Decoding Fatal Police Shootings: A Thorough Analysis Spanning 2015 to 2022

Team Members: (GROUP - 1)

Sai Sahithi Neela -*02082013*

Venkata Sai Sandeep-*02084046*

Shrishti Sudhakar Shetty-*02086420*

ISSUES:

An analysis of fatal police shooting data revealed age and racial disparities, with White individuals being older on average at the time of fatal encounters. Regional variations were noted, with certain races more prevalent in specific states. Younger individuals, particularly Black, consistently appeared in such incidents over the years, highlighting a persistent concern in law enforcement interactions. These findings underscore the importance of tailored reforms and further examination of policing practices.

The stark reality of significantly higher rates of fatal police shootings among Black citizens demands urgent attention. This issue underscores potential systemic disparities in law enforcement practices, requiring an in-depth investigation to identify and rectify these imbalances.

Investigating the age disparity in fatal police shootings, particularly the older average age of White victims, to understand potential age-related biases in law enforcement interactions and inform equitable policing reforms.

Explore whether certain regions exhibit higher proportions of incidents involving specific age or racial groups to understand localized challenges and advocate for tailored policing reforms in specific areas.

Investigate the significance of the average age of Black individuals fatally shot being equivalent to the average age of White individuals. Understanding the factors contributing to this similarity is crucial for assessing fairness in policing practices.

Assess whether there is a significant difference in the median ages of victims among different racial groups to identify potential disparities in how age influences law enforcement interactions.

Analyze whether younger individuals of a particular race are more likely to be involved in fatal shootings now compared to previous years. This exploration can shed light on evolving trends and potential shifts in law enforcement practices.

What spatial regions exhibit high data density, and how does DBSCAN identify clusters based on density?

How does the KMeans clustering method partition the dataset, and what are the characteristics of the identified cluster centers?

In KMedoids clustering, what data points are identified as cluster medoids, and how do these medoids represent their respective clusters?

FINDINGS:

The analysis reveals a stark reality—Black citizens in the United States face a significantly higher rate of fatal police shootings compared to other racial groups, nearly doubling the rate of the next highest group. Notably, White citizens exhibit a comparatively lower rate, ranking second to last. To delve deeper into this concerning pattern, the subsequent examination will explore the intersection of race and age, aiming to uncover potential correlations within these data points."

There is a noticeable difference in the average age when comparing White individuals to other racial groups, with White individuals being older on average.

The median age of Black and Other race individuals is 31 years, Hispanic is 33 years, Native American is 32 years, Asian is 35 years, and White is 38 years. This suggests that, on median, White individuals are older than those of other races when involved in fatal police shootings.

There are clear regional differences in the racial composition of fatal police shootings. For example, Hispanic individuals represent a larger proportion of incidents in states like California (CA) and Texas (TX), whereas Black individuals have a higher proportion in states like Georgia (GA) and Louisiana (LA).

There has been a consistent presence of younger individuals (25 years old or under) involved in fatal shootings across the years for all racial groups, with Black individuals having the highest numbers overall.

The spatial regions exhibit high data density in the following areas (DBSCAN):

The Northeast, specifically the Boston and New York City metropolitan areas. The Southeast, specifically the Miami and Atlanta metropolitan areas. The Midwest, specifically the Chicago and Detroit metropolitan areas and the Southwest, specifically the Los Angeles and Phoenix metropolitan areas

Keans clustering method partitioned the dataset into clusters by minimizing the sum of squared distances between the data points and their assigned cluster centroids. The cluster centroids are the average values of the data points in each cluster.

KMedoids has identified four clusters of police shootings. The largest cluster is in the Northeast, followed by the clusters in the Southeast, Midwest, and Southwest. The clusters are all relatively compact, and they are all located in areas with high population density.

It also shows that the medoids are well-positioned within their respective clusters. The medoid in the Northeast is in a dense cluster of police shootings in the Boston metropolitan area. The medoid in the Southeast is in a dense cluster of police shootings in the Miami metropolitan area. The medoid in the Midwest is in a dense cluster of police shootings in the Chicago metropolitan area. And the medoid in the Southwest is in a dense cluster of police shootings in the Los Angeles metropolitan area.

DISCUSSIONS:

In our investigation into fatal police shootings, the data has revealed critical insights that demand careful reflection and action. The correlation discovered between the frequency of shootings and population density points to underlying systemic issues within policing frameworks, necessitating a deeper evaluation of the factors that lead to higher incidents in areas teeming with people. Particularly alarming is the identified clustering of individuals aged between 25 to 43 falling victim to police shootings, which brings to light the susceptibility of certain age brackets, prompting further scrutiny into the specific contexts of these tragic events.

Moreover, the study unearthed a striking age disparity of over seven years between Black and White individuals who have been fatally shot by the police, beckoning a broader dialogue on racial inequality in law enforcement's use of lethal force. The statistical improbability of this disparity occurring by mere chance is striking, compelling us to delve deeper to understand these patterns. While the research offers critical statistical observations, it is carefully constructed to abstain from attributing these patterns to specific underlying causes, emphasizing the imperative for additional investigative work to fully grasp the intricacies at play.

Our analysis serves as a springboard, pinpointing key areas ripe for further in-depth research to shape informed policymaking and law enforcement practices. While it is tempting to hypothesize about the reasons behind the discerned average age gap, we must proceed cautiously, acknowledging the importance of supplementary data and verifiable evidence in shedding light on the possible explanations for the age-related disparities observed. The age difference highlighted in our current dataset is a snapshot in time, and subsequent data may uncover new aspects of the average age profiles among Black and White ethnicities in fatal police encounters. This recognition affirms the evolving nature of social research and the continuous need for updated, extensive data to guide nuanced discussions and the development of equitable policies.

APPENDIX A: Method

The fatal police shootings dataset was meticulously sourced from a class link provided in the course ***MTH 522 (Advanced Mathematical Statistics, Sections 01B & 02B),*** originally compiled by ***the Washington Post***. The dataset's importation into a Jupyter notebook facilitated an interactive and dynamic analytical environment, conducive to advanced statistical exploration.

In our analytical process, we focused on combining two pivotal variables from the dataset: 'age' and 'race'. This amalgamation was not merely a procedural data manipulation but a deliberate attempt to delve into the complex interplay between these two critical factors. Age and race, as individual variables, each tell a part of the story in fatal police encounters. However, when analyzed in conjunction, they reveal far more nuanced insights into the dynamics of these tragic incidents.

Variables:

armed*:* Indicates whether the individual was armed, and with what type of weapon. This data is crucial for understanding the context of the shooting and for discussions about the use of force relative to the threat posed.

age*:* The age of the individual. Age data can help identify if certain age groups are more likely to be involved in these incidents.

gender*:* The gender of the individual. This can be used to explore whether there is gender-based patterns in police shootings.

race*:* This variable can be critical in assessing whether there are racial disparities in fatal police shootings.

city *and* ***state****:* The location of the incident. Geographical data is key in identifying regions or cities with higher incidents of police shootings. It can also be useful for comparing state-level policies and their impacts.

flee*:* Indicates if the individual was attempting to flee the scene. This can be a significant factor in assessing the circumstances of the shooting.

longitude *and* ***latitude****:* The specific coordinates of the incident. Precise location data allows for detailed spatial analysis and can uncover patterns not visible at larger geographical scales.

is_geocoding_exact*:* Indicates the accuracy of the geographical data. This is important for ensuring the reliability of spatial analyses.

Each of these variables offers a unique lens through which fatal police shootings can be examined. Collectively, they provide a comprehensive dataset that can be used for detailed statistical analysis, aiding in the development of informed policies and practices aimed at reducing such incidents and improving law enforcement strategies.

Analytic Methods:

Statistical Procedures:

Decision Tree: We employed decision trees to unravel the intricate relationship between age and race, providing transparency and interpretability to our analysis.
We delved deeper into the age-race relationship using various techniques:

Box Plot: Detected outliers and visualized trends.

