zones were constructed to represent crime clusters and urban forms.
7. Cluster Visualization: Clusters were visualized by unique colors and evaluated by average proximity to landmarks.
8. Export and Save Results: Final GeoDataFrames and visualizations were exported for use in a written report and for possible conversion to shapefiles or additional GIS applications.

This structured methodology ensures spatial rigor and makes the analysis reproducible and extensible.

Results

This section outlines the spatial and analytical outcomes from each major phase of the project. It includes descriptive interpretations of how geospatial operations relate to real-world crime analysis in Phoenix, Arizona.

1. Cleaning and Preparing the Crime Dataset

The project began by processing a large CSV file of Phoenix crime reports. The dataset included incident number, time of occurrence, address, and ZIP code. Using Pandas, non-informative columns were dropped, and the "Occurred From" field was converted to datetime format. The dataset was sorted by time and the most recent 1000 incidents were selected. Address fields were cleaned and standardized for geocoding using pattern replacement and string manipulation. This made them suitable for spatial geolocation.

This step ensured that only the most relevant and temporally accurate records were included, preparing the data for meaningful spatial analysis. Cleaning addresses enabled more reliable mapping to real-world locations.

2. Geocoding Crime Addresses

Cleaned addresses were geocoded using the Nominatim geolocator from OpenStreetMap. When geocoding failed (common for vague intersections or placeholder addresses), a fallback strategy used ZIP code centroids to approximate location. The result was a dataset of 1000 crimes each with latitude and longitude coordinates.

Geocoding transformed tabular data into mappable spatial data. This enabled the subsequent use of GeoPandas and shapely geometry for spatial operations. ZIP centroid fallback minimized data loss while maintaining geographic context

3. Creating a GeoDataFrame and Projection

The coordinates were converted to Point geometries and wrapped in a GeoDataFrame. The coordinate reference system was transformed to EPSG:3857 (Web Mercator) to ensure compatibility with basemaps and proper distance calculations.

This step converted crime points into true spatial features, making them measurable and visualizable in a real-world coordinate system.

4. Visualizing Crime Distribution

Initial mapping included:

A scatter plot of all crime locations, sample use of MultiPoint and buffered MultiPolygon features for visualization, and overlay of these features on a CartoDB basemap.

The map showed dense clustering in central and southern Phoenix, validating known urban crime hotspots. MultiPoint and MultiPolygon constructs showcased more advanced geometry manipulation and how clusters might envelop urban blocks or districts.

5. DBSCAN Clustering

Density-Based Spatial Clustering of Applications with Noise (DBSCAN) was applied to identify crime hotspots. The model detected several clusters and labeled outlier incidents (noise) that didn't fit into any dense grouping.

DBSCAN highlighted spatially concentrated areas of criminal activity. For example, the largest cluster included over 800 incidents, mostly within a few kilometers of downtown. This method can help city officials prioritize law enforcement resources.

6. Light Rail Proximity Analysis

Light rail stop data was imported and reprojected. The shortest distance from each crime point to a light rail stop was computed. Clusters were compared by their average proximity to rail.

Proximity metrics help test hypotheses about transit-accessible crime. For instance, crimes in cluster 0 occurred significantly closer to rail stops (~6.2 km on average) than those in cluster 2 (~20.7 km). This may reflect correlations between transit hubs and urban density or socioeconomic factors.

7. Gradient Map by Rail Proximity

Crime points were colored based on distance to the nearest light rail stop. A continuous gradient (viridis colormap) was applied.

The gradient map visually emphasized that many crimes occurred closer to rail infrastructure, especially downtown and along major transit corridors.

8. Highway Proximity Analysis

A shapefile of Arizona’s primary and secondary roads was filtered to cover just the Phoenix area. A spatial join computed the distance from each crime to the nearest major highway.

The highway distance gradient showed that some clusters (e.g., cluster 1) were located far from major roads, while others (e.g., cluster 0) were tightly adjacent. This can inform urban planners on how transportation infrastructure overlaps with crime activity.

9. Cluster Center Mapping

Each DBSCAN cluster centroid was computed using the geometric mean of its points. These centers were visualized alongside light rail stops.

Cluster centroids represent approximate 'epicenters' of urban crime patterns. Comparing these to infrastructure locations reveals whether crime is radiating outward from transportation, residential, or commercial centers.

These cumulative results provide both spatially explicit and interpretive insights into the structure and drivers of crime in Phoenix. The next section will summarize findings, discuss limitations, and propose future steps.

