Colab notebook link:
https://colab.research.google.com/drive/1K6dLptT8c7nOMnsGIbtcg93fMS9o-wZJ?usp=sharing

Insights:

Analyzing the dataset:

- There are total 10886 rows and 12 columns.

- Columns are: ['datetime', 'season', 'holiday', 'workingday', 'weather', 'temp', 'atemp', 'humidity', 'windspeed', 'casual', 'registered', 'count']

- Data type of columns: 8 integer type columns, 3 float type columns and 1 object type columns given in the dataset. But we can observe that 'datetime' column is datetime type and also 'season', 'holiday', 'workingday' and 'weather' columns are categorical type.
  So, it's better to change the data types of these 5 columns.

- Replacing the values of category columns to names for better understanding.
  - Season column: 'spring' for 1, 'summer' for 2, 'fall' for 3 and 'winter' for 4.
  - Holiday column: 'holiday' for 1 and 'not holiday' for 0.
  - Workingday column: 'working day' for 1 and 'holiday/weekend' for 0.
  - Weather column: Not changing the values of weather column, the reason is the length of name is too long.

Statistical Summary

- Date time:
  - The data is given in dataset from date: 2011-01-01 to date: 2012-12-19.
  - Total 718 days data is given in the dataset.

- Season:
  - Total 4 seasons in the dataset.
  - Winter season mentioned in the dataset maximum time with 2734 frequencies.

- Holiday:
  - Total 10575 rows have not holiday in dataset.

- Working day:

- Maximum bikes are rented in working day.

- Weather:
  - There are total 4 weather categories mentioned in the dataset.
  - And most of the time weather category 1 means “Clear, Few clouds, partly cloudy, partly cloudy” weather mentioned in dataset.
- Temp:
  - Maximum temp 41 degree Celsius and minimum temp 0.82 degree Celsius recorded in the dataset with mean temp 20 degree Celsius.
  - Noticeable thing is the mean and median are very far away to one another and the value of standard deviation is also high it means that there is high variance in the data of these attributes. Temp column might have outliers.

- Feeling Temp:
  - Maximum feeling temp 45.45 degree Celsius and minimum feeling temp 0.76 degree Celsius as per dataset.
  - Feeling temp column also have outliers because the mean and median is very far away from one another.

- Humidity:
  - As per dataset, mean of all recorded humidity is 61.88.

- Wind speed:
  - Max wind speed recorded in the dataset is 57(approx.)

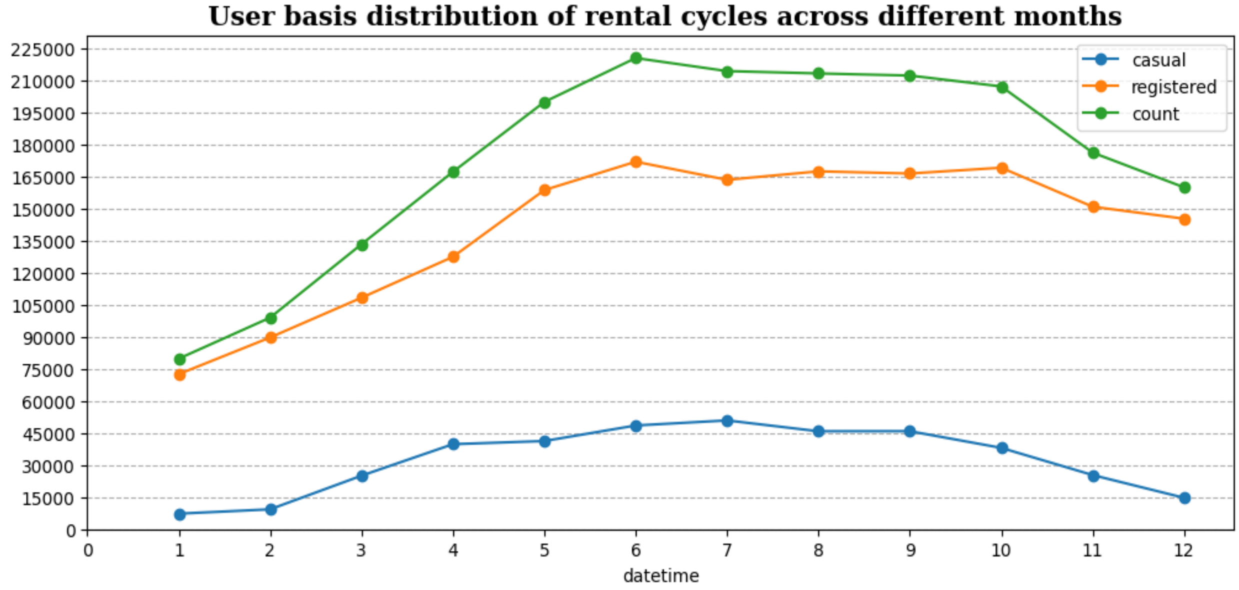
- Casual users:
  - Percentage of casual users is approx. 19% in total users.

- Registered users:
  - Around 81% of total users is registered users.

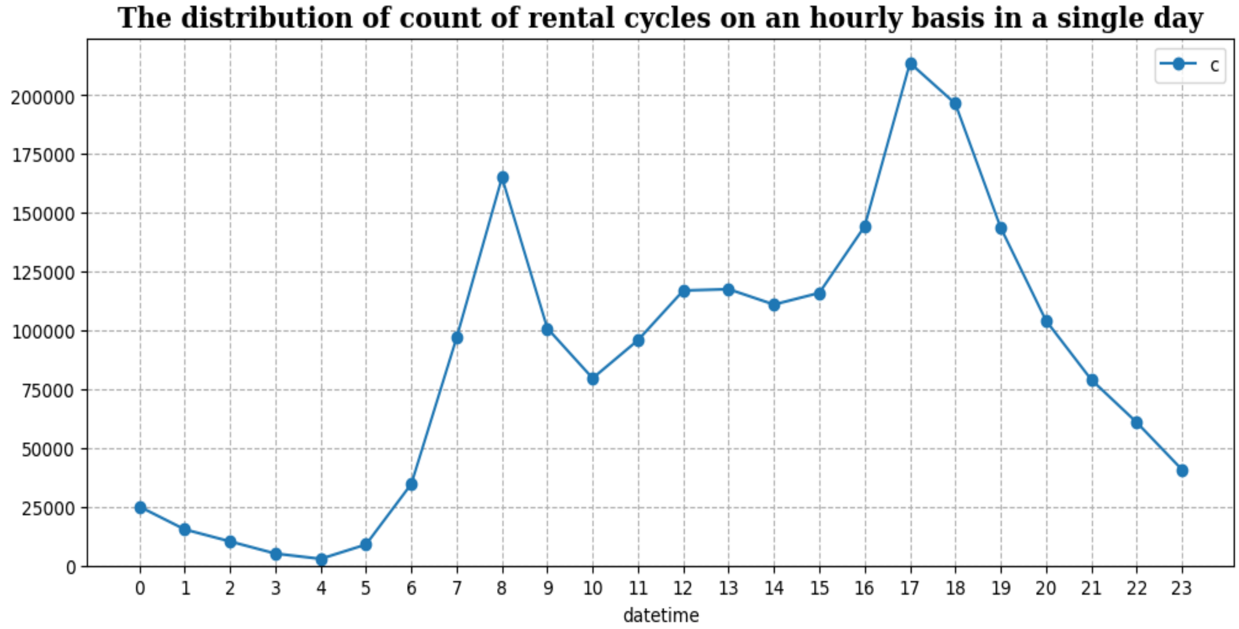
Missing value & Duplicates:

- There are no missing values in the dataset.
- There are no duplicates in the dataset.

Date Time column Analysis:



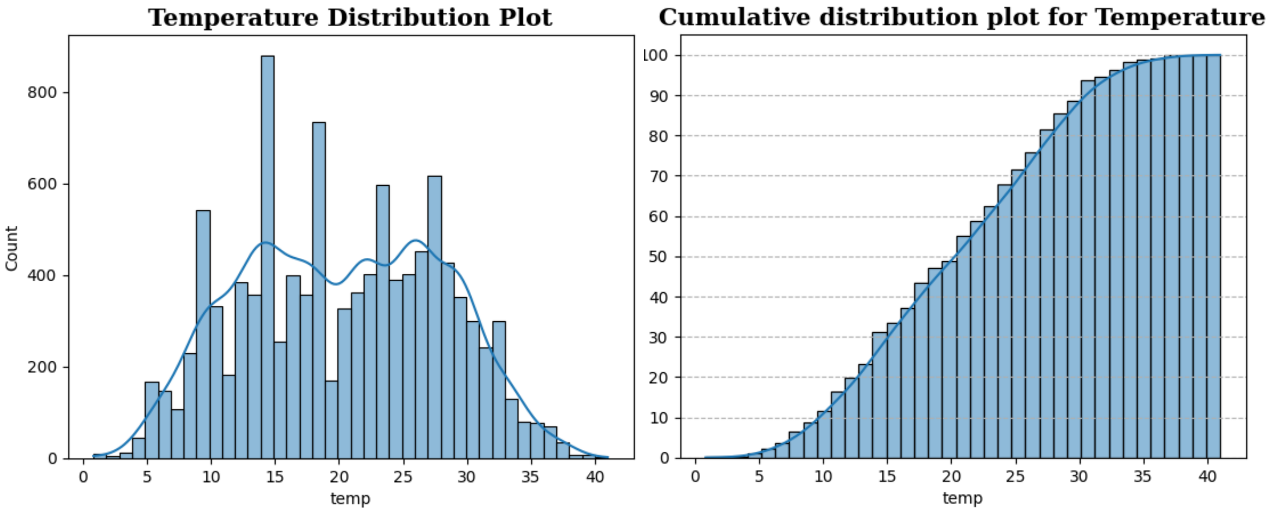
- ***Monthly Basis:***
  - The count of rental cycles shows an increasing trend from January to June
  - From July to September, there is a slight decrease in the count of rental cycles, with negative growth rates
  - The count of rental cycles further declines from October to December, with the largest drop observed between October and November.



- ***Hourly Basis:***
  - The count of rental cycles is the highest at hours 17 (5 PM) followed by hours 18 (6 PM) and hours 8 (8 AM) of the day.
  - The count of rental cycles is the lowest at hours 4 (4 AM) followed by hours 3 (3 AM) and hours 5 (5 AM) of the day.
  - During the early morning hours (hours 0 to 5), there is a significant decrease in the count, with negative growth.
  - However, starting from hour 5 (5 AM), there is a sudden increase in count, with a sharp positive growth.
  - Observed from hour 4 to hour 5 (4 AM to 5 AM), the count continues to rise significantly until reaching its peak at hour 17 (5 PM), compared to the previous hour.
  - After hour 17 (5 PM) means from late evening, there is a gradual decrease in count, with negative growth.
  - These patterns indicate that there is a distinct fluctuation in count throughout the day, with low counts during early morning hours, a sudden increase in the morning, a peak count in the afternoon, and a gradual decline in the late evening and nighttime.

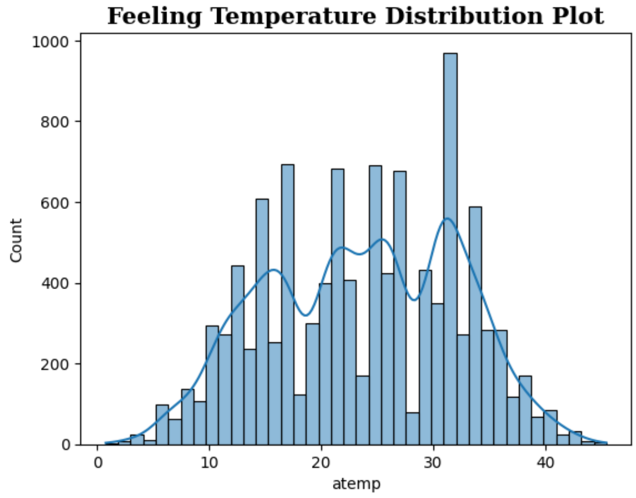
Univariate Analysis:

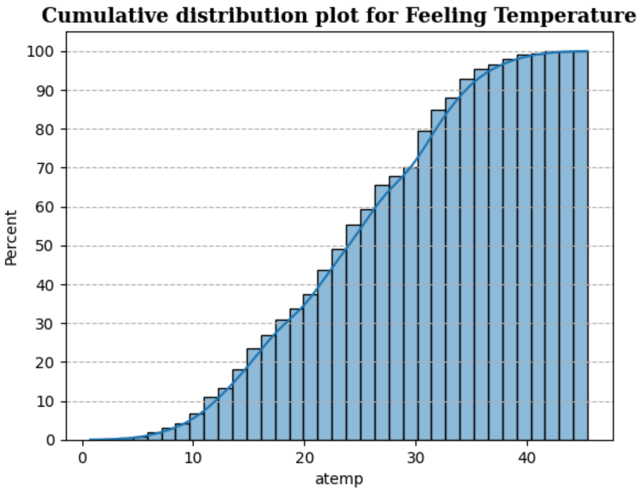
Distribution plot for continuous variables:



- *Temperature:*
  - Temperature distribution graph looks like follows normal distribution.

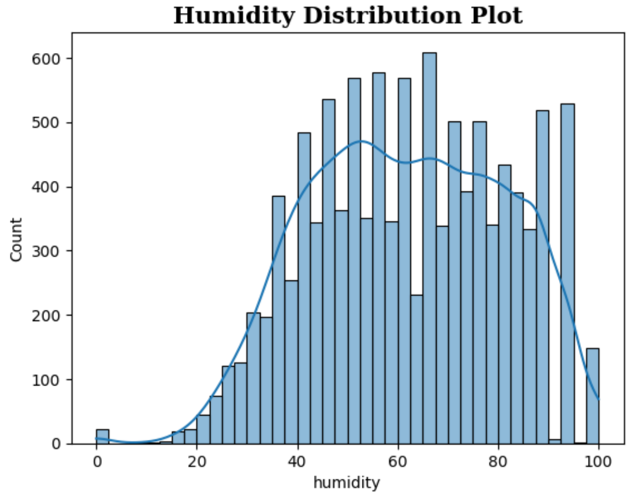
- The mean temperature is 20.23 degree Celsius with standard deviation 7.79 degree Celsius.
- More than 80% of time the temperature is nearly less than 28 degree Celsius.

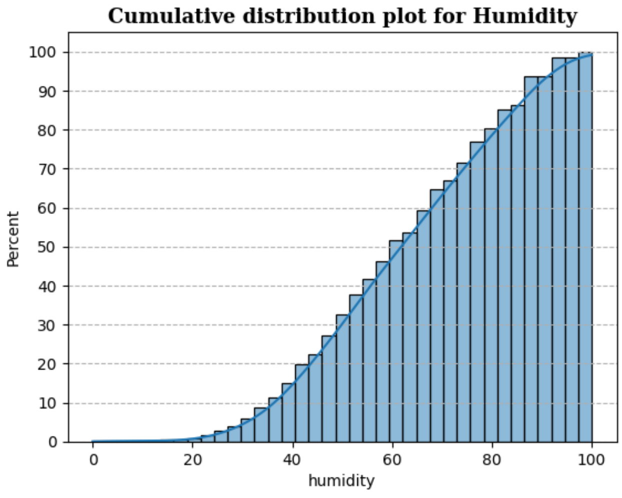




➢ *Feeling Temperature(atemp):*

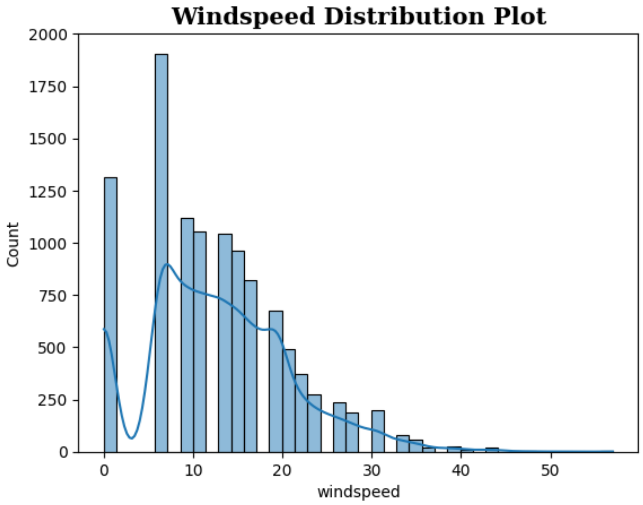
- Feeling temperature chart also looks like follows normal distribution.
- The mean feeling temperature is 23.66 degree Celsius with standard deviation 8.47 degree Celsius.
- Nearly 70% of time the feeling temperature is less than 31 degree Celsius.

