Education Scotland
Foghlam Alba

This resource is supported by *Education Scotland's Enhancing Professional Learning in STEM Grants Programme* through the Scottish Government STEM Education and Training Strategy.

SCIENCE
TECHNOLOGY
ENGINEERING
MATHEMATICS

Education and Training
Strategy for Scotland

Scottish Government
Riaghaltas na h-Alba
gov.scot

For Scotland's Learners, with Scotland's Educators

Working with Data in RStudio

Workbook for the R programming language using RStudio in Noteable

CONTINUOUS LEARNING MATERIALS FOR MATHEMATICS AND STATISTICS TEACHERS FOLLOWING THE SCOTTISH CURRICULUM

Contents of this workbook

This workbook serves to provide information, helpful links and activities that can be used in classroom with Noteable in the R programming language and using the RStudio user interface with Noteable.

Getting Started:.. 2

Accessing RStudio with Noteable 4

Getting started with RStudio in Noteable 5

How to access the in-built RStudio datasets with Noteable 5

Example Activities 6

Activity 1 7

Compare the effect of the different feed types on the weights of chicks 7

Activity 2 10

Compare the effectiveness of insect sprays A to F 10

Activity 3 – Correlation and regression 12

Find any correlation between car speed and stopping distance. Use any model you find to make a prediction. 12

Activity 4 – Testing Hypotheses 15

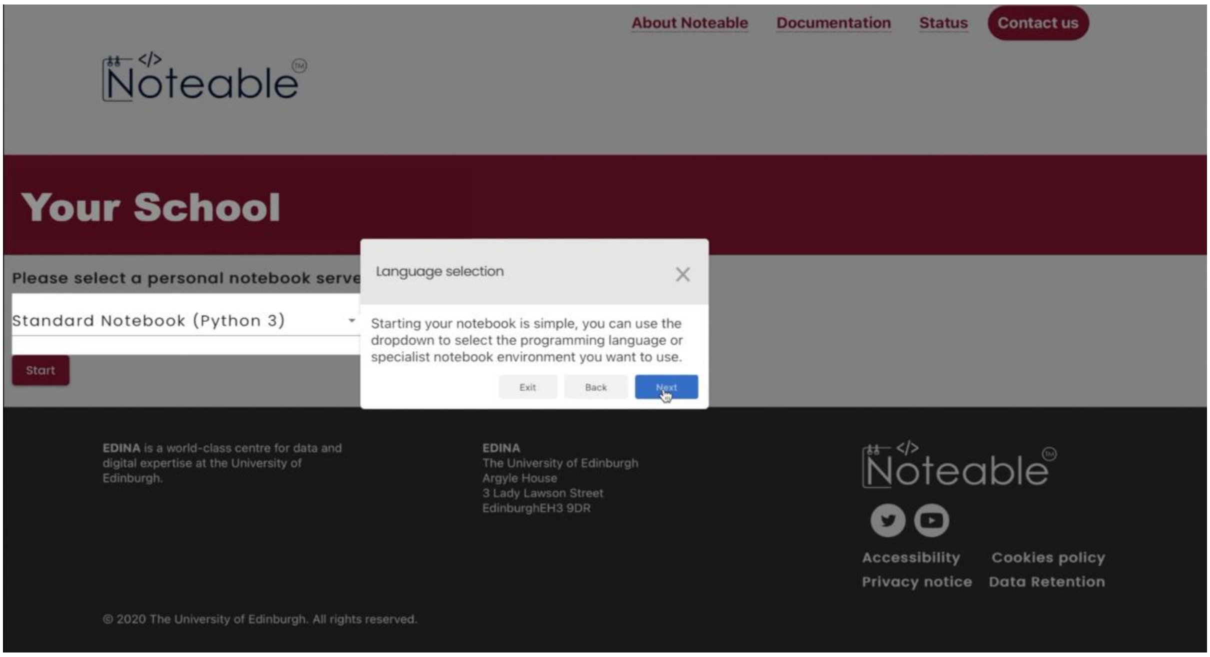
Testing a Hypothesis 15

Getting Started:

You can organise your classroom files, including files for teaching and learning coding with, and associated data files, with Noteable.

To do this, please access Noteable through the GLOW App Library. For information on accessing Noteable through GLOW, please go to the 'Working with Data - ABCs of Coding with Noteable' resource.

