DATA With dprly 60

June 7, 2024

[1]:
```
# Check if you need to install the `tidyverse` library
require("tidyverse")
library(tidyverse)
```

```
Loading required package: tidyverse
Warning message:
"replacing previous import ‘lifecycle::last_warnings’ by ‘rlang::last_warnings’
when loading ‘tibble’"Warning message:
"replacing previous import ‘ellipsis::check_dots_unnamed’ by
‘rlang::check_dots_unnamed’ when loading ‘tibble’"Warning message:
"replacing previous import ‘ellipsis::check_dots_used’ by
‘rlang::check_dots_used’ when loading ‘tibble’"Warning message:
"replacing previous import ‘ellipsis::check_dots_empty’ by
‘rlang::check_dots_empty’ when loading ‘tibble’"  Attaching packages
                     tidyverse 1.3.0
 ggplot2 3.3.0      purrr   0.3.4
 tibble  3.0.1      dplyr   0.8.5
 tidyr   1.0.2      stringr 1.4.0
 readr   1.3.1      forcats 0.5.0
  Conflicts                         tidyverse_conflicts()
 dplyr::filter() masks stats::filter()
 dplyr::lag()    masks stats::lag()
```

[2]:
```
library(readr)
```

[3]:
```
#load the bike-sharing system data from the csv processed in the previous lab
seoul_bike_sharing <- read_csv("raw_seoul_bike_sharing.csv")
```

```
Parsed with column specification:
cols(
  DATE = col_character(),
  RENTED_BIKE_COUNT = col_double(),
  HOUR = col_double(),
  TEMPERATURE = col_double(),
  HUMIDITY = col_double(),
  WIND_SPEED = col_double(),
  VISIBILITY = col_double(),
  DEW_POINT_TEMPERATURE = col_double(),
```

```
  SOLAR_RADIATION = col_double(),
  RAINFALL = col_double(),
  SNOWFALL = col_double(),
  SEASONS = col_character(),
  HOLIDAY = col_character(),
  FUNCTIONING_DAY = col_character()
)
```

[4]:
```
summary(seoul_bike_sharing)
```

```
      DATE           RENTED_BIKE_COUNT      HOUR        TEMPERATURE
 Length:8760        Min.   :   2.0    Min.   : 0.00   Min.   :-17.80
 Class :character   1st Qu.: 214.0    1st Qu.: 5.75   1st Qu.:  3.40
 Mode  :character   Median : 542.0    Median :11.50   Median : 13.70
                    Mean   : 729.2    Mean   :11.50   Mean   : 12.87
                    3rd Qu.:1084.0    3rd Qu.:17.25   3rd Qu.: 22.50
                    Max.   :3556.0    Max.   :23.00   Max.   : 39.40
                    NA's   :295                       NA's   :11
    HUMIDITY       WIND_SPEED      VISIBILITY   DEW_POINT_TEMPERATURE
 Min.   : 0.00   Min.   :0.000   Min.   :  27   Min.   :-30.600
 1st Qu.:42.00   1st Qu.:0.900   1st Qu.: 940   1st Qu.: -4.700
 Median :57.00   Median :1.500   Median :1698   Median :  5.100
 Mean   :58.23   Mean   :1.725   Mean   :1437   Mean   :  4.074
 3rd Qu.:74.00   3rd Qu.:2.300   3rd Qu.:2000   3rd Qu.: 14.800
 Max.   :98.00   Max.   :7.400   Max.   :2000   Max.   : 27.200

 SOLAR_RADIATION     RAINFALL          SNOWFALL         SEASONS
 Min.   :0.0000   Min.   : 0.0000   Min.   :0.00000   Length:8760
 1st Qu.:0.0000   1st Qu.: 0.0000   1st Qu.:0.00000   Class :character
 Median :0.0100   Median : 0.0000   Median :0.00000   Mode  :character
 Mean   :0.5691   Mean   : 0.1487   Mean   :0.07507
 3rd Qu.:0.9300   3rd Qu.: 0.0000   3rd Qu.:0.00000
 Max.   :3.5200   Max.   :35.0000   Max.   :8.80000

   HOLIDAY          FUNCTIONING_DAY
 Length:8760        Length:8760
 Class :character   Class :character
 Mode  :character   Mode  :character
```

[5]:
```
#take a quick look at the dataset
dim(seoul_bike_sharing)
```

1. 8760 2. 14

[6]:
```
colSums(is.na(seoul_bike_sharing))
```

DATE 0 **RENTED_BIKE_COUNT** 295 **HOUR** 0 **TEMPERATURE** 11
HUMIDITY 0 **WIND_SPEED** 0 **VISIBILITY** 0
DEW_POINT_TEMPERATURE 0 **SOLAR_RADIATION** 0 **RAINFALL** 0
SNOWFALL 0 **SEASONS** 0 **HOLIDAY** 0 **FUNCTIONING_DAY** 0

