



University of the  
West of England

**Business Intelligence and Data Visualization**

**UFCFKR-15-M**

## **Personal Reflective Diary**

**Module Leader:**

**Dr. Kamran Soomro**

**Tutor:**

**Dr. Haixia Liu**

**Submitted by:**

**Name : Parashkumar Shah (21051422)**

**Project: 3 (US Crime)**

**Group : 7**

# A Journey of learnings and Improvements

## Background:

My previous business experience and possessing skills like – Understanding Business requirements, problem finding and solving, presentation and storytelling motivated me to become a bridge between technicality and business expertise, which made me choose “Business Intelligence and Data Visualization module”. When looking at the content and the delivery of this module, and some expert suggestions from industry advisors, this proved to be one of my best decisions. Also, A big thanks to our module leader Dr. Kamran Soomro and tutor Dr. Haixia Liu for well-structured deliveries and promptly answering all the queries. The Concept of working in teams gave me a good exposure towards the industry-way of working and meeting awesome people in my team, who are equally hard working and participating as me.

## Getting started with Tableau and Visualization:

Enrolling to a LinkedIn learning Course by (Curt, 2021) along with the theoretical and practical weekly sessions from the first session dated 2-Feb-2022, helped me improve my understanding on tableau basics including attaching file, connecting tables, data cleaning, calculated fields, basic graphs, maps, interactive dashboards and much more. I started working on soft skills after knowing its importance explained by tutor along with technical skill in the same session. The 15 types of presentation skills described by (Matthew, 2021) helped me understand the importance of soft skills such as which I applied in the coursework

The second week session delivered by (Paul, 2022) gave a brief understanding about the twins of visualization – Scientific and Information.

Some further research opened new dimensions on having in-depth understanding for visualization. From the presentation by (Allison, 2016) I gained in-depth understanding of the information visualization, a basic visualization model Data > Visualization > Visual Transfer > Visual Bandwidth, and types of dimensional data. I really feel this helped me built pillars in the field of analysis and data visualization.

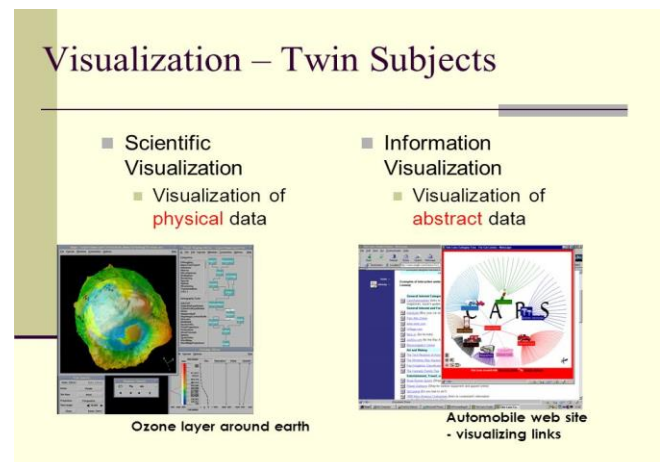


Image Source : [Information Visualization](#)

## Beginning:

I appreciate the idea of working in team for analysis and visualization as developed some most useful teamwork skills, altogether with writing a personal diary encouraged to learn and research more. It started with a team formation where we accumulated people from different background and courses. We choose a project title where we mutually agreed to work on Project-3 US Crime Data on 16-Feb-2022 after thoroughly understanding all datasets where we saw more opportunity to perform well due to its richer data with 633454 records and 24 fields. Here, I knew I can have a good opportunity to research, learn new things and walk towards mastering the art of Data Visualization using a widely accepted, versatile, and free-to-students tool Tableau.

## Finding Missing values:

As guided on the initial step for finding missing values and tools to deal with it in third week session by (Haixia, 2022) I initialized by overviewing of data set using MS excel. I really feel a well explained data description could have been provided for better understanding. This stage helped me to understand the characteristics of data and to trust on what variables at what level in the next stage of analysis and visualization. I got to learn how to change blank to “UNKNOWN” and the importance of rectification of spelling errors which could create additional groups, and how that could be solved in Tableau in 2<sup>nd</sup> Advisory meeting by (Azhar, 2022). This is the summary of my findings in the initial observation stage.

**Initial Observation to Understand Data - Total Records: 638454, Total Fields: 24**

Column No	Column Name	No. of records				Notes
		Unknown	0	Blank	NOISE	
1	Record ID					Checked. Seems all okay.
2	Agency Code					Checked. Difficult to figure any wrong or mis records. Seems okay for now.
3	Agency name	83				Checked. Very few unknown found.
4	Agency Type					Checked. Seems all okay.
5	City					Possibilities for spelling errors. Need to check further
6	State					Checked. Seems all okay.
7	Year					Checked. Seems all okay.
8	Month					Checked. Seems all okay.
9	Incident					what does it mean? Asked professor for clarification.
10	Crime Type					Checked. Seems all okay.
11	Crime Solved					Checked. Seems all okay.
12	Victim Sex	984				Checked. Very Few unknown found.
13	Victim Age		8444		974	974 records found with age 998, and few records with 0 age - Can be unborn or less than 1 year!
14	Victim Race	6676				Checked. Few unknown found.
15	Victim Ethnicity	368303				Checked. Many unknown found. Doubt if it can show accurate result in analysis!
16	Perpetrator Sex	190365				Checked. Many unknown found. Doubt if it can show accurate result in analysis!
17	Perpetrator Age		216327	1		Checked. Many records with 0 age found. Perpetrator cannot logically be of 0 age.!
18	Perpetrator Race	196047				Checked. Many unknown found. Doubt if it can show accurate result in analysis!
19	Perpetrator Ethnicity	446410				Checked. Many unknown found. Doubt if it can show accurate result in analysis!
20	Relationship	273013				Unclear! Victim's relation to perpetrator or perpetrator's relation to victim?
21	Weapon	33192				Checked. Few unknown found.
22	Victim Count		586061			Unclear! Does this count includes or excludes the prime victim mentioned in record? Too many unknown
23	Perpetrator Count		558843			Checked. Too many unknown.
24	Record Source					Checked. Seems all okay.

After Finding the missing values and the anomalies, I researched on how to deal with it, and I found that the two best ways Data Scientists do are – Imputation and Removing Data by (Huan *et al.*, 2013)

- Before deciding which approach to execute, the reason behind the data is found.
  - Missing at Random (MAR)
  - Missing completely at random (MCAR)
  - Missing Not at Random (MNAR)
- While removing data is to delete missing records to reduce bias. It has 3 methods:
  - Listwise
  - Pairwise and
  - Dropping Variables
- Imputation is good to execute when the proportion of missing data is less, if it is more than the natural variation is missing and an effective model cannot be built. The methods to do imputations are:
  - Central Tendency – Mean, Median, Mode
  - Time-Series specific trend
  - Last Observation Carried Forward (LOFC) & Next Observation Carried backward (NOCB)
  - Linear Imputation
  - Seasonal Adjustment with linear interpolation
  - Multiple Imputation
  - KNN (K Nearest Neighbors)

## Data Cleaning:

I started with the spell check in tableau for this I followed this step. At first it appeared to be the same city name but a spelling error, but further research helped me understand about them being different.

The result showed:

- Alleghany is in Virginia
- Allegheny is in Pennsylvania
- Allegany is in New York

The Screenshots has been attached below.

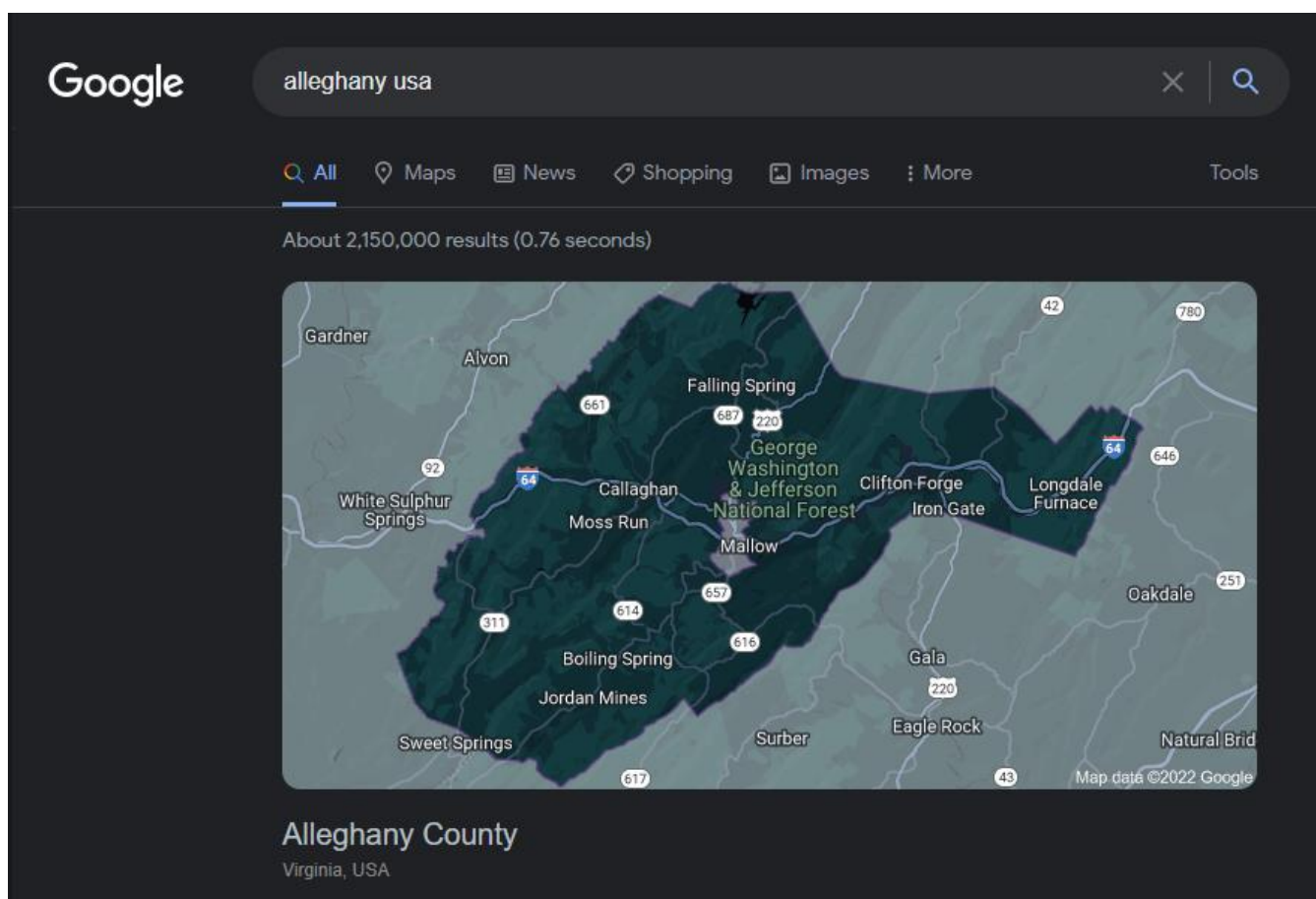
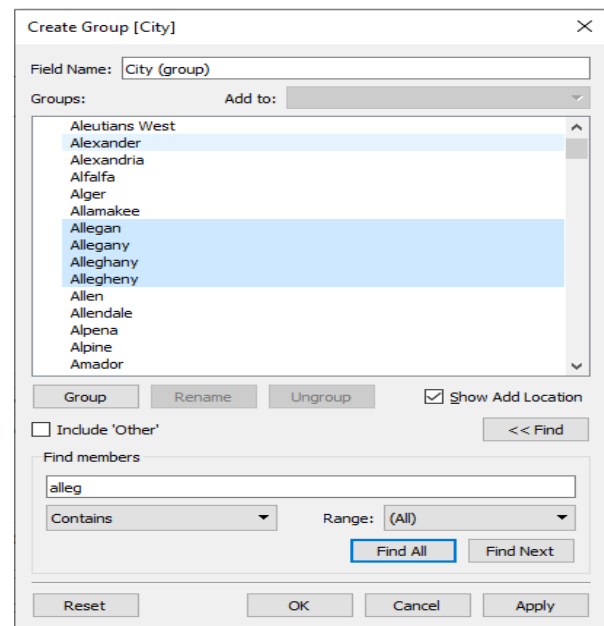
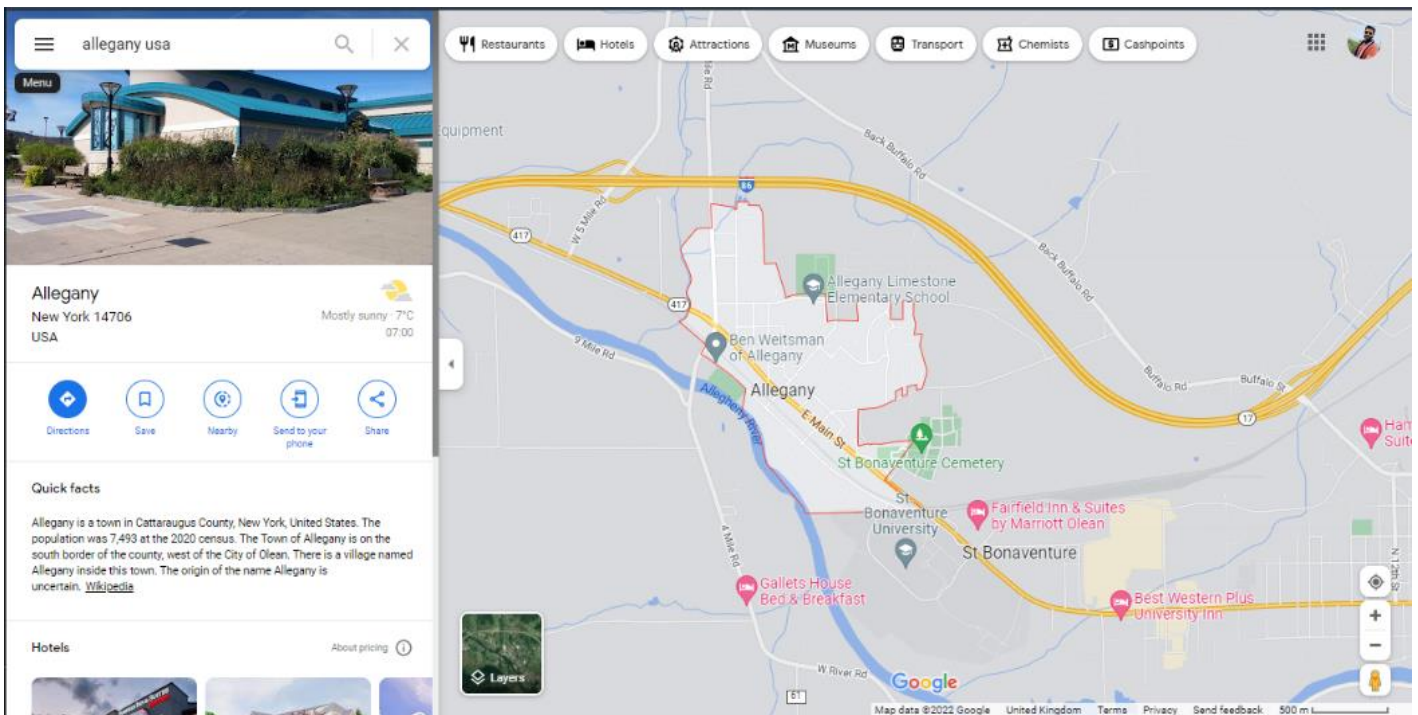
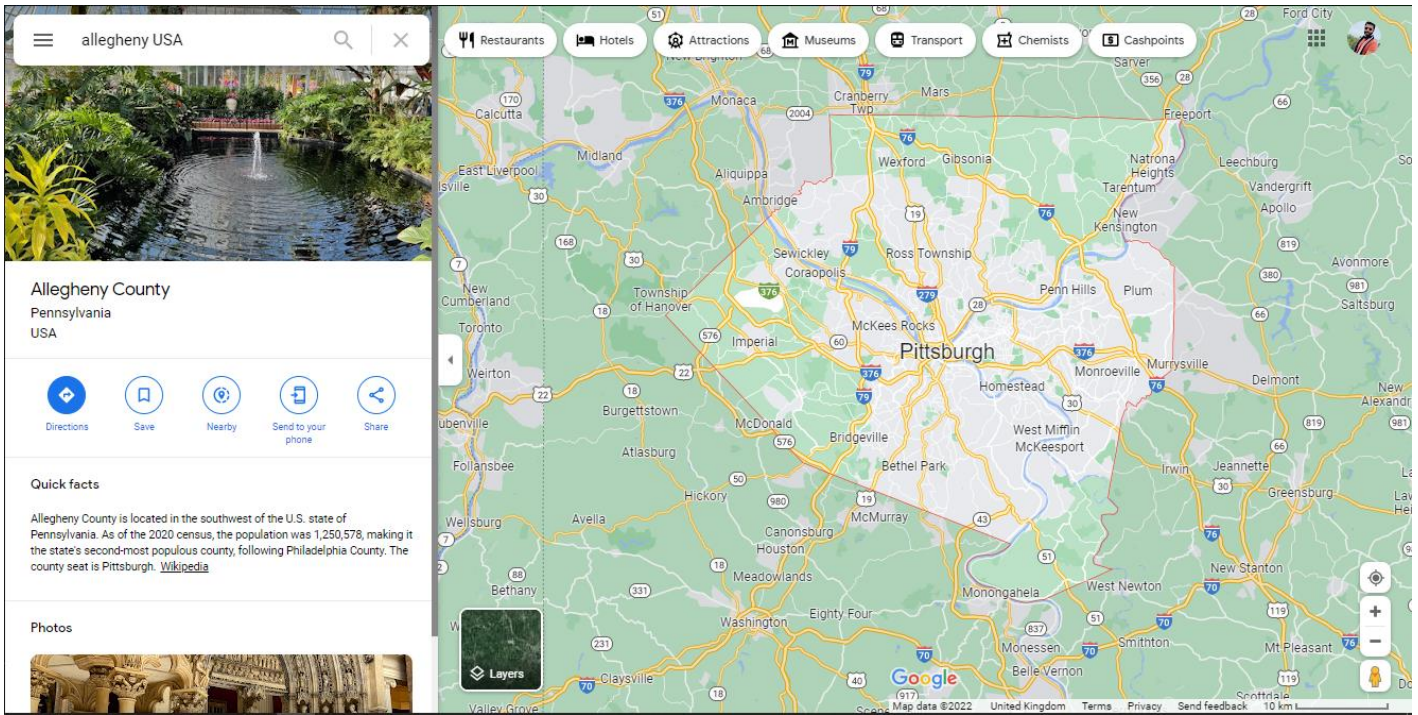


