자전거 이용시간 예측

FASTCAMPUS
DataScience School
신경철

Contents

Exploratory Data Analysis

Continuous Variables

	tripduration	age	Precipitation_In
count	146168	146164	146168
mean	592.985675	37.120495	0.07358
std	731.555415	10.154947	0.182269
min	60.008	18	0
25%	324.20525	30	0
50%	479.152	34	0
75%	693.395	42	0.04
max	27985.884	86	2.2

Exploratory Data Analysis

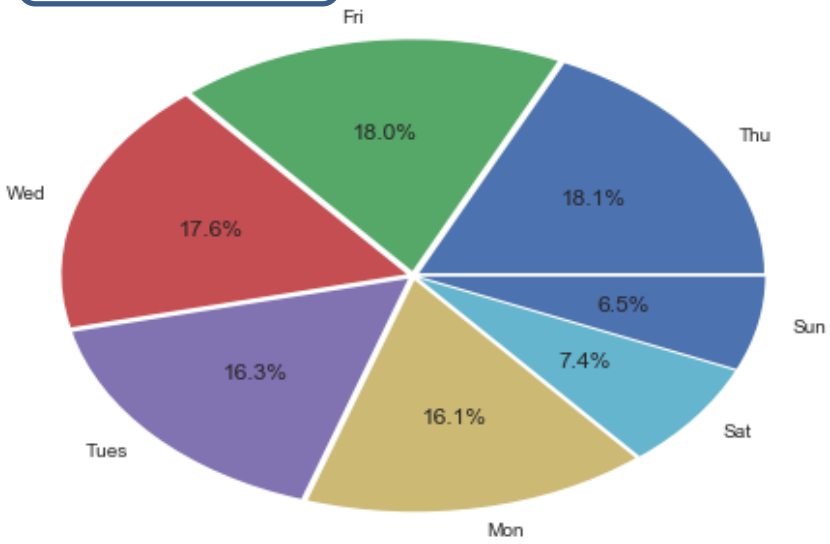
Continuous Variables

	Mean_ Temperature_F	Mean_ Dew_Point_F	Mean_ Humidity	Mean_ Sea_Level_Pressure_In	Mean_ Visibility_Miles	Mean_ Wind_Speed_MPH
count	146122	146168	146168	146168	146168	146168
mean	58.743365	46.065938	66.177775	30.040247	9.530178	4.38542
std	10.143713	7.475085	12.568947	0.176833	1.058423	2.529944
min	33	4	24	29.31	1	0
25%	50	42	57	29.94	10	3
50%	59	47	67	30.04	10	4
75%	66	52	76	30.15	10	6
max	83	59	95	30.81	10	23

