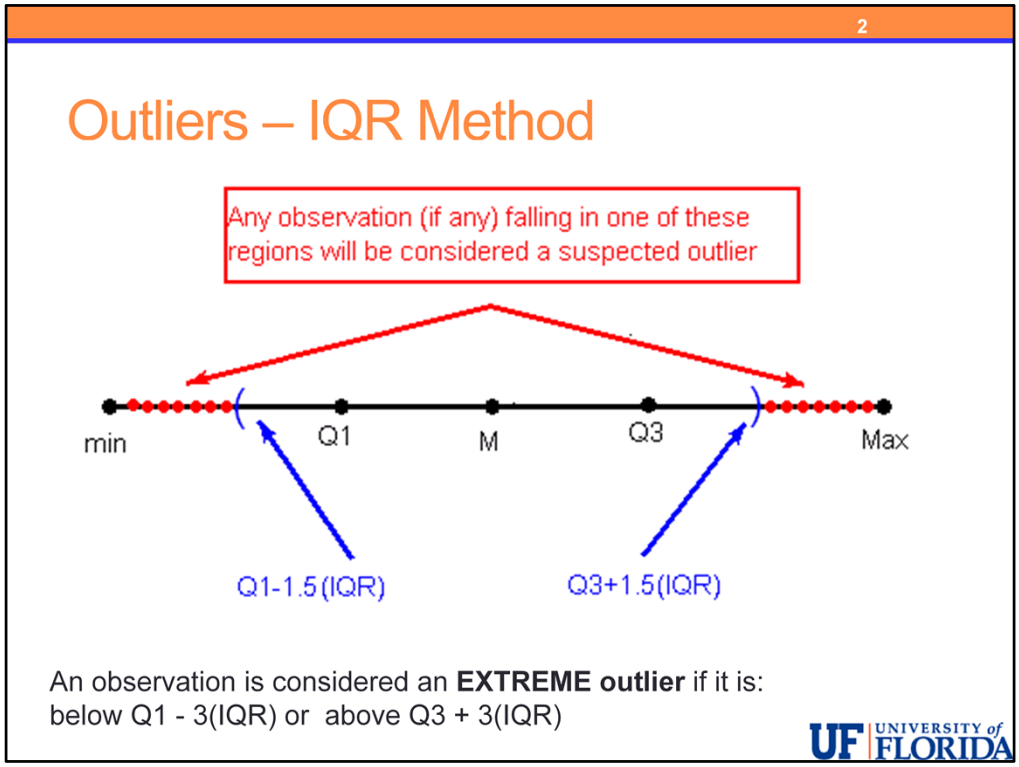
BOXPLOTS

One Quantitative Variable

Unit 1: Exploratory Data Analysis

UF UNIVERSITY of FLORIDA

Boxplots are a graphical display based upon the five-number summary and are very commonly used to compare the distributions of a quantitative variable for multiple groups.

