# Statistical Analysis of Fatal Police Shootings in the United States

****

**Under the guidance of:**

**PROFESSOR DR. KISHEN IYENGER**

**By:**

**Andrew Naveen Kumar (Z1805107)**

**Nayana Chimmian Vellangat (Z1819986)**

**Neha Farya (Z1808969)**

**Vivek Yuvaraj (Z1806698)**

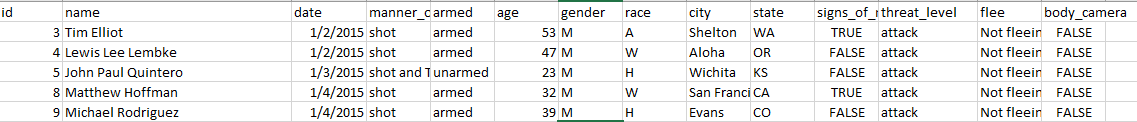
**PROJECT DESCRIPTION:**

Our project aims at finding insights about the fatal police shootings in the United States of America, which is consistently on the rise in the recent years. We spend a couple of weeks to collect the appropriate datasets and cataloguing every fatal shooting nationwide by police officers in the line of duty, from various databases based on news reports, public records, social media and other sources. Each dataset is categorized by various factors including the race of the deceased, the circumstances of the shooting, whether the person was armed and whether the person was experiencing a mental-health crisis etc.

To expand our area of research, and to dive deep into the topic, we combined our collected datasets with the Census data and High School data including fields like county\_id, county level household median income, unemployment rate, poverty rate etc to monitor their economic and education level determining the fatality.

**SAMPLE DATA INTERPRETATION:**

**POLICE KILLING DATASET:**

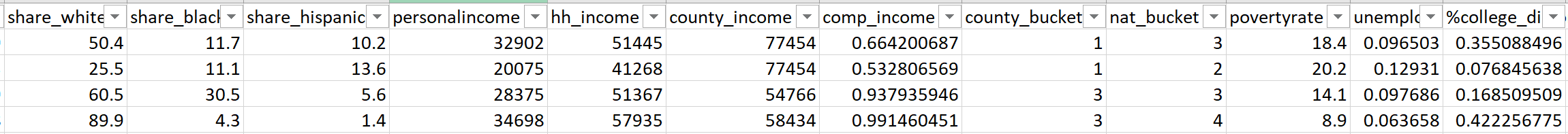


Variables in Police Killing Dataset include: Id, Name of victim, Date of police killing, Manner of death of victim, Armed status of victim (Values – armed/unarmed), Age of victim, Gender of victim (Values – Male/Female), Race of victim (Values A-Asian, W-Caucasian, H-Hispanic, B - Afro-American, N – Native Americans , O – Others), City of Police Killing, State of Police Killing, Signs of mental illness (Values – True/False), Threat levels ( Values – Attacking/Non attacking), Fleeing status of victim (Values – fleeing/not fleeing),Presence of Body camera with police officer (Values – True/False)

**Categorical Variables:** Manner of death of victim, Armed, gender, race, signs of mental illness, threat levels, fleeing status, body camera

