data_clean

gym

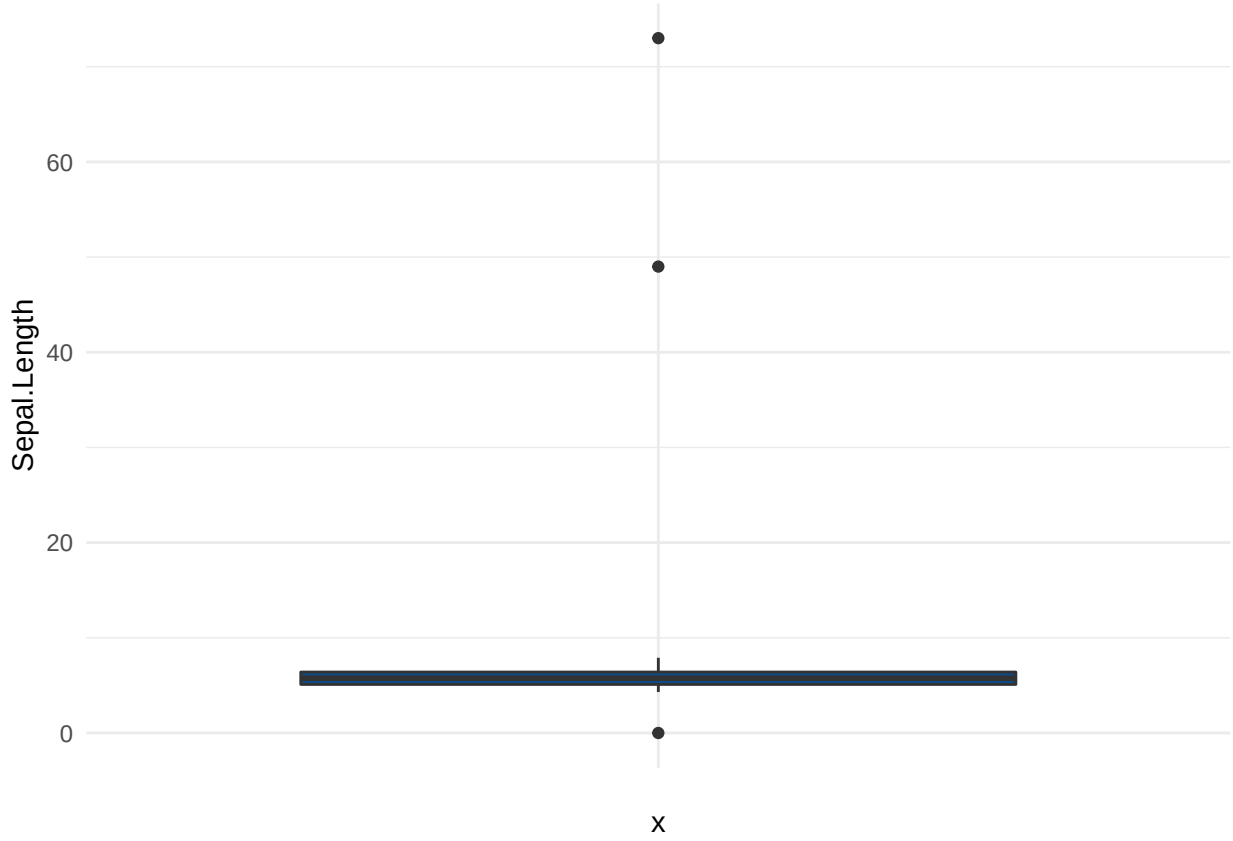
2020/12/24

data_clean_homework

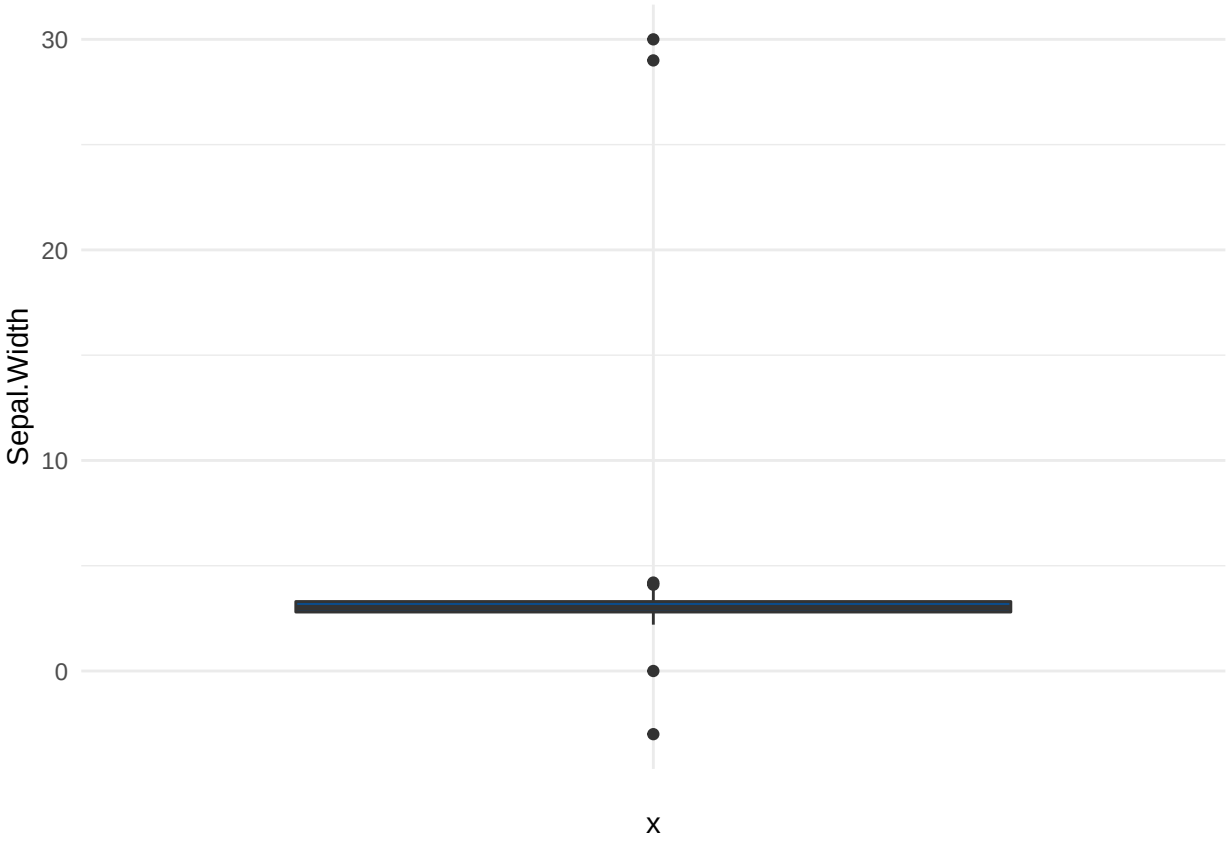
exercise1-reading and manually checking

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union

## Loading required package: igraph

##
## Attaching package: 'igraph'

## The following objects are