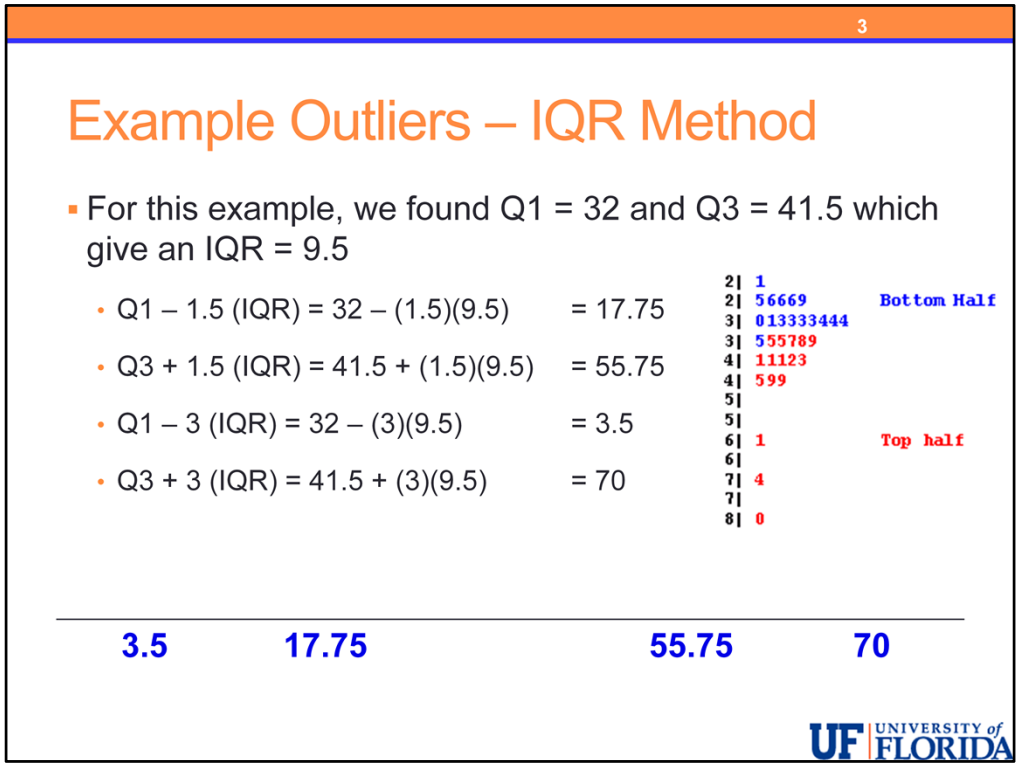


Recall the five-number summary consists of the minimum, Maximum, first and third quartiles (Q1 and Q3) and the median.

These values break down the data into four equal parts of 25% each.

In addition, recall the method for locating the potential and extreme outliers.

We need to know the five-number summary, and any outliers along with their closest non-outlier neighbor before proceeding to sketch a boxplot for a given dataset.



For the age of best actress Oscar winners data, we have the following calculations to determine the outliers.

Once we calculate all four values, we can place them in order on a line representing the data.

For this data our lower extreme outliers would need to be smaller than 3.5 and our lower potential outliers would need to be smaller than 17.75.

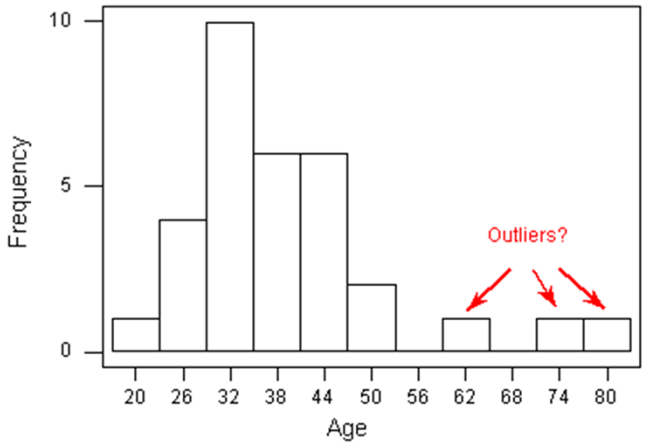
Our upper extreme outliers would need to be larger than 70 and our upper potential outliers would need to be larger than 55.75.

In the stemplot we can see that the smallest or minimum age was 21 and so we don't have any outliers on the lower end for this distribution.

However, on the high end, we have values of 80 and 74 (which would be classified as an extreme outliers) and 61 (which would be classified as a potential outlier).