[7]:
```
###TASK: Detect and handle missing values
# Drop rows with `RENTED_BIKE_COUNT` column == NA
seoul_bike_sharing <- seoul_bike_sharing %>% filter(!is.na(RENTED_BIKE_COUNT))
dim(seoul_bike_sharing)
```

1. 8465 2. 14

[8]:
```
seoul_bike_sharing %>%
  filter(is.na(TEMPERATURE))
```

A spec_tbl_df: 11 × 14

DATE <chr>	RENTED_BIKE_COUNT <dbl>	HOUR <dbl>	TEMPERATURE <dbl>	HUMIDITY <dbl>
07/06/2018	3221	18	NA	57
12/06/2018	1246	14	NA	45
13/06/2018	2664	17	NA	57
17/06/2018	2330	17	NA	58
20/06/2018	2741	19	NA	61
30/06/2018	1144	13	NA	87
05/07/2018	827	10	NA	75
11/07/2018	634	9	NA	96
12/07/2018	593	6	NA	93
21/07/2018	347	4	NA	77
21/08/2018	1277	23	NA	75

[9]:
```
# Calculate the summer average temperature
summer_avg_temp <- seoul_bike_sharing %>%
  filter(SEASONS == "Summer") %>%
  select(TEMPERATURE) %>%
  summarise(mean(TEMPERATURE, na.rm = TRUE)) %>%
  unlist() %>%
  unname()
```

[10]:
```
# Impute missing values for TEMPERATURE column with summer average temperature
seoul_bike_sharing$TEMPERATURE <- replace_na(seoul_bike_sharing$TEMPERATURE,␣
↪summer_avg_temp)
```

[11]:
```
seoul_bike_sharing %>%
  filter(is.na(TEMPERATURE))
```

A spec_tbl_df: 0 × 14

DATE <chr>	RENTED_BIKE_COUNT <dbl>	HOUR <dbl>	TEMPERATURE <dbl>	HUMIDITY <dbl>	WIN <dbl

[12]:
```
# Print the summary of the dataset again to make sure no missing values in all␣
↪columns
summary(seoul_bike_sharing)
```

```
    DATE           RENTED_BIKE_COUNT      HOUR        TEMPERATURE
Length:8465        Min.   :   2.0    Min.   : 0.00   Min.   :-17.80
Class :character   1st Qu.: 214.0    1st Qu.: 6.00   1st Qu.:  3.00
Mode  :character   Median : 542.0    Median :12.00   Median : 13.50
                   Mean   : 729.2    Mean   :11.51   Mean   : 12.77
                   3rd Qu.:1084.0    3rd Qu.:18.00   3rd Qu.: 22.70
                   Max.   :3556.0    Max.   :23.00   Max.   : 39.40
   HUMIDITY       WIND_SPEED      VISIBILITY   DEW_POINT_TEMPERATURE
Min.   : 0.00   Min.   :0.000   Min.   :  27   Min.   :-30.600
1st Qu.:42.00   1st Qu.:0.900   1st Qu.: 935   1st Qu.: -5.100
Median :57.00   Median :1.500   Median :1690   Median :  4.700
Mean   :58.15   Mean   :1.726   Mean   :1434   Mean   :  3.945
3rd Qu.:74.00   3rd Qu.:2.300   3rd Qu.:2000   3rd Qu.: 15.200
Max.   :98.00   Max.   :7.400   Max.   :2000   Max.   : 27.200
SOLAR_RADIATION     RAINFALL          SNOWFALL         SEASONS
Min.   :0.0000   Min.   : 0.0000   Min.   :0.00000   Length:8465
1st Qu.:0.0000   1st Qu.: 0.0000   1st Qu.:0.00000   Class :character
Median :0.0100   Median : 0.0000   Median :0.00000   Mode  :character
Mean   :0.5679   Mean   : 0.1491   Mean   :0.07769
3rd Qu.:0.9300   3rd Qu.: 0.0000   3rd Qu.:0.00000
Max.   :3.5200   Max.   :35.0000   Max.   :8.80000
  HOLIDAY          FUNCTIONING_DAY
Length:8465        Length:8465
Class :character   Class :character
Mode  :character   Mode  :character
```

