School of Informatics & IT
TEMASEK POLYTECHNIC

Specialist Diploma in Big Data Management

AY 2020/2021

Data Visualisation Fundamentals (CBG1C03)

Submitted by

2080548G

_____

Student No

Student Name: Meenakshi Dekshinamurthy

Date of Submission: 27-Dec 2020

Declaration of Work of Originality

Specialist Diploma in Big Data Management
Data Visualisation Fundamentals (CBG1C03)
AY2020/2021
Assignment

Submitted by: 2080548G Meenakshi Dekshinamurthy

Date: signing date in 27/12/2020

"By submitting this work, I am declaring that I am the originator(s) of this work and that all other original sources used in this work has been appropriately acknowledged.

I understand that plagiarism is the act of taking and using the whole or any part of another person's work and presenting it as my own without proper acknowledgement.

I also understand that plagiarism is an academic offence, and that disciplinary action will be taken for plagiarism."

Name and Signature of student: Meenakshi Dekshinamurthy

_____

Table of Contents

1 Analysis on data quality

1.1 Accessible

The data is accessible easily from the published online source from https://data.lacity.org/A-Safe-City/Crime-Data-from-2010-to-2019/63jg-8b9z as excel file easily downloadable. To avoid handling massive data, it has been provided extracted with only four years of data from 2016 to 2019. To provide a Geographical map https://geohub.lacity.org/datasets/lapd-police-stations - The Spatial data for 21 LAPD community police station locations is taken from the link referenced under the column Rpt Dist No.

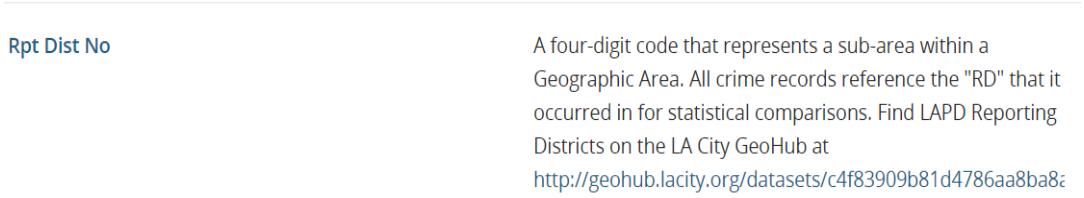


Figure 1: Report District Number

1.2 Useful

The excel file contains data on the crime incidents for the past four years from 2019 which is relevant to the analysis required to be done and contains all relevant columns like the location of the crime occurrence, the time of occurrence, weapons used and type of crime. The data is getting updated weekly.

1.3 Interpretable

Mostly interpretable. All the values of the columns and their codes are described to a major extent except for few anomalies. Some Null values in weapon Description are safely assumed as those with Crime Types that does not require or involve a weapon. The Rpt Dist No. which was necessary to link to the spatial file to create a geographic map was stored as a string in the Crime Table. There was a mismatch when tried to relate to the spatial file. So changed the syntax of "Rpt Dist No." to whole number. But the field "Vict Age" has negative and zero values. The "Vict Sex" has unexplained values like H.

1.4 Believable

The data is owned, updated and provided by the Los Angeles Police Department. In spite of some discrepancies like some missing values in the LAT, LON, the Premise of crime and some NULL values. These discrepancies are in negligible percentage and the main data is not severely impacted. It is accurate to the extent declared and some declaration has been made that some inaccuracies could be present. Some Columns like Crm Cd 2 , Crm Cd 3 , Crm Cd 4 are left with no proper data and the purpose of those Columns are not stated. They are unnecessarily occupying Database Space. The Column Mocodes is not explained clearly.

2 Pre-processing of Data

Data Validation and Fixing Data Model.

2.1 Data Errors

The dirty or invalid data is identified by filtering in excel and by looking under the marks underlying data on some model charts using bar charts and box plots. To a large extent the data is clean.

Some examples were like negative values for Age and undefined values for Gender. They were filtered out in the data source itself.

2.2 Modelling

Table Relations were chosen instead of joins to be efficient and avoid duplicate values. The Matching fields were identified in the tables so that they can be linked to relate the tables. Selected matching fields to define a relationship (no join types). Relationships defer joins to the time and context of analysis. Tableau automatically selects join types based on the fields being used in the visualization. During analysis, Tableau adjusts join types intelligently and preserves the native level of detail in the data.