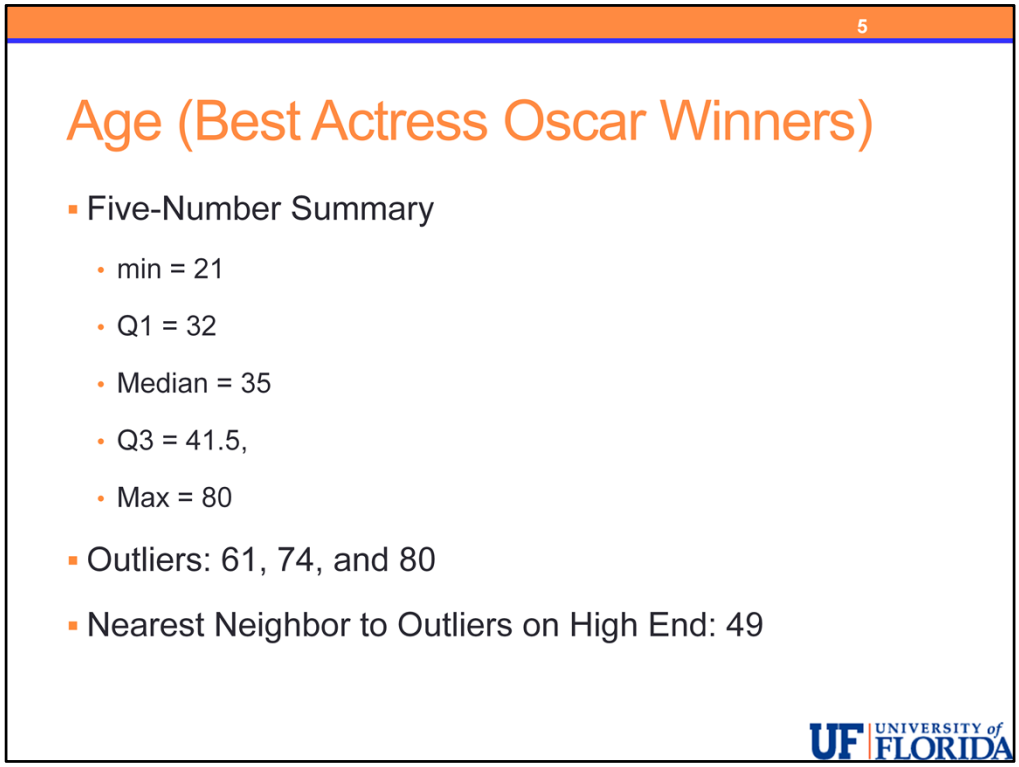
The next value in the data, the next nearest neighbor to the potential outlier of 61, is an age of 49. We will need this value when we create the boxplot.

Outliers – IQR Method



UF UNIVERSITY of FLORIDA

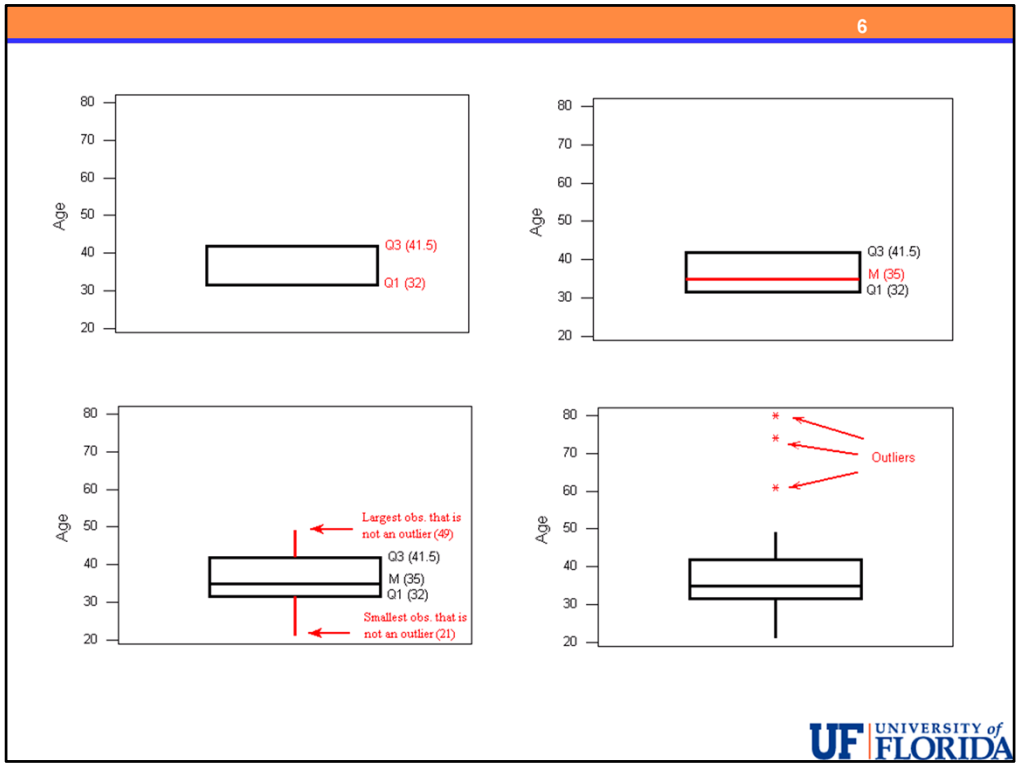
Here is the histogram for this data. The outliers are clearly visible. Our calculations and display seem to agree. This is often but not always the case.



