Name: Mohsin Ali Irfan

Student ID: W1863288

Module: Data Visualization and Dashboarding.

1. **Research Questions:**
   a. **Which Boros has the highest amount of Mass Shootings and which Race was affected the most?**
   b. **Which age group of perpetrator and victims was majorly involved in the mass shooting incidents?**
   c. **Was there an increase or decrease in mass shootings incidents after and during covid?**
   d. **Which race of people was majorly involved and got effected by the mass shootings?**

2. **Data Acquisition:**
   The dataset was downloaded from the pubic repository of the datasets from the city of New York website.
   https://data.cityofnewyork.us/Public-Safety/NYPD-Shooting-Incident-Data-Historic-/833y-fsy8

   The dataset is compiled by the New York City Police Department (NYPD) and provides historical information on shooting incidents in New York City. It includes details like the date, time, location, precinct, borough, victim information, and other relevant details. Since the NYPD compiles the data, we can consider it to be somewhat reliable. However, it's important to be aware of potential limitations and biases. Not all shooting incidents in New York City may be included in the dataset because it relies on the information reported to and recorded by the NYPD. There might also be differences in data quality, reporting methods, or how incidents are categorized.
   When analyzing the data, it's crucial to recognize any potential gaps or biases and interpret the findings accordingly. It's recommended to supplement the dataset with other relevant sources or cross-reference the information with alternative datasets, if possible, to obtain a more comprehensive understanding of shooting incidents in New York City.

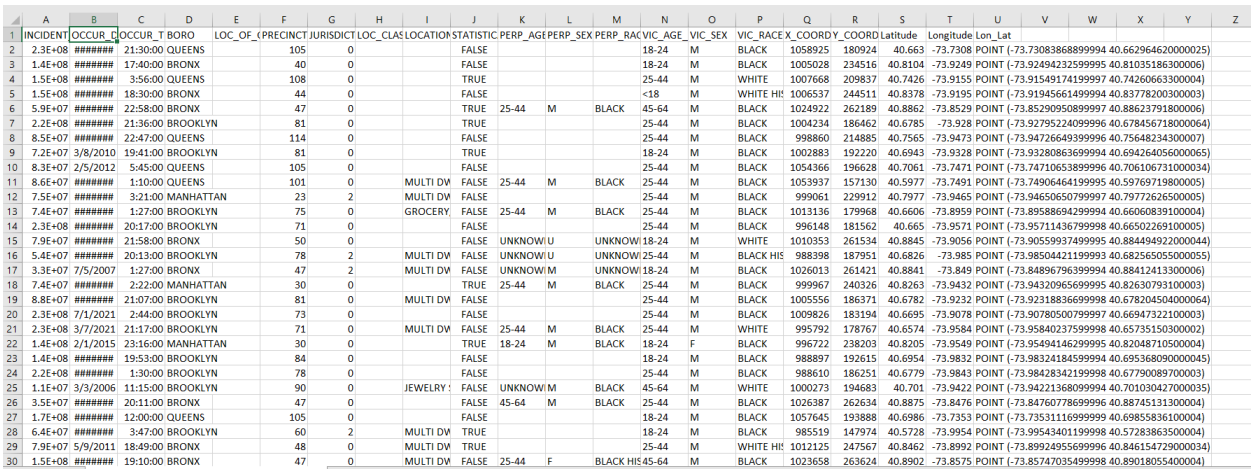
   2.1 Understanding the Dataset.

The Dataset contains the following datasets.
Incident Key: It is a text type column which got the incident keys which are randomly generated as per the description.
OCCUR_DATE: The date of incident reported.
OCCUR_TIME: The exact time of shooting.
BORO: The borough of New York where the incident happened.
PERCINT: Where the shooting incident occurred.
JURISDICTAION_CODE: The jurisdiction refers to the location where the shooting incident took place. In this dataset, jurisdiction codes 0 (Patrol), 1 (Transit), and 2 (Housing) represent the New York City Police Department (NYPD), while codes 3 and higher represent jurisdictions that are not part of the NYPD**.**
LOCATION_DESC: Location of the incident shooting
STATISTICAL_MURDER_FLAG: Whether the victim died in the shooting or not. If yes then it will be considered murder.
PERP_AGE_GROUP: Age group of the Perpetrator.
PERP_SEX: Perpetrator's gender.
PERP_RACE: To which race perpetrator belongs to.
VIC_AGE_GROUP: Victims age within a category.
VIC_SEX: Victims gender.
VIC_RACE: Victim's race
X_COORD_CD: Midblock X coordinate for New York State Plane coordinate system.
Y_COORD_CD: Midblock Y coordinate for New York State Plane coordinate system.
Latitude: The latitude coordinate for the global system.
Longitude: The longitude coordinate for the global system.
Lon_Lat: Longitude and Latitude coordinates for mapping.

3. **Data Preparation:**

Data preparation is one of the most important parts before final visualizations because the data contains a lot of missing values, values that needed to be changed, few features that are not going to help us in the analysis so lets just deal with the data before the final visualizations.

This is how our dataset looks initially.
As we can there are quite a lot of missing values that we need to remove. I removed those missing values using R.