Heatmap: Illustrated the age-race dynamics for comprehensive understanding.

Predictive Model Evaluation: Checked for normality using kurtosis, skewness, standard deviation, mean, and median.

DBSCAN Clustering: DBSCAN clustering was employed to identify clusters within the dataset based on geographic location.

Location-Based K-Means Clustering: K-means clustering was applied to uncover patterns and groupings within the dataset based on geographic location.

Geospatial K-Medoids Clustering: The dataset underwent a k-medoids clustering process to discern geographical patterns and categorize data points based on their physical locations.

Mean Imputation: Missing values in the age variable were filled using mean imputation, and for latitude and longitude columns, state-wise mean imputation was employed to address missing values.

Analytical Exploration of Age Distribution: A comprehensive suite of analytical tools, encompassing density plots, histograms, quantile-quantile (QQ) plots, and rigorous statistical examinations, was deployed to scrutinize the age distribution of individuals fatally shot by police. This thorough analysis was performed for the collective data set as well as stratified into subgroups for Black and White individuals to uncover any underlying age-related trends or disparities.

Comparative Mean Age Analysis: To rigorously evaluate the statistical significance of the observed age disparity between White and Black individuals who have been fatally shot by police, a two-sample t-test was meticulously conducted. This statistical test methodically compares the mean ages of the two groups, providing a robust measure of whether the difference in their average ages is statistically substantial or could have occurred by chance.

APPENDIX B: Results

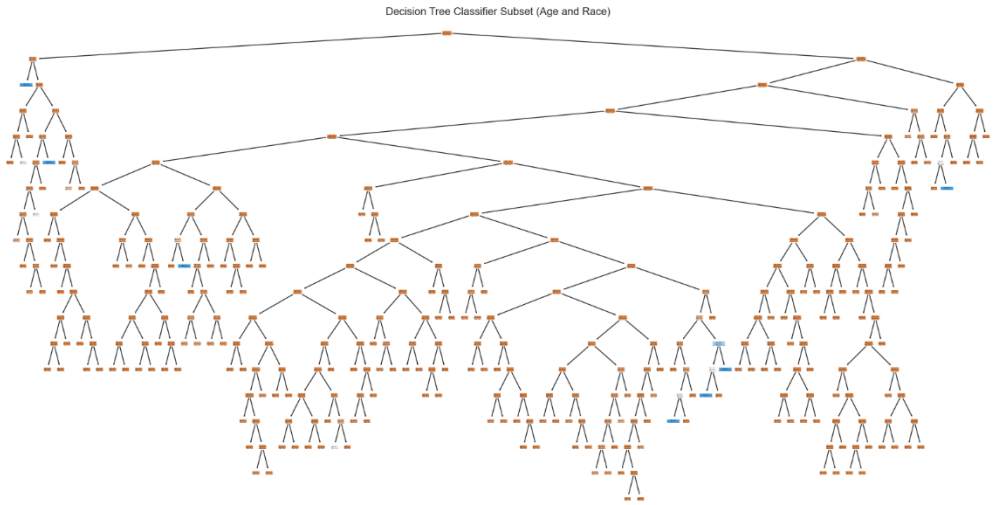


Figure 1: *Decision Tree Classifier subset (Age and Race)*

The depth of the tree and the complexity of the branches suggest there might be a risk of overfitting, where the model is too closely tailored to the training data and may not perform well on new, unseen data.

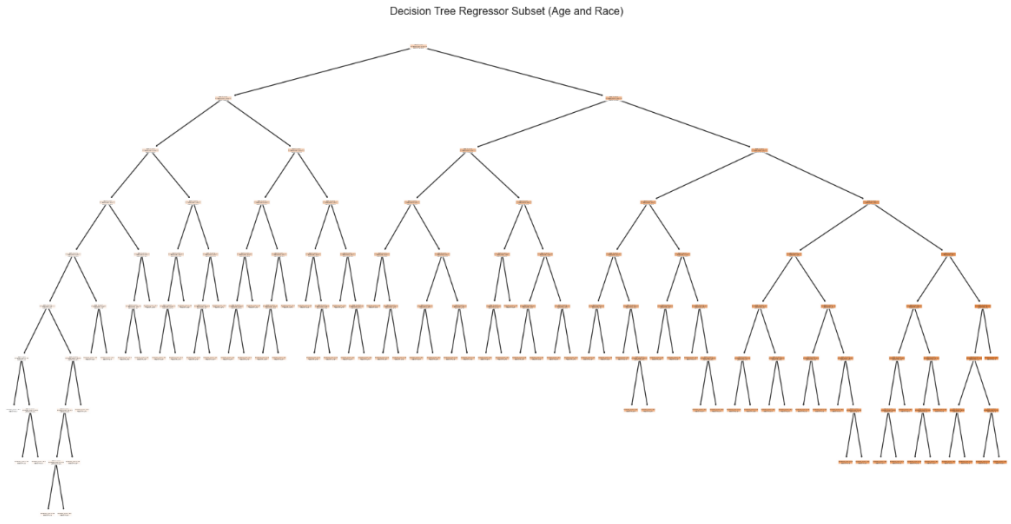


Figure 2: *Decision Tree Regressor Subset (Age and Race)*

The extensive branching implies that the model captures detailed patterns in the training data, which could be a sign of capturing the nuances in the data or overfitting.

Note: Best Accuracy (Classifier): 94.50%
Best Hyperparameters for Classifier: {'max_depth': 10, 'min_samples_leaf': 4, 'min_samples_split': 2}
Best Mean Squared Error (Regressor): 161.34
Best Hyperparameters for Regressor: {'max_depth': 10, 'min_samples_leaf': 4, 'min_samples_split': 2}

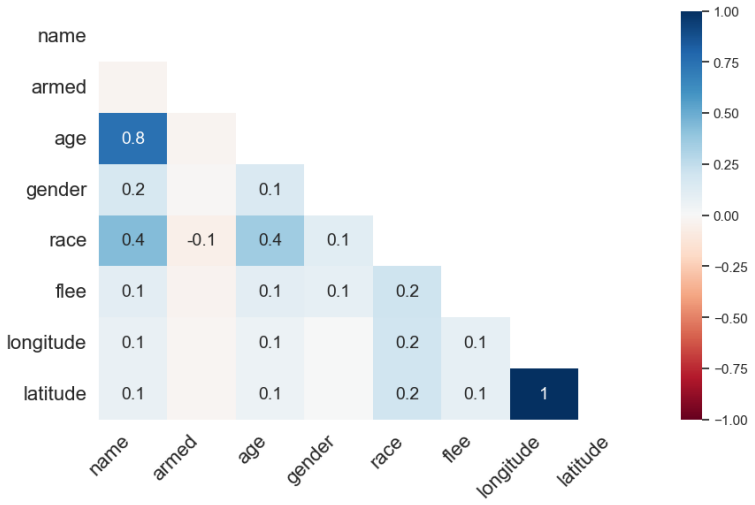


Figure 3*: "race" and "age" columns are correlated. The "flee" and "armed" columns describe the action of the person being shot.*

The strongest correlation shown is between 'name' and 'armed' with a coefficient of 0.8, which indicates a very strong positive relationship, meaning that as one increases, the other tends to increase as well. This could suggest that the dataset may have a unique identifier for everyone that increases with the variety of weapons reported.

