R Notebook

Duplicated Data

1. Identifying Duplicated Data

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

```
## Example
# Question: Identify duplicate data in the given dataframe

# 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
```

Identifying Duplicated Data Code Example 1.1

```
##    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),]

duplicated_rows
```

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

```r
# 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
```

```r
# Now there is another way we can also remove duplicated rows in the dataset or show the unique items;

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
unique_items2
```

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

CHALLENGE 1

```r
# Question: Display and delete the only duplicate records in the iris dataset 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
```

```r
# Deleting duplicate records
# Checking the number of rows before deletion
nrow(iris)
```

```
## [1] 150
```

```
# Checking the number after deletion
iris_deuplicates_dropped <- unique(iris)
nrow(iris_deuplicates_dropped)
```

```
## [1] 149
```

CHALLENGE 2

```
# Question: Drop duplicate records in the iris games dataset from the url

# Importing the data.table
library("data.table")
```

```
## Warning: package 'data.table' was built under R version 4.0.4
```

```
# 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
```

```
# Number of rows
nrow(video_games)
```

```
## [1] 200000
```

```
# Number of rows duplicated
nrow(video_games[duplicated(video_games),])
```

```
## [1] 707
```

```
# Dropping the duplicates
video_games <- unique(video_games)

# Checking the new number of rows
nrow(video_games)
```

```
## [1] 199293
```