EPA 1315
Data Analytics &
Visualization

Final Project
Crime & Public Infrastructure: The Case of Five Major US Cities

Caroline Cullinan
5135737

Leonardo Nicoletti
5138345

November 8, 2019

Contents

1.0 Problem Understanding 3
1.1 Background Information 3
1.2 Organization Objectives 5
1.3 Organization Criteria for Success 5
1.4 Resource Inventory 5
1.5 Requirements, Assumptions, & Constraints 5
1.6 Risks & Contingencies 6
1.7 Terminology 6
1.8 Cost Benefit Analysis 6
1.9 Data Mining Goals 6
1.10 Data Mining Criteria for Success 6
1.11 Project Plan 6
2.0 Data Understanding 7
2.1 Initial Data Collection 7
2.2 Data Description 8
2.3 Data Quality 8
2.4 Exploratory Analysis 9
3.0 Data Preparation 10
3.1 Dataset and Dataset Description 11
3.2 Data Visualization 11
4.0 Modeling 22
4.1 Test Design 24
4.2 Models 24
4.3 Parameter Setting 25
4.4 Model Description 25
4.5 Assessment 26
5.0 Limitations & Recommendations 29
6.0 Acknowledgement of Tasks 30
7.0 References 31
Appendix A 33
Appendix B 34
Appendix C 35

Appendix D 36

Appendix E 43

Appendix F ..... **Error! Bookmark not defined.**

Appendix G ..... **Error! Bookmark not defined.**

Appendix H 47

1.0 Problem Understanding

1.1 Background Information

Internationally, crime devastates countries. Increased crime results in harmful effects for legal economies (Detotto & Otranto, 2010). Investments from foreign and domestic sources are discouraged, and firms become less competitive as a result of crime (Detotto & Otranto, 2010). Crime also threatens the personal security of a country's citizens, thus negatively impacting the overall health and well-being of individuals (Morrall, Marshall, & Macdonald, 2010). Those victimised by crime experience mental-health disorders such as insomnia, depression, and anxiety at increased rates (Morrall, Marshall, & Macdonald, 2010). Therefore, the negative health effects associated with crime can contribute to social disengagement, thus weakening overall social cohesion of communities (Morrall, Marshall, & Macdonald, 2010).

In order to address perpetuators of crime, Wilson & Kelling proposed the "broken window theory". Such theory disputes that regardless of neighborhood, neglected property encourages further neglect and destruction of property (Wilson & Kelling, 1982). Such theory insists that increased incidents of crime are associated with increased instances of community neglect (Wilson & Kelling, 1982). Further argued, the "broken window theory" suggests that minor forms of neglect, such as garbage-littered streets and broken street lamps, have the greatest contributions to criminal activity, for the existence of disorder may allude that the actions of individuals' go unaccounted for (Ceraiti, 2015).

Many cities across the US have based their policing practices off the work of Willson & Kelling (Andersson, 2015; Harcourt, 2009). Law enforcement efforts have been focused on instances of "disorder", resulting in unwarranted police stops and searches in communities deemed "disorderly" in hopes of mitigating crime (Murakawa & Beckett, 2010). While the effectiveness of such policing practices has been questioned, the validity of the overarching theory in which such policies have been based upon deserves further scrutiny. Are aspects of neighborhood "cleanliness" and "order" true indicators of criminal activity?

In order to assist the Organization for Economic Cooperation and Development (OECD) in enhancing world progress, the relationship between crime and elements of the urban fabric that contribute to the "cleanliness" and "care" of streets are assessed across five United States (US) cities; San Francisco, Los Angeles, Boston, Philadelphia, and Washington D.C.. Such analysis examines if trash can availability, street light presence, and the existence of trees in urban communities relate to crime. With the use of a predictive model, such analysis also examines if the existence of such elements (trash cans, street lights, and trees) can predict instances of crime.

With overarching objectives to address grand societal challenges, the OECD has recently launched the "Global Hub on the Governance for the SDGs" (OECD, 2019c). With motivation to advance the 2030 Agenda for Sustainable Development, Goal 9: "Build a resilient infrastructure, promote inclusive and sustainable industrialisation and foster innovation", and Goal 16: "Promote peace and inclusive societies for sustainable development, provide access to justice for all, and build effective, accountable, and inclusive institutions at all levels" are goals of particular interest when examining crime and infrastructure (OECD, 2019c; United Nations, 2019a; United Nations, 2019b). Within the OECD, the Directorate for Employment, Labour and Social Affairs and the Directorate for Public Governance may be further interested in this study. Seeing as crime is a social affair which affects social welfare, the Directorate for Employment, Labour and Social Affairs may be very interested in obtaining further information on crime correlations in order to better address instances of crime through policy (OECD, 2019b). Concerned with budgeting and public expenditures, the Directorate for Public Governance may

also have increased interest in this study, for crime-addressing policies that may influence infrastructure spending could greatly concern the work of this directorate (OECD, 2019a).

In the past, the OECD has incorporated data mining activity to address crime and safety in OECD member states. However, such activity has involved panel data, which illustrates changes over multiple countries in multiple years. Such analysis has also been a-causal in nature. This study, instead, aims to address underlying factors which may perpetuate trends in crime. The intent of this report is to act as a baseline of information for the OECD. With insight from the report, the OECD can then conduct further analysis and implement effective policy-making relative to crime.

The US has been chosen as the location of interest for this study, for a multitude of accurate data sources are available on the US. While results from this study are specific to the US, further analysis and model development could aid the OECD in scaling this study to address crime elsewhere.

For further illustration of the urgency and geographic boundary of the problem at hand, please see Figures 1 which maps crime rates within the contiguous US.

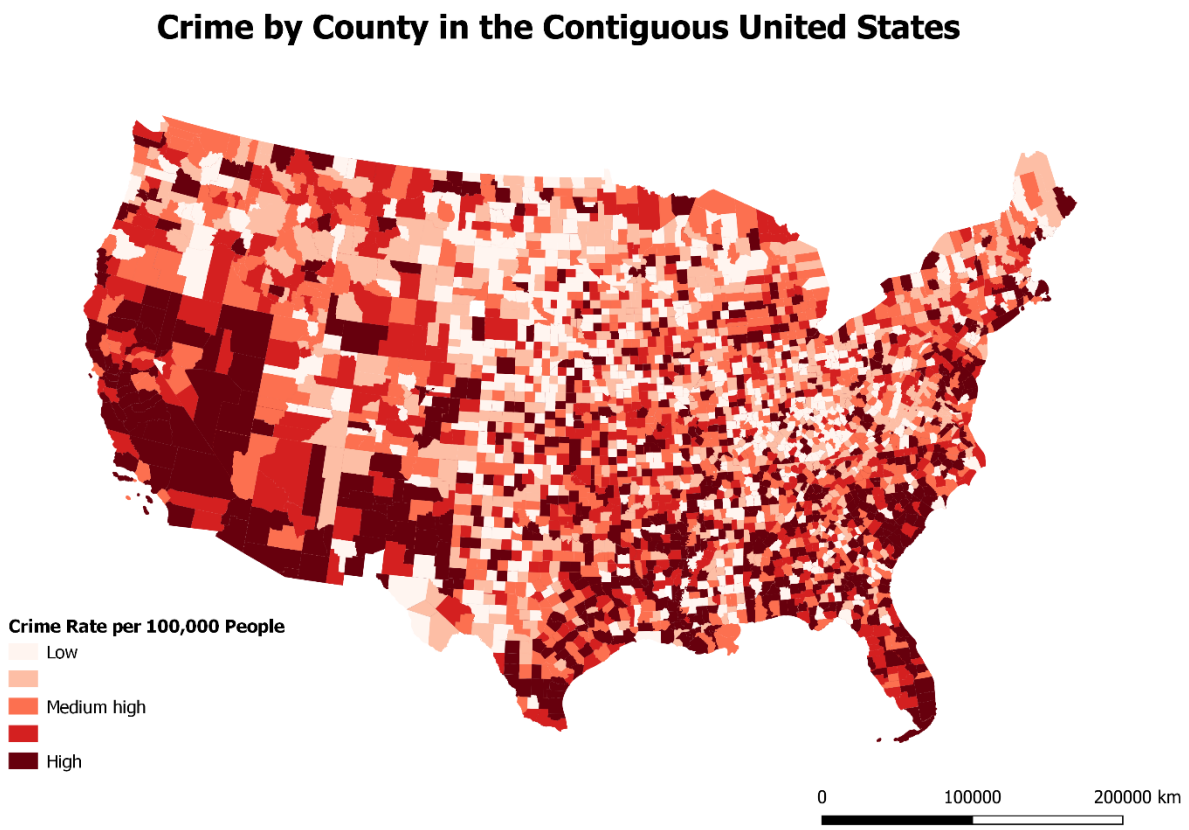

Figure 1: Crime rate per county in the Contiguous United States

1.2 Organization Objectives

The OECD strives to address a multitude of environmental, economic, and social issues through evidence-based solutions for policy (OECD, 2019d). Activity at the OECD aims to confront grand societal challenges. Specifically, the OECD has interest in the United Nations Sustainable Development Goals. With greater insight on perpetuators of crime, the OECD has the objective of appropriately addressing crime with effective policy. In doing so, the OECD can better address Goal 9 and Goal 16 of the 2030 Agenda for Sustainable Development.

1.3 Organization Criteria for Success

In order for this study to be successful, the OECD requires that this analysis addresses the following concerns:

- The study must address a grand societal challenges, per the OECD missions and objectives.
- The study must draw additional attention from various directorates.
- The study must be evidence-based and involve strategic foresight.
- Potential strategies for intervention should be possible given this analysis.
- Fundamental uncertainties should be addressed within this analysis.

1.4 Resource Inventory

A list of resource inventory is provided in Figure 2.

Table 1: List of resource inventory necessary

Resource Inventory		
Hardware & Software	**Data Sources**	**Personnel**
• Internet connection • Computers • Anaconda Navigator • RStudio • JupyterLab • QGIS • Adobe InDesign	• World Bank Open Data Repository • OECD Dataset • DataSF • Los Angeles Open Data • Los Angeles GeoHub • Open Data Philly • City of Philadelphia Metadata Catalog • Open Data DC • Analyze Boston	• Two analytical staff members

1.5 Requirements, Assumptions, & Constraints

Requirements:

- This analysis is scheduled for submission to the Counsellor for Strategic Foresight of the OECD by November 8, 2019.
- In order to aid in the interpretation of this analysis, a written report, and effective graphics and visualisations must be submitted to the Counsellor for Strategic Foresight.
- This analysis must consist of a graphical model and Bayesian analysis.

Assumptions:

- The quality of this analysis is heavily dependent on the quality of available data.
- The results of this analysis will be interpreted and expanded upon by the OECD for effective implementation of crime-addressing policy.

[1] Please note that all figures presented in this report have been created by team members Leonardo Nicoletti and Caroline Cullinan

Constraints:

- Data inventory for relevant factors may not be available.
- Analytical staff members may not have administrative access to all relevant databases and datasets.

1.6 Risks & Contingencies

For an outline of relevant risks and contingencies associated with this analysis, please see Table 2.

Table 2: Project risks and contingencies

Risks and Contingencies		
Risk Type	**Nature of Risk**	**Contingency**
Scheduling	Project activities require more time than expected for completion.	The project is not completed by the mandatory deadline of November 8, 2019.
Financial	Project activities require additional funds.	Allocated analysis funding cannot completely cover cost of this study.
Data	Necessary data is insufficient or unavailable.	The resulting analysis will be of poor quality, thus rendering it useless to the OECD.
Results	Reverberation from the United States as a result of crime and safety portrayal within the country may occur.	The United States may be less likely to willingly cooperate with the OECD in the future.

1.7 Terminology

For clarification of important language, please see a glossary of technical terms in Appendix A.

1.8 Cost Benefit Analysis

In addition to the personal team-based costs associated with this analysis (i.e. time commitments, cost of living, etc.), costs for social media engagement, press reports, and report printing will also arise. However, with this report, benefits are also associated. Through quality analysis and clear documentation, this report could potentially aid the OECD in better addressing crime within the US with appropriate policy. After further development of this analysis, such study could then be scaled to address crime elsewhere, thus also benefiting other OECD member states.

1.9 Data Mining Goals

Goals associated with this data science project include the use of various open data sources, the use of various open software resources, demonstration of the possible relevance of graphical modeling, identification of further data for analysis, and identification of other credible hypotheses for investigation.

1.10 Data Mining Criteria for Success

For associated criteria for data mining success, please see Appendix B.

1.11 Project Plan

Please see Appendix C for an outlined project plan for this analysis.

2.0 Data Understanding

2.1 Initial Data Collection

In order to conduct this analysis, several relevant existing data sources have been considered and are outlined below in Table 3. All collected sources have been merged accordingly. Notably, this analysis does not utilize purchased data, for the amount of open data available for this project is sufficient.

Table 3: Data considered for initial data collection

Initial Data Collection				
Subject of Data	**Dataset Description**	**File Type**	**Dataset Source**	**Link to Dataset**
San Francisco Crime	Police Department Incident Reports from 2018 to present in San Francisco	Shapefile	DataSF	https://data.sfgov.org/Public-Safety/Police-Department-Incident-Reports-2018-to-Present/wg3w-h783
San Francisco Assets	The location of assets, such as trash cans, picnic benches, guard, rails, fences, street lights, etc. in San Francisco	Shapefile	DataSF	https://data.sfgov.org/Culture-and-Recreation/Assets/iyvt-kb7k
San Francisco Street Trees	A list of Department of Public Works maintained street trees in San Francisco	Shapefile	DataSF	https://data.sfgov.org/City-Infrastructure/Street-Tree-List/tkzw-k3nq/data
San Francisco Precinct Boundaries	Voting precincts in San Francisco based on census 2010 geography and as determined by the Department of Elections.	Shapefile	DataSF	https://data.sfgov.org/City-Management-and-Ethics/Election-Precincts-Current-Defined-2012/fhns-n8qp
Los Angeles Crime	Incidents of crime in Los Angeles from 2010 to present	CSV	Los Angeles Open Data	https://data.lacity.org/A-Safe-City/Crime-Data-from-2010-to-Present/63jg-8b9z
Los Angeles Street Lights	Location of street lights in Los Angeles	Shapefile	Los Angeles Open Data	https://data.lacity.org/api/geospatial/9ei6-svt8?method=export&format=Shapefile
Los Angeles Trash Cans	Location of all trash cans located on public sidewalks in Los Angeles	Shapefile	Los Angeles GeoHub	http://geohub.lacity.org/datasets/b998c1a838b4471cb486bb150b2684d9_0
Los Angeles Street Trees	Location of street trees in Los Angeles	Shapefile	Los Angeles GeoHub	http://geohub.lacity.org/datasets/trees-bureau-of-street-services?geometry=-120.303%2C33.600%2C-116.131%2C34.397
Los Angeles Precinct Boundaries	GIS shape/boundaries of voting precincts in Los Angeles	Shapefile	Los Angeles Open Data	https://data.lacity.org/A-Well-Run-City/Precinct-Boundaries/csp5-hary
Boston Crime	Crime incident reports as provided by the Boston Police Department from August 2015 to present in Boston	CSV	Analyze Boston	https://data.boston.gov/dataset/crime-incident-reports-august-2015-to-date-source-new-system
Boston Street Lights	Dataset with over 74,000 city street lights in Boston	Shapefile	Analyze Boston	https://data.boston.gov/dataset/streetlight-locations
Boston Trash Cans	Location of all Big Belly trash bins located in Boston	CSV	Analyze Boston	https://data.boston.gov/dataset/big-belly-locations
Boston Street Trees	Location of street trees in Boston	Shapefile	Analyze Boston	https://data.boston.gov/dataset/trees
Boston Precinct Boundaries	Voting precincts for Boston	Shapefile	Analyze Boston	https://data.boston.gov/dataset/precincts
Philadelphia Crime	Incidents of crime in Philadelphia as provided by the Philadelphia Police Department	Shapefile	Open Data Philly	https://www.opendataphilly.org/dataset/crime-incidents
Philadelphia Street Lights	Location of street lights in Philadelphia	Shapefile	Open Data Philly	https://www.opendataphilly.org/dataset/street-poles
Philadelphia Trash Cans (Big Belly Trash Bins)	Location of all Big Belly brand trash cans in Philadelphia	Shapefile	Open Data Philly	https://www.opendataphilly.org/dataset/big-belly-waste-bins
Philadelphia Trash Cans (Wire Waste Baskets)	Location of non-Big Belly trash cans located in Philadelphia	Shapefile	Open Data Philly	https://www.opendataphilly.org/dataset/wire-waste-baskets
Philadelphia Street Trees	Location of Street Trees in Philadelphia	Shapefile	Open Data Philly	https://www.opendataphilly.org/dataset/philadelphia-street-tree-inventory
Philadelphia Precinct Boundaries	Ward precincts (which function as local election districts) based on U.S. Census Bureau blocks for Philadelphia	Shapefile	City of Philadelphia Metadata Catalog	https://data.cityofchicago.org/Facilities-Geographic-Boundaries/Precincts-current-/uvpq-qeeq
Washington D.C. Crime 2018	Location of incidents of crime in Washington D.C. 2018	CSV	Open Data DC	https://opendata.dc.gov/datasets/crime-incidents-in-2018
Washington D.C. Crime 2019	Location of incidents of crime in Washington D.C. in 2019	CSV	Open Data DC	https://opendata.dc.gov/datasets/crime-incidents-in-2019
Washington D.C. Street Lights	Locations and attributes of street lights in Washington D.C.	Shapefile	Open Data DC	https://opendata.dc.gov/datasets/street-lights?geometry=-77.530%2C38.713%2C-76.496%2C39.087
Washington D.C. Trash Cans	Location of trash cans maintained by the Department of Public Works in Washington D.C.	Shapefile	Open Data DC	https://opendata.dc.gov/datasets/litter-cans
Washington D.C. Street Trees	Location of street trees in Washington D.C.	Shapefile	Open Data DC	https://opendata.dc.gov/datasets/urban-forestry-street-trees
Washington D.C. Precinct Boundaries	Voting precinct boundaries for Washington D.C.	Shapefile	Open Data DC	https://opendata.dc.gov/datasets/44cdf2570d8545dd9c9868e2a2f570c8_27

2.2 Data Description

The data used in this analysis is described below in Table 4.

Table 4: Data description for data utilised for analysis

Data Description						
Subject of Data	Dataset Description	File Type	Amount of Data		Data Types	Conflicting Coding Scheme
			Rows	Columns		
San Francisco Crime	Police Department Incident Reports from 2018 to present in San Francisco	Shapefile	277,205	33	Floating Timestamp, Text, Number, Point	No
San Francisco Assets	The location of assets, such as trash cans, picnic benches, guard, rails, fences, street lights, etc. in San Francisco	Shapefile	5158	22	Floating Timestamp, Text, Number, Point	No
San Francisco Street Trees	A list of Department of Public Works maintained street trees in San Francisco	Shapefile	192,718	33	Floating Timestamp, Text, Number, Point	No
San Francisco Precinct Boundaries	Voting precincts in San Francisco based on census 2010 geography and as determined by the Department of Elections.	Shapefile	605	8	Text, Number, Polygon	No
Los Angeles Crime	Incidents of crime in Los Angeles from 2010 to present	CSV	2,070,686	28	Floating Timestamp, Text, Number, Point	No
Los Angeles Street Lights	Location of street lights in Los Angeles	Shapefile	216,694	12	Text, Point	No
Los Angeles Trash Cans	Location of all trash cans located on public sidewalks in Los Angeles	Shapefile	9,713	28	Floating Timestamp, Text, Point	No
Los Angeles Street Trees	Location of street trees in Los Angeles	Shapefile	635,558	6	Text, Number, Point	No
Los Angeles Precinct Boundaries	GIS shape/boundaries of voting precincts in Los Angeles	Shapefile	1,627	28	Text, Number, Polygon	No
Boston Crime	Crime incident reports as provided by the Boston Police Department from August 2015 to present in Boston	CSV	436,666	17	Floating Timestamp, Text, Number, Point	No
Boston Street Lights	Dataset with over 74,000 city street lights in Boston	Shapefile	74,065	5	Text, Number, Point	No
Boston Trash Cans	Location of all Big Belly trash bins located in Boston	CSV	543	2	Text, Point	No
Boston Street Trees	Location of street trees in Boston	Shapefile	201,845	3	Text, Point	No
Boston Precinct Boundaries	Voting precincts for Boston	Shapefile	255	7	Text, Number, Polygon	No
Philadelphia Crime	Incidents of crime in Philadelphia as provided by the Philadelphia Police Department	Shapefile	303,942	14	Floating Timestamp, Text, Number, Point	No
Philadelphia Street Lights	Location of street lights in Philadelphia	Shapefile	194,740	14	Text, Number, Point	No
Philadelphia Trash Cans (Big Belly Trash Bins)	Location of all Big Belly brand trash cans in Philadelphia	Shapefile	968	5	Text, Number, Point	No
Philadelphia Trash Cans (Wire Waste Baskets)	Location of non-Big Belly trash cans located in Philadelphia	Shapefile	352	4	Text, Number, Point	No
Philadelphia Street Trees	Location of Street Trees in Philadelphia	Shapefile	111,993	5	Number, Point	No
Philadelphia Precinct Boundaries	Ward precincts (which function as local election districts) based on U.S. Census Bureau blocks for Philadelphia	Shapefile	1,703	6	Number, Polygon	No
Washington D.C. Crime 2018	Location of incidents of crime in Washington D.C. 2018	CSV	33,783	24	Floating Timestamp, Text, Number, Point	No
Washington D.C. Crime 2019	Location of incidents of crime in Washington D.C. in 2019	CSV	28,497	24	Floating Timestamp, Text, Number, Point	No
Washington D.C. Street Lights	Locations and attributes of street lights in Washington D.C.	Shapefile	70,829	51	Floating Timestamp, Text, Number, Point	No
Washington D.C. Trash Cans	Location of trash cans maintained by the Department of Public Works in Washington D.C.	Shapefile	6,703	23	Floating Timestamp, Text, Number, Point	No
Washington D.C. Street Trees	Location of street trees in Washington D.C.	Shapefile	187,735	36	Floating Timestamp, Text, Number, Point	No
Washington D.C. Precinct Boundaries	Voting precinct boundaries for Washington D.C.	Shapefile	143	7	Number, Text, Polygon	No

2.3 Data Quality

The data utilised for this report contains characteristics relevant to this analysis. Key attributes per dataset are outlined below in Table 5. Summary statistics on datasets with key attributes have also

been conducted in order to gain greater insight on data quality and attribute relevancy. Please see Appendix D for summary statistics examples.

