General Subjective Question

General Subjective Question

1) Explain the linear regression algorithm in detail?

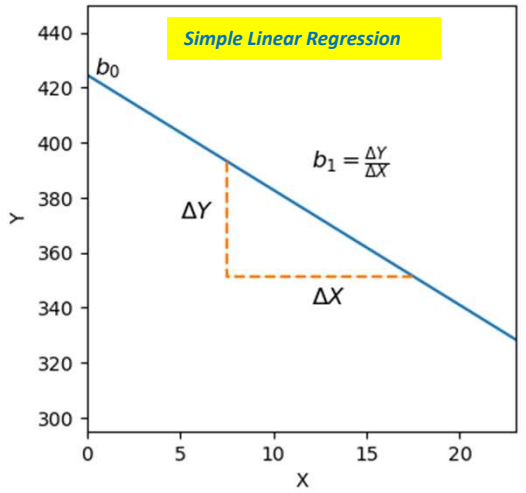
Simple Linear Regression:

Simple Linear regression provides a model of the relationship between the magnitude of one variable and that of as second for Example ,as X increases, Y also increases. Or as X increases, Y decreases. Correlation is another way to measure how two variable are related . The difference between is that while correlation measure the strength of an association between two variables , regression quantifies the nature of the relation ship.

Simple linear regression estimate how much Y will change when X changes by a certain amount. With regression , we are trying to predict the y variable from X using a Linear relationship

$$Y = b_0 + b_1 X$$

We read this as "Y equal b_1 times X ,plus a constant b_0". The symbol b_0 is known as intercept (or constant) , and the symbol b_1 as the slope for X . The Y variable is known as the dependent variable since it depend on X. The X variable known as independent variable.



Fitted value and Residuals:

Fitted values (the predictions) and residuals (predicted errors). In General , the data doesn't fall exactly on a line ,so the regression equation should include an explicit error term ϵ

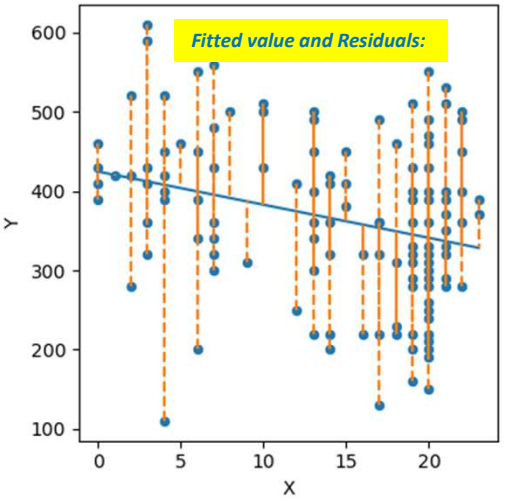
$$Y = b_0 + b_1 X + e$$

The fitted values also referred to as the predicted values ,are typically denoted by $\widehat{Y_i}$

$$\widehat{Y_i} = \widehat{b_0} + \widehat{b_1} X_1$$

we compute the residual $\widehat{e_i}$ by subtracting the predicted values from original data

$$\widehat{e_i} = Y_i - \widehat{Y_i}$$



Least Square:

In practice , the regression line is the estimate that minimizes the sum of square residual values ,also called the residual sum of squares or RSS

$$\text{RSS} = \sum_{I=1}^{n} (Y_i - \widehat{Y_i})^2$$

Multiple Linear Regression :

When there are multiple predictors or independent variable , the equation is simple extended to accommodate them.

$$Y = b_0 + b_1 X_1 + b_2 X_2 + b_3 X_3 + b_4 X_4 + \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots + b_P X_P + \epsilon$$

3) What is Pearson's R?

Exploratory data analysis in many modelling projects involves examining's correlation among predictors, and between predictors and a target variable. Variables X and Y are said to be positively correlated if high values of X go with high values of Y, and low values of X, go with low values of Y. If high values of X, go with low values of Y, and vice versa, the variables are negatively correlated.

