**A PROJECT REPORT**

**ON**

**BIKE RENTAL COUNT PREDICTION USING LINEAR REGRESSION**

Submitted in partial fulfillment for the requirement of the award of

TRAINING

IN

Data Analytics, Machine Learning and AI using Python



*Submitted By*

**Aman Kumar(Dayananda Sagar College Of Engineering,Bangalore)**

*Under the guidance of*

**Mr. Ashish Saini**

**ABSTRACT**

The provided details pertain to modern bike sharing systems, which represent an advancement over traditional bike rentals by automating the entire process from membership registration to bike rental and return. These systems enable users to effortlessly rent bikes from one location and conveniently return them at another. Globally, there are more than 500 bike-sharing programs, collectively featuring over 500,000 bicycles. These systems are currently of significant interest due to their significant contributions to traffic management, environmental concerns, and public health.

Beyond their practical applications in the real world, the data generated by these bike sharing systems possesses characteristics that make them appealing for research purposes. In contrast to other transportation services like buses or subways, these systems capture explicit details such as travel duration and the starting and ending positions. This distinctive aspect transforms bike sharing systems into a virtual network of sensors capable of tracking city mobility. Consequently, it is anticipated that analyzing this data could lead to the detection of various noteworthy events occurring within the city. So we are using linear regression model to count bike rental and accuracy is also tested with the actual data i.e. with reference data and henceforth bike rental count is obtained and predicted on a day basis.

**INTRODUCTION**

The dataset comprises various attributes that provide information about bike-sharing patterns. The "instant" attribute represents a sequential index, while "dteday" records the date. Seasons are categorized into four types: spring, summer, fall, and winter, denoted by values 1 to 4 in the "season" attribute. The "yr" attribute indicates the year, with 0 representing 2011 and 1 representing 2012. Months are represented by numbers 1 to 12 in the "mnth" attribute. The occurrence of holidays is captured by the "holiday" attribute (1 for holidays, 0 for non-holidays), and "weekday" indicates the day of the week. The "workingday" attribute takes the value 1 for workdays and 0 for weekends or holidays.

"Weathersit" describes weather conditions:

* 1: Clear, few clouds, partly cloudy
* 2: Misty, cloudy, broken clouds
* 3: Light snow, light rain, thunderstorm, scattered clouds
* 4: Heavy rain, ice pellets, thunderstorm, mist, snow, fog

The "temp" attribute provides the normalized temperature in Celsius, calculated using a specific formula. Similarly, "atemp" offers the normalized feeling temperature. "Hum" represents normalized humidity divided by 100, and "windspeed" presents normalized wind speed divided by 67.

The dataset also includes user-related counts:

* "casual": Count of casual users
* "registered": Count of registered users
* "cnt": Count of total rental bikes, encompassing both casual and registered users.

**Technology and Concepts**

**Machine Learining**

Learning algorithms are widely used in computer vision applications. Before considering tasks, we are going to have a brief look at basics of machine learning.

Machine learning has emerged as a useful tool for modelling problems that are otherwise difficult to formulate exactly. Classical computer programs are explicitly programmed by hand to perform a task. With machine learning, some portion of the human contribution is replaced by a learning algorithm. As availability of computational capacity and data has increased, machine learning has become more and more practical over the years, to the point of being almost ubiquitous.

It can be used in two ways:

* Supervised Learning
* Unsupervised Learning

**Linear Regression**

Linear regression is a fundamental machine learning technique used for predicting a continuous target variable based on one or more input features. It's a form of supervised learning, where the algorithm learns to establish a linear relationship between the input features and the target variable.

In a simple linear regression, there's only one input feature, and the goal is to fit a straight line that best represents the relationship between the input feature and the target variable. The equation of the line is often represented as:

y=mx+b

Where:

* y is the target variable (output)
* x is the input feature
* m is the slope of the line
* b is the y-intercept

The goal of linear regression is to find the values of m and b that minimize the difference between the predicted values (based on the line) and the actual target values. This difference is typically measured using a loss function, such as mean squared error (MSE), and the algorithm aims to minimize this loss.

In multiple linear regression, there are multiple input features, and the equation becomes:

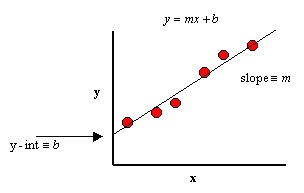
y=b0 +b1 x1 +b2 x2 +…+bnxn

Where:

* y is the target variable (output)
* b0 is the intercept
* b1 ,b2 ,…,bn are the coefficients for each input feature x1 ,x2 ,…,xn

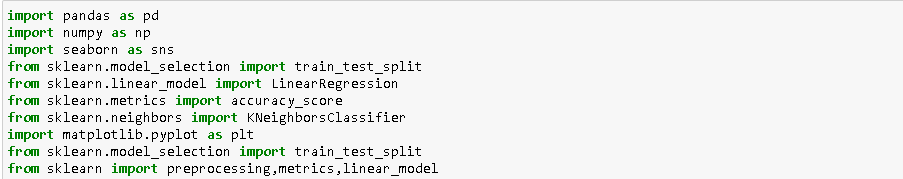
The process of linear regression involves finding the best values for the coefficients that minimize the loss function. This can be done using various optimization techniques, such as gradient descent.

Linear regression is widely used in various fields for tasks like predicting stock prices, housing prices, sales forecasts, and more, whenever there is a perceived linear relationship between input features and the target variable.