Table 5: Data quality with key attributes per dataset

Data Quality			
Subject of Data	**Dataset Description**	**File Type**	**Key Attribute**
San Francisco Crime	Police Department Incident Reports from 2018 to present in San Francisco	Shapefile	analysis_n
			geometry
San Francisco Assets	The location of assets, such as trash cans, picnic benches, guard, rails, fences, street lights, etc. in San Francisco	Shapefile	asset_type
			geometry
San Francisco Street Trees	A list of Department of Public Works maintained street trees in San Francisco	Shapefile	treeid
			xcoord
			ycoord
			geometry
San Francisco Precinct Boundaries	Voting precincts in San Francisco based on census 2010 geography and as determined by the Department of Elections.	Shapefile	neighrep
			prec_2012
			geometry
Los Angeles Crime	Incidents of crime in Los Angeles from 2010 to present	CSV	DATE OCC
			LON
			LAT
Los Angeles Street Lights	Location of street lights in Los Angeles	Shapefile	stlid
			geometry
Los Angeles Trash Cans	Location of all trash cans located on public sidewalks in Los Angeles	Shapefile	LAPDGrid
			geometry
Los Angeles Street Trees	Location of street trees in Los Angeles	Shapefile	TreeID
			geometry
Los Angeles Precinct Boundaries	GIS shape/boundaries of voting precincts in Los Angeles	Shapefile	cnsld
			geometry
Boston Crime	Crime incident reports as provided by the Boston Police Department from August 2015 to present in Boston	CSV	YEAR
			Lat
			Long
			INCIDENT NUMBER
Boston Street Lights	Dataset with over 74,000 city street lights in Boston	Shapefile	objectid
			geometry
Boston Trash Cans	Location of all Big Belly trash bins located in Boston	CSV	Location
			description
Boston Street Trees	Location of street trees in Boston	Shapefile	OBJECTID
			geometry
Boston Precinct Boundaries	Voting precincts for Boston	Shapefile	WARD_PRECI
			geometry
Philadelphia Crime	Incidents of crime in Philadelphia as provided by the Philadelphia Police Department	Shapefile	dispatch_d
			geometry
Philadelphia Street Lights	Location of street lights in Philadelphia	Shapefile	POLE_NUM
			geometry
Philadelphia Trash Cans (Big Belly Trash Bins)	Location of all Big Belly brand trash cans in Philadelphia	Shapefile	geometry
Philadelphia Trash Cans (Wire Waste Baskets)	Location of non-Big Belly trash cans located in Philadelphia	Shapefile	geometry
Philadelphia Street Trees	Location of Street Trees in Philadelphia	Shapefile	OBJECTID
			geometry
Philadelphia Precinct Boundaries	Ward precincts (which function as local election districts) based on U.S. Census Bureau blocks for Philadelphia	Shapefile	DIVISION_N
			geometry
Washington D.C. Crime 2018	Location of incidents of crime in Washington D.C. in 2018	CSV	WARD
			geometry
Washington D.C. Crime 2019	Location of incidents of crime in Washington D.C. in 2019	CSV	WARD
			geometry
Washington D.C. Street Lights	Locations and attributes of street lights in Washington D.C.	Shapefile	WARD
			geometry
Washington D.C. Trash Cans	Location of trash cans maintained by the Department of Public Works in Washington D.C.	Shapefile	LITTERID
			geometry
Washington D.C. Street Trees	Location of street trees in Washington D.C.	Shapefile	WARD
			geometry
Washington D.C. Precinct Boundaries	Voting precinct boundaries for Washington D.C.	Shapefile	NAME
			geometry

2.4 Exploratory Analysis

In order to address data mining goals associated with this analysis, data exploration prior to data-preparation has been utilized. Summary statistics for complete datasets have been conducted and are outlined in Appendix D. Data was also plotted and visualized for further data-understanding prior to data-cleaning. However, some datasets were unable to be plotted during this phase without first

initiating data cleaning. Please refer to Appendix E for example figures of San Francisco relevant dataset visualization prior to data-cleaning.

3.0 Data Preparation

Below, Figure 2 describes the general workflow involved in preparing the data for analysis and modelling.

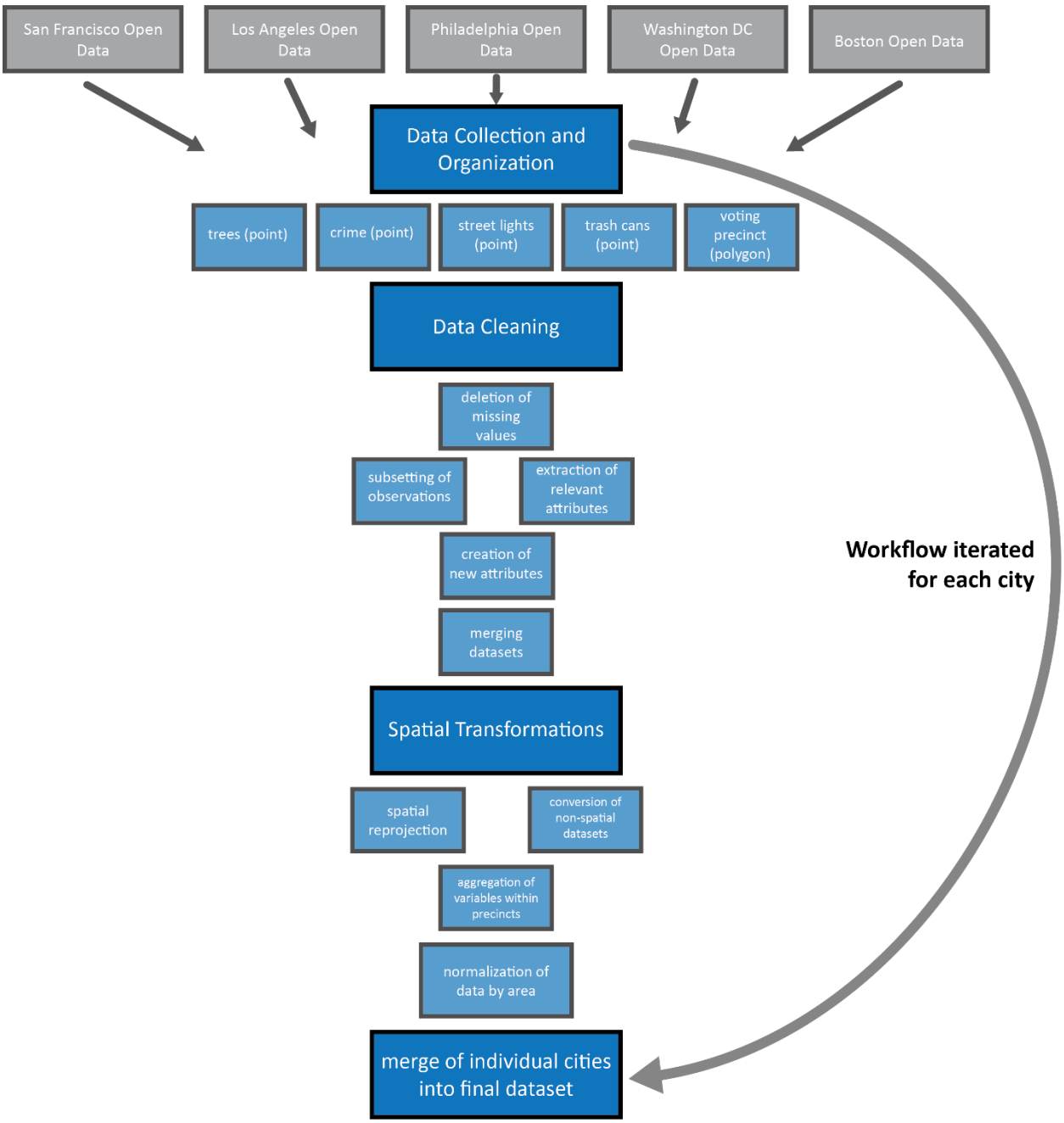

Figure 2: General workflow of project data preparation

As illustrated by Figure 2, a number of steps are involved in preparing the data for analysis. These are described below:

1. Collection of Data and Database Organisation:
First, the data is downloaded from each city's open data websites and organised into relevant folders ("crime", "trees", "TrashCans", "StreetLights", "Precincts"). Each variable and file created during the analysis is named following the general pattern "cityabbreviation_variable". For example, Boston trees are named "BO_trees". The data is read into Jupyter Notebooks (running on an R language kernel) as data frames (.csv files) and spatial objects (.shp files with polygon or point geometry attributes). For each city the data collected is crime (as a point shapefile or .csv file), trees (as a point shapefile or .csv file), trash cans (as a point shapefile or .csv file), and voting precincts (as a multipolygon shapefile).

2. Data Cleaning:
Next, the data is cleaned and prepared for the next step of processing. Rows that contain missing values (NAs) in the geometry attribute are deleted (since all observations must be spatial for the analysis), and relevant attributes are extracted. For trees, trash cans, and street lights the attributes of interest are simply the geometry attribute which indicate each observation's geographic location. For precincts, the attributes of interest are the precinct identifier and the geometry attribute. For cities having one crime dataset for each year (e.g. Washington DC), a merge is performed row-wise. Other datasets are subsetted to contain only relevant observations (e.g. subsetting 2018 and 2019 crime observations for historical crime datasets; e.g. San Francisco trash cans and streetlights must be extracted from the city's "Assets" dataset). Finally, "crimecount", "treecount", "trashcount", and "lightcount" attributes (for each city's crime, trees, trash cans, and street lights) are created respectively.

3. Spatial Transformation:
In this step, non spatial datasets are converted into spatial objects by using their associated longitude and latitude attributes. All spatial datasets are then reprojected into one common coordinate system: WGS-1984. Finally, occurrences of each variable of interest ("crimecount", "treecount", "trashcount", "lightcount") are spatially aggregated and summarised within analysis geometries (voting precincts).

4. Normalization:
All variables are normalized by precinct area (i.e. number of crimes/precinct area).

5. Merging Datasets from Individual Cities into One Dataset:
Once Steps 1-4 have been completed for each city (San Francisco, Los Angeles, Philadelphia, Washington D.C., and Boston), all cities' individual datasets are merged row-wise. Finally, binary dummy columns are created for our categorical variable ("CITY").

Please see appendix F for a complete documentation of the data preparation part, including all R code used.

3.1 Dataset and Dataset Description

Please see Table 2 for a detailed description of each dataset used for this analysis.

3.2 Data Visualization

After data-cleaning, initial visualization of our variables on a map show that each city displays a unique distribution of crime, street trees, street lights, and trash cans. Nonetheless, common to all cities is the concentration of crime in hotspots.

San Francisco

In San Francisco, crime hotspots include Chinatown, the Fisherman's Wharf, the Financial district, the Tenderloin, and Mission District. Meanwhile, trees are concentrated in Western Addition, Haight, and Castro - Noe Valley. Street lights and trash cans are scarcely distributed throughout the city. Overall San Francisco has much fewer street lights and trash cans than the other cities studied. In San Francisco, there seems to be a relationship between the spatial distribution of trees and crime - that is, trees seem to move away from crime and vice versa. Figure 3 illustrates the distribution of crime, street trees, trash cans, and street lights in San Francisco.

Crime, Street Trees, Street Lights and Trash Cans in San Francisco, USA

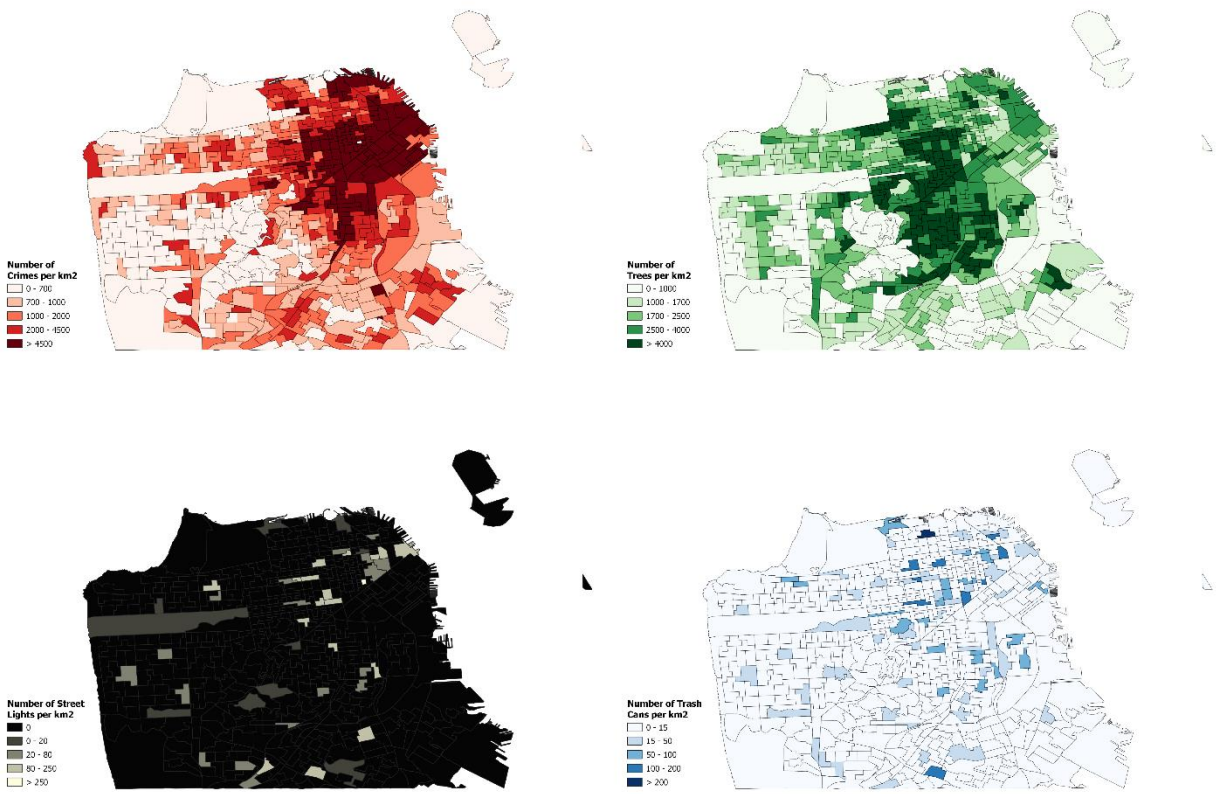

Figure 3: Crime, street trees, street lights, and trash cans in San Francisco, USA

The relationship between crime and our variables of interest is examined further in Figure 4, in which the logarithm of each variable is plotted against the logarithm of number of crimes/km^2. On the y-axis log(number of crimes/km^2) is present and on the x-axis log(number of street trees/km^2) is present. The size of the points is represented by log(number of trash cans/km^2) and the color of the points (from dark to light) is represented by log(number of street lights per km^2). The logarithm of each variable is used to minimize the effect of outliers on the general trends illustrated in Figure 4. In San Francisco, the relationship between trash cans and crime is unclear, due to many outliers. Unexpectedly, there seems to be a positive correlation between trees and crime. The relationship between number of street lights

and crime is unclear from this graph. Possible explanations for such observed trends are explored at the end of this section.

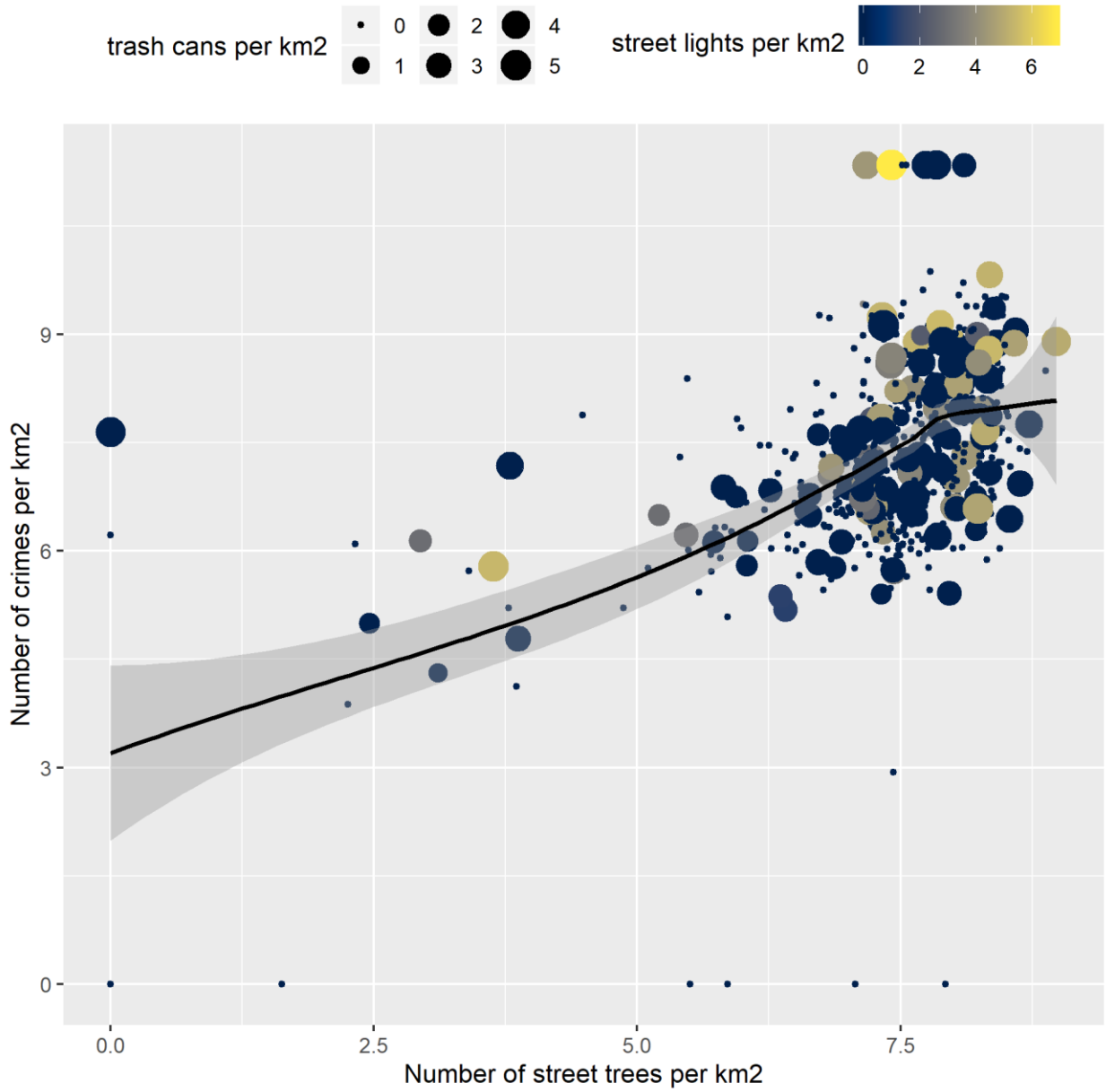

Figure 4: The relationship between number of crimes and each of the variables of interest, street trees, street lights, and trash cans in San Francisco, USA

Los Angeles

In Los Angeles, all variables seem to follow similar spatial trends. Crime hotspots, together with the highest concentrations of streetlights and trash cans, are found in between Southeast LA and Hollywood from (South to North) and between Lincoln Heights and West Los Angeles. While trees also seem to follow this spatial distribution, the tree distribution seems slightly shifted West. Figure 5 illustrates the distribution of crime, street trees, trash cans, and street lights in Los Angeles.

Crime, Street Trees, Street Lights and Trash Cans in Los Angeles, USA

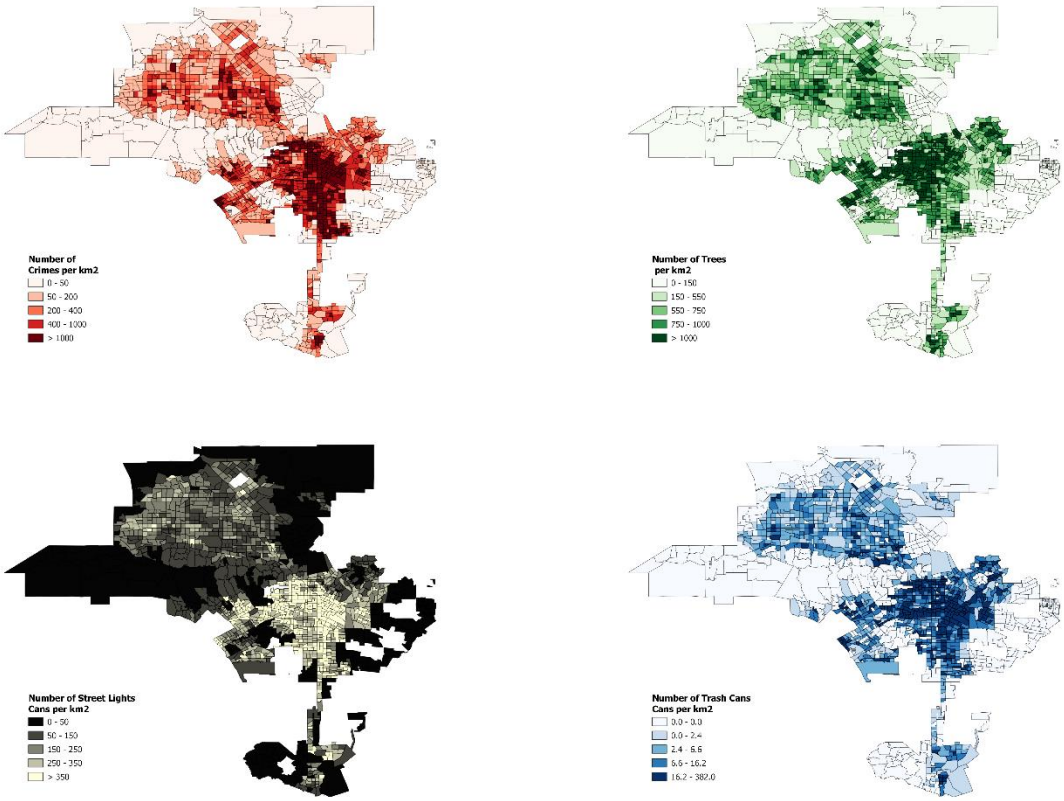

Figure 5: Crime, street trees, street lights, and trash cans in Los Angeles, USA

The relationship between crime and our variables of interest is examined further in Figure 6, in which the logarithm of each variable is plotted against the logarithm of number of crimes/km^2. On the y-axis log(number of crimes/km^2) is present and on the x-axis log(number of street trees/km^2) is present. The size of the points is represented by log(number of trash cans/km^2) and the color of the points (from dark to light) is represented by log(number of street lights per km^2). The logarithm of each variable is used to minimize the effect of outliers on the general trends illustrated in Figure 4. Unexpectedly, there seems to be a fairly positive correlation between trees and crime in Los Angeles, as well as a positive correlation between the number of trash cans and crime. The relationship between number of street lights and crime is slightly less clear. While street lights and crime appear to be positively correlated, it is important to note an abundance of outliers that exist. Crime appears to persist in instances of both high and low street light density. Possible explanations for such observed trends are explored at the end of this section.