Conclusion

This project set out to investigate spatial patterns of crime in Phoenix, Arizona, using Python-based geospatial analysis. Through the application of open-source data, geocoding, spatial clustering (DBSCAN), and proximity analysis to major urban infrastructure such as light rail stops and highways, the project achieved several meaningful insights and established a workflow that could be replicated for other urban areas.

One of the primary conclusions drawn from this work is that crime in Phoenix is not randomly distributed but instead shows measurable clustering behavior. The DBSCAN model revealed several high-density clusters of incidents, particularly in central and western areas of the city. These clusters likely correspond to areas with greater population density, commercial activity, and accessibility, making them important targets for focused community policing or resource deployment.

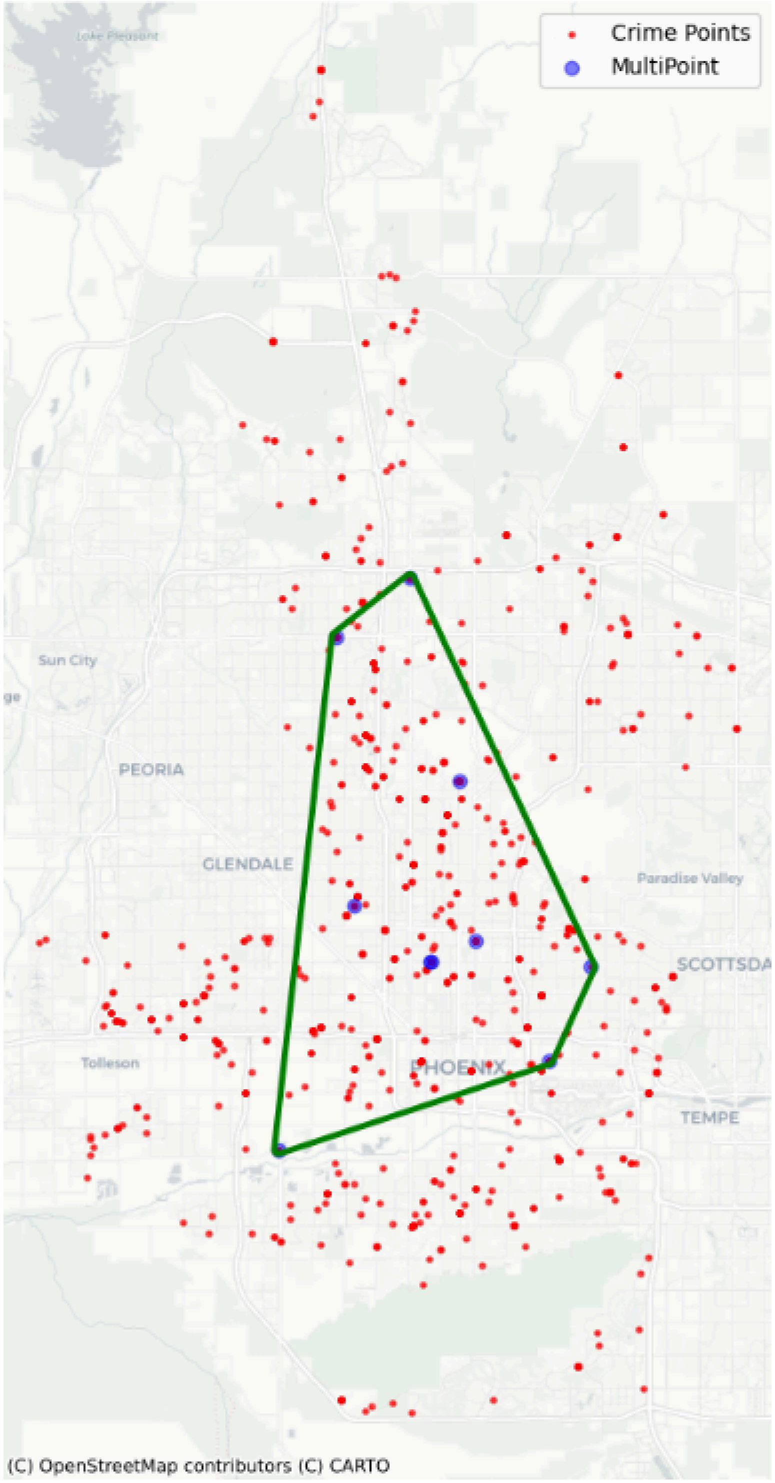
The proximity analyses demonstrated that while some clusters are located within 1–2 km of light rail infrastructure, others—especially those in outlying areas—are significantly more isolated. This spatial disparity suggests a need to consider transit access in urban safety planning. Similarly, measuring crime distance from major highways showed that several clusters are closely situated to high-traffic corridors, which may imply a relationship between vehicular access and opportunistic crimes, particularly theft or drug offenses.

