Multivariate Data Analysis Assignment #1

Multivariate Linear Regression

산업경영공학부 2020170831 민찬홍

[Q1] 데이터셋 선정 및 선정 이유

.-데이터셋: 서울시 2017, 2018년 자전거 대여량

-다운로드 링크: https://archive.ics.uci.edu/dataset/560/seoul+bike+sharing+demand

```
RangeIndex: 8760 entries, 0 to 8759
Data columns (total 14 columns):
 #   Column                  Non-Null Count  Dtype
---  ------                  --------------  -----
 0   Date                    8760 non-null   object
 1   Rented Bike Count       8760 non-null   int64
 2   Hour                    8760 non-null   int64
 3   Temperature             8760 non-null   float64
 4   Humidity(%)             8760 non-null   int64
 5   Wind speed (m/s)        8760 non-null   float64
 6   Visibility (10m)        8760 non-null   int64
 7   Dew point temperature   8760 non-null   float64
 8   Solar Radiation (MJ/m2) 8760 non-null   float64
 9   Rainfall(mm)            8760 non-null   float64
 10  Snowfall (cm)           8760 non-null   float64
 11  Seasons                 8760 non-null   object
 12  Holiday                 8760 non-null   object
 13  Functioning Day         8760 non-null   object
dtypes: float64(6), int64(4), object(4)
memory usage: 958.3+ KB
```

-선정 이유: Instance의 개수가 8760개로 적지 않은 데이터를 가지고 있었고, 입력변수로 총 14개의 attribute가 사용되기에 적합하다고 생각하였다. 또한 입력 변수의 경우 연속형 변수 뿐만 아니라 명목형 변수도 존재하여 1-of-C coding변환도 연습할 수 있을 것이라 생각되었다.

[Q2] 데이터셋 설명

-독립 변수: 자전거 대여량

-종속 변수: 시각, 기온, 습도, 바람의 세기, 가시성, 이슬점, 자외선지수, 강수량, 강설량, 계절, 휴일 , 근무일

독립 변수와 종속 변수를 위와 같이 잡을 수 있다.

이 때, 날짜(Date)의 경우 설명 변수에서 제외시켜 분석을 진행하였다. 날짜(Date)보다 그 날의 기온, 강수량, 강설량 등의 날씨 정보가 결국 실외 활동에 직결되어 자전거 대여량에 영향을 주기에 날짜를 제외한 설명 변수들로 자전거 대여량을 충분히 예측할 수 있을 것이라 생각했기 때문이다. 날짜(Date)를 제외시키면서 나타날 수 있는 문제를 계절(Season)이 어느 정도 커버해줄 수 있을 것이라 판단하기도 하였다.

[Q2-1] 이 데이터는 종속변수와 설명변수들 사이에 실제로 "선형 관계"가 있다고 가정할 수 있겠는가?

자외선지수, 강수량, 강설량의 종속변수의 경우에는 자전거 대여량과 강한 선형 관계가 있다고 가정할 수 있다.

왜냐하면 자전거 대여량은 실외활동 가능 여부 즉, 날씨에 영향을 많이 받기 때문이다.

반면 시각, 기온, 습도, 이슬점(습도가 높으면 이슬점이 높음) 변수의 경우 너무 높거나 낮지 않고 적당한 것이 좋기 때문에 종속변수와의 선형 관계가 있다고 가정하긴 어려울 것이다.

[Q2-2] 이 데이터에서 제공된 설명변수들 중에서 높은 상관관계가 있을 것으로 예상되는 변수들은 어떤 것들이 있는가?

2-1)에서 언급한 자외선지수, 강수량, 강설량 세 설명변수는 높은 상관 관계를 가지고 있을 것이라 예상된다.

왜냐하면 자외선지수, 강수량, 강설량은 직접적으로 자전거 이용에 영향을 끼치는 요소이기 때문이다.

[Q2-3] 제공된 설명변수들 중에서 종속변수를 예측하는데 필요하지 않을 것으로 예상되는 변수들은 어떤 것들이 있는가?

가시성, 이슬점의 경우 날씨와 직접적으로 연관되지 않은 정보기도 하고, 우리가 접하기 쉬운 정보는 아니기에 예측하는데 불필요할 것이라고 생각한다.

[Q3] 개별 입력변수들에 대한 통계량 계산 및 분석

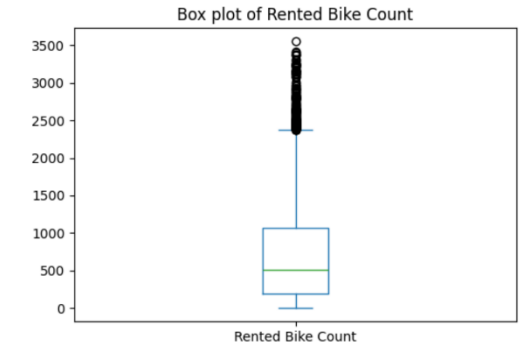
입력변수들에 대한 통계량 계산을 하기 전 우선 명목형 변수인 계절, 휴일, 근무일에 대해서는 1-of-C coding 변환을 하였다. 또한 앞서 언급하였듯이 본래의 데이터셋에서 불필요한 변수인 날짜는 제거하였다.