When you have clicked on the Noteable App tile in the GLOW App Library, you will be taken to the launch page with the Noteable Guided Tour the first time you access, which will look like this:



Starting your notebook is simple, you can use the dropdown to select the programming language or specialist notebook environment you want to use.

EDINA is a world-class centre for data and digital expertise at the University of Edinburgh.

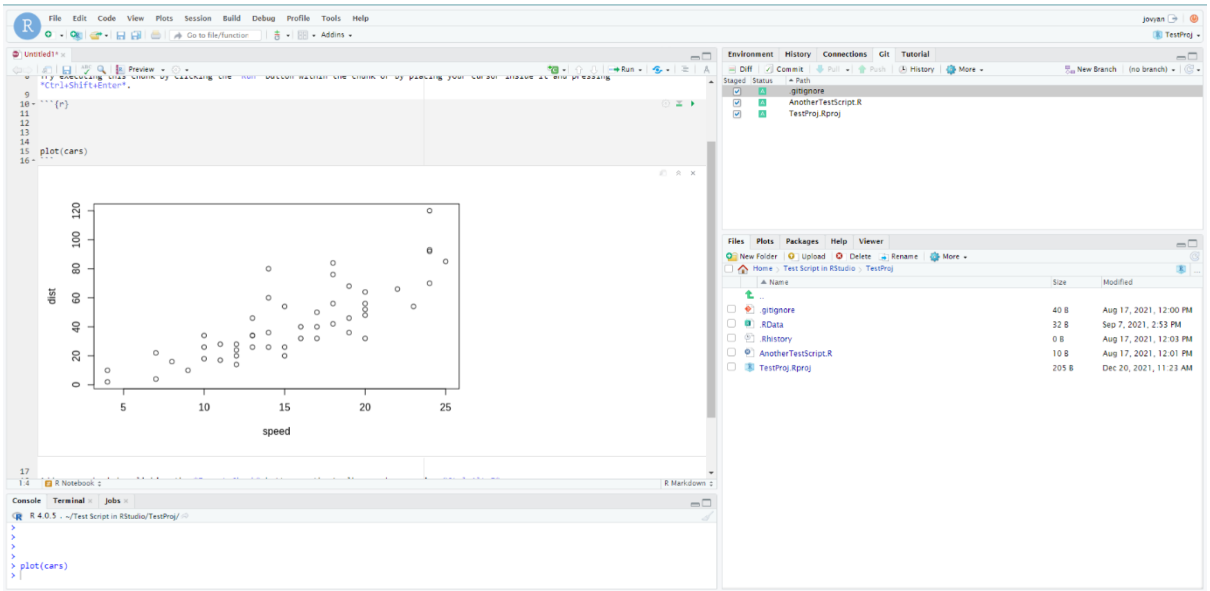
EDINA
The University of Edinburgh
Argyle House
3 Lady Lawson Street
EdinburghEH3 9DR

© 2020 The University of Edinburgh. All rights reserved.

*This workbook focusses on the R programming language, so you will need to choose the 'RStudio Notebook' from the drop-down menu when first logging in.

To launch a new notebook, click on 'New' and then under the subheading ''Notebook' click on **RStudio.** This will launch a new RStudio notebook.

Once you have accessed Noteable, if it is the first time you access RStudio you will find the RStudio interface, which will look like this:



~~~
Try executing this chunk by clicking the *Run* button within the chunk or by placing your cursor inside it and pressing
*Ctrl+Shift+Enter*.

```{r}




plot(cars)
```
~~~

Staged	Status	Path
☑	A	.gitignore
☑	A	AnotherTestScript.R
☑	A	TestProj.Rproj

	Name	Size	Modified
	..		
□	.gitignore	40 B	Aug 17, 2021, 12:00 PM
□	.RData	32 B	Sep 7, 2021, 2:53 PM
□	.Rhistory	0 B	Aug 17, 2021, 12:03 PM
□	AnotherTestScript.R	10 B	Aug 17, 2021, 12:01 PM
□	TestProj.Rproj	205 B	Dec 20, 2021, 11:23 AM

```
R 4.0.5 . ~/Test Script in RStudio/TestProj/
>
>
>
>
> plot(cars)
>
```

There are a number of available activities to get started with RStudio on Noteable.

