# Duplicated Data

## 1. Identifying Duplicated Data

R checks for duplicates across rows through the duplicated() function.

#### Identifying Duplicated Data Code Example 1.1

## Example   
# ---  
# Question: Identify duplicate data in the given dataframe  
# ---  
# OUR CODE GOES BELOW  
#   
  
# Creating our vectors  
# ---  
#   
x1 <- c(2, 4, 5, 6)  
x2 <- c(2, 3, 5, 6)  
x3 <- c(2, 4, 5, 6)  
x4 <- c(2, 4, 5, 6)  
  
# Create a dataframe df from the above vectors  
# ---  
#  
df <- data.frame(rbind(x1, x2, x3, x4))  
  
# Then printing out this dataset   
df

## X1 X2 X3 X4  
## x1 2 4 5 6  
## x2 2 3 5 6  
## x3 2 4 5 6  
## x4 2 4 5 6

# Now lets find the duplicated rows in the dataset df   
# and assign to a variable duplicated\_rows below  
# ---  
#   
duplicated\_rows <- df[duplicated(df),]  
  
# Lets print out the variable duplicated\_rows and see these duplicated rows   
# ---  
# OUR CODE GOES BELOW  
#  
duplicated\_rows

## X1 X2 X3 X4  
## x3 2 4 5 6  
## x4 2 4 5 6

# Removing these duplicated rows in the dataset or   
# showing these unique items and assigning to a variable unique\_items below  
# ---  
#  
unique\_items <- df[!duplicated(df), ]  
  
# What about seeing what these unique items are?  
# ---  
#  
unique\_items

## X1 X2 X3 X4  
## x1 2 4 5 6  
## x2 2 3 5 6

# Now there is another way we can also remove duplicated rows   
# in the dataset or show the unique items;  
# We simply use the unique() function  
# ---  
#  
unique\_items2 <- unique(df)  
  
# After having assigned the unique items to the variable unique\_items2,   
# we will now print out this variable and have a look at these unique items   
# ---  
# OUR CODE GOES BELOW  
#   
unique\_items2

## X1 X2 X3 X4  
## x1 2 4 5 6  
## x2 2 3 5 6

## Challenge   
# ---  
# Question: Display and delete the only duplicate records in the iris dataset below:  
# ---  
# OUR CODE GOES BELOW   
#   
  
# Showing the first 6 records in the iris dataset  
# ---  
#   
head(iris)

## 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

# Deleting duplicate records  
# ---  
# OUR CODE GOES BELOW   
#

# Now lets find the duplicated rows in the dataset df   
# and assign to a variable duplicated\_rows below  
# ---  
#   
duplicated\_rows <- iris[duplicated(iris),]  
  
# Lets print out the variable duplicated\_rows and see these duplicated rows   
# ---  
# OUR CODE GOES BELOW  
#  
duplicated\_rows

## Sepal.Length Sepal.Width Petal.Length Petal.Width Species  
## 143 5.8 2.7 5.1 1.9 virginica

# Now there is another way we can also remove duplicated rows   
# in the dataset or show the unique items;  
# We simply use the unique() function  
# ---  
#  
unique\_items3 <- unique(iris)  
  
# After having assigned the unique items to the variable unique\_items2,   
# we will now print out this variable and have a look at these unique items   
# ---  
# OUR CODE GOES BELOW  
#   
unique\_items3