We used standardized variant approach for correlation. The coefficient , which give an estimate of the correlation between two variables that always lie on the same scale. To compute Pearson's correlation coefficient ,we multiply deviations from the mean for variable 1 time those variable 2 , and divide the product of the standard deviations

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{(n-1)s_x s_y}$$

The Pearson's R correlation coefficient always lies between +1 (perfect positive correlation) and -1(perfect negative correlation); 0 indicates no correlation.

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

Feature scaling is a data preprocessing technique used to transform the values of features or variables in a dataset to a similar scale. The purpose is to ensure that all features contribute equally to the model and to avoid the domination of features with larger values.

When you have a lot of independent variables in a model, a lot of them might be on very different scales which will lead a model with very weird coefficients that might be difficult to interpret. Feature scaling becomes necessary when dealing with datasets containing features that have different ranges, units of measurement, or orders of magnitude. In such cases, the variation in feature values can lead to biased model performance or difficulties during the learning process.

There are several common techniques for feature scaling, including standardization, normalization, and min-max scaling. These methods adjust the feature values while preserving their relative relationships and distributions.

By applying feature scaling, the dataset's features can be transformed to a more consistent scale, making it easier to build accurate and effective machine learning models. Scaling facilitates meaningful comparisons between features, improves model convergence, and prevents certain features from overshadowing others based solely on their magnitude.

Normalization *is a data preprocessing technique used to adjust the values of features in a dataset to a common scale. This is done to facilitate data analysis and modeling, and to reduce the impact of different scales on the accuracy of machine learning models.*
Normalization is a scaling technique in which values are shifted and rescaled so that they end up ranging between 0 and 1. It is also known as Min-Max scaling.
Here's the formula for normalization:

-

$$X' = \frac{X - X_{min}}{X_{max} - X_{min}}$$

S***tandardization*** *is another scaling method where the values are centered around the mean with a unit standard deviation. This means that the mean of the attribute becomes zero, and the resultant distribution has a unit standard deviation.*

Here's the formula for standardization: Mu is the mean of the feature values and σ *is the standard deviation of the feature values. Note that, in this case, the values are not restricted to a particular range.*

$$X' = \frac{X - \mu}{\sigma}$$

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

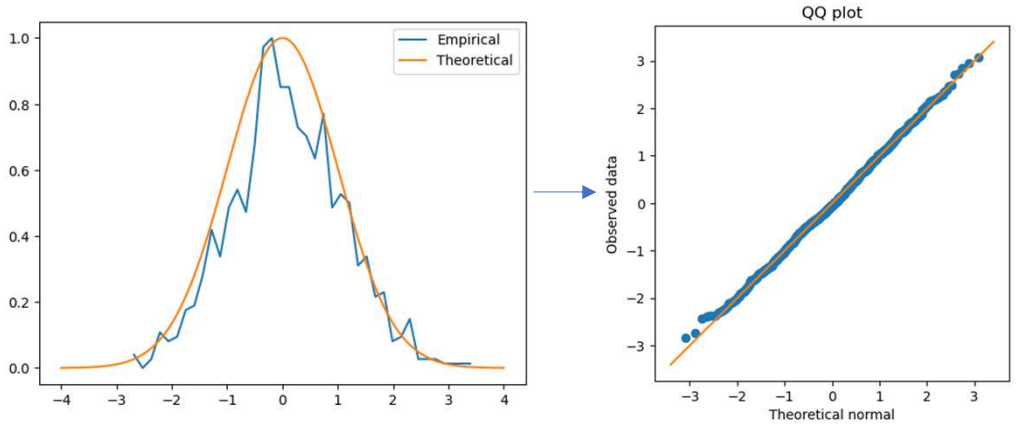
Variance Inflation Factor (VIF):To check this sort of relations between variables, we use VIF. VIF basically helps explaining the relationship of one independent variable with all the other independent variables.

