# Crime and Health in the Mile High City

Correlations in Crime and Health by Neighborhood in Denver, CO

Naureen Bharwani
Computer Science
University of Colorado Boulder
Boulder, CO USA
naureen.bharwani@colorado.edu

Sean Mulligan
Computer Science
University of Colorado Boulder
Boulder, CO USA
sean.mulligan@colorado.edu

Cody Thornton
Computer Science
University of Colorado Boulder
Boulder, CO USA
cody.thornton@colorado.edu

## PROBLEM STATEMENT

We intend to explore the intersection of crime and health in Denver, CO with an emphasis on comparing neighborhoods and their demographic characteristics in the city. In doing so, we hope to shed light on what crimes occur where, and whether there is a measurable correlation between the rates of these crimes and health outcomes as measured by metrics like obesity, and life expectancy. We also wish to look at how crime and health outcomes have changed over time, which may be related to events like the COVID-19 pandemic and the introduction of social programs like Denver's Star Program.

## LITERATURE SURVEY

**The Denver Department of Public Safety [1].** The Denver Department of Public Safety (DOS) launched an initiative called The Denver Opportunity Index. The initiative aimed at identifying areas within the city, where residents' opportunities were less likely than other areas within the city. The Department of Public Safety's object is to focus on these neighborhoods with less opportunity to increase their overall quality of life. Within the initiative the DOS deems quality of life based on financial security, behavioral health(do they have access to health insurance? Does violent crime impact my life?), and people left behind(barriers to employment, housing, education, and access to transportation).

**The US Department of Housing and Urban Development [2].** In 2016, the U.S. Department of Housing and Urban Development published a review of studies related to crime in their publication, *Evidence Matters*. This review notes that violent crime in the United States has dropped by half, a truly "massive" amount, between 1995 and 2014. Decreases in the crime rate have been most dramatic in disadvantaged communities. However, significant disparities still exist in the rate of violent crime between different neighborhoods in American cities. Predominantly African-American communities average five times more violent crimes than predominantly white communities, and predominantly Latino neighborhoods average about two and a half times as many violent crimes as predominantly white neighborhoods. Similarly, low-income people are much more likely than others to experience crime, including violent crime. Furthermore, within neighborhoods, research has indicated that violent crime occurs in a small number of "hot spots" - small geographic areas like street intersections or street segments (two block faces on both sides of a street between two intersections). Policing these hot spots appears to be an effective way to reduce crime. In general, exposure to violence puts youth at significant risk for psychological, social, academic, and physical challenges and also makes them more likely to commit violence themselves. Studies also suggest that violent crime decreases property values.

**The National Institutes of Health [3].** A study of the children of 119 families over three years with assessments occurring one year apart. This study focused on three types of violent crime: marital aggression, parent-to-child aggression, and community violence. The authors of the study ranked cumulative exposure to violence into three bins: Low, Moderate, and High. The greater the cumulative exposure of participants to violence the greater the incidence of somatic complaints, depressive symptoms, anxiety,

aggressive and delinquent behavior, and academic failure. The authors conclude that exposure to violence significantly disrupts adolescents' development.

# PROPOSED WORK

**1) Data Preprocessing.** As with any data mining project, ours will begin by preparing the data itself. The preprocessing phase can be broken down into the 3 sections of Data Discovery, Data Cleaning, and Data Integration. Preprocessing addresses and improves the factors of *data quality.*

**a) Data Discovery**. Initial exploration of the datasets specifically to review the 7 aspects of data quality: accuracy, completeness, consistency, timeliness, believability, and interpretability. Additionally, we will Identify common(crime) and comparative(health) datasets as well as identify shared attribute values between the common and comparative datasets. Crucially, we will also Identify attribute values for potential joining(neighborhood?).

**b) Data Cleaning**. Standard 2 part iterative cleaning process of discrepancy detection and data transformation. Specifically, we will identify and handle missing values with an appropriate technique(global constant/probable value etc) as well as smooth noisy data utilizing binning, regression, or outlier analysis where appropriate.

**c) Data Integration**. Combine like common(crime) datasets. Given the datasets are all from the same source and relevant to the same domain, they have many common attribute values that inherently help with the integration process. However, issues such as data and dimensionality reduction will still need to be addressed. Note that due to dissimilar attribute values, the comparative(health) datasets cannot be easily integrated with one another or with the common(crime) datasets without the application of more advanced mining techniques that will come later in the project.

**2) Initial Analysis.** Once the data has been preprocessed, we will begin by a more thorough standard examination of the data to gain a better understanding of interesting questions and relevant attribute values. This phase will allow us to ensure the quality of the neighborhood level analysis in the next phase.

**a) Common(crime) Datasets**. Using Python and Tableau, we will answer general questions about the crime dataset such as: what is the universe of crimes? What is the frequency of these different crimes? Which of these crimes constitute violent crimes? During this process we will take note of any information that may provide insight into building the neighborhood level analysis.

**b) Comparative(health) Datasets**. Using Python and Tableau, we will answer general questions about the health dataset such as: What is the age life expectancy in Denver? What percentage of the population in Denver is obese? During this process we will take note of any information that may provide insight into building the neighborhood level analysis.

