

Data Cleaning

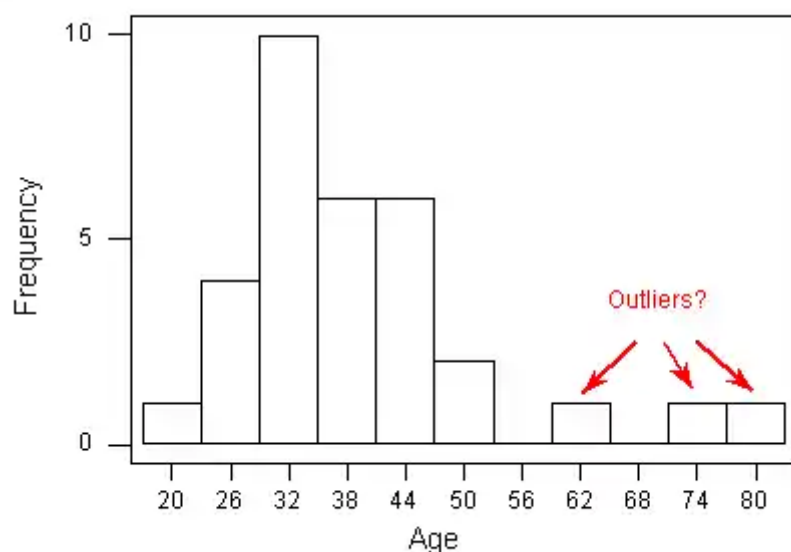
Section 3: Missing Data, Outlier Detection, and Principal Component Analysis (PCA)



LESSON 6: OUTLIER DETECTION

Lesson 6: Histograms and Box Plots

There are some simple graphs that can be used to visually spot outliers. One graph is called a histogram. Think of this like a "bar graph for numbers." What you are looking for in this graph are a few values separated from the majority of the others. For example, here is a histogram showing potential outliers. Notice that there are few values on the right and they seem separated from the bulk of the data on the left.



Histogram with potential outliers

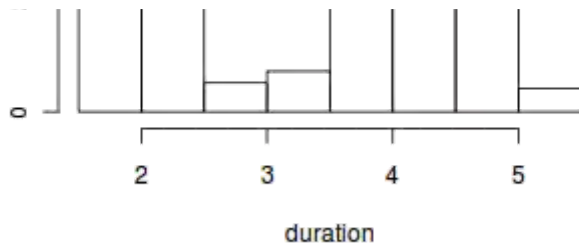
Here is a histogram with no noticeable outliers:



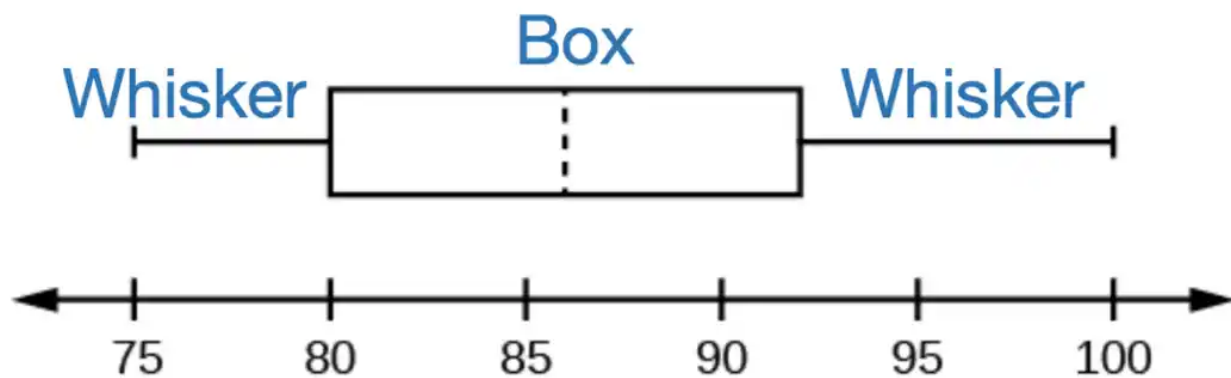
Data Cleaning

Section 3: Missing Data, Outlier Detection, and Principal Component Analysis (PCA)

LESSON 6: OUTLIER DETECTION

*Histogram with no noticeable outliers*

Another graph you can use is a box plot. This graph is sometimes known as a "box and whiskers" plot because it consists of a box and two lines on either side referred to as "whiskers."

