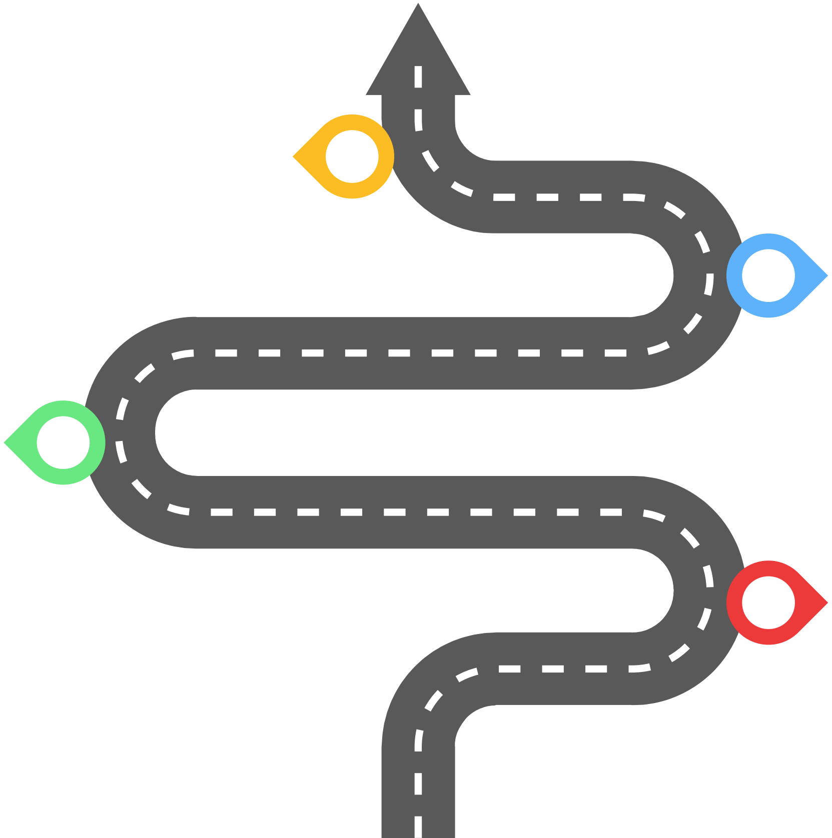
# Chicago Bike Infrastructure Analysis

Hank Snowdon, Daichi Ishikawa,
Christian Piantanida, Toby Chiu



# Executive Summary

**1** **Intro/Problem Definition**

Chicago's Vision Zero Initiative has failed to eliminate all road fatalities - we believe **improving biking accessibility and infrastructure** will help meet this goal

---

**2** **ETL Process**

5 Wrangled Data Sources → OLAP (Snowflake) → Visualizations

---

**3** **Data Analysis**

1. Analyse **biking accessibility/safety** within Chicago
2. Identify improvements towards **integrating biking into Chicago** transportation.

---

**4** **Recommendations**

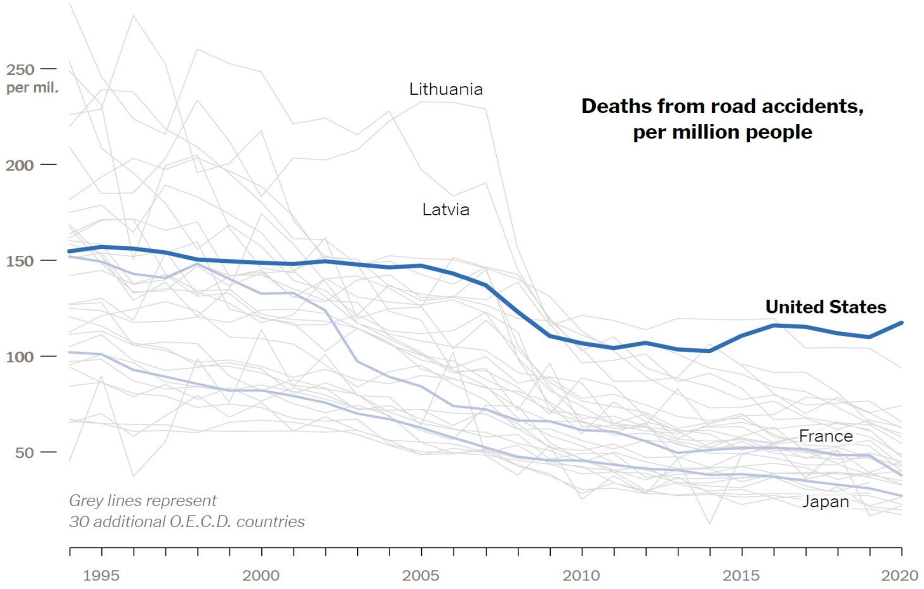
1. **Create safe biking lanes** in high-accident areas
2. **Decrease speed limits and improve infrastructure** on specific arterial roads
3. **Increase Divvy ridership** and accessibility

# Agenda

1. Intro/Problem Definition
2. Data Sources
3. ETL Process
4. Visualization/Insights
5. Recommendations

# Problem Definition

## Biking can present a solution for high traffic incident rates within US/Chicago




Source: Organization for Economic Cooperation and Development • The New York Times

Traffic crashes are a leading source of death and injury in the USA

**Increased bike safety can reduce death and injury, and improve mobility access for non-car owners**

# Introduction

**Premise**

The City of Chicago is engaged in a **Vision Zero Initiative** to decrease road deaths/injuries - Chicago has failed to reduce accidents or meet the zero-death goal

**Reasoning**

Why better bike infrastructure would improve Chicago's roads:

1. **Bikes are less likely to cause road deaths and injuries** as motor vehicles are, as they are far smaller, lighter, and slower and thus cause less damage
2. **Cyclists on certain roads in the city are in high danger of being struck by motor vehicles,** due to a lack of safe, bike-friendly routes

**Target**

**We would like to investigate where and how key investments in bike infrastructure could make the greatest impact on safety and bike ridership**

**Research**

1. Analyse biking accessibility and safety within Chicago
2. Identify improvements towards integrating biking into Chicago transportation.

# Data Sources and Shapes

| | Data | Source | Shape | Description |
|---|---|---|---|---|
| **Crash Data** | Chicago Traffic Crashes | City of Chicago | CSV<br>(23 x 11918) | Chicago biking-related traffic incidents - includes incident type, environmental variables, road type, time. |
| **Biking Data** | Chicago Bike Racks | City of Chicago | CSV<br>(6 x 2198) | Location and type of non-Divvy bike stations in Chicago |
| | Divvy Bike Stations | City of Chicago | CSV<br>(5 x 1420) | Geocoded Divvy bike stations in Chicago. |
| | Divvy Bike Trips | Divvy | CSV<br>(13 x 537114) | Record of all bike trips, - includes timestamp and start/end points. |
| **Demographic Data** | SMART Location Database | US EPA | CSV<br>(30 x 4013) | Chicago Census-Based Demographic Data (Jan 2021). Includes walkability, car-ownership metrics, distance to transit, etc. |

# ETL Steps/Tools

## Multiple Technologies Used to Create Database for Analysis

**Jupyter Notebook**
Data Wrangling

**GCP**
Collaborative
Database Storage

**MySQL**
Database Modelling
Data Pipelining

**Tableau**
Data Visualisation
Storyboarding

# Google Cloud Database

1

## Creating a GCP MySQL instance

```
# Creating a MySQL instance
gcloud services enable sqladmin.googleapis.com
gcloud sql instances create depa-bikes --tier=db-f1-micro --region=us-central1 --database-version=MYSQL_5_
gcloud sql users set-password root --host=% --instance=depa-bikes --password="RootRoot"
gcloud sql instances patch depa-bikes --authorized-networks=0.0.0.0/0
```

2

## Adding an open remote IP to MySQL Instance (for shared database access)

```
# Adding a remote IP to MySQL instance
INSTANCE_NAME="depa-bikes"
NEW_IP="0.0.0.0/0"
CURRENT_IPs=$(gcloud sql instances describe $INSTANCE_NAME --format="value(settings.ipConfiguration.author
UPDATED_IPs="$CURRENT_IPs,$NEW_IP"
gcloud sql instances patch $INSTANCE_NAME --authorized-networks="$UPDATED_IPs"
```

3

## Getting connection IP for GCP instance

```
# Getting connection IP address for GCloud MySQL instance
pip install mysql-connector-python
gcloud sql instances describe "depa-bikes" --format="get(ipAddresses.ipAddress)"
```

# Data Engineering/Transformation

## Ingesting Raw Data into Database Model

1

### Creating Location Lookup Table

- Crash, rack, Divvy trip information is at the latitude/longitude coordinate level in raw data
- Demographic information is at the US Census Block Group level
- To establish a connection between specific locations and demographics of the area, we had to map between locations and their block groups by the following steps:

    1. Extract all coordinates from crashes, bike racks, Divvy trips and stations and round to 4 decimals.
    2. Run each coordinate through the US Census Geocoder API via Python to determine Block Group.
    3. Create and ingest table which maps all relevant locations to Block Groups.

