Street-Level Infrastructure and Crash Risk in Portland: Investigating the Impact on Bicycle and Pedestrian Safety

A Data Summary

Noelle Matthews & Thomas Sato

1. Overview:

How do street-level infrastructure features such as neighborhood greenways, traffic calming devices, and lighting relate to the frequency and severity of bicycle- and pedestrian-involved crashes in Portland?

Our capstone project investigates how street-level infrastructure features relate to the frequency and severity of bicycle and pedestrian-involved crashes in the Portland Metro Area. We focus on data spanning 2007–2022, using publicly available datasets on crashes, infrastructure features (like lighting, calming devices, and greenways), and traffic volume. A key challenge has been harmonizing diverse spatial data sources into a unified relational schema that enables intersection-level analysis. To support this, we created a normalized schema centered on an `intersections` table. All other datasets are joined to this core table, either directly via street name matching or through approximate geospatial alignment using latitude and longitude.

2. Data Ingestion:

The data for this project was sourced from multiple city, state, and regional agencies including the Portland Bureau of Transportation (PBOT), the Oregon Department of Transportation (ODOT), and the Oregon GEOHub. All datasets required targeted preprocessing due to inconsistencies in formats, missing coordinates, and nonstandard key fields. To support a fully normalized schema and enable intersection-level analysis, we built a central `intersections` table that all other tables join to via `intersection_id`. To aid our understanding of available data and appropriate spatial joins, we met with analysts from the City of Portland Vision Zero team and a senior transportation planner from Parametrix with expertise in active transportation and safety. These conversations directly informed our decisions on schema design and data cleaning strategies, and also helped us identify which spatial features and relationships would be most meaningful to query during our analysis.

Oregon Car Crashes:
Crash data from ODOT serves as the foundation of our analysis. This dataset includes over 350,000 crash records from 2007-2022 with fields for date, severity, and crash participants, along with X/Y coordinates and street names. We filtered the data to include only bicycle or pedestrian crashes, dropped records with missing or invalid coordinates, and standardized street names to support downstream joins. This dataset anchors our `crashes` table normalized around a unique crash ID and linked to our `intersections` table `intersection_id` is assigned via matching on `st_full_nm` and `isect_st_full_nm`.

Bike Network:
This dataset catalogs bicycle infrastructure segments, including greenways, lanes, and paths. This dataset additionally includes the year that the infrastructure segment was built. It lacked coordinate fields but included unique segment IDs and street names. Since many fields were in long form or partially incomplete, we standardized facility types and years. Our join logic is based on proximity to intersections along the streets that are described in the dataset. The cleaned output populates the `bicycle_network` table.

Recommended Bike Routes:
Similar to the bike network dataset but with an emphasis on connectivity, this data included from/to streets and connection type. Again, no coordinate columns were present, so we matched explicit intersections by name to our `Intersections` table. This data populates `recommended_bicycle_routes`.

Street Lights:
The street lighting data contained point-level records with X/Y coordinates and descriptive attributes such as owner and power provider. We validated coordinates and transformed them to WGS 84. These records were then spatially joined to the nearest intersection based on proximity thresholds, enabling us to analyze lighting presence at crash locations. This data populates `street_lights`.

Traffic Calming Features:
This file listed point features like speed humps, chicanes, and raised crosswalks. Coordinates were clean but `LocationID` formats varied (e.g., LEG vs. SEG). We filtered out unknown or unplaced entries, then geospatially joined to the nearest intersection. This data is stored in `traffic_calming_features`.

Traffic Volume Counts:
This dataset consisted of multiple traffic counts per location by direction, date, and time of day. It included clean coordinates and detailed volume figures, such as AM, PM, and average daily traffic. We aggregated volumes at the intersection level and matched each record to our `intersection_ID` based on location. This variable will help us account for exposure differences in crash modeling. This data populates `traffic_volume_counts`.

Intersections:
This dataset is derived using the *osmnx* package in Python, which allows users to easily download, model, and analyze OpenStreetMap spatial data, and we used it to create and validate our master intersection list. Since the streets populate two columns, (one for each that makes up an intersection), the data is duplicated to ensure there is no inherent order in which street populates the first versus the second column. This table forms the central `intersections` entity that all other datasets are now normalized around.

