Project Assignment: Data Wrangling

Team: Data Incubator

Part I. Data Quality and Descriptive statistics

a. Document the initial data quality.

The dataset we selected is based on the Boston Crime Incident Report, which can be found at https://data.boston.gov/dataset/crime-incident-reports-august-2015-to-date-source-new-system.

The database summarizes crime incidents that have happened in Boston from August, 2015. Due to the dataset size limit of Trifacta, we took the data starting 2019 as our dataset. There are 17 columns in the dataset that basically specify the type of crime, and time and place they occurred. Each Attribute is displayed below.

Column Name	Type	Values Represented	Description
incident_number	nominal	Ixxxxxxxxx(10 characters)	Number of incidents in sequence
offense_code	numeric	5 digit number	Number based on the type of crimes
offense_code_group	nominal	E.g. fraud,...	Crime type
offense_description	nominal	E.g. fraud- false pretense/scheme,....	Description of crime type

district	nominal	A1, A15, A7, B2, B3, C6, C11, D4, D14, E5, E13, E18	Code of district that each crime took place, for example: D4 represents South End
repoting_area	numeric	E.g. 200, 15,...	Code of the area that each crime took place
ucr_part	nominal	part 1, part 2, part 3, others	Uniform Crime Report that divides crimes into multiple levels based on severity
occured_on_date	nominal	yy/mm/dd, hh:mm AM/PM	Exact date and time that each crime happened
hour	numeric	1-24	Time of crimes happened
year	numeric	2015,2016,2017	Year of crimes happened
month	numeric	1-12	Month of crimes happened
day_of_week	nominal	Monday, Tuesday, Wednesday, Thursday, Friday, Saturday, Sunday	Date of week crimes happened
shooting	nominal	Yes	An indication that whether shootings happened in crimes
street	nominal	E.g. Huntington Ave, …..	Street of crimes happened
lat	nominal	E.g. 42.335269, …..	Latitude of place that crime happened

long	nominal	E.g. -71.101586, …...	Longitude of place that crime happened
location	nominal	E.g. (42.335269, -71.101586),.....	Latitude and longitude of place that crime happened

Overall the dataset quality is not quite good. There are some columns with mismatched values(reporting_area, occured_on_date) and missing value(district, shooting, ucr_part, street, lat, long). The format of a few columns would need adjustment, e.g. occured_on_date. There are also columns containing repetitive information, for example: a combination of Lat and Long columns is basically the Location column. Due to the characteristic of the dataset, there is no outlier.

b. Use Trifacta Wrangler, documenting your process and insights.

Recipe from Trifacta Wrangler:

(1) drop col: incident_number action: Drop
(2) drop col: offense_description action: Drop
(3) drop col: reporting_area action: Drop
(4) drop col: street action: Drop
(5) drop col: location action: Drop

INSIGHT 1: Drop columns like incident_number, offense_description which have thousands of values to avoid model overfitting.

(6) filter type: custom rowType: single row: ISMISSING([long]) action: Delete

Drop row with missing value for “long” column to increase data validitly

(7) filter type: custom rowType: single row: ISMISSING([ucr_part]) action: Delete
(8) filter type: custom rowType: single row: ISMISSING([district]) action: Delete

INSIGHT 2: Drop row with missing value for long, ucr_part and district columns to increase data validity

(9) splitpatterns col: occurred_on_date type: on on: ` ` limit: 1

Split on delimiter to sperate date and time

(10) rename type: manual mapping: [occurred_on_date1,'date
(11) rename type: manual mapping: [occurred_on_date2,'time']

INSIGHT 3: Rename columns to specify column information

(12) set col: shooting value: IFVALID($col, ['String'], 1)
(13) set col: shooting value: IFMISSING($col, 0)
(14) derive type: multiple value: IF(day_of_week == 'Monday', 1, IF(day_of_week == 'Tuesday', 2, IF(day_of_week == 'Wednesday', 3, IF(day_of_week == 'Thursday', 4, IF(day_of_week == 'Friday', 5, IF(day_of_week == 'Saturday', 6, 7)))))) group: day_of_week as: 'day_of_week_numeric'
(15) drop col: day_of_week action: Drop
(16) replacepatterns col: ucr_part with: '' on: `{delim}` global: true
(17) derive type: multiple value: CASE([ucr_part == 'PartOne', 1, ucr_part == 'PartTwo', 2, ucr_part == 'PartThree', 3, 4]) group: ucr_part as: 'ucr_part_numeric'
(18) drop col: ucr_part action: Drop

INSIGHT 4: Convert the nominal attributes, day_of_week, shooting and ucr_part into dummy/numeric variables to allow for the use of numeric estimation. Delete original columns if necessary.

