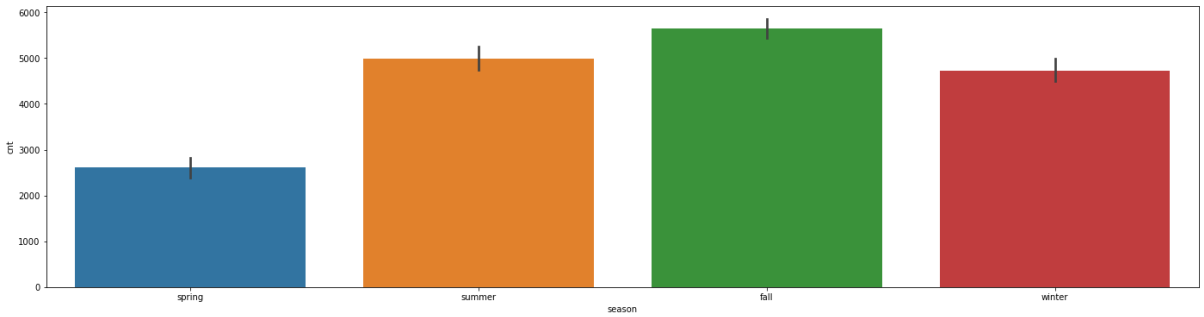
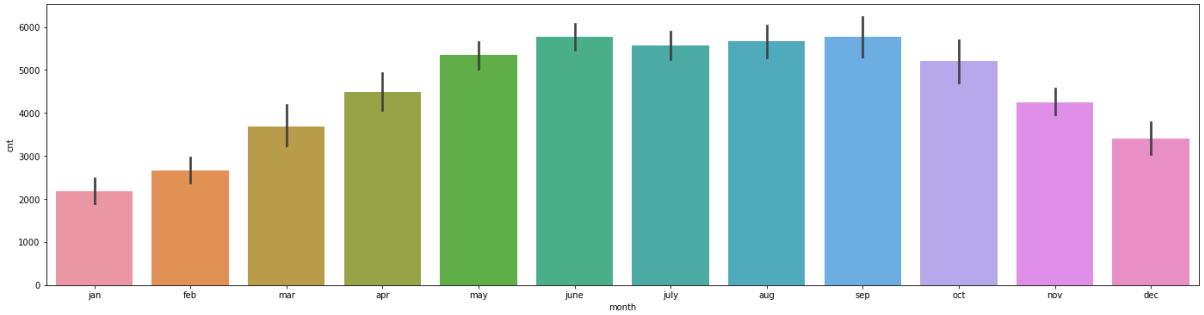
Assignment-based Subjective
Questions 1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?

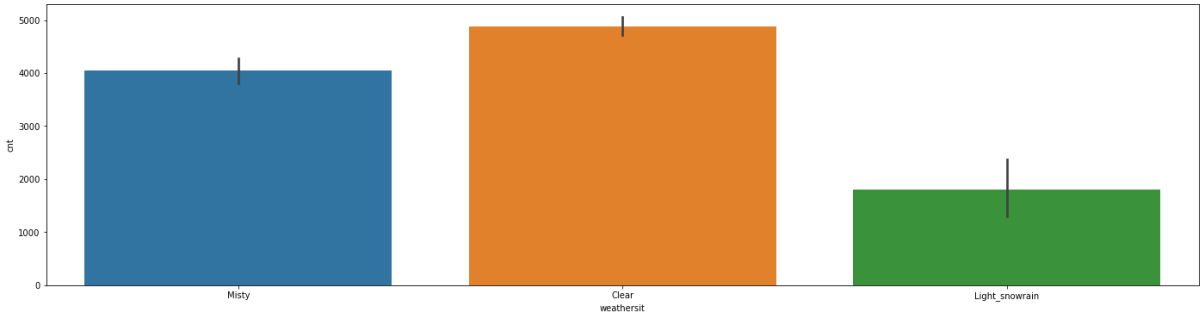
- Fall seasons seems to have more booking compared with other season.



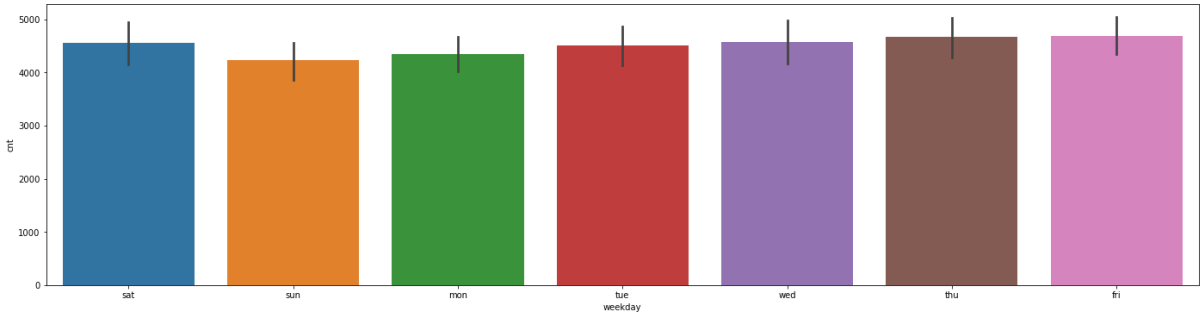
- Most of the bookings are done between month of May-Oct.



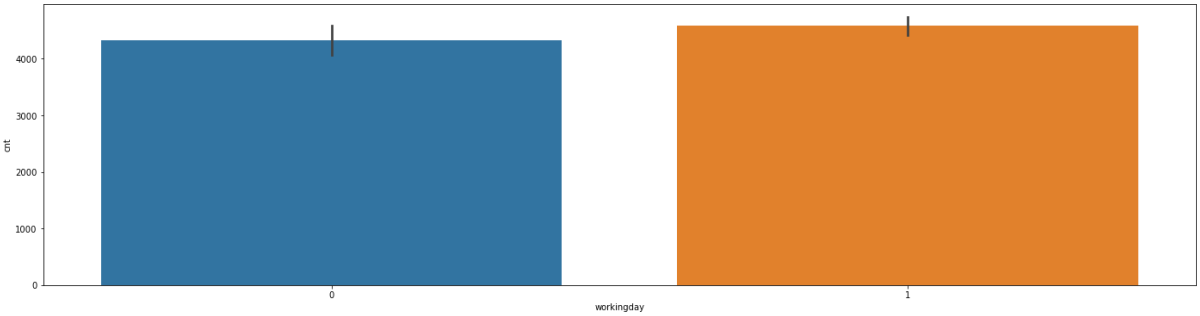
- Clear weather attracts more booking.



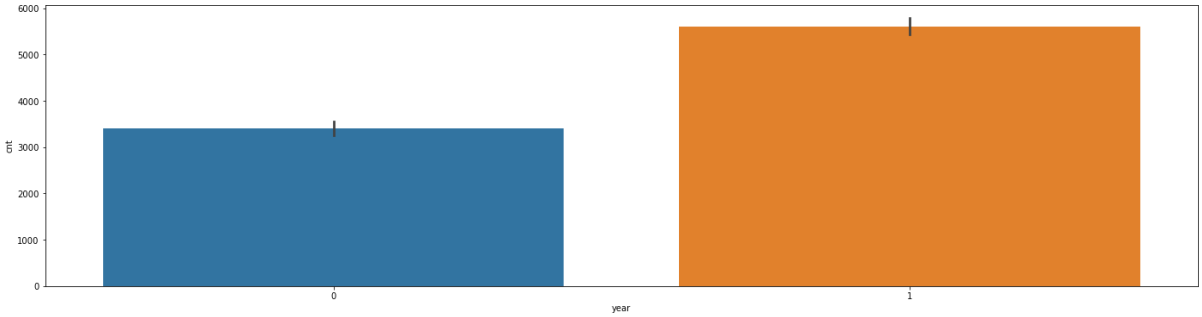
- Thu, Fri, Sat is more booking . Sunday is less , people want to be with family.



