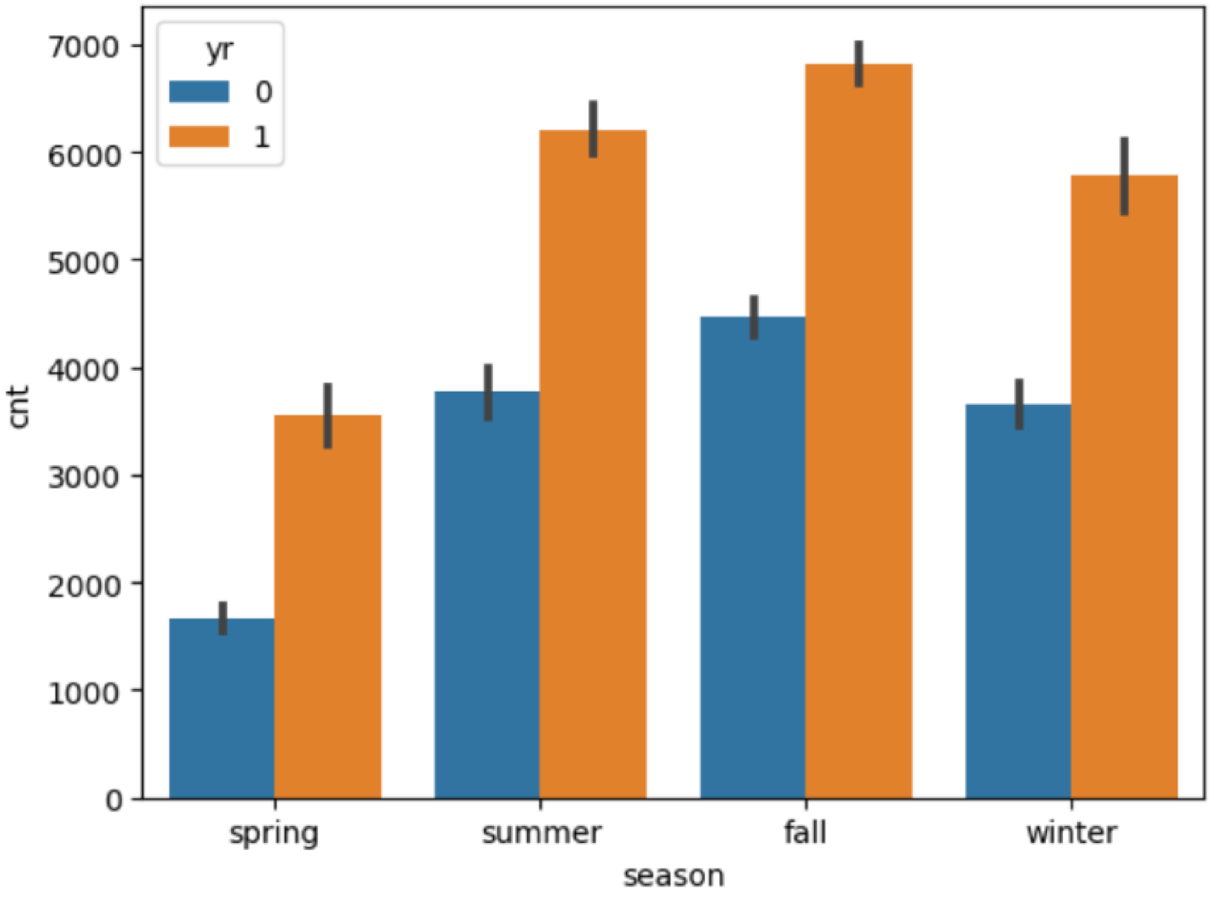
Assignment-based Subjective Questions

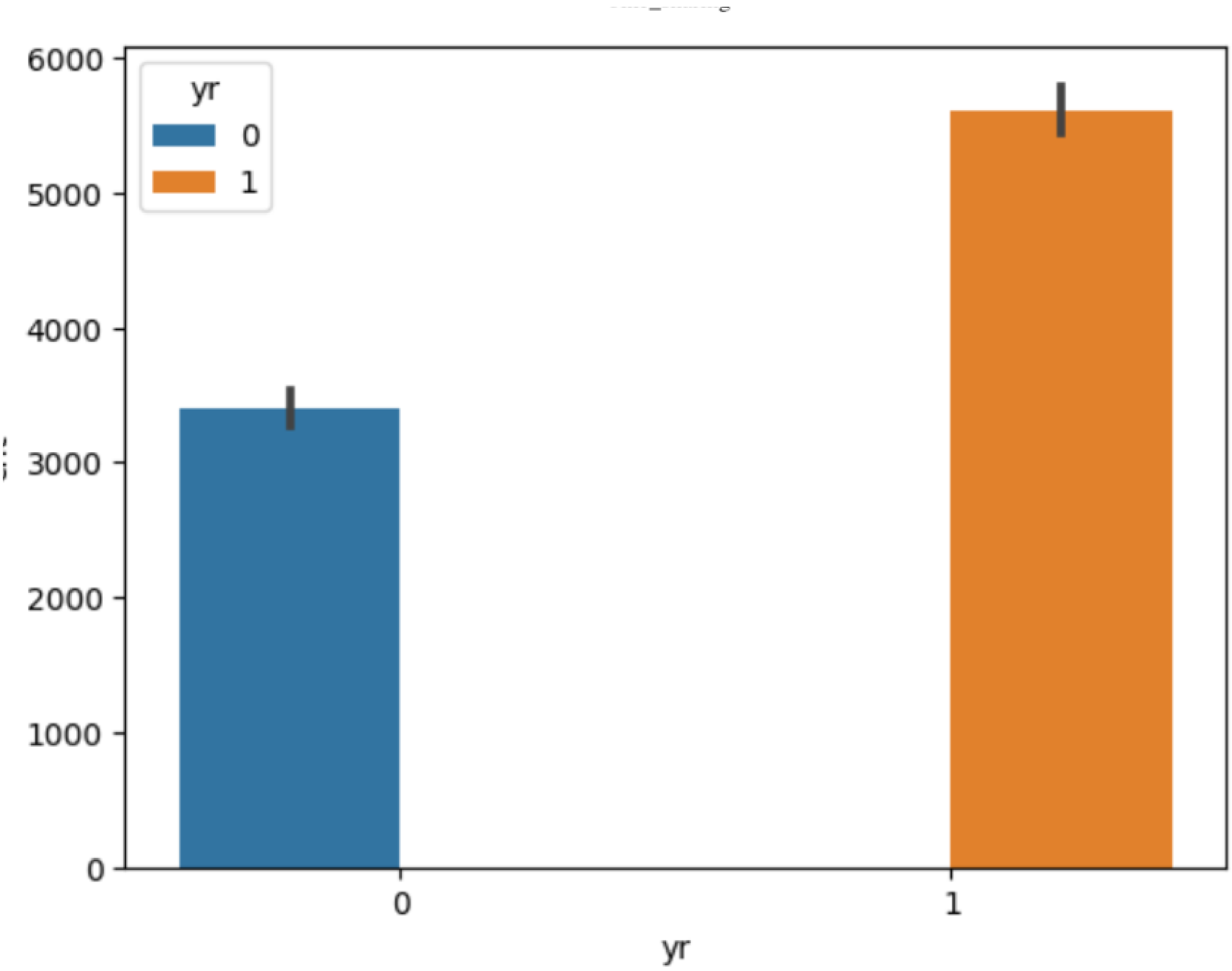
1. **From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?**