These examples make use of the internal data sets that are already available, or built-in, when you launch RStudio on Noteable.

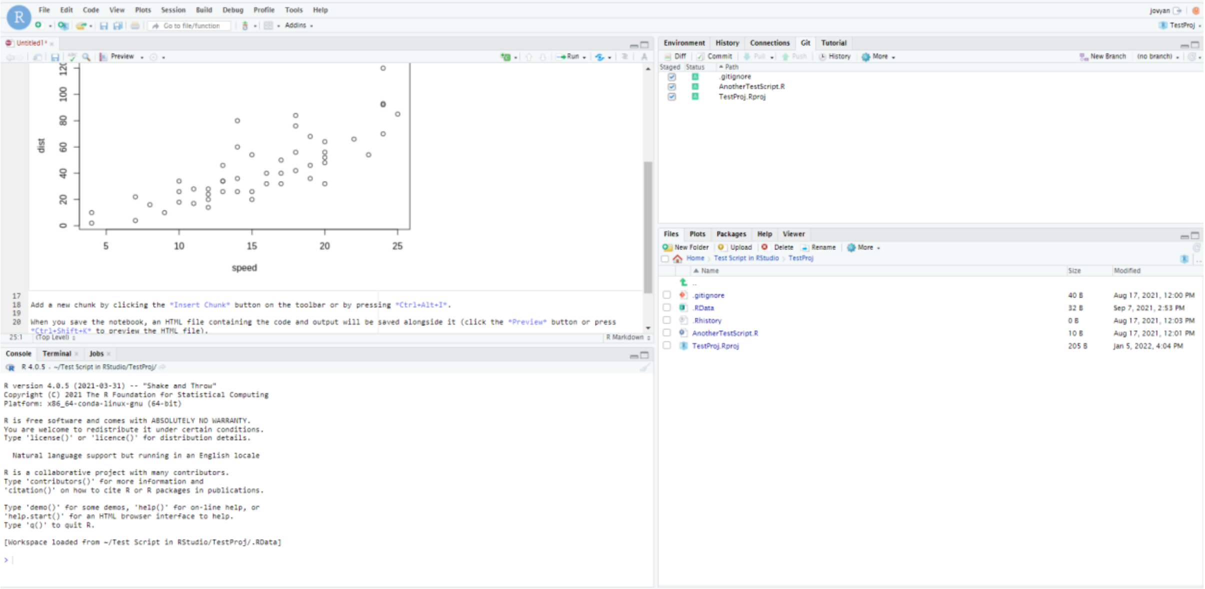
Each example is available as an individual R file, focussing on Higher Apps tasks that use the built-in data sets in RStudio

Accessing RStudio with Noteable

To launch RStudio with Noteable, once you have accessed Noteable through GLOW, you will have access to environments for coding activities, explanations, showing input and output, visualisations and more combined in one file and place. You can select RStudio from the notebook selection dropdown menu. *

When you first access an RStudio notebook with Noteable, you will see 4 panels:

- top-left: scripts and files
- bottom-left: R console
- top-right: objects, history and environment
- bottom-right: tree of folders, graph window, packages, help window, viewer



```
Add a new chunk by clicking the *Insert Chunk* button on the toolbar or by pressing *Ctrl+Alt+I*.

When you save the notebook, an HTML file containing the code and output will be saved alongside it (click the *Preview* button or press
*Ctrl+Shift+K* to preview the HTML file).
```

```
R version 4.0.5 (2021-03-31) -- "Shake and Throw"
Copyright (C) 2021 The R Foundation for Statistical Computing
Platform: x86_64-conda-linux-gnu (64-bit)

R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.

  Natural language support but running in an English locale

R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

[Workspace loaded from ~/Test Script in RStudio/TestProj/.RData]

>
```

Staged	Status	Path
☑	A	.gitignore
☑	A	AnotherTestScript.R
☑	A	TestProj.Rproj

	Name	Size	Modified
	..		
□	.gitignore	40 B	Aug 17, 2021, 12:00 PM
□	.RData	32 B	Sep 7, 2021, 2:53 PM
□	.Rhistory	0 B	Aug 17, 2021, 12:03 PM
□	AnotherTestScript.R	10 B	Aug 17, 2021, 12:01 PM
□	TestProj.Rproj	205 B	Jan 5, 2022, 4:04 PM