- Working and non working booking are almost same



- There are more booking in 2019 compared to previous, means the good progress in business.



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

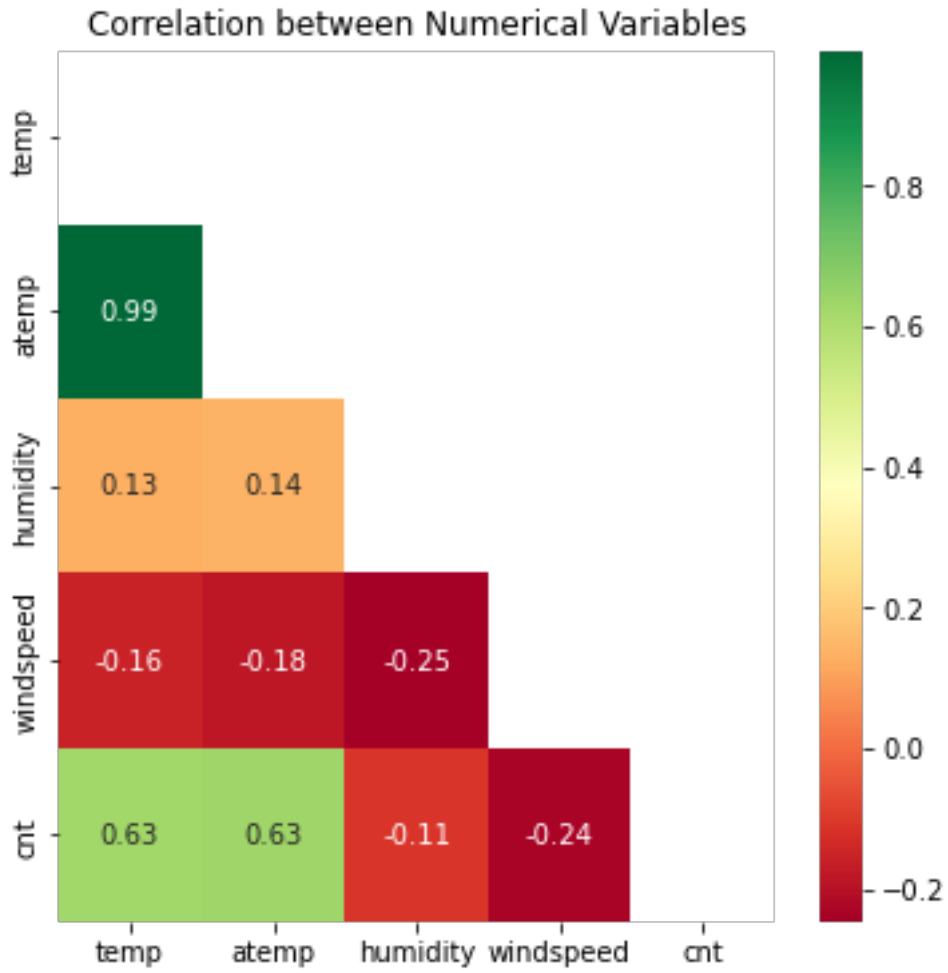
- drop_first=True is important to use, as it helps in reducing the extra column created during dummy variable creation. Hence it reduces the correlations created among dummy variables.
- Let's say we have 3 types of values in Categorical column and we want to create dummy variable for that column. If one variable is not furnished and semi_furnished, then It is obvious unfurnished. So we do not need 3rd variable to identify the unfurnished.

Value	Indicator Variable	
Furnishing Status	furnished	semi-furnished
furnished	1	0
semi-furnished	0	1
unfurnished	0	0

- Hence if we have categorical variable with n-levels, then we need to use n-1 columns to represent the dummy variables.

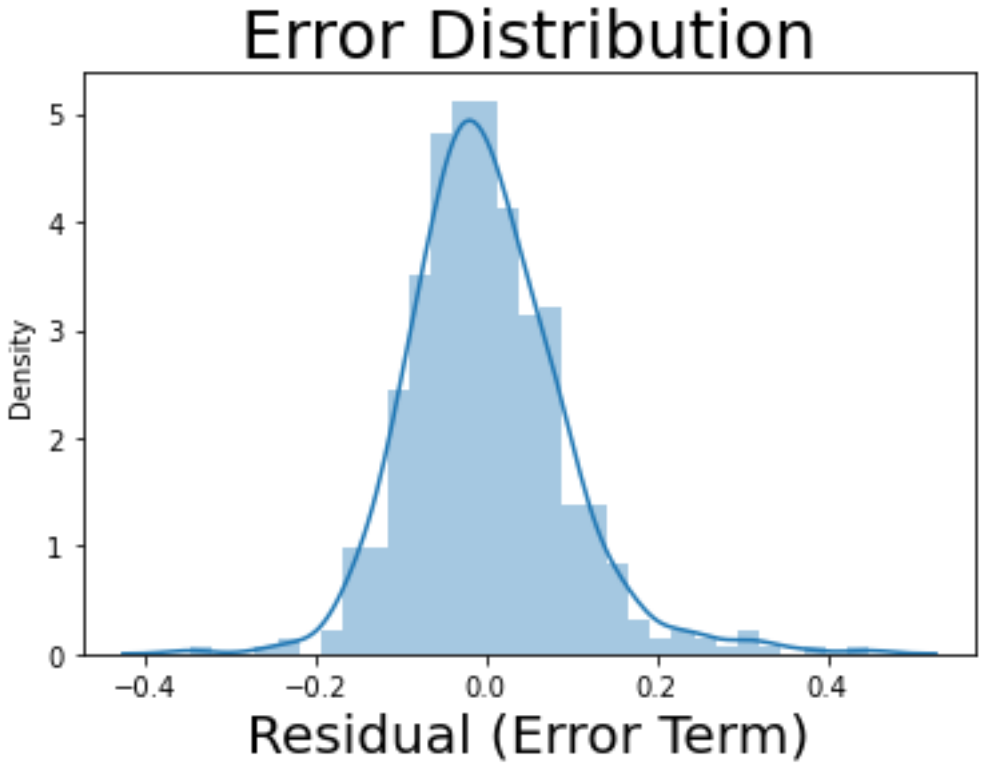
3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?

'temp' and 'atemp' variable has highest corelation.

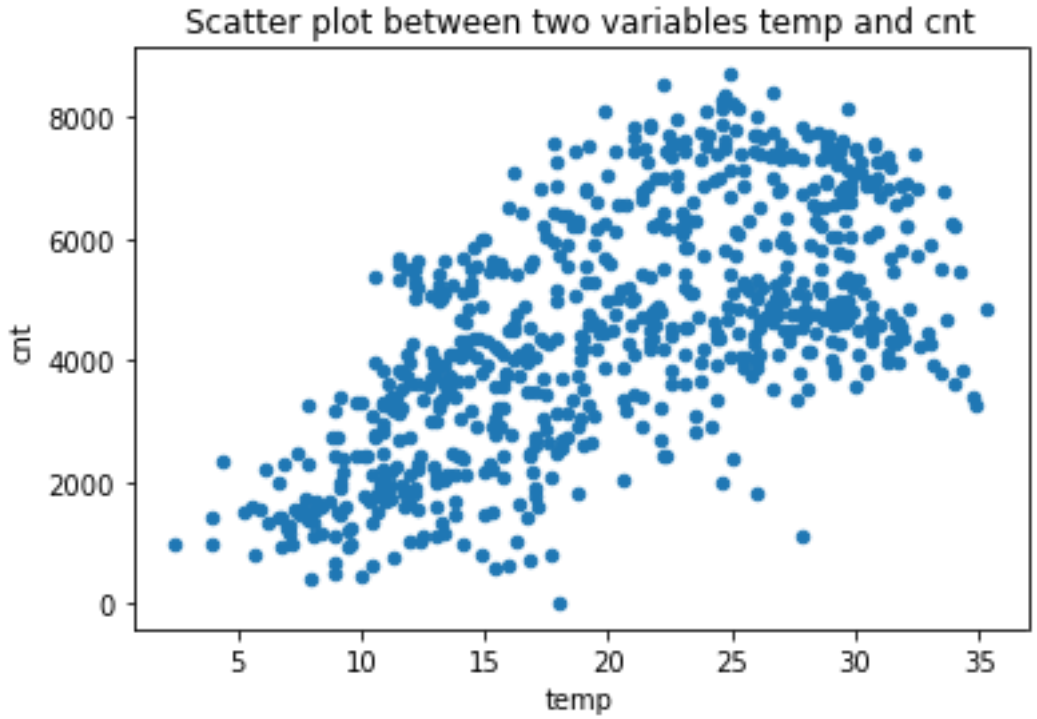


4. How did you validate the assumptions of Linear Regression after building the model on the training set?
I have validated using below assumption