**c) Reevaluation**. Given the bases of 2.a and 2.b, if there are any interesting patterns or questions in the initial layer/level of analysis that warrant any additional analysis or preprocessing? At this point in time we will review all progress made thus far to ensure the data is adequately prepared to provide the knowledge we are seeking to gain.

**3) Neighborhood Analysis.** After the completion of the Initial Analysis, we will begin the neighborhood level analysis of interesting attributes and correlations discovered in the previous phases. This Neighborhood Analysis will be the base of comparison between the common(crime) and comparative(health) datasets. In this phase, we will create a summarizing dataset that allows us to consolidate our findings from Phase 2. Using this summarizing dataset, we will make arguments about correlations between crime and health.

**a) Common(crime) Datasets**. Using Python and Tableau, we will answer specific questions about crime at the neighborhood level such as: What is the frequency of the different crimes in different neighborhoods? Do particular neighborhoods suffer disproportionately from certain kinds of crime?

**b) Comparative(health) Datasets**. Using Python and Tableau, we will answer specific

questions about health at the neighborhood level such as: Which neighborhoods have the most obese population? Which neighborhoods have the least obese population? What neighborhoods have the highest or lowest life expectancy?

**c) Neighborhood Review & Integration**. Upon completion of the first 2 subsections of the Neighborhood Analysis phase, we will audit the analyses to ensure their completeness and accuracy. Then we will combine all of the neighborhood level analyses into a summary dataset. Using this summary dataset we will begin to answer questions that combine our conclusions related to crime and health at the neighborhood level such as: Do neighborhoods with large amounts of violent crime have worse health outcomes than neighborhoods with less violent crime?

**4) Summary Analysis.** Wrap things up (This section depends on progress in prior phases).

**a) Summary Neighborhood Analysis** Initial summary analysis between the common and comparative datasets using Python to explore potential interesting questions.

**b) Visualizations**. With the tools of Tableau, Seaborn, Kibana, Matlplotlib, Pandas and others we will create effective visualization idioms to report our findings.

**c) Predictive Tool**. As many of the standard techniques we will be applying have a predictive nature upon their completion, it may be feasible to create a python based predictive tool for user functionality and the potential processing of dynamic data.

**5) Report & Presentation**. With our findings in hand we will compile them, and our process itself, into our final report, and prepare our presentation.

## DATASETS

For our datasets we used the **City of Denver's Open Data Catalog [4]**. The specific datasets we intend to use are:

• **Crime in the City of Denver**. Criminal offenses in the City and County of Denver, CO for the current year to date as well as for the last five calendar years. Data attributes include a brief description of the criminal offense, the category of the offense, as well as the date and neighborhood where the offense occurred.

https://www.denvergov.org/opendata/dataset/city-and-county-of-denver-crime

• **Hate Crimes in the City of Denver**. Hate crime offenses in the City of Denver, CO for ranging from Jan 2010 to Jan 2021. Dataset explores criminal offenses that target an individual(s) based on the offender's preconception of which group(s) the victim belongs to. Data attributes include a date, time month of year and neighborhood, as well as an offense description and bias type involved in the offense.

https://www.denvergov.org/opendata/dataset/hate-crimes

• **Traffic Accidents in the City of Denver**. Traffic Accidents in the City of Denver, CO for the previous five calendar years plus the current year to date. Relevant attributes include the specific incident address and neighbourhood associated with the accident.

https://www.denvergov.org/opendata/dataset/city-and-county-of-denver-traffic-accidents

• **Police Pedestrian Stops and Vehicle Stops in the City of Denver**. Police pedestrian stops and vehicle stops in the City of Denver, CO for the last four calendar years and the current year to date. Data attributes include an address and neighborhood name, as well as a problem attribute for reason of stop. Data has been cleaned by excluding data without a valid address.

https://www.denvergov.org/opendata/dataset/city-and-county-of-denver-police-pedestrian-stops-and-vehicle-stops

• **American Community Survey in the City of Denver**. Neighborhood level data in the City and County of Denver, CO for a 5 year average, years include 2013 - 2017. Data attributes include a neighborhood name, varying attributes on race, age and education levels as well as varying attributes on poverty levels regionally.

https://www.denvergov.org/opendata/dataset/city-and-county-of-denver-american-community-survey-nbrhd-2013-2017

• **Life Expectancy 2010 - 2015 in the City of Denver**. Average life expectancy in 78 distinct sections of Denver based on U.S. census data

from 2010-2015. Relevant attributes include the Denver area name, life expectancy in years, and Federal Information Processing Standards (FIPS) codes which precisely identify the census tract referred to by the area name.

https://www.denvergov.org/opendata/dataset/city-and-county-of-denver-life-expectancy-2010-2015