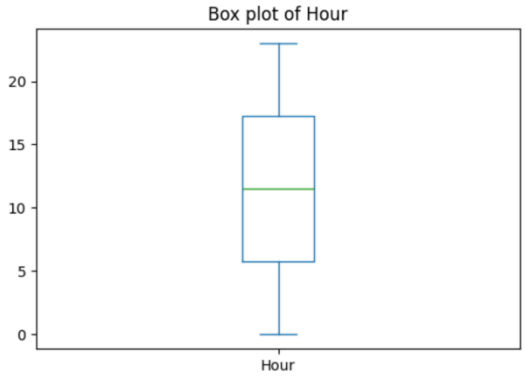
```
RangeIndex: 8760 entries, 0 to 8759
Data columns (total 18 columns):
 #   Column                  Non-Null Count  Dtype
---  ------                  --------------  -----
 0   Rented Bike Count       8760 non-null   int64
 1   Hour                    8760 non-null   int64
 2   Temperature             8760 non-null   float64
 3   Humidity(%)             8760 non-null   int64
 4   Wind speed (m/s)        8760 non-null   float64
 5   Visibility (10m)        8760 non-null   int64
 6   Dew point temperature   8760 non-null   float64
 7   Solar Radiation (MJ/m2) 8760 non-null   float64
 8   Rainfall(mm)            8760 non-null   float64
 9   Snowfall (cm)           8760 non-null   float64
 10  Spring                  8760 non-null   int32
 11  Summer                  8760 non-null   int32
 12  Autumn                  8760 non-null   int32
 13  Winter                  8760 non-null   int32
 14  Holiday                 8760 non-null   int32
 15  No Holiday              8760 non-null   int32
 16  Functioning day         8760 non-null   int32
 17  Not Functioning day     8760 non-null   int32
dtypes: float64(6), int32(8), int64(4)
memory usage: 958.3 KB
None
```

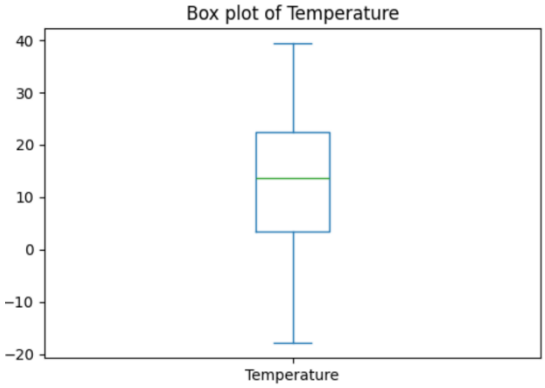
개별 입력 변수들에 대한 단변량 통계량 계산 결과는 다음과 같다.

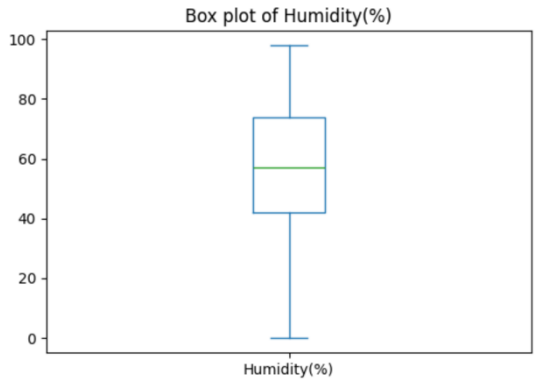
	Mean	Std Dev	Skewness	Kurtosis
Rented Bike Count	704.602055	644.997468	1.153231	0.852215
Hour	11.500000	6.922582	0.000000	-1.204174
Temperature	12.882922	11.944825	-0.198292	-0.837993
Humidity(%)	58.226256	20.362413	0.059569	-0.803785
Wind speed (m/s)	1.724909	1.036300	0.890802	0.726080
Visibility (10m)	1436.825799	608.298712	-0.701666	-0.962116
Dew point temperature	4.073813	13.060369	-0.367236	-0.755683
Solar Radiation (MJ/m2)	0.569111	0.868746	1.503782	1.125105
Rainfall(mm)	0.148687	1.128193	14.530744	284.827774
Snowfall (cm)	0.075068	0.436746	8.439355	93.749107
Spring	0.252055	0.434217	1.142098	-0.695612
Summer	0.252055	0.434217	1.142098	-0.695612
Autumn	0.249315	0.432641	1.158924	-0.656894
Winter	0.246575	0.431042	1.175937	-0.617172
Holiday	0.049315	0.216537	4.162890	15.329651
No Holiday	0.950685	0.216537	-4.162890	15.329651
Functioning day	0.966324	0.180404	-5.170084	24.729765
Not Functioning day	0.033676	0.180404	5.170084	24.729765

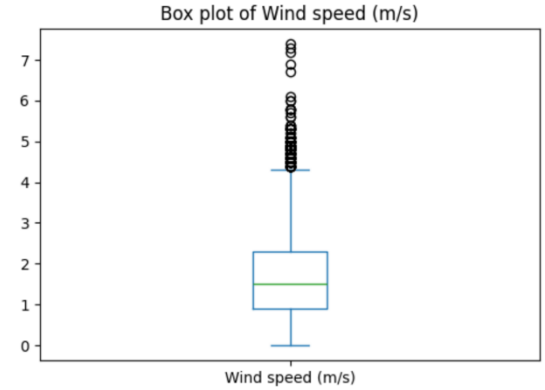
각 numerical 설명 변수들에 대한 box plot은 다음과 같다.