To sketch a boxplot we need

- To know The five-number summary
- To identify any outliers (although it isn't necessary it is good to also know which are extreme)
- And to know the nearest neighbor(s) to any outliers.

Here is a review of the values for this data



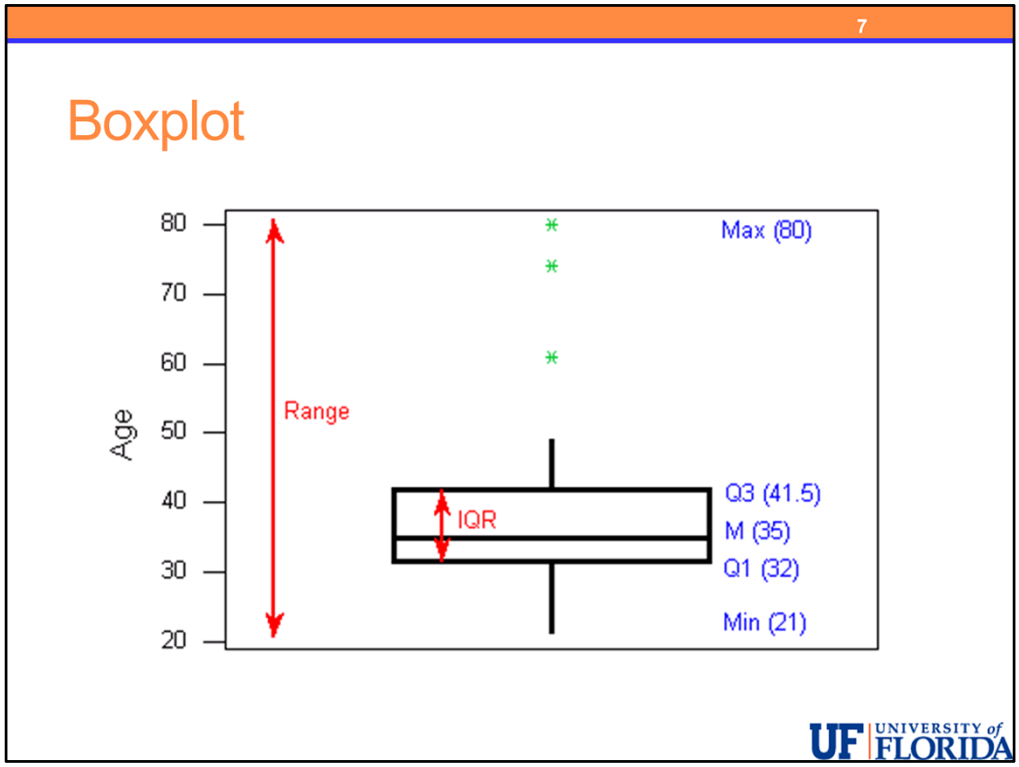
To draw the boxplot:

The central box spans from Q1 to Q3. In our example, the box spans from 32 to 41.5. Note that the width of the box has no meaning.

