0) Name and Email of each team member.

Fiona Fei. Email: fei.90@buckeyemail.osu.edu

Edison Gu. Email: gu.649@buckeyemail.osu.edu

Shay Sun. Email: sun.1983@buckeyemail.osu.edu (GRADUATING)

1) Introduction: **Give an overview of your project: the motivation, the background, the goals, etc.**

In recent years, police shootings have come to the focus of the public. People looked into the statistics and observed something problematic. “Racism”, “Police Brutality” became some of the most popular topics on social media and political debates during 2020, as people’s attention was drawn to the names “George Floyed”, “Breonna Taylor”, etc.

We would like to take a deep dive into this topic by looking at data from 2015 to present. Is there compelling evidence showing that there is a systemic bias towards a certain demographic? Is the problem more widely present than we realized? And how big of a gap do we have in order to overcome this problem?

We will be exploring these questions through analysis and present our results in an interactive dashboard that can be used by news media, lawmakers, and other stakeholders to understand this subject better.

2) The data sets: What are your data sets? Where did you collect them?

The main dataset that we will be using is the US Police Shootings dataset (https://www.kaggle.com/ahsen1330/us-police-shootings)
It was found on kaggle and will be downloaded in a .csv file.

3) Data preprocessing: Did you preprocess your data? If yes, what did you do and why?

Yes. The data was preprocessed by eliminating the first column “id” because it doesn't provide us any useful information here. I also checked if there’s any missing data using msno.matrix(df) function, and discovered that there weren’t any missing values. Lastly, I transformed the type of ‘date’ column to date using df['date'] = pd.to_datetime(df['date']) function.

For some of the temporal analysis, the 2020 data was excluded due to the data incompleting. Year, month, quarter, and week columns were appended to the dataframe for further analysis.

4) Detailed analytical questions: What questions do you ask from the data?

1. Does police shoot a certain demographic more likely than the others? (EG)
2. What's the reason for being shot - were they armed? (EG)
3. Which states/cities have more police shootings than others? What's the race distribution in the top three states that has the most police shooting cases? (FF)
4. Do we see any patterns of police shooting cases from January 2015 to June 2020 that are associated with time? （SS）
    a. We would love to understand how the police shooting cases changed annually, quarterly, monthly, and weekly?
    b. Are there any interaction effects between certain demographic features and temporal features?
    c. Does there exist any interaction effects between spatial features and temporal features?
5. Does police shoot people of certain races even if they do not have "threat level" marked as "attack"? (FF)
6. If police's body cameras were off, will the shooting cases towards the victims who show no attacking behaviors increase? (SS)

5) Visualization requirements: What are the requirements of the visualization in order to answer the questions in (4)?

1. For Gender distribution, Race, and Age distribution, the requirement is simple - we merely have to count the number of individuals in each category. There are 2 variables in each of these graphs.

    However, for the third graph "US Population Demographics vs Victim's Demographic". We calculated the percentage of victims that belong to each race, and also using an external data source to calculate the entire US population demographic.

2. For Weapons, we need to calculate the number of victims who were armed with different weapons. Then sorting is needed to show the relative magnitude of each weapon. We can then draw conclusions on if victims were armed with weapons, and if so, what kind of weapons.
3. The requirement for answering which state or city has more police shooting cases is to show the hierarchical order among all states/cities. The visualization needs to show the order of these states/cities in a clear manner. It also needs to have the ability of letting people know how different races are distributed within the same state.
4. In general, time is a sequential variable with clear direction and order. Different blocks of information within the same temporal dimensions or units (such as hourly, daily, weekly, etc) could be used for comparison. Higher temporal dimensions could be sectioned into sub dimensions (for example, yearly dimensions could be

represented by multiple monthly dimensions). More specific visual requirements for each sub questions of Question 4 are discussing below:

a. In order to understand how the police shooting cases changed annually, quarterly, monthly and weekly, we need to accurately scale the information which is the total number of police shooting cases in this scenario either up or down into the correct temporal dimensions. And the chosen visualization could clearly separate the representations for each temporal dimension. For the purpose of seeing the changes, visualizing the clear comparisons moving in order along the timeline is critical.
b. In order to visualize the demographic and temporal information together, depending on how complex the demographic features are, some categories or values might need to be further grouped. To visualize this sub question, the total number of shooting cases at each demographic level with each temporal dimension should be clearly calculated.
c. In order to visualize the spatial and temporal information together, the base dimension should be carefully chosen. In this scenario, states of America would be our spatial dimension along with four temporal dimensions (weekly, monthly, quarterly, annually). To visualize this sub question, each state with each temporal dimension should be clearly separated.

5. To answer this question, we need to compare two graphs (total cases for threat level being considered as attack and not attack together under the same graph type. Moreover, in order to see the exact number of counts, we need to have a table along with the bar chart.
6. To answer this question, a clear visualized comparison between the number of shooting cases while the camera is on and the number of shooting cases while the camera is off is the key. The summation of which two parts should form into one whole piece.

6) Visualization design: What is the design of your visualization? What are the marks/channels/layouts/interactions used in your visualization and why do you believe they are good choices?

1. For Race distribution at different ages, we had to calculate the number of victims at each age and group the data by race. There are 3 variables - Count, Age, and Race. I used the size of the bar to represent the counts and different bar colors to represent race.

   I chose bar charts because it can most directly show the relative magnitude compared to the other categories. By using color encoding, we can further differentiate the demographic information within each category.

   Note that in the third graph, US population demographic vs victim's demographic, we also made sure that the percentage adds up to 100%. This can give the audience a

more direct comparison between the two groups - Total US demographic and the victim's demographic.

2. For Weapons, we first count the number of victims that were armed with different weapons. Then we sorted the bars to select the top 5. By hovering over each bar the user can see the actual numbers of victims that were shot due to the possession of different weapons. This allows a direct comparison as well.

3. In order to find out the top states and cities with the highest counts for police shootings, we need to first sort out all the states and cities by counts. Next, we will pick out the top ten cities/states and make two bar charts that show the descending order of the cities/states. Finally, we will choose the top three states and make a stacked bar chart to show the proportion of different races. I used the height of the bar to represent the counts of cases, and I used different colors to represent different races. In this case, the area of the bar is the mark, the vertical lengths and horizontal position of the bar charts are the channels.

4. How the police shooting cases changed annually, quarterly, monthly, and weekly?

   A line graph for each temporal dimension will be used to visualize the overall temporal patterns of police shooting cases. Mark is the line. Channels are the x-axis representing years, quarters, months, or weeks and the y-axis representing the total number of police shooting cases. Four line graphs will be arranged in a 2x2 matrix panel.

   A heat map will be used to visualize how the shooting cases changed by months in certain years or by years of the same month. Mark is the squared or rectangular area. Channels are the colors for different levels, x-axis representing one time dimension, and y-axis representing the other time dimension.

   Are there any interaction effects between certain demographic features and temporal features?

   A heat map will be used to visualize the relationship between various the number of shooting cases and different time dimensions on the various demographic levels. Mark is the squared or rectangular area. Channels are the colors for different levels, x-axis representing one time dimension, and y-axis representing levels of demographic features. A dashboard will be applied to organize the individual plots together with the user selection function being provided.

   Does there exist any interaction effects between spatial features and temporal features?

For each temporal dimension, a grouped bar chart will be used to visualize the relationship between the total number of shooting cases and time dimensions. Different states will be distinguished by different colors. Marks are bars (rectangles). Channels are colors and width and height of the bars along the x-axis. A dashboard will be applied to provide users choices to switch between different temporal dimensions. They could also zoom in to the certain temporal section for more detailed information.

5. We will be using two bar charts to compare the total cases of police shooting for the threat level considered as attack and not attack. The marks are 2D areas of the bar chart, and the channels are the vertical lengths and horizontal positions of the bars.

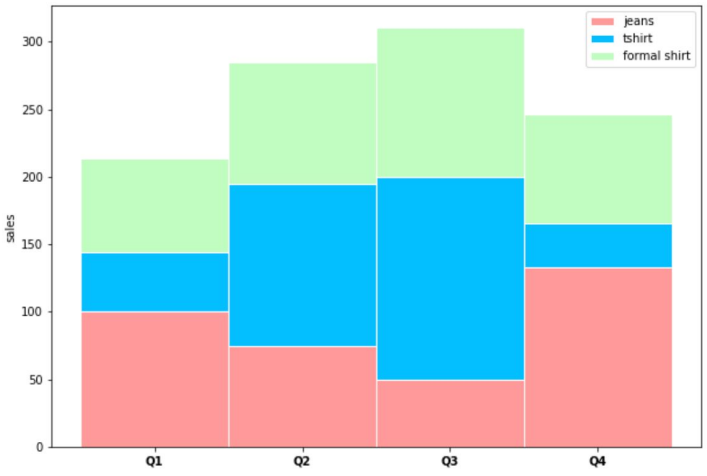
6. A donut chart will be used to visualize the comparison between the total number of police shooting cases while the camera is off and while the camera is on. Mark is the circular band. Channels are colors and proportions.

7) **Software**: Describe your visualization software: the components, the internal structure, the interface, and how to run it. Make sure you include screenshots of your software and some sample visualization. You need to acknowledge all external sources (libraries, code examples/segments done by someone other than the team, the website etc.).