2

### Converting messy fields into consistent formats

- Dates for crashes and Divvy trips were formatted inconsistently within each table.
    - Fix: Using regular expressions to ingest times into DATETIME format depending on different formats within the raw data.
- Longitude and latitude coordinates had inconsistent rounding conventions
    - Fix: Round all coordinates to four decimal places to ensure consistent formatting with generated location table

3

### Remove variables unnecessary for analysis

- Many variables (especially in crash and SMART Location data) were unnecessary for our analysis and were selected for removal upon designing our final data model

# Database Design Rationale

## Create a Snowflake-style OLAP database for easy querying and analysis to accommodate our three fact tables

**Considerations:**

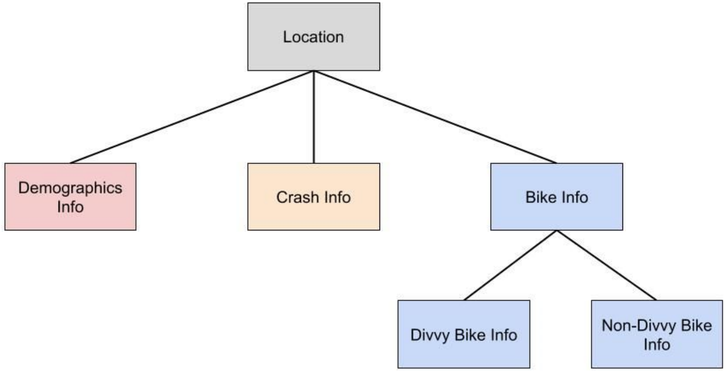
- Our database is wide-reaching and contains tables which relate to bike safety and accessibility in different ways.
  - Bike Crashes
  - Bike Infrastructure (Racks and Divvy Stations)
  - Bike Trips (via Divvy)
- Some tables are very large and fast querying for analysis is important

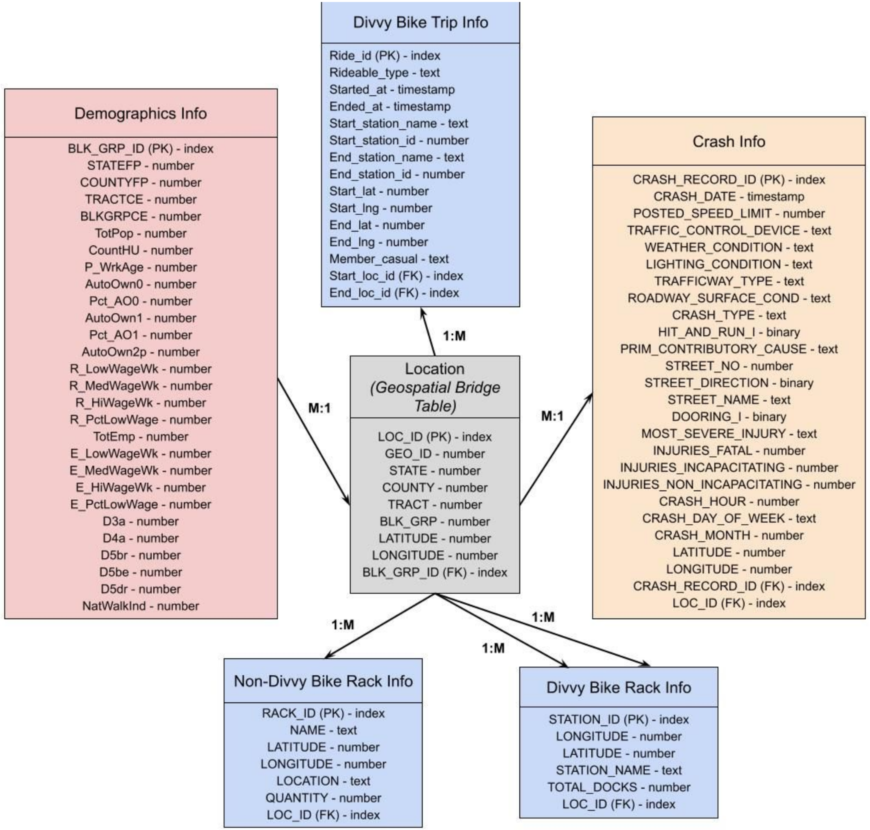
**Solution:**

- Convert from OLTP to OLAP model and establish three main fact tables
- Relate each fact table to location and demographic information
- Normalize by creating important dimension tables which relate to facts:
  - Crash type, road type, weather conditions, time

# Data Models

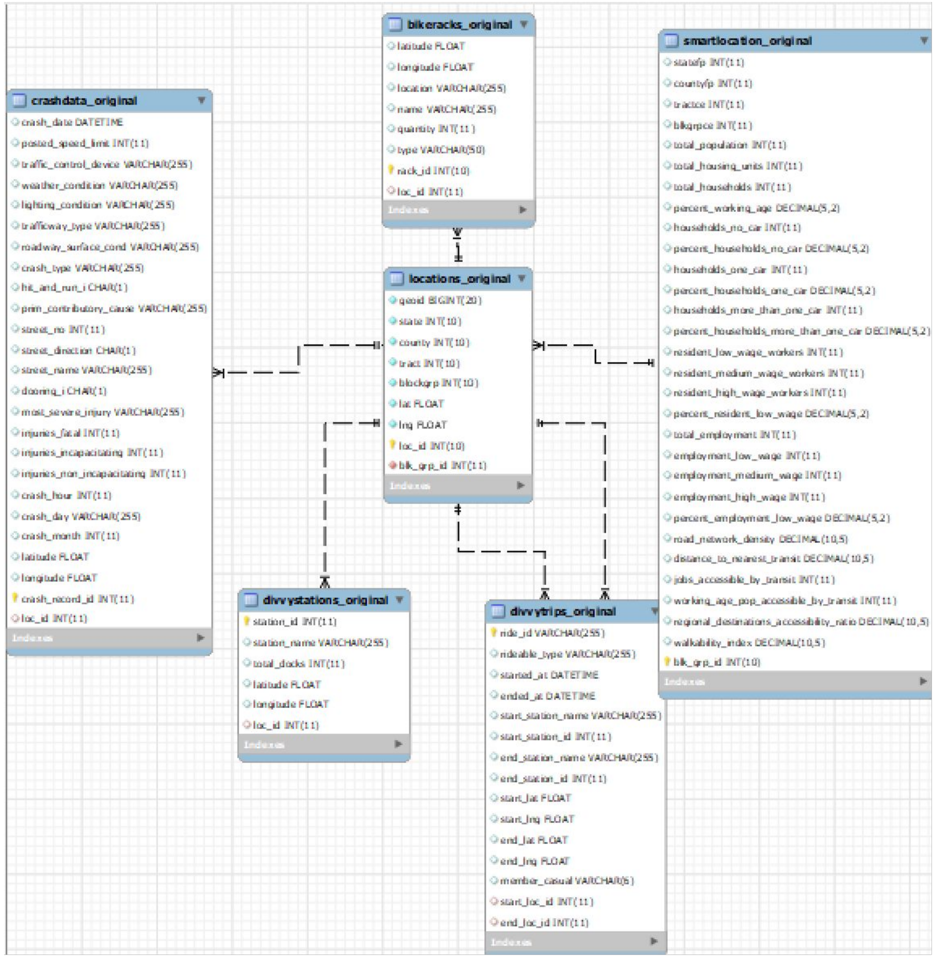
## Conceptual to Logical



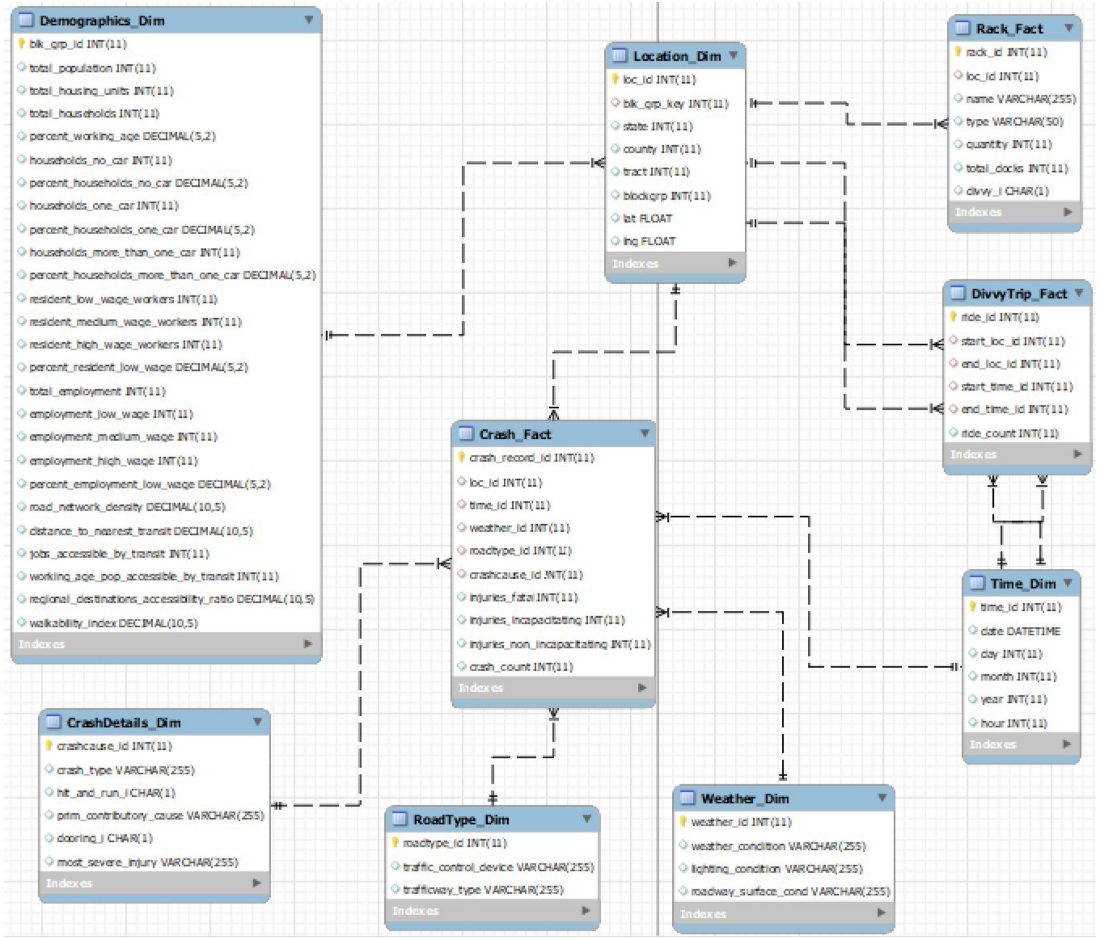
# Data Models

## OLTP Physical Model



# Data Models

## Final OLAP Physical Model



# Database Design

## Final OLAP Model - Snowflake Schema (Fact Tables)

| Table | Table Type | Cardinality | Description |
|---|---|---|---|
| Crash_Fact | Fact | M:1 with 5 dimensions | Records all crashes and injuries in Chicago, stores keys for weather, crash details, road condition, and location data for reference in dimension tables for every crash. |
| Rack_Fact | Fact | M:1 with Location_Dim | Records bike rack info, including type (Divvy/non-Divvy) and capacity, and references location data in a separate dimension |
| DivvyTrip_Fact | Fact | M:1 with 2 dimensions | Records Divvy Trip info, and references trip start/end time, and starting/ending points in separate dimension tables. |

