**Colab NoteBook Link**  
**https://colab.research.google.com/drive/1pz07HFOBszo8XTB3n-IbG-qg7uLobWfo?usp=sharing**  
  
**PDF Notebook Link  
  
https://drive.google.com/file/d/1BcdKNxItlBL7s64NNo40tpGGOLMCy\_i0/view?usp=sharing**

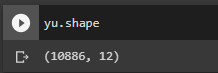
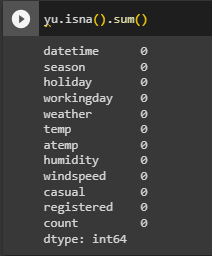
1. **Define Problem Statement and perform Exploratory Data Analysis (10 points)**
   * **Definition of problem (as per given problem statement with additional views)**
   * **Observations on shape of data, data types of all the attributes, conversion of categorical attributes to 'category' (If required) , 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**

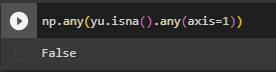
We basically have to identify the factors on which the demand for the shared electric cycles depends in the Indian market.

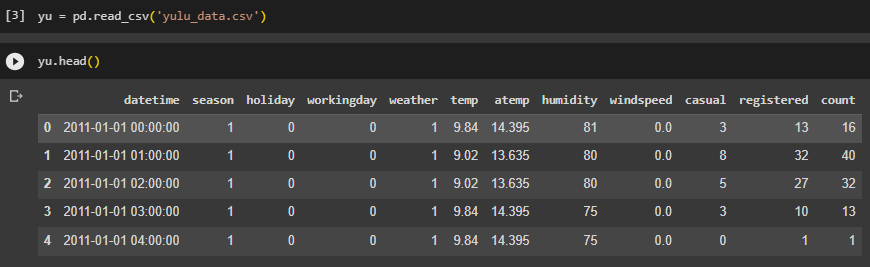
The company wants to know:

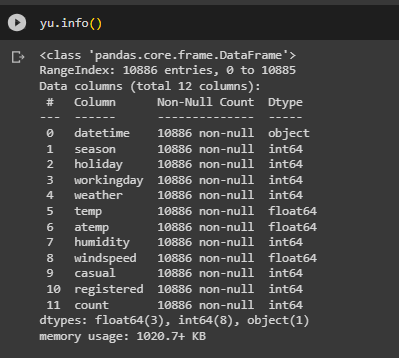
* Which variables are significant in predicting the demand for shared electric cycles in the Indian market?  
    
  To find this we will do hypothesis testing on the dataset available to us, and will try to reach on a conclusion with 5% significance interval. We will also do comparison among the fields available in the data set to see if some of them dependent or not like the season and weather.  
    
  Some of the tests can be whether the **customer count of electric cycles depends on season, weather, particular hour of the day, working/non-working day, holiday/non-holiday etc.**

**Basic Metrics**:

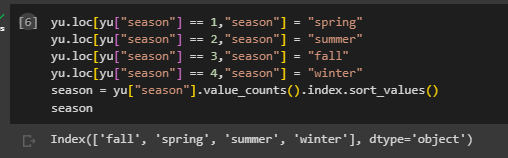
* There are 10886 records of customer booking in the dataset.
* There are 12 columns in the dataset.
* The Shape of the data is (550068, 10).  
  
* There are no-null values present in any column **datetime, season, holiday, workingday, weather, temp, atemp, humidity, windspeed, casual, registered, count.**  
  **yu.isna().sum()  
    
    
    
  np.any(yu.isna().any(axis=1))**

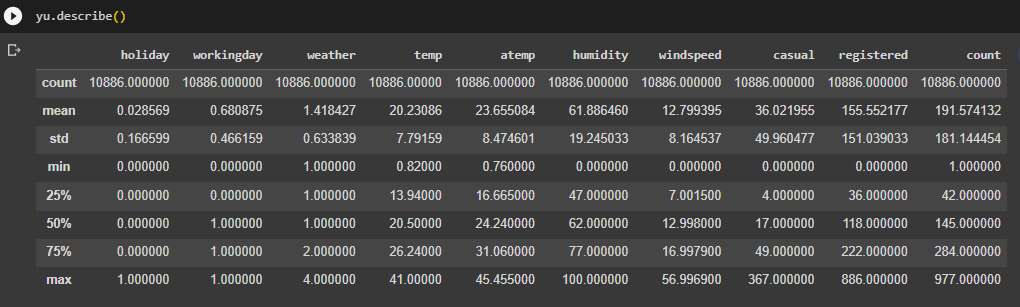
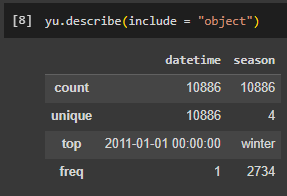


* Top 5 rows in the data set to have a glimpse of data.    
    