Speed Limits:
Though conceptually useful, the speed limit file lacked coordinates and had inconsistent street identifiers. Due to limited joinability and overlap with crash severity proxies, we opted not to integrate this data into the core schema at this stage.

3. Data Organization:

To support intersection-level analysis, we created a normalized relational schema centered on a comprehensive `intersections` table. All other datasets (including crashes, recommended bike routes, street lights, traffic calming features, and traffic volume) join to this table via a shared `intersection_id`. This structure enables consistent spatial alignment and supports flexible querying across time and infrastructure types.

To ensure compatibility across datasets, we standardized all coordinate systems to WGS 84 and used the *osmnx* Python package to generate a cleaned and deduplicated list of Portland intersections. Records from each source were then joined using a combination of street name matching and spatial proximity.

Each resulting table includes only the attributes needed for our analysis and maintains atomic fields and consistent naming conventions. For example, `crashes` is keyed on `crash_id`, and records are linked to their nearest intersection. Other features such as `recommended_bicycle_routes`, `traffic_calming_features`, and `traffic_volume_counts` follow the same pattern.

The resulting relational structure allows each record (whether a crash, street light, or bike feature) to be attributed to a single intersection, enabling aggregation of infrastructure variables by location and facilitating outcome modeling at the intersection level. For example, crash severity can now be analyzed in the context of whether the location features calming devices, high traffic volume, or specific types of bike infrastructure. Figure 1 below shows the resulting structure, data types, and relationships across the created tables.

