IntroBA
Midterm Team Project

Analyzing the data ‘Bike Seoul Sharing’

류승권 장해영 김요섭 남윤서 이주은

Data Analysis Objective

How? Using ‘Seoul Bike Sharing’ Data analyze
What? To find out which variables affect the Number of Bicycles Rented

Data

·Data Source: Kaggle
·Range Index: 8760 entries
·Data columns: Total 14 columns

·Able to identify weather information
: Temperature, Humidity, Windspeed, Visibility, Dewpoint(Temperature), Solar radiation, Rainfall Snowfall

·This data contains information about the number of bikes rented per hour and date information

	Date	Rented Bike Count	Hour	Temperature(° C)	Humidity(%)	Wind speed (m/s)	Visibility (10m)	Dew point temperature(° C)	Solar Radiation (MJ/m2)	Rainfall(mm)	Snowfall (cm)	Seasons	Holiday	Functioning Day
0	01/12/2017	254	0	-5.2	37	2.2	2000	-17.6	0.0	0.0	0.0	Winter	No Holiday	Yes
1	01/12/2017	204	1	-5.5	38	0.8	2000	-17.6	0.0	0.0	0.0	Winter	No Holiday	Yes
2	01/12/2017	173	2	-6.0	39	1.0	2000	-17.7	0.0	0.0	0.0	Winter	No Holiday	Yes
3	01/12/2017	107	3	-6.2	40	0.9	2000	-17.6	0.0	0.0	0.0	Winter	No Holiday	Yes
4	01/12/2017	78	4	-6.0	36	2.3	2000	-18.6	0.0	0.0	0.0	Winter	No Holiday	Yes

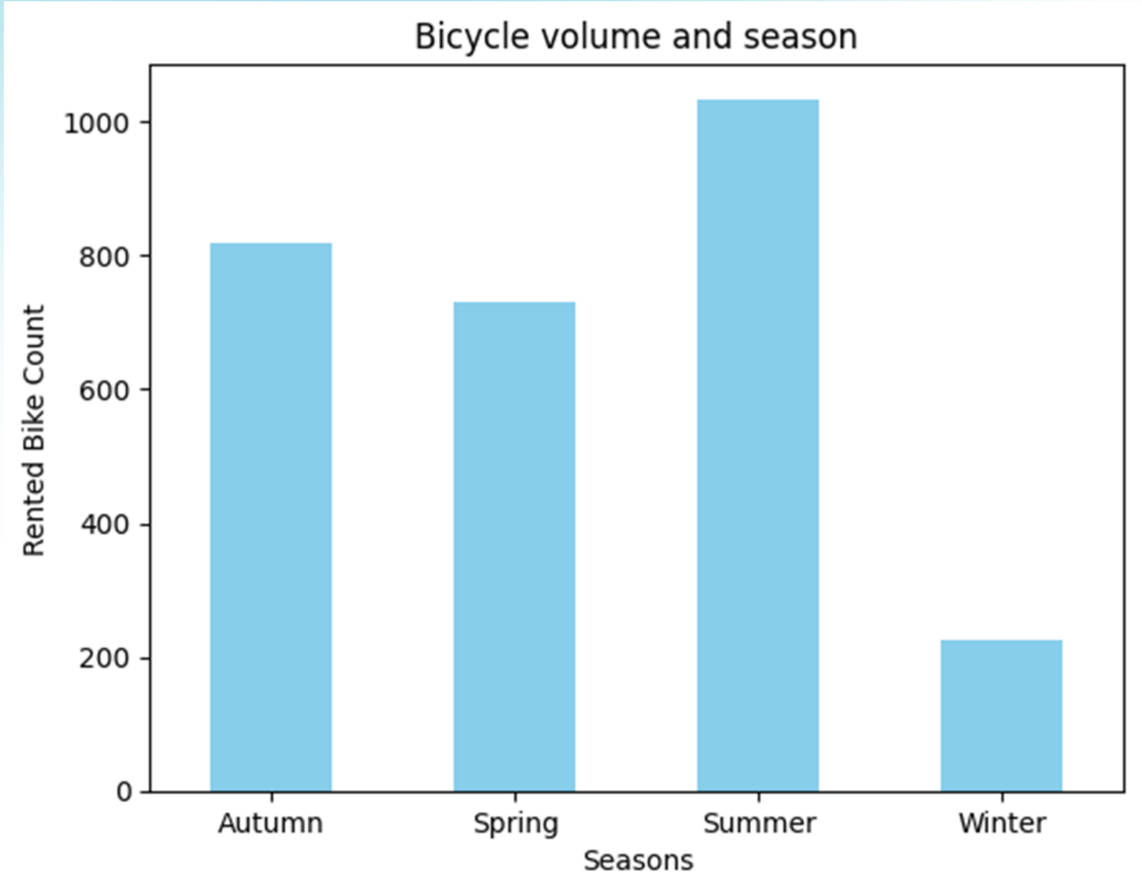
Data Exploration
1. Analysis of the relationship between season and Bike Volume

```
season_order = ['Spring', 'Summer', 'Autumn', 'Winter']

# Average calculation of bicycle volume by season
seasonal_data = df.groupby('Seasons')['Rented Bike Count'].mean()

# Visualize with bar graphs
seasonal_data.plot(kind='bar', color='skyblue')
plt.title("Bicycle volume and season")
plt.xlabel("Seasons")
plt.ylabel("Rented Bike Count")
plt.xticks(rotation=0)
plt.show()
```