Python, NumPy, Pandas, Ploty, Matplotlib, missingno, dash

1. Sample visualization:

https://priteshbgohil.medium.com/stacked-bar-chart-in-python-ddc0781f7d5f



Screenshot of my software:

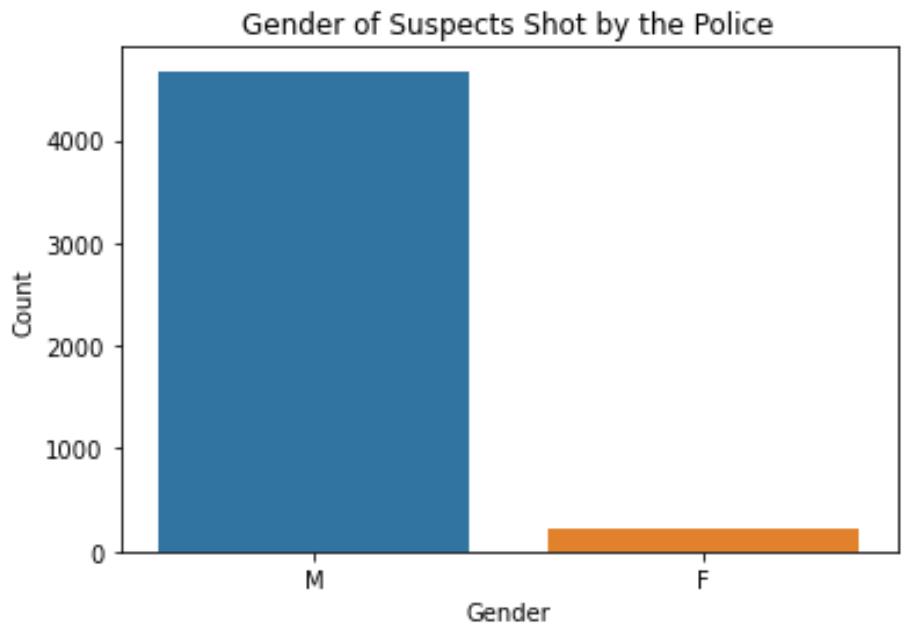
In [76]:
```
# stacked bar plot
# This part of code was modified from https://priteshbgohil.medium.com/stacked-bar-chart-in-python-ddc0781f7d5f

#Get values from the group and categories
state = ["CA", "TX", "FL"]
white = [221, 175, 156]
black = [122, 99, 111, ]
hispanic = [316, 142, 51]
asian = [29,6,1]
native = [4,1,0]

#add colors
colors = ['#ABE6CE', '#DCEDC2','#FFD3B5','#FFAAA6','#FF8C94']
# The position of the bars on the x-axis
r = range(len(state))
barWidth = 1
#plot bars
plt.figure(figsize=(10,10))
ax1 = plt.bar(r, white, color=colors[0], edgecolor='white', width=barWidth, label="white")
ax2 = plt.bar(r, black, bottom=np.array(white), color=colors[1], edgecolor='white', width=barWidth, label='black')
ax3 = plt.bar(r, hispanic, bottom=np.array(white)+np.array(black), color=colors[2], edgecolor='white', width=barWidth,
ax4 = plt.bar(r, asian, bottom=np.array(white)+np.array(black)+np.array(hispanic), color=colors[3], edgecolor='white',
ax5 = plt.bar(r, native, bottom=np.array(white)+np.array(black)+np.array(hispanic)+np.array(asian), color=colors[4], ec
plt.legend()
# Custom x axis
plt.xticks(r, state, fontweight='bold')
plt.ylabel("counts")
for r1, r2, r3, r4, r5 in zip(ax1, ax2, ax3, ax4, ax5):
    h1 = r1.get_height()
    h2 = r2.get_height()
    h3 = r3.get_height()
    h4 = r4.get_height()
    h5 = r5.get_height()
    plt.text(r1.get_x() + r1.get_width() / 2., h1 / 2., "%d" % h1, ha="center", va="center", color="white", fontsize=1
    plt.text(r2.get_x() + r2.get_width() / 2., h1 + h2 / 2., "%d" % h2, ha="center", va="center", color="white", fonts
    plt.text(r3.get_x() + r3.get_width() / 2., h1 + h2 + h3 / 2., "%d" % h3, ha="center", va="center", color="white",
    plt.text(r4.get_x() + r4.get_width() / 2., h1 + h2 + h3 + h4 / 2., "%d" % h4, ha="center", va="center", color="whit
    plt.text(r5.get_x() + r5.get_width() / 2., h1 + h2 + h3 + h4 + h5/ 2., "%d" % h5, ha="center", va="center", color=
plt.savefig("stacked2.png")
plt.show()
# You can replace "%d" % h1 with "{}".format(h1)
```

2. https://github.com/Coding-with-Adam/Dash-by-Plotly This link has been referenced for dashboard making.

8) Results and Evaluation: Provide the answers and visualization to the questions in (4). Explain how your answers are supported by the visualization. Explain whether you think visualization is essential to answer the questions and why.

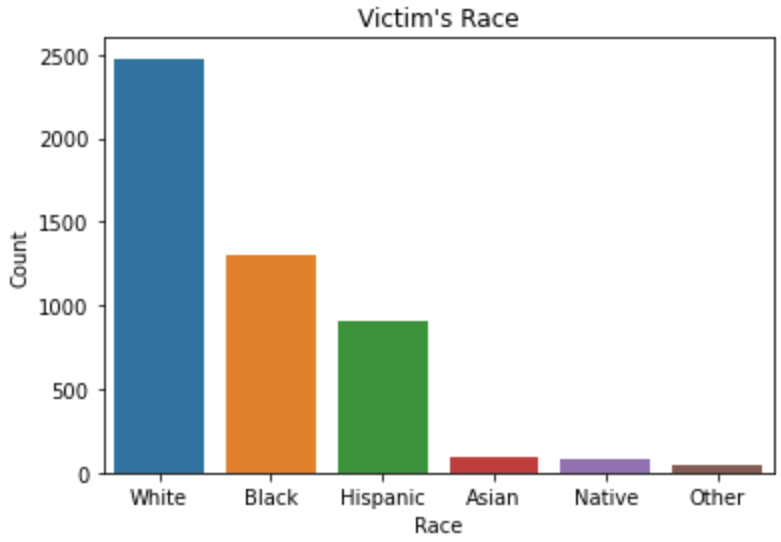
Question 1: Does police shoot a certain demographic more likely than the others?

An obvious demographic trait that stands out in the data set is that almost all of the suspects who were shot by the police were male:

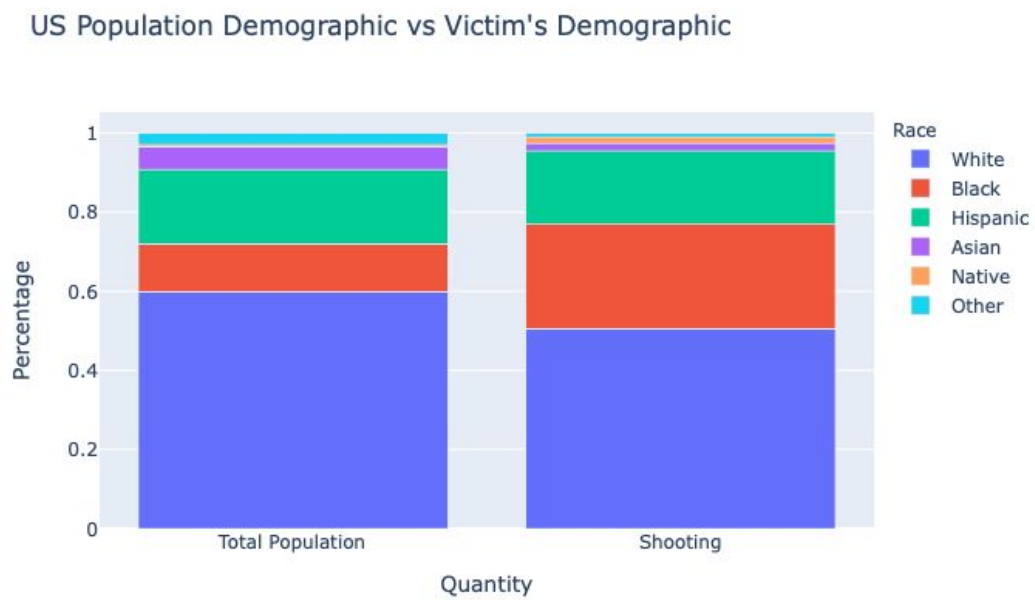


This may be because the police consider male suspects are more likely to attack.

From another point of view, I plotted the Count against Race:

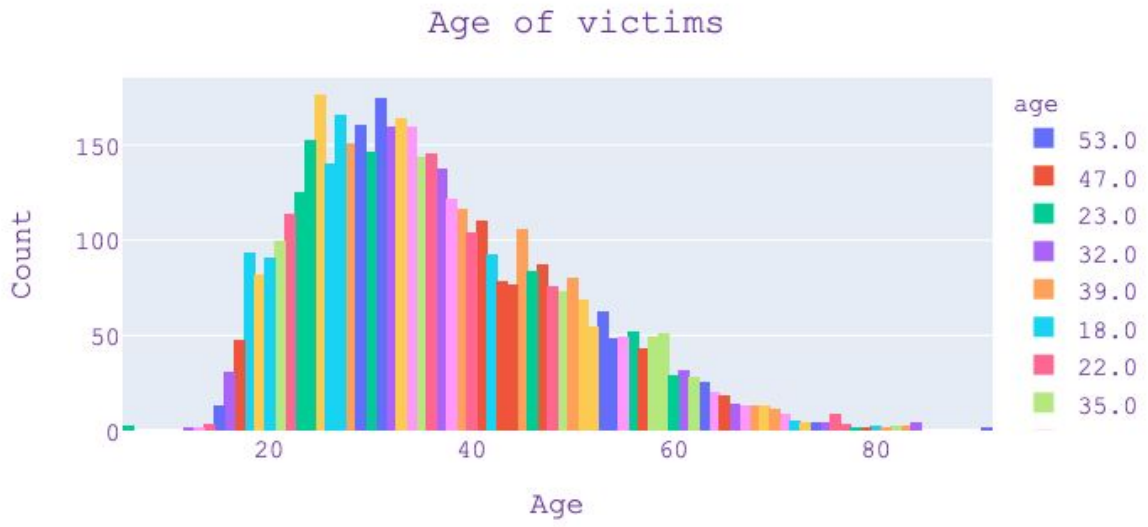


From the first glance, we may say that it looks like White people are being shot by the police the most. However, it is important to be aware that the majority of the US population is white. So, how does this distribution compare to the overall US population demographic?