[13]:
```
write.csv(seoul_bike_sharing, "seoul_bike_sharing.csv")
```

[14]:
```
##TASK: Create indicator (dummy) variables for categorical variables
glimpse(seoul_bike_sharing)
```

```
Rows: 8,465
Columns: 14
$ DATE                 <chr> "01/12/2017", "01/12/2017", "01/12/2017", "01/1…
$ RENTED_BIKE_COUNT    <dbl> 254, 204, 173, 107, 78, 100, 181, 460, 930, 490…
$ HOUR                 <dbl> 0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 1…
$ TEMPERATURE          <dbl> -5.2, -5.5, -6.0, -6.2, -6.0, -6.4, -6.6, -7.4,…
$ HUMIDITY             <dbl> 37, 38, 39, 40, 36, 37, 35, 38, 37, 27, 24, 21,…
$ WIND_SPEED           <dbl> 2.2, 0.8, 1.0, 0.9, 2.3, 1.5, 1.3, 0.9, 1.1, 0.…
$ VISIBILITY           <dbl> 2000, 2000, 2000, 2000, 2000, 2000, 2000, 2000,…
```

```
$ DEW_POINT_TEMPERATURE <dbl> -17.6, -17.6, -17.7, -17.6, -18.6, -18.7, -19.5…
$ SOLAR_RADIATION       <dbl> 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00,…
$ RAINFALL              <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,…
$ SNOWFALL              <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,…
$ SEASONS               <chr> "Winter", "Winter", "Winter", "Winter", "Winter…
$ HOLIDAY               <chr> "No Holiday", "No Holiday", "No Holiday", "No H…
$ FUNCTIONING_DAY       <chr> "Yes", "Yes", "Yes", "Yes", "Yes", "Yes", "Yes"…
```

[15]:
```r
###TASK: Create indicator (dummy) variables for categorical variables
#Convert HOUR column from numeric into character first
# Using mutate() function to convert HOUR column into character type
seoul_bike_sharing <- seoul_bike_sharing %>%
mutate(HOUR = as.character(HOUR))
glimpse(seoul_bike_sharing)
```

```
Rows: 8,465
Columns: 14
$ DATE                  <chr> "01/12/2017", "01/12/2017", "01/12/2017", "01/1…
$ RENTED_BIKE_COUNT     <dbl> 254, 204, 173, 107, 78, 100, 181, 460, 930, 490…
$ HOUR                  <chr> "0", "1", "2", "3", "4", "5", "6", "7", "8", "9…
$ TEMPERATURE           <dbl> -5.2, -5.5, -6.0, -6.2, -6.0, -6.4, -6.6, -7.4,…
$ HUMIDITY              <dbl> 37, 38, 39, 40, 36, 37, 35, 38, 37, 27, 24, 21,…
$ WIND_SPEED            <dbl> 2.2, 0.8, 1.0, 0.9, 2.3, 1.5, 1.3, 0.9, 1.1, 0.…
$ VISIBILITY            <dbl> 2000, 2000, 2000, 2000, 2000, 2000, 2000, 2000,…
$ DEW_POINT_TEMPERATURE <dbl> -17.6, -17.6, -17.7, -17.6, -18.6, -18.7, -19.5…
$ SOLAR_RADIATION       <dbl> 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00,…
$ RAINFALL              <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,…
$ SNOWFALL              <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,…
$ SEASONS               <chr> "Winter", "Winter", "Winter", "Winter", "Winter…
$ HOLIDAY               <chr> "No Holiday", "No Holiday", "No Holiday", "No H…
$ FUNCTIONING_DAY       <chr> "Yes", "Yes", "Yes", "Yes", "Yes", "Yes", "Yes"…
```

[16]:
```r
# Convert SEASONS, HOLIDAY, FUNCTIONING_DAY, and HOUR columns into indicator␣
↪columns.
col <- c("SEASONS", "HOLIDAY", "HOUR")

feature <- function(x) {
  for (x in col) {
    seoul_bike_sharing <<- seoul_bike_sharing %>%
      mutate(dummy = 1) %>%
      spread(key = x, value = dummy, fill = 0)
  }
}
feature()
```