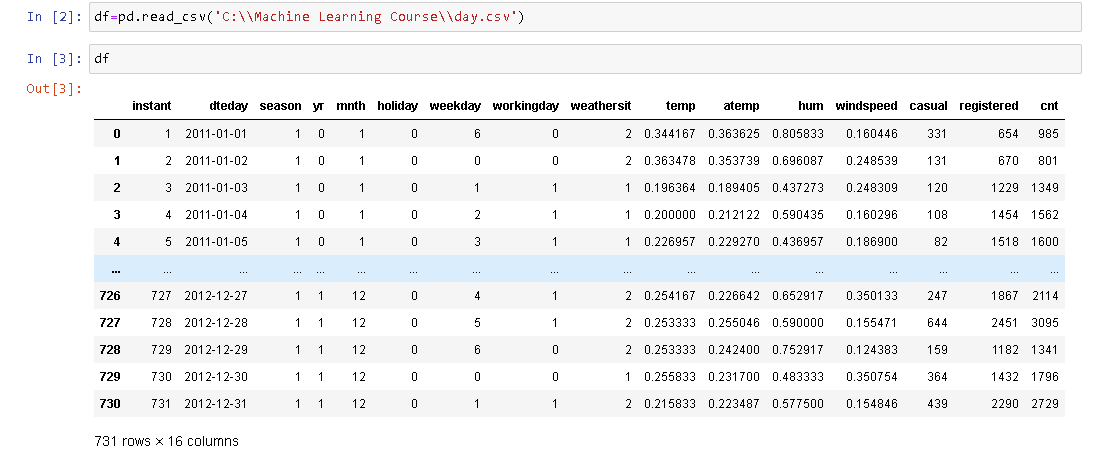


**METHODOLOGY**

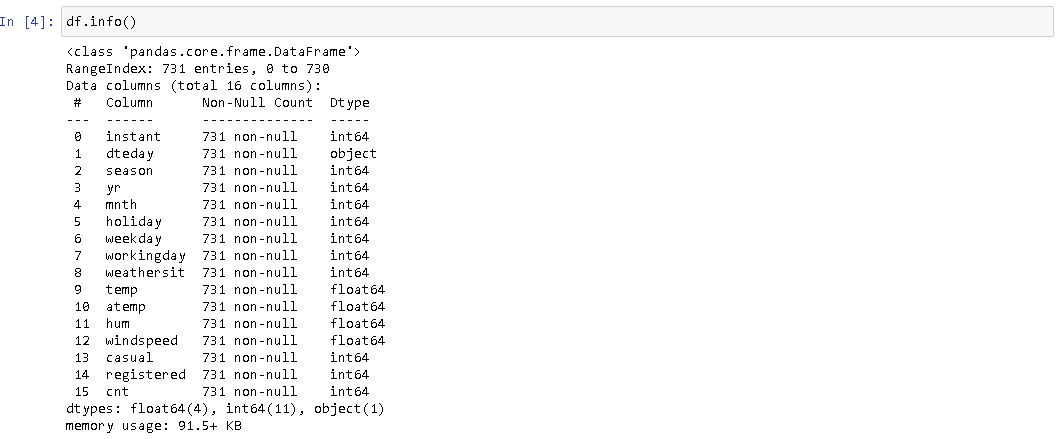
**Step1**.**Importing important and required libraries.**



**Step2**.**Reading Dataset.**

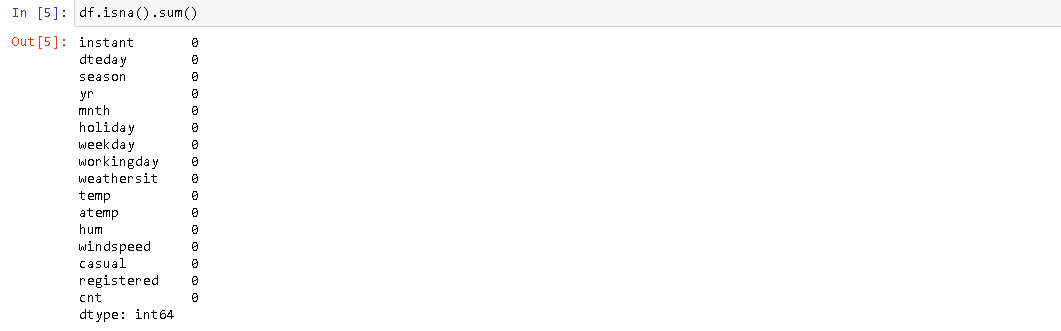


**Step3**.**Getting info about Data Type and Null Value Count.**



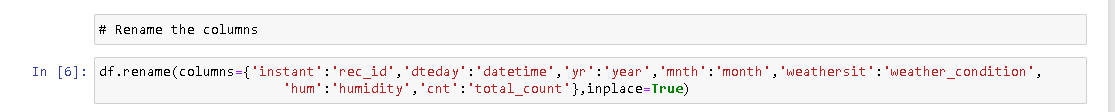
The dataset includes a combination of float,object and integer datatype and has float64(4), int64(11), object(1).The dataset has a total of 731 non-null characters.

**Step4. Missing Value Count**



The dataset has 0 missing value count.