(19) splitpatterns col: date type: on on: '\/' limit: 2
(20) settype col: date3 type: Integer
(21) set col: date3 value: IFVALID($col, ['Integer'], PAD(date3, 2, '0', left))
(22) sort order: date3
(23) set col: date2 value: IFVALID($col, ['Integer'], PAD(date2, 2, '0', left))
(24) merge col: date1,date2,date3 with: '\/' as: 'date_update'

INSIGHT 5: Some rows under data attribute have mismatched data because the date is in the format of mm(single-digit)/dd(single-digit)/yyyy instead of mm(double-digit)/dd(double-digit)/yyyy. To updated the format, split on column based on delimiter into 3 columns; pad ‘0’ in to the left of mouth(date3) and

day(date2) columns if the digits are less than 2. Finally, combine the 3 columns together using '/'.

(25) valuestocols col: district value: 1 default: 0 limit: 12

INSIGHT 6: Create dummy variables for 12 different district to get ready for numeric estimation.

c. Document the final data quality.

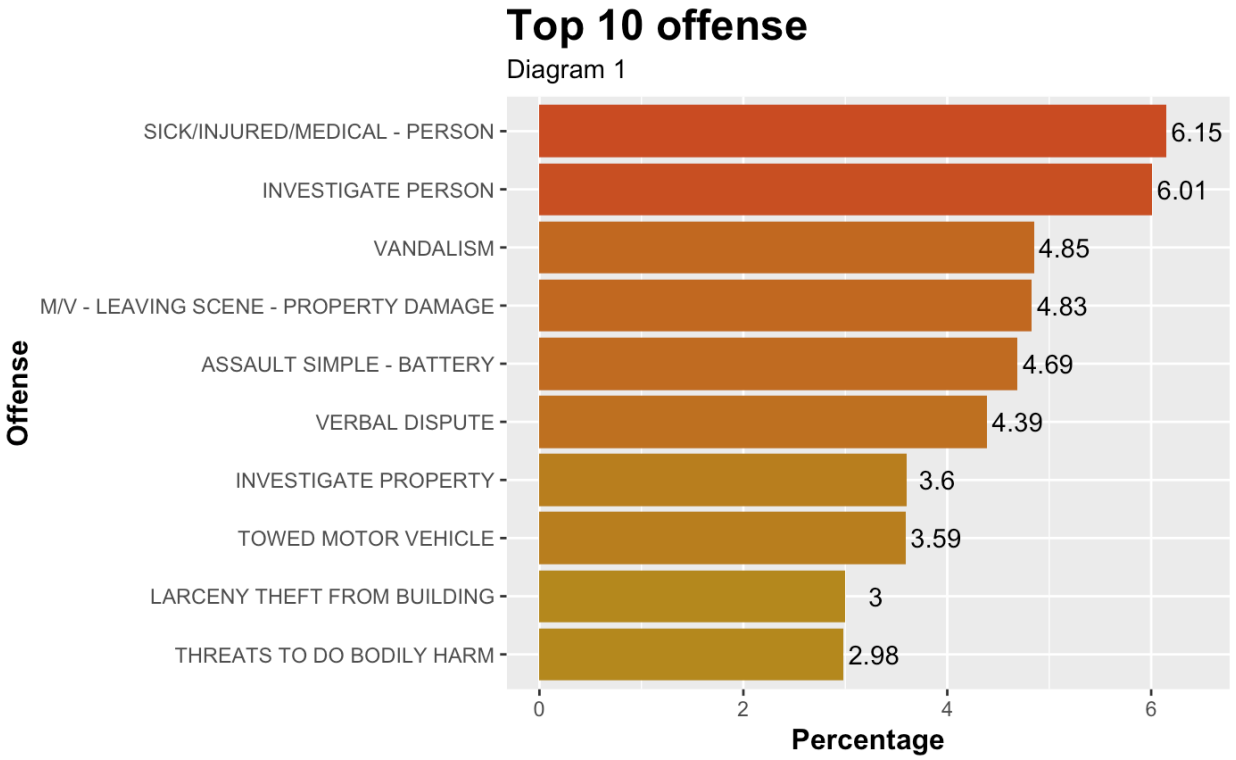
Using both Trifacta Wrangler and RStudio, we cleaned the data and manipulated it for exploratory analysis, and it's ready for further analytics. The detailed validation procedures and screening methods are shown above.

