Ulster University

COM662
Data Analytics

Dr Lu Bai
l.bai@ulster.ac.uk

ulster.ac.uk

Ulster University

Week 3 – Descriptive Statistics and Visualisation Using R

ulster.ac.uk

Week 3
Content

- Descriptive statistics
  - Measurement of central tendency
  - Measure of variability
- Data Visualisation
  - Histogram
  - Line plot
  - Bar plot
  - Scatterplot
  - Box and Whisker plot
- Exploratory data analysis

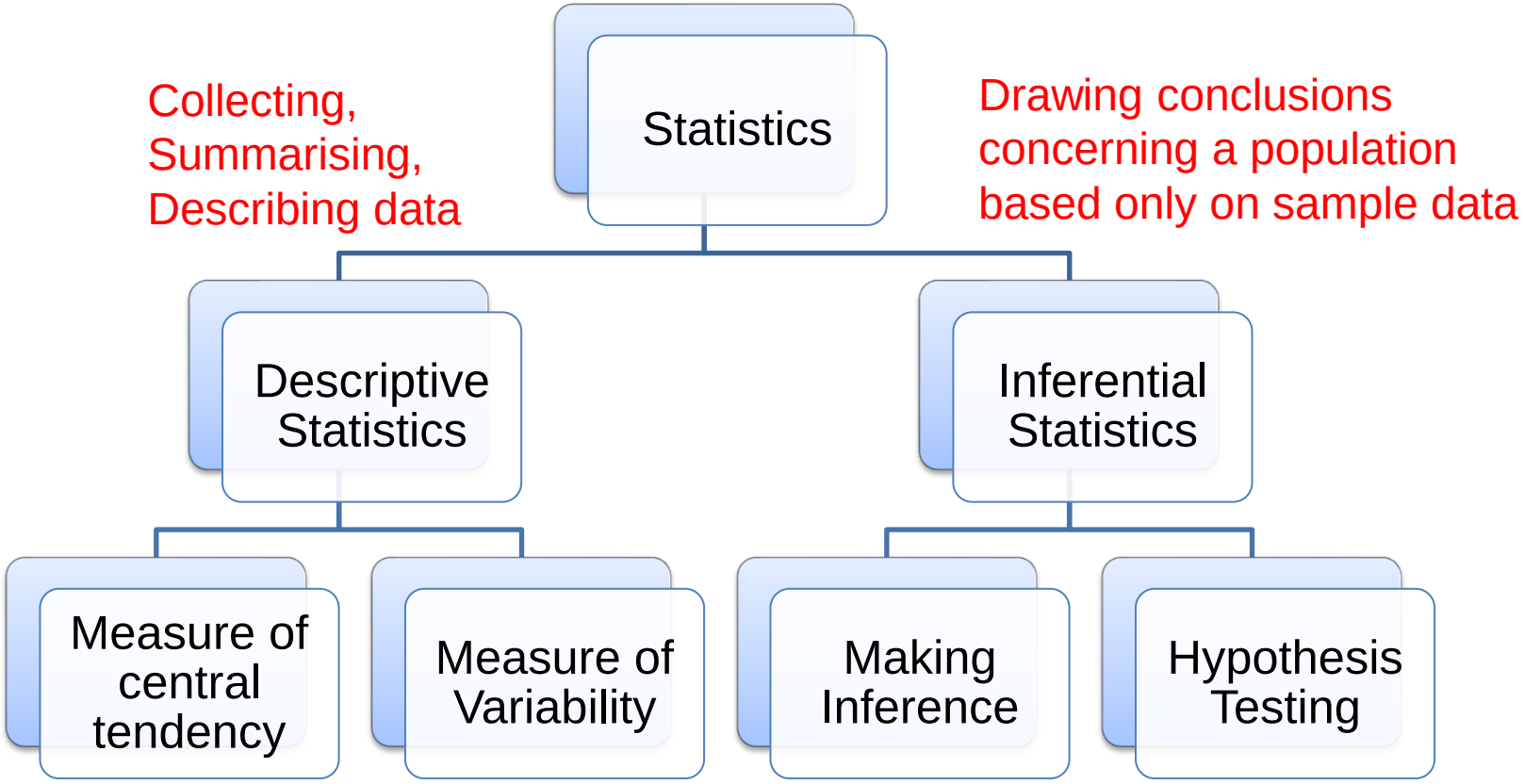
Ulster University

Statistics


Statistics
organization
descriptive
scientific
analysis
samples
collection
interpretation
data
statistical
Inference
theory
mathematical
probability
experiments
science
population
prediction
referring
discipline
pertaining
including
called
often
branch
design
parameters
mathematics
provides
also
concerned
consider
inferential
word
wide
well
statistic
total
used
hypothesis
sampling
presentation
focus probabilities
misleading
new
way
opposite
quality
plural
number
deals
whose
roughly
surveys
tested
describe
leads
without
available
deduce
part
problems
considered
social
working
useful
ways
improve
process
soundness
terms
element
using
modeled
expertise
sciences
people
gained
models
direct
thinking
planning
inferences
companies
usually
collecting
help
methods
consultants
statistician

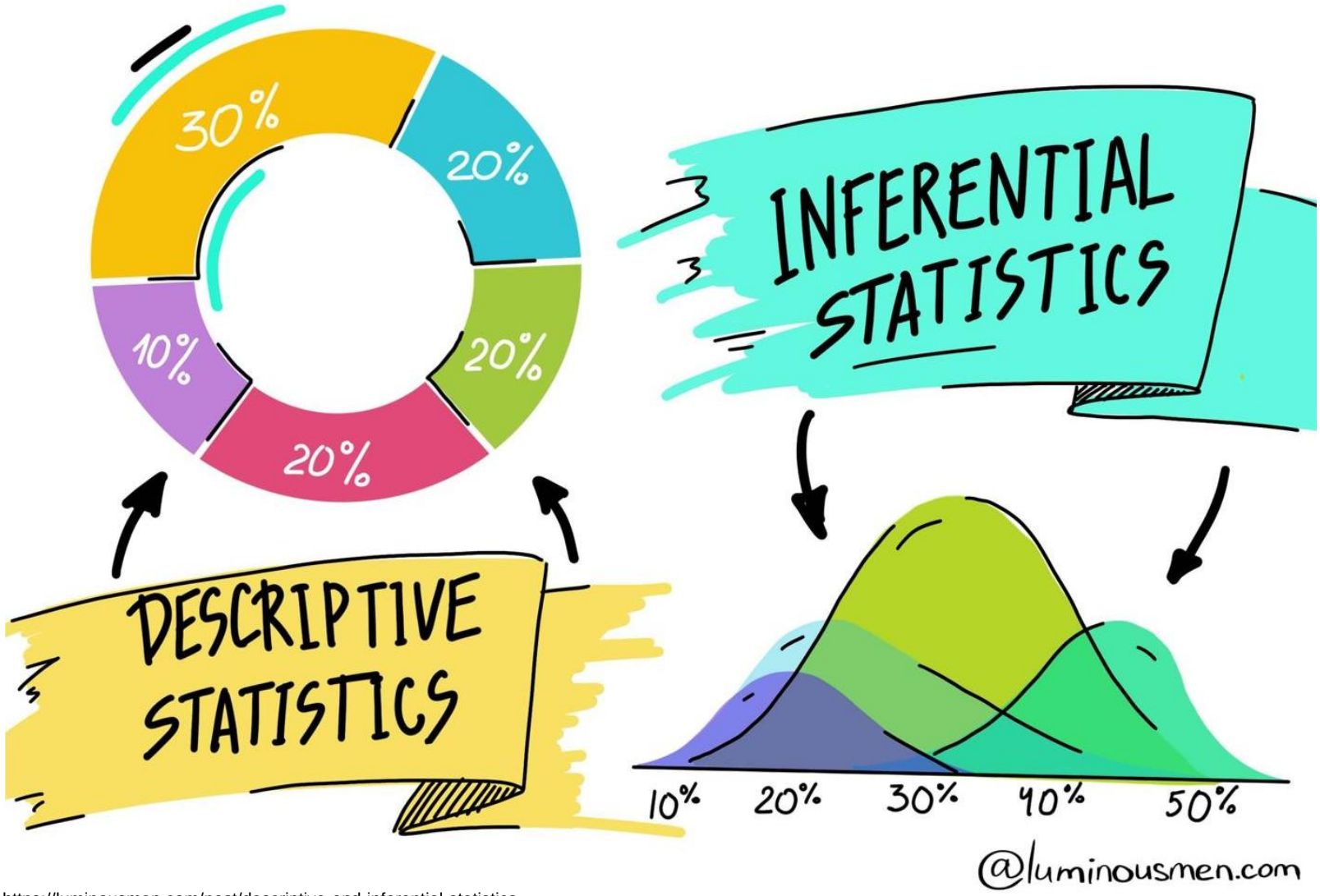
Statistics

Descriptive statistics & Inferential statistics



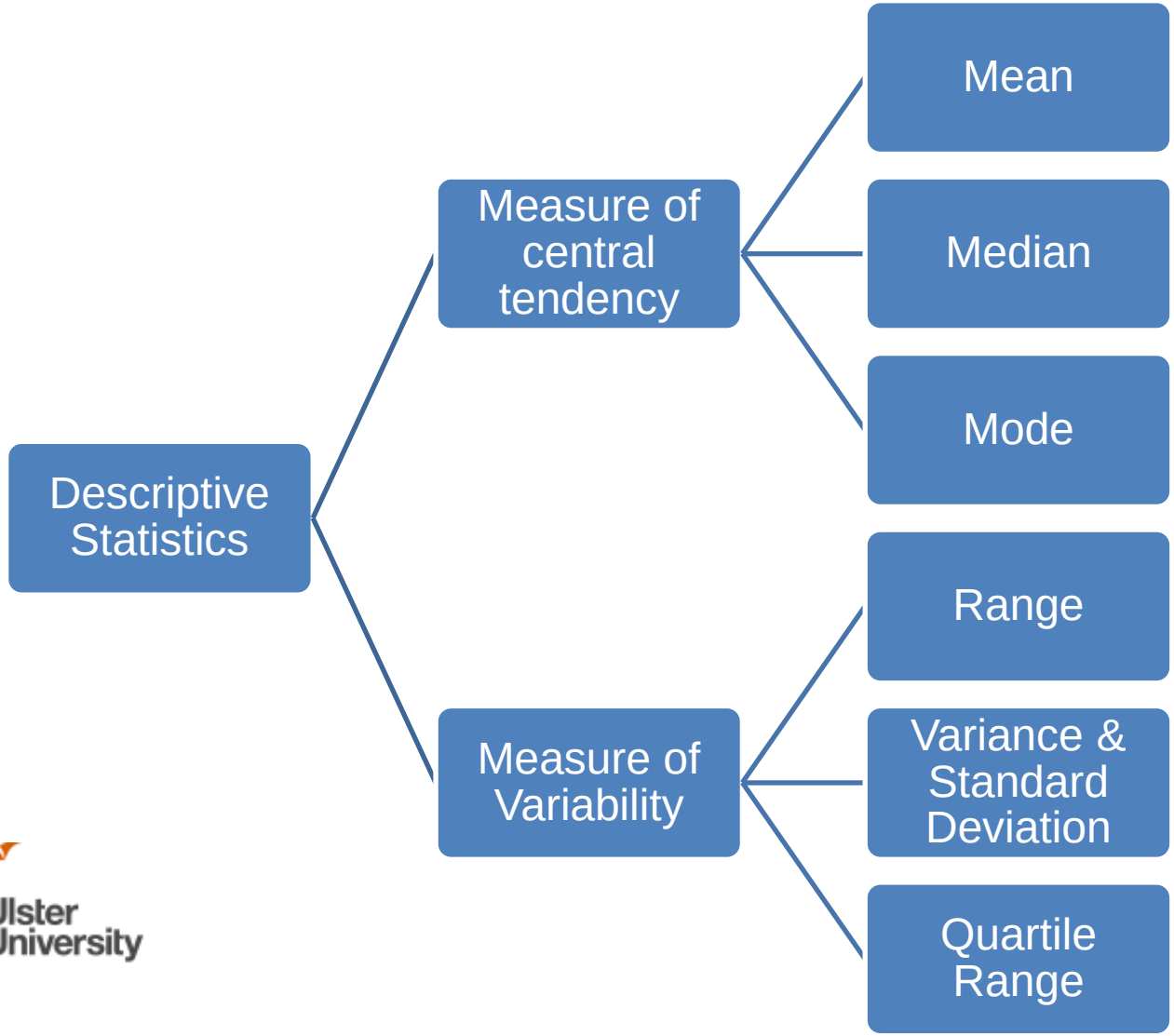
Ulster University

Descriptive vs Inferential statistics



https://luminousmen.com/post/descriptive-and-inferential-statistics

Descriptive statistics



Ulster University

Descriptive statistics

R functions for computing descriptive statistics

- Some R functions for computing descriptive statistics

Description	R function
Mean	mean()
Standard deviation	sd()
Variance	var()
Minimum	min()
Maximum	maximum()
Median	median()
Range of values (minimum and maximum)	range()
Sample quantiles	quantile()
Generic function	summary()
Interquartile range	IQR()

http://www.sthda.com/english/wiki/descriptive-statistics-and-graphics

Descriptive statistics

Measure of central tendency

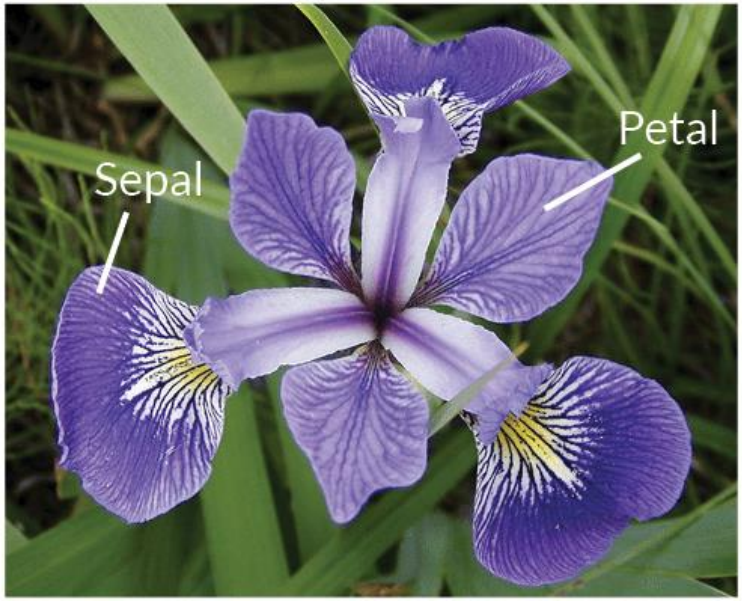
- Mean: the average value
  - R function: mean()
- Median: the middle value
  - R function: median()
- Mode: the most frequent value
  - R function: mfv() [in the modeest R package]

Ulster University

Descriptive statistics

Iris dataset

- Iris dataset contains four features (length and width of sepals and petals) of 50 samples of three species of Iris (Iris setosa, Iris virginica and Iris versicolor)
- http://archive.ics.uci.edu/ml/datasets/Iris



Iris Versicolor

Iris Setosa

Iris Virginica

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

Descriptive statistics
Iris dataset

- The Iris flower data set or **Fisher's Iris data** set is a multivariate data set introduced by the British statistician and biologist **Ronald Fisher** in his 1936 paper.
- The data set consists of **50 samples from each of three species** of Iris (Iris setosa, Iris virginica and Iris versicolor).
- **Four features** were measured **from each sample**: the length and the width of the sepals and petals, in centimeters.
- Based on the combination of these four features, Fisher developed a **linear discriminant model** to distinguish the species from each other.

Ulster University

https://en.wikipedia.org/wiki/Iris_flower_data_set

Descriptive statistics
Iris dataset

- This is one of the best-known database to be found in the pattern recognition literature.
- Fisher's paper is a classic in the field and is referenced frequently to this day.
  - https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1469-1809.1936.tb02137.x
- The data set contains 3 classes of 50 instances each, where each class refers to a type of iris plant.
- Predicted attribute is the class of iris plant. One class is linearly separable from the other 2; the latter are NOT linearly separable from each other.

Ulster University

Descriptive statistics

Measure of central tendency

```
> df <- iris
> head(df)
  Sepal.Length Sepal.Width Petal.Length Petal.Width Species
1          5.1         3.5          1.4         0.2  setosa
2          4.9         3.0          1.4         0.2  setosa
3          4.7         3.2          1.3         0.2  setosa
4          4.6         3.1          1.5         0.2  setosa
5          5.0         3.6          1.4         0.2  setosa
6          5.4         3.9          1.7         0.4  setosa
> mean(df$Sepal.Length) # compute the mean value of Sepal Length
[1] 5.843333
> median(df$Sepal.Width) # compute the median value of Sepal Width
[1] 3
> library('modeest')
> mfv(df$Petal.Length) # compute the mode value of Petal Length
[1] 1.4 1.5
```

Ulster University

Descriptive statistics
Measure of variability

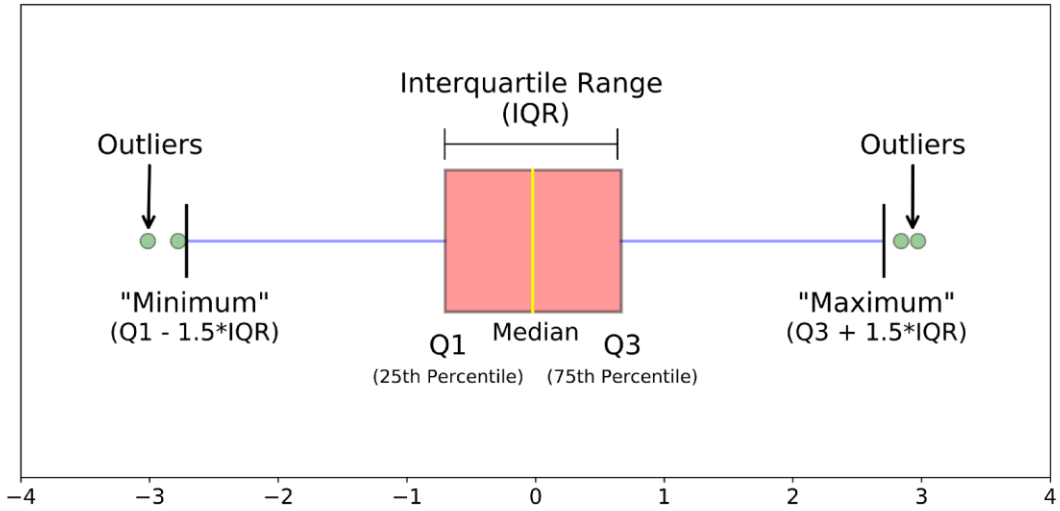
- Range: minimum & maximum: corresponds to biggest value minus the smallest value and gives the full spread of data
  - R function: range() (min(), max())
- Variance: average squared deviation from the mean
- Standard deviation: square root of the variance
  - R function: var() sd()

Ulster University

Descriptive statistics

Measure of variability

- Quartiles: minimum, maximum, and three quartiles (the 0.25, 0.50 and 0.75 quartiles).
  - R function: quantile(),
- Interquartile range (IQR): the difference between the first and third quartiles
  - R function: IQR()



Source: Towards Data Science, on Medium. https://towardsdatascience.com/understanding-boxplots-5e2df7bcbd51

Descriptive statistics

Measure of variability

```
> min(df$Sepal.Length) # Compute the minimum value
[1] 4.3
> max(df$Sepal.Length) # Compute the maximum value
[1] 7.9
> range(df$Sepal.Length) # Compute the range value
[1] 4.3 7.9
> quantile(df$Sepal.Length) # Compute 4 quartiles
  0%  25%  50%  75% 100%
 4.3  5.1  5.8  6.4  7.9
> IQR(df$Sepal.Length) # Compute the interqurtile range
[1] 1.3
>
```