Image Source : [Google](https://www.google.com/maps/@37.25,80.5,12z)





Conversation with tutor helped me understand some of the column's values and to interpret those better.

Reply Reply all Forward Archive Delete Set flag ...

### Some Doubts in Project - 3 US Crime Data Set

**PS** Parashkumar Shah (Student) <Parashkumar2.Shah@live.uwe.ac.uk>  
4/8/2022 11:22 AM

To: Haixia Liu Cc: Kamran Soomro

Hello Professor!

It feels strange not have data description with this data set!  
I have some critical doubts, and they are bothering me to understand data set completely!

**Incident Column** = What does this column means?

**Victim Count** = Is it total Victims (including the prime victim mentioned in the record)? Or is it the other Victims ( Excluding the prime Victim mentioned)

**Relationship** = It is totally unclear who's relation to whom Should it be taken as Victim's Relation to Perpetrator? Or Perpetrator's relation to victim?


Also I need to make some logical assumptions!

**Perpetrator age** = It is obvious that it cannot be 0! But there are significantly **high number** of records. (216327 total to be precise), Thus should I take average age for these or should remove These records?

**Victim Age** = Technically it can be 0! As a victim could be an unborn child! But there are few records with 0. (8445 total to be precise). Even if I consider it to be removed it will not make much difference. Rather I can make logical assumption and use some of these records, based on relation, Where if the victim is son or daughter, it could be considered as unborn child.! But for that I need the clarification of **Relationship** Column.

Yours Truly  
Parashkumar Shah (Parash)  
Student ID : 21051422

Incident: Haixia referred me to this [link](#) and mentioned about API. According to her the incident columns are not showing accurate statistics, but according to my findings I believe Incident column is the most important column for analysis as it shows the repetition of the events.

 **Haixia Liu** <Haixia.Liu@uwe.ac.uk>  
4/16/2022 2:00 AM

To: Parashkumar Shah (Student)

**Incident column**: it looks like it's **incremented/accumulated** based on **month** in a particular location and a particular year.

40	AKASP00	Alaska Sta/State Polic Juneau	Alaska	1980 October	3	Manslaughter by Negligence	Yes	Male
41	AKASP00	Alaska Sta/State Polic Juneau	Alaska	1980 October	4	Manslaughter by Negligence	Yes	Male
42	AKASP00	Alaska Sta/State Polic Juneau	Alaska	1980 November	1	Murder or Manslaughter	Yes	Male
43	AKASP00	Alaska Sta/State Polic Juneau	Alaska	1980 November	2	Murder or Manslaughter	Yes	Male
44	AKASP00	Alaska Sta/State Polic Juneau	Alaska	1980 November	3	Manslaughter by Negligence	Yes	Female
45	AKASP00	Alaska Sta/State Polic Juneau	Alaska	1980 November	4	Manslaughter by Negligence	Yes	Female
46	AKASP00	Alaska Sta/State Polic Juneau	Alaska	1980 December	1	Murder or Manslaughter	Yes	Male
47	AKASP00	Alaska Sta/State Polic Juneau	Alaska	1980 December	2	Manslaughter by Negligence	Yes	Male
48	AL00100	Jefferson Sheriff Jefferson	Alabama	1980 February	1	Murder or Manslaughter	Yes	Male
49	AL00100	Jefferson Sheriff Jefferson	Alabama	1980 March	1	Murder or Manslaughter	Yes	Male
50	AL00100	Jefferson Sheriff Jefferson	Alabama	1980 April	1	Murder or Manslaughter	Yes	Male
51	AL00100	Jefferson Sheriff Jefferson	Alabama	1980 May	1	Murder or Manslaughter	Yes	Male
52	AL00100	Jefferson Sheriff Jefferson	Alabama	1980 June	1	Murder or Manslaughter	Yes	Male
53	AL00100	Jefferson Sheriff Jefferson	Alabama	1980 July	1	Murder or Manslaughter	Yes	Male
54	AL00100	Jefferson Sheriff Jefferson	Alabama	1980 July	2	Murder or Manslaughter	Yes	Male
55	AL00100	Jefferson Sheriff Jefferson	Alabama	1980 July	3	Murder or Manslaughter	Yes	Male

It could confuse us. For me, **Row 43 doesn't** mean that there were 2 crimes in Alaska Nov 1980 with the crimetype murder or manslaughter. So how incident column was created? if you look at the link I sent you several days ago, you would notice that we can get the raw data from their API. And we can 'generate' some numbers from the raw data.

Victim Count, Perpetrator Count: As per my understanding and solution provided by Haixia on 16-April-2022 in email conversation, I came to conclusion that these counts are not inclusive of the main party as almost all the records are 0. With more understanding I found that those are missing values referred in 1<sup>st</sup> [advisor meeting](#) (Azhar, 2022).

Victim Age: 720 records out of 974 are anomalies as mentioned above thus I decided to EXCLUDE those.

Victim Sex: There is a relation between Missing records in “victim sex” and the records with 998 values in “victims age”. As those records are 720 in count, I decided to delete those record or rather EXCLUDE those in tableau while analysis.

Victim Ethnicity, Perpetrator Ethnicity: As more than 60% of the data is missing, I decided to drop these columns out of my analysis.

Perpetrator sex, Perpetrator Age, Perpetrator Race: Quite not sure about what methods to use for imputations as more than of the records are unknown, thus EXCLUDED the missing values.

Relationship: I am not sure as the missing data is more than 40%, For now I decided to Exclude those as imputation seems unreasonable.

## **Research for Potential Stake holders and finding KPI:**

There are some Business questions given with the project. But I am Hungry! I want to analyze this data set from all angles, thus I tried looking for how other people have approached to this data set and found some good and bad analysis and visualizations. Also, I researched on how to improvise.

To get some inspiration on possible visualization and different perspectives on analyzing the data I referred [Tableau Public](#).

- [US Crime Dashboard by Neha chanchani](#): Visually looks decent but there is too much in the Dashboard. The correlation charts on top doesn't have Y-Axis label and without labelling it cannot show the true information and leads to certain assumptions (Mark, 2018). Also, it fails to match the Screen resolution ratio, using scroller makes it more difficult.
- [US State Crime Rates 2010 by Mike Klaczynski](#): It is an extremely poor dashboard and the time series chart showing incorrect information comparing different data with different scale as mentioned by (Stephanie).
- [Crime by location tool by Crime Statistics Agency Victoria](#): Very well story by classifying each category of information on separate page. Overall good color combination, just to be colorblind friendly green and red could not be used together as suggested by (Anthony, 2010). Effective use of male and female symbols.

The analysis of this crime data set can help several parties in many ways.

- Government – Framing policies, Issuing Budget.
- Police forces – level of involvement required for maintaining peace, Most crime active areas.
- Law enforcement agencies – For establishing laws according to the behavior of crime patterns.
- Insurance agencies – To make policies and charge premium accordingly.
- Health sector – The need of creating new hospitals and keeping emergency vehicles.

Key Performance Indicator (KPI)

- Studying the geographics according to state and city – The places with the most and the least number of crimes.
- Studying the age group involved – The most vulnerable age group to the crime and the age group most active in executing crime.



- Time Series analysis – The period with the maximum crime which can be found according to year month and seasons and the trend of crime going up or down for future predictions.
- Crime according to Sex – Which Sex is most affected by the crime.
- Weapons – what are the weapons most used in crimes.
- Relationship – By finding the pattern of crime and relation involved we can figure out most vulnerable relations.

## **Exploratory Data Analysis:**

EDA is a crucial part of the Analysis process and helps understand the summary of the main characteristics in data (Sonal, 2021). It is like asking data what it has got to offer!