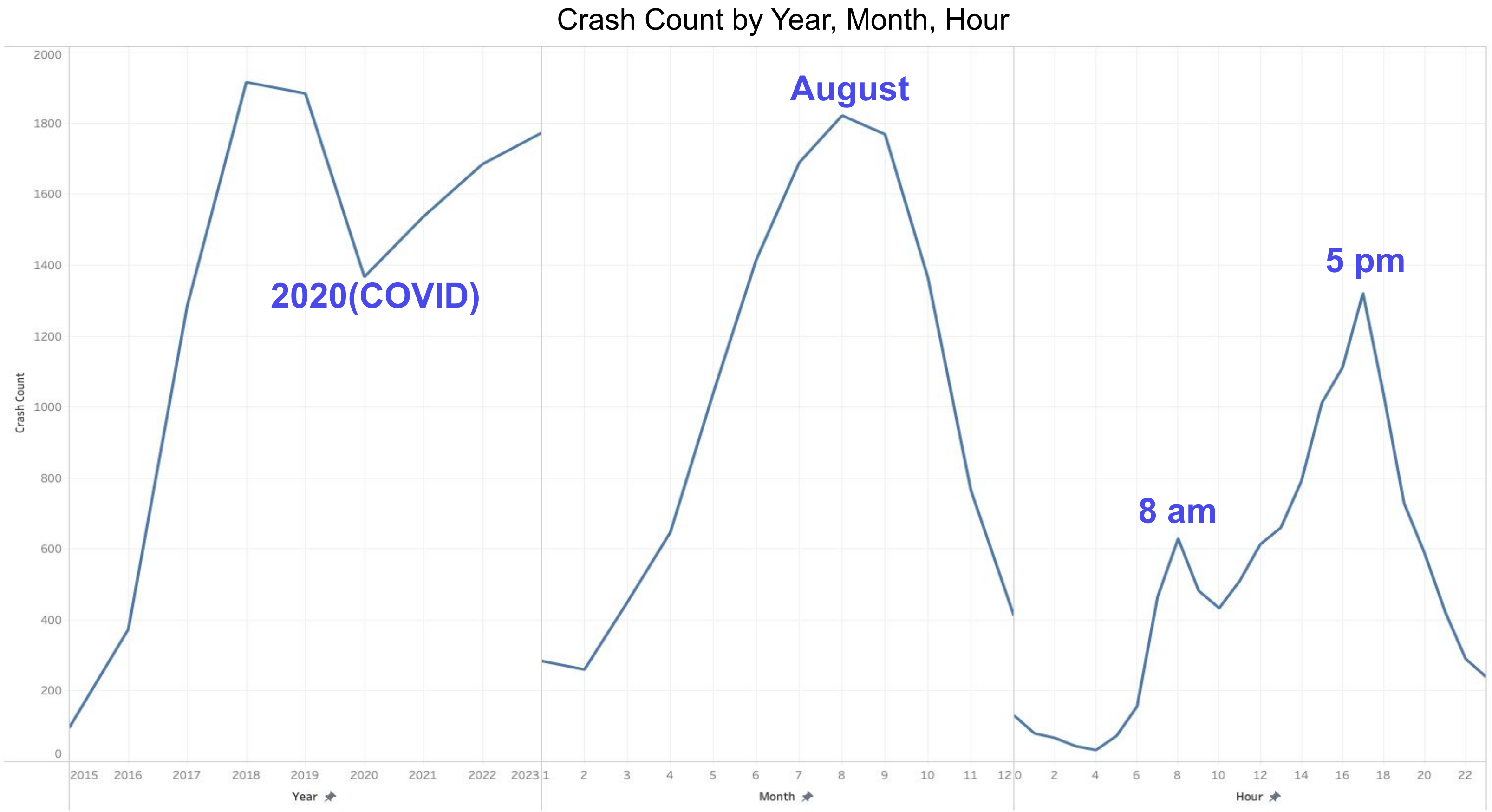
# Database Design

## Final OLAP Model - Snowflake Schema (Dimension Tables)

| Table | Table Type | Cardinality | Description |
|---|---|---|---|
| CrashDetails_Dim | Dimension | 1:M with Crash_Fact | Stores crash-related info, such as type of crash, contributing cause of crash, injuries, etc. |
| Demographics_Dim | Dimension | 1:M with Location_Dim | Stores a wide range of demographic census data for every crash location with tract/block-group granularity, including socioeconomic and accessibility data. |
| RoadType_Dim | Dimension | 1:M with Crash_Fact | Stores info on traffic road conditions for **crashes,** particularly around traffic signalling conditions. |
| Weather_Dim | Dimension | 1:M with Crash_Fact | Stores info on weather and lighting conditions for **crashes,** and how it affects road conditions. |
| Location_Dim | Dimension | 1:M with 3 Fact Tables<br>M:1 with Demographics_Dim | Stores all geocoded/location data, including latitude/longitude coordinates, and county/tract data for every point. Also acts as a bridge various geospatial-based primary keys to allow for comparison. |
| Time_Dim | Dimension | 1:M with 2 Fact Tables | Stores all timestamp data, including specific time of day, date, month, and year data. |

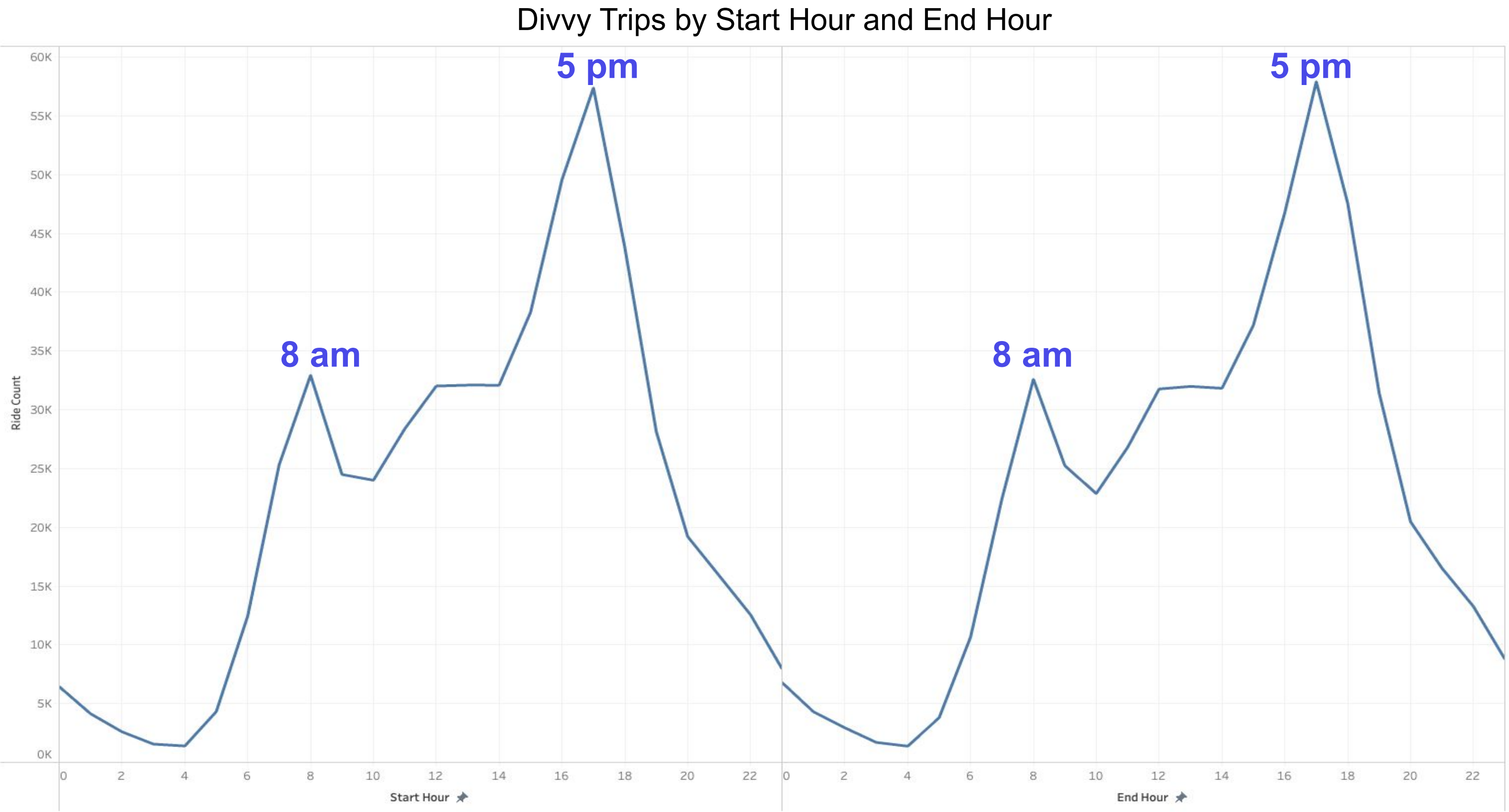
# Crashes by Year, Month, Hour

## Most accidents occur in summer, with peak times being during morning commutes and evening returns



# Divvy Trips by Start Hour and End Hour

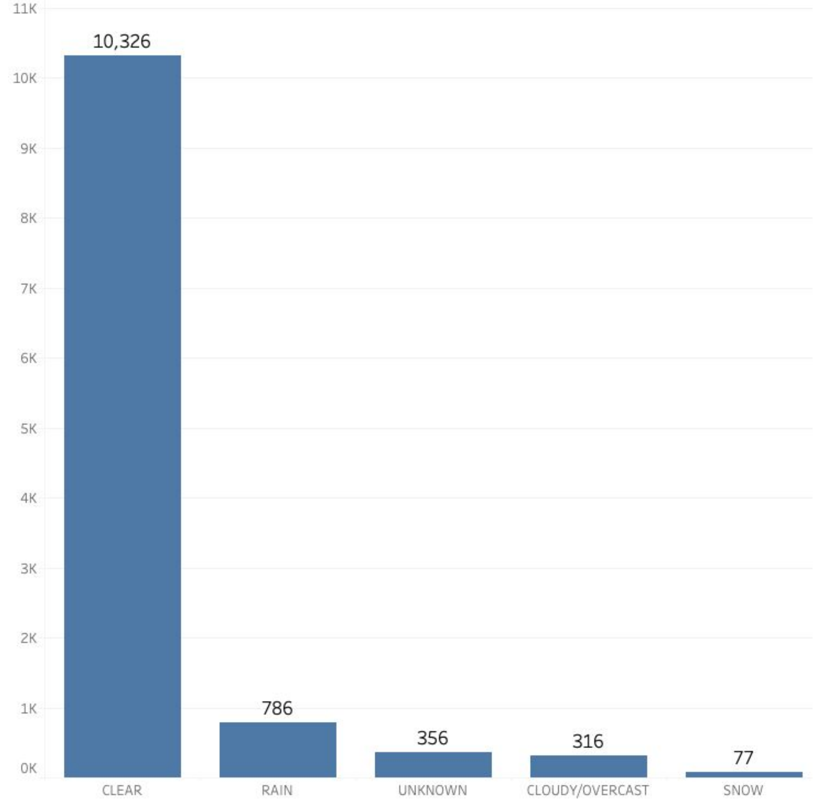
## Peak ride frequency is the same as peak crash frequency