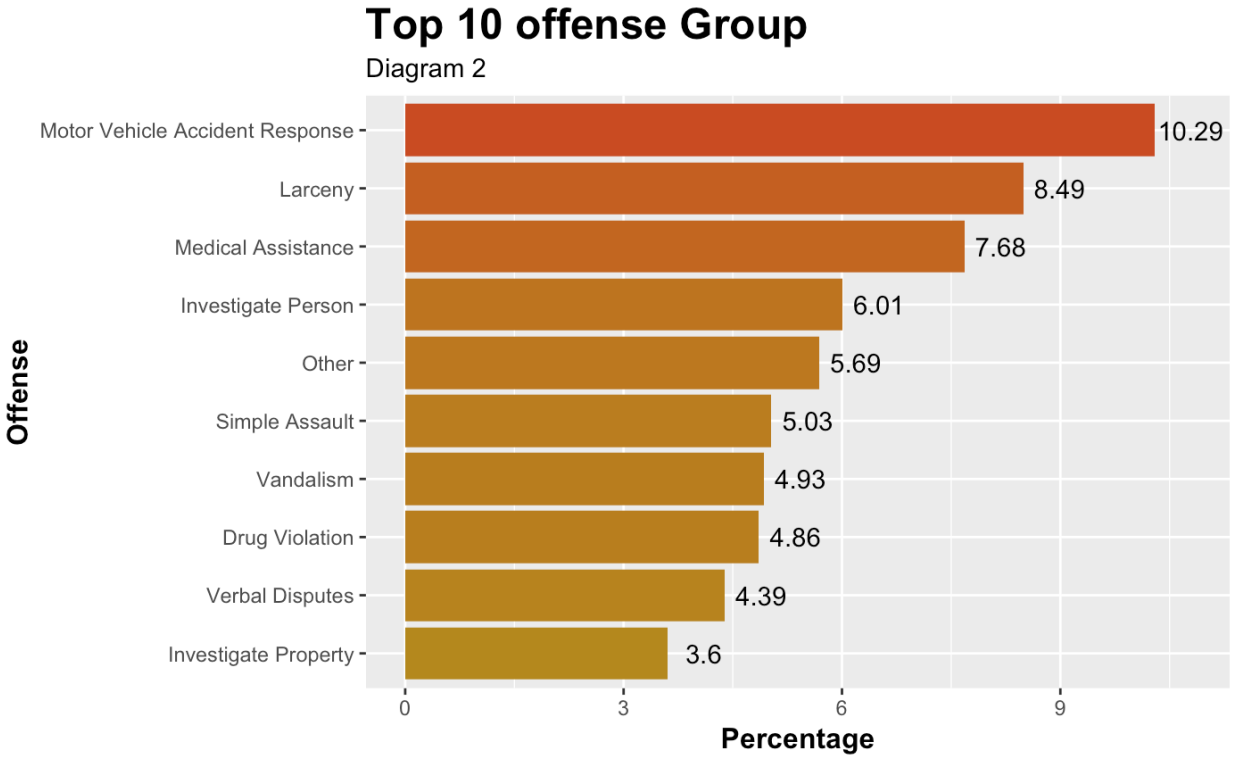
In the data wrangling process, we dealt with some main issues and considered the following key standards of data quality:

(1) Consistency: assigning right formats for mismatched values
(2) Validity: made sure that there are no missing values for key parameters
(3) Precision: dropped variables with repetitive information, and converted categorical variables to dummy/numeric variables to get ready for training regression model; performed statistical summaries to demonstrate patterns of the dataset

The final dataset was lined up in proper columns, with no missing values, no mismatched values, and right formats. Also, there are no outliers in the dataset.

d. Show descriptive statistics (histograms, box-and-whisker plots, etc.) for a few key variables.

