INFS4203/7203 Data Mining
The University of Queensland, Australia
Semester 2, 2018

Tutorial Week 6: Introduction to R for Data Mining (Part II)

Chandra Prasetyo Utomo

c.utomo@uq.edu.au

Objectives

1. To tour some methods in data selection for data mining purposes.
2. To be able to generate and save a plot from data.

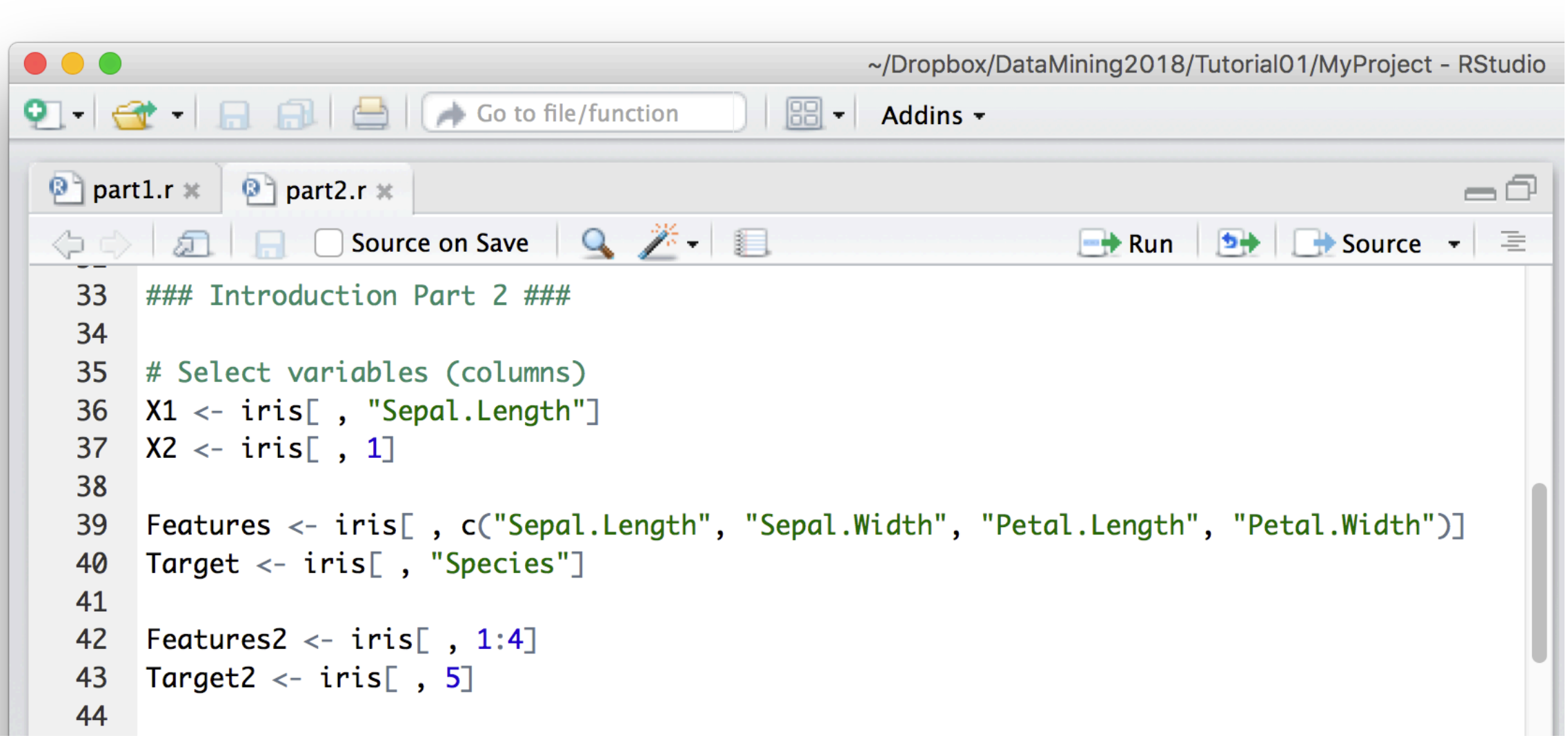
Outline

1. Data Selection (*25 minutes*)
2. Data Visualisation (*25 minutes*)

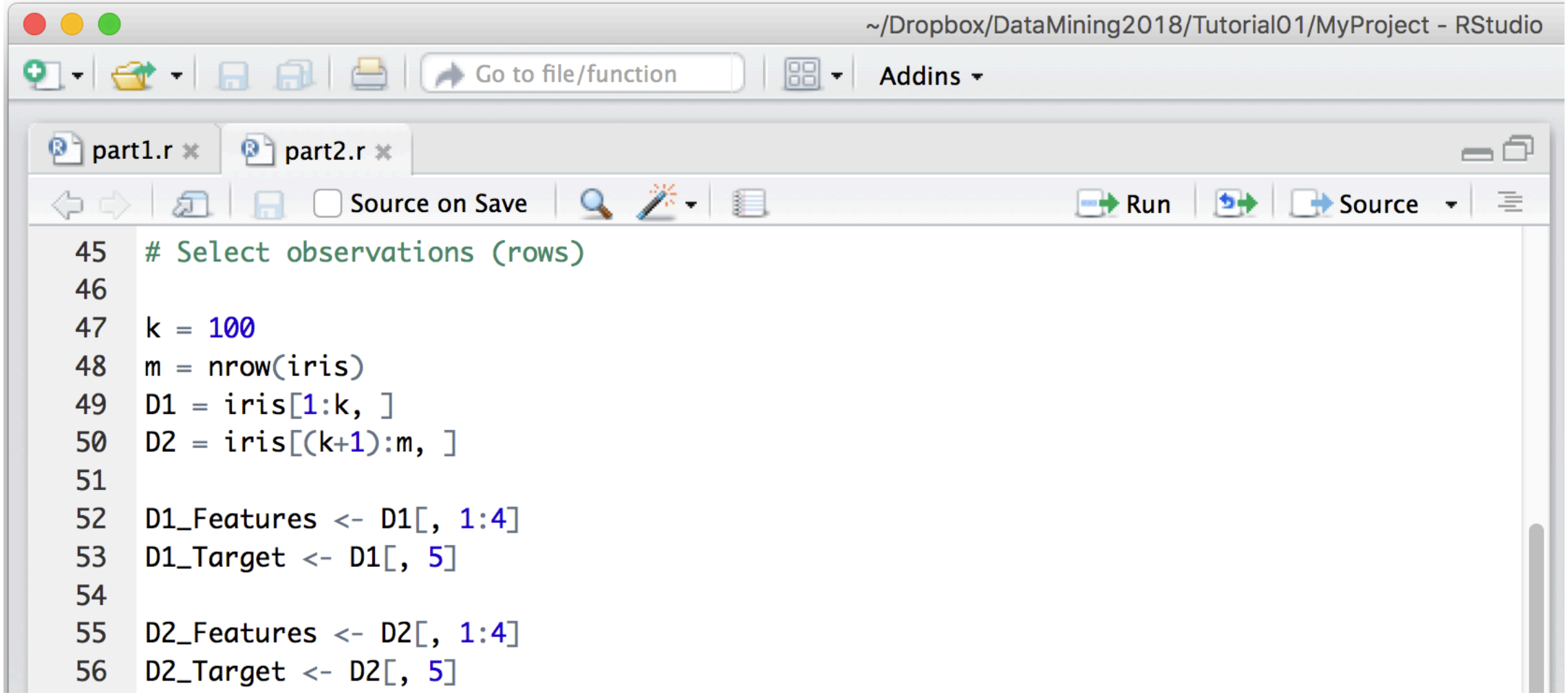
Part 1

Data Selection

Select Variables (columns)

