

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

Tutorial Week 5: Introduction to R for Data Mining

Chandra Prasetyo Utomo
c.utomo@uq.edu.au

Objectives

1. To gain familiarity with basic R syntax and RStudio IDE.
2. To learn some best practices of data mining project.
3. To be able to load external data into R data frame.
4. To tour some methods of exploring a data frame.

Outline

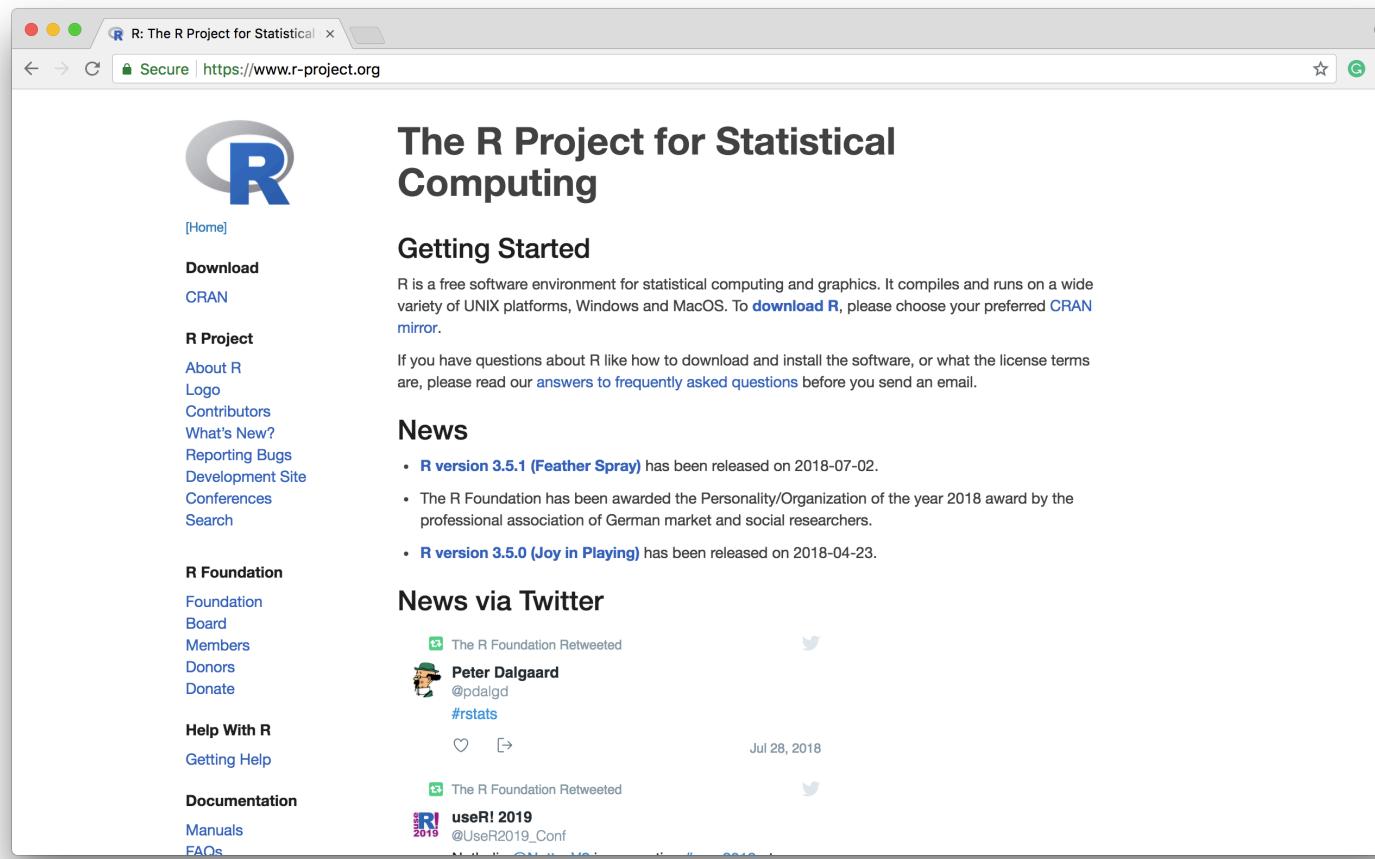
1. Installation and Introduction to RStudio (*10 minutes*)
2. Best Practice Data Mining Project (*10 minutes*)
3. Data Extraction (*10 minutes*)
4. Data Exploration (*20 minutes*)

