**Team 7 EDA Outline**

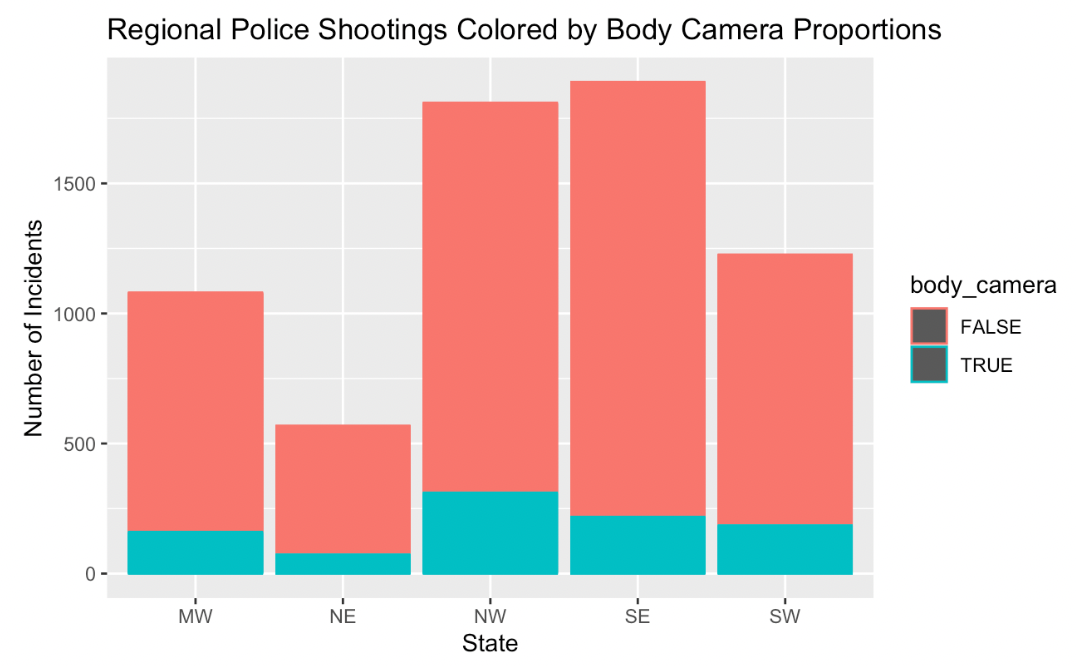
DATS 6101 Intro to Data Science

**Outline**

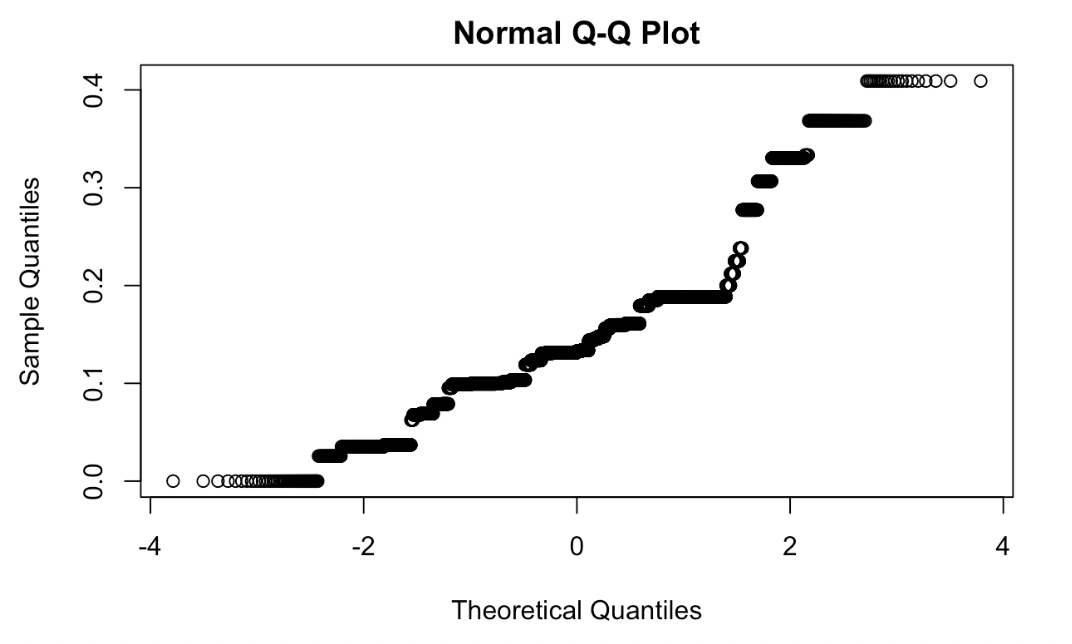
1. **MIDTERM\_EDA.Rmd.** What has already been performed for our EDA (can be found on our GitHub repository).
2. **Fried Chicken - Intro to Data Science Overview.** The example EDA outline that should be used as a model for how to construct our own EDA.
3. **Proposed EDA Outline.** An outline illustrating the proposed structure of our EDA (constructed by Cora, to be amended by Alex and Rajeev as they see fit).
4. **EDA Checklist.** A simple list of analyses already completed and those yet to be completed.