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z
1	INCIDENT	OCCUR_D	OCCUR_T	BORO	LOC_OF_(	PRECINCT	JURISDICT	LOC_CLAS	LOCATION	STATISTIC	PERP_AGE	PERP_SEX	PERP_RAC	VIC_AGE_	VIC_SEX	VIC_RACE	X_COORD	Y_COORD	Latitude	Longitude	Lon_Lat					
2	2.3E+08	########	21:30:00	QUEENS		105	0			FALSE				18-24	M	BLACK	1058925	180924	40.663	-73.7308	POINT (-73.73083868899994 40.662964620000025)					
3	1.4E+08	########	17:40:00	BRONX		40	0			FALSE				18-24	M	BLACK	1005028	234516	40.8104	-73.9249	POINT (-73.92494232599995 40.81035186300006)					
4	1.5E+08	########	3:56:00	QUEENS		108	0			TRUE				25-44	M	WHITE	1007668	209837	40.7426	-73.9155	POINT (-73.91549174199997 40.74260663300004)					
5	1.5E+08	########	18:30:00	BRONX		44	0			FALSE				<18	M	WHITE HIS	1006537	244511	40.8378	-73.9195	POINT (-73.91945661499994 40.83778200300003)					
6	5.9E+07	########	22:58:00	BRONX		47	0			TRUE	25-44	M	BLACK	45-64	M	BLACK	1024922	262189	40.8862	-73.8529	POINT (-73.85290950899997 40.88623791800006)					
7	2.2E+08	########	21:36:00	BROOKLYN		81	0			TRUE				25-44	M	BLACK	1004234	186462	40.6785	-73.928	POINT (-73.92795224099996 40.678456718000064)					
8	8.5E+07	########	22:47:00	QUEENS		114	0			FALSE				25-44	M	BLACK	998860	214885	40.7565	-73.9473	POINT (-73.94726649399996 40.75648234300007)					
9	7.2E+07	3/8/2010	19:41:00	BROOKLYN		81	0			TRUE				18-24	M	BLACK	1002883	192220	40.6943	-73.9328	POINT (-73.93280863699994 40.694264056000065)					
10	8.3E+07	2/5/2012	5:45:00	QUEENS		105	0			FALSE				25-44	M	BLACK	1054366	196628	40.7061	-73.7471	POINT (-73.74710653899996 40.706106731000034)					
11	8.6E+07	########	1:10:00	QUEENS		101	0		MULTI DW	FALSE	25-44	M	BLACK	25-44	M	BLACK	1053937	157130	40.5977	-73.7491	POINT (-73.74906464199995 40.59769719800005)					
12	7.5E+07	########	3:21:00	MANHATTAN		23	2		MULTI DW	FALSE				25-44	M	BLACK	999061	229912	40.7977	-73.9465	POINT (-73.94650650799997 40.79772626500005)					
13	7.4E+07	########	1:27:00	BROOKLYN		75	0		GROCERY	FALSE	25-44	M	BLACK	25-44	M	BLACK	1013136	179968	40.6606	-73.8959	POINT (-73.89588694299994 40.66060839100004)					
14	2.3E+08	########	20:17:00	BROOKLYN		71	0			FALSE				25-44	M	BLACK	996148	181562	40.665	-73.9571	POINT (-73.95711436799998 40.66502269100005)					
15	7.9E+07	########	21:58:00	BRONX		50	0			FALSE	UNKNOWI	U	UNKNOWI	18-24	M	WHITE	1010353	261534	40.8845	-73.9056	POINT (-73.90559937499995 40.884494922000044)					
16	5.4E+07	########	20:13:00	BROOKLYN		78	2		MULTI DW	FALSE	UNKNOWI	U	UNKNOWI	25-44	M	BLACK HIS	988398	187951	40.6826	-73.985	POINT (-73.98504421199993 40.682565055000055)					
17	3.3E+07	7/5/2007	1:27:00	BRONX		47	2		MULTI DW	FALSE	UNKNOWI	M	UNKNOWI	18-24	M	BLACK	1026013	261421	40.8841	-73.849	POINT (-73.84896796399994 40.88412413300006)					
18	7.4E+07	########	2:22:00	MANHATTAN		30	0			TRUE	25-44	M	BLACK	25-44	M	BLACK	999967	240326	40.8263	-73.9432	POINT (-73.94320965699995 40.82630793100003)					
19	8.8E+07	########	21:07:00	BROOKLYN		81	0		MULTI DW	FALSE				25-44	M	BLACK	1005556	186371	40.6782	-73.9232	POINT (-73.92318836699998 40.678204504000064)					
20	2.3E+08	7/1/2021	2:44:00	BROOKLYN		73	0			FALSE				25-44	M	BLACK	1009826	183194	40.6695	-73.9078	POINT (-73.90780500799997 40.66947322100003)					
21	2.3E+08	3/7/2021	21:17:00	BROOKLYN		71	0		MULTI DW	FALSE	25-44	M	BLACK	25-44	M	WHITE	995792	178767	40.6574	-73.9584	POINT (-73.95840237599998 40.65735150300002)					
22	1.4E+08	2/1/2015	23:16:00	MANHATTAN		30	0			TRUE	18-24	M	BLACK	18-24	F	BLACK	996722	238203	40.8205	-73.9549	POINT (-73.95494146299995 40.82048710500004)					
23	1.4E+08	########	19:53:00	BROOKLYN		84	0			FALSE				18-24	M	BLACK	988897	192615	40.6954	-73.9832	POINT (-73.98324184599994 40.695368090000045)					
24	2.2E+08	########	1:30:00	BROOKLYN		78	0			FALSE				25-44	M	BLACK	988610	186251	40.6779	-73.9843	POINT (-73.98428342199998 40.67790089700003)					
25	1.1E+07	3/3/2006	11:15:00	BROOKLYN		90	0		JEWELRY	FALSE	UNKNOWI	M	BLACK	45-64	M	WHITE	1000273	194683	40.701	-73.9422	POINT (-73.94221368099994 40.701030427000035)					
26	3.5E+07	########	20:11:00	BRONX		47	0			FALSE	45-64	M	BLACK	25-44	M	BLACK	1026387	262634	40.8875	-73.8476	POINT (-73.84760778699996 40.88745131300004)					
27	1.7E+08	########	12:00:00	QUEENS		105	0			FALSE				18-24	M	BLACK	1057645	193888	40.6986	-73.7353	POINT (-73.73531116999999 40.69855836100004)					
28	6.4E+07	########	3:47:00	BROOKLYN		60	2		MULTI DW	TRUE				18-24	M	BLACK	985519	147974	40.5728	-73.9954	POINT (-73.99543401199998 40.57283863500004)					
29	7.9E+07	5/9/2011	18:49:00	BRONX		48	0		MULTI DW	TRUE				25-44	M	WHITE HIS	1012125	247567	40.8462	-73.8992	POINT (-73.89924955699996 40.846154729000034)					
30	1.5E+08	########	19:10:00	BRONX		47	0		MULTI DW	FALSE	25-44	F	BLACK HIS	45-64	M	BLACK	1023658	263624	40.8902	-73.8575	POINT (-73.85747035499998 40.89018055400004)					

To deal with the missing values, I first converted them into null values to be clearer about the dataset.

```
> df[df == ""] <- NA
> head(df)
  INCIDENT_KEY OCCUR_DATE OCCUR_TIME     BORO LOC_OF_OCCUR_DESC PRECINCT JURISDICTION_CODE
1    228798151  5/27/2021   21:30:00   QUEENS               <NA>      105                 0
2    137471050  6/27/2014   17:40:00    BRONX               <NA>       40                 0
3    147998800 11/21/2015    3:56:00   QUEENS               <NA>      108                 0
4    146837977  10/9/2015   18:30:00    BRONX               <NA>       44                 0
5     58921844  2/19/2009   22:58:00    BRONX               <NA>       47                 0
6    219559682 10/21/2020   21:36:00 BROOKLYN               <NA>       81                 0
  LOC_CLASSFCTN_DESC LOCATION_DESC STATISTICAL_MURDER_FLAG PERP_AGE_GROUP PERP_SEX PERP_RACE
1               <NA>          <NA>                  FALSE           <NA>      <NA>      <NA>
2               <NA>          <NA>                  FALSE           <NA>      <NA>      <NA>
3               <NA>          <NA>                   TRUE           <NA>      <NA>      <NA>
4               <NA>          <NA>                  FALSE           <NA>      <NA>      <NA>
5               <NA>          <NA>                   TRUE          25-44         M     BLACK
6               <NA>          <NA>                   TRUE           <NA>      <NA>      <NA>
  VIC_AGE_GROUP VIC_SEX       VIC_RACE X_COORD_CD Y_COORD_CD Latitude Longitude
1         18-24       M          BLACK    1058925   180924.0 40.66296 -73.73084
2         18-24       M          BLACK    1005028   234516.0 40.81035 -73.92494
3         25-44       M          WHITE    1007668   209836.5 40.74261 -73.91549
4           <18       M WHITE HISPANIC    1006537   244511.1 40.83778 -73.91946
5         45-64       M          BLACK    1024922   262189.4 40.88624 -73.85291
6         25-44       M          BLACK    1004234   186461.7 40.67846 -73.92795
                                  Lon_Lat
1 POINT (-73.73083868899994 40.662964620000025)
2  POINT (-73.92494232599995 40.81035186300006)
3  POINT (-73.91549174199997 40.74260663300004)
4  POINT (-73.91945661499994 40.83778200300003)
5  POINT (-73.85290950899997 40.88623791800006)
6 POINT (-73.92795224099996 40.678456718000064)
```