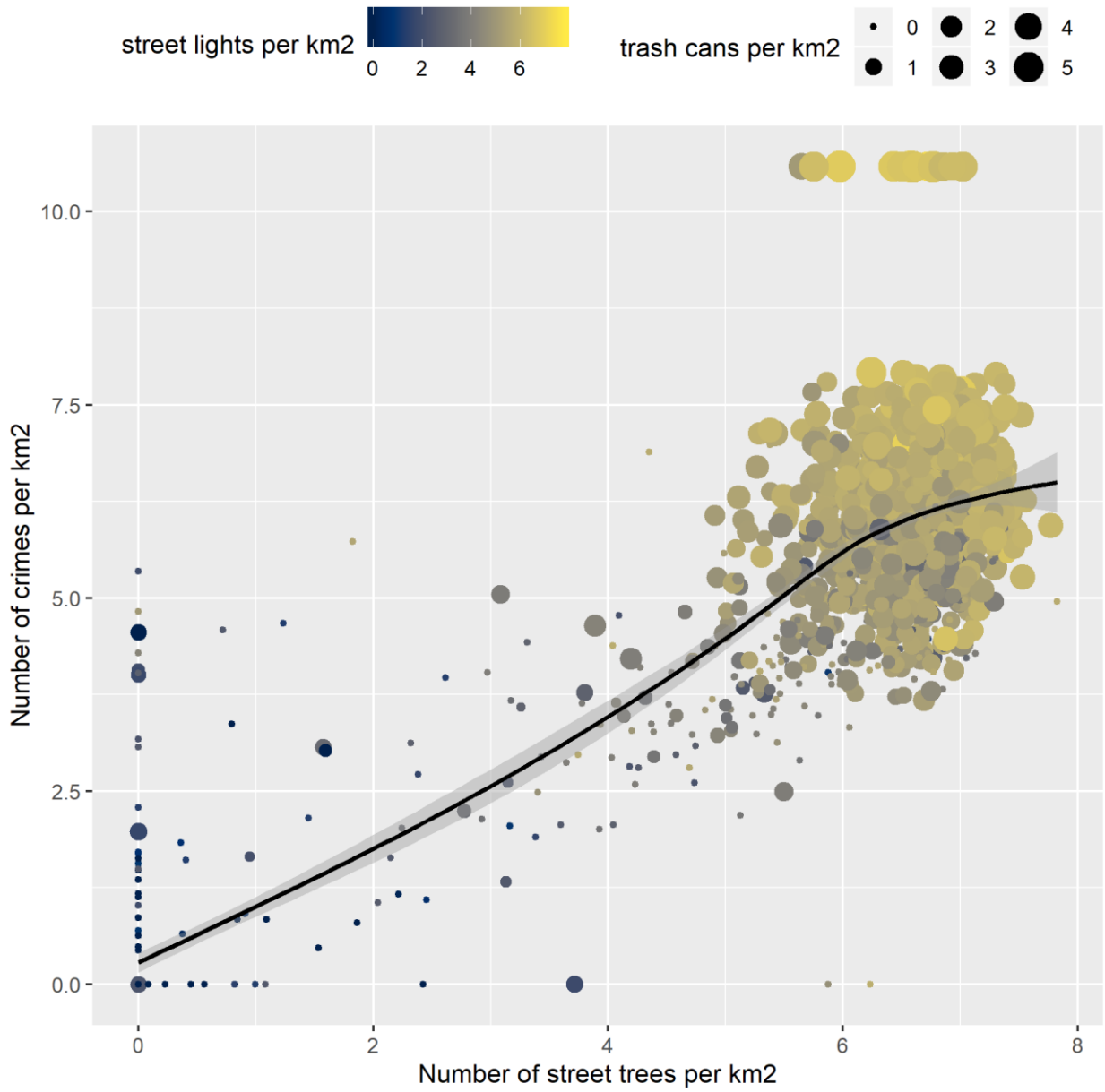

Figure 6: The relationship between number of crimes and each of the variables of interest, street trees, street lights, and trash cans in Los Angeles, USA

Philadelphia

In Philadelphia crime is concentrated around South Philadelphia, Center City, Overbrook, Cobbs Creek, Parkside, Harrowgate, Strawberry Mansion, Swampoodle and Frankford. The spatial distribution of trees and street lights seems to closely follow that of crime with a stronger emphasis, however, on South Philadelphia. Trash cans also seem to have a higher than normal concentration in South Philadelphia, although the trend is less clear. Figure 7 illustrates the distribution of crime, street trees, trash cans, and street lights in Philadelphia.

Crime, Street Trees, Street Lights and Trash Cans in Philadelphia, USA

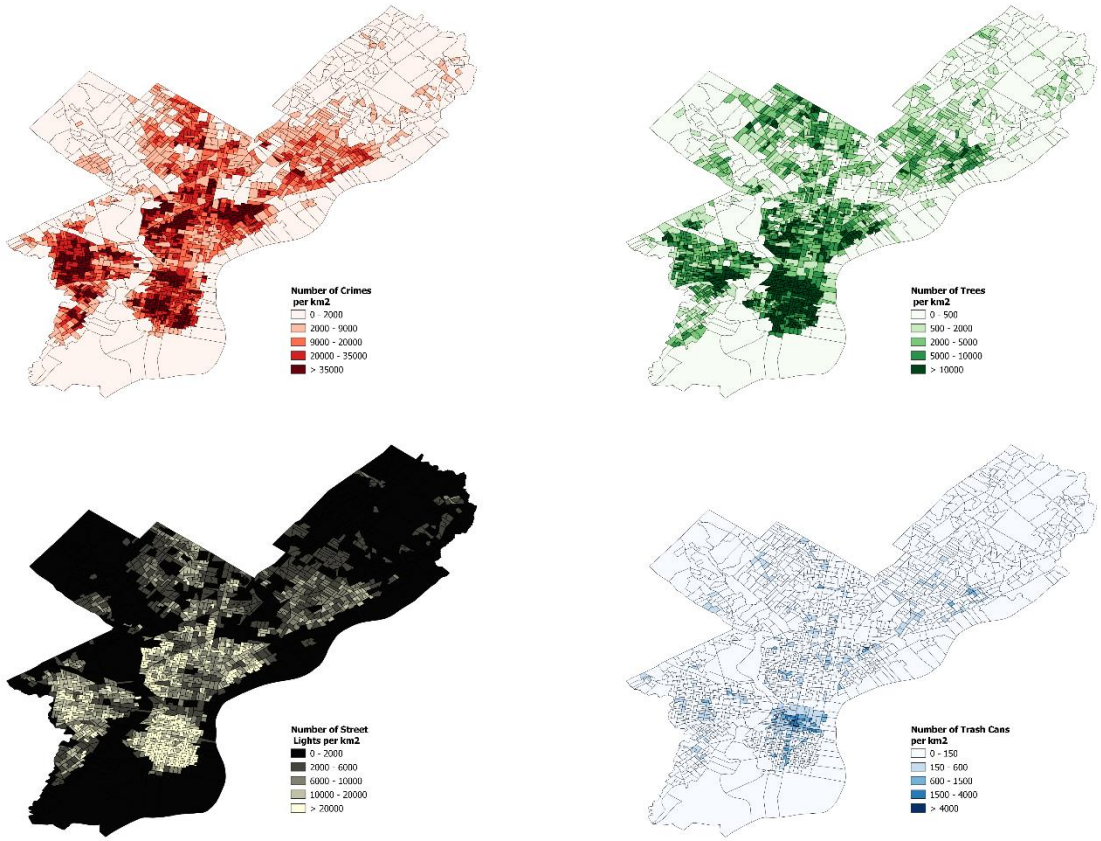

Figure 7: Crime, street trees, street lights, and trash cans in Philadelphia, USA

The relationship between crime and our variables of interest is examined further in Figure 8, in which the logarithm of each variable is plotted against the logarithm of number of crimes/km^2. On the y-axis log(number of crimes/km^2) is present and on the x-axis log(number of street trees/km^2) is present. The size of the points is represented by log(number of trash cans/km^2) and the color of the points (from dark to light) is represented by log(number of street lights per km^2). The logarithm of each variable is used to minimize the effect of outliers on the general trends illustrated in Figure 8. Unexpectedly, there seems to be a positive correlation between trees and crime in Philadelphia, as well as a positive correlation between the number of trash cans and crime. However, the positive correlation between trash cans and crime is less striking, due to the existence of outliers. The relationship between the number of street lights and crime is also positive from this graph. Possible explanations for such observed trends are explored at the end of this section.

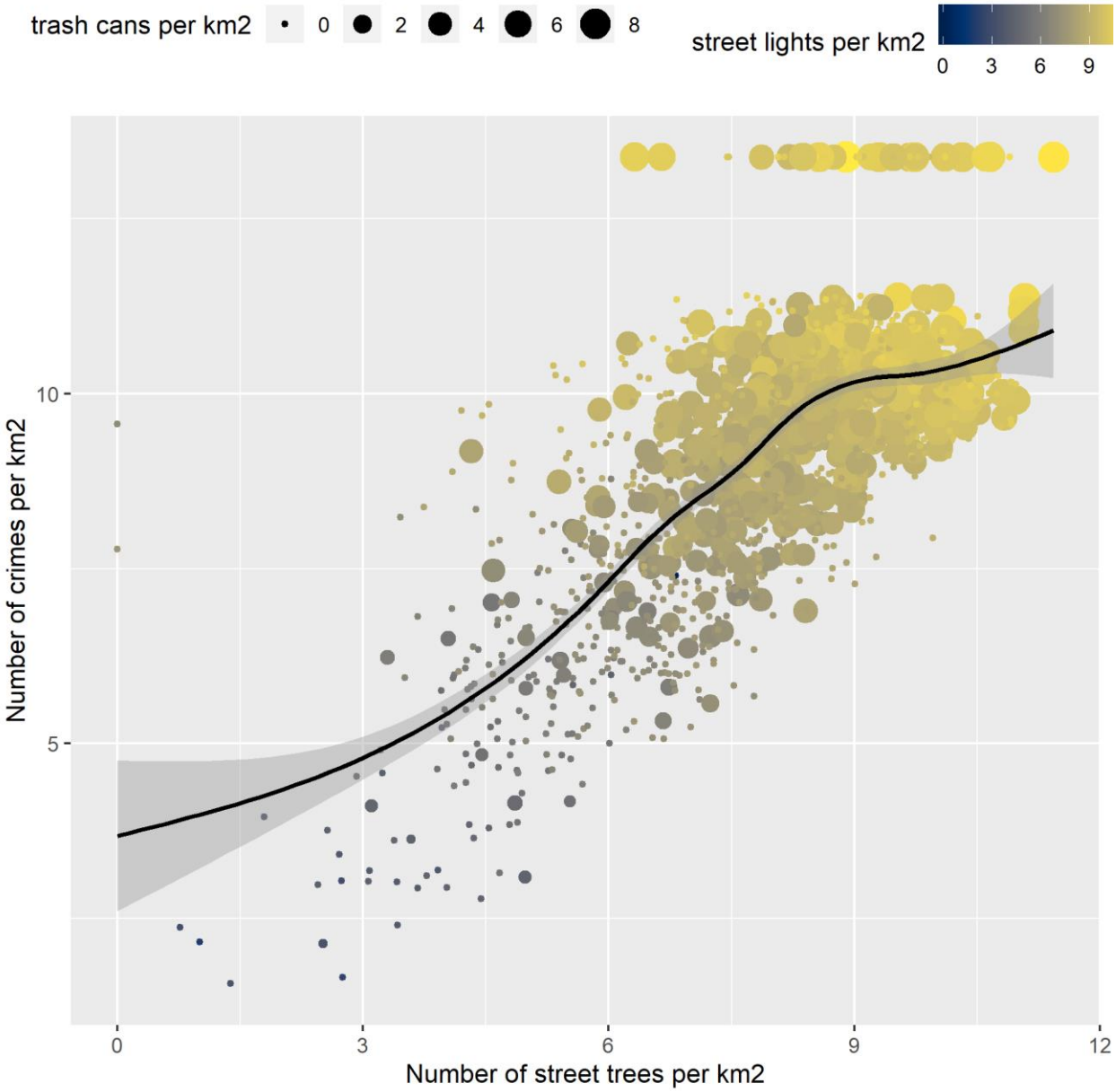

Figure 8: The relationship between number of crimes and each of the variables of interest, street trees, street lights, and trash cans in Philadelphia, USA

Washington D.C.

In Washington D.C., crime, street lights, trash cans and street trees hotspots are all found in the Southside of Banneker City, the Westside of Capitol Hill, and North West of the Southern District. Figure 9 illustrates the distribution of crime, street trees, trash cans, and street lights in Washington DC.

Crime, Street Trees, Street Lights and Trash Cans in Washington DC, USA

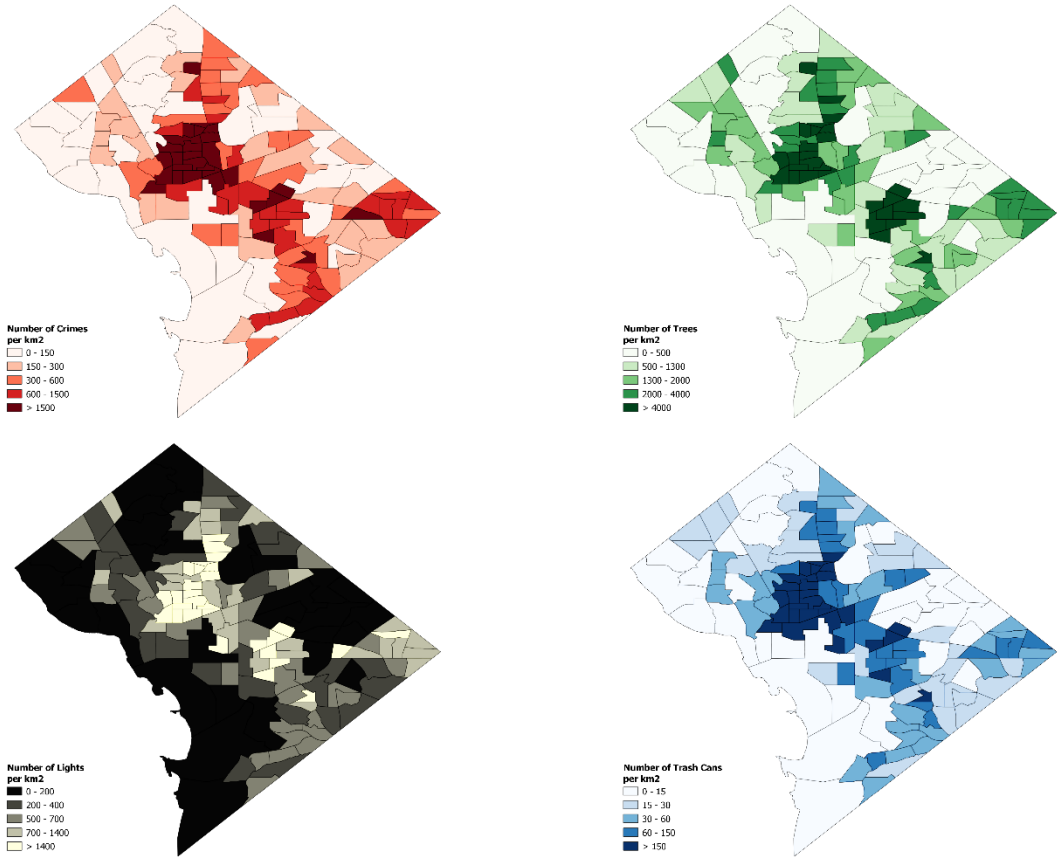

Figure 9: Crime, street trees, street lights, and trash cans in Washington D.C., USA

The relationship between crime and our variables of interest is examined further in Figure 10, in which the logarithm of each variable is plotted against the logarithm of number of crimes/km^2. On the y-axis log(number of crimes/km^2) is present and on the x-axis log(number of street trees/km^2) is present. The size of the points is represented by log(number of trash cans/km^2) and the color of the points (from dark to light) is represented by log(number of street lights per km^2). The logarithm of each variable is used to minimize the effect of outliers on the general trends illustrated in Figure 4. Unexpectedly, there seems to be a fairly positive correlation between trees and crime in Washington D.C., as well as a positive correlation between the number of trash cans and crime. The relationship between number of street lights and crime is also positively correlated. Possible explanations for such observed trends are explored at the end of this section.

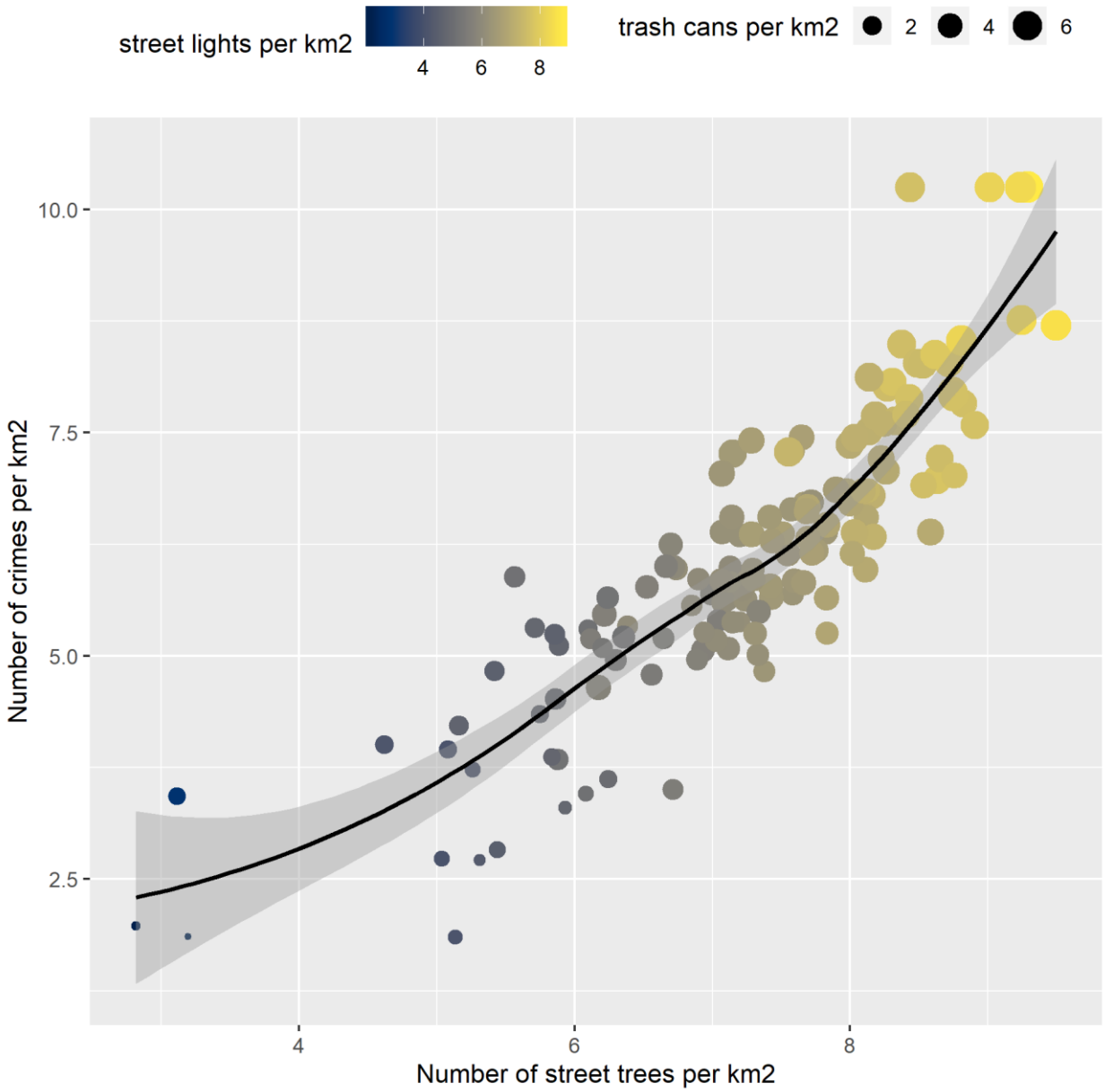

Figure 10: The relationship between number of crimes and each of the variables of interest, street trees, street lights, and trash cans in Washington D.C., USA

Boston

Boston's crime hotspots are located in the South End and North Dorchester. Trees are also concentrated in the South End, but display significant hotspots in Beacon Hill-Black Bay, Allston Brighton and Charlestown as well. Street lights closely mirror the spatial distribution of crime. Finally, there seem to be very few trash cans in Boston, which are mostly concentrated in Central, South End, and Fenway Kenmore. Figure 11 illustrates the distribution of crime, street trees, trash cans, and street lights in Boston.