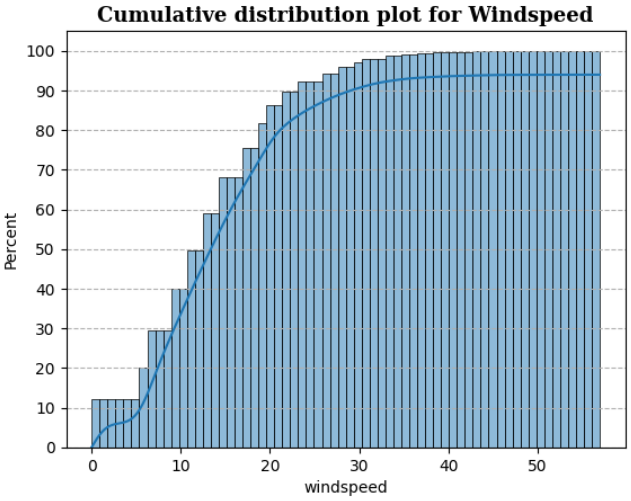




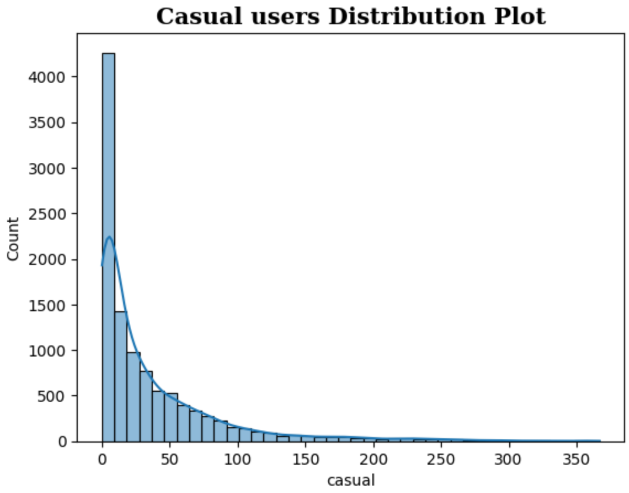
➢ *Humidity:*

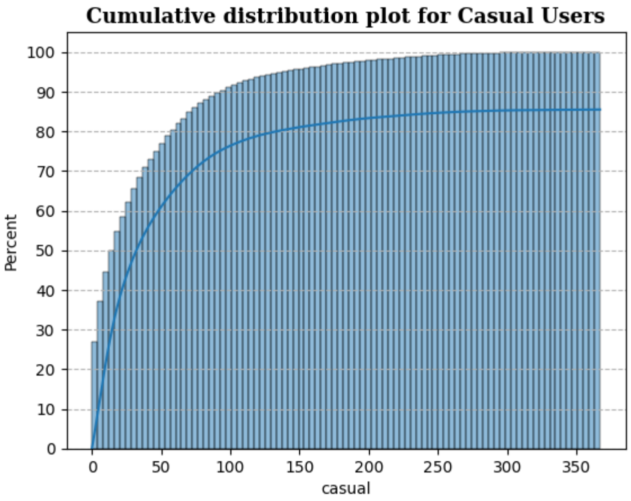
- Humidity chart also looks like follows normal distribution, further we will check it.
- The mean and standard deviation of the humidity is 61.89 and 19.25 respectively.
- Nearly 80 % of the time, the humidity value is greater than 40.



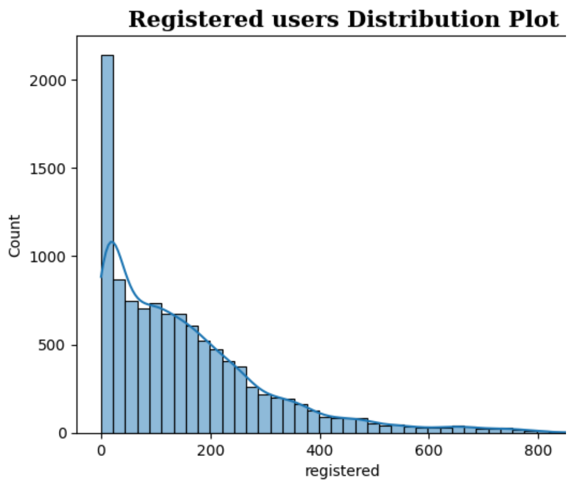


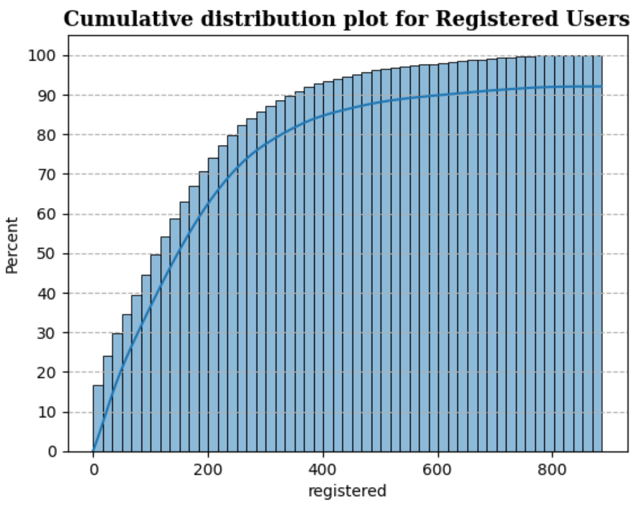
- ➢ *Wind Speed:*
  - Windspeed chart looks like right skewed.
  - The mean and standard deviation of wind speed is 12.8 and 8.16 respectively.
  - Nearly 80 % of the total windspeed data has a value of less than 20.



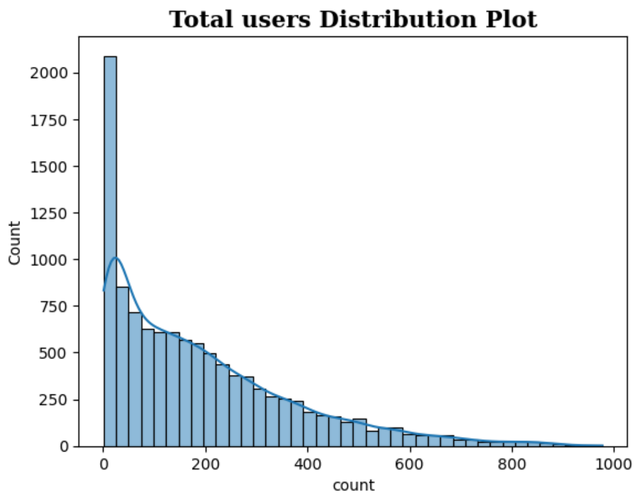


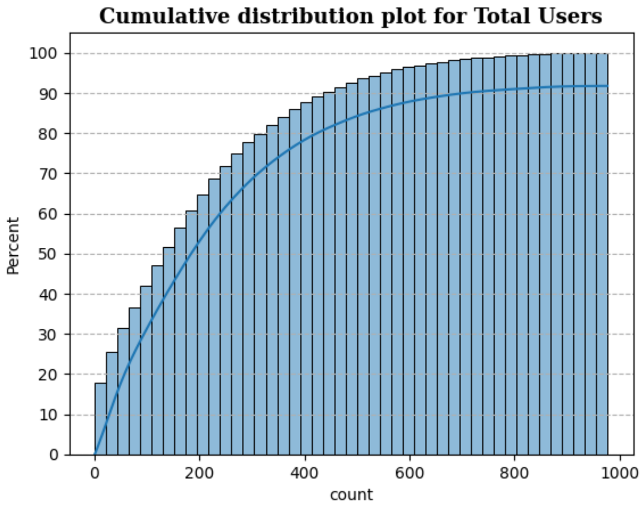
- ➢ *Casual Users:*
  - Count of casual users chart looks like right skewed.
  - The mean value of casual users count is 36.02.
  - Nearly 80% of the time, the count of casual users is less than 60.





- *Registered Users:*
  - Count of registered users chart looks like right skewed.
  - The mean value of count registered users count is 155.55.
  - Nearly 80% of the time, the count of casual users is nearly less than 250.

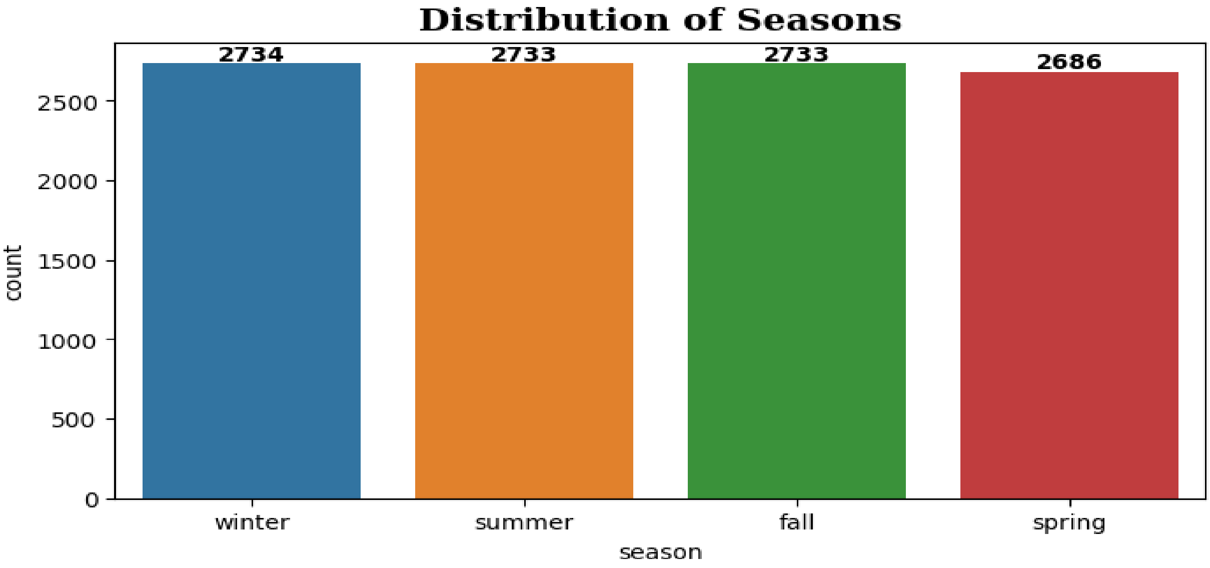




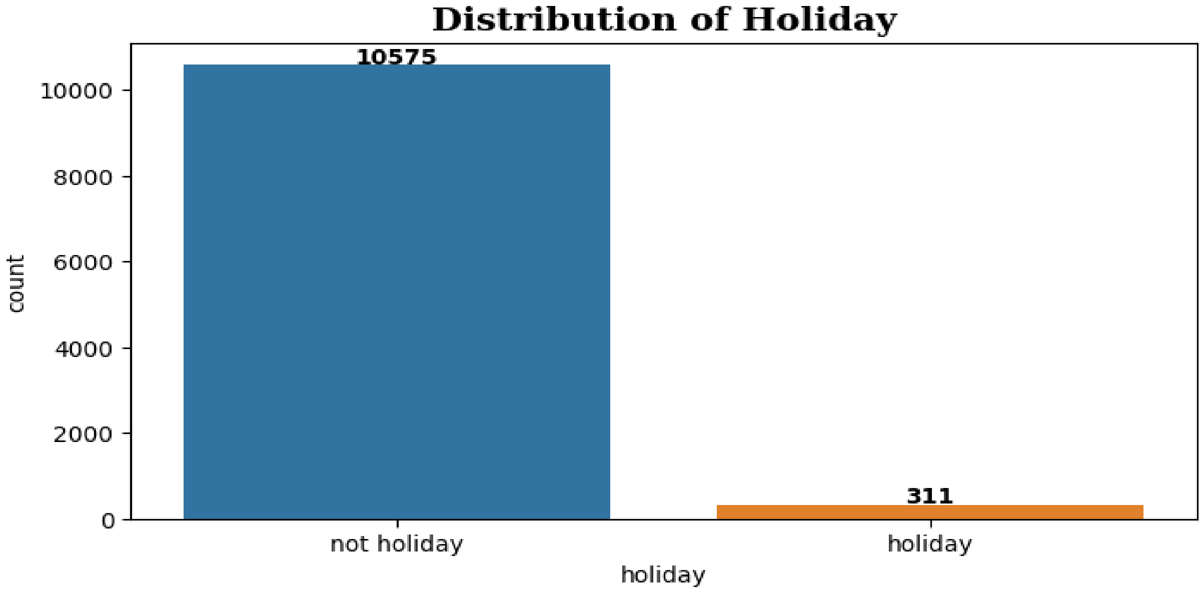
.

- *Total of Casual and Registered Users(count):*
  - The mean value total count of users is 191.574.
  - Nearly 80% of the time, the count of total casual and registered users is nearly less than 300.

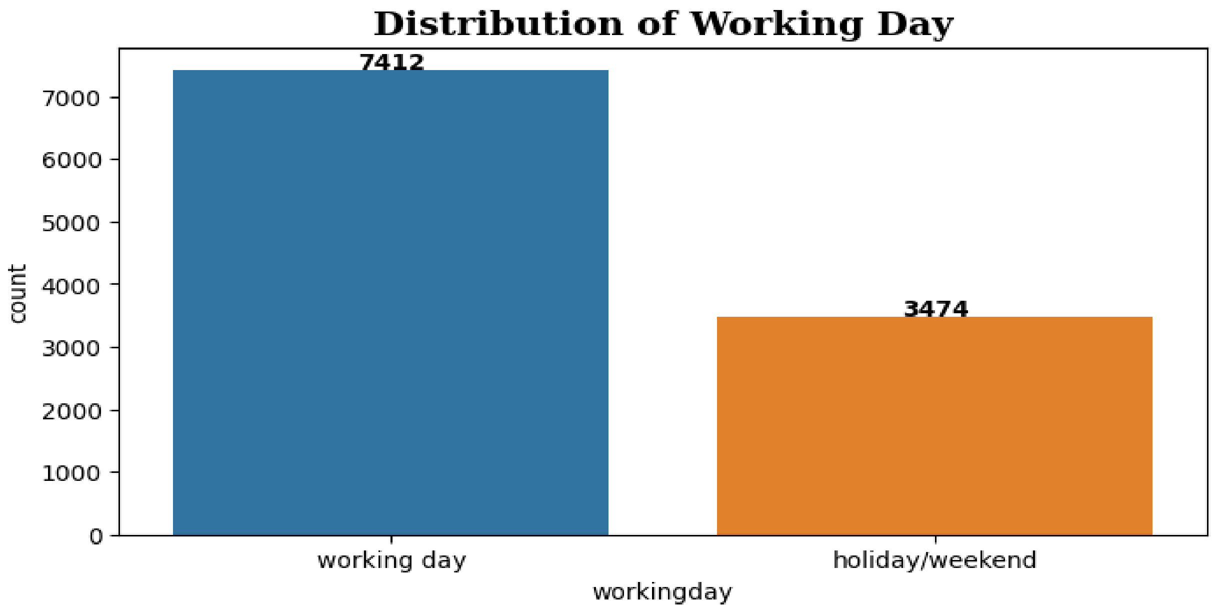
Distribution plot for continuous variables:



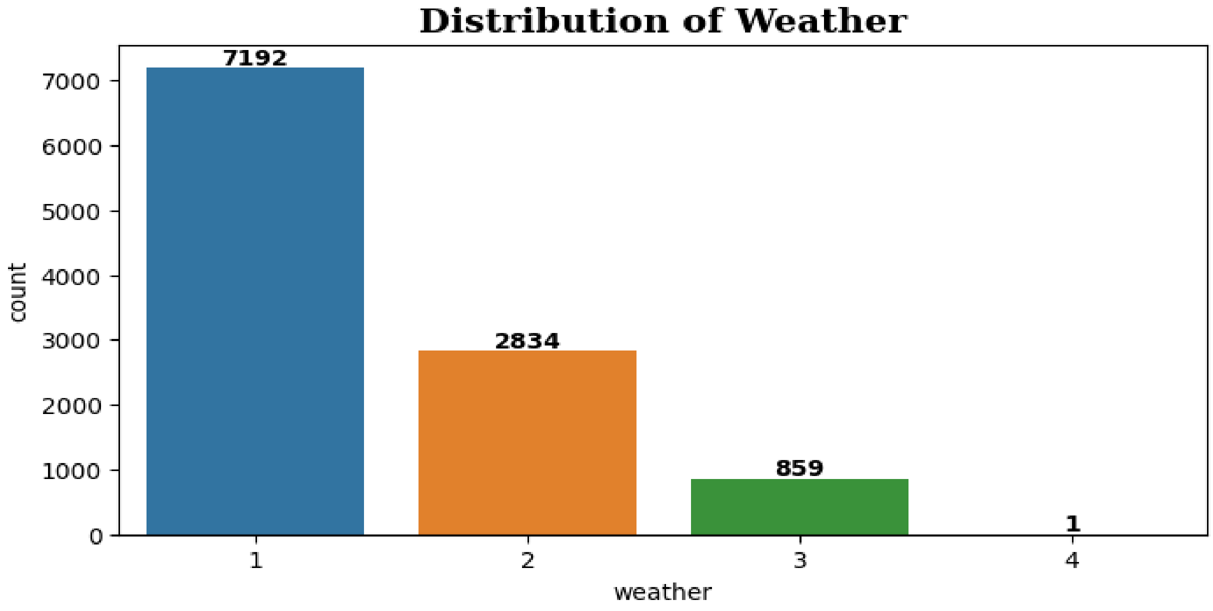
- *Season:*
  - The number of days nearly same for all season.