But first I dropped all the columns that I won't be using in the analysis. This is the final list of the columns that I will be using.

```
 [1] "OCCUR_DATE"              "OCCUR_TIME"              "BORO"
 [4] "PRECINCT"                "STATISTICAL_MURDER_FLAG" "PERP_AGE_GROUP"
 [7] "PERP_SEX"                "PERP_RACE"               "VIC_AGE_GROUP"
[10] "VIC_SEX"                 "VIC_RACE"                "Latitude"
[13] "Longitude"
```

This is how our dataset looks after removing the null values.

```
   OCCUR_DATE OCCUR_TIME      BORO PRECINCT STATISTICAL_MURDER_FLAG PERP_AGE_GROUP PERP_SEX
5   2/19/2009   22:58:00     BRONX       47                    TRUE          25-44        M
10  8/26/2012    1:10:00    QUEENS      101                   FALSE          25-44        M
12  8/29/2010    1:27:00  BROOKLYN       75                   FALSE          25-44        M
14  5/25/2011   21:58:00     BRONX       50                   FALSE        UNKNOWN        U
15 11/9/2008   20:13:00  BROOKLYN       78                   FALSE        UNKNOWN        U
16  7/5/2007    1:27:00     BRONX       47                   FALSE        UNKNOWN        M
   PERP_RACE VIC_AGE_GROUP VIC_SEX       VIC_RACE Latitude Longitude
5      BLACK         45-64       M          BLACK 40.88624 -73.85291
10     BLACK         25-44       M          BLACK 40.59770 -73.74906
12     BLACK         25-44       M          BLACK 40.66061 -73.89589
14   UNKNOWN         18-24       M          WHITE 40.88449 -73.90560
15   UNKNOWN         25-44       M BLACK HISPANIC 40.68257 -73.98504
16   UNKNOWN         18-24       M          BLACK 40.88412 -73.84897
```

I also realized few of our column has got outliers so let's see those outliers and remove them.

```
> print(unique(df$PERP_AGE_GROUP))
 [1] "25-44"   "UNKNOWN" "18-24"   "45-64"   "<18"     "65+"     "940"     "224"     "1020"
[10] "(null)"
```

The PER_AGE_GROUP contains ages that are not possible so lets just remove those rows. After removing, this is how our column looked alike.

```
> print(unique(tempdf$PERP_AGE_GROUP))
[1] "25-44"   "UNKNOWN" "18-24"   "45-64"   "<18"     "65+"
```

The Victims age also have some outliers, lets see those outliers

```
> print(unique(df$VIC_AGE_GROUP))
[1] "45-64"   "25-44"   "18-24"   "65+"     "<18"     "UNKNOWN" "1022"
```

An age of 1022 years is not possible so let's remove it.

On removing the outliers, we get this.

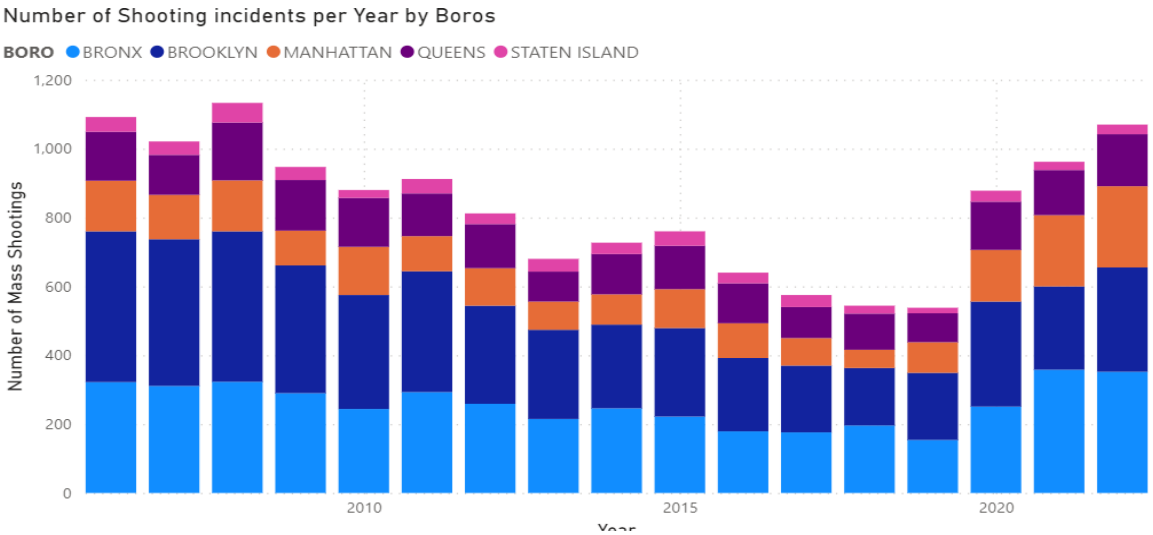
```
> print(unique(tempdf$VIC_AGE_GROUP))
[1] "45-64"   "25-44"   "18-24"   "65+"     "<18"     "UNKNOWN"
```

I have kept the value "UNKNOWN" for PERP_AGE_GROUP and also for the VIC_AGE_GROUP because unknown here simply means that the age is not available so I will deal with this value as some value and not a null value in our dataset and will use it for our visualizations.