**Continuous Variables:** Id, date, age, name, city, state

**CENSUS DATASET:**

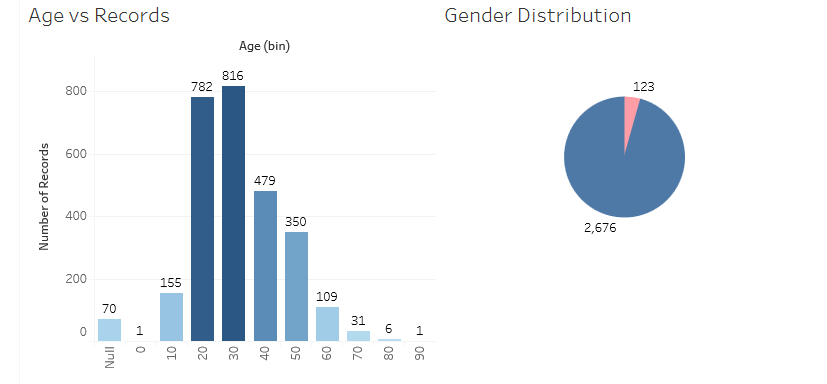
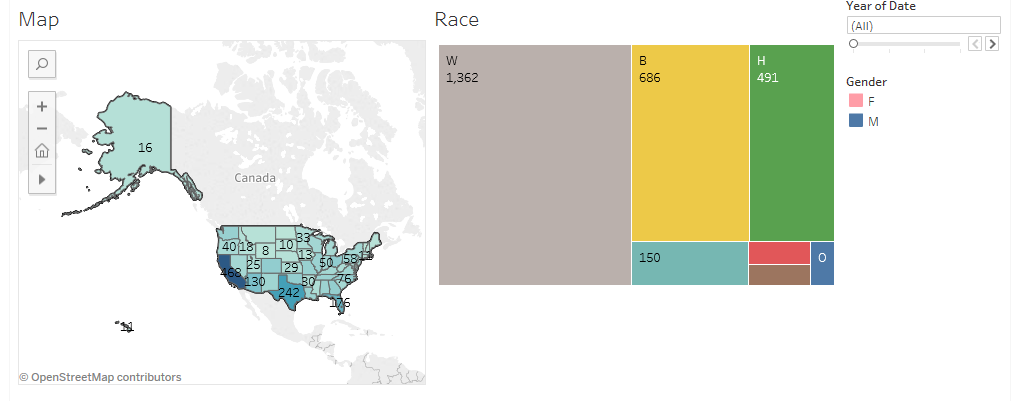


Variables in Census Dataset: Share of Caucasian, Share of afro American, Share of Hispanic, Personal income, Household income, County income, Comp income, County bucket, National bucket, Poverty rate, Unemployment, College diploma

We could combine Police killing dataset with the census dataset using Vlookup method and using county names as primary value.

**DATA VISUALIZATION USING TABLEAU:**

[https://public.tableau.com/profile/vivekraj.yuvaraj#!/vizhome/PolicekillingViz/Dashboard1?publish=yes](https://public.tableau.com/profile/vivekraj.yuvaraj)



The insights which we can get from the above visualization of police kill were as follows:

1. More number of people was killed in southern part of United States where these areas were densely populated.

2. Caucasian was killed more in number when compared to all other races.

3. Higher number of victims who got killed was middle aged men and women.

4. Overall percentage of men who got killed was more than women.

Though we can interpret these results from the Tableau visualization, we are going to find more interesting insights using SAS.

**INSIGHT 1:**

***Which race does most of the victims fall under? Does armament status influence the shootings?***

**TABLE ANALYSIS:**

Variables: Race, Armed

Both the variables are categorical and we wanted to explore if there is a relationship between them. We wanted to know whether the proportions for one variable are different among values of the other variable.

Hypothesis: H0: In the population, the two categorical variables are independent. Ha: In the population, two categorical variables are dependent.

The result is chi-square = 19.32, with degree of freedom = 5 and P = 0.0017 which is less than 0.05, we reject the null and hence conclude the variables are dependent of each other, which means the victim’s Race and their armament status has a relationship between them.

**Interpretations:**

From the graph (Fig1b), we can observe that the people who were armed were the most victims in all the races.

And, the victims all belonged to the race: White.

Even though the killings tend to take place in neighbourhoods that are poorer and blacker than the rest of the U.S. as a whole. About 50% percent of the killings — 236 of the 465 — people who were killed in the year of 2015 were White Americans according to our data. Majority of those killed by police died were armed in some form.

From Fig: 1c, here we are interested in showing that the armed/unarmed groups have different percentages of victims based on the race. Looking at the results, we can conclude by observing that the armed column values have values much higher than the unarmed column values and specifically for the victims belonging to White racial background are significantly armed in high numbers than the other races. Hence we conclude by calculating the percentage of the victims based on the status as armed in all the races combined were 93%.