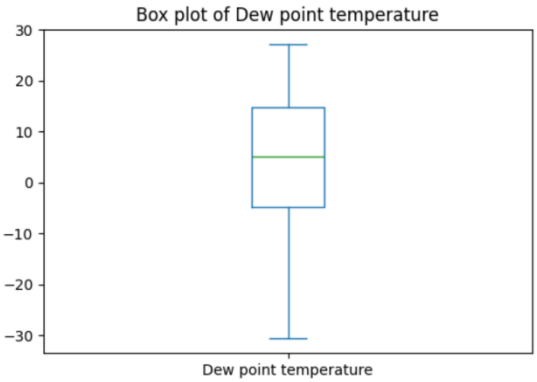


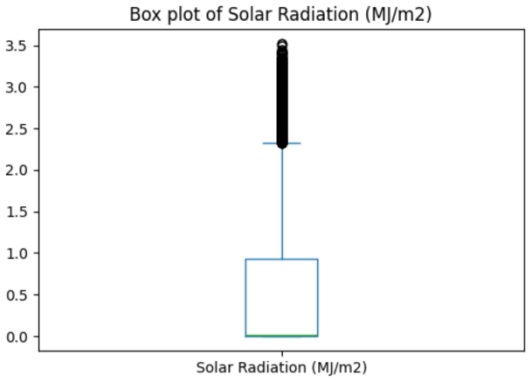


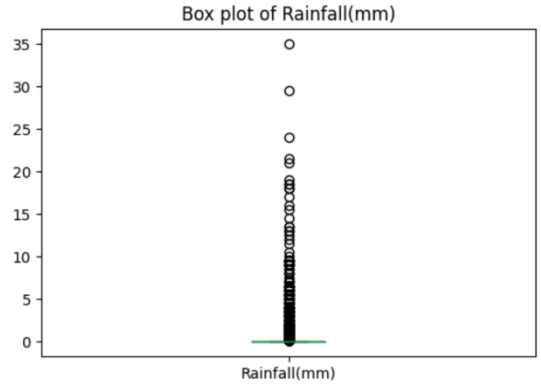


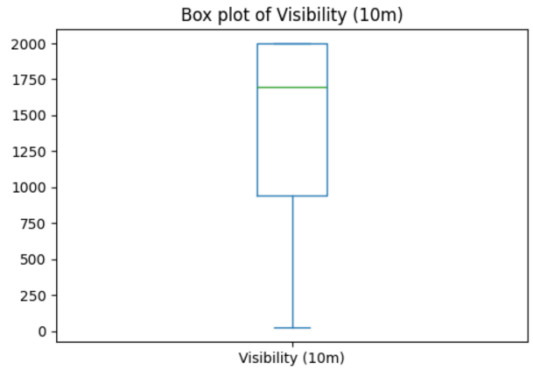












전체 변수들 중에서 정규분포를 따른다고 가정할 수 있는 변수는 습도, 풍속, 가시성, 이슬점이라고 생각했다.

이유는 대부분의 날 들에서 위의 변수들은 평균값을 띄고, 평균값에서 벗어나더라도 그 정도가 다른 변수들에 비해 미미하다고 생각했기 때문이다.

그러나 분포의 비대칭성을 측정하는 값인 왜도와 분포의 뾰족한 정도를 측정하는 값인 첨도를 관찰한 결과 전체 변수들 중에서 정규분포를 따를 수 있다고 가정할 수 있는 변수는 없다고 판단하였다.

이유는 왜도와 첨도가 각각 0과 3에 가까울수록 정규분포에 가깝게 되는데 이 두 조건을 동시에 만족하는 변수는 없었기 때문이다.

-정규성 검정

실제로 설명 변수들이 정규성을 띄는지 확인하기 위해 Shaprio-Wilks-test를 진행하였다.

이 테스트에 대한 귀무가설과 대립가설은 다음과 같다.

H_0: 데이터의 분포가 정규분포를 따른다.

H_1: 데이터의 분포가 정규분포를 따르지 않는다.

이 때, 귀무가설의 기각역 p-value는 0.05로 설정하였다.

	test_statistic	p_value	normality
Rented Bike Count	0.882217	6.957241e-63	False
Hour	0.950960	4.664639e-47	False
Temperature	0.980034	2.798836e-33	False
Humidity(%)	0.982362	1.424195e-31	False
Wind speed (m/s)	0.946810	1.999106e-48	False
Visibility (10m)	0.835091	1.365228e-69	False
Dew point temperature	0.966149	4.496012e-41	False
Solar Radiation (MJ/m2)	0.706303	6.620335e-82	False
Rainfall(mm)	0.112078	2.215472e-108	False
Snowfall (cm)	0.167351	9.700159e-107	False
Spring	0.540303	3.581378e-92	False
Summer	0.540303	3.581378e-92	False
Autumn	0.537982	2.723598e-92	False
Winter	0.535627	2.065333e-92	False
Holiday	0.221915	5.056045e-105	False
No Holiday	0.221915	5.056045e-105	False
Functioning day	0.172179	1.363558e-106	False
Not Functioning day	0.172179	1.363558e-106	False

Shaprio-Wilks-test를 수행한 결과 위와 같이 모든 변수의 p-value값이 0.05보다 작아 귀무가설을 기각한다.

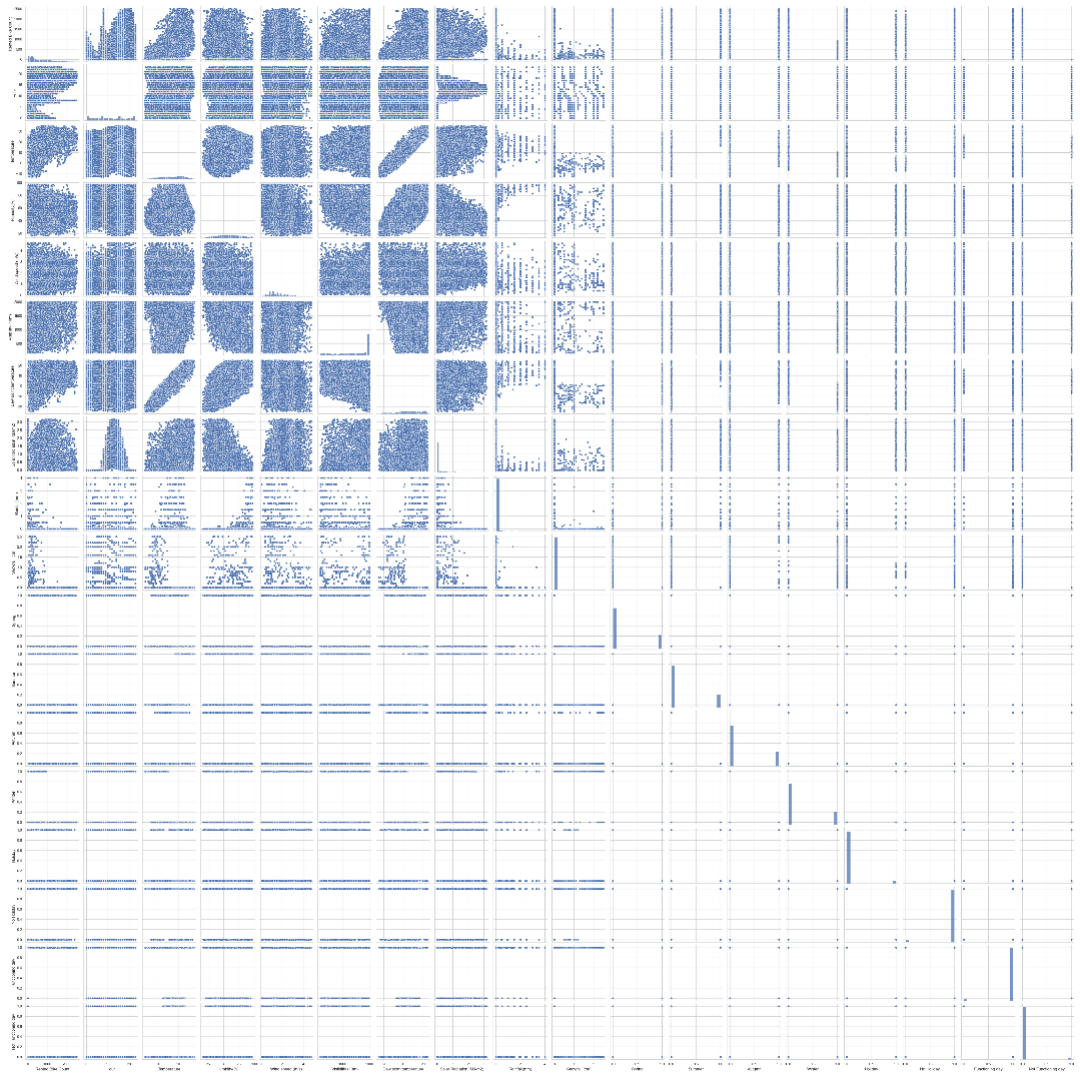
따라서 모든 변수들은 정규분포를 띄지 않는다고 결론 내릴 수 있다.