The Relation of Many to one was chosen from Crime Table to the Dimension tables on Area, Crime and Premise. Apart from these the Spatial data file is also in a Many to one relation with Crime table.

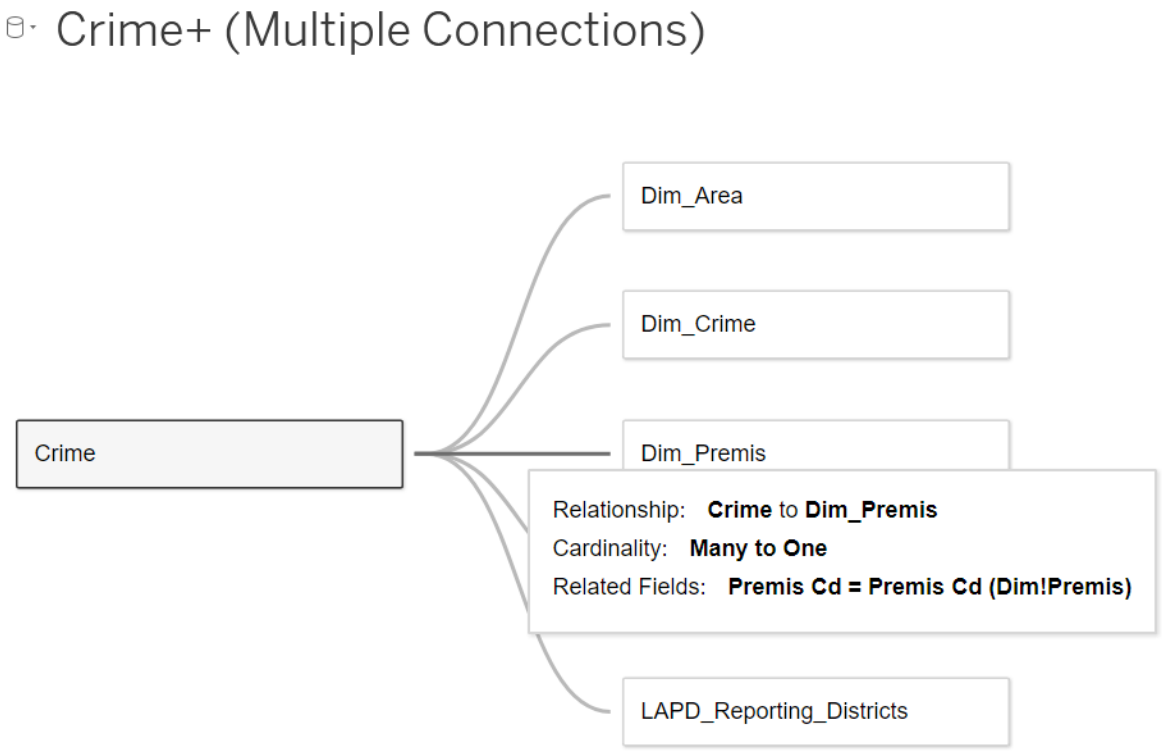


Figure 2: Crime Table relationship

Victim Age was filtered at database level to remove negative and zero values and stored back in the column. Grouping was done to create Age groups of Children, Teens and the rest of the age was grouped in steps of 20 years.

ime)N	Group Ocurence	Group Age Groups
18.227300	7PM - 8PM	20 - 40 years
18.235100	5AM - 6AM	40 - 60 years
18.257100	1AM - 2AM	20 - 40 years
18.245200	7AM - 8AM	40 - 60 years
18.266900	8PM - 9PM	20 - 40 years
18.254000	10PM - 11PM	20 - 40 years
18.260300	12MN - 1AM	20 - 40 years

Figure 3: Victim Age Group

A column Occurrence is created by grouping the time ranges in 24 Hr Military time to AM or PM.

