Data Enrichment (Lab-10 / Part-2)

Bibek Sapkota

Data Enrichment

. ## Deduplication

task1 :Import the library

```
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v readr     2.1.5
## v forcats   1.0.0     v stringr   1.5.1
## v ggplot2   3.5.0     v tibble    3.2.1
## v lubridate 1.9.3     v tidyr     1.3.1
## v purrr     1.0.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become error
```

task 2:create a vector of numbers

```
x <- c(1, 1, 4, 5, 4, 6)
```

task 3: Checking the positions of duplicate elements

```
duplicated(x)
```

```
## [1] FALSE  TRUE FALSE FALSE  TRUE FALSE
```

task 4:Extracting duplicate elements

```
x[duplicated(x)]
```

```
## [1] 1 4
```

task 5: remove duplicated elements, use !duplicated(), where ! is a logical negation.

```
x[!duplicated(x)]
```

```
## [1] 1 4 5 6
```

task 6:Loading Inbulit iris dataset

```r
iris[!duplicated(iris$Sepal.Width), ]
```

```
##    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
## 9           4.4         2.9          1.4         0.2     setosa
## 11          5.4         3.7          1.5         0.2     setosa
## 15          5.8         4.0          1.2         0.2     setosa
## 16          5.7         4.4          1.5         0.4     setosa
## 19          5.7         3.8          1.7         0.3     setosa
## 24          5.1         3.3          1.7         0.5     setosa
## 33          5.2         4.1          1.5         0.1     setosa
## 34          5.5         4.2          1.4         0.2     setosa
## 42          4.5         2.3          1.3         0.3     setosa
## 55          6.5         2.8          4.6         1.5 versicolor
## 58          4.9         2.4          3.3         1.0 versicolor
## 60          5.2         2.7          3.9         1.4 versicolor
## 61          5.0         2.0          3.5         1.0 versicolor
## 63          6.0         2.2          4.0         1.0 versicolor
## 70          5.6         2.5          3.9         1.1 versicolor
## 80          5.7         2.6          3.5         1.0 versicolor
```

Remove duplicate rows in a data frame

task 1: Remove duplicate rows based on all columns:

```r
iris %>% distinct()
```

```
##     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
## 143          6.8         3.2          5.9         2.3  virginica
## 144          6.7         3.3          5.7         2.5  virginica
## 145          6.7         3.0          5.2         2.3  virginica
## 146          6.3         2.5          5.0         1.9  virginica
## 147          6.5         3.0          5.2         2.0  virginica
## 148          6.2         3.4          5.4         2.3  virginica
## 149          5.9         3.0          5.1         1.8  virginica
```

task 2: Remove duplicated rows based on Sepal.Length

```
iris %>% distinct(Sepal.Length, .keep_all = TRUE)
```

```
##    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           4.4         2.9          1.4         0.2     setosa
## 8           4.8         3.4          1.6         0.2     setosa
## 9           4.3         3.0          1.1         0.1     setosa
## 10          5.8         4.0          1.2         0.2     setosa
## 11          5.7         4.4          1.5         0.4     setosa
## 12          5.2         3.5          1.5         0.2     setosa
## 13          5.5         4.2          1.4         0.2     setosa
## 14          4.5         2.3          1.3         0.3     setosa
## 15          5.3         3.7          1.5         0.2     setosa
## 16          7.0         3.2          4.7         1.4 versicolor
## 17          6.4         3.2          4.5         1.5 versicolor
## 18          6.9         3.1          4.9         1.5 versicolor
## 19          6.5         2.8          4.6         1.5 versicolor
## 20          6.3         3.3          4.7         1.6 versicolor
## 21          6.6         2.9          4.6         1.3 versicolor
## 22          5.9         3.0          4.2         1.5 versicolor
## 23          6.0         2.2          4.0         1.0 versicolor
## 24          6.1         2.9          4.7         1.4 versicolor
```

```
## 25          5.6         2.9          3.6         1.3 versicolor
## 26          6.7         3.1          4.4         1.4 versicolor
## 27          6.2         2.2          4.5         1.5 versicolor
## 28          6.8         2.8          4.8         1.4 versicolor
## 29          7.1         3.0          5.9         2.1  virginica
## 30          7.6         3.0          6.6         2.1  virginica
## 31          7.3         2.9          6.3         1.8  virginica
## 32          7.2         3.6          6.1         2.5  virginica
## 33          7.7         3.8          6.7         2.2  virginica
## 34          7.4         2.8          6.1         1.9  virginica
## 35          7.9         3.8          6.4         2.0  virginica
```