[Q4] 변수에 대한 이상치 확인 및 제거

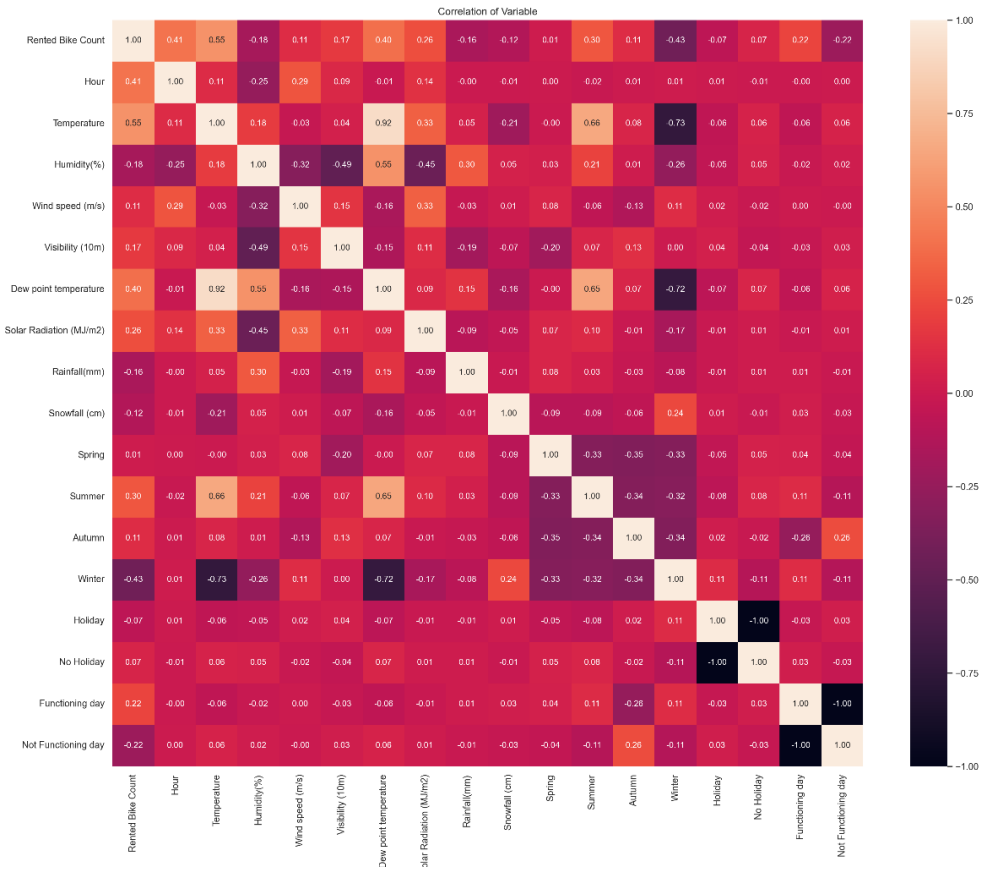
Box plot의 outlier 조건으로 상위 5%, 하위 5% 값을 가지는 데이터를 제거한 결과 기존 8760개의 데이터에서 각각 4027개로 줄어 모델을 충분히 학습하기에는 적합하지 않다는 판단이 들었다. 이에 상위 1%, 하위 1% 값을 가지는 데이터를 제거한 결과 총 7847개의 데이터가 남았고, 이는 적절하다고 판단하였다. 이후 이상치를 제거한 데이터프레임에서 결측치를 제거한 뒤 다음 step인 상관성 분석을 진행하였다.

[Q5] 상관성 분석

-Scatter plot



-Heat map



상관계수는 -1부터 1까지의 값을 가지며, 0에 가까울수록 두 변수의 상관관계가 약하고 상관계수의 절댓값이 1에 가까울수록 상관관계가 크다고 할 수 있다.

상관관계를 분석한 결과 기온(Temperature)와 이슬점(Dew point temperature)의 상관계수가 0.92로 가장 강한 상관관계를 나타냈다. 일반적으로 기온이 올라가면 대기 중의 수증기량이 많아져 그에 따라 이슬점 역시 증가하기에 상관계수가 높게 나온 것이다.

여름과 겨울의 경우 기온, 이슬점과 각각 강한 양의 상관관계, 음의 상관관계를 가진 것으로 나타났다.