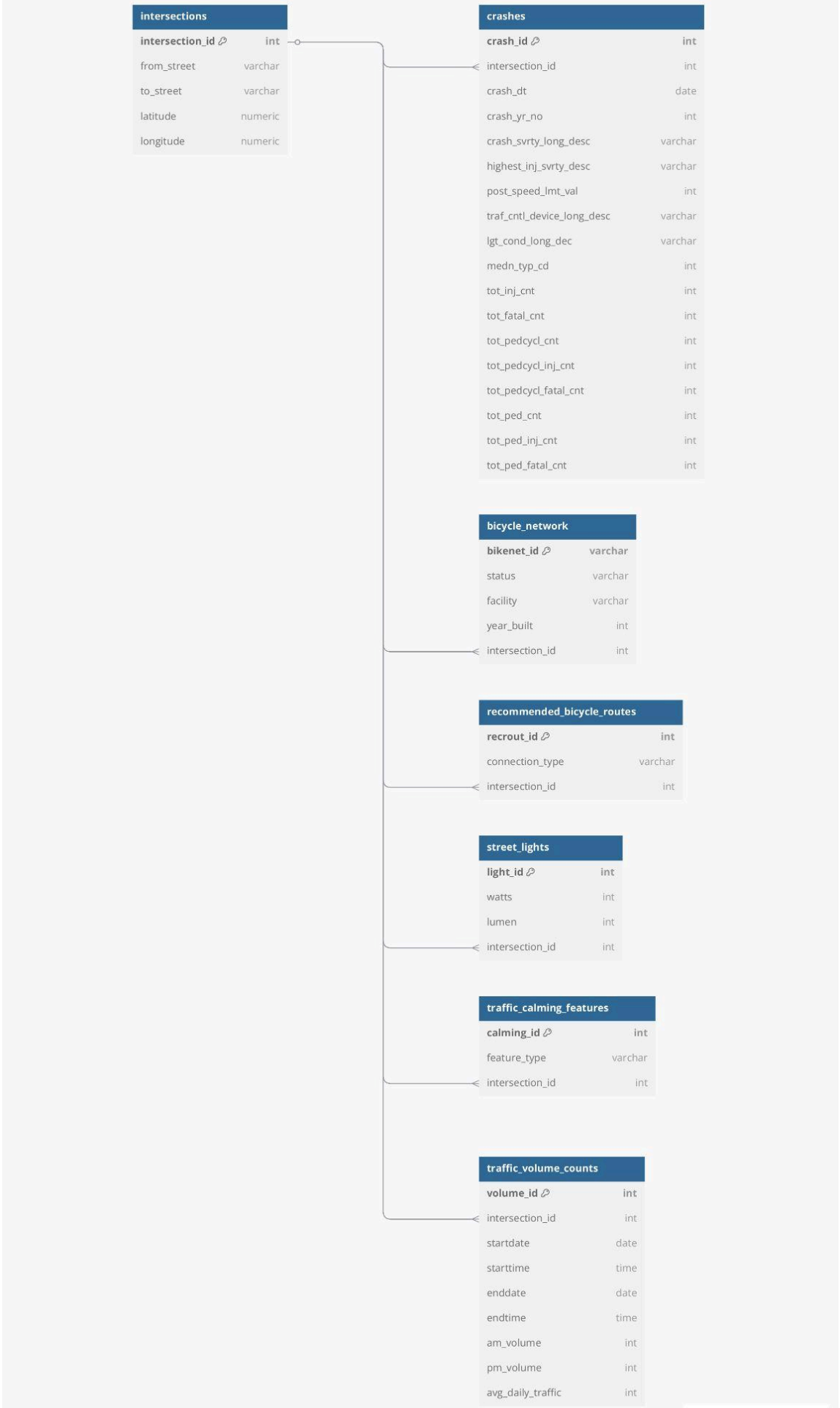

Figure 1: An ER diagram of the created tables and their relationships.

4. Data Compilation:

To demonstrate the joinable structure of our schema and validate our relational approach, we created an exploratory visualization using data from both the `crashes` and `recommended_bicycle_routes` tables. After joining these tables via `intersection_id`, we examined the ratio of crashes involving at least one pedestrian or cyclist across five different connection types: Buffered Bike Lane, Bike Lane, Neighborhood Greenway, Difficult Connection, and Sidewalk Connection.

Figure 2 below displays these ratios, offering an early look at how crash involvement varies by infrastructure type. While we are not aiming to establish causality, this type of analysis helps guide which features may be worth exploring further in modeling and supports our ability to aggregate and compare variables across sources.

