#DataScience
Exploratory Data Analysis in R: Towards Data Understanding

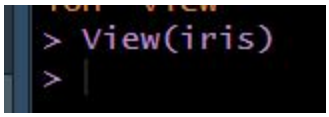
Goal: Learn how to understand data using **exploratory analysis**

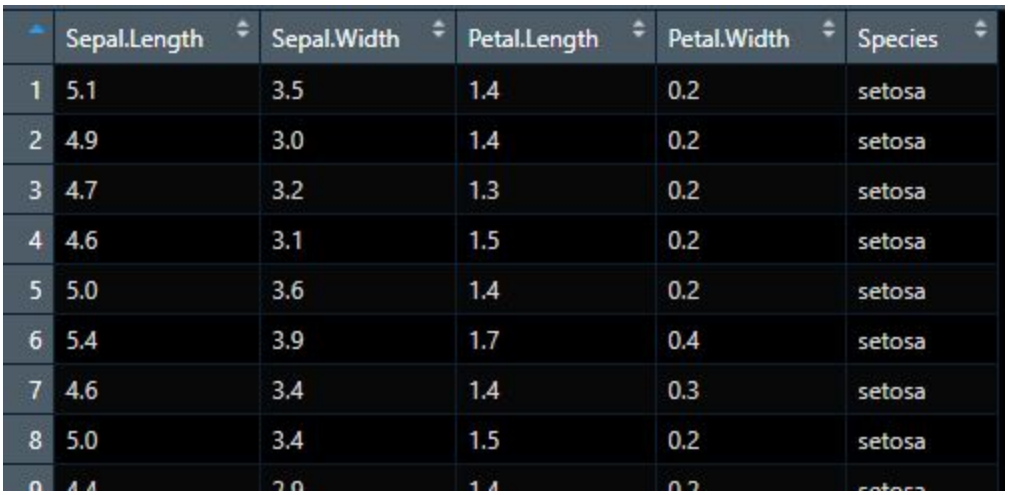
Note: can also use RStudio.cloud

Look at Data Science 101 (Refer to references)

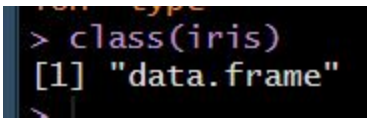
- Will be using the iris data set again
  - Refer to Web App repository and pdf regarding information about this dataset
  - https://github.com/mathstudent97/WebAppsInR_Part2/tree/main/4_WebApp

Viewing the Data & Importing the Data into the Environment

- **View(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
7	4.6	3.4	1.4	0.3	setosa
8	5.0	3.4	1.5	0.2	setosa

- Notice: you get iris data as a dataframe
- **class(iris)**
- 

- Can view data within a certain column (in this case it's Sepal Length; *Sepal.Length*) of the dataset / dataframe
  - So basically the format to retrieve data within a column is **data$column**; you could do this with any column
  - **iris$Sepal.Length**

```
> iris$Sepal.Length
  [1] 5.1 4.9 4.7 4.6 5.0 5.4 4.6 5.0 4.4
 [10] 4.9 5.4 4.8 4.8 4.3 5.8 5.7 5.4 5.1
 [19] 5.7 5.1 5.4 5.1 4.6 5.1 4.8 5.0 5.0
 [28] 5.2 5.2 4.7 4.8 5.4 5.2 5.5 4.9 5.0
 [37] 5.5 4.9 4.4 5.1 5.0 4.5 4.4 5.0 5.1
 [46] 4.8 5.1 4.6 5.3 5.0 7.0 6.4 6.9 5.5
 [55] 6.5 5.7 6.3 4.9 6.6 5.2 5.0 5.9 6.0
 [64] 6.1 5.6 6.7 5.6 5.8 6.2 5.6 5.9 6.1
 [73] 6.3 6.1 6.4 6.6 6.8 6.7 6.0 5.7 5.5
 [82] 5.5 5.8 6.0 5.4 6.0 6.7 6.3 5.6 5.5
 [91] 5.5 6.1 5.8 5.0 5.6 5.7 5.7 6.2 5.1
[100] 5.7 6.3 5.8 7.1 6.3 6.5 7.6 4.9 7.3
[109] 6.7 7.2 6.5 6.4 6.8 5.7 5.8 6.4 6.5
[118] 7.7 7.7 6.0 6.9 5.6 7.7 6.3 6.7 7.2
[127] 6.2 6.1 6.4 7.2 7.4 7.9 6.4 6.3 6.1
[136] 7.7 6.3 6.4 6.0 6.9 6.7 6.9 5.8 6.8
[145] 6.7 6.7 6.3 6.5 6.2 5.9
```