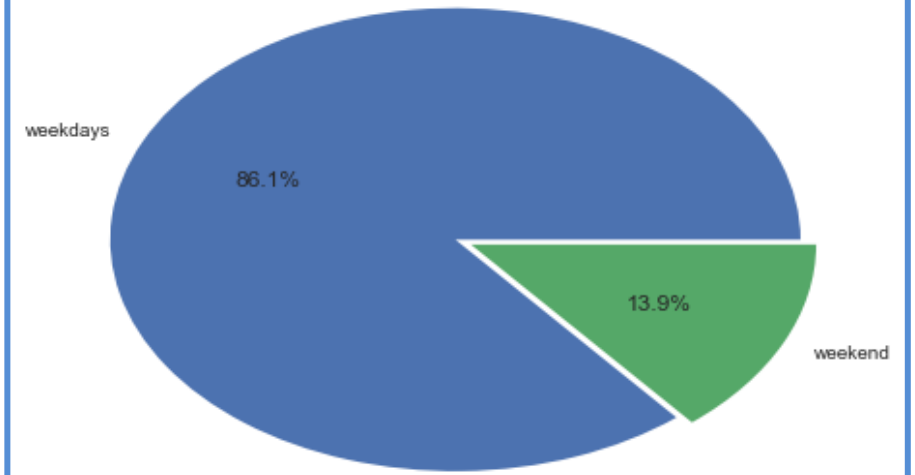
Exploratory Data Analysis

Categorical Variables

Day



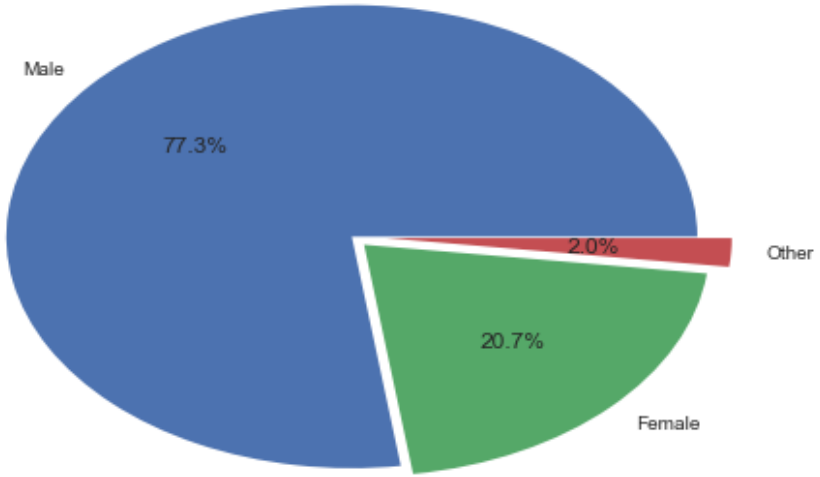
Weekends



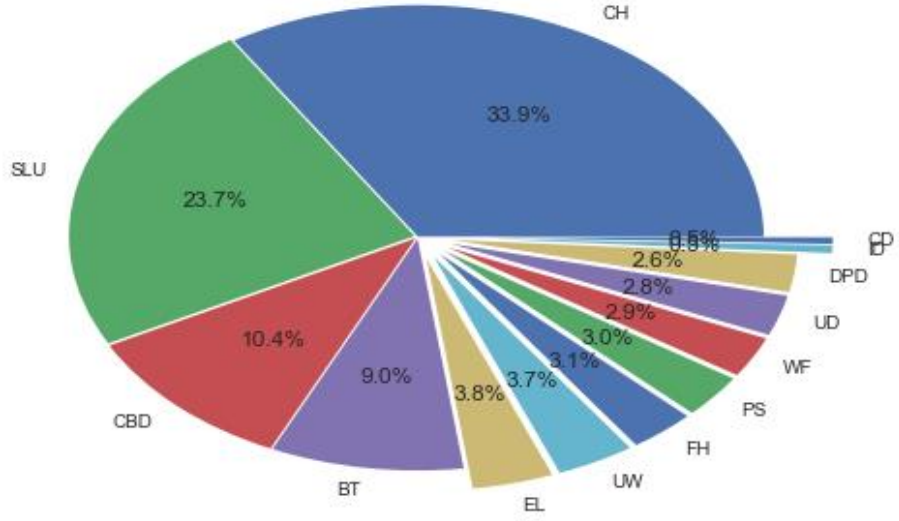
Exploratory Data Analysis

Categorical Variables

Gender



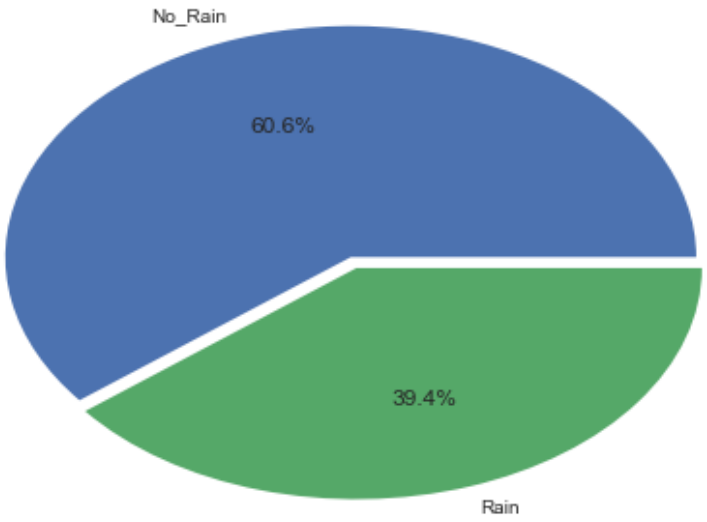
From Station Area



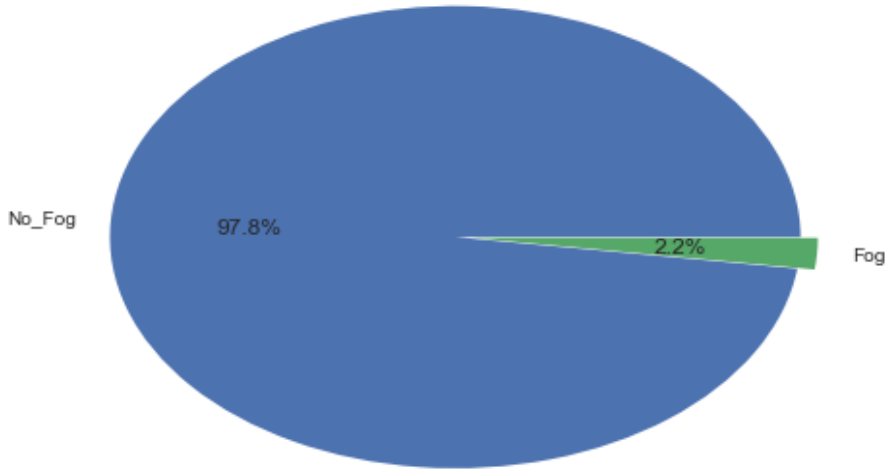
Exploratory Data Analysis

Categorical Variables

Rain



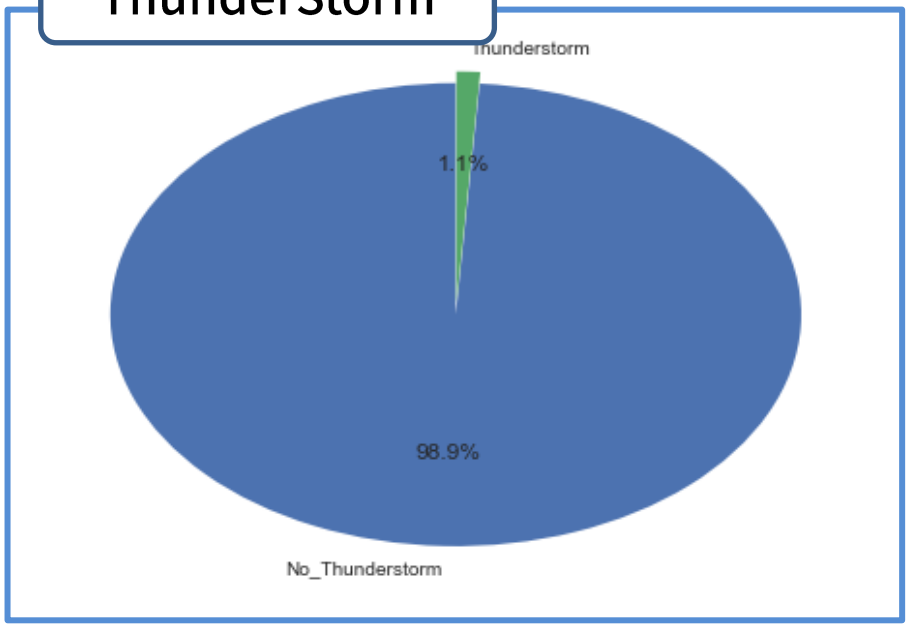
Fog



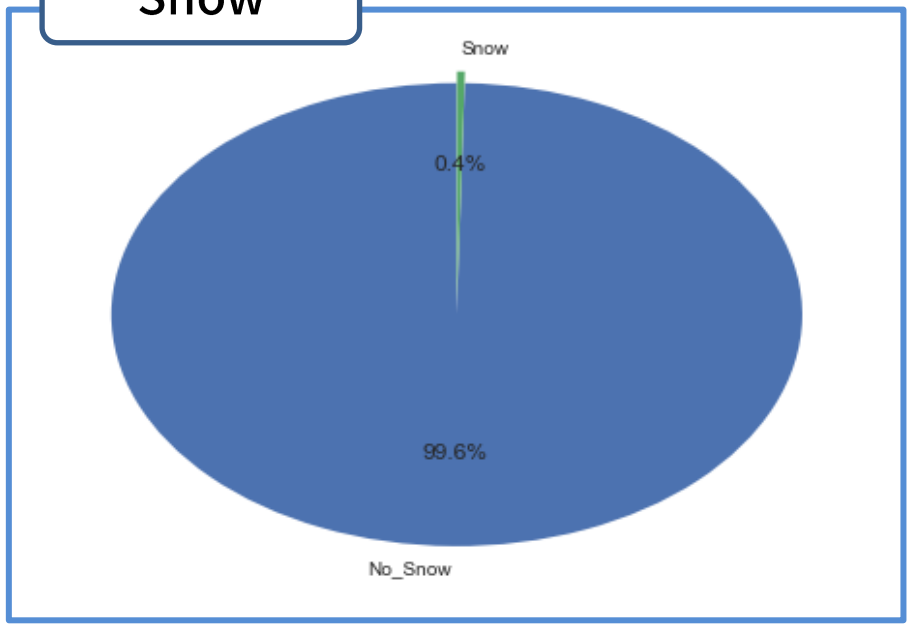
Exploratory Data Analysis

Categorical Variables

ThunderStorm



Snow



Regression

Significant Variable
P < 0.05

'const', 'age', 'Mean_Sea_Level_Pressure_In', 'day', 'weekend', 'Gender', 'from_station_area', 'Rain', 'Mean_Wind_Speed_MPH'

Model Test

- R-squared: 0.072
- Adj. R-squared: 0.071