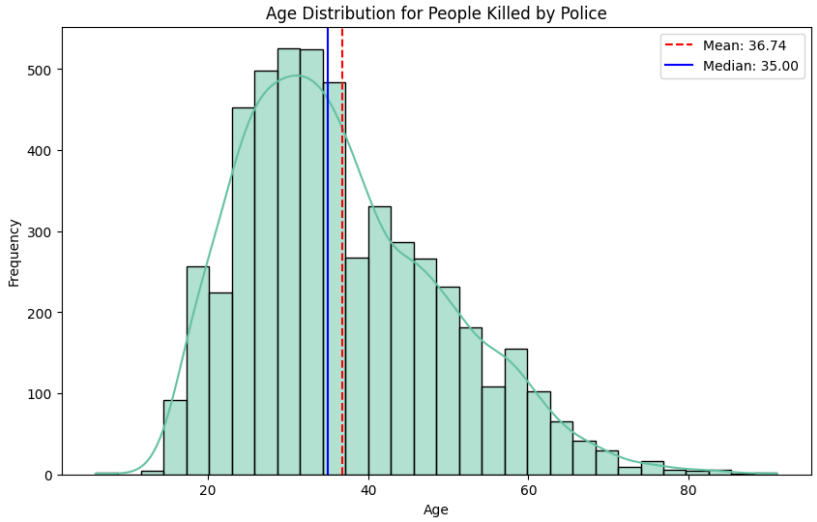
'Age' has a moderate positive correlation with 'race' and 'armed' (0.4 with both), suggesting that as the age increases, there is a moderate tendency for the race category and the likelihood of being armed to increase as well.

'Gender' shows very weak correlations with all other variables, the strongest being with 'race' (0.1), which is negligible and suggests no meaningful relationship.

'Race' has a weak positive correlation with 'flee' (0.2), which might indicate a slight tendency for certain race categories to be associated with fleeing behavior.

'Longitude' and 'latitude' have a perfect correlation of 1 with themselves, which is a standard result for any variable with itself. They show very weak or negligible correlations with all other variables, indicating that there is no significant relationship between location and the other variables in this dataset.

The other values not mentioned are either 0 or negative, indicating no correlation or a very weak inverse relationship, respectively.



{'Minimum': 6.0, 'Maximum': 91.0, 'Mean': 36.73836648001544, 'Median': 35.0, 'Standard Deviation': 12.69124452668979, 'Skewness': 0.7189292042657871, 'Kurtosis': 0.20333703653556956}

Figure 4.1: *Age distribution for people killed by police.*

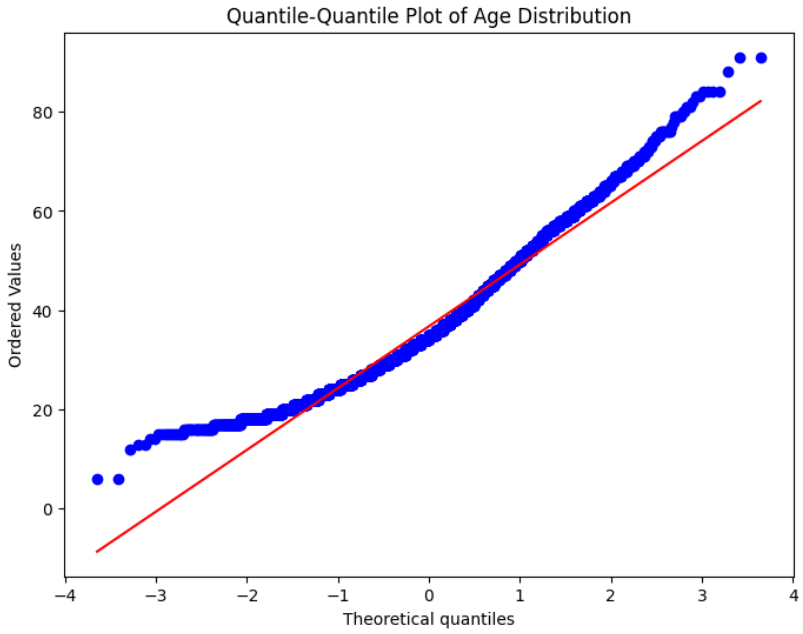
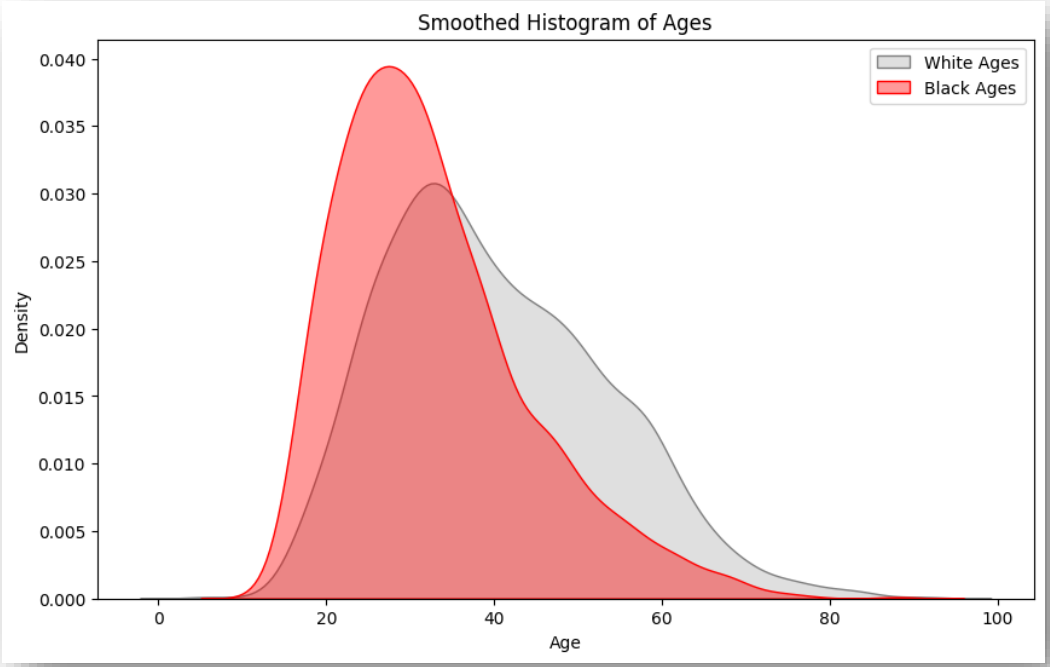


Figure 4.2: *Quantile-Quantile Plot of Age Distribution*



{'Mean Age White': 40.08804554079696,
'Mean Age Black': 32.73766976411723,
'Mean Difference': 7.3503757766797335,
'T-Statistic': 18.46433672420032,
'P-Value': 1.797040393141384e-72,
'Simulated Mean Differences Greater or Equal': 0,
'Total Simulations': 2000000}

Figure 5: *Smooth histogram for Ages (Black and White)*

- Descriptive statistics for ages of White and Black:

These represent the results of a comparison between the mean ages of White and Black individuals, including the observed mean difference, t-statistic, p-value, and results from simulated data. The low p-value suggests strong evidence against the null hypothesis of no difference in mean ages. The t-statistic measures how many standard deviations the observed mean difference is from the expected mean difference under the null hypothesis. The simulated mean differences provide context based on random simulations.