Q. In what season do most bicycle volumes take place?

Kind=‘bar’
X label: “Seasons”
Y label: “Rented Bike Count”

1. Analysis of the relationship between season and Bike Volume



・The relationship between variable ‘Seasons’ and ‘Rented Bike Count’

2. Simple Regression model after identifying correlation between variables

```python
import statsmodels.api as sm

# Set 'Temperature(° C)' as an independent variable
x = df['Temperature(° C)']

# Set 'Rented Bike Count' as a dependent variable
y = df['Rented Bike Count']

x = sm.add_constant(x)

model = sm.OLS(y, x)

results = model.fit()

print(results.summary())
```

- Set Temperature related with seasons as an independent variable

- Set ‘Rented Bike Count’ as a dependent variable

2-1 Regression Analysis Results

OLS Regression Results

Dep. Variable:	Rented Bike Count	R-squared:	0.290
Model:	OLS	Adj. R-squared:	0.290
Method:	Least Squares	F-statistic:	3578.
Date:	Wed, 25 Oct 2023	Prob (F-statistic):	0.00
Time:	22:08:40	Log-Likelihood:	-67600.
No. Observations:	8760	AIC:	1.352e+05
Df Residuals:	8758	BIC:	1.352e+05
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	329.9525	8.541	38.631	0.000	313.210	346.695
Temperature(° C)	29.0811	0.486	59.816	0.000	28.128	30.034

Omnibus:	954.681	Durbin-Watson:	0.271
Prob(Omnibus):	0.000	Jarque-Bera (JB):	1421.965
Skew:	0.817	Prob(JB):	1.68e-309
Kurtosis:	4.108	Cond. No.	25.9

- P-value: 0.000

-> It means that the p-value is smaller than 0.05, so it affects meaningful effect on dependent variable.

- Coef(coefficient) of the 'Temperature' : 29.0811

- R-squared: 0.0290

2-2 Regression Model Visualization

```
import matplotlib.pyplot as plt
import seaborn as sns

plt.figure(figsize=(12, 8))

plt.scatter(df['Temperature(° C)'], df['Rented Bike Count'], label='Real Data', alpha=0.5)
```

