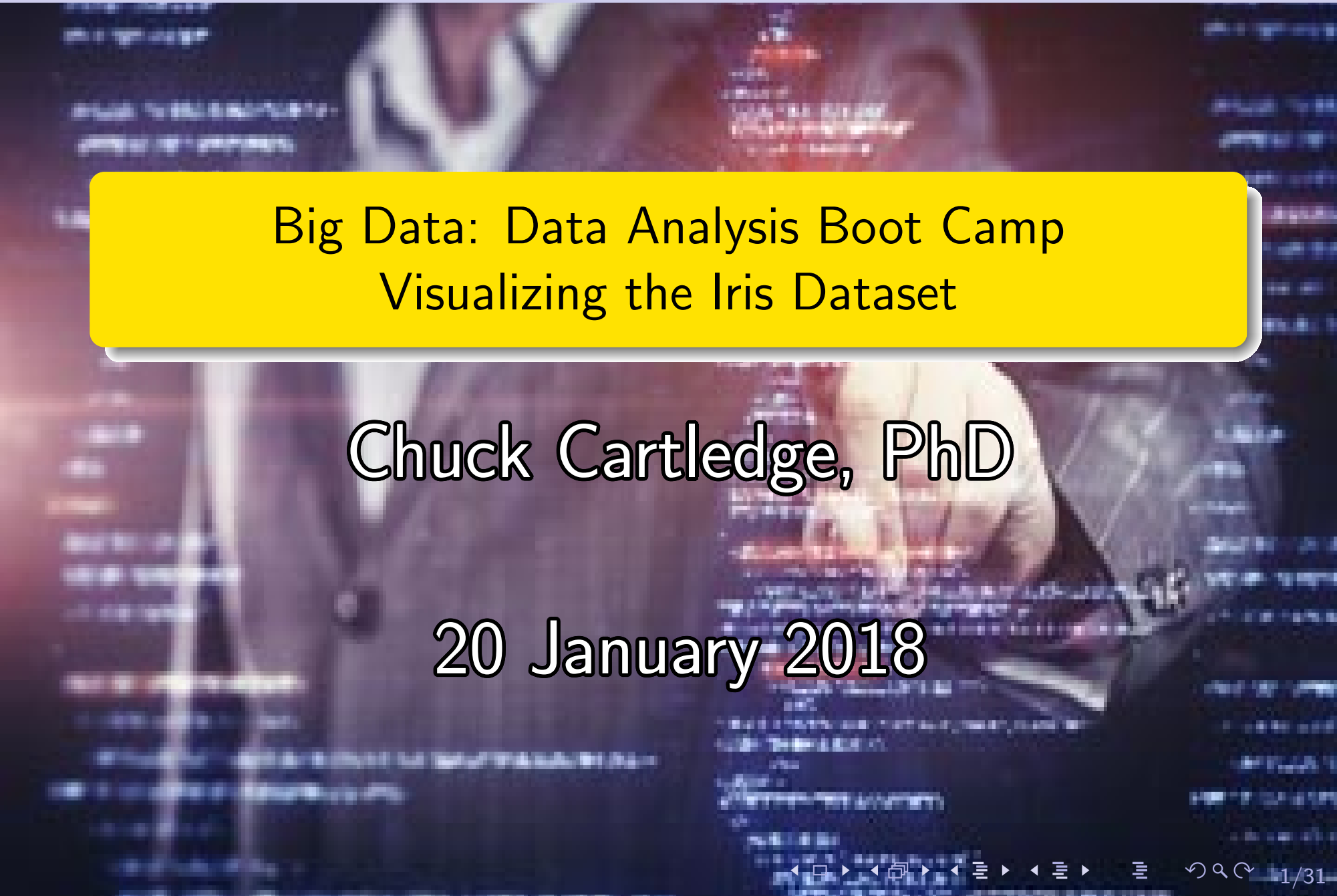


Big Data: Data Analysis Boot Camp
Visualizing the Iris Dataset

Chuck Cartledge, PhD

20 January 2018

Table of contents (1 of 1)

1. Intro.
2. Histograms
   - Background
   - Iris data
3. Scatter plots
   - Iris data
4. Box plots
   - Iris data
5. Outliers
   - iris data
6. Hands-on
7. Q & A
8. Conclusion
9. References
10. Files

What are we going to cover?

We're going to talk about:

- Visually explore the iris dataset.
- See how "messy" data can affect the presentation.

What is a histogram?

"Consider a series of rectangles on equal base c and whose heights are respectively the successive terms of the binomial $(p+q)^n * \frac{\alpha}{c}$*, where* $p+q=1$*."*

K. Pearson [2]

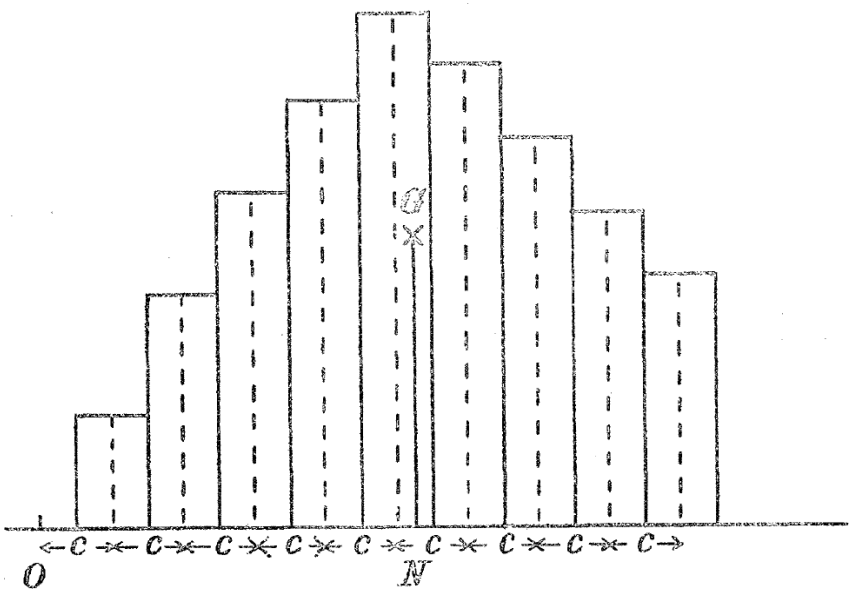


Image from [2].

In layman's terms:

1. Take the range of data and divide into equal range bins
2. Count the number of pieces of data (frequency) in each range bin
3. Plot the count vs. the range bin

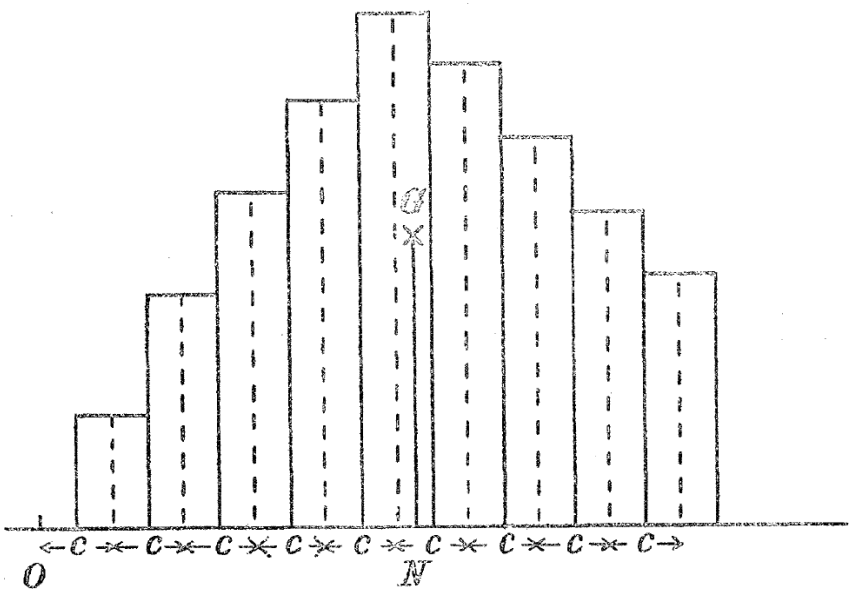
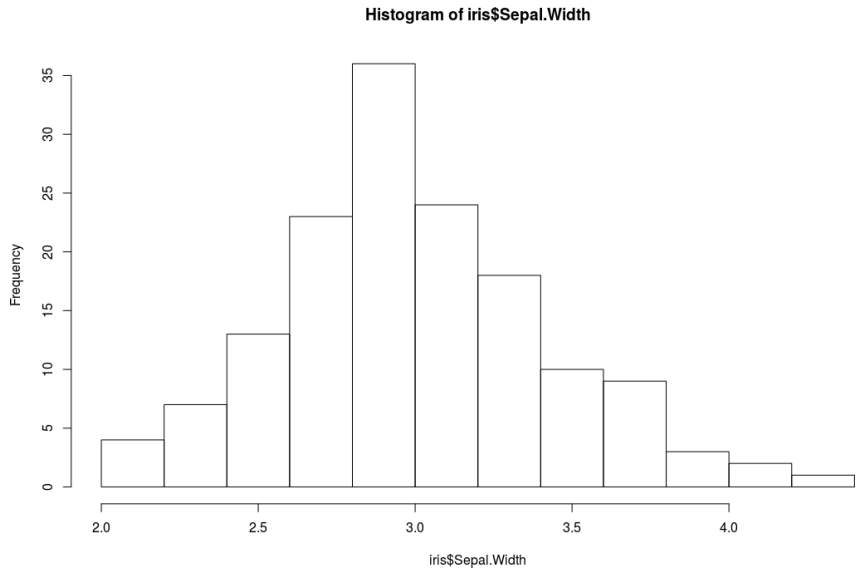


