HAMLINE UNIVERSITY

# Data Mining: Problem Set 1

Suabyi Thao*

September 13, 2023

> The purpose of this document is to simulataneously analyze data on US crime rates and become more familiar with the syntax and abilities of R-markdown to combine code and analysis in a progressional document. Blockquotes look better in HTML typically, but you can see their general effect in any document. The text is highlighted differently in RStudio so you know its part of the block quote. Also, the margins of the text in the final document are narrower to separate the block quote from normal text.

## The Structure of the Data

This data set contains statistics, in arrests per 100,000 residents for assault, murder, and rape in each of the US states in 1973. It has 50 observations, one for each state. As an additional variable, urban population is also accounted for. Urban population is represented as the percent of the population living in urban areas.

```
## 'data.frame':    50 obs. of  4 variables:
##  $ Murder  : num  13.2 10 8.1 8.8 9 7.9 3.3 5.9 15.4 17.4 ...
##  $ Assault : int  236 263 294 190 276 204 110 238 335 211 ...
##  $ UrbanPop: int  58 48 80 50 91 78 77 72 80 60 ...
##  $ Rape    : num  21.2 44.5 31 19.5 40.6 38.7 11.1 15.8 31.9 25.8 ...
```

```
## NULL
```

The data set has 50 observations with 4 columns. These columns are `Murder`, `Assault`, `UrbanPop`, and `Rape`. The `Murder` variable is a numeric data type, as in the `Rape` variable. The `Assault` and `UrbanPop` variables are integers.

*__Email__ sthao19@hamline.edu. __Position__ Analytics Student

## Summary of Features

| Murder | Assault | UrbanPop | Rape |
|---|---|---|---|
| Min. : 0.800 | Min. : 45.0 | Min. :32.00 | Min. : 7.30 |
| 1st Qu.: 4.075 | 1st Qu.:109.0 | 1st Qu.:54.50 | 1st Qu.:15.07 |
| Median : 7.250 | Median :159.0 | Median :66.00 | Median :20.10 |
| Mean : 7.788 | Mean :170.8 | Mean :65.54 | Mean :21.23 |
| 3rd Qu.:11.250 | 3rd Qu.:249.0 | 3rd Qu.:77.75 | 3rd Qu.:26.18 |
| Max. :17.400 | Max. :337.0 | Max. :91.00 | Max. :46.00 |

Across all 50 states, the **mean** of the *Murder* variable is 7.79 arrests for murder per 100,000 people. While the **mean** of *Assault* is 170.76 arrests per 100,000 people. *UrbanPop* has a **mean** of 65.54, and *Rape* has a **mean** of 21.23.

```
# Make sure that this code block shows up in the final document
# and that the resulting plot does also.
library(ggplot2)
library(tidyr)
scaled_data = as.data.frame(sapply(USArrests, scale))
ggplot(gather(scaled_data, cols, value), aes(x = value)) +
       geom_histogram(aes(y=..density..), bins = 10) +
       geom_density(alpha=.2, fill="#FF6666") +
       facet_grid(.~cols) +
       ggtitle("Feature Histograms for the Scaled US Arrests Data")
```

It appears that there may be some slight right -skew to the `Murder` variable and perhaps the `Rape` variable. This tells us that for both of these variables the **mean** is greater than the **median**. `Assault` seems to also have a right-skew because its **mean** is a lot greater than its **median**, with a **mean** of 170.76 and a **median** of 159. `UrbanPop` seems to be a left-skew which tells us that the mean is less than the median, with a **mean** of 65.54 and a **median** of 66.

## Relationships Between Features

The scatter plots shows that there may be a positive correlation between `Murder` and `Assault`. Generally, a positive correlation in this case tells us that as population increase, the rate of `Murder` and `Assault` will also increase. The `Rape` variable also seems to show some sort of linear as well. The scatter plot for the `UrbanPop` variable seems to be nonlinear. This tells us that an increase in population may or may not result an increase or decrease in the rate of crime in each state.