If you are new to RStudio, or would like to refer to this resource for learning how to programme in R, the following link leads to an excellent online book that provides information to learn how to programme in R, from loading data to writing functions: https://rstudio-education.github.io/hopr/preface.html

To get started with importing notebooks and data on Noteable, please refer to the 'Working with Data: the ABCs of Noteable' workbook

Getting started with RStudio in Noteable

The two most important components of the R language are objects, which store data, and functions, which manipulate data. R also uses a host of operators like +, -, *, /, and <- to do basic tasks. As a data scientist, you will use R objects to store data in Noteable's online stored memory.

There is an additional resource with a variety of information about R applied to the curriculum, please refer to the **Teacher's Example Workbook** PDF for this.

R comes with several built-in data sets, which are generally used as demo data for playing with R functions.

To see the list of pre-loaded data available on RStudio in Noteable, you can type the function 'data():'.

How to load and inspect in-built datasets:

- Load a built-in R data set: data("dataset_name")
- Inspect the data set: head(dataset_name)

Below you will find reference to four RStudio files containing Higher Apps tasks that can be carried out with Noteable and use built-in data sets in RStudio.

How to access the in-built RStudio datasets with Noteable

Datasets used across RStudio examples are chickwts, InsectSprays, cars and PlantGrowth.

To get started working with the built-in datasets in RStudio, please do the following:

1. Access RStudio through Noteable
2. In the top-left scripts and files window in RStudio on Noteable, type data() to see the list of in-built datasets. Run this by clicking on 'Run' in the menubar or the Ctrl and Enter buttons at the same time.
3. Then type View(cars) to see the dataset called cars, for example, you can also do this for another one of the built-in datasets such as chickwts PlantGrowth or InsectSprays.
4. Type ?cars to see information about the dataset called cars.
5. Type attach(cars) to use the data in the dataset called cars, which will provide further information on this dataset, namely that it contains data on the Speed and Stopping Distances of Cars.

* Examples of code in R within this workbook will be indented with > at the start and colour coded in blue.

Example Activities

You will find four exercises below. Guidance including steps, prompts for analysis and answers can be found at the end of the workbook.

In the script window when you have opened RStudio, carry out the following exercises:

Activity 1

Compare the effect of the different feed types on the weights of chicks.

Dataset needed: > chickwts

Type of statistical exercise: **using comparative boxplots and other statistical measures** involving generating a boxplot and carrying out statistical measures.

Activity 2

Activity - Compare the effectiveness of insect sprays A to F

Dataset needed: InsectSprays

Type of statistical exercise: **using comparative boxplots** and carrying out statistical measures including **standard deviation**.

Activity 3

Activity - Find any correlation between car speed and stopping distance. Use any model you find to make a prediction.

Dataset needed: cars

Type of statistical exercise: **correlations, model decided by student**

Activity 4

Activity: Carry out a statistical test to determine whether or not a switch to treatment 1 or treatment 2 is warranted.

Dataset needed: PlantGrowth

Type of statistical exercise: **analyse** possible treatments for plant growth and if switching treatments will yield improved growth results.

Activity 1

Compare the effect of the different feed types on the weights of chicks

The dataset used in this example is chickwts.

Datasets used across RStudio examples are chickwts, InsectSprays, cars and PlantGrowth.

To get started working with the built-in datasets in RStudio, please do the following:

1. Access RStudio through Noteable
2. Then type View(chickwts) to see the dataset called chickwts.
3. Type ?chickwts to see information about the dataset called chickwts.
4. Type attach(chickwts) to use the data in the dataset called chickwts, which will provide further information on this dataset.

   * Examples of code in R within this workbook will be indented with > at the start and colour coded in blue.

Activity Exercise

Compare the effect of the different feed types on the weights of chicks.