Ulster University

Descriptive statistics

Overall summary of a variable and an entire data frame

```
> head(df) # first few lines
  Sepal.Length Sepal.Width Petal.Length Petal.Width Species
1          5.1         3.5          1.4         0.2  setosa
2          4.9         3.0          1.4         0.2  setosa
3          4.7         3.2          1.3         0.2  setosa
4          4.6         3.1          1.5         0.2  setosa
5          5.0         3.6          1.4         0.2  setosa
6          5.4         3.9          1.7         0.4  setosa
> summary(df) # summary statistics
  Sepal.Length    Sepal.Width     Petal.Length    Petal.Width          Species
 Min.   :4.300   Min.   :2.000   Min.   :1.000   Min.   :0.100   setosa    :50
 1st Qu.:5.100   1st Qu.:2.800   1st Qu.:1.600   1st Qu.:0.300   versicolor:50
 Median :5.800   Median :3.000   Median :4.350   Median :1.300   virginica :50
 Mean   :5.843   Mean   :3.057   Mean   :3.758   Mean   :1.199
 3rd Qu.:6.400   3rd Qu.:3.300   3rd Qu.:5.100   3rd Qu.:1.800
 Max.   :7.900   Max.   :4.400   Max.   :6.900   Max.   :2.500
> 
```

Ulster University

Descriptive statistics

Overall summary of a variable and an entire data frame

```
> str(df) #show the structure
'data.frame':   150 obs. of  5 variables:
 $ Sepal.Length: num  5.1 4.9 4.7 4.6 5 5.4 4.6 5 4.4 4.9 ...
 $ Sepal.Width : num  3.5 3 3.2 3.1 3.6 3.9 3.4 3.4 2.9 3.1 ...
 $ Petal.Length: num  1.4 1.4 1.3 1.5 1.4 1.7 1.4 1.5 1.4 1.5 ...
 $ Petal.Width : num  0.2 0.2 0.2 0.2 0.2 0.4 0.3 0.2 0.2 0.1 ...
 $ Species     : Factor w/ 3 levels "setosa","versicolor",..: 1 1 1 1 1 1 1 1 1 1 ...
> 
```

Ulster University

Descriptive statistics

Overall summary of a variable and an entire data frame

- sapply() function
  - to apply a particular function over a list or vector.

```
> sapply(df[,-5],median) # Compute the median for each column
Sepal.Length  Sepal.Width Petal.Length  Petal.Width
        5.80         3.00         4.35         1.30
> sapply(df[,-5],quantile) # Compute the quantile for each column
     Sepal.Length Sepal.Width Petal.Length Petal.Width
0%            4.3         2.0         1.00         0.1
25%           5.1         2.8         1.60         0.3
50%           5.8         3.0         4.35         1.3
75%           6.4         3.3         5.10         1.8
100%          7.9         4.4         6.90         2.5
>
```

Ulster University

Data Visualisation

Basic plots

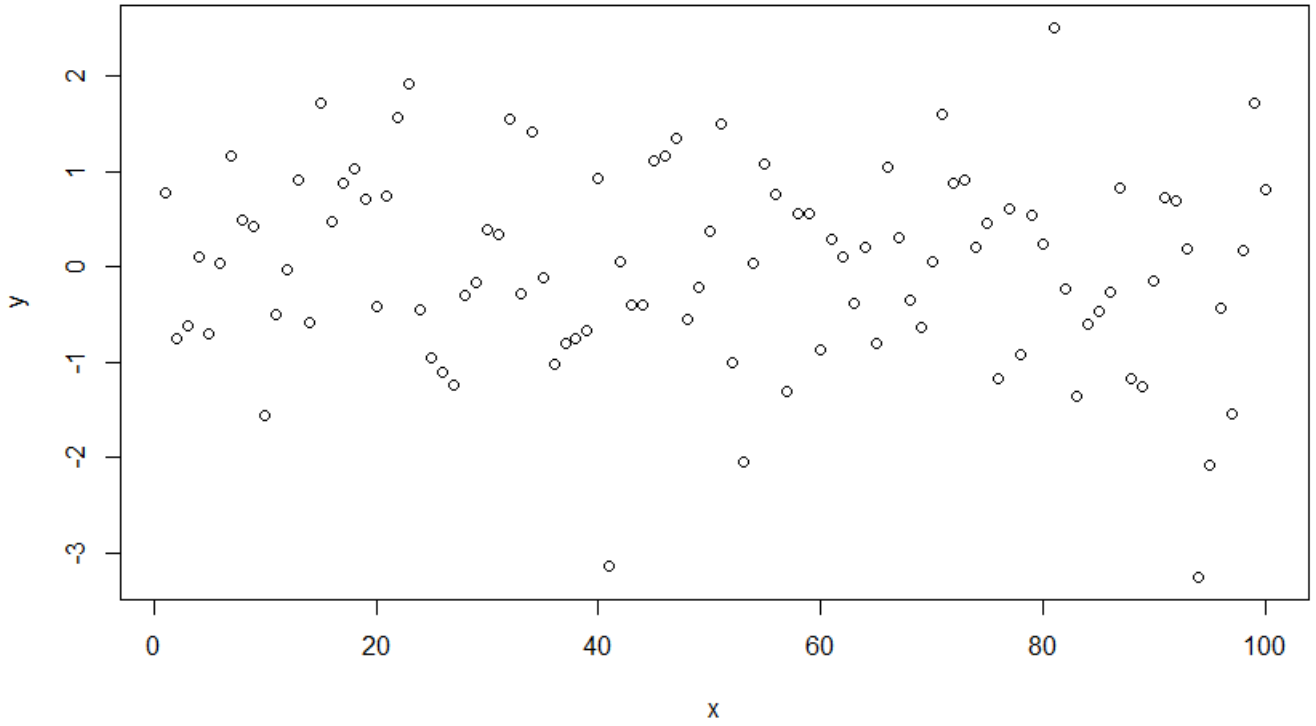
- Scatter plot
- Histogram
- Bar plot
- Line plot
- Box and Whisker plot

Ulster University

Data Visualisation
Basic plots

- Scatter plot: to display values for typically two variables for a set of data.

```
> x <- 1:100
> y <- rnorm(100)
> plot(x,y)
>
```



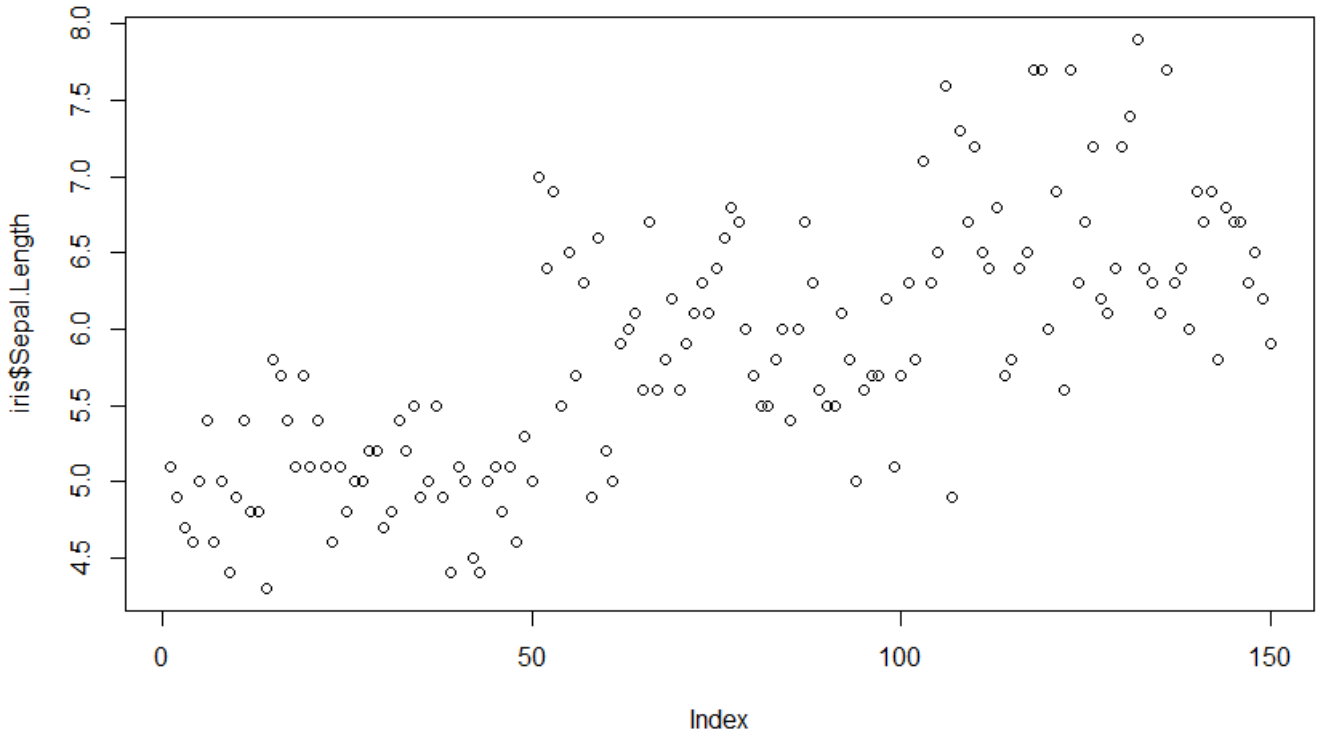
Ulster University

Data Visualisation

Scatter plot

- Simple scatter plot

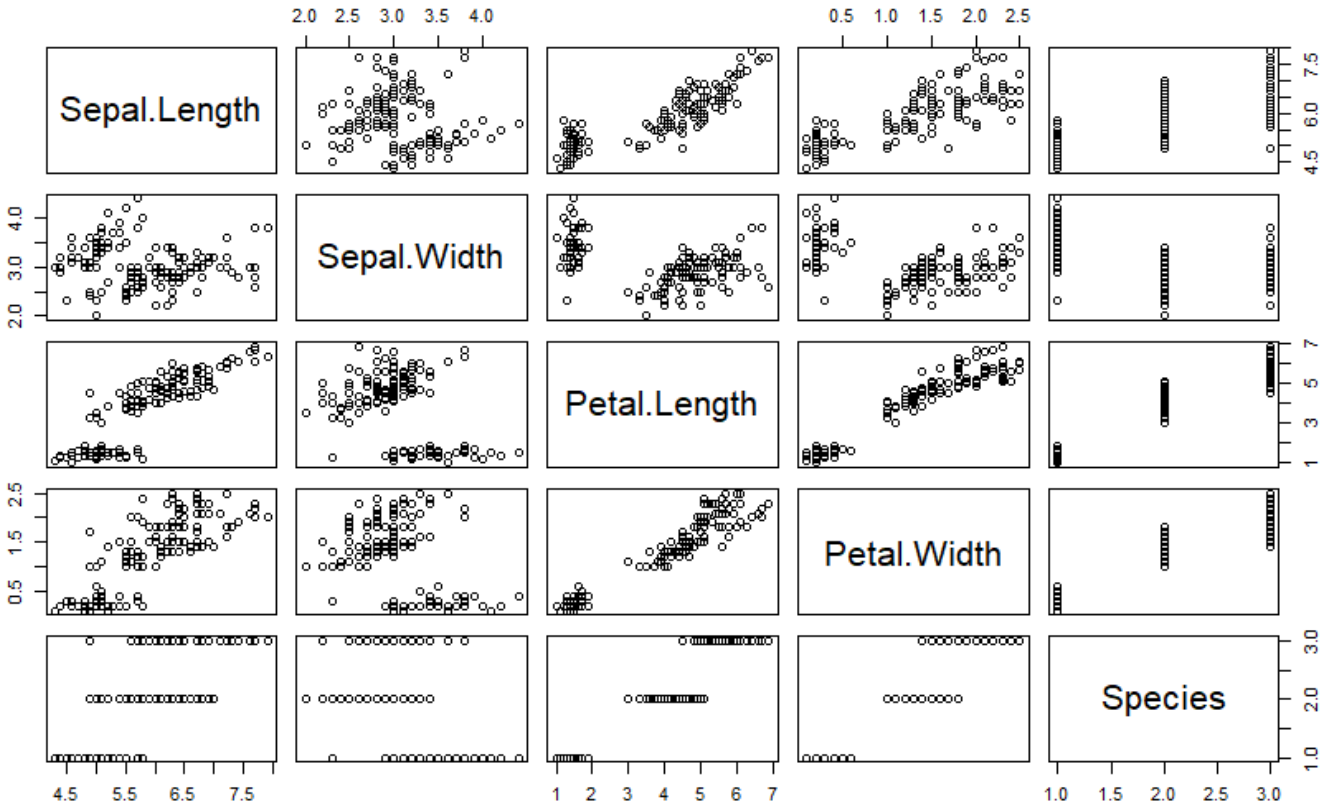
```
> plot(iris$Sepal.Length) # Simple Scatter Plot
>
```



Data Visualisation

Scatter plot

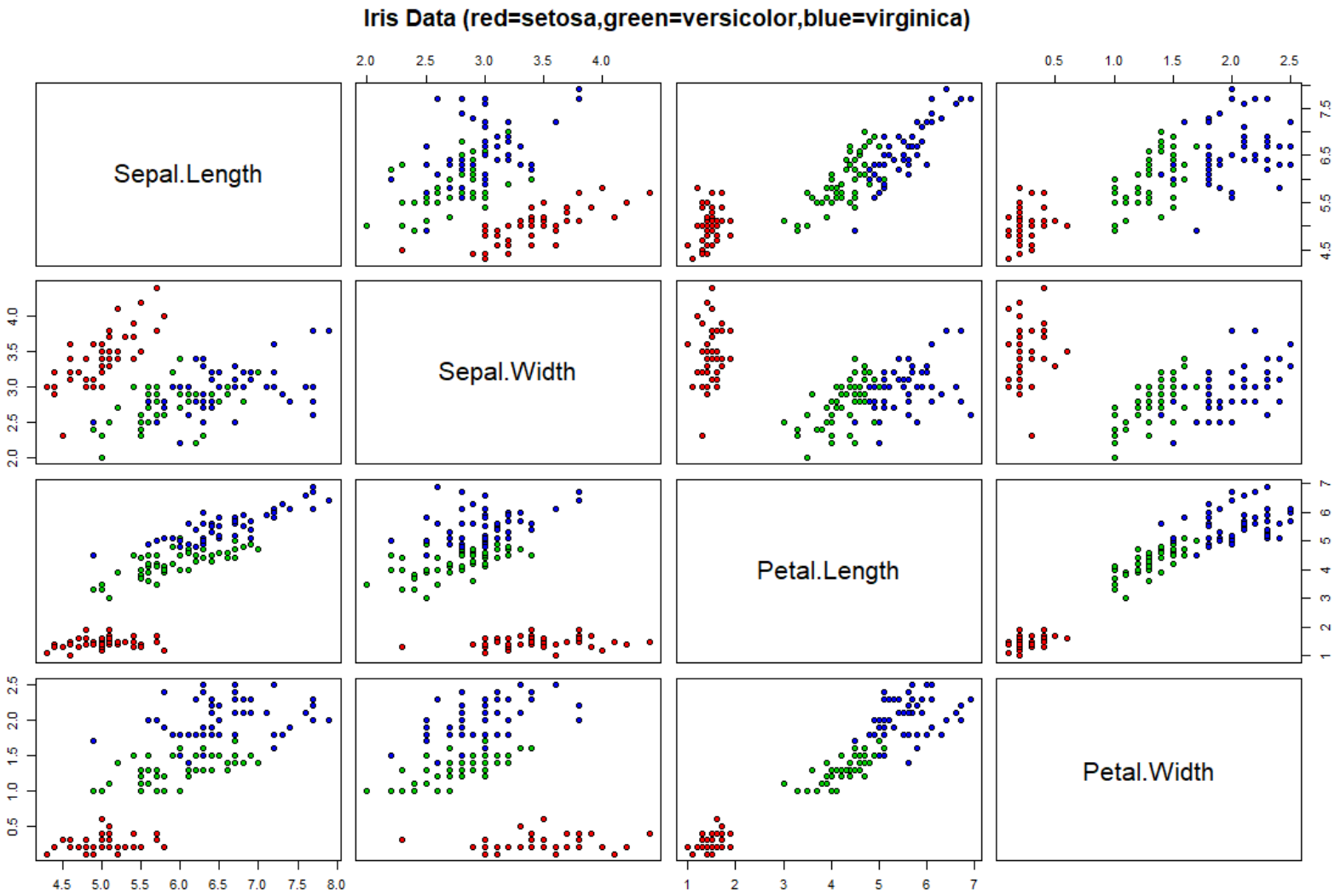
- Scatter plot matrix – pairs(iris)
  - The pairs R function returns a plot matrix, consisting of scatterplots for each variable-combination of a data frame.



Ulster University

Data Visualisation

```
> # Create scatterplots of all pairwise combination of the 4 variables in the dataset
> pairs(iris[1:4], main="Iris Data (red=setosa,green=versicolor,blue=virginica)",
+       pch=21, bg=c("red","green3","blue")[unclass(iris$Species)])
```

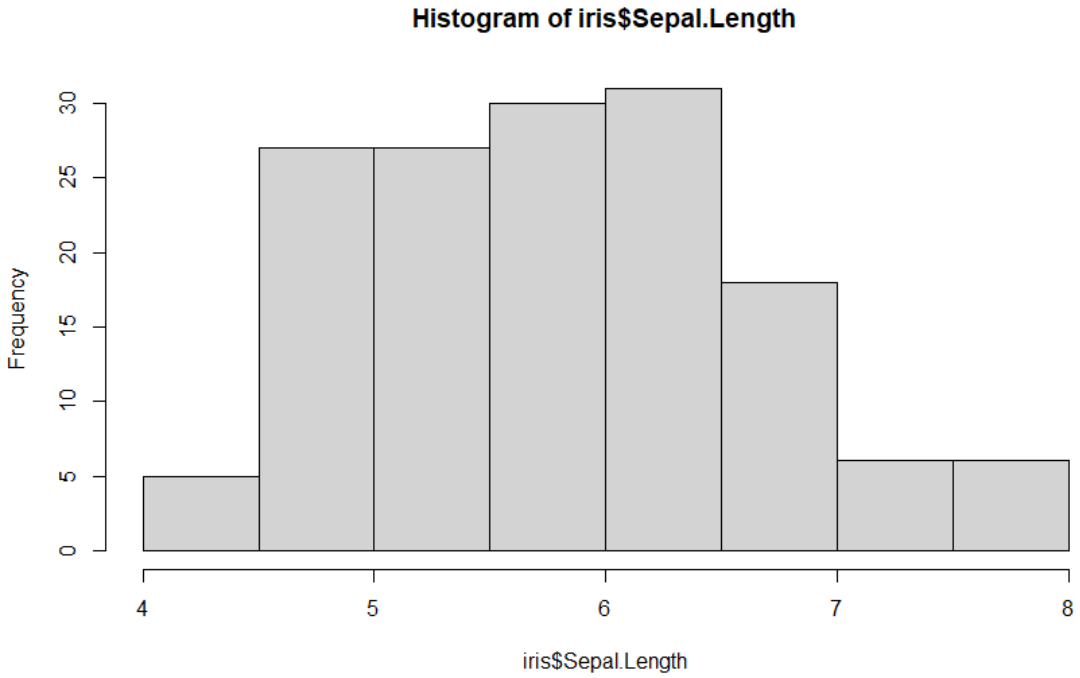


Data Visualisation

Histogram (Frequency distribution plot)

- An approximate representation of the distribution of numerical data.
  - R function: hist()

```
> hist(iris$Sepal.Length) # Base R histogram with default bins
> |
```