**Step5 . Renaming Columns**

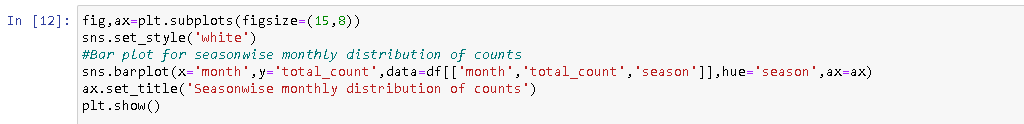


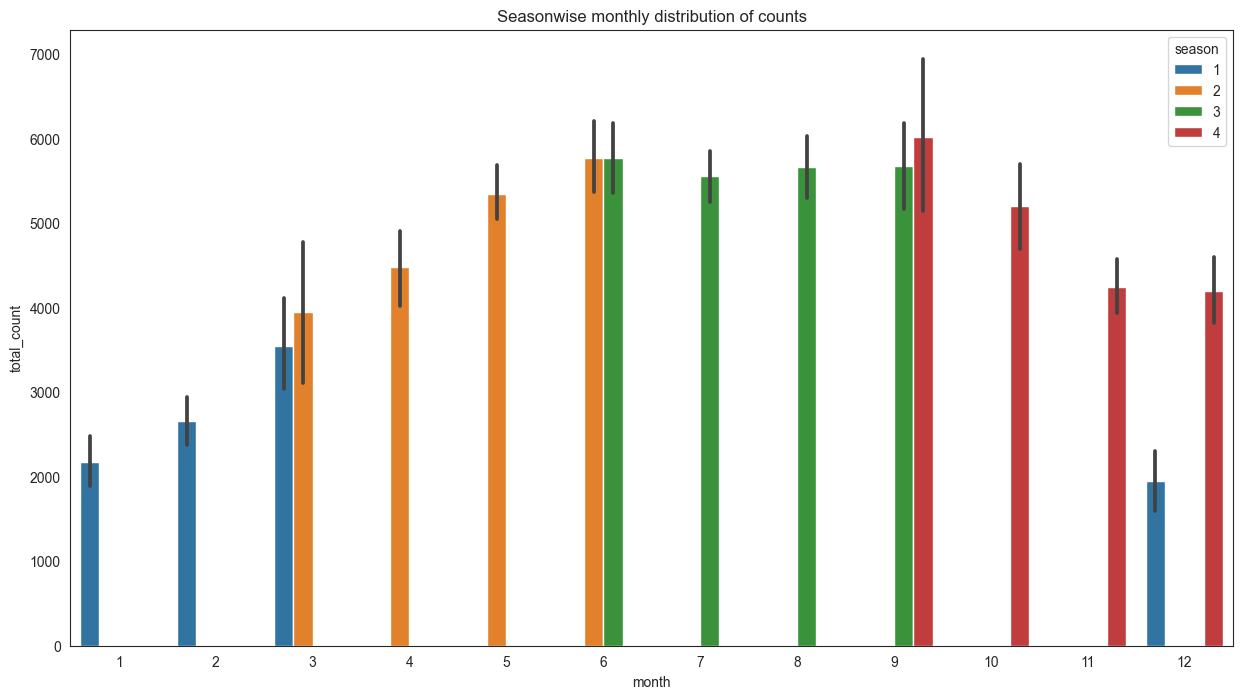
**Step6.** **Typecasting the datetime and numerical attributes.**