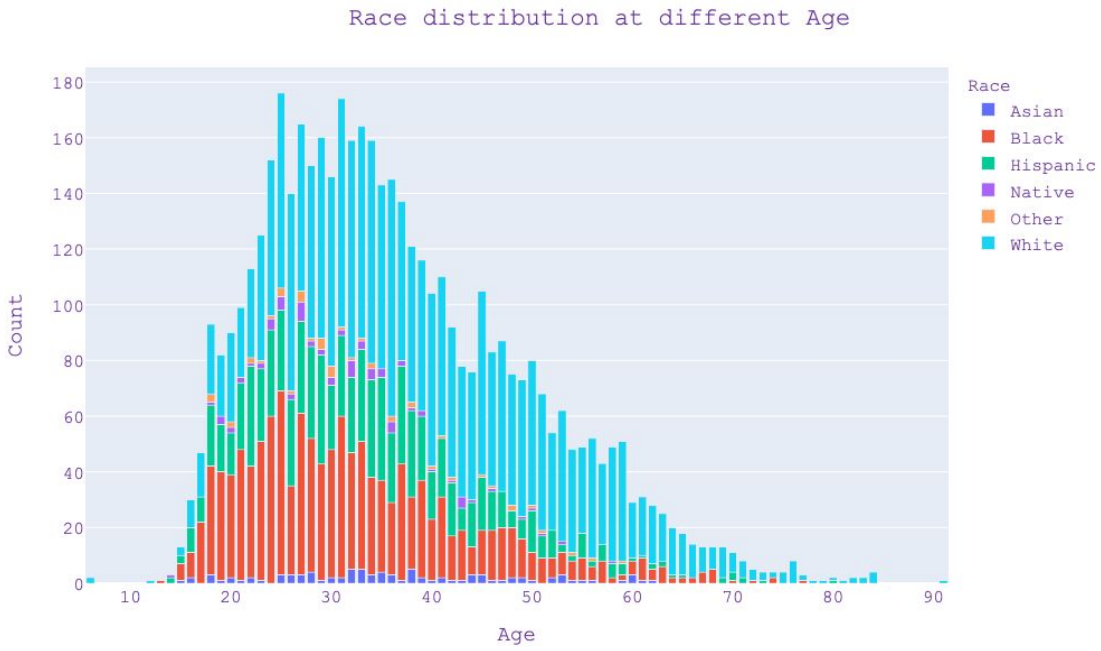


From this graph, we see that a disproportionate percentage of the victims are black, whereas, Hispanic people hold the same percentage, and Whites and Asians are less likely to be shot. This, which can be interpreted as an conditional probability, shows that a black person is more likely to be shot by the police - with 12% of the total US population being black, 27% of the police shooting victims are black.

We can also look at Age as a factor:

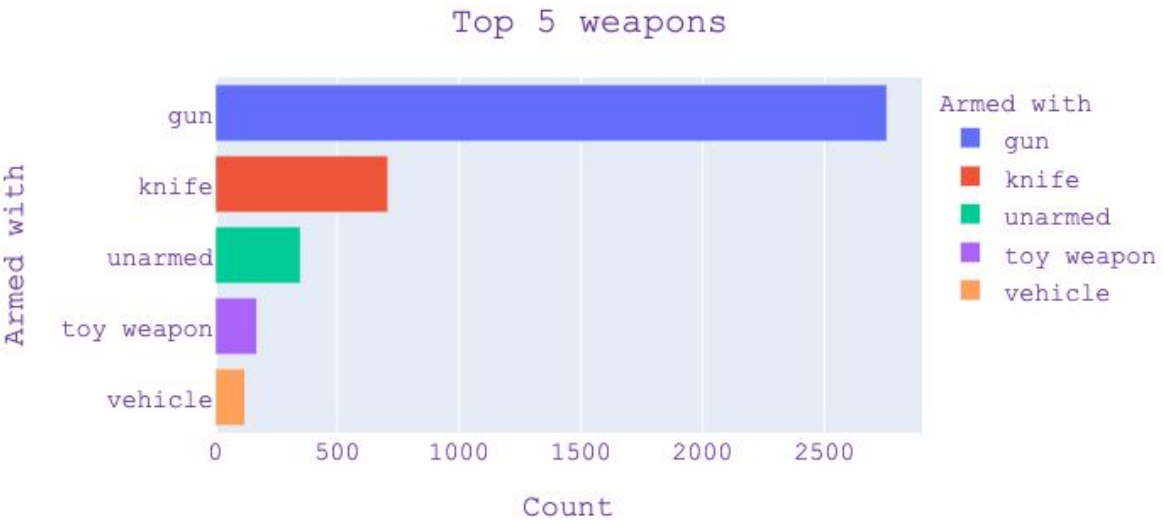


From this graph, we see that the majority of the victims were between 20's and 40's. If we break it down further by age:



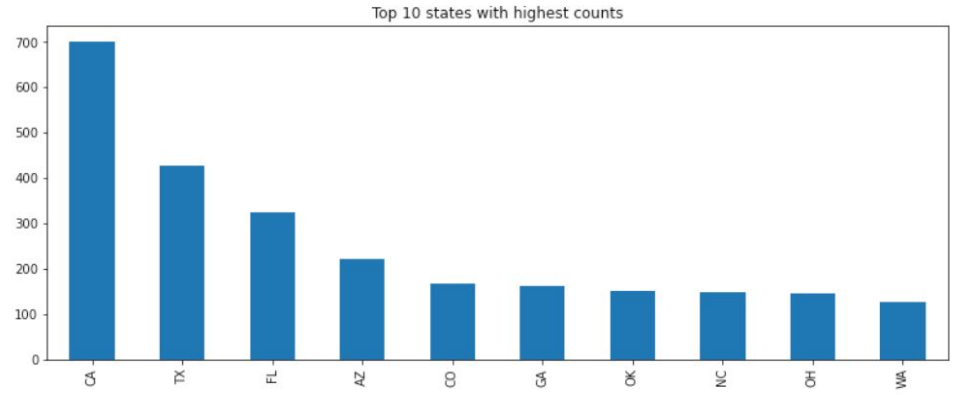
We can see that the majority of younger victims (20 - mid 30's) are black and hispanic, whereas more older victims (40+) are made up of white populations. This poses a serious implication: young black, usually male, victims got shot by the police - they are killing the father of a household - this potentially takes away the economic support of a family, leaving their children without a father at a very young age.

Question 2: What's the reason for being shot - were they armed?



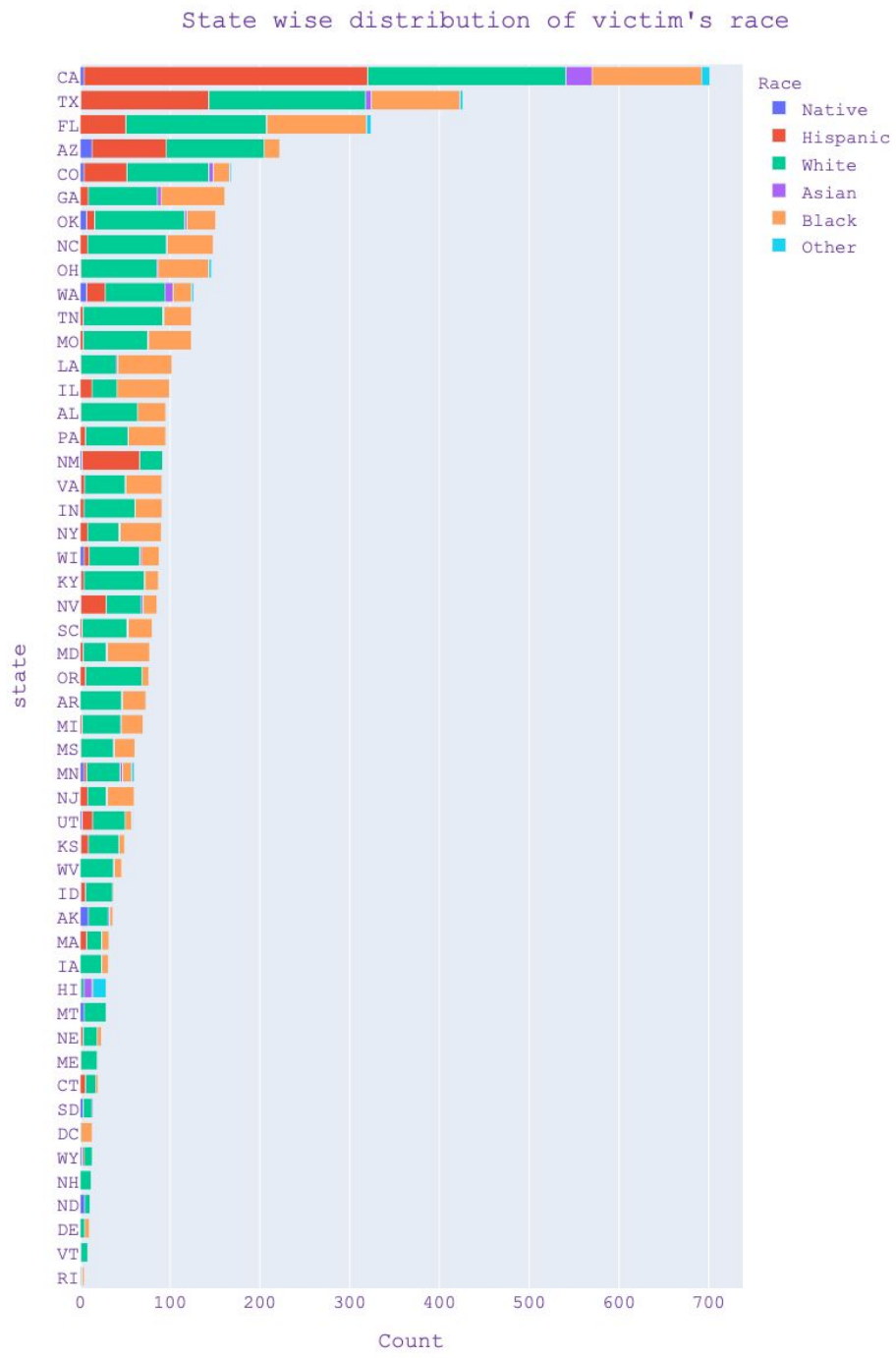
While most victims were armed with guns, many were also killed merely because they were holding a knife - which is less dangerous than a gun. And around 348 were killed unarmed and 171 were killed for possessing toy weapons. These numbers are not small and should sound an alarm