Time interval given in minutes was grouped to create the hourly group

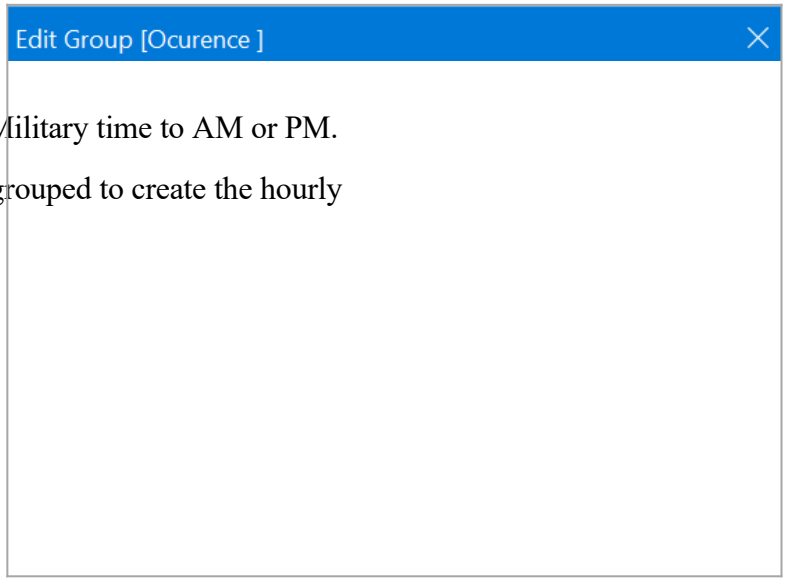


Figure 4: Occurr

Alias was created to make Victim Descent more meaningful. The details were taken from the published site

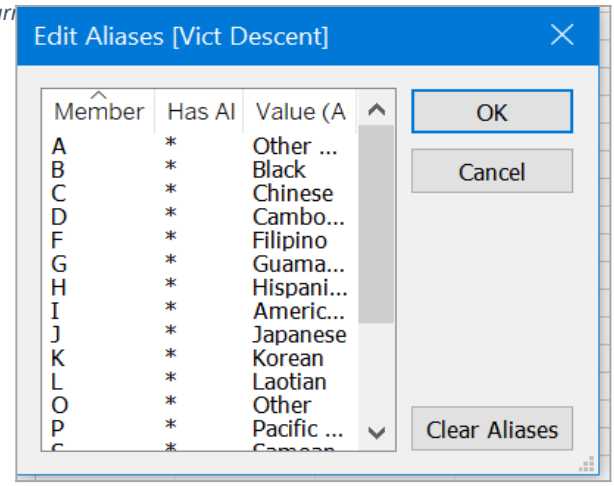


Figure 5: Victim Descent and Aliases

The Rpt Dist No. data type was changed from String to number , so that the table can have a many to one relation with the spatial data table

Crime Rpt Dist No
118
159
185
147
181
131
185

Figure 6: Report Distribution Number

Victim Age was converted to Dimension so that it is not aggregated. Victim Sex was filtered on unknown values like H

The data was prepared by studying the various codes and the valid values. Dimensions of Area, Crime Desc , Weapon Desc and Premise Desc are used to slice and dice the data so that insights can be derived. Tableau by default creates the Measure of the Crime(count) which is nothing but the count of rows. When this count is grouped by Area , we get the no. of crime incidents for that area.

3 Dashboard design considerations

The following questions were set as goals for the Dashboard

1. What are the most common crimes in Los Angeles?
2. Which area has the highest crimes rate?
3. What is the victim demography?
4. Choice of weapon and crime venue
5. Trend of Crime with Time

Good Practises*: Showing related worksheets together, Reset Filter and Current Selection, at the necessary worksheets, use of Master worksheet as filter for the other related worksheets are some of the good practises followed.*

3.1 Summary Analysis

A summary of the crime rate and the most common crime by year was planned. Simple calculation with a single labelled field was used to display the summary details on the no. of crime incidents and the highest common crime. along with a area wise summary presented in a choropleth map. The user can highlight any area to see the no. of crime events in the area highlighted. Clicking outside the highlighted area clears the selection.

Data was analyzed to conclude that **Top 5 hit areas** are a good lot for study of the crime incidents. Focus was shifted to Top 5 hit Areas, so that the data set is not huge. A 'Top 5 Area' dashboard is created.