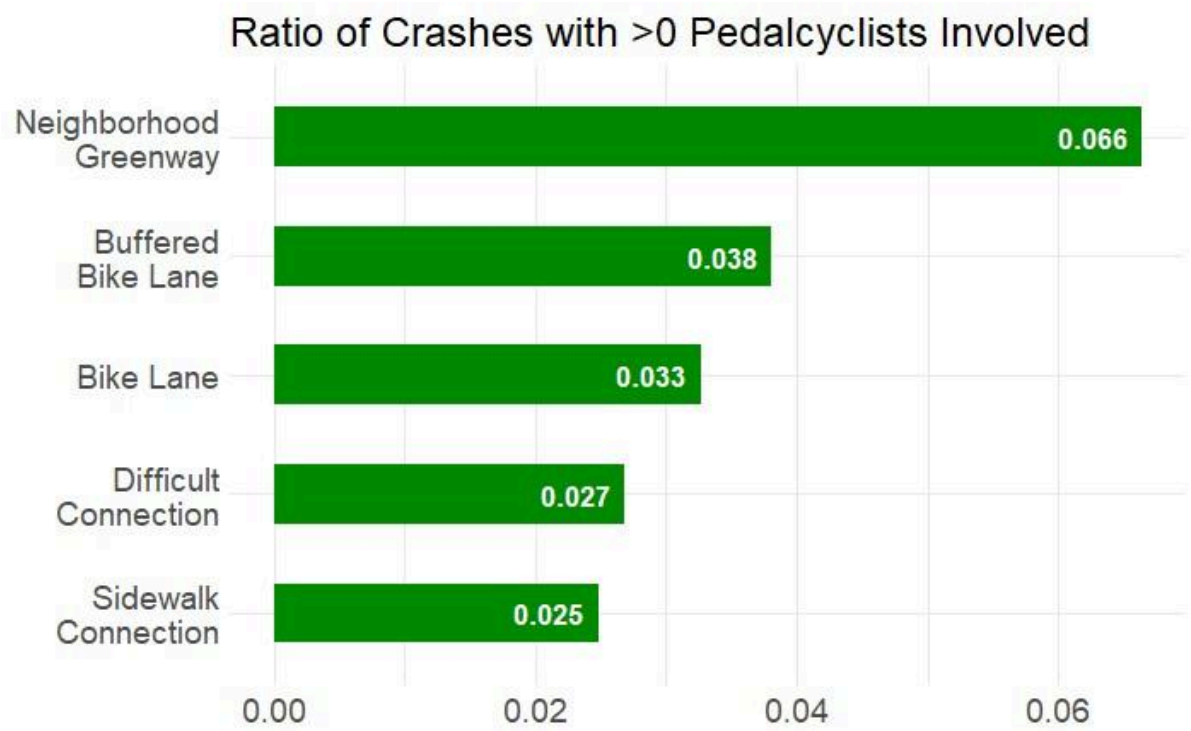

Figure 2. Ratio of Crashes with >0 Pedalcyclists Involved

Below, Figure 3 displays the capabilities of the `bicycle_network` table as joined with the `crashes` table. At each intersection displayed, the number of crashes before a bike lane is built is compared with the number of crashes after a bike lane is built. Of course, the purpose of this is to display the potential power of our database. This graphic does not factor in the time frame of the crash data (2007-2022), so some bike lanes that were built near 2007 will have a higher positive difference. If this were to be controlled for, however, the natural conclusion would be that those intersections with higher numbers of crashes after a bike lane is built become more dangerous, and those that are more negative after a bike lane is built improve safety.

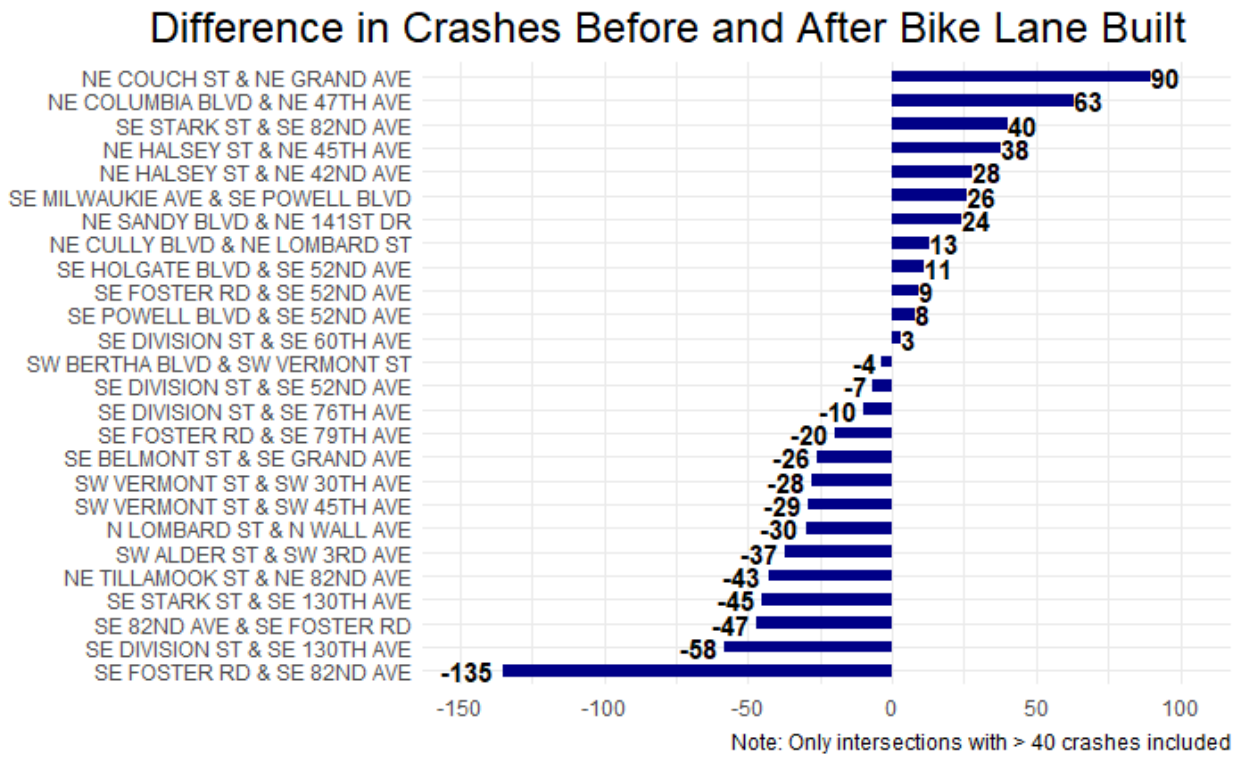

Figure 3. Difference in Crashes Before and After Bike Lane Built