Lab 4 - CycleHub

CycleHub, a bike rental service, recognizes the importance of understanding the factors influencing bike rental demand. This analysis uses a dataset capturing various conditions like weather and holidays to identify these factors. My goal is to provide insights that can aid CycleHub in optimizing its operations and enhancing the customer experience. The following sections will detail my exploratory data analysis, modeling methods, and recommendations.

Data

- Total Observations: 1000
- Variables:
    - bikes_rented: The number of bikes rented per hour.
    - temperature (in Fahrenheit): The ambient temperature.
    - humidity (in %): The relative humidity level.
    - wind_speed (in mph): The speed of the wind.
    - Is_holiday: indicates if the day is a holiday (1 for Yes, 0 for No).

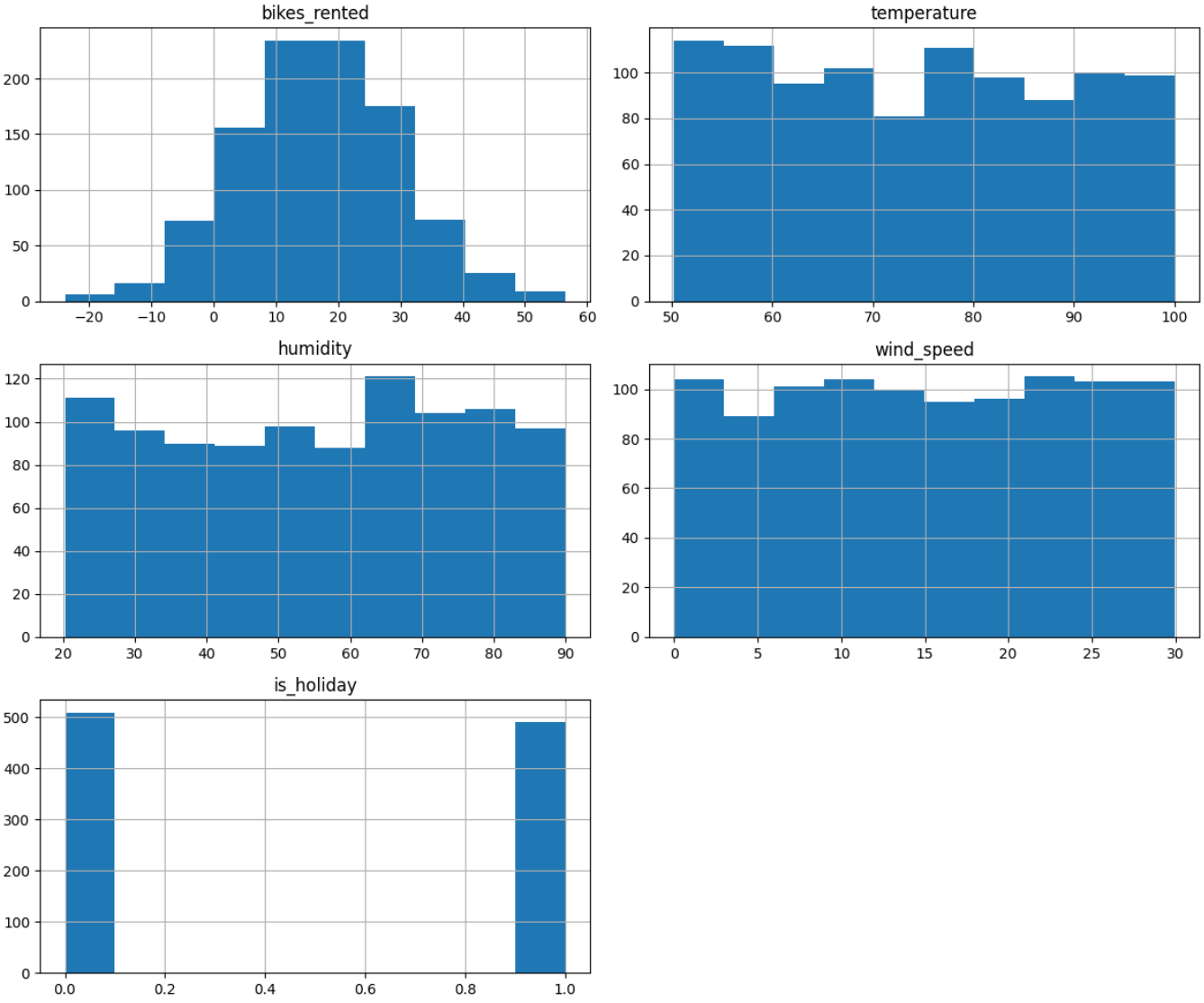
Basic Summary:

Variable	Mean	Min	Max	Std. Deviation
bikes_rented	16.91	-23.86	56.45	12.74
temperature	74.51	50.23	99.99	14.61
humidity	55.49	20.23	89.96	20.45
wind_speed	15.07	0	29.93	8.72
is_holiday	0.49	0	1	0.5

Data Analysis

The Exploratory Data Analysis aims to understand the dataset's patterns, relationships, and more.