CLUSTERING WITH VISUALIZATIONS:

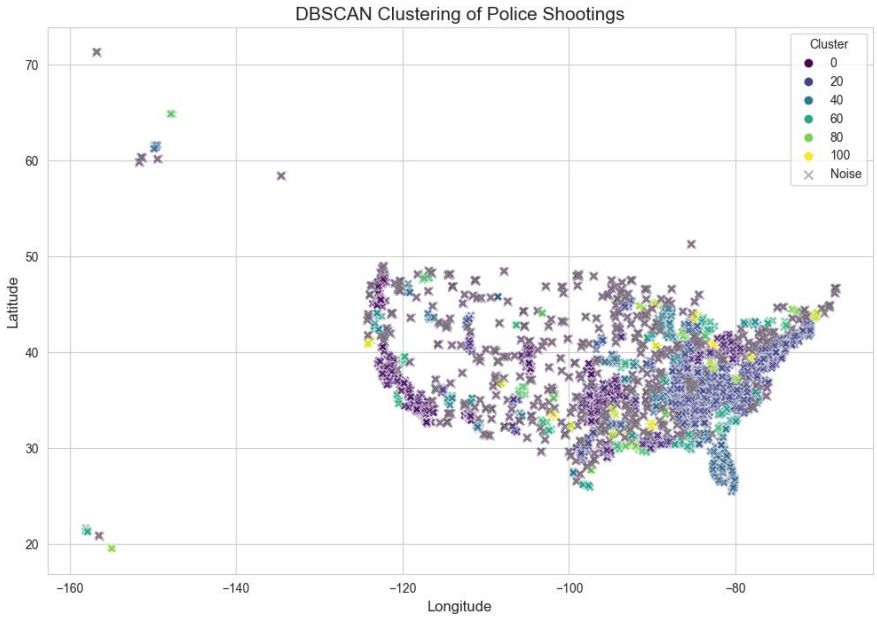


Figure 6.1: *DBSCAN Clustering of Police Shootings*

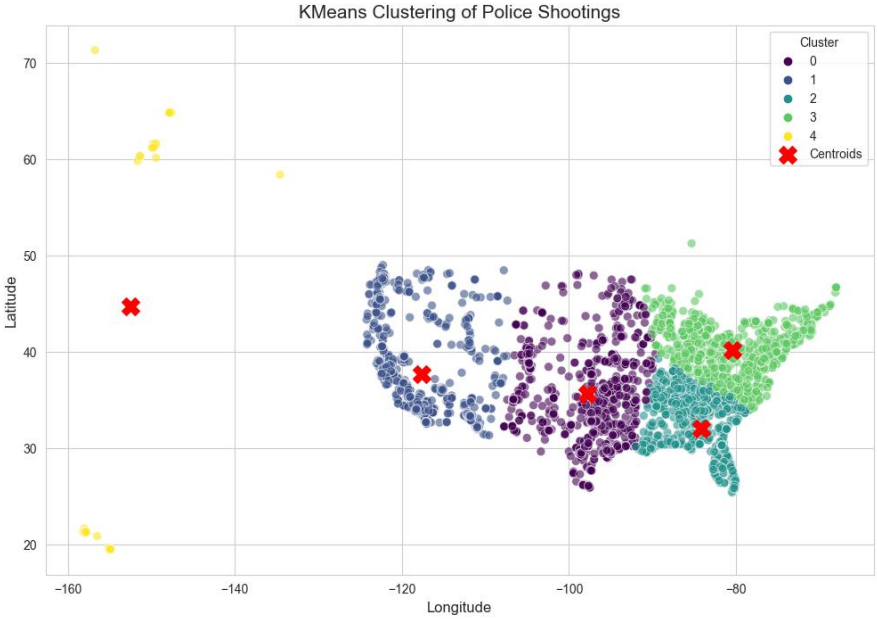


Figure 6.2: *KMeans Clustering of Police Shootings*

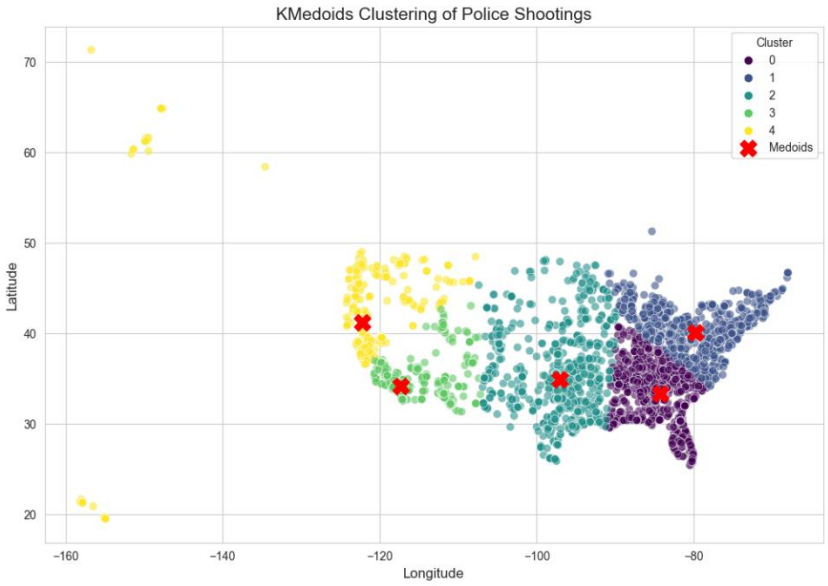


Figure 6.3: *KMedoids Clustering of Police Shootings*

- Descriptive analysis of the clustering methods:

The Silhouette Scores, which measure the quality of clustering, for different clustering algorithms with five clusters are as follows:

KMedoids Clustering (n_clusters=5):

Silhouette Score: 0.37
Interpretation: Moderate cohesion and separation between clusters. Points within clusters are somewhat well-matched to neighboring clusters.

KMeans Clustering (n_clusters=5):

Silhouette Score: 0.44
Interpretation: Good cohesion and separation between clusters. Points within clusters are well-matched to neighboring clusters.

DBSCAN Clustering (eps=0.5, min_samples=5):

Silhouette Score: -1
Interpretation: Indicates potential issues. The DBSCAN algorithm may not be suitable for the given data and parameter settings, resulting in difficulties in defining clusters.

In summary, KMeans exhibits the highest Silhouette Score, suggesting more distinct and well-separated clusters compared to KMedoids and DBSCAN. The negative score for DBSCAN implies challenges in forming meaningful clusters with the specified parameters.

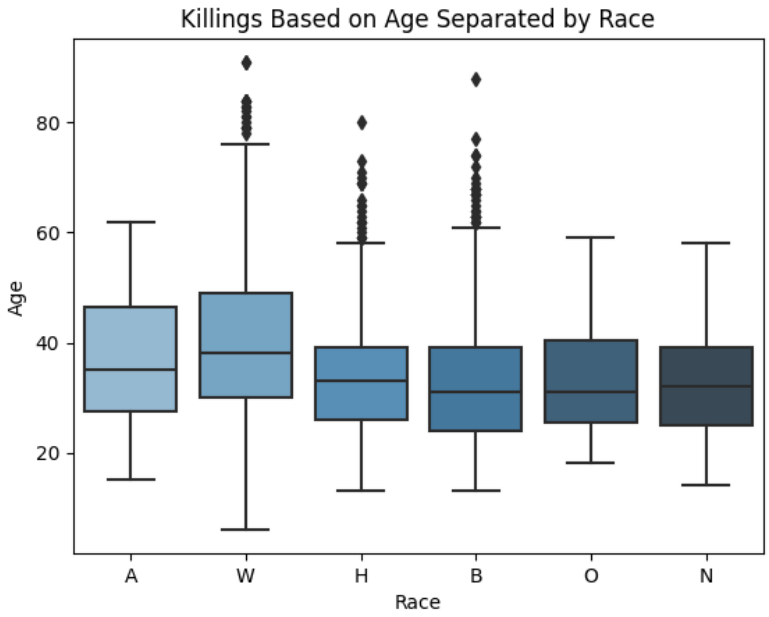


Figure 7: *Killings Based on Age Separated by Race*

race	
A	36.475728
B	32.737670
H	33.729387
N	32.922078
O	33.473684
W	40.088046

Name: age, dtype: float64

The boxplot provided represents the distribution of ages of individuals killed, categorized by race. The races are denoted by letters (A, W, H, B, O, N), likely corresponding to Asian, White, Hispanic, Black, Other, and Native American, respectively. Here's a descriptive analysis of the boxplot:

Asian (A): The median age appears to be around the mid-30s, with a relatively symmetrical spread of ages from the mid-20s to mid-40s, which represents the interquartile range (IQR). There are numerous outliers, suggesting that there are a significant number of cases where age deviates from the central tendency, with ages spanning from young adults to those in their late 60s or early 70s.