task 3:Remove duplicated rows based on Sepal.Length and Petal.Width

```
iris %>% distinct(Sepal.Length, Petal.Width, .keep_all = TRUE)
```

```
##     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            4.4         2.9          1.4         0.2     setosa
## 9            4.9         3.1          1.5         0.1     setosa
## 10           5.4         3.7          1.5         0.2     setosa
## 11           4.8         3.4          1.6         0.2     setosa
## 12           4.8         3.0          1.4         0.1     setosa
## 13           4.3         3.0          1.1         0.1     setosa
## 14           5.8         4.0          1.2         0.2     setosa
## 15           5.7         4.4          1.5         0.4     setosa
## 16           5.1         3.5          1.4         0.3     setosa
## 17           5.7         3.8          1.7         0.3     setosa
## 18           5.1         3.7          1.5         0.4     setosa
## 19           5.1         3.3          1.7         0.5     setosa
## 20           5.0         3.4          1.6         0.4     setosa
## 21           5.2         3.5          1.5         0.2     setosa
## 22           5.2         4.1          1.5         0.1     setosa
## 23           5.5         4.2          1.4         0.2     setosa
## 24           5.0         3.5          1.3         0.3     setosa
## 25           4.5         2.3          1.3         0.3     setosa
## 26           5.0         3.5          1.6         0.6     setosa
## 27           4.8         3.0          1.4         0.3     setosa
## 28           5.3         3.7          1.5         0.2     setosa
## 29           7.0         3.2          4.7         1.4 versicolor
## 30           6.4         3.2          4.5         1.5 versicolor
## 31           6.9         3.1          4.9         1.5 versicolor
## 32           5.5         2.3          4.0         1.3 versicolor
## 33           6.5         2.8          4.6         1.5 versicolor
## 34           5.7         2.8          4.5         1.3 versicolor
## 35           6.3         3.3          4.7         1.6 versicolor
## 36           4.9         2.4          3.3         1.0 versicolor
```

```
## 37          6.6         2.9          4.6         1.3 versicolor
## 38          5.2         2.7          3.9         1.4 versicolor
## 39          5.0         2.0          3.5         1.0 versicolor
## 40          5.9         3.0          4.2         1.5 versicolor
## 41          6.0         2.2          4.0         1.0 versicolor
## 42          6.1         2.9          4.7         1.4 versicolor
## 43          5.6         2.9          3.6         1.3 versicolor
## 44          6.7         3.1          4.4         1.4 versicolor
## 45          5.6         3.0          4.5         1.5 versicolor
## 46          5.8         2.7          4.1         1.0 versicolor
## 47          6.2         2.2          4.5         1.5 versicolor
## 48          5.6         2.5          3.9         1.1 versicolor
## 49          5.9         3.2          4.8         1.8 versicolor
## 50          6.1         2.8          4.0         1.3 versicolor
## 51          6.3         2.5          4.9         1.5 versicolor
## 52          6.1         2.8          4.7         1.2 versicolor
## 53          6.4         2.9          4.3         1.3 versicolor
## 54          6.6         3.0          4.4         1.4 versicolor
## 55          6.8         2.8          4.8         1.4 versicolor
## 56          6.7         3.0          5.0         1.7 versicolor
## 57          6.0         2.9          4.5         1.5 versicolor
## 58          5.7         2.6          3.5         1.0 versicolor
## 59          5.5         2.4          3.8         1.1 versicolor
## 60          5.5         2.4          3.7         1.0 versicolor
## 61          5.8         2.7          3.9         1.2 versicolor
## 62          6.0         2.7          5.1         1.6 versicolor
## 63          5.4         3.0          4.5         1.5 versicolor
## 64          6.7         3.1          4.7         1.5 versicolor
## 65          6.3         2.3          4.4         1.3 versicolor
## 66          5.5         2.6          4.4         1.2 versicolor
## 67          5.7         3.0          4.2         1.2 versicolor
## 68          6.2         2.9          4.3         1.3 versicolor
## 69          5.1         2.5          3.0         1.1 versicolor
## 70          6.3         3.3          6.0         2.5  virginica
## 71          5.8         2.7          5.1         1.9  virginica
## 72          7.1         3.0          5.9         2.1  virginica
## 73          6.3         2.9          5.6         1.8  virginica
## 74          6.5         3.0          5.8         2.2  virginica
## 75          7.6         3.0          6.6         2.1  virginica
## 76          4.9         2.5          4.5         1.7  virginica
## 77          7.3         2.9          6.3         1.8  virginica
## 78          6.7         2.5          5.8         1.8  virginica
## 79          7.2         3.6          6.1         2.5  virginica
## 80          6.5         3.2          5.1         2.0  virginica
## 81          6.4         2.7          5.3         1.9  virginica
## 82          6.8         3.0          5.5         2.1  virginica
## 83          5.7         2.5          5.0         2.0  virginica
## 84          5.8         2.8          5.1         2.4  virginica
## 85          6.4         3.2          5.3         2.3  virginica
## 86          6.5         3.0          5.5         1.8  virginica
## 87          7.7         3.8          6.7         2.2  virginica
## 88          7.7         2.6          6.9         2.3  virginica
## 89          6.9         3.2          5.7         2.3  virginica
## 90          5.6         2.8          4.9         2.0  virginica
```

```
## 91          7.7         2.8          6.7         2.0  virginica
## 92          6.7         3.3          5.7         2.1  virginica
## 93          7.2         3.2          6.0         1.8  virginica
## 94          6.2         2.8          4.8         1.8  virginica
## 95          6.1         3.0          4.9         1.8  virginica
## 96          6.4         2.8          5.6         2.1  virginica
## 97          7.2         3.0          5.8         1.6  virginica
## 98          7.4         2.8          6.1         1.9  virginica
## 99          7.9         3.8          6.4         2.0  virginica
## 100         6.4         2.8          5.6         2.2  virginica
## 101         6.3         3.4          5.6         2.4  virginica
## 102         6.4         3.1          5.5         1.8  virginica
## 103         6.0         3.0          4.8         1.8  virginica
## 104         6.9         3.1          5.4         2.1  virginica
## 105         6.7         3.1          5.6         2.4  virginica
## 106         6.8         3.2          5.9         2.3  virginica
## 107         6.7         3.3          5.7         2.5  virginica
## 108         6.7         3.0          5.2         2.3  virginica
## 109         6.3         2.5          5.0         1.9  virginica
## 110         6.2         3.4          5.4         2.3  virginica
```