masked from 'package:dplyr':
##
##     as_data_frame, groups, union

## The following objects are masked from 'package:stats':
##
##     decompose, spectrum

## The following object is masked from 'package:base':
##
##     union

##
## Attaching package: 'editrules'

## The following objects are masked from 'package:igraph':
##
##     blocks, normalize

## The following object is masked from 'package:dplyr':
##
##     contains

## Loading required package: colorspace

## Loading required package: grid

## VIM is ready to use.

## Suggestions and bug-reports can be submitted at: https://github.com/statistikat/VIM/issues
```

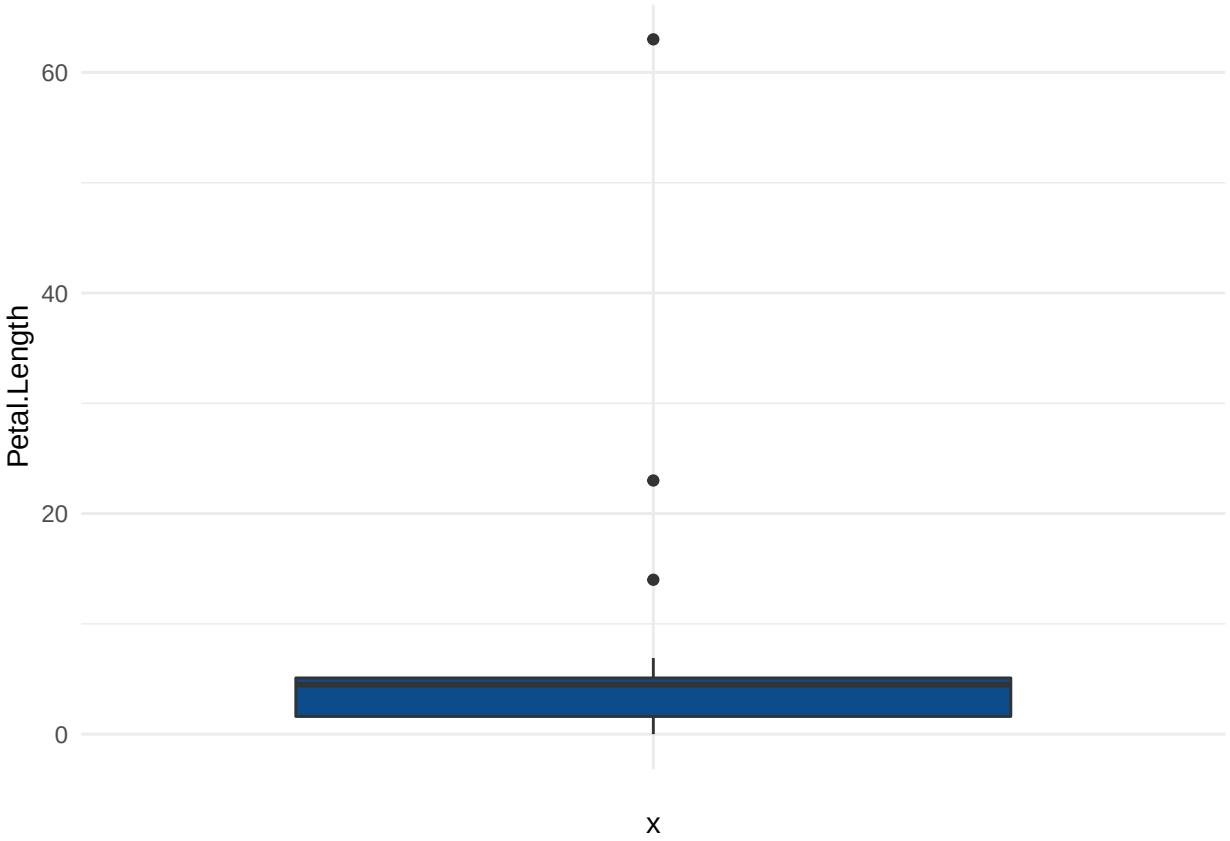