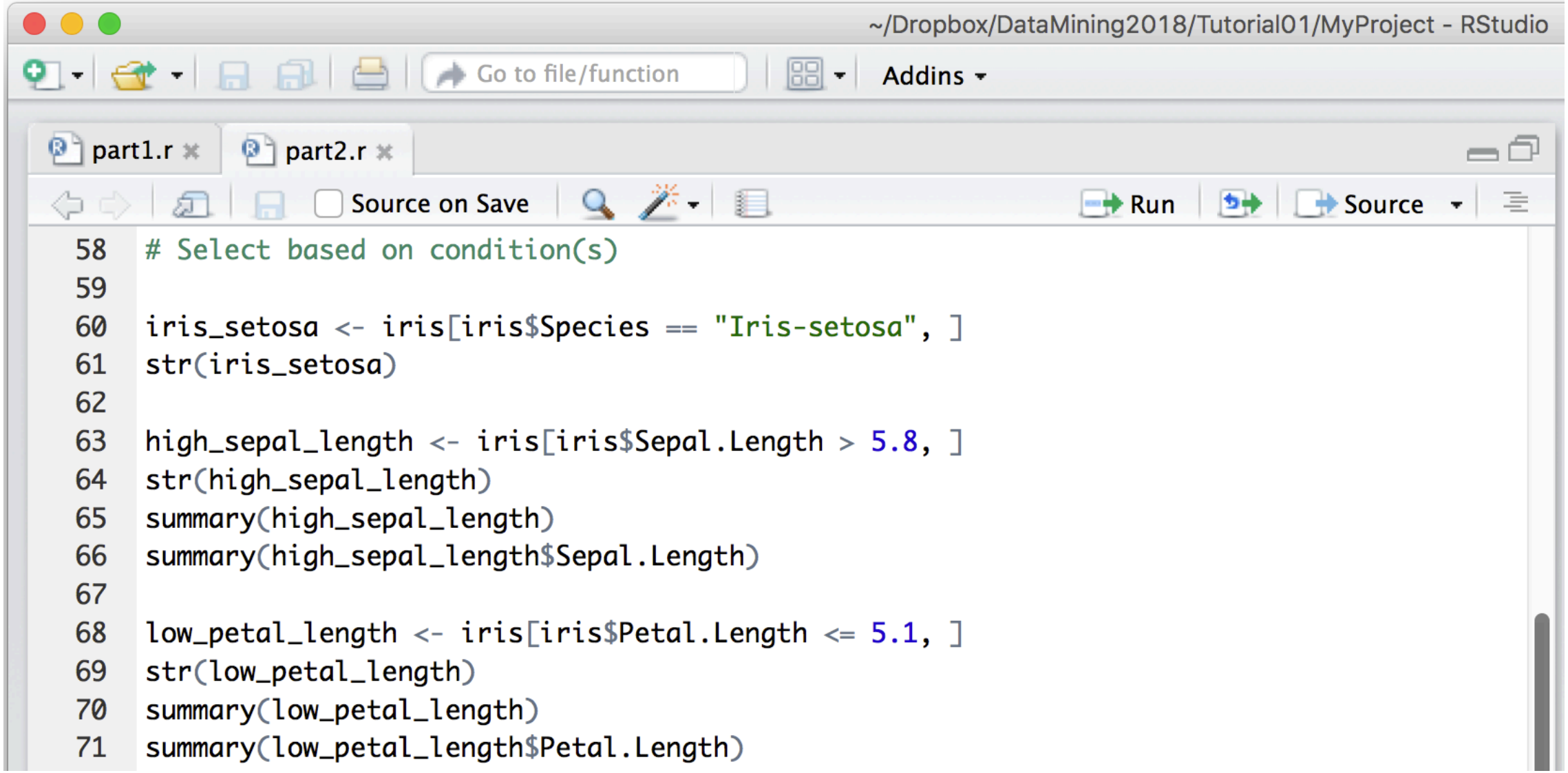


Select Observations (rows)



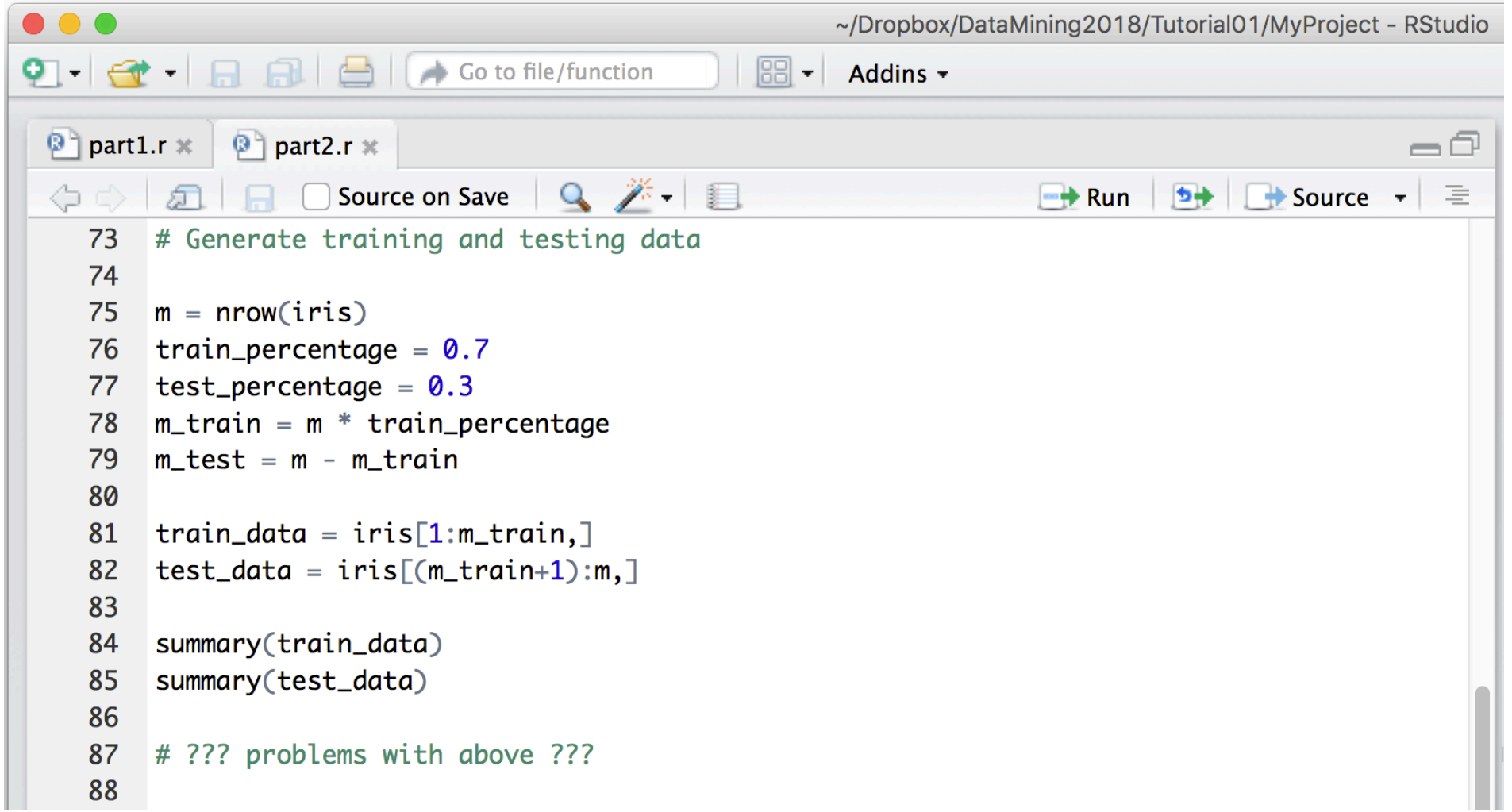

```
45  # Select observations (rows)
46
47  k = 100
48  m = nrow(iris)
49  D1 = iris[1:k, ]
50  D2 = iris[(k+1):m, ]
51
52  D1_Features <- D1[, 1:4]
53  D1_Target <- D1[, 5]
54
55  D2_Features <- D2[, 1:4]
56  D2_Target <- D2[, 5]
```