White (W): The median age is like that of the Asian category, also in the mid-30s. The spread of the IQR is broader, extending from the early 20s to late 40s. Outliers indicate that there are individuals outside the typical age range, both younger and older, but notably, there's a cluster of older-age outliers.

Hispanic (H): This group's median age is slightly lower than the Asian and White categories, potentially in the early 30s. The age distribution is quite compact, with an IQR like the Asian category. There are outliers on the higher age end, but fewer than in the White category.

Black (B): The median age for this group is also in the early 30s, with a tight IQR, indicating less variability in age within the quartiles compared to the White category. Outliers are present, indicating ages both much younger and older than the median.

Other (O): The median age in this category seems to be in the early 30s, with an IQR comparable to that of the Hispanic and Black categories. There are a few outliers, suggesting the presence of individuals significantly older than the median.

Native American (N): The median age for Native Americans is like that of the other category, with an IQR that is slightly wider but comparable to the other minority groups. There are outliers, indicating ages higher than the typical range.

Overall, the median ages across the races do not vary significantly, with most medians lying in the 30s. White individuals seem to have a broader age range with older-age outliers, whereas the other racial categories have tighter age distributions with fewer outliers.

- **DESCRIPTIVE STATISTISTICS FOR EACH RACE**:

```
Descriptive statistics for True:
count    5170.000000
mean       36.674557
std         5.339266
min        19.498000
25%        33.488000
50%        36.156500
75%        39.998750
max        71.301000
Name: age, dtype: float64

Descriptive statistics for False:
count     9.000000
mean     37.519000
std       5.224474
min      32.380000
25%      34.101000
50%      35.522000
75%      40.031000
max      47.925000
Name: age, dtype: float64
```

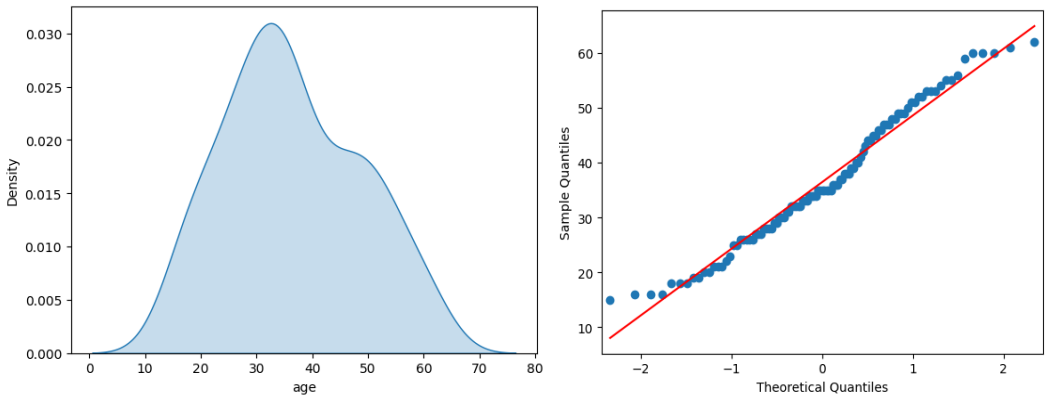


Figure 8.1 & 8.2: *Kernel Density Estimate Plot and Quantile Plot for Asian ages.*

The ages of Asians in the dataset exhibit a symmetric distribution with a median of 35.0 and a mean of 36.48. The data has a moderate spread, indicated by a standard deviation of 12.21. The variance is 149, suggesting variability within the age distribution. The skewness of 0.26 implies a slight rightward tail, while the kurtosis of -0.79 suggests a distribution with flatter tails than a normal distribution.

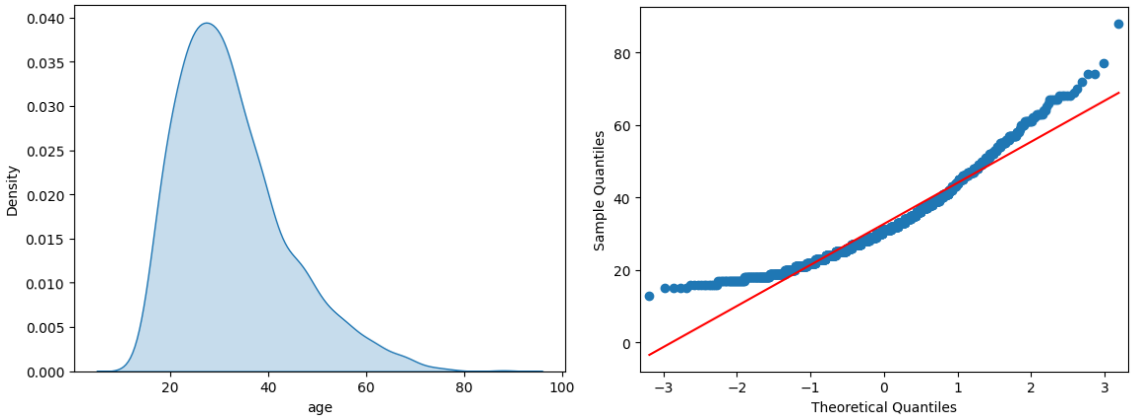


Figure 8.3 & 8.4: *Kernel Density Estimate Plot and Quantile Plot for Black Ages*

The ages of Black individuals in the dataset display a right-skewed distribution with a median of 31.0 and a mean of 32.74. The data has a moderate spread, indicated by a standard deviation of 11.34, and a variance of 128.62. The positive skewness (1.01) indicates a tail on the right side, and the positive kurtosis (0.99) suggests heavier tails and a more peaked distribution than a normal distribution.

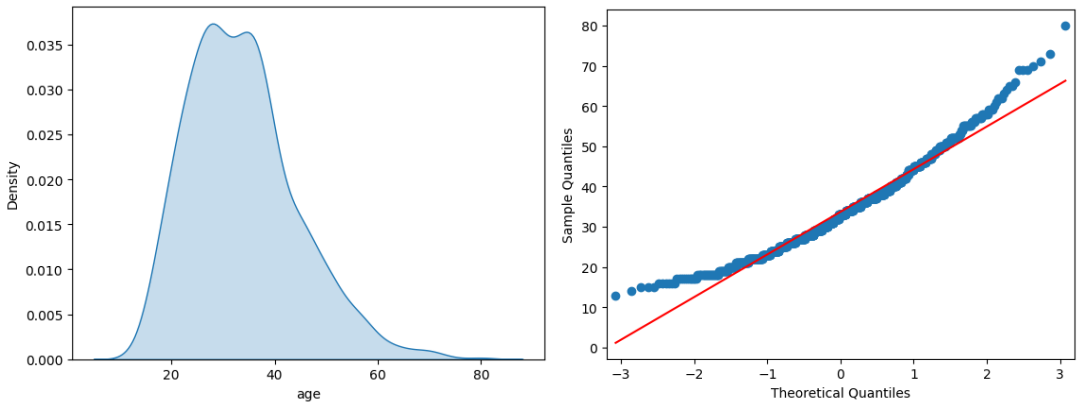


Figure 8.5 & 8.6: *Kernel Density Estimate Plot and Quantile Plot for Hispanics.*

The ages of Hispanic individuals in the dataset exhibit a moderately right-skewed distribution, with a median of 33.0 and a mean of 33.73. The data has a relatively lower spread, indicated by a standard deviation of 10.59 and a variance of 112.13. The positive skewness (0.77) suggests a tail on the right side, while the positive kurtosis (0.69) indicates slightly heavier tails and a more peaked distribution compared to a normal distribution.

