CYBR 520 Lab 1: Introduction to R and

Exploratory Data Analysis (100 points)

# Instructions:

The is a group lab, each team is to submit one submission on eCampus. After the lab is submitted, each group member is to submit a [group member evaluation](https://forms.gle/TvsvercxLsb7sgev9) for each group member (this is worth 15 % of the total grade). We will be using R and Rstudio for this lab. Please read the following the document and provide your answer **below** each question. Keep the formatting of the document as is.

**Dataset: iris**

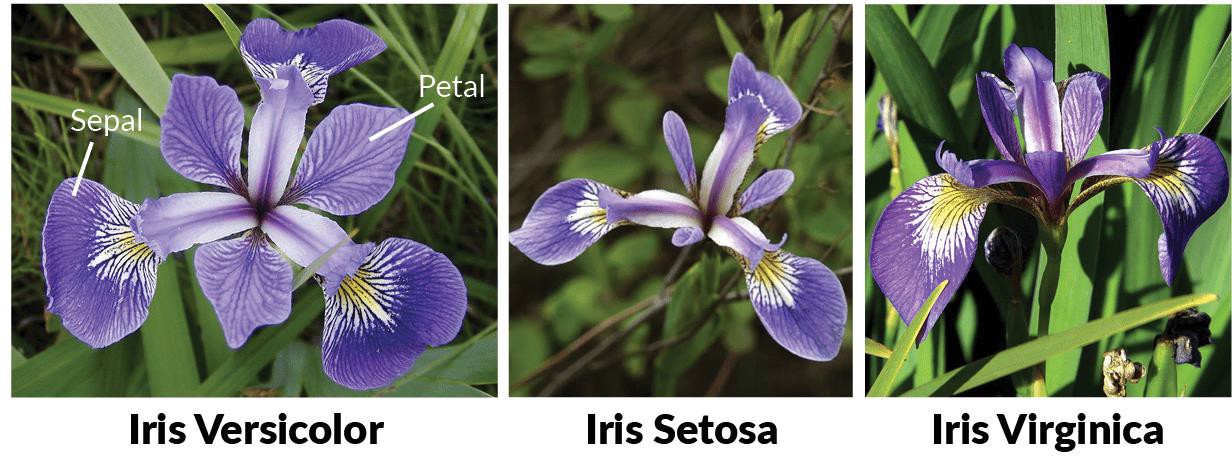
**R Packages: ggplot2**

Background: The iris dataset is a classic example dataset used for machine learning in R. Most, if not all flowers have a sepal and a petal. The sepal (Figure 1) 1 functions as a protector for the flower and support the petals when the flower is in bloom:

The Iris dataset contains measurements (in cm) of petal length and width, and sepal length and width for 3 species of the iris flower: *iris versicolor, iris setosa,* and *iris virginica* (Figure 2)2. There are exactly 50 observations for each species type, bringing the total number of observations to 150. Each observation is also



*Figure 1 - Sepal and Petal*



*Figure 2 - Iris Species*

1 https://en.wikipedia.org/wiki/Petal#/media/File:Petal-sepal.jpg

2 <https://www.datacamp.com/community/tutorials/machine-learning-in-r>

assigned membership to a specific species as well. The membership (aka “class label”) is assigned based upon the name of the species “versicolor”, “setosa”, “virginica”.

The iris dataset is already built into R, so you do not have to import the dataset from an external source or file.

## Load in our Data

Launch R Studio

After R Studio has launched, we’re going to need to load in some basic packages. The first package, datasets contains the iris data we’ll be using. We’re also going to use the ggplot2 package for this lab as well.