• **Adult Obesity 2014-2016 in the City of Denver**. Estimated numbers of obese individuals older than two by neighborhood in Denver, CO. These estimates were arrived at by looking at individuals who sought care at health care institutions participating in the Colorado Department of Public Health and Environment's BMI Monitoring System between 2014 and 2016. As part of their routine care, individuals had their height and weight measured and their BMI calculated from these measurements. Obese individuals in the study are defined as having a BMI of 30 kilograms per meter squared (kg/m2) or greater. Findings are generalized to neighborhood-wide estimates by dividing the total number of individuals in the BMI Monitoring System in a given neighborhood with the total estimated population of that neighborhood as estimated by census data. Relevant attributes include neighborhood, total population in BMI registry, percent obese, number of obese adults, and confidence interval.

https://www.denvergov.org/opendata/dataset/city-and-county-of-denver-adult-obesity-2014-2016

• **Real Property Sales and Transfers in the City of Denver**. Transfers of property ownership via sales or other means from 2010 to the present as recorded by the City and County of Denver Assessment Division. Relevant attributes include sale year, sale price, real estate type, and neighborhood.

https://www.denvergov.org/opendata/dataset/city-and-county-of-denver-real-property-sales-and-transfers

## EVALUATION METHODS

We seek to understand correlations between geography, socioeconomic indicators, crimes, and health outcomes. Initially, we will likely apply correlation analysis to our data. For nominal data we will use the chi-square test. For numeric data we will use the correlation coefficient and covariance. These measurements will help us to understand whether different attributes are dependent or independent variables. We also anticipate developing association rules that describe the relationship between correlated attributes. We will therefore likely rely on measuring support and confidence to evaluate the interestingness of our proposed rules. Finally, naive Bayesian analysis could help us to predict the likelihood of certain crimes and health outcomes occurring in specific neighborhoods. here.

## TOOLS

• **Python.** Python is the base programming language we will use to perform most of our analysis techniques. Many of the other libraries and platforms we will utilize are built on or around python.

• **NumPy.** NumPy is a library used for the python programming language. We will use NumPy as support to explore our large, multidimensional datasets by performing various mathematical operations such as creating a subset array of our data for manipulation, reshaping and combining to explore our results.

• **Pandas.** Pandas is a library used for the python programming language. We will use Pandas to perform data manipulation and analysis on our large, multidimensional datasets. By utilizing Pandas, it will allow us to easily handle missing values within our dataset, and flexibly slice, merge or reshape our data to discover interesting results.

• **matplotlib.** matplotlib is a plotting library for the python programming language, which will employ NumPy for its mathematical extension. We will use matplotlib to explore our large, multidimensional datasets by allowing us to create easily digestible visuals from our dataset. Through these visualizations we will be able to interpret and answer interesting questions that arise throughout our exploration.

• **GitHub.** GitHub is a web-based interface that utilizes Git, which allows for version control and open source code. We will use GitHub to make real-time separate edits, updates and uploads of our source code to collaboratively work on our project as a team.

• **Trello.** Trello is a web-based tool that enables collaboration between teammates to organize a project into storyboards. Within each board a team member will be able to tell what tasks are in process of or have been completed, who is working on or finished which tasks, and milestones that still need to be met.

• **Tableau.** Tableau is an interactive data visualization software. We will use Tableau to explore our large, multidimensional datasets by allowing us to create easily digestible visuals from our dataset. Through these visualizations we will be able to interpret and answer interesting questions that arise throughout our exploration.

## MILESTONES COMPLETED

• **Phase 1.a -** Monday, October 18, 2021. *Data Discovery*: Initial exploration of the datasets specifically to review the 7 aspects of data quality.

• **Phase 1.b** - Monday, October 25, 2021. *Data Cleaning*: Standard 2 part iterative cleaning process of discrepancy detection and data transformation.

• **Project Part 2** - Monday, October 25, 2021. *Project Proposal Paper*.

• **Phase 1.c** - Monday, November 1, 2021. *Data Integration:* Combine like common(crime) datasets. Address data dimensionality and reduction.

**Phase 1 - Data Preprocessing Completed**

• **Phase 2.ab** - Monday, November 8, 2021. *Initial analysis of common(crime) and comparative(health) datasets:* Ask questions such as what are the most common crimes?

• **Phase 2.c** - Wednesday, November 24, 2021. *Reevaluation*: Given the bases of 2.a/b, are there any interesting patterns or questions in the initial layer of analysis that warrant any additional analysis or preprocessing?

**Phase 2 - Initial Analysis Completed**

## MILESTONES TO BE COMPLETED

• **Phases 3.ab** - Monday, November 29, 2021. *Neighborhood analysis of common(crime) and comparative(health) datasets:* Explore a neighborhood level analysis of interesting attributes/correlations. This Neighborhood Analysis will be the base of comparison between the common(crime) and comparative(health) datasets.

• **Phases 3.c** - Thursday, December 2, 2021. *Neighborhood Review & Integration:* Given the bases of 3.a/b, we will audit the analyses to ensure their completeness and accuracy.

• **Project Part 3** - Thursday, December 2, 2021. Progress Report.

**Phase 3 - Neighborhood Analysis Completed**

• **Phase 4.abc** - Monday, December 6, 2021. *Summary Analysis.*

**Phase 4 - Summary Analysis Completed**

• **Final Project** - Friday, December 10, 2021. Final project due; presentation.