Grouping

```r
# First, we have to define which column we are going to use to group.
# In this case we use the species of the plant
my_group <- group_by(iris, Species)
# Now using "summarize_all", we define which function we use to group the values.
# For an example we can use mean.
summarize_all(my_group, funs(mean))
```

```
## Warning: `funs()` was deprecated in dplyr 0.8.0.
## i Please use a list of either functions or lambdas:
##
## # Simple named list: list(mean = mean, median = median)
##
## # Auto named with `tibble::lst()`: tibble::lst(mean, median)
##
## # Using lambdas list(~ mean(., trim = .2), ~ median(., na.rm = TRUE))
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
## generated.
```

```
## # A tibble: 3 x 5
##   Species    Sepal.Length Sepal.Width Petal.Length Petal.Width
##   <fct>             <dbl>       <dbl>        <dbl>       <dbl>
## 1 setosa             5.01        3.43         1.46       0.246
## 2 versicolor         5.94        2.77         4.26       1.33
## 3 virginica          6.59        2.97         5.55       2.03
```

task 2: Download and load the data set.

```
surveys <- read.csv("survey.csv")
surveys
```

```
##    record_id       date plot species sex wgt
## 1      35525 31/12/2002    9      OL   M  26
## 2      25749 10/05/1997    7      PM   F  24
## 3      25848 11/05/1997   13      PP   M  18
## 4      25956 09/06/1997    1      OL   M  16
## 5      26012 09/06/1997    2      PB   F  24
## 6      26068 10/06/1997    8      OT   M  28
## 7      26255 09/07/1997   17      PP   M  18
## 8      26373 09/07/1997   16      PP   M  13
## 9      26562 29/07/1997   24      PE   F  22
## 10     26690 30/07/1997    8      PF   F   7
## 11     26948 28/09/1997   11      DO   F  51
## 12     27118 26/10/1997    6      PE   M  20
## 13     27444 31/01/1998    2      DM   F  41
```

task 3:

```
surveys %>%
group_by(sex) %>%  summarize(mean_weight = mean(wgt, na.rm = TRUE))
```

```
## # A tibble: 2 x 2
##   sex   mean_weight
##   <chr>       <dbl>
## 1 F            28.2
## 2 M            19.9
```

task 4: group by multiple columns

```
surveys %>%
group_by(sex, species) %>%  summarize(mean_weight = mean(wgt, na.rm = TRUE))
```

```
## `summarise()` has grouped output by 'sex'. You can override using the `.groups`
## argument.
```

```
## # A tibble: 10 x 3
## # Groups:   sex [2]
##    sex   species mean_weight
##    <chr> <chr>         <dbl>
##  1 F     DM             41
##  2 F     DO             51
##  3 F     PB             24
##  4 F     PE             22
##  5 F     PF              7
##  6 F     PM             24
##  7 M     OL             21
##  8 M     OT             28
##  9 M     PE             20
## 10 M     PP             16.3
```

```
surveys %>%
group_by(sex, species) %>%  summarize(mean_weight = mean(wgt, na.rm = TRUE),
min_weight = min(wgt, na.rm = TRUE))
```

```
## `summarise()` has grouped output by 'sex'. You can override using the `.groups`
## argument.
```

```
## # A tibble: 10 x 4
## # Groups:   sex [2]
##    sex   species mean_weight min_weight
##    <chr> <chr>         <dbl>      <int>
##  1 F     DM             41           41
##  2 F     DO             51           51
##  3 F     PB             24           24
##  4 F     PE             22           22
##  5 F     PF              7            7
##  6 F     PM             24           24
##  7 M     OL             21           16
##  8 M     OT             28           28
##  9 M     PE             20           20
## 10 M     PP             16.3         13
```