3.2 Top 5 Hit Areas

A geographical map was created with help of spatial file downloaded. The GEOMETRY field from the spatial file is presented as a measure object with a single aggregation function COLLECT(). This aggregation makes a group of polygons coloured, labelled, and highlighted as a single mark. So, the 'Area' is selectable as a whole in the Geographical map and using colour gradient to show the total crime incidents in those areas.

On hovering over each of the areas in the graph, using the tool tip, the topmost 5 crimes in the area are shown in a simple bar chart. This is done by inserting the worksheet containing the bar chart into the Tooltip of the Top 5 hit Area Sheet. Max height and width was adjusted to 500 to accommodate the graph in the tool tip

3.3 Trend over past four years

Each area is represented by a bar color and the length of the bar gives the total crime incidents which is filtered by year by using the year in the page shelf. And animation is set for 2 sec time. As the motion chart is run, an animation of the **bar chart race** shows the trend in the crime incidents. It shows the pattern through the movement of the bars. The chart helps to identify areas with increasing crime incidents over the year and analyze them further.

3.4 Victim Demography by Area

A summary of the Victim Demography over the four years. Two Worksheets with some meaningful relation were grouped together. To increase the effectiveness of the Dashboard, a Master worksheet, Victim Demography, with all the Dimensions like Vict Race, Vict Age and Sex is created as columns of a table. Count of Crime as a Measure is placed in Color and Size shelf. This is then linked to the Area worksheet to study their distribution in the Areas. Having a Master worksheet like this will help to easily build further dashboard for extending the analysis on other dimensions. This Master worksheet is used as a Filter for the other sheets in the dashboard

CrimeType is added as new dimension to study the demography with a simple worksheet of CrimeTypes in Column and Count of Crime as a Measure.

Selecting a cell shape in the Master Worksheet, will update the Area and CrimeType Charts to understand the victim demography on them. Clicking outside the Victim Demography Table will undo the selection. Reset Filter will bring back provide a summary view of the demographic details over all the four years.

3.5 Crime type Distribution

A master sheet with table Column of Year, Quarter and Month / weekday Dimensions and the Count of Crime as Measure in Color Shelf and Size shelf is created. This sheet is used as a filter for the analysis of Time of occurrence of the crime, trend over weekday and the Crime Type Distribution. Only the top 3 crimes are analyzed in the Crime Type Distribution. The No. of crimes by time of occurrence is shown in a text table with color gradient for the top 5 areas in the Occurrence Table.

The Master sheet provides a drill down functionality to study the crime pattern by year, Quarter, Month, and weekday. Box Plot shows that lowest crime happens during middle of week around Wednesday and Mondays and Fridays seem to be on the higher side. Outliers -

These are the major contributors having very high values and show some abnormal incidents that need to be investigated and controlled. A current Selection field is shown via a caption shelf.

The line graph shows the Crime Type occurrence over Month of Year.

3.6 Weapon Analysis by Crime and premise of crime

All the four year Data is combined to create a Tree map of Occurrence Time and Weapon and bar chart of the Premise of Crime are related to study which premise is scene of crime at what time and if any weapon used or not.Weapon and Time of Occurrence Tree map cells act as filter to Premise bar chart. Disable highlighting – prevents the fading of the other dimensions other than the hovered or selected one. Reset Filter button – to bring filters to the initial default setting

3.7 Gender Summary:

A Pareto chart is done to show the areas contributing to the 80% crimes on Female.

The Dashboard designed to give a summary view of all the areas using visual cue in choropleth map