**Phase 5 - Project Completed**

## RESULTS

**See Appendix**

• ***Figure 1a -*** From the outset, we knew that the primary attribute in these datasets that we would join on was going to be neighborhoods. So our first priority was to standardize neighborhood names across our datasets. We did this with a simple program that detected differences in neighborhood values. Issues that we encountered in this process were differences in capitalization and punctuation, and the presence of anomalous NaN values. We resolved these differences to arrive at a single set of 77 standard neighborhoods.

• ***Figure 1b -*** Because the crime dataset is our largest and “messiest” dataset, we decided to reduce this dataset into a simple, clean data warehouse. To do so, we created a program with Python, Pandas, and Numpy that counted the number of crimes by category in each neighborhood. This provides us with a basis for

a data warehouse which will allow us to compare summarized health data with counts of crimes committed by neighborhood in Denver.

• ***Figure 2*** - To explore the data trivally, we wanted to gain information at the Offense Type level for analysis. We examined the different Offense Types that are within our data by counting the total number of each crime based on type. With the count of each Offense Type summarized for our crime dataset, we were able to show what the most common Offense Types were. Using size to encode our count allows the user to easily visualize the most common Offense Type and gives them an opportunity to compare two Offense Types by bubble sizing.

• ***Figures 3a & 3b*** - To obtain primary demographic information at the neighborhood level of analysis, we examined the Neighborhood Values Survey. With the neighborhood on the x axis, the remaining attribute values were grouped into relevant categories, to include sex, race, age, education, income, etc. For example, see figures 3a above and 3b below, which include the general summary and sex tabs. The summary tab again encodes neighborhoods to hue and x axis, with total population, median age, median family income, median home value, and percentage of poverty. Note that several attribute values were created for this visualization using existing attribute values. For example, the percentage male/female was created utilizing the counts of total population, and male/female population. Several additional derived attributes values will be created for the finalized neighborhood level analyses.

• ***Figure 4*** - To obtain primary Offense Type information at the neighborhood level of analysis, we examined the Crime dataset. With the neighborhood encoded by color and each having their own individual tree we can explore the most common Offense Type within that neighborhood. Utilizing the total count of an Offense Type broken down on the neighborhood level allows the user to explore not just most common crimes across neighborhoods, but also shows which neighborhoods have the most crime associated with them.

• ***Figure 5a*** - To obtain primary geographical information at the neighborhood level of analysis, we examined the Crime dataset. We created a generic geo map, where each neighborhood is encoded by color and the individual circles represent a specific type of Offense within that neighborhood. The geo maps goal is to give a user an idea of how the Denver neighborhoods are positioned, i.e. what neighborhoods are next to each other, do they border another county, and gives an overview of Offense Types within neighborhoods.

• ***Figure 5b*** - An extension of Figure 5a, we specifically filtered the geo map on the Offense Type level to obtain all Arson Offense Types separated by neighborhood level. The goal is to allow a user to explore the crime dataset geographical, where the user can filter based on the neighborhood level, the Offense Type and view the incident address. Note: Neighborhood ID is encoded based on color.

• ***Figure 6*** - To obtain primary health information at the neighborhood level of analysis, we examined the adult obesity 2014-2016 in the City of Denver dataset. With the neighborhood plotted on the y axis, they were plotted based on the percentage of obesity within the neighborhood. The size of the dot was encoded based on the sum of the count of how many people were obese within each neighborhood, this helps make the user aware of a potential skew in the dataset.

• ***Figure 7*** - To obtain primary Offense Type information at the Offense Category ID level of analysis, we examined the Crime dataset. We split on the Offense Category first to provide a general breakdown for the user to analyze the data on. The Offense Category ID is encoded by hue. After this we were able to split on the Offense Type next. This summarizes the Offense Category based on Offense Type. For example, from the visual we can easily see that traffic-accident is the most common crime. We are also able to see within the drug-alcohol Offense Category ID the second most common Offense Type is liquor-possession with a count of 4,900.

## ACKNOWLEDGMENTS

This project was completed as part of the Fall 2021 course CSBP 4502 Data Mining at the

University of Colorado at Boulder. The course instructor was Kristy Peterson and the textbook used was ***Data Mining: Concepts and Techniques*** **[5]**. The terminology and methods of analysis used in this project come from this textbook.