```
##
## Attaching package: 'VIM'
## The following object is masked from 'package:datasets':
##
##     sleep
## [1] 96
## [1] 0.64
## [1] "Sepal.Length" "Sepal.Width"  "Petal.Length" "Petal.Width"  "Species"
## Warning: Removed 10 rows containing non-finite values (stat_boxplot).
```



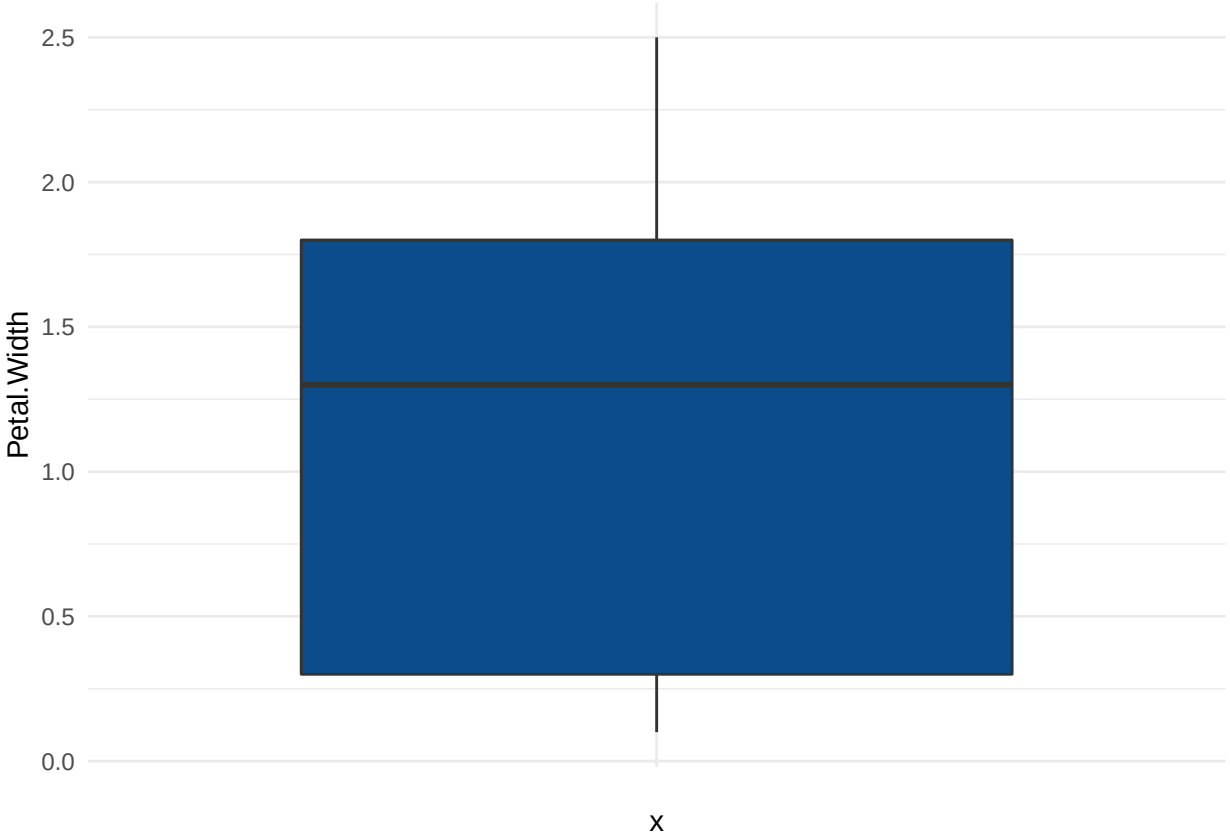