If all the independent variables are orthogonal to each other,(R^2*=0) then VIF = 1.0. If there is perfect correlation between variable*(R^2*=1) , then VIF = infinity.*

$$VIF_i = \frac{1}{1-R_i^2}$$

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

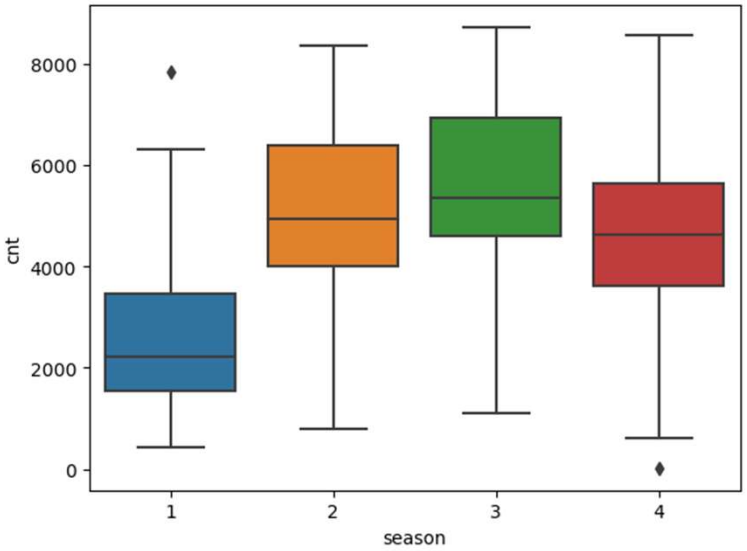
The quantile-quantile plot is a graphical method for determining whether two samples of data came from the same population or not. A q-q plot is a plot of the quantiles of the first data set against the quantiles of the second data set
QQ plots is very useful to determine
•This helps in a scenario of linear regression when we have training and test data set received separately and then we can confirm using Q-Q plot that both the data sets are from populations with same distributions.
•Normal Q-Q Plot: This is used to assess if your residuals are normally distributed. basically what you are looking for here is the data points closely following the straight line at a 45% angle upwards (left to right). Again what to watch here is any patterns that deviate from this - particularly anything that looks curvilinear (bending at either end) or s shaped.