Crime, Street Trees, Street Lights and Trash Cans in Boston, USA

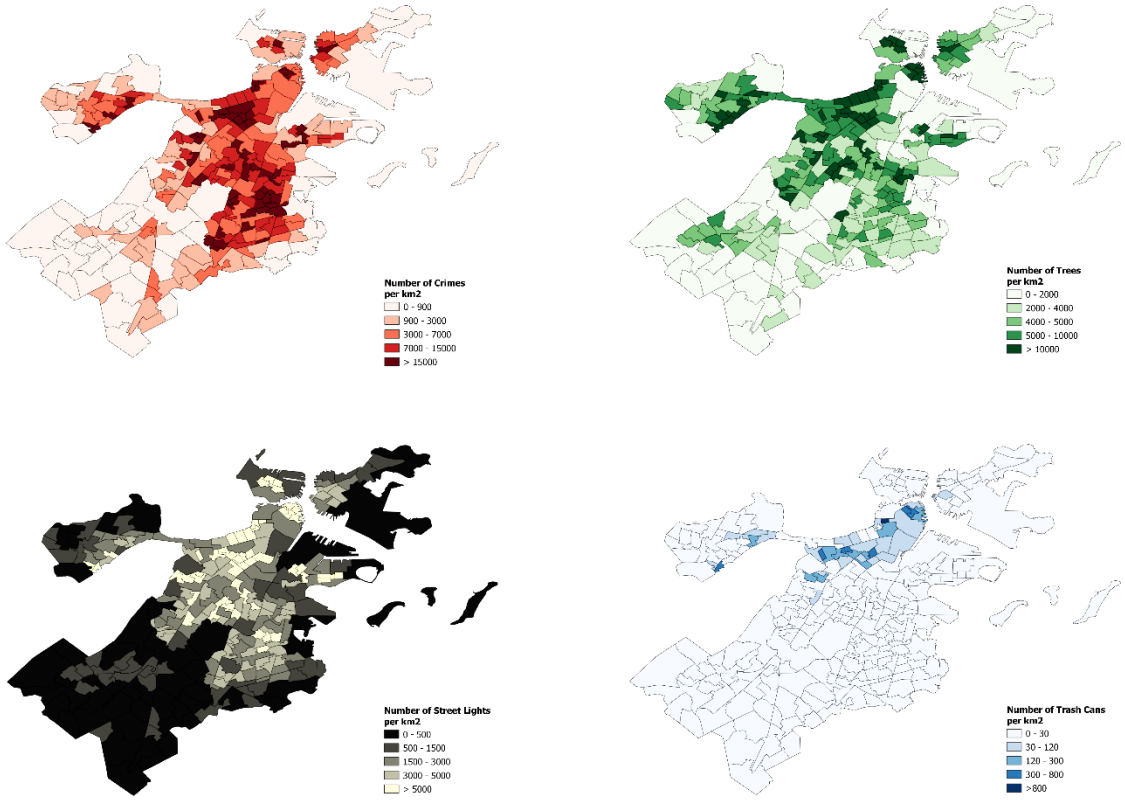

Figure 11: Crime, street trees, street lights, and trash cans in Boston, USA

The relationship between crime and our variables of interest is examined further in Figure 12, in which the logarithm of each variable is plotted against the logarithm of number of crimes/km^2. On the y-axis log(number of crimes/km^2) is present and on the x-axis log(number of street trees/km^2) is present. The size of the points is represented by log(number of trash cans/km^2) and the color of the points (from dark to light) is represented by log(number of street lights per km^2). The logarithm of each variable is used to minimize the effect of outliers on the general trends illustrated in Figure 4. Unexpectedly, there seems to be a positive correlation between trees and crime in Boston, as well as a positive correlation between the number of trash cans and crime. However, the correlation of trash cans with crime is less clear due to the presence of outliers. The relationship between number of street lights and crime is also appears to be positively correlated, however outliers also make this trend less apparent. Possible explanations for such observed trends are explored at the end of this section.

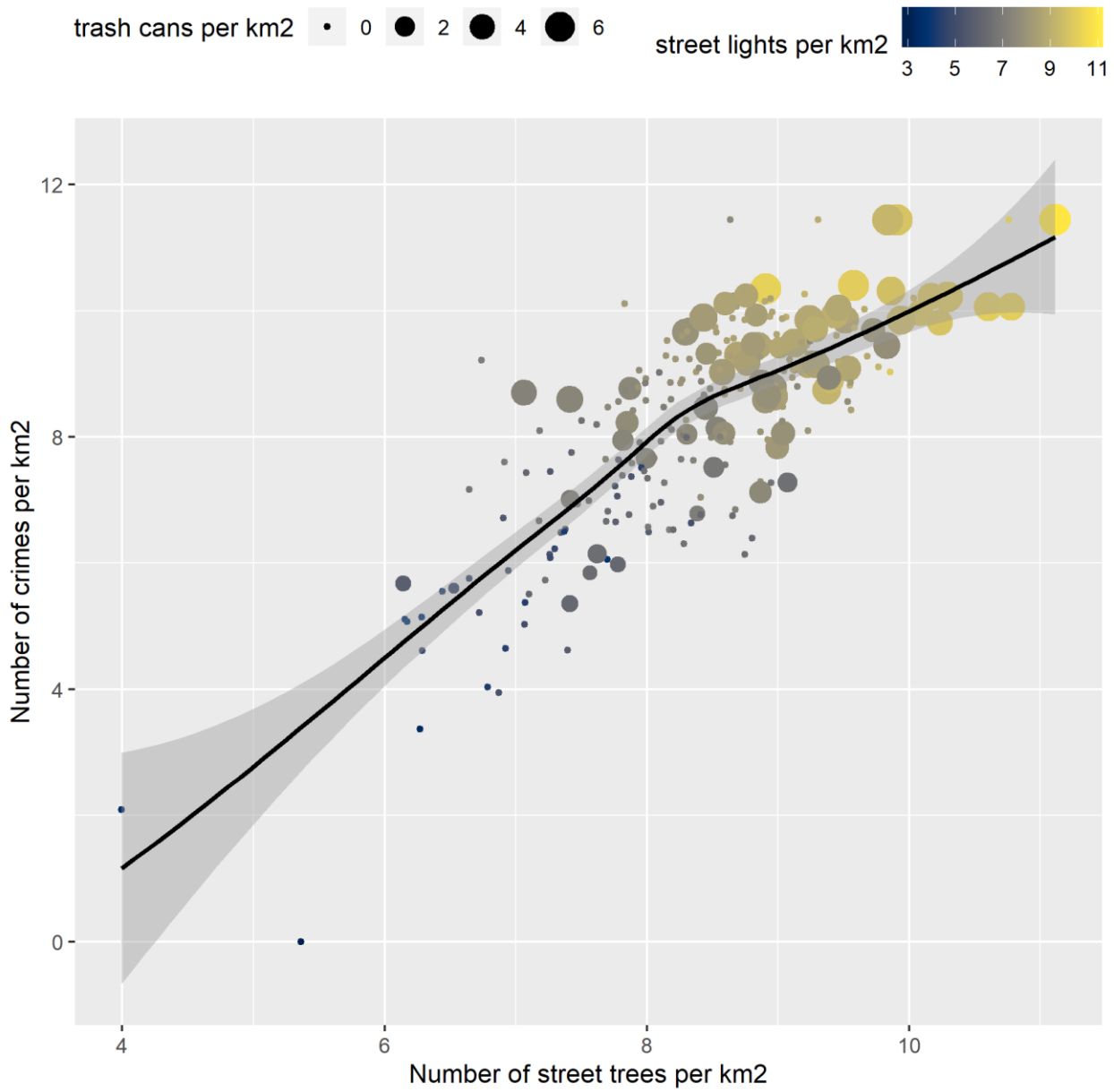

Figure 12: The relationship between number of crimes and each of the variables of interest, street trees, street lights, and trash cans in Boston, USA

Explanations to Observations

In Figures 4, 6, 8, 10, and 13, striking trends are observed. For example, in all cities (San Francisco, Los Angeles, Philadelphia, Washington D.C., and Boston) positive correlations between crime and presence of trees are observed. However, a substantial amount of literature illustrates that the presence of trees and greater tree canopy cover often decrease crime (Bogar & Meyer, 2015; Kuo & Sullivan, 2011; Troy, Grove, & O'Neil-Dunne, 2012). Notably, tree canopy cover is the most accurate measurement for ecosystem services. Such ecosystem services include shade, cooling, physical and mental health benefits for people, and aesthetic value. Thus, one tree with a large canopy may provide a considerably greater amount of ecosystem services than a multitude of young seedlings with small canopies. Due to

data availability, this study, however, utilises the number of trees present within the city limits, as opposed to tree canopy cover. Thus, potential inaccuracies may exist within this analysis, posing as a limitation to this study and serving as a potential explanation for the trends observed between crime and trees.

Additionally, each crime observation has been divided by the area of the precinct for a normalised comparison between observations. More accurate results may have surfaced had crime observations been divided by population instead. Thus, it is also possible that such limitation within this study has resulted in such striking trends observed in Figures 4, 6, 8, 10, & 12.

Specifically, in the contexts of San Francisco and Los Angeles, trends are arguably less apparent. In Figures 4 and 5, while there may be an overall trend observed, the removal of outlying observations could result in a new trend observed within existing clustered observations.

With regard to Washington D.C., the large size of precinct boundaries may make it difficult to appropriately capture differences in crime from street to street within precinct boundaries. It is possible that a large amount of variation exists within each precinct boundary. Thus, this limitation may account for the current appearance of Figure 10. In order to better capture the granularity of crime not only in Washington D.C., but also San Francisco, Los Angeles, Philadelphia, and Boston, census blocks are parcel levels could be utilised for this analysis.

For an outline of all code used for data visualisation purposes, please see Appendix G.

4.0 Modeling

As discussed in the previous section, 3.2 Data Visualisation, the relationships observed from our data are unlikely to be causal and such results indicate that hidden variables that influence crime in US cities may have been left out of this analysis. Nonetheless, the correlations that exist between our variables of interest (street trees, street lights, and trash cans) and our target variable (crime) are strong enough for a model to be developed.

<u>Selecting modelling techniques</u>

Note that the model selected for this study must take into consideration the effects of the following on crime:

- Number of street trees/km^2
- Number of street lights/km^2
- Number of trash cans/km^2
- The fixed effect of changing cities on crime

Thus, in selecting an appropriate model for this study, the following is considered:

1. The data type available for mining is entirely numeric, as listed below:
    - Number of street trees/km^2: this data is inputted into the model as an array of numeric values of type “double”.
    - Number of street lights/km^2: this data is inputted into the model as an array of numeric values of type “double”.
    - Number of trash cans/km^2: this data is inputted into the model as an array of numeric values of type “double”.

- City: the effect of changing cities on crime is inputted into the model as binary dummy columns for each city with a value of 1 when the observation is in a given city and a value of 0 when an observation is not in a given city.

2. The data mining goals for this study are:
   - to estimate the weight of each variable of interest on crime
   - to accurately predict crime based on the variables of interest

3. As observed in the previous section, 3.2 Data Visualisation, the data used in this study contains a significant number of outliers. To minimize the effects of such outliers while still considering them into the analysis, all variables are set to a logarithmic scale.

4. The probability distribution of the target variable (crime) for all cities, as plotted in Figure 13 (total distribution) and Figure 14 (distribution by city).
   -
   - Figure 13 and Figure 14 demonstrate that the data follows a bimodal distribution. However, seeing as this is the first attempt at modelling crime, the model selected for this study assumes that crime is normally distributed.

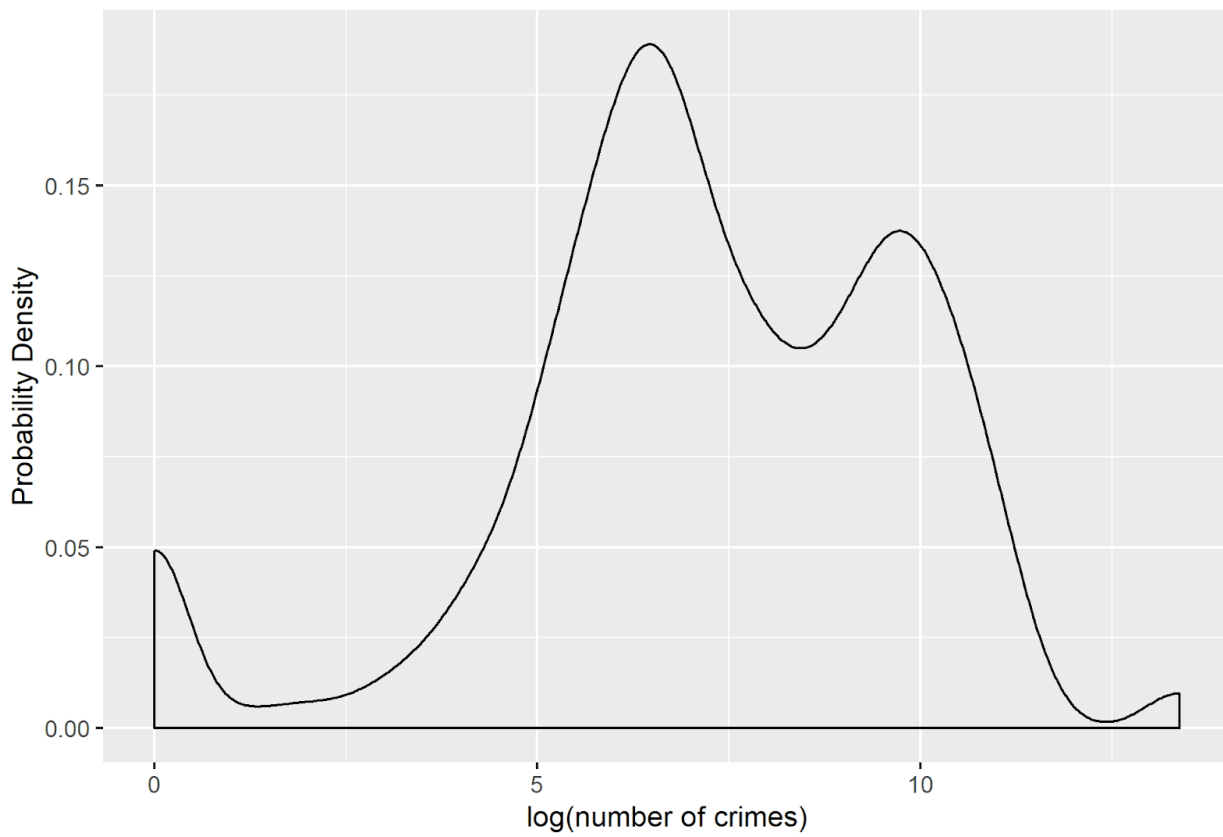

Figure 13: Total probability distribution of the target variable, crime

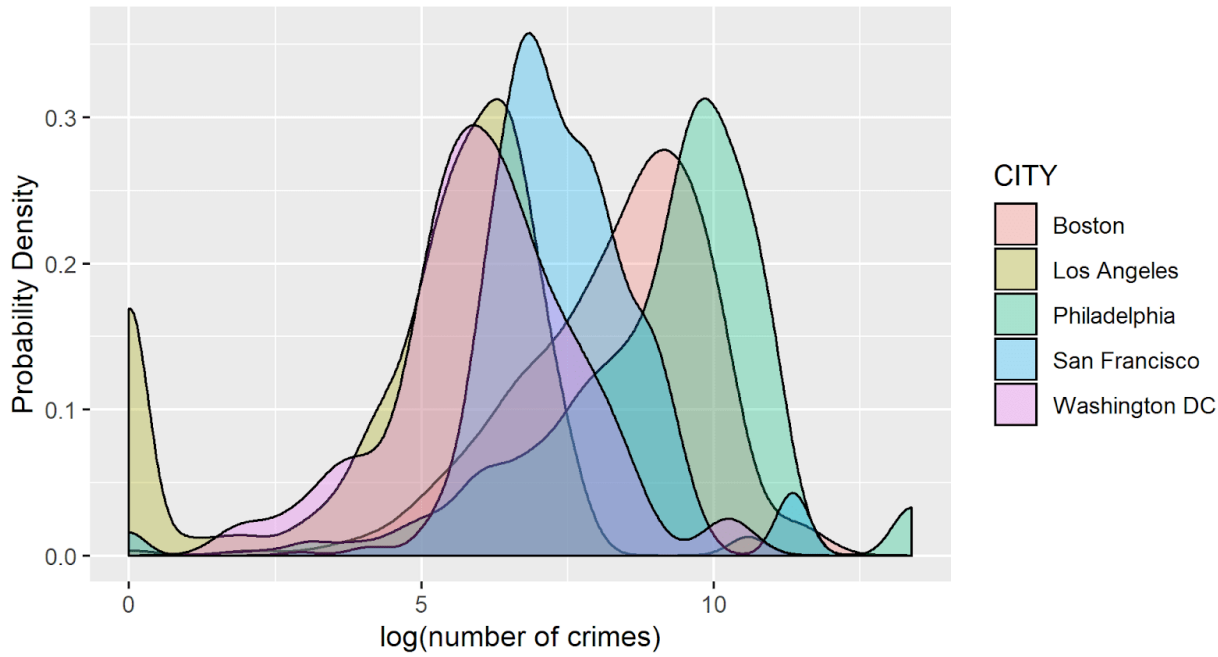

Figure 14: Probability distribution by city of the target variable, crime

Thus, a predictive collider model using Bayesian methodology is selected to explore the effect of each variable of interest in the number of crimes in a given precinct.

4.1 Test Design

The results of the model are assessed through the following methods:

- The distributions of the posteriors are assessed.
- The convergence diagnostics are assessed.
- A comparison of predicted crime using the mean of posterior distributions vs. observed crime is assessed.
- The time it takes for the model to run is assessed.

4.2 Models

The model built for this study is a Bayesian multiple linear regression (mixed) MCMC model. The model is based on the following assumption that the crime data are random samples from normal distributions with equal variance and a mean that can be computed from a linear function of the interest variables (or predictor variables). The model is built using the JAGS package for R.

The model has the following parameters and variables:

- crime[i]: the target variable
- mu[i]: the mean of the target variable that must be modelled/predicted
- sigma: the standard deviation
- inv.var: the prior for the inverse variance of sigma and mu
- beta0: the prior for the y intercept
- beta1: the prior for the coefficient for the “trees” predictor variable
- beta2: the prior for the coefficient for the “street lights” predictor variable
- beta3: the prior for the coefficient for the “trash cans” predictor variable
- beta4: the prior for the coefficient for the “San Francisco” predictor variable

- beta5: the prior for the coefficient for the “Los Angeles” predictor variable
- beta6: the prior for the coefficient for the “Philadelphia” predictor variable
- beta7: the prior for the coefficient for the “Washington DC” predictor variable
- beta8: the prior for the coefficient for the “Boston” predictor variable

The model built for this study models mu[i] with the following linear equation:

$$mu[i] = \beta 0 + \beta 1 * trees[i] + \beta 2 * lights[i] + \beta 3 * trash[i] + \beta 4 * SF[i] + \beta 5 * LA[i] + \beta 6 * PH[i] + \beta 7 * DC[i] + \beta 8 * BO[i]$$

Figure 15: Multivariate linear equation used in the model for estimating the value of mu[i]

4.3 Parameter Setting

The model written in JAGS and R language is illustrated in Figure 16.

```
model_string1 <- "model{

#likelihood
for (i in 1:n){
  crime[i] ~ dnorm(mu[i], inv.var) #insert function that represents distribution of crime
  mu[i] <- beta0 + beta1*trees[i] + beta2*lights[i] + beta3*trash[i] + beta4*SF[i] + beta5*LA[i] + beta6*PH[i] + beta7*DC[i] + beta8*BO[i]
}

#one prior for each beta
beta0 ~ dnorm(0, 0.001)
beta1 ~ dnorm(0, 0.001)
beta2 ~ dnorm(0, 0.001)
beta3 ~ dnorm(0, 0.001)
beta4 ~ dnorm(0, 0.001)
beta5 ~ dnorm(0, 0.001)
beta6 ~ dnorm(0, 0.001)
beta7 ~ dnorm(0, 0.001)
beta8 ~ dnorm(0, 0.001)


#prior for the inverse variance
inv.var ~ dgamma(0.01, 0.0001)
sigma <- 1/sqrt(inv.var)
}"
```

Figure 16: Bayesian multiple linear regression (mixed) MCMC model utilised for this study

As demonstrated by Figure 16, the model parameters are set in the following fashion:

- inv.var: gamma distribution with a mean of 0.01 and a distribution’s precision (1 / variance) of 0.0001
- beta0: normal distribution with a mean of 0 and a distribution’s precision (1 / variance) of 0.001
- beta1: normal distribution with a mean of 0 and a distribution’s precision (1 / variance) of 0.001
- beta2: normal distribution with a mean of 0 and a distribution’s precision (1 / variance) of 0.001
- beta3: normal distribution with a mean of 0 and a distribution’s precision (1 / variance) of 0.001
- beta4: normal distribution with a mean of 0 and a distribution’s precision (1 / variance) of 0.001
- beta5: normal distribution with a mean of 0 and a distribution’s precision (1 / variance) of 0.001
- beta6: normal distribution with a mean of 0 and a distribution’s precision (1 / variance) of 0.001
- beta7: normal distribution with a mean of 0 and a distribution’s precision (1 / variance) of 0.001
- beta8: normal distribution with a mean of 0 and a distribution’s precision (1 / variance) of 0.001

4.4 Model Description

General observations regarding the model used for this study are listed below:

- Model Run Time: the model was relatively fast to execute, considering it generated 20,000 samples for each of 8 priors total.

- The cleaned input data did not generate any errors during any of the model runs. The model performed best when variable arrays were transformed using log1p() as opposed to log(). As many observations fed into the log() function were 0 values, this created many undesirable negative values within the data. Log1p() solves this problem because it takes the natural log of an observation plus one.
- By taking the mean of the predicted beta samples from the model output and plugging them into our linear equation for crime (please see Figure 15), the model was capable of predicting crime values close to those observed. Model results are assessed in section 4.5 Assessment.

4.5 Assessment

In this section the results of the model are presented and briefly evaluated. Below, Figure 17 demonstrates a trace plot with convergence diagnostics with regard to the model runs. In fact, the trajectory of the chain for each beta is consistent over time, and the distributions all look appropriately normal. The model performs well and shows good signs of convergence.