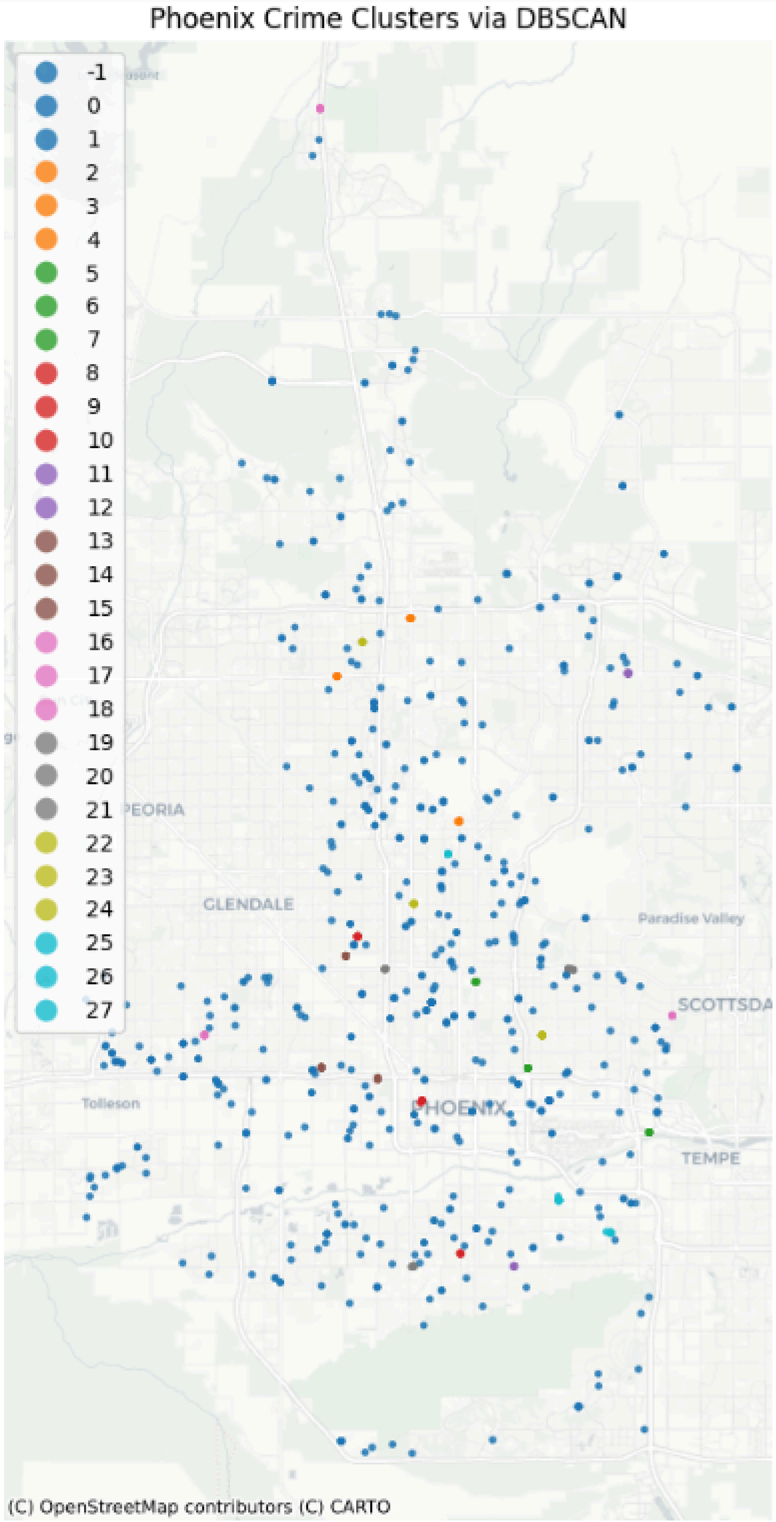
Despite these successes, the project is not without limitations. Geocoding accuracy, especially for intersection-style addresses, posed challenges, and the reliance on open data meant that some entries lacked full location details. Moreover, the DBSCAN clustering approach is sensitive to parameter selection; future work could compare it to hierarchical or density-based methods like HDBSCAN or OPTICS for improved adaptability.