**All analyzed together to answer our question we can conclude that,**

***93% of the victims killed were armed and half of them were White Americans.***

**INSIGHT 2:**

***Are people with mental illness more prone to get killed? Does age & gender influence mental illness?***

We want to check on the behavioural aspect of victim, so we run a test using Binary Logistic Regression model.

**Variables:** Signs of Mental Illness (Dependent variable)

Age, Gender (Independent variable)

**Event of Interest:** 1 (Mental Illness- True)

**Interpretations:**

Going through the results of the test, we can see that the p value is less than 0.001 which is less than alpha value (0.05); hence we can say that the test is significant. Also, the variables age and gender both has p value less than alpha (0.05). We can infer that the variables age and gender does affects the signs of mental illness.

We got the coefficient value of female as 0.7012. Since it is a positive value we, can infer that female victims tend to show more signs of mental illness than male victims

We got the coefficient value of age as 0.0176. Since it is a positive value we, can infer that with increase in the age of victims, signs of mental illness increases. We can see the c value as 57.1 percent by which we can infer that our model is 57.1 percent statistically classified correctly.

* ***Signs of Mental Illness shown by victims are increasing with respect to their age.***
* ***Female victims are observed to have higher rate of mental Illness than Male victims.***

**INSIGHT 3:**

***Does the Age, mental illness of victims and armed status of the victims have impact on Threat Level?***

We want to check on whether the threat level depends upon age, mental illness of victims, armed status of victim, so we run a test using Binary Logistic Regression model.

**Variables:** Threat\_level (Dependent variable)

Armed, Signs of Mental Illness, Age (Independent variable)

**Interpretations:**

The p value is 0.0013 which is less than alpha value (0.05), hence the model is significant. Also, the independent variables such as armed, signs of mental illness and age are having p value less than alpha value (0.05), by which we can infer that the variables do affect the dependent variable.

Armed has the highest positive coefficient value (1.3180), by which we can infer that if the victim is armed, the threat level increases. Age is also having a positive coefficient vale (0.0195) which says that with increase in age the threat level increases.

The negative coefficient value of mental illness as false is -1.2427. These mean victims who are not mentally ill pose less threat to police killing. We can see the c value as 67.4 percent by which we can infer that our model is 67.4 percent statistically classified correctly.

* ***Threat level increases with victims having Mental Illness in the crime scene.***
* ***Threat level is observed to increase with the Victims age.***
* ***Threat Level increases if the victims are armed***

**INSIGHT 4:**

***Does the economic status of an individual affect the fatal rate?***

Now we combine the police killing dataset with census data and want to check the economic status of victims involved in police killings. Here we take the personal income as the variable to find out the inference about the economic status. We are performing Linear Regression test with Personal Income as the dependent variable and age, household income, county income, poverty rate, share of Caucasian as independent variables.

**Variables:** Personal Income (Dependent variables)

Age, hh\_income, county\_income, povertyrate, share\_white (Independent variables)

**Interpretations:**

The p value is less than alpha value (0.05), hence the model is significant. Also, the independent variables such as age and county income are not significant as the p values are 0.0723 and 0.8695 respectively. So, these variables are not related to the personal income of the victim.

Also, hh\_income, povertyrate and sharewhite are having p values less than alpha value (0.05) so we can infer that they do affect the economic status of the victim.

* ***Victims with high personal income come from the households with more number of White Americans and belong to rich households.***
* ***Victims with lower personal income belong to the poor households having high poverty rate.***

**INSIGHT 5:**

***Does the relationship between Unemployment and college education moderated by Poverty rate?***

In this question, we used the census data set to understand if the unemployment and college education of people in a household can be moderated by Poverty rate

**MODEL:** Linear Regression

**Variables:** Personal Income(Dependent variable)

Poverty Rate, Unemployment, College Diploma

**Interpretations:**

It is evident from the regression model that it is significant with F value=265.89.

R-square gives 69.81% of personal income can be explained by independent variables. All the independent variables are significant and the moderator poverty rate is also significant. Since pvalue < Alpha (0.05). Looking at poverty rate\*unemployment combination we can conclude that poverty rate moderates the relationship between unemployment and college education.