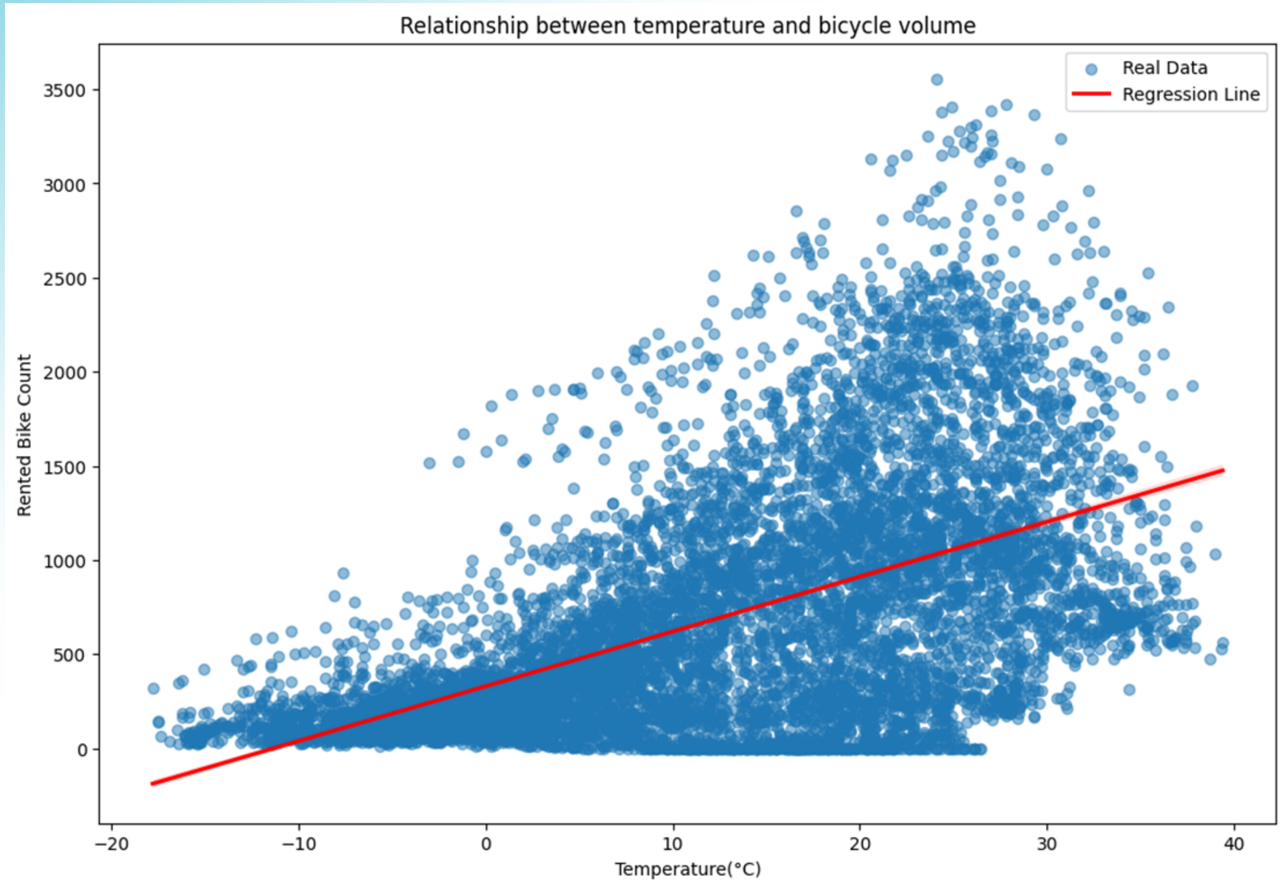
・Preparing visualization of the regression model by using ‘scatter plot’ to show the relationship between ‘Temperature’ and ‘Bicycle Volume’

2-2 Regression Model Visualization

```
# 회귀선 추가
sns.regplot(x='Temperature(° C)', y='Rented Bike Count', data=df, scatter=False, color='red', label='Regression Line')

plt.xlabel('Temperature(° C)')
plt.ylabel('Rented Bike Count')
plt.title('Relationship between temperature and bicycle volume')
plt.legend()
plt.show()
```