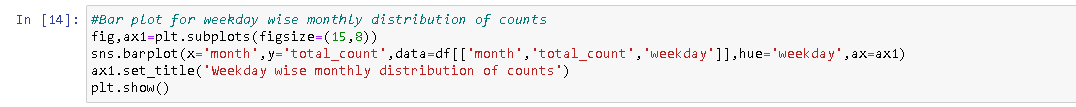


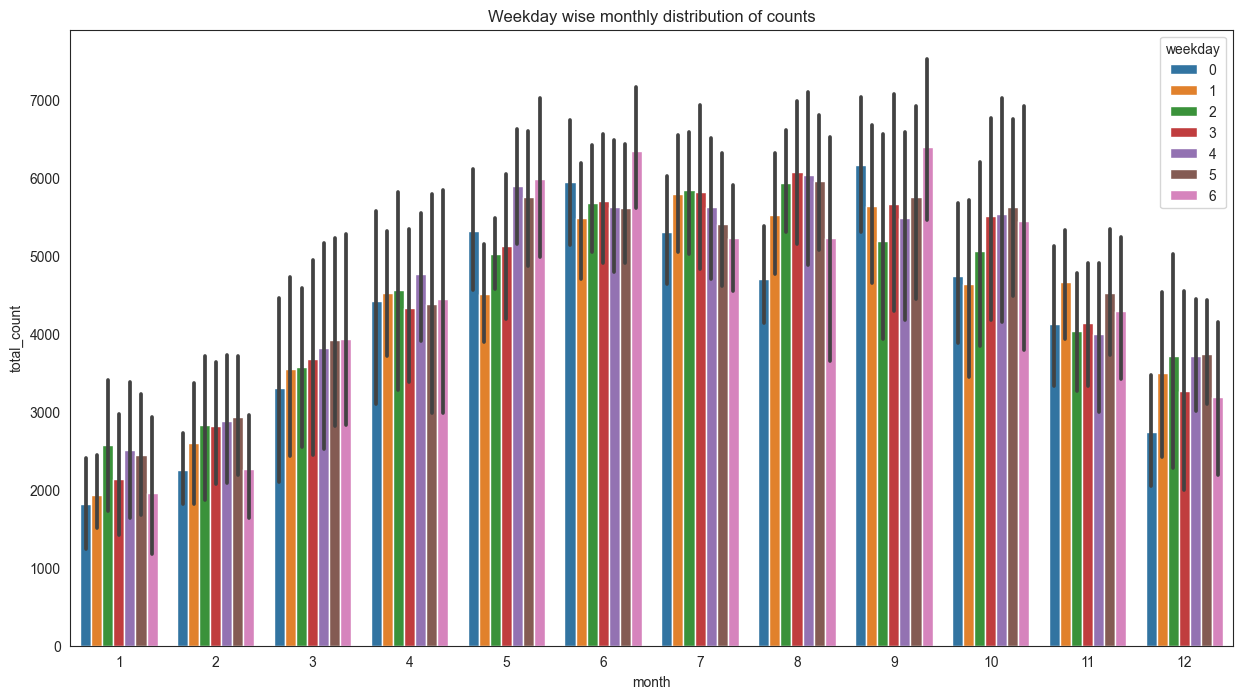
**Step 7. Distribution of Count**

* **Monthly Distribution of Count**



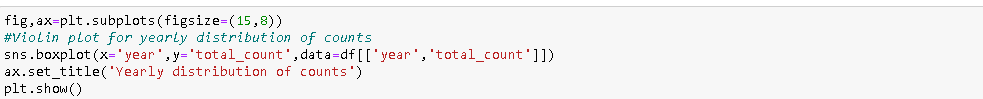


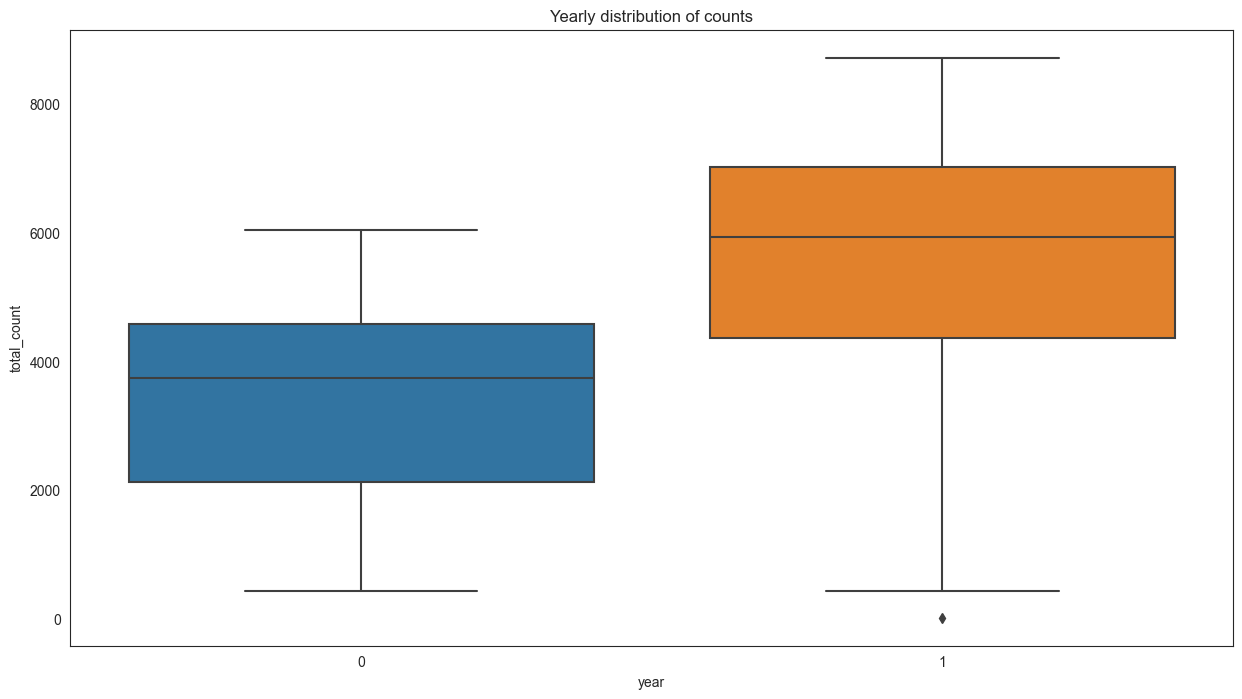




From the plots, we can observed that increasing the bike rental count in springe and summer season and then decreasing the bike rental count in fall and winter season.

* **Yearly Distribution Of Count**



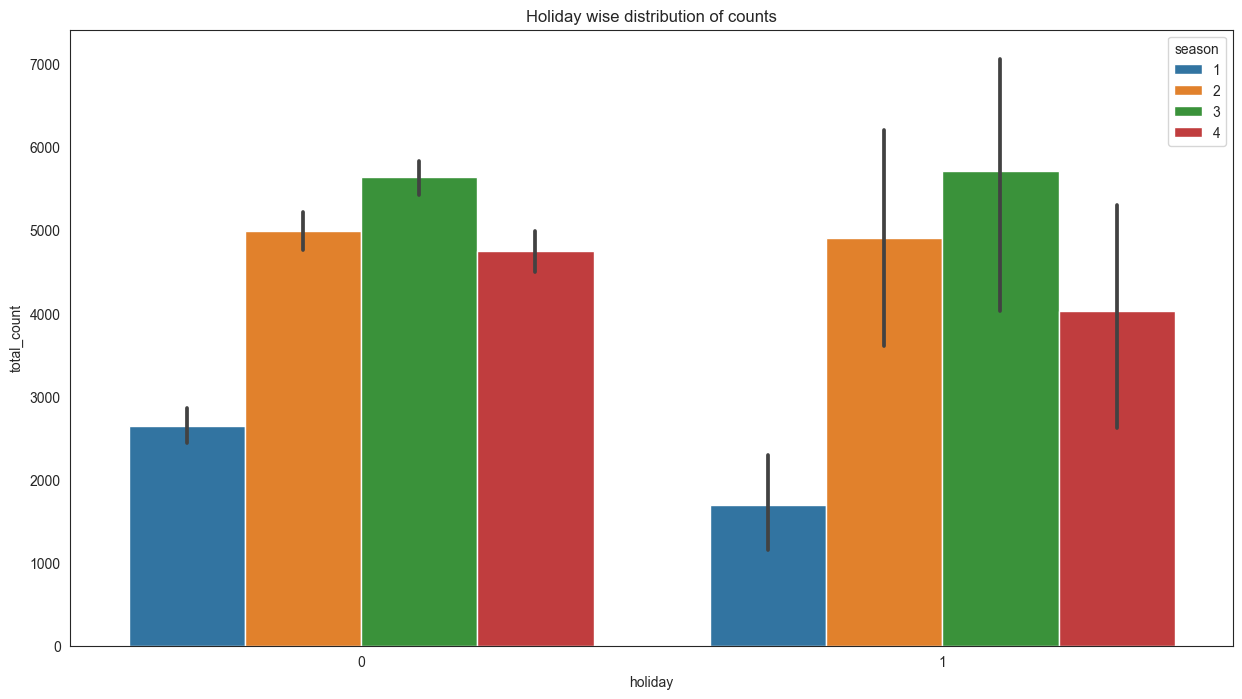


From the plot, we can observed that the bike rental count distribution is highest in year 2012 then in year 2011.

Here,  
year 0-> 2011, year 1-> 2012.

* **Holiday wise distribution of counts**



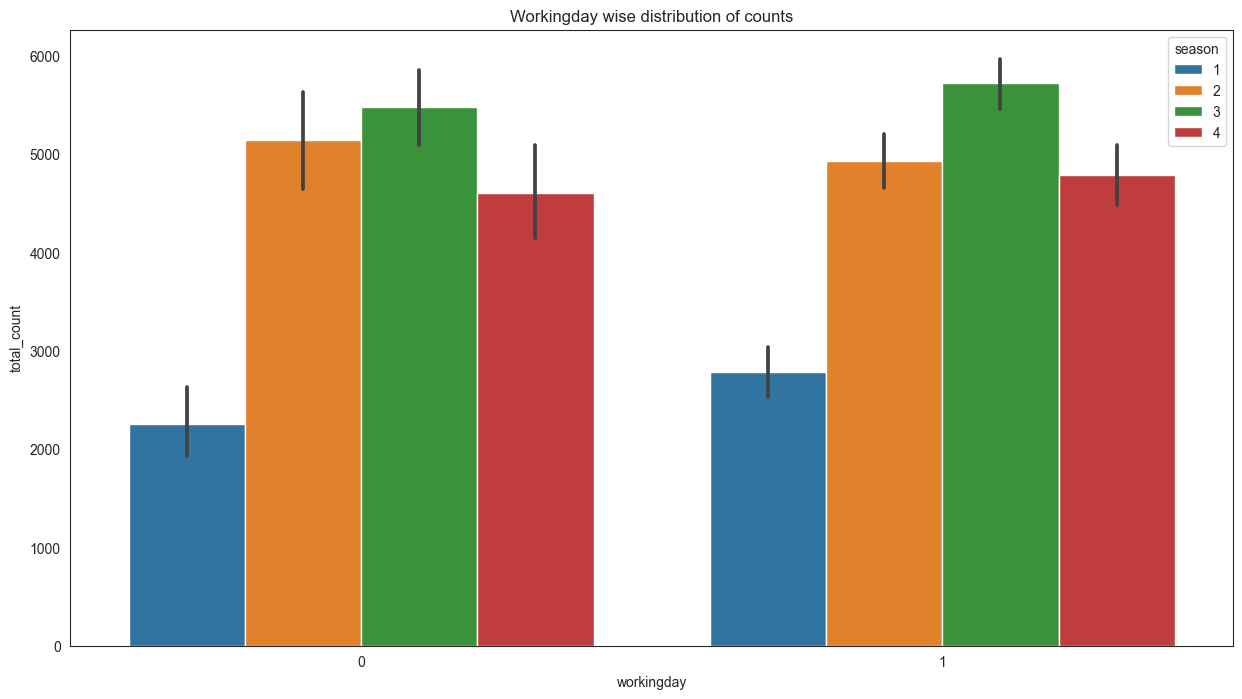


From the bar plot, we can observed that during no holiday the bike rental counts is highest compared to during holiday for different seasons.

Here, 0->No holiday, 1-> holiday

* **Working day wise distribution of count**

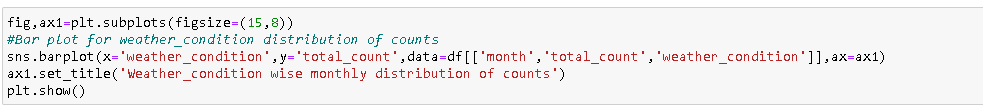


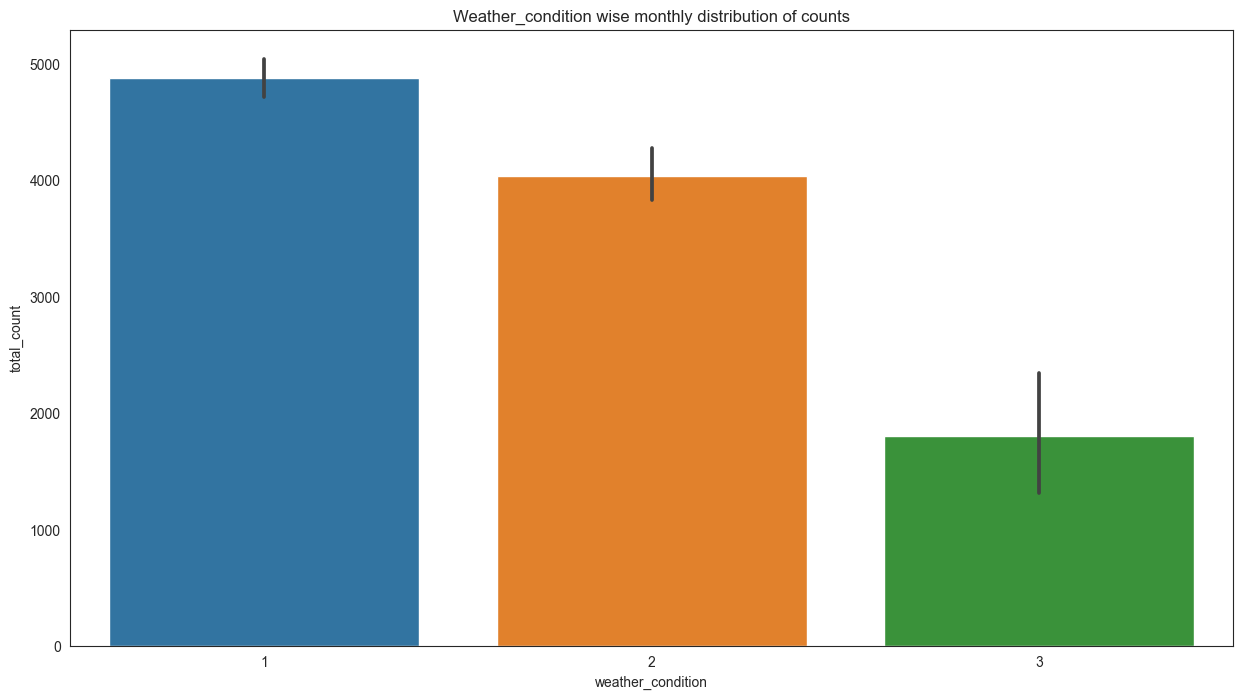


From the bar plot, we can observed that during workingday the bike rental counts is quite highest compared to during no workingday for different seasons.

Here, 0-> No workingday, 1-> workingday

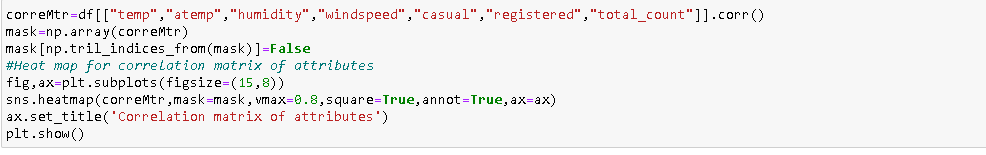
* **Weather condition distribution of counts**



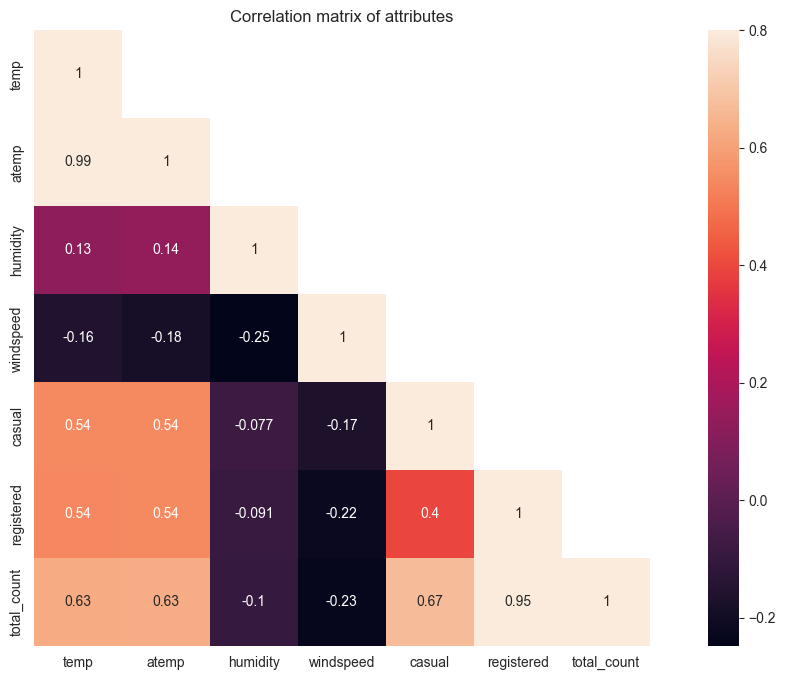


From the bar plot, we can observe that during clear,partly cloudy weather the bike rental count is highest and the second highest is during mist cloudy weather and followed by third highest during light snow and light rain weather.

**Step 8. Correlation Matrix**

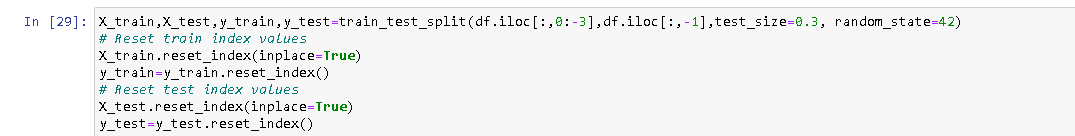


Correlation matrix is tells about linear relationship between attributes and help us to build better models.

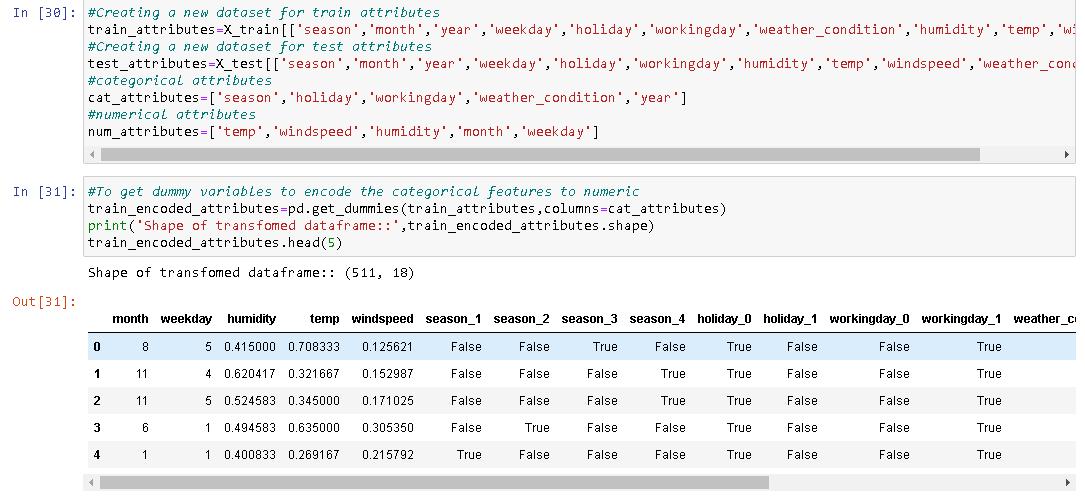


From correlation plot, we can observed that some features are positively correlated or some are negatively correlated to each other. The temp and atemp are highly positively correlated to each other, it means that both are carrying same information.The total\_count,casual and registered are highly positively correlated to each other. So, we are going to ignore atemp,casual and registered variable for further analysis.

**Step 9.Splitting dataset into train , test.**



**Step 10. Splitting the features into categorical and numerical features.**



**Step 11.Training dataset for modelling.**



**Step 12.Model Selection i****.e. linear regression in this case ,training and fitting into it.**

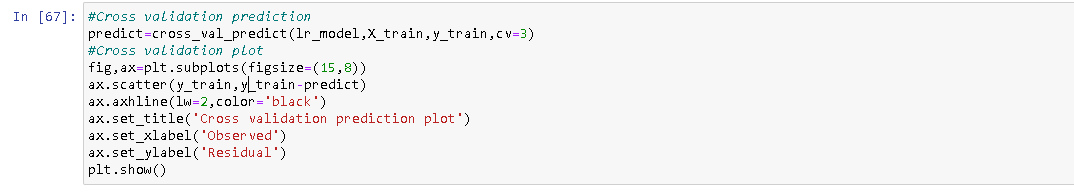


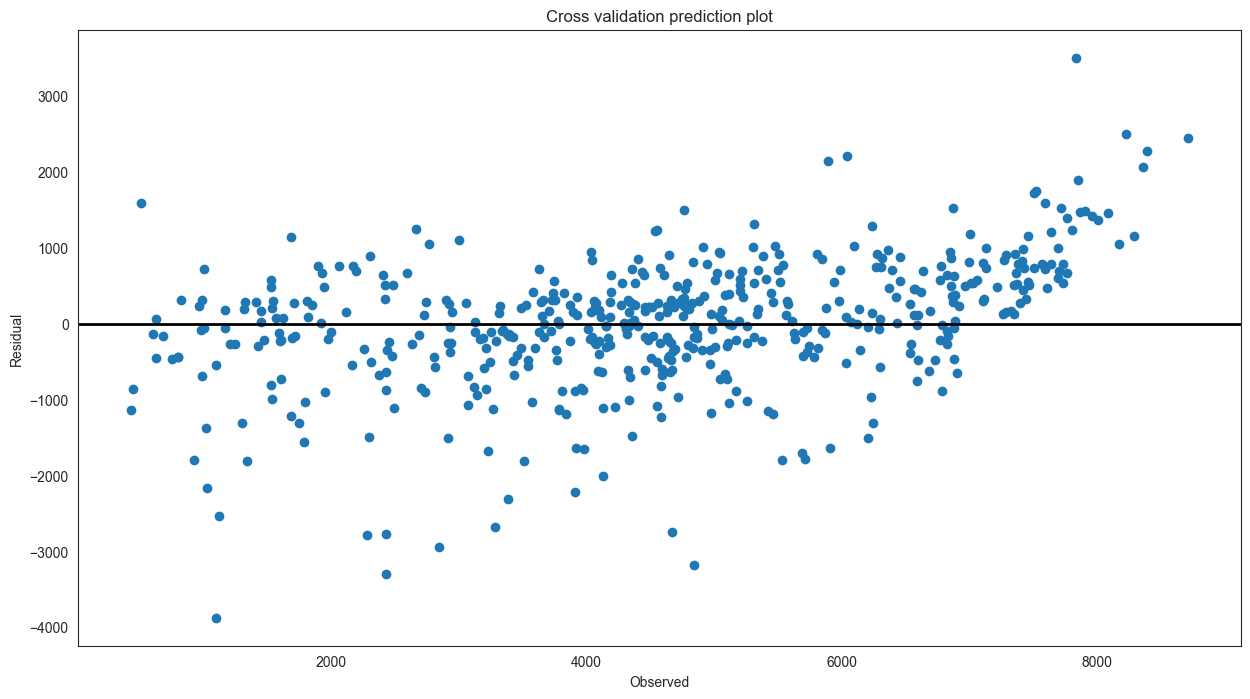
**Step 13.Accuracy of Model**



Accuracy of model in the dataset using linear regression model is81.64644195384456%.

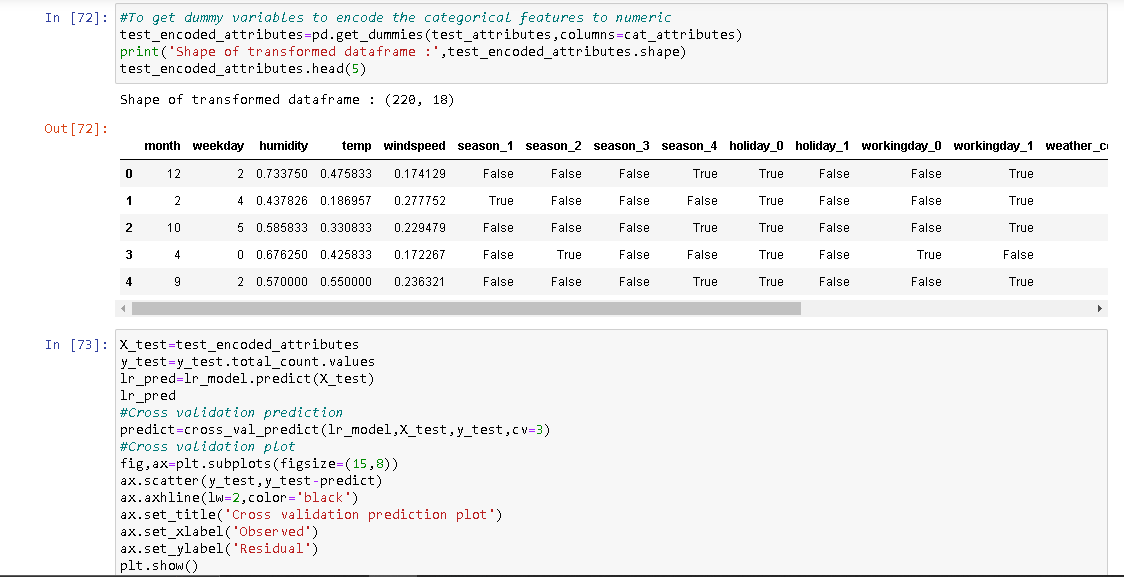
**Step 14.Cross Validation Prediction using Train Dataset**