Bike demand in the fall is the highest and lowest in spring for the years

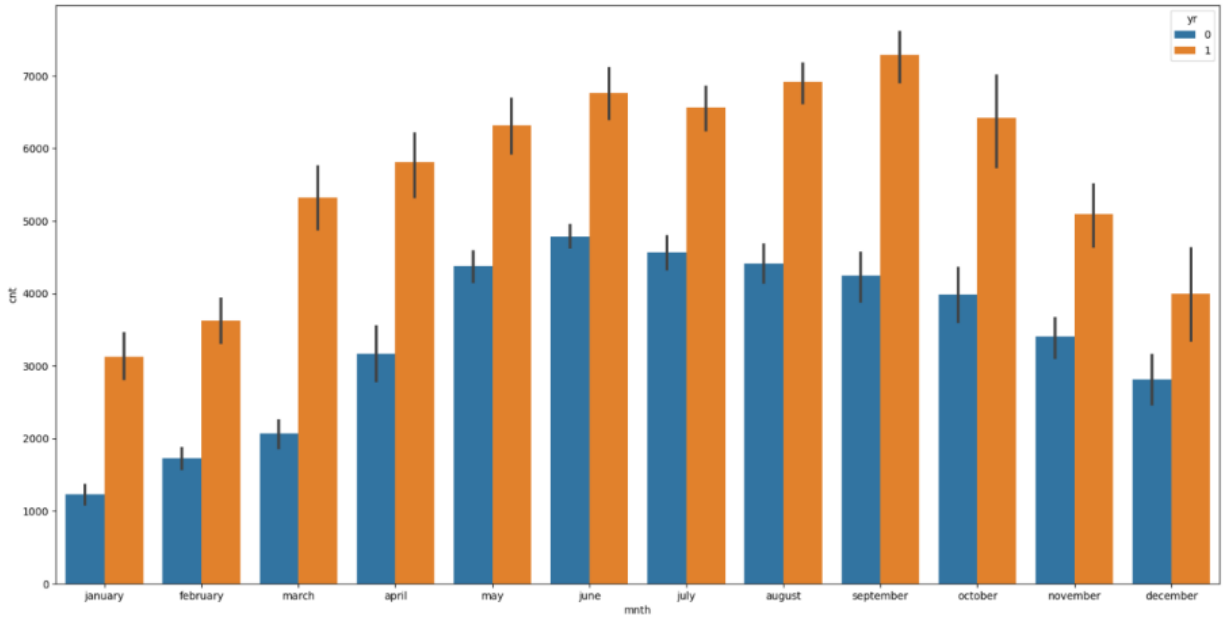


Bike demand in year 2019 is higher as compared to 2018.