4. Data Analysis:

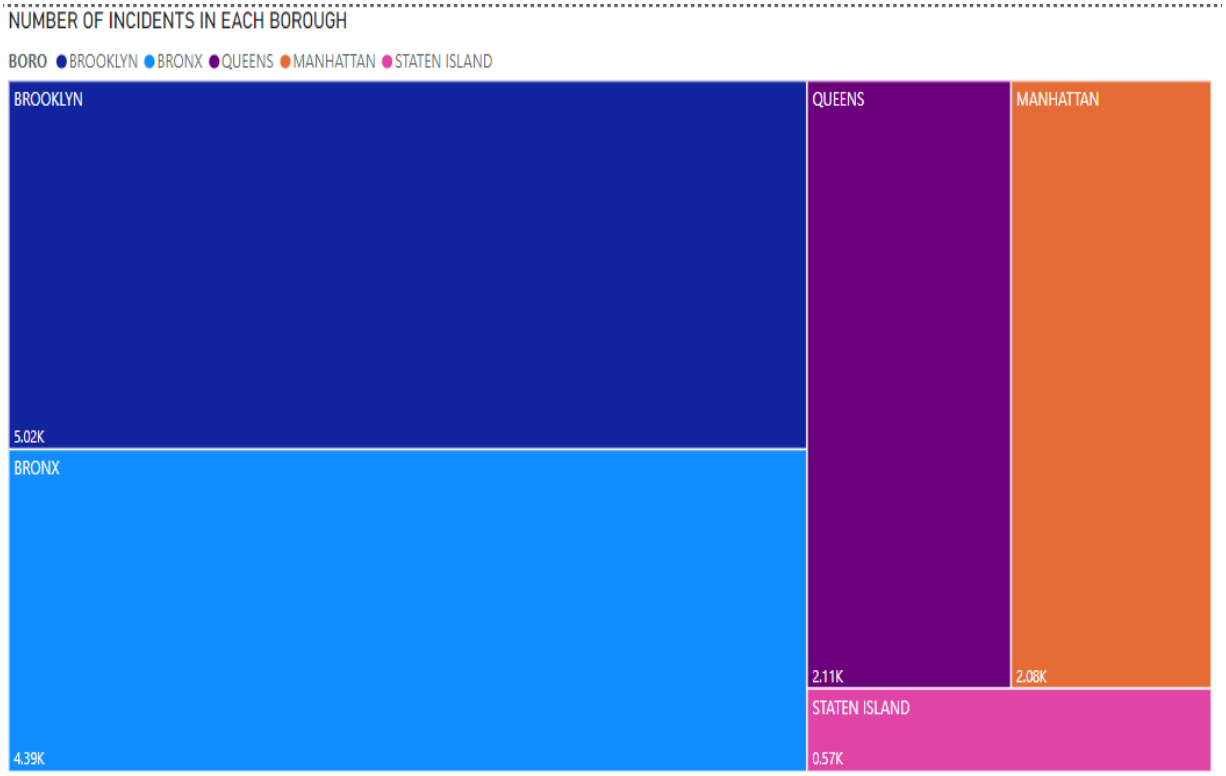
Across all the broughs, MANHATTAN had the most interesting recent trend and started trending up on 2018, rising by 345.28% (183) in 4 years. Count of VIC_AGE_GROUP for BORO BRONX was

trending down between 2006 and 2017 with a drop of 146 but had a significant change in trend and rose by 156 starting 2018.
From the visual it is also clear that there was a spike in total cases reported during and after covid.

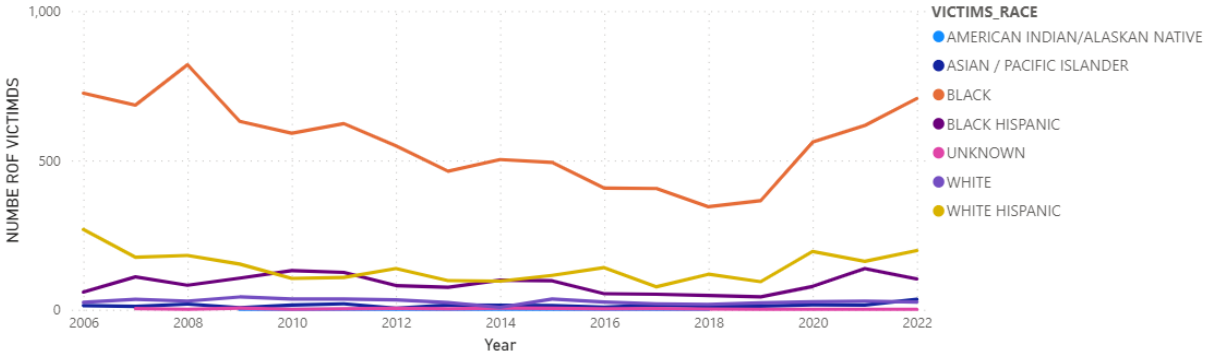


Let's have a better understanding of our data by looking at the total number of mass shooting cases reported in each Boroughs. For this purpose, I have used a Tree Map. From the visualization it is clear that Brooklyn has the highest number incident reported (5.02k) which is 34.14% of the total cases. Bronx being the borough with 2nd highest number of incidents (4.39k) which is 30.98% of the total cases. Then comes Queen, Manhattan, and State Island respectively. State Island being the most peaceful borough amongst all the New York boroughs as it has reported the least number of mass shooting incidents which are only 0.57k from 2006 to 2022.

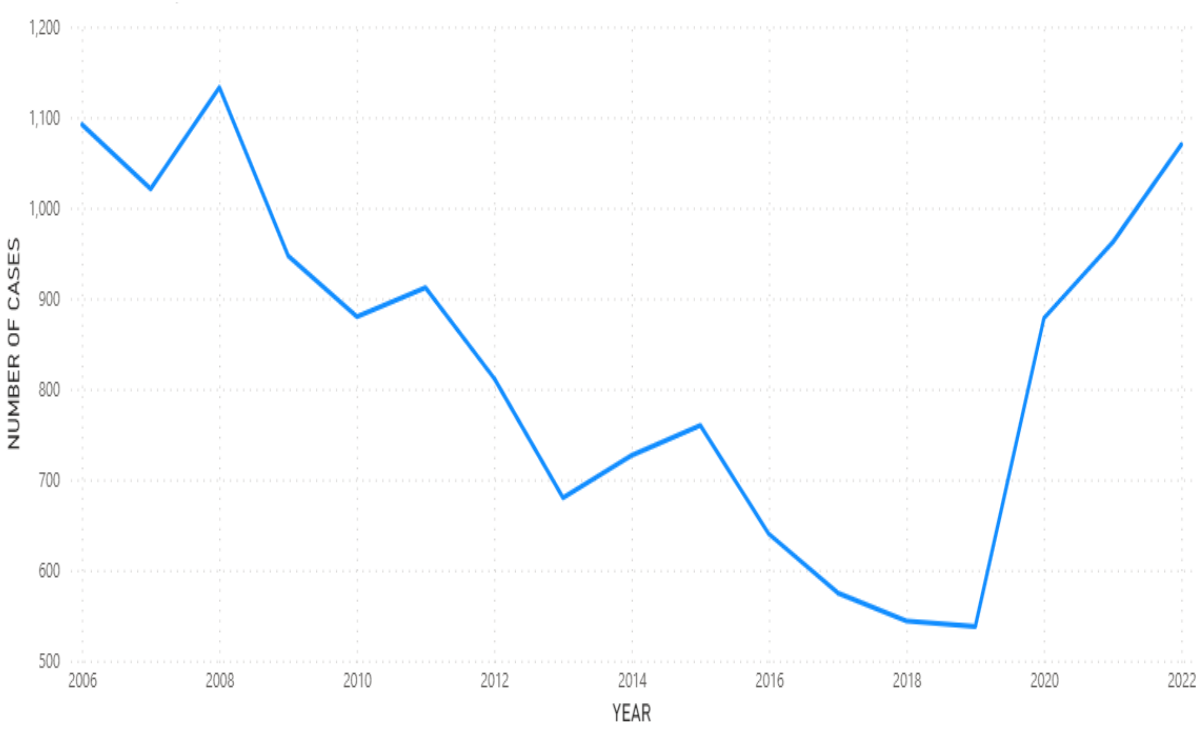
The most unsafe borough is Brooklyn with the greatest number of incident reported. From the visualization we can clearly see that the number of mass shooting incident were deceasing gradually after having a little spike in 2008. From 2013 to 2015 was again a period when there was again a little increase in the cases being reported but later after 2019 especially after covid there is a devastating increase in the cases.