## REFERENCES

[1] City and County of Denver, Department of Public Safety. 2021. Denver Opportunity Index. https://geospatialdenver.maps.arcgis.com/apps/MapSeries/index.html?appid=ff9da000f1a344beb341ce839a720021

[2] Chase Sackett. 2016. Neighborhoods and Violent Crime. In *Evidence Matters* (Summer 2016). Office of Policy Development and Research, Washington, D.C., USA, 13 pages. https://www.huduser.gov/portal/periodicals/em/summer16/highlight2.html

[3] Gayla Margolin, PhD, Katrina A. Vickerman, MA, Pamella H. Oliver, PhD, and Elana B. Gordis, PhD. 2010. Violence Exposure in Multiple Interpersonal Domains: Cumulative and Differential Effects. In J Adolesc Health. PMC 47, 2 (Aug, 2010), 198–205. DOI: 10.1016/j.jadohealth.2010.01.020

[4] City and County of Denver. 2015. City of Denver Open Data Catalog. http://data.denvergov.org Licensed under the Creative Commons Attribution 3.0 license (CC BY 3.0) http://creativecommons.org/licenses/by/3.0

[5] Jiawei Han, Micheline Kamber, and Jian Pei. 2011. *Data Mining: Concepts and Techniques* (3rd ed.). Morgan Kaufmann, Burlington, MA.

# Appendix

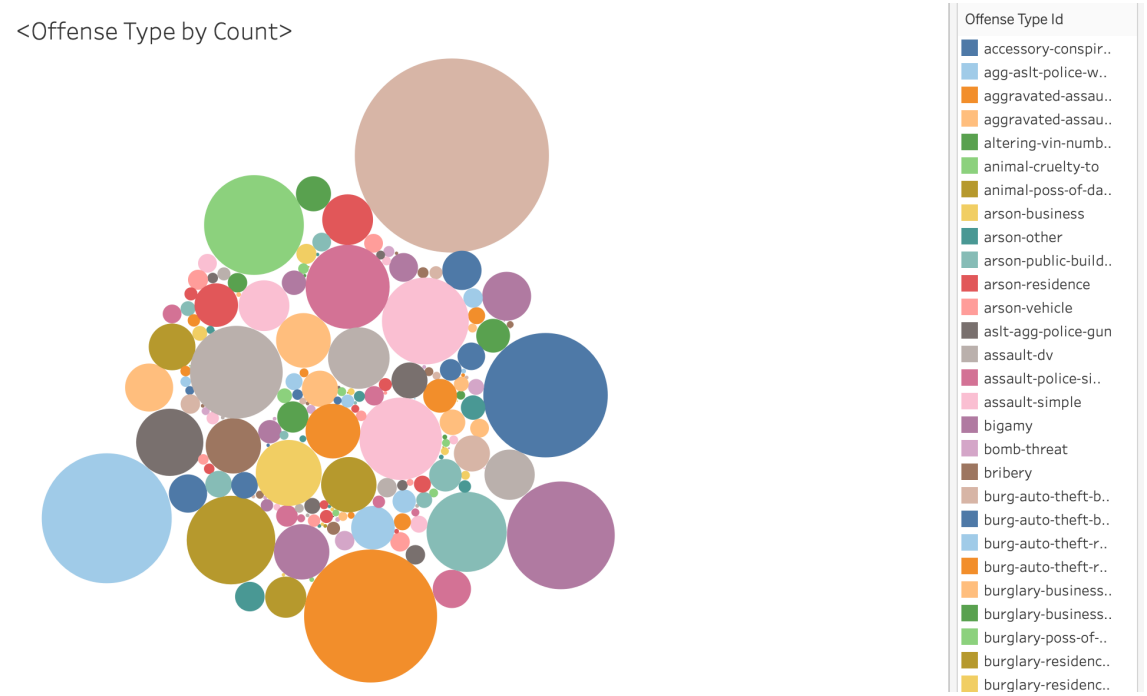
## Figure 1a -

```
differences = []
i = 0
for nbhd in formatted_crime_nbhds:
    if nbhd != obesity_nbhds[i]:
        differences.append(nbhd)
    if nbhd != lexp_nbhds[i]:
        differences.append(nbhd)
    i += 1
print(differences)
```