To carry out the activity, the following steps should be followed in RStudio:

Step 1: input the following code on line 1 of your script window in RStudio on Noteable:

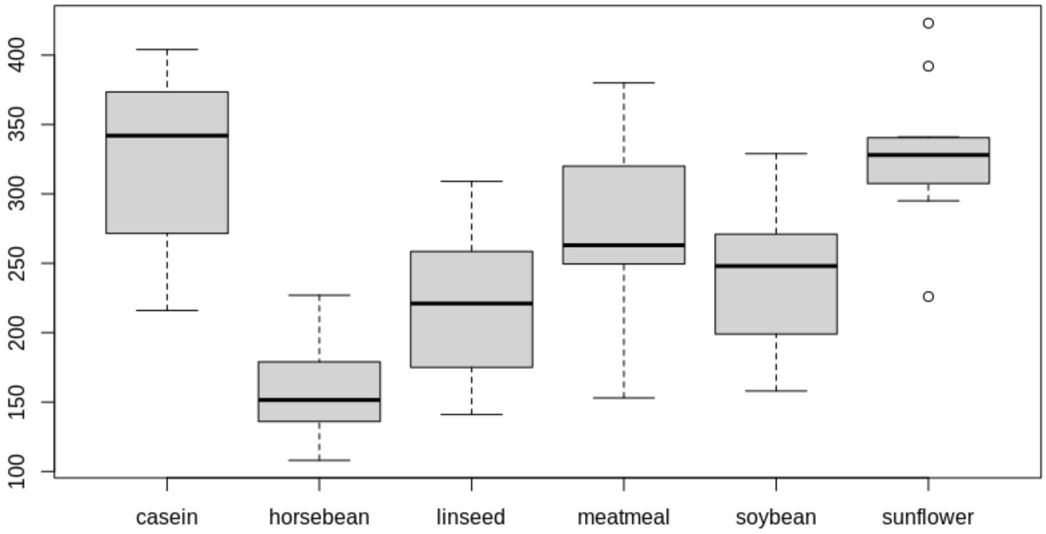
- attach(chickwts)

Step 2: this will open a new tab called ‘R data sets’, you can have a look at the variety of data sets readily available for you to work with in RStudio on Noteable.

Step 3: input the following code in your next script line:

```
> boxplot(split(weight,feed))
```

Step 4: this will generate a new plot analysing the weight of bird chicks, viewable on the bottom-right window of RStudio, under the ‘Plots’ tab. Have a look at this plot graph.



Step 5: input the following code in the next line of your script:

```
> aggregate(weight,list(feed),summary)
```

This will generate a list of the 5 figure summaries for the chick weights associated with each feed type:

```
    Group.1   x.Min. x.1st Qu. x.Median   x.Mean x.3rd Qu.   x.Max.
1    casein 216.0000  277.2500 342.0000 323.5833  370.7500 404.0000
2 horsebean 108.0000  137.0000 151.5000 160.2000  176.2500 227.0000
3   linseed 141.0000  178.0000 221.0000 218.7500  257.7500 309.0000
4  meatmeal 153.0000  249.5000 263.0000 276.9091  320.0000 380.0000
5   soybean 158.0000  206.7500 248.0000 246.4286  270.0000 329.0000
6 sunflower 226.0000  312.7500 328.0000 328.9167  340.2500 423.0000
```

Step 6: input the following code in the next line of your script:

```
> aggregate(weight,list(feed),sd)
```

This will generate a list of the standard deviations associated with each feed type:

```
    Group.1        x
1    casein 64.43384
2 horsebean 38.62584
3   linseed 52.23570
4  meatmeal 64.90062
5   soybean 54.12907
6 sunflower 48.83638
```

Activity Sample Analysis

Analyse the following description/come up with your own analysis of the data:

SAMPLE ANALYSIS BELOW

- The least effective feed seems to be horsebean with by far the lowest mean weight for the chicks that are given this type of feed.
- The most effective feed seems to be sunflower (seeds?) but there are low and high outliers.
- The protein casein also seems to be effective but has much more varied outcomes.

Activity 2

Compare the effectiveness of insect sprays A to F

The dataset used in this example is InsectSprays.

Datasets used across RStudio examples are chickwts, InsectSprays, cars and PlantGrowth.

To get started working with the built-in datasets in RStudio, please do the following:

1. Access RStudio through Noteable
2. Then type View(InsectSprays) to see the dataset called InsectSprays.
3. Type ?InsectSprays to see information about the dataset called InsectSprays.
4. Type attach(InsectSprays) to use the data in the dataset called InsectSprays, which will provide further information on this dataset.