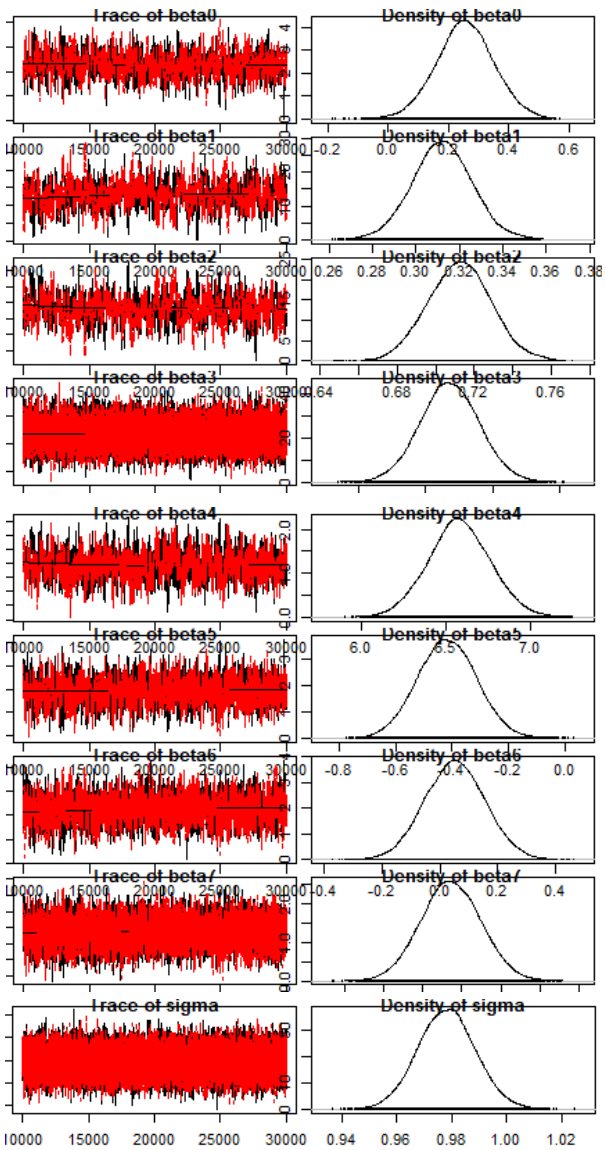

Figure 16: Trace plot with convergence diagnostics with regard to model runs

The model can also be assessed using a Gelman Rubin Diagnostic plot, as illustrated in Figure 18. Gelman Rubin uses "scale reduction factors" to quantify whether there exists a difference between the variance within chains and the variance between chains. The Gelman Rubin diagnostics provide a value for the scale reduction factors for each parameter. The closer the value is to 1, the less difference there is between variance and within chain variance. Such values for this project's parameters are observed in Figure 19 and are all very close to 1, thus showing good signs of convergence for the model.

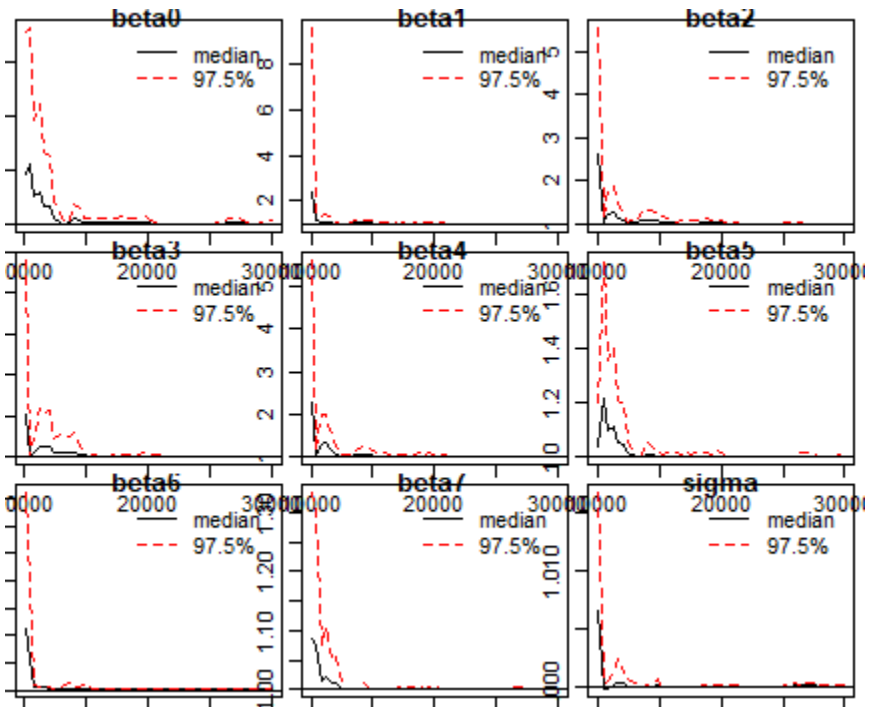

Figure 18: Gelman Rubin Diagnostic plot

```
> gelman.diag(samp)
Potential scale reduction factors:

      Point est. Upper C.I.
beta0          1       1.01
beta1          1       1.00
beta2          1       1.00
beta3          1       1.00
beta4          1       1.00
beta5          1       1.01
beta6          1       1.01
beta7          1       1.00
sigma          1       1.00

Multivariate psrf

1
```

Figure 19: Scale reduction factors from the Gelman Rubin Diagnostic

Finally, autocorrelations can be observed in Figure 20. Figure 21 demonstrates predicted values plotted against observed values.

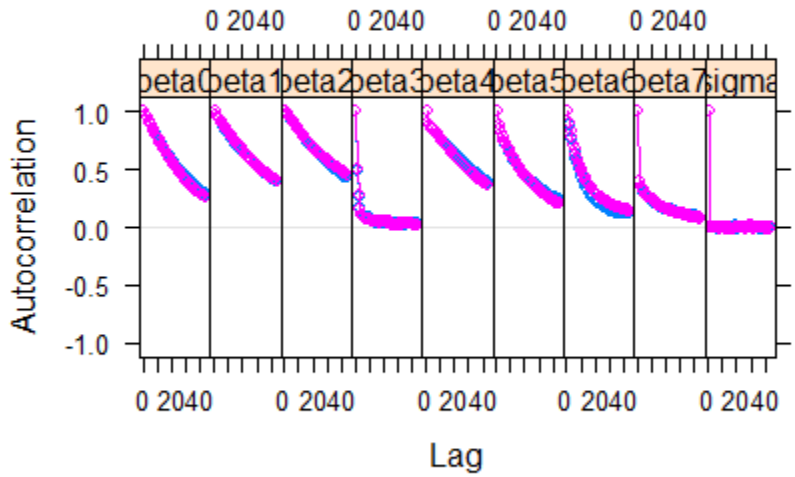

Figure 20: Autocorrelations

By taking the mean of the predicted beta distributions from the model output and plugging them into our linear equation for crime (please see Appendix H), the model was capable of predicting crime values close to those observed. This trend is shown in Figure 21.

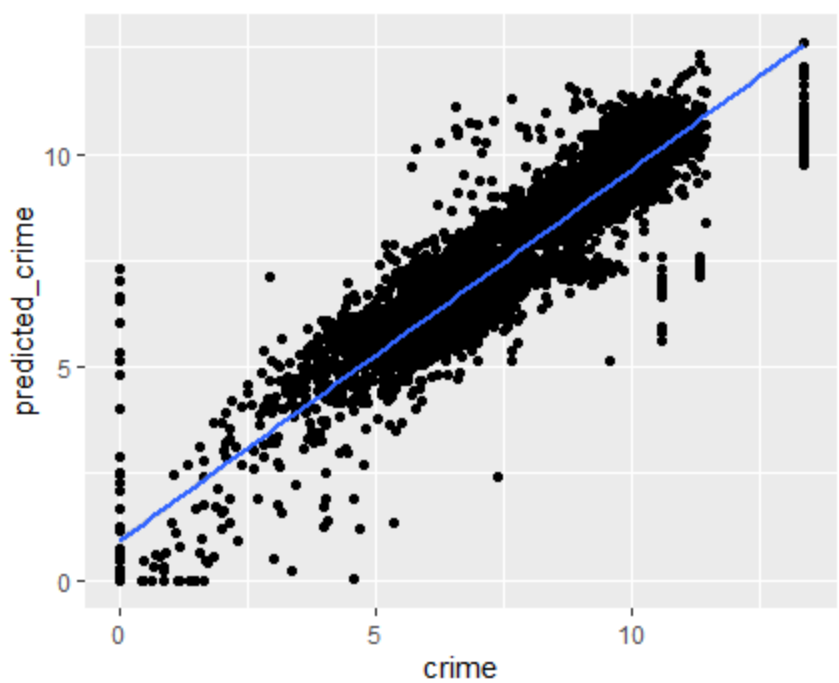

Figure 21: Predicted values plotted against observed values

5.0 Limitations & Recommendations

This study on crime and public infrastructure among US cities provides new insight into the potential relationship of minor elements of the urban fabric with crime. Post collection of data from a variety of publicly accessible sources, occurrences of crime, street trees, trash cans, and street lights were aggregated by voting precinct for five coastal US cities. The spatial distribution of these variables was mapped for each city, correlations between crime and public infrastructures were investigated and visualized, and finally a Bayesian multiple linear regression (mixed) MCMC predictive model was constructed. This study serves as a baseline for the OECD, and thus limitations are presented throughout this analysis.

Limitations:

- Data was collected entirely from government sources.
- Tree count has been utilised instead of tree canopy cover (please see further explanation in section 3.2 Data Visualization).
- Some datasets utilised in this study were lacking for some cities (e.g. trash can data for Boston was severely limited).
- Population density was not taken into account.
- This study was conducted at the precinct scale, and other scales, such as census block level or parcel level could have proved more appropriate (please see further explanation in section 3.2 Data Visualization).

Thus, given the limitations listed above, it is strongly recommended that the OECD considers the following points and improves upon them for further development of this analysis. Further recommendations for future analysis development include:

- Further analysis should normalise all variables for population as opposed to area.
- Explore additional variables, such as socio-economic variables, accessibility to public transportation, and accessibility to public services, in order to address potential hidden variables.
- Further explore historical trends in urban planning within the study.
- Explore the use of novel, open source data sources, such as:
    - OpenStreetMap for more data across more cities
    - Social media channels such as Twitter and Facebook
- Given the appropriate data, expanding the scope of this study to additional cities from other parts of the world, such as Central and South America, Europe, and Southeast Asia, could add immense value to this study and potentially generate greater insight into global crime patterns

6.0 Acknowledgement of Tasks

Working as a group of two, our team decided that each member was to be responsible for all aspects of this project (Problem Understanding, Data Understanding, Data Preparation, Analysis, and Reporting). Tasks were discussed, agreed upon, and completed by each team member as outlined in Table 6.

Table 6: Acknowledgement of tasks for team work related to this project

Acknowledgement of Tasks	
Task	**Team Member**
Group Brainstorm and Problem Identification	Caroline & Leonardo
Data Collection	Caroline & Leonardo
Written Code for Efficient Data Cleaning	Leonardo
Data Cleaning	Caroline & Leonardo
Creation of Data Visualisations	Caroline & Leonardo
Model Research	Caroline
Model Drafting	Caroline & Leonardo
Model Finalisation	Leonardo
Reporting	Caroline & Leonardo

7.0 References

Andersson, J. (2015). 'Wilding' in the West Village: Queer Space, Racism and Jane Jacobs Hagiography. *International Journal of Urban and Regional Research, 39*(2). https://doi.org/10.1111/1468-2427.12188

Bogar, S., & Meyer, K. M. (2015). Green Space, Violence, and Crime: A Systematic Review. *Cochrane database of systematic reviews, 8(8).* https://doi.org/10.1177/1524838015576412

Ceraiti, M. K. (2015). Cleaner streets may be safer streets. *The World Bank.* Retrieved from https://www.worldbank.org/en/news/feature/2015/06/09/por-que-las-calles-mas-limpias-pueden-ser-mas-seguras

Detotto, C., & Otranto, E. (2010) Does Crime Affect Economic Growth? *Kyklos, 63(3).* Retrieved from https://doi.org/10.1111/j.1467-6435.2010.00477.x

European Commission. (2019). *The 2030 Agenda for Sustainable Development and the SDGs* Retrieved from https://ec.europa.eu/environment/sustainable-development/SDGs/index_en.htm

Harcourt, B. E. (2009). *Illusions of order: The false promise of broken windows policing.* Harvard University Press. Retrieved from https://books.google.nl/books?hl=en&lr=&id=mSWDy1kGUGAC&oi=fnd&pg=PP11&dq=broken+windows+policing+US&ots=znS4a1f9ny&sig=uTD274jmGVSNfxcVQqAub5ndKvY&redir_esc=y#v=onepage&q=broken%20windows%20policing%20US&f=false

Kuo, F. E., & Sullivan, W. C. (2011). Environment and Crime in the Inner City: Does Vegetation Reduce Crime?. *Acoustics, Speech, and Signal Processing Newsletter, IEEE, 33(3).* https://doi.org/10.1177/0013916501333002

Moffatt, M. (2019). *What is Panel Data?: The Definition and Relevance of Panel Data in Economic Research.* Retrieved from https://www.thoughtco.com/panel-data-definition-in-economic-research-1147034

Morrall, P., Marshall, P., Pattison, S., & Macdonald, G. (2010). Crime and health: a preliminary study into the effects of crime on the mental health of UK university students. *Journal of Psychiatric and Mental Health Nursing, 17(1).* Retrieved from https://doi.org/10.1111/j.1365-2850.2010.01594.x

Murakawa, N., & Beckett, K. (2010). The Penology of Racial Innocence: The Erasure of Racism in the Study and Practice of Punishment. *Law and Society Review, 44*(3-4). https://doi.org/10.1111/j.1540-5893.2010.00420.x

OECD. (2019a). *Budgeting and Public Expenditures in OECD Countries in 2019.* Retrieved from http://www.oecd.org/gov/budgeting-and-public-expenditures-in-oecd-countries-2018-9789264307957-en.htm

OECD. (2019b) *Directorate for Employment, Labour and Social Affairs.* Retrieved from https://www.oecd-ilibrary.org/docserver/4663640b-en.pdf?expires=1573144692&id=id&accname=guest&checksum=4203082ED81C0075CE69A53885EF6A53

OECD. (2019c). *Launch of the Global Hub on the Governance for the SDGs.* Retrieved from https://www.oecd.org/gov/pcsd/launchoftheglobalhubonthegovernanceforthesdgs.htm

OECD. (2019d). *Who we are.* Retrieved from http://www.oecd.org/about/

Rouse, M. (2019a). *Data Mining.* Retrieved from https://searchsqlserver.techtarget.com/definition/data-mining

Rouse, M. (2019b). *Predictive Modeling.* Retrieved from https://searchenterpriseai.techtarget.com/definition/predictive-modeling

Troy, A., Grove, J. M., & O'Neil-Dunne, J. (2012). The relationship between tree canopy and crime rates across an urban-rural gradient in the greater Baltimore region. *Landscape and Urban Planning, 106(3), 262-270.* https://doi.org/10.1016/j.landurbplan.2012.03.010

United Nations. (2019a). #Envision2030 Goal 9: Industry, Innovation, and Infrastructure. *Department of Economic and Social Affairs.* Retrieved from https://www.un.org/development/desa/disabilities/envision2030-goal9.html

United Nations. (2019b). #Envision2030 Goal 16: Peace, Justice, and Strong Institutions. *Department of Economic and Social Affairs.* Retrieved from https://www.un.org/development/desa/disabilities/envision2030-goal16.html

Weisstein, E. W. (2019). *Bayesian Analysis.* Retrieved from http://mathworld.wolfram.com/BayesianAnalysis.html

Wilson, J. Q., & Kelling, G. L. (1982). *Broken Windows: The police and neighborhood safety.* Retrieved from http://www.lakeclaire.org/docs/BrokenWindows-AtlantaicMonthly-March82.pdf

Appendix A

Glossary of technical terms

Table A1: Glossary of technical terms

Glossary of Technical Terms	
Bayesian analysis	A statistical procedure for parameter estimation of an underlying distribution based on the observed distribution[1].
Data mining	The process of sorting through datasets for pattern identification and relationship establishment for problem-solving via data analysis[2].
Directorate	One of the major bureaus of the OECD, centred around a long-term and functional goal of the OECD.
Grand societal challenges	Grand societal challenges are a form of mission-oriented funding which build enthusiasm for both funders and researchers to work on common problems.
Graphical model	A style of machine learning which grew out of classical statistics and encompasses elements of Bayesian statistics and deep learning.
Member state	One of the thirty-six participating nations in the OECD.
Panel data	Data derived from a number of observations made over time on a number of cross-sectional units[3].
Predictive model	A model process that utilizes data mining and probability for outcome forecasting[4].
2030 Agenda for Sustainable Development	A commitment, adopted by the Heads of State and Government at a special UN summit, to achieve sustainable development world-wide by 2030[5].

*Definition sources are attributed as follows: 1 (Weisstein, 2019); 2 (Rouse, 2019a); 3 (Moffatt, 2019); 4 (Rouse, 2019b); 5 (European Commission, 2019)

Appendix B

Criteria for data mining success

Table B1: Criteria for data mining success

	5 (or less)	6	7	8	9 (or more)
Task 1. Problem Understanding	A problem is not selected; the data and graphics are not adapted to the needs of the decision-maker	A problem is selected, but poorly described	A problem is compellingly described. Links are made by the appropriate selection of data or the appropriate selection of graphics	A problem is compellingly described. Both data and graphics are appropriately used.	A problem is compellingly described Both data and graphics are compellingly used.
Task 2. Graphics and Visualization	Not presented	Charts are presented using base R	Charts are presented and customized; the insight from the chart is described in text	Custom graphics are included using ggplot or other packages; the graphics are tailored to for decision-makers, analysts or stakeholders	Multiple effective graphics are shown throughout the report
Task 3. Data and Preparation	Data was not selected or prepared.	Data is selected, but its appropriateness for the problem is not discussed.	Data is selected and appropriately used. Inappropriate assumptions about the data are introduced.	Appropriate cleaning was performed and discussed. This is fully evidenced using the CRISP-DM reporting mechanisms.	Exceptional efforts were made to collect additional data or to fuse multiple data sets.
Task 4 Analysis	A graphical model in R is not presented.	The model selected is not discussed	The model selected is discussed	The model selected is strongly motivated; the Bayesian assumptions are fully discussed	The model selected is strongly motivated and additional insight is provided
Task 5. Reporting	The report is incomplete.	Discrepancies or limitations of reporting given the CRISP-DM framework are discussed.	The reporting is complete; the CRISP-DM framework is used in a productive manner.	The reporting is both concise and complete.	The report invites clear next steps for additional analysis or decision-making.

Appendix C

Project plan

Table C1: Proposed project plan

Date	Phase	Task	Description
Tue, 15 October 2019	Problem Understanding	Case	Case Selection, Select an appropriate case
Wed, 16 October 2019		Graphic (optional)	Create a graphic clarifying the situation.
Fri, 18 October 2019	Data Understanding	Design (optional)	Select an appropriate database
Mon, 21 October 2019		Data	Select appropriate variables
Wed, 23 October 2019		Reporting	Complete the data understanding report
Thurs, 24 October 2019		Graphic	Create graphics showing key relationships in the data
Monday, 28 October 2019	Data Preparation	Data	Prepare the data
Tuesday, 29 October 2019		Data (optional)	Enrich the data in light of exploration
Wednesday, 30 October 2019		Reporting	Complete the data preparation section
Thursday, 31 October 2019	Analysis	Modeling	Select and run an appropriate graphical model
Friday, 1 November 2019		Modeling (optional)	Elaborate on your initial model
Monday, 4 November 2019		Reporting	Complete the analysis section
Tue, 5 November 2019		Graphics (optional)	Explore the posterior predictive distribution
Tue, 5 November 2019			Final report due

Appendix D

Summary Statistics for datasets prior to data-cleaning

San Francisco

```
> summary(SF_crime)
                          analysis_n         cad_number          cnn          filed_onli                incident_c
 Mission                        : 30465   190531525:    35   34016000:  2014   ?:217910   Larceny Theft      : 88456
 Tenderloin                     : 27497   999999999:    32   24904000:  1717   T: 59295   Other Miscellaneous: 21190
 Financial District/South Beach: 24770   180393444:    26   23914000:  1649              Non-Criminal       : 17209
 South of Market                : 22944   180681633:    20   24170000:  1437              Assault            : 16646
 Bayview Hunters Point          : 15047   180750395:    14   33719000:  1207              Malicious Mischief : 16274
 (Other)                        :140553   (Other)  :213532   (Other) :253252              (Other)            :117410
 NA's                           : 15929   NA's     : 63546   NA's    : 15929              NA's               :    20
  incident_c.1      date_incid                time_incid      date_incid.1              time_incid.1        incident_d
 06244  : 38463   Min.   :2018-01-01   00:00:00.000:277205   Min.   :2018-01-01   00:00:00.000:  6973   Friday   :42108
 71000  : 10137   1st Qu.:2018-06-15                         1st Qu.:2018-06-15   12:00:00.000:  6662   Monday   :39428
 06372  :  8490   Median :2018-11-21                         Median :2018-11-21   18:00:00.000:  5695   Saturday :39561
 04134  :  8277   Mean   :2018-11-27                         Mean   :2018-11-27   20:00:00.000:  5403   Sunday   :36408
 28150  :  8081   3rd Qu.:2019-05-17                         3rd Qu.:2019-05-17   19:00:00.000:  5245   Thursday :39190
 64020  :  6822   Max.   :2019-10-31                         Max.   :2019-10-31   17:00:00.000:  5025   Tuesday  :39318
 (Other):196935                                                                  (Other)     :242202   Wednesday:41192
                                   incident_d.1      incident_i          incident_n                 incident_s
 Theft, From Locked Vehicle, >$950        : 38463   626010 :     4   190071345:    34   Larceny - From Vehicle: 51339
 Lost Property                            : 10137   626157 :     4   180354292:    20   Other                 : 38338
 Theft, Other Property, $50-$200          :  8490   628517 :     4   190129518:    20   Larceny Theft - Other : 21190
 Battery                                  :  8277   631597 :     4   190176490:    18   Vandalism             : 15882
 Malicious Mischief, Vandalism to Property:  8081   632466 :     4   190312945:    18   Simple Assault        : 10774
 Mental Health Detention                  :  6822   633984 :     4   190014486:    17   (Other)               :139662
 (Other)                                  :196935   (Other):277181   (Other)  :277078   NA's                  :    20
```