```
## Warning: Removed 17 rows containing non-finite values (stat_boxplot).
```




```
## Warning: Removed 19 rows containing non-finite values (stat_boxplot).
```




```
## Warning: Removed 13 rows containing non-finite values (stat_boxplot).
```



exercise2-rules

```
rules <- editfile('R_test.txt', type = 'all')
rules
```

```
##
## Edit set:
## NULL :
```

exercise3-correcting

```
names(dirty_data)
```

```
## [1] "Sepal.Length" "Sepal.Width"  "Petal.Length" "Petal.Width"  "Species"
```

```
#help(package = "deducorrect")
u <- correctionRules(expression(
  if ( is.na(Petal.Width) ) Petal.Width <- Inf,
  if ( Petal.Width == 'Inf' ) Petal.Width <- NA
))
correctWithRules(u,dirty_data)
```

```
## $corrected
##     Sepal.Length Sepal.Width Petal.Length Petal.Width    Species
## 1            6.4         3.2        4.500         1.5 versicolor
## 2            6.3         3.3        6.000         2.5  virginica
## 3            6.2          NA        5.400         2.3  virginica
```

```
## 4           5.0         3.4        1.600         0.4     setosa
## 5           5.7         2.6        3.500         1.0 versicolor
## 6           5.3          NA           NA         0.2     setosa
## 7           6.4         2.7        5.300          NA  virginica
## 8           5.9         3.0        5.100         1.8  virginica
## 9           5.8         2.7        4.100         1.0 versicolor
## 10          4.8         3.1        1.600         0.2     setosa
## 11          5.0         3.5        1.600         0.6     setosa
## 12          6.0         2.7        5.100         1.6 versicolor
## 13          6.0         3.0        4.800          NA  virginica
## 14          6.8         2.8        4.800         1.4 versicolor
## 15           NA         3.9        1.700         0.4     setosa
## 16          5.0          NA        3.500         1.0 versicolor
## 17          5.5          NA        4.000         1.3 versicolor
## 18          4.7         3.2        1.300         0.2     setosa
## 19           NA         4.0           NA         0.2     setosa
## 20          5.6          NA        4.200         1.3 versicolor
## 21          4.9         3.6           NA         0.1     setosa
## 22          5.4          NA        4.500         1.5 versicolor
## 23          6.2         2.8           NA         1.8  virginica
## 24          6.7         3.3        5.700         2.5  virginica
## 25           NA         3.0        5.900         2.1  virginica
## 26          4.6         3.2        1.400         0.2     setosa
## 27          4.9         3.1        1.500         0.1     setosa
## 28           NA          NA           NA          NA  virginica
## 29          6.5         3.2        5.100         2.0  virginica
## 30           NA         2.8        0.820         1.3 versicolor
## 31          4.4         3.2           NA         0.2     setosa
## 32          5.9         3.2        4.800          NA versicolor
## 33          5.7         2.8        4.500         1.3 versicolor
## 34          6.2         2.9           NA         1.3 versicolor
## 35          6.6         2.9           NA         1.3 versicolor
## 36          4.8         3.0        1.400         0.1     setosa
## 37          6.5         3.0        5.500         1.8  virginica
## 38          6.2         2.2        4.500         1.5 versicolor
## 39          6.7         2.5        5.800         1.8  virginica
## 40          5.0         3.0        1.600         0.2     setosa
## 41          5.0          NA        1.200         0.2     setosa
## 42          5.8         2.7        3.900         1.2 versicolor
## 43           NA          NA        1.300         0.4     setosa
## 44          5.8         2.7        5.100         1.9  virginica
## 45          5.5         4.2        1.400         0.2     setosa
## 46          7.7         2.8        6.700         2.0  virginica
## 47          5.7          NA           NA         0.4     setosa
## 48          7.0         3.2        4.700         1.4 versicolor
## 49          6.5         3.0        5.800         2.2  virginica
## 50          6.0         3.4        4.500         1.6 versicolor
## 51          5.5         2.6        4.400         1.2 versicolor
## 52          4.9         3.1           NA         0.2     setosa
## 53          5.2         2.7        3.900         1.4 versicolor
## 54          4.8         3.4        1.600         0.2     setosa
## 55          6.3         3.3        4.700         1.6 versicolor
## 56          7.7         3.8        6.700         2.2  virginica
## 57          5.1         3.8        1.500         0.3     setosa
```