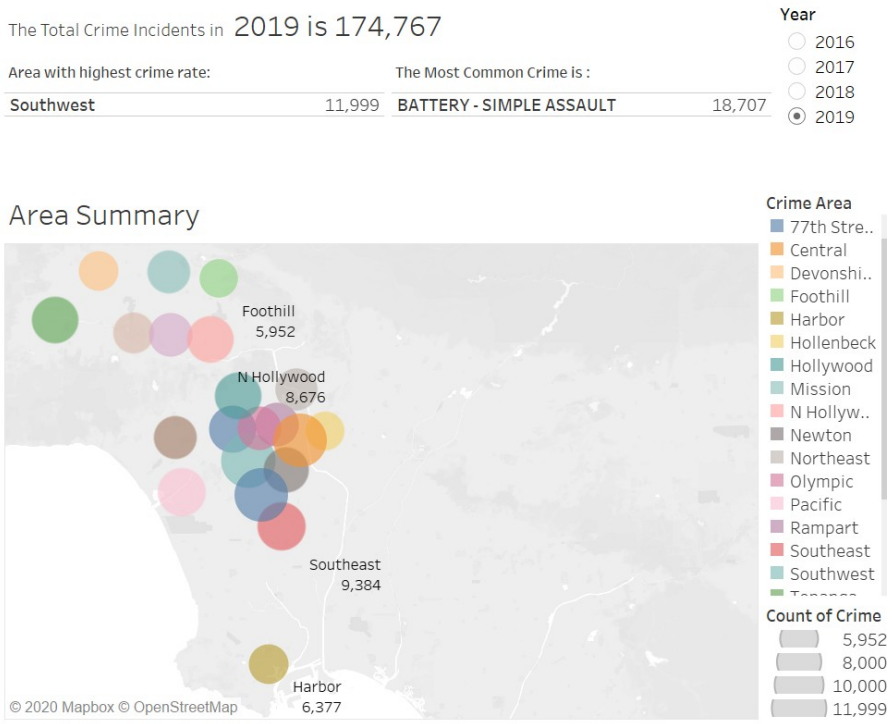


Figure 7: Area Summary

The highest crime rate area in each year and the most common crime in the areas is summarized along with a choropleth map to give a quick summary analysis of the areas crime incidents. The most basic form of visualization to map the data to color and shape is used to represent area's crime rate.

Color saturation and shapes are used to show the Metric of count of crime in Top 5 areas of the highest crime rate. Selecting all the areas will limit the visual to the top 5 areas. The bar chart

is shown from the worksheet Count Crime by Area where the top 5 crime types are listed by Area.

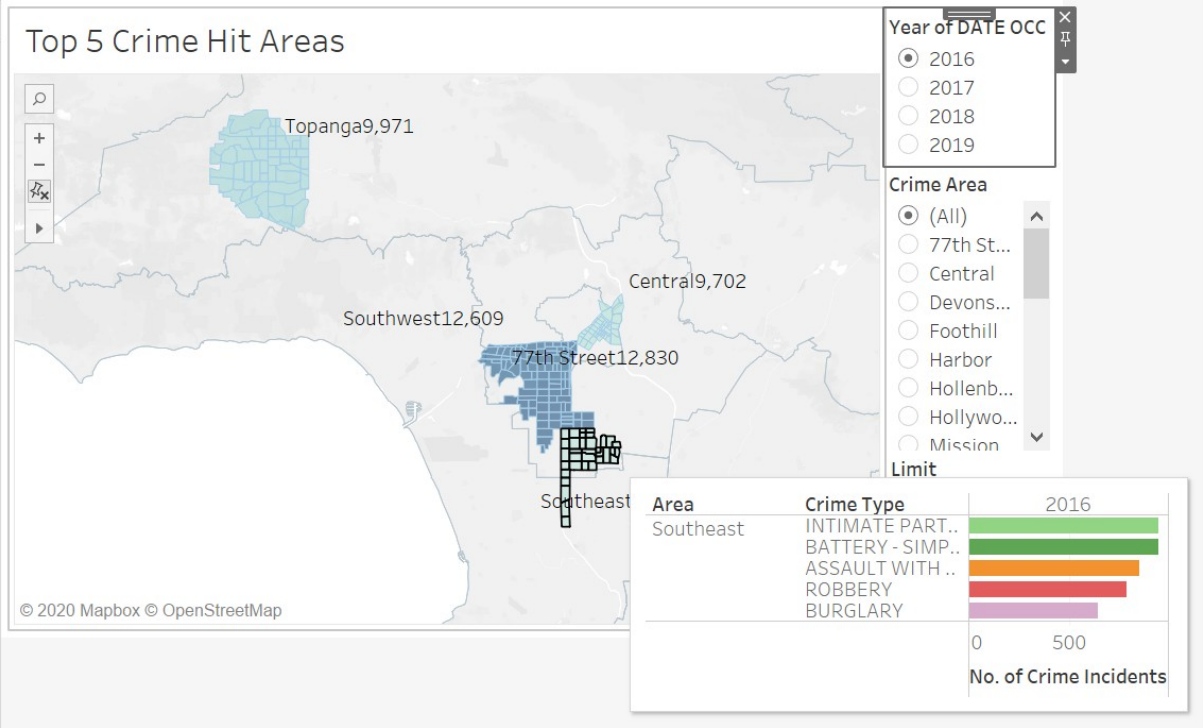


Figure 8: Top 5 Crime Areas

An Animation trend of crime rate is created to analyze the past 4 years of crime. The Count of Crime was ranked by Dimension of Area and paginated by the year on the page shelf.

