````
---
title: Problem Set 1 |
  ![](long_logo.png){width=4.5in}
author: Isaac Baron                        [**ibaron01@hamline.edu** **-** **Student** ]
date: "9/15/2023"
format: html
---

```{=html}
<!--
In markdown, the dashes, brackets and exclamation points marking the beginning
and end of this block of text represent comments. They will not be included as
text or as code in the document you generate. This can be a handy way to leave
yourself, teammates, coworkers, etc. important information that travels with the
document without becoming part of the final output. I will use these comment
blocks to provide directions to you in this assignment.
-->
```
```{r setup, include=FALSE}
rm(list = ls())
knitr::opts_chunk$set(echo = FALSE, warning = FALSE, message = FALSE)
# You may need to run: install.packages("tidyverse")
# before loading the library.
library(tidyverse)
```

```{=html}
<!-- The paragraph below is written as a "block quote" as a sort
     sort of abstract for the document.
-->
```
> 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

```{=html}
<!-- You are going to discuss the data we are about to analyze.
  * In the lower-right pane of RStudio, click on the Help tab.
    In the help tab search box, type in USArrests and hit enter.
    Read the provided help file.
  * Write a short paragraph discussing what the data set is about.
-->
```
This data set contains statistics, in arrests per 100,000 residents for assault, murder, and rape in each of the 50 US states in 1973. Also given is the percent of the population living in urban areas.This is a data frame with 50 observations with 4 variables. As an additional variable, urban population is also accounted for. Urban population is represented by a percent of the population.

```{r include=FALSE}
# Make sure that neither the code nor any of its
# output is included in the final document.

# Load the data into memory
data(USArrests)
```

```{r echo=FALSE}
# Make sure the code is NOT included, but that the
````
````

```
# output of the code is included in the final document.
# Print out information about the "structure" of the dataset.
print(str(USArrests))
```

```{=html}
<!-- Write a paragraph discussing the structure of the data such as:
* How many observations do we have?
* How many columns and what are they, how do we interpret their numbers?
* What kind of data types do we have for each column?
* Whenever you mention a column name, like "Assault", in your paragraph, surround the word
with single back-ticks such as `Assault`. This will change the font for that word to
monospace and make it look like code.
-->
```
The data set has 50 observations with four columns. These columns a `Murder`, `Assault`,`UrbanPop`, and `Rape`. The `Murder` variable is a numeric data type, as is the `Rape` variable. The `Assault` and `UrbanPop` variables and integers.

Summary of Features

```{r}
# This code should NOT be included, but its output should be.
knitr::kable(summary(USArrests))
```

```{=html}
<!-- Discuss the summary.
* Quickly discuss the mean of each column and interpret the mean values
  based on the definition of the column in the help file.
* In this paragraph, each time you type a column name, like "Murder"
  surround it in single stars *Murder* so that it will be italicized.
* In this paragraph, each time you type the word "mean", surround it
  with double stars **mean** so it will be bolded.
-->
```
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 170.8 arrests per 100,000 people. The **mean** of *Rape* is 21.23 arrests per 100,000 people. While the **mean** of *UrbanPop* is 65.54 per 100,000 people.

```{r echo=TRUE}
# 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")
```

```{=html}
<!-- Scaling the data centered the features at zero
and allows features to deviate above and below.  Write
a paragraph describing whether you see any slight skew
in the distributions of the features and include it below
-->
```
Murder is right-skewed. *Assault* is right-skewed. *UrbanPop* is approximately symmetric. *Rape* is right-skewed.

Relationships Between Features

```{r fig.cap="Facet Grid of Scatter Plots"}
# We can set options to make the plot result into a figure in the text.
# This allows it to be numbered, labeled, referred to etc.
# Add a caption to the figure with fig.cap="..."
# Make sure the output plot shows up, but make sure the code
# does not show up in the final document.
plot(USArrests,
     main="Scatter Plots of Crime Rates and Urban Population")
```

```{=html}
<!-- Write a paragraph describing whether you see any relationships
in terms of correlation between the features of the dataset. Do your
best to interpret any of these relationships and what they may or may
not mean.
-->
```
There appears to be a positivly correlated relationship between *Murder* and *Assault*. The appears to also be a positivly correlated relationship between *UrbanPop* and the other three arrest variables *Murder*, *Assault*, and *Rape*.

```{=html}
<!--
Finally, create a table of the mean values.
In markdown, we can specify tables using some basic
text formatting and it will be turned into a nice table.
For each feature, replace the ___ marks with inline R code,
you know the `r ` that will insert the mean value of each feature
in the table. You can get the mean using,
mean(USArrests$Murder). For the remaining features, replace
the Murder part with the feature name as spelled in the dataset.
-->
```
Variable	**Mean**
Murder	`r mean(USArrests$Murder)`
Assault	`r mean(USArrests$Assault)`
UrbanPop	`r mean(USArrests$UrbanPop)`
Rape	`r mean(USArrests$Rape)`

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?

1. Data Collection
2. Data Reprocessing
3. Splitting the Data
4. Model Selection
5. Model Training
6. Model Evaluation
7. Model Deployment

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

The bias-variance trade-off: it refers to the balance that you need to strike between two sources of error when building a model: Bias: Bias is the error introduced by approximating a real-world problem, which may be complex, by a simplified model. High bias can lead to under fitting, where the model is too simple to capture the underlying patterns in the data. It results in poor performance on both the training and validation

sets. Variance: Variance is the error introduced by the models sensitivity to small fluctuations or noise in the training data. High variance can lead to over fitting, where the model becomes too complex and fits the training data too closely. It performs well on the training data but poorly on the validation or test data. The goal in machine learning is to find a model that achieves a balance between bias and variance. This is because reducing bias often increases variance, and vice versa. The challenge is to select the right complexity of the model and fine-tune its parameters to minimize both bias and variance, resulting in a model that generalizes well to unseen data.

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

Cross-validation is essential machine because: Cross-validation provides a more robust estimate of a models performance compared to a single train-test split. It helps in assessing how well a model generalizes to different subsets of the data. Maximizing Data Utilization: By rotating through different subsets of the data as training and validation sets, cross-validation ensures that all available data is used for both training and evaluation, which is particularly important when the dataset is limited.