[17]:
```r
# Print the dataset summary again to make sure the indicator columns are␣
↪created properly
```

```
summary(seoul_bike_sharing)
```

```
    DATE           RENTED_BIKE_COUNT  TEMPERATURE        HUMIDITY
Length:8465        Min.   :   2.0    Min.   :-17.80   Min.   : 0.00
Class :character   1st Qu.: 214.0    1st Qu.:  3.00   1st Qu.:42.00
Mode  :character   Median : 542.0    Median : 13.50   Median :57.00
                   Mean   : 729.2    Mean   : 12.77   Mean   :58.15
                   3rd Qu.:1084.0    3rd Qu.: 22.70   3rd Qu.:74.00
                   Max.   :3556.0    Max.   : 39.40   Max.   :98.00
  WIND_SPEED      VISIBILITY   DEW_POINT_TEMPERATURE SOLAR_RADIATION
Min.   :0.000   Min.   :  27   Min.   :-30.600       Min.   :0.0000
1st Qu.:0.900   1st Qu.: 935   1st Qu.: -5.100       1st Qu.:0.0000
Median :1.500   Median :1690   Median :  4.700       Median :0.0100
Mean   :1.726   Mean   :1434   Mean   :  3.945       Mean   :0.5679
3rd Qu.:2.300   3rd Qu.:2000   3rd Qu.: 15.200       3rd Qu.:0.9300
Max.   :7.400   Max.   :2000   Max.   : 27.200       Max.   :3.5200
   RAINFALL          SNOWFALL       FUNCTIONING_DAY        Autumn
Min.   : 0.0000   Min.   :0.00000   Length:8465        Min.   :0.0000
1st Qu.: 0.0000   1st Qu.:0.00000   Class :character   1st Qu.:0.0000
Median : 0.0000   Median :0.00000   Mode  :character   Median :0.0000
Mean   : 0.1491   Mean   :0.07769                      Mean   :0.2288
3rd Qu.: 0.0000   3rd Qu.:0.00000                      3rd Qu.:0.0000
Max.   :35.0000   Max.   :8.80000                      Max.   :1.0000
    Spring           Summer           Winter          Holiday
Min.   :0.0000   Min.   :0.0000   Min.   :0.0000   Min.   :0.0000
1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.0000
Median :0.0000   Median :0.0000   Median :0.0000   Median :0.0000
Mean   :0.2552   Mean   :0.2608   Mean   :0.2552   Mean   :0.0482
3rd Qu.:1.0000   3rd Qu.:1.0000   3rd Qu.:1.0000   3rd Qu.:0.0000
Max.   :1.0000   Max.   :1.0000   Max.   :1.0000   Max.   :1.0000
  No Holiday           0                 1                 10
Min.   :0.0000   Min.   :0.00000   Min.   :0.00000   Min.   :0.0000
1st Qu.:1.0000   1st Qu.:0.00000   1st Qu.:0.00000   1st Qu.:0.0000
Median :1.0000   Median :0.00000   Median :0.00000   Median :0.0000
Mean   :0.9518   Mean   :0.04158   Mean   :0.04158   Mean   :0.0417
3rd Qu.:1.0000   3rd Qu.:0.00000   3rd Qu.:0.00000   3rd Qu.:0.0000
Max.   :1.0000   Max.   :1.00000   Max.   :1.00000   Max.   :1.0000
      11               12               13               14
Min.   :0.0000   Min.   :0.0000   Min.   :0.0000   Min.   :0.0000
1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.0000
Median :0.0000   Median :0.0000   Median :0.0000   Median :0.0000
Mean   :0.0417   Mean   :0.0417   Mean   :0.0417   Mean   :0.0417
3rd Qu.:0.0000   3rd Qu.:0.0000   3rd Qu.:0.0000   3rd Qu.:0.0000
Max.   :1.0000   Max.   :1.0000   Max.   :1.0000   Max.   :1.0000
      15               16               17               18
Min.   :0.0000   Min.   :0.0000   Min.   :0.0000   Min.   :0.0000
1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.0000
```

```
Median :0.0000   Median :0.0000   Median :0.0000   Median :0.0000
Mean   :0.0417   Mean   :0.0417   Mean   :0.0417   Mean   :0.0417
3rd Qu.:0.0000   3rd Qu.:0.0000   3rd Qu.:0.0000   3rd Qu.:0.0000
Max.   :1.0000   Max.   :1.0000   Max.   :1.0000   Max.   :1.0000
      19               2                20               21
Min.   :0.0000   Min.   :0.00000   Min.   :0.0000   Min.   :0.0000
1st Qu.:0.0000   1st Qu.:0.00000   1st Qu.:0.0000   1st Qu.:0.0000
Median :0.0000   Median :0.00000   Median :0.0000   Median :0.0000
Mean   :0.0417   Mean   :0.04158   Mean   :0.0417   Mean   :0.0417
3rd Qu.:0.0000   3rd Qu.:0.00000   3rd Qu.:0.0000   3rd Qu.:0.0000
Max.   :1.0000   Max.   :1.00000   Max.   :1.0000   Max.   :1.0000
      22               23               3                4
Min.   :0.0000   Min.   :0.0000   Min.   :0.00000   Min.   :0.00000
1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.00000   1st Qu.:0.00000
Median :0.0000   Median :0.0000   Median :0.00000   Median :0.00000
Mean   :0.0417   Mean   :0.0417   Mean   :0.04158   Mean   :0.04158
3rd Qu.:0.0000   3rd Qu.:0.0000   3rd Qu.:0.00000   3rd Qu.:0.00000
Max.   :1.0000   Max.   :1.0000   Max.   :1.00000   Max.   :1.00000
      5                6                7                8
Min.   :0.00000   Min.   :0.00000   Min.   :0.0000   Min.   :0.0000
1st Qu.:0.00000   1st Qu.:0.00000   1st Qu.:0.0000   1st Qu.:0.0000
Median :0.00000   Median :0.00000   Median :0.0000   Median :0.0000
Mean   :0.04158   Mean   :0.04158   Mean   :0.0417   Mean   :0.0417
3rd Qu.:0.00000   3rd Qu.:0.00000   3rd Qu.:0.0000   3rd Qu.:0.0000
Max.   :1.00000   Max.   :1.00000   Max.   :1.0000   Max.   :1.0000
      9
Min.   :0.0000
1st Qu.:0.0000
Median :0.0000
Mean   :0.0417
3rd Qu.:0.0000
Max.   :1.0000
```