여름은 기온이 높아 그에 따라 이슬점 역시 높아져 나타나는 결과이고, 반대로 겨울은 기온이 낮아 그에 따라 이슬점 역시 낮아져 나타나는 결과라고 해석할 수 있다

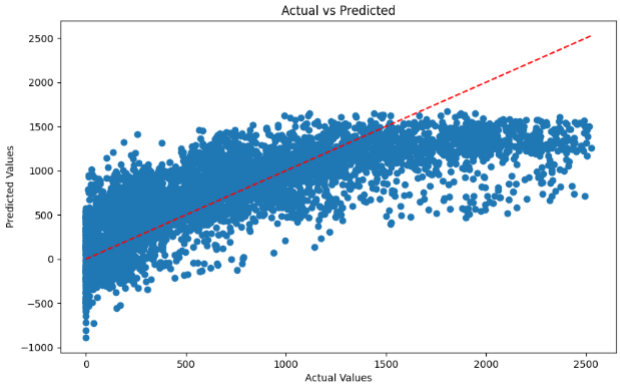
습도(Humidity)와 가시성(Visibility)의 경우 상관계수가 -0.49로 어느 정도 음의 상관관계를 띈다고 해석할 수 있다.

이는 습도가 올라갈수록 안개와 같은 가시성을 방해하는 요소가 발생할 가능성이 높아지기 때문이라고 해석할 수 있다.

[Q6] MLR 모델 학습

MLR 모델 학습을 위해 명목형 변수의 원핫인코딩, 이상치 제거, 결측치 제거를 함으로써 데이터를 가공하였다. 시간(Hour)에 따른 편차를 방지하고자 행을 기준으로 데이터를 랜덤하게 섞은 후 70%의 학습 데이터와 30%의 테스트 데이터로 나뉜 뒤 모델 학습을 진행하였다.

훈련한 모델을 통해 예측한 Predicted value값과 Actual value 값의 선형성을 확인하기 위해 plot을 그린 결과 다음과 같이 Predicted value값이 음수가 나오는 것을 확인할 수 있다.



선형회귀의 경우 종속변수의 범위에 대한 제한이 없으니 설명변수 값의 조합에 따라 음수가 나올 수 있기 때문이다.

그러나 빌린 자전거의 대수는 현실에서는 결코 음수가 될 수 없다. 모델링은 잘 되어있기에 모델링 부분에서 문제가 없다고 판단하였다. 이에 자전거의 대수가 음수가 나오는 경우는 그 만큼 자전거를 빌릴 가능성이 낮다는 것을 의미하기에 예측값을 0으로 조정하는 작업을 거쳤다.

[Q6-1] Adjusted R2을 통한 데이터의 선형성 판단

```
                            OLS Regression Results
==============================================================================
Dep. Variable:      Rented Bike Count   R-squared:                       0.573
Model:                            OLS   Adj. R-squared:                  0.572
Method:                 Least Squares   F-statistic:                     525.5
Date:                Mon, 01 Apr 2024   Prob (F-statistic):               0.00
Time:                        22:15:41   Log-Likelihood:                -40725.
No. Observations:                5492   AIC:                         8.148e+04
Df Residuals:                    5477   BIC:                         8.158e+04
Df Model:                          14
Covariance Type:            nonrobust
================================================================================================
                                   coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------------------------
const                          224.9769     67.408      3.338      0.001      92.830     357.123
Hour                            27.2946      0.852     32.025      0.000      25.624      28.965
Temperature                      6.8881      5.873      1.173      0.241      -4.626      18.402
Humidity(%)                    -12.6559      1.684     -7.515      0.000     -15.958      -9.354
Wind speed (m/s)                10.1610      6.316      1.609      0.108      -2.222      22.544
Visibility (10m)                -0.0125      0.012     -1.080      0.280      -0.035       0.010
Dew point temperature           21.3923      6.231      3.433      0.001       9.176      33.608
Solar Radiation (MJ/m2)        -57.2429      9.162     -6.248      0.000     -75.204     -39.282
Rainfall(mm)                  -225.7531     16.215    -13.922      0.000    -257.541    -193.965
Snowfall (cm)                   69.2221     23.863      2.901      0.004      22.441     116.003
```

```
Spring                         93.9919     20.275      4.636      0.000      54.245     133.739
Summer                         46.1902     24.330      1.898      0.058      -1.507      93.887
Autumn                        232.1934     19.897     11.670      0.000     193.188     271.199
Winter                       -147.3985     23.486     -6.276      0.000    -193.439    -101.358
Holiday                        63.9240     37.079      1.724      0.085      -8.766     136.614
No Holiday                    161.0529     34.964      4.606      0.000      92.510     229.596
Functioning day               593.1326     35.969     16.490      0.000     522.619     663.646
Not Functioning day          -368.1557     38.351     -9.600      0.000    -443.339    -292.973
==============================================================================
Omnibus:                      622.301   Durbin-Watson:                   1.990
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              940.273
Skew:                           0.833   Prob(JB):                    6.64e-205
Kurtosis:                       4.154   Cond. No.                     3.67e+19
==============================================================================
```

Adjusted R-Squared 값이 0.572로 설명변수와 종속변수 간에 강하지는 않지만 어느 정도의 선형 관계가 있다고 할 수 있다.

[Q6-2] Ordinary Least Square 방식의 Solution이 만족해야 하는 가정

Ordinary Least Square 방식의 solution은 아래의 가정들을 만족해야한다.

1. 오차항 ε 이 정규분포를 따른다.

2. 설명변수와 종속변수 사이에 선형관계가 성립한다.

3. 각 관측치들은 서로 독립이다.