Figure D1: San Francisco crime summary statistics before data-cleaning

```
> summary(SF_assets)
    asset_id                 asset_subt            asset_type                 comments      date_creat
 Min.   : 20715   Picnic Table     : 157   Bench         :2041   .                    :  29   Min.   :2013-05-22
 1st Qu.:409395   Park Sign        :  99   Trash Can     : 909   Need Photo           :  17   1st Qu.:2013-05-22
 Median :983326   Basketball Hoop  :  28   Lamp Post     : 639   No Plaque            :   5   Median :2014-11-13
 Mean   :775096   30 gallon        :  23   Table         : 546   .<div><br /></div>:   3   Mean   :2015-03-01
 3rd Qu.:984888   Parcourse Station:  14   Play Structure: 224   Field Verify         :   3   3rd Qu.:2016-11-22
 Max.   :986526   (Other)          :  47   (Other)       : 795   (Other)              : 134   Max.   :2018-09-13
                  NA's             :4790   NA's          :   4   NA's                 :4967
       time_creat                    created_us      facility_i         floor_id        functional
 17:10:03.000:2037   rpdgis                :2209   Min.   :      5   Min.   : 1000   Min.   :      9
 18:43:32.000: 967   RPDGIS@RPDGIS.SFGOV.ORG :1915   1st Qu.:    657   1st Qu.: 1003   1st Qu.:    761
 18:44:01.000: 913   pkercher              : 225   Median :  12480   Median : 1051   Median :  18479
 16:29:15.000:  91   jwoodard              : 184   Mean   :  34908   Mean   : 4275   Mean   :  93847
 18:44:05.000:  30   dmolnar@rpdgis.sfgov.org: 163   3rd Qu.:  36291   3rd Qu.: 1051   3rd Qu.:  86007
 13:56:37.000:  28   szhang@rpdgis.sfgov.org :  85   Max.   :1257753   Max.   :89462   Max.   :1258319
 (Other)     :1092   (Other)               : 377   NA's   :3         NA's   :5119    NA's   :136
                                globalid      date_last_          time_last_                        last_edite
 {0029FB45-89BA-47CE-856A-E36BC31463F9}:   1   Min.   :2018-07-02   18:47:56.000:  54   RPDGIS@RPDGIS.SFGOV.ORG    :3454
 {00394147-B466-4DF6-8DDB-B69FEE46C5B9}:   1   1st Qu.:2018-07-02   22:37:32.000:  42   EDOWDAKIN@RPDGIS.SFGOV.ORG :1325
 {00431511-8700-4E34-97F1-D1E4B387FD46}:   1   Median :2018-07-02   22:37:31.000:  32   kcabrera                   : 184
 {004C7CEB-FFDF-48E9-9E6D-4B245455A4A7}:   1   Mean   :2018-07-27   19:26:35.000:  14   DHALVERSON@RPDGIS.SFGOV.ORG: 106
 {005388BC-3B0B-4E92-9EFC-35F2981412FC}:   1   3rd Qu.:2018-08-10   20:18:11.000:  13   ntinclair                  :  44
 {0074E205-FF9D-43D8-BD8B-5216D27D2009}:   1   Max.   :2019-03-27   22:37:33.000:  13   rpdgis@rpdgis.sfgov.org    :  26
 (Other)                                :5152                        (Other)     :4990   (Other)                    :  19
```

Figure D2: San Francisco assets summary statistics before data-cleaning

```
> summary(SF_trees)
      dbh             latitude     location_a    location_c    location_s    location_z      longitude
 Min.   :   0.00   Min.   :37.51   NA's:192718   NA's:192718   NA's:192718   NA's:192718   Min.   :-138.3
 1st Qu.:   3.00   1st Qu.:37.74                                                           1st Qu.:-122.5
 Median :   7.00   Median :37.76                                                           Median :-122.4
 Mean   :   9.97   Mean   :37.77                                                           Mean   :-122.4
 3rd Qu.:  12.00   3rd Qu.:37.78                                                           3rd Qu.:-122.4
 Max.   :9999.00   Max.   :47.27                                                           Max.   :-122.4
 NA's   :41687     NA's   :2823                                                            NA's   :2823
                permitnote       date_plant                time_plant           planttype           plotsize
 Permit Number 45761:   300   Min.   :1955-09-19   00:00:00.000: 66555   Landscaping:   279   Width 3ft:36298
 Permit Number 44451:   226   1st Qu.:1995-01-07   14:11:00.000:    10   tree       :     3   3x3      :28992
 Permit Number 43922:   200   Median :2001-05-23   05:26:00.000:     9   Tree       :192436   Width 0ft:17010
 Permit Number 43835:   184   Mean   :2000-09-20   07:34:00.000:     8                        Width 4ft:13701
 Permit Number 46305:    65   3rd Qu.:2008-08-11   10:33:00.000:     8                        3x3      :12076
 (Other)            : 51197   Max.   :2019-12-09   (Other)     :  1627                        (Other)  :34602
 NA's               :140546   NA's   :124501       NA's        :124501                        NA's     :50039
                    qaddress                    qcareassis        qcaretaker                          qlegalstat
 100X Cargo Way          :   134   FUF                    : 22782   Private :161742   DPW Maintained              :141569
 700 Junipero Serra Blvd:   125   DPW for City Agency    :   797   DPW     : 27286   Permitted Site              : 39745
 1000 San Jose Ave       :   112   DPW                    :   260   SFUSD   :  1088   Undocumented                :  8053
 1200 Sunset Blvd        :   110   Private                :   149   Rec/Park:   726   Significant Tree            :  1606
 1600 Sunset Blvd        :   102   Cleary Bros. Landscape:   117   Port    :   723   Planning Code 138.1 required:   936
 (Other)                 :190645   (Other)                :   188   PUC     :   261   (Other)                     :   753
 NA's                    :  1490   NA's                   :168425   (Other) :   892   NA's                        :    56
```

Figure D3: San Francisco tree summary statistics before data-cleaning

```
> summary(SF_prcnts)
 assemdist bartdist   congdist         neighrep     prec_2010     prec_2012      supdist            geometry
 17:349    7   : 43   12:529   MAR/PAC HTS : 49   2001   :  4   1101   :  1   5      : 72   MULTIPOLYGON :605
 19:256    8   :268   13:  1   W TWIN PKS  : 48   1102   :  3   1102   :  1   8      : 70   epsg:4326    :  0
           9   :293   14: 75   SUNSET      : 47   2005   :  3   1103   :  1   7      : 63   +proj=long...:  0
           NA's:  1            RICHMOND    : 40   3501   :  3   1104   :  1   2      : 62
                               WST ADDITION: 38   3924   :  3   1105   :  1   9      : 53
                               MISSION     : 37   GG Park:  3   (Other):598   (Other):284
                               (Other)     :346   (Other):586   NA's   :  2   NA's   :  1
```

Figure D4: San Francisco voting precincts summary statistics before data-cleaning

Los Angeles

```
> summary(LA_crime)
     DR_NO                              Date.Rptd
 Min.   :       210   08/07/2017 12:00:00 AM:    796
 1st Qu.:121317043   12/18/2017 12:00:00 AM:    786
 Median :150519436   04/09/2018 12:00:00 AM:    775
 Mean   :146981780   07/24/2017 12:00:00 AM:    774
 3rd Qu.:171224054   11/26/2018 12:00:00 AM:    772
 Max.   :910220366   05/21/2018 12:00:00 AM:    765
                     (Other)                :2066018
                   DATE.OCC           TIME.OCC          AREA
 01/01/2010 12:00:00 AM:   2236   Min.   :   1   Min.   : 1.00
 01/01/2011 12:00:00 AM:   2159   1st Qu.: 930   1st Qu.: 6.00
 01/01/2012 12:00:00 AM:   1771   Median :1430   Median :11.00
 01/01/2013 12:00:00 AM:   1594   Mean   :1361   Mean   :11.09
 01/01/2014 12:00:00 AM:   1466   3rd Qu.:1900   3rd Qu.:16.00
 01/01/2015 12:00:00 AM:   1422   Max.   :2359   Max.   :21.00
 (Other)                :2060038
       AREA.NAME        Rpt.Dist.No      Part.1.2         Crm.Cd
 77th Street: 142502   Min.   : 100   Min.   :1.000   Min.   :110.0
 Southwest  : 132790   1st Qu.: 643   1st Qu.:1.000   1st Qu.:330.0
 N Hollywood: 111602   Median :1195   Median :1.000   Median :442.0
 Pacific    : 110082   Mean   :1156   Mean   :1.446   Mean   :507.1
 Southeast  : 109250   3rd Qu.:1668   3rd Qu.:2.000   3rd Qu.:626.0
 Mission    : 101638   Max.   :2199   Max.   :2.000   Max.   :956.0
 (Other)    :1362822
```

Figure D5: Los Angeles crime summary statistics before data-cleaning

```
> summary(LA_trashcans)
   OBJECTID_1      OBJECTID                              Date      RedCurb    Mounted       RecType      ED8                          Org
 Min.   :   1   Min.   :    5   2016-07-11T00:00:00.000Z: 745   NO :1140   NO :3858   ALB    :2131   NO :8546   LASAN                :2546
 1st Qu.:2429   1st Qu.: 2464   2016-06-27T00:00:00.000Z: 690   YES:8573   YES:5855   CEMENT : 365   YES:1167   MARTIN OUTDOOR MEDIA:1956
 Median :4857   Median : 4931   2016-07-06T00:00:00.000Z: 684                         METAL  :6947               JC DECAUX            :1555
 Mean   :4857   Mean   : 4941   2016-06-23T00:00:00.000Z: 631                         OTHER  :  84               CD10                 : 526
 3rd Qu.:7285   3rd Qu.: 7395   2016-07-05T00:00:00.000Z: 621                         SOLAR-D:  48               UNKNOWN              : 337
 Max.   :9713   Max.   :10294   (Other)                 :6326                         SOLAR-S: 138               DOWNTOWN CENTER BID : 181
                                NA's                    :  16                                                        (Other)              :2612
       CD            LAPDGrid              geometry
 Min.   : 0.000   Min.   :    0   POINT        :9713
 1st Qu.: 5.000   1st Qu.:  521   epsg:4326    :   0
 Median : 9.000   Median :  943   +proj=long...:   0
 Mean   : 8.612   Mean   : 1031
 3rd Qu.:13.000   3rd Qu.: 1496
 Max.   :15.000   Max.   :13655
```

Figure D6: Los Angeles trash can summary statistics before data-cleaning

```
> summary(LA_streetlights)
     lampa              lampb              lampc              lampd              lampe              lampf           lastplan          postdesc
 70W LED: 30731   36W LED  :   547   64W LED :   167   40W FLU :    66   1000W HPS:     3   1000W HPS:     3   M2412  :  2690   CD851A : 39377
 82W LED: 21260   300W ATCH:   465   40W FLU :   111   27W LED :    38   1000W MH :    13   1000W MH :    13   LED0143:   731   CD953C : 38925
 68W LED: 20813   34W LED  :   391   60W LED :   105   400W MH :    22   195W LED :     7   NA's     :216678   M2423  :   702   CD951A : 16051
 52W LED: 16857   82W LED  :   374   103W INC:    96   100W HPS:    21   27W LED  :    38                      LED0170:   700   CD851B : 10330
 35W LED: 15523   100W HPS :   369   55W LED :    88   195W LED:    17   40W FLU  :    20                      LED0096:   652   CD855  :  9939
 (Other):111509   (Other)  :  6903   (Other) :   272   (Other) :    31   40W LED  :     4                      (Other):211207   (Other):102071
 NA's   :     1   NA's     :207645   NA's    :215855   NA's    :216499   NA's     :216609                      NA's   :    12   NA's   :     1
      slid                     status            stlid                geometry
 Min.   :     1   AsBuilt           :207515   D14721-2:     8   POINT        :216694
 1st Qu.: 58781   Energized         :  7315   D14721-3:     8   epsg:4326    :     0
 Median :116390   Not BSL Maintained:  1863   D14721-1:     7   +proj=long...:     0
 Mean   :116911   NA's              :     1   D14721-4:     7
 3rd Qu.:173630                               D14721-5:     5
 Max.   :999999                               D14721-6:     5
                                              (Other) :216654
```

Figure D7: Los Angeles street light summary statistics before data-cleaning

```
> summary(LA_trees)
   OBJECTID          TreeID             Type
 Min.   :     1   Min.   : 224036   Min.   :1.00
 1st Qu.:158890   1st Qu.: 644688   1st Qu.:1.00
 Median :317780   Median : 972411   Median :1.00
 Mean   :317780   Mean   : 899630   Mean   :1.16
 3rd Qu.:476669   3rd Qu.:1134754   3rd Qu.:1.00
 Max.   :635558   Max.   :1301010   Max.   :4.00

                                                
```

```
> summary(BO_streetlights)
     type            long            lat           objectid              geometry
 0785  :    1   Min.   :-71.18   Min.   :42.23   Min.   :    1   POINT        :74065
 1795  :    1   1st Qu.:-71.10   1st Qu.:42.31   1st Qu.:18517   epsg:4326    :    0
 light :    5   Median :-71.07   Median :42.34   Median :37033   +proj=long...:    0
 LIGHT :74016   Mean   :-71.08   Mean   :42.33   Mean   :37033
 LIGHT`:    1   3rd Qu.:-71.06   3rd Qu.:42.35   3rd Qu.:55549
 LIGHTS:    2   Max.   :-70.93   Max.   :42.40   Max.   :74065
 NA's  :   39
```

Figure D12: Boston street lights summary statistics before data-cleaning

```
> summary(BO_trees)
    OBJECTID                          TYPE                    geometry
 Min.   :     1   Public Street Tree:148528   POINT        :201845
 1st Qu.: 50462   Tree in Park      : 53317   epsg:4326    :     0
 Median :100923                               +proj=long...:     0
 Mean   :100923
 3rd Qu.:151384
 Max.   :201845
```

Figure D13: Boston tree summary statistics before data-cleaning

```
> summary(BO_prcnts)
    OBJECTID        PRECINCT     WARD_PRECI  Plan_Dist    Shape__Are          Shape__Len              geometry
 Min.   :  1.0   01     : 22   0101   :  1   NA's:255   Min.   :5.920e-06   Min.   :0.01188   MULTIPOLYGON :255
 1st Qu.: 64.5   02     : 22   0102   :  1              1st Qu.:2.154e-05   1st Qu.:0.02378   epsg:4326    :  0
 Median :128.0   03     : 22   0103   :  1              Median :3.477e-05   Median :0.03172   +proj=long...:  0
 Mean   :128.0   04     : 22   0104   :  1              Mean   :5.379e-05   Mean   :0.03705
 3rd Qu.:191.5   05     : 22   0105   :  1              3rd Qu.:5.912e-05   3rd Qu.:0.04138
 Max.   :255.0   06     : 21   0106   :  1              Max.   :6.857e-04   Max.   :0.22637
                 (Other):124   (Other):249
```

Figure D14: Boston voting precincts summary statistics before data-cleaning

Philadelphia

```
> summary(PH_crime)
    point_y         objectid           dc_dist       psa          dispatch_d
 Min.   :39.88   Min.   :21873294   24     : 24625   1: 93254   Min.   :2018-01-02
 1st Qu.:39.96   1st Qu.:26611686   19     : 22288   2:110369   1st Qu.:2018-06-24
 Median :39.99   Median :31073146   15     : 21960   3: 84791   Median :2018-12-07
 Mean   :39.99   Mean   :31126671   22     : 20590   4: 14311   Mean   :2018-12-09
 3rd Qu.:40.03   3rd Qu.:35883819   35     : 19582   A:  1217   3rd Qu.:2019-06-01
 Max.   :40.14   Max.   :38937728   25     : 17833              Max.   :2019-11-01
 NA's   :2540                       (Other):177064
      dispatch_1         hour_              dc_key                                location_b
 2019-09-18:   615   Min.   : 0.00   199814060305:     1   5200 BLOCK FRANKFORD AVE      :   876
 2018-08-06:   613   1st Qu.:10.00   201801000080:     1   1400 BLOCK JOHN F KENNEDY BLVD:   531
 2019-10-02:   609   Median :14.00   201801000119:     1   5900 BLOCK MARKET ST          :   527
 2018-05-07:   608   Mean   :13.46   201801000120:     1   5900 BLOCK N BROAD ST         :   492
 2018-06-15:   599   3rd Qu.:18.00   201801000132:     1   1300 BLOCK MARKET ST          :   482
 2019-09-26:   599   Max.   :23.00   201801000147:     1   0 BLOCK N 52ND ST             :   446
 (Other)   :300299   NA's   :3214    (Other)     :303936   (Other)                       :300588
```

Figure D15: Philadelphia crime summary statistics before data-cleaning

```
> summary(PH_trashcans_bb)
    OBJECTID        SERIALNUM                                   DESCRIPTIO  RECYCLER          geometry
 Min.   :  1.0   1400004:  1   4910 Botanic Ave                  :  2   no :599   POINT        :968
 1st Qu.:242.8   1400007:  1   4910 Botanic Ave (Unknown Location):  2   yes:369   epsg:4326    :  0
 Median :484.5   1400012:  1   4910 Botanic Ave. (Unknown Station):  2             +proj=long...:  0
 Mean   :484.5   1400013:  1   Ne 4th/Chestnut                   :  2
 3rd Qu.:726.2   1400014:  1   Se Front/Chestnut                 :  2
 Max.   :968.0   1400015:  1   Sw 15th/Pine                      :  2
                 (Other):962   (Other)                           :956
```

Figure D16: Philadelphia Big Belly trash can summary statistics before data-cleaning

```
> summary(PH_trashcans_w)
    OBJECTID                                ADDRESS     CORNERSIDE          geometry
 Min.   :  1.0   4300 Kensington Ave.        :  4   N W :  1   POINT        :352
 1st Qu.:111.8   4600 Frankford Ave.         :  4   NE  : 80   epsg:4326    :  0
 Median :199.5   4700 Frankford Ave.         :  4   NW  :107   +proj=long...:  0
 Mean   :199.3   5000 Frankford Ave.         :  4   S   :  1
 3rd Qu.:287.2   5th & Cayuga St.            :  4   SE  : 87
 Max.   :386.0   Huntingdon & Germantown Ave.:  4   SW  : 64
                 (Other)                     :328   NA's: 12
```

Figure D17: Philadelphia wire waste basket trash can summary statistics before data-cleaning

```
> summary(PH_streetlights)
    OBJECTID          OID        POLE_NUM          TYPE            NLUMIN           LUM_SIZE         HEIGHT
 Min.   :     1   Min.   :0   Min.   :100001   WP     :102263   Min.   :   0.00   Min.   :  0.0   Min.   : 0.00
 1st Qu.: 48688   1st Qu.:0   1st Qu.:223548   SNP    : 24349   1st Qu.:   1.00   1st Qu.:100.0   1st Qu.:25.00
 Median : 97376   Median :0   Median :415330   AEL    : 18110   Median :   1.00   Median :100.0   Median :25.00
 Mean   : 97376   Mean   :0   Mean   :412728   C20    :  8553   Mean   :   1.14   Mean   :120.5   Mean   :23.37
 3rd Qu.:146063   3rd Qu.:0   3rd Qu.:600240   SLA    :  6432   3rd Qu.:   1.00   3rd Qu.:120.0   3rd Qu.:25.00
 Max.   :194750   Max.   :0   Max.   :886047   C13    :  6093   Max.   :3614.00   Max.   :400.0   Max.   :45.00
                                               (Other): 28950   NA's   :68412     NA's   :68412   NA's   :68412
                    POLE_DATE                          UP_DATE                OWNER            TAP_ID            BLOCK
 2003-06-24T00:00:00.000Z:  8792   2003-06-24T00:00:00.000Z:  8792   Streets:168451   Min.   :     0   27W32NC:    12
 1997-11-13T00:00:00.000Z:  1433   1997-11-13T00:00:00.000Z:  1433   PECO   : 23540   1st Qu.:     0   65W11NB:    12
 1997-11-12T00:00:00.000Z:  1380   1997-11-12T00:00:00.000Z:  1380   PennDOT:   952   Median :220788   27E43NH:    11
 1997-07-30T00:00:00.000Z:  1338   1997-07-30T00:00:00.000Z:  1338   Private:   841   Mean   :270312   20E46NB:    10
 1997-09-03T00:00:00.000Z:  1330   1997-09-03T00:00:00.000Z:  1330   CCD    :   312   3rd Qu.:509252   01E66GA:     9
 (Other)                  :178972   (Other)                  :178972   (Other):   640   Max.   :883341   (Other): 18053
 NA's                     :  1505   NA's                     :  1505   NA's   :    14   NA's   :5676     NA's   :176643
     PLATE                    geometry
 3A-2   :  1414   POINT        :194750
 5A-4   :  1193   epsg:4326    :     0
 3B-1   :  1127   +proj=long...:     0
 4--2   :  1102
 3--4   :   966
 (Other): 12305
 NA's   :176643
```

Figure D18: Philadelphia street light summary statistics before data-cleaning

```
> summary(PH_trees)
    OBJECTID       SPECIES         STATUS          DBH                    geometry
 Min.   :     1   NA's:111993   NA's:111993   NA's:111993   POINT        :111993
 1st Qu.: 27999                                             epsg:4326    :     0
 Median : 55997                                             +proj=long...:     0
 Mean   : 55997
 3rd Qu.: 83995
 Max.   :111993
```

Figure D19: Philadelphia tree summary statistics before data-cleaning

```
> summary(PH_prcnts)
    OBJECTID        SHORT_DIV_     DIVISION_N     Shape__Are         Shape__Len            geometry
 Min.   :   1.0   02      :  66   0101   :   1   Min.   :   19911   Min.   :  631   POLYGON      :1703
 1st Qu.: 426.5   03      :  66   0102   :   1   1st Qu.:  120201   1st Qu.: 1585   epsg:4326    :   0
 Median : 852.0   04      :  66   0103   :   1   Median :  177517   Median : 1995   +proj=long...:   0
 Mean   : 852.0   05      :  66   0104   :   1   Mean   :  369655   Mean   : 2481
 3rd Qu.:1277.5   06      :  66   0105   :   1   3rd Qu.:  323165   3rd Qu.: 2766
 Max.   :1703.0   07      :  66   0106   :   1   Max.   :25833264   Max.   :26923
                  (Other):1307   (Other):1697
```

Figure D20: Philadelphia voting precincts summary statistics before data-cleaning

