--------------------------------------------

## ATTACH & DETACH FUNCTIONS

--------------------------------------------

- find what attach() and detach() commands do???

--------------------------------------------

## HEAD & TAIL FUNCTIONS

--------------------------------------------

- find what head() and tail() commands do
- Use head() and tail() commands to display sample observations of mtcars dataset
- Use head() command to Print first 10 observations
- Use tail() commands to Print last 15 observations

--------------------------------

## SORTING

--------------------------------

- Sort the observations of the dataset “mtcars” in increasing order based on the values in the column "mpg"
- Sort the observations of the dataset “mtcars” in decreasing order based on the values in the column "cyl"
- Sort the observations of the dataset “mtcars” in increasing order based on the values in the columns both "mpg" and "cyl"
- Sort the observations of the dataset “mtcars” in decreasing order based on the values in the columns both "mpg" and "cyl”
- Sort the observations of the dataset “mtcars” by column “mpg” in increasing order and column “cyl” in decreasing order

**Data Cleaning:**

---------------------------------------------------

DETERMINE DUPLICATE ELEMENTS

---------------------------------------------------

**Finding Duplicate Values:**

*duplicated()* determines which elements of a vector or data frame are duplicates of elements with smaller subscripts, and returns a logical vector indicating which elements (rows) are duplicates.

*anyDuplicated()* returns the index of the first duplicate value if any, otherwise 0. *anyDuplicated()* is a “generalized” more efficient shortcut for *any(duplicated())*

---------------------------------------------

EXERCISES

---------------------------------------------

- Create a vector as x <- c(9:20, 1:5, 3:7, 0:8)
- Use *duplicated()* function to print the logical vector indicating the duplicate values present in x
- Observe the output of *duplicated(x, fromLast = TRUE)*
- What is the difference between *duplicated(x)* and *duplicated(x,fromLast=TRUE)*
- Extract duplicate elements from x
- Extract unique elements from x
- Print duplicate elements from x in different order (**Hint:** Use *duplicated(x, fromLast = TRUE)*)
- Extract unique elements from x in different order (**Hint:** Use *duplicated(x, fromLast = TRUE)*)
- Print the indices of duplicate elements
- Print the indices of unique elements
- How many unique elements are in x
- How many duplicate elements are in x
- Create a dataframe df :

```
a <- c(rep("A", 3), rep("B", 3), rep("C",2))

b <- c(1,1,2,4,1,1,2,2)

df <-data.frame(a,b)
```

- Use *duplicated()* function to print the logical vector indicating the duplicate values present in dataframe “df”
- Extract duplicate elements from dataframe “df”
- Extract unique elements from dataframe “df”
- Print the indices of duplicate elements
- Print the indices of unique elements
- How many unique elements are in dataframe "df"
- How many duplicate elements are in dataframe "df"

---------------------------------------------

DATASET – INTRODUCTION

---------------------------------------------

## Fisher’s Iris Dataset

Iris Flowers

## Description

This famous (Fisher's or Anderson's) iris data set gives the measurements in centimeters of the variables sepal length and width and petal length and width, respectively, for 50 flowers from each of 3 species of iris. The species are Iris setosa, versicolor, and virginica.

## Format

iris is a data frame with 150 cases (rows) and 5 variables (columns) named Sepal.Length, Sepal.Width, Petal.Length, Petal.Width, and Species

---------------------------------------------

EXERCISES

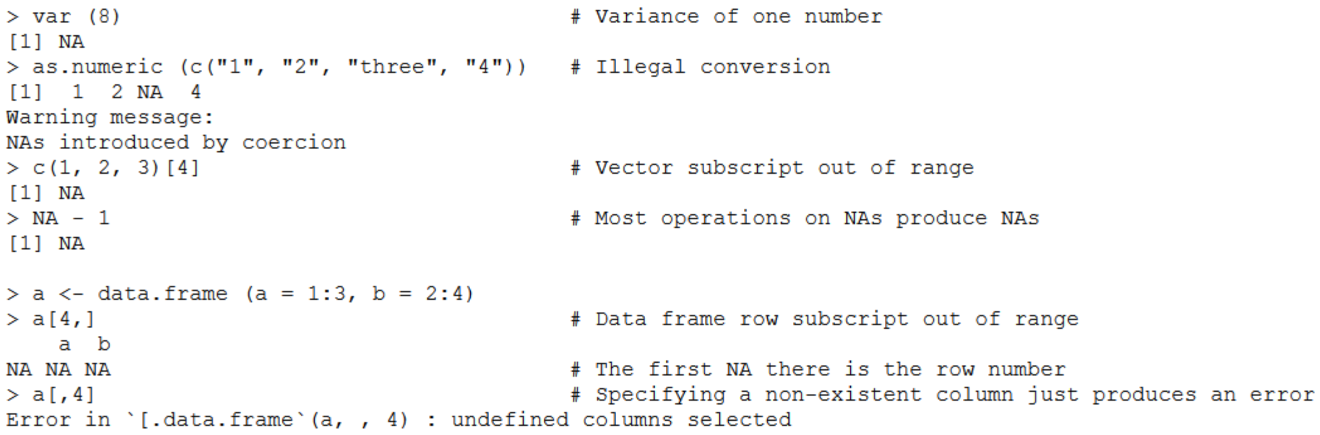
---------------------------------------------

- Print the dataset *iris*
- Print the structure of the dataset *iris*
- Print the summary of all the variables of the dataset *iris* (**Hint:** Use function *summary()*)
- How many of the variables (columns) are in the dataset *iris*
- How many observations (rows) are in the dataset *iris*
- Use *duplicated()* function to print the logical vector indicating the duplicate values present in the dataset *iris*
- Extract duplicate elements from the dataset *iris*
- Extract unique elements from the dataset *iris*
- Print the indices of duplicate elements in the dataset *iris*
- Print the indices of unique elements in the dataset *iris*
- How many unique elements are in the dataset *iris*
- How many duplicate elements are in the dataset *iris*

## Missing Values:

- A missing value is one whose value is unknown.
- Missing values are represented in R by the NA symbol.
- NA is a special value whose properties are different from other values.
- NA is one of the very few reserved words in R: you cannot give anything this name.
- Missing values are often legitimate: values really are missing in real life.
- NAs can arise when you read in a Excel spreadsheet with empty cells, for example.
- You will also see NA when you try certain operations that are illegal or don't make sense.

Here are some examples of operations that produce NA's.



```
> var (8)                                  # Variance of one number
[1] NA
> as.numeric (c("1", "2", "three", "4"))   # Illegal conversion
[1]  1  2 NA  4
Warning message:
NAs introduced by coercion
> c(1, 2, 3)[4]                            # Vector subscript out of range
[1] NA
> NA - 1                                   # Most operations on NAs produce NAs
[1] NA

> a <- data.frame (a = 1:3, b = 2:4)
> a[4,]                                    # Data frame row subscript out of range
    a  b
NA NA NA                                   # The first NA there is the row number
> a[,4]                                    # Specifying a non-existent column just produces an error
Error in `[.data.frame`(a, , 4) : undefined columns selected
```

----------------------------------------------

EXERCISES

----------------------------------------------

- Practice above examples that generate NA values
- Create NA values by some illegal operations
- Practice exercises in lecture slide
- What happens when we try to sort the data with NA values
- How to find the length of a vector with NA values