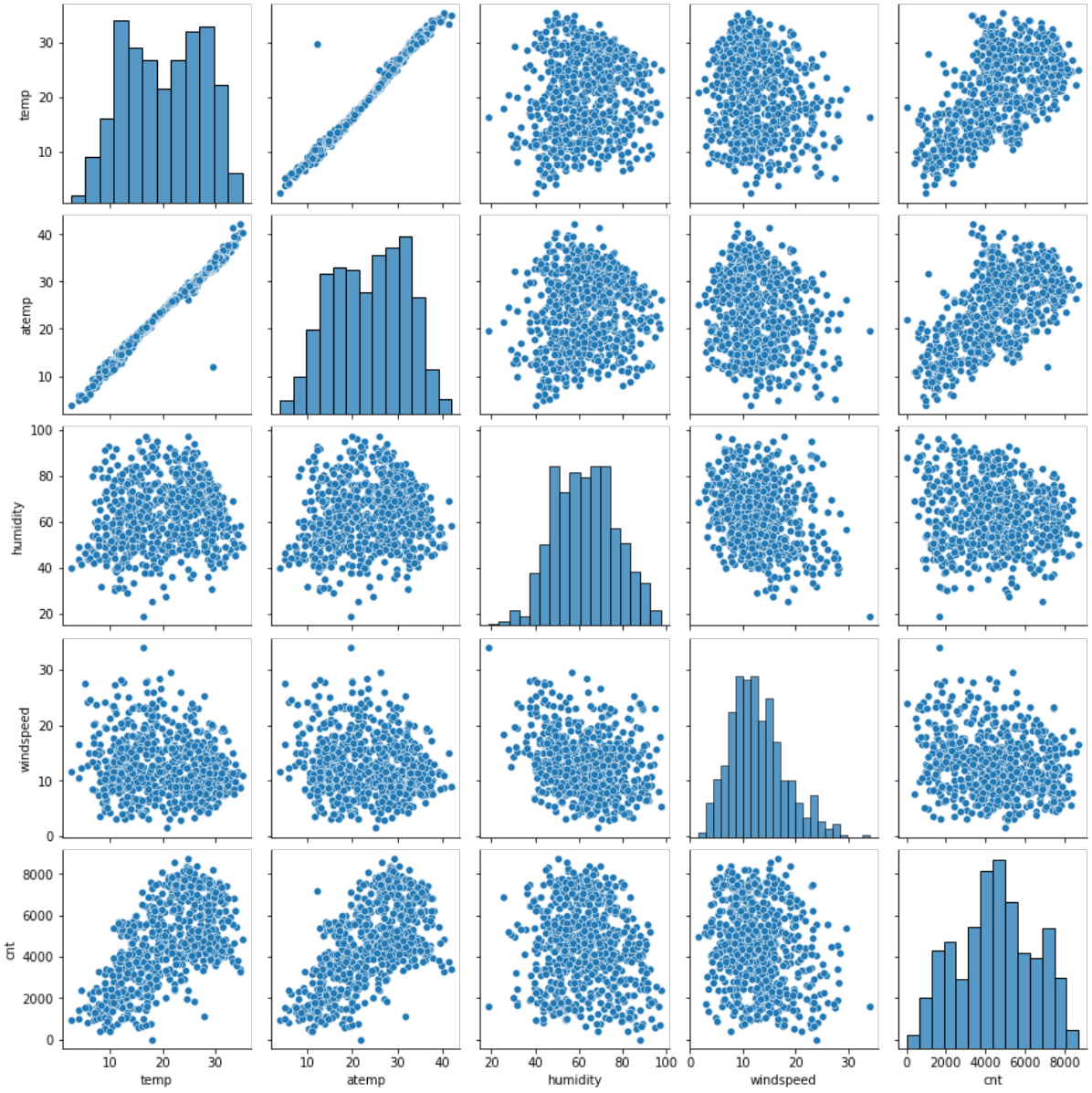
- Normal distribution of error terms: Here error is normal distributed



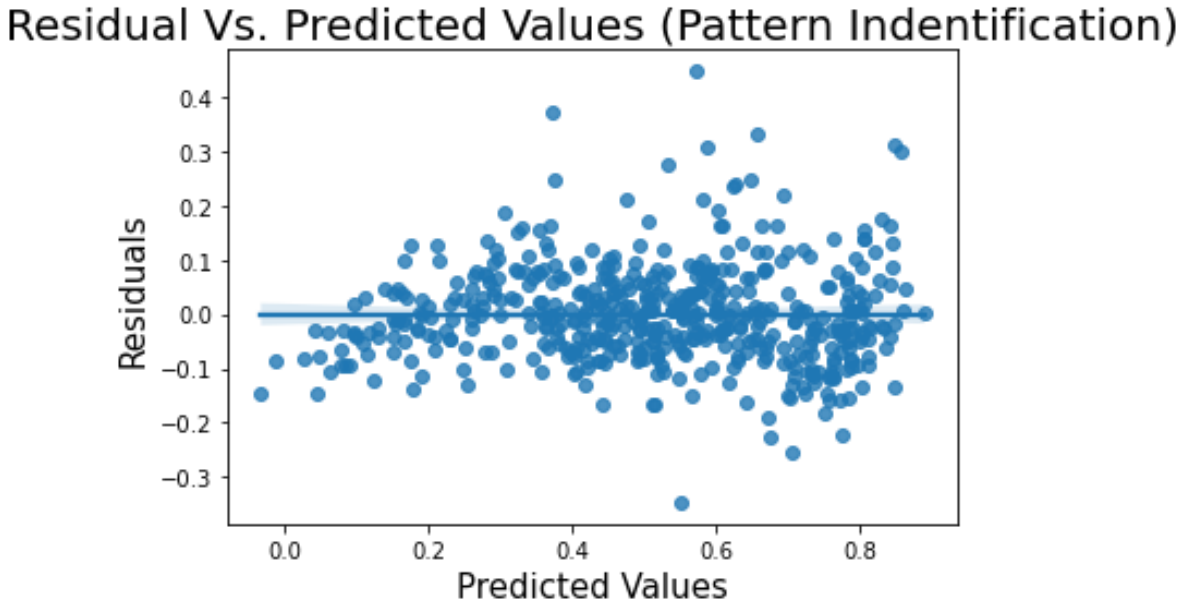
- Linear Relationship between the features and target.There is direct positive dependent on variables



- Little or no Multicollinearity between the features: no two variables are collinearity



- Homoscedasticity Assumption: There is no pattern and all values are equally scattered.