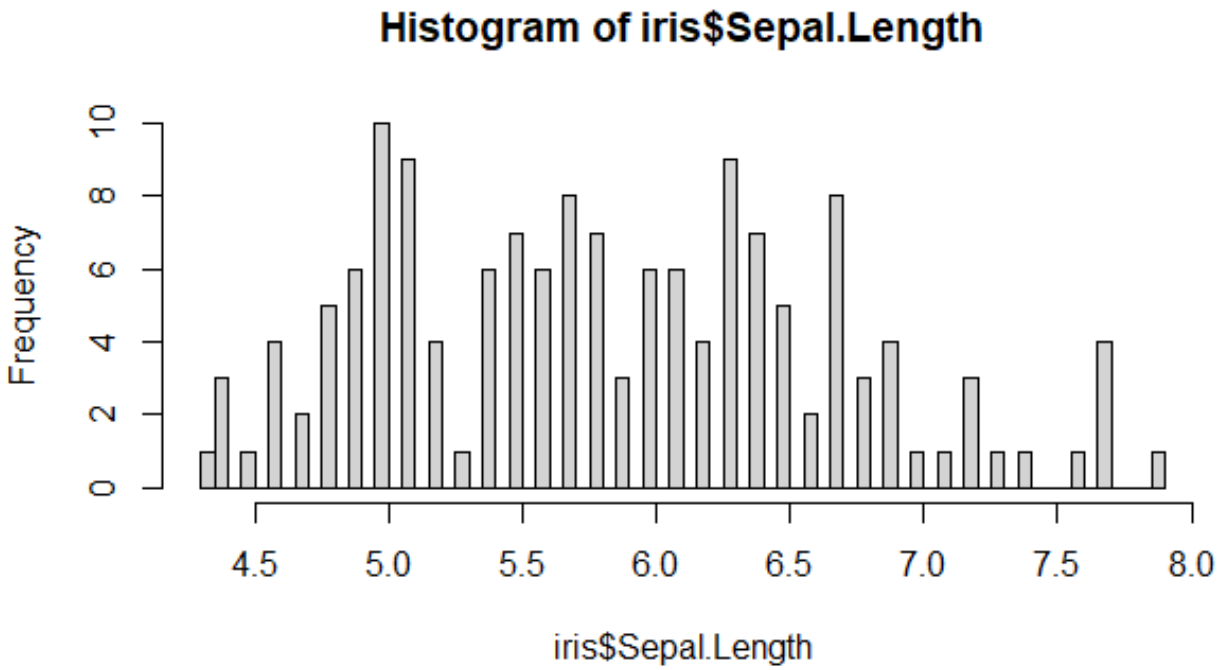


Ulster University

Data Visualisation
Histogram

- # breaks argument of the hist function is to increase or decrease the width of the bars.

```
> hist(iris$Sepal.Length, breaks = 100) # Base R histogram with manual bins
>
```

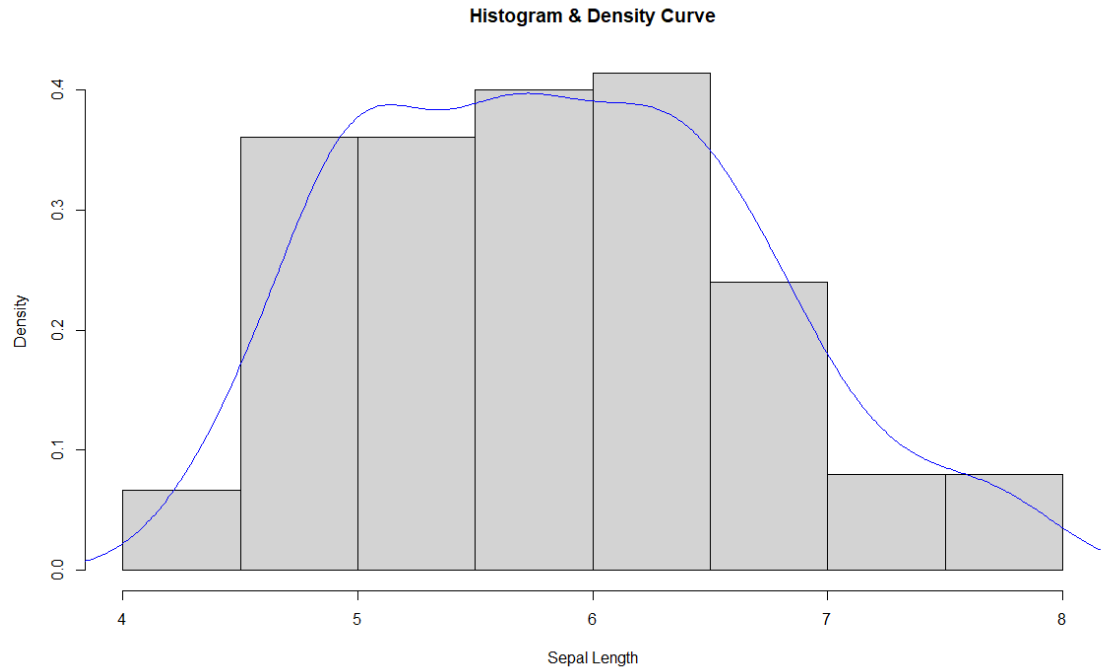


Ulster University

Data Visualisation
Histogram

- Histogram & Density Curve

```
> d <- density(iris$Sepal.Length)
> hist(iris$Sepal.Length, prob=TRUE, xlab='Sepal Length', main='Histogram & Density Curve')
> lines(d,col='blue')
>
```



Ulster University

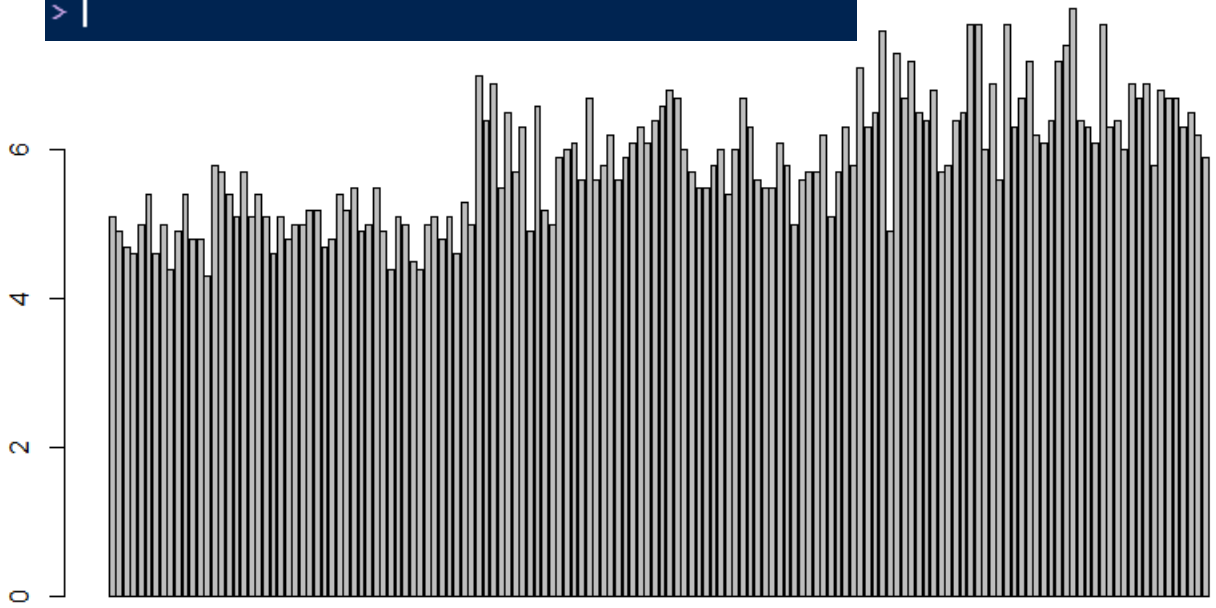
https://www.publichealth.columbia.edu/sites/default/files/media/fdawg_ggplot2.html

Data Visualisation

Barplot

- Data is represented in the form of rectangular bars
  - R function: barplot()

```
> barplot(iris$Sepal.Length,xlab = 'Sepal Length')
>
```

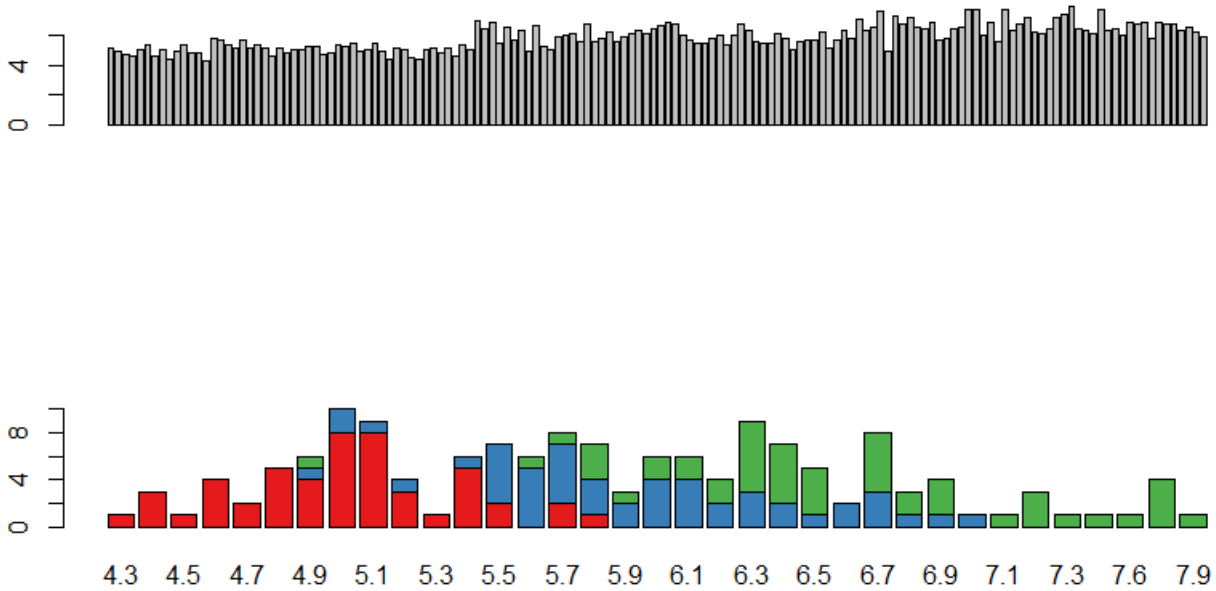


Data Visualisation

Bar plot

- Arrange plots with par(mfrow)
- ? par()

```
> par(mfrow=c(2,1))
> barplot(iris$Sepal.Length)  # creating simple bar graph
> library(RColorBrewer)
> barplot(table(iris$Species,iris$Sepal.Length),col = brewer.pal(3,"Set1")) # stacked plot
>
```

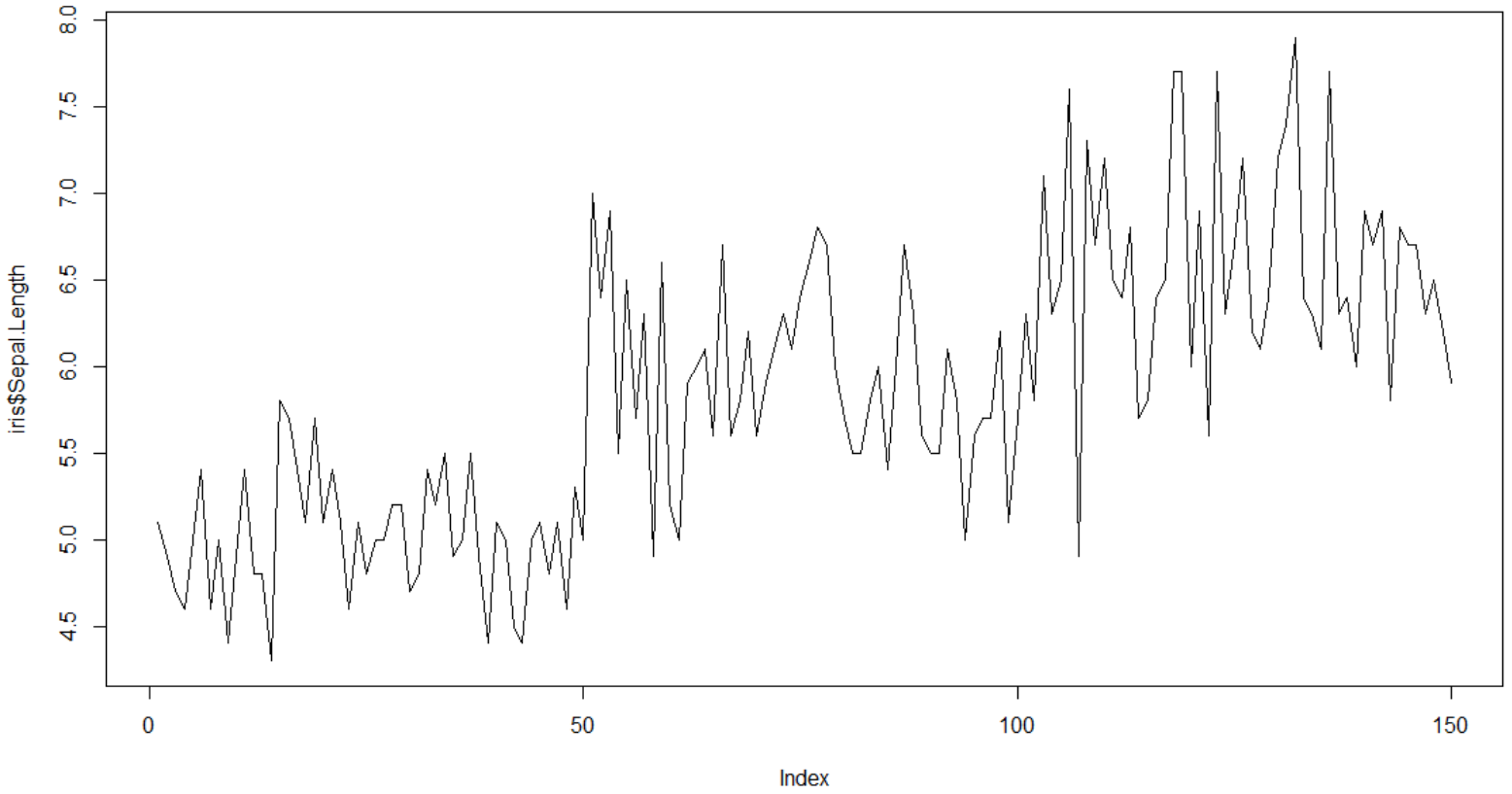


Ulster University

Data Visualisation

Line plot

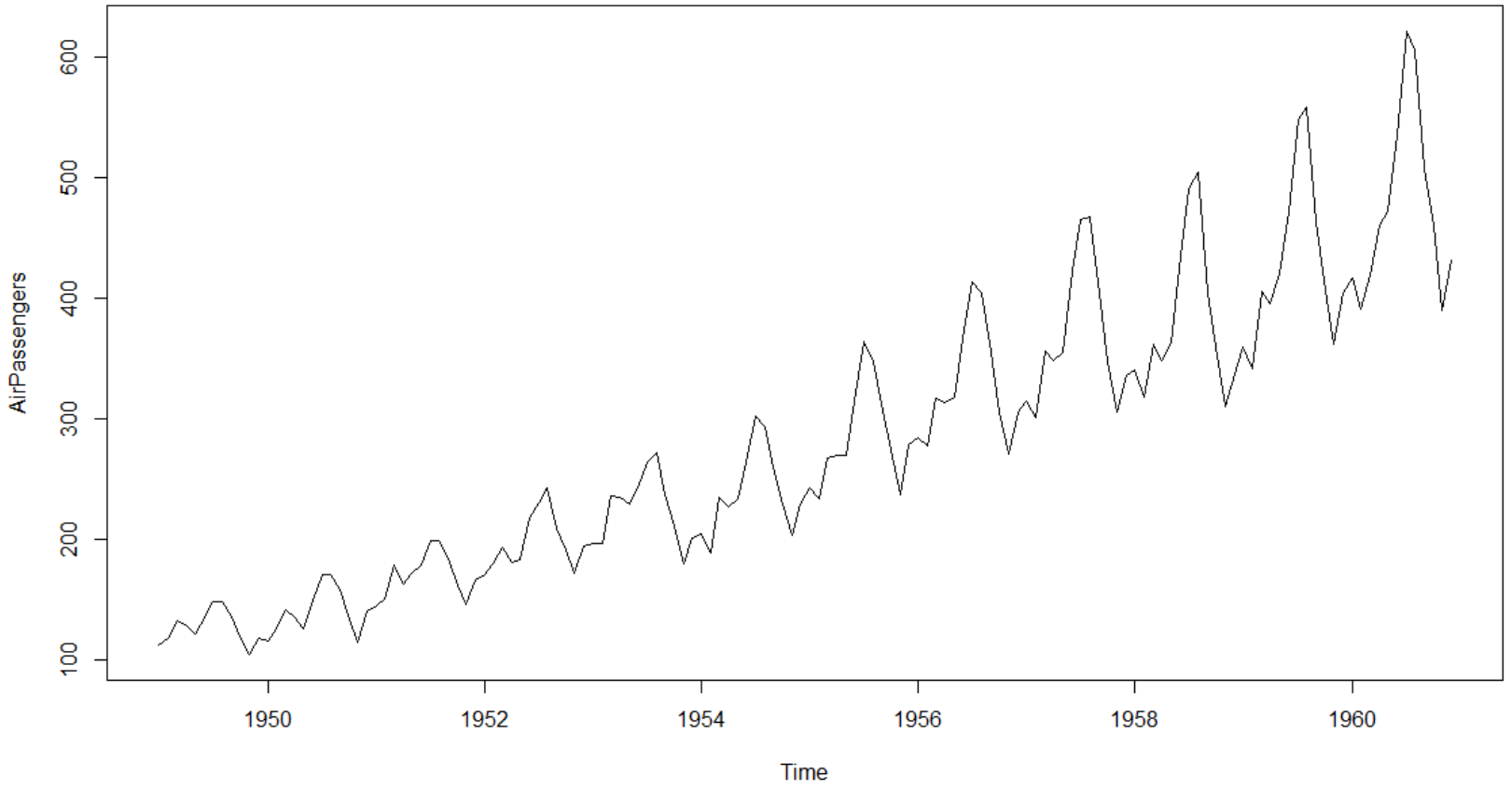
```
> plot(iris$Sepal.Length,type='l')
>
```



Data Visualisation

Line plot

```
> plot(AirPassengers,type='l')
>
```



Data Visualisation

Box and Whisker plot

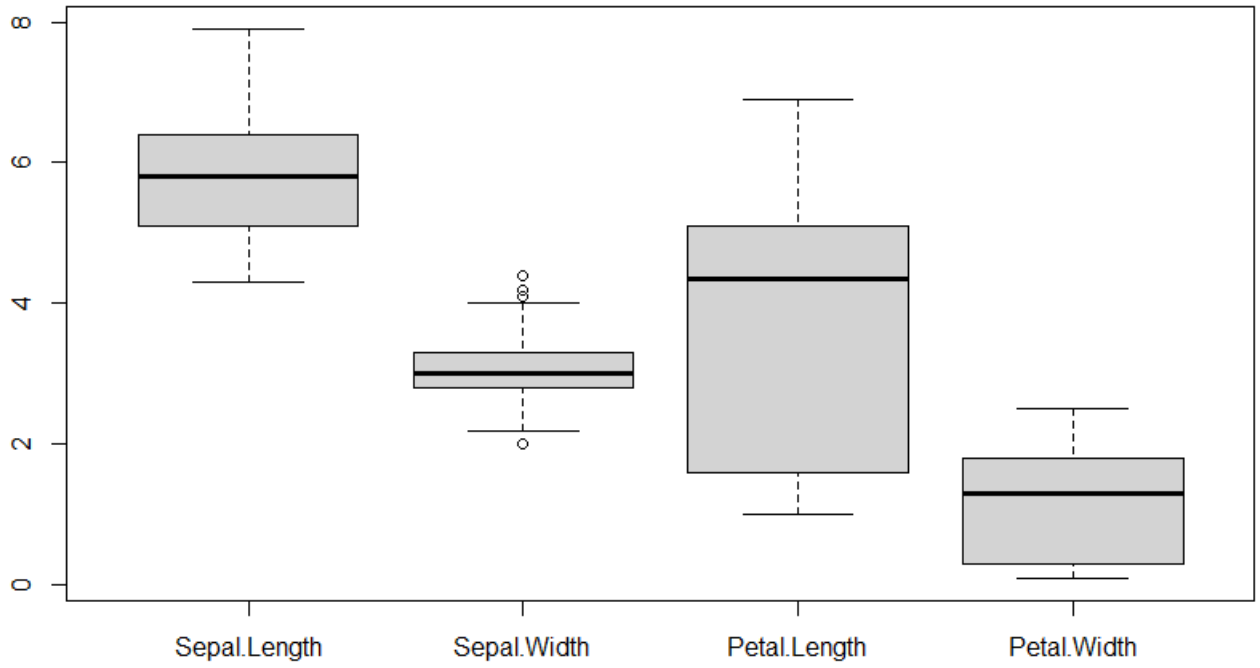
- **Box plot**: graphically depicting groups of numerical data through their quartiles.
- **Box-and-whisker plot**: lines extending from the boxes (whiskers) indicating variability outside the upper and lower quartiles.