-

	Sepal.Length
1	5.1
2	4.9
3	4.7
4	4.6
5	5.0

-

Summary Statistics

- Simply typing in iris provides you with the dataset as output
- **iris**

```
124          4.9          1.8  virginica
125          5.7          2.1  virginica
126          6.0          1.8  virginica
127          4.8          1.8  virginica
128          4.9          1.8  virginica
129          5.6          2.1  virginica
130          5.8          1.6  virginica
131          6.1          1.9  virginica
```

-

-

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

    - These **4 variables are the independent variables** that will allow the prediction model to learn the characteristics of the different types of flowers.
        - There are three types of flowers: virginica, versicolor, setosa
        - So, on the **basis of the 4 characteristics, the model will be able to predict / determine the type of flower**
            - (I made a web app for this in my other repository: https://github.com/mathstudent97/WebAppsInR_Part2/tree/main/4_WebApp)
- **head(iris, 5)**
-

```
> head(iris, 5)
  Sepal.Length Sepal.Width Petal.Length
1          5.1         3.5          1.4
2          4.9         3.0          1.4
3          4.7         3.2          1.3
4          4.6         3.1          1.5
5          5.0         3.6          1.4
  Petal.Width Species
1         0.2  setosa
2         0.2  setosa
3         0.2  setosa
4         0.2  setosa
5         0.2  setosa
>
```

    - First 5 lines / rows of the dataset
- **tail(iris, 5)**
-

```
    Sepal.Length Sepal.Width Petal.Length
146          6.7         3.0          5.2
147          6.3         2.5          5.0
148          6.5         3.0          5.2
149          6.2         3.4          5.4
150          5.9         3.0          5.1
    Petal.Width   Species
146         2.3 virginica
147         1.9 virginica
148         2.0 virginica
149         2.3 virginica
150         1.8 virginica
>
```

    - Last 5 lines / rows of the dataset
- **summary(iris)**

```
> 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   :3.057   Mean   :3.758   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
 setosa    :50
 versicolor:50
 virginica :50
```

- Summary of the statistics within the dataset
  - Shows number of flowers for each species

- **summary(iris$Sepal.Length)**

```
> summary(iris$Sepal.Length)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  4.300   5.100   5.800   5.843   6.400   7.900
>
```

- Summary stats of a certain column / variable

Summation of missing data

- **sum(is.na(iris))**

```
> sum(is.na(iris))
[1] 0
>
```

-
- this is a good way to check if missing data exists within the dataset
  - *0* as output means no missing values exists within the dataset

- **skim(iris)**

```
> skim(iris) # This displays a more expanded summary of stats re
ng the dataset.
-- Data Summary ------------------------
                           Values
Name                       iris
Number of rows             150
Number of columns          5
___________________________
Column type frequency:
  factor                   1
  numeric                  4
___________________________
Group variables            None

-- Variable type: factor ---------------------------------------
# A tibble: 1 x 6
  skim_variable n_missing complete_rate ordered n_unique
* <chr>              <int>         <dbl> <lgl>      <int>
1 Species                0             1 FALSE          3
  top_counts
* <chr>
1 set: 50, ver: 50, vir: 50

-- Variable type: numeric --------------------------------------
# A tibble: 4 x 11
  skim_variable n_missing complete_rate  mean    sd    p0   p25
* <chr>              <int>         <dbl> <dbl> <dbl> <dbl> <dbl>
1 Sepal.Length           0             1  5.84 0.828   4.3   5.1
2 Sepal.Width            0             1  3.06 0.436   2     2.8
3 Petal.Length           0             1  3.76 1.77    1     1.6
4 Petal.Width            0             1  1.20 0.762   0.1   0.3
    p50   p75  p100 hist
* <dbl> <dbl> <dbl> <chr>
1  5.8    6.4   7.9 ▆▇▇▅▂
2  3      3.3   4.4 ▁▆▇▂▁
3  4.35   5.1   6.9 ▇▁▆▇▂
4  1.3    1.8   2.5 ▇▁▆▁▂
```

