**Project ETL - Extract, Transform, Load**

**Team members**

Swati Swain

Emi Rivera

Nick Prentkowski

Terrence Cummings

**Minneapolis Police Incidents, Use of Force, and Demographics**

Project Objective:

To create a database that will facilitate an analysis of the relationship between police incidents, the use of force by police in those incidents, and the demographic characteristics of the neighborhood in which the incidents occurred such as race and income.

The team will extract data from various sources, clean and normalize the data, and load the data into a database. Such data can then be used by applications to allows analysis of relationships and trends between these factors.

The target data, for the city of Minneapolis, includes:

1. Police incident data by neighborhood
2. Police “use of force” data by neighborhood
3. Income levels by neighborhood
4. Race distribution by neighborhood

The follow sections describe the sections described the detailed steps in the source, extraction, transformation, and loading of this data into a PostgreSQL database. A relational database was selected for ease of joining disparate datasets on a common key such as neighborhood ID.

The overall entity relationship diagram (ERD) is as below:

[insert ERD screenshot]

Table of Contents

[Police Incidents 3](#_Toc42025696)

[Description 3](#_Toc42025697)

[Data sources 3](#_Toc42025698)

[Data extraction 3](#_Toc42025699)

[Data transformation 3](#_Toc42025700)

[Data loading 3](#_Toc42025701)

[Police Use of Force 4](#_Toc42025702)

[Description 4](#_Toc42025703)

[Data sources 4](#_Toc42025704)

[Data extraction 4](#_Toc42025705)

[Data transformation 4](#_Toc42025706)

[Data loading 4](#_Toc42025707)

[Neighborhood Race Demographics 5](#_Toc42025708)

[Description 5](#_Toc42025709)

[Data sources 5](#_Toc42025710)

[Data extraction 5](#_Toc42025711)

[Data transformation 5](#_Toc42025712)

[Data loading 5](#_Toc42025713)

[Neighborhood Income Demographics 6](#_Toc42025714)

[Description 6](#_Toc42025715)

[Data sources 6](#_Toc42025716)

[Data extraction 6](#_Toc42025717)

[Data transformation 6](#_Toc42025718)

[Data loading 6](#_Toc42025719)

[Web Application 7](#_Toc42025720)

# Police Incidents

## Description

## Data sources

## Data extraction

## Data transformation

## Data loading

# Police Use of Force

## Description

## Data sources

## Data extraction

## Data transformation

## Data loading

# Neighborhood Race Demographics

## Description

The objective is to obtain data regarding the racial mix of Minneapolis neighborhoods and communities.

## Data sources

Data is obtained from MINNESOTA COMPASS (mncompass.org). We need to scrape data from the following endpoints:

1. Scrape links to Minneapolis neighborhood-specific webpage on mncompass.org found on:
   1. <http://www.mncompass.org/profiles/neighborhoods/minneapolis-saint-paul>
2. Scrape links to Minneapolis community-specific webpage on mncompass.org found on:
   1. <http://www.mncompass.org/profiles/neighborhoods/minneapolis-saint-paul>
3. Scrape race data for each Minneapolis neighborhood links obtained in step 1. For example, Armatage neighborhood at:
   1. <http://www.mncompass.org/profiles/neighborhoods/minneapolis/armatage>
4. Scrape race data for each Minneapolis community links obtained in step 2. For example, Camden at:
   1. <http://www.mncompass.org/profiles/communities/minneapolis/camden>

## Data extraction

Selenium webdriver (from selenium import webdriver) was used to scrape data at the URL. This is because the data is populated by Javascript and therefore not accessible by Splinter.

Extraction followed the following process:

1. Scrape the individual neighborhood and community links and store in lists of URLs.
2. Send the webdriver to each link in the lists and scrape the race data from each page.

The neighborhood and community race data is then stored in a Pandas dataframe and written to csv files.

## Data transformation

Transformation (cleaning) involved the following steps:

1. Read in the csv’s from extraction as Pandas dataframes.
2. The scraped data contained the word ‘suppressed’ in some table cells. Replace this with NaN so all missing data is represented by NaN.
3. Convert text-styled numbers into numeric type.
4. Add a ‘total’ column as the sum of the individual race columns.
5. Use pd.merge to bring in neighborhood and community ID’s that will be used in PostgreSQL keys.
6. Delete extraneous columns.
7. Reorder columns for presentability.

## Data loading

Steps:

1. Create tables in PostgreSQL using the SQL script based on the ERD presented at the beginning of this document.
2. Use Sqlalchemy (from sqlalchemy import create\_engine) to connect to PostgreSQL database.
3. Use Pandas df.to\_sql to populate PostgreSQL tables with Pandas dataframe values.

# Neighborhood Income Demographics

## Description

## Data sources

## Data extraction

## Data transformation

## Data loading

# Web Application