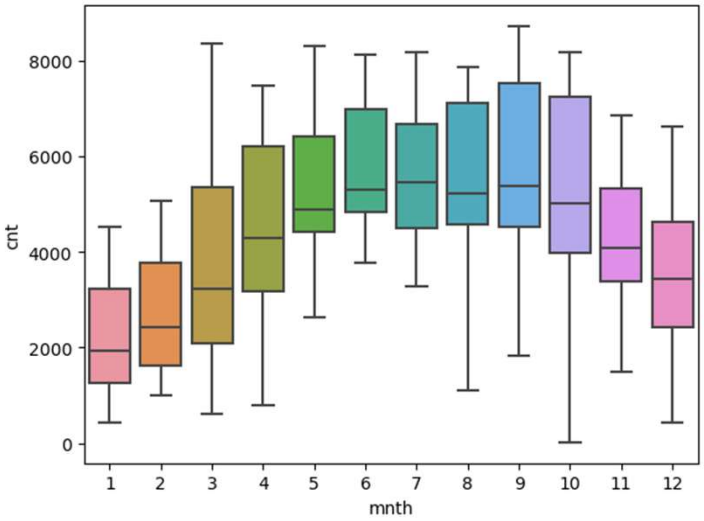
Assignment-based Subjective Questions

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

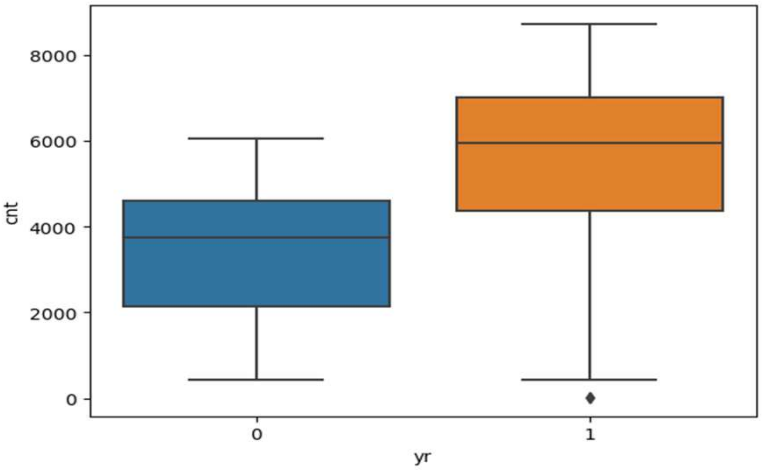
Rental bikes demand increases during summer and fall



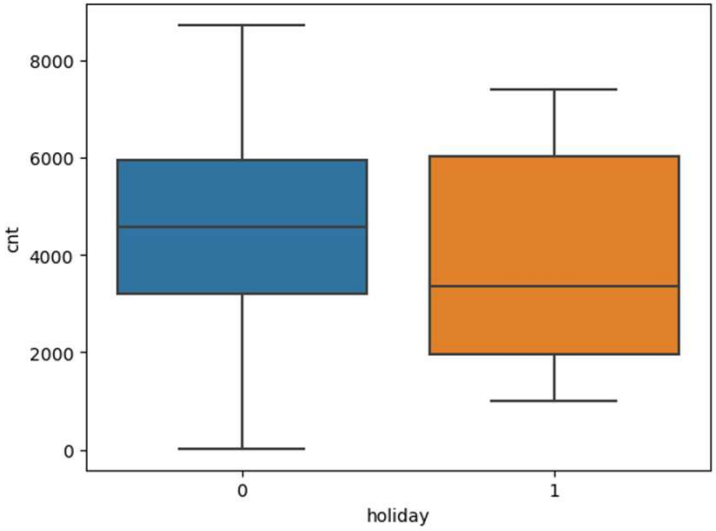
Rental bikes demand is high between May and oct months



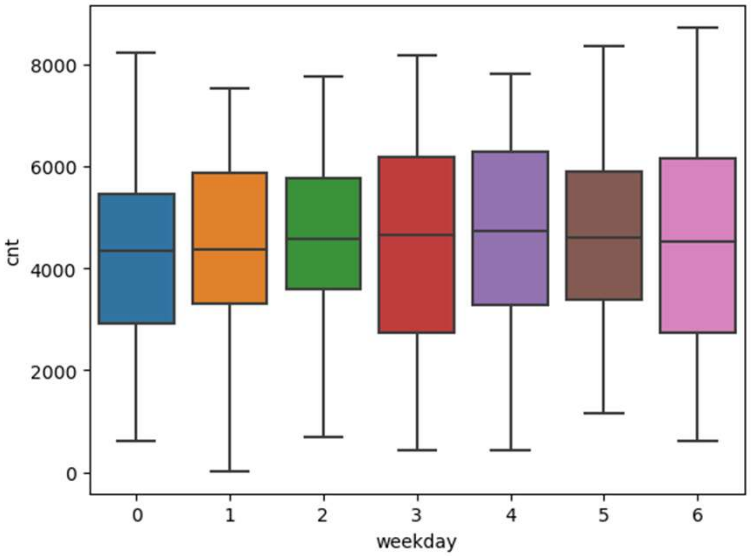
Rental bikes sales increases yr on yr bases



