**Analysis of Police Homicides in America (2015)**

**Summary of Research Questions:**

1. Which state experienced the highest number of police killings as compared to other states in 2015?

This question will be addressed by comparing the raw numbers of killings and killings per capita for each state. The per capita measurements will help understand which states contribute at a higher rate or fraction of their total population. Since state populations are variable, per capita measurement will help standardize the data and allow for a more accurate comparison and representation of the situation. The answer to this question is important because it can help understand which state contributed the most to the total number of police killings in 2015.

1. From the data, can it be inferred that out of the people who were killed, people who were armed were more likely to be victims of police homicide in 2015 when compared to people who were unarmed?

To answer this question, the likelihood of being a victim of police homicide will be calculated for both sections of the population of people who were victims to police homicide in 2015 – armed and unarmed. This answer can help understand a police officer’s decision to shoot and can play an important role while debating if an officer’s decision to shoot was justified.

1. From the data, can it be inferred that out of the total unarmed people killed, people from a particular section of the American population (race/ethnicity) was more likely to be victims of police homicide?

To answer this question, raw number of killings and killings (victim unarmed) in percent for each race will be compared. Why unarmed? Armed can be seen as a confounding factor because. An officer may shoot or kill an armed person while defending, regardless of the person’s race. Hence, using ‘unarmed’ data will eliminate this confounding variable and help provide a more accurate representation. The answer will help understand if any racial bias was involved in the decision to shoot.

1. Is the current trend (2016) of police killings different from the trend in 2015?

To answer this question, the total number of police killings in 2016 by month will be compared with the total number of police killings in 2015 by month.

**Motivation and Background**

Police in the US has often been accused of brutality and abuse of authority. Rallies and protests have been carried out against police brutality. However, according to the Guardian, the US government has no comprehensive record of the number of people killed by law enforcement. This lack of basic data has been glaring amid the protests, riots and worldwide debate set in motion by the fatal police shooting of Michael Brown, an unarmed 18-year-old, in Ferguson, Missouri, in August 2014. Analysis of this data will help uncover trends which can be used as evidence and hence contribute to an informed public discussion about the use of force by the police. Analyzing and visualizing data for the highest number of police killings by state can uncover information about a particular state that has been contributing the most to the total number of police killings and thus, should take steps to prevent such incidents in the future. Police brutality is also associated with racial profiling. Some sections of the US population perceive victims of police homicide as relatively powerless groups such as racial or cultural minorities. If the analysis of the data shows that a particular section of the American population (race / ethnicity) is more likely to be a victim of police homicide, then this data can be used by law enforcement agencies to take steps to protect this particular section of the US population. A look at the number of police killings for armed vs. unarmed can help address the issue of justified killing. If there is a low difference between the number of killings (armed vs. unarmed) per capita, then it can be questioned if the actions of the officers were justified. A comparison of the police homicide trend in 2015 and 2016 can help make conclusions about preventive and public safety measures taken based on the difference in the number of killings between 2015 and 2016.

**Dataset**

The program utilizes 2 data files:

1. Guardian’s interactive database of *Americans killed by police – The Counted* (<http://www.theguardian.com/us-news/ng-interactive/2015/jun/01/the-counted-police-killings-us-database)> - A .zip file containing two comma-separated values (CSV) files – ‘the-counted-2015.csv’ and ‘the-counted-2016.csv’ – can be downloaded from http://www.theguardian.com/us-news/ng-interactive/2015/jun/01/about-the-counted. However, I have made both these files available via Dropbox so that the program can access them directly:
   1. the-counted-2015.csv: <https://www.dropbox.com/s/ohknei1ivvogq7z/the-counted-2015.csv?dl=0>
   2. the-counted-2016.csv: <https://www.dropbox.com/s/h9kff40kamgut69/the-counted-2016.csv?dl=0>

The data set includes demographic information on those killed, such as age, sex and race, as well as address of where the incident occurred and cause of death.

1. National, State, and Puerto Rico Commonwealth Totals Datasets: April 1, 2010 to July 1, 2015 - I have obtained this data from the United States Census Bureau’s website. The data can be accessed in CSV format from the following link - <http://www.census.gov/popest/data/national/totals/2015/files/NST-EST2015-alldata.csv>. I will be using this dataset to extract the individual state populations (2015) in the US. The dataset contains population, population change, and estimated components of population change by state and year – change in total resident population, net migration, etc.

**Methodology:**

The program will access all 3 data files directly via the web. Data will be obtained and stored after performing data cleaning.

1. First question – To answer this question, the program will access Guardian’s ‘The Counted’ data file (the-counted-2015.csv) and the ‘National, State, and Puerto Rico Commonwealth Totals Dataset’ – state populations - data file.
   1. The program will utilize ‘the-counted-2015.csv’ data file to obtain the total number of killings for each state. The total number of killings for a particular state will be mapped to that state.
   2. The total population of each state will be mapped to that state and this data will be accessed from the state populations data file.
   3. The program will compute the number of killings per capita for each state and then map the data to the respective state.
   4. A histogram for State vs. Population will be created and saved as ‘state\_population.png’.
   5. A histogram for State vs. total number of killings will be created and saved as ‘state\_total\_kills.png’ and the name of the state with the highest number of killings will be printed.
   6. A histogram for state vs. killings per capita will be created and saved as ‘state\_kills\_percapita.png’ and the name of the state with the largest number of per capita killings will be printed.
2. Second Question: ‘the-counted-2015.csv’ data file will be required to help address this question.
   1. The total number of killings will be mapped to the respective weapon or unarmed.
   2. A pie chart will be created displaying the number of killings (in percent) for each armed / weapon and the number of killings (in percent) for unarmed. The pie chart will be saved as ‘armed\_unarmed\_stats.png’
   3. However, all weapons will be displayed in the same color (belong to armed section) and with the weapon name as the label. The unarmed section will be in a different color.
3. Third Question: The program will utilize ‘the-counted-2015.csv’ data file to answer this question.
   1. The total number of killings for which the victim was unarmed by race (obtained from ‘the-counted-2015.csv’) will be obtained.
   2. A histogram for race vs. total number of killings (unarmed) will be created and saved as ‘race\_unarmed\_killings.png’.
   3. A histogram for race vs. the probability of a victim being from that race will be created and saved as ‘race\_unarmed\_probabilty.png’.
   4. Unarmed kills by race, Total kills by race and probabilities will be printed.
4. Fourth Question:
   1. Required data files – ‘the-counted-2015.csv’ and ‘the-counted-2015.csv’.
   2. Two Lists of total number of killings for each month will be created for each year – 2015 and 2016.
   3. A month vs. number of police killings line graph with two trend lines (different colors) for each year will be created and saved as ‘trend\_compare.png’