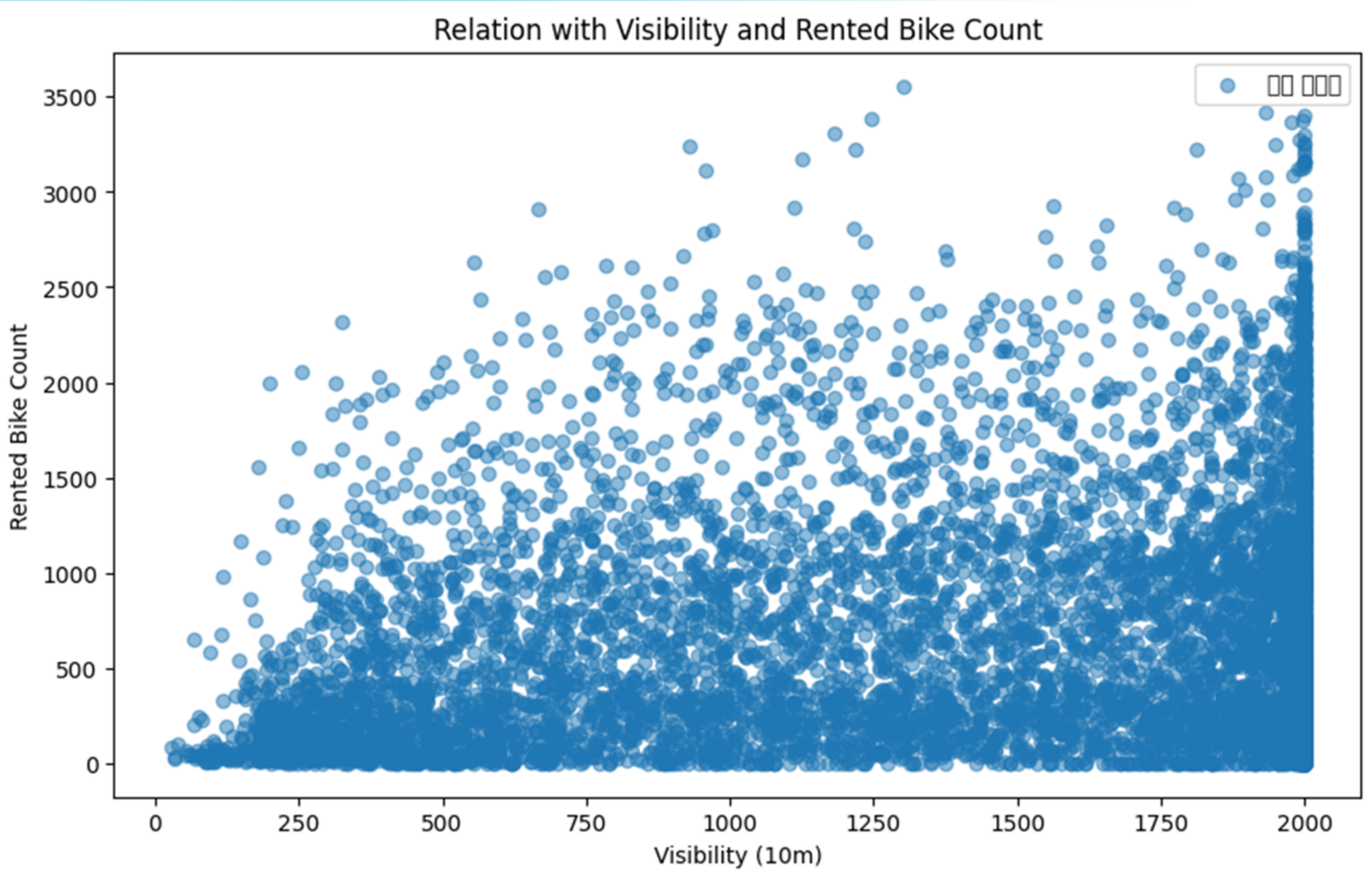
- Adding Regression Line to show how ‘Bicycle Volume’ changes as the ‘Temperature’ increases.

- **The Relationship between ‘Temperature’ and ‘Rented Bike Count’**



- The higher the ‘Temperature’, the higher the ‘Rented Bike Count’
- This Scatter Plot shows the relationship between ‘Temperature’ and ‘Bicycle volume’

- **The Relationship between ‘Visibility’ and ‘Rented Bike Count’**



- Scatter Plot
- X label: Visibility(10m)
- Y label: Rented Bike Count

3 Regression Analysis

- Set ‘Visibility’ as an independent variable
- Set ‘Rented Bike Count’ as a dependent variable