Selection based on Condition(s)

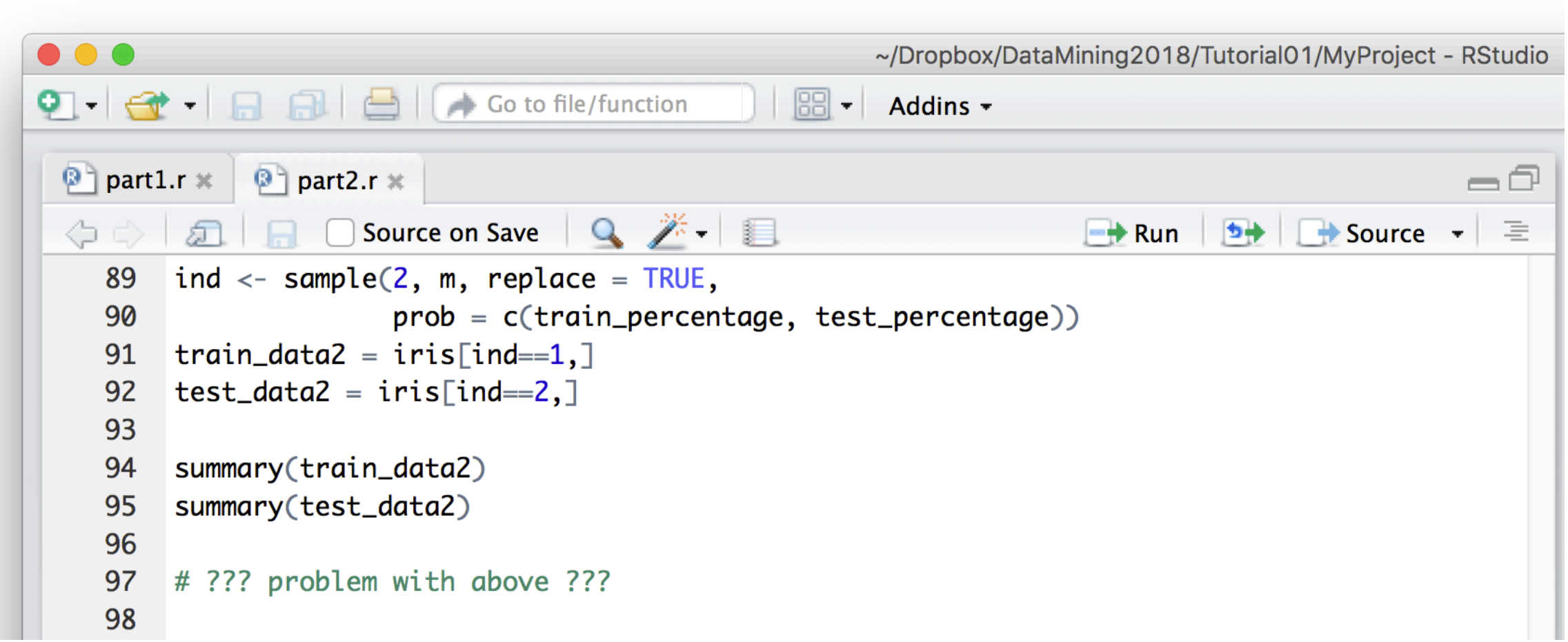



```
58  # Select based on condition(s)
59
60  iris_setosa <- iris[iris$Species == "Iris-setosa", ]
61  str(iris_setosa)
62
63  high_sepal_length <- iris[iris$Sepal.Length > 5.8, ]
64  str(high_sepal_length)
65  summary(high_sepal_length)
66  summary(high_sepal_length$Sepal.Length)
67
68  low_petal_length <- iris[iris$Petal.Length <= 5.1, ]
69  str(low_petal_length)
70  summary(low_petal_length)
71  summary(low_petal_length$Petal.Length)
```