* ***Poverty rate, Unemployment and College education factors have impact on personal income of the victims.***
* ***The relationship between Unemployment and personal income is observed to be moderated by poverty rate.***

**INSIGHT 6:**

***Is Unemployment status largely observed in a particular race?  
Does the educational background influence the unemployment rate?***

**Unemployment** affects the economic status of an individual.

In this question, we used the census data set to understand the relationship between unemployment and ethnicity of people in a household

**MODEL:** Linear Regression

**Variables:** Unemployment- (Dependent variable)

%of whites, %ofblack, %oofhispanic, %of college diploma (Independent variables)

**Interpretations:**

The regression model is significant with F value=53.03.

R-square gives 40.09% of unemployment can be explained by independent variables. All variables except age is found to be significant with pvalue < alpha (0.05).

All variables except %of black have negative correlation with the dependent variable unemployment.

**INFERENCE:**

* ***Victims being unemployed are high, if they belong to households with more number of African American population.***
* ***Victims being unemployed are high, if they belong to households with low rate college diploma.***

**INSIGHT 7:**

***Does the socio-economic status among the two most predominant races have any significance between them?***

**MODEL:** Correlation AnalysisIf the Share of African Americans and White Americans population in the household suspects belong to, have any correlation with their personal income, household income, poverty rate, unemployment rate and college diploma rate.

**Interpretations:**

In this correlation analysis, all the variables are dependent each other and have correlation between them. Since the p values obtained are significant.

* ***Suspects coming from households with more number of African Americans***

***Have lower personal income***

***Less educated***

***Belong to households, with lower household income***

***Belong to tract with high poverty rate and unemployment rate***

* ***Suspects coming from households with more number of White Americans***

***Have high personal income***

***Well educated and have college degree.***

***Belong to households, with higher household income***

***Belong to tract with low poverty rate and unemployment rate***

**INSIGHT 8:**

***How the victim population is segregated across the communities based on their ethnicity?***

**MODEL:** One way Anova

We have used one way Anova to get insights for this questions because It involves continuous dependent variable and categorical independent variable.  
**Variables:** Share\_white, share\_black, share\_hispanic, Race Ethnicity of the victims

Our interpretation from the result is that P value less than alpha therefore all three of our model is significant.

**Interpretations:**

From the graph shown above, we can observe that the people who belong to a particular ethnicity belong to the county where their ethnic population is higher.

* ***Race of the victims and the household they live, have the same ethnic population.***
* ***Share of a particular ethnic population in a household is observed high and is same as the race of the victim.***

**INSIGHT 9:**

***Does the victim fleeing the scene are armed and considered as a threat?***

Both the variables are categorical and we wanted to explore if there is a relationship between them. We wanted to know whether the proportions for one variable are different among values of the other variable.

Hypothesis: H0: In the population, the two categorical variables are independent. Ha: In the population, two categorical variables are dependent.

The P value is calculated based on the Observed frequency and the expected frequency used in this formula,

**Model:** Table Analysis: Armed and Flee

The result is chi-square = 8.19, with degree of freedom = 1 and P = 0.0042 which is less than 0.05, we reject the null and hence conclude the variables are dependent of each other, which means the victim’s Flee status and their armament status has a relationship between them.

**Interpretations:**

From the graph shown above, we can observe that the people who were armed were the most who is not fleeing from the scene. And, the victims who were not armed were likely to be fleeing from the spot.

* ***People who are not fleeing from the spot were more likely to be armed.***

**INSIGHT 10:**

***Does the economic disadvantage in low income counties influence the fatal rates?***

Both the variables are categorical and we wanted to explore if there is a relationship between them. We wanted to know whether the proportions for one variable are different among values of the other variable.

Hypothesis: H0: In the population, the two categorical variables are independent. Ha: In the population, two categorical variables are dependent.

The P value is calculated based on the Observed frequency and the expected frequency used in this formula,

The result is chi-square = 40.69, with degree of freedom =24 and P = 0.018 which is less than 0.05, we reject the null and hence conclude the variables are dependent of each other, which means the victim’s Flee status and the raceethinicity by county bucket has a relationship between them.