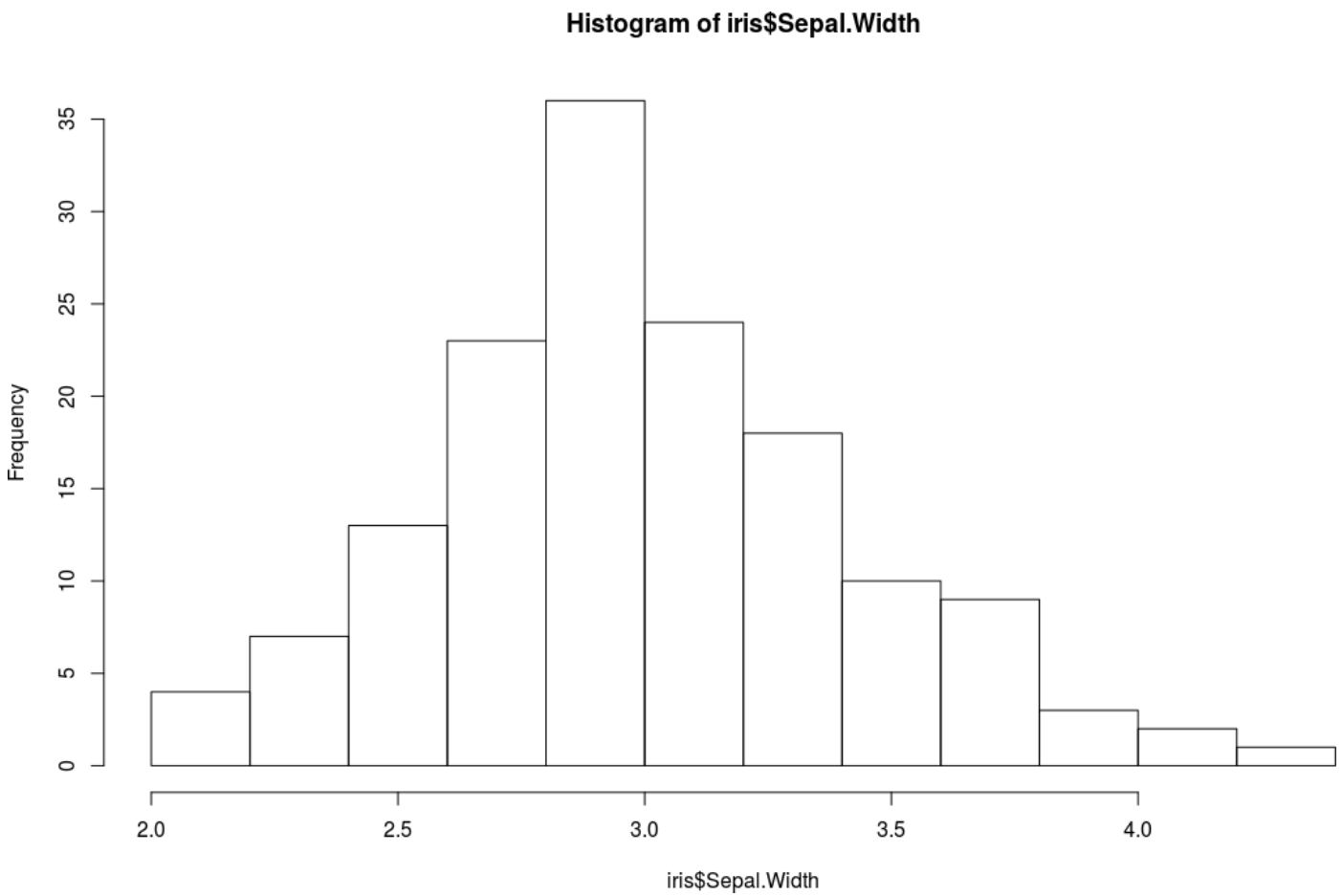
Image from [2].

Looking at sepal widths

```
data(iris)
hist(iris$Sepal.Width)
```

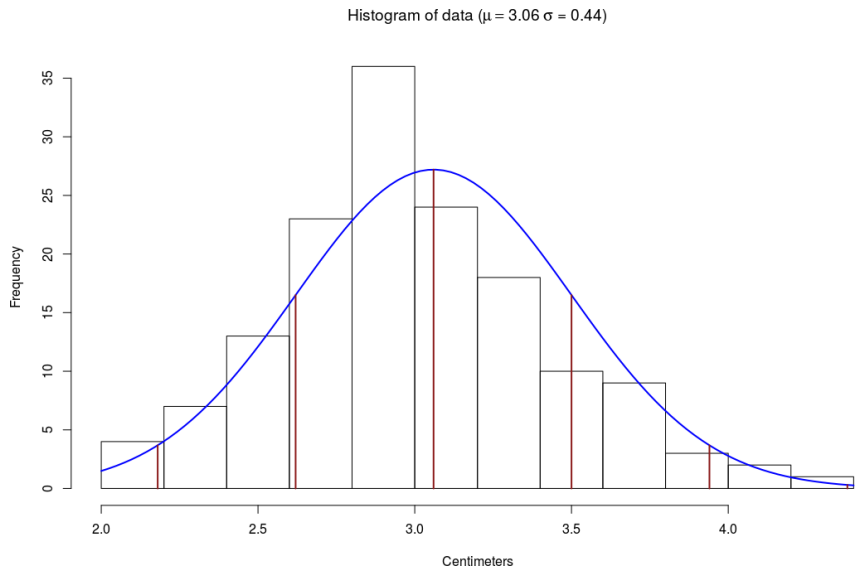


Same image.


Histogram of iris$Sepal.Width
Frequency
0
5
10
15
20
25
30
35
2.0
2.5
3.0
3.5
4.0
iris$Sepal.Width