Part 1

Installation & Introduction to RStudio

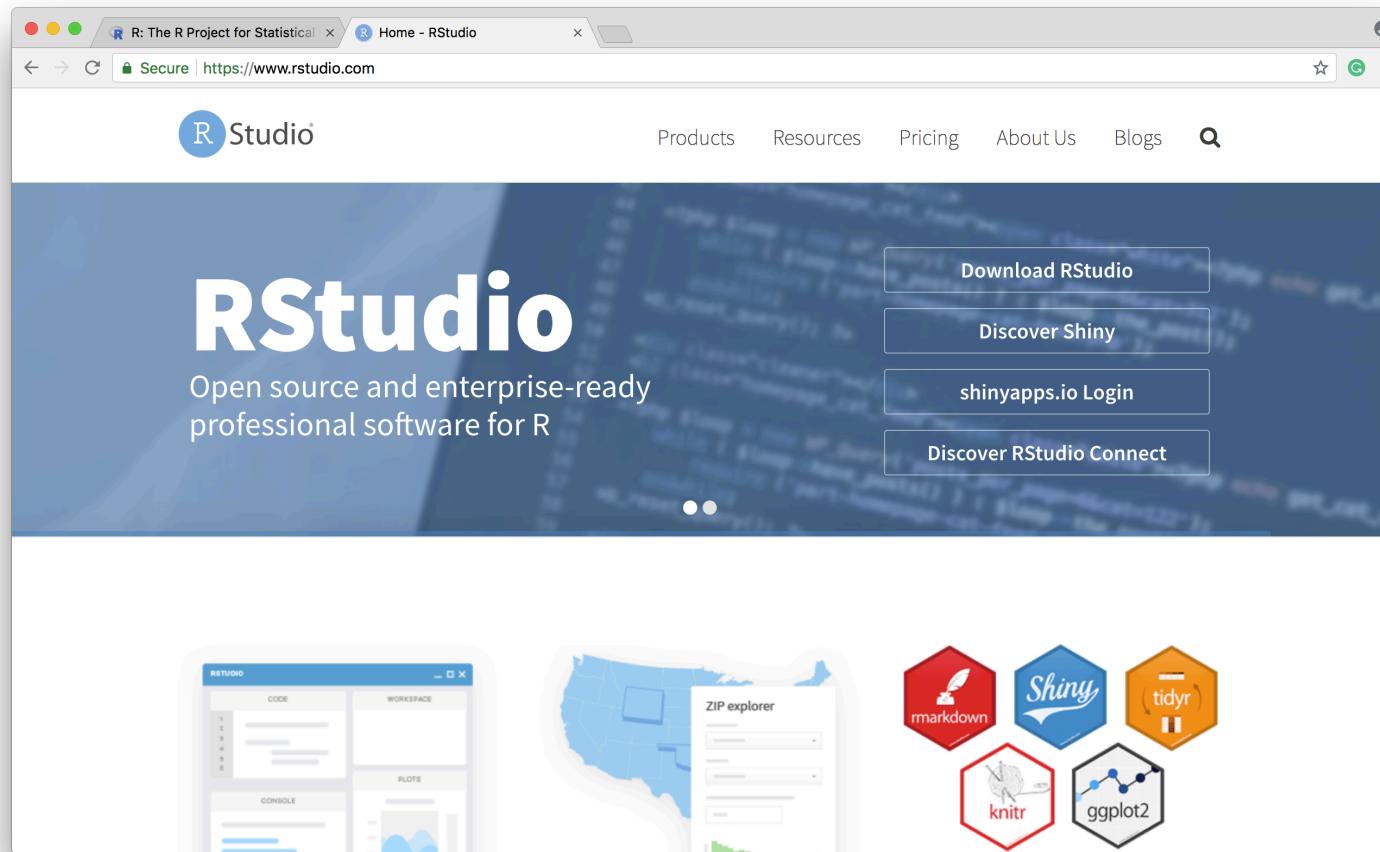
Install R

- <https://www.r-project.org/>