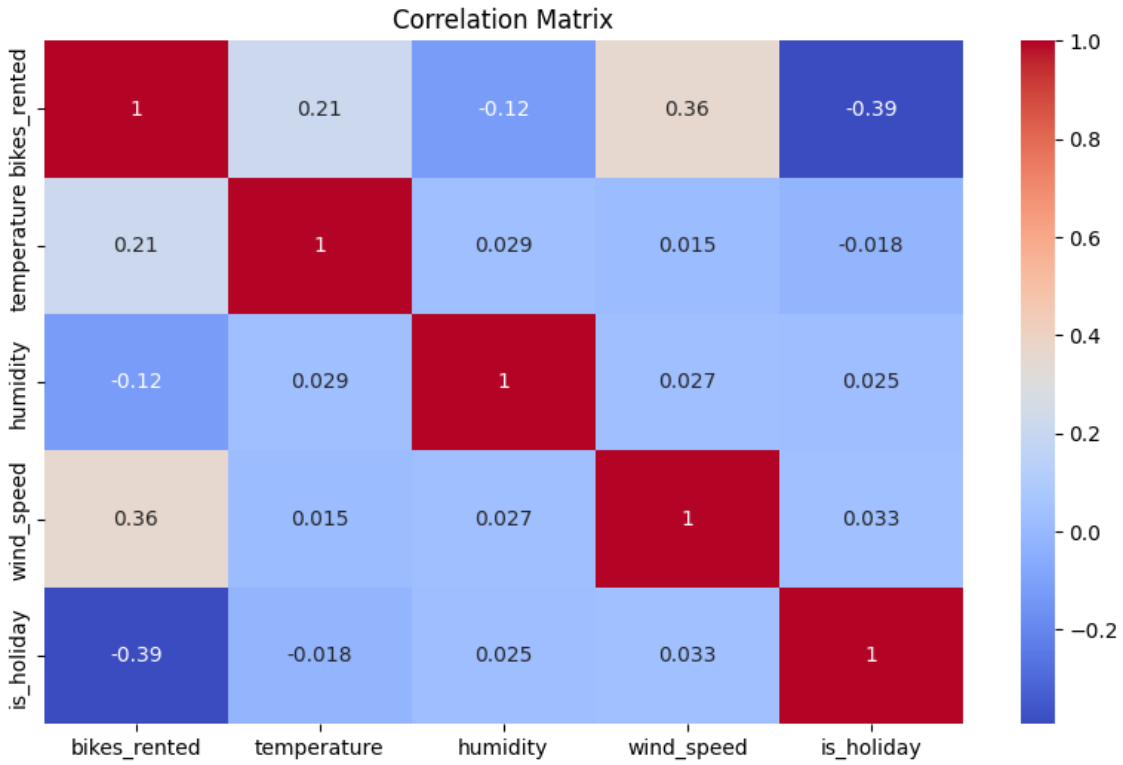
The histograms provide a quick snapshot of the distribution of each variable:



Observations:

- bikes_rented: The distribution shows that the bike rentals mostly lie in the range of 10 to 30 bikes per hour.
- temperature: The temperature seems fairly distributed but slightly biased towards warmer temperatures.
- humidity: The data is spread across, indicating varied humidity levels on different days.
- wind_speed: Wind speed shows a fairly even distribution across the range.
- is_holiday: Most data points are on non-holiday days, indicating that CycleHub operates primarily on regular days.

  Correlation Matrix gives insights into the linear relationships between the variables.



Observations:

- bikes_rented and temperature show a positive correlation, suggesting that the number of bike rentals also tends to increase as the temperature rises.
- humidity has a slightly negative correlation with bikes_rented, meaning bike rentals might decrease on days with higher humidity.
- wind_speed and bikes_rented have a moderate positive correlation.

- is_holiday shows a negative correlation with bikes_rented, indicating fewer rentals on holidays.

Regression Analysis

I did a linear regression analysis to understand how the variables relate to the number of bikes rented.