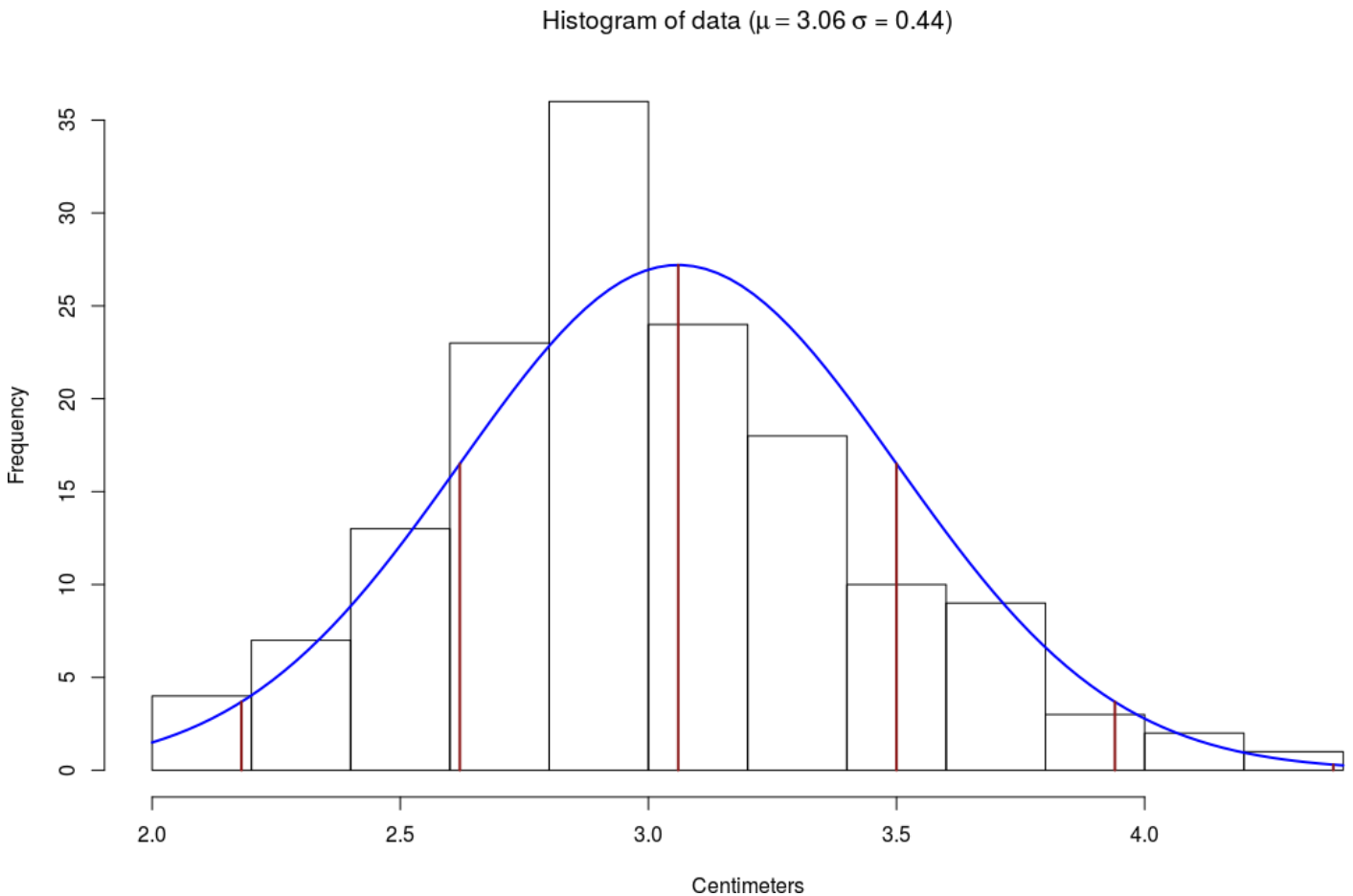
An annotated look at sepal widths

1. Compute the mean and standard deviation
2. Add a "normal" (a.k.a., Gaussian) distribution curve
3. Add $\pm 3\sigma$ vertical lines



Attached file.

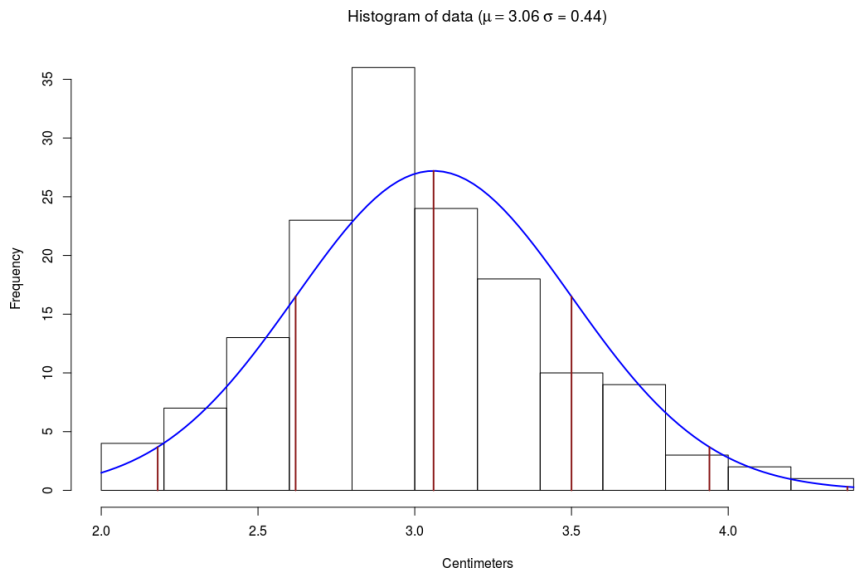
Same image.



Attached file.

What is a "normal" distribution?

1. Ideas and base equations attributed to Carl Friedrich Gauss and Abraham de Moivre[1] (de Moivre is more general than Gauss')
2. Based on the idea of a central value μ and a variation from that value σ^2
3. Equation:
$$P(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$
4. The probability of x is dependent on μ and σ



Sigmas (σ) are important

Likelihood that a value exists based on a "normal" distribution is:

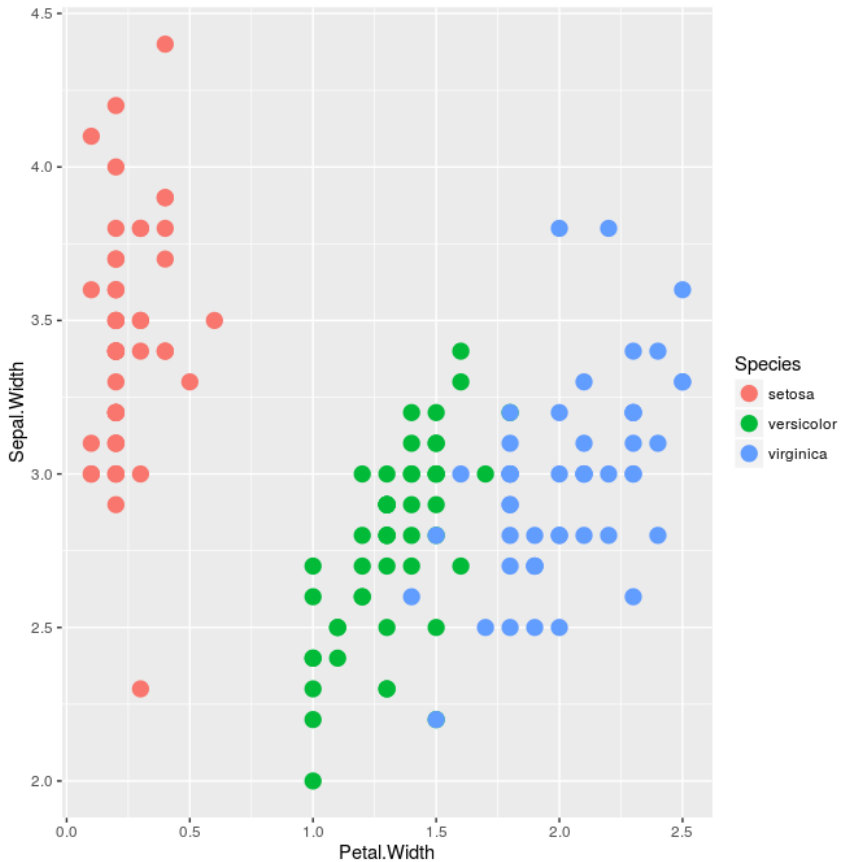
Range $\pm\sigma$	**Expected population within range**
0.5	38.29
1.0	68.26
1.5	86.63
2.0	95.44
2.5	98.75
3.0	99.73
3.5	99.95
4.0	99.99

