HAMLINE UNIVERSITY

Data Mining: Problem Set 1

Xander C*

09-18-2023

> 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

> The USArrests dataset contains the number of arrests for assault, murder, and rape per 100,000 people across all 50 states in 1973. It also contains the percentage of people living in urban areas vs rural 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 USArrests dataset contains 200 observations with 50 observations over 4 variables. The 4 variables being `Murder`, `Assault`, `UrbanPop`, and `Rape`. This allows us to measure the number of `Murder`, `Assault`, and `Rape` arrests compared to the `Urban Population`. These variables can tell us if there is a trend in the amount of arrests to the size of the population, or if specific states are outliers in certain crime types.

Summary of Features

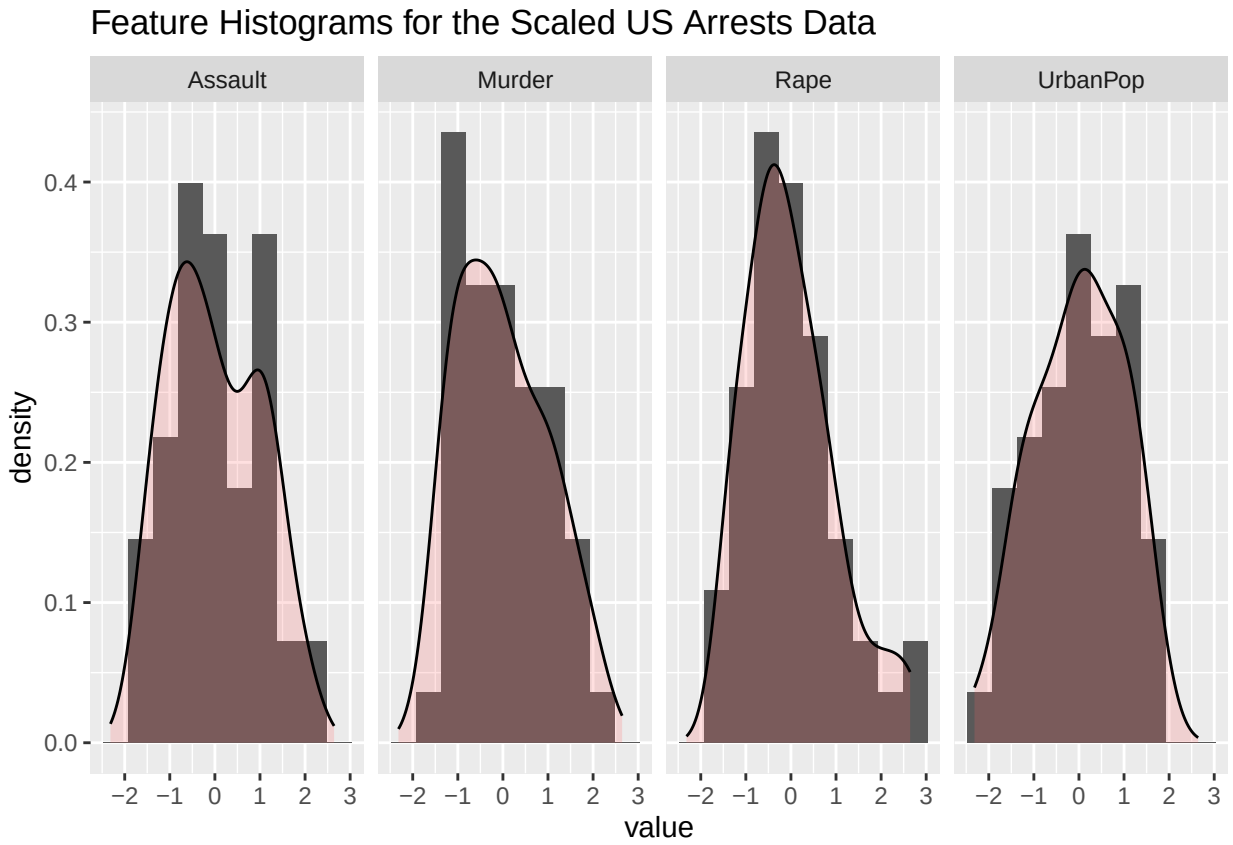
*__Email__ achapman03@hamline.edu. __Position__ Student

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

To summarize this data, we can see that per 100,000 people, there are around 7 murder arrests, 171 assault arrests, and 21 rape arrests.