Next steps could involve incorporating more layers of socioeconomic or demographic data to explore correlations between crime and social vulnerability. Adding time series analysis—such as seasonality or day-of-week trends—could yield even deeper insights. Finally, deploying the methodology in a dashboard format would make it more accessible to urban planners, law enforcement, and the public.

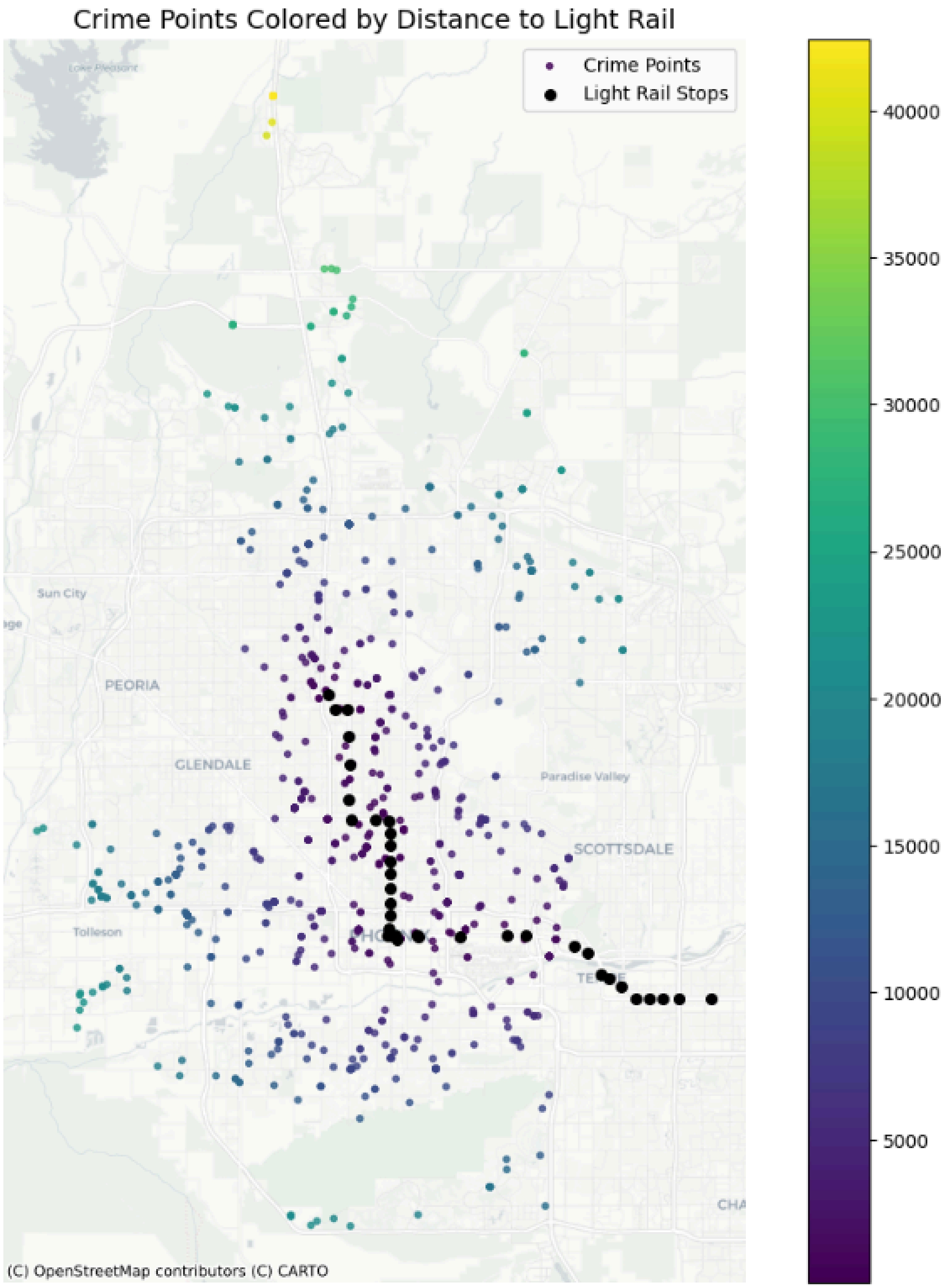
In summary, this project successfully met its goals: spatializing and analyzing crime data to uncover meaningful geographic trends. The tools and techniques used provide a replicable framework for civic data analysis, and the results have implications for real-world crime prevention, infrastructure planning, and urban accessibility.