A line in the box marks the median M, which in our case is 35.

Lines extend from the edges of the box to the smallest and largest observations that were not classified as suspected outliers (using the 1.5xIQR criterion). In our example, we have no low outliers, so the bottom line goes down to the smallest observation, which is 21. Since we have three high outliers (61,74, and 80), the top line extends only up to 49, which is the largest observation that has not been flagged as an outlier.

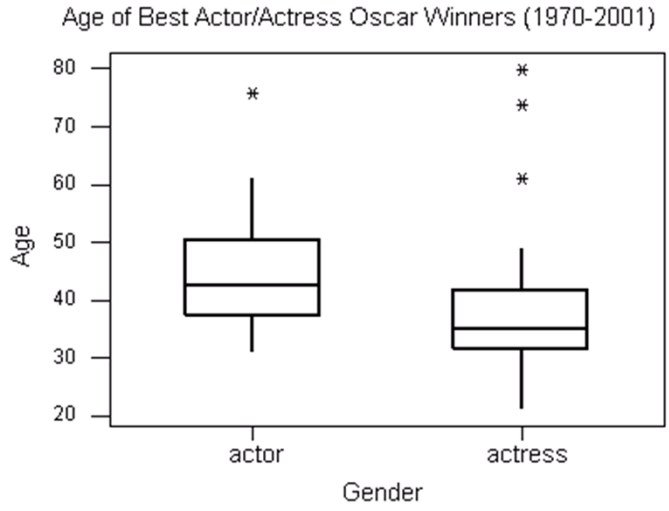
outliers are marked with asterisks (*).



To summarize: the following information is visually depicted in the boxplot:

- the five number summary (blue)
- the range and IQR (red)
- outliers (green)

Side By Side Boxplots



UF UNIVERSITY of FLORIDA

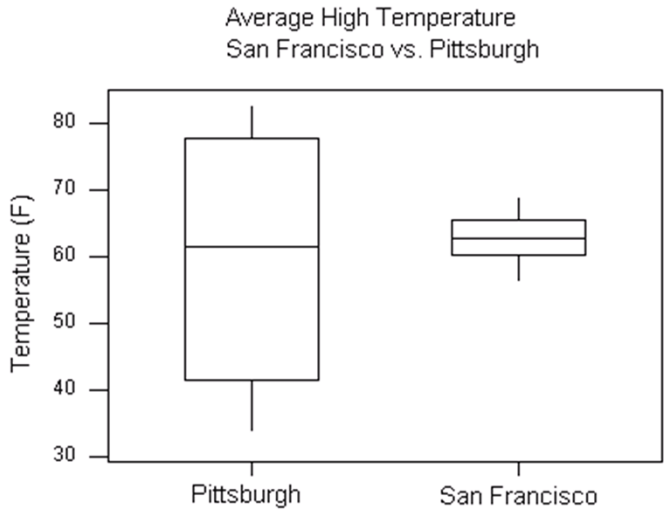
We mentioned that boxplots are very useful for comparing groups.

Here is a comparison of the distributions of ages of best actor and actress Oscar winners.

The primary difference is that overall, the distribution of ages is higher for actors than actresses.

