

Seoul Bike Dataset

Andrew Danda, 구민우, 방설화

2024-04-21

This is a dataset containing Seoul Bike sharing ridership from December 1, 2017 to November 30, 2018. The objective is to predict the number of riders on at any given time across the year given the data shown in this dataset

Dependent variable	
<i>RentedBikeCount</i>	The number of bikes rented
Independent variables	
1. Time Variables	
<i>Date</i>	The Date (dd/mm/yyyy)
<i>Hour</i>	The Hour (integer between 1 and 24)
<i>Holiday</i>	Dummy variable if that day is a holiday or not
<i>FunctionalDay</i>	Dummy variable if the bikes were functional or not
<i>SeasonsSpring</i>	Dummy variable if the season is Spring or not
<i>SeasonsSummer</i>	Dummy variable if the season is Summer or not
<i>SeasonsAutumn</i>	Dummy variable if the season is Autumn or not
<i>SeasonsWinter</i>	Dummy variable if the season is Winter or not
2. Weather variables	
<i>Temperature</i>	Temperature in Celsius
<i>Humidity</i>	Humidity (%)
<i>Windspeed</i>	Wind speed in meters per second
<i>Visibility</i>	Visibility in Kilometers
<i>DewpointTemperature</i>	Dew Point Temperature in Celsius
<i>SolarRadiation</i>	Solar Radiation in MJ/m2
<i>Rainfall</i>	Rainfall in millimeters
<i>Snowfall</i>	Snowfall in centimeters

The Dataset

```
library(tidyverse)
library(dplyr)
library(fastDummies)
```

```
bikeData <- read.csv("SeoulBikeData.csv", stringsAsFactors=FALSE, fileEncoding="latin1")

# Clean dataset
```

```

# rename columns
bikeData <- bikeData %>%
  rename("Rented_Bike_Count" = "Rented.Bike.Count",
         "Temperature" = "Temperature..C.",
         "Humidity" = "Humidity...",
         "Wind_Speed" = "Wind.speed..m.s.",
         "Visibility" = "Visibility..10m.",
         "Dew_Point_Temperature" = "Dew.point.temperature..C.",
         "Solar_Radiation" = "Solar.Radiation..MJ.m2.",
         "Rainfall" = "Rainfall.mm.",
         "Snowfall" = "Snowfall..cm.",
         "Functioning_Day" = "Functioning.Day")

#divide the visibility by 100 to change it's units from 10s of meters to kilometers
bikeData$Visibility <- bikeData$Visibility / 100

# Dummy variables
bikeData$Holiday <- ifelse(bikeData$Holiday == "No Holiday", 0, 1)
bikeData$Functioning_Day <- ifelse(bikeData$Functioning_Day == "Yes", 1, 0)

#Holiday Dummies
bikeData <- bikeData %>% dummy_cols(select_columns = c("Seasons"))

summary(bikeData)

```

Date	Rented_Bike_Count	Hour	Temperature
Length:8760	Min. : 0.0	Min. : 0.00	Min. : -17.80
Class :character	1st Qu.: 191.0	1st Qu.: 5.75	1st Qu.: 3.50
Mode :character	Median : 504.5	Median : 11.50	Median : 13.70
	Mean : 704.6	Mean : 11.50	Mean : 12.88
	3rd Qu.: 1065.2	3rd Qu.: 17.25	3rd Qu.: 22.50
	Max. : 3556.0	Max. : 23.00	Max. : 39.40
Humidity	Wind_Speed	Visibility	Dew_Point_Temperature
Min. : 0.00	Min. : 0.000	Min. : 0.27	Min. : -30.600
1st Qu.: 42.00	1st Qu.: 0.900	1st Qu.: 9.40	1st Qu.: -4.700
Median : 57.00	Median : 1.500	Median : 16.98	Median : 5.100
Mean : 58.23	Mean : 1.725	Mean : 14.37	Mean : 4.074
3rd Qu.: 74.00	3rd Qu.: 2.300	3rd Qu.: 20.00	3rd Qu.: 14.800
Max. : 98.00	Max. : 7.400	Max. : 20.00	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	Seasons_Autumn	Seasons_Spring
Min. :0.00000	Min. :0.0000	Min. :0.0000	Min. :0.0000
1st Qu.:0.00000	1st Qu.:1.0000	1st Qu.:0.0000	1st Qu.:0.0000
Median :0.00000	Median :1.0000	Median :0.0000	Median :0.0000
Mean :0.04932	Mean :0.9663	Mean :0.2493	Mean :0.2521
3rd Qu.:0.00000	3rd Qu.:1.0000	3rd Qu.:0.0000	3rd Qu.:1.0000
Max. :1.00000	Max. :1.0000	Max. :1.0000	Max. :1.0000
Seasons_Summer	Seasons_Winter		
Min. :0.0000	Min. :0.0000		
1st Qu.:0.0000	1st Qu.:0.0000		
Median :0.0000	Median :0.0000		
Mean :0.2521	Mean :0.2466		
3rd Qu.:1.0000	3rd Qu.:0.0000		
Max. :1.0000	Max. :1.0000		