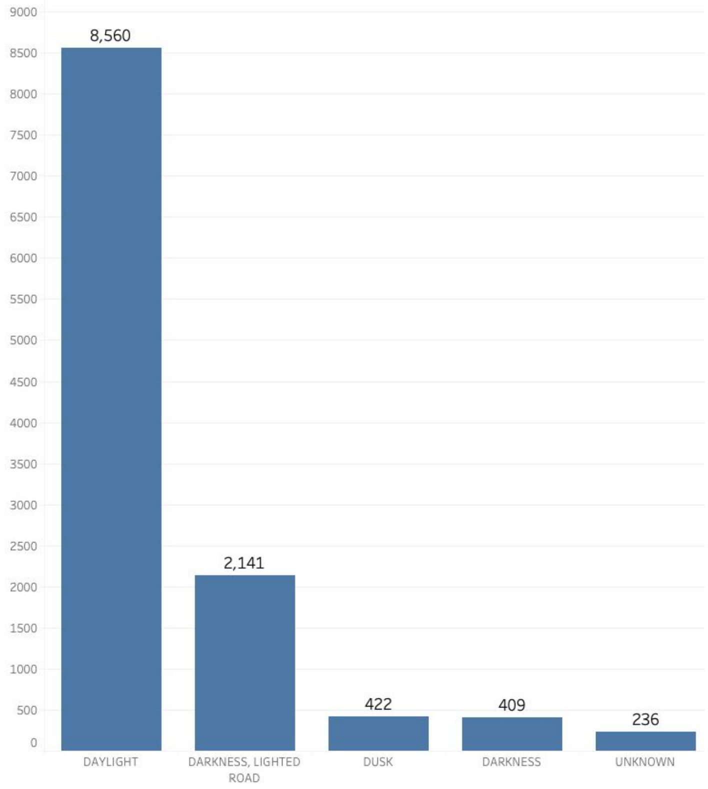
# Crashes by Weather and Lighting Conditions

## Weather conditions and lighting don't seem to be major factors

Crash Count by Weather conditions (Top 5)



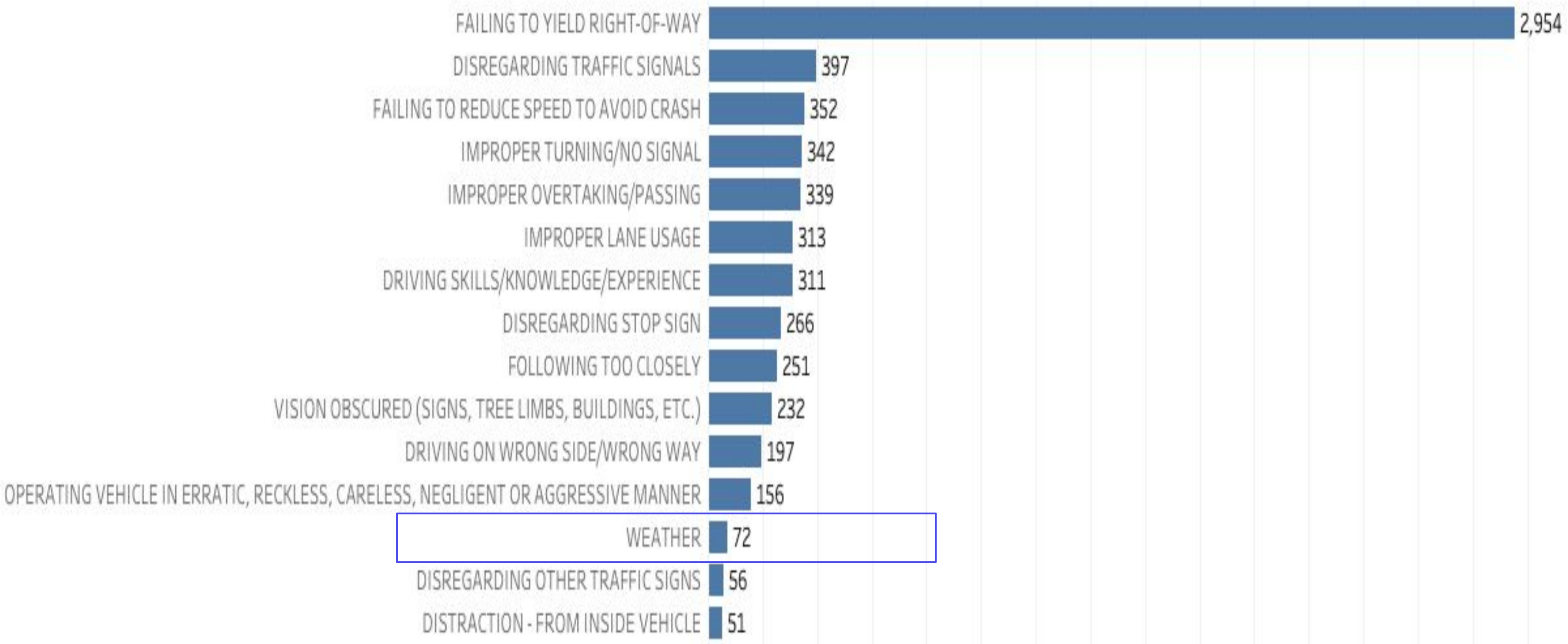

Crash Count by Lighting conditions (Top 5)



# Crash Cause

## Causes are mostly due to improper driving or rule violations

Crash Count by Cause (Top 15)

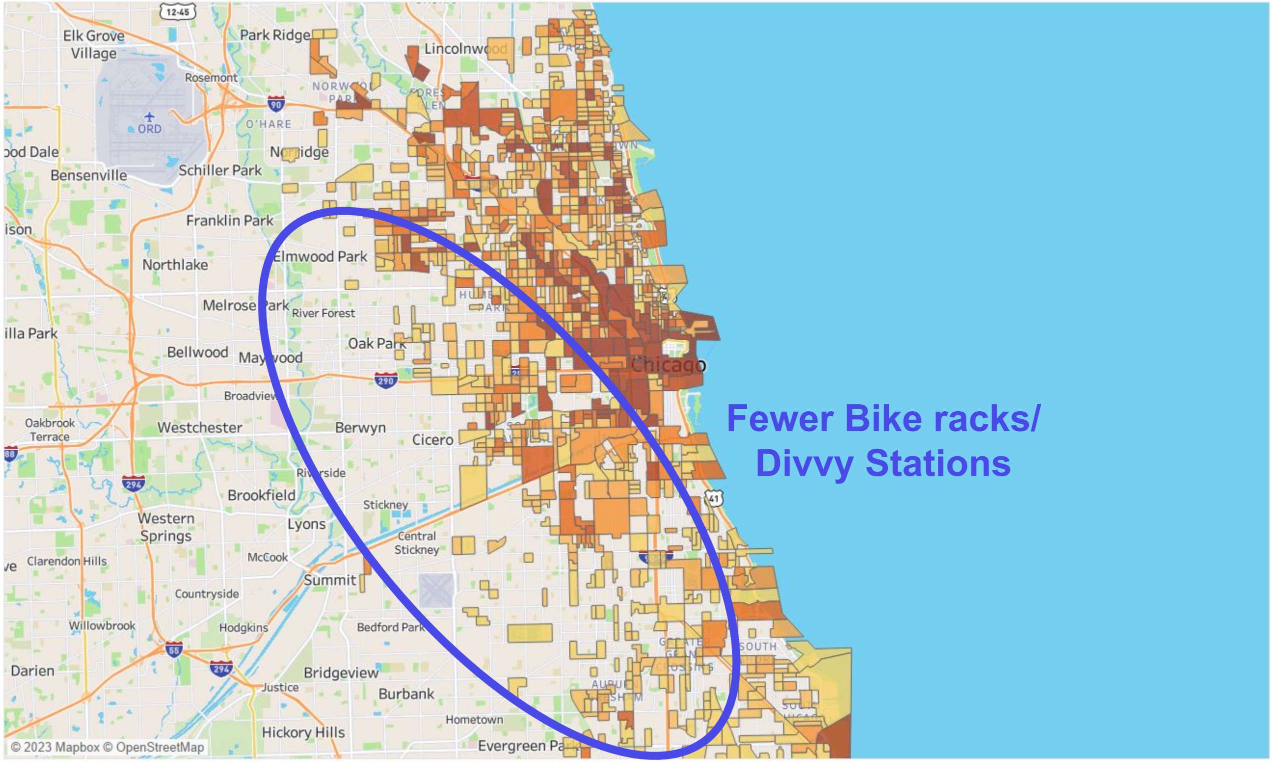



*Unable to determine and Not Applicable are excluded

# Bike Racks and Divvy Station Distribution

## Not spread evenly across the city

Density of Divvy stations and city racks by block group



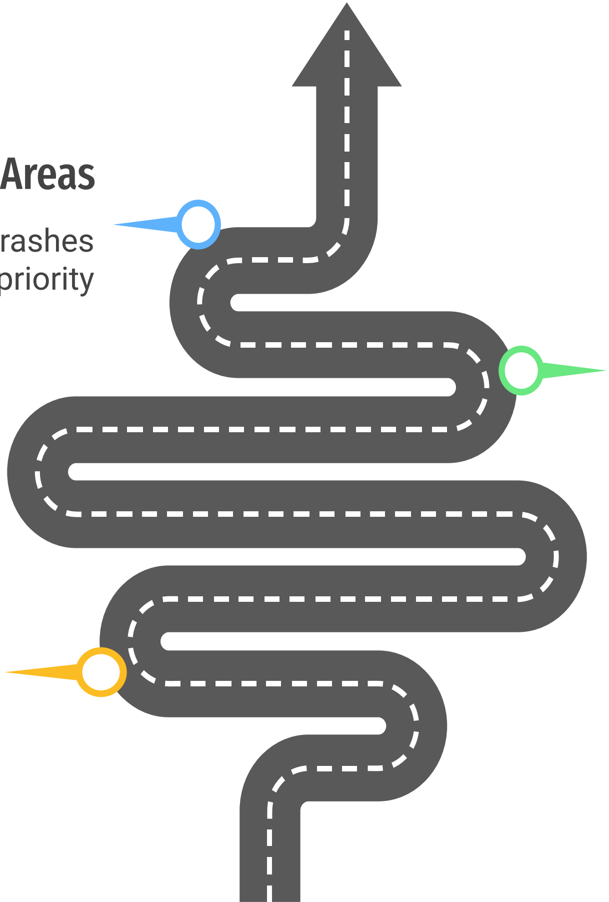
# EDA Takeaways:

Crashes are mainly caused by driver error, not by conditions or cyclist mistakes

Physical infrastructure (protected bike lanes) is the best way to physically separate bikes from dangerous cars

Existing bike infrastructure varies across the city

# Key areas of infrastructure intervention



## High-Volume Crash Areas

Parts of the city where crashes occur often must be the first priority

## Critical areas with low bike traffic

Many neighborhoods with low bike traffic need lower-cost options than expensive driving

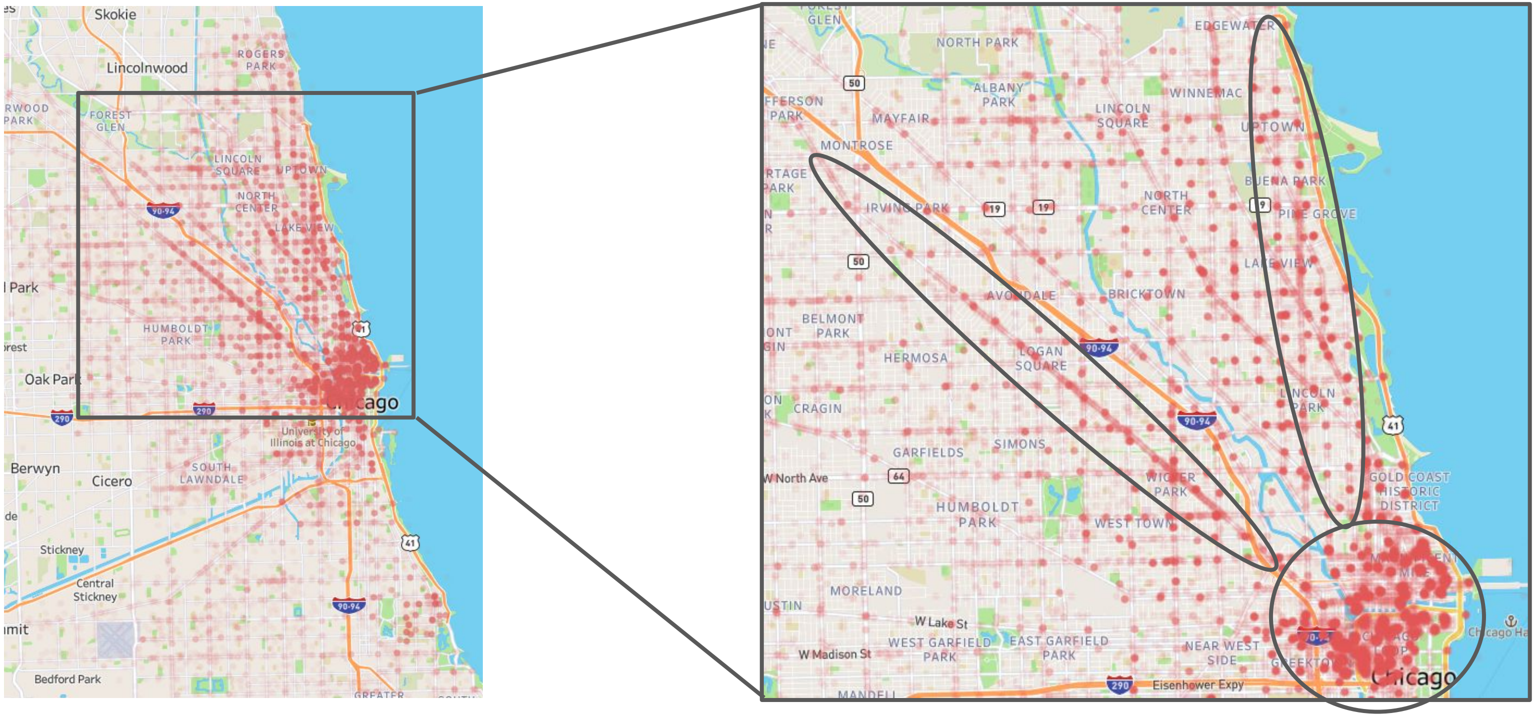
## Limited Divvy/bike storage accessibility

