Big Data Analytics Project

Crime Data from 2010 to 2017 (Los Angeles)

Gayatri Nambiar 12/18/17 AIT-580

• Briefly describe the dataset (size, data types)

Crime Data from 2010 to Present (Data.gov, 2017)

This dataset contains crime data in the City of Los Angeles from 2010 to 2017. This data is copied from crime reports originally written on paper. The size of the dataset is 357 MB with 1.62 million rows and 26 rows where each row is a crime incident. The types of data in this dataset are nominal, ordinal and ratio (Data.gov, 2017).

The data in this dataset includes the following fields. This data is obtained from the "data.lacity.org" website.

Table 1 About the columns in the dataset (Eric Garcetti #dataLa, 2017)

COLUMN NAME	Description
DR NUMBER	Division of Records Number: Official file number made up of a 2 digit year, area ID, and 5 digits
DATE REPORTED	MM/DD/YYYY
DATE OCCURRED	MM/DD/YYYY
TIME OCCURRED	In 24 hour military time.
AREA ID	The LAPD has 21 Community Police Stations referred to as Geographic Areas within the department. These Geographic Areas are sequentially numbered from 1-21
AREA NAME	The 21 Geographic Areas or Patrol Divisions are also given a name designation that references a landmark or the surrounding community that it is responsible for. For example 77th Street Division is located at the intersection of South Broadway and 77th Street, serving neighborhoods in South Los Angeles.
REPORTING DISTRICT	A four-digit code that represents a sub-area within a Geographic Area. All crime records reference the "RD" that it occurred in for statistical comparisons. Find LAPD Reporting Districts on the LA City GeoHub at http://geohub.lacity.org/datasets/c4f83909b81d4786aa8ba8a74a4b4db1_4
CRIME CODE	Indicates the crime committed. (Same as Crime Code 1)
CRIME CODE DESCRIPTION	Defines the Crime Code provided.
MO CODES	Modus Operandi: Activities associated with the suspect in commission of the crime. See attached PDF for list of MO Codes in numerical order. https://data.lacity.org/api/views/y8tr-7khq/files/3a967fbd-f210-4857-bc52-60230efe256c?download=true&filename=MO%20CODES%20(numerical%20order).pdf
VICTIM AGE	Two character numeric
VICTIM SEX	F - Female M - Male X - Unknown

VICTIM DESCENT	Descent Code: A - Other Asian B - Black C - Chinese D - Cambodian F - Filipino G - Guamanian H - Hispanic/Latin/Mexican I - American Indian/Alaskan Native J - Japanese K - Korean L - Laotian O - Other P - Pacific Islander S - Samoan U - Hawaiian V - Vietnamese W - White X - Unknown Z - Asian Indian
PREMISE CODE	The type of structure, vehicle, or location where the crime took place.
PREMISE DESCRIPTION	Defines the Premise Code provided.
WEAPON USED CODE	The type of weapon used in the crime.
WEAPON DESCRIPTION	Defines the Weapon Used Code provided.
STATUS CODE	Status of the case. (IC is the default)
STATUS DESCRIPTION	Defines the Status Code provided.
CRIME CODE 1	Indicates the crime committed. Crime Code 1 is the primary and most serious one. Crime Code 2, 3, and 4 are respectively less serious offenses. Lower crime class numbers are more serious.
CRIME CODE 2	May contain a code for an additional crime, less serious than Crime Code 1.
CRIME CODE 3	May contain a code for an additional crime, less serious than Crime Code 1.
CRIME CODE 4	May contain a code for an additional crime, less serious than Crime Code 1.
ADDRESS	Street address of crime incident rounded to the nearest hundred block to maintain anonymity.
CROSS STREET	Cross Street of rounded Address.
LOCATION	The location where the crime incident occurred. Actual address is omitted for confidentiality. XY coordinates reflect the nearest 100 block.

• Who (company, agency, organization) collected the data?

The data is provided by Los Angeles Police Department and published by “data.lacity.org” on Data.gov (Data.Gov, 2017).

• Who they are, what do they do?

The Data.gov website is hosted by U.S government for open data. This website was launched in the year 2009 by the then Federal Chief Information Officer of the United States, Vivek Kundra (Data.gov, “About Data.gov”, 2017).
They have a wide variety of datasets on topics including Agriculture, Climate, Consumer, Education, Energy, Finance, Health, Local Government, Manufacturing, Maritime, Ocean, Public Safety, Science and Research (Data.gov ,“Crime Data from 2010 to Present-A Safe City”, 2017).

• What is their role/purpose?

The purpose of Data.gov is to improve public access to high value, machine readable datasets. By making government data open -more accessible, and discoverable, the government aims to impact research and scientific discoveries, cost savings, efficiency, fuel for business, and increase public participation in the democratic dialogue (Data.gov, "Impact", 2017).
Apart from that, open data powers software applications to help people to make informed decisions (Data.gov, "Applications", 2017).

Need
• Why did they collect this data?