- Little or No autocorrelation in the residuals:

Durbin-Watson values approx. 2 which shpws thate there is no auto correlation in residual

OLS Regression Results

Dep. Variable:	cnt	R-squared:	0.837
Model:	OLS	Adj. R-squared:	0.832
Method:	Least Squares	F-statistic:	195.6
Date:	Sat, 12 Nov 2022	Prob (F-statistic):	1.23e-185
Time:	19:02:39	Log-Likelihood:	501.12
No. Observations:	510	AIC:	-974.2
Df Residuals:	496	BIC:	-915.0
Df Model:	13		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	0.2379	0.032	7.373	0.000	0.174	0.301
year	0.2350	0.008	28.586	0.000	0.219	0.251
holiday	-0.0907	0.026	-3.460	0.001	-0.142	-0.039
temp	0.4248	0.036	11.755	0.000	0.354	0.496
windspeed	-0.1591	0.025	-6.263	0.000	-0.209	-0.109
dec	-0.0434	0.018	-2.429	0.015	-0.078	-0.008
jan	-0.0522	0.018	-2.824	0.005	-0.089	-0.016
nov	-0.0393	0.019	-2.040	0.042	-0.077	-0.001
sep	0.0823	0.016	4.994	0.000	0.050	0.115
Light_snowrain	-0.2926	0.025	-11.803	0.000	-0.341	-0.244
Misty	-0.0787	0.009	-8.994	0.000	-0.096	-0.061
spring	-0.0597	0.021	-2.814	0.005	-0.101	-0.018
summer	0.0496	0.015	3.400	0.001	0.021	0.078
winter	0.0988	0.018	5.628	0.000	0.064	0.133

Omnibus:	73.264	Durbin-Watson:	2.057
Prob(Omnibus):	0.000	Jarque-Bera (JB):	179.059
Skew:	-0.742	Prob(JB):	1.31e-39
Kurtosis:	5.495	Cond. No.	18.8

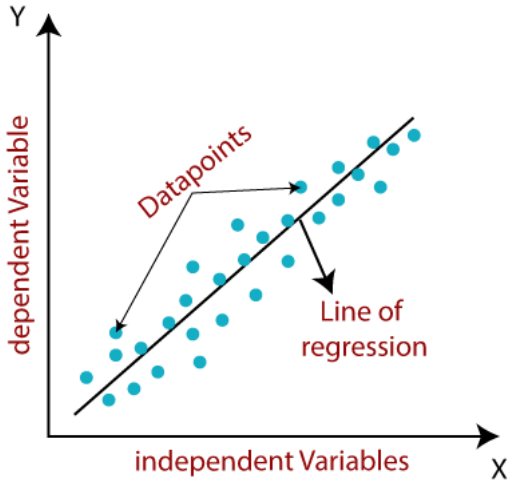
5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (2 marks)

- Temp
- Sep
- Winter

General Subjective Questions

1. Explain the linear regression algorithm in detail. (4 marks)

- Linear regression algorithm shows a linear relationship between a dependent (y) and one or more independent (y) variables, hence called as linear regression. Since linear regression shows the linear relationship, which means it finds how the value of the dependent variable is changing according to the value of the independent variable.
- The linear regression model provides a sloped straight line representing the relationship between the variables. Consider the below image:



Mathematically, we can represent a linear regression as:

y= a0+a1x

Y= Dependent Variable (Target Variable)

X= Independent Variable (predictor Variable)

a0= intercept of the line (Gives an additional degree of freedom)

a1 = Linear regression coefficient (scale factor to each input value).

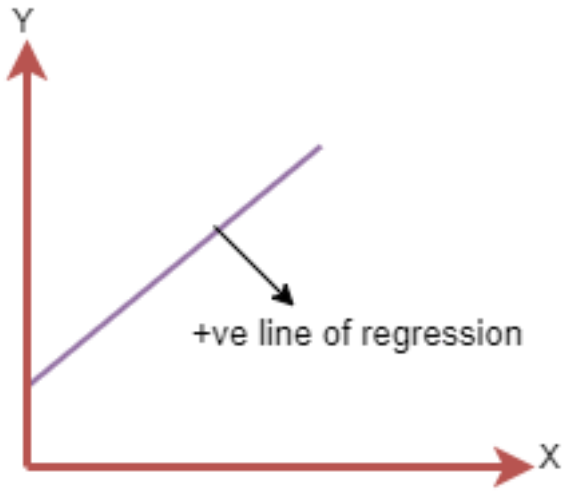
Types of Linear Regression

Linear regression can be further divided into two types of the algorithm:

- **Simple Linear Regression:**
  If a single independent variable is used to predict the value of a numerical dependent variable, then such a Linear Regression algorithm is called Simple Linear Regression.

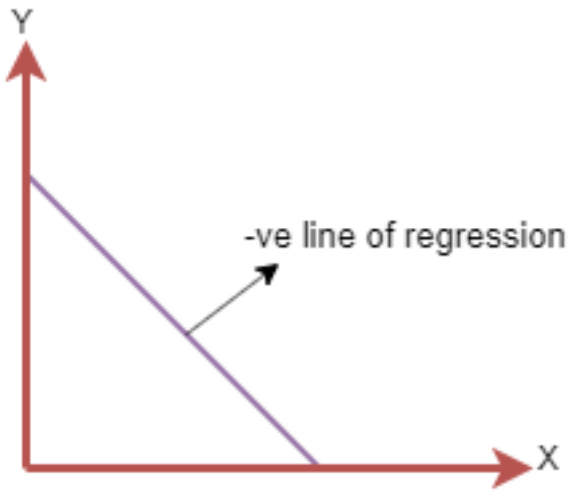
- **Multiple Linear regression:**
  If more than one independent variable is used to predict the value of a numerical dependent variable, then such a Linear Regression algorithm is called Multiple Linear Regression.

- **Positive Linear Relationship:**
  If the dependent variable increases on the Y-axis and independent variable increases on X-axis, then such a relationship is termed as a Positive linear relationship.



The line equation will be: $Y= a_0+a_1x$

- **Negative Linear Relationship:**
  If the dependent variable decreases on the Y-axis and independent variable increases on the X-axis, then such a relationship is called a negative linear relationship.