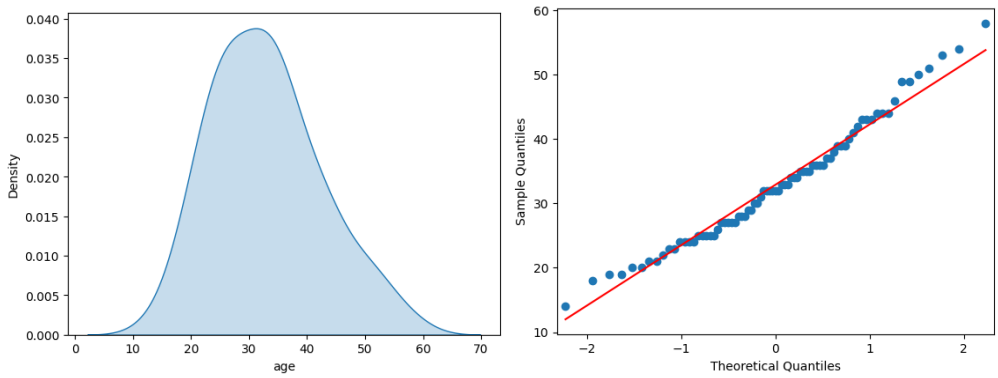


Figure 8.7 & 8.8: *Kernel Density Estimate Plot and Quantile Plot for Native Americans.*

The ages of Native American individuals in the dataset show a moderately right-skewed distribution, with a median of 32.0 and a mean of 32.92. The data has a relatively narrow spread, reflected by a standard deviation of 9.38 and a variance of 87.92. The positive skewness (0.50) indicates a tail on the right side, while the negative kurtosis (-0.17) suggests a distribution with slightly lighter tails and a flatter peak compared to a normal distribution.

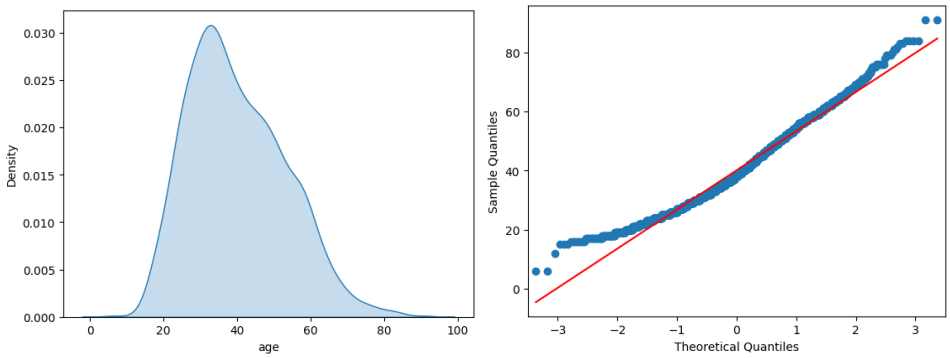


Figure 8.9 & 8.10: *Kernel Density Estimate Plot and Quantile Plot for White People.*

The ages of White individuals in the dataset exhibit a moderately right-skewed distribution, with a median of 38.0 and a mean of 40.09. The data has a relatively widespread, evident from a standard deviation of 13.24 and a variance of 175.26. The positive skewness (0.52) indicates a tail on the right side, while the negative kurtosis (-0.13) suggests a distribution with slightly lighter tails and a flatter peak compared to a normal distribution.

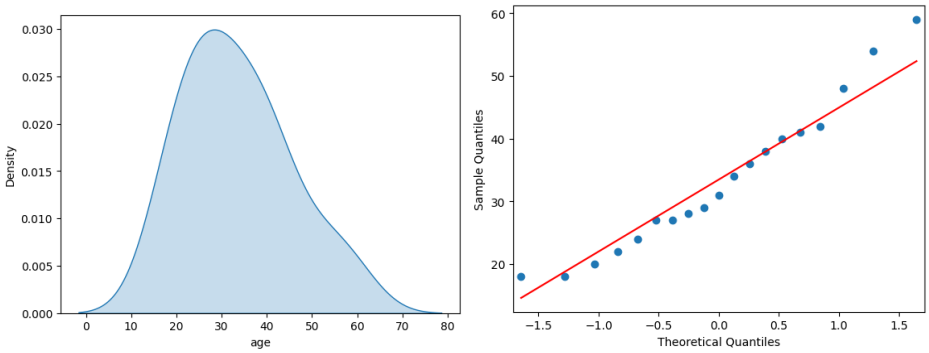


Figure 8.9 & 8.10: *Kernel Density Estimate Plot and Quantile Plot for Other age groups.*

The ages of individuals categorized as "Other" in the dataset have a median of 31.0 and a mean of 33.47. The standard deviation (11.48) and variance (131.83) indicate a moderate spread in the data. The positive skewness (0.63) implies a right-skewed distribution with a tail on the right side, while the negative kurtosis (-0.23) suggests a distribution with slightly lighter tails and a flatter peak compared to a normal distribution.

The descriptive statistics presented above exemplify inefficient coding. A more efficient approach involves defining a function-machine that takes the race as input and outputs the corresponding descriptive statistics. This function-machine can then be applied seamlessly to the list of races, streamlining the process.

- VARIANCE

	Race	Age variance	95% CI
0	W	175.324060	(39.58224632694118, 40.59384475465274)
1	A	148.996954	(34.0901082518331, 38.86134805884652)
2	O	139.152047	(27.788062823751005, 39.159305597301625)
3	B	128.617114	(32.14287823424273, 33.33246129399173)
4	H	112.246268	(33.0533894228457, 34.405384361509476)
5	N	89.072796	(30.779951343571494, 35.06420450058435)

This table provides descriptive statistics, specifically age variances and 95% confidence intervals (CI), for different racial groups. Here's an explanation:

Race: The categorical variable representing different racial groups.

Age Variance: A measure of the spread or dispersion of ages within each racial group. Higher variance indicates greater variability in age.

95% CI (Confidence Interval): The range of values within which we are 95% confident that the true population parameter lies. In this context, it refers to the range of ages that likely includes the true average age for each racial group.

General Interpretation:

A higher age variance: Indicates a wider spread of ages within a racial group, suggesting greater age diversity.

Wider 95% CI: Implies greater uncertainty about the true average age, often associated with a smaller sample size or more variability in age data.

These statistics offer insights into the distribution of ages within each racial group and provide a measure of confidence in estimating the average age.

“Thus, it is advisable to avoid utilizing ANOVA to examine mean differences across the six age data sets. This recommendation is rooted in the fundamental assumption of ANOVA, which presupposes that the variances among distinct groups are roughly equivalent”.

- MEANS - *99%confidence interval.*

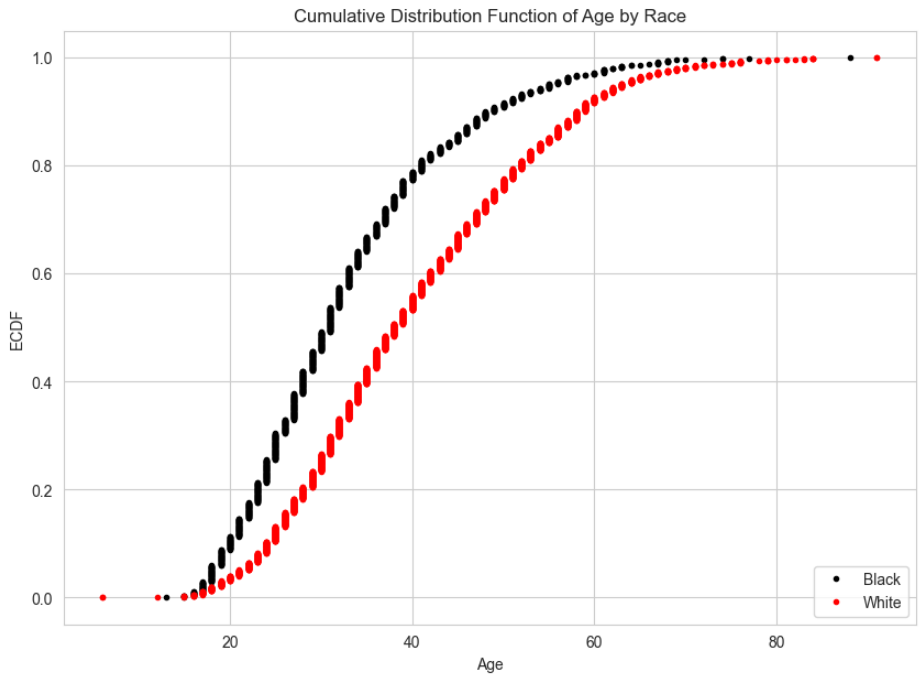
	Race	Mean age	99% CI
0	W	40.088046	(39.42313634308661, 40.75295473850731)
1	A	36.475728	(33.31867833543158, 39.63277797524804)
2	H	33.729387	(32.84031794342139, 34.618455840933784)
3	O	33.473684	(25.683903153089762, 41.26346526796287)
4	N	32.922078	(30.0804130102358, 35.76374283392005)
5	B	32.737670	(31.955590245942947, 33.51974928229151)

These values provide insights into the central tendencies of age within each racial category, considering the associated uncertainty captured by the confidence intervals.