* Examples of code in R within this workbook will be indented with > at the start and colour coded in blue.

Activity Exercise

Analyse the following description/come up with your own analysis of the data:

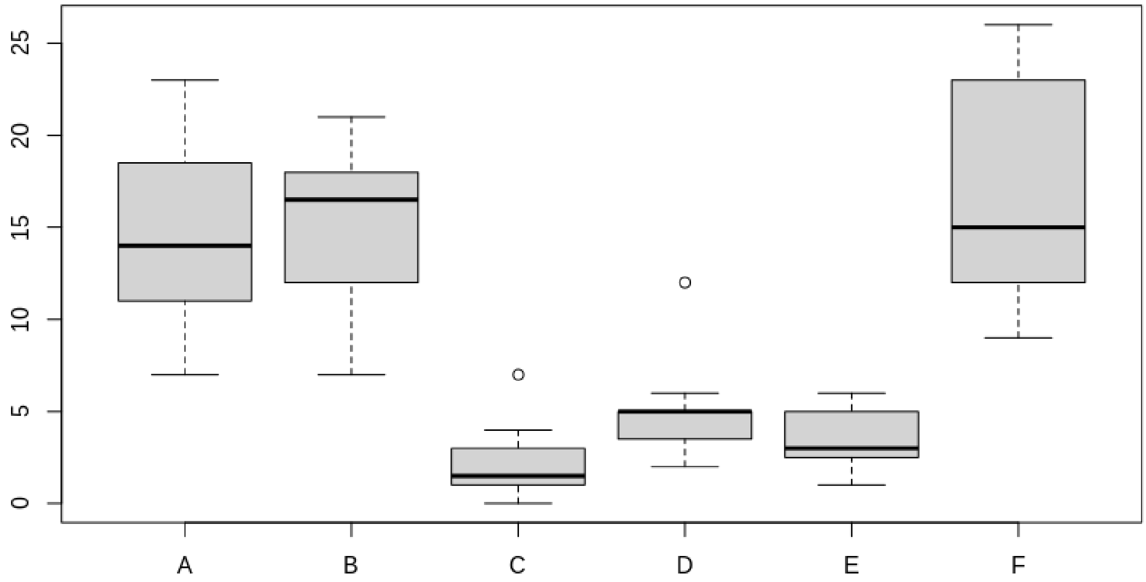
To carry out the activity, the following steps should be followed in RStudio:

Step 1: input the following code on line 1 of your script window in RStudio on Noteable:

```
➢ attach(InsectSprays)
  > boxplot(split(count,spray))
  > aggregate(count, list(spray),summary)
```

Step 2: this will open a list and visualise a boxplot of the data:

```
  Group.1    x.Min. x.1st Qu.  x.Median    x.Mean x.3rd Qu.    x.Max.
1       A  7.000000 11.500000 14.000000 14.500000 17.750000 23.000000
2       B  7.000000 12.500000 16.500000 15.333333 17.500000 21.000000
3       C  0.000000  1.000000  1.500000  2.083333  3.000000  7.000000
4       D  2.000000  3.750000  5.000000  4.916667  5.000000 12.000000
5       E  1.000000  2.750000  3.000000  3.500000  5.000000  6.000000
6       F  9.000000 12.500000 15.000000 16.666667 22.500000 26.000000
```



Step 3: input the following code to calculate the standard deviation of the insect sprays being explored, and analyse what the results show:

-
  ```
  aggregate(count, list(spray),sd)
    Group.1        x
  1       A 4.719399
  2       B 4.271115
  3       C 1.975225
  4       D 2.503028
  5       E 1.732051
  6       F 6.213378
  ```

Activity Sample Analysis

Analyse the following description/come up with your own analysis of the data:

SAMPLE ANALYSIS BELOW

- Sprays C, D and E seem to be significantly more effective than A, B and F at reducing insect count as they have significantly lower mean numbers of insects counted.
- Spray C would seem to be most effective with the lowest insect count.
- The most varied results were from spray F, and the most consistent from spray E based on standard deviations.

Activity 3 – Correlation and regression

Find any correlation between car speed and stopping distance. Use any model you find to make a prediction.

The dataset used in this example is cars.

Datasets used across RStudio examples are chickwts, InsectSprays, cars and PlantGrowth.