```
## 58           NA        2.9        4.500        1.5 versicolor
## 59          6.4        2.8        5.600         NA  virginica
## 60          6.4        2.8        5.600        2.1  virginica
## 61          5.0        2.3        3.300         NA versicolor
## 62          7.4        2.8        6.100        1.9  virginica
## 63          4.3        3.0        1.100        0.1     setosa
## 64          5.0        3.3        1.400        0.2     setosa
## 65          7.2        3.0        5.800        1.6  virginica
## 66          6.3        2.5        4.900        1.5 versicolor
## 67          5.1        2.5           NA        1.1 versicolor
## 68           NA        3.2        5.700        2.3  virginica
## 69          5.1        3.5           NA         NA     setosa
## 70          5.0        3.5        1.300        0.3     setosa
## 71          6.1        3.0        4.600        1.4 versicolor
## 72          6.9        3.1        5.100        2.3  virginica
## 73          5.1        3.5        1.400        0.3     setosa
## 74          6.5         NA        4.600        1.5 versicolor
## 75          5.6        2.8        4.900        2.0  virginica
## 76          4.9        2.5        4.500         NA  virginica
## 77          5.5        3.5        1.300        0.2     setosa
## 78          7.6        3.0        6.600        2.1  virginica
## 79          5.1        3.8           NA        0.2     setosa
## 80          7.9        3.8        6.400        2.0  virginica
## 81          6.1        2.6        5.600        1.4  virginica
## 82          5.4        3.4        1.700        0.2     setosa
## 83          6.1        2.9        4.700        1.4 versicolor
## 84          5.4        3.7        1.500        0.2     setosa
## 85          6.7        3.0        5.200        2.3  virginica
## 86          5.1        3.8        1.900         NA     setosa
## 87          6.4        2.9        4.300        1.3 versicolor
## 88          5.7        2.9        4.200        1.3 versicolor
## 89          4.4        2.9        1.400        0.2     setosa
## 90          6.3        2.5        5.000        1.9  virginica
## 91          7.2        3.2        6.000        1.8  virginica
## 92          4.9         NA        3.300        1.0 versicolor
## 93          5.2        3.4        1.400        0.2     setosa
## 94          5.8        2.7        5.100        1.9  virginica
## 95          6.0        2.2        5.000        1.5  virginica
## 96          6.9        3.1           NA        1.5 versicolor
## 97          5.5        2.3        4.000        1.3 versicolor
## 98          6.7         NA        5.000        1.7 versicolor
## 99          5.7        3.0        4.200        1.2 versicolor
## 100         6.3        2.8        5.100        1.5  virginica
## 101         5.4        3.4        1.500        0.4     setosa
## 102         7.2        3.6           NA        2.5  virginica
## 103         6.3        2.7        4.900         NA  virginica
## 104         5.6        3.0        4.100        1.3 versicolor
## 105         5.1        3.7           NA        0.4     setosa
## 106         5.5         NA        0.925        1.0 versicolor
## 107         6.5        3.0        5.200        2.0  virginica
## 108         4.8        3.0        1.400         NA     setosa
## 109         6.1        2.8           NA        1.3 versicolor
## 110         4.6        3.4        1.400        0.3     setosa
## 111         6.3        3.4           NA        2.4  virginica
```