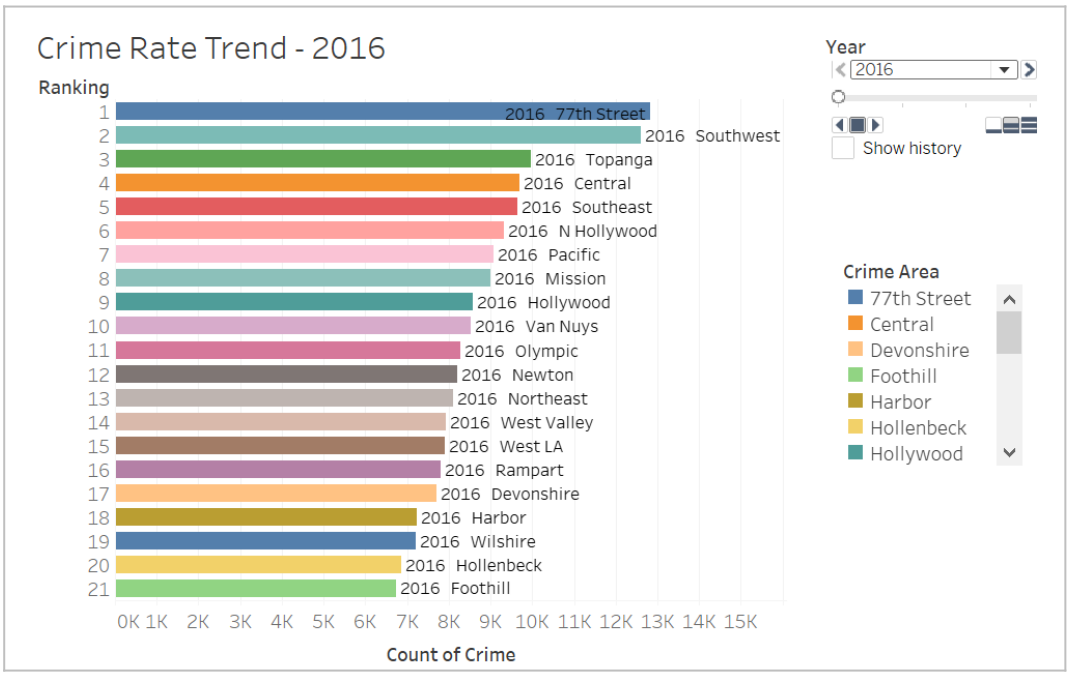
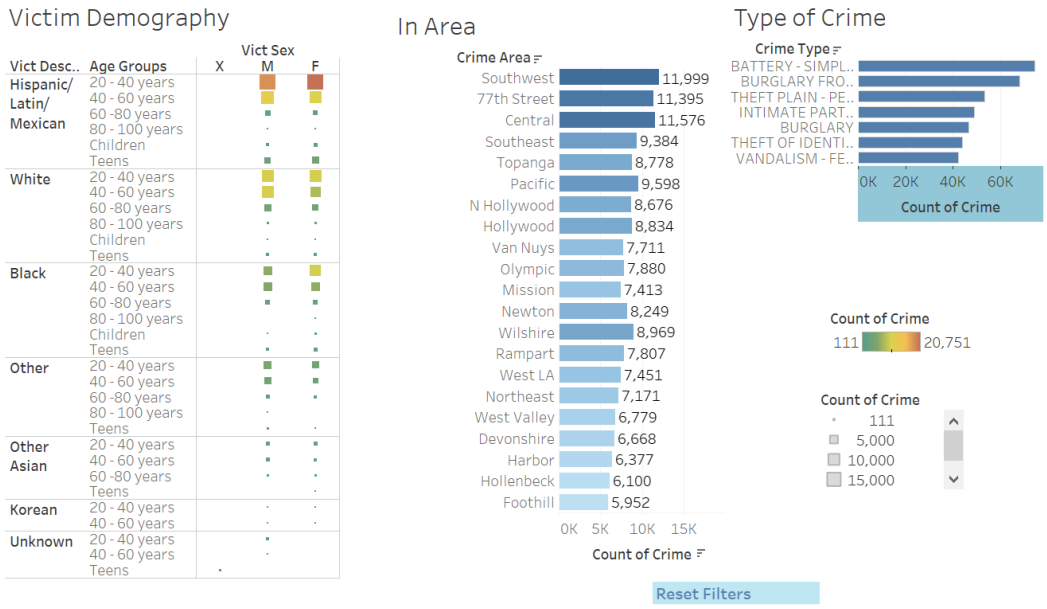