- F-statistic: 395.1
- Prob (F-statistic): 0.00

Residual diagnostics

- Durbin-Watson: 1.993
- Prob(Omnibus): 0.000
- Prob(JB): 0.00
- Cond. No. : 6.77e+16
- Prob(LM-test): 2.44e-165

mean_absolute_error : 266.89

Lasso Regression + Adaboost

Coef = 0
No Adaboost,
Only Lasso

'Mean_Sea_Level_Pressure_In', 'Precipitation_In',
'Gender_O', 'from_station_area[CD]',
'from_station_area[DPD]', 'Snow', 'Fog', 'Thunderstorm'

Lasso Regression
+
Adaboost

Alpha

0.00068578132181214193

mean_absolute_error : 269.47

Decision Tree

Feature Importance by Extremly Random Trees

Information Gain을 최대로 하는 9개의 변수를 선택

Feature_Importance	
Age	0.391617
Mean_Sea_Level_Pressure_In	0.090345
Mean_Humidity	0.081532
Mean_Temperature_F	0.073519
MeanDew_Point_F	0.072018
Mean_Wind_Speed_MPH	0.067069
Day	0.051533
Precipitation_In	0.046344
Mean_Visibility_Miles	0.029952

Decision Tree

Input Variable	'Mean_Sea_Level_Pressure_In', 'Mean_Humidity', 'Mean_Temperature_F', 'Mean_Visibility_Miles', 'MeanDew_Point_F', 'Mean_Wind_Speed_MPH', 'day', 'Precipitation_In', 'Rain' 'age',

Many Selected Features	Mean Absolute Error
'Mean_Temperature_F', 'Mean_Humidity', 'age', 'Mean_Visibility_Miles'	288.68

Conclusion

- 낮은 Performance(266.89 ~ 288.68)
- Regression에서의 문제점
  - 기본적으로 독립변수와 종속변수 간 선형상관관계가 매우 약함.
    - 낮은 상관계수, R-squared: 0.072, Adj. R-squared: 0.071
  - 잔차가 정규분포가 아니며 이분산성을 띄고 있기 때문에, 분석 결과의 신뢰도가 떨어짐
    - Prob(Omnibus): 0.000, Prob(JB): 0.00, Prob(LM-test): 2.44e-165
  - 3333444
- Decision Tree에서의 문제점
  - 유용한 변수가 적음
    - Feature Importance에서의 낮은 Information Gain값
    - Tree Graph에서 age 변수만 너무 자주 이용됨.