- *Holiday:*
  - As per given data, no. of holiday is 311 only and non holiday is 10575, which 34 times more than holiday.

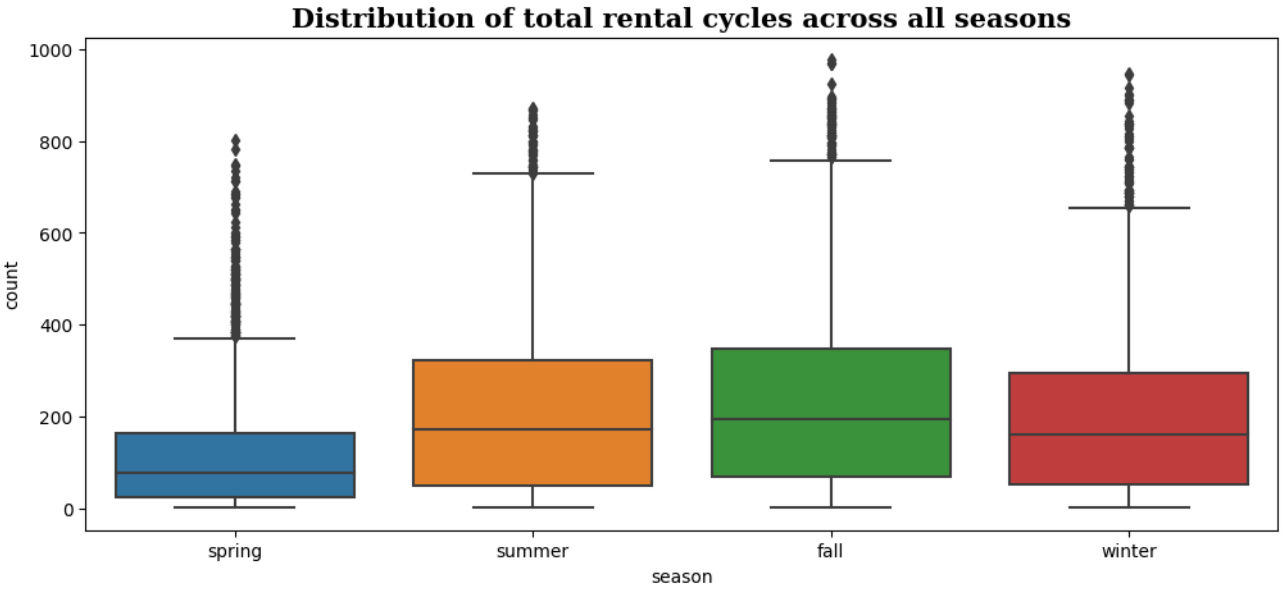


- *Working Day:*
  - As per given data, most of the days are working days.
  - If we compare the holiday chart and working day chart then we can find out that total no of weekend is (3474-311) = 3163 days.



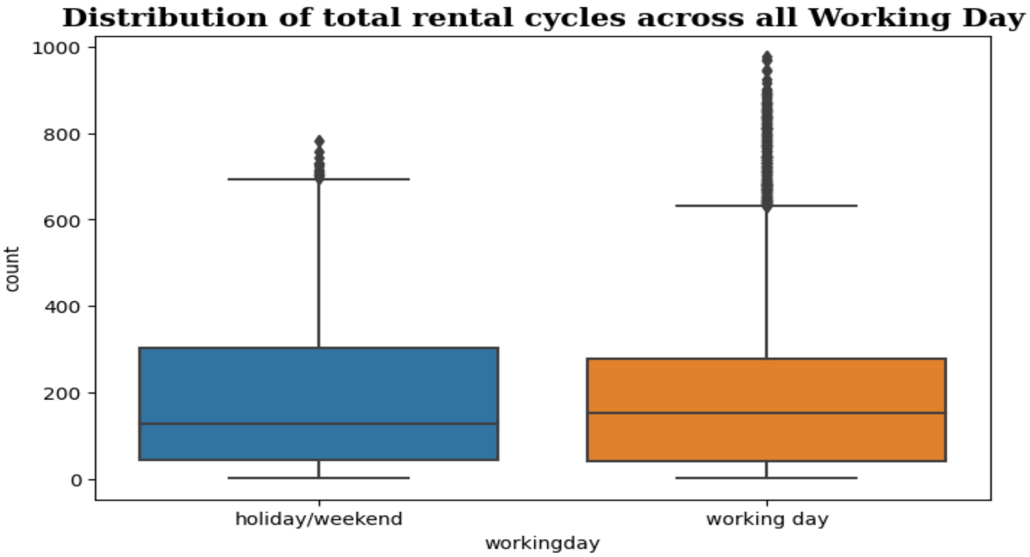
- *Weather:*
  - Most of the days weather is 1 means (clear, few clouds, partly cloudy)

Bivariate Analysis for Important variables:



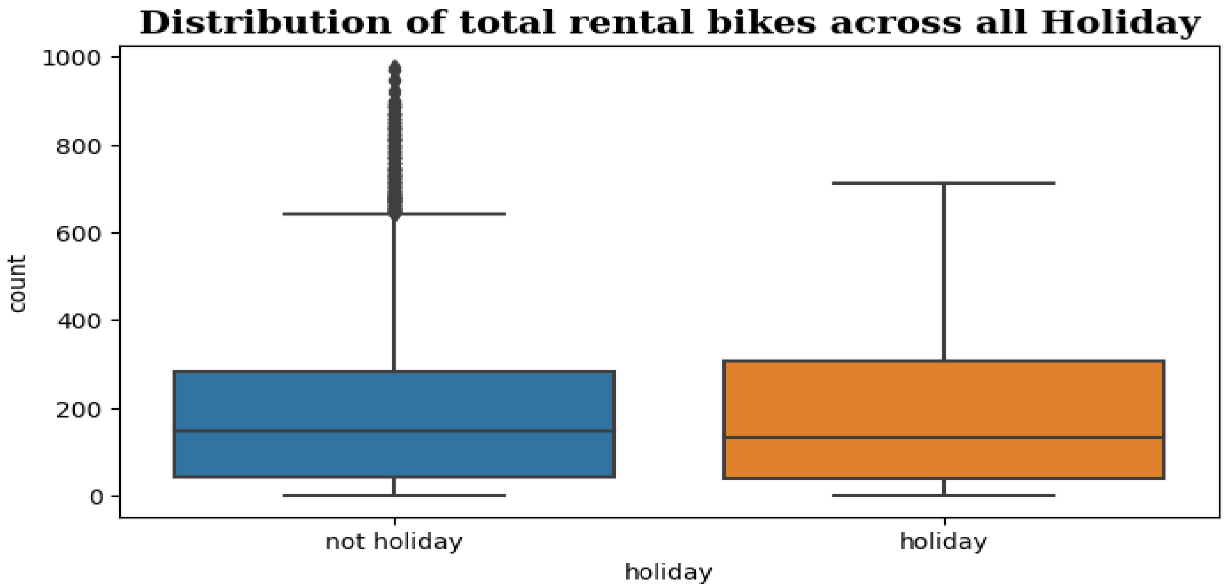
➢ *Season vs. Count:*

- The total rental cycles are higher in the 'fall' season as compare to the other seasons.
- In second position 'summer' season, this is followed by 'winter' seasons.
- It is generally low in the spring season.

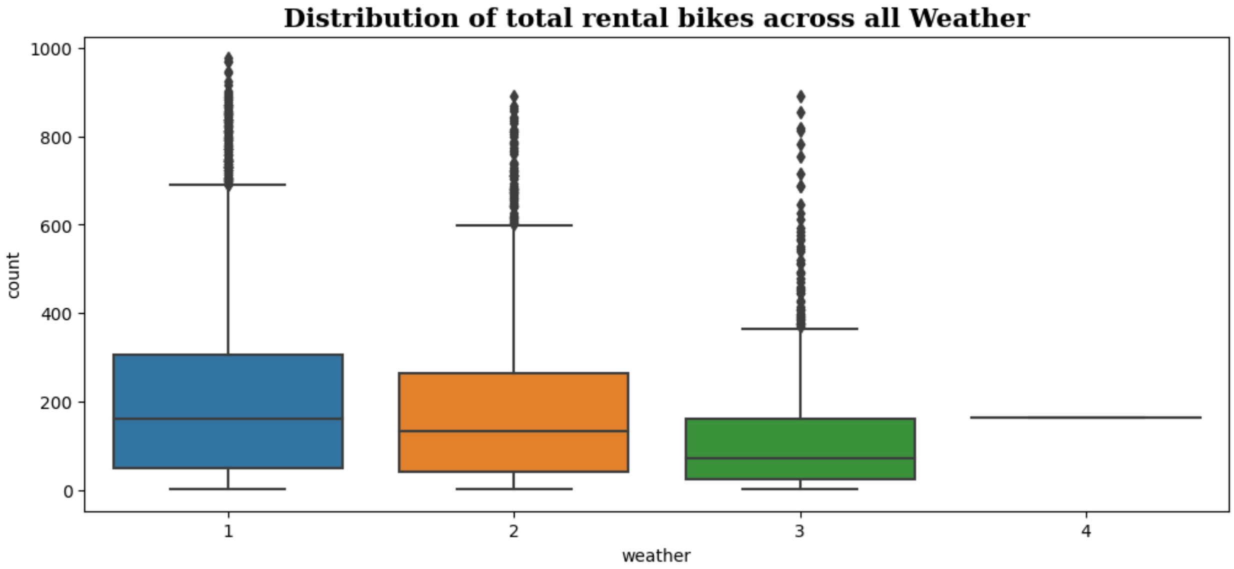


➢ *Working day vs. Count:*

- From the chart, we can analyze, in holiday or weekend slightly more cycles were rented.

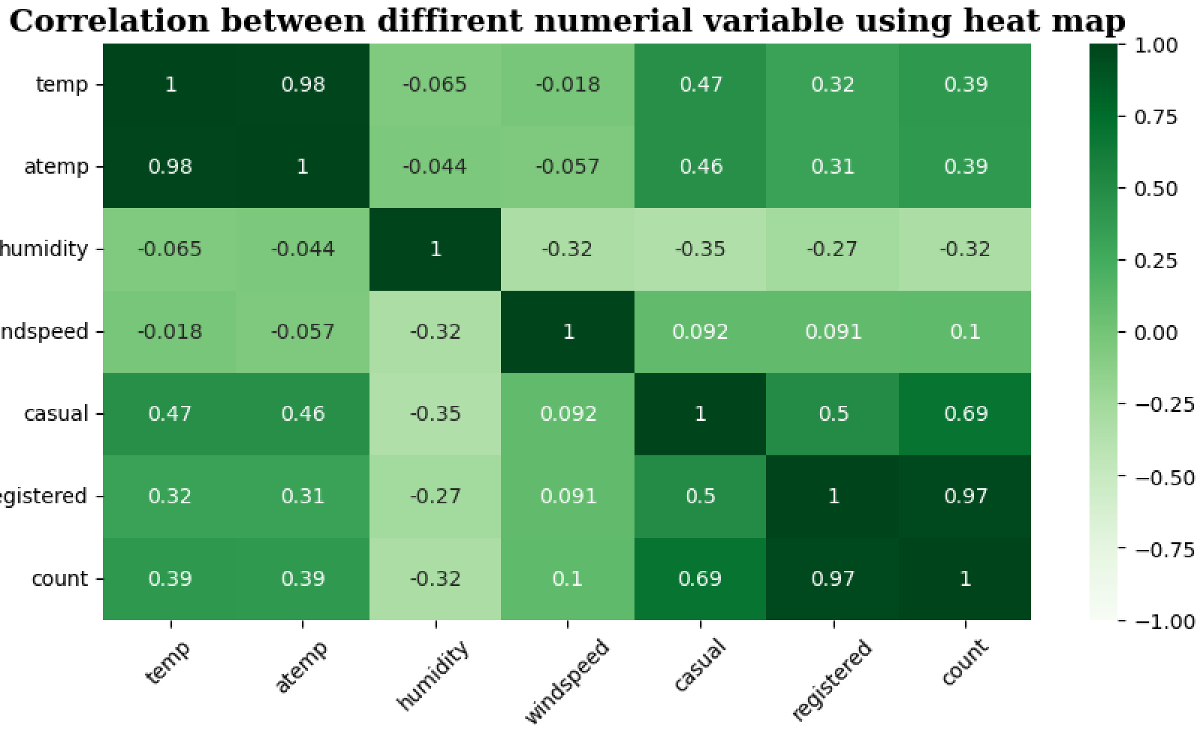


- *Holiday vs. Count:*
  - In holiday also, a good number of cycles were rented.



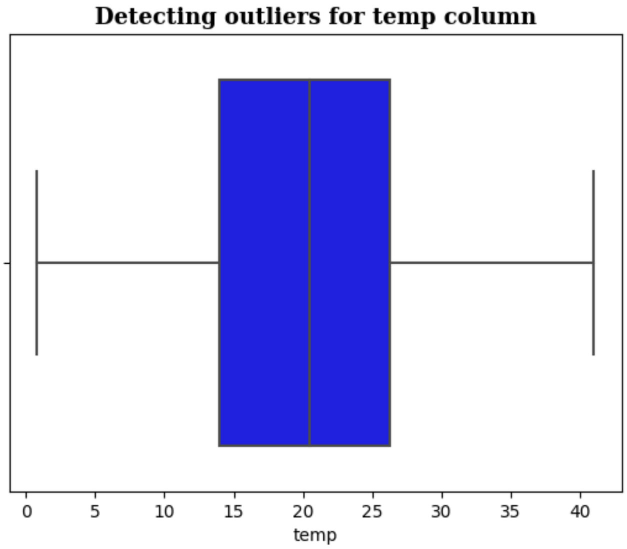
- *Weather vs. Count:*
  - From the chart, it is clear that with weather the number of cycles rent drastically changed.
  - In good weather means clear, few clouds, partly cloudy weather, the highest number of cycles are rented.
  - And in bad weather means heavy rain, ice pallets, thunderstorm, mist, snow or fog weather, we can say no one taking rent the cycle.