  
* Column**datetime is** **object**type rest all other columns are**Integer types/ Float types.**



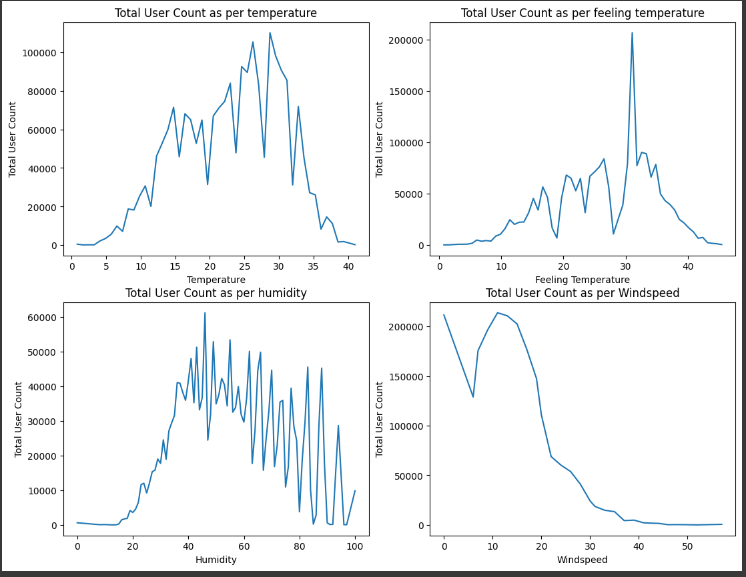
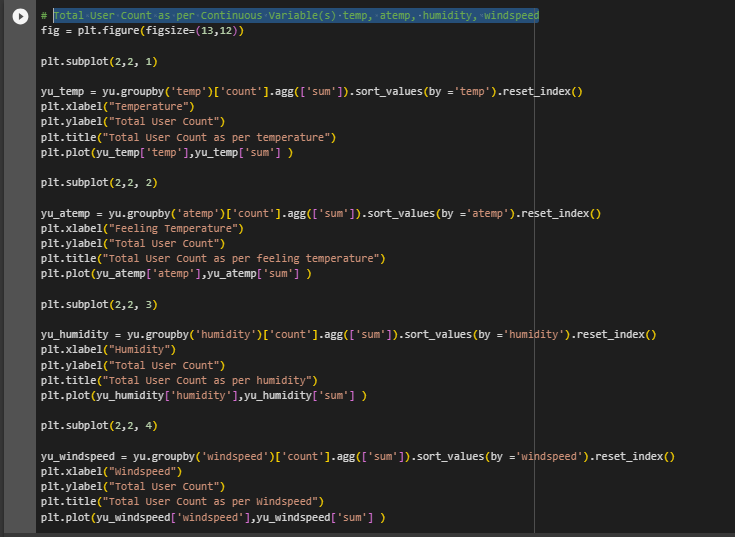
* Converted season from int64 to object type (string) for better view in graphs.



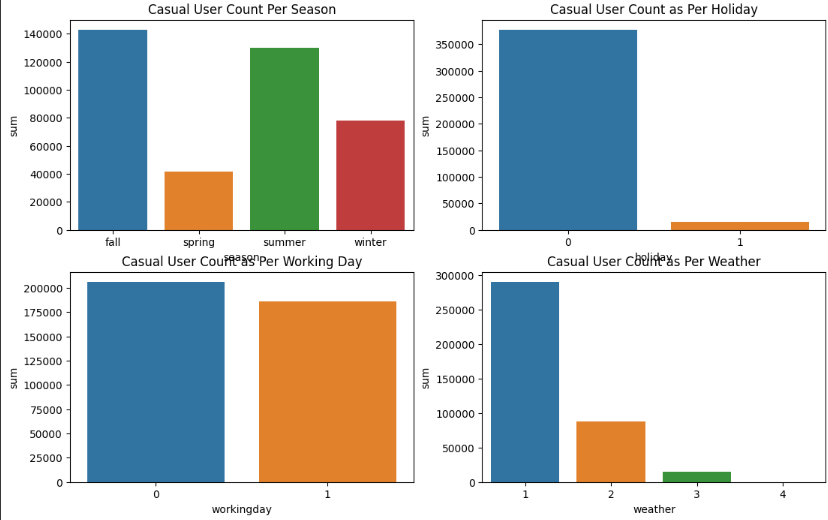
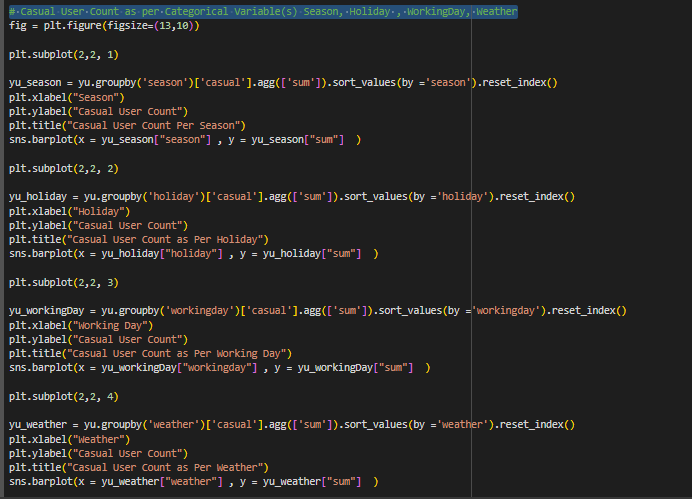
* The statistical summary can be found using **describe**method mentioned below.  
    