This crime data is collected to analyze crime trends on a city level. It provides annual statistics on the number of crimes, date and time of occurrence, gender, sex and descent of victims targeted, kind of weapons used, types of crimes and in what areas in the city. This will help the local police department to take precautionary actions or by law enforcement agencies for strategic decision making and operational or tactical purposes (Lauritsen, 2016).
An open crime data also helps the public to stay informed about public safety as well as gain transparency into what local police is doing.

• What potential questions could be answered by studying this data?
§ List some specific questions, and be sure to answer them in your analysis

1. Are there any patterns in the occurrence of crimes based on date or time?
2. Are these crimes increasing or decreasing over an annual timeframe?
3. Locations associated with specific crime.
4. Does gender, age or descent play a role in getting victimized?
5. Effect of factors such as per capita income, or population on crime statistics?
6. Locations where crimes occur and if there is a relationship between those places.
7. What are the most commonly used weapons for crime?

• Are there any privacy, quality, or other issues with this data?

As this is an open dataset, to maintain the privacy of data, address and location fields are rounded off to nearest 100 blocks to maintain privacy. Quality wise there is are some missing data and also since the data is transcribed from paper-based records, there may be some inaccuracies within the data (Eric Garcetti #dataLA, "Crime Data from 2010 to Present-A Safe City", 2017).

Requirements, resources needed
• **What software and hardware resources are needed to study this data?**

Software resources: PyCharm, SQL Server Management Studio, Tableau
Hardware: Personal Laptop

Present the Results/Findings

To analyze the crime data, I have divided the process into following steps:

I. Diagnose data for cleaning.

This step is performed to examine the structure of data, number of rows and columns, column names and data types. Diagnosing the data is an important step as it gives insight into the data and helps to detect any problems in the data (Nelson, 2017).

Code:

```
import pandas as pd

df = pd.read_csv('Crime_Data_from_2010_to_Present_Original.csv', parse_dates=['Date Occurred'])
pd.set_option('display.max_columns', None)                              #displays all columns

print("==================Head========================")
print(df.head())
#
print("==================Tail========================")
print(df.tail())

print("==================Columns========================")
print(df.columns)

print("==================Shape========================")
print(df.shape)

print("==================Info========================")
print(df.info())
```

Figure 1 Code for cleaning data

Output:

```
=================Shape======================
(1621438, 26)
=================Info=======================
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1621438 entries, 0 to 1621437
Data columns (total 26 columns):
DR Number                 1621438 non-null int64
Date Reported             1621438 non-null object
Date Occurred             1621438 non-null datetime64[ns]
Time Occurred             1621438 non-null int64
Area ID                   1621438 non-null int64
Area Name                 1621438 non-null object
Reporting District        1621438 non-null int64
Crime Code                1621438 non-null int64
Crime Code Description    1620987 non-null object
MO Codes                  1445608 non-null object
Victim Age                1490225 non-null float64
Victim Sex                1472590 non-null object
Victim Descent            1472557 non-null object
Premise Code              1621361 non-null float64
Premise Description       1617969 non-null object
Weapon Used Code          537190 non-null float64
Weapon Description        537189 non-null object
Status Code               1621436 non-null object
Status Description        1621438 non-null object
Crime Code 1              1621432 non-null float64
Crime Code 2              102509 non-null float64
Crime Code 3              2273 non-null float64
Crime Code 4              74 non-null float64
Address                   1621438 non-null object
Cross Street              269375 non-null object
Location                  1621429 non-null object
dtypes: datetime64[ns](1), float64(7), int64(5), object(13)
memory usage: 321.6+ MB
None
```

Figure 2 Output of shape and info command from figure 1

Figure 2 shows that there are 1621438 columns and 26 rows in the dataset. It also shows the data types associated with each column. The numbers prefixing the data types is the count of values in each column, therefore we can derive the number of missing values from this. For example: DR Number has count 1621438, but Victim Age has count only 1490225.

```
                          Cross Street             Location
1621433                            NaN  (34.1649, -118.6027)
1621434                            NaN  (34.1923, -118.6104)
1621435                            NaN  (34.1883, -118.6274)
1621436                            NaN  (34.1905, -118.6059)
1621437  SHERMAN                    WY   (34.232, -118.6006)
```

Figure 3 Output of the Tail command from figure 1

Figure 3. is a snapshot of the output of tail command. As you can see in figure 3, we have some NaN values and the location field contains longitude and latitude combined into one column which would make it difficult to analyze further and visualize.

```
==================Columns========================
Index(['DR Number', 'Date Reported', 'Date Occurred', 'Time Occurred',
       'Area ID', 'Area Name', 'Reporting District', 'Crime Code',
       'Crime Code Description', 'MO Codes', 'Victim Age', 'Victim Sex',
       'Victim Descent', 'Premise Code', 'Premise Description',
       'Weapon Used Code', 'Weapon Description', 'Status Code',
       'Status Description', 'Crime Code 1', 'Crime Code 2', 'Crime Code 3',
       'Crime Code 4', 'Address', 'Cross Street', 'Location '],
      dtype='object')
```

Figure 4 Output of columns command from figure 1

By viewing the names of the column in Figure 4, we notice that the Location column has a space suffixing.