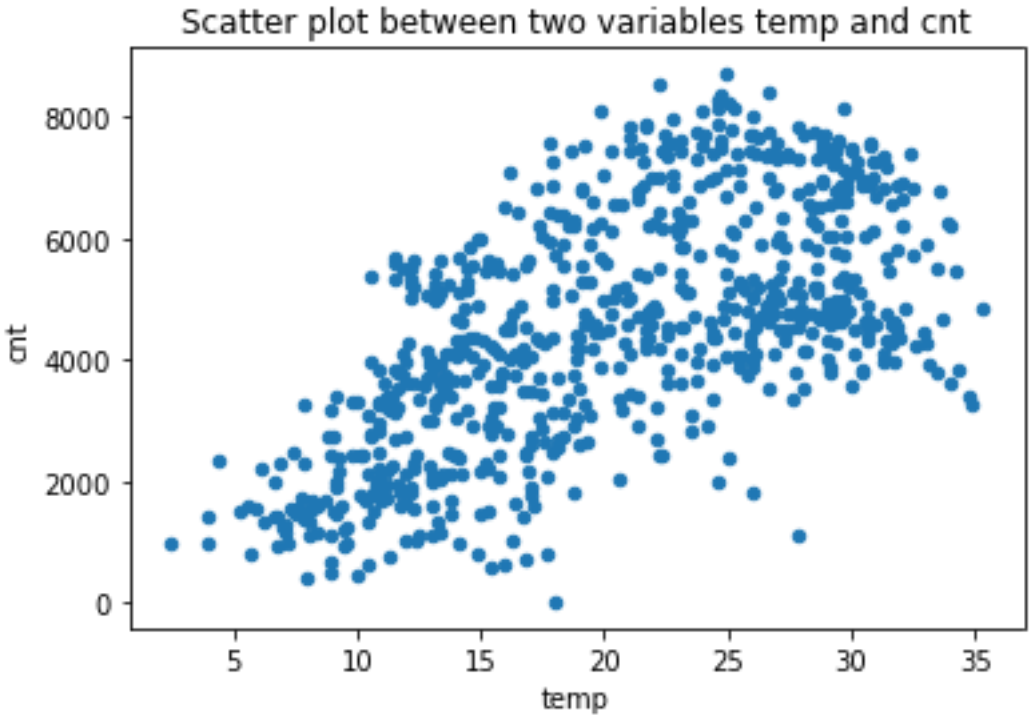


The line of equation will be: $Y= -a_0+a_1x$

Assumptions of Linear Regression

1. **Linear Relationship between the features and target:**

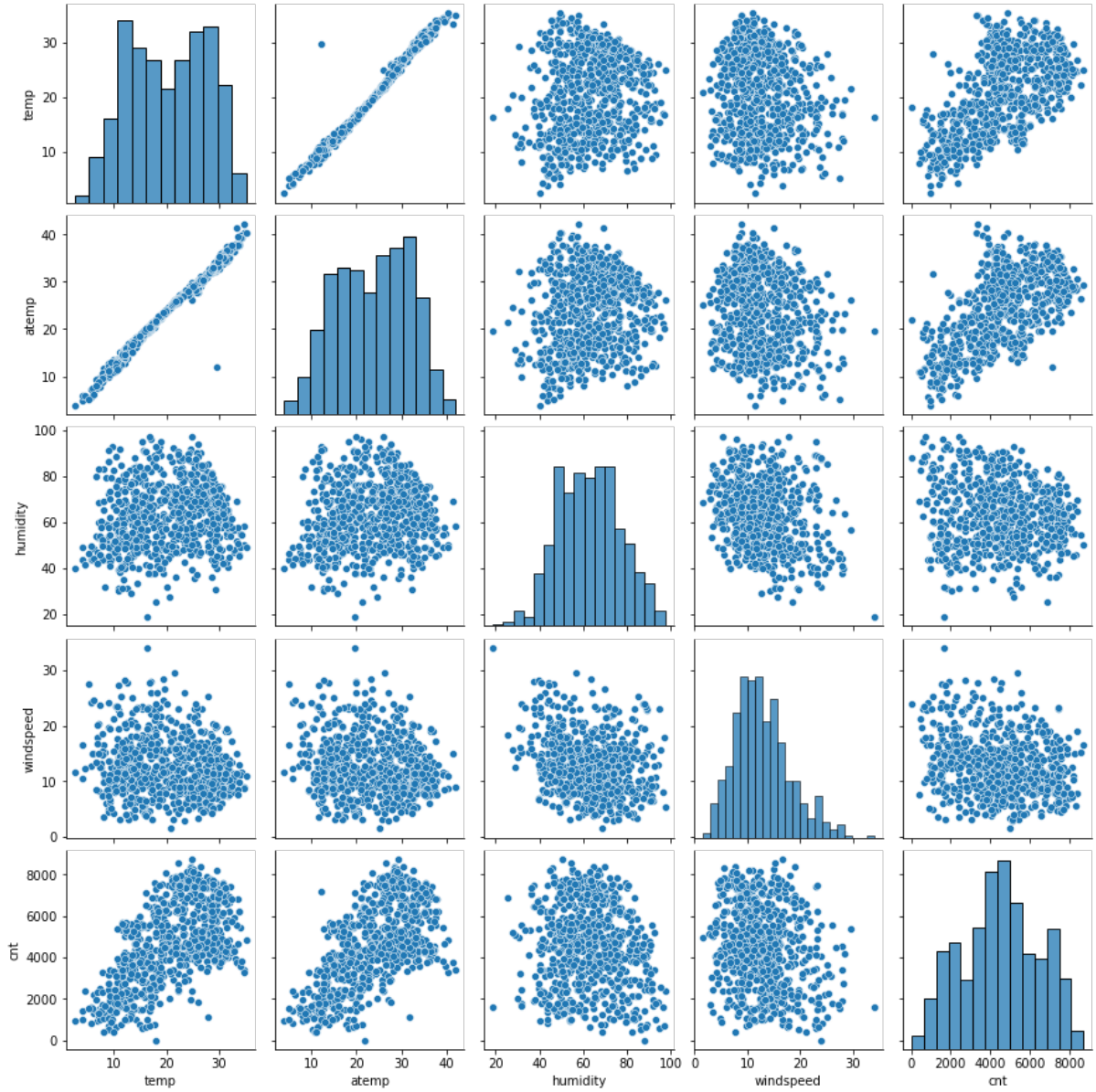
According to this assumption there is linear relationship between the features and target. Linear regression captures only linear relationship. This can be validated by plotting a scatter plot between the features and the target

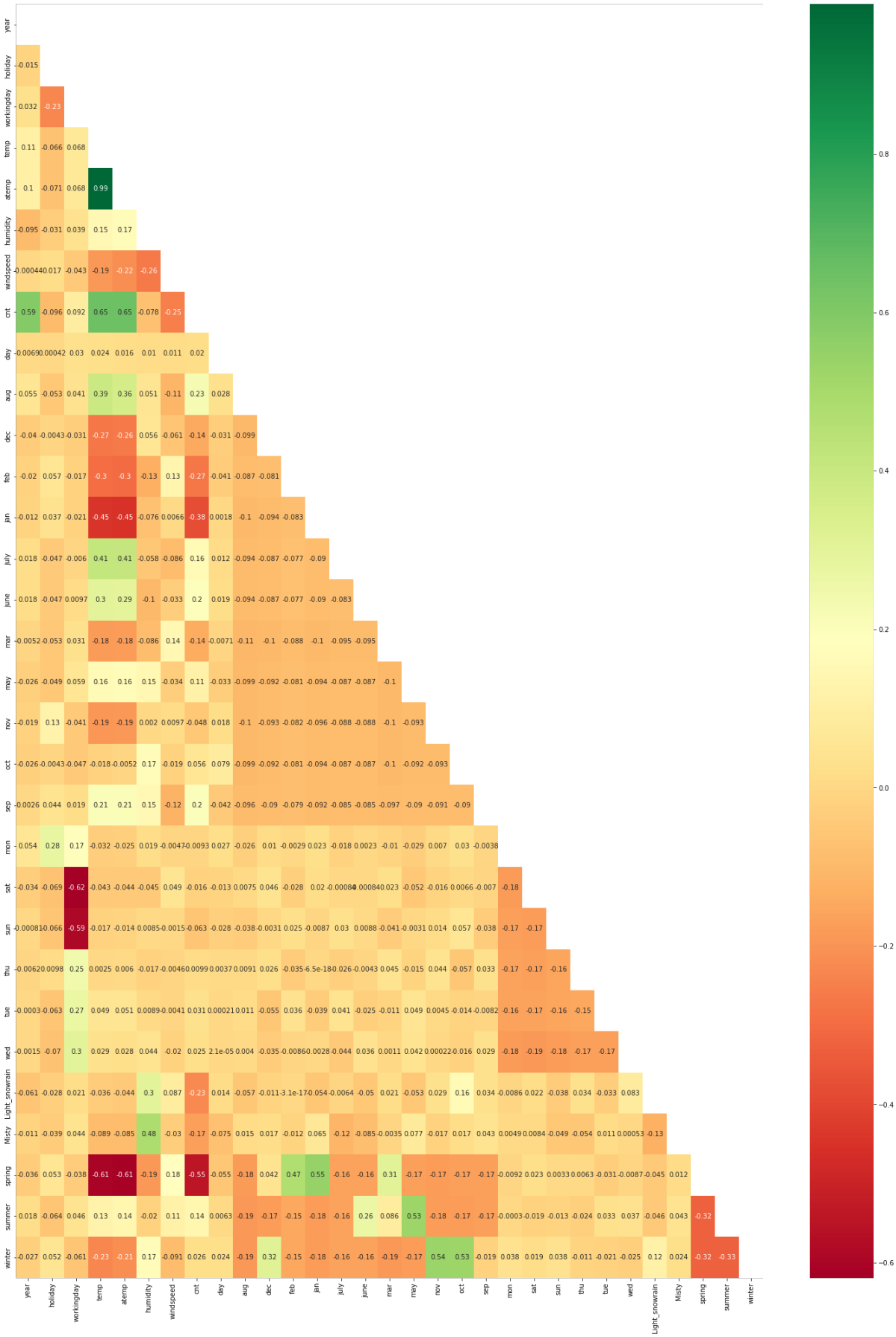


scatter plot of the feature temp vs cnt tells us that as the temp is increased number of booking also increased linearly.

2. Little or no Multicollinearity between the features:

Multicollinearity is a state of very high inter-correlations or inter-associations among the independent variables. It is therefore a type of disturbance in the data if present weakens the statistical power of the regression model. Pair plots and heatmaps(correlation matrix) can be used for identifying highly correlated features.