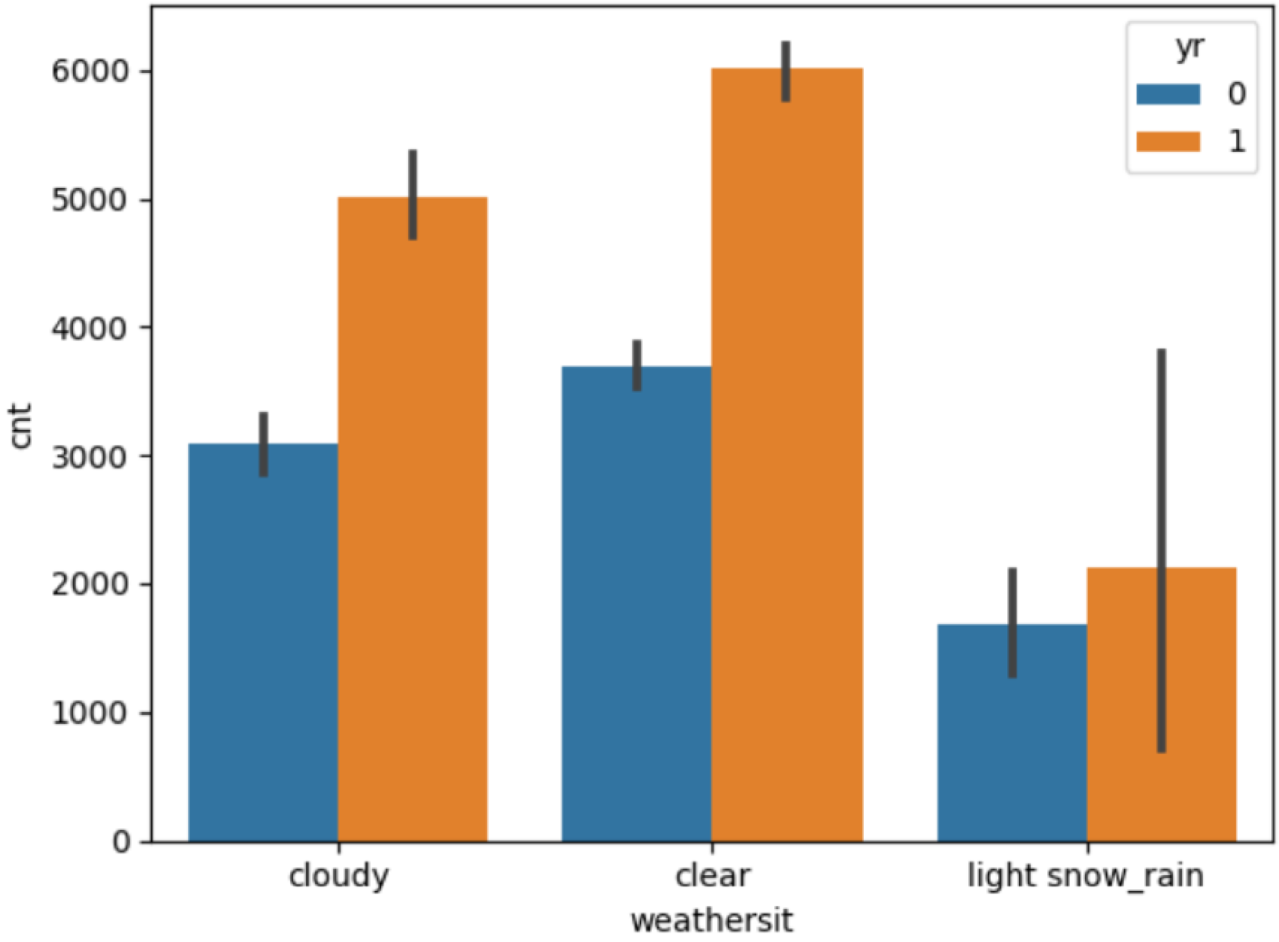
Note- 0: 2018 , 1:2019



Bike demand is high in the months from May to October.



Bike demand is high if weather is clear or with mist cloudy while it is low when there is light rain or light snow.