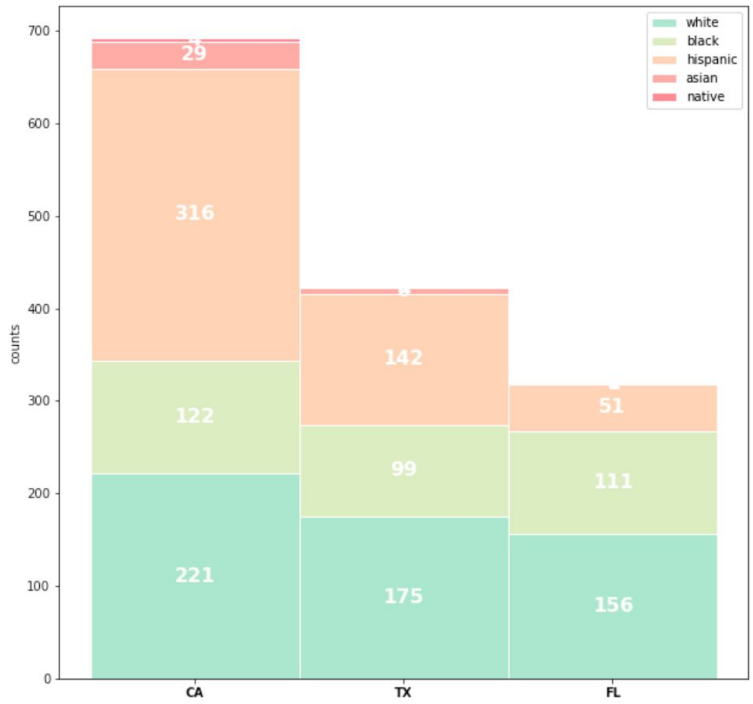
Question 3: Which states/cities have more police shootings than others? What's the race distribution in the top three states that has the most police shooting cases?



.

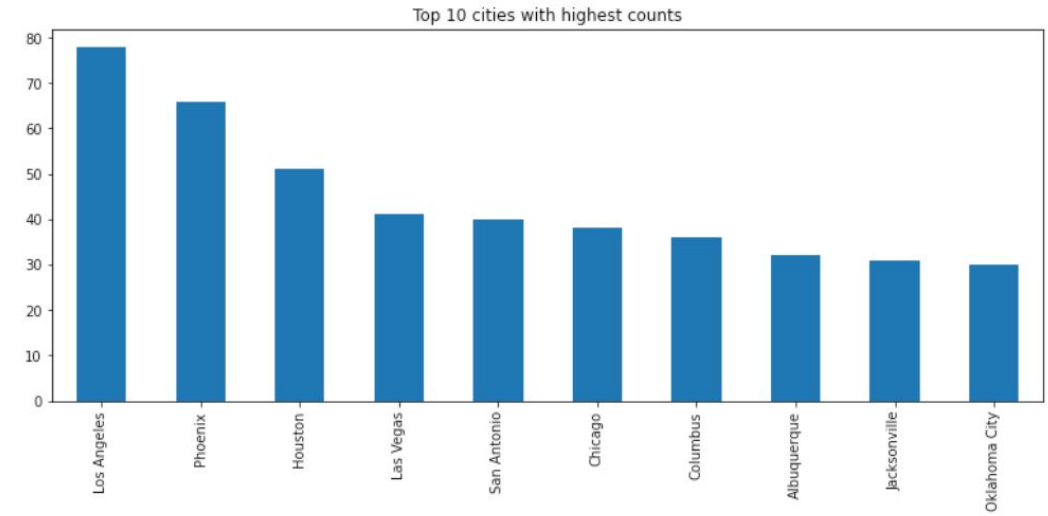
By looking at the top 10 states with highest counts, we can discover that California, Texas, and Florida are the top three states that have the highest counts of police shooting cases. Thus, we choose to take out the three states and take a look at the race distribution of these cases.


State wise distribution of victim's race
Race
Native
Hispanic
White
Asian
Black
Other
state
CA
TX
FL
AZ
CO
GA
OK
NC
OH
WA
TN
MO
LA
IL
AL
PA
NM
VA
IN
NY
WI
KY
NV
SC
MD
OR
AR
MI
MS
MN
NJ
UT
KS
WV
ID
AK
MA
IA
HI
MT
NE
ME
CT
SD
DC
WY
NH
ND
DE
VT
RI
0
100
200
300
400
500
600
700
Count



From the stacked bar plot, we can see that hispanic is the race with the highest proportion in the state California, followed by white, black, asian and native. As for Texas, the cases are mostly white people, followed by hispanic, black, asian, and native. Lastly, white is the race with the highest counts of cases in Florida, followed by black, hispanic, asian, and native.

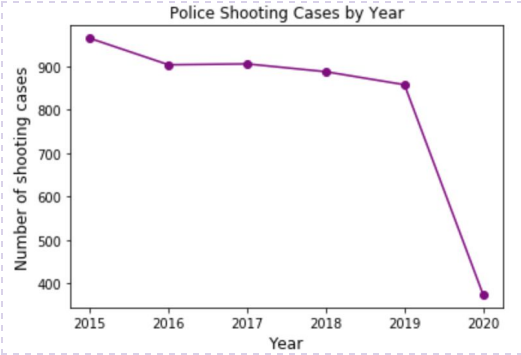
Next, we are going to check which are the top 10 cities with highest counts.

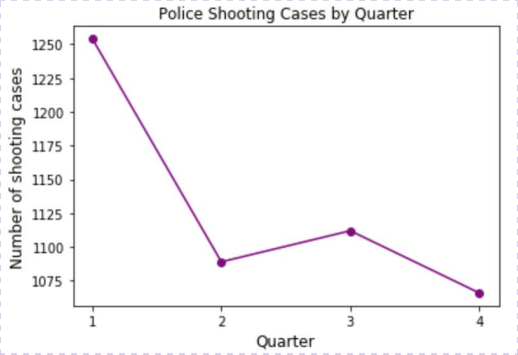


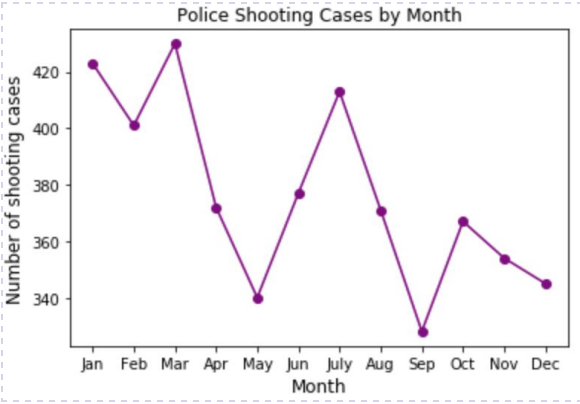
We can see from the plot above that the top three cities with highest counts are Los Angeles, Phoenix, and Houston.

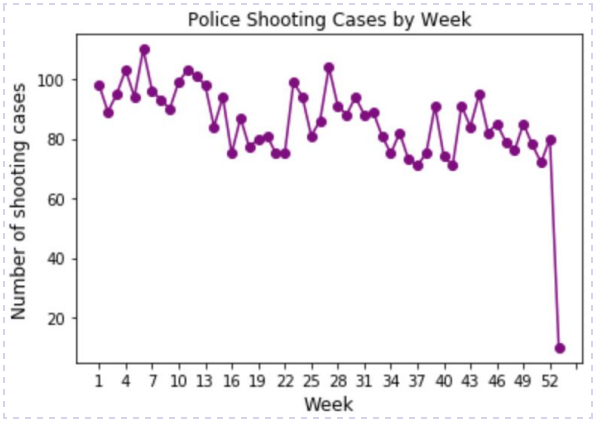
Questions 4: Do we see any temporal patterns associated with police shooting cases

- How the police shooting cases changed annually, quarterly, monthly, and weekly?