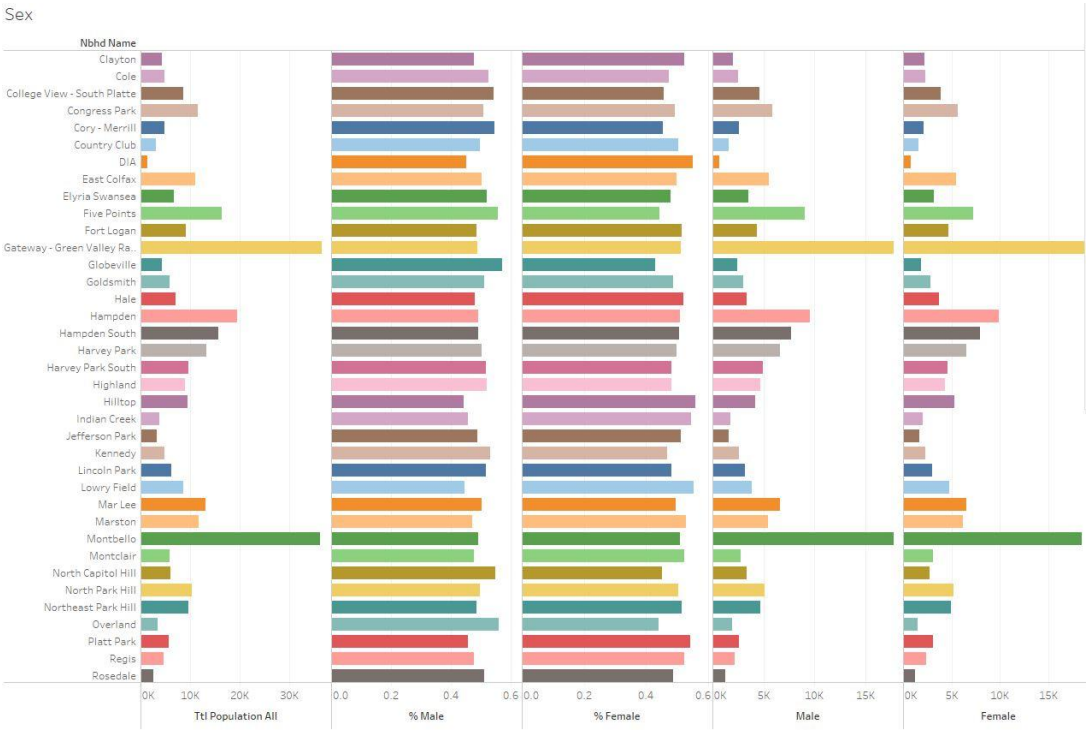
## Figure 1b -

```
# initialize list of lists
warehouse = ""
row = []

# Check length of neighborhood lists
#     to be used in indexing the following 'for' loop
assert( len(crime_nbhds) == len(formatted_crime_nbhds) )

#
# Create list of lists corresponding to desired DataFrame
#
for i in range(len(crime_nbhds)):
    row.clear()
    row.append(formatted_crime_nbhds[i])
    nbhd = crime_nbhds[i]
    for j in range(len(crimes)):
        numInstances = crime_data.query(f'NEIGHBORHOOD_ID == "{nbhd}" a
        assert(numInstances[5]==numInstances[16])
        row.append(numInstances[5])
    warehouse += f"{row}, "

print(warehouse)
```

# Figure 2 -

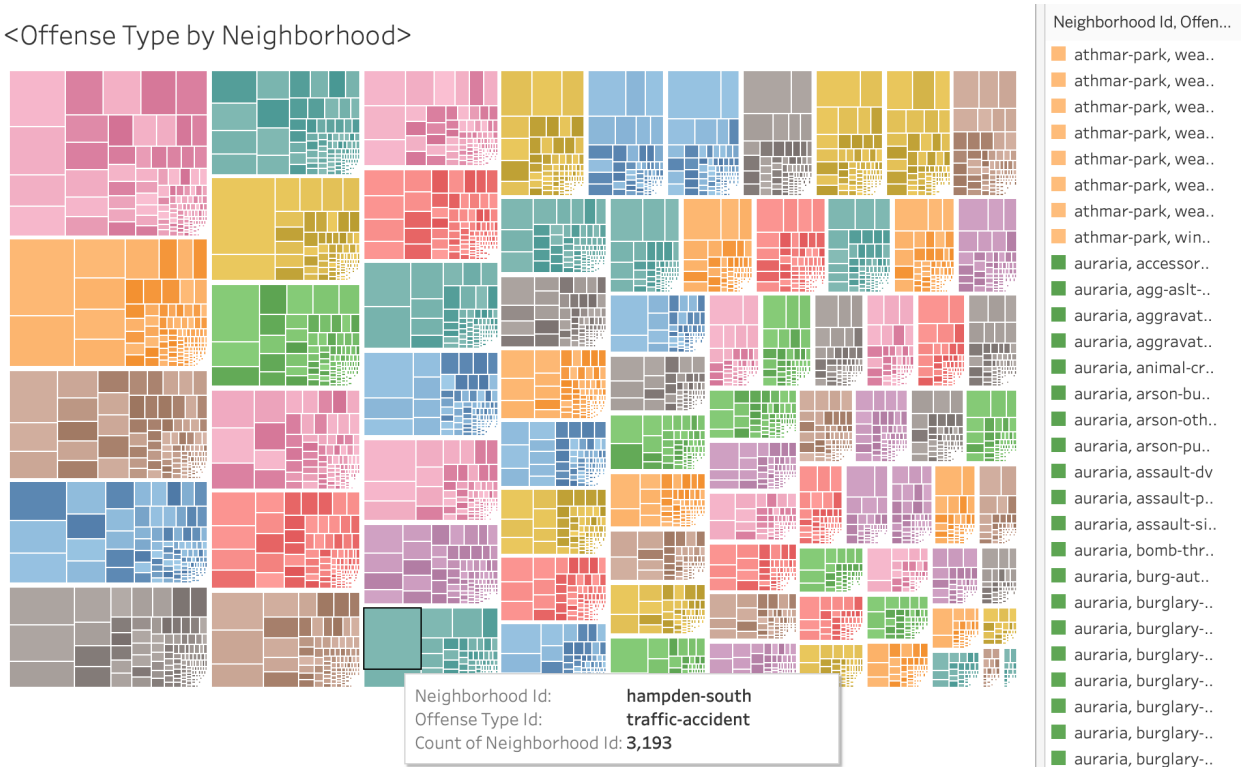

<Offense Type by Count>
Offense Type Id
accessory-conspir..
agg-aslt-police-w..
aggravated-assau..
aggravated-assau..
altering-vin-numb..
animal-cruelty-to
animal-poss-of-da..
arson-business
arson-other
arson-public-build..
arson-residence
arson-vehicle
aslt-agg-police-gun
assault-dv
assault-police-si..
assault-simple
bigamy
bomb-threat
bribery
burg-auto-theft-b..
burg-auto-theft-b..
burg-auto-theft-r..
burg-auto-theft-r..
burglary-business..
burglary-business..
burglary-poss-of-..
burglary-residenc..
burglary-residenc..