OLS Regression Results

Dep. Variable:	Rented Bike Count	R-squared:	0.040
Model:	OLS	Adj. R-squared:	0.040
Method:	Least Squares	F-statistic:	362.2
Date:	Wed, 25 Oct 2023	Prob (F-statistic):	3.67e-79
Time:	22:08:42	Log-Likelihood:	-68923.
No. Observations:	8760	AIC:	1.378e+05
Df Residuals:	8758	BIC:	1.379e+05
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	400.9966	17.324	23.147	0.000	367.038	434.955
Visibility (10m)	0.2113	0.011	19.031	0.000	0.190	0.233

Omnibus:	1323.474	Durbin-Watson:	0.202
Prob(Omnibus):	0.000	Jarque-Bera (JB):	2014.445
Skew:	1.096	Prob(JB):	0.00
Kurtosis:	3.848	Cond. No.	4.00e+03

- P-value: 0.000

-> It means that the p-value is smaller than 0.05, so it affects meaningful effect on dependent variable.

- Coef(coefficient) of the Visibility: 0.2113

- R-squared: 0.040

3-1 The relationship between ‘Visibility’ and ‘Humidity’

·Predicted Humidity was the most influential factor in visibility

```
correlation = df['Visibility (10m)'].corr(df['Humidity(%)'])
print(f"Pearson Correlation between Visibility and Humidity: {correlation}")
```

```
Pearson Correlation between Visibility and Humidity: -0.5430903446558321
```

-> In fact, there was a negative correlation between the two variables

3-2 Multiple Regression Analysis

·After analyzing the relationship between Visibility and Humidity, we did Multiple Regression Analysis setting Visibility and Humidity as independent variables.

```
import statsmodels.api as sm

# 가시거리와 습도를 독립 변수로 선택
x = df[['Visibility (10m)', 'Humidity(%)']]

x['Visibility*Humidity'] = df['Visibility (10m)'] * df['Humidity(%)']

y = df['Rented Bike Count']

x = sm.add_constant(x)

# 다중 회귀 모델 생성
model = sm.OLS(y, x)

results = model.fit()
```

```
print(results.summary())
```

OLS Regression Results

Dep. Variable:	Rented Bike Count	R-squared:	0.071
Model:	OLS	Adj. R-squared:	0.070
Method:	Least Squares	F-statistic:	221.9
Date:	Wed, 25 Oct 2023	Prob (F-statistic):	1.01e-138
Time:	22:08:42	Log-Likelihood:	-68779.
No. Observations:	8760	AIC:	1.376e+05
Df Residuals:	8756	BIC:	1.376e+05
Df Model:	3		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	1599.8335	73.556	21.750	0.000	1455.647	1744.020
Visibility (10m)	-0.4073	0.043	-9.553	0.000	-0.491	-0.324
Humidity(%)	-15.8695	0.960	-16.532	0.000	-17.751	-13.988
Visibility*Humidity	0.0080	0.001	13.394	0.000	0.007	0.009

Omnibus:	1370.843	Durbin-Watson:	0.214
Prob(Omnibus):	0.000	Jarque-Bera (JB):	2129.853
Skew:	1.110	Prob(JB):	0.00
Kurtosis:	3.952	Cond. No.	9.47e+05

・P-value: 0.000
-> It means that the p-value is smaller than 0.05, so it affects meaningful effect on dependent variables.

・Coef(coefficient) of the Visibility*Humidity : 0.0080
-> Interaction between two variables