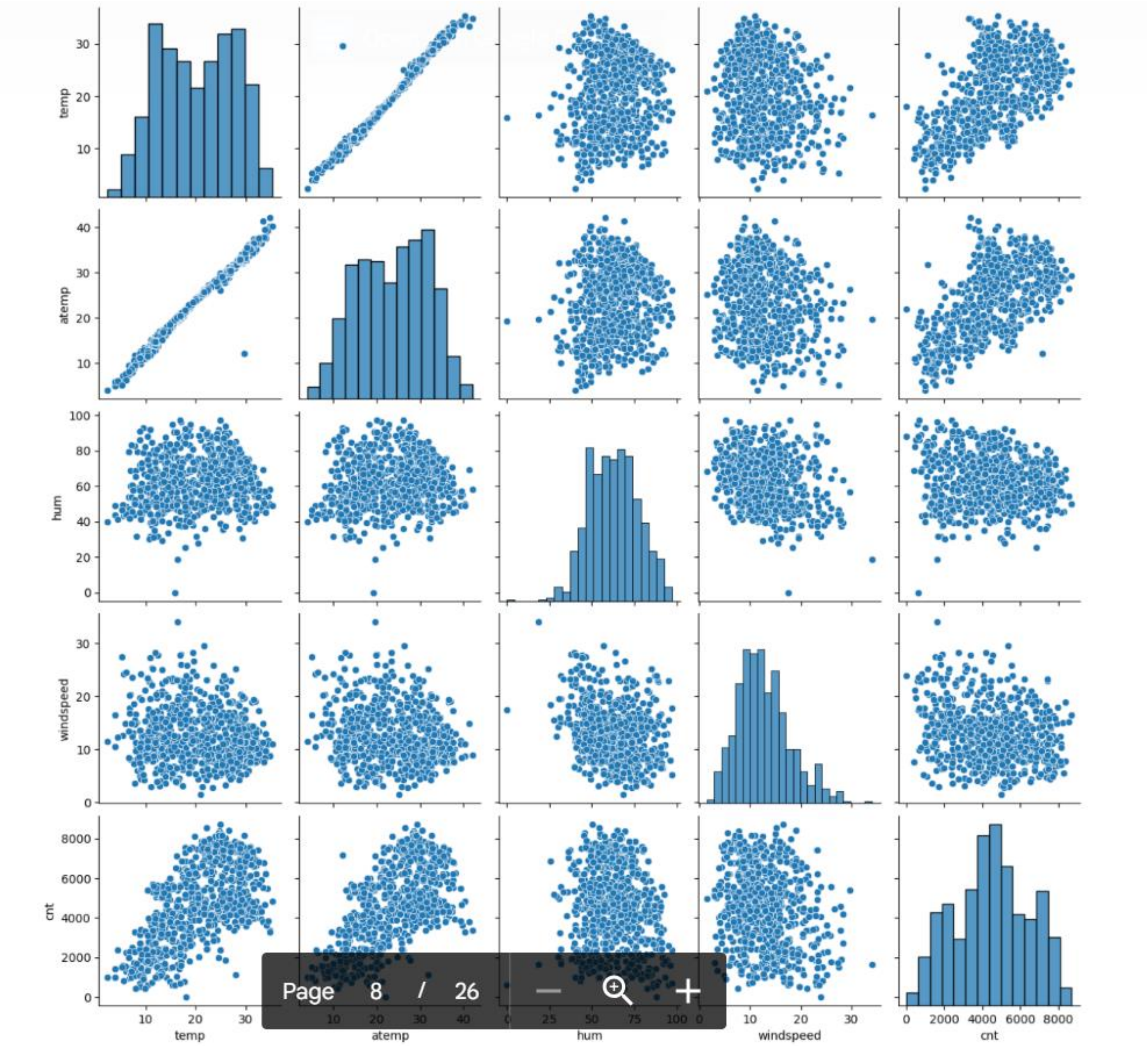
The demand of bike is similar throughout the weekdays