```
## 112          5.0         3.4        1.500         0.2     setosa
## 113          5.1         3.4        1.500         0.2     setosa
## 114           NA         3.3        5.700         2.1  virginica
## 115          6.7         3.1        4.700         1.5 versicolor
## 116          7.7         2.6        6.900         2.3  virginica
## 117          6.3          NA        4.400         1.3 versicolor
## 118          4.6         3.1        1.500         0.2     setosa
## 119           NA         3.0        5.500         2.1  virginica
## 120           NA         2.8        4.700         1.2 versicolor
## 121          5.9         3.0           NA         1.5 versicolor
## 122          4.5         2.3        1.300         0.3     setosa
## 123          6.4         3.2        5.300         2.3  virginica
## 124          5.2         4.1        1.500         0.1     setosa
## 125           NA          NA           NA         2.0     setosa
## 126          5.6         2.9        3.600         1.3 versicolor
## 127          6.8         3.2        5.900         2.3  virginica
## 128          5.8          NA        5.100         2.4  virginica
## 129          4.6         3.6           NA         0.2     setosa
## 130          5.7          NA        1.700         0.3     setosa
## 131          5.6         2.5        3.900         1.1 versicolor
## 132          6.7         3.1        4.400         1.4 versicolor
## 133          4.8          NA        1.900         0.2     setosa
## 134          5.1         3.3        1.700         0.5     setosa
## 135          4.4         3.0        1.300          NA     setosa
## 136          7.7         3.0           NA         2.3  virginica
## 137          4.7         3.2        1.600         0.2     setosa
## 138           NA         3.0        4.900         1.8  virginica
## 139          6.9         3.1        5.400         2.1  virginica
## 140          6.0         2.2        4.000         1.0 versicolor
## 141          5.0          NA        1.400         0.2     setosa
## 142          5.5          NA        3.800         1.1 versicolor
## 143          6.6         3.0        4.400         1.4 versicolor
## 144          6.3         2.9        5.600         1.8  virginica
## 145          5.7         2.5        5.000         2.0  virginica
## 146          6.7         3.1        5.600         2.4  virginica
## 147          5.6         3.0        4.500         1.5 versicolor
## 148          5.2         3.5        1.500         0.2     setosa
## 149          6.4         3.1           NA         1.8  virginica
## 150          5.8         2.6        4.000          NA versicolor
##
## $corrections
##    row    variable old new                                          how
## 1    7 Petal.Width  NA Inf  if (is.na(Petal.Width)) Petal.Width <- Inf
## 2    7 Petal.Width Inf  NA if (Petal.Width == "Inf") Petal.Width <- NA
## 3   13 Petal.Width  NA Inf  if (is.na(Petal.Width)) Petal.Width <- Inf
## 4   13 Petal.Width Inf  NA if (Petal.Width == "Inf") Petal.Width <- NA
## 5   28 Petal.Width  NA Inf  if (is.na(Petal.Width)) Petal.Width <- Inf
## 6   28 Petal.Width Inf  NA if (Petal.Width == "Inf") Petal.Width <- NA
## 7   32 Petal.Width  NA Inf  if (is.na(Petal.Width)) Petal.Width <- Inf
## 8   32 Petal.Width Inf  NA if (Petal.Width == "Inf") Petal.Width <- NA
## 9   59 Petal.Width  NA Inf  if (is.na(Petal.Width)) Petal.Width <- Inf
## 10  59 Petal.Width Inf  NA if (Petal.Width == "Inf") Petal.Width <- NA
## 11  61 Petal.Width  NA Inf  if (is.na(Petal.Width)) Petal.Width <- Inf
## 12  61 Petal.Width Inf  NA if (Petal.Width == "Inf") Petal.Width <- NA
```

```
## 13  69 Petal.Width  NA Inf  if (is.na(Petal.Width)) Petal.Width <- Inf
## 14  69 Petal.Width Inf  NA if (Petal.Width == "Inf") Petal.Width <- NA
## 15  76 Petal.Width  NA Inf  if (is.na(Petal.Width)) Petal.Width <- Inf
## 16  76 Petal.Width Inf  NA if (Petal.Width == "Inf") Petal.Width <- NA
## 17  86 Petal.Width  NA Inf  if (is.na(Petal.Width)) Petal.Width <- Inf
## 18  86 Petal.Width Inf  NA if (Petal.Width == "Inf") Petal.Width <- NA
## 19 103 Petal.Width  NA Inf  if (is.na(Petal.Width)) Petal.Width <- Inf
## 20 103 Petal.Width Inf  NA if (Petal.Width == "Inf") Petal.Width <- NA
## 21 108 Petal.Width  NA Inf  if (is.na(Petal.Width)) Petal.Width <- Inf
## 22 108 Petal.Width Inf  NA if (Petal.Width == "Inf") Petal.Width <- NA
## 23 135 Petal.Width  NA Inf  if (is.na(Petal.Width)) Petal.Width <- Inf
## 24 135 Petal.Width Inf  NA if (Petal.Width == "Inf") Petal.Width <- NA
## 25 150 Petal.Width  NA Inf  if (is.na(Petal.Width)) Petal.Width <- Inf
## 26 150 Petal.Width Inf  NA if (Petal.Width == "Inf") Petal.Width <- NA
```

exercise4-Imputing

```
cleaned_data <- kNN(dirty_data)
View(cleaned_data)
```