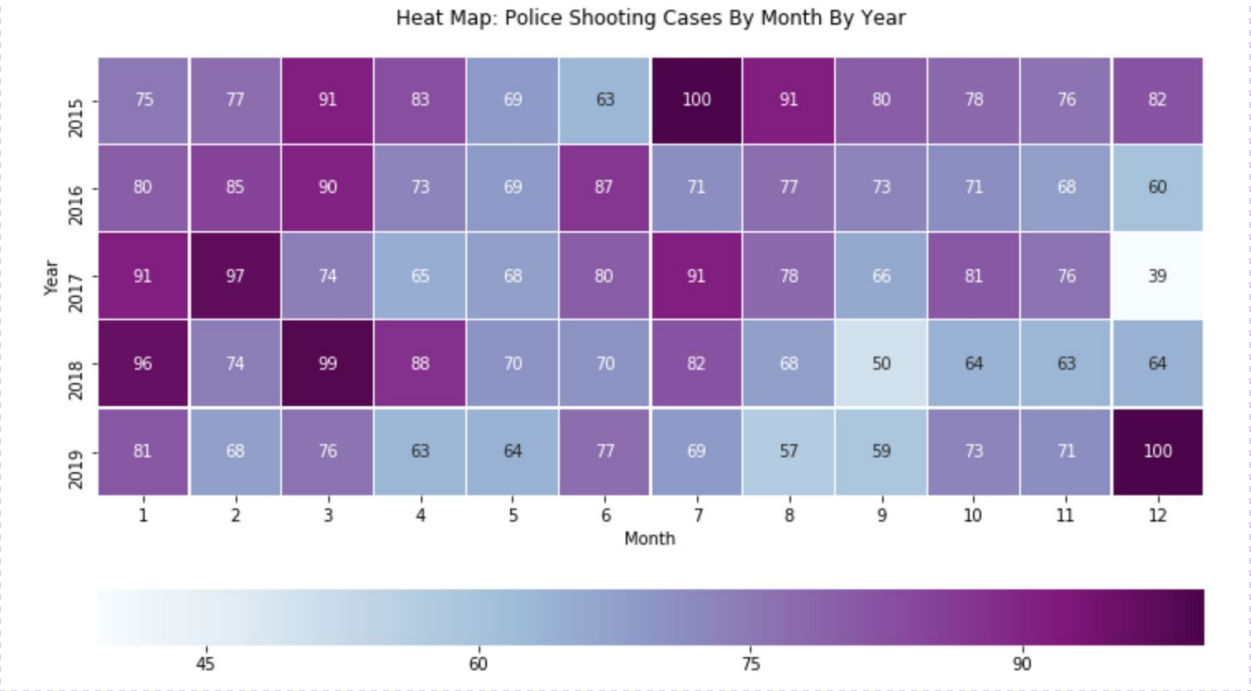






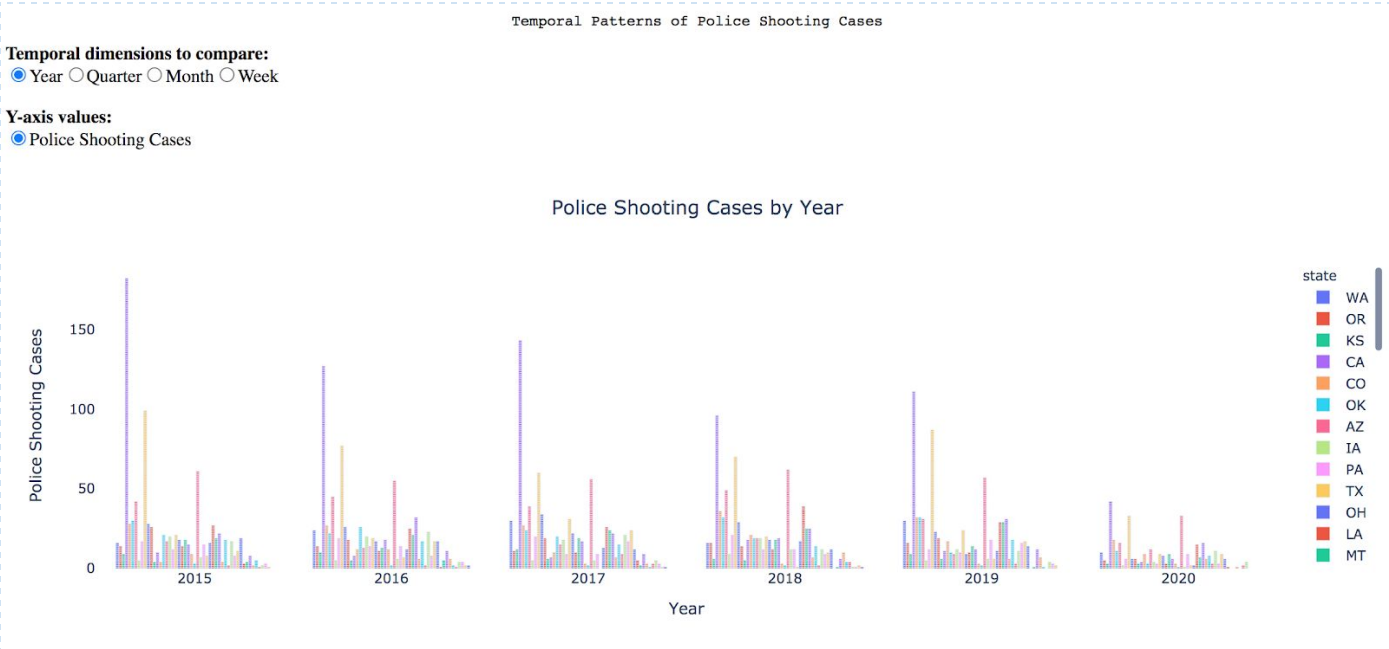
Since the data only covers until June 2020, then data in 2020 will be excluded in the cases that matters. In the line charts above, 2020 data was excluded except for visualizing the annually pattern. From an annual perspective, the total number of shooting cases decreases year by year. Due to the missing data for year 2020, we see 2020 has the lowest cases which should be ignored. The first quarter is associated with the highest shooting cases which is consistent with what is being shown in the monthly plot. We could also observe a huge spike in July. This might be interesting to further investigate. Police shooting cases reached the lowest point in the fourth quarter which might be attributed to the rapid drop in week 52. Overall, the shooting cases wend down along the timeline.

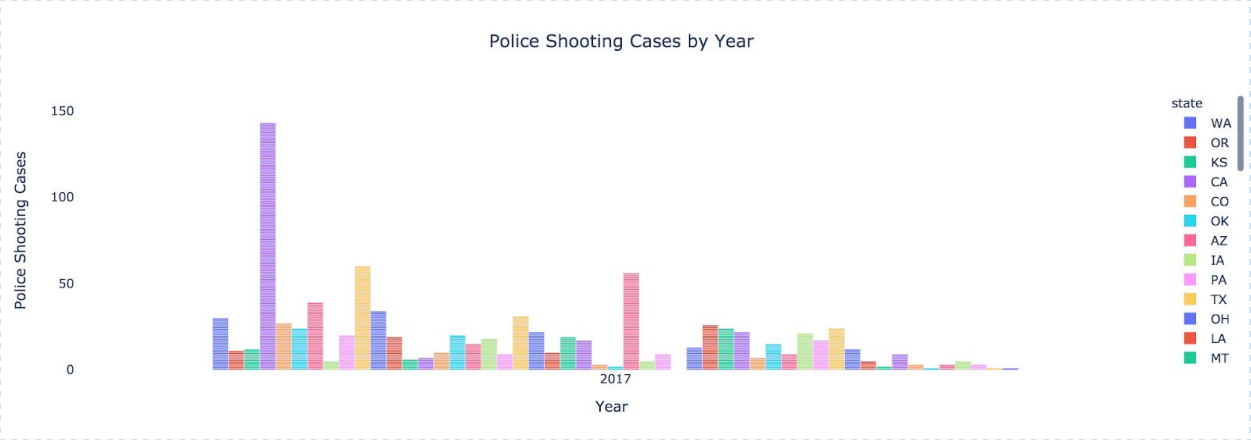
- Police shooting cases by year across months



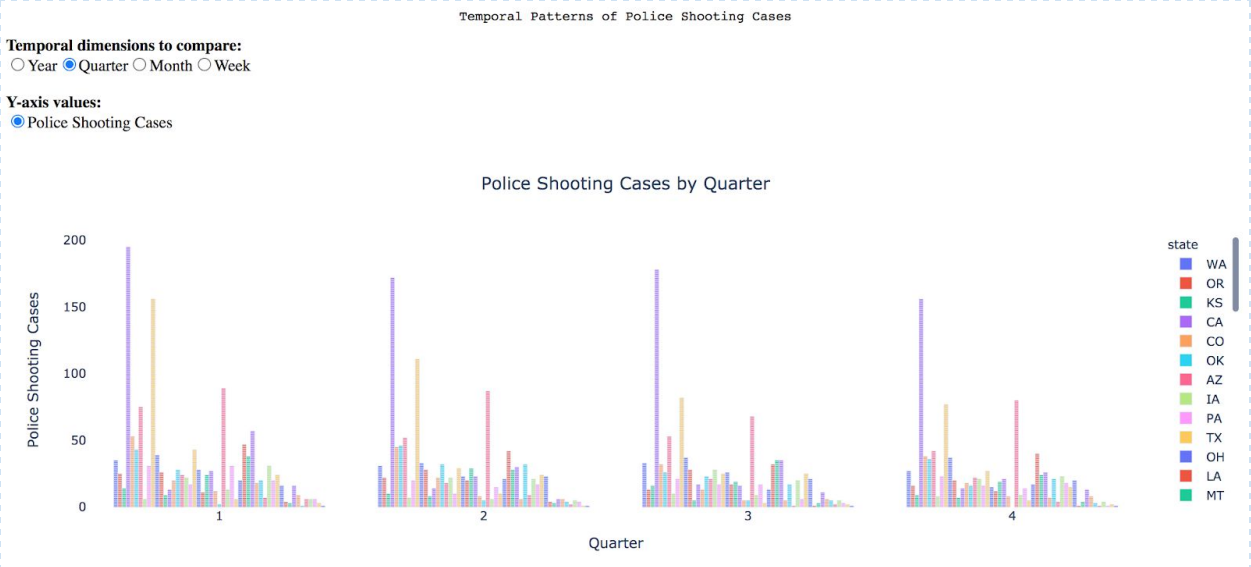
The result from the heat map above is consistent with the results from the four line graphs. We could observe an overall decrease in the first half of each year until June or July followed by a huge spike. In the second half of each year, the shooting cases decrease again.