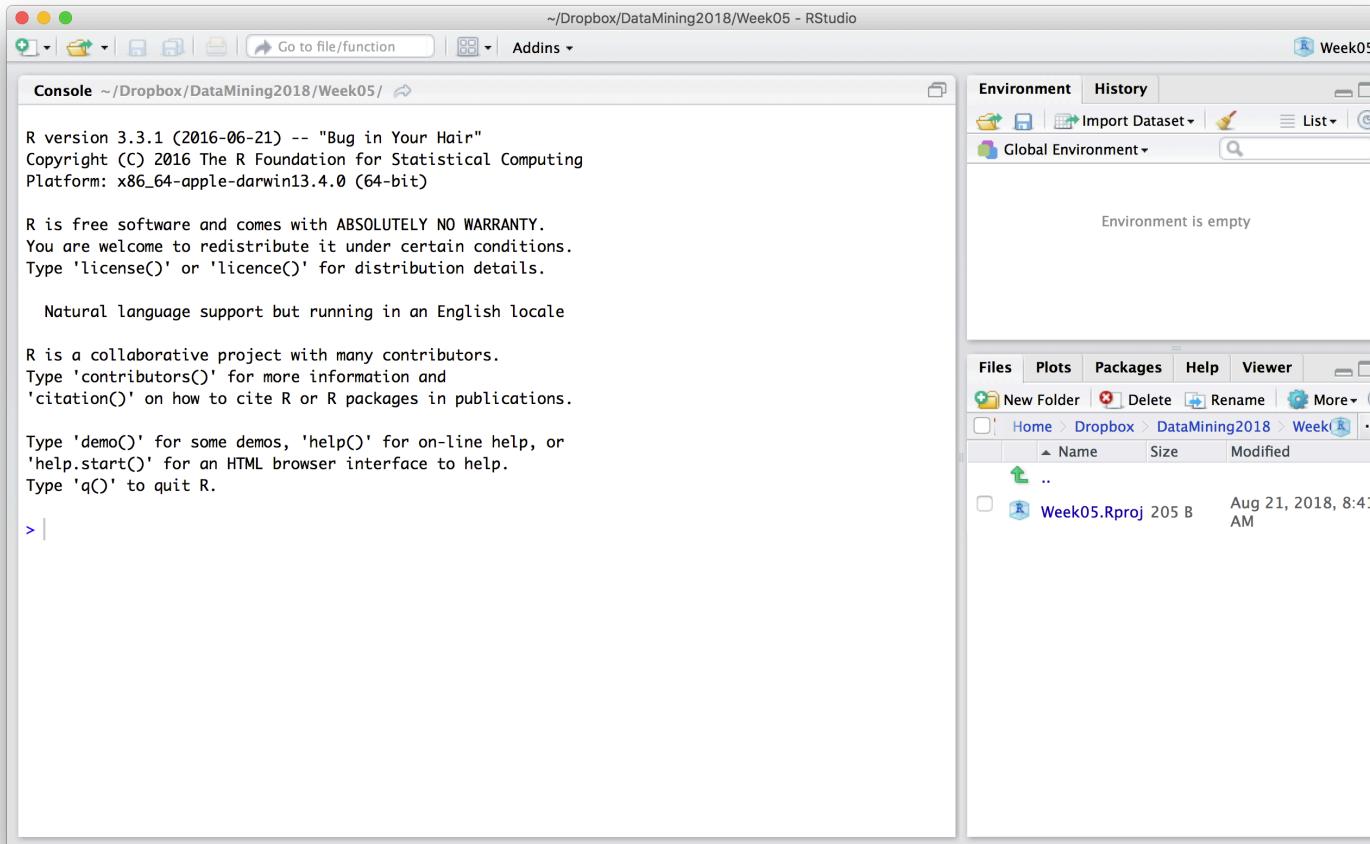


Install RStudio

- <https://www.rstudio.com/>



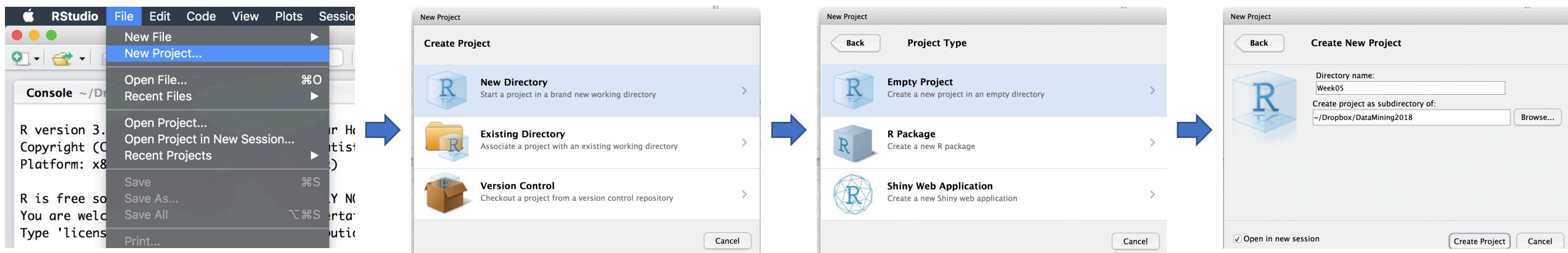
RStudio



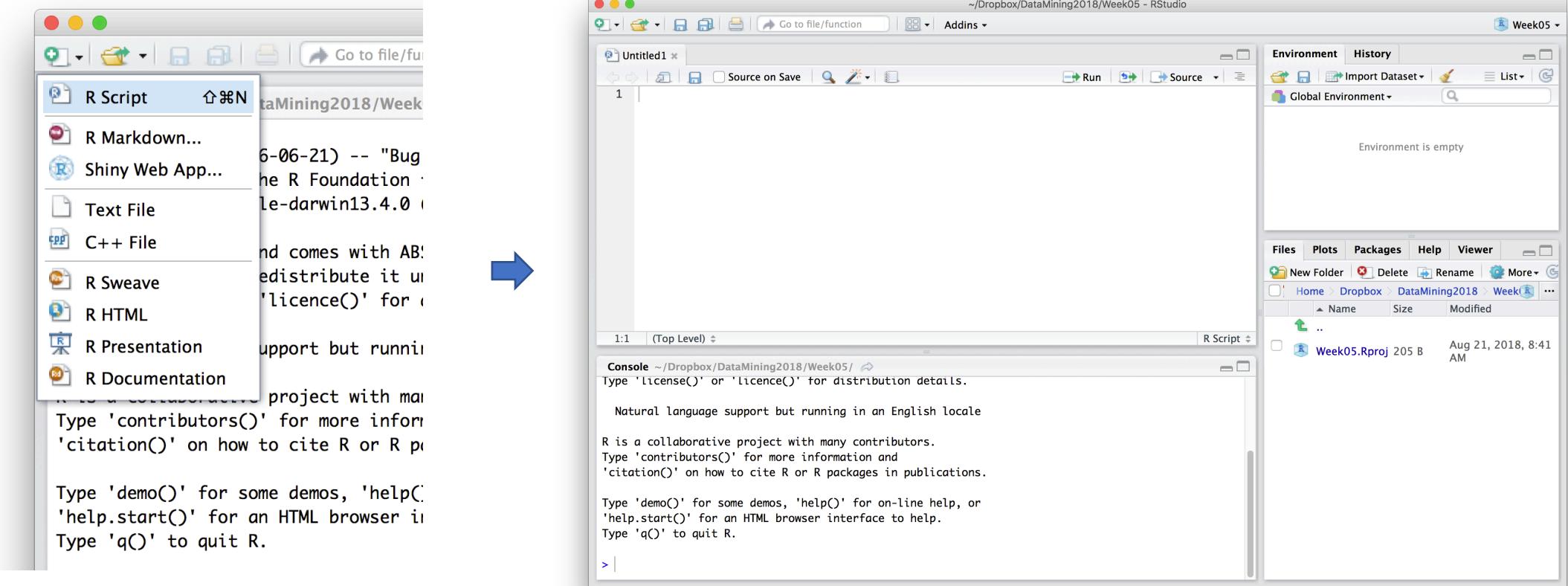