[18]:
```
# drop the new added column
seoul_bike_sharing <- seoul_bike_sharing[, -1]
```

[19]:
```
# Print the summary of the dataset again to make sure the numeric columns range␣
 ↪between 0 and 1
summary(seoul_bike_sharing)
```

```
RENTED_BIKE_COUNT  TEMPERATURE        HUMIDITY       WIND_SPEED
Min.   :   2.0    Min.   :-17.80   Min.   : 0.00   Min.   :0.000
1st Qu.: 214.0    1st Qu.:  3.00   1st Qu.:42.00   1st Qu.:0.900
Median : 542.0    Median : 13.50   Median :57.00   Median :1.500
Mean   : 729.2    Mean   : 12.77   Mean   :58.15   Mean   :1.726
3rd Qu.:1084.0    3rd Qu.: 22.70   3rd Qu.:74.00   3rd Qu.:2.300
Max.   :3556.0    Max.   : 39.40   Max.   :98.00   Max.   :7.400
  VISIBILITY   DEW_POINT_TEMPERATURE SOLAR_RADIATION     RAINFALL
```

```
Min.   :  27   Min.   :-30.600       Min.   :0.0000   Min.   : 0.0000
1st Qu.: 935   1st Qu.: -5.100       1st Qu.:0.0000   1st Qu.: 0.0000
Median :1690   Median :  4.700       Median :0.0100   Median : 0.0000
Mean   :1434   Mean   :  3.945       Mean   :0.5679   Mean   : 0.1491
3rd Qu.:2000   3rd Qu.: 15.200       3rd Qu.:0.9300   3rd Qu.: 0.0000
Max.   :2000   Max.   : 27.200       Max.   :3.5200   Max.   :35.0000
   SNOWFALL       FUNCTIONING_DAY        Autumn           Spring
Min.   :0.00000   Length:8465        Min.   :0.0000   Min.   :0.0000
1st Qu.:0.00000   Class :character   1st Qu.:0.0000   1st Qu.:0.0000
Median :0.00000   Mode  :character   Median :0.0000   Median :0.0000
Mean   :0.07769                      Mean   :0.2288   Mean   :0.2552
3rd Qu.:0.00000                      3rd Qu.:0.0000   3rd Qu.:1.0000
Max.   :8.80000                      Max.   :1.0000   Max.   :1.0000
    Summer           Winter          Holiday         No Holiday
Min.   :0.0000   Min.   :0.0000   Min.   :0.0000   Min.   :0.0000
1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:1.0000
Median :0.0000   Median :0.0000   Median :0.0000   Median :1.0000
Mean   :0.2608   Mean   :0.2552   Mean   :0.0482   Mean   :0.9518
3rd Qu.:1.0000   3rd Qu.:1.0000   3rd Qu.:0.0000   3rd Qu.:1.0000
Max.   :1.0000   Max.   :1.0000   Max.   :1.0000   Max.   :1.0000
      0                1                10               11
Min.   :0.00000   Min.   :0.00000   Min.   :0.0000   Min.   :0.0000
1st Qu.:0.00000   1st Qu.:0.00000   1st Qu.:0.0000   1st Qu.:0.0000
Median :0.00000   Median :0.00000   Median :0.0000   Median :0.0000
Mean   :0.04158   Mean   :0.04158   Mean   :0.0417   Mean   :0.0417
3rd Qu.:0.00000   3rd Qu.:0.00000   3rd Qu.:0.0000   3rd Qu.:0.0000
Max.   :1.00000   Max.   :1.00000   Max.   :1.0000   Max.   :1.0000
      12               13               14               15
Min.   :0.0000   Min.   :0.0000   Min.   :0.0000   Min.   :0.0000
1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.0000
Median :0.0000   Median :0.0000   Median :0.0000   Median :0.0000
Mean   :0.0417   Mean   :0.0417   Mean   :0.0417   Mean   :0.0417
3rd Qu.:0.0000   3rd Qu.:0.0000   3rd Qu.:0.0000   3rd Qu.:0.0000
Max.   :1.0000   Max.   :1.0000   Max.   :1.0000   Max.   :1.0000
      16               17               18               19
Min.   :0.0000   Min.   :0.0000   Min.   :0.0000   Min.   :0.0000
1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.0000
Median :0.0000   Median :0.0000   Median :0.0000   Median :0.0000
Mean   :0.0417   Mean   :0.0417   Mean   :0.0417   Mean   :0.0417
3rd Qu.:0.0000   3rd Qu.:0.0000   3rd Qu.:0.0000   3rd Qu.:0.0000
Max.   :1.0000   Max.   :1.0000   Max.   :1.0000   Max.   :1.0000
      2                20               21               22
Min.   :0.00000   Min.   :0.0000   Min.   :0.0000   Min.   :0.0000
1st Qu.:0.00000   1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.0000
Median :0.00000   Median :0.0000   Median :0.0000   Median :0.0000
Mean   :0.04158   Mean   :0.0417   Mean   :0.0417   Mean   :0.0417
3rd Qu.:0.00000   3rd Qu.:0.0000   3rd Qu.:0.0000   3rd Qu.:0.0000
Max.   :1.00000   Max.   :1.0000   Max.   :1.0000   Max.   :1.0000
```

```
      23                3                 4                 5
 Min.   :0.0000   Min.   :0.00000   Min.   :0.00000   Min.   :0.00000
 1st Qu.:0.0000   1st Qu.:0.00000   1st Qu.:0.00000   1st Qu.:0.00000
 Median :0.0000   Median :0.00000   Median :0.00000   Median :0.00000
 Mean   :0.0417   Mean   :0.04158   Mean   :0.04158   Mean   :0.04158
 3rd Qu.:0.0000   3rd Qu.:0.00000   3rd Qu.:0.00000   3rd Qu.:0.00000
 Max.   :1.0000   Max.   :1.00000   Max.   :1.00000   Max.   :1.00000
      6                 7                 8                 9
 Min.   :0.00000   Min.   :0.0000   Min.   :0.0000   Min.   :0.0000
 1st Qu.:0.00000   1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.0000
 Median :0.00000   Median :0.0000   Median :0.0000   Median :0.0000
 Mean   :0.04158   Mean   :0.0417   Mean   :0.0417   Mean   :0.0417
 3rd Qu.:0.00000   3rd Qu.:0.0000   3rd Qu.:0.0000   3rd Qu.:0.0000
 Max.   :1.00000   Max.   :1.0000   Max.   :1.0000   Max.   :1.0000
```