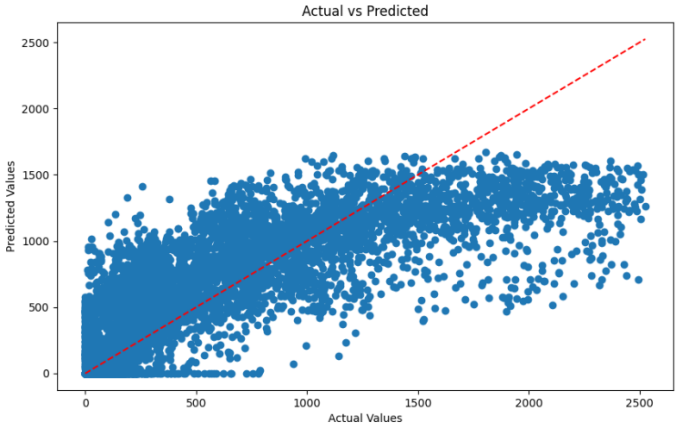
4. 종속변수 Y에 대한 오차항(residual)은 설명변수 값의 범위에 관계없이 일정하다 (homoscedasticity)

위의 가정 중 3번 가정은 현실 데이터의 경우 위배되는 경우가 많고, 확인하기 어렵다.

따라서 1번, 2번, 4번 가정을 데이터가 만족하는지 정성적으로 판단하겠다.

1번 가정은 Residual distribution, QQ plot을 통해 4번 과정은 Residual plot을 그려 확인하도록 하겠다.

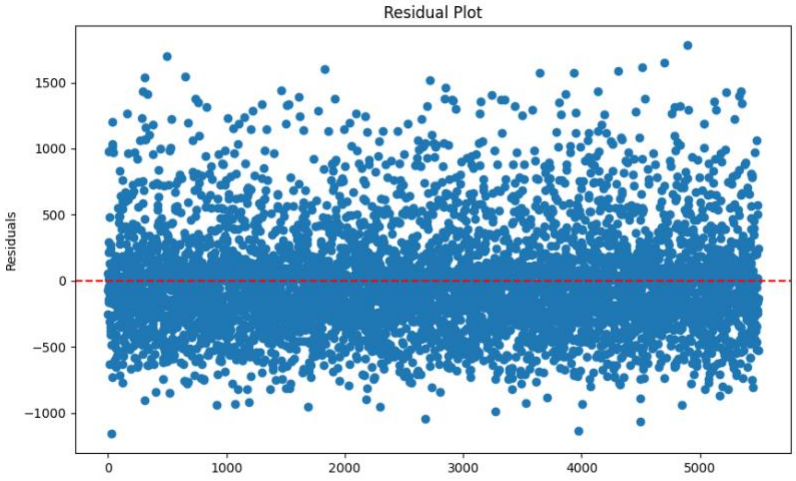
-설명 변수와 종속 변수의 선형성

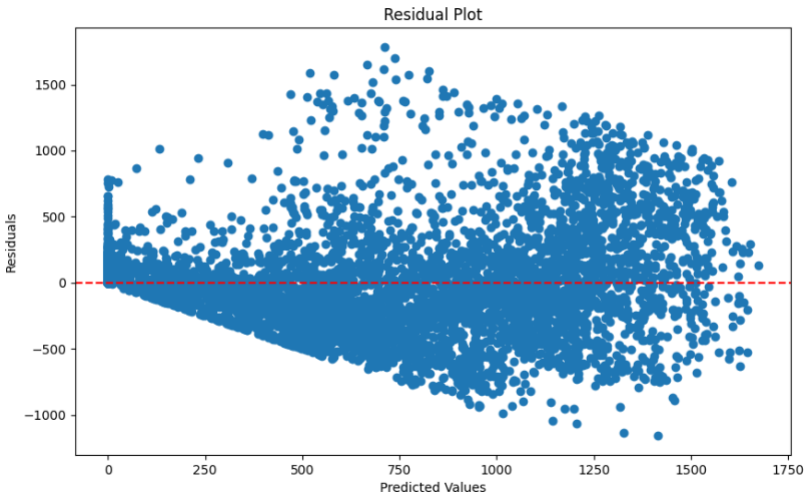


위 plot을 보면 알 수 있듯이 설명변수와 종속변수간 linearity가 있다는 것을 확인할 수 있다.

특히 실제 자전거 빌린 대수는 0이상이고, 이에 따라 예측값이 음수인 경우 의미가 없다고 판단해 0으로 처리하여 위와 같은 그림의 plot이 나오게 되었다.

-Residual plot

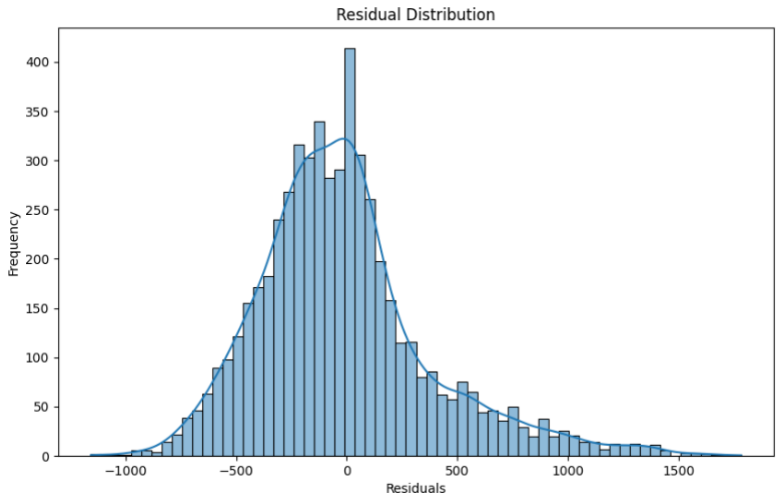


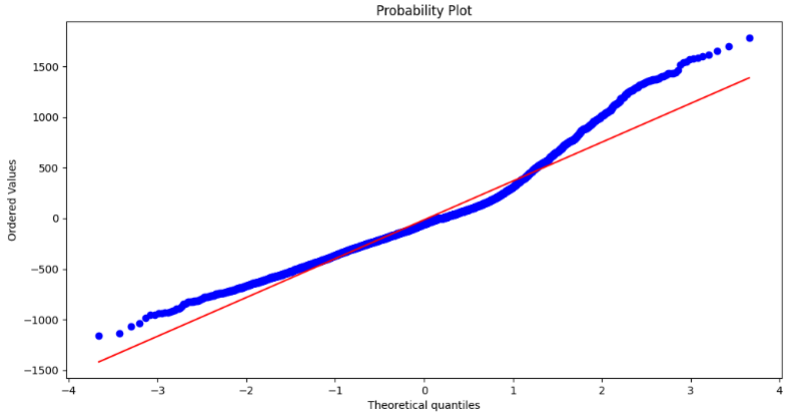


Residual plot을 보면 알 수 있듯이 잔차들이 특정 분포를 띄는 것이 아닌 random하게 분포하는 것을 알 수 있다.