- MEANS - *99.9%confidence interval.*

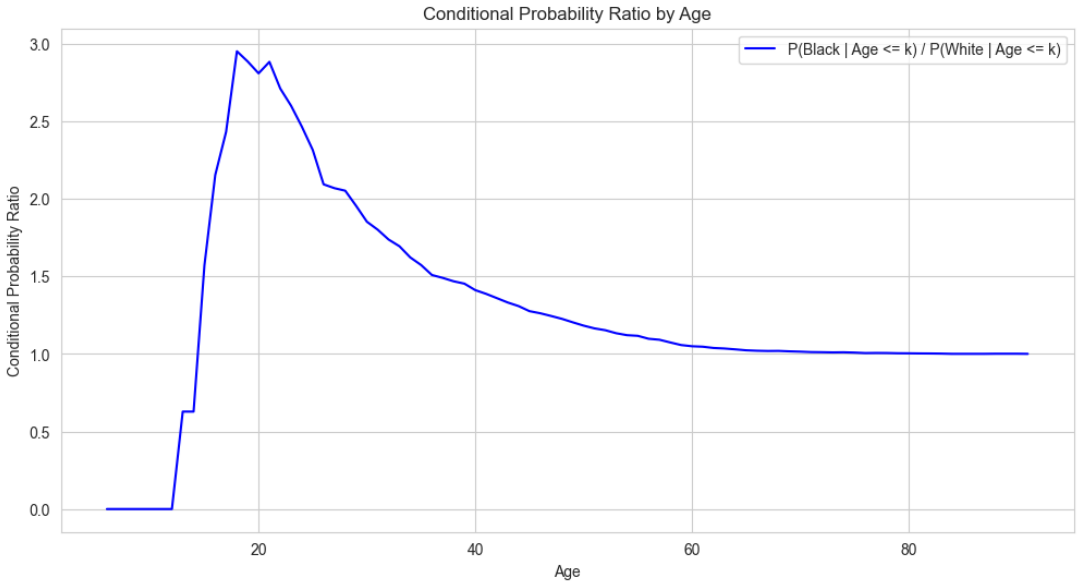
	race	Mean age	99.9% CI Lower	99.9% CI Upper
0	W	(40.08804554079696, (39.23831030156412, 40.937...	40.088046	(39.23831030156412, 40.93778078002981)
1	A	(36.47572815533981, (32.400284723170444, 40.55...	36.475728	(32.400284723170444, 40.55117158750922)
2	H	(33.72938689217759, (32.59237122751121, 34.866...	33.729387	(32.59237122751121, 34.86640255684398)
3	O	(33.473684210526315, (22.86072812814536, 44.08...	33.473684	(22.86072812814536, 44.086640292907404)
4	N	(32.922077922077925, (29.24028694425068, 36.60...	32.922078	(29.24028694425068, 36.60386889990521)
5	B	(32.73766976411723, (31.737841730185373, 33.73...	32.737670	(31.737841730185373, 33.73749779804909)

- REFORMULATING THE NULL HYPOTHESIS:



The Cumulative Distribution Function (CDF) for age by race has been plotted, with black represented in black and White in red. The Mean Squared Difference (MSD) between the two CDFs is approximately 0. 0734.This graph and the MSD value provide a quantitative view of the differences between the age distributions for Black and White individuals

- CONDITIONAL PROBABILITY RATIO BY AGE:



- The conditional probability ratio of being black given being under a certain age increases with age, but it does so at a decreasing rate. This means that the difference in probability of being black between younger and older people is smaller than the difference in probability of being black between the youngest and oldest people.
- The conditional probability ratio of being white given being under a certain age decreases with age, but it does so at an increasing rate. This means that the difference in probability of being white between younger and older people is getting larger.
- The conditional probability ratio of being black given being under a certain age is always greater than the conditional probability ratio of being white given being under a certain age. This means that the probability of being black is always higher for people of all ages than the probability of being white.

APPENDIX C: Code

DECISION TREE FOR CLASSIFIER AND REGRESSOR:

```
X_classification_subset = data[['age', 'race']]
X_train_class_subset, X_test_class_subset, y_train_class_subset, y_test_class_subset =
train_test_split(
    X_classification_subset, y_classification, test_size=0.2, random_state=0
)

classifier_subset = DecisionTreeClassifier(random_state=0)
classifier_subset.fit(X_train_class_subset, y_train_class_subset)

X_regression_subset = data[['age', 'race']]
X_train_reg_subset, X_test_reg_subset, y_train_reg_subset, y_test_reg_subset = train_test_split(
    X_regression_subset, y_regression, test_size=0.2, random_state=0
)

regressor_subset = DecisionTreeRegressor(random_state=0)
regressor_subset.fit(X_train_reg_subset, y_train_reg_subset)

plt.figure(figsize=(20, 10))
plot_tree(classifier_subset, filled=True, feature_names=['age', 'race'],
class_names=class_names_list, rounded=True)
plt.title("Decision Tree Classifier Subset (Age and Race)")
plt.show()

plt.figure(figsize=(20, 10))
plot_tree(regressor_subset, filled=True, feature_names=['age', 'race'], rounded=True)
plt.title("Decision Tree Regressor Subset (Age and Race)")
plt.show()
```

CLUSTERING METHODS DBSCAN, KMEANS AND KMEDOIDS:

```
def dbscan_clustering_with_visualization(coordinates, eps=0.5, min_samples=5):

    dbscan = DBSCAN(eps=eps, min_samples=min_samples)
    clusters = dbscan.fit_predict(coordinates)


    plt.figure(figsize=(12, 8))
    sns.scatterplot(x=coordinates['longitude'], y=coordinates['latitude'], hue=clusters, marker='X',
palette='viridis', s=50, alpha=0.6, edgecolor='w')

    noise_points = coordinates[clusters == -1]
    plt.scatter(noise_points['longitude'], noise_points['latitude'], color='grey', s=50, alpha=0.6,
label='Noise', marker='x')
```

```
    plt.title('DBSCAN Clustering of Police Shootings', fontsize=15)
    plt.xlabel('Longitude', fontsize=12)
    plt.ylabel('Latitude', fontsize=12)

    plt.legend(title='Cluster', loc='upper right')
    plt.show()

def kmeans_clustering_with_visualization(coordinates, n_clusters=5):

    kmeans = KMeans(n_clusters=n_clusters, random_state=42)
    clusters = kmeans.fit_predict(coordinates)

    plt.figure(figsize=(12, 8))
    sns.scatterplot(x=coordinates['longitude'], y=coordinates['latitude'], hue=clusters,
palette='viridis', s=50, alpha=0.6, edgecolor='w')
    plt.scatter(kmeans.cluster_centers_[:, 1], kmeans.cluster_centers_[:, 0], s=200, color='red',
marker='X', label='Centroids')
    plt.title('KMeans Clustering of Police Shootings', fontsize=15)
    plt.xlabel('Longitude', fontsize=12)
    plt.ylabel('Latitude', fontsize=12)
    plt.legend(title='Cluster', loc='upper right')
    plt.show()

def kmedoids_clustering_with_visualization(coordinates, n_clusters=5):
    kmedoids = KMedoids(n_clusters=n_clusters, random_state=42)
    clusters = kmedoids.fit_predict(coordinates)


    plt.figure(figsize=(12, 8))
    sns.scatterplot(x=coordinates['longitude'], y=coordinates['latitude'], hue=clusters,
palette='viridis', s=50, alpha=0.6, edgecolor='w')
    plt.scatter(kmedoids.cluster_centers_[:, 1], kmedoids.cluster_centers_[:, 0], s=200, color='red',
marker='X', label='Medoids')
    plt.title('KMedoids Clustering of Police Shootings', fontsize=15)
    plt.xlabel('Longitude', fontsize=12)
    plt.ylabel('Latitude', fontsize=12)
    plt.legend(title='Cluster', loc='upper right')
    plt.show()

coordinates = df[['latitude', 'longitude']].dropna()


dbscan_clustering_with_visualization(coordinates, eps=0.5, min_samples=5)
kmeans_clustering_with_visualization(coordinates, n_clusters=5)
kmedoids_clustering_with_visualization(coordinates, n_clusters=5)
```