Simple Linear Regression with 'temperature' as the predictor for 'bikes_rented':

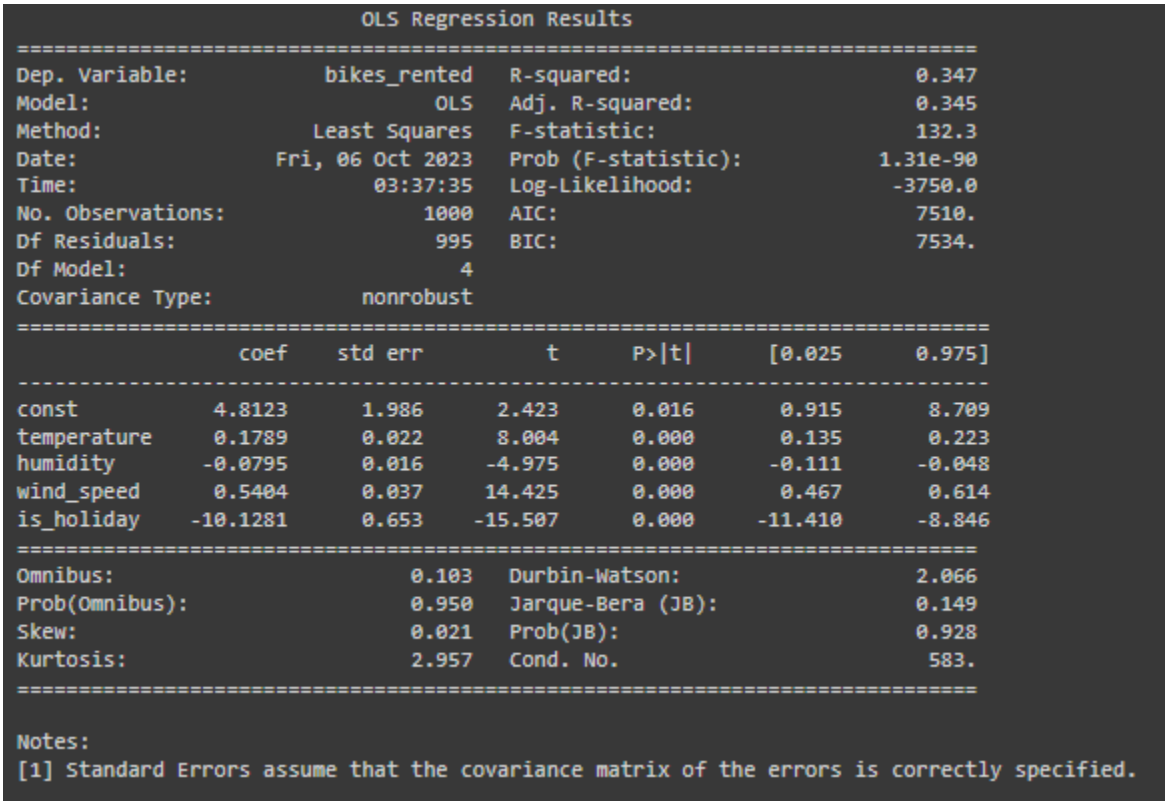
```
                            OLS Regression Results
==============================================================================
Dep. Variable:           bikes_rented   R-squared:                       0.046
Model:                            OLS   Adj. R-squared:                  0.045
Method:                 Least Squares   F-statistic:                     47.93
Date:                Fri, 06 Oct 2023   Prob (F-statistic):           7.91e-12
Time:                        03:28:48   Log-Likelihood:                -3939.7
No. Observations:                1000   AIC:                             7883.
Df Residuals:                     998   BIC:                             7893.
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const          2.9958      2.048      1.463      0.144      -1.022       7.014
temperature    0.1867      0.027      6.923      0.000       0.134       0.240
==============================================================================
Omnibus:                        0.780   Durbin-Watson:                   2.056
Prob(Omnibus):                  0.677   Jarque-Bera (JB):                0.664
Skew:                           0.050   Prob(JB):                        0.718
Kurtosis:                       3.077   Cond. No.                         395.
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
```

Observations:

- The temperature coefficient is positive, indicating that the number of bikes rented also tends to increase as temperature increases.
- The R-squared value of 0.046 indicates that about 4.6% of the variation in bike rentals can be explained by the temperature alone.