Ulster University

Data Visualisation

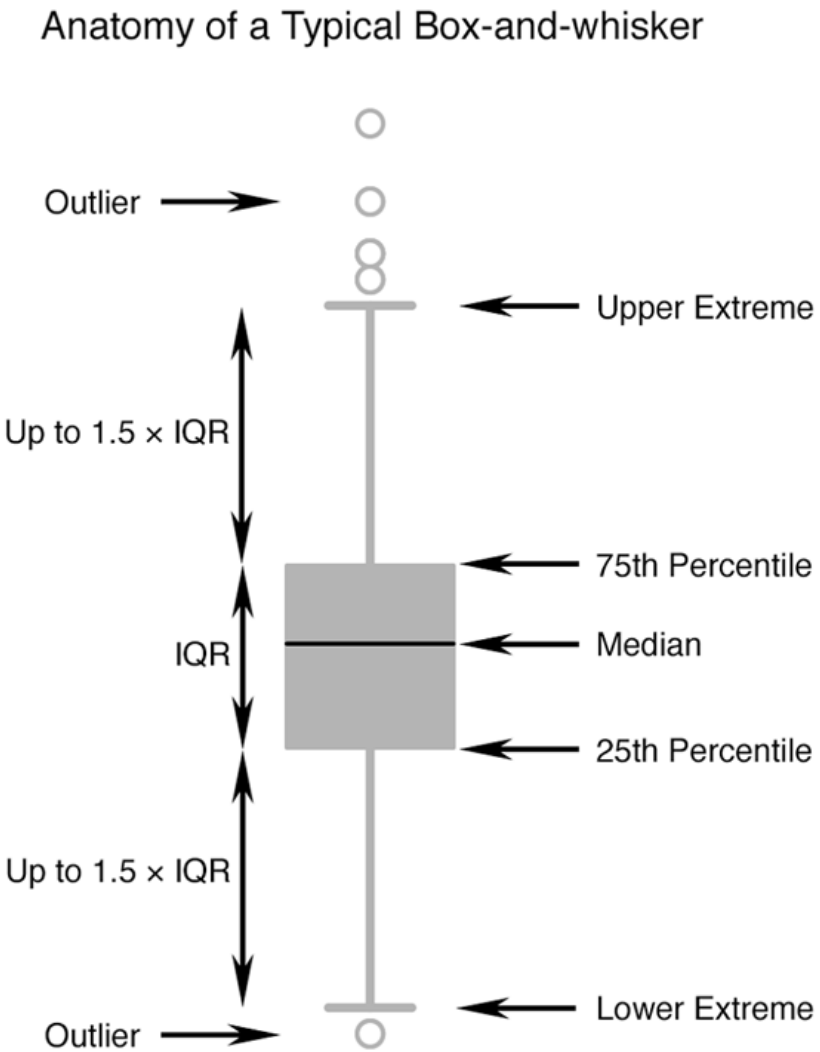
Box and Whisker plot

```
> boxplot(iris[,-5])
>
```



Data Visualisation

Box and Whisker plot



https://www.infragistics.com/community/blogs/b/tim_brock/posts/demystifying-box-and-whisker-plots-part-1

Data Visualisation

Sophisticated Visualisation - ggplot2

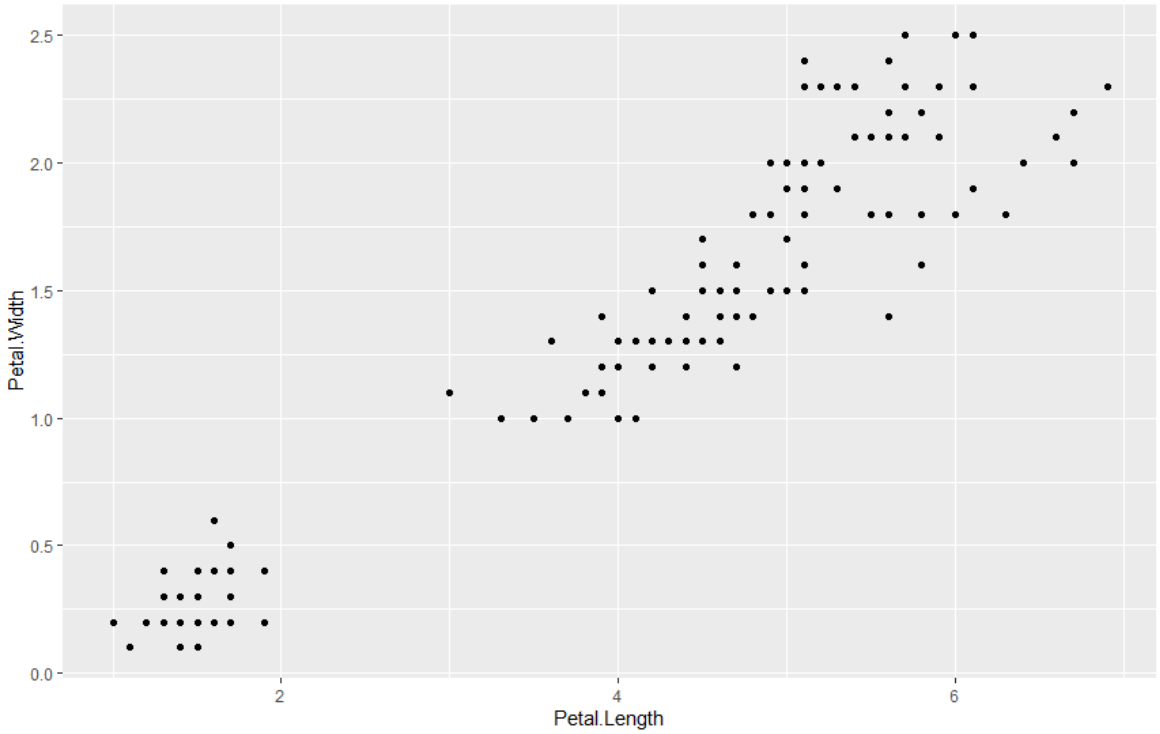
- One of the most widely used visualisation packages in R
- Install ggplot2 package
  - install.packages("ggplot2")
- Include ggplot2 library
  - library(ggplot2)
- http://docs.ggplot2.org/current/

Ulster University

Data Visualisation
ggplot

- Scatter plot

```
> library(ggplot2)
> g <- ggplot(data=iris, aes(x=Petal.Length, y=Petal.Width))
> g + geom_point()
>
```



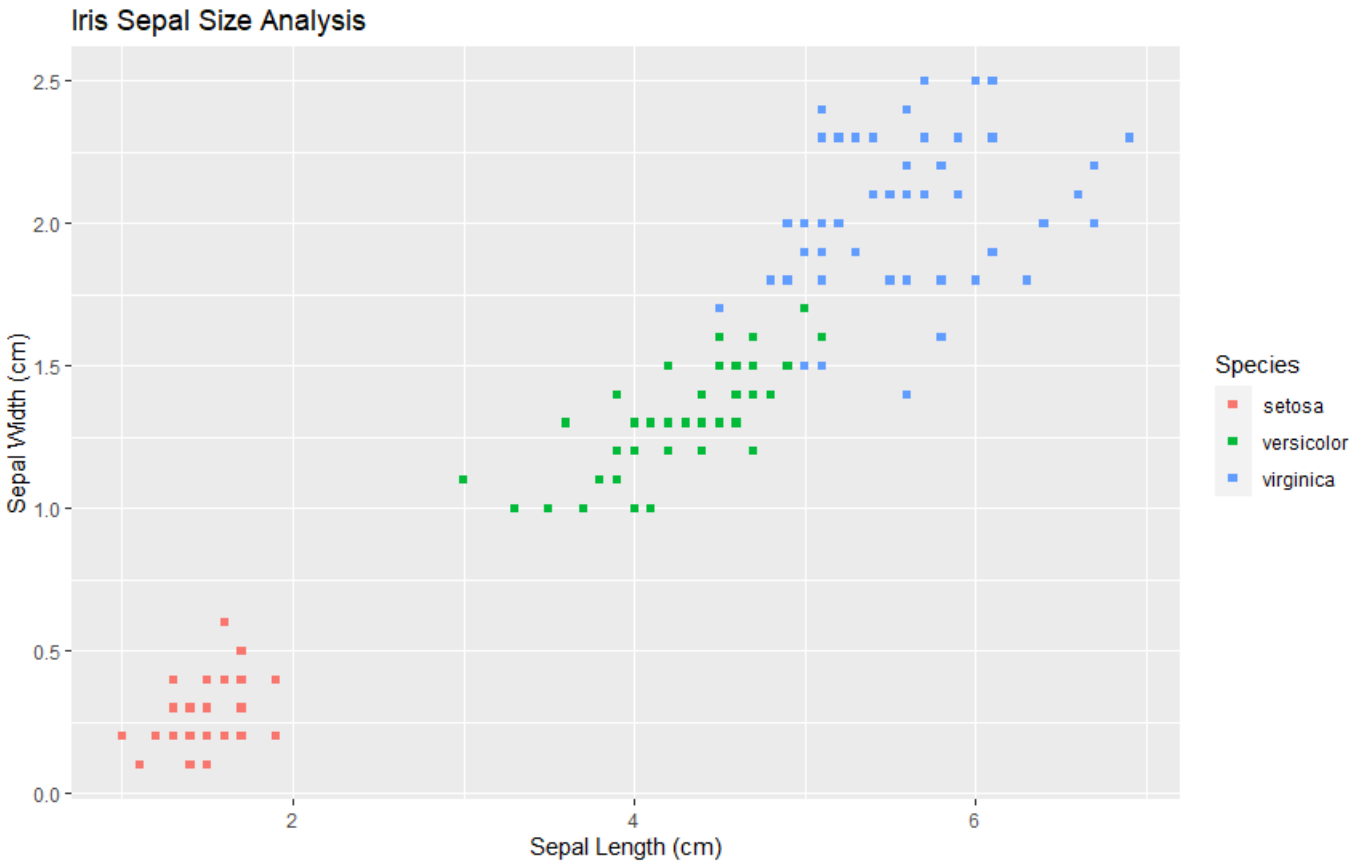
Ulster University

Data Visualisation

ggplot

- Scatter plot

```
> g <- ggplot(data=iris, aes(x=Petal.Length, y=Petal.Width))
> g + geom_point(aes(col = Species), shape=15, size=1.5) +
+     ggtitle('Iris Sepal Size Analysis') +
+     labs(x='Sepal Length (cm)', y='Sepal Width (cm)')
> |
```

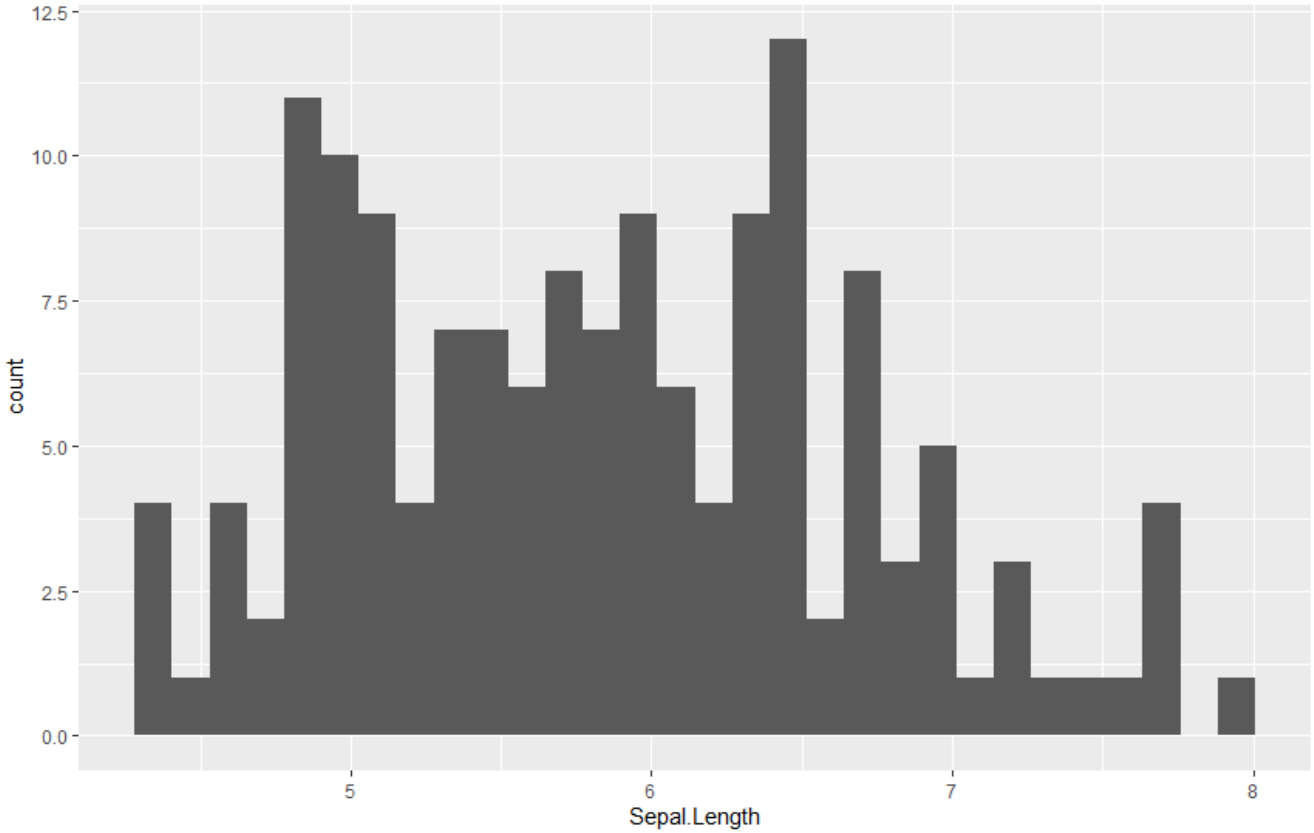


https://rstudio-pubs-static.s3.amazonaws.com/469868_06fe036bb54942d5af4801ad88b49b53.html

Data Visualisation
ggplot

- Histogram

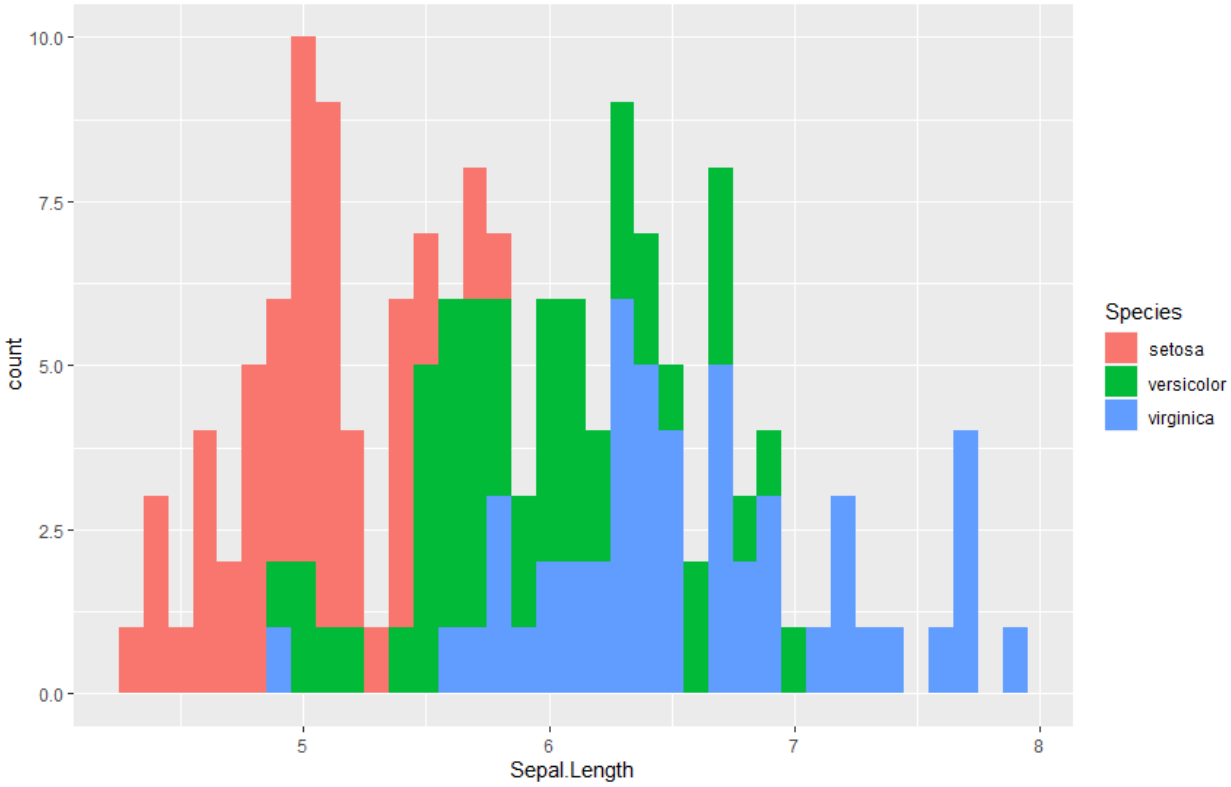
```
> ggplot(iris, aes(x=Sepal.Length)) + geom_histogram()
`stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
>
```



Data Visualisation
ggplot

- Histogram

```
> ggplot(iris, aes(x=Sepal.Length, fill=Species)) +
+     geom_histogram(binwidth = 0.1)
> |
```