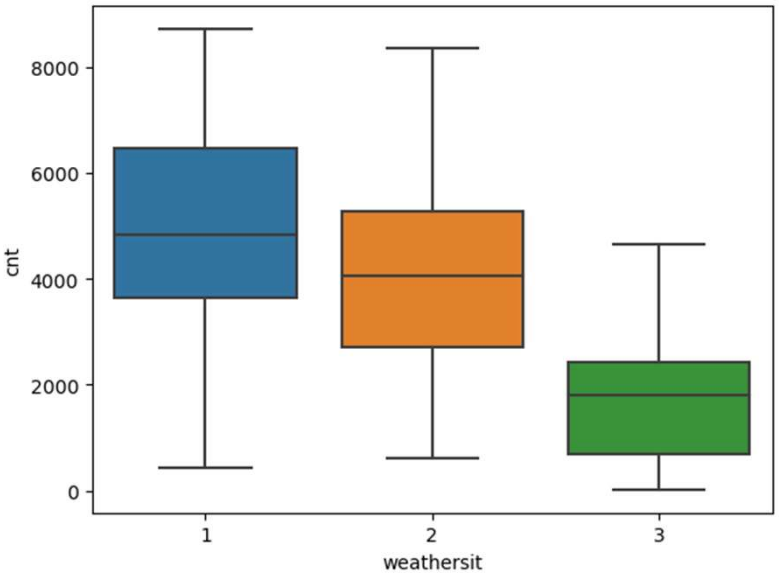
Rental bikes demand is high during holiday



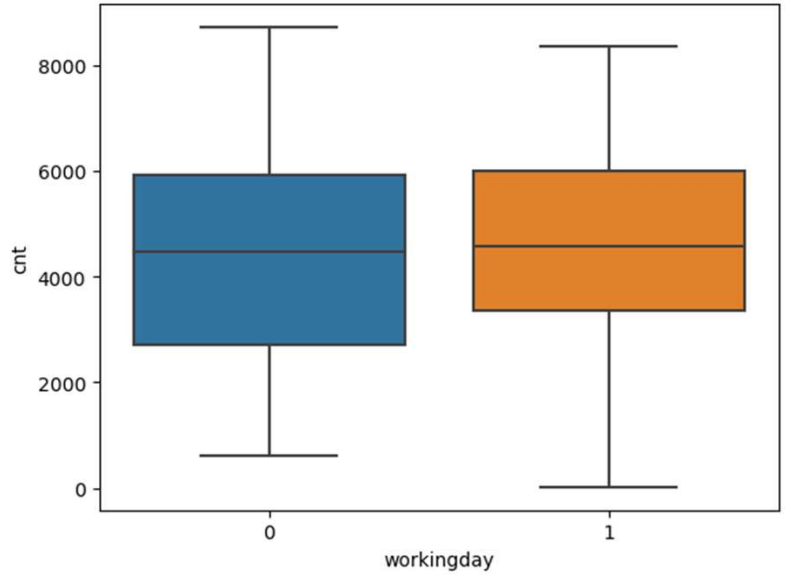
Rental bikes demand is almost similar throughout the weekdays.



Rental bikes demand is high if weather is Clear, Few clouds, Partly cloudy.



Rental bikes demand is doesn't change whether day is working day or not



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

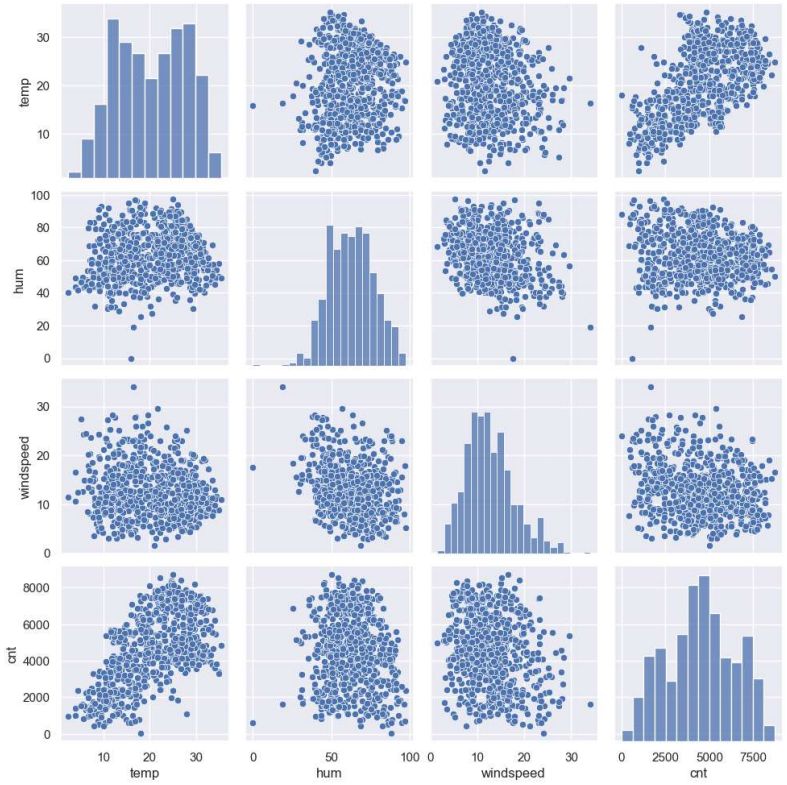
In python ,we convert categorical variables to dummies using pandas method get_dummies:

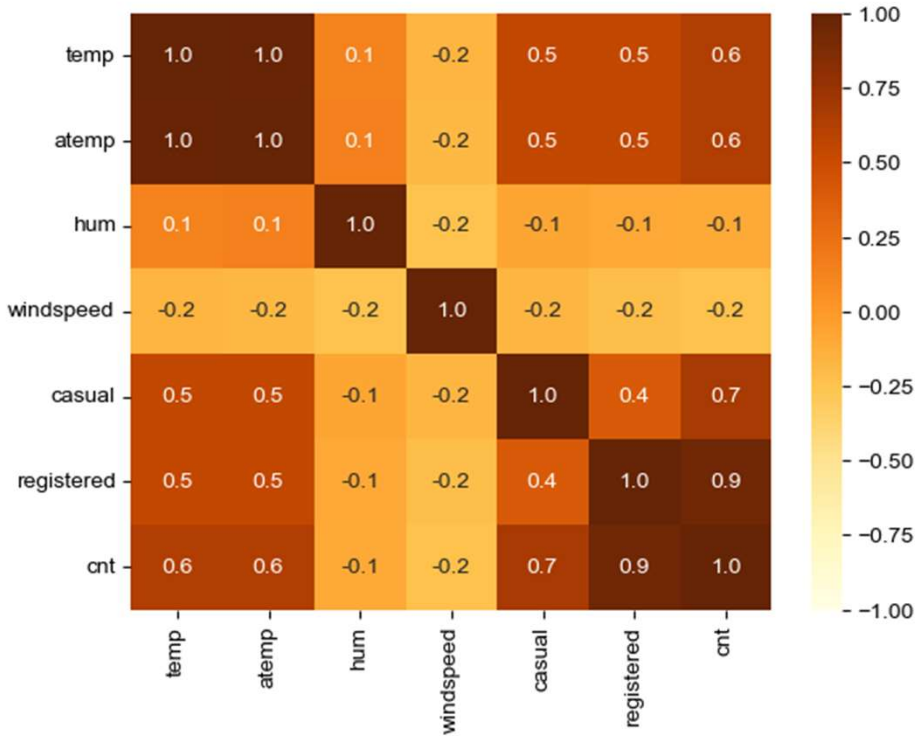
pd.get_dummies(data,drop_first=True)

The keyword argument drop_first will return P-1 columns. Use this to avoid the problem of multicollinearity. In the regression setting, a factor variable with P distinct levels is usually represented by a matrix with only P-1 columns.This is because a regression model typically includes an intercept term. With an intercept, once you have defined the values for P-1 binaries, the value of Pth is known and could be considered redundant . Adding the Pth column will cause a multicollinearity error.