II. Cleaning and handling Missing Data (Pandas documentation, n.d.)

Given the size of the dataset (357 MB with 1.62 million rows), it is impossible to edit the dataset as an excel sheet as it would cause data loss. Splitting the CSV file would cause discrepancy in data because the data contained tabs, space and commas and missing values within columns. The following code is written in python using pandas package to fill the NaNs, predict missing data and normalize data types.

```python
import pandas as pd

pd.set_option('display.max_columns', None)

df = pd.read_csv('Crime_Data_from_2010_to_Present_Original.csv', parse_dates=['Date Occurred'])

new_df = df.fillna({
'Crime Code Description' : 'no description',
'MO Codes' : "0",
'Victim Sex' : "U",
'Victim Descent' : "X",
'Premise Code' : 0,
'Premise Description' : "no data",
'Weapon Used Code' : 0,
'Weapon Description' : "no data",
'Status Code' : "0",
'Status Description' : "no data",
'Crime Code 1' : 0,
'Crime Code 2' : 0,
'Crime Code 3' : 0,
'Crime Code 4' : 0,
'Address': "no data",
'Cross Street': "no data"
})


# Create a groupby object: by_sex_descent
by_sex_descent = new_df.groupby(['Victim Sex', 'Victim Descent'])
```

```python
# Write a function that imputes median
def impute_median(series):
    return series.fillna(series.median())

# Impute age
new_df['Victim Age'] = by_sex_descent['Victim Age'].transform(impute_median)



new_df['Longitude'], new_df['Latitude'] = new_df['Location '].str.split(',', 1).str


new_df['Longitude'] = new_df['Longitude'].str.strip()
new_df['Longitude'] = new_df['Longitude'].str.replace(r"^\D", '', case=False)

new_df['Latitude'] = new_df['Latitude'].str.strip()
new_df['Latitude'] = new_df['Latitude'].str.replace(r"\D$", '', case=False)


#interpolate lon and lat

new_df['Longitude'] = new_df['Longitude'].interpolate()
new_df['Latitude'] = new_df['Latitude'].interpolate()



new_df.to_csv('Clean_CrimeDataset.csv', sep='\t', encoding='utf-8')
```

In the data we have, there are columns such as “MO codes”,”Crime Code”, “Code Description”, “Weapon description” and similar columns which have NaN values that cannot be predicted. Therefore we have filled such columns using “fillna” command to replace with 0 or “no data” based on the data type.
The Vixtims’s age can be predicted using other attributes that describe the victim such as Descent and Sex. Hence we have grouped by the Victim’s Sex and Descent and imputed the age using a median function.
Next, we split the Location column into two columns “Longitude” and “Latitude”, then strip and replace unnecessary characters and space. Having two separate columns for Longitude and Latitude will facilitate visualization in tableau by acting as a measure and not simply a string value.
Since the XY coordinates of the Location reflect the nearest 100 block, interpolation for the Longitude and Latitude appeared to be a good option.
Also, I have normalized the data type of “Date Occurred “ so that the numbers will read in as “Date” and not a “String”.

III. Import to SQL

The First step is to create a database on the SQL Server. The CSV with the clean data is then imported into the database using SQL Server Management Studio. The data has been imported using the SQL Server Import and Export Wizard. Following is the screenshot of the design of the table created which shows the column, data type, and Null check:

Column Name	Data Type	Allow Nulls
[Date Reported]	date	☑
[Date Occurred]	date	☑
[Time Occurred]	int	☑
[Area ID]	int	☑
[Area Name]	varchar(50)	☑
[Reporting District]	int	☑
[Crime Code]	int	☑
[Crime Code Description]	varchar(50)	☑
[MO Codes]	varchar(50)	☑
[Victim Age]	float	☑
[Victim Sex]	varchar(50)	☑
[Victim Descent]	varchar(50)	☑
[Premise Code]	float	☑
[Premise Description]	varchar(50)	☑
[Weapon Used Code]	float	☑
[Weapon Description]	varchar(50)	☑
[Status Code]	varchar(50)	☑
[Status Description]	varchar(50)	☑
[Crime Code 1]	float	☑
[Crime Code 2]	float	☑
[Crime Code 3]	float	☑
[Crime Code 4]	float	☑
Address	varchar(50)	☑
[Cross Street]	varchar(50)	☑
[Location]	varchar(50)	☑
Longitude	varchar(50)	☑
Latitude	varchar(50)	☑

IV. Analysis & Visualization

To begin with, a general analysis of the data is performed. The results of which are used to dive deep into the specific analysis.