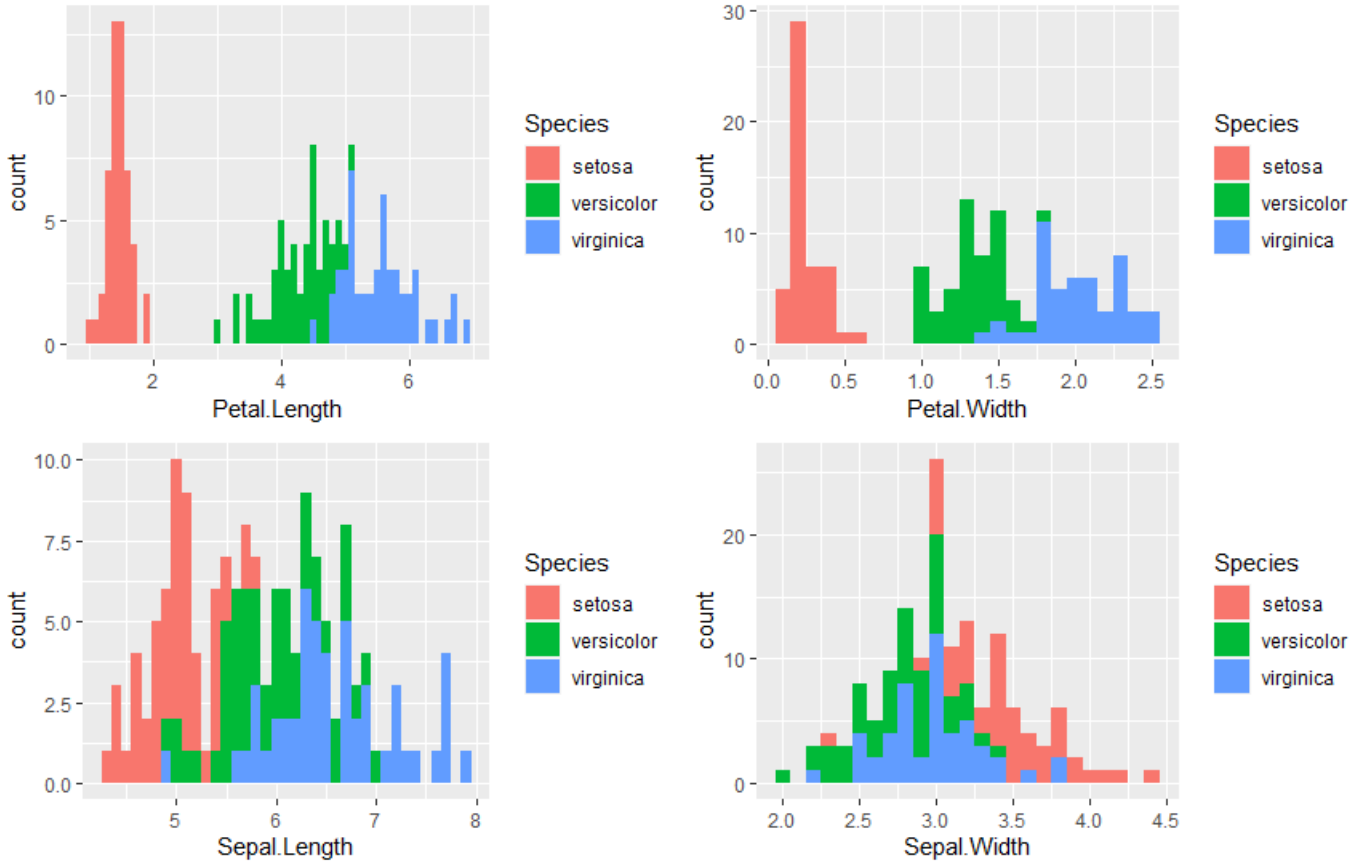


Data Visualisation

ggplot

- Histogram
- (for all four variables)

```
> library(gridExtra)
> plen <- ggplot(iris, aes(x=Petal.Length, fill=Species)) +
+     geom_histogram(binwidth = 0.1)
> pwth <- ggplot(iris, aes(x=Petal.Width, fill=Species)) +
+     geom_histogram(binwidth = 0.1)
> slen <- ggplot(iris, aes(x=Sepal.Length, fill=Species)) +
+     geom_histogram(binwidth = 0.1)
> swth <- ggplot(iris, aes(x=Sepal.Width, fill=Species)) +
+     geom_histogram(binwidth = 0.1)
>
> grid.arrange(plen, pwth, slen, swth, nrow = 2)
> |
```

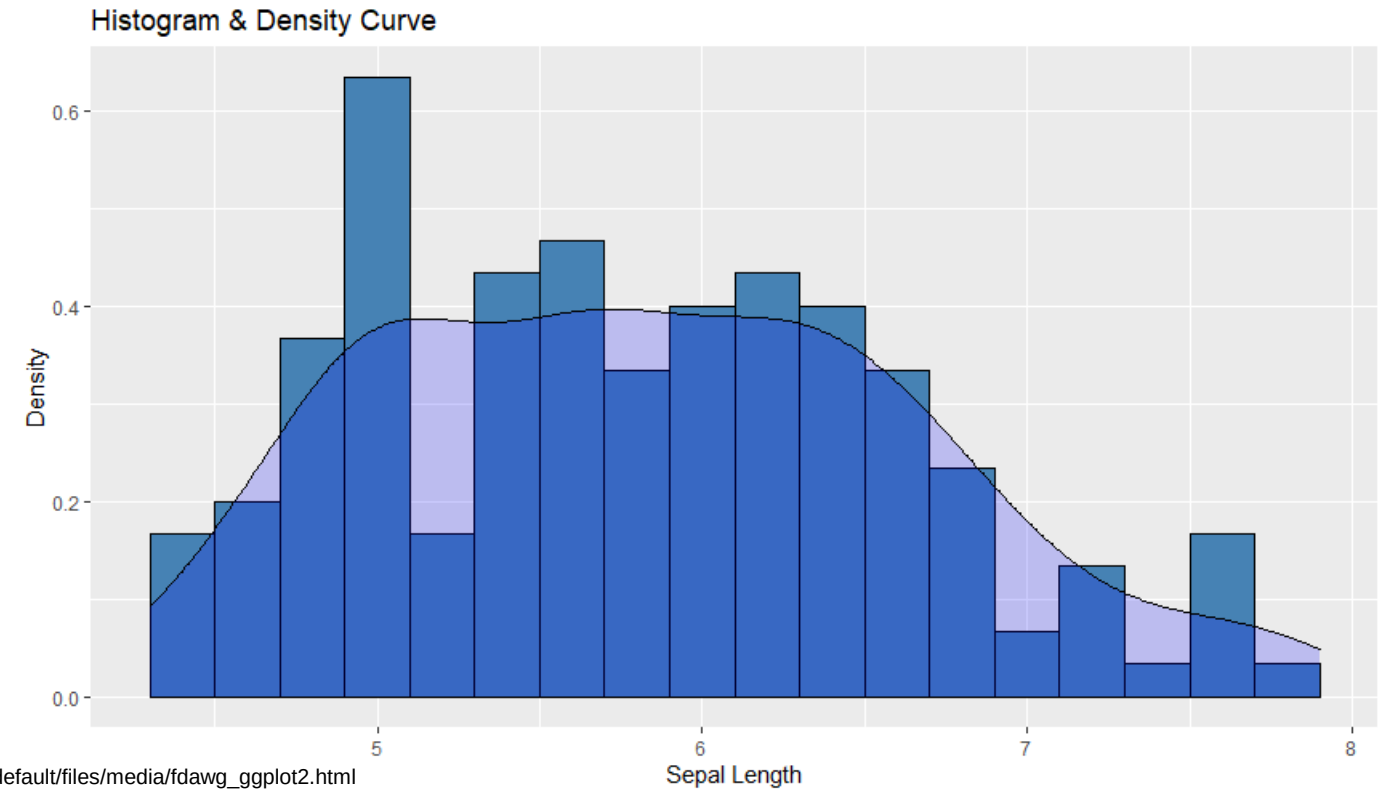


Ulster University

Data Visualisation
ggplot

- Density Curve

```
> density <- ggplot(data=iris, aes(x=Sepal.Length))
> density + geom_histogram(binwidth=0.2, color="black", fill="steelblue", aes(y=..density..)) +
+     geom_density(stat="density", alpha=I(0.2), fill="blue") +
+     xlab("Sepal Length") +  ylab("Density") + ggtitle("Histogram & Density Curve")
> |
```



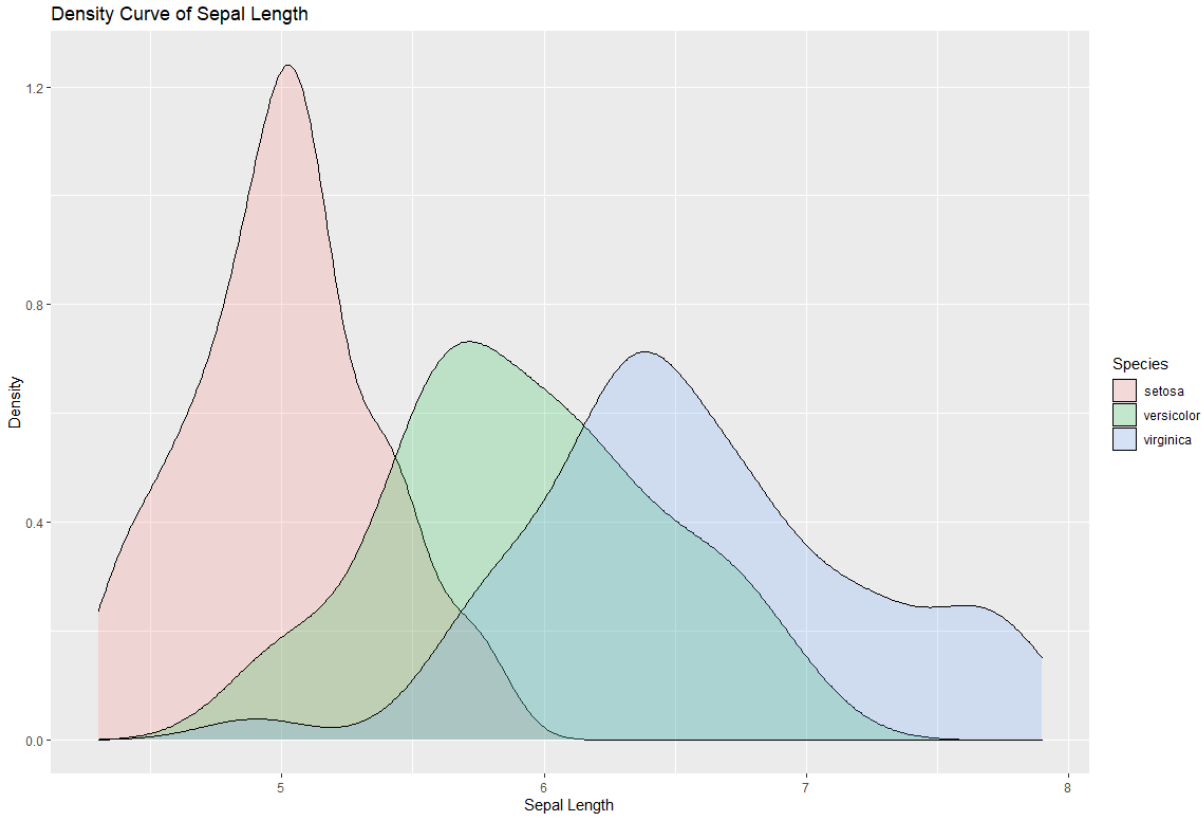
Ulster University

https://www.publichealth.columbia.edu/sites/default/files/media/fdawg_ggplot2.html

Data Visualisation

ggplot

```
> density2 <- ggplot(data=iris, aes(x=Sepal.Length, fill=Species))
> density2 + geom_density(stat='density', alpha=(0.2)) +
+     xlab('Sepal Length') + ylab('Density') + ggtitle('Density Curve of Sepal Length')
>
```



Ulster University

https://www.publichealth.columbia.edu/sites/default/files/media/fdawg_ggplot2.html

Data Visualisation
ggplot

- Bar plot

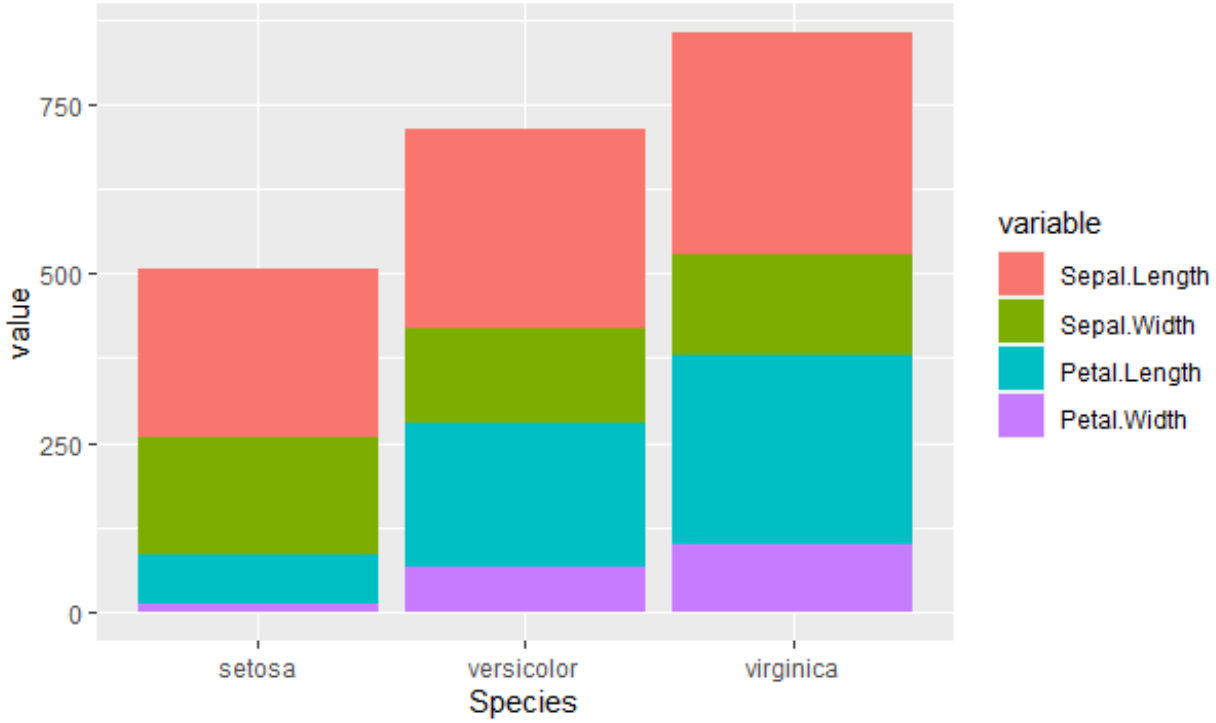
```
> library(reshape2)
> df2 <- melt(iris, id.vars="Species") # use melt() function to restructure dataset
> head(df2)
  Species     variable value
1  setosa Sepal.Length   5.1
2  setosa Sepal.Length   4.9
3  setosa Sepal.Length   4.7
4  setosa Sepal.Length   4.6
5  setosa Sepal.Length   5.0
6  setosa Sepal.Length   5.4
>
```

Ulster University

Data Visualisation

ggplot

```
> library(ggplot2)
> bar1 <- ggplot(data=df2, aes(x=Species, y=value, fill=variable))
> bar1 + geom_bar(stat="identity")
>
```

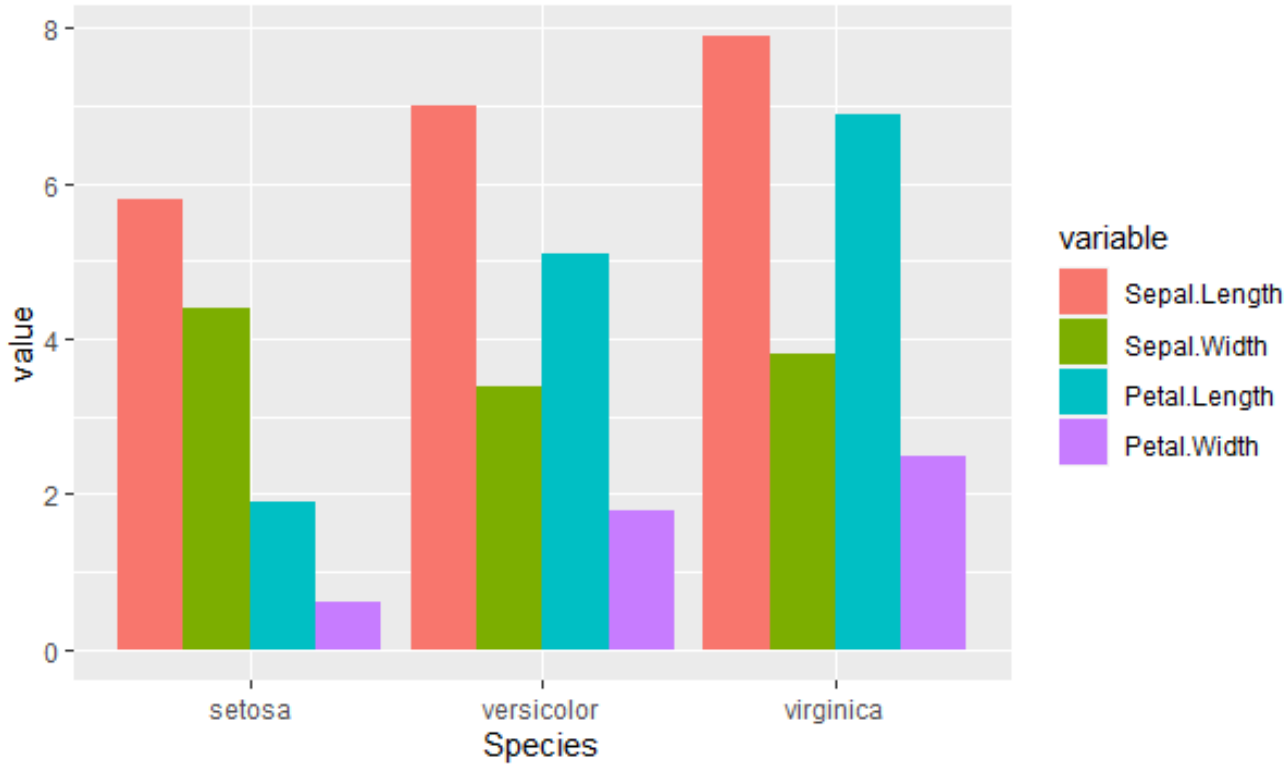


Ulster University

Data Visualisation

ggplot

```
> # Use position=position_dodge()
> library(ggplot2)
> bar1 <- ggplot(data=df2, aes(x=Species, y=value, fill=variable))
> bar1 + geom_bar(stat="identity", position = position_dodge())
>
```

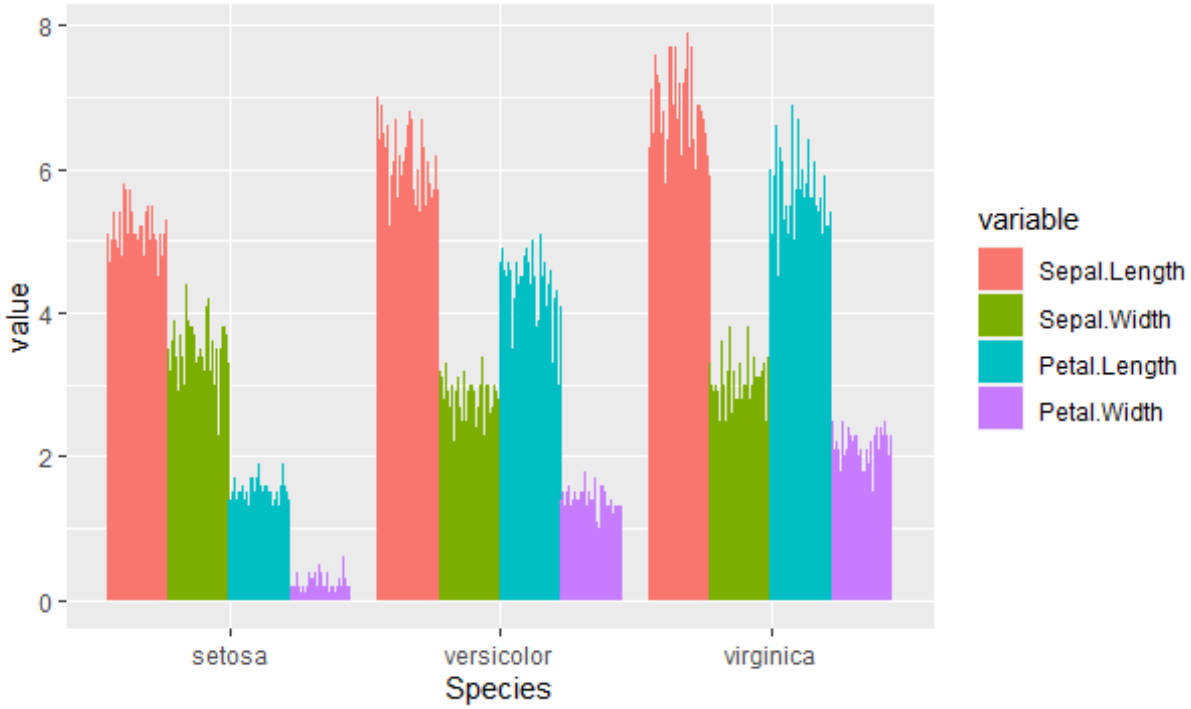


Ulster University

Data Visualisation

ggplot

```
> # Use position=position_dodge2()
> library(ggplot2)
> bar1 <- ggplot(data=df2, aes(x=Species, y=value, fill=variable))
> bar1 + geom_bar(stat="identity", position = position_dodge2())
>
```