To get started working with the built-in datasets in RStudio, please do the following:

1. Access RStudio through Noteable
2. Then type View(cars) to see the dataset called cars.
3. Type ?cars to see information about the dataset called cars.
4. Type attach(cars) to use the data in the dataset called cars, which will provide further information on this dataset, namely that it contains data on the Speed and Stopping Distances of Cars.

   * Examples of code in R within this workbook will be indented with > at the start and colour coded in blue.

Activity Exercise

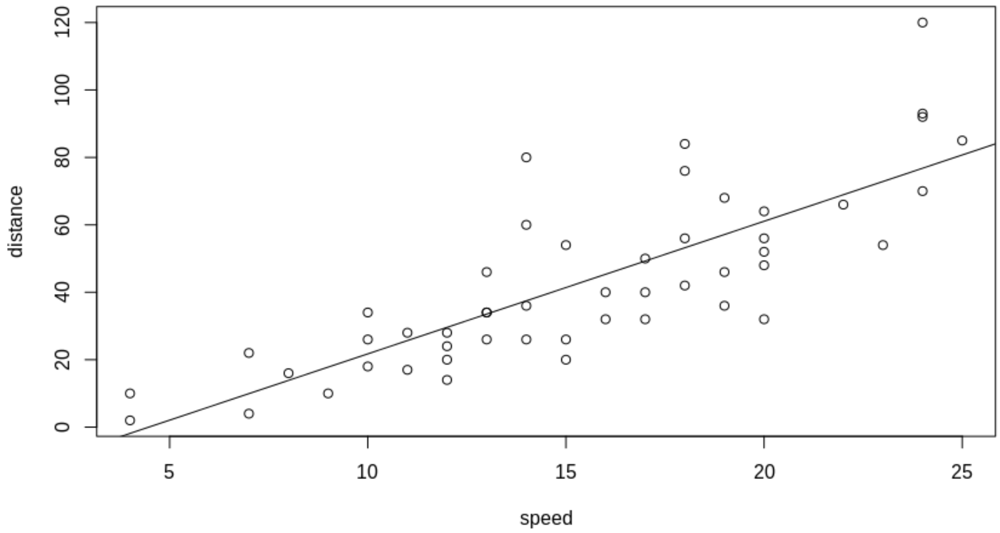
Find any correlation between car speed and stopping distance. Use any model you find to make a prediction.

To carry out the activity, the following steps should be followed in RStudio:

Step 1: input the following code on line 1 of your script window in RStudio on Noteable:

```
attach(cars)
> plot(speed,dist, xlab = "speed",ylab = "distance")
> cor.test(speed, dist)
```

Step 2: this will open a new tab in the Help tab of the bottom right window, providing a description of the data being called and a plot of the data will be shown under the Plots tab.



Step 3: analyse the data inputting the following code:

```
> abline(lm(dist~speed))

> lm(dist~speed)
> predict(lm(dist~speed), newdata =
data.frame(speed=50),interval="pred")
```

Activity Sample Analysis

Find any correlation between car speed and stopping distance. Use any model you find to make a prediction.

SAMPLE ANALYSIS BELOW

```
The following is an example of Pearson's product-moment correlation

data:
```

-
  ```
  speed and dist
  t = 9.464, df = 48, p-value = 1.49e-12
  alternative hypothesis: true correlation is not equal to 0
  95 percent confidence interval:
   0.6816422 0.8862036
  sample estimates:
        cor
  0.8068949
  ```

p value <0.05 so there is a correlation

- `lm(dist~speed)`

```
Call:
lm(formula = dist ~ speed)
```

```
Coefficients:
(Intercept)        speed
    -17.579        3.932
```

So stopping distance (ft) = -17.579 + 3.932 x speed (mph)

-
  ```
  predict(lm(dist~speed), newdata =
  data.frame(speed=50),interval="pred")
         fit      lwr      upr
  1 179.0413 136.4865 221.5962
  ```

At 50 mph the predicted stopping distance will be 179 feet within a 95% confidence interval ranging from 136.5 to 221.6 feet.

Activity 4 – Testing Hypotheses

Testing a Hypothesis

The dataset used in this example is PlantGrowth.

Datasets used across RStudio examples are chickwts, InsectSprays, cars and PlantGrowth.