**MIDTERM\_EDA.Rmd**

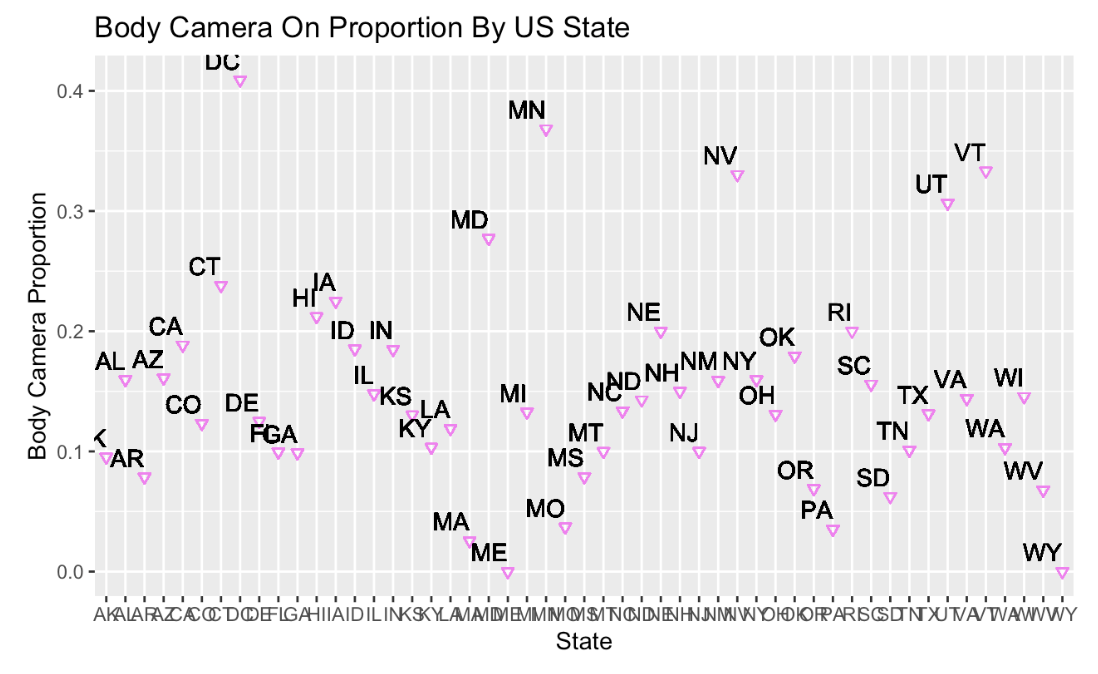
1. **Number of observations.** The number of rows in this data set is 6,574.
2. **Maximum date.** The final date in this data set is September 9, 2022.
3. **Mean age.** The average age in this data set is 37.2 years.
4. **Median age.** The median age in this data set is 35 years.
5. **Number of NE observations.** The number of observations that took place in the Northeast United States is 1,810.
6. **Number of SE observations.** The number of observations that took place in the Southeast United States is 1,226.
7. **Number of MW observations.** The number of observations that took place in the Midwest United States is 1,080.
8. **Number of SW observations.** The number of observations that took place in the Southwest United States is 1,890.
9. **Number of NW observations.** The number of observations that took place in the Northwest United States is 568.
10. **Bar graph I.** Shows the number of fatal police shootings and proportions of body cameras being on by region.