Washington D.C.

```
> summary(DC_crime_2018)
       CCN                          REPORT_DAT          SHIFT          METHOD                                  OFFENSE
 18001724:    2   2018-05-03T00:00:00.000Z:    3   DAY     :12150   GUN    : 1598   THEFT/OTHER                :14227
 18066272:    2   2018-05-04T00:00:00.000Z:    3   EVENING :14394   KNIFE :   772   THEFT F/AUTO               :11609
 18067093:    2   2018-05-08T00:00:00.000Z:    3   MIDNIGHT: 7239   OTHERS:31413   MOTOR VEHICLE THEFT        : 2393
 18067757:    2   2018-09-23T00:00:00.000Z:    3                                   ROBBERY                    : 2024
 18070920:    2   2018-09-25T00:00:00.000Z:    3                                   ASSAULT W/DANGEROUS WEAPON: 1672
 18085163:    2   2018-01-04T01:03:43.000Z:    2                                   BURGLARY                   : 1419
 (Other) :33771   (Other)                 :33766                                   (Other)                    :  439
                                   BLOCK            XBLOCK           YBLOCK            WARD            ANC            DISTRICT
 3100 - 3299 BLOCK OF 14TH STREET NW      :  258   Min.   :390262   Min.   :127432   2      :6259   1B     : 2285   2      :6492
 900 - 999 BLOCK OF RHODE ISLAND AVENUE NE:  180   1st Qu.:397228   1st Qu.:136303   6      :5637   2B     : 1789   3      :6243
 1 - 99 BLOCK OF H STREET NW              :  168   Median :398809   Median :137731   5      :5010   5C     : 1524   5      :4918
 1200 - 1229 BLOCK OF 1ST STREET NE       :  145   Mean   :399228   Mean   :137802   1      :4632   1A     : 1502   1      :4748
 2800 - 2999 BLOCK OF ALABAMA AVENUE SE   :  132   3rd Qu.:401157   3rd Qu.:139677   7      :4076   2C     : 1375   6      :4442
 600 - 699 BLOCK OF H STREET NE           :  127   Max.   :407806   Max.   :147307   8      :3125   6B     : 1350   (Other):6935
 (Other)                                  :32773                                    (Other):5044   (Other):23958   NA's   :   5
      PSA            NEIGHBORHO       BLOCK_GROU       CENSUS_TRA          VOTING_PRE        LATITUDE        LONGITUDE
 305    : 1292   Cluster 2 : 2457   005800 1:  886   005800 : 1065   Precinct 129: 1525   Min.   :38.81   Min.   :-77.11
 207    : 1286   Cluster 8 : 2366   004400 2:  759   004400 : 1057   Precinct 17 : 1062   1st Qu.:38.89   1st Qu.:-77.03
 206    : 1192   Cluster 25: 1889   010600 2:  680   010700 :  777   Precinct 83 : 1001   Median :38.91   Median :-77.01
 208    : 1134   Cluster 3 : 1767   010700 1:  522   010600 :  770   Precinct 72 :  823   Mean   :38.91   Mean   :-77.01
 302    : 1024   Cluster 6 : 1726   008803 1:  513   011100 :  621   Precinct 22 :  734   3rd Qu.:38.92   3rd Qu.:-76.99
 (Other):27850   (Other)   :23200   (Other) :30351   (Other):29421   Precinct 37 :  673   Max.   :38.99   Max.   :-76.91
 NA's   :    5   NA's      :  378   NA's    :   72   NA's   :   72   (Other)     :27965
```

Figure D21: Washington D.C. crime 2018 summary statistics before data-cleaning

```
> summary(DC_crime_2019)
       CCN                          REPORT_DAT          SHIFT          METHOD                                  OFFENSE
 19015227:    3   2019-01-26T00:00:00.000Z:    4   DAY     : 9597   GUN    : 1350   THEFT/OTHER                :12960
 19002844:    2   2019-07-20T00:00:00.000Z:    4   EVENING :12406   KNIFE :   574   THEFT F/AUTO               : 8917
 19007330:    2   2019-01-06T00:00:00.000Z:    3   MIDNIGHT: 6494   OTHERS:26573   MOTOR VEHICLE THEFT        : 1943
 19086306:    2   2019-01-13T00:00:00.000Z:    3                                   ROBBERY                    : 1899
 19114552:    2   2019-10-09T00:00:00.000Z:    3                                   ASSAULT W/DANGEROUS WEAPON: 1358
 19119050:    2   2019-01-05T00:00:00.000Z:    2                                   BURGLARY                   : 1092
 (Other) :28484   (Other)                 :28478                                   (Other)                    :  328
                                   BLOCK            XBLOCK           YBLOCK            WARD            ANC            DISTRICT
 3100 - 3299 BLOCK OF 14TH STREET NW      :  195   Min.   :390103   Min.   :127956   2      :6004   1B     : 2004   2      :6147
 900 - 999 BLOCK OF RHODE ISLAND AVENUE NE:  151   1st Qu.:397174   1st Qu.:136400   6      :4823   2B     : 1716   3      :5537
 1000 - 1249 BLOCK OF BRENTWOOD ROAD NE   :  130   Median :398529   Median :137708   5      :4236   2C     : 1452   5      :4306
 2800 - 2999 BLOCK OF ALABAMA AVENUE SE   :  120   Mean   :399113   Mean   :137748   1      :4036   5C     : 1336   1      :4003
 1200 - 1229 BLOCK OF 1ST STREET NE       :  119   3rd Qu.:401090   3rd Qu.:139501   7      :3154   1A     : 1259   6      :3459
 5300 - 5399 BLOCK OF WISCONSIN AVENUE NW :  119   Max.   :407806   Max.   :147307   8      :2402   2F     : 1154   (Other):5023
 (Other)                                  :27663                                    (Other):3842   (Other):19576   NA's   :  22
      PSA            NEIGHBORHO       BLOCK_GROU       CENSUS_TRA          VOTING_PRE        LATITUDE        LONGITUDE
 207    : 1232   Cluster 8 : 2146   005800 1:  929   005800 : 1044   Precinct 129: 1641   Min.   :38.82   Min.   :-77.11
 305    : 1191   Cluster 2 : 2051   004400 2:  730   004400 :  969   Precinct 17 : 1043   1st Qu.:38.90   1st Qu.:-77.03
 208    : 1187   Cluster 25: 1724   010600 2:  566   010700 :  790   Precinct 83 :  825   Median :38.91   Median :-77.02
 206    :  991   Cluster 6 : 1638   010700 1:  523   010600 :  635   Precinct 72 :  744   Mean   :38.91   Mean   :-77.01
 209    :  972   Cluster 3 : 1631   009102 2:  481   009102 :  560   Precinct 22 :  602   3rd Qu.:38.92   3rd Qu.:-76.99
 (Other):22902   (Other)   :18941   (Other) :25219   (Other):24450   Precinct 37 :  551   Max.   :38.99   Max.   :-76.91
 NA's   :   22   NA's      :  366   NA's    :   49   NA's   :   49   (Other)     :23091
```

Figure D22: Washington D.C. crime 2019 summary statistics before data-cleaning

```
> summary(DC_trashcans)
    LITTERID      LITTERID_1        MAPX            MAPY            LINK_ID                TYPE         BUSSTOP            NAME
 Min.   :   0   Min.   :   0   Min.   :     0   Min.   :     0   Min.   :       0   1       :5636   Min.   :0.0000   TQ     :5821
 1st Qu.:   0   1st Qu.:   0   1st Qu.:     0   1st Qu.:     0   1st Qu.:       0   2       : 147   1st Qu.:0.0000   MBR    : 174
 Median :1673   Median :1615   Median :397153   Median :136644   Median :    9549   3       :   8   Median :1.0000   YXF    : 146
 Mean   :2004   Mean   :1978   Mean   :277224   Mean   : 95998   Mean   :  172802   Concrete:  62   Mean   :0.5216   KEO    : 130
 3rd Qu.:3632   3rd Qu.:3620   3rd Qu.:399661   3rd Qu.:139213   3rd Qu.:   20873   Steel   : 614   3rd Qu.:1.0000   BXM    : 115
 Max.   :5790   Max.   :5790   Max.   :407747   Max.   :147074   Max.   :45089020   NA's    : 236   Max.   :1.0000   (Other): 287
                                                                                                                        NA's   :  30
                   CREATEDAT                                 GlobalID                        LINERCHG        OWNER
 2004-02-24T00:00:00.000Z: 283   {001AEF24-9B3F-4AB7-95C2-FDA202C991AB}:   1   2015-11-24T00:00:00.000Z: 214   1      :4799
 2004-03-01T00:00:00.000Z: 264   {00277216-A6F0-4C40-A0F8-4DCA6E7E510C}:   1   2017-01-11T00:00:00.000Z:  50   10     : 296
 2004-02-23T00:00:00.000Z: 262   {0027826C-CFF0-408A-BC93-13809A311A48}:   1   2015-12-01T00:00:00.000Z:  41   P      : 202
 2004-03-03T00:00:00.000Z: 251   {00367BE9-8CA2-4C50-9240-1BB9C6027F11}:   1   2017-01-12T00:00:00.000Z:  25   7      : 201
 2004-03-02T00:00:00.000Z: 249   {00424BFE-0AAE-49FE-AC80-3F89561F4EAF}:   1   2015-06-24T00:00:00.000Z:  23   11     : 138
 (Other)                  :3402   {0042FBA7-F0C3-4186-94F4-E19422C01926}:   1   (Other)                 :  53   (Other): 322
 NA's                     :1992   (Other)                               :6697   NA's                    :6297   NA's   : 745
 LINEREA                LCAN_Repai                    Rat_Cover                      Powerwash                LCAN_Damag
 1   : 337   bolted/tightened:   4   2017-01-12T00:00:00.000Z:  10   2015-07-23T00:00:00.000Z:   1   broken bolts       :   7
 2   :  20   door replaced   :   3   2017-01-18T00:00:00.000Z:  10   2015-07-28T00:00:00.000Z:   1   dented             :  19
 3   :   3   welded          :   4   2017-01-11T00:00:00.000Z:   4   2015-09-03T00:00:00.000Z:   1   missing/broken bars:   2
 4   : 100   NA's            :6692   1899-12-30T00:00:00.000Z:   1   2017-01-12T00:00:00.000Z:   1   NA's               :6675
 NA's:6243                           2015-07-23T00:00:00.000Z:   1   2017-01-18T00:00:00.000Z:   1
                                     (Other)                  :   4   NA's                     :6698
                                     NA's                     :6673
```

Figure D23: Washington D.C. trash can summary statistics before data-cleaning

```
> summary(DC_streetlights)
          ADDEDBY                          ADDTOGIS       ARMLENGTH1      ARMLENGTH2              ARMSTYLE_D     ASSETSTATU
 Osmose       :62441   2007-05-09T00:00:00.000Z: 1235   8      :29033   N/A    :70076   Regular Alley  :14797   Active:70829
 Tugril DBI   : 5135   2007-05-15T00:00:00.000Z: 1149   3      :16996   8      :  371   WP Arm with Guy:14378
 SEVERN       : 1623   2007-04-11T00:00:00.000Z: 1097   N/A    :13841   4      :  153   N/A            :13855
 JBLEVINS     :  888   2007-05-17T00:00:00.000Z: 1093   6      :10013   3      :  115   Simple Pendant :12323
 W. A. Chester:  668   2007-05-01T00:00:00.000Z: 1083   12     :  503   6      :   53   Simple WP      : 6205
 TNGUYEN      :   52   (Other)                  :65160   4      :  165   (Other):   58   Wood pole Truss: 4438
 (Other)      :   22   NA's                     :   12   (Other):  278   NA's   :    3   (Other)        : 4833
            ASSETTYPE      CCT_DESC                          CONDITION_      CROSSSTREE      FACILITYID     FEEDMANHOL
 Street Light    :54076   2700:   47   Good Condition            : 6193   14TH ST: 1505   60588  :    2   DDOT :  989
 Alley Light     :14828   3000: 1633   Need Immediate Replacement:   20   13TH ST: 1201   71591DC:    2   N/A  :35209
 Bridge Light    :  987   4000: 4122   Need Inspection in 2 years:  637   16TH ST: 1045   066E201:    1   PEPCO:   43
 Bike Trail Light:  453   4500:   32   Replace Within a Year     :   21   7TH ST :  972   066E203:    1   NA's :34588
 OHGS            :  144   5000: 1687   NA's                      :63958   4TH ST :  915   066E205:    1
 Pedestrian Light:  134   N/A :62926                                       17TH ST:  875   (Other):70816
 (Other)         :  207   NA's:  382                                       (Other):64316   NA's   :    6
                FIXTURESTY                                        GLOBALID        HOUSENO                          INOPERATIO
 Cobrahead Cutoff     :24724   {000118AF-1633-4D57-B084-357AA26CDEB7}:    1   0      :16668   2017-09-01T00:00:00.000Z:  208
 Cobrahead Semi Cutoff:18427   {00013D3B-9F6E-4668-91EE-CF3F106C1988}:    1   200    :  267   2013-10-01T00:00:00.000Z:  155
 Posttop              :13294   {00019D43-2C2B-4D74-89F7-2F5CEC6B102D}:    1   100    :  257   2013-05-03T00:00:00.000Z:  103
 Standard LED         : 5868   {000222B6-997C-4188-81BE-456B963AFE7F}:    1   600    :  247   2010-11-11T00:00:00.000Z:   84
 Teardrop             : 3666   {00025E79-C3C0-4B9A-9F8F-4F76FF4D78E8}:    1   300    :  231   2010-01-05T00:00:00.000Z:   80
 (Other)              : 4843   {0002FF89-C2ED-4636-99D4-E1A08C5C27C1}:    1   900    :  218   (Other)                  : 4130
 NA's                 :    7   (Other)                                :70823   (Other):52941   NA's                     :66069
```

Figure D24: Washington D.C. street light summary statistics before data-cleaning

```
> summary(DC_trees)
    OBJECTID                          FACILITYID                                                    VICINITY          WARD
 Min.   :36593856   10340-155-3001-0095-000:   126   Lansburgh Park 1098 Delaware Ave SW        :   212   Min.   :   1.000
 1st Qu.:36640950   35778-290-3005-0057-000:    94   OPP 1200 MISSISSIPPI AVE SE (OXON RUN PARK):   164   1st Qu.:   3.000
 Median :36687883   2                      :    60   Unit Blk Potomac Ave SW                    :   163   Median :   5.000
 Mean   :36687883   24875-000-3001-0000-000:    48   4200 DALECARLIA PKWY NW                    :   158   Mean   :   4.768
 3rd Qu.:36734816   15349-260-3001-0214-000:    32   IRVING TRAFFIC CIRCLE (SW)                 :   153   3rd Qu.:   6.000
 Max.   :36781750   (Other)                :152446   (Other)                                    :183004   Max.   :9006.000
                    NA's                   : 34929   NA's                                       :  3881   NA's   :573
     TBOX_L            TBOX_W                WIRES                CURB              SIDEWALK          TBOX_STAT
 Min.   :    0.00   Min.   :    0.00   Both        : 45916   Flexipave:     1   Flexipave:   205   Plant   :161246
 1st Qu.:   99.00   1st Qu.:    4.00   High Voltage:   627   None     : 27449   None     : 39448   Open    :  9579
 Median :   99.00   Median :    5.00   Low Voltage : 10068   Permanent:154757   Permanent:143581   Retired :  7327
 Mean   :   80.87   Mean   :   15.64   None        :127183   Temporary:  1690   Temporary:   515   Proposed:  5374
 3rd Qu.:   99.00   3rd Qu.:    6.00   NA's        :  3941   NA's     :  3838   NA's     :  3986   Conflict:  2419
 Max.   :10000.00   Max.   :10000.00                                                              (Other) :  1688
 NA's   :17278      NA's   :17241                                                                 NA's    :   102
                    RETIREDDT                    SCI_NM                    CMMN_NM                    DATE_PLANT
 1899-12-30T00:00:00.000Z:   256   Acer rubrum      : 12662   Red maple        : 12485   1899-12-30T00:00:00.000Z:   419
 2010-02-03T00:00:00.000Z:    30   Quercus phellos  :  9863   Willow oak       :  9725   2005-11-01T00:00:00.000Z:   363
 2016-03-21T00:00:00.000Z:    30   Quercus palustris:  8986   Other (See Notes):  9677   2006-03-01T00:00:00.000Z:   304
 2015-12-11T00:00:00.000Z:    25   Other (See Notes):  6831   Pin oak          :  8969   2006-02-01T00:00:00.000Z:   186
 2019-08-01T16:09:57.000Z:    23   Quercus rubra    :  5882   American elm     :  5813   2016-04-01T00:00:00.000Z:   147
 (Other)                  :  4843   (Other)          :125326   (Other)          :120820   (Other)                  : 74091
 NA's                     :182528   NA's             : 18185   NA's             : 20246   NA's                     :112225
```

Figure D25: Washington D.C. tree summary statistics before data-cleaning

```
> summary(DC_prcnts)
  OBJECTID_1       OBJECTID                NAME                                                                  WEB_URL      SHAPE_Leng
 Min.   :  1.0   Min.   :  1.0   Precinct 1  :  1   http://www.dcboee.org/voter_info/find_pollingplace/:143   Min.   : 1864
 1st Qu.: 36.5   1st Qu.: 36.5   Precinct 10 :  1                                                            1st Qu.: 3485
 Median : 72.0   Median : 72.0   Precinct 100:  1                                                            Median : 4463
 Mean   : 72.0   Mean   : 71.6   Precinct 101:  1                                                            Mean   : 4983
 3rd Qu.:107.5   3rd Qu.:106.5   Precinct 102:  1                                                            3rd Qu.: 5554
 Max.   :143.0   Max.   :142.0   Precinct 103:  1                                                            Max.   :24090
                                 (Other)     :137
   SHAPE_Area                 geometry
 Min.   :  169201   MULTIPOLYGON :143
 1st Qu.:  587368   epsg:4326    :  0
 Median :  865910   +proj=long...:  0
 Mean   : 1240952
 3rd Qu.: 1361879
 Max.   :10101460
```

Figure D26: Washington D.C. voting precincts summary statistics before data-cleaning

Appendix E

Sample figures of dataset visualization prior to data-cleaning of San Francisco

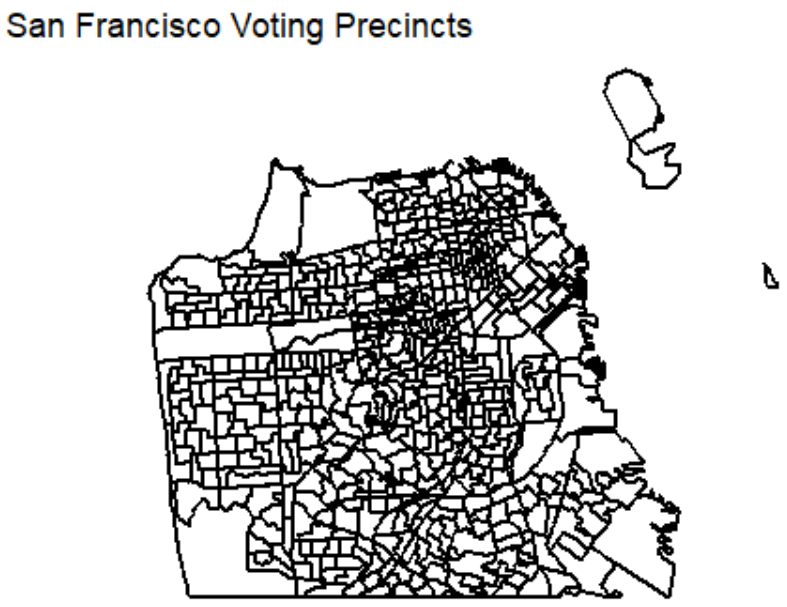

Figure E1: San Francisco by voting precincts before data-cleaning

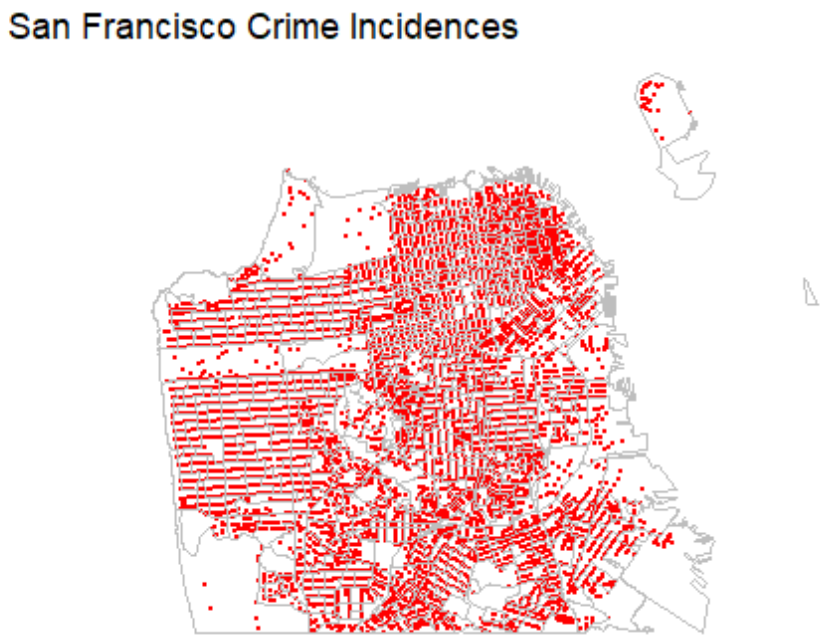

Figure E2: San Francisco crime by voting precincts before data-cleaning

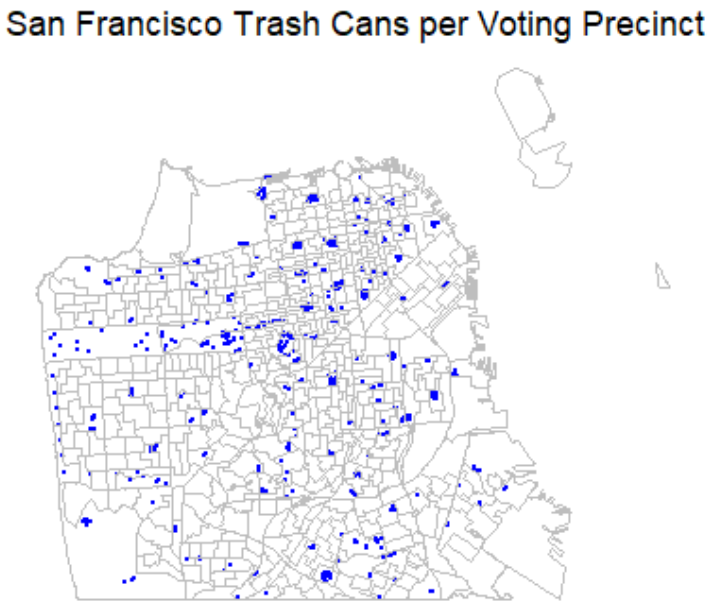

Figure E3: San Francisco trash cans by voting precincts before data-cleaning

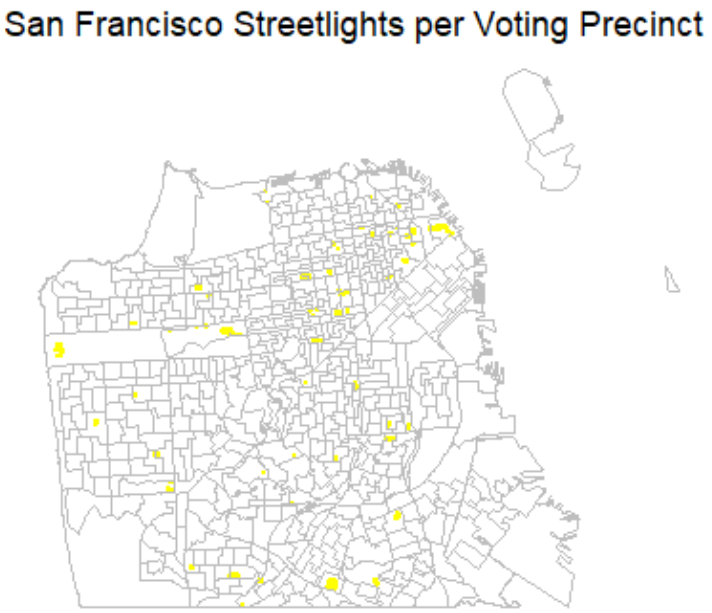

Figure E4: San Francisco street lights by voting precincts before data-cleaning

Crime and public infrastructure in coastal US cities

To what extent do elements of the urban fabric influence crime rates in major US cities?