From the following graph we can see how different races are being affected. Black people are the who get affected most by the mass shootings. If we have a close look at the line, we can clearly see the violence against the black people rose up after 2019 and in covid after gradually decreasing for previous years. BLACK HISPANIC started trending up in 2020, rising by 32.05% in 2 years.
The BLACK HISPANIC started trending up on 2020, rising by 32.05% (25) in 2 years. The number of BLACK HISPANIC who faced this incident were trending down between 2013 and 2019 with a drop of 32 but had a significant change in trend and rose by 25 starting 2020.
Count for ASIAN / PACIFIC ISLANDER started trending up on 2017, rising by 118.75% (19) in 5 years.
The Count for WHITE HISPANIC started trending up on 2017, rising by 157.14% (121) in 5 years.



From the following line graph we can have a clear understanding that after 2008 the mass shooting incidents started decreasing gradually but then the incidents had a spike 2013 to 2015 but after 2015 the number of cases started decreasing again. After 2019 and during covid, the cases had a great spike and till the data of 2022 the numbers were increasing rapidly.



Lets see the different races being attacked in different boroughs. From the following graph we can see that for each borough, the races that was the victim of mass shootings the most was Black. After Black the 2nd most affected race after the black were black Hispanic.

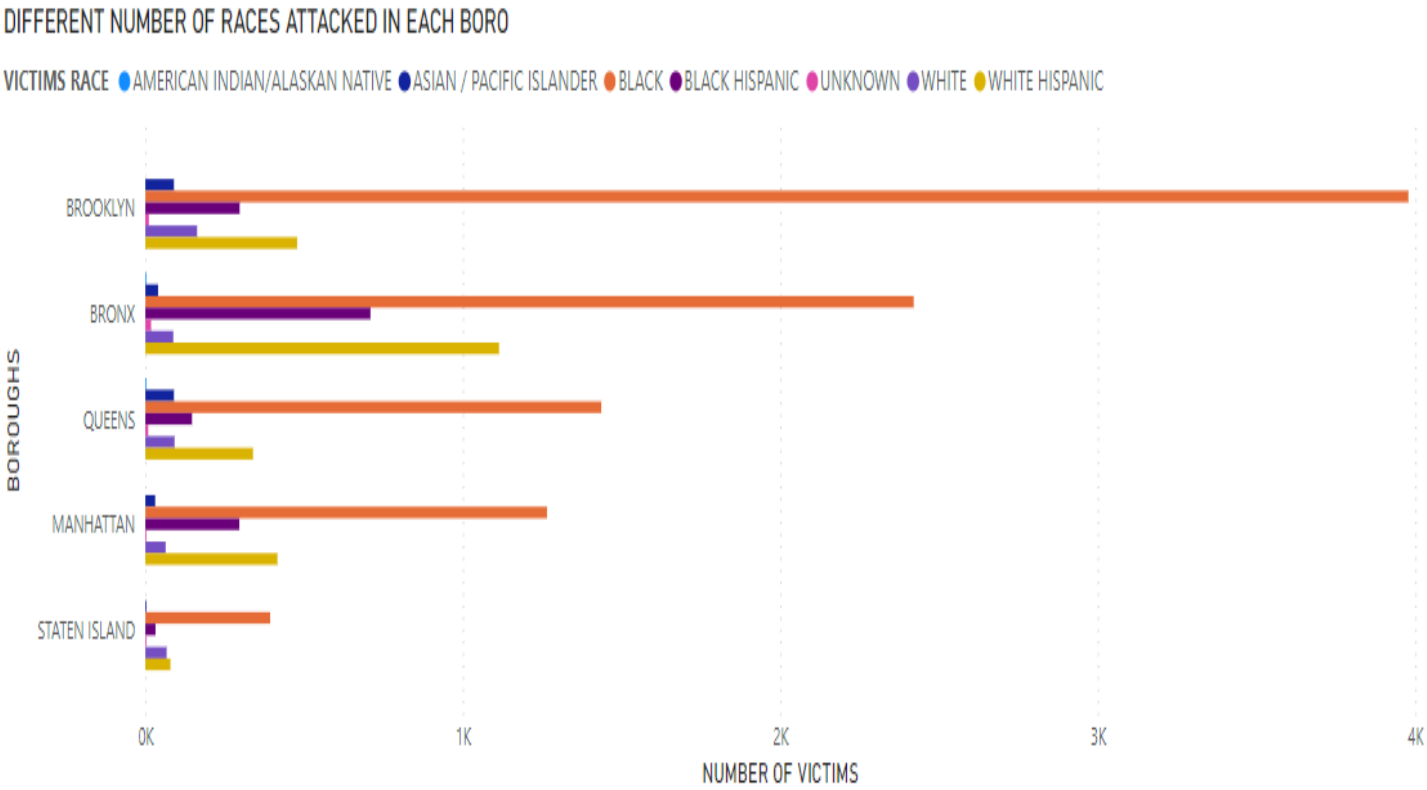
In Brooklyn the 3rd most affected race was white Hispanic and then white. Asian were considered to be the safest race in Brooklyn with least number of victims.

In Bronx the 2nd most affected race as White Hispanic and the 3rd was Black Hispanic people. 2nd safest people after Asians were white.

Queens has the same ratio of incident per race apart from Asian/Pacific Islanders and White as both has same number of incidents reported so both can be considered as the safest races in Queens.