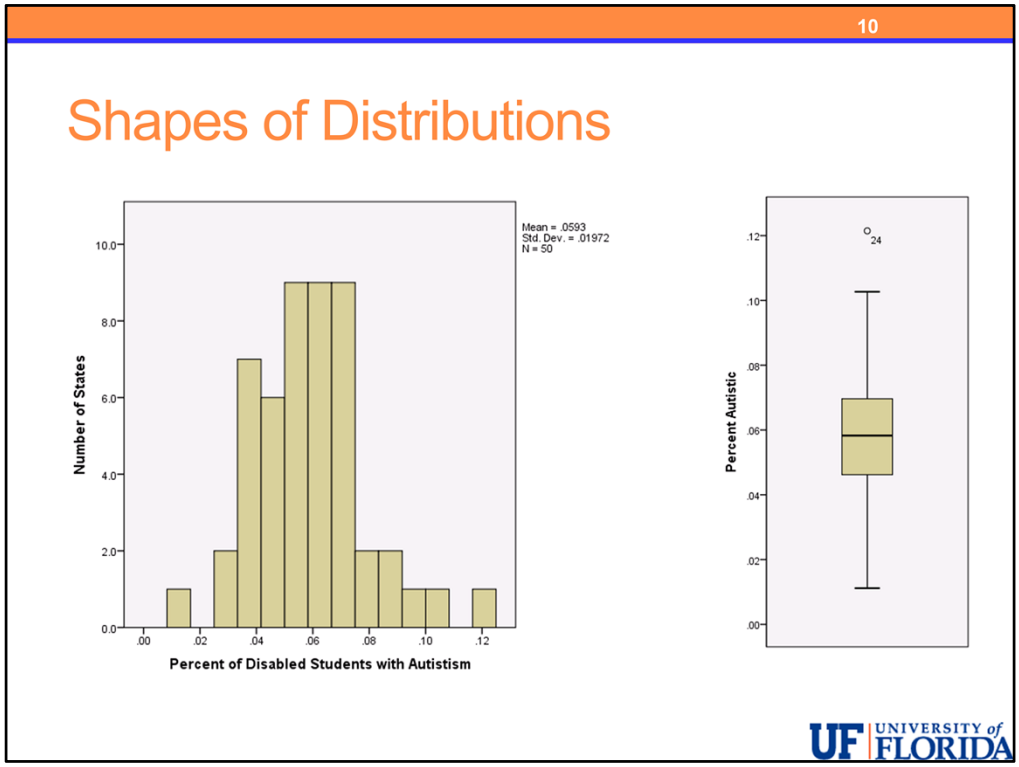
The variation is somewhat similar. Overall the females range is larger, however their IQR is slightly smaller than that of males. If we ignored the outliers, the variation in the two distributions would be fairly similar.

Side By Side Boxplots



UF UNIVERSITY of FLORIDA

Here, the primary difference is in the variation. In Pittsburgh, the weather varies greatly, however in San Francisco, the temperature is much more stable.



Before moving on we would like to point out a few comments about shape.

It is difficult if not impossible to tell whether a distribution is unimodal from a boxplot. However, we can see aspects of skewness and symmetry.

A perfectly symmetric distribution would have a boxplot in which the length of the whiskers (outer quartiles), including any outliers, are exactly the same length and the median splits the box exactly in half.

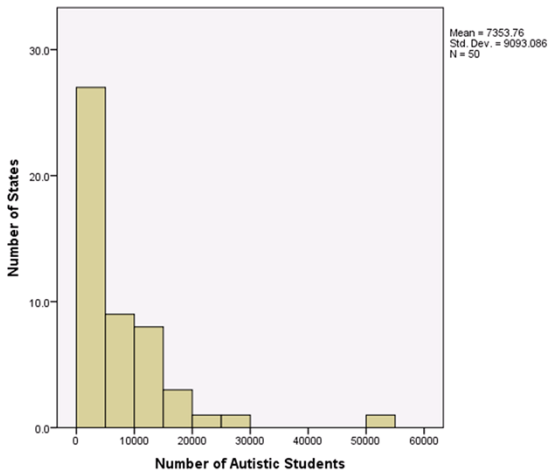
The closer the boxplot comes to satisfying those requirements, the closer the distribution is to being symmetric.