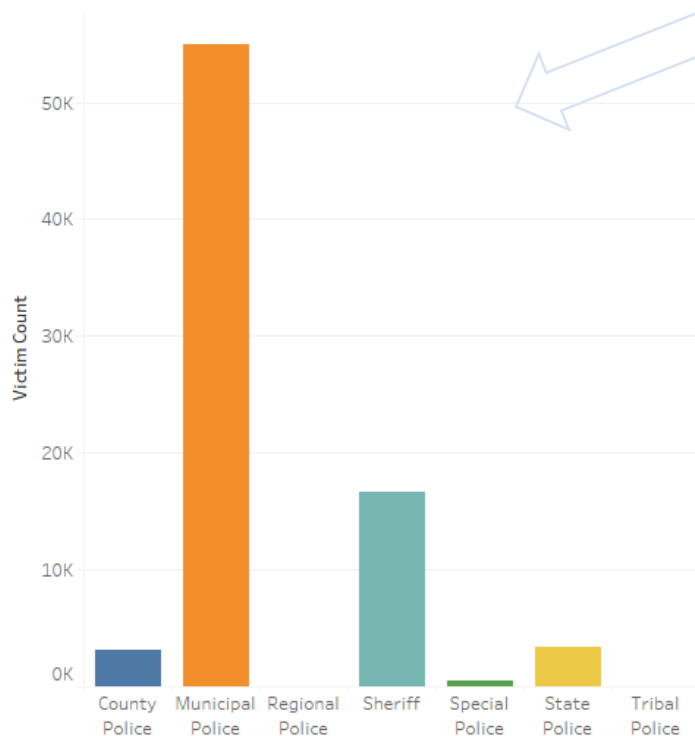
	UNIVARIATE	MUTIVARIATE
<b>Graphical</b>	<ul style="list-style-type: none"> <li>• <b>Quantitative Variable:</b> <ul style="list-style-type: none"> <li>• Histogram</li> <li>• Boxplots</li> <li>• Normal QQ-plot</li> </ul> </li> <li>• <b>Categorical Variable:</b> Bar Charts</li> <li>• Time data – Line Plot</li> </ul>	<ul style="list-style-type: none"> <li>• <b>One Categorical and One Quantitative Variable:</b> <ul style="list-style-type: none"> <li>• Side-by-side Boxplots</li> </ul> </li> <li>• <b>Two or More Categorical Variables:</b> <ul style="list-style-type: none"> <li>• Grouped Bar Chart</li> </ul> </li> <li>• <b>Two or More Quantitative Variables:</b> <ul style="list-style-type: none"> <li>• Scatterplot</li> <li>• Correlation Heatmap</li> <li>• Pairplot</li> </ul> </li> <li>• Missing Data Detection</li> </ul>
<b>Non-Graphical</b>	<ul style="list-style-type: none"> <li>• <b>Categorical Variable:</b> tabular representation of frequency (or relative frequency)</li> <li>• <b>Quantitative Variable:</b> <ul style="list-style-type: none"> <li>• Location (mean, median)</li> <li>• Spread (IQR, std dev, range)</li> <li>• Modality (mode)</li> <li>• Shape (skewness, kurtosis)</li> <li>• Outliers</li> </ul> </li> <li>• Missing Data Detection</li> </ul>	<ul style="list-style-type: none"> <li>• <b>One Categorical and One Quantitative Variable:</b> <i>standard univariate nongraphical statistics for the quantitative variables separately for each level of the categorical variable.</i> <ul style="list-style-type: none"> <li>• Mean</li> <li>• Median</li> <li>• Range and Spread measures</li> </ul> </li> <li>• <b>Two or More Quantitative Variables:</b> <ul style="list-style-type: none"> <li>• Correlation</li> <li>• Covariance</li> <li>• Descriptive stat per</li> </ul> </li> <li>• Missing Data Detection</li> </ul>

Image Source : [EDA Cheat sheet](#)

Based on the above cheat sheet, I carried out following EDA on Crime Data Set. I felt it is impossible to do accurate analysis without going through this stage. This gave me a direction to look forward into the direction that data tells. While carrying out this task I faced some problems for which I referred some resources and there are some insights that I found from the data set. Both the above details are mentioned below with the graphs.

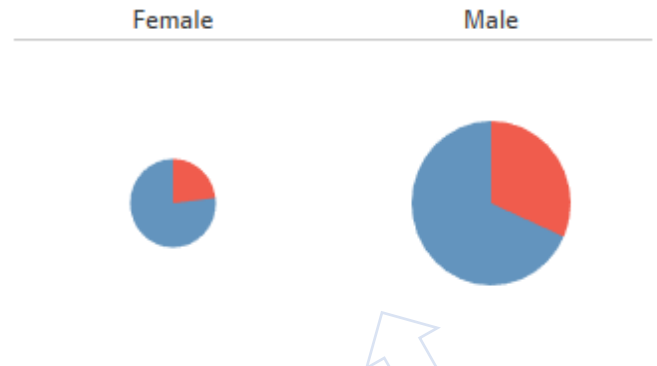


Number Of Crimes according to Agency Type



The Simple bar graph shows that Municipal policies has highest number of recorded cases. It is more than twice compared to the total of all other agency types.

Crime solved Based on Victim Sex



Pie chart at the bottom shows that the crime solving ratio of females is better than crime solving ratio of Male victims. With the help of Design Elements in visual literacy the use of color specifies the positive and negative results.

In the below table, I found that the most crimes happen between same races. Where Black and White Races are significantly higher in crimes than others.

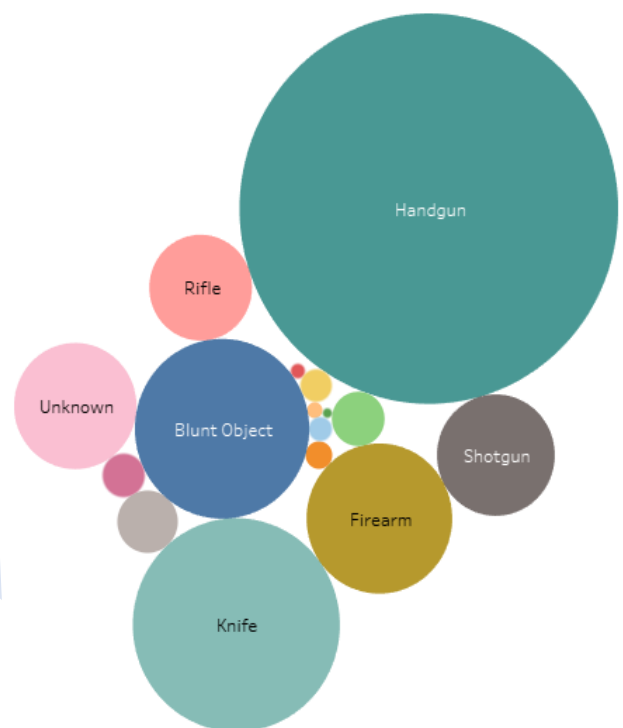
Grand Total in Analysis tab, Explained by Tutor helped improvise the chart.

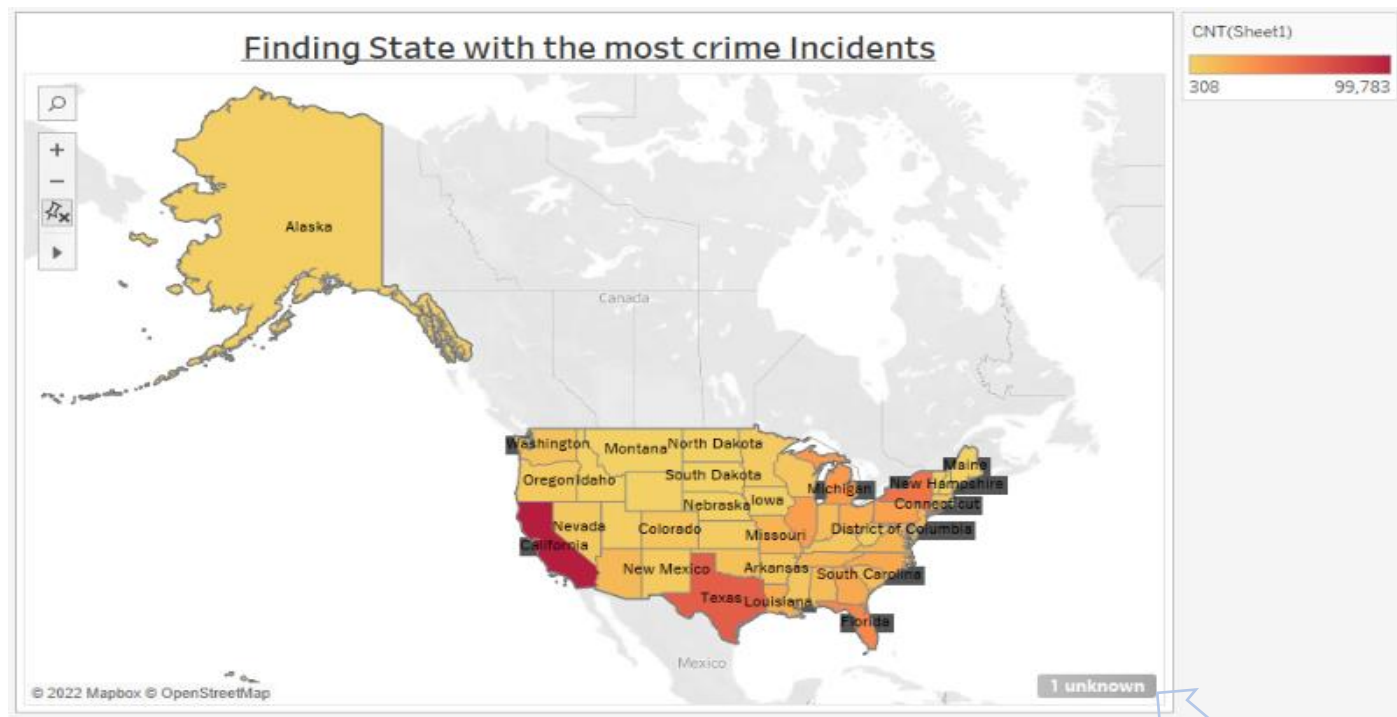
Relation between Victim Race and Perpetrator Race

Perpetrator Race	Victim Race				Grand Total
	Asian/ Pacific Isl..	Black	Native American..	White	
Asian/Pacific Islander	3,859	561	32	1,532	5,984
Black	1,249	179,258	411	32,635	213,553
Native American/Alaska N..	34	191	2,006	1,350	3,581
White	1,708	17,004	1,158	197,179	217,049
Grand Total	6,850	197,014	3,607	232,696	440,167

From this chart, I can find that the most frequent crime object is handgun, Followed by Knife, Blunt Object, and firearm respectively. Design Element - Size