Type:



We also need to install the ggplot package. You can either go the packages tab and click install. Type in ggplot2 under packages and click install. Or you can type:



Now we need to load the data. Type:



We can look at some rows in the irisi dataset. Type:



Notice that the iris data should now show up in your global environment. Let’s now look at this dataset by typing:



**Question 1: What data is shown by this output? What information can you get about this dataset? (10 points)**

**Observation:**

**The summary() command statistically summarizes the data included in the dataset passed to the command (e.g., the ‘iris’ dataset, in this case). More specifically, the minimum value, 1st quantile, median, mean, 3rd quantile, and maximum value are displayed for each data field in the data set. The data fields summarized include Sepal.Length, Sepal.Width, Petal.Length, Petal.Width, and Species. In addition, the count of each species in the dataset is displayed. Please refer to the screenshot below:**

Text

Description automatically generated with medium confidence

## Exploratory Data Analysis – Iris Sataset

Now let’s do some exploratory data analysis. Exploratory data analysis is a process of performing an initially investigation of the dataset using analysis and visualization techniques to discover patterns, spot outliers and anomalies, and relationships between variables in the data. This is an often skipped, but critical step in any analysis. It can also help you develop a more informed hypothesis for scientific inquiries. In essence, we’re going to try to gain some insight on this dataset.

Let’s load the ggplot package. Type:



Now let’s visualize our data. In R, it’s best that we store our visualization as a variable. To store any function or command, or output as a variable we type an arrow “ <- “ pointing to a variable.

Type:



Now let’s display the plot. Type:



**Question 2: Post a screenshot of the plot below. You can use the export function or a snipping/screen capture tool.**

**Observation:**

**Please refer to the screenshot below:**

Chart, scatter chart

Description automatically generated

**Question 3: What observations (if any) can you make from this graph?**

**Observation:**

**From the graph above, the following observations can be made:**

1. **For all species of Iris in this dataset, there appears to be a *linear positive relationship* between the Sepal Length and Sepal Width variables. In other words, when Sepal Width increases, so does Sepal Length.**
2. **The Sepal Width and Sepal Length variables appear to be more strongly correlated in the Setosa species than in the Versicolor and Virginica species. The correlation coefficient (r) of the variables for the Setosa species would likely be around 0.7 – 0.8. However, r of the variables for the Versicolor and Virginica species would likely be closer to 0.5.**
3. **The closeness of the Versicolor and Virginica species’ points on the scatterplot suggests that these species’ sepals dimensions are more similar to each other than they are to the dimensions of the Setosa species’ sepals. Because of this, it is difficult to differentiate between the Versicolor and Virginica species when Sepal.Length is plotted as a function of Sepal.Width.**
4. **Sepal Length of the Setosa species appears to be generally smaller than that of the Versicolor and Virginica species. However, Setosa Sepal Width appears to be generally higher than that of the Versicolor and Virginica species.**

Now notice that in our graph we plotted Sepal.Length as a function of Sepal.Width.

Modify the code above to plot petal length as a function of petal width, and store that graph as a new variable. HINT: use the summary(iris) command first to identify the other two petal variables

**Question 4: Paste your modified code below and a screenshot of your new graph showing petal length and petal width.**

**Observation:**

**Please refer to the updated code below. Notice that a new variable (irispetalplot) was created and entered into the prompt to generate the scatterplot:**

**> irispetalplot <- ggplot(data=iris, aes(x = Petal.Length, y = Petal.Width)) + geom\_point(aes(color=Species, shape=Species))**

**> irispetalplot**

Chart, scatter chart

Description automatically generated

**Question 5:**

1. **What observations (if any) can you make based upon this visualization?**

**Observation: From the scatterplot above, the following observations can be made:**

1. **For all species in this dataset, there appears to be a *linear positive relationship* between the Petal Length and Petal Width variables.**
2. **The correlation strength between the Petal Length and Petal Width variables appears to be similar in all 3 species of Iris (r = ~0.7 – 0.8).**
3. **The Setosa species generally has the smallest petal dimensions (0.2 – 0.5 for width and 0.3 – 1.9 for length). The Versicolor species has the mid-range petal dimensions (1.0 – 1.7 for width and 3.0 – 5.1 for length). The Virginica species generally has the largest petal dimensions (1.4 – 2.5 for width and 4.5 – 6.9 for length)**
4. **Which of the previous visualizations shows a clear distinction between all three species?**