따라서 등분산성(homoscedasticity을 위배하지 않음을 확인할 수 있다.

--Residual distribution, QQ plot





Residual distribution은 residual의 분포를 나타내고, QQ plot은 데이터가 얼마나 정규 분포를 띄는지 확인할 수 있는 그래프이다.

Residual distribution의 경우 왼쪽으로 약간 skewed된 분포를 띔을 알 수 있었다.

QQ plot에서는 데이터가 평균에서 멀어짐에 따라 직선에서 크게 벗어나는 형태를 띄고 있다.

이를 통해 잔차는 정규성을 띄지 않는다고 결론 내릴 수 있다.

따라서 Residual plot을 통해 등분산성을 만족함을 확인하였지만, Residual distribution과 QQ Plot을 통해 잔차의 정규성은 확인할 수 없었다.

[Q7] 유의미한 변수 판단

Q6에 첨부한 statsmodels 라이브러리에서 제공한 model의 Summary 결과를 참고하면 유의미하지 않은 변수는 p-value 값이 0.241인 기온(Temperature)와 0.108인 풍속(Wind speed)가 있다. 이 외에도 p-value 값이 0.280인 가시성(Visibility), 0.058인 여름(Summer), 0.085인 휴일(Holiday)가 유의미하지 않다. 그 외의 변수들은 p-value값이 0.01보다 작기에 유의미한 변수들이라고 할 수 있다.

유의미한 변수들 중 coef 값이 양의 값을 가지면 종속 변수와 양의 상관관계를 음의 값을 가지면 음의 상관관계가 있다고 할 수 있다. 이를 정리하면 아래와 같다.

-양의 상관관계

'Hour', 'Dew point temperature', 'Snowfall (cm)', 'Spring', 'Autumn', 'No Holiday', 'Functioning day'

-음의 상관관계:

'Humidity(%)', 'Solar Radiation (MJ/m2)', 'Rainfall(mm)', 'Winter', 'Not Functioning day'

[Q8] MLR 모델 성능 평가

[Q8] Test 데이터셋에 대하여 MAE, MAPE, RMSE를 계산하고 그에 대한 해석을 해 보시오.

	RMSE	MAE	MAPE
Seoul bike rent	393.5341	292.2926	2.255769e+16

각 평가지표는 유효숫자 4자리로 통일하였으나 MAPE는 불가능하여 위와 같이 표기하였다.

MAE란 평균 절대 오차로 실제 값과 예측 값 사이의 절대적인 오차의 평균을 측정한 것을 말한다. MAE 값이 292.2926의 값으로 실제 값과 예측 값 사이에 약 292회의 오차가 있는 것을 확인할 수 있다.

MAPE란 평균 절대 비율 오차로 위와 같이 매우 큰 비율이 나왔다. 이는 실제 자전거 대여량이 0인 경우가 존재하기에 발생하는 현상이라고 해석할 수 있다.

이 모델에서는 MAE와 MAPE 중 MAE를 우선적으로 고려해야 한다. 상대적 차이보다 절대적 차이가 중요한 분야이기도 하고, Not Functional Day일 경우 자전거 대여량이 0이기 때문에 MAPE값이 매우 크게 나오기 때문이다.

RMSE란 예측값과 실제값의 차이를 제곱한 뒤에 그것들은 평균 내어 루트를 씌운 값을 말한다. 오차에 대해 제곱 연산을 수행하기에 더 큰 오차에 대해 더 큰 penalty를 부여하는 것을 알 수 있다. RMSE의 값은 393.5341로 약 394회의 오차가 있다고 할 수 있다.

[Q9] 변수 재선정

우선 p value 값이 0.01을 넘어가는 풍속(Wind speed)과 가시성(Visibility)을 제외한다. 이후 Q5에서 알 수 있듯이 기온(Temperature)과 이슬점(Dew point temperature)은 강한 상관관계를 가졌기에 이슬점(Dew point temperature) 역시 제외한다. 여름, 겨울의 경우에도 기온(Temperature) 및 이슬점(Dew point temperature)과 강한 상관관계를 가진 것을 확인할 수 있었다. 따라서 동일한 이유로 여름, 겨울도 제외한다. 이 때, 사계절에 따라 변화하는 요소가 바론 기온(Temperature)이므로 나머지 계절인 봄, 가을을 제외시키더라도 기온(Temperature)으로 충분히 설명할 수 있을 것이라 판단하여 사계절 모두 제외시켰다.

Functioning Day, Holiday, Rainfall, Snowfall 변수도 제외시켰다. 이유는 EDA(Exploratory Data Analysis) 과정에서 Functioning Day, Holiday 데이터 비중이 각각 약 96.6%, 4.9%로 값이 한 쪽으로 지나치게 편향되어 의미가 없다고 판단하였기 때문이다.

```
Rainfall이 0인 비율: 0.9397260273972603
Snowfall이 0인 비율: 0.9494292237442923
```

동일한 이유로 Rainfall과 Snowfall도 제외시켰다. 실제 Snowfall을 제외시키기 전에는 앞서 살펴봤듯이 상식과 다른 양의 상관관계를 띄는 것을 확인할 수 있기도 하였다.

이후 앞서 언급한 10개 변수를 제외한 나머지 변수들에 대해 다중공산성을 확인하였다. 다중공산성을 확인하는 이유는 회귀 모델에서 설명 변수 간의 상관 관계로 인해 오차 항의 분산이 증가하고 모델의 해석이 어려워질 수 있기 때문이다. 일반적으로 VIF 값이 10보다 큰 경우 다중공산성이 있다고 말한다. 계산한 다중공산성은 다음과 같다.