https://en.wikipedia.org/wiki/68%E2%80%9395%E2%80%9399.7_rule

Choose any two values and see what they look like

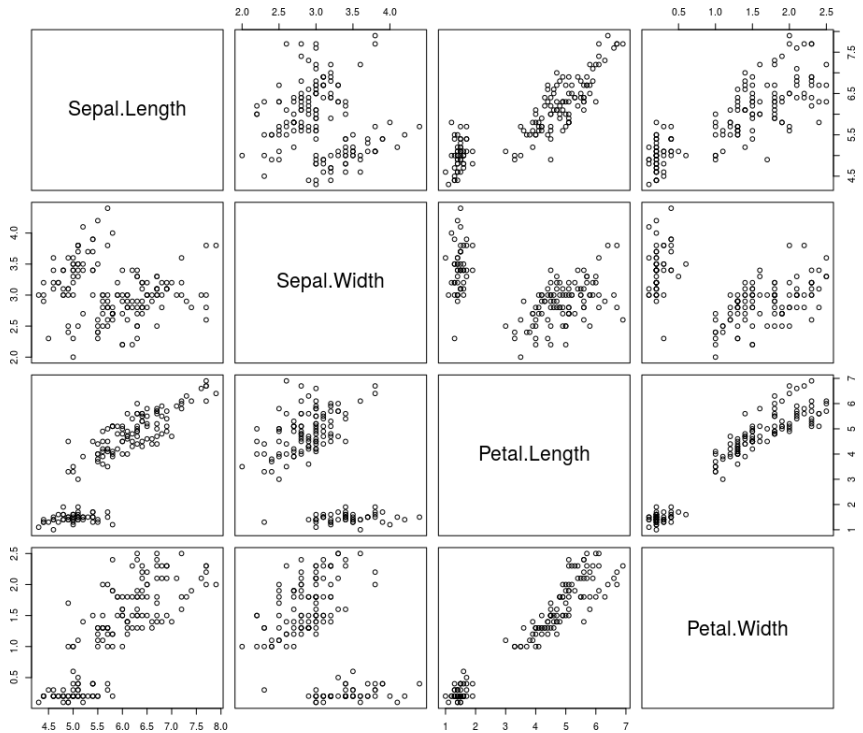
- Part of the data exploration toolset
- Used to visually identify, or verify correlation between attributes

```
1  library(ggplot2)
2  qplot(Petal.Width, Sepal.Width,
        data=iris, colour=Species,
        size=I(4))
```



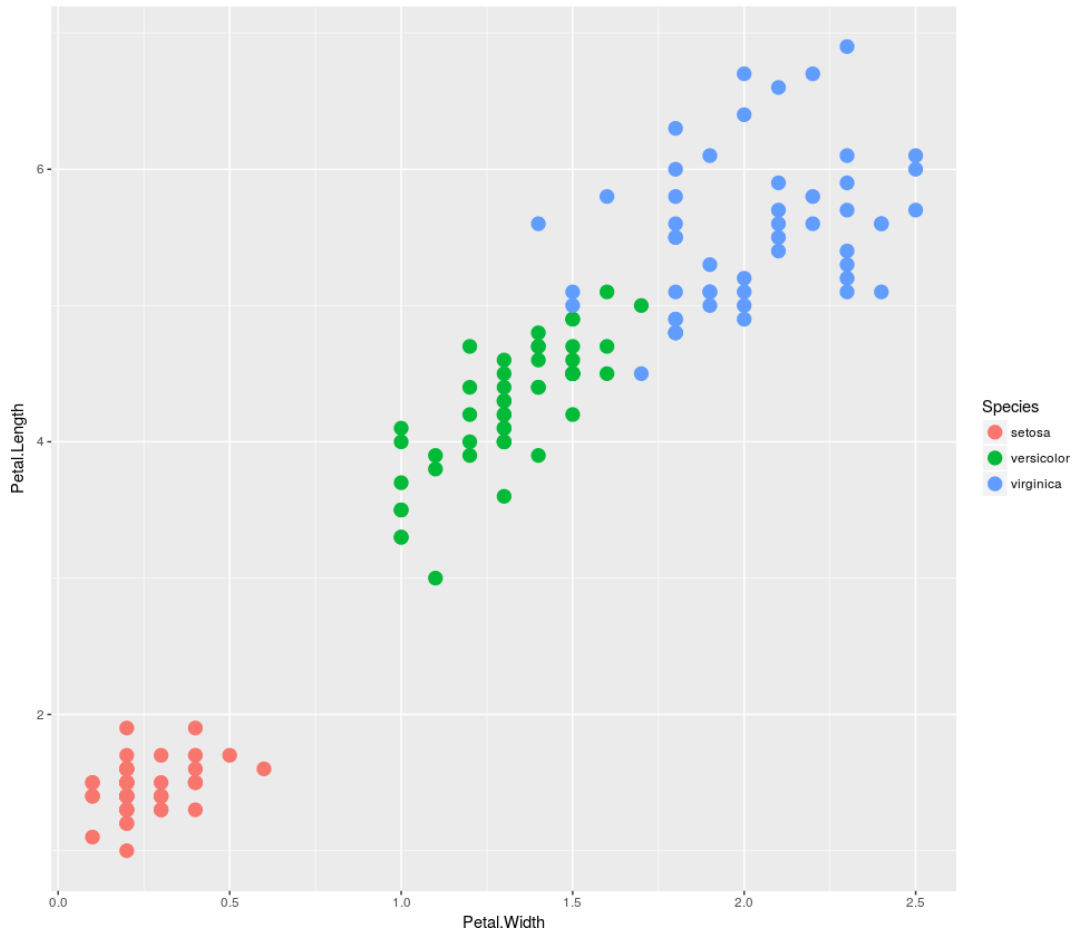
Sometimes you don't know which attribute to choose

- 2D plots are easy to understand
- 3D plots are harder to understand
- $> 3D$ requires special training
- How to choose which attributes are interesting?




```
pairs(~Sepal.Length+Sepal.Width+Petal.Length+Petal.Width,
data=iris)
```

Sometimes there are new insights




```
library(ggplot2); qplot(Petal.Width,Petal.Length, data=iris,colour=Species, size=I(4))
```

Background

Some terminology

- **Quartiles** of a ranked set of data values, divide the data set into four equal groups, each group comprising a quarter of the data
- **Q1**: splits off the lowest 25% of data from the highest 75%
- **Q2**: cuts the dataset in half (median)
- **Q3**: splits off the highest 25% of data from the lowest 75%
- **IQR**: Interquartile range $= Q3 - Q1$
- **Lower fence** $= Q1 - 1.5 * IQR$
- **Upper fence** $= Q3 + 1.5 * IQR$

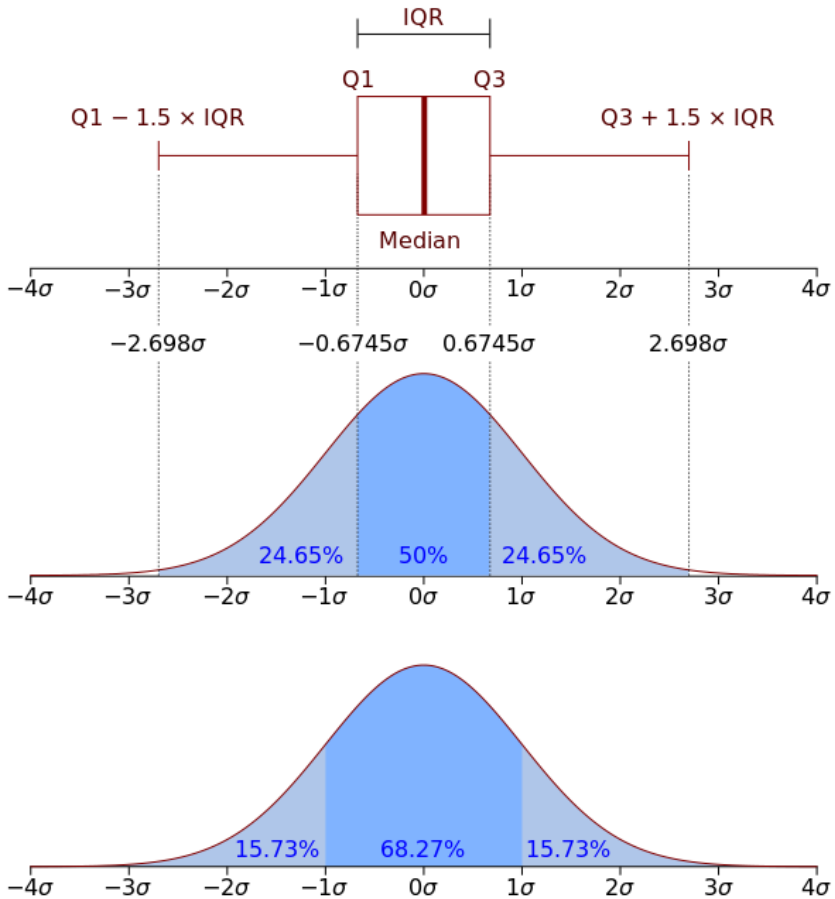


Image from [4].