Most used Object in Crime

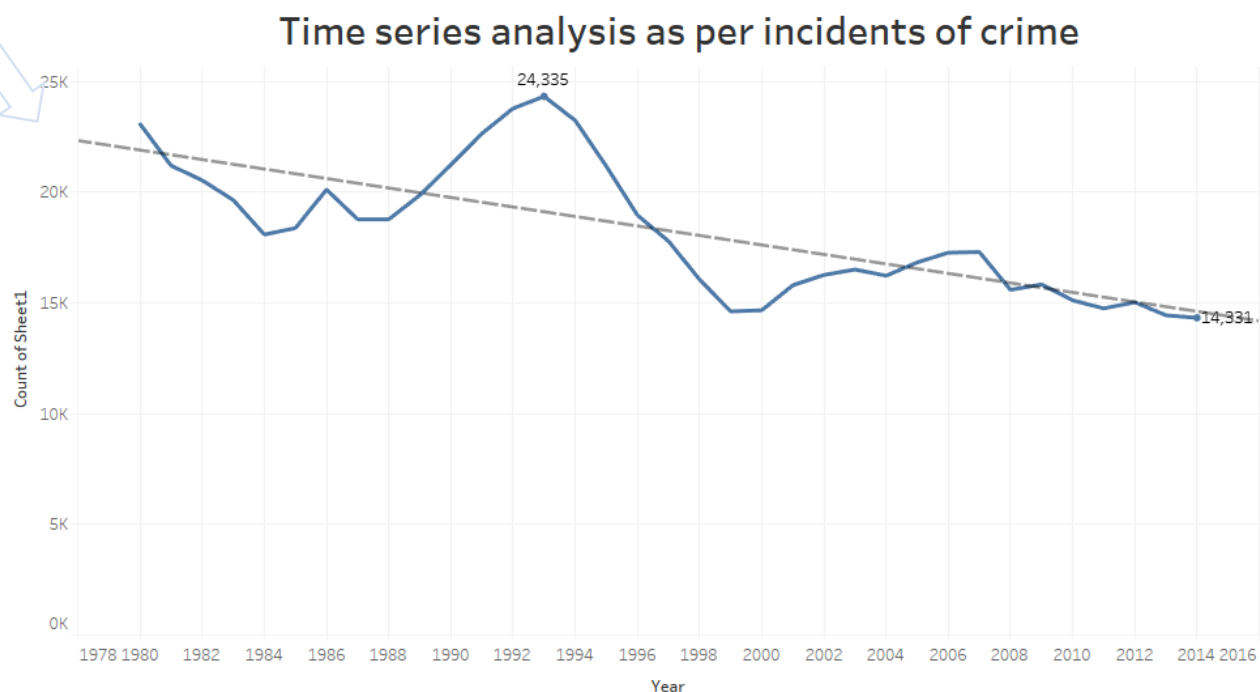




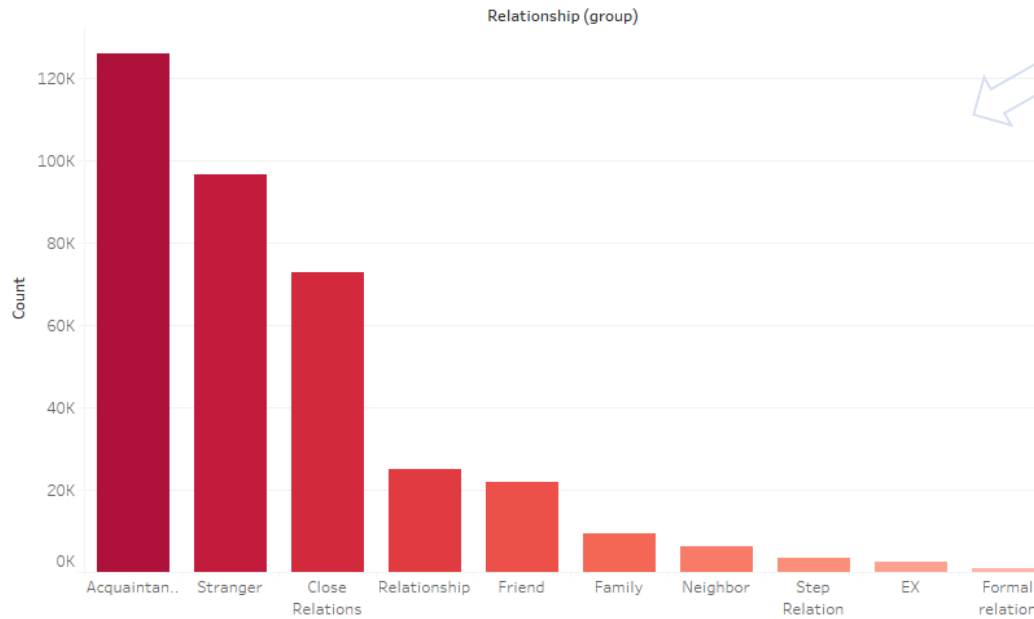
I found Heat Map handy When I wanted to find the region/s with the minimum and maximum crimes. The Unknown values are the one not recognized by tableau, in certain cases it is better to filter out unknowns from the map - week 5 by (Haixia, 2022). Initial analysis shows that California has highest crime followed by Texas and New York.

Time series analysis are helpful in extracting meaningful insights from the data (Wikipedia, 2022). Forecasting models can be applied to predict future events. My findings shows that 1993 was the year with highest crimes and a major turning point which gave a downward movement to crimes.

I referred (Tableau.com,2022) to learn how to plot the trend line of time series. Analysis > Trend Lines > Show Trend Lines



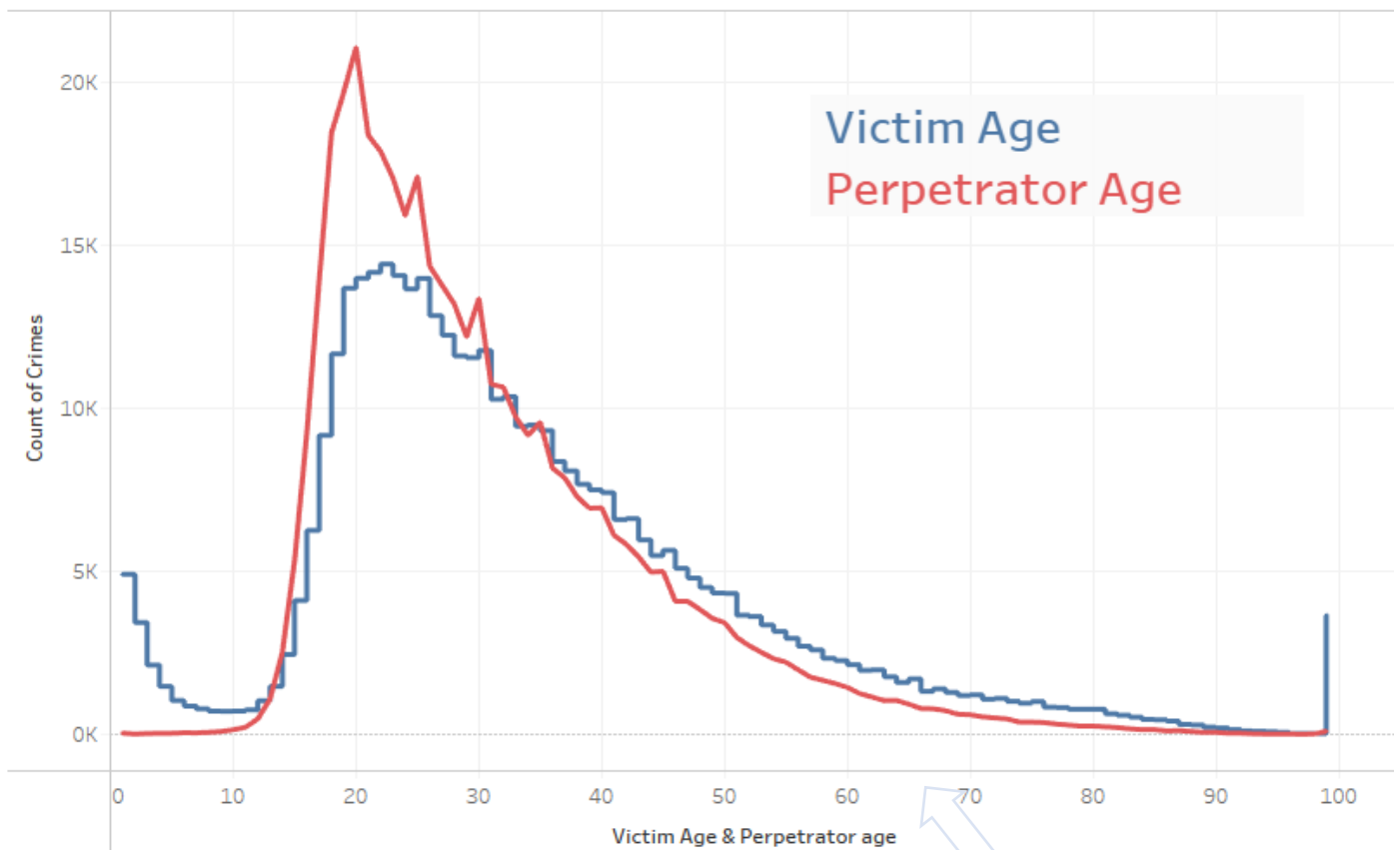
## Finding the most common Relationship between Perpetrator and Victim



I have created groups according to the closeness of the relation. For example, Relationship includes Boyfriend, Girlfriend and Boyfriend/Girlfriend.

I have used color pallet to show the reducing effect.

Finally, the use of Bar graph to for comparison of the categories of relationship, as the change is significantly large after first three classes leaned from (Graphing tutorial).



For find participating age in victim and perpetrator I referred [tableau.com](https://tableau.com) and learned that it can be done by Right click on 2<sup>nd</sup> tab > Dual Axis. After removing missing values my analysis shows that the age group of 17-21 years is highly active in perpetrator, while age group of 17-35 is most likely to be targeted in victim. I find annotate tool very useful as it helps make my Graphs look more detailed.

I realized the importance of Visual literacy which was a wonderful session in week 4 on 23-Feb-2022 by (Haixia, 2022) when I started creating my own charts. Here I came to know how use Design elements – Line, Color, Texture, Shape and Space. Also, the identification of Design Principles – Emphasis, Balance, Harmony, Variety, Movement, Proportion, Rhythm, Unity in visualization helps make a good deliverable. After the EDA Stage we had a group meeting on 24-Feb-2022, Where we Discussed our findings and further planned to research on business questions.

## Further Analysis and Visualization

In the [Google Data Analytics Certificate](#) I developed some interesting skills in tableau. The very initial lesson I learned is to ask 5 Why's to get to the root cause which helped improve my analytical skills. I learned about the types of Data Analysis - Descriptive, Diagnostic, Predictive, Prescriptive, and Advanced in Week 4 session on 23-February by (Haixia, 2022). From that I felt Descriptive and Diagnostic analysis in US Crime data set would be perfect as the dataset consisted mostly numerical data and by analyzing the past data of 34 years, solution of some root cause can be found out.

- Firstly, I looked up into the EDA in map. The figures of crimes seemed inappropriate an incomplete without a second referential figure. Because a population of the state at a particular time period might influence the ratio. Thus I downloaded the US Census Data from 1980 – 2020 from [Wikipedia](#) and averaged the difference for the middle years as the census is calculated after every 10 years. I used XLOOKUP in Excel to populate the new data as according to states and particular years. I than connected the new Excel file to the current one in Tableau. Further, I calculated the crimes per 100000 population with the following calculation, I got this idea from the analysis report by (Rebecca, 2022).

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
1		1,980	1,981	1,982	1,983	1,984	1,985	1,986	1,987	1,988	1,989	1,990	1,991	1,992	1,993	1,994
2	California	23,667,902	24,220,517	24,773,131	25,325,746	25,878,360	26,430,975	26,983,589	27,536,204	28,088,818	28,641,433	29,194,047	29,746,661	30,299,276	30,851,891	31,404,506
3	New York	17,558,072	17,601,310	17,644,549	17,687,787	17,731,025	17,774,264	17,817,502	17,860,740	17,903,978	17,947,217	17,990,455	18,033,693	18,076,932	18,120,170	18,163,408
4	Texas	14,229,191	14,504,923	14,780,655	15,056,387	15,332,119	15,607,851	15,883,582	16,159,314	16,435,046	16,710,778	16,986,510	17,262,242	17,537,974	17,813,706	18,089,438
5	Pennsylvania	11,863,895	11,865,670	11,867,445	11,869,219	11,870,994	11,872,769	11,874,544	11,876,319	11,878,093	11,879,868	11,881,643	11,883,418	11,885,193	11,886,968	11,888,743
6	Florida	9,746,324	10,065,484	10,384,644	10,703,805	11,022,965	11,342,125	11,661,285	11,980,445	12,299,605	12,618,766	12,937,926	13,257,086	13,576,246	13,895,406	14,214,566
7	Illinois	11,426,518	11,426,926	11,427,335	11,427,743	11,428,152	11,428,560	11,428,968	11,429,377	11,429,785	11,430,194	11,430,602	11,431,010	11,431,418	11,431,826	11,432,234
8	Ohio	10,797,630	10,802,579	10,807,527	10,812,476	10,817,424	10,822,373	10,827,321	10,832,270	10,837,218	10,842,167	10,847,115	10,852,063	10,857,011	10,861,959	10,866,907
9	Michigan	9,262,078	9,265,400	9,268,722	9,272,044	9,275,366	9,278,688	9,282,009	9,285,331	9,288,653	9,291,975	9,295,297	9,298,619	9,301,941	9,305,263	9,308,585
10	New Jersey	7,364,823	7,401,360	7,437,896	7,474,433	7,510,969	7,547,506	7,584,042	7,620,579	7,657,115	7,693,652	7,730,188	7,766,725	7,803,262	7,839,798	7,876,335
11	North Carolina	5,881,766	5,956,453	6,031,140	6,105,827	6,180,514	6,255,202	6,329,889	6,404,576	6,479,263	6,553,950	6,628,637	6,703,324	6,778,011	6,852,698	6,927,385
12	Georgia	5,463,105	5,564,616	5,666,127	5,767,638	5,869,149	5,970,661	6,072,172	6,173,683	6,275,194	6,376,705	6,478,216	6,579,727	6,681,238	6,782,749	6,884,260
13	Virginia	5,346,818	5,430,872	5,514,926	5,598,980	5,683,034	5,767,088	5,851,142	5,935,196	6,019,250	6,103,304	6,187,358	6,271,412	6,355,466	6,439,520	6,523,574
14	Massachusetts	5,737,037	5,764,976	5,792,915	5,820,853	5,848,792	5,876,731	5,904,670	5,932,609	5,960,547	5,988,486	6,016,425	6,044,364	6,072,303	6,100,242	6,128,181
15	Indiana	5,490,224	5,495,618	5,501,011	5,506,405	5,511,798	5,517,192	5,522,585	5,527,979	5,533,372	5,538,766	5,544,159	5,549,552	5,554,946	5,560,339	5,565,733
16	Washington	4,132,156	4,205,610	4,279,063	4,352,517	4,425,970	4,499,424	4,572,878	4,646,331	4,719,785	4,793,238	4,866,692	4,940,145	5,013,598	5,087,051	5,160,504
17	Missouri	4,916,686	4,936,725	4,956,763	4,976,802	4,996,841	5,016,880	5,036,918	5,056,957	5,076,996	5,097,034	5,117,073	5,137,112	5,157,151	5,177,190	5,197,229
18	Wisconsin	4,705,767	4,724,367	4,742,967	4,761,568	4,780,168	4,798,768	4,817,368	4,835,968	4,854,569	4,873,169	4,891,769	4,910,369	4,928,969	4,947,569	4,966,169
19	Tennessee	4,591,120	4,619,727	4,648,333	4,676,940	4,705,546	4,734,153	4,762,759	4,791,366	4,819,972	4,848,579	4,877,185	4,905,792	4,934,399	4,962,999	4,991,606
20	Maryland	4,216,975	4,273,424	4,329,874	4,386,323	4,442,772	4,499,222	4,555,671	4,612,120	4,668,569	4,725,019	4,781,468	4,837,917	4,894,366	4,950,815	4,999,264
21	Arizona	2,718,215	2,812,916	2,907,618	3,002,319	3,097,020	3,191,722	3,286,423	3,381,124	3,475,825	3,570,527	3,665,228	3,760,000	3,854,771	3,949,543	4,044,314
22	Minnesota	4,075,970	4,105,883	4,135,796	4,165,709	4,195,622	4,225,535	4,255,447	4,285,360	4,315,273	4,345,186	4,375,099	4,405,012	4,434,925	4,464,838	4,494,751
23	Louisiana	4,205,900	4,207,307	4,208,715	4,210,122	4,211,529	4,212,937	4,214,344	4,215,751	4,217,158	4,218,566	4,219,973	4,221,380	4,222,787	4,224,194	4,225,601
24	Alabama	3,893,888	3,908,558	3,923,228	3,937,898	3,952,568	3,967,238	3,981,908	3,996,577	4,011,247	4,025,917	4,040,587	4,055,257	4,069,927	4,084,597	4,099,267
25	Colorado	2,889,964	2,930,407	2,970,850	3,011,293	3,051,736	3,092,179	3,132,622	3,173,065	3,213,508	3,253,951	3,294,394	3,334,837	3,375,280	3,415,723	3,456,166
26	Kentucky	3,660,777	3,663,229	3,665,681	3,668,133	3,670,585	3,673,037	3,675,488	3,677,940	3,680,392	3,682,844	3,685,296	3,687,748	3,690,200	3,692,652	3,695,104
27	South Carolina	3,121,820	3,158,308	3,194,797	3,231,285	3,267,773	3,304,262	3,340,750	3,377,238	3,413,726	3,450,215	3,486,703	3,523,191	3,559,679	3,596,167	3,632,655
28	Connecticut	3,107,576	3,125,530	3,143,484	3,161,438	3,179,392	3,197,346	3,215,300	3,233,254	3,251,208	3,269,162	3,287,116	3,305,070	3,323,024	3,340,978	3,358,932
29	Oklahoma	3,025,290	3,037,320	3,049,349	3,061,379	3,073,408	3,085,438	3,097,467	3,109,497	3,121,526	3,133,556	3,145,585	3,157,615	3,169,644	3,181,674	3,193,703
30	Oregon	2,633,105	2,654,027	2,674,948	2,695,870	2,716,791	2,737,713	2,758,635	2,779,556	2,800,478	2,821,399	2,842,321	2,863,243	2,884,164	2,905,086	2,926,007
31	Iowa	2,913,808	2,900,103	2,886,397	2,872,692	2,858,987	2,845,282	2,831,576	2,817,871	2,804,166	2,790,460	2,776,755	2,763,050	2,749,345	2,735,640	2,721,935
32	Mississippi	2,520,638	2,525,896	2,531,154	2,536,411	2,541,669	2,546,927	2,552,185	2,557,443	2,562,700	2,567,958	2,573,216	2,578,474	2,583,732	2,588,990	2,594,248
33	Kansas	2,363,679	2,375,069	2,386,459	2,397,848	2,409,237	2,420,627	2,432,016	2,443,406	2,454,795	2,466,185	2,477,574	2,488,964	2,500,353	2,511,743	2,523,132
34	Arkansas	2,286,435	2,292,864	2,299,293	2,305,722	2,312,151	2,318,580	2,325,009	2,331,438	2,337,867	2,344,296	2,350,725	2,357,154	2,363,583	2,369,999	2,376,428
35	Utah	1,461,037	1,487,218	1,513,400	1,539,581	1,565,762	1,591,944	1,618,125	1,644,306	1,670,487	1,696,669	1,722,850	1,749,031	1,775,212	1,801,393	1,827,574
36	Nevada	800,493	840,627	880,761	920,895	961,029	1,001,163	1,041,297	1,081,431	1,121,565	1,161,699	1,201,833	1,241,967	1,282,101	1,322,235	1,362,369
37	New Mexico	1,302,894	1,324,112	1,345,329	1,366,547	1,387,764	1,408,982	1,430,199	1,451,417	1,472,634	1,493,852	1,515,069	1,536,287	1,557,504	1,578,721	1,599,938