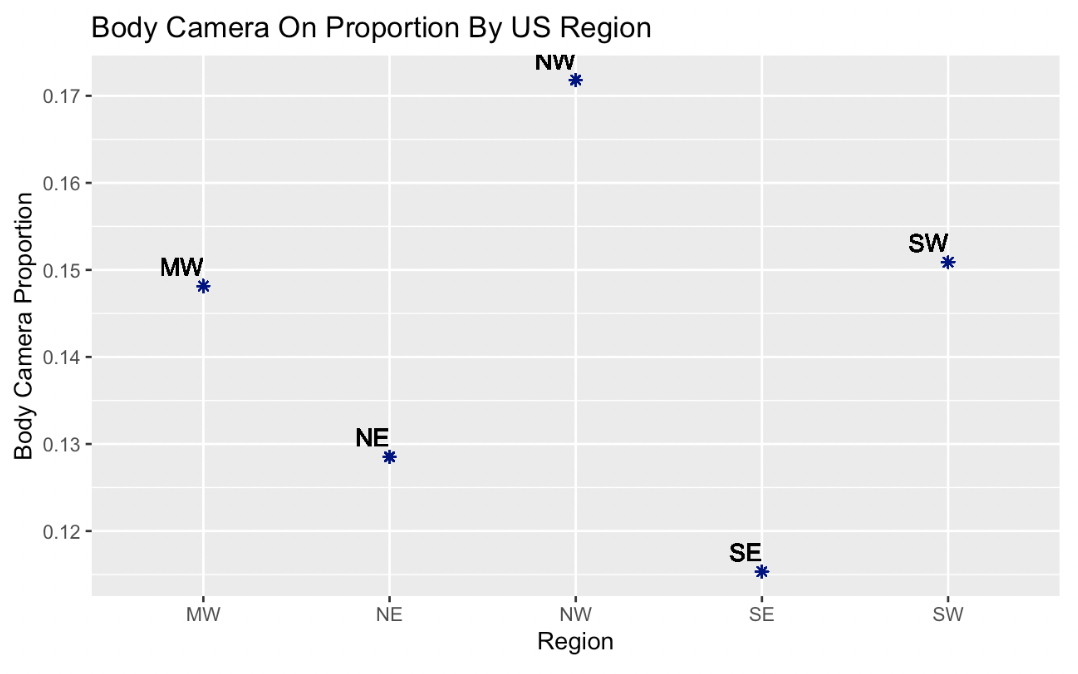
1. **QQ-plot I.** Normality check of proportions of body cameras being on by state.



1. **Scatter plot I.** Shows the proportion of body cameras being on by state.



1. **Scatter plot II.** Shows the proportion of body cameras being on by region.



1. **Analysis of variance I.** Not complete yet.

**Fried Chicken - Intro to Data Science Overview**

1. **Basic statistics.**
   1. **Mean.** The average of a data set.
   2. **Standard deviation.** Measures the dispersion of a dataset relative to its mean and is calculated as the square root of the variance. The standard deviation is calculated as the square root of variance by determining each data point's deviation relative to the mean.
   3. **Median.** The middle number in a sorted, ascending or descending list of numbers.
   4. **Range.** The spread of your data from the lowest to the highest value in the distribution.
2. **Simple correlations and tests.**
3. **Correlation matrix.** A table showing correlation coefficients between variables. Each cell in the table shows the correlation between two variables. A correlation matrix is used to summarize data.
4. **Z-test.** Used to determine whether two population means are different when the variances are known and the sample size is large. A z-test is a hypothesis test in which the z-statistic follows a normal distribution. A z-statistic, or z-score, is a number representing the result from the z-test.
5. **T-test.** Used to compare the means of two groups. It is often used in hypothesis testing to determine whether a process or treatment actually has an effect on the population of interest, or whether two groups are different from one another.
6. **Anova test.** Analysis of variance, or ANOVA, is a statistical method that separates observed variance data into different components to use for additional tests. A one-way ANOVA is used for three or more groups of data, to gain information about the relationship between the dependent and independent variables.
7. **Chi-squared test.** A statistical test used to examine the differences between categorical variables from a random sample in order to judge goodness of fit between expected and observed results.
8. **Normality.** 
   1. **QQ-plot.** A probability plot, a graphical method for comparing two probability distributions by plotting their quantiles against each other. A point on the plot corresponds to one of the quantiles of the second distribution plotted against the same quantile of the first distribution.
   2. **Boxplot.** Displays the five-number summary of a set of data. The five-number summary is the minimum, first quartile, median, third quartile, and maximum. In a box plot, we draw a box from the first quartile to the third quartile. A vertical line goes through the box at the median.
   3. **Histogram.** A graphical representation of data points organized into user-specified ranges. Similar in appearance to a bar graph, the histogram condenses a data series into an easily interpreted visual by taking many data points and grouping them into logical ranges or bins.
   4. **Shapiro-Wilk test.** A measure of how well the ordered and standardized sample quantiles fit the standard normal quantiles. The statistic will take a value between 0 and 1 with 1 being a perfect match.

