Practice 9: k-Nearest Neighbor

Loading data into R

```
bike_rental_df <- read.csv('SeoulBikeData.csv')
```

bike rental dataset Currently Rental bikes are introduced in many urban cities for the enhancement of mobility comfort.

It is important to make the rental bike available and accessible to the public at the right time as it lessens the waiting time.

Eventually, providing the city with a stable supply of rental bikes becomes a major concern.

The crucial part is the prediction of bike count required at each hour for the stable supply of rental bikes.

The dataset contains weather information (Temperature, Humidity, Windspeed, Visibility, Dewpoint, Solar radiation, Snowfall, Rainfall), the number of bikes rented per hour and date information.

The dataset was collected from 2017 DEC 1st to 2018 NOV 30th for one year.

이 데이터는 매시간 서울시의 자전거 렌탈 수요를 예측하기 위한 데이터로 해당 날짜/시간에 해당하는 날씨 정보가 함께 포함되어있다.

자전거 렌탈 수요를 정확하게 예측할 수 있다면 자전거가 필요한 시민들이 기다리는 시간 없이 보다 쉽게 자전거를 대여할 수 있도록 시에서는 필요한 곳에 필요한 수의 자전거를 준비할 수 있을 것이다.

- **Attribute Information**:

variable	unit measurement
Date	day/month/year
Rented Bike Count	Count of bikes rented at each hour
Hour	Hour of he day
Temperature	Temperature in Celsius
Humidity	%
Wind speed	m/s
Visibility	10m
Dew point	Celsius
Solar Radiation	MJ/m2
Rainfall	mm
Snowfall	cm
Seasons	Winter, Spring, Summer, Autumn
Holiday	Holiday/No holiday
Functioning Day	NoFunc(Non Functional Hours), Fun(Functional hours)

Data Prep.

Question 1-1

Split the dataset into two for training and testing. For training, use the dataset from 2017 DEC 1st to 2018 SEP 30. For testing, use the rest from 2018 OCT 1st.

학습데이터와 검증데이터로 분할 하여라.

학습데이터는 2018년 9월 이전까지, 검증 데이터는 2018년 10월 이후 데이터를 사용하여라.

```
dim(train_bike)
```

```
## [1] 7296   14
```

```
dim(test_bike)
```

```
## [1] 1464   14
```

Question 1-2

Compare distribution of each variable in both training and testing dataset.

Compare distribution and range of each variable.

Explain how similar or different the distribution is for both training and testing dataset.

학습 데이터와 검증 데이터에 대해 각 변수들의 분포를 비교하여라.

각 변수들의 분포와 범위를 비교하여라.

각 변수들이 학습 데이터와 검증 데이터에서 얼마나 비슷하게 (혹은 다르게) 분포하는지 설명하여라

Question 1-3

We will use k-Nearest Neighbor(kNN) method to predict the hourly number of rented bikes for test dataset.

Perform any necessary pre-processing you need before use kNN method.

Explain what they are and why you need them.

kNN 방법을 사용해서 검증 데이터의 자전거 대여 수요를 예측하고자 한다.

필요하다고 생각하는 전처리 과정이 있다면 수행하여라

어떤 전처리를 수행하였는지 그 이유는 무엇인지 설명하여라.

Applying kNN method

Question 2

Predict the hourly number of rented bikes for test dataset using kNN method (pick any k value you like to use).

Show the RMSE and R^2 of your result and explain what you can infer for your result.

kNN 방법을 사용하여 시간별 자전거 대여 수를 예측하여라 (k는 임의로 선정).

예측 결과에 대해 RMSE와 R^2를 계산하여라.

예측 결과에 대해서 어떻게 평가할 수 있는가?

Question 3

Try a range of **k**s to find the best k setting for prediction bike rental demand.

Draws graphs show how RMSE and R^2 changes over k values.