  - *skimr* library provides a more in-depth summary of stats regarding the data
  - This displays:
    - The name of the dataset, its number of rows & columns
    - The number and type of column types:

- In this case, it shows the number of factor(s) types; in this case there is only one (*Species*) and 4 numeric types (*Sepal length* & *width*, *Petal length* & *width*)
- Group variables
    - None in this case
- Missing values
- Mean values, sd, various quantiles within the data set (p =0, 0.25, 0.50, 75, etc)
- Mini histogram / rough distribution of the data
    - Notice there are 2 populations for petal length and petal width, since the bars are separated

Using skim() by grouping

- Say you want to use skim() by grouping the data in terms of species b/c there are three flower types
    - Ex: This will help answer the Q: What is the mean value of each characteristic for each flower type?

```
-- Data Summary ------------------------
                           Values
Name                       Piped data
Number of rows             150
Number of columns          5
_______________________
Column type frequency:
  numeric                  4
________________________
Group variables            Species

-- Variable type: numeric ---------------------------------------------------------
# A tibble: 12 x 12
   skim_variable Species    n_missing complete_rate  mean    sd    p0   p25   p50
 * <chr>         <fct>          <int>         <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>
 1 Sepal.Length  setosa             0             1 5.01  0.352   4.3  4.8   5
 2 Sepal.Length  versicolor         0             1 5.94  0.516   4.9  5.6   5.9
 3 Sepal.Length  virginica          0             1 6.59  0.636   4.9  6.22  6.5
 4 Sepal.Width   setosa             0             1 3.43  0.379   2.3  3.2   3.4
 5 Sepal.Width   versicolor         0             1 2.77  0.314   2    2.52  2.8
 6 Sepal.Width   virginica          0             1 2.97  0.322   2.2  2.8   3
 7 Petal.Length  setosa             0             1 1.46  0.174   1    1.4   1.5
 8 Petal.Length  versicolor         0             1 4.26  0.470   3    4     4.35
 9 Petal.Length  virginica          0             1 5.55  0.552   4.5  5.1   5.55
10 Petal.Width   setosa             0             1 0.246 0.105   0.1  0.2   0.2
11 Petal.Width   versicolor         0             1 1.33  0.198   1    1.2   1.3
12 Petal.Width   virginica          0             1 2.03  0.275   1.4  1.8   2
    p75  p100 hist
 * <dbl> <dbl> <chr>
 1  5.2    5.8
 2  6.3    7
 3  6.9    7.9
 4  3.68   4.4
 5  3      3.4
 6  3.18   3.8
 7  1.58   1.9
 8  4.6    5.1
 9  5.88   6.9
10  0.3    0.6
11  1.5    1.8
12  2.3    2.5
```

- Ex. The mean value for the sepal length of the setosa flower is 5.01

- This helps with comparisons
    - Ex. for sepal length, the virginica flower appears to have a higher mean value (6.59) compared to setosa (5.01) and versicolor (5.94)

References

https://www.youtube.com/watch?v=JW5Ug6NQexg&list=PLtqF5YXg7GLk9QRC5kS5Am4Ijo4S9gqk_

https://rstudio.cloud/

https://www.youtube.com/watch?v=7XdoaQYwTeA&list=PLtqF5YXg7GLn0WWB_wQx7wHrIvbs0EH2e

https://github.com/mathstudent97/WebAppsInR_Part2/tree/main/4_WebApp