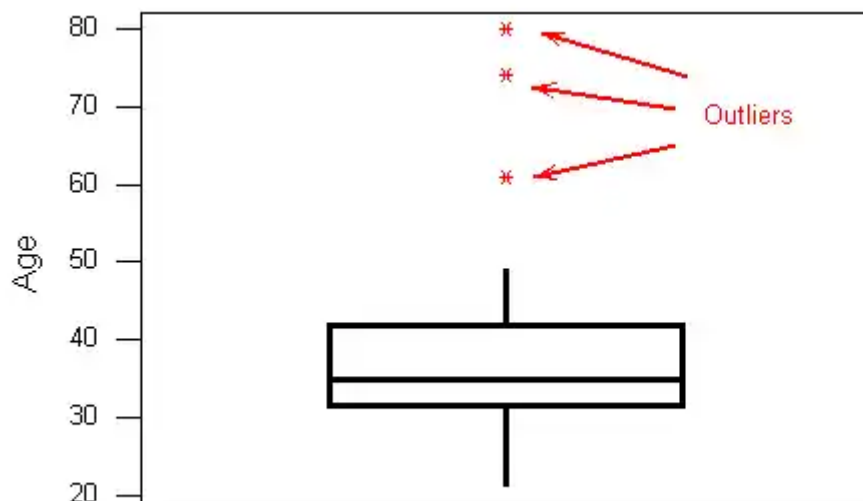
*Box and whiskers plot*

Data Cleaning

Section 3: Missing Data, Outlier Detection, and Principal Component Analysis (PCA)



LESSON 6: OUTLIER DETECTION



A box and whiskers plot with outlier dots

In this graph, there are three dots for certain ages: 60, 75, and 80. Although 80 is the biggest age, a 50 is considered the biggest typical age (based on the majority of the data). Anything beyond this value is considered an outlier. An age of 22 is considered to be the smallest age, and this is considered typical, so the line extends there with no outliers. There are some minor math formulas that help you determine what is typical versus unusual. You will learn about these in the next course.

Within R, the most used visualization tool for exploratory data analysis is ggplot2.

Data Cleaning

Section 3: Missing Data, Outlier Detection, and Principal Component Analysis (PCA)



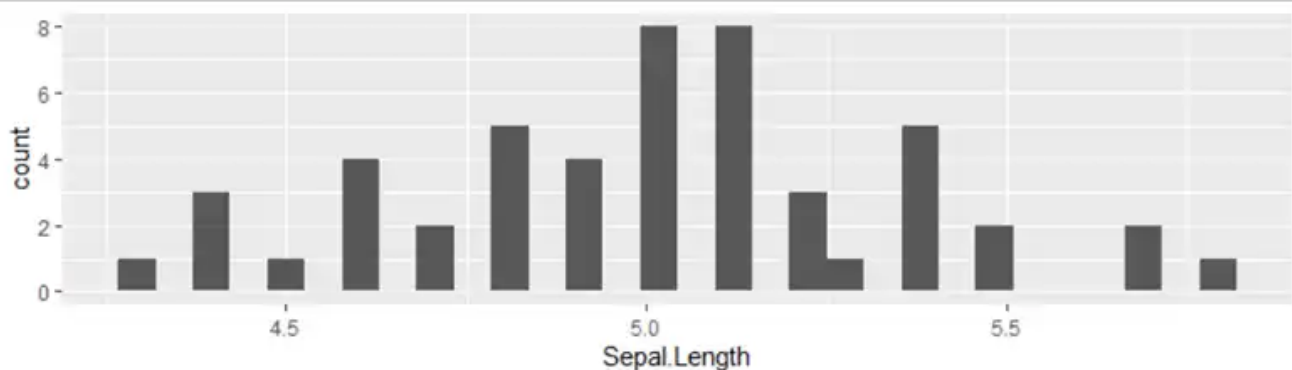
LESSON 6: OUTLIER DETECTION

R Input

```
# Histogram and box plot for the setosa species of the data set  
hist_plt <- ggplot(data %>% filter(Species == 'setosa'), aes(Sepal.Length)) + geom_histogram()  
  
box_plt <- ggplot(data %>% filter(Species == 'setosa'), aes(Sepal.Length)) + geom_boxplot()  
  
print(hist_plt)  
print(box_plt)
```

R code to plot a data set

[Download the R code and output \(opens new tab\)](#)