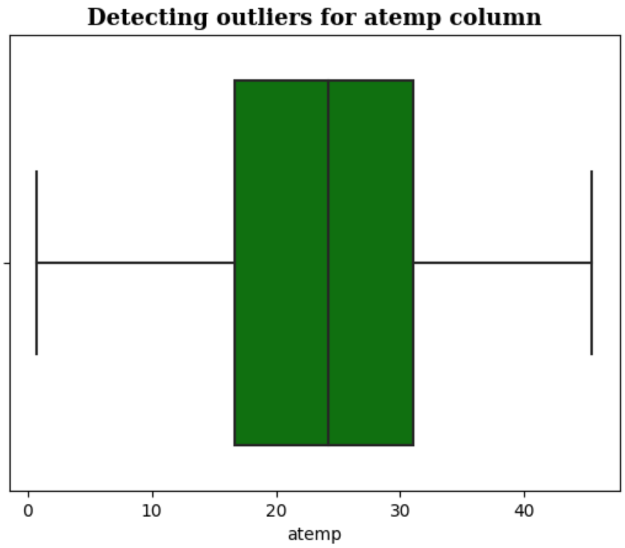
Multivariate Analysis:



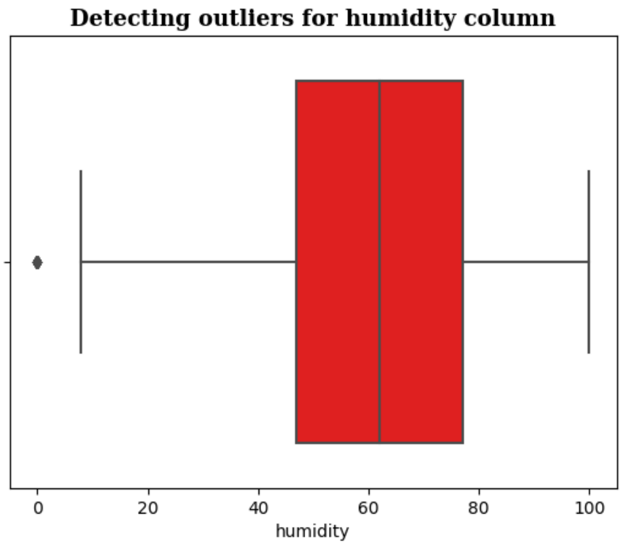
- *Correlation between different numerical variable using Heat-map:*
  - Very high correlation means greater than 0.9 exists between columns atemp-temp, and count-registered.
  - Moderate positive correlation means 0.5-0.7 exists between columns casual-count and casual-registered.
  - Low positive correlation means 0.3-0.5 exists between columns count-temp, count-atemp and casual-atemp.
  - Negative correlation exists between all other combinations of columns like humidity-count, humidity-registered, humidity-casual etc.

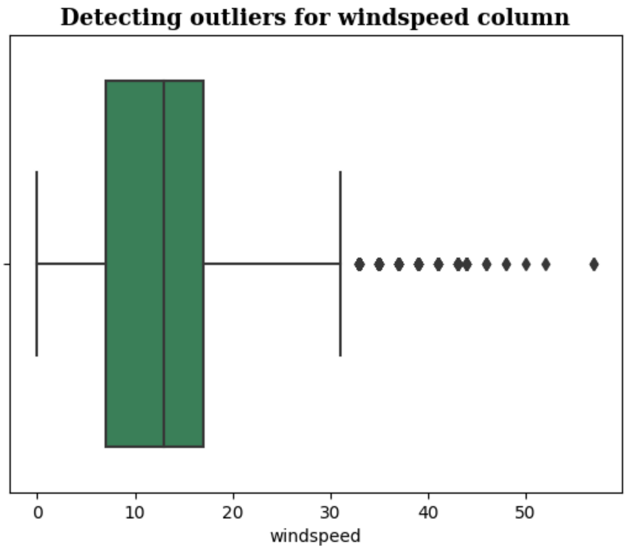
Outliers Detection:


Detecting outliers for temp column

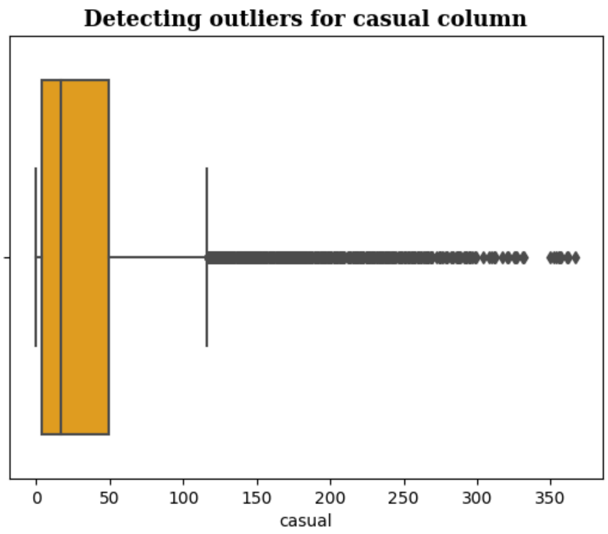

Detecting outliers for atemp column

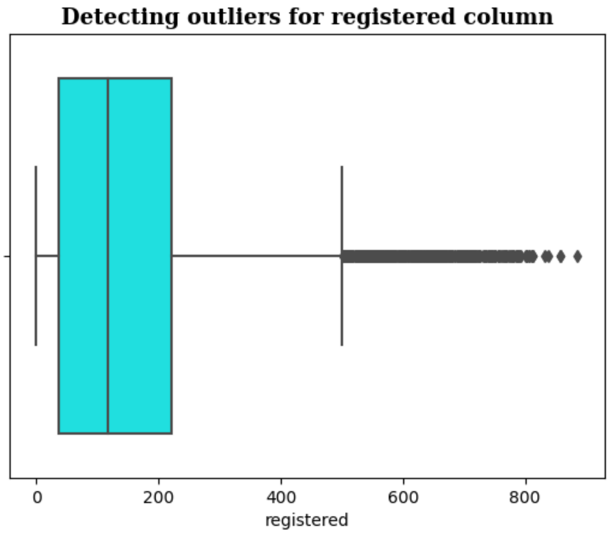
- ➢ *temp column*: No outliers in temp column.
- ➢ *atemp column*: No outliers in atemp column.


Detecting outliers for humidity column

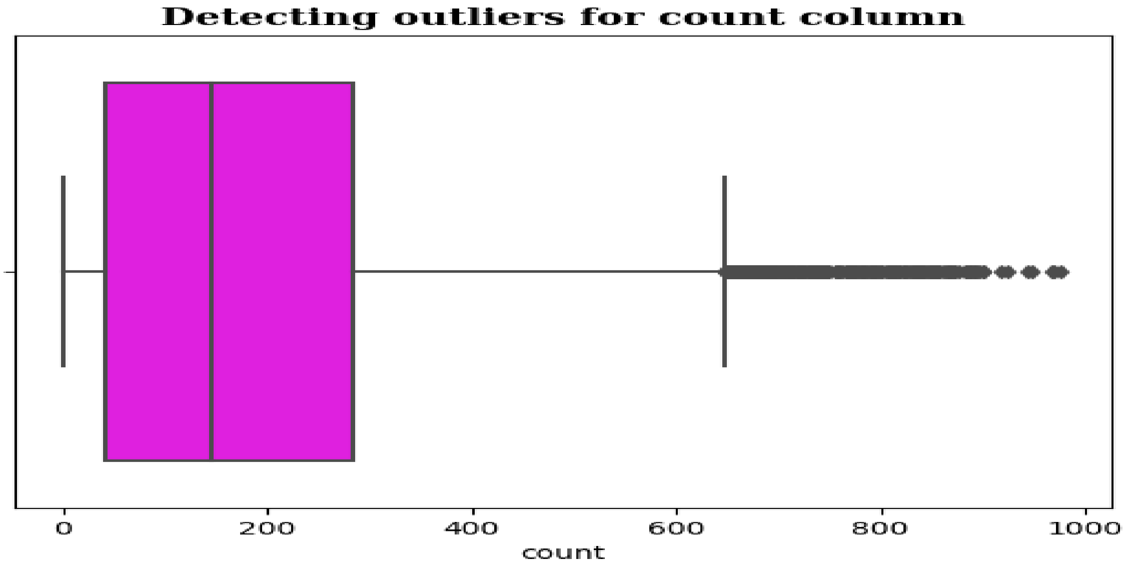

Detecting outliers for windspeed column

- ➢ *humidity column*: Very few outliers in humidity column.
- ➢ *windspeed column*: Many outliers in windspeed column





- *casual column*: Many outliers in casual column.
- *registered column:* Many outliers in registered column.



- *count column*: Many outliers in count column.

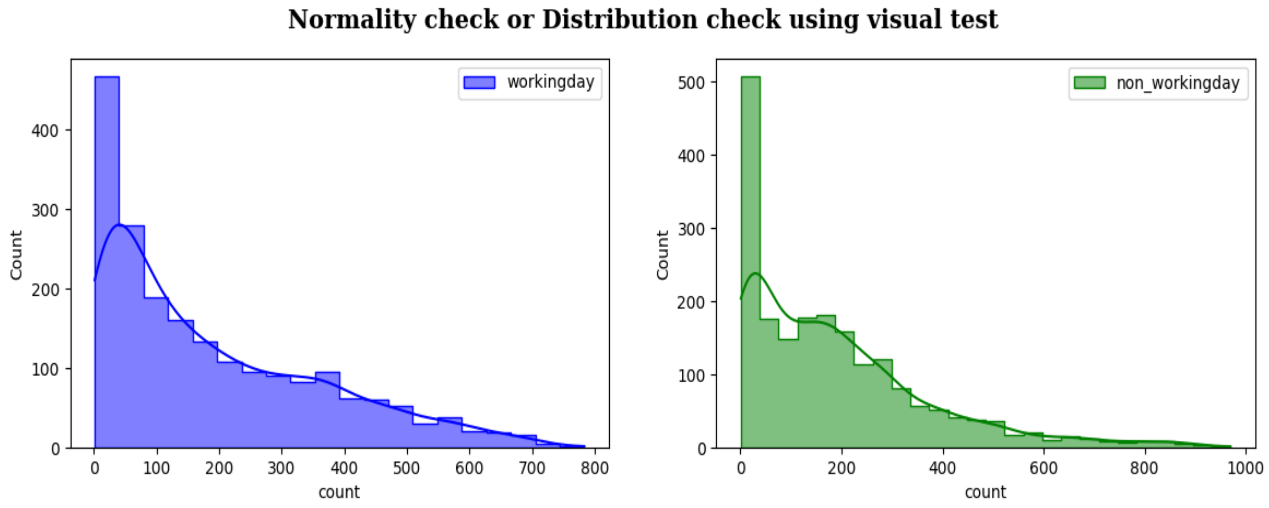
Hypothesis Testing:

Our target column is ‘count’ column.

Working Day vs. Count

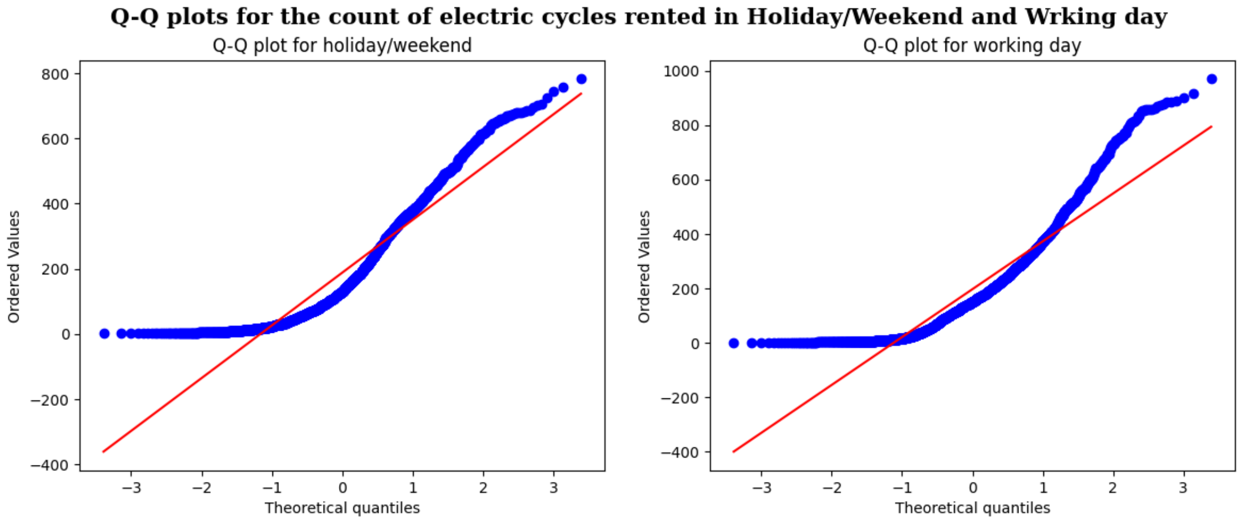
- **Step 1: Setup Null Hypothesis**
  - Null Hypothesis (Ho): Working Day does not have any effect on the number of electric cycles rented.
  - Alternative Hypothesis (Ha): Working Day has effect on the number of electric cycles rented.

- **Step 2: Checking the assumption for the test.**



Normality check or Distribution check using visual test

  - *Normality or Distribution check using histogram or visual test:*
    - We can visualize from the above histogram plot, that the distribution doesn’t follow normal distribution.



Q-Q plots for the count of electric cycles rented in Holiday/Weekend and Wrking day

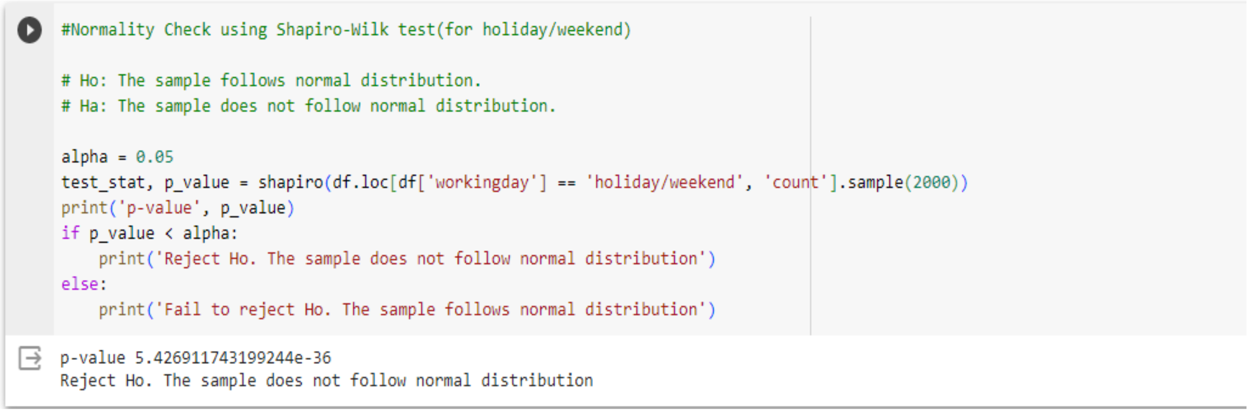
Q-Q plot for holiday/weekend

Q-Q plot for working day

- *Normality or Distribution check using Q-Q plot test:*
    - We can figure it from the above Q-Q plot that the distribution doesn't follow normal distribution.

From the above plots we have seen that the samples do not come from normal distribution. Now we are applying another test Shapiro-wilk test for normality.

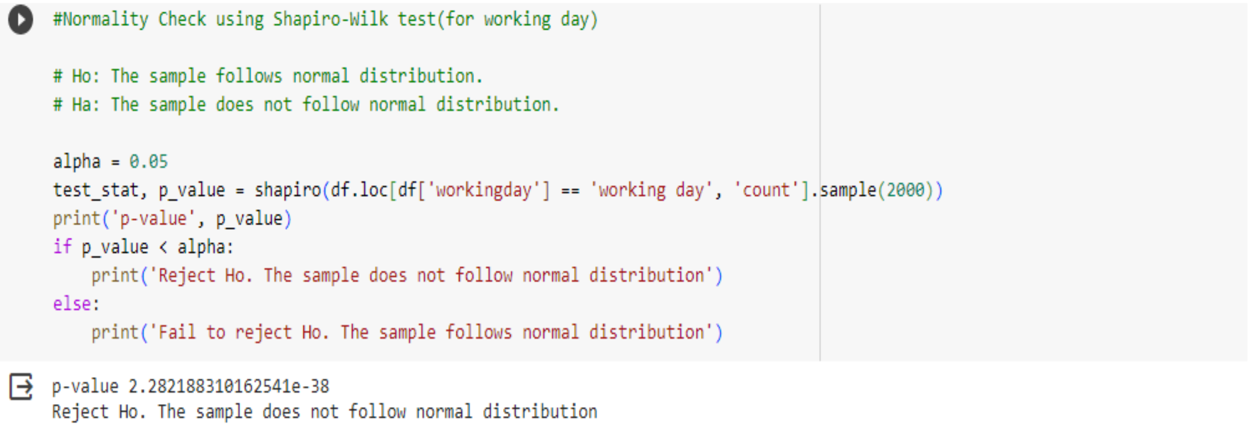
- *Normality check using Shapiro-wilk test:*



```
#Normality Check using Shapiro-Wilk test(for holiday/weekend)

# Ho: The sample follows normal distribution.
# Ha: The sample does not follow normal distribution.

alpha = 0.05
test_stat, p_value = shapiro(df.loc[df['workingday'] == 'holiday/weekend', 'count'].sample(2000))
print('p-value', p_value)
if p_value < alpha:
    print('Reject Ho. The sample does not follow normal distribution')
else:
    print('Fail to reject Ho. The sample follows normal distribution')
```

```
p-value 5.426911743199244e-36
Reject Ho. The sample does not follow normal distribution
```

- For holiday/weekend, the test result is: p-value 5.426911743199244e-36, the sample does not follow normal distribution.



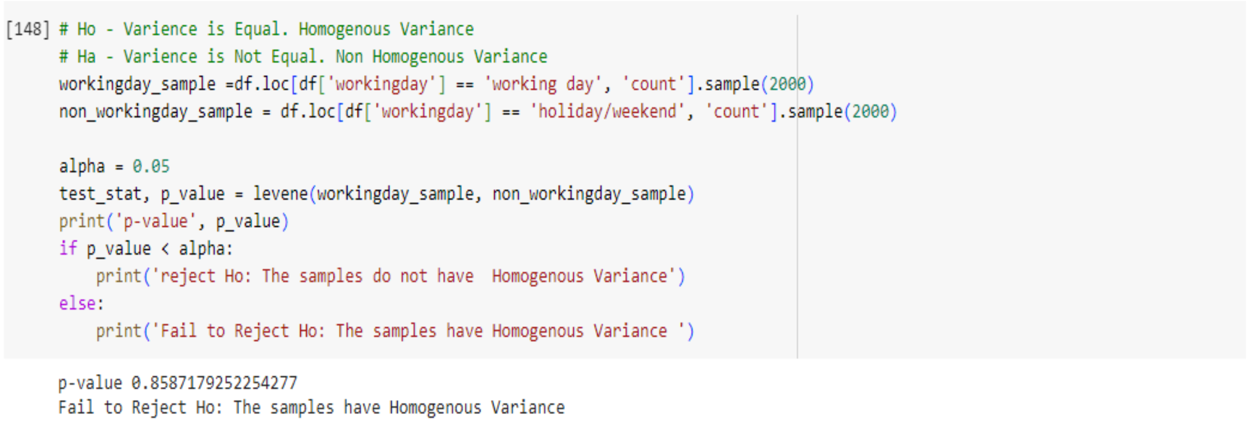
```
#Normality Check using Shapiro-Wilk test(for working day)

# Ho: The sample follows normal distribution.
# Ha: The sample does not follow normal distribution.

alpha = 0.05
test_stat, p_value = shapiro(df.loc[df['workingday'] == 'working day', 'count'].sample(2000))
print('p-value', p_value)
if p_value < alpha:
    print('Reject Ho. The sample does not follow normal distribution')
else:
    print('Fail to reject Ho. The sample follows normal distribution')
```

```
p-value 2.282188310162541e-38
Reject Ho. The sample does not follow normal distribution
```

- For working day, the test result is: p-value 2.282188310162541e-38, the sample does not follow normal distribution.

Even after applying Shapiro-wilk test, still we find out that the distribution of the "workingday" and "holiday/weekend" data, the samples do not follow normal distribution.

- *Variance Check using Levene's test:*



```
[148] # Ho - Varience is Equal. Homogenous Variance
      # Ha - Varience is Not Equal. Non Homogenous Variance
      workingday_sample =df.loc[df['workingday'] == 'working day', 'count'].sample(2000)
      non_workingday_sample = df.loc[df['workingday'] == 'holiday/weekend', 'count'].sample(2000)

      alpha = 0.05
      test_stat, p_value = levene(workingday_sample, non_workingday_sample)
      print('p-value', p_value)
      if p_value < alpha:
          print('reject Ho: The samples do not have  Homogenous Variance')
      else:
          print('Fail to Reject Ho: The samples have Homogenous Variance ')

      p-value 0.8587179252254277
      Fail to Reject Ho: The samples have Homogenous Variance
```

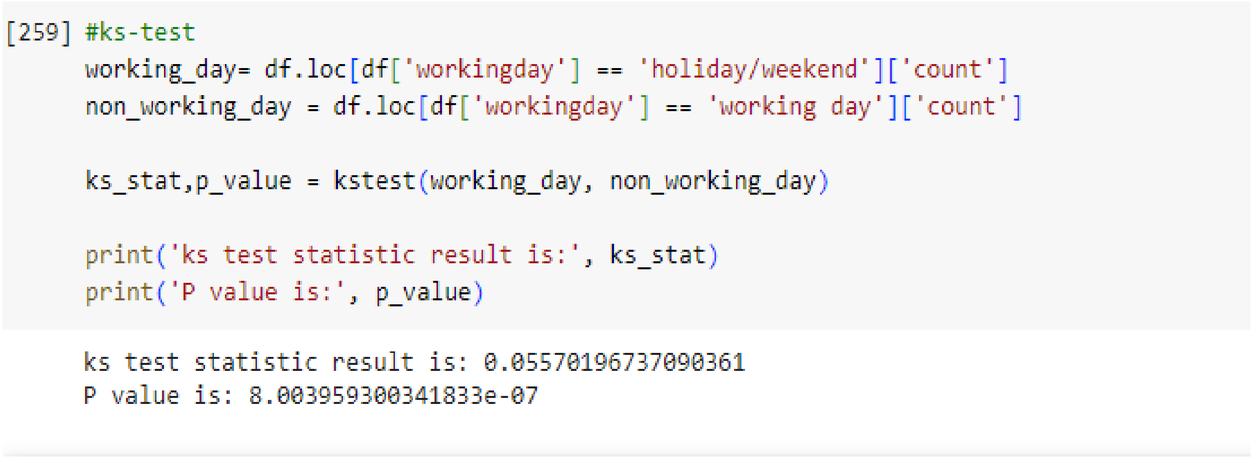
  - The test result is: p-value 0.8587179252254277, Fail to Reject Ho: The samples have Homogenous Variance

➢ **Step 3: Set a significance level (alpha).**

- We set our alpha to be 0.05.

➢ **Step 4: Calculate test statistics.**

- Standard deviation of the population is not known. So, T-test is right choice for checking the statistics.
- But**,** we have seen in previously (using histogram plot, Q-Q plot, Shapiro-wilk test) that the distribution is not normal. And in variance test (using Levene's test), we have seen that the variance is homogeneous.
- Since the samples are not normally distributed. So, T-test is couldn't give us proper statistics result, it probably increase the risk of errors.
- We can perform non-parametric test. i.e; ks- test, ks-test doesn't depend on the distribution.

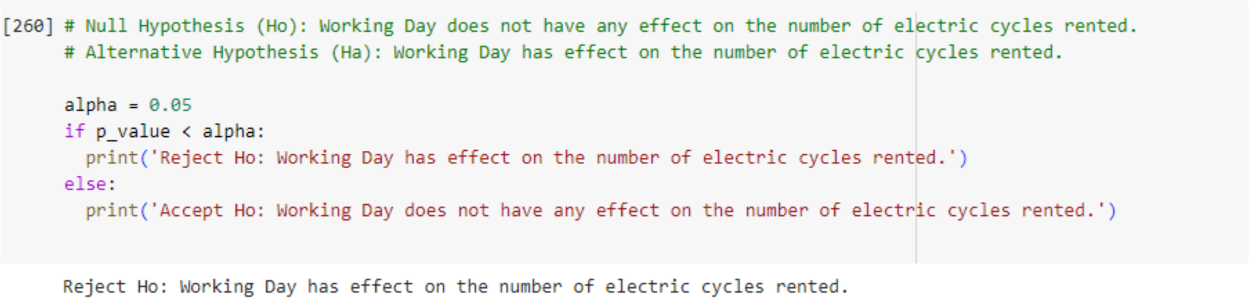


```
[259] #ks-test
      working_day= df.loc[df['workingday'] == 'holiday/weekend']['count']
      non_working_day = df.loc[df['workingday'] == 'working day']['count']

      ks_stat,p_value = kstest(working_day, non_working_day)

      print('ks test statistic result is:', ks_stat)
      print('P value is:', p_value)

      ks test statistic result is: 0.05570196737090361
      P value is: 8.003959300341833e-07
```

- ***ks-Test:***
  - The result of ks-test: ks test statistic result is: 0.05570196737090361, P value is: 8.003959300341833e-07

➢ **Step 5: Decision to accept or reject null hypothesis.**

- Based on P value, we accept the null hypothesis.
  - If P value < significance level (alpha) then reject null hypothesis.
  - If P value > significance level (alpha) then accept null hypothesis.



```
[260] # Null Hypothesis (Ho): Working Day does not have any effect on the number of electric cycles rented.
      # Alternative Hypothesis (Ha): Working Day has effect on the number of electric cycles rented.

      alpha = 0.05
      if p_value < alpha:
        print('Reject Ho: Working Day has effect on the number of electric cycles rented.')
      else:
        print('Accept Ho: Working Day does not have any effect on the number of electric cycles rented.')

      Reject Ho: Working Day has effect on the number of electric cycles rented.
```

➢ **Step 6: Inference from the analysis.**

- Therefore, the count of total rental cycles is statistically different for both working and non-working days.

Question: Is there any effect of Working Day on the number of electric cycles rented?
Answer: Final conclusion from the above analysis that the working day has an effect on total number of rental cycles.

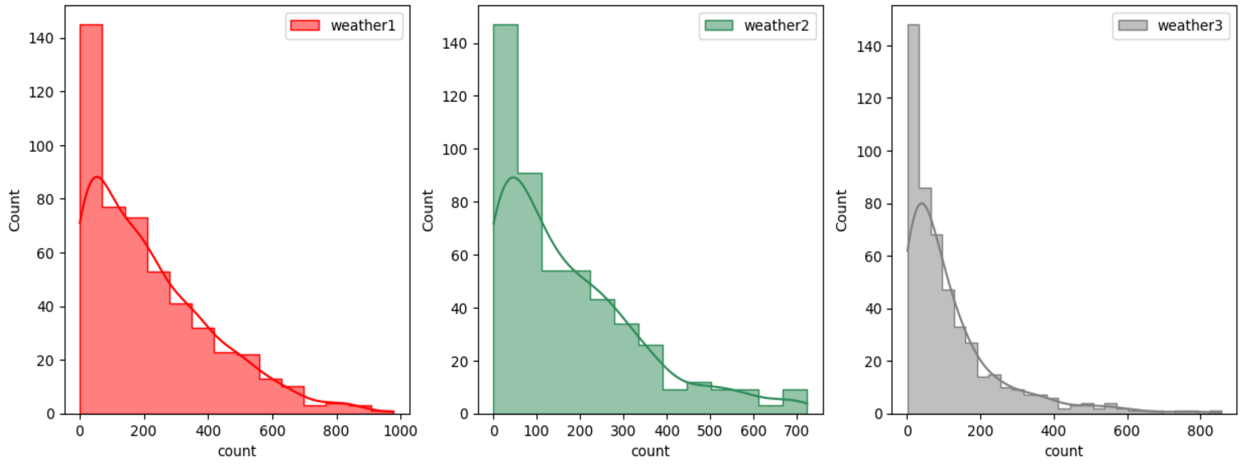
Weather vs. Count

➢ **Step 1: Setup Null Hypothesis**

- Null Hypothesis (Ho): Mean of cycle rented is same for weather 1, 2 and 3. (We are not considering weather 4 as there in only 1 data point for weather 4 and we cannot perform a ANOVA test with a single data point for a group)
- Alternative Hypothesis (Ha): Mean of cycle rented is not same for weather 1, 2, 3 and 4. Or mean of cycle rented is not same for any of two weather.

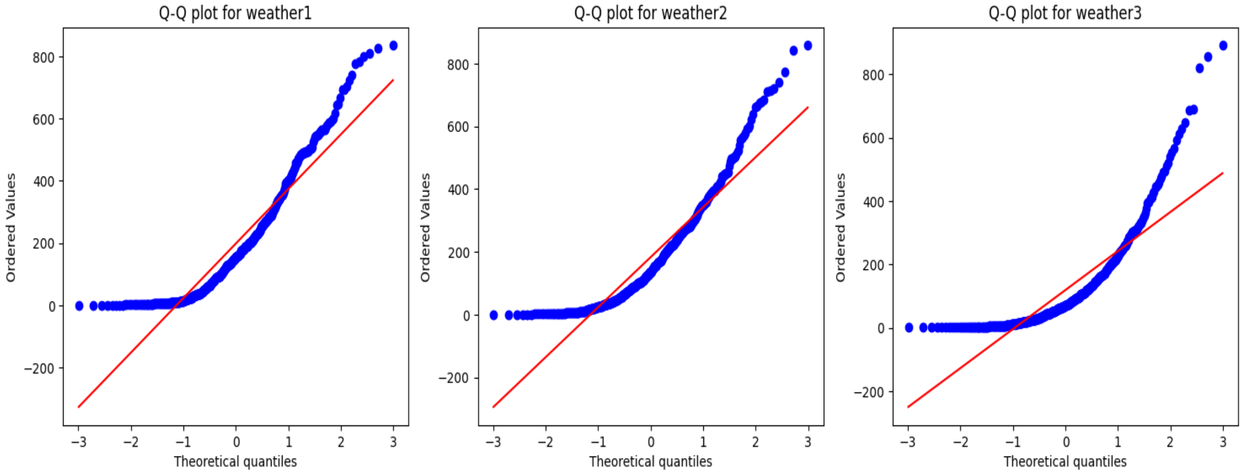
- **Step 2: Checking the assumption for the test.**

Normality check or Distribution check using visual test



- *Normality or Distribution check using histogram or visual test:*
  - We can visualize from the above histogram plot, that the distributions doesn't follow normal distribution.

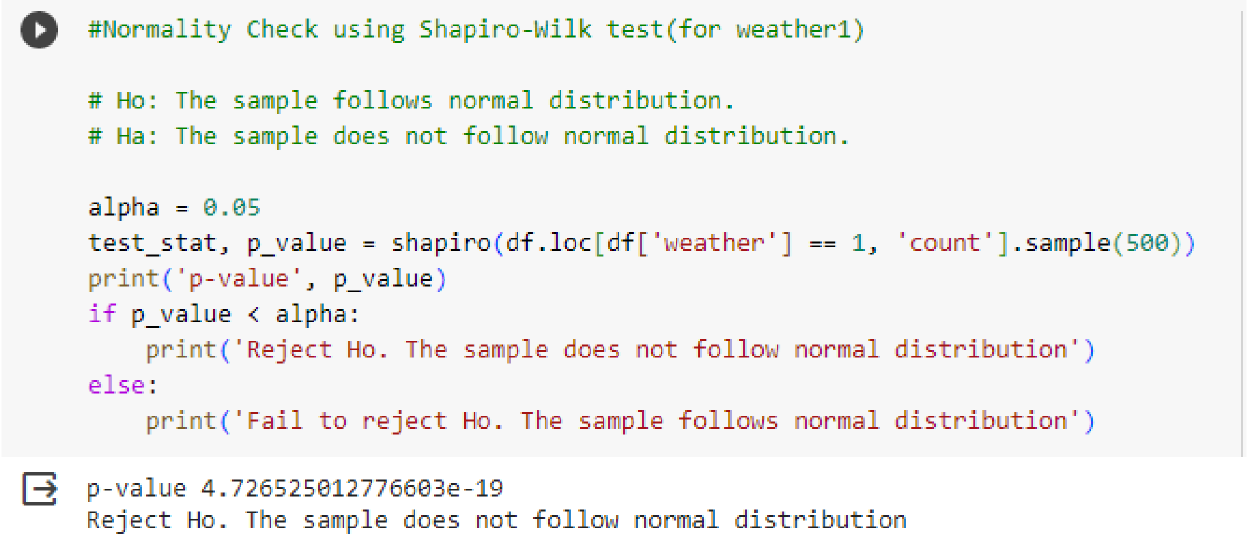
Q-Q plots for the count of electric cycles rented in different weathers



- *Normality or Distribution check using Q-Q plot test:*
  - We can figure it from the above Q-Q plot that the distribution doesn't follow normal distribution.

From the above plots we have seen that the samples do not come from normal distribution. Now we are applying another test Shapiro-wilk test for normality.

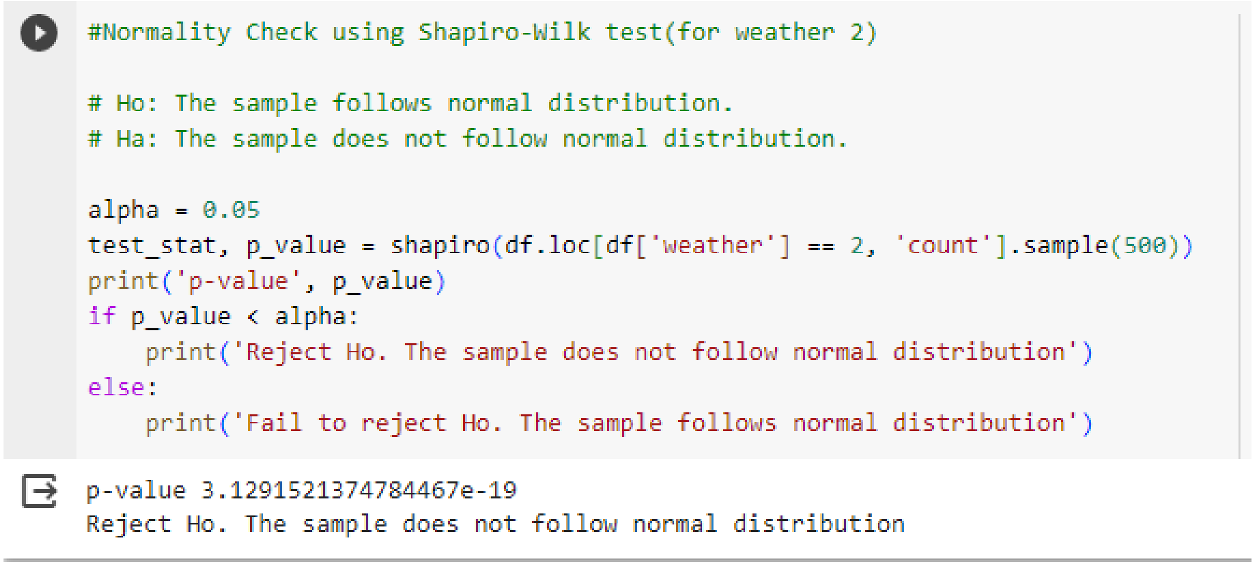
- *Normality check using Shapiro-wilk test:*



```
#Normality Check using Shapiro-Wilk test(for weather1)

# Ho: The sample follows normal distribution.
# Ha: The sample does not follow normal distribution.

alpha = 0.05
test_stat, p_value = shapiro(df.loc[df['weather'] == 1, 'count'].sample(500))
print('p-value', p_value)
if p_value < alpha:
    print('Reject Ho. The sample does not follow normal distribution')
else:
    print('Fail to reject Ho. The sample follows normal distribution')
```

```
p-value 4.726525012776603e-19
Reject Ho. The sample does not follow normal distribution
```

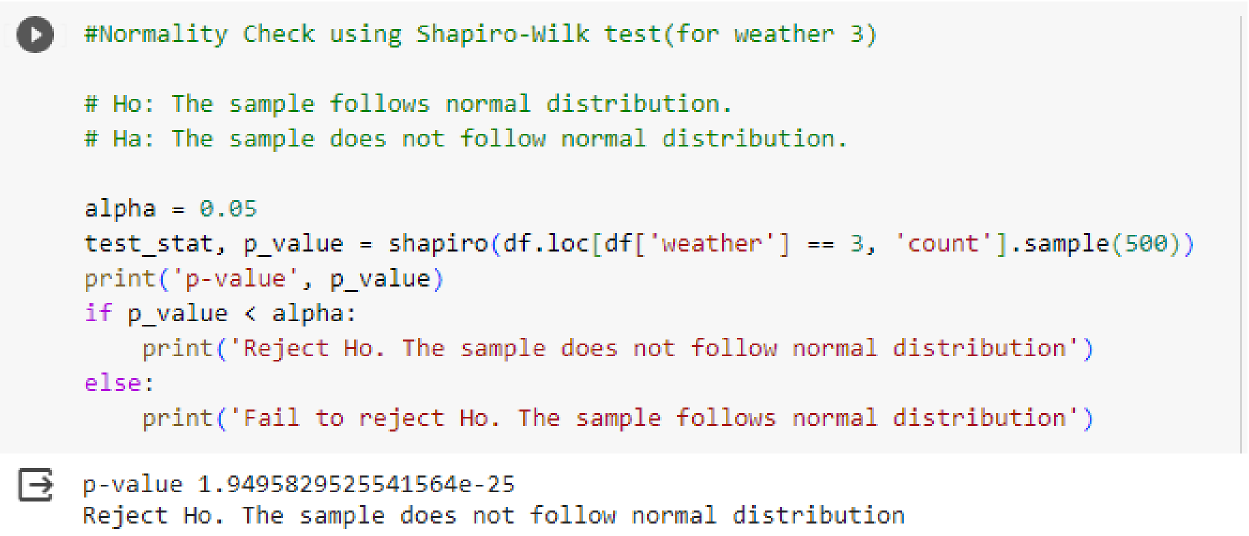
  - For weather 1, the test result is: p-value 4.726525012776603e-19, the sample does not follow normal distribution.



```
#Normality Check using Shapiro-Wilk test(for weather 2)

# Ho: The sample follows normal distribution.
# Ha: The sample does not follow normal distribution.

alpha = 0.05
test_stat, p_value = shapiro(df.loc[df['weather'] == 2, 'count'].sample(500))
print('p-value', p_value)
if p_value < alpha:
    print('Reject Ho. The sample does not follow normal distribution')
else:
    print('Fail to reject Ho. The sample follows normal distribution')
```

```
p-value 3.1291521374784467e-19
Reject Ho. The sample does not follow normal distribution
```

  - For weather 2, the test result is: p-value 3.1291521374784467e-19, the sample does not follow normal distribution.



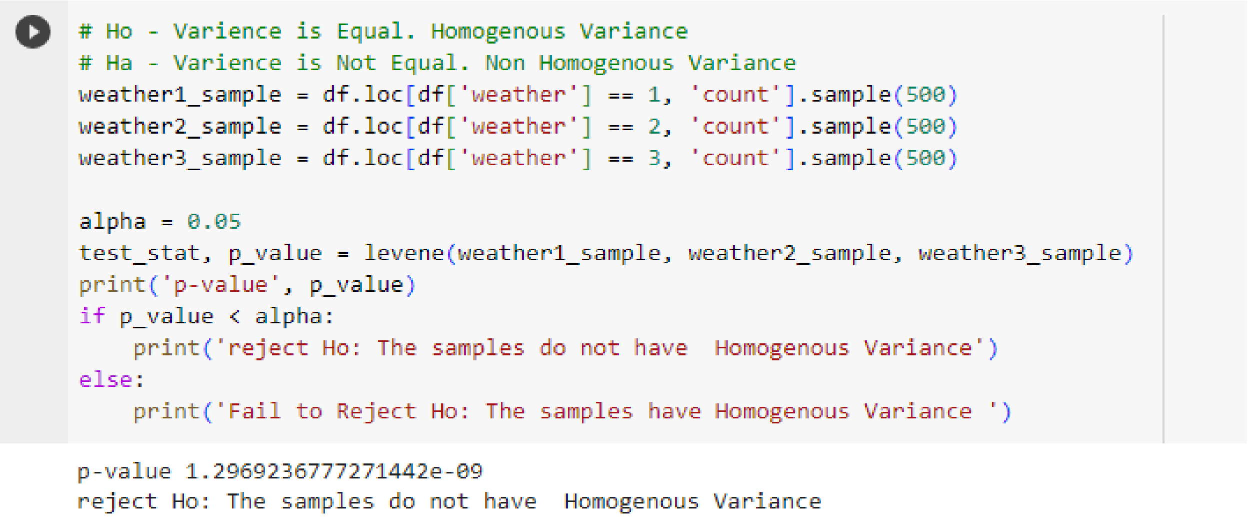
```
#Normality Check using Shapiro-Wilk test(for weather 3)

# Ho: The sample follows normal distribution.
# Ha: The sample does not follow normal distribution.

alpha = 0.05
test_stat, p_value = shapiro(df.loc[df['weather'] == 3, 'count'].sample(500))
print('p-value', p_value)
if p_value < alpha:
    print('Reject Ho. The sample does not follow normal distribution')
else:
    print('Fail to reject Ho. The sample follows normal distribution')
```

```
p-value 1.9495829525541564e-25
Reject Ho. The sample does not follow normal distribution
```

- For weather 3, the test result is: p-value 1.9495829525541564e-25, the sample does not follow normal distribution.

Even after applying Shapiro-wilk test, still we find out that the distribution of the "weather 1", “weather 2” and "weather 3" data, the samples do not follow normal distribution.

- *Variance Check using Levene's test:*



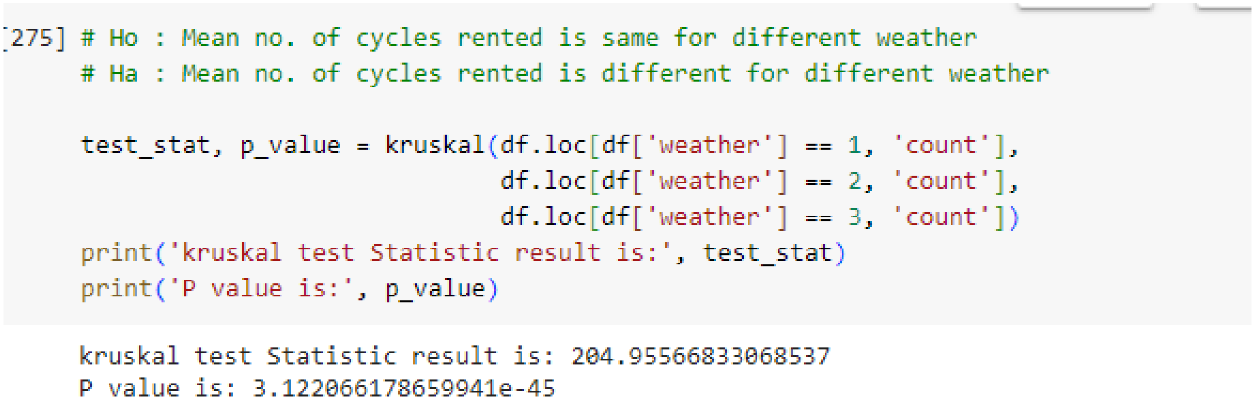
```
# Ho - Varience is Equal. Homogenous Variance
# Ha - Varience is Not Equal. Non Homogenous Variance
weather1_sample = df.loc[df['weather'] == 1, 'count'].sample(500)
weather2_sample = df.loc[df['weather'] == 2, 'count'].sample(500)
weather3_sample = df.loc[df['weather'] == 3, 'count'].sample(500)

alpha = 0.05
test_stat, p_value = levene(weather1_sample, weather2_sample, weather3_sample)
print('p-value', p_value)
if p_value < alpha:
    print('reject Ho: The samples do not have  Homogenous Variance')
else:
    print('Fail to Reject Ho: The samples have Homogenous Variance ')
```

```
p-value 1.2969236777271442e-09
reject Ho: The samples do not have  Homogenous Variance
```

- The test result is: p-value 1.2969236777271442e-09, Reject Ho: The samples do not have Homogenous Variance

- **Step 3: Set a significance level (alpha).**
  - We set our alpha to be 0.05.

- **Step 4: Calculate test statistics.**
  - We have more than 2 categories. So, ANOVA test is right choice.
  - But, we have seen in previously (using histogram plot, Q-Q plot, Shapiro-wilk test) that the distribution is not normal. And in variance test (using Levene's test), we have seen that also variance is not equal.
  - Even the samples are not normally distributed and also they do not have same variance. So, ANOVA-test is couldn't give us proper statistics result, it probably increase the risk of errors.
  - We need to perform non-parametric equivalent test i.e; kruskal-wallis H independent test.

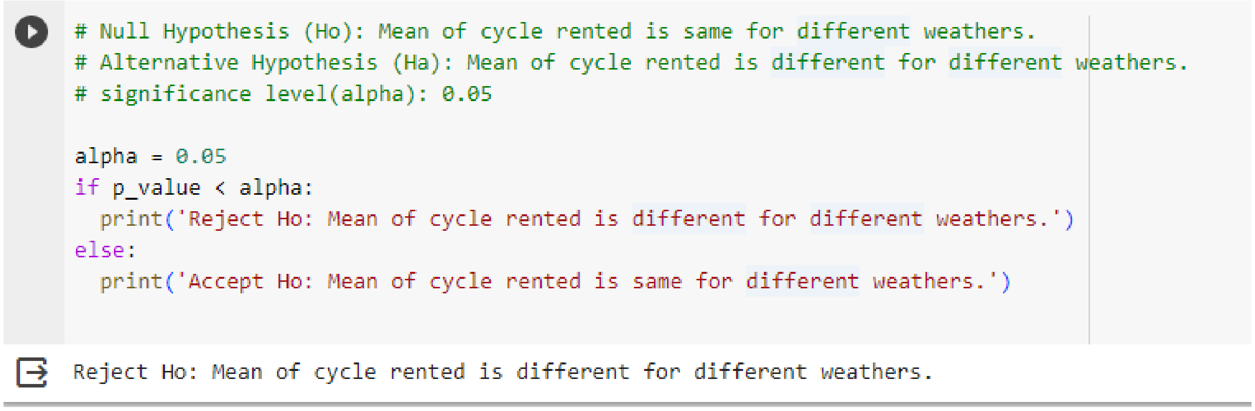


```
[275] # Ho : Mean no. of cycles rented is same for different weather
      # Ha : Mean no. of cycles rented is different for different weather

      test_stat, p_value = kruskal(df.loc[df['weather'] == 1, 'count'],
                                   df.loc[df['weather'] == 2, 'count'],
                                   df.loc[df['weather'] == 3, 'count'])
      print('kruskal test Statistic result is:', test_stat)
      print('P value is:', p_value)
```

```
kruskal test Statistic result is: 204.95566833068537
P value is: 3.122066178659941e-45
```

  - ***Kruskal-wallis Test:***
    - The result of kruskal-test: statistic result is: 204.95566833068537, P value is: 3.122066178659941e-45.

- **Step 5: Decision to accept or reject null hypothesis.**
  - Based on P value, we accept the null hypothesis.
    - If P value < significance level (alpha) then reject null hypothesis.
    - If P value > significance level (alpha) then accept null hypothesis.



```
# Null Hypothesis (Ho): Mean of cycle rented is same for different weathers.
# Alternative Hypothesis (Ha): Mean of cycle rented is different for different weathers.
# significance level(alpha): 0.05

alpha = 0.05
if p_value < alpha:
  print('Reject Ho: Mean of cycle rented is different for different weathers.')
else:
  print('Accept Ho: Mean of cycle rented is same for different weathers.')
```

```
Reject Ho: Mean of cycle rented is different for different weathers.
```

- **Step 6: Inference from the analysis.**
  - Therefore, the average number of rental cycles is statistically different for different weathers.

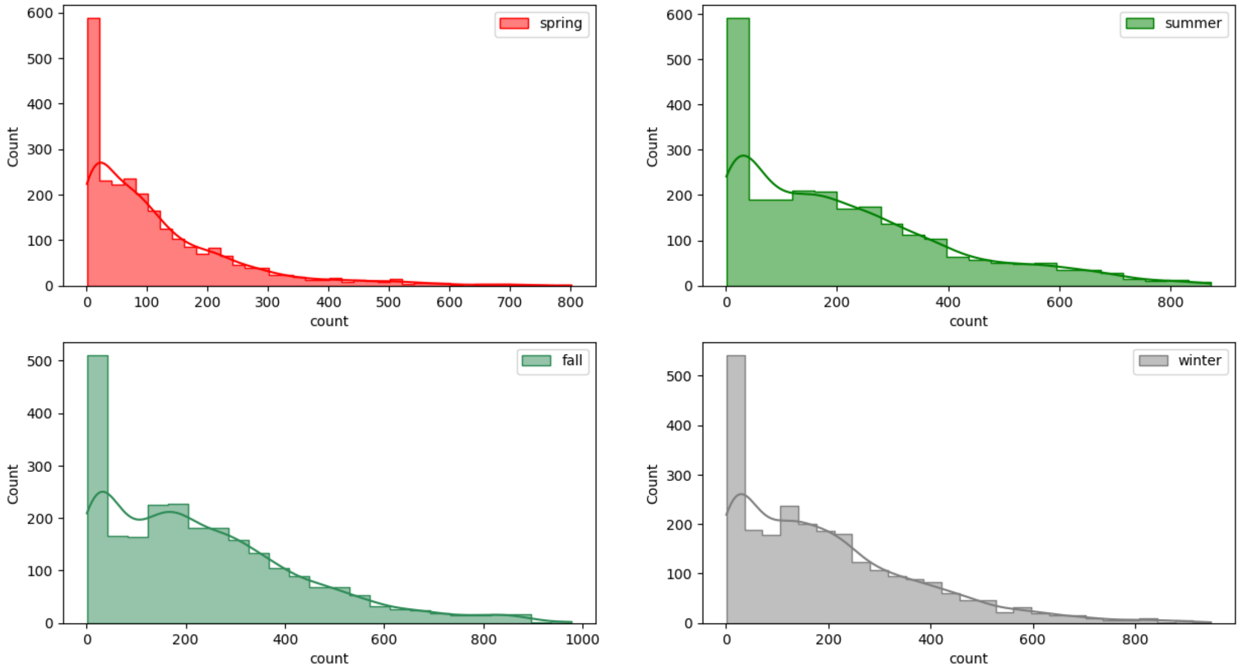
Question: Is the number of cycles rented is similar or different in different weather?
Answer: Final conclusion from the above analysis that the number of cycles is different in different weathers.

Season vs. Count

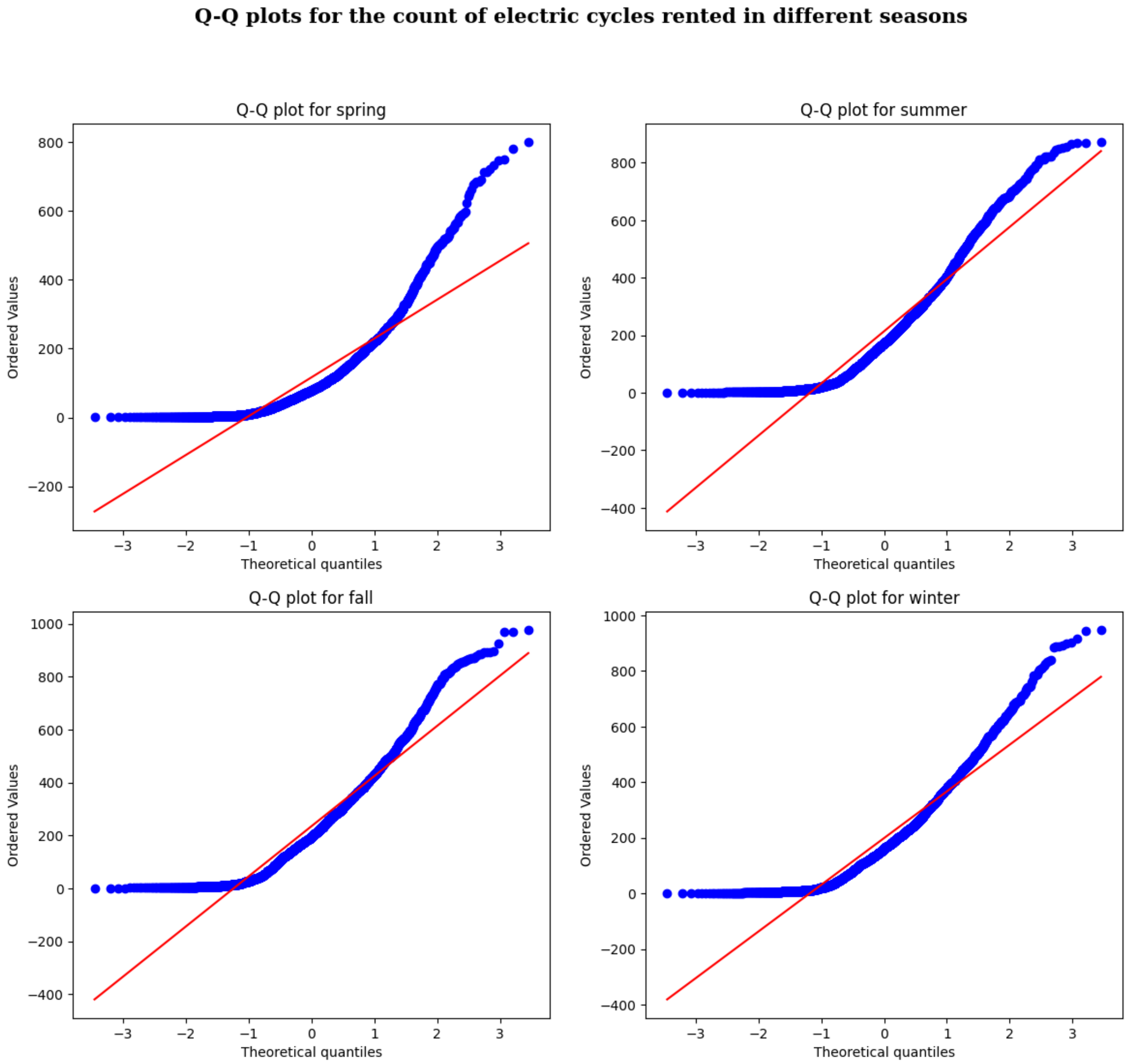
- **Step 1: Setup Null Hypothesis**
  - Null Hypothesis (Ho): Mean of cycle rented is same for all different seasons (spring, summer, fall, winter)
  - Alternative Hypothesis (Ha): Mean of cycle rented is not same for spring, summer, fall and winter seasons. Or mean of cycle rented is different for any of these two seasons.

- **Step 2: Checking the assumption for the test.**

Normality check or Distribution check using visual test



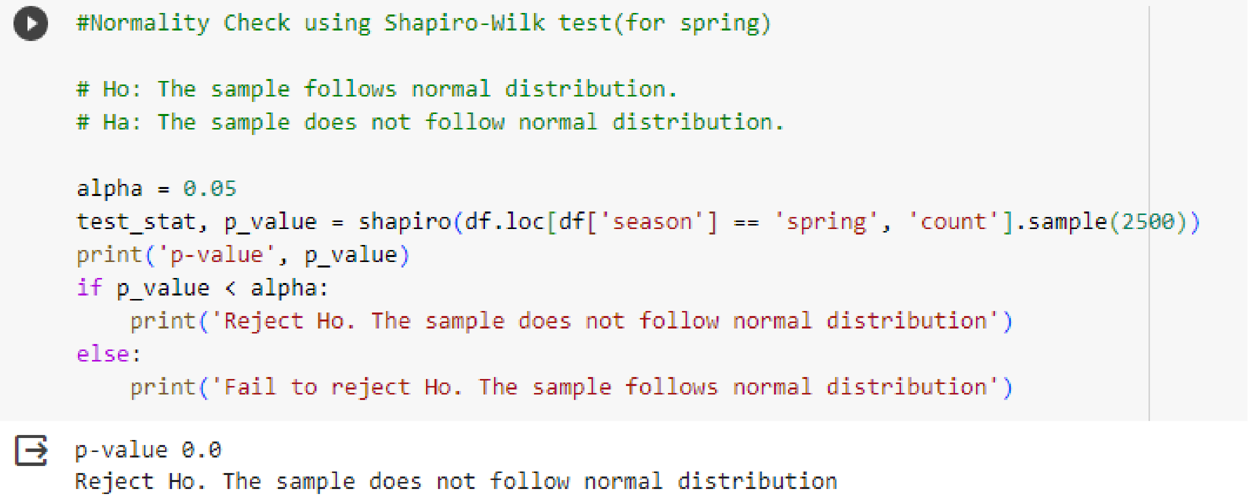
- *Normality or Distribution check using histogram or visual test:*
    - We can visualize from the above histogram plot, that the distributions doesn't follow normal distribution.



- *Normality or Distribution check using Q-Q plot test:*
    - We can figure it from the above Q-Q plot that the distribution doesn't follow normal distribution.

From the above plots we have seen that the samples do not come from normal distribution. Now we are applying another test Shapiro-wilk test for normality.

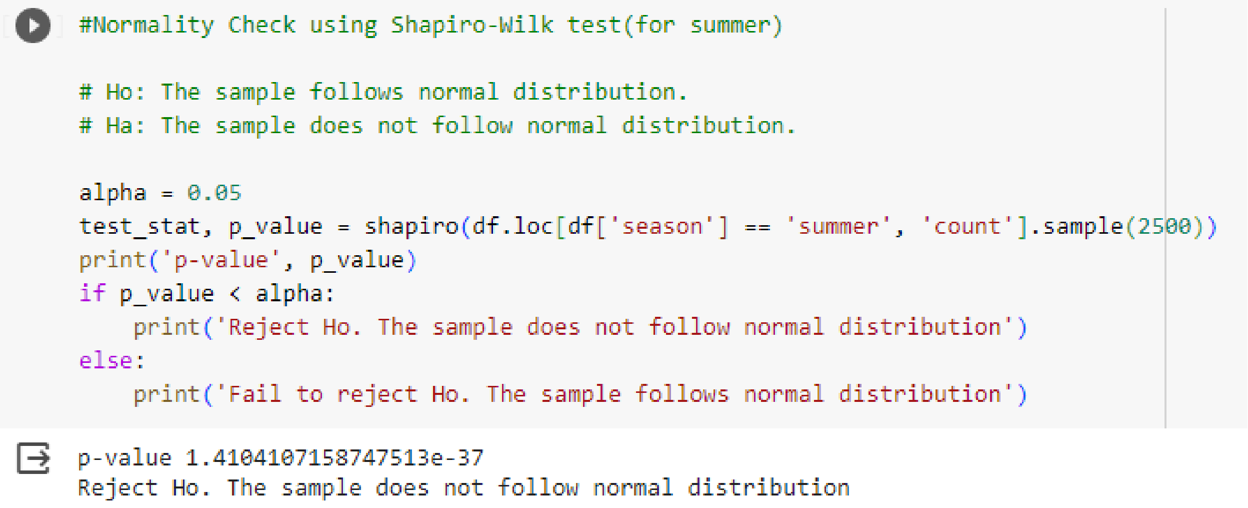
- *Normality check using Shapiro-wilk test:*



```
#Normality Check using Shapiro-Wilk test(for spring)

# Ho: The sample follows normal distribution.
# Ha: The sample does not follow normal distribution.

alpha = 0.05
test_stat, p_value = shapiro(df.loc[df['season'] == 'spring', 'count'].sample(2500))
print('p-value', p_value)
if p_value < alpha:
    print('Reject Ho. The sample does not follow normal distribution')
else:
    print('Fail to reject Ho. The sample follows normal distribution')
```

```
p-value 0.0
Reject Ho. The sample does not follow normal distribution
```

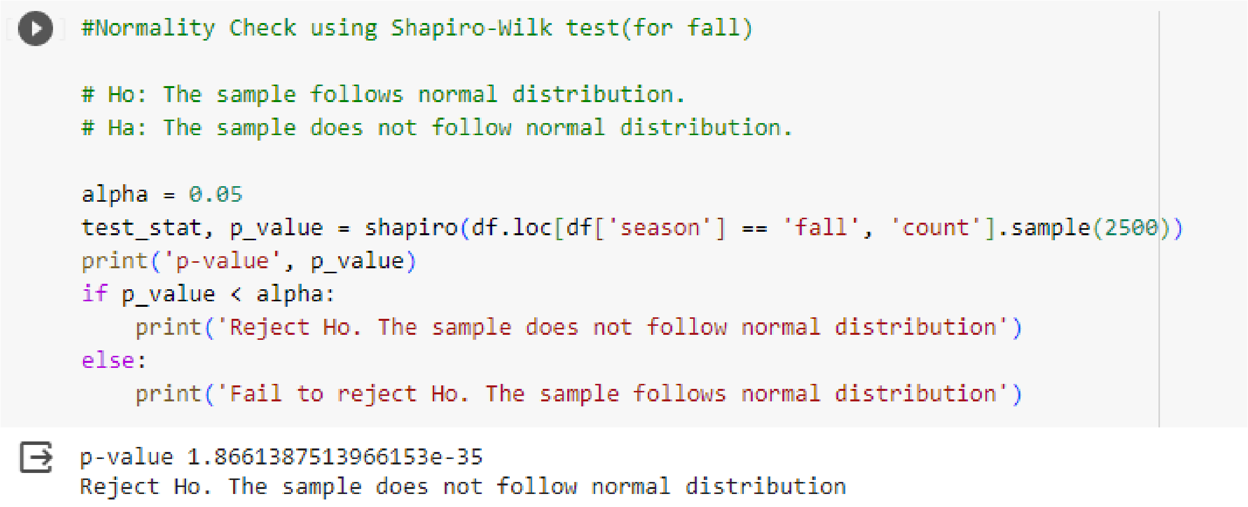
  - For spring season, the test result is: p-value 0.0, the sample does not follow normal distribution.



```
#Normality Check using Shapiro-Wilk test(for summer)

# Ho: The sample follows normal distribution.
# Ha: The sample does not follow normal distribution.

alpha = 0.05
test_stat, p_value = shapiro(df.loc[df['season'] == 'summer', 'count'].sample(2500))
print('p-value', p_value)
if p_value < alpha:
    print('Reject Ho. The sample does not follow normal distribution')
else:
    print('Fail to reject Ho. The sample follows normal distribution')
```

```
p-value 1.4104107158747513e-37
Reject Ho. The sample does not follow normal distribution
```

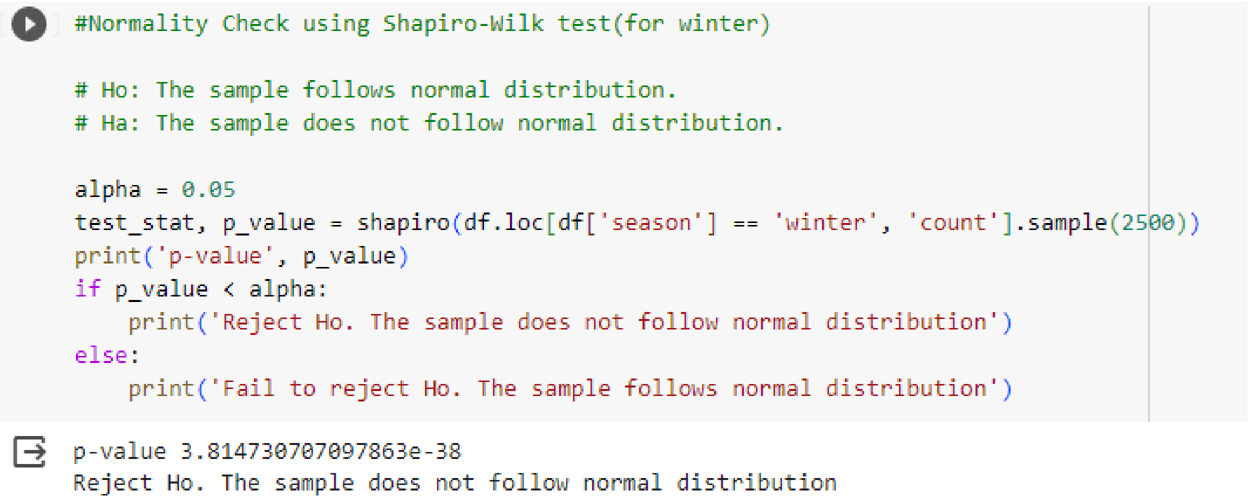
  - For summer season, the test result is: p-value 1.4104107158747513e-37, the sample does not follow normal distribution.



```
#Normality Check using Shapiro-Wilk test(for fall)

# Ho: The sample follows normal distribution.
# Ha: The sample does not follow normal distribution.

alpha = 0.05
test_stat, p_value = shapiro(df.loc[df['season'] == 'fall', 'count'].sample(2500))
print('p-value', p_value)
if p_value < alpha:
    print('Reject Ho. The sample does not follow normal distribution')
else:
    print('Fail to reject Ho. The sample follows normal distribution')
```

```
p-value 1.8661387513966153e-35
Reject Ho. The sample does not follow normal distribution
```

- For fall season, the test result is: p-value 1.8661387513966153e-35, the sample does not follow normal distribution.