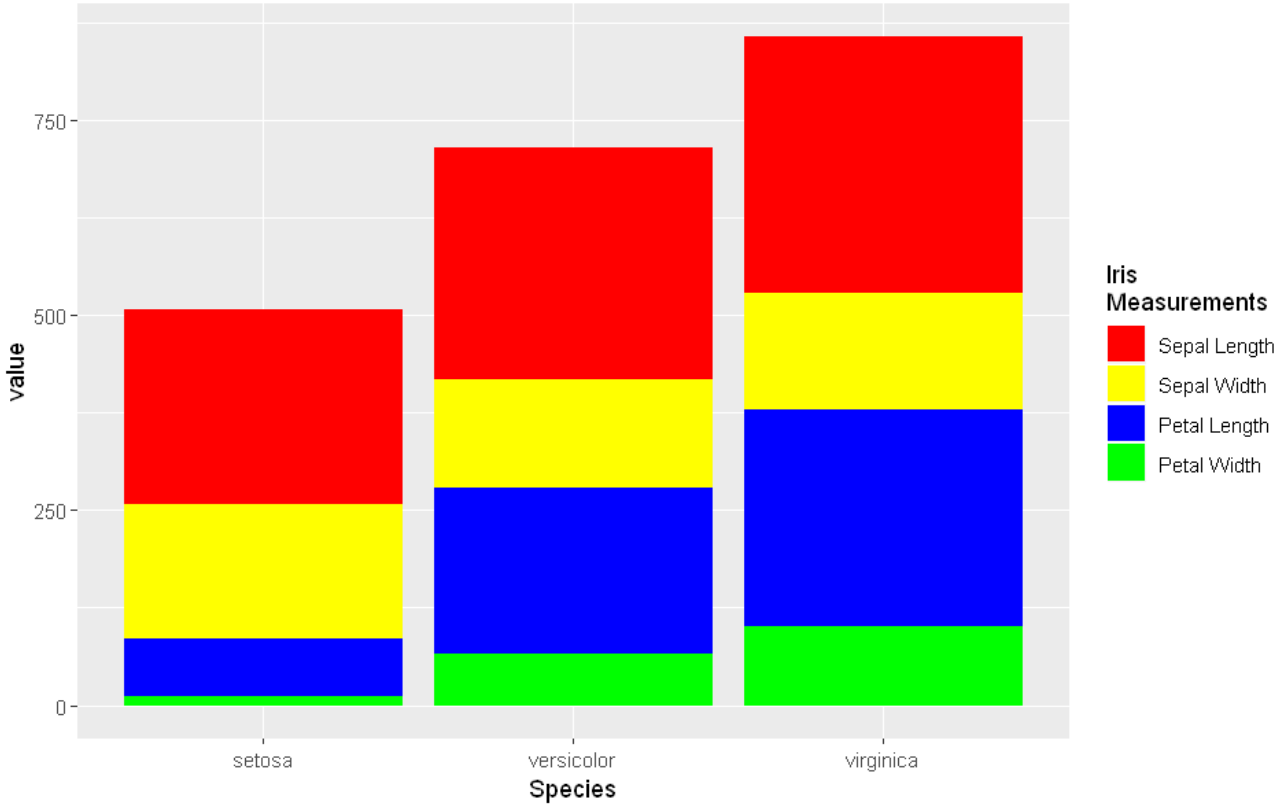


Ulster University

Data Visualisation

ggplot

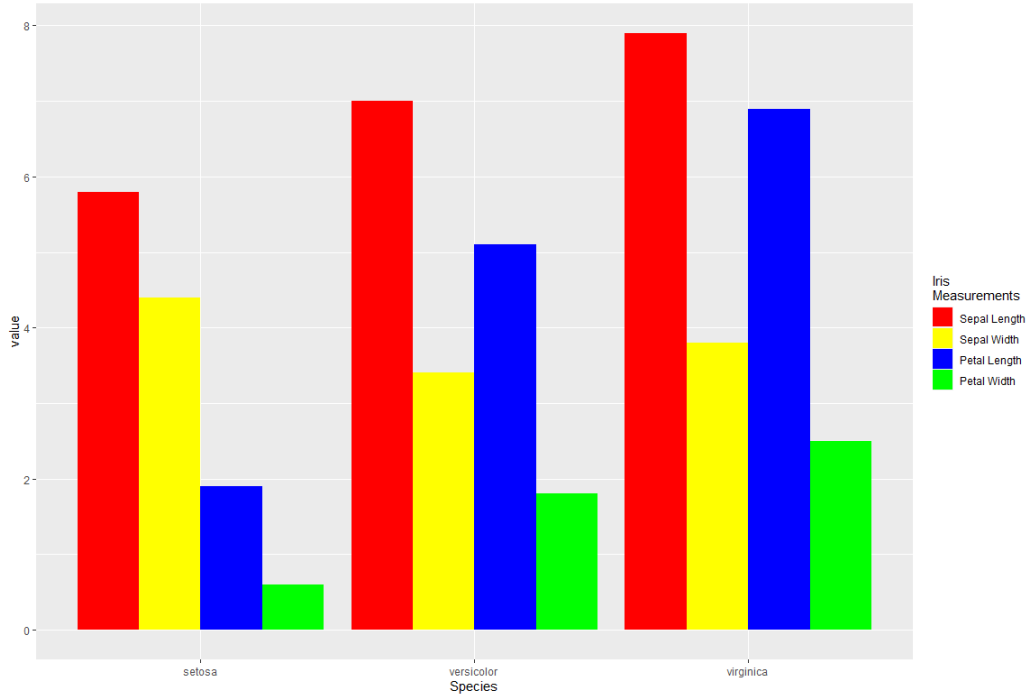
```
> library(ggplot2)
> bar1 <- ggplot(data=df2, aes(x=Species, y=value, fill=variable))
> bar1 + geom_bar(stat="identity")+
+     scale_fill_manual(values=c("red", "yellow", "blue", "green"),
+                       name="Iris\nMeasurements",
+                       breaks=c("Sepal.Length", "Sepal.Width", "Petal.Length", "Petal.Width"),
+                       labels=c("Sepal Length", "Sepal Width", "Petal Length", "Petal Width"))
> |
```



Data Visualisation
ggplot

- Use position=position_dodge()

```
> library(ggplot2)
> bar1 <- ggplot(data=df2, aes(x=Species, y=value, fill=variable))
> bar1 + geom_bar(stat="identity", position = position_dodge()) +
+     scale_fill_manual(values=c("red", "yellow", "blue", "green"),
+                       name="Iris\nMeasurements",
+                       breaks=c("Sepal.Length", "Sepal.Width", "Petal.Length", "Petal.Width"),
+                       labels=c("Sepal Length", "Sepal Width", "Petal Length", "Petal Width"))
> |
```



Data Visualisation
ggplot

- Line plot

```
> ?Orange
> head(Orange)
Grouped Data: circumference ~ age | Tree
  Tree  age circumference
1    1  118            30
2    1  484            58
3    1  664            87
4    1 1004           115
5    1 1231           120
6    1 1372           142
> |
```

Orange {datasets}

Growth of Orange Trees

Description

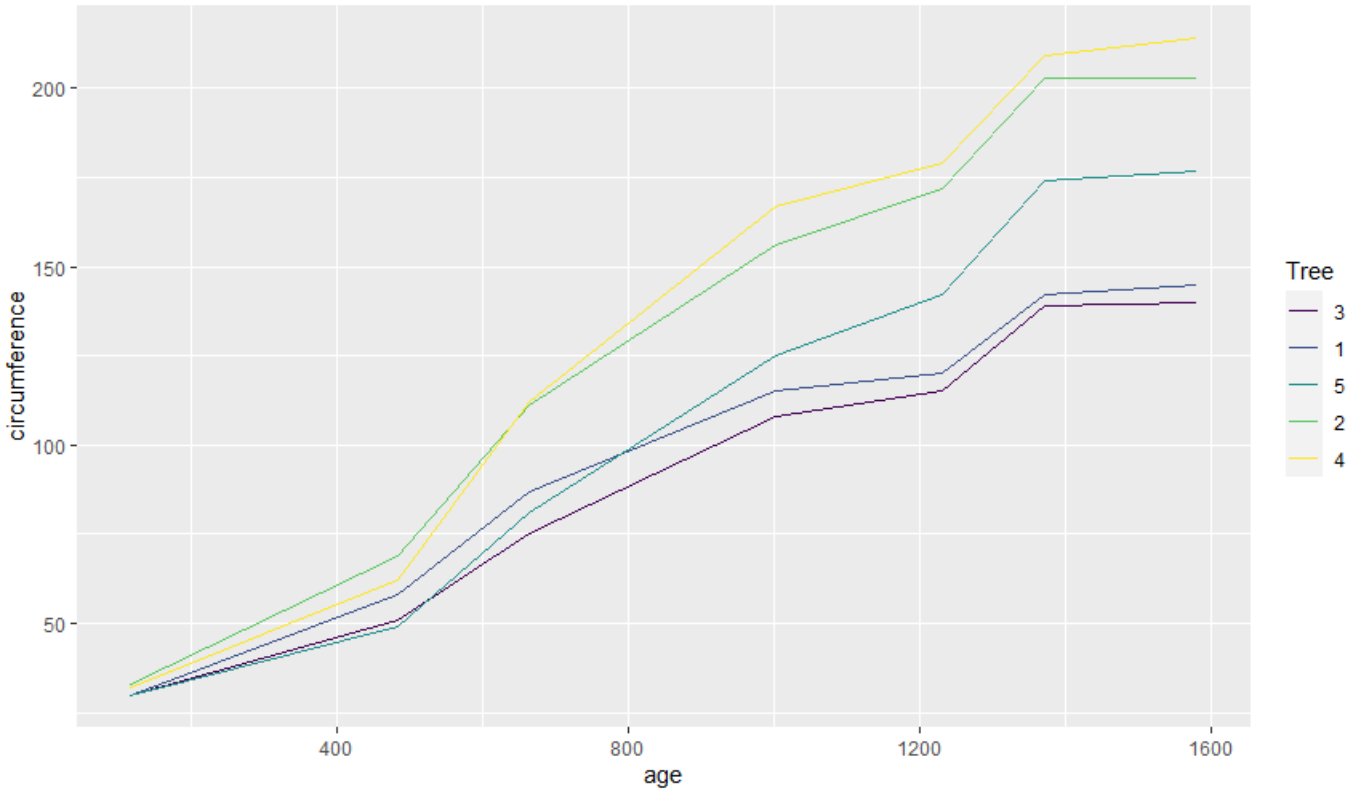
The `Orange` data frame has 35 rows and 3 columns of records of the growth of orange trees.

Ulster University

Data Visualisation
ggplot

- Line plot

```
> ggplot(Orange) + geom_line(aes(x = age, y = circumference, color = Tree))
>
```



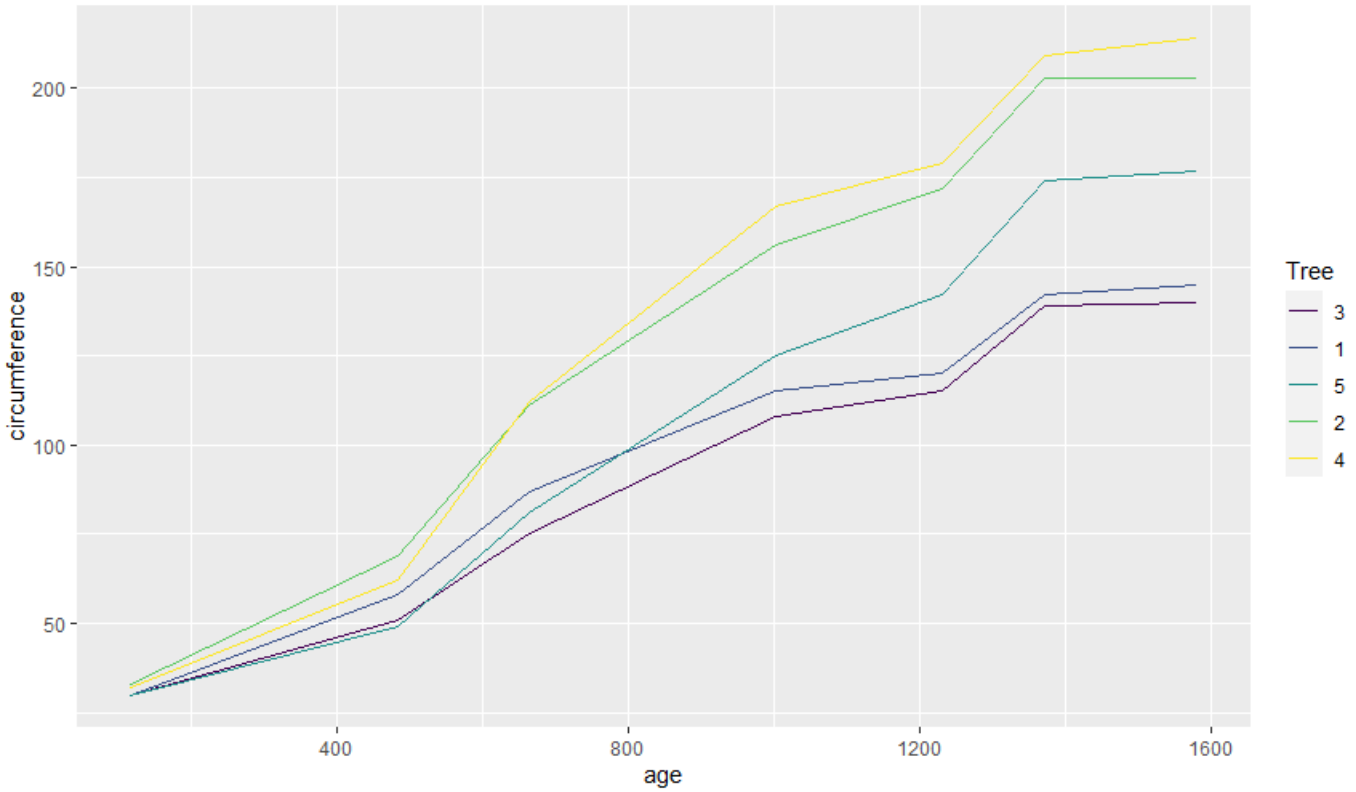
Ulster University

Data Visualisation

ggplot

- Line plot

```
> ggplot(data = Orange, aes(x = age, y = circumference, color = Tree)) + geom_line()
> 
```

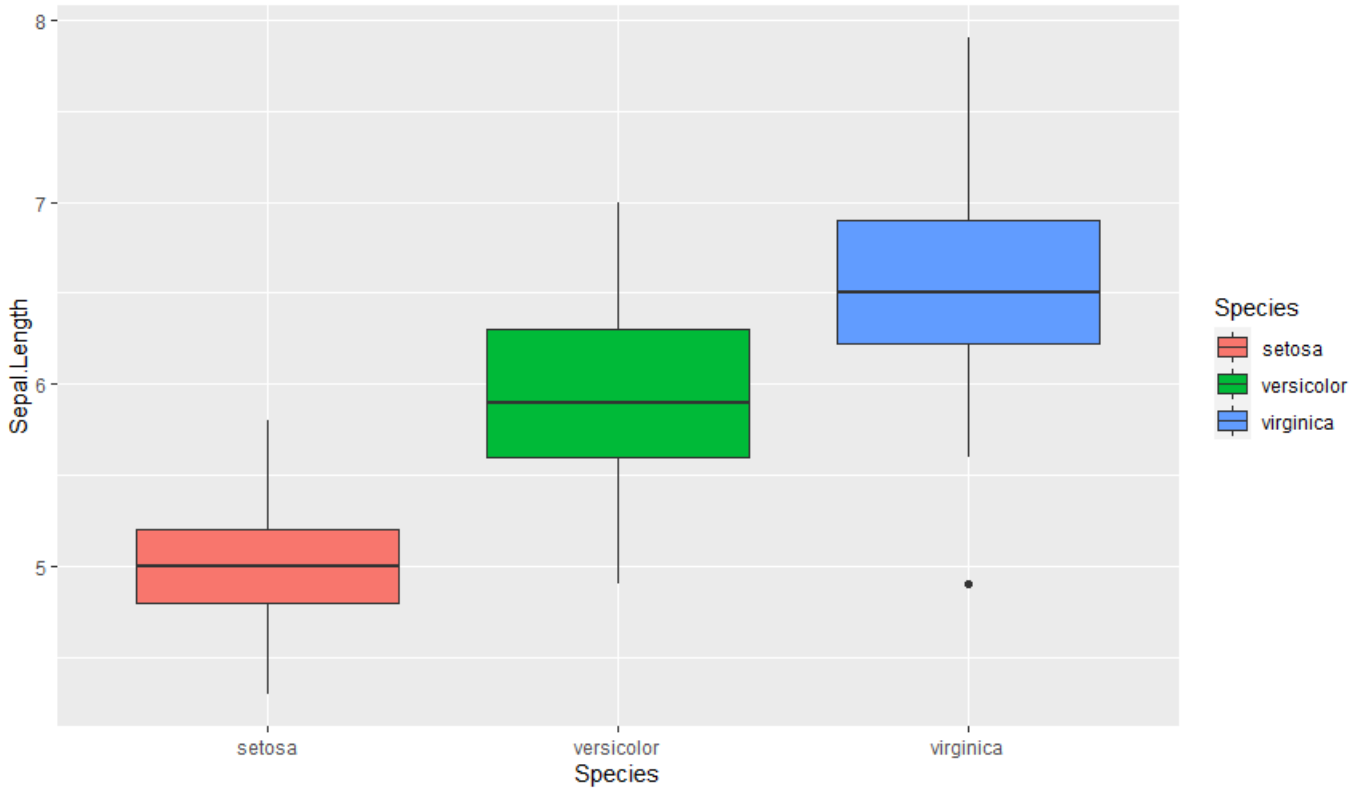


Ulster University

Data Visualisation
ggplot

- Box plot

```
> box <- ggplot(data=iris, aes(x=Species, y=Sepal.Length))
> box + geom_boxplot(aes(fill=Species))
>
```



Ulster University

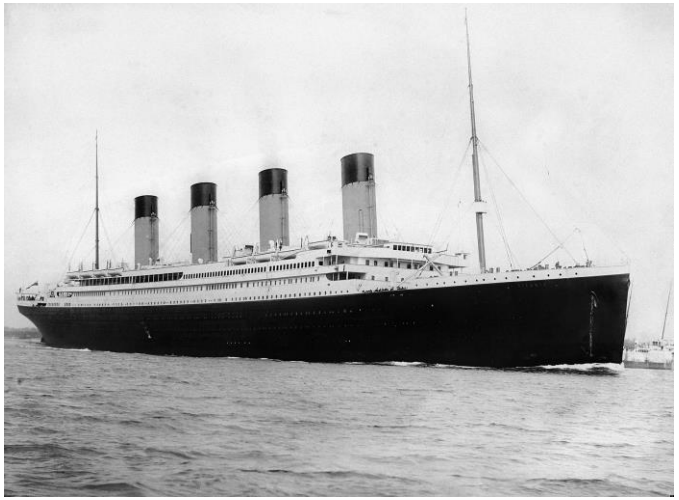
Exploratory data analysis

- Cleaning and Pre-Processing the data
- Calculating descriptive statistics
- Visualizing the data
- Density plot / Correlation plot

Ulster University

Exploratory data analysis
Exercise

- Studying the Titanic passenger survival
- Titanic dataset from Kaggle https://www.kaggle.com/c/titanic
- Carry out exploratory data analysis



RMS Titanic, By F.G.O. Stuart (1843–1923)

https://medium.com/analytics-vidhya/a-beginners-guide-to-learning-r-with-the-titanic-dataset-a630bc5495a8

Ulster University

Exploratory data analysis
Exercise – preparation

- Download dataset
- Load dataset

```
> df <- read.csv('train.csv',na.strings = '')  # empty values are read as NA values
>
```

- Install packages: install.packages()
  - visdat, dplyr, ggplot2, corrplot
- Load packages: library()