Part 2

Best Practice Data Mining Project

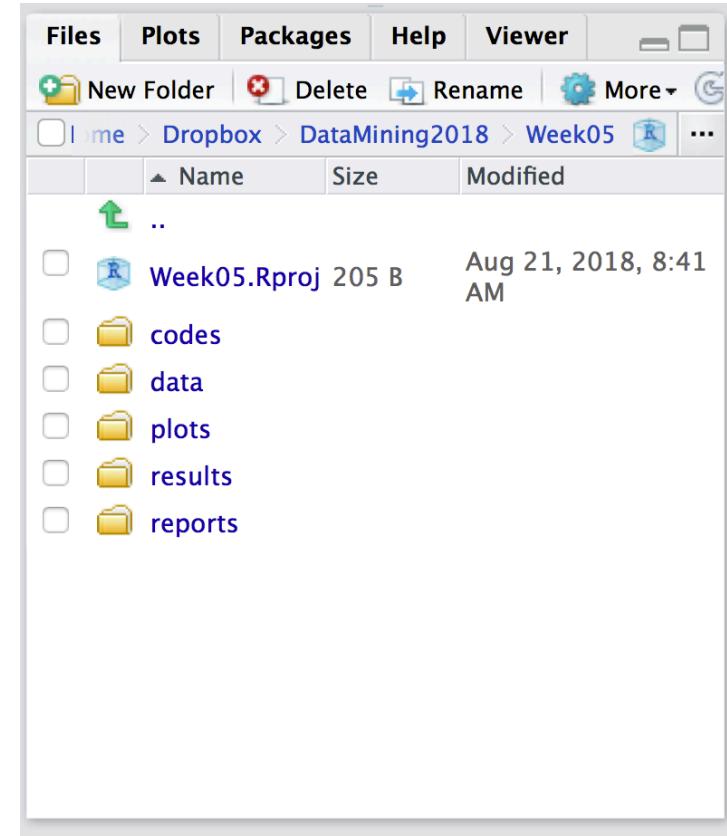
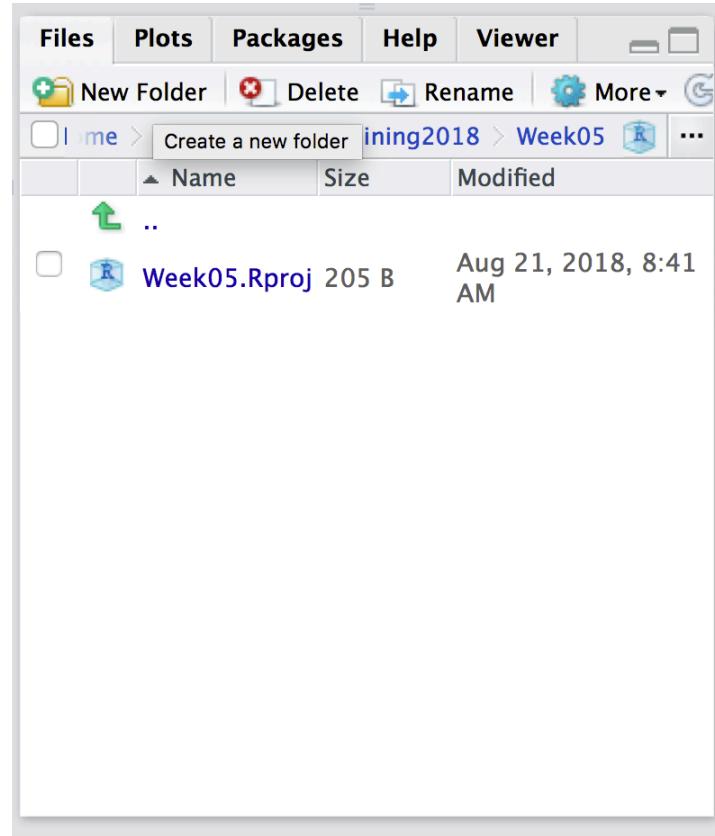
New Project