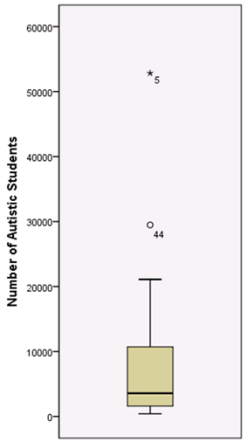
In this case, we have a fairly symmetric distribution where the lengths of the whiskers are about the same and the median splits the box near the center.

Notice that the tighter the central 50% in the box, the more gathered the data will be near the center.

It is an excellent exercise to view the histograms and boxplots for a wide variety of shapes of data to see how they are connected.

Shapes of Distributions





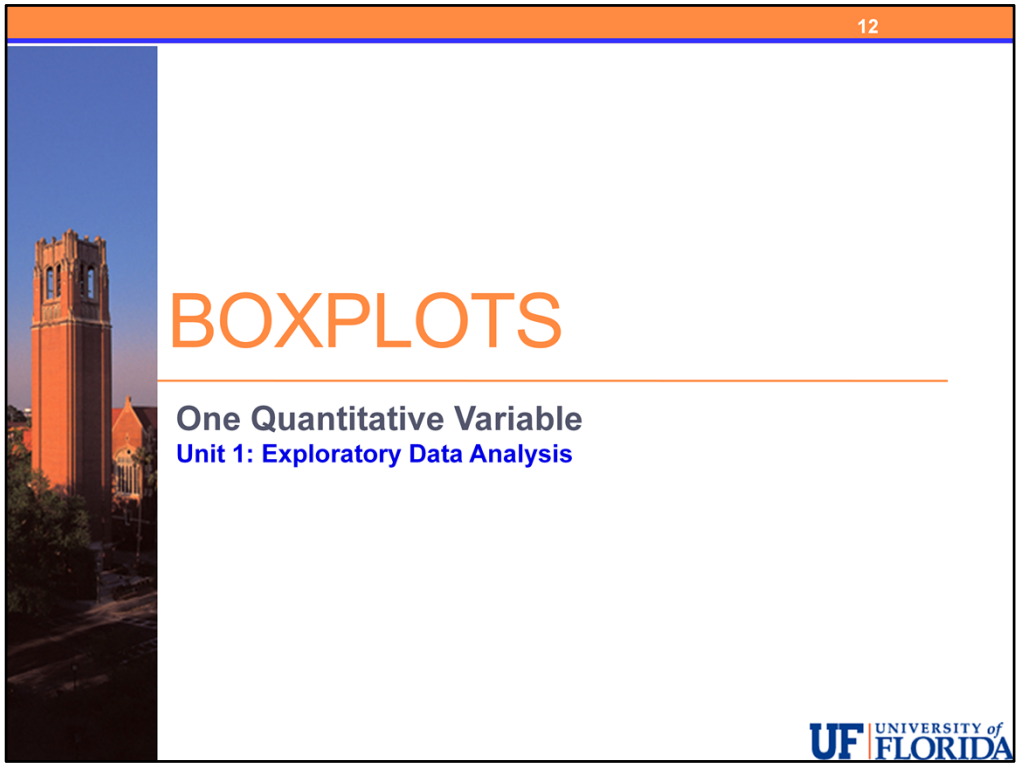
UF UNIVERSITY of FLORIDA

Here we have a skewed right distribution. The lower tail and lower half of the box are much smaller than their upper counter parts.

Each division represents 25% of the data and so the long upper half of the box and the long upper tail with two outliers shows the data in those areas are more stretched out than for the lower half of the box and lower whisker.

50% of the data is below the median in a very small range and 50% of the data are between the median and the maximum – a MUCH larger range.

Learning to understand the information displayed in these two types of displays is critical to exploratory data analysis for one quantitative variable.



This concludes the main content of our discussion on exploratory data analysis for one quantitative variable.

Before moving on to relationships between variables we will briefly discuss a related and important special topic, normal distributions.