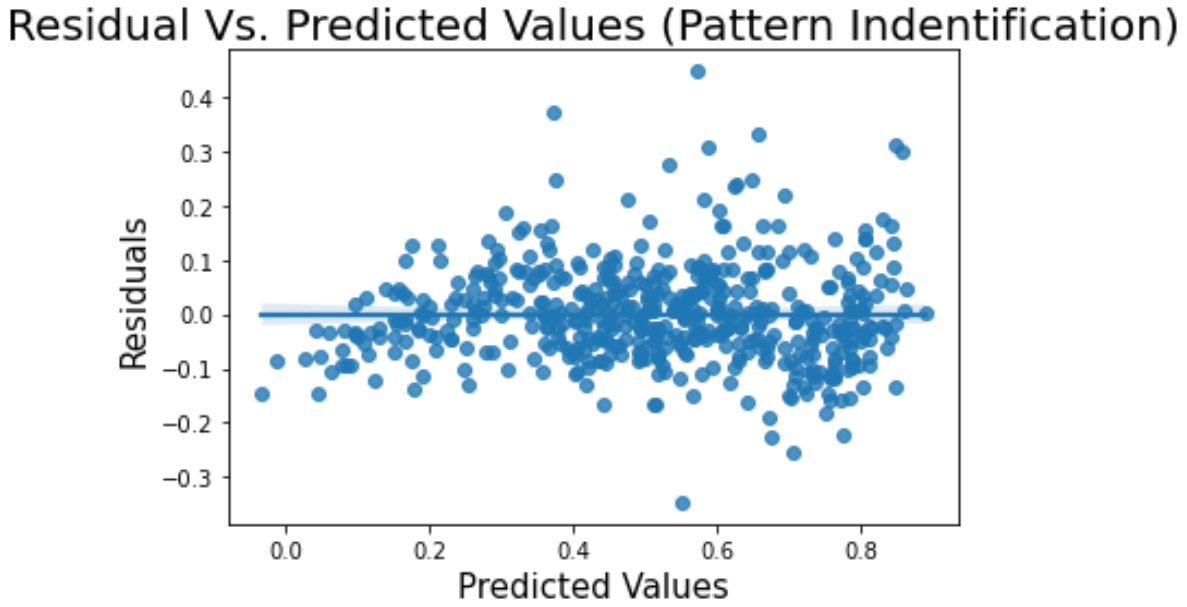


As temp and atemp is highly corelated, we only considered one variable.

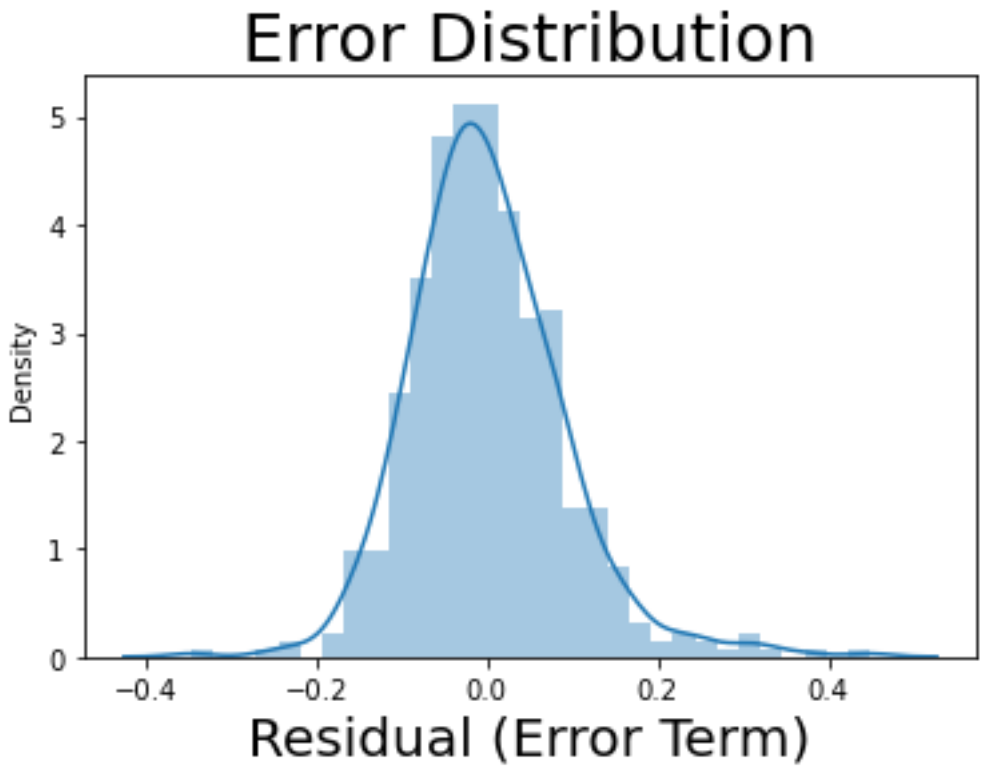
The interpretation of a regression coefficient is that it represents the mean change in the target for each unit change in an feature when you hold all of the other features constant. However, when features are correlated, changes in one feature in turn shifts another feature/features. The stronger the correlation, the more difficult it is to change one feature without changing another. It becomes difficult for the model to estimate the relationship between each feature and the target independently because the features tend to change in unison.

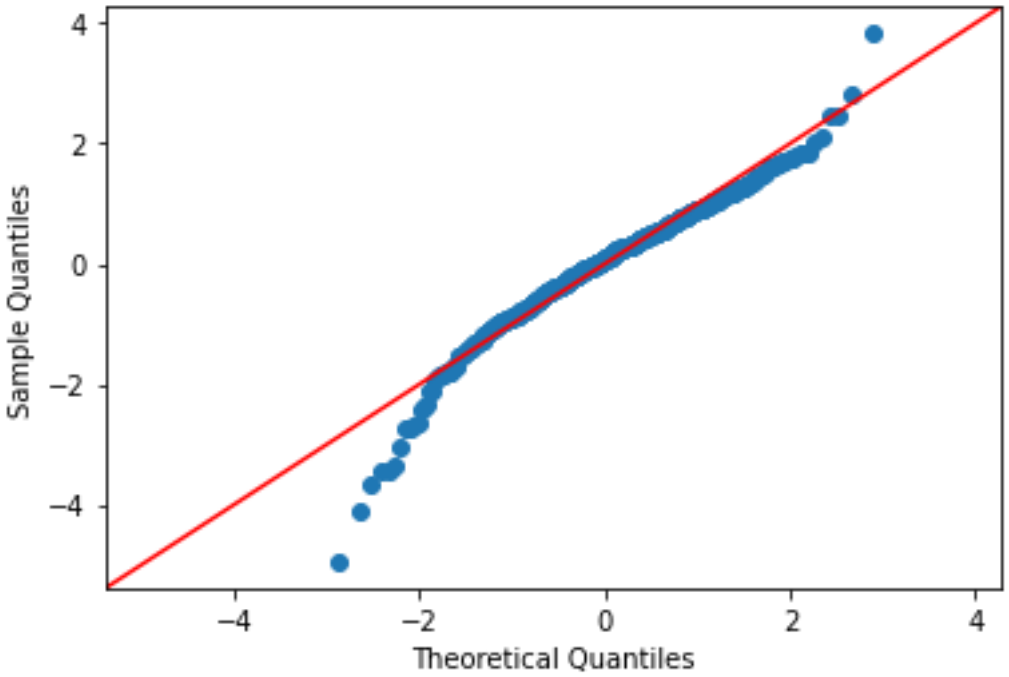
Homoscedasticity Assumption:

Homoscedasticity describes a situation in which the error term (that is, the "noise" or random disturbance in the relationship between the features and the target) is the same across all values of the independent variables. A scatter plot of residual values vs predicted values is a goodway to check for homoscedasticity. There should be no clear pattern in the distribution



Normal distribution of error terms:





The q-q plot of the advertising data set shows that the errors(residuals) are fairly normally distributed.The histogram plot in the “Error(residuals) vs Predicted values” in assumption .also shows that the errors are normally distributed with mean close to 0.