    
    
  

**Visual Analysis – Univariate, Bivariate, Outliers after pre-processing of the data**

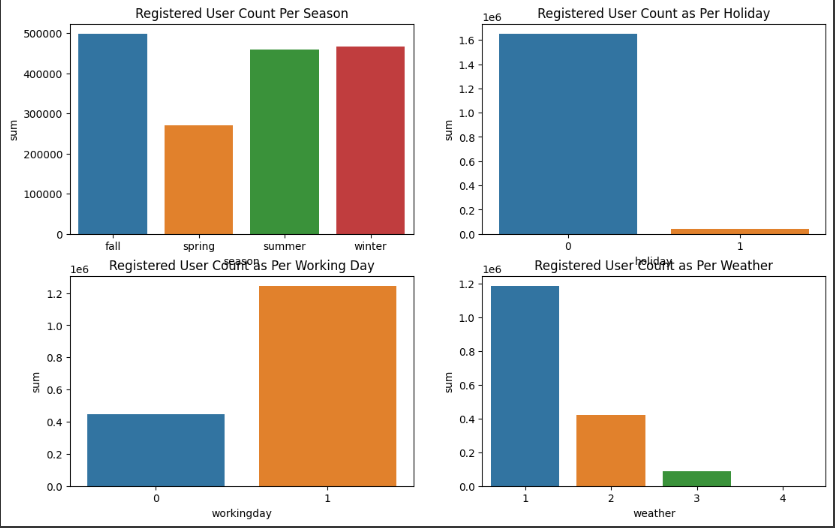
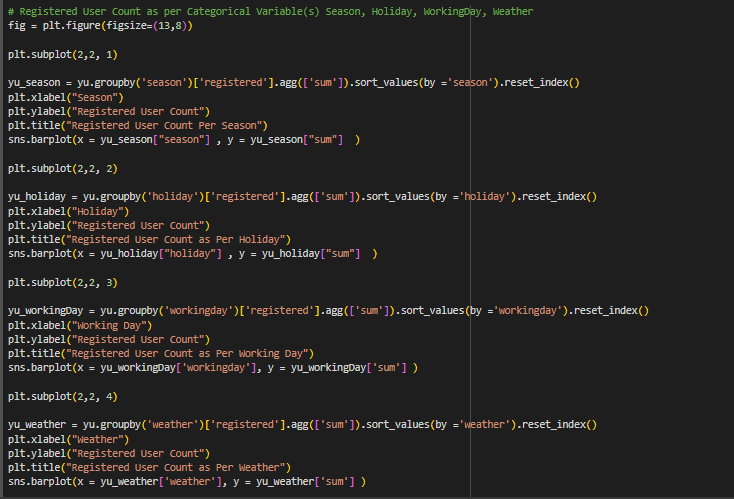
* **Total User Count** as **per Continuous Variable(s) temp, atemp, humidity, windspeed.**There is peak of customers when the feeling temperature is around 30 -32 degree Celsius.  
  Also, customers are more when the windspeed is in between 10-20.