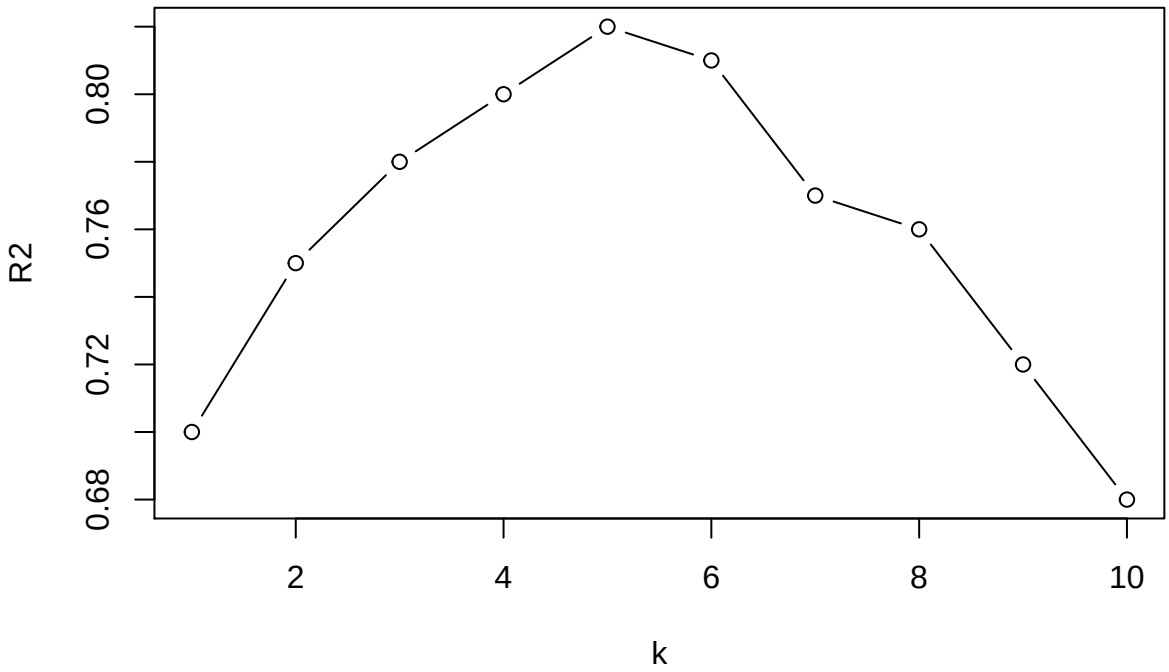
You can use R, Excel spread sheets, or any other SW to draw a graph. Or you can just hand-draw the graph as well. what is the best k?

다양한 k를 시도해보고 자전거 대여 수요를 가장 잘 예측하는 k를 찾아보시오.

k를 변화하면서 RMSE와 R^2가 변화하는 것을 보기 위해 그래프를 그리시오.

그래프는 R이나 엑셀이나 어떤 SW 사용해도 되고, 손으로 그려도 됩니다. 가장 성능이 좋은 k는 무엇입니까?

- **sample graph**



Question 4

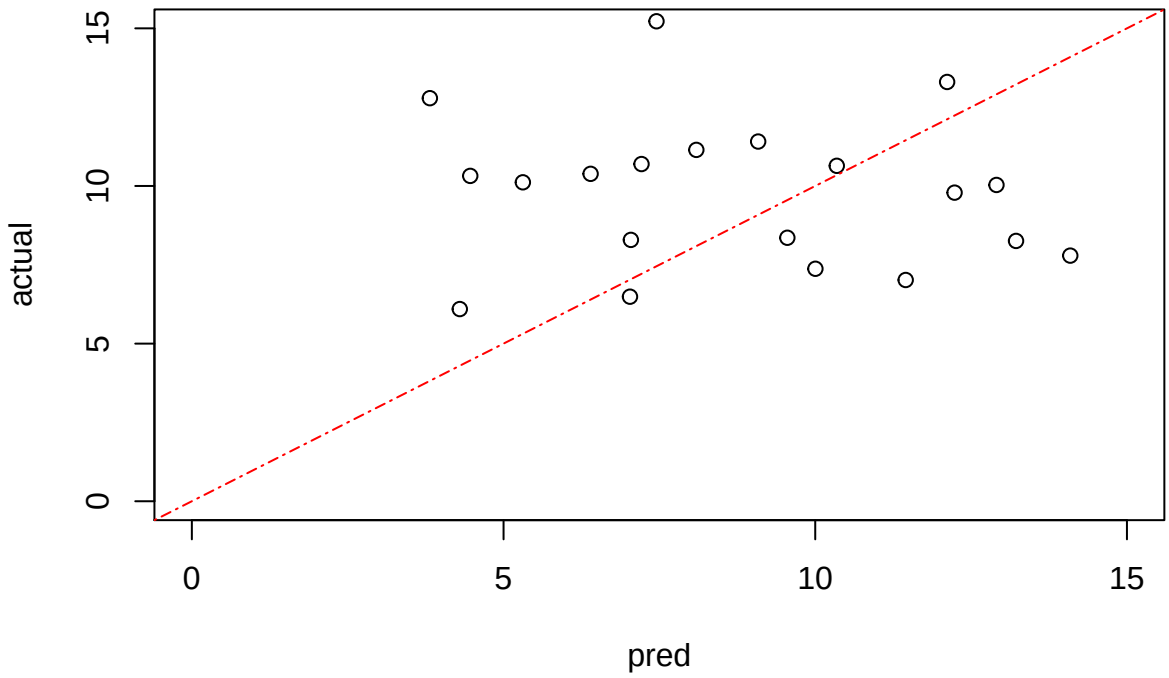
For the best k you have found in Q. 3, draw a scatterplot of your result where x-axis represents predicted bike-rents and y-axis for actual number of rented bikes with line of y = x.

What do the points above the line implies? the points below the line? what about points on the line?

앞서 발견한 최적의 k를 사용한 결과 대해서, x축에는 예측 값, y축에는 실제 자전거 대여수를 표현하는 산점도를 그리라. 그래프와 함께 y=x 선도 함께 그리라.

선 위에 나타난 점들과 선 아래 점들, 선에 가깝거나 겹치게 나타난 점들이 의미하는 것은 무엇인가?

- **sample graph**



Seoul bike rental dataset from UCI data repository