Little or No autocorrelation in the residuals:

Autocorrelation occurs when the residual errors are dependent on each other. The presence of correlation in error terms drastically reduces model's accuracy. This usually occurs in time series models where the next instant is dependent on previous instant.
Autocorrelation can be tested with the help of Durbin-Watson test.The null hypothesis of the test is that there is no serial correlation. The Durbin-Watson test statistics is defined as:

$$\sum_{t=2}^{T}((e_t - e_{t-1})^2) / \sum_{t=1}^{T} e_t^2$$

The test statistic is approximately equal to 2*(1-r) where r is the sample autocorrelation of the residuals. Thus, for r == 0, indicating no serial correlation, the test statistic equals 2. This statistic will always be between 0 and 4. The closer to 0 the statistic, the more evidence for positive serial correlation. The closer to 4, the more evidence for negative serial correlation.

OLS Regression Results

Dep. Variable:	cnt	**R-squared:**	0.837
Model:	OLS	**Adj. R-squared:**	0.832
Method:	Least Squares	**F-statistic:**	195.6
Date:	Sat, 12 Nov 2022	**Prob (F-statistic):**	1.23e-185
Time:	15:51:35	**Log-Likelihood:**	501.12
No. Observations:	510	**AIC:**	-974.2
Df Residuals:	496	**BIC:**	-915.0
Df Model:	13		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	0.2379	0.032	7.373	0.000	0.174	0.301
year	0.2350	0.008	28.586	0.000	0.219	0.251
holiday	-0.0907	0.026	-3.460	0.001	-0.142	-0.039
temp	0.4248	0.036	11.755	0.000	0.354	0.496
windspeed	-0.1591	0.025	-6.263	0.000	-0.209	-0.109
dec	-0.0434	0.018	-2.429	0.015	-0.078	-0.008
jan	-0.0522	0.018	-2.824	0.005	-0.089	-0.016
nov	-0.0393	0.019	-2.040	0.042	-0.077	-0.001
sep	0.0823	0.016	4.994	0.000	0.050	0.115
Light_snowrain	-0.2926	0.025	-11.803	0.000	-0.341	-0.244
Misty	-0.0787	0.009	-8.994	0.000	-0.096	-0.061
spring	-0.0597	0.021	-2.814	0.005	-0.101	-0.018
summer	0.0496	0.015	3.400	0.001	0.021	0.078
winter	0.0988	0.018	5.628	0.000	0.064	0.133

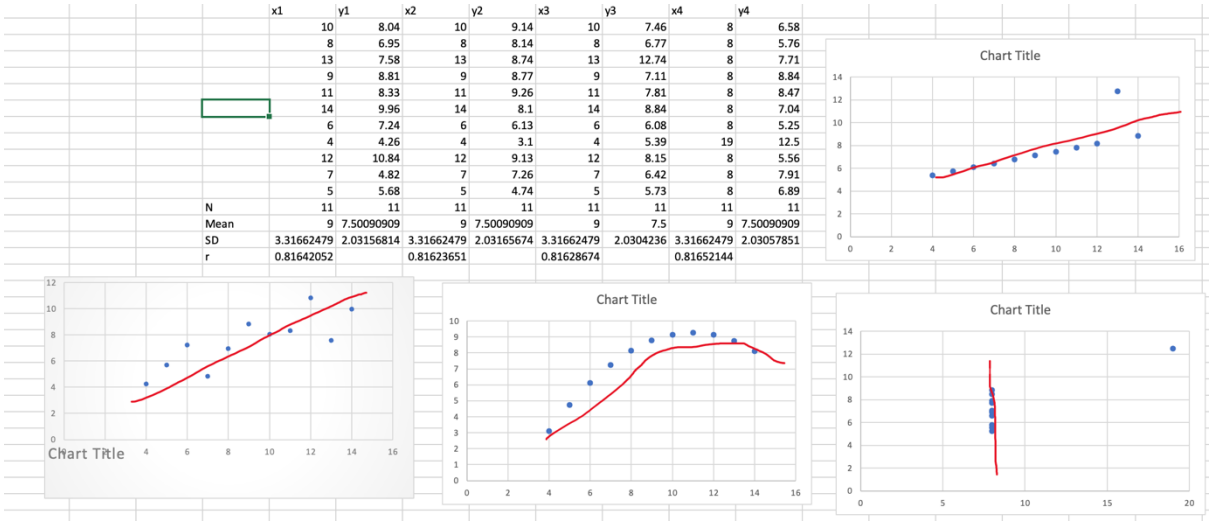
Omnibus:	73.264	**Durbin-Watson:**	2.057
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	179.059
Skew:	-0.742	**Prob(JB):**	1.31e-39
Kurtosis:	5.495	**Cond. No.**	18.8

From the above summary note that the value of Durbin-Watson test is 1.885 quite close to 2 as said before when the value of Durbin-Watson is equal to 2, r takes the value 0 from the equation 2*(1-r),which in turn tells us that the residuals are not correlated.

2. Explain the Anscombe's quartet in detail. (3 marks)

Anscombe's Quartet can be defined as a group of four data sets which are nearly identical in simple descriptive statistics, but there are some peculiarities in the dataset that fools the regression model if built. They have very different distributions and appear differently when plotted on scatter plots.
It was constructed in 1973 by statistician Francis Anscombe to illustrate
the importance of plotting the graphs before analyzing and model building, and the effect of other observations on statistical properties.There are these four data set plots which have nearly same statistical observations, which provides same statistical information that involves variance, and mean of all x,y points in all four datasets.
This tells us about the importance of visualising the data before applying various algorithms out there to build models out of them which suggests that the data features must be plotted in order to see the distribution of the samples that can help you identify the various anomalies present in the data like outliers, diversity of the data, linear separability of the data, etc. Also, the Linear Regression can be only be considered a fit for the data with linear relationships and is incapable of handling any other kind of datasets. These four plots can be defined as follows:

The statistical information for all these four datasets are approximately similar and can be computed as follows:



	x1	y1	x2	y2	x3	y3	x4	y4
	10	8.04	10	9.14	10	7.46	8	6.58
	8	6.95	8	8.14	8	6.77	8	5.76
	13	7.58	13	8.74	13	12.74	8	7.71
	9	8.81	9	8.77	9	7.11	8	8.84
	11	8.33	11	9.26	11	7.81	8	8.47
	14	9.96	14	8.1	14	8.84	8	7.04
	6	7.24	6	6.13	6	6.08	8	5.25
	4	4.26	4	3.1	4	5.39	19	12.5
	12	10.84	12	9.13	12	8.15	8	5.56
	7	4.82	7	7.26	7	6.42	8	7.91
	5	5.68	5	4.74	5	5.73	8	6.89
N	11	11	11	11	11	11	11	11
Mean	9	7.50090909	9	7.50090909	9	7.5	9	7.50090909
SD	3.31662479	2.03156814	3.31662479	2.03165674	3.31662479	2.0304236	3.31662479	2.03057851
r	0.81642052		0.81623651		0.81628674		0.81652144	

The four datasets can be described as:
Dataset 1: this fits the linear regression model pretty well.
Dataset 2: this could not fit linear regression model on the data quite well as the data is non-linear.
Dataset 3: shows the outliers involved in the dataset which cannot be handled by linear regression model
Dataset 4: shows the outliers involved in the dataset which cannot be handled by linear regression model

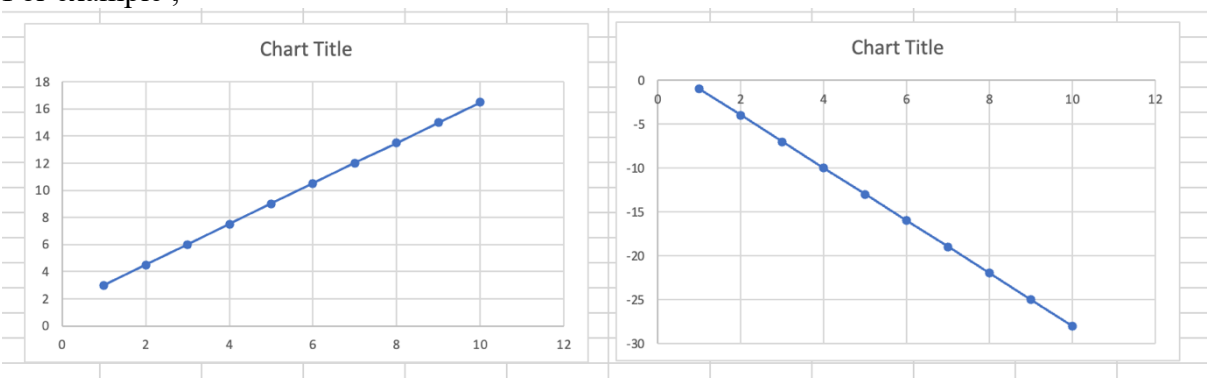
Conclusion:
Here the four datasets that were intentionally created to describe the importance of data visualisation and how any regression algorithm can be fooled by the same. Hence, all the important features in the dataset must be visualised before implementing any machine learning algorithm on them which will help to make a good fit model.

3. What is Pearson's R? (3 marks)

The Pearson correlation coefficient (r) is the most common way of measuring a linear correlation. It is a number between –1 and 1 that measures the strength and direction of the relationship between two variables.

Pearson correlation coefficient (*r*)	Correlation type	Interpretation	Example
Between 0 and 1	Positive correlation	When one variable changes, the other variable changes in the **same direction**.	Baby length & weight: The longer the baby, the heavier their weight.
0	No correlation	There is **no relationship** between the variables.	Car price & width of windshield wipers: The price of a car is not related to the width of its windshield wipers.
Between 0 and –1	Negative correlation	When one variable changes, the other variable changes in the **opposite direction**.	Elevation & air pressure: The higher the elevation, the lower the air pressure.

For example ,



Here first fig, if x increased then y will also increased but in fig 2 if x increased positively then y will decrease (negative slope)
4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?

Normalization typically means rescales the values into a range of [0,1]. Standardization typically means rescales data to have a mean of 0 and a standard deviation of 1 (unit variance).

	instant	season	yr	mnth	holiday	weekday	workingday	weathersit	temp	atemp	hum	windspeed	
count	730.000000	730.000000	730.000000	730.000000	730.000000	730.000000	730.000000	730.000000	730.000000	730.000000	730.000000	730.000000	73
mean	365.500000	2.498630	0.500000	6.526027	0.028767	2.997260	0.683562	1.394521	20.319259	23.726322	62.765175	12.763620	84
std	210.877136	1.110184	0.500343	3.450215	0.167266	2.006161	0.465405	0.544807	7.506729	8.150308	14.237589	5.195841	68
min	1.000000	1.000000	0.000000	1.000000	0.000000	0.000000	0.000000	1.000000	2.424346	3.953480	0.000000	1.500244	
25%	183.250000	2.000000	0.000000	4.000000	0.000000	1.000000	0.000000	1.000000	13.811885	16.889713	52.000000	9.041650	31
50%	365.500000	3.000000	0.500000	7.000000	0.000000	3.000000	1.000000	1.000000	20.465826	24.368225	62.625000	12.125325	71
75%	547.750000	3.000000	1.000000	10.000000	0.000000	5.000000	1.000000	2.000000	26.880615	30.445775	72.989575	15.625589	109
max	730.000000	4.000000	1.000000	12.000000	1.000000	6.000000	1.000000	3.000000	35.328347	42.044800	97.250000	34.000021	341

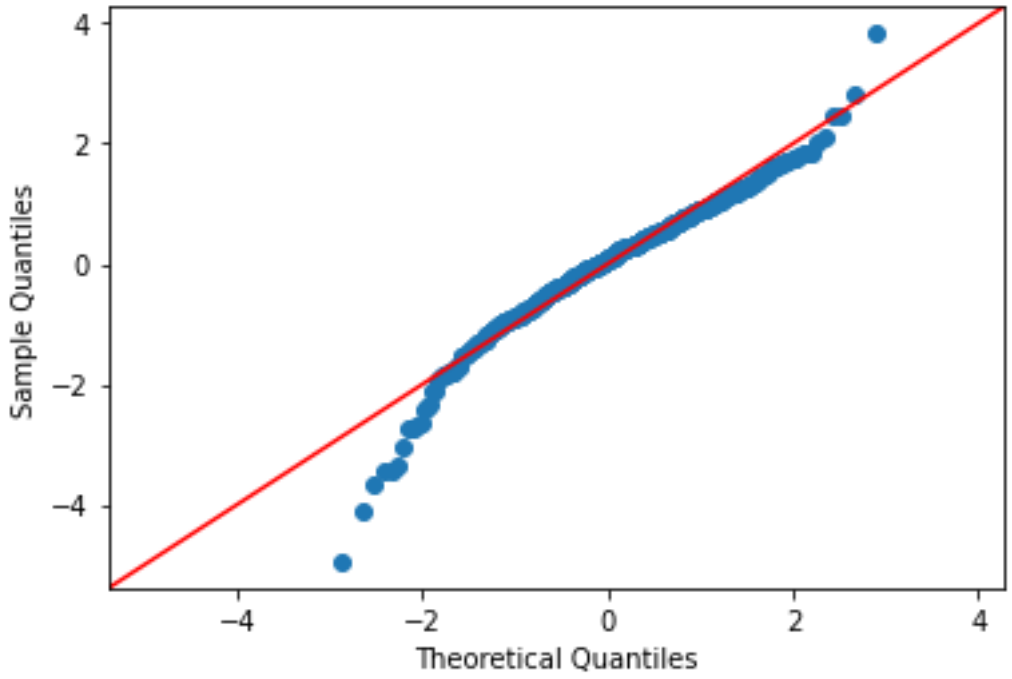
As we can observed, many independent variable has varying units and range, so scaling required independent variables to normalize the data within a particular range. It also helps in speeding up the calculations in an algorithm. Most of the times, collected data set contains features highly varying in magnitudes, units and range

S.NO.	Normalisation	Standardisation
1	Minimum and maximum value of features are used for scaling	Mean and standard deviation is used for scaling.
2	It is used when features are of different scales.	It is used when we want to ensure zero mean and unit standard deviation.
3	Scales values between [0, 1] or [-1, 1].	It is not bounded to a certain range.
4	It is really affected by outliers.	It is much less affected by outliers.
5	Scikit-Learn provides a transformer called `MinMaxScaler` for Normalization.	Scikit-Learn provides a transformer called `StandardScaler` for standardization.
6	This transformation squishes the n-dimensional data into an n-dimensional unit hypercube.	It translates the data to the mean vector of original data to the origin and squishes or expands.
7	It is useful when we don't know about the distribution	It is useful when the feature distribution is Normal or Gaussian.
8	It is a often called as Scaling Normalization	It is a often called as Z-Score Normalization.

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

If there is perfect correlation, then VIF = infinity. This shows a perfect correlation between two independent variables. In the case of perfect correlation, we get R2 =1, which lead to 1/(1-R2) infinity. To solve this problem we need to drop one of the variables from the dataset which is causing this perfect multicollinearity.
An infinite VIF value indicates that the corresponding variable may be expressed exactly by a linear combination of other variables.
In our data set, temp and atemp columns are best example.

6.What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression. (3 marks)
Q Q Plots (Quantile-Quantile plots) are plots of two quantiles against each other. A quantile is a fraction where certain values fall below that quantile. For example, the median is a quantile where 50% of the data fall below that point and 50% lie above it. The purpose of Q Q plots is to find out if two sets of data come from the same distribution. A 45 degree angle is plotted on the Q Q plot; if the two data sets come from a common distribution, the points will fall on that reference line.



The q-q plot of the advertising data set shows that the errors(residuals) are fairly normally distributed.The histogram plot in the “Error(residuals) vs Predicted values” in assumption .also shows that the errors are normally distributed with mean close to 0.