[20]:
```
# Save the dataset as `seoul_bike_sharing_converted.csv`
write_csv(seoul_bike_sharing, "seoul_bike_sharing_converted.csv")
```

[21]:
```
###TASK: Normalize data
#TODO: Apply min-max normalization on RENTED_BIKE_COUNT, TEMPERATURE,
#HUMIDITY, WIND_SPEED, VISIBILITY, DEW_POINT_TEMPERATURE, SOLAR_RADIATION,
#RAINFALL, SNOWFALL
minmax_scale <- (seoul_bike_sharing$RENTED_BIKE_COUNT -␣
 ↪min(seoul_bike_sharing$RENTED_BIKE_COUNT)) /
                (max(seoul_bike_sharing$RENTED_BIKE_COUNT) -␣
 ↪min(seoul_bike_sharing$RENTED_BIKE_COUNT))
head(minmax_scale)
```

1. 0.0709060213843557 2. 0.0568373663477772 3. 0.0481148002250985 4. 0.0295441755768149 5. 0.0213843556555993 6. 0.0275745638716939

[22]:
```
minmax_scale1 <- (seoul_bike_sharing$TEMPERATURE -␣
 ↪min(seoul_bike_sharing$TEMPERATURE)) /
                (max(seoul_bike_sharing$TEMPERATURE) -␣
 ↪min(seoul_bike_sharing$TEMPERATURE))
head(minmax_scale1)
```

1. 0.22027972027972 2. 0.215034965034965 3. 0.206293706293706 4. 0.202797202797203 5. 0.206293706293706 6. 0.199300699300699

[23]:
```
minmax_scale2 <- (seoul_bike_sharing$HUMIDITY -␣
 ↪min(seoul_bike_sharing$HUMIDITY)) /
                (max(seoul_bike_sharing$HUMIDITY) -␣
 ↪min(seoul_bike_sharing$HUMIDITY))
head(minmax_scale2)
```

1. 0.377551020408163 2. 0.387755102040816 3. 0.397959183673469 4. 0.408163265306122 5. 0.36734693877551 6. 0.377551020408163

[24]:
```
minmax_scale3 <- (seoul_bike_sharing$WIND_SPEED -␣
 ↪min(seoul_bike_sharing$WIND_SPEED)) /
                (max(seoul_bike_sharing$WIND_SPEED) -␣
 ↪min(seoul_bike_sharing$WIND_SPEED))
head(minmax_scale3)
```

1. 0.297297297297297 2. 0.108108108108108 3. 0.135135135135135 4. 0.121621621621622 5. 0.310810810810811 6. 0.202702702702703

[25]:
```
minmax_scale4 <- (seoul_bike_sharing$VISIBILITY -␣
 ↪min(seoul_bike_sharing$VISIBILITY)) /
                (max(seoul_bike_sharing$VISIBILITY) -␣
 ↪min(seoul_bike_sharing$VISIBILITY))
head(minmax_scale4)
```

1. 1 2. 1 3. 1 4. 1 5. 1 6. 1

[26]:
```
minmax_scale5 <- (seoul_bike_sharing$DEW_POINT_TEMPERATURE -␣
 ↪min(seoul_bike_sharing$DEW_POINT_TEMPERATURE)) /
                (max(seoul_bike_sharing$DEW_POINT_TEMPERATURE) -␣
 ↪min(seoul_bike_sharing$DEW_POINT_TEMPERATURE))
head(minmax_scale5)
```