Multiple Linear Regression using all significant variables:



OLS Regression Results

Dep. Variable:	bikes_rented	R-squared:	0.347
Model:	OLS	Adj. R-squared:	0.345
Method:	Least Squares	F-statistic:	132.3
Date:	Fri, 06 Oct 2023	Prob (F-statistic):	1.31e-90
Time:	03:37:35	Log-Likelihood:	-3750.0
No. Observations:	1000	AIC:	7510.
Df Residuals:	995	BIC:	7534.
Df Model:	4		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	4.8123	1.986	2.423	0.016	0.915	8.709
temperature	0.1789	0.022	8.004	0.000	0.135	0.223
humidity	-0.0795	0.016	-4.975	0.000	-0.111	-0.048
wind_speed	0.5404	0.037	14.425	0.000	0.467	0.614
is_holiday	-10.1281	0.653	-15.507	0.000	-11.410	-8.846

Omnibus:	0.103	Durbin-Watson:	2.066
Prob(Omnibus):	0.950	Jarque-Bera (JB):	0.149
Skew:	0.021	Prob(JB):	0.928
Kurtosis:	2.957	Cond. No.	583.

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

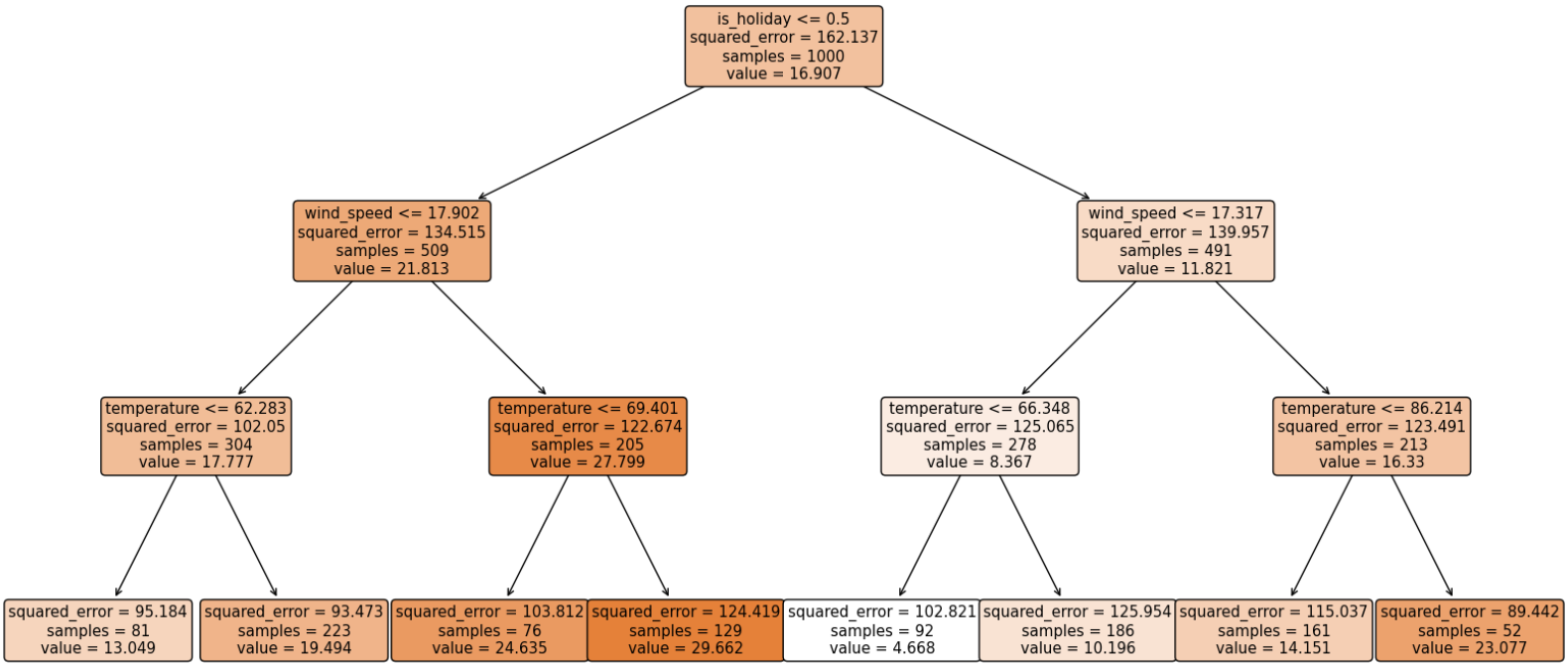
<u>Observations:</u>

- The R-squared value has increased, indicating a better fit with multiple variables.
- Temperature still has a positive coefficient, but other variables like humidity negatively affect bike rentals. This implies that as humidity increases, bike rentals tend to decrease. Similarly, interpretations can be made for wind_speed and is_holiday.

In conclusion, this analysis provides a clear understanding of the patterns and relationships within the data. This analysis offers valuable insights into how CycleHub operates.

Regression Tree Model

Regression trees are used for predicting continuous target variables. They divide the predictor space into non-overlapping regions. For every observation that falls into one region, the model predicts the mean of the target values of the training observations in that region.



