How to begin:

Import the dataset and do usual exploratory data analysis steps like checking the structure & characteristics of the dataset

Try establishing a relation between the dependent and independent variable (Dependent “Count” & Independent: Workingday, Weather, Season etc)

Select an appropriate test to check whether:

Working Day has effect on number of electric cycles rented

No. of cycles rented similar or different in different seasons

No. of cycles rented similar or different in different weather

Weather is dependent on season (check between 2 predictor variable)

Set up Null Hypothesis (H0)

State the alternate hypothesis (H1)

Check assumptions of the test (Normality, Equal Variance). You can check it

using Histogram, Q-Q plot or statistical methods like levene’s test, Shapiro-wilk test (optional)

Please continue doing the analysis even If some assumptions fail (levene’s test

or Shapiro-wilk test) but double check using visual analysis and report

wherever necessary

Set a significance level (alpha)

Calculate test Statistics.

Decision to accept or reject null hypothesis.

Inference from the analysis

Evaluation Criteria (50 Points):

Define Problem Statement and perform Exploratory Data Analysis (10 points)

Definition of problem (as per given problem statement with additional views)

Observations on shape of data, data types of all the attributes, conversion of categorical attributes to 'category' (If required) , missing value detection, statistical summary.

Univariate Analysis (distribution plots of all the continuous variable(s) barplots/countplots of all the categorical variables)

Bivariate Analysis (Relationships between important variables such as workday and count, season and count, weather and count.

Illustrate the insights based on EDA

Comments on range of attributes, outliers of various attributes

Comments on the distribution of the variables and relationship between them

Comments for each univariate and bivariate plots

Hypothesis Testing (30 Points):

1. 2- Sample T-Test to check if Working Day has an effect on the number of electric cycles rented (10 points)
2. ANNOVA to check if No. of cycles rented is similar or different in different 1. weather 2. season (10 points)
3. Chi-square test to check if Weather is dependent on the season (10 points) Notebook Quality (10 points):

Structure & Flow

Well commented code

What good looks like (distribution of 10 points):

Visual analysis (1)

Hypothesis formulation (1)

Select the appropriate test (1)

Check test assumptions (2)

Find the p-value(1)

Conclusion based on the p-value (2)

```
#Downloading Dataset
!gdown "https://d2beiqkhq929f0.cloudfront.net/public_assets/assets/000/001/428/original/bike_
```

```
Downloading...
From: https://d2beiqkhq929f0.cloudfront.net/public_assets/assets/000/001/428/original/bi
To: /content/bike_sharing.csv?1642089089
100% 648k/648k [00:00<00:00, 7.40MB/s]
```

```
import pandas as pd
import seaborn as sns
import numpy as np, matplotlib.pyplot as plt

from scipy.stats import ttest_ind,f_oneway,chi2_contingency,chi2
```

```
df = pd.read_csv("/content/bike_sharing.csv?1642089089")
```

```
#Preview of Dataset
df.head()
```

	datetime	season	holiday	workingday	weather	temp	atemp	humidity	windspeed	ca
0	2011-01-01 00:00:00	1	0	0	1	9.84	14.395	81	0.0	
1	2011-01-01 01:00:00	1	0	0	1	9.02	13.635	80	0.0	
	2011-01-									

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 10886 entries, 0 to 10885
Data columns (total 12 columns):
 #   Column      Non-Null Count  Dtype  
---  ------      --------------  -----  
 0   datetime    10886 non-null  object 
 1   season      10886 non-null  int64  
 2   holiday     10886 non-null  int64  
 3   workingday  10886 non-null  int64  
 4   weather     10886 non-null  int64  
 5   temp        10886 non-null  float64
 6   atemp       10886 non-null  float64
 7   humidity    10886 non-null  int64  
 8   windspeed   10886 non-null  float64
 9   casual      10886 non-null  int64  
 10  registered  10886 non-null  int64  
 11  count       10886 non-null  int64  
dtypes: float64(3), int64(8), object(1)
memory usage: 1020.7+ KB
```

```
#Changing Datatype of Datetime
df['datetime'] = df['datetime'].astype('datetime64')
```

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 10886 entries, 0 to 10885
Data columns (total 12 columns):
 #   Column      Non-Null Count  Dtype         
---  ------      --------------  -----         
 0   datetime    10886 non-null  datetime64[ns]
 1   season      10886 non-null  int64         
 2   holiday     10886 non-null  int64         
 3   workingday  10886 non-null  int64         
 4   weather     10886 non-null  int64         
 5   temp        10886 non-null  float64       
 6   atemp       10886 non-null  float64       
 7   humidity    10886 non-null  int64         
 8   windspeed   10886 non-null  float64       
 9   casual      10886 non-null  int64         
 10  registered  10886 non-null  int64         
 11  count       10886 non-null  int64         
dtypes: datetime64[ns](1), float64(3), int64(8)
memory usage: 1020.7 KB
```

```
df.describe()
```

	season	holiday	workingday	weather	temp	atemp
count	10886.000000	10886.000000	10886.000000	10886.000000	10886.00000	10886.000000
mean	2.506614	0.028569	0.680875	1.418427	20.23086	23.655084
std	1.116174	0.166599	0.466159	0.633839	7.79159	8.474601
min	1.000000	0.000000	0.000000	1.000000	0.82000	0.760000
25%	2.000000	0.000000	0.000000	1.000000	13.94000	16.665000
50%	3.000000	0.000000	1.000000	1.000000	20.50000	24.240000
75%	4.000000	0.000000	1.000000	2.000000	26.24000	31.060000
max	4.000000	1.000000	1.000000	4.000000	41.00000	45.455000

From above tab, we can observe that:

1. Total number of Counts in each columns are same, so no null values.
2. Mean of Season is around 2.5, in this we can conclude that most of the data lies in Summer season. Holiday has mean near to 0, so most of the days are non-Holidays / Working days. Weather is mostly clear or partly cloud, and in this we can see that in rainy season i.e thunderstorm, rainy weather the bicycle renting goes down. Mean Temperature and Humidity is near to 21 and 62, which also depicts that most of data recorded in Summer Season.
3. Observing Maximum columns, we can see that currently unique registered count is 866, Maximum temperature recorded in dataset is near to 100.

```
#Checking the dimension of Dataset
```

```
df.shape
```

```
(10886, 12)
```

```
df.isnull().sum()
```

```
datetime      0
season        0
holiday       0
workingday    0
weather       0
temp          0
atemp         0
humidity      0
windspeed     0
casual        0
registered    0
count         0
dtype: int64
```

```
df.columns
```

```
Index(['datetime', 'season', 'holiday', 'workingday', 'weather', 'temp',
       'atemp', 'humidity', 'windspeed', 'casual', 'registered', 'count'],
      dtype='object')
```

```
cols_store = []
cols = ['season', 'holiday', 'workingday', 'weather', 'temp','atemp', 'humidity', 'windspeed'
for i in cols:
  print(i,": ",df[i].nunique())
```

```
season :  4
holiday :  2
workingday :  2
weather :  4
temp :  49
atemp :  60
humidity :  89
windspeed :  28
casual :  309
registered :  731
count :  822
```

From above we can conclude that Categorical Columns are: Season, Holiday, Working_day and Weather.

```
cat_cols= ['season', 'holiday', 'workingday', 'weather']
```

```
cols1 = []
for i in cat_cols:
  print(i,":",df[i].unique())
```

```
season : [1 2 3 4]
holiday : [0 1]
workingday : [0 1]
weather : [1 2 3 4]
```

1. Season : [1: spring, 2: summer, 3: fall, 4: winter]

2. Weather:-

   1: Clear, Few clouds, partly cloudy.

   2: Mist + Cloudy, Mist + Broken clouds, Mist + Few clouds, Mist

   3: Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds

   4: Heavy Rain + Ice Pallets + Thunderstorm + Mist, Snow + Fog

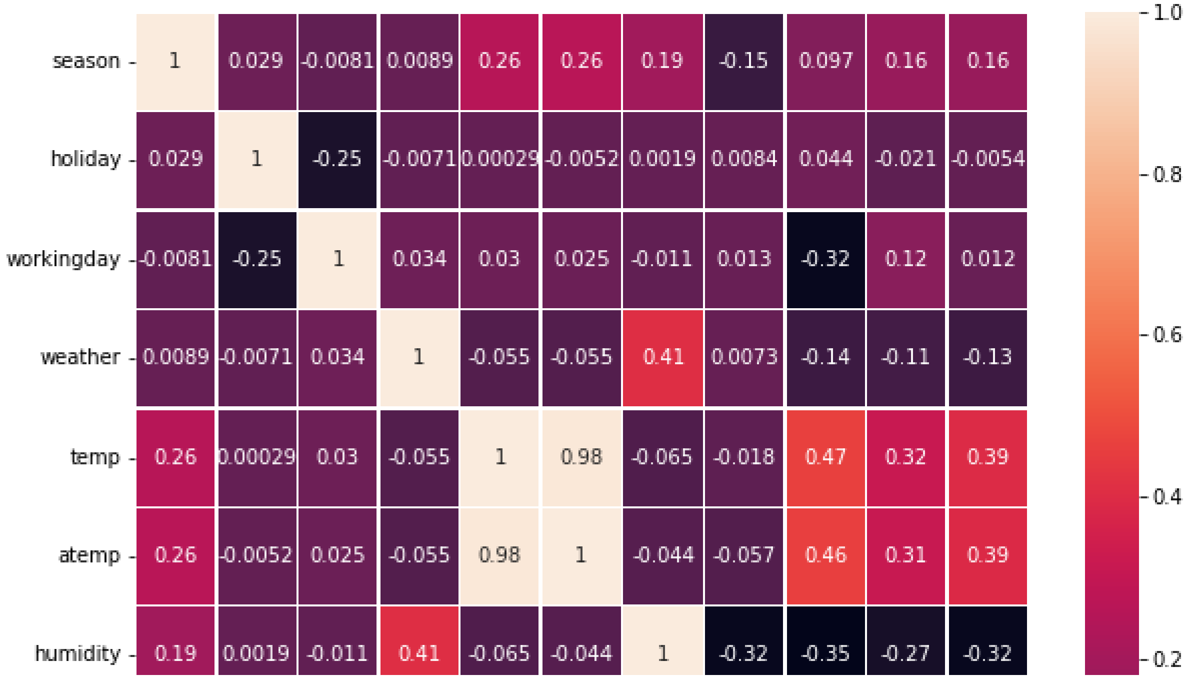
```
# Duration of Dataset:

df['datetime'].min(),df['datetime'].max()
```

```
(Timestamp('2011-01-01 00:00:00'), Timestamp('2012-12-19 23:00:00'))
```

From above we can see that the DataFrame is from 1st Jan 2011 to 19th Dec 2012, which is very old Data.

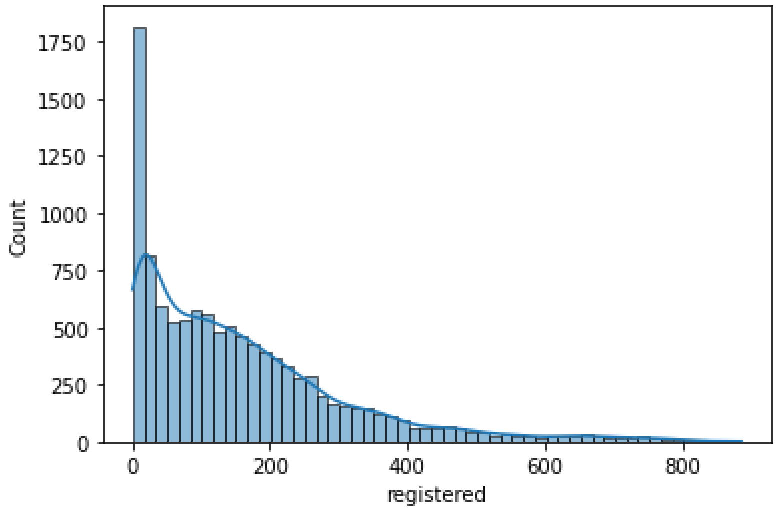
```
fig, ax = plt.subplots(figsize=(10,10))
sns.heatmap(df.corr(),annot = True,linewidths=.5, ax=ax)
plt.show()
```



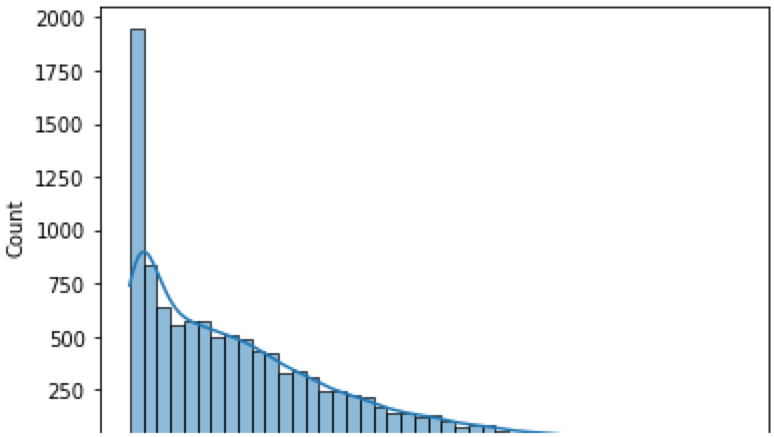
From HeatMap we can conclude that Registered and Count columns are closely co-related and other columns are mostly independent of Other columns.

Univariate Analysis

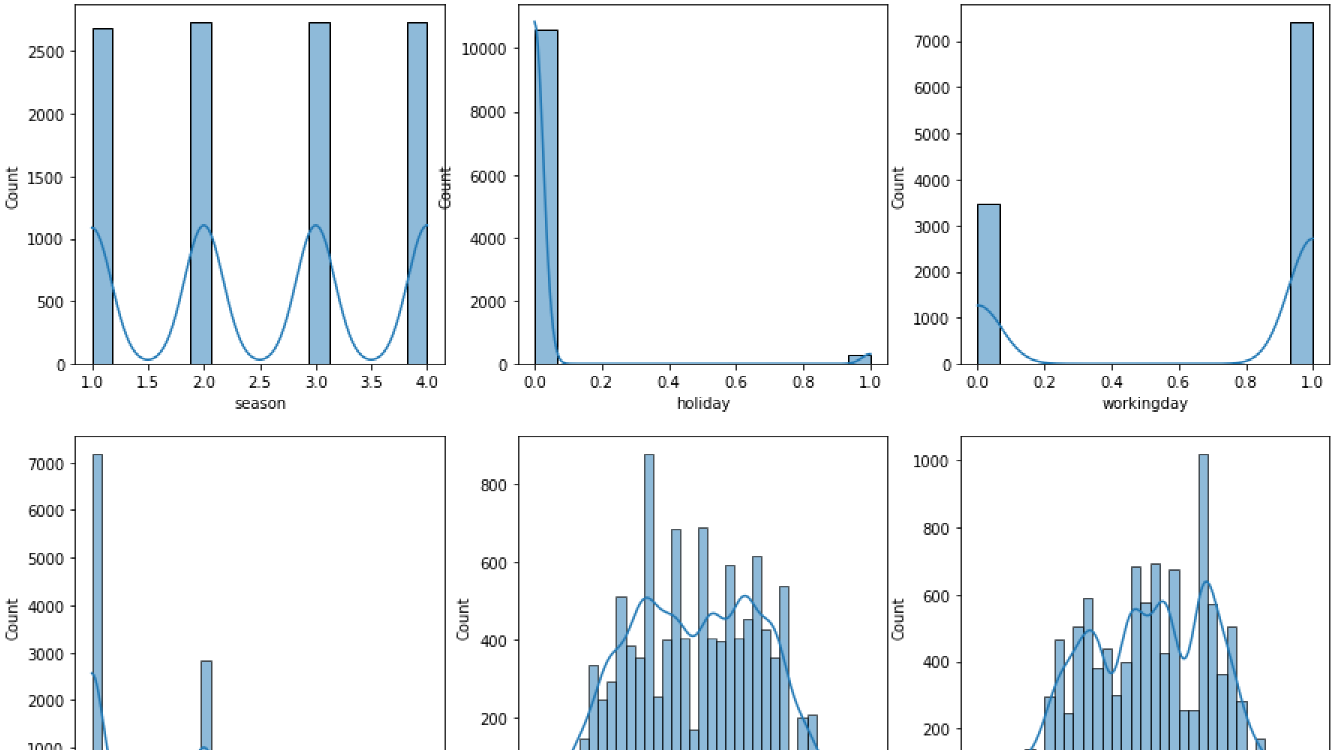
```
sns.histplot(data = df,x='registered',kde = True)
plt.show()
```



```
sns.histplot(data = df,x='count',kde = True)
plt.show()
```



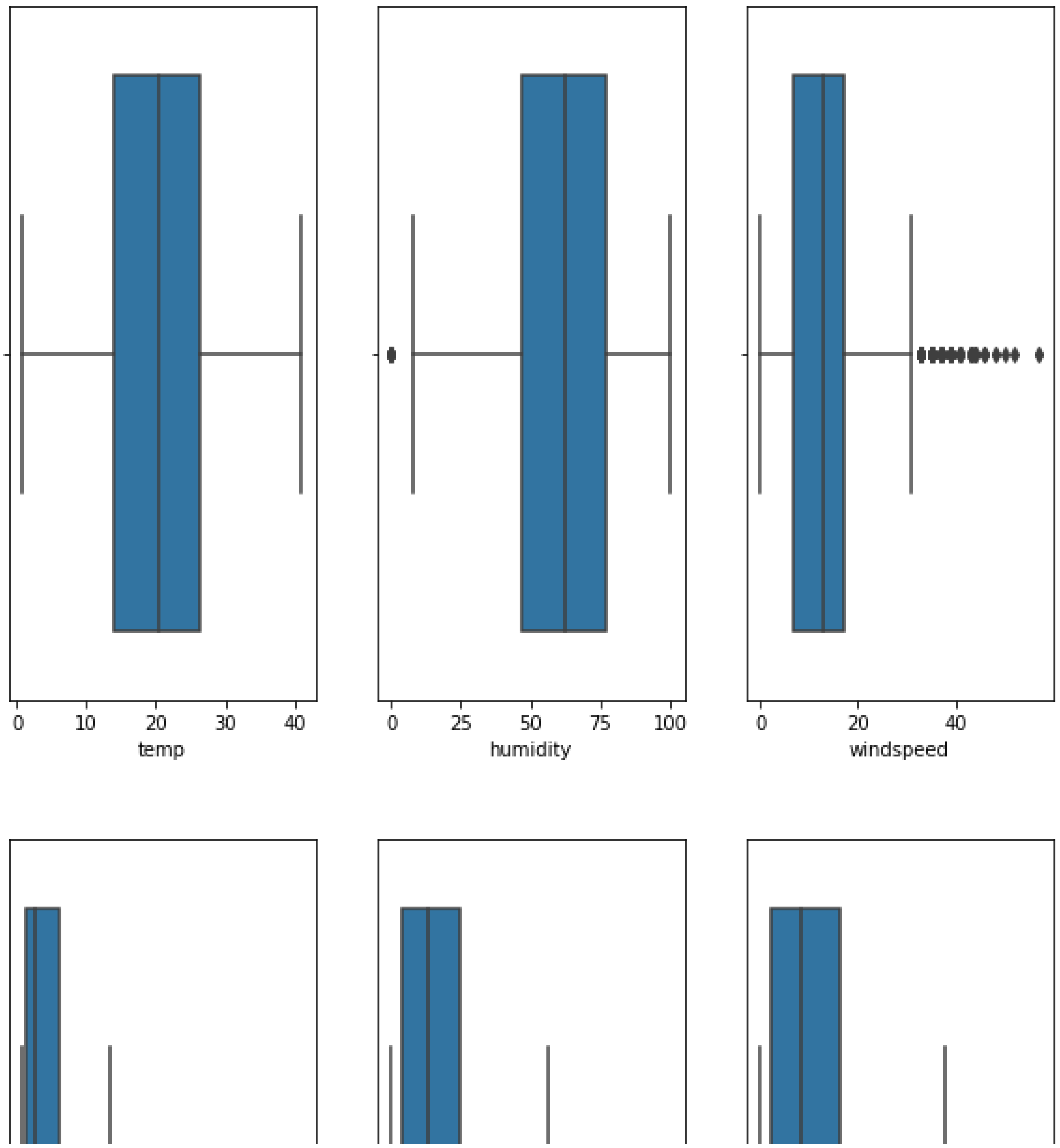
```
fig,axes = plt.subplots(nrows = 3,ncols = 3,figsize=(15,15))
cols2 = ['season', 'holiday', 'workingday', 'weather', 'temp','atemp', 'humidity', 'windspeed
sns.histplot(data = df,x='season',ax = axes[0,0],kde = True)
sns.histplot(data = df,x='holiday',ax = axes[0,1],kde = True)
sns.histplot(data = df,x='workingday',ax = axes[0,2],kde = True)
sns.histplot(data = df,x='weather',ax = axes[1,0],kde = True)
sns.histplot(data = df,x='temp',ax = axes[1,1],kde = True)
sns.histplot(data = df,x='atemp',ax = axes[1,2],kde = True)
sns.histplot(data = df,x='humidity',ax = axes[2,0],kde = True)
sns.histplot(data = df,x='windspeed',ax = axes[2,1],kde = True)
sns.histplot(data = df,x='casual',ax = axes[2,2],kde = True)
plt.show()
```



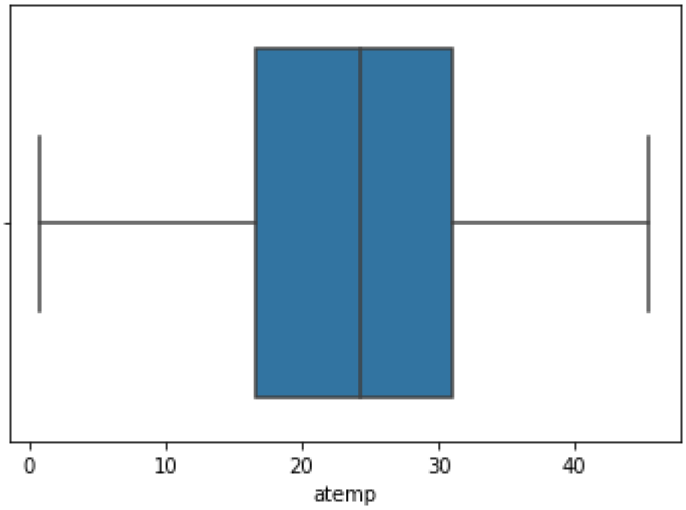
Categorical Columns are easily identified, and other than that Casual, Count, Registered seems to have Outliers.

Checking for Outliers:

```
fig,axes = plt.subplots(nrows = 2,ncols = 3,figsize=(10,15))
cols2 = ['temp','atemp', 'humidity', 'windspeed', 'casual', 'registered', 'count']
sns.boxplot(data = df,x='temp',ax = axes[0,0])
sns.boxplot(data = df,x='humidity',ax = axes[0,1])
sns.boxplot(data = df,x='windspeed',ax = axes[0,2])
sns.boxplot(data = df,x='casual',ax = axes[1,0])
sns.boxplot(data = df,x='registered',ax = axes[1,1])
sns.boxplot(data = df,x='count',ax = axes[1,2])
plt.show()
```



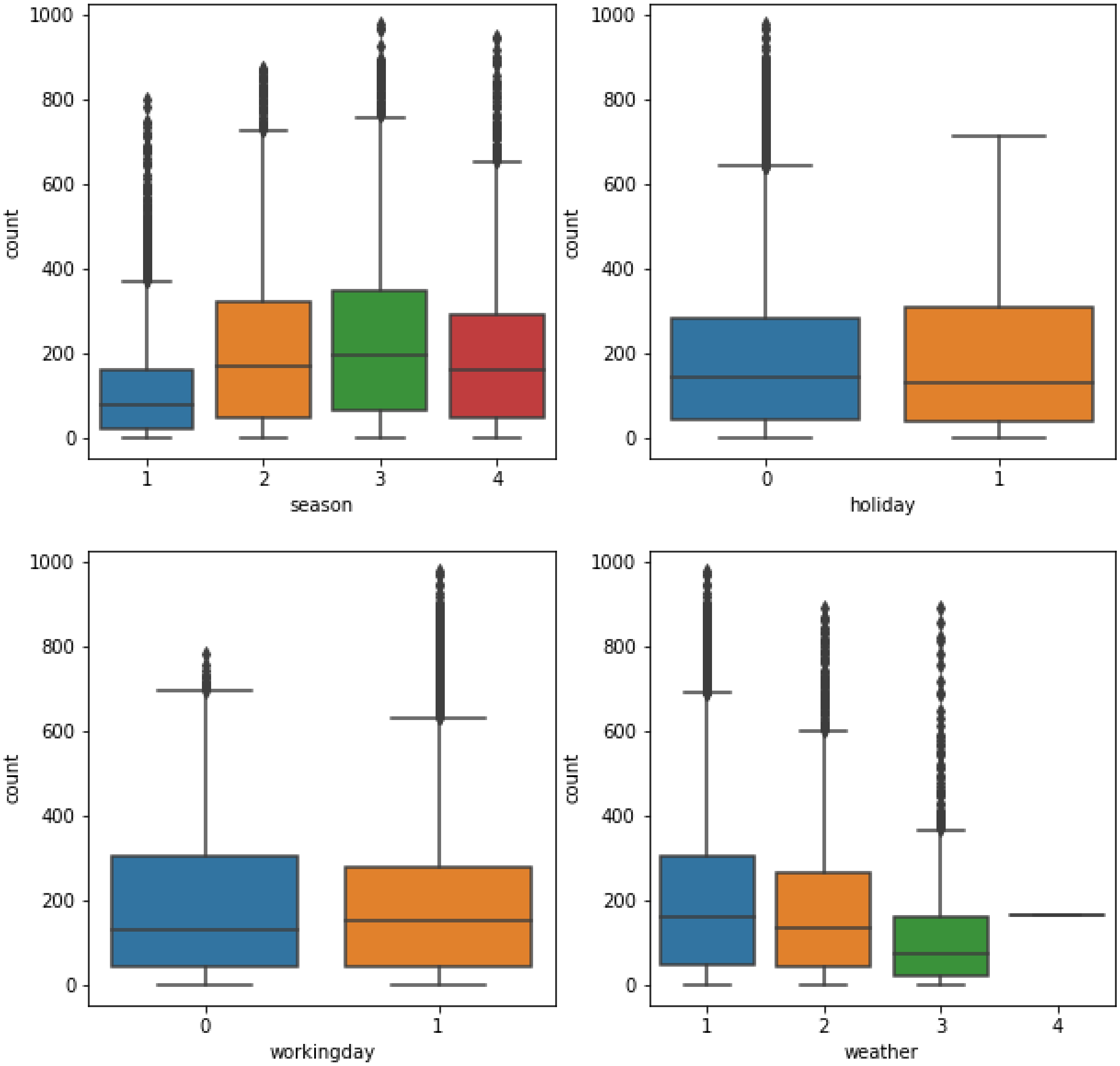
```
sns.boxplot(data = df,x='atemp')
plt.show()
```



From above we can see that, there are outliers in Windspeed, Casual, Registered and Count columns.

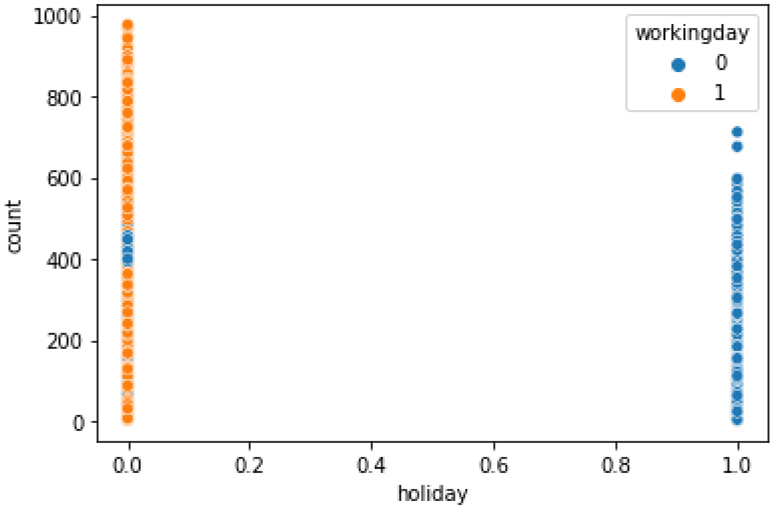
Bi-Variate Analysis:

```
fig, axes = plt.subplots(nrows=2, ncols=2, figsize=(10, 10))
iterator = 0
for row in range(2):
  for col in range(2):
    sns.boxplot(data = df,x=cat_cols[iterator],y='count',ax = axes[row,col])
    iterator += 1
plt.show()
```