**Work Plan**

1. Write and execute some rough code to access the required data from the web and perform some data cleaning to understand how required data should be stored. Try to answer research questions by performing some data analysis in rough. (no more than 1 hour).
2. Begin by addressing each research question.
3. For each question, identify the data needed to address the question, execute data cleaning steps and identify and store the data in a suitable data structure such that its available for future use. (Not more than 90 minutes per question).
4. Create and run tests to validate data. If any issues occur, address those issues. (Not more than 20 minutes per question).
5. Once the data extracted and stored has been validated and is available for use, for each question, use tools to visualize the stored data. (Not more than 45 minutes for each question)
6. Once code needed to address a research question is ready, add the needed methods to the main method and run the program.
7. Identify redundant code and create methods to split long procedures into steps. Add documentation for each method created. (Not more than an hour)
8. Run the final program and check if everything is working properly.

**Results**

1. Research Question 1:

From the ‘state\_total\_kills.png’ bar graph, it can be observed that California experienced the highest number of killings i.e. 210 killings. All other states have experienced less that 110 killings. However, from the ‘state\_population.png’ bar graph, it can be seen that California has the largest total population i.e. 39,144,818 which may be a cause for the high number of police killings as compared to killings in other states. Therefore, by observing the ‘state\_kills\_percapita.png’ bar graph, it can be seen that Wyoming, New Mexico and Oklahoma have the highest per capita kills. From this data, it can be inferred that these states contribute to the total number of police kills in the US at a higher rate as compared to other states, even though California experienced the most number of killings in 2015.

1. Research Question 2:

From the ‘armed\_unarmed\_stats.png’ pie chart, it can be observed that around 20% of the people killed very unarmed, whereas the 80% were armed. Out of those who were armed, around 48.6% had a firearm. The probability that a police homicide victim was armed is 0.8 and unarmed is 0.196. From this data it can be inferred that an armed person was more likely to be a victim of police homicide than an unarmed person. A reason for this can be that officers tend to shoot at or kill a person more often in the process of defending themselves if the person is armed, especially when the person has a firearm.

1. Research Question 3:

From the ‘race\_unarmed.png’ bar graph, it can be seen that unarmed Whites, experienced the highest number police killings i.e. 102, followed by Blacks and Hispanics/Latinos. However, from the ‘race\_killings.png’ bar graph, it can be observed that Whites also experienced the highest total number of police killings and this can be a cause for the highest number of unarmed killings. From the ‘race\_unarmed\_prob.png’ bar graph which it can be seen that out of the top three races (White, Black and Hispanic/Latino), it can be observed that the probability that a Black victim of police homicide is unarmed is 0.25, followed by the probability for Hispanic/Latino i.e. 0.19, and then White i.e. 0.18. (Probabilities for Arab-American, Native American, Unknown and Asian/Pacific Islander should not be taken into consideration because the total number of people killed belonging to this race is very small and the probabilities do not provide an accurate representation of the situation). Although, White people experienced the highest number of total and unarmed killings, the probability that a Black victim was unarmed is much higher. Hence, from this data it can be inferred that some racial bias is involved in an officer’s decision to shoot a person.

1. Research Question 4:

Complete data of police killings for 2016 is not available i.e. data after the month of February. Hence, the trend for 2015 and the difference in trend till March will be observed. From the ‘trend\_compare.png’ graph, it can be observed that there was a slow decline in the number of total number of killings from late March, 2015 to early June, 2015. This is also the period during which ‘Black Lives Matter’ protests were carried out in 2015. However, police killings increased at a high rate from early June to mid-July and reached their peak in the month of July.

Comparing the 2015 and 2016 trend-lines shows that, unlike 2015, police killings from January to March have increased. This comparison also questions if law enforcement agencies have implemented preventive and safety measures.

**Reproducing Results**

Datasets are accessed directly from the web and do not need to be downloaded. The code does not need to be run from the command line and should be instead run idle from Canopy. Methods / functions required to produce results are available in the main() function and have been divided into parts. Documentation has been used to state what each part does. The program can be run part by part by commenting out functions within different parts. However, this is not required and all functions can be run at once. Graphs / plots will be saved in the same location as the Python file. Figures with descriptions will also appear when code is run.

1. Research Question 1:

Methods / functions required to produce results are available in the main() function and have been divided into parts. Documentation has been used to state what each part does.

**Work Plan Evaluation**

I followed the steps in the work plan however each section took more time than I had estimated. This is also because I had to make to change to research question 3 and decide to remove a dataset for race populations in US. The Counted’s per million estimates for killings by race and race population data are based on data miscellaneous sources such as police records and voter registration data. The figures didn’t match up with the total population estimates in dataset 2 and I had to make some changes to my question.

**Video Presentation**

**No collaboration.**