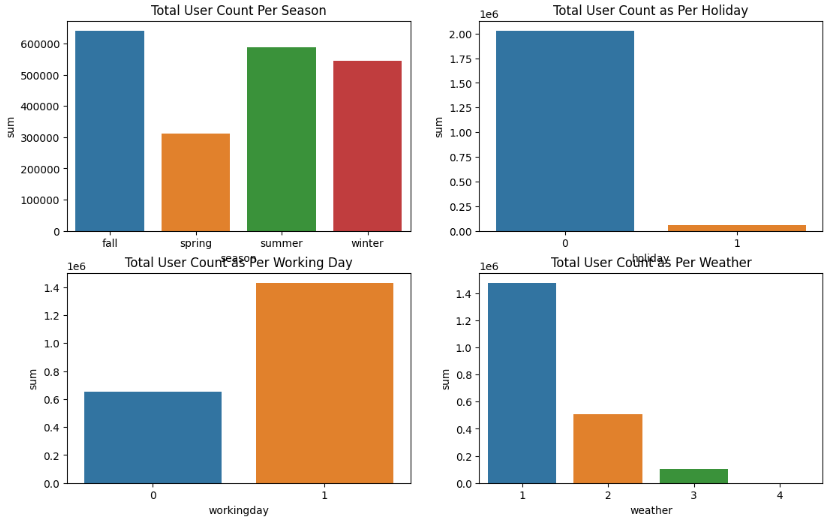
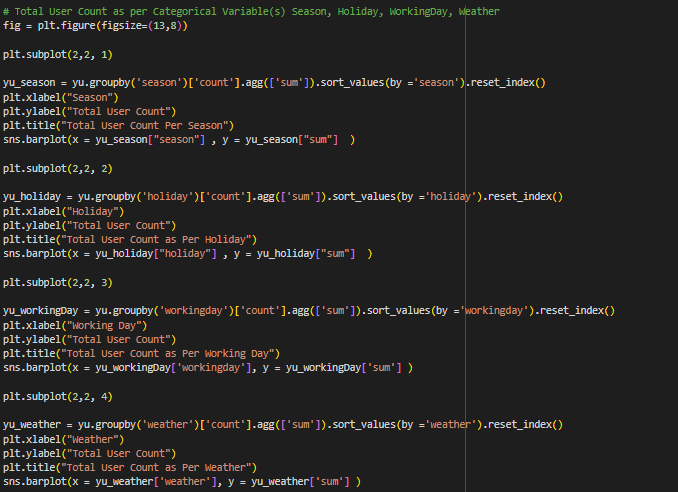
* **Casual User Count as per Categorical Variable(s) Season, Holiday, WorkingDay, Weather**

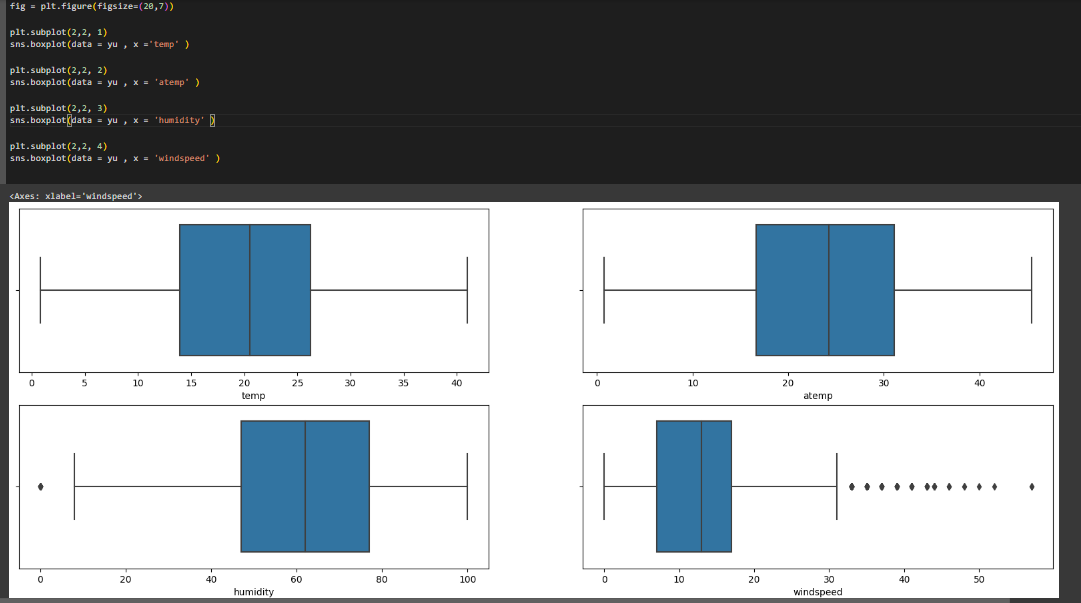
1. Casual Customers are more in the season of fall and summer followed by winter and spring.
2. Casual customers are relatively more when it’s not a holiday.
3. Casual customers are relatively more on a non-working day than a working day
4. Casual Customers are more in Clear, Few clouds, partly cloudy, partly cloudy weather.  
     
     
     
     
   

* **Registered User Count as per Categorical Variable(s) Season, Holiday, WorkingDay, Weather**

1. Registered Customers are mostly same in the season of fall, summer, winter followed by spring.
2. Registered customers are relatively more when it’s not a holiday.
3. Registered customers are relatively more on a working day than a non-working day.
4. Registered Customers are more in Clear, Few clouds, partly cloudy, partly cloudy weather.  
     
     
     
   

* **Total User Count as per Categorical Variable(s) Season, Holiday, WorkingDay, Weather**
  1. Total Customers are mostly same in the season of fall, summer, winter followed by spring.
  2. Total customers are relatively more when it’s not a holiday.
  3. Total customers are relatively more on a working day than a non-working day.
  4. Total Customers are more in Clear, Few clouds, partly cloudy, partly cloudy weather.

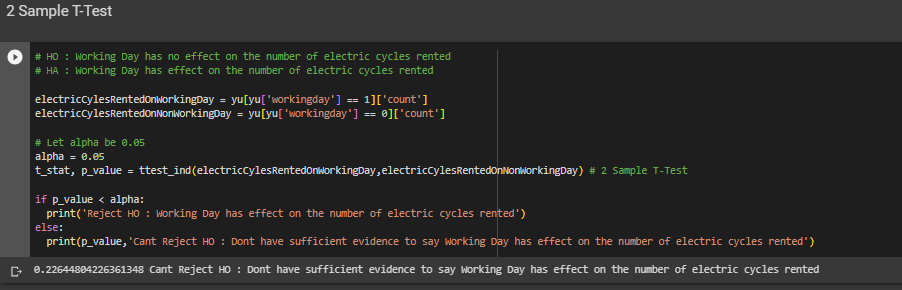
  
  
****

* **Outliers Detection.  
    
  **

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

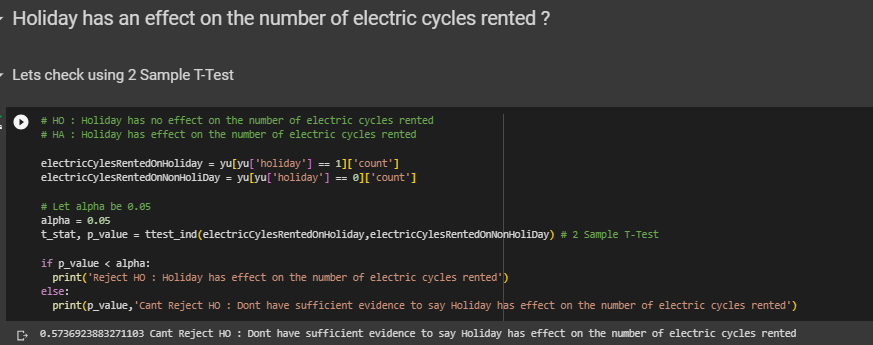
2.1 **Working Day has an effect on the number of electric cycles rented ?**

Cant Reject HO : Don’t have sufficient evidence to say Working Day has effect on the number of electric cycles rented



**p\_value = 0.22644804226361348**

2.2 **Holiday has an effect on the number of electric cycles rented ?**

****

**p\_value = 0.5736923883271103, Cant Reject HO** : Dont have sufficient evidence to say Holiday has effect on the number of electric cycles rented

2.3 **No. of cycles rented similar or different in different seasons?**

Reject HO: No. of cycles rented different in different seasons.

**p\_value : 6.164843386499654e-149**

# HO : No. of cycles rented similar in different seasons

# HA : No. of cycles rented different in different seasons

# Let alpha be 0.05

alpha = 0.05

electricCylesRentedInSpringSeason = yu[yu['season'] == 'spring']['count']

electricCylesRentedInSummerSeason = yu[yu['season'] == 'summer']['count']

electricCylesRentedInFallSeason = yu[yu['season'] == 'fall']['count']

electricCylesRentedInWinterSeason = yu[yu['season'] == 'winter']['count']

# Performed ANNOVA Test

t\_stat, p\_value = f\_oneway(electricCylesRentedInSpringSeason,electricCylesRentedInSummerSeason,electricCylesRentedInFallSeason, electricCylesRentedInWinterSeason )

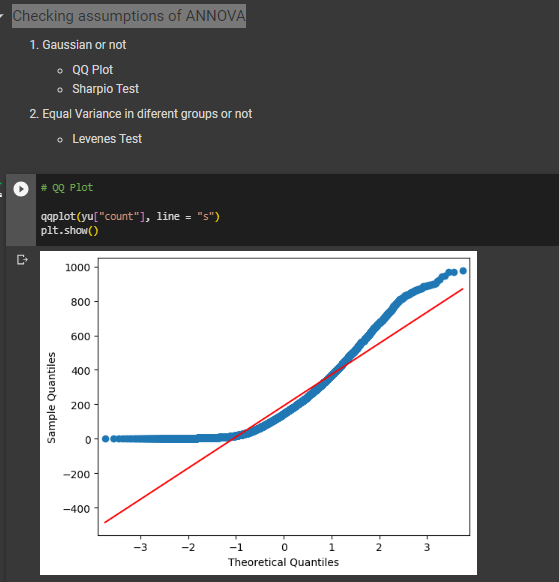
if p\_value < alpha:

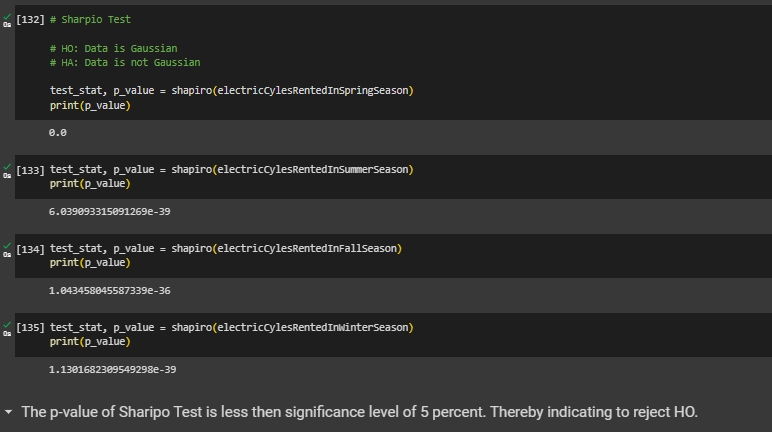
  print('Reject HO : No. of cycles rented different in different seasons')

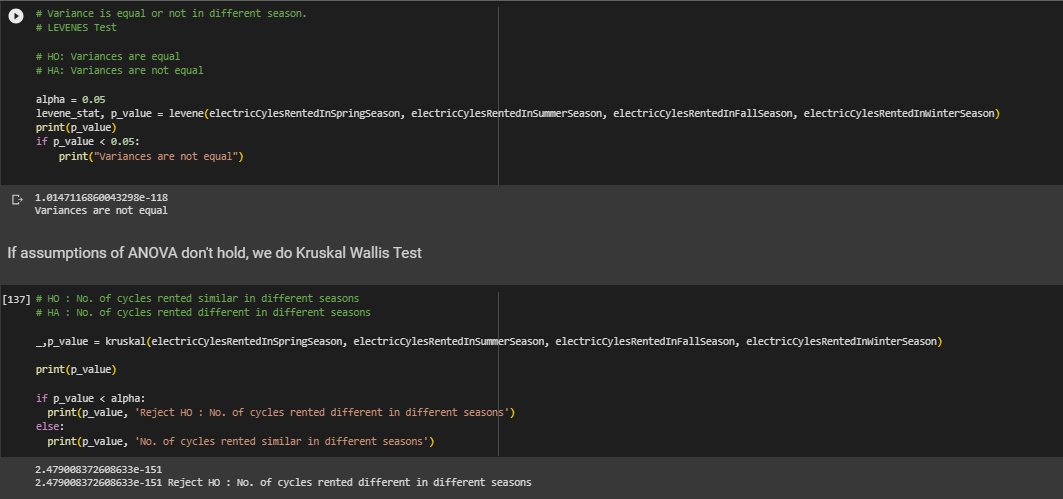
else:

  print('No. of cycles rented similar in different seasons')

**Checking assumptions of ANNOVA**

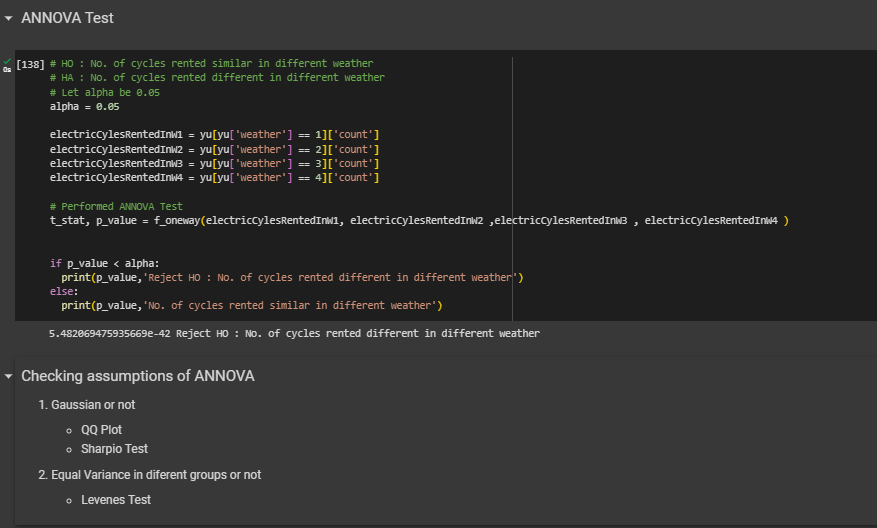
****

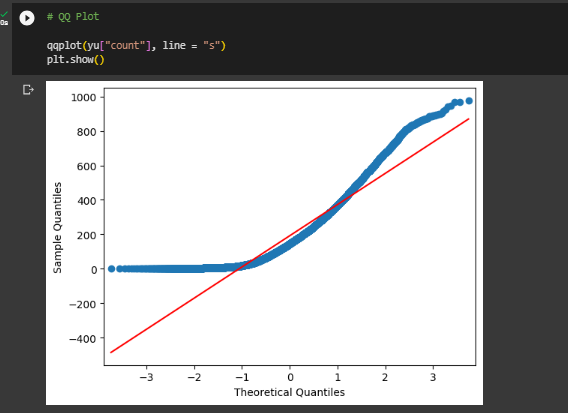
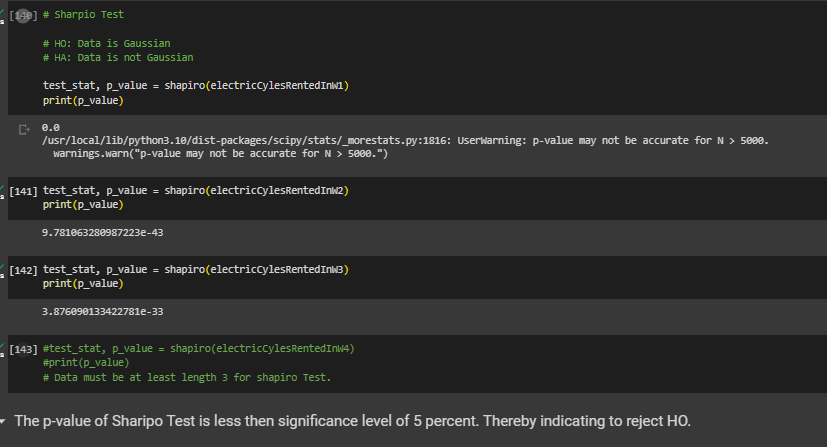
****

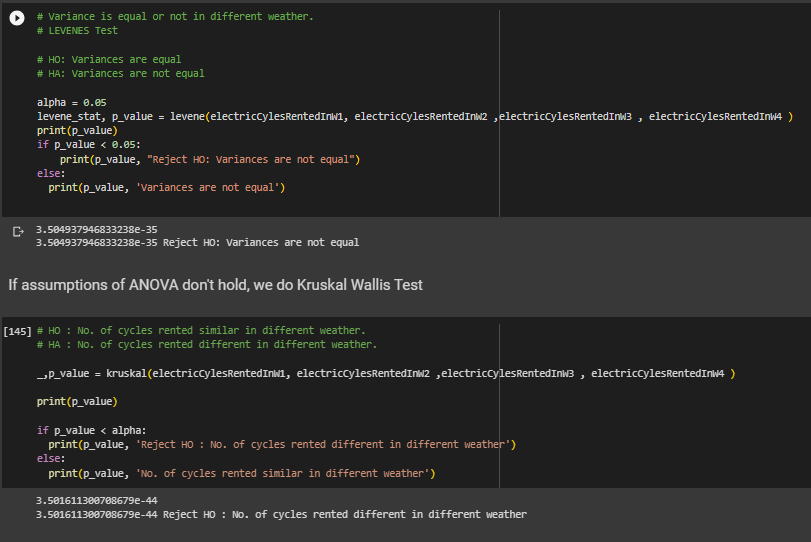
****

2.4 **No. of cycles rented similar or different in different weather ?**

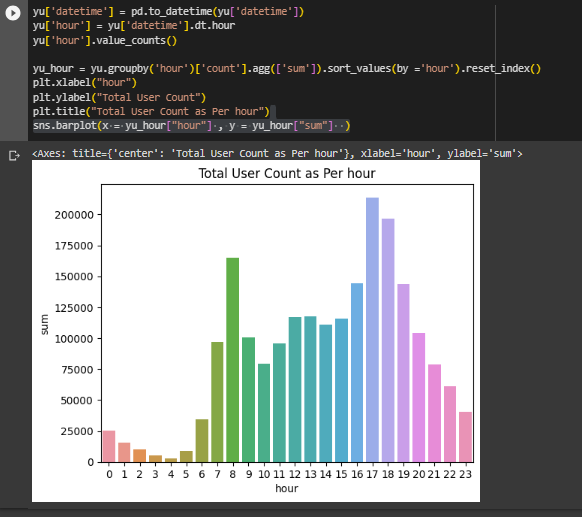
**p\_value = 5.482069475935669e-42, Reject HO : No. of cycles rented different in different weather**