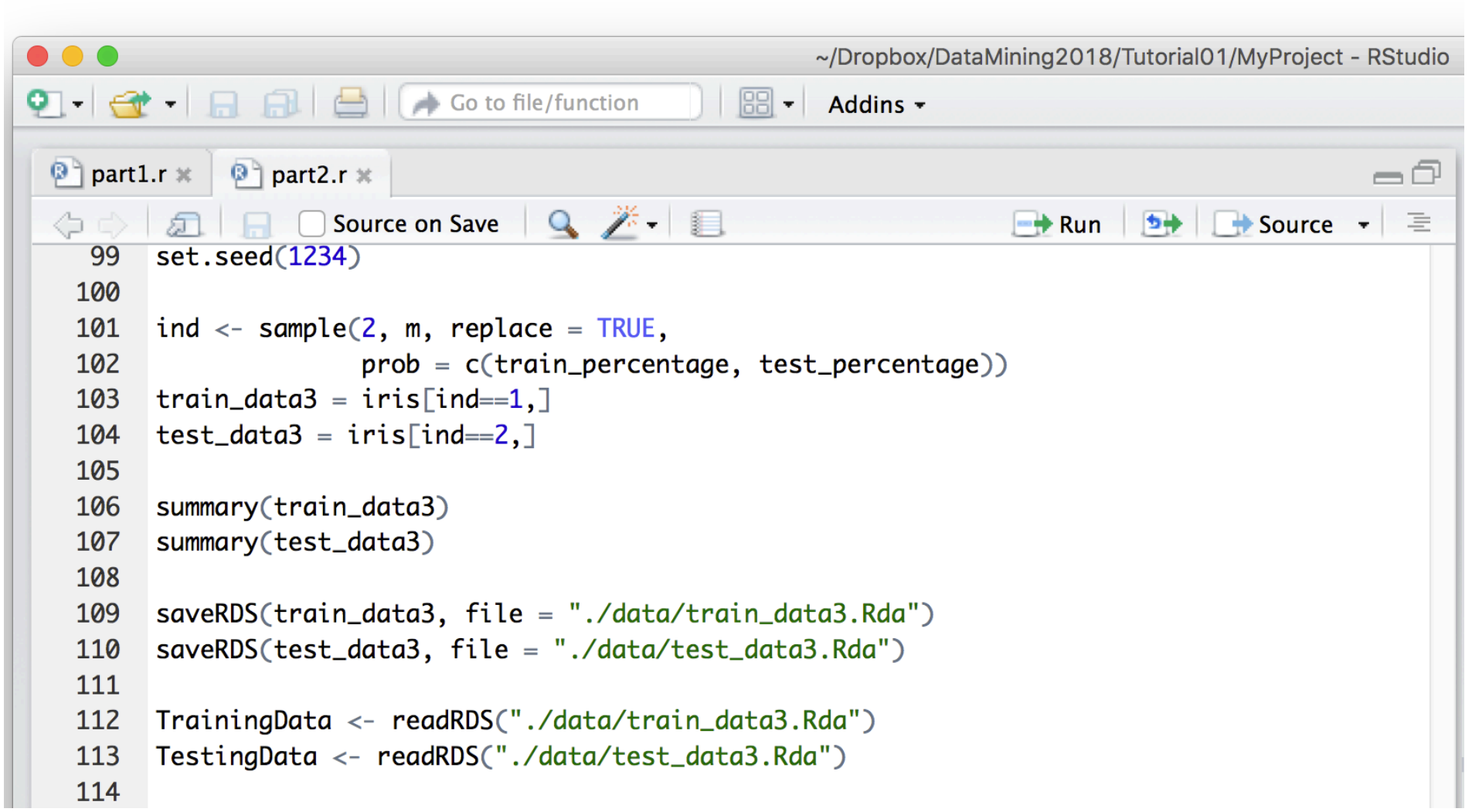
Generate Training and Testing Data 1



Generate Training and Testing Data 2



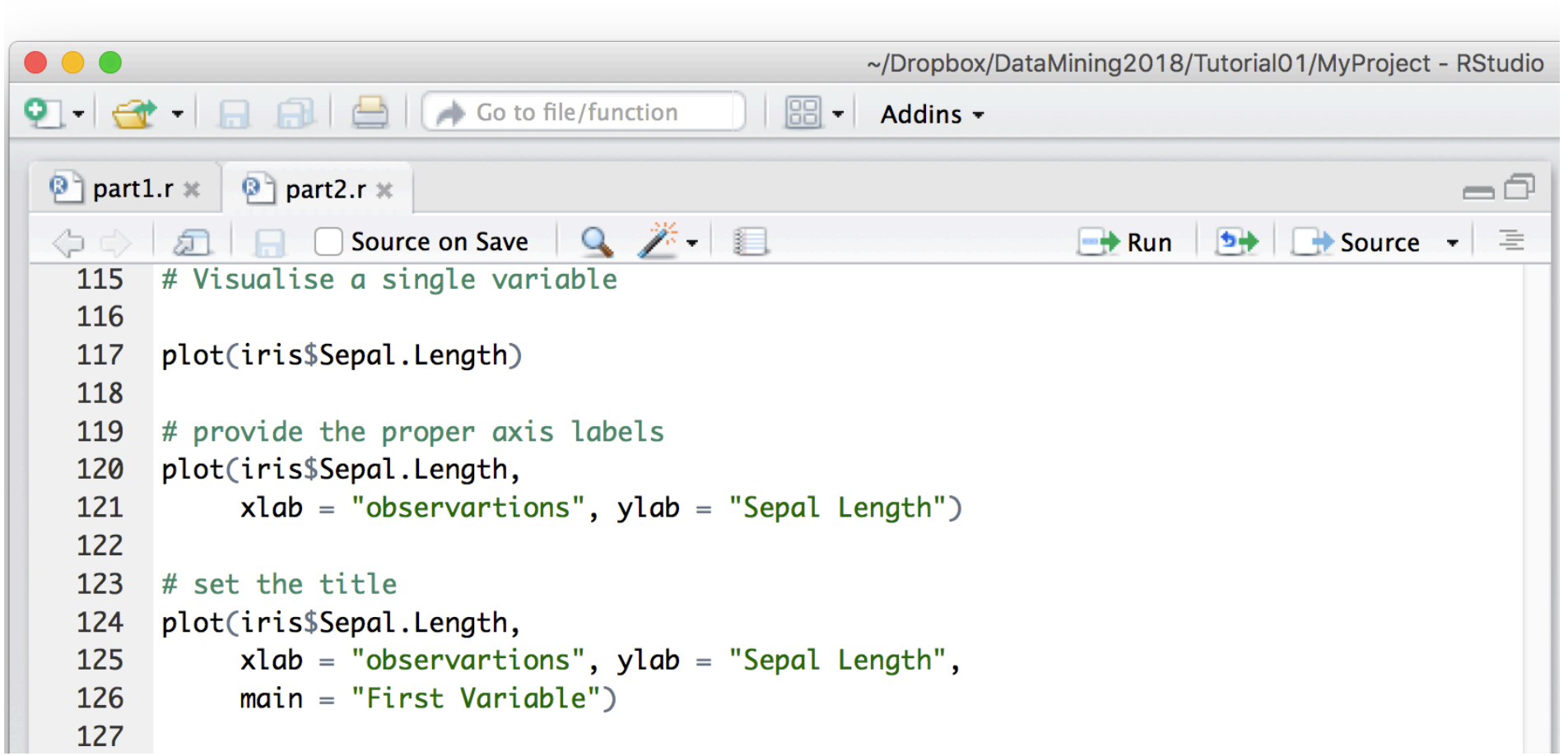
Generate Training and Testing Data 3



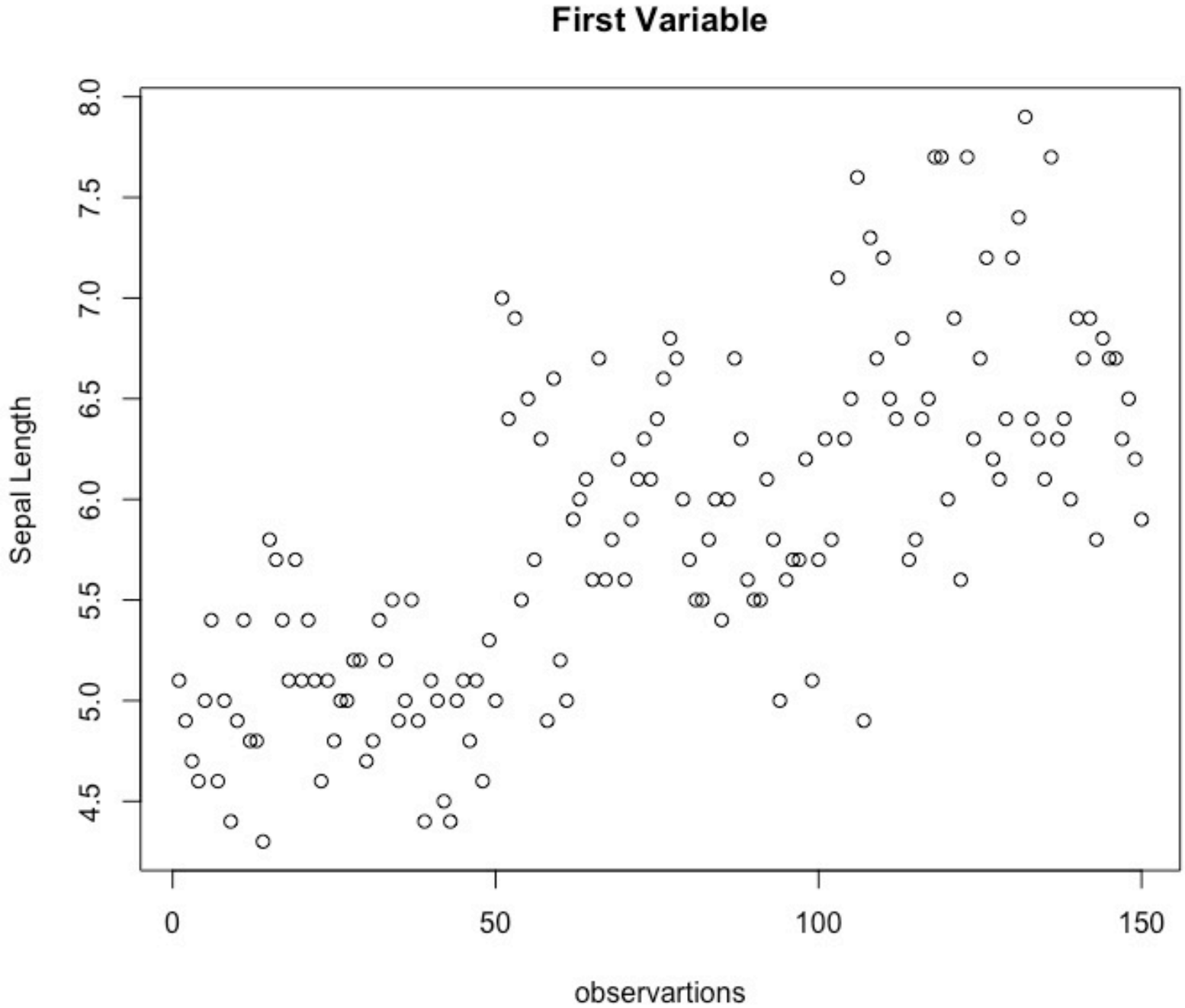
Part 2

Data Visualisation

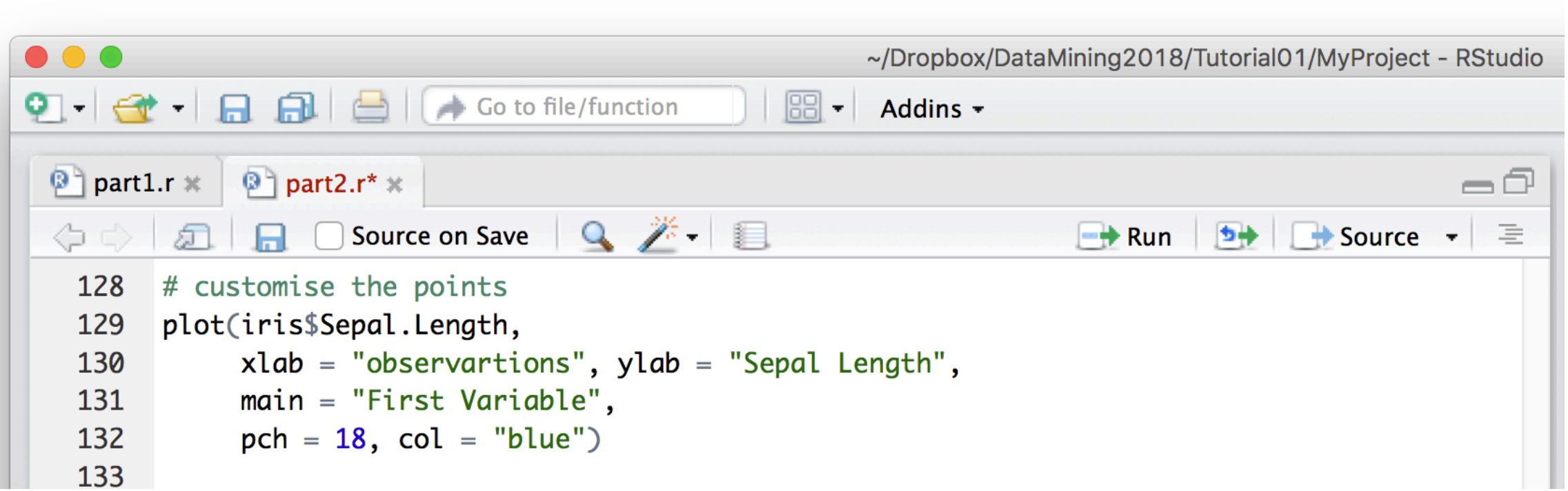
Visualise a Single Variable

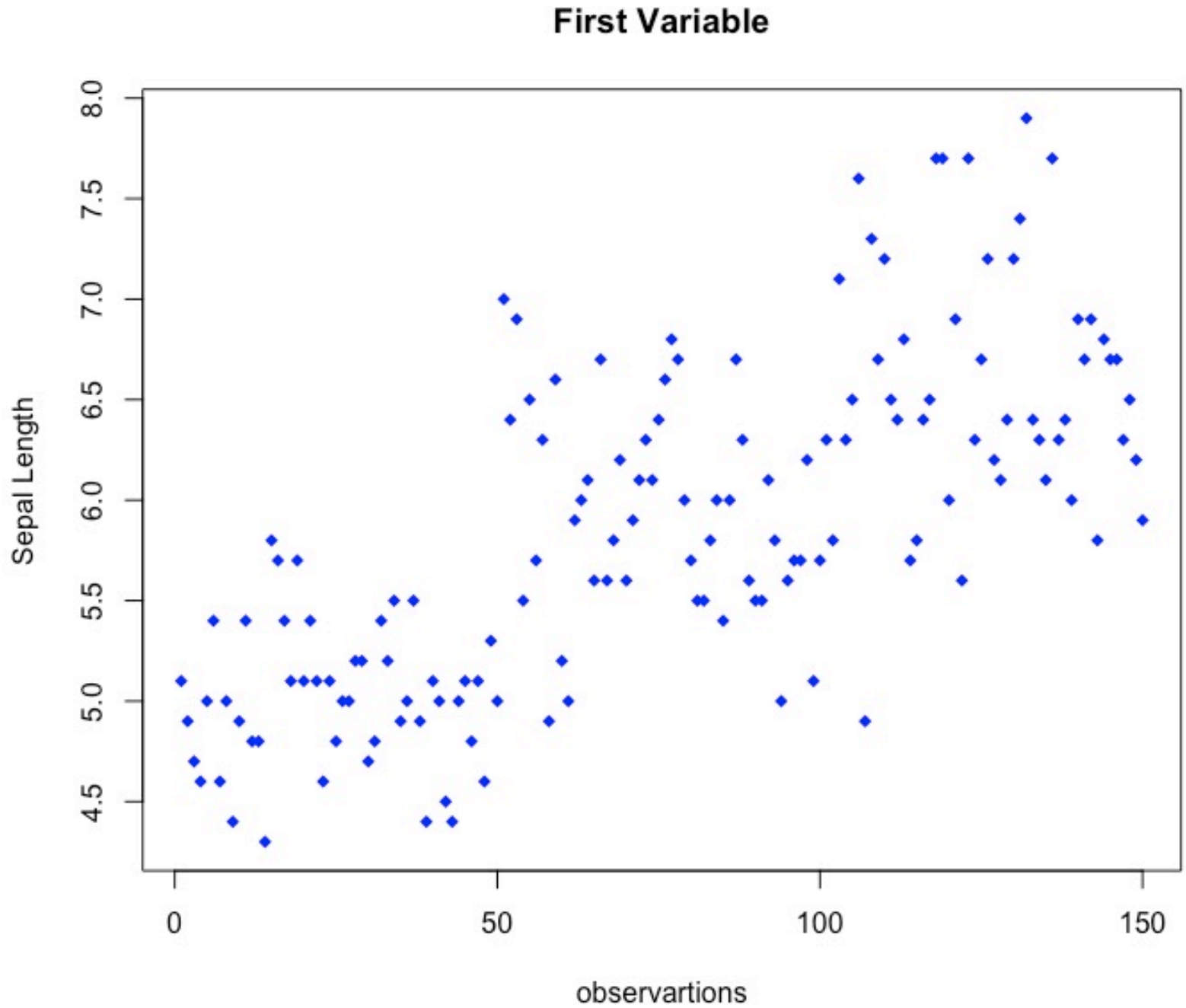



```
115  # Visualise a single variable
116
117  plot(iris$Sepal.Length)
118
119  # provide the proper axis labels
120  plot(iris$Sepal.Length,
121       xlab = "observartions", ylab = "Sepal Length")
122
123  # set the title
124  plot(iris$Sepal.Length,
125       xlab = "observartions", ylab = "Sepal Length",
126       main = "First Variable")
127
```