## 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  
## 7 4.6 3.4 1.4 0.3 setosa  
## 8 5.0 3.4 1.5 0.2 setosa  
## 9 4.4 2.9 1.4 0.2 setosa  
## 10 4.9 3.1 1.5 0.1 setosa  
## 11 5.4 3.7 1.5 0.2 setosa  
## 12 4.8 3.4 1.6 0.2 setosa  
## 13 4.8 3.0 1.4 0.1 setosa  
## 14 4.3 3.0 1.1 0.1 setosa  
## 15 5.8 4.0 1.2 0.2 setosa  
## 16 5.7 4.4 1.5 0.4 setosa  
## 17 5.4 3.9 1.3 0.4 setosa  
## 18 5.1 3.5 1.4 0.3 setosa  
## 19 5.7 3.8 1.7 0.3 setosa  
## 20 5.1 3.8 1.5 0.3 setosa  
## 21 5.4 3.4 1.7 0.2 setosa  
## 22 5.1 3.7 1.5 0.4 setosa  
## 23 4.6 3.6 1.0 0.2 setosa  
## 24 5.1 3.3 1.7 0.5 setosa  
## 25 4.8 3.4 1.9 0.2 setosa  
## 26 5.0 3.0 1.6 0.2 setosa  
## 27 5.0 3.4 1.6 0.4 setosa  
## 28 5.2 3.5 1.5 0.2 setosa  
## 29 5.2 3.4 1.4 0.2 setosa  
## 30 4.7 3.2 1.6 0.2 setosa  
## 31 4.8 3.1 1.6 0.2 setosa  
## 32 5.4 3.4 1.5 0.4 setosa  
## 33 5.2 4.1 1.5 0.1 setosa  
## 34 5.5 4.2 1.4 0.2 setosa  
## 35 4.9 3.1 1.5 0.2 setosa  
## 36 5.0 3.2 1.2 0.2 setosa  
## 37 5.5 3.5 1.3 0.2 setosa  
## 38 4.9 3.6 1.4 0.1 setosa  
## 39 4.4 3.0 1.3 0.2 setosa  
## 40 5.1 3.4 1.5 0.2 setosa  
## 41 5.0 3.5 1.3 0.3 setosa  
## 42 4.5 2.3 1.3 0.3 setosa  
## 43 4.4 3.2 1.3 0.2 setosa  
## 44 5.0 3.5 1.6 0.6 setosa  
## 45 5.1 3.8 1.9 0.4 setosa  
## 46 4.8 3.0 1.4 0.3 setosa  
## 47 5.1 3.8 1.6 0.2 setosa  
## 48 4.6 3.2 1.4 0.2 setosa  
## 49 5.3 3.7 1.5 0.2 setosa  
## 50 5.0 3.3 1.4 0.2 setosa  
## 51 7.0 3.2 4.7 1.4 versicolor  
## 52 6.4 3.2 4.5 1.5 versicolor  
## 53 6.9 3.1 4.9 1.5 versicolor  
## 54 5.5 2.3 4.0 1.3 versicolor  
## 55 6.5 2.8 4.6 1.5 versicolor  
## 56 5.7 2.8 4.5 1.3 versicolor  
## 57 6.3 3.3 4.7 1.6 versicolor  
## 58 4.9 2.4 3.3 1.0 versicolor  
## 59 6.6 2.9 4.6 1.3 versicolor  
## 60 5.2 2.7 3.9 1.4 versicolor  
## 61 5.0 2.0 3.5 1.0 versicolor  
## 62 5.9 3.0 4.2 1.5 versicolor  
## 63 6.0 2.2 4.0 1.0 versicolor  
## 64 6.1 2.9 4.7 1.4 versicolor  
## 65 5.6 2.9 3.6 1.3 versicolor  
## 66 6.7 3.1 4.4 1.4 versicolor  
## 67 5.6 3.0 4.5 1.5 versicolor  
## 68 5.8 2.7 4.1 1.0 versicolor  
## 69 6.2 2.2 4.5 1.5 versicolor  
## 70 5.6 2.5 3.9 1.1 versicolor  
## 71 5.9 3.2 4.8 1.8 versicolor  
## 72 6.1 2.8 4.0 1.3 versicolor  
## 73 6.3 2.5 4.9 1.5 versicolor  
## 74 6.1 2.8 4.7 1.2 versicolor  
## 75 6.4 2.9 4.3 1.3 versicolor  
## 76 6.6 3.0 4.4 1.4 versicolor  
## 77 6.8 2.8 4.8 1.4 versicolor  
## 78 6.7 3.0 5.0 1.7 versicolor  
## 79 6.0 2.9 4.5 1.5 versicolor  
## 80 