Box plot visual

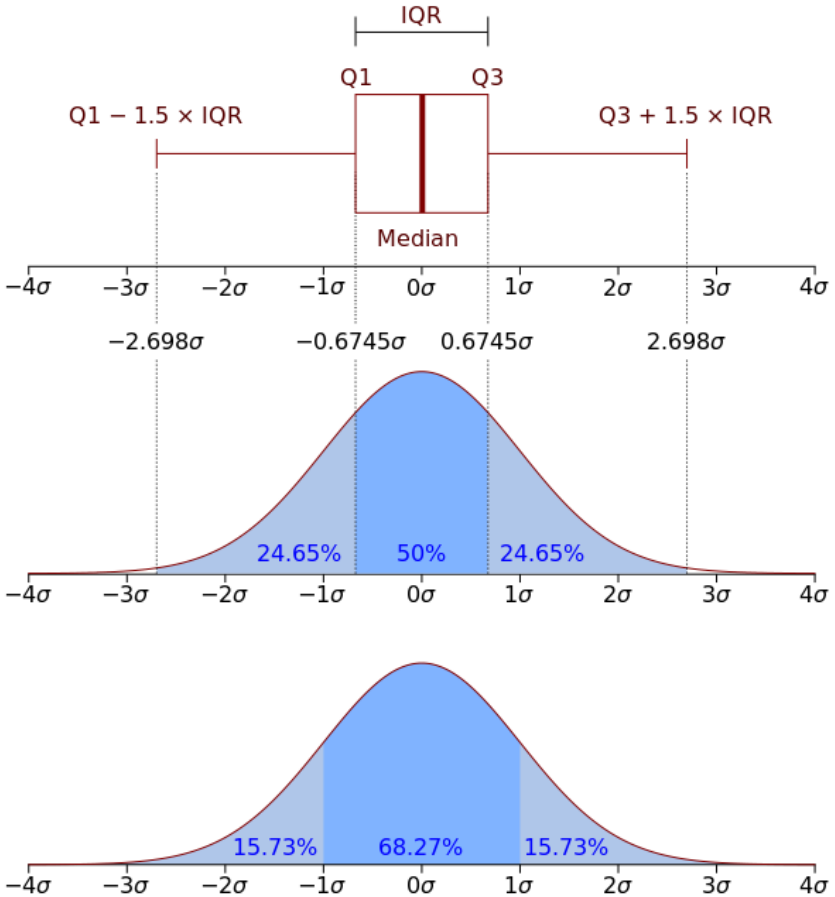
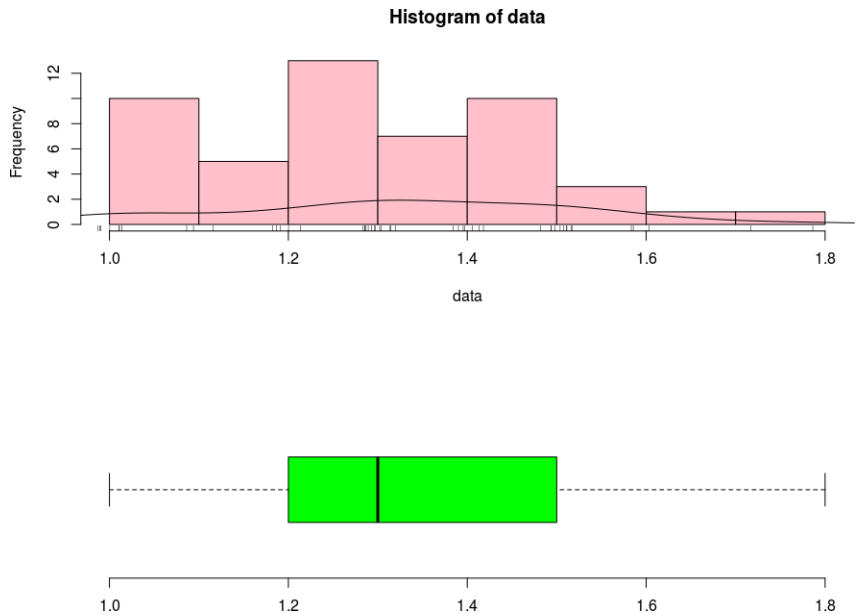


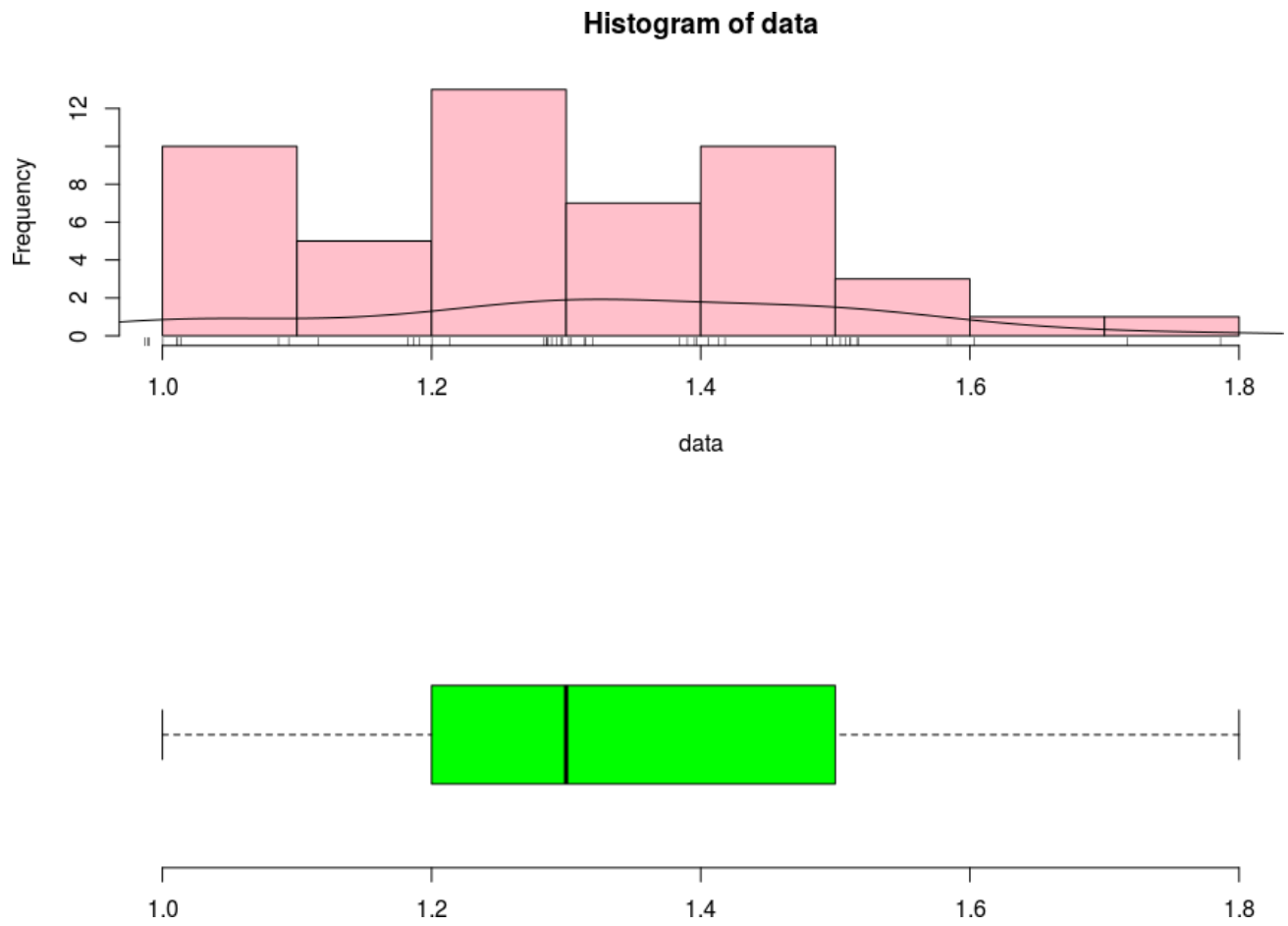
Image from [4].

Looking at versicolor petal widths

```
par(mfrow=c(2,1))
data <- iris[which(iris$Species ==
"versicolor"),"Petal.Width"]
hist(data,xlim=range(data), col = "pink", freq =
TRUE)
lines(density(data))
rug(jitter(data))
boxplot(data, horizontal=TRUE,
outline=TRUE,ylim=range(data), frame=FALSE, col
= "green1")
```



Same image.