1. 0.224913494809689 2. 0.224913494809689 3. 0.22318339100346 4. 0.224913494809689 5. 0.207612456747405 6. 0.205882352941177

[27]:
```
minmax_scale6 <- (seoul_bike_sharing$SOLAR_RADIATION -␣
 ↪min(seoul_bike_sharing$SOLAR_RADIATION)) /
                (max(seoul_bike_sharing$SOLAR_RADIATION) -␣
 ↪min(seoul_bike_sharing$SOLAR_RADIATION))
head(minmax_scale6)
```

1. 0 2. 0 3. 0 4. 0 5. 0 6. 0

[28]:
```
minmax_scale7 <- (seoul_bike_sharing$RAINFALL -␣
 ↪min(seoul_bike_sharing$RAINFALL)) /
                (max(seoul_bike_sharing$RAINFALL) -␣
 ↪min(seoul_bike_sharing$RAINFALL))
head(minmax_scale7)
```

1. 0 2. 0 3. 0 4. 0 5. 0 6. 0

[29]:
```
minmax_scale8 <- (seoul_bike_sharing$SNOWFALL -␣
 ↪min(seoul_bike_sharing$SNOWFALL)) /
                (max(seoul_bike_sharing$SNOWFALL) -␣
 ↪min(seoul_bike_sharing$SNOWFALL))
head(minmax_scale8)
```

1. 0 2. 0 3. 0 4. 0 5. 0 6. 0

[30]:
```
# Print the summary of the dataset again to make sure the numeric columns range␣
↪between 0 and 1
summary(seoul_bike_sharing)
```

```
RENTED_BIKE_COUNT  TEMPERATURE        HUMIDITY       WIND_SPEED
Min.   :   2.0    Min.   :-17.80   Min.   : 0.00   Min.   :0.000
1st Qu.: 214.0    1st Qu.:  3.00   1st Qu.:42.00   1st Qu.:0.900
Median : 542.0    Median : 13.50   Median :57.00   Median :1.500
Mean   : 729.2    Mean   : 12.77   Mean   :58.15   Mean   :1.726
3rd Qu.:1084.0    3rd Qu.: 22.70   3rd Qu.:74.00   3rd Qu.:2.300
Max.   :3556.0    Max.   : 39.40   Max.   :98.00   Max.   :7.400
  VISIBILITY   DEW_POINT_TEMPERATURE SOLAR_RADIATION     RAINFALL
Min.   :  27   Min.   :-30.600       Min.   :0.0000   Min.   : 0.0000
1st Qu.: 935   1st Qu.: -5.100       1st Qu.:0.0000   1st Qu.: 0.0000
Median :1690   Median :  4.700       Median :0.0100   Median : 0.0000
Mean   :1434   Mean   :  3.945       Mean   :0.5679   Mean   : 0.1491
3rd Qu.:2000   3rd Qu.: 15.200       3rd Qu.:0.9300   3rd Qu.: 0.0000
Max.   :2000   Max.   : 27.200       Max.   :3.5200   Max.   :35.0000
   SNOWFALL       FUNCTIONING_DAY        Autumn           Spring
Min.   :0.00000   Length:8465        Min.   :0.0000   Min.   :0.0000
1st Qu.:0.00000   Class :character   1st Qu.:0.0000   1st Qu.:0.0000
Median :0.00000   Mode  :character   Median :0.0000   Median :0.0000
Mean   :0.07769                      Mean   :0.2288   Mean   :0.2552
3rd Qu.:0.00000                      3rd Qu.:0.0000   3rd Qu.:1.0000
Max.   :8.80000                      Max.   :1.0000   Max.   :1.0000
    Summer           Winter           Holiday          No Holiday
Min.   :0.0000   Min.   :0.0000   Min.   :0.0000   Min.   :0.0000
1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:1.0000
Median :0.0000   Median :0.0000   Median :0.0000   Median :1.0000
Mean   :0.2608   Mean   :0.2552   Mean   :0.0482   Mean   :0.9518
3rd Qu.:1.0000   3rd Qu.:1.0000   3rd Qu.:0.0000   3rd Qu.:1.0000
Max.   :1.0000   Max.   :1.0000   Max.   :1.0000   Max.   :1.0000
      0                 1                10               11
Min.   :0.00000   Min.   :0.00000   Min.   :0.0000   Min.   :0.0000
1st Qu.:0.00000   1st Qu.:0.00000   1st Qu.:0.0000   1st Qu.:0.0000
Median :0.00000   Median :0.00000   Median :0.0000   Median :0.0000
Mean   :0.04158   Mean   :0.04158   Mean   :0.0417   Mean   :0.0417
3rd Qu.:0.00000   3rd Qu.:0.00000   3rd Qu.:0.0000   3rd Qu.:0.0000
Max.   :1.00000   Max.   :1.00000   Max.   :1.0000   Max.   :1.0000
      12                13                14               15
Min.   :0.0000   Min.   :0.0000   Min.   :0.0000   Min.   :0.0000
1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.0000
Median :0.0000   Median :0.0000   Median :0.0000   Median :0.0000
Mean   :0.0417   Mean   :0.0417   Mean   :0.0417   Mean   :0.0417
3rd Qu.:0.0000   3rd Qu.:0.0000   3rd Qu.:0.0000   3rd Qu.:0.0000
Max.   :1.0000   Max.   :1.0000   Max.   :1.0000   Max.   :1.0000
      16                17                18               19
```

```
Min.   :0.0000   Min.   :0.0000   Min.   :0.0000   Min.   :0.0000
1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.0000
Median :0.0000   Median :0.0000   Median :0.0000   Median :0.0000
Mean   :0.0417   Mean   :0.0417   Mean   :0.0417   Mean   :0.0417
3rd Qu.:0.0000   3rd Qu.:0.0000   3rd Qu.:0.0000   3rd Qu.:0.0000
Max.   :1.0000   Max.   :1.0000   Max.   :1.0000   Max.   :1.0000
      2                20               21               22
Min.   :0.00000   Min.   :0.0000   Min.   :0.0000   Min.   :0.0000
1st Qu.:0.00000   1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.0000
Median :0.00000   Median :0.0000   Median :0.0000   Median :0.0000
Mean   :0.04158   Mean   :0.0417   Mean   :0.0417   Mean   :0.0417
3rd Qu.:0.00000   3rd Qu.:0.0000   3rd Qu.:0.0000   3rd Qu.:0.0000
Max.   :1.00000   Max.   :1.0000   Max.   :1.0000   Max.   :1.0000
      23               3                4                5
Min.   :0.0000   Min.   :0.00000   Min.   :0.00000   Min.   :0.00000
1st Qu.:0.0000   1st Qu.:0.00000   1st Qu.:0.00000   1st Qu.:0.00000
Median :0.0000   Median :0.00000   Median :0.00000   Median :0.00000
Mean   :0.0417   Mean   :0.04158   Mean   :0.04158   Mean   :0.04158
3rd Qu.:0.0000   3rd Qu.:0.00000   3rd Qu.:0.00000   3rd Qu.:0.00000
Max.   :1.0000   Max.   :1.00000   Max.   :1.00000   Max.   :1.00000
      6                7                8                9
Min.   :0.00000   Min.   :0.0000   Min.   :0.0000   Min.   :0.0000
1st Qu.:0.00000   1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.0000
Median :0.00000   Median :0.0000   Median :0.0000   Median :0.0000
Mean   :0.04158   Mean   :0.0417   Mean   :0.0417   Mean   :0.0417
3rd Qu.:0.00000   3rd Qu.:0.0000   3rd Qu.:0.0000   3rd Qu.:0.0000
Max.   :1.00000   Max.   :1.0000   Max.   :1.0000   Max.   :1.0000
```