Ulster University

Exploratory data analysis

Exercise – data cleaning

- View and summaries the dataset
  - head(), summary()

```
> head(df)
  PassengerId Survived Pclass                                                 Name    Sex Age SibSp Parch            Ticket
1           1        0      3                              Braund, Mr. Owen Harris   male  22     1     0         A/5 21171
2           2        1      1 Cumings, Mrs. John Bradley (Florence Briggs Thayer) female  38     1     0          PC 17599
3           3        1      3                               Heikkinen, Miss. Laina female  26     0     0 STON/O2. 3101282
4           4        1      1         Futrelle, Mrs. Jacques Heath (Lily May Peel) female  35     1     0            113803
5           5        0      3                             Allen, Mr. William Henry   male  35     0     0            373450
6           6        0      3                                     Moran, Mr. James   male  NA     0     0            330877
     Fare Cabin Embarked
1  7.2500  <NA>        S
2 71.2833   C85        C
3  7.9250  <NA>        S
4 53.1000  C123        S
5  8.0500  <NA>        S
6  8.4583  <NA>        Q
> summary(df)
  PassengerId       Survived         Pclass          Name               Sex                 Age             SibSp
 Min.   :  1.0   Min.   :0.0000   Min.   :1.000   Length:891         Length:891         Min.   : 0.42   Min.   :0.000
 1st Qu.:223.5   1st Qu.:0.0000   1st Qu.:2.000   Class :character   Class :character   1st Qu.:20.12   1st Qu.:0.000
 Median :446.0   Median :0.0000   Median :3.000   Mode  :character   Mode  :character   Median :28.00   Median :0.000
 Mean   :446.0   Mean   :0.3838   Mean   :2.309                                         Mean   :29.70   Mean   :0.523
 3rd Qu.:668.5   3rd Qu.:1.0000   3rd Qu.:3.000                                         3rd Qu.:38.00   3rd Qu.:1.000
 Max.   :891.0   Max.   :1.0000   Max.   :3.000                                         Max.   :80.00   Max.   :8.000
                                                                                        NA's   :177
     Parch           Ticket               Fare            Cabin              Embarked
 Min.   :0.0000   Length:891         Min.   :  0.00   Length:891         Length:891
 1st Qu.:0.0000   Class :character   1st Qu.:  7.91   Class :character   Class :character
 Median :0.0000   Mode  :character   Median : 14.45   Mode  :character   Mode  :character
 Mean   :0.3816                      Mean   : 32.20
 3rd Qu.:0.0000                      3rd Qu.: 31.00
 Max.   :6.0000                      Max.   :512.33
```

Exploratory data analysis
Exercise – data cleaning

- Missing values
  - identify count of NAs in data frame

```
> sum(is.na(df))
[1] 866
>
```

  - compute the total missing values in each column

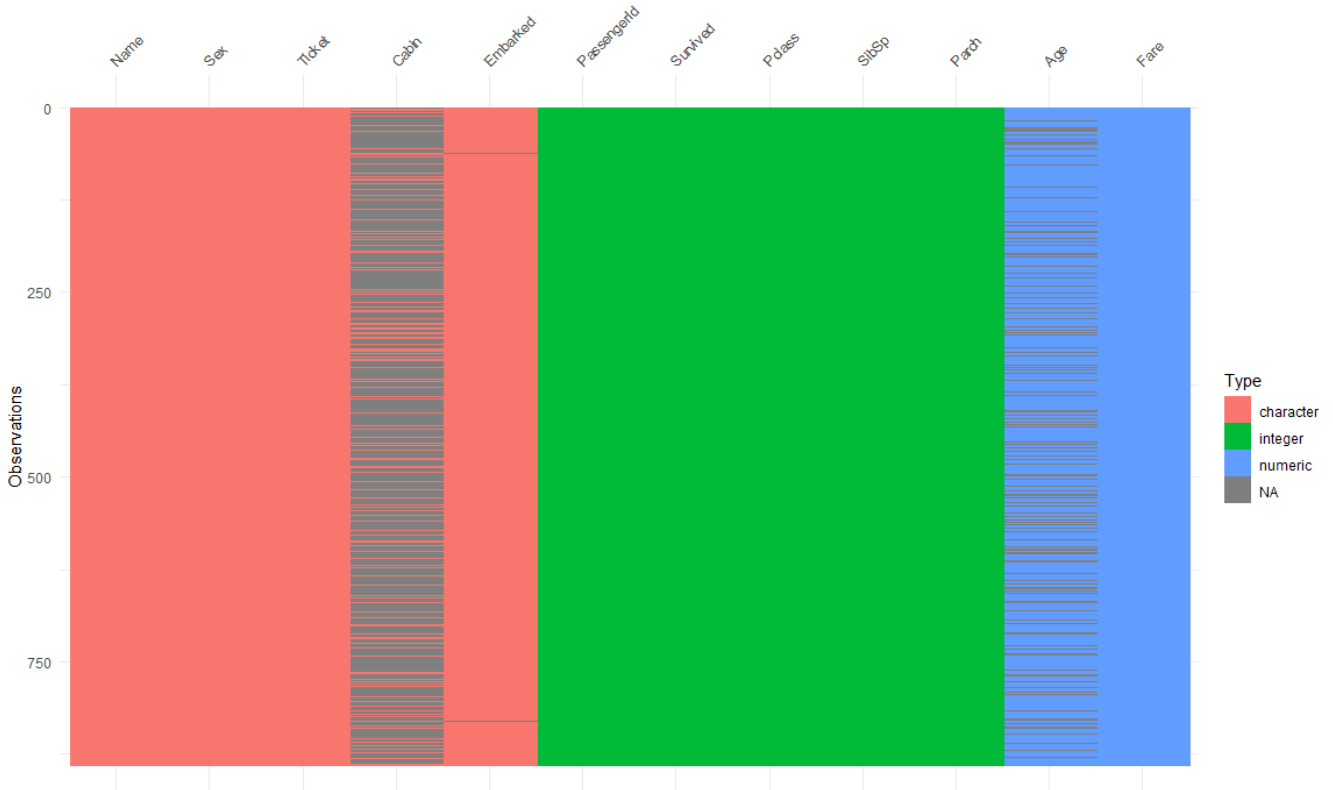
```
> colSums(is.na(df))
PassengerId    Survived      Pclass        Name         Sex         Age       SibSp       Parch      Ticket        Fare       Cabin    Embarked
          0           0           0           0           0         177           0           0           0           0         687           2
>
```

Ulster University

Exploratory data analysis
Exercise – data cleaning

- Missing values
  - Visulising missing values
  - library('visdat')

```
> library('visdat')
> vis_dat(df)
>
```



Exploratory data analysis
Exercise – data cleaning

- Handle the missing values
  - Remove missing values
  - Recode missing values using mean/median (numerical variables), using most frequent category (categorical variables)
  - other imputation methods: LOCF (last observation carried forward)

Ulster University

Exploratory data analysis

Exercise – data cleaning

- Select the useful columns and drop the columns with NA values: select ()
- Drop the rows with NA values: na.omit()

```
> df <- read.csv('train.csv',na.strings = '')
> str(df)
'data.frame':   891 obs. of  12 variables:
 $ PassengerId: int  1 2 3 4 5 6 7 8 9 10 ...
 $ Survived   : int  0 1 1 1 0 0 0 0 1 1 ...
 $ Pclass     : int  3 1 3 1 3 3 1 3 3 2 ...
 $ Name       : chr  "Braund, Mr. Owen Harris" "Cumings, Mrs. John Bradley (Florence Briggs Thayer)" "Heikkinen, Miss. Laina" "Futrelle, Mrs. Jacques Heath (Lily May Pee
l)" ...
 $ Sex        : chr  "male" "female" "female" "female" ...
 $ Age        : num  22 38 26 35 35 NA 54 2 27 14 ...
 $ SibSp      : int  1 1 0 1 0 0 0 3 0 1 ...
 $ Parch      : int  0 0 0 0 0 0 0 1 2 0 ...
 $ Ticket     : chr  "A/5 21171" "PC 17599" "STON/O2. 3101282" "113803" ...
 $ Fare       : num  7.25 71.28 7.92 53.1 8.05 ...
 $ Cabin      : chr  NA "C85" NA "C123" ...
 $ Embarked   : chr  "S" "C" "S" "S" ...
>
```

```
> library(dplyr)
> df <- select(df, Survived, Pclass, Age, Sex, SibSp, Parch)
> df <- na.omit(df)
> str(df)
'data.frame':   714 obs. of  6 variables:
 $ Survived: int  0 1 1 1 0 0 0 1 1 1 ...
 $ Pclass  : int  3 1 3 1 3 1 3 3 2 3 ...
 $ Age     : num  22 38 26 35 35 54 2 27 14 4 ...
 $ Sex     : chr  "male" "female" "female" "female" ...
 $ SibSp   : int  1 1 0 1 0 0 3 0 1 1 ...
 $ Parch   : int  0 0 0 0 0 0 1 2 0 1 ...
 - attr(*, "na.action")= 'omit' Named int [1:177] 6 18 20 27 29 30 32 33 37 43 ...
  ..- attr(*, "names")= chr [1:177] "6" "18" "20" "27" ...
>
```

Exploratory data analysis
Exercise – data cleaning

- Convert integer values to categorical values

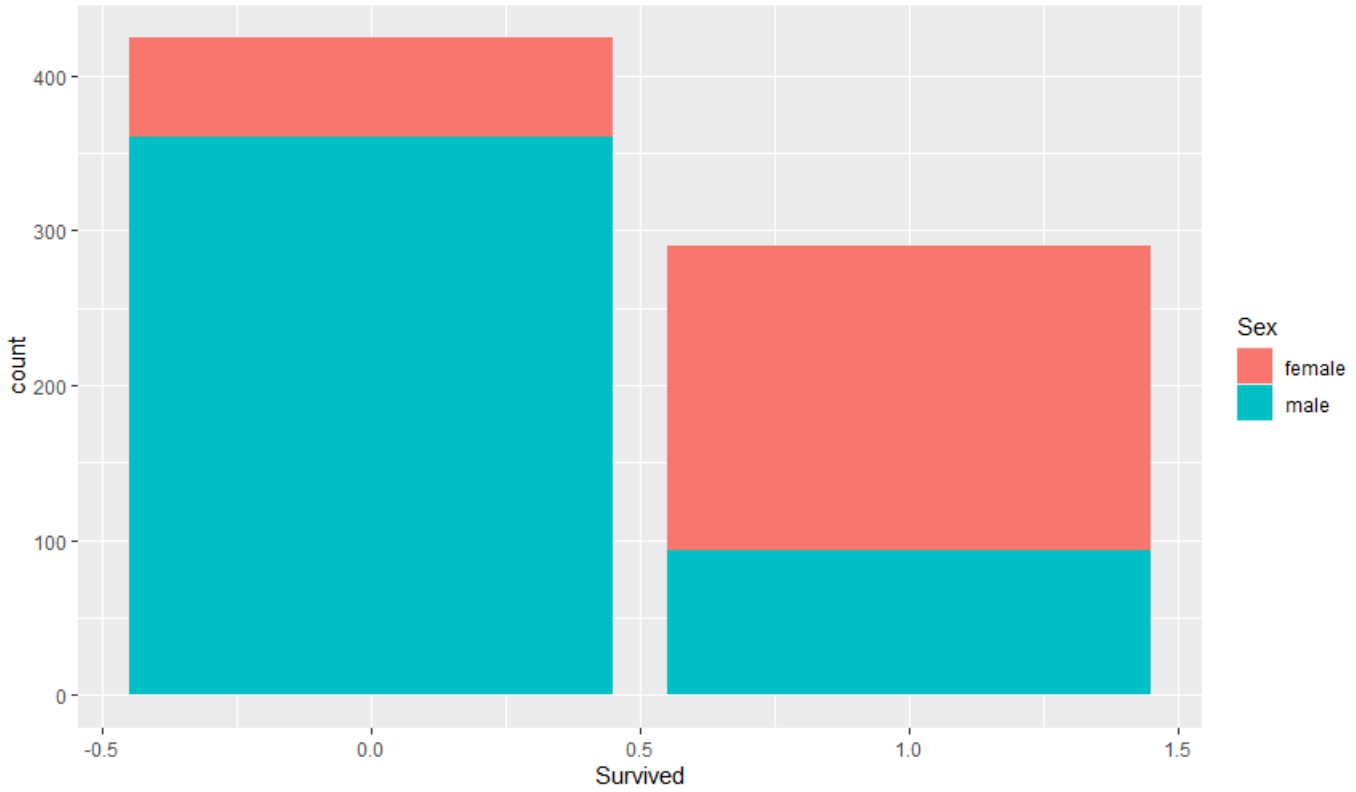
```
> df$Survived = factor(df$Survived)
> df$Pclass = factor(df$Pclass)
>
> str(df)
'data.frame':   714 obs. of  6 variables:
 $ Survived: Factor w/ 2 levels "0","1": 1 2 2 2 1 1 1 2 2 2 ...
 $ Pclass  : Factor w/ 3 levels "1","2","3": 3 1 3 1 3 1 3 3 2 3 ...
 $ Age     : num  22 38 26 35 35 54 2 27 14 4 ...
 $ Sex     : chr  "male" "female" "female" "female" ...
 $ SibSp   : int  1 1 0 1 0 0 3 0 1 1 ...
 $ Parch   : int  0 0 0 0 0 0 1 2 0 1 ...
 - attr(*, "na.action")= 'omit' Named int [1:177] 6 18 20 27 29 30 32 33 37 43 ...
  ..- attr(*, "names")= chr [1:177] "6" "18" "20" "27" ...
> 
```

Ulster University

Exploratory data analysis
Exercise - visualisation

- Visualisation
  - Survival by Sex

```
> ggplot(df, aes(x = Survived, fill=Sex)) +
+     geom_bar()
> |
```

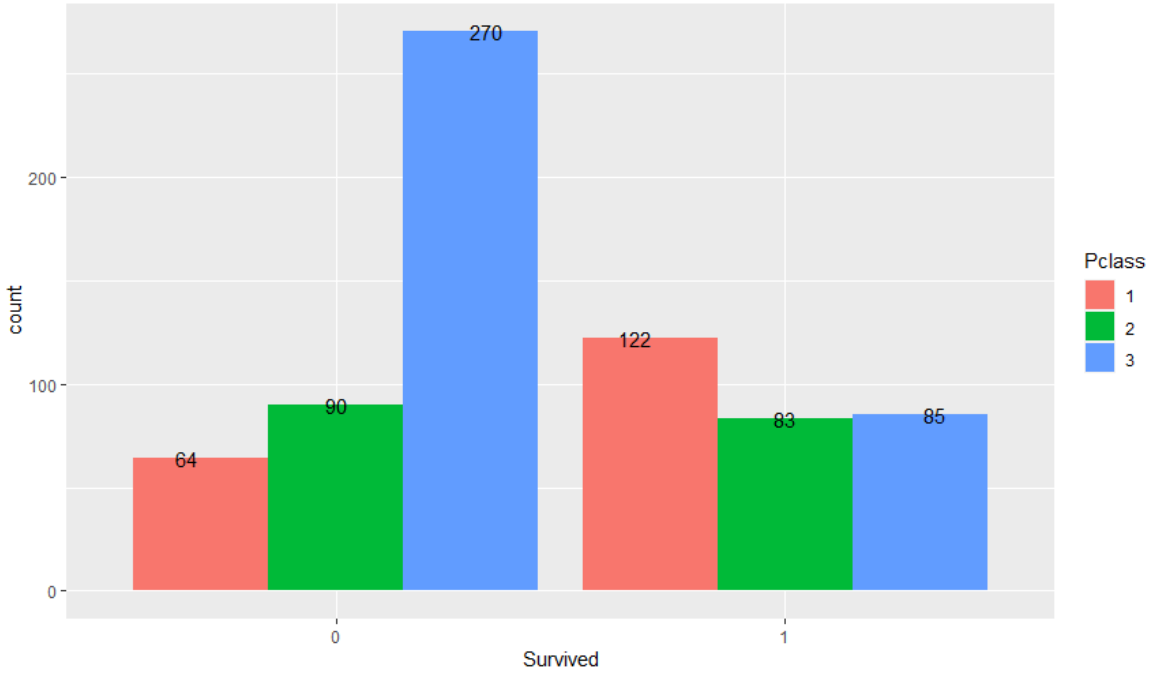


Ulster University

Exploratory data analysis
Exercise - visualisation

- Visualisation
  - Survival by Pclass

```
> ggplot(df, aes(x=Survived, fill=Pclass))+
+     geom_bar(position = position_dodge())+
+     geom_text(stat='count',
+                   aes(label=stat(count)),
+                   position = position_dodge(width = 1))
>
```



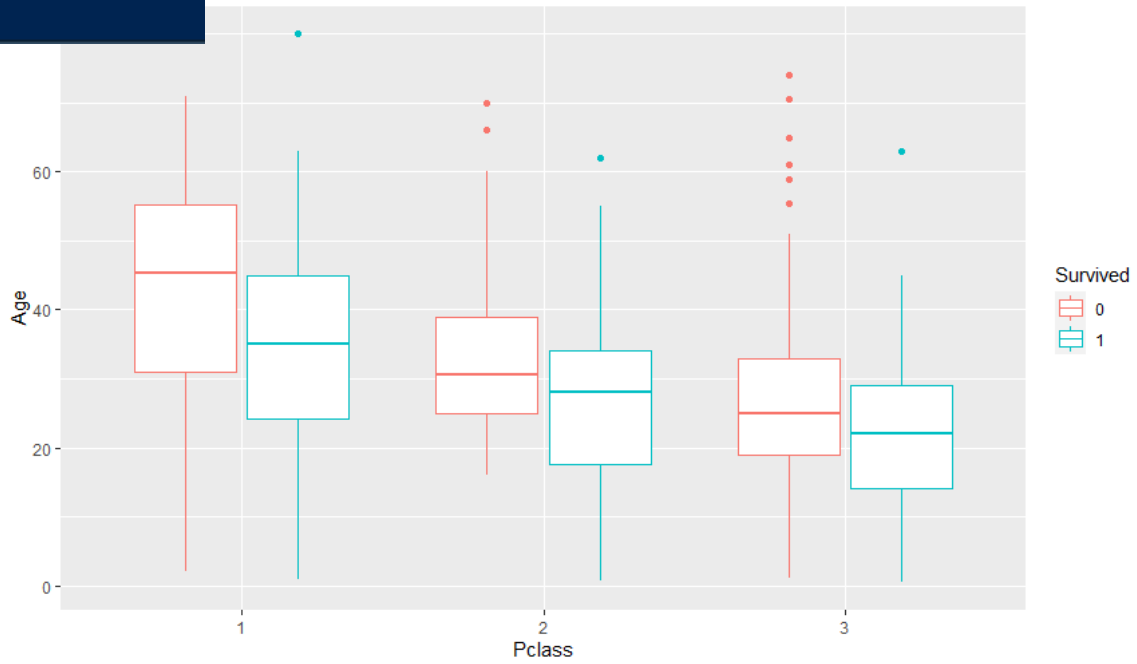
Ulster University

Exploratory data analysis

Exercise - visualisation

- Pclass vs Age
  - To compare a categorical variable like Pclass with a continuous variable like Age there are several useful visualizations. Here we will use a boxplot.

```
> ggplot(df, aes(Pclass, Age, colour = Survived)) +
+     geom_boxplot()
>
```