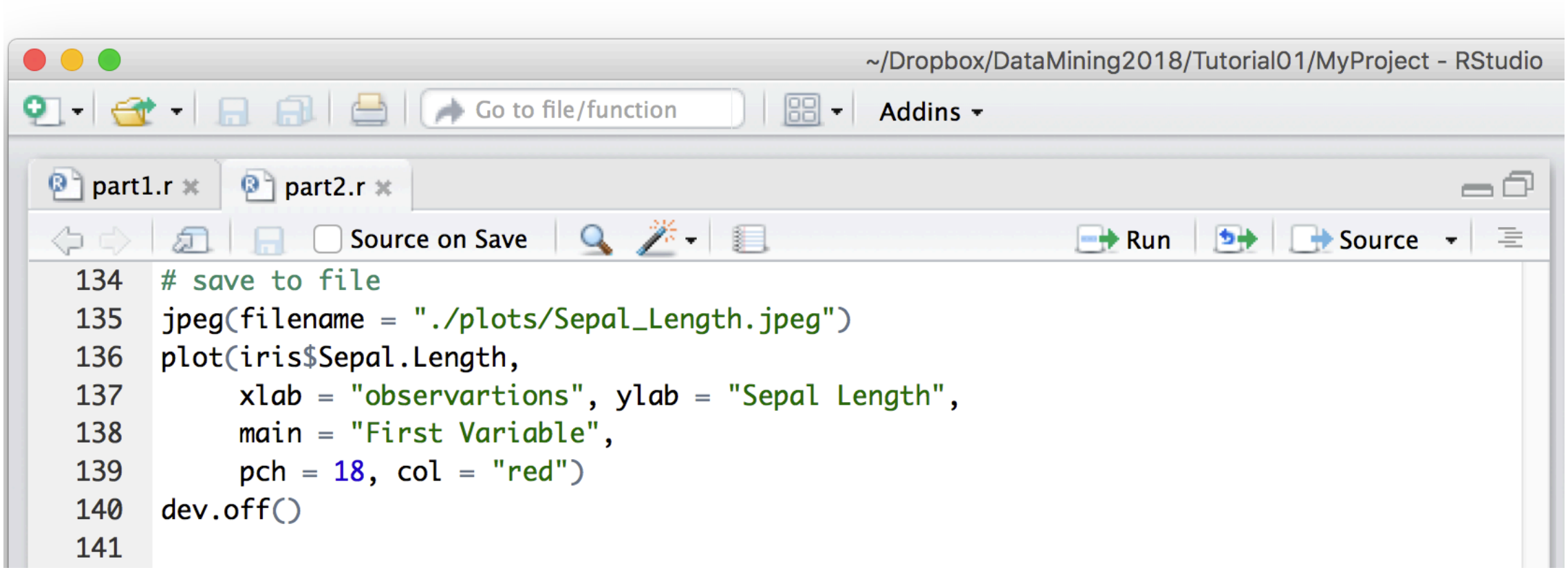

First Variable
Sepal Length
4.5
5.0
5.5
6.0
6.5
7.0
7.5
8.0
0
50
100
150
observartions

Customise the points

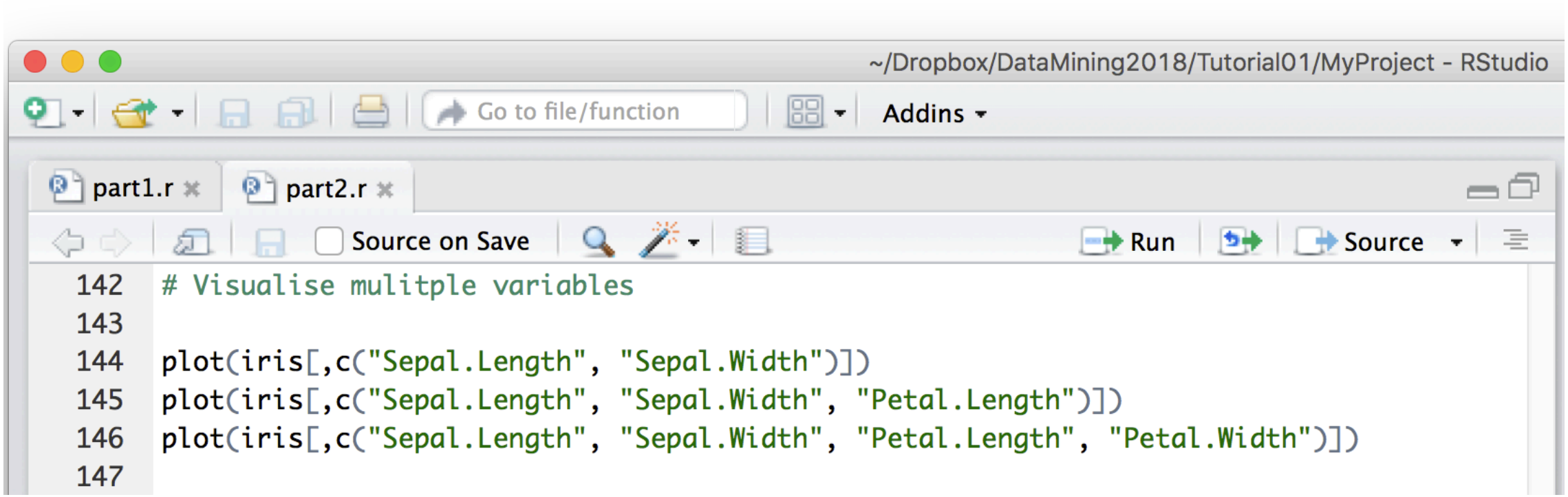


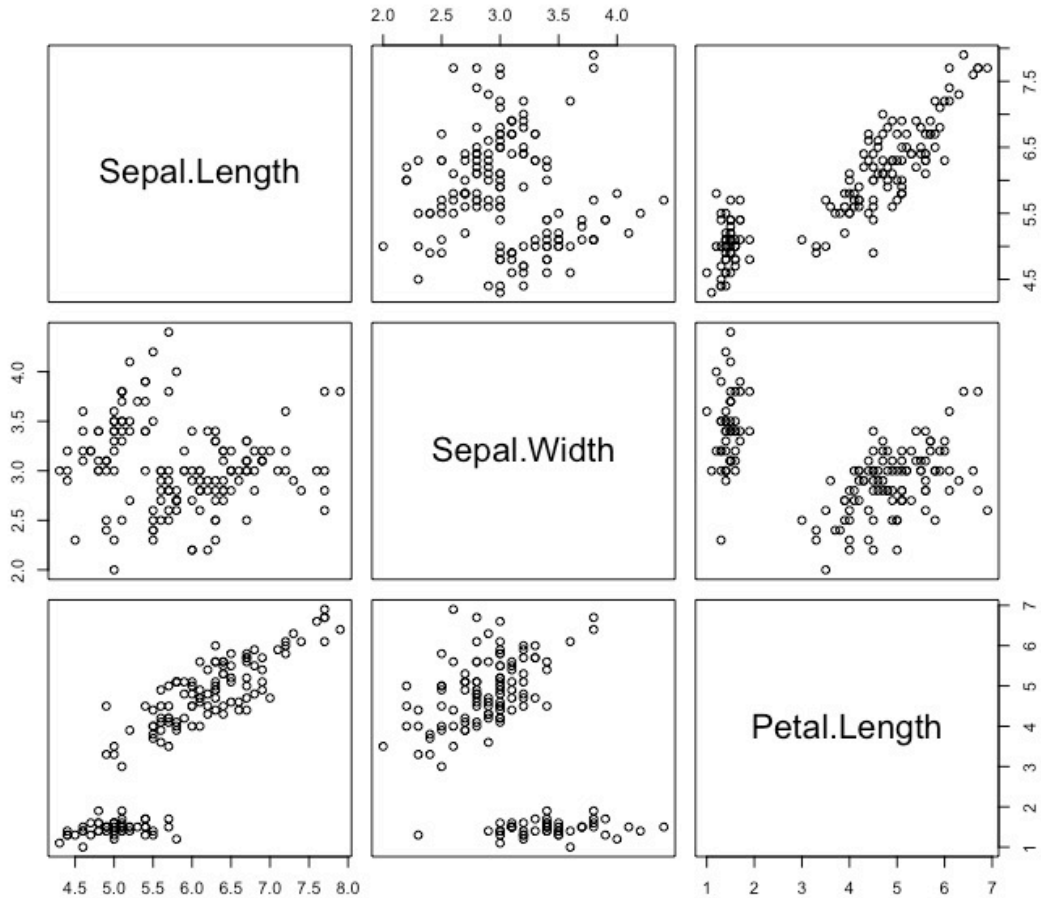

First Variable
Sepal Length
4.5
5.0
5.5
6.0
6.5
7.0
7.5
8.0
0
50
100
150
observartions