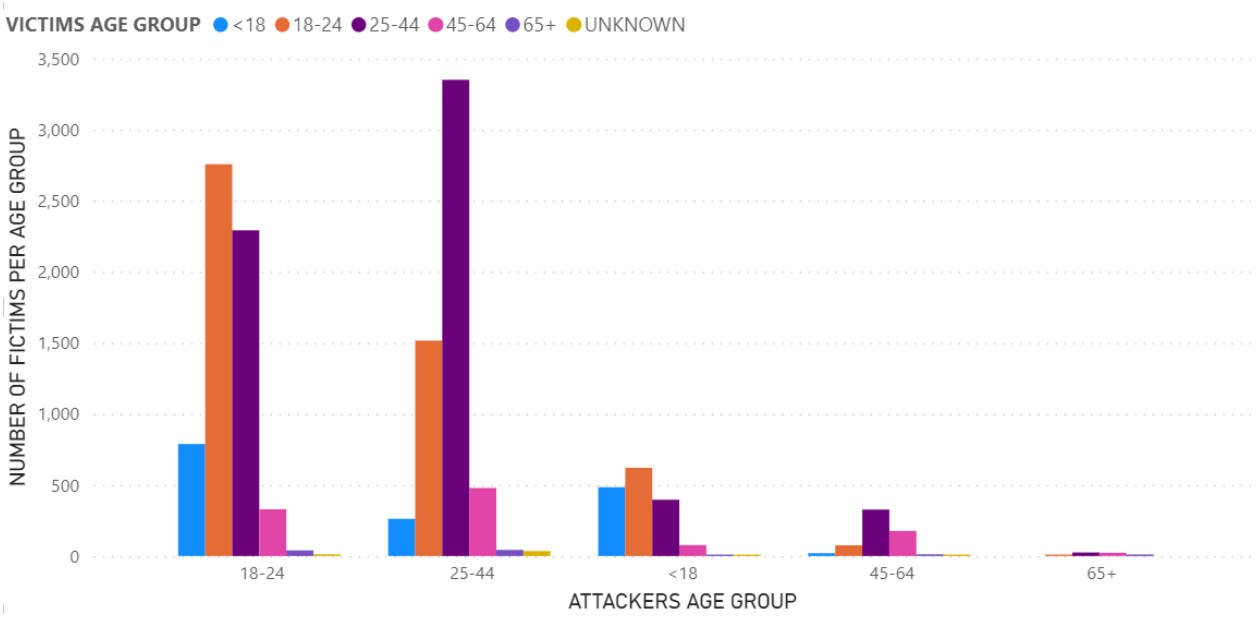
Again, in Manhattan, the safest race Is Asian/Pacific Islander.

Staten Island is the safest state for Asian/Pacific Islander. The crime against all other races was also very low. Black race being the most victimized race in Staten Island has less then 500 cases which is around only 10% of cases reported in Brooklyn for black people.
American Indian/Alaskan Native were the safest race in all of the boroughs of New York as having almost no cases reported in each borough.



Lets see if there is a relation between the attacker's age and victims age.
The attacker aged between 18 to 24 has affected victims from age 18 to 24 the most with around 2700 cases reported. The 2nd most affected age group was 25 to 44 from around 2300 cases reported.
The attacker aged between 25 to 44 has affected victims aged from 25 to 44 with around 2900 cases reported. The 2nd most affected group was aged from 18 to 24 with 1500 cases reported.
The attacker aged less than 18 has affected victims aged from 18 to 24 the most with around 650 cases reported. The 2nd most affected group was aged less then 18 with around 500 cases reported.

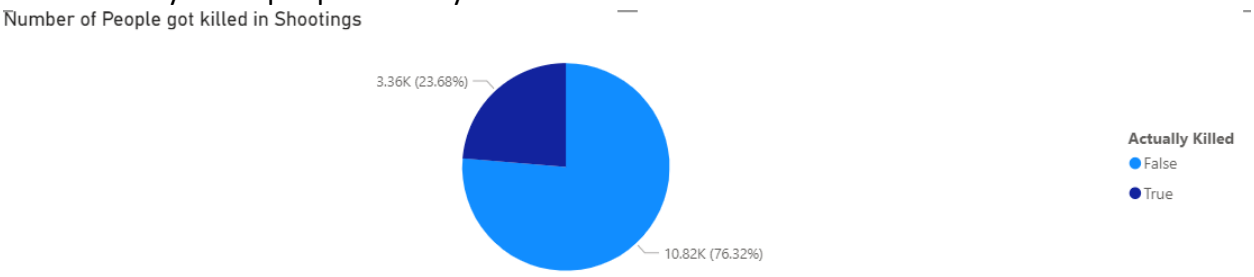
The attacker aged between 45 to 64 has affected victims aged from 25 to 44 with around 300 cases reported. The 2nd most affected group was aged from 45 to 64 with 200 cases reported.
The attacker with more than age 65 has affected very little amount of people which are around 100.



Mass shootings incidents were reported every month so let's see which month has the highest number of mass shootings incidents in each boroughs.
For Bronx, the highest number of May, June, July while August being the month with highest number of cases reported which are 500.
For Brooklyn, the increase in the number of cases was reported from May to August with July being the month with the greatest number of cases reported which is 661.
For Manhattan, the increase in the number of cases was reported from May to August with August being the month with the greatest number of cases reported which is 220.
For Queen, the increase in the number of cases was reported from May to August with July being the month with the greatest number of cases reported, which is 214.

For Staten Island, the increase in the number of cases was reported from May to August with May and July being the months with the greatest number of cases reported, which is 62.

From the following Pie chart, we can see that not all of the people got killed during the shooting incidents. Out of 14.18k people 10.82k people didn't die fortunately but unfortunately 3.36k people actually died due the incidents.



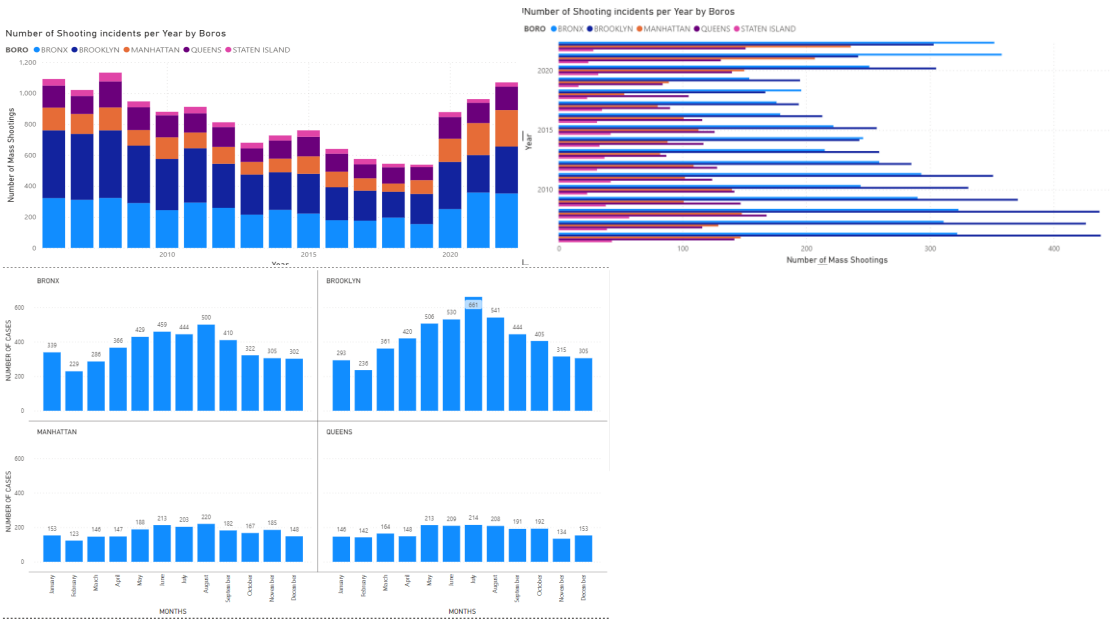
Conclusions:

- The Borough with the highest number of cases reported was Brooklyn while the safest borough was Staten Island.
- The most victimized race was Black according to the visualizations on the dataset.
- The number of mass shootings were gradually decreasing but during and after covid a rapid increase can be seen in the cases till 2022.
- The race which was involved in the mass shooting incident the most was black according to the dataset.
- The age group of attackers and victims that were mainly involved in the shootings was 25-44.
- The months with the greatest number of cases reported were July and August.
- The summer season has the most number pf shooting happened.

5. Visualizations:

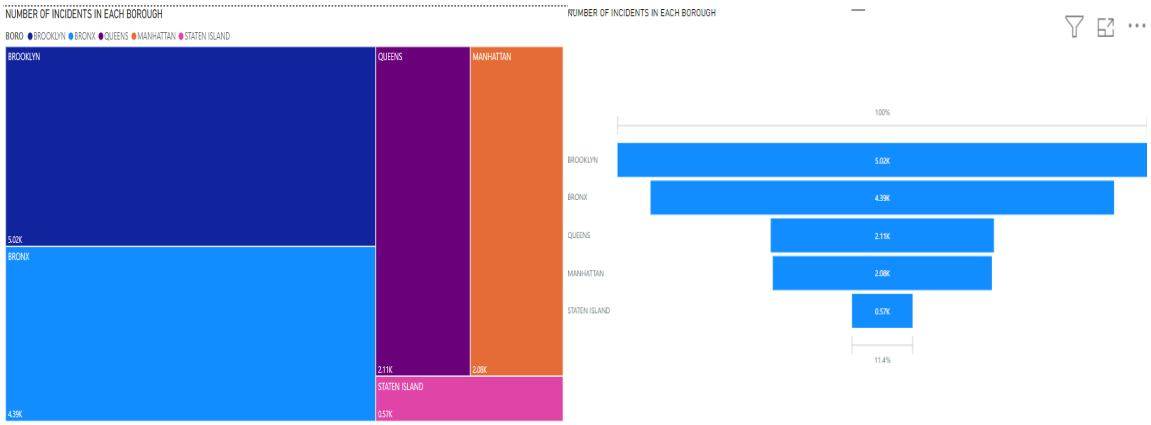
Column Chart:

Using stacked column charts was a better option in some cases like shown in the example below. From the graphs below we can clearly see that using stacked column chart was a better option. Both the graphs represent the same data, but the visualization looks much easier to understand using a stacked column chart.



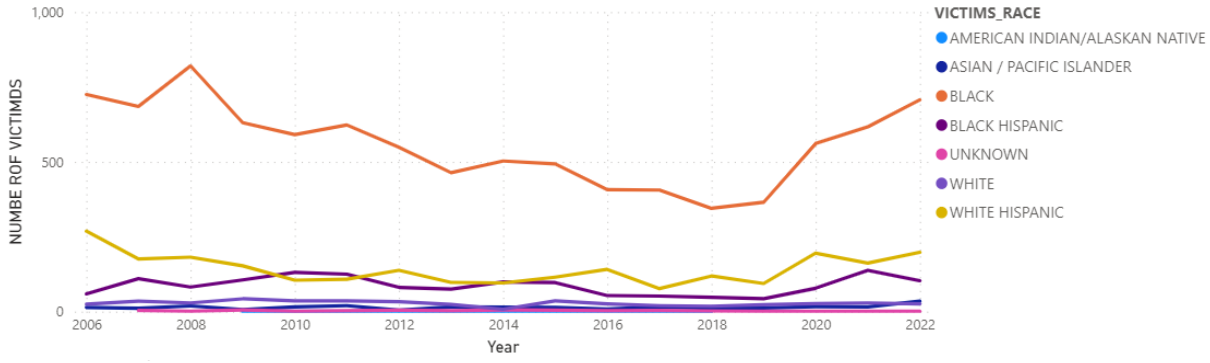
TREEMAP:

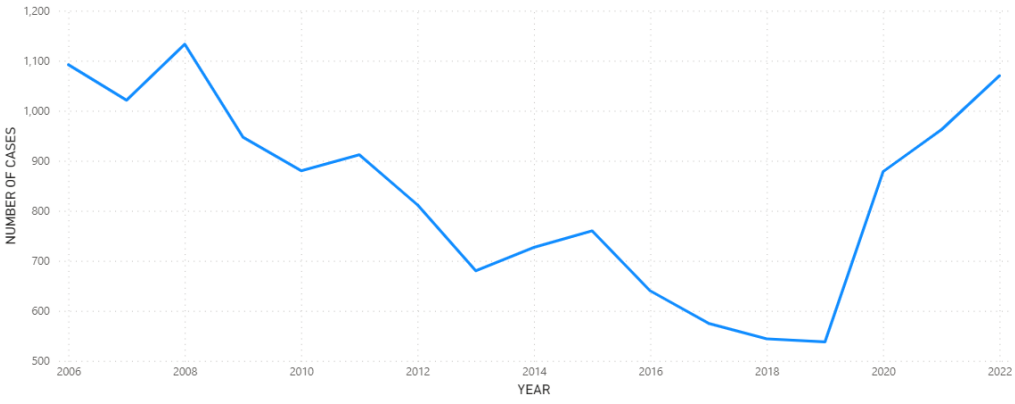
Used a tree map to see the total number of cases reported in each of the boroughs of New York. From the following comparison we can see that using a Funnel could be a better option for this kind of data as it is clearer and easier to interpret.