Crime by Population

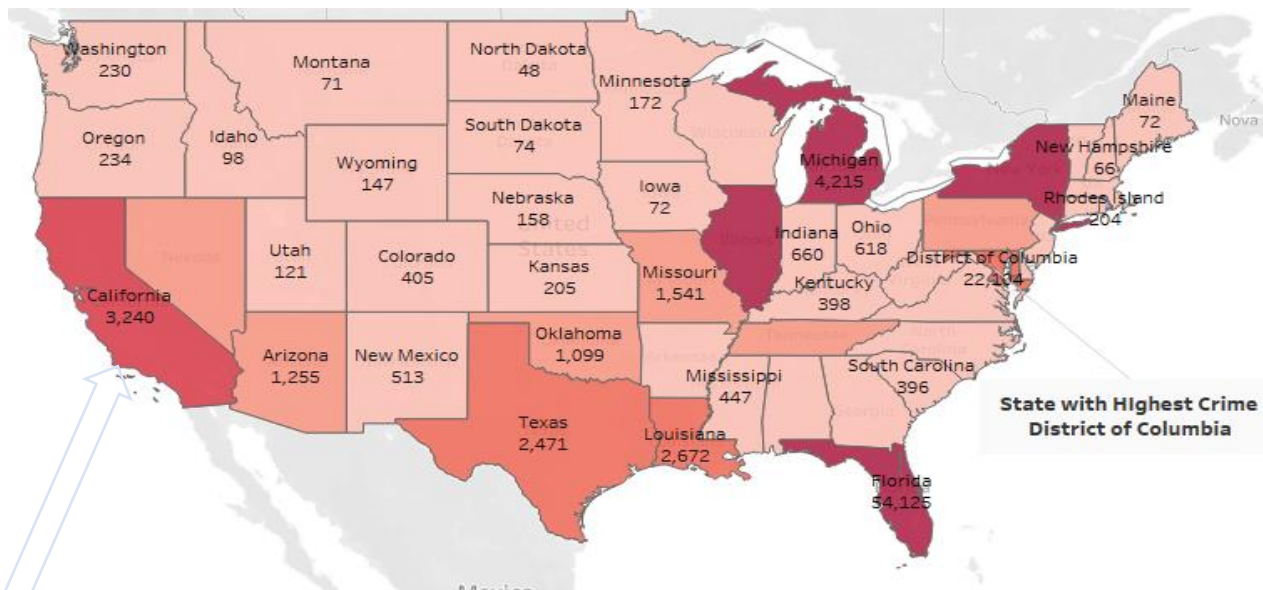
Sheet1 (Project 3 - US Crime Dataset)



`ROUND((SUM([Incident]) / AVG([Population]))*100000)`

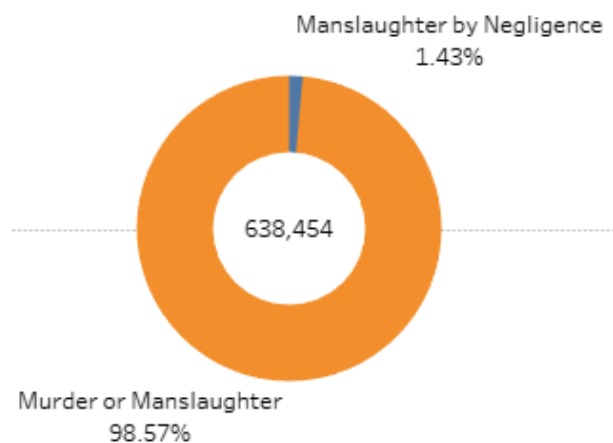
I leaned the application of this formula from (BI the Trend)



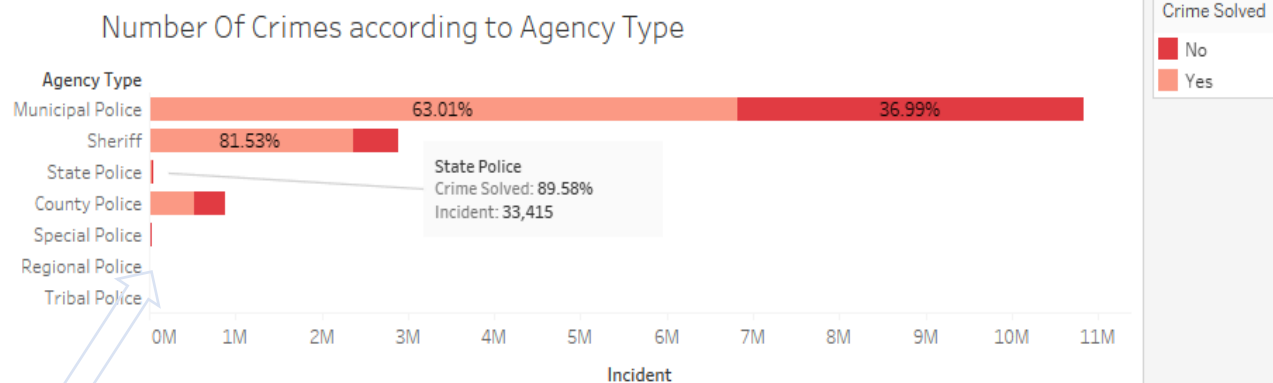


With the new information. The results Drastically changed, now showing the true picture of "District of Columbia" having the highest ratio of crime and "California" is effectively having much lesser crime rate.

- Donut Chart/ Pie chart are good representative when data is divided into 7 or less categories. I learned to create a Donut chart from a video by (Matthew, 2021). I the most frequent types of crimes, as there were just categories, it represented in this way. This helped me analyze the type of crime.



- For measuring the performance of the Agency types, I classified total incidents by agency type and made a stacked bar chart.



Municipal records highest number of crimes but the crime solving ratio of State Police is best with almost 90%.

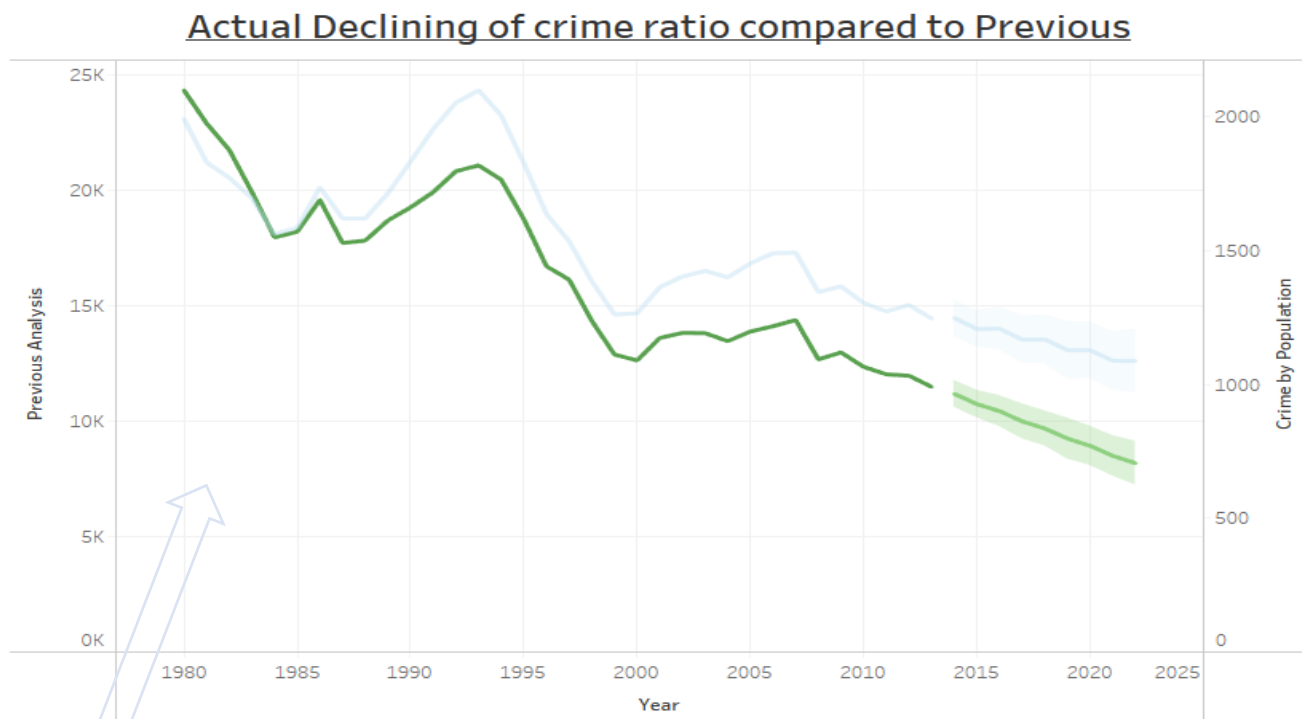
- Actual analysis of the Declining ratio of crime changed because of comparison with population was introduced. Here I wanted to forecast the prediction for next 8 years. I used the feature, but it was unclear to me what is the logic behind it. My doubt was solved by the module leader on 6-April-2022, who explained me that the trend shows the upward or downward motion in chart and season shows the ups and downs in the chart, these trends can be used according to the situations. Haixia Cleared my doubt on prediction interval, which means the possible range where the prediction might alter. These lessons are very crucial for my future journey in data science where I might need to make a lot of predictions.

Forecast Model

Custom

Trend: Additive
Season: Multiplicative

☒ Show prediction intervals 50%



The Green line shows the true effect on change in crime which is showing that crime has reduced just by 40% over the period of 34 years, but that is just the number of crimes. When we take the population growth into account, the actual reduction in the crime is around 55%. Which is really a massive change. I did some literature review on the reason behind a steep drop in Murder and Manslaughter crimes after mid 90's and found some possible reasons by (Cohn et al., 2013).