R Script



Folders



Part 3

Data Extraction

Write and Read Data

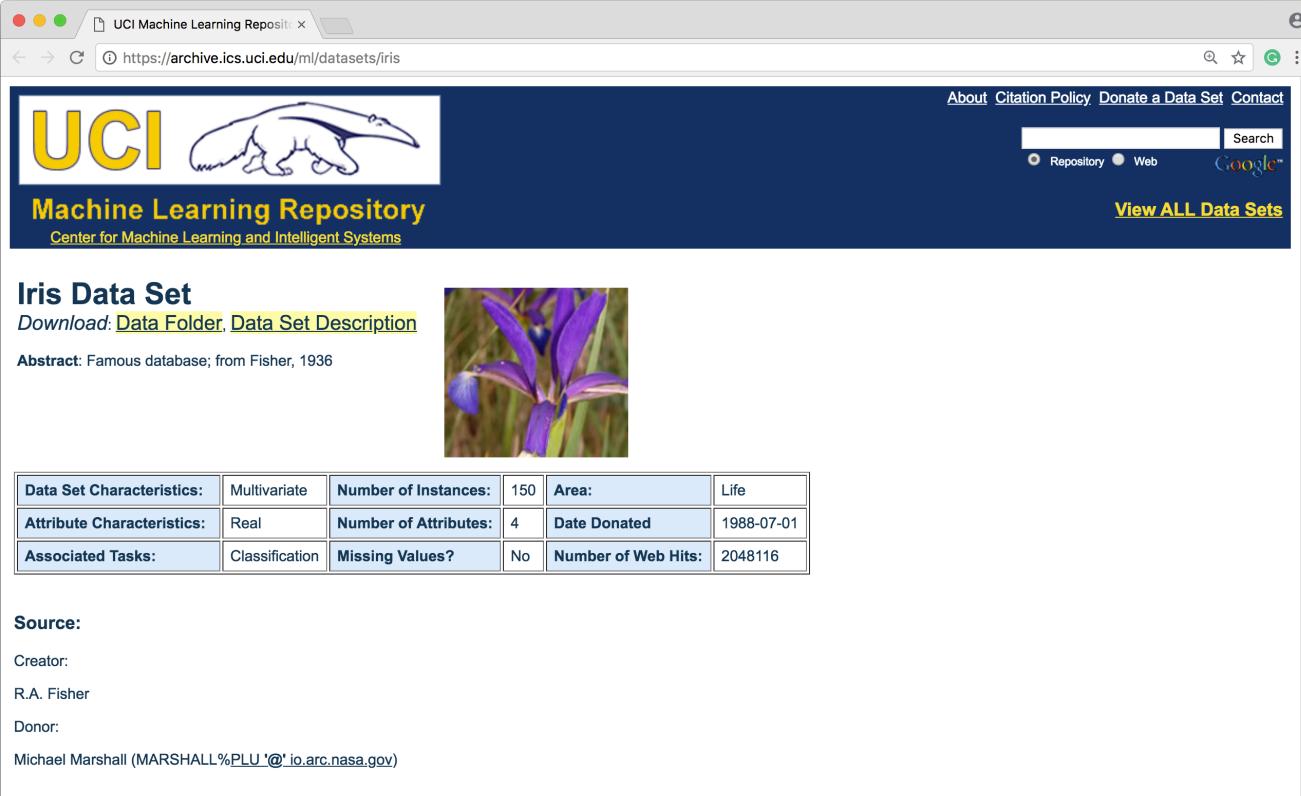
The screenshot shows the RStudio interface with the following components:

- Title Bar:** ~/Dropbox/DataMining2018/Week05 - RStudio
- Source Editor:** Displays R code for generating data frames and writing them to a CSV file, then reading them back.
- Console:** Shows the output of the R code, including the creation of vectors var1, var2, and var3, their conversion into a data frame df1, and the resulting data frame structure.
- Environment Tab:** Shows the global environment with objects df1, df2, var1, var2, and var3.
- Data Tab:** Shows the data frames df1 and df2 with their respective structures.
- Values Tab:** Shows the contents of the variables var1, var2, and var3.
- Files Tab:** Shows the project structure with folders codes, data, plots, reports, results, and the Week05.Rproj file.
- Plots, Packages, Help, and Viewer tabs:** Standard RStudio navigation tabs.

```
> var1 <- 1:5
> var1
[1] 1 2 3 4 5
> var2 <- (1:5)/10
> var2
[1] 0.1 0.2 0.3 0.4 0.5
> var3 <- c("R", "and", "Data Mining", "Examples", "Case Studies")
> var3
[1] "R"           "and"         "Data Mining"   "Examples"     "Case Studies"
> df1 <- data.frame(var1, var2, var3)
> df1
  var1 var2      var3
1    1  0.1        R
2    2  0.2      and
3    3  0.3 Data Mining
4    4  0.4   Examples
5    5  0.5 Case Studies
> names(df1) <- c("VariableInt", "VariableReal", "VariableChar")
> df1
  VariableInt VariableReal VariableChar
1            1          0.1             R
2            2          0.2           and
3            3          0.3 Data Mining
4            4          0.4   Examples
5            5          0.5 Case Studies
> write.csv(df1, "./data/dummyData.csv", row.names = FALSE)
> df2 <- read.csv("./data/dummyData.csv")
> df2
  VariableInt VariableReal VariableChar
1            1          0.1             R
2            2          0.2           and
3            3          0.3 Data Mining
4            4          0.4   Examples
5            5          0.5 Case Studies
>
```

Download Dataset

- Iris Dataset: <https://archive.ics.uci.edu/ml/datasets/iris>
- File “iris.data”: <https://archive.ics.uci.edu/ml/machine-learning-databases/iris/>
- Save to data folder



The screenshot shows a web browser displaying the UCI Machine Learning Repository. The URL in the address bar is <https://archive.ics.uci.edu/ml/datasets/iris>. The page features the UCI logo and a banner for the Machine Learning Repository. The main content area is titled "Iris Data Set" and includes links for "Data Folder" and "Data Set Description". A photograph of an Iris flower is displayed. Below the title, there is an abstract stating "Famous database; from Fisher, 1936". A table provides dataset characteristics:

Data Set Characteristics:	Multivariate	Number of Instances:	150	Area:	Life
Attribute Characteristics:	Real	Number of Attributes:	4	Date Donated:	1988-07-01
Associated Tasks:	Classification	Missing Values?	No	Number of Web Hits:	2048116

Below the table, sections for "Source" and "Creator" are present, along with the name "R.A. Fisher". The "Source" section also lists "Michael Marshall (MARSHALL%PLU '@' io.arc.nasa.gov)".

Extract Data

The screenshot shows the RStudio interface with the following components:

- Top Bar:** Shows the path `~/Dropbox/DataMining2018/Week05 - RStudio` and a tab bar with `myCode.R` and `dummyData.csv`.
- Code Editor:** Displays the following R code:

```
1 # Extract Data
2 iris <- read.table("./data/iris.data", sep = ',')
3
4 # Assign name to variables.
5 names(iris) <- c("Sepal.Length", "Sepal.Width", "Petal.Length", "Petal.Width", "Species")
6
7
```
- Environment Panel:** Shows the `iris` dataset with the following structure:

	150 obs. of 5 variables
Sepal.Length	num 5.1 4.9 4.7 4.6 5 5...
Sepal.Width	num 3.5 3 3.2 3.1 3.6 3...
Petal.Length	num 1.4 1.4 1.3 1.5 1.4...
Petal.Width	num 0.2 0.2 0.2 0.2 0.2...
Species	Factor w/ 3 levels "Iris-setosa", "Iris-versicolor", "Iris-virginica"
- File Browser:** Shows the project structure:

```
Week05
├── codes
└── data
```

Part 4

Data Exploration

Explore Dataset

The screenshot shows the RStudio interface with the following components:

- Code Editor:** Displays the script file `myCode.R` which includes code to load the Iris dataset and inspect its structure.
- Console:** Shows the output of running the script, including the dimensions of the dataset, its names, structure, and row names.
- Environment:** Shows the loaded objects, specifically the `iris` dataset, which contains 150 observations of 5 variables: Sepal.Length, Sepal.Width, Petal.Length, Petal.Width, and Species.
- Files:** Shows the project structure with a folder named `Week05` containing subfolders for codes, data, plots, reports, and results.

```
myCode.R x dummyData.csv x
6
7 # Have a look at data
8 dim(iris)
9 names(iris)
10 str(iris)
11 attributes(iris)
12
10:10 (Top Level) R Script
Console ~ /Dropbox/DataMining2018/Week05/
> dim(iris)
[1] 150 5
> names(iris)
[1] "Sepal.Length" "Sepal.Width" "Petal.Length" "Petal.Width"
[5] "Species"
> str(iris)
'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 "Iris-setosa",...: 1 1 1 1 1 1 1 1 1 1 ...
..
> attributes(iris)
$names
[1] "Sepal.Length" "Sepal.Width" "Petal.Length" "Petal.Width"
[5] "Species"

$class
[1] "data.frame"

$row.names
[1] 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16
[17] 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32
[33] 33 34 35 36 37 38 39 40 41 42 43 44 45 46 47 48
[49] 49 50 51 52 53 54 55 56 57 58 59 60 61 62 63 64
[65] 65 66 67 68 69 70 71 72 73 74 75 76 77 78 79 80
[81] 81 82 83 84 85 86 87 88 89 90 91 92 93 94 95 96
[97] 97 98 99 100 101 102 103 104 105 106 107 108 109 110 111 112
[113] 113 114 115 116 117 118 119 120 121 122 123 124 125 126 127 128
[129] 129 130 131 132 133 134 135 136 137 138 139 140 141 142 143 144
[145] 145 146 147 148 149 150

> |
```