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

From pair plot temperature having correlation with target variable(dropping atemp, Casual & registered columns)





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

Assumptions of Linear Regression

One of the most important assumptions is that a linear relationship is said to exist between the dependent and the independent variables

No auto-correlation or independence:

The residuals (error terms) are independent of each other. In other words, there is no correlation between the consecutive error terms of the time series data. The presence of correlation in the error terms drastically reduces the accuracy of the model. If the error terms are correlated, the estimated standard error tries to deflate the true standard error.
Measurement : Conduct a Durbin-Watson (DW) statistic test. The values should fall between 0-4. If DW=2, no auto-correlation

OLS Regression Results

Dep. Variable:	cnt	R-squared:	0.823
Model:	OLS	Adj. R-squared:	0.820
Method:	Least Squares	F-statistic:	242.0
Date:	Wed, 20 Dec 2023	Prob (F-statistic):	6.11e-207
Time:	21:43:45	Log-Likelihood:	549.04
No. Observations:	584	AIC:	-1074.
Df Residuals:	572	BIC:	-1022.
Df Model:	11		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	0.0699	0.019	3.759	0.000	0.033	0.106
temp	0.5486	0.019	28.477	0.000	0.511	0.586
yr	0.2297	0.008	28.925	0.000	0.214	0.245
season_winter	0.1169	0.012	9.886	0.000	0.094	0.140
weathersit_LightSnow	-0.2867	0.023	-12.425	0.000	-0.332	-0.241
weathersit_Mist	-0.0770	0.008	-9.073	0.000	-0.094	-0.060
season_summer	0.0934	0.010	9.272	0.000	0.074	0.113
mnth_sep	0.1022	0.015	6.800	0.000	0.073	0.132
windspeed	-0.1093	0.022	-4.945	0.000	-0.153	-0.066
mnth_oct	0.0553	0.017	3.339	0.001	0.023	0.088
weekday_Sat	0.0632	0.014	4.525	0.000	0.036	0.091
workingday	0.0516	0.011	4.907	0.000	0.031	0.072

Omnibus:	70.844	Durbin-Watson:	2.067
Prob(Omnibus):	0.000	Jarque-Bera (JB):	46.192
Skew:	-0.697	Prob(JB):	1.80e-32
Kurtosis:	5.016	Cond. No.	11.6

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

Durbin-Watson (DW)

No Multicollinearity :

Determine the VIF (Variance Inflation Factor). VIF<=5 implies no multicollinearity, whereas VIF>=10 implies serious multicollinearity.

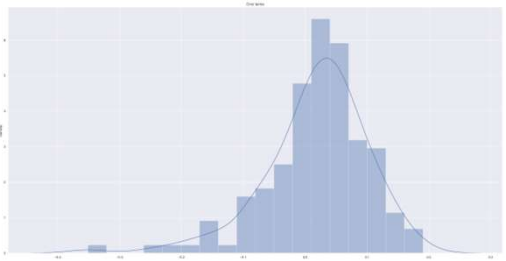
	Features	VIF
0	temp	5.15
7	windspeed	4.00
10	workingday	4.00
2	season_winter	2.03
1	yr	1.99
8	mnth_oct	1.67
5	season_summer	1.63
9	weekday_Sat	1.61
4	weathersit_Mist	1.54
6	mnth_sep	1.24
3	weathersit_LightSnow	1.11

Variance Inflation Factor

Homoscedasticity:

Homoscedasticity means the residuals have constant variance at every level of x. The absence of this phenomenon is known as heteroscedasticity
Create a scatter plot that shows residual vs fitted value

Normal distribution of error terms



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

Coefficient for the variables are:

```
const                   0.070
temp                    0.549
yr                      0.230
season_winter           0.117
weathersit_LightSnow   -0.287
weathersit_Mist        -0.077
season_summer           0.093
mnth_sep                0.102
windspeed              -0.109
mnth_oct                0.055
weekday_Sat             0.063
workingday              0.052
dtype: float64
```

Based on final model top three features contributing significantly towards explaining the demand are:

- *Temperature (0.549)*
- *weathersit : Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds(-0.287)*
- *year (0.230)*