5.7 2.6 3.5 1.0 versicolor  
## 81 5.5 2.4 3.8 1.1 versicolor  
## 82 5.5 2.4 3.7 1.0 versicolor  
## 83 5.8 2.7 3.9 1.2 versicolor  
## 84 6.0 2.7 5.1 1.6 versicolor  
## 85 5.4 3.0 4.5 1.5 versicolor  
## 86 6.0 3.4 4.5 1.6 versicolor  
## 87 6.7 3.1 4.7 1.5 versicolor  
## 88 6.3 2.3 4.4 1.3 versicolor  
## 89 5.6 3.0 4.1 1.3 versicolor  
## 90 5.5 2.5 4.0 1.3 versicolor  
## 91 5.5 2.6 4.4 1.2 versicolor  
## 92 6.1 3.0 4.6 1.4 versicolor  
## 93 5.8 2.6 4.0 1.2 versicolor  
## 94 5.0 2.3 3.3 1.0 versicolor  
## 95 5.6 2.7 4.2 1.3 versicolor  
## 96 5.7 3.0 4.2 1.2 versicolor  
## 97 5.7 2.9 4.2 1.3 versicolor  
## 98 6.2 2.9 4.3 1.3 versicolor  
## 99 5.1 2.5 3.0 1.1 versicolor  
## 100 5.7 2.8 4.1 1.3 versicolor  
## 101 6.3 3.3 6.0 2.5 virginica  
## 102 5.8 2.7 5.1 1.9 virginica  
## 103 7.1 3.0 5.9 2.1 virginica  
## 104 6.3 2.9 5.6 1.8 virginica  
## 105 6.5 3.0 5.8 2.2 virginica  
## 106 7.6 3.0 6.6 2.1 virginica  
## 107 4.9 2.5 4.5 1.7 virginica  
## 108 7.3 2.9 6.3 1.8 virginica  
## 109 6.7 2.5 5.8 1.8 virginica  
## 110 7.2 3.6 6.1 2.5 virginica  
## 111 6.5 3.2 5.1 2.0 virginica  
## 112 6.4 2.7 5.3 1.9 virginica  
## 113 6.8 3.0 5.5 2.1 virginica  
## 114 5.7 2.5 5.0 2.0 virginica  
## 115 5.8 2.8 5.1 2.4 virginica  
## 116 6.4 3.2 5.3 2.3 virginica  
## 117 6.5 3.0 5.5 1.8 virginica  
## 118 7.7 3.8 6.7 2.2 virginica  
## 119 7.7 2.6 6.9 2.3 virginica  
## 120 6.0 2.2 5.0 1.5 virginica  
## 121 6.9 3.2 5.7 2.3 virginica  
## 122 5.6 2.8 4.9 2.0 virginica  
## 123 7.7 2.8 6.7 2.0 virginica  
## 124 6.3 2.7 4.9 1.8 virginica  
## 125 6.7 3.3 5.7 2.1 virginica  
## 126 7.2 3.2 6.0 1.8 virginica  
## 127 6.2 2.8 4.8 1.8 virginica  
## 128 6.1 3.0 4.9 1.8 virginica  
## 129 6.4 2.8 5.6 2.1 virginica  
## 130 7.2 3.0 5.8 1.6 virginica  
## 131 7.4 2.8 6.1 1.9 virginica  
## 132 7.9 3.8 6.4 2.0 virginica  
## 133 6.4 2.8 5.6 2.2 virginica  
## 134 6.3 2.8 5.1 1.5 virginica  
## 135 6.1 2.6 5.6 1.4 virginica  
## 136 7.7 3.0 6.1 2.3 virginica  
## 137 6.3 3.4 5.6 2.4 virginica  
## 138 6.4 3.1 5.5 1.8 virginica  
## 139 6.0 3.0 4.8 1.8 virginica  
## 140 6.9 3.1 5.4 2.1 virginica  
## 141 6.7 3.1 5.6 2.4 virginica  
## 142 6.9 3.1 5.1 2.3 virginica  
## 144 6.8 3.2 5.9 2.3 virginica  
## 145 6.7 3.3 5.7 2.5 virginica  
## 146 6.7 3.0 5.2 2.3 virginica  
## 147 6.3 2.5 5.0 1.9 virginica  
## 148 6.5 3.0 5.2 2.0 virginica  
## 149 6.2 3.4 5.4 2.3 virginica  
## 150 5.9 3.0 5.1 1.8 virginica