Line Charts:

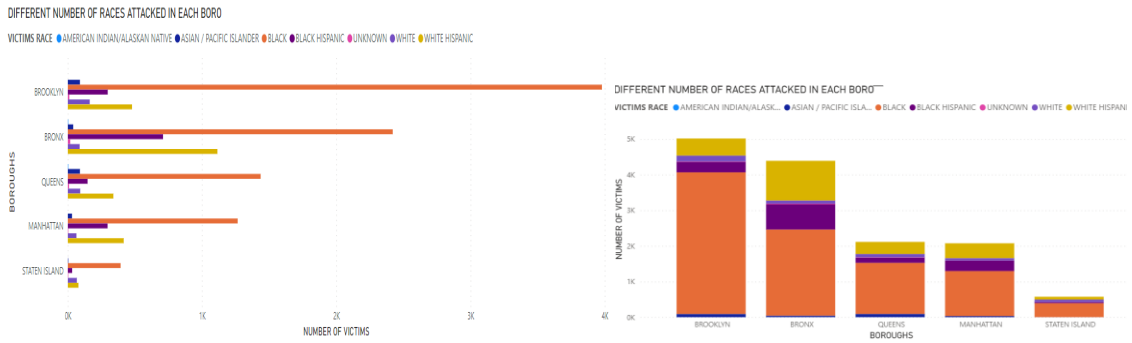
Line graphs are quite useful when we need to tell changes over time. The following line charts make it easier to understand the change over time. We can clearly see the which how the trend changes over time

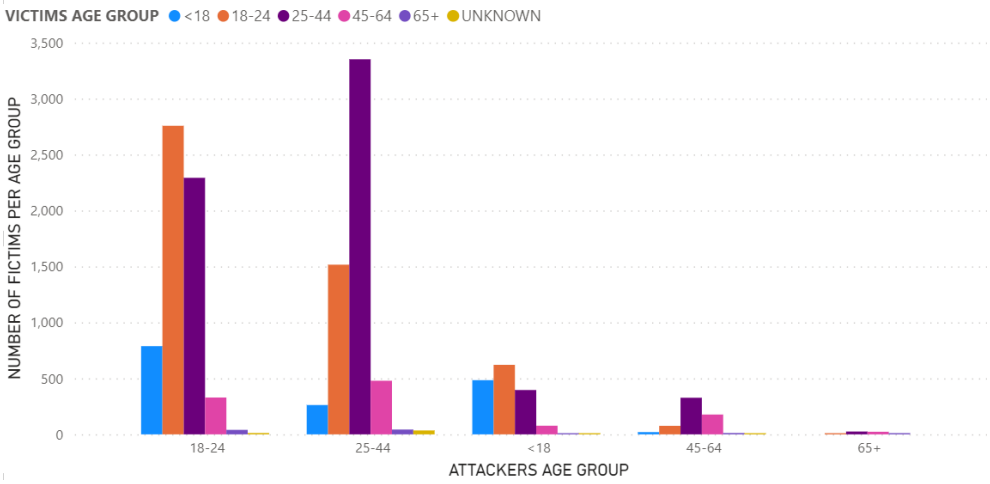




BAR CHART:

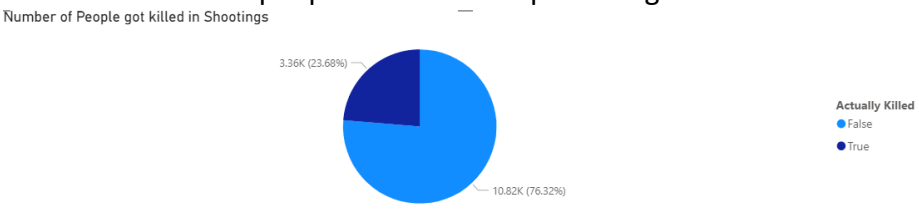
I have used the bar chart here to represent the categorical data. Using bar chart here allows to have a better understanding of data. Using the column chart here instead of bar chart would end up having a bad visualization because it is difficult or almost impossible able to visualize the categories that have a minimal value.





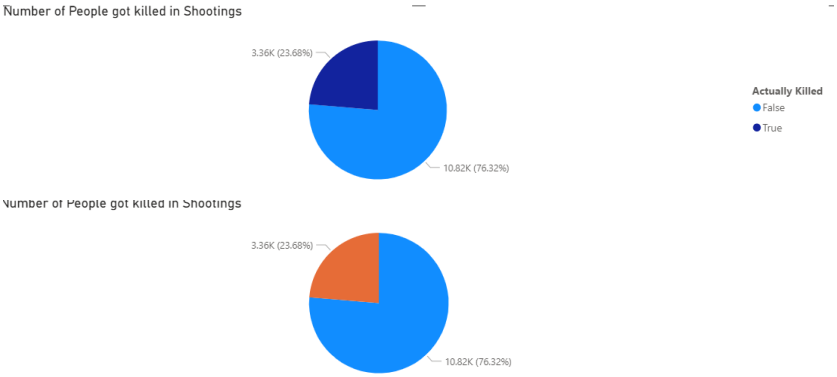
PIE CHART:

I Have used Pie chart here as I just needed to tell the difference between the people who got killed and who didn't die. Using a pie chart here is the best option as it clearly tells the number of people and the total percentage too who lived and died.



COLORS:

Using the color contrasts for color blind people is an important task in data visualization. In the following example, my pie chart was initially blue and dark blue but changing the dark blue to orange will result in better visualization as there is a difference between the contrast of orange and sky blue.



Report title	ANALYSING THE NEW YORK MASS SHOOTINGS
Name/Id of student seeking feedback	W1863288
Name/Id of student giving feedback	W1736783
Reviewer's comments (what is good, what could be improved?)	
The Bar Charts used for the categorical data are very help in understanding different number of races attacked in each borough and which age group of attackers has attacked which age group the most. Using a funnel chart instead of tree map for representing the total amount of cases reported in each brough.	
Feedback given on (date)	14/05/2023
Reviewee's comments (What – if anything – did you change in your report after the feedback?)	
Changed the tree map into a funnel chart.	

Report title	ANALYSING THE NEW YORK MASS SHOOTINGS
Name/Id of student seeking feedback	W1863288
Name/Id of student giving feedback	W1764460
Reviewer's comments (what is good, what could be improved?)	
Overall graphs are good but the line chart made it clearer to see the change in trend over time especially it could be seen that the number of mass shooting has increased after covid. Try changing the column chart into bar chart, it might have a better visualization.	
Feedback given on (date)	14/05/2023
Reviewee's comments (What – if anything – did you change in your report after the feedback?)	
I did not change anything.	

Report title	ANALYSING THE NEW YORK MASS SHOOTINGS
Name/Id of student seeking feedback	W1863288
Name/Id of student giving feedback	w1702010
Reviewer's comments (what is good, what could be improved?)	
I really liked the visualization as all of were easy to understand. Why don't you try changing the pie chart as it is not preferable everywhere. Try to change it into a column chart or something	
Feedback given on (date)	15/05/2023
Reviewee's comments (What – if anything – did you change in your report after the feedback?)	
Didn't change anything	