Relationships Between Features

```
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")
```



In the scaled data above, we can see a small amount of skew to the left in the arrest features, and a right skew in the population feature.

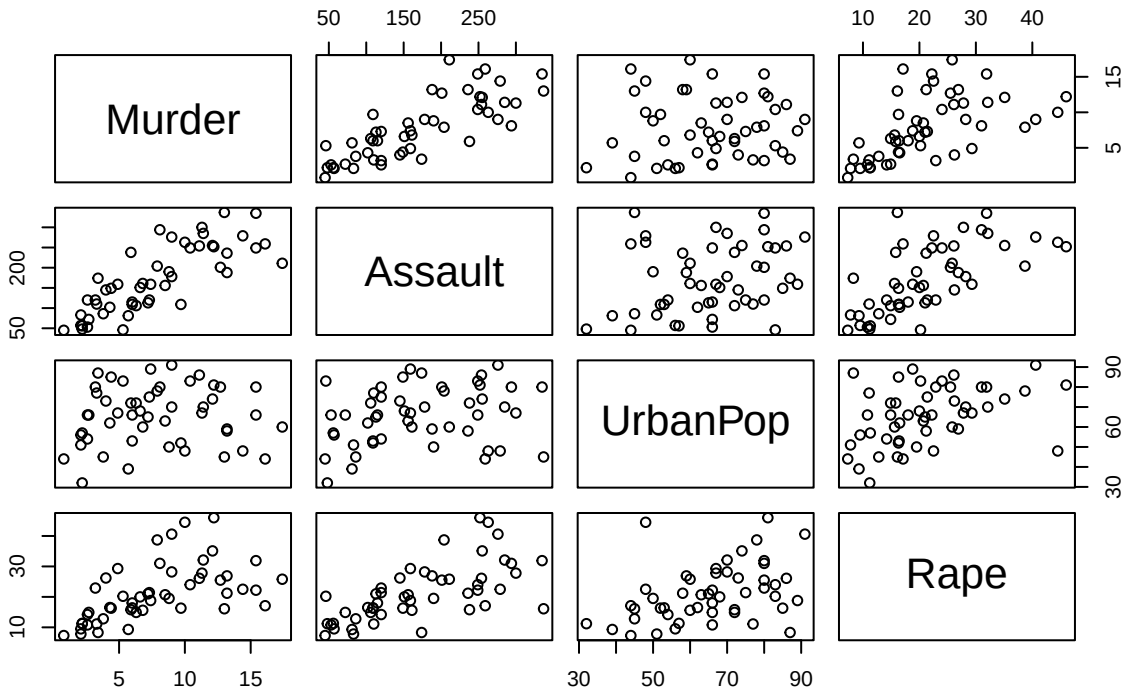


Figure 1: Facet Grid of Scatter Plots

When looking for relationships in the scatter plots above, we can see that there is some linaer relationship between the features of the dataset. This tells us that the more of one type of arrest or higher population generally indicates that there are going to be more arrests of other types.

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 of machine learning are Gathering Data, Data Preparation, Choosing a model, Training, Evaluation, Parameter turning, and prediciton.

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

The bias-variance tradeoff in supervised machine learning is give and take of overfitting and underfitting your maching learning model. Bias would be that your model is underfitting and variance would be that your model is overfitting. As you increase a model's complexity, you fit the training data better, increase the variance, and reduce the bias of the model. As you decrease a model's complexity, you give the algorithim more bias, decrease the variance, but also won't fit the training data better.

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

Cross validation is useful in Machine Learning becasue it can be used to ensure your model is well tuned. Cross validation can be used for model assessment, understanding the bias-variance tradeoff, parameter turning, and overall robustness of your model.