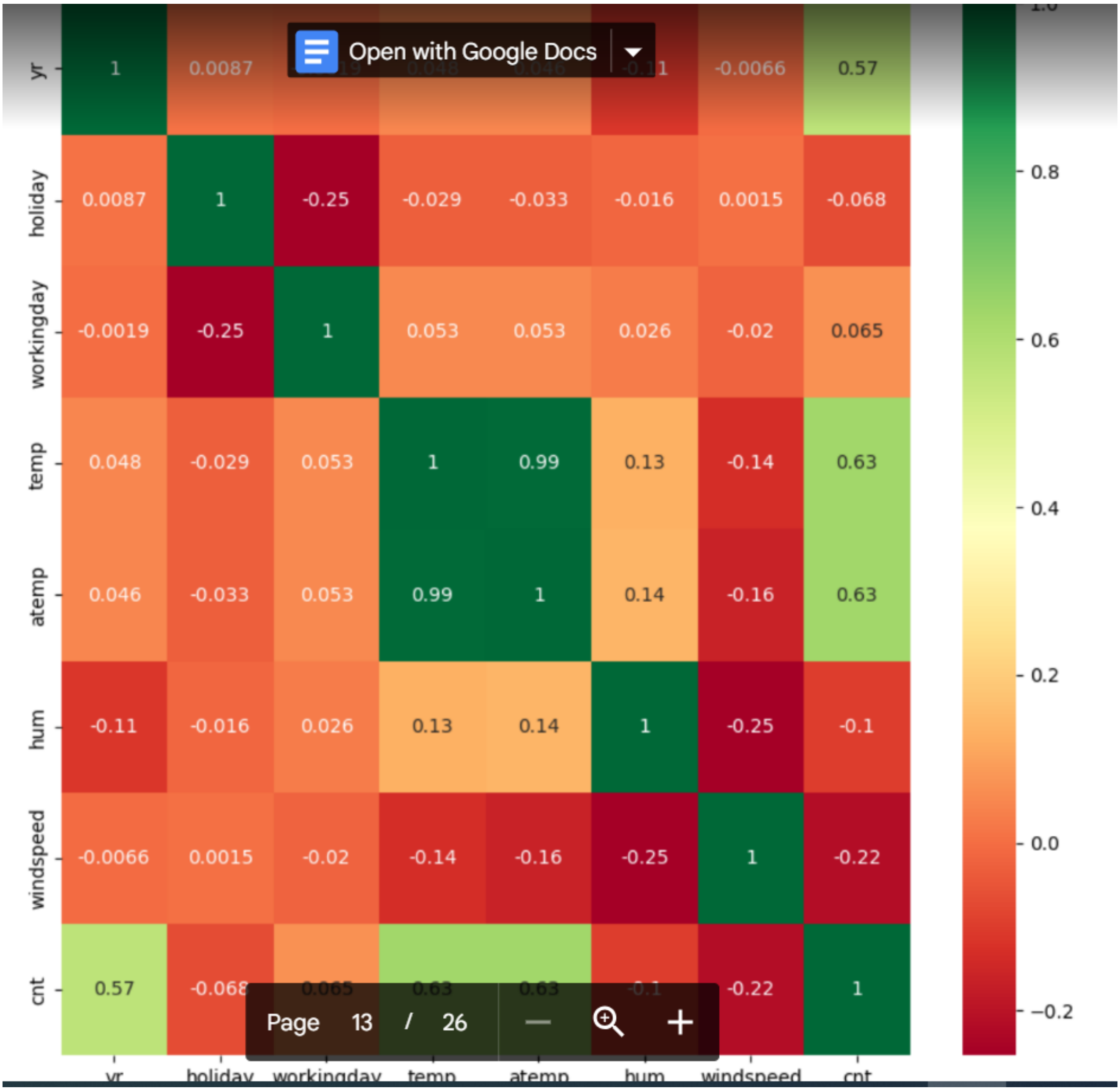
2. **Why is it important to use drop_first=True during dummy variable creation?**

drop_first=True argument drops the extra column that has been generated during the dummy Variable creation.As a rule fo thumb if a categorical column has n values then it is advisable to have only n-1 dummy variables for it.The argument drop_first helps in dropping that extra column