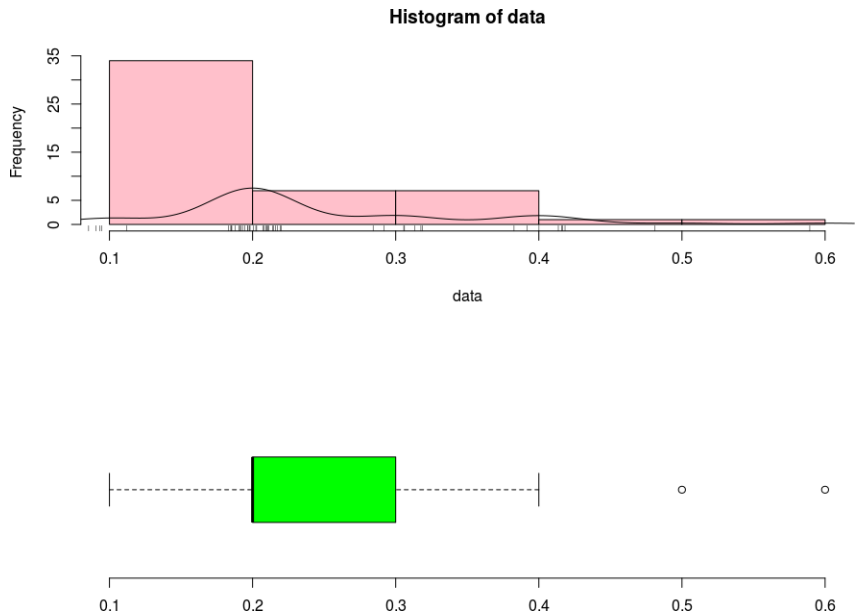

Histogram of data
Frequency
0
2
4
6
8
12
1.0
1.2
1.4
1.6
1.8
data
1.0
1.2
1.4
1.6
1.8

Looking at setosa petal widths

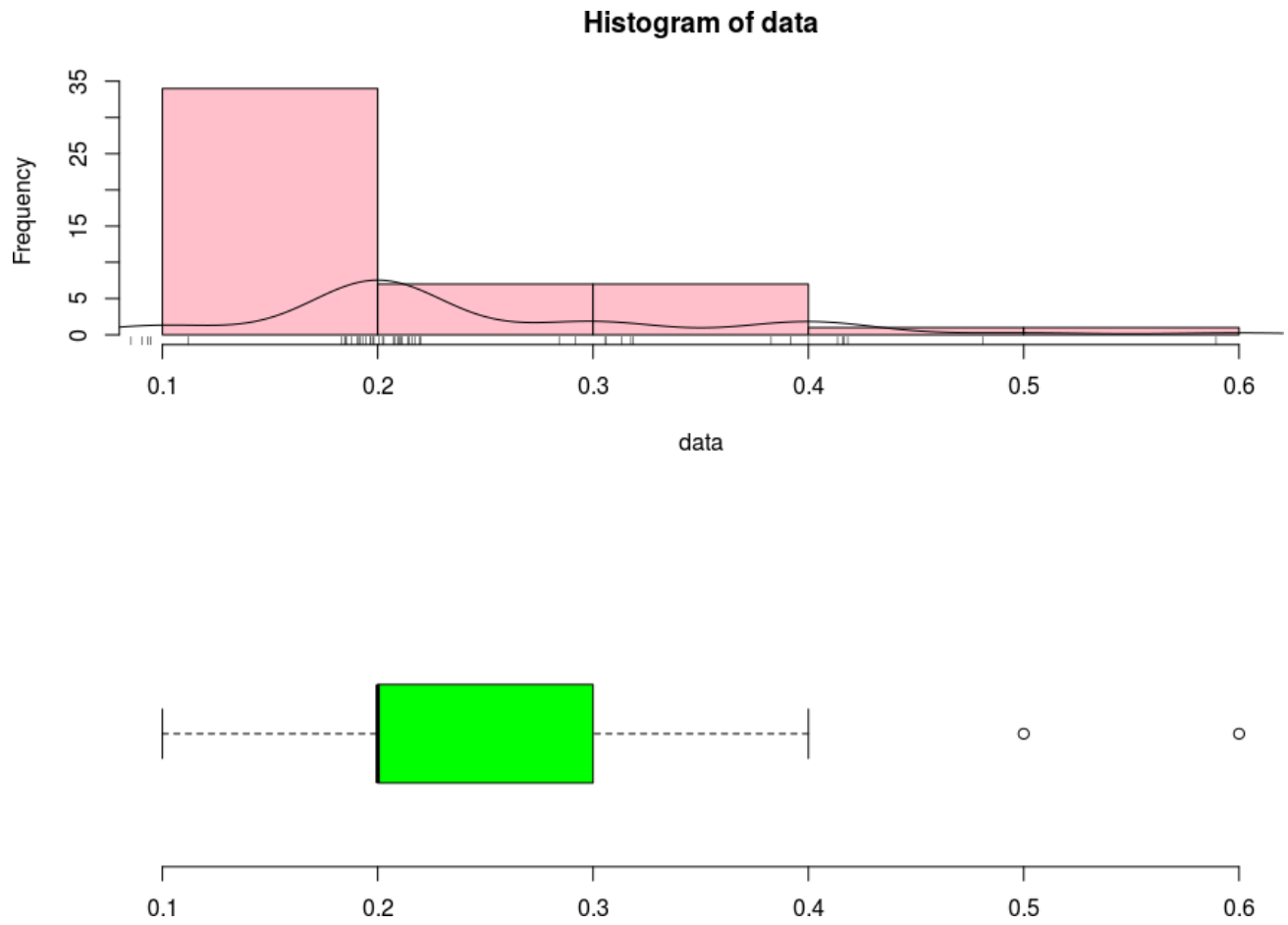
Change one line, and replot.

```
data <- iris[which(iris$Species ==
"setosa"),"Petal.Width"]
```

Now we see data points beyond the upper fence. They are generically known as "outliers."



Same image.



What are outliers?

A definition:

> *"Outliers are observations that do not follow the pattern of the majority of the data. Outliers in a multivariate point cloud can be hard to detect, especially when the dimension p exceeds 2, because then we can not longer rely on visual perception."*
>
> *Rousseeuw and Van Zomeren [3]*

The difficulty is defining a pattern in the data, and then defining what it means to not follow the pattern.

Bringing things together.

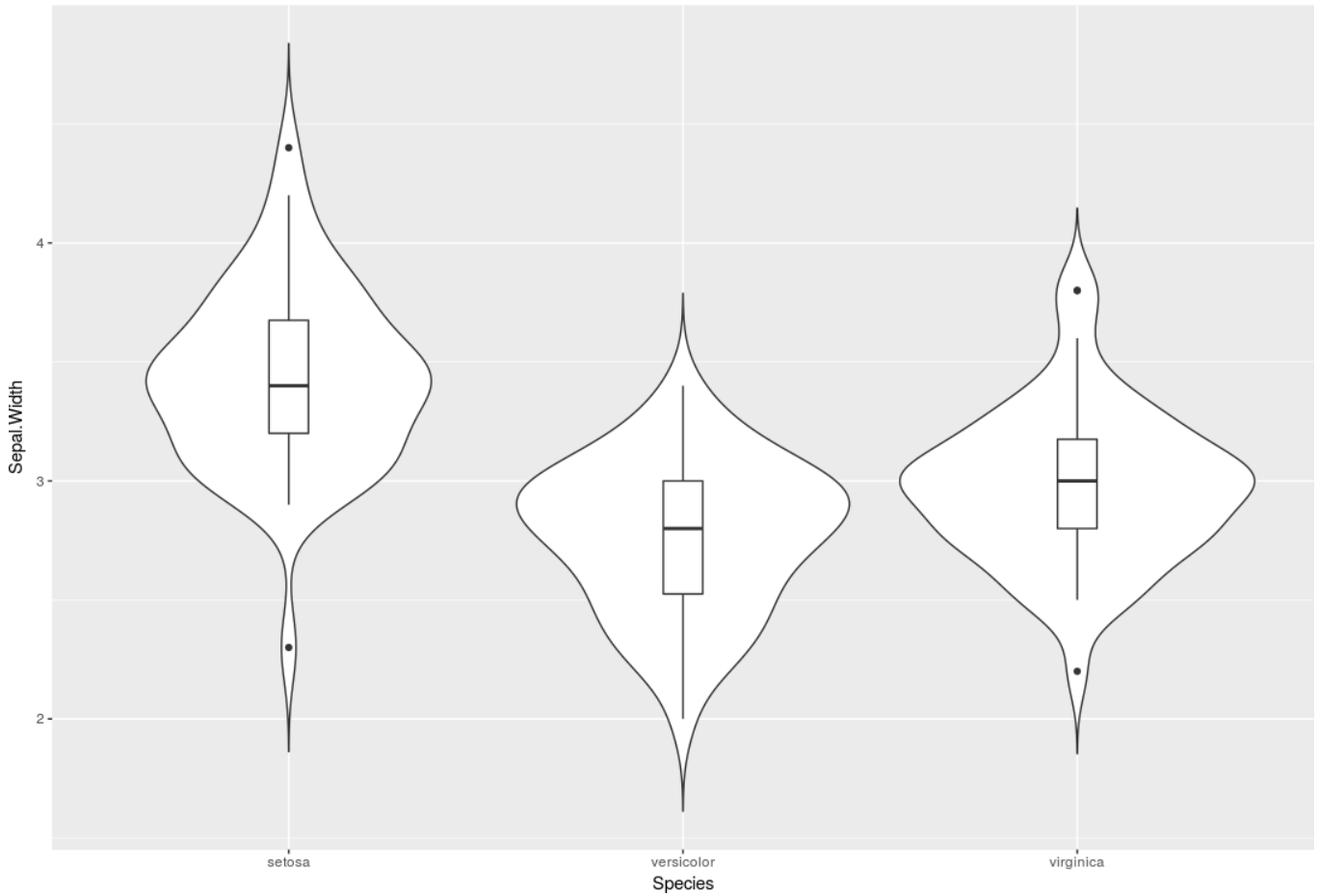
Ideas that come together into simple visualizations:

1. Density plot (curved lines on the previous histograms)
2. Density plots that may not be "normal" or "Gaussian" in shape
3. Box plots showing where the bulk of the data points are
4. Outliers are points that don't fit a pattern

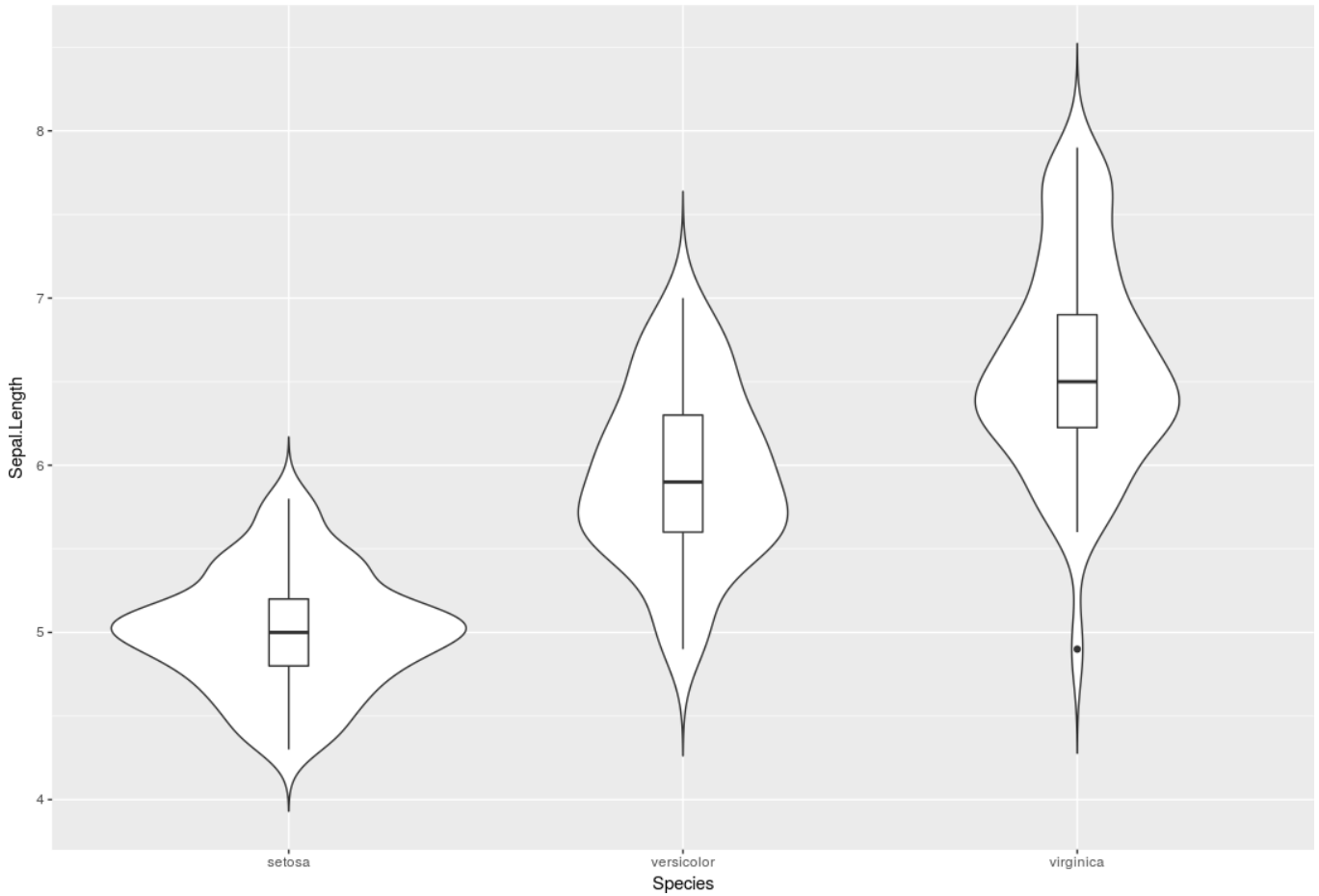
```
1 library(ggplot2)
2 ggplot(iris, aes(x=Species, y=Sepal.Width)) + geom_violin(trim=FALSE) + geom_
      boxplot(width=0.1)
```

Change the y value.