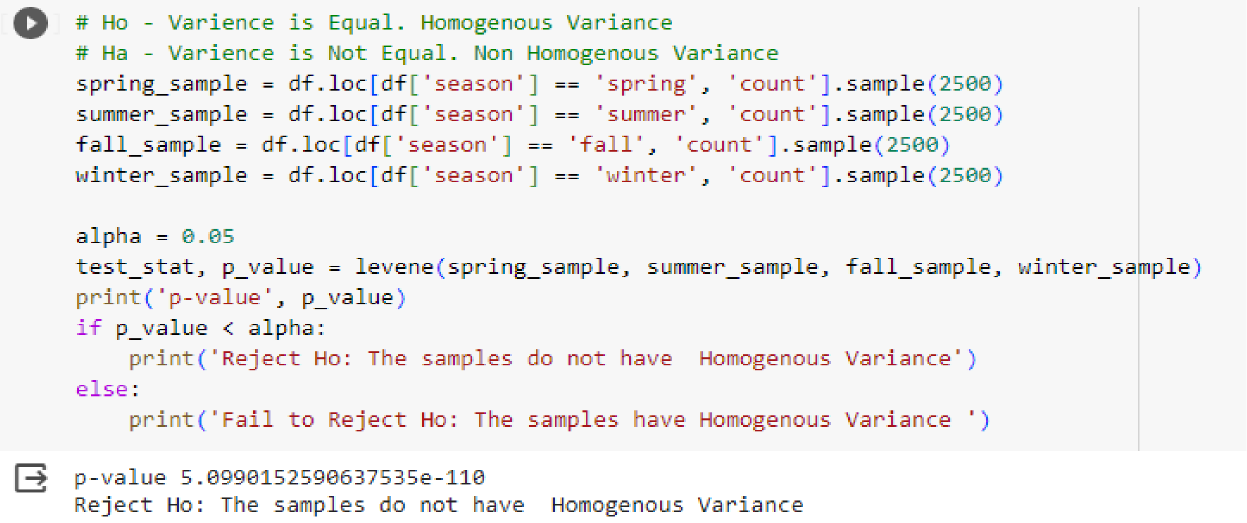
```
#Normality Check using Shapiro-Wilk test(for winter)

# Ho: The sample follows normal distribution.
# Ha: The sample does not follow normal distribution.

alpha = 0.05
test_stat, p_value = shapiro(df.loc[df['season'] == 'winter', 'count'].sample(2500))
print('p-value', p_value)
if p_value < alpha:
    print('Reject Ho. The sample does not follow normal distribution')
else:
    print('Fail to reject Ho. The sample follows normal distribution')
```

```
p-value 3.814730707097863e-38
Reject Ho. The sample does not follow normal distribution
```

- For winter season, the test result is: p-value 3.814730707097863e-38, the sample does not follow normal distribution.

Even after applying Shapiro-wilk test, still we find out that the distribution of the "spring", “summer”, “fall” and "winter" season data, the samples do not follow normal distribution.

- *Variance Check using Levene's test:*



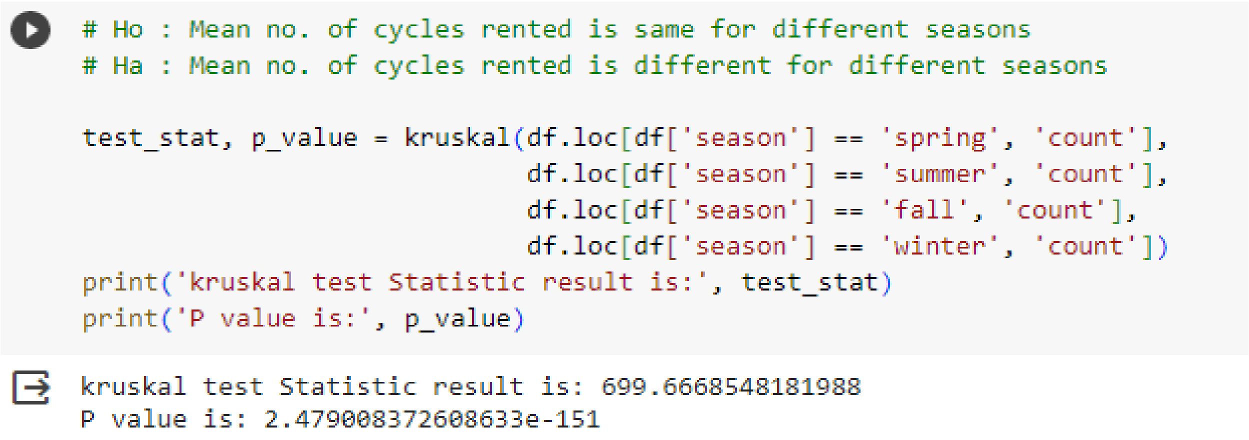
```
# Ho - Varience is Equal. Homogenous Variance
# Ha - Varience is Not Equal. Non Homogenous Variance
spring_sample = df.loc[df['season'] == 'spring', 'count'].sample(2500)
summer_sample = df.loc[df['season'] == 'summer', 'count'].sample(2500)
fall_sample = df.loc[df['season'] == 'fall', 'count'].sample(2500)
winter_sample = df.loc[df['season'] == 'winter', 'count'].sample(2500)

alpha = 0.05
test_stat, p_value = levene(spring_sample, summer_sample, fall_sample, winter_sample)
print('p-value', p_value)
if p_value < alpha:
    print('Reject Ho: The samples do not have  Homogenous Variance')
else:
    print('Fail to Reject Ho: The samples have Homogenous Variance ')
```

```
p-value 5.0990152590637535e-110
Reject Ho: The samples do not have  Homogenous Variance
```

  - The test result is: p-value 5.09901525900637535e-110, Reject Ho: The samples do not have Homogenous Variance

- **Step 3: Set a significance level (alpha).**
  - We set our alpha to be 0.05.

- **Step 4: Calculate test statistics.**
  - We have more than 2 categories. So, ANOVA test is right choice.
  - But**,** we have seen in previously (using histogram plot, Q-Q plot, Shapiro-wilk test) that the distribution is not normal. And in variance test (using Levene's test), we have seen that also variance is not equal.
  - Even the samples are not normally distributed and also they do not have same variance. So, ANOVA-test is couldn't give us proper statistics result, it probably increase the risk of errors.
  - We need to perform non-parametric equivalent test i.e; kruskal-wallis H independent test.

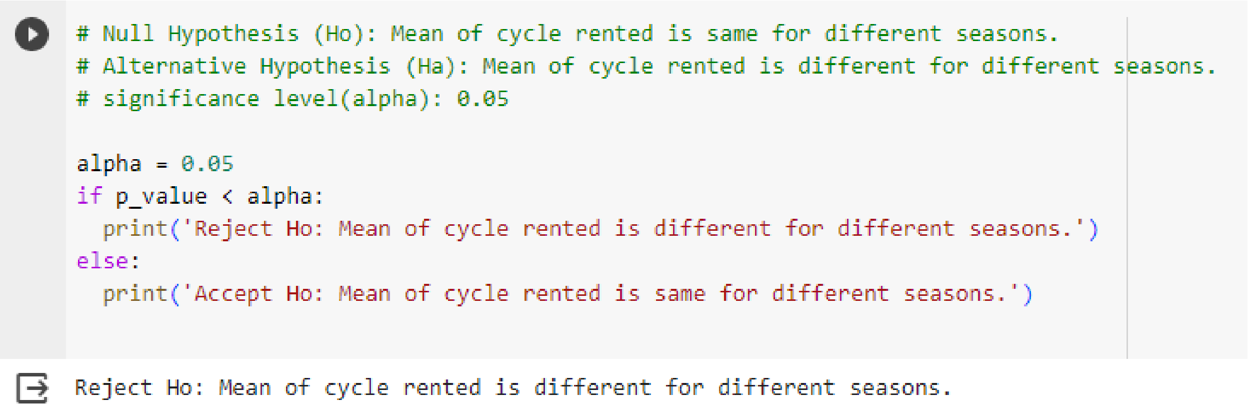


```
# Ho : Mean no. of cycles rented is same for different seasons
# Ha : Mean no. of cycles rented is different for different seasons

test_stat, p_value = kruskal(df.loc[df['season'] == 'spring', 'count'],
                             df.loc[df['season'] == 'summer', 'count'],
                             df.loc[df['season'] == 'fall', 'count'],
                             df.loc[df['season'] == 'winter', 'count'])
print('kruskal test Statistic result is:', test_stat)
print('P value is:', p_value)
```

```
kruskal test Statistic result is: 699.6668548181988
P value is: 2.479008372608633e-151
```

- ***Kruskal-wallis Test:***
    - The result of kruskal-test: statistic result is: 699.6668548181988, P value is: 2.479008372608633e-151.

- **Step 5: Decision to accept or reject null hypothesis.**
    - Based on P value, we accept the null hypothesis.
        - If P value < significance level (alpha) then reject null hypothesis.
        - If P value > significance level (alpha) then accept null hypothesis.



```
# Null Hypothesis (Ho): Mean of cycle rented is same for different seasons.
# Alternative Hypothesis (Ha): Mean of cycle rented is different for different seasons.
# significance level(alpha): 0.05

alpha = 0.05
if p_value < alpha:
  print('Reject Ho: Mean of cycle rented is different for different seasons.')
else:
  print('Accept Ho: Mean of cycle rented is same for different seasons.')
```

```
Reject Ho: Mean of cycle rented is different for different seasons.
```

- **Step 6: Inference from the analysis.**
    - Therefore, the average number of rental cycles is statistically different for different seasons.

Question: Is the number of cycles rented is similar or different in different seasons?
Answer: Final conclusion from the above analysis that the number of cycles is different in different seasons.

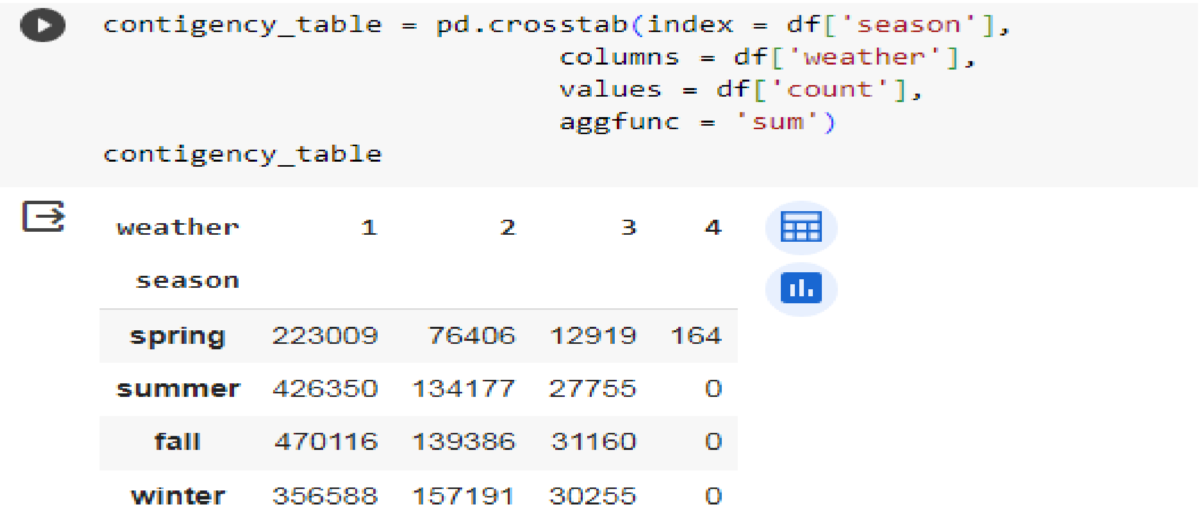
Weather vs. Season

- **Step 1: Setup Null Hypothesis**
    - Null Hypothesis (Ho): Weather is not dependent on season.
    - Alternative Hypothesis (Ha): Weather is dependent on season.

- **Step 2: Define the test statistics.**
  - ‘weather’ and ‘season’ both are categorical in nature. So, Chi-square test is applicable here.
  - The Chi-square test of independence checks whether two variables are likely to be related or not.

- **Step 3: Set a significance level (alpha).**
  - We set our alpha to be 0.05.

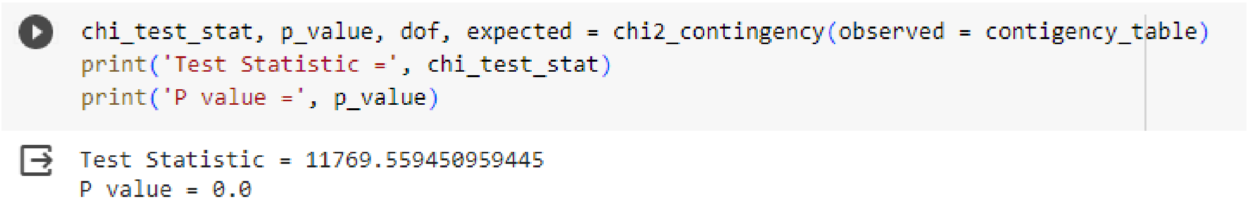
- **Step 4: Calculate test statistics.**



```
contigency_table = pd.crosstab(index = df['season'],
                               columns = df['weather'],
                               values = df['count'],
                               aggfunc = 'sum')

contigency_table
```

weather	1	2	3	4
season				
spring	223009	76406	12919	164
summer	426350	134177	27755	0
fall	470116	139386	31160	0
winter	356588	157191	30255	0

  - First, we create the contingency table such that each value is the total number of rented cycles for a particular weather and season.

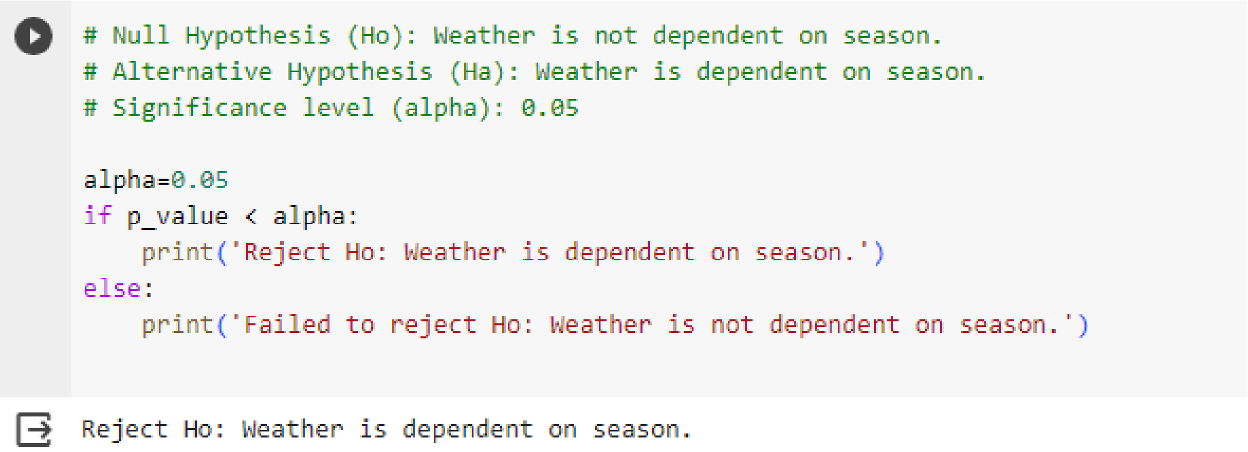


```
chi_test_stat, p_value, dof, expected = chi2_contingency(observed = contigency_table)
print('Test Statistic =', chi_test_stat)
print('P value =', p_value)
```

```
Test Statistic = 11769.559450959445
P value = 0.0
```

  - ***Chi-square Test:***
    - The test result of chi-square is: test statistics is 11769.559450959445 and P value is 0.0

- **Step 5: Decision to accept or reject null hypothesis.**
  - Based on P value, we accept the null hypothesis.
    - If P value < significance level (alpha) then reject null hypothesis.
    - If P value > significance level (alpha) then accept null hypothesis.



```
# Null Hypothesis (Ho): Weather is not dependent on season.
# Alternative Hypothesis (Ha): Weather is dependent on season.
# Significance level (alpha): 0.05

alpha=0.05
if p_value < alpha:
    print('Reject Ho: Weather is dependent on season.')
else:
    print('Failed to reject Ho: Weather is not dependent on season.')
```

```
Reject Ho: Weather is dependent on season.
```

- **Step 6: Inference from the analysis.**
  - Therefore, there is statistically dependency of weather and season based on the number of cycles rented.

Question: Is weather dependent on the season?
Answer: Final conclusion from the above analysis that weather is dependent on the season.

Colab notebook link:
https://colab.research.google.com/drive/1K6dLptT8c7nOMnsGIbtcg93fMS9o-wZJ?usp=sharing