- 1) Because of industrial revolution a new vibrant economy emerged, due to which even low-skilled were able to get jobs.
- 2) It's a fact from our analysis that people in age group of 15-25 years are more active in crime. According to one hypothesis, abortion was legalized after 1973 which resulted in fewer unwanted births and unwanted child is more likely to involve into criminal activities. The effect of this measure resulted after mid 90's.
- 3) Changing demographic - The shrinking population of adolescent age after mid 90's all around the western world brought a significant reduction in all sorts of crimes including murder and manslaughter.

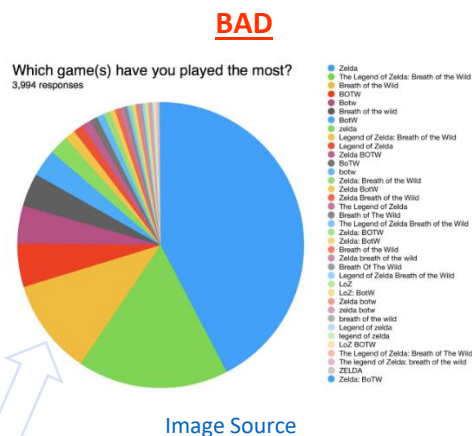
I learned some very good methods to practice data analysis with the help of visualization, I cultivated the habit of asking more Why's? of anything that data was showing on the charts, and to find the answer behind it I was keen to do some research and review some literatures. I wish to stick to this way of working as this could benefit me in my data science journey. This lesson I learned from (Dan, 2022) - Senior BI Analyst Performance at UWE, who gave a wonderful presentation on 9-Mar-2022 and motivated me with his visualization on how explained about what challenges are involved with implementing BI. Some tips given by him like effective communication within the team and data collecting person, can actually prove to be beneficial in getting a good quality data while analysis.

## Making Charts Beautiful and Meaningful:

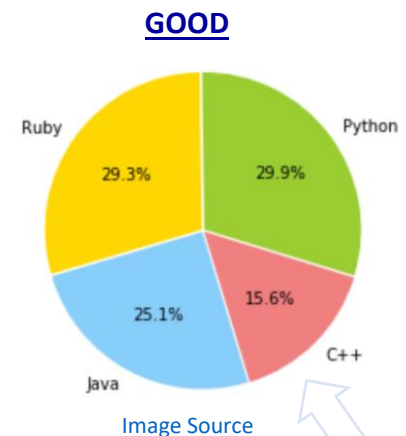
Understanding of many different types of charts for different purposes to make visualization look interesting from [Datavizproject.com](https://datavizproject.com) and connecting all the learnings from sessions I researched more on how to improvise charts.

Iconic memory or Visual sensor memory plays a crucial role for creating graphs, charts, or presentation - I learnt this in week 5 session about "Draw your audience attention" on 2-March (Haixia, 2022). As iconic memory is shorter than a short memory, it tends to remember 3-4 details at the maximum for an average person proven by an exercise by (Brian). It was interesting to learn about because in a truly short time we need to present the maximum and most important information to stakeholders. For doing so, I learnt that pre-attentive attribute such as text and color could be used accordingly. Also making sure to design the visualizations according to color blind people, where not to put red and green together was a particularly good tip. 25 tips to Instantly improve your Data Visualization Design by (Katy) was very informative to know some insights about good charts such as including a zero-base line, not to over explain, using single color, consistently ordering etc.

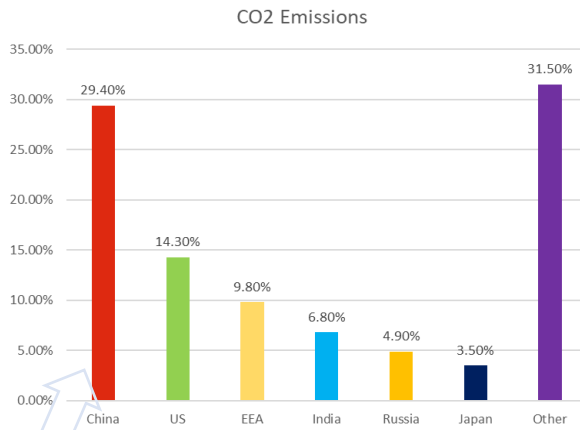
Some additional research on what good and bad charts look like.



There is a Wrong selection of type of chart. Pie charts are not good for more than 5-7 classes. Here it looks very messy, and the information is not clearly understandable. And 3D pie charts are not good.

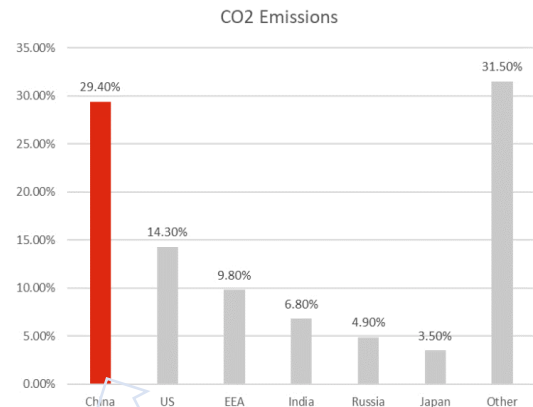


Here few details are mentioned and with percentages. A good color combination is used which makes chart look good and understandable



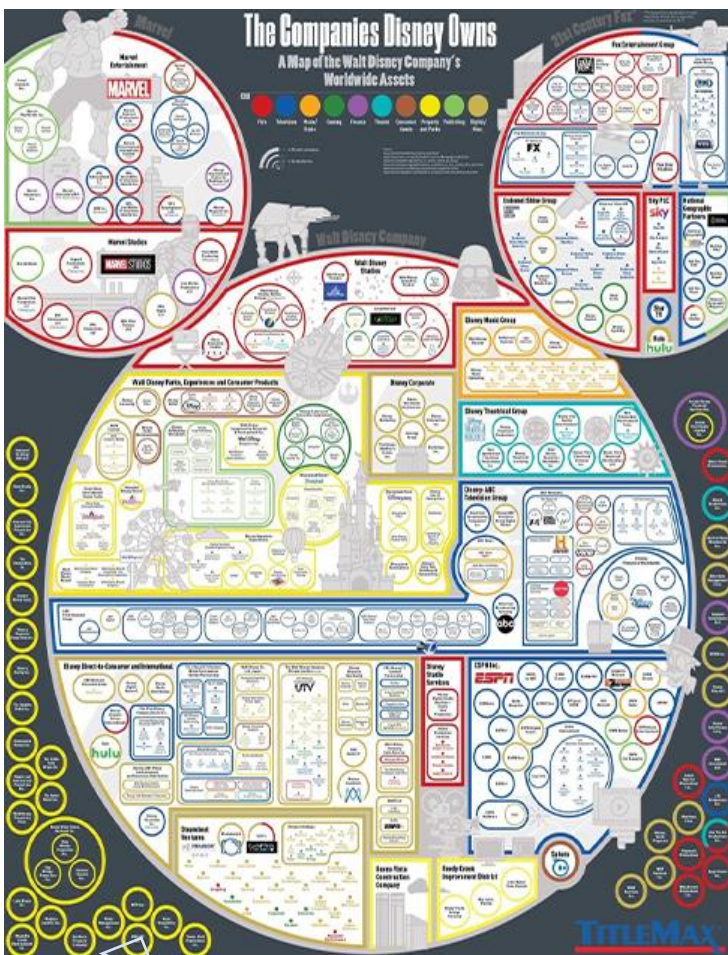
[Image Source](#)

Here the graph looks beautiful with a lot of good color, but it fails to deliver the message.



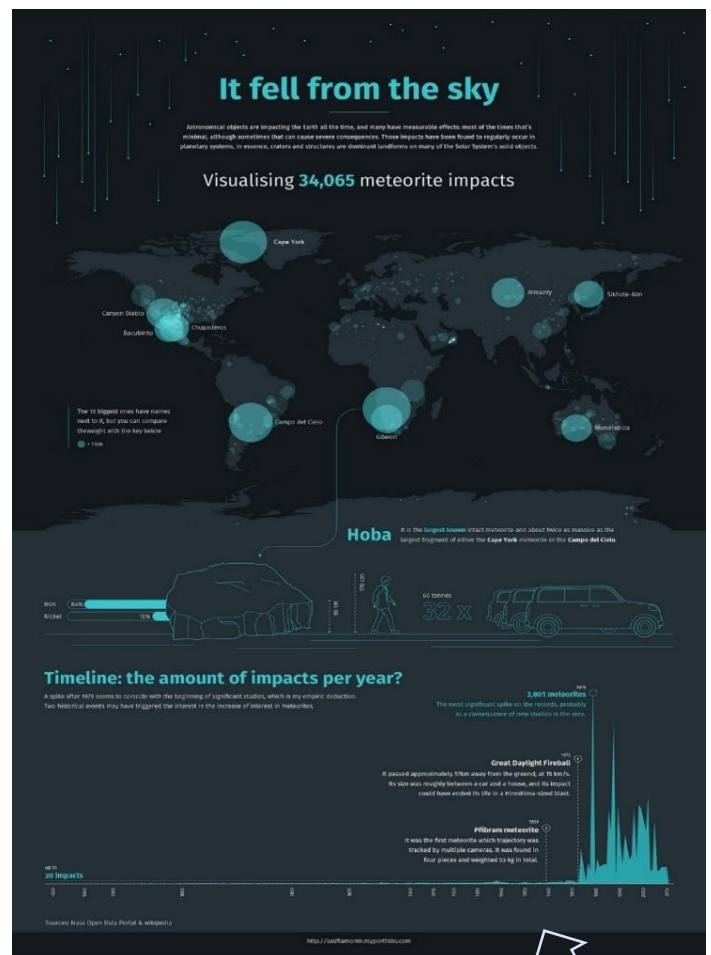
[Image Source](#)

It highlights accurate information showing that China is the country with highest Co2 Emissions.



[Image Source](#)

As Disney owns a lot of companies and assets, this chart is very messy and with too much information. A simple article or a table could have solved the purpose of representation.



[Image Source](#)

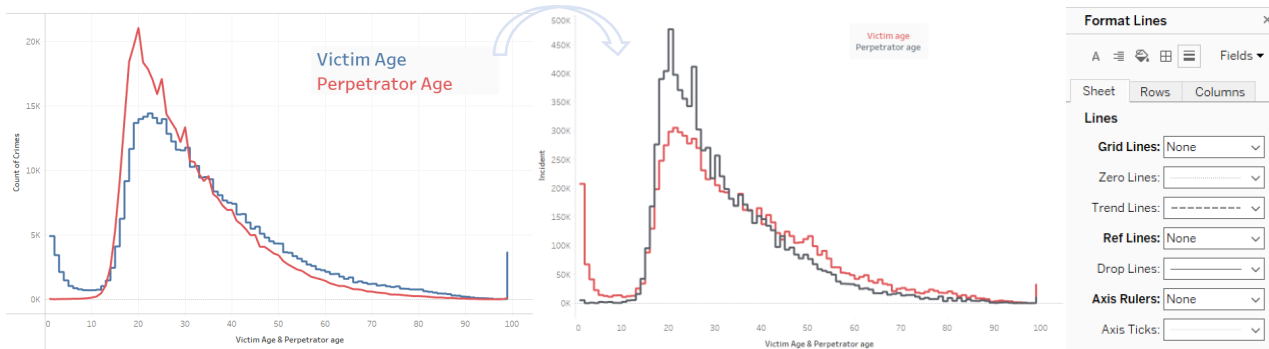
This has a very Stylish Graphics, All the information is clear and separated, it shows versatile information from timeline, biggest meteorite, and places where they struck. A good use of title and icons



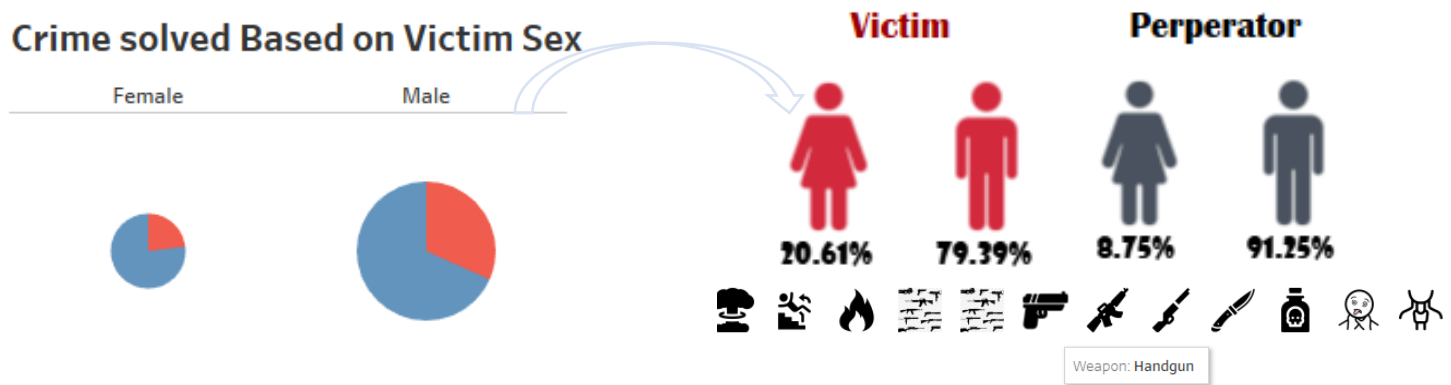
Moreover, further research on visualizations helped me uncover what are Good, Bad and Biased visualizations which includes reading the paper which explained the problems associated with 3-D charts, Wrong scaling which represents false information, and an innovative approach towards representation of data in a unique way with examples of superposition and explicit Encoding Diagrams by (Szafir, 2018)

### Some Improvements made on charts after learning.

- Making graphs free from unwanted lines like grid lines, dividers, etc. in the background.