Figure 9: Crime Rate Trend

Figure 10: Victim Demography



Slicing of the crime incidents by Victim Gender and Race is done to understand better which race is affected most and how law can be enforced to protect them.

Victim Demography Sheet is set to be “use as filter”. Selecting a box representing, the Descent, Age group and Sex filters the areas and type of crimes in which they are victims. All areas are listed, and type of crimes is filtered by top 5 for easy visualization. To reset filter to show all

values click on Reset filter, Tableau does not provide option to undo the highlight, the workaround is to click just outside the selection which performs the reset of the filter

A box plot on occurrence of crime on weekdays helps to identify the peak days of crime Distribution by year:

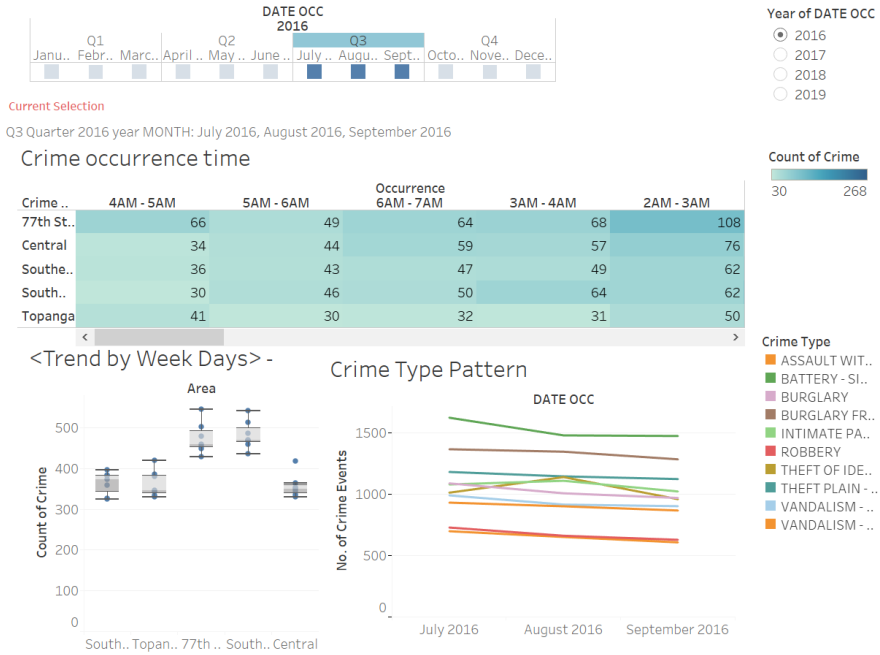


Figure 11: Crime Pattern

Weapon and Premises is analyzed to find out what aids the occurrence of crime and how it can be prevented. The availability of weapons and an apt location is helping the rime?