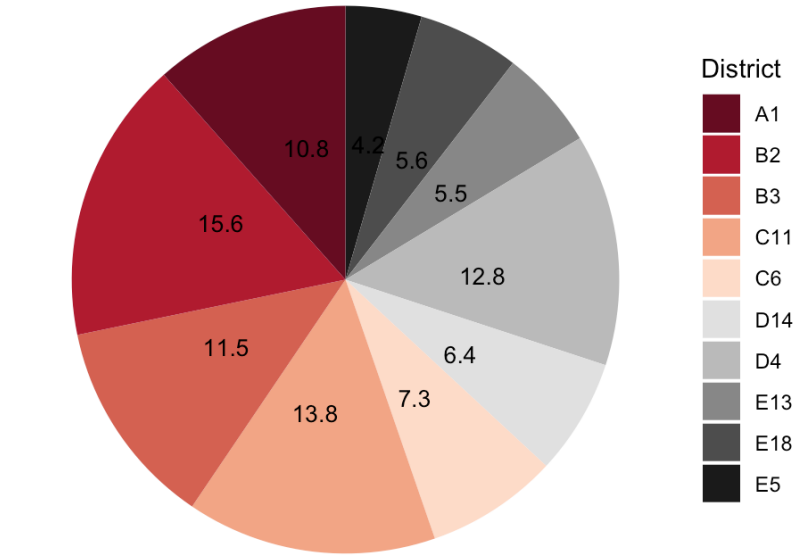




From Diagram 1 & 2, we can find that the top offense cases are not significantly serious but relatively trivial and less serious. 6.15% of the total cases that the Boston Police Department (BPD) respond to are medical cases (Diagram 1). Only 2.98% of the cases are about threats to do bodily harm (Diagram 1). And most of the crime is about property (motor or other) and larceny (Diagram 2). Therefore, Boston is a relatively safe city with not much major serious crimes.