**Model**: Table Analysis: Race and County Bucket

**Interpretations:**

From the graph shown above, we can observe that the majority of people who is in the county bucket 5 were white Americans. The majority of people who were in the county bucket 1 is African Americans.

Which means that White Americans economy was higher when compared to all the other races but still white Americans were killed more in number.

* ***Among all the races number of African Americans killed fall under the low income county bucket.***
* ***Number of white Americans killed are higher than all other races***
* ***So the economic disadvantage in low income counties does not influence the fatal rate***

**CONCLUSION:**

The myth is that violence as posed by the victim is the cause of these deaths is absolutely questionable after our analysis in depth. We started our project with a central question that if there is an economic disadvantage among these victims and are these disadvantages segregated across various races or confined to any particular race. However the largely the victims were belonging to White, when we analyzed and found that 40 percent of those killed were from the African American community, despite only comprising 20 percent of the entire population. In all, we can conclude that the Economic disadvantages and high armament status among all the races in segregated communities is at the core of the police violence problem.

**REFERENCES:**

1. <https://github.com/>

2. <https://www.sas.com/en_us/learn/academic-programs/students.html>

3. <https://github.com/fivethirtyeight/data/tree/master/police-killings>

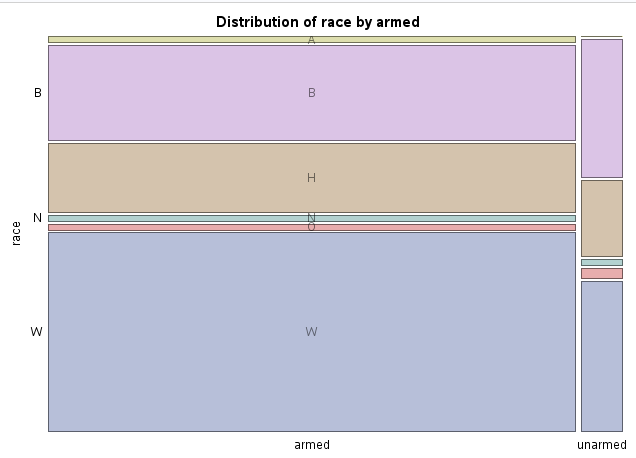
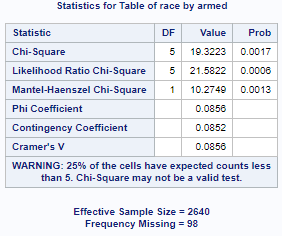
4. <https://www.washingtonpost.com/graphics/national/police-shootings-2017/>

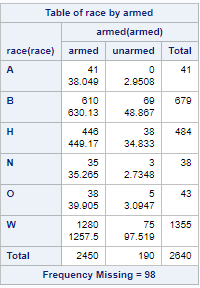
5. <https://mappingpoliceviolence.org/aboutthedata/>

6. CENSUS: https://www.census.gov/

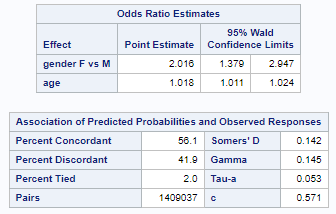
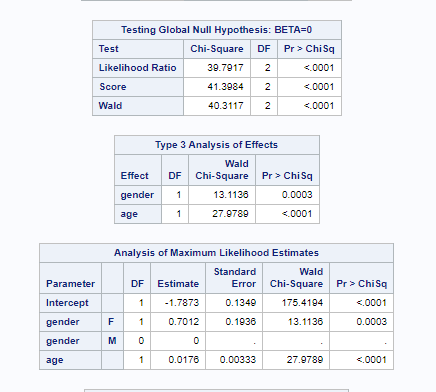
Appendix

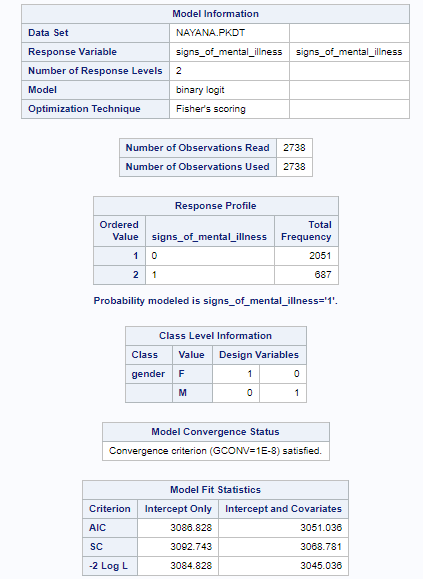
**Insight 1**



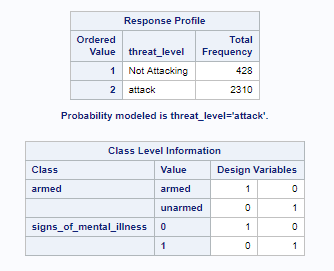
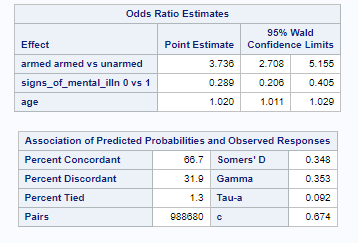
****