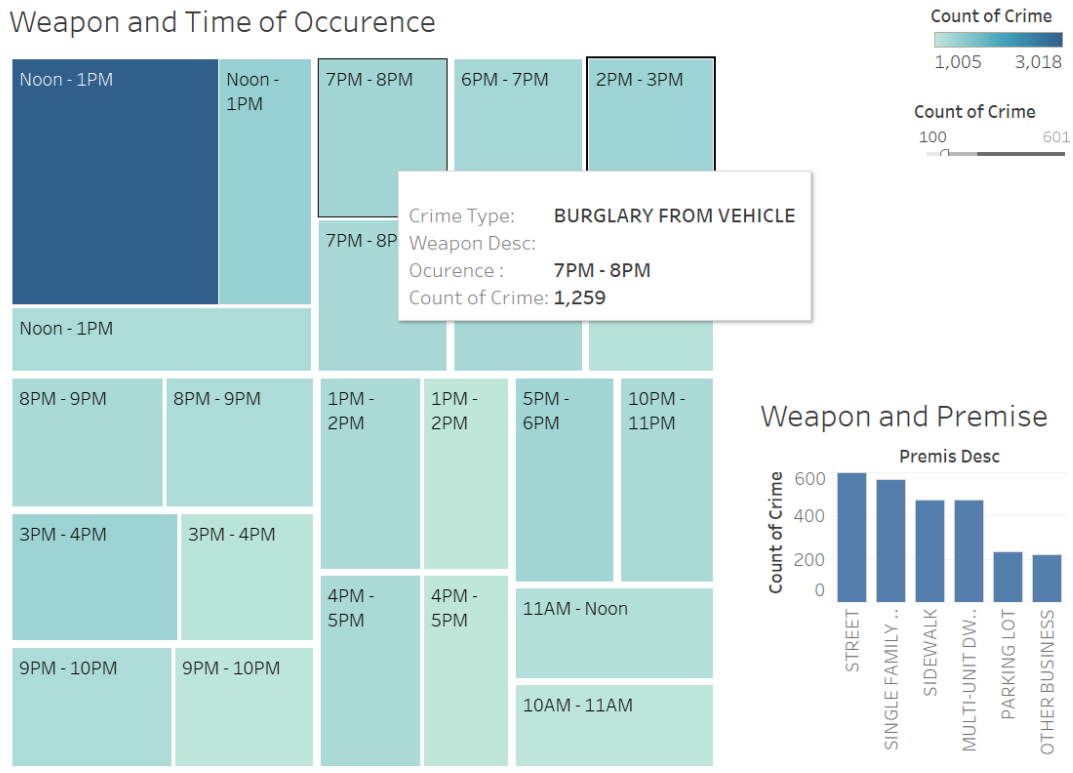


Figure 12: Premise, Weapon and Time of Occurrence

Reset filter button designed to reset filters to default value

Trend animations were designed using bar Ranking Calculation

The Victim Demography is analyzed Gender and Race. More than 80% of the crimes against male and female was analyzed to find if there is any target to gender or race.

4 Findings and recommendations

Based on the analysis of the data following are the preliminary findings:

1. Top 5 Areas having highest crime rate over the four years: SouthEast, 77th Street, Central, SouthWest and Topanaga
   They seem to have a larger proportion of victim people from Hispanic/Latin Mexican origin and Black. So, the Police Force can focus their patrols and take safety precautions in their areas. The residents should be alerted to and trained to protect themselves against these crime incidents.
2. From the Victim Demography study on the Crime Type, it is found that INTIMATE PART ASSAULT victims are Female mostly from the Hispanic/Latin Mexican origin and Black. The Age group is around 20 – 40 years

3. Single Family dwellings seem to be the premise of crime in the top 5 areas with the inclusion of streets, sidewalks and to some extent on parking lots and multifamily dwellings, with a time of occurrence around the mid afternoon
4. Most of the crime incidents seem to be associated with hand fist and bodily force fights— which indicates a likely development of verbal dispute escalating to bodily fight. Not much of Gun shots and other deadly weapons seen

Consistently Battery Simple Assault has been reported too high in all the areas

And mostly Fridays and Mondays seem to be hit with maximum crime incidents

Patrol can be increased around Mid-day in the single dwelling areas with Mexican and black population or strategies can be made to forecast scenarios of crimes in those areas and provide protection.

More Stricter Fines can be charged for the battery and assault crimes and Safety Measures can be improved through employing security guards, CCTVs for watching out on the premises of crime. A separate helpline can be announced to help victims of the Battery Simple Assault, Burglary, Vandalism. For Identity threats, the people can be educated on the precautions needed for sharing personal information like Ids etc…