3. **Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?**

Temp and Atemp have the highest correlation

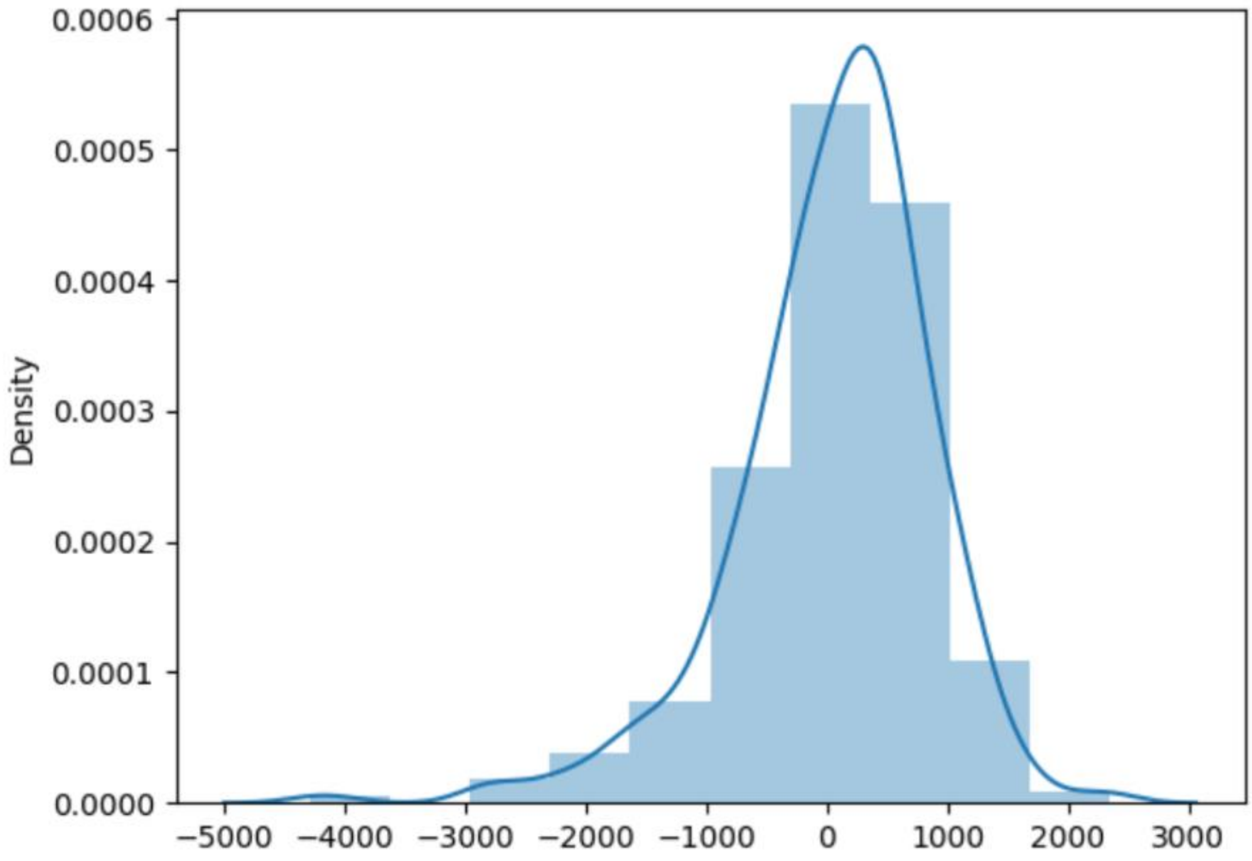




	yr	holiday	workingday	temp	atemp	hum	windspeed	cnt
yr	1	0.0087	-0.0019	0.048	0.046	-0.11	-0.0066	0.57
holiday	0.0087	1	-0.25	-0.029	-0.033	-0.016	0.0015	-0.068
workingday	-0.0019	-0.25	1	0.053	0.053	0.026	-0.02	0.065
temp	0.048	-0.029	0.053	1	0.99	0.13	-0.14	0.63
atemp	0.046	-0.033	0.053	0.99	1	0.14	-0.16	0.63
hum	-0.11	-0.016	0.026	0.13	0.14	1	-0.25	-0.1
windspeed	-0.0066	0.0015	-0.02	-0.14	-0.16	-0.25	1	-0.22
cnt	0.57	-0.068	0.065	0.63	0.63	-0.1	-0.22	1

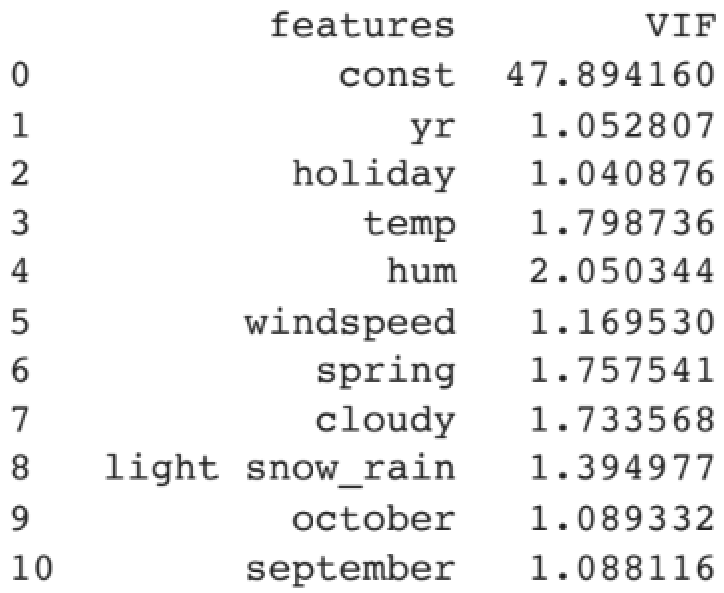
4. How did you validate the assumptions of Linear Regression after building the model on the training set?

Validated the error distribution distribution is normal by plotting a distplot between y_actual and y_predicted.