# Figure 3a -

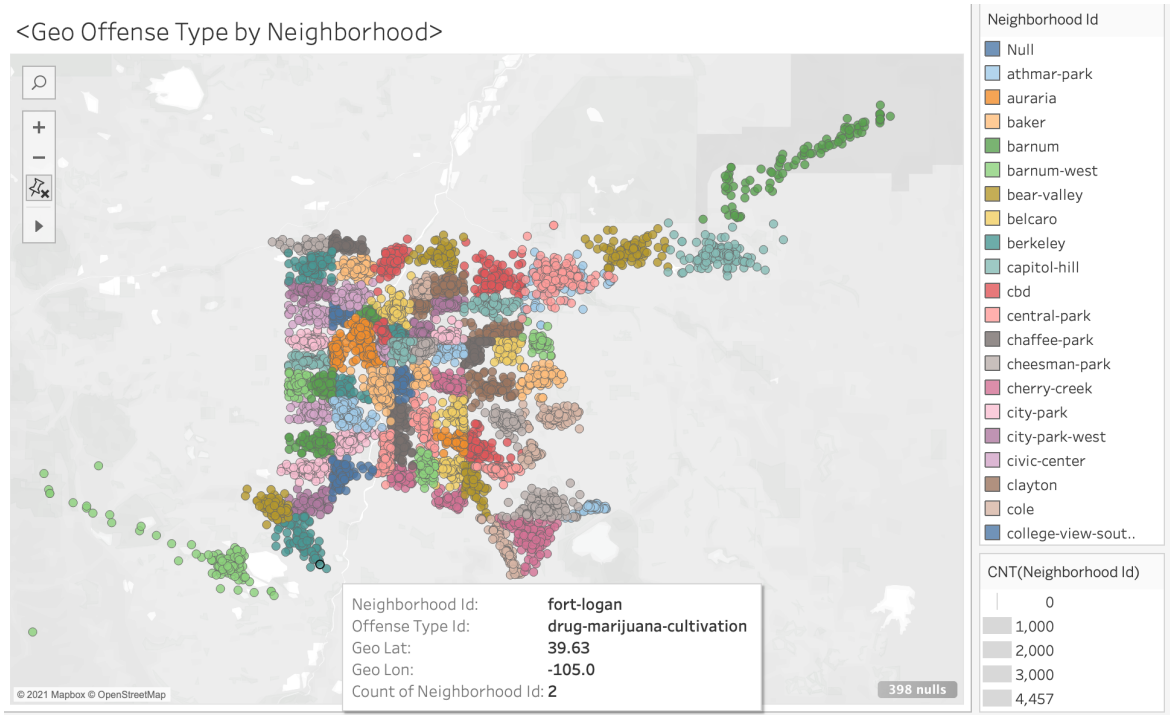



# Figure 3b -

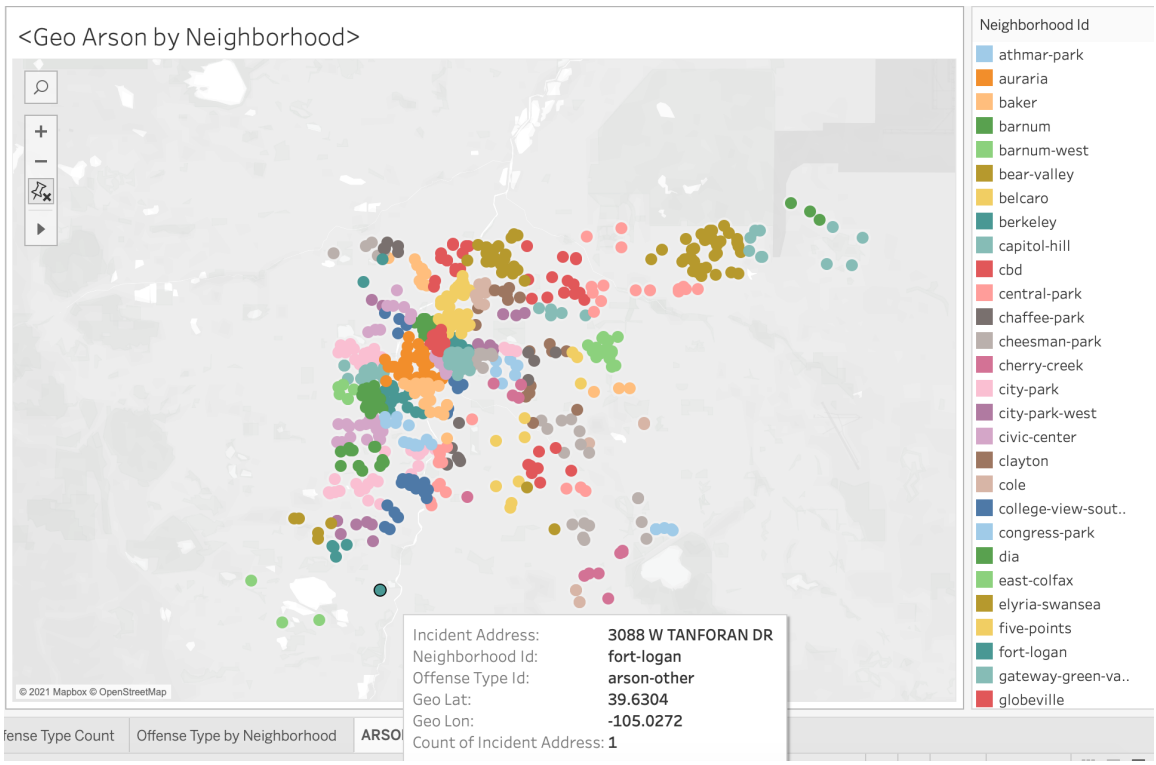
# Figure 4 -


<Offense Type by Neighborhood>
Neighborhood Id, Offen...
athmar-park, wea..
athmar-park, wea..
athmar-park, wea..
athmar-park, wea..
athmar-park, wea..
athmar-park, wea..
athmar-park, wea..
athmar-park, win..
auraria, accessor..
auraria, agg-aslt-..
auraria, aggravat..
auraria, aggravat..
auraria, animal-cr..
auraria, arson-bu..
auraria, arson-oth..