BOX PLOT FOR KILLINGS BASED ON AGE SEPARATED BY RACE:

```
 x = df.groupby('race')
 print(x['age'].mean())
 sns.boxplot(x='race', y='age', data=df, palette='Blues_d')
 plt.xlabel("Race")
 plt.ylabel("Age")
plt.title("Killings Based on Age Separated by Race")
```

CUMULATIVE DISTRIBUTION FUNCTION:

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt

file_path=('C:/Users/User1/Documents/MTH-project2/fatal-police-shootings-data_cleaned.xlsx')
data = pd.read_excel(file_path)

data_b = data[data['race'] == 'B']['age'].dropna()
data_w = data[data['race'] == 'W']['age'].dropna()


    n = len(data)

    x = np.sort(data)

    y = np.arange(1, n+1) / n
    return x, y

x_b, y_b = ecdf(data_b)
x_w, y_w = ecdf(data_w)

plt.figure(figsize=(10, 7))
plt.plot(x_b, y_b, marker='.', linestyle='none', color='black')
plt.plot(x_w, y_w, marker='.', linestyle='none', color='red')

plt.xlabel('Age')
plt.ylabel('ECDF')
plt.legend(['Black', 'White'], loc='lower right')
plt.title('Cumulative Distribution Function of Age by Race')

plt.grid(True)
plt.show()

cdf_diff = np.abs(y_b - y_w[:len(y_b)])  # Match the shorter array length

msd = np.mean(cdf_diff ** 2)

msd
```

CONDITIONAL PROBABILITY RATIO BY AGE:

```
def calc_conditional_probs(data):
    # Calculate the total number of Black and White individuals
    total_black = data[data['race'] == 'B'].shape[0]
    total_white = data[data['race'] == 'W'].shape[0]

    # Initialize dictionaries to store the probabilities
    prob_black_given_age = {}
    prob_white_given_age = {}

    # Calculate the probabilities for each age k
    for k in range(data['age'].min(), data['age'].max() + 1):
        prob_black_given_age[k] = data[(data['age'] <= k) & (data['race'] == 'B')].shape[0] /
total_black
        prob_white_given_age[k] = data[(data['age'] <= k) & (data['race'] == 'W')].shape[0] /
total_white

    return prob_black_given_age, prob_white_given_age

# Calculate the conditional probabilities using the actual dataset
prob_black_given_age, prob_white_given_age = calc_conditional_probs(data)

# Calculate the ratio of these probabilities for all k
prob_ratio = {k: prob_black_given_age[k] / prob_white_given_age[k] for k in
prob_black_given_age}

# Extract the k and probability ratio values for plotting
k_vals = list(prob_ratio.keys())
ratio_vals = list(prob_ratio.values())

# Plot the probability ratio P(Black | Age <= k) / P(White | Age <= k)
plt.figure(figsize=(12, 6))
plt.plot(k_vals, ratio_vals, color='blue', label='P(Black | Age <= k) / P(White | Age <= k)')
plt.xlabel('Age')
plt.ylabel('Conditional Probability Ratio')
plt.title('Conditional Probability Ratio by Age')
plt.legend()
plt.grid(True)
plt.show()
```

KDE AND Q-Q PLOT FOR ASIAN AGES:

```
df = pd.read_excel('C:/Users/User1/Documents/MTH-project2/fatal-police-shootings-data_cleaned.xlsx')

asian_ages = df[df['race'] == 'A']['age']

print("Median Asian ages =", asian_ages.median())
print("Mean Asian ages =", asian_ages.mean())
print("Stddev Asian ages =", asian_ages.std())
print("Variance Asian ages =", asian_ages.var())
print("Skewness Asian ages =", skew(asian_ages, bias=False))
print("Kurtosis Asian ages =", kurtosis(asian_ages, bias=False))

sns.kdeplot(asian_ages, fill=True)
plt.show()

from statsmodels.graphics.gofplots import qqplot

qqplot(asian_ages, line='s')
plt.show()
```

NOTE: Similarly, we plotted for other races.

SMOOTH HISTOGRAM OF AGES:

```
ages_white = df[df['race'] == 'W']['age'].dropna()
ages_black = df[df['race'] == 'B']['age'].dropna()

# Calculate mean and standard deviation for white individuals
mean_white = ages_white.mean()
std_dev_white = ages_white.std()

# Calculate mean and standard deviation for black individuals
mean_black = ages_black.mean()
std_dev_black = ages_black.std()

# Calculate the mean difference for visualization
mean_difference = mean_white - mean_black

# Perform an independent t-test
from scipy import stats
t_stat, p_value = stats.ttest_ind(ages_white, ages_black, equal_var=False)

# Monte Carlo method for estimating the p-value
np.random.seed(0)  # For reproducibility
n_simulations = 2000000
```

```
mean_diffs = []
for _ in range(n_simulations):
    pooled = np.concatenate([ages_white, ages_black])
    np.random.shuffle(pooled)
    sample_white = pooled[:len(ages_white)]
    sample_black = pooled[len(ages_white):]
    mean_diffs.append(sample_white.mean() - sample_black.mean())

# Count how many simulations have a mean difference larger than the observed mean difference
count = np.sum(np.array(mean_diffs) >= mean_difference)

# Visualizing the distributions of ages for white and black individuals
plt.figure(figsize=(10, 6))
sns.kdeplot(ages_white, color='gray', shade=True, label='White Ages')
sns.kdeplot(ages_black, color='red', shade=True, alpha=0.4, label='Black Ages')
plt.title('Smoothed Histogram of Ages')
plt.xlabel('Age')
plt.ylabel('Density')
plt.legend()
plt.show()

# Output the results
results = {
    'Mean Age White': mean_white,
    'Mean Age Black': mean_black,
    'Mean Difference': mean_difference,
    'T-Statistic': t_stat,
    'P-Value': p_value,
    'Simulated Mean Differences Greater or Equal': count,
    'Total Simulations': n_simulations
}

results
```

References:

[1]: *MTH 522 (Advanced Mathematical Statistics, sections 01B & 02B)*. (n.d.). MTH 522 (Advanced Mathematical Statistics, Sections 01B & 02B). https://mth522.wordpress.com/

[2]: Step-by-Step Working of Decision Tree Algorithm: https://www.analyticsvidhya.com/blog/2023/01/step-by-step-working-of-decision-tree-algorithm/

[3]: Washington Post https://www.washingtonpost.com/graphics/investigations/police-shootings-database/