1. Generic analysis

```python
print("Total number of crime incidents: " + str(len(df)))
print("Number of columns: " + str(len(df.columns)))
print(df.shape)
print('\n')

print("Total unique areas")
print(df['Area Name'].nunique())

print('\n')


print("Number of crimes by areas")
print(df.groupby('Area Name').size().sort_values(ascending=False))


print('\n')

print("Most Common crimes")
print(df.groupby('Crime Code Description').size().sort_values(ascending=False).head(10))


print('\n')


print("Date on which most number of crimes occurred")
print(df[].groupby('Date Occurred').size().sort_values(ascending=False).head(10))
#

# print('\n')

print("Date on which least number of crimes occurred")
print(df.groupby('Date Occurred').size().sort_values(ascending=True).head(10))

for x in range(2010, 2018):
    a = str(x)
    print("Most common crime on 01/01/" + a)
    print(df._where(df['Date Occurred'] == '01/01/' + a).groupby('Crime Code Description').size().sort_values(ascending=F

print('\n')
```

Figure 5 Python code for analyzing data

Results:

1. **Total number of rows (crime incidents)**
   1621438

2. **Total number of columns**
   26

3. **Total unique areas**
   21

4. Top 10 most Common crime

Crime Code Description	Count
BATTERY - SIMPLE ASSAULT	148950
BURGLARY FROM VEHICLE	124315
VEHICLE - STOLEN	124238
BURGLARY	117223
THEFT PLAIN - PETTY ($950 & UNDER)	116109
THEFT OF IDENTITY	102559
INTIMATE PARTNER - SIMPLE ASSAULT	87821
VANDALISM - FELONY ($400 & OVER, ALL CHURCH VANDALISMS) 0114	81523
VANDALISM - MISDEAMEANOR ($399 OR UNDER)	72862
ASSAULT WITH DEADLY WEAPON, AGGRAVATED ASSAULT	69395

5. Number of crimes by area (sorted by count)

Area Name	Count
77th Street	113061
Southwest	104611
N Hollywood	88374
Pacific	85692
Southeast	85452
Mission	81974
Northeast	78254
Van Nuys	77346
Newton	76283
Devonshire	75656
Topanga	75092
Hollywood	74131
Harbor	72390
Olympic	71914
West Valley	69313
Central	69031
Rampart	68525
West LA	67829
Wilshire	65212
Foothill	62291
Hollenbeck	59007

6. **Top 10 Dates on which most number of crimes occurred**

Date Occurred	Count
01/01/2010	2143
01/01/2011	2055
01/01/2012	1647
01/01/2013	1467
01/01/2014	1333
01/01/2015	1271
01/01/2016	1145
06/01/2012	997
06/01/2010	943
04/01/2011	939

From the above analysis, we observe that:
1. The most common crime in LA is BATTERY - SIMPLE ASSAULT.
2. The most number of crimes every year has occurred on new year's day. Let's do a deeper analysis to check for a pattern in such crimes.

2. Specific Analysis

Lets take a look at the top 2 crimes that has happened every year on new years :

Most common crime on 01/01/2010	
Crime Code Description	**Number of Occurrence**
THEFT OF IDENTITY	**997**
CRM AGNST CHLD (13 OR UNDER) (14-15 & SUSP 10 YRS OLDER)0060	**168**
Most common crime on 01/01/2011	
Crime Code Description	
THEFT OF IDENTITY	**1053**
CRM AGNST CHLD (13 OR UNDER) (14-15 & SUSP 10 YRS OLDER)0060	**128**
Most common crime on 01/01/2012	
Crime Code Description	
THEFT OF IDENTITY	**666**
BATTERY - SIMPLE ASSAULT	**92**
Most common crime on 01/01/2013	
Crime Code Description	
THEFT OF IDENTITY	**531**
BATTERY - SIMPLE ASSAULT	**93**
Most common crime on 01/01/2014	
Crime Code Description	
THEFT OF IDENTITY	**536**
BATTERY - SIMPLE ASSAULT	**78**
Most common crime on 01/01/2015	
Crime Code Description	

THEFT OF IDENTITY	**453**
BATTERY - SIMPLE ASSAULT	**64**
Most common crime on 01/01/2016	
Crime Code Description	
THEFT OF IDENTITY	**295**
BATTERY - SIMPLE ASSAULT	**76**
Most common crime on 01/01/2017	
Crime Code Description	
THEFT OF IDENTITY	**151**
BATTERY - SIMPLE ASSAULT	**74**

From the above results, we derive that:
1. The crimes committed on the new year's are mainly "Theft of Identity", "Battery – Simple Assault" and "crime against children that are 13 or under".

Trend of Battery – Simple Assault from 2010-2017 in LA:

```
declare @i int
set @i=2010
while(@i<=2017)
begin
select count(*) from Clean_CrimeDataset
where [Crime Code Description] like '%BATTERY - SIMPLE ASSAULT%'
and year(convert(date,[Date Occurred])) = @i;
set @i=@i+1
end
```

Figure 6 SQL query to obtain number of Batter- Simple Assault by year

	(No column na...
1	20519

	(No column na...
1	19902

	(No column na...
1	19815

	(No column na...
1	18924

	(No column na...
1	18403

	(No column na...
1	17583

	(No column na...