- Using Icons wherever possible to have more interesting visualization.



I have applied the maximum of what I have learned in making my charts beautiful and informative. After spending ample time on analysis, we had another group meeting on 30-Mar-2022 where we discussed about improvising our work and creating final Dashboards and stories.

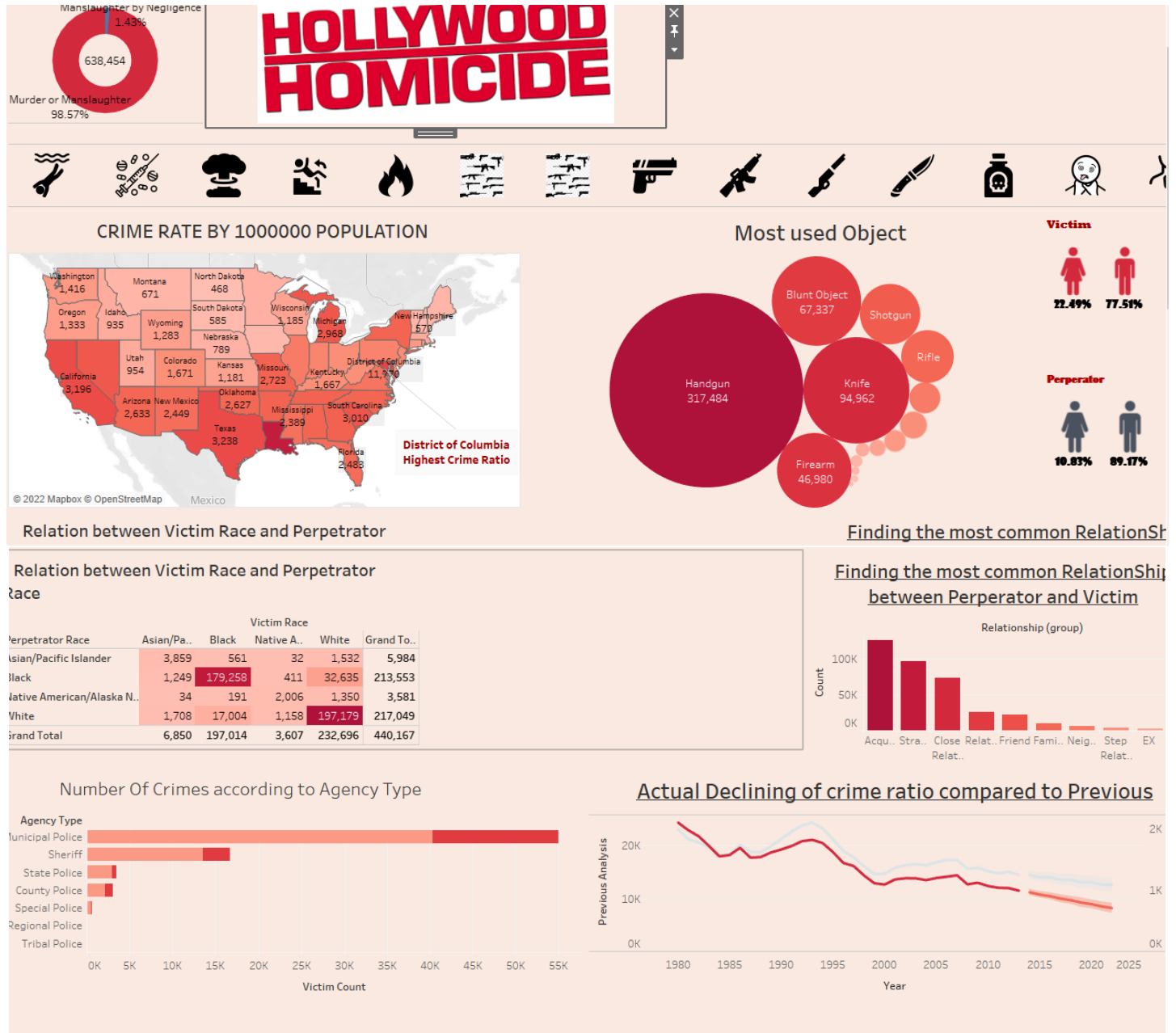
## Combining All together

After doing all the cooking It was time to put it on a plate to serve! I started with a video lesson by (sqlbelle, 2021) for creating a wonderful Dashboard. This helped me learn many new things like how to add blank spaces, using filters, publishing workbook, adding external icons and images, etc. I applied all my learnings into my work.

- I started with adding some vertical and horizontal objects in the dashboard. I learned from above video that it is a good practice to add a blank space on top and bottom, which gives convenience in creating dashboard.
- I then added the working sheets on those objects and started arranging them.
- I Downloaded some logo for weapon type and sex from [Flatcoin.com](https://flatcoin.com) for a better visual representation. Then followed the steps to save it and load it in tableau from above learning.
- According to the lessons taught in Week 8 on 23-Mar by (Haixia, 2022) about visualizing according to audience and always keep in consideration the presentation screen size, I designed according to my laptop size as we are supposed to record a video and as the data set is about Homicide, I chose to work on a red colored theme.

- As accordingly I changed the color and background of all visualizations. I then removed all the backgrounds and frames in all charts along with removing unwanted borders from format tab inspired by (DataCated).
- Considering the Affordances and its 3 categories – Perceptible, hidden, and false explained by tutor in week 8, I designed my dashboard accordingly.

After few steps of working this is how my initial dashboard looked.



We had our next group meeting on 7-April-2022 for sharing our individual work and getting feedbacks. There were some good points and bad points in each one's work. This was a good team learning experience as I got to know the strength and weaknesses of our team and myself. I got some feedback like my dashboard was visually genuinely nice and the color saturation was fantastic, also the idea of connecting population to derive true result was amazing, however the dashboard was little crowdy. We agreed to make some final dashboard in the Easter holidays including the best features of each and everyone's work and come up with a good story with it.

Some of the suggestions from tutor helped me improvise on my dashboard.

- For showing the truthfulness of the data I worked on tutors' recommendation and added all unknown data in percentage on my dashboard.

#### Re: Some Doubts in Project - 3 US Crime Data Set



Haixia Liu <Haixia.Liu@uwe.ac.uk>

4/19/2022 2:21 AM

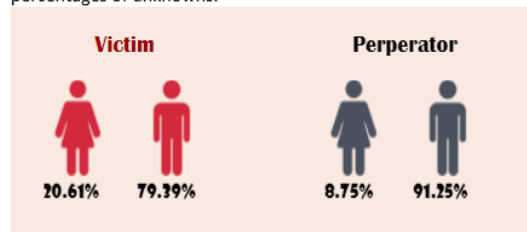


To: Parashkumar Shah (Student)

Well-done for your dashboard! Take my comments critically. You do not have to agree with me. But hopefully, my comments can offer you some topics to discuss in your diary.

- Truthful:

How did you deal with the unknown gender? I saw some crime report that put footnotes about the percentages of unknowns.

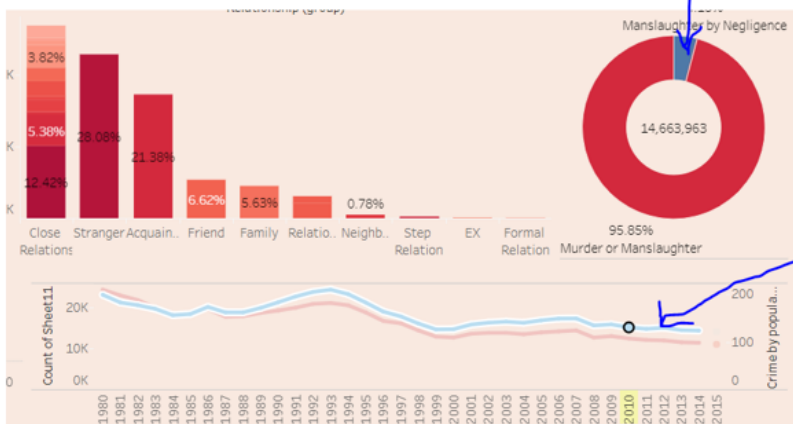


Unknown Data			
Perpetrator Sex	33.92%	Relation	72.34%
Perpetrator Race	34.77%	Weapon	5.899%

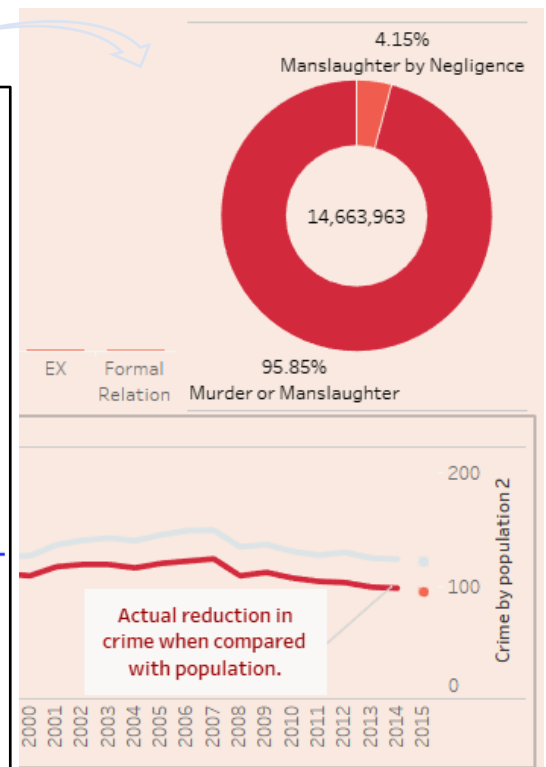
- Another good suggestion by tutor on choosing color so that the audience might not get wrong or misguided information. And to add a note on time series for the comparison.

- Accessible:

1. Color were chosen carefully. Good job. A little comment: the blue1 and blue2 could make some audience think that they represent the same dimension but it's not what you meant.

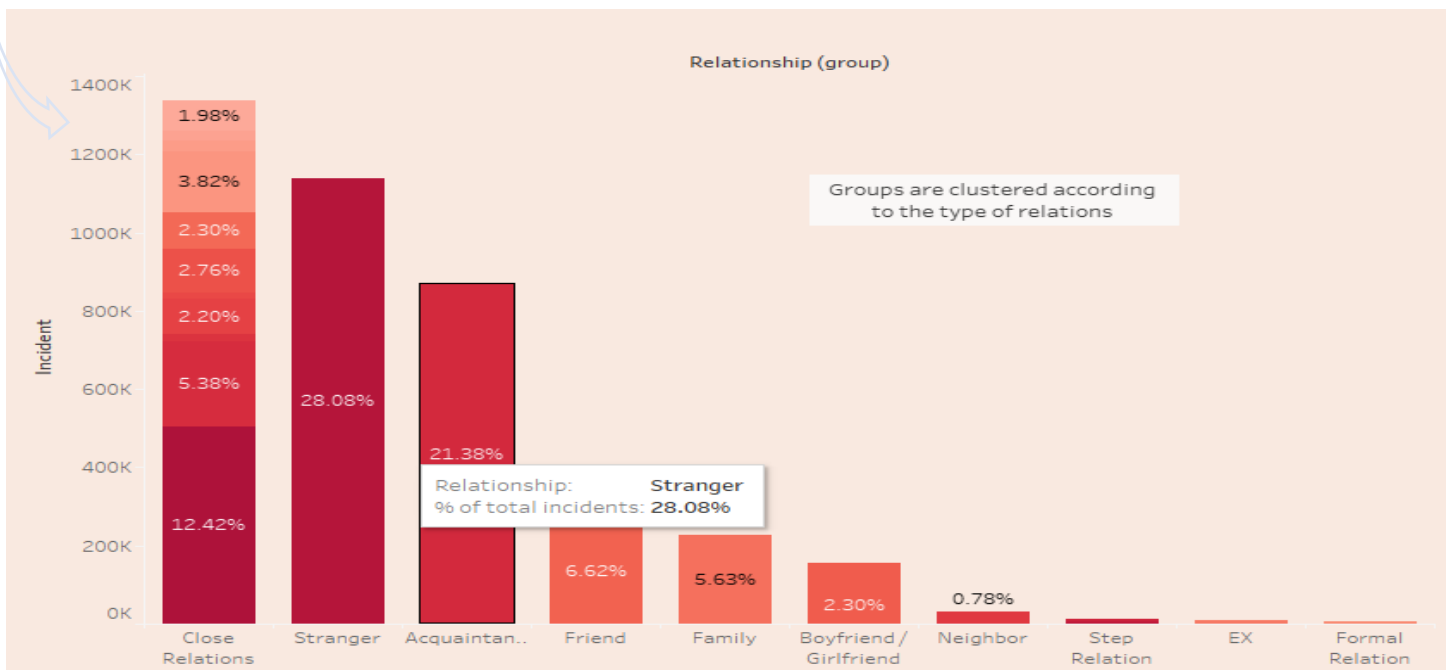


2. Look at the above lines: Better label these 2 lines (one is considering population?) directly in the end of the lines. The 2 y-axis should be avoided because they have different scales.