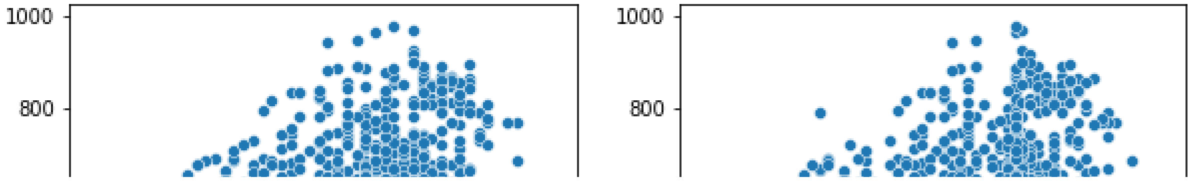


```
sns.scatterplot(data = df,x='holiday',y='count',hue='workingday')
```

```
plt.show()
```



```
fig, axes = plt.subplots(nrows=2, ncols=2, figsize=(10, 10))
iterator = 0
for row in range(2):
  for col in range(2):
    sns.scatterplot(data = df,x=cols2[iterator],y='count',ax = axes[row,col])
    iterator += 1
plt.show()
```



From Uni-Variate / Bi-Variate Analysis we can conclude that:-

1. In Summer season and Fall Seasons, the Bike renting is high.
2. Whenever there is Holiday more Bikes are rented irrespective of Working Day.
3. It can be observed that whenever there is a working day, the Bikes renting count is increased w.r.t Holidays/weekends.
4. We can observe there are countable days in Rainy weather as very few people have rented bikes on these days. Company needs to come up with innovation in these cases.
5. Clear/partly cloud weather or foggy weather more bikes are rented compared to other season.
6. We can observe Scenarios where bike renting is declined and the scenarios are:
    1. Whenever the Temperature is below 10 degree Celius.
    2. Whenever Humidity is less than 20.
    3. Whenever Windspeed is Above 40.

```
# Select an appropriate test to check whether:

# Working Day has effect on number of electric cycles rented
# No. of cycles rented similar or different in different seasons
# No. of cycles rented similar or different in different weather
# Weather is dependent on season (check between 2 predictor variable)

# Set up Null Hypothesis (H0)
# State the alternate hypothesis (H1)
# Check assumptions of the test (Normality, Equal Variance). You can check it
# using Histogram, Q-Q plot or statistical methods like levene’s test,
# Shapiro-wilk test (optional)
# Please continue doing the analysis even If some assumptions fail (levene’s test
# or Shapiro-wilk test) but double check using visual analysis and report
# wherever necessary
# Set a significance level (alpha)
# Calculate test Statistics.
# Decision to accept or reject null hypothesis.
# Inference from the analysis



# Hypothesis Testing (30 Points):
# 2- Sample T-Test to check if Working Day has an effect on the number of electric cycles ren
# ANNOVA to check if No. of cycles rented is similar or different in different 1. weather 2.
# Chi-square test to check if Weather is dependent on the season (10 points) Notebook Quality
```

Hypothesis Testing - I:

2- Sample T-Test to check if Working Day has an effect on the number of electric cycles rented (10 points)

Setting Up Null Hypothesis:

H0 :- Working Day has no effect on Electric Cycles rented.

H1 (Alternate Hypothesis) :- Working Day has effect on Electric Cycle rented.

Significance Level :- 0.05

Test Statistics :- 2-Sample T-Test as only two categorical groups are to be tested.

```
working_day = df[df['workingday'] == 1]['count'].values
holiday = df[df['workingday']==0]['count'].values
```

```
working_day
```

```
     array([  5,   2,   1, ..., 168, 129,  88])
```

```
holiday
```

```
     array([ 16,  40,  32, ..., 106,  89,  33])
```

Performing 2 sample t-test on above data.

```
ttest_ind(working_day,holiday)
```

```
     Ttest_indResult(statistic=1.2096277376026694, pvalue=0.22644804226361348)
```

As p-value comes out to be 0.2264 which is greater than significance value 0.05 so we **FAIL TO REJECT NULL HYPOTHESIS**. And from above plots we cannot come to conclusion that they are dependent/independent factors.

Hypothesis Testing-2

No. of cycles rented similar or different in different seasons.

Null Hypothesis (H0) :- Number of cycles rented is similar in different weather and seasons.

Alternate Hypothesis (H1) :- Number of cycles rented is not similar in different weather and season.

Significance level (alpha) :- 0.05

Test Statistics :- ANOVA (As there are more than two independent groups)

```
print("Season: ",df['season'].nunique())
print("Weather: ",df['weather'].nunique())
```

```
Season:  4
Weather:  4
```

```
print("Season: ",df['season'].value_counts(),"Weather: ",df['weather'].value_counts())
```

```
Season:  4    2734
2    2733
3    2733
1    2686
Name: season, dtype: int64 Weather:  1    7192
2    2834
3     859
4       1
Name: weather, dtype: int64
```

As both the columns have 4 unique values of so need to be extracted individually from the dataframe.

1. season: season (1: spring, 2: summer, 3: fall, 4: winter)
2. weather:

   1: Clear, Few clouds, partly cloudy, partly cloudy

   2: Mist + Cloudy, Mist + Broken clouds, Mist + Few clouds, Mist

   3: Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds

   4: Heavy Rain + Ice Pallets + Thunderstorm + Mist, Snow + Fog

```
spring = df[df['season'] == 1]['count'].values
summer = df[df['season'] == 2]['count'].values
fall = df[df['season'] == 3]['count'].values
winter = df[df['season'] == 4]['count'].values
```

```
clear = df[df['season'] == 1]['count'].values
mist = df[df['season'] == 2]['count'].values
snowy = df[df['season'] == 3]['count'].values
rainy = df[df['season'] == 4]['count'].values
```

```
f_oneway(spring,summer,fall,winter,clear,mist,snowy,rainy)
```

```
F_onewayResult(statistic=203.09718069456093, pvalue=3.099184115882506e-293)
```

As p-value is less than alpha value we can **Reject Null Hypothesis** and can conclude that Number of Cycles are rented are not similar to different weather and Seasons.

Hypothesis Testing - 3

To check if Weather is dependent on the season.

Null Hypothesis (H0) :- Weather is independent of Season.

Alternate Hypothesis (H1) :- Weather is dependent of Season.

Significance Level (Alpha) :- 0.05

Test Statistics :- Chi-Square Test

```
cross_table = pd.crosstab(df['season'], df['weather'])
cross_table
```

weather	**1**	**2**	**3**	**4**
season				
1	1759	715	211	1
2	1801	708	224	0
3	1930	604	199	0
4	1702	807	225	0

```
val = chi2_contingency(cross_table)
```

```
nrows, ncols = 4, 4
dof = (nrows-1)*(ncols-1)
print("degrees of freedom: ", dof)
alpha = 0.05
chi_sqr = sum([(o-e)**2/e for o, e in zip(cross_table.values,expected_values)])
chi_sqr_statistic = chi_sqr[0] + chi_sqr[1]
print("chi-square test statistic: ", chi_sqr_statistic)
critical_val = chi2.ppf(q=1-alpha, df=dof)
print(f"critical value: {critical_val}")
p_val = 1-chi2.cdf(x=chi_sqr_statistic, df=dof)
print(f"p-value: {p_val}")
```

```
degrees of freedom:  9
```

```
chi-square test statistic:  44.09441248632364
critical value: 16.918977604620448
p-value: 1.3560001579371317e-06
```

As p-value is less than significance level, we **Reject Null Hypothesis** .

And also we can conclude that, Weather is dependent on Season.

Observations:

1. In Summer season and Fall Seasons, the Bike renting is high.
2. Whenever there is Holiday more Bikes are rented irrespective of Working Day.
3. It can be observed that whenever there is a working day, the Bikes renting count is increased w.r.t Holidays/weekends.
4. We can observe there are countable days in Rainy weather as very few people have rented bikes on these days. Company needs to come up with innovation in these cases.
5. Clear/partly cloud weather or foggy weather more bikes are rented compared to other season.
6. We can observe Scenarios where bike renting is declined and the scenarios are:
    1. Whenever the Temperature is below 10 degree Celius.
    2. Whenever Humidity is less than 20.
    3. Whenever Windspeed is Above 40.

Recommendations:

1. As we have observed that in Summer and Fall Seasons more bikes are rented at different region, so using Regions data we can conclude to re-shuffling of Bikes from one Place to Other Place (High Demanding). And similarly, as we can see drastic declined in renting bikes of other season, company can come-up with different offers or some innovations in Bikes that can be more suitable at those Seasons.
2. And Similarly no the company should preserve there Bikes accordingly as during High Windspeed day, Low temperature and Humidity day the Bikes renting are down, so the Company might store them and used them in other weathers/seasons when there is high amount of Demand of Bikes.
3. From Hypothesis Testing - 1, we can conclude that Working Day might have some effect on renting Bikes but due to lack of evidences we cannot come to conclusion.
4. From Hypothesis Testing - 2, we can conclude that In Each different Seasons and Weathers there is always some distinguish amount of difference can be observed and they are not

similar always changing.

5. From Hypothesis Testing - 3: we can conclude that Weather is Dependent on Seasons.

✓ 4s completed at 2:33 PM ● ×