Figure 7 Output of Figure 6

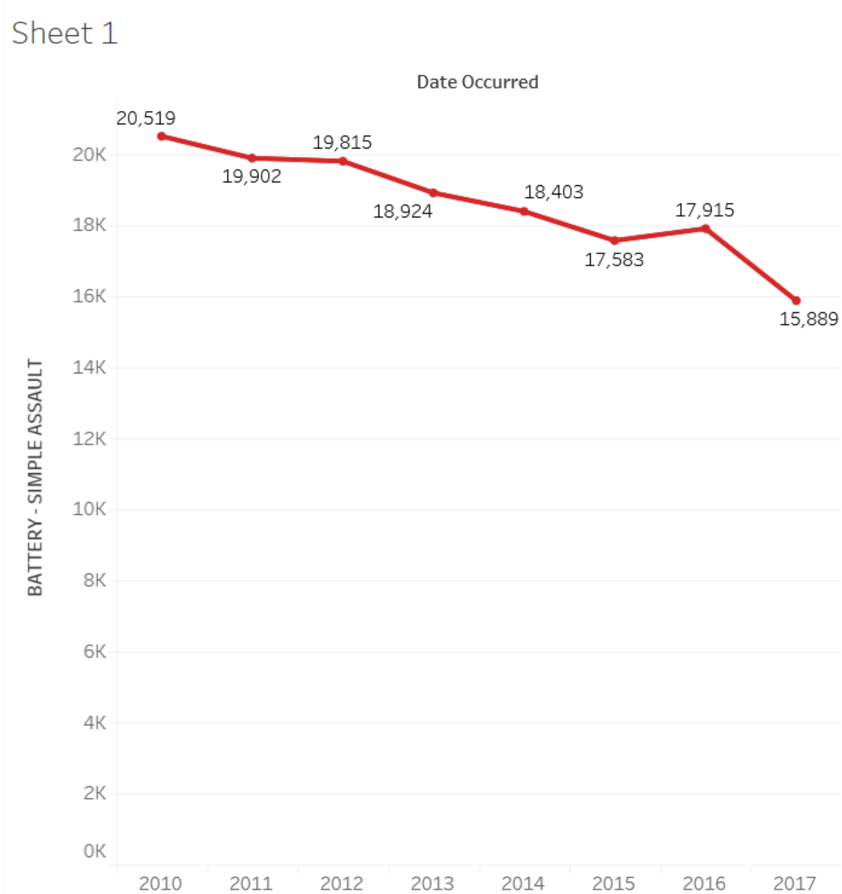


Figure 8 Graph for trend of Battery- Simple Assault

We observe that the graph of "BATTERY - SIMPLE ASSAULT" is going down, which means its good news, the crime is decreasing.

Crime statistics from 2010-2017 in LA :

```
declare @i int
set @i=2010
while(@i<=2017)
begin
select count(Clean_CrimeDataset.[Crime Code]) as total from Clean_CrimeDataset
where year(convert(date,[Date Occurred])) = @i
set @i=@i+1
end
```