South/West parts of Chicago have poor Divvy station and bike rack coverage

# Crash-Prone Areas:
## Critically dangerous for bikers

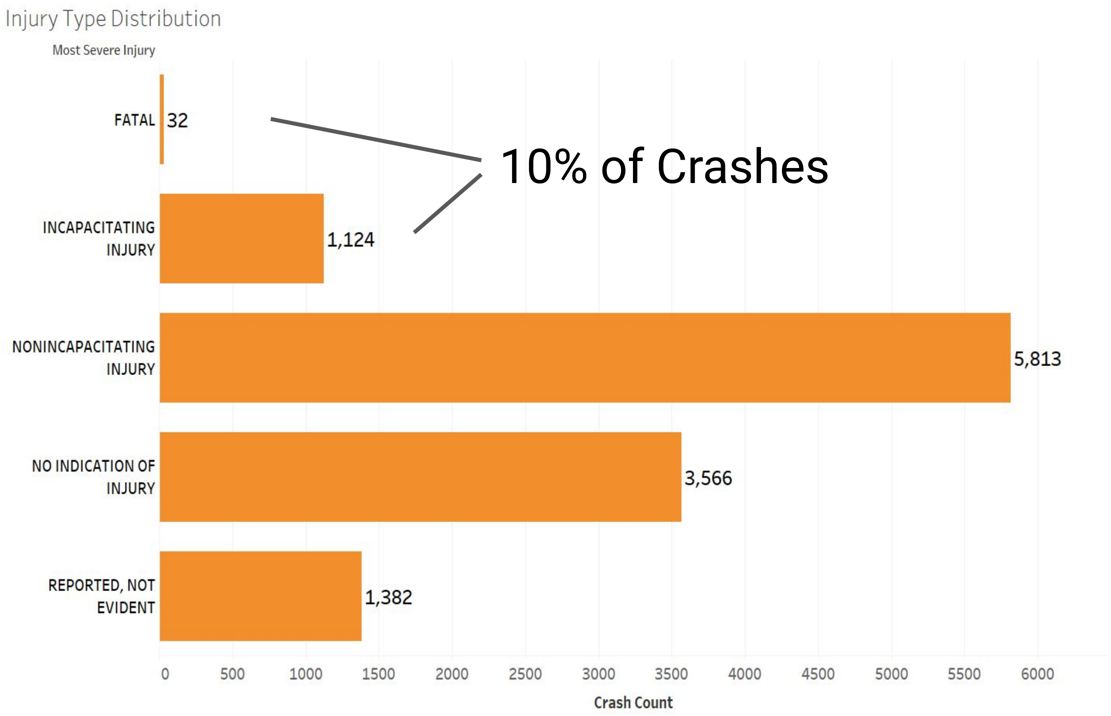
# Crash Heat Map

## Accidents are most frequent in downtown areas and along two roads stretching north: Milwaukee Ave and Clark St

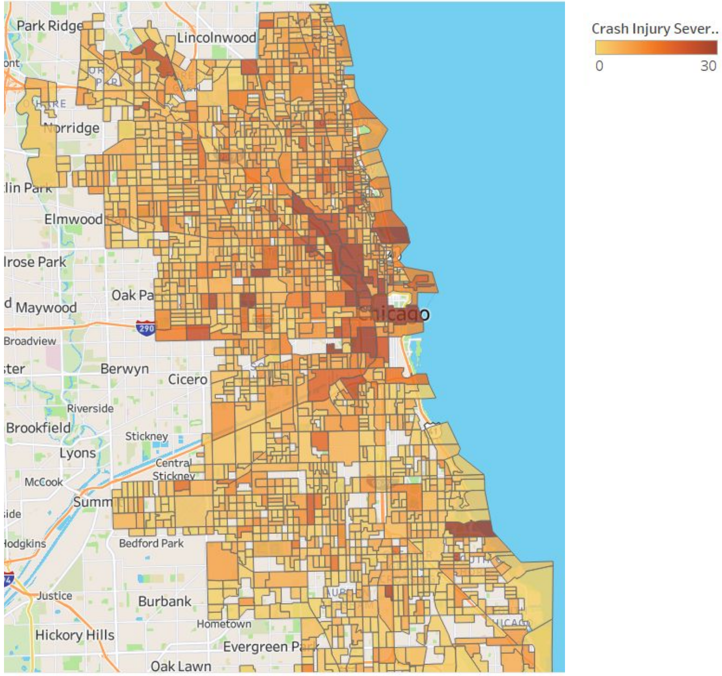


# Injury Type Distribution

## While most injuries in bike crashes are non incapacitating, many are serious injuries



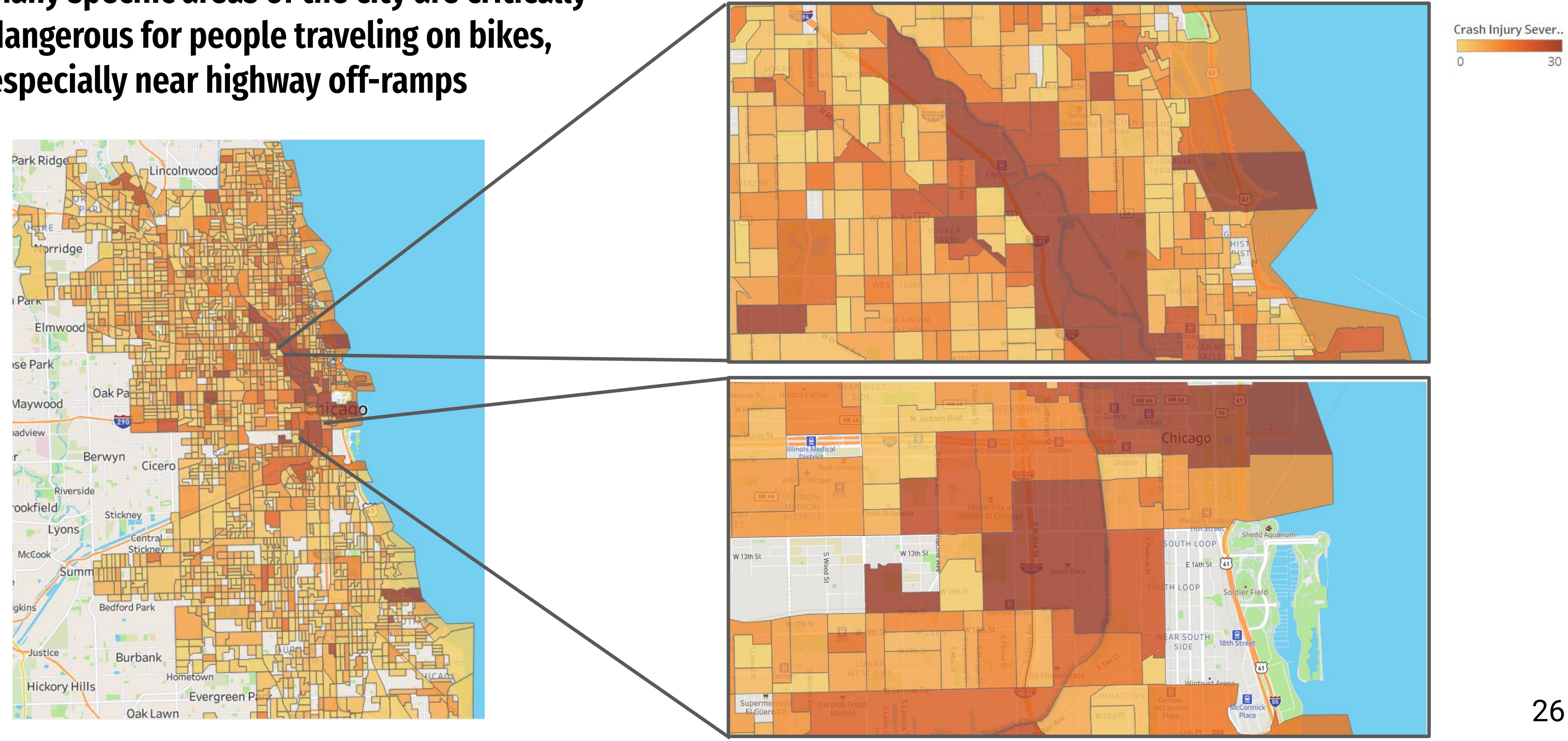

Sum of Crash Count for each Most Severe Injury. The marks are labeled by sum of Crash Count.




d Latitude (generated). Color shows sum of Crash Injury Severity. Details are shown for GEOID (tl 2019 17 bg.shp1). The data is filtered on Action (Most Severe Injury), which keeps 5 members. The view is filtered on GEOID (tl 2019 17 bg.shp1), which keeps 1,852 of 2,104 members.

# Injuries weighted by severity

**Many specific areas of the city are critically dangerous for people traveling on bikes, especially near highway off-ramps**

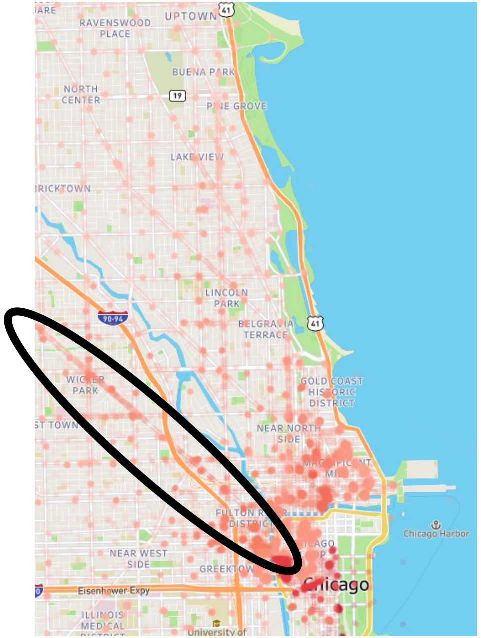


# Crashes and Demographic info

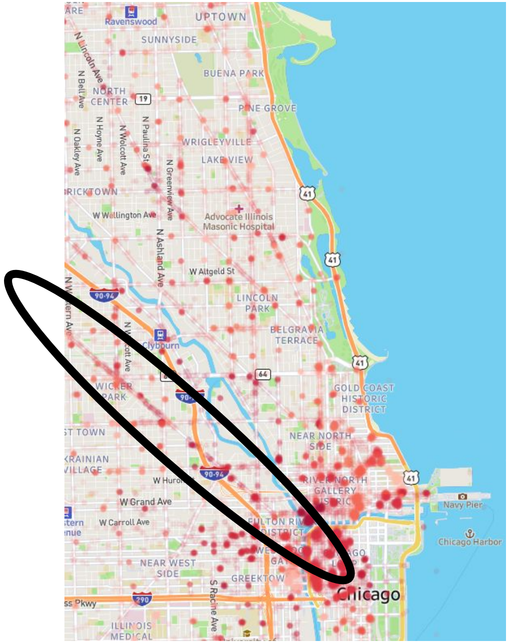
**The accessibility situation on Clark St is worse than on Milwaukee Ave:**

- **Clark lacks access to rail transit unlike Milwaukee (situated on Blue Line)**
- **Need better bikeability West from Clark to access Red/Purple/Brown Lines**

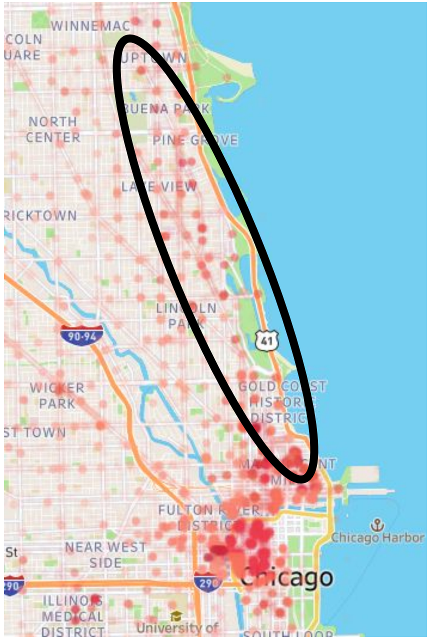
Destination Accessibility Index




Walkability Index




% of Households w/ No Car



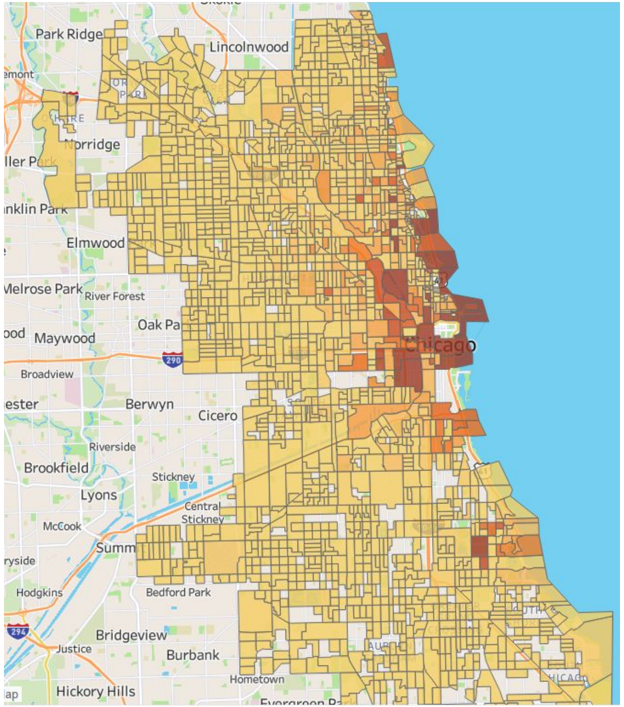
# Low-Traffic Bike Areas:

## In many parts of Chicago very few people bike, despite it being much more cost-effective than driving

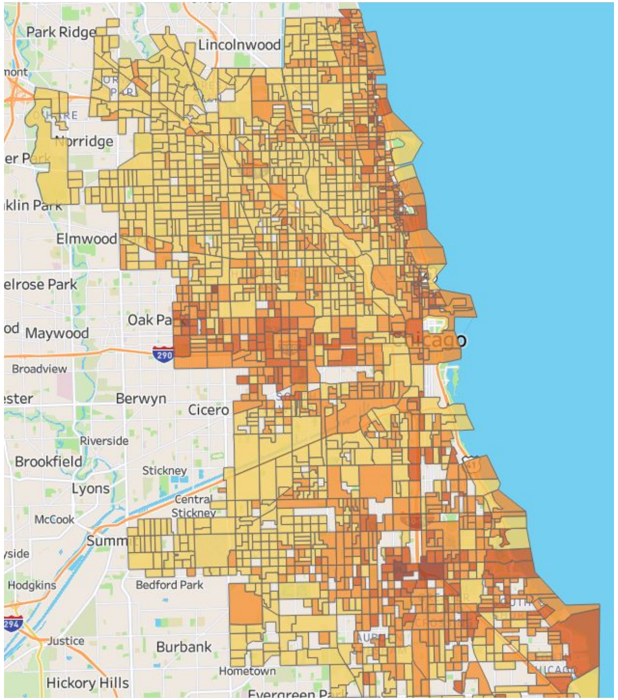
# Lack of Bike Traffic in Areas With Worse Access to Transportation

## Divvy bike traffic is primarily focused Downtown and on the North Side, with very low traffic on South/West Sides
## The areas with low traffic are the same areas where few people own cars

Divvy Trip Frequency by Block Group




% Households with no car by Block Group



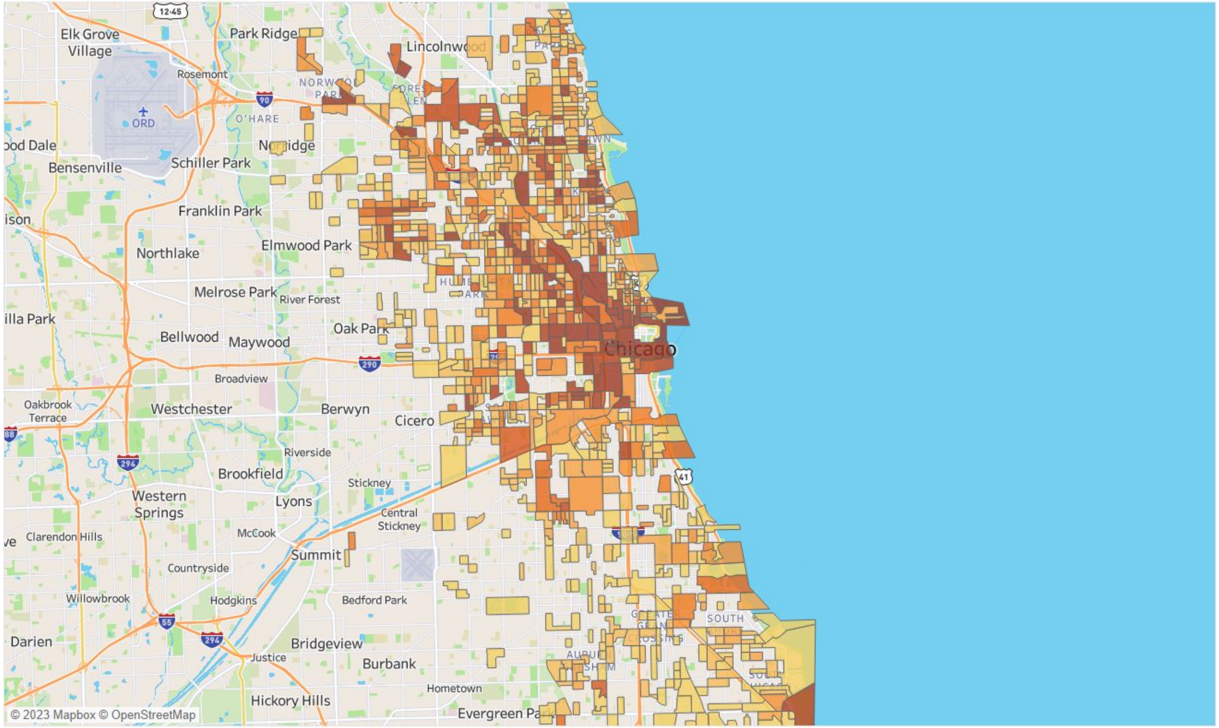
# Limited Divvy/bike storage accessibility:

## The same areas of the city lack Divvy infrastructure and city-owned bike racks, and thus hurt biking accessibility

# Map of Bike Racks and Divvy Stations

## Certain areas of the city have significantly fewer racks and stations to enable bikers

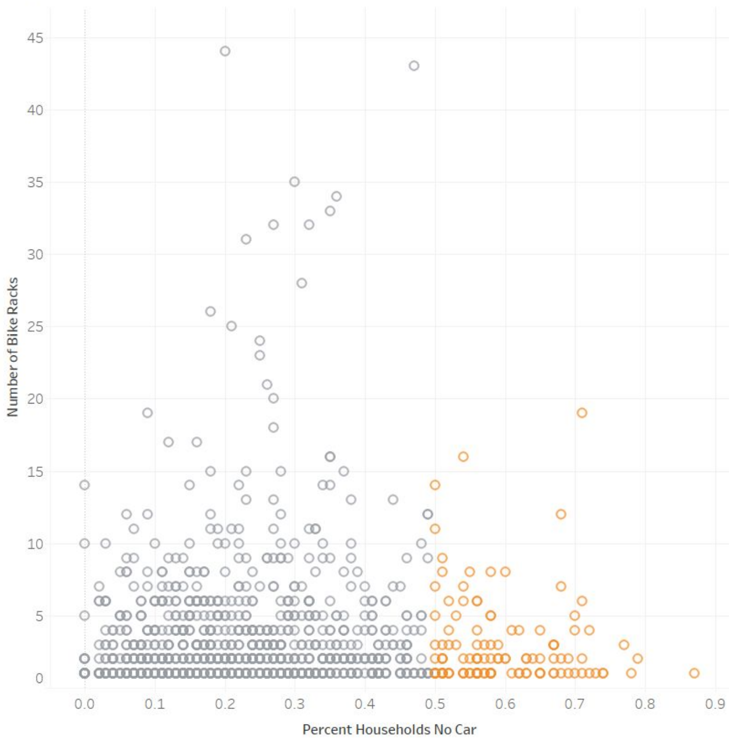
Rack Map



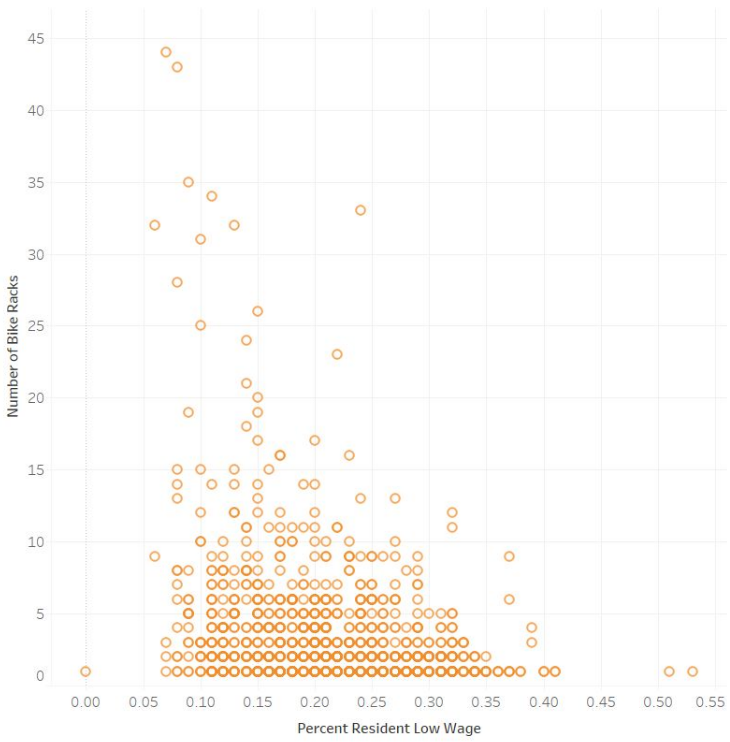

Map based on Longitude (generated) and Latitude (generated). Color shows sum of New_Quantity. Details are shown for Geoid. The view is filtered on Geoid, which keeps 1,034 of 2,104 members.

# Bike Rack Distributions by Demographics

## Areas with fewer racks are more likely to be poorer and have lower car ownership




Sum of Percent Households No Car vs. sum of New_Quantity. Color shows details about Flag. Details are shown for Geoid. The view is filtered on Geoid, which excludes Null and 170318391001.




Sum of Percent Resident Low Wage vs. sum of New_Quantity. Details are shown for Geoid. The view is filtered on Geoid, which excludes Null and 170318391001.

# Recommendations

**Decrease the number of crashes and encourage bike ridership**

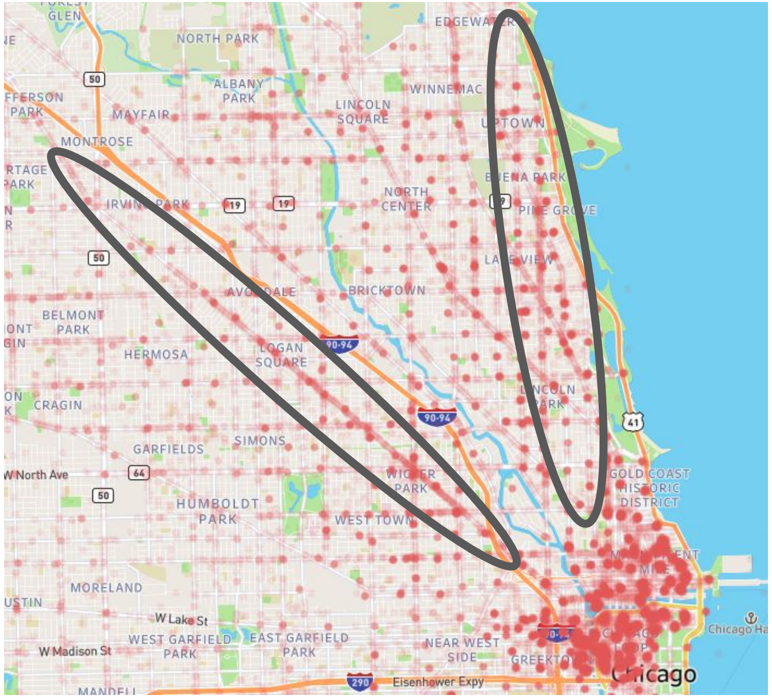
1. **To decrease the number of crashes**
    a. Protected bike lanes and lower speed limits on Clark St and Milwaukee Ave
    b. Protected bike routes West from Lakefront to increase safe access to transit
    c. Create safe bike paths in specific dangerous crash areas (Clybourn, etc.)

2. **Encourage further ridership**
    a. Protected bike lanes in South Side
    b. Influx of Divvy stations/bike racks in South/West sides

# Recommendations

**Protected bike lanes and lower speed limits on Clark St and Milwaukee Ave**

- Reduces crashes on most crash-prone thoroughfares




# Recommendations

## Protected bike routes East-West from Lakefront toward transit options

- Increases safety of transit access and connections via bike

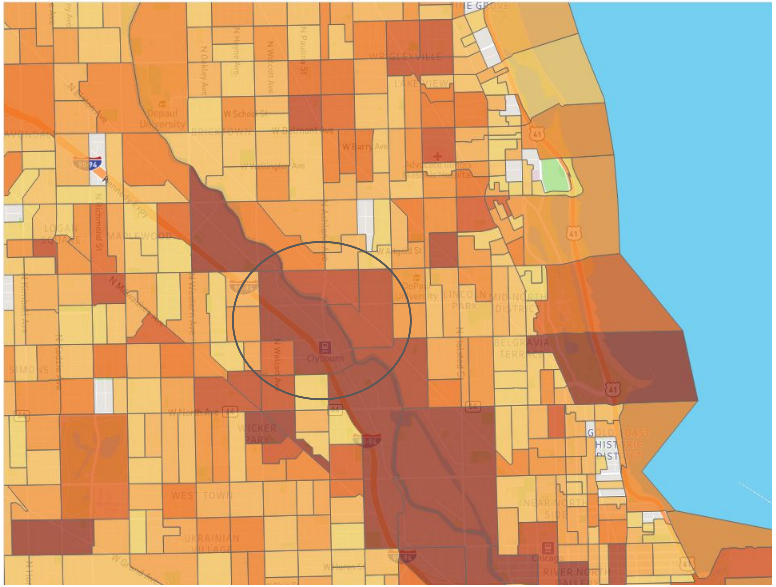
Clark St
CTA


# Recommendations

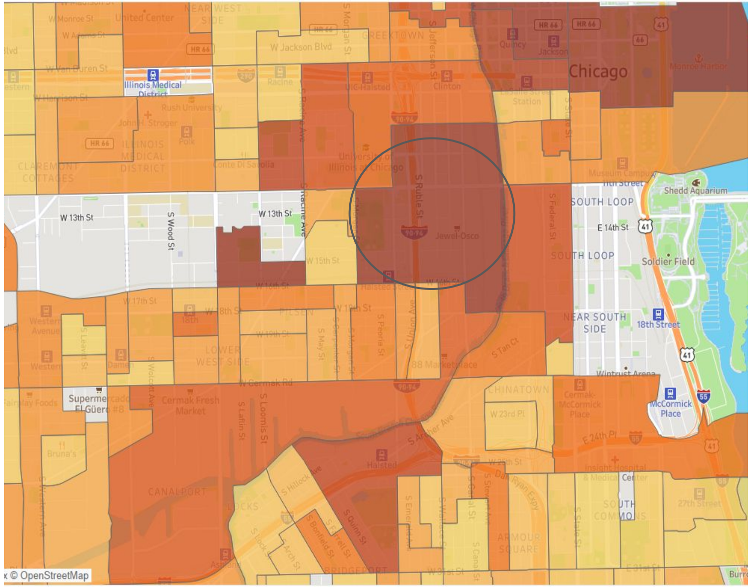
## Safe bike paths in specific dangerous crash areas