**Observation: The second visualization, which plotted the Petal Length as a function of Petal Width, demonstrates a clear difference in dimensions between the 3 species of Iris. Refer to the answer for part a. above for a more detailed comparison of the dimensions.**

1. **Based upon the visualizations, which variables would be most useful for identifying or distinguishing all three species from each other?**

**Observation: Petal Length and Petal Width would be the most useful variables in distinguishing between the 3 species of iris, due to the clear delineation observed between the 3 species on the scatterplot generated when comparing petal length and width.**

Take a look at some ggplot documentation ([https://datacarpentry.org/R-ecology-lesson/04-](https://datacarpentry.org/R-ecology-lesson/04-visualization-ggplot2.html) [visualization-ggplot2.html](https://datacarpentry.org/R-ecology-lesson/04-visualization-ggplot2.html)).

**Question 6:**

1. **Create a histogram and boxplot visualization of any two variables. See if you can get creative with the coloring and visualization.**
2. **Do you think either of the plots are useful for exploring the data? Paste a copy of your visualizations below and the code you used to create them.**

**Observation:**

**Please refer below to the screenshots and code for the histograms and boxplots our team created. In addition, we provide an analysis of the utility of each visualization:**

**Histogram:**

Chart, bar chart, histogram

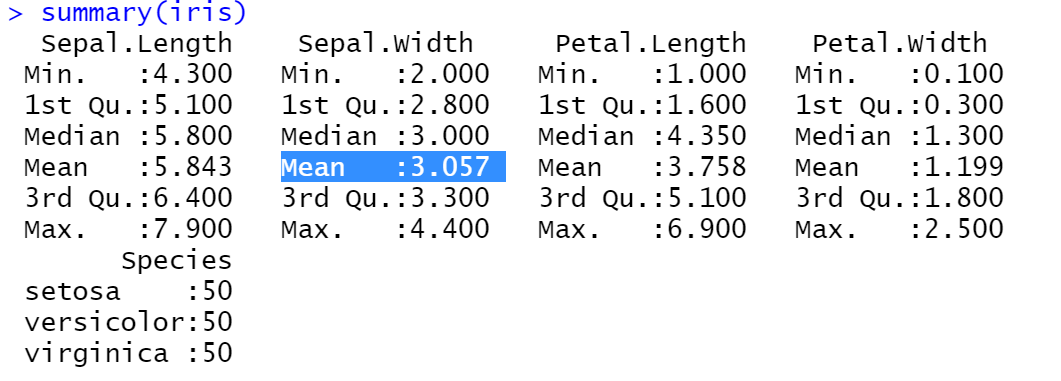
Description automatically generated

**Code:**

**> irisHist <- ggplot(data=iris, aes(x = Sepal.Width, color = Species)) + geom\_histogram(aes(y=..density..), bins=15, color="cyan", fill="darkred") + geom\_density(alpha=0.2,fill="blue") + theme(legend.position = "right") + theme(plot.title = element\_text(hjust = 0.5)) + labs(title="Distribution of Sepal Length") + xlab("Sepal Width") + ylab("Density")**

**> irisHist**

**Histograms are useful in analyzing datasets in general because they display the number of data points across a range of values for a given variable. Because the iris dataset consists of three species of iris wherein there is a degree of commonality between certain variables, the viewer may find it difficult to determine which species of iris contribute to the count in a particular range on the graph. The histogram above utilizes kernel density estimation (KDE) curves to mitigate this issue by plotting a continuous curve for each species. This is helpful in exploring the iris dataset because it helps the viewer visualize the portions of the distribution that can be attributed to each separate species. This histogram method can be beneficial in cases such as this where the data from multiple categories (i.e., species of iris) are being represented on the same chart, which results in an overlay of the chart series. From the histogram above, it can be gleaned that the most common Sepal Width amongst the 3 species is somewhere around 3.0 – 3.1, which agrees with the mean value reported from the summary(iris) command, as shown below:**