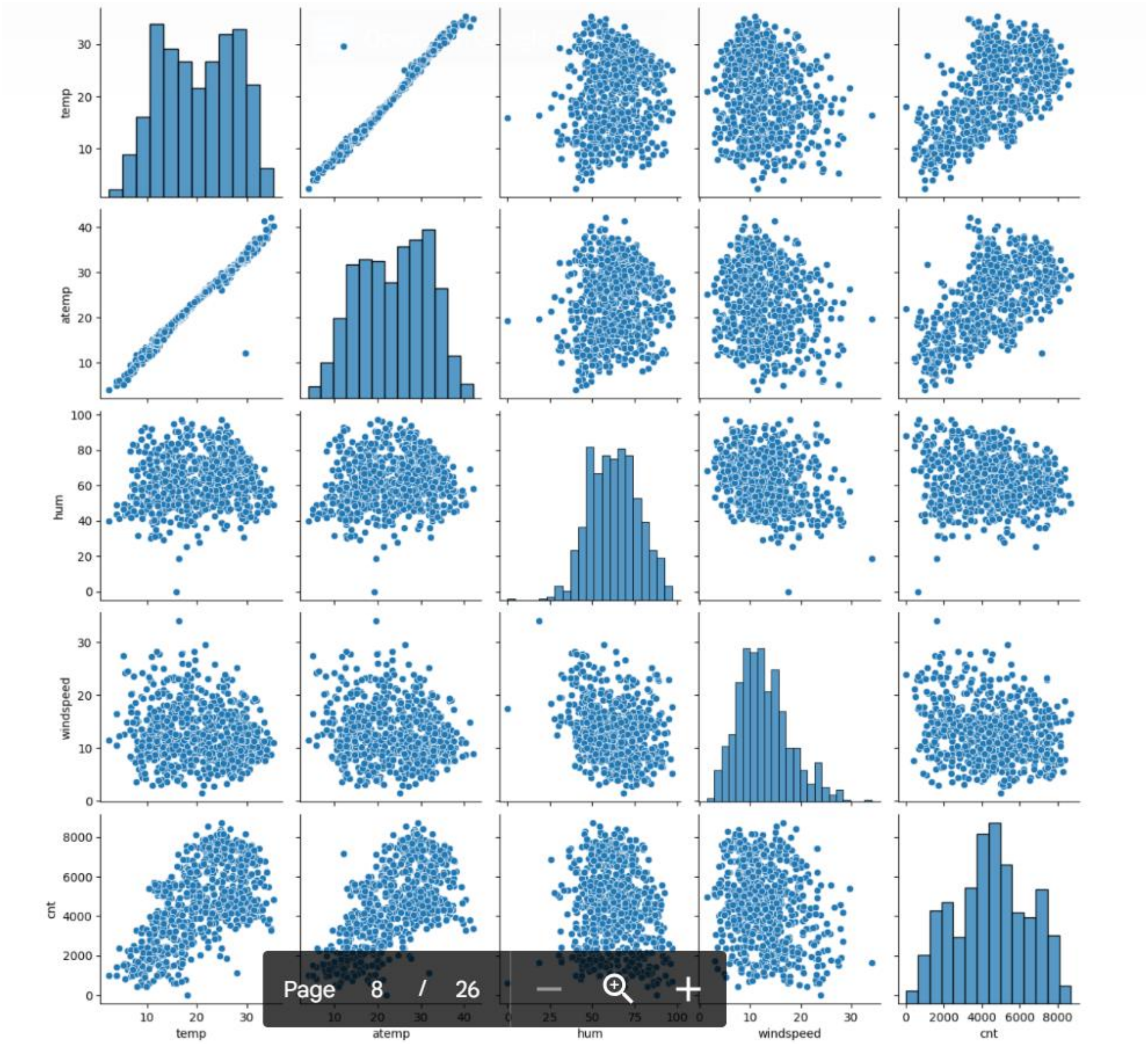
Validated the mean of error is 0 by (y_actual-y_predicted).mean()

Calculated VIF of the variables to check if any Multi-collinearity existed or not between them



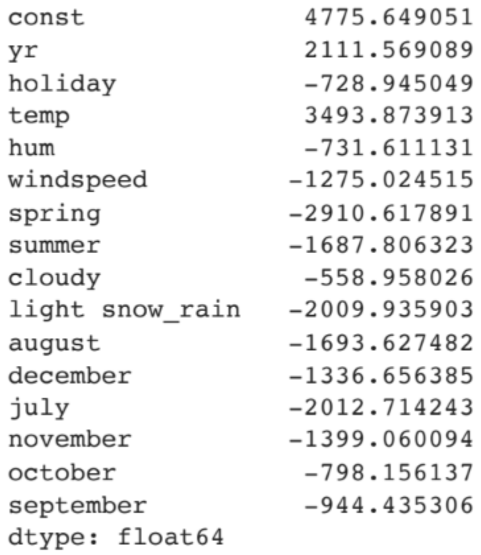
```
              features        VIF
0                const  47.894160
1                   yr   1.052807
2              holiday   1.040876
3                 temp   1.798736
4                  hum   2.050344
5            windspeed   1.169530
6               spring   1.757541
7               cloudy   1.733568
8    light snow_rain   1.394977
9              october   1.089332
10           september   1.088116
```

Checked if there is a linear relationship between target and the independent variable.