District Crime Rate

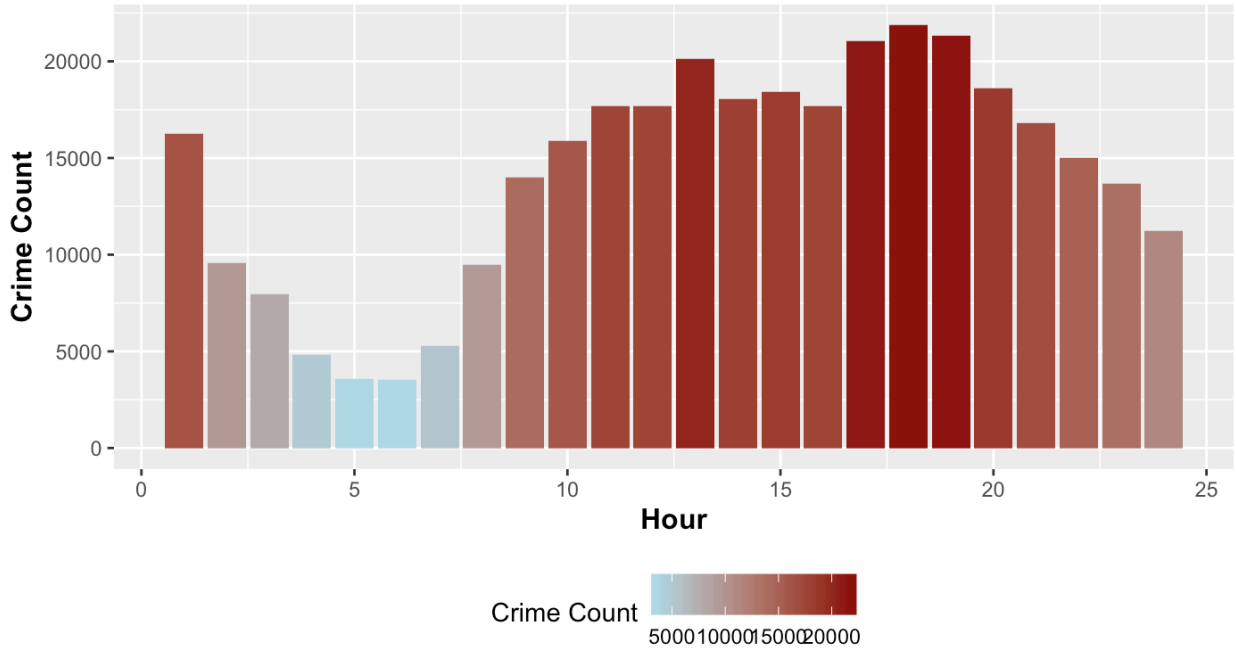
Diagram 3



From Diagram 3, the district with the highest crime rate is A1 with 15.6% and C11 comes later with 13.8%. District E (E5, E13, E18) are comparably safer than other districts.

High Crime Case Hour

Diagram 4



Offense Group	Hour	Count
Motor Vehicle Accident Response	1	1128
Simple Assault	2	838
Motor Vehicle Accident Response	3	898
Motor Vehicle Accident Response	4	563
Medical Assistance	5	406
Motor Vehicle Accident Response	6	461
Motor Vehicle Accident Response	7	812
Towed	8	1331
Motor Vehicle Accident Response	9	1715
Motor Vehicle Accident Response	10	1670
Motor Vehicle Accident Response	11	1639
Larceny	12	1711
Larceny	13	2198
Larceny	14	1990

Larceny	15	2164
Larceny	16	2024
Motor Vehicle Accident Response	17	2269
Motor Vehicle Accident Response	18	2398
Motor Vehicle Accident Response	19	2191
Motor Vehicle Accident Response	20	1832
Motor Vehicle Accident Response	21	1642
Motor Vehicle Accident Response	22	1519
Motor Vehicle Accident Response	23	1482
Motor Vehicle Accident Response	24	1223

From Diagram 4, we can easily find out that 1:00, 13:00, 17:00-19:00 are high crime hours, while 4:00-7:00 are safe hours. Interestingly, combined with the table that shows the most offense group, during 17:00-19:00 which is commute hour, most of the offense case are motor vehicle accident. And from 20:00 to 1:00, most of the motor vehicle accident may due to tiredness and alcohol. While during 12:00 to 16:00, most of the cases are about larceny. This may because during office hour people are not at home which is a good time for housebreaking and larceny.

Part II: One page summary of reflections on how Trifacta Wrangler could help us in general, including the following:

a. in the modeling process

Trifacta Wrangler allows us to do data cleaning and data gathering easily so that we will have actionable and aggregated data for our modeling process. The pro version supports data import from diverse sources including external databases and cloud storage, which makes the data acquisition process more convenient. For data cleansing process, Trifacta Wrangler provides statistical information including missing value, mismatch value, and data distribution for each column, and offers data cleaning suggestions and a preview of our operating results. It eliminates the pain of syntax correction and time-consuming rework. Joining data in Trifacta Wrangler is also handy. The “flows” page gives us a clear visualization of our data-editing and data-joining workflow. If we discover any errors in data during the data utilizing process, then we can trace the flow to find its origin and correct it.

b. in combination with other software tools

Trifacta Wrangler is a good data preparation tool, and it’s easy to combine it with other software tools like R studio. We used R to show descriptive statistics of our key variables. We can output the edited dataset to a .csv file and input the .csv file to R and do further data exploration.

c. in a collaborative organizational workflow

Trifacta can improve the work efficiency within a team. When working in a team for data analysis, it is likely to happen that everyone within the team has different preferences for programming languages. Therefore sometimes it can be hard and time-consuming to understand other teammates’ work. However, since Trifacta is so easy to use that even a person with no programming background can easily get used to it. In general, Trifacta allows the whole team to speak one language and therefore improve work efficiency.

d. in the current or future industry where we (wish to) work.

Take the marketing industry as an example. As a marketing analyst, you might need to manipulate data some times. However, instead of writing complex code using R or Python, Trifacta allows you to work faster and easily produce reports. In addition, Trifacta can to some extent eliminate the misunderstanding problem between non-tech and technical personnel, since the non-tech personnel can easily understand what those data mean.