To get started working with the built-in datasets in RStudio, please do the following:

The dataset used in this example is PlantGrowth.

1. Access RStudio through Noteable

   * Examples of code in R within this workbook will be indented with > at the start and colour coded in blue.

2. Then type View(PlantGrowth) to see the dataset called PlantGrowth.
3. Type ?PlantGrowth to see information about the dataset called PlantGrowth.
4. Type attach(PlantGrowth) to use the data in the dataset called PlantGrowth, which will provide further information on this dataset.

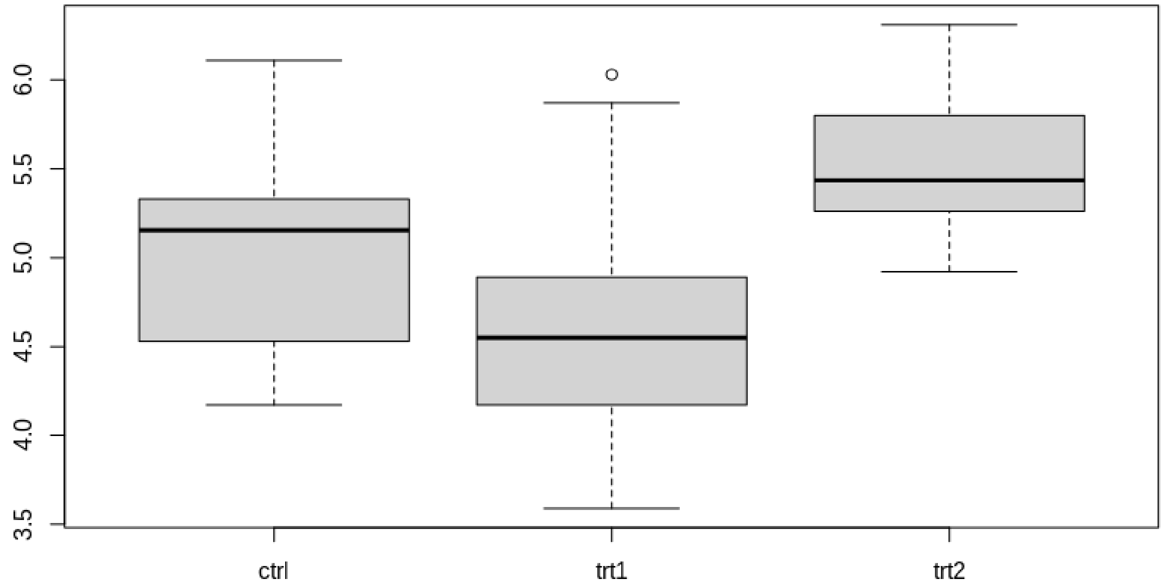
* Examples of code in R within this workbook will be indented with > at the start and colour coded in blue.

<u>Activity Exercise</u>

<u>Carry out a statistical test to determine whether or not a switch to treatment 1 or treatment 2 is warranted.</u>

Step 1: input the following code on line 1 of your script window in RStudio on Noteable:

```
attach(PlantGrowth)
> boxplot(split(weight,group))
```



Step 2: input the following code and analyse the results

```
      > t.test(weight [group=="ctrl"], weight [group=="trt1"])

Welch Two Sample t-test
data:  weight[group == "ctrl"] and weight[group == "trt1"]
t = 1.1913, df = 16.524, p-value = 0.2504
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.2875162  1.0295162
sample estimates:
mean of x mean of y
    5.032     4.661

      > t.test(weight [group=="ctrl"], weight [group=="trt2"])

Welch Two Sample t-test
data:  weight[group == "ctrl"] and weight[group == "trt2"]
t = -2.134, df = 16.786, p-value = 0.0479
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.98287213 -0.00512787
sample estimates:
mean of x mean of y
    5.032     5.526
```

Activity Sample Analysis

Analyse the following description/come up with your own analysis of the data:

SAMPLE ANALYSIS BELOW

- Treatment 1 seems to inhibit plant growth compared to the control group. Would not recommend this choice.
- Treatment 2 gives greater plant growth than the control group (greater mean weight) and as the t test p value <0.05 we can reject the null hypotheses and conclude there is a true difference in the mean weights.
- We can therefore recommend using Treatment 2 to increase plant growth.