- A Very useful suggestion to chane Toolpits to make the information more redeable and also to add some other details for making graphs more informative.

Relationship chart: on clicking part of the bar, audience may would like to see other infor changing accordingly to discover insights. Also, please document your logic of clustering/grouping the relations.



## Story Telling and presentation

Week 9 and 10 sessions on 30-March and 6-April by (Haixia, 2022) were informative on knowing that while presentation one needs to be true and to tell the story from within heart to connect audience with emotions. Further research on public speaking made me learn not to ramble, maintain good eye contact with audience and keeping slides simple by (Alex, 2018).

Amongst the two ways of delivering presentation by (Paul, 2018) I discovered my interest and ability to execute method-2 on this data set where some important points such as engaging into story and findings extracted from analysis, Giving hints to audience and allow them to make guesses, and the flow of story must be kept as Story > Conclusion > Recommendations from method-2 influenced more to me when compared with method-1, which is a different approach where all figures and facts are first told to audience and giving a pause to understand, analyze and solve the problem.



For a good story telling it is very important to find reason behind it recommended by (Azhar, 2020) in 2<sup>nd</sup> advisor meeting on 21 March, which is similar to the Google concept of asking 5 why's to find the root cause of the problem. With all these learnings, I started narrating my story around the findings from the data set.

Tutor recommendation on dashboard being crowded helped me improvise on my story telling by creating the story points of my dashboard which further helped me to add few charts and I found it more convenient for presentation for my future Stories.

## Re: Some Doubts in Project - 3 US Crime Data Set



Haixia Liu <Haixia.Liu@uwe.ac.uk>

4/21/2022 10:40 AM




To: Parashkumar Shah (Student)

if you feel your dashboard is too busy, you can 'cluster' your visuals into storypoint?  
I believe you know how to create story:

[Introduction](#) [Homicide](#) [Homicide History](#) [Weapon Stats](#) [Perpetrators and Victim Stats](#) [Police Performance](#) [Re...](#)

Business Intelligence and Data Visualization  
UFCFKR-15-M

 University of the West of England

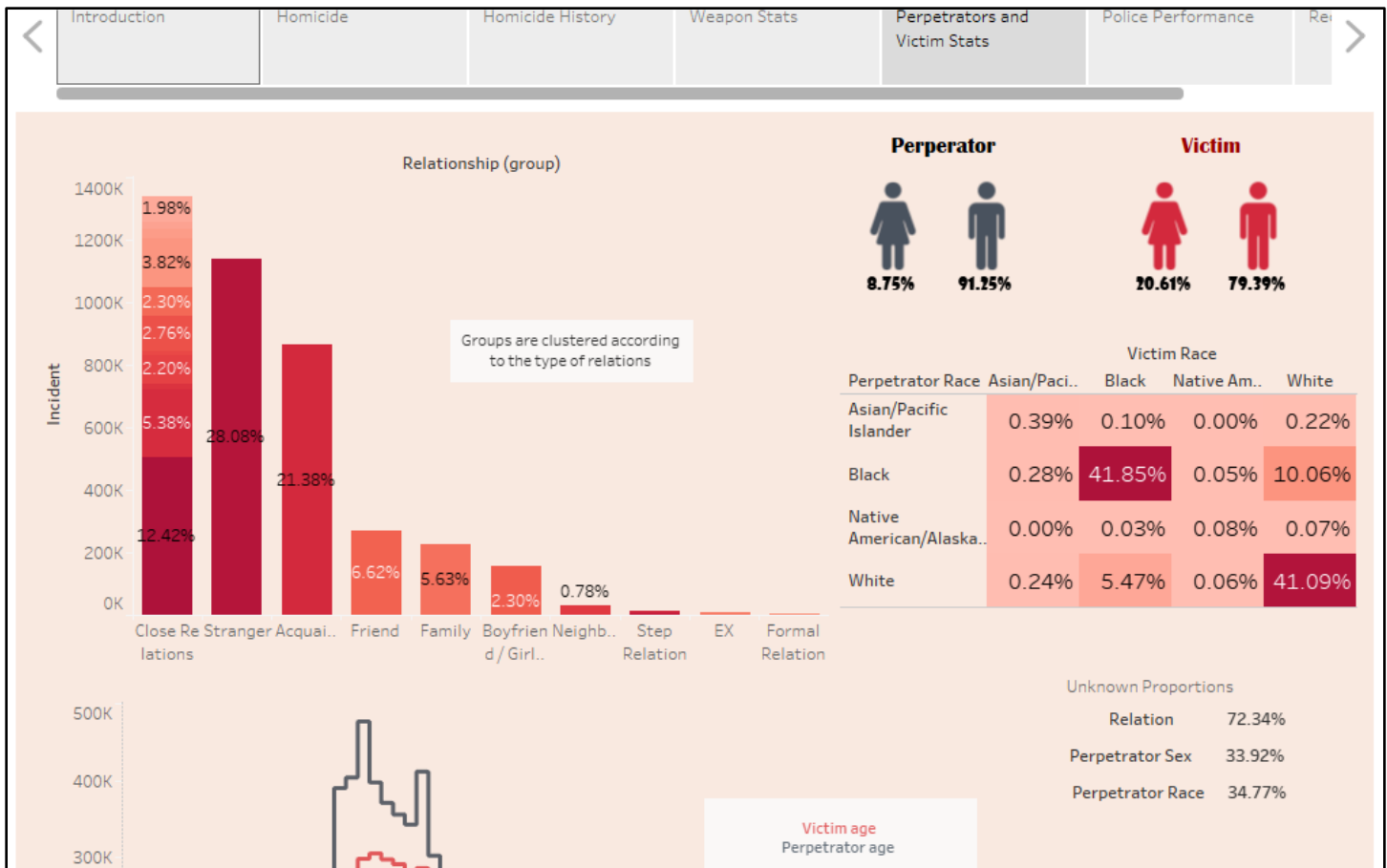
# Hollywood Homicide

## Based on: Project 3 (US Crime)

MAY 2022

BY.  
GROUP: 7  
OKEY  
TEMI  
SEE  
PARASHKUMAR SHAH (21051422)





The final Story prepared by me could be found [here](#)

## Recommendations

After making in depth analysis, I found some useful insights on basis of which I can recommend some solutions to the stake holder.

- 1) In almost 50% of the known weapons, "Handgun" has been used to execute crime. Thus, more preventive measure like strict licensing should be taken to make it less accessible for the perpetrators.
- 2) The highest proportion amongst all perpetrators, Adults between 17-25 years are very actively. This is an age of high adrenaline, which might affect their behavior making them more aggressive. Through planning should be done by Govt agencies to provide them appropriate guidance.
- 3) There are still 33% unsolved crimes amongst 34 years of data, which means that those perpetrators are still free and can harm other people. As 28% of total victim-perpetrator relations are strangers. More CCTV cameras around the cities should be planted to have better data for solving crimes.

## Conclusion

The Whole Project and the way of working it proved to be a very learning experience. From installing the tableau to making story points and presenting the data, I learned dozens of new skills. The feedbacks and solutions provided by tutor and module acted as catalyst in the process of learning. However, I felt that Advisor sessions could have been more interactive and frequent. Overall, I loved the deliveries and it motivated me to learn more about BI and Data Visualization in more details.

## References

1. Frye, C. (2021). Tableau Essential Training. *LinkedIn Learning* [Video Series]. 06 August. Available From: <https://www.linkedin.com/learning/tableau-essential-training-13964003/download-tableau?autoplay=true&resume=false&u=56744785> [Accessed February 2022]
2. Zane, M. (2021). Most important presentation skills. *Zippia* [Online]. 25 January. Available From: <https://www.zippia.com/advice/presentation-skills/> [Accessed 6 February 2022]
3. Matthews, P. (2022). Visual Perception, Cognition and Design. *Business Intelligence and Data Visualization* [Online]. 09 February. Available From: <https://uwe.cloud.panopto.eu/Panopto/Pages/Viewer.aspx?id=7bfadc60-4994-4dc8-ac94-ae3700de6f2a&instance=Blackboard> [Accessed 09 February 2022]
4. Sabina, A. (2016). Information Visualization. Slideplayer [online]. Available from: <https://slideplayer.com/slide/9281541/> [Accessed 13 February 2022]
5. Liu, H. (2022). Context and declutter. *Business Intelligence and Data Visualization*. 16 February. Available From: <https://uwe.cloud.panopto.eu/Panopto/Pages/Viewer.aspx?id=5bbf9e7c-e3f0-4f82-9165-ae3e00d83316&instance=Blackboard>
6. Kang, H et al. (2013). The prevention and handling of missing data. *PubMed Central*. Available from: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3668100/> [Accessed 20 February 2022]
7. Kennan, M. (2018). Proper Way to Label a Graph. *Science.com*. 25 April. Available from: <https://sciencing.com/proper-way-label-graph-5195234.html> [Accessed 25 February 2022]
8. Glen, S. Misleading Graphs in Real life. *Statisticshowto*. Available from: <https://sciencing.com/proper-way-label-graph-5195234.html> [Accessed 25 February 2022]
9. Anthony. (2010). Why you Should Never Use Pair Green and Red on the Web. *UXMovement*. 24 December. Available from: <https://uxmovement.com/content/why-you-should-never-pair-green-and-red-on-the-web/#:~:text=Color%20Blind%20Users%20See%20No%20Difference&text=The%20most%20common%20form%20of,color%20blind%20users%20can%20distinguish>. [Accessed 25 February 2022]
10. Sonal. (2021). Importance of Exploratory Data Analysis Before ML Modeling. *Eduonix*. 25 September. Available from: <https://blog.eduonix.com/bigdata-and-hadoop/importance-exploratory-data-analysis-ml-modelling/> [Accessed 1 March 2022]
11. Wikipedia. (2022). Time Series. Available from: [https://en.wikipedia.org/wiki/Time\\_series](https://en.wikipedia.org/wiki/Time_series) [Accessed 1 March 2022]
12. Edward, R. (2022). Our Ranking Explained. *Safewise*. 29 March. Available from: <https://www.safewise.com/blog/the-safewise-safest-cities-our-methodology/> [Accessed 4 March 2022]
13. Wikipeda. (2022). 2020 United States Census. Available from: [https://en.wikipedia.org/wiki/2020\\_United\\_States\\_census#:~:text=By%20July%2014%2C%202020%2C%20the,ate%20over%20a%2099.9%25%20rate](https://en.wikipedia.org/wiki/2020_United_States_census#:~:text=By%20July%2014%2C%202020%2C%20the,ate%20over%20a%2099.9%25%20rate) [Accessed 4 March 2022]
14. Anon. (no date). Tableau – Table Calculations How to divide two different columns using aggregate. Youtube [Video]. Available from: <https://www.youtube.com/watch?v=E11H40n05Sg&t=567s> [Accessed 5 March 2022]
15. Heffron, M. (2021) Tableau Tutorial – Donut Charts. Youtube [Video]. Available from: <https://www.youtube.com/watch?v=dLhojoAuiEI> [Accessed 5 March 2022]
16. Cohn, D et al. (2013). Gun Homicide Rate Down 49% Since 1993 Peak; Public Unaware. *Pew Research Centre*. 7 May. Available from: <https://www.pewresearch.org/social-trends/2013/05/07/gun-homicide-rate-down-49-since-1993-peak-public-unaware/> [Accessed 7 March 2022]
17. Bougourd, D. (2022). Business intelligence at UWE. *Business Intelligence and Data visualization*. March. Available from: <https://uwe.cloud.panopto.eu/Panopto/Pages/Viewer.aspx?id=24414926-ad1c-489b-af64-ae530100d137>
18. Gendron, B. (2019). Memory – Encoding & Iconic Memory. Youtube [Video]. Available from: <https://www.youtube.com/watch?v=OfUS4DmYbFc> [Accessed 2 March 2022]
19. Szafir, D. (2018). The Good The Bad and The Biased: Five way A Visualizations can Mislead. *Interactions.ACM.org*. July-August. Available at: <https://cmci.colorado.edu/visualab/papers/p26-szafir.pdf> [Accessed 15 March 2022]



20. Sqlbelle. (2021). Tableau Dashboard Tutorial – step by step, from start to finish – Part 2. *Youtube* [Video]. Available from: <https://www.youtube.com/watch?v=pVOJ1LM5cX0> [Accessed 1 April 2022]
21. Datacated. (2018). Tableau Background Formatting Tips. *Youtube* [Video]. Available from: <https://www.youtube.com/watch?v=rCgwB4XudC4&t=52s> [Accessed 10 April 2022]
22. Lyon, A. (2018). Public Speaking for Beginners. Communication coach. *Youtube* [Video]. Available from: <https://www.youtube.com/watch?v=i5mYphUoOCs> [Accessed 22 April 2022]
23. Smith, P. (2018). Telling Stories with Data – method 1 (The “How we got here” method). *Youtube* [Video]. Available from: <https://www.youtube.com/watch?v=g-rCSinAZf0&t=168s> [Accessed 30 March 2022]
24. Smith, P. (2018). Telling Stories with Data – method 2 (The Discover Journey Method). *Youtube* [Video]. Available from: <https://www.youtube.com/watch?v=YKilPOLG4CY&t=211s> [Accessed 30 March 2022]