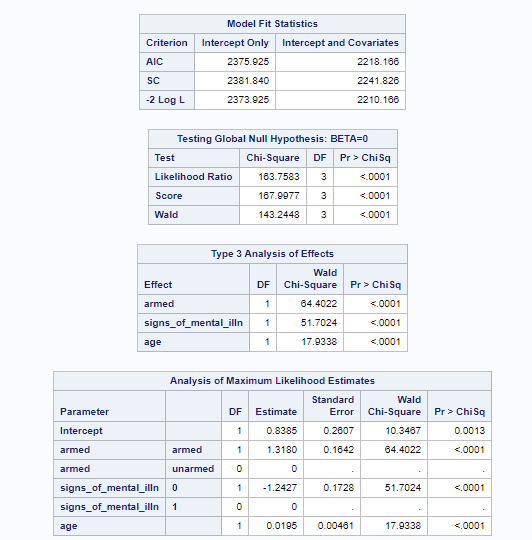
**Insight -2**



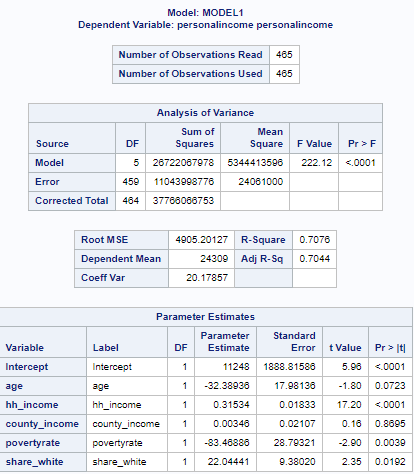
****

**Insight -3**

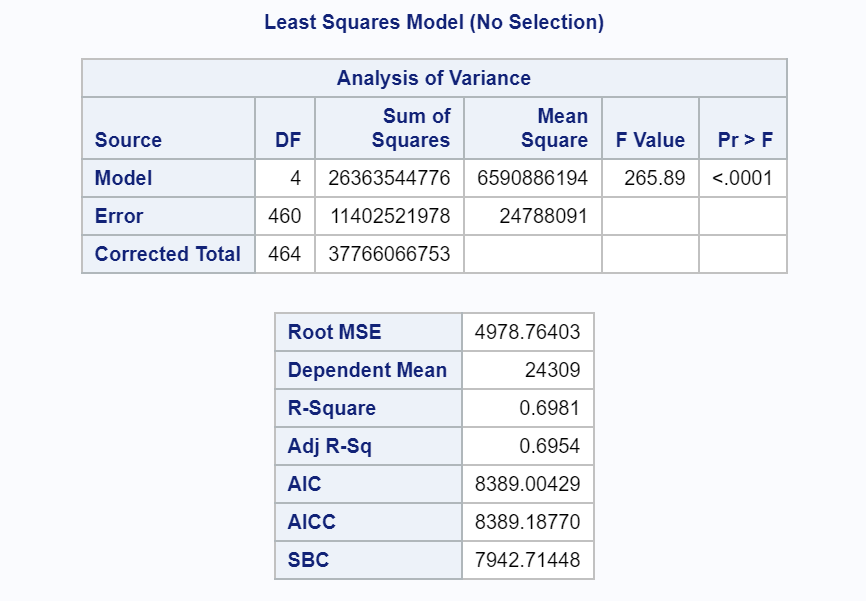
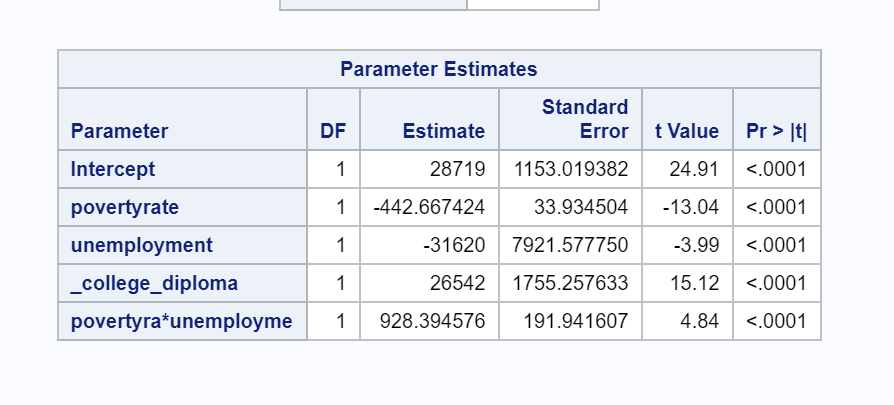
 



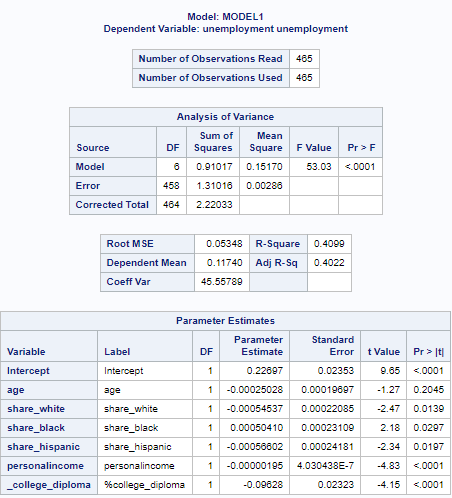
**Insight-4**



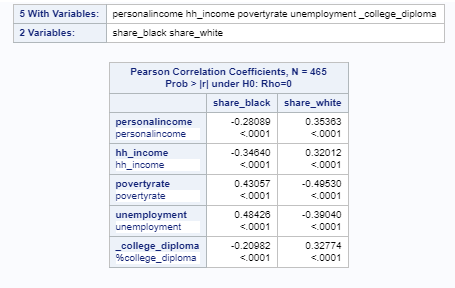
**Insight-5**

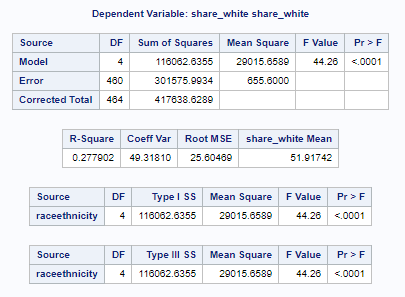
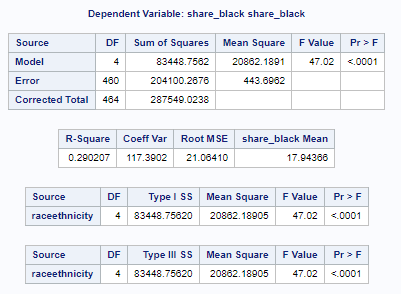
**Insight-6**