- Reduce serious injuries in high-risk areas

I-90/94 exit near Clybourn Metra Stop



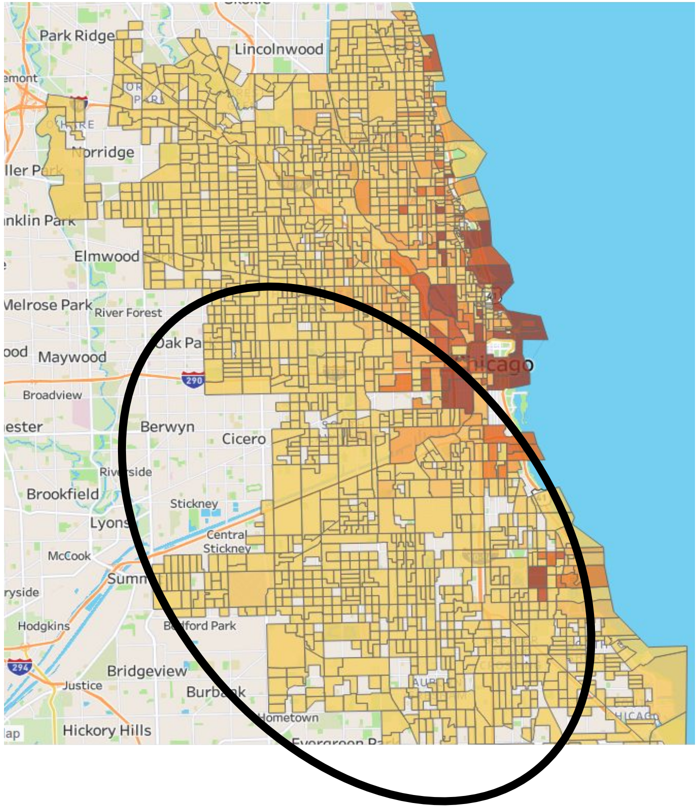

I-90/94 exit near UIC



# Recommendations

## More Divvy stations, bike racks, protected bike lanes in South and West Side

- Encourage further ridership in vulnerable communities

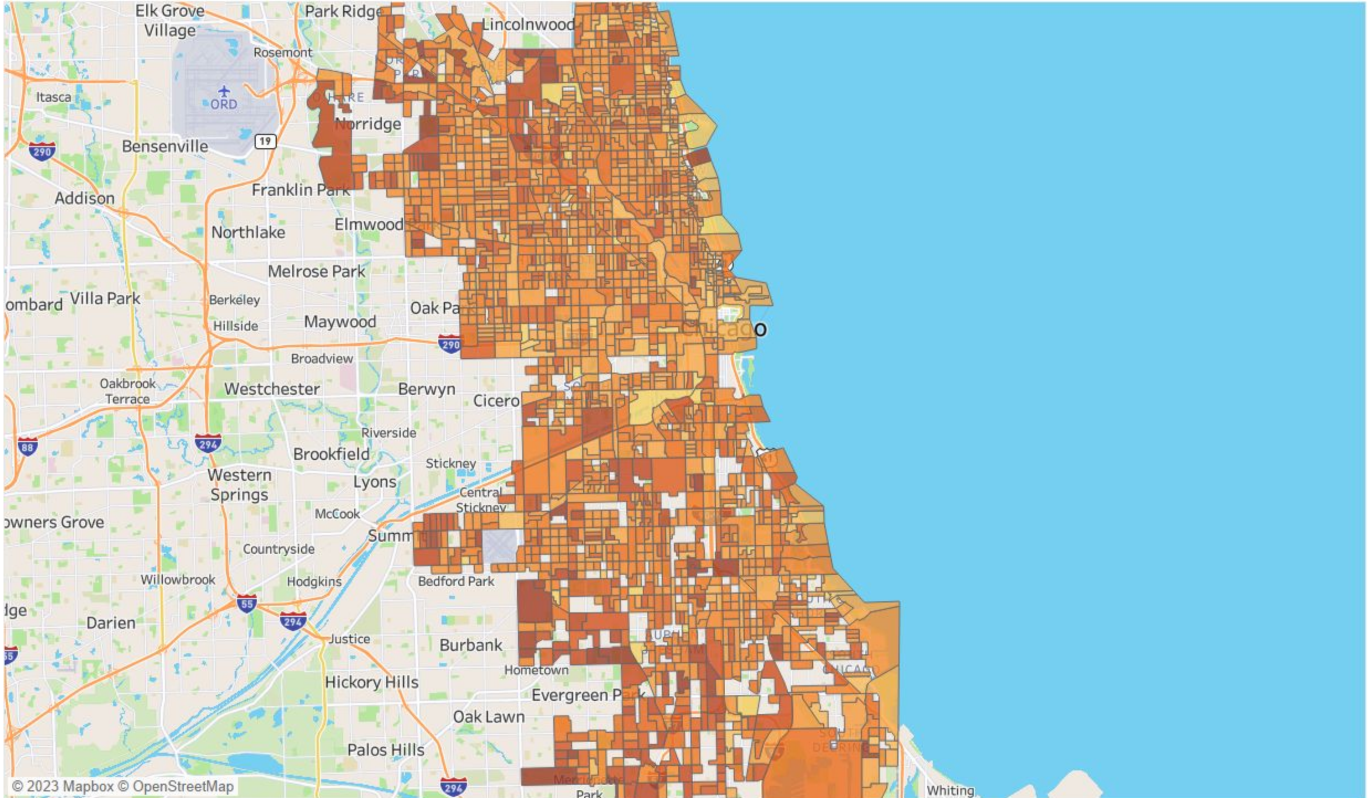



# Lessons Learned

1. Using Raw Data from Government Sources is Difficult
2. Collaborating on a Cloud Database Required Careful Planning
3. Working with Geospatial Data is Strenuous

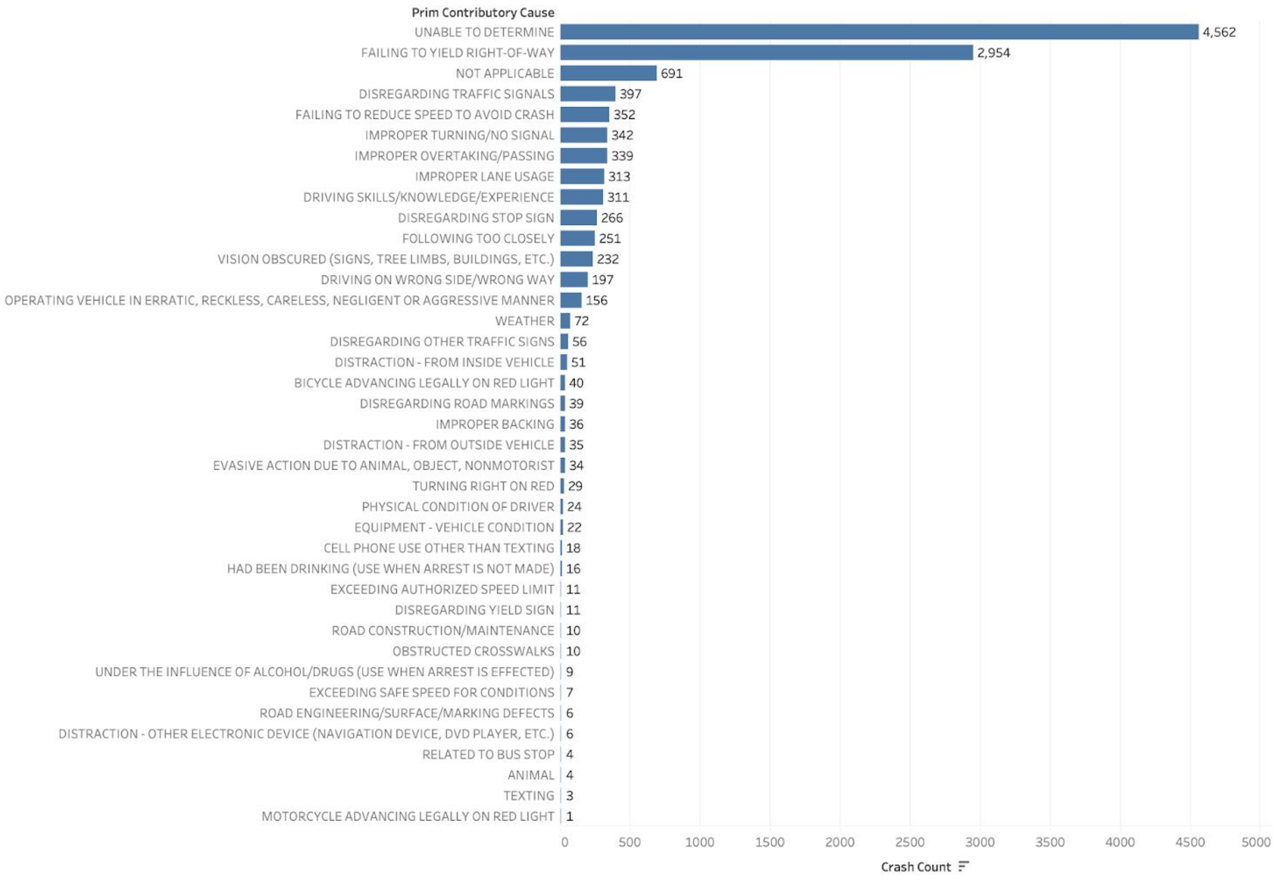
# Appendix

## Distance to Transit




*Distance to nearest available transit by block group*

Crash Count by Cause




Count of Crash Record Id for each Prim Contributory Cause. The data is filtered on Crash Type, which keeps INJURY AND / OR TOW DUE TO CRASH and NO INJURY / DRIVE AWAY.

*Full count of contributing causes for road crashes (condensed on slide 19)*