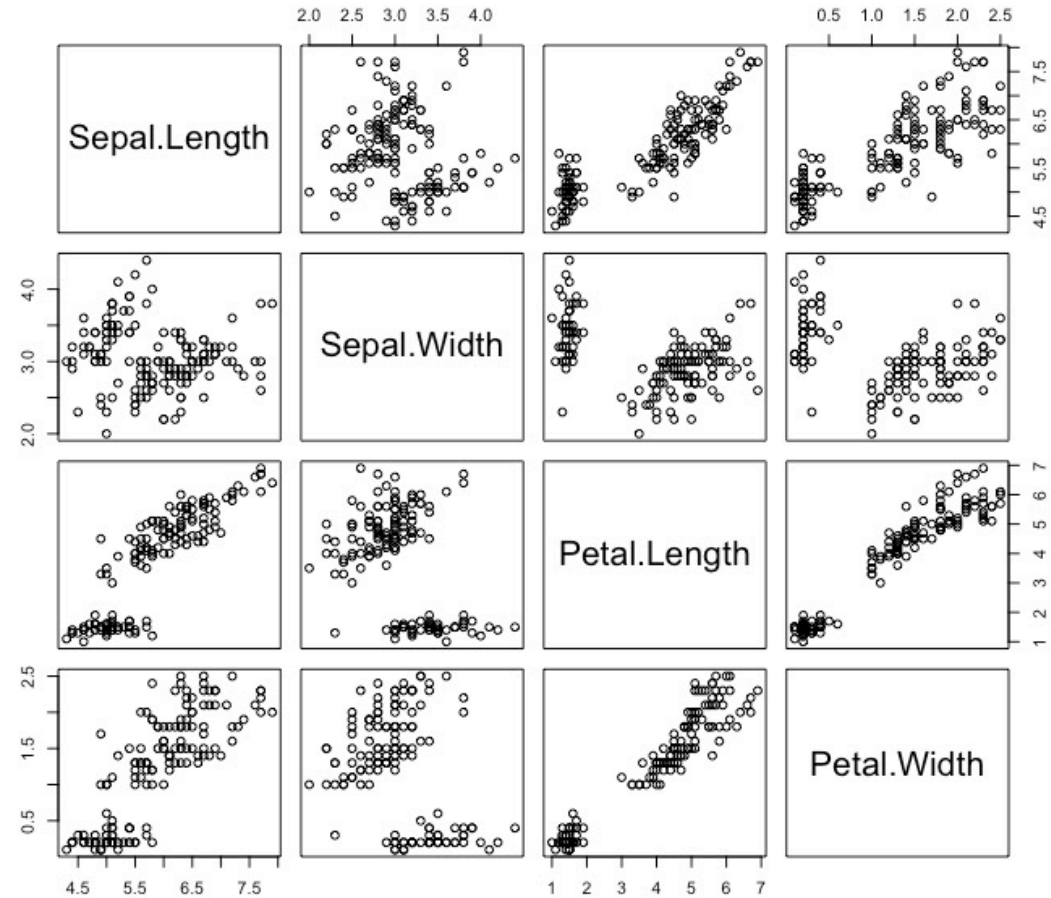
Save to File


~/Dropbox/DataMining2018/Tutorial01/MyProject - RStudio
Go to file/function
Addins
part1.r
part2.r
Source on Save
Run
Source
134 # save to file
135 jpeg(filename = "./plots/Sepal_Length.jpeg")
136 plot(iris$Sepal.Length,
137 xlab = "observartions", ylab = "Sepal Length",
138 main = "First Variable",
139 pch = 18, col = "red")
140 dev.off()
141