Internationally, crime statistics consistently show that coastal cities have higher crime rates than inland cities. In particular, from Google trend datasets across six Mediterranean and three Atlantic coastal cities, we found that the coastal cities are more likely to be associated with crime than inland cities. Although this study does not include any inland cities, it is worth noting that the coastal cities in the Mediterranean Sea and Atlantic Ocean are all famous for being tourist destinations, and that the Atlantic coastal cities are also famous for being cultural centres. Therefore, the hypothesis that tourism affects coastal crime rate comes from a broad stereotypical view of tourists being more vulnerable to crime than locals.

In order to better understand if crime rates are influenced by tourism, we ask the question "Is there a relationship between the number of tourists and crime rate in coastal US cities?". Since the data for tourist numbers in US cities is not readily available, we use the number of visitors to US national parks as a proxy for tourist numbers. Since these cities have the highest number of visits are associated with a tourist destination, we believe that the number of visitors to US national parks is a good proxy for the number of tourists in US coastal cities.

In order to assist the Organisation for Economic Cooperation and Development (OECD) in enhancing anti-crime policies, this study also investigates the relationship between crime rate and other factors such as the number of police officers, the number of restaurants, the number of hotels, the number of bars, the number of cafes, Gross Domestic Product (GDP), and the number of public transport stations. The study focuses on the cities of San Francisco, Los Angeles, and New York City, since these cities have the highest number of visits and contribute the most to the overall tourist count, which allows us to examine the relationship between tourism and crime rate.

After combining and cleaning the national park visitation data, the OECD crime statistics, and the Yelp data for each city, we found that there is a positive correlation between the number of tourists and the crime rate in the cities of San Francisco and New York City. In contrast, we found that the number of police officers, the number of restaurants, the number of hotels, the number of bars, the number of cafes, GDP, and the number of public transport stations do not have a strong correlation with crime rate. Although we do not have a causal explanation for the observed correlations, we believe that the results suggest that tourism may be a factor in the crime rate of coastal US cities. Therefore, we recommend that the Gross Office and the Secretaries for Public Governance may further investigate the relationship between tourism and crime rate in coastal US cities. Although we do not have a causal explanation for the observed correlation, we believe that the results suggest that tourism may be a factor in the crime rate of coastal US cities. Therefore, we recommend that the Gross Office and the Secretaries for Public Governance may further investigate the relationship between tourism and crime rate in coastal US cities. Although we do not have a causal explanation for the observed correlation, we believe that the results suggest that tourism may be a factor in the crime rate of coastal US cities. Therefore, we recommend that the Gross Office and the Secretaries for Public Governance may further investigate the relationship between tourism and crime rate in coastal US cities.

In order to better understand the relationship between tourism and crime rate, we propose that future research should include other factors that may be associated with crime rate, such as the number of tourists in other cities, the number of police officers in other cities, the number of restaurants in other cities, the number of hotels in other cities, the number of bars in other cities, the number of cafes in other cities, GDP in other cities, and the number of public transport stations in other cities. With the use of a potential model, such analysis also examines if the relations of such elements hold true, which would be a good topic for further research.

Although we do not have a causal explanation for the observed correlation between tourism and crime rate in coastal US cities, we believe that the results suggest that tourism may be a factor in the crime rate of coastal US cities. Therefore, we recommend that the Gross Office and the Secretaries for Public Governance may further investigate the relationship between tourism and crime rate in coastal US cities. Although we do not have a causal explanation for the observed correlation, we believe that the results suggest that tourism may be a factor in the crime rate of coastal US cities. Therefore, we recommend that the Gross Office and the Secretaries for Public Governance may further investigate the relationship between tourism and crime rate in coastal US cities.

Part 1: Data Preprocessing

```
!pip install -q -U google-cloud-bigquery
!pip install -q -U google-cloud-bigquery-magics
!pip install -q -U pandas-gbq
!pip install -q -U geopandas
!pip install -q -U shapely
!pip install -q -U pyproj
!pip install -q -U fiona
!pip install -q -U rtree
!pip install -q -U pygeos
!pip install -q -U geopandas-view
!pip install -q -U folium
!pip install -q -U matplotlib-venn
!apt-get -qq install -q libspatialindex-dev
!pip install -q -U descartes
!pip install -q -U mapclassify
!pip install -q -U keplergl
!pip install -q -U pydeck
```

1.1 San Francisco

```
# Set up environments
import os
os.environ['GOOGLE_APPLICATION_CREDENTIALS'] = '/content/drive/MyDrive/cmdl-data-fall-2021-ucla-f7ef1c7a8ef7.json'
os.environ['GCLOUD_PROJECT'] = 'cmdl-data-fall-2021-ucla'
os.environ['GOOGLE_CLOUD_PROJECT'] = 'cmdl-data-fall-2021-ucla'
os.environ['GCS_BUCKET'] = 'cmdl-data-fall-2021-ucla-bucket'
os.environ['BIGQUERY_STORAGE_PROJECT'] = 'cmdl-data-fall-2021-ucla'
os.environ['BIGQUERY_STORAGE_DATASET'] = 'cmdl_data_fall_2021_ucla'
os.environ['GCS_PREFIX'] = 'cmdl-data-fall-2021-ucla/data'
```

```
# Authenticate
from google.colab import auth
auth.authenticate_user()
```

```
# Load the extension
%load_ext google.cloud.bigquery
%load_ext google.cloud.bigquery_magics
%load_ext rpy2.ipython
%load_ext google.colab.data_table
%load_ext google.colab.drive
%load_ext google.colab.files
%load_ext google.colab.auth
%load_ext google.colab.output
%load_ext google.colab.patches
%load_ext google.colab.widgets
%load_ext google.colab.vertex_ai
%load_ext google.colab.gsheets
%load_ext google.colab.kernel
%load_ext google.colab.interactive_table
%load_ext google.colab.data_table_extension
%load_ext google.colab.dataflow
%load_ext google.colab.dataproc
%load_ext google.colab.bigquery
%load_ext google.colab.bigquery_magics
%load_ext google.colab.bigquery_magics_v2
%load_ext google.colab.bigquery_magics_v3
%load_ext google.colab.bigquery_magics_v4
%load_ext google.colab.bigquery_magics_v5
%load_ext google.colab.bigquery_magics_v6
%load_ext google.colab.bigquery_magics_v7
%load_ext google.colab.bigquery_magics_v8
%load_ext google.colab.bigquery_magics_v9
%load_ext google.colab.bigquery_magics_v10
%load_ext google.colab.bigquery_magics_v11
%load_ext google.colab.bigquery_magics_v12
%load_ext google.colab.bigquery_magics_v13
%load_ext google.colab.bigquery_magics_v14
%load_ext google.colab.bigquery_magics_v15
%load_ext google.colab.bigquery_magics_v16
%load_ext google.colab.bigquery_magics_v17
%load_ext google.colab.bigquery_magics_v18
%load_ext google.colab.bigquery_magics_v19
%load_ext google.colab.bigquery_magics_v20
%load_ext google.colab.bigquery_magics_v21
%load_ext google.colab.bigquery_magics_v22
%load_ext google.colab.bigquery_magics_v23
%load_ext google.colab.bigquery_magics_v24
%load_ext google.colab.bigquery_magics_v25
%load_ext google.colab.bigquery_magics_v26
%load_ext google.colab.bigquery_magics_v27
%load_ext google.colab.bigquery_magics_v28
%load_ext google.colab.bigquery_magics_v29
%load_ext google.colab.bigquery_magics_v30
%load_ext google.colab.bigquery_magics_v31
%load_ext google.colab.bigquery_magics_v32
%load_ext google.colab.bigquery_magics_v33
%load_ext google.colab.bigquery_magics_v34
%load_ext google.colab.bigquery_magics_v35
%load_ext google.colab.bigquery_magics_v36
%load_ext google.colab.bigquery_magics_v37
%load_ext google.colab.bigquery_magics_v38
%load_ext google.colab.bigquery_magics_v39
%load_ext google.colab.bigquery_magics_v40
%load_ext google.colab.bigquery_magics_v41
%load_ext google.colab.bigquery_magics_v42
%load_ext google.colab.bigquery_magics_v43
%load_ext google.colab.bigquery_magics_v44
%load_ext google.colab.bigquery_magics_v45
%load_ext google.colab.bigquery_magics_v46
%load_ext google.colab.bigquery_magics_v47
%load_ext google.colab.bigquery_magics_v48
%load_ext google.colab.bigquery_magics_v49
%load_ext google.colab.bigquery_magics_v50
%load_ext google.colab.bigquery_magics_v51
%load_ext google.colab.bigquery_magics_v52
%load_ext google.colab.bigquery_magics_v53
%load_ext google.colab.bigquery_magics_v54
%load_ext google.colab.bigquery_magics_v55
%load_ext google.colab.bigquery_magics_v56
%load_ext google.colab.bigquery_magics_v57
%load_ext google.colab.bigquery_magics_v58
%load_ext google.colab.bigquery_magics_v59
%load_ext google.colab.bigquery_magics_v60
%load_ext google.colab.bigquery_magics_v61
%load_ext google.colab.bigquery_magics_v62
%load_ext google.colab.bigquery_magics_v63
%load_ext google.colab.bigquery_magics_v64
%load_ext google.colab.bigquery_magics_v65
%load_ext google.colab.bigquery_magics_v66
%load_ext google.colab.bigquery_magics_v67
%load_ext google.colab.bigquery_magics_v68
%load_ext google.colab.bigquery_magics_v69
%load_ext google.colab.bigquery_magics_v70
%load_ext google.colab.bigquery_magics_v71
%load_ext google.colab.bigquery_magics_v72
%load_ext google.colab.bigquery_magics_v73
%load_ext google.colab.bigquery_magics_v74
%load_ext google.colab.bigquery_magics_v75
%load_ext google.colab.bigquery_magics_v76
%load_ext google.colab.bigquery_magics_v77
%load_ext google.colab.bigquery_magics_v78
%load_ext google.colab.bigquery_magics_v79
%load_ext google.colab.bigquery_magics_v80
%load_ext google.colab.bigquery_magics_v81
%load_ext google.colab.bigquery_magics_v82
%load_ext google.colab.bigquery_magics_v83
%load_ext google.colab.bigquery_magics_v84
%load_ext google.colab.bigquery_magics_v85
%load_ext google.colab.bigquery_magics_v86
%load_ext google.colab.bigquery_magics_v87
%load_ext google.colab.bigquery_magics_v88
%load_ext google.colab.bigquery_magics_v89
%load_ext google.colab.bigquery_magics_v90
%load_ext google.colab.bigquery_magics_v91
%load_ext google.colab.bigquery_magics_v92
%load_ext google.colab.bigquery_magics_v93
%load_ext google.colab.bigquery_magics_v94
%load_ext google.colab.bigquery_magics_v95
%load_ext google.colab.bigquery_magics_v96
%load_ext google.colab.bigquery_magics_v97
%load_ext google.colab.bigquery_magics_v98
%load_ext google.colab.bigquery_magics_v99
%load_ext google.colab.bigquery_magics_v100
```

```
# Load the data
```

Appendix 3

Crime and public infrastructure in coastal US cities

To what extent do elements of the urban fabric influence crime rates in major US cities?

Part 2: Data Visualization

```
In [1]: # Basic event by setting set directory and importing relevant packages
        setwd("C:/Users/Aniruddha/Desktop/Ani_Work/DS_Data/DDS-Crime/Crime/")
        # Importing relevant packages for data manipulation, "ggplot2", "tibble"
        # Alternative required for this project
        library(ggplot2)
        library(dplyr)
        library(magrittr)
        library(tibble)
        library(tidyverse)
        library(tidyr)
        library(ggmap)
        library(RCurl)
        library(RJSONIO)
        library(stringr)
        library(plyr)
        library(ggsci)
        library(ggthemes)
        library(maps)
        library(mapproj)
        library(mapdata)
        library(rvest)
        library(XML)
        library(jsonlite)
        library(scales)
        library(grid)
        library(Matrix)
        library(mvtnorm)

        Loading required package: ggplot2
        Loading required package: dplyr
        Loading required package: magrittr
        Loading required package: tibble
        Loading required package: tmap
        Loading required package: grid
        ggmap: version 3.0.0, ggmap devel@a9ec8ac
        Google Maps API Terms of Service: https://developers.google.com/maps/terms.
        Please cite ggmap if you use it: see citation("ggmap") for details.
        Loading required package: RJSONIO
        ----------------------------------------------------------------------
        rvest: version 0.3.2, requires XML >= 3.98-1.3
        ----------------------------------------------------------------------
        Attaching package: 'rvest'
        The following object is masked from 'package:XML':
        ----------------------------------------------------------------------
        Attaching package: 'plyr'
        The following objects are masked from 'package:dplyr':
        ----------------------------------------------------------------------
        Attaching package: 'scales'
        The following object is masked from 'package:plyr':
        The following objects are masked from 'package:grid':
        ----------------------------------------------------------------------
        Attaching package: 'Matrix'
        The following object is masked from 'package:tidyr':
        ----------------------------------------------------------------------
        Attaching package: 'mvtnorm'
        The following object is masked from 'package:dplyr':
        ----------------------------------------------------------------------
        Attaching package: 'tidyverse'
        The following object is masked from 'package:ggmap':
        The following object is masked from 'package:ggplot2':
        The following object is masked from 'package:dplyr':
        The following object is masked from 'package:plyr':
        The following object is masked from 'package:tmap':
        The following object is masked from 'package:grid':
        The following object is masked from 'package:Matrix':
        The following object is masked from 'package:mvtnorm':
        The following object is masked from 'package:tibble':
        The following object is masked from 'package:stringr':
        The following object is masked from 'package:tmap':
        The following object is masked from 'package:ggplot2':
        The following object is masked from 'package:scales
```

Appendix H

Model

```
library(rjags)

#let's start by setting our directory and importing relevant packages
setwd("C:/Users/Leonardo/Documents/TU_Delft/EPA_Year1/EPA1315/FinalProject")

#load dataset
SF_LA_PH_DC_prcnts_vars <- read.csv('Data/AllCities/Tables/SF_LA_PH_DC_BO_prcnts_vars.csv')
#panel regression with fixed effects. Do not normalize variables,
#instead we set all variables on logarithmic scale
crime <- log1p(SF_LA_PH_DC_prcnts_vars$crimecount)
trees <- log1p(SF_LA_PH_DC_prcnts_vars$treecount)
lights <- log1p(SF_LA_PH_DC_prcnts_vars$lightcount)
trash <- log1p(SF_LA_PH_DC_prcnts_vars$trashcount)
SF <- log1p(SF_LA_PH_DC_prcnts_vars$CITY_San.Francisco)
LA <- log1p(SF_LA_PH_DC_prcnts_vars$CITY_Los.Angeles)
PH <- log1p(SF_LA_PH_DC_prcnts_vars$CITY_Philadelphia)
DC <- log1p(SF_LA_PH_DC_prcnts_vars$CITY_Washington.DC)
BO <- log1p(SF_LA_PH_DC_prcnts_vars$CITY_Boston)
n <- nrow(SF_LA_PH_DC_prcnts_vars)

model_string1 <- "model{

#likelihood
for (i in 1:n){
  crime[i] ~ dnorm(mu[i], inv.var) #insert function that represents distribution of crime
  mu[i] <- beta0 + beta1*trees[i] + beta2*lights[i] + beta3*trash[i] + beta4*SF[i] + beta5*LA[i] + beta6*PH[i] + beta7*DC[i] + beta8*BO[i]
}

#one prior for each beta
beta0 ~ dnorm(0, 0.001)
beta1 ~ dnorm(0, 0.001)
beta2 ~ dnorm(0, 0.001)
beta3 ~ dnorm(0, 0.001)
beta4 ~ dnorm(0, 0.001)
beta5 ~ dnorm(0, 0.001)
beta6 ~ dnorm(0, 0.001)
beta7 ~ dnorm(0, 0.001)
beta8 ~ dnorm(0, 0.001)


#prior for the inverse variance
inv.var ~ dgamma(0.01, 0.0001)
sigma <- 1/sqrt(inv.var)
}"
#initialize model
model <- jags.model(textConnection(model_string1),
                    data = list(crime=crime,trees=trees,
                                lights=lights,trash=trash,SF=SF,
                                LA=LA,PH=PH,DC=DC, n=n), n.chains = 2)

update(model, 10000, progress.bar="none"); # Burnin for 10000 samples
samp <- coda.samples(model,variable.names=c("beta0","beta1","beta2","beta3","beta4",
                                            "beta5","beta6", "beta7", "sigma"),
                                            n.iter=20000, progress.bar="text")
#get summary of posterior samples
summary(samp)

traceplot(samp)
# sometimes the gelman plot won't fit on a screen
# we have to reduce the margins
par(mar=c(.4,.4,.4,.4))
gelman.plot(samp)
gelman.diag(samp)
densplot(samp)
acfplot(samp)

#visualization of the complete posterior distributions
#with traceplots for each parameter, we first convert the model object to class mcmc
samp_mcmc <- as.mcmc(samp)
plot(samp)

# get the effective sample size
effectiveSize(samp)
```

Figure H1: R script for Model part 1

```
# take the mean of each predicted coefficient
coeff_preds <- as.matrix(samp)
coeff_preds <- as.data.frame(coeff_preds)
beta0_pred <- mean(coeff_preds$beta0)
beta1_pred <- mean(coeff_preds$beta1)
beta2_pred <- mean(coeff_preds$beta2)
beta3_pred <- mean(coeff_preds$beta3)
beta4_pred <- mean(coeff_preds$beta4)
beta5_pred <- mean(coeff_preds$beta5)
beta6_pred <- mean(coeff_preds$beta6)
beta7_pred <- mean(coeff_preds$beta7)
sigma_pred <- mean(coeff_preds$sigma)
#compute predicted crime with model betas
crime_predictor <- function(var1, var2, var3, var4, var5, var6, var7, var8){
  # vars: trees, lights, trash, SF, LA, PH, DC, n
  #empty dataframe to be filled with predicted crime values
  crime_pred <- data.frame(matrix(ncol = 1, nrow = n))
  x <- c("predicted_crime")
  colnames(crime_pred) <- x
    #function for crime
  crime_pred$predicted_crime <- beta0_pred + beta1_pred*var1 + beta2_pred*var2 +
    beta3_pred*var3 + beta4_pred*var4 + beta5_pred*var5 +
    beta6_pred*var6 + beta7_pred*var7
  return(crime_pred)
}

#for (i in 1:n){
#crime_pred$predicted_crime[i] <- beta0_pred + beta1_pred*var1[i] + beta2_pred*var2[i] +
#  beta3_pred*var3[i] + beta4_pred*var4[i] + beta5_pred*var5[i] +
#  beta6_pred*var6[i] + beta7_pred*var7[i]

crime_predictions <- crime_predictor(trees, lights, trash, SF, LA, PH, DC, n)
crime_obs_pred <- cbind(crime_predictions, crime)

library('ggplot2')
# plot predicted crime versus observed crime
ggplot(crime_obs_pred, aes(x=crime, y=predicted_crime)) +
  geom_point() +
  geom_smooth(method=lm)

plot(crime_obs_pred$crime, crime_obs_pred$predicted_crime, ylab='Predicted Crime', xlab='Observed Crime')
lines(seq(1,3500),seq(1,3500), col='red', lwd=2, cex=3)


model_output<-as.matrix(samp)
saveRDS(samp,"ModelResults/modelrun1.RDS")
write.csv(model_output,"modelruns2.csv",row.names = FALSE)
```

Figure H2: R script for Model part 2

```
Iterations = 10001:30000
Thinning interval = 1
Number of chains = 2
Sample size per chain = 20000

1. Empirical mean and standard deviation for each variable,
   plus standard error of the mean:

          Mean       SD  Naive SE Time-series SE
beta0  0.25571 0.095357 4.768e-04      3.983e-03
beta1  0.31216 0.014016 7.008e-05      6.770e-04
beta2  0.71237 0.016465 8.232e-05      8.584e-04
beta3  0.06721 0.008764 4.382e-05      1.017e-04
beta4  6.57801 0.183796 9.190e-04      8.880e-03
beta5 -0.41758 0.105534 5.277e-04      3.674e-03
beta6  0.05211 0.103912 5.196e-04      2.764e-03
beta7 -1.65546 0.152847 7.642e-04      3.180e-03
sigma  0.97820 0.010482 5.241e-05      5.241e-05

2. Quantiles for each variable:

          2.5%      25%      50%     75%    97.5%
beta0  0.06598  0.19365  0.25605  0.3188  0.44318
beta1  0.28475  0.30297  0.31208  0.3213  0.34047
beta2  0.68035  0.70133  0.71252  0.7233  0.74515
beta3  0.05004  0.06132  0.06715  0.0731  0.08451
beta4  6.21791  6.45733  6.57609  6.6996  6.94388
beta5 -0.62266 -0.48951 -0.41831 -0.3467 -0.21022
beta6 -0.14820 -0.01858  0.05189  0.1217  0.25625
beta7 -1.95329 -1.75829 -1.65661 -1.5521 -1.35448
sigma  0.95783  0.97108  0.97816  0.9851  0.99916
```

Figure H3: Model summary

```
> effectiveSize(samp)
    beta0      beta1      beta2      beta3      beta4      beta5      beta6      beta7      sigma
 572.8092   428.9641   367.2109  7473.1015   430.8856   824.3085  1416.7090  2320.9285 40000.0000
```

Figure H4: Effective sample size