**Proposed EDA Outline**

1. **Introduction.** Explain the data set, the purpose of the EDA, and its relevance and contribution to scholarly research.
2. **Basic Statistics.** Create figures with all relevant statistical expressions:
3. **Table I.** Basic statistical analyses using the number of fatal shootings in each state, grouped by region.

| **Fatal shootings by state** | **NE** | **SE** | **MW** | **SW** | **NW** |
| --- | --- | --- | --- | --- | --- |
| **Min** |  |  |  |  |  |
| **Median** |  |  |  |  |  |
| **Mean** |  |  |  |  |  |
| **Max** |  |  |  |  |  |
| **Range** |  |  |  |  |  |

1. **Table II.** Basic statistical analyses using the number of fatal shootings in each state when the police body camera was on, grouped by region.

| **Fatal shootings by state with body camera on** | **NE** | **SE** | **MW** | **SW** | **NW** |
| --- | --- | --- | --- | --- | --- |
| **Min** |  |  |  |  |  |
| **Median** |  |  |  |  |  |
| **Mean** |  |  |  |  |  |
| **Max** |  |  |  |  |  |
| **Range** |  |  |  |  |  |

1. **Standard deviation I.** A simple bell curve graph that uses data points created from each state’s quantity of shootings.
2. **Simple correlation and tests.** Checking correlation between region and shootings when the body camera was on.
3. **Correlation matrix I.** Checking the correlation between several different variables in the data set:
4. **Age.** Age of the victim.
5. **Gender.** Gender of the victim.
6. **Race.** Race of the victim.
7. **State.** The state where the shooting took place.
8. **Region.** The region where the shooting took place.
9. **Body camera on.** Whether the body camera was on.
10. **Anova test.** Alex is currently working on the development of an anova test.
11. **Z- and T-test table III.** I propose the following table:

| **The proportion of body camera on fatal shootings to body camera off fatal shootings in one region compared to the national proportion** | **Z-test** | **T-test** |
| --- | --- | --- |
| **NE** |  |  |
| **SE** |  |  |
| **MW** |  |  |
| **SW** |  |  |
| **NW** |  |  |

1. **Chi-squared test.** If possible, let’s do a statistical test from a random sample to determine whether we can predict whether a body camera will be on during a shooting in a given region? I’m not sure about this, so let’s discuss.
2. **Normality.**
3. **Bar graph I.** Shows the number of fatal police shootings and proportions of body cameras being on by region.
4. **QQ-plot I.** Normality check of proportions of body cameras being on by state.
5. **QQ-plot II.** Normality check of proportions of body cameras being on by region.
6. **Scatter plot I.** Shows the proportion of body cameras being on by region.
7. **Boxplot I.** Boxes showing frequency of shootings when body cameras were on for the NE, SE, MW, SW, and NW regions.
8. **Shapiro-Wilk test.** Perform this test on the NE, SE, MW, SW, and NW regions (using the quantity of shootings when body cameras were on).
9. **Findings.** Discuss the results of the EDA.
10. **Conclusion.** Decide where to go from here. What further research is needed to run with these findings? What are the next steps?

**EDA Checklist**

* Table I (basic statistical analyses using the number of fatal shootings in each state, grouped by region)
* Table II (basic statistical analyses using the number of fatal shootings in each state when the police body camera was on, grouped by region)
* Standard deviation bell curve
* Correlation matrix I
* Anova test
* Z- and T-test table III
* Chi-squared test.
* ~~Bar graph I. (the number of fatal police shootings and proportions of body cameras being on by region)~~
* ~~QQ-plot I (normality check of proportions of body cameras being on by state)~~
* QQ-plot II. (normality check of proportions of body cameras being on by region)
* ~~Scatter plot I (the proportion of body cameras being on by region)~~
* Boxplot I (boxes showing frequency of shootings when body cameras were on for the NE, SE, MW, SW, and NW regions)
* Shapiro-Wilk tests