Figure 9 SQL query to obtain crime statistics

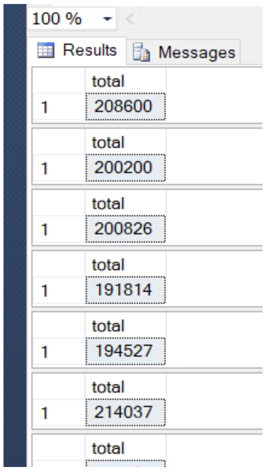


Figure 10 Output of Figure 8

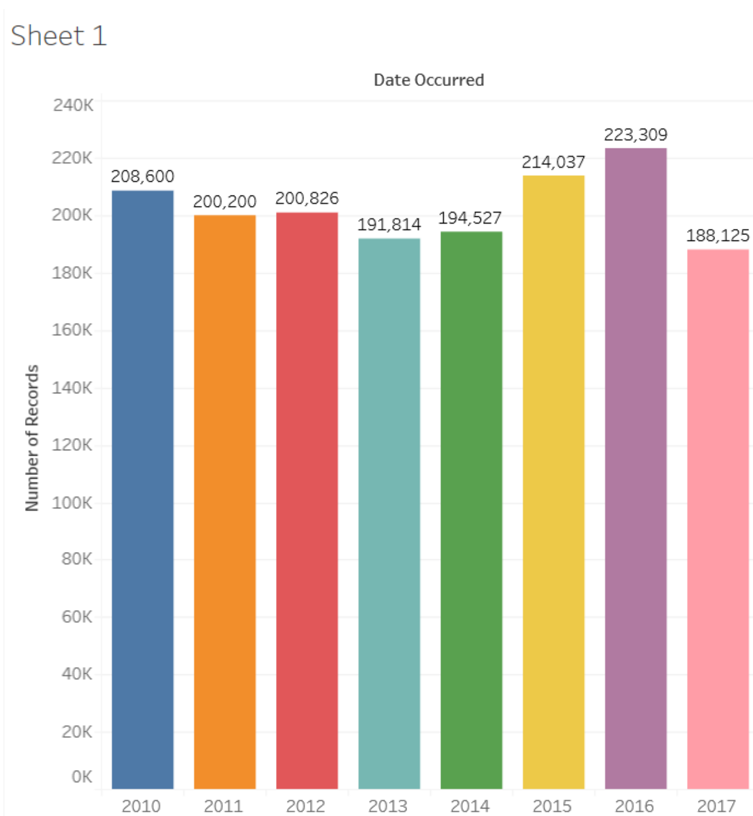


Figure 11 Graph for number of crimes by year in LA

Influence of per capita income on crimes in LA:

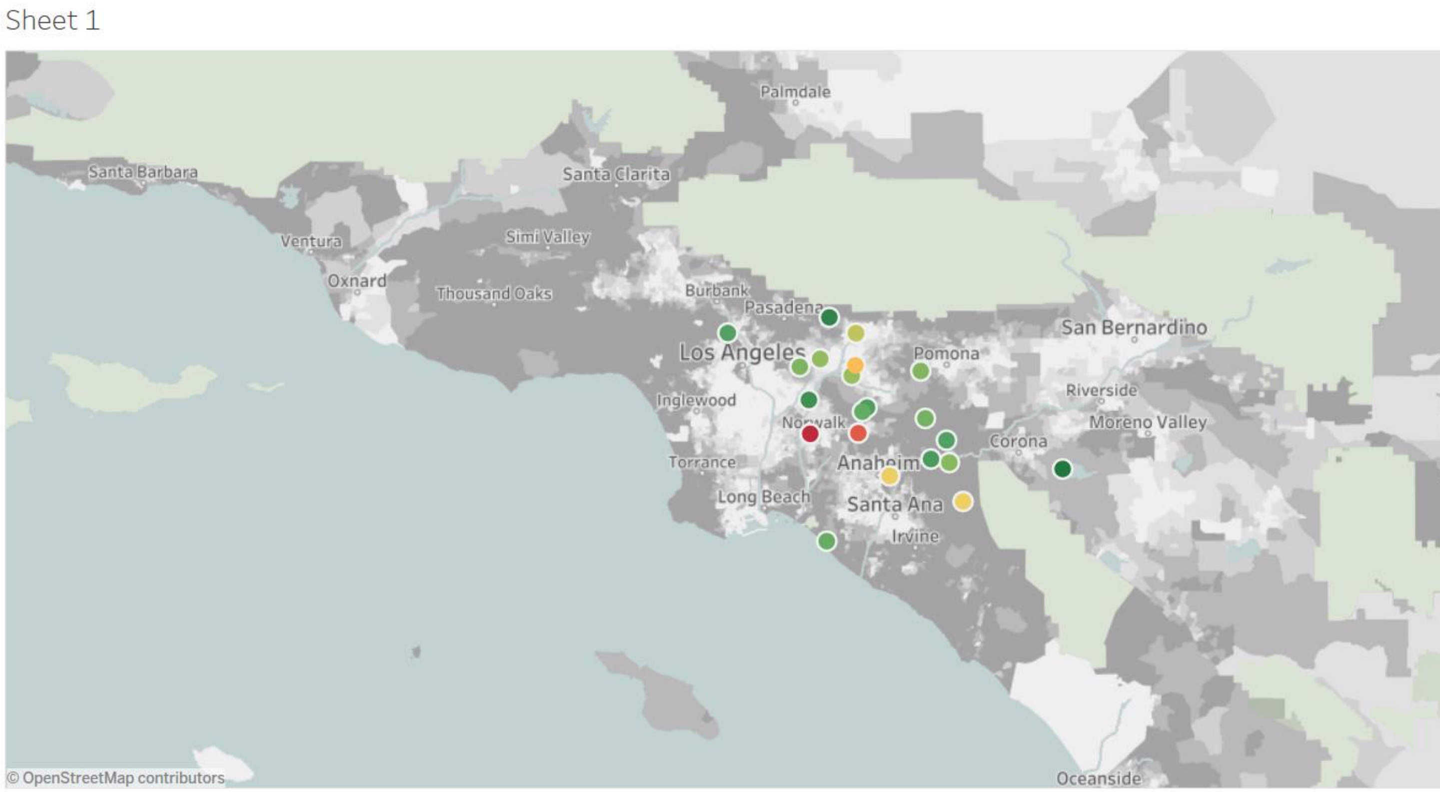


Figure 12 Map for influence of per capita income on crimes in LA

From the map, we observe that the crimes are more concentrated on the low per capita income areas of Los Angeles.