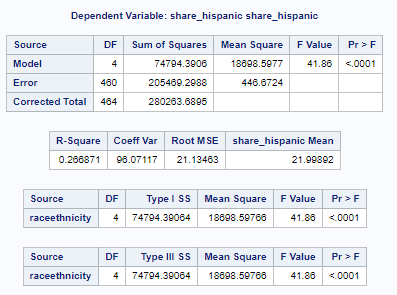


**Insight-7**

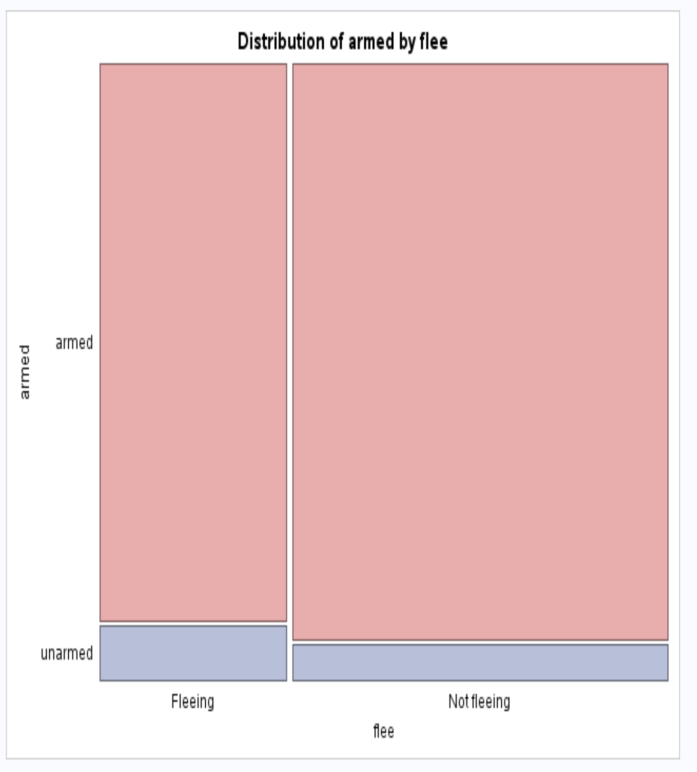
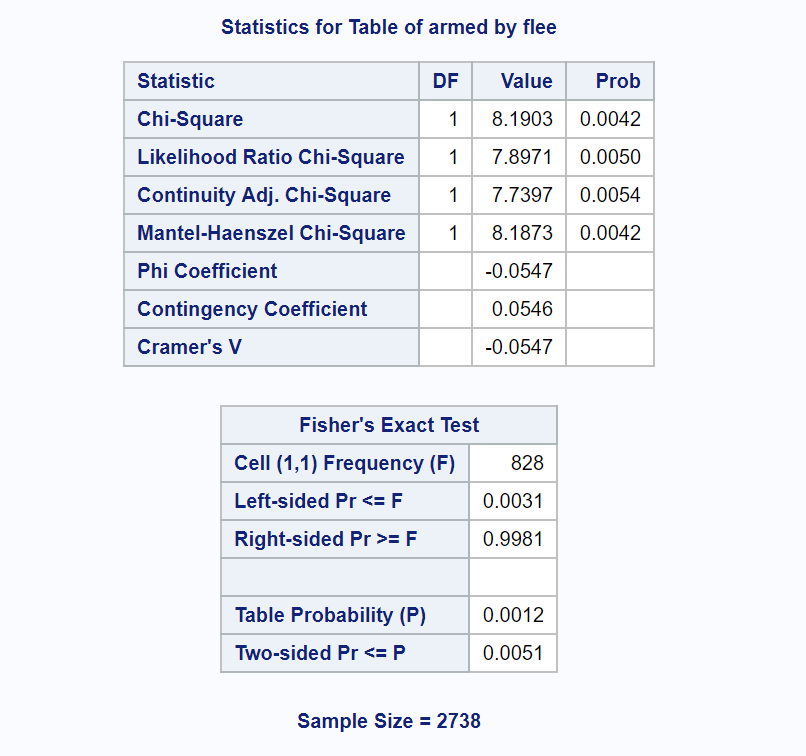


**Insight-8**



**Insight-9**



**Insight-10**