Histogram of setosa flower

Data Cleaning

Section 3: Missing Data, Outlier Detection, and Principal Component Analysis (PCA)



LESSON 6: OUTLIER DETECTION

Box plot of setosa flower

In the box plot, there are no outliers present. Here is an example using the car data built into R (mpg):

```
p <- ggplot(mpg, aes(class, hwy))
```

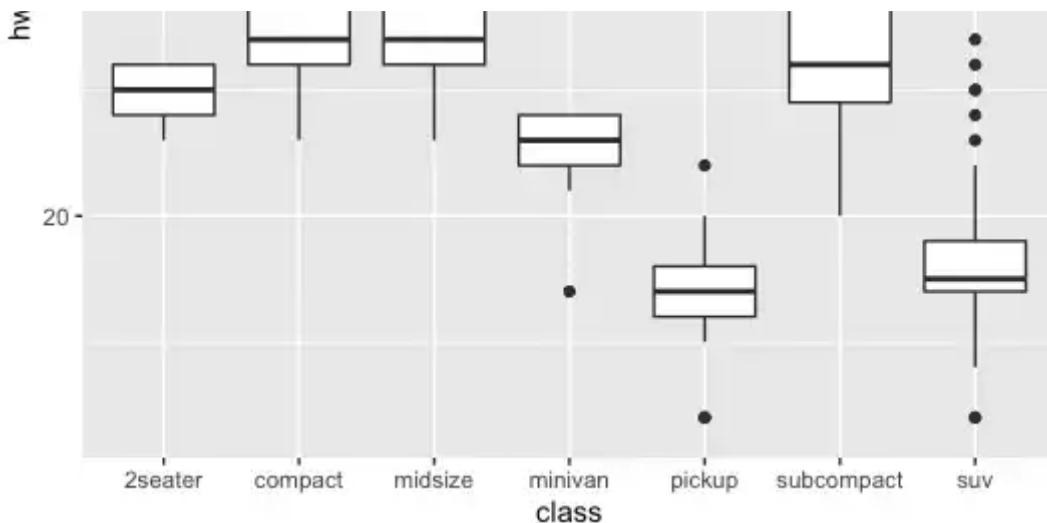
```
p + geom_boxplot()
```

Data Cleaning

Section 3: Missing Data, Outlier Detection, and Principal Component Analysis (PCA)



LESSON 6: OUTLIER DETECTION



Multiple box plots with outliers

Here there are numerous instances of outliers broken up by class of car.

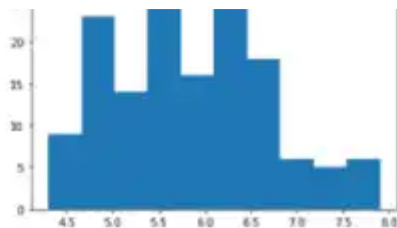
In Python, you will use the Matplotlib package to help visualize data. In these examples, you did not filter by the species. Instead, you have just taken all the values of the column to create the histogram and box plot. You can see how to create the same scatterplot as above in R with each color representing a species of iris. For more information about Matplotlib, check out <https://matplotlib.org> for examples on how to create some really great visuals.

Data Cleaning

Section 3: Missing Data, Outlier Detection, and Principal Component Analysis (PCA)

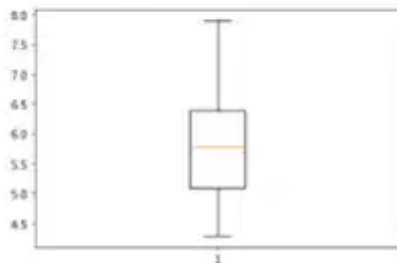


LESSON 6: OUTLIER DETECTION



```
In [4]: #Box plot
boxplt, ex2 = plt.subplots()
ex2.boxplot(iris_df.iloc[:,0])

Out[4]: {'whiskers': [matplotlib.lines.Line2D at 0x15c1d6dc790],
          'caps': [matplotlib.lines.Line2D at 0x15c1d6dca0],
          'boxes': [matplotlib.lines.Line2D at 0x15c1d6dc430],
          'medians': [matplotlib.lines.Line2D at 0x15c1d6eb550],
          'fliers': [matplotlib.lines.Line2D at 0x15c1d6eb850],
          'means': []}
```



Python code creating histogram and box plots

[PREVIOUS](#)[NEXT](#)