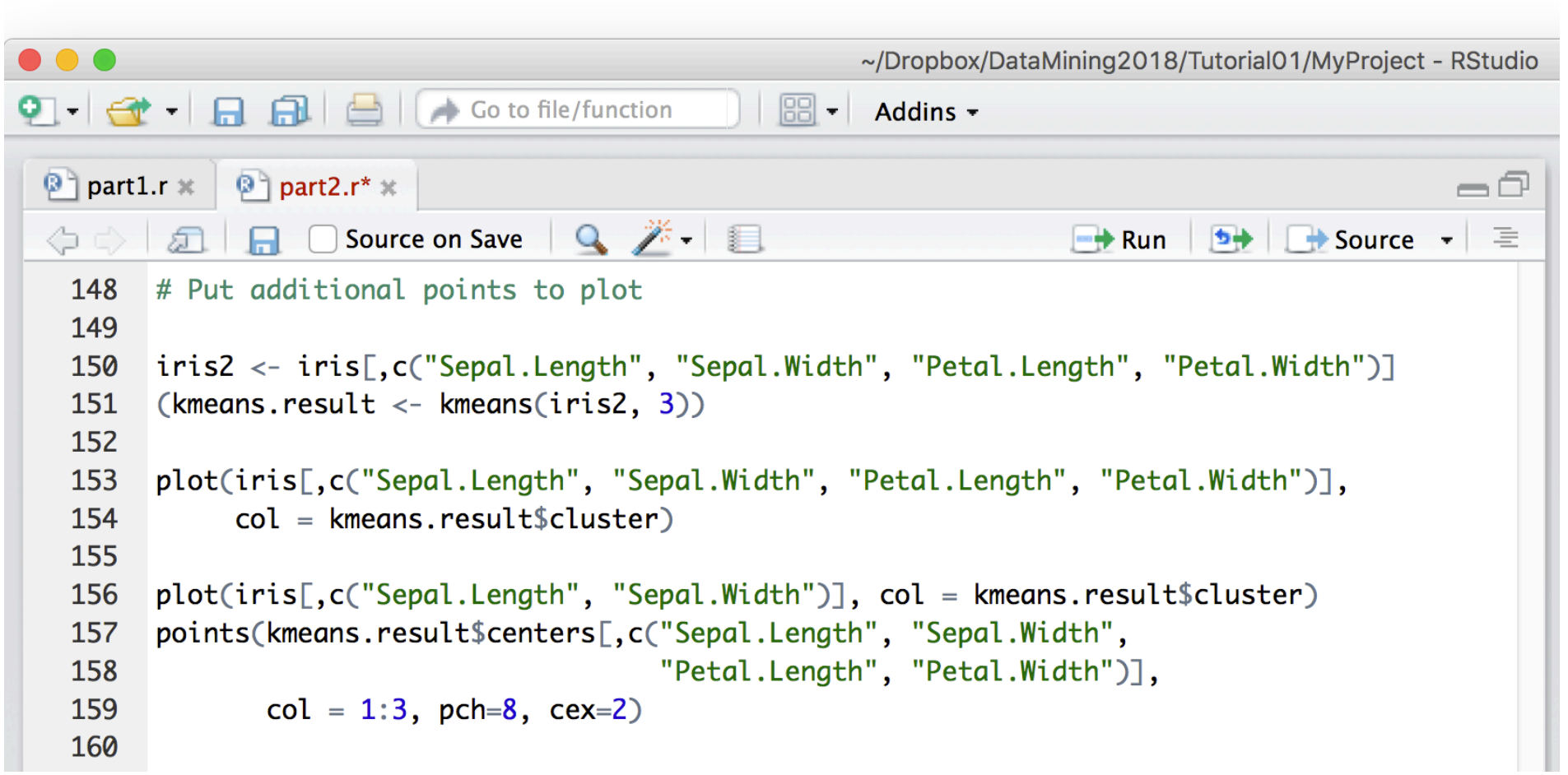
Visualise Multiple Variables



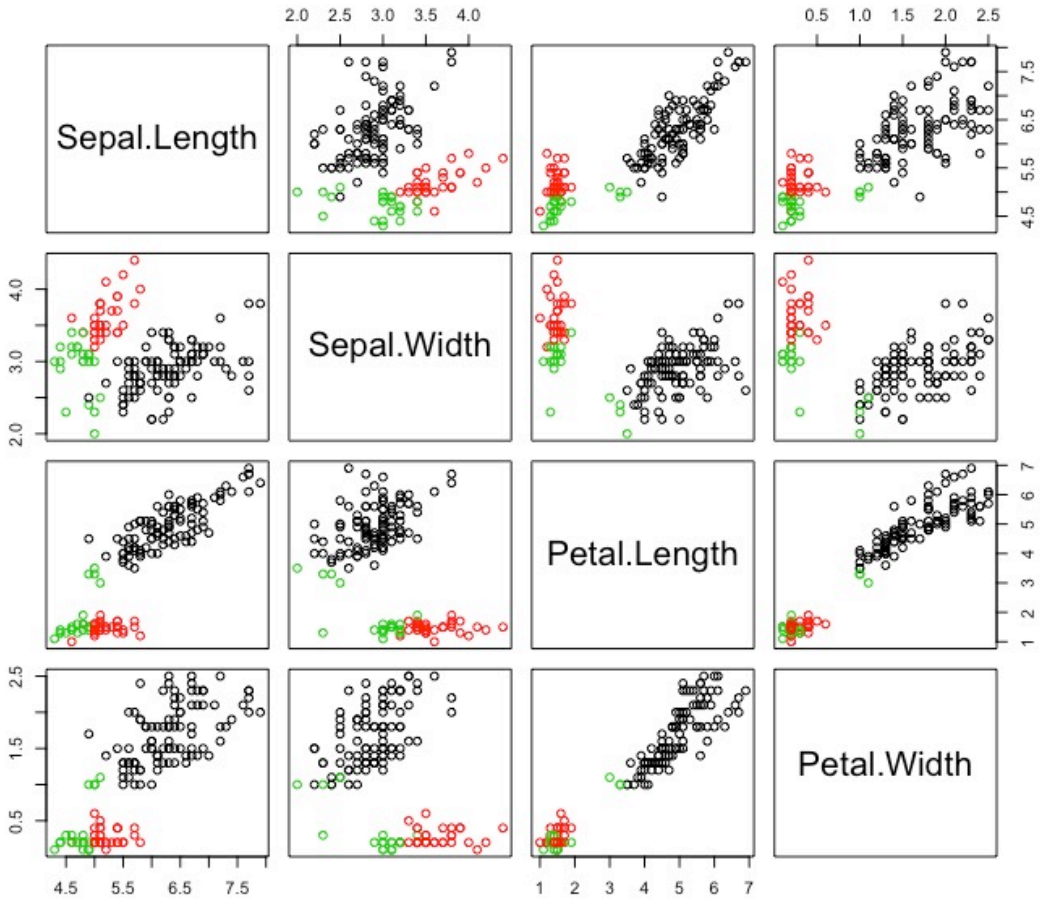

Sepal.Length
Sepal.Width
Petal.Length

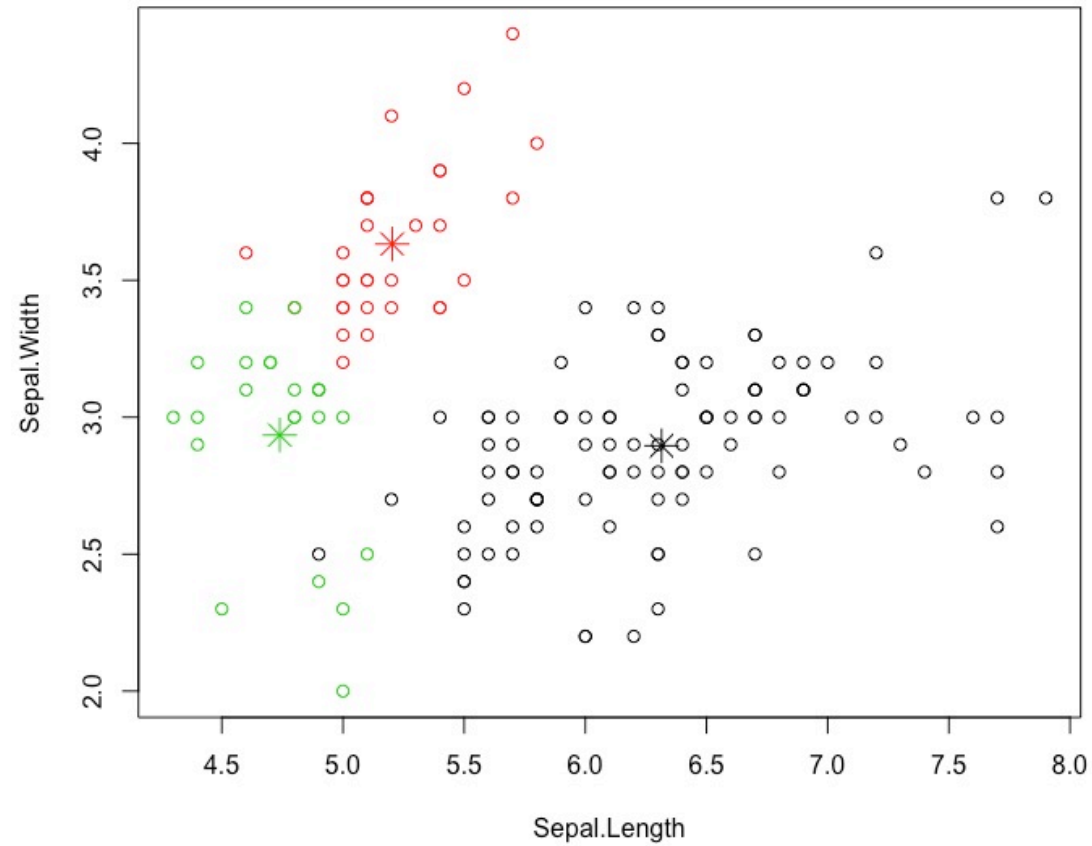

Sepal.Length
Sepal.Width
Petal.Length
Petal.Width

Add points to plot



Don’t worry about k-means clustering at the moment.
We will go deeper and discuss more detail about it in the next tutorial.


Sepal.Length
Sepal.Width
Petal.Length
Petal.Width


Sepal.Width
Sepal.Length

Reference:

- R and Data Mining. Yangchan Zhao. Academic Press 2012. https://www.sciencedirect.com/book/9780123969637/r-and-data-mining