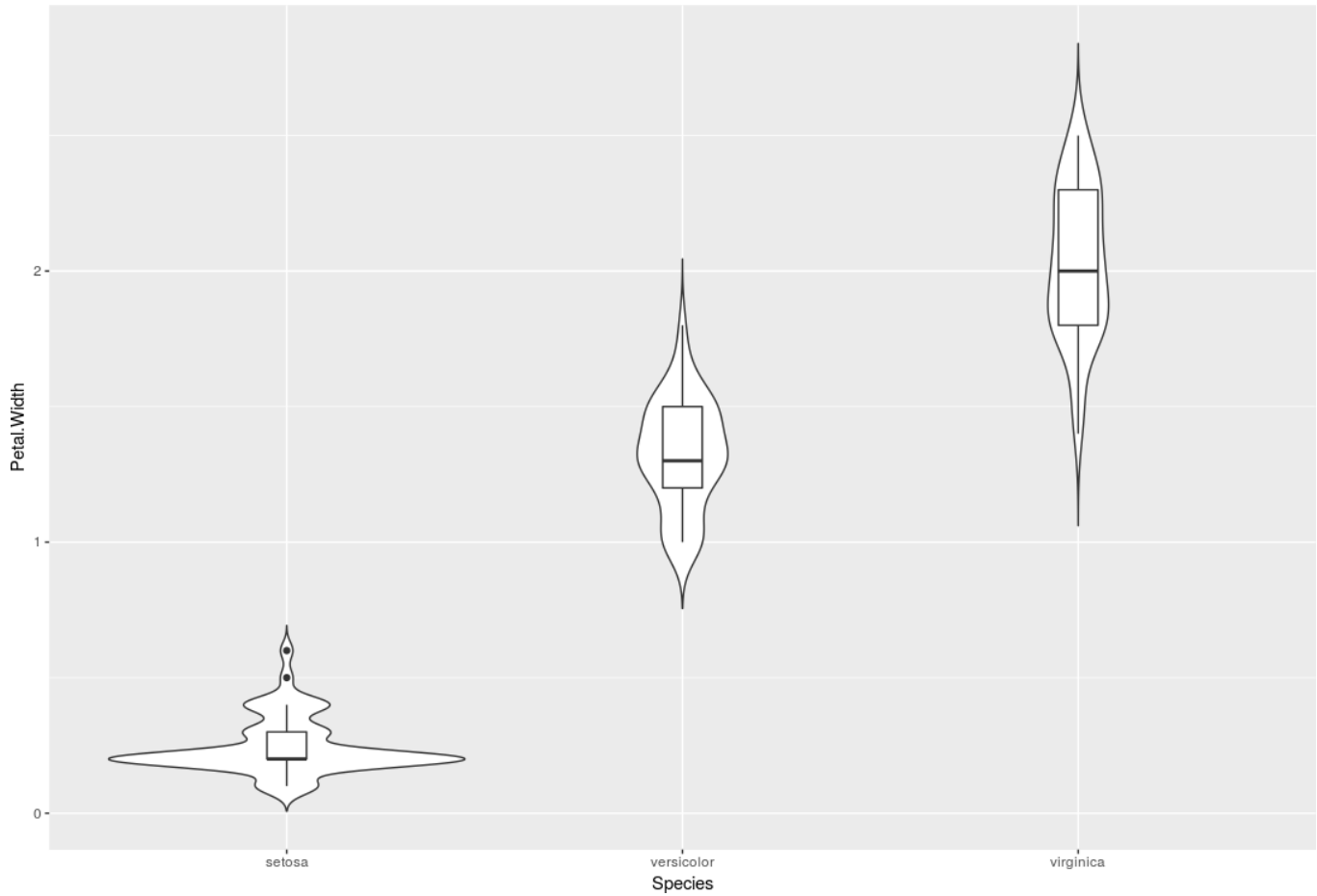
Violin plot of iris Sepal Width



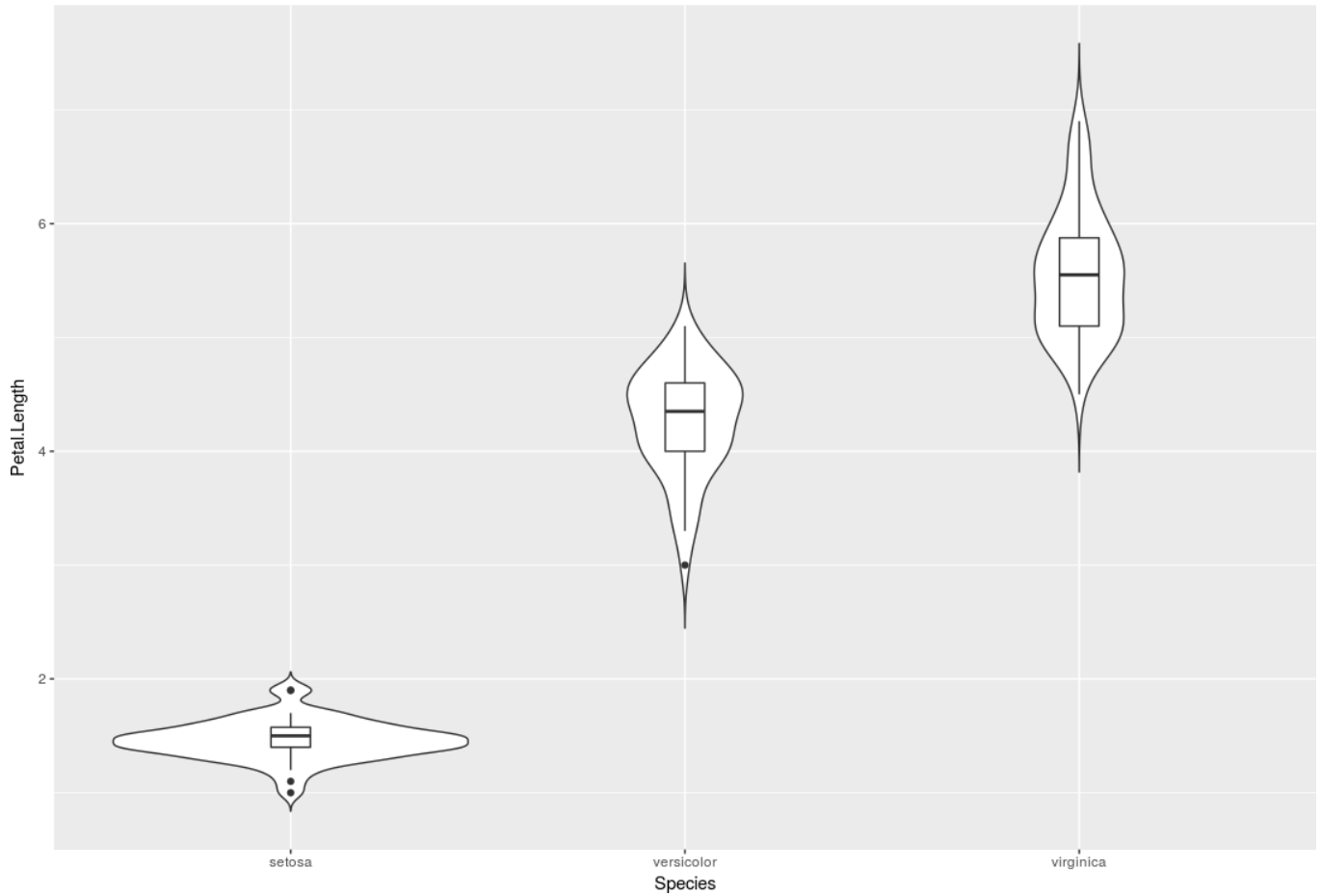
Violin plot of iris Sepal Length



Violin plot of iris Petal Width



Violin plot of iris Petal Length



Some simple exercises to get familiar with data visualization

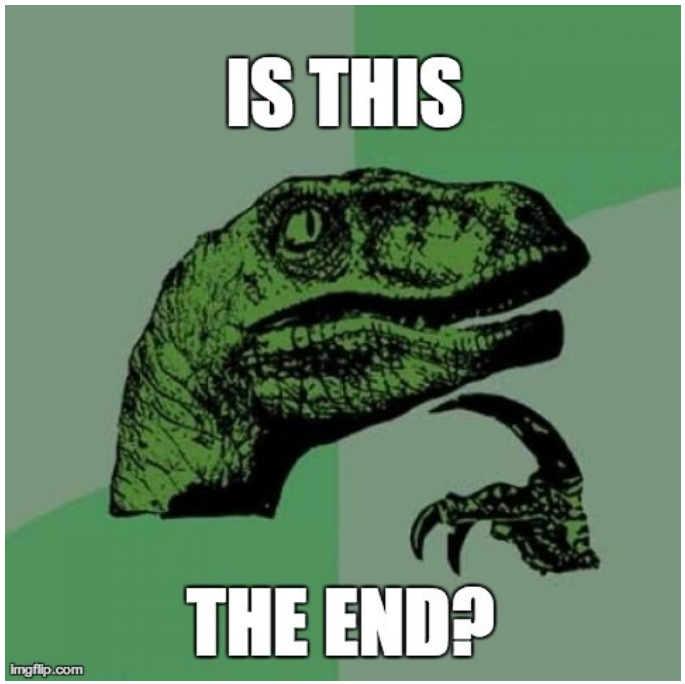
1. Does a histogram of iris petal length support a "normal" distribution (qualitative vice quantitative)
2. What does the scatterplot on page 12 say about using widths as a classification criteria?
3. Which combination of sepal and petal, lengths and widths is best?
4. What is the purpose of the `rug` function on page 17?
5. How do the other iris species petal lengths compare to Versicolor on page 17?
6. Create a geom_violin plot of the built in mtcars dataset that show the relationship between number of gears and mpg

Q & A time.

Q: What's Dr. Presume's full name?
A: Dr. Livingston I. Presume.

What have we covered?

- Basic ways to visualize data
  1. Histograms
  2. Scatter plots
  3. Box plots
- Outliers and how they can affect data
- Looked at iris data using basic plotting functions and a little of the ggplot library



Next: LPAR Chapter 3, data visualization with Lattice

References (1 of 1)

[1] Abraham De Moivre, The doctrine of chances: or, a method of calculating the probabilities of events in play, vol. 1, Chelsea Publishing Company, 1756.

[2] Karl Pearson, Contributions to the Mathematical Theory of Evolution, Philosophical Transactions of the Royal Society of London. A **185** (1894), 71–110.

[3] Peter J Rousseeuw and Bert C Van Zomeren, Unmasking Multivariate Outliers and Leverage Points, Journal of the American Statistical Association **85** (1990), no. 411, 633–639.

[4] Wiki Staff, Quartile, `https://en.wikipedia.org/wiki/Quartile`, 2017.

Files of interest

1. Create annotated iris histogram 
2. Calculus to derive equation for "normal" distribution 
3. YouTube video about deriving the "normal" distribution: `https://www.youtube.com/watch?v=ebewBjZmZTw`
4. R library script file