Proportion of crime committed based on Victim's descent:

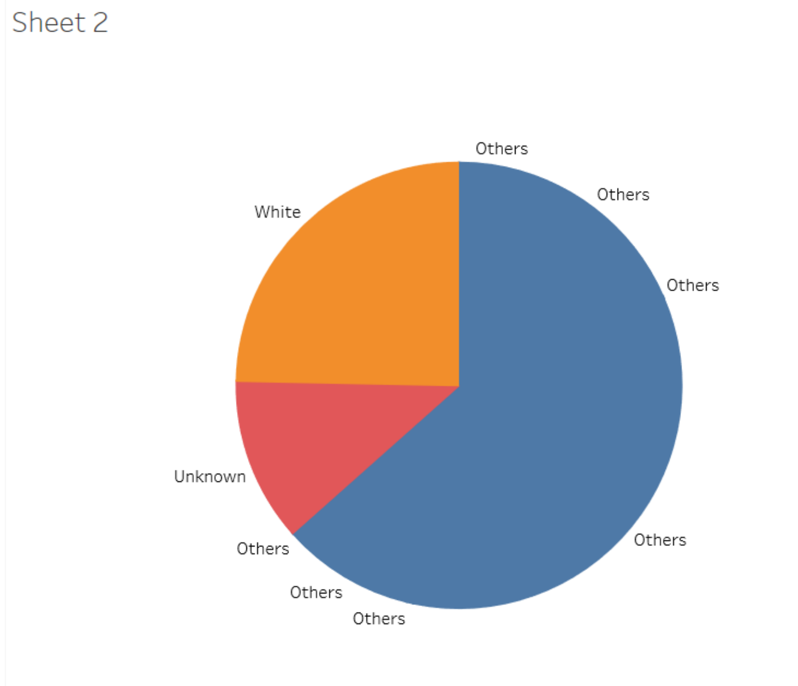
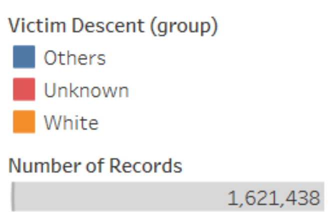


Figure 13 Pie chart depicting the proportion of crime in LA based on Victim's Descent




```
select count(*) from Clean_CrimeDataset
where [Victim Descent]!='W'   and   [Victim Descent] != 'X'

select count(*) from Clean_CrimeDataset
where [Victim Descent]='X'

select count(*) from Clean_CrimeDataset
where [Victim Descent]='W'
```

Figure 14 SQL queries for proportion of crimes committed based on Victim's Descent (W- White, X- Unknown)

	(No column na...
1	1028745

Figure 15 Output of Fig 13: Others

	(No column na...
1	400222

Figure 16 Output of Fig 13: White

	(No column na...
1	192471

Figure 7 Output of Fig 13: Unknown

As we see, 63.4% of victims are other than whites, 24.6% are whites, and remaining 11.8% are unkown. We could derive how many of these crimes were hate crimes, but for that, we would need more data from another source and perform a combined analysis.

Commonly used weapons for crimes in Los Angeles:

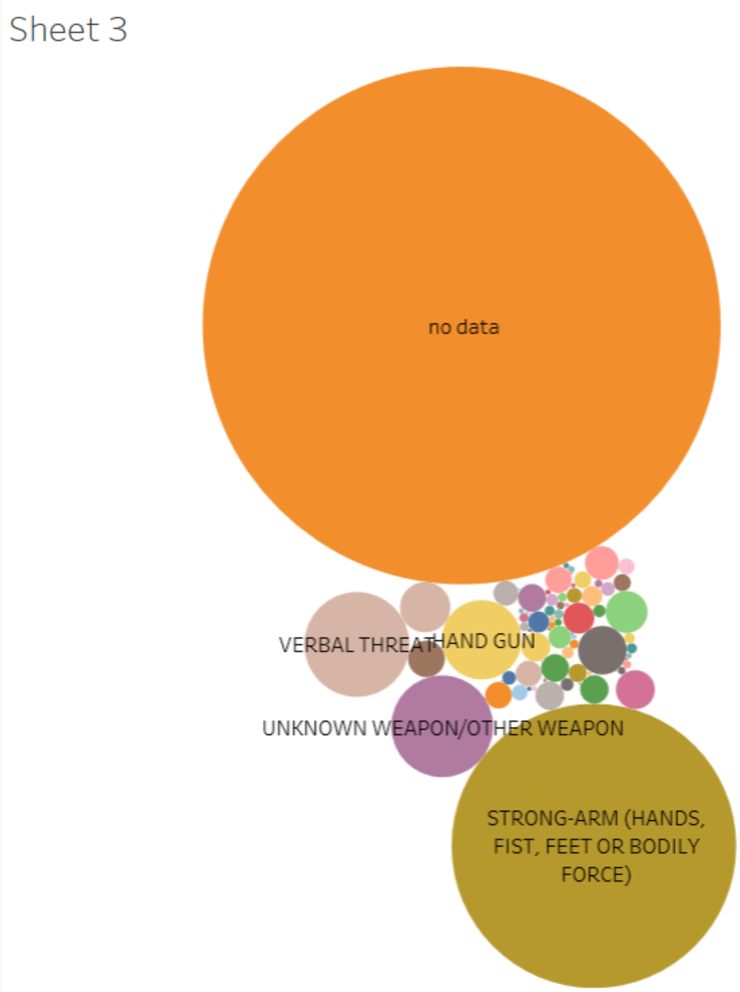


Figure 17 Bubble chart depicting the ratio of Weapons used for crime in LA

```
select [Weapon Description], count(*) as total from Clean_CrimeDataset
group by [Weapon Description]
order by total desc
```

Figure 18 SQL query for commonly used weapons

	Weapon Description	total
1	no data	1084249
2	STRONG-ARM (HANDS, FIST, FEET OR BODILY FORCE)	327235
3	VERBAL THREAT	44761
4	UNKNOWN WEAPON/OTHER WEAPON	41917
5	HAND GUN	25897

Figure 19 Output of Figure 16

As we see from the analysis, there are a lot of missing data (1084249) on weapon description, therefore we cannot be 100% sure on what is the maximum used weapons for crimes in Los Angeles. From the data we have, majority crimes involve strong arm and verbal threat more than any other weapon.

• Interpret the results; what conclusions can be supported?

Summarizing key findings:

1. Every year, the most number of crimes are committed on the new year's day. These crimes are particularly "Theft of Identity", "Battery – Simple Assault" and "crime against children that are 13 or under".
2. Most common crime in Los Angeles is BATTERY - SIMPLE ASSAULT, but we see a decreasing trend in this crime since 2010. However, the overall crime statistics in LA has remained pretty much the same, 202679.75 on an average of recorded incidents every year.
3. Crimes in LA are more concentrated in areas such as 77th street and Southwest where the per capita income is lower compared to surrounding areas.
4. Statistics report that people from an ethnic background other than whites are victimized more than whites.
5. Strong arm and verbal threat are involved in more number of crimes than weapons.

Explain/define terms

1. Data cleaning
   Data cleaning is the process of dealing with incorrect data, improperly formatted, incomplete or duplicate data (Margaret Rouse, 2010).
2. Data Normalization
   It is the process of organizing data into a table by a thorough investigation of data to avoid duplication of data and insert, delete and update anomaly (Tomar, 2011).
3. Simple battery
   Simple battery is the term used by the court to refer to the unauthorized or unlawful use of force to the body of another person resulting in injury or offensive touching (Vukovic, 2017).

• References

Data.Gov. (2017, November 7). Crime Data from 2010 to Present - A Safe City [Comma Separated Values file]. Eric Garcetti #dataLA [distributer] Retrieved December 18, 2017, from https://catalog.data.gov/dataset/crime-data-from-2010-to-present

Data.gov. (2017). Crime Data from 2010 to Present | Los Angeles - Open Data Portal. Retrieved December 18, 2017, from https://data.lacity.org/A-Safe-City/Crime-Data-from-2010-to-Present/y8tr-7khq

Data.gov. (2017). Impact. Retrieved December 18, 2017, from https://data.lacity.org/impact/

Data.gov. (2017). Applications. Retrieved December 18, 2017, from https://data.lacity.org/applications/

Data.gov. (2017). About Data.gov . Retrieved December 18, 2017, from https://data.lacity.org/about/

Eric Garcetti #dataLA (2017). Crime Data from 2010 to Present | Los Angeles - Open Data Portal. Retrieved December 18, 2017, from https://data.lacity.org/A-Safe-City/Crime-Data-from-2010-to-Present/y8tr-7khq

Lauritsen, J. L., & Cork, D. L. (2016). Modernizing crime statistics: Report 1 -- Defining and classifying crime. Washington, DC: National Academies Press.

Nelson, A. (2017, August 17). Cleaning Dirty Data with Pandas & Python. Retrieved December 18, 2017, from http://www.developintelligence.com/blog/2017/08/data-cleaning-pandas-python/

Pandas documentatio. (n.d.). Working with missing data¶. Retrieved December 18, 2017, from https://pandas.pydata.org/pandas-docs/stable/missing_data.html

Rouse, M. (2010). What is data scrubbing (data cleansing)? - Definition from WhatIs.com. Retrieved December 18, 2017, from http://searchdatamanagement.techtarget.com/definition/data-scrubbing

Tomar, N. (2011, March 15). Normalization and its Types. Retrieved December 18, 2017, from http://www.c-sharpcorner.com/uploadfile/nipuntomar/normalization-and-its-types/

Vukovic, A. (2017, May 25). Find the Right Lawyer Now! Choose Your Legal Category:. Retrieved December 18, 2017, from https://www.legalmatch.com/law-library/article/simple-battery.html