Observations:

- The primary factor splitting the data is the is_holiday variable.
- Depending on is_holiday, subsequent splits are made based on wind_speed and temperature.
- The depth of the tree indicates the hierarchy of feature importance in determining bike rentals.

Special Case prediction

Given the conditions (temperature of 72F, humidity of 45%, wind speed of 10 mph, and a non-holiday), the Regression Tree predicts 19.49 = 19 or 20 bikes will be rented.

Recommendation

CycleHub can enhance its bike-sharing service. They should consider implementing pricing adjustments based on weather conditions and offering holiday promotions. Weather forecasts can also inform marketing strategies and maintenance planning.

Code

```
# -*- coding: utf-8 -*-
"""STAT220_Lab4_jgallar.ipynb

Automatically generated by Colaboratory.

Original file is located at
   https://colab.research.google.com/drive/1R1ZHcPU7msuVYi3cPbf9a8-R18GA8Djd
"""
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
import statsmodels.api as sm
from sklearn.tree import DecisionTreeRegressor, plot_tree

#1

url = "https://richardson.byu.edu/220/bike_sharing_data.csv"
data = pd.read_csv(url)

print(data.head())

#2
# Basic Summary
print(data.describe())

# Check for missing values
print(data.isnull().sum())

# Scatter Plots
features = ['temperature', 'humidity', 'wind_speed', 'is_holiday']
for feature in features:
   plt.figure(figsize=(10, 6))
   sns.scatterplot(data=data, x=feature, y='bikes_rented')
   plt.title(f'Scatter Plot of bikes_rented vs {feature}')
   plt.show()

# Histograms
data.hist(figsize=(12, 10))
plt.tight_layout()
plt.show()

# Boxplots
for feature in features:
   plt.figure(figsize=(10, 6))
   sns.boxplot(data[feature])
   plt.title(f'Boxplot of {feature}')
   plt.show()
```

```
# Correlation Matrix
correlation = data.corr()
plt.figure(figsize=(10, 6))
sns.heatmap(correlation, annot=True, cmap='coolwarm')
plt.title('Correlation Matrix')
plt.show()

#3
# a) choosen variable 'temperature'
# b)
X = sm.add_constant(data['temperature'])
y = data['bikes_rented']

model = sm.OLS(y, X).fit()

#c)
print(model.summary())

#4
# a)b) Prepare the data for regression
X = data[['temperature', 'humidity', 'wind_speed', 'is_holiday']]
X = sm.add_constant(X)
y = data['bikes_rented']

model = sm.OLS(y, X).fit()

#c)
print(model.summary())

#5
X = data[['temperature', 'humidity', 'wind_speed', 'is_holiday']]
y = data['bikes_rented']

# Regression tree
reg_tree = DecisionTreeRegressor(max_depth=3)
reg_tree.fit(X, y)

# Graph
plt.figure(figsize=(20,10))
plot_tree(reg_tree, filled=True, feature_names=X.columns, rounded=True)
plt.show()

#6
# Given coefficients and values from models
simple_regression_coef = 0.1867
simple_regression_const = 2.9958

multiple_regression_coef_temperature = 0.1789
```

```
multiple_regression_coef_wind_speed = 0.5404
multiple_regression_const = 4.8123

# Given weather forecast
temperature = 72
humidity = 45
wind_speed = 10
is_holiday = 0

# prediction
simple_regression_prediction = simple_regression_const +
simple_regression_coef * temperature

# Multiple Linear Regression prediction
multiple_regression_prediction = (multiple_regression_const +
multiple_regression_coef_temperature * temperature +
multiple_regression_coef_wind_speed * wind_speed)

# From the tree
if is_holiday <= 0.5:
   if wind_speed <= 17.9:
       if temperature <= 62.28:
           regression_tree_prediction = 13.05
       else:
           regression_tree_prediction = 19.49
   else:
       if temperature <= 69.4:
           regression_tree_prediction = 24.64
       else:
           regression_tree_prediction = 29.66
else:
   if wind_speed <= 17.32:
       if temperature <= 66.35:
           regression_tree_prediction = 4.67
       else:
           regression_tree_prediction = 10.20
   else:
       if temperature <= 86.21:
           regression_tree_prediction = 14.15
       else:
           regression_tree_prediction = 23.08

#
print(f"SLR Prediction: {simple_regression_prediction:.2f} bikes")
print(f"MLR Prediction: {multiple_regression_prediction:.2f} bikes")
print(f"RT Prediction: {regression_tree_prediction:.2f} bikes")
```