5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?

Temprature, Spring,Year



```
const                4775.649051
yr                   2111.569089
holiday              -728.945049
temp                 3493.873913
hum                  -731.611131
windspeed           -1275.024515
spring              -2910.617891
summer              -1687.806323
cloudy               -558.958026
light snow_rain     -2009.935903
august              -1693.627482
december            -1336.656385
july                -2012.714243
november            -1399.060094
october              -798.156137
september            -944.435306
dtype: float64
```

General Subjective Question

1. Explain the linear regression algorithm in detail.

Linear Regression is one of the most simple and powerful Supervised Machine Learning Predictive Algorithm that is used to predict Future Values based on some Variables .There are 2 types of Linear Regression Algorithm

- Simple Linear Regression Algorithm
- Multiple Linear Regression Algorithm

Liner Regression is represented by the equation for Simple Linear Regression.

y=mx+c

Where,

c=Intercept made on the Y Axis

m=Slope of the Line

x=Independent Variable

y=Dependent Variable

The line y=mx+c is called the regression line .Relationship between the Dependent and Independent Variable can be positive as well as negative.In case of a positive relationship the value of dependent variable increases with the increase in Independent variable and for negative relationship the value of dependent variable decrease with increase in Independent variable.

To get the performance of the Linear Regression we calculate R2 score. R2 score is calculated as the summation of the difference between abs(Y_actual-Y_predicted)

Below are the assumption for Linear Regression

- Error Values are normally distributed
- There exist a linear relationship between the dependent variable and independent variable
- There is no multi-collinearity between the variables used in prediction.We calculate this using VIF.
- There is no correlation between error terms

2. Explain the Anscombe’s quartet in detail

Anscombe's quartet is a set of four datasets that have nearly identical simple descriptive statistics, yet have very different distributions and appear very different when graphed. Each dataset consists of eleven (x,y) points.The quartet is still often used to illustrate the importance of looking at a set of data graphically before starting to analyze according to a particular type of relationship, and the inadequacy of basic statistic properties for describing realistic datasets.

3.What is Pearson’s R?

The Pearson's correlation coefficient varies between -1 and +1 where:

r = 1 means the data is perfectly linear with a positive slope (i.e., both variables tend to change in the same direction)
r = -1 means the data is perfectly linear with a negative slope (i.e., both variables tend to change in different directions)
r = 0 means there is no linear association
r > 0 < 5 means there is a weak association
r > 5 < 8 means there is a moderate association
r > 8 means there is a strong association
Pearson's r is a numerical summary of the strength of the linear association between the variables. If the variables tend to go up and down together, the correlation coefficient will be positive. If the variables tend to go up and down in opposition with low values of one variable associated with high values of the other, the correlation coefficient will be negative.

4.What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?

It is a step of data Pre-Processing which is applied to independent variables to normalize the data within a particular range. It also helps in speeding up the calculations in an algorithm.

Most of the times, collected data set contains features highly varying in magnitudes, units and range. If scaling is not done then algorithm only takes magnitude in account and not units hence incorrect modelling. To solve this issue, we have to do scaling to bring all the variables to the same level of magnitude.

Normalization/Min-Max Scaling:

It brings all of the data in the range of 0 and 1. **sklearn.preprocessing.MinMaxScaler** helps to implement normalization in python.

Standardization Scaling:

Standardization replaces the values by their Z scores. It brings all of the data into a standard normal distribution which has mean (**μ**) zero and standard deviation one (**σ**).

5.You might have observed that sometimes the value of VIF is infinite. Why does this happen?

VIF is an index that provides a measure of how much the variance of an estimated regression coefficient increases due to collinearity. In order to determine VIF, we fit a regression model between the independent variables.

If all the independent variables are orthogonal to each other, then VIF = 1.0. If there is perfect correlation, then VIF = infinity. A large value of VIF indicates that there is a correlation between the variables.

6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in inear regression.sdn

It is a plot, or quantile-quantile plot, is a graphical tool that can be used to assess the distribution of a set of data. It is a probability plot that compares the quantiles of a sample distribution to the quantiles of a theoretical distribution. The most common theoretical distribution used for Q-Q plots is the normal distribution.

In linear regression, a Q-Q plot can be used to assess the assumption of normality of the residuals. The residuals are the differences between the observed values and the predicted values from the regression model. If the residuals are normally distributed, then the Q-Q plot will be a straight line. However, if the residuals are not normally distributed, then the Q-Q plot will deviate from a straight line.