Phoenix Crime Map with MultiPoint and MultiPolygon





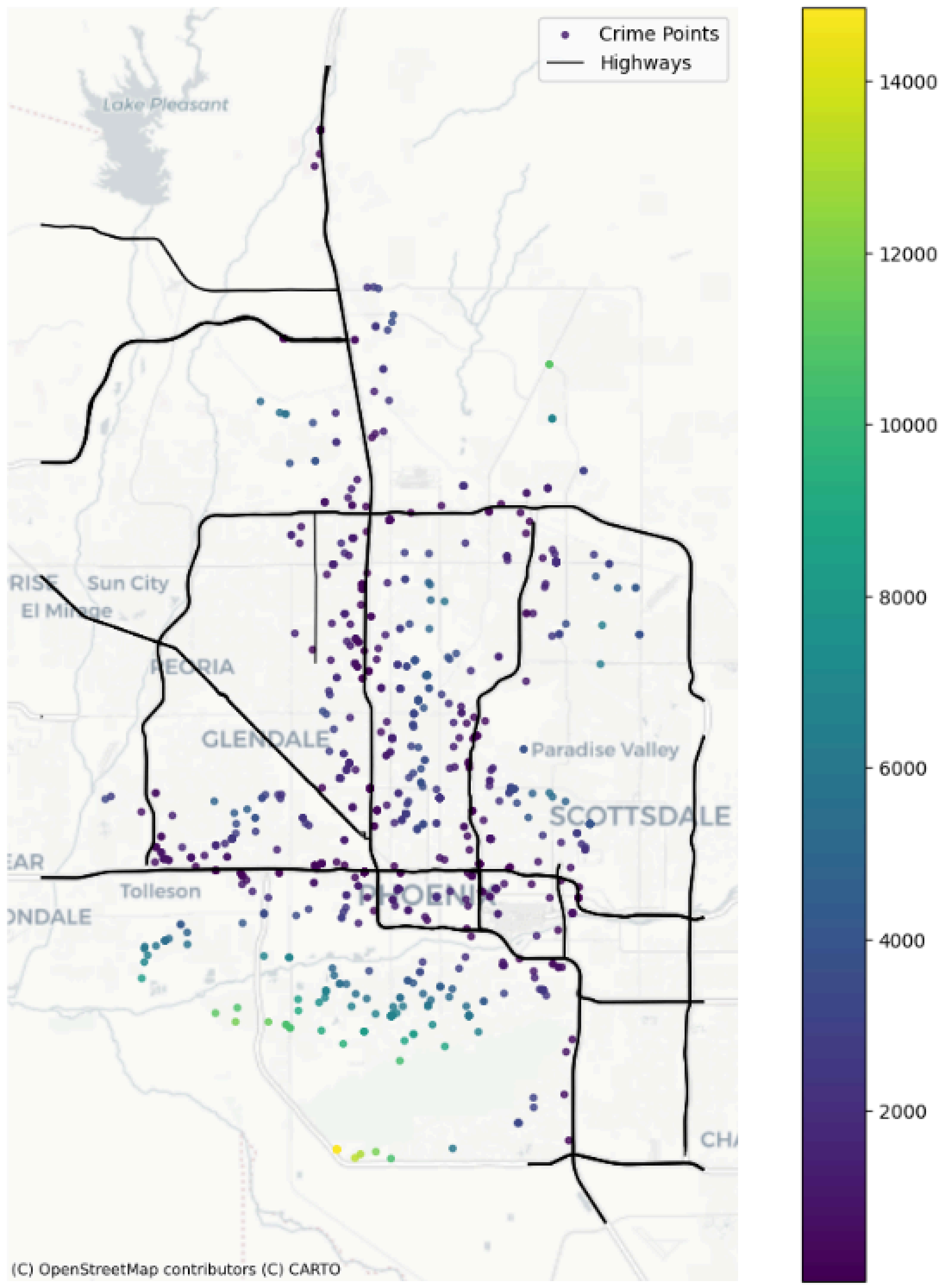
(C) OpenStreetMap contributors (C) CARTO



Crime Points Colored by Distance to Light Rail

(C) OpenStreetMap contributors (C) CARTO

Phoenix Crime Points Colored by Distance to Highways


Crime Points
Highways
Lake Pleasant
Sun City
El Mirage
PEORIA
GLENDALE
Paradise Valley
SCOTTSDALE
Tolleson
PHOENIX
CH
(C) OpenStreetMap contributors (C) CARTO
14000
12000
10000
8000
6000
4000
2000