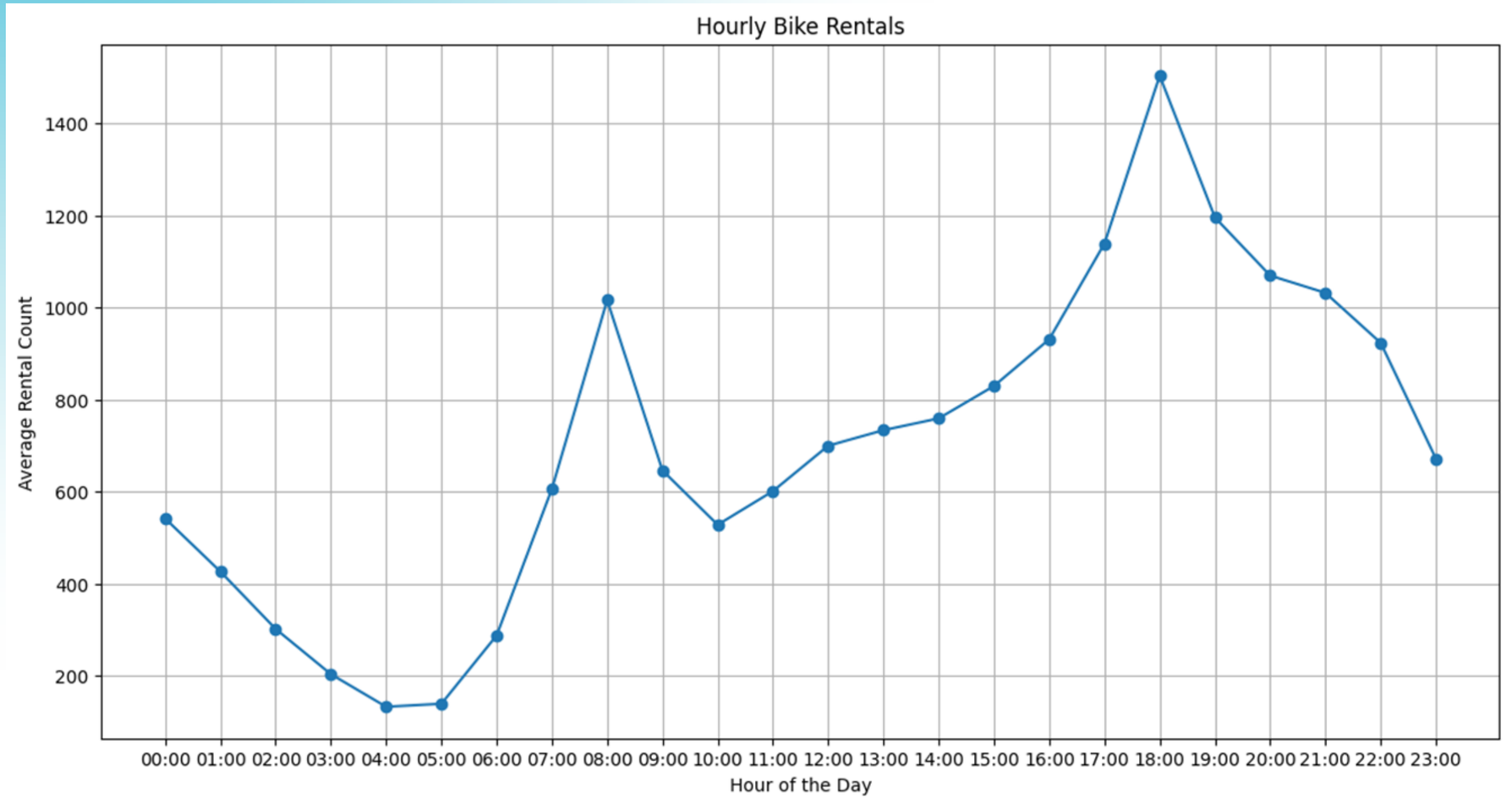
・R-squared: 0.071

4 Analysis of Rented Bike Count by time zone

- Grouped based on the "Hour" column to calculate the average bicycle load over time

```
import matplotlib.pyplot as plt

# 시간대별 대여량의 평균 계산
hourly_rentals = df.groupby('Hour')['Rented Bike Count'].mean()

plt.figure(figsize=(14, 7))
plt.plot(hourly_rentals.index, hourly_rentals.values, marker='o')
plt.title('Hourly Bike Rentals')
plt.xlabel('Hour of the Day')
plt.ylabel('Average Rental Count')
plt.grid(True)

plt.xticks(hourly_rentals.index)

plt.xticks(hourly_rentals.index, [f"{hour:02}:00" for hour in hourly_rentals.index])

plt.show()
```

4 Analysis of Rented Bike Count by time zone



・This graph is intended to visualize the average of bicycle traffic by time zone to determine when bicycle traffic is high and low during the day.

・In this case, it can be observed that the demand for bicycles increases during rush hour.

Summary

We analyzed with...

- Season - Rented Bike Count
- Temperature - Rented Bike Count
- Visibility - Rented Bike Count
- Visibility & Humidity - Rented Bike Count
- Time zone - Rented Bike Count

Through the analysis, a meaningful correlation between these variables could be confirmed.

QnA