- Temporal patterns by states
    - By Year

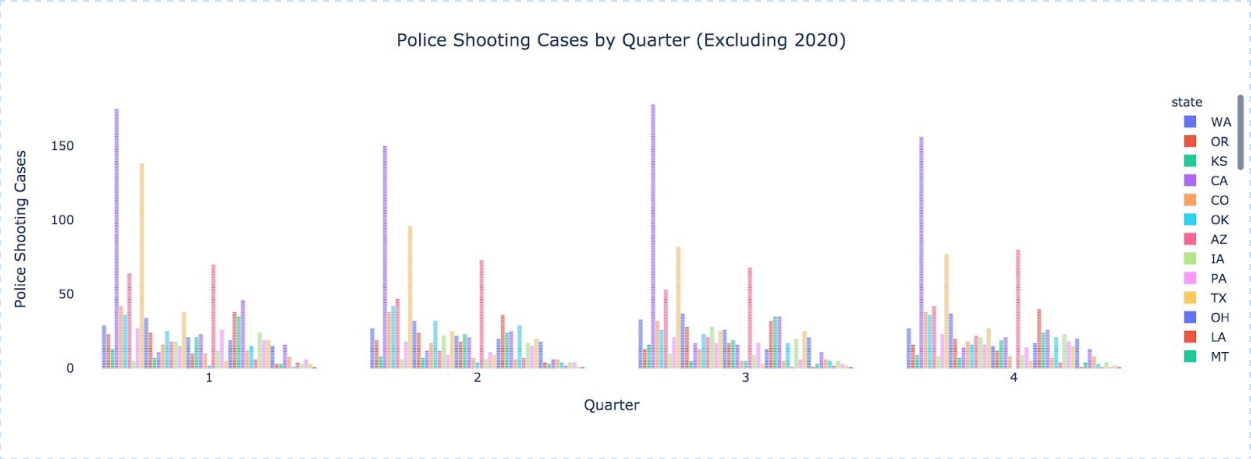


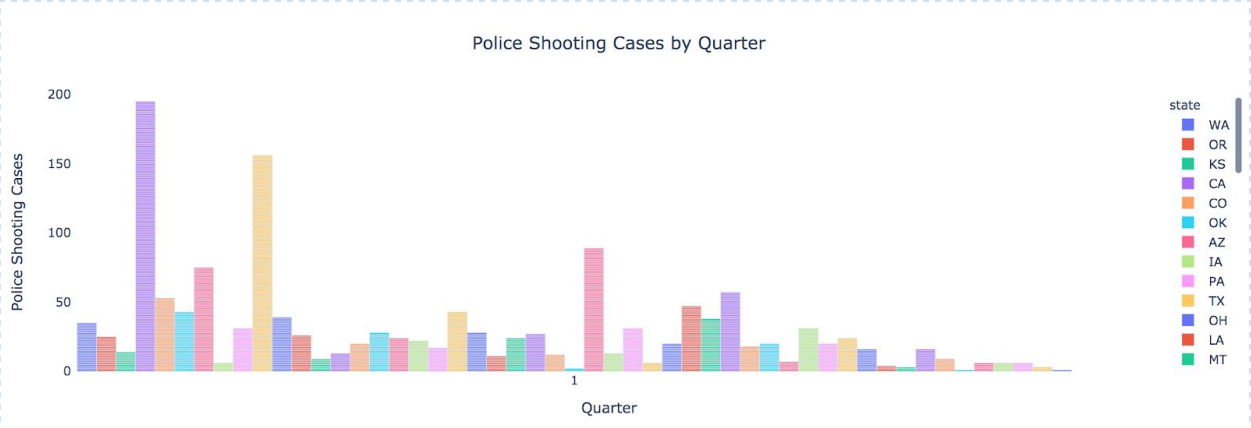


- By Quarter (Missing data for quarter 3 & 4)

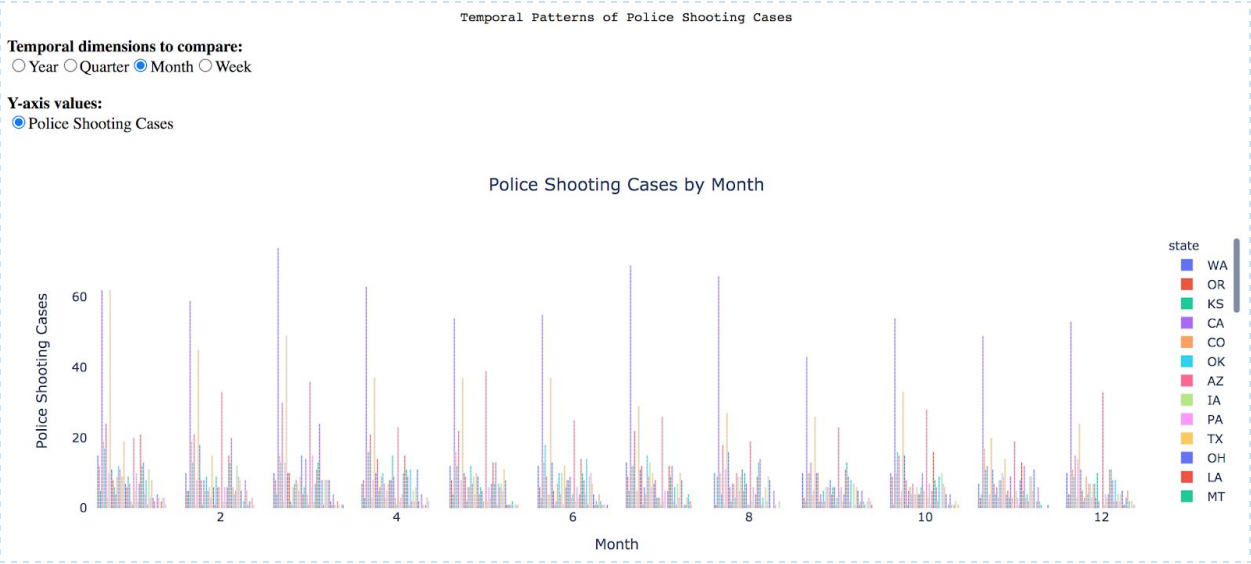


- By Quarter (Excluding 2020 data)

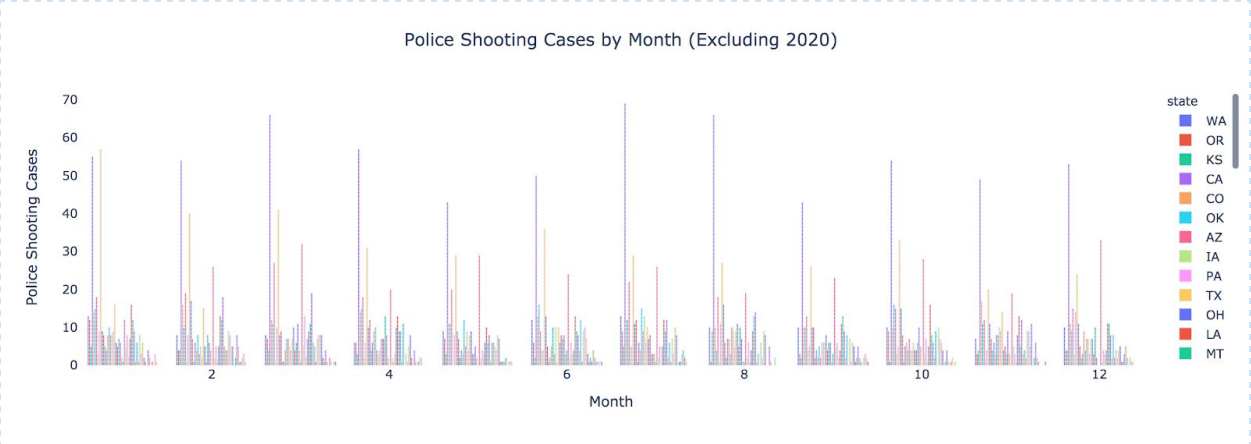




- By Month (Missing data from 2020.7-2020.12)



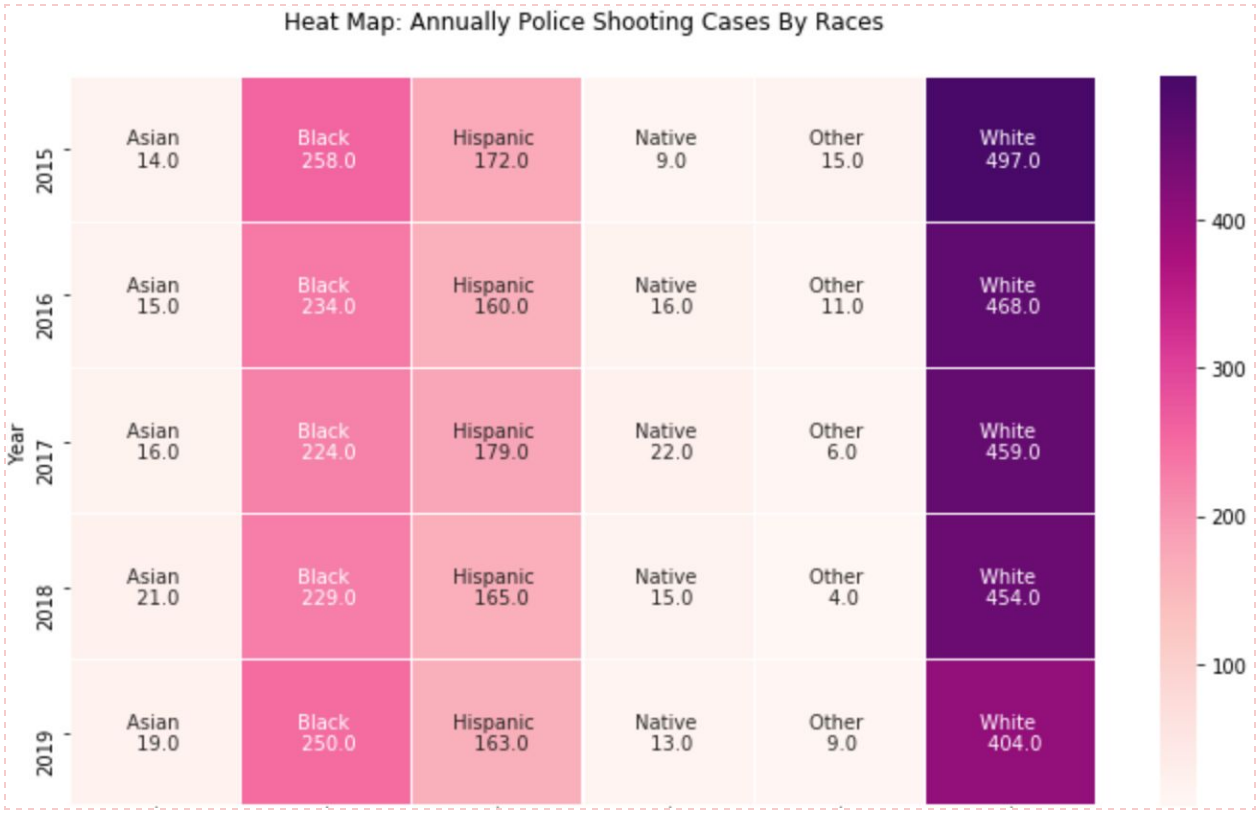
- By Month (Excluding 2020 data)