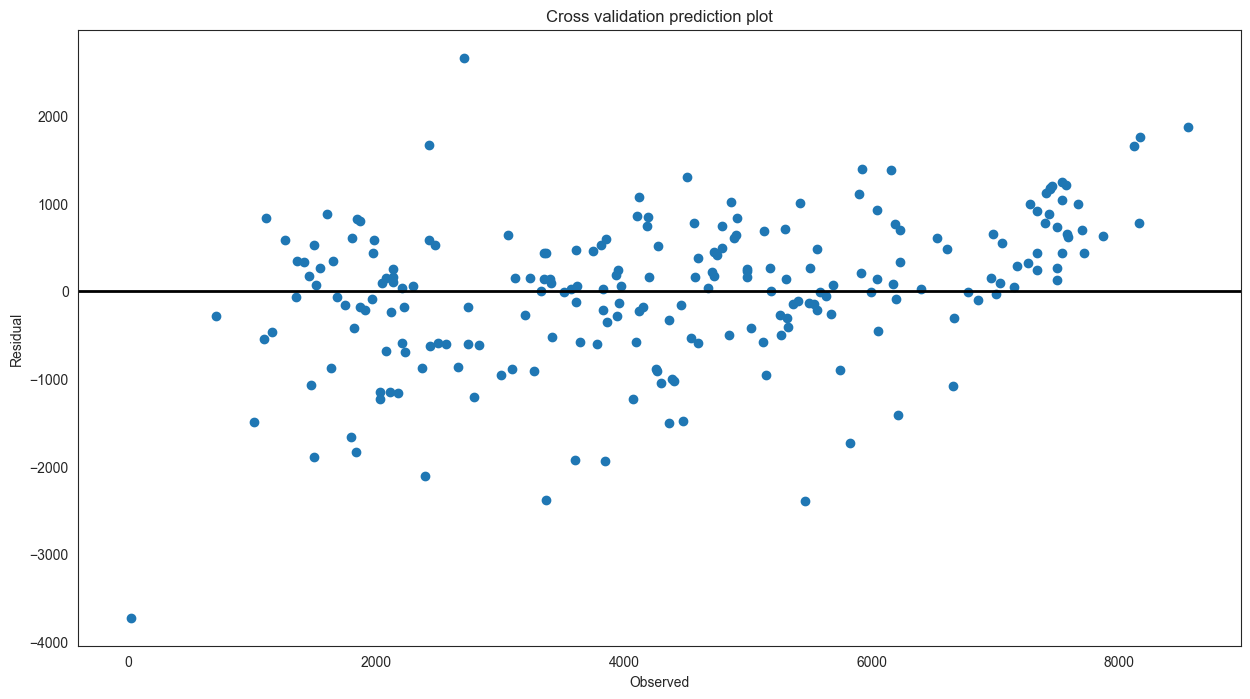




Cross validation prediction plot tells about finite variance between actual target value and predicted target value. In this plot, some data points are have same finite variance between them and for some are not have it.

**Step 14.Cross Validation Prediction using Test Dataset**





Residual plot tells about finite variance between actual target value and predicted target value.In this plot,very less data points are have same finite variance between them and for most are not have it.

**Conclusion**

In this project, bike rental count prediction I have used Linear Regression to train the model with the dataset encompasses a range of attributes that provide comprehensive insights into bike-sharing patterns. The attributes include sequential indices denoted by "instant," date information captured by "dteday," and seasonal distinctions marked by values 1 to 4 in the "season" attribute. The "yr" attribute indicates the year, with 0 representing 2011 and 1 representing 2012, while months are represented numerically in the "mnth" attribute. "Holiday" and "weekday" attributes record holiday occurrences and days of the week, respectively, with "workingday" distinguishing between workdays and weekends/holidays.

For all the above data given graphs are plotted for the analysis individually and been shown as bar plot .This gives a better understanding and which data should be used for prediction and also seen as a whole.

On seeing with respect to weather it is seen that on spring and summer season ,rental count is comparitively more when compared with rest other season. Also,count is more on no holidays and less on holidays with respect to any season and can been seen same with respect to working day and no working day respectively.In 2012, the count is highest or can be said more than 2011. Also, during clear,partly cloudy weather the bike rental count is highest and the second highest is during mist cloudy weather and followed by third highest during light snow and light rain weather.

For training a correlation matrix is set up and the temp and atemp are highly positively correlated to each other, it means that both are carrying same information.The total\_count,casual and registered are highly positively correlated to each other. So, we are going to ignore atemp,casual and registered variable for further analysis.

Then the model is fitted into linear regression model and when accuracy is predicted,shows a accuracy of 81.64644195384456%.Cross Validation prediction for both test and train data is predicted using the Linear Regression Model and graphs are plotted respectively.They give the finite variance between actual target value and predicted target value.