Python Bike Data Project

February 20, 2023

#Analyzing the Impact of Weather Conditions on Bike Rentals in a Specific City

[144]:
```
# 1. Introduction

#As part of my data analysis project, I decided to analyze the impact of␣
↪weather conditions on bike rentals in a specific city.
#I downloaded the bike-sharing dataset(https://www.kaggle.com/marklvl/
↪bike-sharing-dataset), which contains information on the number of bike␣
↪rentals, weather conditions, and other
#factors such as time of day and day of the week. My goal was to examine the␣
↪relationship between weather conditions and bike
#rentals, and answer the question: How do different weather conditions affect␣
↪the number of bike rentals in a specific city?

#I started by stating my primary research question: How do different weather␣
↪conditions affect the number of bike rentals in a
#specific city? My hypothesis was that there would be a correlation between␣
↪certain weather conditions and the number of bike
#rentals. For example, I expected to see more rentals on days with pleasant␣
↪weather (e.g. moderate temperature, low humidity,
#low windspeed) and fewer rentals on days with extreme weather (e.g. very hot␣
↪or very cold temperatures, high windspeed, heavy rain).

#The findings of this project could have significant implications for␣
↪bike-sharing companies, transportation planners, and city
#governments. By understanding how weather conditions impact bike rentals,␣
↪stakeholders can better plan and allocate resources
#to support and promote bike-sharing programs. Additionally, my project could␣
↪serve as a case study for other cities to learn
#from and adapt to their own unique circumstances.
```

[145]:
```
# 2. Download, Import, and Check

# First, I downloaded the bike-sharing dataset from Kaggle, a platform for data␣
↪science projects.
# I imported the dataset into Jupyter Notebook and used the Pandas library to␣
↪convert it into a dataframe.
```

```python
import pandas as pd

# Load the dataset into a Pandas dataframe
df = pd.read_csv("C:/Users/tmq94/Desktop/Data Portfolio 2-19-2023/2. Bike␣
 ↪Project/bike-sharing-dataset/day.csv")

# Next, I inspected the dataset for any issues such as missing data or␣
 ↪duplicates.
# I used various Pandas functions to explore the data and check for any␣
 ↪anomalies.

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

# Check for duplicates
print(df.duplicated().sum())

# In this case, there were no missing values or duplicates in the dataset.␣
 ↪Therefore, I proceeded to the next step of the project: cleaning and␣
 ↪preparing the dataset for analysis.
```

```
instant       0
dteday        0
season        0
yr            0
mnth          0
holiday       0
weekday       0
workingday    0
weathersit    0
temp          0
atemp         0
hum           0
windspeed     0
casual        0
registered    0
cnt           0
dtype: int64
0
```

[146]:
```python
# 3. Data Cleaning and Preparation

# First, I dropped the unnecessary columns from the dataframe.
# I decided to drop the 'instant' column, which contained a unique ID for each␣
 ↪row, as well as the 'dteday' column, which contained the date in a␣
 ↪non-standard format that would not be useful for analysis.
```

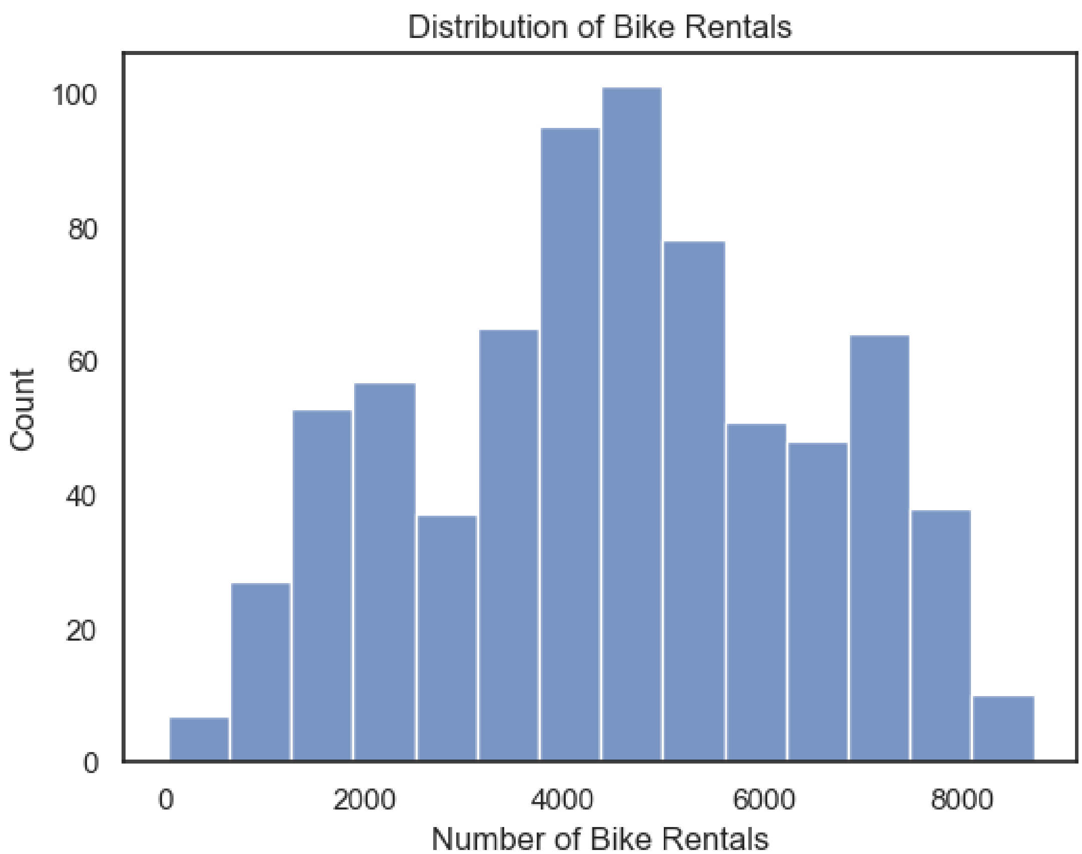
```
df.drop(['instant', 'dteday'], axis=1, inplace=True)

# Next, I converted the 'season' and 'weathersit' columns from numerical values␣
↪to categorical values to make them more interpretable.

df['season'] = df['season'].map({1: 'spring', 2: 'summer', 3: 'fall', 4:␣
↪'winter'})
df['weathersit'] = df['weathersit'].map({1: 'clear', 2: 'misty/cloudy', 3:␣
↪'light snow/rain', 4: 'heavy snow/rain'})

# I also created dummy variables for the categorical columns to use them in the␣
↪regression model.

season_dummies = pd.get_dummies(df['season'], prefix='season', drop_first=True)
weathersit_dummies = pd.get_dummies(df['weathersit'], prefix='weathersit',␣
↪drop_first=True)
df = pd.concat([df, season_dummies, weathersit_dummies], axis=1)
df.drop(['season', 'weathersit'], axis=1, inplace=True)

# Finally, I standardized the continuous variables to put them on the same␣
↪scale.

from sklearn.preprocessing import StandardScaler

scaler = StandardScaler()
continuous_vars = ['temp', 'atemp', 'hum', 'windspeed']
df[continuous_vars] = scaler.fit_transform(df[continuous_vars])
```

[148]:
```
# 4.Data Exploration and Visualization,



import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt

# Load the dataset
df = pd.read_csv("C:/Users/tmq94/Desktop/Data Portfolio 2-19-2023/2. Bike␣
↪Project/bike-sharing-dataset/day.csv")

# Check the column names
print("Column Names:")
print(df.columns)
print()

# Create a histogram of bike rentals
```

```python
print("Histogram of Bike Rentals:")
print("Summary Statistics:")
print(df['cnt'].describe())
sns.histplot(x='cnt', data=df, kde=False)
plt.title('Distribution of Bike Rentals')
plt.xlabel('Number of Bike Rentals')
plt.show()

# Create a scatter plot of bike rentals by temperature
print("Scatter plot of Bike Rentals by Temperature:")
print("Summary Statistics:")
print(df[['temp', 'cnt']].describe())
plt.figure(figsize=(10, 6))
sns.scatterplot(x='temp', y='cnt', data=df)
plt.title('Scatter plot of Bike Rentals by Temperature')
plt.xlabel('Temperature')
plt.ylabel('Number of Bike Rentals')
plt.show()

# Box plot of bike rentals by season
print("Box plot of Bike Rentals by Season:")
print("Summary Statistics:")
print(df.groupby('season')['cnt'].describe())
plt.figure(figsize=(10, 6))
sns.boxplot(x='season', y='cnt', data=df)
plt.title('Distribution of Bike Rentals by Season')
plt.xlabel('Season')
plt.ylabel('Number of Bike Rentals')
plt.show()

# Create a correlation matrix of variables
print("Correlation Matrix:")
corr = df.corr()
print(corr)
plt.figure(figsize=(12, 10))
sns.heatmap(corr, cmap='coolwarm', annot=True, fmt='.2f', annot_kws={"size": 8})
plt.title('Correlation Matrix')
plt.show()
print(corr.describe())
```

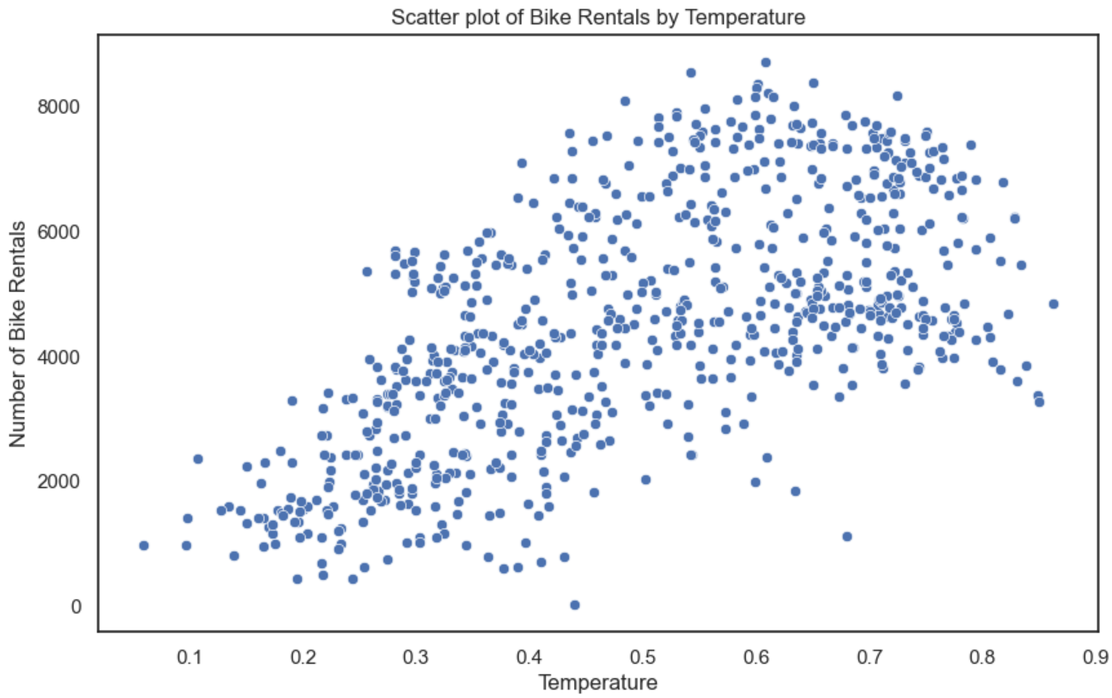
```
Column Names:
Index(['instant', 'dteday', 'season', 'yr', 'mnth', 'holiday', 'weekday',
       'workingday', 'weathersit', 'temp', 'atemp', 'hum', 'windspeed',
       'casual', 'registered', 'cnt'],
      dtype='object')

Histogram of Bike Rentals:
Summary Statistics:
```

```
count     731.000000
mean     4504.348837
std      1937.211452
min        22.000000
25%      3152.000000
50%      4548.000000
75%      5956.000000
max      8714.000000
Name: cnt, dtype: float64
```

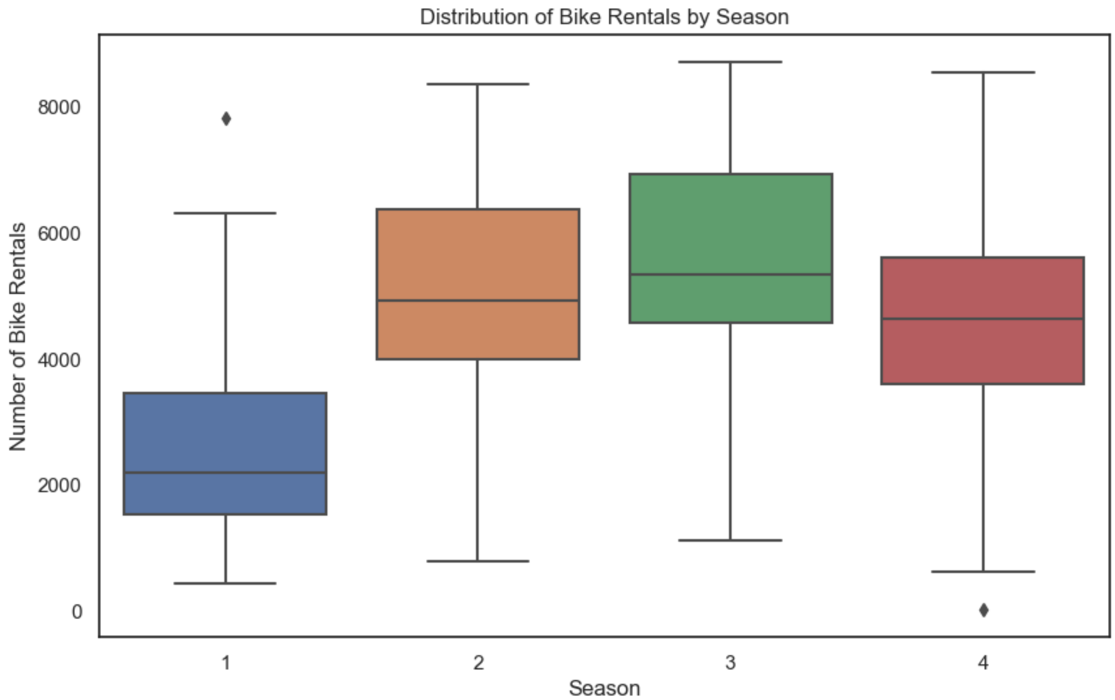



```
Scatter plot of Bike Rentals by Temperature:
Summary Statistics:
             temp          cnt
count  731.000000   731.000000
mean     0.495385  4504.348837
std      0.183051  1937.211452
min      0.059130    22.000000
25%      0.337083  3152.000000
50%      0.498333  4548.000000
75%      0.655417  5956.000000
```

```
max       0.861667  8714.000000
```




```
Box plot of Bike Rentals by Season:
Summary Statistics:
        count         mean          std     min     25%     50%      75%  \
season
1       181.0  2604.132597  1399.942119   431.0  1538.0  2209.0  3456.00
2       184.0  4992.331522  1695.977235   795.0  4003.0  4941.5  6377.00
3       188.0  5644.303191  1459.800381  1115.0  4586.5  5353.5  6929.25
4       178.0  4728.162921  1699.615261    22.0  3615.5  4634.5  5624.50

           max
season
1       7836.0
2       8362.0
3       8714.0
4       8555.0
```



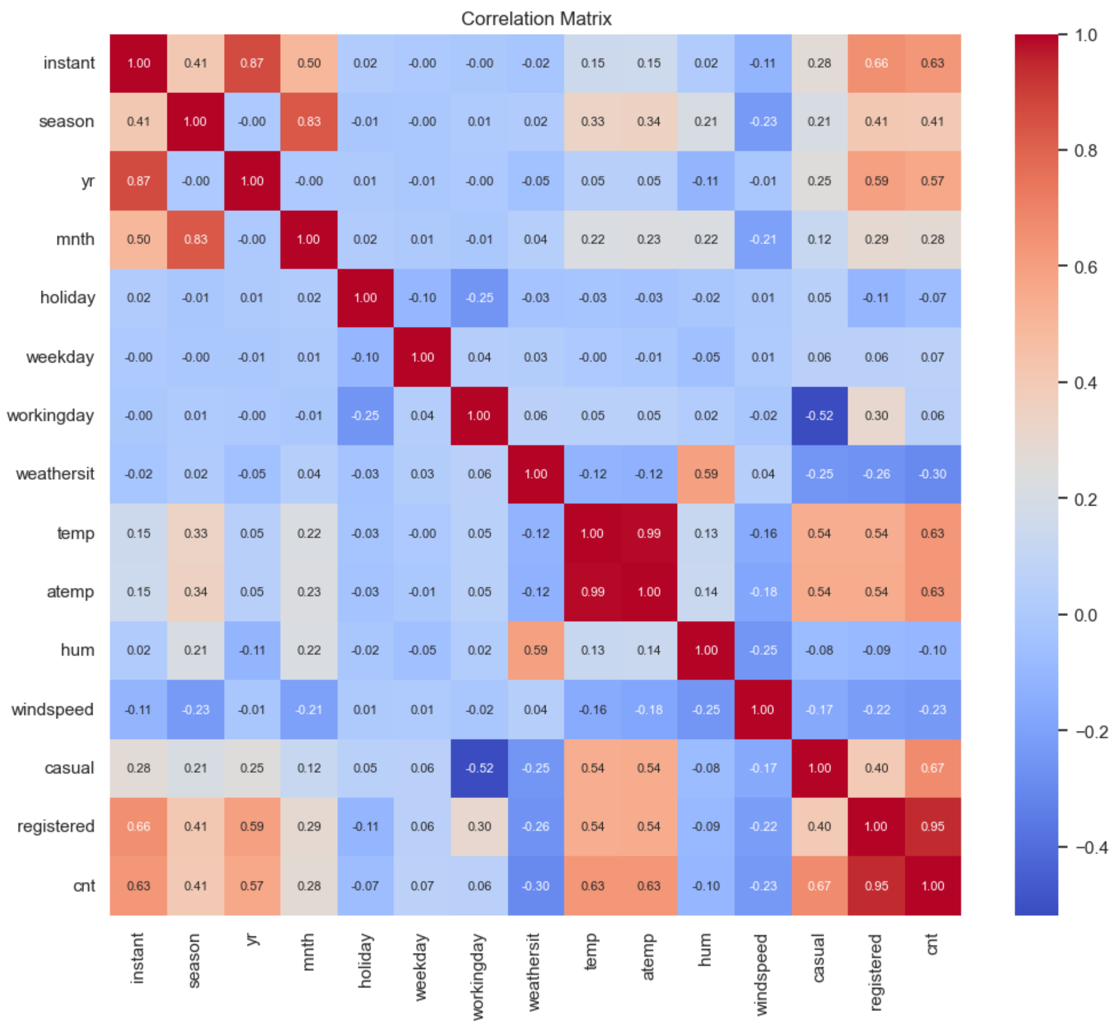
Correlation Matrix:

	instant	season	yr	mnth	holiday	weekday	\
instant	1.000000	0.412224	0.866025	0.496702	0.016145	-0.000016	
season	0.412224	1.000000	-0.001844	0.831440	-0.010537	-0.003080	
yr	0.866025	-0.001844	1.000000	-0.001792	0.007954	-0.005461	
mnth	0.496702	0.831440	-0.001792	1.000000	0.019191	0.009509	
holiday	0.016145	-0.010537	0.007954	0.019191	1.000000	-0.101960	
weekday	-0.000016	-0.003080	-0.005461	0.009509	-0.101960	1.000000	
workingday	-0.004337	0.012485	-0.002013	-0.005901	-0.253023	0.035790	
weathersit	-0.021477	0.019211	-0.048727	0.043528	-0.034627	0.031087	
temp	0.150580	0.334315	0.047604	0.220205	-0.028556	-0.000170	
atemp	0.152638	0.342876	0.046106	0.227459	-0.032507	-0.007537	
hum	0.016375	0.205445	-0.110651	0.222204	-0.015937	-0.052232	
windspeed	-0.112620	-0.229046	-0.011817	-0.207502	0.006292	0.014282	
casual	0.275255	0.210399	0.248546	0.123006	0.054274	0.059923	
registered	0.659623	0.411623	0.594248	0.293488	-0.108745	0.057367	
cnt	0.628830	0.406100	0.566710	0.279977	-0.068348	0.067443	

	workingday	weathersit	temp	atemp	hum	windspeed	\
instant	-0.004337	-0.021477	0.150580	0.152638	0.016375	-0.112620	
season	0.012485	0.019211	0.334315	0.342876	0.205445	-0.229046	
yr	-0.002013	-0.048727	0.047604	0.046106	-0.110651	-0.011817	
mnth	-0.005901	0.043528	0.220205	0.227459	0.222204	-0.207502	
holiday	-0.253023	-0.034627	-0.028556	-0.032507	-0.015937	0.006292	

weekday	0.035790	0.031087	-0.000170	-0.007537	-0.052232	0.014282
workingday	1.000000	0.061200	0.052660	0.052182	0.024327	-0.018796
weathersit	0.061200	1.000000	-0.120602	-0.121583	0.591045	0.039511
temp	0.052660	-0.120602	1.000000	0.991702	0.126963	-0.157944
atemp	0.052182	-0.121583	0.991702	1.000000	0.139988	-0.183643
hum	0.024327	0.591045	0.126963	0.139988	1.000000	-0.248489
windspeed	-0.018796	0.039511	-0.157944	-0.183643	-0.248489	1.000000
casual	-0.518044	-0.247353	0.543285	0.543864	-0.077008	-0.167613
registered	0.303907	-0.260388	0.540012	0.544192	-0.091089	-0.217449
cnt	0.061156	-0.297391	0.627494	0.631066	-0.100659	-0.234545

	casual	registered	cnt
instant	0.275255	0.659623	0.628830
season	0.210399	0.411623	0.406100
yr	0.248546	0.594248	0.566710
mnth	0.123006	0.293488	0.279977
holiday	0.054274	-0.108745	-0.068348
weekday	0.059923	0.057367	0.067443
workingday	-0.518044	0.303907	0.061156
weathersit	-0.247353	-0.260388	-0.297391
temp	0.543285	0.540012	0.627494
atemp	0.543864	0.544192	0.631066
hum	-0.077008	-0.091089	-0.100659
windspeed	-0.167613	-0.217449	-0.234545
casual	1.000000	0.395282	0.672804
registered	0.395282	1.000000	0.945517
cnt	0.672804	0.945517	1.000000



	instant	season	yr	mnth	holiday	weekday	\
count	15.000000	15.000000	15.000000	15.000000	15.000000	15.000000	
mean	0.302397	0.262774	0.212993	0.236768	0.029974	0.073663	
std	0.354657	0.330447	0.361033	0.324238	0.278330	0.259917	
min	-0.112620	-0.229046	-0.110651	-0.207502	-0.253023	-0.101960	
25%	0.008064	0.005320	-0.003737	0.014350	-0.051487	-0.004270	
50%	0.152638	0.210399	0.007954	0.220205	-0.015937	0.009509	
75%	0.562766	0.408862	0.407628	0.286732	0.012049	0.046579	
max	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	

	workingday	weathersit	temp	atemp	hum	windspeed	\
count	15.000000	15.000000	15.000000	15.000000	15.000000	15.000000	
mean	0.053440	0.042229	0.288503	0.288453	0.108685	-0.048625	
std	0.315927	0.336145	0.374154	0.377230	0.316413	0.307738	
min	-0.518044	-0.297391	-0.157944	-0.183643	-0.248489	-0.248489	
25%	-0.005119	-0.121093	0.023717	0.019285	-0.084048	-0.212475	

```
50%      0.024327  -0.021477  0.150580  0.152638  0.016375  -0.157944
75%      0.056908   0.041520  0.541648  0.544028  0.172716  -0.002763
max      1.000000   1.000000  1.000000  1.000000  1.000000   1.000000

          casual  registered        cnt
count  15.000000   15.000000  15.000000
mean    0.207775    0.337839   0.345744
std     0.389079    0.396470   0.420206
min    -0.518044   -0.260388  -0.297391
25%    -0.011367   -0.016861  -0.003596
50%     0.210399    0.395282   0.406100
75%     0.469284    0.569220   0.629948
max     1.000000    1.000000   1.000000
```

[ ]:

```
#5. Interpreting the Results: Bike Rental Data Analysis


#1. Histogram of Bike Rentals: Summary Statistics and Pattern

#The mean number of bike rentals is 4,504 with a standard deviation of 1,937,␣
↪indicating that the data is somewhat spread out around the mean.
#The minimum number of rentals is 22 and the maximum is 8,714, showing a wide␣
↪range of values in the data.
#The data is slightly skewed to the right, with more values towards the higher␣
↪end of the rental counts.
#The distribution of bike rentals appears to be roughly normal with a peak␣
↪around the mean and tapering off towards the extremes.

#Overall, this suggests that bike rentals are a popular and widely used␣
↪transportation option in the area, with a fairly consistent demand␣
↪throughout the given time period.


# 2. Exploring the Relationship between Temperature and Bike Rentals: A Scatter␣
↪Plot Analysis and Summary Statistics
# There is a positive correlation between temperature and bike rentals,␣
↪indicating that as temperature increases, so does the number of bike rentals.
# The mean temperature value is 0.495385, while the mean bike rentals value is␣
↪4504.348837, which supports the positive correlation between these two␣
↪variables.
# The standard deviation for both temperature and bike rentals is relatively␣
↪large, indicating a wide range of values for both variables.
# The minimum value for bike rentals is 22, while the maximum value is 8714.␣
↪The minimum temperature value is 0.059130, while the maximum temperature␣
↪value is 0.861667.
# The 25th percentile for bike rentals is 3152, and the 75th percentile is␣
↪5956, suggesting that the majority of data points fall within this range.
```

```
# These findings suggest that temperature is an important factor to consider when predicting bike rentals, and that warmer temperatures are associated with increased bike rentals.



#3. Summary statistics for Graph 3:Box Plot of Distribution of Bike Rentals by Season

#The box plot of bike rentals by season shows that the mean number of bike rentals increases from season 1 to season 3 and then decreases in season 4. Season 2 has the highest mean number of bike rentals.

#The box plot also shows that there is a wider range of bike rentals in seasons 2 and 3 compared to seasons 1 and 4. In season 2, there are more outliers in the upper range of bike rentals, suggesting that there were some days with exceptionally high bike rentals during that season.

#The standard deviation is highest in season 2, indicating that the data points are more spread out from the mean, while season 1 has the lowest standard deviation, indicating that the data points are more tightly clustered around the mean.

#Overall, the box plot suggests that bike rentals are generally higher in the warmer seasons (seasons 2 and 3) and lower in the colder seasons (seasons 1 and 4), with season 2 having the highest overall demand for bike rentals.



#4.Correlation Matrix: Relationships between Variables in Bike Sharing Dataset


#Based on the observations, we can make the following logical deductions:

#Instant, which represents the index of the record, has a strong positive correlation with year and a moderate positive correlation with cnt and registered. This indicates that as the years pass, the number of bike rentals increases, and as the record index increases, the number of rentals also tends to increase.

#Season has a moderate positive correlation with month and a weak positive correlation with temp and atemp. This suggests that the season is related to the month, with the warmer seasons occurring in the middle of the year, and that temperature and apparent temperature have some influence on the season.
```

```
#Casual has a moderate positive correlation with temp and atemp, and a weak␣
↪positive correlation with season and month. This implies that the␣
↪temperature has a greater impact on casual bike rentals compared to␣
↪registered rentals. The warmer months and seasons also seem to have a slight␣
↪influence on the number of casual rentals.

#Registered has a moderate positive correlation with temp and atemp, and a weak␣
↪positive correlation with season and month. This indicates that temperature␣
↪and apparent temperature have some impact on registered bike rentals. The␣
↪month and season also have a slight influence on the number of registered␣
↪rentals.

#Cnt has a moderate positive correlation with temp and atemp, and a weak␣
↪positive correlation with season and month. This suggests that temperature␣
↪and apparent temperature have some impact on the overall bike rental count.␣
↪The month and season also have a slight influence on the total number of␣
↪rentals.

#Year has a strong positive correlation with instant and a moderate positive␣
↪correlation with cnt and registered. This indicates that as the years pass,␣
↪the index of the record and the number of bike rentals increase. The␣
↪relationship between year and cnt and registered rentals is stronger than␣
↪the relationship between year and casual rentals.
```