## Challenge   
# ---  
# Question: Drop duplicate records in the iris games dataset from the url   
# ---  
# OUR CODE GOES BELOW  
#   
  
# Importing the data.table  
# ---  
#   
library("data.table")  
  
# Reading our dataset  
# ---  
#   
video\_games <- fread('http://bit.ly/VideoGamesDataset')  
  
# Previewing the first 6 records of the video games dataset  
# ---  
#   
head(video\_games)

## V1 V2 V3 V4 V5  
## 1: 151603712 The Elder Scrolls V Skyrim purchase 1.0 0  
## 2: 151603712 The Elder Scrolls V Skyrim play 273.0 0  
## 3: 151603712 Fallout 4 purchase 1.0 0  
## 4: 151603712 Fallout 4 play 87.0 0  
## 5: 151603712 Spore purchase 1.0 0  
## 6: 151603712 Spore play 14.9 0

# Cleaning our dataset   
# ---  
# OUR CODE GOES BELOW  
#

# Now lets find the duplicated rows in the dataset df   
# and assign to a variable duplicated\_rows below  
# ---  
#   
duplicated\_rows <- video\_games[duplicated(video\_games),]  
  
# Lets print out the variable duplicated\_rows and see these duplicated rows   
# ---  
# OUR CODE GOES BELOW  
#  
duplicated\_rows

## V1 V2 V3 V4 V5  
## 1: 11373749 Sid Meier's Civilization IV purchase 1 0  
## 2: 11373749 Sid Meier's Civilization IV Beyond the Sword purchase 1 0  
## 3: 11373749 Sid Meier's Civilization IV Warlords purchase 1 0  
## 4: 56038151 Grand Theft Auto San Andreas purchase 1 0  
## 5: 56038151 Grand Theft Auto Vice City purchase 1 0  
## ---   
## 703: 39146470 Sid Meier's Civilization IV Warlords purchase 1 0  
## 704: 48666962 Crysis 2 purchase 1 0  
## 705: 112845094 Grand Theft Auto San Andreas purchase 1 0  
## 706: 112845094 Grand Theft Auto Vice City purchase 1 0  
## 707: 112845094 Grand Theft Auto III purchase 1 0

# Removing these duplicated rows in the dataset or   
# showing these unique items and assigning to a variable unique\_items below  
# ---  
#  
unique\_items <- video\_games[!duplicated(video\_games), ]  
  
# What about seeing what these unique items are?  
# ---  
#  
unique\_items

## V1 V2 V3 V4 V5  
## 1: 151603712 The Elder Scrolls V Skyrim purchase 1.0 0  
## 2: 151603712 The Elder Scrolls V Skyrim play 273.0 0  
## 3: 151603712 Fallout 4 purchase 1.0 0  
## 4: 151603712 Fallout 4 play 87.0 0  
## 5: 151603712 Spore purchase 1.0 0  
## ---   
## 199289: 128470551 Titan Souls play 1.5 0  
## 199290: 128470551 Grand Theft Auto Vice City purchase 1.0 0  
## 199291: 128470551 Grand Theft Auto Vice City play 1.5 0  
## 199292: 128470551 RUSH purchase 1.0 0  
## 199293: 128470551 RUSH play 1.4 0

# Now there is another way we can also remove duplicated rows   
# in the dataset or show the unique items;  
# We simply use the unique() function  
# ---  
#  
unique\_items2 <- unique(video\_games)  
  
# After having assigned the unique items to the variable unique\_items2,   
# we will now print out this variable and have a look at these unique items   
# ---  
# OUR CODE GOES BELOW  
#   
unique\_items2

## V1 V2 V3 V4 V5  
## 1: 151603712 The Elder Scrolls V Skyrim purchase 1.0 0  
## 2: 151603712 The Elder Scrolls V Skyrim play 273.0 0  
## 3: 151603712 Fallout 4 purchase 1.0 0  
## 4: 151603712 Fallout 4 play 87.0 0  
## 5: 151603712 Spore purchase 1.0 0  
## ---   
## 199289: 128470551 Titan Souls play 1.5 0  
## 199290: 128470551 Grand Theft Auto Vice City purchase 1.0 0  
## 199291: 128470551 Grand Theft Auto Vice City play 1.5 0  
## 199292: 128470551 RUSH purchase 1.0 0  
## 199293: 128470551 RUSH play 1.4 0