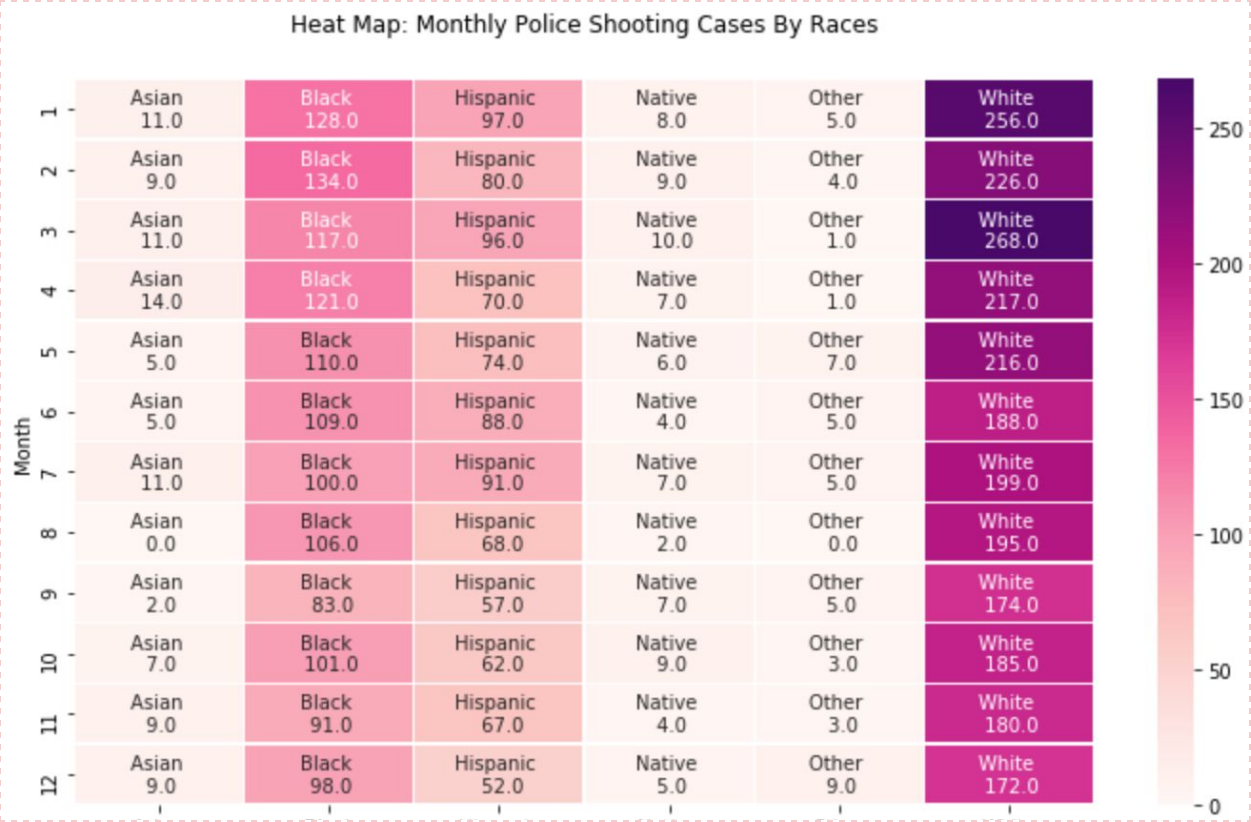
- By Week

Regardless of the different states, the overall temporal patterns, no matter if it is by years, months, quarter, or weeks, are consistent with the results we found previously. Comparing among states, CA, TX, FL took up the largest number of cases in each temporal dimension. This is consistent with the geographical results shown below.

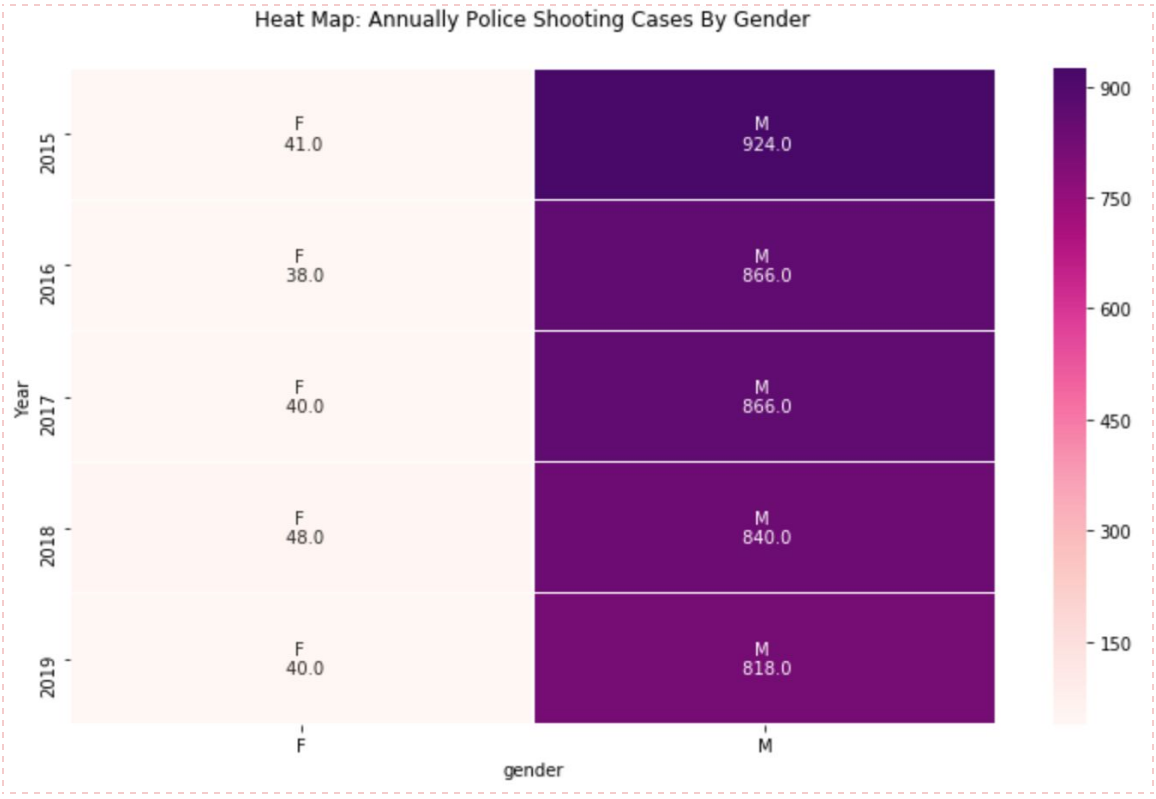
- Are there any interaction effects between certain demographic features and temporal features?