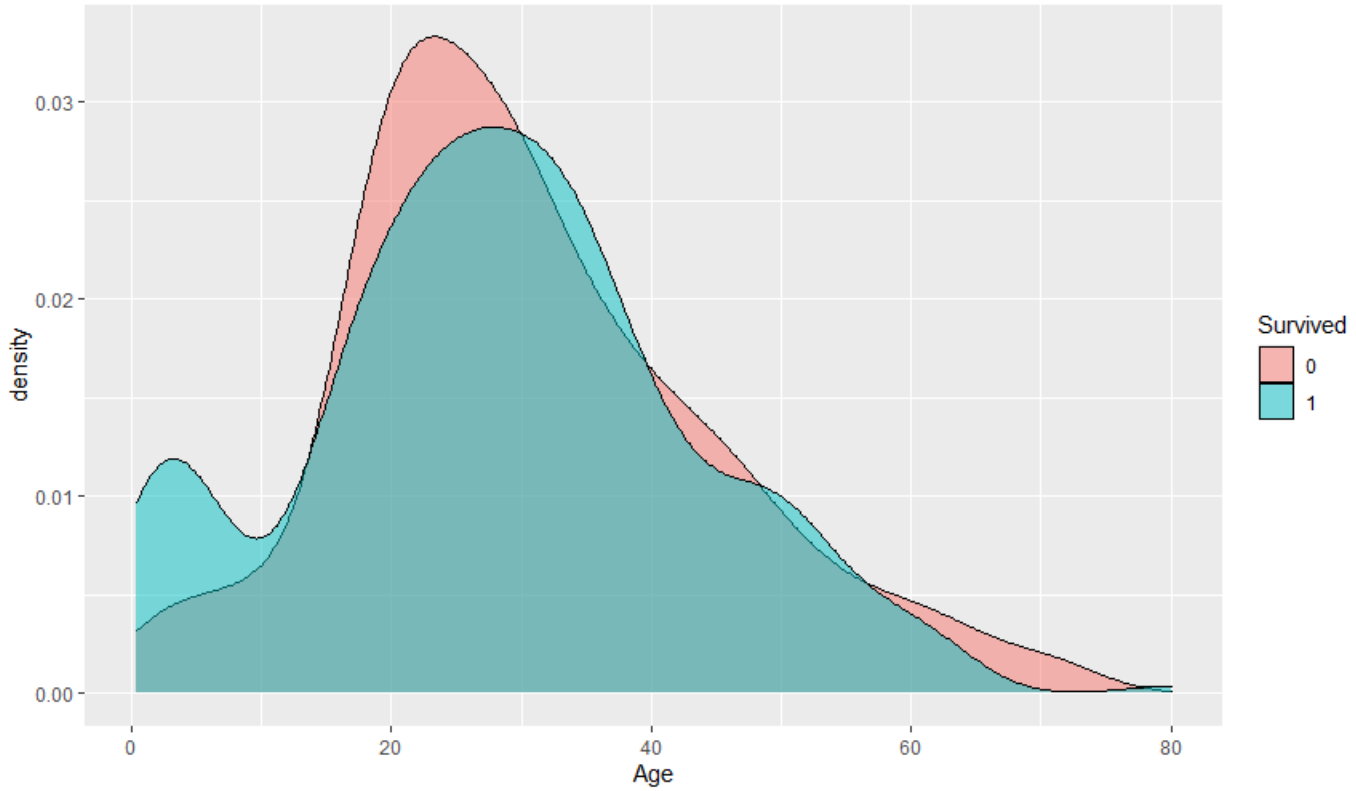
Ulster University

Exploratory data analysis

Exercise - density plot

- Density plot – Survived by Age

```
> ggplot(df, aes(x=Age))+
+ geom_density(aes(fill = Survived), alpha = 0.5)
>
```

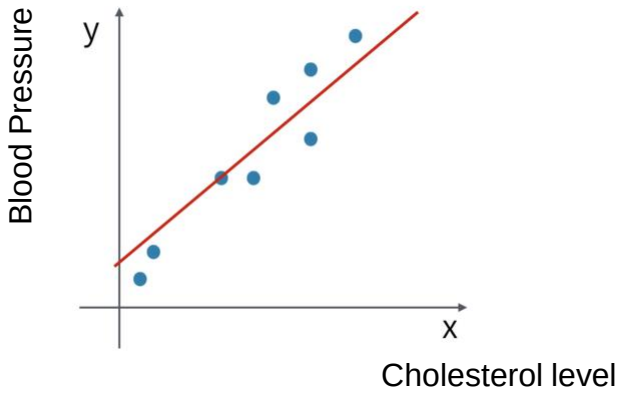


Ulster University

Exploratory data analysis
Correlation analysis

- Correlation analysis explores a relationship between two or more variables.
  - However, seeing two variables moving together does not necessarily mean we know whether one variable causes the other to occur.
  - This is why we commonly say “correlation does not imply causation.”

https://www.jmp.com/en_gb/statistics-knowledge-portal/what-is-correlation/correlation-vs causation.html#:~:text=Correlation%20tests%20for%20a%20relationship,correlation%2(does%20not%20imply%20causation.%E2%80%9D



Exploratory data analysis
Correlation formula

- Pearson correlation formula
- Spearman correlation formula
- Kendall correlation formula

$$r = \frac{\sum (x - m_x)(y - m_y)}{\sqrt{\sum (x - m_x)^2 \sum (y - m_y)^2}}$$

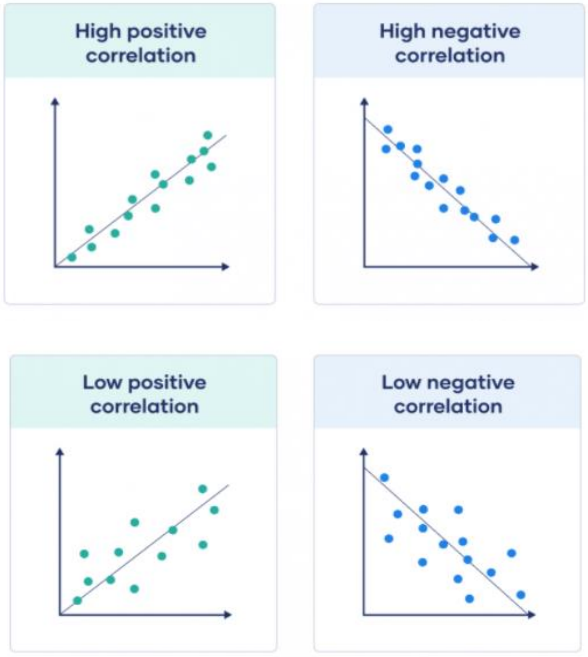
$$rho = \frac{\sum (x' - m_{x'})(y'_i - m_{y'})}{\sqrt{\sum (x' - m_{x'})^2 \sum (y' - m_{y'})^2}}$$

$$tau = \frac{n_c - n_d}{\frac{1}{2} n(n-1)}$$

http://www.sthda.com/english/wiki/correlation-test-between-two-variables-in-r

Correlation coefficient	Correlation strength	Correlation type
-.7 to -1	Very strong	Negative
-.5 to -.7	Strong	Negative
-.3 to -.5	Moderate	Negative
0 to -.3	Weak	Negative
0	None	Zero
0 to .3	Weak	Positive
.3 to .5	Moderate	Positive
.5 to .7	Strong	Positive
.7 to 1	Very strong	Positive

https://www.scribbr.com/statistics/correlation-coefficient/



Exploratory data analysis
Exercise - correlation plot

- Correlation plot
  - Covert categorical values to numerical values

```
> df_corr <- df
> df_corr$Survived <- as.numeric(df_corr$Survived)
> df_corr$Pclass <- as.numeric(df_corr$Pclass)
> df_corr$Sex <- ifelse(df_corr$Sex=='male', 1, 0)
> str(df_corr)
'data.frame':   714 obs. of  6 variables:
 $ Survived: num  1 2 2 2 1 1 1 2 2 2 ...
 $ Pclass  : num  3 1 3 1 3 1 3 3 2 3 ...
 $ Age     : num  22 38 26 35 35 54 2 27 14 4 ...
 $ Sex     : num  1 0 0 0 1 1 1 0 0 0 ...
 $ SibSp   : int  1 1 0 1 0 0 3 0 1 1 ...
 $ Parch   : int  0 0 0 0 0 0 1 2 0 1 ...
 - attr(*, "na.action")= 'omit' Named int [1:177] 6 18 20 27 29 30 32 33 37 43 ...
  ..- attr(*, "names")= chr [1:177] "6" "18" "20" "27" ...
> 
```

Ulster University

Exploratory data analysis
Correlation matrix

- The R function **cor()** can be used to compute a **correlation matrix**
  - **cor(df_corr)**

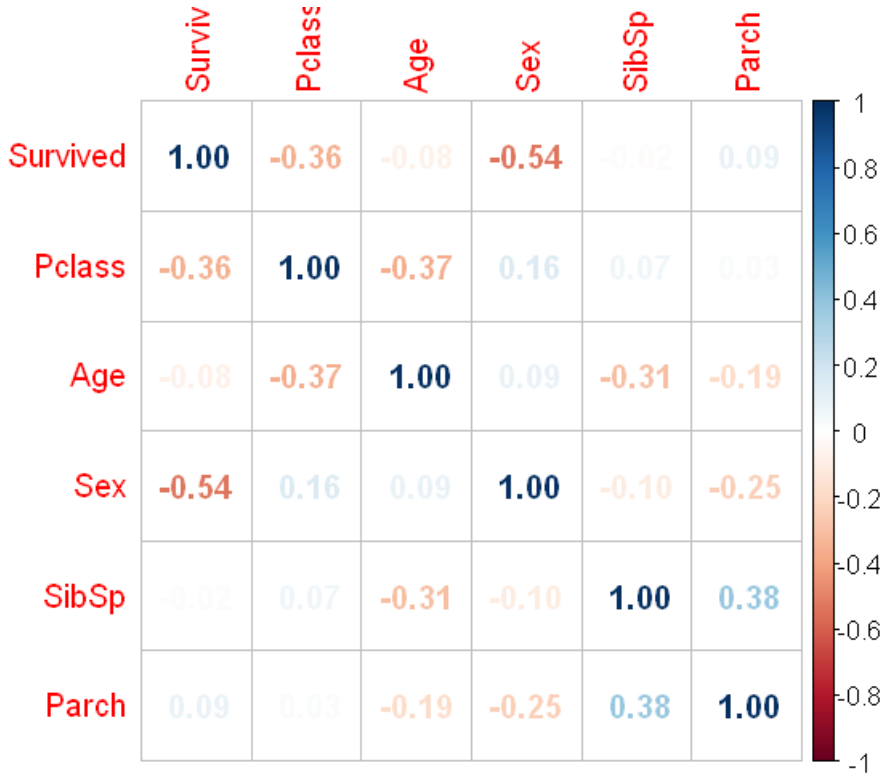
	Survived	Pclass	Age	Sex	SibSp	Parch
Survived	1.00000000	-0.35965268	-0.07722109	-0.53882559	-0.01735836	0.09331701
Pclass	-0.35965268	1.00000000	-0.36922602	0.15546030	0.06724737	0.02568307
Age	-0.07722109	-0.36922602	1.00000000	0.09325358	-0.30824676	-0.18911926
Sex	-0.53882559	0.15546030	0.09325358	1.00000000	-0.10394968	-0.24697204
SibSp	-0.01735836	0.06724737	-0.30824676	-0.10394968	1.00000000	0.38381986
Parch	0.09331701	0.02568307	-0.18911926	-0.24697204	0.38381986	1.00000000

Ulster University

Exploratory data analysis
Exercise - correlation plot

- Correlation plot – visualising correlation matrix

```
> M <- cor(df_corr)
> corrplot(M, method = 'number')
>
```

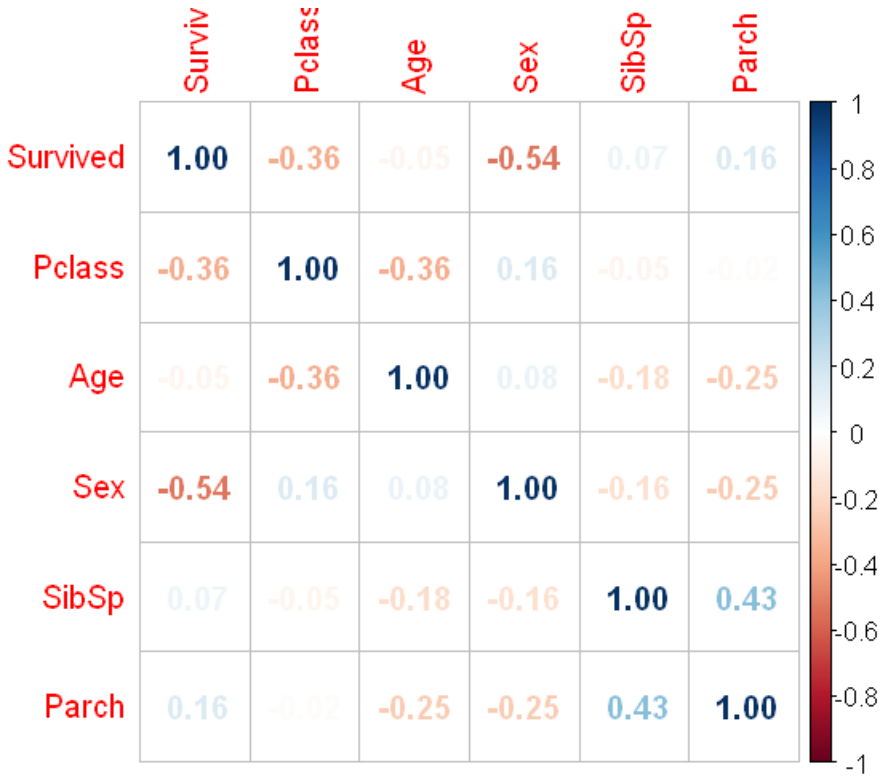


Ulster University

Exploratory data analysis
Exercise - correlation plot

- Correlation plot – visualising correlation matrix

```
M <- cor(df_corr, method = "spearman")
corrplot(M, method = 'number')
```



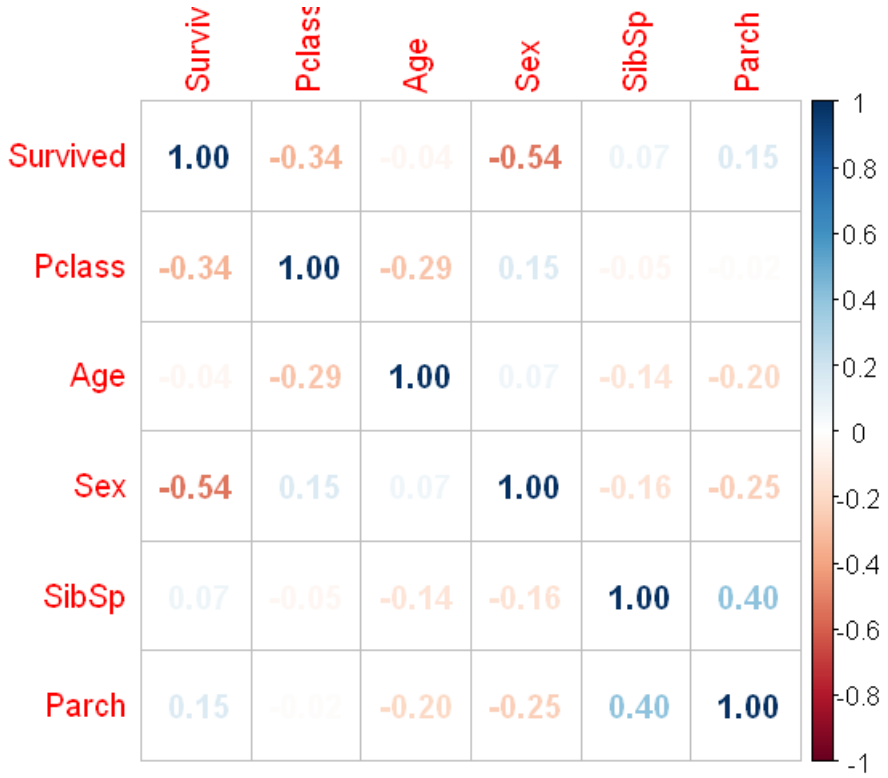
Ulster University

Exploratory data analysis
Exercise - correlation plot

- Correlation plot – visualising correlation matrix

  M <- cor(df_corr, method = "kendall")

  corrplot(M, method = 'number')



Ulster University

Exploratory data analysis

Exercise - correlation plot

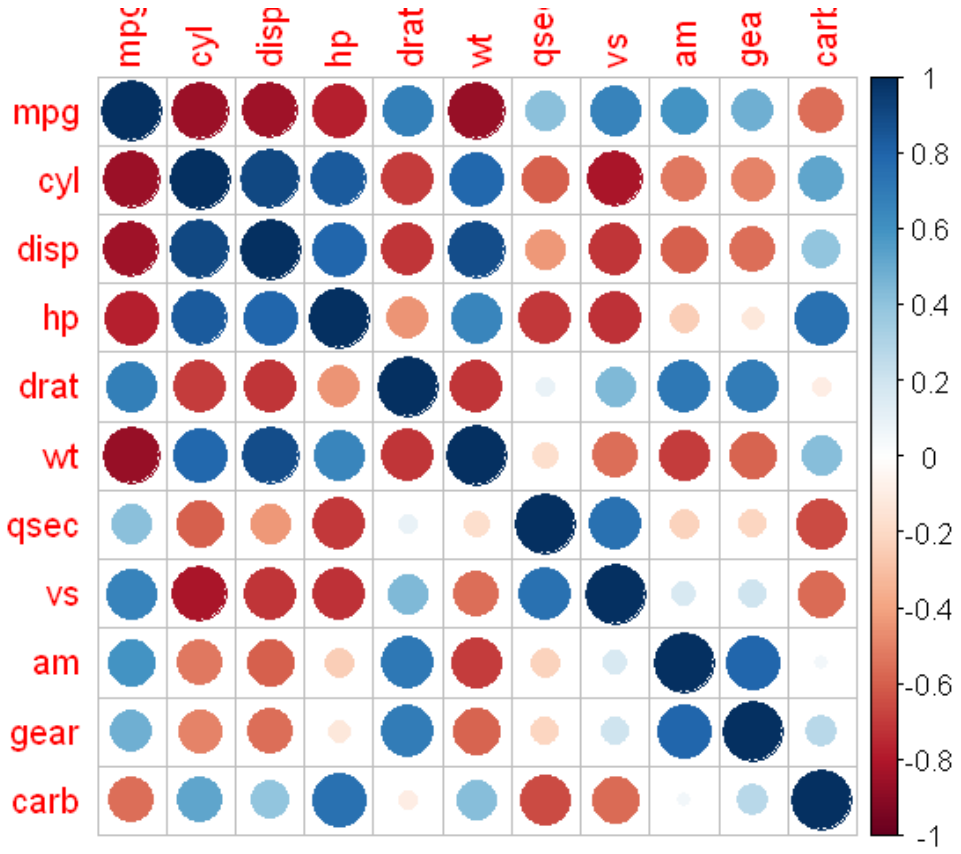
- corrplot(cor(mtcars))

```
> str(mtcars)
'data.frame':   32 obs. of  11 variables:
 $ mpg : num  21 21 22.8 21.4 18.7 18.1 14.3 24.4 22.8 19.2 ...
 $ cyl : num  6 6 4 6 8 6 8 4 4 6 ...
 $ disp: num  160 160 108 258 360 ...
 $ hp  : num  110 110 93 110 175 105 245 62 95 123 ...
 $ drat: num  3.9 3.9 3.85 3.08 3.15 2.76 3.21 3.69 3.92 3.92 ...
 $ wt  : num  2.62 2.88 2.32 3.21 3.44 ...
 $ qsec: num  16.5 17 18.6 19.4 17 ...
 $ vs  : num  0 0 1 1 0 1 0 1 1 1 ...
 $ am  : num  1 1 1 0 0 0 0 0 0 0 ...
 $ gear: num  4 4 4 3 3 3 3 4 4 4 ...
 $ carb: num  4 4 1 1 2 1 4 2 2 4 ...
> 
```

Ulster University

Exploratory data analysis
Exercise - correlation plot

- corrplot(cor(mtcars))



Ulster University

Open Datasets

- UCI machine learning Repository
  - https://archive.ics.uci.edu/ml/index.php
- Kaggle
  - https://www.kaggle.com/datasets
- OPEN Data NI
  - https://www.opendatani.gov.uk/

Ulster University

Further reading

- Gareth, J., Daniela, W., Trevor, H. and Robert, T., 2013. *An introduction to statistical learning: with applications in R*. Spinger. (Chapter 3)
- Lantz, Brett. Machine learning with R. Packt Publishing Ltd, 2013 (Chapter 2)
- R Visulisation
  - https://www.publichealth.columbia.edu/sites/default/files/media/fdawg_ggplot2.html
  - http://www.sthda.com/english/wiki/data-visualization
- Correlation:
  - https://www.scribbr.com/statistics/correlation-coefficient/
  - An Introduction to corrplot Package
    https://cran.r-project.org/web/packages/corrplot/vignettes/corrplot-intro.html
- Exploratory data analysis
  - https://bookdown.org/steve_midway/DAR/exploratory-data-analysis.html

Ulster University