	Variable	VIF
0	const	55.540062
1	Rented Bike Count	1.857437
2	Hour	1.335932
3	Temperature	2.141695
4	Humidity(%)	2.429305
5	Wind speed (m/s)	1.255493
6	Visibility (10m)	1.473123
7	Solar Radiation (MJ/m2)	1.860341

기존 17개의 변수에서 10개의 변수를 제외하고 계산한 다중 공산성은 위와 같다. Q7에서 p-value 값이 0.241인 기온(Temperature)을 제외하지 않은 이유는 기온(Temperature)이 일반적으로 생각하였을 때 가장 직접적으로 실외 활동에 영향을 미치는 요소라고 생각했기 때문이다. 실제로 변수를 제거하기 전과 후의 기온(Temperature)의 다중공산성은 약 152.95에서 2.14로 낮아지는 것을 확인할 수 있는데 이는 기온(Temperature)과 직접적으로 관련된 계절(Seasons), 이슬점(Dew point temperature) 등을 변수에서 제외시켰기 때문에 나타나는 현상이라고 해석할 수 있다.

[Q10] MLR 모델 재학습 및 성능평가

Q9에서 선택한 변수들을 가지고 MLR 모델을 재학습한 결과는 아래와 같다.

```
                            OLS Regression Results
==============================================================================
Dep. Variable:      Rented Bike Count   R-squared:                       0.459
Model:                            OLS   Adj. R-squared:                  0.458
Method:                 Least Squares   F-statistic:                     775.6
Date:                Tue, 02 Apr 2024   Prob (F-statistic):               0.00
Time:                        21:59:54   Log-Likelihood:                -41336.
No. Observations:                5492   AIC:                         8.269e+04
Df Residuals:                    5485   BIC:                         8.273e+04
Df Model:                           6
Covariance Type:            nonrobust
================================================================================================
                                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------------------------
const                        470.4007     44.355     10.605      0.000     383.447     557.354
Hour                          25.5864      0.949     26.973      0.000      23.727      27.446
Temperature                   31.5927      0.656     48.174      0.000      30.307      32.878
Humidity(%)                   -7.8220      0.475    -16.482      0.000      -8.752      -6.892
Wind speed (m/s)              -0.2371      6.979     -0.034      0.973     -13.919      13.444
Visibility (10m)               0.0152      0.012      1.241      0.215      -0.009       0.039
Solar Radiation (MJ/m2)      -61.7679      9.951     -6.207      0.000     -81.277     -42.259
==============================================================================
Omnibus:                      363.674   Durbin-Watson:                   1.994
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              594.448
```

Adj R-squared 값은 기존 0.572에서 0.458로 소폭 감소한 것으로 확인할 수 있다. 그리고 Wind speed와 Visibility의 경우 p-value값을 0.01로 설정하였을 때 이보다 큰 값을 가지기에 유의미한 변수가 아닌 것을 알 수 있다.

	RMSE	MAE	MAPE
Seoul bike rent	449.3066	336.5744	1.325941e+17

새로운 데이터 셋으로 훈련한 모델의 RMSE, MAE, MAPE값은 위와 같다.

MAPE는 앞서 살펴봤듯이 실제 자전거 대여량이 0인 경우가 존재하여 분모가 0으로 가기 때문에 무한대에 가까운 값을 가져 발생하는 현상이라고 해석할 수 있다. RMSE와 MSE의 값 모두 기존 393.5341, 292.2926에서 증가한 것을 확인할 수 있다.

종합하여 말하자면 설명 변수의 개수가 감소함에 따라 설명력이 감소하고, 오차가 늘어나기 때문에 발생하는 현상이라고 할 수 있다. 그러나 설명 변수 16개 중 10개를 줄였음에도 불구하고 Adj R-squared 값은 기존 0.572에서 0.458로 소폭 감소한 것을 통해 어느 정도 유의미한 변수를 추출하였다고 해석할 수 있다.

[Extra Question] 이 외 해당 데이터셋을 통해 MLR 관점에서 가능한 추가적인 분석을 웹에서 검색해서 수행하고 그 결과를 해석해 보시오

추가적으로 선형 회귀 모델에 규제를 추가한 LASSO 회귀 분석을 해봤다. 이유는 특성의 개수가 17개로 다소 많다고 생각하여 훈련 세트에 과대 적합될 가능성이 있다고 생각했기 때문이다. 처음 하이퍼파라미터인 alpha 값을 0.01 max_iter를 1000으로 설정하고 LASSO 회귀 분석을 진행한 결과 아래와 같은 경고 메시지가 발생하게 되었다.

ConvergenceWarning: Objective did not converge. You might want to increase the number of iterations, check the scale of the features or consider increasing regularisation.

이는 LASSO 모델이 최적의 계수를 찾기 위해 반복적인 계산을 수행하는데, 지정한 반복 횟수가 부족할 때 다음과 같은 메시지가 발생한다고 한다. 그래서 max_iter 값을 10000으로 변경하였으나 여전히 동일한 경고 메시지가 발생하였고, 이에 규제 강도인 alpha 값을 1.0으로 변경하여 문제를 해결할 수 있었다.

	RMSE	MAE	MAPE
Seoul bike rent	396.749	301.8521	4.068180e+16

LASSO 모델을 훈련시켜 구한 RMSE, MAE, MAPE 값은 위와 같았다. 실제 데이터 값이 0이기 때문에 수치가 지나치게 높은 MAPE는 해석에서 제외하겠다. RMSE, MAE 값은 선형 회귀 방식으로 구한 393.5341, 292.2926보다 소폭 증가한 값을 가지고 있었다.

하이퍼파라미터 값인 알파를 임의로 1로 지정하였는데 훈련 세트 그래프와 테스트 세트 그래프의 알파 값에 따른 R^2 값의 변화 추이를 보고 최적의 알파 값을 찾아 모델을 다시 훈련한다면 결과가 더 좋게 나올 것이라고 생각된다.