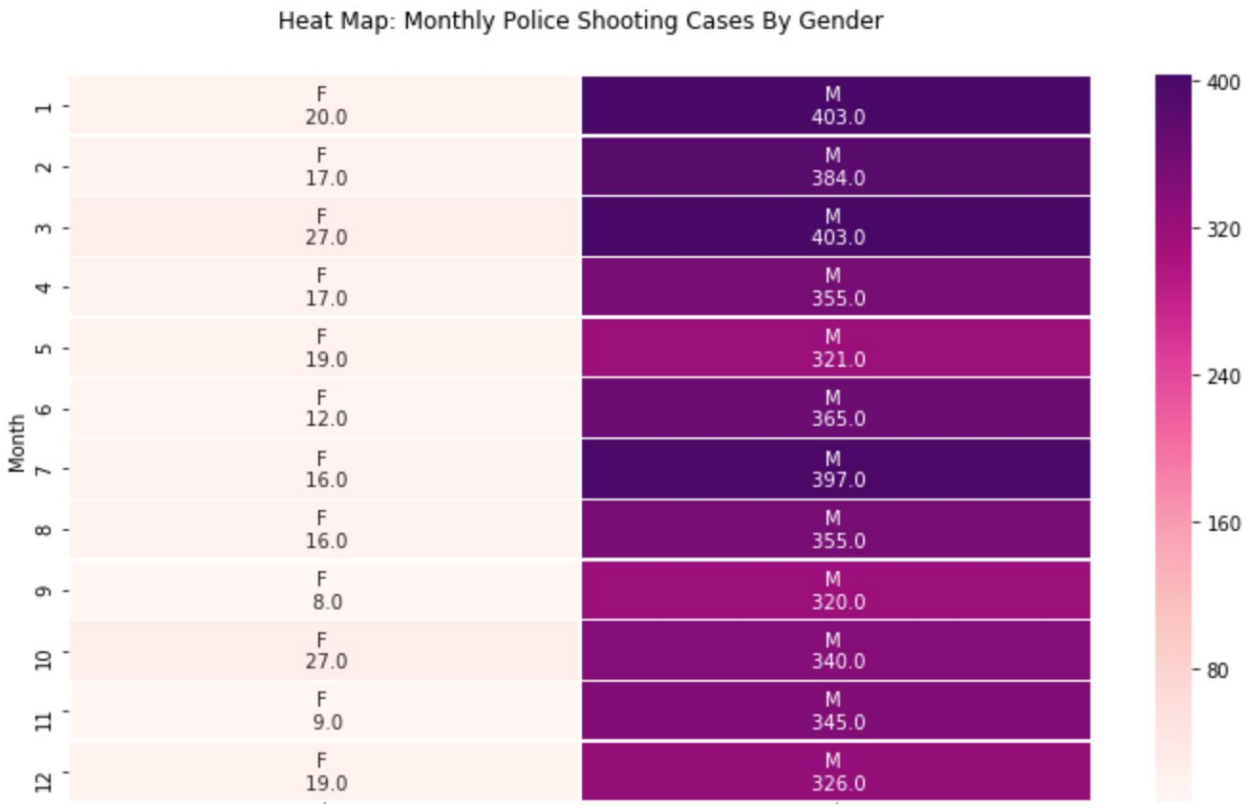
Across each year, we did not observe too many changes for Asian, Hispanic, Native, and Other races. Black people are associated with the largest police shooting cases in year 2015 and year 2019. Whereas the total shooting cases for white people decreased since 2015. Across different races, white people are associated with the highest number of shooting cases followed by black people and hispanic people.



A similar pattern could be observed by month in the heatmap above compared to the one by year. Overall, the total shooting cases associated with white, black, and hispanic are all decreasing across months from January to December.

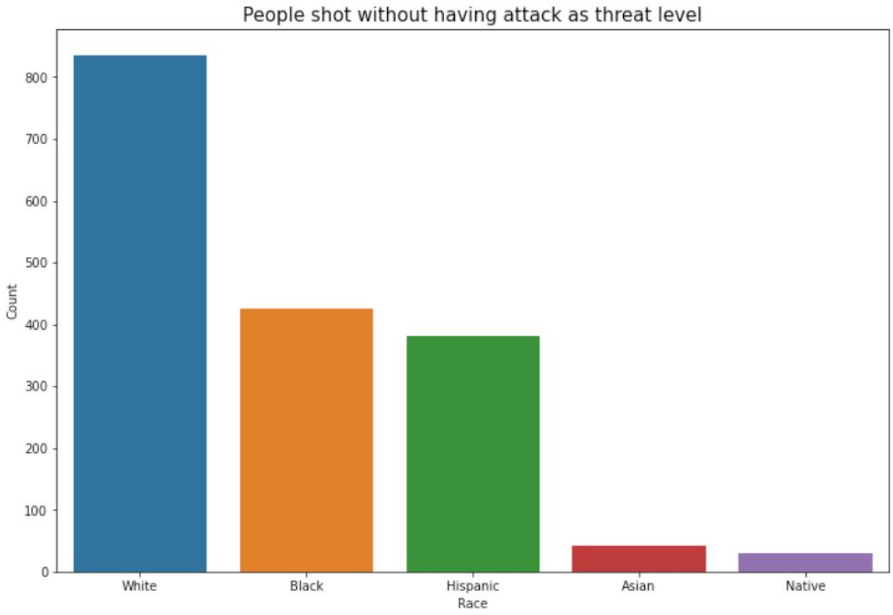


No significantly different changes were observed for the total shooting cases associated with females across years, whereas a decrease pattern could be observed for males across years. Overall, males are associated with the significantly larger number of police shooting cases than females.

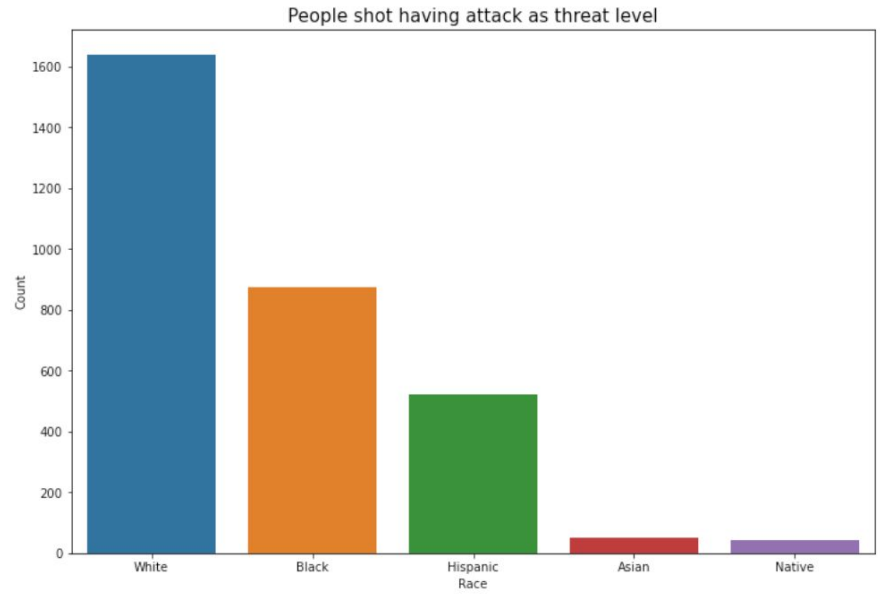


No significantly different changes were observed for the total shooting cases associated with females across months, whereas a decrease pattern could be observed for males across months. Overall, males are associated with the significant larger number of police shooting cases than females.

Question #5: Does police shoot people of certain races who do not have "threat level" marked as "attack" more often than the others?



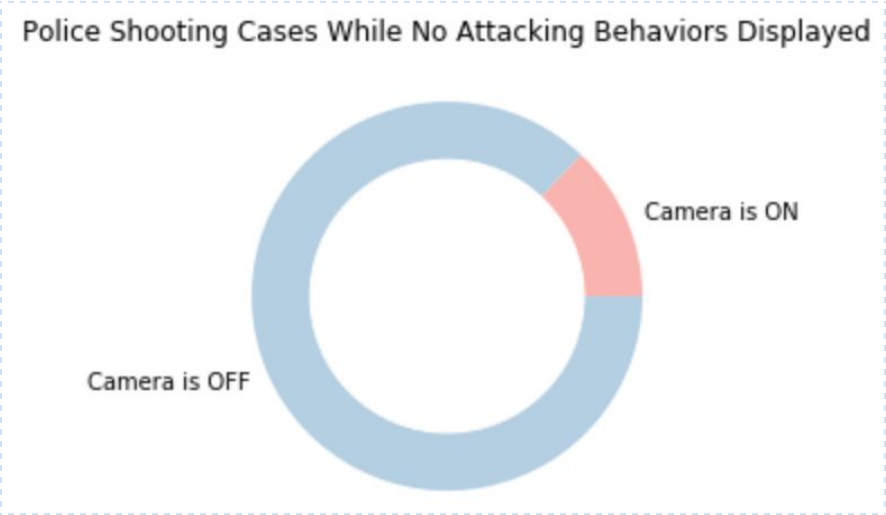
	Race	Count
0	White	836
1	Black	425
2	Hispanic	381
3	Asian	42
4	Native	30



	Race	Count
0	White	1640
1	Black	873
2	Hispanic	521
3	Asian	51
4	Native	44

From the table and bar chart above, we are able to compare the difference between people shot having attack as threat level, and people shot without having attack as threat level. Usually, people will be less likely to be shot when they are not considered as "attack". However, we can see that the proportion of hispanic people getting shot has increased from "attack" to "not attack".

Questions 6:



There is a significant difference between the proportion of police shooting cases while the camera is on versus off. We predict that police will care less about shooting people while the camera is off even the victims showed no attacking behaviors. The result is consistent with our prediction.

9) **Project log:** Describe how the final is done from beginning to the end: what is done and when

10.29 Team Orientation

11.10 Individual brainstorming about project topic and datasets

11.15 First group meeting: choose the project topic and discuss about the datasets

11.16 Choose the main software to use: Python, Plotly dash

11.22 Second group meeting: Finalize the dataset to use, and assign work to each member

11.23 Work on the project proposal and initial data exploration

11.27 Third group meeting: Came up with questions to ask the professor.

11.28 Work on the presentation powerpoint and practice the presentation individually

12.3 Fourth group meeting: Presentation practice

12.4 Work on the assigned part of project individually

10) **Teamwork**: Provide the names of all team members. Explain how the team works together, and describe what each member has done for the final product. Your report requires a signature from each of the team members. You can refuse to sign if you do not agree with what is written here. In this case, you can email Prof. Shen to explain.

Fiona Fei:

The whole project has been distributed evenly and our team works great together with each team member eager to put effort in this project. Everyone is very active during group meetings for sharing thoughts about the project. Everyone has completed the assigned parts well and in a timely manner although we are in different time zones.

Signature:

Edison Gu:

The team worked together through regular zoom meetings and group chats. Everyone is very responsive to messages and are able to communicate expectations clearly. Though 3 of us are located in the different time zones, we were able to work together smoothly and take equal parts in creating visualizations as well as writing the final report.

Signature:

Shay Sun:

All team members have evenly contributed to the project for the entire time. Each time's discussion and communication were efficient and clear. The project timeline was set up reasonably and each team member was always engaged, active, and willing to take responsibilities. Each of us took equal parts of questions to address.

Signature:

Shay Sun