Explore Subset of Dataset

The screenshot shows the RStudio interface with the following components:

- Code Editor:** Displays an R script named "myCode.R" containing code to explore a subset of the Iris dataset. The code includes `dim(iris)`, `names(iris)`, `str(iris)`, `attributes(iris)`, and various subset operations like `iris[1:5,]`, `head(iris)`, `tail(iris)`, `iris[1:10, "Sepal.Length"]`, and `iris\$Sepal.Length[1:10]`.
- Console:** Shows the output of the R code. It displays the first 5 rows of the Iris dataset, the head of the dataset, the tail of the dataset, and the first 10 Sepal.Length values.
- Environment:** Shows the global environment with the "iris" dataset loaded. The "iris" dataset is described as having 150 observations and 5 variables: Sepal.Length, Sepal.Width, Petal.Length, Petal.Width, and Species.
- File Explorer:** Shows the project structure under "Week05". The project contains a "Week05.Rproj" file and folders for "codes", "data", "plots", "reports", and "results".

Explore Individual Variable

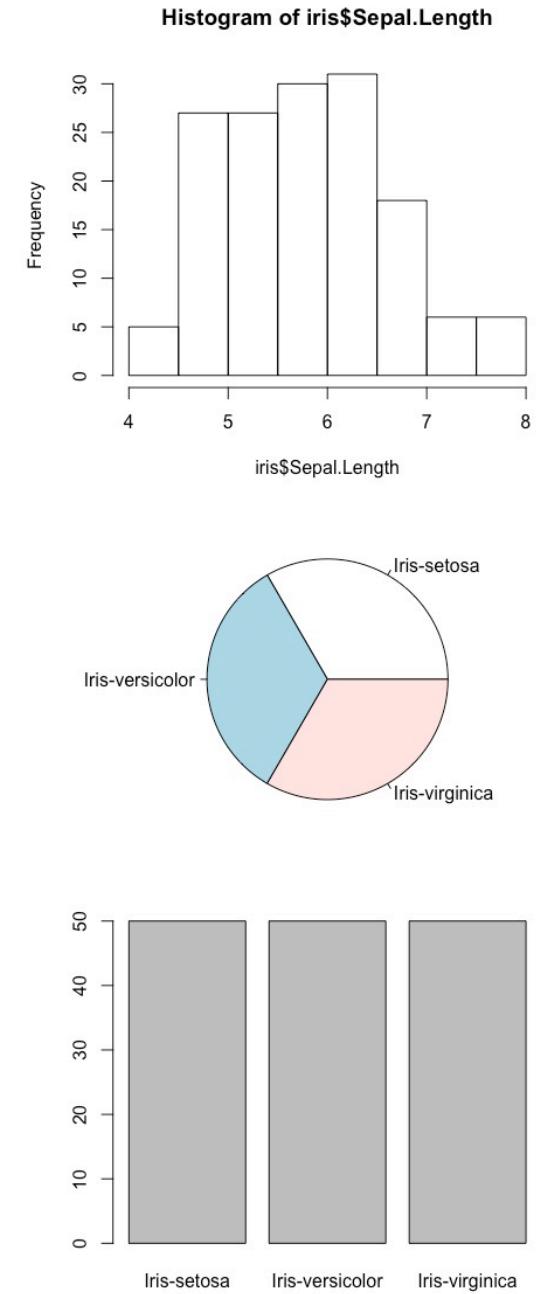
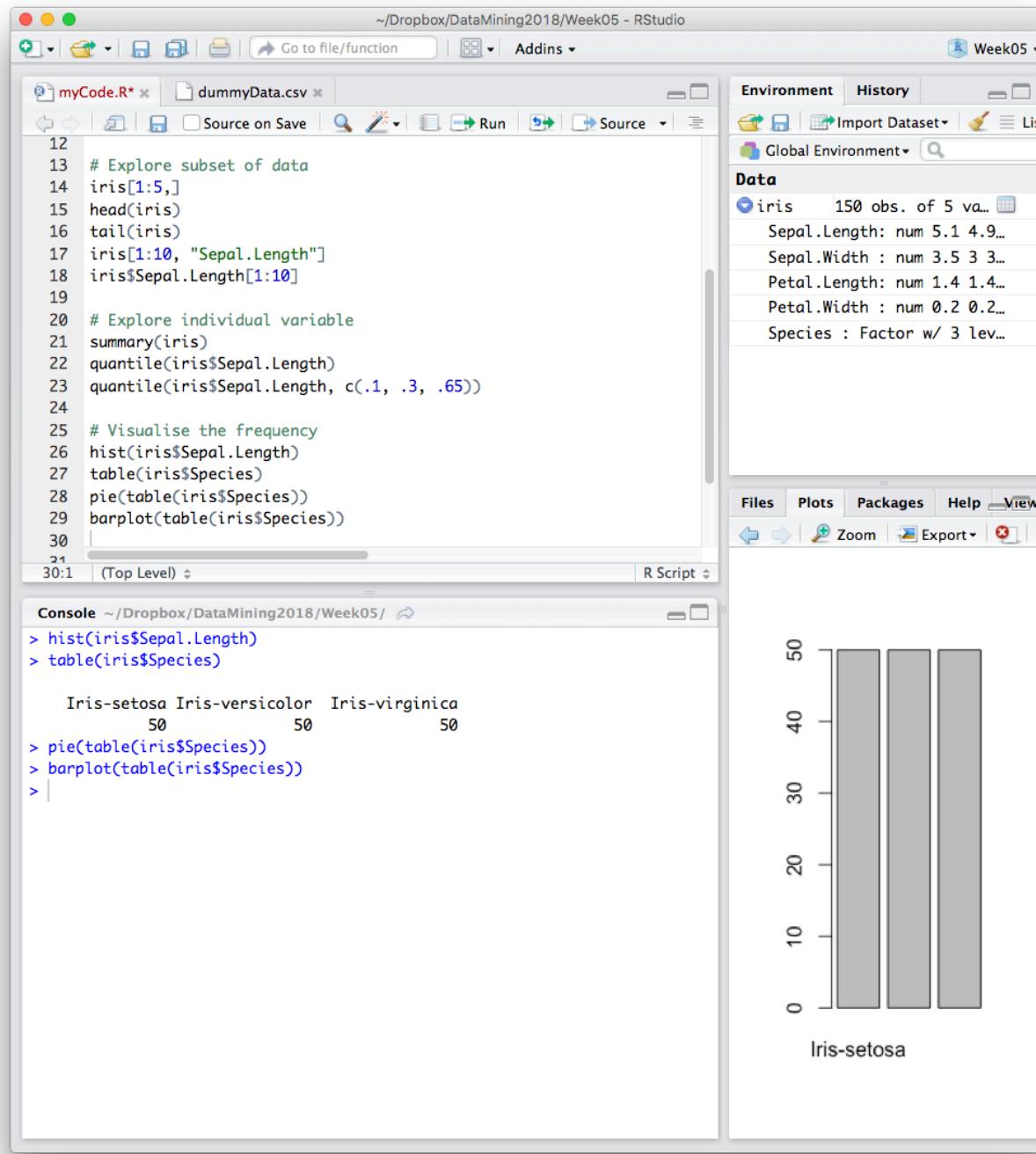
The screenshot shows the RStudio interface with the following components:

- Code Editor:** Displays the file `myCode.R*` containing R code to explore the `iris` dataset.
- Console:** Shows the output of running the code, including the `summary(iris)` command which provides descriptive statistics for each variable, and quantile calculations for Sepal.Length.
- Environment:** Shows the `iris` dataset in the Global Environment, detailing its structure (150 observations, 5 variables) and types (Sepal.Length: numeric, Sepal.Width: numeric, Petal.Length: numeric, Petal.Width: numeric, Species: Factor).
- File Browser:** Shows the project structure under `DataMining2018 > Week05`, including files like `Week05.Rproj` and folders `codes`, `data`, `plots`, `reports`, and `results`.

```
myCode.R*  dummyData.csv
7 # Have a look at data
8 dim(iris)
9 names(iris)
10 str(iris)
11 attributes(iris)
12
13 # Explore subset of data
14 iris[1:5,]
15 head(iris)
16 tail(iris)
17 iris[1:10, "Sepal.Length"]
18 iris$Sepal.Length[1:10]
19
20 # Explore individual variable
21 summary(iris)
22 quantile(iris$Sepal.Length)
23 quantile(iris$Sepal.Length, c(.1, .3, .65))
24
25
20:30 (Top Level) R Script
Console ~ /Dropbox / DataMining2018 / Week05 /
> summary(iris)
   Sepal.Length   Sepal.Width   Petal.Length   Petal.Width
Min.    :4.300   Min.    :2.000   Min.    :1.000   Min.    :0.100
1st Qu.:5.100   1st Qu.:2.800   1st Qu.:1.600   1st Qu.:0.300
Median :5.800   Median :3.000   Median :4.350   Median :1.300
Mean   :5.843   Mean   :3.054   Mean   :4.375   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
Species
Iris-setosa      :50
Iris-versicolor:50
Iris-virginica :50

> quantile(iris$Sepal.Length)
  0%  25%  50%  75% 100%
4.3  5.1  5.8  6.4  7.9
> quantile(iris$Sepal.Length, c(.1, .3, .65))
  10%  30%  65%
4.80  5.27  6.20
>
```

Visualise the frequency



Reference:

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