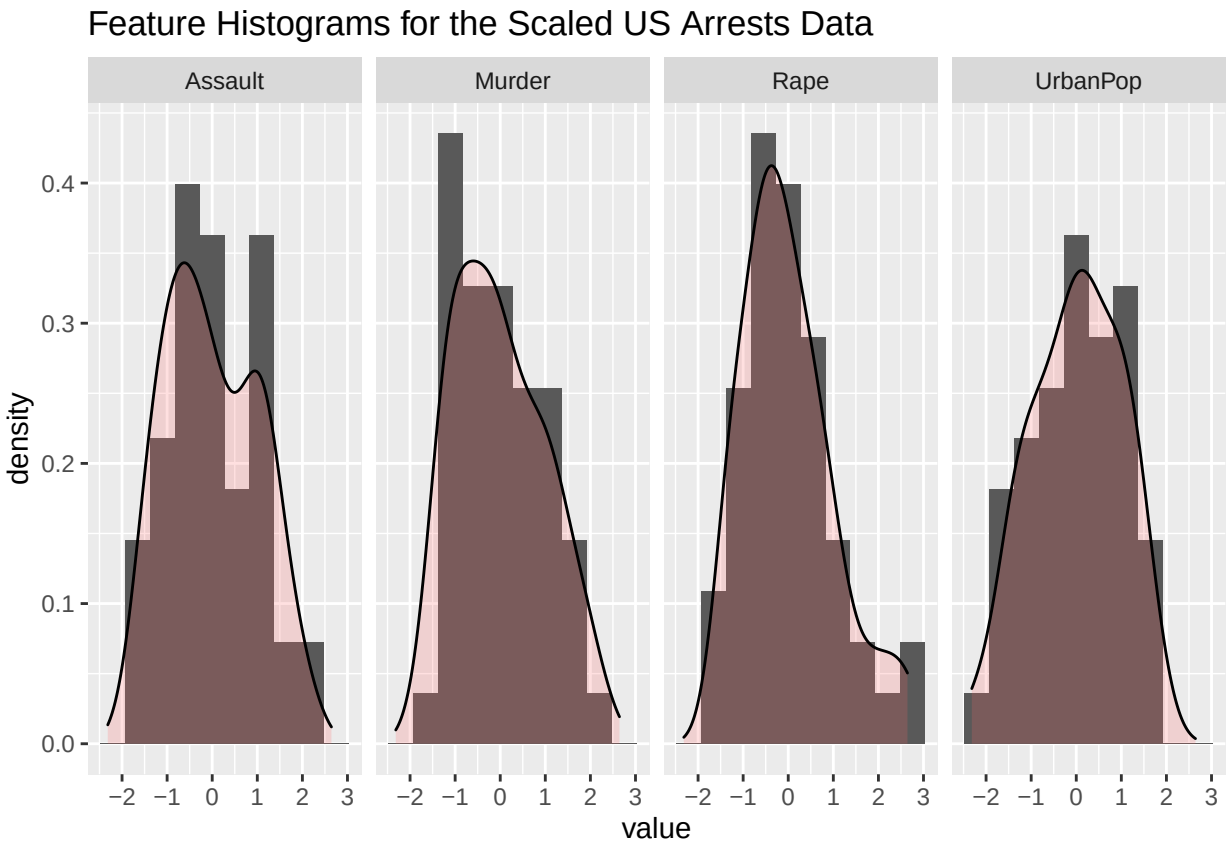


Figure 1: Histogram of Scaled Data

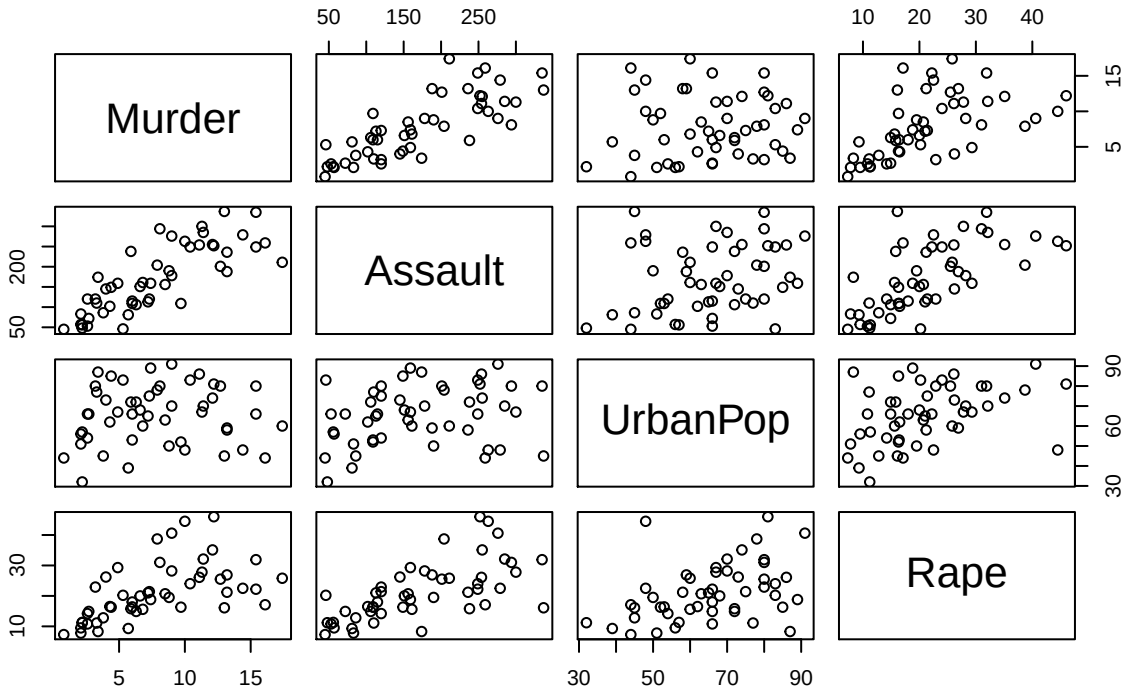


Figure 2: Facet Grid of Scatter Plots

| Variable | Mean |
|---|---|
| Murder | 7.788 |
| Assault | 170.76 |
| UrbanPop | 65.54 |
| Rape | 21.232 |

# Machine Learning Questions

In this section, you will type your paragraph answers to the following questions presented below. Do your best to answer the questions after reading chapter 1 of the textbook and watching the assigned videos.

## What are the 7 basic steps of machine learning?

The 7 basic steps are: gathering data, data preparation, choosing a model, training, evaluation, parameter tuning, and prediction.

## In your own words, please explain the bias-variance tradeoff in supervised machine learning and make sure to include proper terminology.

You need bias-variance tradeoff when you try to minimize one error but another error occurs. This is when you will need to try to find the right amount of balance between the bias and variance in order to create an accurate model which is known as the bias-variance tradeoff in supervised machine learning.

## Explain, in your own words, why cross-validation is important and useful.

Cross-validation is important because it improves the accuracy of your model. It is a way to reduce overfitting and trying to figure out the parameters that will result in less errors.