****

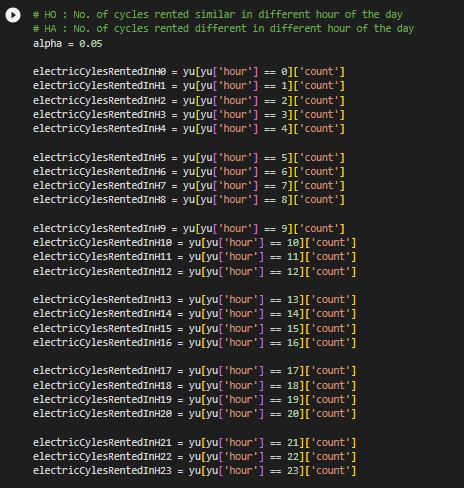
****

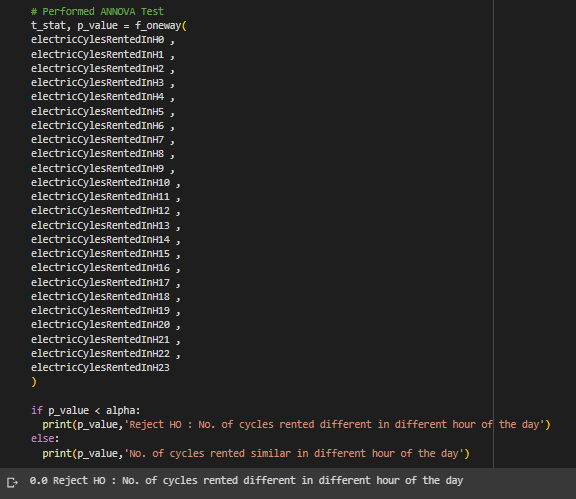


2.5 **No. of cycles rented similar or different in different hour of the day ?  
  
p\_value = 0.0, Reject HO : No. of cycles rented different in different hour of the day**

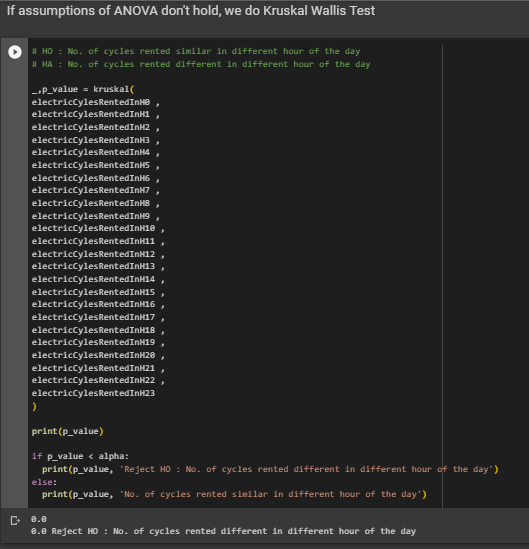


**Check using ANNOVA Test**

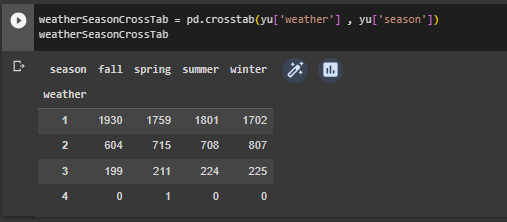
****

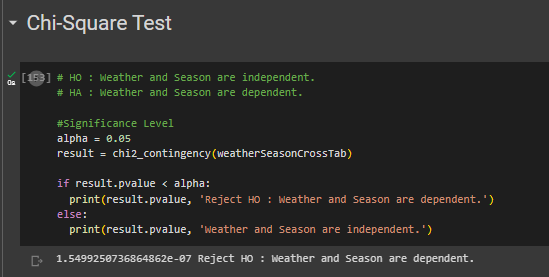
****

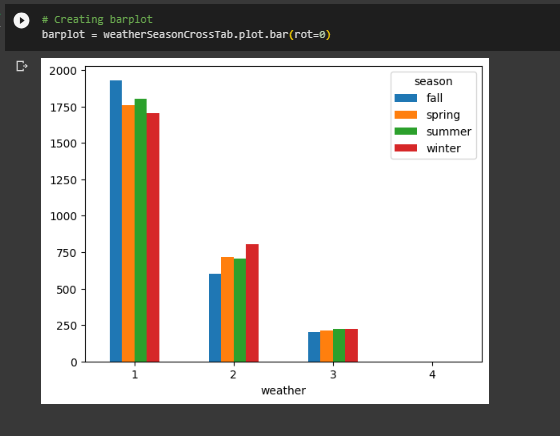
****

2.6 **Weather is dependent on season ?**







**Conclusion**

* With 5% Significance Interval No. of cycles rented depends on **Weather**.
* With 5% Significance Interval No. of cycles rented depends on **Season**.
* With 5% Significance Interval No. of cycles rented depends on **Hour of the day** it is rented.
* With 5% Significance Interval **Season and Weather are dependent**.