auraria, arson-pu..
auraria, assault-dv
auraria, assault-p..
auraria, assault-si..
auraria, bomb-thr..
auraria, burg-aut..
auraria, burglary-..
auraria, burglary-..
auraria, burglary-..
auraria, burglary-..
auraria, burglary-..
auraria, burglary-..
auraria, burglary-..
Neighborhood Id: hampden-south
Offense Type Id: traffic-accident
Count of Neighborhood Id: 3,193

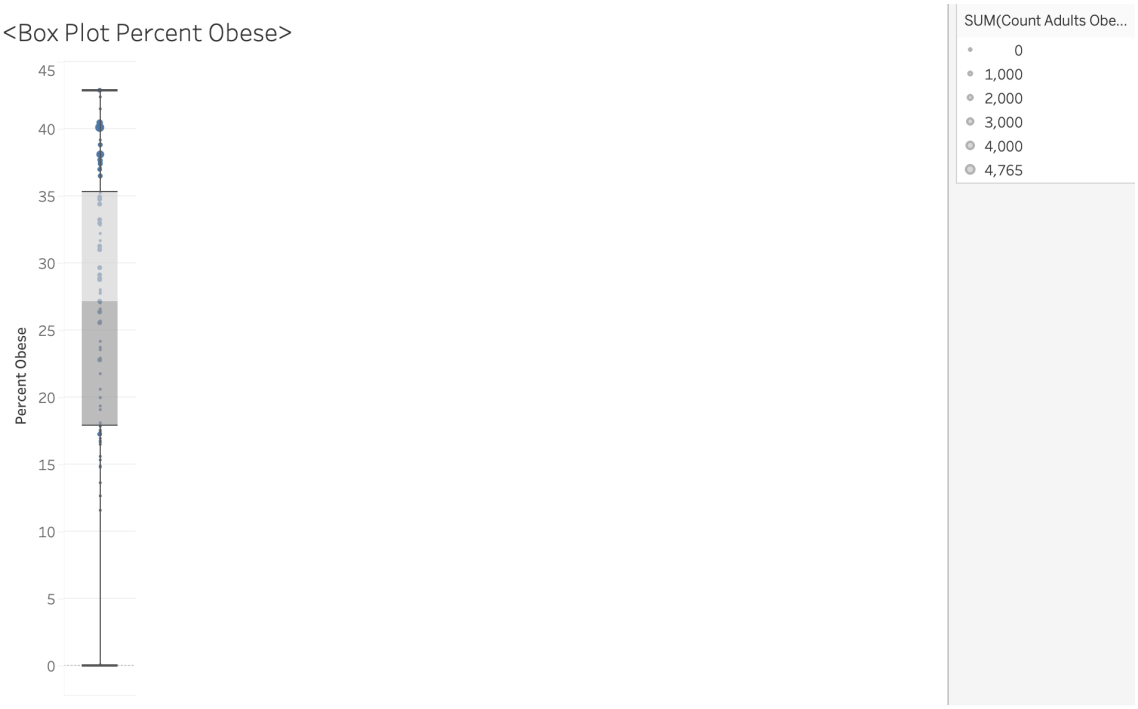
# Figure 5a -




# Figure 5b -



# Figure 6 -




# Figure 7 -