[31]:
```
# Save the dataset as `seoul_bike_sharing_converted_normalized.csv`
write_csv(seoul_bike_sharing, "seoul_bike_sharing_converted_normalized.csv")
```

[32]:
```
# Dataset list
dataset_list <- c('seoul_bike_sharing.csv', 'seoul_bike_sharing_converted.csv',␣
 ↪'seoul_bike_sharing_converted_normalized.csv')

for (dataset_name in dataset_list){
    # Read dataset
    dataset <- read_csv(dataset_name)
    # Standardized its columns:
    # Convert all columns names to uppercase
    names(dataset) <- toupper(names(dataset))
    # Replace any white space separators by underscore, using str_replace_all␣
 ↪function
    names(dataset) <- str_replace_all(names(dataset), " ", "_")
    # Save the dataset back
    write.csv(dataset, dataset_name, row.names=FALSE)
}
```

```
Warning message:
"Missing column names filled in: 'X1' [1]"Parsed with column specification:
cols(
  X1 = col_double(),
  DATE = col_character(),
  RENTED_BIKE_COUNT = col_double(),
  HOUR = col_double(),
  TEMPERATURE = col_double(),
  HUMIDITY = col_double(),
  WIND_SPEED = col_double(),
  VISIBILITY = col_double(),
  DEW_POINT_TEMPERATURE = col_double(),
  SOLAR_RADIATION = col_double(),
  RAINFALL = col_double(),
  SNOWFALL = col_double(),
  SEASONS = col_character(),
  HOLIDAY = col_character(),
  FUNCTIONING_DAY = col_character()
)
Parsed with column specification:
cols(
  .default = col_double(),
  FUNCTIONING_DAY = col_character()
)
See spec(…) for full column specifications.
Parsed with column specification:
cols(
  .default = col_double(),
  FUNCTIONING_DAY = col_character()
)
See spec(…) for full column specifications.
```

[ ]: