Assignment

read and View data in R

```r
library(readr)
fish_data <- read_csv("fish_data.csv")
```

```
## Rows: 2000 Columns: 8
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (2): habitat, color
## dbl (5): id, average_length, average_weight, ph_of_water, life_span
## lgl (1): Gender
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```r
View(fish_data)
print(fish_data)
```

```
## # A tibble: 2,000 x 8
##       id average_length average_weight habitat    ph_of_water color          Gender
##    <dbl>          <dbl>          <dbl> <chr>            <dbl> <chr>          <lgl>
##  1     1          14.7            5.87 ponds              6.2 Reddish_Ora~ FALSE
##  2     2           1.32           3.86 idlewater          6.8 Calico         TRUE
##  3     3          14.2           12.1  lakes              7.9 Reddish_Ora~ TRUE
##  4     4           2.54           3.2  rivers             6.7 White          FALSE
##  5     5          13.1            9.81 lakes              7.8 Orange         TRUE
##  6     6          15.2            8.99 lakes              7.8 White          FALSE
##  7     7          16.2            5.08 ponds              6.3 Red_and_Sil~ FALSE
##  8     8          13.7           13.0  rivers             6.7 White          FALSE
##  9     9          13.2            5.22 lakes              7.6 Black_and_O~ FALSE
## 10    10          19.0           15.5  rivers             6.7 Calico         TRUE
## # i 1,990 more rows
## # i 1 more variable: life_span <dbl>
```

The first, second and third, fifth, eighth columns are numerical, fourth and sixth columns are character and seventh colum is logical

The dimension of the dataset is 2000x8. This means the dataset has 2000 rows and 8 columns

```
fish_data[ , 'ph_of_water' ]
```

Selecting a column using square brackets

```
## # A tibble: 2,000 x 1
##    ph_of_water
##          <dbl>
##  1         6.2
##  2         6.8
##  3         7.9
##  4         6.7
##  5         7.8
##  6         7.8
##  7         6.3
##  8         6.7
##  9         7.6
## 10         6.7
## # i 1,990 more rows
```

```
fish_data[ fish_data$ph_of_water > 7 , 'ph_of_water']
```

Selecting a column using logical statements

```
## # A tibble: 969 x 1
##    ph_of_water
##          <dbl>
##  1         7.9
##  2         7.8
##  3         7.8
##  4         7.6
##  5         7.2
##  6         7.6
##  7         7.9
##  8         7.3
##  9         7.1
## 10         7.8
## # i 959 more rows
```

```
summary(fish_data)
```

summary of data

```
##        id          average_length   average_weight     habitat
##  Min.   :   1.0   Min.   : 1.000   Min.   : 2.000   Length:2000
##  1st Qu.: 500.8   1st Qu.: 5.857   1st Qu.: 6.138   Class :character
##  Median :1000.5   Median :10.660   Median :10.455   Mode  :character
##  Mean   :1000.5   Mean   :10.557   Mean   :10.449
##  3rd Qu.:1500.2   3rd Qu.:15.172   3rd Qu.:14.665
##  Max.   :2000.0   Max.   :20.000   Max.   :18.960
##   ph_of_water        color               Gender          life_span
##  Min.   :6.000   Length:2000        Mode :logical   Min.   : 1.00
##  1st Qu.:6.500   Class :character   FALSE:1007      1st Qu.: 7.80
##  Median :7.000   Mode  :character   TRUE :969       Median :14.40
##  Mean   :7.015                      NA's :24        Mean   :14.37
##  3rd Qu.:7.500                                      3rd Qu.:20.90
##  Max.   :8.000                                      Max.   :28.00
```

Summery of this data gives a simple statistics of each column.The statistics includes Max, Median, Mean, Min, 1st Quartile and 3rd quartile of 1st, 2nd, 3rd, 5th, 8th colums; length, class, mode of 4th and 6th colums; mode, false true and NA's of 7th colum.

```r
fish_data$double_ph_of_water = fish_data$ph_of_water * 2
head(fish_data)
```

Calculation of data with adding a column

```
## # A tibble: 6 x 9
##      id average_length average_weight habitat ph_of_water color Gender life_span
##   <dbl>          <dbl>          <dbl> <chr>         <dbl> <chr> <lgl>      <dbl>
## 1     1          14.7            5.87 ponds           6.2 Redd~ FALSE       10.9
## 2     2           1.32           3.86 idlewa~         6.8 Cali~ TRUE         5.2
## 3     3          14.2           12.1  lakes           7.9 Redd~ TRUE        25.3
## 4     4           2.54           3.2  rivers          6.7 White FALSE       16.4
## 5     5          13.1            9.81 lakes           7.8 Oran~ TRUE         3.2
## 6     6          15.2            8.99 lakes           7.8 White FALSE       21.6
## # i 1 more variable: double_ph_of_water <dbl>
```

```r
aggregate(ph_of_water ~ habitat, data = fish_data, FUN = mean)
```

Use base R to aggregate data

```
##            habitat ph_of_water
## 1       idlewater    6.983117
## 2           lakes    7.014115
## 3           ponds    7.039163
## 4          rivers    7.032405
## 5 slowmovingwaters    7.004545
```

```r
aggregate( average_length ~ habitat, data = fish_data, FUN = mean)
```

```
##            habitat average_length
## 1        idlewater       10.40330
## 2            lakes       10.64957
## 3            ponds       10.44638
## 4           rivers       11.21332
## 5 slowmovingwaters       10.06803
```

```r
aggregate( average_length ~ habitat, data = fish_data, FUN = sum)
```

```
##            habitat average_length
## 1        idlewater        4005.27
## 2            lakes        4451.52
## 3            ponds        4241.23
## 4           rivers        4429.26
## 5 slowmovingwaters        3986.94
```

```r
aggregate( average_length ~ habitat, data = fish_data, FUN = max)
```

```
##            habitat average_length
## 1        idlewater          19.96
## 2            lakes          20.00
## 3            ponds          19.97
## 4           rivers          20.00
## 5 slowmovingwaters          19.95
```

```r
aggregate( average_length ~ habitat, data = fish_data, FUN = min)
```

```
##            habitat average_length
## 1        idlewater           1.00
## 2            lakes           1.00
## 3            ponds           1.07
## 4           rivers           1.03
## 5 slowmovingwaters           1.01
```