**Boxplot:**

A picture containing box and whisker chart

Description automatically generated

**Code:**

**> irisBox <- ggplot(data=iris, mapping = aes(x = Sepal.Length, y = Species)) + geom\_boxplot(color="dark blue", fill="gold") + stat\_boxplot(geom="errorbar") + xlab("Sepal Length") + labs(title="Sepal Length Across Species") + theme(plot.title = element\_text(hjust = 0.5)) + geom\_jitter(alpha=0.5, color="dark red")**

**> irisBox**

**Similar to a histogram, a boxplot provides a visual representation of the distribution of a given variable with respect to a specific variable. A boxplot is especially useful in showing data skew and variation within the data set, which are important statistical characteristics that can be more difficult to visualize in other graphs, in a linear fashion. More specifically, from the boxplot above it is arguably easier to visualize the median (blue line in the box) and maximum / minimum values than it is to determine these values from a histogram. The box plot above includes whiskers to show which data points are outliers, as well as jitter points showing where the data points of each species lies within the range of the Sepal Length variable.**

# Exploratory Data Analysis - Freeform

There are several additional preloaded datasets in R. Explore and choose one of these datasets and conduct an exploratory data analysis.

**Question 7:**

* 1. **Write a brief 1-2 page exploratory data analysis summary of the dataset of your choice, and your exploratory data analysis of that sample dataset.**
  2. **Include a useful visualization or two of the datasets of your choice as well. There are also several cybersecurity related datasets that can be easily loaded into R including the spambase dataset that I’ve loaded on eCampus. Feel free to explore that dataset as your example as well.**

To load a csv file into R, you can use the read.csv command below:



Notice that in R, we use forward slashes for directory navigation, even in Windows.

**Observation:**

**For the freeform exploratory data analysis portion of this lab, we have chosen to analyze the USArrests dataset included in R. The USArrests dataset includes the number of criminal arrests for murder, assault, and rape per 100,000 persons across all 50 states in 1973. In addition, the dataset reports for all 50 states the percentage of the population that lives in Urban areas. For a sample view of the dataset, please refer to the output of the head() function for USArrests below:**

Table

Description automatically generated

**It is also helpful to view a summary of the USArrests dataset to ascertain the statistical characteristics of the data. Please refer to the output of the summary() function below:**

Text

Description automatically generated with medium confidence

**Given the data above, it would be helpful to determine whether a correlation exists between crime type and the percentage of the population that lives in Urban areas. We examine this using the plots below:**

Chart, calendar

Description automatically generatedChart, scatter chart

Description automatically generated

A picture containing text, white

Description automatically generated

**From the charts above, it can be gleaned that there is approximately 0 correlation between Urban Population percentage and the number of arrests for Murder and Assault. However, a slight correlation in the positive direction appears to exist between the Urban Population percentage and the number of Rape arrests.**

**Next, it would be helpful to determine if there is a correlation between any two crime types (or all of them). We modify the charts above as follows:**

Chart, calendar, scatter chart

Description automatically generatedCalendar

Description automatically generated

Chart, scatter chart

Description automatically generated

**From the charts above, the viewer can see there is a positive relationship between Rape and Murder, Assault and Murder, and Assault and Rape. Is should be noted, however, that this does not mean any of the crimes cause any of the other crimes. In other words, there is a correlation but not necessarily a causal relationship between the crime type variables.**

**To conclude our study, we would like to visualize statistical characteristics from the distributions of data, such as variance and skew, utilizing histograms. Refer to these visualizations for each crime below:**

Chart, histogram

Description automatically generatedChart, histogram

Description automatically generated

Chart, histogram

Description automatically generated

**In conclusion, we have demonstrated that each crime type appears to be correlated to one another. In addition, the distributions of each crime type are skewed to the right, and the Assault distribution is bimodal.**